Ε Θ Ν Ι Κ Ο Μ Ε Τ Σ Ο Β Ι Ο Π Ο Λ Υ Τ Ε Χ Ν Ε Ι Ο
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Τομέας Μαθηματικών
Πολυτεχνειούπολη – Ζωγράφου ΑΘΗΝΑ - 157 80
ΤΗΛ. : 772 1774
FAX : 772 1775

ΜΑΘΗΜΑ: *Ανάλυση Δεδομένων με Η/Υ (6ο εξάμηνο)*
ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΕΡΓΑΣΙΑ 1η

Θέμα Εργασίας: Εισαγωγή στην R και Περιγραφική Στατιστική

Άσκηση 1

Μια ερευνήτρια θέλει να εξετάσει τους παράγοντες που επηρεάζουν τον χρόνο που περνούν οι ενήλικες στο τηλέφωνό τους ανά ημέρα. Μέσω ενός ερωτηματολογίου, συλλέχθηκαν πληροφορίες από 70 διαφορετικούς ενήλικες. Οι μεταβλητές που συλλέχθηκαν στο δείγμα για κάθε ενήλικα είναι οι εξής: (α) η ηλικία (**Age**) σε έτη, (β) το φύλο **(Gender)** οπού παρατηρήθηκαν οι κλάσεις "M" (αρσενικό) και "F" (θηλυκό), (γ) το λογισμικό που χρησιμοποιούν **(Operating_System)** στο κινητό τους τηλέφωνο, οπού παρατηρήθηκαν οι κλάσεις "IOS" και "Android", (δ) ο τύπος εργασίας τους **(Occupation)** χωρισμένος σε κλάσεις "Self_Employed" (αυτοαπασχολούμενος), "Wage Worker" (μισθωτός) και "Other" (άλλο), (ε) η ηλικία του κινητού τους σε ημέρες **(Phone_Age_in_Days)**, (στ) ο χρόνος σε ώρες που περάσαν στο κινητό για λόγους ευχαρίστησης (μη εργασιακούς) **(Screen_Time_Leasure)** την προηγούμενη μέρα, (ζ) ο χρόνος σε ώρες που περάσαν στο κινητό για λόγους εργασιακούς **(Screen_Time_Business)** την προηγούμενη μέρα.

Τα δεδομένα βρίσκονται στο αρχείο:

http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/Phone.txt

Στα δεδομένα μας, με τον χαρακτήρα "$" συμβολίζουμε τις αγνοούμενες τιμές. Εισάγετε τα δεδομένα στην R με χρήση της εντολής `read.table`, αλλάζοντας κατάλληλα το σύμβολο για τις αγνοούμενες τιμές (για αυτόματη αλλαγή, ελέγξτε στο

μενού `help` της R τα δυνατά ορίσματα της εντολής `read.table`). Ποια είναι η δομή του αντικειμένου που δημιουργείται από την παραπάνω εντολή; Στη συνέχεια αφαιρέστε οποιαδήποτε γραμμή περιέχει αγνοούμενη τιμή.

i. Δώστε μια περιγραφική ανάλυση, για τις τιμές ή κλάσεις (κατηγορίες) κάθε μίας από τις 7 μεταβλητές ξεχωριστά, η οποία να αποτελείται από κατάλληλες αριθμητικές και γραφικές μεθόδους και σχολιάστε τα ευρήματα σας.
ii. Με τη βοήθεια κατάλληλου διαγράμματος εξετάστε περιγραφικά αν ο χρόνος που περάσαν οι ενήλικες του δείγματος στο κινητό για λόγους ευχαρίστησης (μη εργασιακούς) **(Screen_Time_Leasure)** διαφοροποιείται ανάλογα με το λογισμικό (**Operating_System**). Υλοποιήστε παρόμοιες συγκρίσεις, με χρήση διαγραμμάτων, μεταξύ των τιμών της μεταβλητής **Screen_Time_Leasure** και των τιμών ή κλάσεων καθεμιάς από τις υπόλοιπες μεταβλητές. Τι συμπεραίνετε;
iii. Να κατασκευαστεί ο πίνακας συχνοτήτων και σχετικών συχνοτήτων για τα δεδομένα που αφορούν την ηλικία των ατόμων που συμμετέχουν στην έρευνα με τη χρήση 3 κλάσεων: [18-30), [30-50), [50 και άνω). Δώστε κατάλληλα ονόματα στις κατηγορίες της νέας αυτής μεταβλητής την οποία ονομάστε **f_age**. Εν συνεχεία, κατασκευάστε μια ακόμα κατηγορική μεταβλητή, με όνομα **f_PA** και κλάσεις $[0,q_1)$, $[q_1,q_2)$, $[q_2,q_3)$, $[q_3$ και άνω), όπου q_i $(i = 1, 2, 3)$ είναι το i-στο τεταρτημόριο των τιμών της μεταβλητής **Phone_Age_in_Days**. Κατασκευάστε έναν δισδιάστατο πίνακα συχνοτήτων των μεταβλητών **f_PA** και **f_age** στο δείγμα. Δώστε τις σχετικές συχνότητες κελιών και σχολιάστε τα αποτελέσματα. Δημιουργείστε ένα στοιβαγμένο ραβδόγραμμα και σχολιάστε τα αποτελέσματα.

Άσκηση 2

α) Υπολογιστής Δείκτη Μάζας Σώματος (BMI)

Γράψτε μια συνάρτηση που υπολογίζει τον Δείκτη Μάζας Σώματος (BMI) ενός ατόμου και τον κατηγοριοποιεί ανάλογα με την τιμή του. Πιο συγκεκριμένα, ορίστε στην R μια συνάρτηση `bmi(weight, height)` που υπολογίζει το BMI χρησιμοποιώντας τον τύπο:

$$\text{BMI} = \frac{\text{weight(Kg)}}{\text{height(m)}^2},$$

όπου weight είναι το βάρος σε κιλά και height το ύψος σε μέτρα. Στη συνάρτησή σας, κατηγοριοποιήστε το BMI ως εξής: (α) **Underweight**: BMI < 18.5, (β) **Normal weight**: 18.5 ≤ BMI < 24.9, (γ) **Overweight**: 24.9 ≤ BMI < 29.9 και (δ) **Obese**: BMI ≥ 29.9. Η συνάρτηση πρέπει να επιστρέφει τόσο την τιμή του BMI όσο και την κατηγορία της. Δοκιμάστε τη συνάρτηση με διάφορες τιμές βάρους και ύψους, όπως (70, 1.75), (50, 1.60), και (90, 1.80).

β) Υπολογιστής Δόσης Δανείου

Γράψτε μια συνάρτηση στην R που υπολογίζει τη μηνιαία εξέλιξη αποπληρωμής ενός δανείου, δεδομένου του αρχικού κεφαλαίου (principal), του ετήσιου επιτοκίου (annual_rate) και της διάρκειας του δανείου σε χρόνια (years). Το ετήσιο επιτόκιο (annual_rate) να δίνεται στη συνάρτηση ως ένας θετικός αριθμός (π.χ. 3.5) όπου διαιρεμένος με το 100 θα επιστρέφει το ετήσιο ποσοστό επιτοκίου (π.χ. 3.5%). Ονομάστε τη συνάρτησή σας `loan_payment(principal, annual_rate, years)` και υπολογίστε τη μηνιαία δόση M χρησιμοποιώντας τον τύπο

$$M = P\frac{r(1+r)^n}{(1+r)^n - 1},$$

όπου P εκφράζει το αρχικό κεφάλαιο (principal), r εκφράζει το μηνιαίο ποσοστό επιτοκίου (δηλαδή (ετήσιο επιτόκιο (annual_rate)/100) / 12) και n εκφράζει τον συνολικό αριθμό μηνών (χρόνια (years) × 12) της διάρκειας του δανείου. Αφού υπολογίσετε το M στη συνέχεια για κάθε μήνα της διάρκειας του δανείου (ξεκινώντας από τον πρώτο και ολοκληρώνοντας στον τελευταίο μήνα n) θα πρέπει

i. Να υπολογίσετε το ποσό που πρέπει να καταβληθεί τον εν λόγω μήνα για τον τοκισμό (interest) του υπολειπόμενου κεφαλαίου (balance) του δανείου (τον πρώτο μήνα το υπολειπόμενο κεφάλαιο είναι το συνολικό αρχικό κεφάλαιο (principal)), πολλαπλασιάζοντας το μηνιαίο σταθερό ποσοστό επιτοκίου (r) με το υπολειπόμενο κεφάλαιο (balance).
ii. Να υπολογίσετε το κεφάλαιο αποπληρωμής (principal_paid) ως την μηνιαία δόση M μείον τον τοκισμό (interest).
iii. Να υπολογίσετε το νέο κεφάλαιο (balance) του δανείου το ήδη υπάρχων κεφάλαιο (balance) μείον το κεφάλαιο αποπληρωμής (principal_paid).

Η συνάρτηση πρέπει να επιστρέφει ένα πλαίσιο δεδομένων (`results`) με n γραμμές και δύο στήλες. Στην πρώτη στήλη θα αναγράφεται ο μήνας (ξεκινώντας από τον πρώτο) και στην δεύτερη το νέο κεφάλαιο (balance) όπως υπολογίστηκε στο iii. παραπάνω, κρατώντας δύο μόνο δεκαδικά ψηφία.

Δοκιμάστε τη συνάρτηση με διάφορες τιμές, όπως:

- `loan_payment(10000, 5, 2)` (Δάνειο €10,000, 5% ετήσιο επιτόκιο, 2 χρόνια)
- `loan_payment(50000, 3.5, 5)`

Οδηγίες

- Η εργασία θα πρέπει **να παραδοθεί ηλεκτρονικά** στον **Σωτήρη Ζαμπέλη** στο email του, szampelis.emp@gmail.com
- Η εργασία που θα παραδώσετε πρέπει να είναι **σε pdf μορφή**. Παρακαλώ χρησιμοποιήστε τον **ακόλουθο τίτλο στο pdf αρχείο σας**: Surname-Name-Ex1.pdf, όπου Surname είναι το επώνυμό σας (με λατινικούς χαρακτήρες) και Name το όνομα σας (με λατινικούς χαρακτήρες). Π.χ. αν παρέδιδα εγώ εργασία θα την ονόμαζα ως εξής: Fouskakis-Dimitris-Ex1.pdf.
- Παρακαλώ χρησιμοποιήστε **ένα εξώφυλλο στο pdf αρχείο σας**, στο οποίο να αναγράφεται ο τίτλος της εργασίας (Εισαγωγή στην R και Περιγραφική Στατιστική), **το ονοματεπώνυμό σας, το email σας, η Σχολή σας και ο αριθμός μητρώου σας**.
- Θα πρέπει να **αποστείλετε ένα μόνο αρχείο**. Η εργασία θα πρέπει να περιλαμβάνει τους κώδικες της R, όχι σε παράρτημα αλλά στην απάντηση του κάθε ερωτήματος.
- Η εργασία θα πρέπει να αποσταλεί **μέχρι την Παρασκευή 28/03/2025 στις 13:00. Καμιά εργασία δεν θα γίνει δεκτή μετά την ώρα αυτή.**
- Η εργασία θα πρέπει να είναι σε μορφή **αναφοράς** και να περιλαμβάνει τους κώδικες της R με πλήρη επεξήγηση, γραφήματα και πίνακες με κατάλληλους τίτλους και πλήρη επεξήγηση των αποτελεσμάτων. Επίσης, δε θα πρέπει να υπερβαίνει τις **15 σελίδες** με μέγεθος γραμματοσειράς **12**.

- Θα δοθεί ιδιαίτερη σημασία **στην παρουσίαση της εργασίας**. Η εργασία πρέπει να είναι κατανοητή και να περιγράφει οτιδήποτε χρησιμοποιήσατε πειστικά για κάποιον που δεν γνωρίζει πολλά για το αντικείμενο.

Εύχομαι Επιτυχία