



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Σχεδίαση και υλοποίηση Jupyter Notebook για την
παρουσίαση και εκμάθηση της βιβλιοθήκης
οπτικοποίησης δεδομένων Seaborn

ΒΑΣΙΛΙΚΗ Β. ΒΑΛΙΑΤΖΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Αντώνης Δαδαλιάρης
Επίκουρος Καθηγητής

Λαμία 2025



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Σχεδίαση και υλοποίηση Jupyter Notebook για την
παρουσίαση και εκμάθηση της βιβλιοθήκης
οπτικοποίησης δεδομένων Seaborn

ΒΑΣΙΛΙΚΗ Β. ΒΑΛΙΑΤΖΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Αντώνης Λαδαλιάρης
Επίκουρος Καθηγητής

Λαμία, 2025



UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

MASTER OF SCIENCE IN EDUCATIONAL APPLICATIONS WITH STEM
EPISTEMOLOGY

Design and implementation of a Jupyter Notebook for
the presentation and learning of the Seaborn data
visualization library

VASILIKI V. VALIATZA

FINAL THESIS

ADVISOR

Antonios N. Dadaliaris

Assistant Professor

Lamia 2025

Υπεύθυνη Δήλωση Συγγραφέα:

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις ⁽¹⁾, που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.
2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.
3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια
4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: 02 / 06 /2025

Θ – Η Δηλ.

1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

,

Αφιέρωση

Αφιερώνω τη διπλωματική μου στον κ. Ευθύμη Βαϊόπουλο, τον καθηγητή χωρίς αίθουσα, που μου έδειξε ότι τα δεδομένα κρύβουν ιστορίες που περιμένουν να τις αφηγηθούμε. Οι καινοτόμες ιδέες του όχι μόνο διαμόρφωσαν αυτή την εργασία, αλλά εφαρμόστηκαν με επιτυχία στην τάξη μου. Αυτή η εργασία είναι η απόδειξη ότι δάσκαλοι δεν βρίσκονται μόνο στις αίθουσες διδασκαλίας, αλλά παντού όπου υπάρχει όραμα, καινοτομία και πάθος για πρόοδο.

Τον ευχαριστώ που μου έμαθε να "ζωγραφίζω" με αριθμούς και να μεταφράζω την πολυπλοκότητα σε απλές αλήθειες.

Περίληψη

Σκοπός της διπλωματικής εργασίας είναι η σχεδίαση και υλοποίηση σε περιβάλλον Jupyter Notebook ενός οδηγού μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn. Στο παρόν κείμενο παρουσιάζεται, σε αδρές γραμμές, ο οδηγός μελέτης και το περιβάλλον εκτέλεσής του (Jupyter Notebook).

Το Jupyter Notebook είναι ένα διαδικτυακό εργαλείο που επιτρέπει τη δημιουργία και κοινή χρήση αλληλεπιδραστικών εγγράφων με ζωντανό κώδικα, εικόνες και αφήγηση κειμένου, καθιστώντας το ιδανικό για ανάλυση δεδομένων και εκπαίδευση. Χάρη στον ενεργό κώδικα, τα έγγραφα που συντάσσουμε με το Jupyter Notebook είναι υπολογιστικά και διαδραστικά, τα αποκαλούμε, δε, υπολογιστικά σημειωματάρια ή απλά σημειωματάρια (notebooks).

Ο οδηγός μελέτης οπτικοποίησης δεδομένων με τη βιβλιοθήκη Seaborn έχει υλοποιηθεί στο περιβάλλον Jupyter Notebook σαν υπολογιστικό σημειωματάριο, με όνομα αρχείου "Data Visualization with Seaborn.ipynb". Οι ενδιαφερόμενοι μπορούν να μελετήσουν λειτουργίες οπτικοποίησης δεδομένων της βιβλιοθήκης Seaborn αλληλεπιδρώντας με το σημειωματάριο μελέτης (τον οδηγό μελέτης) σε περιβάλλον Jupyter Notebook. Το διαδραστικό υλικό του οδηγού μελέτης μπορεί να χρησιμοποιηθεί ελεύθερα είτε μέσω προσωπικής, τοπικής εγκατάστασης του περιβάλλοντος Jupyter Notebook είτε μέσα από τις διαθέσιμες κοινόχρηστες εγκαταστάσεις στο νέφος (όπως το Google Colab). Για διευκόλυνση των χρηστών περιλαμβάνουμε στο παρόν κείμενο συνοπτική περιγραφή των βασικών δυνατοτήτων του περιβάλλοντος Jupyter Notebook για δημιουργία και χρήση σημειωματαρίων (notebooks).

Η βιβλιοθήκη Seaborn είναι ένα ισχυρό εργαλείο οπτικοποίησης δεδομένων. Αλληλεπιδρώντας με τον οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn

ο χρήστης έχει τη δυνατότητα να εντρυφήσει σε έννοιες και μεθόδους οπτικοποίησης δεδομένων και να εξασκηθεί σε βασικές τεχνικές και είδη γραφημάτων που υποστηρίζει η Seaborn, όπως διαγράμματα διασποράς παρατηρήσεων, διαγράμματα συσχέτισης και κατανομής μεταβλητών αλλά και κατηγορικά διαγράμματα (διαγράμματα διασποράς, κατανομών και κεντρικής τάσης) και διαγράμματα εξέλιξης.

Στον οδηγό μελέτης δίνεται έμφαση στις ρυθμίσεις των λειτουργιών (συναρτήσεων) σχεδίασης της βιβλιοθήκης Seaborn, που ελέγχουν την απεικόνιση στατιστικής πληροφορίας με γραφικά. Οι απεικονίσεις της βιβλιοθήκης περιλαμβάνουν γραφικά στοιχεία όπως καμπύλες σχέσεων και κατανομών, ράβδους ιστογραμμάτων, σημεία κεντρικής τάσης και μπάρες ή ζώνες σφάλματος/αβεβαιότητας. Παράλληλα, μελετώνται οι ρυθμίσεις των συναρτήσεων σχεδίασης που επιτρέπουν την άμεση άντληση των μεταβλητών ενδιαφέροντος από σύνολα δεδομένων και τη σημασιολογική διάκριση των γραφικών για επιμέρους ομάδες παρατηρήσεων μέσω διαφορετικών χρωμάτων και διαγραμμάτων. Προτεραιότητα του οδηγού μελέτης αποτελεί η επεξήγηση των εικονιζόμενων γραφικών και η ερμηνεία των απεικονίσεων.

Ο οδηγός μελέτης περιλαμβάνει πρακτικά παραδείγματα με ζωντανό κώδικα, που διευκολύνουν την κατανόηση των λειτουργιών της βιβλιοθήκης και των σχεδιαστικών αρχών που διέπουν την οπτικοποίηση δεδομένων. Προτείνεται ο διαμοιρασμός του οδηγού μελέτης μέσα από συνεργατικά περιβάλλοντα, όπως το Google Colab, που επιτρέπουν την εκτέλεση του κώδικα στο cloud και υποστηρίζουν τη συνεργασία και την ομαδική μελέτη.

Η αξιοποίηση του περιβάλλοντος Jupyter Notebook και της βιβλιοθήκης Seaborn επέτρεψε τη σύνθεση ενός ολοκληρωμένου οδηγού για εκμάθηση και εφαρμογή τεχνικών οπτικοποίησης δεδομένων, που ευελπιστούμε θα αποτελέσει εργαλείο μελέτης για ερευνητές και επαγγελματίες στο χώρο της επιστήμης δεδομένων (data science).

,

Abstract

The purpose of this thesis is to design and implement a study guide for the Seaborn data visualization library in the Jupyter Notebook environment. This text presents, in broad terms, the study guide and its execution environment (Jupyter Notebook).

Jupyter Notebook is an online tool that allows the creation and sharing of interactive documents with live code, images and text narration, making it ideal for data analysis and education. Thanks to the active code, the documents we develop with Jupyter Notebook are computational and interactive, and we call them computational notebooks or simply notebooks.

The study guide for data visualization with the Seaborn library has been implemented in the Jupyter Notebook environment as a computational notebook, with the file name "Data Visualization with Seaborn.ipynb". Interested parties can study data visualization functions of the Seaborn library by interacting with the study notebook (the study guide) in the Jupyter Notebook environment. The interactive material of the study guide can be used freely either through a personal, local installation of the Jupyter Notebook environment or through the available shared installations in the cloud (such as Google Colab). For the convenience of users, we include in this text a brief description of the basic capabilities of the Jupyter Notebook environment for creating and using notebooks.

The Seaborn library is a powerful data visualization tool. By interacting with the study guide of the Seaborn data visualization library, the user has the opportunity to delve into concepts and methods of data visualization and to practice basic techniques and types of graphs supported by Seaborn, such as scatter plots, regression plots, distribution plots, as well as categorical plots (scatter, distribution and central tendency plots) and evolution (line) plots.

The study guide emphasizes the parameters of the plotting functions that control the visualization of statistical information with graphs. The library's visualizations include graphics such as regression and distribution curves, histogram bars, points of central tendency, and error/uncertainty bars or bands. At the same time, the parameters of the plotting functions that allow the extraction of variables of interest from data sets and the semantic distinction of the graphics for separate groups of observations (through different colors and diagrams) are studied. The study guide's priority is the explanation of the depicted graphics and the interpretation of the visualizations.

At the same time, the settings of the design functions that allow the direct extraction of the variables of interest from data sets and the semantic distinction of the graphics for individual groups of observations through different colors and diagrams are studied.

The study guide includes practical examples with active code, which facilitate the understanding of the library's functions and the design principles that govern data visualization. It is recommended that the study guide be shared through collaborative environments, such as Google Colab, which allow code execution in the cloud and support collaboration and group study.

The use of the Jupyter Notebook environment and the Seaborn library allowed the composition of a comprehensive guide for learning and applying data visualization techniques, which we hope will become a study tool for researchers and professionals in the field of data science.

Πίνακας περιεχομένων

Περιεχόμενα

Περίληψη.....	3
Abstract.....	5
Πίνακας περιεχομένων	7
Κατάλογος Εικόνων	9
Εισαγωγή.....	14
A. Jupyter Notebook	18
Απαραίτητες βιβλιοθήκες.....	22
Είσοδος στο περιβάλλον Jupyter Notebook	23
Εργασία στο περιβάλλον Jupyter Notebook	24
Δημιουργία σημειωματάριου.....	26
Η καρτέλα Running	27
Κεφαλίδα	27
Κύριο Μέρος	28
Συγγραφή σημειωματάριου	29
Εμπλουτισμός περιεχομένου στα κελιά.....	31
Εξαγωγή notebook	31
B. Βιβλιοθήκη οπτικοποίησης δεδομένων Seaborn	33
Σχέσεις μεταβλητών	35
Σχέση δύο ποσοτικών μεταβλητών	36
Σχέση κατηγορικής-ποσοτικής μεταβλητής - κατηγορικά διαγράμματα	43
Εξέλιξη μεταβλητής με το χρόνο	46
Κατανομές μεταβλητών	49
Κατανομή κατηγορικής μεταβλητής	49
Κατανομές ποσοτικών μεταβλητών.....	57
Εξερεύνηση δεδομένων	80
Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (joint plot)	81
Συνδυαστική οπτικοποίηση μεταβλητών κατά ζεύγη (pair plot).....	86
Γ. Οδηγός Μελέτης.....	93
Ψηφιακά αρχεία	93
1. Οδηγός μελέτης "Data Visualization with Seaborn.ipynb"	93
2. Οδηγός μελέτης "Data Visualization with Seaborn.pdf"	93

Σύνοψη 94

Βιβλιογραφία..... 99

 Δεδομένα που χρησιμοποιούνται 99

Κατάλογος Εικόνων

Εικόνα 1 Εκτελεσμένο υπολογιστικό κελί	18
Εικόνα 2 Γραμμή εντολών	24
Εικόνα 3 Εκτελέσιμο αρχείο	24
Εικόνα 4 Εκκίνηση Jupyter Notebook	25
Εικόνα 5 Πλήκτρα επιλογών	25
Εικόνα 6 Δημιουργία νέου σημειωματάριου	26
Εικόνα 7 Επιλογή Kernel	26
Εικόνα 8 Καρτέλα Running	27
Εικόνα 9 Κουμπί τερματισμού λειτουργίας	27
Εικόνα 10 Κεφαλίδα	28
Εικόνα 11 Κελιά	29
Εικόνα 12 Κώδικας	30
Εικόνα 13 Εκτέλεση κώδικα	30
Εικόνα 14 Εργαλεία μέσα στο κελί	30
Εικόνα 15 Εξαγωγή αρχείων	32
Εικόνα 16 Κατηγοριοποίηση των συναρτήσεων σχεδίασης της βιβλιοθήκης Seaborn	33
Εικόνα 17 Η συνάρτηση load_dataset	35
Εικόνα 18 Κλήση της συνάρτησης σχεδίασης διαγραμμάτων διασποράς	36
Εικόνα 19 Διαγράμματα διασποράς για γεύματα και δείπνα	37
Εικόνα 20 Διαγράμματα διασποράς με διαφοροποίηση σε διάφορες ομάδες	38
Εικόνα 21 Κλήση της συνάρτησης σχεδίασης συσχέτισης	40
Εικόνα 22 Διαγράμματα συσχέτισης για γεύματα και δείπνα	40
Εικόνα 23 Διαγράμματα κατανάλωσης και ιπποδύναμης (μή γραμμική σχέση)	42
Εικόνα 24 Διαγράμματα συσχέτισης και αποκλίσεων όπου φαίνεται μη γραμμική σχέση	42
Εικόνα 25 Κλήση της συνάρτησης σχεδίασης διασποράς των αποκλίσεων (residplot)	43
Εικόνα 26 Διαγράμματα διασποράς του μεγέθους παρέας ('size') με το ποσό λογαριασμού ('total_bill')	44
Εικόνα 27 Κλήση της συνάρτησης σχεδίασης συσχέτισης	44
Εικόνα 28 Διαγράμματα διασποράς του μεγέθους παρέας ('size') με το ποσό λογαριασμού ('total_bill')	45

Εικόνα 29 Κλήση της συνάρτησης σχεδίασης κατηγορικού διαγράμματος διασποράς	46
Εικόνα 30 Κλήση της συνάρτησης σχεδίασης relplot (με kind='line')	47
Εικόνα 31 Διαγράμματα εξέλιξης του προσδόκιμου ζωής για άνδρες και γυναίκες	47
Εικόνα 32 Διάγραμμα εξέλιξης του προσδόκιμου ζωής σε αναπτυσσόμενες χώρες	48
Εικόνα 33 Κλήση της συνάρτησης σχεδίασης displot	50
Εικόνα 34 Κατανομή πιθανότητας επιβίωσης στο ναυάγιο του Τιτανικού. Η οπτικοποίηση περιέχει τρία διαγράμματα ένα για κάθε θέση (μεταβλητή 'class')	52
Εικόνα 35 Κατανομή πιθανότητας επιβίωσης στο ναυάγιο του Τιτανικού. Η οπτικοποίηση περιέχει τρία διαγράμματα ένα για κάθε θέση (μεταβλητή 'class')	53
Εικόνα 36 Πιθανότητα επιβίωσης (1) και αποβίωσης (0) για τους επιβάτες κάθε θέσης στο ναυάγιο του Τιτανικού	54
Εικόνα 37 Αλλαγή των επιμέρους ομάδων με αλλαγή των σημασιολογικών ρυθμίσεων	55
Εικόνα 38 Δημιουργία επιμέρους ομάδων σε διαφορετικά διαγράμματα και με διαφορετικά χρώματα	56
Εικόνα 38β Διαφορετικές κατανομές για διαφορετικά διαγράμματα και με διαφορετικά χρώματα	56
Εικόνα 39 Κλήση της συνάρτησης σχεδίασης κατανομών (displot)	58
Εικόνα 40 Ιστόγραμμα όπου απεικονίζουμε την κατανομή βάρους σώματος σε γραμμάρια από σύνολο δεδομένων με στοιχεία πιγκουίνων	60
Εικόνα 41 Κατανομή βάρους σώματος των πιγκουίνων με binwidth=200 που οδηγεί σε οπτικοποίηση με «θόρυβο»	62
Εικόνα 42 Συνδυαστική κατανομή δύο ποσοτικών μεταβλητών	63
Εικόνα 43 Ιστόγραμμα κατανομής του βάρους των πιγκουίνων με ράβδους διαφορετικών χρωμάτων που είναι τοποθετημένες η μία πάνω από την άλλη	65
Εικόνα 44 Ιστόγραμμα κατανομής του βάρους των πιγκουίνων με ράβδους διαφορετικών χρωμάτων που είναι τοποθετημένες η μία πίσω από την άλλη	66
Εικόνα 45 Ιστογράμματα κατανομής του βάρους των πιγκουίνων με ράβδους διαφορετικών χρωμάτων που είναι τοποθετημένες η μία πίσω από την άλλη για κάθε φύλο	67
Εικόνα 46 Ιστόγραμμα και καμπύλες κατανομής	68
Εικόνα 47 Έξι καμπύλες κατανομών. Στη οπτικοποίηση έχουμε ξεχωριστό διάγραμμα για τη σημασιολογική μεταβλητή φύλο των πιγκουίνων μέσα στο οποίο διακρίνονται με διαφορετικό χρώμα τα τρία είδη των πιγκουίνων (2x3=6)	69
Εικόνα 48 Κατανομή, κεντρική τάση, διασπορά	71

Εικόνα 49 Μέσοι όροι περιθώριο σφάλματος για κάθε είδους πιγκουίνου	71
Εικόνα 50 Κλήση της συνάρτησης σχεδίασης κατηγορικών διαγραμμάτων	72
Εικόνα 51 Δύο διαγράμματα εκτίμησης πυκνότητας με 6 κατανομές για τα είδη και το φύλο των πιγκουίνων	73
Εικόνα 52 Διάγραμμα σωρευτικής κατανομής πιθανότητας	75
Εικόνα 53 Κατανομή πιθανότητας δύο ποσοτικών μεταβλητών με ιστόγραμμα	77
Εικόνα 54 Πυκνότητα παρατηρήσεων (διασπορά) δύο ποσοτικών μεταβλητών	78
Εικόνα 55 Κατανομή πιθανότητας ζεύγους ποσοτικών μεταβλητών με εκτίμηση πυκνότητας	79
Εικόνα 56 Κλήση συνάρτησης σχεδίασης ζεύγους ποσοτικών μεταβλητών (jointplot) ..	81
Εικόνα 57 Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (jointplot): διασπορά και περιθώριες κατανομές	83
Εικόνα 58 Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (jointplot): κατανομές (στο κέντρο/περιθώρια)	84
Εικόνα 59 Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (jointplot): διασπορά, συσχέτιση και περιθώριες κατανομές	85
Εικόνα 60 Κλήση συνάρτησης σχεδίασης ζευγών ποσοτικών μεταβλητών (pair plot) ..	86
Εικόνα 61 Εξερεύνηση πολλών μεταβλητών με pair plot (διασπορά)	89
Εικόνα 62 Εξερεύνηση πολλών μεταβλητών με pair plot (κατανομές)	90
Εικόνα 63 Εξερεύνηση πολλών μεταβλητών με pair plot (συσχέτιση)	91

Εισαγωγή

Παρουσιάζουμε έναν οδηγό μελέτης στατιστικής οπτικοποίησης δεδομένων (statistical data visualization) με τη βιβλιοθήκη Seaborn. Ο οδηγός είναι υλοποιημένος σαν υπολογιστικό σημειωματάριο (computational notebook), στο περιβάλλον Jupyter Notebook. Ένα υπολογιστικό σημειωματάριο είναι ένα έγγραφο όπου μπορούμε να συνδυάζουμε κομμάτια κώδικα, με τα οποία διενεργούμε υπολογισμούς, και κείμενο, όπου καταγράφουμε συνοδευτική τεκμηρίωση. Συχνά, ένα υπολογιστικό σημειωματάριο το αποκαλούμε απλά σημειωματάριο (notebook). Στο σημειωματάριο του οδηγού μελέτης οι χρήστες έχουν τη δυνατότητα να δοκιμάζουν σενάρια οπτικοποίησης δεδομένων με διαδραστικό τρόπο, παραλλάσσοντας και εκτελώντας το διαθέσιμο κώδικα, ερμηνεύοντας τα αποτελέσματα και επαναλαμβάνοντας τη διαδικασία, ενώ παράλληλα συμβουλευονται το σχετικό, επεξηγηματικό κείμενο.

Οι χρήστες που επιλέγουν να μελετήσουν τον οδηγό στον Η/Υ τους, πρέπει να έχουν τοπική εγκατάσταση του περιβάλλοντος Jupyter Notebook (ή ενός πιο εξεζητημένου περιβάλλοντος ανάπτυξης σημειωματαρίων και σχετικού λογισμικού, του JupyterLab). Εναλλακτικά, ένας χρήστης μπορεί να επιλέξει να μελετήσει τον οδηγό μελέτης στο διαδίκτυο, σε συνεργατικά περιβάλλοντα νέφους όπως το Google Collab (που είναι, συνήθως, παραλλαγές του JupyterLab). Ένα συνεργατικό περιβάλλον στο διαδίκτυο προσφέρει στους χρήστες το πλεονέκτημα του διαμοιρασμού της εργασίας και των ευρημάτων τους είτε με τροποποιήσεις στο σημειωματάριο μελέτης της οπτικοποίησης δεδομένων είτε με επεκτάσεις σε δικό τους σημειωματάριο. Ανώτερος στόχος της παρούσας εργασίας είναι να συνεισφέρει στην έρευνα στην οπτικοποίηση δεδομένων, που είναι κλάδος της επιστήμης δεδομένων, (VanderPlas, J. (2016). Υπηρετώντας το στόχο της ανοικτής

Οπτικοποίηση δεδομένων με τη βιβλιοθήκη Seaborn
επιστήμης, σκοπεύουμε να ανεβάσουμε τα παραδοτέα της εργασίας σε αποθετήριο στο GitHub.

Η βιβλιοθήκη Seaborn είναι ένα εργαλείο σχεδίασης στατιστικών οπτικοποιήσεων από διαθέσιμα δεδομένα (statistical data visualization). Οι οπτικοποιήσεις που παίρνουμε με τη βιβλιοθήκη Seaborn είναι οπτικοποιήσεις δεδομένων (data visualizations), γιατί τις δημιουργούμε με βάση διαθέσιμα (παρατηρησιακά ή πειραματικά) δεδομένα. Είναι, επίσης, στατιστικές οπτικοποιήσεις (statistical visualizations), γιατί έχουν τη δυνατότητα (και αυτό είναι η δύναμή τους) να εμφανίζουν, στα διαγράμματα που περιέχουν, συγκεντρωτική πληροφορία για τις μεταβλητές που καταγράφονται στα δεδομένα με τη μορφή κατανομών, συσχετίσεων, στατιστικών μεγεθών.

Στην παρούσα εργασία χρησιμοποιούμε τη λέξη οπτικοποίηση ή απεικόνιση εναλλακτικά, για να αναφερθούμε σε στατιστικές οπτικοποιήσεις δεδομένων. Οι λέξεις οπτικοποίηση και απεικόνιση είναι, λοιπόν, ταυτόσημες και αναφέρονται σε εικόνες που περιέχουν ένα ή περισσότερα διαγράμματα με στατιστική πληροφορία για τυχαίες μεταβλητές. Οι λέξεις διάγραμμα και γράφημα έχουν την ίδια σημασία (είναι ταυτόσημες). Τυχαίες μεταβλητές είναι μεταβλητές που η τιμή που ενδέχεται να πάρουν κάθε φορά δε μπορεί να προσδιοριστεί με ακρίβεια αλλά έχει ένα περιθώριο σφάλματος (αβεβαιότητα). Αν δεν αναφέρεται διαφορετικά, η λέξη μεταβλητή σημαίνει τυχαία μεταβλητή. Όταν αναλύουμε και απεικονίζουμε παρατηρησιακά δεδομένα τότε αναφερόμαστε σε παρατηρήσεις, ενώ όταν πρόκειται για πειραματικά δεδομένα τότε αναφερόμαστε σε μετρήσεις. Στην εργασία, όμως, θεωρούμε τις δύο λέξεις, παρατήρηση και μέτρηση, ταυτόσημες.

Συχνά, εμφανίζουμε στις απεικονίσεις που σχεδιάζουμε με τη βιβλιοθήκη Seaborn τις πρωτογενείς παρατηρήσεις για να πάρουμε μία πρώτη αίσθηση της διασποράς των

Οπτικοποίηση δεδομένων με τη βιβλιοθήκη Seaborn
δεδομένων και των μοτίβων (patterns) που ενδεχόμενα σχηματίζουν. Η ουσιαστική, όμως, συνεισφορά της βιβλιοθήκης Seaborn είναι το ευρύ φάσμα δυνατοτήτων που προσφέρει για εξαγωγή και εμφάνιση στατιστικής (συγκεντρωτικής) πληροφορίας για τα δεδομένα, που εξηγεί/ερμηνεύει τη διασπορά των παρατηρήσεων. Για την απεικόνιση της πληροφορίας (πρωτογενούς ή στατιστικής) η βιβλιοθήκη διαθέτει γκάμα εξειδικευμένων γραφικών, όπως σημεία (dots), γραμμές (lines), μπάρες (bars), διαστήματα (ranges), κά.

Μία οπτικοποίηση της Seaborn μπορεί να περιέχει διαγράμματα (ένα ή περισσότερα) διαφόρων τύπων, όπως διαγράμματα διασποράς, διαγράμματα συσχέτισης, διαγράμματα εξέλιξης, κατηγορικά διαγράμματα και διαγράμματα κατανομών μίας ή ζεύγους μεταβλητών. Η βιβλιοθήκη Seaborn διαθέτει γκάμα δυνατοτήτων σχεδίασης διαγραμμάτων και οπτικοποιήσεων, που υλοποιούνται είτε με τις συναρτήσεις σχεδίασης της βιβλιοθήκης ή με τα αντικείμενα σχεδίασης της βιβλιοθήκης. Οι δύο προσεγγίσεις μπορούν να χρησιμοποιηθούν συμπληρωματικά για επίτευξη μέγιστου αποτελέσματος. Για την υλοποίηση του σημειωματάριου μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn έχουν επιλεγεί οι συναρτήσεις σχεδίασης της βιβλιοθήκης, ως ωριμότερη, πιο ολοκληρωμένη και πιο διαισθητική λειτουργικότητα.

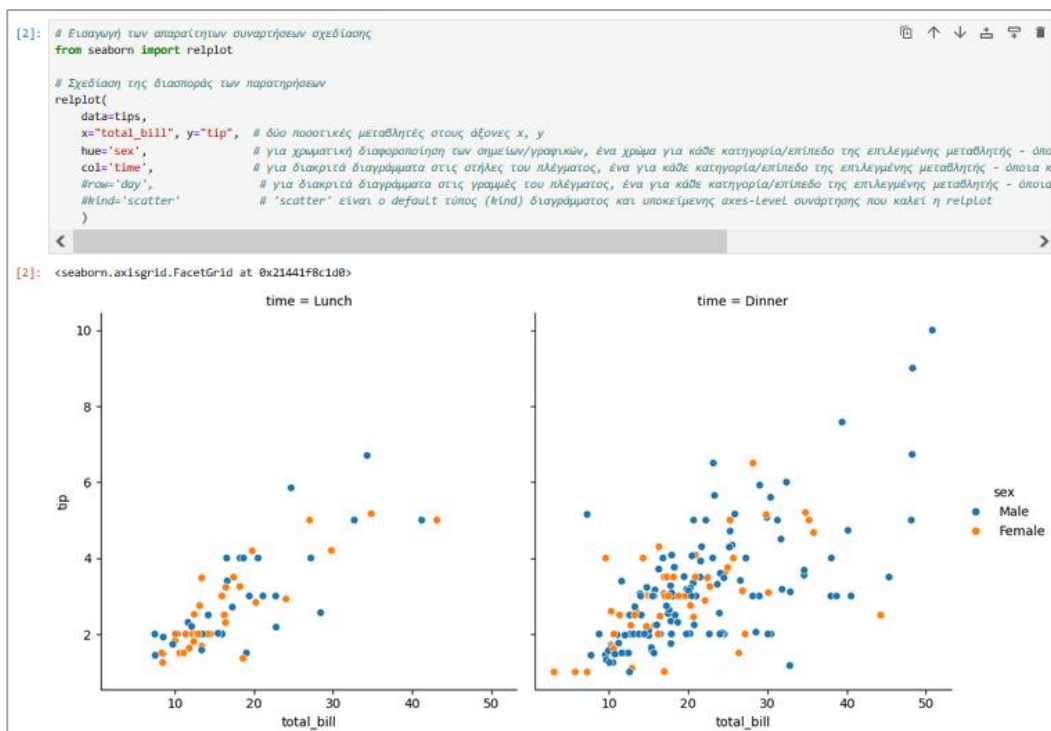
Λέξεις κλειδιά: οδηγός μελέτης, σημειωματάριο μελέτης, υπολογιστικό σημειωματάριο ή απλά σημειωματάριο, διαδραστική χρήση, περιβάλλον Jupyter Notebook/JupyterLab, συνεργατικό περιβάλλον, τοπικό/διαδικτυακό περιβάλλον, GitHub, αποθετήριο, διαμοιρασμός (εργασίας, ιδεών), ανοικτή επιστήμη, επιστήμη δεδομένων, οπτικοποίηση δεδομένων, στατιστική οπτικοποίηση, οπτικοποίηση ή απεικόνιση, διάγραμμα ή γράφημα, παρατήρηση ή μέτρηση, μεταβλητή, τυχαία μεταβλητή, διασπορά παρατηρήσεων,

Οπτικοποίηση δεδομένων με τη βιβλιοθήκη Seaborn
στατιστικό ή στατιστικό μέγεθος, κατανομή, συσχέτιση ή συμμεταβολή,
στατιστική/συγκεντρωτική πληροφορία, συναρτήσεις σχεδίασης, αντικείμενα σχεδίασης,
γραφικά.

A. Jupyter Notebook

Ο οδηγός μελέτης οπτικοποίησης δεδομένων με τη βιβλιοθήκη Seaborn που παρουσιάζεται στην παρούσα εργασία έχει υλοποιηθεί σαν υπολογιστικό σημειωματάριο (ή απλά σημειωματάριο ή notebook). Ένα σημειωματάριο (όπως ο οδηγός μελέτης) μπορεί να αναγνωστεί και να εκτελεστεί στο περιβάλλον Jupyter Notebook. Ο χρήστης, αφού ανοίξει ένα σημειωματάριο μέσα στο περιβάλλον Jupyter Notebook, έχει τη δυνατότητα να εκτελεί τα κομμάτια κώδικα (scripts) που είναι διαθέσιμα στα υπολογιστικά κελιά (είτε αυτούσια είτε τροποποιώντας τα) ενώ παράλληλα συμβουλευείται το συνοδευτικό κείμενο στα αντίστοιχα κελιά κειμένου.

Στην ακόλουθη εικόνα (Εικ.1) φαίνεται (με γκρι χρώμα) ένα εκτελεσμένο υπολογιστικό κελί του οδηγού μελέτης και η οπτικοποίηση που έχει προκύψει από την εκτέλεσή του.



Εικόνα 1 Εκτελεσμένο υπολογιστικό κελί

Σε κάθε υπολογιστικό κελί του οδηγού μελέτης ο χρήστης μπορεί να δοκιμάζει τροποποιήσεις στον κώδικα και να αξιολογεί τα αποτελέσματα που προκύπτουν. Μπορεί να επαναλαμβάνει αυτή τη διαδικασία δοκιμών όσες φορές το επιθυμεί. Χάρη στο Jupyter Notebook, επομένως, ο χρήστης διατηρεί διαδραστική επικοινωνία με το σημειωματάριο μελέτης οργανώνοντας την εργασία του σε ακολουθία βημάτων (σενάριο) της αρεσκείας του.

Το Jupyter Notebook είναι διαδικτυακή εφαρμογή, που λειτουργεί μέσω προγράμματος περιήγησης (browser). Το Jupyter Notebook προσφέρει το διαδραστικό περιβάλλον για την επεξεργασία και χρήση σημειωμάτων με τα οποία πετυχαίνουμε ανάπτυξη, τεκμηρίωση και εκτέλεση κώδικα, εξερεύνηση και οπτικοποίηση δεδομένων, επικοινωνία των αποτελεσμάτων, διαμοιρασμό εργασίας και ιδεών.

Στα σημειωμάτια (notebooks) καταγράφεται ολόκληρη η διαδικασία υπολογισμού: ανάπτυξη, τεκμηρίωση, εκτέλεση κώδικα και επικοινωνία των αποτελεσμάτων. Πρόκειται για έγγραφα που συνδυάζουν κώδικα, που μπορεί να χρησιμοποιήσει το ευρύ φάσμα των εγκατεστημένων βιβλιοθηκών λογισμικού (της εγκατεστημένης λειτουργικότητας) και τα (τοπικά ή στο διαδίκτυο) διαθέσιμα δεδομένα και κείμενο που μπορεί να περιέχει τεκμηρίωση, επεξηγήσεις, οδηγίες, σκοπούς, συμπεράσματα, θεωρία, μαθηματικές εξισώσεις, σχήματα, εικόνες, πίνακες, υπερσυνδέσμους, κ.ά. Ο υπολογιστικός χαρακτήρας των σημειωμάτων είναι αυτό που τους έδωσε τον χαρακτηρισμό “υπολογιστικό” (υπολογιστικό σημειωματάριο).

Στον οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn, για παράδειγμα, ο χρήστης μπορεί να εκτελέσει ένα κελί κώδικα που καλεί μία συνάρτηση σχεδίασης της βιβλιοθήκης και εμφανίζει σχετική οπτικοποίηση και, αφού αξιολογήσει το αποτέλεσμα, να κάνει στον κώδικα τις αλλαγές που θα του επιτρέψουν να διερευνήσει

εναλλακτικές περιπτώσεις σχεδίασης και να τον εκτελέσει ξανά. Αυτή τη διαδικασία δοκιμών μπορεί να την επαναλάβει αρκετές φορές μέχρι να πάρει το αποτέλεσμα που επιθυμεί. Την εργασία τους και τα αποτελέσματα οι χρήστες μπορούν να τα κοινοποιούν και να τα επεξεργάζονται από κοινού μέσα από συνεργατικά περιβάλλοντα (όπως το Google Collab), (Google, 2025) ή μέσα από αποθετήρια (όπως το GitHub).

Το περιβάλλον του Jupyter Notebook παρέχει, λοιπόν, τρία πλεονεκτήματα:

- **Διάδραση:** Διαδραστικό περιβάλλον ανάπτυξης και χρήσης (υπολογιστικών) σημειωματάρων, όπου ο χρήστης έχει τη δυνατότητα να δοκιμάσει κομμάτια κώδικα (σε διάφορες γλώσσες προγραμματισμού) γρήγορα και αλληλεπιδραστικά, αξιολογώντας κάθε φορά τα αποτελέσματα της εκτέλεσης των υπολογισμών και κάνοντας τις διορθώσεις που επιθυμεί, ενώ παράλληλα προσαρμόζει το συνοδευτικό κείμενο με τη σχετική τεκμηρίωση.
- **Δυναμικοί υπολογισμοί:** Ο κώδικας στα υπολογιστικά κελιά του σημειωματάρου συντάσσεται, εκτελείται, διορθώνεται και επανεκτελείται σε πραγματικό χρόνο, αξιοποιώντας τη λειτουργικότητα των εγκατεστημένων, διαθέσιμων βιβλιοθηκών (για παράδειγμα, εγκατεστημένων python packages). Μπορούν να δοκιμαστούν συναρτήσεις που εκτελούν εξειδικευμένους υπολογισμούς, σχεδιάζουν πλούσιες απεικονίσεις, εκπαιδεύουν μοντέλα, κá. Τα κελιά κώδικα συμπληρώνονται με κελιά κειμένου, όπου σε απλή γλώσσα Markdown (ή HTML σε εξεζητημένες περιπτώσεις) καταγράφεται σχετική τεκμηρίωση εμπλουτισμένη με συνδέσμους, εικόνες, μαθηματικές εξισώσεις, πίνακες, κá.
- **Διαμοιρασμός:** Τα (υπολογιστικά) σημειωματάρια μπορούν να διαμοιραστούν, ώστε να τύχουν κοινής επεξεργασίας, ενισχύοντας έτσι τη συνεργασία ερευνητών και

μηχανικών. Ένα ευρέως διαδεδομένο συνεργατικό περιβάλλον που χρησιμοποιεί το περιβάλλον του Jupyter Notebook είναι το [Google Collab](#).

Με απλά λόγια, ένα jupyter-notebook είναι ένα αλληλεπιδραστικό έγγραφο που συνδυάζει κείμενο και κομμάτια κώδικα (scripts) σε αντίστοιχα κελιά κειμένου (Markdown) και υπολογιστικά κελιά (Code).

Στην παρούσα εργασία επιλέξαμε να παρουσιάσουμε τις δυνατότητες οπτικοποίησης δεδομένων της βιβλιοθήκης seaborn με ένα jupyter-notebook (υπολογιστικό σημειωματάριο που έχει συνταχθεί στο περιβάλλον του Jupyter Notebook) γιατί εξασφαλίζει τη μελέτη και εκμάθηση της σχεδίασης στατιστικών οπτικοποιήσεων δεδομένων με τα πλεονεκτήματα που προσφέρει το διαδραστικό περιβάλλον του Jupyter Notebook και ο υπολογιστικός χαρακτήρας των (υπολογιστικών) σημειωμάτων.

Μελετώντας ο χρήστης το συνημμένο jupyter-notebook σημειωματάριο μπορεί να κάνει δύο παράλληλα πράγματα:

- Να εκτελεί τα ενσωματωμένα κομμάτια κώδικα (scripts), τα οποία καλούν τις υπό μελέτη δυνατότητες σχεδίασης οπτικοποιήσεων (συναρτήσεις σχεδίασης) και να εμφανίζει τις σχετικές οπτικοποιήσεις. Ο χρήστης έχει, επιπλέον, τη δυνατότητα να παρεμβαίνει τροποποιώντας τον κώδικα, ώστε να διερευνά εναλλακτικές περιπτώσεις σχεδίασης της εκάστοτε απεικόνισης.
- Να διαβάζει, παράλληλα, κείμενο με θεωρία, επεξηγήσεις, οδηγίες, σκοπούς και ερμηνεία σχετικά με τον κώδικα και την παραγόμενη οπτικοποίηση.

Τα notebook ως έγγραφα αποθηκεύονται με την .ipynb επέκταση.

Τα σημειωματάρια μπορούν να εξαχθούν σε μια σειρά στατικών μορφών, όπως HTML (για παράδειγμα, για αναρτήσεις ιστολογίου), reStructuredText, LaTeX, PDF και παρουσιάσεις διαφανειών, μέσω της εντολής [nbconvert].

Απαραίτητες βιβλιοθήκες

Για να αξιοποιήσουμε τον οδηγό μελέτης της οπτικοποίησης δεδομένων με τη βιβλιοθήκη Seaborn απαιτούνται κάποιες βιβλιοθήκες λογισμικού (python packages). Επομένως, αν επιθυμούμε να δουλέψουμε σε τοπικό περιβάλλον (στον Η/Υ μας), τότε πρέπει να έχει προηγηθεί τοπική εγκατάσταση των απαραίτητων βιβλιοθηκών. Αυτές είναι οι βιβλιοθήκες που εξασφαλίζουν το περιβάλλον Jupyter Notebook, η βιβλιοθήκη οπτικοποιήσεων Seaborn και κάποιες ακόμη βιβλιοθήκες από τις οποίες εξαρτάται η λειτουργία του Jupyter Notebook και της Seaborn.

Στο περιβάλλον Jupyter Notebook διαβάζουμε, εκτελούμε κώδικα και γενικά αλληλεπιδρούμε με σημειωματάρια, επομένως και με τον οδηγό μελέτης οπτικοποίησης δεδομένων με τη βιβλιοθήκη Seaborn, που είναι, επίσης, σημειωματάριο. Τη βιβλιοθήκη Seaborn τη χρησιμοποιούμε στον κώδικα του οδηγού μελέτης για να σχεδιάσουμε στατιστικές οπτικοποιήσεις δεδομένων.

Σε τοπικό επίπεδο προτείνεται εγκατάσταση των βιβλιοθηκών σε ένα εικονικό περιβάλλον python (python virtual environment).

Είσοδος στο περιβάλλον Jupyter Notebook

Εκκινούμε το περιβάλλον Jupyter Notebook από τη γραμμή εντολών του λειτουργικού συστήματος. Για παράδειγμα, στα Windows εκκινούμε το Jupyter Notebook μέσα από το PowerShell.

Στη γραμμή εντολών κάνουμε δύο ενέργειες:

1. Αλλάζουμε τον κατάλογο εργασίας στον κατάλογο όπου βρίσκονται τα σημειωματάρια (notebooks) με τα οποία θέλουμε να εργαστούμε. Για να το κάνουμε αυτό χρησιμοποιούμε την εντολή `cd`.

Για παράδειγμα, για να ορίσουμε ως κατάλογο εργασίας τον κατάλογο που βρίσκεται στη διαδρομή `C:\Dev\visualization\learnSeaborn` πληκτρολογούμε:

```
cd C:\Dev\visualization\learnSeaborn
```

2. Εκκινούμε το περιβάλλον Jupyter Notebook με την εντολή `"jupyter notebook"` (αν δεν έχουμε ενεργοποιήσει το εικονικό περιβάλλον όπου είναι εγκατεστημένες οι βιβλιοθήκες και το εκτελέσιμο `jupyter.exe`, περιλαμβάνουμε τη διαδρομή προς το εκτελέσιμο, για παράδειγμα, αν το εικονικό περιβάλλον έχει όνομα `'venv'` και βρίσκεται μέσα στον κατάλογο εργασίας, τότε χρησιμοποιούμε τη σχετική διαδρομή `.\venv\Scripts\jupyter.exe` ή απλά `.\venv\Scripts\jupyter`)

```
.\venv\Scripts\jupyter notebook
```

Η εκτέλεση των δύο παραπάνω ενεργειών φαίνεται στο παρακάτω snapshot:


```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

PS C:\Users\user> cd C:\Dev\visualization\learnSeaborn
PS C:\Dev\visualization\learnSeaborn> .\venv\Scripts\jupyter.exe notebook
[I 2025-05-06 19:41:00.091 ServerApp] jupyter_lsp | extension was successfully linked.
[I 2025-05-06 19:41:00.107 ServerApp] jupyter_server_terminals | extension was successfully linked.
[I 2025-05-06 19:41:00.122 ServerApp] jupyterlab | extension was successfully linked.
[I 2025-05-06 19:41:00.122 ServerApp] notebook | extension was successfully linked.
[I 2025-05-06 19:41:00.724 ServerApp] notebook_shim | extension was successfully linked.
[I 2025-05-06 19:41:00.780 ServerApp] notebook_shim | extension was successfully loaded.
[I 2025-05-06 19:41:00.793 ServerApp] jupyter_lsp | extension was successfully loaded.
[I 2025-05-06 19:41:00.793 ServerApp] jupyter_server_terminals | extension was successfully loaded.
[I 2025-05-06 19:41:00.793 LabApp] JupyterLab extension loaded from C:\Dev\visualization\learnSeaborn\venv\Lib\site-packages\jupyterlab
[I 2025-05-06 19:41:00.793 LabApp] JupyterLab application directory is C:\Dev\visualization\learnSeaborn\venv\share\jupyter\lab
[I 2025-05-06 19:41:00.793 LabApp] Extension Manager is 'pypi'.
[I 2025-05-06 19:41:01.349 ServerApp] jupyterlab | extension was successfully loaded.
[I 2025-05-06 19:41:01.357 ServerApp] notebook | extension was successfully loaded.
[I 2025-05-06 19:41:01.357 ServerApp] Serving notebooks from local directory: C:\Dev\visualization\learnSeaborn
[I 2025-05-06 19:41:01.357 ServerApp] Jupyter Server 2.14.2 is running at:
[I 2025-05-06 19:41:01.357 ServerApp] http://localhost:8888/tree?token=19f9e64ad88d8a3c3955e4226950de4d1ac1750b5734084a
[I 2025-05-06 19:41:01.357 ServerApp] http://127.0.0.1:8888/tree?token=19f9e64ad88d8a3c3955e4226950de4d1ac1750b5734084a
[I 2025-05-06 19:41:01.357 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).

To access the server, open this file in a browser:
file:///C:/Users/user/AppData/Roaming/jupyter/runtime/jpserver-7160-open.html
Or copy and paste one of these URLs:
http://localhost:8888/tree?token=19f9e64ad88d8a3c3955e4226950de4d1ac1750b5734084a
http://127.0.0.1:8888/tree?token=19f9e64ad88d8a3c3955e4226950de4d1ac1750b5734084a
[I 2025-05-06 19:41:01.530 ServerApp] Skipped non-installed server(s): bash-language-server, dockerfile-language-server-nodejs, javascript-type
language-server, pyright, python-language-server, python-lsp-server, r-languageserver, sql-language-server, texlab, typescript-language-server, u
-bins, vscode-html-languageserver-bin, vscode-json-languageserver-bin, yamll-language-server
```

Εικόνα 2 Γραμμή εντολών

Για να μην εκτελούμε τις παραπάνω εντολές (Εικ. 2) κάθε φορά που θέλουμε να ξεκινήσουμε το Jupyter Notebook έχουμε δημιουργήσει ένα batch αρχείο εκκίνησης του Jupyter Notebook (Εικ. 3). Εκκινούμε το batch αρχείο, όπως κάθε εκτελέσιμο αρχείο με διπλό κλικ πάνω σε αυτό. Οι εντολές που περιλαμβάνει το batch αρχείο είναι οι παρακάτω:

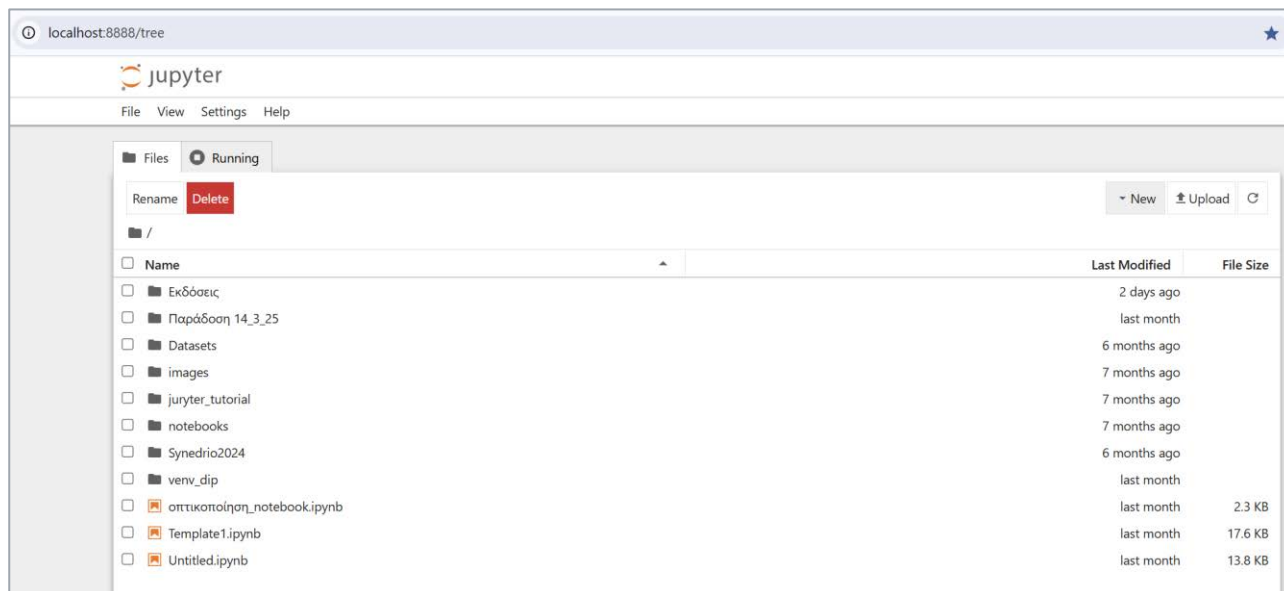
```
1 @echo off
2 echo Starting Jupyter ...
3 cd C:\Dev\visualization\learnSeaborn
4 .\venv\Scripts\jupyter notebook
5
6 echo Press enter to close Jupyter (if open) and deactivate the environment (if active).
7 pause > nul
8 taskkill /IM ".\venv\Scripts\jupyter-notebook.exe" /F
9
```

Εικόνα 3 Εκτελέσιμο αρχείο

Εργασία στο περιβάλλον Jupyter Notebook

Όταν ξεκινάμε το Jupyter Notebook στον Η/Υ μας η εφαρμογή ανοίγει σε μία νέα καρτέλα, στο default πρόγραμμα περιήγησης (browser) με διεύθυνση url: <https://localhost:8888/tree>. Αρχικά εμφανίζεται η λίστα των καταλόγων και αρχείων που

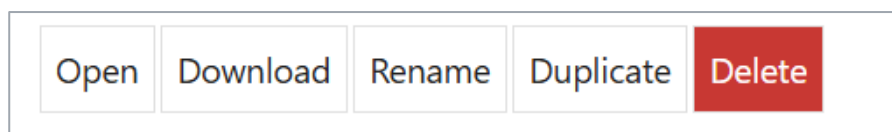
βρίσκεται μέσα στον κατάλογο εργασίας που ορίσαμε κατά την εκκίνηση του Jupyter Notebook (Εικ. 4).



Εικόνα 4 Εκκίνηση Jupyter Notebook

Ανοίγουμε έναν από τους διαθέσιμους καταλόγους στην παραπάνω λίστα κάνοντας διπλό κλικ πάνω σε αυτόν. Ανοίγουμε ένα σημειωματάριο που εντοπίζουμε μέσα σε ένα κατάλογο με διπλό κλικ πάνω σε αυτό ή με επιλογή του και πάτημα (από τα κουμπιά που εμφανίζονται) του πλήκτρου Open. Με αυτόν τον τρόπο ανοίγουμε και τον οδηγό μελέτης της βιβλιοθήκης οπτικοποιήσεων δεδομένων Seaborn, που, όπως είπαμε, είναι ένα σημειωματάριο.

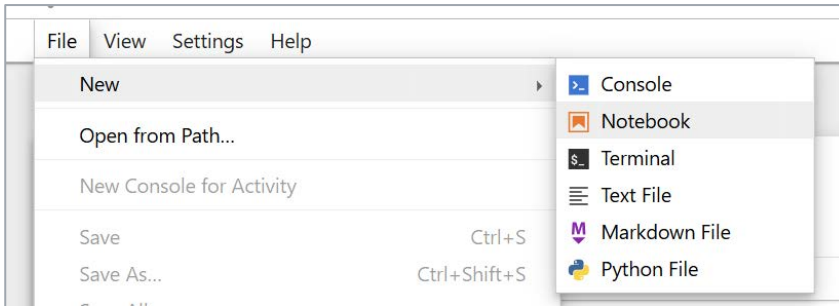
Όταν επιλέγουμε ένα σημειωματάριο (για παράδειγμα τον οδηγό μελέτης, 'Data Visualization with Seaborn.ipynb'), τότε εκτός από το πλήκτρο Open, εμφανίζονται και άλλα πλήκτρα, όπως το Download και το Rename, που η λειτουργία τους είναι αυτονόητη. Ιδιαίτερη προσοχή με το πλήκτρο Delete (Εικ. 5).



Εικόνα 5 Πλήκτρα επιλογών

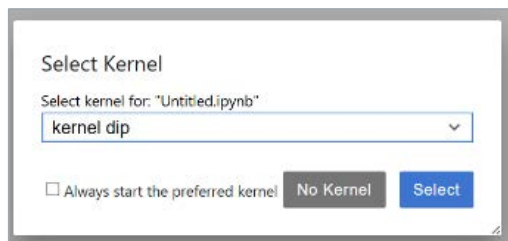
Δημιουργία σημειωματρίου

Δημιουργούμε νέο σημειωματάριο κάνοντας κλικ στο αναπτυσσόμενο κουμπί Νέο στο επάνω μέρος της λίστας (Εικ. 6). Ένα σημειωματάριο είναι ένα αρχείο με προέκταση `.ipynb`.



Εικόνα 6 Δημιουργία νέου σημειωματρίου

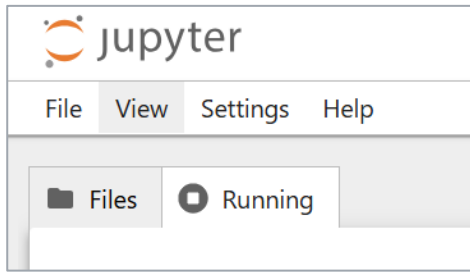
Κατά τη δημιουργία ενός νέου σημειωματρίου μας ζητείται να ορίσουμε πυρήνα (kernel). Επιλέγουμε έναν από τους διαθέσιμους πυρήνες (Εικ. 7). Μπορούμε, ανά πάσα στιγμή να αλλάξουμε τον πυρήνα του σημειωματρίου, κατά την επεξεργασία του, όπως θα δούμε στη σχετική ενότητα.



Εικόνα 7 Επιλογή Kernel

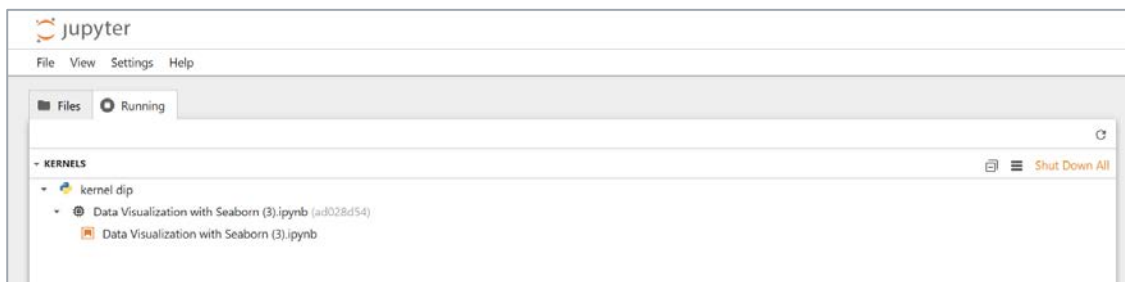
Μπορούμε να δώσουμε στο σημειωματάριο ένα αντιπροσωπευτικό όνομα επιλέγοντας το και πατώντας μετά το πλήκτρο Rename. Ανά πάσα στιγμή μπορούμε να μετονομάσουμε ένα σημειωματάριο κατά την επεξεργασία του, όπως θα δούμε στη σχετική ενότητα.

Η καρτέλα Running



Εικόνα 8 Καρτέλα Running

Η καρτέλα Running εμφανίζει (Εικ. 8) τα τρέχοντα notebooks και μας βοηθά να παρακολουθούμε σε ποια notebooks εργαζόμαστε. Το κλείσιμο της σελίδας του notebook δεν επαρκεί για τον τερματισμό ενός kernel. Για τον τερματισμό του kernel που συσχετίζεται με το notebook που εργαζόμαστε χρησιμοποιούμε το πορτοκαλί κουμπί τερματισμού λειτουργίας Shut Down All (Εικ. 9).



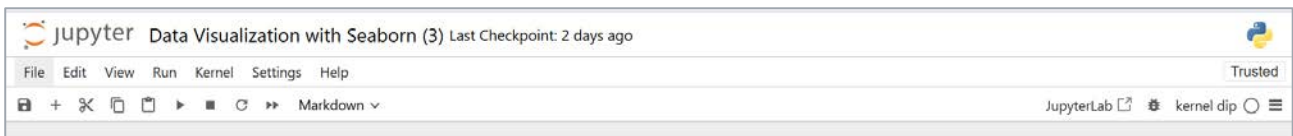
Εικόνα 9 Κουμπί τερματισμού λειτουργίας

Κάθε ανοικτό σημειωματάριο φιλοξενείται σε μία σελίδα σε ξεχωριστή καρτέλα του φυλλομετρητή (browser). Η σελίδα αυτή, όπου επεξεργαζόμαστε ένα σημειωματάριο, έχει μενού και εργαλεία με τα οποία κάνουμε την επεξεργασία του σημειωματάριου.

Κεφαλίδα

Στο επάνω μέρος του εγγράφου του σημειωματαρίου υπάρχει μια κεφαλίδα που περιέχει τον τίτλο του σημειωματαρίου, μια γραμμή μενού και μια γραμμή εργαλείων (Εικ.

10). Αυτή η κεφαλίδα παραμένει σταθερή στο επάνω μέρος της οθόνης, ακόμα και όταν γίνεται κύλιση στο κυρίως μέρος του σημειωματαρίου. Ο τίτλος μπορεί να υποστεί επεξεργασία επί τόπου (γεγονός που μετονομάζει το αρχείο του σημειωματαρίου) και η γραμμή μενού και η γραμμή εργαλείων περιέχουν μια ποικιλία ενεργειών που ελέγχουν την πλοήγηση στο σημειωματάριο και τη δόμηση του εγγράφου.



Εικόνα 10 Κεφαλίδα

Κύριο Μέρος

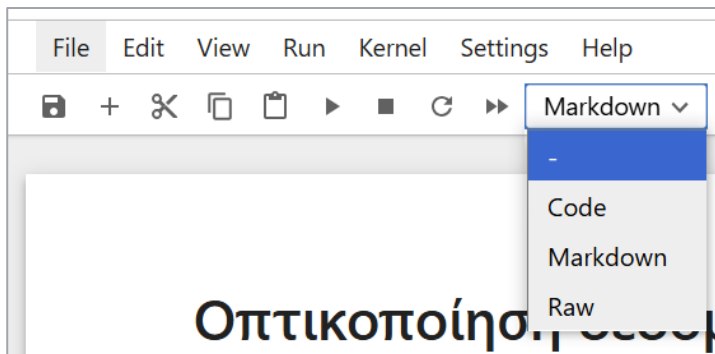
Το κύριο μέρος ενός σημειωματαρίου αποτελείται από κελιά. Κάθε κελί περιέχει είτε markdown, εισαγωγή κώδικα, έξοδο κώδικα ή ακατέργαστο κείμενο. Τα κελιά μπορούν να συμπεριληφθούν με οποιαδήποτε σειρά και να υποστούν επεξεργασία κατά βούληση, επιτρέποντας μεγάλη ευελιξία για τη δημιουργία μιας αφήγησης.

Κελιά Markdown - Αυτά χρησιμοποιούνται για τη δημιουργία μιας όμορφα μορφοποιημένης αφήγησης γύρω από τον κώδικα στο έγγραφο. Το μεγαλύτερο μέρος αυτού του μαθήματος αποτελείται από κελιά markdown (Εικ. 11).

Υπολογιστικά κελιά (code) - Αυτά χρησιμοποιούνται για τον ορισμό του υπολογιστικού κώδικα στο έγγραφο. Διατίθενται σε δύο μορφές: το κελί εισόδου όπου ο χρήστης πληκτρολογεί τον κώδικα που θα εκτελεστεί και το κελί εξόδου που είναι η αναπαράσταση του εκτελεσμένου κώδικα. Ανάλογα με τον κώδικα, αυτή η αναπαράσταση μπορεί να είναι

μία απλή βαθμωτή τιμή ή κάτι πιο περίπλοκο, όπως ένα γράφημα ή ένα διαδραστικό γραφικό στοιχείο (Εικ. 11).

Ακατέργαστα κελιά (row)- Αυτά χρησιμοποιούνται όταν το κείμενο πρέπει να συμπεριληφθεί σε ακατέργαστη μορφή, χωρίς εκτέλεση ή μετασχηματισμό (Εικ. 11).



Εικόνα 11 Κελιά

Συγγραφή σημειωματρίου

Ο οδηγός μελέτης στατιστικής οπτικοποίησης δεδομένων με τη βιβλιοθήκη Seaborn που παρουσιάζεται με την παρούσα εργασία έχει υλοποιηθεί σαν σημειωματάριο (notebook), το οποίο μπορεί να αναγνωστεί και να εκτελεστεί στο περιβάλλον Jupyter Notebook. Ο χρήστης, αφού ανοίξει το σημειωματάριο στο περιβάλλον, μπορεί να εκτελεί τα κομμάτια κώδικα (scripts) που είναι διαθέσιμα στα σχετικά κελιά του σημειωματαρίου (είτε αυτούσια είτε παραλλάσσοντάς τα) ενώ παράλληλα διαβάζει τα κομμάτια κειμένου στα αντίστοιχα κελιά. Μπορεί να εκτελέσει όποια ακολουθία σεναρίων οπτικοποίησης επιθυμεί, παραλλάσσοντας τον κώδικα και να αξιολογήσει τα αποτελέσματα.

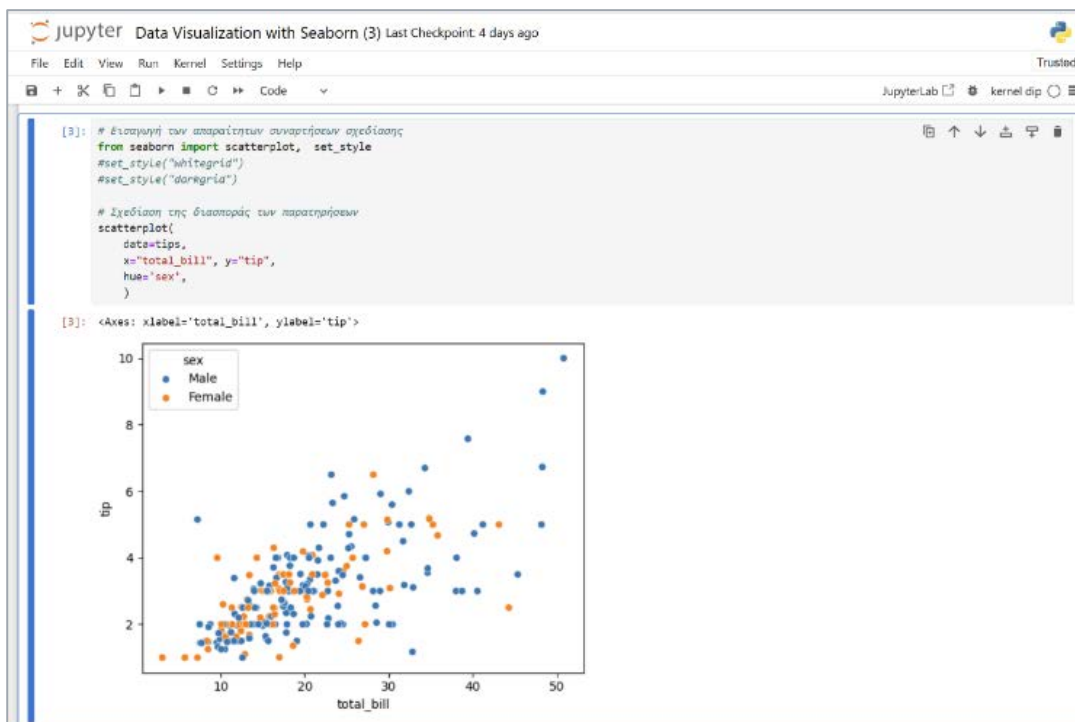
Για παράδειγμα εάν προσπαθήσουμε να τρέξουμε τον κώδικα της Εικ. 12 στο παρακάτω κελί του οδηγού μελέτης:

```
# Εισαγωγή των απαραίτητων συναρτήσεων σχεδίασης
from seaborn import scatterplot, set_style
#set_style("whitegrid")
#set_style("darkgrid")

# Σχεδίαση της διαφοράς των παρατηρήσεων
scatterplot(
    data=tips,
    x="total_bill", y="tip",
    hue='sex',
)
```

Εικόνα 12 Κώδικας

Η εκτέλεση του κώδικα δίνει το αποτέλεσμα που προκύπτει -παρακάτω (Εικ. 13):



Εικόνα 13 Εκτέλεση κώδικα

Μέσα στο κελί εργαλεία που τα χρησιμοποιούμε για να αλληλοεπιδρούμε με το σημειωματάριο άμεσα και φαίνονται εδώ (Εικ. 14):



Εικόνα 14 Εργαλεία μέσα στο κελί

- Το κουμπί με το οποίο διπλασιάζουμε το καλί που εργαζόμαστε
- Υπάρχουν τα βελάκια που εξυπηρετούν για την μετακίνηση πάνω ή κάτω του τρέχοντος κελιού
- Το Insert κουμπί χρησιμοποιείται για την εισαγωγή νέων κελιών πάνω ή κάτω από το κελί που εργαζόμαστε
- Καθώς και το κουμπί με το οποίο διαγράφουμε το κελί που εργαζόμαστε

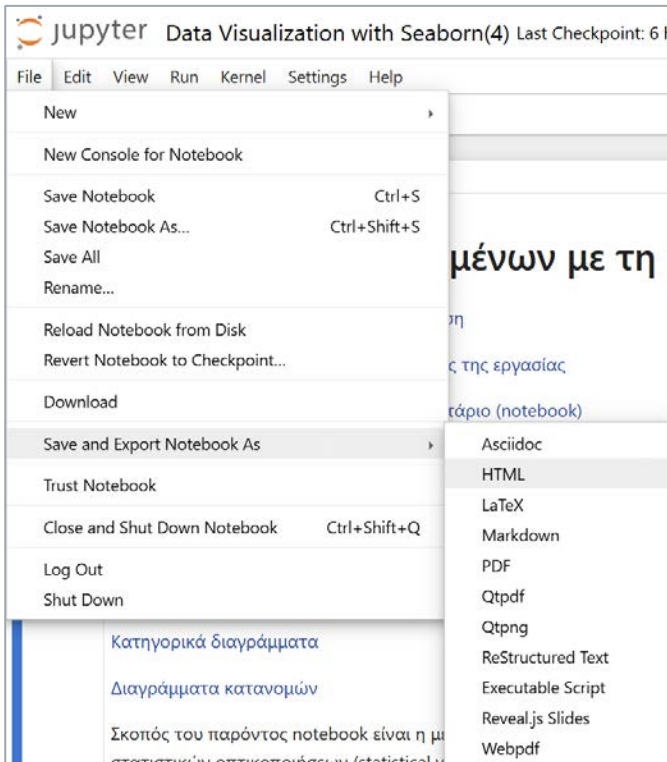
Τα Notebooks αποθηκεύονται αυτόματα αρκετά συχνά, επομένως σπάνια χάνονται δεδομένα. Παρόλα, αυτά όταν χρειάζεται να τερματίσουμε το διακομιστή φροντίζουμε να αποθηκεύουμε τα δεδομένα.

Εμπλουτισμός περιεχομένου στα κελιά

Το Jupyter Notebook υποστηρίζει τη γλώσσα μορφοποίησης κειμένου Markdown. Αυτό υπήρξε ιδιαίτερα χρήσιμο για τον οδηγό μελέτης της Seaborn. Συγκεκριμένα δημιουργήθηκαν επεξηγηματικά κείμενα ανάμεσα στα κελιά κώδικα που με τη σειρά τους πρόσθεσαν τεκμηριωμένη ανάλυση πάνω στις οπτικοποιήσεις οι οποίες αποτέλεσαν το αντικείμενο μελέτης μας. Παράλληλα το notebook εμπλουτίστηκε με τίτλους, επικεφαλίδες

Εξαγωγή notebook

Τα αρχεία notebook (.ipynb) μπορούν να μετατραπούν ή να εξαχθούν σε διάφορες μορφές όπως φαίνεται στην παρακάτω εικόνα (Εικ. 15) προκειμένου να διαμοιρασθούν.



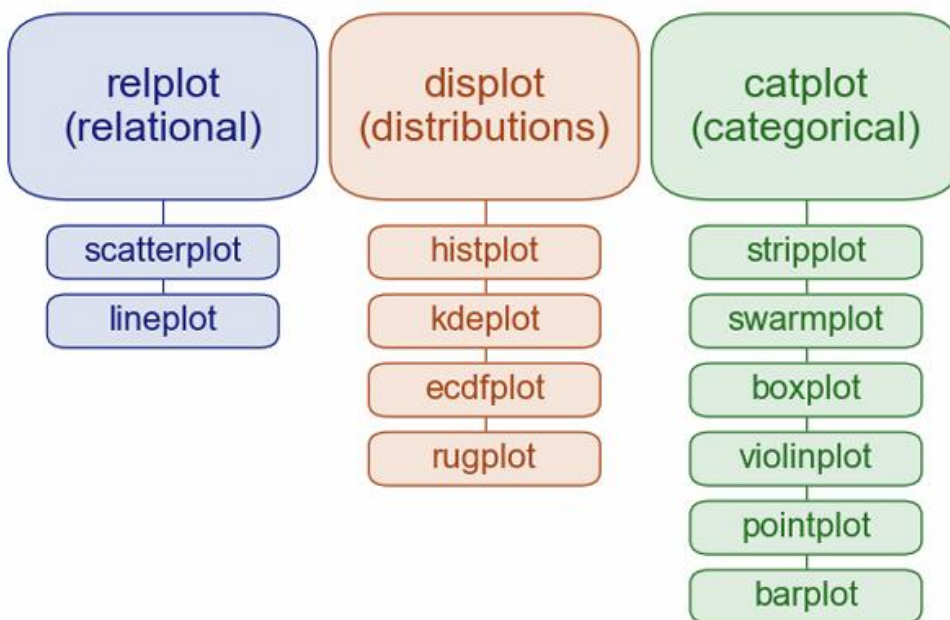
Εικόνα 15 Εξαγωγή αρχείων

Στον οδηγό που έχουμε δημιουργήσει για την εμβάθυνση στις οπτικοποιήσεις με seaborn αρκετά χρήσιμη υπήρξε η εξαγωγή σε html. Αντίθετα, η εξαγωγή του αρχείου με το ελληνικό κείμενο απευθείας σε pdf απαιτούσε περεταίρω εγκαταστάσεις και δεν επιχειρήθηκε.

B. Βιβλιοθήκη οπτικοποίησης δεδομένων Seaborn

Στην ακόλουθη ενότητα παρουσιάζονται με απλό τρόπο βασικές λειτουργίες οπτικοποίησης δεδομένων (data visualization functions) με τη βιβλιοθήκη οπτικοποίησης δεδομένων Seaborn. Τις λειτουργίες οπτικοποίησης δεδομένων τις αποκαλούμε αλλιώς λειτουργίες σχεδίασης ή συναρτήσεις σχεδίασης (plotting functions). Το αποτέλεσμα της εκτέλεσης μίας λειτουργίας σχεδίασης είναι μία οπτικοποίηση (απεικόνιση), με ένα ή περισσότερα διαγράμματα (γραφήματα), (Seaborn development team, 2024).

Στην ακόλουθη εικόνα (Εικ. 16) εμφανίζεται μία κατηγοριοποίηση των συναρτήσεων σχεδίασης της βιβλιοθήκης Seaborn (Seaborn development team, 2024). Η κατηγοριοποίηση, αν και όχι επικαιροποιημένη, περιέχει τις βασικές κατηγορίες των σχέσεων και των κατανομών.



Εικόνα 16 Κατηγοριοποίηση των συναρτήσεων σχεδίασης της βιβλιοθήκης Seaborn

Εις βάθος εφαρμογή και ανάλυση των λειτουργιών οπτικοποίησης δεδομένων της βιβλιοθήκης Seaborn θα βρείτε στον οδηγό μελέτης, "[Data Visualization with Seaborn.ipynb](#)".

Ο οδηγός μελέτης οπτικοποίησης δεδομένων με τη βιβλιοθήκη Seaborn έχει υλοποιηθεί σαν σημειωματάριο (notebook) στο περιβάλλον Jupyter Notebook και σε αυτόν μπορείτε να δοκιμάσετε παραλλαγές λειτουργιών οπτικοποίησης δεδομένων διαβαθμισμένης πολυπλοκότητας. Για να χρησιμοποιήσετε τον οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn θα πρέπει να τον ανοίξετε μέσα από το περιβάλλον Jupyter Notebook.

Οργανώνουμε την οπτικοποίηση δεδομένων σε τρεις μεγάλες κατηγορίες:

- Σχεδίαση και διαγράμματα σχέσεων μεταβλητών (relational plots & models plots),
- Σχεδίαση και διαγράμματα κατανομών μεταβλητών (distribution plots),
- Σχεδίαση και διαγράμματα εξερεύνησης μεταβλητών κατά ζεύγη (pair plot & joint plot).

Η τρίτη κατηγορία λειτουργιών σχεδίασης συνδυάζει διαγράμματα σχέσεων και κατανομών μεταβλητών, που σχεδιάζονται για ζεύγη μεταβλητών, και τα τακτοποιεί όλα σε μία οπτικοποίηση με σκοπό την εξερεύνηση των διαθέσιμων δεδομένων. Στον οδηγό μελέτης έχουμε, επιπλέον, περιλάβει μία τέταρτη κατηγορία σχεδίασης διαγραμμάτων, τα κατηγορικά διαγράμματα, τα οποία ουσιαστικά, εξυπηρετούν άλλοτε τη σχεδίαση σχέσεων και άλλοτε τη σχεδίαση κατανομών μεταβλητών.

Οι οπτικοποιήσεις (απεικονίσεις) που παίρνουμε με τις λειτουργίες σχεδίασης της βιβλιοθήκης Seaborn βασίζονται σε δεδομένα. Με την έννοια αυτή είναι εμπειρικές, αφού προκύπτουν ως εκτιμήσεις της πραγματικότητας από τα διαθέσιμα κάθε φορά δεδομένα.

Για να εισάγουμε τα διαθέσιμα δεδομένα σε μία συνάρτηση σχεδίασης της βιβλιοθήκης θα πρέπει να τα έχουμε σε μορφή πίνακα, όπου κάθε στήλη περιέχει μία υποψήφια μεταβλητή και κάθε γραμμή μία παρατήρηση (μέτρηση). Τον πίνακα του συνόλου δεδομένων που θα χρησιμοποιήσουμε κάθε φορά τον φορτώνουμε σε ένα pandas dataframe (που έχει την απαιτούμενη μορφή πίνακα), (McKinney, W. & pandas development team, 2025). Φορτώνουμε τα δεδομένα από διαδικτυακές πηγές, όπως το αποθετήριο της βιβλιοθήκης Seaborn στο Github ή άλλα αποθετήρια στο διαδίκτυο, (Waskom, M. 2025).

Για παράδειγμα, για να φορτώσουμε ένα σύνολο δεδομένων από το αποθετήριο της βιβλιοθήκης Seaborn χρησιμοποιούμε τη συνάρτηση `load_dataset` της Seaborn (Εικ. 17).

```
from seaborn import load_dataset

# Φόρτωση συνόλου δεδομένων (από το αποθετήριο της seaborn στο διαδίκτυο)
tips = load_dataset("tips")

# Εμφάνιση του συνόλου δεδομένων
print('Στήλες στο σύνολο δεδομένων (dataframe) tips: ', tips.columns.to_list(), '\n')
tips
```

Εικόνα 17 Η συνάρτηση `load_dataset`

Σχέσεις μεταβλητών

Δύο μεταβλητές σχετίζονται (ή συσχετίζονται) όταν μεταβολή στη μία μεταβλητή συνεπάγεται μεταβολή στην άλλη. Δηλαδή, όταν μεταβάλλεται η μία μεταβλητή κατά ένα ποσό, τότε ακολουθεί και η άλλη και μεταβάλλεται και αυτή κατά κάποιο σχετιζόμενο ποσό. Συχνά αναφερόμαστε στη συσχέτιση δύο μεταβλητών με τη λέξη συμμεταβολή.

Οι γραφικές παραστάσεις (διαγράμματα) μας βοηθούν να κατανοήσουμε τη σχέση δύο μεταβλητών και να εκτιμήσουμε τη συσχέτιση (συμμεταβολή) τους (πόσο μεταβάλλεται η

“εξαρτημένη” μεταβλητή για δεδομένη μεταβολή της “ανεξάρτητης μεταβλητής”), ιδιαίτερα όταν οι μεταβλητές είναι ποσοτικές.

Σχέση δύο ποσοτικών μεταβλητών

Διαγράμματα διασποράς

Με ένα διάγραμμα διασποράς (scatter plot) δύο ποσοτικών μεταβλητών μπορούμε να διαπιστώσουμε αν οι μεταβλητές συσχετίζονται (συμμεταβάλλονται). Τα διαγράμματα διασποράς που ακολουθούν παράγονται στον οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn με το ακόλουθο κομμάτι κώδικα (script) (Εικ. 18).

```
# Εισαγωγή της συνάρτησης σχεδίασης
from seaborn import relplot

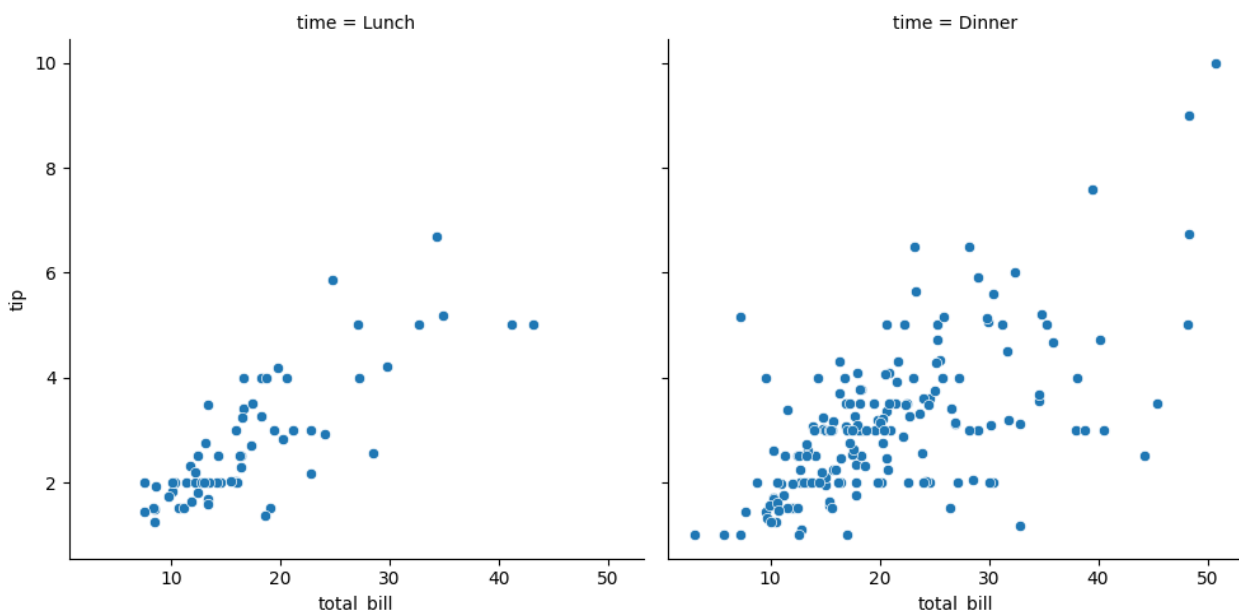
# Σχεδίαση της διασποράς των παρατηρήσεων
relplot(
    data=tips,
    x="total_bill", y="tip", # δύο ποσοτικές μεταβλητές στους άξονες x, y
    #x="size", y="total_bill", # εναλλακτικό ζεύγος αριθμητικής-ποσοτικής μεταβλητής στους άξονες x, y
    #x="size", y="tip", # εναλλακτικό ζεύγος αριθμητικής-ποσοτικής μεταβλητής στους άξονες x, y
    hue='sex', # χρωματική διαφοροποίηση σημείων παρατηρήσεων - όποια κατηγορική ή διακριτή αριθμητική (ή, ακόμη, ποσοτική)
    #style='smoker', # διαφοροποίηση σχήματος σημείων παρατηρήσεων - όποια κατηγορική ή διακριτή αριθμητική
    size='size', # διαφοροποίηση μεγέθους σημείων παρατηρήσεων - όποια κατηγορική ή διακριτή αριθμητική
    col='time', # διακριτά διαγράμματα στις στήλες του πλέγματος, για κατηγορίες της επιλεγμένης μεταβλητής - όποια κατηγορική
    row='day', # διακριτά διαγράμματα στις γραμμές του πλέγματος, για κατηγορίες της επιλεγμένης μεταβλητής - όποια κατηγορική
    #kind='scatter' # επιλογές: 'scatter' (για διάγραμμα διασποράς - default, μπορεί να παραληφθεί), 'line' (για διάγραμμα εξέλιξης)
)
```

Εικόνα 18 Κλήση της συνάρτησης σχεδίασης διαγραμμάτων διασποράς

Για να πάρουμε οπτικοποίηση με ένα ή περισσότερα διαγράμματα διασποράς χρησιμοποιούμε τη figure-level συνάρτηση σχεδίασης relplot (relation plot), θέτοντας στο όρισμα kind την τιμή “scatter” (αν και μπορεί να παραληφθεί, γιατί είναι το default). Ο χαρακτηρισμός figure-level σημαίνει ότι η συνάρτηση παράγει την εικόνα ολόκληρης της οπτικοποίησης, δυνητικά με πολλά διαγράμματα, όλα του ίδιου τύπου (στην προκειμένη περίπτωση, διασποράς). Αν επιθυμούμε να πάρουμε ένα μόνο διάγραμμα (ώστε να το

τοποθετήσουμε, για παράδειγμα, σε υπάρχουσα οπτικοποίηση) χρησιμοποιούμε την axes-level λειτουργία σχεδίασης scatterplot.

Στην ακόλουθη οπτικοποίηση (Εικ. 19) χρησιμοποιούμε δεδομένα διαθέσιμα από το διαδίκτυο για να διερευνήσουμε με ένα διάγραμμα διασποράς αν υπάρχει σχέση μεταξύ των ποσοτικών μεταβλητών ποσού λογαριασμού ('total_bill') και φιλοδωρήματος ('tip'). Ο πίνακας δεδομένων (pandas dataframe) που χρησιμοποιούμε έχει όνομα tips και οι στήλες του είναι: 'total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size'. Δηλώνουμε τον πίνακα δεδομένων από τις στήλες του οποίου αντλούμε τις μεταβλητές της απεικόνισης στο όρισμα data και τις μεταβλητές των αξόνων, 'total_bill' και 'tip', στα ορίσματα x και y. Όλες οι συναρτήσεις σχεδίασης της βιβλιοθήκης Seaborn διαθέτουν τα ορίσματα data, x, y (Stack Overflow, 2021).

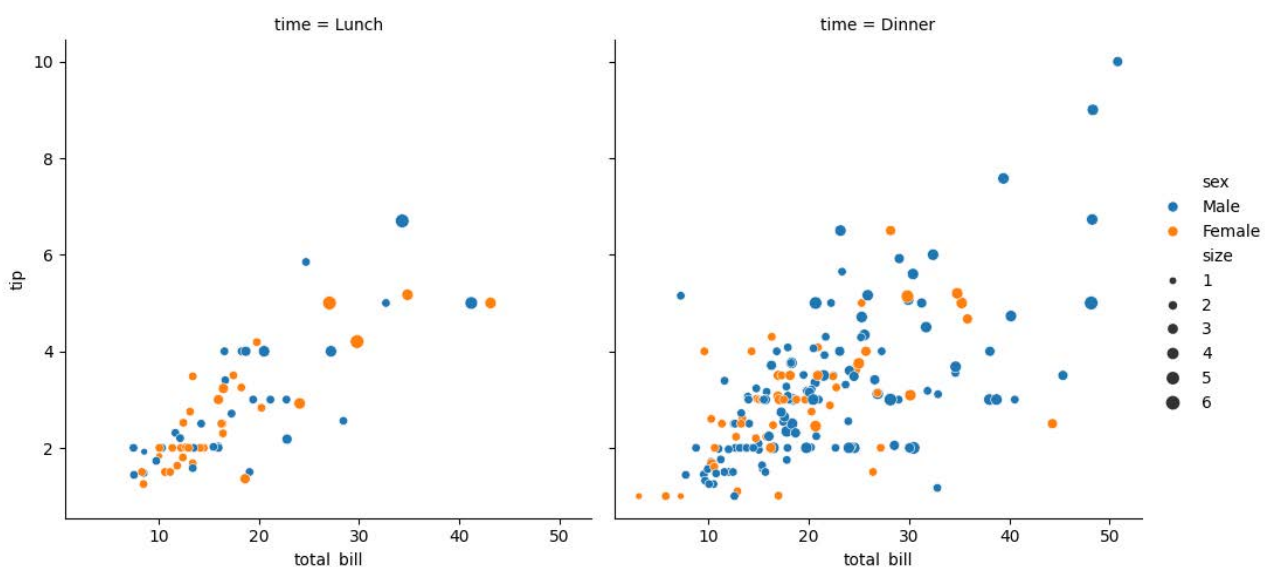


Εικόνα 19 Διαγράμματα διασποράς για γεύματα και δείπνα

Στην απεικόνιση παίρνουμε δύο διαγράμματα διασποράς, ένα για γεύματα (Lunch) και άλλο για δείπνα (Dinner), επειδή μας ενδιαφέρει να εξετάσουμε τη σχέση χωριστά στις δύο περιπτώσεις. Θα μπορούσαμε να εμφανίσουμε όλες τις παρατηρήσεις σε ένα διάγραμμα

διασποράς. Στη βιβλιοθήκη Seaborn έχουμε τη δυνατότητα να εμφανίσουμε περισσότερα του ενός διαγράμματα μέσα στο πλέγμα της απεικόνισης (σε γραμμές ή/και στήλες) με τα ορίσματα `col` και `row`, που είναι διαθέσιμα στην πλειονότητα των συναρτήσεων σχεδίασης.

Οι λειτουργίες σχεδίασης της βιβλιοθήκης Seaborn διαθέτουν δυνατότητες για να ξεχωρίζουμε τα σημεία των παρατηρήσεων, είτε με χρώμα (εδώ διαφορετική απόχρωση με βάση το φύλο, γυναίκες - άνδρες), είτε με το μέγεθος (εδώ με βάση το μέγεθος της παρέας), ακόμη και με το στυλ του σημείου (Εικ. 20). Η διαφοροποίηση με το χρώμα, που επιτυγχάνεται με το όρισμα `hue`, είναι βασική και υποστηρίζεται από την πλειονότητα των λειτουργιών σχεδίασης της βιβλιοθήκης Seaborn για την πλειονότητα των γραφικών (σημεία, γραμμές/καμπύλες, ράβδοι, διαστήματα, κά) που παράγονται από αυτές (τις λειτουργίες σχεδίασης).



Εικόνα 20 Διαγράμματα διασποράς με διαφοροποίηση σε διάφορες ομάδες

Για να μπορείτε να πειραματιστείτε με τη δημιουργία διαγραμμάτων σε στήλες και γραμμές μέσα στο πλέγμα της οπτικοποίησης, με τη χρήση χρωμάτων, μεγέθους, στυλ για τη διάκριση των σημείων των παρατηρήσεων ανατρέξτε στον οδηγό μελέτης, βρείτε τον

κώδικα με τον οποίο σχεδιάζουμε τα διαγράμματα διασποράς (ο κώδικας στην Εικ. 18) και δοκιμάστε να τον τροποποιήσετε και να εκτελέσετε παραλλαγές του.

Και στα δύο διαγράμματα διασποράς της απεικόνισης διαπιστώνουμε ένα ανοδικό μοτίβο/τάση στη διασπορά των παρατηρήσεων, δηλαδή μία ανοδική ζώνη διασποράς, που αποτελεί ένδειξη για συσχέτιση μεταξύ των δύο μεταβλητών. Η σχέση, μάλιστα, δείχνει να είναι γραμμική, αφού η ζώνη διασποράς δε φαίνεται να καμπυλώνει.

Διαγράμματα συσχέτισης

Η διασπορά των παρατηρήσεων δείχνει, λοιπόν, αν υπάρχει κάποιο μοτίβο συσχέτισης (συμμεταβολής) μεταξύ δύο ποσοτικών μεταβλητών. Τα παραπάνω διαγράμματα διασποράς δείχνουν σχέση και μάλιστα γραμμική και για τις δύο περιπτώσεις, δηλαδή και για τα γεύματα και για τα δείπνα.

Μπορούμε να σχεδιάσουμε εκτίμηση αυτής της σχέσης με μία βέλτιστη καμπύλη και μάλιστα διαφορετική καμπύλη για κάθε διάγραμμα (γεύμα ή δείπνο) και κάθε χρώμα (γυναίκες ή άνδρες), αν το επιθυμούμε, επιλέγοντας τις ίδιες ρυθμίσεις (col, row, hue) με αυτές στα διαγράμματα διασποράς. Τα διαγράμματα με τα οποία απεικονίζουμε την εκτίμηση της σχέσης μεταξύ δύο ποσοτικών μεταβλητών τα ονομάζουμε διαγράμματα συσχέτισης.

Για να πάρουμε απεικόνιση με ένα ή περισσότερα διαγράμματα συσχέτισης δύο μεταβλητών χρησιμοποιούμε τη figure-level συνάρτηση σχεδίασης `lmplot` (linear model plot). Όπως η συνάρτηση `relplot` που χρησιμοποιήσαμε για να σχεδιάσουμε διαγράμματα διασποράς έτσι και η συνάρτηση `lmplot` είναι figure-level, που σημαίνει ότι παράγει την εικόνα ολόκληρης της οπτικοποίησης με, δυνητικά, πολλά διαγράμματα, όλα του ίδιου τύπου

(στην προκειμένη περίπτωση, διαγράμματα συσχέτισης). Αν επιθυμούμε να πάρουμε ένα μόνο διάγραμμα (ώστε να το τοποθετήσουμε, για παράδειγμα, σε υπάρχουσα οπτικοποίηση) χρησιμοποιούμε την axes-level λειτουργία σχεδίασης `regplot` (regression plot).

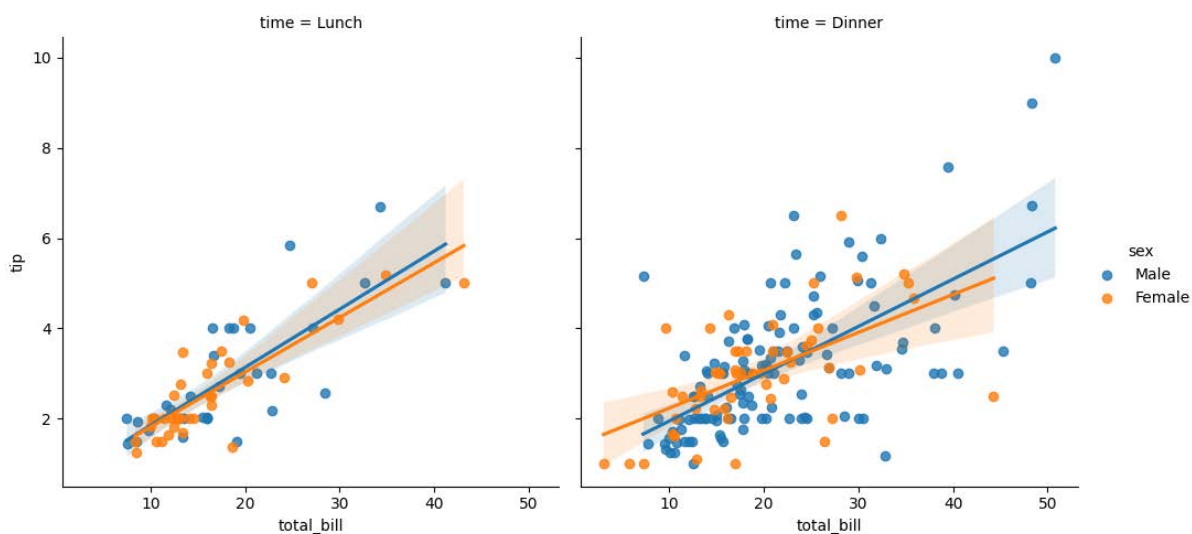
Εντοπίστε στον οδηγό μελέτης οπτικοποίησης δεδομένων το ακόλουθο κομμάτι κώδικα που μας επιτρέπει να σχεδιάσουμε απεικόνιση με διαγράμματα συσχέτισης.

```
# Εισαγωγή συνάρτησης σχεδίασης
from seaborn import lmplot

# Σχεδίαση της θέλτιστης καμπύλης συσχέτισης (με ή χωρίς τις παρατηρήσεις)
lmplot(
    data=tips,
    x="total_bill", y="tip", # δύο ποσοτικές μεταβλητές στους άξονες x, y
    #x="size", y="total_bill", # εναλλακτικό ζεύγος αριθμητικής-ποσοτικής μεταβλητής στους άξονες x, y
    #x="size", y="tip", # εναλλακτικό ζεύγος αριθμητικής-ποσοτικής μεταβλητής στους άξονες x, y
    hue='sex', # χρωματική διαφοροποίηση σημείων παρατηρήσεων - όποια κατηγορική ή διακριτή αριθμητική (ή, ακόμη, ποσοτική)
    col='time', # διακριτά διαγράμματα στις στήλες του πλέγματος, για κατηγορίες της επιλεγμένης μεταβλητής - όποια κατηγορική
    #row='day', # διακριτά διαγράμματα στις γραμμές του πλέγματος, για κατηγορίες της επιλεγμένης μεταβλητής - όποια κατηγορική
    #order=2, # επιλογές: γραμμική σχέση (order=1 ή απενεργοποίηση (σχολιασμός)) ή ανώτερης τάξης (order>1)
    #scatter=False, # αν θα εμφανίσει τη διασπορά των παρατηρήσεων (scatter=True, default) ή όχι (scatter=False)
    #fit_reg=False, # αν θα εκτιμήσει και θα εμφανίσει την καμπύλη συσχέτισης μεταξύ των μεταβλητών x και y (fit_reg=True, default) ή όχι
)
```

Εικόνα 21 Κλήση της συνάρτησης σχεδίασης συσχέτισης

Εκτελώντας αυτόν τον κώδικα μπορείτε να πάρετε την ακόλουθη απεικόνιση (Εικ. 22). Μπορείτε να δοκιμάσετε δικές σας παραλλαγές εκτέλεσης, σχολιάζοντας και αποσχολιάζοντας ρυθμίσεις ορισμάτων στον κώδικα, ώστε να ελέγξετε τα εναλλακτικά αποτελέσματα.



Εικόνα 22 Διαγράμματα συσχέτισης για γεύματα και δείπνα

Η οπτικοποίηση περιέχει δύο διαγράμματα συσχέτισης (ένα για γεύματα και ένα για δείπνα) και τέσσερις διαφορετικές καμπύλες, για κάθε διάγραμμα (γεύμα ή δείπνο) και κάθε χρώμα (γυναίκες ή άνδρες). Παίρνουμε διακριτές καμπύλες σε κάθε διάγραμμα για κάθε χρώμα επειδή ζητάμε χρωματική διαφοροποίηση για γυναίκες - άνδρες.

Επειδή το μοτίβο της διασποράς δείχνει ευθύγραμμο, αφήνουμε στη συνάρτηση σχεδίασης τη default ρύθμιση, `order=1`, για να εκτιμήσει ευθεία γραμμή (την οποία και εμφανίζει για κάθε περίπτωση, γεύματα - δείπνα, άνδρες - γυναίκες, συνολικά 4 βέλτιστες καμπύλες). Γύρω από την κάθε καμπύλη εμφανίζεται έγχρωμη μία ζώνη που δείχνει την αβεβαιότητα (το περιθώριο σφάλματος) στην εκτίμηση της καμπύλης.

Εκτός από τις καμπύλες εμφανίζεται η διασπορά των παρατηρήσεων, (default ρύθμιση, `scatter=True`). Μπορούμε να επιλέξουμε να μην εμφανίζεται η διασπορά θέτοντας `scatter=False`. Ή να εμφανίζεται μόνο η διασπορά των παρατηρήσεων (όχι οι καμπύλες) επιλέγοντας `fit_reg=False` (φυσικά, default ρύθμιση είναι να εμφανίζονται οι καμπύλες, `fit_reg=True`, όπως και η διασπορά, `scatter=True`).

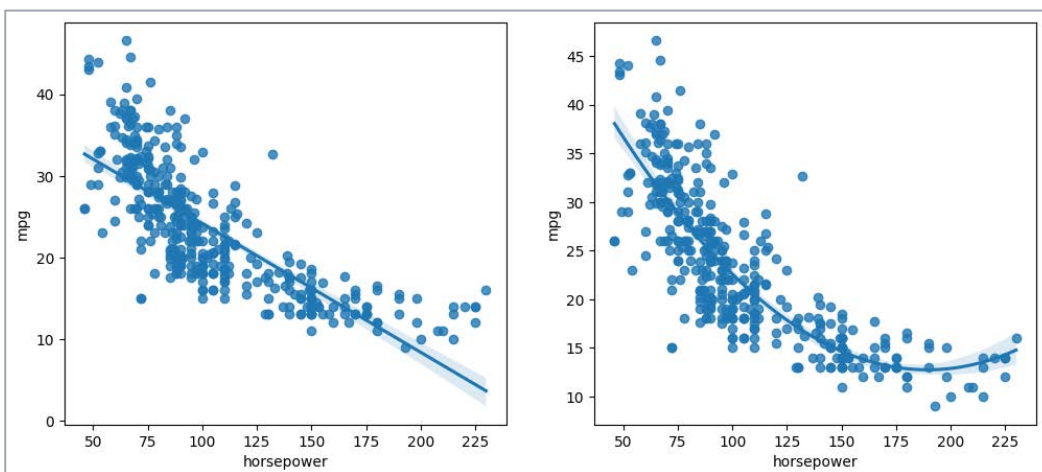
Μη γραμμική σχέση - διάγραμμα αποκλίσεων

Ας εξετάσουμε μία περίπτωση μη γραμμικής (μη ευθύγραμμης) σχέσης δύο ποσοτικών μεταβλητών. Μη γραμμική σχέση είναι αυτή της κατανάλωσης (miles per gallon, 'mpg') και ιπποδύναμης ('horsepower'). Χρησιμοποιούμε ένα σύνολο δεδομένων από το αποθετήριο της Seaborn με όνομα mpg, που περιέχει στοιχεία για παλιά αυτοκίνητα.

Χρησιμοποιούμε το κομμάτι κώδικα (script) που παρουσιάσαμε προηγουμένως (Εικ. 21), το οποίο μας επιτρέπει να σχεδιάσουμε απεικόνιση με διαγράμματα συσχέτισης. Παίρνουμε

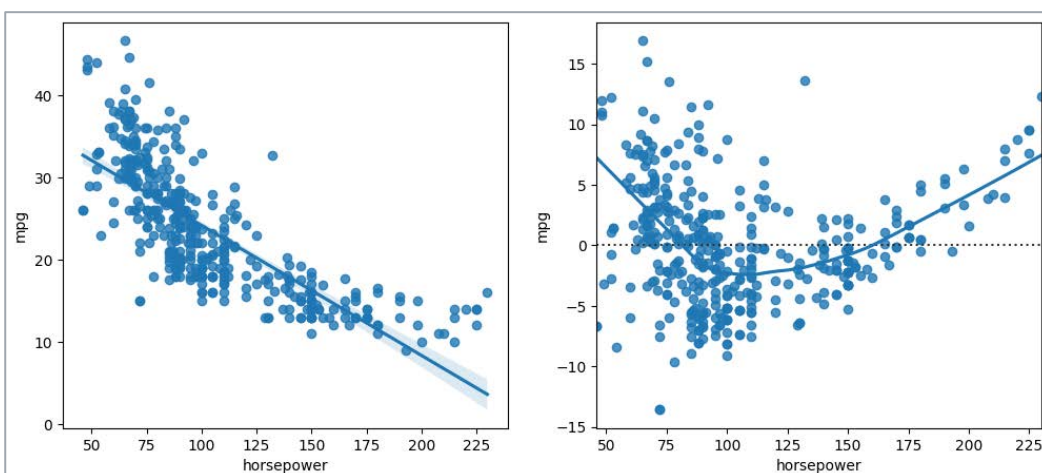
διασπορά παρατηρήσεων που σχηματίζει καμπυλόγραμμη ζώνη. Η γραμμική εκτίμηση ($\text{order}=1$ ή παράλειψη, αφού είναι η default ρύθμιση) δεν ταιριάζει στα δεδομένα (underfits the data).

Αυτό σημαίνει ότι πρέπει να ζητήσουμε από τη συνάρτηση σχεδίασης `lmplot` ανώτερης τάξης εκτίμηση. Παρατηρούμε, ότι αν ζητήσουμε καμπύλη δεύτερης τάξης ($\text{order}=2$), αυτή δείχνει να ταιριάζει γάντι στα δεδομένα (Εικ. 23).



Εικόνα 23 Διαγράμματα κατανάλωσης και ιπποδύναμης (μή γραμμική σχέση)

Αλλά και το διάγραμμα αποκλίσεων (των παρατηρήσεων από την καμπύλη) στο δεξιό μέρος της παρακάτω εικόνας (Εικ. 24) δείχνει ξεκάθαρα μη γραμμική σχέση.



Εικόνα 24 Διαγράμματα συσχέτισης και αποκλίσεων όπου φαίνεται μη γραμμική σχέση

Παίρνουμε το διάγραμμα αποκλίσεων εκτελώντας την axes-level συνάρτηση σχεδίασης `residplot` με το ακόλουθο κομμάτι κώδικα (Εικ. 25). Δείτε στον οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn για περισσότερες λεπτομέρειες.

```
# Σχεδίαση της διασποράς των αποκλίσεων των μετρήσεων (των residuals) για εκτίμηση (γραμμική ή διωνυμική)
residplot(
    data=mpg,
    x="horsepower", y="mpg", # δύο ποσοτικές μεταβλητές στους άξονες x, y
    #order=2,               # γραμμική σχέση (order=1, default) ή ανώτερης τάξης (order>1)
    lowess=True,            # αν θα προστεθεί non-parametric καμπύλη εξομάλυνσης του σχηματισμού (ζώνης)
    ax=axes2,
)
```

Εικόνα 25 Κλήση της συνάρτησης σχεδίασης διασποράς των αποκλίσεων (`residplot`)

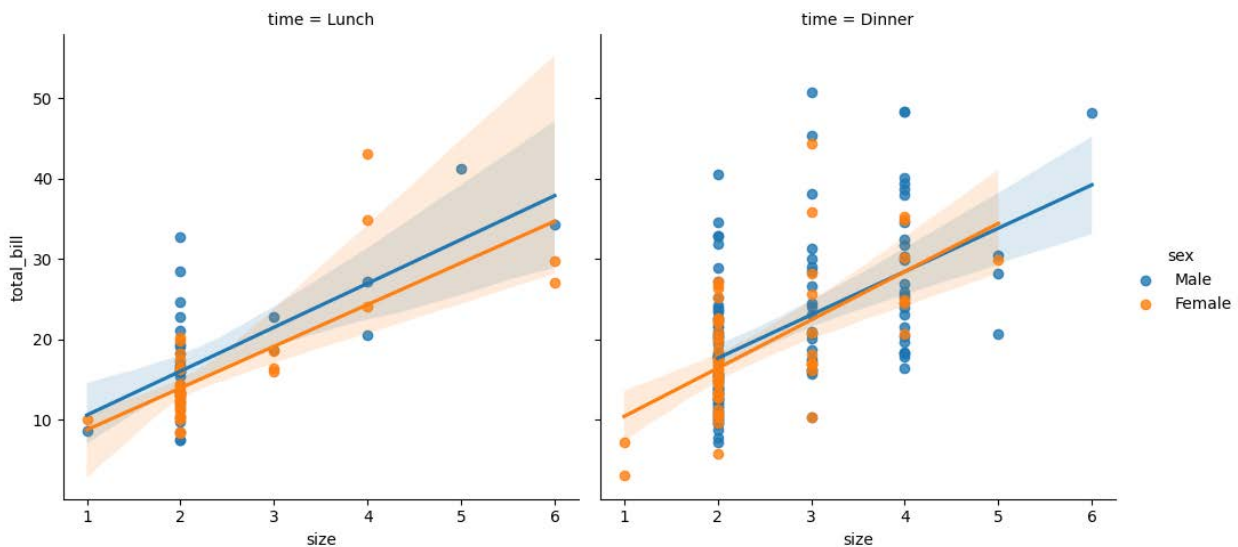
Αν ζητήσουμε από τη συνάρτηση σχεδίασης `residplot` γραμμική σχέση (`order=1` ή παράλειψη, αφού είναι η default ρύθμιση), τότε τα σημεία των αποκλίσεων, δημιουργούν ξεκάθαρο σχηματισμό (μοτίβο), διαπίστωση που μας οδηγεί να απορρίψουμε την υπόθεση γραμμικότητας (Εικ. 24). Αντίθετα, αν υποθέσουμε μη γραμμική σχέση (`order=2`), τότε ο σχηματισμός χάνεται.

Σχέση κατηγορικής-ποσοτικής μεταβλητής - κατηγορικά διαγράμματα

Ας διερευνήσουμε τη σχέση μίας κατηγορικής με μία ποσοτική μεταβλητή (Εικ. 26). Ας πάρουμε διαγράμματα διασποράς του μεγέθους παρέας ('size') με το ποσό λογαριασμού ('total_bill') ή το φιλοδώρημα ('tip'). Αντλούμε τις μεταβλητές από το σύνολο δεδομένων (pandas dataframe) `tips` (με στήλες: 'total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size').

Το μέγεθος παρέας ('size') είναι αριθμητική μεταβλητή και παίρνει μερικές ακέραιες τιμές. Συμπεριφέρεται, επομένως, σαν κατηγορική μεταβλητή. Μαζί με τη διασπορά των παρατηρήσεων μεγέθους παρέας και ποσού λογαριασμού ας εμφανίσουμε τη καμπύλη

συσχέτισης. Η συσχέτιση δείχνει να είναι θετική: μεγαλύτερες παρέες τείνουν να κάνουν μεγαλύτερους λογαριασμούς.



Εικόνα 26 Διαγράμματα διασποράς του μεγέθους παρέας ('size') με το ποσό λογαριασμού ('total_bill')

Διασπορά και συσχέτιση παίρνουμε με τον κώδικα που μας επιτρέπει να σχεδιάσουμε απεικόνιση με διαγράμματα συσχέτισης με την figure-level λειτουργία σχεδίασης `lmplot` (Εικ. 27).

```
# Εισαγωγή συνάρτησης σχεδίασης
from seaborn import lmplot

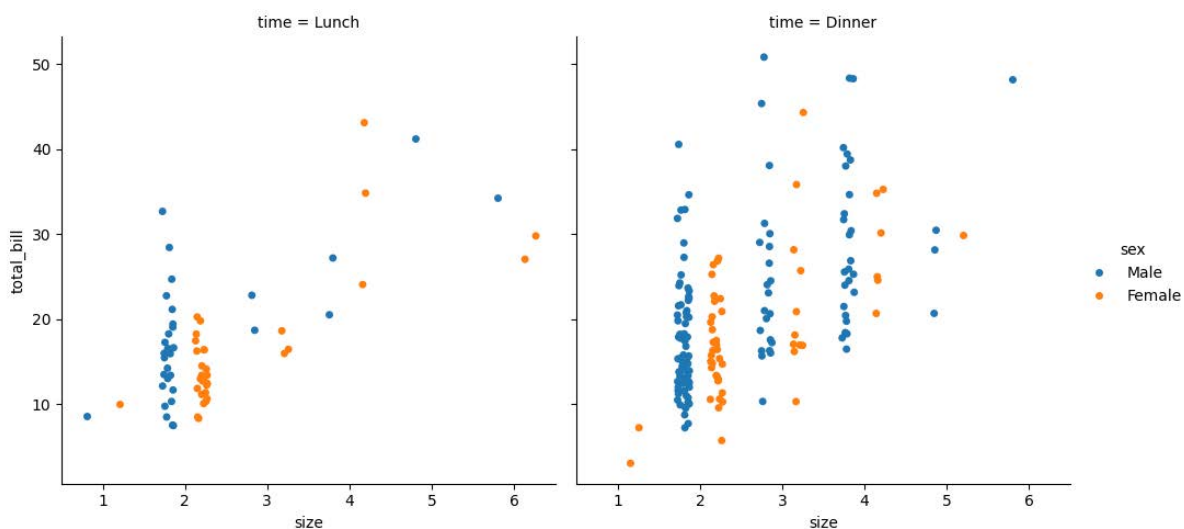
# Σχεδίαση της θέλτιστης καμπύλης συσχέτισης (με ή χωρίς τις παρατηρήσεις)
lmplot(
    data=tips,
    #x="total_bill", y="tip", # δύο ποσοτικές μεταβλητές στους άξονες x, y
    x="size", y="total_bill", # εναλλακτικό ζεύγος αριθμητικής-ποσοτικής μεταβλητής στους άξονες x, y
    #x="size", y="tip", # εναλλακτικό ζεύγος αριθμητικής-ποσοτικής μεταβλητής στους άξονες x, y
    hue='sex', # χρωματική διαφοροποίηση σημείων παρατηρήσεων - όποια κατηγορική ή διακριτή αριθμητική (ή, ακόμη, ποσοτική)
    col='time', # διακριτά διαγράμματα στις στήλες του πλέγματος, για κατηγορίες της επιλεγμένης μεταβλητής - όποια κατηγορική
    #row='day', # διακριτά διαγράμματα στις γραμμές του πλέγματος, για κατηγορίες της επιλεγμένης μεταβλητής - όποια κατηγορική
    #order=2, # επιλογές: γραμμική σχέση (order=1 ή απενεργοποίηση (σχολιασμός)) ή ανώτερης τάξης (order>1)
    #scatter=False, # αν θα εμφανίσει τη διασπορά των παρατηρήσεων (scatter=True, default) ή όχι (scatter=False)
    #fit_reg=False, # αν θα εκτιμήσει και θα εμφανίσει την καμπύλη συσχέτισης μεταξύ των μεταβλητών x και y (fit_reg=True, default) ή όχι
)
```

Εικόνα 27 Κλήση της συνάρτησης σχεδίασης συσχέτισης

Η διασπορά των παρατηρήσεων συμβαίνει κατά τη διεύθυνση της ποσοτικής μεταβλητής (την κατακόρυφη διεύθυνση στη συγκεκριμένη απεικόνιση) για κάθε τιμή της κατηγορικής μεταβλητής. Παρατηρούμε ότι η διασπορά των παρατηρήσεων πάνω στη διεύθυνση της

ποσοτικής μεταβλητής σχηματίζει πυκνώματα και αραιώματα. Μπορούμε να δώσουμε έμφαση στην πυκνότητα των παρατηρήσεων με ένα κατηγορικό διάγραμμα διασποράς.

Στην ακόλουθη απεικόνιση (Εικ. 28) παίρνουμε κατηγορικά διαγράμματα διασποράς του μεγέθους παρέας ('size') με το ποσό λογαριασμού ('total_bill'). Στα κατηγορικά διαγράμματα διασποράς τα σημεία των παρατηρήσεων αποκλίνουν τεχνητά, εκατέρωθεν της διεύθυνσης της ποσοτικής μεταβλητής (σχηματίζοντας λωρίδες, strip plot με ρύθμιση jitter, ή σμήνη, swarm plot), ώστε να διακρίνονται ευκολότερα τα πυκνώματα και τα αραιώματα.



Εικόνα 28 Διαγράμματα διασποράς του μεγέθους παρέας ('size') με το ποσό λογαριασμού ('total_bill')

Η παρατήρηση ότι η διασπορά των παρατηρήσεων μίας ποσοτικής μεταβλητής, όπως το ποσό λογαριασμού, δεν παρουσιάζει την ίδια πυκνότητα σε όλα τα σημεία, μας οδηγεί με διαισθητικό τρόπο στην έννοια της πυκνότητας πιθανότητας και της κατανομής πιθανότητας. Για την απεικόνιση της κατανομής πιθανότητας ποσοτικών και κατηγορικών μεταβλητών χρησιμοποιούμε διαγράμματα κατανομών, όπως ιστογράμματα και διαγράμματα εκτίμησης πυκνότητας (το τελευταίο μόνο για ποσοτικές μεταβλητές). Επίσης, χρησιμοποιούμε

κατηγορικά διαγράμματα κατανομών και κεντρικής τάσης, που συμπληρώνουν τα κατηγορικά διαγράμματα διασποράς.

Κατηγορικά διαγράμματα διασποράς παίρνουμε με το ακόλουθο κομμάτι κώδικα (Εικ. 29).

```
catplot(
    data=tips,
    #x="day", y="total_bill", # δύο μεταβλητές στους άξονες x, y, μία κατηγορική και μία ποσοτική
    #x="day", y="tip",       # εναλλακτική περίπτωση ποσοτικής μεταβλητής, σχετιζόμενης με την κατηγορική
    x="size", y="total_bill", # εναλλακτική περίπτωση κατηγορικής μεταβλητής, η αριθμητική 'size', με διακριτές τιμές (σαν κατηγορική)
    #x="size", y="tip",       # εναλλακτική περίπτωση "κατηγορικής" και ποσοτικής μεταβλητής
    hue='sex',               # για χρωματική διαφοροποίηση των σημείων/γραφικών, ένα χρώμα για κάθε κατηγορία/επίπεδο της επιλεγμένης μεταβλητής - όπ
    #hue='smoker',           # εναλλακτική μεταβλητή για χρωματική διαφοροποίηση των σημείων/γραφικών
    col='time',              # (μόνο για figure-level) για διακριτά διαγράμματα στις στήλες του πλέγματος, ένα για κάθε κατηγορία/επίπεδο της επιλεγμ
    #row='size',             # (μόνο για figure-level) για διακριτά διαγράμματα στις γραμμές του πλέγματος, ένα για κάθε κατηγορία/επίπεδο της επιλε
    dodge=True,             # σημεία κάθε κατηγορίας σε μία στήλη (False, default) ή σε πολλές (True), όσες τα επίπεδα της μεταβλητής hue
    #kind="swarm",          # (μόνο για τη catplot) μπορούμε να παραλείψουμε αυτό το όρισμα, αν η επιλογή είναι η 'strip' (η default)
)
```

Εικόνα 29 Κλήση της συνάρτησης σχεδίασης κατηγορικού διαγράμματος διασποράς

Μπορείτε να δοκιμάσετε παραλλαγές του συγκεκριμένου κώδικα (Εικ. 29) για κατηγορικά διαγράμματα διασποράς, αλλά και άλλες προτάσεις (scripts), στον οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn.

Εξέλιξη μεταβλητής με το χρόνο

Για να απεικονίσουμε τη μεταβολή ενός μεγέθους (μεταβλητής) με την πάροδο του χρόνου ή, γενικότερα, σε συνάρτηση με μία μεταβλητή που αυξάνεται ομαλά προς την ίδια πάντα κατεύθυνση, όπως ο χρόνος, χρησιμοποιούμε την axes-level συνάρτηση σχεδίασης `lineplot` (για ένα διάγραμμα) ή τη figure-level συνάρτηση σχεδίασης `relplot` με ρύθμιση `kind='line'` (για δυνητικά πολλά διαγράμματα εξέλιξης).

Με το ακόλουθο κομμάτι κώδικα (Εικ. 30) παίρνουμε την εξέλιξη του προσδόκιμου ζωής ("OBS_VALUE") με την πάροδο των ετών ("TIME_PERIOD"), (από το αποθετήριο UNICEF

Data Warehouse. Data) σε μία οπτικοποίηση που σχεδιάζουμε με τη figure level συνάρτηση σχεδίασης relplot (με kind='line').

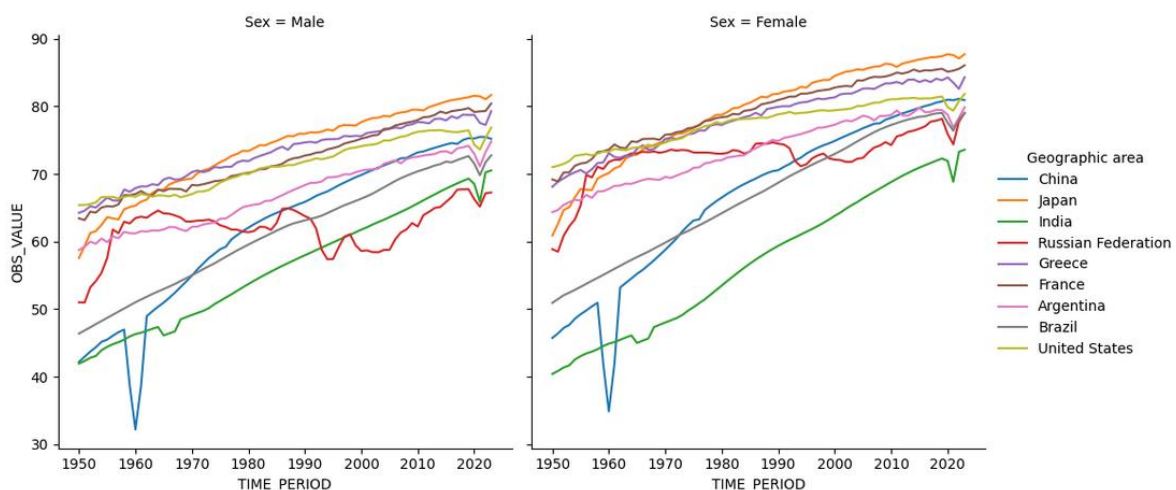
```
from seaborn import relplot, lineplot

# Σχεδίαση των γραφημάτων εξέλιξης
relplot(
    data=life_exp_filtered,          # χρησιμοποιούμε το φιλτραρισμένο σύνολο, με τις επιλεγμένες χώρες
    x="TIME_PERIOD", y="OBS_VALUE", # το έτος στον άξονα x και το προσδόκιμο ζωής στον άξονα y
    hue="Geographic area",          # η χώρα για χρωματική διαφοροποίηση
    col="Sex",                      # διακριτά διαγράμματα με βάση το φύλο
    kind = 'line',                 # διαγράμματα εξέλιξης
)
```

Εικόνα 30 Κλήση της συνάρτησης σχεδίασης relplot (με kind='line')

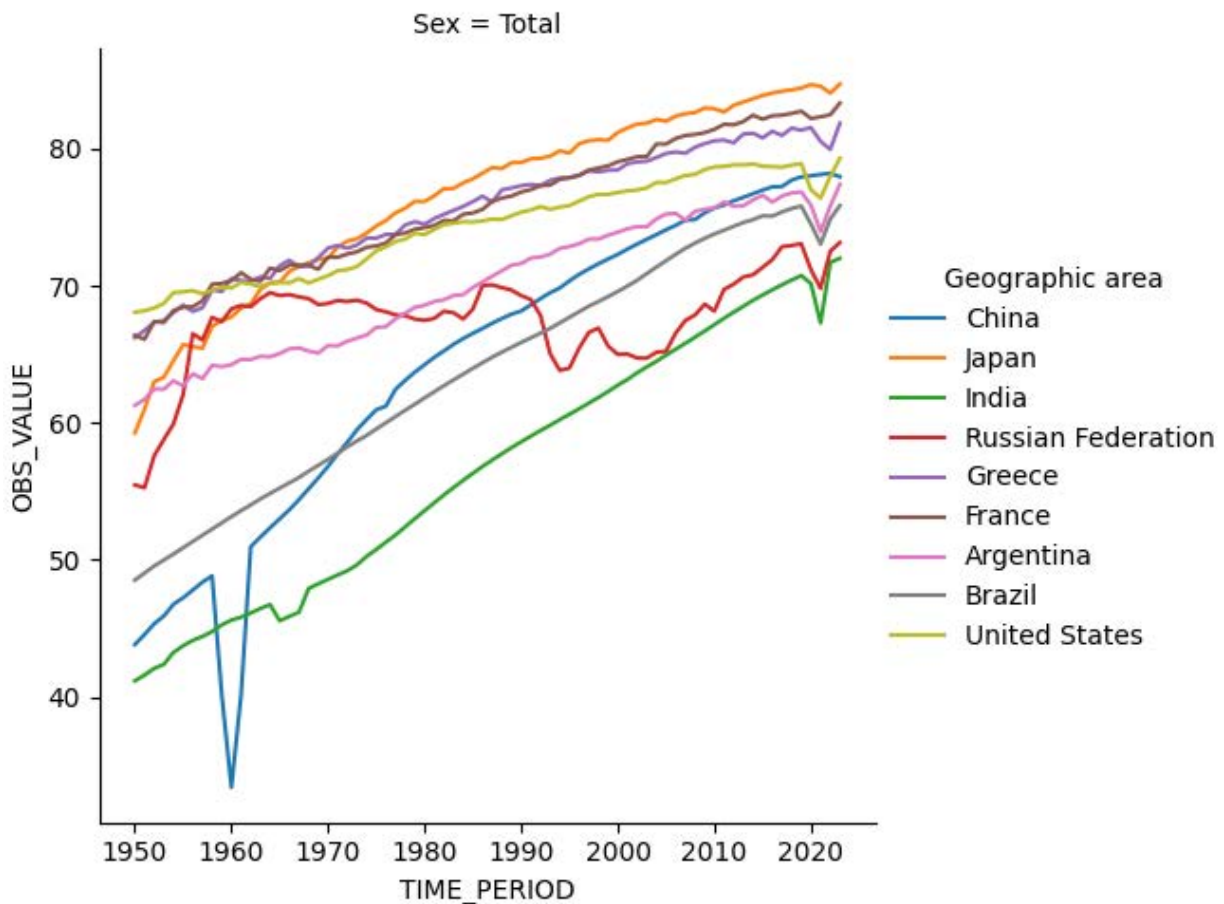
Για να δοκιμάσετε παραλλαγές του κώδικα και να ελέγξετε πως φορτώνουμε και φιλτράρουμε το σύνολο δεδομένων, πριν το χρησιμοποιήσουμε κατευθυνθείτε στον οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn.

Παίρνουμε διαφορετικά διαγράμματα, για γυναίκες και άνδρες θέτοντας col="Sex". Σε κάθε διάγραμμα της Εικ. 31 παίρνουμε, με διαφορετικό χρώμα, μία γραμμή εξέλιξης του προσδόκιμου ζωής για κάθε χώρα (από τις επιλεγμένες στο φιλτράρισμα του συνόλου δεδομένων), δηλώνοντας hue="Geographic area".



Εικόνα 31 Διαγράμματα εξέλιξης του προσδόκιμου ζωής για άνδρες και γυναίκες

Όπως αναμένεται η γυναίκες ζουν περισσότερο. Ας πάρουμε και ένα συνολικό διάγραμμα. Παρατηρούμε μικρό αλλά σταθερό ρυθμό αύξησης του προσδόκιμου ζωής για αναπτυγμένες χώρες, όπως η Γαλλία και η Ιαπωνία, αλλά ραγδαία αύξηση του προσδόκιμου ζωής για αναπτυσσόμενες χώρες, όπως η Κίνα, η Βραζιλία, η Ινδία (Εικ. 32).



Εικόνα 32 Διάγραμμα εξέλιξης του προσδόκιμου ζωής σε αναπτυσσόμενες χώρες

Κατανομές μεταβλητών

Απεικονίζουμε την κατανομή πιθανότητας μίας κατηγορικής μεταβλητής με ιστόγραμμα (hist plot).

Απεικονίζουμε την κατανομή πιθανότητας μίας ή ζεύγους ποσοτικών μεταβλητών με ιστόγραμμα (hist plot) ή διάγραμμα εκτίμησης πυκνότητας πιθανότητας (kernel density estimate or kde plot).

Για την απεικόνιση της κατανομής μίας ποσοτικής μεταβλητής υπάρχει και το διάγραμμα σωρευτικής πυκνότητας (empirical cumulative density function or ecdf plot).

Όλες οι απεικονίσεις της κατανομής μίας ή δύο μεταβλητών αποτελούν εκτιμήσεις (προσεγγίσεις) της κατανομής με βάση τα διαθέσιμα δεδομένα. Επομένως, οι απεικονίσεις κατανομών είναι εμπειρικές, καθώς προκύπτουν από δεδομένα.

Κατανομή κατηγορικής μεταβλητής

Για να απεικονίσουμε την κατανομή πιθανότητας μίας κατηγορικής μεταβλητής έχουμε ένα μοναδικό τρόπο, με ιστογράμματα (hist plots).

Σχεδίαση ιστογράμματος κατηγορικής μεταβλητής

Κώδικας. Παίρνουμε ιστογράμματα κατηγορικής μεταβλητής εκτελώντας τη figure-level συνάρτηση σχεδίασης `displot` της βιβλιοθήκης Seaborn, με `kind='hist'`. Στο παρακάτω κομμάτι κώδικα (Εικ. 33) περιλαμβάνονται κάποιες βασικές ρυθμίσεις της συνάρτησης, Mulani, S. (2022,).

Ανάλογα με το αποτέλεσμα που επιθυμούμε κάθε φορά κάνουμε στις ρυθμίσεις τις κατάλληλες τροποποιήσεις. Στο σχεδιασμό ιστογραμμάτων παραλείπουμε τη ρύθμιση `kind='hist'` (το βλέπετε σχολιασμένο), γιατί το ιστόγραμμα είναι ο default τύπος διαγράμματος της `displot`.

```
displot(
    data=titanic,          # το σύνολο δεδομένων (υπό μορφή pandas dataframe) από τις στήλες του οποίου αντλούνται οι μεταβλητές της οπτικοποίησης
    x="survived",          # η κατηγορική μεταβλητή στον άξονα x, για κατακόρυφο προσανατολισμό των ράβδων (ή, σπανιότερα, στον άξονα y)
    hue='sex',             # για χρωματική διαφοροποίηση των σημείων/γραφικών, ένα χρώμα για κάθε κατηγορία/επίπεδο της επιλεγμένης μεταβλητής - όπω
    hue='class',           # (μόνο για figure-level) για διακριτά διαγράμματα στις στήλες του πλέγματος, ένα για κάθε κατηγορία/επίπεδο της επιλεγμ
    #col='class',          # (μόνο για figure-level) για διακριτά διαγράμματα στις γραμμές του πλέγματος, ένα για κάθε κατηγορία/επίπεδο της επιλεγμ
    #row='sex',            # το στατιστικό που θέλουμε να μετράται με το ύψος ή το εμβαδό των ράβδων (default: 'count')
    stat='percent',        # επιλογή μίας μόνο ρύθμισης stat: percent (bar heights normalize to 100%) ή 'probability'/'proportion'
    #stat='proportion',    # δήλωση ότι πρόκειται για κατηγορική ή αριθμητική μεταβλητή (με διακριτές τιμές)
    discrete=True,        # συμπίεση του εύρους των ράβδων ώστε να μην εφάπτονται και να ξεχωρίζουν (οπτικό αποτέλεσμα)
    shrink=0.8,           # ράβδοι σε στρώσεις ('layer', default), ή ξεχωριστά ('dodge'), ή στοιβαγμένες ('stack')
    multiple='dodge',      # κανονικοποίηση σε όλο το σύνολο (common_norm=True, default) ή στις επιμέρους (σημασιολογικές) ομάδες (common_norm=False
    #common_norm=False,   # η ρύθμιση μπορεί να παραληφθεί, γιατί είναι η default για τη συνάρτηση displot
    kind='hist',
)
```

Εικόνα 33 Κλήση της συνάρτησης σχεδίασης `displot`

Μπορείτε να βρείτε το συγκεκριμένο κώδικα (script) της Εικ. 33 και να δοκιμάσετε παραλλαγές εκτέλεσής του στο σημειωματάριο του οδηγού μελέτης, "Data Visualization with Seaborn.ipynb". Θυμίζουμε ότι ένα σημειωματάριο (όπως είναι ο οδηγός μελέτης) μπορούμε να το εκτελέσουμε μέσα σε περιβάλλον Jupyter Notebook.

Συναρτήσεις σχεδίασης. Η `displot`, σαν figure-level συνάρτηση σχεδίασης που είναι, έχει δυνατότητα παραγωγής οπτικοποίησης με πολλά διαγράμματα, ιστογράμματα στην προκειμένη περίπτωση, γιατί `kind='hist'`. Εναλλακτικά, αν θέλουμε ένα ιστόγραμμα, για να το ενσωματώσουμε σε υπάρχουσα οπτικοποίηση, μπορούμε να χρησιμοποιήσουμε την axes-level συνάρτηση σχεδίασης `histplot`.

Σύνολο δεδομένων. Το σύνολο δεδομένων που χρησιμοποιούμε για να μελετήσουμε τη σχεδίαση απεικονίσεων κατηγορικής μεταβλητής αφορά το ναυάγιο του Τιτανικού. Περιέχει αρκετές κατηγορικές μεταβλητές, όπως το φύλο του επιβάτη ('sex'), τη θέση του εισιτηρίου ('class'), αν επιβίωσε ή όχι ('survived'). Το σύνολο δεδομένων το δηλώνουμε στο

όρισμα `data`, `data=titanic`. Μπορείτε εύκολα να δοκιμάσετε σύνολα δεδομένων που αντιπροσωπεύουν δικά σας προβλήματα και μεταβλητές.

Να θυμίσουμε ότι ένα σύνολο δεδομένων φορτώνεται σε ένα `pandas dataframe` και δηλώνεται στο όρισμα `data` με το όνομα που έχουμε δώσει στο `dataframe` (για παράδειγμα, `data=titanic`). Στο `pandas dataframe` έχουμε ένα πίνακα δεδομένων με μία υποψήφια μεταβλητή σε κάθε στήλη και μία παρατήρηση σε κάθε γραμμή. Τις στήλες του συνόλου δεδομένων τις χρησιμοποιούμε στις συναρτήσεις σχεδίασης της βιβλιοθήκης Seaborn για να δηλώσουμε μεταβλητές στους άξονες (ορίσματα `x`, `y`) και στις σημασιολογικές ρυθμίσεις (ορίσματα `hue`, `col` και `row`).

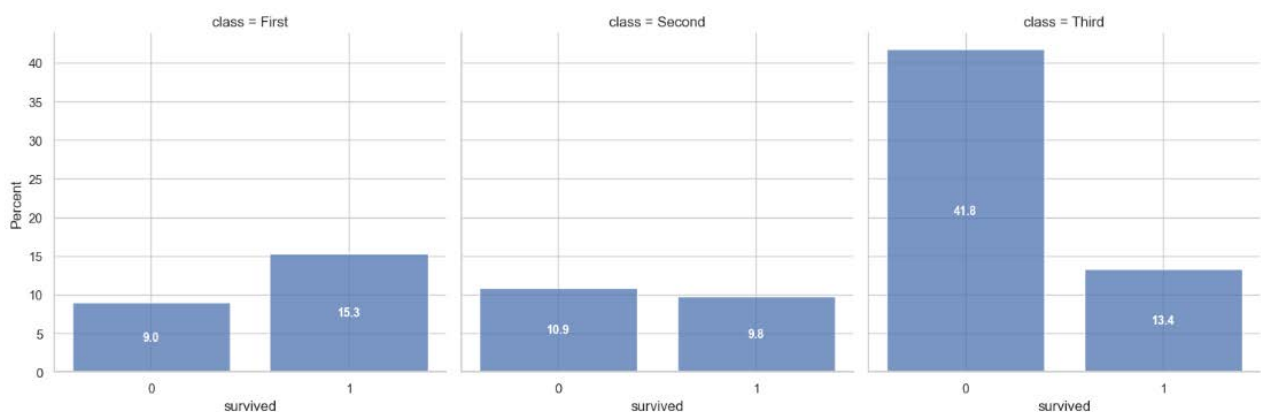
Ρυθμίσεις ιστογράμματος κατηγορικής μεταβλητής

Χρησιμοποιούμε το όρισμα `discrete` για να δηλώσουμε ότι πρόκειται για ιστογράμματα κατηγορικής μεταβλητής, δηλαδή μεταβλητής με διακριτά επίπεδα (κατηγορίες). Θέτουμε `discrete=True`.

Στο ιστόγραμμα κατηγορικής μεταβλητής έχουμε μία ράβδο (ιστό) για κάθε κατηγορία (επίπεδο) της μεταβλητής. Το ύψος/μήκος της κάθε ράβδου εκφράζει την πιθανότητα μία νέα παρατήρηση να ανήκει στην κατηγορία που αντιστοιχεί η ράβδος.

Στην ακόλουθη απεικόνιση (Εικ. 34) για το ναυάγιο του Τιτανικού, εκφράζουμε τις πιθανότητες επιβίωσης (1) και αποβίωσης (0) (μεταβλητή `'survived'`) για τους επιβάτες κάθε θέσης, 1ης, 2ης και 3ης ως ποσοστά (percent). Η οπτικοποίηση περιέχει τρία διαγράμματα ένα για κάθε θέση (μεταβλητή `'class'`), γιατί έχουμε ορίσει `col='class'`.

Στα ιστογράμματα κατηγορικών μεταβλητών η πιθανότητα κάθε κατηγορίας εκφράζεται σαν ποσοστό ή αναλογία του πλήθους των παρατηρήσεων της κατηγορίας επί του συνόλου των παρατηρήσεων (που συμμετέχουν στον υπολογισμό της κατανομής). Αυτό γίνεται με ρύθμιση του ορίσματος `stat`, που σημαίνει στατιστικό (statistic). Θέτουμε `stat='percent'` ή `stat='proportion'`.



Εικόνα 34 Κατανομή πιθανότητας επιβίωσης στο ναυάγιο του Τιτανικού. Η οπτικοποίηση περιέχει τρία διαγράμματα ένα για κάθε θέση (μεταβλητή 'class')

Αν η πιθανότητα εκφράζεται σαν ποσοστό, τότε το άθροισμα των μηκών των ράβδων στην κατανομή είναι 100, αν εκφράζεται σαν αναλογία τότε το άθροισμα των μηκών των ράβδων στην κατανομή είναι 1.

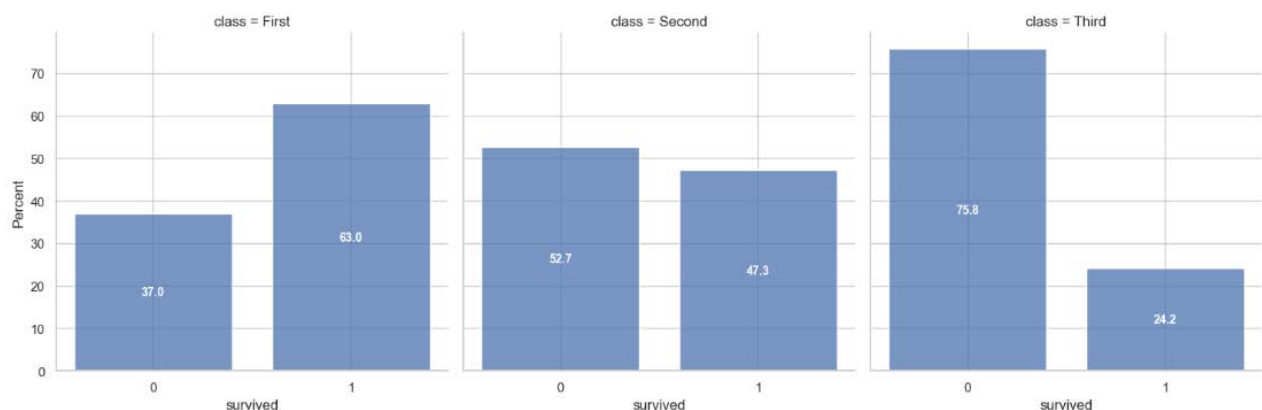
Ενιαία (συνδυαστική) κατανομή και υπό συνθήκη κατανομές

Μία κατανομή. Στην προηγούμενη απεικόνιση έχουμε μία κατανομή των δύο μεταβλητών ('survived', 'class'), που εκτείνεται και στα τρία διαγράμματα. Αυτό σημαίνει ότι παίρνουμε άθροισμα 100 (ή 1 στην περίπτωση αναλογίας) αν αθροίσουμε όλες τις πιθανότητες, δηλαδή τα μήκη όλων των ράβδων (και των έξι ράβδων, από όλα τα διαγράμματα).

Διασφαλίζουμε ότι θα πάρουμε μία κατανομή με ρύθμιση του ορίσματος `common_norm`, που σημαίνει κοινή κανονικοποίηση, δηλαδή κανονικοποίηση/άθροιση σε ένα, κοινό σύνολο, για να πάρουμε μία ενιαία κατανομή. Για να πάρουμε κοινή κανονικοποίηση και μία ενιαία κατανομή θέτουμε `common_norm=True`.

Υπό συνθήκη κατανομές. Μπορούμε να πάρουμε διαφορετικές κατανομές πιθανότητας για επιμέρους ομάδες στις παρατηρήσεις μας, για παράδειγμα, μία κατανομή της μεταβλητής 'survived' για κάθε θέση, 'class' (υπό συνθήκη κατανομές, δεδομένης της θέσης). Προϋπόθεση είναι η ύπαρξη τουλάχιστον μίας από τις ρυθμίσεις `hue`, `col` ή `row`, ώστε να ορίζονται επιμέρους ομάδες μέσα στις οποίες θα γίνει ο υπολογισμός των διακριτών κατανομών.

Στην ακόλουθη απεικόνιση (Εικ. 35) έχουμε ορίσει `col='class'` για να πάρουμε ξεχωριστό διάγραμμα για κάθε θέση.



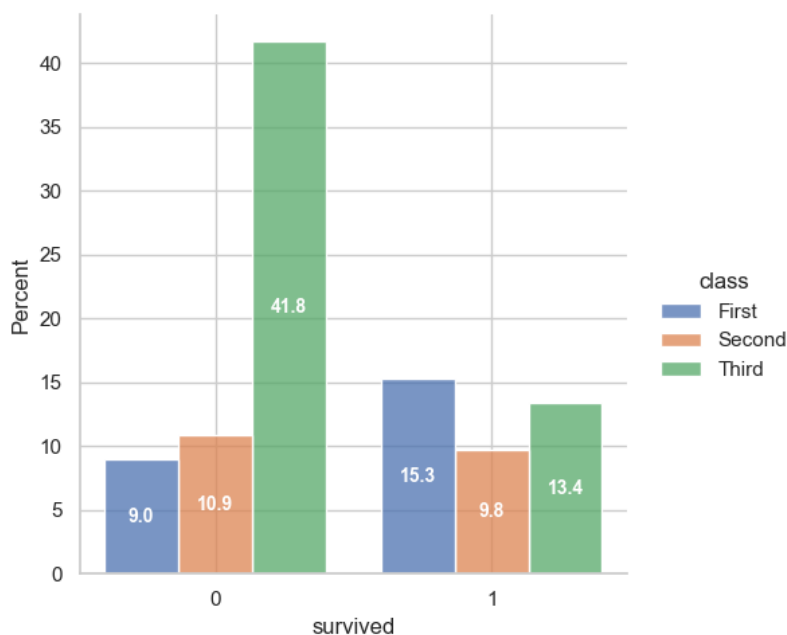
Εικόνα 35 Κατανομή πιθανότητας επιβίωσης στο ναυάγιο του Τιτανικού. Η οπτικοποίηση περιέχει τρία διαγράμματα ένα για κάθε θέση (μεταβλητή 'class')

Επειδή τώρα επιθυμούμε κανονικοποίηση/άθροιση, όχι στο σύνολο των διαθέσιμων παρατηρήσεων, αλλά στις παρατηρήσεις στην επιμέρους ομάδα κάθε θέσης, ώστε να πάρουμε διακριτές κατανομές της μεταβλητής 'survived', θέτουμε `common_norm=False`.

Κάθε κατανομή πιθανότητας πρέπει να έχει άθροισμα πιθανοτήτων ίσο με 100%, αν εκφράζεται σαν ποσοστό ή 1, αν εκφράζεται σαν αναλογία, επομένως αναμένουμε το άθροισμα των μηκών των ράβδων μέσα στο κάθε διάγραμμα (στην κάθε κατανομή) να είναι 100 ή 1, αντίστοιχα.

Μπορούμε να πάρουμε την ίδια πληροφορία όχι σε απεικόνιση με διαφορετικό διάγραμμα για κάθε θέση, αλλά σε ένα διάγραμμα, με διαφορετική απόχρωση για κάθε θέση (με ρύθμιση του ορίσματος `hue='class'`, αντί του ορίσματος `col='class'`).

Στην ακόλουθη απεικόνιση έχουμε ένα διάγραμμα με την πιθανότητα (ποσοστό) επιβίωσης (1) και αποβίωσης (0) (μεταβλητή `'survived'`) για τους επιβάτες κάθε θέσης, 1ης, 2ης και 3ης, όπου κάθε θέση (μεταβλητή `'class'`) συμβολίζεται με ξεχωριστό χρώμα (`hue='class'`).

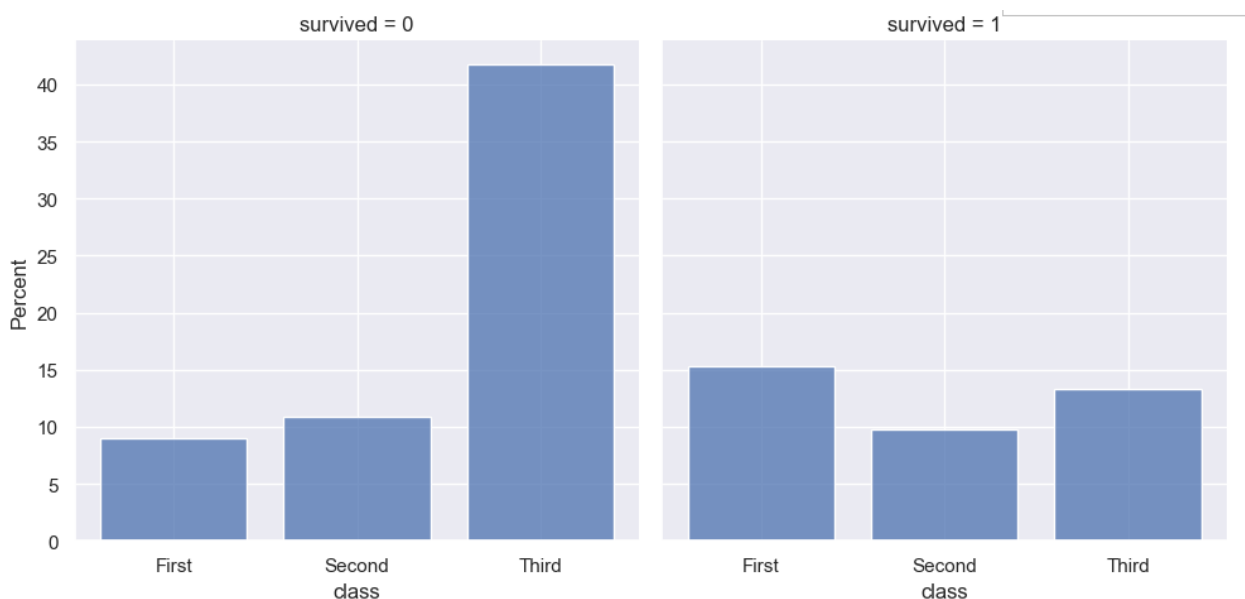


Εικόνα 36 Πιθανότητα επιβίωσης (1) και αποβίωσης (0) για τους επιβάτες κάθε θέσης στο ναυάγιο του Τιτανικού

Στο διάγραμμα έχουμε μία κατανομή όλων των μεταβλητών με μήκη ράβδων ίδια με τα μήκη των αντίστοιχων ράβδων στην πρώτη απεικόνιση με τα τρία διαγράμματα (ένα για κάθε θέση), αυτή που οι ράβδοι απαρτίζουν μία, ενιαία κατανομή.

Επιμέρους ομάδες

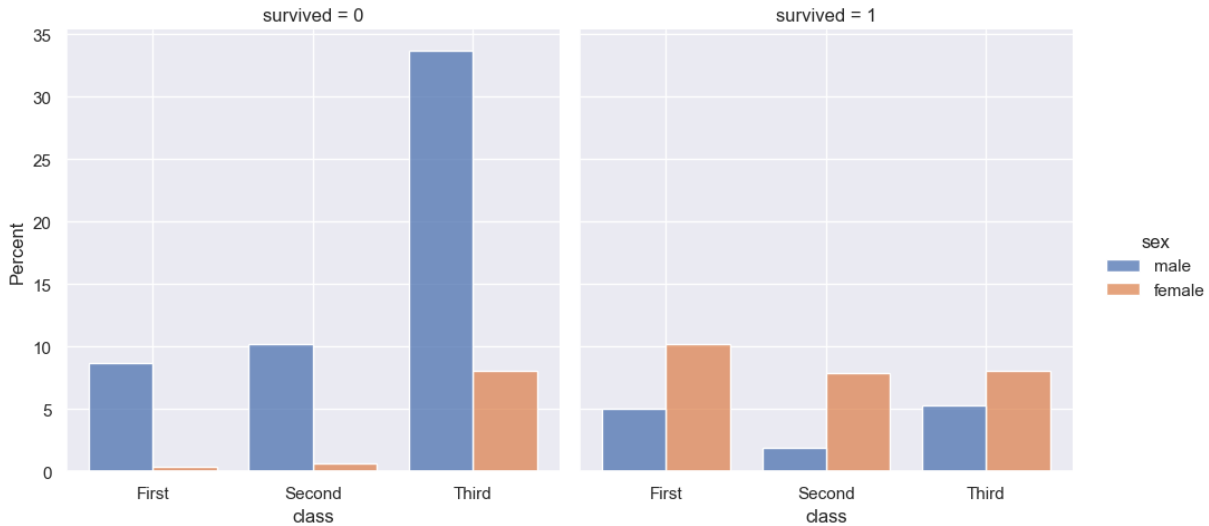
Μπορούμε να εναλλάξουμε τις μεταβλητές μεταξύ των ορισμάτων `x` και `hue` ή `x` και `col`, ώστε να έχουμε `x='class'` και `col='survived'` ή `x='class'` και `hue='survived'`. Συνεχίσουμε να παίρνουμε την ίδια πληροφορία για την ενιαία κατανομή των δύο μεταβλητών (αν `common_norm=True`) ή τις δύο υπό συνθήκη κατανομές της μεταβλητής `'class'` που ορίζουν οι επιμέρους ομάδες για τα επίπεδα της μεταβλητής `'survived'` (αν `common_norm=False`) (Εικ. 37).



Εικόνα 37 Αλλαγή των επιμέρους ομάδων με αλλαγή των σημασιολογικών ρυθμίσεων

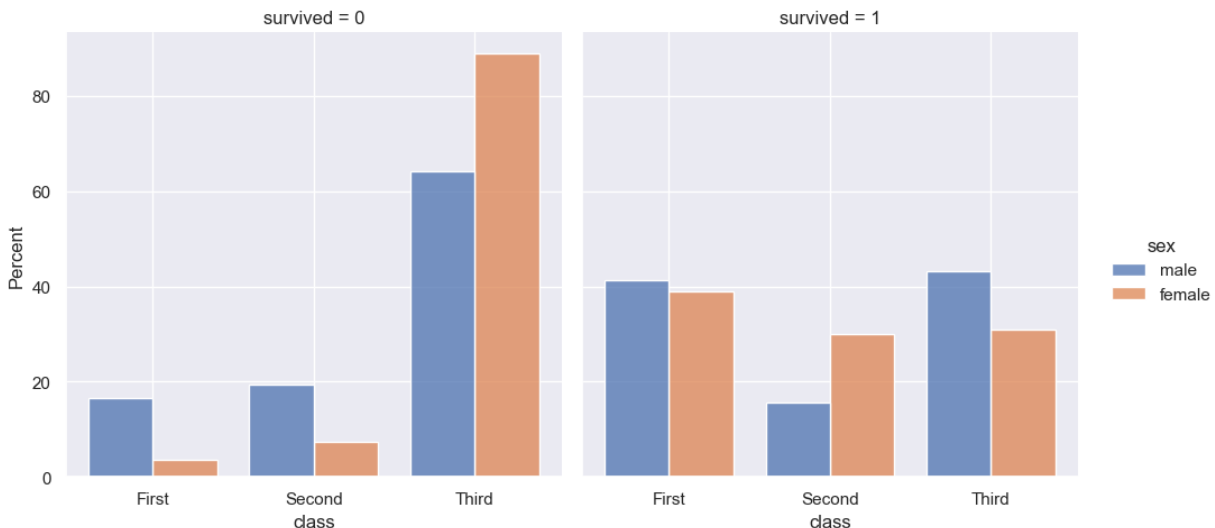
Μπορούμε, ακόμη όπως φαίνεται στην Εικόνα 38, να χρησιμοποιήσουμε τρεις κατηγορικές μεταβλητές, μία στον οριζόντιο άξονα, `x='class'` και δύο για να ορίσουμε 4 επιμέρους ομάδες, `col='survived'` (δύο διακριτά διαγράμματα) και `hue='sex'`, (δύο διαφορετικά χρώματα). Αν θέσουμε, `common_norm = True`, τότε παίρνουμε ενιαία

συνδυαστική κατανομή όλων των μεταβλητών με 12 εισόδους (12 συνδυασμούς κατηγοριών) και αντίστοιχες ράβδους με συνολικό άθροισμα υψών 100 (ή 1).



Εικόνα 38 Δημιουργία επιμέρους ομάδων σε διαφορετικά διαγράμματα και με διαφορετικά χρώματα

Αν, όμως, θέσουμε, `common_norm=False`, τότε παίρνουμε 4 υπό συνθήκη κατανομές της μεταβλητής 'class' για τις 4 επιμέρους ομάδες που εισάγουν οι ρυθμίσεις `col='survived'` (δύο διακριτά διαγράμματα) και `hue='sex'`, (δύο διαφορετικά χρώματα) (Εικ. 40).



Εικόνα 38β Διαφορετικές κατανομές για διαφορετικά διαγράμματα και με διαφορετικά χρώματα

Κατανομές ποσοτικών μεταβλητών

Απεικονίζουμε την κατανομή πιθανότητας μίας ή ζεύγους ποσοτικών μεταβλητών με ιστογράμμο (hist plot) ή διάγραμμα εκτίμησης πυκνότητας πιθανότητας (kernel density estimate or kde plot).

Για την απεικόνιση της κατανομής μίας ποσοτικής μεταβλητής υπάρχει και τρίτο είδος γραφήματος, το διάγραμμα σωρευτικής πυκνότητας (empirical cumulative density function or ecdf plot). Αυτός ο τύπος διαγράμματος δε μπορεί να οπτικοποιηθεί για συνδυαστικές κατανομές δύο ποσοτικών μεταβλητών.

Όλες οι απεικονίσεις της κατανομής μίας ή δύο ποσοτικών μεταβλητών αποτελούν εκτιμήσεις (προσεγγίσεις) της κατανομής με βάση τα διαθέσιμα δεδομένα. Επομένως, οι απεικονίσεις κατανομών είναι εμπειρικές, καθώς προκύπτουν από δεδομένα.

Ιστογράμματα και διαγράμματα εκτίμησης πυκνότητας

Κώδικας. Παίρνουμε απεικονίσεις κατανομών πιθανότητας μίας ή ζεύγους ποσοτικών μεταβλητών με ιστογράμματα, διαγράμματα εκτίμησης πυκνότητας ή διαγράμματα σωρευτικής πυκνότητας εκτελώντας τη figure-level συνάρτηση σχεδίασης `displot` της βιβλιοθήκης Seaborn, με `kind='hist'`, `kind='kde'`, ή `kind='ecdf'`, αντίστοιχα. Στο ακόλουθο κομμάτι κώδικα βλέπουμε τις βασικές ρυθμίσεις.

Μπορείτε να βρείτε το συγκεκριμένο κώδικα (script) (Εικ. 39) και να δοκιμάσετε παραλλαγές εκτέλεσής του στο σημειωματάριο του οδηγού μελέτης, “Data Visualization with Seaborn.ipynb”. Θυμίζουμε ότι ένα σημειωματάριο (όπως είναι ο οδηγός μελέτης) το εκτελούμε μέσα σε περιβάλλον Jupyter Notebook.

```

displot(
    data=penguins,          # το σύνολο δεδομένων (υπό μορφή pandas dataframe) από τις στήλες του οποίου αντλούνται οι μεταβλητές της οπτικοποίησης
    x="body_mass_g",        # η ποσοτική μεταβλητή στον άξονα x, για κατακόρυφο προσανατολισμό των ράβδων (ή, σπανιότερα, στον άξονα y)
    #x="flipper_length_mm", # εναλλακτική περίπτωση ποσοτικής μεταβλητής
    hue='species',          # για χρωματική διαφοροποίηση των σημείων/γραφικών, ένα χρώμα για κάθε κατηγορία/επίπεδο της επιλεγμένης μεταβλητής - όπο
    #col='sex',             # (μόνο για figure-level) για διακριτά διαγράμματα στις στήλες του πλέγματος, ένα για κάθε κατηγορία/επίπεδο της επιλεγμ
    #row='island',          # (μόνο για figure-level) για διακριτά διαγράμματα στις γραμμές του πλέγματος, ένα για κάθε κατηγορία/επίπεδο της επιλεγμ
    stat='percent',         # stat: το στατιστικό που θέλουμε να μετράται με το ύψος ή το εμβαδό των ράβδων (default: 'count')
    #stat='proportion',    # percent: bar heights normalize to 100%, 'probability'/'proportion': bar heights normalize to 1
    #stat='density',       # density: bar area normalize to 1
    bins=20,               # πλήθος διακριτών περιοχών (διαστημάτων) τιμών στις οποίες ζητάμε να χωριστεί το συνεχές φάσμα τιμών της ποσοτικής μεταβ
    #binwidth=100,         # εναλλακτικά αντί για πλήθος περιοχών ζητάμε η κατάτμηση του φάσματος τιμών ορίζοντας το εύρος της κάθε περιοχής
    #kde=True,             # προσθήκη καμπύλης πυκνότητας πιθανότητας στο ιστόγραμμα
    rug=True,              # προσθήκη πυκνωμάτων και αραιωμάτων στον άξονα της ποσοτικής μεταβλητής
    multiple='stack',      # γραφικά σε στρώσεις ('layer', default), ή ξεχωριστά ('dodge'), ή στοιβαγμένα ('stack')
    #common_norm=False,    # κανονικοποίηση σε όλο το σύνολο (common_norm=True, default) ή στις επιμέρους (σημασιολογικές) ομάδες (common_norm=False)
    #kind='hist',          # η ρύθμιση μπορεί να παραληφθεί, γιατί είναι η default για τη συνάρτηση displot
)

```

Εικόνα 39 Κλήση της συνάρτησης σχεδίασης κατανομών (displot)

Ανάλογα με το αποτέλεσμα που επιθυμούμε κάθε φορά, κάνουμε στις ρυθμίσεις τις κατάλληλες τροποποιήσεις. Αν σχεδιάζουμε απεικόνιση με ιστογράμματα, για παράδειγμα, τότε παραλείπουμε τη ρύθμιση `kind='hist'`, γιατί το ιστόγραμμα είναι ο default τύπος διαγράμματος για τη `displot`.

Συναρτήσεις σχεδίασης. Σαν figure-level συνάρτηση σχεδίασης, η `displot` έχει τη δυνατότητα για σχεδίαση οπτικοποίησης με πολλά διαγράμματα, που μπορεί να είναι είτε ιστογράμματα, αν η ρύθμιση είναι `kind='hist'` (default μπορεί να αγνοηθεί), είτε διαγράμματα εκτίμησης πυκνότητας αν η ρύθμιση είναι `kind='kde'`, είτε διαγράμματα σωρευτικής πυκνότητας αν η ρύθμιση είναι `kind='ecdf'`. Εναλλακτικά, αν θέλουμε ένα μόνο διάγραμμα προκειμένου να το ενσωματώσουμε σε υπάρχουσα οπτικοποίηση, μπορούμε να χρησιμοποιήσουμε τις αντίστοιχες axes-level συνάρτηση σχεδίασης `histplot`, `kdeplot` ή `ecdfplot`.

Σύνολα δεδομένων. Ένα σύνολο δεδομένων που χρησιμοποιούμε στη μελέτη των κατανομών πιθανότητας ποσοτικών μεταβλητών αφορά πιγκουίνους. Στο συγκεκριμένο σύνολο δεδομένων καταγράφονται σωματομετρικά χαρακτηριστικά των πιγκουίνων, όπως το βάρος του σώματος ή διάφορες διαστάσεις στο σώμα τους, σε συνδυασμό με κατηγορικές μεταβλητές, όπως το είδος και το φύλο του κάθε πουλιού. Το σύνολο δεδομένων το

δηλώνουμε στο όρισμα `data`, `data=penguins`. Φυσικά, μπορείτε να δοκιμάσετε σύνολα δεδομένων που σχετίζονται με δικά σας προβλήματα και μεταβλητές.

Να θυμίσουμε ότι ένα σύνολο δεδομένων φορτώνεται σε ένα `pandas dataframe` και δηλώνεται στο όρισμα `data` με το όνομα που έχουμε δώσει στο `dataframe` (για παράδειγμα, `data=penguins`). Στο `pandas dataframe` έχουμε ένα πίνακα δεδομένων με μία υποψήφια μεταβλητή σε κάθε στήλη και μία παρατήρηση σε κάθε γραμμή. Τις στήλες του συνόλου δεδομένων τις χρησιμοποιούμε στις συναρτήσεις σχεδίασης της βιβλιοθήκης Seaborn για να δηλώσουμε μεταβλητές στους άξονες (ορίσματα `x`, `y`) και στις σημασιολογικές ρυθμίσεις (ορίσματα `hue`, `col` και `row`).

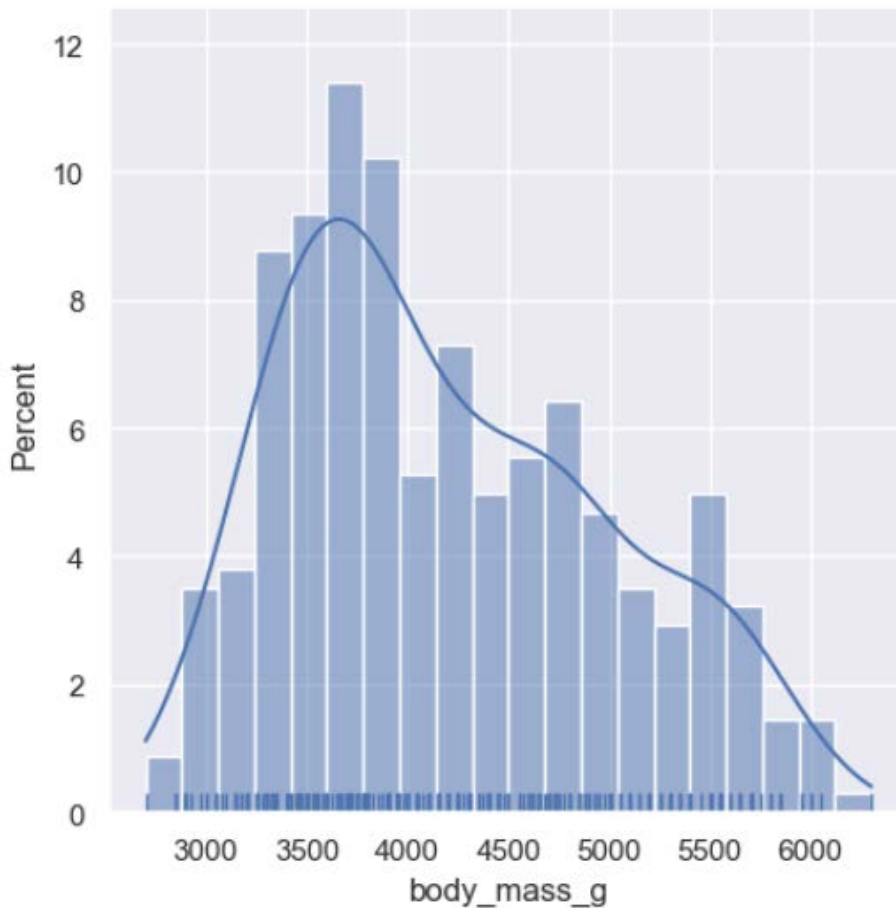
Ρυθμίσεις ιστογράμματος ποσοτικών μεταβλητών

Μία ποσοτική μεταβλητή. Το ιστόγραμμα είναι η συχνότερα χρησιμοποιούμενη απεικόνιση της κατανομής πιθανότητας μίας ποσοτικής μεταβλητής.

Με το ακόλουθο ιστόγραμμα απεικονίζουμε την κατανομή της ποσοτικής μεταβλητής `"body_mass_g"` (βάρος σώματος σε γραμμάρια) από το σύνολο δεδομένων με στοιχεία πιγκουίνων (Εικ. 40).

Περιοχές τιμών (bins). Στο ιστόγραμμα μίας ποσοτικής μεταβλητής χωρίζουμε το συνεχές εύρος τιμών της μεταβλητής σε διακριτές περιοχές (διαστήματα) τιμών, επιχειρώντας **διακριτοποίηση (discretization ή binning)** του συνεχούς φάσματος τιμών της ποσοτικής μεταβλητής. Στο παρακάτω ιστόγραμμα το συνεχές φάσμα τιμών της ποσοτικής μεταβλητής εκτείνεται στον άξονα `x`, όπου έχουμε προσθέσει χαραγές με τις

πρωτογενείς τιμές (παρατηρήσεις) της μεταβλητής, ρυθμίζοντας το όρισμα `rug`, `rug=True` (Εικ. 40).



Εικόνα 40 Ιστόγραμμα όπου απεικονίζουμε την κατανομή βάρους σώματος σε γραμμάρια από σύνολο δεδομένων με στοιχεία πιγκουίνων

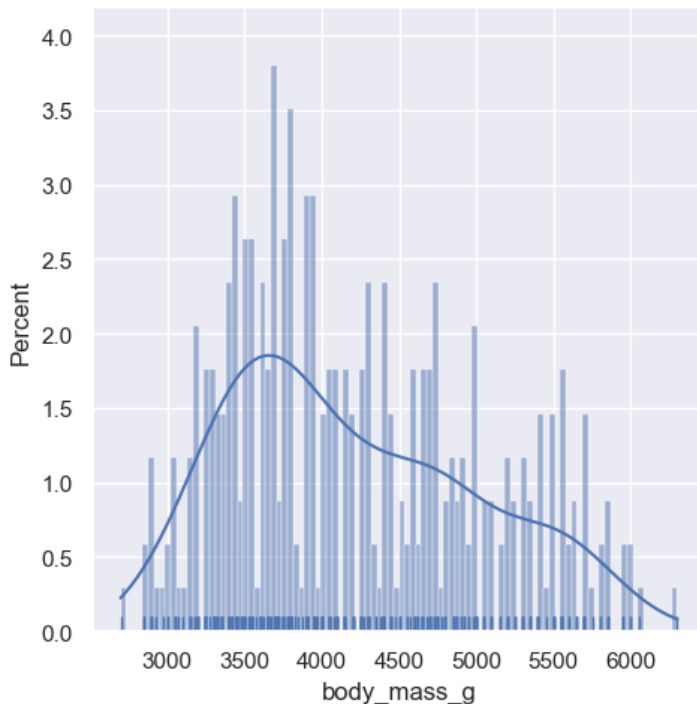
Οι περιοχές είναι κάτι αντίστοιχο των κατηγοριών της κατηγορικής μεταβλητής. Αντί του ορίσματος `discrete` που ρυθμίζουμε στα ιστογράμματα κατηγορικής μεταβλητής, στην περίπτωση ποσοτικών μεταβλητών ρυθμίζουμε το όρισμα `bins`, με το οποίο δηλώνουμε το επιθυμητό αριθμό των περιοχών. Εναλλακτικά, ρυθμίζουμε το όρισμα `binwidth`, για να δηλώσουμε το κοινό εύρος των περιοχών (αν υπάρχουν και τα δύο ορίσματα το όρισμα `binwidth` έχει προτεραιότητα έναντι του ορίσματος `bins`).

Αντιστοιχίζουμε σε κάθε περιοχή τιμών της ποσοτικής μεταβλητής μία ράβδο (ιστό), το ύψος/μήκος (ή το εμβαδό) της οποίας εκφράζει την πυκνότητα πιθανότητας στην περιοχή, ή απλά την πιθανότητα μία νέα παρατήρηση να πέσει στην περιοχή.

Θα πρέπει να τονιστεί ότι όποια αλλαγή του αριθμού (ή του εύρους) των ράβδων προκαλεί αλλαγή στο ύψος/μήκος τους και στο γενικότερο σχήμα της κατανομής. Στα ιστογράμματα που βλέπετε εδώ, για παράδειγμα, έχουμε κάνει την επιλογή, `bins=20` (ή την παραπλήσια επιλογή, `binwidth=200`).

Η επιλογή του αριθμού (ή, εναλλακτικά, του πλάτους) των περιοχών είναι πολύ σημαντική για τη σωστή απεικόνιση του σχήματος της κατανομής (κορυφές, διασπορά, συμμετρία/λοξότητα, ακραίες τιμές), το οποίο μας δίνει, επιπλέον, μία ένδειξη για χαρακτηριστικά στατιστικά μεγέθη της κατανομής και του πληθυσμού (κεντρική τάση, διακύμανση, αβεβαιότητα ή περιθώριο σφάλματος). Μεγαλύτερο πλήθος περιοχών (ή, εναλλακτικά, μικρότερο πλάτος περιοχών) γεμίζει την οπτικοποίηση με θόρυβο, που οφείλεται σε τυχαίους παράγοντες και θολώνει τη γενική εικόνα της κατανομής. Αντίθετα, μικρότερο πλήθος περιοχών (ή, εναλλακτικά, μεγαλύτερο πλάτος περιοχών) προκαλεί απώλεια ουσιαστικής πληροφορίας (χαρακτηριστικών) της κατανομής.

Στο ακόλουθο ιστόγραμμα (Εικ. 41) έχουμε εσκεμμένα επιλέξει πολύ μεγάλο πλήθος περιοχών (ή πολύ μικρό εύρος περιοχών). Διαπιστώνουμε την ύπαρξη θορύβου που αλλοιώνει το σχήμα και δυσκολεύει την εκτίμηση της διακύμανσης της κατανομής.



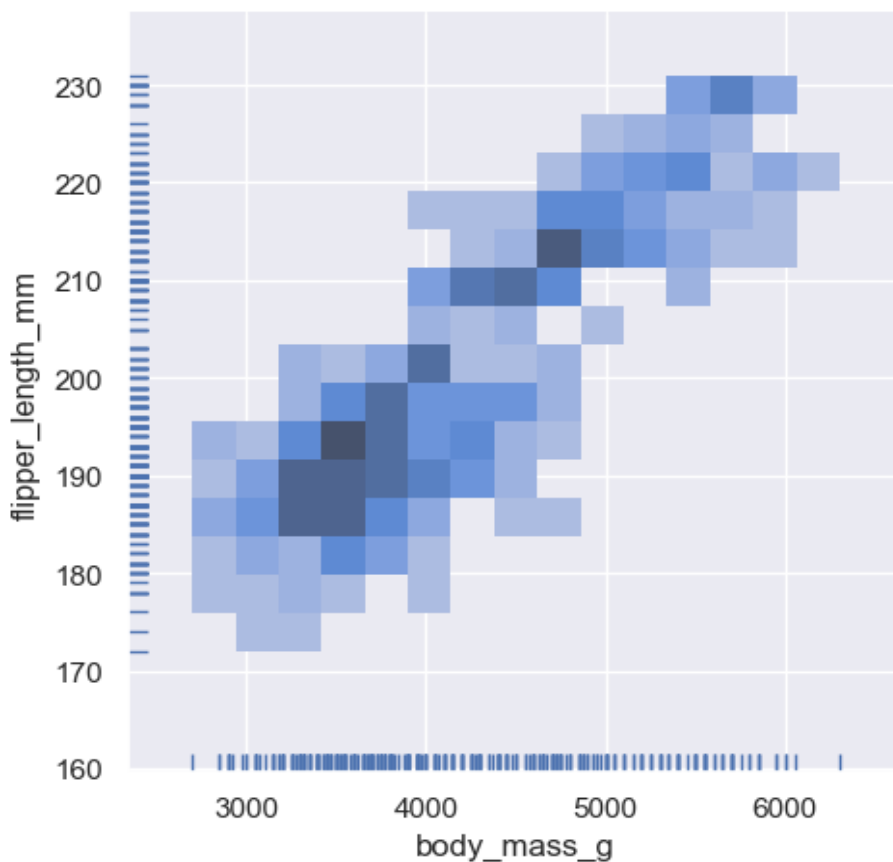
Εικόνα 41 Κατανομή βάρους σώματος των πιγκουίνων με $binwidth=200$ που οδηγεί σε οπτικοποίηση με «θόρυβο»

Στο ιστόγραμμα του βάρους σώματος έχουμε επιπλέον, ζητήσει μία εξομαλυμένη καμπύλη πυκνότητας πιθανότητας, με τη ρύθμιση $kde=True$. Το επικρατέστερο βάρος (ψηλότερη ράβδος) είναι γύρω στα 3700 με πιθανότητα (ύψος ράβδου=ποσοστό) λίγο κάτω από 12%. Σε αυτή τη ράβδο και στις γειτονικές της διαπιστώνουμε μεγαλύτερη πυκνότητα παρατηρήσεων, όπως φαίνεται από τις χαραγές πάνω στον άξονα της ποσοτικής μεταβλητής (οι χαραγές εμφανίζονται χάρη στη ρύθμιση $rug=True$). Παρατηρούμε μία δεύτερη χαμηλότερη κορυφή (δεύτερη επικρατούσα τιμή) γύρω στα 4700, που, και πάλι, επαληθεύεται από την πυκνότητα των χαραγών στην περιοχή της ράβδου. Αυτή η δεύτερη κορυφή είναι υπεύθυνη και για την ασυμμετρία και τη λοξότητα προς τα δεξιά της κατανομής, γεγονός που αντικατοπτρίζεται και στην εκτιμώμενη καμπύλη πυκνότητας πιθανότητας. Γενικά παρατηρούμε ότι, για κάθε ράβδο, η πυκνότητα πιθανότητας, όπως

εκφράζεται με το ύψος/μήκος της ράβδου, συμφωνεί με την πυκνότητα των παρατηρήσεων, όπως φαίνεται από τις χαραγές πάνω στον άξονα της ποσοτικής μεταβλητής (Εικ. 41)..

Ζεύγος ποσοτικών μεταβλητών. Στην περίπτωση δύο ποσοτικών μεταβλητών οι περιοχές είναι διδιάστατες, συνήθως τετράγωνα. Αντί να “εκτείνουμε” ράβδους στην τρίτη διάσταση και να μετρούμε/συγκρίνουμε την πιθανότητα με βάση το ύψος κάθε ράβδου υπό προοπτική, χρωματίζουμε τις περιοχές στις δύο διαστάσεις (με ίδιο χρώμα) και μετρούμε/συγκρίνουμε την πιθανότητα με βάση την ένταση του χρώματος σε κάθε περιοχή.

Το ακόλουθο ιστόγραμμα απεικονίζει τη συνδυαστική κατανομή δύο ποσοτικών μεταβλητών. Οι δύο μεταβλητές είναι σωματομετρικά χαρακτηριστικά των πιγκουίνων, συγκεκριμένα το βάρος σώματος και το μήκος πτερυγίου.



Εικόνα 42 Συνδυαστική κατανομή δύο ποσοτικών μεταβλητών

Διακρίνουμε δύο σχηματισμούς (δυνητικές κατανομές) με αντίστοιχες κορυφές, μία κάτω αριστερά στο ιστόγραμμα και μία δεύτερη πάνω δεξιά. Οι συντεταγμένες της περιοχής όπου εκτείνεται, για παράδειγμα, ο κάτω αριστερά σχηματισμός είναι: βάρος σώματος από 3000+ έως περίπου 4000 και μήκος πτερυγίου από 180+ έως περίπου 200.

Στατιστικό ιστογράμματος. Στα ιστογράμματα μίας ή δύο ποσοτικών μεταβλητών η πιθανότητα κάθε περιοχής εκφράζεται, συνήθως, σαν ποσοστό ή αναλογία του πλήθους των παρατηρήσεων της περιοχής επί του συνόλου των παρατηρήσεων (που συμμετέχουν στον υπολογισμό της κατανομής). Αυτό γίνεται με ρύθμιση του ορίσματος `stat`, που σημαίνει στατιστικό (statistic). Θέτουμε `stat='percent'` ή `stat='proportion'`.

Αν η πιθανότητα εκφράζεται σαν ποσοστό, τότε το άθροισμα των μηκών των ράβδων (ή, στην περίπτωση δύο ποσοτικών μεταβλητών, της χρωματικής έντασης των περιοχών) στην κατανομή είναι 100, αν εκφράζεται σαν αναλογία τότε το άθροισμα των μηκών των ράβδων (ή της χρωματικής έντασης των περιοχών) στην κατανομή είναι 1.

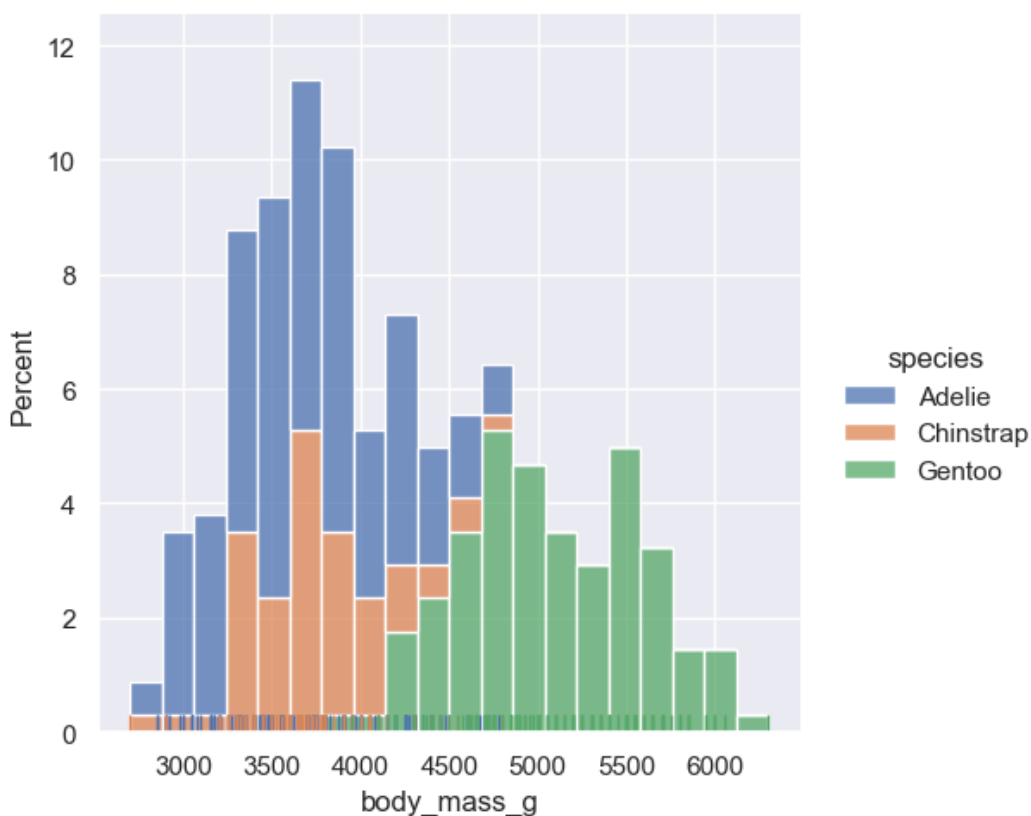
Επιμέρους ομάδες

Μπορούμε να εκμεταλλευτούμε τη δυνατότητα για σημασιολογικές ρυθμίσεις των συναρτήσεων σχεδίασης της βιβλιοθήκης Seaborn και ειδικότερα της figure-level συνάρτησης `displot`, που ειδικεύεται στο σχεδιασμό ιστογραμμάτων και διαγραμμάτων εκτίμησης πυκνότητας πιθανότητας, ώστε να διακρίνουμε και να συγκρίνουμε επιμέρους ομάδες στα δεδομένα μας.

Συγκεκριμένα, μπορούμε να αξιοποιήσουμε τις σημασιολογικές ρυθμίσεις με τα ορίσματα `hue`, `col` και `row` για να ξεχωρίσουμε με χρώμα τα τρία είδη πιγκουίνων (`hue="species"`), με

δύο διακριτά διαγράμματα τα αρσενικά από τα θηλυκά πουλιά (`col='sex'`) και με περισσότερα, ακόμη, διαγράμματα το νησί όπου ζει το κάθε είδος (`row='island'`).

Στην ακόλουθη απεικόνιση παίρνουμε σε ένα ιστόγραμμα του βάρους σώματος μία ενιαία κατανομή, όπου βλέπουμε με διαφορετικό χρώμα τη συμμετοχή κάθε είδους στο μήκος της ράβδου κάθε περιοχής (Εικ. 43).. Σε κάθε περιοχή υψώνεται, επομένως, μία πολυχρωματική ράβδος με τα κομμάτια διαφορετικού χρώματος στοιβαγμένα, το ένα πάνω στο άλλο. Η στοίβα επιτυγχάνεται με τη ρύθμιση `multiple='stack'`.

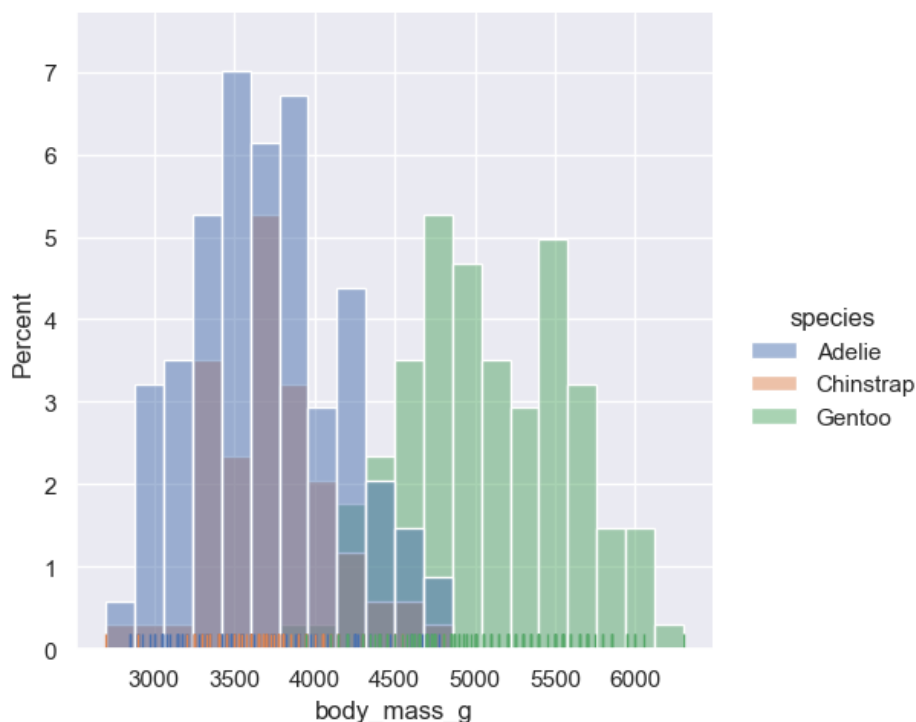


Εικόνα 43 Ιστόγραμμα κατανομής του βάρους των πιγκουίνων με ράβδους διαφορετικών χρωμάτων που είναι τοποθετημένες η μία πάνω από την άλλη

Κάθε πολύχρωμη ράβδος στο παραπάνω ιστόγραμμα έχει το ίδιο ύψος με την αντίστοιχη μονοχρωματική ράβδο στο αντίστοιχο μονοχρωματικό ιστόγραμμα του βάρους σώματος. Για παράδειγμα, η ψηλότερη ράβδος έχει και στα δύο ιστογράμματα ύψος κάτω από 12% και η

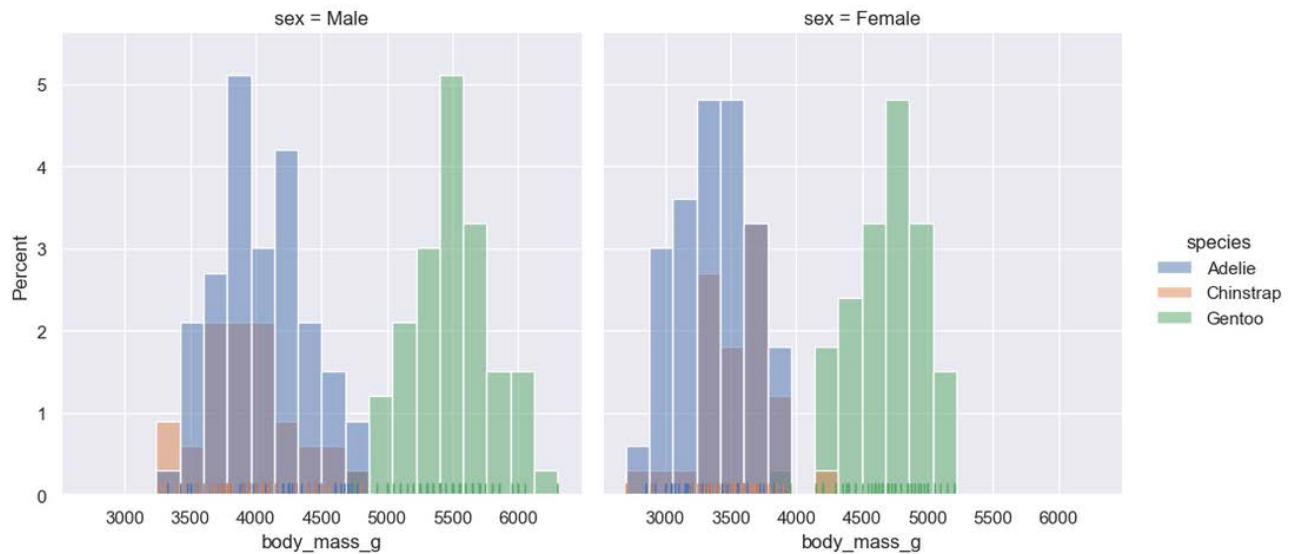
δεύτερη ψηλότερη έχει ύψος λίγο πάνω από 10%. Επιλέγουμε κοινή κανονικοποίηση ορίζοντας `common_norm=True`, επομένως έχουμε μία κατανομή με το άθροισμα των μηκών όλων των (πολύχρωμων) ράβδων να είναι 100, στην περίπτωση που η πιθανότητα εκφράζεται σαν ποσοστό ή 1, στην περίπτωση που εκφράζεται σαν αναλογία.

Στην ακόλουθη απεικόνιση έχουμε ιστόγραμμα μίας κατανομής, μόνο που βλέπουμε τη συμμετοχή κάθε χρώματος με ράβδους που είναι τοποθετημένες η μία πίσω από την άλλη (`multiple='layer', default`) (Εικ. 44)..



Εικόνα 44 Ιστόγραμμα κατανομής του βάρους των πιγκουίνων με ράβδους διαφορετικών χρωμάτων που είναι τοποθετημένες η μία πίσω από την άλλη

Αν, εκτός από τη ρύθμιση του χρώματος, `hue="species"` των ειδών των πιγκουίνων, προσθέσουμε τη ρύθμιση `col='sex'`, τότε παίρνουμε την ακόλουθη απεικόνιση (Εικ. 45). με δύο διακριτά διαγράμματα ένα για τα αρσενικά και δεύτερο για τα θηλυκά πουλιά (αν, επιπλέον, προσθέσουμε τη ρύθμιση `row='island'`, τότε παίρνουμε επιπλέον διαγράμματα, σε επιπλέον γραμμές, με το νησί όπου ζει το κάθε είδος).



Εικόνα 45 Ιστογράμματα κατανομής του βάρους των πιγκουίνων με ράβδους διαφορετικών χρωμάτων που είναι τοποθετημένες η μία πίσω από την άλλη για κάθε φύλο

Και στις δύο παραπάνω οπτικοποιήσεις χρησιμοποιούμε διαστρωμάτωση των ράβδων (με τη ρύθμιση `multiple='layer'`) και επιλέγουμε κοινή κανονικοποίηση (`common_norm=True`). Η ρύθμιση κοινής κανονικοποίησης έχει σαν αποτέλεσμα να πάρουμε μία ενιαία κατανομή, με το άθροισμα των μηκών όλων των ράβδων (για όλα τα χρώματα και σε όλα τα διαγράμματα) να είναι 100, στην περίπτωση που η πιθανότητα εκφράζεται σαν ποσοστό ή 1, στην περίπτωση που εκφράζεται σαν αναλογία.

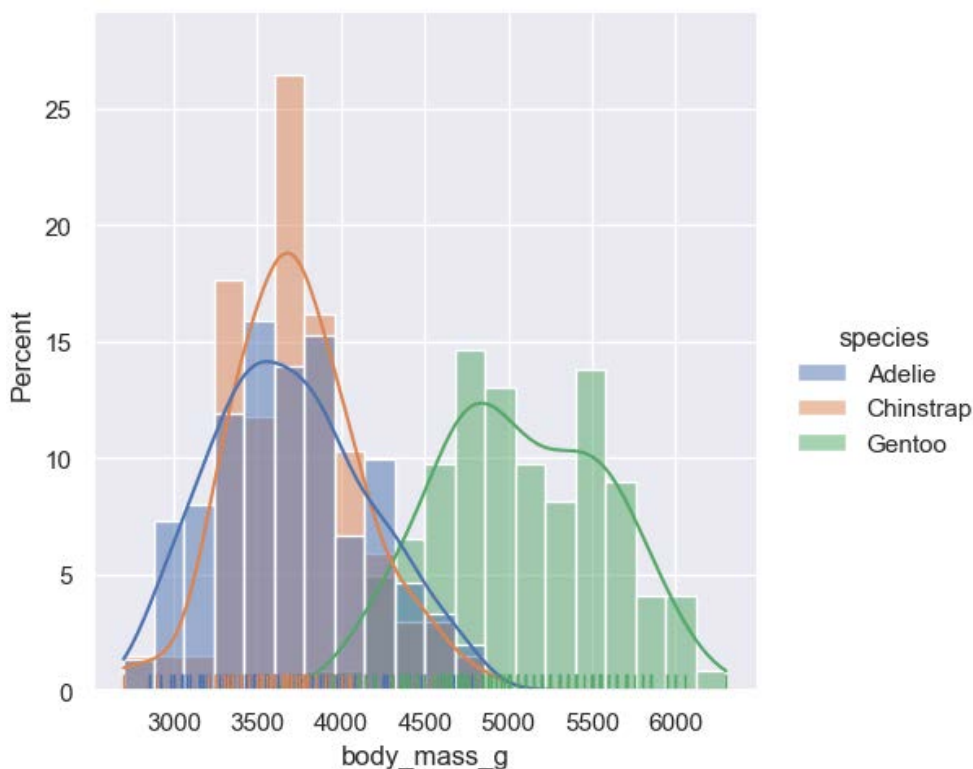
Η διαστρωμάτωση των ράβδων (`multiple='layer'`) μας επιτρέπει να διακρίνουμε το σχήμα των "επιμέρους ιστογραμμάτων" για το κάθε είδος πουλιού, δηλαδή, για το κάθε χρώμα (αναφερόμαστε σε "επιμέρους ιστογράμματα" και όχι κατανομές, γιατί η κατανομή είναι μία).

Υπό συνθήκη κατανομές

Διατηρώντας τη ρύθμιση του χρώματος, `hue="species"`, για τα είδη των πιγκουίνων, μπορούμε να πάρουμε τρεις διαφορετικές κατανομές πιθανότητας, μία για κάθε είδος πιγκουίνου (υπό συνθήκη κατανομές, δεδομένου του είδους), αρκεί να ακυρώσουμε την

κοινή κανονικοποίηση (δηλαδή, να θέσουμε `common_norm=False`). Χρησιμοποιούμε διαστρωμάτωση των ράβδων (`multiple='layer'`), που μας επιτρέπει εύκολα να διακρίνουμε το σχήμα της κάθε κατανομής.

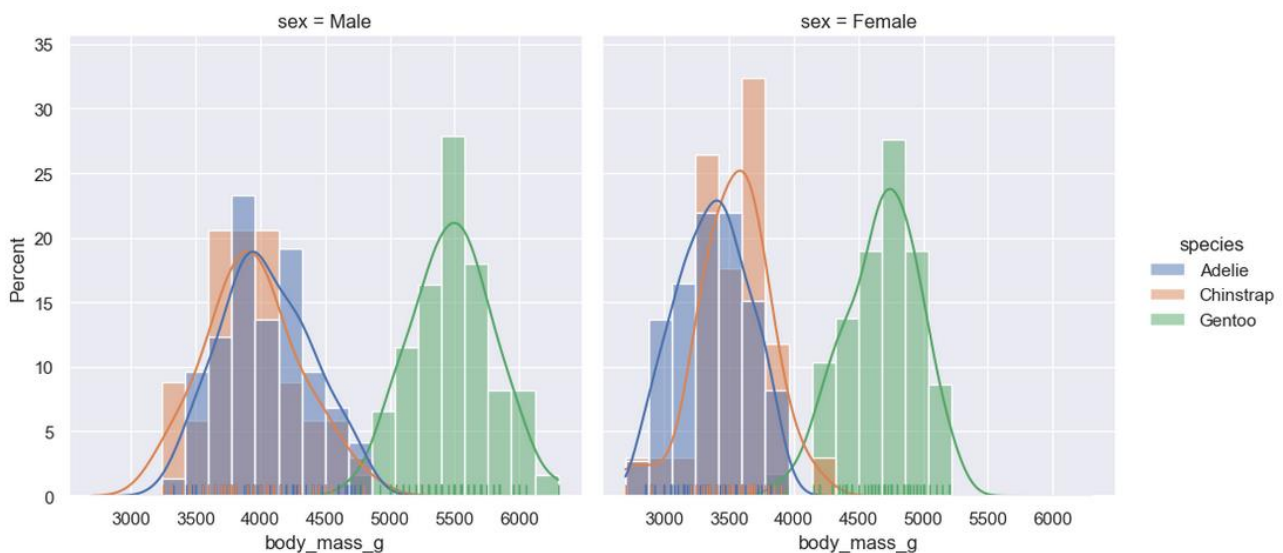
Στο ακόλουθο ιστόγραμμα (Εικ. 46) εικονίζονται τρεις κατανομές (τρία ιστογράμματα), μία για κάθε ένα από τα τρία διαφορετικά χρώματα που αντιπροσωπεύουν τα τρία είδη. Με τη ρύθμιση `kde=True` έχουμε προσθέσει στο ιστόγραμμα μία εξομαλυμένη καμπύλη πυκνότητας πιθανότητας για κάθε κατανομή.



Εικόνα 46 Ιστόγραμμα και καμπύλες κατανομής

Επειδή κάθε κατανομή πιθανότητας πρέπει να έχει άθροισμα πιθανοτήτων ίσο με 1 (ή με 100%), θα πρέπει το άθροισμα των μηκών των ράβδων στο κάθε χρώμα (στην κάθε κατανομή) να είναι 1 (αν επιλέγαμε να εκφράσουμε την πιθανότητα σαν αναλογία) ή 100 (αν επιλέγαμε να εκφράσουμε την πιθανότητα σαν ποσοστό).

Εφόσον διατηρούμε τη ρύθμιση του χρώματος, `hue="species"`, για τα είδη των πιγκουίνων, και, επιπλέον, εισάγουμε τη ρύθμιση για διακριτά διαγράμματα, `col='sex'`, για το φύλο των πουλιών και δεν ζητούμε κοινή κανονικοποίηση (`common_norm=False`), τότε μπορούμε να πάρουμε 6 διαφορετικές κατανομές πιθανότητας, μία για κάθε συνδυασμό είδους και φύλου πιγκουίνου (υπό συνθήκη κατανομές, δεδομένου του είδους και του φύλου) (Εικ. 47). Χρησιμοποιούμε, και πάλι, διαστρωμάτωση των ράβδων (με τη ρύθμιση `multiple='layer'`), που επιτρέπει την εύκολη αναγνώριση του σχήματος της κάθε κατανομής.



Εικόνα 47 Έξι καμπύλες κατανομών. Στη οπτικοποίηση έχουμε ξεχωριστό διάγραμμα για τη στατιστική μεταβλητή φύλο των πιγκουίνων μέσα στο οποίο διακρίνονται με διαφορετικό χρώμα τα τρία είδη των πιγκουίνων (2x3=6)

Κάθε μία από τις 6 κατανομές πιθανότητας (Εικ. 47) πρέπει να έχει άθροισμα πιθανοτήτων ίσο με 1 (ή με 100%), οπότε το άθροισμα των μηκών των ράβδων στην κάθε κατανομή (στο κάθε διάγραμμα και στο κάθε χρώμα μέσα στο διάγραμμα) θα πρέπει να είναι 1 (αν επιλέγουμε να εκφράσουμε την πιθανότητα σαν αναλογία) ή 100 (αν επιλέγουμε να εκφράσουμε την πιθανότητα σαν ποσοστό).

Η απεικόνιση υπό συνθήκη κατανομών, δυνητικά με διαφορετικά χρώματα και σε διακριτά διαγράμματα, επιτρέπει τη σύγκριση του σχήματος των κατανομών και

χαρακτηριστικών του σχήματος όπως οι κορυφές, η διασπορά, η συμμετρία/λοξότητα, οι ακραίες τιμές.

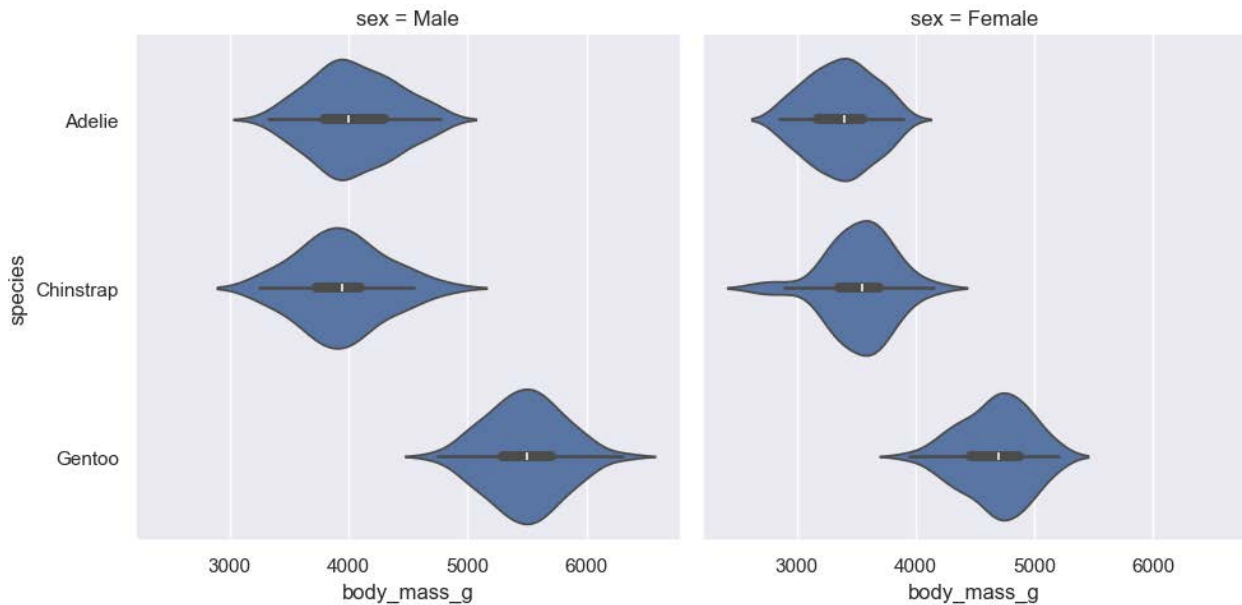
Στα ιστογράμματα που απεικονίζουν υπο συνθήκη κατανομές φροντίζουμε για την προσθήκη καμπυλών πυκνότητας πιθανότητας, μία για κάθε κατανομή, για να έχουμε, εκτός από το ιστόγραμμα και μία εξομαλυμένη εικόνα της κάθε κατανομής.

Συνδυαστική χρήση και ερμηνεία ιστογραμμάτων και κατηγορικών διαγραμμάτων

Μπορούμε να συνδυάσουμε και να συγκρίνουμε ιστογράμματα που εικονίζουν υπο συνθήκη κατανομές με κατηγορικά διαγράμματα κατανομών και κεντρικής τάσης για να συνδυάσουμε πληροφορίες για το σχήμα της κατανομής (κορυφές, διασπορά, συμμετρία/λοξότητα, ακραίες τιμές) με πληροφορίες για στατιστικά μεγέθη της κατανομής και του πληθυσμού (κεντρική τάση, διακύμανση, αβεβαιότητα ή περιθώριο σφάλματος).

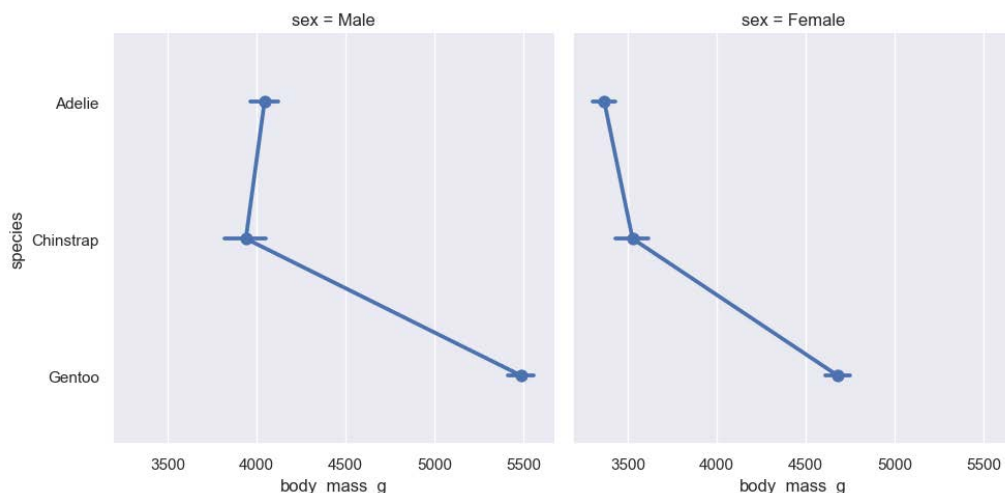
Αντλούμε πληροφορίες για το σχήμα των κατανομών από τα ιστογράμματα και πληροφορίες για στατιστικά μεγέθη των κατανομών και των αντίστοιχων πληθυσμών από τα κατηγορικά διαγράμματα κατανομών και κεντρικής τάσης, (Seaborn development team, χ.χ.).

Για παράδειγμα, αντιπαραβάλλουμε τα προηγούμενα ιστογράμματα όπου απεικονίζονται 6 κατανομές για τους συνδυασμούς είδους και φύλου πιγκουίνων με το ακόλουθο κατηγορικό διάγραμμα βιολιού, που επίσης έχει πληροφορία για τις συγκεκριμένες 6 κατανομές (Εικ. 48).



Εικόνα 48 Κατανομή, κεντρική τάση, διασπορά

Από το διάγραμμα βιολιού (Εικ. 48) παρατηρούμε, για παράδειγμα, λίγο μεγαλύτερη διάμεσο (στατιστικό κεντρικής τάσης) για το είδος Adelie από το είδος Chinstrap, στοιχείο που συνάδει με τη μικρή λοξότητα προς τα δεξιά της κατανομής του πρώτου είδους (διαπιστώνουμε τη λοξότητα και από το ιστόγραμμα αλλά και από το διάγραμμα βιολιού).



Εικόνα 49 Μέσοι όροι περιθώριο σφάλματος για κάθε είδους πιγκουίνου

Όπως φαίνεται από το διάγραμμα σημείου στην Εικ. 49. (απεικονίζει 6 μέσους όρους και σχετικά περιθώρια σφάλματος, ένα για κάθε κατανομή), ο μέσος όρος ακολουθεί το ίδιο μοτίβο διαφοροποίησης για τα δύο είδη, (Seaborn development team, χ.χ.).

Τα παραπάνω δύο κατηγορικά διαγράμματα τα πήραμε με τον ακόλουθο κώδικα (Εικ. 50).

```
# Εισαγωγή της συνάρτησης σχεδίασης
from seaborn import catplot

catplot(
    data=penguins,
    x='body_mass_g', y='species', # δύο μεταβλητές στους άξονες x, y, μία κατηγορική και μία ποσοτική
    #x='flipper_length_mm', y='species', # εναλλακτική περίπτωση ποσοτικής μεταβλητής, σχετιζόμενης με την κατηγορική
    #hue='island', # για χρωματική διαφοροποίηση των σημείων/γραφικών, ένα χρώμα για κάθε κατηγορία/επίπεδο της επιλεγμένης μεταβλητής
    col="sex", # (μόνο για figure-level) για διακριτά διαγράμματα στις στήλες του πλέγματος, ένα για κάθε κατηγορία/επίπεδο της μεταβλητής
    #split=True, # αντικρυστές ή διακριτές κατανομές με διαφορετικό χρώμα
    #inner="quart", # προστίθεται στο εσωτερικό του violinplot πληροφορία: 'box' (default) ή τεταρτημόρια ('quart') ή οι παρατηρήσεις
    #inner='point',
    kind="violin", # κατηγορικό διάγραμμα(τα) κατανομών
)
```

Εικόνα 50 Κλήση της συνάρτησης σχεδίασης κατηγορικών διαγραμμάτων

Για να πάρουμε κατηγορικά διαγράμματα κατανομών (όπως το διάγραμμα βιολιού) ή κεντρικής τάσης (όπως το διάγραμμα σημείου) χρησιμοποιούμε τη figure-level συνάρτηση σχεδίασης catplot της βιβλιοθήκης Seaborn με kind='violin' ή kind='point', αντίστοιχα, (Yi, M., χ.χ.).

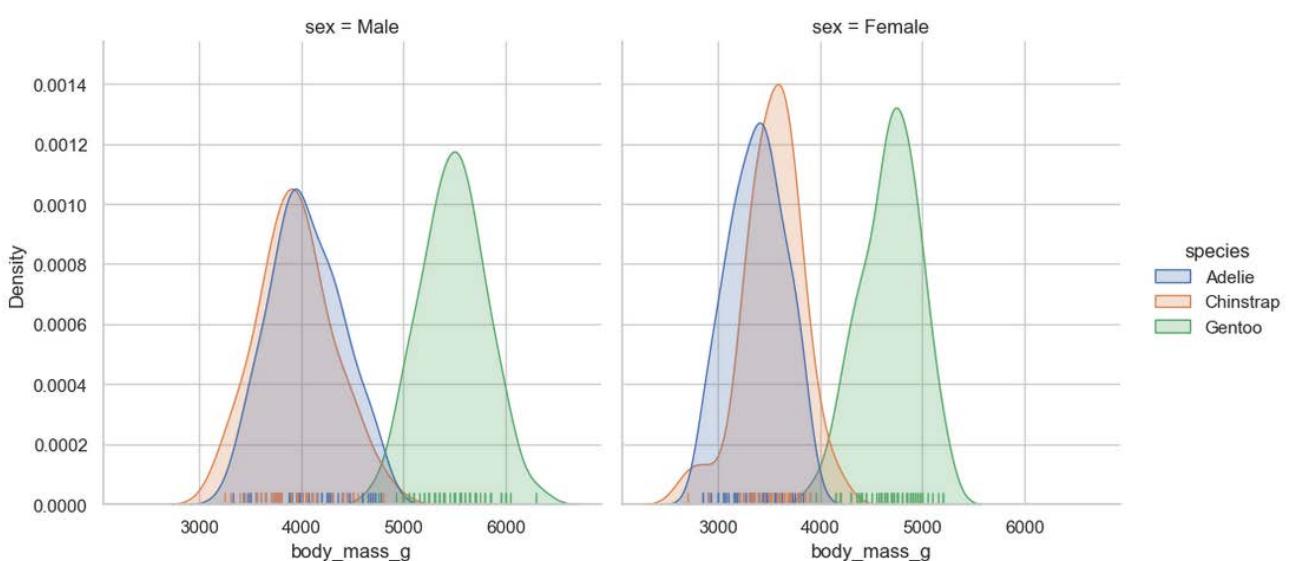
Μπορείτε να δοκιμάσετε παραλλαγές του συγκεκριμένου κώδικα (script) και να βρείτε περισσότερες λεπτομέρειες για τα κατηγορικά διαγράμματα και τη χρήση τους (είτε ανεξάρτητα είτε σε συνδυασμό με διαγράμματα κατανομών, όπως ιστογράμματα και διαγράμματα εκτίμησης πυκνότητας) στο σημειωματάριο του οδηγού μελέτης, “Data Visualization with Seaborn.ipynb”. Θυμίζουμε ότι ένα σημειωματάριο (όπως είναι ο οδηγός μελέτης) το εκτελούμε μέσα σε περιβάλλον Jupyter Notebook.

Διαγράμματα εκτίμησης πυκνότητας (πιθανότητας)

Στα ιστογράμματα συχνά προσθέτουμε καμπύλες πυκνότητας πιθανότητας, για να έχουμε, εκτός από το ιστόγραμμα, μία ομαλή εικόνα της κατανομής. Αν έχουμε μία απεικόνιση με πολλές (υπό συνθήκη) κατανομές (μία για κάθε επιμέρους ομάδα παρατηρήσεων), τότε έχουμε αντίστοιχες καμπύλες πυκνότητας.

Προσθέτουμε καμπύλες πυκνότητας στα ιστογράμματα με τη ρύθμιση `kde=True` στην κλήση της συνάρτησης σχεδίασης `displot`. Εναλλακτικά, μπορούμε να πάρουμε ανεξάρτητα διαγράμματα με καμπύλες πυκνότητας πιθανότητας (διαγράμματα εκτίμησης πυκνότητας) αν περάσουμε τη ρύθμιση `kind='kde'` στη συνάρτηση `displot`.

Αξιοποιώντας τον ίδιο κώδικα με αυτόν που χρησιμοποιήσαμε για σχεδίαση ιστογραμμάτων μπορούμε να πάρουμε απεικόνιση με διαγράμματα εκτίμησης πυκνότητας. Τα ακόλουθα δύο διαγράμματα εκτίμησης πυκνότητας περιέχουν τις ίδιες 6 κατανομές που πήραμε σε ιστογράμματα για τα είδη και το φύλο των πιγκουίνων (`hue="species", col='sex', common_norm=False, kind='kde'`) (Εικ. 51).



Εικόνα 51 Δύο διαγράμματα εκτίμησης πυκνότητας με 6 κατανομές για τα είδη και το φύλο των πιγκουίνων

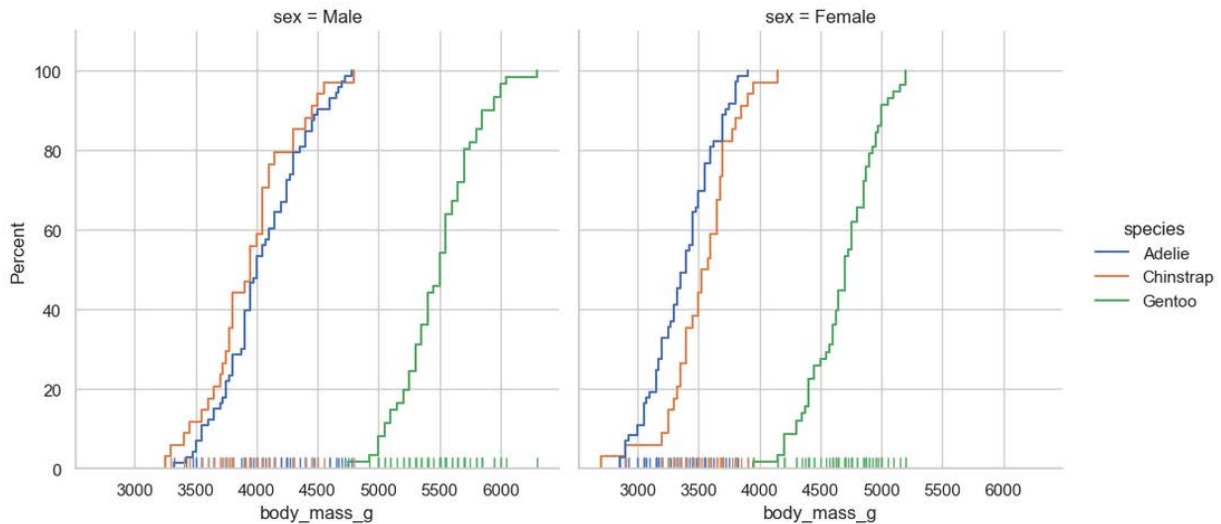
Παραλείπουμε τα ορίσματα για στατιστικό και περιοχές (stat και bins ή binwidth), καθώς δεν έχουν εφαρμογή στα διαγράμματα εκτίμησης πυκνότητας. Προσθέτουμε τη ρύθμιση fill=True, για να γεμίσουμε το εμβαδό της κάθε καμπύλης με το χρώμα της καμπύλης.

Είναι οι ίδιες καμπύλες πυκνότητας πιθανότητας που πήραμε επιπρόσθετα μέσα στα ιστογράμματα, αλλά τώρα στον κατακόρυφο άξονα y μετρούμε την πυκνότητα πιθανότητας (density), ώστε το εμβαδό κάτω από την καμπύλη κάθε μίας από τις 6 κατανομές να είναι 1. Αυτή τη δυνατότητα την έχουμε και στη σχεδίαση ιστογραμμάτων με τη ρύθμιση stat='density'.

Σωρευτική πυκνότητα

Μπορούμε να πάρουμε διαγράμματα σωρευτικών καμπυλών πυκνότητας (πιθανότητας). Η σχεδίαση σωρευτικών καμπυλών στηρίζεται στη συσσώρευση των πρωτογενών παρατηρήσεων στην κάθε καμπύλη μέχρι το άθροισμα να φθάσει το όριο: 100% (στην περίπτωση που εκφράζουμε την πιθανότητα σαν ποσοστό), ή 1 (στην περίπτωση που εκφράζουμε την πιθανότητα σαν αναλογία).

Τα ακόλουθα δύο διαγράμματα σωρευτικών καμπυλών πυκνότητας περιέχουν τις ίδιες 6 κατανομές που πήραμε σε ιστογράμματα ή διαγράμματα εκτίμησης πυκνότητας για τα είδη και το φύλο των πιγκουίνων (hue="species", col='sex', stat='percent', kind='ecdf').



Εικόνα 52 Διάγραμμα σωρευτικής κατανομής πιθανότητας

Το διάγραμμα σωρευτικής κατανομής πιθανότητας διευκολύνει τις συγκρίσεις μεταξύ κατανομών αλλά δεν κάνει εύκολο τον εντοπισμό κορυφών και κεντρικής τάσης, ούτε την αίσθηση της διασποράς/διακύμανσης της κατανομής.

Στην παραπάνω οπτικοποίηση διαπιστώνουμε ότι μεγαλύτερη κλίση της σωρευτικής καμπύλης αντιστοιχεί σε σημεία με πυκνώματα παρατηρήσεων και, άρα, σε περιοχή όπου παρατηρείται κορυφή στη καμπύλη πυκνότητας και, ενδεχομένως, κεντρική τάση. Πράγματι, αυτό επαληθεύεται με σύγκριση με αντίστοιχα ιστογράμματα ή διαγράμματα εκτίμησης πυκνότητας. Αντίθετα μικρότερη κλίση υπονοεί αραιώματα παρατηρήσεων και άρα περιοχές όπου παρατηρείται μεγαλύτερη διασπορά/διακύμανση.

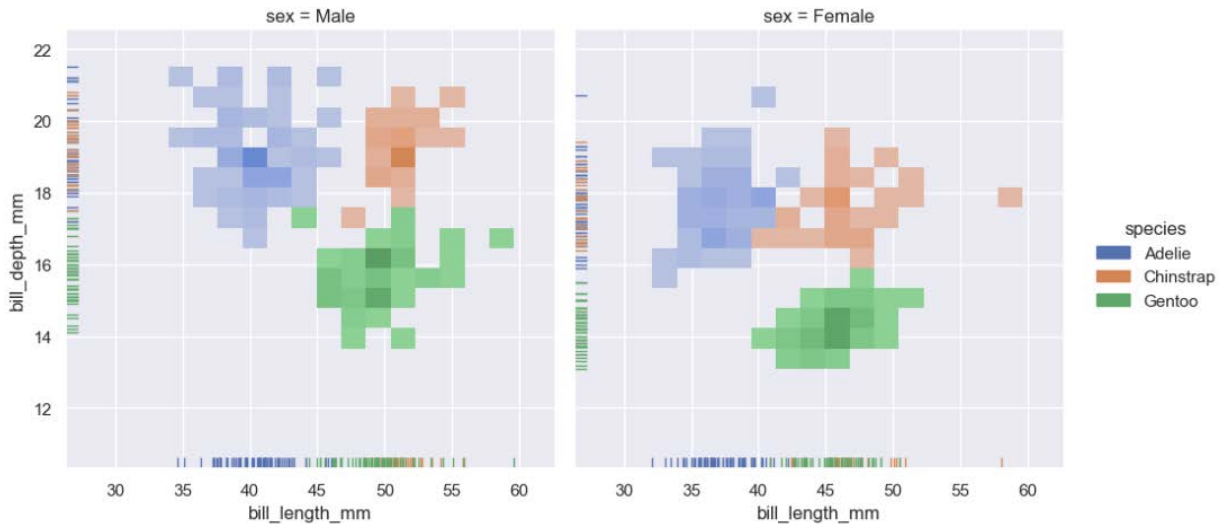
Κατανομές δύο ποσοτικών μεταβλητών

Είδαμε ότι μπορούμε να αξιοποιήσουμε ιστογράμματα, διαγράμματα εκτίμησης πυκνότητας (πιθανότητας), ακόμη και κατηγορικά διαγράμματα για να συγκρίνουμε διαφορετικές κατηγορίες στα δεδομένα μας (διαφορετικές ομάδες παρατηρήσεων), όπως, για παράδειγμα, τα τρία είδη των πιγκουίνων.

Μέχρι τώρα περιοριστήκαμε σε κατανομές μίας ποσοτικής μεταβλητής, όπου συγκρίνοντας σχήμα και διαστάσεις κατανομών μίας ποσοτικής μεταβλητής (κορυφές, διασπορά, συμμετρία/λοξότητα, ακραίες τιμές, κεντρική τάση, διακύμανση, περιθώριο σφάλματος) διαπιστώνουμε διαφορές μεταξύ κατηγοριών (ομάδων παρατηρήσεων). Η οπτικοποίηση των κατανομών δύο ποσοτικών μεταβλητών υπερτερεί στην προσπάθεια σύγκρισης και διάκρισης μεταξύ διαφορετικών κατηγοριών (ομάδων) στα δεδομένα μας.

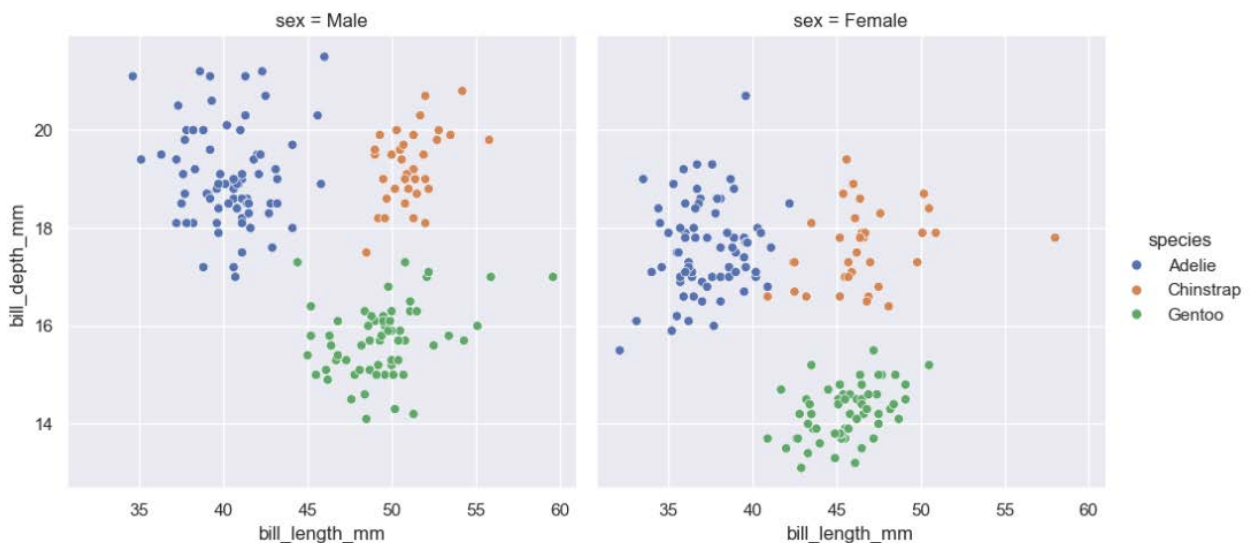
Για να πάρουμε κατανομές δύο ποσοτικών μεταβλητών ομοίως χρησιμοποιούμε ιστογράμματα ή διαγράμματα εκτίμησης πυκνότητας (πιθανότητας) που παίρνουμε με τη figure-level συνάρτηση σχεδίασης `displot` της βιβλιοθήκης Seaborn, με `kind='hist'` ή `kind='kde'`, αντίστοιχα. Προσθέτουμε τη δεύτερη ποσοτική μεταβλητή στον άξονα `y` και οι περιοχές τιμών (που ρυθμίζονται και πάλι με το όρισμα `bins` ή `binwidth`) είναι πλέον διδιάστατες (τετράγωνα) και το στατιστικό (που ρυθμίζεται και πάλι με το όρισμα `stat`, σαν ποσοστό ή αναλογία) εκφράζεται με την ένταση του χρώματος σε κάθε τετράγωνο.

Στην ακόλουθη απεικόνιση έχουμε δύο ιστογράμματα, για αρσενικούς και θηλυκούς πιγκουίνους, όπου απεικονίζουμε, με διαφορετικό χρώμα για κάθε είδος πουλιού, την κατανομή πιθανότητας δύο ποσοτικών μεταβλητών, του μήκους και του βάθους του ράμφους σε χιλιοστά (`bill_length_mm`, `bill_depth_mm`).



Εικόνα 53 Κατανομή πιθανότητας δύο ποσοτικών μεταβλητών με ιστόγραμμα

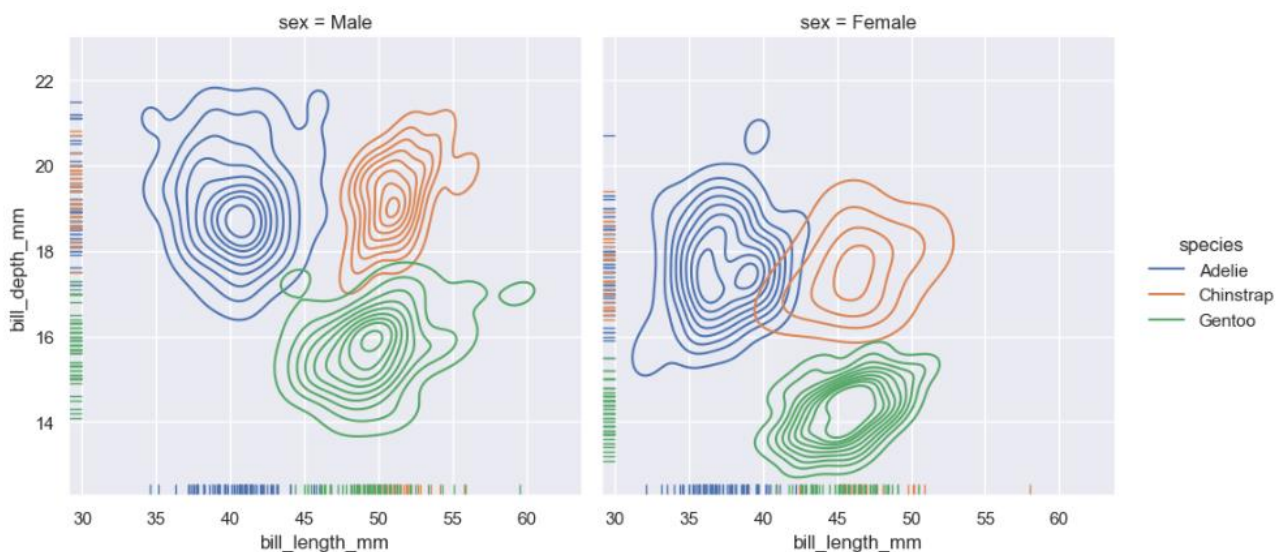
Ο συνδυασμός των επιλεγμένων μεταβλητών αποδεικνύεται κατάλληλος για τη διάκριση των τριών ειδών των πιγκουίνων. Η θέσεις στα ιστογράμματα των περιοχών με εντονότερο χρώμα, που υποδηλώνουν σημεία όπου παρατηρούνται κορυφές των κατανομών, συμφωνούν με τις θέσεις σε αντίστοιχα διαγράμματα διασποράς όπου παρατηρούνται μέγιστες συγκεντρώσεις των παρατηρήσεων. Το ακόλουθο διάγραμμα διασποράς το σχεδιάζουμε με τη figure-level συνάρτηση σχεδίασης relplot (με τη default ρύθμιση kind='scatter').



Εικόνα 54 Πυκνότητα παρατηρήσεων (διασπορά) δύο ποσοτικών μεταβλητών

Τα διαγράμματα εκτίμησης πυκνότητας (πιθανότητας) δίνουν, ίσως, μία καλύτερη εικόνα του κεντρικής τάσης και της διασποράς των κατανομών από τα ιστογράμματα και, επομένως, παρέχουν ένα, ενδεχομένως, καλύτερο εργαλείο διάκρισης μεταξύ διαφορετικών κατηγοριών (ομάδων) στα δεδομένα μας.

Στα ακόλουθα διαγράμματα εκτίμησης πυκνότητας εικονίζονται οι ισοϋψείς καμπύλες της πυκνότητας. Τις αποκαλούμε ισοϋψείς γιατί όλα τα σημεία πάνω σε κάθε καμπύλη αντιστοιχούν σε συγκεκριμένη πυκνότητα πιθανότητας. Στην απεικόνιση φαίνεται ξεκάθαρα η σφαιρική συμμετρία των καμπυλών κάθε κατανομής (κάθε είδους, δηλαδή σε κάθε χρώμα και διάγραμμα) και η συγκέντρωσή τους προς μία κεντρική τάση. Από το “άνοιγμα” των ισοϋψών φαίνεται, επίσης, η διασπορά/διακύμανση της κάθε κατανομής.



Εικόνα 55 Κατανομή πιθανότητας ζεύγους ποσοτικών μεταβλητών με εκτίμηση πυκνότητας

Αξιοσημείωτη είναι στις τρεις παραπάνω οπτικοποιήσεις η απομόνωση των κατανομών για τα τρία είδη των πουλιών (όπως διαπιστώνουμε από την έκταση που καταλαμβάνει κάθε χρώμα στις περιοχές του ιστογράμματος ή στη διασπορά παρατηρήσεων, ή στις καμπύλες εκτίμησης πυκνότητας), γεγονός που επιτρέπει τη διάκριση μεταξύ των ειδών. Η

συγκεκριμένη επιλογή των δύο σωματομετρικών μεταβλητών (του μήκους και του βάθους του ράμφους) αποδεικνύεται ιδανική για τη διάκριση των τριών ειδών των πιγκουίνων.

Εξερεύνηση δεδομένων

Είδαμε ότι η αντιπαραβολή των πυκνωμάτων της διασποράς ή των κατανομών μίας ή δύο ποσοτικών μεταβλητών που ορίζονται για κατηγορίες (ομάδες) δεδομένων επιτρέπει τη σύγκριση των κατηγοριών αυτών και την εύρεση διακριτικών στοιχείων μεταξύ τους.

Για παράδειγμα, συγκρίνοντας τη διασπορά (scatterplot) ή τις κατανομές (εκφρασμένες σε ιστογράμματα, histplot ή διαγράμματα εκτίμησης πυκνότητας, kdeplot) δύο ποσοτικών μεταβλητών (στο σύνολο ή σε επιμέρους κατηγορίες (ομάδες) των διαθέσιμων δεδομένων) μπορούμε να διακρίνουμε περιοχές συγκέντρωσης των δεδομένων, να προσδιορίσουμε για κάθε μία περιοχή την τοποθεσία της (κεντρική τάση, όπως μέσος όρος, διάμεσος, κορυφή) και την έκταση της (διασπορά, διακύμανση, αβεβαιότητα) και να την αντιστοιχίσουμε σε μία από τις κατηγορίες των δεδομένων.

Επίσης, είδαμε ότι όταν από τη διασπορά των παρατηρήσεων υπάρχει ένδειξη συσχέτισης (συμμεταβολής) δύο ποσοτικών μεταβλητών, τότε μπορούμε να πάρουμε μία εκτίμηση της σχέσης των δύο μεταβλητών με παλινδρόμηση (regression, regplot).

Η θεωρία της οπτικοποίησης δεδομένων προβλέπει δύο ισχυρά εργαλεία οπτικοποιήσεων διασποράς, σχέσεων και κατανομών ποσοτικών μεταβλητών κατά ζεύγη, τη συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (joint plot) και τη συνδυαστική οπτικοποίηση ζευγών ποσοτικών μεταβλητών (pair plot). Η βιβλιοθήκη οπτικοποίησης δεδομένων Seaborn υλοποιεί τα δύο αυτά εργαλεία με τις figure-level συναρτήσεις σχεδίασης jointplot και pairplot.

Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (joint plot)

Κώδικας. Παίρνουμε απεικονίσεις (οπτικοποιήσεις) ζεύγους ποσοτικών μεταβλητών εκτελώντας τη figure-level συνάρτηση σχεδίασης `jointplot` της βιβλιοθήκης Seaborn. Στο ακόλουθο κομμάτι κώδικα (Εικ. 56) βλέπουμε τις βασικές ρυθμίσεις της συνάρτησης `jointplot`. Ανάλογα με το αποτέλεσμα που επιθυμούμε κάθε φορά, κάνουμε στις ρυθμίσεις της συνάρτησης τις κατάλληλες τροποποιήσεις.

```
jointplot(
    data=penguins,          # το σύνολο δεδομένων (υπό μορφή pandas dataframe) από τις στήλες του οποίου αντλούνται οι μεταβλητές
    #x="body_mass_g",        # η ποσοτική μεταβλητή στον άξονα x
    #y="flipper_length_mm",  # η ποσοτική μεταβλητή στον άξονα y
    x='bill_length_mm',     # εναλλακτική ποσοτική μεταβλητή στον άξονα x
    y='bill_depth_mm',     # εναλλακτική ποσοτική μεταβλητή στον άξονα y
    hue='species',          # για χρωματική διαφοροποίηση των σημείων/γραφικών, ένα χρώμα για κάθε κατηγορία/επίπεδο της επιλεγμένης
    #kind='kde',             # ο τύπος του κεντρικού διαγράμματος: 'scatter' (default), 'kde' or 'hist' or 'hex' or 'reg' or 'resid'
    #kind='hist',
    #kind='hex',
    joint_kws=None,
    marginal_kws=dict(fill=True),
)
```

Εικόνα 3656 Κλήση συνάρτησης σχεδίασης ζεύγους ποσοτικών μεταβλητών (`jointplot`)

Μπορείτε να βρείτε το συγκεκριμένο κώδικα (script) και να δοκιμάσετε παραλλαγές εκτέλεσής του στο σημειωματάριο του οδηγού μελέτης, “Data Visualization with Seaborn.ipynb”. Θυμίζουμε ότι ένα σημειωματάριο (όπως είναι ο οδηγός μελέτης) το εκτελούμε μέσα σε περιβάλλον Jupyter Notebook

Συναρτήσεις σχεδίασης και διαγράμματα. Η figure-level συνάρτηση σχεδίασης `jointplot` παράγει μία απεικόνιση ζεύγους ποσοτικών μεταβλητών. Αυτή είναι μία οπτικοποίηση με ένα κεντρικό διάγραμμα και δύο διαγράμματα στο περιθώριο.

Στο κεντρικό διάγραμμα παίρνουμε διάγραμμα διασποράς ή κατανομή/ές (ιστόγραμμα ή διάγραμμα εκτίμησης πυκνότητας) ή διάγραμμα συσχέτισης (ή αποκλίσεων). Στο κάθε ένα από τα δύο περιθώρια διαγράμματα παίρνουμε κατανομή/ές της αντίστοιχης ποσοτικής μεταβλητής, που, ανάλογα με τις ρυθμίσεις των ορισμάτων `kind` και `hue` (περιγράφονται στη

συνέχεια), μπορεί να είναι ιστογράμματα ή διαγράμματα εκτίμησης πυκνότητας (δείτε στον οδηγό μελέτης, για λεπτομερή τεκμηρίωση).

Ρυθμίσεις της συνάρτησης `jointplot`. Η σημαντικότερη ρύθμιση είναι το όρισμα `kind`. Αν επιθυμούμε στο κεντρικό διάγραμμα απεικόνιση της διασποράς των παρατηρήσεων, τότε θέτουμε `kind='scatter'` (default, δεν είναι απαραίτητη η ρύθμιση), αν επιθυμούμε ιστογράμματα, τότε ορίζουμε `kind='hist'` (ή `kind='hex'`, για εξαγώνες περιοχές τιμών), αν επιθυμούμε εκτίμηση πυκνότητας, τότε θέτουμε `kind='kde'`, αν επιθυμούμε καμπύλη συσχέτισης, τότε ορίζουμε `kind='reg'` (ή `kind='resid'` για διάγραμμα αποκλίσεων). Εσωτερικά, η συνάρτηση `jointplot` καλεί τις αντίστοιχες axes-level συναρτήσεις, `scatterplot`, `histplot`, `kdeplot` ή `regplot`.

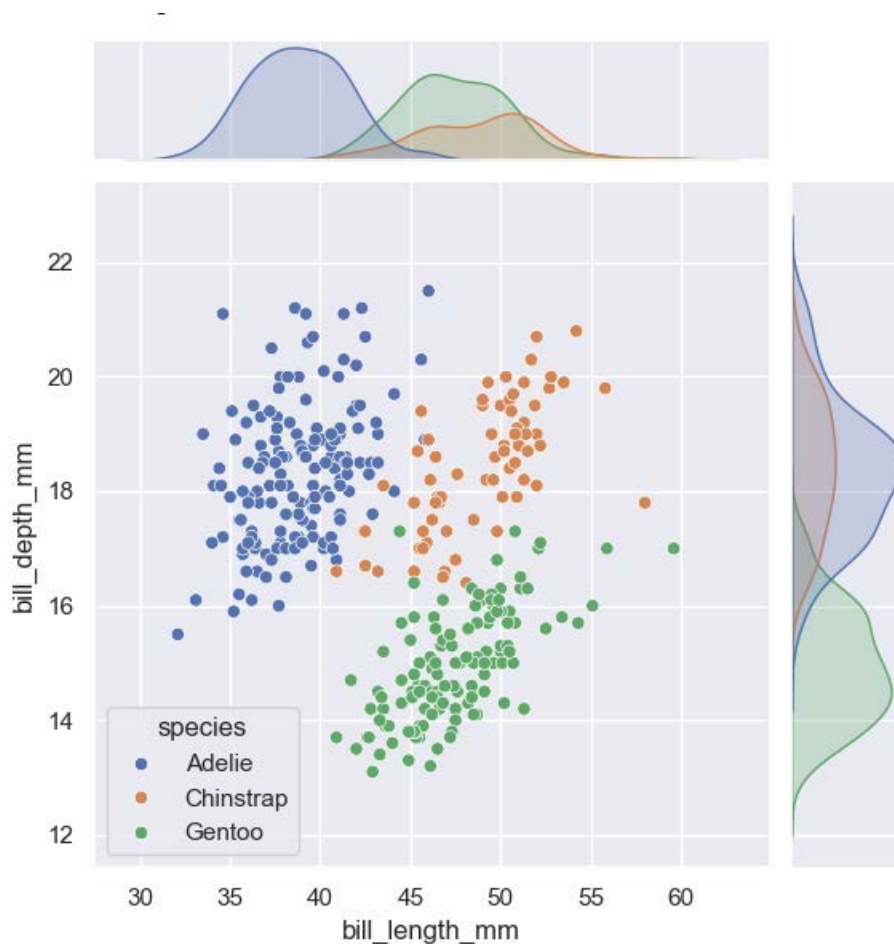
Με τη γνωστή ρύθμιση `'hue'` επιτυγχάνουμε χρωματική διαφοροποίηση στο κεντρικό και στα δύο περιθώρια διαγράμματα (εκτός αν `'kind='hex'` ή `'kind='reg'` ή `'kind='resid'`, επιλογές που δεν υποστηρίζουν προς το παρόν τη ρύθμιση `'hue'` και, επομένως, χρωματική διαφοροποίηση).

Η figure-level συνάρτηση σχεδίασης `jointplot` δε διαθέτει τα ορίσματα `col` και `row`, γιατί το πλέγμα της είναι δεδομένο, με τρία διαγράμματα, ένα στο κέντρο και δύο στα περιθώρια. Για λεπτομερή τεκμηρίωση των ρυθμίσεων της συνάρτησης `jointplot` συμβουλευτείτε τον οδηγό μελέτης.

Διασπορά και κατανομές. Στις ακόλουθες συνδυαστικές οπτικοποιήσεις ζεύγους ποσοτικών μεταβλητών επεκτείνουμε τα διαγράμματα διασποράς ή κατανομών δύο ποσοτικών μεταβλητών από το σύνολο δεδομένων των πιγκουίνων (Εικ. 53-55), όπου απεικονίζουμε, με διαφορετικό χρώμα για κάθε είδος πουλιού, τη διασπορά ή το ιστόγραμμα

ή την εκτίμηση πυκνότητας του μήκους και του βάθους του ράμφους σε χιλιοστά (bill_length_mm, bill_depth_mm).

Ας εμφανίσουμε πρώτα απεικόνιση με κεντρικό διάγραμμα διασποράς (`kind='scatter'`) και περιθώρια διαγράμματα εκτίμησης πυκνότητας (Εικ. 57).

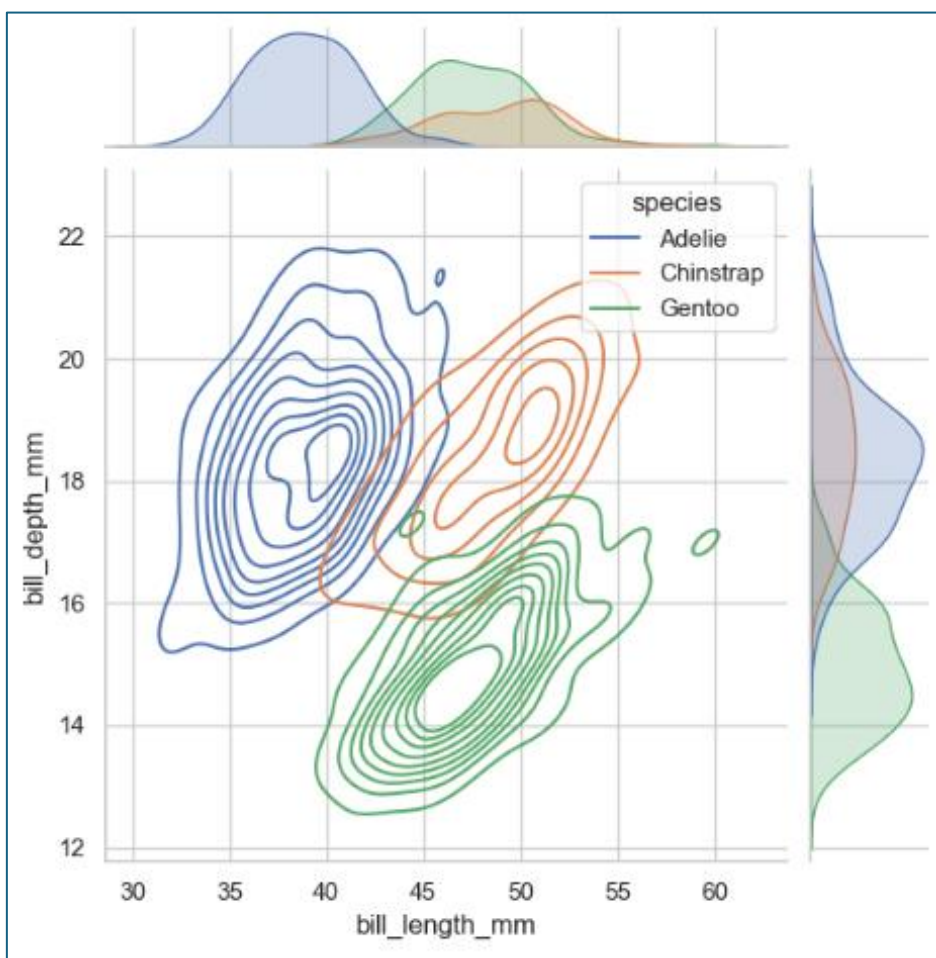


Εικόνα 57 Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (jointplot): διασπορά και περιθώριες κατανομές

Στο διάγραμμα διασποράς διαπιστώνουμε ότι οι παρατηρήσεις των δύο ποσοτικών μεταβλητών συγκεντρώνονται σε διακριτές περιοχές του επιπέδου κάτι που καθιστά τις δύο επιλεγμένες μεταβλητές ιδανικές για διάκριση των ειδών. Μπορούμε, επιπλέον να εκτιμήσουμε το κέντρο και το βαθμό διασποράς και να συγκρίνουμε τις

συγκεντρώσεις/πυκνώματα στο επίπεδο με τις συγκεντρώσεις/πυκνώματα στις περιθώριες κατανομές.

Αν ζητήσουμε σε νέα οπτικοποίηση (Εικ. 58) διάγραμμα εκτίμησης πυκνότητας (`kind='kde'`), τότε η γεωμετρία των ισοϋψών καμπυλών πυκνότητας συμφωνούν με τη διασπορά των παρατηρήσεων (στην Εικ. 57) και τις περιθώριες κατανομές ως προς την τοποθεσία (κεντρική τάση) και την έκταση (διασπορά/διακύμανση) των κατανομών.

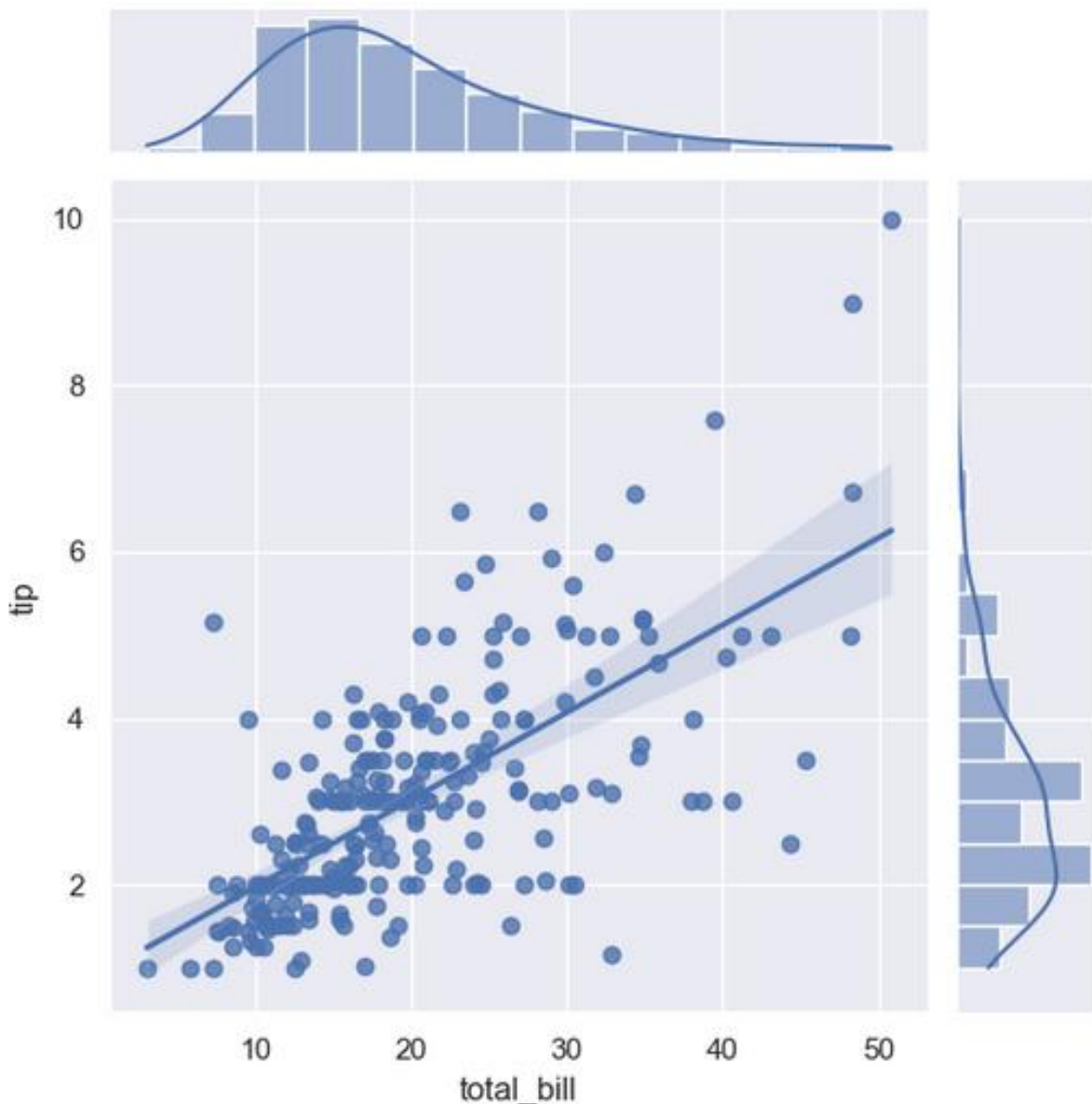


Εικόνα 58 Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (jointplot): κατανομές (στο κέντρο/περιθώρια)

Η εξερεύνηση των δύο ποσοτικών μεταβλητών, του μήκους και του βάθους του ράμφους σε χιλιοστά (bill_length_mm, bill_depth_mm) ως προς την ικανότητα τους για διάκριση των

τριών ειδών οδηγεί στο συμπέρασμα ότι το ζεύγος των εν λόγω μεταβλητών έχει συνδυαστική κατανομή που το καθιστά ιδανικό για αυτό το ρόλο.

Διασπορά και συσχέτιση. Ας εξετάσουμε τώρα τη συσχέτιση δύο ποσοτικών μεταβλητών, του ποσού λογαριασμού και του φιλοδωρήματος (`total_bill` και `tip`) από το σύνολο δεδομένων `tips` και πως αυτή συνδυάζεται με τις περιθώριες κατανομές (Εικ. 59).



Εικόνα 59 Συνδυαστική οπτικοποίηση ζεύγους ποσοτικών μεταβλητών (*jointplot*): διασπορά, συσχέτιση και περιθώριες κατανομές

Από τη σύγκριση στην Εικ. 59 του κεντρικού διαγράμματος συσχέτισης με τις περιθώριες κατανομές διαπιστώνουμε ότι το μικρότερο περιθώριο σφάλματος (αβεβαιότητα) στην εκτίμηση της συσχέτισης σημειώνεται σε περιοχές τιμών όπου παρατηρείται μεγαλύτερη πυκνότητα (δηλαδή, στην περιοχή του επιπέδου όπου οι παρατηρήσεις σχηματίζουν πυκνώματα ή στην περιοχή γύρω από την κορυφή στην κάθε περιθώρια κατανομή).

Συνδυαστική οπτικοποίηση μεταβλητών κατά ζεύγη (pair plot)

Κώδικας. Παίρνουμε απεικονίσεις (οπτικοποιήσεις) πολλών ζευγών ποσοτικών μεταβλητών εκτελώντας τη figure-level συνάρτηση σχεδίασης pairplot της βιβλιοθήκης Seaborn. Στο ακόλουθο κομμάτι κώδικα (Εικ. 60) βλέπουμε τις βασικές ρυθμίσεις της συνάρτησης pairplot. Ανάλογα με το αποτέλεσμα που επιθυμούμε κάθε φορά, κάνουμε στις ρυθμίσεις της συνάρτησης τις κατάλληλες τροποποιήσεις.

```
pairplot(
    data=penguins,          # το σύνολο δεδομένων (υπό μορφή pandas dataframe) από τις στήλες του οποίου αντλούνται οι ποσοτικές μεταβλητές
    #vars=['bill_length_mm', 'bill_depth_mm', 'flipper_length_mm'], # επιλογή υποσυνόλου των ποσοτικών μεταβλητών
    #x_vars=['bill_length_mm', 'bill_depth_mm', 'flipper_length_mm'], # επιλογή υποσυνόλου ποσοτικών μεταβλητών στον άξονα x (μη τετραγωνικό πλέγμα)
    #y_vars=['flipper_length_mm', 'body_mass_g'], # επιλογή υποσυνόλου ποσοτικών μεταβλητών στον άξονα y (μη τετραγωνικό πλέγμα)
    hue='species',          # για χρωματική διαφοροποίηση των σημείων/γραφικών, ένα χρώμα για κάθε κατηγορία/επίπεδο της επιλεγμένης μεταβλητής - όποι
    kind='hist',            # ο τύπος των διαγραμμάτων εκτός διαγώνιου: 'scatter' (default) or 'kde' or 'hist' or 'reg'
    diag_kind='hist',       # ο τύπος των διαγραμμάτων στη διαγώνιο (κατανομές): 'auto' (default, 'kde' if hue, or 'hist') or 'hist' or 'kde' or None
    plot_kws=dict(bins=15), # ορίσματα για προσαρμογή (ρυθμίσεις) των διαγραμμάτων εκτός διαγώνιου
    diag_kws=dict(bins=20, kde=True), # ορίσματα για προσαρμογή των διαγραμμάτων της διαγώνιου (εναλλακτική σύνταξη: diag_kws={'bins':20, 'kde':True})
    #markers=None,
    #height=2.0, aspect=1,
    #corner=True,
    #dropna=True,
)
```

Εικόνα 60 Κλήση συνάρτησης σχεδίασης ζευγών ποσοτικών μεταβλητών (pair plot)

Μπορείτε να βρείτε το συγκεκριμένο κώδικα (script) και να δοκιμάσετε παραλλαγές εκτέλεσής του στο σημειωματάριο του οδηγού μελέτης, “Data Visualization with Seaborn.ipynb”. Θυμίζουμε ότι ένα σημειωματάριο (όπως είναι ο οδηγός μελέτης) το εκτελούμε μέσα σε περιβάλλον Jupyter Notebook.

Συναρτήσεις σχεδίασης, διαγράμματα και ρυθμίσεις. Η figure-level συνάρτηση σχεδίασης `pairplot` συνδυάζει σε ζεύγη όλες τις ποσοτικές μεταβλητές από τις στήλες του συνόλου δεδομένων (`dataframe`) που δηλώνουμε στο όρισμα `data`, για να παράξει μία οπτικοποίηση που περιέχει στο πλέγμα της ένα διάγραμμα για κάθε δυνατό ζευγάρι αυτών των ποσοτικών μεταβλητών.

Μπορούμε να περιορίσουμε τη λίστα των ποσοτικών μεταβλητών από τις οποίες θα σχηματιστούν τα ζεύγη με το όρισμα `vars`. Μπορούμε, ακόμη, να φιλτράρουμε τις μεταβλητές που θα χρησιμοποιήσουμε στην κάθε μία από τις διευθύνσεις του πλέγματος, οριζόντια και κατακόρυφη (ουσιαστικά στον κάθε άξονα, x και y) με δύο ακόμη ορίσματα, τα `x_vars` και `y_vars`. Περισσότερα για την επιλογή των ποσοτικών μεταβλητών στον οδηγό μελέτης της βιβλιοθήκης Seaborn.

Στη διαγώνιο του πλέγματος (όπου κάθε μεταβλητή ζευγαρώνει με τον εαυτό της) η συνάρτηση σχεδίασης `pairplot` τοποθετεί διαγράμματα κατανομών, είτε ιστογράμματα (αν η σχετική ρύθμιση είναι `diag_kind='hist'`, οπότε εσωτερικά καλείται η axes-level συνάρτηση `histplot`) είτε διαγράμματα εκτίμησης πυκνότητας (αν η σχετική ρύθμιση είναι `diag_kind='kde'`, οπότε εσωτερικά καλείται η axes-level συνάρτηση `kdeplot`).

Στις θέσεις του πλέγματος εκτός της διαγωνίου η συνάρτηση σχεδίασης `pairplot` τοποθετεί διαγράμματα για τα ζεύγη των ποσοτικών μεταβλητών. Όλα τα διαγράμματα είναι ενός συγκεκριμένου τύπου, είτε διασποράς (που παίρνουμε με τη ρύθμιση `kind='scatter'` (default), ώστε εσωτερικά να κληθεί η axes-level συνάρτηση `scatterplot`) είτε συσχέτισης (που παίρνουμε με τη ρύθμιση `kind='reg'`, ώστε εσωτερικά να κληθεί η axes-level συνάρτηση `regplot`) είτε κατανομής (που παίρνουμε με τη ρύθμιση `kind='hist'` ή `kind='kde'`, ώστε εσωτερικά να κληθεί η axes-level συνάρτηση `histplot` ή `kdeplot`).

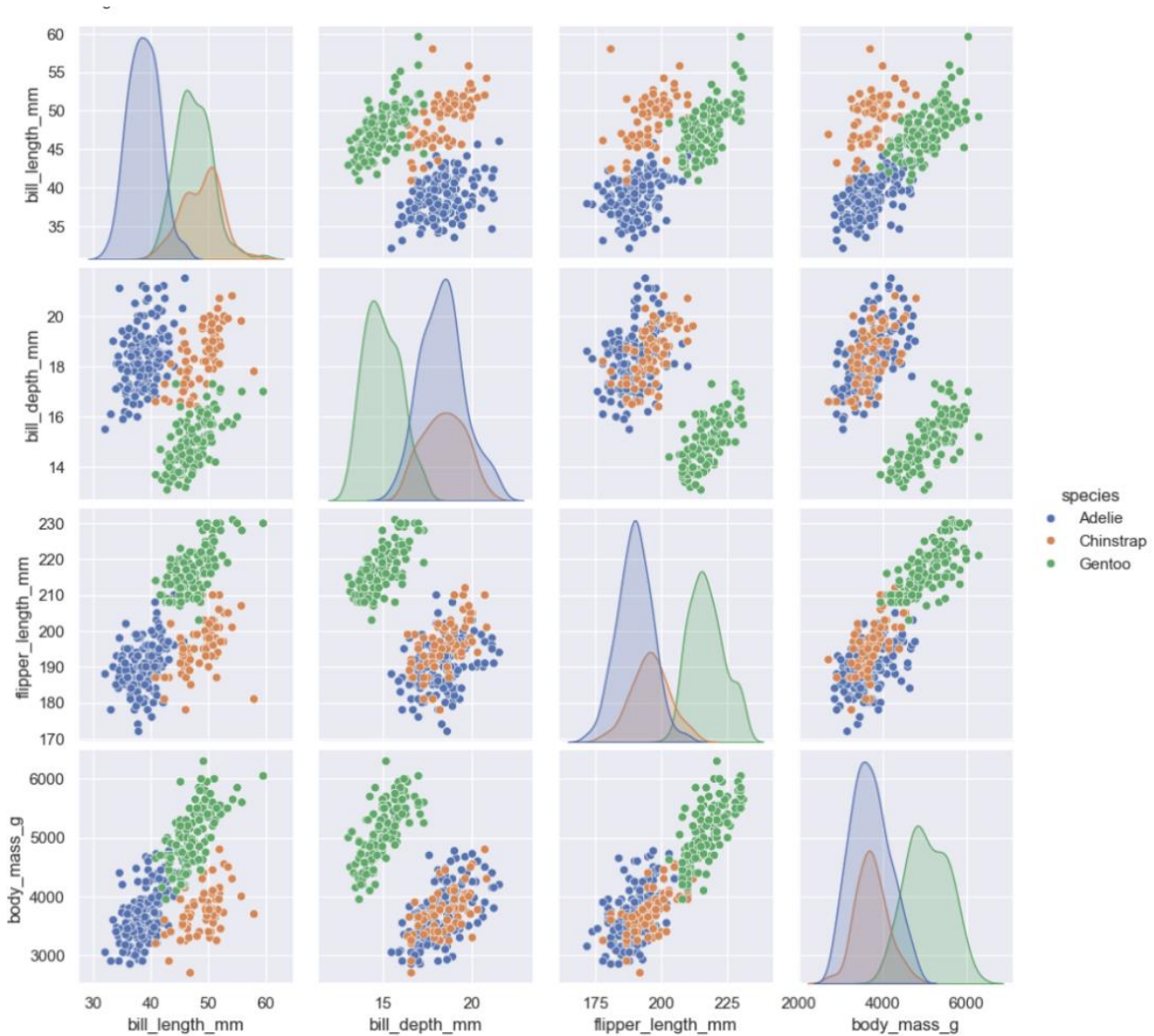
Αν δεν έχουμε κάνει ρύθμιση στα ορίσματα `diag_kind` και `kind`, τότε χρησιμοποιούνται οι εξ' ορισμού (default) ρυθμίσεις. Για τον τύπο διαγράμματος στη διαγώνιο η default ρύθμιση είναι `diag_kind='kde'`, αν έχει γίνει ρύθμιση `hue` (διάγραμμα εκτίμησης πυκνότητας πιθανότητας), αλλιώς είναι `diag_kind='hist'` (ιστόγραμμα). Για τον τύπο διαγράμματος εκτός διαγωνίου η default ρύθμιση είναι `kind='scatter'`.

Μπορούμε να θεωρήσουμε τη συνάρτηση σχεδίασης `pairplot` σαν γενίκευση της συνάρτησης σχεδίασης `jointplot`, όπου οι δύο περιθώριες θέσεις του πλέγματος στη `jointplot` γενικεύονται στις θέσεις της διαγωνίου στη `pairplot` και το κέντρο του πλέγματος της `jointplot` γενικεύεται στις θέσεις εκτός διαγωνίου της `pairplot`.

Διασπορά και κατανομές. Ας επεκτείνουμε τις συνδυαστικές οπτικοποιήσεις ζεύγους ποσοτικών μεταβλητών (`joint plots`), που πήραμε στην προηγούμενη ενότητα από το σύνολο δεδομένων με σωματομετρικά και κατηγορικά στοιχεία πιγκουίνων (Εικ. 57-58) σε συνδυαστικές οπτικοποιήσεις πολλών ζευγαριών ποσοτικών μεταβλητών (`pair plots`).

Ας υποθέσουμε ότι θέλουμε να διερευνήσουμε αν στο σύνολο δεδομένων, με σωματομετρικά στοιχεία πιγκουίνων, υπάρχει ζευγάρι ποσοτικών μεταβλητών με συνδυαστική κατανομή που επιτρέπει τη σαφή διάκριση των τριών ειδών των πιγκουίνων.

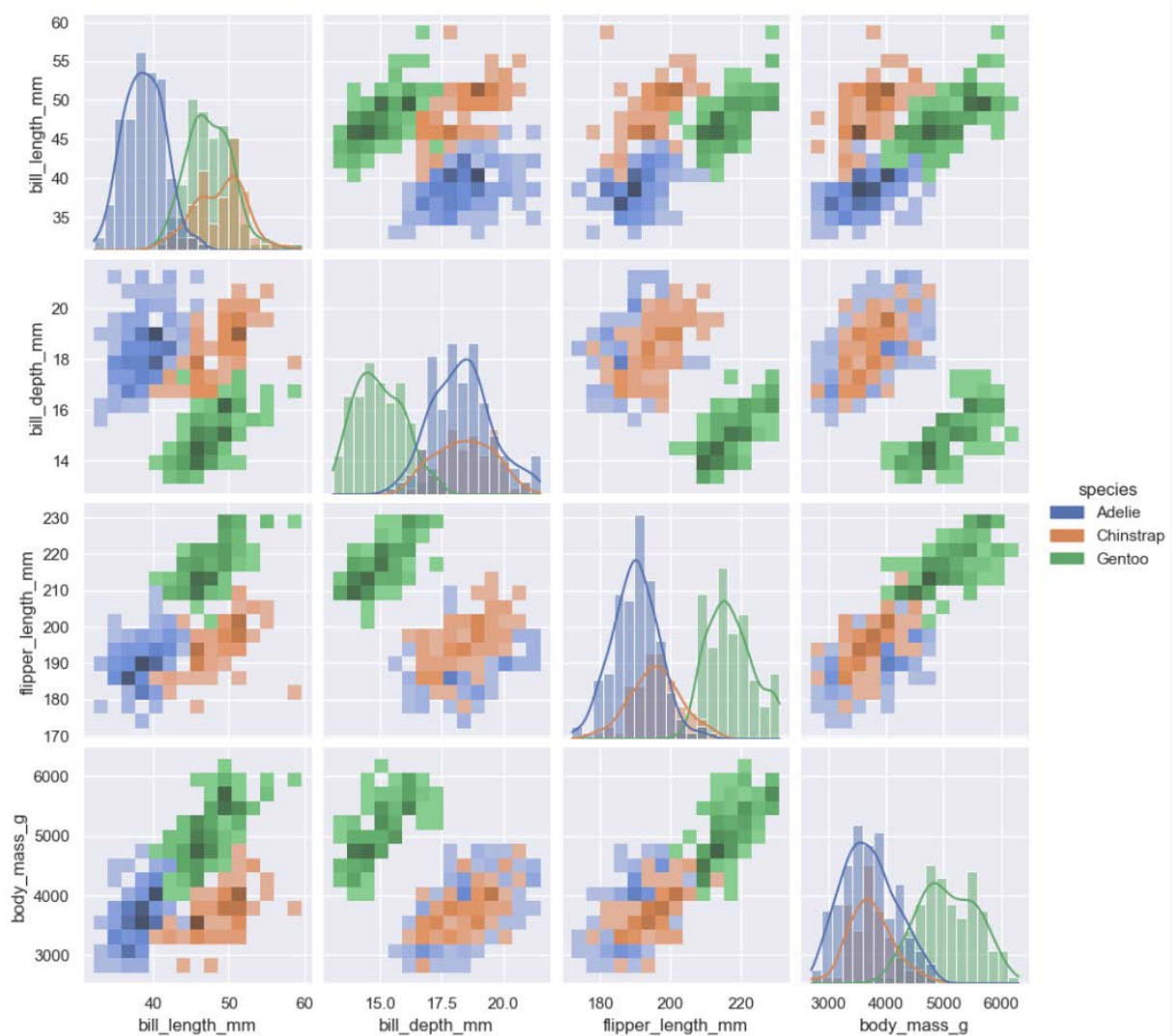
Στην Εικ. 61 εμφανίζουμε για τις ποσοτικές μεταβλητές στο σύνολο δεδομένων των πιγκουίνων διαγράμματα διασποράς κάθε ζεύγους εκτός της διαγωνίου (`kind='scatter'`) και διαγράμματα εκτίμησης πυκνότητας κάθε μεταβλητής στη διαγώνιο (`diag_kind='kde'`).



Εικόνα 61 Εξερεύνηση πολλών μεταβλητών με pair plot (διασπορά)

Η figure-level συνάρτηση σχεδίασης pairplot παράγει οπτικοποιήσεις που υποστηρίζουν την εξερεύνηση των δεδομένων. Συγκρίνοντας τα εκτός διαγωνίου διαγράμματα διασποράς διαπιστώνουμε ότι καλύτερα ζεύγη μεταβλητών είναι αυτά που σχηματίζονται από τη μεταβλητή του μήκους του ράμφους σε χιλιοστά (bill_length_mm) και μία από τις τρεις άλλες μεταβλητές.

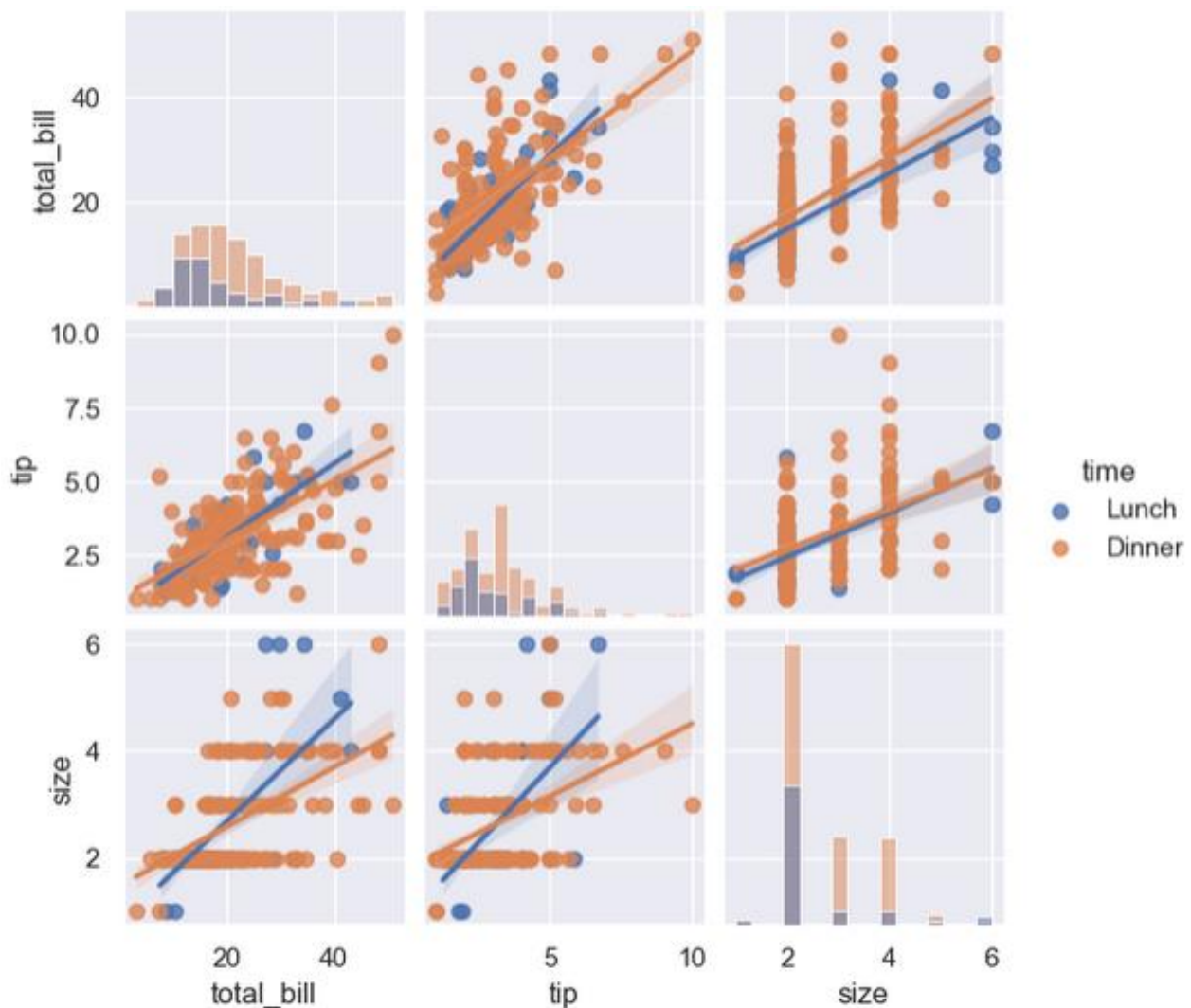
Στο ίδιο συμπέρασμα οδηγούμαστε αν δημιουργήσουμε οπτικοποίηση (Εικ. 62) με ιστογράμματα ή διαγράμματα εκτίμησης πυκνότητας για κάθε ζεύγος μεταβλητών εκτός της διαγωνίου (`kind='hist'` ή `kind='kde'`).



Εικόνα 62 Εξερεύνηση πολλών μεταβλητών με *pair plot* (κατανομές)

Στην παραπάνω οπτικοποίηση επιλέγουμε να εμφανίσουμε ιστογράμματα για κάθε ζεύγος μεταβλητών εκτός της διαγωνίου, όμοια, όμως, πληροφορία παίρνουμε με διαγράμματα εκτίμησης πυκνότητας. Κατευθυνθείτε στον οδηγό μελέτης για να κάνετε τις δικές σας δοκιμές και να επαληθεύσετε ή απορρίψετε τα παραπάνω συμπεράσματα.

Διασπορά και συσχέτιση. Με τη συνάρτηση `pairplot` μπορούμε, επίσης, να διερευνήσουμε τη συσχέτιση ζευγών ποσοτικών μεταβλητών εκτός διαγωνίου (`kind='reg'`), σε συνδυασμό με τις περιθώριες πιθανότητες των μεταβλητών, στη διαγώνιο (`diag_kind='hist'`).



Εικόνα 63 Εξερεύνηση πολλών μεταβλητών με *pair plot* (συσχέτιση)

Στην Εικ. 63, για κάθε ζεύγος ποσοτικών μεταβλητών στο σύνολο δεδομένων φιλοδωρημάτων (tips), εμφανίζουμε στα διαγράμματα εκτός διαγωνίου μαζί τη διασπορά των παρατηρήσεων και την εκτιμώμενη γραμμή συσχέτισης. Με μία ματιά επαληθεύουμε τη διαίσθηση ότι υπάρχει αύξουσα σχέση μεταξύ φιλοδωρήματος (tip) και ποσού λογαριασμού

(total_bill). Επίσης, ότι υπάρχει αύξουσα σχέση μεταξύ φιλοδωρήματος ή ποσού λογαριασμού και μεγέθους παρέας (size).

Γ. Οδηγός Μελέτης

Ψηφιακά αρχεία

1. Οδηγός μελέτης "Data Visualization with Seaborn.ipynb"

Το αρχείο "Data Visualization with Seaborn.ipynb" περιέχει τον αλληλεπιδραστικό οδηγό μελέτης της βιβλιοθήκης οπτικοποίησης δεδομένων Seaborn. Ο οδηγός μελέτης περιλαμβάνει παραδείγματα με διαδραστικό κώδικα και επεξηγηματικό κείμενο για την εκμάθηση της οπτικοποίησης δεδομένων με τη βιβλιοθήκη Seaborn.

2. Οδηγός μελέτης "Data Visualization with Seaborn.pdf"

Στο αρχείο αυτό περιλαμβάνεται ο οδηγός μελέτης σε στατική μορφή PDF (125 σελίδες).

Σύνοψη

Η παρούσα διπλωματική εργασία είχε ως στόχο τη σχεδίαση και υλοποίηση ενός ολοκληρωμένου εκπαιδευτικού οδηγού (οδηγού μελέτης) για τη βιβλιοθήκη οπτικοποίησης δεδομένων Seaborn, με τη μορφή ενός διαδραστικού σημειωματαρίου (σε περιβάλλον Jupyter Notebook). Μέσα από αυτή τη διαδικασία, επιτεύχθηκε η δημιουργία ενός δομημένου και προσβάσιμου εκπαιδευτικού υλικού που καλύπτει προοδευτικά, από τις βασικές, έως πιο προχωρημένες λειτουργίες της Seaborn.

Το παραδοτέο αποτελείται από δύο βασικά μέρη: ένα κείμενο-παρουσίαση που περιλαμβάνει, σε αδρές γραμμές, την εισαγωγή στο περιβάλλον Jupyter Notebook και την ανάλυση των δυνατοτήτων της Seaborn και τον πρακτικό οδηγό μελέτης σε μορφή διαδραστικού σημειωματαρίου.

Η κατηγοριοποίηση του περιεχομένου σε τρεις βασικές ενότητες, σχέσεις μεταβλητών, κατανομές μεταβλητών, και εξερεύνηση ζευγαριών μεταβλητών, αποδείχθηκε αποτελεσματική για την προοδευτική εκμάθηση της βιβλιοθήκης. Αυτή η δομή επιτρέπει στους χρήστες να κατανοήσουν σταδιακά τις δυνατότητες της Seaborn, από τις βασικές λειτουργίες σχεδίασης έως τις πιο σύνθετες αναλύσεις.

Η επιλογή του Jupyter Notebook ως πλατφόρμας παρουσίασης αποδείχθηκε ιδιαίτερα επιτυχής, καθώς προσφέρει ένα διαδραστικό και ευέλικτο περιβάλλον μάθησης. Ενσωματώνει αρμονικά κείμενο, εκτελέσιμο κώδικα και οπτικοποιήσεις, επιτρέποντας στους χρήστες να πειραματιστούν άμεσα και να λαμβάνουν ανατροφοδότηση σε πραγματικό χρόνο. Αυτό ενισχύει την κατανόηση και διευκολύνει τη βιωματική προσέγγιση στην εκμάθηση.

Στο πλαίσιο αυτό, ο οδηγός αξιοποιεί πλήρως τις δυνατότητες της βιβλιοθήκης Seaborn, καλύπτοντας ένα ευρύ φάσμα τύπων οπτικοποιήσεων. Η βιβλιοθήκη Seaborn προσφέρει ελκυστικές και όμορφες απεικονίσεις δεδομένων μέσω ενός διαισθητικού και λειτουργικού περιβάλλον. Η Seaborn απευθύνεται κυρίως σε αναλυτές δεδομένων και επιστήμονες που επιδιώκουν την αποτελεσματική επικοινωνία πληροφοριών και τη βελτίωση της διαδικασίας λήψης αποφάσεων. Επιτρέπει την αναγνώριση μοτίβων για παρακολούθηση τάσεων και την ανάλυση σχέσεων μεταξύ των δεδομένων, μετατρέποντας χρώματα αντί αριθμών σε κατανοητές οπτικές αναπαραστάσεις, ώστε οι ακατέργαστοι αριθμοί να μετατρέπονται σε αφήγημα.

Η βιβλιοθήκη Seaborn προσφέρει προηγμένες τεχνικές στατιστικής απεικόνισης, παρέχοντας σαφή εικόνα για τα δεδομένα και αναδεικνύοντας στοιχεία όπως η κεντρική τάση και άλλα στατιστικά μέτρα. Οι πλούσιες οπτικοποιήσεις που δημιουργεί οδηγούν σε βαθύτερη γνώση των δεδομένων, καθώς βοηθούν το ακροατήριο να διαβάσει γρήγορα τις πληροφορίες, κατευθύνουν το ακροατήριο προς τα σημαντικότερα ευρήματα και διευκολύνουν τη διαδικασία με την οποία συγκρίνουμε τα δεδομένα.

Στον οδηγό μελέτης παρουσιάζονται διαγράμματα διασποράς, διαγράμματα συσχέτισης, κατηγορικά διαγράμματα, χρονοσειρές και διαγράμματα εξέλιξης, ιστογράμματα και κατανομές, καθώς και εξερεύνηση δεδομένων με χρήση joint plots και pair plots. Αξιοποιώντας τον οδηγό μελέτης οι χρήστες εξοικειώνονται με τις λειτουργίες της βιβλιοθήκης, ενισχύοντας παράλληλα τις δεξιότητές τους στην ανάλυση και παρουσίαση δεδομένων.

Το παραγόμενο έργο διαθέτει σημαντική εκπαιδευτική και πρακτική αξία, καθιστώντας το ένα χρήσιμο εργαλείο για ευρύ φάσμα χρηστών. Από τη μία πλευρά, μπορεί να αξιοποιηθεί

εκπαιδευτικά από φοιτητές που εισάγονται στην ανάλυση και οπτικοποίηση δεδομένων, από επαγγελματίες που επιδιώκουν την ενίσχυση των δεξιοτήτων τους στη χρήση της βιβλιοθήκης Seaborn, αλλά και από εκπαιδευτές ως υποστηρικτικό υλικό για τη διδασκαλία σχετικών μαθημάτων.

Η πρακτική του αξία ενισχύεται περαιτέρω από τη χρήση πραγματικών συνόλων δεδομένων και την παρουσίαση παραδειγμάτων που ανταποκρίνονται σε ρεαλιστικά σενάρια ανάλυσης. Αυτό καθιστά τον οδηγό άμεσα εφαρμόσιμο σε πραγματικές ανάγκες του χώρου των δεδομένων.

Το έργο αυτό θέτει τις βάσεις για μια σειρά από μελλοντικές επεκτάσεις. Σε τεχνικό επίπεδο, μπορούν να προστεθούν πιο προχωρημένες τεχνικές οπτικοποίησης, διαδραστικά στοιχεία μέσω widgets, καθώς και διασύνδεση με άλλες βιβλιοθήκες όπως η matplotlib και η plotly. Παράλληλα, υπάρχουν περιθώρια για εκπαιδευτική βελτίωση, μέσω της ενσωμάτωσης ασκήσεων και quiz για αυτο-αξιολόγηση, δημιουργίας συνοδευτικών βίντεο-tutorials και ανάπτυξης διαφορετικών επιπέδων δυσκολίας για ποικίλα κοινά.

Η διαδικασία δημιουργίας αυτού του εκπαιδευτικού οδηγού αποτέλεσε μια πολύπλευρη μαθησιακή εμπειρία. Αρχικά, η εμβάθυνση στις λειτουργίες της Seaborn οδήγησε σε βαθύτερη κατανόηση των αρχών της οπτικοποίησης δεδομένων και της σημασίας της επιλογής του κατάλληλου τύπου γραφήματος για κάθε περίπτωση. Παράλληλα, η εργασία με το Jupyter Notebook ανέδειξε τις δυνατότητες των διαδραστικών εκπαιδευτικών εργαλείων και την αξία της άμεσης εκτέλεσης κώδικα στη διδακτική διαδικασία. Η εμπειρία αυτή διεύρυνε την κατανόησή μου για τα σύγχρονα εργαλεία ανάπτυξης και τις δυνατότητές τους στην εκπαίδευση.

Κατά τη διάρκεια της εργασίας, αντιμετωπίστηκαν διάφορες προκλήσεις που συνέβαλαν στην προσωπική και επαγγελματική ανάπτυξη. Στον τεχνικό τομέα, η εξοικείωση με την πολυπλοκότητα ορισμένων λειτουργιών της Seaborn απαιτήσε εκτενή μελέτη της τεκμηρίωσης και πειραματισμό, ενώ η διαχείριση διαφορετικών τύπων δεδομένων και η προσαρμογή των οπτικοποιήσεων ανάλογα με τη φύση τους αποτέλεσαν επιπλέον προκλήσεις. Επίσης, η επίλυση προβλημάτων συμβατότητας μεταξύ διαφορετικών εκδόσεων βιβλιοθηκών απαιτήσε προσεκτική διαχείριση του περιβάλλοντος ανάπτυξης.

Από παιδαγωγική άποψη, η εύρεση της κατάλληλης ισορροπίας μεταξύ απλότητας και πληρότητας στην παρουσίαση αποδείχθηκε ιδιαίτερα απαιτητική. Η δημιουργία παραδειγμάτων που να είναι ταυτόχρονα εκπαιδευτικά και ενδιαφέροντα, καθώς και η οργάνωση του περιεχομένου με τρόπο που να διευκολύνει την προοδευτική εκμάθηση, απαιτήσαν συνεχή αναθεώρηση και βελτίωση της προσέγγισης.

Σε οργανωτικό επίπεδο, η διαχείριση χρόνου για την ολοκλήρωση ενός τόσο εκτενούς έργου και η διατήρηση συνέπειας στο ύφος και τη δομή σε όλο το υλικό αποτέλεσαν σημαντικές προκλήσεις που βοήθησαν στην ανάπτυξη οργανωτικών δεξιοτήτων.

Η ολοκλήρωση αυτής της διπλωματικής εργασίας συνέβαλε σημαντικά στην προσωπική μου ανάπτυξη σε πολλαπλούς τομείς. Στον τεχνικό τομέα, αναπτύχθηκε βαθιά κατανόηση της οπτικοποίησης δεδομένων και των εργαλείων Python, ενώ παράλληλα ενισχύθηκαν οι ερευνητικές ικανότητες μέσω της ανάπτυξης μεθοδολογίας για την οργάνωση και παρουσίαση τεχνικού υλικού.

Επιπλέον, η εμπλοκή με την παιδαγωγική διάσταση του έργου οδήγησε σε κατανόηση των αρχών της αποτελεσματικής διδασκαλίας τεχνολογικών θεμάτων και ανάπτυξη ευαισθησίας στις ανάγκες των μαθητευόμενων. Τέλος, η ολοκλήρωση ενός τόσο μακροπρόθεσμου και πολύπλοκου έργου συνέβαλε στην ανάπτυξη υπομονής, επιμονής και της ικανότητας διαχείρισης μεγάλων έργων με πολλαπλές συνιστώσες.

Η ολοκλήρωση αυτής της διπλωματικής εργασίας απέδειξε ότι η δημιουργία ποιοτικού εκπαιδευτικού υλικού για τεχνολογικά θέματα απαιτεί συνδυασμό τεχνικής γνώσης, παιδαγωγικής προσέγγισης και πρακτικής εμπειρίας. Το Jupyter Notebook αποτελεί ιδανικό εργαλείο για τη διδασκαλία data science, ενώ η Seaborn παραμένει μία από τις πιο προσιτές και ισχυρές βιβλιοθήκες οπτικοποίησης. Το παραγόμενο υλικό συνεισφέρει στην εκπαιδευτική κοινότητα παρέχοντας έναν ολοκληρωμένο, δομημένο και άμεσα χρήσιμο οδηγό που μπορεί να χρησιμοποιηθεί τόσο για αυτόνομη μελέτη όσο και ως βοηθητικό υλικό σε ακαδημαϊκά και επαγγελματικά περιβάλλοντα.

Βιβλιογραφία

Η βιβλιογραφία και τα δεδομένα που ακολουθούν αφορούν τόσο το παρών κείμενο όσο και τον οδηγό μελέτης.

GeeksforGeeks. (n.d.). *Box plot visualization with Pandas and Seaborn*. GeeksforGeeks. <https://www.geeksforgeeks.org/box-plot-visualization-with-pandas-and-seaborn/>

Google. (2025). Google Colaboratory: A hosted Jupyter notebook service. [Online]. Διαθέσιμο στο: <https://colab.research.google.com/> [Ημερομηνία πρόσβασης: 29 Απριλίου 2025]

McKinney, W. & pandas development team. (2025). pandas: powerful Python data analysis toolkit (Έκδοση 2.2.3) [Computer software]. <https://pandas.pydata.org/>

Mulani, S. (2022,). *Seaborn Distplot: A Comprehensive Guide*. DigitalOcean. <https://www.digitalocean.com/community/tutorials/seaborn-distplot>

Seaborn development team. (2024). Function categorization diagram. *seaborn documentation*. [Online]. Διαθέσιμο στο: https://seaborn.pydata.org/tutorial/function_overview.html [Ημερομηνία πρόσβασης: 29 Απριλίου 2025]

Seaborn development team. (χ.χ.). *Statistical estimation and error bars*. PyData. Ανακτήθηκε από https://seaborn.pydata.org/tutorial/error_bars.html

Stack Overflow. (2021, Απριλίου 28). *Python 3.x - What estimators does seaborn support* [Δημοσίευση σε online φόρουμ]. <https://stackoverflow.com/questions/67305620/what-estimators-does-seaborn-support>

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media [Online]. Διαθέσιμο στο: <https://jakevdp.github.io/PythonDataScienceHandbook/> [Ημερομηνία πρόσβασης: 29 Απριλίου 2025]

Yi, M. (χ.χ.). *Box Plot: The complete guide*. Atlassian. Ανακτήθηκε στις 29 Απριλίου 2025 από <https://www.atlassian.com/data/charts/box-plot-complete-guide>

Δεδομένα που χρησιμοποιούνται

UNICEF Data Warehouse. *Data on Life Expectancy (DM_LIFE_EXP)*. Ανακτήθηκε από https://sdmx.data.unicef.org/ws/public/sdmxapi/rest/data/UNICEF,DM,1.0/.DM_LIFE_EXP..?format=csv

Waskom, M. (2025). Mpg dataset. In *seaborn-data: Data repository for seaborn examples*. [Online]. Διαθέσιμο στο: <https://github.com/mwaskom/seaborn-data>
[seaborn.load_dataset — seaborn 0.13.2 documentation](https://seaborn.pydata.org/seaborn-load_dataset.html) [Ημερομηνία πρόσβασης: 29 Απριλίου 2025]

Waskom, M. (2025). Penguins dataset. In *seaborn-data: Data repository for seaborn examples*. [Online]. Διαθέσιμο στο: https://github.com/mwaskom/seaborn-data/seaborn.load_dataset — seaborn 0.13.2 documentation [Ημερομηνία πρόσβασης: 29 Απριλίου 2025]

Waskom, M. (2025). Tips dataset. In *seaborn-data: Data repository for seaborn examples*. [Online]. Διαθέσιμο στο: https://github.com/mwaskom/seaborn-data/seaborn.load_dataset — seaborn 0.13.2 documentation [Ημερομηνία πρόσβασης: 29 Απριλίου 2025]

Waskom, M. (2025). Titanic dataset. In *seaborn-data: Data repository for seaborn examples*. [Online]. Διαθέσιμο στο: https://github.com/mwaskom/seaborn-data/seaborn.load_dataset — seaborn 0.13.2 documentation [Ημερομηνία πρόσβασης: 29 Απριλίου 2025]