Online Academic Data Analysis Bootcamp Using Open-Access Program R: Essentials

Erick and Patrick

Aug 10th, 2023

# Session 2 Exercises

# Prerequisites

1. You should have installed the following packages by running the following code. You should be connected to a stable internet during the installation (Note: Remove the # comment symbols to run the code line)

#install.packages(“writexl”, dependencies = TRUE); #install.packages(“foreign”, dependencies = TRUE); #install.packages(“haven”, dependencies=TRUE)

1. Download and save all the files sent in the session 2 email
2. Set your working directory to the folder you have saved today’s files sent via email

# Exercise 1

1. Use the c() / concatenate function to create a “neurosis” dataframe consisting of the following variables. The datapoints in the variables are arranged in the same order for each case (name)

* Variable “name” “Ben”, “Martin”,“Andy”,“Paul”, “Graham”,“Carina”,“Karina”,“Doug”,“Mark”, “Zoe”
* Variable “birth\_date”

“1977-07-03”, “1969-05-24”, “1973-06-21”, “1970-07-16”, “1949-10-10”, “1983-11-05”, “1987-10-08”, “1989-09-16”, “1973-05-20”, “1984-11-12”

* Variable “marriage\_date” “1997-04-03”, “1999-04-24”, “2003-12-25”, “1996-03-11”, “1976-05-07”, “2010-03-02”, “2012-02-01”, “2011-10-05”, “1996-03-12”, “2009-02-01”
* Variable “friends” 5,2,0,4,1,10,12,15,12, 17
* Variable “alcohol” 10,15,20,5,30,25,20,16,17,18
* Variable “initial\_income” 20000,40000,35000,22000,50000,5000,100,3000,10000,10
* Variable “neurotic” 10,17,14,13,21,7,13,9,14,13

1. You collect another income variable after 5 years. Please add this variable to the “neurosis” dataset using the c() function. The new income variable is as follows:

* Variable “current\_income” 30000,55000,50000,35000,65000,15000,1500,5500,23000,510

1. Calculate and add a new variable “income\_increase” to the “neurosis” dataset which expresses the percent increase in income since employment Hint: percent increase is ((initial\_income - current\_income)/ initial\_income) \* 100
2. Calculate the age at marriage for each of the study participants
3. Compute and add a coding /factor variable “sex” where the first half of the participants are female and the next half are male.

# Exercise 2

The data *airquality* which is already in your R environment presents daily air quality measurements in New York, May to September 1973. As is typical of many data collection exercises, this data contains missing values.

1. Calculate the mean Ozone concentration
2. List rows of the data that have missing values
3. Create a new dataset “Ozone\_2” without missing data and use it to calculate the mean and standard deviation of the wind speed, temperature and ozone concentration

# Exercise 3

The dataset “Soil\_composition.xlsx” consists of soil characteristics that were measured on samples from three types of contours (Top, Slope, and Depression) and at four depths (0-10cm, 10-30cm, 30-60cm, and 60-90cm). The area was divided into 4 blocks, in a randomized block design.

1. Save the dataset as a *Comma delimited file* (with a .csv file extension) extension and as a *Text file* (Tab Delimited with .txt file extension)
2. Import and view (using View()) the two datasets from a) above into R and name them Soil\_compositions\_csv (for csv) and Soil\_compositions\_txt (for text)
3. Import and view (using View()) the dataset directly into R naming it Soil\_compositions\_xlsx

# Exercise 4

The dataset “Monitoring\_the\_future.sav” presents surveys of 8th- and 10th-grade students are part of a series that explores changes in important values, behaviors, and lifestyle orientations of contemporary American youth. You are presented with this data after entry in SPSS and need to analyze it in R.

1. Import the dataset into R
2. Examining the first and last 5 rows in the dataset.
3. Examining the data by listing the variable names, looking at the structure of the data, number of rows and columns.

# Exercise 5

Export the dataset created in *Exercise 1* to:

1. A tab delimited text file
2. An Excel spreadsheet

# Exercise 6

This exercise is optional depending on availability of own data

Repeat Exercise 4 using your own dataset and if necessary compute and append any necessary new variables as done in Exercise 1