Quiz 1

Your name here!!!!

Due: 4:30pm, Jul 19, 2021

Contents

1 Overview

This is a 30 minute, **individual** quiz. It is an open book exercise. Feel free to copy any R-chunks from the lectures or homework. Ask your TF for help if you get stuck somewhere.

Submit the following two files to the canvas:

- quiz1.rmd
- quiz1.html or .pdf (Only 1 compiled one)
- We will keep the submission open until 6:00pm

Objectives:

We will focus on the following aspects:

- Read data into R
- Make sense out of the data through EDA
- Be able to use R-Markdown
- Be able to submit your work to Canvas
- No need to beautify your R-markdown .html or .pdf
- Solutions/R functions might not be unique
- Data needed for this quiz: `data/car_04_regular_sub.csv`

2 Case study: are Asian cars more efficient?

Asian cars, especially Japanese and Korean cars have been loved by families due to their reputation for fuel efficiency and relatively cheaper prices. We will use a subset of the full data `car_04_regular.csv` called `data/car_04_regular_sub.csv` to perform some preliminary analyses.

2.1 Question 1: Read the data

Read the data first. To make sure you get the data into R we have included the following R-chuck. Notice this chunk only works if you have stored the data `car_04_regular_sub.csv` in a sub-folder called `data`.

```
car_data <- read.csv("data/car_04_regular_sub.csv", header=TRUE)
```

   i. How many variables are in this data? And how many car models are there?

```
num_variables <- length(names(car_data))
num_variables
```

```
## [1] 13
```

```
num_models <- length(unique(car_data$Make.Model))
num_models
```

```
## [1] 225
```

   ii. The variable `Continent` is a categorical variable for labeling the region a car is made in. How many continents are reported on in this data set?

```
num_continents <- length(unique(car_data$Continent))
num_continents
```

```
## [1] 3
```

    iii. Do you spot anything unusual in this data set?

```
summary(car_data)
```

```
##    Make.Model          Continent            MPG_City        MPG_Hwy
##  Length:225          Length:225          Min.   :11.0   Min.   :14.0
##  Class :character    Class :character    1st Qu.:17.0   1st Qu.:23.0
##  Mode  :character    Mode  :character    Median :19.0   Median :26.0
##                                          Mean   :19.2   Mean   :25.8
##                                          3rd Qu.:22.0   3rd Qu.:29.0
##                                          Max.   :32.0   Max.   :38.0
##    Horsepower         Weight            Length          Width          Seating
##  Min.   :103    Min.   :2.19    Min.   :143    Min.   :65.7    Min.   :2.00
##  1st Qu.:160    1st Qu.:3.16    1st Qu.:178    1st Qu.:69.7    1st Qu.:5.00
##  Median :201    Median :3.57    Median :187    Median :71.7    Median :5.00
##  Mean   :212    Mean   :3.69    Mean   :186    Mean   :72.1    Mean   :5.12
##  3rd Qu.:252    3rd Qu.:4.06    3rd Qu.:194    3rd Qu.:74.4    3rd Qu.:5.00
##  Max.   :390    Max.   :5.82    Max.   :224    Max.   :80.2    Max.   :8.00
##    Cylinders     Displacement        Make             Transmission
##  Min.   :2.0    Min.   :1.30    Length:225          Length:225
##  1st Qu.:4.0    1st Qu.:2.40    Class :character    Class :character
##  Median :6.0    Median :3.20    Mode  :character    Mode  :character
##  Mean   :5.7    Mean   :3.22
##  3rd Qu.:6.0    3rd Qu.:4.00
##  Max.   :8.0    Max.   :6.00
```

```
view(car_data)
```

After viewing the data table itself and its summary, I was unable to find anything out of the ordinary

2.2 Question 2: EDA

Fuel efficiency can be measured with `MPG_Hwy` (mileage per gallon on highways). This is the variable we use to measure the fuel efficiency for a car.

    i. What is the sample mean and sample standard deviation of `MPG_Hwy` for all cars?

```
Hwy_eff <- car_data$MPG_Hwy
sampleSD <- sd(Hwy_eff)
sampleMean <- mean(Hwy_eff)
sampleSD
```
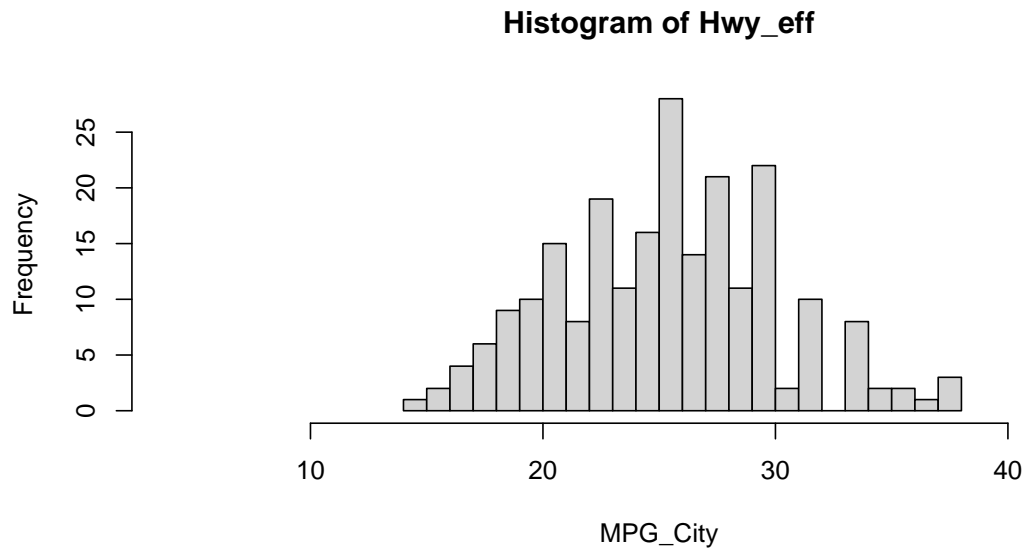
```
## [1] 4.71
```

```
sampleMean
```

```
## [1] 25.8
```

    ii. Show the histogram of `MPG_Hwy` for all cars. Does `MPG_Hwy` appear to be a normal variable? Why or why not?

```
hist(Hwy_eff, breaks = 18, xlim = range(4:46), xlab = "MPG_City", plot = T)
```

Histogram of Hwy_eff

The histogram of MPG_Hwy suggests that it follows an approximately normal pattern

2.3 Question 3: Comparison

    i. What is the sample mean of `MPG_Hwy` for American cars?

```
american <- car_data %>% filter(Continent == "Am")
asian <- car_data %>% filter(Continent == "As")
American_mean_eff <- mean(american$MPG_Hwy)
American_mean_eff
```

```
## [1] 25.5
```

    ii. What is the sample mean of `MPG_Hwy` for Asian cars?

```
Asian_mean_eff <- mean(asian$MPG_Hwy)
Asian_mean_eff
```

```
## [1] 26.4
```

    iii. By comparing the two sample means above, do we have some evidence that Asian cars are more efficient than American cars? (A larger value for `MPG_Hwy` indicates a more efficient car.)

Part 3iii: Yes, we have *some* evidence. On highways, asian cars have a mpg of 26.4, compared to the slightly lesser value of 25.5 mpg for american cars. However, this does not paint the entire picture. Hypothesis testing should be done to see the probabilities of us getting this result even if their true means were the same, and other variables should be looked at, such as MPG_City.

    iv. Is it fair to only compare the sample means from the two continents? What are some other statistics that would be useful for this analysis?

As alluded to before, we should conduct hypothesis testing and include other variables. To make better conclusions, even looking at something like standard deviation by itself would help, to perhaps help form a conclusion that, while asian cars are on average more efficient, they vary more... etc, etc. (Not the actual conclusion, just making a point). We should also include MPG_City. If we wished to go even further, we should also be comparing cars with a similar purpose. An American truck compared to a small japanese sedan is like comparing apples to oranges. We could categorize cars by horsepower intervals, or even cylinders to help with that. I'm sure there's more that could be done, but those were some of my ideas.

End of the quiz

3 Solution to come