STA 5207: Homework 1

13th Sept 2023

```
                                                  STA 5207: Homework 1
```

Include your R code as an appendix at the end of your homework. Do not include your code in your answers unless the question explicitly tells you to include your code. Your answers to each exercise should be self-contained without code so that the grader can determine your solution without reading your code or deciphering its output.

Exercise 1 (Exploratory Data Analysis (EDA) for `Diabetes` Data) [70 points]

For this exercise, we will use the `diabetes` dataset from the `faraway` package, which is the package for the recommend textbook for this course *Linear Models with R* by Julian J. Faraway.

1. (5 points) Install and load the `faraway` package. **Do not** include the installation command in your .Rmd file. (If you do it will install the package every time you knit your file.). **Do** include the command to load the package into your environment in an `R` chunk that is not executed as you answer to this question below.

   Ans :- library(faraway) faraway::diabetes

How many observations are in the `diabetes` dataset? How many variables? Who are the individuals in this dataset?

Ans :- Observations in the Diabetes Dataset is 403 and variables are 19.Induviduals in this dataset are the diabetes patients

(9 points) How many individuals have HDL levels (High Density Lipoprotein) that are missing, that is, have a value of `NA`? What are the row numbers of the missing individuals? What is the mean HDL level (High Density Lipoprotein) of individuals in this sample?

Hint: The mean should be calculated with `NA` values removed. Use `?mean` to determine an argument that removes missing values before calculating the mean.

Ans :-

- Only 1 individuals HDL levels is missing
- 50.44527 is the mean HDL of individuals in this sample.

(5 points) What is the standard deviation of total cholesterol of individuals in this sample after removing missing values?

Hint: The standard deviation should be calculated with NA values removed. Use ?sd to determine an argument that removes missing values before calculating the standard deviation.

Ans :- 44.44556 is the standard deviation for the above dataset

5. (6 points) What is the range of ages of individuals in this sample?

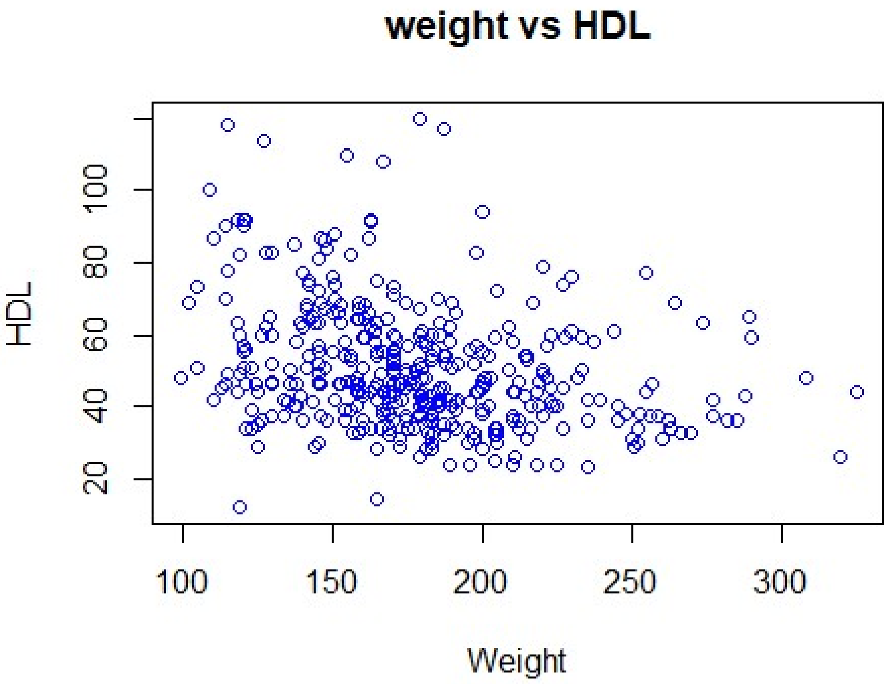
   Ans :- The age ranges from 19 to 92 in this particular sample

6. (6 points) What is the mean HDL of males in this sample?

   Ans :- 48.125 is the mean HDL of males in this sample

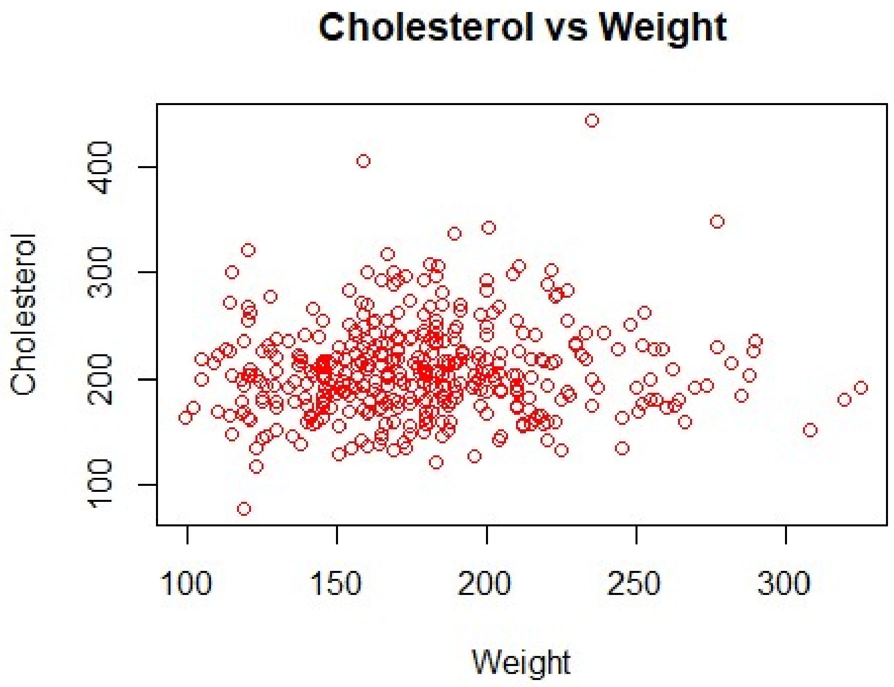
7. (10 points) Create a scatter plot of HDL (y-axis) vs weight (x-axis). Use a non-default color for the points. Also, be sure to give the plot a title and label the axes appropriately. Based on the scatter plot, does there seem to be a relationship between the two variables? Briefly explain.

   Ans : - Based on the ScatterPlot we can say that the plot is going downwards and it has a negative trend so we can say that the correlation will be negative too.



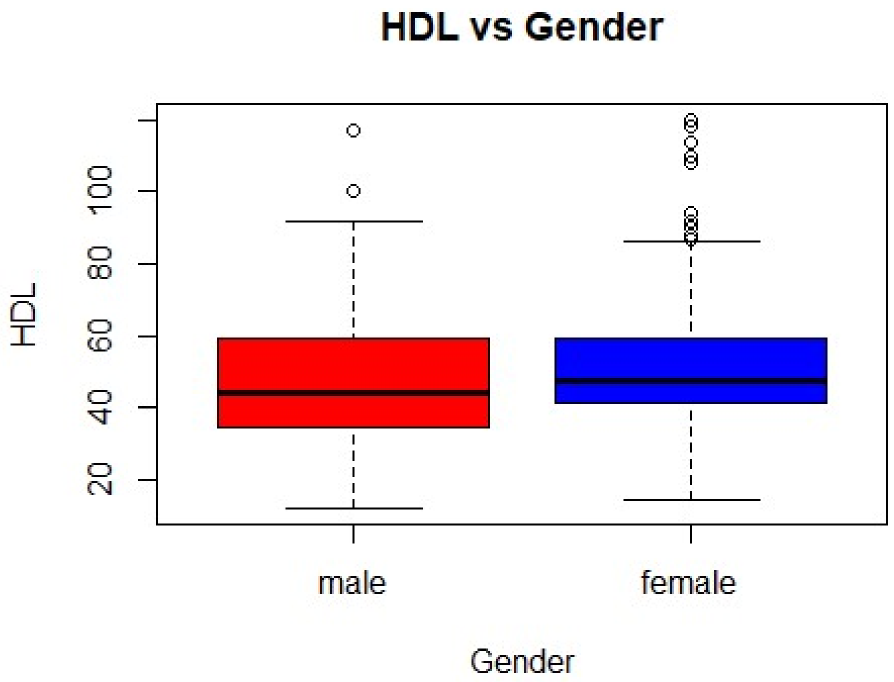
8. (10 points) Create a scatter plot of total cholesterol (y-axis) vs weight (x-axis). Use a non-default color for the points. Also, be sure to give the plot a title and label the axes appropriately. Based on the scatter plot, does there seem to be a relationship between the two variables? Briefly explain.

Ans : - Based on the generated scatter plot we can say that there is no relationship between Cholesterol and Weight and there is no correlation.



9. (10 points) Create side-by-side boxplots for HDL by gender. Use non-default colors for the plot. Also, be sure to give the plot a title and label the axes appropriately. Based on the boxplot, does there seem to be a difference between HDL level and gender? Briefly explain.

Ans : - Based on the generated Boxplots we can say that there is a slight difference in Median, outliers, and IQR1 and IQR3.



Exercise 2 (EDA for the `Salary` Dataset) [30 points]

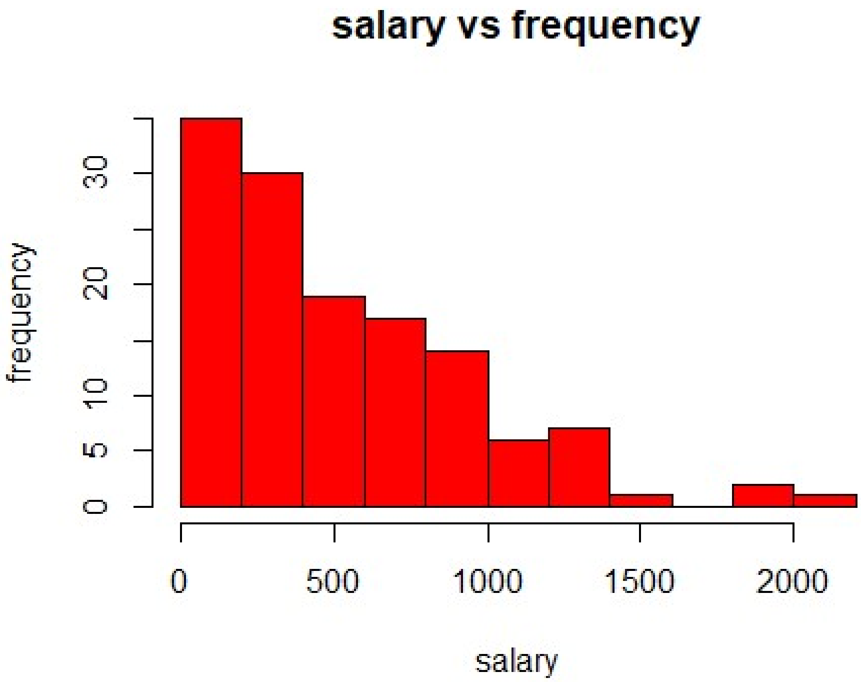
This exercise will use the `hitters` data set, which you can find in `hitters.csv` on Canvas. This is a version of the `Hitters` data set from the `ISLR2` package with missing observations removed. The data set contains 132 observations of Major League Baseball (MLB) players from the 1986 and 1987 seasons. The data set contains the following 20 variables

- `AtBat`: Number of times at bat in 1986
- `Hits`: Number of hits in 1986
- `HmRuns`: Number of home runs in 1986
- `Runs`: Number of runs in 1986
- `RBI`: Number of runs batted in in 1986
- `Walks`: Number of walks in 1986
- `Years`: Number of years in the major leagues
- `CAtBat`: Number of times at bat during his career
- `CHits`: Number of hits during his career
- `CHmRun`: Number of home runs during his career
- `CRuns`: Number of runs during his career
- `CRBI`: Number of runs batted in during his career
- `CWalks`: Number of walks during his career
- `League`: The player's league. (League = 1 for American, League = 0 for National)
- `Division`: The player's division. (Division = 1 for West, Division = 0 for East)
- `PutOuts`: Number of put outs in 1986

- `Assists`: Number of assists in 1986
- `Errors`: Number of errors in 1986
- `Salary`: 1987 annual salary on opening day in thousands of dollars
- `NewLeague`: The player's league at the beginning of 1987 (League = 1 for National, League = 0 for American)

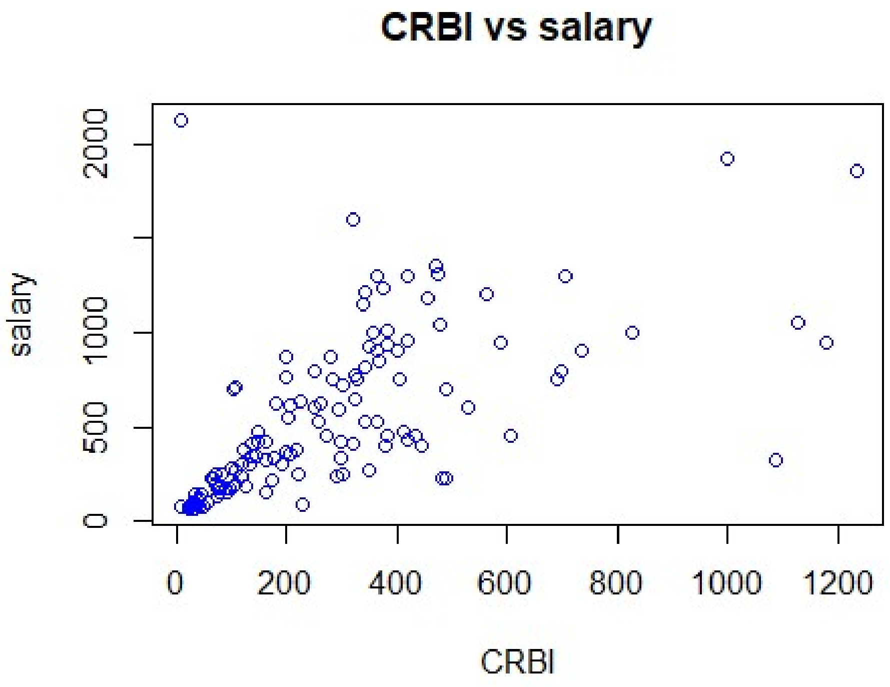
1. (10 points) Create a histogram of `Salary`. Do not modify R's default bin selection. Be sure to give the plot a title and label the axes appropriately. Describe the shape of the histogram and how it relates to the distribution of player's salary in the MLB.

   Ans : - From the Histogram we can say that the plot has Right Skewed Hstogram. Histogram show that most of the people were lies on 0-1000 as graph has more peak frequency on this salary range.



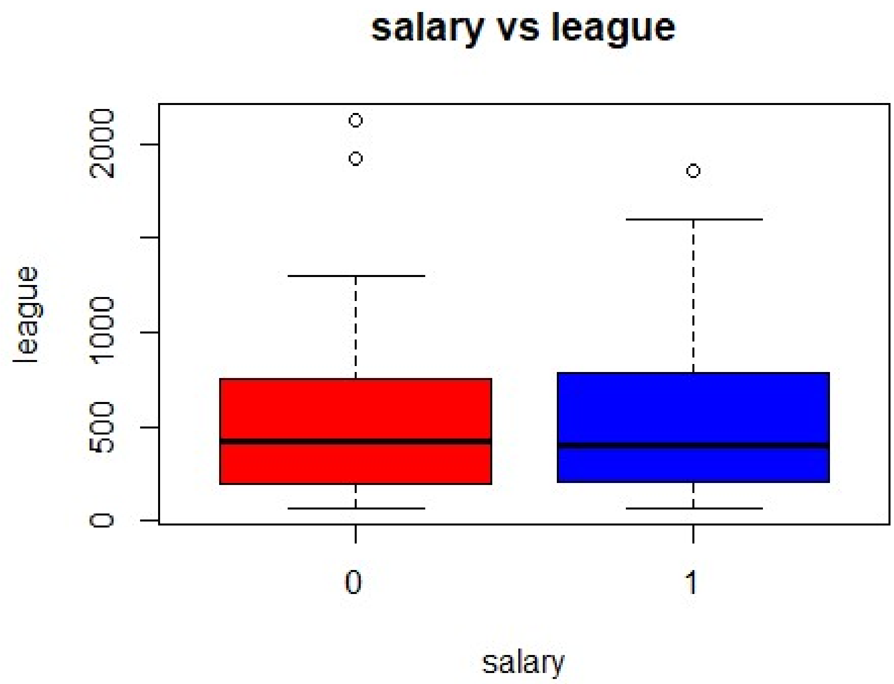
2. (10 points) Create a scatter plot of salary (y-axis) vs CRBI (x-axis). Use a non-default color for the points. Also, be sure to give the plot a title and label the axes appropriately. Do you notice any trends?

   Ans : - From the scatterplot we can say that there is a relation between Salary and CRBI . We can say that the plot has positive correlation.



3. (10 points) Create side-by-side boxplots for Salary by league. Use non-default colors for the plot. Also, be sure to give the plot a title and label the axes appropriately. Based on the boxplot, does there seem to be a difference between Salary by league? Briefly explain.

   Ans : - We can see from the below generated Boxplot that the Median, First IQR, and Third IQR has no difference.



Code Appendix

```
## Exercise 1
## Code for Problem 1, Question 1
library(faraway)
faraway::diabetes
nrow(diabetes)
ncol(diabetes)

sum(is.na(diabetes$hdl))
which((is.na(diabetes$hdl)))
mean(diabetes$hdl,na.rm = TRUE)

sd(diabetes$chol,na.rm = TRUE)

min(diabetes$age)
max(diabetes$age)
manju = diabetes[diabetes$gender == "male", c('hdl')]
mean(manju, na.rm = TRUE)
library(faraway)
plot(x=diabetes$weight, y=diabetes$hdl, col = "blue", xlab = "Weight", ylab =
"HDL", main = 'weight vs HDL')
plot(x=diabetes$weight, y=diabetes$chol, col = "red", xlab = "Weight", ylab =
"Cholesterol", main = "Cholesterol vs Weight")

boxplot(diabetes$hdl ~ diabetes$gender, col = c("red",'blue'), data =
diabetes, xlab = "Gender", ylab = "HDL", main = 'HDL vs Gender')
```

```
   hitters=read.csv("D:/hitters.csv")
   hist(hitters$Salary, col = "red", xlab = "salary", ylab = "frequency",
main = 'salary vs frequency' )

plot(hitters$CRBI, hitters$Salary, col = "blue", xlab = "CRBI", ylab =
"salary", main = 'CRBI vs salary')
boxplot(hitters$Salary ~ hitters$League, col = c("red",'blue'), data =
hitters, xlab = "salary", ylab = "league", main = 'salary vs league')
```