HW 1

Enter your name and EID here: Austine Do (ahd589)

You will submit this homework assignment as a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Part 1

The dataset `mtcars` was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and other aspects of automobile design and performance for different cars (1973-74 models). Look up the documentation for this data frame with a description of the variables by typing **`?mtcars` in the console** pane.

Question 1: (1 pt)

Take a look at the first 6 rows of the dataset by using an `R` function in the code chunk below. Have you heard about any (or all) of these cars?

```
# Information on the 'mtcars' built-in data set
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  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

None of these cars look familiar to me, although I do recognize the 'Mazda' brand name.

Question 2: (2 pts)

How many rows and columns are there in this data frame in total?

```
# Dimensions of the data frame for 'mtcars'
dim(mtcars)
```

```
## [1] 32 11
```

There are 32 rows and 11 columns in this data frame.

Question 3: (1 pt)

It is always a good practice to make a local copy of the dataset in your environment. Save `mtcars` in your environment and name it as your `eid`. From now on, use this new object instead of the built-in dataset.

```
# saved 'mtcars' data frame under a variable name 'ahd589'
ahd589 <- mtcars
```

Question 4: (2 pts)

When is your birthday? Using indexing, grab the row of `mpg` that corresponds to the day of your birthday (the latter should be a number between 1 and 31).

```
# From the column 'mpg' we grabbed row 30's information on mpg
ahd589$mpg[30]
```

```
## [1] 19.7
```

My birthday is on January 30th, 2002. The row that corresponds with the day of my birthday from the column 'mpg' contains the information on the Ferrari Dino's mpg which is 19.7 mpg.

Question 5: (2 pts)

Using logical indexing, count the number of rows in the dataset where the variable `mpg` takes on values greater than 30.

```
# counts the number of observations in the column 'mpg' where the value of mpg is greater than 30
sum(ahd589$mpg > 30)
```

```
## [1] 4
```

There are 4 rows in which the car's mpg is greater than 30.

Question 6: (2 pts)

Let's create a new variable called `kpl` which converts the fuel efficiency `mpg` in kilometers per liter. Knowing that 1 mpg corresponds to 0.425 kpl, what is the maximum value of `kpl` in the dataset?

```r
# adds another column variable to the dataset called 'kpl'
ahd589$kpl <- ahd589$mpg * 0.425
max(ahd589$kpl)
```

```
## [1] 14.4075
```

The max value found in the 'kpl' column is 14.4075.

Part 2

Let's quickly explore another built-in dataset: `airquality` which contains information about daily air quality measurements in New York, May to September 1973.

Question 7: (2 pts)

Calculate the mean `Ozone` (in ppb) using the `mean()` function. Why does it make sense to get this answer? *Hint: take a look at the column `Ozone` in the dataset.*

```r
# looks at the basic structure of the dataset and calculates the mean of the ozone column in the datase
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```r
mean(airquality$Ozone)
```

```
## [1] NA
```

The mean() function for the ozone column returns NA and this makes sense because there are many rows that don't contain a value for ozone.

Question 8: (1 pt)

Look at the documentation for the function `mean()` by running **`?mean` in the console**. What argument should be used to find the mean value that we were not able to get in the previous question? What type of values does that argument take?

We should use the 'na.rm' argument and it should be set to false to dismiss all NA values found in the dataset while calculating the mean for that column. The argument 'na.rm' takes the values T or F.

Question 9: (2 pts)

Sometimes the R documentation does not feel complete. We wish we had more information or more examples. Find a post online (include the link) that can help you use that argument in the `mean()` function. Then finally find the mean ozone!

```
# Calculate the mean of ozone while ignoring the NA values in the data set
mean(airquality$Ozone, na.rm = T)
```

```
## [1] 42.12931
```

This link leads to a website that provides documentation and examples for the mean() function: https://sparkbyexamples.com/r-programming/calculate-mean-or-average-in-r/

The mean of the Ozone column in the data set is 42.12931

Part 3

The Internet clothing retailer Stitch Fix wants to develop a new model for selling clothes to people online. Their basic approach is to send people a box of 5–6 items of clothing and allow them to try the clothes on. Customers keep (and pay for) what they like while mailing back the remaining clothes. Stitch Fix then sends customers a new box of clothes a month later.

A critical question for Stitch Fix to consider is "Which clothes should we send to each customer?" Since customers do not request specific clothes, Stitch Fix has to come up with 5–6 items on its own that it thinks the customers will like (and therefore buy). In order to learn something about each customer, they administer an **intake survey** when a customer first signs up for the service. The survey has about 20 questions and the data is then used to predict what kinds of clothes customers will like. In order to use the data from the intake survey, a statistical algorithm must be built in order to process the customer data and make clothing selections.

Suppose you are in charge of building the intake survey and the algorithm for choosing clothes based on the intake survey data.

Question 10: (2 pts)

What kinds of questions do you think might be useful to ask of a customer in the intake survey in order to better choose clothes for them? Come up with 4 questions to ask customers, with 2 questions leading to numeric data and 2 questions leading to categorical data. *Make sure to indicate which question is which type.*

1) What color do you typically look for in clothes? (earth tones, pastel, neutral, etc) [categorical data]

2) What style of clothes are looking to add to your wardrobe? (seasonal, casual, business/formal, etc.) [categorical data]

3) What is your height in feet and inches? [numerical data]

4) What is a comfortable price range per clothing item for your budget? [numerical data]

Question 11: (2 pts)

In addition to the technical challenges of collecting the data and building this algorithm, you must also consider the impact the algorithm may have on the people involved. What potential negative impact might the algorithm have on the customers who are submitting their data? Consider both the data being submitted as well as the way in which the algorithm will be used when answering this question.

Some potential negative impact and concerns that this algorithm can introduce is obviously some data security concerns regarding the customer, data accuracy based on the customer's input (which might be incorrect and therefore leading the algorithm to incorrectly recommend items), and changing personal style of the customer that the algorithm may not consider since this survey is only submitted before the use of the service. These are some of the many issue that can arise while using consumer data and an algorithm to predict customer preference.

Formatting: (1 pt)

Knit your file! You can knit into html and once it knits in html, click on `Open in Browser` at the top left of the window that pops out. **Print** your html file into pdf from your browser.

Is it working? If not, try to decipher the error message (look up the error message, consult websites such as stackoverflow or crossvalidated.

Finally, remember to select pages for each question when submitting your pdf to Gradescope.