# Prep Exercise (PE02) Dataframes and Modeling

### General Instructions

For this exercise you will upload this word file to blackboard. You can find the submission link by going to Weekly Content, Week 2, and then clicking on the PE02 DropBox near the lower 1/3 of your blackboard screen. Be sure to read the Homework guideline that defines what your first 3 lines of code (comments in this case) will look like.

A few highlights to pay close attention to:

* We expect that you will benefit from the code presented in Introduction to Data Science, as well as examples provided by the instructors, the Internet, and other sources. But you must always, always, always give credit to your sources. For example, if you find a line of code someone else wrote that helps you with developing a solution, by all means borrow that line of code, but make sure to use a comment to document who wrote the code. The same is true from helpful videos or tutorials from internet and YouTube.
* Questions regarding the data, code, or steps conducted within the Prep Exercise should be answered directly within this word document, attaching screenshots where necessary.
* # IST 687, Standard Homework Heading
* #
* # Student name: Thadhani Hitesh Chandrakumar
* # Homework number: PE02
* # Date due: Wed 4th Sep 2019 11:59PM
* #
* # Attribution statement: (choose the statements that are true)
* # 1. I did this work by myself, with help from the book and the professor
* # 2. I did this homework with help from the book and the professor and these Internet sources: <provide the urls>
* # 3. I did this homework with coaching from <Name of another student> but did not cut and paste any code

### Prep Exercise

1. **Create a new dataframe within R Studio using the ‘USArrests’ dataset (‘USArrests’ is preloaded into R Studio). ﻿**This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.
   1. Creating a dataframe using a given dataset:

**myArrests <- USArrests**

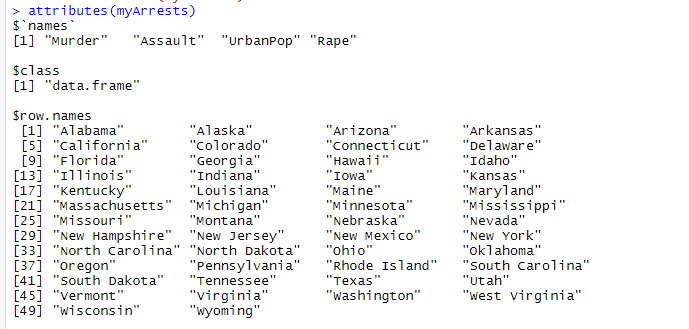
1. **Verify that your new dataframe was properly populated with the USArrests dataset.**
   1. Viewing the contents of a dataframe:

**View(myArrests)**

The output is pasted at the end of this question tabular representation of observations and variables 50 and 4 respectively.

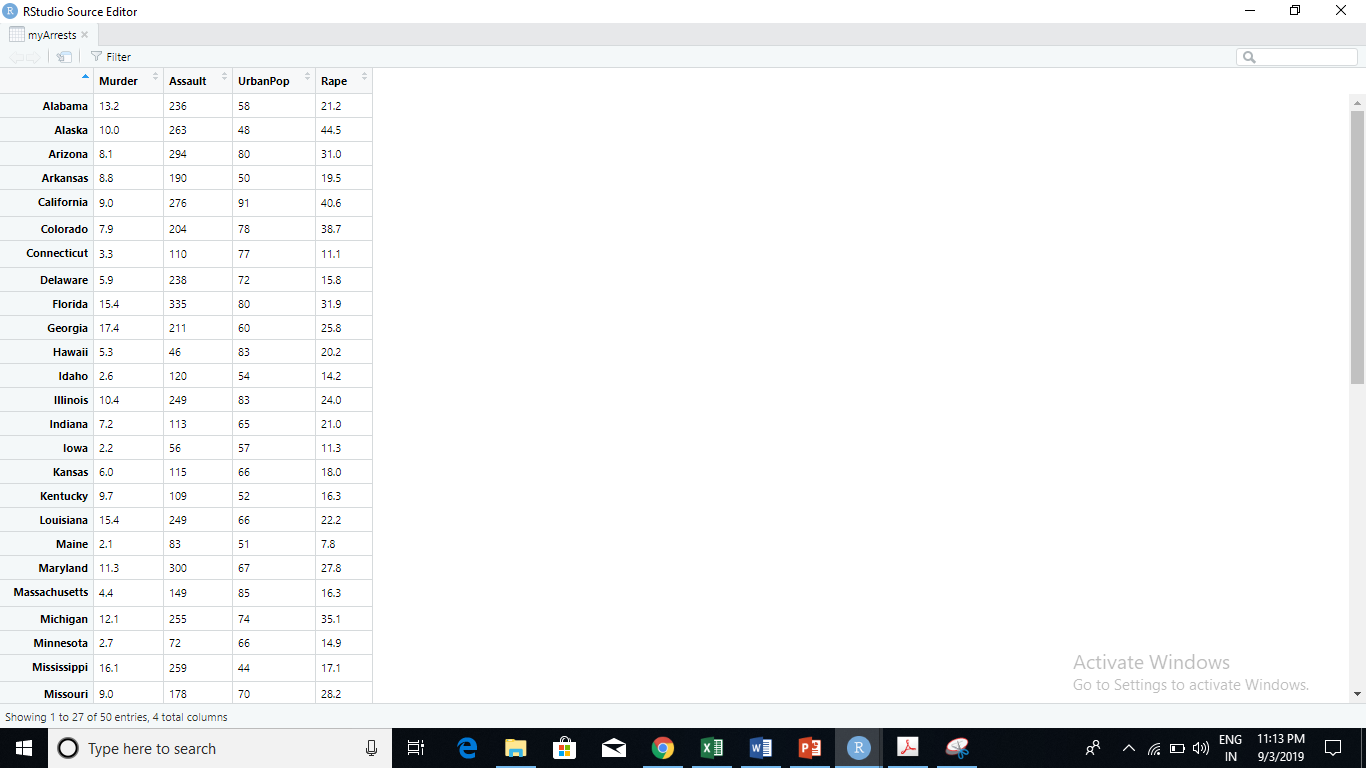
* 1. List the attributes of the dataframe:

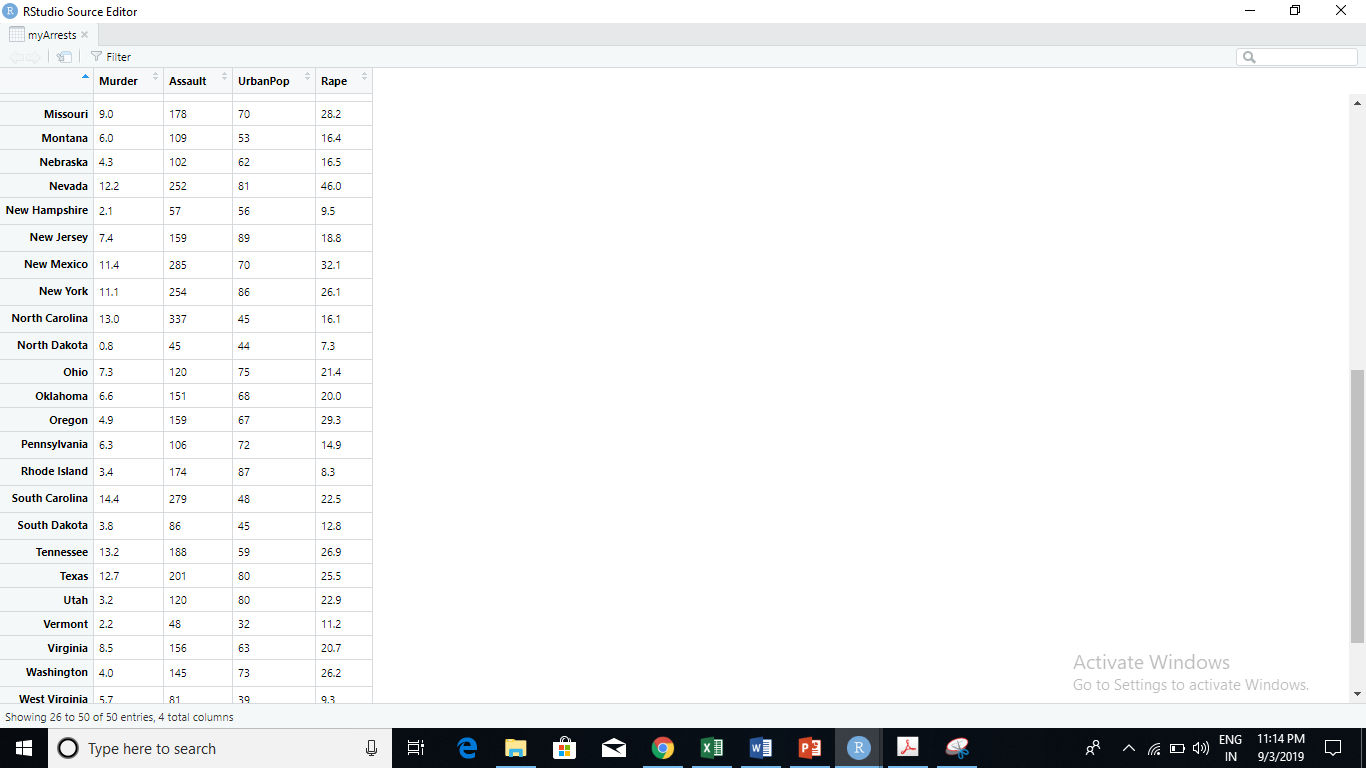
**Murder, Assault, UrbanPop, Rape (4 Variables/attributes)**

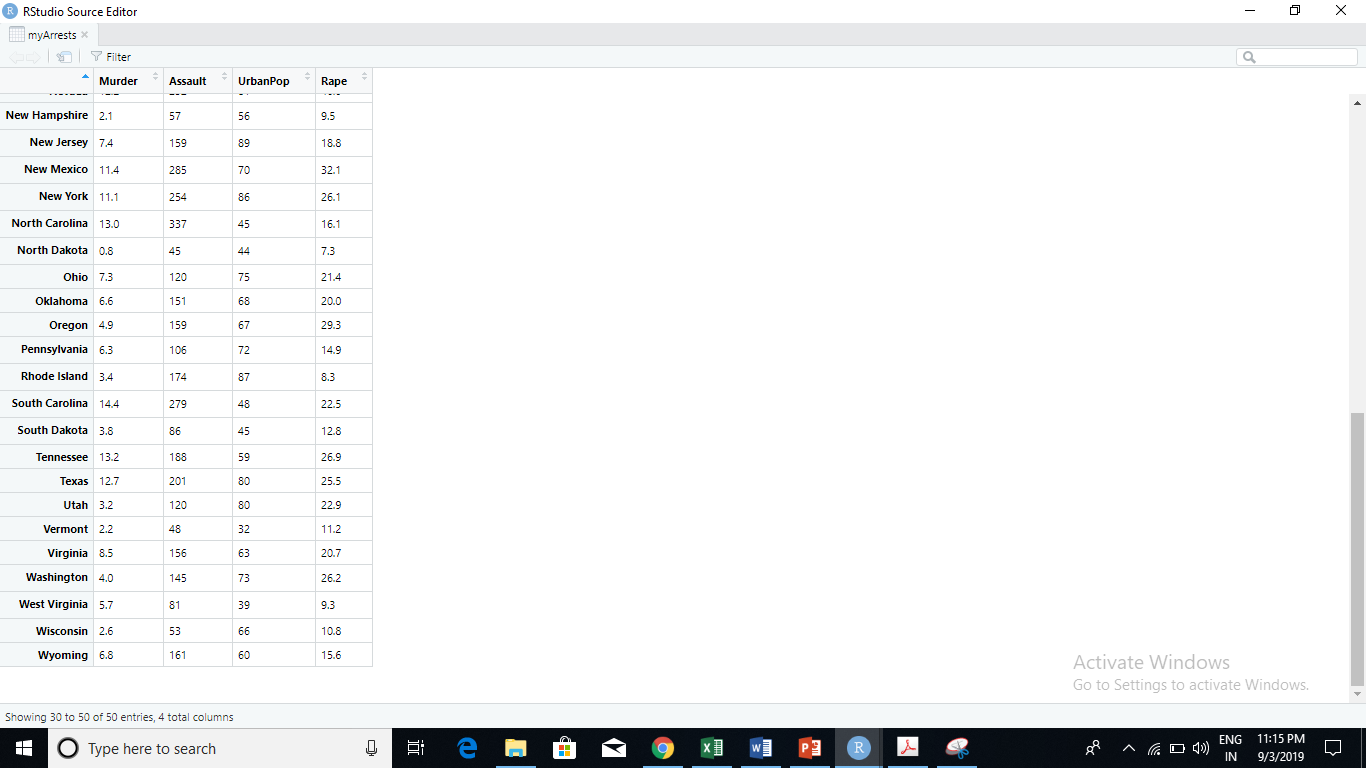


* 1. Describe the instances within the dataframe:

**States of USA (50 observations/cases/instances) with attributes such as murder, assault, urbanpop and rape incidents for each state.**



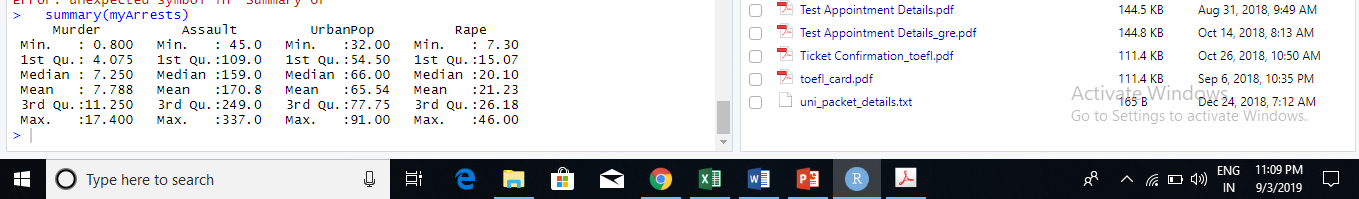




1. **Obtain a summary of the newly created dataframe using one R Studio command. Summary of the dataframe includes Min, 1st Quartile, Median, Mean, 3rd Quartile, and Max.** 
   1. Summarize your new dataset:

**summary(myArrests)**

* 1. Place a screen shot below of the output from summary(myArrests)



* 1. Why is the Summary command useful to a data scientist?

Summary command is used to provide more information (detailed) about the Numerical Variables like Min, Max, Mean, Median for descriptive statistics can be found out using this command at one go helpful to see how spread the data is, any extreme/unusual data and number of times variables are repeated.

1. **Now create your own dataframe with information about your family!**
   1. Create a column of names for your family dataframe:

**myFamilyNames <- c("Mom", "Dad", "Brother", "Sister")**

* 1. Create a vector for family member ages, entering the ages in the same order that would correspond to your family members, i.e. if you entered mom first in the previous step then your mom’s age should be the first entered within the vector.

**myFamilyAges <- c("48", "52" ,"9", "20")**

* 1. Add a eye color attribute to your dataframe by defining each family members eye color in a vector:

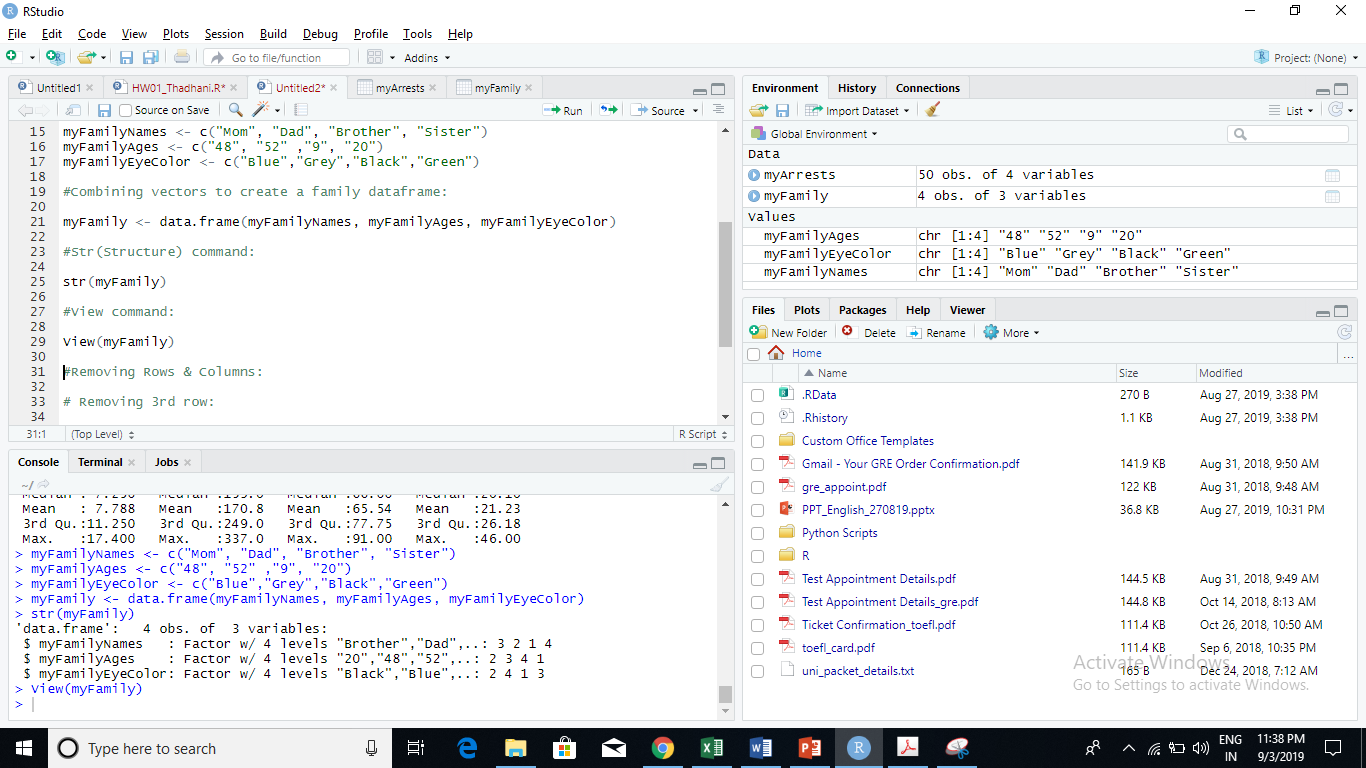
**(Hint: Refer to step B)**

**myFamilyEyeColor <- c("Blue","Grey","Black","Green")**

* 1. Combine the vectors into a dataframe:

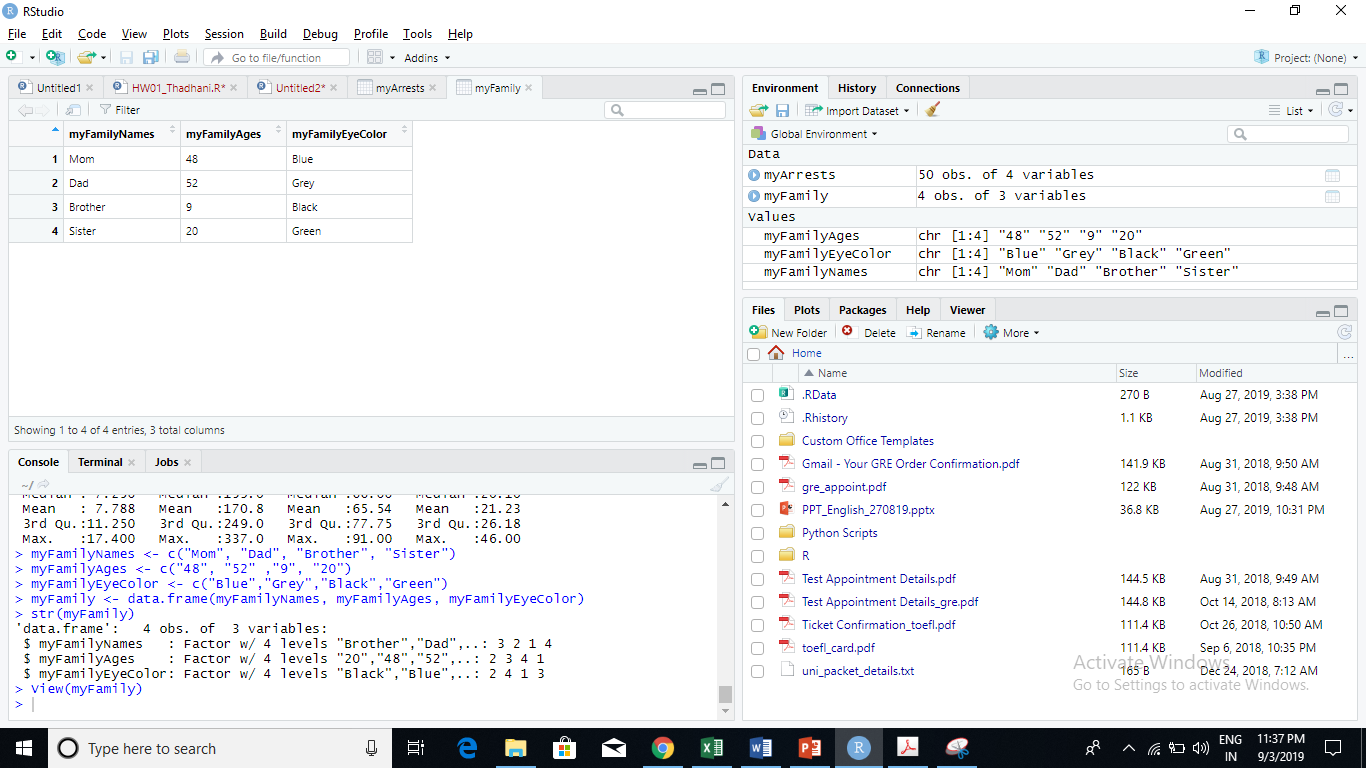
**myFamily <- data.frame(myFamilyNames, myFamilyAges, myFamilyEyeColor)**

* 1. Run the structure (**str()**) command on your newly created dataframe and record your observations of what the command did:

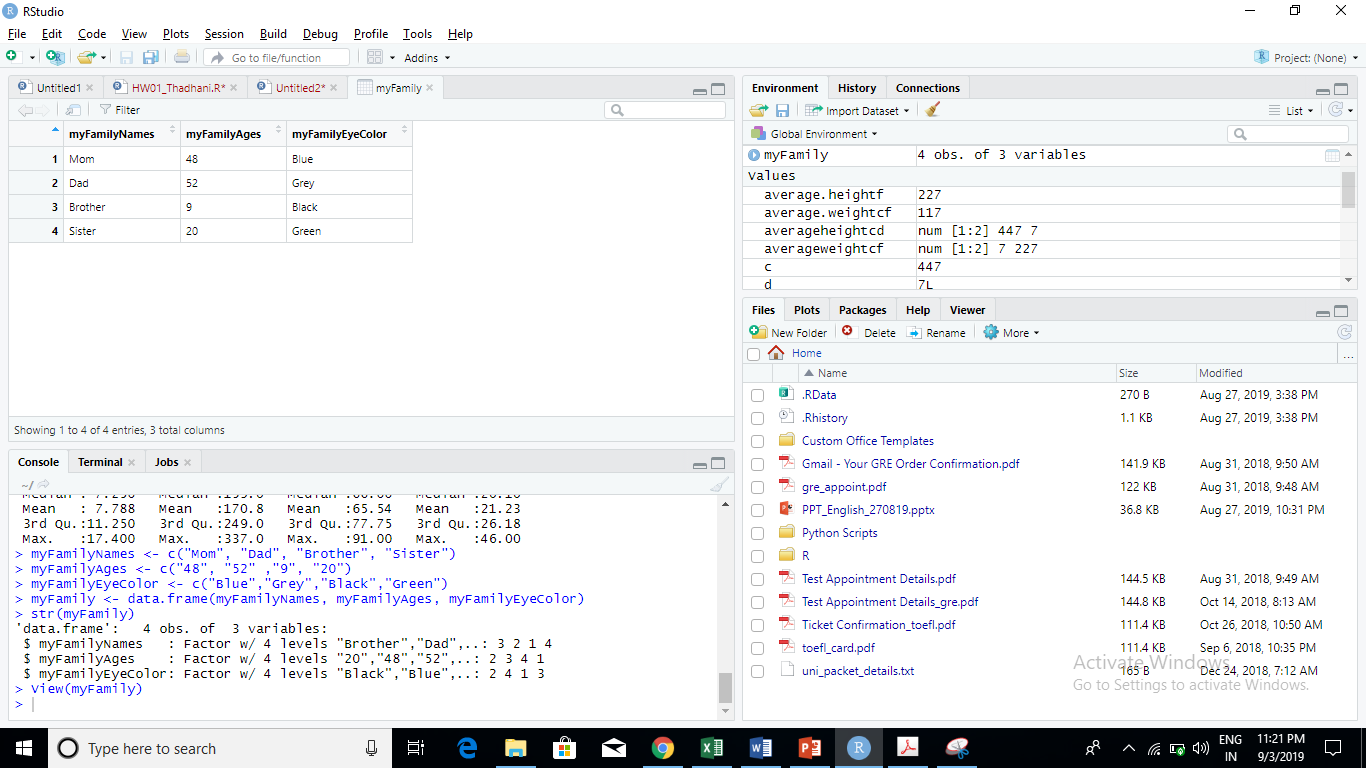


* 1. Now run the (**View()**)command. How does the output of the View command correlate with the structurecommand?

View() command correlates with structure command showing the same number of observations, variables and helps to understand the structure of the dataframe clearly.

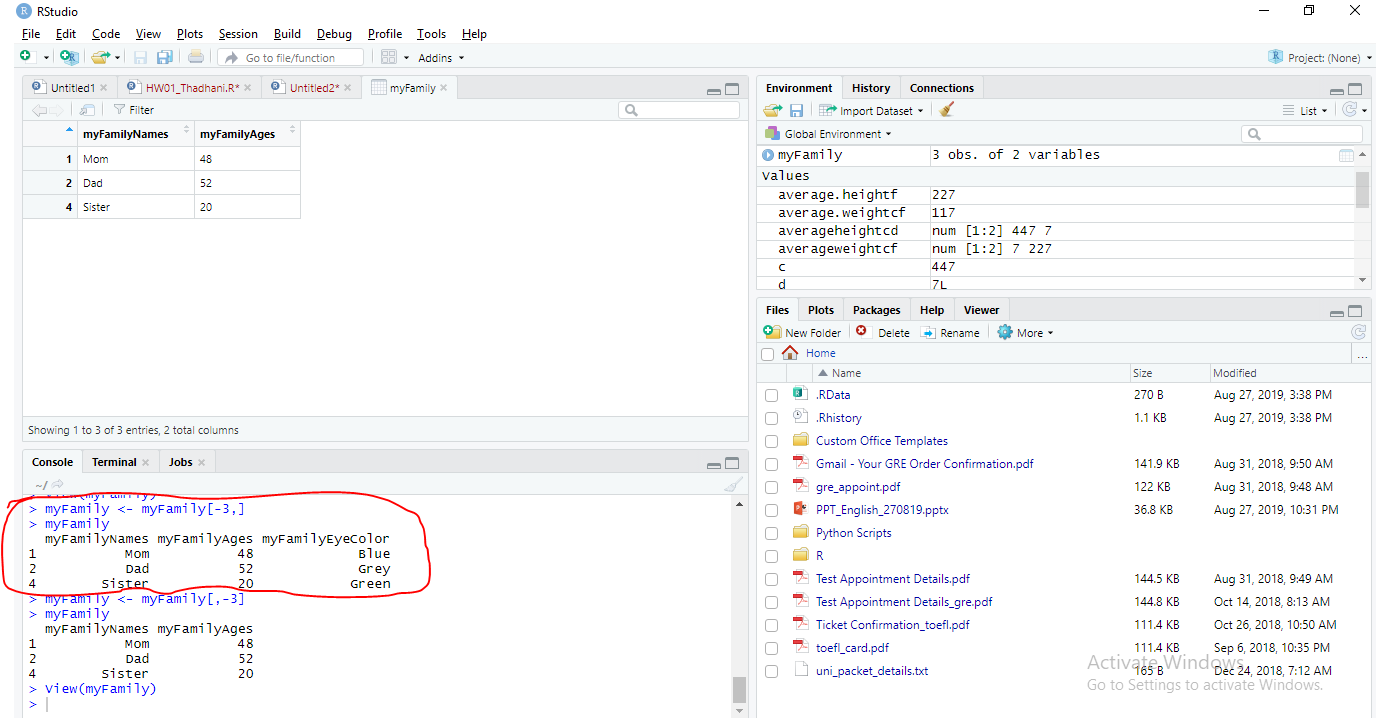


* 1. Place a screenshot below of the output from the View() command.



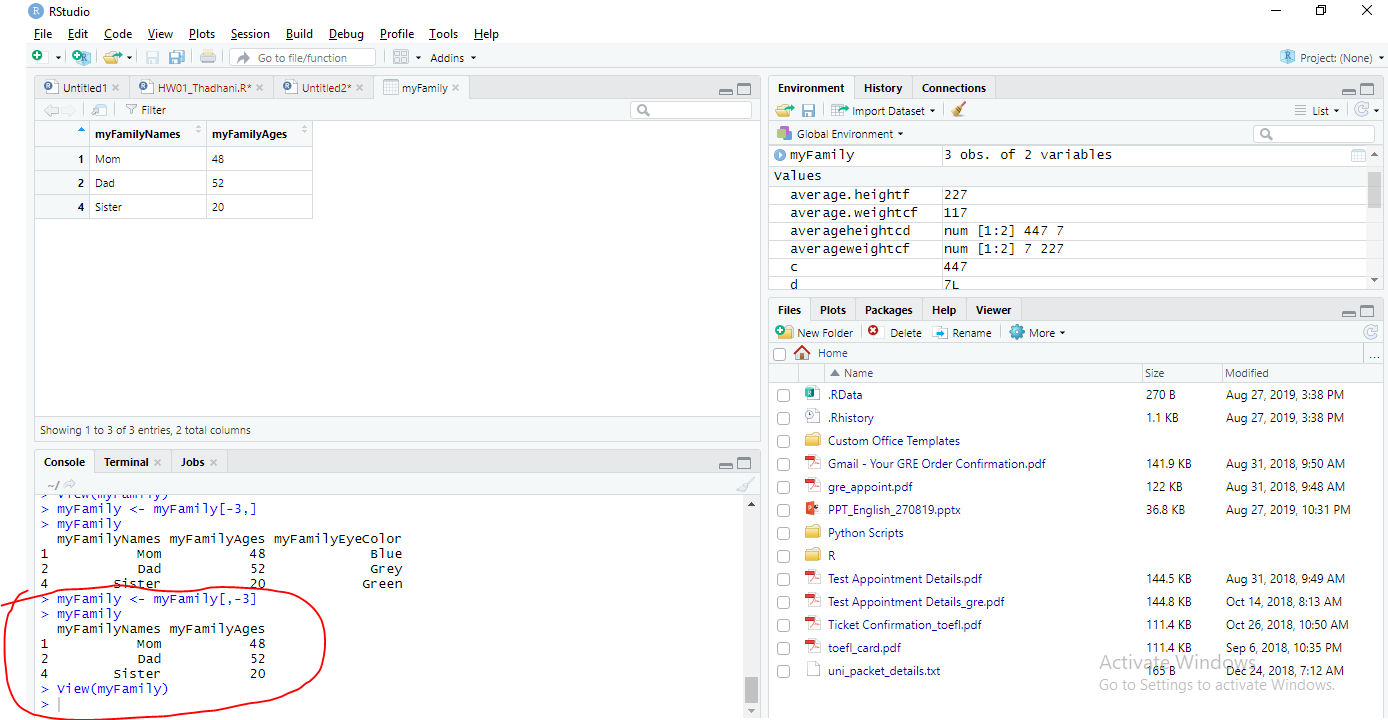
1. **Removing Rows and Columns within a dataframe.**
   1. Remove the third row in your newly created dataframe:

**myFamily <- myFamily[-3,]**



* 1. Remove the eye color column in your newly created dataframe:

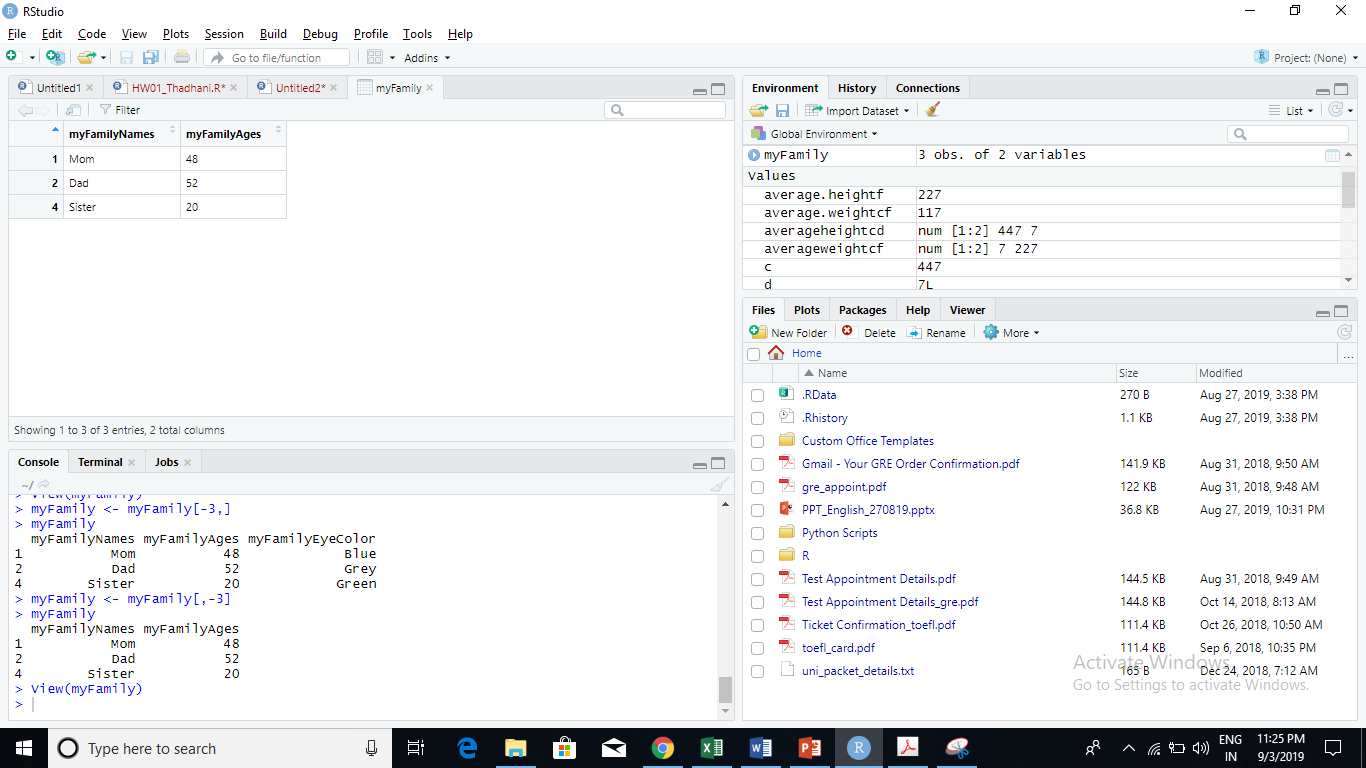
**myFamily <- myFamily[,-3]**



* 1. Assuming a data scientist wishes to analyze quantitative data within the dataframe, how is removing the eye color column an example of munging?

Eye color is a categorical variable. Munging is removing the unwanted data to clean the dataset before actually going for descriptive analyses which will be done on numeric data like age to get spread, average, make visualization so removing eye color would help to work better on quantitative data involving numbers rather than category and is an example of munging.

* 1. Place a screenshot below of your final dataframe. (Hint: use the View() command)



1. **List any additional resources you used here.**

None read the chapter 5 of Intro to Data Science by Saltz, Stanton. Everything is covered in the book.

1. **Be sure to save your work as this will become the starting code for your homework.**

***You must submit all Prep Exercises to blackboard prior to the deadline specified for each assignment.*** PE assignments are due on the evening prior to the lecture class. Late PE assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*