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

ΟΝΟΜΑΤΕΠΩΝΥΜΟ .....

ΕΞΑΜΗΝΟ .....

Ημερομηνία Παράδοσης **: 13/12/2022**

ΘΕΜΑΤΙΚΗ ΕΡΓΑΣΙΑ #3

Σκοπός: Η παρούσα θεματική εργασία αποσκοπεί στην εξοικείωση σας με τις βασικές εντολές εισαγωγής στο λογισμικό R αρχείων δεδομένων από διαφορετικές πηγές, μορφότυπους και δομές. Η προσπάθεια θα επικεντρωθεί αρχικά στην χρήση ενδεικτικών τυπικών μικρών αρχείων, και σε ένα δεύτερο μέρος στη χρήση πιο αντιπροσωπευτικών αρχείων δεδομένων για τυπικές εφαρμογές με το R.

ΠΡΟΚΑΤΑΡΚΤΙΚΑ – Για τους σκοπούς της συγκεκριμένης θεματικής εργασίας στις ιστοσελίδες του μαθήματος έχουν αναρτηθεί ενδεικτικά αρχεία δεδομένων τα οποία εμπεριέχονται στο ντοσιέ ***myrdata***. Για τα συγκεκριμένα σετ δεδομένων δεν έχετε πρόσβαση για άμεση οπτική επιθεώρηση του περιεχομένου τους, αλλά αυτά μπορούν να μεταφορτωθούν στον Η/Υ σας (μόνο με κατάλληλες εντολές μέσα από το R, βλ. παρακάτω) από τον σύνδεσμο:
http://portal.survey.ntua.gr/main/labs/hgeod/ddeli/myrdata/

Αν κάποιο αρχείο λείπει εκ παραδρομής, ειδοποιήστε τον διδάσκοντα με ένα email.

Για παράδειγμα εάν, για τους σκοπούς της άσκησης, πρέπει να μεταφορτώσετε (download) στον Η/Υ σας, από το ντοσιέ ***myrdata***, κάποιο σετ δεδομένων που περιέχονται σε ένα αρχείο με την πλήρη ονομασία ***thefiletodownload.ext*** (συμπεριλαμβανομένης της προέκτασης ***.ext***) πρέπει να κάνετε χρήση της συνάρτησης ***download.file()*** με τον συγκεκριμένο σύνδεσμο (URL): http://portal.survey.ntua.gr/main/labs/hgeod/ddeli/myrdata/thefiletodownload.ext. Σημειώστε ότι σε κάποια ερωτήματα της θεματικής εργασίας, θα σας ζητηθεί να χρησιμοποιήσετε συγκεκριμένα αρχεία δεδομένων απευθείας μέσω του εκάστοτε συνδέσμου μεταφόρτωσής τους (δηλ. ακολουθώντας το παράδειγμα της προαναφερόμενης διεύθυνσης URL).

Η πλήρης σύνταξη για την κλήση της συνάρτησης *download.file()* είναι:

```
download.file(url, destfile, method, quiet = FALSE, mode = "w",
        cacheOK = TRUE, extra = getOption("download.file.extra"), headers = NULL, ...)
```

Προκειμένου να ασκηθείτε με τη χρήση διαφορετικών συναρτήσεων που επιτρέπουν την ανάγνωση αρχείων σε μορφότυπους ***.txt, .dat, .csv, .tsv, .xls, .xlsx*** θα χρησιμοποιήσετε, κάθε φορά, κατάλληλες συναρτήσεις ανάγνωσης των εκάστοτε αρχείων οι οποίες θα επιλέγονται από τουλάχιστον δύο από τις ακόλουθες κατηγορίες συναρτήσεων:

A. Ενσωματωμένες συναρτήσεις στον πυρήνα του R (R build-in functions)
   - ***read.table(), read.csv(), read.csv2(), read.delim(), read.delim2(), scan()***

B. Συναρτήσεις από το πακέτο ***gdata*** του R
   - ***read.xls(), xls2csv(), xls2tab(), xls2tsv(), xls2sep()***

Πιθανά για κάποιες από αυτές τις συναρτήσεις θα χρειαστεί επίσης να εγκαταστήσετε κάποια βασικά εργαλεία για το περιβάλλον της γλώσσας Perl, π.χ. το ***ActivePerl*** ή το ***Strawberry Perl*** , ακολουθώντας το σύνδεσμο https://www.perl.org/get.html ή τους επιμέρους συνδέσμους

- https://www.activestate.com/products/perl/ , ή
- https://strawberryperl.com/

C. Συναρτήσεις από το πακέτο ***readr*** του R
   - ***read_csv(), read_tsv(), read_delim(), read_fwf(), read_table(), read_log()***

D. Συναρτήσεις από το πακέτο ***readxl*** του R
   - ***read_excel(), excel_sheets(), read_delim(), read_fwf(), read_table(), read_log()***

E. Συναρτήσεις από το πακέτο ***xlsx*** του R
   - ***read.xlsx(), read.xlsx2(), write.xlsx(), write.xlsx2()***

F. Συναρτήσεις από τα πακέτα ***openxlsx*** και ***writexl*** του R
   - ***read.xlsx(), write.xlsx(), ..., write_xlsx()***

G. Συναρτήσεις από το πακέτο ***XLConnect*** του R
   - ***loadworkbook(), ...***

Κατεβάστε από τους συνδέσμους https://www.povertyactionlab.org/sites/default/files/r-cheat-sheet.pdf και https://tinyurl.com/Rimportdata-cheatsheet χρήσιμα βοηθήματα κοινότυπων εντολών *(cheat sheets)* που ενδέχεται να χρειαστείτε.

Για κάθε αρχείο που εισάγετε στο R τα δεδομένα συνήθως θα αποθηκεύονται ως πλαίσιο δεδομένων ή ανάλογα με τη συνάρτηση που χρησιμοποιείται μπορεί να αποθηκεύονται και με άλλες μορφές αντικειμένων (π.χ. διανύσματα τιμών, λίστες, κ.ά.). Κάθε φορά που εισάγετε ένα αρχείο δεδομένων συνιστάται κατ' ελάχιστον να κάνετε χρήση εντολών όπως: ***head()*** και ***tail()*** για να επιθεωρήσετε στην οθόνη σας τα δεδομένα στην αρχή και το τέλος του εκάστοτε αρχείου, ***str ()*** για την εμφάνιση της δομής του αντικειμένου του R στο οποίο αποθηκεύονται τα δεδομένα, και την εντολή ***summary()*** προκειμένου να υπολογιστούν περιλήψεις περιγραφής διαφόρων στατιστικών μέτρων που αφορούν μια, όλες ή επιλεγμένες μεταβλητές του εκάστοτε πλαισίου δεδομένων. Μπορείτε (και συνίσταται) επίσης να χρησιμοποιήσετε οποιεσδήποτε άλλες διαχειριστικές εντολές τις επιλογής σας που θα σας επιτρέπουν για να εξάγετε άλλες διαχειριστικές πληροφορίες για τη δομή και τα πεδία των δεδομένων και των μεταβλητών που συμπεριλαμβάνονται στο εκάστοτε αρχείο, π.χ. την κλάση των αντικειμένων που αποθηκεύει το R σε κάθε περίπτωση, να ελέγξετε το μέγεθος/πλήθος, διαστάσεις (π.χ., αριθμό των στηλών και γραμμών) του εκάστοτε συνόλου των δεδομένων κλπ. Για το σκοπό αυτό, δοκιμάστε ενδεικτικά όσες περισσότερες από τις εντολές περιλαμβάνονται στο ενδεικτικό αρχείο εντολών ***typical_commands_data handling.pdf*** που θα βρείτε στην ιστοσελίδα των ασκήσεων του μαθήματος, και συγκεκριμένα συμβουλευτείτε τις ενότητες των εντολών *'Exploring the data'* και *'Check for missing data'*.

Τέλος, για κάθε ένα από τα μέρη της Θεματικής Εργασίας καθορίσετε στον Η/Υ σας ένα **ξεχωριστό ντοσιέ εργασίας (working directory) για την εκάστοτε τρέχουσα συνεδρία σας στο R**. Αντίστοιχα, με την ολοκλήρωση κάθε μέρους της θεματικής εργασίας δημιουργήστε ένα ***.Rhistory*** αρχείο το οποίο θα πρέπει να υποβάλλετε με την Τεχνική Έκθεσή σας.

ΕΠΙΜΕΡΟΥΣ ΟΜΑΔΕΣ ΑΡΧΕΙΩΝ ΔΕΔΟΜΕΝΩΝ

```
unzip(zipfile, files = NULL, list = FALSE, overwrite = TRUE,
      junkpaths = FALSE, exdir = ".", unzip = "internal", setTimes = FALSE)
```

Η παράμετρος εισόδου 'zipfile' μπορεί να περιέχει το όνομα του συμπιεσμένου αρχείου (εάν αυτό βρίσκεται στο χώρο εργασίας της τρέχουσας συνεδρίας του R) ;ή να περιέχει την πλήρη διαδρομή του αρχείου zip.

1.(a) Κατεβάστε, με την εντολή ***download.file()***, από το ντοσιέ ***myrdata*** στον ιστοχώρο του μαθήματος, το συμπιεσμένο αρχείο ***faithful_datafiles.zip*** και με την εντολή ***unzip()*** αποσυμπιέστε το, απευθείας μέσα από το R, και ανακτήστε στο χώρο εργασίας σας ή οπουδήποτε αλλού επιθυμείτε στον Η/Υ σας, όλα τα περιεχόμενα επιμέρους αρχεία δεδομένων με τις ονομασίες *faithful…* που θα χρειαστείτε παρακάτω. Τα εν λόγω αρχεία προέρχονται από το προ-φορτωμένο σύνολο δεδομένων του R 'faithful' που περιλαμβάνει τους χρόνους αναμονής και διάρκειας εκτοξεύσεων (αναπήδησης) θερμού νερού από τον ονομαστό θερμοπίδακα *Old Faithful* στο Εθνικό Πάρκο Yellowstone, Wyoming, USA.

Εξοικειωθείτε με τη χρήση των συναρτήσεων ***read.table()***, ***read.csv()***, ή ***scan()*** προκειμένου να εισάγετε στο R τα δεδομένα από τα ακόλουθα αρχεία και να δημιουργήσετε αντίστοιχα αντικείμενα με τις παρακάτω αναφερόμενες ονομασίες ***(→ αντικείμενο του R***). Χρησιμοποιήστε καθεμία από τις εν λόγω συναρτήσεις για καθένα από τα παρακάτω αρχεία προκειμένου να δείτε τις μεταξύ τους διαφορές:

- ✓ Χρησιμοποιήστε τα αναφερόμενα αρχεία απευθείας από το ντοσιέ εργασίας που έχετε ορίσει για την τρέχουσα συνεδρία σας στο R (εφόσον έχετε ήδη κατευθύνει τις εντολές χρήσης των συναρτήσεων ***download()*** και ***unzip()*** να κατεβάσουν και να αποσυμπιέσουν εκεί τα ζητούμενα αρχεί δεδομένων):

  1. ***faithful_commasep.csv*** – δεδομένα με οριοθέτες κόμματα (comma separated file), ονόματα μεταβλητών (στηλών) στην πρώτη γραμμή του αρχείου (header) → αποθηκεύστε τα δεδομένα στο αντικείμενο με την ονομασία ***faithful_commasep***
  2. ***faithful_commasep_missing.csv*** – τα ίδια δεδομένα με το προηγούμενο ερώτημα, αλλά με μερικές τιμές παρατηρήσεων να λείπουν (missing values) → ***faithful_commasep_missing***
  3. ***faithful_semi.csv*** – τα ίδια προηγούμενα δεδομένα, όπως στο (1) αλλά με οριοθέτες τον χαρακτήρα ';' → ***faithful_semi***
  4. ***faithful_zsep.csv*** – τα ίδια προηγούμενα δεδομένα, όπως στο (1) αλλά με οριοθέτες τον χαρακτήρα 'z' → ***faithful_zsep***
  5. Επαναλάβατε το κάθε ένα από τα προαναφερόμενα βήματα (1.) έως (4.) παραλείποντας να αναγνώσετε την πρώτη γραμμή του αρχείου, και δίνοντας αντί αυτού τα ονόματα 'eruption' και 'waiting' των στηλών με διαφορετικό τρόπο εκχώρησης.

- ✓ Αναγνώστε τα προαναφερόμενα ίδια αρχεία απευθείας, χρησιμοποιώντας τις συναρτήσεις ***file.choose()*** ή ***choose.files()*** προκειμένου να επιλέξετε διαδραστικά το εκάστοτε αρχείο από το ντοσιέ όπου αυτό βρίσκεται στον Η/Υ σας. Εκτελέστε τα ίδια βήματα (1.) έως (4.).

- ✓ Αναγνώστε πάλι τα προαναφερόμενα αρχεία, αυτή τη φορά απευθείας από το ντοσιέ ***myrdata***, χρησιμοποιώντας τον κατάλληλο URL σύνδεσμο μεταφόρτωσής τους ως όρισμα εισόδου στην εκάστοτε εντολή ανάγνωσης από τον ιστοχώρο του μαθήματος (π.χ. http://portal.survey.ntua.gr/main/labs/hgeod/myrdata/faithfulfile.extension/). Εκτελέστε πάλι τα ίδια βήματα (1.) έως (4.).

- ✓ Τέλος δοκιμάστε να διαβάσετε το ίδιο αρχείο δεδομένων 'faithful' από τον ιστοχώρο http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat όπου το αρχείο ***faithful.dat*** περιέχει στην αρχή του πολλαπλές γραμμές επεξηγήσεων (multi-line header).

Έχοντας δημιουργήσει τα αντικείμενα ***faithful_commasep, faithful_commasep_missing, faithful_semi, faithful_zsep*** δοκιμάστε στο καθένα από αυτά μερικές δικές σας επιπλέον εντολές που θα σας βοηθήσουν να μετατρέψετε τα δεδομένα σας σε διαφορετικούς τύπους αντικειμένων.

Επιπλέον, χρησιμοποιώντας ένα από τα προηγούμενα αντικείμενα δεδομένων που δημιουργήσατε υπολογίστε τα ακόλουθα βασικά στατιστικά μεγέθη για τα χαρακτηριστικά του θερμοπίδακα Old Faithful:

- Χρησιμοποιήστε τη γενική συνάρτηση ***summary()*** για να δώσετε περιληπτικά αποτελέσματα βασικών στατιστικών μέτρων (μέσες τιμές, διασπορά, τεταρτημόρια κ.ά.) που αφορούν στους χρόνους των εκρήξεων και αναμονής του θερμοπίδακα.
- Κάνοντας χρήση της συνάρτησης ***hist()***, δημιουργήστε ένα ιστόγραμμα της μεταβλητής των εκρήξεων (eruptions) που να δείχνει τον αριθμό των εκρήξεων του θερμοπίδακα Old Faithful ταξινομημένων σύμφωνα με τη διάρκεια τους. Μπορείτε να συμβουλευτείτε τυπικά παραδείγματα χρήσης της συνάρτησης hist() στο σύνδεσμο https://www.datamentor.io/r-programming/histogram/ .

Θεωρείστε το κάθε ένα ζεύγος των τιμών των εκρήξεων και αναμονής ως μια παρατήρηση με *συντεταγμένες* (*x,* y) τις εν λόγω τιμές. Στη συνέχεια, δώστε μια προεπισκόπηση των ζευγών των τιμών των δεδομένων των εκρήξεων και αναμονής με τη βοήθεια της συνάρτησης ***cbind()*** και ακολούθως δημιουργήστε ένα γράφημα διασποράς (scatterplot) χρησιμοποιώντας τη συνάρτηση ***plot()***. Διερευνήσετε περαιτέρω το συγκεκριμένο σύνολο δεδομένων, προσπαθώντας σχεδιάσετε μόνο τους χρόνους αναμονής ή τη διάρκεια της εκτόξευσης νερού – Σημειώστε ότι η συνάρτηση ***plot()*** σας επιτρέπει να ορίσετε μια ονομαστική στήλη από το σύνολο δεδομένων και να αγνοήσετε το υπόλοιπο ή/και να σχεδιάσει ένα διάγραμμα γραμμής *(line chart)* αντί να σχεδιάσετε ζεύγη σημείων (scatterplot). Σχολιάστε εάν διακρίνετε από τα τελευταία γραφήματα κάποια μοτίβα στα δεδομένα που απεικονίσατε; Εάν είναι δύσκολο να διακρίνετε κάποια μοτίβα χρησιμοποιώντας το σύνολο των δεδομένων (δηλ. από τα στοιχεία 272 παρατηρήσεων), χρησιμοποιήστε την εντολή ***head ()*** ώστε να επιλέξετε τις πρώτες ***n*** σειρές των δεδομένων, και να δημιουργήσετε αντίστοιχα διαγράμματα γραμμής όπως πριν, αλλά περιορίζοντας την ποσότητα των δεδομένων που σχεδιάστηκαν. Δοκιμάστε να αλλάξετε την τιμή του ***n*** για να χρησιμοποιήσετε περισσότερα ή λιγότερα δεδομένα.

- Ανακτήστε από τον ιστοχώρο των δεδομένων της παρούσης εργασίας το R Script ***oldfaithful.R*** και εκτελέστε βήμα-βήμα τις εντολές που περιέχονται σε αυτό προκειμένου να δείτε μια πληρέστερη ανάλυση των δεδομένων για τον θερμοπίδακα *Old Faithful*.

Στα επόμενα ερωτήματα δοκιμάστε μερικές υποθέσεις σχετικά με τα διαθέσιμα δεδομένα για τον θερμοπίδακα *Old Faithful*. Καταρχή, διαχωρίστε το πλαισιο των πρωτογενων δεδομένων σε δύο χωριστά πλαίσια δεδομένων: (1) εκείνο με τις καταχωρήσεις των χρόνων αναπήδησης (eruptions) μικρότερες από 3 λεπτά και (2) εκείνο με τις καταχωρήσεις με χρόνους αναπήδησης μεγαλύτερους ή ίσους με 3 λεπτά. Απαντήστε τα παρακάτω ερωτήματα σχετικά με τον χρόνο αναμονής (waiting):

(a) Για τις καταχωρίσεις με μικρούς χρόνους αναπήδησης, υπολογίστε πόσες φορές η αναπήδηση σχετιζόταν με αναμονή που διαρκεί λιγότερο από 60 λεπτά;
(b) Αντίστοιχα, για τις καταχωρίσεις με μεγάλους χρόνους διάρκειας αναπήδησης, υπολογίστε πόσες φορές ο σχετικός χρόνος αναμονής διαρκεί λιγότερο από 80 λεπτά.
(c) Ποιός είναι ο μέσος όρος της διάρκειας αναπήδησης των προαναφερόμερων συμβάντων (a) και (b).

1.(b) Υπάρχουν διάφοροι τρόποι για να εξακριβώσετε τα σύνολα δεδομένων που περιλαμβάνονται στο R, π.χ., βλ. https://vincentarelbundock.github.io/Rdatasets/datasets.html.

Θυμηθείτε ότι για να δείτε ποια είναι τα προ-φορτωμένα σύνολα δεδομένων μπορείτε να χρησιμοποιείστε τη συνάρτηση ***data()*** από το πακέτο utils.

Χρησιμοποιήστε τη συνάρτηση ***data()*** με κατάλληλα ορίσματα στην εντολή κλήσης της, ώστε το R να σας δώσει μια λίστα με όλα τα σύνολα δεδομένων στα διαθέσιμα πακέτα (δηλαδή και τα μη φορτωμένα στον Η/Υ).

Ακολούθως, χρησιμοποιήστε πάλι τη συνάρτηση ***data()*** με κατάλληλα ορίσματα στην εντολή κλήσης της, ώστε το R να σας δώσει μια λίστα με όλα τα σύνολα δεδομένων στο πακέτο ***boot***.

Συγκεκριμένα, για τους σκοπούς της άσκησης από τα ακόλουθα σύνολα δεδομένων, επιλέξτε τρία σύνολα δεδομένων, το καθένα από τα αναφερόμενα διαφορετικά πακέτα:

- ***sunspot.month, airquality, nottem, quakes (από το πακέτο datasets),***
- ***bomsoi2001, rainforest (από το πακέτο DAAG)***,
- ***salinity (από το πακέτο robustbase)***.

(1) Αφού προηγουμένως επιθεωρήσετε τη δομή και τα πεδία των δεδομένων και των μεταβλητών που συμπεριλαμβάνονται στο εκάστοτε αρχείο, π.χ. την κλάση των αντικειμένων που αποθηκεύει το R σε κάθε περίπτωση, ακολούθως για κάθε ένα από τα σύνολα δεδομένων που έχετε επιλέξει, κάτω από τα ακόλουθα ενδεικτικά σχόλια εισάγετε τις κατάλληλες εντολές προκειμένου να επιτύχετε το αναφερόμενο αποτέλεσμα με το εκάστοτε επιλεγμένο αρχείο των δεδομένων σας:

```
# Get size in rows by columns.
# Get the names of variables in the dataset.
# See the internal structure of the dataset
# See the first and last 10 lines of the dataset

# Retrieve a column vector of the dataset.
# Then retrieve the same column vector by its name.
# Also, retrieve it with the "$" operator instead of the double square bracket operator.

# Extract some elements from one of the columns of the dataset.
# Retrieve the data elements of a single row or of multiple (consecutive or otherwise) rows.
# Retrieve some rows by their names, and furthermore using a logical index vector.

# Retrieve a data frame slice with the names of any two columns of the dataset used in an
# index vector inside the single square bracket operator.

# Sort the data frame in ascending or descending order of one of its variables, and put it in a
# new data.frame called 'dataset.sorted', όπου dataset είναι η ονομασία του
# χρησιμοποιούμενου συνόλου δεδομένων

# See the first and last 20 lines of the sorted dataframe

# Write the elements of the sorted data.frame into a .txt file
```

a) Από το ντοσιέ ***myrdata*** τις ιστοσελίδες του μαθήματος μεταφορτώστε στο χώρο εργασίας σας στο R το συμπιεσμένο αρχείο ***datasets_atm.zip*** το οποίο περιέχει τα επιμέρους αρχεία ***'Athina0_atm.txt'***, ***'Athina1_atm.txt'***, ***'Athina2_atm.txt'***, ***'Athina_atm.txt'***. Αυτά περιλαμβάνουν τα ίδια ατμοσφαιρικά δεδομένα από το παλαιό αεροδρόμιο της Αθήνας στις 12:00, 15/11/2016, αλλά διαδοχικά τα αρχεία είναι χωρίς header record, με header record μιας γραμμής (με τα μετρούμενα μεγέθη), με header record δύο γραμμών (δηλ. και μιας επιπλέον γραμμής με τις μονάδες των μετρούμενων μεγεθών), και με πλήρη header πολλαπλών (36) γραμμών για την επεξήγηση των χαρακτηριστικών του σταθμού και των μετρήσεων. Παρόμοια ατμοσφαιρικά δεδομένα, με πλήρη header, περιλαμβάνονται στα αρχεία ***Trapani_atm.txt*** και ***Vienna_atm.txt*** τα οποία επίσης συμπεριλαμβάνονται στο προαναφερόμενο συμπιεσμένο αρχείο.

i. Για κάθε ένα από τα παραπάνω .txt αρχεία εκτυπώστε το αντίστοιχο header και δώστε τις κατάλληλες εντολές προκειμένου να επιθεωρήσετε στην οθόνη σας τα περιεχόμενα του αρχείου στην αρχή και το τέλος των δεδομένων, καθώς επίσης και τον αριθμό των στηλών και γραμμών των δεδομένων.

ii. Εισαγάγετε στο R τα δεδομένα από το τοπικό αρχείο '***Athina0_atm.txt***' (εκείνο που δεν περιέχει header). Όταν εκτυπώνετε στην οθόνη σας τα περιεχόμενα του αρχείου στην αρχή και το τέλος των δεδομένων, το R χρησιμοποιεί για ετικέτες των στηλών τις ονομασίες μεταβλητών **V1, V2, …** κλπ.

   - Χρησιμοποιείστε την κατάλληλη εντολή με την παράμετρο επιλογής ***col.names*** προκειμένου να χρησιμοποιήσετε τις ονομασίες **'PRES HGHT TEMP DWPT RELH MIXR DRCT SKNT THTA THTE THTV'** και τυπώστε πάλι τα περιεχόμενα του αρχείου στην αρχή και το τέλος των δεδομένων, ώστε να βεβαιωθείτε ότι το R χρησιμοποιεί τις επιθυμητές ετικέτες για κάθε στήλη.

   - Επιλέξτε και τυπώστε τις γραμμές του εν λόγω πλαισίου δεδομένων για τις οποίες οι τιμές της θερμοκρασίας είναι θετικές. Ακολούθως εξάγετε το υποσύνολο των εν λόγω γραμμών των δεδομένων σε ένα νέο αρχείο '***output_Athina0_atm.csv***'.

   - Χρησιμοποιήστε τις κατάλληλες εντολές προκειμένου

     - Να βρείτε τη μέγιστη τιμή της υγρασίας της ατμόσφαιρας (η παράμετρος στην 5η στήλη του πλαισίου)
     - Να βρείτε το ύψος στο οποίο μετρήθηκε η εν λόγω τιμή, καθώς και την αντίστοιχη θερμοκρασία που μετρήθηκε εκεί.
     - Να απομονώσετε τις γραμμές του πλαισίου των δεδομένων, για τις οποίες η τιμή στα κελιά της στήλης 7 είναι 270.

2.(a) Θεωρείστε τα πρόσφατα μετεωρολογικά δεδομένα 56 ημερών από έξι (6) μετεωρολογικούς σταθμούς σε αεροδρόμια της χώρας (Αθήνα, Κοζάνη, Σούδα, Σύρος, Σαντορίνη, Κάρπαθος) στα επιμέρους αρχεία σε μορφότυπο ***.xls*** και ονομασίες της μορφής ***wxobservations_CODE_Location56_Hgtm***, όπου ***CODE*** είναι 4ψηφιος κωδικός του μετεωρολογικού σταθμού, ***Location*** είναι η τοποθεσία του σταθμού και το 56 υποδηλώνει 56 ημέρες δεδομένων, ***Hgtm*** δίνει το υψόμετρο του σταθμού σε μέτρα πάνω από τη Μέση Στάθμη της Θάλασσας (ΜΣΘ). Συγκεκριμένα, τα επιμέρους αρχεία ενδιαφέροντος είναι:

- wxobservations_LGAV_ELVEN56_94m (Αθήνα, Αεροδρόμιο Ελ. Βενιζέλος)
- wxobservations_LGKZ_KOZANI56_628m (Κοζάνη)
- wxobservations_LGSA_SOUDA56_149m (Σούδα, Κρήτης)
- wxobservations_LGSO_SYROS56_72m (Σύρος)
- wxobservations_LGSR_Santorini56_38m (Σαντορίνη)
- wxobservations_LGKP_Karpathos56_20m (Κάρπαθος)

βρίσκονται, ξεχωριστά το καθένα, στο ντοσιέ ***myrdata*** στο σχετικό ιστοχώρο του μαθήματος, σε ένα συμπιεσμένο αρχείο με την ονομασία ***wxobservations.zip***.

Αρχικά, χρησιμοποιώντας την εντολή ***download()***, μέσα από το R, μεταφορτώστε το αρχείο ***wxobservations.zip*** στον Η/Υ σας και με την εντολή ***unzip()*** αποσυμπιέστε τα επιμέρους αρχεία των δεδομένα στο ντοσιέ εργασίας σας για το συγκεκριμένο τμήμα της άσκησης. Με την αποσυμπίεση, θα πρέπει να έχετε στο ντοσιέ εργασίας σας όλα τα παραπάνω αρχεία τόσο σε μορφότυπο ***.xls***, όσο και σε ***.csv***.

✓ Αναγνώστε τα δεδομένα από δύο αρχεία τύπου ***.txt*** (για δύο από τις παραπάνω τοποθεσίες) χρησιμοποιώντας κατάλληλες συναρτήσεις από τις κατηγορίες των συναρτήσεων A (R build-in functions) ή C (συναρτήσεις του πακέτου ***readr***) και με κατάλληλες εντολές να διερευνήσετε το περιεχόμενο και τη δομή των δεδομένων και των μεταβλητών που συμπεριλαμβάνονται στο εκάστοτε αρχείο.

✓ Αναγνώστε τα δεδομένα από δύο αρχεία τύπου ***.xls*** για δυο διαφορετικές τοποθεσίες και επαναλάβετε την ίδια διαδικασία χρησιμοποιώντας τις συναρτήσεις του πακέτου ***xlsx***.

✓ Αναγνώστε τα δεδομένα από δύο αρχεία τύπου ***.xlsx*** για τις δυο τελευταίες τοποθεσίες και επαναλάβετε την ίδια διαδικασία χρησιμοποιώντας τις συναρτήσεις του πακέτου ***openxlsx***.

- ✓ Αναγνώστε από τον ιστοχώρο του μαθήματος τα δεδομένα του αρχείου ***wxobservations_6sites.xls*** το οποίο περιέχει 6 φύλλα εργασίας με ακριβώς τα ίδια προηγούμενα δεδομένα των 6 μετεωρολογικών σταθμών. Εγκαταστήστε και φορτώστε το πακέτο ***XLConnect*** του R και χρησιμοποιώντας τις κατάλληλες συναρτήσεις με αντίστοιχες εντολές διερευνήσετε το περιεχόμενο και τη δομή των δεδομένων και των μεταβλητών που συμπεριλαμβάνονται στα φύλλα εργασίας που αντιστοιχούν σε δύο (της επιλογής σας) από τις τοποθεσίες των μετεωρολογικών σταθμών.

2.(b) Από το ντοσιέ ***myrdata*** στο σχετικό ιστοχώρο του μαθήματος κατεβάστε στον Η/Υ σας το συμπιεσμένο αρχείο ***gmsl.zip*** το οποίο περιέχει επιμέρους αρχεία ***.csv*** με δεδομένα που αναφέρονται στις "σωρευτικές" μεταβολές *(cumulative changes)* στη στάθμη της θάλασσας για τους ωκεανούς του κόσμου από το 1880 μέχρι πρόσφατα, με βάση ένα συνδυασμό μακροχρόνιων (από το 1880) επίγειων μετρήσεων παλιρροιογράφων σε παράκτιες και νησιωτικές περιοχές και πρόσφατων (από το 1993) μετρήσεων από δορυφόρους αλτιμετρίας. Τα δεδομένα προέρχονται από τους οργανισμούς προστασίας του Περιβάλλοντος των ΗΠΑ και της Αυστραλίας (EPA και CSIRO αντίστοιχα).

Ζητείται, με κατάλληλο τρόπο της επιλογής σας, να αναγνώσετε τα εν λόγω αρχεία στο R και να εκτελέσετε σειρά βασικών στατιστικών αναλύσεων (π.χ. να υπολογίσετε διάφορα στατιστικά μέτρα, όπως μέσες τιμές, τυπική απόκλιση κλπ. ή/και να απεικονίσετε με απλά γραφήματα τυχόν διαφαινόμενες τάσεις στις εν λόγω μεταβολές, π.χ. να εκτελέσετε κάποιες γραμμικές παλινδρομήσεις). Ο στόχος των αναλύσεων σας θα πρέπει να εστιάσει στην προσπάθεια να απαντήσετε σε τυπικά ερωτήματα όπως:

- Ποιος είναι ο εκτιμώμενος ρυθμός ανόδου της ΜΣΘ από τα επίγεια δεδομένα και, αντίστοιχα, από τα δορυφορικά δεδομένα, κατά το διάστημα 1993-2009;
- Ποια είναι η παγκόσμια μέση αύξηση της στάθμης της θάλασσας από το 1880 μέχρι το 2009;
- Ποια είναι η γραμμική τάση από το 1900 έως το 2009; και αντίστοιχα, από το 1961 και μετά;
- Διαφαίνεται από τα δεδομένα εάν υπάρχει σημαντική μεταβλητότητα στον ρυθμό αύξησης κατά τον εικοστό αιώνα;
- Διαφαίνεται από τα δεδομένα κάποια μεταβολή της στάθμης της θάλασσας από το 1990 έως το 1993, πιθανότατα ως αποτέλεσμα της ηφαιστειακής έκρηξης του Mount Pinatubo το 1991;

Στο ίδιο συμπιεσμένο αρχείο ***gmsl.zip*** περιλαμβάνονται τα δεδομένα των προηγούμενων αρχείων σε έναν διαφορετικό μορφότυπο αρχείων τύπου ***.json*** *(JavaScript Object Notation)*, των οποίων η ανάγνωση και ο χειρισμός μπορεί να γίνει με εντολές του πακέτου ***jsonlite***, βλ. τα παραδείγματα στους συνδέσμους

- https://cran.r-project.org/web/packages/jsonlite/vignettes/json-aaquickstart.html
- https://cran.r-project.org/web/packages/jsonlite/jsonlite.pdf
- https://www.tutorialspoint.com/r/r_json_files.htm

Ακολουθώντας τα προηγούμενα παραδείγματα δοκιμάστε να αναγνώσετε μερικά από τα διαθέσιμα αρχεία ***.json*** και υπολογίστε αντίστοιχα στατιστικά μέτρα για τις παραμέτρους στις οποίες αναφέρονται τα δεδομένα.

3. Από το διαδίκτυο ανακτήστε τα ακόλουθα αρχεία θερμοκρασίας που βρίσκονται στους συνδέσμους:
https://github.com/datasets/global-temp/blob/master/data/annual.csv ,
https://github.com/datasets/global-temp/blob/master/data/monthly.csv ,

ή χρησιμοποιώντας τους ακόλουθους συνδέσμους και τη συνάρτηση ***scan()***
https://raw.githubusercontent.com/datasets/global-temp/master/data/annual.csv
https://raw.githubusercontent.com/datasets/global-temp/master/data/monthly.csv

Τα παρεχόμενα στοιχεία προέρχονται από τη NASA *(GISS Surface Temperature (GISTEMP) Analysis)* και το ερευνητικό πρόγραμμα *Global Component of Climate at a Glance* (GCAG) και περιέχουν δύο σύνολα δεδομένων:

- *Global monthly mean,*
- *Annual mean temperature anomalies in degrees Celsius from 1880 to the present.*

Κάνετε το ίδιο με τα ακόλουθα αρχεία ανωμαλιών θερμοκρασίας που βρίσκονται στους συνδέσμους:

https://github.com/datasets/global-temp-anomalies/blob/master/data/global-temp-5yr.csv
https://github.com/datasets/global-temp-anomalies/blob/master/data/global-temp-annual.csv ,

ή αντίστοιχα χρησιμοποιώντας τους ακόλουθους συνδέσμους και τη συνάρτηση ***scan()***
https://raw.githubusercontent.com/datasets/global-temp-anomalies/master/data/global-temp-5yr.csv
https://raw.githubusercontent.com/datasets/global-temp-anomalies/master/data/global-temp-annual.csv

Τα παρεχόμενα στοιχεία προέρχονται από τη NASA (GISS Surface Temperature (GISTEMP) Analysis) και περιέχουν τα δεδομένα

- Global Annual Temperature Anomalies (Land) 1880-2014,
- Global Annual Temperature Anomalies (Land and Ocean) 1880-2014,
- Hemispheric Temperature Anomalies (Land+ Ocean) 1880-2014 and
- Annual Temperature anomalies (Land + Ocean) for three latitude bands that cover 30%, 40% and 30% of the global area, respectively, 1900-2014.

Με κατάλληλες εντολές προσπαθήστε να εξακριβώσετε τα ακόλουθα:

- Πόσο θερμός συγκαταλέγεται ο Οκτώβριος του 2017 σε σύγκριση με τον ίδιο μήνα στα 137 χρόνια σύγχρονης δειγματοληψίας τιμών από τη μηνιαία ανάλυση των παγκόσμιων θερμοκρασιών; Ποιος ήταν ο θερμότερος Οκτώβριος μεταξύ όλων των ετών;
- Πόσο θερμός συγκαταλέγεται ο Ιούνιος του 2017 σε σύγκριση με τον ίδιο μήνα των υπόλοιπων ετών;
- Πόσο διαφέρουν οι τιμές της παγκόσμιας θερμοκρασίας για τους 12 μήνες από τον Δεκέμβριο του 2001 μέχρι τον Νοέμβριο του 2002, σε σχέση με τη μέση τιμή για την περίοδο 1951-1980; Πόσο διαφέρει η μέση τιμή των εν λόγω 12 μηνών, από τη μέση τιμή για την περίοδο 1951-1980;
- Πόσο θερμό ήταν το καλοκαίρι (μήνες Ιούνιος-Ιούλιος-Αύγουστος) του 2010, σε σχέση με τα 137 χρόνια καταγραφών και αντίστοιχα σε σχέση με το καλοκαίρι του 2009;
- Ποιες ήταν οι τάσεις/χαρακτηριστικά της παγκόσμιας θερμοκρασίας το 2004, σε σχέση με τον κλιματολογικό μέσο της περιόδου 1951-1980;
- Ποιο έτος ήταν το ψυχρότερο μετά από το 2000; Πόσο θερμό συγκαταλέγεται το ίδιο έτος, σε σχέση με τα άλλα έτη από το 1880; Ποια ήταν τα πέντε θερμότερα έτη από το 1890;

4. (a) Επιλέξτε δύο διαφορετικά πλαίσια δεδομένων από εκείνα που έχετε ήδη επεξεργαστεί στο ερώτημα **1.(β)** και διαδοχικά εξάγετε τα αποθηκευμένα στοιχεία τους σε ένα αρχείο ***.csv*** και σε ένα αρχείο κειμένου ***.txt*** χρησιμοποιώντας τις συναρτήσεις ***write.csv***, ***write.table*** και ***write.delim()***. Δοκιμάστε διαφορετικές επιλογές προκειμένου να εξαιρέσετε τα ονόματα γραμμών και στηλών, να καθορίσετε τι να χρησιμοποιήσετε για τις τιμές που λείπουν, να προσθέσετε ή να αφαιρέσετε χαρακτήρες σε ονοματα στηλών ή γραμμών κ.λπ.

(b) Εξαγάγετε το ενσωματωμένο στο R σύνολο δεδομένων ***mtcars*** στον Η/Υ σας. Δημιουργήστε έναν υποφάκελο με τον τίτλο "exportdata" στον κατάλογο εργασίας σας. Τώρα αποθηκεύστε τα δεδομένα του ***mtcars*** σε ένα αρχείο ***.csv*** στον υποφάκελο.

Εξαγάγετε το ενσωματωμένο συνόλο δεδομένων ***iris*** σε ένα αρχείο ***.csv*** με κωδικοποίηση UTF-8 χρησιμοποιώντας τη συνάρτηση ***write_excel_csv***.

Εξαγάγετε το ενσωματωμένο σύνολο δεδομένων ***USArrests*** ως αρχείο δεδομένων διαχωρισμένων με tabs.

Δημιουργήστε μια λίστα με την ονομασία ***lista*** που να περιέχει τα πλαίσια δεδομένων ***iris*** και ***mtcars***, και ακολούθως χρησιμοποιήστε τη συνάρτηση ***write.xlsx(x, file, …)*** **από το πακέτο openxlsx** προκειμένου να δημιουργήσετε (i) ένα απλό αρχείο excel με δύο φύλλα excel με τις ονομασίες ***iris*** και ***mtcars***, και (ii) ένα αρχείο excel με δύο αντίστοιχα excel workbooks με τις ονομασίες ***iris*** και ***mtcars***, των οποίων κάθε στήλη του αντίστοιχου πίνακα έχει ενεργοποιημένο το φιλτράρισμα στη γραμμή της κεφαλίδας της στήλης, ώστε να μπορείτε να φιλτράρετε ή να ταξινομήσετε γρήγορα τα δεδομένα του πίνακα τιμών του workbook.

(c) Δημιουργήστε ένα κενό βιβλίο εργασίας (excel workbook) και αποθηκεύστε τα πλαίσια δεδομένων των αρχείων ***annual.csv*** και ***monthly.csv*** από το ερώτημα (3) σε ξεχωριστά φύλλα εργασίας μέσα στο ίδιο (αυτό που δημιουργήσατε) βιβλίο εργασίας. **Χρησιμοποιήστε για αυτό το σκοπό το πακέτο xlsx.**

Επιπλέον, στο ίδιο excel workbook, προσθέστε δύο ακόμα φύλλα εργασίας που να περιέχουν αντίστοιχα τα σύνολα δεδομένων του ***mtcars*** *και* δεδομένων του ***iris.*** Δοκιμάστε, χρησιμοποιώντας τα σύνολα δεδομένων ***mtcars*** *και* δεδομένων ***iris,*** να δημιουργήσετε μερικά επιπλέον ξεχωριστά φύλλα εργασίας στα οποία να κάνετε κάποια δική σας μορφοποίηση (αλλαγές) στα ονόματα γραμμών και στηλών, ή επιχειρώντας να επεξεργαστείτε τίτλους, υπότιτλους, χρώμα γραμματοσειρών σε επιλεγμένα κελιά, να θέσετε προδιαγραφές για το πλάτος στήλης ή στηλών κλπ.

ΣΗΜΕΙΩΣΗ – Με τα παραδοτέα σας θα πρέπει να συμπεριληφθούν τα εκάστοτε ***.Rhistory*** αρχεία που θα προκύψουν από τις ξεχωριστές συνεδρίες του R που διεκπεραιώσατε για την υλοποίηση κάθε μέρους της Θεματικής Εργασίας.