# STA404/504 HW01

Ethan Gutknecht

09/01/2022

**Learning Objectives:**

- Access and read data in R
- Identifying basic R object characteristics
- Basic data manipulation and computation

**Submission Requirement (Apply to all the HW)**

Please upload the following files separately to the canvas website. Please **do not** upload a zip file. Files do not follow the requirements will receive points deduction.

(1) An **Rmarkdown** file that can reproduce your knitted file. Please be sure to properly document and organize your code so that the data manipulation and plot creation processes are clear. Please also mark the questions clearly and have necessary answers and discussions.

(2) A knitted word/pdf/html file that containing the code, the outputs (the plots) and appropriate labels, discussion if applicable.

**Note**

The tutorial of using Rmarkdown can be found **here** (click here).

---

**Question 1.**

The first question is to practice using R built-in datasets. R built-in datasets are contained in R packages, and are generally used as demo data for playing with R functions. To see the list of pre-loaded data, run code **data()**. One of the datasets is "mtcars". To access this data, simply run the code **mtcars**, or run code **data(mtcars)**. Answer the following sub questions using this data.

**(1) Read the description of the "mtcars" using help() function, and understand the meaning of each variable. (This is just for your reference; you don't need to paste the R help file here). Then use the function we discussed in class to figure out what data types each variable in the "mtcars" dataset belongs to and report it.**

[, 1] mpg Miles/(US) gallon

[, 2] cyl Number of cylinders

[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower

[, 5] drat Rear axle ratio

[, 6] wt Weight (1000 lbs)

[, 7] qsec 1/4 mile time

[, 8] vs Engine (0 = V-shaped, 1 = straight)

[, 9] am Transmission (0 = automatic, 1 = manual)

[,10] gear Number of forward gears

[,11] carb Number of carburetors

**(2) How many vehicles have a V-shaped engine? How many vehicles have a straight engine?**

```
# Take a sum of all the cars that have V-Shaped engines
sum(with(cardata,vs == 0))
```

```
## [1] 18
```

```
sum(with(cardata,vs == 1))
```

```
## [1] 14
```

```
# "8]   vs  Engine (0 = V-shaped, 1 = straight)"
#
# Website referenced for help:
# https://www.delftstack.com/howto/r/count-observations-in-r/
```

There are 18 cars that have an engine that is V-Shaped and 14 cars that have a straight engine

**(3) Calculate and report the average miles/(US) gallon for vehicles with different number of cylinders.**

```r
# Take Average
mtcars %>%
  group_by(mtcars$cyl) %>%
  summarise(Average = mean(mpg))
```

```
## # A tibble: 3 x 2
##   `mtcars$cyl` Average
##          <dbl>   <dbl>
## 1            4    26.7
## 2            6    19.7
## 3            8    15.1
```

```r
# References:
# https://www.geeksforgeeks.org/group-by-function-in-r-using-dplyr/
```

**(4) Change the variable "am" to a factor variable which shows "Automatic" and "Manual" directly, and store it as a new variable in the dataset. Show the head of the edited dataset.**

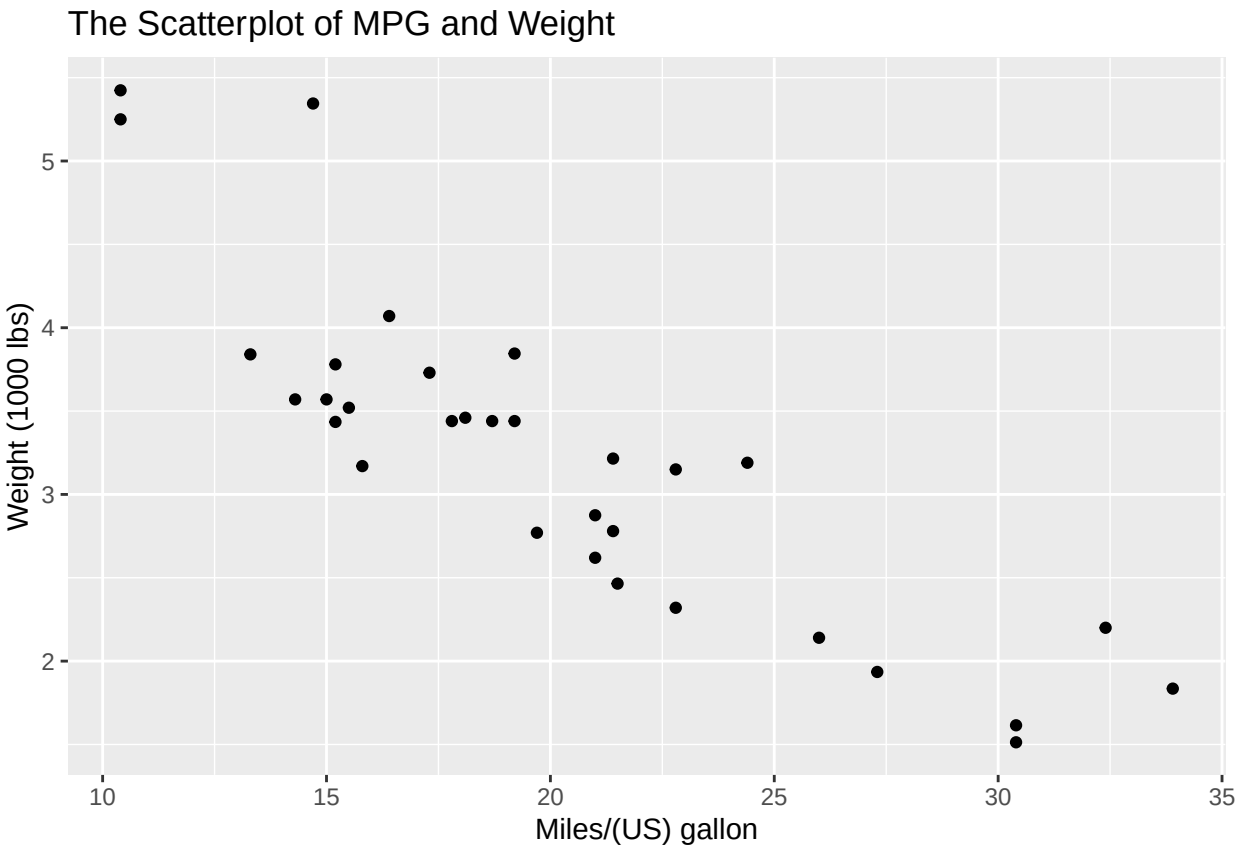
```r
# Create factor labels and replace
newFactorLabels <- factor(mtcars$am, labels = c("Automatic","Manual"))
mtcars$am <- newFactorLabels

# Show first lines of data
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs        am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0    Manual    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0    Manual    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1    Manual    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 Automatic    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 Automatic    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1 Automatic    3    1
```

**(5) Explore the data, make a plot showing the relationship between two continuous variables that you pick, describe the relationship reflected from the graph. (Please use ggplot() to generate this graph. Make sure the aesthetic settings of this graphs are appropriate.)**

```r
ggplot() +
  geom_point(aes(x=mpg, y=wt),data=mtcars) +
  labs(x="Miles/(US) gallon", y="Weight (1000 lbs)") +
  ggtitle("The Scatterplot of MPG and Weight")
```




As the MPG increases, the weight of the car decreases. This is because it takes less power to move a car with less weight then it would with a higher weight.

---

**Question 2.**

The dataset "student-por.csv" on canvas is about student's achievement in a Portugue language course in secondary education of two Portuguese schools (Cortez and Silva, 2008). The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. This is one of the available datasets. You may refer to this website for more information about the attributes and the data: https://archive.ics.uci.edu/ml/datasets/student+performance.

First, you may run the following code to read the data into R. After running it, R will show the column specifications.

#if you set the working directory to where you store the data
**por <- read.table("student-por.csv",sep=";",header=TRUE)**
or
#if you did not set the working directory to where you store the data
**por <- read.table("the file path on your computer/student-por.csv ", sep=";",header=TRUE)**

check whether "por" is a data frame, by **is.data.frame(por)**

Then, answer the following questions:

**(1) What percent of the students choose a specific school because the school is close to their home?**

```
por <- read.table("student-por.csv",sep=";",header=TRUE)
```

There were 149 students that choose the school because it was close to their home.

**(2) For all the students with a second period grade for Portuguese language course greater than 16, what percent of them are female?**

```
# (G2 > 16 && sex == F) / (G2 > 16)
filteredDataQ2 <- filter(por, G2 > 16)
filteredDataQ2Female <- filter(por, G2 > 16 & sex == "F")

count(filteredDataQ2Female) / count(filteredDataQ2)
```

```
##           n
## 1 0.7142857
```

```
# References
# https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter
```

71% of people that have a second period grade for Portuguese language course greater than 16 are female.

**(3) For the students who study two to five hours per week and has extra-curricular activities, what is the average and standard deviation of their first period grade for Portuguese language course?**

```
# filteredData <- (studytime == 2 && activities == 1)
filteredDataQ3 <- filter(por, studytime == 2 & activities == "yes")

mean(filteredDataQ3$G1)
```

```
## [1] 11.79739
```

```
sd(filteredDataQ3$G1)
```

```
## [1] 2.737074
```

The students who study 2-5 hours per week and have extra curricular activities have an average of 11.80 and a standard deviation of 2.73