ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
Σχολή Επιστημών Υγείας
Πανεπιστήμιο Θεσσαλίας

Πρόγραμμα Μεταπτυχιακών Σπουδών (ΠΜΣ)

«Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική»

SPSS ΣΤΗΝ ΙΑΤΡΙΚΗ

Ανάλυση ιατρικών δεδομένων με τη χρήση του στατιστικού πακέτου SPSS

Νικόλαος Γεωργίου
Πτυχιούχος Ιατρικής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Λάρισα 2024

No	Student's Name	Thesis
1	**ΓΕΩΡΓΙΟΥ ΝΙΚΟΛΑΟΣ** ngeorgiou63@gmail.com	**Ανάλυση ιατρικών δεδομένων με τη χρήση του στατιστικού πακέτου SPSS** 1. Ιατρικά δεδομένα-Συλλογή και επεξεργασία αυτών 2. Η μέθοδος clustering 3. Η μέθοδος διαχωριστικής ανάλυσης 4. Η μέθοδος πολλαπλής παλινδρόμησης 5. Χρησιμότητα και συμπεράσματα της στατιστικής επεξεργασίας ιατρικών δεδομένων

Supervisor: Θεόδωρος Μπρότσης

Εντεταλμένος Διδάσκων του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας (tmprotsis@uth.gr)

Co-supervisor 1: Σωτήριος Κωτσιαντής,

Αναπληρωτής Καθηγητής του Τμήματος Μαθηματικών του Πανεπιστήμιου Πατρών (sotos@math.upatras.gr)

Co-supervisor 2: Χρυσούλα Δοξάνη

Εντεταλμένη Διδάσκουσα του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας (doxani@med.uth.gr)

Πρόλογος

Η παρούσα Διπλωματική Εργασία εκπονήθηκε κατά την θερινή περίοδο του Ακαδημαϊκού Έτους 2023-2024 στο Τμήμα Ιατρικής του Πανεπιστημίου Θεσσαλίας στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών (ΠΜΣ) με τίτλο **«ΜΕΘοΔοΛοΓΙΑ ΒΙοΪΑΤΡΙΚΗΣ ΈΡΕΥΝΑΣ, ΒΙοΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΚΛΙΝΙΚΗ ΒΙοΠΛΗΡοΦοΡΙΚΗ»**. Η εργασία πραγματοποιήθηκε υπό την επίβλεψη των καθηγητών κου Θεόδωρου Μπρότση, Εντεταλμένου Διδάσκοντα του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας, κου Σωτήρη Κωτσιαντή, Αναπληρωτή Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Πατρών και της κας Χρυσούλας Δοξάνη, Εντεταλμένης Διδάσκουσας του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας. Αντικείμενο της εργασίας αποτελεί η ανάλυση ιατρικών δεδομένων με την χρήση του στατιστικού πακέτου SPSS.

Οφείλω να εκφράσω τις θερμές μου ευχαριστίες προς τους επιβλέποντες καθηγητές της εργασίας, για την καθοδήγησή τους και την πολύτιμη βοήθεια που προσέφεραν σε κάθε στάδιο εκπόνησης της διπλωματικής μου εργασίας. Επίσης, θερμά ευχαριστώ τους γονείς μου για τη συμπαράσταση και τη συνεχή βοήθειά τους στην προσπάθειά μου να ολοκληρώσω τις μεταπτυχιακές μου σπουδές στο Τμήμα Ιατρικής του Πανεπιστημίου Θεσσαλίας.

Νικόλαος Γεωργίου
Πτυχιούχος Ιατρικής

Περίληψη

Στην διπλωματική εργασία μελετάμε και αναλύουμε ιατρικά δεδομένα που λαμβάνονται, συλλέγονται ή και αποθηκεύνται σε κλινικές, νοσοκομεία και ερευνητικές ομάδες υπό το πρίσμα της στατιστικής επεξεργασίας του λογισμικού SPSS. Ειδικότερα, παρουσιάζοντας αρχικά έναν οδηγό με τα διάφορα είδη ιατρικών δεδομένων και τον ρόλο της στατιστικής επεξεργασίας τους με τη χρήση διαφόρων στατιστικών πακέτων στην Ιατρική Επιστήμη, περιγράφουμε μεθόδους μελέτης και ανάλυσης τέτοιων δεδομένων μέσα από το SPSS.

Περιγράφομε την μέθοδο clustering (ομαδοποίηση) και τύπους μεθόδων ομαδοποίησης, δίνοντας ιδιαίτερη έμφαση στη σπουδαιότητά της στην Ιατρική. Ως εφαρμογές της μεθόδου αυτής, μελετάμε στο SPSS την k-means ομαδοποίηση και την hierarchical (ιεραρχική) ομαδοποίηση, περιγράφοντας μέσα από παραδείγματα το περιβάλλον εντολών και επεξεργασίας του SPSS. Στη συνέχεια, περιγράφουμε τη διαχωριστική ανάλυση και τον ρόλο της στη μελέτη ιατρικών μελετών. Μέσα από παραδείγματα, περιγράφουμε την επεξεργασία ιατρικών δεδομένων στο SPSS χρησιμοπιώντας αυτή την μέθοδο. Επιπλέον, στην παρούσα εργασία, παρουσιάζουμε και την γραμμική παλινδρόμηση ως μία τεχνική με ποικίλες πρακτικές εφαρμογές στην Ιατρική. Στο SPSS μελετάμε το περιβάλλον δημιουργίας γραφικών παραστάσεων, μελέτης του συντελεστή συσχέτισης Pearson r και προσδιορισμού γραμμικών μοντέλων. Η λογιστική παλινδρόμηση συμπληρώνει την αντίστοιχη μελέτη παλινδρόμησης στο SPSS. Ολοκληρώνουμε την εργασία μελετώντας στο SPSS την μέθοδο Kaplan-Meier περί ανάλυσης επιβίωσης πέραν δεδομένων χρονικών σημείων.

Abstract

In the diploma thesis we study and analyze medical data obtained, collected or stored in clinics, hospitals and research teams in the prism of the statistical processing of the SPSS software. In particular, initially presenting a guide to the various types of medical data and the role of their statistical processing using various statistical packages in Medical Science, we describe methods of studying and analyzing such data through SPSS.

We describe the clustering method and types of clustering methods, emphasizing its importance in Medicine. As applications of this method, we study the k-means clustering and the hierarchical clustering in SPSS, describing through examples the command and processing environment of SPSS. Next, we describe the discriminant analysis and its role in the study of medical studies. Through examples, we describe the processing of medical data in SPSS using this method. In addition, in this work, we also present the linear regression as a technique with various practical applications in Medicine. In SPSS we study the environment for creating graphical representations, studying the Pearson correlation coefficient r and determining linear models. Logistic regression complements the corresponding regression study in SPSS. We complete the work by studying in SPSS the Kaplan-Meier method of survival analysis beyond given time points.

Περιεχόμενα

Εισαγωγή

Το αντικείμενο της παρούσας διπλωματικής εργασίας είναι η ανάλυση ιατρικών δεδομένων με χρήση του Στατιστικού Πακέτου SPSS. Στόχος της εργασίας είναι:

- Η ανασκόπηση και η παρουσίαση των ιατρικών δεδομένων.
- Η αναφορά στην έννοια της στατιστικής επεξεργασίας δεδομένων και παρουσίαση γνωστών στατιστικών πακέτων για τη στατιστική επεξεργασία δεδομένων.
- Τα συμπεράσματα και τα οφέλη μιας στατιστικής επεξεργασίας ιατρικών δεδομένων για την Ιατρική.
- Η ανάλυση της χρησιμότητας της έννοιας "clustering" στην Ιατρική μέσου του στατιστικού πακέτου SPSS.
- Η παρουσίαση της Διαχωριστικής Ανάλυσης και η χρησιμότητα αυτής στην διάγνωση ασθενειών στην Ιατρική μέσου του στατιστικού πακέτου SPSS.
- Η μελέτη, η παρουσίαση και η χρησιμότητα της πολλαπλής γραμμικής παλινδρόμησης στην Ιατρική μέσου του στατιστικού πακέτου SPSS.

Η διπλωματική εργασία χωρίζεται σε δύο μέρη, ένα θεωρητικό που αφορά την παρουσίαση του βασικού θεωρητικού υπόβαθρου και ένα πειραματικό στάδιο, που αφορά την μελέτη εφαρμογής του καταμερισμού στην περιοχή μελέτης με τη βοήθεια του στατιστικού πακέτου επεξεργασίας δεδομένων SPSS.

Μετά την οριστικοποίηση των στόχων, διενεργήθηκε βιβλιογραφική ανασκόπηση στη διεθνή και εγχώρια βιβλιογραφία με σκοπό την κάλυψη βασικών εννοιών στην επιστήμη των ιατρικών δεδομένων και της στατιστικής επεξεργασίας αυτών με χρήση του στατιστικού πακέτου SPSS. Δίνεται έμφαση στην συλλογή και παρουσίαση των ιατρικών δεδομένων και ιδιαίτερα στο κομμάτι εκείνο που αφορά περισσότερο την στατιστική επεξεργασία με την χρήση του στατιστικού πακέτου SPSS. Ιδιαίτερη σημασία δίνεται στο θέμα της πρόγνωσης ασθενειών στην Ιατρική. Ακολούθως, γίνεται συλλογή στοιχείων και δεδομένων για την περιοχή μελέτης, που είναι απαραίτητα για το πειραματικό στάδιο που ακολουθεί. Για τη συλλογή των στοιχείων έγινε αναζήτηση τόσο σε παλαιότερες μελέτες όσο και σε σχετική βιβλιογραφία και στο διαδίκτυο.

Η διπλωματική εργασία χωρίζεται σε έξι διακριτές μεταξύ τους ενότητες.

- Η πρώτη ενότητα αποτελεί την εισαγωγή, όπου περιγράφεται το αντικείμενο, ο στόχος της εργασίας, η διαδικασία εκπόνησής της και τέλος η δομή της μελέτης.
- Στη δεύτερη ενότητα (Κεφάλαιο 1) γίνεται εισαγωγή και παρουσίαση των ιατρικών δεδομένων. Ιδιαίτερα, εμβαθύνουμε στην χρησιμότητα της Στατιστικής Επεξεργασίας στην Ιατρική.
- Η τρίτη ενότητα (Κεφάλαιο 2) αναφέρεται στην έννοια του "clustering" και τη σημασία και την αξία της μεθόδου αυτής για την Ιατρική.
- Στην τέταρτη ενότητα (Κεφάλαιο 3) γίνεται περιγραφή και συνοπτική παρουσίαση της Διαχωριστικής Ανάλυσης στην Ιατρική.
- Στην πέμπτη ενότητα (Κεφάλαιο 4) παρουσιάζεται η πολλαπλή γραμμική παλινδρόμηση και η σημασία αυτής στην μελέτη ιατρικών δεδομένων.
- Στην έκτη ενότητα (Κεφάλαιο 5) γίνεται, μέσα από ένα απλό παράδειγμα, η παρουσίαση της Ανάλυσης επιβίωσης (μέθοδος Kaplan-Meier) στο πρόγραμμα SPSS.
- Τέλος, στην έβδομη ενότητα γίνεται η παρουσίαση την Βιβλιογραφίας (Ελληνική και Ξενόγλωσση) που έπαιξε σημαντικό ρόλο για την συγγραφή της διπλωματικής εργασίας.

ΚΕΦΑΛΑΙΟ 1ο
Ιατρικά δεδομένα και στατιστική επεξεργασία δεδομένων στην υγεία

Στο κεφάλαιο αυτό γίνεται παρουσίαση των ιατρικών δεδομένων, της στατιστικής επεξεργασίας δεδομένων, των βασικών στατιστικών πακέτων επεξεργασίας δεδομένων και της χρησιμότητας της Στατιστικής στην εξαγωγή χρήσιμων συμπεράσματων τόσο στην διαγνωση ασθενειών όσο και στην Ιατρική έρευνα.

1.1 Ιατρικά Δεδομένα

Τα ιατρικά δεδομένα περιέχουν πληροφορίες για την κατάσταση της υγείας ενός ατόμου και την ιατρική περίθαλψη που έχει λάβει. Τέτοιες πληροφορίες μπορούν να καταγραφούν σε χαρτί ή σε ψηφιακά αρχεία και βάσεις δεδομένων. Τα δεδομένα αυτά είναι το κλειδί για την ανάπτυξη τεχνολογίας υγειονομικής περίθαλψης, τη δημιουργία καλύτερων θεραπειών και τη βελτίωση της φροντίδας των ασθενών.

Τα ιατρικά δεδομένα συλλέγονται και αποθηκεύονται σε κλινικές και νοσοκομεία αλλά μπορούν επίσης να κοινοποιηθούν σε δημόσιες αρχές υγείας και άλλες δημόσιες αρχές όπως για παράδειγμα, την αστυνομία ή ένα δικαστήριο. Τα δεδομένα που αφορούν την υγεία μπορούν επίσης να χρησιμοποιηθούν στην έρευνα της ιατρικής επιστήμης. Ωστόσο, σε αυτή την περίπτωση, πρέπει να πληρούνται αυστηρά κριτήρια και κατά γενικό κανόνα απαιτείται η συναίνεση των ατόμων.

Παρακάτω παρουσιάζουμε έναν οδηγό με τα διάφορα είδη δεδομένων από ιατρικά αρχεία έως προσωπικές πληροφορίες υγείας, όπως αυτά καταγράφηκαν σε σχετικές αναζητήσεις μας σε βιβλιογραφικές παραπομπές (βλ. για παράδειγμα, [8-9], [11-12], [22-23], [26-27], [30]).

1.1.1 Ηλεκτρονικοί Φάκελοι Υγείας και Ηλεκτρονικοί Ιατρικοί Φάκελοι

Τα ηλεκτρονικά αρχεία υγείας (EHR) και τα ηλεκτρονικά ιατρικά αρχεία (EMR) είναι δύο βασικοί τύποι δεδομένων υγειονομικής περίθαλψης με τους οποίους οι πάροχοι υγειονομικής περίθαλψης διαχειρίζουν τις πληροφορίες των ασθενών.

- Ο **Ηλεκτρονικός Ιατρικός Φάκελος Ασθενών (EMR)** είναι μία ισχυρή, ασφαλής και επεκτάσιμη πλατφόρμα. Παρέχει εύκολη και άμεση πρόσβαση στα δεδομένα των ασθενών. Προσφέρει στους επαγγελματίες υγείας τις πληροφορίες του ιατρικού φακέλου του ασθενούς. Η πρόσβαση επιτρέπεται από οποιαδήποτε φυσική θέση και συσκευή στο διαδίκτυο, αρκεί να εξασφαλίζονται τα απαιτούμενα ειδικά δικαιώματα πρόσβασης και χρήσης. Με αυτόν τον τρόπο έχουμε ψηφιακή έκδοση του διαγράμματος του ασθενούς με όλες τις πληροφορίες αποθηκευμένες στο σύστημα του υπολογιστή. Γνωρίζουμε άμεσα για κάθε ασθενή το ιατρικό ιστορικό του, τις εργαστηριακές

εξετάσεις του, τις διαγνώσεις κ.λπ.. Η αποθήκευση γίνεται στο σύστημα και όχι με τη μορφή ογκωδών αρχείων χαρτιού. οι πληροφορίες αυτές όμως δεν μπορούν να πάνε έξω από τις εγκαταστάσεις του οργανισμού.

οι κλινικοί γιατροί μπορούν να χρησιμοποιήσουν ηλεκτρονικά ιατρικά αρχεία για:

- Να προσδιορίσουν εάν οι ασθενείς χρειάζονται προληπτικούς ελέγχους ή εξετάσεις.
- Να εξετάσουν την απόδοση των ασθενών σε συγκεκριμένα μέτρα, όπως καρδιαγγειακά νοσήματα ή ανοσοποιήσεις.
- Να αξιολογήσουν και να βελτιώσουν την ποιότητα των υπηρεσιών υγειονομικής περίθαλψης σε μια κλινική.

Ωστόσο, τα δεδομένα στο EMR δεν φεύγουν εύκολα από τον οργανισμό. Το EMR δεν διαφέρει από αυτή την άποψη από τα έντυπα αρχεία.

- Τα **δεδομένα EHR**, από την άλλη πλευρά, προωθούν αυτήν την ιδέα ένα βήμα παραπέρα, προσφέροντας μια πιο ολοκληρωμένη και ολιστική άποψη της υγείας του ασθενούς. Τα EHR περιλαμβάνουν ζωτικά σημεία, ιατρικό ιστορικό του παρελθόντος, διαγνώσεις, σημειώσεις προόδου, φάρμακα, αλλεργίες, δεδομένα εργαστηρίου, ημερομηνίες εμβολιασμού και αναφορές απεικόνισης. Αυτές οι πληροφορίες μπορούν να ταξιδέψουν και εκτός των εγκαταστάσεων του οργανισμού. Έχουν σχεδιαστεί για να μοιράζονται πληροφορίες με άλλους παρόχους υγειονομικής περίθαλψης και οργανισμούς - όπως εργαστήρια, ειδικούς, εγκαταστάσεις ιατρικής απεικόνισης, φαρμακεία, εγκαταστάσεις έκτακτης ανάγκης και κλινικές- διασφαλίζοντας ότι οι πληροφορίες υγείας ενός ασθενούς είναι προσβάσιμες και με ασφάλεια κοινοποιούνται σε διαφορετικά περιβάλλοντα υγειονομικής περίθαλψης. Αυτή η απρόσκοπτη ανταλλαγή πληροφοριών όχι μόνο βελτιώνει την ακρίβεια των διαγνώσεων και των σχεδίων θεραπείας, αλλά ενισχύει επίσης τη συνολική ποιότητα της περίθαλψης καθιστώντας το ιστορικό υγείας του ασθενούς εύκολα προσβάσιμο σε όλους τους εξουσιοδοτημένους παρόχους που εμπλέκονται στη φροντίδα τους.

  Τα δεδομένα EHR είναι τα δεδομένα που συλλέγονται όταν βλέπουμε έναν γιατρό, παίρνουμε μια συνταγή από το φαρμακείο ή ακόμα και από μια επίσκεψη στον οδοντίατρο. Επίσης, τα δεδομένα EHR μπορούν να προέρχονται από πολλές διαφορετικές πηγές στον σημερινό κόσμο: συσκευές παρακολούθησης καρδιακών παλμών, ακτινογραφίες και άλλες ακτινολογικές σαρώσεις, ανιχνευτές φυσικής κατάστασης κ.λπ. Αυτά τα δεδομένα γίνονται ακόμη πιο κρίσιμα όταν μπορούμε να τα έχουμε διαχρονικά για την υγεία ενός ασθενούς και τα συλλέγουμε με τρόπο που επιτρέπει στους επαγγελματίες υγείας και τους επιστήμονες δεδομένων να κάνουν ουσιαστικές και ακριβείς προβλέψεις.

  Γενικά τα δεδομένα EHR συμβάλουν σε καλύτερη διάγνωση και έλεγχο. Τονίζουμε ότι οι εταιρείες αξιοποιούν τα μεγάλα σύνολα δεδομένων σε πληθυσμούς για να βρουν τρόπους πρόβλεψης ασθενειών και καταστάσεων πολύ πριν εμφανιστούν, βλέποντας τάσεις που μόνο αυτά τα μεγάλα σύνολα δεδομένων μπορούν να παρέχουν.

  Τέλος, το EHR ξεκλειδώνει πιο ισχυρές θεραπείες συνδυάζοντας δεδομένα EHR με άλλες πηγές δεδομένων, όπως δεδομένα γενετικής και απεικόνισης, για τη δημιουργία ακόμη πιο ισχυρών και εξατομικευμένων θεραπειών. οι εταιρείες χρησιμοποιούν δεδομένα EHR για να προσδιορίσουν πιο αποτελεσματικά τις κοόρτες ασθενών και τις κλινικές δοκιμές και να προβλέψουν τους κινδύνους για διαφορετικά φάρμακα με βάση δεδομένα πληθυσμού.

1.1.2 Διοικητικά Στοιχεία

Τα διοικητικά δεδομένα υγειονομικής περίθαλψης δημιουργούνται κατά τη διάρκεια κάθε αλληλεπίδρασης με το σύστημα υγειονομικής περίθαλψης, είτε πρόκειται για επίσκεψη στο ιατρείο, μια διαγνωστική διαδικασία, εισαγωγή στο νοσοκομείο, αλληλεπιδράσεις με εγκαταστάσεις φροντίδας, κατ' οίκον φροντίδα ή συμπλήρωση φαρμακευτικών συνταγών. Αυτά τα δεδομένα συλλέγονται για λόγους διαχείρισης ή χρέωσης, αλλά μπορούν επίσης να χρησιμοποιηθούν για την ανάλυση της παροχής υγειονομικής περίθαλψης, των πλεονεκτημάτων, των μειονεκτημάτων και της σχέσης κόστους-αποτελεσματικότητας. Σε ορισμένες περιοχές, τέτοια δεδομένα περιλαμβάνουν λεπτομέρειες σχετικά με τις πληροφορίες εξιτηρίου από το νοσοκομείο, οι οποίες αργότερα θα αναφερθούν στην κυβέρνηση ή σε άλλους φορείς.

1.1.3 Στοιχεία Διοικητικής Υγείας

Αυτοί οι τύποι δεδομένων υγειονομικής περίθαλψης καλύπτουν κυρίως δεδομένα χρηματοοικονομικής και παρακολούθησης δραστηριοτήτων, τα οποία χρησιμεύουν ως βασικά στοιχεία για διάφορες στρατηγικές παρακολούθησης.

Τα διοικητικά δεδομένα είναι τα δεδομένα που παράγονται ως αποτέλεσμα των καθηκόντων και λειτουργιών ρουτίνας που εκτελούνται από οργανισμούς, εταιρείες, κυβερνητικές υπηρεσίες και άλλες οντότητες. Περιλαμβάνει πληροφορίες που συλλέγονται και διατηρούνται συστηματικά κατά την τακτική λειτουργία αυτών των φορέων για την υποστήριξη των διοικητικών και λειτουργικών αναγκών τους.

Αυτός ο τύπος δεδομένων δεν συλλέγεται συνήθως για ερευνητικούς σκοπούς, αλλά μάλλον εξυπηρετεί εσωτερικές λειτουργίες όπως η τήρηση αρχείων, η διαχείριση και η λήψη αποφάσεων. Παραδείγματα περιλαμβάνουν αρχεία υγειονομικής περίθαλψης, δεδομένα εγγραφής στην εκπαίδευση, κρατικά φορολογικά αρχεία, επιχειρηματικές οικονομικές συναλλαγές και πληροφορίες ανθρώπινου δυναμικού.

Τα διοικητικά δεδομένα μπορούν να αξιοποιηθούν για τη βελτιστοποίηση της ανάπτυξης οργανισμών και επιχειρήσεων. Χρησιμεύουν ως εργαλείο για τον προσδιορισμό της αποτελεσματικότητας των εφαρμοζόμενων στρατηγικών και τον προσδιορισμό του εάν ευθυγραμμίζονται με τους καθορισμένους στόχους.

Συμπερασματικά, τα διοικητικά δεδομένα χρησιμεύουν ως πυξίδα για τους παρέχοντας κρίσιμες γνώσεις για τη λειτουργική υγεία και την πρόοδό τους προς τους στόχους. Αξιοποιώντας τη δύναμη αυτών των δεδομένων, οι οργανισμοί μπορούν να λαμβάνουν τεκμηριωμένες αποφάσεις, να προσδιορίζουν τομείς προς βελτίωση και να διασφαλίζουν ότι οι στρατηγικές τους ευθυγραμμίζονται με τους πρωταρχικούς στόχους.

1.1.4 Στοιχεία αξιώσεων

Τα δεδομένα αξιώσεων περιλαμβάνουν τις χρεώσιμες αλληλεπιδράσεις μεταξύ των ασφαλισμένων ασθενών και του συστήματος παροχής υγειονομικής περίθαλψης, κυρίως για κάλυψη ή αποζημίωση. οι αξιώσεις περιέχουν πληροφορίες σχετικά με τις διαγνώσεις ασθενών, τις διαδικασίες και τις δοκιμές

που πραγματοποιήθηκαν, τις ημερομηνίες παροχής υπηρεσιών, το κόστος και τον τόπο παροχής των υπηρεσιών.

1.1.5 Μητρώα ασθενών / ασθενειών

Τα μητρώα ασθενών/ασθενειών είναι συλλογές δευτερογενών δεδομένων που σχετίζονται με ασθενείς που έχουν διαγνωστεί με μια συγκεκριμένη πάθηση, ασθένεια ή διαδικασία. Αυτά τα δεδομένα είναι απαραίτητα για την παρακολούθηση των φαρμακευτικών προϊόντων μετά την κυκλοφορία.

Τα συστήματα κλινικών πληροφοριών παρακολουθούν βασικά δεδομένα για χρόνιες παθήσεις όπως το Αλτσχάιμερ, ο καρκίνος, ο διαβήτης, οι καρδιακές παθήσεις και το άσθμα. Αυτές οι πηγές πληροφοριών προσφέρουν κρίσιμες πληροφορίες για το χειρισμό και την παρακολούθηση των καταστάσεων των ασθενών.

Στην κατηγορία αυτή είναι δυνατό να συναντήσουμε και τον ιατρικό φάκελο του ασθενούς. ο ιατρικός φάκελος είναι ένα ιστορικό της υγείας κάποιου. Ένας ιατρικός φάκελος περιλαμβάνει πληροφορίες σχετικά με τα εξής:

- ηλικία, φύλο και εθνικότητα,
- ύψος και βάρος,
- ατρικά προβλήματα (όπως άσθμα, επιληψία ή διαβήτης),
- ζητήματα ψυχικής υγείας (όπως άγχος ή κατάθλιψη),
- αποτελέσματα ιατρικών εξετάσεων (από εργαστηριακές εξετάσεις, ακτινογραφίες κ.λπ.),
- φάρμακα, συμπεριλαμβανομένων των δόσεων και της συχνότητας λήψης του φαρμάκου,
- αλλεργίες σε φάρμακα (συνταγογραφούμενα και μη), τσιμπήματα και τσιμπήματα εντόμων, τρόφιμα και οποιεσδήποτε άλλες ουσίες,
- Νοσηλεία,
- εμβολιασμοί (εμβόλια),
- τιμολόγηση και ασφάλιση.

1.1.6 Έρευνες Υγείας

Εθνικές έρευνες για την υγεία αξιολογούν την υγεία του πληθυσμού και εκτιμούν την συχνότητα μιας νόσου και την εξέλιξή της σ' ένα ορισμένο διάστημα (τον επιπολασμό της νόσου). Αυτές οι πηγές δεδομένων επιμελούνται για ερευνητικούς στόχους και είναι ευρέως διαθέσιμες.

Οι έρευνες για την υγεία βρίσκονται στην πρώτη γραμμή στη λήψη αποφάσεων για τη διαμόρφωση σχεδίων υγείας, προσφέροντας ακριβή δεδομένα για την επιδημιολογική κατάσταση, τα πρότυπα υγείας, τις συμπεριφορές στον τρόπο ζωής και την εμπειρία των ασθενών με τις υπηρεσίες υγειονομικής περίθαλψης. Οι ερευνητές δημόσιας υγείας χρησιμοποιούν δεδομένα ερευνών υγείας για να αναλύσουν συμπεριφορές που σχετίζονται με την υγεία και την ψυχολογική ευεξία. οι έρευνες για την υγεία είναι πολύτιμες για τον εντοπισμό συγκεκριμένων καταστάσεων υγείας ή παραγόντων κινδύνου εντός της κοινότητας κοντά σε εγκαταστάσεις υγειονομικής περίθαλψης, συμπεριλαμβανομένης της χρήσης καπνού και αλκοόλ, ανθυγιεινών προγραμμάτων διατροφής και σωματικής αδράνειας.

1.1.7 Δεδομένα κλινικών δοκιμών

Τα δεδομένα κλινικών δοκιμών προέρχονται από ερευνητικές μελέτες που αξιολογούν ιατρικές, χειρουργικές ή θεραπείες σε ασθενείς. Αυτά τα δεδομένα χρησιμεύουν πάρα πολύ στους ερευνητές κυρίως για:

- Εύρεση μεθόδων για την έγκαιρη ανίχνευση της νόσου, συχνά πριν από τα συμπτώματα.
- Ανακάλυψη στρατηγικών για τη λήψη προληπτικών ενεργειών για συγκεκριμένα θέματα υγείας, ακόμη και για φαινομενικά υγιείς ανθρώπους.
- Βελτίωση της ποιότητας ζωής των ασθενών που αντιμετωπίζουν σοβαρή ασθένεια ή χρόνιες παθήσεις.
- Βελτίωση της απόδοση των επαγγελματιών υγείας και των φροντιστών αυτής.

Τα δεδομένα κλινικών δοκιμών μπορούν να χρησιμοποιηθούν σε διάφορες δημοσιεύσεις. Υπάρχουν διάφορα μητρώα που παρέχουν πρόσβαση σε βάσεις δεδομένων κλινικών δοκιμών, όπως το ClinicalTrials.gov και το OpenTrials.

1.1.8 Γονιδιωματικά Δεδομένα

Τα γονιδιωματικά δεδομένα περιέχουν τη διάταξη και τη λειτουργία του γονιδιώματος ενός οργανισμού, το οποίο περιλαμβάνει όλα τα κυτταρικά δεδομένα που είναι απαραίτητα για την ανάπτυξη και τη λειτουργία του. οι γονιδιωματικές πληροφορίες χρησιμεύουν ως βάση για την επιστήμη των γονιδιωματικών δεδομένων, τη συγχώνευση μελετών γενετικής και υπολογιστικής βιολογίας με στατιστική ανάλυση δεδομένων και επιστήμη των υπολογιστών. Για παράδειγμα, οι ερευνητές γονιδιωματικών δεδομένων χρησιμοποιούν δεδομένα αλληλουχίας DNA για να διερευνήσουν ασθένειες και να αποκαλύψουν νέες θεραπείες και φάρμακα. Αυτά τα δεδομένα συλλέγονται για να δείξουν πώς τα γενετικά δεδομένα επηρεάζουν την ανάπτυξη και τη λειτουργία των οργανισμών, βοηθώντας έτσι την έρευνα στις βιοεπιστήμες, τη διάγνωση γενετικών ασθενειών και την ανάπτυξη φαρμάκων.

1.2 Γενικά σχόλια για τα ιατρικά δεδομένα

1. Τα σύνολα δεδομένων που χρησιμοποιούνται από ιατρικούς ερευνητές προέρχονται από μια ποικιλία διαφορετικών πηγών. Τα πιο συνηθισμένα είναι:

- Δεδομένα ασθενών από διοικητικές λεπτομέρειες νοσοκομείων και πρωτοβάθμιας περίθαλψης έως τις ιδιαιτερότητες της θεραπείας, τις ιατρικές και διαγνωστικές διαδικασίες.
- Έρευνα για την ευημερία των πληθυσμών, είτε επικεντρώνεται σε μια συγκεκριμένη πάθηση υγείας όπως ο καρκίνος είτε σε παράγοντες που επηρεάζουν τη δημόσια υγεία όπως το κάπνισμα.
- Γενετικές πληροφορίες που μπορούν να εξαχθούν από δείγματα αίματος ή ιστών.
- Εικόνες πλούσιες σε πληροφορίες όπως ακτινογραφίες, μαγνητική τομογραφία και αξονική τομογραφία.
- Φορητές συσκευές για υγεία και φυσική κατάσταση που παρέχουν δεδομένα σχετικά με μετρήσεις όπως ο καρδιακός ρυθμός, η σωματική δραστηριότητα και οι θερμίδες.

2. Υπάρχουν περισσότεροι από 40 οργανισμοί ανάπτυξης προτύπων που δραστηριοποιούνται στον τομέα της υγειονομικής περίθαλψης των ΗΠΑ. Μεταξύ των ευρέως αναγνωρισμένων προτύπων είναι:

- FHIR - Fast Healthcare Information Resources

- HL7 - Health Level 7 International
- NCDPD - Εθνικό Συμβούλιο για Προγράμματα Συνταγογραφούμενων Φαρμάκων
- IHTSDO - Διεθνείς οργανισμοί Ανάπτυξης Προτύπων ορολογίας Υγείας
- Πρότυπα DirectTrust
- CDISC - Κοινοπραξία προτύπων ανταλλαγής κλινικών δεδομένων.

3. Τα δεδομένα υγειονομικής περίθαλψης αποθηκεύονται και εμφανίζονται σε μία από τις ακόλουθες μορφές:

- **Δομημένα δεδομένα:** Τα δομημένα δεδομένα αναφέρονται σε πληροφορίες που συλλέγονται και αποθηκεύονται με καλά οργανωμένο, συχνά τυποποιημένο τρόπο. Τα δομημένα δεδομένα διαδραματίζουν ζωτικό ρόλο για τη σύγκριση και την ενοποίηση στοιχείων δεδομένων μεταξύ ασθενών, παρόχων υγειονομικής περίθαλψης και πληρωτών.
- **Μη δομημένα δεδομένα:** Δωρεάν δεδομένα κειμένου ή εικόνας, όπως αναφορές προόδου, κλινικές σημειώσεις και ακτινολογικές εικόνες αποθηκεύονται συχνά ως μη δομημένα δεδομένα σε ιατρικά αρχεία. Το μεγαλύτερο μέρος των δεδομένων υγειονομικής περίθαλψης αποθηκεύεται σε αυτή τη μορφή, θέτοντας προκλήσεις για ενοποίηση και συγκρίσεις μεταξύ συστημάτων.

1.3 Στατιστική επεξεργασία δεδομένων στην υγεία

Ο απώτερος στόχος του ιατρικού ερευνητή μέσου της έρευνάς του είναι να απαντήσει σε ερωτήσεις και να λύσει προβλήματα που σχετίζονται με την ανθρώπινη υγεία. Για να το πετύχει αυτό, πρέπει να συλλέξει και να αναλύσει δεδομένα με ουσιαστικό τρόπο. Εκεί μπαίνει η Στατιστική Ανάλυση (βλ. για παράδειγμα, [4], [7], [13-14], [18-19], [20], [25], [28]).

Η Στατιστική Ανάλυση είναι η διαδικασία χρήσης μαθηματικών και υπολογιστικών τεχνικών για την ανάλυση δεδομένων που συλλέγονται από πειράματα ή έρευνες. Στην ιατρική έρευνα, η Στατιστική Ανάλυση χρησιμοποιείται για την αξιολόγηση των σχέσεων μεταξύ των μεταβλητών, τον προσδιορισμό της σημασίας των αποτελεσμάτων και την εξαγωγή συμπερασμάτων σχετικά με τα δεδομένα.

Υπάρχουν διάφοροι λόγοι για τους οποίους η Στατιστική Ανάλυση είναι κρίσιμη στην ιατρική έρευνα.

- Πρώτον, βοηθά στη μείωση του αντίκτυπου της τυχαίας μεταβλητότητας στα δεδομένα και αυξάνει την ακρίβεια των αποτελεσμάτων.
- Δεύτερον, επιτρέπει στους ερευνητές να δοκιμάσουν τις υποθέσεις τους και να κάνουν προβλέψεις για μελλοντικά αποτελέσματα.
- Τρίτον, παρέχει έναν τρόπο λήψης αποφάσεων σχετικά με τα δεδομένα με βάση τα στοιχεία και την κοινοποίηση των αποτελεσμάτων σε άλλους.

Τα βήματα στη Στατιστική Ανάλυση στην ιατρική έρευνα περιλαμβάνουν:

- Καθορισμός του ερευνητικού ερωτήματος και προσδιορισμός των μεταβλητών.
- Συλλογή και καθαρισμός των δεδομένων.
- Διερευνητική ανάλυση δεδομένων.
- Επιλογή των κατάλληλων στατιστικών δοκιμών.
- Διενέργεια των στατιστικών δοκιμών και ερμηνεία των αποτελεσμάτων.
- Εξαγωγή συμπερασμάτων και διατύπωση συστάσεων με βάση τα αποτελέσματα.
- Τύποι στατιστικών δοκιμών στην ιατρική έρευνα: *Υπάρχουν πολλοί διαφορετικοί τύποι στατιστικών δοκιμών που μπορούν να χρησιμοποιηθούν στην ιατρική έρευνα, συμπεριλαμβανομένων των τεστ t,*

ANOVA, παλινδρόμησης και τεστ χ-τετράγωνο. Η επιλογή του τεστ που θα χρησιμοποιηθεί εξαρτάται από το ερευνητικό ερώτημα, το είδος των δεδομένων και τους στόχους της μελέτης.

Τα αποτελέσματα των στατιστικών δοκιμών μπορεί να είναι πολύπλοκα και δυσνόητα, αλλά είναι σημαντικό να ερμηνευτούν σωστά. Τα αποτελέσματα πρέπει να παρουσιάζονται με σαφή και συνοπτικό τρόπο και να συνοδεύονται από κατάλληλα γραφήματα και πίνακες. Είναι επίσης σημαντικό να ληφθούν υπόψη οι περιορισμοί της μελέτης και οι επιπτώσεις των αποτελεσμάτων για μελλοντική έρευνα.

Συμπερασματικά, η Στατιστική Ανάλυση είναι ένα κρίσιμο συστατικό της ιατρικής έρευνας. Παρέχει έναν τρόπο αξιολόγησης των σχέσεων μεταξύ μεταβλητών, προσδιορισμού της σημασίας των αποτελεσμάτων και εξαγωγής συμπερασμάτων σχετικά με τα δεδομένα. Χρησιμοποιώντας κατάλληλες στατιστικές δοκιμές και ερμηνεύοντας σωστά τα αποτελέσματα, οι ιατρικοί ερευνητές μπορούν να υποστηρίξουν τα ευρήματά τους και να συμβάλλουν στην πρόοδο της ανθρώπινης υγείας.

1.4 Ο ρόλος της Στατιστικής στην κλινική έρευνα

Η κλινική έρευνα περιλαμβάνει τη διερεύνηση προτεινόμενων ιατρικών θεραπειών, την αξιολόγηση των σχετικών οφελών των ανταγωνιστικών θεραπειών και τη δημιουργία βέλτιστων συνδυασμών θεραπείας (βλ. για παράδειγμα, [3-4], [7], [10], [18], [27]). Η κλινική έρευνα επιχειρεί να απαντήσει σε ερωτήματα όπως για παράδειγμα: *Πρέπει ένας άνδρας με καρκίνο του προστάτη να υποβληθεί σε ριζική προστατεκτομή ή ακτινοβολία ή να περιμένει προσεκτικά; Είναι η συχνότητα σοβαρών ανεπιθύμητων ενεργειών μεταξύ των ασθενών που λαμβάνουν μια νέα αναλγητική θεραπεία μεγαλύτερη από τη συχνότητα σοβαρών ανεπιθύμητων ενεργειών σε ασθενείς που λαμβάνουν την τυπική θεραπεία;* Πριν από την ευρεία χρήση των πειραματικών δοκιμών, οι κλινικοί γιατροί προσπάθησαν να απαντήσουν σε τέτοιες ερωτήσεις γενικεύοντας από τις εμπειρίες μεμονωμένων ασθενών στον πληθυσμό γενικότερα.

Η χρήση στατιστικών μεθόδων επιτρέπει στους κλινικούς ερευνητές να αντλούν λογικά και ακριβή συμπεράσματα από τις συλλεγόμενες πληροφορίες και να λαμβάνουν ορθές αποφάσεις παρουσία αβεβαιότητας. Η κλινική και η στατιστική λογική είναι και οι δύο κρίσιμες για την πρόοδο στην ιατρική. οι κλινικοί ερευνητές πρέπει να γενικεύουν από τους λίγους σε πολλούς και να συνδυάζουν τα εμπειρικά στοιχεία με τη θεωρία. Τόσο στις ιατρικές όσο και στις στατιστικές επιστήμες, η εμπειρική γνώση παράγεται από παρατηρήσεις και δεδομένα. Η στατιστική θεωρία προέρχεται από μαθηματικά και πιθανολογικά μοντέλα.

Μια κλινική δοκιμή είναι στην πραγματικότητα ένα πείραμα που δοκιμάζει ιατρικές θεραπείες σε ανθρώπους. ο κλινικός ερευνητής ελέγχει παράγοντες που συμβάλλουν στη μεταβλητότητα και την προκατάληψη, όπως η επιλογή των υποκειμένων, η εφαρμογή της θεραπείας, η αξιολόγηση του αποτελέσματος και οι μέθοδοι ανάλυσης. Η διάκριση μιας κλινικής δοκιμής από άλλους τύπους ιατρικών μελετών είναι η πειραματική φύση της δοκιμής και η εμφάνισή της σε ανθρώπους. Ο σχεδιασμός είναι η διαδικασία ή η δομή που απομονώνει τους παράγοντες ενδιαφέροντος. Αν και ο ερευνητής σχεδιάζει μια δοκιμή για τον έλεγχο της μεταβλητότητας λόγω παραγόντων διαφορετικών από τη θεραπεία που τον ενδιαφέρει, υπάρχει εγγενώς μεγαλύτερη μεταβλητότητα στην έρευνα που περιλαμβάνει ανθρώπους από ό,τι σε μια ελεγχόμενη εργαστηριακή κατάσταση.

Ο όρος «κλινική δοκιμή» προτιμάται από τον όρο «κλινικό πείραμα», επειδή ο τελευταίος όρος μπορεί να υποδηλώνει ασέβεια για την αξία της ανθρώπινης ζωής.

Οι κλινικές δοκιμές χρησιμοποιούνται για την ανάπτυξη και δοκιμή παρεμβάσεων σε όλους σχεδόν τους τομείς της ιατρικής και της δημόσιας υγείας. Σε πολλές χώρες, η έγκριση για την εμπορία νέων φαρμάκων εξαρτάται από την αποτελεσματικότητα και την ασφάλεια των αποτελεσμάτων από κλινικές δοκιμές. Παρόμοιες απαιτήσεις υπάρχουν για την εμπορία εμβολίων. Οι χειρουργικές επεμβάσεις θέτουν μοναδικές προκλήσεις, δεδομένου ότι οι χειρουργικές προσεγγίσεις συνήθως αναλαμβάνονται για ασθενείς με καλή πρόγνωση και ενδέχεται να μην επιδέχονται τυχαιοποίηση ή απόκρυψη ερευνητών και ασθενών στην παρέμβαση (ή άλλων συνθηκών που μπορεί να οδηγήσουν σε προκαταλήψεις). Οι κλινικές δοκιμές είναι χρήσιμες για την απόδειξη της αποτελεσματικότητας και της ασφάλειας διαφόρων ιατρικών θεραπειών, προληπτικών μέτρων και διαγνωστικών διαδικασιών, εάν η θεραπεία μπορεί να εφαρμοστεί ομοιόμορφα και υπάρχει η δυνατότητα να ελεγχθούν πιθανές προκαταλήψεις.

1.5 Γνωστά προγράμματα Στατιστικής Επεξεργασίας

Τα τελευταία χρόνια, λόγω της συνεχής εξέλιξης της τεχνολογίας ο κλάδος της Στατιστικής Ανάλυσης Δεδομένων (Data Analysis) εξελίσσεται συνεχώς. Το τελευταίο διάστημα δημιουργούνται νέα στατιστικά πακέτα και τα ήδη υπάρχοντα εξελίσσονται συνεχώς. Τα πιο διαδεδομένα στατιστικά πακέτα την τελευταία δεκαετία είναι το SPSS, η R, το Minitab, το SAS και το STATA (βλ. για παράδειγμα, [1-2], [5-6], [15-17], [21], [24-33]).

Το στατιστικό πακέτο SPSS

Το στατιστικό πρόγραμμα SPSS αποτελεί μια ισχυρή στατιστική πλατφόρμα, που παρέχει ένα πλήρες σύνολο λειτουργιών που επιτρέπουν μέσω των υπαρχόντων πληροφοριών να εξάγονται συμπεράσματα. Τα πλεονεκτήματα του πακέτου αυτού είναι :

- ✓ Εύκολη Χρήση.
- ✓ Φιλικός Επεξεργαστής Δεδομένων.
- ✓ Δυνατότητα Ανάλυσης μεικτών δεδομένων.

Το στατιστικό πακέτο R

Η R αποτελεί ταυτόχρονα ένα περιβάλλον και μια γλώσσα προγραμματισμού για στατιστικούς υπολογισμούς και γραφικές παραστάσεις. Τα πλεονεκτήματα χρήσης του πακέτου R είναι:

- ✓ Μεγάλη γκάμα στατιστικών και γραφικών τεχνικών.
- ✓ Επεκτάσιμη Γλώσσα.
- ✓ Παρέχει την δυνατότητα για έρευνα ανοιχτού κώδικα.
- ✓ Ελεύθερο Λογισμικο.

Το στατιστικό πακέτο Minitab

Το Minitab συνίσταται για στατιστικές αναλύσεις σε εφαρμογές ελέγχου ποιότητας, επομένως προσφέρει μεγάλη γκάμα στατιστικών μεθόδων ανάλυσης. Τα πλεονεκτήματα χρήσης αυτού του πακέτου είναι:

- ✓ Αρκετές εκπαιδευτικές λειτουργίες.
- ✓ Ύπαρξη Βοηθού [Assistant].
- ✓ Άμεση αποστολή των αποτελεσμάτων στο Word ή στο PowerPoint.

Το στατιστικό πακέτο SAS

Το SAS είναι ένα στατιστικό πακέτο που απευθύνεται κυρίως σε εξειδικευμένους χρήστες και η

χρήση του απαιτεί συγγραφή κώδικα για την ανάλυση των δεδομένων. Τα πλεονεκτήματα χρήσης SAS είναι:

- ✓ Ικανότητα διασύνδεσης με βάσεις δεδομένων και η εκτέλεση ερωτημάτων μέσω της γλώσσας Structured Query Language (SQL).
- ✓ Δυνατότητα χειρισμού τεράστιου όγκου δεδομένων.
- ✓ Ισχυρά γραφικά εργαλεία.

Το στατιστικό πακέτο STATA

Το Stata είναι ένα στατιστικό πακέτο που απευθύνεται τόσο σε αρχάριους όσο και σε έμπειρους χρήστες, αφού είναι αρκετά εύκολο στην εκμάθηση και ταυτόχρονα πολύ ισχυρό. Τα πλεονεκτήματα χρήσης STATA είναι:

- ✓ Δυνατότητα πραγματοποίησης πολλών ισχυρών λειτουργιών ταυτόχρονα.
- ✓ Δυνατότητα χειρισμού τεράστιου όγκου δεδομένων.
- ✓ Ευκολία υλοποίησης παλινδρόμησης και λογιστικής παλινδρόμησης.

Excel

Αποτελεί μέρος του πακέτου Office της εταιρίας Microsoft. Προσφέρει ένα μεγάλο αριθμό δυνατοτήτων για τη στατιστική ανάλυση δεδομένων. Συνοψίζει τα δεδομένα με προεπισκοπήσεις και προσφέρει ποικίλους τρόπους για την παρουσίασή τους.

1.6 Γενικά συμπεράσματα για την συμβολή της Στατιστικής στην Ιατρική

Η Στατιστική αποτελεί αναπόσπαστο μέρος της ιατρικής έρευνας. Χρησιμοποιείται από ερευνητές για τη διεξαγωγή δοκιμών για τον προσδιορισμό των αποτελεσμάτων από πειράματα, κλινικές δοκιμές και συμπτώματα ασθενειών. Η εφαρμογή στατιστικών στον ιατρικό κλάδο δίνει τη δυνατότητα στο κοινό να κατανοήσει καλύτερα τους κινδύνους που ενέχουν οι ασθένειες, καθώς και συμπεριφορές που μπορούν να προκαλέσουν καρδιοπάθειες και καρκίνο (βλ. για παράδειγμα, [9], [12], [18], [22], [27-28]).

Διαφορετικοί κλάδοι της Στατιστικής βοηθούν στην ενημέρωση των τομέων της ιατρικής. Ενδεικτικά αναφέρουμε:

- Η ποσοτική έρευνα καθοδηγεί τους υπεύθυνους λήψης αποφάσεων για την υγειονομική περίθαλψη που χρησιμοποιούν αριθμητικά δεδομένα που συλλέγονται από μετρήσεις που περιγράφουν τα χαρακτηριστικά συγκεκριμένων δειγμάτων πληθυσμού.
- Η περιγραφική στατιστική είναι ένας άλλος κλάδος της στατιστικής που συνοψίζει τη χρησιμότητα, την αποτελεσματικότητα και το κόστος των ιατρικών αγαθών και υπηρεσιών.
- Ένας μεγάλος αριθμός οργανισμών υγειονομικής περίθαλψης χρησιμοποιούν στατιστικές αναλύσεις για να μετρήσουν τα επίπεδα απόδοσής τους. Εφαρμόζουν προγράμματα βελτίωσης της ποιότητας που βασίζονται σε δεδομένα βάσει αυτής της ανάλυσης.
- Οι στατιστικές συμπερασμάτων χρησιμοποιούνται για τον εντοπισμό των αιτιών των ασθενειών. οι επαγγελματίες στις φαρμακευτικές, ιατροδικαστικές και βιολογικές επιστήμες βασίζονται σε μεγάλο βαθμό σε στατιστικές για πληροφορίες σχετικά με την υγεία και την ιατρική.

Ας ρίξουμε μια εστιασμένη άποψη σε ορισμένους άλλους σημαντικούς τρόπους χρήσης των στατιστικών στον ιατρικό τομέα:

1. Συλλογή δεδομένων. Τα δεδομένα για δείγματα ανθρώπινου πληθυσμού μπορούν να συγκεντρωθούν με οργανωμένο και αποτελεσματικό τρόπο χρησιμοποιώντας επιστημονικές μεθόδους. Χρησιμοποιώντας αυτές τις πληροφορίες, ο κλάδος της υγειονομικής περίθαλψης αποκτά μια εικόνα για τα χαρακτηριστικά της καταναλωτικής αγοράς, όπως η ηλικία, το φύλο, η φυλή, το εισόδημα και οι αναπηρίες. Αυτά τα δημογραφικά στοιχεία μπορούν να χρησιμοποιηθούν για να προβλέψουν τους τύπους υπηρεσιών που χρησιμοποιούν οι άνθρωποι και το επίπεδο φροντίδας που μπορούν να αντέξουν οικονομικά.

2. Κατανομή πόρων. Οι ιατρικοί πόροι είναι σπάνιοι και ακριβοί, ειδικά σε μια αναπτυσσόμενη χώρα όπως η δική μας. οι στατιστικές διαδραματίζουν βασικό ρόλο στη συνετή κατανομή των ιατρικών πόρων. Χρησιμοποιώντας στατιστικά εργαλεία είναι δυνατό να προσδιοριστεί ποιος συνδυασμός αγαθών και υπηρεσιών πρέπει να παραχθεί και ποιοι πόροι θα πρέπει να διατεθούν για την παραγωγή τους. Οι στατιστικές υγειονομικής περίθαλψης επηρεάζουν επίσης την αποδοτικότητα της παραγωγής. Οι αποφάσεις κατανομής θα περιλαμβάνουν πάντα συμβιβασμούς, οι οποίοι απαιτούν τον υπολογισμό του κόστους των χαμένων ευκαιριών. Με αξιόπιστες στατιστικές πληροφορίες ο κίνδυνος αυτών των αντισταθμίσεων ελαχιστοποιείται.

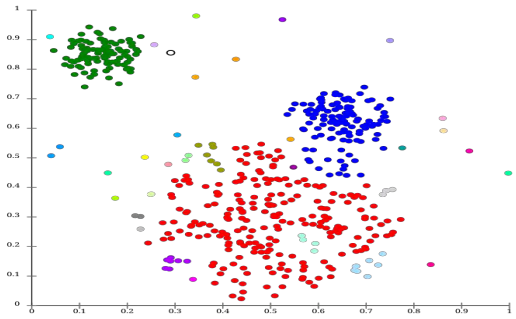
3. Βελτίωση Ποιότητας. Οι πάροχοι υγειονομικής περίθαλψης εργάζονται πάντα για να παράγουν αγαθά και υπηρεσίες με τρόπο που είναι αποτελεσματικός. οι εταιρείες υγειονομικής περίθαλψης χρησιμοποιούν στατιστικά στοιχεία για να μετρήσουν την επιτυχία ή την αποτυχία της απόδοσης. Με πρόσβαση σε τέτοια δεδομένα, οι εταιρείες μπορούν να καθιερώσουν δείκτες αναφοράς ή πρότυπα αριστείας. Αυτό προωθεί τις πρωτοβουλίες βελτίωσης της ποιότητας και βοηθά στην μέτρηση των μελλοντικών αποτελεσμάτων. οι αναλυτές χαρτογραφούν τη συνολική ανάπτυξη και βιωσιμότητα μιας εταιρείας υγειονομικής περίθαλψης χρησιμοποιώντας στατιστικά δεδομένα που συγκεντρώθηκαν με την πάροδο του χρόνου.

ΚΕΦΑΛΑΙΟ 2ο
Η μέθοδος clustering στο πρόγραμμα SPSS

Στο κεφάλαιο αυτό παρουσιάζεται η μέθοδος clustering (ομαδοποίηση), τύποι μεθόδων ομαδοποίησης ενώ γίνεται ιδιαίτερη μνεία σε ποικίλες εφαρμογές ομαδοποίησης σε διάφορους κλάδους με σημαντικό αντίκτυπο στις επιταγές της σύγχρονης κοινωνίας καθώς και στην Ιατρική Επιστήμη. Επίσης, παρουσιάζεται η μέθοδος ομαδοποίησης μέσα από το Στατιστικό Πακέτο SPSS.

2.1 Περιγραφή της μεθόδου Clustering

Η ομαδοποίηση ή η ανάλυση συμπλέγματος είναι η μέθοδος ομαδοποίησης των οντοτήτων με βάση ομοιότητες. Αποκαλύπτει υποομάδες στα διαθέσιμα ετερογενή σύνολα δεδομένων, έτσι ώστε κάθε μεμονωμένο σύμπλεγμα να έχει μεγαλύτερη ομοιογένεια από το σύνολο. Με πιο απλά λόγια, αυτά τα συμπλέγματα είναι ομάδες όμοιων αντικειμένων που διαφέρουν από τα αντικείμενα σε άλλα συμπλέγματα. Στην ομαδοποίηση, το μηχάνημα μαθαίνει μόνο του τα χαρακτηριστικά και τις τάσεις χωρίς καμία παρεχόμενη αντιστοίχιση εισόδου-εξόδου. οι αλγόριθμοι ομαδοποίησης εξάγουν μοτίβα και συμπεράσματα από τον τύπο των αντικειμένων δεδομένων και στη συνέχεια δημιουργούν διακριτές κατηγορίες ομαδοποίησης τους κατάλληλα (βλ. Εικόνα 2.1).



Εικόνα 2.1

2.1.1 Εφαρμογές Cluster Analysis

Η ανάλυση συστάδων εφαρμόζεται σε διάφορα πεδία για να αποκαλύψει διακριτές ομάδες με βάση τις ομοιότητες στα δεδομένα. Ακολουθούν μερικά παραδείγματα:

Ιατρική: Χρησιμοποιείται για την αναγνώριση διαγνωστικών συστάδων αναλύοντας τα συμπτώματα των ασθενών σε ομάδες παρόμοιων περιπτώσεων. Για παράδειγμα, στην ψυχολογία, η ανάλυση συστάδων μπορεί να ταξινομήσει ασθενείς με παρόμοια πρότυπα συμπεριφοράς, βοηθώντας σε στοχευμένα σχέδια θεραπείας.

Μάρκετινγκ: Βοηθά στην κατάτμηση της αγοράς αναλύοντας δεδομένα καταναλωτών όπως ανάγκες, στάσεις, δημογραφικά στοιχεία και συμπεριφορές. Αυτό μπορεί να ενημερώσει για στοχευμένες στρατηγικές μάρκετινγκ εντοπίζοντας τμήματα πελατών με παρόμοια χαρακτηριστικά.

Εκπαίδευση: Προσδιορίζει ομάδες μαθητών που απαιτούν συγκεκριμένες εκπαιδευτικές παρεμβάσεις αναλύοντας δεδομένα ψυχολογίας, ικανοτήτων και επιδόσεων. Αυτό μπορεί να βοηθήσει στην ανάπτυξη εξατομικευμένων εκπαιδευτικών προγραμμάτων και υποστήριξης.

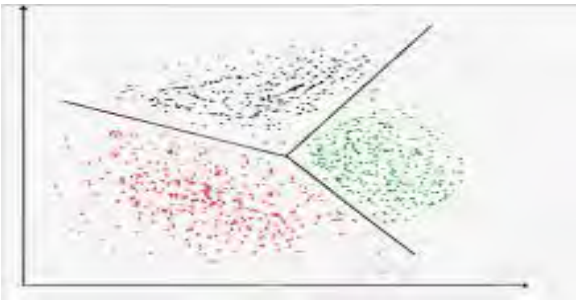
Βιολογία: Βοηθά στη δημιουργία ταξινομιών ομαδοποιώντας είδη με βάση φαινοτυπικά χαρακτηριστικά. Αυτό μπορεί να οδηγήσει σε καλύτερη κατανόηση της βιοποικιλότητας και των σχέσεων του οικοσυστήματος.

2.2 Τύποι μεθόδων ομαδοποίησης

Οι μέθοδοι ομαδοποίησης χωρίζονται ευρέως σε Hard clustering (σημείο δεδομένων που ανήκει μόνο σε μία ομάδα) και Soft Clustering (τα σημεία δεδομένων μπορούν επίσης να ανήκουν σε άλλη ομάδα). Υπάρχουν όμως και άλλες διάφορες προσεγγίσεις ομαδοποίησης. Ακολουθούν οι κύριες μέθοδοι ομαδοποίησης που χρησιμοποιούνται στη Μηχανική μάθηση:

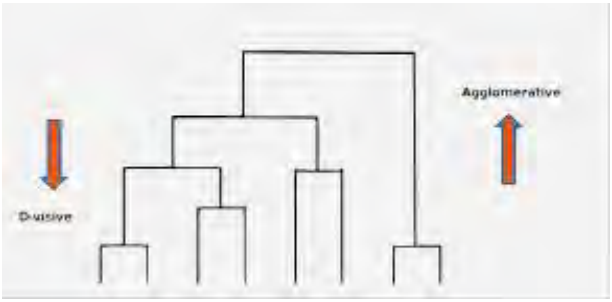
- Μερική ομαδοποίηση (Partitional clustering).
- Ιεραρχική ομαδοποίηση (Hierarchical Clustering).
- ομαδοποίηση με βάση την πυκνότητα (Density-based clustering).
- Ασαφής ομαδοποίηση (Fuzzy clustering).

Μερική ομαδοποίηση (Partitional clustering): Αυτή η μέθοδος ομαδοποίησης χωρίζει τις πληροφορίες σε πολλαπλές ομάδες με βάση τα χαρακτηριστικά και τις ομοιότητες των δεδομένων (βλ. Εικόνα 2.2). Δεδομένου ενός συνόλου δεδομένων N σημείων, μια μέθοδος καταμερισμού κατασκευάζει K ($N \geq K$) κατατμήσεις των δεδομένων, με κάθε διαμέρισμα να αντιπροσωπεύει ένα σύμπλεγμα. Το K πρέπει να επιλέγεται χειροκίνητα σύμφωνα με την ειδικότητα των δεδομένων και την περιοχή τομέα. ο πιο δημοφιλής αλγόριθμος τμηματικής ομαδοποίησης είναι το K-means.



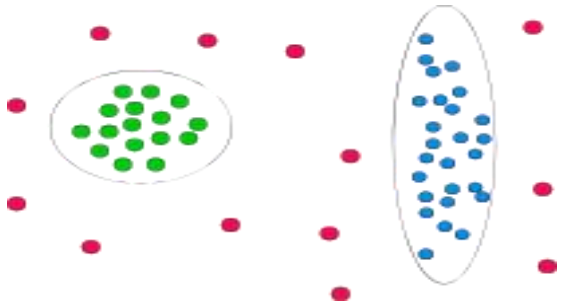
Εικόνα 2.2

Ιεραρχική ομαδοποίησηη (Hierarchical Clustering): Σε αυτόν τον τύπο ομαδοποίησης, ο στόχος είναι να δημιουργηθεί μια δομή που μοιάζει με δέντρο από ένθετα συμπλέγματα, όπου κάθε σύμπλεγμα μπορεί να περιέχει μεμονωμένα σημεία δεδομένων ή άλλα συμπλέγματα (βλ. Εικόνα 2.3). Χρησιμοποιώντας αυτή τη δομή που μοιάζει με δέντρο, μπορούμε να καταλάβουμε με ποια σειρά ακριβώς τα σημεία συγχωνεύονται. Η ιεραρχική ομαδοποίηση μπορεί περαιτέρω να χωριστεί σε δύο υποτύπους: αθροιστική ομαδοποίηση και διαιρετική ομαδοποίηση.



Εικόνα 2.3

Ομαδοποίηση με βάση την πυκνότητα (Density-based clustering): Αυτός ο τύπος ομαδοποίησης προσδιορίζει συστάδες με βάση την πυκνότητα των σημείων δεδομένων στον χώρο χαρακτηριστικών (βλ. Εικόνα 2.4). ο στόχος της ομαδοποίησης με βάση την πυκνότητα είναι να βρεθούν περιοχές υψηλής πυκνότητας που χωρίζονται από περιοχές χαμηλής πυκνότητας. οι πιο δημοφιλείς αλγόριθμοι ομαδοποίησης με βάση την πυκνότητα είναι οι DBSCAN και Mean-shift.



Εικόνα 2.4

Ασαφής ομαδοποίηση (Fuzzy clustering): Η Fuzzy Clustering είναι ένας τύπος αλγορίθμου ομαδοποίησης στη μηχανική εκμάθηση που επιτρέπει σε ένα σημείο δεδομένων να ανήκει σε περισσότερα από ένα σύμπλεγμα με διαφορετικούς βαθμούς συμμετοχής. Σε αντίθεση με τους παραδοσιακούς αλγόριθμους ομαδοποίησης, όπως η k-means ή η ιεραρχική ομαδοποίηση, που εκχωρούν κάθε σημείο δεδομένων σε ένα μόνο σύμπλεγμα, η ασαφής ομαδοποίηση εκχωρεί έναν βαθμό μέλους μεταξύ 0 και 1 για κάθε σημείο δεδομένων για κάθε σύμπλεγμα.

- **Εφαρμογές σε διάφορους τομείς της ασαφούς ομαδοποίησης**

Τμηματοποίηση εικόνας: Η ασαφής ομαδοποίηση μπορεί να χρησιμοποιηθεί για την τμηματοποίηση εικόνων ομαδοποιώντας pixel με παρόμοιες ιδιότητες μεταξύ τους, όπως χρώμα ή υφή.

Αναγνώριση προτύπων: Η ασαφής ομαδοποίηση μπορεί να χρησιμοποιηθεί για τον εντοπισμό μοτίβων σε μεγάλα σύνολα δεδομένων ομαδοποιώντας παρόμοια σημεία δεδομένων μαζί.

Μάρκετινγκ: Η ασαφής ομαδοποίηση μπορεί να χρησιμοποιηθεί για την τμηματοποίηση των πελατών με βάση τις προτιμήσεις και την αγοραστική τους συμπεριφορά, επιτρέποντας πιο στοχευμένες καμπάνιες μάρκετινγκ.

Ιατρική διάγνωση: Η ασαφής ομαδοποίηση μπορεί να χρησιμοποιηθεί για τη διάγνωση ασθενειών ομαδοποιώντας ασθενείς με παρόμοια συμπτώματα.

Περιβαλλοντική παρακολούθηση: Η ασαφής ομαδοποίηση μπορεί να χρησιμοποιηθεί για τον εντοπισμό περιοχών περιβαλλοντικής ανησυχίας ομαδοποιώντας περιοχές με παρόμοια επίπεδα ρύπανσης ή άλλους περιβαλλοντικούς δείκτες.

Ανάλυση ροής κυκλοφορίας: Η ασαφής ομαδοποίηση μπορεί να χρησιμοποιηθεί για την ανάλυση μοτίβων ροής κυκλοφορίας ομαδοποιώντας παρόμοια μοτίβα κυκλοφορίας, επιτρέποντας καλύτερη διαχείριση και σχεδιασμό της κυκλοφορίας.

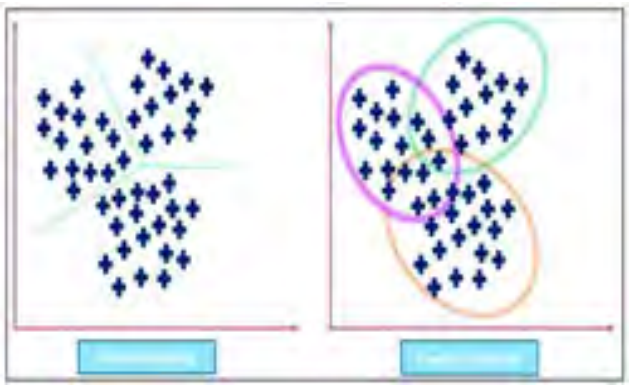
Εκτίμηση κινδύνου: Η ασαφής ομαδοποίηση μπορεί να χρησιμοποιηθεί για τον εντοπισμό και την ποσοτικοποίηση των κινδύνων σε διάφορους τομείς, όπως τα χρηματοοικονομικά, οι ασφάλειες και η μηχανική.

➢ Πλεονεκτήματα της ασαφούς ομαδοποίησης

Ευελιξία: Η ασαφής ομαδοποίηση επιτρέπει επικαλυπτόμενες συστάδες, οι οποίες μπορεί να είναι χρήσιμες όταν τα δεδομένα έχουν πολύπλοκη δομή ή όταν υπάρχουν διφορούμενα ή επικαλυπτόμενα όρια κλάσεων.

Ισχυρότητα: Η ασαφής ομαδοποίηση μπορεί να είναι πιο ανθεκτική σε ακραίες τιμές και θόρυβο στα δεδομένα, καθώς επιτρέπει μια πιο σταδιακή μετάβαση από το ένα σύμπλεγμα στο άλλο.

Ερμηνευσιμότητα: Η ασαφής ομαδοποίηση παρέχει μια πιο λεπτή κατανόηση της δομής των δεδομένων, καθώς επιτρέπει μια πιο λεπτομερή αναπαράσταση των σχέσεων μεταξύ σημείων δεδομένων και συστάδων (βλ. Εικόνα 2.5).



Εικόνα 2.5

2.3 Η μέθοδος ομαδοποίησης (clustering) μέσα απο το στατιστικό πακέτο SPSS

Το SPSS προσφέρει τρεις μεθόδους για την ανάλυση συστάδων:

- K-Means Cluster,
- Hierarchical Cluster και
- Two-Step cluster.

Η **K-means cluster** είναι μια μέθοδος γρήγορης ομαδοποίησης μεγάλων συνόλων δεδομένων. ο ερευνητής ορίζει τον αριθμό των συστάδων εκ των προτέρων. Αυτό είναι χρήσιμο για τη δοκιμή διαφορετικών μοντέλων με διαφορετικό υποθετικό αριθμό συστάδων.

Η **ιεραρχική συστάδα (Hierarchical Cluster)** είναι η πιο κοινή μέθοδος. Η ιεραρχική ομαδοποίηση, γνωστή και ως ιεραρχική ανάλυση συστάδων, είναι ένας αλγόριθμος που ομαδοποιεί παρόμοια αντικείμενα σε ομάδες που ονομάζονται συστάδες. Το τελικό σημείο είναι ένα σύνολο συστάδων, όπου κάθε σύμπλεγμα είναι ξεχωριστό από το άλλο σύμπλεγμα και τα αντικείμενα μέσα σε κάθε σύμπλεγμα είναι σε γενικές γραμμές παρόμοια μεταξύ τους.

Η **ανάλυση συστάδων δύο βημάτων (Two-Step cluster)** προσδιορίζει τις ομαδοποιήσεις εκτελώντας πρώτα την προ-ομαδοποίηση και μετά εκτελώντας ιεραρχικές μεθόδους. Επειδή χρησιμοποιεί έναν γρήγορο αλγόριθμο συμπλέγματος εκ των προτέρων, μπορεί να χειριστεί μεγάλα σύνολα δεδομένων που θα χρειαζόταν πολύ χρόνο για να υπολογιστούν με μεθόδους ιεραρχικής συμπλέγματος. Από αυτή την άποψη, είναι ένας συνδυασμός των δύο προηγούμενων προσεγγίσεων. Η ομαδοποίηση δύο βημάτων μπορεί να χειριστεί δεδομένα κλίμακας και τακτικότητας στο ίδιο μοντέλο και επιλέγει αυτόματα τον αριθμό των συστάδων.

Στην παρούσα Διπλωματική Εργασία θα ασχοληθούμε με την περιγραφή των K-Means Cluster και Hierarchical Cluster που είναι αρκετά εύχρηστες στο στατιστικό πακέτο SPSS και δίνουν καλά αποτελέσματα.

2.3.1 Hierarchical Cluster στο Στατιστικό Πακέτο SPSS

Η ιεραρχική ομαδοποίηση (Hierarchical Cluster) είναι μια τεχνική ανάλυσης δεδομένων που έχει σχεδιαστεί για να ταξινομεί τα σημεία δεδομένων σε συστάδες ή ομάδες, με βάση ένα σύνολο παρόμοιων χαρακτηριστικών. Η ιεραρχική ομαδοποίηση λειτουργεί δημιουργώντας ένα «δέντρο» συστάδας, όπου τα συμπλέγματα αρχίζουν μεγαλύτερα και στη συνέχεια διασπώνται σε μικρότερες ομάδες σε κάθε σημείο διακλάδωσης του δέντρου.

Για παράδειγμα, αν έχουμε να κατηγοριοποιήσουμε μια ομάδα ασθενών, τότε το κορυφαίο (μεγαλύτερο) σύμπλεγμα μπορεί να είναι ολόκληρο το σύνολο των ασθενών. Κάτω από αυτό, τα δεδομένα μας ενδέχεται να διακλαδιστούν σε μικρότερες ομάδες ανά κατηγορία ασθενών (ασθενείς με διαβήτη, ασθενείς με καρκίνο κ.λπ.) και στη συνέχεια να διασπαστούν περαιτέρω σε ομάδες ασθενών (ασθενείς με διαφορετικούς τύπους διαβήτη, ασθενείς με διάφορους τύπους καρκίνου κ.λπ.) και καθώς κατεβαίνουμε πιο μακριά κάτω από το δέντρο συστάδας. Η ιεραρχική ομαδοποίηση χρησιμοποιείται συνήθως για ιεραρχικά δεδομένα, δηλαδή δεδομένα που υποκατηγοριοποιούνται εύκολα σε διαφορετικά επίπεδα εξειδίκευσης.

Η ιεραρχική ομαδοποίηση χρησιμοποιείται σε πολλές εφαρμογές. Για παράδειγμα μπορεί να χρησιμοποιηθεί στα ακόλουθα περιβάλλοντα:

Τμηματοποίηση πελατών: Οι οργανισμοί βασίζονται στην ιεραρχική ομαδοποίηση για να ομαδοποιήσουν τους πελάτες με βάση την τοποθεσία, τη συμπεριφορά και τα δημογραφικά στοιχεία. Αυτό βοηθά στην πιο αποτελεσματική στόχευση εκστρατειών μάρκετινγκ.

Ανάλυση εγκλήματος: Οι υπηρεσίες επιβολής του νόμου μπορούν να εφαρμόσουν ιεραρχική ομαδοποίηση για να κατηγοριοποιήσουν διαφορετικούς τύπους εγκλημάτων ή να βρουν πρότυπα εγκληματικής δραστηριότητας. Αυτή η ανάλυση βοηθά στην κατανομή των πόρων και στις προληπτικές στρατηγικές.

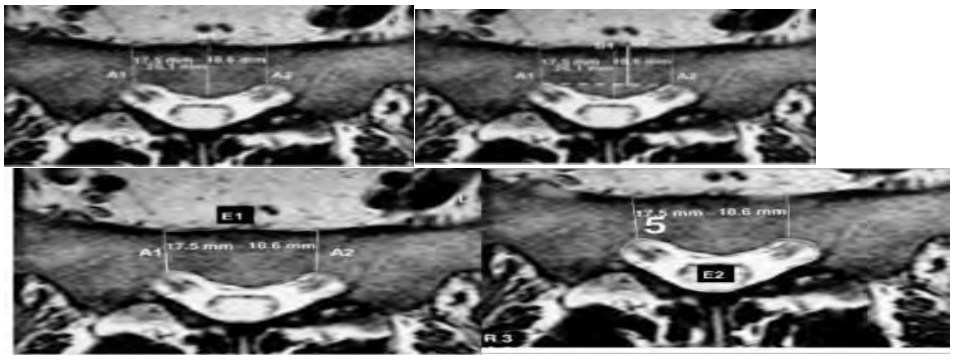
Υγειονομική περίθαλψη και ιατρική διάγνωση: Αυτή η τεχνική μπορεί να ομαδοποιήσει ασθενείς με παρόμοια συμπτώματα ή ιατρικό ιστορικό, βοηθώντας στη διάγνωση και την ποιότητα της περίθαλψης.

Επεξεργασία φυσικής γλώσσας: Η τεχνική αυτή μπορεί να χρησιμοποιηθεί για την ομαδοποίηση λέξεων ή φράσεων με παρόμοια σημασία. Αυτή η τεχνική μηχανικής εκμάθησης και ανάκτησης πληροφοριών αναλύει κείμενο και προσδιορίζει τις τάσεις.

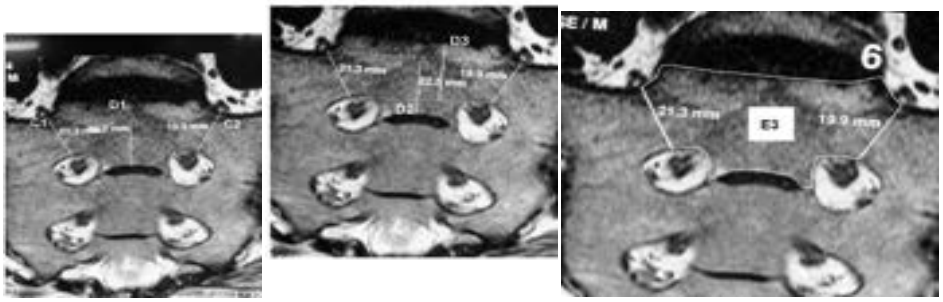
Συστήματα συστάσεων: Μπορούν να βοηθήσουν στην ομαδοποίηση παρόμοιων προϊόντων με βάση τα χαρακτηριστικά των καταναλωτών, ενισχύοντας την ακρίβεια των συστάσεων.

Παρακάτω θα δούμε την μέθοδο Hierarchical Cluster στο Στατιστικό Πακέτο SPSS μέσου απλών παραδειγμάτων.

Παράδειγμα 1. Ελήφθησαν 12 μορφομετρικές μετρήσεις ιερού οστού από 20 άτομα. οι μετρούμενες μεταβλητές είναι: A1, A2, B1, B2, E1, E2 (Επίπεδο-1), G1, G2, D1, D2, D3, E3 (Επίπεδο-2) (βλ. ακτίνες X). Θα προσδιορίσουμε τις ομάδες ατόμων με παρόμοιο μορφολογικό ιερό οστό (βλ. Εικόνα 2.6 και Εικόνα 2.7) κάνοντας χρήση της μεθόδου Hierarchical Cluster.



Εικόνα 2.6: Transverse plane, level 1: A1 - anteroposterior (εμπροοπίσθια) diameter of right pedicle (εγκάρσια απόφυση), A2 - anteroposterior diameter of left pedicle, B1 - anteroposterior diameter of vertebra body, B2 - line between anterior aspect of vertebra body and the level of the lateral recess, E1 - area of vertebra body, E2 - area of spinal canal.

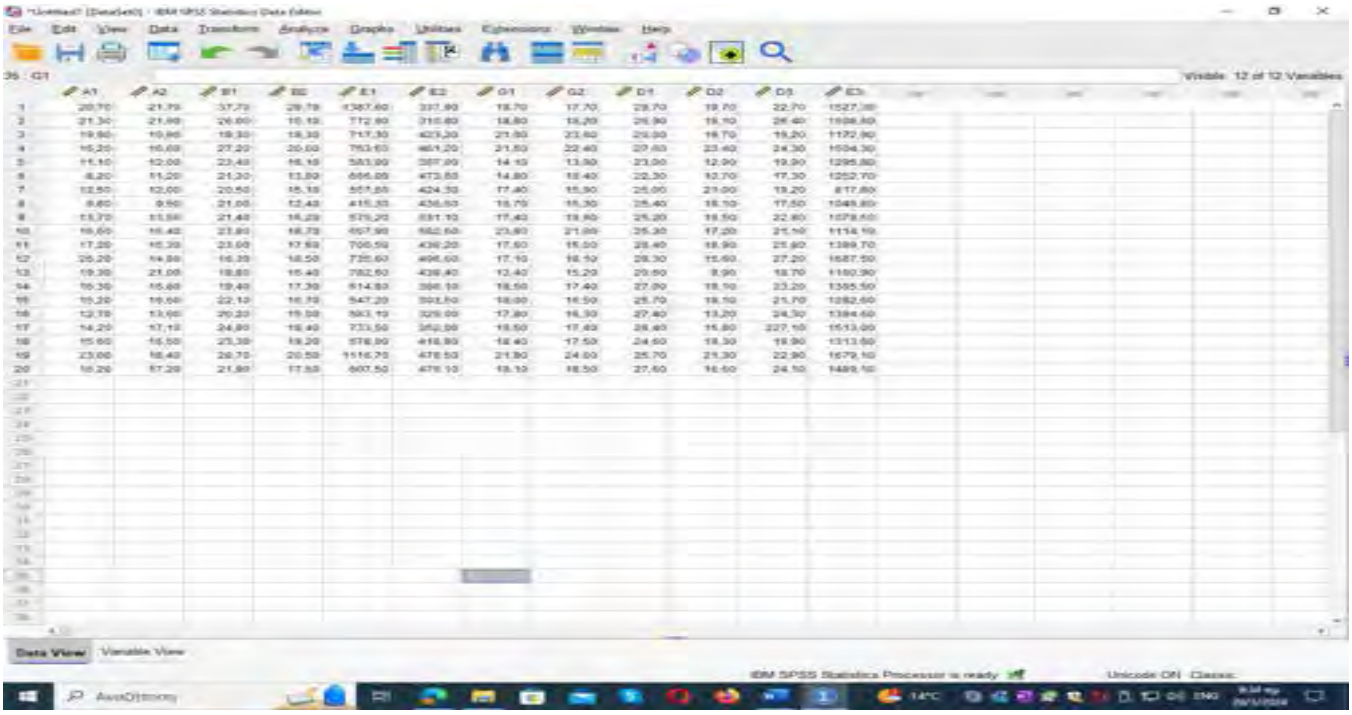


Εικόνα 2.7: Coronal plane, level 2: C1 - superior-inferior diameter of right pedicle (κάθετη απόφυση), C2 - superior-inferior diameter of left pedicle, D1 - superior-inferior diameter of S1 vertebra body, D2 - line between the level of the ala and the inferior aspect of vertebra body, D3 - line between superior aspect of vertebra body and the level of roof of the foramen, and E3 - area of vertebra body (σώμα σπονδύλου).

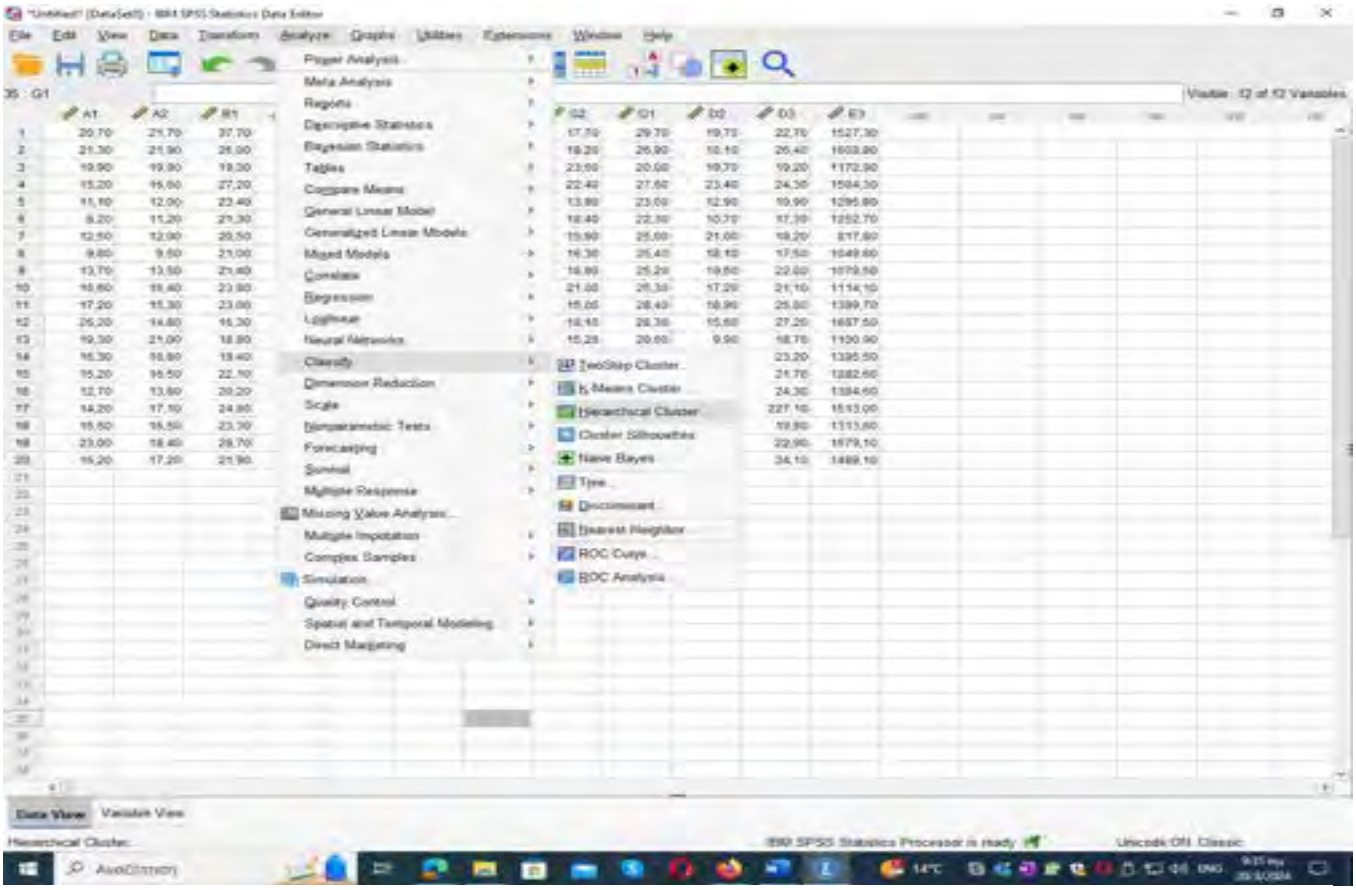
Τα δεδομένα είναι τα παρακάτω:

A/A	A1	A2	B1	B2	E1	E2	G1	G2	D1	D2	D3	E3
1	20,7	21,7	37,7	29,7	1387,6	337,9	19,7	17,7	29,7	19,7	22,7	1527,3
2	21,3	21,9	26,0	15,1	772,9	310,8	18,8	18,2	26,9	18,1	26,4	1608,8
3	19,9	19,9	19,3	19,3	717,3	423,2	21,0	23,6	20,0	19,7	19,2	1172,9
4	15,2	16,6	27,2	20,0	753,5	461,2	21,5	22,4	27,6	23,4	24,3	1504,3
5	11,1	12,0	23,4	16,1	583,0	397,0	14,1	13,9	23,0	12,9	19,9	1295,8
6	8,2	11,2	21,3	13,9	666,0	473,8	14,8	18,4	22,3	10,7	17,3	1252,7
7	12,5	12,0	20,5	15,1	557,5	424,3	17,4	15,9	25,0	21,0	19,2	817,8
8	9,8	9,5	21,0	12,4	415,3	436,6	16,7	16,3	25,4	18,1	17,5	1049,8
9	13,7	13,5	21,4	16,2	575,2	331,1	17,4	18,9	25,2	19,5	22,8	1078,5
10	16,6	16,4	23,8	18,7	657,9	582,6	23,9	21,0	25,3	17,2	21,1	1114,1
11	17,2	15,3	23,0	17,6	700,5	436,2	17,6	15,0	28,4	18,9	25,8	1399,7
12	26,2	14,8	16,3	18,5	735,6	496,6	17,1	18,1	28,3	15,6	27,2	1687,5
13	19,3	21,0	18,8	16,4	782,6	439,4	13,4	15,2	20,6	9,9	18,7	1100,9
14	16,3	16,8	19,4	17,3	614,8	380,1	18,5	17,4	27,0	19,1	23,2	1395,5
15	15,2	16,5	22,1	16,7	547,2	303,5	18,0	16,5	25,7	18,1	21,7	1282,6
16	12,7	13,6	20,2	15,0	583,1	329,0	17,8	16,3	27,4	13,2	24,3	1394,6
17	14,2	17,1	24,8	18,4	733,5	352,0	16,5	17,4	28,4	16,8	227,1	1513,0
18	15,6	16,5	23,3	19,2	578,0	418,8	18,4	17,5	24,6	19,3	19,9	1313,6
19	23,0	18,4	28,7	20,5	1116,7	478,5	21,8	24,0	25,7	21,3	22,9	1679,1
20	16,2	17,2	21,9	17,6	607,5	476,1	18,1	18,5	27,6	16,6	24,1	1489,1

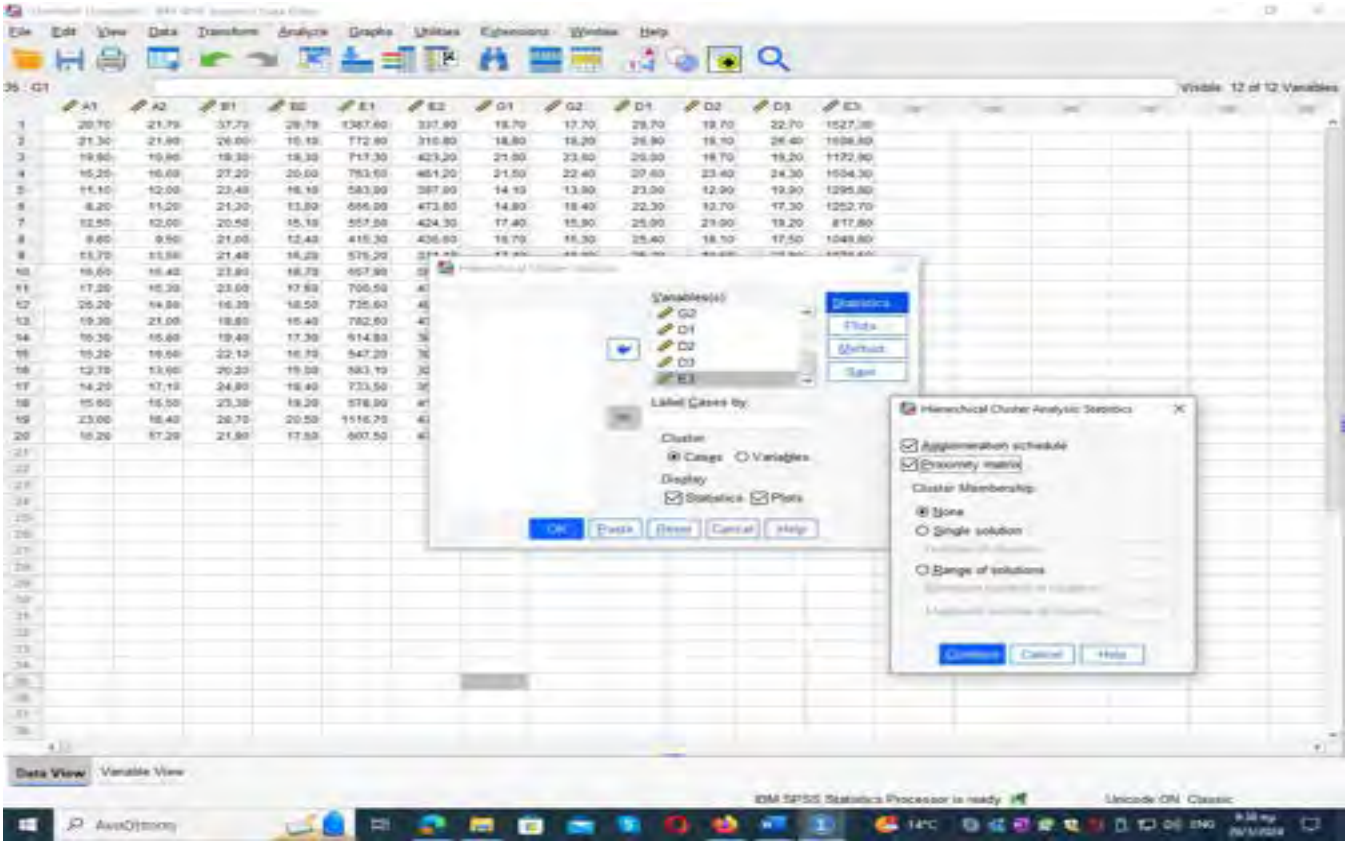
Εισάγοντας τα παραπάνω δεδομένα στο SPSS ακολουθούμε τα παρακάτω βήματα όπως αυτά φαίνονται στις εικόνες που ακολουθούν (Εικόνα 2.8-Εικόνα 2.15).



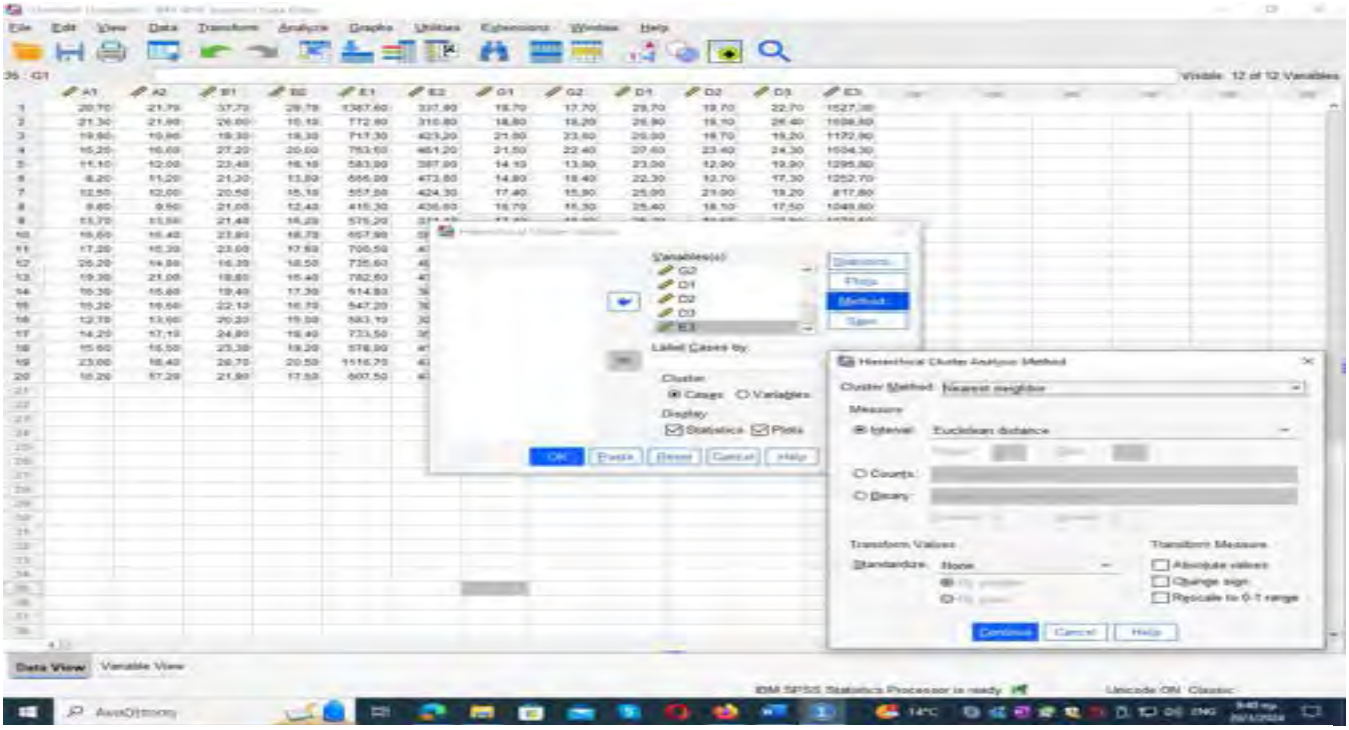
Εικόνα 2.8



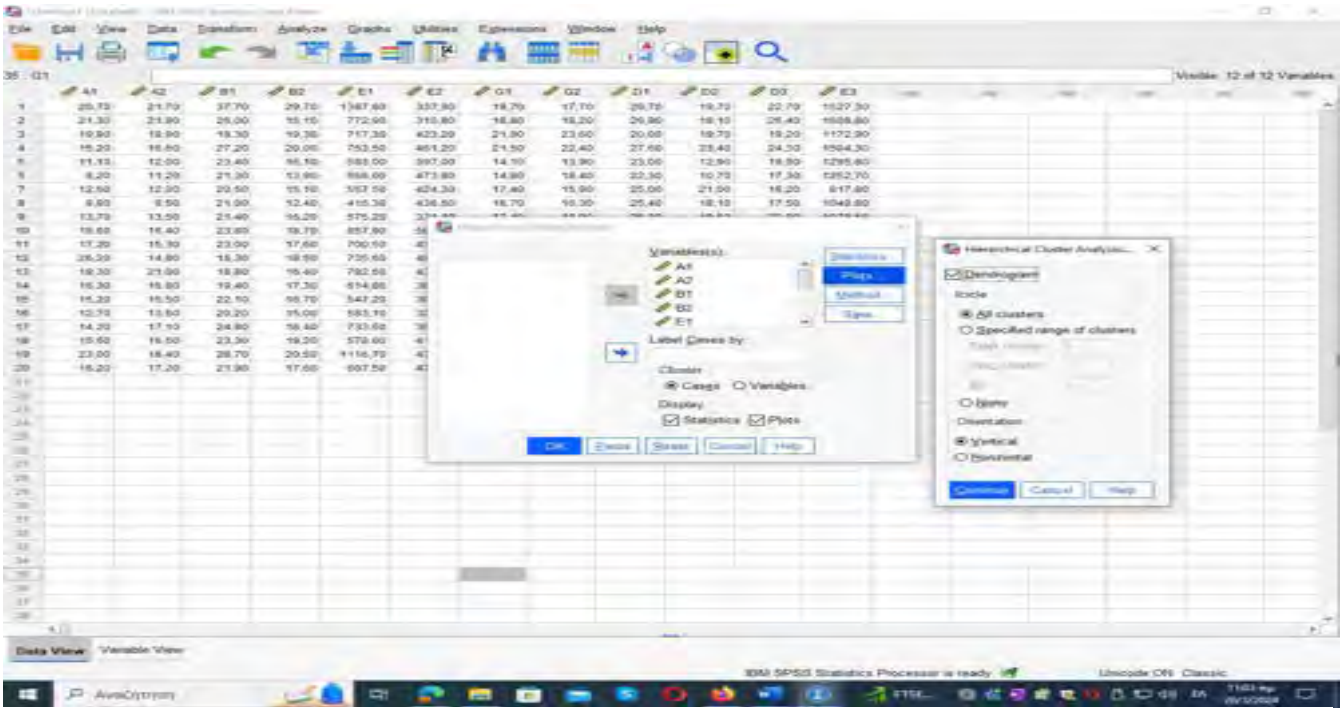
Εικόνα 2.9



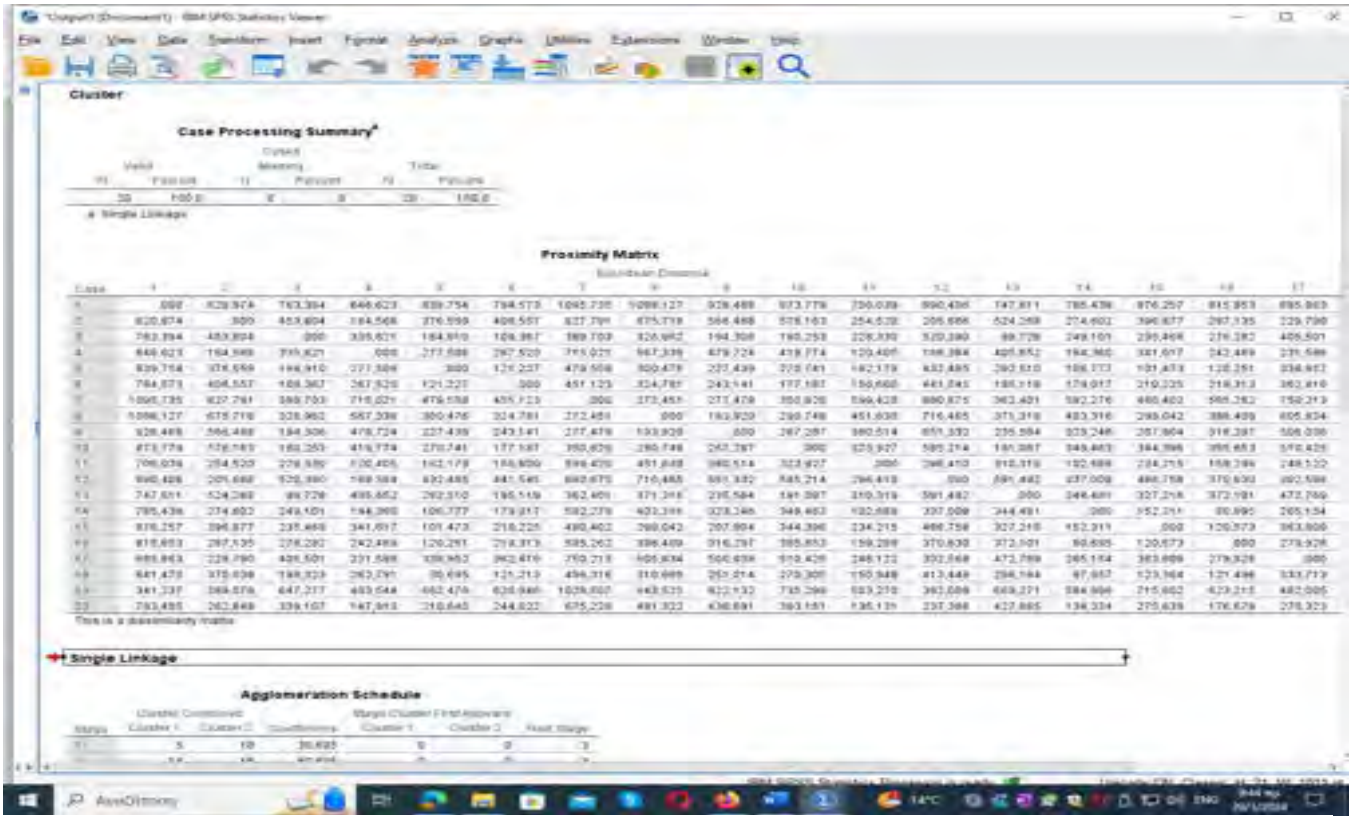
Εικόνα 2.10



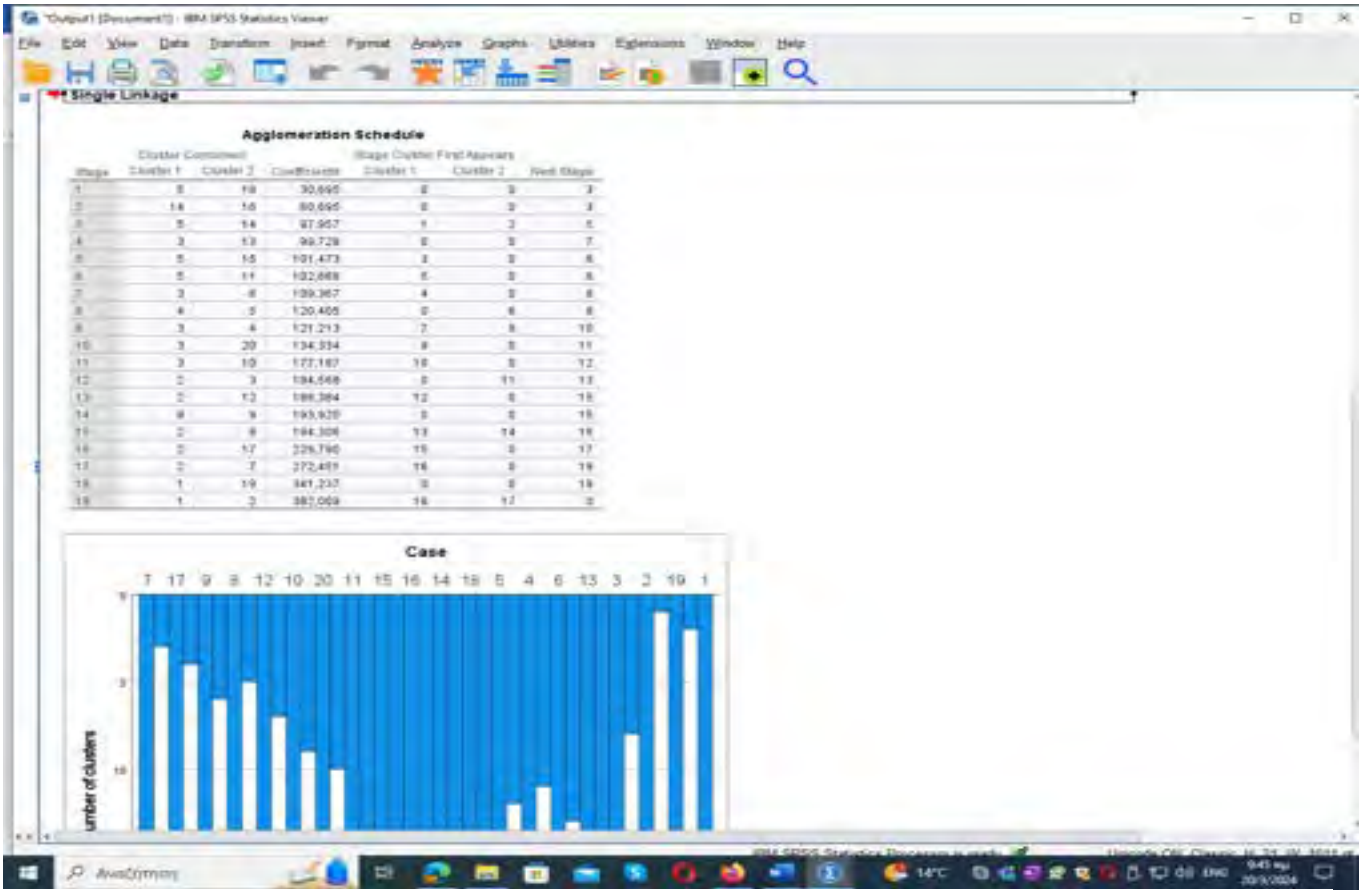
Εικόνα 211



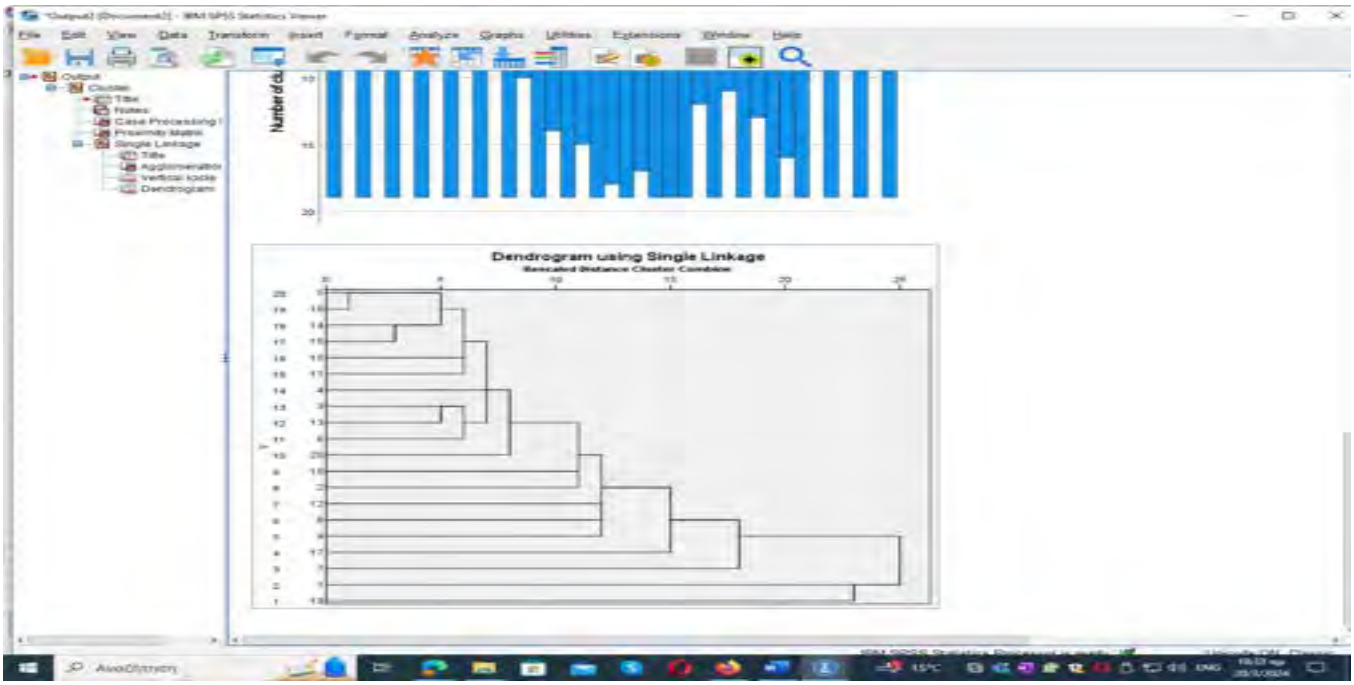
Εικόνα 2.12



Εικόνα 2.13



Εικόνα 2.14



Εικόνα 2.15

Συνεπώς, λαμβάνοντας υπόψη τα παραπάνω αποτελέσματα του «δέντρου» συστάδας, όπου ο διαχωρισμός βασίζεται στο χαρακτηριστικό της απόστασης, διαπιστώνουμε ότι ο διαχωρισμός με το μικρότερο δυνατό σύμπλεγμα ομάδων είναι δυνατόν να οδηγήσει σε καλύτερα συμεπράσματα καθώς η μελέτη επικεντρώνεται σε όσο το δυνατόν μικρότερο αριθμό ομάδων. Ο διαχωρισμός με μεγαλύτερο αριθμό ομάδων είναι δυνατόν να οδηγήσει σε εσφαλμένα συμπεράσματα καθώς ο ερευνητής έρχεται αντιμέτωπος με περισσότερα χαρακτηριστικά ατόμων προκειμένου να συμπεράνει τη ομοιότητα ή μη αυτών.

- Η μικρότερη απόσταση μεταξύ των ασθενών είναι 30,695, μεταξύ του 5 και 18. Οι ασθενείς 5 και 18 δημιουργούν μια ομάδα. Αρα στην απόσταση των 30,695 έχουμε 19 ομάδες:
  (5,18), (1), (2), (3), (4), (6), (7), (8), (9), (10), (11),
  (12), (13), (14), (15), (16), (17), (19), (20).

- Η επόμενη μικρότερη απόσταση μεταξύ των ασθενών είναι 60,695, μεταξύ του 14 και 16. Οι ασθενείς 14 και 16 δημιουργούν μια ομάδα. Αρα στην απόσταση των 60,695 έχουμε 18 ομάδες:
  (5,18), (14, 16), (1), (2), (3), (4), (6), (7), (8), (9), (10),
  (11), (12), (13), (15), (17), (19), (20).

- Η επόμενη μικρότερη απόσταση μεταξύ των ασθενών είναι 97,957, μεταξύ του 5 και 4. Αρα στην απόσταση των 97,957 έχουμε 17 ομάδες:
  (5,18,14, 16), (1), (2), (3), (4), (6), (7), (8), (9), (10),
  (11), (12), (13), (15), (17), (19), (20).

- Η επόμενη μικρότερη απόσταση μεταξύ των ασθενών είναι 99,728, μεταξύ του 3 και 13. Οι ασθενείς 3 και 13 δημιουργούν μια ομάδα. Αρα στην απόσταση των 99,728 έχουμε 16 ομάδες:
  (5,18, 14, 16), (3, 13), (1), (2), (4), (6), (7), (8), (9),
  (10), (11), (12), (15), (17), (19), (20).

- Η επόμενη μικρότερη απόσταση μεταξύ των ασθενών είναι 101,473, μεταξύ του 5 και 15. Αρα στην απόσταση των 101,473 έχουμε 15 ομάδες:
  (5,18, 14, 15,16), (3, 13), (1), (2), (4), (6), (7), (8), (9), (10), (11), (12), (17), (19), (20).

- Η επόμενη μικρότερη απόσταση μεταξύ των ασθενών είναι 102,668, μεταξύ του 5 και 11. Αρα στην απόσταση των 102,668 έχουμε 14 ομάδες:
  (5,18, 14, 15,16,11), (3, 13), (1), (2), (4), (6), (7), (8), (9), (10), (12), (17), (19), (20).

- Η επόμενη μικρότερη απόσταση μεταξύ των ασθενών είναι 109,367, μεταξύ του 3 και 6. Αρα στην απόσταση των 109,367 έχουμε 13 ομάδες:
  (5,18, 14, 15,16,11), (3, 6,13), (1), (2), (4), (7), (8), (9), (10), (12), (17), (19), (20).

- Η επόμενη μικρότερη απόσταση μεταξύ των ασθενών είναι 120,405, μεταξύ του 4 και 5. Αρα στην απόσταση των 120,405 έχουμε 12 ομάδες:
  (4, 5, 18, 14, 15,16,11), (3, 6,13), (1), (2), (7), (8), (9), (10), (12), (17), (19), (20).

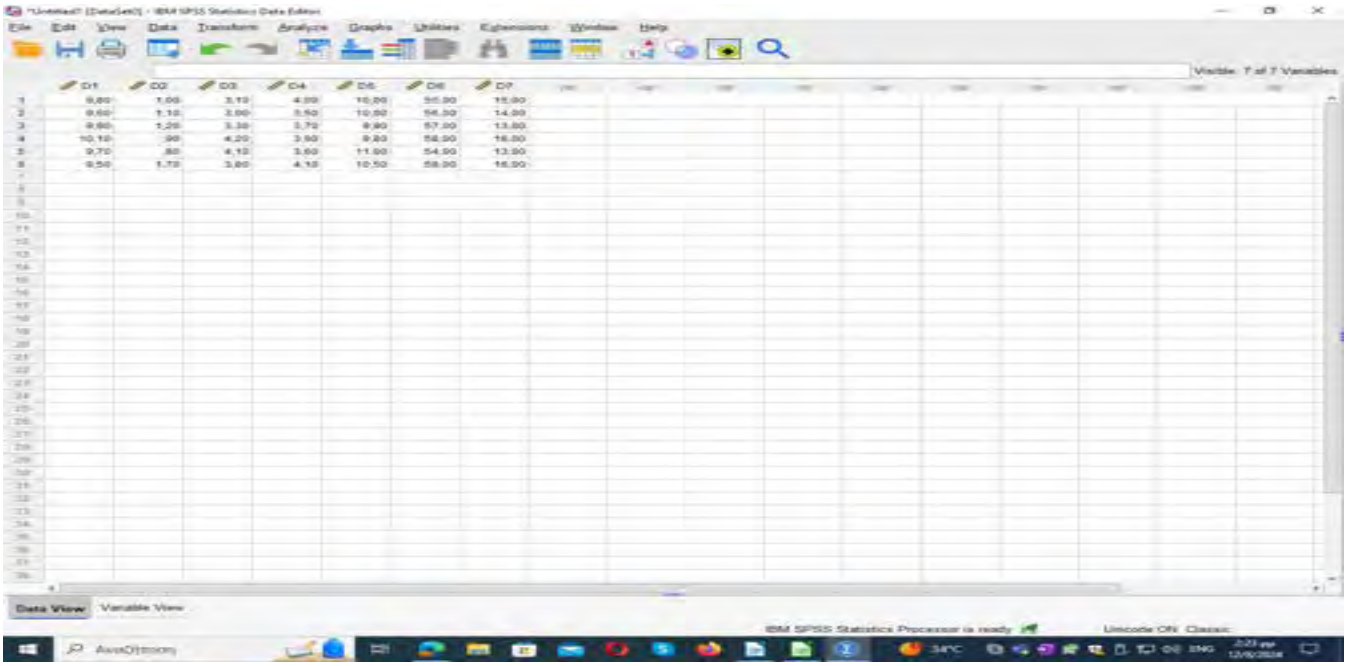
- Συνεχίζοντας την παραπάνω διαδικασία στην απόσταση 121,213 έχουμε τις

παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13), (1), (2), (7), (8), (9), (10), (12), (17), (19), (20).
■ Στην απόσταση 134,334 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6,13,20), (1), (2), (7), (8), (9), (10), (12), (17), (19).
■ Στην απόσταση 177,187 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10), (1), (2), (7), (8), (9), (12), (17), (19).
■ Στην απόσταση 184,568 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10,2), (1), (7), (8), (9), (12), (17), (19).
■ Στην απόσταση 188,384 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10, 2, 12), (1), (7), (8), (9), (17), (19).
■ Στην απόσταση 193,90 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10, 2, 12), (8, 9), (1), (7), (17), (19).
■ Στην απόσταση 194,306 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10, 2, 12, 8, 9), (1), (7), (17), (19).
■ Στην απόσταση 229,790 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10, 2, 12, 8, 9,17), (1), (7), (19).
■ Στην απόσταση 272,451 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10, 2, 12, 8, 9,17,7), (1), (19).
■ Στην απόσταση 341,237 έχουμε τις παρακάτω ομάδες: (4, 5, 18, 14, 15,16,11, 3, 6, 13, 20, 10, 2, 12, 8, 9,17,7), (1, 19).

Παράδειγμα 2. Στον παρακάτω πίνακα δίνονται οι τιμές 7 δεικτών που χαρακτηρίζουν το σχήμα των κρανίων έξι προϊστορικών πληθυσμών. Με βάση αυτόν τον πίνακα θα ελέγξουμε πιθανές συγγένειες μεταξύ των πληθυσμών.

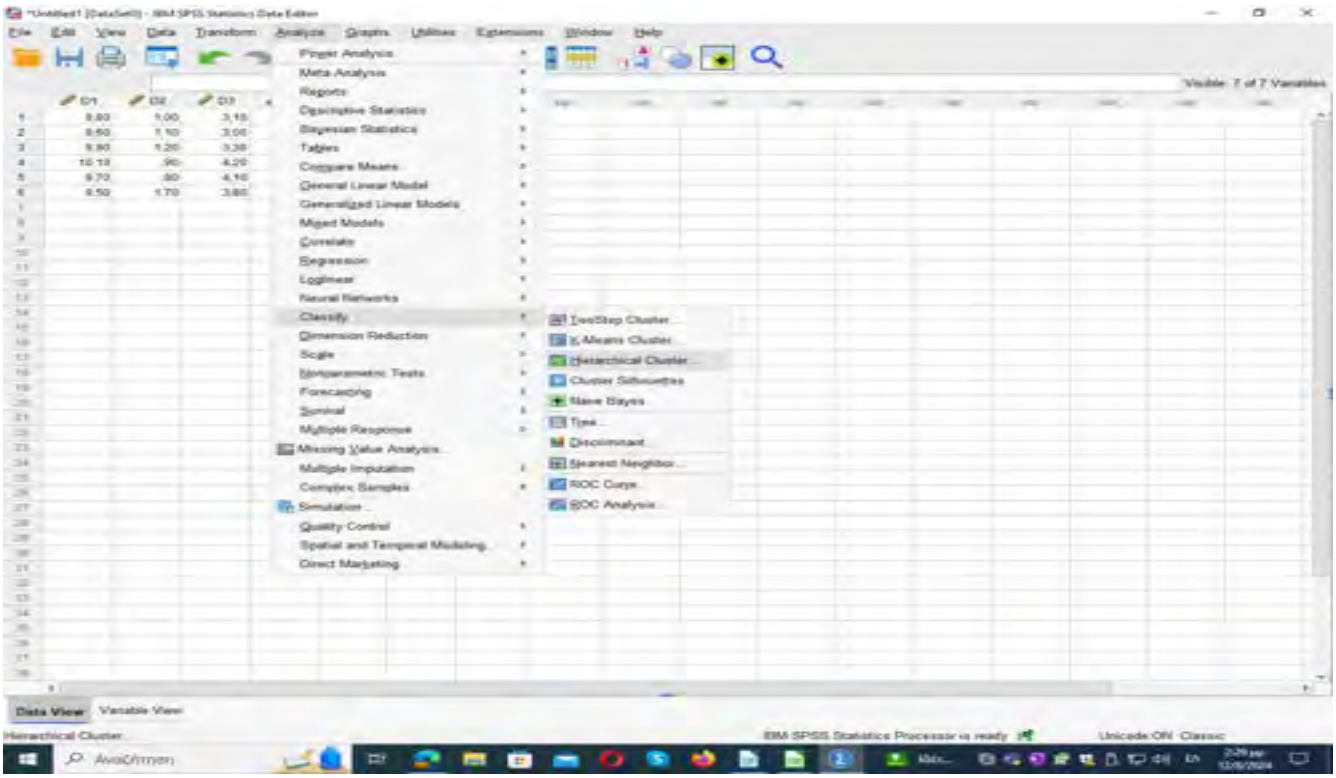
	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**
I	9.8	1	3.1	4	10	55	15
II	9.6	1.1	3.0	3.5	10	56	14
III	9.9	1.2	3.3	3.7	9.9	57	13
IV	10.1	0.9	4.2	3.9	9.8	58	16
V	9.7	0.8	4.1	3.6	11	54	13
VI	9.5	1.7	3.8	4.1	10.5	59	16

Το πρώτο βήμα που κάνουμε στο Στατιστικό Πακέτο SPSS είναι να εισάγουμε τα δεδομένα στο πρόγραμμα όπως φαίνεται παρακάτω (Εικόνα 2.16):



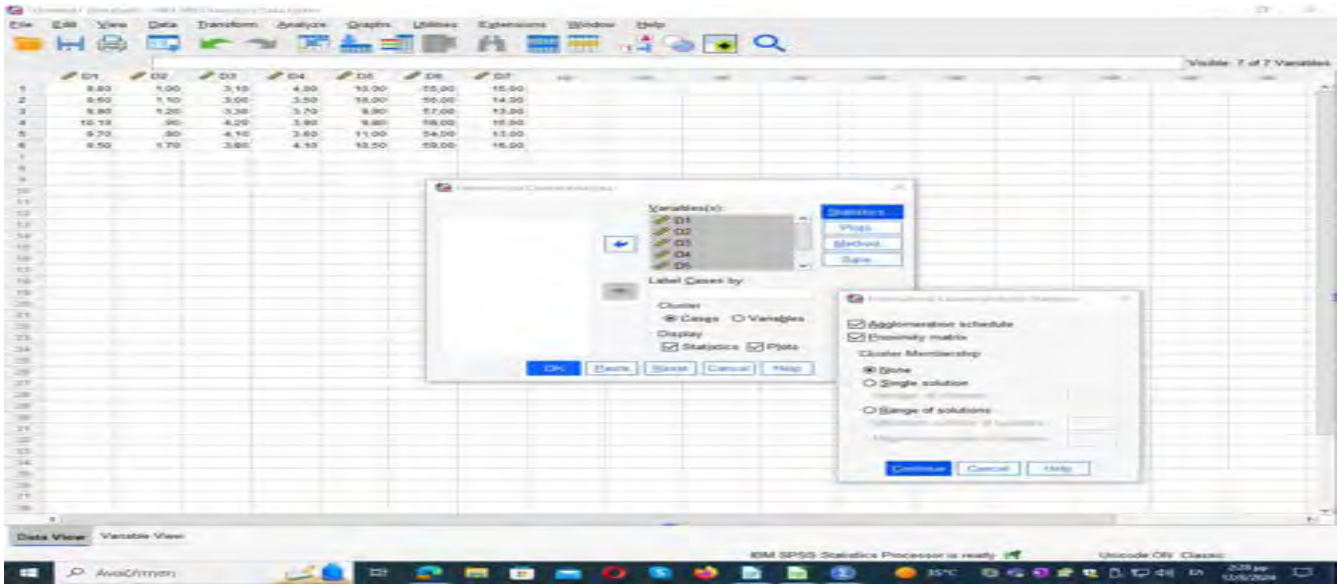
Εικόνα 2.16

Το δεύτερο βήμα που κάνουμε στο Στατιστικό Πακέτο SPSS είναι να επιλέξουμε: **Analyze → Classify → Hierarchical Cluster**. οπότε έχουμε την παρακάτω Εικόνα 2.17:



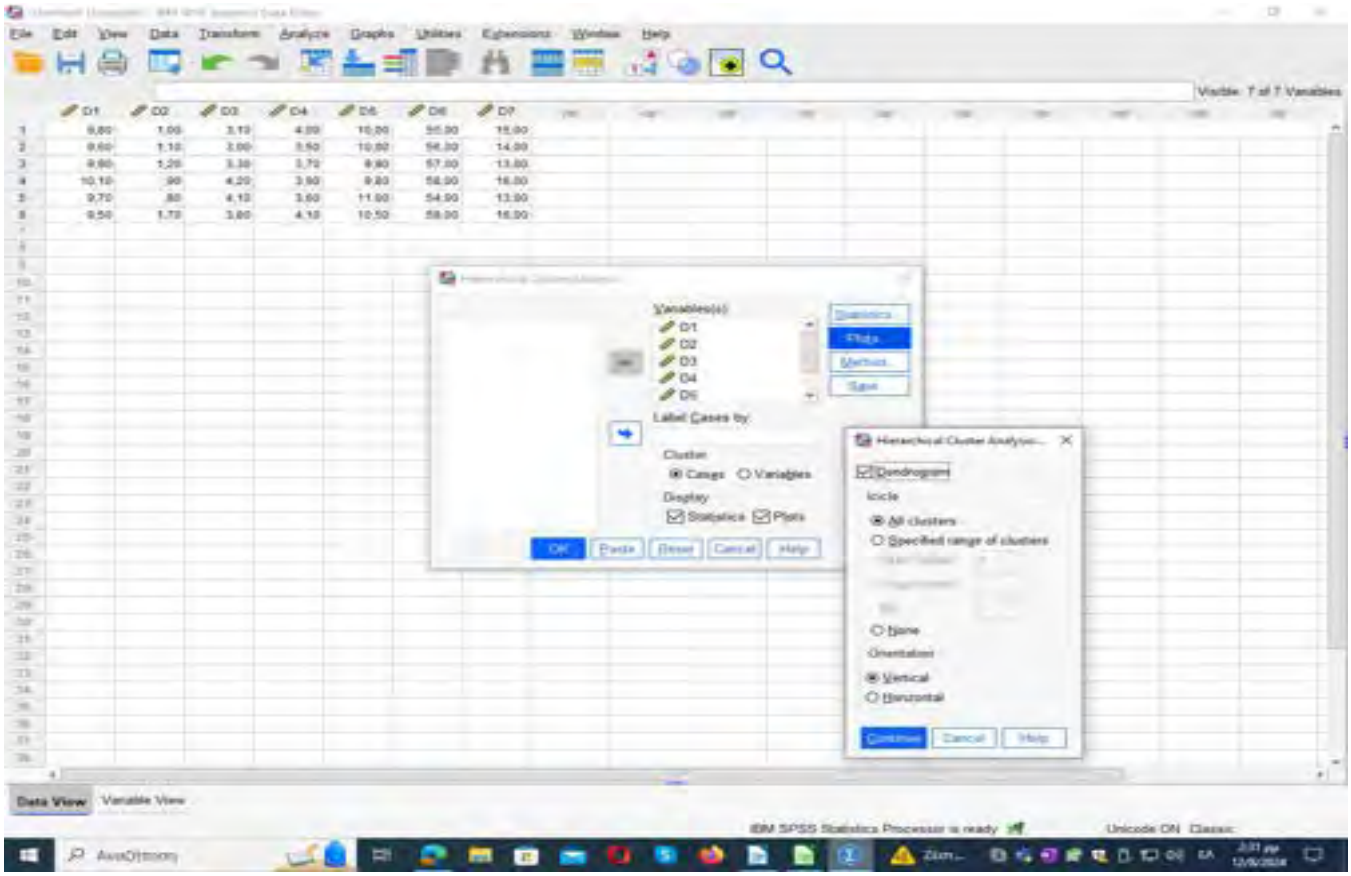
Εικόνα 2.17

Στη συνέχεια ως τρίτο βήμα κάνουμε τις παρακάτω ρυθμίσεις όπως στις παρακάτω εικόνες (Εικόνα 2.18-Εικόνα 2.22).

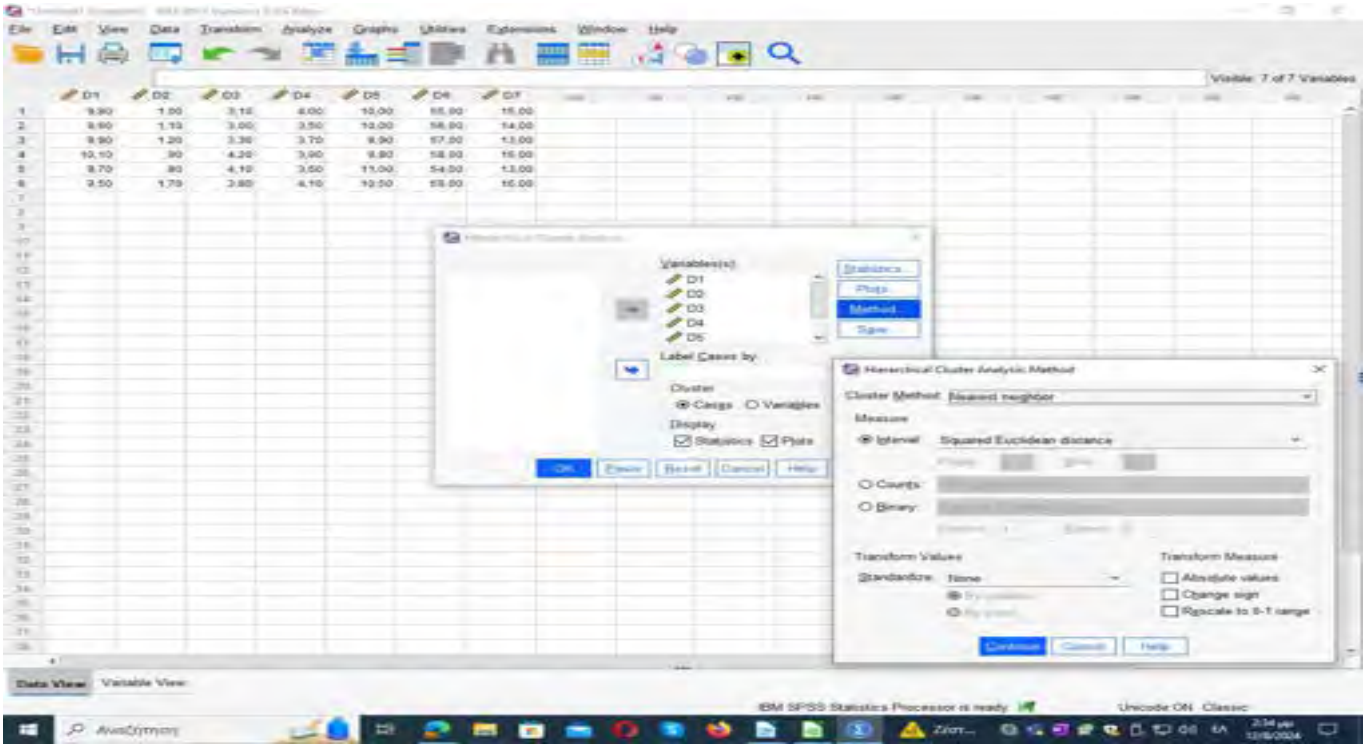


:

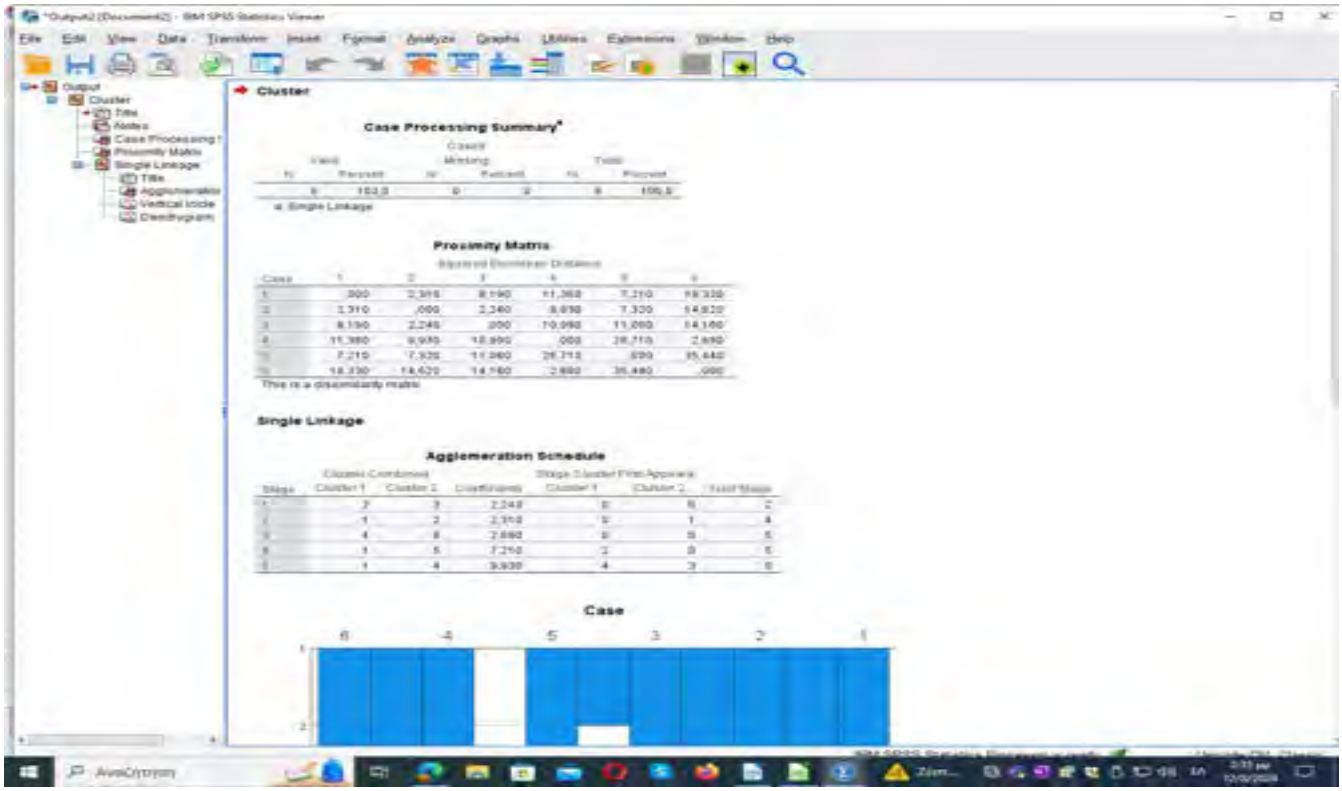
Εικόνα 2.18



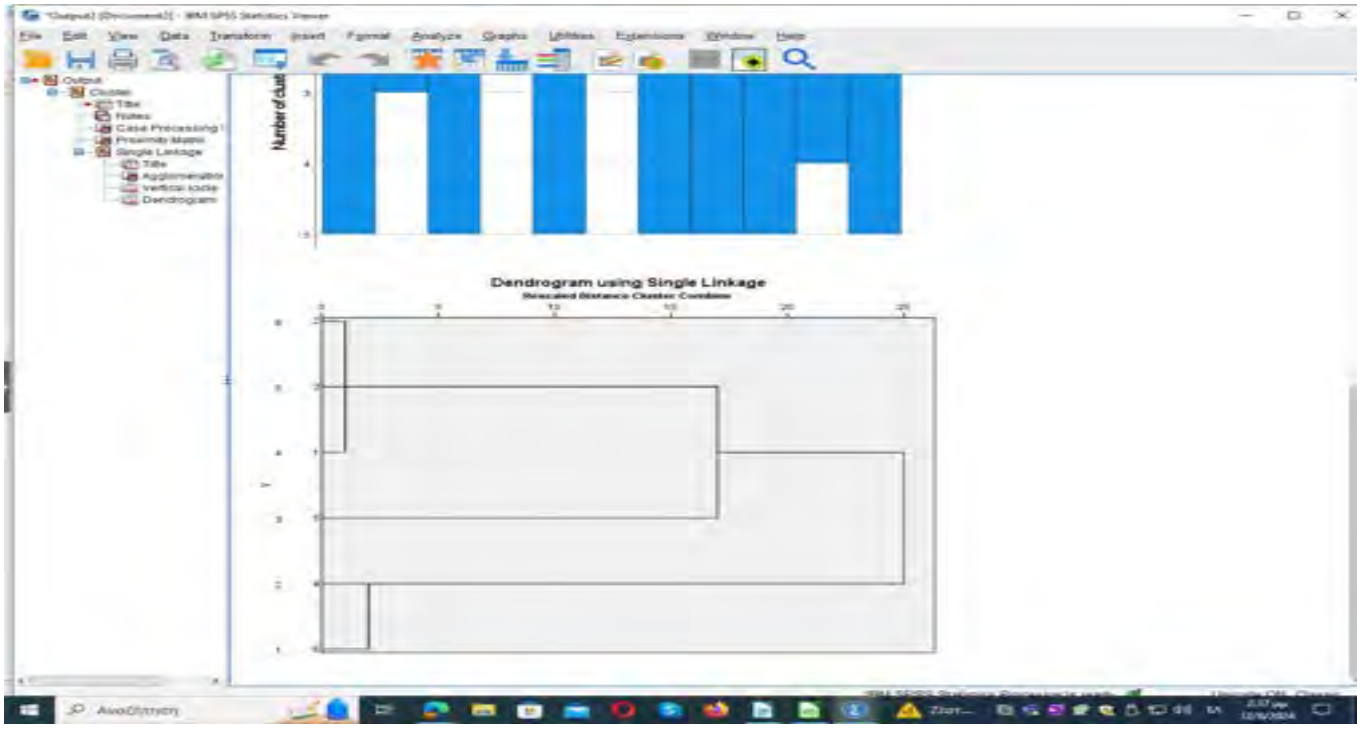
Εικόνα 2.19



Εικόνα 2.20



Εικόνα 2.21



Εικόνα 2.22

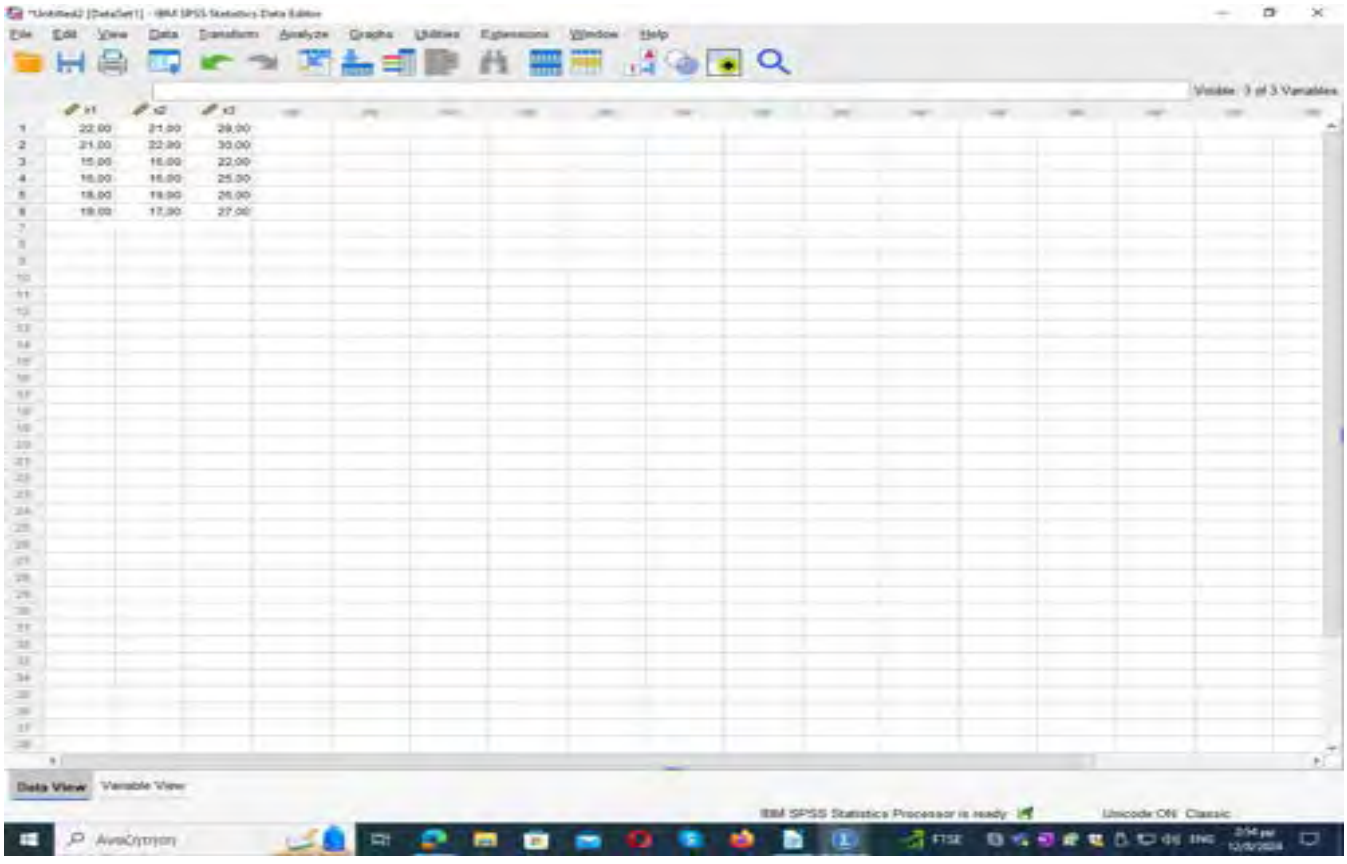
Συνεπώς, λαμβάνοντας υπόψη τα παραπάνω αποτελέσματα του «δέντρου» συστάδας, όπου ο διαχωρισμός βασίζεται στο χαρακτηριστικό της απόστασης, διαπιστώνουμε ότι ο διαχωρισμός των πληθυσμών με το μικρότερο δυνατό σύμπλεγμα ομάδων είναι δυνατόν να οδηγήσει σε καλύτερα συμπεράσματα μελέτης συγγένειας των αντίστοιχων πληθυσμών, καθώς η μελέτη επικεντρώνεται σε όσο το δυνατόν μικρότερο αριθμό ομάδων. Ο διαχωρισμός με μεγαλύτερο αριθμό ομάδων είναι δυνατόν να οδηγήσει σε εσφαλμένα συμπεράσματα καθώς ο ερευνητής έρχεται αντιμέτωπος με περισσότερα χαρακτηριστικά πληθυσμών προκειμένου να συμπεράνει τη συγγένεια ή μη αυτών.

■ Η μικρότερη απόσταση είναι 2,240 μεταξύ του 2 και 3. οπότε οι 2 και 3 δημιουργούν μια ομάδα. Αρα στην απόσταση των 2,240 έχουμε 5 ομάδες: (2,3), (1), (4), (5), (6).
■ Η επόμενη μικρότερη απόσταση είναι 2,310 μεταξύ του 1 και 2. οπότε οι 1, 2 και 3 δημιουργούν μια ομάδα. Αρα στην απόσταση των 2,310 έχουμε 4 ομάδες: (1,2,3), (4), (5), (6).
■ Η επόμενη μικρότερη απόσταση είναι 2,690 μεταξύ του 4 και 6. Αρα στην απόσταση των 2,690 έχουμε 3 ομάδες: (1, 2, 3), (4, 6), (5).
■ Η επόμενη μικρότερη απόσταση είναι 7,210 μεταξύ του 1 και 5. Αρα στην απόσταση των 7,210 έχουμε 2 ομάδες: (1,2, 3, 5), (4, 6).

Παράδειγμα 3. Σε 6 ασθενείς με τραυματισμούς πυελικού δακτυλίου καταγράφηκαν οι ακόλουθες μορφολογικές μεταβλητές του εγκάρσιου επιπέδου του ιερού οστού:
X1 = διάμετρος του δεξιού αυχένα,
X2 = διάμετρος του αριστερού αυχένα και
X3 = διάμετρος του σπονδυλικού σώματος
Μας ενδιαφέρει η διερεύνηση ομάδων ασθενών με παρόμοιο σωματομετρικό προφίλ στο ιερό οστό. Η ταυτοποίηση αυτών των ομάδων μπορεί να μας οδηγήσει στην επιλογή της κατάλληλης χειρουργικής τεχνικής σε ασθενείς με τραυματισμούς πυελικού δακτυλίου.

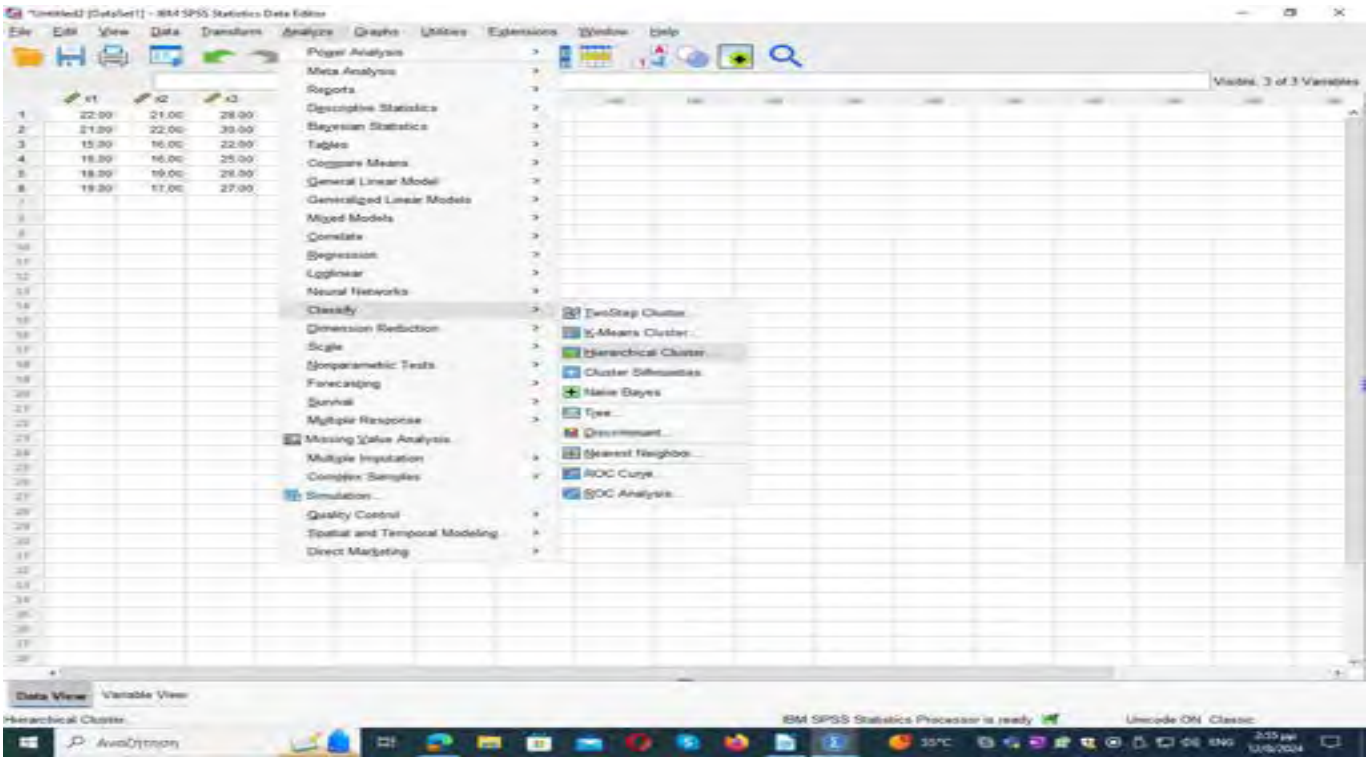
	X1	**X2**	**X3**
I	22	21	28
II	21	22	30
III	15	16	22
IV	16	16	25
V	18	19	26
VI	19	17	27

Το πρώτο βήμα που κάνουμε στο Στατιστικό Πακέτο SPSS είναι να εισάγουμε τα δεδομένα στο πρόγραμμα όπως φαίνεται παρακάτω (Εικόνα 2.23):



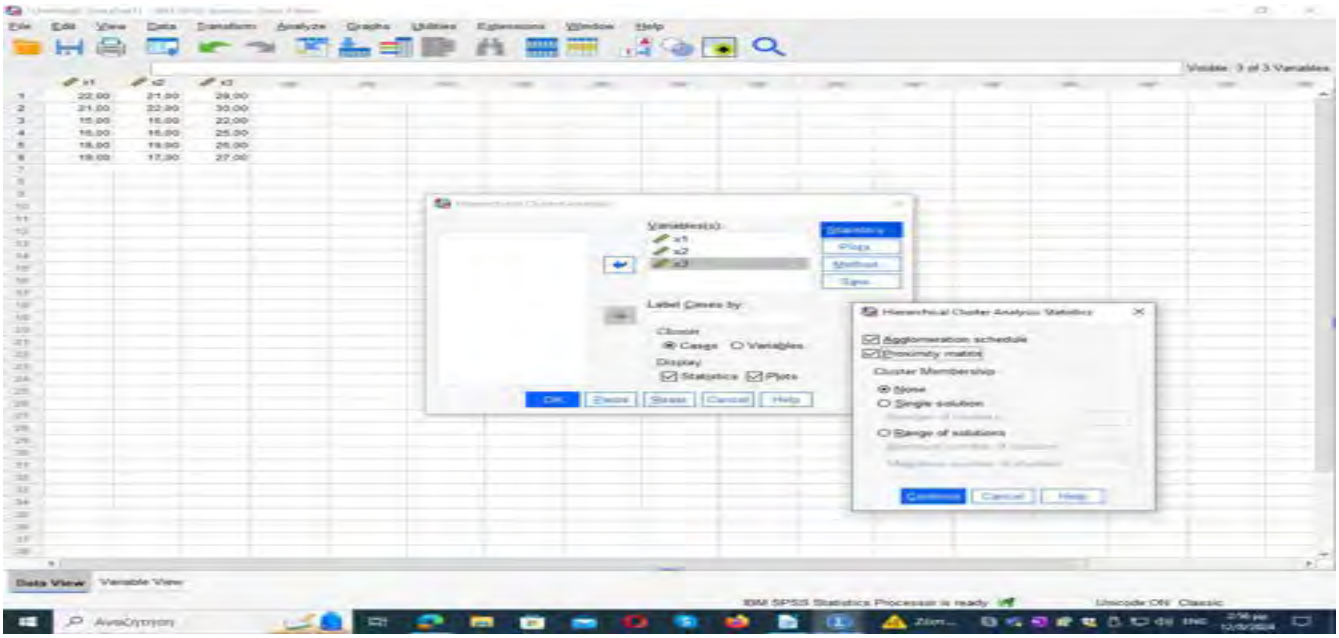
Εικόνα 2.23

Το δεύτερο βήμα που κάνουμε στο Στατιστικό Πακέτο SPSS είναι να επιλέξουμε: **Analyze → Classify → Hierarchical Cluster**. Οπότε έχουμε την παρακάτω εικόνα (Εικόνα 2.24):

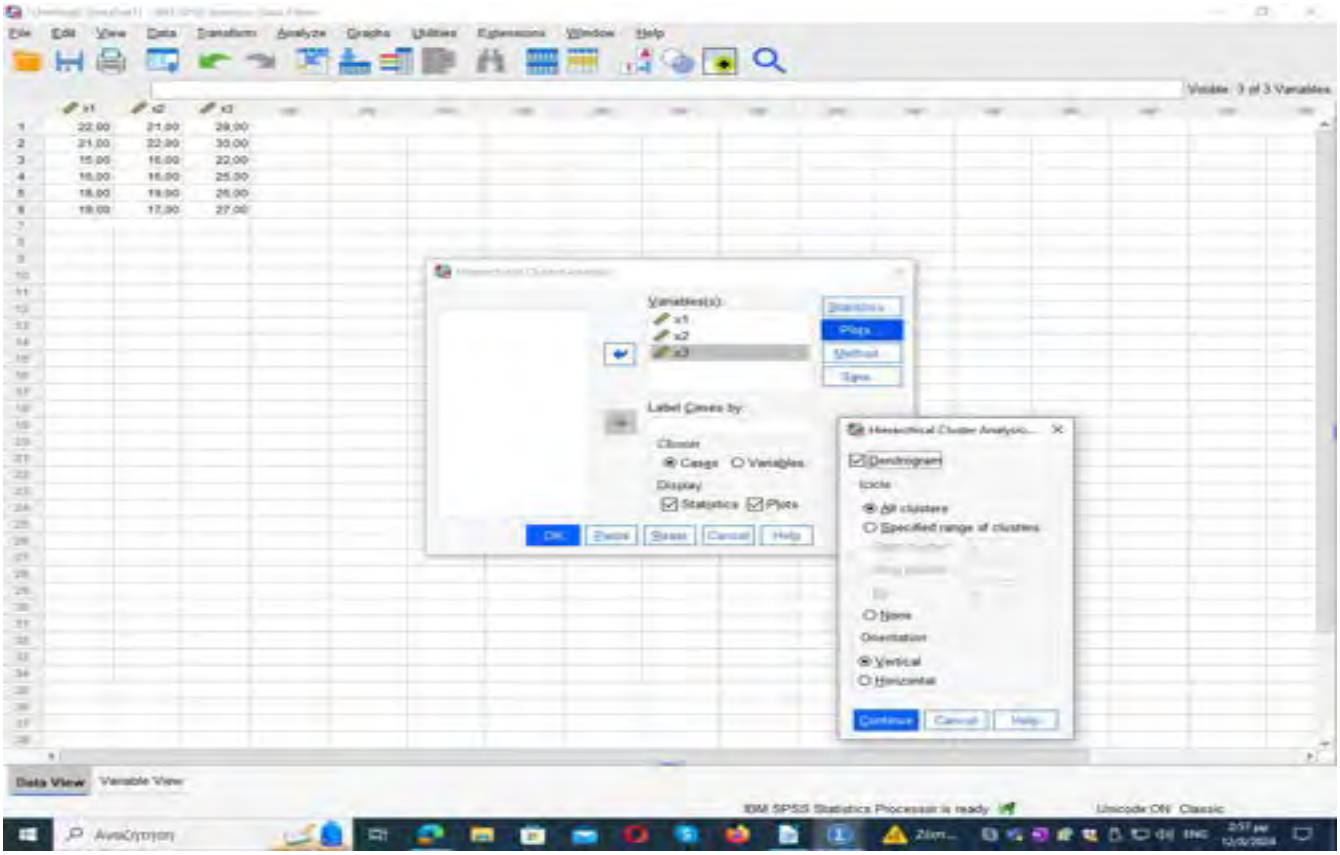


Εικόνα 2.24

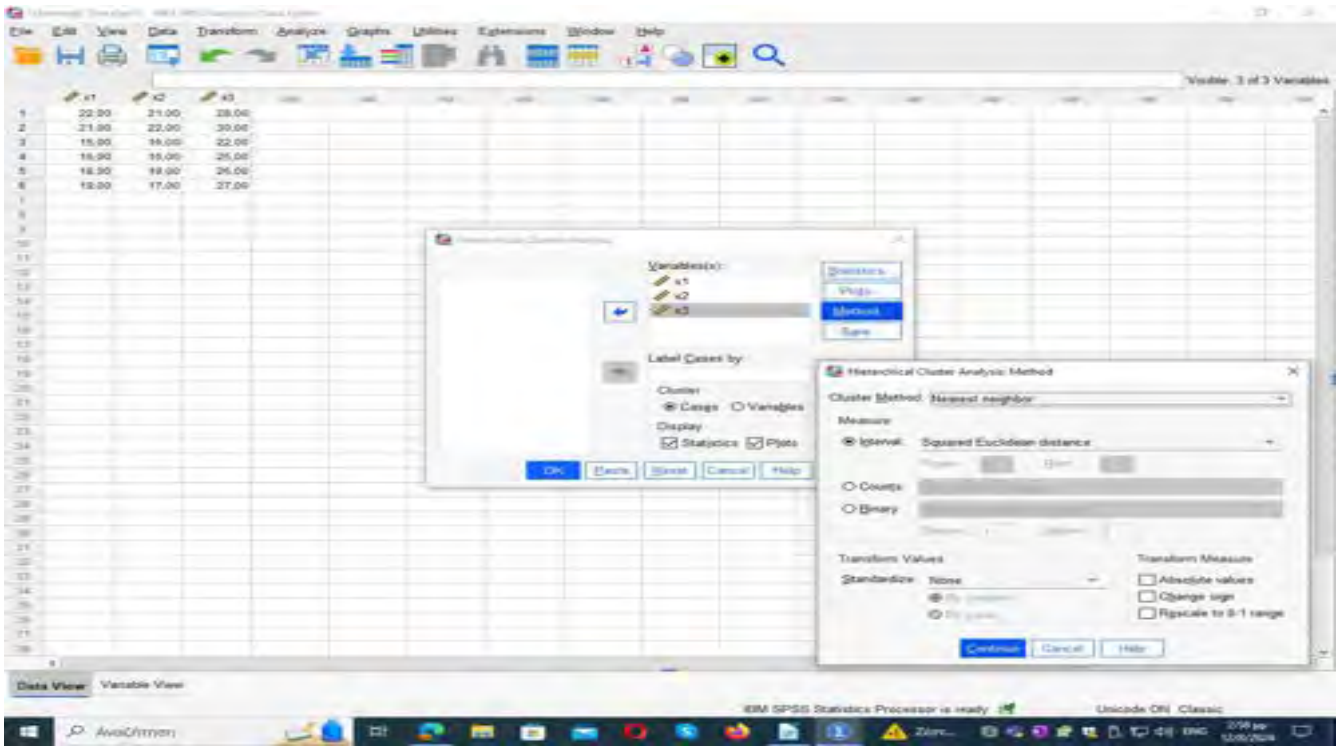
Στη συνέχεια ως τρίτο βήμα κάνουμε τις παρακάτω ρυθμίσεις όπως στις παρακάτω εικόνες (Εικόνα 2.25-Εικόνα 2.28):



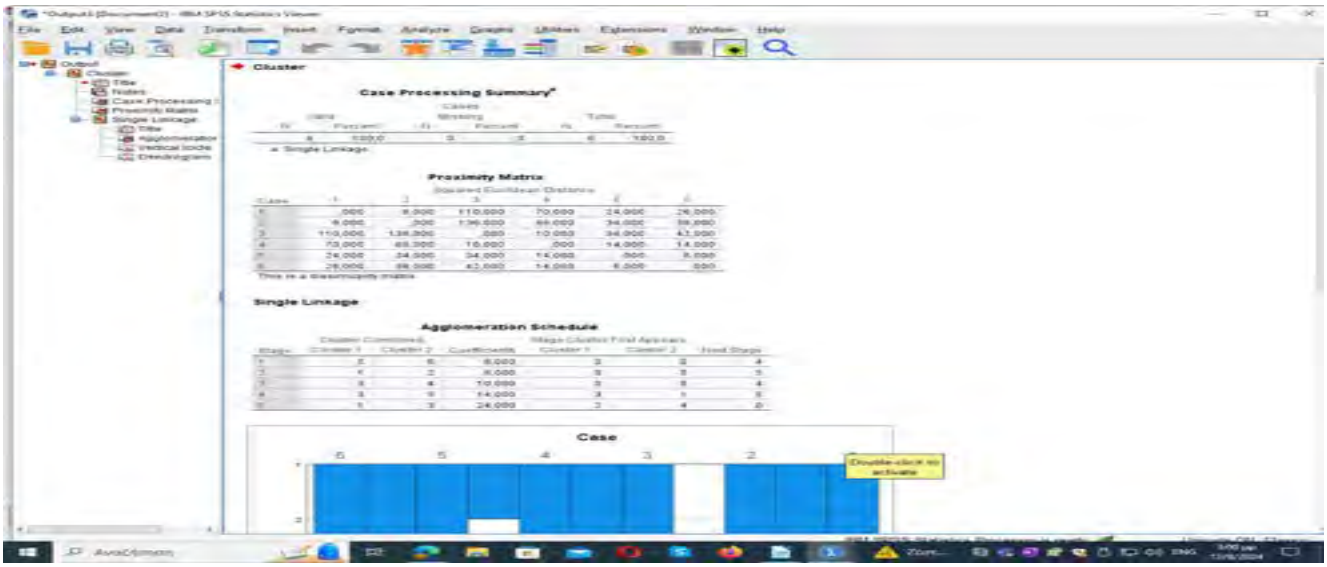
Εικόνα 2.25



Εικόνα 2.26



Εικόνα 2.27



Εικόνα 2.28

Συνεπώς, λαμβάνοντας υπόψη τα παραπάνω αποτελέσματα του «δέντρου» συστάδας, όπου ο διαχωρισμός βασίζεται στο χαρακτηριστικό της απόστασης διαπιστώνουμε ότι ο διαχωρισμός των ασθενών με το μικρότερο δυνατό σύμπλεγμα ομάδων είναι δυνατόν να οδηγήσει σε καλύτερα συμπεράσματα μελέτης, καθώς η μελέτη επικεντρώνεται σε όσο το δυνατόν μικρότερο αριθμό ομάδων. Ο διαχωρισμός με μεγαλύτερο αριθμό ομάδων είναι δυνατόν να οδηγήσει σε εσφαλμένα συμπεράσματα καθώς ο ερευνητής έρχεται αντιμέτωπος με περισσότερα χαρακτηριστικά ασθενών προκειμένου να μελετήσει το σωματομετρικό προφίλ τους στο ιερό οστό και κατ' επέκταση να προτείνει στους ιατρούς ή και να εφαρμόσει ό ίδιος την κατάλληλη χειρουργική τεχνική σε ασθενείς με τραυματισμούς πυελικού δακτυλίου.

■ Η μικρότερη απόσταση είναι 6,000 μεταξύ του 5 και 6. Επίσης 6,000 είναι και η απόσταση μεταξύ των 1 και 2. Αρα στην απόσταση των 6,000 έχουμε 4 ομάδες: (1,2), (5,6), (3), (4).
■ Η επόμενη μικρότερη απόσταση είναι 10 μεταξύ του 3 και 4. οπότε οι 3και 4 δημιουργούν μια ομάδα. Αρα στην απόσταση των 10 έχουμε 3 ομάδες: (1,2), (5,6), (3,4).
■ Η επόμενη μικρότερη απόσταση είναι 14 μεταξύ του 3 και 5. Αρα στην απόσταση των 14 έχουμε 2 ομάδες: (1,2), (3,4,5,6).
■ Η επόμενη μικρότερη απόσταση είναι 24 μεταξύ του 1 και 3. Αρα στην απόσταση των 24 έχουμε 1 ομάδα: (1, 2,3,4,5,6).

2.3.2 k-means Cluster στο Στατιστικό Πακέτο SPSS

Το k-means είναι ένας από τους πιο δημοφιλείς αλγόριθμους «ομαδοποίησης». Διατηρεί k κεντροειδή για τον ορισμό συστάδων. Ένα σημείο ανήκει σε ένα σύμπλεγμα εάν είναι πιο κοντά στο κέντρο αυτού του συμπλέγματος από άλλα.

Γενικά η k-means ομαδοποίηση είναι ένας αλγόριθμος ομαδοποίησης που έχει σχεδιαστεί για να διαχωρίζει δεδομένα σε έναν ορισμένο αριθμό (αυτός είναι το "K") διακριτών ομαδοποιήσεων. Με άλλα λόγια, η k-means ομαδοποίηση βρίσκει παρατηρήσεις που μοιράζονται σημαντικά χαρακτηριστικά και τις ταξινομεί μαζί σε συστάδες (βλ. Εικόνα 2.29). Είναι μια μέθοδος που χρησιμοποιείται για να διαιρέσει μια δέσμη σημείων δεδομένων σε διακριτές ομάδες, διασφαλίζοντας ότι κάθε σημείο βρίσκεται στην ομάδα που βρίσκεται πιο κοντά σε αυτό.

Εικόνα 2.29

Η διαδικασία ανάλυσης συστάδων k-means επιχειρεί να εντοπίσει σχετικά ομοιογενείς ομάδες περιπτώσεων με βάση επιλεγμένα χαρακτηριστικά, χρησιμοποιώντας έναν αλγόριθμο που μπορεί να χειριστεί μεγάλους αριθμούς δεδομένων. ο αλγόριθμος απαιτεί να καθορίσουμε τον αριθμό των συμπλεγμάτων. Μπορούμε να καθορίσουμε αρχικά κέντρα συμπλέγματος εάν γνωρίζουμε αυτές τις πληροφορίες. Η ανάλυση συστάδων k-means είναι ένα εργαλείο που έχει σχεδιαστεί για να εκχωρεί περιπτώσεις σε έναν σταθερό αριθμό ομάδων (συστάδες) των οποίων τα χαρακτηριστικά δεν είναι ακόμη γνωστά αλλά βασίζονται σε ένα σύνολο καθορισμένων μεταβλητών. Είναι χρήσιμη όταν θέλουμε να ταξινομήσουμε έναν μεγάλο αριθμό περιπτώσεων. Μια καλή ανάλυση συστάδων είναι αποτελεσματική όταν χρησιμοποιεί όσο το δυνατόν λιγότερα clusters και καταγράφει όλα τα στατιστικά και εμπορικά σημαντικά συμπλέγματα.
Η ομαδοποίηση k-means μπορεί να εφαρμοστεί σε πολλά διαφορετικά πεδία, καθένα από τα οποία επωφελείται από την ικανότητα του αλγόριθμου να ομαδοποιεί δεδομένα σε έναν προκαθορισμένο αριθμό συστάδων. οι εφαρμογές ποικίλλουν ευρέως. Παρακάτω αναφέρουμε μερικές χαρακτηριστικές εφαρμογές της ομαδοποίησης k-means.

Επιχειρήσεις: Οι εταιρείες μπορούν να χρησιμοποιήσουν την ομαδοποίηση k-means για να τμηματοποιήσει πελάτες με παρόμοιες συμπεριφορές ή προτιμήσεις. Με τις πληροφορίες αυτές προσαρμόζουν τις στρατηγικές μάρκετινγκ, με σκοπό να βελτιώσουν την εξυπηρέτηση πελατών ή να στοχεύσουν συγκεκριμένες ομάδες. Πολλές φορές οι επιχειρήσεις χρησιμοποιούν την ομαδοποίηση k-means προκειμένου να κατηγοριοποίησουν τα αποθέματά τους και να εντοπίσουν ανωμαλίες.

Μηχανική μάθηση και τεχνητή νοημοσύνη: Στον κλάδο της τεχνολογίας, τα k-means βοηθούν στην απλοποίηση πολύπλοκων συνόλων δεδομένων. Η μέθοδος επιτρέπει στα μοντέλα μηχανικής μάθησης να μαθαίνουν και να κάνουν προβλέψεις πιο εύκολα.

Ψυχολογία: Οι ερευνητές εφαρμόζουν ομαδοποίηση k-means για να κατανοήσουν τα πρότυπα στην ανθρώπινη συμπεριφορά ή τις κοινωνικές τάσεις. Για παράδειγμα, η ομαδοποίηση των

χαρακτηριστικών των μαθητών μπορεί να αποκαλύψει τις τάσεις απόδοσης και να βοηθήσει στην ανάπτυξη μοντέλων πρόβλεψης για μελλοντικές ακαδημαϊκές επιδόσεις.

Βιολογία: Στις βιοεπιστήμες, η ομαδοποίηση k-means μπορεί να βοηθήσει στην ομαδοποίηση ιατρικών δεδομένων, όπως υποτύπους καρκίνου, με παρόμοια μοτίβα έκφρασης, τα οποία μπορούν να βοηθήσουν στην κατανόηση του κινδύνου ασθένειας. Χρησιμοποιείται επίσης σε οικολογικές μελέτες για την ταξινόμηση παρόμοιων οικοτόπων ή ειδών με βάση περιβαλλοντικά δεδομένα.

➢ Πλεονεκτήματα της ομαδοποίησης k-means

Απλότητα: Η k-means ομαδοποίηση είναι εύκολα κατανοητή γιατί ομαδοποιούνται τα δεδομένα με βάση το πόσο παρόμοιο είναι κάθε σημείο με την πλησιέστερη τιμή κέντρου.

Ταχύτητα: Η k-means ομαδοποίηση είναι γρήγορη και αποτελεσματική. Μπορεί να οργανώσει γρήγορα τα δεδομένα σε συμπλέγματα.

Κλιμάκωση σε μεγάλα σύνολα δεδομένων: Ο αλγόριθμος k-means γενικά κλιμακώνεται καλά σε μεγαλύτερα σύνολα δεδομένων, κάτι που είναι ωφέλιμο για μεγάλα σύνολα δεδομένων σε κλάδους όπως η υγειονομική περίθαλψη, το μάρκετινγκ και οι μεταφορές.

➢ Μειονεκτήματα της ομαδοποίησης k-means

Απαιτούνται συνεχείς μεταβλητές: Η k-means ομαδοποίηση λειτουργεί μόνο όταν όλες οι μεταβλητές είναι συνεχείς. Δεν λειτουργεί καλά με κατηγορικά δεδομένα (όπως ονόματα φύλου ή χωρών), γεγονός που περιορίζει την εφαρμογή του σε συγκεκριμένους τύπους συνόλων δεδομένων.

Ο αριθμός των συμπλεγμάτων είναι υποκειμενικός: Ο αριθμός συστάδων πριν ξεκινήσουμε τον αλγόριθμο k-means πρέπει να επιλεγεί με βάση τη γνώση του θέματος και τη δομή των δεδομένων μας. Επειδή είναι μια υποκειμενική επιλογή, θα μπορούσε ενδεχομένως να καταλήξουμε σε σφάλματα σε ορισμένες περιπτώσεις.

Ευαίσθητο σε ακραίες τιμές: Ο αλγόριθμος k-means είναι ευαίσθητος σε ακραίες τιμές, που σημαίνει ότι τα αποτελέσματα μπορεί να είναι λιγότερο ακριβή εάν τα δεδομένα μας έχουν υψηλή μεταβλητότητα.

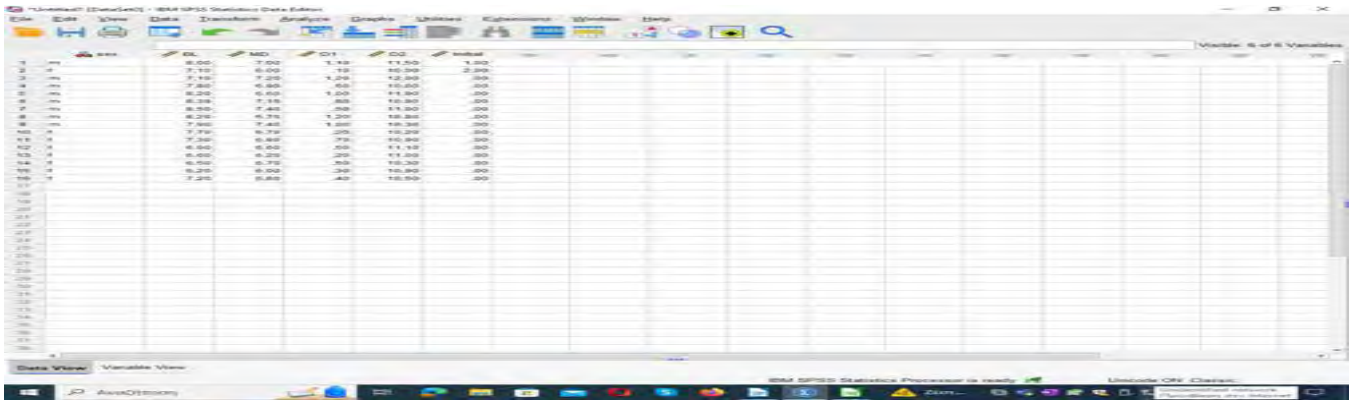
Παράδειγμα 1. Στον παρακάτω πίνακα δίνονται τέσσερεις χαρακτηριστικοί δείκτες των δοντιών ενηλίκων ανδρών και γυναικών. Θεωρούμε ότι η πρώτη στήλη είναι άγνωστη, δηλαδή δεν γνωρίζουμε ποια δείγματα είναι ανδρών (m) και ποια γυναικών (f). Γνωρίζουμε όμως ότι από τα δείγματα αυτά η περίπτωση 1 είναι χαρακτηριστική των ανδρών και η 2 των γυναικών. Με βάση αυτή την πληροφορία θα εκτιμήσουμε ποια δείγματα είναι ανδρικά και ποια γυναικεία.

	BL	MD	D1	D2	Initial
m	8	7	1,10	11,50	1
f	7,10	6	0,10	10	2
m	7,10	7,20	1,20	12	0
m	7,80	6,80	0,5	10	0
m	8,20	6,60	1	11,90	0
m	8,39	7,10	0,8	10,80	0
m	8,50	7,40	0,50	11	0

m	8,20	6,70	1,20	10,80	0
m	7,90	7,40	1	10,30	0
f	7,70	6,70	0,20	10,20	0
f	7,30	6,80	0,70	10,90	0
f	6,90	6,60	0,50	11,10	0
f	6,60	6,20	0,20	11	0
f	6,50	6,70	0,50	10,30	0
f	6,20	6,00	0,30	10,90	0
f	7,20	5,80	0,40	10,50	0

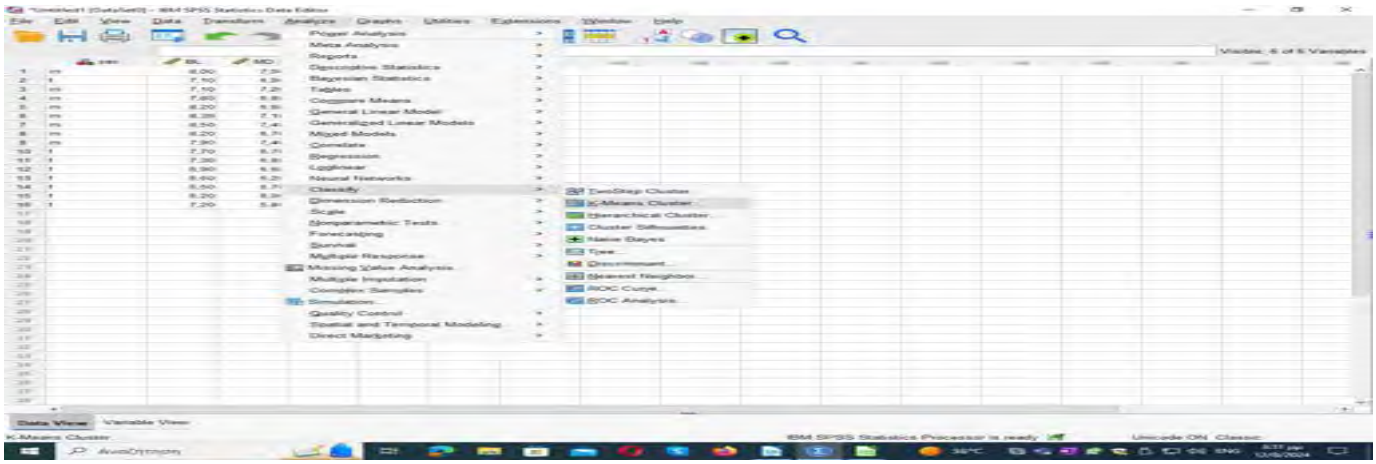
Στο SPSS το πρόβλημα αυτό λύνεται με τη μέθοδο K-Means Cluster. Πρώτα όμως κάνουμε τις εξής ενέργειες:

1η Ενέργεια
Στην έκτη στήλη γράφουμε τον τίτλο Initial και τη συμπληρώνουμε με 0. Στη γραμμή 1 το μηδέν το κάνουμε 1 και στη 2 το 0 γίνεται 2, εφόσον αυτές οι περιπτώσεις είναι χαρακτηριστικές των ανδρών και γυναικών, αντίστοιχα (Εικόνα 2.30).



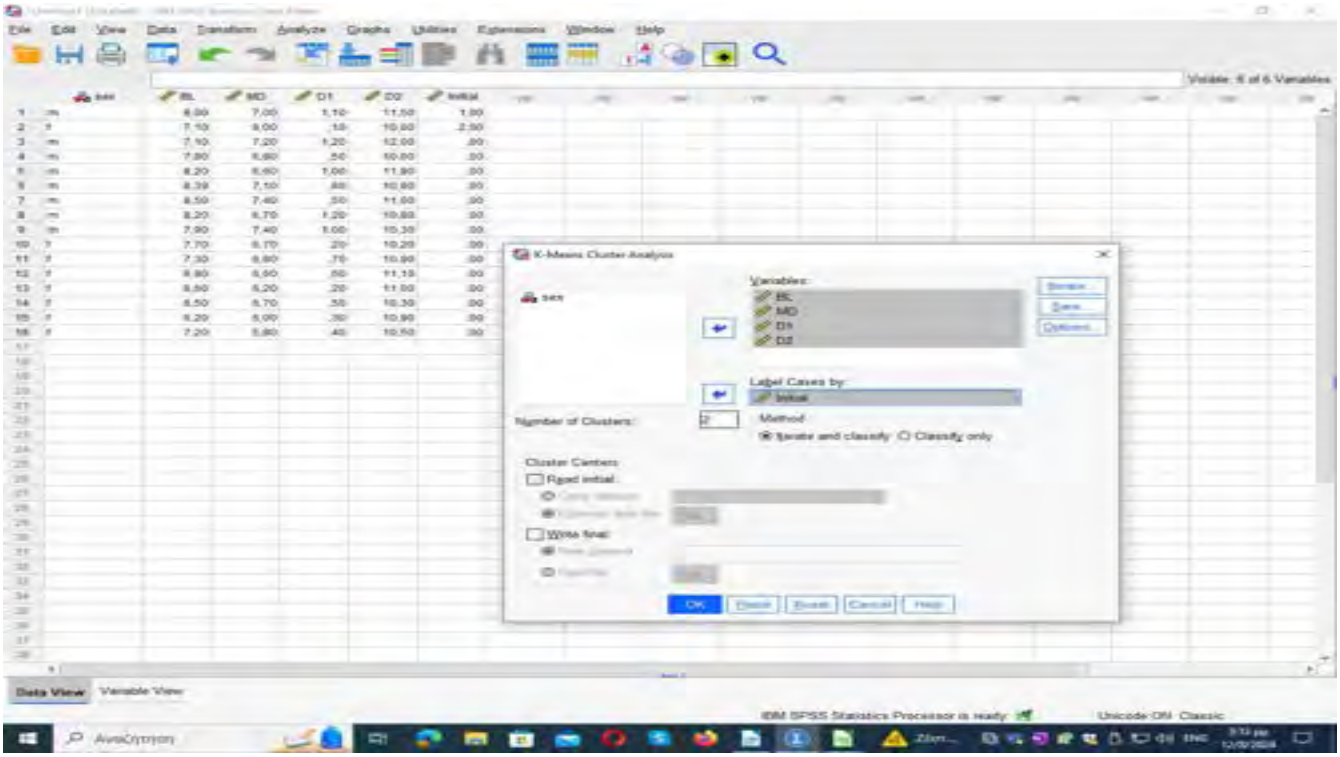
Εικόνα 2.30

2η Ενέργεια
Στη συνέχεια ακολουθούμε την πορεία: **Analyze → Classify → K-Means Cluster** (Εικόνα 2.31).



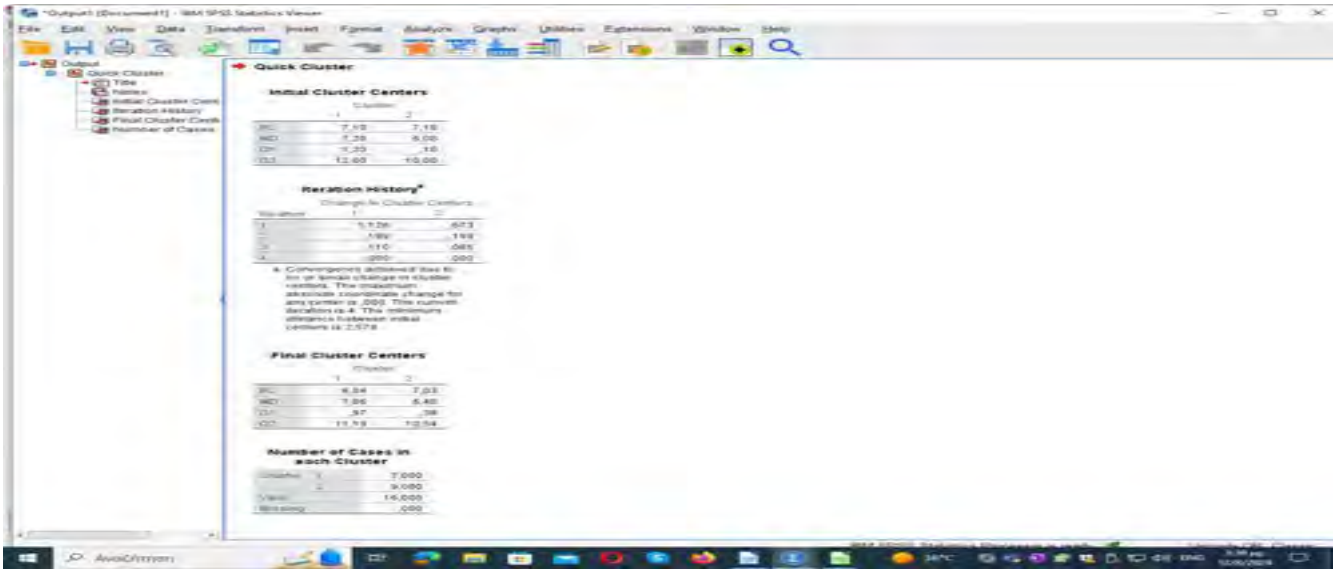
Εικόνα 2.31

Στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές BL, MD, D1 και D2 στο πλαίσιο **Variables** ενώ τη μεταλητή Initial στο **Label Cases by** (Εικόνα 2.32).



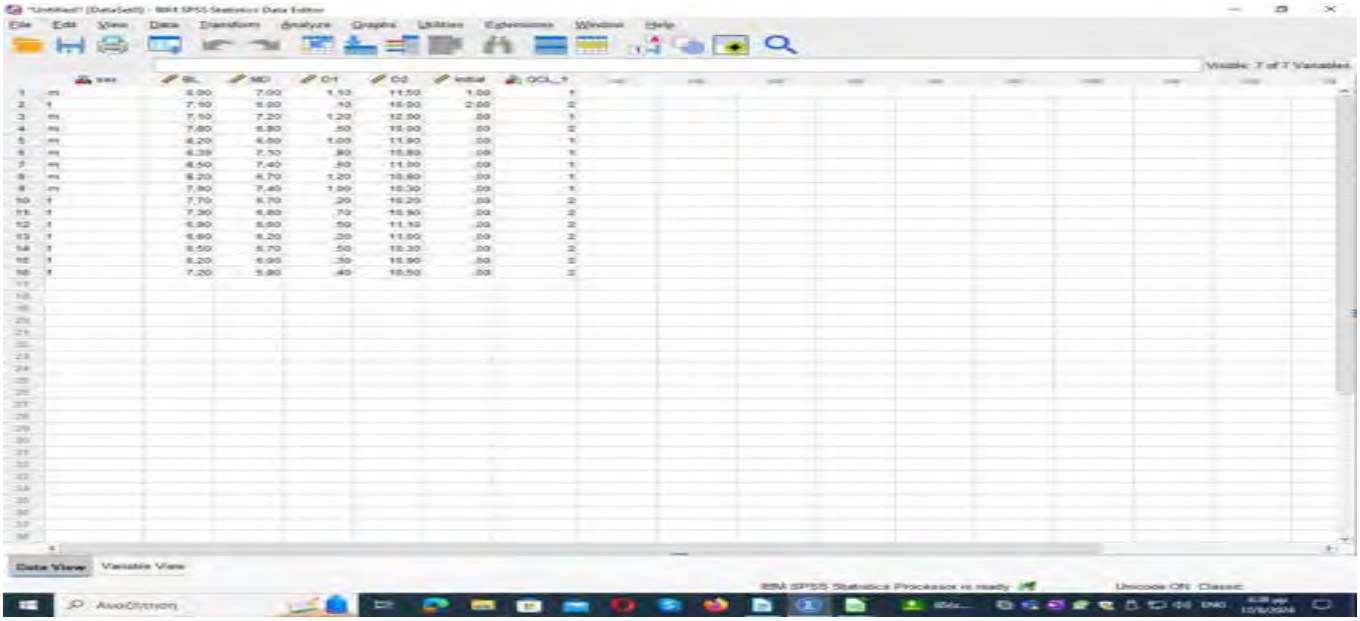
Εικόνα 2.32

Στη συνέχεια κάνουμε κλικ στο **Save** και επιλέγουμε **Cluster membership**. Ολοκληρώνουμε με κλικ στο **Continue** και στο **OK**. οπότε έχουμε τους παρακάτω πίνακες (Εικόνα 2.33):



Εικόνα 2.33

Τα αποτελέσματα της μεθόδου δίνονται σε μια νέα στήλη που προστίθεται στο αρχικό φύλλο εργασίας με τίτλο QCL_1. Στη στήλη αυτή με 1 δηλώνονται τα ανδρικά δείγματα και με 2 τα γυναικεία, εφόσον αυτούς τους αριθμούς χρησιμοποιήσαμε στη στήλη Initial για να ξεχωρίσουμε τα ανδρικά από τα γυναικεία δείγματα (βλ. Εικόνα 2.34).



Εικόνα 2.34

Με αυτό τον τρόπο μπορούμε να ομαδοποιήσουμε τα δεδομένα μας σε ανδρικά (με χαρακτηριστική 1) και σε γυναικεία (με χαρακτηριστική 2).

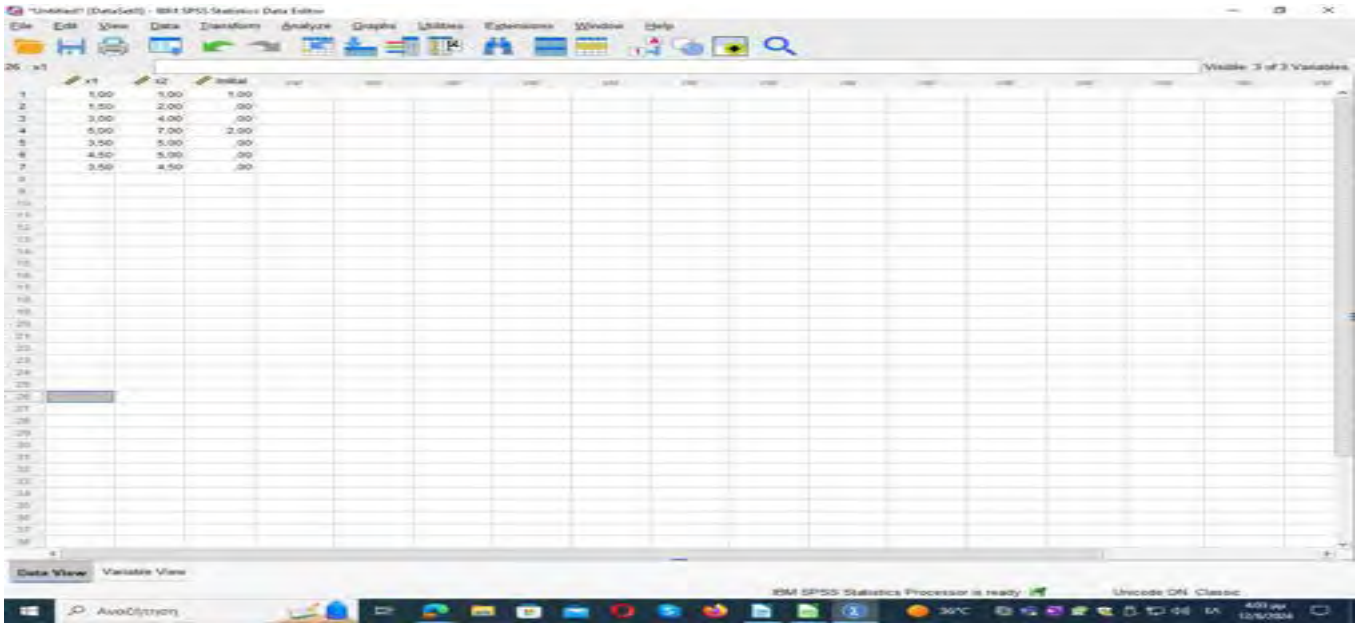
Παράδειγμα 2. Στον παρακάτω πίνακα δίνονται δύο χαρακτηριστικοί δείκτες για χαρακτηριστικά δυο διαφορετικών ζώων. Γνωρίζουμε όμως ότι από τα δείγματα αυτά η περίπτωση 1 είναι χαρακτηριστική του ζώου Α και η 4 του ζώου Β. Με βάση αυτή την πληροφορία θα εκτιμήσουμε ποια δείγματα είναι του ζώου Α και ποια του ζώου Β.

x1	**x2**
1	1
1,5	2
3	4
5	7
3,5	5
4,5	5
3,5	4,5

Στο SPSS το πρόβλημα αυτό λύνεται με τη μέθοδο K-Means Cluster. Πρώτα όμως κάνουμε τις εξής ενέργειες:

1η Ενέργεια

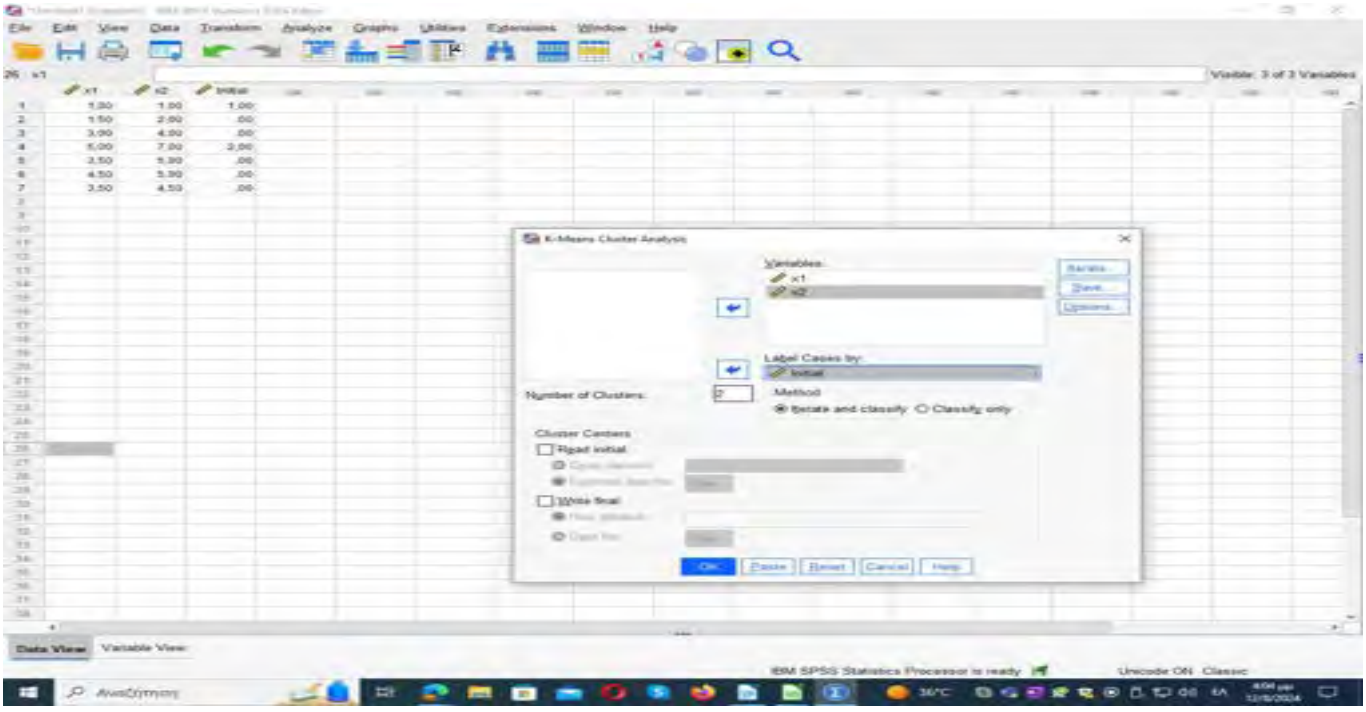
Στην τρίτη στήλη γράφουμε τον τίτλο Initial και τη συμπληρώνουμε με 0. Στη γραμμή 1 το μηδέν το κάνουμε 1 και στη 4 το 0 γίνεται 2, εφόσον αυτές οι περιπτώσεις είναι χαρακτηριστικές των ζώων Α και ζώων Β, αντίστοιχα (Εικόνα 2.35).



Εικόνα 2.35

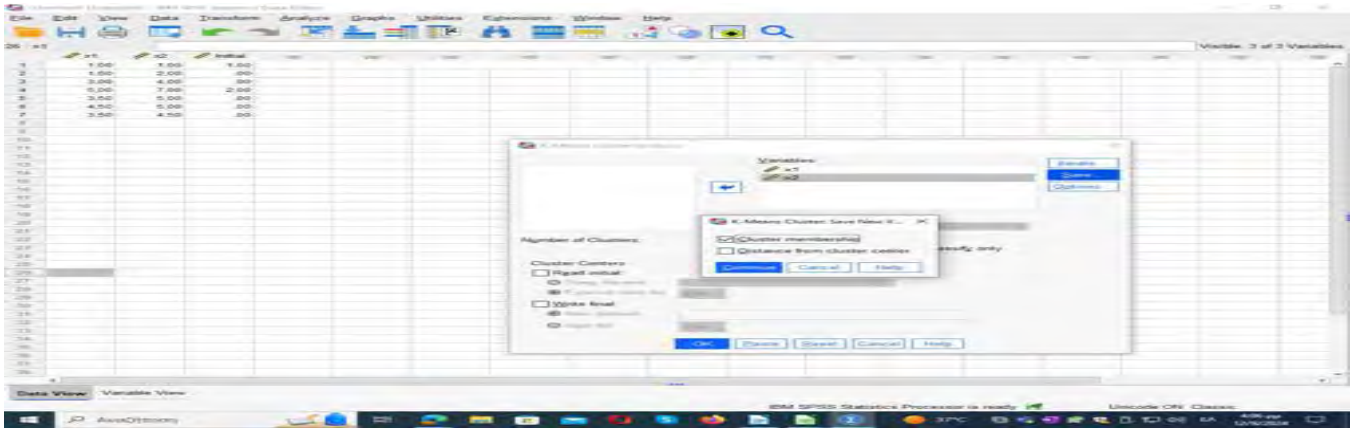
2η Ενέργεια

Στη συνέχεια ακολουθούμε την πορεία: **Analyze → Classify → K-Means Cluster** (Εικόνα 2.36).



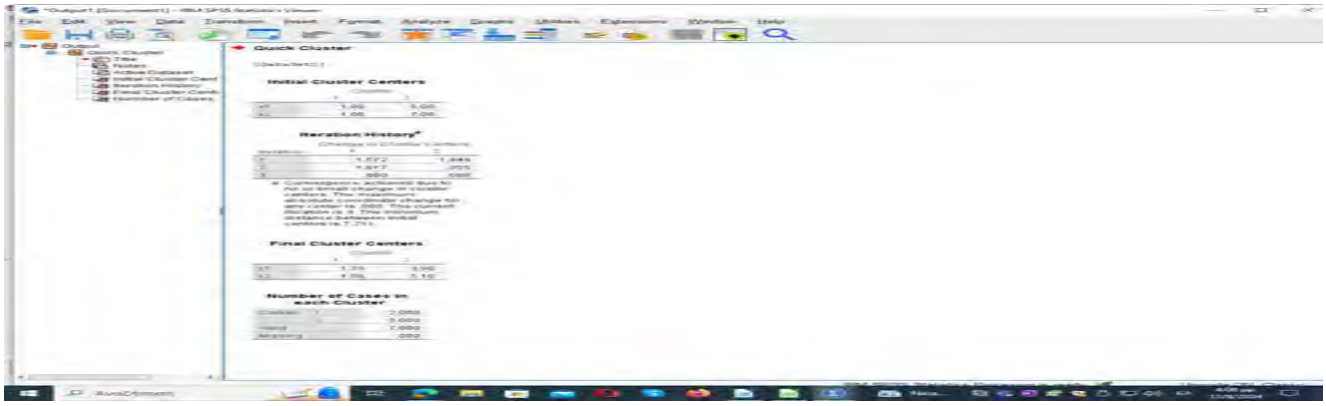
Εικόνα 2.36

Στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές x1 και x2 στο πλαίσιο **Variables** ενώ τη μεταβλητή Initial στο **Label Cases by** (Εικόνα 2.37).



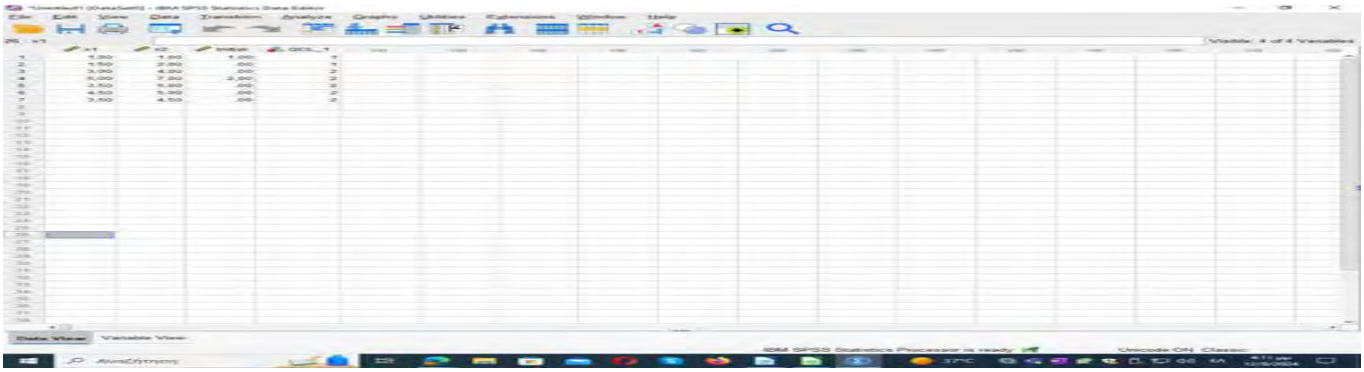
Εικόνα 2.37

Οπότε έχουμε τον παρακάτω πίνακα (Εικόνα 2.38):



Εικόνα 2.38

Τα αποτελέσματα της μεθόδου δίνονται σε μια νέα στήλη που προστίθεται στο αρχικό φύλλο εργασίας με τίτλο QCL_1. Στη στήλη αυτή με 1 δηλώνονται τα δείγματα του ζώου Α και με 2 τα δείγματα του ζώου Β, εφόσον αυτούς τους αριθμούς χρησιμοποιήσαμε στη στήλη Initial για να ξεχωρίσουμε τα δείγματα του ζώου Α από τα δείγματα του ζώου Β (βλ. Εικόνα 2.39).



Εικόνα 2.39

ΚΕΦΑΛΑΙΟ 3ο
Η Διαχωριστική Ανάλυση στο πρόγραμμα SPSS

Στο κεφάλαιο αυτό παρουσιάζεται η τεχνική της διαχωριστικής ανάλυσης, ποικίλες εφαρμογές αυτής καθώς και η σύνδεσή της με το στατιστικό πακέτο SPSS μέσα από παραδείγματα της Ιατρικής Επιστήμης.

3.1 Περιγραφή της Διαχωριστικής Ανάλυσης

Η διαχωριστική ανάλυση είναι μία στατιστική τεχνική, η οποία έχει δύο στόχους:
1. Την διάκριση ενός πληθυσμού σε ευδιάκριτα σύνολα.
2. Την ταξινόμηση παρατηρήσεων στα παραπάνω σύνολα (χρησιμοποιώντας έναν κανόνα - σχέση).

Η Διαχωριστική Ανάλυση είναι μια στατιστική τεχνική διαχωρισμού των παρατηρήσεων. Προϋποθέτει τον διαχωρισμό των δεδομένων σε δύο ή περισσότερες ομάδες (οι ομάδες είναι γνωστές). Αποτελεί μια μέθοδο ταξινόμησης νέων παρατηρήσεων στις αρχικές ομάδες ανάλογα με τα χαρακτηριστικά τους. Μας βοηθάει να λάβουμε αποφάσεις για το μέλλον (βλ. για παράδειγμα, [7], [9], [14], [23], [25]). Τα αποτελέσματά μας για να είναι καλά και αξιόπιστα πρέπει να έχουμε:

- Πολυ-μεταβλητή κανονικότητα .
- Αποφυγή πολύ-συγγραμικότητας.
- Μεταξύ των ανεξάρτητων μεταβλητών πρέπει να υπάρχει γραμμική σχέση.
- Να τονίσουμε ότι οι μεταβλητές που θα χρησιμοποιήσουμε πρέπει να είναι αυστηρά ποσοτικές, αριθμητικές και συγκεκριμένα συνεχείς ή έστω διακριτές με πολλές τιμές όμως. Δηλαδή γνώση της ομάδας αίματος, των πολιτικών πεποιθήσεων και γενικά κατηγορικές μεταβλητές δεν επιτρέπονται.
- Τα δεδομένα μας πρέπει να ακολουθούν Κανονική Κατανομή.

Όλα τα παραπάνω μπορούμε να τα ελέξουμε με την χρήση του στατιστικού πακέτου SPSS. Η διαχωριστική ανάλυση (discriminant analysis) έχει πάρα πολλές εφαρμογές και είναι ιδιαίτερα χρήσιμη στην Ιατρική για διαγνώσεις ασθενειών μέσου υπολογιστή. Ας περιγράψουμε τώρα μερικά παράδειγμα όπου η διαχωριστική ανάλυση παίζει σημαντικό ρόλο.

Παράδειγμα 1. Υποθέτουμε ότι έχουμε μετρήσεις για ασθενείς μιας κλινικής οπως η ηλικία, το βάρος, το ύψος, οι εξετάσεις αίματος και άλλες ιατρικές εξετάσεις. Για τους ασθενείς αυτούς γνωρίζουμε ποιοι από αυτούς έχουν μια ασθένεια και ποιοι όχι. Σκοπός είναι να δούμε αν μπορούμε να διαχωρίσουμε τους ασθενείς σε δυο κατηγορίες: σε αυτούς που έχουν την ασθένεια και σε αυτούς που δεν έχουν την ασθένεια, με βάση μόνο τις μετρήσεις που έχουμε στη διάθεση μας για τους ασθενείς αυτούς.

Παράδειγμα 2. Ας υποθέσουμε ότι έχουμε μετρήσεις από μία ή περισσότερες μεταβλητές και επίσης ξέρουμε και τις διάφορες ομάδες του δείγματος. Δηλαδή, μπορεί να έχουμε την ηλικία, το βάρος, το ύψος του μισθού, τα χρόνια εκπαίδευσης και τα χρόνια προϋπηρεσίας για κάποιους εργαζομένους σε μία εταιρεία και επίσης ξέρουμε και το φύλο τους. Σκοπός είναι να δούμε αν μπορούμε να διαχωρίσουμε τους άντρες από τις γυναίκες με βάση τις μεταβλητές που έχουμε στη διάθεση μας.

Παράδειγμα 3. Ξέρουμε ότι τα αυτοκίνητα είναι είτε Ευρωπαϊκά, είτε Αμερικάνικα, είτε Ιαπωνικά. Με βάση λοιπόν κάποιες πληροφορίες που έχουμε για αυτά θα προσπαθήσουμε να τα διαχωρίσουμε. Γνωρίζουμε το βάρος, την επιτάχυνση, την κατανάλωση, τον κυβισμό και τον αριθμό των κυλίνδρων τους

Η Discriminant Analysis έχει πολλές εφαρμογές στην έρευνα, την Ιατρική και τις επιχειρήσεις, όπως για παράδειγμα:

Τμηματοποίηση πελατών (Μάρκετινγκ): Η Discriminant Analysis βοηθά στον εντοπισμό των χαρακτηριστικών των πελατών που αγοράζουν ένα συγκεκριμένο προϊόν ή υπηρεσία, κάτι που μπορεί να βοηθήσει στην τμηματοποίηση της αγοράς και στη στόχευση συγκεκριμένων ομάδων πελατών.

Πιστωτική βαθμολογία: Η Discriminant Analysis χρησιμοποιείται για την αξιολόγηση της πιστοληπτικής ικανότητας ατόμων με βάση τα οικονομικά και δημογραφικά τους δεδομένα.

Ιατρική έρευνα: Η Discriminant Analysis μπορεί να βοηθήσει στον εντοπισμό των παραγόντων που διαφοροποιούνται μεταξύ υγιών και ασθενών ατόμων ή μεταξύ διαφορετικών σταδίων μιας ασθένειας.

Ποιοτικός έλεγχος: Η Discriminant Analysis μπορεί να βοηθήσει στον εντοπισμό των παραγόντων που επηρεάζουν την ποιότητα ενός προϊόντος και να διασφαλίσει ότι πληροί τα απαιτούμενα πρότυπα.

Ιατρική διάγνωση: Η Discriminant Analysis μπορεί να χρησιμοποιηθεί για τη διάγνωση ασθενειών ταξινομώντας τους ασθενείς σε ομάδες ασθενειών ή μη με βάση τα συμπτώματά τους ή τα αποτελέσματα των δοκιμών τους. Η διάγνωση της ασθένειας κάποιου ασθενή με βάση τα συμπτώματα που έχει αυτός είναι μια σημαντική εφαρμογή της Discriminant Analysis. Δεδομένου πως τα συμπτώματα κάθε ασθένειας είναι γνωστά, σκοπός μας είναι η δημιουργία ενός κανόνα που θα κάνει τη διάγνωση για έναν καινούριο ασθενή, λαμβάνοντας υπόψη τα συμπτώματά της ασθένειας αλλά και τη γνώση μας για τα συμπτώματα ενός συνόλου ασθενειών.

Ποινικό προφίλ: Η ανάλυση διακρίσεων μπορεί να χρησιμοποιηθεί για τον εντοπισμό των χαρακτηριστικών των εγκληματιών με βάση τα εγκληματικά τους πρότυπα ή άλλα χαρακτηριστικά.

Οικολογία: Η διακριτική ανάλυση μπορεί να χρησιμοποιηθεί για τον προσδιορισμό των οικοτόπων διαφορετικών ειδών με βάση τα περιβαλλοντικά χαρακτηριστικά τους.

Εκπαίδευση: Η ανάλυση διακρίσεων μπορεί να χρησιμοποιηθεί για να προβλέψει ποιοι μαθητές είναι πιο πιθανό να επιτύχουν σε ένα συγκεκριμένο πρόγραμμα με βάση τα χαρακτηριστικά τους.

Γενικά η Discriminant Analysis είναι ένα ισχυρό στατιστικό εργαλείο που μπορεί να μας βοηθήσει να ταξινομήσουμε τα δεδομένα μας σε διακριτές ομάδες με βάση ένα σύνολο ανεξάρτητων μεταβλητών. Έχει διάφορες εφαρμογές στην έρευνα και τις επιχειρήσεις και μπορεί να βοηθήσει στην τμηματοποίηση πελατών, τη βαθμολογία πιστώσεων, την ιατρική έρευνα και τον ποιοτικό έλεγχο.

Στη διαχωριστική ανάλυση κύριο μέλημα μας είναι η κατασκευή ενός κανόνα που θα μας βοηθήσει να λάβουμε αποφάσεις στο μέλλον, ενώ στην ανάλυση κατά συστάδες ο κύριος στόχος μας είναι να δημιουργήσουμε ομοειδείς ομάδες με σκοπό την κατανόηση των ήδη υπαρχόντων στοιχείων και τη μείωση της διασποράς σε επιμέρους ομάδες.

3.2. Διαχωριστική Ανάλυση και SPSS

Παρακάτω θα αναφέρουμε προβλήματα Διαχωριστικής Ανάλυσης η αντιμετώπιση των οποίων θα γίνει με χρήση του στατιστικού πακέτου SPSS.

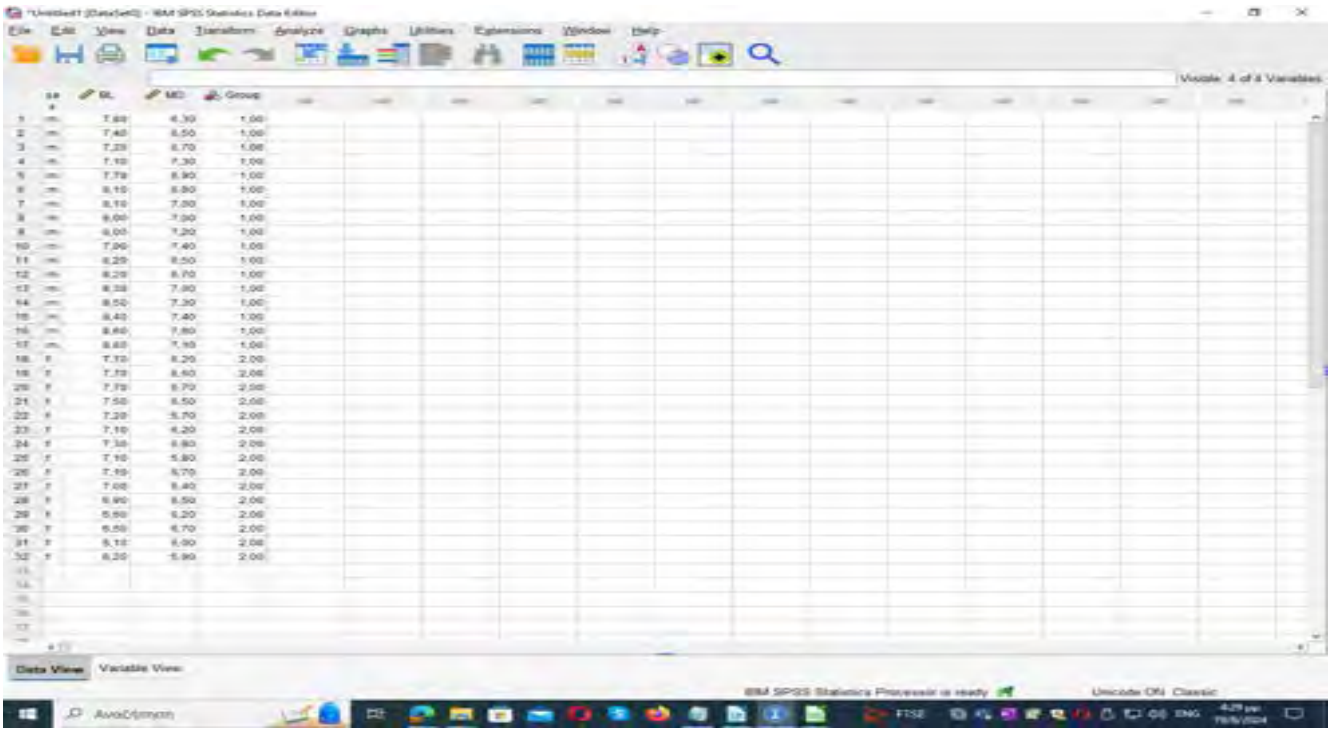
Παράδειγμα 1. Στον παρακάτω πίνακα δίνονται οι δείκτες BL και MD των δοντιών ενήλικων ανδρών και γυναικών.

m/f	BL	MD	m/f	BL	MD
m	7,8	6,3	m	8,8	7,1
m	7,4	6,5	f	7,7	6,2
m	7,2	6,7	f	7,7	6,6
m	7,1	7,3	f	7,7	6,7
m	7,7	6,9	f	7,5	6,5
m	8,1	6,8	f	7,2	5,7
m	8,1	7	f	7,1	6,2
m	8	7	f	7,3	6,8
m	8	7,2	f	7,1	5,8
m	7,9	7,4	f	7,1	6,7
m	8,2	6,5	f	7	6,4
m	8,2	6,7	f	6,9	6,5
m	8,3	7	f	6,6	6,2
m	8,5	7,3	f	6,5	6,7
m	8,4	7,4	f	6,1	6
m	8,6	7,6	f	6,2	5,9

Σε πρώτη φάση αυτό που θέλουμε και ζητούμε απο το πρόγραμμα SPSS στα πλαίσια της διαχωριστικής ανάλυσης είναι να δούμε αν τα δεδομένα μας σχηματίζουν δύο διακριτές κατηγορίες. Δηλαδή αυτό που θα δούμε σε πρώτη φάση είναι αν με βάση τις μετρήσεις που έχουμε στην διάθεσή μας υπάρχει ικανοποιητικός διαχωρισμός των δύο κατηγοριών (άνδρες και γυναίκες). Σε δεύτερη φάση θα προσπαθήσουμε να προσδιορίσουμε αν τα δείγματα (BL,MD)=(8, 6.7) και (BL, MD)= (7, 6.5) ανήκουν σε άνδρα και γυναίκα.

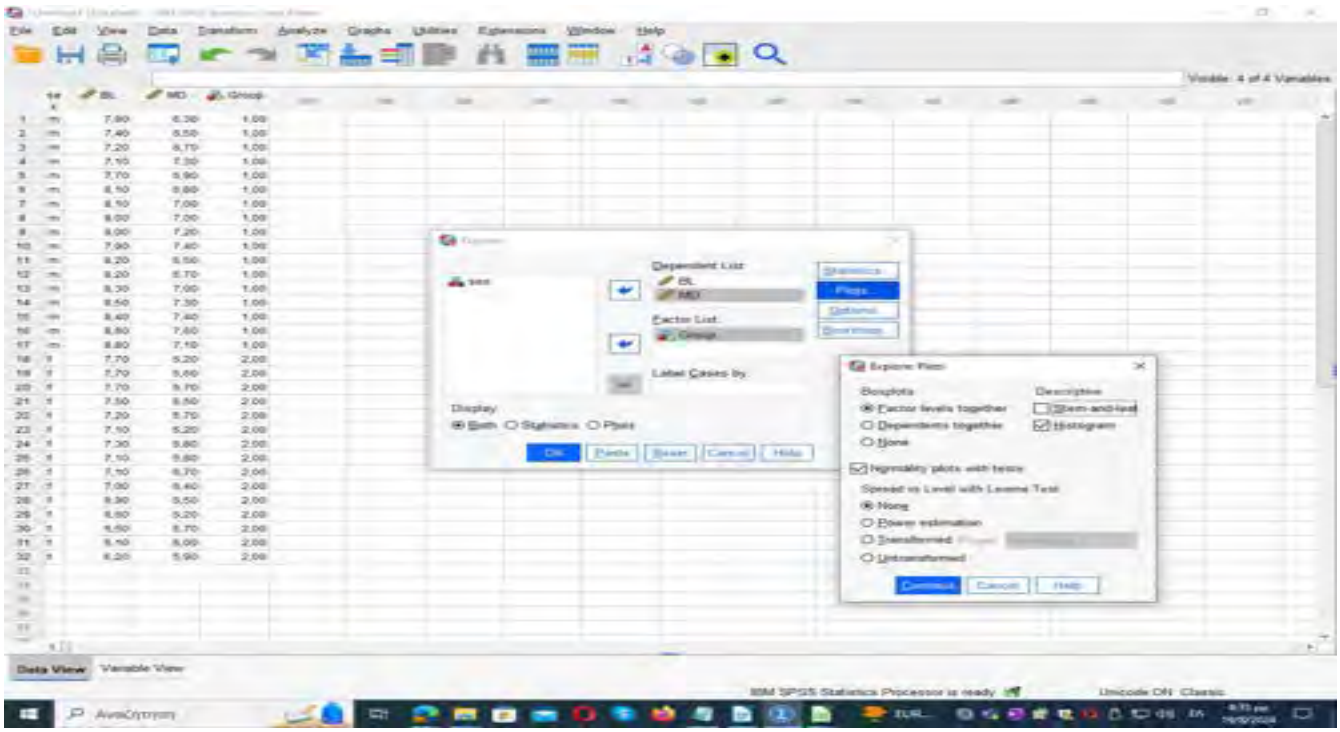
Ας δούμε έναν-έναν τους προβληματισμούς που θέσαμε παραπάνω με την χρήση του στατιστικού πακέτου SPSS.

Βήμα 1. Μεταφέρουμε τα δεδομένα στο SPSS όπως φαίνεται στο παρακάτω σχήμα και δημιουργούμε μια νέα στήλη με όνομα Group της οποίας η μεταβλητή παίρνει την τιμή 1 όταν αντιστοιχεί στον άνδρα (m) και την τιμή 2 όταν αντιστοιχεί σε γυναίκα (f) (βλ. Εικόνα 3.1).



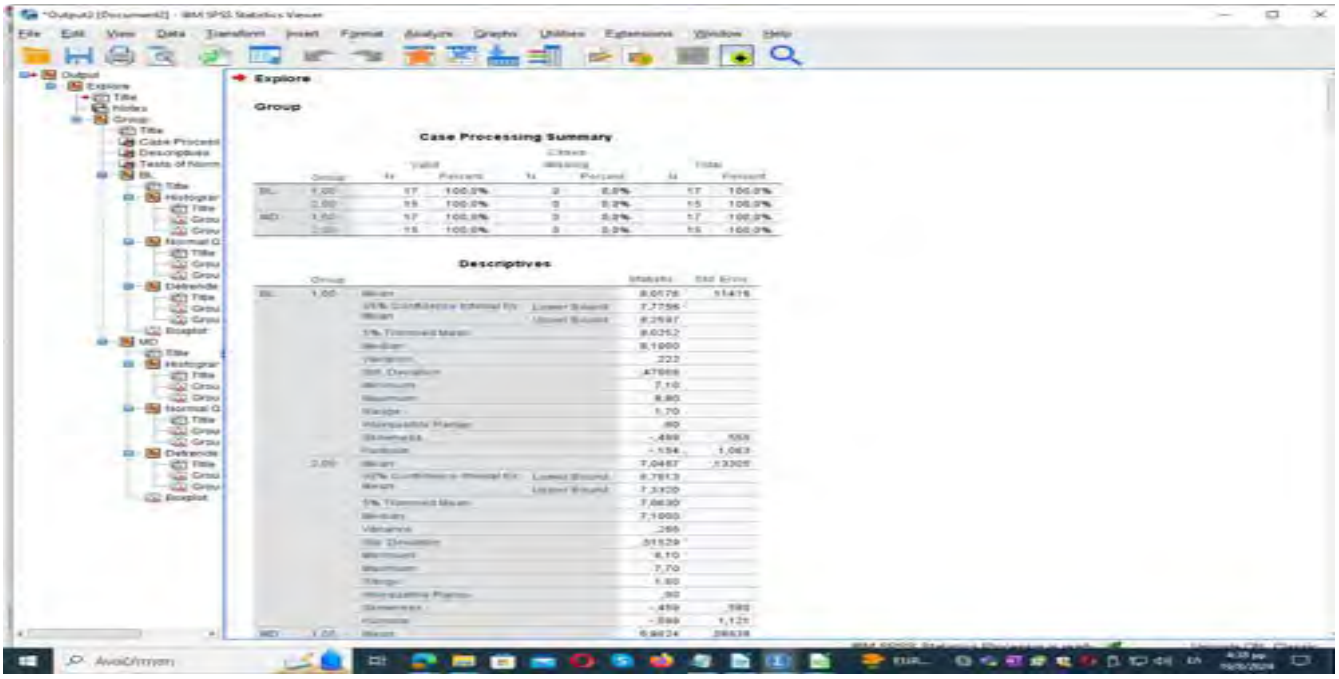
Εικόνα 3.1

Βήμα 2. Κάνουμε έλεγχο κανονικότητας. Επιλέγουμε: **Analyze → Descriptive Statistics → Explore**…(Εικόνα 3.2).

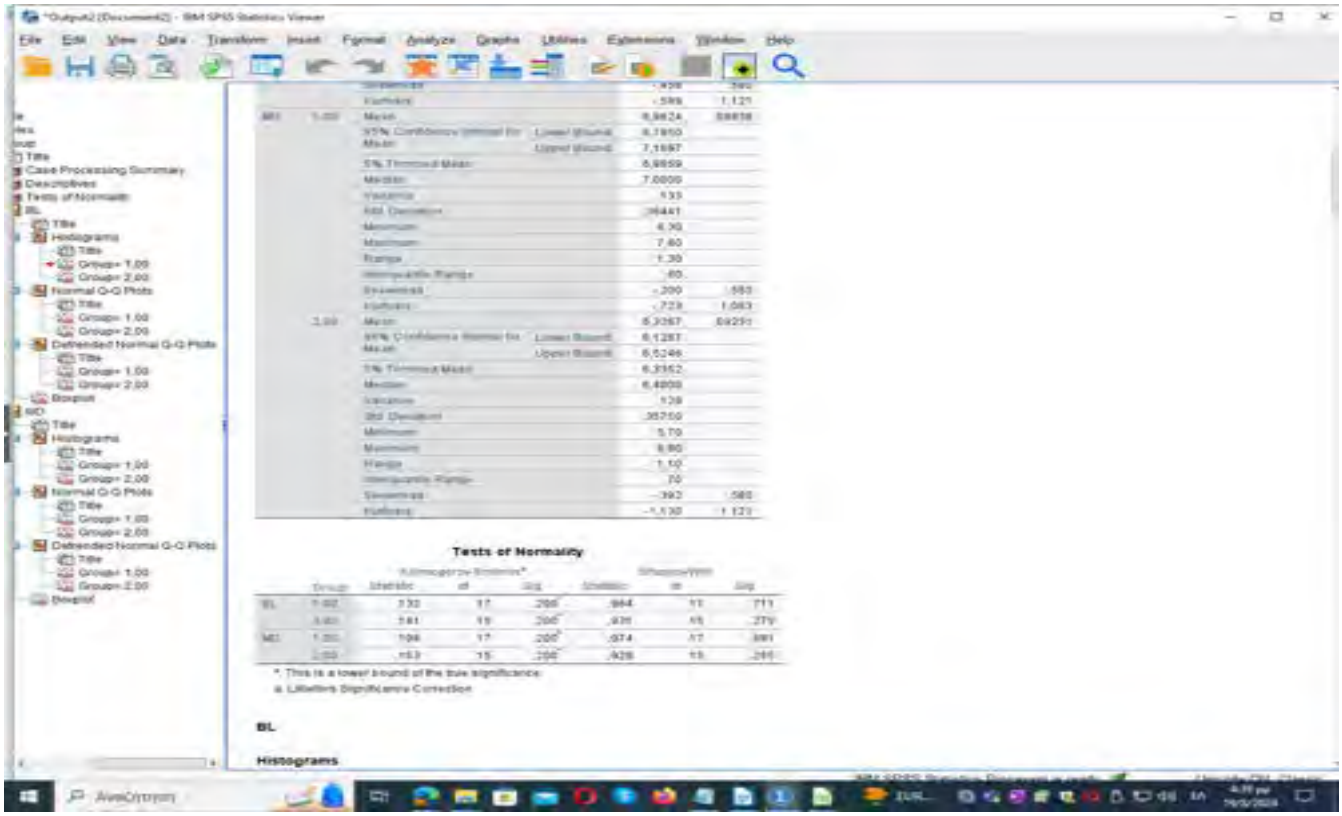


Εικόνα 3.2

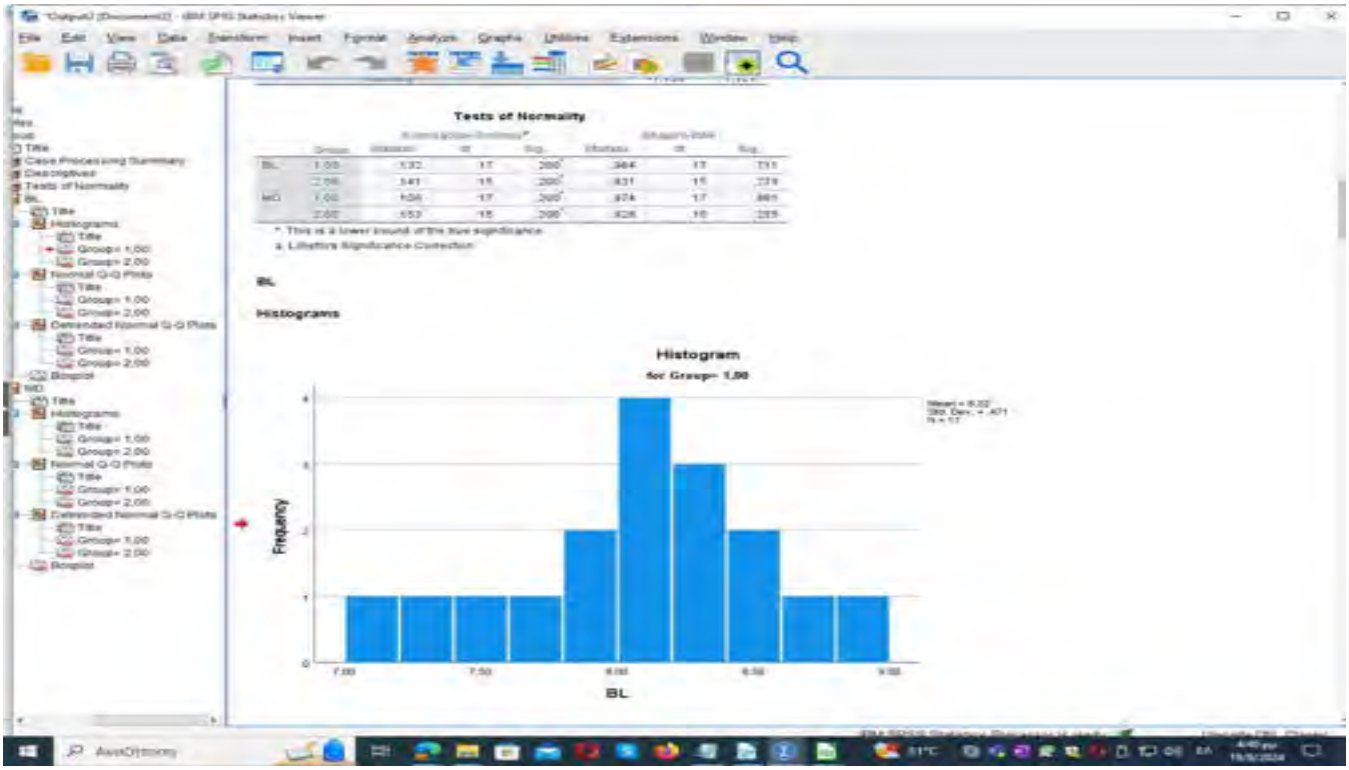
Οπότε με την επιλογή **OK** έχουμε τα παρακάτω αποτελέσματα (Εικόνα 3.3-Εικόνα 3.8):



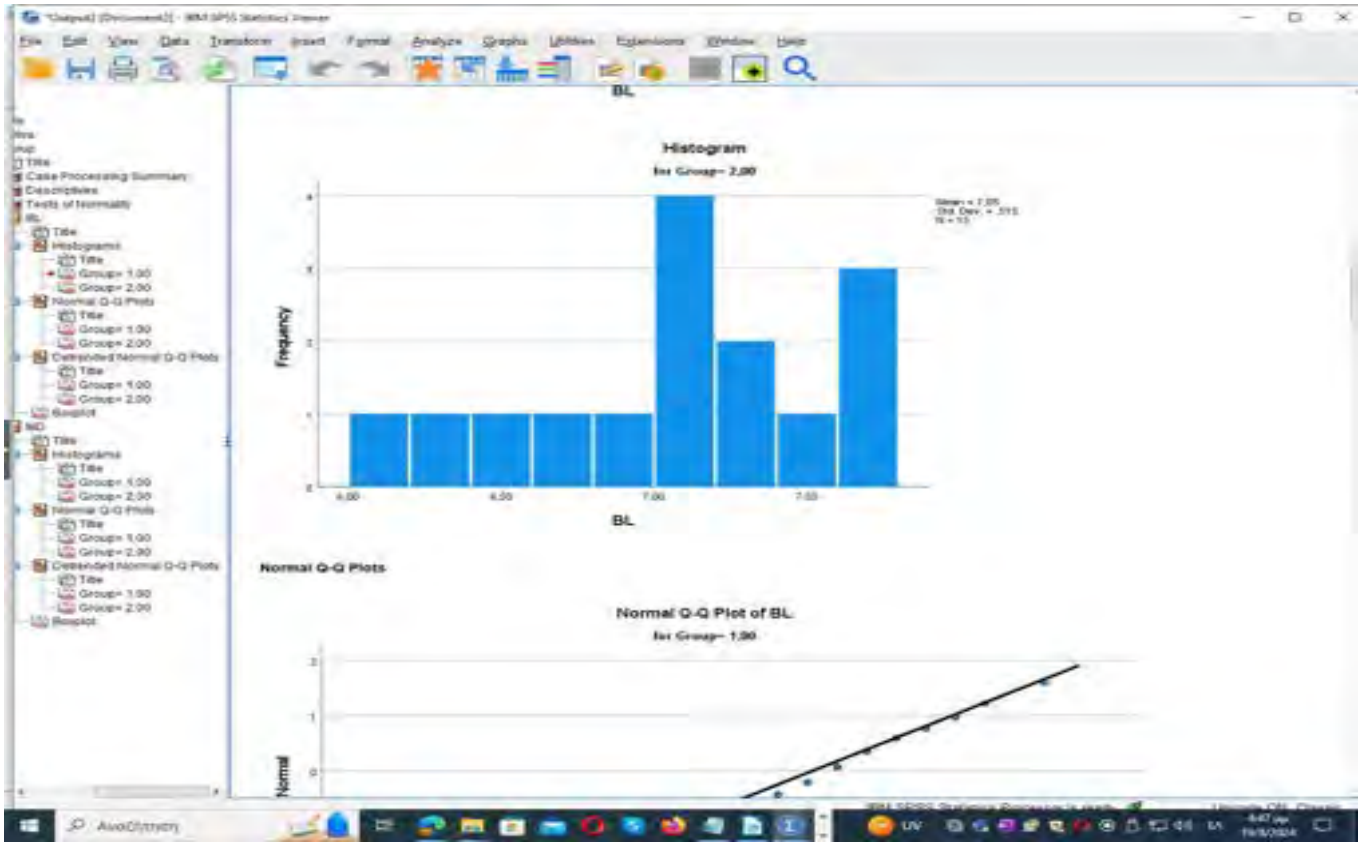
Εικόνα 3.3



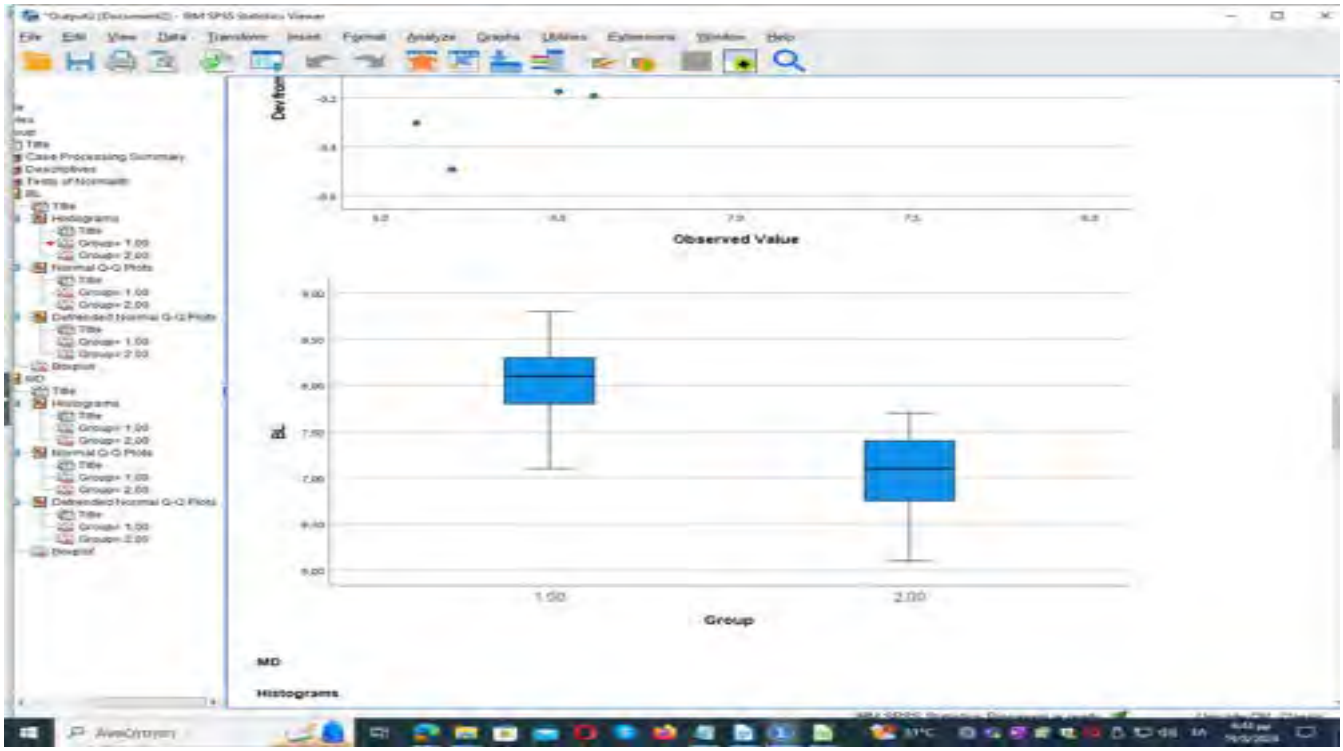
Εικόνα 3.4



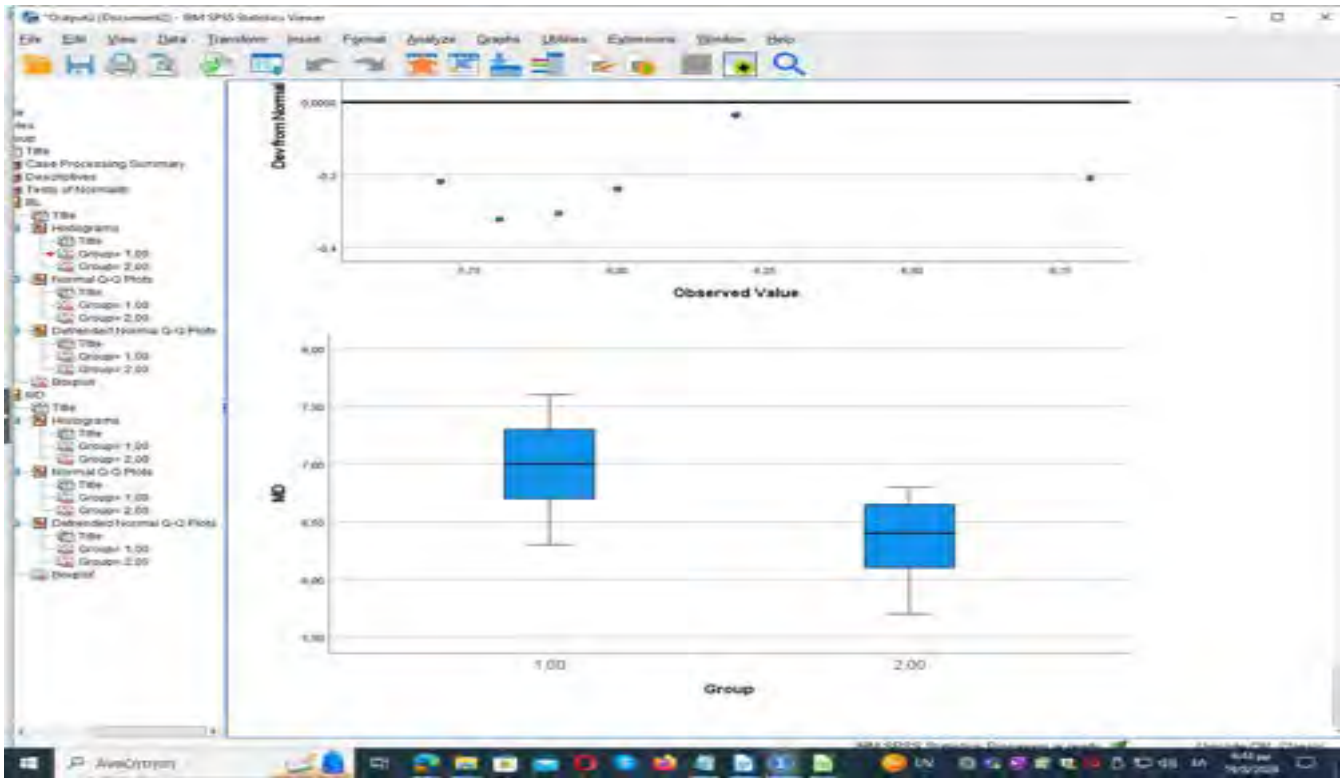
Εικόνα 3.5



Εικόνα 3.6

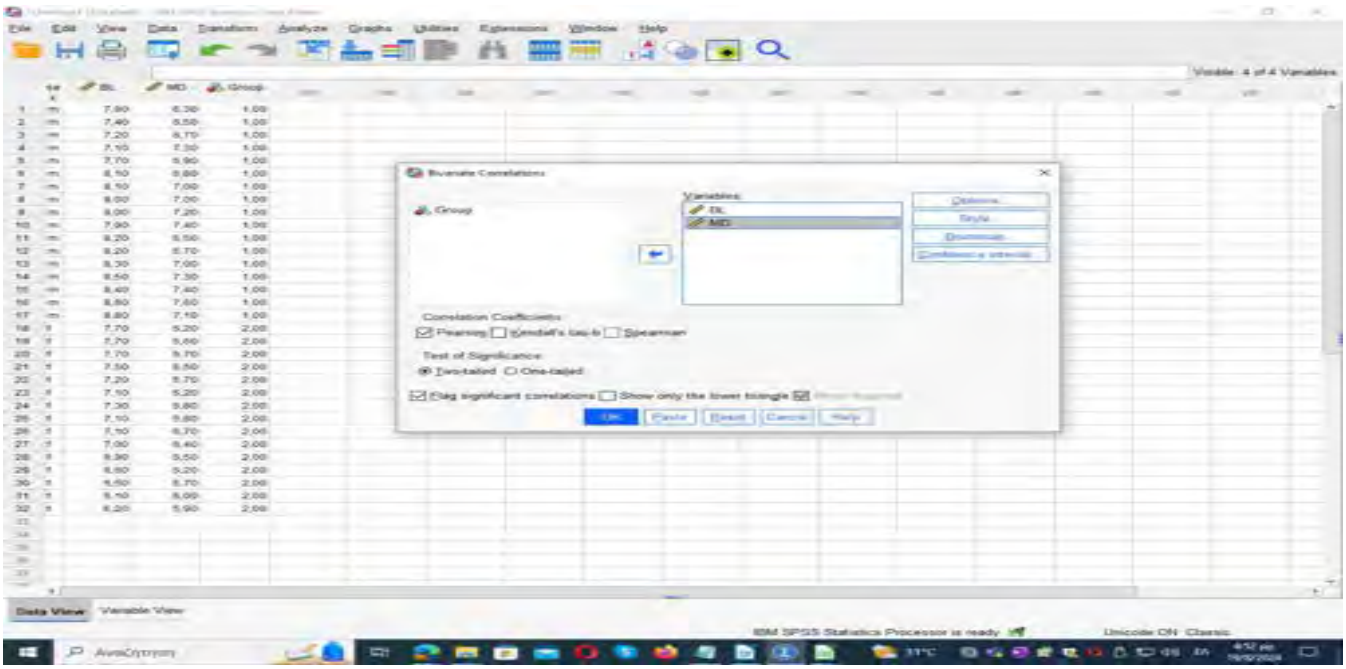


Εικόνα 3.7



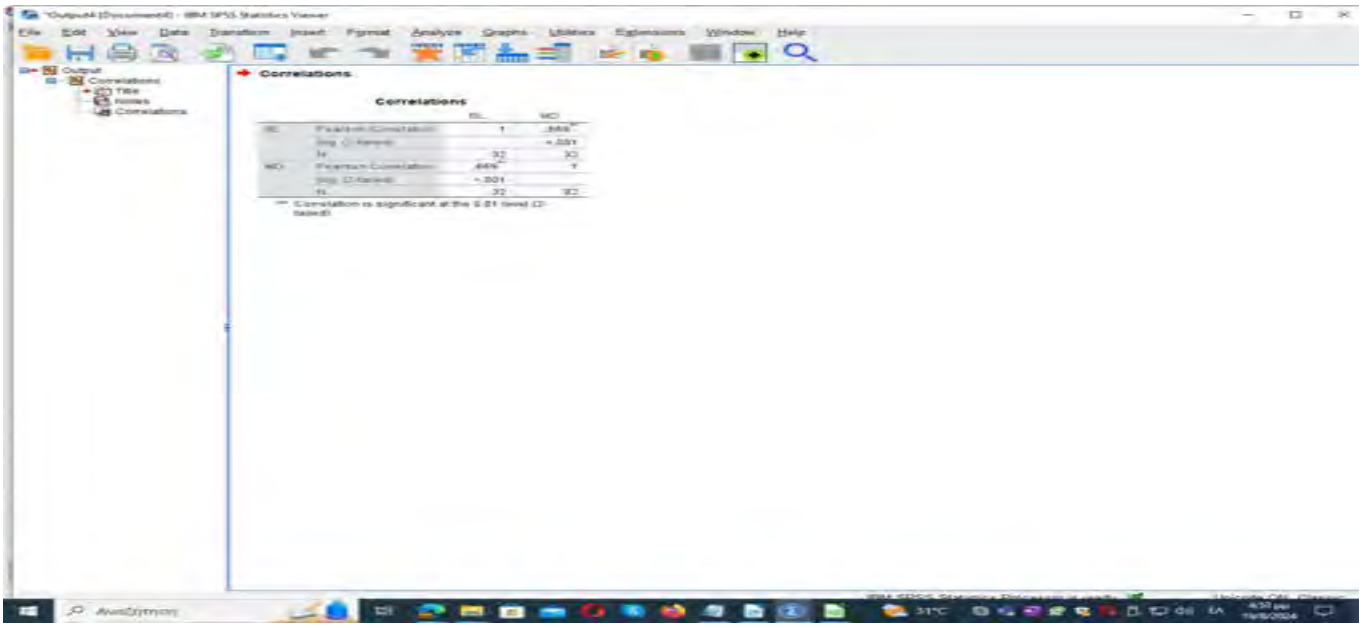
Εικόνα 3.8

Βήμα 3. Κάνουμε έλεγχο πολυσυγγραμικότητας (multicollinearity). Επιλέγουμε: **Analyze** → **Correlate** → **Bivariate** (Εικόνα 3.9).



Εικόνα 3.9

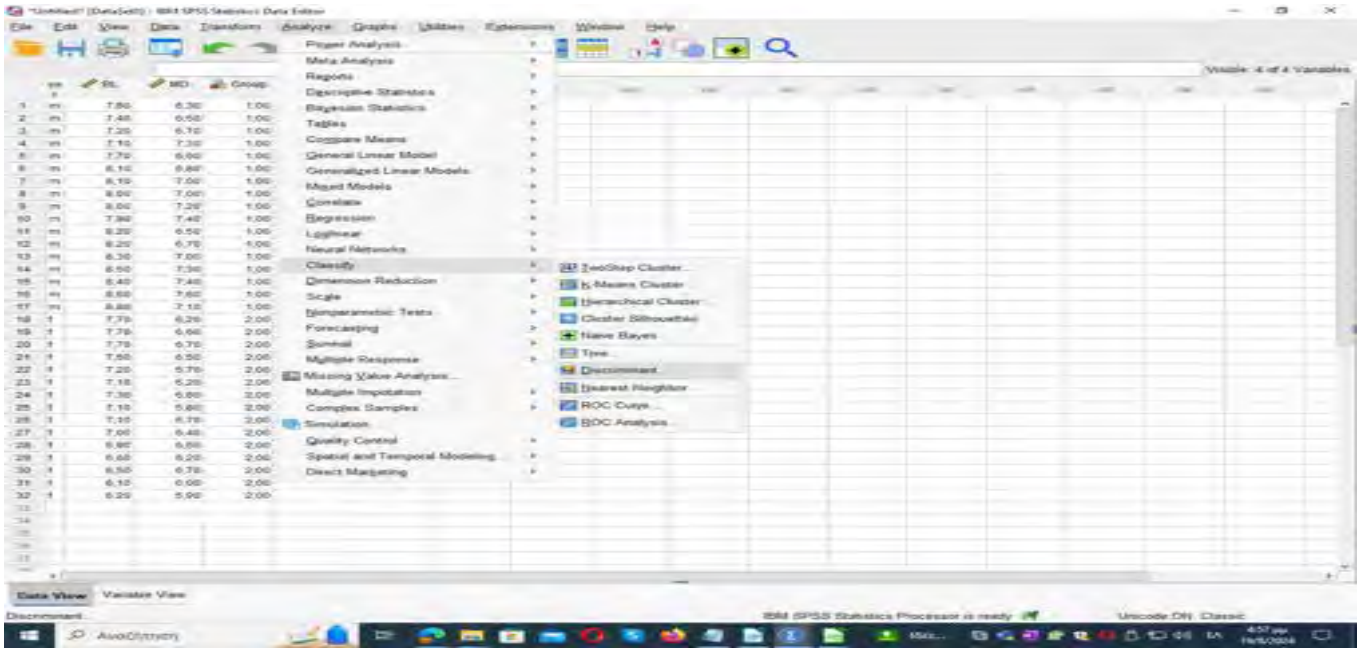
Πατάμε **OK** και έχουμε τα παρακάτω αποτελέσματα (Εικόνα 3.10):



Εικόνα 3.10

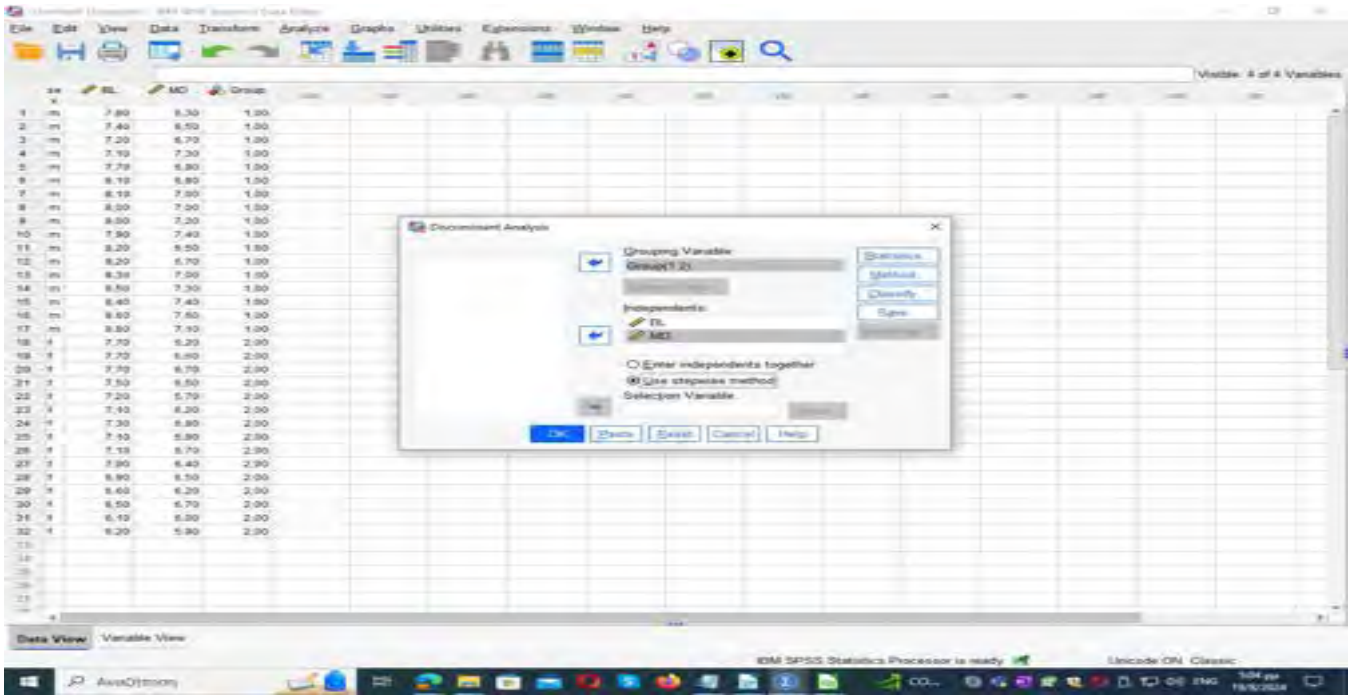
Μία τιμή συσχέτισης κάτω από 0.8 ή 0.9 είναι ικανοποιητική.

Βήμα 4. Στην συνέχεια επιλέγουμε: **Analyze → Classify → Discriminant** (Εικόνα 3.11).



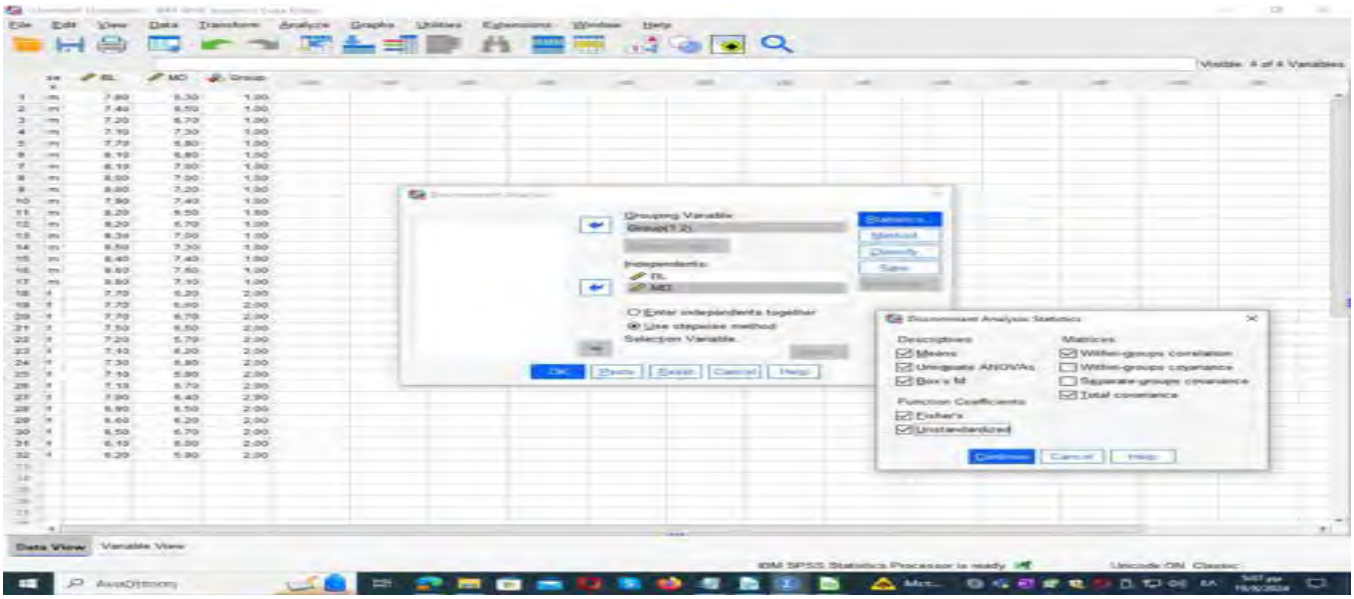
Εικόνα 3.11

Ορίζουμε την εξαρτημένη μεταβλητή **Group** και τις ανεξάρτητες (independents) **(BL,MD)**. Στην συνέχεια πατάμε **Define Range** και εισάγουμε στο **Minimum 1** και στο **Maximum 2,** ανάλογα με τις τιμές (ομάδες) της μεταβλητής **Group** (Εικόνα 3.12).



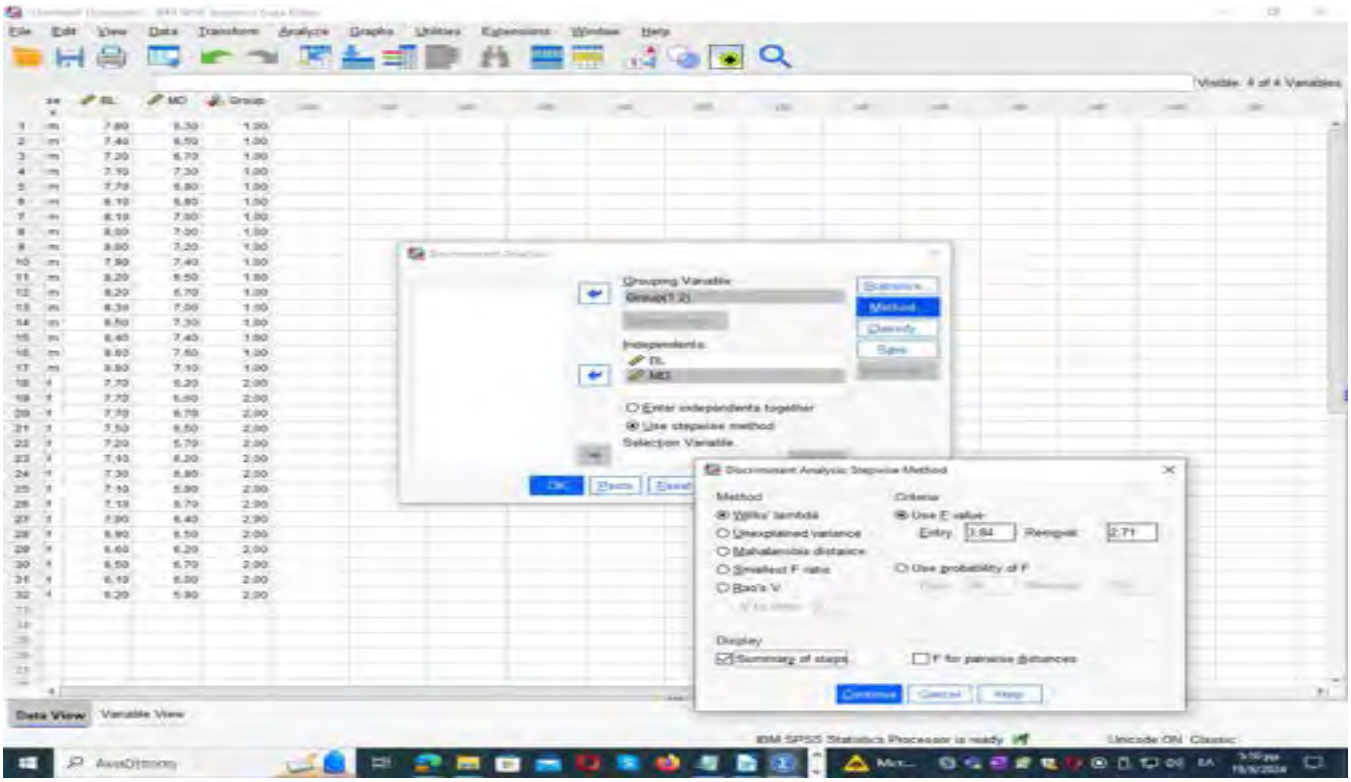
Εικόνα 3.12

Enter independents together: Χρήση όλων των ανεξάρτητων μεταβλητών.
Use stepwise method: Χρήση κλιμακωτών μεθόδων επιλογής ανεξάρτητων μεταβλητών.
Επιλέγουμε **Statistics** και κάνουμε τις ρυθμίσεις όπως στην παρακάτω Εικόνα 3.13:



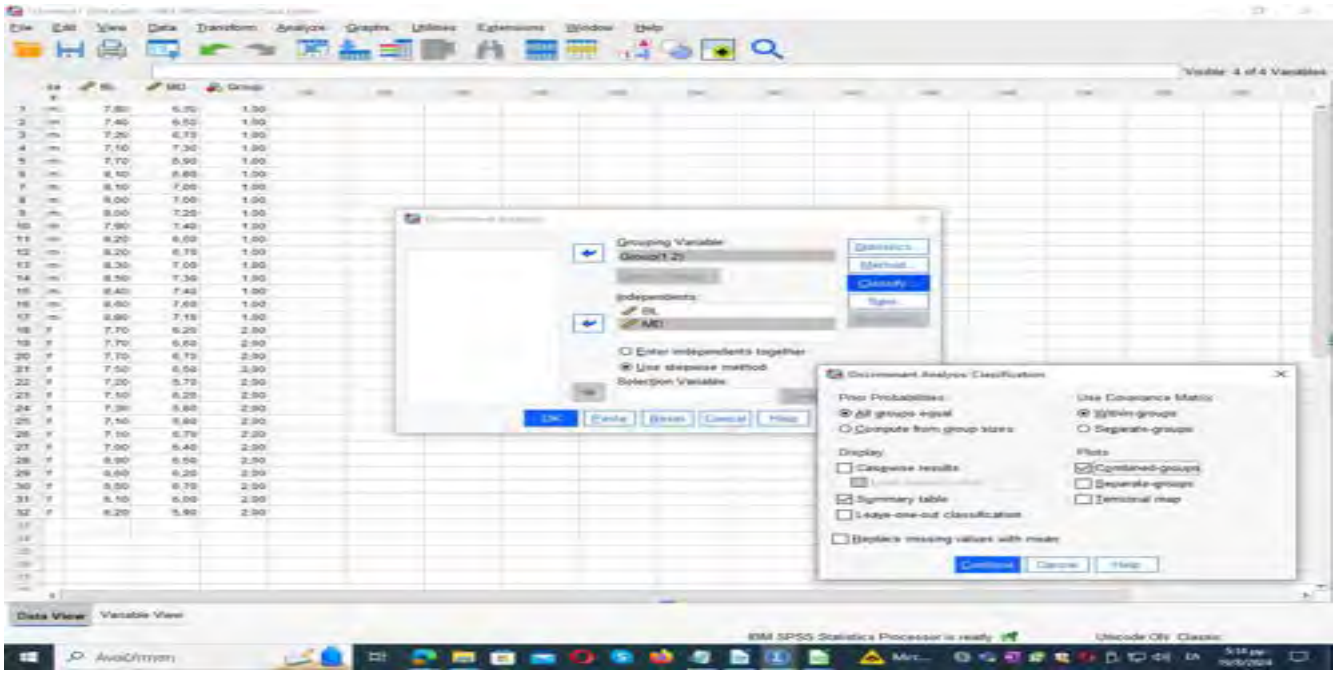
Εικόνα 3.13

Επιλέγουμε **Methods** και κάνουμε τις ρυθμίσεις όπως στην παρακάτω Εικόνα 3.14:



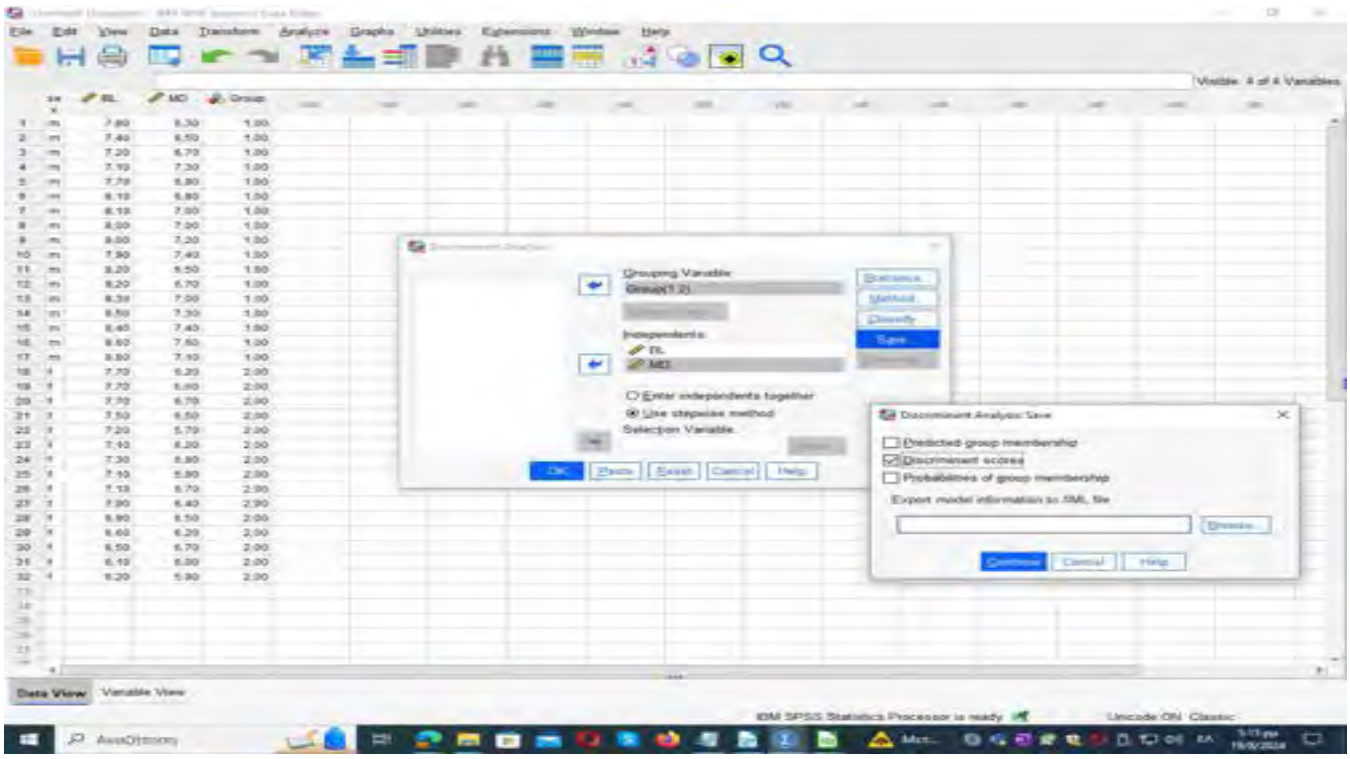
Εικόνα 3.14

Επιλέγουμε **Classify…** και κάνουμε τις ρυθμίσεις όπως στην παρακάτω Εικόνα 3.15:



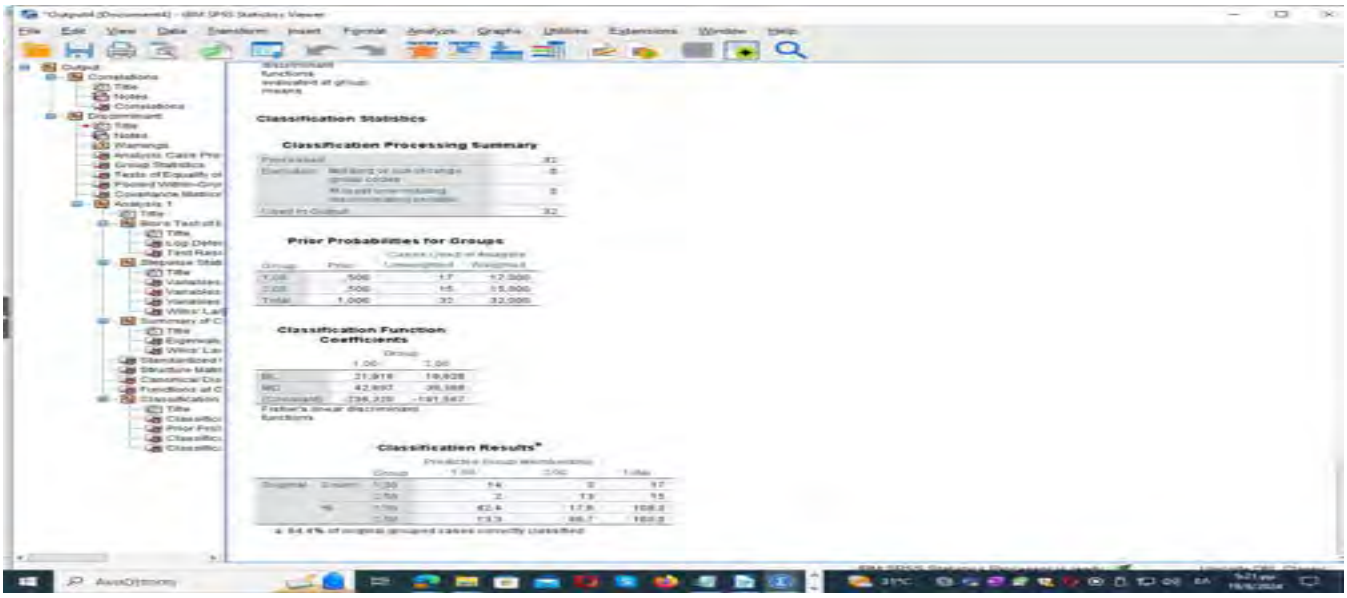
Εικόνα 3.15

Τέλος επιλέγουμε **Save…** και κάνουμε και εδώ τις ρυθμίσεις όπως στην παρακάτω Εικόνα 3.16:



Εικόνα 3.16

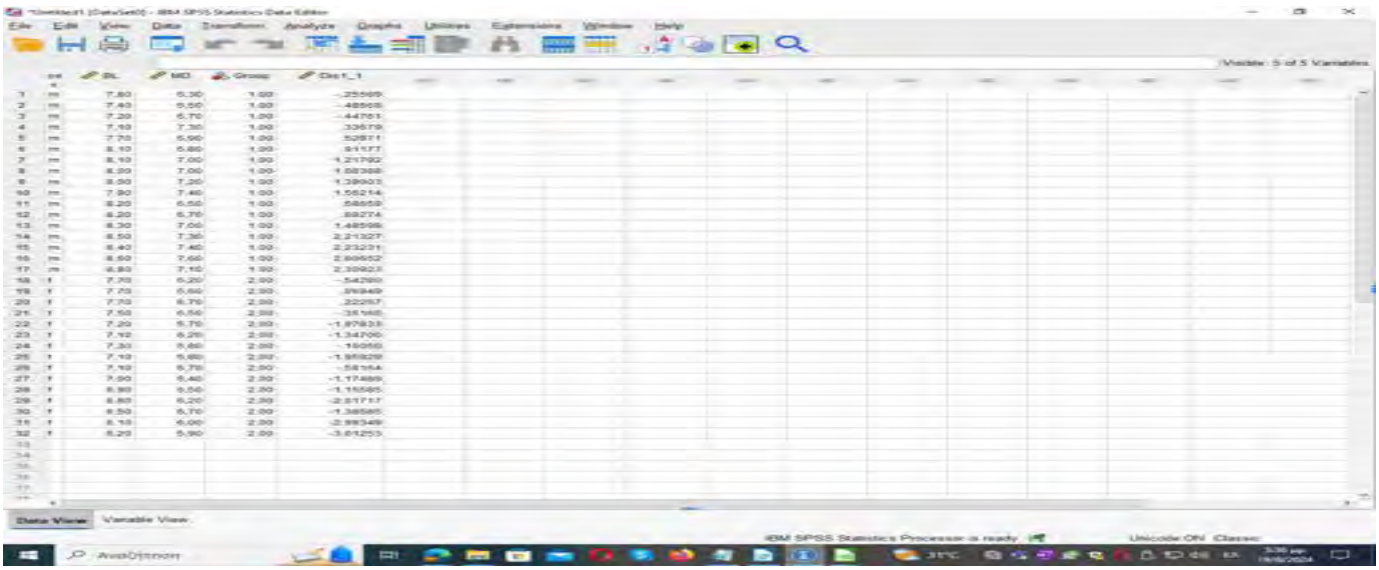
Στην συνέχεια παίρνουμε διάφορα αποτελέσματα με τα πιο σημαντικά να περιγράφονται στον παρακάτω πίνακα της Εικόνας 3.17 (Classification results). Ο πίνακας αυτός δείχνει πόσες από τις αρχικές παρατηρήσεις έχουν ταξινομηθεί σωστά.



Εικόνα 3.17

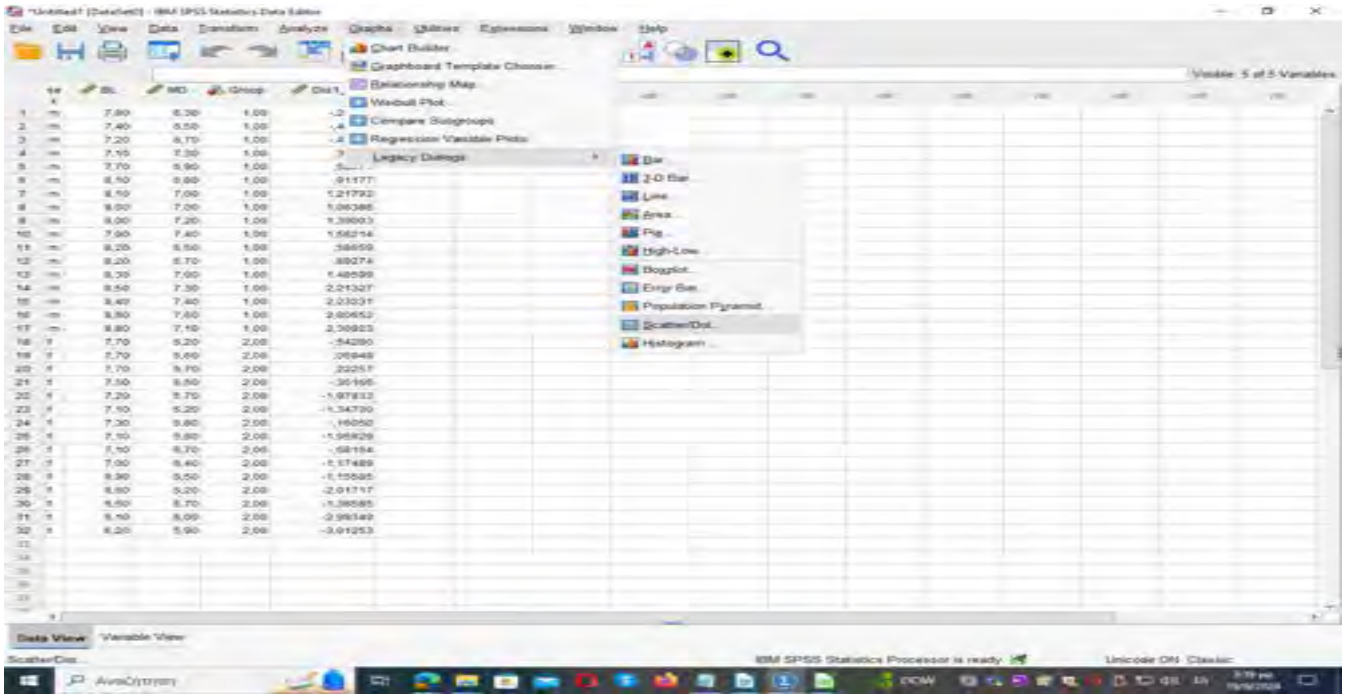
Από τον παραπάνω πίνακα διαπιστώνουμε τα εξής**:**
Από τις 17 παρατηρήσεις της ομάδας 1, οι 14 έχουν καταταγεί στην ομάδα 1 και 3 στην ομάδα 2.
Από τις 15 παρατηρήσεις της ομάδας 2, οι 13 έχουν καταταγεί στην ομάδα 2 και 2 στην ομάδα 1.
Στην ομάδα 1 η ορθή ταξινόμηση είναι 82,4% και στη ομάδα 2 η ορθή ταξινόμηση είναι επίσης 86,7%

Μετά την εκτέλεση της διαχωριστικής ανάλυσης δημιουργήθηκε μια στήλη με το όνομα Dis1_1 που αναφέρει διακριτικά score (Εικόνα 3.18):



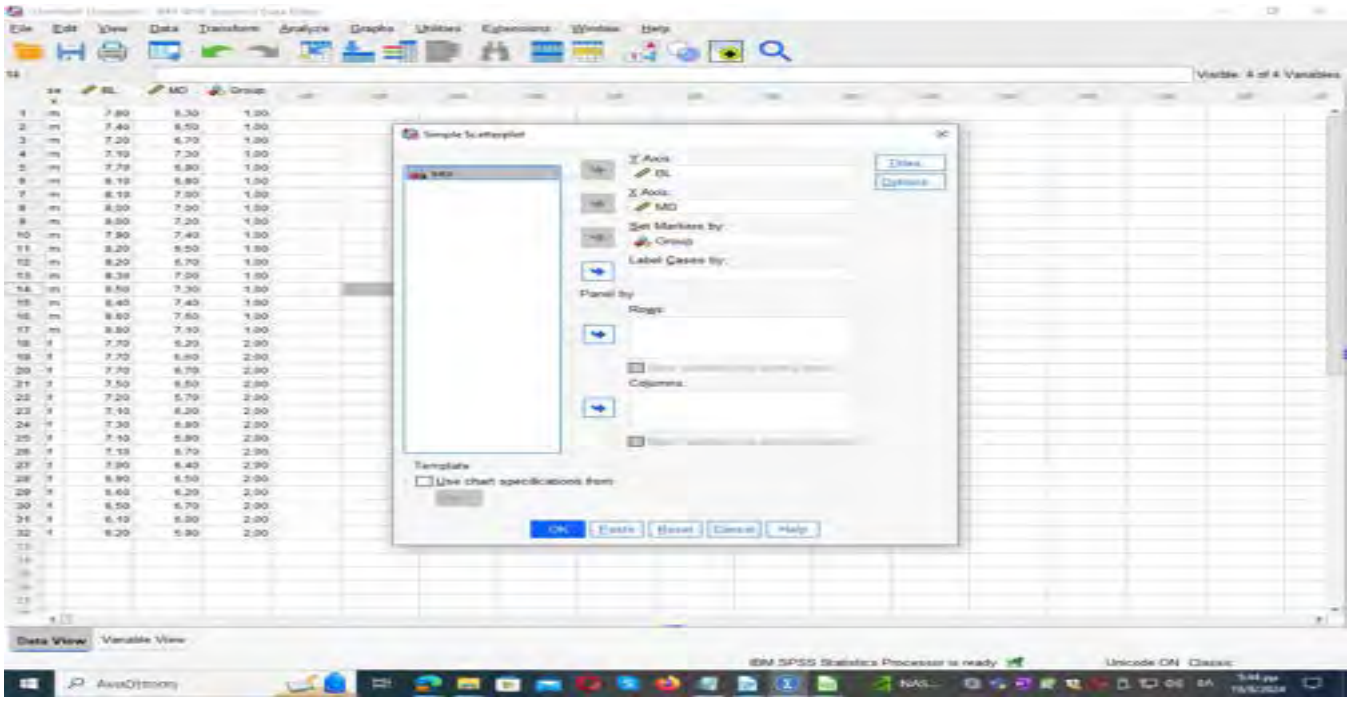
Εικόνα 3.18

Παρατήρηση. Αν θέλουμε να δούμε μια γεωμετρική εικόνα των παραπάνω βλέπουμε την Εικόνα 3.19:

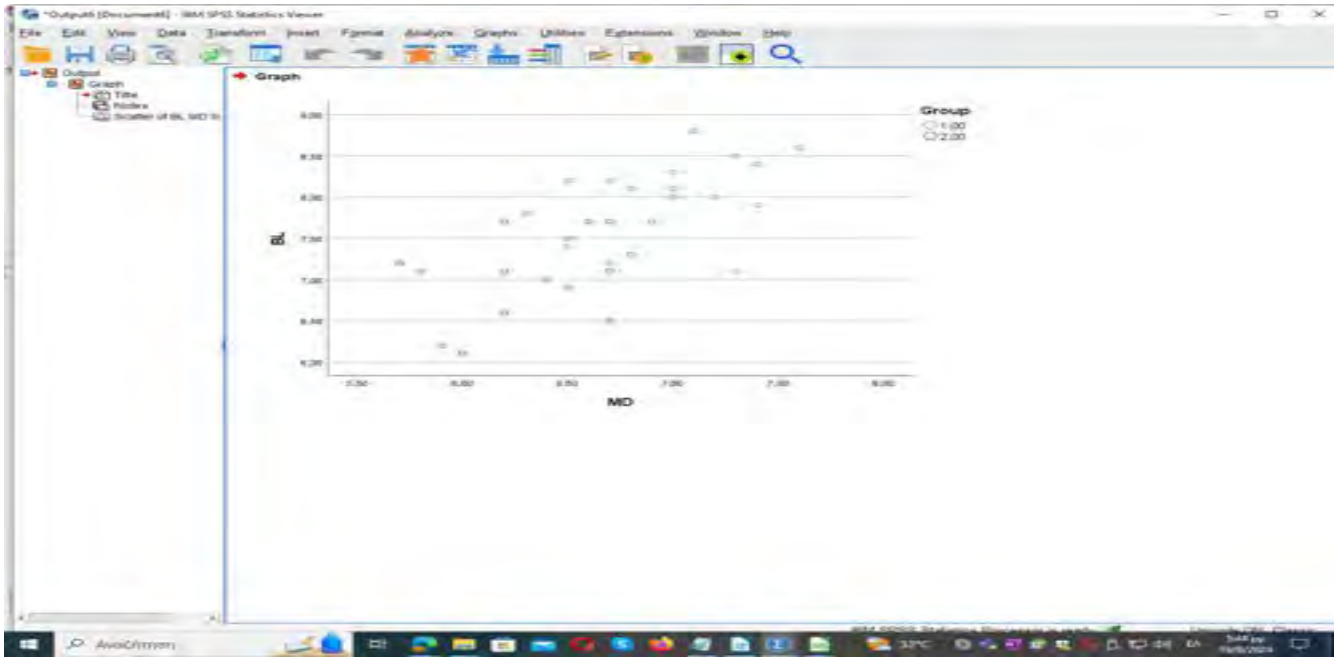


Εικόνα 3.19

Οπότε παίρνουμε τα παρακάτω αποτελέσματα (Εικόνα 3.20-Εικόνα 3.21):



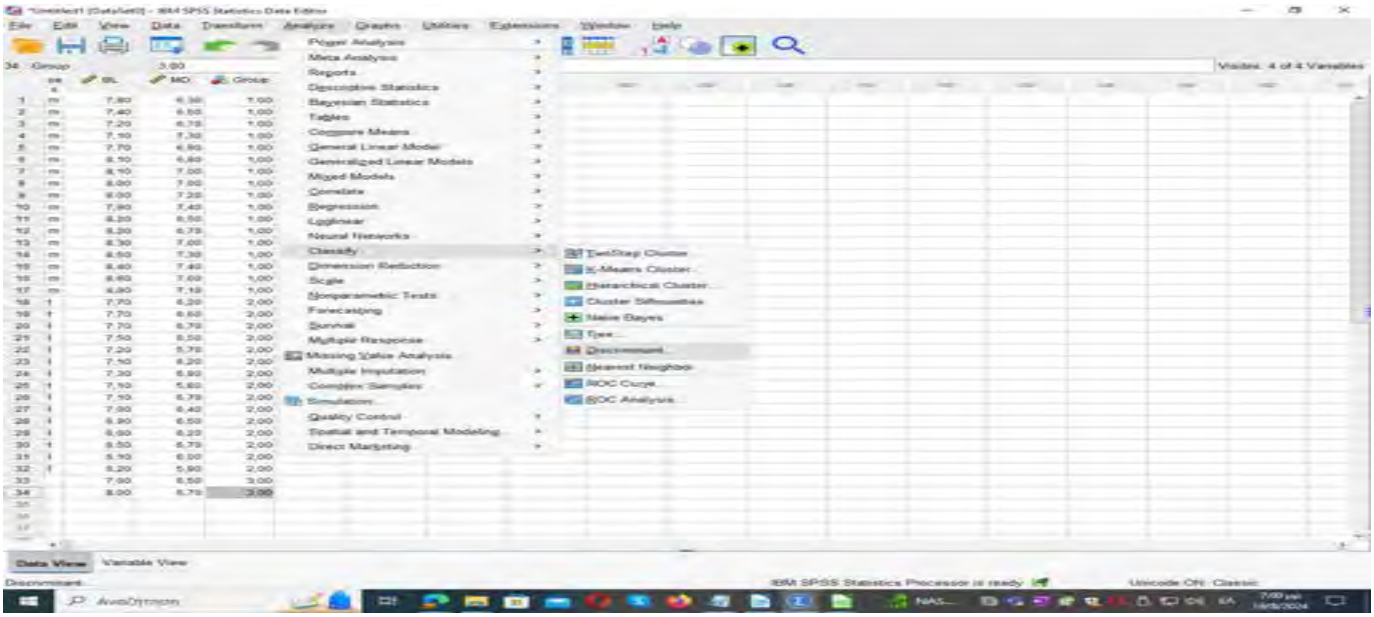
Εικόνα 3.20



Εικόνα 3.21

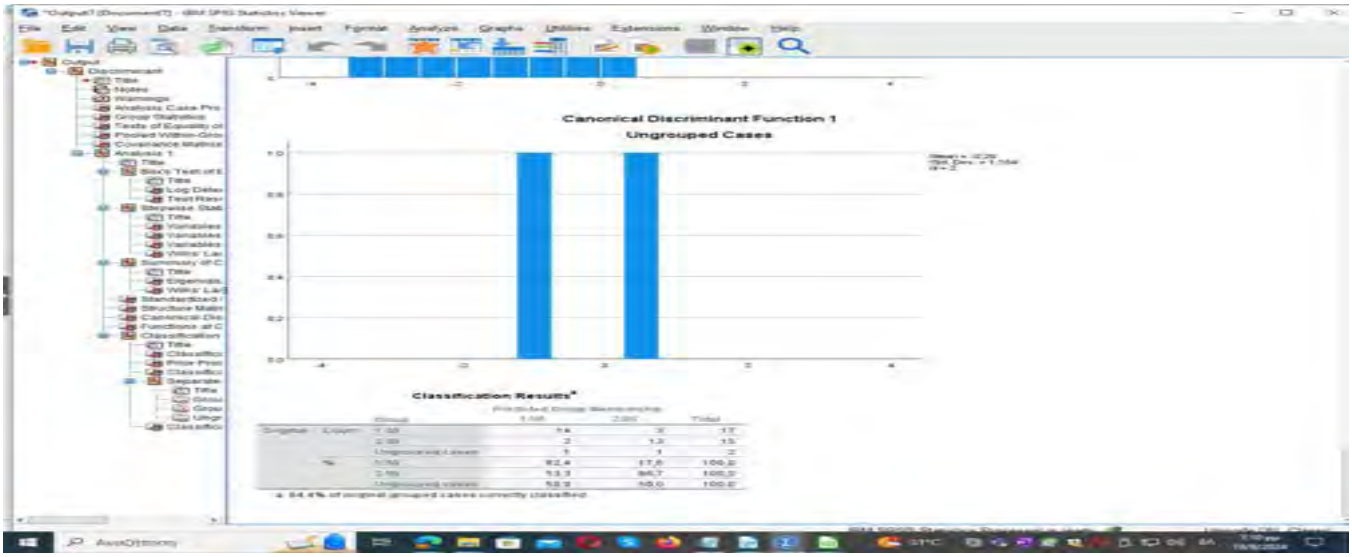
Σε δεύτερη φάση τώρα θα προσπαθήσουμε να προσδιορίσουμε αν τα δείγματα (BL,MD)=(8, 6.7) και (BL, MD)= (7, 6.5) ανήκουν σε άνδρα και γυναίκα. Για την λύση αυτού του προβλήματος εργαζόμαστε όπως παρακάτω.

Προσθέτουμε στον πίνακα του SPSS τα δυο δείγματα που έχουμε (BL,MD)=(8, 6.7) και (BL, MD)= (7, 6.5) και στην στήλη Group βάζουμε το νούμερο 3 όπως φαίνεται στον παρακάτω πίνακα. Στην συνέχεια ακολουθούμε την διαδικασία που φαίνεται στην παρακάτω Εικόνα 3.22:

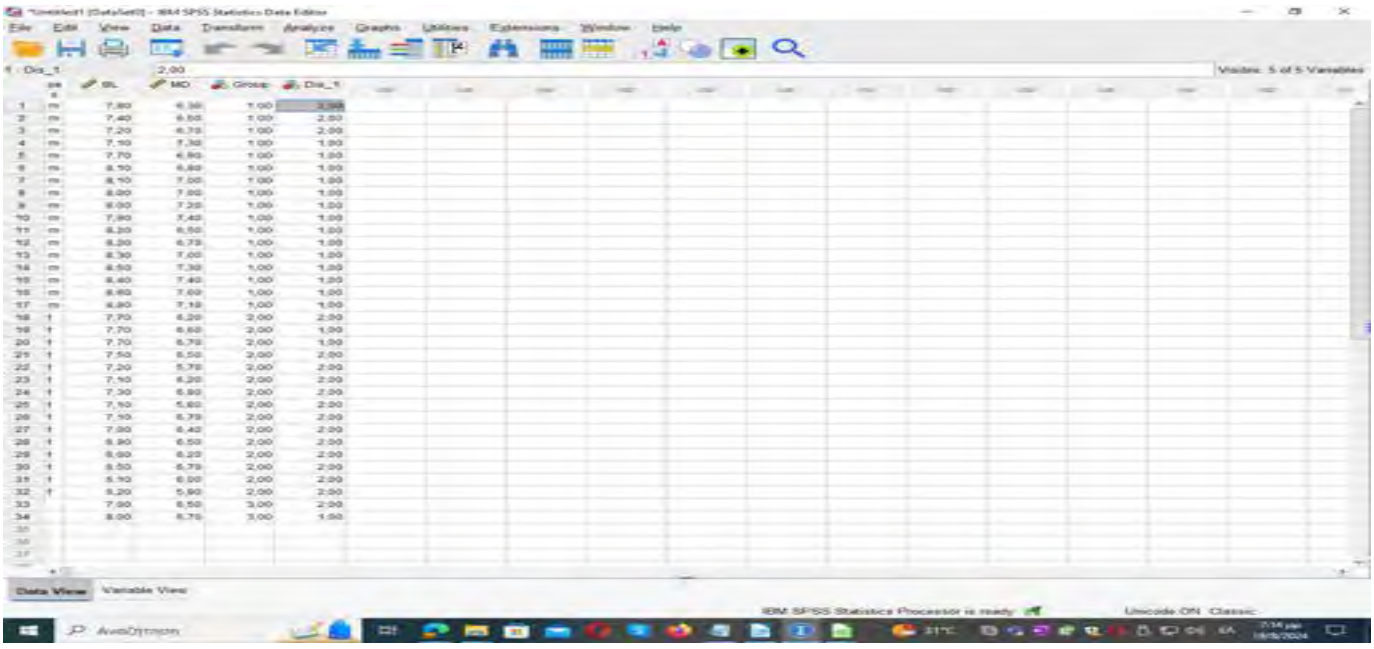


Εικόνα 3.22

Οπότε ακολουθούμε: **Analyze** → **Classify** → **Discriminant** και στο παράθυρο που ανοίγει μεταφέρουμε τις μεταβλητές BL, MD στο πλαίσιο **Independent** και τη μεταβλητή Group στο **Grouping Variable**. Με κλικ στο **Define Range** εισάγουμε στο **Minimum** την τιμή 1 και στο **Maximum** την τιμή 2 (όχι την 3). Κάνουμε κλικ στο **Continue** και στο **Save** επιλέγουμε Predicted group membership και Probabilities of group membership. Επίσης, στο **Classify** επιλέγουμε το Summary table και ολοκληρώνουμε με κλικ στο **Continue** και στο **OK**. Από τους πίνακες που παίρνουμε ενδιαφέρον παρουσιάζει ο Classification Results και η στήλη στα δίπλα από τα αρχικά δεδομένα μας (Εικόνα 3.23-Εικόνα 3.24).



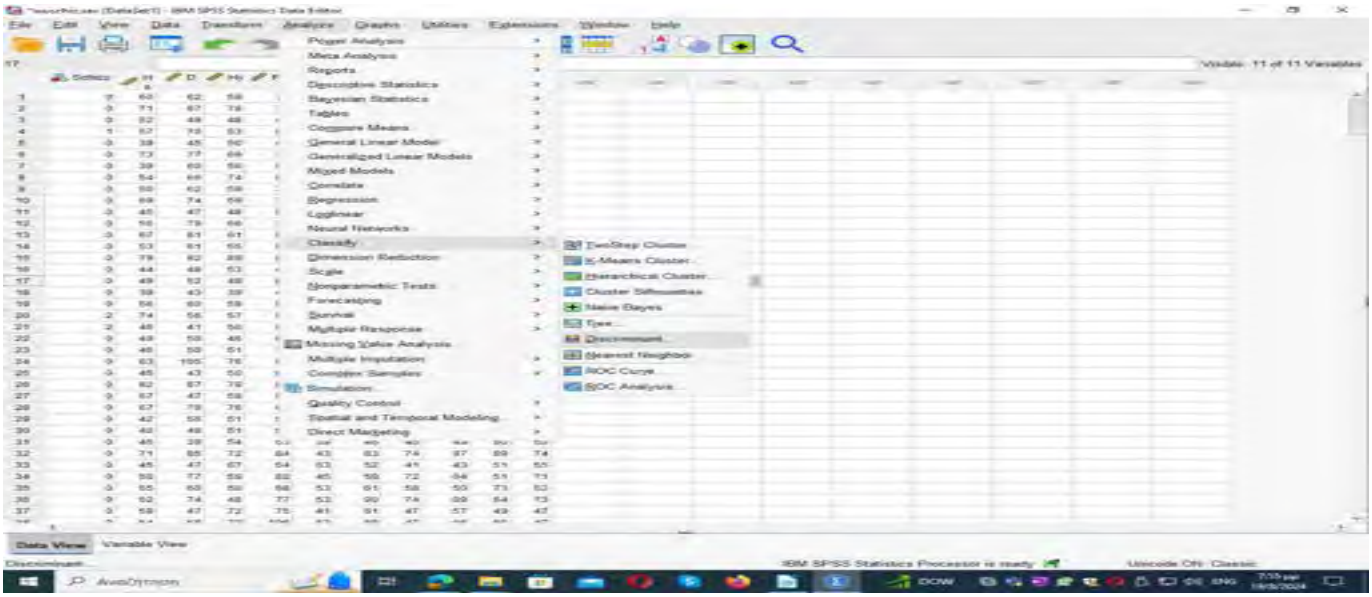
Εικόνα 3.23



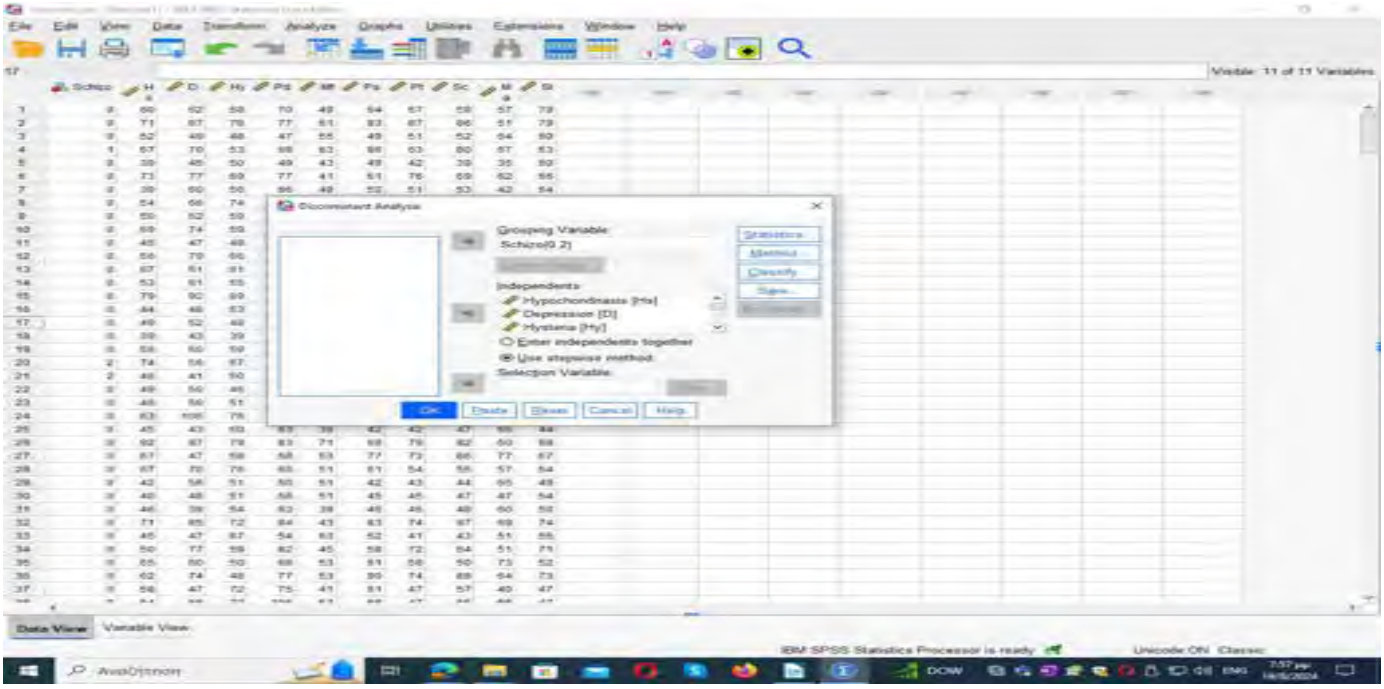
Εικόνα 3.24

Όπως βλέπουμε παραπάνω το πρώτο άγνωστο δείγμα ανήκει σε γυναίκα και το δεύτερο άγνωστο δείγμα ανήκει σε άνδρα.

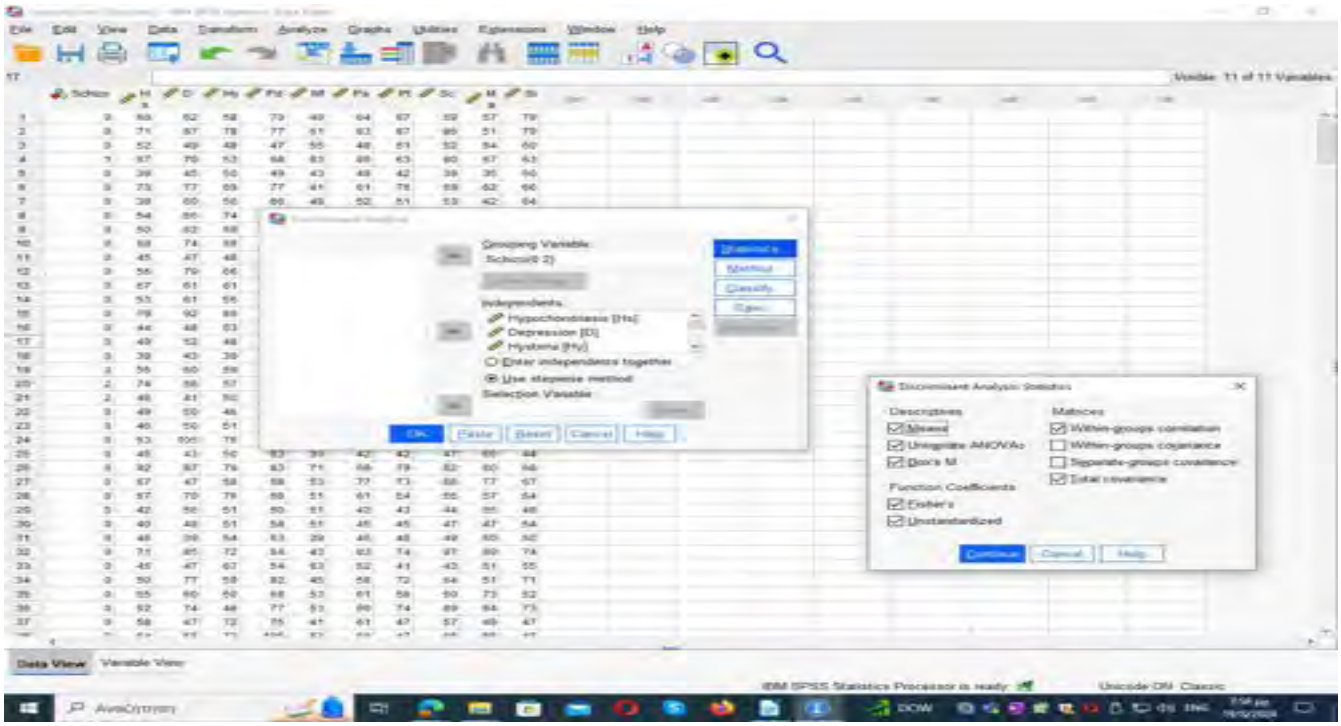
Παράδειγμα 2. Το παράδειγμα που ακολουθεί αφορά την εκτίμηση του επιπέδου της σχιζοφρένειας σε μια κλινική. Για τις τιμές της μεταβλητής σχιζοφρένειας έχουμε:
0 = consensus diagnosis outside of the schizophrenia spectrum
1 = consensus diagnosis of schizophrenia spectrum disorder
2 = consensus diagnosis of definite of probable schizophrenia
Ακολουθούμε την διαδικασία που είδαμε στο πρώτο παράδειγμα και έχουμε τις Εικόνες 3.25- 3.31:



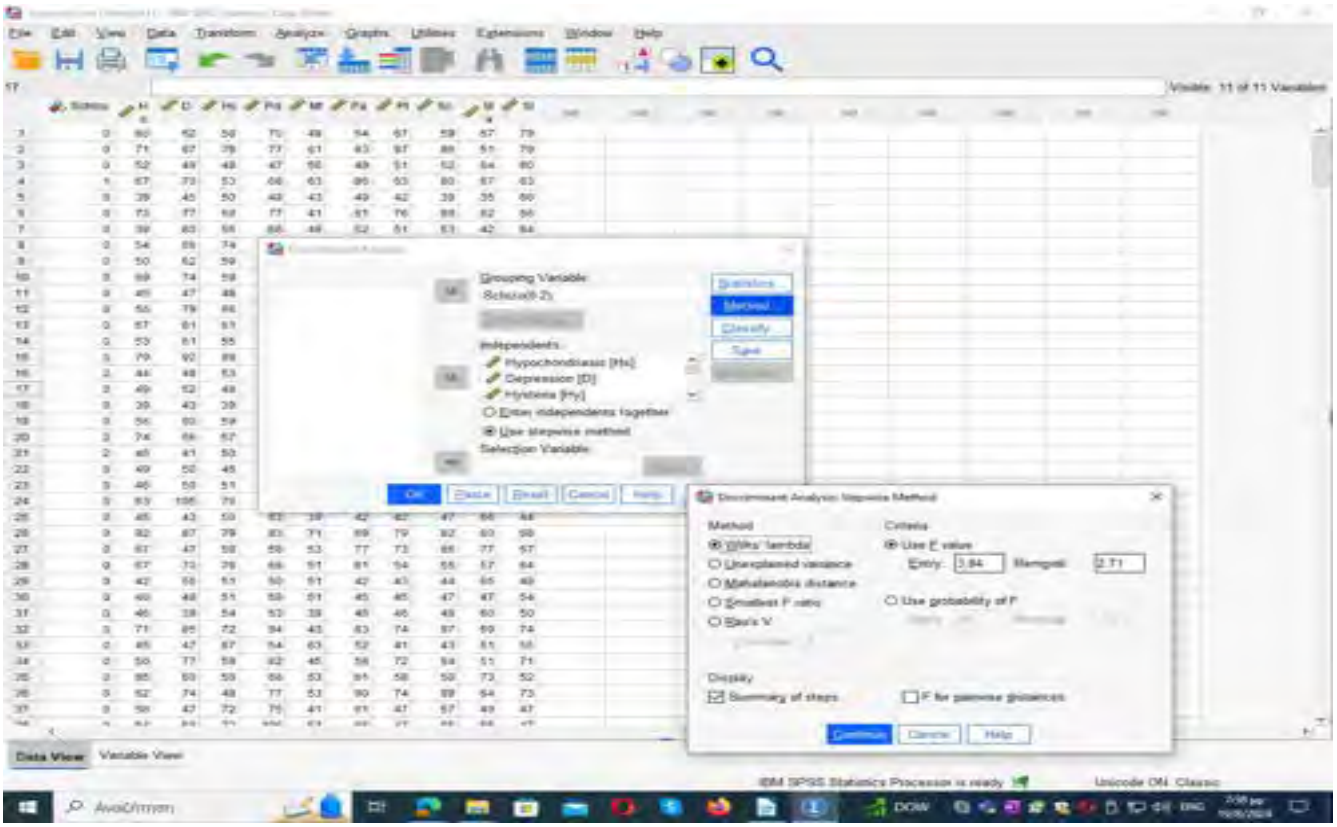
Εικόνα 3.25



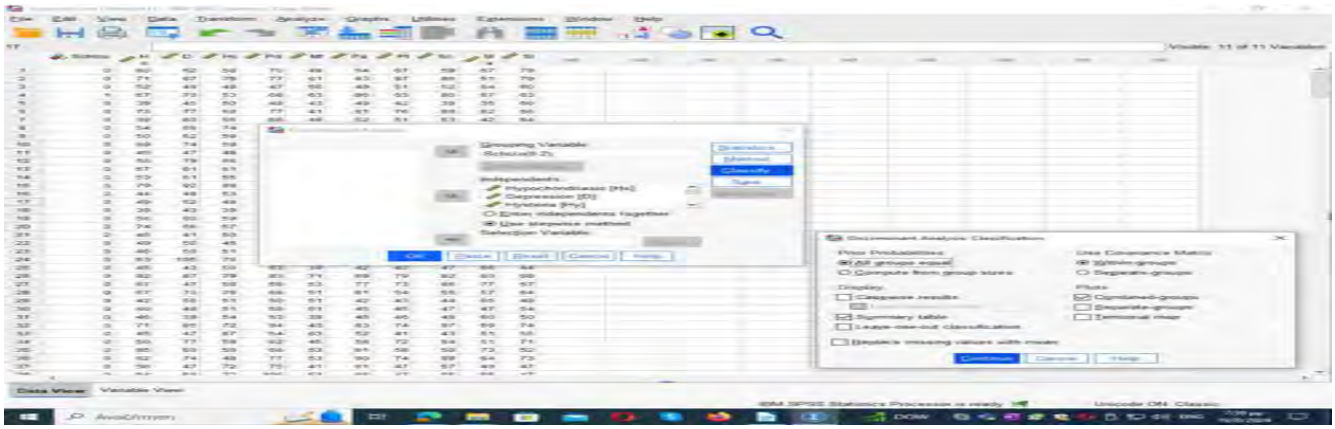
Εικόνα 3.26



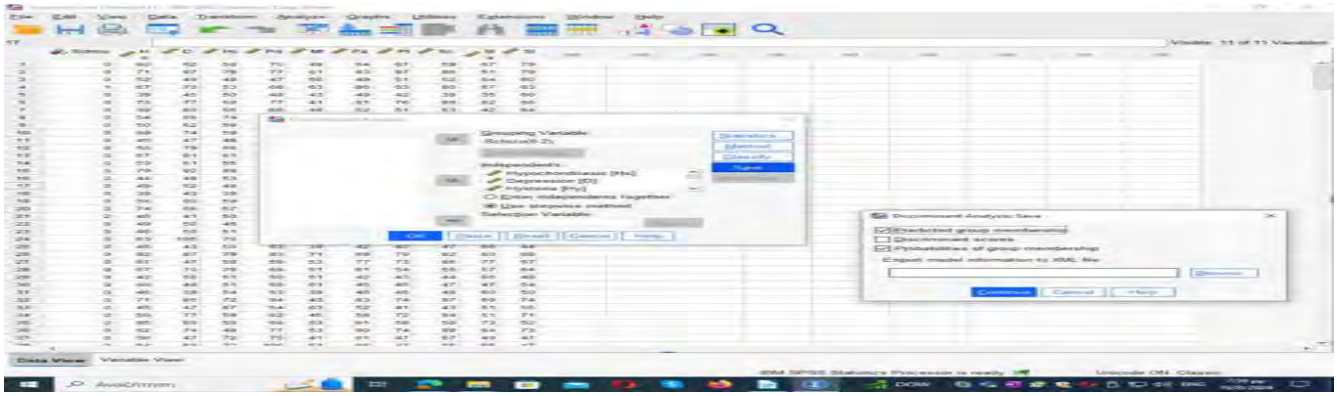
Εικόνα 3.27



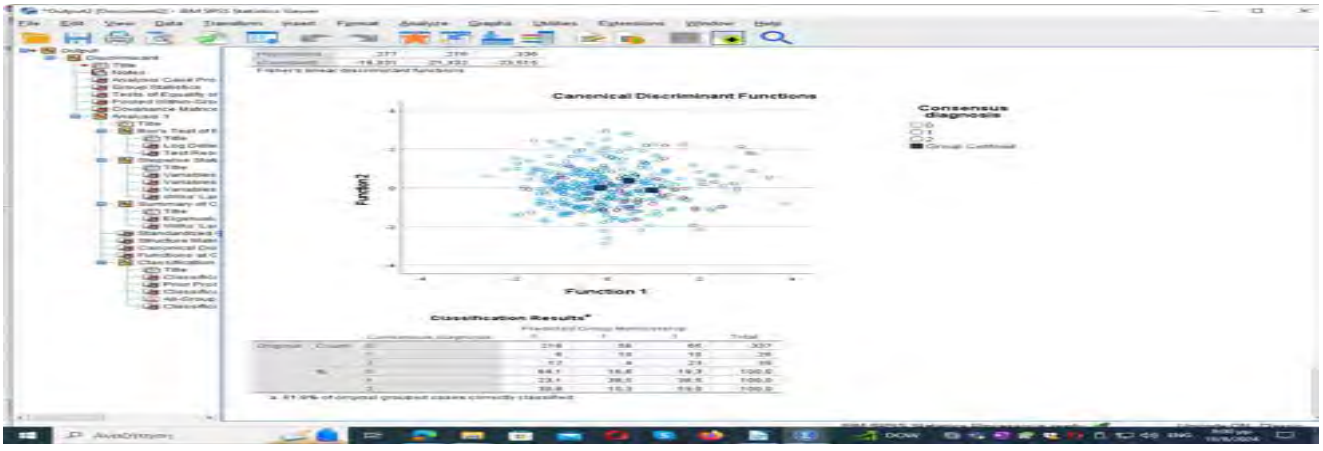
Εικόνα 3.28



Εικόνα 3.29



Εικόνα 3.30



Εικόνα 3.31

Ο παραπάνω πίνακας πίνακας δείχνει πόσες από τις αρχικές παρατηρήσεις έχουν ταξινομηθεί σωστά. Συγκεκριμένα από τις 337 παρατηρήσεις της ομάδας 0, 56 έχουν καταταγεί στην ομάδα 1 και 65 στην ομάδα 2. Από τις 26 παρατηρήσεις της ομάδας 1, 16 δεν έχουν ταξινομηθεί σωστά, ενώ από τις 39 παρατηρήσεις της ομάδας 3, οι 23 έχουν ταξινομηθεί σωστά. Στην ομάδα 0 η ορθή ταξινόμηση είναι 64.1%, στην ομάδα 1 η ορθή ταξινόμηση είναι 38.5% και στη ομάδα 2 η ορθή ταξινόμηση είναι επίσης 59.0 %.

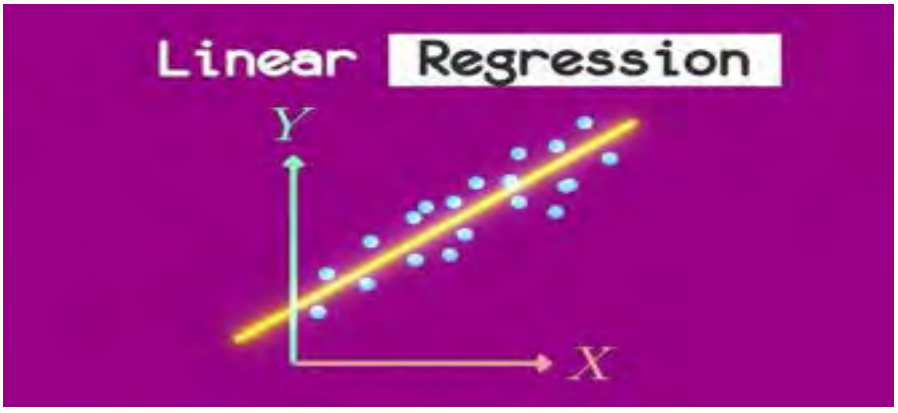
ΚΕΦΑΛΑΙΟ 4ο
Η Πολλαπλή Παλινδρόμηση στο πρόγραμμα SPSS

Στο κεφάλαιο αυτό παρουσιάζεται η γραμμική παλινδρόμηση και το πλαίσιο γύρω από το οποίο μπορούμε να αναπτύξουμε τη συλλογιστική μας προκειμένου να εξάγουμε ασφαλή και σημαντικά συμπεράσματα για ποικίλα ζητήματα της Ιατρικής Επιστήμης. Επιπλέον, η γραμμική παλινδρόμηση μέσα από το στατιστικό πακέτο SPSS προσδίδει στη μελέτη αυτού του κεφαλαίου μία διαφορετική οπτική στην ιατρική έρευνα.

4.1 Η Γραμμική Παλινδρόμηση - Γενικά

Η γραμμική παλινδρόμηση είναι ένας αλγόριθμος που παρέχει μια γραμμική σχέση μεταξύ μιας ανεξάρτητης μεταβλητής και μιας εξαρτημένης μεταβλητής για την πρόβλεψη του αποτελέσματος μελλοντικών γεγονότων. Είναι μια στατιστική μέθοδος που χρησιμοποιείται στην επιστήμη δεδομένων και τη μηχανική μάθηση για προγνωστική ανάλυση.

Η ανεξάρτητη μεταβλητή είναι επίσης η προγνωστική ή επεξηγηματική μεταβλητή που παραμένει αμετάβλητη λόγω της αλλαγής σε άλλες μεταβλητές. Ωστόσο, η εξαρτημένη μεταβλητή αλλάζει με τις διακυμάνσεις της ανεξάρτητης μεταβλητής. Το μοντέλο παλινδρόμησης προβλέπει την τιμή της εξαρτημένης μεταβλητής, η οποία είναι η μεταβλητή απόκρισης ή αποτελέσματος που αναλύεται ή μελετάται. Έτσι, η γραμμική παλινδρόμηση είναι ένας εποπτευόμενος αλγόριθμος μάθησης που προσομοιώνει μια μαθηματική σχέση μεταξύ μεταβλητών και κάνει προβλέψεις για συνεχείς ή αριθμητικές μεταβλητές όπως πωλήσεις, μισθός, ηλικία, τιμή προϊόντος κ.λπ. Μια κεκλιμένη ευθεία αντιπροσωπεύει το μοντέλο γραμμικής παλινδρόμησης (Εικόνα 4.1).



Εικόνα 4.1

Στο παρεαπάνω σχήμα το Χ είναι η ανεξάρτητη μεταβλητή και το Υ η εξαρτημένη μεταβλητή. Γενικά η γραμμική παλινδρόμηση είναι ένας τύπος εποπτευόμενου αλγορίθμου μηχανικής μάθησης που υπολογίζει τη γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής και ενός ή περισσοτέρων ανεξάρτητων χαρακτηριστικών προσαρμόζοντας μια γραμμική εξίσωση στα παρατηρούμενα δεδομένα. Όταν υπάρχει μόνο ένα ανεξάρτητο χαρακτηριστικό, είναι γνωστή ως Απλή Γραμμική Παλινδρόμηση και όταν υπάρχουν περισσότερα από ένα χαρακτηριστικά, είναι γνωστή ως Πολλαπλή Γραμμική Παλινδρόμηση.

Η εξίσωση του μοντέλου της Γραμμικής Παλινδρόμησης παρέχει σαφείς συντελεστές που διευκρινίζουν την επίδραση κάθε ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή, διευκολύνοντας τη βαθύτερη κατανόηση της υποκείμενης δυναμικής. Η απλότητά της είναι μια αρετή, καθώς η γραμμική παλινδρόμηση είναι διαφανής, εύκολη στην εφαρμογή και χρησιμεύει ως θεμελιώδης ιδέα για πιο σύνθετους αλγόριθμους. Η γραμμική παλινδρόμηση δεν είναι απλώς ένα εργαλείο πρόβλεψης. αποτελεί τη βάση για διάφορα προηγμένα μοντέλα. Επιπλέον, η γραμμική παλινδρόμηση είναι ένας ακρογωνιαίος λίθος στη δοκιμή υποθέσεων, επιτρέποντας στους ερευνητές να επικυρώσουν βασικές υποθέσεις σχετικά με τα δεδομένα.

Απλή Γραμμική Παλινδρόμηση

Αυτή είναι η απλούστερη μορφή γραμμικής παλινδρόμησης και περιλαμβάνει μόνο μία ανεξάρτητη μεταβλητή και μία εξαρτημένη μεταβλητή. Η εξίσωση για την απλή γραμμική παλινδρόμηση είναι:

$$\mathbf{Y=\beta_0+\beta_1 X},$$

όπου το Υ είναι η εξαρτημένη μεταβλητή, το Χ είναι η ανεξάρτητη μεταβλητή,το β_0 είναι η σταθερά και το β_1 είναι η κλίση της ευθείας.

Πολλαπλή Γραμμική Παλινδρόμηση

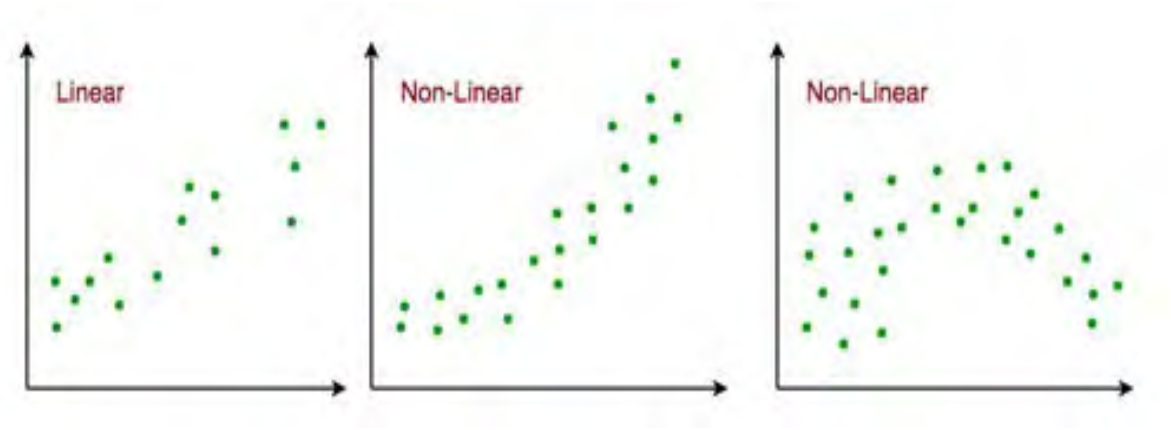
Αυτή περιλαμβάνει περισσότερες από μία ανεξάρτητες μεταβλητές και μία εξαρτημένη μεταβλητή. Η εξίσωση για την πολλαπλή γραμμική παλινδρόμηση είναι:

$$\mathbf{Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots..+\beta_n X_n},$$

όπου:το Υ είναι η εξαρτημένη μεταβλητή, X_1, X_2, …, X_n είναι οι ανεξάρτητες μεταβλητές, το β_0 είναι η σταθερά και οι β_1,β_2, …,β_n είναι οι συντελεστές που καθορίζουν την κλίση.

Η γραμμική παλινδρόμηση είναι ένα ισχυρό εργαλείο για την κατανόηση και την πρόβλεψη της συμπεριφοράς μιας μεταβλητής, ωστόσο, πρέπει να πληροί ορισμένες προϋποθέσεις για να είναι ακριβείς και αξιόπιστες λύσεις. Για την **απλή γραμμική παλινδρόμηση** απαιτούμε να έχουμε:

Γραμμικότητα: Οι ανεξάρτητες και εξαρτημένες μεταβλητές έχουν γραμμική σχέση μεταξύ τους. Αυτό σημαίνει ότι οι αλλαγές στην εξαρτημένη μεταβλητή ακολουθούν αυτές στην ανεξάρτητη μεταβλητή με γραμμικό τρόπο. Αυτό σημαίνει ότι θα πρέπει να υπάρχει μια ευθεία γραμμή που μπορεί να “τραβηχτεί” μέσω των σημείων δεδομένων (Εικόνα 4.2). Εάν η σχέση δεν είναι γραμμική, τότε η γραμμική παλινδρόμηση δεν θα είναι ακριβές μοντέλο.



Εικόνα 4.2

Ανεξαρτησία: **Οι** παρατηρήσεις στο σύνολο δεδομένων είναι ανεξάρτητες η μία από την άλλη. Αυτό σημαίνει ότι η τιμή της εξαρτημένης μεταβλητής για μια παρατήρηση δεν εξαρτάται από την τιμή της εξαρτημένης μεταβλητής για μια άλλη παρατήρηση. Εάν οι παρατηρήσεις δεν είναι ανεξάρτητες, τότε η γραμμική παλινδρόμηση δεν θα είναι ακριβές μοντέλο.

Ομοσκεδαστικότητα: Σε όλα τα επίπεδα των ανεξάρτητων μεταβλητών, η διακύμανση των σφαλμάτων είναι σταθερή. Αυτό υποδηλώνει ότι το ποσό της ανεξάρτητης μεταβλητής δεν επηρεάζει τη διακύμανση των σφαλμάτων. Εάν η διακύμανση των υπολειμμάτων δεν είναι σταθερή, τότε η γραμμική παλινδρόμηση δεν θα είναι ακριβές μοντέλο.

Κανονικότητα: Τα υπολείμματα πρέπει να κατανέμονται κανονικά. Αυτό σημαίνει ότι τα υπολείμματα πρέπει να ακολουθούν μια καμπύλη σε σχήμα καμπάνας.

Για την **πολλαπλή γραμμική παλινδρόμηση** απαιτούμε τα εξής: Για την πολλαπλή γραμμική παλινδρόμηση, ισχύουν και οι τέσσερις προυποθέσεις που αναφέραμε για την απλή γραμμική παλινδρόμηση. Εκτός όμως από αυτές, παρακάτω αναφέρουμε μερικές επιπλόν ακόμη:

Χωρίς πολυσυγγραμμικότητα: Δεν υπάρχει υψηλή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Αυτό δείχνει ότι υπάρχει μικρή ή καθόλου συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών. Η πολυσυγγραμμικότητα εμφανίζεται όταν δύο ή περισσότερες ανεξάρτητες μεταβλητές συσχετίζονται σε μεγάλο βαθμό μεταξύ τους, γεγονός που μπορεί να δυσκολέψει τον προσδιορισμό της μεμονωμένης επίδρασης κάθε μεταβλητής στην εξαρτημένη μεταβλητή. Εάν υπάρχει πολυσυγγραμμικότητα, τότε η πολλαπλή γραμμική παλινδρόμηση δεν θα είναι ακριβές μοντέλο.

Προσθετικότητα: Το μοντέλο υποθέτει ότι η επίδραση των αλλαγών σε μια μεταβλητή πρόβλεψης στη μεταβλητή απόκρισης είναι συνεπής ανεξάρτητα από τις τιμές των άλλων μεταβλητών. Αυτή η υπόθεση συνεπάγεται ότι δεν υπάρχει αλληλεπίδραση μεταξύ των μεταβλητών ως προς τις επιδράσεις τους στην εξαρτημένη μεταβλητή.

Επιλογή χαρακτηριστικών: Στην πολλαπλή γραμμική παλινδρόμηση, είναι απαραίτητο να επιλέξουμε προσεκτικά τις ανεξάρτητες μεταβλητές που θα συμπεριληφθούν στο μοντέλο. Η συμπερίληψη άσχετων ή περιττών μεταβλητών μπορεί να οδηγήσει σε υπερβολική προσαρμογή και να περιπλέξει την ερμηνεία του μοντέλου.

➢ Πλεονεκτήματα της Γραμμικής Παλινδρόμησης

- Η γραμμική παλινδρόμηση είναι ένας σχετικά απλός αλγόριθμος, που καθιστά εύκολη την κατανόηση και την εφαρμογή του. οι συντελεστές του μοντέλου γραμμικής παλινδρόμησης μπορούν να ερμηνευθούν ως η αλλαγή στην εξαρτημένη μεταβλητή για μια μεταβολή μιας μονάδας στην ανεξάρτητη μεταβλητή, παρέχοντας πληροφορίες για τις σχέσεις μεταξύ των μεταβλητών.
- Η γραμμική παλινδρόμηση είναι υπολογιστικά αποδοτική και μπορεί να χειριστεί αποτελεσματικά μεγάλα σύνολα δεδομένων. Μπορεί να εκπαιδευτεί γρήγορα σε μεγάλα σύνολα δεδομένων, καθιστώντας το κατάλληλο για εφαρμογές σε πραγματικό χρόνο.
- Η γραμμική παλινδρόμηση είναι σχετικά ισχυρή σε ακραίες τιμές σε σύγκριση με άλλους αλγόριθμους μηχανικής μάθησης. οι ακραίες τιμές ενδέχεται να έχουν μικρότερο αντίκτυπο στη συνολική απόδοση του μοντέλου.
- Η γραμμική παλινδρόμηση συχνά χρησιμεύει ως ένα καλό βασικό μοντέλο για σύγκριση με πιο σύνθετους αλγόριθμους μηχανικής μάθησης.
- Η γραμμική παλινδρόμηση είναι ένας καθιερωμένος αλγόριθμος με πλούσιο ιστορικό και είναι ευρέως διαθέσιμος σε διάφορες βιβλιοθήκες μηχανικής εκμάθησης και πακέτα λογισμικού.

➢ Μειονεκτήματα της Γραμμικής παλινδρόμησης

- Η γραμμική παλινδρόμηση προϋποθέτει μια γραμμική σχέση μεταξύ των εξαρτημένων και των ανεξάρτητων μεταβλητών. Εάν η σχέση δεν είναι γραμμική, το μοντέλο μπορεί να μην έχει καλή απόδοση.
- Η γραμμική παλινδρόμηση είναι ευαίσθητη στην πολυσυγγραμμικότητα, η οποία συμβαίνει όταν υπάρχει υψηλή συσχέτιση μεταξύ ανεξάρτητων μεταβλητών. Η πολυσυγγραμμικότητα μπορεί να διογκώσει τη διακύμανση των συντελεστών και να οδηγήσει σε ασταθείς προβλέψεις μοντέλων.
- Η γραμμική παλινδρόμηση είναι ευαίσθητη τόσο στην υπερπροσαρμογή όσο και στην υποπροσαρμογή. Η υπερπροσαρμογή συμβαίνει όταν το μοντέλο μαθαίνει πολύ καλά τα δεδομένα εκπαίδευσης και αποτυγχάνει να γενικεύσει σε μη ορατά δεδομένα. Η υποπροσαρμογή συμβαίνει όταν το μοντέλο είναι πολύ απλό για να καταγράψει τις υποκείμενες σχέσεις στα δεδομένα.
- Η γραμμική παλινδρόμηση παρέχει περιορισμένη επεξηγηματική ισχύ για πολύπλοκες σχέσεις μεταξύ μεταβλητών. Πιο προηγμένες τεχνικές μηχανικής εκμάθησης μπορεί να είναι απαραίτητες για βαθύτερες πληροφορίες.

Με βάση τα παραπάνω στοιχεία θα μπορούσαμε να πούμε ότι η γραμμική παλινδρόμηση είναι ένας θεμελιώδης αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται ευρέως εδώ και πολλά χρόνια λόγω της απλότητας, της ερμηνευσιμότητας και της αποτελεσματικότητάς του. Είναι ένα πολύτιμο εργαλείο για την κατανόηση των σχέσεων μεταξύ μεταβλητών και την πραγματοποίηση προβλέψεων σε μια ποικιλία εφαρμογών. Ωστόσο, είναι σημαντικό να γνωρίζουμε τους περιορισμούς του, όπως η υπόθεση της γραμμικότητας και η ευαισθησία στην πολυσυγγραμμικότητα. Όταν αυτοί οι περιορισμοί εξετάζονται προσεκτικά, η γραμμική παλινδρόμηση μπορεί να είναι ένα ισχυρό εργαλείο για την ανάλυση και την πρόβλεψη δεδομένων. Παρακάτω αναφέρουμε σημαντικούς τομείς που έχει εφαρμογή η γραμμική παλινδρόμηση.

<u>Προγνωστική Μοντελοποίηση και Πρόβλεψη:</u>

- Πρόβλεψη πωλήσεων.
- Πρόβλεψη ζήτησης.
- Πρόβλεψη τιμών μετοχών.

- Πρόγνωση καιρού.

Μάρκετινγκ και Ανάλυση πελατών:
- Μοντελοποίηση απόκρισης της αγοράς.
- Πρόβλεψη αξίας διάρκειας ζωής πελάτη.
- Ανάλυση μεριδίου αγοράς.

Χρηματοοικονομική Ανάλυση:
- Πιστωτική βαθμολόγηση και αξιολόγηση κινδύνου.
- Διαχείριση χαρτοφυλακίου.
- Ανάλυση χρηματοοικονομικής αγοράς.
- Ανίχνευση απάτης.

Υγειονομική και Ιατρική Έρευνα:
- Πρόβλεψη και διάγνωση ασθενειών.
- Πρόβλεψη της έκβασης του ασθενούς.
- Ανάλυση αποτελεσματικότητας φαρμάκων.
- Εκτίμηση κινδύνου για την υγεία.

Ποιοτικός Έλεγχος και Βελτιστοποίηση Διαδικασιών:
- Βελτιστοποίηση διαδικασίας παραγωγής.
- Έλεγχος ποιότητας προϊόντος.
- Βελτιστοποίηση εφοδιαστικής αλυσίδας.
- Ανίχνευση ανωμαλιών.

Κοινωνικές Επιστήμες και Ανάλυση Συμπεριφοράς:
- Ανάλυση κοινωνικών και οικονομικών επιπτώσεων.
- Εξόρυξη γνώμης και ανάλυση συναισθήματος.
- Εκπαιδευτική έρευνα.
- Δημογραφική ανάλυση

Ενέργεια και Υπηρεσίες κοινής ωφέλειας:
- Πρόβλεψη κατανάλωσης ενέργειας.
- Πρόβλεψη φορτίου.
- Μοντελοποίηση τιμών ενέργειας.
- Βελτιστοποίηση ανανεώσιμων πηγών ενέργειας.

Sports Analytics:
- Ανάλυση απόδοσης παίκτη.
- Πρόβλεψη αποτελέσματος.
- Βελτιστοποίηση σύνθεσης ομάδας.
- Λήψη αποφάσεων εντός του παιχνιδιού.

Περιβαλλοντική Ανάλυση:
- Κλιματική μοντελοποίηση.
- Πρόβλεψη ρύπανσης.
- Εκτίμηση περιβαλλοντικών επιπτώσεων.
- Διαχείριση φυσικών πόρων.

Ανάλυση χρονοσειρών:

- Ανάλυση τάσεων.
- Προσδιορισμός εποχιακών προτύπων.
- Πρόβλεψη οικονομικών δεικτών.
- Ανάλυση χρηματιστηρίου.

Πρακτικές Εφαρμογές στην Ιατρική

Η γραμμική παλινδρόμηση έχει ένα ευρύ φάσμα εφαρμογών στην ιατρική, όπως:

Πρόβλεψη αποτελεσμάτων ασθενών: Η γραμμική παλινδρόμηση μπορεί να χρησιμοποιηθεί για την πρόβλεψη των αποτελεσμάτων των ασθενών με βάση διάφορους παράγοντες.

Προσδιορισμός παραγόντων κινδύνου: Οι πάροχοι υγειονομικής περίθαλψης μπορούν να χρησιμοποιήσουν γραμμική παλινδρόμηση για να εντοπίσουν παράγοντες κινδύνου για ορισμένες ασθένειες.

Αποτελεσματικότητα θεραπείας: Οι ερευνητές μπορούν να χρησιμοποιήσουν τη γραμμική παλινδρόμηση για να αξιολογήσουν την αποτελεσματικότητα διαφορετικών θεραπειών. Αναφέρουμε για παράδειγμα ότι οι ιατρικοί ερευνητές μπορούν να χρησιμοποιήσουν την γραμμική παλινδρόμηση για να προσδιορίσουν τη σχέση μεταξύ ανεξάρτητων χαρακτηριστικών, όπως η ηλικία και το σωματικό βάρος, και εξαρτημένων χαρακτηριστικών, όπως η αρτηριακή πίεση. Αυτό μπορεί να βοηθήσει στην αποκάλυψη των παραγόντων κινδύνου που σχετίζονται με ασθένειες.

Σε ότι ακολουθεί αναφέρουμε παραδείγματα γραμμικής παλινδρόμησης.

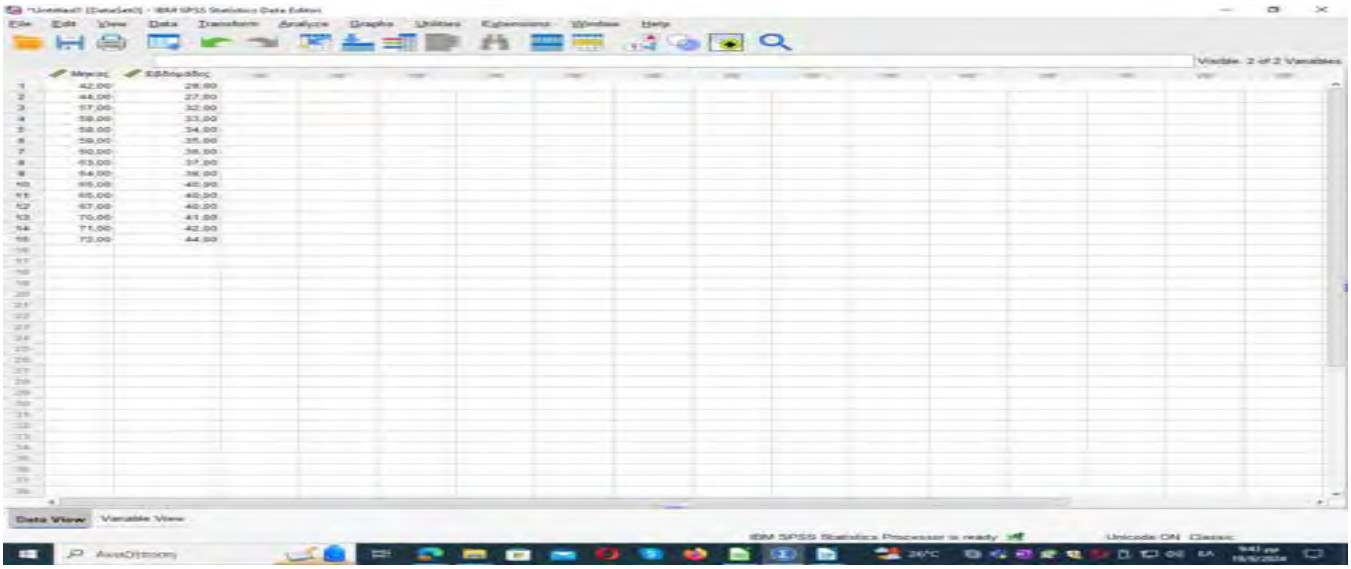
Παράδειγμα 1. Στον πίνακα που ακολουθεί δίνεται η μεταβολή του μήκους του βραχίονα νηπίων σε mm με το χρόνο σε εβδομάδες.

Μήκος (mm)	**Εβδομάδες (weeks)**
42	28
44	27
57	32
58	33
58	34
59	35
60	36
63	37
64	38
65	40
65	40
67	40
70	41
71	42
72	44

Με την χρήση του στατιστικού πακέτου SPSS θα γίνει η γραφική παράσταση weeks-mm.

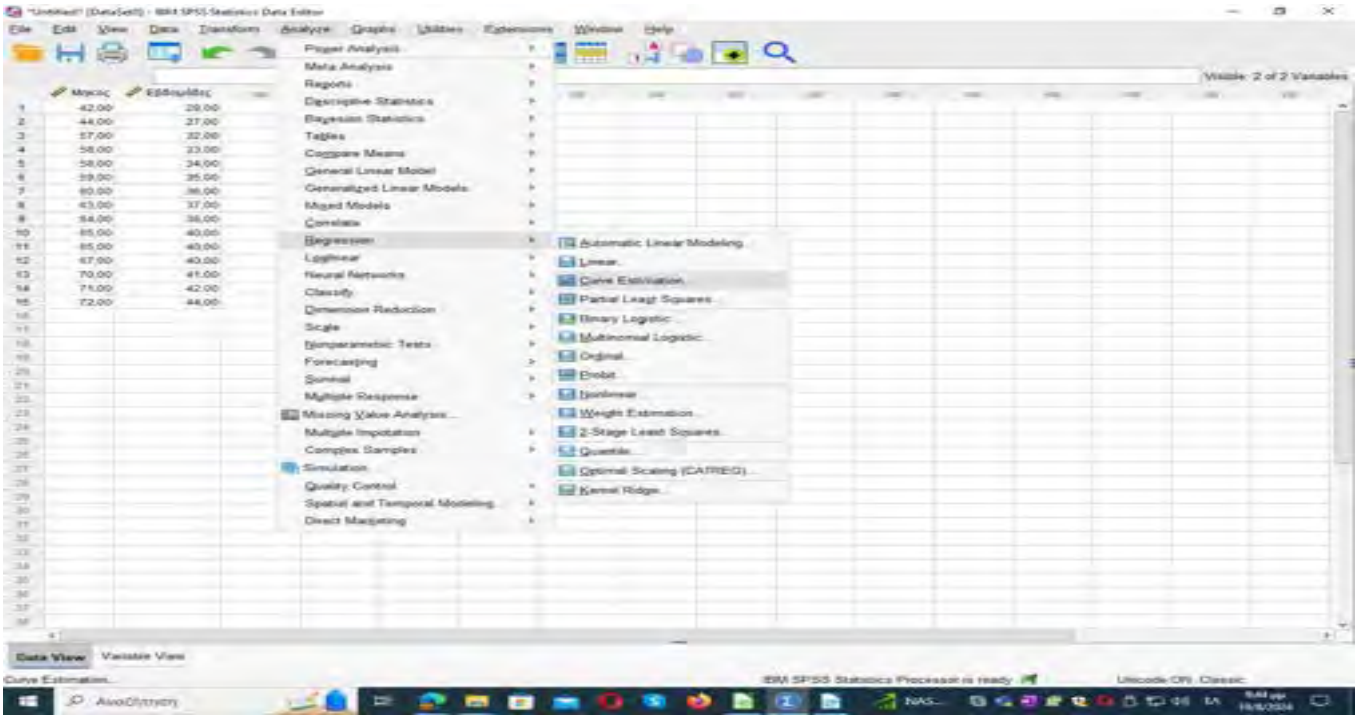
Ακολουθούμε τα παρακάτω βήματα:

Βήμα 1. Μεταφέρουμε τα δεδομένα σ' ένα φύλλο εργασίας του SPSS και αποφασίζουμε ποια μεταβλητή θα είναι ανεξάρτητη και ποια εξαρτημένη. Στην περίπτωσή μας ως ανεξάρτητη μεταβλητή θα οριστεί το μήκος του βραχίονα και εξαρτημένη η ηλικία (Εικόνα 4.3).



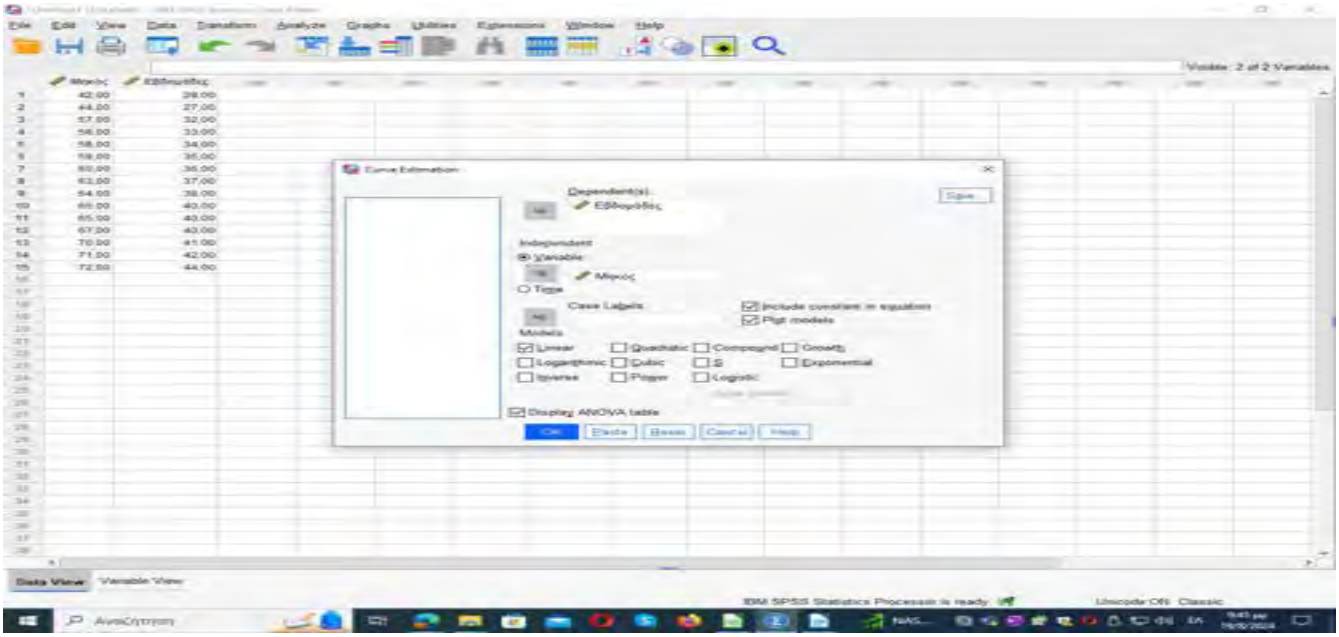
Εικόνα 4.3

Βήμα 2. Στην συνέχεια επιλέγουμε: **Analyze → Regression → Curve Estimation** (Εικόνα 4.4).



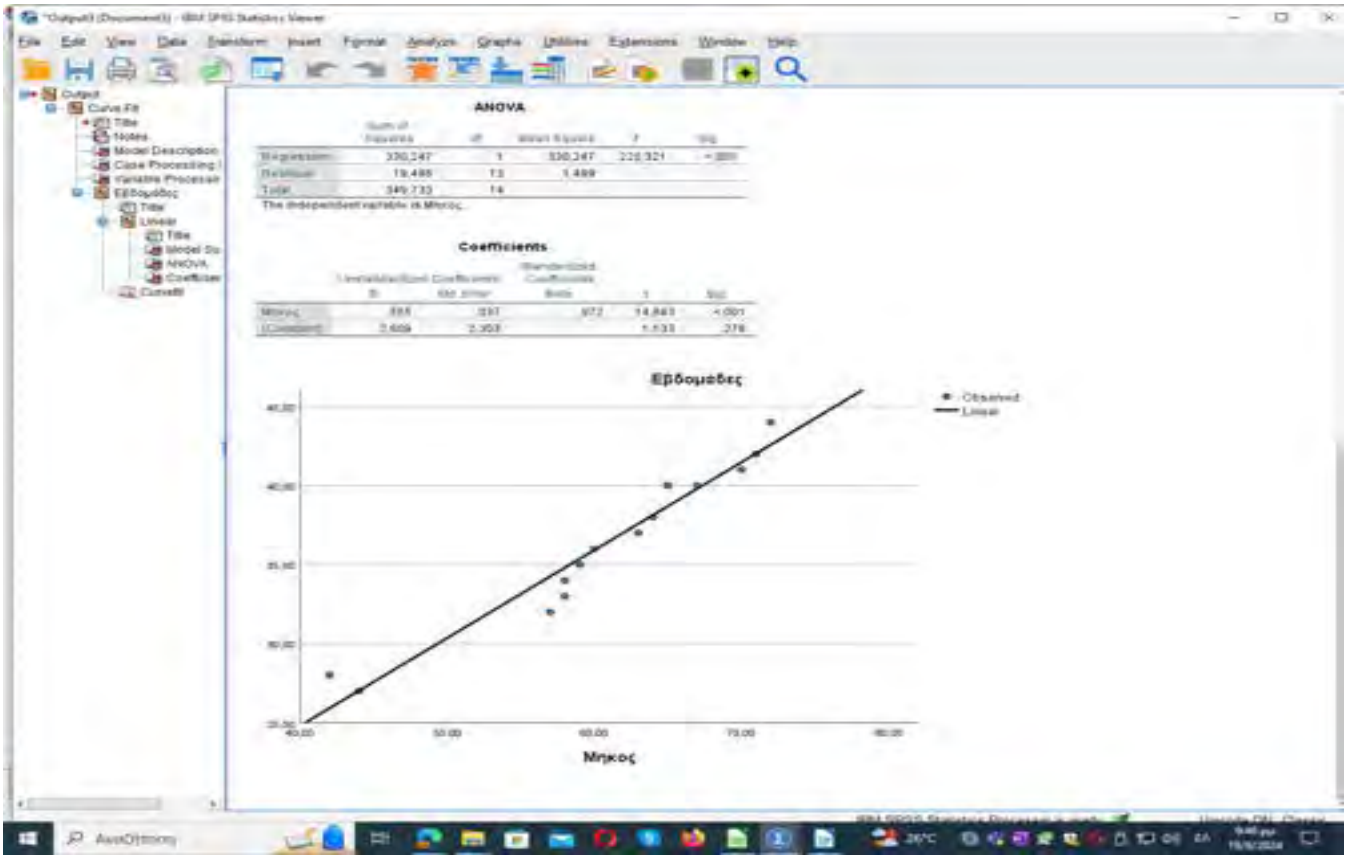
Εικόνα 4.4

Εισάγουμε τη μεταβλητή Εβδομάδες στο πλαίσιο **Variable(s)** και τη μεταβλητή mm στο Independent **Variable**, όπως φαίνεται παρακάτω. Επίσης, επιλέγουμε **Display ANOVA table**, **Linear**, **Plot models** και **Include constant in equation** (Εικόνα 4.5).



Εικόνα 4.5

Οπότε παίρνουμε τα παρακάτω αποτελέσματα της Εικόνας 4.6:



Εικόνα 4.6

Από τον πίνακα προκύπτει ότι η εξίσωση της ευθείας είναι:

$$y = 2{,}609 + 0.555x.$$

Η τυπική απόκλιση της σταθεράς β_0 είναι 2.609 και της κλίσης β_1 είναι 0.037. Η τελευταία στήλη μας ενημερώνει αν μια σταθερά τ είναι στατιστικά σημαντική. Πρέπει η τιμή Sig. να είναι μικρότερη από 0.05. Γενικά αν μια σταθερά είναι στατιστικά μη σημαντική μπορεί να απαλειφθεί από τη μελέτη, εκτός κι αν υπάρχουν ισχυροί λόγοι να παραμείνει.

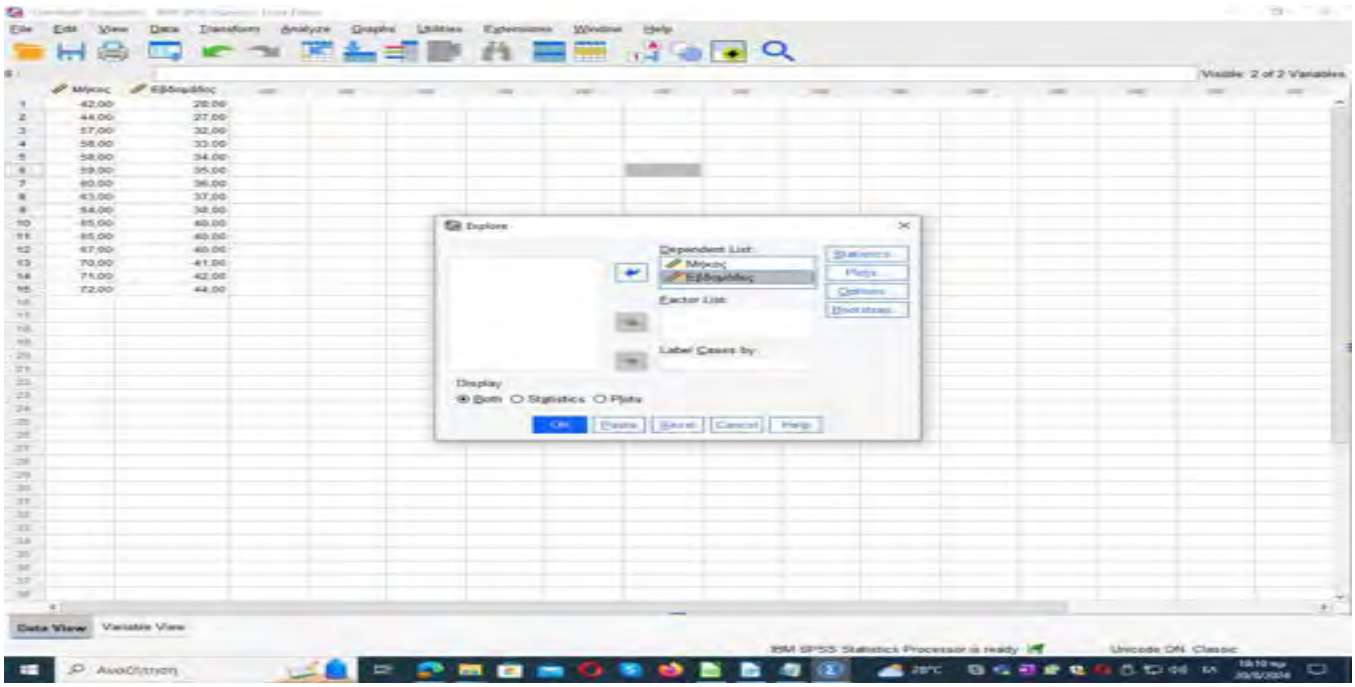
Γνωρίζοντας την παραπάνω εξίσωση μπορούμε για παράδειγμα να προβλέψουμε την ηλικία ενός νηπίου αν γνωρίζουμε για παράδειγμα το μήκος του βραχίονα. Έτσι για παράδειγμα αν x=53, έχουμε:

$$y=2{,}609+0{,}555\ 53=32{,}024 \text{ εβδομάδες.}$$

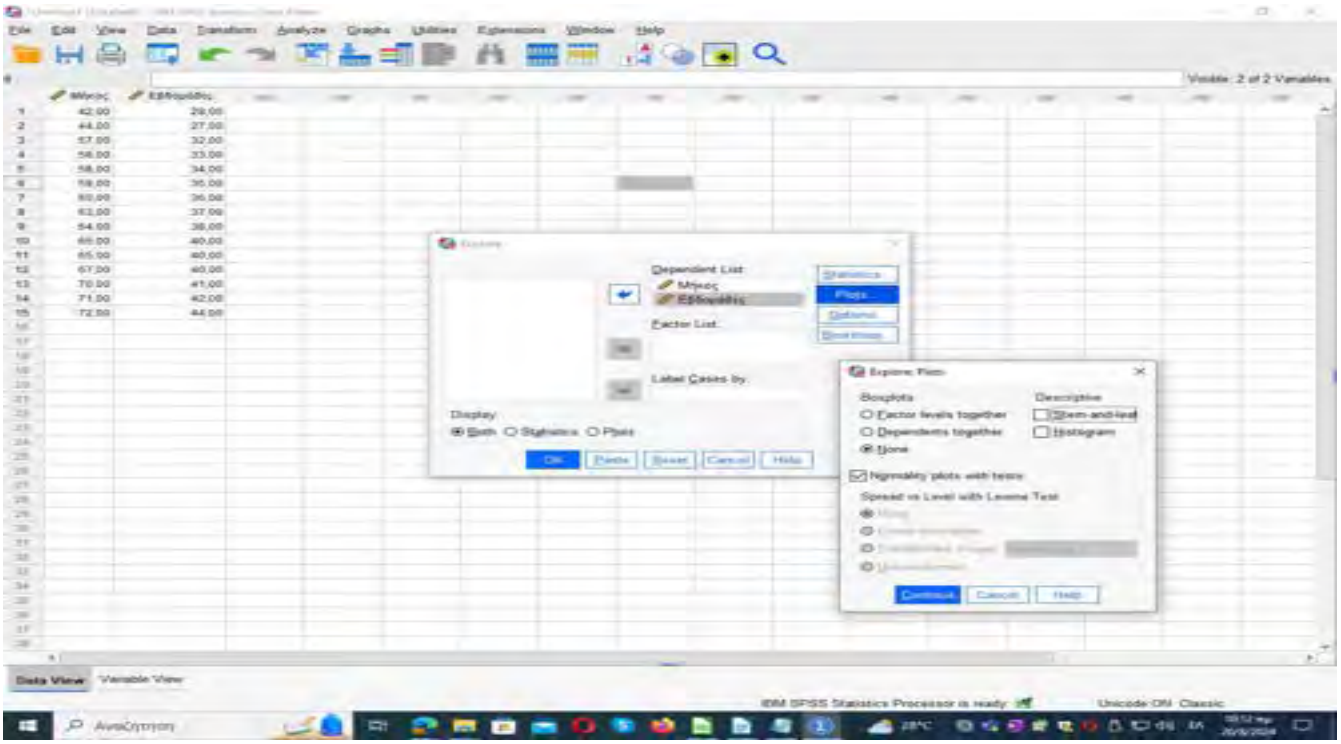
Παρατήρηση. Ένα θέμα που σχετίζεται με την παλινδρόμηση είναι το πρόβλημα της συσχέτισης (correlation) δύο μεταβλητών. Σε αρκετές περιπτώσεις είναι καλό να γνωρίζουμε αν δύο τυχαίες μεταβλητές σχετίζονται ή όχι. Αν δηλαδή η μεταβολή της μιας μεταβάλει και την άλλη. Για να ελέγξουμε αν δύο μεταβλητές, x και y σχετίζονται, υπολογίζουμε συνήθως τον συντελεστή Pearson r. Ο συντελεστής r παίρνει τιμές στο διάστημα από –1 έως 1. Αρνητικές τιμές του r σημαίνουν ότι όταν η μεταβλητή x αυξάνει, η y ελαττώνεται και το αντίστροφο. r = 0 σημαίνει παντελή έλλειψη συσχέτισης και r θετικό σημαίνει ότι όταν η μια μεταβλητή αυξάνει, αυξάνει και η άλλη.

Παρακάτω θα δούμε βήμα - βήμα τον υπολογισμό του συντελεστή Pearson r για το παράδειγμα που αναφέραμε παραπάνω.

Επιλέγουμε: **Analyze → Descriptive Statistics → Explore** (Εικόνα 4.7-Εικόνα 4.8).

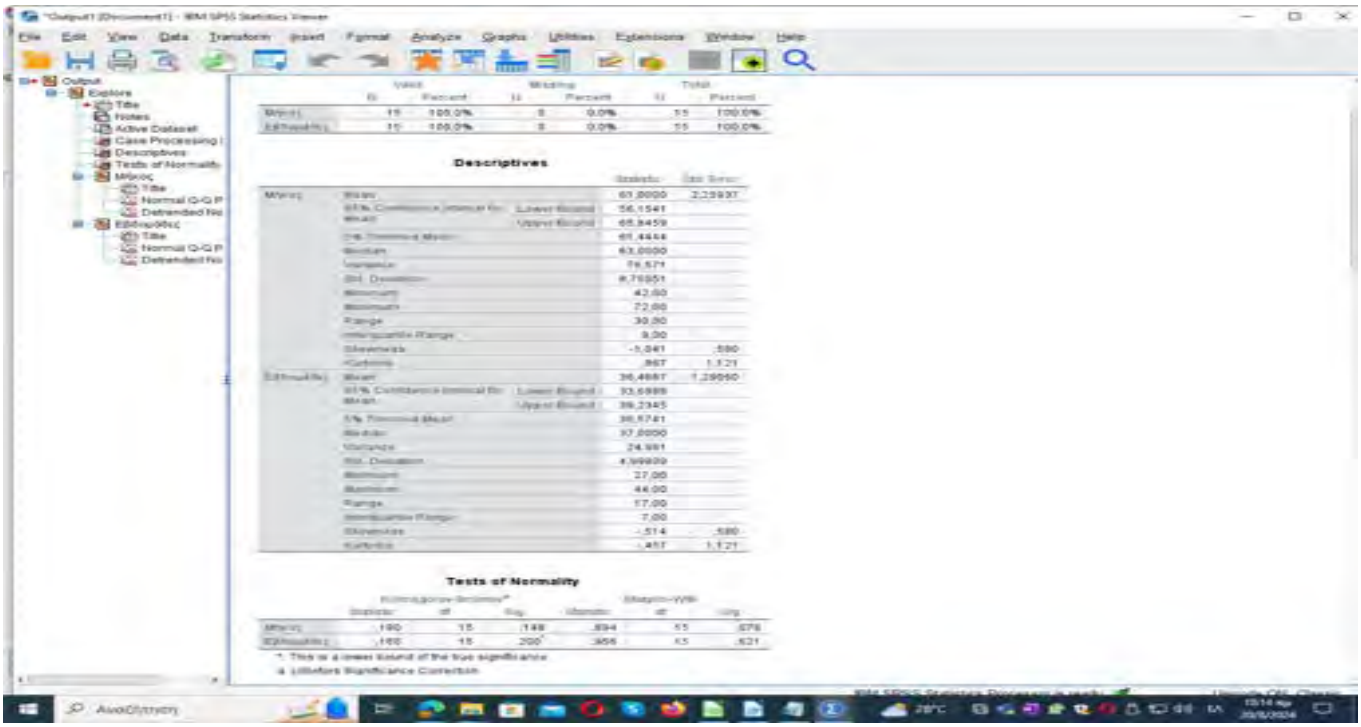


Εικόνα 4.7



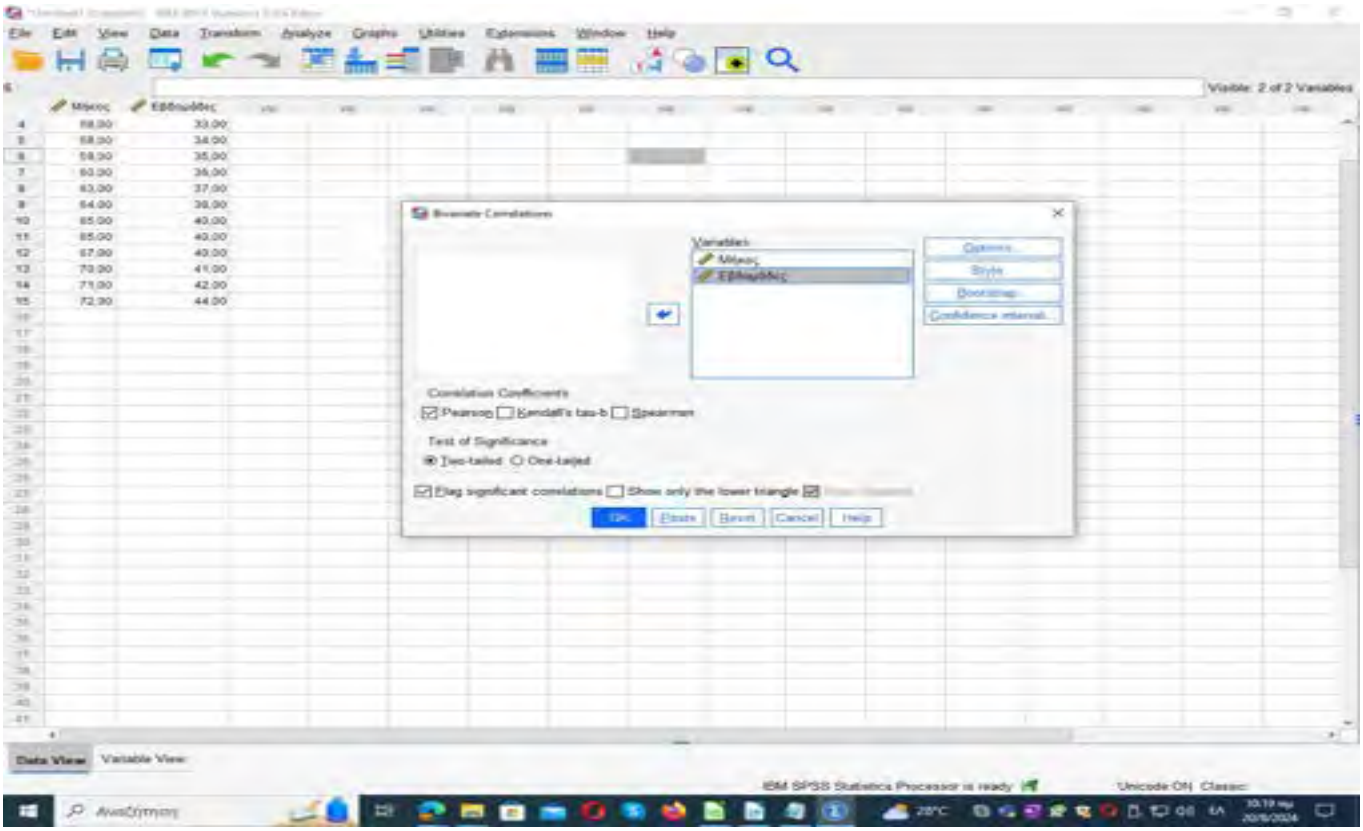
Εικόνα 4.8

Βήμα 2. Οπότε από τον παρακάτω πίνακα της Εικόνας 4.9 βλέπουμε αν ακολουθούν οι μεταβλητές την κανονική κατανομή.



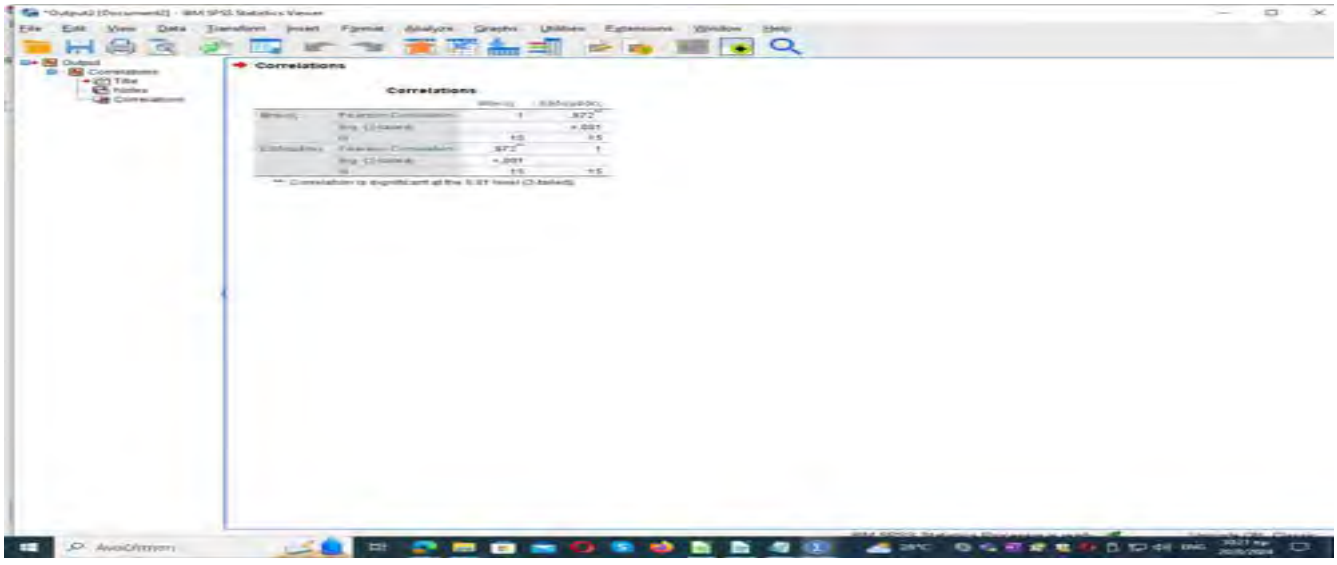
Εικόνα 4.9

Βήμα 3. Τώρα υπολογίζουμε τον συντελεστή Pearson r. Επιλέγουμε: **Analyze** → **Correlate** → **Bivariate** (Εικόνα 4.10).



Εικόνα 4.10

Οπότε παίρνουμε τα παρακάτω αποτελέσματα της Εικόνας 4.11.



Εικόνα 4.11

Βλέπουμε ότι υπάρχει θετική συσχέτιση των μεταβλητών και μάλιστα υψηλή συσχέτιση r = 0,972.

Παράδειγμα 2. Σε μια κλινική έγινε έρευνα για τους δείκτες μιας ασθένειας και προέκυψε ο παρακάτω πίνακας:

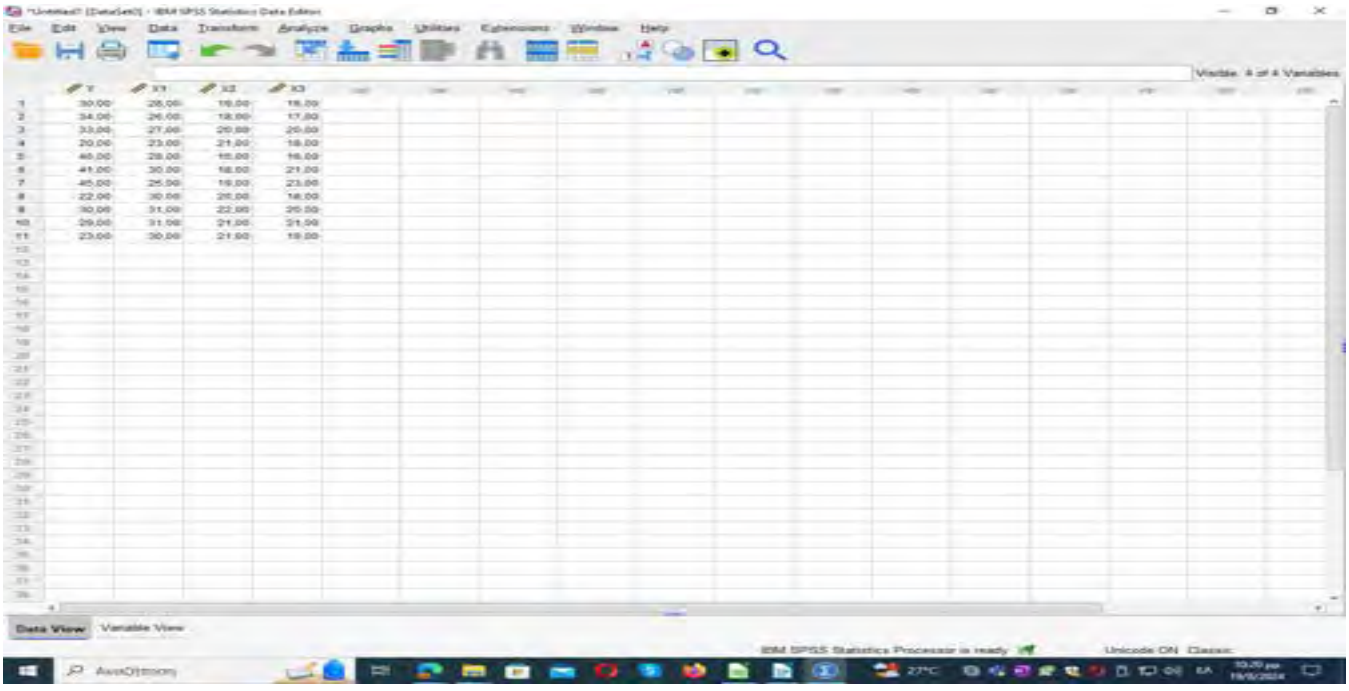
Σοβαρή Ένδειξη για την υγεία του ασθενούς Y	Μέτρηση x1	Μέτρηση x2	Μέτρηση x3
30	28	19	16
34	26	18	17
33	27	20	20
20	23	21	18
46	28	15	16
41	30	18	21
45	25	19	23
22	30	20	18
30	31	22	20
29	31	21	21
23	30	21	19

Θα προσδιορίσουμε το γραμμικό μοντέλο, δηλαδή τη συνάρτηση:

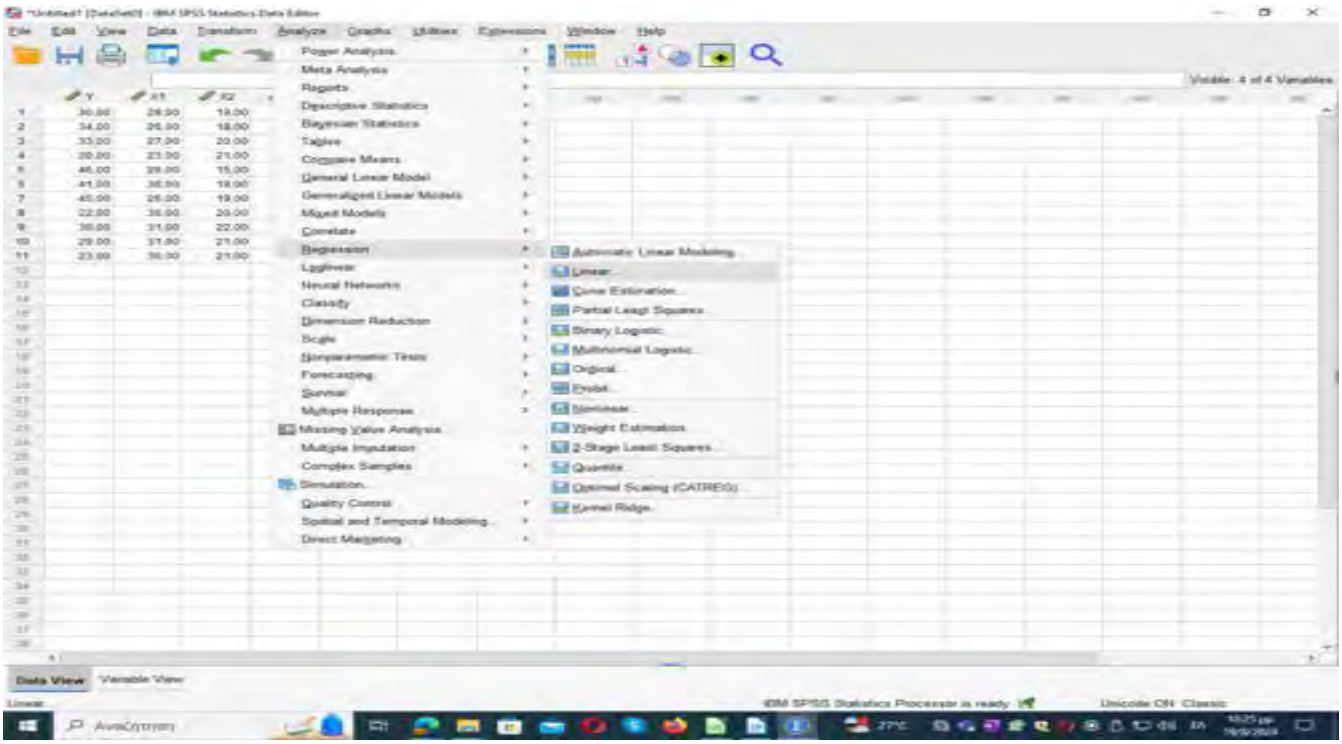
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Για τον προσδιορισμό των συντελεστών β_0, β_1, β_2 και β_3 ακολουθούμε τα παρακάτω βήματα:

Βήμα 1. Μεταφέρουμε τα δεδομένα στο SPSS και ακολουθούμε την πορεία: **Analyze → Regression → Linear** (Εικόνα 4.12-Εικόνα 4.13).

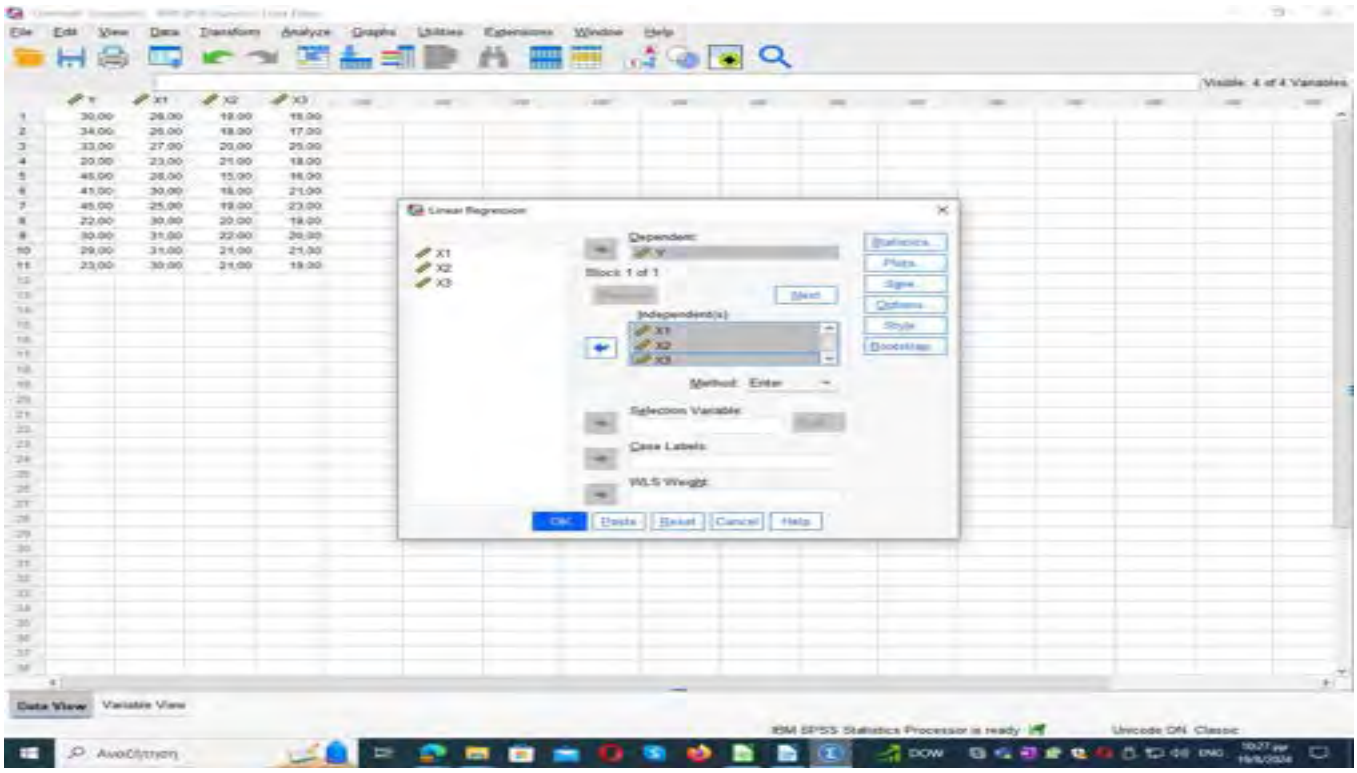


Εικόνα 4.12



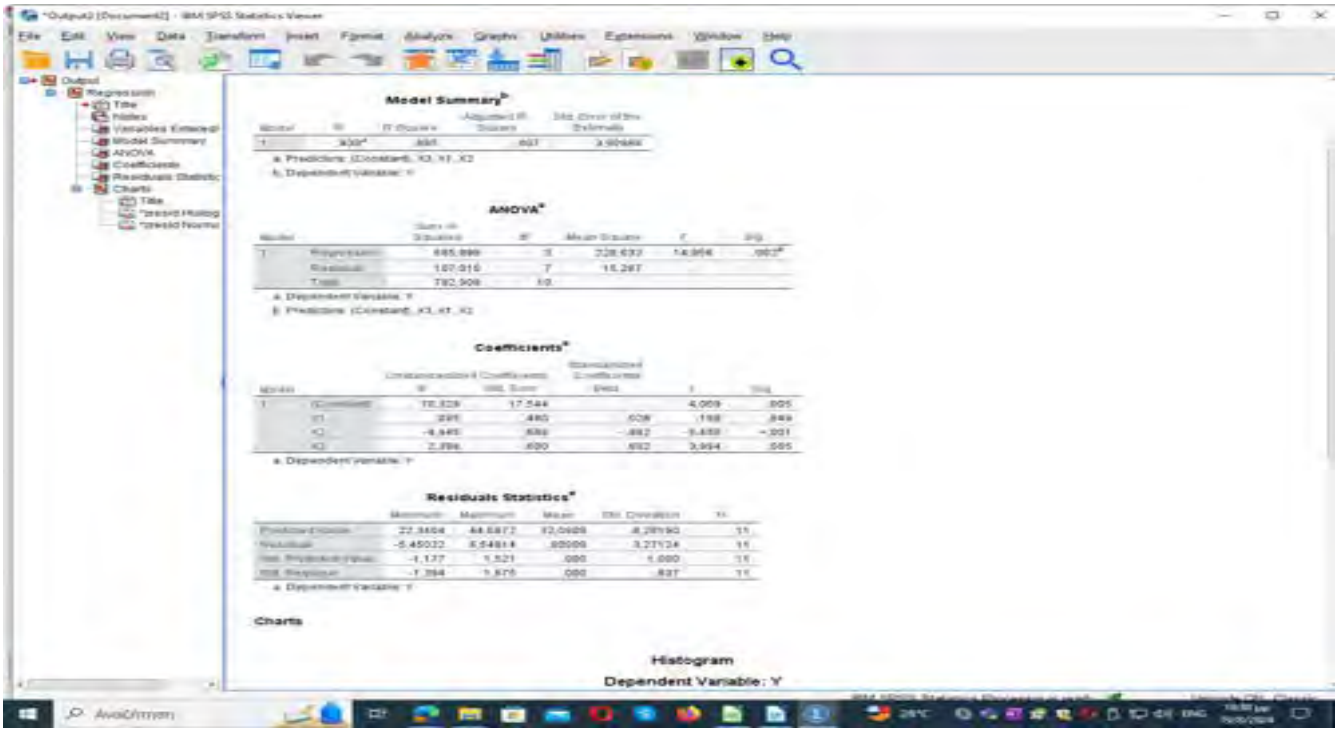
Εικόνα 4.13

Βήμα 2. Εισάγουμε τη μεταβλητή y στο πλαίσιο **Dependen**t και τις μεταβλητές X_1, X_2 και X_3 στο **Independent(s)** (Εικόνα 4.14).



Εικόνα 4.14

Πατώντας **OK** παίρνουμε τα παρακάτω αποτελέσματα της Εικόνας 4.15.



Εικόνα 4.15

Από τον παραπάνω πίνακα προκύπτει ότι η ζητούμενη εξίσωση είναι:

$$y= 70,329+ 0,095\ x_1 - 4,445\ x2 + 2,398\ x3.$$

4.2 Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση είναι η κατάλληλη ανάλυση παλινδρόμησης που πρέπει να διεξαχθεί όταν η εξαρτημένη μεταβλητή είναι διχοτομική (δυαδική). Όπως όλες οι αναλύσεις παλινδρόμησης, η λογιστική παλινδρόμηση είναι μια προγνωστική ανάλυση. Η λογιστική παλινδρόμηση χρησιμοποιείται για να περιγράψει δεδομένα και να εξηγήσει τη σχέση μεταξύ μιας εξαρτημένης δυαδικής μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών ονομαστικής, τακτικής, διαστήματος ή επιπέδου αναλογίας.

Η λογιστική παλινδρόμηση είναι μια ειδική περίπτωση ανάλυσης παλινδρόμησης και χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι ονομαστικά κλιμακούμενη. Αυτό συμβαίνει, για παράδειγμα, με τη μεταβλητή απόφαση αγοράς με τις δύο τιμές "αγοράζει ένα προϊόν" και "δεν αγοράζει ένα προϊόν". Με την λογιστική παλινδρόμηση, είναι πλέον δυνατό να εξηγηθεί η εξαρτημένη μεταβλητή ή να εκτιμηθεί η πιθανότητα εμφάνισης των κατηγοριών της μεταβλητής.

Παρακάτω αναφέρουμε μερικά χαρακτηριστικά παραδείγματα εφαρμογής της Λογιστικής Παλινδρόμησης.

Επιχειρηματικό παράδειγμα: Για έναν διαδικτυακό έμπορο λιανικής, πρέπει να προβλέψουμε ποιο προϊόν είναι πιο πιθανό να αγοράσει ένας συγκεκριμένος πελάτης. Για αυτό, λαμβάνουμε ένα σύνολο δεδομένων με προηγούμενους επισκέπτες και τις αγορές τους από τον διαδικτυακό πωλητή λιανικής.

Ιατρικό παράδειγμα: Θέλουμε να διερευνήσουμε εάν ένα άτομο είναι επιρρεπές σε μια συγκεκριμένη ασθένεια ή όχι. Για το σκοπό αυτό, λαμβάνουμε ένα σύνολο δεδομένων με άρρωστα και μη άτομα καθώς και άλλες ιατρικές παραμέτρους.

Πολιτικό παράδειγμα: Θα ψήφιζε κάποιος για το κόμμα Α αν γίνονταν εκλογές το επόμενο Σαββατοκύριακο;

Προσδιορισμός πιθανότητας καρδιακών προσβολών: Με τη βοήθεια ενός λογιστικού μοντέλου, οι ιατροί μπορούν να προσδιορίσουν τη σχέση μεταξύ μεταβλητών όπως το βάρος, η άσκηση κ.λπ και ενός ατόμου και να το χρησιμοποιήσουν για να προβλέψουν εάν το άτομο θα υποφέρει από καρδιακή προσβολή ή οποιαδήποτε άλλη ιατρική επιπλοκή.

Προσδιορισμός ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου: Τα εισερχόμενα email φιλτράρονται για να προσδιοριστεί εάν η επικοινωνία μέσω email είναι διαφημιστική/ανεπιθύμητη, κατανοώντας τις μεταβλητές πρόβλεψης και εφαρμόζοντας έναν αλγόριθμο λογιστικής παλινδρόμησης ελέγχουμε την αυθεντικότητά τους.

Εξίσωση Logistic Regression

Έστω ότι θέλουμε να κάνουμε παλινδρόμηση (απλή ή πολλαπλή) και η εξαρτημένη μεταβλητή παίρνει δύο τιμές μόνο, 0 ή 1 (επιτυχία ή αποτυχία). Παραδείγματα δυαδικών απαντήσεων θα μπορούσαν να περιλαμβάνουν την επιτυχία ή την αποτυχία ενός τεστ, την απάντηση ναι ή όχι σε μια έρευνα και την υψηλή ή χαμηλή αρτηριακή πίεση. Σε αυτήν την περίπτωση η κλασσική παλινδρόμηση που αναφέραμε δεν μπορεί να εφαρμοστεί (γιατί ή εξαρτημένη μεταβλητή δεν ακολουθεί την κανονική κατανομή). Σε μια τέτοια περίπτωση πρέπει να εφαρμόσουμε την λογιστική παλινδρόμηση.

Στον πυρήνα της Δυαδικής Λογιστικής Παλινδρόμησης βρίσκεται η Εξίσωση Λογιστικής Παλινδρόμησης, η οποία είναι ζωτικής σημασίας για την κατανόηση της σχέσης μεταξύ των μεταβλητών πρόβλεψης και του δυαδικού αποτελέσματος. Η εξίσωση μπορεί να εκφραστεί ως εξής:

$$\mathbf{\ln(p/p-1)= b_0 + b_1X_1 + b_2X_2 + \ldots.+ b_k X_k,}$$

όπου το p αντιπροσωπεύει την πιθανότητα του δυαδικού αποτελέσματος (π.χ. επιτυχία ή αποτυχία), το b_0 είναι ο σταθερός όρος και $b_1,b_2,\ldots,b_k$ είναι οι συντελεστές για τις ανεξάρτητες μεταβλητές X_1, X_2, …, X_k, αντίστοιχα. Επίσης το ln υποδηλώνει τον φυσικό λογάριθμο.

Παραδείγματα εφαρμογής της Λογιστικής Παλινδρόμησης

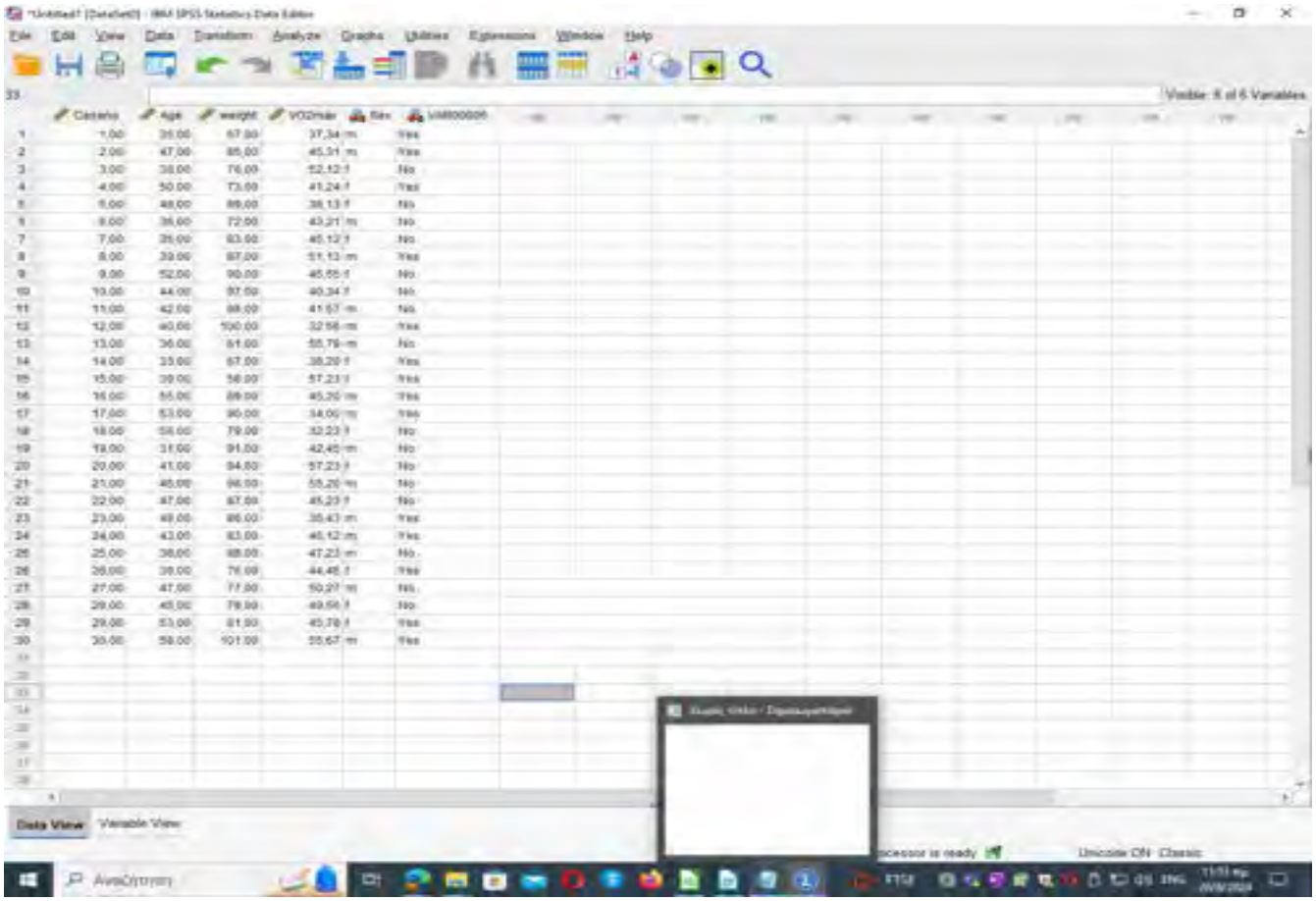
- Ένα κατάστημα θα ήθελε να αξιολογήσει τους παράγοντες που οδηγούν σε επιστροφή/μη επιστροφή του πελάτη.
- Ένα Πειραματικό Σχολείο θα ήθελε να αξιολογήσει την εισαγωγή (αποδοχή / μη αποδοχή) ενός μαθητή με βάση την ηλικία, το βαθμό, τα αποτελέσματα του τεστ ικανότητας.
- Αξιολόγηση εάν ένας συγκεκριμένος υποψήφιος κερδίσει/χάσει τις εκλογές με βάση το χρόνο που πέρασε στην εκλογική περιφέρεια, προηγουμένως εκλεγμένος, όχι. των ζητημάτων που επιλύθηκαν.

- Ένας ερευνητής ανθρώπινου δυναμικού θα ήθελε να εξακριβώσει πώς παράγοντες όπως η εμπειρία, τα χρόνια εκπαίδευσης, ο προηγούμενος μισθός, η κατάταξη του πανεπιστημίου επηρεάζουν την επιλογή ενός υποψηφίου σε μια συνέντευξη εργασίας.
- Ένας μελετητής θα ήθελε να προβλέψει την επιλογή της τράπεζας (Δημόσιας ή Ιδιωτικής) με βάση ανεξάρτητες μεταβλητές που περιλαμβάνουν την Τεχνολογία, τα Επιτόκια, τις Υπηρεσίες Προστιθέμενης Αξίας, τους αντιληπτούς κινδύνους, τη φήμη και άλλα.

Παράδειγμα στο SPSS. Ένας ερευνητής υγείας θέλει να είναι σε θέση να προβλέψει εάν η «επίπτωση της καρδιακής νόσου» μπορεί να προβλεφθεί με βάση την «ηλικία», το «βάρος», το «φύλο» και το «VO2max» (όπου το VO2max αναφέρεται στη μέγιστη αερόβια ικανότητα, ένας δείκτης της φυσικής κατάστασης και της υγείας).

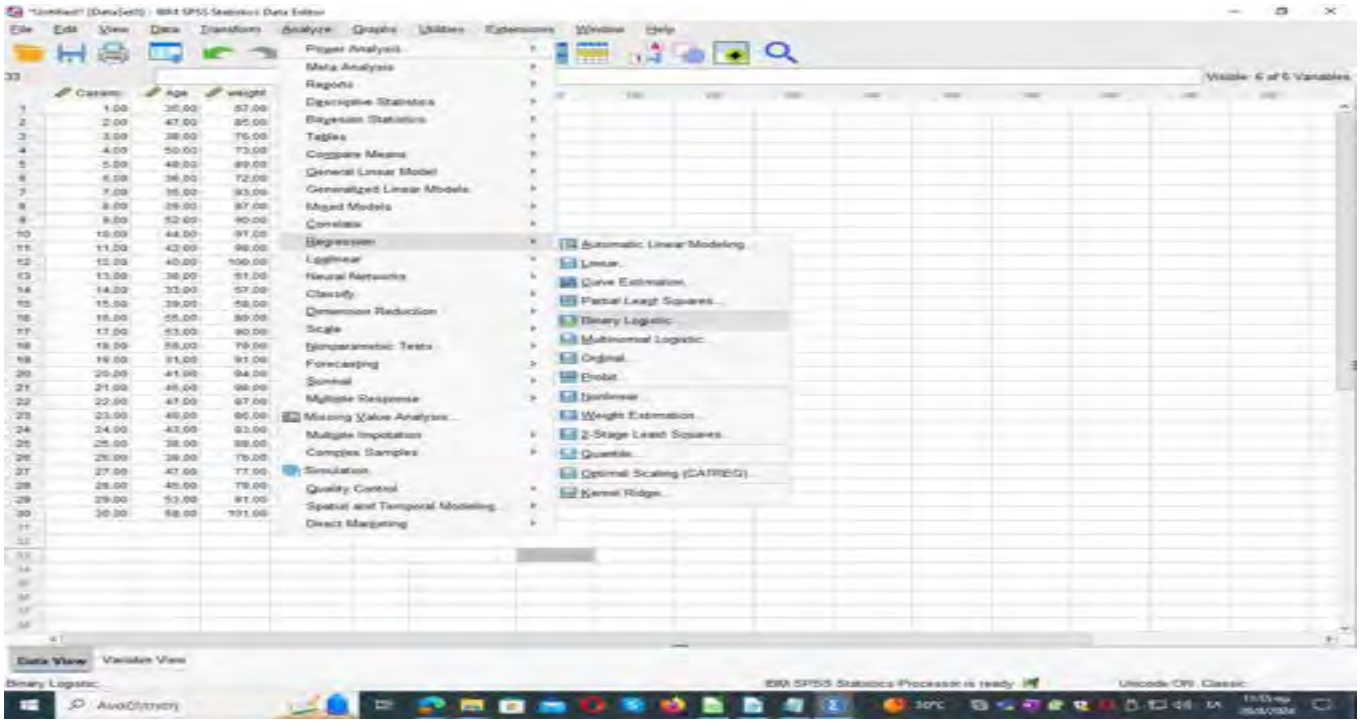
Σε αυτό το παράδειγμα, υπάρχουν έξι μεταβλητές: (1) καρδιακή_ασθένεια, δηλαδή εάν ο συμμετέχων έχει καρδιακή νόσο: "ναι" ή "όχι" (δηλαδή, η εξαρτημένη μεταβλητή). (2) VO2max, που είναι η μέγιστη αερόβια ικανότητα. (3) ηλικία, η οποία είναι η ηλικία του συμμετέχοντος. (4) βάρος, το οποίο είναι το βάρος του συμμετέχοντος (τεχνικά, είναι η «μάζα» του). και (5) το φύλο, που είναι το φύλο του συμμετέχοντος (δηλαδή, οι ανεξάρτητες μεταβλητές). και (6) caseno, που είναι ο αριθμός υπόθεσης. Στο παράδειγμα αυτό θα δούμε βήμα - βήμα στο SPSS την Λογιστική Παλινδρόμηση.

Βήμα 1. Περνάμε τα δεδομένα που έχουμε στο Λογιστικό φύλλο του SPSS (Εικόνα 4.16).



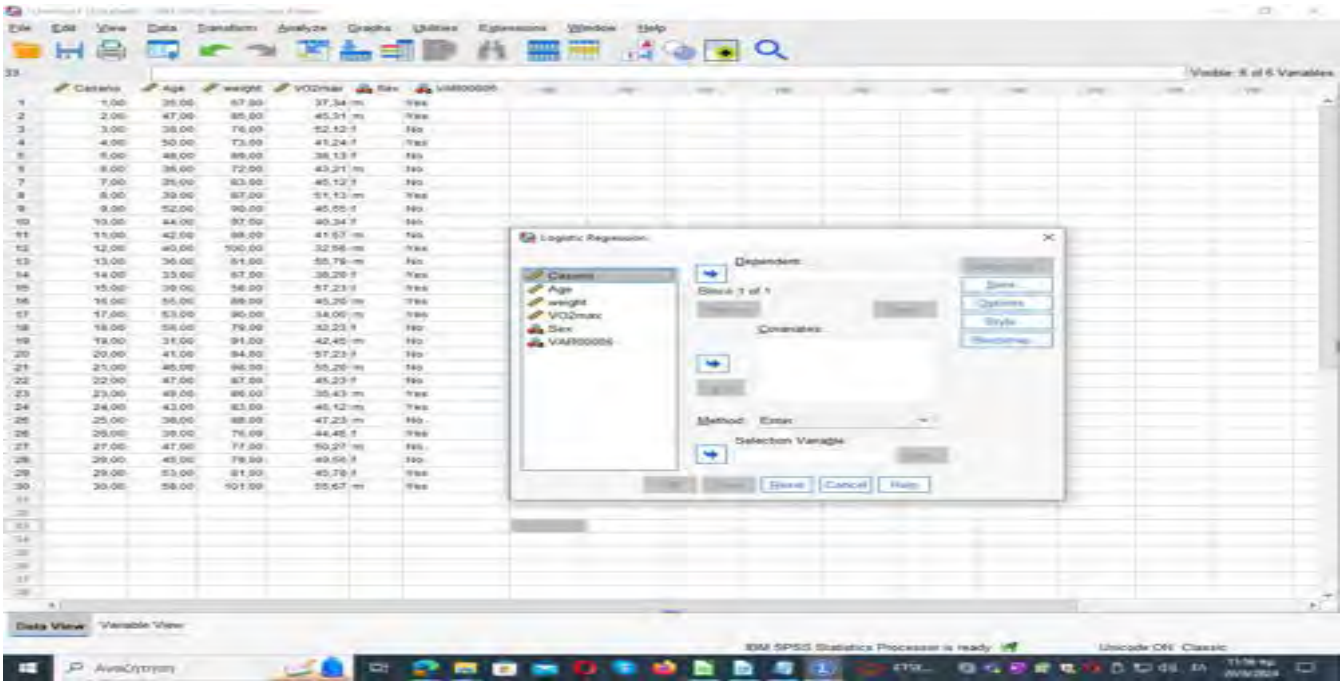
Εικόνα 4.16

Στην συνέχεια επιλέγουμε: **Analyze → Regression → Binary Logistic...**(Εικόνα 4.17).



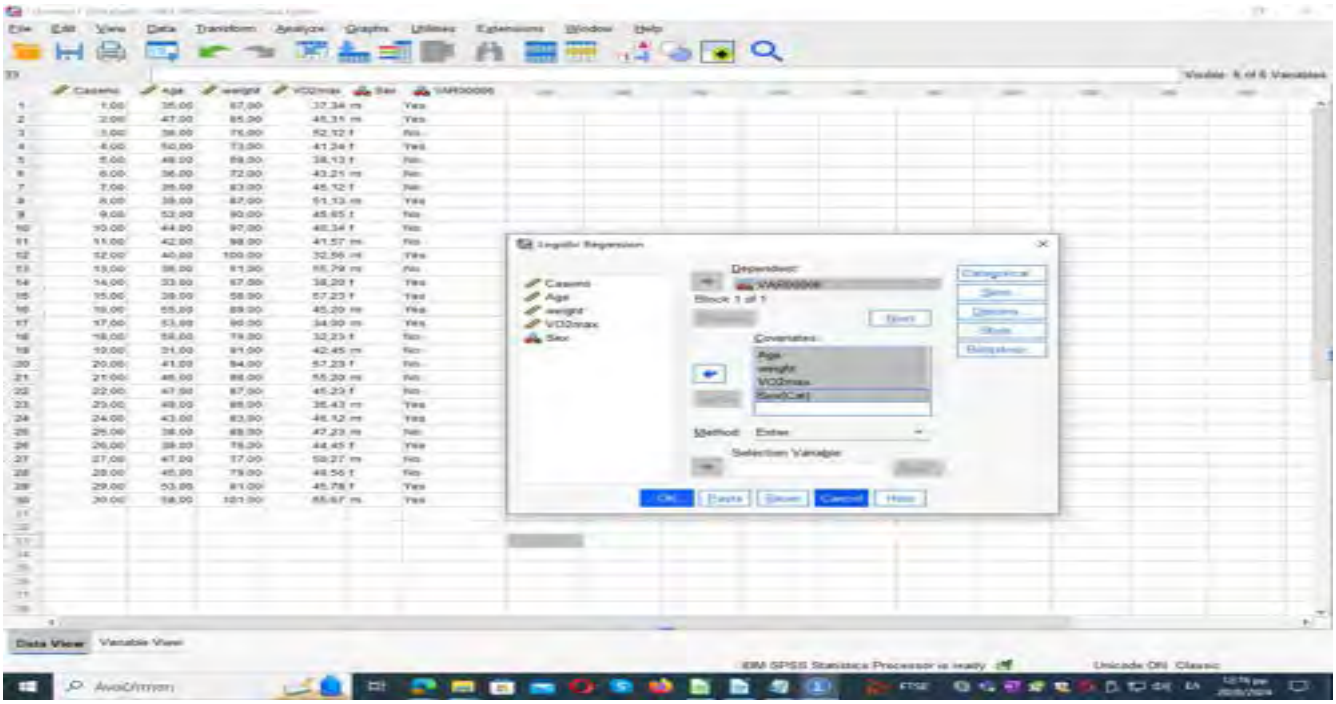
Εικόνα 4.17

Οπότε θα πάρουμε την παρακάτω Εικόνα 4.18:



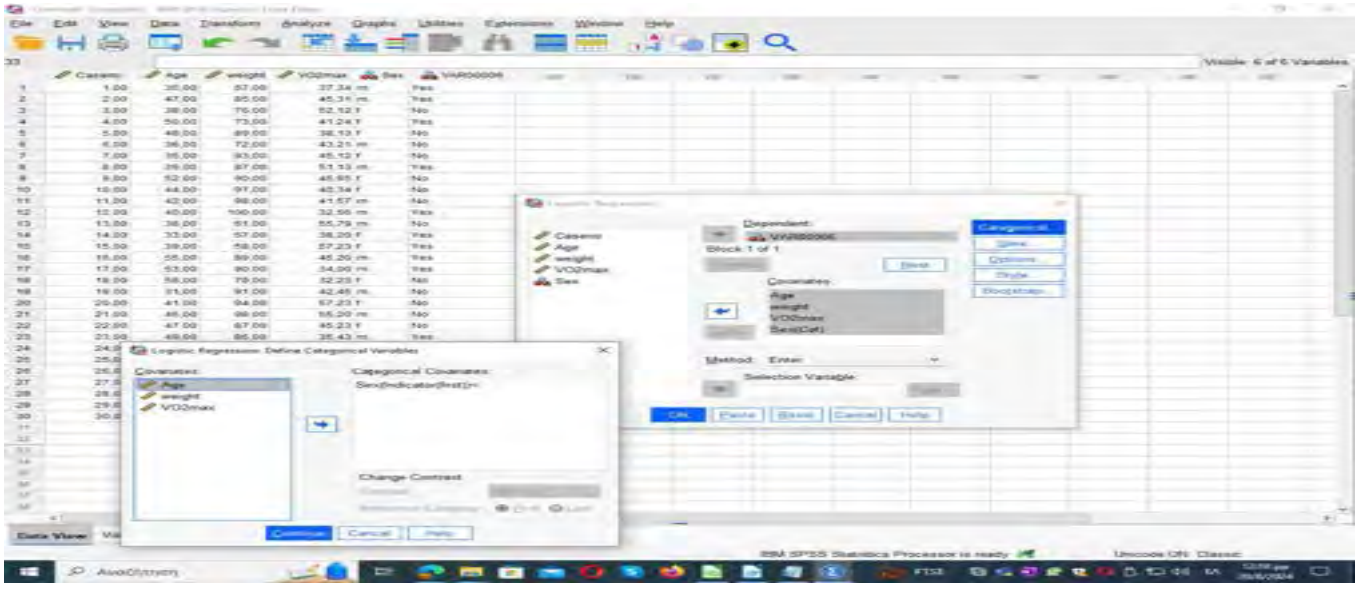
Εικόνα 4.18

Βήμα 2. Στην συνέχεια τοποθετούμε τις μεταβλητές μας όπως φαίνεται στην παρακάτω Εικόνα 4.19. Σημειώνουμε ότι η μεταβλητή με το όνομα VAR00006 δηλώνει καρδιακή_ασθένεια, δηλαδή εάν ο συμμετέχων έχει καρδιακή νόσο: "ναι" ή "όχι" (δηλαδή, η εξαρτημένη μεταβλητή).



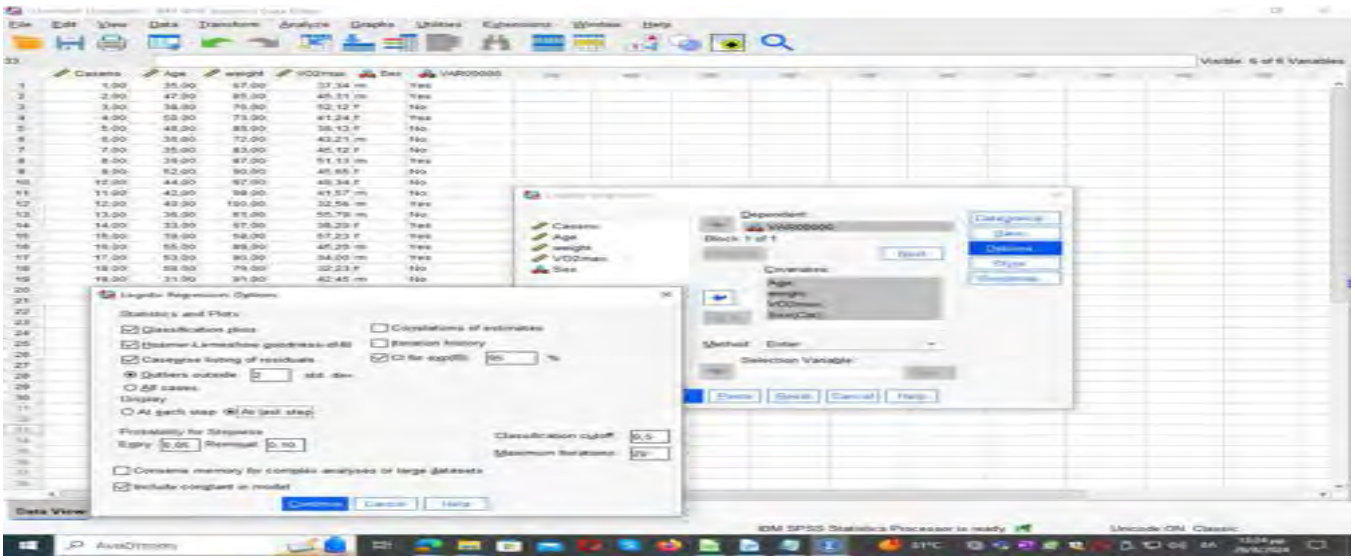
Εικόνα 4.19

Βήμα 3. Στην συνέχεια κάνουμε κλικ στο κουμπί **Categorical…**. οπότε, θα εμφανιστεί το πλαίσιο διαλόγου **Logistic Regression: Define Categorical Variables**, όπως φαίνεται παρακάτω (Εικόνα 4.20):



Εικόνα 4.20

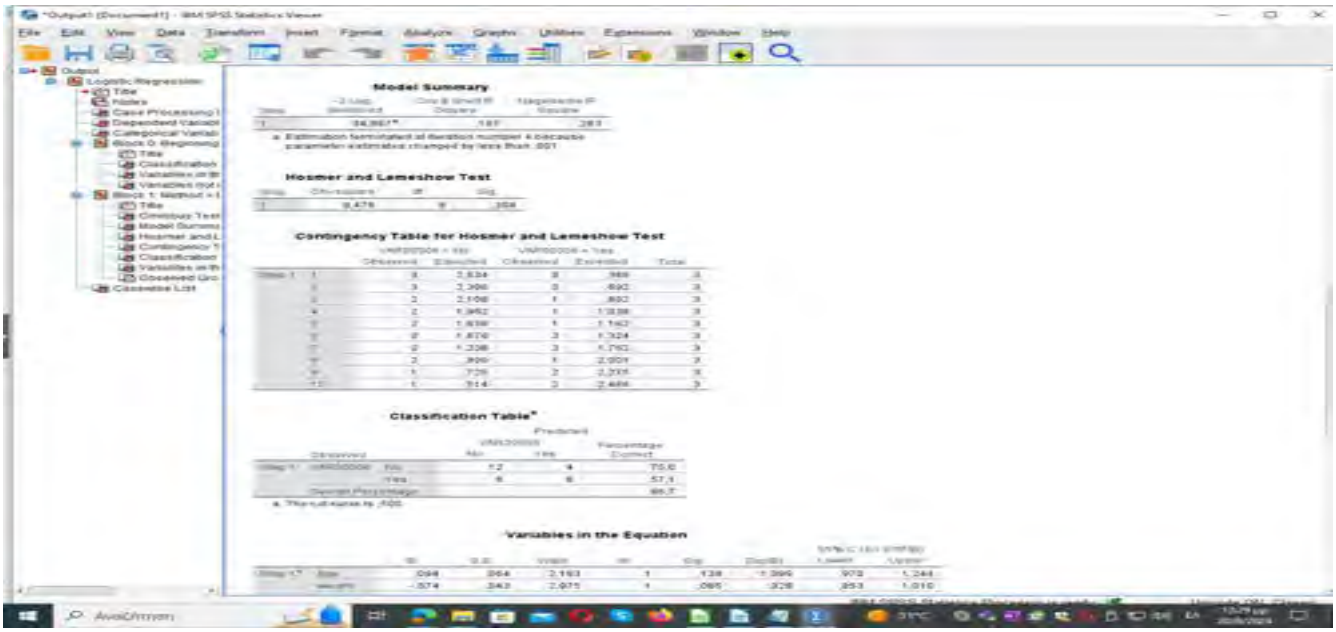
Στην συνέχεια επιλέγουμε **Options** και θα εμφανιστεί το πλαίσιο διαλόγου **Logistic Regression: Options**, όπως φαίνεται παρακάτω (Εικόνα 4.21):



Εικόνα 4.21

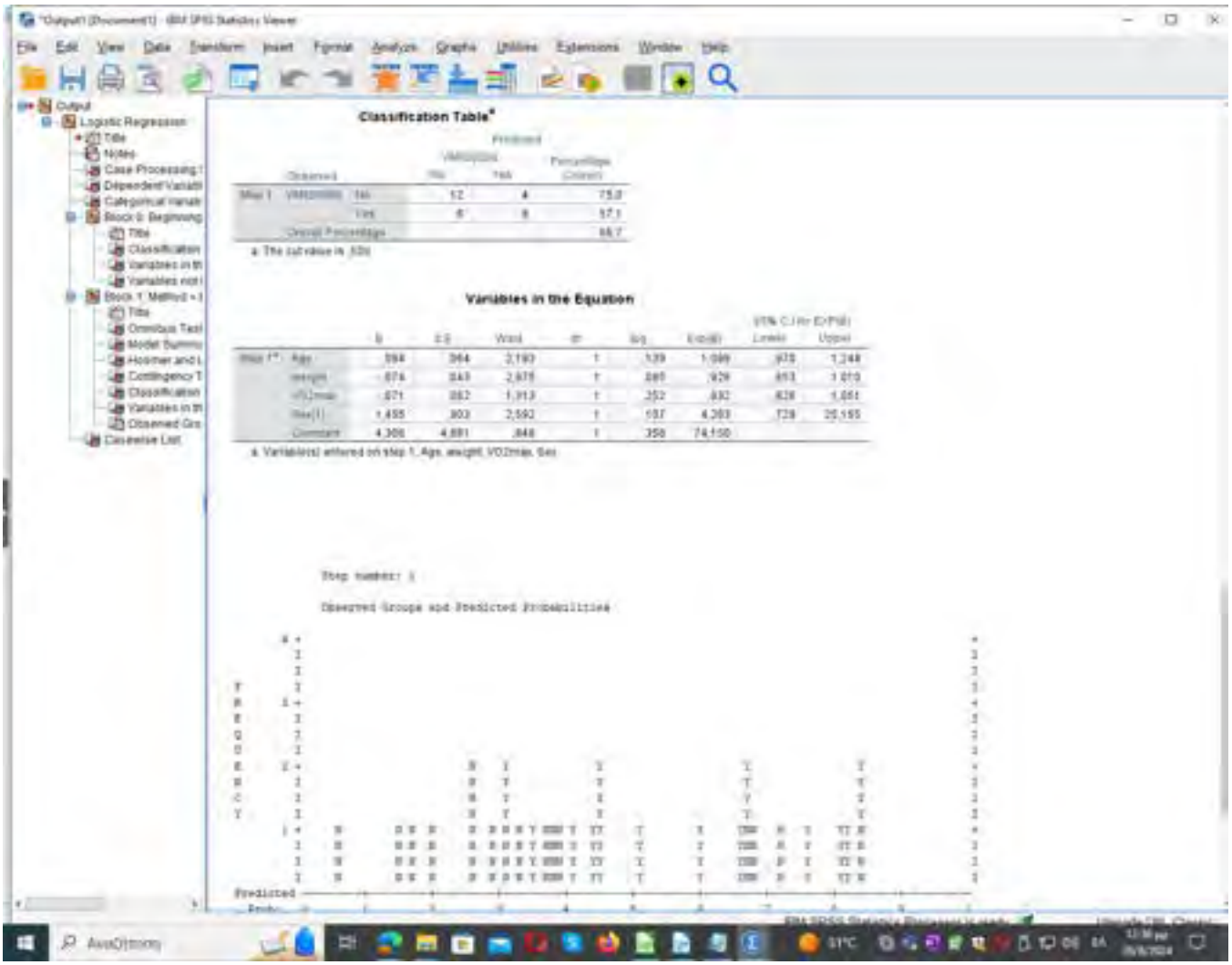
Και τέλος πατάμε το κουμπί **OK**.

Το SPSS Statistics δημιουργεί πολλούς πίνακες εξόδου κατά την εκτέλεση διωνυμικής λογιστικής παλινδρόμησης. Σε αυτήν την ενότητα, δείχνουμε μόνο τους τρεις κύριους πίνακες που απαιτούνται για την κατανόηση των αποτελεσμάτων από τη διαδικασία διωνυμικής λογιστικής παλινδρόμησης (Εικόνα 4.22).



Εικόνα 4.22

Ο πίνακας **Model Summary** περιέχει τις τιμές Cox & Snell R Square και Nagelkerke R Square. Στην συνέχεια ενδιαφέρον παρουσιάζει ο παρακάτω πίνακας (Εικόνα 4.23):



Εικόνα 4.23

Στον πίνακα **Classification Table** βλέπουμε την ταξινόμηση των 30 ασθενών και τις αντίστοιχες πιθανότητες και ο πίνακας **Variables in the Equation** δείχνει τη συμβολή κάθε ανεξάρτητης μεταβλητής στο μοντέλο και τη στατιστική της σημασία. Αυτός ο πίνακας φαίνεται παραπάνω. Η στατιστική σημασία του τεστ βρίσκεται στη Sig. στήλη. Από αυτά τα αποτελέσματα μπορούμε να εξάγουμε συμπεράσματα για τη σύνδεση της καρδιακής νόσου με την ηλικία (λαμβάνοντας p = 0,139), το φύλο (λαμβάνοντας p = 0,107), το VO2max (λαμβάνοντας p = 0,252) και το βάρος (λαμβάνοντας p = 0,085).

ΚΕΦΑΛΑΙΟ 5ο
Ανάλυση επιβίωσης (μέθοδος Kaplan-Meier) στο πρόγραμμα SPSS

Η διπλωματική εργασία ολοκληρώνεται με την παρουσίαση της μεθόδου Kaplan-Meier μέσα από το στατιστικό πακέτο SPSS. Η περιγραφή της σύνδεσης αυτής της μεθόδου με το SPSS μέσα από ένα παράδειγμα της Ιατρικής μας δίνει τη σπουδαιότητα τόσο της μεθόδου όσο και του SPSS για την εξαγωγή σημαντικών και χρήσιμων συμπερασμάτων.

5.1 Η Ανάλυση επιβίωσης (μέθοδος Kaplan-Meier) - Γενικά

Η μέθοδος Kaplan-Meier είναι μια μη παραμετρική μέθοδος που χρησιμοποιείται για την εκτίμηση της πιθανότητας επιβίωσης πέραν δεδομένων χρονικών σημείων.

Για παράδειγμα, σε μια μελέτη σχετικά με την επίδραση της δόσης του φαρμάκου στην επιβίωση του καρκίνου, μπορούμε να χρησιμοποιήσουμε τη μέθοδο Kaplan-Meier για να κατανοήσουμε την κατανομή επιβίωσης (με βάση το χρόνο μέχρι το θάνατο) που λαμβάνουν μία από τέσσερις διαφορετικές δόσεις φαρμάκου.

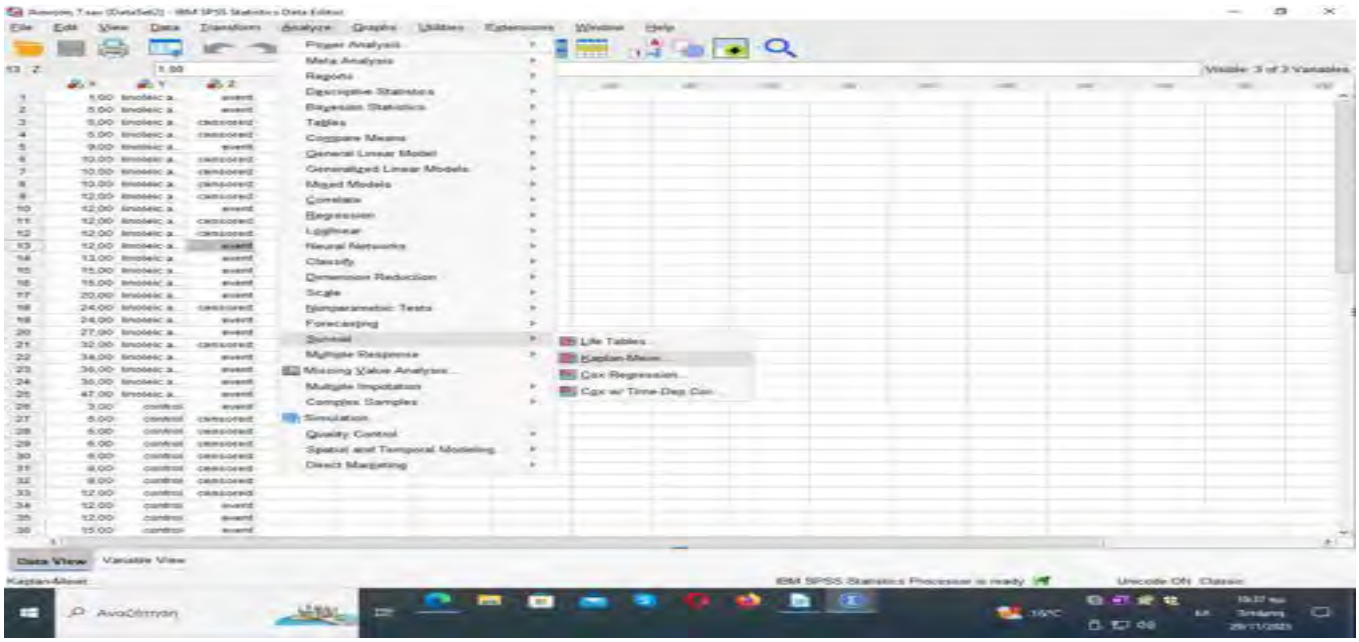
Επίσης με την μέθοδο Kaplan-Meier θα μπορούσαμε να δούμε αν μια νέα θεραπεία για το AIDS έχει κάποιο θεραπευτικό όφελος στην επέκταση της επιβίωσης του ασθενούς. Θα μπορούσαμε να πραγματοποιήσουμε μια μελέτη χρησιμοποιώντας δύο ομάδες ασθενών με AIDS, μία από τις οποίες λαμβάνει παραδοσιακή θεραπεία και η άλλη που λαμβάνει την πειραματική θεραπεία. Κατασκευάζοντας ένα μοντέλο Kaplan-Meier από τα δεδομένα θα μας επιτρέψει να συγκρίνουμε τα συνολικά ποσοστά επιβίωσης μεταξύ των δύο ομάδων και να προσδιορίσουμε εάν η πειραματική θεραπεία είναι μια βελτίωση σε σχέση με την παραδοσιακή θεραπεία.

5.2. Η Ανάλυση επιβίωσης (μέθοδος Kaplan-Meier) μέσα από ένα παράδειγμα

69 ασθενείς συμμετείχαν σε μια κλινική δοκιμή για τη θεραπεία του καρκίνου του παχέος εντέρου του Dukes, η οποία περιγράφεται από τους Mcllmurray και Turkie (a). Δίνονται τα δεδομένα για τις δύο θεραπείες. Θα εξετάσουμε με τη χρήση του Στατιστικού Προγράμματος SPSS αν υπάρχει σημαντική διαφορά στην επιβίωση μεταξύ των δύο θεραπειών. Σ' ό,τι ακολουθεί θα έχουμε τους παρακάτω συμβολισμούς:

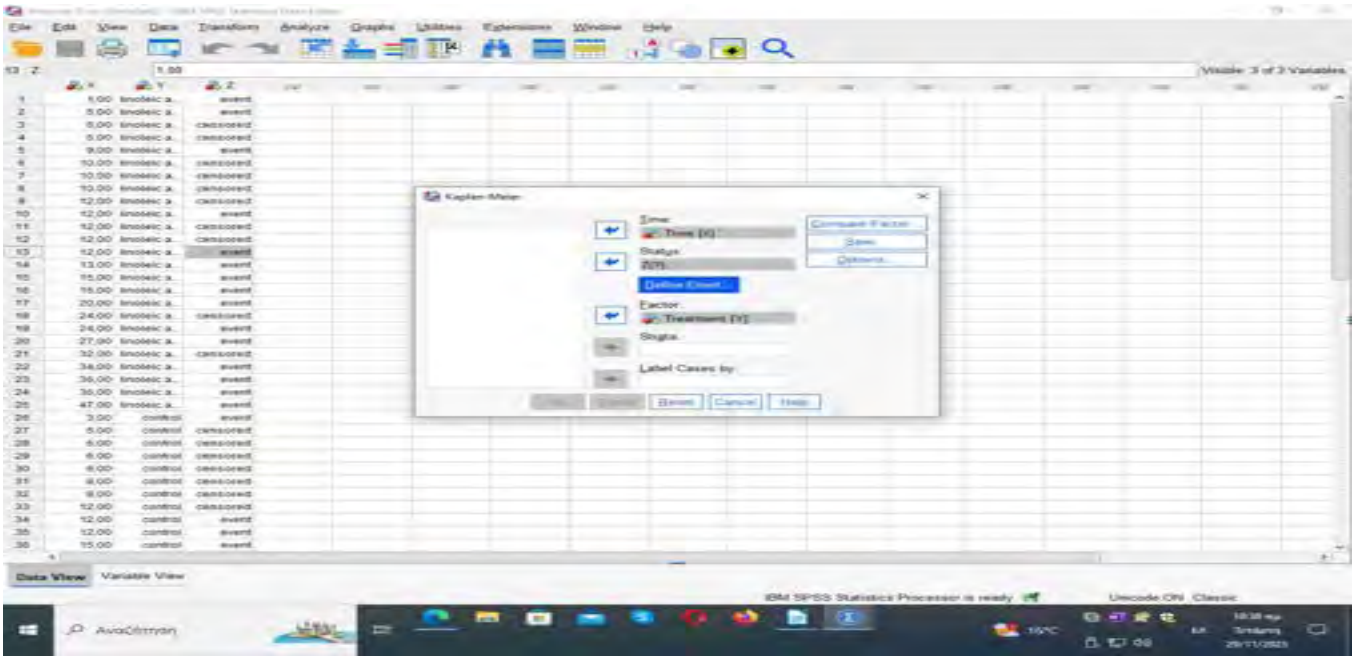
X: Χρόνος επιβίωσης σε μήνες.
Y: οι δυο μέθοδοι (βάζουμε 0 και 1 για να τις ξεχωρίζουμε)
Z: Όσοι επιβίωσαν.

Επιλέγουμε: **Analyze → Survival → Kaplan-Meier...**(Εικόνα 5.1):

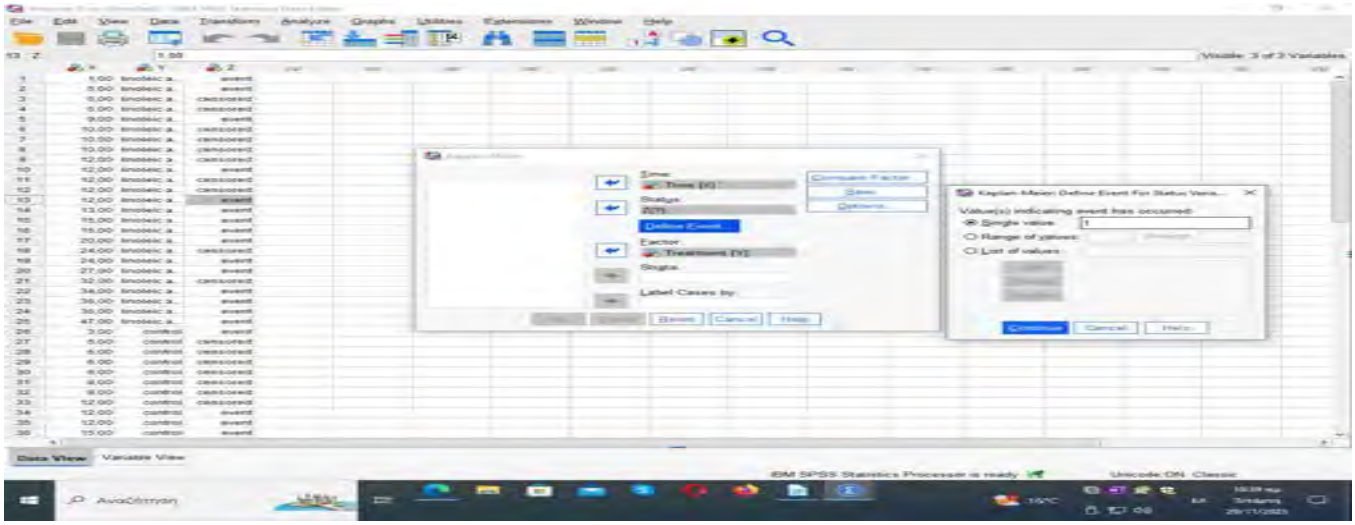


Εικόνα 5.1

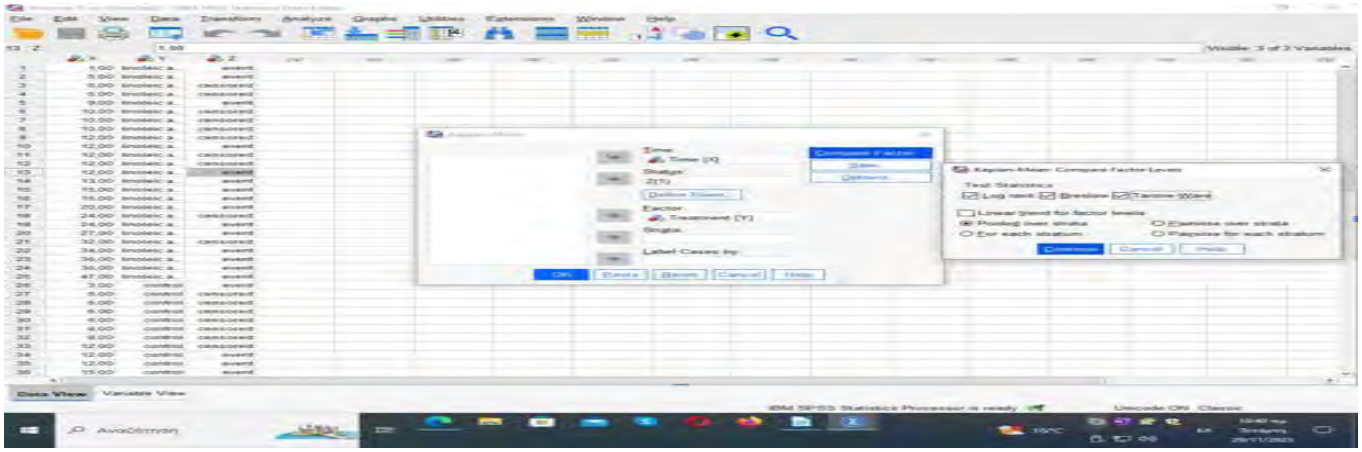
Στη συνέχεια κάνουμε τις επιλογές που φαίνονται στις παρακάτω εικόνες (Εικόνα 5.2-Εικόνα 5.4):



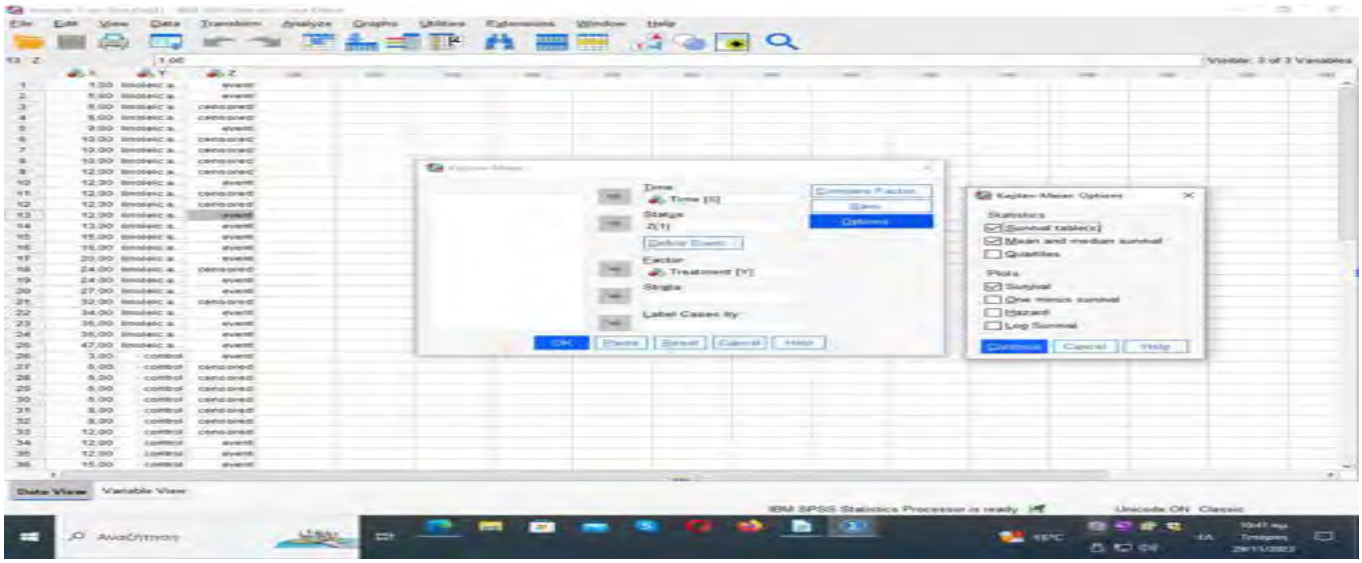
Εικόνα 5.2



Εικόνα 5.3

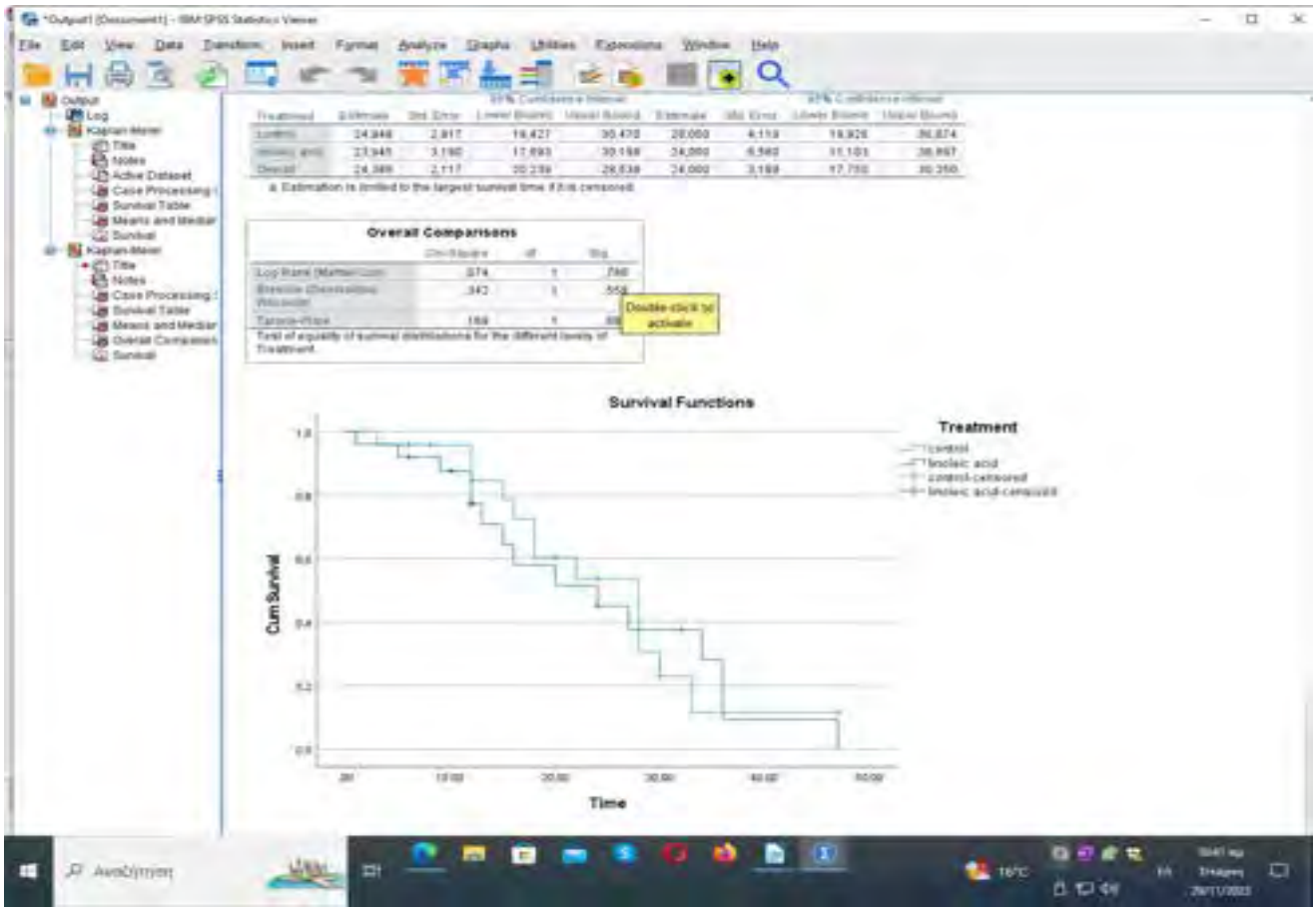


Εικόνα 5.4



Εικόνα 5.5

Τέλος πατάμε **OK** οπότε έχουμε τα παρακάτω αποτελέσματα της Εικόνας 5.6:



Εικόνα 5.6

Η παραπάνω γραφική παράσταση μας βοηθάει να κατανοήσουμε πώς συγκρίνονται οι κατανομές επιβίωσης μεταξύ των ομάδων. Μια χρήσιμη λειτουργία της γραφικής παράστασης είναι να απεικονίσει εάν οι καμπύλες επιβίωσης διασταυρώνονται μεταξύ τους (δηλαδή, εάν υπάρχει μια «αλληλεπίδραση» μεταξύ των κατανομών επιβίωσης. Επιπλέον, μπορούμε να δούμε εάν οι καμπύλες επιβίωσης έχουν παρόμοιο σχήμα, ακόμα κι αν είναι πάνω ή κάτω η μία από την άλλη.

Μια καμπύλη επιβίωσης ομάδας που εμφανίζεται «πάνω» από την καμπύλη επιβίωσης μιας άλλης ομάδας συνήθως θεωρείται ότι επιδεικνύει ευεργετική/πλεονεκτική επίδραση.

Επίσης πολύ σημαντικός είναι και ο πίνακας της παραπάνω εικόνας που εμφανίζεται πάνω από τις καμπύλες με τον τίτλο **Overall Comparisons**. Εάν στον πίνακα αυτό το $p < 0,05$, τότε έχουμε ένα στατιστικά σημαντικό αποτέλεσμα και μπορούμε να συμπεράνουμε ότι οι κατανομές επιβίωσης των διαφορετικών τύπων παρέμβασης δεν είναι ίσες στον πληθυσμό (δηλαδή, δεν είναι όλες ίδιες). Από την άλλη πλευρά, εάν $p > 0,05$, τότε δεν έχουμε στατιστικά σημαντικό αποτέλεσμα και δεν μπορούμε να συμπεράνουμε ότι οι κατανομές επιβίωσης είναι διαφορετικές στον πληθυσμό (δηλαδή, είναι όλες ίδιες/ίσες). Σε αυτό το παράδειγμα, αφού $p=0,786>0,05$ δεν υπάρχει σημαντική διαφορά στον χρόνο επιβίωσης μεταξύ των δυο μεθόδων.

Συμπεράσματα μελέτης

ΙΑΤΡΙΚΗ

Η **Ιατρική** είναι η επιστήμη που ασχολείται με την έρευνα και την εφαρμογή μεθόδων και τεχνικών για την πρόληψη, τη διάγνωση και τη θεραπεία των ασθενειών. Αναντίρρητα, αποτελεί έναν από τους θεμέλιους λίθους της ανθρωπότητας από την αρχαιότητα έως και τη σύγχρονη εποχή με τα επιτεύγματά της να βρίσκουν εφαρμογές στην ζωή των ανθρώπων προσπαθώντας να βελτιώσουν την καθημερινότητά τους.

Ωστόσο, πολλά είναι τα δεδομένα που καλούνται οι ιατροί και οι ερευνητές να αντιμετωπίσουν και να καταγράψουν σε καθημερινή βάση, τα οποία είναι δυνατό να τα συναντούν στον διαχωρισμό ασθενειών και φύλου (άνδρας-γυναίκα), στην ηλικία και στο βάρος ασθενών, στο ιατρικό ιστορικό τους, στην ομάδα αίματός τους, στην ασθένεια που έχει να αντιμετωπίσει ο κάθε ένας και η κάθε μία ασθενής κ.ο.κ. Συνεπώς, η ταξινόμησή τους και αργότερα η μελέτη τους από ερευνητικές ιατρικές ομάδες για την εξαγωγή συμπερασμάτων προκειμένου αυτά τα συμπεράσματα να χρησιμοποιηθούν:

- για την παραγωγή νέων φαρμάκων,
- τη βελτίωση της αποτελεσματικότητας των ήδη υπάρχοντων φαρμάκων,
- τη βελτίωση διαγνωστικών ελέγχων διαφόρων ασθενειών,
- τη δημιουργία νέων θεραπειών και
- τη βελτίωση της αποτελεσματικότητας ήδη υπάρχουσων θεραπειών

κρίνεται επιτακτική.

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

Συγχρόνως, στον κλάδο των Μαθηματικών, η **Στατιστική** γνωρίζει ιδιαίτερη άνθηση εντάσσοντας στους τομείς έρευνάς της νέα τεχνολογικά εργαλεία που έρχονται να εναρμονιστούν πλήρως με τις σύγχρονες επιταγές μίας τεχνολογικά αναπτυσσόμενης κοινωνίας.

Μπορούμε να πούμε ότι η Στατιστική δίνει τα απαραίτητα εργαλεία για την επίτευξη των παραπάνω στόχων της Ιατρικής. οι πρακτικές της Στατιστικής και ειδικότερα της Στατιστικής Ανάλυσης στην Ιατρική Επιστήμη βασίζονται:

- στο σχεδιασμό μελετών,
- στη συλλογή δεδομένων,
- στην επεξεργασία και ανάλυση δεδομένων,
- στην εξαγωγή συμπερασμάτων και τελικώς
- στην ερμηνεία αυτών των αποτελεσμάτων.

Έτσι λοιπόν, η Στατιστική αποτελεί έναν από τους πυρήνες της ιατρικής, αλλά και κάθε άλλης επιστημονικής, μεθοδολογίας.

SPSS ΣΤΗΝ ΙΑΤΡΙΚΗ

Στην **διπλωματική αυτή εργασία** κατορθώνουμε να συγκεράσουμε την Ιατρική με τη Στατιστική Ανάλυση αναδεικνύοντας την άρρηκτη σύνδεσή τους, η οποία είναι δυνατό να προσφέρει στη σύγχρονη Ιατρική νέες τεχνολογικές μεθόδους μελέτης ιατρικών θεμάτων και την εξαγωγή ασφαλών συμπερασμάτων για την εξέλιξη και πρόοδο ιατρικών ερευνών.

Το στατιστικό εργαλείο που έχουμε στη διάθεσή μας είναι το, ευρέως γνωστό και χρησιμοποιούμενο λογισμικό Στατιστικής Ανάλυσης, SPSS. Μελετάμε στατιστικές μεθόδους επεξεργασίας ιατρικών δεδομένων και ειδικότερα:

- την μέθοδο clustering (ομαδοποίηση) και τύπους μεθόδων ομαδοποίησης,

- την τεχνική της διαχωριστικής ανάλυσης,
- τη γραμμική και λογιστική παλινδρόμηση και
- την μέθοδο Kaplan-Meier περί εκτίμησης της πιθανότητας επιβίωσης πέραν δεδομένων χρονικών σημείων.

μέσα από το στατιστικό πακέτο SPSS.

Με αυτόν τον τρόπο πετυχαίνουμε να μελετήσουμε ιατρικά δεδομένα, ποικίλης φύσης, με μία διαφορετική οπτική. Διαπιστώνουμε ότι η καταγραφή των δεδομένων μας στο SPSS και στη συνέχεια, η μελέτη τους μέσα από τα εργαλεία που μας παρέχει το λογισμικό αυτό προσφέρει πιο ασφαλή συμπεράσματα από μία συνήθη καταγραφή τέτοιων δεδομένων ενδεχομένως σε, παρωχημένες για την εποχή μας, μεθόδους που ο όγκος των δεδομένων μπορεί να οδηγήσει σε εσφαλμένες εκτιμήσεις για μελλοντικές θεραπείες ή διαγνώσεις.

Συμπερασματικά, οι ιατρικές μελέτες που βασίζονται σε δεδομένα τα οποία μελετώνται στο SPSS οδηγούν σε καλύτερα αποτελέσματα για την επίτευξη στόχων που θα βελτιώσουν τις ήδη υπάρχουσες συνθήκες ιατρικών ερευνών περί φαρμάκων, θεραπειών, πρόληψης και διάγνωσης ασθενειών. Άλλωστε, στις μέρες μας, όλοι αντιλαμβανόμαστε πως πλέον η σύγχρονη κοινωνία και κάθε είδους δραστηριότητα που την περικλείει, από τις Επιστήμες έως και την καθημερινότητά μας, εξελίσσεται τεχνολογικά. Η εξέλιξη αυτή αντανακλάται και στις ιατρικές μελέτες με σύγχρονα λογισμικά να έρχονται να συνδυαστούν με τη γνώση ιατρών και ερευνητών για να επιτευχθούν τα καλύτερα αποτελέσματα για τους ανθρώπους, σεβόμενοι πάντα την ανθρώπινη αξία.

Βιβλιογραφία

[1] Βιτωράτου Σ. & Λύκου Α., *Σημειώσεις σεμιναρίου SPSS*.

[2] Δάρας Τ., Αποστολάκη Ι., Σταμούλη Μ.Α., *Ασκήσεις SPSS στην Υγεία*, Εκδόσεις Παπαζήση.

[3] Δαφέρμος Β., *Κοινωνική στατιστική με το SPSS*, Εκδόσεις Ζήτη (2005).

[4] Δαφέρμος Β., *Επαναληπτικές στατιστικές μετρήσεις στις κοινωνικές επιστήμες*, Εκδόσεις Leader Books. Αθήνα (2002).

[5] Ζαφειρόπουλος Κ., Μυλωνάς Ν., *Στατιστική με SPSS*, Εκδόσεις Τζιόλα (2017).

[6] Ζιντζαράς Η., *Εισαγωγή στο SPSS*, Πανεπιστημιακές Σημειώσεις (Αυγουστος 2024).

[7] Καλαματιανού Α., *Κοινωνική Στατιστική, Μέθοδοι Μονοδιάστατης Ανάλυσης,* Εκδόσεις Παπαζήση (2003).

[8] Μπερσίμης Σ., *Διδακτικές σημειώσεις προγράμματος επιμόρφωσης στην ιατρική στατιστική* (2007).

[9] Ντζούφρας Ι., *Εισαγωγή στη Βιοστατιστική και την Επιδημιολογία,* Πανεπιστημιακές σημειώσεις (2005).

[10] Ξεκαλάκη Ε., *Μη Παραμετρική Στατιστική*, Αθήνα (2001).

[11] Ξεκαλάκη Ε., *Τεχνικές Δειγματοληψία,* Εκδόσεις οικονομικού Πανεπιστημίου Αθηνών (2004).

[12] Παπαϊωάννου Τ., Φερεντίνος Κ., *Ιατρική Στατιστική και Στοιχεία Βιομαθηματικών*, Βελτιωμένη Έκδοση, Εκδόσεις Σταμούλη (2000).

[13] Παυλόπουλος Β., *Εισαγωγή στη Στατιστική Επεξεργασία Δεδομένων με το SPSS για Windows* (2006).

[14] Παυλόπουλος Β., *Μοντέλα Ανάλυσης Διακύμανσης* (2008).

[15] Ρούσσος Π., Ευσταθίου Γ., *Σύντομο Εγχειρίδιο SPSS 16.0*, Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (2008).

[16] Συμεωνάκη Μ., *Στατιστική για όλους με το SPSS*, Εκδόσεις Σοφία (2015).

[17] Τσαγρής Μ., *Ανάλυση Διακύμανσης στο SPSS,* Διδακτικές σημειώσεις, (2006).

[18] Χαλικιάς Ι., *Στατιστική: Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις,* 2η Έκδοση, Rosili, Αθήνα (2003).

[19] Barton Belinda, Medical *Statistics: A Guide to SPSS, Data Analysis and Critical Appraisal Paperback,* 2nd Edition (2014).

[20] Bryman Alan, Cramer Duncan, *Quantitative Data Analysis with IBM SPSS 17, 18 and 19: A Guide for Social Scientists,* New York: Routledge (2011).

[21] Field A. P. , *Discovering statistics using SPSS,* Second Edition, London: Sage (2005).

[22] Gunarto Hary, *Parametric & Nonparametric Data Analysis for Social Research: IBM SPSS,* LAP Academic Publishing (2019).

[23] Hejase A.J., Hejase, H.J., *Research Methods, A Practical Approach for Business Students,* 2nd edn. Philadelphia, PA, USA: Masadir Inc. (2013).

[24] James Allen, Kellie Bennett, *SPSS for the health & behavioural sciences*, Publisher: Thomson, Australia (2008).

[25] Landau S., Everitt B.S., *A Handbook of Statistical Analyses using SPSS,* Chapman and Hall (2004).

[26] Levesque R., *SPSS Programming and Data Management: A Guide for SPSS and SAS Users,* 4th ed. Chicago, Illinois: SPSS Inc. (2007).

[27] Louise Marston, *Introductory Statistics for Health and Nursing Using SPSS*, SAGE Publications Ltd (2010).

[28] Nie Norman H, Bent Dale H., Hadlai Hull C, *SPSS: Statistical package for the social sciences,* McGraw-Hill (1970).

[29] Nie Norman H., *SCSS: A User's Guide to the SPSS Conversational Statistical System,* McGraw-Hill (1980).

[30] Pallant J., *SPSS: οδηγός ανάλυσης δεδομένων με το IBM SPSS*, Εκδόσεις Κλειδάριθμος, Έκδοση 7.

[31] Sachdev Ameet, *IBM's $1.2 billion bid for SPSS Inc helps resolve trademark dispute,* Chicago Tribune, (27 September, 2009).

[32] Vijay Gupta, *SPSS for Beginners*, Published by VJBooks Inc. (1999).

[33] *What's New in SPSS Statistics 29*,. community.ibm.com (12 September, 2022).