



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΕΦΑΡΜΟΣΜΕΝΗ ΟΙΚΟΝΟΜΙΚΗ ΜΕ ΚΑΤΕΥΘΥΝΣΗ ΤΗΝ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΑΝΑΛΥΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΜΠΕΙΡΙΚΗ ΔΙΕΡΕΥΝΗΣΗ ΜΕΤΡΩΝ ΣΥΣΧΕΤΙΣΗΣ ΜΕ ΤΕΧΝΙΚΕΣ
ΚΑΙ ΕΡΓΑΛΕΙΑ ΑΝΑΛΥΤΙΚΗΣ ΔΕΔΟΜΕΝΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟΣ ΦΟΙΤΗΤΗΣ: ΞΑΝΘΟΠΟΥΛΟΣ ΘΕΟΔΩΡΟΣ

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ: ΤΣΙΛΙΚΑ ΚΥΡΙΑΚΗ

Υποβλήθηκε ως αιτούμενο για την απόκτηση του μεταπτυχιακού
διπλώματος στην Εφαρμοσμένη Οικονομική.

ΙΟΥΝΙΟΣ 2025

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες σε όλους όσους με στήριξαν στην εκπόνηση της παρούσας διπλωματικής εργασίας. Ιδιαίτερες ευχαριστίες οφείλω στην επιβλέπουσα καθηγήτρια μου, κα. Κυριακή Τσιλίκα, για την αμέριστη καθοδήγηση, τη συνεχή υποστήριξη και τις πολύτιμες επιστημονικές συμβουλές της. Η άμεση ανταπόκριση της σε κάθε ερώτημα ή απορία μου, καθώς και η διάθεση της να με καθοδηγεί με σαφήνεια και υπομονή σε κάθε στάδιο της εργασίας, υπήρξαν καθοριστικοί παράγοντες για την ολοκλήρωση της παρούσας εργασίας. Επίσης, ευχαριστώ θερμά την οικογένεια μου για την αδιάκοπη στήριξη, την υπομονή και την ενθάρρυνση σε όλη τη διάρκεια των σπουδών μου. Τέλος, εκφράζω την ευγνωμοσύνη μου στους φίλους και συμφοιτητές μου για την ενθάρρυνση, την υποστήριξη και τη δημιουργική συνεργασία καθ' όλη τη διάρκεια αυτής της ακαδημαϊκής πορείας.

Πίνακας περιεχομένων

ΕΥΧΑΡΙΣΤΙΕΣ.....	2
Πίνακας περιεχομένων.....	3
ΠΕΡΙΛΗΨΗ.....	4
ABSTRACT.....	5
ΕΙΣΑΓΩΓΗ.....	6
ΚΕΦΑΛΑΙΟ 1: ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ	7
1.1 Εισαγωγή στην Έννοια της Συσχέτισης	7
1.2 Ιστορική Εξέλιξη	8
ΚΕΦΑΛΑΙΟ 2: ΜΙΑ ΓΕΝΙΚΗ ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ ΣΥΣΧΕΤΙΣΗΣ	9
2.1.1 Συντελεστές συσχέτισης για Ποσοτικές Μεταβλητές.....	9
2.1.2 Συντελεστές Συσχέτισης για Ποιοτικές Μεταβλητές	19
2.1.3 Πολυμεταβλητή Ανάλυση Συσχέτισης.....	22
2.1.4 Σύνοψη.....	28
ΚΕΦΑΛΑΙΟ 3: ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΣΥΣΧΕΤΙΣΗΣ ΣΤΗΝ ΕΡΕΥΝΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ.....	31
3.1 Βιβλιογραφική ανασκόπηση.....	31
3.2 Εμπειρική διερεύνηση με εργαλεία της αναλυτικής δεδομένων.....	44
3.2.1 Υπολογιστική Προσέγγιση των μέτρων συσχέτισης μέσω γλώσσα προγραμματισμού R.....	44
3.2.2 Υπολογιστική προσέγγιση μέσω πλατφόρμας αναλυτικής δεδομένων KNIME.....	74
3.2.3 Υπολογισμός μέτρων συσχέτισης μέσω ενσωμάτωσης της γλώσσας προγραμματισμού R στην πλατφόρμα αναλυτικής δεδομένων KNIME.....	81
3.3 Οπτικοποίηση των μοντέλων συσχέτισης.....	83
ΚΕΦΑΛΑΙΟ 4: ΣΥΜΠΕΡΑΣΜΑΤΑ.....	97
4.1 Περιορισμοί και προτάσεις για περαιτέρω έρευνα.....	99

ΠΕΡΙΛΗΨΗ

Η ανάλυση συσχέτισης περιλαμβάνει έναν ευρύ αριθμό τεχνικών και μέτρων συσχέτισης, ανάλογα με το είδος των δεδομένων που εξετάζονται. Πέρα από τους κλασικούς συντελεστές συσχέτισης Pearson, Spearman, Kendall's Tau, Hoeffding's D, νέες μετρικές, όπως η συσχέτιση απόστασης (Distance Correlation) και ο μέγιστος συντελεστής πληροφορίας (Maximal Information Coefficient) έχουν αναπτυχθεί για την ανάλυση πολύπλοκων και μη γραμμικών σχέσεων.

Με τη διάδοση των ελεύθερων λογισμικών προγραμματισμού, οι τεχνικές συσχέτισης έχουν εξελιχθεί υπολογιστικά. Η εύληπτη παρουσίαση των πινάκων συσχέτισης με εικόνες και η πολυμεταβλητή ανάλυση συσχέτισης σε ένα output, είναι από τα πλεονεκτήματα που μας παρέχει η χρήση εργαλείων αναλυτικής δεδομένων. Στην παρούσα εργασία τέτοια αποτελέσματα παρουσιάζονται στη γλώσσα προγραμματισμού R, ενώ στην πλατφόρμα αναλυτικής δεδομένων KNIME, διερευνώνται οι υπολογιστικές προσεγγίσεις της συσχέτισης μέσω οπτικού προγραμματισμού και, παρουσιάζεται η ενσωμάτωση της R στο KNIME για την παραγωγή συνδυαστικών αποτελεσμάτων.

Λέξεις Κλειδιά: ανάλυση συσχέτισης, γλώσσα υπολογιστικής στατιστικής R, οπτικός προγραμματισμός, πλατφόρμα αναλυτικής δεδομένων KNIME.

ABSTRACT

Correlation analysis encompasses a wide range of techniques and measures, depending on the type of data under examination. Beyond the classical correlation coefficients, such as Pearson, Spearman, Kendall's Tau, and Hoeffding's D new metrics like Distance Correlation and the Maximal Information Coefficient (MIC) have been developed to analyze complex and nonlinear relationships.

With the rise of open-source programming software, correlation techniques have evolved computationally, offering capabilities for clear presentation of correlation matrices, multivariate analysis, and data visualization. In this study, such results are presented using the R programming language, while computational approaches to correlation through visual programming are explored within the KNIME analytics platform, including the integration of R into KNIME for the production of combined outputs.

Keywords: correlation analysis, R statistical programming language, visual programming, KNIME analytics platform.

ΕΙΣΑΓΩΓΗ

Η ανάλυση συσχέτισης αποτελεί μία από τις πιο θεμελιώδεις και διαδεδομένες τεχνικές στην ποσοτική έρευνα, με ευρεία εφαρμογή στους τομείς της Οικονομικής Επιστήμης, της Στατιστικής, των Επιχειρήσεων και της Ανάλυσης Δεδομένων. Η ουσία της συσχέτισης έγκειται στη μελέτη του τρόπου με τον οποίο δύο ή περισσότερες μεταβλητές σχετίζονται μεταξύ τους, είτε θετικά είτε αρνητικά, είτε με γραμμικό είτε με μη γραμμικό τρόπο. Η κατανόηση αυτών των σχέσεων επιτρέπει την καλύτερη ερμηνεία οικονομικών φαινομένων, τη διατύπωση υποθέσεων, τη λήψη αποφάσεων και την ανάπτυξη προβλεπτικών μοντέλων.

Παραδοσιακά, η συσχέτιση μελετάται μέσω γνωστών στατιστικών συντελεστών, όπως ο Συντελεστής Συσχέτισης Pearson, που εξετάζει τη γραμμική σχέση μεταξύ δύο ποσοτικών μεταβλητών, ή οι μη παραμετρικοί συντελεστές Spearman και Kendall's Tau, που βασίζονται σε κατατάξεις. Παράλληλα, άλλες τεχνικές όπως ο Hoeffdings's D προσφέρουν τη δυνατότητα ανίχνευσης πιο σύνθετων μορφών εξάρτησης. Ωστόσο, με την αυξανόμενη πολυπλοκότητα των δεδομένων και την ανάγκη ανάλυσης μη γραμμικών ή και πολυδιάστατων σχέσεων, αναπτύχθηκαν νέες μετρικές, όπως η Συσχέτιση Απόστασης (Distance Correlation) και ο Μέγιστος Συντελεστής Πληροφορίας (Maximal Information Coefficient – MIC), που ξεπερνούν τους περιορισμούς των παραδοσιακών μεθόδων.

Η εξέλιξη των τεχνικών συσχέτισης συνοδεύτηκε από την εξίσου σημαντική ανάπτυξη λογισμικών και εργαλείων ανάλυσης δεδομένων. Οι πλατφόρμες ανοικτού κώδικα, όπως η γλώσσα προγραμματισμού R, προσφέρουν ένα ισχυρό και ευέλικτο περιβάλλον για την εφαρμογή σύνθετων αναλυτικών μεθόδων, την παραμετροποίηση πινάκων συσχέτισης και την οπτικοποίηση αποτελεσμάτων. Παράλληλα, πλατφόρμες οπτικού προγραμματισμού όπως το KNIME Analytics Platform, παρέχουν τη δυνατότητα εκτέλεση υπολογισμών με διαδραστικό και γραφικό τρόπο, ενώ υποστηρίζουν και την ενσωμάτωση της R για τη δημιουργία συνδυαστικών λύσεων.

Στο πλαίσιο αυτό, η παρούσα διπλωματική εργασία επιδιώκει να διερευνήσει ποιοι είναι οι βασικοί συντελεστές συσχέτισης που χρησιμοποιούνται στην ανάλυση οικονομικών δεδομένων και ποια είναι τα χαρακτηριστικά τους, σε ποιες περιπτώσεις οι παραδοσιακοί συντελεστές όπως ο Pearson ή ο Spearman ενδέχεται να είναι ανεπαρκείς, πως διαφοροποιούνται τα αποτελέσματα με την εφαρμογή νέων μετρικών όπως η Distance Correlation και ο Maximal Information Coefficient (MIC), και ποια είναι τα πλεονεκτήματα και οι περιορισμοί της εφαρμογής αυτών των μεθόδων σε σύγχρονα υπολογιστικά

περιβάλλοντα. Επιπλέον επιχειρείται να αποτιμηθεί πως συμβάλλουν οι σύγχρονες υπολογιστικές τεχνικές στη βελτίωση της ερμηνείας των σχέσεων μεταξύ μακροοικονομικών και επιχειρηματικών μεταβλητών.

Για την απάντηση στα παραπάνω ερευνητικά ερωτήματα χρησιμοποιούνται δύο εργαλεία: η γλώσσα προγραμματισμού R και η πλατφόρμα αναλυτικής δεδομένων KNIME. Μέσω της R πραγματοποιείται ο υπολογισμός και η οπτικοποίηση των συντελεστών συσχέτισης, ενώ μέσω του KNIME διερευνώνται εναλλακτικές προσεγγίσεις μέσω οπτικού διαγράμματος ροής, αξιοποιώντας ταυτόχρονα τη δυνατότητα ενσωμάτωσης της R για τη χρησιμότητα των σύγχρονων τεχνικών συσχέτισης και να προσφέρει μια εμπειρική σύγκριση των παραδοσιακών νεότερων μεθόδων, συμβάλλοντας έτσι στην καλύτερη κατανόηση της σχέσης μεταξύ μεταβλητών στο πεδίο της εφαρμοσμένης οικονομικής ανάλυσης.

ΚΕΦΑΛΑΙΟ 1: ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

1.1 Εισαγωγή στην Έννοια της Συσχέτισης

Καθώς η διπλωματική εργασία έχει ως θέμα «Εμπειρική Διερεύνηση Μέτρων Συσχέτισης με Τεχνικές και Εργαλεία Αναλυτικής Δεδομένων» είναι σημαντικό να εξετάσουμε την έννοια της συσχέτισης, μια από τις πιο θεμελιώδεις στατιστικές μεθόδους που χρησιμοποιούνται για την αξιολόγηση των σχέσεων μεταξύ των μεταβλητών. Θα αναλυθεί η σημασία της στην ανάλυση δεδομένων και η ευρεία εφαρμογή της σε διάφορους τομείς. Ο λόγος της επιλογής του συγκεκριμένου θέματος είναι γιατί η ανάλυση συσχέτισης χρησιμοποιείται ευρέως για την αξιολόγηση σχέσεων μεταξύ διαφορετικών οικονομικών παραγόντων. Για παράδειγμα η ανάλυση των σχέσεων μεταξύ τιμών προϊόντων και της διαθέσιμης προσφοράς μπορεί να βοηθήσει στην κατανόηση των διακυμάνσεων των τιμών (Tinbergen, 1947). Η κατανόηση της συσχέτισης είναι κρίσιμη για την ανάπτυξη μοντέλων πρόβλεψης και τη λήψη αποφάσεων βασισμένων σε δεδομένα. Σε έρευνες οικονομικού περιεχομένου, η συσχέτιση βοηθά να αποκαλυφθούν οι κρυφές σχέσεις μεταξύ των δεδομένων, καθιστώντας δυνατή την ανάπτυξη ακριβών και αξιόπιστων οικονομικών μοντέλων (Tinbergen, 1947).

Η διερεύνηση της συσχέτισης είναι μία θεμελιώδης στατιστική ανάλυση που χρησιμοποιείται για την αξιολόγηση της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών. Αυτή η έννοια χρησιμοποιείται ευρέως σε διάφορους τομείς, όπως τα οικονομικά, η βιολογία,

η ιατρική και η κοινωνιολογία για να κατανοηθεί η δυναμική των δεδομένων και να αποκαλυφθούν πιθανά πρότυπα. Σε αρκετές περιπτώσεις όταν μελετάμε δύο ή περισσότερες μεταβλητές X , Y και Z δεν είμαστε σε θέση να εξακριβώσουμε ποια μπορεί να θεωρηθεί ως ανεξάρτητη και ποια ως εξαρτημένη μεταβλητή, επειδή δεν έχουμε την δυνατότητα να ελέγξουμε καμία. Ωστόσο όμως μπορούμε να μετρήσουμε την ένταση της συμμεταβολής που υφίσταται μεταξύ των τιμών των τυχαίων μεταβλητών, χρησιμοποιώντας τον συντελεστή συσχέτισης. (Καραπιστόλη, 2015)

1.2 Ιστορική Εξέλιξη

Η κατανόηση της συσχέτισης δεν μπορεί να είναι πλήρης χωρίς μια αναδρομή στην ιστορική εξέλιξη. Σε αυτή την ενότητα, θα εξετάσουμε πώς η έννοια της συσχέτισης αναπτύχθηκε. Η ανάπτυξη και η εφαρμογή των μέτρων συσχέτισης έχουν μακρά ιστορία, από τις πρώτες θεωρητικές ανακαλύψεις έως τις σύγχρονες εφαρμογές στατιστικής ανάλυσης. Η έννοια της συσχέτισης ξεκίνησε να διαμορφώνεται τον 19^ο αιώνα με τον Francis Galton, ο οποίος εισήγαγε την έννοια της "παλινδρόμησης προς τη μέση" (Asuero et al., 2006) και διερεύνησε τη σχέση μεταξύ των χαρακτηριστικών των γονέων και των απογόνων τους. Το έργο του Galton αποτέλεσε τη βάση για την ανάπτυξη της γραμμικής συσχέτισης (Asuero et al., 2006).

Ο συντελεστής συσχέτισης Pearson, ανεπτυγμένος από τον Karl Pearson στις αρχές του 20ού αιώνα, είναι ένας από τους ευρέως χρησιμοποιούμενους δείκτες συσχέτισης λόγω της απλότητας στον υπολογισμό και την ερμηνεία του. Ωστόσο, ο συντελεστής συσχέτισης Pearson έχει σχεδιαστεί για να χρησιμοποιείται με κανονικά κατανεμημένες μεταβλητές. Παρ' όλα αυτά, πολλές φορές χρησιμοποιείται λανθασμένα σε οποιοδήποτε είδος δεδομένων (Forthofer et al., 2007). Λόγω αυτών των περιορισμών, το 1904 ο Charles Spearman, ένας Βρετανός ψυχομετρικός, πρότεινε μια τροποποίηση του Pearson, ώστε να μπορεί να εφαρμοστεί σε μη κανονικά δεδομένα. Ο Spearman πρότεινε τη μετατροπή των τιμών σε τάξεις (ranks), αντί της χρήσης των πραγματικών τιμών, και στη συνέχεια την εφαρμογή του τύπου Pearson στα ranks. Το αποτέλεσμα αυτής της διαδικασίας είναι ο Spearman Rank Correlation Coefficient (r_s) (Forthofer et al., 2007). Ο Spearman rank correlation μπορεί να χρησιμοποιηθεί όχι μόνο με μη κανονικά συνεχή, δεδομένα, αλλά και διατακτικές (ordinal) μεταβλητές, γεγονός που τον καθιστά ιδιαίτερα χρήσιμο σε κοινωνικές επιστήμες. Όταν προκύπτουν ισοβαθμίες στις κατατάξεις, αυτές λαμβάνουν τον ίδιο μέσο όρο των θέσεων και αυτό δεν

επηρεάζει τον υπολογισμό spearman (Forthofer et al., 2007).

Παράλληλα, ο Maurice Kendall ανέπτυξε το Kendall's Tau, ένα μη παραμετρικό μέτρο που αξιολογεί τη σχετική κατάταξη των δεδομένων. Αργότερα, ο Harold Hotelling εισήγαγε την Canonical Correlation Analysis (CCA) το 1936, επιτρέποντας την ανάλυση της συσχέτισης μεταξύ δύο συνόλων μεταβλητών. Το 1948 ο Wassily Hoeffding's ανέπτυξε έναν μη παραμετρικό συντελεστή που είναι ιδιαίτερα χρήσιμος για την ανίχνευση πολύπλοκων και μη γραμμικών συσχετίσεων, καθιστώντας τον κατάλληλο για περιπτώσεις όπου οι μέθοδοι Pearson, Spearman και Kendall δεν αποκαλύπτουν σχέσεις μεταξύ των δεδομένων. Με την έλευση των ελεύθερων περιβαλλόντων προγραμματισμού και των Big Data, οι τεχνικές συσχέτισης έχουν εξελιχθεί ραγδαία. Νέες μέθοδοι, όπως η Distance Correlation και ο Maximal Information Coefficient (MIC), έχουν αναπτυχθεί για την ανάλυση πολύπλοκων και μη γραμμικών σχέσεων (Clark, 2013). Ενώ τα τελευταία χρόνια έχουν αναπτυχθεί ανθεκτικότερα μέτρα συσχέτισης όπως ο Biweight mid, ο Percentage Bend, Ο Shepherds Pi και ο συντελεστής συσχέτισης Gamma.

ΚΕΦΑΛΑΙΟ 2: ΜΙΑ ΓΕΝΙΚΗ ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ ΣΥΣΧΕΤΙΣΗΣ

2.1.1 Συντελεστές συσχέτισης για Ποσοτικές Μεταβλητές

Σε αυτή την ενότητα, θα εξετάσουμε τους διάφορους συντελεστές συσχέτισης που χρησιμοποιούνται για τη μέτρηση της σχέσης μεταξύ ποσοτικών μεταβλητών.

Οι συντελεστές συσχέτισης για ποσοτικές μεταβλητές χρησιμοποιούνται για να μετρήσουν τη δύναμη και την κατεύθυνση της σχέσης μεταξύ δύο συνεχών ποσοτικών μεταβλητών. Αυτοί οι συντελεστές παρέχουν μια αριθμητική τιμή που εκφράζει πόσο κοντά βρίσκονται οι δύο μεταβλητές σε μία γραμμική ή μη γραμμική συσχέτιση. Οι κύριοι τύποι συντελεστών συσχέτισης για ποσοτικές μεταβλητές περιλαμβάνουν:

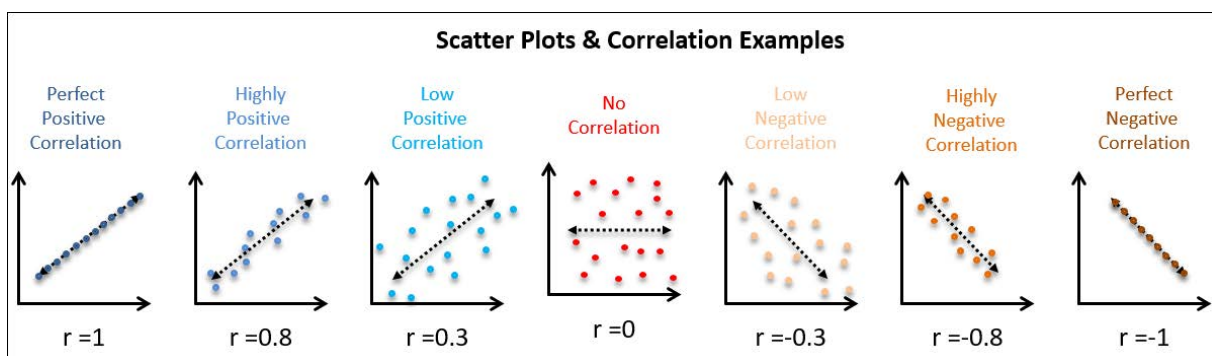
Συντελεστές Συσχέτισης για Γραμμικές Σχέσεις:

Όταν λέμε γραμμική σχέση, εννοούμε μια σχέση μεταξύ δύο μεταβλητών που μπορεί να ερμηνευτεί από μια ευθεία γραμμή. Δηλαδή, η μία μεταβλητή αλλάζει με έναν σταθερό ρυθμό σε σχέση με την άλλη. Αν σχεδιάσουμε τα δεδομένα σε ένα διάγραμμα διασποράς (Scatter Plot), τα σημεία τείνουν να σχηματίζουν μια ευθεία γραμμή, είτε με θετική είτε με

αρνητική κλίση. Συντελεστής συσχέτισης για γραμμικές σχέσεις είναι ο συντελεστής Pearson καθώς και νέα ανθεκτικότερα μέτρα όπως ο Biweight Mid και ο Percentage Bend.

Συντελεστής Συσχέτισης Pearson:

Ο συντελεστής συσχέτισης Pearson, συχνά αναφερόμενος ως **Pearson's r** , που αναπτύχθηκε από τον Karl Pearson στις αρχές του 20ού αιώνα, είναι ένας από τους πιο διαδεδομένους και βασικούς μετρητές συσχέτισης. Ορίζεται ως το πηλίκο της διακύμανσης των προϊόντων των δύο μεταβλητών προς το γινόμενο των τυπικών αποκλίσεων τους (Lind et al., 2012). Είναι ένα μέτρο που μετράει αποκλειστικά τη γραμμική συσχέτιση μεταξύ δύο συνεχών μεταβλητών και κυμαίνεται από -1 έως 1. Ένας συντελεστής συσχέτισης +1 δηλαδή $r = 1$ υποδηλώνει τέλεια θετική γραμμική συσχέτιση, ενώ $r = -1$ υποδηλώνει τέλεια αρνητική γραμμική συσχέτιση. Ένας συντελεστής $r = 0$ υποδηλώνει ότι δεν υπάρχει γραμμική σχέση μεταξύ των δύο μεταβλητών (Clark, 2013). Στην πράξη, τα δεδομένα από μια συγκεκριμένη χρονική περίοδο μπορούν να δώσουν έναν Pearson $r = 0.85$, υποδεικνύοντας μια ισχυρή θετική συσχέτιση. Στο σχήμα 1 απεικονίζονται οι διάφορες μορφές που μπορεί να πάρει η γραμμική συσχέτιση Pearson.



Εικόνα 1. Τίτλος: Παραδείγματα Γραμμικών Συσχετίσεων (Pearson)

Πηγή: CQE Academy¹

Ο συντελεστής συσχέτισης Pearson εδραιώνει μια σχέση μεταξύ των δύο μεταβλητών με βάσει πέντε υποθέσεις. Αυτές είναι:

- a) **Η σχέση είναι γραμμική:** Υποθέτει ότι η σχέση μεταξύ δύο μεταβλητών είναι γραμμική, δηλαδή ότι οι αλλαγές στη μία μεταβλητή προκαλούν αναλογικές αλλαγές στην άλλη. Όταν

¹<https://cqeacademy.com/cqe-body-of-knowledge/continuous-improvement/quality-control-tools/the-scatter-plot-linear-regression/>

αυτή η παραδοχή δεν ισχύει, τα αποτελέσματα μπορεί να είναι παραπλανητικά (Gogtay & Thatte, 2017).

- b) Οι μεταβλητές είναι ανεξάρτητες μεταξύ τους:** Οι παρατηρήσεις πρέπει να είναι ανεξάρτητες μεταξύ τους. Η παραβίαση αυτής της παραδοχής μπορεί να οδηγήσει σε ανακριβείς εκτιμήσεις του συντελεστή συσχέτισης (Gogtay & Thatte, 2017).
- c) Οι μεταβλητές είναι κανονικά κατανεμημένες:** Οι μεταβλητές θα πρέπει να ακολουθούν την κανονική κατανομή. Εάν αυτή η παραδοχή δεν πληρείται, ο συντελεστής Pearson μπορεί να μην είναι η κατάλληλη μέθοδος συσχέτισης (Gogtay & Thatte, 2017).
- d) Απουσία Ακραίων τιμών:** Είναι σημαντικό να αντιμετωπιστούν οι ακραίες τιμές πριν την ανάλυση καθώς η παρουσία τους επηρεάζουν σημαντικά τον συντελεστή συσχέτισης (Gogtay & Thatte, 2017).
- e) Η φύση των Μεταβλητών:** Οι μεταβλητές πρέπει να είναι ποσοτικές και συνεχείς για να είναι έγκυρη η ανάλυση συσχέτισης Pearson (Clark, 2013).

Ο τύπος του συντελεστή Pearson είναι:

$$r = \frac{\sum(X - \bar{X}) \cdot (Y - \bar{Y})}{(n-1)s_X s_Y} \quad (1)$$

Όπου:

- X και Y είναι οι τιμές των δύο μεταβλητών,
- $\bar{X}$ και $\bar{Y}$ είναι οι μέσες τιμές των X και Y ,
- s_X και s_Y είναι οι τυπικές αποκλίσεις των X και Y , και
- n είναι το μέγεθος του δείγματος.

Ενώ ο Pearson είναι κατάλληλος για γραμμικές σχέσεις, υπάρχουν περιπτώσεις όπου οι σχέσεις μεταξύ μεταβλητών δεν είναι γραμμικές. Σε αυτές τις περιπτώσεις, χρησιμοποιείται ο συντελεστής συσχέτισης Spearman.

Συντελεστής Global Correlation Coefficient (gk):

Ο Global Correlation Coefficient (gk) με βάση τους James και Roos (1975), είναι ένα χρήσιμο εργαλείο για την ανάλυση της ισχύος των συσχετίσεων μεταξύ μιας μεταβλητής και του γραμμικού συνδυασμού όλων των άλλων μεταβλητών. Ο συντελεστής αυτός υπολογίζεται από τον πίνακα συσχετίσεων και εκφράζει την υψηλότερη δυνατή συσχέτιση της μεταβλητής με τον συνδυασμό των υπόλοιπων μεταβλητών (Baak et al., 2020). Η τιμή gk κυμαίνεται

μεταξύ 0 και 1. Αυτός ο συντελεστής είναι χρήσιμος σε εφαρμογές όπου θέλουμε να εκτιμήσουμε πόσο καλά μπορεί να προβλεφθεί μια μεταβλητή με βάση τις υπόλοιπες μεταβλητές, ανεξάρτητα από τον τύπο μεταβλητών αυτών (π.χ. κατηγορικές, διατακτικές, ποσοτικές) (Baak et al., 2020). Επομένως ο τύπος του Global Correlation Coefficient είναι ο εξής:

$$g_k = \sqrt{1 - [V_{kk} \cdot (V^{-1})_{kk}]}$$
 (2)

Όπου:

- V_{kk} : το διαγώνιο στοιχείο του πίνακα διακύμανσης V για τη μεταβλητή k ,
- $(V^{-1})_{kk}$: το αντίστοιχο στοιχείο του αντίστροφου πίνακα διακύμανσης.

Συντελεστής Μερικής Συσχέτισης (Partial Correlation Coefficient):

Ο συντελεστής μερικής συσχέτισης (Partial Correlation Coefficient) είναι ένας γραμμικός συντελεστής καθώς χρησιμοποιεί μεθόδους που αφαιρούν τις μη γραμμικές εξαρτήσεις από τρίτες μεταβλητές για να προσφέρει μια πιο ολοκληρωμένη ανάλυση. Συγκεκριμένα ο συντελεστής μερικής συσχέτισης χρησιμοποιείται για να μετρήσει τη συσχέτιση μεταξύ δύο μεταβλητών, ενώ αφαιρεί την επίδραση μίας ή περισσότερων άλλων μεταβλητών (γνωστές ως συγχέουσες μεταβλητές). Οι παραδοσιακοί μερικοί συντελεστές συσχέτισης είναι σχεδιασμένοι για να επιτρέπουν στους ερευνητές να εστιάζουν στη σχέση μεταξύ δύο μεταβλητών αφού αφαιρεθεί η γραμμική επίδραση της τρίτης μεταβλητής (Vinod, 2017). Ο τύπος για τον υπολογισμό του παραδοσιακού μερικού συντελεστή είναι ο εξής:

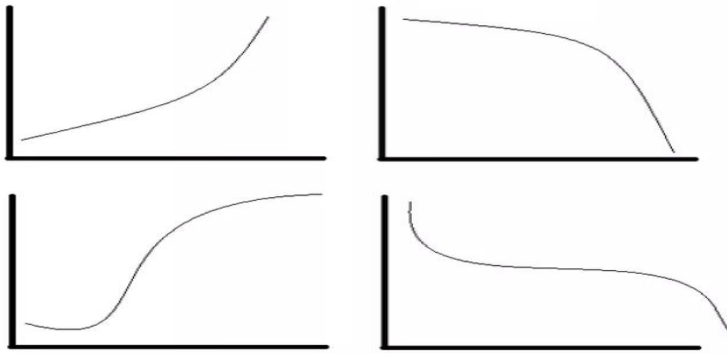
$$r_{12;3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$
 (3)

Όπου:

- r_{12} : είναι ο συντελεστής συσχέτισης μεταξύ των μεταβλητών X_1 και X_2
- r_{13} : είναι ο συντελεστής συσχέτισης μεταξύ X_1 και της συγχέουσας μεταβλητής X_3
- r_{23} : είναι ο συντελεστής συσχέτισης μεταξύ X_2 και X_3

Συντελεστές Συσχέτισης για Μη γραμμικές Σχέσεις:

Με τον όρο μη γραμμικές σχέση, εννοούμε μια σχέση μεταξύ δύο μεταβλητών που δεν μπορεί να παρασταθεί με ευθεία γραμμή. Σε αντίθεση με τις γραμμικές σχέσεις, στις μη γραμμικές σχέσεις ο ρυθμός της αλλαγής δεν είναι σταθερός και μπορεί να ποικίλει ανάλογα με την τιμή των μεταβλητών. Στο σχήμα 2 βλέπουμε παραδείγματα μη γραμμικών σχέσεων.



Εικόνα 2. Τίτλος: Παραδείγματα μη γραμμικών σχέσεων.

Πηγή: <https://el.economy-pedia.com/11035533-non-linear-relationship>

Για τη μελέτη μη γραμμικών σχέσεων, που δεν μπορούν να αναλυθούν με τον συντελεστή Pearson, έχουν αναπτυχθεί άλλοι συντελεστές, όπως η Συσχέτιση Απόστασης (Distance Correlation) και ο Μέγιστος Συντελεστής Πληροφορίας (Maximal Information Coefficient- MIC). Αυτές οι μέθοδοι είναι κατάλληλες για πιο πολύπλοκες σχέσεις, όπου η συσχέτιση δεν είναι απαραίτητα γραμμική. Αξίζει να σημειωθεί ότι οι συντελεστές Spearman, Kendall και Hoeffding αποδίδουν μη γραμμικές σχέσεις, ωστόσο θα αναφερθούν και θα αναλυθούν στην συνέχεια στην ενότητα «Συντελεστές Συσχέτισης για Μονοτονικές Σχέσεις».

Συντελεστής Συσχέτισης Αποστάσεων (Distance Correlation Coefficient):

Ο συντελεστής συσχέτισης των αποστάσεων είναι ένα μέτρο που χρησιμοποιεί τις αποστάσεις μεταξύ των παρατηρήσεων για να μετρήσει τη δύναμη της σχέσης μεταξύ δύο μεταβλητών. Έχει αποδειχθεί ότι είναι αποτελεσματική σε περιπτώσεις όπου οι παραδοσιακές μέθοδοι αποτυγχάνουν να ανιχνεύσουν συσχετίσεις, καθιστώντας την ιδιαίτερα χρήσιμη για τη μελέτη πολύπλοκων, μη γραμμικών σχέσεων (Székely et al., 2007). Ο τύπος της Συσχέτισης Απόστασης ορίζεται ως εξής (Lind et al., 2012) :

$$d\text{Cov}^2(X,Y) = \frac{1}{n^2} \sum_{ij=1}^n A_{ij}B_{ij} \quad (4)$$

Όπου:

- A_{ij} και B_{ij} είναι οι αποστάσεις των παρατηρήσεων X και Y αντίστοιχα,
- n είναι το πλήθος των παρατηρήσεων.
- Η τιμή του συντελεστή είναι μηδέν όταν οι μεταβλητές είναι ανεξάρτητες. Σε κάθε άλλη περίπτωση η τιμή του είναι θετική, καθώς ανιχνεύει κάθε μορφή εξάρτησης.

Μέγιστος Συντελεστής Πληροφορίας (MIC):

Ο Μέγιστος Συντελεστής Πληροφορίας (MIC) αναπτύχθηκε ως ένα "μέτρο συσχέτισης για τον 21ο αιώνα", βασισμένος σε έννοιες από τη θεωρία της πληροφορίας. Είναι ένα σύγχρονο μέτρο συσχέτισης το οποίο αναπτύχθηκε στο πλαίσιο της οικογένειας μέτρων MINE (Maximal Information-based Nonparametric Exploration). Είναι ιδιαίτερα αποτελεσματικός στην ανίχνευση ισχυρών σχέσεων μεταξύ των μεταβλητών, ανεξάρτητα από τη φύση τους, καθιστώντας τον χρήσιμο για την ανάλυση μεγάλων και πολύπλοκων δεδομένων καθώς υπερβαίνει τους περιορισμούς των κλασικών μέτρων συσχέτισης και μπορεί να ανιχνεύσει τόσο γραμμικές όσο και μη γραμμικές σχέσεις, προσφέροντας μεγαλύτερη ευελιξία στην ανάλυση δεδομένων (Clark, 2013). Ο τύπος του MIC ορίζεται ως εξής (Lind et al., 2012):

$$MIC(X, Y) = \max\left(\frac{I(X; Y)}{\log_2(\min(|X|, |Y|))}\right) \quad (5)$$

Όπου:

- $I(X; Y)$ είναι η αμοιβαία πληροφορία μεταξύ των X και Y ,
- $\min(|X|, |Y|)$ είναι το ελάχιστο μέγεθος των δύο συνόλων παρατηρήσεων για τις μεταβλητές X και Y
- Η αμοιβαία πληροφορία $I(X; Y)$ μετρά την ποσότητα της πληροφορίας που μοιράζονται οι δύο μεταβλητές και μπορεί να ανιχνεύσει οποιοδήποτε είδος σχέσης μεταξύ τους, είτε γραμμική είτε μη γραμμική (Clark, 2013).

Για μια πιο ολοκληρωμένη εικόνα των σχέσεων μεταξύ των μεταβλητών ο συντελεστής συσχέτισης (MIC) συνοδεύεται από τρία επιπλέον μέτρα που προσφέρουν διαφορετικές οπτικές στη δομή των δεδομένων:

- *Maximum Asymmetry Score (MAS)*: Μετρά τη μη συμμετρία στη σχέση μεταξύ δύο μεταβλητών, ανιχνεύοντας ετεροσκελείς σχέσεις.

- Maximum Edge Value (MEV): Αξιολογεί την πολυπλοκότητα της σχέσης, παρέχοντας πληροφορίες για τη δυσκολία μοντελοποίησης της.
- Minimum Cell Number (MCN): Δείχνει τον ελάχιστο αριθμό κελιών που απαιτούνται για την ανίχνευση σχέσεων στα δεδομένα, υποδεικνύοντας τη λεπτομέρεια που απαιτείται στην ανάλυση.

Σε αντίθεση με τα κλασικά μέτρα συσχέτισης όπως ο Pearson ή ο Spearman, το MIC και τα συνοδευτικά μέτρα MAS, MEV και MCN παρέχουν μια πιο ολοκληρωμένη εικόνα των σχέσεων μεταξύ των μεταβλητών, επιτρέποντας την ανίχνευση πολύπλοκων δομών και εξαρτήσεων στα δεδομένα. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε εφαρμογές μεγάλων δεδομένων και μηχανικής μάθησης, όπου οι σχέσεις μεταξύ των μεταβλητών μπορεί να είναι μη προφανείς και μη γραμμικές.

Συντελεστές Συσχέτισης για Μονοτονικές Σχέσεις:

Εκτός από τις γραμμικές σχέσεις, υπάρχουν και περιπτώσεις όπου οι σχέσεις μεταξύ των μεταβλητών είναι μονοτονικές. Οι μονοτονικές σχέσεις αναφέρονται σε σχέσεις μεταξύ δύο μεταβλητών, στις οποίες η μεταβολή της μία μεταβλητής συνοδεύεται πάντα από τη μεταβολή της άλλης προς την ίδια ή την αντίθετη κατεύθυνση, χωρίς απαραίτητα η σχέση να είναι γραμμική. Με άλλα λόγια, μια μονοτονική σχέση είναι μια σχέση που διατηρεί την κατεύθυνση αλλαγής², αλλά μπορεί να μην έχει σταθερό ρυθμό. Για την ανάλυση αυτών των σχέσεων, χρησιμοποιούνται συγκεκριμένοι συντελεστές συσχέτισης, όπως ο συντελεστής Spearman's Rho, ο συντελεστής Kendall's Tau και ο συντελεστής Hoeffding's D.

Συντελεστής Συσχέτισης Spearman:

Ο συντελεστής συσχέτισης Spearman, γνωστός ως **Spearman's rank correlation coefficient (rs)** αναπτύχθηκε από τον Charles Spearman, είναι μία μη παραμετρική μέθοδος που βασίζεται σε κατατάξεις (rankings) και μπορεί να χρησιμοποιηθεί για την ανάλυση μη γραμμικών σχέσεων. Οι προϋποθέσεις για την ορθή χρήση του συντελεστή Spearman (rs) περιλαμβάνουν τα εξής:

² Η διατήρηση της κατεύθυνσης αλλαγής είναι το χαρακτηριστικό που κάνει τη σχέση να είναι μονοτονική, ακόμα κι αν δεν είναι γραμμική (δηλαδή η αύξηση ή η μείωση μπορεί να γίνεται με μεταβλητό ρυθμό, αλλά πάντα προς την ίδια κατεύθυνση).

- a) **Μη Κανονική Κατανομή:** Ο συντελεστής Spearman είναι μη παραμετρικός, επομένως δεν απαιτεί κανονικότητα στη διανομή των δεδομένων. Είναι ιδιαίτερα χρήσιμος όταν τα δεδομένα δεν πληρούν τις παραδοχές της κανονικότητας (Clark, 2013).
- b) **Ανθεκτικότητα σε Ακραίες Τιμές (Outliers):** Σε αντίθεση με τον Pearson, ο Spearman είναι πιο ανθεκτικός στις ακραίες τιμές, επειδή βασίζεται στην κατάταξη και όχι στις αρχικές τιμές. Ωστόσο, αν υπάρχουν πολλές δεσμευμένες τιμές (tied ranks), αυτό μπορεί να επηρεάσει την ακρίβεια του αποτελέσματος (Clark, 2013).
- c) **Μονοτονική Σχέση:** Ο Spearman's (rs) μετράει τη δύναμη της μονοτονικής (όχι απαραίτητα γραμμικής) συσχέτισης μεταξύ δύο μεταβλητών. Αυτές οι προϋποθέσεις καθιστούν τον συντελεστή Spearman κατάλληλο για πολλές περιπτώσεις όπου η γραμμική συσχέτιση δεν είναι κατάλληλη λόγω της μη γραμμικής σχέσης των δεδομένων (Clark, 2013).
- d) **Διαχείριση Ισοβαθμιών (Tied Ranks):** Σε περιπτώσεις όπου δύο ή περισσότερες τιμές είναι ίδιες, λαμβάνουν την ίδια μέση κατάταξη. Αυτό δεν δημιουργεί πρόβλημα στον υπολογισμό του Spearman, καθώς χρησιμοποιεί την τροποποιημένη εκδοχή του τύπου Pearson για τις κατατάξεις (Looney & Hagan, 2007).
- e) **Εφαρμογή σε Ταξινομικά Δεδομένα (Ordinal Data):** Ο συντελεστής Spearman μπορεί να χρησιμοποιηθεί όχι μόνο σε συνεχή μη κανονικά δεδομένα, αλλά και σε κατηγορικά δεδομένα με ταξινομική κλίμακα (ordinal). Αυτή η ιδιότητα τον καθιστά ιδιαίτερα χρήσιμο σε μετρήσεις που γίνονται συχνά με βάση κατατάξεις (Looney & Hagan, 2007).

Ο τύπος του συντελεστή Spearman είναι: (Lind et al., 2012)

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad (6)$$

Όπου:

- r_s είναι ο συντελεστής Spearman
- d είναι η διαφορά στις τάξεις για κάθε ζεύγος.
- n είναι ο αριθμός των παρατηρήσεων

Συντελεστής Συσχέτισης Kendall's Tau:

Ο συντελεστής Kendall's Tau είναι ένας μη παραμετρικός δείκτης που χρησιμοποιείται για την αξιολόγηση της μονοτονικής συσχέτισης μεταξύ δύο μεταβλητών. Εξετάζει επίσης τις

σχέσεις με βάση τις κατατάξεις, και είναι χρήσιμος όταν οι δεδομένες σχέσεις δεν είναι γραμμικές (Clark, 2013). Οι προϋποθέσεις που πρέπει να πληρούν τα δεδομένα περιλαμβάνουν τα εξής:

- a) **Μονοτονική Σχέση:** Ο συντελεστής Kendall's Tau μετρά τη δύναμη και την κατεύθυνση μιας μονοτονικής σχέσης μεταξύ δύο μεταβλητών. Δεν απαιτείται γραμμική συσχέτιση, αλλά οι μεταβλητές πρέπει να έχουν μια συνολικά αυξανόμενη ή φθίνουσα τάση (Lind et al., 2012).
- b) **Ανθεκτικότητα σε Ακραίες Τιμές (Outliers):** Ο Kendall's Tau είναι πιο ανθεκτικός στις ακραίες τιμές σε σύγκριση με τον συντελεστή Pearson. Αυτό σημαίνει ότι τα αποτελέσματα του συντελεστή Kendall's Tau δεν επηρεάζονται σημαντικά από τις ακραίες τιμές, καθιστώντας τον κατάλληλο όταν υπάρχουν ακραίες τιμές στα δεδομένα (Lind et al., 2012).
- c) **Αντιμετώπιση Δεσμευμένων Τιμών (Ties):** Ο Kendall's Tau σε αντίθεση με τον συντελεστή Spearman έχει παραλλαγές και διαφορετική προσέγγιση (όπως το Tau-b) για τις ισοβαθμίες. Ειδικότερα ο Kendall's Tau-b είναι μία τροποποιημένη εκδοχή που λαμβάνει υπόψη τις δεσμευμένες τιμές και τις προσαρμόζει στον υπολογισμό της συσχέτισης. Αυτή η προσαρμογή καθιστά τον Kendall's Tau-b πιο κατάλληλο όταν τα δεδομένα περιέχουν μεγάλο αριθμό ισότιμων (Lind et al., 2012).

Τέλος αξίζει να αναφερθεί ότι τόσο ο συντελεστής Spearman όσο και ο συντελεστής Kendall's Tau εφαρμόζονται σε ποσοτικές και σε ποιοτικές μεταβλητές ανάλογα με την περίπτωση. Και οι δύο συντελεστές έχουν τη θέση τους στην ανάλυση δεδομένων, ανάλογα με τη φύση των δεδομένων και τις συγκεκριμένες απαιτήσεις της ανάλυσης. Σε αυτό το σημείο είναι σημαντικό να αναφέρουμε την διάκριση μεταξύ των συντελεστών Spearman και Kendall. Πιο συγκεκριμένα:

- ο Spearman μετρά τη μονοτονική συσχέτιση με βάση τις κατατάξεις των δεδομένων, και είναι πιο ευαίσθητος σε γραμμικές και μη γραμμικές σχέσεις. Ο Kendall Tau-b υπολογίζει την πιθανότητα συμφωνίας ή διαφωνίας μεταξύ ζευγών παρατηρήσεων.
- Ο Kendall Tau-b είναι λιγότερο ευαίσθητος σε ακραίες τιμές σε σύγκριση με τον Spearman, καθιστώντας τον καλύτερο για δεδομένα με τέτοιες ιδιότητες.
- Για μικρά δείγματα, ο Kendall Tau-b θεωρείται πιο ακριβής, ενώ ο Spearman είναι προτιμότερος όταν η συσχέτιση είναι πιο περίπλοκη.

- Ο Spearman είναι πιο δημοφιλής λόγω της ευκολίας χρήσης του και της ικανότητας του να ανιχνεύει μονοτονικές σχέσεις, είτε γραμμικές είτε μη γραμμικές. Ο Kendall Tau-b είναι κατάλληλος όταν θέλουμε ακριβέστερη σύγκριση δύο μεταβλητών ειδικά σε δεδομένα με ισχυρή εξάρτηση ή μικρό μέγεθος δείγματος.

Επομένως η επιλογή του συντελεστή εξαρτάται από το μέγεθος του δείγματος, την παρουσία δεσμευμένων τιμών, και την ανάγκη για ακρίβεια στα αποτελέσματα. Ο τύπος του συντελεστή Kendall's Tau είναι (Lind et al., 2012):

$$T = \frac{C-D}{C+D} \quad (7)$$

Όπου:

- C είναι ο αριθμός των σύμφωνων ζευγών, δηλαδή σύμφωνο ζεύγος (C) είναι ζεύγος παρατηρήσεων (X_i, Y_i) και (X_j, Y_j) για το οποίο, εάν $X_i < X_j$, τότε και $Y_i < Y_j$
- D είναι ο αριθμός των αντίθετων ζευγών, δηλαδή αντίθετο ζεύγος (D) είναι ένα ζεύγος παρατηρήσεων όπου η σχετική σειρά των δύο μεταβλητών είναι αντίστροφη $X_i < X_j$ αλλά $Y_i > Y_j$

Εκτός από τους συντελεστές Spearman και Kendall, οι οποίοι προϋποθέτουν τη μονοτονική σχέση μεταξύ των μεταβλητών, υπάρχει και ο συντελεστής Hoeffding's D, ο οποίος επιτρέπει την ανίχνευση πιο πολύπλοκων σχέσεων.

Συντελεστής Συσχέτισης Hoeffding's D:

Ο Hoeffding's D συντελεστής συσχέτισης είναι ένα μέτρο που αξιολογεί τη σχέση μεταξύ μεταβλητών, λαμβάνοντας υπόψη γραμμικές, μονοτονικές και μη μονοτονικές σχέσεις (Hoeffding, 1948). Ο Hoeffding's D κυμαίνεται στο διάστημα $[-0.5, 1]$ όταν δεν υπάρχουν ίδιες κατατάξεις, με τις υψηλότερες τιμές να υποδεικνύουν ισχυρότερη συσχέτιση μεταξύ των μεταβλητών (Clark, 2013). Σε αντίθεση με τους συντελεστές Spearman, Kendall και Pearson, ο Hoeffding's D μπορεί να ανιχνεύσει πιο πολύπλοκες δομές εξάρτησης πέρα από τη μονοτονική συσχέτιση (Halkos et al., 2022). Όταν ο συντελεστής Spearman για μια μεταβλητή είναι κοντά στο 0, αλλά ο Hoeffding έχει υψηλή τιμή, αυτό υποδεικνύει μία μη μονοτονική σχέση. Παρομοίως, εάν ο συντελεστής Pearson είναι κοντά στο 0 και ο Hoeffding's υψηλός, αυτό δείχνει μία μη γραμμική σχέση. Για την ανάδειξη κρυφών μη μονοτονικών σχέσεων, ο

συντελεστής Hoeffding D υπολογίζεται για μεταβλητές με χαμηλές τιμές Spearman και Pearson. Ο τύπος του Hoeffding's D είναι ο εξής (Hoeffding, 1948):

$$D = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)} \quad (8)$$

Όπου:

- n το μέγεθος του δείγματος.
- A : Είναι το συνολικό άθροισμα που προκύπτει από τον συνδυασμό κατατάξεων για όλες τις δυνατές ομάδες πέντε παρατηρήσεων μέσα στο δείγμα.
- B : Υπολογίζεται από το προϊόν κάποιων από τις κατατάξεις και σχετίζεται με τον εντοπισμό συγκεκριμένων συσχετίσεων μέσα στα δεδομένα.
- C : Αυτός ο όρος αφορά το πλήθος των κοινών στοιχείων που μοιράζονται συγκεκριμένες ιδιότητες μεταξύ των δεδομένων.

2.1.2 Συντελεστές Συσχέτισης για Ποιοτικές Μεταβλητές

Οι συντελεστές συσχέτισης για ποιοτικές μεταβλητές μετρούν τη δύναμη της σχέσης μεταξύ κατηγορικών δεδομένων. Αυτοί οι συντελεστές είναι κατάλληλοι για δεδομένα που δεν είναι αριθμητικά, αλλά κατηγοριοποιημένα, όπως οι δημογραφικές πληροφορίες, οι κατηγορίες προϊόντων ή οι κλίμακες απαντήσεων σε ερωτηματολόγια. Για δεδομένα κατάταξης (διατακτικές μεταβλητές), χρησιμοποιούνται και οι συντελεστές Spearman και Kendall που έχουμε αναφέρει.

Πιο συγκεκριμένα η διαφορά στη χρήση των Spearman και Kendall για ποιοτικές μεταβλητές είναι οι εξής:

- Spearman (rho): Διατακτικά δεδομένα (ordinal data) + γραμμική σχέση (linear relationship)
- Kendall's Tau: Διατακτικά δεδομένα (ordinal data) + μονοτονική σχέση (monotonic relationship)

και στους δύο συντελεστές Spearman και Kendall's Tau θα πρέπει τα δεδομένα να είναι σε κατατάξεις (ranks). Δηλαδή τα δεδομένα να είναι σε κλίμακα κατατάξεων ή να μπορούν να μετατραπούν σε κατατάξεις. Και οι δύο συντελεστές συσχέτισης χρησιμοποιούν τους ίδιους τύπους τόσο σε ποσοτικές μεταβλητές όσο και ποιοτικές μεταβλητές. Εφόσον αναφέραμε τους

συντελεστές συσχέτισης Spearman και Kendall νωρίτερα θα εξετάσουμε και άλλους τύπους συντελεστών συσχέτισης για ποιοτικές μεταβλητές:

Συντελεστής Συσχέτισης Cramér's V:

Ο συντελεστής Cramér's V είναι ένα μέτρο συσχέτισης που χρησιμοποιείται για κατηγορικές μεταβλητές και βασίζεται στον πίνακα συχνοτήτων. Δηλαδή πίνακα που χρησιμοποιείται για να παρουσιάσει την κατανομή των συχνοτήτων (counts) δύο ή περισσότερων κατηγορικών μεταβλητών. Αποτελείται από γραμμές και στήλες όπου κάθε κελί του πίνακα αντιπροσωπεύει τον αριθμό των περιπτώσεων που εμπίπτουν σε συγκεκριμένους συνδυασμούς κατηγοριών των μεταβλητών (Asuero et al., 2006). Κυμαίνεται από 0 έως 1, όπου 0 υποδηλώνει «καμία συσχέτιση» και 1 υποδηλώνει «τέλεια συσχέτιση» (Asuero et al., 2006). Πρόκειται για μια τροποποίηση του συντελεστή X^2 (chi-square) και χρησιμοποιείται όταν οι μεταβλητές είναι κατηγορικές (nominal) και θέλουμε να αξιολογήσουμε αν υπάρχει κάποια μορφή συσχέτισης μεταξύ τους (Yule, 1922). Ο συντελεστής υπολογίζεται ως εξής:

$$V = \sqrt{\frac{x^2}{n * \min(k-1, r-1)}} \quad (9)$$

Πηγή: (Yule, 1922)

Όπου:

- x^2 είναι η τιμή του στατιστικού chi-square.
- n είναι το μέγεθος του δείγματος.
- k είναι ο αριθμός των στηλών του πίνακα συνάφειας.
- r είναι ο αριθμός των γραμμών του πίνακα συνάφειας.

Συντελεστής συσχέτισης Φ (Phi):

Ο συντελεστής Φ είναι ένα μέτρο συσχέτισης για κατηγορικές μεταβλητές, ειδικά όταν και οι δύο μεταβλητές είναι δυαδικές δηλαδή έχουν μόνο δύο κατηγορίες (π.χ., ναι/όχι, αληθές/ψευδές) (Yule, 1922). Ο συντελεστής αυτός είναι ιδιαίτερα χρήσιμος σε περιπτώσεις όπου τα δεδομένα μπορούν να παρουσιαστούν σε έναν πίνακα 2X2. Επίσης κυμαίνεται από -1 έως 1, με τις ίδιες ερμηνείες όπως ο συντελεστής Pearson (Asuero et al., 2006). Ο τύπος για τον υπολογισμό του Φ είναι:

$$\Phi = \sqrt{\frac{\chi^2}{n}} \quad (10)$$

Πηγή: (Yule, 1922)

Όπου:

- χ^2 είναι η τιμή του στατιστικού chi-square.
- n είναι το συνολικό μέγεθος του δείγματος

Ο συντελεστής Φ είναι ουσιαστικά η τετραγωνική ρίζα της τιμής του χ^2 , διαιρεμένη με το μέγεθος του δείγματος. Ο Φ Coefficient είναι αριθμός μεταξύ -1 και 1, είναι μια ειδική περίπτωση του Cramer's V, όταν εφαρμόζονται σε πίνακες 2x2. Ο συντελεστής Cramer's V μπορεί να χρησιμοποιηθεί για μεγαλύτερους πίνακες συνάφειας (με περισσότερες από δύο κατηγορίες), ενώ ο Φ εφαρμόζεται αποκλειστικά σε διχοτομικούς πίνακες (Yule, 1922).

Συντελεστής Συσχέτισης Φ_K :

Ο συντελεστής Φ_K (Phi-k Coefficient) είναι ένα μέτρο συσχέτισης που χρησιμοποιείται για να εκτιμήσει τη δύναμη της συσχέτισης μεταξύ κατηγορικών μεταβλητών. Είναι ένας νέος συντελεστής συσχέτισης, βασισμένος στο τεστ ανεξαρτησίας χ^2 , ο οποίος είναι πιο χρήσιμος για μεταβλητές μικτών τύπων (διατακτικές και κατηγορικές). Το Φ_K αντιμετωπίζει κάθε μεταβλητή ως κατηγορική. Αυτό είναι χρήσιμο όταν μελετάμε σύνολα δεδομένων με μικτούς τύπους μεταβλητών, όπου ορισμένες από τις μεταβλητές είναι κατηγορικές. Θα μπορούσαμε να πούμε ότι είναι μια γενίκευση του Φ (Phi) Coefficient, όπου σε σύγκριση με τον συντελεστή συσχέτισης Φ του Cramer, που είναι η μοναδική βιώσιμη εναλλακτική, η ισχύς του Φ_K είναι κατασκευαστικά πολύ παρόμοια με το r του Pearson, καθιστώντας τον εύκολο στην ερμηνεία. Ο υπολογισμός του είναι σχετικά ανθεκτικός στην ομαδοποίηση συνεχών μεταβλητών σε διαστήματα (binning) και το Φ_K περιλαμβάνει μια διόρθωση θορύβου έναντι των στατιστικών διακυμάνσεων, καθιστώντας το πιο σταθερό έναντι των αλλαγών στο μέγεθος του δείγματος. Ο τύπος είναι ο εξής (Yule, 1922):

$$\Phi_K = \sqrt{\frac{\chi^2}{n(k-1)}} \quad (11)$$

Όπου:

- χ^2 είναι η τιμή του στατιστικού chi-square από τον πίνακα συνάφειας.

- n είναι το συνολικό μέγεθος του δείγματος.
- k είναι ο αριθμός των κατηγοριών.

Αφού αναλύσαμε τις συσχετίσεις μεταξύ δύο μεταβλητών είναι απαραίτητο να προχωρήσουμε στην πολυμεταβλητή ανάλυση, όπου η συνδυασμένη επίδραση πολλών μεταβλητών μελετάται ταυτόχρονα. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε πιο σύνθετα δεδομένα, όπου οι σχέσεις μεταξύ περισσότερων από δύο μεταβλητών είναι σημαντικές.

2.1.3 Πολυμεταβλητή Ανάλυση Συσχέτισης

Η πολυμεταβλητή ανάλυση συσχέτισης επιτρέπει την ταυτόχρονη ανάλυση της επίδρασης πολλών ανεξάρτητων μεταβλητών σε μία εξαρτημένη μεταβλητή. Σύμφωνα με τον Tinbergen το 1947, η πολυμεταβλητή συσχέτιση είναι απαραίτητη σε περιπτώσεις όπου οι διακυμάνσεις σε μια μεταβλητή επηρεάζονται από περισσότερες από μία μεταβλητές. Η ανάλυση αυτή γίνεται συνήθως μέσω της μέθοδο των ελάχιστων τετραγώνων, η οποία εκτιμά τους συντελεστές παλινδρόμησης που μετρούν την επίδραση κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη. (Clark, 2013)

Στην ανάλυση αυτή, η σχέση μεταξύ της εξαρτημένης μεταβλητής Y και των ανεξάρτητων μεταβλητών X_i μπορεί να εκφραστεί μέσω μια εξίσωσης παλινδρόμησης (Lind et al., 2012):

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k \quad (12)$$

Όπου:

- Y είναι η εξαρτημένη μεταβλητή.
- a είναι η σταθερά (συντελεστής τομής).
- $b_1, b_2, \dots, b_k$ είναι οι συντελεστές παλινδρόμησης για τις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_k$
- k είναι ο αριθμός των ανεξάρτητων μεταβλητών.

Η πολυμεταβλητή ανάλυση συσχέτισης παρέχει τα εργαλεία για την κατανόηση της σχέσης μεταξύ πολλαπλών μεταβλητών ταυτόχρονα. Στα πλαίσια αυτής της ανάλυσης, είναι σημαντικό να ελέγχεται η πολυσυγγραμμικότητα (multicollinearity), καθώς μπορεί να επηρεάσει σημαντικά την ακρίβεια των εκτιμήσεων (Halkos & Tsilika, 2018).

Στην πολυμεταβλητή ανάλυση, η πολυσυγγραμμικότητα είναι ιδιαίτερα προβληματική γιατί εμποδίζει την ορθή ερμηνεία των σχέσεων μεταξύ των μεταβλητών. Οι εκτιμήσεις των συντελεστών παλινδρόμησης δεν είναι αξιόπιστες όταν υπάρχει πολυσυγγραμμικότητα, ενώ η εξαγωγή σημαντικών συμπερασμάτων καθίσταται δύσκολη (Imdadullah et al., 1983). Η πολυσυγγραμμικότητα προκύπτει όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές σε ένα μοντέλο παλινδρόμησης είναι πολύ ισχυρά συσχετισμένες μεταξύ τους. Αυτό δημιουργεί πρόβλημα στις εκτιμήσεις των συντελεστών της παλινδρόμησης, διότι καθιστά δύσκολο να διαχωριστεί η επίδραση της κάθε μεταβλητής στη μεταβλητή απόκρισης. Αυτό μπορεί να οδηγήσει σε:

1. Υψηλή αστάθεια στους συντελεστές της παλινδρόμησης: οι διακυμάνσεις και τα τυπικά σφάλματα των συντελεστών παλινδρόμησης μεγαλώνουν.
2. Μείωση της ακρίβειας των εκτιμήσεων: τα μεγέθη των εκτιμήσεων των συντελεστών παλινδρόμησης μπορεί να είναι διαφορετικά από αυτά που αναμένονται.
3. Δυσκολία στην ερμηνεία των συντελεστών: Τα πρόσημα των εκτιμήσεων των συντελεστών παλινδρόμησης μπορεί να είναι αντίθετα από αυτά που αναμένονται.
4. Η εισαγωγή ή αφαίρεση μεταβλητών προκαλεί μεγάλες αλλαγές στις εκτιμήσεις των συντελεστών ή στα πρόσημα τους.
5. Η αφαίρεση ενός σημείου δεδομένων προκαλεί μεγάλες αλλαγές στις εκτιμήσεις των συντελεστών ή στα πρόσημα τους.
6. Σε κάποιες περιπτώσεις, ο λόγος F είναι σημαντικός αλλά κανένας από τους λόγους t δεν είναι σημαντικός

Για την αντιμετώπιση της πολυσυγγραμμικότητας, χρησιμοποιούνται διάφορες μετρικές, όπως ο συντελεστής διόγκωσης της διακύμανσης (Variance Inflation Factor - VIF), οι ιδιοτιμές (Eigenvalues), η ανοχή (Tolerance -TOL) και ο δείκτης κατάστασης (Condition Index - CI).

Συντελεστής Διόγκωσης της Διακύμανσης (VIF):

Ο Συντελεστής Διόγκωσης της Διακύμανσης (Variance Inflation Factor VIF) είναι μια μετρική που χρησιμοποιείται για την ανίχνευση της πολυσυγγραμμικότητας. Ο VIF μετράει πόσο πολύ η διακύμανση ενός συντελεστή παλινδρόμησης έχει αυξηθεί λόγω της πολυσυγγραμμικότητας (Imdadullah et al., 1983). Μία υψηλή τιμή VIF υποδεικνύει ισχυρή συσχέτιση μεταξύ ανεξάρτητων μεταβλητών, η οποία μπορεί να επηρεάσει τα αποτελέσματα του μοντέλου. Ο VIF υπολογίζεται με βάση το συντελεστή προσδιορισμού R^2 (R-squared) της

παλινδρόμησης της συγκεκριμένης ανεξάρτητης μεταβλητής στις υπόλοιπες. Ένας VIF μεγαλύτερος από 10 συχνά υποδηλώνει σοβαρή πολυσυγγραμμικότητα, που μπορεί να επηρεάσει αρνητικά το μοντέλο (Chan et al., 2022). Μάλιστα οι τιμές του Variance Inflation Factor κατηγοριοποιούνται ως εξής:

- **0-5:** Θεωρείται ότι δεν υπάρχει πρόβλημα πολυσυγγραμμικότητας. Το επίπεδο συσχέτισης είναι χαμηλό και δεν επηρεάζει την αξιοπιστία των εκτιμήσεων.
- **5-10:** Υπάρχει κάποια ανησυχία για μέτρια πολυσυγγραμμικότητα. Ενδέχεται να επηρεαστεί η ακρίβεια των εκτιμήσεων, αλλά το πρόβλημα δεν είναι ιδιαίτερα σοβαρό.
- **>10:** Υποδεικνύεται υψηλή πολυσυγγραμμικότητα, η οποία ενδέχεται να προκαλέσει αστάθεια στις εκτιμήσεις και χρειάζεται να αντιμετωπιστεί (Belsley, 1991).

Ο μαθηματικός τύπος του Variance Inflation Factor δίνεται από την εξίσωση (Hanck et al., 2024):

$$VIF(X_i) = \frac{1}{1-R_i^2} \quad (13)$$

Όπου:

- R_i^2 είναι ο συντελεστής προσδιορισμού (coefficient of determination) που προκύπτει από την παλινδρόμηση της X_i ως συνάρτηση όλων των άλλων ανεξάρτητων μεταβλητών στο μοντέλο (Hanck et al., 2024).

Ιδιοτιμές (Eigenvalues):

Οι ιδιοτιμές (Eigenvalues) στη συσχέτιση είναι θεμελιώδης έννοια στην πολυμεταβλητή ανάλυση, ιδιαίτερα όταν πρόκειται για την κατανόηση της δομής των δεδομένων και της συσχέτισης μεταξύ των μεταβλητών. Η ανάλυση των ιδιοτιμών χρησιμοποιείται για να εντοπιστούν γραμμικές εξαρτήσεις μεταξύ των ανεξάρτητων μεταβλητών. Πολύ χαμηλές ιδιοτιμές υποδεικνύουν ισχυρή πολυσυγγραμμικότητα (Imdadullah et al., 1983).

Οι ιδιοτιμές των πινάκων συνδιακύμανσης ή συσχέτισης χρησιμοποιούνται επίσης για τον εντοπισμό της πολυσυγγραμμικότητας. Μικρές ιδιοτιμές υποδεικνύουν πιθανή πολυσυγγραμμικότητα, καθώς δείχνουν ότι οι ανεξάρτητες μεταβλητές είναι γραμμικά εξαρτημένες (Chan et al., 2022). Για τις ιδιοτιμές (Eigenvalues) ισχύει (Lin et al., 2021):

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0 \quad (14)$$

Όπου:

- λ_j είναι οι ιδιοτιμές του πίνακα συσχέτισης, και χρησιμοποιούνται για να διαγνώσουν την πολυδιάστατη συσχέτιση. Η μεγαλύτερη και η μικρότερη ιδιοτιμή αξιοποιούνται για την κατασκευή δεικτών όπως ο Condition Index (βλ. σχ. (16)).

Δείκτης Ανοχής (Tolerance):

Η ανοχή (Tolerance - TOL) είναι το αντίστροφο του VIF και μετρά το βαθμό στον οποίο μια ανεξάρτητη μεταβλητή δεν εξαρτάται από τις υπόλοιπες. Μία μικρή τιμή TOL, συνήθως μικρότερη από 0.1, υποδηλώνει υψηλή πολυσυγγραμμικότητα (Chan et al., 2022). Ο τύπος της ανοχής είναι (Ταμπάκης et al., 2013):

$$TC_{(i)} = \frac{1}{VIF} \quad (15)$$

Όπου:

- i υποδηλώνει τη συγκεκριμένη μεταβλητή για την οποία υπολογίζεται το VIF.

Δείκτης Κατάστασης (Condition Index - CI):

Ο Condition Index (CI) υπολογίζεται από τις ιδιοτιμές και χρησιμοποιείται για να ανιχνεύσει την πολυσυγγραμμικότητα ειδικά όταν οι ιδιοτιμές έχουν μεγάλο εύρος. Χρησιμοποιείται για την αξιολόγηση της σταθερότητας των παλινδρομήσεων σε περιπτώσεις όπου οι μεταβλητές έχουν υψηλό επίπεδο συσχέτισης (Salmerón et al., 2018). Ένας CI μεγαλύτερος από 15 δείχνει πιθανή πολυσυγγραμμικότητα, ενώ τιμές πάνω από 30 υποδηλώνουν σοβαρή πολυσυγγραμμικότητα (Chan et al., 2022). Πιο συγκεκριμένα προσφέρει μια κλίμακα που παρέχει μια σαφή καθοδήγηση για το πότε η πολυσυγγραμμικότητα μπορεί να είναι προβληματική, επιτρέποντας έτσι να λάβουμε τα απαραίτητα μέτρα. Η κλίμακα είναι ως εξής:

- **CI<10:** Υποδηλώνει ότι η πολυσυγγραμμικότητα είναι χαμηλή και δεν αναμένεται να επηρεάσει σημαντικά τα αποτελέσματα του μοντέλου.
- **10≤CI<30:** Υποδηλώνει μέτρια έως ισχυρή πολυσυγγραμμικότητα, όπου το πρόβλημα αρχίζει να γίνεται αισθητό και μπορεί να επηρεάσει τις εκτιμήσεις των παραμέτρων.

- **CI>30:** Υποδεικνύει υψηλή πολυσυγγραμμικότητα, γεγονός που μπορεί να επηρεάσει σοβαρά την ακρίβεια και τη σταθερότητα των εκτιμήσεων του μοντέλου (Callaghan & Chen, 2008).

Ο τύπος είναι ο εξής (Ταμπάκης, et al., 2013):

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \quad (16)$$

Όπου:

- λ_{max} : Η **μέγιστη ιδιοτιμή** είναι η μεγαλύτερη ιδιοτιμή που προκύπτει από τον πίνακα συσχέτισης ή συνδιακύμανσης. Αντιπροσωπεύει τη συνιστώσα που εξηγεί τη μεγαλύτερη ποσότητα διασποράς στα δεδομένα.
- λ_{min} : Η **ελάχιστη ιδιοτιμή** είναι η μικρότερη ιδιοτιμή (συνήθως κοντά στο μηδέν αν υπάρχει πολυδιάστατη συσχέτιση) και αντιπροσωπεύει τη συνιστώσα που εξηγεί τη λιγότερη ποσότητα διασποράς στα δεδομένα.

Αφού αναλύσαμε τις διάφορες μεθόδους συσχέτισης που μπορούν να χρησιμοποιηθούν σε αναλύσεις δύο ή περισσότερων μεταβλητών, είναι σημαντικό να εξετάσουμε πιο προηγμένες τεχνικές που επιτρέπουν τη μελέτη των γραμμικών σχέσεων μεταξύ δύο πολυδιάστατων συνόλων μεταβλητών. Μια τέτοια τεχνική είναι η Canonical Correlation Analysis (CCA), η οποία επιτρέπει την ανάλυση σύνθετων σχέσεων και την εξαγωγή γραμμικών συνδυασμών που ερμηνεύουν τη σχέση μεταξύ των συνόλων δεδομένων.

Canonical Correlation Analysis (CCA):

Πριν αναφερθούμε στην Canonical Correlation Analysis CCA είναι σημαντικό να αναφερθούμε στο θεμελιωτή και να αναφέρουμε πως οδηγήθηκε σε αυτήν την ανάπτυξη. Ο Harold Hotelling είναι ο πρωτοπόρος της Canonical Correlation Analysis (CCA), μιας στατιστικής μεθόδου που εισήγαγε το 1936. Η μέθοδος αυτή αναπτύχθηκε για να επιλύσει προβλήματα που περιλαμβάνουν την ανάλυση της σχέσης μεταξύ δυο ομάδων μεταβλητών. Ο Hotelling, μέσα από την εργασία "The Most Predictable Criterion", εστίασε στην εύρεση γραμμικών συνδυασμών από δύο ομάδες μεταβλητών που μεγιστοποιούν τη συσχέτιση μεταξύ τους (Hotelling, 1935). Πριν την ανάπτυξη της CCA, υπήρχαν μέθοδοι όπως η απλή γραμμική παλινδρόμηση, που περιοριζόταν στην ανάλυση σχέσεων μεταξύ μια εξαρτημένης και πολλών ανεξάρτητων μεταβλητών. Όμως, για πολλά πρακτικά προβλήματα, όπως η πρόβλεψη της

ακαδημαϊκής επιτυχίας ή του γενικού επιπέδου τιμών, αυτό δεν ήταν αρκετό. Ο Hotelling αναγνώρισε ότι υπήρχε ανάγκη για μια πιο γενική μέθοδο που να επιτρέπει την ταυτόχρονη ανάλυση δύο ομάδων πολλών μεταβλητών και να βρίσκει τους γραμμικούς συνδυασμούς που μεγιστοποιούν τη συσχέτιση μεταξύ αυτών των ομάδων. Η CCA στοχεύει στο να βρει ζεύγη γραμμικών συνδυασμών (U και V) από δύο ομάδες μεταβλητών, έτσι ώστε η συσχέτιση μεταξύ αυτών των γραμμικών συνδυασμών να είναι η μέγιστη δυνατή. Για παράδειγμα, αν έχουμε p μεταβλητές πρόβλεψης (predictors) και q κριτήρια (criteria), η μέθοδος βρίσκει τις γραμμικές συναρτήσεις αυτών των δύο ομάδων που είναι περισσότερο προβλέψιμες μεταξύ τους. Ο Hotelling βασίστηκε στις προηγούμενες εργασίες του στην "Ανάλυση Κύριων Συνιστωσών" (Principal Component Analysis – PCA), μια μέθοδο που ανέπτυξε για τη μείωση των διαστάσεων και την κατανόηση της δομής των δεδομένων. Στην πραγματικότητα, η CCA μπορεί να θεωρηθεί επέκταση της PCA, όπου αντί να επικεντρωθούμε στην εξήγηση της διακύμανσης σε μία ομάδα μεταβλητών, προσπαθούμε να εξηγήσουμε τη συσχέτιση μεταξύ δύο ομάδων μεταβλητών. Στην εργασία του "The Most Predictable Criterion", ο Hotelling (1935) παρουσίασε μια μεθοδολογία που περιλαμβάνει την επίλυση συστήματος εξισώσεων με τη βοήθεια των πολλαπλών συσχετίσεων και τη χρήση των ίδιων τιμών και των ίδιων διανυσμάτων. Αυτό επιτρέπει τον υπολογισμό των γραμμικών συνδυασμών των κριτηρίων και των μεταβλητών πρόβλεψης που έχουν την μέγιστη συσχέτιση μεταξύ τους (Hotelling, 1935). Όπως αναφερθήκαμε η Canonical Correlation Analysis (CCA) είναι μια πολυμεταβλητή στατιστική μέθοδος που μελετά τη γραμμική σχέση μεταξύ δύο πολυδιάστατων συνόλων μεταβλητών. Σε αντίθεση με τις παραδοσιακές μεθόδους συσχέτισης, η CCA μπορεί να αναλύσει πολλαπλές εξαρτήσεις ταυτόχρονα, καθιστώντας την ιδανική για την ανάλυση σύνθετων σχέσεων μεταξύ συνόλων δεδομένων όπως οι μακροοικονομικές μεταβλητές και οι αποδόσεις μετοχών (Mazuruse, 2014). Η CCA στοχεύει στην εύρεση γραμμικών συνδυασμών δύο συνόλων μεταβλητών που μεγιστοποιούν τη συσχέτιση μεταξύ τους. Έστω ότι έχουμε δύο σύνολα μεταβλητών:

Η CCA επιδιώκει να βρει γραμμικούς συνδυασμούς των μεταβλητών U και V

- $U = a_1X_1 + a_2X_2 + \dots + a_pX_p$
- $V = b_1Y_1 + b_2Y_2 + \dots + b_qY_q$

Οι συντελεστές a_i και b_j είναι αυτοί που θα πρέπει να εκτιμηθούν ώστε να μεγιστοποιείται η συσχέτιση μεταξύ των συνδυασμών U και V. Ο Hotelling (1936) υπολογίζει τη συσχέτιση μεταξύ των γραμμικών συνδυασμών U και V χρησιμοποιώντας τις συνδιακυμάνσεις των

μεταβλητών. Για να βρει τους βέλτιστους συντελεστές χρησιμοποιεί την εξίσωση (Hotelling, 1936):

$$R^2 = \frac{(r_x^T r_{xy} a)^2}{(a^T r_{xx} a)} \quad (\text{Υπολογισμός των Συσχετίσεων και των συντελεστών}) \quad (17)$$

Όπου:

- r_{xx} είναι ο πίνακας συσχέτισης των μεταβλητών πρόβλεψης.
- r_{yy} είναι ο πίνακας συσχέτισης των κριτηρίων³.
- r_{xy} είναι ο πίνακας διασταυρούμενων συσχετίσεων μεταξύ των μεταβλητών πρόβλεψης και των κριτηρίων.

Για να βρεθούν οι συντελεστές που μεγιστοποιούν τη συσχέτιση, ο Hotelling λύνει το παρακάτω σύστημα εξισώσεων (Hotelling, 1936):

$$(W_{a\beta} - \lambda r_{a\beta})a_a = 0 \quad (\text{Εξίσωση Κανονικών Συντελεστών}) \quad (18)$$

Όπου:

- $W_{a\beta}$ είναι ένας πίνακας που περιλαμβάνει τις συσχετίσεις μεταξύ των κριτηρίων και των μεταβλητών πρόβλεψης.
- λ είναι ένας πολλαπλασιαστής Lagrange

Η επίλυση αυτών των εξισώσεων γίνεται με τη εύρεση των ίδιων τιμών και ίδιων διανυσμάτων των πινάκων. Η μεγαλύτερη ιδιοτιμή δίνει τον συντελεστή που μεγιστοποιεί την κανονική συσχέτιση (Hotelling, 1936).

2.1.4 Σύνοψη

Μετά την ανάλυση των διαφόρων ειδών συσχέτισης, είναι σημαντικό να συνοψίσουμε τις κύριες μεθόδους και τα χαρακτηριστικά τους. Στις προηγούμενες ενότητες, εξετάστηκαν οι παραδοσιακοί συντελεστές συσχέτισης, όπως ο Pearson, ο Spearman ο Kendall's Tau, και πιο σύγχρονες μέθοδοι όπως η πολυμεταβλητή συσχέτιση. Κάθε μέθοδος έχει τα δικά της

³ Ο όρος κριτήρια αναφέρεται σε μεταβλητές που είναι εξαρτημένες ή στόχοι μιας ανάλυσης. Στο πλαίσιο της Canonical Correlation Analysis (CCA), τα κριτήρια είναι οι μεταβλητές που θέλουμε να εξηγήσουμε ή να προβλέψουμε. Επομένως, όταν λέμε "πίνακας συσχέτισης κριτηρίων", εννοούμε έναν πίνακα που δείχνει πως οι διάφορες εξαρτημένες μεταβλητές συσχετίζονται μεταξύ τους.

πλεονεκτήματα και μειονεκτήματα, ανάλογα με τον τύπο των δεδομένων και τις σχέσεις που πρέπει να αναλυθούν. Στον πίνακα 1 παρουσιάζονται συγκεντρωτικά οι κύριοι τύποι συσχέτισης, παρέχοντας μια συνοπτική επισκόπηση της καταλληλότητας τους, καθώς και των βασικών πλεονεκτημάτων και μειονεκτημάτων τους.

ΠΙΝΑΚΑΣ 1: Συνοπτική Παρουσίαση των Συντελεστών Συσχέτισης

Είδος Συσχέτισης	Έτος	Στόχος	Πλεονεκτήματα	Μειονεκτήματα
Συντελεστής Συσχέτισης Pearson	1895	Μέτρηση γραμμικής σχέσης μεταξύ δύο ποσοτικών μεταβλητών	Εύκολη κατανόηση και εφαρμογή, κατάλληλη για γραμμικές σχέσεις	Ευαισθησία σε ακραίες τιμές και απαιτεί κανονικότητα
Συντελεστής Συσχέτισης Spearman	1904	Μέτρηση μονοτονικής σχέσης με χρήση των τάξεων δεδομένων	Ανθεκτική σε ακραίες τιμές, κατάλληλη για μη γραμμικές σχέσεις	Χαμηλότερη ακρίβεια σε σχέση με άλλες μεθόδους όταν πληρούνται οι παραδοχές της κανονικότητας
Συντελεστής Συσχέτισης Kendall's Tau	1938	Μέτρηση τάξης μονοτονικών σχέσεων	Περισσότερο ανθεκτική σε ακραίες τιμές από τον Spearman	Περιορίζεται σε μικρές σχέσεις τάξης
Hoeffding's D⁴	1948	Μέτρηση γραμμικών, μονοτονικών και μη μονοτονικών σχέσεων	Ανίχνευση πιο πολύπλοκων δομών εξάρτησης πέρα από τις μονοτονικές	Δυσκολία στην ερμηνεία των αποτελεσμάτων λόγω της πολυπλοκότητας
Distance Correlation⁵	2007	Μέτρηση της συσχέτισης μέσω των αποστάσεων	Κατάλληλη για μη γραμμικές και σύνθετες σχέσεις	Περίπλοκος υπολογισμός και αυξημένες

⁴ Hoeffding, W. (1948). A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics*, 19(4), 546–557. DOI: 10.1214/aoms/1177730150.

⁵ Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794. DOI: 10.1214/009053607000000505.

		μεταξύ παρατηρήσεων		απαιτήσεις σε υπολογιστικούς πόρους
Maximal Information Coefficient⁶	2011	Ανίχνευση ισχυρών συσχετίσεων ανεξάρτητα από τη μορφή τους	Κατάλληλη για μεγάλες και πολύπλοκες σχέσεις δεδομένων	Δύσκολη ερμηνεία και υψηλή πολυπλοκότητα στον υπολογισμό
Συντελεστής Συσχέτισης Φ (Phi Coefficient)	1900	Μέτρηση συσχέτισης μεταξύ δύο δυναμικών μεταβλητών	Κατάλληλη για δυναμικές μεταβλητές, εύκολη υπολογιστικά	Περιορισμένη σε δυναμικά δεδομένα, δεν εφαρμόζεται σε περισσότερες κατηγορίες
Συντελεστής Συσχέτισης Τάξης Cramer's v⁷	1946	Μέτρηση συσχέτισης μεταξύ κατηγορικών μεταβλητών	Ιδανική για κατηγορικές μεταβλητές με πολλά επίπεδα	Περιορίζεται σε κατηγοριοποιημένα δεδομένα, δεν μπορεί να εφαρμοστεί σε συνεχείς μεταβλητές
Συντελεστής Συσχέτισης Πολλαπλών Συσχετίσεων (Multiple Correlation Coefficient)	20ός αιώνας	Μέτρηση της συσχέτισης μεταξύ μιας εξαρτημένης και πολλαπλών ανεξάρτητων μεταβλητών	Επιτρέπει την ανάλυση της συνδυασμένης επίδρασης πολλαπλών ανεξάρτητων μεταβλητών	Δεν λαμβάνει υπόψη τις αλληλεπιδράσεις μεταξύ των ανεξάρτητων μεταβλητών
Συντελεστής Συσχέτισης	1936	Μέτρηση συσχέτισης μεταξύ δύο	Ιδανική για πολυδιάστατες συσχετίσεις, αναλύει σχέσεις	Υψηλή πολυπλοκότητα απαιτεί μεγάλη ποσότητα

⁶ Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., ... & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062), 1518–1524.

⁷ Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Canonical Correlation⁸		συνόλων μεταβλητών	μεταξύ συνόλων μεταβλητών	δεδομένων (10 έως 20 παρατηρήσεις ανά μεταβλητή)
--	--	-----------------------	---------------------------------	--

ΚΕΦΑΛΑΙΟ 3: ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΣΥΣΧΕΤΙΣΗΣ ΣΤΗΝ ΕΡΕΥΝΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

3.1 Βιβλιογραφική ανασκόπηση

Στο παρόν κεφάλαιο, θα εξετάσουμε τις εφαρμογές των υπολογιστικών μεθόδων συσχέτισης στην έρευνα της οικονομικής επιστήμης. Οι μέθοδοι συσχέτισης αποτελούν ένα ισχυρό εργαλείο για την κατανόηση των πολύπλοκων σχέσεων μεταξύ μεταβλητών, όπως οι τιμές προϊόντων, τα επιτόκια, η απόδοση επενδύσεων και η επίδραση οικονομικών πολιτικών. Μέσα από την ανάλυση αυτών των σχέσεων, μπορούμε να εξάγουμε σημαντικά συμπεράσματα που επηρεάζουν τη λήψη αποφάσεων, τη χάραξη στρατηγικών, και τη βελτίωση της απόδοσης σε διάφορους τομείς της οικονομίας.

Οι συντελεστές συσχέτισης, που αναφέραμε στο προηγούμενο κεφάλαιο έχουν ευρεία εφαρμογή στα οικονομικά δεδομένα. Σε αυτό το κεφάλαιο, θα παρουσιάσουμε πώς αυτοί οι συντελεστές χρησιμοποιούνται για την ανάλυση των σχέσεων μεταξύ διαφορετικών οικονομικών παραγόντων. Ιδιαίτερα θα αναλυθούν συγκεκριμένες περιπτώσεις από τη βιβλιογραφία, όπου η χρήση αυτών των μεθόδων συνέβαλε στην κατανόηση και πρόβλεψη κρίσιμων οικονομικών φαινομένων.

Στη συνέχεια, θα παρουσιαστούν παραδείγματα υπολογιστικών εφαρμογών των μεθόδων συσχέτισης στην οικονομική επιστήμη, περιλαμβάνοντας τη μελέτη των συσχετίσεων μεταξύ επενδύσεων και κερδοφορίας, καθώς και την ανάλυση των παραγόντων που επηρεάζουν τη ζήτηση προϊόντων και υπηρεσιών. Μέσα από τις αναλύσεις αυτές, θα δούμε πώς οι διάφορες υπολογιστικές τεχνικές σε μέτρα συσχέτισης μπορούν να συμβάλουν στη λήψη στρατηγικών αποφάσεων και στη βελτιστοποίηση της αποδοτικότητας των επιχειρηματικών μοντέλων. Στο πλαίσιο αυτό, πραγματοποιήθηκε αναζήτηση εντός πέντε διακεκριμένων περιοδικών που εντάσσονται στη θεματική περιοχή των Υπολογιστικών Οικονομικών. Καθώς

⁸ Hotelling, H. (1936). *Relations Between Two Sets of Variates*. *Biometrika*, 28(3/4), 321–377. DOI: 10.2307/2333955.

η εργασία ασχολείται με μέτρα συσχέτισης, με τεχνικές αναλυτικής δεδομένων εστιάζουμε σε περιοδικά με υπολογιστικές μελέτες και όχι σε θεωρητικές μελέτες. Οι αναζητήσεις πραγματοποιήθηκαν στην αγγλική γλώσσα και διεξήχθησαν στη βάση ακαδημαϊκών δεδομένων Scopus. Η βάση Scopus είναι ένα από τα καταλληλότερα εργαλεία βάσης δεδομένων για τους σκοπούς του άρθρου, καθώς σύμφωνα με του δημιουργούς της έχει πάνω από 14.000 τίτλους STM (επιστήμης, τεχνολογίας και ιατρικής) και κοινωνικών επιστημών από 4.000 εκδότες, αναφέροντας ότι είναι η μεγαλύτερη μεμονωμένη βάση δεδομένων με ευρετήρια και περιλήψεις που έχει ποτέ δημιουργηθεί. Εκτός από αμερικανικά περιοδικά, περιλαμβάνει ευρωπαϊκή και ασιατική βιβλιογραφία τόσο στα αγγλικά όσο και σε άλλες γλώσσες, καλύπτοντας δημοσιεύσεις από το 1788 έως και σήμερα, καθιστώντας την μία από τις μεγαλύτερες επιμελημένες βάσεις δεδομένων βιβλιογραφικών περιλήψεων και αναφορών στον κόσμο.

Η ανάλυση εφαρμόστηκε στα δεδομένα του Scopus, όπως προέκυψαν από μια σύνθετη αναζήτηση η οποία αποτελείται από το πεδίο “Source title” (όπου βάλαμε καθένα από τα πέντε επιστημονικά περιοδικά) σε συνδυασμό με το πεδίο “keywords” (όπου βάλαμε τη λέξη κλειδί “Correlations”). Η χρονική περίοδος της αναζήτησης καλύπτει τα έτη 1994-2024. Τα αποτελέσματα της σύνθετης αναζήτησης συνοπτικά παρουσιάζονται στον Πίνακα 2.

Τα άρθρα αυτά επιλέχθηκαν βάσει της συμβολής τους στην κατανόηση των συσχετίσεων που αφορούν τομείς όπως η ανάλυση χρηματοοικονομικών στοιχείων, η πρόβλεψη απόδοσης επενδύσεων και η αξιολόγηση στρατηγικών διοίκησης. Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα αυτής της βιβλιογραφικής ανασκόπησης, όπου περιλαμβάνονται σημαντικές εφαρμογές των συντελεστών συσχέτισης για την ανάλυση σύνθετων οικονομικών και διοικητικών φαινομένων. Σε αυτό το σημείο θα κάνουμε μια σύντομη αναφορά στα 5 περιοδικά απ’ όπου αντλήσαμε την βιβλιογραφία από το 1994 έως το 2024. Στον Πίνακα 3 παρουσιάζονται ενδελεχώς τα αποτελέσματα αυτής της βιβλιογραφικής ανασκόπησης για το ημερολογιακό έτος 2024⁹.

Computational Economics (Springer)

Το περιοδικό Computational Economics επικεντρώνεται στη χρήση υπολογιστικών μεθόδων για τη κατανόηση και επίλυση σύνθετων οικονομικών προβλημάτων. Είναι ένα επιστημονικό περιοδικό που συνδυάζει υπολογιστικές τεχνικές με κλάδους της οικονομίας. Το

⁹ Η αναζήτηση που έδωσε τα αποτελέσματα του πίνακα 2 και πίνακα 3 έγινε στις 3/11/2024.

περιοδικό δημοσιεύει ερευνητικά άρθρα, κριτικές για λογισμικό, καθώς και ειδικές εκδόσεις αφιερωμένες σε συγκεκριμένα θέματα. Είναι επίσημο περιοδικό της Society for Computational Economics, και αναφέρεται επίσημα ως (Comput Econ) το οποίο ενσωματώνει την υπολογιστική επιστήμη σε όλες τις πτυχές της οικονομικής έρευνας, και προσφέρει μια πλατφόρμα για καινοτόμες προσεγγίσεις στην ανάλυση δεδομένων.

Journal of Applied Econometrics (Wiley)

Είναι ένα διεθνές επιστημονικό περιοδικό που δημοσιεύει εργασίες υψηλής ποιότητας, οι οποίες εφαρμόζουν υπάρχουσες αλλά και νέες οικονομετρικές τεχνικές σε ένα ευρύ φάσμα προβλημάτων της οικονομίας. Το περιοδικό επικεντρώνεται κυρίως στις πρακτικές εφαρμογές της οικονομετρίας παρά σε θεωρητικά ζητήματα. Μία από τις σημαντικές του απαιτήσεις είναι η διαφάνεια των δεδομένων. Από το 1994, απαιτεί από τους συγγραφείς να καταθέτουν πλήρη σύνολα δεδομένων (αν δεν είναι εμπιστευτικά) στο αρχείο δεδομένων του περιοδικού, έτσι ώστε να επιτρέπεται η αναπαραγωγή των αποτελεσμάτων από άλλους ερευνητές. Το περιοδικό ιδρύθηκε το 1986, δημοσιεύεται από τον εκδοτικό οίκο John Wiley & Sons και εκδίδεται 7 φορές το χρόνο.

Journal of Economic Dynamics and Control (JEDC) (Elsevier)

Το Journal of Economic Dynamics and Control (JEDC) είναι ένα επιστημονικό περιοδικό με κριτές που εστιάζει στην υπολογιστική οικονομική, στα δυναμικά οικονομικά μοντέλα, και στα μακροοικονομικά. Ξεκίνησε την έκδοση το 1979 και δημοσιεύεται από τον εκδοτικό οίκο Elsevier. Το περιοδικό ασχολείται με θέματα όπως οι υπολογιστικές μέθοδοι και τα δυναμικά στοχαστικά γενικής ισορροπίας υποδείγματα (DSGE models), αλλά και με την εφαρμογή στατιστικής φυσικής στα οικονομικά χρηματοοικονομικά. Το περιοδικό μερικές φορές αφιερώνει ειδικά τεύχη σε συγκεκριμένα θέματα, όπως «Πολυπλοκότητα στα Οικονομικά και Χρηματοοικονομικά», «Εφαρμογές της Στατιστικής Φυσικής στα Οικονομικά και Χρηματοοικονομικά»

Στοχεύει στη δημοσίευση τόσο θεωρητικών όσο και εμπειρικών μελετών, ενώ δίνει ιδιαίτερη έμφαση στη χρήση προηγμένων υπολογιστικών μεθόδων για την επίλυση προβλημάτων στα οικονομικά και τα χρηματοοικονομικά.

Journal of Economic Interaction and Coordination (Springer)

Το Journal of Economic Interaction and Coordination είναι ένα επιστημονικό περιοδικό, που, από το 2006 που ιδρύθηκε επικεντρώνεται στις κοινωνικές επιστήμες και οικονομικές.

Περιλαμβάνει δηλαδή πειραματικά οικονομικά, οικονομοφυσική, μη γραμμική δυναμική, πολύπλοκα προσαρμοστικά συστήματα. Οι δημοσιεύσεις του περιοδικού εξετάζουν φαινόμενα όπως η ανάδυση συμπεριφορών και η συλλογική δράση σε οικονομικά και κοινωνικά συστήματα.

Το περιοδικό φιλοξενεί δημοσιεύσεις από ποικίλους κλάδους όπως η οικονομία, η φυσική, και η επιστήμη των υπολογιστών, εστιάζοντας στην προσομοίωση και ανάλυση των αλληλεπιδράσεων μεταξύ πρακτόρων και στη συμπεριφορά αγορών με ετερογενείς παράγοντες. Επίσης, ενθαρρύνονται οι δημοσιεύσεις που παρουσιάζουν πειραματικές μελέτες και έρευνες που βασίζονται σε θεωρητικά μοντέλα.

ACM Transactions on Economics and Computation (TEAC) (Association for Computing Machinery (ACM))

Το περιοδικό ACM Transactions on Economics and Computation (TEAC) εστιάζει στη διασταύρωση της επιστήμης των υπολογιστών και της οικονομίας. Καλύπτει θέματα όπως οι υπολογιστικές πτυχές των οικονομικών αλληλεπιδράσεων, οι οικονομικές αγορές και οι μηχανισμοί δημοπρασιών, καθώς και η θεωρία παιγνίων και οι αλγόριθμοι κατανομής πόρων ενδιαφέρον για το περιοδικό είναι οποιοδήποτε θέμα σχετίζεται τόσο με οικονομολόγους όσο και με επιστήμες υπολογιστών. Το περιοδικό δημοσιεύει έρευνες που χρησιμοποιούν υπολογιστικά εργαλεία και οικονομικές ιδέες για να κατανοήσουν και να βελτιώσουν τις κοινωνικές και οικονομικές αλληλεπιδράσεις από το 2013 έως και σήμερα.

Τα άρθρα TEAC συχνά ασχολούνται με την οικονομική θεωρία, χρησιμοποιώντας επίσημους μαθηματικούς συλλογισμούς και υπολογιστικά μοντέλα για να αναλύσουν προβλήματα στην οικονομία και τη διαχείριση πόρων. Το περιοδικό συνδέει τους τομείς της υπολογιστικής, των μαθηματικών, της οικονομίας και της χρηματοοικονομικής για να προσφέρει καινοτόμες λύσεις σε προβλήματα κατανομής πόρων και κοινωνικής οργάνωσης.

ΠΙΝΑΚΑΣ 2: Αποτελέσματα Σύνθετης Αναζήτησης (Scopus)

The literature search results ¹⁰			
Scientific publishing knowledge base	Keywords	Source title	Search results (1980-2024)
		Computational Economics	158 Documents (1994-2024)

¹⁰ Ημερομηνία αναζήτησης για τα αποτελέσματα του πίνακα 2: 03/11/2024

Scopus	Correlations	Journal of Applied Econometrics	85 Documents (1988-2024)
		Journal of Economic Dynamics and Control	125 Documents (1980-2024)
		Journal of Economic Interaction and Coordination	26 Documents (2006-2024)
		ACM Transactions on Economics and Computation	2 Documents (2020-2023)

ΠΙΝΑΚΑΣ 3: Αποτελέσματα στην ανάλυση συσχέτισης, subject area: (Economics, Econometrics and Finance, Computer Science, Business, Management and Accounting) Scopus 2024¹¹.

Τίτλος άρθρου	Είδος συσχέτισης	Πεδίο εφαρμογής	Υπολογιστικό εργαλείο εκτίμησης συσχέτισης
Computational Economics			
Zhao, Q., Wang, Z. & Song, Y. Systematic Research on Multi-dimensional and Multiple Correlation Contagion Networks of Extreme Risk in China's Banking Industry. <i>Comput Econ</i> 64, 1137–1162 (2024).	Πολλαπλή ανάλυση συσχέτισης (Multiple Correlation Analysis)	Τραπεζικός κλάδος	Γλώσσα προγραμματισμού Matlab
Lucchetti, R., Pedini, L. The Spherical Parametrisation for Correlation Matrices and its Computational Advantages. <i>Comput Econ</i> 64, 1023–1046 (2024)	Pearson correlation	Οικονομετρία και μαθηματική ανάλυση	Gretl ¹²
The Relationship Between Non-additivity Valuations, Cash Flows and Sales Growth (2024)	Non-addictivity correlation	Χρηματοοικονομική ανάλυση	Γλώσσα προγραμματισμού Matlab
Musaev, A., Makshanov, A. & Grigoriev, D. Multi-regression Forecast in Stochastic Chaos. <i>Comput Econ</i> 64, 137–160 (2024)	Πολλαπλή ανάλυση συσχέτισης (Multiple Correlation Analysis)	Χρηματοοικονομική ανάλυση (Forex)	Γλώσσα προγραμματισμού R
Ghosh, S., Tiwari, A.K., Doğan, B. <i>et al.</i> The Dynamic Relationship Between Gas and Crude Oil Markets	Cross-Quantilogram	Μάκρο-Μίκρο οικονομικός κλάδος	Δεν αναφέρεται

¹¹ Ημερομηνία αναζήτησης πίνακα 3: 03/11/2024

¹² Gretl: Ανοιχτό λογισμικό για οικονομετρική ανάλυση, το οποίο χρησιμοποιήθηκε για την εκτίμηση των μοντέλων και την αριθμητική βελτιστοποίηση που παρουσιάζεται.

and the Causal Impact of US Shale Gas. <i>Comput Econ</i> 63 , 2501–2524 (2024)	Correlation (CQC) ¹³		
Sahiner, M. Volatility Spillovers and Contagion During Major Crises: An Early Warning Approach Based on a Deep Learning Model. <i>Comput Econ</i> 63 , 2435–2499 (2024)	Dynamic Conditional Correlation (DCC) ¹⁴	Χρηματοοικονομική ανάλυση	Γλώσσα προγραμματισμού Python
Su, E., Mak, VV. & So, PY. Exploring Three-style Return Comovements and Contagion Using a Correlation Decomposition GARCH Model. <i>Comput Econ</i> 63 , 2271–2305 (2024)	Dynamic Conditional Correlation (DCC)	Χρηματοοικονομική ανάλυση	Γλώσσα προγραμματισμού R με το πακέτο quantmod
Mesquita, C.M., Valle, C.A. & Pereira, A.C.M. Scenario Generation for Financial Data with a Machine Learning Approach Based on Realized Volatility and Copulas. <i>Comput Econ</i> 63 , 1879–1919 (2024)	Kendall's T correlation	Χρηματοοικονομική ανάλυση	Γλώσσα προγραμματισμού R και Python με τις βιβλιοθήκες Keras, Tensorflow και scikit-learn
He, H., Li, H. A New Boosting Algorithm for Online Portfolio Selection Based on dynamic Time Warping and Anti-correlation. <i>Comput Econ</i> 63 , 1777–1803 (2024)	Αντί συσχέτισης (anti-correlation) ¹⁵ και της μέσης επαναφοράς (mean reversion) ¹⁶	Χρηματοοικονομική ανάλυση	Δεν αναφέρεται

¹³ Cross-Quantilogram Correlation (CQC): είναι στατιστικό εργαλείο που χρησιμοποιείται για να εξετάσει τη συσχέτιση ή τη σχέση εξάρτησης μεταξύ δύο χρονικών σειρών σε διάφορα ποσοστιαία σημεία της κατανομής (quantiles). Σε αντίθεση με τις παραδοσιακές μεθόδους συσχέτισης το CQC επιτρέπει την ανάλυση της σχέσης σε διαφορετικά μέρη της κατανομής, αντί μόνο για τη μέση τιμή.

¹⁴ Dynamic Conditional Correlation (DCC): Επιτρέπει στις συσχετίσεις να μεταβάλλονται με την πάροδο του χρόνου. Έτσι αναγνωρίζει ότι οι συσχετίσεις μεταξύ χρηματοοικονομικών μεταβλητών δεν είναι σταθερές, αλλά εξαρτώνται από τις συνθήκες της αγοράς.

¹⁵ Αντί συσχέτιση (anti-correlation): Στατιστική μέτρηση που περιγράφει τη σχέση μεταξύ δύο μεταβλητών όταν κινούνται σε αντίθετες κατευθύνσεις.

¹⁶ Μέση επαναφορά (mean reversion): είναι μια οικονομική και στατιστική θεωρία που υποστηρίζει ότι οι τιμές των περιουσιακών στοιχείων, όπως οι μετοχές ή οι αγορές, τείνουν να επιστρέφουν στη μακροχρόνια μέση ή φυσιολογική τους τιμή μετά από περιόδους αποκλίσεων.

Jo, G., Kim, H., Kim, H. et al. Fuzzy Portfolio Selection Using Stochastic Correlation. <i>Comput Econ</i> 63, 1493–1509 (2024).	Στοχαστική συσχέτιση (stochastic correlation)	Χρηματοοικονομική ανάλυση	Γλώσσα προγραμματισμού Matlab
Daglis, T. The Tourism Industry's Performance During the Years of the COVID-19 Pandemic. <i>Comput Econ</i> 63, 1173–1189 (2024)	Cross Correlation	Τουριστικό τομέα	Δεν αναφέρεται
Liang, K., Li, S., Zhang, W. et al. Evolution of Complex Network Topology for Chinese Listed Companies Under the COVID-19 Pandemic. <i>Comput Econ</i> 63, 1121–1136 (2024)	Pearson Correlation Coefficient	Χρηματοοικονομικό τομέα	Γλώσσα προγραμματισμού Python
Rahmani, A.M., Hosseini Mirmahaleh, S.Y. An Intelligent Algorithm to Predict GDP Rate and Find a Relationship Between COVID-19 Outbreak and Economic Downturn. <i>Comput Econ</i> 63, 1001–1020 (2024)	Spearman Correlation Coefficient	Μακροοικονομικό τομέα	Δεν αναφέρεται
Lee, H., Ha, H., Lee, G. et al. Two-Asset Double Barrier Options. <i>Comput Econ</i> (2024)	Δεν αναφέρεται άμεση συσχέτιση χρησιμοποιεί Probabilistic Methods για pricing options	Χρηματοοικονομικό τομέα	Γλώσσα προγραμματισμού R

Ullah, M., Sohag, K., Doroshenko, S. <i>et al.</i> Examination of Bitcoin Hedging, Diversification and Safe-Haven Ability During Financial Crisis: Evidence from Equity, Bonds, Precious Metals and Exchange Rate Markets. <i>Comput Econ</i> (2024)	GARCH, E-GARCH VAR, CVAR	Χρηματοοικονομικό τομέα	Γλώσσα προγραμματισμού R
Oprea, SV., Bâra, A. Assessing the Dual Impact of the Social Media Platforms on Psychological Well-being: A Multiple-Option Descriptive-Predictive Framework. <i>Comput Econ</i> (2024)	Pearson Correlation Coefficient	Κοινωνικό-ψυχολογικό τομέα	Γλώσσα προγραμματισμού Python
Yue, J., Wang, MH., Huang, NJ. <i>et al.</i> Asset Prices with Investor Protection in the Cross-Sectional Economy. <i>Comput Econ</i> (2024)	Pearson Correlation Coefficient	Χρηματοοικονομικό τομέα	Γλώσσα προγραμματισμού Matlab
Lucchetti, R., Pedini, L. Correction to: The Spherical Parametrisation for Correlation Matrices and its Computational Advantages. <i>Comput Econ</i> (2024)	Spherical Parametrisation for Correlation Matrice ¹⁷	Μαθηματικά-Οικονομικό τομέα	Δεν αναφέρεται
Gong, X., Jia, G. The Impact of Financial Stress on New Energy Vehicles Industry from Cross-correlation to Explainable Machine Learning: Proof from China. <i>Comput Econ</i> (2024)	Multifractal Detrended Cross-Correlation Analysis (MFDCCA), SHAP	Χρηματοοικονομικό τομέα	Γλώσσα προγραμματισμού Python

¹⁷ Spherical Parametrisation for Correlation Matrices: Είναι μια μέθοδος που χρησιμοποιείται για την αναπαράσταση πινάκων συσχέτισης με παραμέτρους. Σκοπός είναι να διασφαλίσει ότι οι πίνακες συσχέτισης που δημιουργούνται είναι έγκυροι.

Vo, M. Measuring and Forecasting Stock Market Volatilities with High-Frequency Data. <i>Comput Econ</i> (2024)	Heterogeneous Autoregressive Models (HAR)	Χρηματοοικονομικό τομέα	Δεν αναφέρεται
Jumamyradov, M., Craig, B.M., Greene, W.H. <i>et al.</i> Comparing the Mixed Logit Estimates and True Parameters under Informative and Uninformative Heterogeneity: A Simulated Discrete Choice Experiment. <i>Comput Econ</i> (2024)	Correlation in Mixed Logit Models ¹⁸	Επιχειρηματικό τομέα	Δεν αναφέρεται
Cui, H., Wang, X. & Chu, X. Stock Returns Prediction Based on Implied Volatility Spread Under Network Perspective. <i>Comput Econ</i> (2024)	Pearson Correlation Coefficient	Χρηματοοικονομικό τομέα	Δεν αναφέρεται
Lee, S. Measuring Interdependence of Inflation Uncertainty. <i>Comput Econ</i> (2024)	Pearson Correlation Coefficient, Spearman Correlation Coefficient, Kendall's Tau Correlation	Μακροοικονομικό τομέα	Γλώσσα προγραμματισμού R
Arslan, S. Bitcoin Price Prediction Using Sentiment Analysis and Empirical Mode Decomposition. <i>Comput Econ</i> (2024)	Δεν αναφέρεται συγκεκριμένη συσχέτιση, χρησιμοποιεί Empirical Mode Decomposition (EMD) και Long Short-Term Memory (LSTM)	Χρηματοοικονομικό τομέα (Κρυπτονομίσματα)	Γλώσσα προγραμματισμού Python (με TextBlob, LSTM)

¹⁸ Correlation in Mixed Logit Models: Αναφέρεται στη συσχέτιση μεταξύ των παραμέτρων ενός μοντέλου μικτής λογιστικής παλινδρόμησης (Mixed Logit).

Touli, E.F., Nguyen, H. & Bodnar, O. Monitoring the Dynamic Networks of Stock Returns with an Application to the Swedish Stock Market. <i>Comput Econ</i> (2024)	Pearson Correlation	Χρηματοοικονομικό τομέα (αγορές μετοχών)	Δεν αναφέρεται
Ling, J., Zhong, Z. & Wei, H. Copper Price Forecasting Based on Improved Least Squares Support Vector Machine with Butterfly Optimization Algorithm. <i>Comput Econ</i> (2024)	Principal Component Analysis (PCA)	Οικονομικό τομέα (Εμπόρευμα Χαλκού)	Δεν αναφέρεται
Moudud-Ul-Huq, S., Rahman, M.S. Stock Market Efficiency of the BRICS Countries Pre-, During, and Post Covid-19 Pandemic: A Multifractal Detrended Fluctuation Analysis. <i>Comput Econ</i> (2024)	Multifractal Detrended Fluctuation Analysis (MF-DFA), Hurst Exponen	Χρηματοοικονομικό τομέα (αγορές μετοχών)	Γλώσσα προγραμματισμού R Studio
Momparler, A., Carmona, P. & Climent, F. Catalyzing Sustainable Investment: Revealing ESG Power in Predicting Fund Performance with Machine Learning. <i>Comput Econ</i> (2024)	Δεν αναφέρει άμεσα συσχέτιση	Χρηματοοικονομικό τομέα	Γλώσσα προγραμματισμού R με πακέτα (XGBoost, h2o, DALEX)
Is the European refugee crisis a potential threat to public health? Evidence from Italy	NO ACCESS		
Wang, J., Jiang, W., Huang, M. <i>et al.</i> Cross-Correlation Analysis of Crude Oil-Related Stock Markets in China Caused by the Conflict Between Russia and Ukraine. <i>Comput Econ</i> (2024)	Cross-Correlation Analysis	Χρηματοοικονομικό τομέα	Γλώσσα προγραμματισμού Matlab

de Salles, D.S., Gea, C., Mello, C.E. <i>et al.</i> Multi-Scale Event Detection in Financial Time Series. <i>Comput Econ</i> (2024)	Δεν αναφέρεται συγκεκριμένη συσχέτιση (χρησιμοποιεί CEEMD για ανάλυση) ¹⁹	Χρηματοοικονομικό τομέα (Χρηματιστήριο)	Δεν αναφέρεται
Niveditha, P.S. Identifying Safe Haven Assets: Evidence from Fractal Market Hypothesis. <i>Comput Econ</i> (2024)	Wavelet Quantile Correlation (WQC), Continuous Wavelet Transform (CWT)	Χρηματοοικονομικά τομέα (Ασφάλειας)	Δεν αναφέρεται
Xing, K., Li, H. Market Ecology: Trading Strategies and Market Volatility. <i>Comput Econ</i> (2024)	Trend-following and Value Strategies	Χρηματοοικονομικό τομέα (Στρατηγικές συναλλαγών)	Δεν αναφέρεται
Ekinci, A., Sen, S. Forecasting Bank Failure in the U.S.: A Cost-Sensitive Approach. <i>Comput Econ</i> (2024)	αλγόριθμοι ML όπως CSForest, XGBoost, Logistic Regression	Τραπεζικός τομέας	Γλώσσα προγραμματισμού WEKA (για τα περισσότερα μοντέλα) και R (για XGBoost)
Wang, Z., Wang, S. & Yan, Y. Sieve Bootstrap for Fixed- <i>b</i> Phillips–Perron Unit Root Test. <i>Comput Econ</i> (2024)	Serial Correlation Analysis ²⁰	Μακροοικονομικός τομέας	Γλώσσα προγραμματισμού Matlab
Aiolfi, R., Moreni, N., Bianchetti, M. <i>et al.</i> Learning Bermudans. <i>Comput Econ</i> (2024)	Interest Rate, Swaptions	Χρηματοοικονομικό τομέα (Παράγωγα προϊόντα swaptions)	Γλώσσα προγραμματισμού Python (scikit-learn, TensorFlow, Keras)
Neděla, D., Ortobelli Lozza, S. & Tichý, T. Dynamic Return Scenario Generation Approach for Large-Scale Portfolio Optimisation Framework. <i>Comput Econ</i> (2024)	Pearson Correlation	Χρηματοοικονομικό τομέα (Βελτιστοποίηση Χαρτοφυλακίου)	Γλώσσα προγραμματισμού Matlab

¹⁹ Complete Ensemble Empirical Mode Decomposition with Adaptive Noise): Βελτιωμένη παραλλαγή της Empirical Mode Decomposition, που χρησιμοποιείται για την ανάλυση μη γραμμικών και μη στατικών σημάτων.

²⁰ Serial Correlation Analysis: Ανάλυση σχέσης ανάμεσα στις τιμές μιας χρονοσειράς με τις προηγούμενες τιμές της.

Su, K., Yao, Y., Zheng, C. <i>et al.</i> Portfolio Selection Based on EMD Denoising with Correlation Coefficient Test Criterion. <i>Comput Econ</i> 63 , 391–421 (2024)	Correlation Coefficient Test ²¹	Χρηματοοικονομικό τομέα	Δεν αναφέρεται
Müller, F.M., Righi, M.B. Comparison of Value at Risk (VaR) Multivariate Forecast Models. <i>Comput Econ</i> 63 , 75–110 (2024)	DCC-GARCH, GO-GARCH, Copulas (C-Vine, D-Vine, R-Vine)	Χρηματοοικονομικός τομέας	Γλώσσα προγραμματισμού R (για copulas), MATLAB (για GARCH)
Mikhaylov, A., Bhatti, I.M., Dinçer, H. <i>et al.</i> Integrated decision recommendation system using iteration-enhanced collaborative filtering, golden cut bipolar for analyzing the risk-based oil market spillovers. <i>Comput Econ</i> 63 , 305–338 (2024)	Pearson Correlation	Χρηματοοικονομικό τομέα	Δεν αναφέρεται
Afuecheta, E., Okorie, I.E., Nadarajah, S. <i>et al.</i> Forecasting Value at Risk and Expected Shortfall of Foreign Exchange Rate Volatility of Major African Currencies via GARCH and Dynamic Conditional Correlation Analysis. <i>Comput Econ</i> 63 , 271–304 (2024)	DCC-GARCH, Copula	Χρηματοοικονομικό τομέα (Συναλλαγματικές ισοτιμίες)	Γλώσσα προγραμματισμού R
Li, Z., Yang, Y. Directed association network analysis on the Standard and Poor's 500 Index. <i>Comput Econ</i> 63 , 111–127 (2024)	Asymmetric Correlation Coefficient (Chatterjee, 2020)	Χρηματοοικονομικά, Χρηματιστήριο	Δεν αναφέρεται

²¹ Correlation Coefficient Test Criterion: Είναι ένα στατιστικό κριτήριο που χρησιμοποιείται για να αξιολογηθεί η ύπαρξη και η σημασία της συσχέτισης μεταξύ δύο ή περισσότερων μεταβλητών.

Journal of Applied Econometrics			
Statistical identification in panel structural vector autoregressive models based on independence criteria	Structural Vector Autoregression (SVAR) με χρήση ανάλυσης ανεξαρτησίας (independent component analysis)	Χρηματοοικονομικά (νομισματική πολιτική στην Ευρωζώνη)	Γλώσσα προγραμματισμού R
Sample selection in linear panel data models with heterogeneous coefficients	Linear Panel Data Models με ετερογενείς συντελεστές	Οικονομετρικά μοντέλα για την ανάλυση επιλογής δειγμάτων σε δεδομένα panel	Γλώσσα προγραμματισμού R (Στατιστικές τεχνικές δεδομένων πάνελ)
Journal of Economic Dynamics and Control			
A contagion test with unspecified heteroscedastic errors	Διατομεακές συσχετίσεις (cross-sectional correlations)	Οικονομική δυναμική και έλεγχοι για χρηματοοικονομική μόλυνση, Διεθνής κρίσεις	Γλώσσα προγραμματισμού R
A quantitative theory of the new life cycle of women's employment	Reduced-form Correlation (θετική συσχέτιση μεταξύ των αποφάσεων απασχόλησης και γονιμότητας) Δεν αναφέρει κάποιο συγκεκριμένο μέτρο συσχέτισης.	Μακροοικονομική ανάλυση	Δεν αναφέρεται
Multinational entry and exit, technology transfer, and international business cycles	Global-International Correlation	Διεθνής επιχειρηματικοί κύκλοι	Γλώσσα προγραμματισμού Python
Missing growth and economic fluctuations: Empirical evidence from Korea	Phillips correlation ²²	Μακροοικονομική ανάλυση των οικονομικών διακυμάνσεων στην Κορέα	Στατιστικό πακέτο Stata

²² Ο Phillips Correlation δεν είναι ένας επίσημος στατιστικός όρος, μερικές φορές χρησιμοποιείται ανεπίσημα για να περιγράψει τη σχέση που καταγράφεται από την καμπύλη Phillips δηλαδή, τη συσχέτιση μεταξύ πληθωρισμού και ανεργίας ή άλλων οικονομικών μεταβλητών.

Do unemployment benefit extensions explain the emergence of jobless recoveries	Beveridge curve ²³	Οικονομετρική ανάλυση	Dynare
Journal of Economic Interaction and Coordination			
Zema, S.M., Fagiolo, G., Squartini, T. <i>et al.</i> Mesoscopic structure of the stock market and portfolio optimization. <i>J Econ Interact Coord</i> (2024). https://doi.org/10.1007/s11403-024-00426-y	Μεσοσκοπική συσχέτιση (mesoscopic correlation)	Χρηματοοικονομικές αγορές και χαρτοφυλάκια	Γλώσσα προγραμματισμού R.

3.2 Εμπειρική διερεύνηση με εργαλεία της αναλυτικής δεδομένων

Μετά την παρουσίαση της βιβλιογραφικής ανασκόπησης η οποία περιλαμβάνει σημαντικές εφαρμογές των συντελεστών συσχέτισης στην ανάλυση σύνθετων οικονομικών και διοικητικών φαινομένων, είναι σημαντικό να προχωρήσουμε στην υπολογιστική προσέγγιση των μέτρων συσχέτισης μέσω της γλώσσας προγραμματισμού R και στην πλατφόρμα αναλυτικής KNIME, δύο εργαλεία ανάλυσης δεδομένων ανοικτού κώδικα. Αυτά τα εργαλεία επιτρέπουν την εκτέλεση υπολογισμών και γραφικών παραστάσεων, δίνοντας μια σαφή εικόνα της σχέσης μεταξύ διάφορων ειδών μεταβλητών.

Στόχος της ανάλυσης είναι η εφαρμογή των συντελεστών συσχέτισης που αναφέραμε στο κεφάλαιο 2, μέσω διαφορετικών πακέτων της R, αλλά και μέσα από την πλατφόρμα αναλυτικής δεδομένων KNIME ώστε να εξεταστούν οι εναλλακτικοί τρόποι υπολογισμού και γραφικής απεικόνισης. Κάθε μέτρο συσχέτισης θα διερευνηθεί με πολλαπλά πακέτα και με διαφορετικές μεθόδους γραφικής απεικόνισης, ώστε να αναδειχθούν οι διαφορές μεταξύ των προσεγγίσεων.

3.2.1 Υπολογιστική Προσέγγιση των μέτρων συσχέτισης μέσω της γλώσσας υπολογιστικής στατιστικής R

Στην παρούσα υπό ενότητα εφαρμόζονται τα μέτρα συσχέτισης που παρουσιάστηκαν στο θεωρητικό μέρος, αξιοποιώντας την ευελιξία και την υπολογιστική ισχύ της γλώσσας

²³ Η καμπύλη Beveridge εξετάζει την αρνητική συσχέτιση μεταξύ της ανεργίας και των κενών θέσεων εργασίας, αλλά δεν είναι στατιστικό μέτρο συσχέτισης όπως τα μέτρα συσχέτισης που αναφέραμε στο δεύτερο κεφάλαιο.

προγραμματισμού R. Μέσα από τα κατάλληλα πακέτα, πραγματοποιείται η στατιστική ανάλυση και η απεικόνιση των σχέσεων μεταξύ των μεταβλητών.

Συντελεστής Συσχέτισης Pearson:

Στην ανάλυση του Pearson Correlation Coefficient, θα εξετάσουμε ενδεικτικά τις σχέσεις μεταξύ των βασικών οικονομικών μεταβλητών (Ακαθάριστου Εγχώριου Προϊόν, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Όγκος λιανικών πωλήσεων) της Ευρωπαϊκής Ένωσης. Τα δεδομένα συλλέχθηκαν από την βάση δεδομένων FRED (Federal Reserve Bank)²⁴, η οποία παρέχει αξιόπιστα μακροοικονομικά δεδομένα. Η γλώσσα προγραμματισμού R επιτρέπει την άμεση πρόσβαση στα δεδομένα μέσω της εντολής `'fimport'`. Οι μεταβλητές αφορούν την περίοδο 1/1/2005 έως 1/10/2022 και τα δεδομένα που έχουμε συλλέξει είναι τριμηνιαία. Με την εντολή `str(df)` όπως μπορούμε να δούμε και στην εικόνα 3, το data frame αποτελείται από 69 παρατηρήσεις (γραμμές) και 6 μεταβλητές (στήλες). Επίσης εμφανίζεται ο τύπος δεδομένων της κάθε μεταβλητής, ο οποίος είναι αριθμητικός (numeric), καθώς και τις πρώτες τιμές που περιέχει κάθε μεταβλητή. Επιπλέον στην εικόνα 4 παρατηρούμε τις πρώτες γραμμές του data frame (df). Στην περίπτωση μας εμφανίζονται οι τιμές για κάθε μεταβλητή, οργανωμένες ανά ημερομηνία. Αυτό μπορεί να μας βοηθήσει στο να αποκτήσουμε μια πρώτη εικόνα των δεδομένων και να ελέγξουμε αν το data frame έχει διαμορφωθεί σωστά. Συγκεκριμένα βλέπουμε τα δεδομένα των μεταβλητών (ΑΕΠ, Ανεργία, Εισαγωγές, Εξαγωγές) για την Ευρωπαϊκή Ένωση.

²⁴ Τα δεδομένα προήλθαν από το FRED- Federal Reserve Economic Data και είναι διαθέσιμα στον σύνδεσμο (<https://fred.stlouisfed.org>)

CLVMEURSCAB1GQEA19=GDP_EU #Real Gross Domestic Product for Euro Area (19 Countries)
LRUN64TTEZQ156S=UNEMPLOYMENT_EU
NAEXCP07EZQ189S=IMPORTS_EU
NAEXCP06EZQ189=EXPORTS_EU
NAEXKP03EZQ652S=INVESTMENT_EU
SLRTO01EZQ661S=RETAIL_SALES_EU
CLVMNACSCAB1GQEL=GDP_GR
LRUN64TTGRQ156=UNEMPLOYMENT_GR
NAEXKP07GRQ652S=IMPORTS_GR
NAEXKP06GRQ652S=EXPORTS_GR
NAEXKP03GRQ657S=HOUSEHOLD_CONSUMPTION_GR

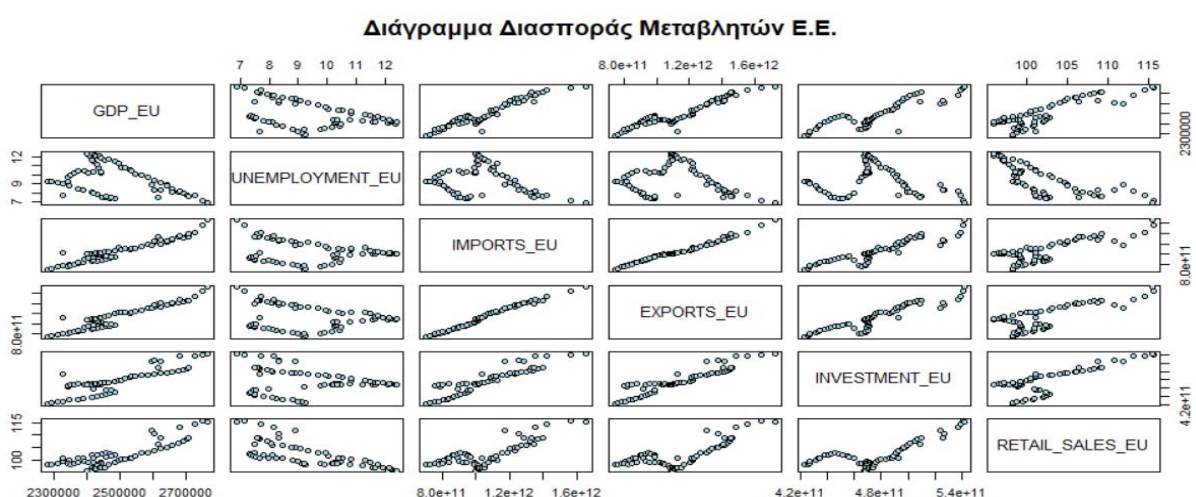
```
> str(df)
'data.frame': 69 obs. of 6 variables:
 $ GDP_EU      : num  2279118 2293174 2310520 2326340 2347951 ...
 $ UNEMPLOYMENT_EU: num   9.21 9.23 9.1 8.98 8.83 ...
 $ IMPORTS_EU    : num  7.01e+11 7.26e+11 7.53e+11 7.80e+11 8.08e+11 ...
 $ EXPORTS_EU    : num  7.42e+11 7.60e+11 7.84e+11 8.04e+11 8.29e+11 ...
 $ INVESTMENT_EU : num  4.23e+11 4.26e+11 4.27e+11 4.27e+11 4.32e+11 ...
 $ RETAIL_SALES_EU: num  98.2 98.3 98.7 99.4 99.8 ...
```

ΕΙΚΟΝΑ 3: Επισκόπηση δομής data frame με την εντολή str(df) στη γλώσσα προγραμματισμού R.

```
> head(df)
      GDP_EU UNEMPLOYMENT_EU  IMPORTS_EU  EXPORTS_EU INVESTMENT_EU RETAIL_SALES_EU
2005-01-01 2279118          9.212245 700599548350 742181573906 423178873549      98.24152
2005-04-01 2293174          9.229578 725570339249 759743050700 425833786685      98.34153
2005-07-01 2310520          9.095285 753117581886 783639903360 426693080953      98.74156
2005-10-01 2326340          8.981615 779522944747 804336477590 427332743638      99.44162
2006-01-01 2347951          8.829857 807832535782 828567572819 431907457370      99.84165
2006-04-01 2373848          8.462947 827309913125 852525483529 433204110629     100.70839
```

ΕΙΚΟΝΑ 4: Επισκόπηση δομής πλαισίου δεδομένων (data frame) με την εντολή head(df) στη γλώσσα προγραμματισμού R.

Πριν προβούμε στον υπολογισμό του συντελεστή συσχέτισης Pearson, δημιουργήσαμε ένα διάγραμμα διασποράς (scatterplot matrix), όπως φαίνεται στην εικόνα 5 ώστε να αποκτήσουμε μια εικόνα των σχέσεων μεταξύ των μεταβλητών. Τα διαγράμματα αυτά παρουσιάζουν κάθε πιθανό ζεύγος μεταβλητών του πίνακα δεδομένων και απεικονίζουν την κατανομή και τη μορφή της συσχέτισης τους. Το διάγραμμα διασποράς μας βοηθά να διακρίνουμε πιθανά πρότυπα, γραμμικότητες ή και ακραίες τιμές, ώστε να επιβεβαιώσουμε αν η συσχέτιση Pearson είναι κατάλληλη για τις μεταβλητές μας.



ΕΙΚΟΝΑ 5: Διάγραμμα Διασποράς (Scatterplot Matrix) μεταξύ βασικών οικονομικών μεταβλητών της Ελλάδας και της Ευρωπαϊκής Ένωσης (ΑΕΠ, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Όγκος λιανικών πωλήσεων) στη γλώσσα προγραμματισμού R.

Ο υπολογισμός του συντελεστή συσχέτισης Pearson θα γίνει με τα εξής πακέτα:

Συνάρτηση cor(): Είναι ο πιο απλός τρόπος υπολογισμού του συντελεστή συσχέτισης Pearson στην R είναι μέσω της ενσωματωμένης συνάρτησης cor(), η οποία υπολογίζει τον πίνακα συσχέτισης μεταξύ των μεταβλητών. Το αποτέλεσμα μπορούμε να το δούμε και στην εικόνα 6, που παρουσιάζει τον πίνακα συσχέτισης Pearson, ο οποίος δείχνει τον βαθμό γραμμικής σχέσης μεταξύ των μεταβλητών που επιλέξαμε.

```
> print(pearson_cor)
      GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU      1.0000000      -0.4243271    0.9540744    0.9478579    0.8592764
UNEMPLOYMENT_EU -0.4243271      1.0000000    -0.2909802   -0.2292004   -0.1966536
IMPORTS_EU      0.9540744    -0.2909802    1.0000000    0.9924605    0.9136212
EXPORTS_EU      0.9478579   -0.2292004    0.9924605    1.0000000    0.9211930
INVESTMENT_EU    0.8592764   -0.1966536    0.9136212    0.9211930    1.0000000
RETAIL_SALES_EU 0.8507759   -0.7201406    0.7888951    0.7603720    0.7618905
      RETAIL_SALES_EU
GDP_EU      0.8507759
UNEMPLOYMENT_EU -0.7201406
IMPORTS_EU      0.7888951
EXPORTS_EU      0.7603720
INVESTMENT_EU    0.7618905
RETAIL_SALES_EU    1.0000000
```

ΕΙΚΟΝΑ 6: Υπολογισμός Pearson Correlation Coefficient με την συνάρτηση cor στην γλώσσα προγραμματισμού R.

Πακέτο Hmisc: Η εικόνα 7 παρουσιάζει τον πίνακα συσχέτισης Pearson που υπολογίστηκε με το πακέτο Hmisc. Το συγκεκριμένο πακέτο παρέχει παρόμοια αποτελέσματα με την συνάρτηση cor(), καθώς επίσης δίνει μια επιπλέον πληροφορία που είναι η εξαγωγή των p-values. Τα p-values χρησιμοποιούνται για να διαπιστωθεί αν η συσχέτιση είναι στατιστικά σημαντική. Οι συντελεστές συσχέτισης με τιμές των p-values μικρότερες από 0.05 είναι στατιστικά σημαντικές.

```
> cor_results
      GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU RETAIL_SALES_EU
GDP_EU      1.00      -0.42      0.95      0.95      0.86      0.85
UNEMPLOYMENT_EU -0.42      1.00      -0.29      -0.23      -0.20      -0.72
IMPORTS_EU      0.95      -0.29      1.00      0.99      0.91      0.79
EXPORTS_EU      0.95      -0.23      0.99      1.00      0.92      0.76
INVESTMENT_EU    0.86      -0.20      0.91      0.92      1.00      0.76
RETAIL_SALES_EU 0.85      -0.72      0.79      0.76      0.76      1.00

n= 69

P
      GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU RETAIL_SALES_EU
GDP_EU      0.0003      0.0000      0.0000      0.0000      0.0000      0.0000
UNEMPLOYMENT_EU 0.0003      0.0153      0.0000      0.0582      0.1053      0.0000
IMPORTS_EU      0.0000      0.0153      0.0000      0.0000      0.0000      0.0000
EXPORTS_EU      0.0000      0.0582      0.0000      0.0000      0.0000      0.0000
INVESTMENT_EU    0.0000      0.1053      0.0000      0.0000      0.0000      0.0000
RETAIL_SALES_EU 0.0000      0.0000      0.0000      0.0000      0.0000      0.0000
```

ΕΙΚΟΝΑ 7: Υπολογισμός Pearson Correlation Coefficient με το πακέτο Hmisc (Frank, 2025στη γλώσσα προγραμματισμού R.

Πακέτο Psych: Η εικόνα 8 παρουσιάζει τον πίνακα συσχέτισης Pearson, που υπολογίστηκε με το πακέτο psych. Το συγκεκριμένο πακέτο παρέχει παρόμοια αποτελέσματα με την συνάρτηση

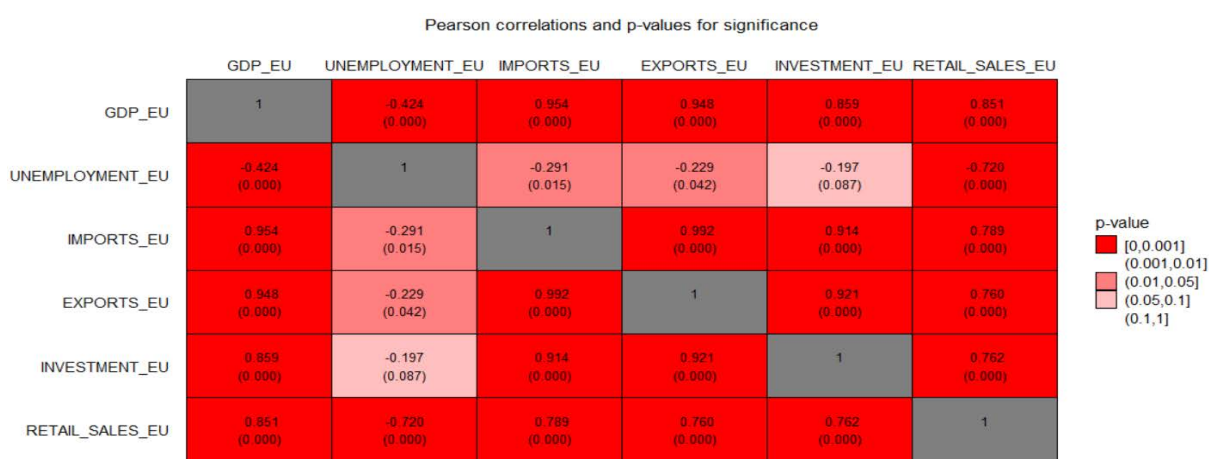
cor(), ίδια αποτελέσματα με το πακέτο Hmisc καθώς επίσης δίνει μια επιπλέον πληροφορία που είναι η εξαγωγή των p-values.

```
> cor_results
Call:corr.test(x = df, method = "pearson")
Correlation matrix
      GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU RETAIL_SALES_EU
GDP_EU      1.00      -0.42      0.95      0.95      0.86      0.85
UNEMPLOYMENT_EU -0.42      1.00     -0.29     -0.23     -0.20     -0.72
IMPORTS_EU      0.95     -0.29      1.00      0.99      0.91      0.79
EXPORTS_EU      0.95     -0.23      0.99      1.00      0.92      0.76
INVESTMENT_EU   0.86     -0.20      0.91      0.92      1.00      0.76
RETAIL_SALES_EU 0.85     -0.72      0.79      0.76      0.76      1.00
Sample Size
[1] 69
Probability values (Entries above the diagonal are adjusted for multiple tests.)
      GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU RETAIL_SALES_EU
GDP_EU      0      0.00      0.00      0.00      0.00      0
UNEMPLOYMENT_EU 0      0.00      0.05      0.12      0.12      0
IMPORTS_EU      0      0.02      0.00      0.00      0.00      0
EXPORTS_EU      0      0.06      0.00      0.00      0.00      0
INVESTMENT_EU   0      0.11      0.00      0.00      0.00      0
RETAIL_SALES_EU 0      0.00      0.00      0.00      0.00      0

To see confidence intervals of the correlations, print with the short=FALSE option
```

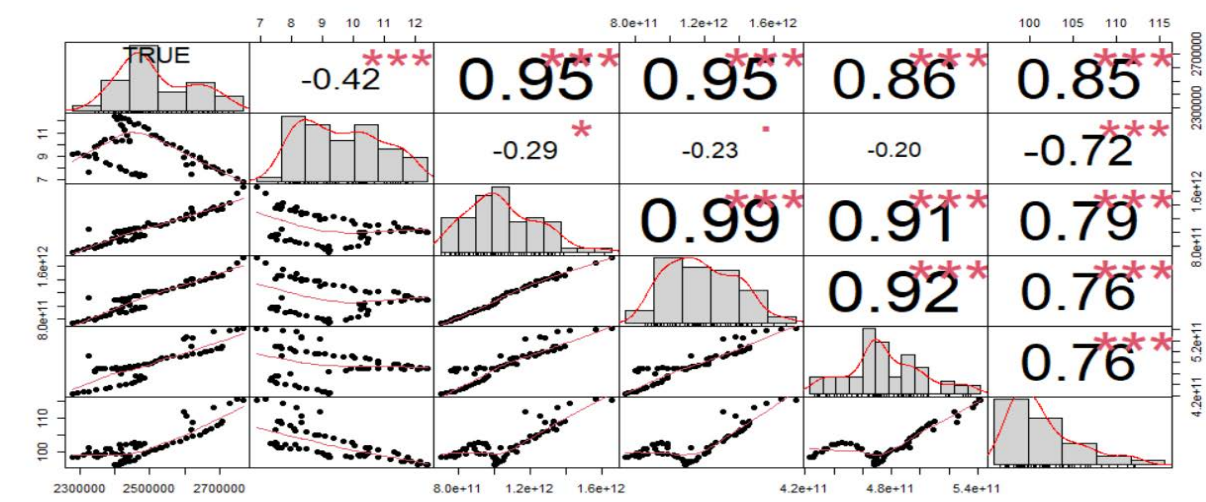
EIKONA 8: Υπολογισμός Pearson Correlation Coefficient με το πακέτο Psych (William, 2025) στη γλώσσα προγραμματισμού R.

Πακέτο Testcorr: Η εικόνα 9 παρουσιάζει τον πίνακα συσχέτισης Pearson που υπολογίστηκε με το πακέτο testcorr μέσω της συνάρτησης rcorr.test(). Το πακέτο υπολογίζει τους συντελεστές συσχέτισης Pearson και παράλληλα παρέχει τα αντίστοιχα p-values, διευκολύνοντας την αξιολόγηση της στατιστικής σημαντικότητας κάθε συσχέτισης. Επιπλέον, προσφέρει οπτική αναπαράσταση των αποτελεσμάτων μέσω διαβαθμισμένων πινάκων που ενισχύουν την κατανόηση των σχέσεων μεταξύ των μεταβλητών. Το testcorr αποτελεί ένα ιδιαίτερο εργαλείο όταν εξετάζονται πολλές μεταβλητές ταυτόχρονα, καθώς συγκεντρώνει σε έναν πίνακα τόσο τα αποτελέσματα των συσχετίσεων όσο και των στατιστικών ελέγχων.



EIKONA 9: Συνοπτική Απεικόνιση Συσχετίσεων Pearson και Επιπέδων Σημαντικότητας με το πακέτο testcorr (Dalla et al., 2020) στη γλώσσα προγραμματισμού R.

Chart Correlation (PerformanceAnalytics): Η εικόνα 10 απεικονίζει την συσχέτιση των μεταβλητών χρησιμοποιώντας τη συνάρτηση `chart.Correlation()` από το πακέτο `PerformanceAnalytics`. Οι τιμές που είναι πάνω από την κύρια διαγώνιο είναι οι συντελεστές συσχέτισης Pearson και τα αστεράκια δείχνουν τις στατιστικά σημαντικές σχέσεις. Η κύρια διαγώνιος δείχνει την κατανομή συχνότητας κάθε μεταβλητής. Κάτω από την κύρια διαγώνιο μας δείχνει οπτικά τη σχέση μεταξύ δύο μεταβλητών. Αν η διασπορά των σημείων σχηματίζει ευθεία γραμμή, τότε η σχέση είναι γραμμική.



ΕΙΚΟΝΑ 10: Γραφική παράσταση του Pearson με `chart correlation` (Peterson & Carl, 2022) στη γλώσσα προγραμματισμού R.

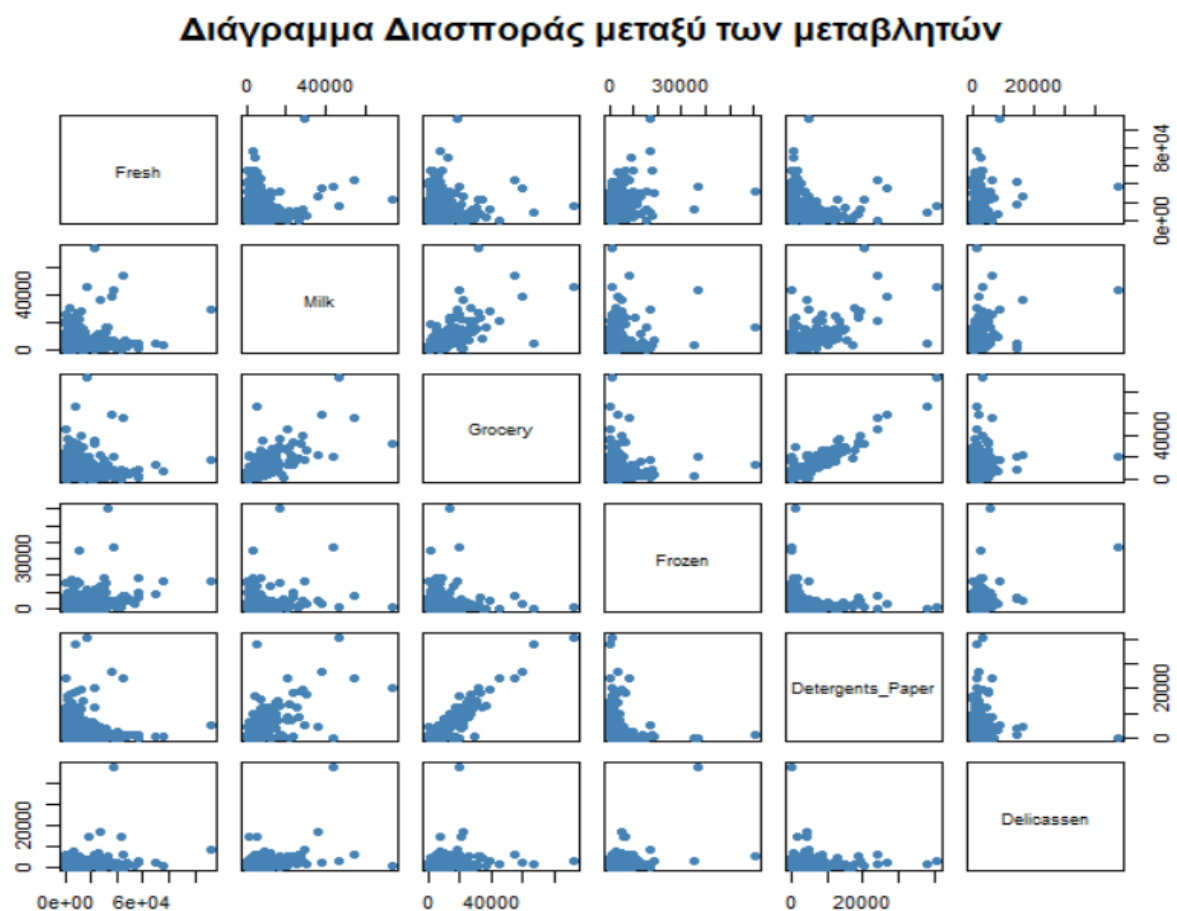
Συντελεστής Συσχέτισης *Biweight mid (Bicor)*:

Για την ενίσχυση της αξιοπιστίας των αποτελεσμάτων, εφαρμόσαμε τον συντελεστή συσχέτισης *Biweight Midcorrelation (bicor)*, ο οποίος αποτελεί μια ανθεκτική (robust) εναλλακτική του Pearson, λιγότερο ευαίσθητη στην παρουσία ακραίων τιμών (outliers). Για την εφαρμογή του συγκεκριμένου συντελεστή χρησιμοποιήσαμε το dataset “Wholesale customers”²⁵ από το UCI Machine Learning Repository, το οποίο περιλαμβάνει πληροφορίες για τις ετήσιες αγορές πελατών σε διάφορες κατηγορίες προϊόντων. Επιλέχθηκαν έξι ποσοτικές μεταβλητές: Fresh, Milk, Grocery, Frozen, Detergents_Paper και Delicassen, καθώς το συγκεκριμένο σύνολο δεδομένων παρουσιάζει έντονες ακραίες τιμές και αποκλίσεις. Η ανάλυση πραγματοποιήθηκε στη γλώσσα R με τη χρήση του πακέτου `correlation` και της

²⁵ Πηγή δεδομένων: UCI Machine Learning Repository – Wholesale customers Data ser.
Διαθέσιμο στον σύνδεσμο: <https://archive.ics.uci.edu/dataset/292/wholesale+customers>.

εντολής `correlation(df, method = "biweight")`, η οποία επιστρέφει τις τιμές συσχέτισης και τα αντίστοιχα επίπεδα σημαντικότητας.

Πριν την παρουσίαση των αποτελεσμάτων παρατίθεται το διάγραμμα διασποράς μεταξύ των μεταβλητών Fresh, Milk Grocery Frozen, Detergents_Paper και Delicassen. Το διάγραμμα αυτό μας δίνει μια πρώτη οπτική εκτίμηση της μορφής των σχέσεων μεταξύ των μεταβλητών, επιτρέποντας μας να εντοπίσουμε ενδεχόμενα μοτίβα ή αποκλίσεις που μπορεί να επηρεάσουν τον υπολογισμό του συντελεστή. Η γραφική αυτή αναπαράσταση συμβάλει στην καλύτερη κατανόηση της κατανομής και της συσχέτισης των μεταβλητών πριν προχωρήσουμε στην εφαρμογή του συντελεστή συσχέτισης Biweight mid.



EIKONA 11: Διάγραμμα Διασποράς (Scatterplot Matrix) μεταξύ μεταβλητών των μεταβλητών (Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen) από dataset “Wholesale Customers”

Στην εικόνα 12 παρουσιάζονται τα αποτελέσματα του Biweight Midcorrelation Coefficient.

```
> bicor_result <- correlation(df, method = "biweight", p_adjust = "none")
> bicor_result
# Correlation Matrix (biweight-method)
```

Parameter1	Parameter2	r	95% CI	t(438)	p
Fresh	Milk	-0.08	[-0.17, 0.02]	-1.61	0.108
Fresh	Grocery	-0.10	[-0.19, 0.00]	-2.01	0.046*
Fresh	Frozen	0.29	[0.20, 0.37]	6.27	< .001***
Fresh	Detergents_Paper	-0.07	[-0.16, 0.02]	-1.47	0.143
Fresh	Delicassen	0.22	[0.12, 0.30]	4.61	< .001***
Milk	Grocery	0.72	[0.67, 0.76]	21.63	< .001***
Milk	Frozen	-0.11	[-0.20, -0.01]	-2.23	0.026*
Milk	Detergents_Paper	0.44	[0.36, 0.51]	10.25	< .001***
Milk	Delicassen	0.28	[0.19, 0.37]	6.14	< .001***
Grocery	Frozen	-0.14	[-0.23, -0.05]	-3.02	0.003**
Grocery	Detergents_Paper	0.57	[0.50, 0.63]	14.43	< .001***
Grocery	Delicassen	0.24	[0.15, 0.33]	5.24	< .001***
Frozen	Detergents_Paper	-0.12	[-0.21, -0.03]	-2.51	0.012*
Frozen	Delicassen	0.14	[0.04, 0.23]	2.89	0.004**
Detergents_Paper	Delicassen	0.17	[0.07, 0.26]	3.52	< .001***

```
p-value adjustment method: none
Observations: 440
```

EIKONA 12: Υπολογισμός του συντελεστή συσχέτισης Biweight Mid Correlation Coefficient με το πακέτο correlation (Makowski et al., 2020).

Συντελεστής Συσχέτισης *Percentage Bend*:

Για την διερεύνηση της γραμμικής σχέσης μεταξύ των μεταβλητών, εφαρμόσαμε τον συντελεστή Percentage Bend Correlation χρησιμοποιώντας και εδώ τις ίδιες μακροοικονομικές μεταβλητές με εκείνη της ανάλυσης συσχέτισης με τον συντελεστή Biweight. Πρόκειται για μια ανθεκτική (robust) παραλλαγή του Pearson, η οποία μειώνει την επιρροή των πιο ακραίων παρατηρήσεων, βασισμένη σε ένα ποσοστιαίο όριο "καμπής" των τιμών. Έτσι, το 10% των χαμηλότερων και το 10% των υψηλότερων τιμών δεν επηρεάζουν ουσιαστικά τον υπολογισμό της συσχέτισης, γεγονός που προσφέρει πιο αξιόπιστα αποτελέσματα σε περιπτώσεις όπου τα δεδομένα παρουσιάζουν θόρυβο ή αποκλίσεις από την κανονικότητα. Η ένταξη του συντελεστή συσχέτισης Percentage Bend στην μελέτη των μέτρων συσχέτισης αναδεικνύει ότι υπάρχουν πολλαπλές μεθοδολογικές προσεγγίσεις για την εκτίμηση γραμμικών συσχετίσεων.

```
> percentage_bend_results <- correlation(df, method = "percentage", p_adjust = "none")
> percentage_bend_results
# Correlation Matrix (percentage-method)
```

Parameter1	Parameter2	r	95% CI	t(438)	p
Fresh	Milk	-0.11	[-0.20, -0.02]	-2.34	0.020*
Fresh	Grocery	-0.14	[-0.23, -0.05]	-2.92	0.004**
Fresh	Frozen	0.35	[0.26, 0.43]	7.73	< .001***
Fresh	Detergents_Paper	-0.26	[-0.35, -0.17]	-5.71	< .001***
Fresh	Delicassen	0.23	[0.14, 0.32]	5.03	< .001***
Milk	Grocery	0.78	[0.74, 0.81]	26.13	< .001***
Milk	Frozen	-0.10	[-0.19, -0.01]	-2.13	0.034*
Milk	Detergents_Paper	0.70	[0.64, 0.74]	20.29	< .001***
Milk	Delicassen	0.34	[0.26, 0.42]	7.65	< .001***
Grocery	Frozen	-0.17	[-0.26, -0.08]	-3.65	< .001***
Grocery	Detergents_Paper	0.86	[0.83, 0.88]	34.63	< .001***
Grocery	Delicassen	0.29	[0.20, 0.37]	6.30	< .001***
Frozen	Detergents_Paper	-0.26	[-0.34, -0.17]	-5.57	< .001***
Frozen	Delicassen	0.22	[0.13, 0.31]	4.68	< .001***
Detergents_Paper	Delicassen	0.16	[0.07, 0.25]	3.41	< .001***

p-value adjustment method: none
Observations: 440

EIKONA 13: Υπολογισμός του συντελεστή συσχέτισης Percentage Bend με το πακέτο correlation (Makowski et al., 2020).

Partial Correlation Coefficient:

Στον συντελεστή συσχέτισης Partial θα εξετάσουμε την σχέση μεταξύ των ίδιων βασικών οικονομικών μεταβλητών της Ευρωπαϊκής Ένωσης (Ακαθάριστο Εγχώριο Προϊόν, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Όγκος Λιανικών Πωλήσεων), με στόχο να απομονώσουμε την επίδραση των υπόλοιπων μεταβλητών και να αξιολογήσουμε τις καθαρές μεταξύ τους σχέσεις. Για λόγους συνέπειας και σύγκρισης με τα προηγούμενα αποτελέσματα, χρησιμοποιούμε τις ίδιες μεταβλητές που εξετάστηκαν στους συντελεστές συσχέτισης Pearson, Biweight και Percentage. Τα δεδομένα προέρχονται από τη βάση FRED (Federal Reserve Economic Data), καλύπτουν την περίοδο από 1/1/2005 έως 1/10/2022 σε τριμηνιαία βάση. Η ανάλυση πραγματοποιείται με τη χρήση της γλώσσας προγραμματισμού R, η οποία επιτρέπει την εύκολη εισαγωγή και διαχείριση μακροοικονομικών δεδομένων από τη βάση FRED μέσω της εντολής 'fImport'. Να σημειωθεί ότι η δομή του πίνακα δεδομένων str(df) και η προεπισκόπηση των πρώτων γραμμών head(df) παραμένουν ακριβώς ίδιες με εκείνες που παρουσιάστηκαν στον συντελεστή Pearson Correlation Coefficient.

Πακέτο psych: Στην εικόνα 14 βλέπουμε τον πίνακα μερικής συσχέτισης (Partial Correlation) που υπολογίστηκε με το πακέτο psych. Στο συγκεκριμένο πακέτο θα εξετάσουμε τις ίδιες μεταβλητές με αυτές που εξετάστηκαν ο συντελεστής συσχέτισης Pearson, καθώς ο μερικός

συντελεστής συσχέτισης θα μπορέσει να αφαιρέσει την επίδραση των υπόλοιπων μεταβλητών και να ανιχνεύσει την πραγματική σχέση των ζευγών.

```
> print(partial_cor_results)
```

	GDP_EU	UNEMPLOYMENT_EU	IMPORTS_EU	EXPORTS_EU
GDP_EU	1.000000000	-0.1642631	0.001165318	0.37699557
UNEMPLOYMENT_EU	-0.164263130	1.0000000	-0.328814080	0.40045785
IMPORTS_EU	0.001165318	-0.3288141	1.000000000	0.89065424
EXPORTS_EU	0.376995570	0.4004578	0.890654243	1.000000000
INVESTMENT_EU	-0.231324030	0.4378169	0.054035923	0.18392159
RETAIL_SALES_EU	0.242167872	-0.7974413	-0.085289353	0.05110311

	INVESTMENT_EU	RETAIL_SALES_EU
GDP_EU	-0.23132403	0.24216787
UNEMPLOYMENT_EU	0.43781695	-0.79744131
IMPORTS_EU	0.05403592	-0.08528935
EXPORTS_EU	0.18392159	0.05110311
INVESTMENT_EU	1.00000000	0.57489380
RETAIL_SALES_EU	0.57489380	1.00000000

EIKONA 14: Υπολογισμός Partial Correlation με το πακέτο Psych (William, 2025)

Συσχέτιση Απόστασης (Distance Correlation):

Στο παρόν στάδιο της ανάλυσης, θα εφαρμόσουμε τον συντελεστή συσχέτιση απόστασης (Distance Correlation Coefficient) χρησιμοποιώντας τις ίδιες μεταβλητές με εκείνες που χρησιμοποιήθηκαν στον υπολογισμό των συντελεστών συσχέτισης Pearson, Biweight, Percentage και Partial (Ακαθάριστο Εγχώριο Προϊόν, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Όγκος Λιανικών Πωλήσεων) για την Ευρωπαϊκή Ένωση. Τα δεδομένα προέρχονται από τη βάση FRED (Federal Reserve Bank), και καλύπτουν την περίοδο 1/1/2005 έως 1/10/2022 που μπορέσαμε να τα εισάγουμε με την εντολή 'fImport'. Η επιλογή των ίδιων μεταβλητών μας επιτρέπει να συγκρίνουμε τα αποτελέσματα του συντελεστή Distance Correlation με αυτά του Pearson Correlation, ώστε να διαπιστώσουμε κατά πόσο εντοπίζονται και μη γραμμικές συσχετίσεις που δεν καταγραφήκαν προηγουμένως. Η εντολή str(df) και η head(df) είναι ακριβώς ίδιες με εκείνες που παρουσιάστηκαν στην ανάλυση του Pearson Correlation Coefficient. Ο συντελεστής Distance Correlation θα υπολογιστεί με τα εξής πακέτα:

Πακέτο energy: Το πακέτο energy είναι ένα στατιστικό πακέτο στην R που περιλαμβάνει εργαλεία για μη παραμετρικές μεθόδους στατιστικής ανάλυσης. Ένα από τα βασικά του χαρακτηριστικά είναι ο Distance Correlation (dCor), καθώς αποτελεί ένα ισχυρό εργαλείο για σύνθετα δεδομένα που μπορεί να υπάρχουν κρυφές μη γραμμικές σχέσεις. Εάν η τιμή του Distance Correlation είναι παρόμοια με τον Pearson Correlation, τότε η σχέση είναι πιθανότατα

γραμμική. Ωστόσο εάν η τιμή του Distance Correlation είναι σημαντικά μεγαλύτερη από τον Pearson Correlation τότε υπάρχει πιθανότητα ύπαρξης μη γραμμικής σχέσης. Από την εικόνα 15 αποτυπώνονται τα αποτελέσματα για τον Distance Correlation και μπορούμε να κάνουμε μια συγκριτική ανάλυση των δύο μεθόδων (Distance και Pearson)

```
> print(distance_cor_df)
              GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU          1.0000000      0.4972952   0.9511045   0.9393840    0.8725139
UNEMPLOYMENT_EU 0.4972952      1.0000000   0.4830633   0.4541880    0.5251090
IMPORTS_EU       0.9511045      0.4830633   1.0000000   0.9929044    0.9155767
EXPORTS_EU       0.9393840      0.4541880   0.9929044   1.0000000    0.9175377
INVESTMENT_EU    0.8725139      0.5251090   0.9155767   0.9175377    1.0000000
RETAIL_SALES_EU 0.8805699      0.7331166   0.8267687   0.8019500    0.8362160
              RETAIL_SALES_EU
GDP_EU          0.8805699
UNEMPLOYMENT_EU 0.7331166
IMPORTS_EU       0.8267687
EXPORTS_EU       0.8019500
INVESTMENT_EU    0.8362160
RETAIL_SALES_EU 1.0000000
```

ΕΙΚΟΝΑ 15: Υπολογισμός Distance Correlation με το πακέτο energy (Rizzo & Székely, 2019).

ΠΙΝΑΚΑΣ 4: Σύγκριση αποτελεσμάτων Pearson Correlation VS Distance Correlation

Μεταβλητές	Pearson Correlation (psych package)	Distance Correlation (energy package)	Συμπεράσματα
GDP_EU- UNEMPLOYMENT_EU	-0.42	0.49	Η σχέση είναι κυρίως γραμμική.
GDP_EU-IMPORTS_EU	0.95	0.95	Η σχέση είναι κυρίως γραμμική.
GDP_EU – EXPORTS_EU	0.94	0.93	Η σχέση είναι γραμμική.
GDP_EU – INVESTMENT_EU	0.85	0.87	Η σχέση είναι κυρίως γραμμική.
GDP – RETAIL_SALES_EU	0.85	0.88	Η σχέση είναι κυρίως γραμμική.
UNEMPLOYMENT_EU – IMPORTS_EU	-0.29	0,48	Περιλαμβάνει μη γραμμική σχέση.
UNEMPLOYMENT_EU – EXPORTS_EU	-0.22	0.45	Περιλαμβάνει μη γραμμική σχέση που ο Pearson δεν ανιχνεύει.

UNEMPLOYMENT_EU – INVESTMENT_EU	-0.19	0.52	Περιλαμβάνει μη γραμμική σχέση που ο Pearson δεν ανιχνεύει.
UNEMPLOYMENT_EU – RETAIL_SALES_EU	-0.72	0.73	Περιλαμβάνει μη γραμμική σχέση.
IMPORTS_EU EXPORTS_EU	0.99	0.99	Η σχέση είναι γραμμική.
IMPORTS_EU INVESTMENT_EU	0.91	0.91	Η σχέση είναι γραμμική.
IMPORTS_EU RETAIL_SALES_EU	0.78	0.82	Η σχέση είναι κυρίως γραμμική.
EXPORTS_EU INVESTMENT_EU	0.92	0.91	Η σχέση είναι γραμμική.
EXPORTS_EU RETAIL_SALES_EU	0.76	0.80	Η σχέση είναι κυρίως γραμμική.
INVESTMENT_EU RETAIL_SALES_EU	0.76	0.83	Η σχέση είναι κυρίως γραμμική.

Μέγιστος Συντελεστής Πληροφορίας (Maximal Information Coefficient):

Ο Μέγιστος Συντελεστής Πληροφορίας (Maximal Information Coefficient) είναι ένα μέτρο συσχέτισης που μπορεί να ανιχνεύσει τόσο τις γραμμικές όσο και τις μη γραμμικές σχέσεις μεταξύ των μεταβλητών. Στο πλαίσιο της παρούσας ανάλυσης, χρησιμοποιήθηκε το dataset “Wholesale customers”, το ίδιο σύνολο δεδομένων που χρησιμοποιήσαμε και στους συντελεστές συσχέτισης Biweight Midcorrelation και Percentage Bend Correlation. Η επιλογή αυτών των μεταβλητών έγινε με στόχο να διερευνήσουμε την ικανότητα του MIC να ανιχνεύει σύνθετες συσχετίσεις που ενδεχομένως να μην εντοπίζονται από τα κλασικά γραμμικά μέτρα, όπως Pearson Correlation Coefficient. Στην εικόνα 16 παρουσιάζεται η δομή του dataframe `df_mic` μέσω της εντολής `str()`, όπου διαπιστώνουμε ότι το dataset περιλαμβάνει 440 παρατηρήσεις και 6 μεταβλητές αριθμητικού τύπου (integer). Αντίστοιχα, στην εικόνα 17 εμφανίζονται οι πρώτες γραμμές του πίνακα δεδομένων `head(df_mic)`, προσφέροντας μια πρώτη εποπτική ματιά στο περιεχόμενο του dataset. Με τον τρόπο αυτό διασφαλίζεται η καταλληλότητα των δεδομένων πριν την εφαρμογή του MIC. Στην εικόνα 16 μπορούμε να παρατηρήσουμε τις πρώτες γραμμές του data frame (`df_selected`) που μας βοηθάει να αποκτήσουμε μια πρώτη εικόνα των δεδομένων καθώς μας δείχνει τις πρώτες 6 γραμμές.

```
> str(df_mic)
'data.frame': 440 obs. of 6 variables:
 $ Fresh      : int 12669 7057 6353 13265 22615 9413 12126 7579 5963 6006 ...
 $ Milk       : int 9656 9810 8808 1196 5410 8259 3199 4956 3648 11093 ...
 $ Grocery    : int 7561 9568 7684 4221 7198 5126 6975 9426 6192 18881 ...
 $ Frozen     : int 214 1762 2405 6404 3915 666 480 1669 425 1159 ...
 $ Detergents_Paper: int 2674 3293 3516 507 1777 1795 3140 3321 1716 7425 ...
 $ Delicassen : int 1338 1776 7844 1788 5185 1451 545 2566 750 2098 ...
```

EIKONA 16: Επισκόπηση δομής data frame με την εντολή `str(df_selected)`.

```
> head(df_mic)
  Fresh Milk Grocery Frozen Detergents_Paper Delicassen
1 12669 9656   7561    214           2674         1338
2  7057 9810   9568   1762           3293         1776
3  6353 8808   7684   2405           3516         7844
4 13265 1196   4221   6404            507         1788
5 22615 5410   7198   3915           1777         5185
6  9413 8259   5126    666           1795         1451
```

EIKONA 17: Επισκόπηση δομής data frame με την εντολή `head(df_selected)`.

Πακέτο minerva: Στην εικόνα 18 παρουσιάζονται τα αποτελέσματα του Μέγιστου Συντελεστή Πληροφορίας (MIC) για τις έξι ποσοτικές μεταβλητές του dataset “Wholesale customers”. Παρατηρούμε ότι οι περισσότερες συσχετίσεις μεταξύ των μεταβλητών έχουν χαμηλές τιμές, γεγονός που υποδηλώνει ασθενή συσχέτιση. Παρόλα αυτά, η σχέση μεταξύ Grocery και Detergents_Paper (MIC = 0.623) εμφανίζει μέτρια ισχυρή συσχέτιση, γεγονός που ενδεχομένως οφείλεται στο ότι πρόκειται για κατηγορίες προϊόντων με παρόμοια καταναλωτική χρήση. Επίσης, το ζεύγος Milk – Grocery (MIC = 0.584) παρουσιάζει επίσης μια σχετικά αξιοσημείωτη σχέση, ενώ οι υπόλοιπες συσχετίσεις παραμένουν ασθενείς. Τα αποτελέσματα

```
> library(minerva)
> mic_results <- mine(df_mic)
> mic_results
$MIC
      Fresh      Milk  Grocery  Frozen Detergents_Paper Delicassen
Fresh      1.0000000 0.1697354 0.1741883 0.2662339      0.2328209 0.1755982
Milk        0.1697354 1.0000000 0.5840044 0.1714520      0.4806618 0.2828202
Grocery     0.1741883 0.5840044 1.0000000 0.2056580      0.6230436 0.2457021
Frozen      0.2662339 0.1714520 0.2056580 1.0000000      0.1799851 0.2018061
Detergents_Paper 0.2328209 0.4806618 0.6230436 0.1799851      1.0000000 0.1927976
Delicassen   0.1755982 0.2828202 0.2457021 0.2018061      0.1927976 1.0000000
```

EIKONA 18: Υπολογισμός Maximal Information Coefficient με το πακέτο Minerva (\$MIC) (Albanese et al., 2013).

Συντελεστής Συσχέτισης *Spearman's Rho*:

Για την εφαρμογή του συντελεστή συσχέτισης Spearman, επιλέξαμε διαφορετικά δεδομένα από εκείνα που χρησιμοποιήθηκαν στην περίπτωση του συντελεστή συσχέτισης Pearson με στόχο να επεκταθεί η ανάλυση σε μεταβλητές που δεν ακολουθούν απαραίτητα γραμμική σχέση. Συγκεκριμένα, χρησιμοποιήθηκε το dataset “learning”²⁶ από το πακέτο `peopleanalyticsdata` το οποίο περιλαμβάνει αξιολογήσεις συμμετεχόντων σε προγράμματα μάθησης. Οι μεταβλητές που εξετάστηκαν είναι οι εξής:

- `rel`: Σχετικότητα προγράμματος με την εργασία τους
- `fun`: Πόσο ευχάριστο βρήκαν το πρόγραμμα
- `clar`: Πόσο ξεκάθαρη ήταν η παρουσίαση
- `home`: Ποιότητα εργασιών
- `class`: Ποιότητα των συμμετεχόντων
- `fac`: Ποιότητα εισηγητών

Για να κατανοήσουμε καλύτερα τη δομή των δεδομένων που χρησιμοποιούμε στην ανάλυση στην εικόνα 19 και 20 εφαρμόσαμε τις συναρτήσεις `str()` και `head()`. Η συνάρτηση `str(learning)` εμφανίζει τη βασική δομή του dataset, περιγράφοντας τον αριθμό παρατηρήσεων και μεταβλητών, καθώς και τον τύπο δεδομένων κάθε μεταβλητής. Όπως φαίνεται από τα αποτελέσματα της εντολής, το dataset περιλαμβάνει 4974 παρατηρήσεις και 8 μεταβλητές, εκ των οποίων κάποιες είναι διατακτικές (`rel`, `fun`, `clar`, `home`, `class`, `fac`). Επιπλέον, η συνάρτηση `head(learning)` εμφανίζει τις πρώτες έξι γραμμές του dataset, παρέχοντας μια γρήγορη επισκόπηση των δεδομένων που περιέχει. Από τα πρώτα αποτελέσματα παρατηρούμε ότι η μεταβλητή `job` περιλαμβάνει επαγγέλματα όπως `management`, `technician`, `entrepreneur`, ενώ η μεταβλητή `education` περιλαμβάνει τιμές όπως `tertiary`, `secondary` και `unknown`.

```
To see confidence intervals of the correlations, print with the short=FALSE option
> str(learning)
'data.frame':   4974 obs. of  8 variables:
 $ idcode: int   176304 176304 176304 176304 176304 176304 176304 176304 176304 176304
 $ rec   : int    0 0 0 1 0 0 0 0 1 1 ...
 $ rel   : int    3 4 3 4 3 3 4 3 5 4 ...
 $ fun   : int    5 4 4 5 5 5 5 5 4 5 ...
 $ clar  : int    3 5 4 5 5 3 4 3 5 5 ...
 $ home  : int    4 4 3 4 4 4 5 5 4 4 ...
 $ class : int    4 4 4 5 5 2 3 4 4 4 ...
 $ fac   : int    4 1 1 5 4 2 1 3 4 4 ...
```

ΕΙΚΟΝΑ 19: Επισκόπηση δομής data frame με την εντολή `str(df_selected)`.

²⁶ Πηγή δεδομένων: Πακέτο R “peopleanalyticsdata” – Dataset “learning”
Διαθέσιμο στο CRAN: <https://cran.r-project.org/package=peopleanalyticsdata>

```
> head(learning)
  idcode rec rel fun clar home class fac
1 176304  0  3  5   3   4   4   4
2 176304  0  4  4   5   4   4   1
3 176304  0  3  4   4   3   4   1
4 176304  1  4  5   5   4   5   5
5 176304  0  3  5   5   4   5   4
6 176304  0  3  5   3   4   2   2
```

EIKONA 20: Επισκόπηση δομής data frame με την εντολή head(df_selected).

Η χρήση του συγκεκριμένου dataset μας επιτρέπει να εξετάσουμε τη συσχέτιση των μεταβλητών και να μας δώσει επιπλέον πληροφορίες σχετικά με τη δομή των δεδομένων και την πιθανή ύπαρξη μη γραμμικών σχέσεων. Η εφαρμογή του συντελεστή συσχέτισης Spearman θα γίνει με διάφορα πακέτα για να συγκρίνουμε τα αποτελέσματα και να αξιολογήσουμε τη σταθερότητα των μετρήσεων.

Συνάρτηση cor():

```
> spearman_cor <- cor(df_numeric, method = "spearman", use = "complete.obs")
>
> # Εκτύπωση
> print(spearman_cor)
      rel      fun      clar      home      class      fac
rel  1.0000000 0.3423212 0.4995744 0.3100115 0.4240065 0.2582038
fun  0.3423212 1.0000000 0.4397107 0.3949383 0.3326681 0.2728178
clar 0.4995744 0.4397107 1.0000000 0.4345585 0.5390834 0.3537829
home 0.3100115 0.3949383 0.4345585 1.0000000 0.3500643 0.2331640
class 0.4240065 0.3326681 0.5390834 0.3500643 1.0000000 0.4410828
fac  0.2582038 0.2728178 0.3537829 0.2331640 0.4410828 1.0000000
```

EIKONA 21: Υπολογισμός Spearman Correlation Coefficient με το πακέτο cor.

Πακέτο Hmisc: Το συγκεκριμένο πακέτο τα αποτελέσματα απεικονίζονται στην εικόνα 22, αναφέραμε και στην διερεύνηση των προηγούμενων μεταβλητών παρέχει παρόμοια αποτελέσματα με την συνάρτηση cor(), καθώς το μαθηματικό μοντέλο που βρίσκετε πίσω από το πακέτο δεν αλλάζει. Επίσης δίνει μια επιπλέον πληροφορία είναι η εξαγωγή των p-values.

```
> print(spearman_hmisc$r)
      rel      fun      clar      home      class      fac
rel  1.0000000 0.3272258 0.5025998 0.3186134 0.4189728 0.2459454
fun  0.3272258 1.0000000 0.4243673 0.3913999 0.3345627 0.2622015
clar 0.5025998 0.4243673 1.0000000 0.4347839 0.5398592 0.3441174
home 0.3186134 0.3913999 0.4347839 1.0000000 0.3539196 0.2288592
class 0.4189728 0.3345627 0.5398592 0.3539196 1.0000000 0.4358839
fac  0.2459454 0.2622015 0.3441174 0.2288592 0.4358839 1.0000000

> print(spearman_hmisc$p)
      rel fun clar home class fac
rel    NA  0    0    0    0    0
fun    0  NA    0    0    0    0
clar   0  0   NA    0    0    0
home   0  0    0   NA    0    0
class  0  0    0    0   NA    0
fac    0  0    0    0    0   NA
```

EIKONA 22: Υπολογισμός Spearman Correlation Coefficient με το πακέτο Hmisc (Frank E, 2025).

Πακέτο wdm: Το πακέτο wdm (Weighted Dependency Measures) στην R είναι ιδιαίτερα χρήσιμο γιατί επιτρέπει τον υπολογισμό σε μεγάλους όγκους δεδομένων καθώς υποστηρίζει την δυνατότητα missing values (NA). Στην εικόνα 23 βλέπουμε ότι τα αποτελέσματα είναι ίδια με την συνάρτηση cor της R και παρόμοια με το πακέτο Hmisc.

```
> print(spearman_wdm)
      rel      fun      clar      home      class      fac
rel  1.0000000 0.3272258 0.5025998 0.3186134 0.4189728 0.2459454
fun  0.3272258 1.0000000 0.4243673 0.3913999 0.3345627 0.2622015
clar 0.5025998 0.4243673 1.0000000 0.4347839 0.5398592 0.3441174
home 0.3186134 0.3913999 0.4347839 1.0000000 0.3539196 0.2288592
class 0.4189728 0.3345627 0.5398592 0.3539196 1.0000000 0.4358839
fac  0.2459454 0.2622015 0.3441174 0.2288592 0.4358839 1.0000000
```

EIKONA 23: Υπολογισμός Spearman Correlation Coefficient με το πακέτο wdm (Nagler, 2025).

Πακέτο psych: Το πακέτο psych δίνει ακριβώς τα ίδια αποτελέσματα με το πακέτο Hmisc καθώς επιστρέφει και p-values αλλά δεν μπορεί να χειριστεί των υπολογισμό δεδομένων με Missing values.

```
> print(spearman_psych$r)
      rel      fun      clar      home      class      fac
rel  1.0000000 0.3272258 0.5025998 0.3186134 0.4189728 0.2459454
fun  0.3272258 1.0000000 0.4243673 0.3913999 0.3345627 0.2622015
clar 0.5025998 0.4243673 1.0000000 0.4347839 0.5398592 0.3441174
home 0.3186134 0.3913999 0.4347839 1.0000000 0.3539196 0.2288592
class 0.4189728 0.3345627 0.5398592 0.3539196 1.0000000 0.4358839
fac  0.2459454 0.2622015 0.3441174 0.2288592 0.4358839 1.0000000
>
> # 8. Εμφάνιση των p-values
> cat("\nP-values:\n")

P-values:
> print(spearman_psych$p)
      rel      fun      clar      home      class
rel  0.000000e+00 5.263942e-119 1.782699e-301 4.151343e-107 9.336508e-183
fun  8.773236e-120 0.000000e+00 2.038266e-204 1.024130e-164 2.190596e-112
clar 1.273356e-302 1.698555e-205 0.000000e+00 3.047937e-206 0.000000e+00
home 1.037836e-107 1.137922e-165 2.344567e-207 0.000000e+00 2.459505e-123
class 9.336508e-184 4.381193e-113 0.000000e+00 3.513579e-124 0.000000e+00
fac  1.412403e-66 7.764156e-75 3.824789e-130 2.815557e-54 3.669107e-198
      fac
rel  2.824805e-66
fun  2.329247e-74
clar 3.059832e-129
home 2.815557e-54
class 4.036017e-197
fac  0.000000e+00
```

EIKONA 24: Υπολογισμός Spearman Correlation Coefficient με το πακέτο psych (William Revelle, 2025).

Συντελεστής Συσχέτισης Shepherd's Pi:

Ο συντελεστής συσχέτισης Shepherd's Pi υπολογίστηκε με τις ίδιες μεταβλητές που εξετάστηκαν και στον συντελεστή συσχέτισης Spearman's Rho (rel, fun, clar, home, class, fac).

Το συγκεκριμένο μέτρο βασίζεται στην ίδια λογική κατάταξη όπως και ο Spearman, ωστόσο εφαρμόζει ένα πρόσθετο στάδιο ελέγχου για ακραίες τιμές. Συγκεκριμένα ο συντελεστής Shepherd's Pi αφαιρεί τις επιρρεπείς εξωτερικές τιμές και διπλασιάζει τη τιμή p- τιμή προκειμένου να ενσωματώσει τη διαδικασία καθαρισμού των δεδομένων. Αυτό καθιστά τον δείκτη ιδιαίτερα ανθεκτικό σε ακραίες παρατηρήσεις, προσφέροντας πιο σταθερές και αξιόπιστες εκτιμήσεις συσχέτισης.

```
> result_shepherd <- correlation(df_selected, method = "shepherd", p_adjust = "none")
> result_shepherd
# Correlation Matrix (shepherd-method)
```

Parameter1	Parameter2	rho	95% CI	S	p
rel	fun	0.36	[0.34, 0.39]	6.50e+09	< .001***
rel	clar	0.56	[0.54, 0.59]	4.36e+09	< .001***
rel	home	0.39	[0.36, 0.41]	6.24e+09	< .001***
rel	class	0.45	[0.42, 0.48]	5.30e+09	< .001***
rel	fac	0.29	[0.26, 0.32]	6.43e+09	< .001***
fun	clar	0.47	[0.45, 0.50]	5.35e+09	< .001***
fun	home	0.46	[0.43, 0.48]	5.51e+09	< .001***
fun	class	0.43	[0.40, 0.45]	5.59e+09	< .001***
fun	fac	0.33	[0.30, 0.35]	6.95e+09	< .001***
clar	home	0.50	[0.48, 0.53]	5.17e+09	< .001***
clar	class	0.55	[0.53, 0.58]	4.42e+09	< .001***
clar	fac	0.45	[0.42, 0.47]	5.30e+09	< .001***
home	class	0.44	[0.41, 0.47]	5.59e+09	< .001***
home	fac	0.23	[0.20, 0.26]	7.48e+09	< .001***
class	fac	0.46	[0.43, 0.48]	5.09e+09	< .001***

p-value adjustment method: none
Observations: 4120

EIKONA 25: Υπολογισμός του συντελεστή συσχέτισης Συντελεστής Συσχέτισης Shepherd's Pi με το πακέτο correlation (Makowski et al., 2020).

Συντελεστής Συσχέτισης Kendall's Tau:

Για την εφαρμογή του συντελεστή συσχέτισης Kendall's Tau, επιλέξαμε το ίδιο σύνολο μεταβλητών με τους συντελεστές συσχέτισης Spearman's Rho και Shepherd's Pi προκειμένου να υπάρξει μια σύγκριση των αποτελεσμάτων έτσι ώστε να μπορέσουμε να διακρίνουμε τις διαφορές των συντελεστών συσχέτισης. Τα δεδομένα προέρχονται από το dataset "learning" από το πακέτο "reopleanalyticsdata" το οποίο περιλαμβάνει αξιολογήσεις συμμετεχόντων σε προγράμματα μάθησης. Οι μεταβλητές που επιλέξαμε καταγράφουν τη σχετική γνώμη των συμμετεχόντων ως προς την εμπειρία τους στα προγράμματα.

Συνάρτηση cor(): Στην εικόνα 26 που ο υπολογισμός έγινε με την ενσωματωμένη συνάρτηση cor της R με την μέθοδο Kendall's Tau βλέπουμε ότι υπάρχουν αλλαγές σε σχέση με τις τιμές που έδωσε ο συντελεστής spearman. Πιο συγκεκριμένα ο συντελεστής Spearman δίνει υψηλότερες τιμές συσχέτισης από τον Kendall. Ο συντελεστής συσχέτισης Kendall's Tau είναι

πιο συντηρητικός στη μέτρηση. Από την θεωρία γνωρίζουμε ότι μεταξύ των δύο μεταβλητών ο συντελεστής Spearman είναι κατάλληλος για μεγάλες βάσεις δεδομένων, καθώς είναι πιο εύκολος στον υπολογισμό ενώ ο Kendall's Tau είναι κατάλληλος για μικρότερα σύνολα δεδομένων.

```
> df_selected <- learning[, c("rel", "fun", "clar", "home", "class", "fac")]
> kendall_matrix <- cor(df_selected, method = "kendall", use = "pairwise.complete.obs")
> kendall_matrix
```

	rel	fun	clar	home	class	fac
rel	1.0000000	0.2965826	0.4594348	0.2884112	0.3771814	0.2183736
fun	0.2965826	1.0000000	0.3877512	0.3577382	0.3005376	0.2354541
clar	0.4594348	0.3877512	1.0000000	0.3991680	0.4915325	0.3075734
home	0.2884112	0.3577382	0.3991680	1.0000000	0.3201587	0.2038082
class	0.3771814	0.3005376	0.4915325	0.3201587	1.0000000	0.3891370
fac	0.2183736	0.2354541	0.3075734	0.2038082	0.3891370	1.0000000

EIKONA 26: Υπολογισμός Kendall's Tau Correlation με την συνάρτηση cor.

Πακέτο wdm:

```
> library(wdm)
> df_selected <- learning[, c("rel", "fun", "clar", "home", "class", "fac")]
> kendall_wdm <- wdm::wdm(df_selected, method = "kendall")
> kendall_wdm
```

	rel	fun	clar	home	class	fac
rel	1.0000000	0.2965826	0.4594348	0.2884112	0.3771814	0.2183736
fun	0.2965826	1.0000000	0.3877512	0.3577382	0.3005376	0.2354541
clar	0.4594348	0.3877512	1.0000000	0.3991680	0.4915325	0.3075734
home	0.2884112	0.3577382	0.3991680	1.0000000	0.3201587	0.2038082
class	0.3771814	0.3005376	0.4915325	0.3201587	1.0000000	0.3891370
fac	0.2183736	0.2354541	0.3075734	0.2038082	0.3891370	1.0000000

EIKONA 27: Υπολογισμός Kendall's Tau Correlation με πακέτο wdm (Thomas, 2025).

Πακέτο Psych:

```
> library(psych)
> df_selected <- learning[, c("rel", "fun", "clar", "home", "class", "fac")]
>
> kendall_psych <- psych::corr.test(df_selected, method = "kendall")
> kendall_psych
Call:psych::corr.test(x = df_selected, method = "kendall")
Correlation matrix
```

	rel	fun	clar	home	class	fac
rel	1.00	0.30	0.46	0.29	0.38	0.22
fun	0.30	1.00	0.39	0.36	0.30	0.24
clar	0.46	0.39	1.00	0.40	0.49	0.31
home	0.29	0.36	0.40	1.00	0.32	0.20
class	0.38	0.30	0.49	0.32	1.00	0.39
fac	0.22	0.24	0.31	0.20	0.39	1.00

```
Sample Size
```

	rel	fun	clar	home	class	fac
rel	4850	4784	4751	4543	4332	4764
fun	4784	4788	4715	4528	4305	4706
clar	4751	4715	4756	4509	4301	4677
home	4543	4528	4509	4547	4200	4476
class	4332	4305	4301	4200	4336	4283
fac	4764	4706	4677	4476	4283	4774

```
Probability values (Entries above the diagonal are adjusted for multiple tests.)
```

	rel	fun	clar	home	class	fac
rel	0	0	0	0	0	0
fun	0	0	0	0	0	0
clar	0	0	0	0	0	0
home	0	0	0	0	0	0
class	0	0	0	0	0	0
fac	0	0	0	0	0	0

EIKONA 28: Υπολογισμός Kendall's Tau Correlation με πακέτο psych (William, 2025).

Συντελεστής Συσχέτισης *Blomqvist's β*:

Για τον συντελεστή συσχέτισης *Blomqvist's β* χρησιμοποιήθηκαν οι ίδιες μεταβλητές που εξετάστηκαν και με τους συντελεστές συσχέτισης *Spearman's Rho*, *Shepherd's Pi* και *Kendall's Tau*, δηλαδή τις μεταβλητές (*rel*, *fun*, *clar*, *home*, *class* και *fac*) που με λίγα λόγια αποτυπώνουν τις αντιλήψεις των φοιτητών για τη μαθησιακή διαδικασία. Ο συγκεκριμένος συντελεστής υπολογίζει τη γραμμική συσχέτιση με βάση τη διάμεσο των μεταβλητών, προσφέροντας ανθεκτικότητα σε ακραίες τιμές και σταθερότητα σε ασύμμετρες κατανομές. Η ανάλυση πραγματοποιήθηκε με τη γλώσσα R με το πακέτο *correlation*, με επιλογή *method = "blomqvist"*. Η χρήση του συντελεστή συσχέτισης *Blomqvist's β* και του *Kendall's Tau* αναδεικνύει τη δυνατότητα πολλαπλής προσέγγισης των ordinal δεδομένων. Στην εικόνα 29 μπορούμε να δούμε τα αποτελέσματα του συγκεκριμένου συντελεστή συσχέτισης.

```
> result_blomqvist <- correlation(df_selected, method = "blomqvist", p_adjust = "none")
> result_blomqvist
# Correlation Matrix (blomqvist-method)
```

Parameter1	Parameter2	r	95% CI	t	df	p
rel	fun	0.22	[0.19, 0.24]	15.28	4782	< .001***
rel	clar	0.42	[0.40, 0.45]	32.35	4749	< .001***
rel	home	0.26	[0.23, 0.28]	17.85	4541	< .001***
rel	class	0.33	[0.31, 0.36]	23.16	4330	< .001***
rel	fac	0.22	[0.19, 0.25]	15.47	4762	< .001***
fun	clar	0.34	[0.32, 0.37]	25.17	4713	< .001***
fun	home	0.37	[0.34, 0.39]	26.44	4526	< .001***
fun	class	0.08	[0.06, 0.11]	5.58	4303	< .001***
fun	fac	-0.03	[-0.05, 0.00]	-1.81	4704	0.071
clar	home	0.40	[0.37, 0.42]	29.17	4507	< .001***
clar	class	0.39	[0.37, 0.42]	27.84	4299	< .001***
clar	fac	0.22	[0.20, 0.25]	15.70	4675	< .001***
home	class	0.24	[0.21, 0.27]	15.98	4198	< .001***
home	fac	0.11	[0.08, 0.14]	7.43	4474	< .001***
class	fac	0.46	[0.44, 0.49]	34.22	4281	< .001***

```
p-value adjustment method: none
Observations: 4200-4784
```

EΙΚΟΝΑ 29: Υπολογισμός του συντελεστή συσχέτισης *Blomqvist's β* με το πακέτο *correlation* (Makowski et al., 2020).

Συντελεστής Συσχέτισης *Gamma*:

Για την περαιτέρω διερεύνηση των μεταβλητών σε κλίμακα *Likert* εφαρμόστηκε και ο συντελεστής συσχέτισης *Gamma*, χρησιμοποιώντας τις ίδιες μεταβλητές που εξετάστηκαν στις αναλύσεις με τους συντελεστές συσχέτισης *Spearman*, *Shepherd's Pi*, *Kendall's Tau* και *Blomqvist's Correlation*. Ο συντελεστής *Gamma* είναι κατάλληλος για ordinal δεδομένα και παρουσιάζει ιδιαίτερη ανθεκτικότητα σε περιπτώσεις με πολλές ισοβαθμίες (*ties*). Η ανάλυση πραγματοποιήθηκε στην γλώσσα R, μέσω του πακέτου *Correlation()* με επιλογή *method = "gamma"*. Στην εικόνα 30 μπορούμε να δούμε τα αποτελέσματα του συντελεστή συσχέτισης *Gamma*.

```
> result_gamma <- correlation(df_selected, method = "gamma", p_adjust = "none")
> result_gamma
# Correlation Matrix (gamma-method)
```

Parameter1	Parameter2	r	95% CI	t	df	p
rel	fun	0.45	[0.42, 0.47]	34.53	4782	< .001***
rel	clar	0.66	[0.64, 0.67]	59.78	4749	< .001***
rel	home	0.43	[0.40, 0.45]	31.95	4541	< .001***
rel	class	0.55	[0.52, 0.57]	42.90	4330	< .001***
rel	fac	0.33	[0.30, 0.35]	23.79	4762	< .001***
fun	clar	0.57	[0.55, 0.59]	47.15	4713	< .001***
fun	home	0.53	[0.51, 0.55]	41.68	4526	< .001***
fun	class	0.45	[0.42, 0.47]	32.65	4303	< .001***
fun	fac	0.35	[0.33, 0.38]	25.88	4704	< .001***
clar	home	0.57	[0.55, 0.59]	46.44	4507	< .001***
clar	class	0.69	[0.67, 0.71]	62.63	4299	< .001***
clar	fac	0.45	[0.43, 0.47]	34.60	4675	< .001***
home	class	0.46	[0.44, 0.49]	33.89	4198	< .001***
home	fac	0.30	[0.28, 0.33]	21.40	4474	< .001***
class	fac	0.56	[0.54, 0.58]	43.78	4281	< .001***

```
p-value adjustment method: none
Observations: 4200-4784
```

EIKONA 30: Υπολογισμός του συντελεστή συσχέτισης Gamma με το πακέτο correlation(Makowski et al., 2020).

Συντελεστής Συσχέτισης Gaussian Rank:

Για τον συντελεστή συσχέτισης Gaussian Rank, χρησιμοποιήθηκαν οι ίδιες μεταβλητές που εξετάστηκαν και με τους συντελεστές συσχέτισης Kendall's Tau, Blomqvist's και Gamma, δηλαδή οι μεταβλητές (rel, fun, clar, home class και fac), οι οποίες καταγράφουν τις απόψεις των συμμετεχόντων σχετικά με τη μαθησιακή τους εμπειρία. Ο συγκεκριμένος συντελεστής βασίζεται στις κανονικές ποσοστιαίες θέσεις των ταξινομημένων τιμών (ranks), προσφέροντας μία πιο σταθερή εκτίμηση σε σχέση με κλασικές μεθόδους, ειδικά σε περιπτώσεις με μη κανονικές κατανομές. Η ανάλυση πραγματοποιήθηκε με τη γλώσσα R μέσω του πακέτου correlation, με επιλογή method= "gaussian". Η χρήση του συντελεστή Gaussian Rank σε συνδυασμό με τον συντελεστή Kendall's Tau και τον Blomqvist's αναδεικνύει τη σημασία της εφαρμογής εναλλακτικών μεθόδων συσχέτισης για ordinal δεδομένα. Στην εικόνα 31 παρουσιάζονται τα αποτελέσματα του συγκεκριμένου συντελεστή.

```
> correlation(df_selected, method = "gaussian", p_adjust = "none")
# Correlation Matrix (gaussian-method)
```

Parameter1	Parameter2	r	95% CI	t	df	p
rel	fun	0.35	[0.32, 0.37]	25.81	4782	< .001***
rel	clar	0.52	[0.50, 0.54]	41.55	4749	< .001***
rel	home	0.33	[0.31, 0.36]	23.83	4541	< .001***
rel	class	0.44	[0.42, 0.47]	32.38	4330	< .001***
rel	fac	0.26	[0.23, 0.28]	18.21	4762	< .001***
fun	clar	0.44	[0.41, 0.46]	33.45	4713	< .001***
fun	home	0.40	[0.37, 0.42]	29.15	4526	< .001***
fun	class	0.36	[0.33, 0.39]	25.30	4303	< .001***
fun	fac	0.27	[0.24, 0.30]	19.15	4704	< .001***
clar	home	0.44	[0.42, 0.47]	33.29	4507	< .001***
clar	class	0.56	[0.54, 0.58]	44.59	4299	< .001***
clar	fac	0.36	[0.33, 0.38]	25.99	4675	< .001***
home	class	0.38	[0.35, 0.40]	26.40	4198	< .001***
home	fac	0.23	[0.20, 0.26]	15.92	4474	< .001***
class	fac	0.44	[0.41, 0.46]	31.91	4281	< .001***

```
p-value adjustment method: none
Observations: 4200-4784
```

EIKONA 31: Υπολογισμός του συντελεστή συσχέτισης Gaussian Rank με το πακέτο correlation (Makowski et al., 2020).

Συντελεστής Συσχέτισης Hoeffding's D:

Ο συντελεστής συσχέτισης Hoeffding's D γνωρίζουμε από την θεωρία ότι είναι ένα μέτρο συσχέτισης που αξιολογεί την γενική εξάρτηση μεταξύ δύο μεταβλητών ανεξάρτητα αν η σχέση τους είναι γραμμική ή μη γραμμική. Για τη συγκεκριμένη ανάλυση, επιλέξαμε να εξετάσουμε τις ίδιες μεταβλητές που χρησιμοποιήθηκαν στον Pearson Correlation καθώς και σε άλλους συντελεστές συσχέτισης (Ακαθάριστου Εγχώριου Προϊόν, Ανεργία, Εισαγωγές, Εξαγωγές, Επενδύσεις και Όγκος λιανικών πωλήσεων) της Ευρωπαϊκής Ένωσης. Ο λόγος αυτής της επιλογής είναι να ανιχνεύσουμε αν υπάρχουν σχέσεις μεταξύ των μεταβλητών που δεν είναι αποκλειστικά γραμμικές, αλλά περιλαμβάνουν πιο σύνθετες μη γραμμικές εξαρτήσεις, τις οποίες ο συντελεστής Pearson δεν μπορεί να εντοπίσει. Οι μεταβλητές συλλέχθηκαν από την βάση δεδομένων FRED (Federal Reserve Bank) όπου η γλώσσα προγραμματισμού R μπορεί εύκολα να τραβήξει τα δεδομένα με μια απλή εντολή. Το χρονικό διάστημα που καλύπτουν τα δεδομένα είναι από 1/1/2005 έως 1/10/2022. Για τον υπολογισμό και την γραφική παράσταση του συντελεστή Hoeffding's D, θα χρησιμοποιηθούν πολλαπλά πακέτα της R, όπως το Hmisc, το wdm και το corrplot. Το Hmisc θα χρησιμοποιηθεί για τον βασικό υπολογισμό του συντελεστή, το wdm για εναλλακτικές μεθόδους υπολογισμού, ενώ το corrplot θα χρησιμοποιηθεί για τη γραφική απεικόνιση των αποτελεσμάτων.

Πακέτο (Hmisc):

```
> as.matrix(hoeff_result$D)
      GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU INVESTMENT_EU
GDP_EU      1.00000000      0.09610248      0.5352971      0.5411528      0.4353577
UNEMPLOYMENT_EU 0.09610248      1.00000000      0.1054800      0.1108233      0.1006545
IMPORTS_EU      0.53529711      0.10548001      1.0000000      0.9322013      0.6014366
EXPORTS_EU      0.54115282      0.11082334      0.9322013      1.0000000      0.6051537
INVESTMENT_EU   0.43535773      0.10065451      0.6014366      0.6051537      1.0000000
RETAIL_SALES_EU 0.30804423      0.26516514      0.2209413      0.2355879      0.2706968
      RETAIL_SALES_EU
GDP_EU      0.3080442
UNEMPLOYMENT_EU 0.2651651
IMPORTS_EU      0.2209413
EXPORTS_EU      0.2355879
INVESTMENT_EU   0.2706968
RETAIL_SALES_EU 1.0000000
> as.matrix(hoeff_result$P)
      GDP_EU UNEMPLOYMENT_EU IMPORTS_EU EXPORTS_EU
GDP_EU      NA      1.091557e-06 1.000000e-08 1.000000e-08
UNEMPLOYMENT_EU 1.091557e-06      NA 3.21078e-07 1.598814e-07
IMPORTS_EU      1.000000e-08 3.210780e-07      NA 1.000000e-08
EXPORTS_EU      1.000000e-08 1.598814e-07 1.000000e-08      NA
INVESTMENT_EU 1.000000e-08 6.026674e-07 1.000000e-08 1.000000e-08
RETAIL_SALES_EU 1.000000e-08 1.000000e-08 1.000000e-08 1.000000e-08
      INVESTMENT_EU RETAIL_SALES_EU
GDP_EU      1.000000e-08      1e-08
UNEMPLOYMENT_EU 6.026674e-07      1e-08
IMPORTS_EU      1.000000e-08      1e-08
EXPORTS_EU      1.000000e-08      1e-08
INVESTMENT_EU      NA      1e-08
RETAIL_SALES_EU 1.000000e-08      NA
```

EIKONA 32: Υπολογισμός Hoeffding's D Correlation με πακέτο Hmisc (Frank, 2025).

Πακέτο (wdm):

```
> new_labels <- c("GDP_EU", "UNEMPLOYMENT_EU", "IMPORTS_EU", "EXPORTS_EU", "INVE  
> rownames(hoeff_result) <- new_labels  
> colnames(hoeff_result) <- new_labels  
> round(hoeff_result, 3)
```

	GDP_EU	UNEMPLOYMENT_EU	IMPORTS_EU	EXPORTS_EU	INVESTMENT_EU
GDP_EU	1.000	0.096	0.535	0.541	0.435
UNEMPLOYMENT_EU	0.096	1.000	0.105	0.111	0.101
IMPORTS_EU	0.535	0.105	1.000	0.932	0.601
EXPORTS_EU	0.541	0.111	0.932	1.000	0.605
INVESTMENT_EU	0.435	0.101	0.601	0.605	1.000
RETAIL_SALES_EU	0.309	0.264	0.221	0.236	0.271

	RETAIL_SALES_EU
GDP_EU	0.309
UNEMPLOYMENT_EU	0.264
IMPORTS_EU	0.221
EXPORTS_EU	0.236
INVESTMENT_EU	0.271
RETAIL_SALES_EU	1.000

EIKONA 33: Υπολογισμός Hoeffding's D με πακέτο wdm (Thomas Nagler, 2025).

Συντελεστής Συσχέτισης Cramér's V:

Όπως αναφέραμε και στην θεωρία ο συντελεστής συσχέτισης Cramér's V είναι ένα μέτρο συσχέτισης που χρησιμοποιείται για την εκτίμηση της δύναμης της σχέσης μεταξύ δύο κατηγορικών μεταβλητών. Ο συντελεστής συσχέτισης Cramér's V είναι ειδικά σχεδιασμένος για να αναλύει ποιοτικές μεταβλητές. Για τον υπολογισμό του Cramér's V, χρησιμοποιήσαμε το σύνολο δεδομένων "Bank Marketing"²⁷ από το UC Irvine Machine Learning Repository. Επιλέξαμε συγκεκριμένες ποιοτικές μεταβλητές, οι οποίες περιλαμβάνουν:

- Job (εργασία)
- Marital (οικογενειακή κατάσταση)
- Education (εκπαίδευση)
- Default (μη πληρωμή της πιστωτικής κάρτας ασυνέπεια)
- Housing (στεγαστικό δάνειο)
- Loan (καταναλωτικό δάνειο)
- Month (μήνας επικοινωνίας με τον πελάτη)

Για να κατανοήσουμε καλύτερα τα δεδομένα στην εικόνα 34 και 35 αντίστοιχα εφαρμόσαμε τις συναρτήσεις str() και head(). Η συνάρτηση str(df_selected) εμφανίζει την δομή του dataset

²⁷ Πηγή δεδομένων: UC Irvine Machine Learning Repository – Dataset "Bank Marketing"
Διαθέσιμο στο: <https://archive.ics.uci.edu/ml/datasets/bank+marketing>

που επιλέξαμε, τον αριθμό παρατηρήσεων και μεταβλητών, καθώς και τον τύπο κάθε μεταβλητής. Τα αποτελέσματα του dataset περιλαμβάνουν 45.211 παρατηρήσεις και 7 μεταβλητές όπου όλες είναι κατηγορικές μεταβλητές για το συγκεκριμένο μέτρο συσχέτισης οι μεταβλητές δεν θα μπορούσαν να είναι διαφορετικό είδος καθώς ο συντελεστής Cramér's V χρησιμοποιείται σε κατηγορικές μεταβλητές. Επιπλέον η συνάρτηση `head(df_selected)` εμφανίζει τις πρώτες έξι γραμμές του dataset παρέχοντας μια γρήγορη επισκόπηση των δεδομένων.

```
> df_selected <- data_bank[, c("job", "marital", "education", "default", "housing", "$
> str(df_selected)
'data.frame': 45211 obs. of 7 variables:
 $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
 $ marital  : chr  "married" "single" "married" "married" ...
 $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
 $ default  : chr  "no" "no" "no" "no" ...
 $ housing  : chr  "yes" "yes" "yes" "yes" ...
 $ loan     : chr  "no" "no" "yes" "no" ...
 $ month    : chr  "may" "may" "may" "may" ...
```

EIKONA 34: Επισκόπηση δομής data frame με την εντολή `str(df_selected)`.

```
> head(df_selected)
      job marital education default housing loan month
1  management married  tertiary      no      yes  no   may
2  technician single   secondary     no      yes  no   may
3 entrepreneur married  secondary     no      yes yes   may
4 blue-collar married   unknown      no      yes  no   may
5    unknown single    unknown      no      no  no   may
6  management married  tertiary      no      yes  no   may
```

EIKONA 35: Επισκόπηση δομής data frame με την εντολή `head(df_selected)`.

Πακέτο rcompanion: Στην εικόνα 36 βλέπουμε παρατηρούμε τα αποτελέσματα των συσχετίσεων με τον συντελεστή Cramér's V οι πιο σημαντικές συσχετίσεις είναι οι εξής:

Job-Loan (0.106): Υπάρχει ασθενής σχέση δηλώνοντας ότι το επάγγελμα δεν αποτελεί καθοριστικό παράγοντα για την λήψη δανείου.

Job-Month (0.110): Υπάρχει ασθενής σχέση μεταξύ της εργασίας και του μήνα επικοινωνίας.

Education-Housing (0.119): Υπάρχει ασθενής συσχέτιση που θα μπορούσε να συνδέεται με υψηλότερη εκπαίδευση θα υπάρχει και καλύτερη οικονομική σταθερότητα.

Marital-Education (0.121): Υπάρχει ασθενής σχέση μεταξύ της οικογενειακής κατάστασης και της εκπαίδευσης.

Loan-Month (0.182): Υπάρχει ασθενής σχέση μεταξύ καταναλωτικό δάνειο και του μήνα επικοινωνίας.

Job-Marital (0.206): Υπάρχει ασθενής σχέση μεταξύ της εργασίας και της οικογενειακής κατάστασης.

Job-Housing (0.282): Υπάρχει ασθενής σχέση ανάμεσα στην δουλειά και το στεγαστικό δάνειο καθώς, ορισμένα επαγγέλματα προσφέρουν οικονομική σταθερότητα και διευκολύνουν την απόκτηση σπιτιού.

Job-Education (0.458): Υπάρχει μέτρια σχέση, καθώς η εκπαίδευση συνδέεται άμεσα με το είδος της εργασίας που μπορεί να έχει ένα άτομο.

Housing-Month (0.504): Υπάρχει ισχυρή σχέση που δείχνει ότι η περίοδος του έτους επηρεάζει τη λήψη στεγαστικού δανείου, πιθανώς λόγω εποχικών οικονομικών συνθηκών.

```
> cramers_v_df <- as.data.frame(cramers_v_matrix)
>
> cramers_v_df
```

	job	marital	education	default	housing	loan	month
job	1.00000	0.20600	0.45830	0.036530	0.281700	0.10650	0.11020
marital	0.20600	1.00000	0.12160	0.019230	0.020690	0.05194	0.07232
education	0.45830	0.12160	1.00000	0.015900	0.119300	0.08028	0.11010
default	0.03653	0.01923	0.01590	1.000000	0.006025	0.07723	0.05867
housing	0.28170	0.02069	0.11930	0.006025	1.000000	0.04132	0.50420
loan	0.10650	0.05194	0.08028	0.077230	0.041320	1.00000	0.18280
month	0.11020	0.07232	0.11010	0.058670	0.504200	0.18280	1.00000

EIKONA 36: Υπολογισμός Cramer's V με το πακέτο rcompanion (Salvatore S. Mangiafico, 2025).

Πακέτο vcd: Στην εικόνα 37 αποτυπώνονται τα αποτελέσματα των συσχετίσεων με το μετρό συσχέτισης Cramer's V χρησιμοποιώντας το πακέτο vcd. Αυτό που μπορούμε να αναφέρουμε είναι ότι τόσο το πακέτο vcd όσο και το πακέτο rcompanion δίνουν τα ίδια αποτελέσματα χωρίς καμία απόκλιση στις τιμές τους.

```
> cramers_v_df
```

	job	marital	education	default	housing	loan	month
job	1.00000	0.20600	0.45830	0.036530	0.281700	0.10650	0.11020
marital	0.20600	1.00000	0.12160	0.019230	0.020690	0.05194	0.07232
education	0.45830	0.12160	1.00000	0.015900	0.119300	0.08028	0.11010
default	0.03653	0.01923	0.01590	1.000000	0.006025	0.07723	0.05867
housing	0.28170	0.02069	0.11930	0.006025	1.000000	0.04132	0.50420
loan	0.10650	0.05194	0.08028	0.077230	0.041320	1.00000	0.18280
month	0.11020	0.07232	0.11010	0.058670	0.504200	0.18280	1.00000

EIKONA 37: Υπολογισμός Cramer's V με το πακέτο vcd (Meyer et al.,2024).

Συντελεστή Συσχέτισης Φ (Phi Coefficient):

Ο συντελεστής συσχέτισης Φ (Phi Coefficient) είναι κατάλληλος για την μέτρηση ποιοτικών μεταβλητών και πόσο μάλλον όταν αυτές οι μεταβλητές είναι δυαδικές (binary). Δηλαδή κατάλληλες μεταβλητές για τον Phi Coefficient είναι αυτές που έχουν δύο κατηγορίες (π.χ. "Ναι/Όχι"). Για το μέτρο συσχέτισης Phi Coefficient θα συσχετίσουμε τις μεταβλητές που αφορούν:

- Default (αν ο πελάτης έχει αθετήσει την πληρωμή)
- Housing (αν ο πελάτης έχει στεγαστικό δάνειο)
- Loan (αν ο πελάτης έχει καταναλωτικό δάνειο)
- Y (αν ο πελάτης ανταποκρίθηκε θετικά στην καμπάνια)

Οι μεταβλητές αντλήθηκαν από το Bank Marketing Dataset από το UC Irvine Machine Learning Repository. Στην εικόνα 38 και 39 έχουμε εφαρμόσει εξίσου τις συναρτήσεις `str()` και `head()`. Με την συνάρτηση `str(df_selected)` βλέπουμε ότι το dataset περιλαμβάνει 45.211 παρατηρήσεις και 4 μεταβλητές όπου και οι 4 μεταβλητές είναι κατηγορικές και μάλιστα δυαδικές. Με την συνάρτηση `head(df_selected)` εμφανίζονται οι πρώτες έξι γραμμές παρέχοντας μια γρήγορη επισκόπηση.

```
> df_selected <- data_bank[, c("default", "housing", "loan", "y")]
>
> str(df_selected)
'data.frame': 45211 obs. of 4 variables:
 $ default: chr "no" "no" "no" "no" ...
 $ housing: chr "yes" "yes" "yes" "yes" ...
 $ loan : chr "no" "no" "yes" "no" ...
 $ y : chr "no" "no" "no" "no" ...
```

EIKONA 38: Επισκόπηση δομής data frame με την εντολή `str(df_selected)`.

```
> head(df_selected)
  default housing loan  y
1      no     yes  no no
2      no     yes  no no
3      no     yes yes no
4      no     yes  no no
5      no      no  no no
6      no     yes  no no
```

EIKONA 39: Επισκόπηση δομής data frame με την εντολή `head(df_selected)`.

Ο υπολογισμός του Συντελεστή Φ πραγματοποιήθηκε χρησιμοποιώντας διαφορετικά πακέτα στην R, ώστε να εξετάσουμε τυχόν διαφορές στις μεθόδους υπολογισμού.

Πακέτο psych: Από τα αποτελέσματα της συσχέτισης Φ στην εικόνα 40 παρατηρούμε ότι οι τιμές των συντελεστών αναδεικνύουν ασθενεί ή σχεδόν μηδενικές συσχετίσεις κάτι που σημαίνει ότι τα δεδομένα αυτά δεν δείχνουν σαφείς συνδέσεις μεταξύ δανείων (καταναλωτικό, στεγαστικό) αθέτησής πληρωμών και ανταπόκριση στις τραπεζικές καμπάνιες.

```
> phi_df <- as.data.frame(phi_matrix)
> phi_df
      default housing loan y
default 1.00000 -0.00603 0.0772 -0.0224
housing -0.00603 1.00000 0.0413 -0.1390
loan 0.07720 0.04130 1.0000 -0.0682
y -0.02240 -0.13900 -0.0682 1.0000
```

EIKONA 40: Υπολογισμός Phi Coefficient με το πακέτο psych (Revelle, 2025).

Συνάρτηση Cor: Χρησιμοποιούμε τη συνάρτηση cor για τον υπολογισμό του συντελεστή Φ (Phi), καθώς ο Φ είναι ουσιαστικά ο συντελεστής συσχέτισης Pearson μεταξύ δυο δυαδικών (binary) μεταβλητών. Στην εικόνα 41 μπορούμε να διαπιστώσουμε την ότι τα αποτελέσματα της συνάρτησης Cor είναι ήδια με τα αποτελέσματα του πακέτου psych.

```
> phi_matrix <- cor(df_selected, method = "pearson")
>
> phi_matrix
      default housing loan y
default 1.000000000 -0.006025218 0.07723424 -0.02241897
housing -0.006025218 1.000000000 0.04132287 -0.13917270
loan 0.077234241 0.041322866 1.000000000 -0.06818503
y -0.022418966 -0.139172702 -0.06818503 1.000000000
```

EIKONA 41: Υπολογισμός Phi Coefficient με την συνάρτηση Cor.

Συντελεστή Διόγκωσης της Διακύμανσης (VIF):

Στην ανάλυση του συντελεστή διόγκωσης της διακύμανσης VIF (Variance Inflation Factor), χρησιμοποιούμε τα δεδομένα που παρουσιάζονται στο βιβλίο του Gujarati (2003, σ. 290), τα οποία αφορούν τις ετήσιες αγορές πλαστικού από έναν σημαντικό πελάτη κατασκευαστικής εταιρίας τηλεφωνικών καλωδίων για την περίοδο 1968-1983²⁸. Οι μεταβλητές που χρησιμοποιούνται ως ανεξάρτητες είναι το Ακαθάριστο Εθνικό Προϊόν (GNP), η έναρξη οικοδομικών έργων (Housing Starts), το ποσοστό ανεργίας (unemployment), το βασικό επιτόκιο (Prime Rate), και τα κέρδη σε τηλεφωνικές γραμμές του πελάτη (Line Gains), ενώ η εξαρτημένη μας μεταβλητή είναι οι συνολικές αγορές πλαστικού (Plastic Purchases). Η ανάλυση πραγματοποιείται στη γλώσσα προγραμματισμού R, με τη χρήση της εντολής vif()

²⁸ Πηγή δεδομένων: Gujarati, D. N. (2003). Basic Econometrics (4th ed.). McGraw-Hill.

Διαθέσιμο στο: https://www.academia.edu/40263427/BASIC_ECONOMETRICS_FOURTH_EDITION.

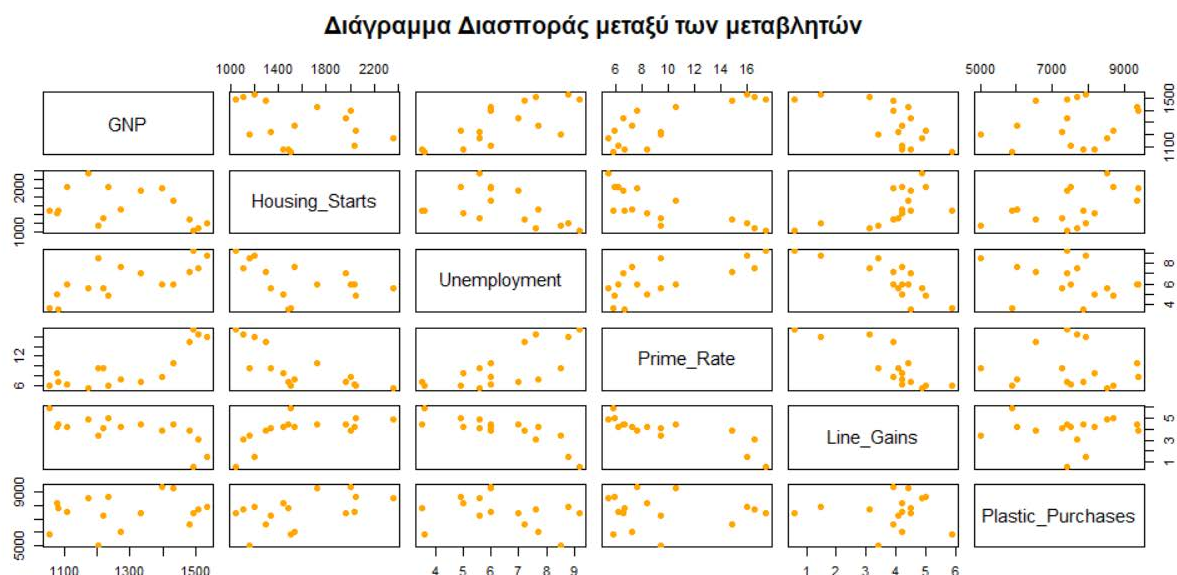
από το πακέτο car, το οποίο υπολογίζει τον συντελεστή VIF για κάθε ανεξάρτητη μεταβλητή. Ο συντελεστής αυτός μετράει τον βαθμό πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών, προσδιορίζοντας κατά πόσο η κάθε μία εξηγείται γραμμικά από τις υπόλοιπες. Στην εικόνα 42 παρουσιάζεται η δομή του πλαισίου δεδομένων df_model, η οποία περιλαμβάνει 16 παρατηρήσεις και 6 αριθμητικές μεταβλητές, όλες numeric. Στην εικόνα 43 παρουσιάζονται οι πρώτες έξι γραμμές του πίνακα δεδομένων, όπου φαίνονται ενδεικτικά οι τιμές για κάθε μεταβλητή.

```
> str(df_model)
'data.frame': 16 obs. of 6 variables:
 $ GNP : num 1052 1079 1075 1108 1171 ...
 $ Housing_Starts : num 1504 1487 1435 2036 2361 ...
 $ Unemployment : num 3.6 3.5 5 6 5.6 4.9 5.6 8.5 7.7 7 ...
 $ Prime_Rate : num 5.8 6.7 8.4 6.2 5.4 5.9 9.4 9.4 7.2 6.6 ...
 $ Line_Gains : num 5.9 4.5 4.2 4.2 4.9 5 4.1 3.4 4.2 4.5 ...
 $ Plastic_Purchases: num 5873 7852 8189 7497 8534 ...
```

EIKONA 42: Επισκόπηση δομής data frame με την εντολή str(df_model).

```
> head(df_model)
   GNP Housing_Starts Unemployment Prime_Rate Line_Gains Plastic_Purchases
1 1051.8      1503.6         3.6         5.8         5.9         5873
2 1078.8      1486.7         3.5         6.7         4.5         7852
3 1075.3      1434.8         5.0         8.4         4.2         8189
4 1107.5      2035.6         6.0         6.2         4.2         7497
5 1171.1      2360.8         5.6         5.4         4.9         8534
6 1235.0      2043.9         4.9         5.9         5.0         8688
```

EIKONA 43: Επισκόπηση δομής data frame με την εντολή head(df_model).



EIKONA 44: Πολλαπλό Διάγραμμα διασποράς μεταξύ των μεταβλητών GNP, Housing Starts, Unemployment, Prime Rate, Line Gains και Plastic Purchase.

Στην εικόνα 45 βλέπουμε ότι οι μεταβλητές GNP, Housing Starts, Unemployment και Line Gains παρουσιάζουν τιμές VIF μικρότερες του 10, υποδεικνύοντας αποδεκτά επίπεδα συσχέτισης. Ωστόσο, η μεταβλητή Prime Rate παρουσιάζει τιμή VIF ίση 14.68, ξεπερνώντας το κρίσιμο όριο του 10, το οποίο αποτελεί ένδειξη ισχυρής πολυσυγγραμμικότητας.

```
> model_vif <- lm(Plastic_Purchases ~ ., data = df_model)
>
> # Υπολογισμός VIF
> vif(model_vif)
```

	GNP	Housing_Starts	Unemployment	Prime_Rate	Line_Gains
	6.905160	4.344945	3.967917	14.683046	5.423499

ΕΙΚΟΝΑ 45: Υπολογισμός συντελεστή Variance Inflation Factor με το πακέτο car (Fox J & Weisberg S, 2019) .

Ιδιοτιμές (Eigenvalues):

Στο πλαίσιο της διάγνωσης πολυσυγγραμμικότητας, πραγματοποιούμε ανάλυση ιδιοτιμών (Eigenvalues) του πίνακα των ανεξάρτητων μεταβλητών. Χρησιμοποιούμε τις ίδιες μεταβλητές με εκείνες που χρησιμοποιήθηκαν στην ανάλυση του συντελεστή VIF, ώστε να διατηρηθεί η αναλυτική συνέπεια και να είναι δυνατή η άμεση σύγκριση των αποτελεσμάτων. Συγκεκριμένα εξετάζονται οι μεταβλητές GNP, Housing Starts, Unemployment, Prime Rate και Line Gains από το παράδειγμα του Gujarati (2003, σ.290) οι οποίες αφορούν οικονομικά και επιχειρηματικά χαρακτηριστικά σχετιζόμενα με τις ετήσιες αγορές πλαστικού από έναν σημαντικό πελάτη τηλεπικοινωνιακής εταιρίας για την χρονική περίοδο 1968 έως 1983.

Το άθροισμα των eigenvalues είναι ίσο με το άθροισμα των μεταβλητών που συσχετίσαμε. Δηλαδή $\text{Eigenvalues} = 3.71684128 + 0.71055658 + 0.35341055 + 0.17958812 + 0.03960348 = 5$. Επομένως μπορούμε να δούμε το ποσοστό διακύμανση που έχει κάθε συνιστώσα.

1. Συνιστώσα (Eigenvalue): Ποσοστό Διακύμανσης $= \frac{3.71684128}{5} \times 100 = 74,33\%$.
Βλέπουμε ότι η πρώτη συνιστώσα συγκεντρώνει το 74,33% της συνολικής διακύμανσης εξηγεί, δηλαδή το μεγαλύτερο ποσοστό διακύμανσης.
2. Συνιστώσα (Eigenvalue): Ποσοστό Διακύμανσης: $= \frac{0.71055658}{5} \times 100 = 14.21\%$. Η δεύτερη συνιστώσα συγκεντρώνει το 14,21% της συνολικής διακύμανσης.
3. Συνιστώσα (Eigenvalue): Ποσοστό Διακύμανσης $= \frac{0.35341055}{5} \times 100 = 7,06\%$. Η Τρίτη συνιστώσα συγκεντρώνει το 7,06% της συνολικής διακύμανσης.

4. Συνιστώσα (Eigenvalue): Ποσοστό Διακύμανσης = $\frac{0.17958812}{5} \times 100 = 3,59\%$. Η Τρίτη συνιστώσα συγκεντρώνει το 1.41% της συνολικής διακύμανσης.

5. Συνιστώσα (Eigenvalue): Ποσοστό Διακύμανσης = $\frac{0.03960348}{5} \times 100 = 0.79\%$. Η Τρίτη συνιστώσα συγκεντρώνει το 0.79% της συνολικής διακύμανσης.

Η πρώτη και η δεύτερη συνιστώσα εξηγούν συνολικά το 88,54% της συνολικής διακύμανσης καθιστώντας τες σημαντικές για την συγκεκριμένη ανάλυση.

Συνάρτηση eigen():

```
> X_vars <- df_model[, c("GNP", "Housing_Starts", "Unemployment", "Prime_Rate", "Line_G$
> cor_matrix <- cor(X_vars)
> print(cor_matrix)
      GNP Housing_Starts Unemployment Prime_Rate Line_Gains
GNP      1.0000000    -0.3496099    0.7265714    0.8098949   -0.6743040
Housing_Starts -0.3496099    1.0000000   -0.4633229   -0.7486925    0.6061996
Unemployment  0.7265714   -0.4633229    1.0000000    0.7009449   -0.8088545
Prime_Rate    0.8098949   -0.7486925    0.7009449    1.0000000   -0.8308159
Line_Gains   -0.6743040    0.6061996   -0.8088545   -0.8308159    1.0000000
> eig_vals <- eigen(cor_matrix)$values
> print(eig_vals)
[1] 3.71684128 0.71055658 0.35341055 0.17958812 0.03960348
```

Εικόνα 46: Αποτελέσματα Eigenvalues με την συνάρτηση eigen() της R.

Συντελεστή Ανοχής Tolerance Coefficient (TOL):

Για την αξιολόγηση της πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών, υπολογίσαμε τους συντελεστές Tolerance, χρησιμοποιώντας τις ίδιες μεταβλητές όπως και στις προηγούμενες αναλύσεις με τον VIF και τις ιδιοτιμές (Eigenvalues), προκειμένου να διατηρηθεί η αναλυτική συνέπεια και να είναι δυνατή η σύγκριση των αποτελεσμάτων. Συγκεκριμένα, χρησιμοποιήθηκαν οι μεταβλητές GNP, Housing Starts, Unemployment, Prime Rate και Line Gains, όπως παρουσιάζονται στο παράδειγμα του Gujirati (2003, σ. 290), για την περίοδο 1968-1983. Η ανάλυση πραγματοποιήθηκε στη γλώσσα προγραμματισμού R, με τον υπολογισμό του Tolerance ως το αντίστροφο του VIF. Τα αποτελέσματα παρουσιάζονται στην εικόνα 47 και δείχνουν ότι η μεταβλητή Prime Rate έχει συντελεστή Tolerance ίσο με 0.068, τιμή μικρότερη από το κρίσιμο όριο του 0.1. Η τιμή αυτή υποδεικνύει ισχυρή γραμμική εξάρτηση της συγκεκριμένης μεταβλητής και συνεπώς σαφή ένδειξη πολυσυγγραμμικότητας. Αντίθετα οι υπόλοιπες μεταβλητές παρουσιάζουν Tolerance μεγαλύτερο του 0.1, γεγονός που υποδεικνύει αποδεκτό επίπεδο ανεξαρτησίας μεταξύ τους.

```
> model_vif <- lm(Plastic_Purchases ~ ., data = df_model)
> vif_values <- vif(model_vif)
> tolerance_values <- 1 / vif_values
> print(tolerance_values)
```

GNP	Housing_Starts	Unemployment	Prime_Rate	Line_Gains
0.14481923	0.23015249	0.25202143	0.06810576	0.18438282

EIKONA 47: Αποτελέσματα του συντελεστή Tolerance.

Δείκτης κατάστασης (Condition Index - CI)

Στο πλαίσιο της εξερεύνησης της πολυσυγγραμμικότητας, υπολογίσαμε τον συντελεστή Condition Index (CI), χρησιμοποιώντας τις ίδιες μεταβλητές που χρησιμοποιήσαμε στον συντελεστή VIF, Eigenvalues και Tolerance. Η διαδικασία πραγματοποιήθηκε στην γλώσσα R, με τα αποτελέσματα να απεικονίζονται στην εικόνα 58. Η μέγιστη τιμή CI (9.68) που εμφανίζεται στην μεταβλητή Line Gains δεν υπερβαίνει το κρίσιμο όριο του 10 και αυτό σημαίνει ότι δεν υπάρχει σοβαρή πολυσυγγραμμικότητα. Τα αποτελέσματα από τον συντελεστή Condition Index αποδεικνύει ότι είναι ένα χρήσιμο εργαλείο στον εντοπισμό της πολυσυγγραμμικότητας σε συνδυασμό με τους δείκτες VIF, Tolerance, Eigenvalues.

```
> X_vars <- df_model[, c("GNP", "Housing_Starts", "Unemployment", "Prime_Rate", "Line_G$
> cor_matrix <- cor(X_vars)
> eig_vals <- eigen(cor_matrix)$values
> condition_index <- sqrt(max(eig_vals) / eig_vals)
> condition_index
[1] 1.000000 2.287113 3.243002 4.549338 9.687692
```

EIKONA 48: Αποτελέσματα του Δείκτη κατάστασης.

Canonical Correlation Analysis (CCA)

Ο συντελεστής συσχέτισης που δίνει η Canonical Correlation Analysis αποτελεί ένα σημαντικό μέτρο καθώς χρησιμοποιείται για να μετρήσει τη σχέση μεταξύ δύο σύνολα μεταβλητών, σε αντίθεση με τα συνηθισμένα μέτρα που συσχετίζουν δύο μεμονωμένες μεταβλητές. Για το μέτρο συσχέτισης της Canonical Correlation Analysis θα συσχετίσουμε την ομάδα 1 που περιλαμβάνει τις μεταβλητές ΑΕΠ, ανεργία, εισαγωγές και εξαγωγές της Ελλάδας με την ομάδα 2 που περιλαμβάνει αντίστοιχα το ΑΕΠ, την Ανεργία, τις εισαγωγές και τις εξαγωγές της Ευρώπης. Τα δεδομένα συλλέχθηκαν από την βάση δεδομένων FRED που αφορούν την Ελλάδα από το 1/1/2005 έως το 1/1/2024. Από την εικόνα 49 βλέπουμε ότι:

Πακέτο cca:

```
> print(cca_result$cor) # Αυτό δείχνει τις canonical correlations
[1] 0.9798217 0.9615387 0.6180374 0.2105565
>
> # Εμφάνιση και των canonical weights (φορτίσεις)
> print(cca_result$xcoef) # Βάρη για την ομάδα 1
      [,1]      [,2]      [,3]      [,4]
GDP      -1.1519420  0.1635944  2.584401  0.08155593
UNEMPLOYMENT -0.3911602  0.8386208  1.121205 -2.37084934
IMPORTS     0.3660348 -0.2942927 -1.466451 -2.79652520
EXPORTS     0.4166524 -0.1911752  1.639251  1.19304971
> print(cca_result$ycoef) # Βάρη για την ομάδα 2
      [,1]      [,2]      [,3]      [,4]
GDP_EU      -0.6425947 -0.4882678  3.845568 -1.5257798
UNEMPLOYMENT_EU -0.4050495  0.8425866  1.178648  0.2889705
IMPORTS_EU     -1.6246946 -0.8856990  3.348819  8.8483662
EXPORTS_EU      3.0912156  1.6558311 -6.480153 -7.1629946
```

Εικόνα 49: Αποτελέσματα της συσχέτισης Canonical Correlation Analysis μεταξύ το σύνολο των μεταβλητών της ομάδας 1 με το σύνολο των μεταβλητών της ομάδας 2 με το πακέτο CCA (Ignacio González & Déjean Maintainer Sébastien Déjean, 2025).

Πακέτο ccp:

```
> print(cca_result$cor) # Οι canonical συσχετίσεις
[1] 0.9798217 0.9615387 0.6180374 0.2105565
> print(cca_result$xcoef) # Συντελεστές για το df1
      [,1]      [,2]      [,3]      [,4]
GDP      -1.1519420  0.1635944  2.584401  0.08155593
UNEMPLOYMENT -0.3911602  0.8386208  1.121205 -2.37084934
IMPORTS     0.3660348 -0.2942927 -1.466451 -2.79652520
EXPORTS     0.4166524 -0.1911752  1.639251  1.19304971
> print(cca_result$ycoef) # Συντελεστές για το df2
      [,1]      [,2]      [,3]      [,4]
GDP_EU      -0.6425947 -0.4882678  3.845568 -1.5257798
UNEMPLOYMENT_EU -0.4050495  0.8425866  1.178648  0.2889705
IMPORTS_EU     -1.6246946 -0.8856990  3.348819  8.8483662
EXPORTS_EU      3.0912156  1.6558311 -6.480153 -7.1629946
> n <- nrow(df1_std) # Παρατηρήσεις
> p <- ncol(df1_std) # Μεταβλητές Ομάδας 1
> q <- ncol(df2_std) # Μεταβλητές Ομάδας 2
>
> # P-values
> p_values <- p.asym(cca_result$cor, n, p, q, tstat = "Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
  stat      approx df1      df2      p.value
1 to 4:  0.001780112 81.148263  16 186.9957 0.000000e+00
2 to 4:  0.044558094 43.472870   9 151.0423 0.000000e+00
3 to 4:  0.590629963  9.487631   4 126.0000 9.751855e-07
4 to 4:  0.955665955  2.969007   1  64.0000 8.970275e-02
```

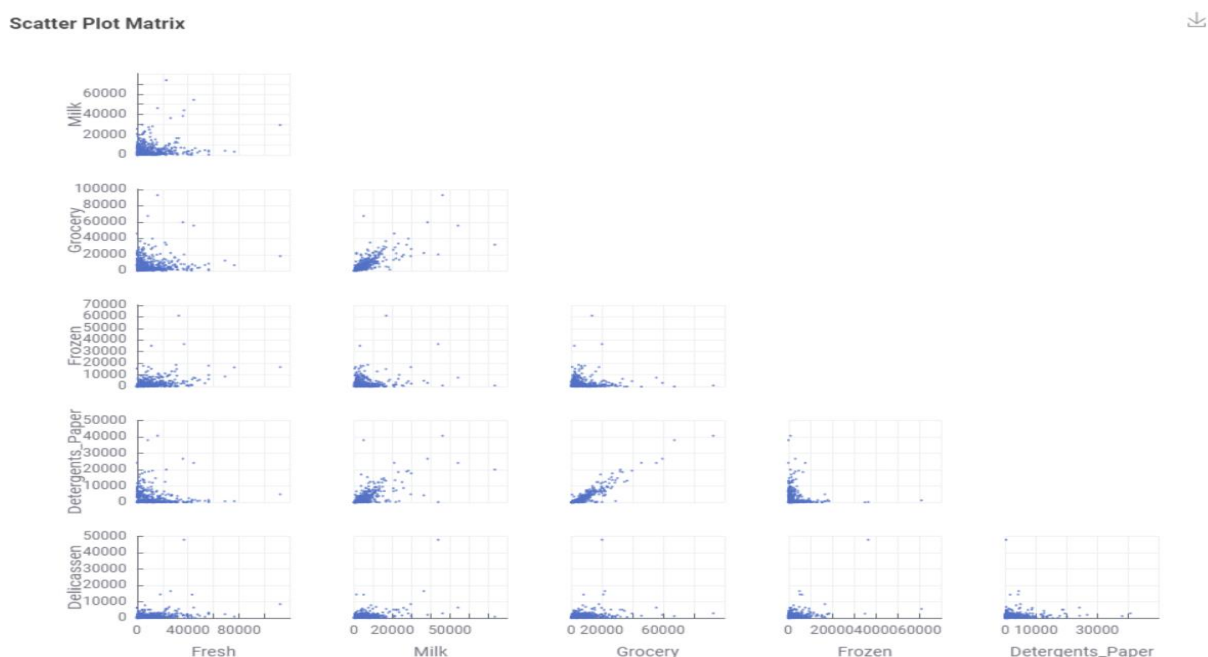
Εικόνα 50: Αποτελέσματα της συσχέτισης Canonical Correlation Analysis μεταξύ το σύνολο των μεταβλητών της ομάδας 1 με το σύνολο των μεταβλητών της ομάδας 2 με το πακέτο CCP (Uwe Menzel, 2022).

3.2.2 Υπολογιστική προσέγγιση μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

Στην υπό ενότητα αυτή, η ανάλυση επεκτείνεται με τη χρήση της πλατφόρμας KNIME, όπου μέσω των κατάλληλων κόμβων (Fillbrunn et al., 2017), υπολογίζονται επιλεγμένα μέτρα συσχέτισης παρέχοντας μια εναλλακτική, αλλά ισοδύναμα αξιόπιστη προσέγγιση στην εμπειρική διερεύνηση των μεταβλητών.

Συντελεστής Συσχέτισης Pearson:

Πριν την παρουσίαση των αποτελεσμάτων, αξίζει να σημειωθεί ότι η πλατφόρμα KNIME Analytics Platform παρέχει τη δυνατότητα γραφικής αναπαράστασης της σχέσης μεταξύ μεταβλητών μέσω του κόμβου Scatter Plot Matrix. Όπως φαίνεται και στην εικόνα 51, η πλατφόρμα KNIME μπορεί να δημιουργήσει διάγραμμα διασποράς το ίδιο ισοδύναμο με τη γλώσσα προγραμματισμού R, προσφέροντας άμεση απεικόνιση των μοτίβων συσχέτισης μεταξύ των μεταβλητών.



ΕΙΚΟΝΑ 51: Διάγραμμα Διασποράς (Scatterplot Matrix) μεταξύ των μεταβλητών (Fresh, Milk, Grocery, Frozen, Detergents_Paper και Delicassen) μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

Η ανάλυση που πραγματοποιήθηκε μέσω της πλατφόρμας KNIME Analytics Platform επιβεβαιώνει ότι οι υπολογισμοί μπορούν να εκτελεστούν εξίσου αξιόπιστα και σε εναλλακτικά υπολογιστικά περιβάλλοντα, πέραν της γλώσσας R. Συγκεκριμένα χρησιμοποιήθηκε το σύνολο δεδομένων “Wholesale customers”²⁹ από το UCI Machine Learning Repository, το οποίο περιλαμβάνει πληροφορίες για της ετήσιες δαπάνες πελατών σε διάφορες κατηγορίες προϊόντων. Η ανάλυση περιορίστηκε στις έξι ποσοτικές μεταβλητές Fresh, Milk, Grocery, Frozen, Detergents_Paper και Delicassen, καθώς αυτές πληρούν τις προϋποθέσεις για εφαρμογή του Pearson Correlation Coefficient. Όπως αποτυπώνεται στην εικόνα 52, ο

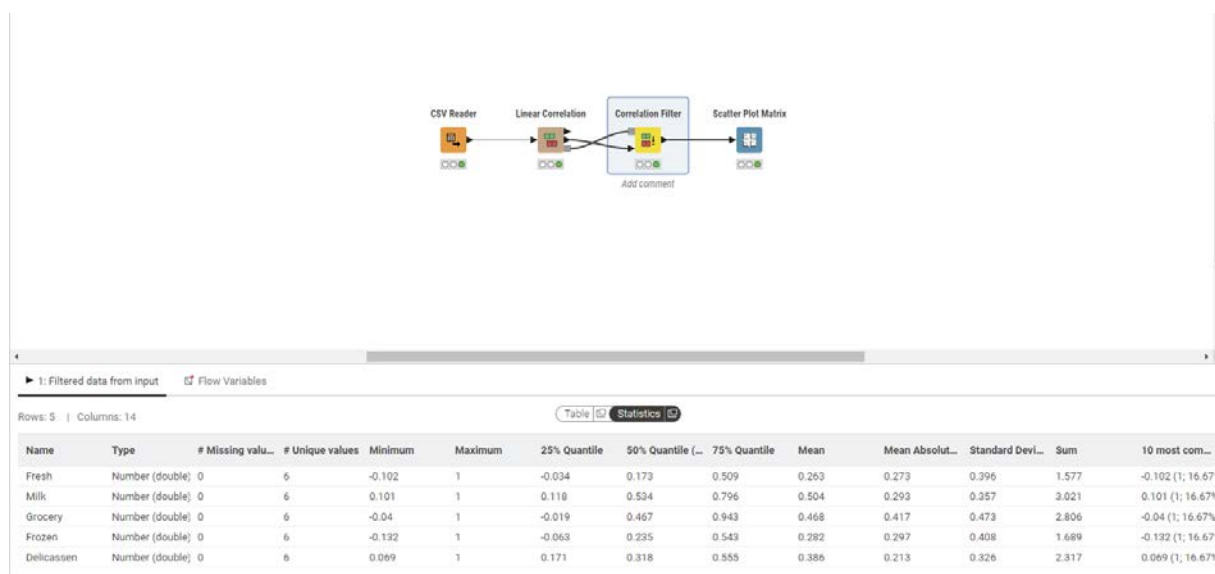
²⁹ Πηγή δεδομένων: <https://archive.ics.uci.edu/dataset/292/wholesale+customers>.

υπολογισμός πραγματοποιήθηκε με τον κόμβο Linear Correlation, ο οποίος παρέχει έναν πλήρη πίνακα γραμμικών συσχετίσεων μεταξύ των μεταβλητών.

#	RowID	Fresh Number (double)	Milk Number (double)	Grocery Number (double)	Frozen Number (double)	Detergents_Paper Number (double)	Delicassen Number (double)
1	Fresh	1	0.101	-0.012	0.346	-0.102	0.245
2	Milk	0.101	1	0.728	0.124	0.662	0.406
3	Grocer	-0.012	0.728	1	-0.04	0.925	0.205
4	Frozen	0.346	0.124	-0.04	1	-0.132	0.391
5	Deterg	-0.102	0.662	0.925	-0.132	1	0.069
6	Delicas	0.245	0.406	0.205	0.391	0.069	1

ΕΙΚΟΝΑ 52: Υπολογισμός Pearson Correlation Coefficient μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

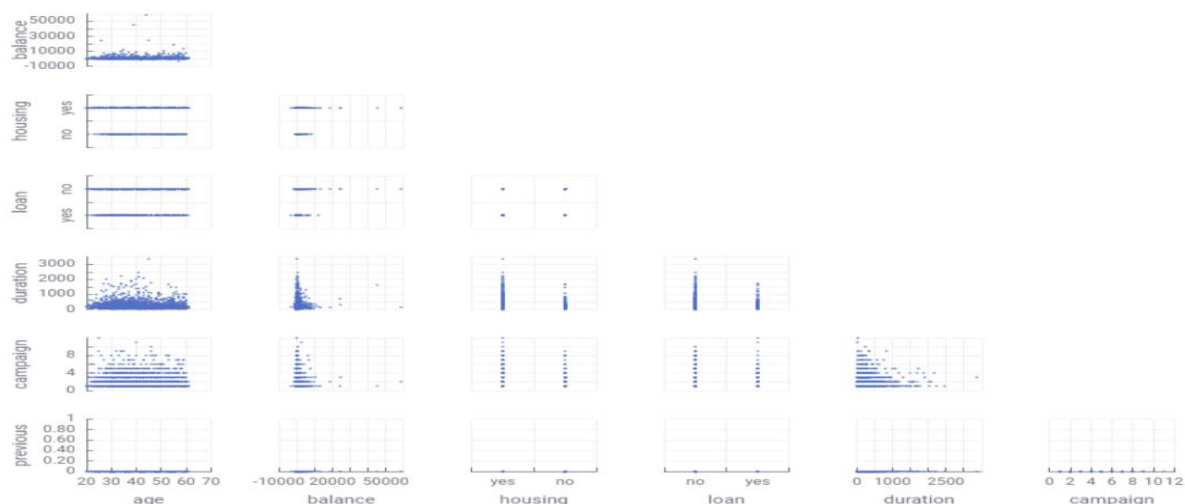
Στο πλαίσιο της ανάλυσης με το KNIME Analytics Platform, εφαρμόστηκε αρχικά ο συντελεστής συσχέτισης Pearson, όπως παρουσιάστηκε και στην εικόνα 52, μέσω του κόμβου Linear Correlation, με σκοπό να εντοπιστούν ισχυρές γραμμικές σχέσεις μεταξύ των ποσοτικών μεταβλητών του συνόλου Wholesale Customers. Στην συνέχεια, για την βελτιστοποίηση της ανάλυσης και τη μείωση της πολυπλοκότητας των δεδομένων, αξιοποιήθηκε ο κόμβος Correlation filter (threshold 0.75) όπως αποτυπώνεται στην εικόνα 53. Με λίγα λόγια ο συγκεκριμένος κόμβος εντοπίζει τις μεταβλητές που εμφανίζουν συσχέτιση μεγαλύτερη μιας τιμής που ο χρήστης ορίζει ($|r| > 0.75$) και διατηρεί μόνο μία αντιπροσωπευτική στήλη από κάθε ομάδα συσχετίσεων, απομακρύνοντας τις υπόλοιπες. Στο τελικό σύνολο δεδομένων περιλαμβάνει τις μεταβλητές Fresh, Milk, Grocery, Frozen και Delicassen, καθώς μόνο η μεταβλητή Detergents_Paper απομακρύνθηκε λόγω υψηλής συσχέτισης με το Grocery.



ΕΙΚΟΝΑ 53: Εφαρμογή Pearson Correlation Coefficient με Correlation Filter (threshold 0.75) μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

Συντελεστής Συσχέτισης *Spearman's Rho*:

Για τον συντελεστή συσχέτισης θα ακολουθήσουμε ακριβώς την ίδια μεθοδολογία δηλαδή πριν την ανάλυση αποτελεσμάτων προχωράμε στην γραφική αναπαράσταση των μεταβλητών μέσω του κόμβου Scatter Plot Matrix που μπορούμε να δούμε στην εικόνα 54.



ΕΙΚΟΝΑ 54: Διάγραμμα Διασποράς (Scatterplot Matrix) μεταξύ των μεταβλητών (age, previous, campaign, duration, loan, housing και balance) μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

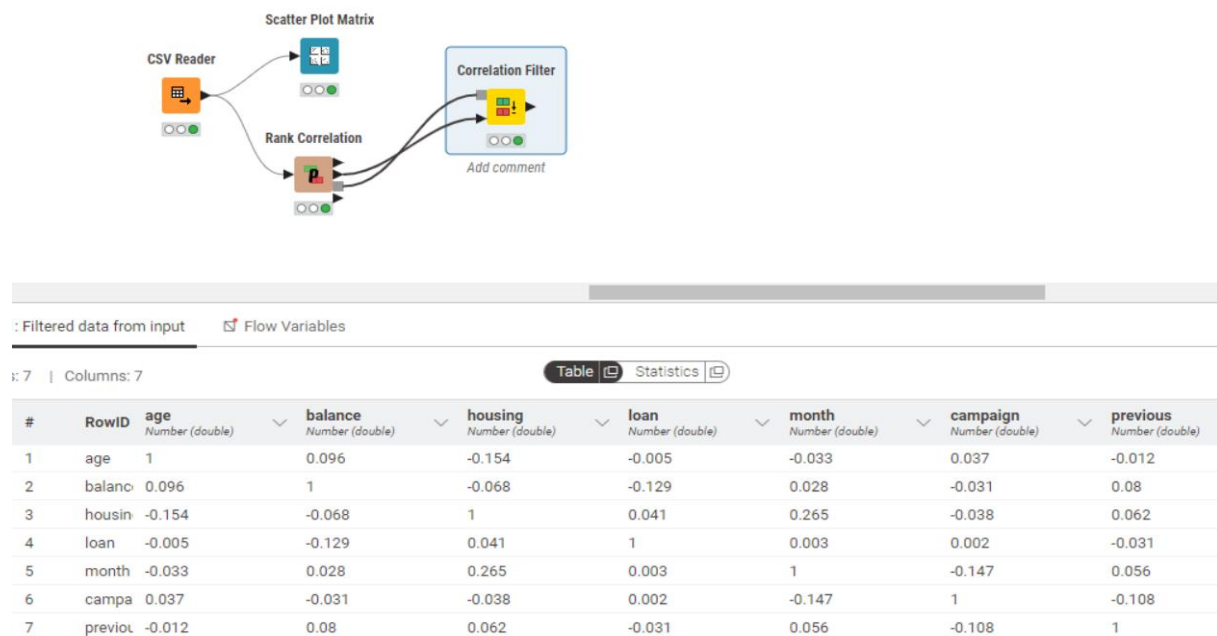
Στην εικόνα 59 βλέπουμε ότι ο συντελεστής συσχέτισης *Spearman* υπολογίστηκε και μέσω του KNIME Analytics Platform. Μέσω του κόμβου Rank Correlation, εφαρμόστηκε η μέθοδος *Spearman* στις ίδιες μεταβλητές του Bank Marketing Dataset, όπως και στη R. Διαπιστώνουμε ότι ακόμη και με διαφορετικά υπολογιστικά περιβάλλοντα προκύπτουν τα ίδια στατιστικά συμπεράσματα.

RowID	age Number (double)	balance Number (double)	housing Number (double)	loan Number (double)	month Number (double)	campaign Number (double)	previous Number (double)
age	1	0.096	-0.154	-0.005	-0.033	0.037	-0.012
balanc	0.096	1	-0.068	-0.129	0.028	-0.031	0.08
housin	-0.154	-0.068	1	0.041	0.265	-0.038	0.062
loan	-0.005	-0.129	0.041	1	0.003	0.002	-0.031
month	-0.033	0.028	0.265	0.003	1	-0.147	0.056
campa	0.037	-0.031	-0.038	0.002	-0.147	1	-0.108
previou	-0.012	0.08	0.062	-0.031	0.056	-0.108	1

ΕΙΚΟΝΑ 55: Υπολογισμός *Spearman Correlation Coefficient* στη γλώσσα προγραμματισμού KNIME.

Αφού εφαρμόστηκε ο συντελεστής συσχέτισης *Spearman* που είδαμε στην εικόνα 54 στην συνέχεια θα βελτιστοποιήσουμε την ανάλυση μας. Στην εικόνα 55 χρησιμοποιούμε τον

κόμβο Correlation filter (threshold 0.75), που εντοπίζει τις μεταβλητές που εμφανίζουν συσχέτιση $|r| > 0.75$ έτσι ώστε να μπορέσει να τις απομακρύνει. Η διαδικασία αυτή είναι πολύ χρήσιμη καθώς αναδεικνύει τη σημασία της επεξεργασίας των δεδομένων.

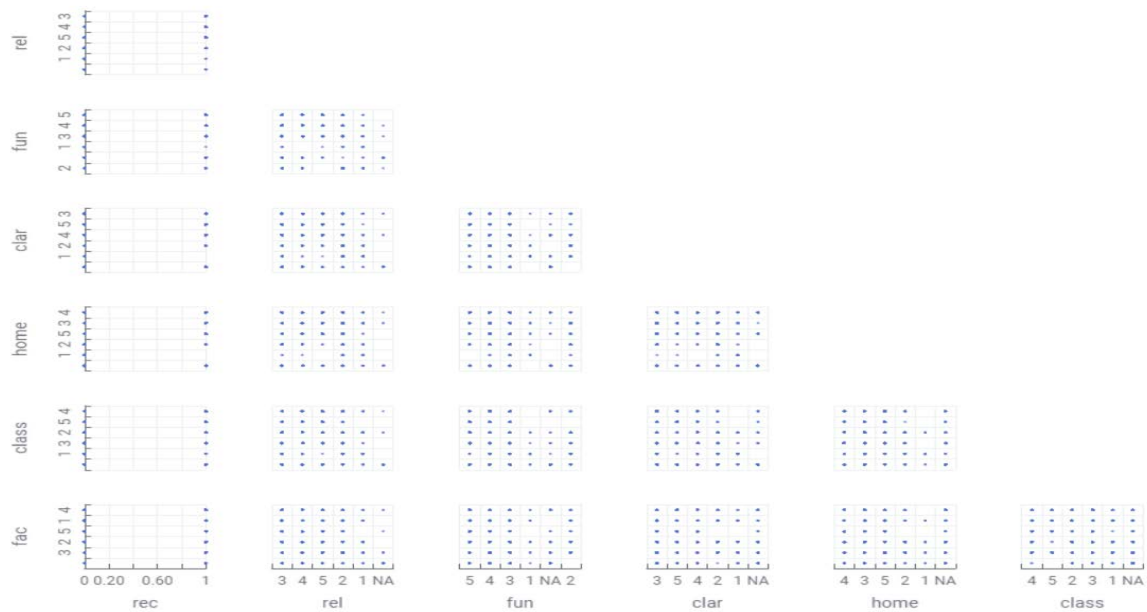


EIKONA 56:Εφαρμογή Spearman Correlation Coefficient με Correlation Filter (threshold 0.75) στη γλώσσα προγραμματισμού KNIME.

Συντελεστής Συσχέτισης Kendall's Tau:

Στην εικόνα 57 απεικονίζεται το διάγραμμα διασποράς που έγινε με την πλατφόρμα KNIME. Οι μεταβλητές μας είναι ταξινομημένες σε κλίμακα Likert (1-5) και προέρχονται από το σύνολο δεδομένων learning. Μέσω του κόμβου Scatter Plot Matrix, γίνεται οπτική διερεύνηση των μεταξύ τους συσχετίσεων, επιτρέποντας μια πρώτη εικόνα της μορφής των σχέσεων.

Scatter Plot Matrix



EIKONA 57: Διάγραμμα Διασποράς (Scatterplot Matrix) μεταξύ των μεταβλητών (rec, rel, fun, clar, home, class, fac) μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

Η ανάλυση του συντελεστή Kendall's Tau που πραγματοποιήθηκε μέσω της αναλυτικής πλατφόρμας KNIME επιβεβαιώνει ότι οι υπολογισμοί μπορούν να εκτελεστούν εξίσου αξιόπιστα και σε εναλλακτικά υπολογιστικά περιβάλλοντα, πέραν της γλώσσας προγραμματισμού R. Χρησιμοποιήθηκαν οι ίδιες μεταβλητές από το dataset learning που επιλέχθηκαν και για την ανάλυση στην R, οι οποίες βασίζονται σε κλίμακα Likert και αποτυπώνουν τη γνώμη των συμμετεχόντων σχετικά με την εμπειρία τους σε προγράμματα μάθησης. Όπως αποτυπώνεται και στην εικόνα 58, η συσχέτιση υπολογίστηκε με τον κόμβο Rank Correlation, επιλέγοντας ως μέθοδο τον Kendall's Tau.

#	RowID	rec Number (double)	rel Number (double)	fun Number (double)	clar Number (double)	home Number (double)	class Number (double)	fac Number (double)
1	rec	1	0.272	0.109	0.225	0.078	0.177	0.164
2	rel	0.272	1	0.282	0.395	0.249	0.283	0.226
3	fun	0.109	0.282	1	0.355	0.319	0.26	0.23
4	clar	0.225	0.395	0.355	1	0.349	0.375	0.279
5	home	0.078	0.249	0.319	0.349	1	0.311	0.182
6	class	0.177	0.283	0.26	0.375	0.311	1	0.287
7	fac	0.164	0.226	0.23	0.279	0.182	0.287	1

EIKONA 58: Υπολογισμός Kendall's Tau Correlation Coefficient μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

Αφού εφαρμόστηκε ο συντελεστής συσχέτισης Kendall's Tau που είδαμε στην εικόνα 58 στην συνέχεια θα βελτιστοποιήσουμε την ανάλυση μας. Στην εικόνα 59 χρησιμοποιούμε τον κόμβο Correlation filter (threshold 0.75), που εντοπίζει τις μεταβλητές που εμφανίζουν

συσχέτιση $|r| > 0.75$ έτσι ώστε να μπορέσει να τις απομακρύνει. Η διαδικασία αυτή είναι πολύ χρήσιμη καθώς αναδεικνύει τη σημασία της επεξεργασίας των δεδομένων.

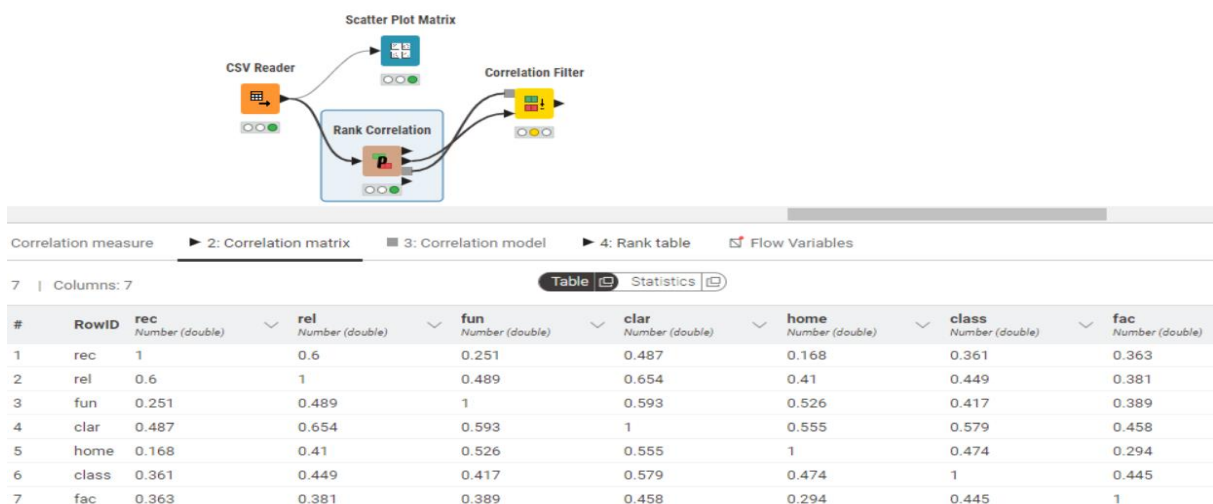
Rows: 7 | Columns: 7

☐	#	RowID	rec Number (double)	rel Number (double)	fun Number (double)	clar Number (double)	home Number (double)	class Number (double)	fac Number (double)
☐	1	rec	1	0.272	0.109	0.225	0.078	0.177	0.164
☐	2	rel	0.272	1	0.282	0.395	0.249	0.283	0.226
☐	3	fun	0.109	0.282	1	0.355	0.319	0.26	0.23
☐	4	clar	0.225	0.395	0.355	1	0.349	0.375	0.279
☐	5	home	0.078	0.249	0.319	0.349	1	0.311	0.182
☐	6	class	0.177	0.283	0.26	0.375	0.311	1	0.287
☐	7	fac	0.164	0.226	0.23	0.279	0.182	0.287	1

ΕΙΚΟΝΑ 59: Εφαρμογή Kendall's Tau Correlation Coefficient με Correlation Filter (threshold 0.75) μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

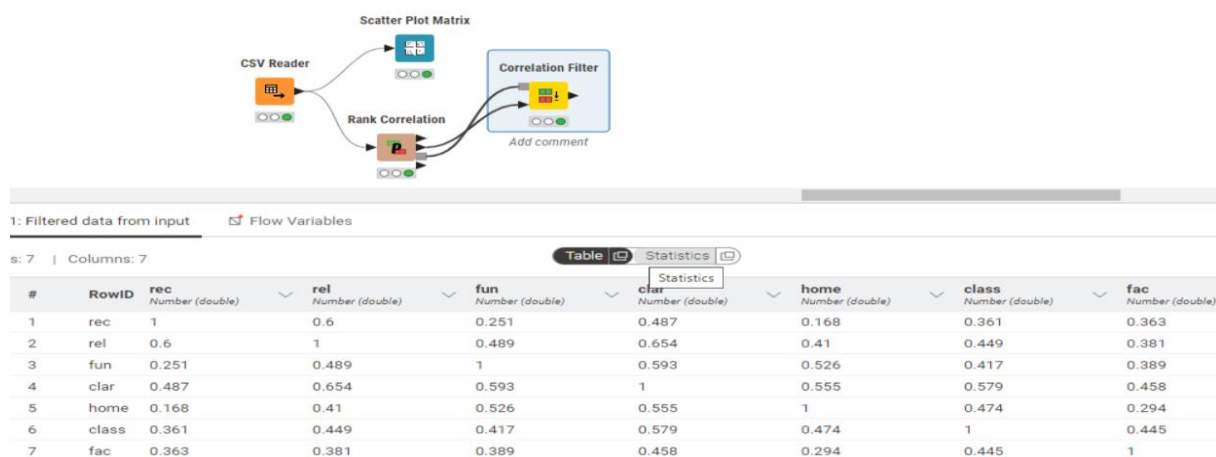
Συντελεστής Συσχέτισης Gamma:

Η ανάλυση του συντελεστή συσχέτισης Gamma πραγματοποιήθηκε μέσω της αναλυτικής πλατφόρμας KNIME αξιοποιώντας το ίδιο σύνολο μεταβλητών από το dataset learning, όπως και στην περίπτωση Kendall's Tau. Όπως παρουσιάζεται και στην εικόνα 60 ο υπολογισμός έγινε μέσω του κόμβου Rank Correlation, επιλέγοντας αυτή την φορά ως μέτρο συσχέτισης τον Gamma correlation. Ο συντελεστής αυτός είναι ιδιαίτερα κατάλληλος όταν υπάρχουν πολλές ισοβαθμίες (ties) στα δεδομένα, όπως συμβαίνει συχνά με ordinal μεταβλητές. Η χρήση του KNIME ενισχύει την ανάλυση χωρίς την ανάγκη κώδικα, προσφέροντας έναν απλό αλλά ισοδύναμο τρόπο προσέγγισης.



ΕΙΚΟΝΑ 60: Υπολογισμός Gamma Correlation Coefficient μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

Αφού προηγουμένως εφαρμόστηκε ο συντελεστής Gamma, όπως παρουσιάστηκε στην εικόνα 60, θα προχωρήσουμε στην βελτιστοποίηση της ανάλυσης με σκοπό τη μείωση της πλεονάζουσας πληροφορίας. Στην εικόνα 61 χρησιμοποιήθηκε ο κόμβος Correlation Filter με μέγιστο όριο συσχέτισης 0.75, ο οποίος εντοπίζει και απορρίπτει μεταβλητές που παρουσιάζουν υψηλή μεταξύ τους συσχέτιση $|r| > 0.75$. Η διαδικασία αυτή ενισχύει τη διαγνωστική ισχύ του μοντέλου και διευκολύνει την ερμηνεία των αποτελεσμάτων, αναδεικνύοντας τον κρίσιμο ρόλο του Correlation Filter στις αναλύσεις συσχέτισης με ordinal δεδομένα.



ΕΙΚΟΝΑ 61: Εφαρμογή Kendall's Tau Correlation Coefficient με Correlation Filter (threshold 0.75) μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME.

3.2.3 Υπολογισμός μέτρων συσχέτισης μέσω ενσωμάτωσης της γλώσσας προγραμματισμού R στην πλατφόρμα αναλυτικής δεδομένων KNIME.

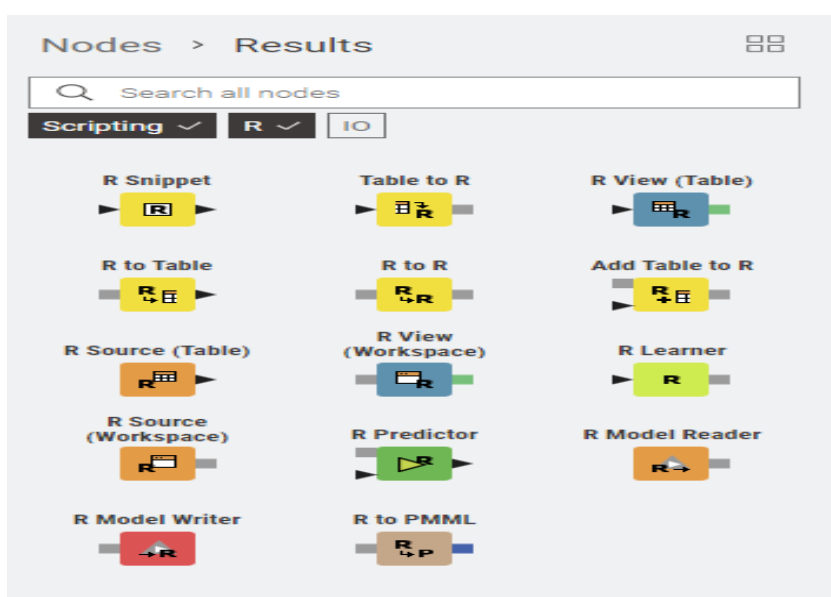
Παρότι η πλατφόρμα KNIME παρέχει έτοιμους κόμβους για την εφαρμογή βασικών μεθόδων συσχέτισης όπως ο Pearson, υπάρχουν μέτρα που δεν υποστηρίζονται από την πλατφόρμα αναλυτικής δεδομένων KNIME. Για την επέκταση αυτών των δυνατοτήτων, αξιοποιείται η δυνατότητα ενσωμάτωσης της γλώσσας προγραμματισμού R εντός του περιβάλλοντος KNIME.

Μέσω της χρήσης των κόμβων R Snippet υπάρχει δυνατότητα να καλέσουμε εξειδικευμένες συναρτήσεις της R υπολογίζοντας έτσι συσχετίσεις που δεν υποστηρίζονται από το εγγενές περιβάλλον του KNIME. Η δυνατότητα αυτή καθιστά το KNIME ένα υβριδικό εργαλείο, που συνδυάζει την ευχρηστία του οπτικού προγραμματισμού με την ισχύ και την ευελιξία της γλώσσας προγραμματισμού R. Στην εικόνα 62 αποτυπώνονται οι διαθέσιμοι R

κόμβοι που προσφέρει η πλατφόρμα KNIME, μέσω της εγκατεστημένης επέκτασης «KNIME Interactive R Statistics Integration» Οι κόμβοι αυτοί καλύπτουν πλήθος λειτουργιών, όπως:

- Την αμφίδρομη μεταφορά δεδομένων μεταξύ KNIME και R (Table to R, R to Table)
- Την εκτέλεση R κώδικα (R Snippet, R Source),
- Την οπτικοποίηση (R View) και
- Την εφαρμογή μοντέλων πρόβλεψης (R Learner, R Predictor).

Η παρουσίαση των κόμβων αυτών ενισχύει την λειτουργικότητα του KNIME επιτρέπει την υποστήριξη προηγμένων αναλύσεων που βασίζονται σε εξειδικευμένα στατιστικά πακέτα της R.

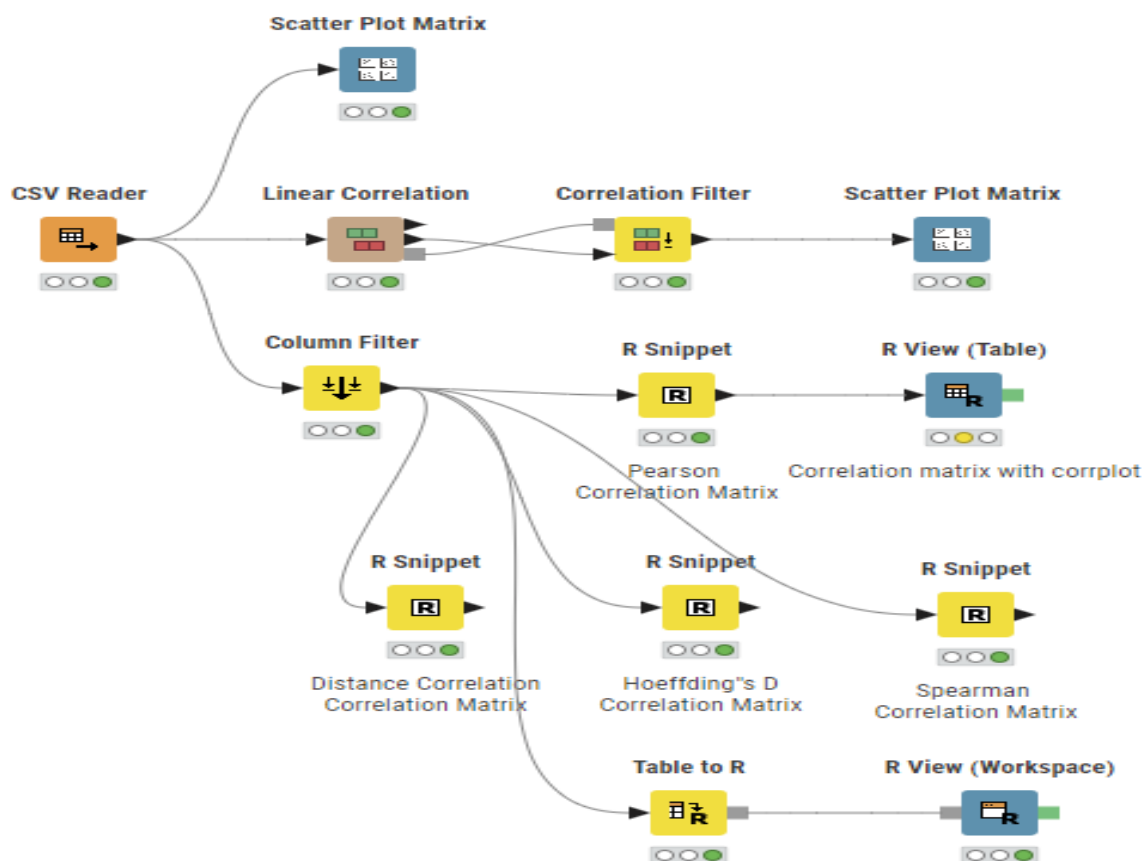


ΕΙΚΟΝΑ 62: Κόμβοι για ενσωμάτωση της R στο KNIME.

Στην εικόνα 63 παρουσιάζεται μια ολοκληρωμένη ροή εργασιών (workflow) στην πλατφόρμα αναλυτικής δεδομένων KNIME, στην οποία εφαρμόζονται ποικίλα μέτρα συσχέτισης μέσω της ενσωμάτωσης της γλώσσας R. Οι μεταβλητές εισάγονται από csv αρχεία με τον κόμβο CSV Reader φιλτράρονται κατάλληλα και στη συνέχεια εφαρμόζονται μέτρα συσχέτισης μέσω διαδοχικών κόμβων R Snippet. Συγκεκριμένα, υπολογίζονται:

- Ο συντελεστής συσχέτισης Pearson, με χρήση συνάρτησης `cor()` της R και απεικόνιση μέσω `corrplot`.
- Ο συντελεστής συσχέτισης Spearman, για την αξιολόγηση μονοτονικών σχέσεων.
- Ο Hoeffding's D, για την ανίχνευση γενικότερων εξαρτήσεων μεταξύ μεταβλητών.
- Η συσχέτιση Απόστασης (Distance Correlation), μέσω της συνάρτησης `dcor()` από το πακέτο `energy`.

Η χρήση του R View επιτρέπει την οπτική απεικόνιση των πινάκων συσχέτισης, ενώ τα δεδομένα εξόδου παρουσιάζονται στον πίνακα στο κάτω μέρος της εικόνας. Το παράδειγμα αυτό αποδεικνύει ότι μέσω της ενσωμάτωσης της R, η πλατφόρμα KNIME μετατρέπεται σε ένα ισχυρό εργαλείο προηγμένης αναλυτικής, ικανό να υποστηρίξει την εφαρμογή ακόμα και των πιο εξειδικευμένων μεθόδων.



ΕΙΚΟΝΑ 63: Πλήρης ροή εργασιών στο KNIME με ανάλυση συσχετίσεων με ενσωμάτωση της R.

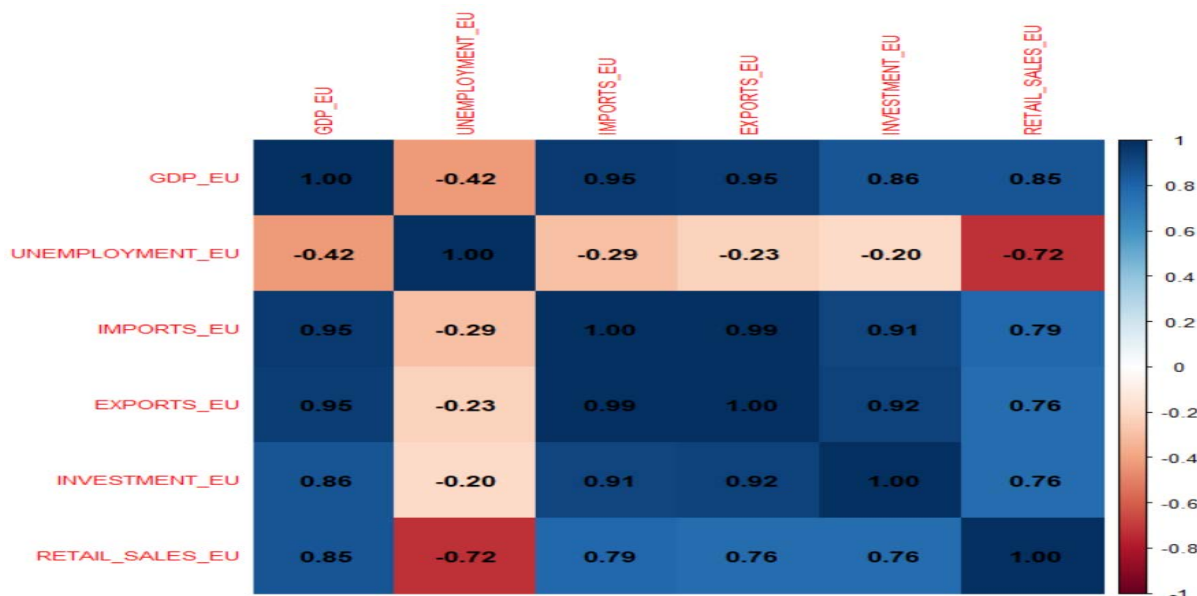
3.3 Οπτικοποίηση των μοντέλων συσχέτισης.

Μετά την ολοκλήρωση της υπολογιστικής διερεύνησης των μέτρων συσχέτισης στις προηγούμενες ενότητες τόσο μέσω της γλώσσας προγραμματισμού R, όσο και μέσω της πλατφόρμας αναλυτικής δεδομένων KNIME η παρούσα ενότητα επικεντρώνεται στην οπτική αναπαράσταση των αποτελεσμάτων. Η γραφική αναπαράσταση αποτελεί σημαντικό στάδιο στην ανάλυση δεδομένων, καθώς ενισχύει τη δυνατότητα κατανόησης και ερμηνείας των συσχετίσεων που προέκυψαν από τα επιλεγμένα μοντέλα. Στο πλαίσιο αυτό, χρησιμοποιήθηκαν διαγράμματα συσχέτισης (corrplots) για τους συντελεστές συσχέτισης που

υποστηρίζουν γραφική απεικόνιση, προσφέροντας μια πιο άμεση και κατανοητή αποτύπωση των συσχετίσεων μεταξύ των μεταβλητών. Επιπλέον, εφαρμόστηκε ιεραρχική ομαδοποίηση μεταβλητών με στόχο την ανάδειξη συστάδων (clusters) μεταβλητών που παρουσιάζουν υψηλό βαθμό συσχέτισης μεταξύ τους. Τα δένδροδιαγράμματα που προκύπτουν από αυτή τη διαδικασία υποστηρίζουν την αποκωδικοποίηση της δομής της συσχέτισης του dataset.

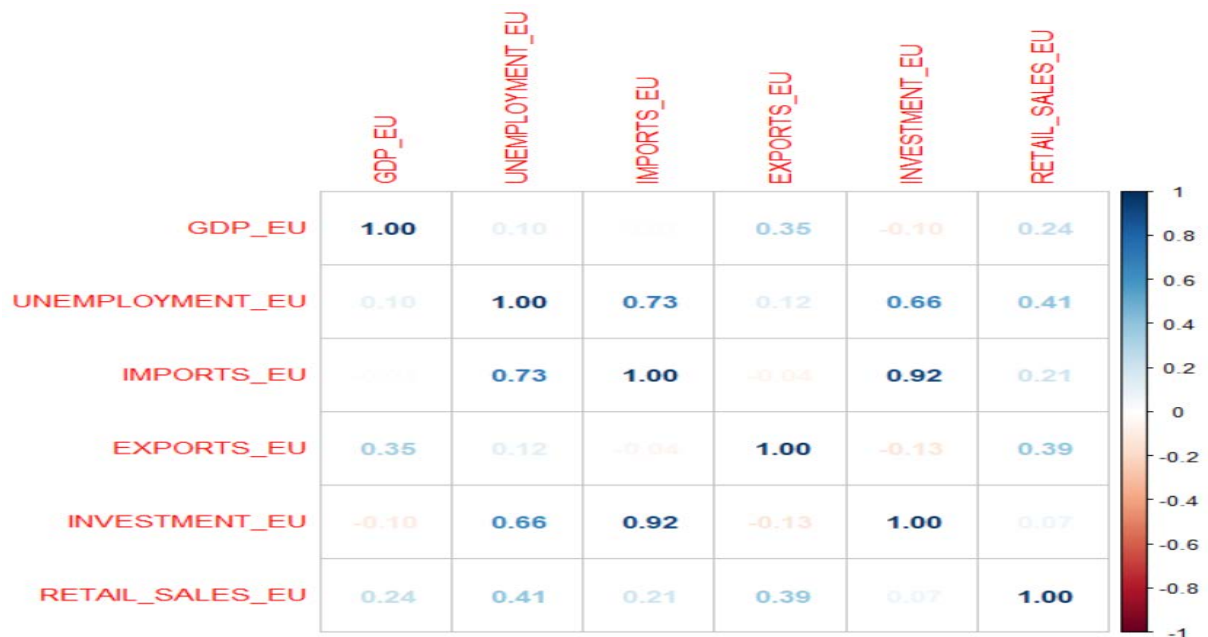
Συντελεστής Συσχέτισης Pearson:

Η εικόνα 64 απεικονίζει τον θερμικό χάρτη του συντελεστή συσχέτισης Pearson των βασικών μακροοικονομικών μεταβλητών της Ευρωπαϊκής Ένωσης μέσω της βιβλιοθήκης `corrplot` χρησιμοποιώντας την επιλογή `method=color`, η οποία αποδίδει κάθε τιμή συσχέτισης με κατάλληλη χρωματική απόχρωση.



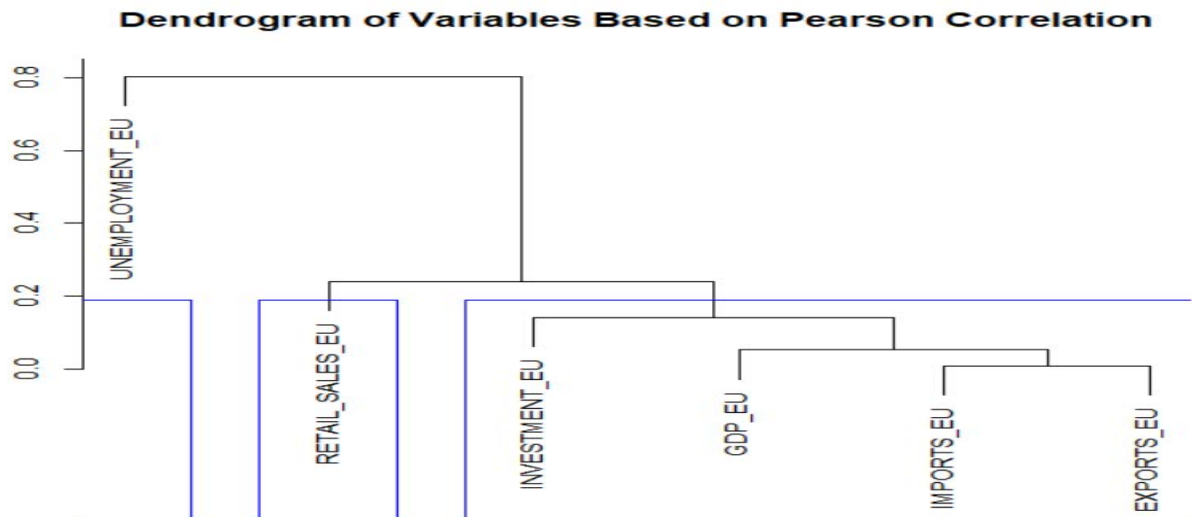
ΕΙΚΟΝΑ 64: Γραφική απεικόνιση του Pearson με `Corrplot method=color`. Σκούρο μπλε υποδηλώνει ισχυρή θετική συσχέτιση (τιμές κοντά στο +1), ανοιχτό μπλε υποδηλώνει ασθενέστερη θετική συσχέτιση, ανοιχτό κόκκινο υποδηλώνει ασθενή αρνητική συσχέτιση και σκούρο κόκκινο υποδηλώνει ισχυρή αρνητική συσχέτιση (τιμές κοντά στο -1).

Στην εικόνα 65 παρουσιάζεται ένας θερμικός χάρτης συσχέτισης μέσω της βιβλιοθήκης `corrplot` χρησιμοποιώντας την επιλογή `method=number`, στον οποίο αποτυπώνονται τόσο οι αριθμητικές τιμές των συντελεστών συσχέτισης όσο και η χρωματική κωδικοποίηση της έντασης πάνω στις τιμές τις συσχέτισης.



EIKONA 65: Γραφική απεικόνιση Pearson Correlation Coefficient με Corrplot method=number.

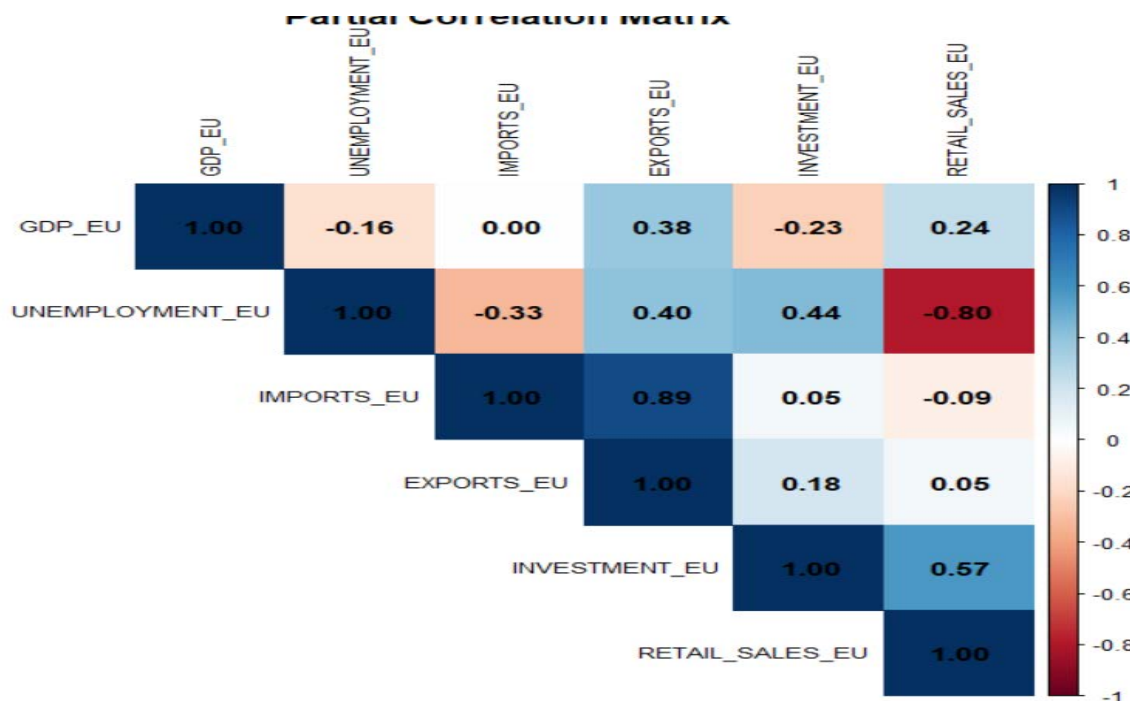
Στην εικόνα 66 παρουσιάζεται το δένδροδιάγραμμα των μεταβλητών βάσει της συσχέτισης Pearson, το οποίο προέκυψε με τη μέθοδο ιεραρχικής ομαδοποίησης (hierarchical clustering). Ο κάθε κλάδος του διαγράμματος απεικονίζει την συσχέτιση εγγύτητα μεταξύ των μεταβλητών με λίγα λόγια όσο πιο κοντά συνδέονται δύο μεταβλητές, τόσο ισχυρότερη είναι η θετική τους συσχέτιση. Στο συγκεκριμένο παράδειγμα, παρατηρείται ότι οι μεταβλητές EXPORTS_EU και IMPORTS_EU σχηματίζουν ένα πρώτο υποσύνολο με πολύ υψηλή συσχέτιση, το οποίο ενώνεται σταδιακά με τις GDP_EU και INVESTMENT_EU. Η RETAIL_SALES_EU βρίσκεται απομονωμένη, ενώ η UNEMPLOYMENT_EU διαχωρίζεται νωρίς, γεγονός που υποδηλώνει αρνητική ή χαμηλή συσχέτιση με τις υπόλοιπες μεταβλητές. Η χρήση της εντολής addrect=3 στην R επιτρέπει την εμφάνιση τριών ομάδων μεταβλητών (clusters), όπως φαίνεται με τα μπλε πλαίσια στο γράφημα



ΕΙΚΟΝΑ 66: Δενδροδιάγραμμα του Pearson Correlation Coefficient.

Partial Correlation Coefficient:

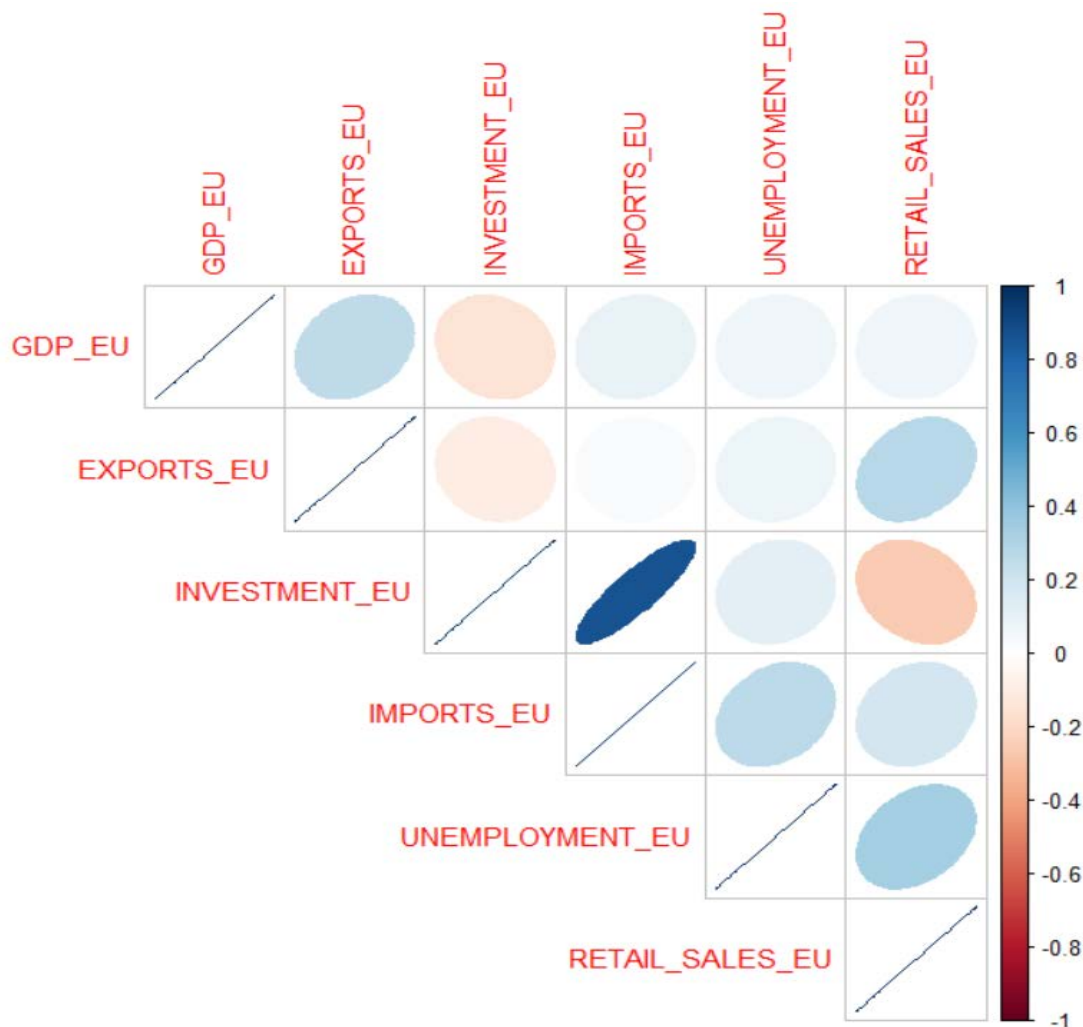
Η εικόνα 67 απεικονίζει τον πίνακα της μερικής συσχέτισης (Partial Correlation) χρησιμοποιώντας τη βιβλιοθήκη corrplot. Η γραφική αναπαράσταση απεικονίζει την σχέσεων μεταξύ των μεταβλητών σε τρίγωνο άνω πίνακα συσχέτισης. Το χρώμα και η διαβάθμιση του χρώματος των κελιών αντιπροσωπεύουν την κατεύθυνση και την ένταση της συσχέτισης αντίστοιχα.



ΕΙΚΟΝΑ 67: Γραφική απεικόνιση του Partial Correlation Coefficient με Corrplot

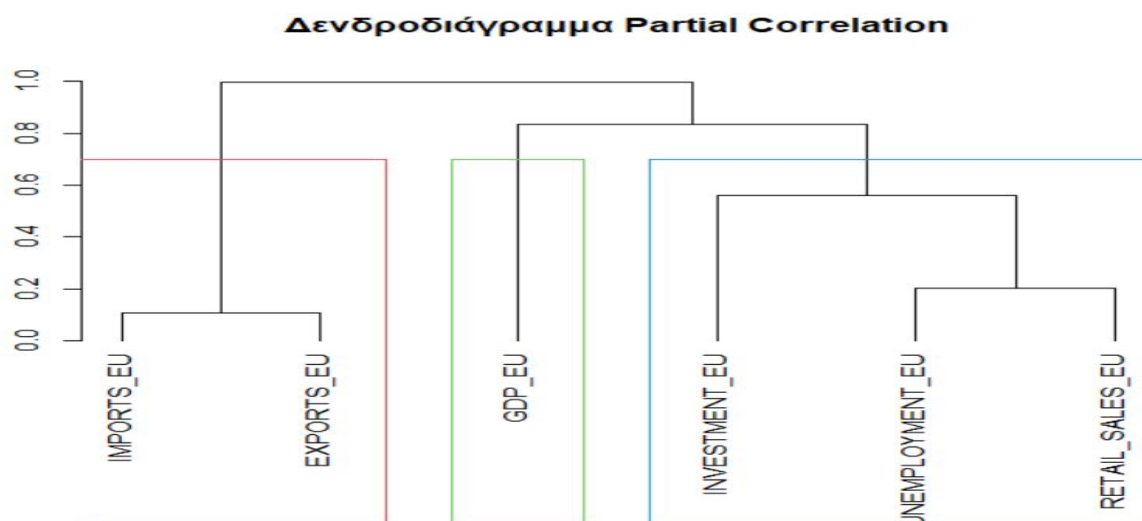
method=color. Σκούρο μπλε υποδηλώνει ισχυρή θετική συσχέτιση (τιμές κοντά στο +1), ανοιχτό μπλε υποδηλώνει ασθενέστερη θετική συσχέτιση, ανοιχτό κόκκινο υποδηλώνει ασθενή αρνητική συσχέτιση και σκούρο κόκκινο υποδηλώνει ισχυρή αρνητική συσχέτιση (τιμές κοντά στο -1).

Στην εικόνα 68 παρουσιάζεται η γραφική απεικόνιση του πίνακα Partial Correlation Coefficient. Το γράφημα δημιουργήθηκε με την μέθοδο “ellipse”

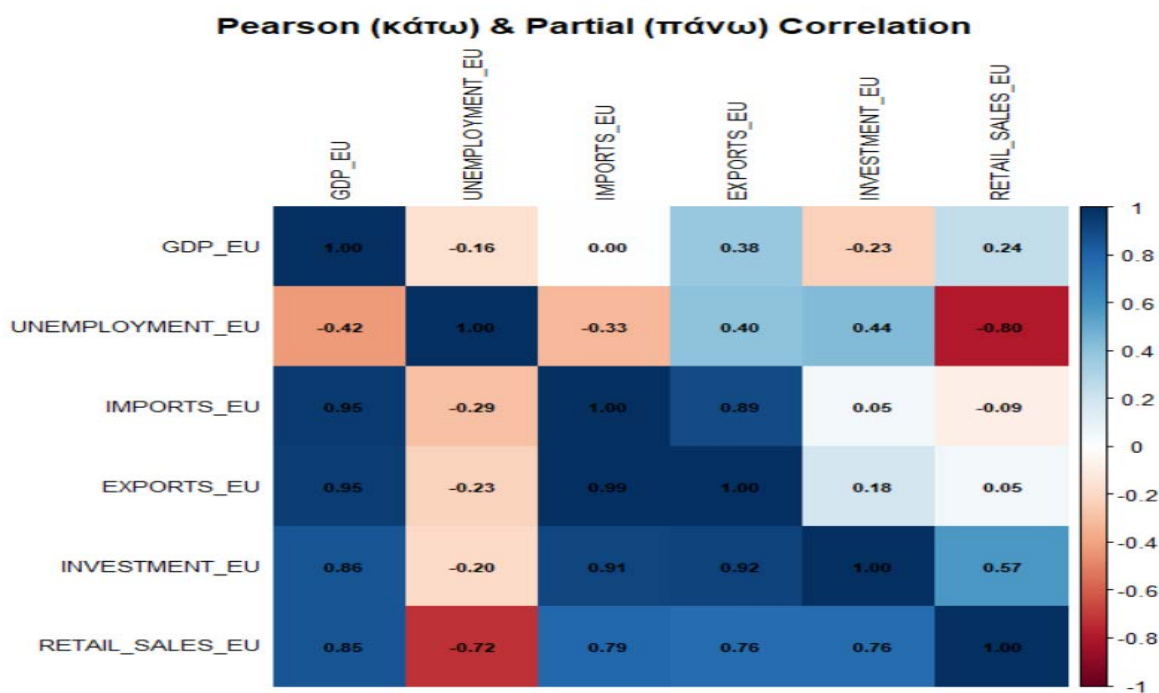


ΕΙΚΟΝΑ 68: Γραφική απεικόνιση του Partial Correlation Coefficient με Corrplot method=ellipse. Η κατεύθυνση της έλλειψης δείχνει το πρόσημο της συσχέτισης (δεξιά=θετική, αριστερά=αρνητική). Η κλίση και η συμπίεση της έλλειψης υποδηλώνει την κατεύθυνση και την ένταση της συσχέτισης αντίστοιχα. Το χρώμα είναι όπως αναφέραμε και στον συντελεστή Pearson σκούρο μπλε=ισχυρή θετική συσχέτιση και σκούρο κόκκινο ισχυρή αρνητική συσχέτιση.

Στην εικόνα 69 βλέπουμε ότι στο δένδροδιάγραμμα οι μεταβλητές IMPORTS_EU και EXPORTS_EU συνδέονται μεταξύ τους και αποτελούν μια ομάδα, ενώ η μεταβλητή GDP εμφανίζεται μόνη, οι μεταβλητές INVESTMENT_EU, UNEMPLOYMENT_EU και RETAIL_SALES_EU συγκροτούν την Τρίτη ομάδα με πιθανές εσωτερικές εξαρτήσεις.



ΕΙΚΟΝΑ 69: Δενδροδιάγραμμα Partial Correlation Coefficient.



ΕΙΚΟΝΑ 69: Pearson Correlation (Lower) & Partial Correlation (Upper)

Συντελεστής Συσχέτισης Απόστασης (Distance Correlation)

Στην εικόνα 70 παρουσιάζουμε την γραφική απεικόνιση του συντελεστή συσχέτισης απόστασης. Σύμφωνα με την θεωρία ο Distance Correlation Coefficient είναι ικανός να ανιχνεύει τόσο γραμμικές όσο και μη γραμμικές σχέσεις μεταξύ των μεταβλητών και να λαμβάνει τιμή από το 0 έως το 1. Επομένως σε αντίθεση με άλλους συντελεστές συσχέτισης όπως του Pearson, δεν παρατηρούνται αρνητικές τιμές, και κατά συνέπεια δεν εμφανίζεται χρωματική διαφοροποίηση με κόκκινες αποχρώσεις, οι οποίες σε προηγούμενα γραφήματα αντιστοιχούσαν σε αρνητικές συσχετίσεις.



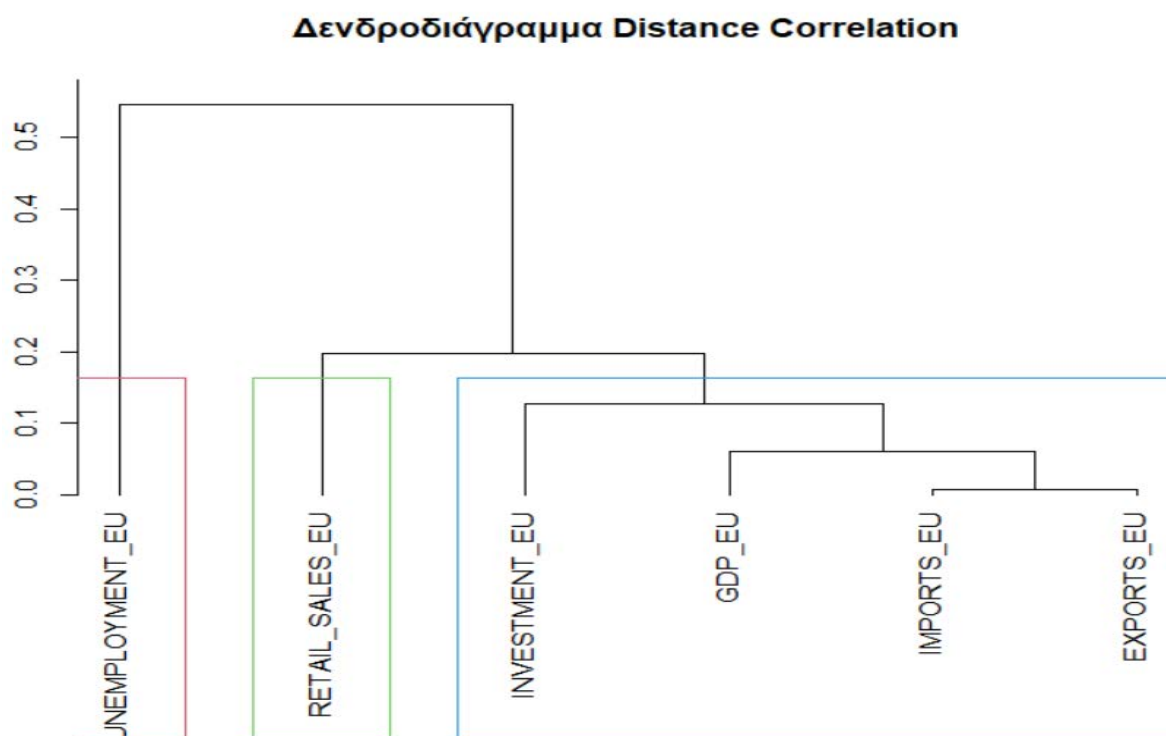
ΕΙΚΟΝΑ 70: Γραφική παράσταση του Distance Correlation Coefficient με αριθμητικές τιμές και χρωματική απεικόνιση έντασης. Οι τιμές εκφράζουν τη δύναμη εξάρτησης μεταξύ ζευγών μεταβλητών, ανεξαρτήτως της μορφής της σχέσης (γραμμική ή μη).

Στην εικόνα 71 παρουσιάζεται ο πίνακας Distance Correlation με γραφική αναπαράσταση μέσω κυκλικών σχημάτων από την κύρια διαγώνιο και άνω και τιμές συσχέτισης από την κύρια διαγώνιο και κάτω. Οι κύκλοι ποικίλλουν ως προς το μέγεθος και την ένταση του μπλε χρώματος, αποτυπώνοντας την ισχύ της εξάρτησης, ενώ οι αριθμητικές τιμές επιτρέπουν την ακριβή ποσοτική εκτίμηση της συσχέτισης ανάμεσα στις μεταβλητές.



ΕΙΚΟΝΑ 71: Γραφική παράσταση του Distance Correlation Coefficient με διπλή αναπαράσταση. Άνω μέρος μέσω κυκλικών σχημάτων, όπου το μέγεθος και η ένταση του μπλε χρώματος αντιστοιχούν στη δύναμη της συσχέτισης. Στο κάτω μέρος παρουσιάζονται ακριβείς αριθμητικές.

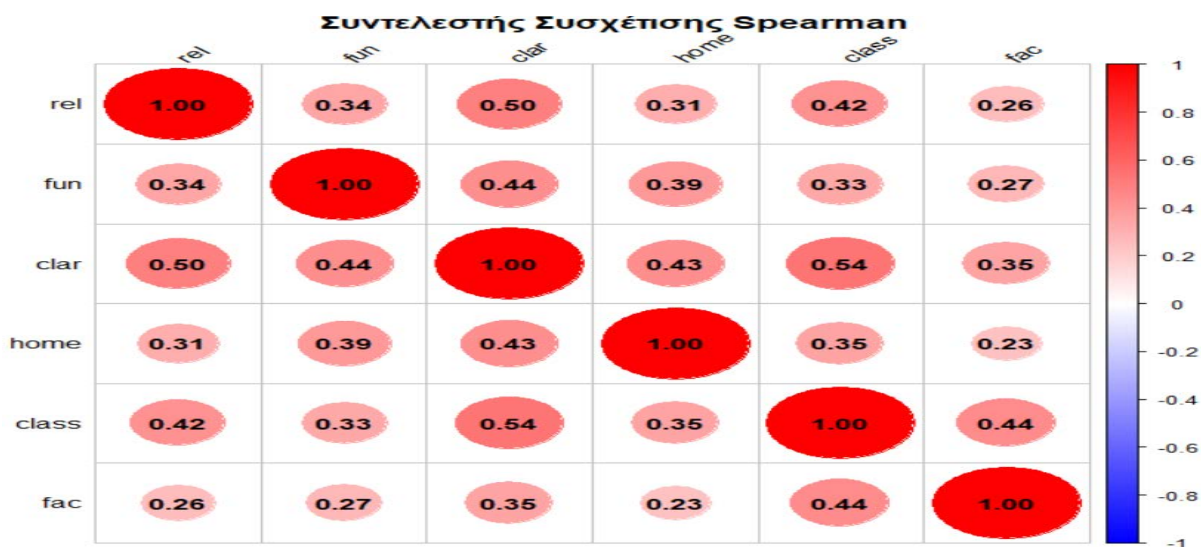
Στην εικόνα 72 παρουσιάζεται το δένδροδιάγραμμα βάσει του συντελεστή Distance Correlation, το οποίο προέκυψε μέσω ιεραρχικής ομαδοποίησης (hierarchical clustering). Το γράφημα αυτό επιτρέπει την ομαδοποίηση μεταβλητών με αποτέλεσμα να ανιχνεύσει βαθύτερες συσχετίσεις. Από την ανάλυση προκύπτουν τρεις κύριες ομάδες. Η πρώτη ομάδα περιλαμβάνει την μεταβλητή UNEMPLOYMENT_EU, η οποία διαφοροποιείται έντονα από τις υπόλοιπες μεταβλητές. Η δεύτερη ομάδα περιλαμβάνει την RETAIL_SALES_EU, η οποία παραμένει και αυτή ξεχωριστή. Η τρίτη ομάδα περιλαμβάνει τις μεταβλητές INVESTMENT_EU, GDP_EU, IMPORTS_EU και EXPORTS_EU, οι οποίες παρουσιάζουν μεταξύ τους υψηλές μη γραμμικές συσχετίσεις, στοιχείο που επιβεβαιώνεται και από τον αντίστοιχο πίνακα Distance Correlation.



ΕΙΚΟΝΑ 72: Δενδροδιάγραμμα Distance Correlation Coefficient.

Spearman's Rho Correlation Coefficient:

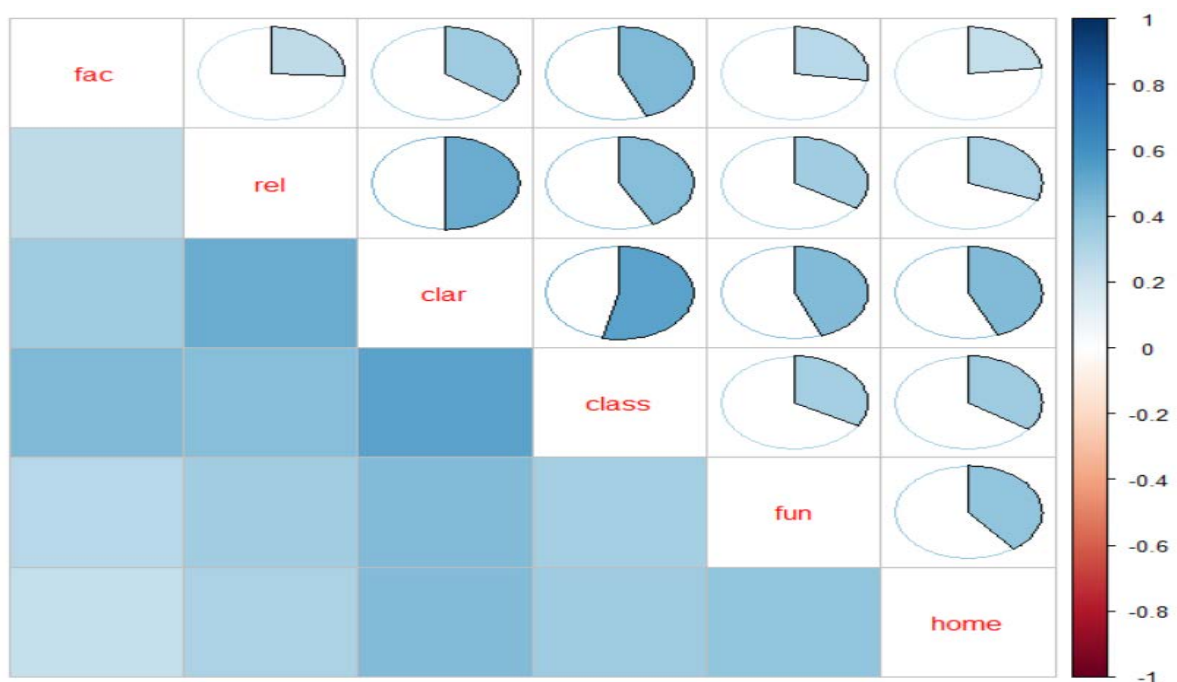
Στην εικόνα 73 απεικονίζεται ο συντελεστής συσχέτισης Spearman's Rho μεταξύ έξι ποιοτικών (ordinal) μεταβλητών από το dataset "learning". Στο διάγραμμα μπορούμε να παρατηρήσουμε ότι το μέγεθος των κύκλων αποτυπώνει το μέτρο της συσχέτισης και η ένταση του κόκκινου χρώματος δηλώνει ισχυρή θετική συσχέτιση.



ΕΙΚΟΝΑ 73: Γραφική απεικόνιση Spearman Correlation Coefficient μέσω κυκλικών

σημάτων. Το μέγεθος και η ένταση του χρώματος κάθε κύκλου εκφράζουν τη δύναμη της συσχέτισης μεταξύ των μεταβλητών.

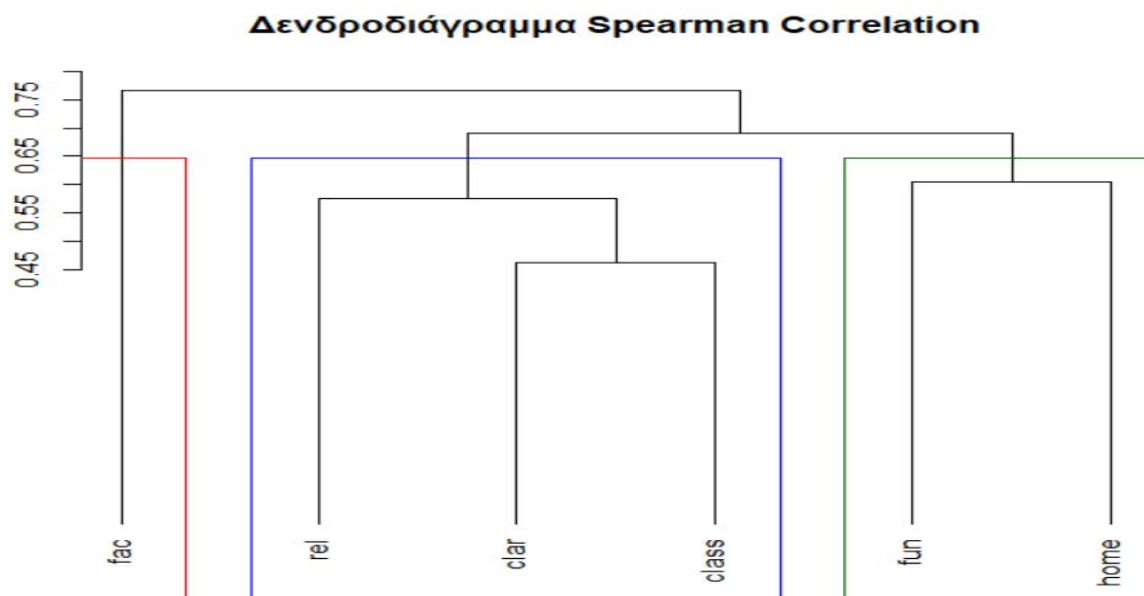
Στην εικόνα 74 απεικονίζεται ο συντελεστής συσχέτισης Spearman με συνδυασμένη αναπαράσταση. Από την κύρια διαγώνιο και άνω, κάθε κελί αναπαρίσταται σαν πίτα (pie-type), ενώ από την κάτω διαγώνιο, οι συσχετίσεις αποδίδονται με χρωματική ένταση, όπου οι πιο σκούρες αποχρώσεις του μπλε υποδηλώνουν ισχυρότερες θετικές συσχετίσεις. Η συνδυασμένη απεικόνιση προσφέρει διπλό πλεονέκτημα. Την αίσθηση της έντασης της σχέσης (μέσω χρώματος) και τη οπτικά άμεση σύγκριση μεταξύ όλων των ζευγών μεταβλητών.



EIKONA 74: Γραφική απεικόνιση Spearman's Rho Correlation Coefficient με συνδυασμένη απεικόνιση. Πάνω από τη διαγώνιο κυκλικά διαγράμματα (pie-type) και κάτω από τη διαγώνιο η ένταση της συσχέτισης.

Στην εικόνα 75 παρουσιάζεται το δένδροδιάγραμμα βάσει του συντελεστή συσχέτισης Spearman, το οποίο προέκυψε μέσω ιεραρχικής ομαδοποίησης (hierarchical clustering) όπου ομαδοποιεί μεταβλητές με παρόμοια συμπεριφορά, χωρίς να απαιτείται γραμμική σχέση. Από την ανάλυση διακρίνονται τρεις ομάδες μεταβλητών. Η πρώτη ομάδα (rel, clar, class) αποτελεί ένα σύνολο μεταβλητών που συνδέονται με την ποιότητα του μαθήματος και την κατανόηση

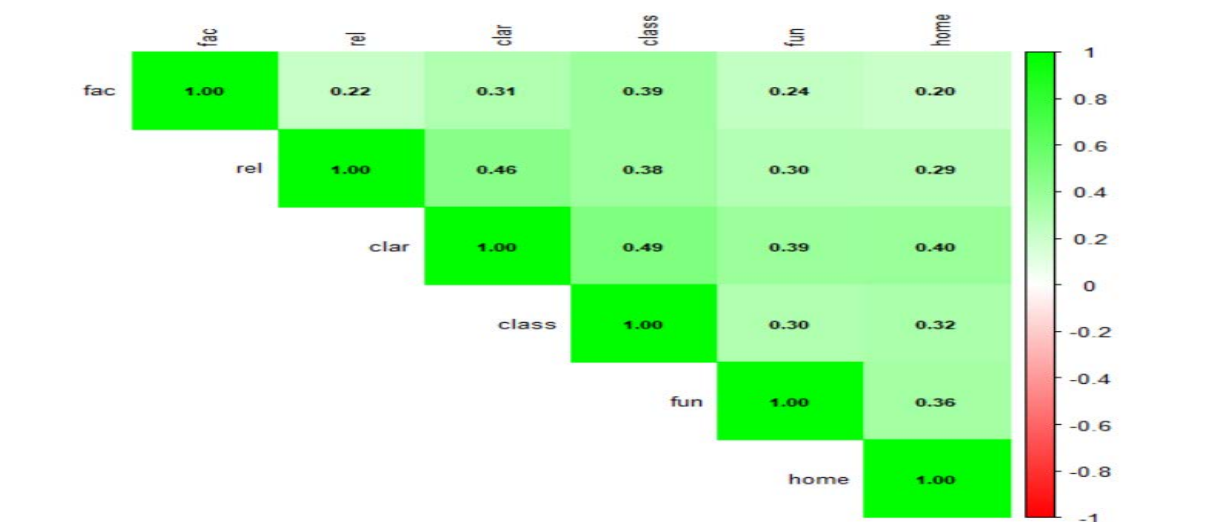
του. Η δεύτερη ομάδα (fun, home) περιλαμβάνει παραμέτρους σχετικές με την γενική ευχαρίστηση από τη μάθηση. Η Τρίτη ομάδα fac εμφανίζεται απομονωμένη από τις υπόλοιπες.



ΕΙΚΟΝΑ 75: ΔΕΝΔΡΟΔΙΑΓΡΑΜΜΑ Spearman Correlation Coefficient.

Συντελεστής Συσχέτισης Kendall's Tau:

Στην εικόνα 76 παρουσιάζεται η γραφική απεικόνιση του συντελεστή συσχέτισης Kendall's Tau, ο οποίος είναι ιδιαίτερα χρήσιμος σε περιπτώσεις που τα δεδομένα περιέχουν ισοβαθμίες ή είναι μικρά σε όγκο.



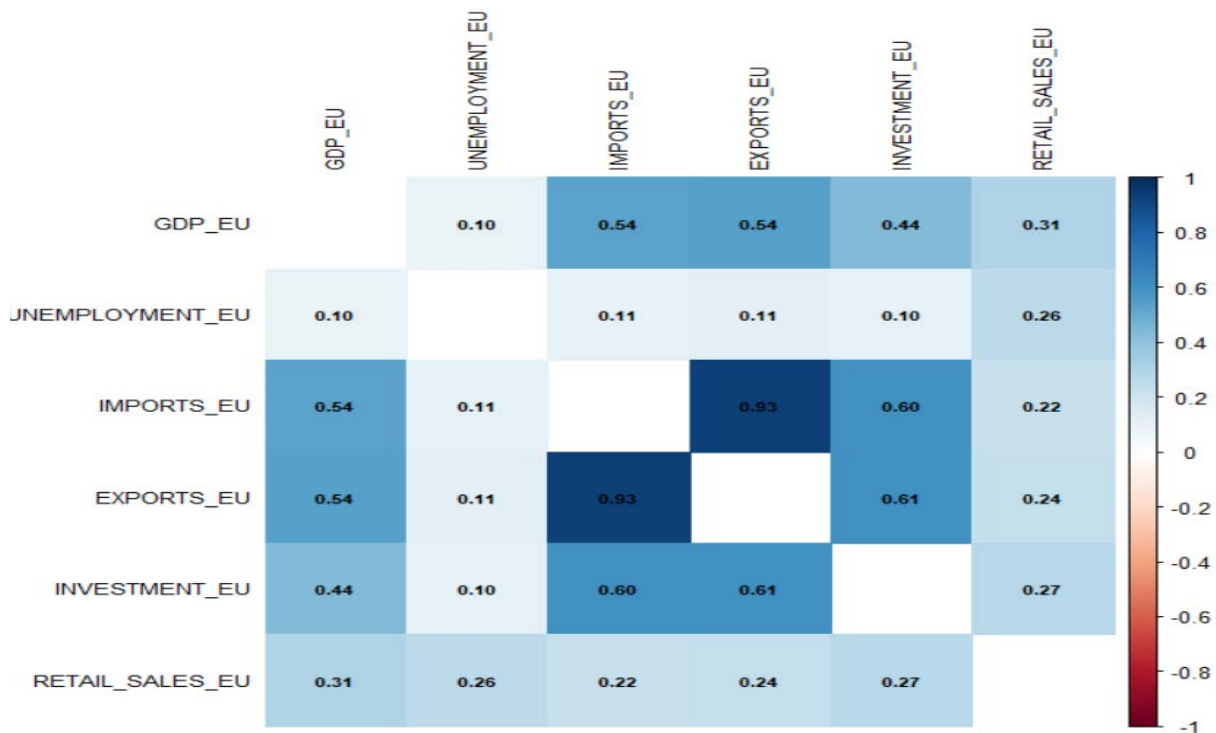
ΕΙΚΟΝΑ 76: Γραφική παράσταση Kendall's Tau correlation Coefficient. Το χρώμα αντιστοιχεί στην ένταση και το πρόσημο της σχέσης μεταξύ των μεταβλητών.

Στην εικόνα 77 παρουσιάζεται το δένδροδιάγραμμα του συντελεστή συσχέτισης Kendall's Tau. Όπως φαίνεται, η δομή του διαγράμματος είναι ταυτόσημη με αυτή που προέκυψε από τη συσχέτιση Spearman, επιβεβαιώνοντας ότι και οι δύο συντελεστές καταγράφουν παρόμοιες σχέσεις μεταξύ των μεταβλητών.



ΕΙΚΟΝΑ 77: ΔΕΝΔΡΟΔΙΑΓΡΑΜΜΑ Kendall's Tau Correlation Coefficient.

Συντελεστής Συσχέτισης *Hoeffding's D*:



ΕΙΚΟΝΑ 78: Γραφική απεικόνιση Hoeffding's D Correlation Coefficient με Corplot.

Στην εικόνα 79 παρουσιάζεται η γραφική απεικόνιση του συντελεστή συσχέτισης Hoeffding's D, με κάθε κελί να περιέχει οριζόντιες μπάρες (bar segments) που αντιπροσωπεύει την σχετική ένταση της εξάρτησης μεταξύ δύο μεταβλητών. Όσο μεγαλύτερο το μήκος και πιο σκουρόχρωμη η γραμμή, τόσο ισχυρότερη είναι η εξάρτηση που καταγράφεται από τον συντελεστή Hoeffding's D. Η απεικόνιση δεν παρέχει αριθμητικές τιμές, αλλά αναδεικνύει ποιοτικά τις πιο ισχυρές σχέσεις. Για παράδειγμα παρατηρείται ότι τα ζεύγη IMPORTS_EU-EXPORTS_EU, GDP_EU-INVESTMENT_EU και EXPORTS_EU-INVESTMENT_EU διαθέτουν σαφώς μεγαλύτερες μπάρες, ένδειξη ισχυρής εξάρτησης.



ΕΙΚΟΝΑ 79: Ποιοτική απεικόνιση Hoeffding's D Correlation Coefficient με οριζόντιες μπάρες. Το μήκος και το χρώμα της γραμμής αντανακλούν τη δύναμη της εξάρτησης μεταξύ μεταβλητών.

Στην εικόνα 80 παρουσιάζεται το δένδροδιάγραμμα του συντελεστή συσχέτισης Hoeffding's D. Από την γραφική απεικόνιση εντοπίζονται τρεις κύριες ομάδες:



ΕΙΚΟΝΑ 80: ΔΕΝΔΡΟΔΙΑΓΡΑΜΜΑ Hoeffding's D Correlation Coefficient.

Συντελεστής Συσχέτισης Phi:

Στην εικόνα 81 παρουσιάζεται η γραφική απεικόνιση του συντελεστή συσχέτισης Φ για τέσσερις δυαδικές (binary) μεταβλητές (default, housing, loan, y). Ο πίνακας είναι συμμετρικός με την κύρια διαγώνιο να περιλαμβάνει τιμές 1, καθώς κάθε μεταβλητή συσχετίζεται πλήρως με τον εαυτό της. Οι τιμές εκτός της διαγώνιο είναι γενικά μικρές, γεγονός που δείχνει ασθενείς συσχετίσεις μεταξύ των μεταβλητών.



ΕΙΚΟΝΑ 81: Γραφική παράσταση Phi Coefficient για δυαδικές μεταβλητές. Η χρωματική ένταση εκφράζει τη δύναμη και το πρόσημο της συσχέτισης μεταξύ των μεταβλητών.

ΚΕΦΑΛΑΙΟ 4: ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία είχε ως στόχο τη διερεύνηση και σύγκριση των μέτρων συσχέτισεων, εφαρμόζοντας τεχνικές και εργαλεία αναλυτικής δεδομένων. Μέσα από την ανάλυση, αναδείχθηκε η σημασία της κατάλληλης επιλογής συντελεστή συσχέτισης, ανάλογα με τη φύση των δεδομένων και τη μορφή της σχέσης μεταξύ των μεταβλητών. Είναι σημαντικό να τονιστεί ότι τα ευρέως χρησιμοποιούμενα στατιστικά λογισμικά, όπως το SPSS και το MINITAB, παρέχουν περιορισμένες δυνατότητες ως προς την εφαρμογή της ανάλυσης συσχέτισης. Οι διαθέσιμοι συντελεστές περιορίζονται στους πλέον κλασικούς: Pearson, Spearman, Kendall's Tau, Cramer's V, Phi (Φ), καθώς και σε μετρικές που αφορούν την πολυσυγγραμμικότητα, όπως οι VIF, Tolerance, Eigenvalues και ο Condition Index. Ωστόσο, η ανάγκη για ανίχνευση πιο σύνθετων ή μη γραμμικών σχέσεων οδηγεί στην αναγκαιότητα χρήσης προηγμένων συντελεστών που δεν υποστηρίζονται από τα παραπάνω λογισμικά. Με τη χρήση εργαλείων της αναλυτικής δεδομένων, η ανάλυση συσχέτισης εμπλουτίζεται σημαντικά με τη δυνατότητα εύληπτων απεικονίσεων και γραφημάτων.

Η συνεισφορά της εργασίας είναι ότι αναδεικνύει τρόπους υπολογισμού για συντελεστές όπως ο Distance Correlation, ο Maximal Information Coefficient (MIC) και ο Hoeffding's D, οι οποίοι επιτρέπουν την ανίχνευση πολύπλοκων ή μη γραμμικών εξαρτήσεων που δεν μπορούν να αποτυπωθούν μέσω παραδοσιακών συντελεστών Pearson, Spearman, Kendall. Οι πιο εξειδικευμένοι συντελεστές που προσεγγίστηκαν υπολογιστικά στην ανάλυση της εργασίας αποδείχθηκαν πολύ σημαντικοί καθώς μπορούν να εντοπίζουν μη γραμμικές σχέσεις, ιδιαίτερα όταν τα δεδομένα παρουσιάζουν αποκλίσεις από την γραμμικότητα. Αξιοσημείωτο είναι επίσης το γεγονός ότι συντελεστές όπως ο Biweight Mid Correlation και ο Percentage Bend Correlation ακολουθούν την ίδια λογική με τον Pearson ως προς την ανάλυση γραμμικών σχέσεων, αλλά είναι πιο ανθεκτικοί σε ακραίες τιμές, καθώς προσαρμόζουν την επίδραση των ακραίων τιμών.

Επιπλέον, η εφαρμογή και γραφική απεικόνιση του συντελεστή μερικής συσχέτισης Partial Correlation Coefficient παρέχει την δυνατότητα να απομονώσει την καθαρή σχέση μεταξύ δυο μεταβλητών, χωρίς να λαμβάνει την επίδραση τρίτων μεταβλητών. Η προσέγγιση αυτή ήταν ιδιαίτερα χρήσιμη στη μελέτη πολυπαραγοντικών συστημάτων, όπου οι συσχετίσεις μεταξύ μεταβλητών είναι συνηθισμένες.

Ο συντελεστής συσχέτισης Spearman's Rho και ο Kendall's Tau είναι κατάλληλοι για την μελέτη ποιοτικών μεταβλητών, διατακτικών (ordinal) μεταβλητών, όπως αυτές που προκύπτουν από ερωτηματολόγια με κλίμακες Likert. Ιδιαίτερη έμφαση δόθηκε στις παραλλαγές του Kendall's Tau, όπου η έκδοση Tau-b θεωρείται προτιμότερη σε περιπτώσεις ισοβαθμιών, καθώς τις προσαρμόζει στην ανάλυση, σε αντίθεση με την παραλλαγή Tau-a, η οποία τις αγνοεί οδηγώντας σε υπερεκτιμήσεις και υψηλές τιμές συσχέτισης.

Ένας άλλος σημαντικός συντελεστής συσχέτισης είναι ο Cramer's V όπου είναι ένα μη παραμετρικό μέτρο συσχέτισης ειδικά σχεδιασμένο για την ανάλυση μόνο ποιοτικών μεταβλητών όπου αποτελεί μια τροποποιημένη εκδοχή του Chi-Square. Είναι σημαντικό να αναφερθούμε και στον συντελεστή συσχέτισης Φ όπου είναι ισοδύναμος με τον συντελεστή συσχέτισης Pearson correlation σε 2*2 πίνακα συνάφειας που χρησιμοποιείται για την εκτίμηση της σχέσης μεταξύ δύο δυαδικών (binary) μεταβλητών.

Στο πλαίσιο της πολυμεταβλητής ανάλυσης, ιδιαίτερη έμφαση δόθηκε στο φαινόμενο της πολυσυγγραμμικότητας που αποτελεί το πιο συχνό πρόβλημα στη πολυμεταβλητή ανάλυση καθώς δημιουργεί αστάθεια στις εκτιμήσεις των συντελεστών παλινδρόμησης, μειώνει την ακρίβεια των εκτιμήσεων και δυσκολεύει την ερμηνεία των αποτελεσμάτων. Για την διάγνωση της παρουσίας του φαινομένου αυτού, χρησιμοποιήθηκαν διαγνωστικά μετρικές (δείκτες): Variance Inflation Factor (VIF), Tolerance Coefficient, Eigenvalues και Condition Index, τα οποία εντοπίζουν ισχυρές γραμμικές σχέσεις μεταξύ ανεξάρτητων μεταβλητών.

Τέλος, η Canonical Correlation Analysis (CCA) αποτέλεσε μια σημαντική μεθοδολογική προσέγγιση για την μελέτη της σχέσης δύο ομάδων μεταβλητών, Είναι ιδιαίτερα χρήσιμο όταν θέλουμε να δούμε πώς οικονομικές μεταβλητές επηρεάζονται συνολικά από χρηματοοικονομικές μεταβλητές.

Η ολοκληρωμένη ανάλυση πραγματοποιήθηκε μέσω της γλώσσας προγραμματισμού R, η οποία προσφέρει πρόσβαση σε ένα ευρύ φάσμα πακέτων για εξειδικευμένες μεθόδους, αλλά και της πλατφόρμας αναλυτικής δεδομένων KNIME, η οποία πέρα από την φιλικότητα του οπτικού προγραμματισμού (Berthold et al., n.d.), επιτρέπει και την ενσωμάτωση της R για την υλοποίηση πιο προχωρημένων αναλύσεων. Ο συνδυασμός των δύο εργαλείων απέδειξε πως είναι δυνατή μια ευέλικτη, διαφανής και επιστημονικά τεκμηριωμένη προσέγγιση στην μελέτη της συσχέτισης. Συνοψίζοντας, η εργασία μας δείχνει ότι η επιλογή του κατάλληλου μέτρου συσχέτισης εξαρτάται από:

- Το είδος των μεταβλητών (ποσοτικές, ποιοτικές, κατάταξης, κλπ.)

- Τη μορφή της σχέσης (γραμμική, μονοτονική, μη γραμμική).
- Και τους στόχους της ανάλυσης.

Ο συνδυασμός διαφορετικών συντελεστών σε υπολογιστικά ευέλικτα περιβάλλοντα αποτελεί πλέον μια απαραίτητη στρατηγική για την αξιόπιστη μελέτη της συσχέτισης στην εποχή της αναλυτικής δεδομένων.

4.1 Περιορισμοί και προτάσεις για περαιτέρω έρευνα

Η παρούσα εργασία, παρότι προσφέρει μια ξεκάθαρη εξερεύνηση στη μελέτη των μέτρων συσχέτισης, υπόκειται σε ορισμένους περιορισμούς που αξίζει να αναφερθούν:

1. Η ανάλυση επικεντρώθηκε σε μικρά σύνολα δεδομένων που προέρχονται κυρίως από τον οικονομικό και επιχειρηματικό τομέα. Αυτό ήταν σκοπός για να διασφαλιστεί η εφαρμοσιμότητα των υπολογιστικών μεθόδων στην οικονομική επιστήμη. Η χρήση μεγάλων δεδομένων (big data) δεν καλύπτεται στην παρούσα εργασία.
2. Κατά την εφαρμογή μεθόδων που περιλαμβάνουν μεγάλο αριθμό μεταβλητών, παρατηρήθηκε ότι ορισμένοι συντελεστές όπως ο συντελεστής συσχέτισης Φκ και ο Global Correlation Coefficient στα υπολογιστικά εργαλεία που χρησιμοποιήσαμε δεν διαθέτουν βιβλιοθήκες που να υποστηρίζουν πολυμεταβλητές συσχετίσεις, περιορίζοντας τη χρηστικότητα τους.

Κατόπιν της παρουσίασης των βασικών περιορισμών που εντοπίστηκαν κατά την εκπόνηση της παρούσας εργασίας, καθίσταται σαφές ότι υπάρχουν περιθώρια περαιτέρω διερεύνησης και εμβάθυνσης. Στο πλαίσιο αυτό, παρατίθενται ορισμένες προτάσεις που μπορούν να αποτελέσουν τη βάση για περαιτέρω μελέτη και εμπλουτισμό της υπάρχουσας γνώσης.

1. Μια συστηματική ανάλυση για το πώς επηρεάζονται τα αποτελέσματα των μεθόδων από ακραίες τιμές, η θόρυβο, θα πρόσφερε χρήσιμα συμπεράσματα για την πρακτική εφαρμογή τους.

2. Μια ιδιαίτερα ενδιαφέρουσα κατεύθυνση για μελλοντική έρευνα αποτελεί η εστιασμένη μελέτη μέτρων συσχέτισης που έχουν αναπτυχθεί πρόσφατα και παρουσιάζουν αυξημένη ανθεκτικότητα στην παρουσία ακραίων τιμών. Τέτοια μέτρα μπορούν να προσφέρουν νέα ερμηνευτικά εργαλεία, τα οποία δεν έχουν μελετηθεί επαρκώς στο πλαίσιο της οικονομικής και επιχειρησιακής ανάλυσης.
3. Μια πολλά υποσχόμενη προσέγγιση θα ήταν ο συνδυασμός διαφορετικών τύπων συσχέτισης σε ένα ενιαίο διαγνωστικό πλαίσιο, προκειμένου να αποτυπωθούν πληρέστερα οι σχέσεις μεταξύ των μεταβλητών.
4. Καθώς, κάποιες μέθοδοι, όπως ο Μέγιστος Συντελεστής Πληροφορίας (MIC) ή ο Distance Correlation, είναι υπολογιστικά απαιτητικές, προτείνεται η αξιολόγηση τους σε μεγάλα δεδομένα (big data).

ΑΝΑΦΟΡΕΣ

- Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., & Furlanello, C. (2013). Minerva and minepy: A C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics*, 29(3). <https://doi.org/10.1093/bioinformatics/bts707>
- Asuero, A. G., Sayago, A., & González, A. G. (2006). The correlation coefficient: An overview. In *Critical Reviews in Analytical Chemistry* (Vol. 36, Issue 1, pp. 41–59). <https://doi.org/10.1080/10408340500526766>
- Baak, M., Koopman, R., Snoek, H., & Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics and Data Analysis*, 152. <https://doi.org/10.1016/j.csda.2020.107043>
- Belsley, D. A. (1991). A Guide to Using the Collinearity Diagnostics. In *Computer Science in Economics and Management* (Vol. 4).
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (n.d.). *KNIME-The Konstanz Information Miner Version 2.0 and Beyond*. <http://labs.knime.org>
- Callaghan, K., & Chen, J. (2008). *Revisiting the Collinear Data Problem: An Assessment of Estimator "Ill-Conditioning" in Linear Regression - Practical Assessment, Research & Evaluation*. 13(5).
- Chan, J. Y. Le, Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. In *Mathematics* (Vol. 10, Issue 8). MDPI. <https://doi.org/10.3390/math10081283>
- Clark, M. (2013). *A Comparison of Correlation Measures*. Center for Social Research, University of Notre Dame. <https://www3.nd.edu/~mclark19/learn/PDFs/ComparisonCorrelation.pdf>
- Dalla, V., Mary, Q., & Phillips, P. C. B. (2020). *The testcorr Package Liudas Giraitis*.
- Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). KNIME for reproducible cross-domain analysis of life science data. In *Journal of*

Biotechnology (Vol. 261, pp. 149–156). Elsevier B.V.
<https://doi.org/10.1016/j.jbiotec.2017.07.028>

Forthofer, R. N. (n.d.). *BIostatISTICS: A GUIDE TO DESIGN, ANALYSIS, AND DISCOVERY*.

Fox J, & Weisberg S. (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>. In *Thousand Oaks CA: Sage*.

Frank E. (2025). *Hmisc: Harrell Miscellaneous*. (Version 5.2-3) [R package manual]. CRAN
<https://hbiostat.org/R/Hmisc/>

G. UDNY YULE, C. B. E., M. A., F. R. S., (1922). *AN INTRODUCTION TO THE THEOEY OF STATISTICS*.

Gogtay, N. J., & Thatte, U. M. (2017). Principles of Correlation Analysis. In *Journal of The Association of Physicians of India* ■ (Vol. 65).

Halkos, G. E., & Tsilika, K. D. (2018). Programming Correlation Criteria with free CAS Software. *Computational Economics*, 52(1), 299–311. <https://doi.org/10.1007/s10614-016-9604-1>

Halkos, G., Nomikos, S., & Tsilika, K. (2022). Evidence for Novel Structures Relating CSR Reporting and Economic Welfare: Environmental Sustainability—A Continent-Level Analysis. *Computational Economics*, 59(2), 415–444. <https://doi.org/10.1007/s10614-020-10091-5>

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2024). *Introduction to Econometrics with R*.

Hoeffding, W. (1948). A Non-Parametric Test of Independence. In *Source: The Annals of Mathematical Statistics* (Vol. 19, Issue 4).

Hotelling, H. (1935). *THE MOST PREDICTABLE CRITERION*. *Journal of Educational Psychology*, 26(2), pp. 139-42.

Hotelling, H. (1936). *Relations between two sets of variables*. 1–58.

Ignacio González, A., & Déjean Maintainer Sébastien Déjean, S. (2025). *Package “CCA” Title Canonical Correlation Analysis Version 1.2.2*. www.lsp.ups-tlse.fr/CCA

Imdadullah, M., Aslam, M., & Altaf, S. (1983). *mctest: An R Package for Detection of Collinearity among Regressors*. Johnston.

Lin, C., Wang, K., & Mueller, S. (2021). MCVIS: A New Framework for Collinearity Discovery, Diagnostic, and Visualization. *Journal of Computational and Graphical Statistics*, 30(1), 125–132. <https://doi.org/10.1080/10618600.2020.1779729>

Lind, D. A., Marchal, W. G., & Wathen, S. Adam. (2012). *Statistical techniques in business & economics*. McGraw-Hill / Irwin.

Looney, S. W., & Hagan, J. L. (2007). 4 Statistical Methods for Assessing Biomarkers and Analyzing Biomarker Data. In *Handbook of Statistics* (Vol. 27, pp. 109–147). [https://doi.org/10.1016/S0169-7161\(07\)27004-X](https://doi.org/10.1016/S0169-7161(07)27004-X)

Makowski, D., Ben-Shachar, M., Patil, I., & Lüdtke, D. (2020). Methods and Algorithms for Correlation Analysis in R. *Journal of Open Source Software*, 5(51). <https://doi.org/10.21105/joss.02306>

Mazuruse, P. (2014). Canonical correlation analysis: Macroeconomic variables versus stock returns. *Journal of Financial Economic Policy*, 6(2), 179–196. <https://doi.org/10.1108/JFEP-09-2013-0047>

Meyer D + Zeileis A + Hornik K. (2024). *Package “vcd” Title Visualizing Categorical Data*. <https://orcid.org/0000-0002-3237-0941>

Peterson, B. G., & Carl, P. (2022). PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis. R package version 2.0.4. *R Package Version*, 1(3).

Rizzo, M. L., & Székely, G. J. (2019). energy: E-Statistics: Multivariate Inference via the Energy of Data. In *R package version*.

- Salmerón, R., García, C. B., & García, J. (2018). Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation*, 88(12), 2365–2384. <https://doi.org/10.1080/00949655.2018.1463376>
- Salvatore S. Mangiafico. (2025). *{rcompanion}: Functions to Support Extension Education Program Evaluation*.
- Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6), 2769–2794. <https://doi.org/10.1214/009053607000000505>
- Thomas Nagler. (2025). Package ‘wdm.’ In *An Introduction to Copulas*. Springer New York. <https://doi.org/10.1007/0-387-28678-0>
- Tinbergen, J. (1947). The Use of Correlation Analysis in Economic Research. In *Arg* (Vol. 49).
- Uwe Menzel. (2022). *Package “CCP” Title Significance Tests for Canonical Correlation Analysis (CCA)*.
- Vinod, H. D. (2017). Generalized correlation and kernel causality with applications in development economics. *Communications in Statistics: Simulation and Computation*, 46(6), 4513–4534. <https://doi.org/10.1080/03610918.2015.1122048>
- William Revelle. (2025). *Package “psych” Title Procedures for Psychological, Psychometric, and Personality Research*. <https://personality-project.org/r/psych/>
- Καραπιστόλης, Δ. Ν. (2013). *Στατιστική Επιχειρήσεων*. Θεσσαλονίκη: Εκδόσεις Αθανάσιος Αλτιντζής.
- Ταμπάκης, Ν. & Χαψά, Ξ. (2013). *Εφαρμοσμένη Στατιστική: Εργαστηριακές Ασκήσεις*. Θεσσαλονίκη: Εκδόσεις Ζυγός.