ΜΑΘΗΜΑ: "Εισαγωγή στον Προγραμματισμό"

ΕΡΓΑΣΤΗΡΙΟ 13: ΕΠΑΝΑΛΗΠΤΙΚΕΣ ΑΣΚΗΣΕΙΣ ΣΤΗΝ R, ΜΕΡΟΣ ΙΙ

Άσκηση 1 (Παραλλαγή Ντζούφρας & Καρλής 2016, Άσκηση 6, σελ. 306): Ένας φοιτητής περνάει στο επόμενο εξάμηνο εάν

(α) Έχει σε όλα τα μαθήματα πάνω από τη βάση (50)

(β) Εάν έχει σε ένα κάτω από τη βάση, να σημειώνει μέσο όρο βαθμολογίας πάνω από 60

(γ) Εάν έχει δύο κάτω από τη βάση αλλά πάνω από 40, να σημειώνει μέσο όρο βαθμολογίας πάνω από 70

Να γράψετε μια έκφραση (συνάρτηση, ονομάστε τη myfun1) που να ελέγχει, αν η βαθμολογία ενός φοιτητή σε 4 μαθήματα τον προάγει στο επόμενο εξάμηνο. Να επιστρέφει σε λίστα το αποτέλεσμα PASS/FAIL, το κριτήριο που ικανοποιείται (σε περίπτωση προαγωγής), το διάνυσμα των βαθμών και το μέσο όρο τους.

Άσκηση 2 (Παραλλαγή Ντζούφρας & Καρλής 2016, Άσκηση 8, σελ. 307): Έστω ότι στα x και y αντίστοιχα, είναι αποθηκευμένοι οι βαθμοί στις εργασίες και τη γραπτή εξέταση ενός μαθήματος για έναν συγκεκριμένο φοιτητή. Οι βαθμοί είναι στην κλίμακα 0-10.

(α) Να δοθεί έκφραση στην R για τον υπολογισμό της τελικής βαθμολογίας του φοιτητή αν αυτή υπολογίζεται ως 0.5*ΕΡΓΑΣΙΑ + 0.5*(ΓΡΑΠΤΗ ΕΞΕΤΑΣΗ).

(β) Να δοθεί συνάρτηση στην R (ονομάστε τη myfun2) η οποία να έχει ως όρισμα τους βαθμούς σε εργασίες και γραπτή εξέταση (δηλ. τα x, y). Να επιστρέφει μήνυμα για την τελική βαθμολογία στο μάθημα σύμφωνα με τα παρακάτω κριτήρια: Αν κάποιος έχει βαθμό μικρότερο από 7 στις εργασίες, τότε αποτυγχάνει (FAIL) και ο τελικός βαθμός είναι 4. Αν ο βαθμός στις εργασίες είναι >8, τότε η τελική βαθμολογία είναι ο βαθμός της τελικής εξέτασης, προσαυξημένος κατά 1 μονάδα. Διαφορετικά, είναι βαθμός της τελικής εξέτασης.

(γ) Να δοθεί έκφραση στην R για τον υπολογισμό της τελικής βαθμολογίας του φοιτητή αν αυτή υπολογίζεται ως ο μικρότερος από τους βαθμούς της εργασίας και του γραπτού.

(δ) Να δοθεί συνάρτηση στην R (ονομάστε τη myfun3) η οποία να έχει ως όρισμα τους βαθμούς σε εργασίες και γραπτή εξέταση (δηλ. τα x, y). Να επιστρέφει μήνυμα για την τελική βαθμολογία στο μάθημα σύμφωνα με τα παρακάτω κριτήρια:

i. Αν οι εργασίες είχαν βαθμό τουλάχιστον 5, ο βαθμός του γραπτού και της εργασίας με ίσο βάρος, ενώ

ii. Αν οι εργασίες ήταν κάτω από 5 μόνο ο βαθμός του γραπτου.

Άσκηση 3 (Φουσκάκης 2013, Άσκηση 3.6, σελ. 110): Θεωρήστε τα δεδομένα Arthritis της βιβλιοθήκης vcd της R (με χρήση library(vcd), αφού πρώτα το εγκαταστήσετε). Πρόκειται για δεδομένα κλινικής δοκιμής με στόχο τη διερεύνηση της αποτελεσματικότητας τη νέας θεραπείας (φαρμάκου) για τη ρευματοειδή αρθρίτιδα. Οι μεταβλητές είναι οι εξής: ID = μητρώο ασθενούς, Treatment = είδος θεραπείας (Placebo: εικονικό φάρμακο, Treated: πραγματικό φάρμακο), Sex = φύλο (Female: γυναίκα, Male: άνδρας), Age = ηλικία (σε έτη) και Improved = ένδειξη αποτελεσματικότητας θεραπείας (None: καθόλου, Some: μερική, Marked: σημαντική).

i. Ποιο ποσοστό των ασθενών του δείγματος υποβλήθηκε πράγματι σε θεραπεία; Κατασκευάστε το κυκλικό διάγραμμα (pie chart) της μεταβλητής Treatment, δίνοντας κατάλληλο τίτλο.
ii. Κατασκευάστε τον πίνακα συνάφειας απόλυτων και σχετικών συχνοτήτων των μεταβλητών Treatment και Improved. Ποιο ποσοστό των ασθενών έδειξε σημαντική βελτίωση ανεξαρτήτως θεραπείας; Δώστε την εντολή στην R για την κατασκευή του κατάλληλου πίνακα (σχετικών συχνοτήτων).
iii. Κατασκευάστε το ραβδόγραμμα συχνοτήτων της μεταβλητής Improved, δίνοντας κατάλληλο τίτλο. *Homework:* Κατασκευάστε και το ραβδόγραμμα σχετικών συχνοτήτων της Improved.
iv. Ποιο το ποσοστό των ασθενών που δεν υποβλήθηκαν στη νέα θεραπεία και παρουσίασαν σημαντική βελτίωση;
v. Ποιο ποσοστό των ασθενών που δεν παρουσίασαν καμία βελτίωση, είχε υποβληθεί σε νέα θεραπεία; **Σημείωση:** Είναι δεδομένο/γνωστό το ότι δεν παρουσίασαν καμία βελτίωση.
vi. Κατασκευάστε ένα στοιβαγμένο ραβδόγραμμα συχνοτήτων της μεταβλητής Improved με τις στοίβες ορισμένες ως προς το είδος της θεραπείας με κατάλληλη λεζάντα.
vii. Κατασκευάστε το ομαδοποιημένο ραβδόγραμμα συχνοτήτων και σχετικών συχνοτήτων της μεταβλητής Improved δοθέντος του φύλου των ασθενών με κατάλληλη λεζάντα.

Άσκηση 4 (Φουσκάκης 2013, Άσκηση 3.9, σελ. 112): Θεωρήστε τα δεδομένα anorexia της βιβλιοθήκης MASS της R (με χρήση library(MASS)). Τα δεδομένα αφορούν 72 νεαρές γυναίκες που έπασχαν από νευρική ανορεξία και περιέχουν τιμές για τρεις μεταβλητές: το είδος θεραπείας (Treat) που ακολούθησαν (Cont: καμία θεραπεία, CBT: γνωσιακή ψυχοθεραπεία και FT: θεραπεία βασισμένη στην οικογενειακή παρέμβαση), το βάρος των γυναικών πριν τη θεραπεία (Prewt) και το βάρος των γυναικών μετά τη θεραπεία (Postwt) σε λίβρες (lb).

i. Αφού διαιρέσετε τις τιμές των βαρών με το 2.2 ώστε να μετατρέψετε τις λίβρες σε κιλά, συγκρίνετε το εύρος του βάρους των γυναικών πριν και μετά τη θεραπεία.
ii. Δημιουργήστε μια νέα μεταβλητή με όνομα Changewt για τη μεταβολή του βάρους των γυναικών μετά το τέλος της θεραπείας σε σχέση με πριν. Ποια είναι η μέση και ποια η διάμεση

μεταβολή βάρους των γυναικών του δείγματος; Ποια είναι η τυπική απόκλιση της μεταβολής βάρους στο δείγμα;

iii. Ποιες είναι οι τιμές της μεταβολής βάρους που αντιστοιχούν στο κάτω 25% και στο άνω 25% των γυναικών του δείγματος; Προσπαθήστε να τις ερμηνεύσετε.

iv. Δώστε τις απόλυτες και τις σχετικές συχνότητες των διαφόρων ειδών θεραπείας που ακολούθησαν οι γυναίκες του δείγματος;

v. Κατασκευάστε σε ένα παράθυρο τα θηκοδιαγράμματα του βάρους των γυναικών πριν και μετά τη θεραπεία δίνοντας κατάλληλα ονόματα.

vi. Κατασκευάστε στο ίδιο παράθυρο το θηκόγραμμα της μεταβολής βάρους ξεχωριστά για κάθε είδος θεραπείας. Ποιο είδος θεραπείας φαίνεται να είναι πιο αποτελεσματικό;

Άσκηση 5 (Παραλλαγή Φουσκάκης 2013, Άσκηση 3.11 + 4.16, σελ. 113, 154): Θεωρήστε τα δεδομένα του ακόλουθου συνδέσμου http://www.math.ntua.gr/~fouskakis/Rbook/carfuel.txt (υπάρχουν και στο eclass του μαθήατος).

Τα δεδομένα αφορούν μετρήσεις της κατανάλωσης καυσίμου κίνησης. Συγκεκριμένα, το τυχαίο δείγμα αποτελείται από 150 αυτοκίνητα επτά γνωστών εταιρειών κατασκευής τα οποία σχεδιάστηκαν και προωθήθηκαν στην αγορά το έτος 1999 ή το έτος 2008. Συλλέχθηκαν οι ακόλουθες μεταβλητές: Manufacturer = κατασκευαστική εταιρεία (1: Audi, 2: Chevrolet, 4: Ford, 6: Huyndai, 11: Nissan, 14: Toyota, 15: Volkswagen), Engine_size = εκτόπισμα μηχανής (σε λίτρα – lt), Year = έτος σχεδίασης μοντέλου (1999 ή 2008), Cylinders = αριθμός κυλίνδρων οχήματος (4, 6, 8) και City_mpg = κατανάλωση καυσίμου εντός πόλης (σε μίλια ανά γαλόνι – mpg).

i. Στα δεδομένα, με τον χαρακτήρα '*' συμβολίζονται οι ελλιπείς (ελλείπουσες) τιμές. Εισάγετε τα δεδομένα την R, αλλάζοντας κατάλληλα το σύμβολο για τις ελλιπείς τιμές (για την αυτόματη αλλαγή, ελέγξτε στο μενού help της R τα ορίσματα της εντολής scan()) και δημιουργήστε ένα πλαίσιο δεδομένων δίνοντας ονόματα στις πέντε μεταβλητές.

ii. Η μεταβλητή Manufacturer είναι στην πραγματικότητα κατηγορική. Μετατρέψτε τη μεταβλητή σε κατηγορική χρησιμοποιώντας την εντολή as.factor() και στη συνέχεια αντικαταστήστε τις αριθμητικές τιμές με τις αντίστοιχες ονομασίες (εταιρίες κατασκευής) των κατηγοριών. Όμοια, μετατρέψτε τη μεταβλητή Year σε κατηγορική.

iii. Δώστε μια περιγραφική (στατιστική) ανάλυση για κάθε μεταβλητή ξεχωριστά, η οποία να αποτελείται από κατάλληλες αριθμητικές και γραφικές μεθόδους. Σχολιάστε (εν συντομία) τα ευρήματά σας.

iv. Με τη βοήθεια κατάλληλων γραφημάτων, δείτε αν στο δείγμα οι τιμές της κατανάλωσης καυσίμου εντός πόλης διαφοροποιούνται ανάλογα με την κατασκευάστρια εταιρεία, το έτος σχεδίασης και τον αριθμό κυλίνδρων του αυτοκινήτου. Χωριστά διαγράμματα για κάθε μεταβλητή. Να δοθούν ονόματα στους άξονες, στις μεταβλητές.

v. Να κατασκευαστεί ο πίνακας συχνοτήτων και σχετικών συχνοτήτων για τα δεδομένα που αφορούν το εκτόπισμα της μηχανής (Engine_size) με τη χρήση 3 κλάσεων, όπου μηχανές μικρότερες από 2.5 lt θεωρούνται ελαφρές (light), μηχανές από 2.5 lt έως και 3.9 lt θεωρούνται κανονικές (normal) και μηχανές από 4 lt και άνω θεωρούνται μεγάλες (large). Δώστε αντίστοιχα ονόματα στις κατηγορίες της νέας αυτής κατηγορικής μεταβλητής. Στη συνέχεια, περιγράψτε κατάλληλα τη μεταβλητή City_mpg ξεχωριστά για κάθε κατηγορία εκτοπίσματος της μηχανής (συμπεριλάβετε γραφικές και αριθμητικές μεθόδους). Συγκρίνετε και σχολιάστε τα αποτελέσματά σας.

vi. Κατασκευάστε τον πίνακα συνάφειας της μεταβλητής Cylinders (γραμμές) και της νέας κατηγορικής μεταβλητής για το εκτόπισμα της μηχανής που δημιουργήσατε στο προηγούμενο ερώτημα (στήλες). Δώστε τις σχετικές συχνότητες κελιών, γραμμών και στηλών και σχολιάστε τα αποτελέσματα. Επίσης, κατασκευάστε το στοιβαγμένο ραβδόγραμμα συχνοτήτων με βάση τον προηγούμενο πίνακα συχνοτήτων κατά στήλες και δώστε κατάλληλη λεζάντα.

vii. (Παλινδρόμηση με Πίνακες) Ορίστε τον παρακάτω πίνακα A: Η $1^{η}$ στήλη να είναι μια στήλη μόνο με μονάδες, η $2^{η}$ να είναι η στήλη που αντιστοιχεί στη μεταβλητή Citympg και η $3^{η}$ να είναι η στήλη που αντιστοιχεί στη μεταβλητή EngineSize. Προσοχή, να φροντίσετε να μην υπάρχουν τιμές NA στον πίνακα που φτιάχνετε (και προφανώς οι γραμμές του πίνακα θα είναι τόσες, όσες και το πλήθος των παρατηρήσεων χωρίς NA). Στη συνέχεια, να κάνετε την πράξη (πινάκων)

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

όπου ο πίνακας $\mathbf{X}$ είναι οι 2 πρώτες στήλες του A και το διάνυσμα $\mathbf{y}$ είναι η $3^{η}$ στήλη. Το αποτέλεσμα να είναι διάνυσμα (έστω $\mathbf{b}$). Χρησιμοποιήστε τις τιμές του διανύσματος b και υπολογίστε τις τιμές $b[1] + b[2] \cdot$ Citympg. Να τις καταχωρίσετε σε ένα διάνυσμα $\mathbf{y1}$. Φτιάξτε το διάγραμμα διασποράς των μεταβλητών Citympg (άξονας x), EngineSize (άξονας y). Απεικονίστε στο ίδιο διάγραμμα την ευθεία $b[1] + b[2] \cdot$ Citympg (να είναι μπλε χρώματος). Δοκιμάστε να ξεχωρίσετε τα σημεία (δίνοντας) διαφορετικό χρώμα ως προς το εκτόπισμα της μηχανής (κατηγορική μεταβλητή EngineSizeCat). Να υπάρχει και κατάλληλη λεζάντα.

viii. Βρείτε τις διαφορές EngineSize – $\mathbf{y1}$ και να τις καταχωρίσετε σε ένα διάνυσμα y2. Στη συνέχεια να κάνετε το διάγραμμα διασποράς των τιμών y1 (οριζόντιος άξονας) και y2 (κάθετος άξονας). Να βάλετε και την ευθεία y=0 (με διακεκομμένη γραμμή) στο ίδιο γράφημα.

ix. Να γραφεί συνάρτηση στην R (ονομάστε τη myfun4) η οποία να υπολογίζει την παρακάτω ποσότητα

$$\frac{1}{(n-k)s^2}\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t+k}-\bar{x})$$

όπου $\bar{x}, s^2$ είναι η δειγματική μέση τιμή και η δειγματική διασπορά ενός διανύσματος με τιμές $x_1, x_2, \dots, x_n$. Στη συνέχεια να εφαρμόσετε την παραπάνω συνάρτηση στις τιμές του διανύσματος y2 (δείτε προηγούμενο ερώτημα) και υπολογίστε τις τιμές της για $k = 1,2, \dots .20$. Να τις απεικονίσετε στο επίπεδο (χρησιμοποιήστε την plot) με type="h". Στο γράφημα να τοποθετηθούν διακεκομμένες μπλε γραμμές στις τιμές $-\frac{1}{n} \pm \frac{2}{\sqrt{n}}$.. Εδώ το n θα είναι το πλήθος του διανύσματος y2.