Εργαστήριο Ανώτερης Γεωδαισίας
Μεταπτυχιακό Πρόγραμμα ΓΕΩΠΛΗΡΟΦΟΡΙΚΗΣ
«Αναλυτικές Μέθοδοι στη Γεωπληροφορική»
(Ακαδ. Έτος 2022-23)

ΟΝΟΜΑΤΕΠΩΝΥΜΟ .....

ΕΞΑΜΗΝΟ .....

Ημερομηνία Παράδοσης **: 18/1/2023**

Αρχεία δεδομένων που χρειάζονται για τους σκοπούς της Θεματικής Εργασίας και ζητείται να ανακτηθούν από τις ιστοσελίδες του μαθήματος, θα βρίσκονται εντός του ντοσιέ δεδομένων ***'askdata'***, στο σύνδεσμο http://portal.survey.ntua.gr/main/labs/hgeod/ddeli/analmgeo/askdata/ , π.χ. ένα αρχείο ***testdata.txt*** θα μπορεί να ανακτηθεί απευθείας από το σύνδεσμο http://portal.survey.ntua.gr/main/labs/hgeod/ddeli/analmgeo/askdata/testdata.txt.

ΘΕΜΑΤΙΚΗ ΕΡΓΑΣΙΑ #5

Σκοπός: Η θεματική εργασία #5 αποσκοπεί στην εξοικείωση σας με την πρακτική χρήση του R για δύο βασικές ανάγκες:

(a) Τον υπολογισμό βασικών στατιστικών μέτρων που χρησιμοποιούνται για να συνοψίσουν ένα σύνολο (συνήθως μεγάλου αριθμού) παρατηρήσεων, προκειμένου να αντληθούν από τα δεδομένα όσο το δυνατόν απλούστερες πληροφορίες, τόσο από αριθμητικά δεδομένα, όσο και από κατάλληλα γραφήματα.

(b) Στη χρήση των εργαλείων που προσφέρει το λογισμικό R για τη διεξαγωγή στατιστικών δοκιμών που αποσκοπούν στον έλεγχο υποθέσεων ή/και τη διερεύνηση της στατιστικής αξιοπιστίας υπό διερεύνηση παραμέτρων ενδιαφέροντος.

1. Στη Στατιστική ένα δείγμα δεδομένων καλείται **ποιοτικό** ή **κατηγορικό** , αν οι τιμές του ανήκουν σε μια συλλογή γνωστών καθορισμένων κατηγοριών που δεν αλληλεπικαλύπτονται.

Στην ιστοσελίδα του μαθήματος βρίσκεται το αρχείο ***NA_Border_Crossings_Entry_Data.csv*** ή εναλλακτικά, σε μορφή φύλλου excel το αρχείο ***NA_Border_Crossings_Entry_Data.xls*** που και τα δύο περιέχουν τα ίδια δεδομένα από μια συστηματική μελέτη, η οποία ανανεώνεται κάθε τρίμηνο από την Υπηρεσία Τελωνείων και Προστασίας Συνόρων των ΗΠΑ και αφορά τη διερχόμενη κίνηση μεταφορικών μέσων και ατόμων στους συνοριακούς σταθμούς των ΗΠΑ με το Μεξικό και τον Καναδά. Ειδικότερα στην ποιοτική μεταβλητή ***'measure'*** περιέχονται οι ακόλουθοι τύποι πληροφοριών που φαίνονται στην παραπλεύρως λίστα 12 ποιοτικών χαρακτηριστικών για τις συνοριακές γραμμές***US-Canada*** και ***US-Mexico***. Σημειώστε ότι σε αυτό το σύνολο δεδομένων δεν μετράται ο αριθμός των μοναδικών οχημάτων, επιβατών ή πεζών, αλλά απλά καταγράφεται ο αριθμός για τις εισερχόμενες διελεύσεις (π.χ., το ίδιο φορτηγό μπορεί να πηγαινοέρχεται στα σύνορα πολλές φορές την ημέρα, αλλά θα συλλέγονται δεδομένα για κάθε φορά).

Personal Vehicle Passengers
Personal Vehicles
Trucks
Bus Passengers
Buses
Rail Containers Empty
Truck Containers Empty
Rail Containers Full
Truck Containers Full
Pedestrians
Train Passengers
Trains

1. Φορτώστε το χώρο εργασία σας του R, ένα από τα παραπάνω αναφερόμενα αρχεία δεδομένων και εκχωρήστε τα περιεχόμενα στοιχεία σε ένα αντικείμενο με την ονομασία ***border_crossing***.

Αρχικά, χρησιμοποιήστε τη βασική συνάρτηση **View()** του R για την προβολής των δεδομένων του αρχείου σε στυλ υπολογιστικού φύλλου. Ακολούθως, εξακριβώστε ότι τα δεδομένα αποθηκεύθηκαν στο R ως πλαίσιο δεδομένων και εκτυπώστε τη δομή του αποθηκευμένου αντικειμένου, καθώς και λίγες σειρές στην αρχή και στο τέλος του αρχείου των δεδομένων. Επιβεβαιώστε ότι στα δεδομένα περιέχονται 65535 καταγραφές (γραμμές στο πλαίσιο δεδομένων ***border_crossing***) σε οκτώ στήλες στοιχείων. Επιπλέον, χρησιμοποιώντας τη συνάρτηση ***summary()***, εκτυπώστε συνοπτικά στατιστικά μέτρα που αφορούν τα στοιχεία σε κάθε στήλη του πλαισίου των δεδομένων.

Από την απλή επιθεώρηση των εκτυπωμένων συνοπτικών δεδομένων (π.χ., στη στήλη ***Value***), παρατηρήστε ότι σε κάποιους συνοριακούς σταθμούς δεν υπάρχουν καταγραφές για όλους τους τύπους διελεύσεων. Οι εμφανιζόμενες μηδενικές καταγραφές υποδηλώνουν ότι εάν προσπαθήσετε να υπολογίσετε (π.χ., μέσω των συναρτήσεων ***min(), max(), median(), mean()***, …) στατιστικά μέτρα των διελεύσεων από όλους τους συνοριακούς σταθμούς για όλη τη χρονική περίοδο που καλύπτουν τα διαθέσιμα δεδομένα στο αντικείμενο ***border_crossing*** τα αποτελέσματα θα είναι λανθασμένα, εκτός εάν προηγουμένως, με κατάλληλες εντολές, παραλείψετε στους υπολογισμούς τις τυχόν μηδενικές τιμές.

Επίσης παρατηρήστε ότι τα κελιά με την ονομασία ***Location*** εμπεριέχουν συνδυαστικά τόσο την τοποθεσία, όσο και τις γεωγραφικές συντεταγμένες του εκάστοτε συνοριακού σταθμού.

Για να έχετε ομοιομορφία στις ετικέτες (ονομασίες) των στηλών του πλαισίου δεδομένων ***border_crossing*** συνιστάται να αλλάξετε τις ονομασίες `Port Name`, `State Abbreviation` και `Port Code` αντίστοιχα σε `PortName`, `StateAbbreviation` και `PortCode`.

Η συνάρτηση separate() από το πακέτο *tidyr* του R μπορεί να χρησιμοποιηθεί για τον διαχωρισμό των στοιχείων μιας στήλης ενός πλαισίου ή ενός πίνακα δεδομένων σε διαφορετικές πολλαπλές στήλες σύμφωνα με τη εντολή της μορφής:

separate(data, col, into, sep = "[^[:alnum:]]+", remove = TRUE, convert = FALSE)

Οι ακόλουθες συναρτήσεις

substr(text, start, stop)
nchar(x, type = "chars", allowNA = FALSE, keepNA = NA)

επιτρέπουν αντίστοιχα την εξαγωγή των απαιτούμενων χαρακτήρων από μια συμβολοσειρά ή/και να αντικατασταθούν συγκεκριμένες τιμές σε μια συμβολοσειρά

Για τις λεπτομέρειες συμβουλευτείτε τη βοήθεια του R.

2. Επιθεωρήστε/εκτυπώστε μερικές γραμμές από τα στοιχεία της τελευταίας στήλης (με την ετικέτα ***Location***) του πλαισίου ***border_crossing*** και παρατηρήστε ότι κάθε καταγραφή περιλαμβάνει σε ξεχωριστές υπο-γραμμές (συμβολικά ***\r\n***) το όνομα της πολιτείας και, τις συντεταγμένες γεωγραφικού πλάτους και μήκους του συνοριακού σταθμού. Συγκεκριμένα, τα στοιχεία της στήλης θα πρέπει να είναι χαρακτήρες στη μορφή π.χ., ***"NEW YORK\r\n(44.35, -75.98)"***.

Δημιουργήστε ένα νέο πλαίσιο δεδομένων με την ονομασία ***border.crossing.location*** που να περιλαμβάνει όλα τα υπόλοιπα στοιχεία του αντικειμένου ***border_crossing*** εκτός της αρχικής στήλης με την ονομασία ***Location***, η οποία αντίστοιχα να αντικατασταθεί στο αντικείμενο ***border.crossing.location*** με τρεις νέες στήλες, με τις ονομασίες ***"State", "Latıtude"*** και ***"Longitude"***. Για αυτό το σκοπό θα χρειαστείτε να κάνετε χρήση των συναρτήσεων ***separate()*** από το πακέτο ***tidyr*** και τις ενσωματωμένες συναρτήσεις του R ***substring()***, και ***nchar()***. Επιβεβαιώστε τη νέα δομή των δεδομένων εκτυπώνοντας μερικές από τις σειρές των δεδομένων στην αρχή και στο τέλος του νέου αντικειμένου ***border.crossing.location***. Εάν οι ζητούμενες στήλες δημιουργήθηκαν σωστά θα περιέχουν στοιχεία της μορφής

```
…$ State          : chr  "ALASKA" "ALASKA" "ALASKA" "ALASKA" ...
…$ Latitude       : chr  "62.61" "62.61" "62.61" "62.61" ...
…$ Longitude      : chr  "-141.0" "-141.0" "-141.0" "-141.0" ...
```

Αντιγράψτε το αρχικό πλαίσιο δεδομένων ***border.crossing.location*** σε ένα νέο αντικείμενο με την ονομασία ***border.crossing.datestamp***. Ακολούθως, στο ***border.crossing.datestamp*** αντικαταστήστε τα στοιχεία του στη στήλη Date, από τη μορφή ***01-01-99 0:00*** *(μήνα-μέρα-έτος με 2 ψηφία και ώρες:λεπτά)* στη μορφή ***1999-01-01*** *(έτος με 4 ψηφία-μήνας-μέρα)*. Για το σκοπό αυτό μπορείτε να χρησιμοποιήσετε τη συνάρτηση ***as.Date()*** με μια εντολή της μορφής

... as.Date(..., format = "%m-%d-%y")

Επιθεωρήστε/εκτυπώστε μερικές γραμμές από τα στοιχεία του πλαισίου ***border.crossing.datestamp*** και παρατηρήστε ότι οι καταγραφές στη στήλη ***Date*** είναι στην προαναφερόμενη επιθυμητή μορφή *YYYY-MM-DD*.

Η συνάρτηση filter() από το πακέτο *dplyr* του R μπορεί να χρησιμοποιηθεί για την παραγωγή (φιλτράρισμα) ενός υποσυνόλου ενός πλαισίου δεδομένων (dataframe), διατηρώντας όλες τις σειρές που ικανοποιούν τις καθορισμένες συνθήκες (conditions):

είτε απευθείας με εντολές της μορφής

- ***filter(dataframe, conditions)***

είτε με εντολές της μορφής

- ***dataframe %>% filter (conditions)***

όπου οι καθορισμένες συνθήκες μπορούν να περιλαμβάνουν τελεστές σύγκρισης (==, >, >=) , λογικούς τελεστές (&, |, !, xor()) , τελεστές περιοχής (*between(), near()*) καθώς και έλεγχο τιμών NA έναντι των τιμών μιας στήλης. Η παραπάνω αναφερόμενη 2η προσέγγιση προτιμάται συνήθως για χρήση με ένα μεγάλο σύνολο συνθηκών, όπου ο τελεστής %>% παρέχει έναν μηχανισμό για την προώθηση μιας αλυσίδας εντολών ή το αποτέλεσμα μιας έκφρασης, στην επόμενη κλήση/έκφραση μιας συνάρτησης.

Για τις λεπτομέρειες συμβουλευτείτε τη βοήθεια του R.

3. Δημιουργήστε ένα αντίγραφο του αντικειμένου ***border.crossing.datestamp*** με την ονομασία ***border.crossing.clean*** το οποίο να περιλαμβάνει μόνο τις σειρές που περιέχουν μη μηδενικές τιμές στα κελιά της στήλης *"Value"* (δηλ. οι καταγραφές των διελεύσεων).

   Για αυτό το σκοπό θα χρειαστείτε να κάνετε χρήση της συνάρτησης ***filter()*** από το πακέτο ***dplyr***, η οποία εκχωρεί σε ένα νέο αντικείμενο το υποσύνολο ενός πλαισίου δεδομένων διατηρώντας όλες τις σειρές που ικανοποιούν κάποιες καθορισμένες (από το χρήστη) προϋποθέσεις (π.χ., εν προκειμένω οι καταγραφές να αφορούν μη μηδενικές τιμές στη συγκεκριμένη στήλη *Value*).

   Εξακριβώστε τις νέες διαστάσεις (γραμμές και στήλες) του νέου πλαισίου των δεδομένων και εκτυπώστε συνοπτικά στατιστικά μέτρα που αφορούν τα στοιχεία σε κάθε στήλη του νέου αυτού πλαισίου των δεδομένων.

Από το αντικείμενο ***border.crossing.clean*** δημιουργήστε ένα νέο πλαίσιο δεδομένων με την ονομασία ***canada.crossings*** που να περιλαμβάνει μόνο όλες τις καταγραφές που αφορούν δεδομένα από σταθμούς της συνοριακής γραμμής *US-Canada* (και, εν προκειμένω, θα περιέχει μόνο τις σειρές που περιέχουν μη μηδενικές τιμές στα κελιά της στήλης **"Value"**).

Για τη συγκεκριμένη συνοριακή γραμμή υπολογίστε τυπικά στατιστικά μέτρα (π.χ., ελάχιστη, μέγιστη, μέση, διάμεση τιμή) τόσο (i) για όλους τους τύπους διελεύσεων, όσο και (ii) για μόνο επιβατικά οχήματα *(Personal Vehicles)* και (iii) για μόνο τραίνα πλήρως φορτωμένα *(Rail Containers Full)*.

Με παρόμοιο τρόπο, από το αντικείμενο ***border.crossing.clean*** δημιουργήστε ένα νέο πλαίσιο δεδομένων με την ονομασία ***mexico.crossings*** που επίσης να περιλαμβάνει μόνο όλες τις μη μηδενικές καταγραφές που αφορούν δεδομένα από σταθμούς της συνοριακής γραμμής *US-Mexico*. Για τη συγκεκριμένη συνοριακή γραμμή υπολογίστε τυπικά στατιστικά μέτρα συνολικά *για όλους τους τύπους διελεύσεων με τραίνα* (δηλ. του τύπου ***Rail Containers Full*** *και* ***Rail Containers Empty*** *και* ***Train Passengers*** *και* ***Trains***).

Με το ίδιο σκεπτικό θεωρείστε ένα αντικείμενο ***canada.border.railCE*** που να περιλαμβάνει μόνο τις καταγραφές που αφορούν δεδομένα από σταθμούς της συνοριακής γραμμής *US-Canada* και μόνο για μη μηδενικές *διελεύσεις τραίνων με κενά βαγόνια* (δηλ. του τύπου ***Rail Containers Empty***). Ακολούθως υπολογίστε για τη συγκεκριμένη κατηγορία διελεύσεων το σύνολο των καταγραφών (ασχέτως συνοριακού σταθμού) και για όλες τις περιπτώσεις που για το σύνολο των συνοριακών σταθμών (α) παρατηρήθηκαν λιγότερες από 20 διελεύσεις τραίνων με κενά βαγόνια και (β) παρατηρήθηκαν αντίστοιχα περισσότερες από 999 τέτοιες διελεύσεις.

Δημιουργήστε ένα αντικείμενο δεδομένων με την ονομασία ***mexico.crossings.pedestrians*** που να περιλαμβάνει μόνο τις καταγραφές που αφορούν δεδομένα από όλους τους σταθμούς της συνοριακής γραμμής *US-Mexico* και μόνο για *διελεύσεις πεζών* (δηλ. του τύπου ***Pedestrians***), και ακολούθως υπολογίστε την ελάχιστη (μη μηδενική) και τη μέγιστη τιμή των εν λόγω διελεύσεων πεζών από τη συγκεκριμένη συνοριακή γραμμή (δηλ. ανεξαρτήτως συνοριακού σταθμού). Πως συγκρίνεται η εν λόγω μέγιστη τιμή διελεύσεων πεζών με την αντίστοιχα παρατηρούμενη μέγιστη τιμή των διελεύσεων πεζών από τους σταθμούς της συνοριακής γραμμής *US-Canada*;

Από το πακέτο *dplyr*, η συνάρτηση *group_by()* χρησιμοποιείται για την ομαδοποίηση επιμέρους κοινών στοιχείων σε ένα πλαίσιο δεδομένων, η οποία θα πρέπει να ακολουθείται από τη συνάρτηση *summarise()* με μια κατάλληλη ενέργεια προς εκτέλεση. Οι αντίστοιχες εντολές είναι της μορφής:

group_by(col,..) %>% summarise(action)

Για τις λεπτομέρειες συμβουλευτείτε τη βοήθεια του R.

4. Για να πάρετε μια εποπτική εικόνα των δεδομένων δημιουργήστε μερικά ενδεικτικά γραφήματα, αφού προηγουμένως συμβουλευτείτε π.χ. τις ιστοσελίδες

- https://www.statology.org/grouped-barplot-in-r/
- https://mq-software-carpentry.github.io/r-ggplot-extension/02-categorical-data/index.html

όπου αναφέρονται αντιπροσωπευτικά παραδείγματα κατασκευής σχετικών γραφημάτων με τη βοήθεια του πακέτου ***ggplot2*** του R, τα οποία θα μπορούσατε να προσαρμόσετε για τις ανάγκες οπτικοποίησης των συγκεκριμένων δεδομένων της Θεματικής Εργασίας.

Συγκεκριμένα, χρησιμοποιώντας τη συνάρτηση ***select()*** από το πακέτο ***dplyr*** του R, προκειμένου να εξάγετε από το πλαίσιο δεδομένων ***border.crossing.clean*** μόνο τις στήλες που αντιστοιχούν στις μεταβλητές *Measure, Date,* και *Value* και δημιουργώντας ένα νέο αντικείμενο δεδομένων με την ονομασία ***plot.crossings***. Για τους σκοπούς της δημιουργίας απλών γραφημάτων από τα δεδομένα του ***plot.crossings*** είναι απαραίτητο να αλλάξετε τα στοιχεία της στήλης *Date* από τη μορφή πλήρους ημερομηνίας π.χ. 1999-01-01 στην μορφή μόνο αναφοράς του έτους, π.χ. αντίστοιχα 1999, χρησιμοποιώντας τις συναρτήσεις ***as.factor()*** και ***substring()*** μια εντολή της μορφής

plot.crossings$Date <- as.factor(substring(...$Date,1,4))

Προκειμένου να απεικονίσετε τις ετήσιες καταγραφές των διαφόρων τύπων διελεύσεων είναι απαραίτητο να γίνει μια σύμπτυξη των δεδομένων του ***plot.crossings*** χρησιμοποιώντας τις συναρτήσεις ***group_by()*** και ***summarise()***. Για παράδειγμα, εκτελέστε τις ενδεικτικές εντολές

```
plot.crossings <- plot.crossings %>% group_by(Date, Measure) %>%
  summarise(sum=sum(Value))
```

Εν προκειμένω, με με τον τρόπο αυτό δημιουργείται ένα αναθεωρημένο πλαίσιο δεδομένων ***plot.crossings***, τα στοιχεία του οποίου εξάγονται από την κλήση της συνάρτησης ***group_by()*** ομαδοποιώντας τα στοιχεία των στηλών με τις ονομασίες ***Date*** *και* ***Measure*** στο αρχικό αντικείμενο ***plot.crossings*** και, από την κλήση της συνάρτησης ***summarise()*** υπολογίζεται το πλήθος των συνοριακών διελεύσεων για τις επιμέρους "ομάδες" (εν προκειμένω, ανά έτος και ανά τύπο διέλευσης).

Συγκεκριμένα, ως παράδειγμα, μπορείτε να δημιουργήσετε ένα γράφημα που να απεικονίζει στοιβαγμένες ράβδους για κάθε έτος καταγραφών και για κάθε τύπο διελεύσεων με το ύψος των διαφορετικών έγχρωμων υπο-ράβδων να αντιστοιχεί στην αναλογία του εκάστοτε τύπου διέλευσης που συνεισφέρει στις ετήσιες καταγραφές. Η συνάρτηση ***ggplot()*** παρέχει μια εύκολη λύση χρησιμοποιώντας το όρισμα *"fill=…"* στο τον καθορισμό της αισθητικής του γραφήματος.

Εναλλακτικά, προσπαθήστε να δημιουργήσετε ένα αντίστοιχο γράφημα που να απεικονίζει δίπλα-δίπλα τις ράβδους που αντιστοιχούν σε κάθε τύπο διέλευσης, χρησιμοποιώντας το όρισμα *position* στην συνάρτηση ***geom_bar()*** του ggplot2 και ρυθμίζοντας το στην ένδειξη *"dodge"*.

Για να εξοικειωθείτε περαιτέρω με τη χρήση των συναρτήσεων ***group_by()*** και ***summarise()*** εκτελέστε τις ακόλουθες εντολές:

```
# Number of crossings by Border
crossings.by.location <- border.crossing.clean %>%
    group_by(Border, State) %>%
    summarise(sumPortCode=n_distinct('PortCode')) %>%
    as.data.frame()
crossings.by.location
```

Εν προκειμένω, με τις εν λόγω εντολές δημιουργείται ένα πλαίσιο δεδομένων ***crossings.by.location***, τα στοιχεία του οποίου εξάγονται από την κλήση της συνάρτησης ***group_by()*** ομαδοποιώντας τα στοιχεία των στηλών με τις ονομασίες ***Border*** *και* ***State*** στο αρχικό αντικείμενο ***border.crossing.clean*** και, από την κλήση της συνάρτησης ***summarise()*** υπολογίζεται το πλήθος των συνοριακών σταθμών για τις επιμέρους "ομάδες" (εν προκειμένω, τους συνοριακούς σταθμούς ανά πολιτεία).

Κατά παρόμοιο τρόπο (με μια διαφορετική ομαδοποίηση) δημιουργήστε ένα πλαίσιο δεδομένων ***crossings.by.location***, τα στοιχεία του οποίου θα δημιουργηθούν από την κλήση της συνάρτησης ***group_by()*** που ενεργεί στο αντικείμενο ***border.crossing.clean*** επί των στηλών του με τις ονομασίες ***Port.Name*** *και* ***State*** και το μετατρέπει σε ένα ομαδοποιημένο πίνακα τιμών όπου η συνάρτηση ***summarise()*** εκτελείται "ανά ομάδες" της μεταβλητής ***PortCode*** υπολογίζοντας το συνολικό άθροισμα των συνοριακών σταθμών από τις επιμέρους αντίστοιχες τιμές κάθε ομάδας.

Κατά παρόμοιο τρόπο (με μια διαφορετική ομαδοποίηση) υπολογίστε την κατανομή των συνοριακών σταθμών στις δύο συνοριακές γραμμές των ΗΠΑ για να δείτε την επιμέρους εικόνα των δεδομένων. Τα αποτελέσματα της νέας ομαδοποίησης θα πρέπει να σας δώσουν ότι 78 συνοριακοί σταθμοί είναι στη συνοριακή γραμμή *US-Canada* και οι υπόλοιποι 25 στη συνοριακή γραμμή *US-Mexico.*

Κατά παρόμοιο τρόπο (με μια διαφορετική ομαδοποίηση στο αντικείμενο ***border.crossing.clean***) υπολογίστε την κατανομή κάθε τύπου διέλευσης σε όλους τους συνοριακούς σταθμούς κάθε πολιτείας των ΗΠΑ. Τα αποτελέσματα της νέας ομαδοποίησης θα πρέπει να σας δώσουν αποτελέσματα της μορφής

	Measure	State	sumPortCode
1	Bus Passengers	ALASKA	3
2	Bus Passengers	ARIZONA	5
3	Bus Passengers	CALIFORNIA	6

Με παρόμοιο τρόπο, υπολογίστε την ελάχιστη, μέγιστη, διάμεσο και μέση τιμή των διελεύσεων από όλους τους συνοριακούς σταθμούς ξεχωριστά (δηλ. για τους 78 και 25 σταθμούς) σε καθεμία από τις πολιτείες που εκτείνονται κατά μήκος των συνοριακών γραμμών, για όλη τη χρονική περίοδο που

καλύπτουν τα διαθέσιμα δεδομένα στο αντικείμενο ***border.crossing.clean***. Σημειώστε ποιες είναι οι 6 πολιτείες με τον υψηλότερο αριθμό διελεύσεων (ανεξαρτήτως τύπου) – θα χρειαστείτε αυτή την πληροφορία αγότερα.

5. Από τα διαθέσιμα δεδομένα στο αντικείμενο ***border.crossing.clean*** δημιουργήστε ένα νέο πλαίσιο δεδομένων με την ονομασία ***actual.crossings*** που θα περιέχει μόνο τις καταγραφές από τις στήλες **"Border", "Measure"**, και **"Value",** δηλαδή για όλους τους τύπους διέλευσης σε όλους τους σταθμούς και των δύο συνοριακών γραμμών χωρίς μηδενικές καταγραφές στα κελιά της στήλης για τη μεταβλητή ***Value***. Εκτυπώστε συνοπτικά στοιχεία των δεδομένων που περιέχονται στο αντικείμενο ***actual.crossings***.

Όπως και στο προηγούμενο ερώτημα, χρησιμοποιώντας τη συνάρτηση ***table()***, υπολογίστε, από πόσες συνολικά διελεύσεις κάθε τύπου έχουν συλλεχθεί αντίστοιχα στοιχεία ξεχωριστά για τους σταθμούς των συνοριακών γραμμών *US-Canada* και *US-Mexico, και συνολικά για κάθε τύπο διέλευσης και στις δύο συνοριακές γραμμές*. Εκχωρήστε το αποτέλεσμα της καταμέτρησης σε ένα αντικείμενο με την ονομασία ***counts_actual_border*** και εκτυπώστε τα επιμέρους (δηλ. κατά συνοριακή γραμμή και τύπο διέλευσης) και το συνολικό αποτέλεσμα (δηλ. για κάθε τύπο διέλευσης και από τις δύο συνοριακές γραμμές). Επιβεβαιώστε το αποτέλεσμα της εν λόγω καταμέτρησης με ένα απλό γράφημα ράβδων, κάνοντας χρήση της συνάρτησης ***barplot()***. Επιπρόσθετα, δημιουργήστε ένα αντίστοιχο (βελτιωμένης αισθητικής) γράφημα κάνοντας χρήση της συνάρτησης ***ggplot()*** του πακέτου ***ggplot2()***.

6. Από τα δεδομένα του αντικειμένου ***canada.crossings*** δημιουργήστε ένα νέο αντικείμενο με την ονομασία ***crossings.byport.canada*** που να περιλαμβάνει τις καταγραφές από τους σταθμούς της συνοριακής γραμμής US-Canada, ομαδοποιήστε τα δεδομένα για τις μεταβλητές ***PortName, State, PortCode, Border,*** και ***Value*** και υπολογίστε το άθροισμα των διελεύσεων, ανεξαρτήτως τύπου, για κάθε συνοριακό σταθμό. Ακολούθως, κάνοντας χρήση της συνάρτησης ***order()***, εξακριβώστε ποιοί είναι οι 5 συνοριακοί σταθμοί στους οποίους γίνονται (i) οι περισσότερες διελεύσεις ανεξαρτήτως τύπου, (ii) οι λιγότερες διελεύσεις ανεξαρτήτως τύπου. Αντίστοιχα, με παρόμοιο τρόπο εξακριβώστε ποιές είναι οι 5 πολιτείες κατά μήκος αυτής της συνοριακής γραμμής στις οποίες γίνονται (iii) οι περισσότερες διελεύσεις ανεξαρτήτως τύπου και (iv) οι λιγότερες διελεύσεις ανεξαρτήτως τύπου.

Κατά παρόμοιο τρόπο, από τα δεδομένα του αντικειμένου ***mexico.crossings*** δημιουργήστε ένα νέο αντικείμενο με την ονομασία ***crossings.byport.buses.mexico*** που να περιλαμβάνει τις καταγραφές από τους σταθμούς της συνοριακής γραμμής US-Mexico μόνο για διελεύσεις με λεωφορεία (δηλ. εστιάστε στη μεταβλητή ***Buses***), ομαδοποιήστε τα δεδομένα για τις μεταβλητές ***PortName, State, PortCode, Border,*** και ***Value*** και υπολογίστε το άθροισμα των διελεύσεων με λεωφορεία, για κάθε συνοριακό σταθμό. Ακολούθως, εξακριβώστε ποιοί είναι οι 5 συνοριακοί σταθμοί στους οποίους γίνονται (i) οι περισσότερες διελεύσεις με λεωφορεία. Αντίστοιχα, με παρόμοιο τρόπο εξακριβώστε ποιές είναι οι 5 πολιτείες κατά μήκος αυτής της συνοριακής γραμμής στις οποίες γίνονται (ii) οι περισσότερες διελεύσεις με ιδιωτικά αυτοκίνητα με επιβάτες (δηλ. εστιάστε στη μεταβλητή ***Personal Vehicle Passengers***).

2. Η συμπεριφορά των κλιματικών φαινομένων El Nino και El Nina εκφράζεται με τους λεγόμενους ***δείκτες ONI*** *(Oceanic Nino Index)* που βασίζονται στις ανωμαλίες (αποκλίσεις) της θερμοκρασίας της θαλάσσιας επιφάνειας *(SSTA, Sea Surface Temperature Anomalies)* από τις μέσες συνθήκες συνήθως σε μια περίοδο 30 ετών κατά μήκος 4 συγκεκριμένων περιοχών του Νότιου Ειρηνικού. Οι τέσσερεις περιοχές είναι γνωστές ως **"Nino 1+2"** (φ:0-12^{o} S, λ:90^{o} – 120^{o} W), **"Nino 3"** (5N-5S, 150W-90W), **"Nino 3.4"** (5N-5S, 170W-120W) και **"Nino 4"** (5N-5S, 160E-150W). Ειδικά ο δείκτης ONI από την περιοχή Nino 3.4 συνήθως χρησιμοποιείται για να χαρακτηριστούν έντονα El Niño και La Niña συμβάντα. Για τους σκοπούς της άσκησης θα χρειαστείτε το σχετικό αρχείο ***elnino.for*** το οποίο ακολουθεί τον μορφότυπο *fixed width format*. Για τη διευκόλυνσή σας, τα πρώτα στοιχεία του αρχείου είναι:

```
Weekly SST data starts week centered on 3Jan1990

                  Nino1+2       Nino3         Nino34        Nino4
 Week           SST SSTA      SST SSTA      SST SSTA      SST SSTA
```

```
03JAN1990     23.4-0.4     25.1-0.3     26.6 0.1     28.6 0.5
10JAN1990     23.4-0.8     25.2-0.3     26.6 0.1     28.6 0.5
17JAN1990     24.2-0.3     25.3-0.3     26.5-0.1     28.6 0.5
24JAN1990     24.4-0.4     25.5-0.4     26.5-0.1     28.4 0.3
31JAN1990     25.1-0.1     25.8-0.2     26.7 0.1     28.4 0.3
```

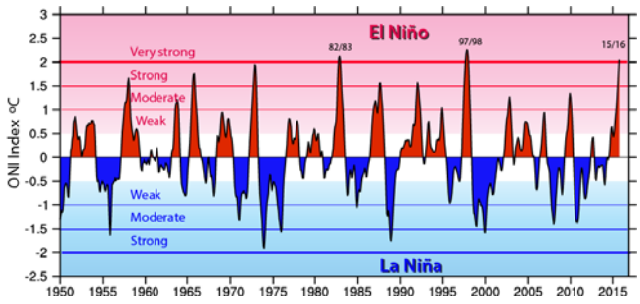
- Συγκεκριμένα, κατεβάστε στο χώρο εργασίας σας του R, χρησιμοποιώντας τη σχετική εντολή ***download.file()***, **μόνο τα δεδομένα** από τις στήλες 'Week' και 'Nino34/SST SSTA', εκχωρώντας τα δεδομένα σε μια μεταβλητή ***area_34***. Σημειώστε ότι κάτω από τις ετικέτες των στηλών '…Nino' ακολουθούν δύο στήλες δεδομένων, SST (Sea Surface Temperatures, σε ^{o}C) και SSTA (Sea Surface Temperature Anomalies, επίσης σε ^{o}C). Εκχωρήστε στα δεδομένα τις ετικέτες "week", "sst", "ssta".

- Με τις κατάλληλες εντολές στο R, δώστε εκτυπώστε τη δομή του συνόλου των δεδομένων που έχει αποθηκευθεί από το R, καθώς και μια περίληψη των επιμέρους δεδομένων που περιέχονται στο συγκεκριμένο αντικείμενο του R. Ακολούθως εκτελέστε τις ακόλουθες εντολές εξαγωγής υποσυνόλων των δεδομένων:
    - Εκτυπώστε τις πρώτες και τις τελευταίες 20 σειρές των δεδομένων
    - Ποιο είναι το πλήθος των δεδομένων SST ή/και SSTA για την περιοχή Nino 3.4
    - Εξάγετε τη στήλη των δεδομένων SSTA για την περιοχή Nino 3.4 χρησιμοποιώντας την ονομασία της στήλης.
    - Εξάγετε τη 3^{η} , 5^{η} και 12^{η} γραμμή από τις στήλες SST ή/και SSTA για την περιοχή Nino 3.4
    - Μετατρέψτε τις ενδείξεις του χρόνου των δεδομένων (π.χ. 03JAN1990) , στη μορφή
        - ως ημερομηνία (yyyy_dy_mo), π.χ. 1990-01-03, και ακολούθως ως Έτος(year), Μήνα(month), Ημέρα(day) και τέλος ως έτος σε δεκαδική μορφή (yr_frac). Τυπώστε μερικές από αυτές τις ενδείξεις στην αρχική μορφή τους και κατόπιν των σχετικών μετατροπών τους.
    - Δημιουργήστε ένα νέο πλαίσιο δεδομένων ***nino34*** που να περιλαμβάνει ως στήλες του τα δεδομένα ***yyyy_dy_mo, yr_frac,*** και ***sst, stta*** από την περιοχή Nino 3.4. Επιβεβαιώστε με τις κατάλληλες εντολές τον τύπο και τη δομή του εν λόγω αντικειμένου. Πόσες γραμμές δεδομένων περιλαμβάνονται σε αυτό; Εκτυπώστε τα δεδομένα από την τελευταία χρονική εποχή.
    - Δημιουργήστε, από το προηγούμενο αντικείμενο του R, δύο διανύσματα δεδομένων, και εκτυπώστε μερικές από τις πρώτες και τελευταίες τιμές των δεδομένων εκάστου διανύσματος
        - ***sst_v***, τα στοιχεία με όλες τις τιμές sst
        - ***ssta_v***, τα στοιχεία με όλες τις τιμές ssta
    - Ομοίως, δημιουργήστε δύο υποσύνολα δεδομένων, εξακριβώστε τον τύπο και τη δομή τους στο R, και εκτυπώστε το πλήθος των στοιχείων τους, καθώς μερικές από τις πρώτες και τελευταίες τιμές των δεδομένων εκάστου διανύσματος:
        - ***la_nina***, τα στοιχεία με τις τιμές ***ssta<0***
        - ***el_nino***, τα στοιχεία με τις τιμές ***ssta≥0***.
    - Χρησιμοποιώντας τη βασική συνάρτηση ***plot()*** του R, δημιουργήστε δύο ξεχωριστά γραφήματα, όπου στο καθένα:
        - στον άξονα των x είναι οι χρονικές εποχές με ετικέτα 'Years', και
        - στον άξονα των y, για τα αντίστοιχα γραφήματα τα δεδομένα ssta που περιλαμβάνονται στα αντικείμενα
            - ***la_nina*** (χρησιμοποιήστε μπλέ χρώμα), και ετικέτα στον άξονα 'La Nina, SSTA in oC' στο ένα και
            - ***el_nino*** (χρησιμοποιήστε κόκκινο χρώμα), και ετικέτα στον άξονα 'El Nino, SSTA in oC' στο άλλο.

- Χρησιμοποιώντας κατάλληλες εντολές κλήσης των βασικών συναρτήσεων ***plot()*** και ***points()*** του R, δημιουργήστε ένα ενιαίο γράφημα στο οποίο
    - στον άξονα των x είναι οι χρονικές εποχές με ετικέτα 'Years',
    - στον άξονα των y να υπάρχει ετικέτα 'SST Anomalies, in oC',
    - ο τίτλος στην κεφαλή του γραφήματος να είναι 'Nino 3.4 – Weekly SST Anomalies since 1990'
    - να απεικονίζονται στο ίδιο γράφημα τα προηγούμενα δεδομένα,
        - ***la_nina*** (με μπλέ χρώμα), και
        - ***el_nino*** (με κόκκινο χρώμα).

Για την διευκόλυνσή σας, το ενιαίο γράφημα τυπικά θα πρέπει να φαίνεται όπως στην ενδεικτική παραπλεύρως απεικόνιση. **Είναι προαιρετικό για τις ανάγκες της άσκησης, αλλά συνίσταται να δοκιμάσετε να δημιουργήσετε ένα αντίστοιχο ενιαίο γράφημα χρησιμοποιώντας το πακέτο *ggplot2***.

- Θεωρείστε ότι το διάνυσμα τιμών ***sst_v*** εκφράζει το θεωρητικό πληθυσμό θερμοκρασιών της θαλάσσιας επιφάνειας στην περιοχή Nino 3.4. Υπολογίστε τη μέση τιμή και τη διασπορά του εν λόγω πληθυσμού. Ακολούθως δημιουργήστε από τον πληθυσμό τυχαία δείγματα ***sst_10, sst_20, sst_50, sst_200, sst_600*** και ***sst_900*** το καθένα αντίστοιχα με 10, 20, 50, 200, 600 και 900 στοιχεία και υπολογίστε τη δειγματική μέση τιμή και τη διασπορά καθενός από τα εν λόγω δείγματα. Επιθεωρήστε, με γραφικό τρόπο (το ίδιο θα μπορούσε να γίνει αναλυτικά με την εκτέλεση ανάλογων στατιστικών δοκιμών με τη χρήση της συνάρτησης ***t.test()***), εάν η δειγματική μέση τιμή και η διασπορά καθενός από τα δείγματα με αυξανόμενο πλήθος στοιχείων συγκλίνει προς τη μέση τιμή και τη διασπορά του πληθυσμού, για κάθε περίπτωση δημιουργήστε ένα γραφήματα κουτιού ή θηκόγραμμα (boxplot). Για λόγους αισθητικής τοποθετήστε τα εν λόγω 6 επιμέρους γραφήματα σε ένα ενιαίο γράφημα-κανάβου 3x2.



- επιβεβαιώστε εάν η δειγματική μέση τιμή και η διασπορά ενός δείγματος με αυξανόμενο πλήθος στοιχείων συγκλίνει προς τη μέση τιμή και τη διασπορά του πληθυσμού. Οπτικά αυτό μπορεί να εξεταστεί
    - Εξακριβώστε, με γραφικό τρόπο, αλλά και υπολογίζοντας (με χρήση της συνάρτησης ***moment()*** του πακέτου **e1071** του R) τις στατιστικές ροπές μ_1 (μέση τιμή), μ_2 (διακύμανση), μ_3 (λοξότητα) και μ_4 (κύρτωση) των δεδομένων ***sst_v*** και ***ssta_v***, αν σε αυτά υπάρχει λοξότητα (0 – κανονική κατανομή, <0 – λοξή προς τα δεξιά ή >0 – λοξή προς τα αριστερά) και κύρτωση (3 - μεσόκυρτη κατανομή, <3 - λεπτόκυρτη κατανομή, >3 - πλατύκυρτη κατανομή) στην κατανομή τους.
    - Εξακριβώστε, γραφικά και αναλυτικά χρησιμοποιώντας τις συναρτήσεις ***plot()*** και ***cor()***, αν υπάρχει συσχέτιση μεταξύ των δεδομένων ***sst_v*** και ***ssta_v***.

- Στις κλιματολογικές μελέτες συνηθίζεται τα συμβάντα El Nino και El Nina να κατατάσσονται ως ***Ασθενή (Weak), Μέτρια (Moderate), Ισχυρά (Strong)***, και ***Ακραία (Extreme ή Very Strong)*** ανάλογα με τις τιμές των ανωμαλιών της θερμοκρασίας της θάλασσας (δηλ. το μέγεθος της παραμέτρου SSTA). Από τα διαθέσιμα δεδομένα δημιουργήστε 8 αντίστοιχα πλαίσια δεδομένων τα οποία να περιλαμβάνουν τις εποχές (στη μορφή π.χ. 2012-03-14 ή/και YEAR, MONTH, DAY), SST και SSTA για τις ακόλουθες κατηγορίες συμβάντων:

    - **el_nino_weak** (0≤ SSTA <1.0), **el_nino_mod** (1≤ SSTA <1.5), **el_nino_strong** (1.5≤ SSTA <2.0), **el_nino_super** (2.0≤ SSTA), και
    - **la_nina_weak** (-1.0 ≤ SSTA < 0), **la_nina _mod** (-1.5 ≤ SSTA < -1.0), **la_nina _strong** (-2.0 ≤ SSTA < -1.5), **la_nina _super** (-2.0 ≤ SSTA).
        - Για τα δεδομένα κάθε κατηγορίας συμβάντων υπολογίστε τα κύρια στατιστικά χαρακτηριστικά τους.

- Εξακριβώστε, γραφικά και αναλυτικά, εάν τα συμβάντα ισχυρών (strong) El Nino και El Nina αναπτύσσονται κυρίως κατά τους μήνες Απρίλιο-Ιούνιο.
- Εξακριβώστε, γραφικά και αναλυτικά, εάν τα συμβάντα ισχυρών (strong) El Nino και El Nina τείνουν να φθάνουν στη μέγιστή ισχύ τους κατά τους μήνες Οκτώβριο-Φεβρουάριο.
- Εξακριβώστε, γραφικά και αναλυτικά, εάν τα συμβάντα El Nino και El Nina τυπικά διαρκούν από 9-12 μήνες, αν και ενίοτε μπορεί να διαρκούν μέχρι και 2 έτη.
- Εξακριβώστε, γραφικά και αναλυτικά, εάν τα συμβάντα ισχυρών (strong) El Nino ή El Nina τυπικά επαναλαμβάνονται κάθε 2 με 7 έτη.

ΣΗΜΕΙΩΣΤΕ - Το θεωρητικό υπόβαθρο που είναι συναφές για την εκπόνηση του παρακάτω τρίτου μέρους της Θεματικής Εργασίας θα αποτελέσει μέρος της διάλεξης του μαθήματος στις 10/1/2023.

3. Φορτώστε στο χώρο εργασίας σας του R, π.χ. με την ονομασία ***CO2***, το σύνολο δεδομένων *CO2* (προσοχή με κεφαλαία γράμματα) που είναι άμεσα διαθέσιμο στο R και αναφέρεται στα δεδομένα μιας έρευνας για την ποσότητα του διοξειδίου του άνθρακα (μεταβλητή ***uptake***) που απορροφάται από έξι τύπους επιφανειακής βλάστησης σε δυο διαφορετικές περιοχές (***Quebec*** και ***Mississippi***) και σε διάφορα επίπεδα συγκέντρωσης *CO2* (μεταβλητή ***conc***) στο περιβάλλον. Για τη μελέτη καταγράφηκαν μετρήσεις μετά από συνθήκες ψύχους (μεταβλητή ***chilled***) ή χωρίς ψύχος (μεταβλητή ***nonchilled***) κατά τη διάρκεια της νύχτας. Το περιεχόμενο των στηλών του πλαισίου δεδομένων ***CO2*** αντιστοιχούν στις μεταβλητές ομαδοποίησης **Type** (Quebec, Mississippi) και **Treatment** (chilled, nonchilled).

Αρχικά, επιβεβαιώστε ότι τα δεδομένα έχουν αποθηκευθεί από το R ως πλαίσιο δεδομένων, και με κατάλληλες εντολές εξακριβώστε εάν υπάρχουν άδεια κελιά ή/και τον αριθμό των άδειων κελιών (εάν υπάρχουν) στο συγκεκριμένο πλαίσιο δεδομένων. Ακολούθως, χρησιμοποιήστε κατάλληλες συναρτήσεις (π.χ. ***str(), head(), summary()***, κ.ά.) προκειμένου να πάρετε μια πρώτη εικόνα της δομής και του περιεχομένου του πλαισίου δεδομένων *CO2*. Χρησιμοποιήστε κατάλληλα την συνάρτηση ***View()*** προκειμένου να προβάλλετε το πλαίσιο δεδομένων *CO2* στην μορφή ενός υπολογιστικού φύλλου.

Στο R η συνάρτηση ***subset()*** διευκολύνει να εξαχθούν από το συνολικό σετ των δεδομένων, συγκεκριμένες υπο-ομάδες δεδομένων. Για τη διευκόλυνσή σας, ως παράδειγμα, εκτελέστε τις ακόλουθες ενδεικτικές εντολές:

```
mississippi_chilled <- subset(CO2, Type == "Mississippi" &
Treatment=="chilled")
mississippi_chilled
plot(mississippi_chilled$uptake)
```

προκειμένου να εξάγετε μόνο τα δεδομένα από την περιοχή *Mississippi* και μόνο για τις μετρήσεις μετά από ψύξη *(chilled)*, τα οποία μπορείτε επίσης να τα απεικονίσετε σε ένα απλό γράφημα.

Με παρόμοιο τρόπο, δημιουργήστε 4 ξεχωριστά αντικείμενα π.χ., με ονομασίες ***quebec_nonchilled, quebec_chilled, …*** που θα περιέχουν αντίστοιχα τα ομαδοποιημένα υποσύνολα δεδομένων ***"Quebec & nonchilled", "Quebec & chilled", "Mississippi & nonchilled", "Mississippi & chilled"***. Επιβεβαιώστε ότι τα νέα αυτά αντικείμενα είναι πλαίσια δεδομένων. Υπολογίστε, για κάθε ένα από αυτά, τα βασικά στατιστικά μέτρα (π.χ., mean(), sd(), median(), …) για τις μεταβλητές ***'conc'*** και ***'uptake'***.

Συνεχίζοντας με τη χρήση της συνάρτησης ***subset()***, εξάγετε σε ένα νέο αντικείμενο, π.χ. με την ονομασία ***quebec_nonchilled.ge.500***, το οποίο θα περιέχει, από την περιοχή Quebec, μόνο δεδομένα συγκεντρώσεων CO2 *(*μεταβλητή ***conc****)* με τιμές μεγαλύτερες ή ίσες με 500, από μετρήσεις μετά από ψύξη των φυτών *(*μεταβλητή ***chilled****)*. Αντίστοιχα, από το σύνολο των δεδομένων και στις δύο περιοχές, εξάγετε σε ένα νέο αντικείμενο, π.χ. με την ονομασία ***treatment_chilled***, τα δεδομένα των ψυχρών δειγμάτων που αντιστοιχούν σε συγκέντρωση CO2 ίση με 250 *(conc=250)*.

Ένας εναλλακτικός τρόπος για να διαχωριστούν τα δεδομένα ενός πλαισίου δεδομένων σε υποσύνολα, παρέχεται από τη χρήση της συνάρτησης ***aggregate()*** η οποία υπολογίζει συνοπτικά στατιστικά στοιχεία για κάθε υποσύνολο και επιστρέφει το αποτέλεσμα σε μια βολική φόρμα.

Χρησιμοποιήστε την εν λόγω συνάρτηση επί των δεδομένων του πλαισίου CO2, για τους συνδυασμούς των δεδομένων ***uptake & Treatment*** και ***conc & Treatment*** ή ακολουθώντας τον συμβολισμό του R (***uptake~Treatment*** και ***conc~Treatment***), και παράμετρο εισόδου FUN=summary.

Χρησιμοποιώντας τις συναρτήσεις δημιουργίας γραφημάτων ***scatterplot()*** και ***hist()*** δημιουργήστε αντίστοιχα γραφήματα για τα ποσοστά της πρόσληψης CO2 (μεταβλητή *uptake*). Αντίστοιχα, με τη χρήση της συνάρτησης ***boxplot()*** δημιουργήστε ένα αντίστοιχο γράφημα για τα ποσοστά της πρόσληψης CO2 ως συνάρτηση των μετρήσεων μετά από ψύξη (chilled) ή χωρίς ψύξη (nonchilled).

Επιπλέον, προκειμένου να υπολογίσετε με συνοπτικό τρόπο διάφορα στατιστικά μέτρα των υπο-ομάδων δεδομένων του πλαισίου CO2, χρησιμοποιήστε (επί του πλαισίου δεδομένων CO2) τη συνάρτηση ***summarise()*** από το πακέτο ***dplyr*** και εκχωρήστε σε ένα νέο αντικείμενο, π.χ. με την ονομασία ***CO2_groups_stats***, την μέση τιμή (mean), τη διάμεσο (median) και την τυπική απόκλιση (sd) των δεδομένων CO2 για τις 4 υπο-ομάδες κατανεμημένες ανά **Type** και **Treatment**. Εξακριβώστε τις διαφορές στους υπολογισμούς που προκύπτουν αντίστοιχα από τη χρήση της βασικής συνάρτησης ***summary()*** επί των αντικειμένων ***CO2_groups_stats*** και **CO2**. Δώστε ένα σύντομο σχόλιο (# ….) στον κώδικα του R, πως ερμηνεύετε τις παρατηρούμενες διαφορές.

Θεωρήστε ότι θέλετε να ελέγξετε τις ακόλουθες περιπτώσεις στατιστικών υποθέσεων:

(a) **H0**: Η θερμοκρασία (δηλ., η επεξεργασία (Treatment) με ψύξη ή χωρίς ψύξη) **δεν επηρεάζει** την πρόσληψη CO2

H1: Η θερμοκρασία (δηλ., η επεξεργασία (Treatment) με ψύξη ή χωρίς ψύξη) **επηρεάζει** την πρόσληψη CO2

(b) **H0**: Η τοποθεσία προέλευσης του φυτού (Type) **δεν επηρεάζει** την πρόσληψη CO2

H1: Η τοποθεσία προέλευσης του φυτού (Type) **επηρεάζει** την πρόσληψη CO2

(c) **H0**: Η τοποθεσία προέλευσης του φυτού (Type) και η θερμοκρασία (η επεξεργασία (Treatment) με ψύξη ή χωρίς ψύξη) **δεν επηρεάζουν** την πρόσληψη CO2

H1: Η τοποθεσία προέλευσης του φυτού (Type) και η θερμοκρασία (η επεξεργασία (Treatment) με ψύξη ή χωρίς ψύξη) **επηρεάζουν** την πρόσληψη CO2

Επειδή έχουμε να κάνουμε με πολυδιάστατα δεδομένα (δηλ. δύο ή περισσότερες ομάδες), αυτό μπορεί να επιτευχθεί με τη λεγόμενη διαδικασία **μονόπλευρης (one-way)** ή/και **παραγοντικής (two-way) ανάλυσης της διακύμανσης** (**ANOVA, analysis of variance**), η οποία είναι μια επέκταση της δοκιμής **t-test** για τη σύγκριση των μέσων δύο ανεξάρτητων δειγμάτων.

Εν προκειμένω, για παράδειγμα, η μονόπλευρη ανάλυση δοκιμής t-test για τις μεταβλητές ***uptake*** και ***Treatment*** επιτυγχάνεται με τις ακόλουθες ενδεικτικές εντολές χρήσης των συναρτήσεων ***lm()*** και ***anova()***:

```
# Perform one-way t test on uptake & Treatment _ CASE (1)
fit_Treatment = lm(uptake~Treatment, data=CO2)
anova(fit_Treatment)
```

Στο αποτέλεσμα επιθεωρήστε τη ζητούμενη κρίσιμη τιμή πιθανότητας *p-value = Pr(>F)*, η οποία οδηγεί στην αποδοχή (εάν p-value ≥ a=0.05) ή απόρριψη (εάν p-value < a=0.05) της Μηδενικής Υπόθεσης της δοκιμής, σύμφωνα με το εκάστοτε καθορισμένο επίπεδο εμπιστοσύνης *a*.

Εκτελέστε με παρόμοιο τρόπο την αντίστοιχη ανάλυση δοκιμής t-test για τις μεταβλητές ***uptake*** και ***Type.***

Για την δοκιμή των υποθέσεων στην περίπτωση (c) των στατιστικών υποθέσεων εκτελέστε μια δίπλευρη δοκιμή (two-way ANOVA) για τη μεταβλητή ***uptake*** σε συνδυασμό με τις μεταβλητές ομαδοποίησης **Type** και **Treatment (**συμβολικά **Type*Treatment** στην κλίση της συνάρτησης *lm()* για τη συμβολική περιγραφή του μοντέλου στη μορφή ***response ~ terms*****).** Για λεπτομέρειες ως προς τη χρήση της συνάρτησης ***lm()*** συμβουλευτείτε το σύνδεσμο https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm.

Από τα αποτελέσματα των δοκιμών ANOVA για τις τρεις προαναφερόμενες περιπτώσεις, δώστε στον κώδικα R, μερικές σύντομες γραμμές σχολίων όπου καλείστε να συνοψίσετε τα συμπεράσματα σας για την αποδοχή ή μη της εκάστοτε περίπτωσης ελέγχου των προαναφερόμενων υποθέσεων.

Χρησιμοποιήστε τη συνάρτηση ***aov()*** του R για να διενεργήσετε μια σύντομη ανάλυση της διασποράς, για κάθε μια από τις τρεις προαναφερόμενες περιπτώσεις ***(uptake~Treatment , uptake~Type , uptake~Type*Treatment)***. Επιπλέον, υπολογίστε τα υπόλοιπα του εκάστοτε μοντέλου (*response ~ terms*) κάνοντας χρήση της συνάρτησης ***residuals(object = …)*** και ακολούθως εκτελέστε τη δοκιμή κανονικότητας Shapiro-Wilk (για τον έλεγχο εάν τα εκάστοτε υπό εξέταση δεδομένα ακολουθούν μια κανονική κατανομή).