CADS Assignment 3 - Importing Data into R from my Computer and Visualization using ggplot2

As mentioned in Assignment #2. It is usually more efficient to have your data input in a spreadsheet like Excel. You can do some manipulation there to prepare the data and save it. You may also have data sets that are already prepared and ready for use. There are many formats in which a data set file may be saved. We will only focus on a few. Once you understand this, you can google and find the templates or syntax for using other files.

A: Importing Data into R from my computer

1. Use datasets available in libraries within R. The datasets are available immediately. You can experiment with them. You need to install packages and load as follows:

   Install.packages()
   Library (dplyr)

   There are datasets that are already built into R. That means these data sets are readily available for us to use. Once we install the packages and load the dplyr, we simply call the dataset by name.
   Caution: **These could be very large data sets. So always avoid printing**. Instead use head(), tail(), view().

   Exercise1: Select two datasets below and explore (We already used USArrests, so pick any two

   Explore the four of the following,
   How many rows and columns. Use dim()
   What are the column names?
   Identify the numeric variables.

   mtcars
   USArrests
   iris
   msleep
   world_bank_pop
   wordphones
   economics
   mpg
   morley

nhtemp
nottem
npk
women
sleep
state
occupationalStatus
islands
USPersonalExpenditure
UCBAdmissions
Titanic
CO2
AirPassengers
ChickWeight
UKDriveDeaths
etc.

Import Data from a file saved on your computer.
First, your data must be saved in a particular location for this to work.
Write the following command in R Studio:

getwd() # This tells you exactly where you must place your file to make it accessible.
For now, **let us avoid** using setwd() which could introduce complications.

So let's make a data file in Excel and save as ".csv", ".xlxs" or "txt"-These options are available to you when you save a file in Excel. Files in other formats like SAS, SPSS …can also be imported. Just do a search on the web and you will find the appropriate library function to use.

Please use the data file shown below to practice using these different files.

a. Read as a ".csv" file

```
library(dplyr)
read.csv("Path where csv file is stored\\fileName.csv")
# You can get the path by using:getwd  #This gets you your directory
# If your data is not there, you should go and open the file and save it there.
# You can then copy and paste into your code. No need to type.
#In my case, I get
read.csv("Path\\my_file.csv")
```

Exercise 1(10points): Create this data file in Excel and save as "MyFriends.csv"

Name	Age	Height	Married
Nicolas	27	180	TRUE
Thierry	25	170	FALSE
Bernard	29	185	TRUE
Jerome	26	169	TRUE

Import the saved file into R following instructions above.

Call the file "MyFriends.csv". Make sure the file is saved in the "Path" specified above. Place it in the appropriate directory and import into R. Call the data frame MyFriendsDf

b. Import xlxs files

```
# Install the package
install.packages("readxl")
# Loading the library
library(readxl))
# Using library("readxl") with read_excel ("Path\\my_file.xlxs, sheetIndex) should also work
# Notice that in R, double backslashes (\\) are used.
My_data<-read_excel("Path\\my_file.xlxs", sheetIndex, header=TRUE)
```

Exercise 2(10points): Make a file and save it in Excel as "MyTeams.xlxs".

Team	Points	Assists	Home
A	78	12	TRUE
B	85	20	FALSE
C	93	23	TRUE
D	90	8	TRUE

Import the file into R following the instructions above.

Place it in the appropriate directory and import into R. Call the data frame MyTeamDF

c. Import a txt file

# Read a txt from my saved folder
My_data <- read.table ("Path\\my_file.txt", header=TRUE").
Inside the R Studio, go to the bottom right pane. You should find your file there. You can bring it into your Script pane for work.

Exercise 3 (10points): Create a file in Excel but save as a txt file with the data below and save as "ForSaleHouse.txt".

Place it in the appropriate directory and import into R.

Room	SqFoot	Windows	Flooring
Family	600	6	Marble
Master	400	4	Marble
Living	500	4	Wood
Dining	225	2	Wood

Following the instructions above, import the file into R-Studio. Call the data frame ForSaleHouseDf.

2. You can bring a dataset directly from the web.
   This is probably the most challenging but it should work.
   I will post a guide for this in Canvas and you can explore on your own.
   In fact, this is a whole topic by itself called "web scraping" where one gathers data together from web sources into a usable form like a csv file for analysis.

3. What about writing a data file back into your computer?

   Export data frame as csv

   i) write.csv(df, "Path\\MyFile.csv", row.names=FALSE)

   # df is the name of your data frame you want to export. When using this method, specify

   # row.names=FALSE if you don't want R to export the row names.

   ii) library(readr)

   write_csv(df, "Path\\MyFile.csv")

Notice that once you export the file, you can open it in Excel and you can open it with a notepad to get a txt file.

Exercise 4 (10points): Take the data frame that you created above MyFriensDf and write it back but without the names-we won't confuse it with the original data that we created and saved.

B. Visualization with ggplot2

Base R has some plotting capability.

But the best facility for plotting in R requires an added library.

install.packages()

library(ggplot2)

I will encourage you to work through the posted link below or the Lc Carpentry that I already posted into Modules for you. You don't need to do everything or understand everything.

You need to be able to do the main types of plots:

1. Box plot
2. Bar plot
3. Line plot
4. Combining of line and plots
5. Histogram
6. Density plot
7. Combination of histogram and density.

Exercise 5. (60 points)

Using the ggplot2, produce visualization for the Father in the FatherandSon.txt that is uploaded for you.

Hint: Download file and save into the working directory.

Use getwd(), copy and paste;

Use read.table('Path.FatherandSon.txt",header=TRUE) # header=T tells R that the first line is a header.

Assign the name for the data frame as FatherandSon by using:

FatherandSon <- read.table('Path.FatherandSon.txt",header=TRUE)

Now check to make sure your data frame is ready. Try header(), dim() etc.

a) Now, using ggplot2, do a visualization (use the suitable plots from the list below) of the FatherandSon. Produce a visualization for the Father variable and then the Son variable.
b) Do a visualization of a scatter plot of x variable as Father and y variable as Son.
c) Calculate the correlation coefficient (see DataCamp pdf on Correlation posted in CADS Module in Canvas).
d) Draw a Regression line (See DataCamp pdf notes on Regression posted in CADS Module in Canvas).