Stat 291 - Recitation 5

Orçun Oltulu

26 / 11 / 2021

Last Week:

Recall the last exercise from Recitation 4.

Load 'Credit' data set from ISLR package. Read the document for the data set; '?Credit'.

```
library(ISLR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Part A.

Check the first 10 observations of Credit data set.

Part B.

Create a subset, new_credit1, for Asian married females.

Part C.

From new_credit1, select only numeric values and create new data frame, new_credit2.

Part D.

Find the mean of each numeric value in new_credit2.

Part F.

Now, find minimum and maximum Income for Asian males and females separately,

- Filter Asian people,
- Select only Gender and Income variables,
- Group them by Gender,
- Summarize using min and max functions.

Reading data files into R:

Exercise 1:

Read the file 'Table0.txt';

(a) Assign names to the columns to Name, Age, Height, Weight and Sex.

(b) Change the row names so that they are the same as Name, and remove the variable Name.

Exercise 2:

Read the file 'Table1.txt';

(a) How many rows and columns does it have?

(b) Reread the file and make the variable Name be the row names. Make sure you read the variable as characters and not as factors.

Exercise 3:

Read the file 'Table2.txt';

(a) What is the problem with the first and last columns ?

```
##          Name Age Height Weight Sex
## 1     /Alex/  25    177     57 /F/
## 2    /Lilly/  31    163     69 /F/
## 3     /Mark/  23    190     83 /M/
## 4   /Oliver/  52    179     75 /M/
## 5   /Martha/  76    163     70 /F/
## 6    /Lucas/  49    183     83 /M/
## 7 /Caroline/  26    164     53 /F/
```

(b) How can you fix that problem ?

Exercise 4:

Read the file 'Table3.txt';

(a) How many missing value does this data set have?

(b) Assign NA to each 'weird' value.

(c) Reread the data set but this time make sure you only have 'NA' values rather than {'*','**', "_"}

Exercise 5:

Read the file 'Table4.txt';

Watch out for the missing values and the decimal separator.

Exercise 6:

Read the file 'Table5.txt';

Watch out for the missing values and the decimal separator and the separator.

Exercise 7:

Read the file 'Table6.txt';

Check out the file first. Notice that the information is repeated, we only want the first non-repeated ones. Make sure to create only characters not factors this time around. Lastly, we don't want the comments.

Exercise 8:

Read the file 'states1.csv';

(a) The names of the states should be the row names.

(b) Check the dimensions of both 'df8' and 'df8.a' data.

Exercise 9:

Read the file 'states2.csv';

The names of the states should be the row names, watch out for the decimal separator and the separator.

Exercise 10:

Read the file 'states3.csv';

Watch out for the same as the last exercise plus the missing values. Add to the previous data set, column-wise.

Exercise 11:

Read 'iris.Rdata' into R.

Writing data files:

Exercise 12:

Using following commands create a data frame and write it to 'data1.txt' file.

```
set.seed(291)
vec1 <- rnorm(15, mean = 5, sd = 2)
vec2 <- sample(100, size = 15)
vec3 <- runif(15)
vec4 <- sample(c("Male","Female"),
               size = 15, replace = T)

data <- data.frame(vec1, vec2, vec3, vec4)
```

Exercise 13:

Write the same data set to 'data1.csv' file.

Add a new column c(15,25,0.5,"Female") to 'data1.csv' file.

Exercise 14:

Create an .RData file with same data set, name it 'data1.Rdata'.

Reading Data from Web:

Exercise 15:

Read .csv data from from github.

Go to the github repo with the following URL:

https://github.com/oltuluorcun/Stat291

Read the "data1.csv" file in the repo.

Hint: You have to use the **"Raw"** version of the data set and copy the raw versions URL.

Exercise 16:

Web scraping with "rvest" package.

Go to the following web page where you will see the current F1 standings for both Drivers and Constructors.

https://www.statsf1.com/en/2021.aspx

Then, using Selector Gagdet, find the 'location' of the tables.

Finally, do the magic and extract the both tables and read into R.

Check out for further info for usage of 'rvest' package:

https://www.kaggle.com/orcunoltulu1/web-scraping-in-r-rvest-package