````
---
title: "DS311 - R Lab Assignment"
author: "Fatima Hararah"
date: "`r Sys.Date()`"
output:
  pdf_document: default
  html_document:
    theme: united
    highlight: tango
    df_print: paged
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Assignment 1

* In this assignment, we are going to apply some of the build in data set in R for descriptive statistics analysis.
* To earn full grade in this assignment, students need to complete the coding tasks for each question to get the result.
* After finished all the questions, knit the document into HTML format for submission.

### Question 1

Using the **mtcars** data set in R, please answer the following questions.

```{r}
# Loading the data
data(mtcars)

# Head of the data set
head(mtcars)
```

a. Report the number of variables and observations in the data set.

```{r}
# Enter your code here!


nrow(mtcars)


# Answer:
print("There are total of __11___ variables and __32___ observations in this data set.")
```



b. Print the summary statistics of the data set and report how many discrete and continuous variables are in the data set.

```{r}
# Enter your code here!
summary(mtcars)

# Answer:
print("There are __5___ discrete variables and ___6__ continuous variables in this data set.")
```
````

c. Calculate the mean, variance, and standard deviation for the variable **mpg** and assign them into variable names m, v, and s. Report the results in the print statement.

```{r}
# Enter your code here!
m <- mean(mtcars$mpg)
v <- var(mtcars$mpg)
s <- sd(mtcars$mpg)

cat("Mean of mpg variable:", m, "\n")
cat("Variance of mpg variable:", v, "\n")
cat("Standard deviation of mpg variable:", s, "\n")

# print(paste("The average of Mile Per Gallon from this data set is ", 20.09062  , " with variance ", 36.3241 , " and standard deviation", 6.026948 , "."))
```

d. Create two tables to summarize 1) average mpg for each cylinder class and 2) the standard deviation of mpg for each gear class.

```
install.packages("dplyr")
library(dplyr)

mpg_cyl_table <- mtcars %>%
  group_by(cyl) %>%
  summarize(avg_mpg = mean(mpg))

mpg_cyl_table

mpg_gear_table <- mtcars %>%
  group_by(gear) %>%
  summarize(sd_mpg = sd(mpg))

mpg_gear_table
```

```

e. Create a crosstab that shows the number of observations belong to each cylinder and gear class combinations. The table should show how many observations given the car has 4 cylinders with 3 gears, 4 cylinders with 4 gears, etc. Report which combination is recorded in this data set and how many observations for this type of car.

```{r}
# Enter your code here!

cyl_gear_table <- table(mtcars$cyl, mtcars$gear)

cyl_gear_table


most_common <- which.max(cyl_gear_table)
cyl <- rownames(cyl_gear_table)[floor(most_common/nrow(cyl_gear_table))+1]
gear <- colnames(cyl_gear_table)[most_common %% ncol(cyl_gear_table)]
count <- as.vector(cyl_gear_table[most_common])
```
```

```
cat("The most common car type in this data set is car with", cyl, "cylinders and", gear,
"gears. There are a total of", count, "cars belonging to this specification in the data
set.")


print("The most common car type in this data set is car with _6___ cylinders and __4__
gears. There are total of _12____ cars belong to this specification in the data set.")
```

***

Question 2

Use different visualization tools to summarize the data sets in this question.

a. Using the **PlantGrowth** data set, visualize and compare the weight of the plant in the three separated group. Give labels to the title, x-axis, and y-axis on the graph. Write a paragraph to summarize your findings.

```
data("PlantGrowth")

# Head of the data set
head(PlantGrowth)

# Enter your code here!
install.packages("ggplot2")
library(ggplot2)

ggplot(PlantGrowth, aes(x = group, y = weight)) +
  geom_boxplot() +
  labs(title = "Weight of Plants in Three Groups",
       x = "Group",
       y = "Weight")

```

Result:

 => Report a paragraph to summarize your findings from the plot!

Based on the plot above, trt1 has the lowest weight, ctrl has the second highest weight, and trt2 has the heightest weight overall.

b. Using the **mtcars** data set, plot the histogram for the column **mpg** with 10 breaks. Give labels to the title, x-axis, and y-axis on the graph. Report the most observed mpg class from the data set.

```
ggplot(mtcars, aes(x = mpg)) +
  geom_histogram(binwidth = 3, color = "black", fill = "blue") +
  labs(title = "Distribution of MPG in mtcars",
       x = "Miles per gallon",
       y = "Frequency")
print("Most of the cars in this data set are in the class of ___15___ mile per gallon.")
```

c. Using the **USArrests** data set, create a pairs plot to display the correlations between the variables in the data set. Plot the scatter plot with **Murder** and **Assault**. Give labels to the title, x-axis, and y-axis on the graph. Write a paragraph to summarize your results from both plots.

```
data("USArrests")

# Head of the data set
head(USArrests)

# Enter your code here!
pairs(USArrests,
      main = "Pairs Plot of USArrests Data",
      labels = c("Murder", "Assault", "UrbanPop", "Rape"))


ggplot(USArrests, aes(x = Murder, y = Assault)) +
  geom_point() +
  labs(title = "Scatter Plot of Murder and Assault",
       x = "Murder",
       y = "Assault")
```

Result:

=> Report a paragraph to summarize your findings from the plot!

Based off the plot above, we can see how the states with higher murder rates also have higher assualt rates and vice versa.

***

Question 3

Download the housing data set from www.jaredlander.com and find out what explains the housing prices in New York City.

Note: Check your working directory to make sure that you can download the data into the data folder.

```{r, echo=FALSE}
# Load and clean the housing data set
download.file(url='https://www.jaredlander.com/data/housing.csv',
              destfile='data/housing.csv', mode='wb')
housingData <- read.csv('data/housing.csv')
housingData <- subset(housingData,
                      select = c("Neighborhood", "Market.Value.per.SqFt", "Boro",
"Year.Built"))
housingData <- na.omit(housingData)
```

a. Create your own descriptive statistics and aggregation tables to summarize the data set and find any meaningful results between different variables in the data set.

```
head(housingData)

# Enter your code here!
library(dplyr)
install.packages("tidyr")
library(tidyr)

housingData %>%
  select_if(is.numeric) %>%
  summarize_all(list(mean = mean, median = median, sd = sd, min = min, max = max))
```

```
housingData %>%
  group_by(type_of_dwelling) %>%
  summarize(avg_sale_price = mean(sale_price))

housingData %>%
  group_by(type_of_dwelling, num_bedrooms) %>%
  summarize(avg_sale_price = mean(sale_price))
```

b. Create multiple plots to demonstrates the correlations between different variables. Remember to label all axes and give title to each graph.

```{r}
# Enter your code here!
library(ggplot2)

ggplot(housingData, aes(x = Year.Built, y = Market.Value.per.SqFt, color = Boro)) +
  geom_point() +
  labs(title = "Market Value per SqFt vs. Year Built, Colored by Borough",
       x = "Year Built",
       y = "Market Value per SqFt")

ggplot(housingData, aes(x = Boro, y = Market.Value.per.SqFt)) +
  geom_boxplot() +
  labs(title = "Market Value per SqFt by Borough",
       x = "Borough",
       y = "Market Value per SqFt")




```


```
ggplot(housingData, aes(x = Market.Value.per.SqFt, fill = Boro)) +
  geom_density(alpha = 0.4) +
  labs(title = "Density of Market Value per SqFt, by Borough",
       x = "Market Value per SqFt",
       y = "Density")
```

c. Write a summary about your findings from this exercise.

=> Enter your answer here!

Based on the findings of the data above, Staten Island has the highest density of markert value with lower market value sqft, and mandattan has the highest overall value of market value per SqFt