Assignment 1

1 Recitation Exercises

Que1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.
→In this case **flexible method will perform bette**r than the inflexible method as sample size n is very high, due to which the performance of the model will be high and have low variance.

(b) The number of predictors p is extremely large, and the number of observations n is small.
→In this case **flexible method will perform worse** than inflexible method as sample size n is small, due to which the performance of the model will be low which may result in overfitting of model.

(c) The relationship between the predictors and the response is highly non-linear.
→In this case **flexible method will perform better** than inflexible method as flexible methods are good at developing models which have a non-linear relationship. However, inflexible methods have difficulty in developing such models which may result in underfitting model.

(d) The variance of the error terms, i.e. σ2 = Var(), is extremely high.
→In this case **flexible method will perform worse** than inflexible method as high variance of error terms means a lot of noise is present in dataset. which may result in overfitting of model in case of flexible method.

Que 2. Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(a) We collect a set of data on the top 500 firms in the US. For each firm, we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
→This is a **regression problem** because the CEO salary is a continuous variable which depends on the independent variables like profit number of employees and industry. Also, this is an **inference problem** because here we are interested in finding how independent variables impact the salary of CEO. Here **n=500** and **p=3**.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product, we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
→This is a **classification problem** because here we want to find the success or failure of the firm. Also, this **is a prediction problem** because we are mainly interested in the success or failure of the company without concerning much on how vastly different parameters impact the firm. Here **n=20** and **p=13**.

(c) We are interested in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.
→This is a **regression problem** as the % change in the US dollar is a dependent variable which depends on independent variables like the % change in the US market, the % change in the British market, and the % change in the German market. Also, this is a **prediction problem** because here we are mainly interested in the % change in dollar not on the % impact of other parameters on % dollar change. Here **n=52** and **p=3**.

Que 3. You will now think of some real-life applications for statistical learning.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.
→

Assignment 1

I. Classification model will be useful to classify whether a person should buy a specific product, based on parameters like a requirement, quality, price, offers.

Response: buy/not buy
Predictors: Requirement, Quality, Price, Offers
Goal: prediction because here we are interested in finding whether he should buy the product or not.

II. Classification model will be useful to classify whether a person is eligible to take loan from the bank, based on different parameters like gender, age, salary, marital status, and monthly expenses.

Response: eligible/not eligible
Predictors: Gender, Age, Salary, Marital status, Monthly expenses
Goal: Prediction because here we are interested in finding whether a person is eligible for loan, not concerning much about the impact of parameters on eligibility of loan.

III. Classification model will be useful to classify whether a person should purchase a car to travel to his office, based on following parameters like distance to the office, travel time, regularly spent on public service transportation, salary.

Response: purchase/not purchase
Predictors: distance to office, travel time, regularly spent on public service transportation and salary
Goal: Prediction because here we are interested in finding whether a person should purchase a car or not.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

→

I. The regression model will be useful to predict insurance premium for different people based on age, deductibles applied, accident history, salary.

Response: Insurance premium amount (in $)
Predictors: age, deductibles applied, accident history, salary
Goal: Prediction because here we are interested in finding insurance premium for different people.

II. The regression model will be useful to find eligibility for % offers on Ventra passes for different people based on their occupation, age, especially abled.

Response: %offer (0% to 25%)
Predictors: occupation, age, especially abled
Goal: prediction because here we are interested in finding the %offer on Ventra for different people without concerning much about how much each parameter contribute to that offer.

III. The regression model will be useful to predict the % chances for confirmation of waiting list ticket in Indian railway system. It is based on following parameters like waiting list number, class of ticket, month of travel, source & destination.

Response: % chance_of_confirmation (0 to 100%)
Predictors: Waiting list number, Class of ticket, Month of travel and Source & Destination.
Goal: prediction because here we are interested in predicting the % chances for confirmation of ticket.

(c) Describe three real-life applications in which cluster analysis might be useful.

→

I. Based on individual preferences, people can be clustered into different groups having similar behavior and recommendations can sent to different groups based on their preferences.

II. In different areas, people with similar requirements (like Internet speed, usage, budget) can be grouped together and based on their requirements Internet service providers can be recommended in different groups.

III. People can be clustered into different groups based on their liking about different genres of movies (like action, horror, comedy, etc.) and accordingly new upcoming movies can be promoted by recommending them in a group.

Que 4. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

→Parametric methods make some assumption about the functional form which simplifies the problem of estimating function because it is generally much easier to estimate a set of parameters and predefine model then it is to fit an entirely arbitrary function. The potential disadvantage of the parametric approach is that the model we choose will usually not match the true unknown form. If the chosen models are too far from the truth, then our estimate will be poor. On the other hand, non-parametric methods do not make any external assumption about the functional form, instead they seek the estimate function that gets as close to the data points as possible without being too rough. such approaches can have a major advantage over parametric approaches by avoiding the assumption of particular functional form. They have the potential to accurately fit a wider range of possible shapes, but non-parametric approaches do suffer from major disadvantages since they do not reduce the problem of estimating function to a small number of parameters, a very large number of observations is it required in order to obtain an accurate estimate function.

Linear models such as linear regression and classification have various advantages of parametric method like we have a fixed number of parameters, so the complexity of model is less. While in non-parametric methods the complexity of model grows with the number of samples.

Que 5. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	**X1**	**X2**	**X3**	**Y**
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors

(a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

→

Obs.	**X1**	**X2**	**X3**	**Euclidean distance**
1	0	3	0	3
2	2	0	0	2
3	0	1	3	3.16
4	0	1	2	2.24
5	-1	0	1	1.41
6	1	1	1	1.73

(b) What is our prediction with K = 1? Why?

→for k=1, we consider only single nearest neighbor of point (0,0,0). In our case, it is Obs. 5 (with distance 1.41). Now, as the Obs. 5 has green color. So, model will predict green color for point (0,0,0).

(c) What is our prediction with K = 3? Why?
→for k=3, we consider three nearest neighbors of point (0,0,0). In our case it is Obs. 5 (with distance 1.41), Obs. 6(with distance 1.73) and Obs. 2(with distance 2). Now, as the Obs. 2 and Obs.6 have Red color. So, the model will predict Red color for point (0,0,0).

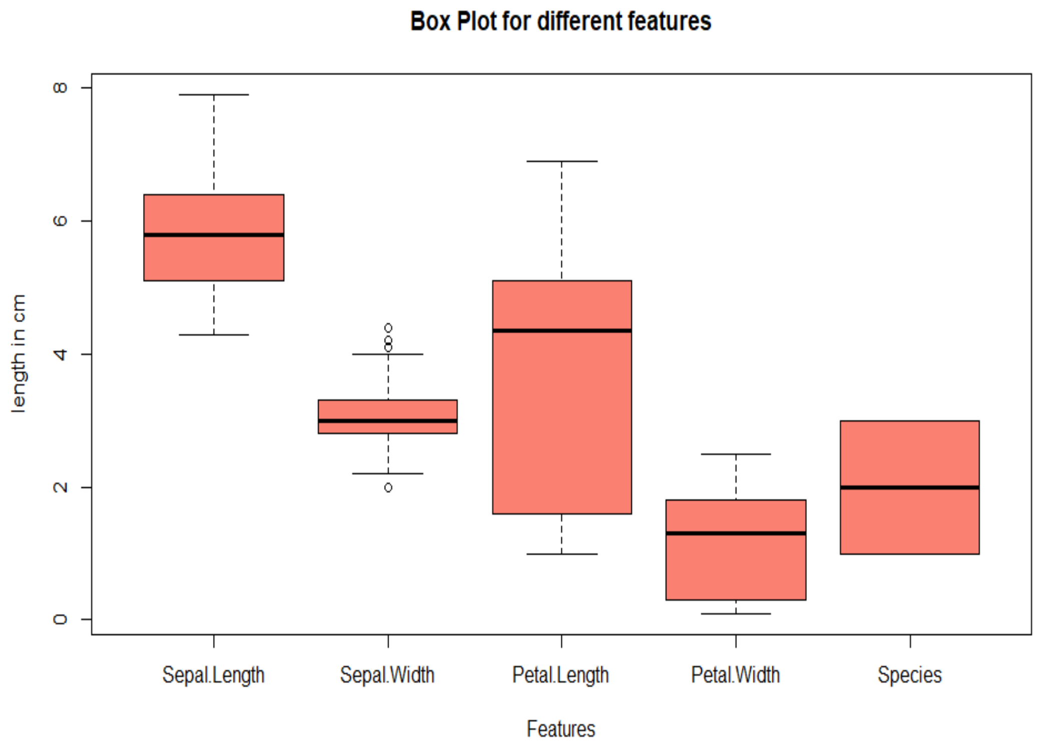
(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?
→for the higher value of k, the Bayes boundary will become almost linear. However, in this case Bayes boundary is highly noncolinear which suggest that the value for K should be small.

==

2. Practicum Problems

PROBLEM 1

```
#Load Libraries
library(gridExtra)
library(ggplot2)


#load iris dataset
iris <-data.frame(iris)

#boxplot for attributes:
boxplot(iris,main="Box Plot for different features",xlab="Features",ylab="length in cm",
col = "salmon")
```



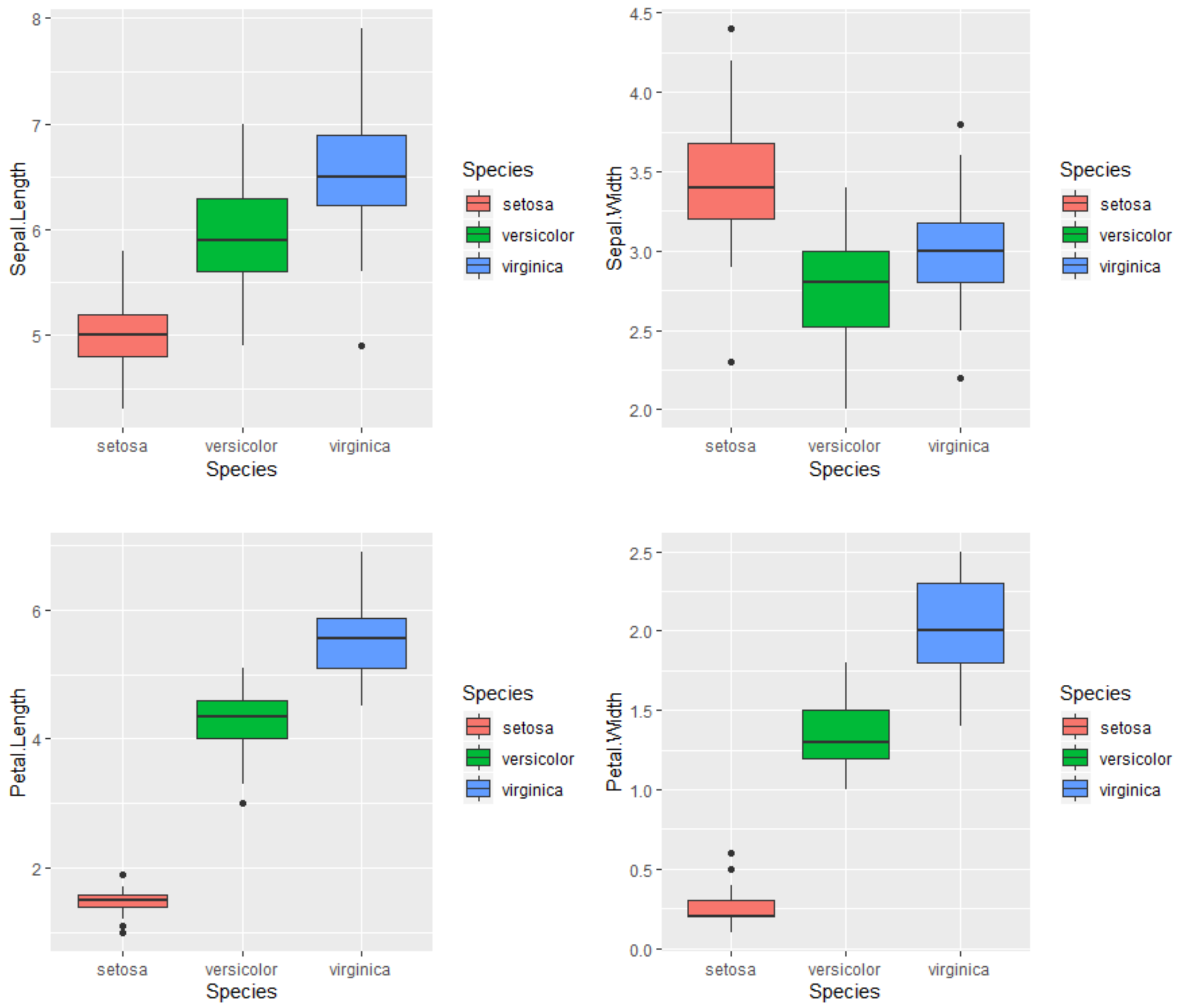
Assignment 1

```
#Inter Quartile range calculation:

# for Sepal length:
IQR(iris$Sepal.Length)
## [1] 1.3

# for Sepal Width:
IQR(iris$Sepal.Width)
## [1] 0.5

# for Petal length:
IQR(iris$Petal.Length)
## [1] 3.5

# for Petal Width:
IQR(iris$Petal.Width)
## [1] 1.5
```

From the above boxplots we can conclude that as the **petal Length** has maximum length boxplot. So, Petal length will have maximum Inter quartile range. Also, on calculating Inter quartile range for each attribute we found the same results.

```

#standard-deviation calculation

# for sepal length:
sd(iris$Sepal.Length)
## [1] 0.8280661

# for sepal Width:
sd(iris$Sepal.Width)
## [1] 0.4358663

# for Petal Length:
sd(iris$Petal.Length)
## [1] 1.765298

# for Petal Width:
sd(iris$Petal.Length)
## [1] 1.765298

#Boxplot of attributes with class species
library(ggplot2)
a <- ggplot(iris, aes(Species, Sepal.Length,  fill=Species)) + geom_boxplot()
b <-ggplot(iris, aes(Species, Sepal.Width,  fill=Species)) +  geom_boxplot()
c <-ggplot(iris, aes(Species, Petal.Length,  fill=Species))  +   geom_boxplot()
d <-ggplot(iris, aes(Species, Petal.Width,  fill=Species))  +   geom_boxplot()

grid.arrange(a + ggtitle(""),
             b + ggtitle(""),
             c + ggtitle(""),
             d + ggtitle(""),
             nrow = 2)
```

Assignment 1

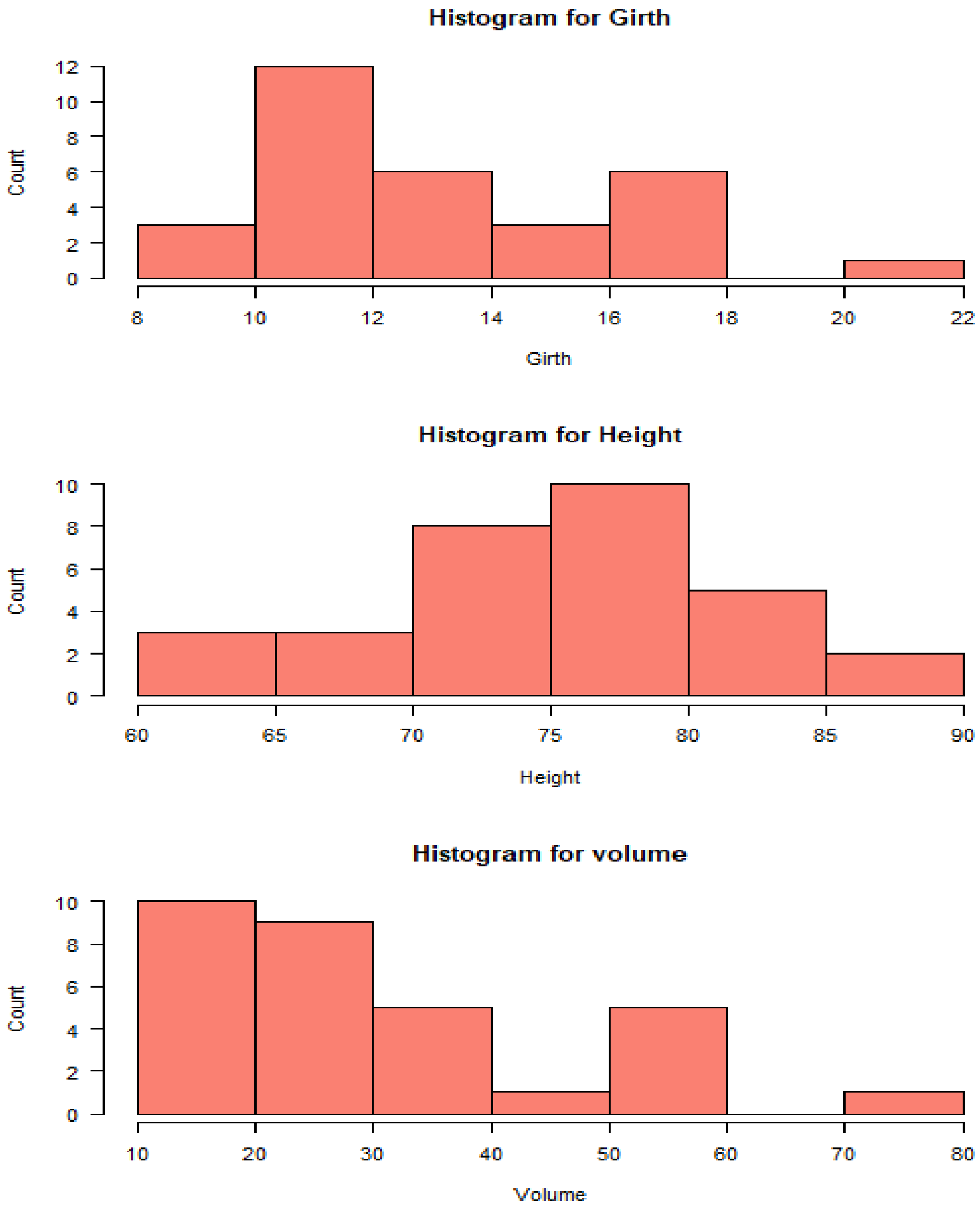


From the above boxplots we can conclude that on separating each class species of different attributes class **Setosa** has comparatively different values of petal length and width.

==

PROBLEM 2

```
#load library
library(moments)

#load tree sample dataset
trees <-data.frame(trees)

#summary of attributes
summary(trees)

##      Girth          Height       Volume
##  Min.   : 8.30   Min.   :63   Min.   :10.20
##   1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##   Median :12.90   Median :76   Median :24.20
##   Mean   :13.25   Mean   :76   Mean   :30.17
##   3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##   Max.   :20.60   Max.   :87   Max.   :77.00

#Histogram of attributes
par(mfrow=c(3,1))
hist(trees$Girth,main="Histogram for Girth", xlab = "Girth", ylab = "Count",border="black
", col="salmon",las=1)
hist(trees$Height,main="Histogram for Height", xlab = "Height", ylab = "Count",border="bl
ack", col="salmon",las=1)
hist(trees$Volume,main="Histogram for volume", xlab = "Volume", ylab = "Count",border="bl
ack", col="salmon",las=1)
```



From the above histograms, we can conclude that **Height** attribute appears to have close to **normal distribution** with **negative skewness**. **Girth** attribute appears to have **positive skewness** and **Volume** attribute appears to have **highly positive skewness**.

```
#Import library skewness
library(moments)

#Skewness Calculation

#for Girth
skewness(trees$Girth, na.rm = FALSE)
## [1] 0.5263163

#for Height
skewness(trees$Height, na.rm = FALSE)
## [1] -0.374869

#for Volume
skewness(trees$Volume, na.rm = FALSE)
## [1] 1.064357
```

On calculating skewness of each attribute, we found the same results as on visual inspection. Here the skewness of height attribute is in range from -0.5 to +0.5. Hence, it is close to normal distribution with negative skewness as it has negative skewness value. Skewness of girth attribute is in range from 0.5 to 1. Hence, it is moderately positive skewed, and the skewness of volume attribute is greater than 1. Hence, is highly positive skewed.

===

PROBLEM 3

```
#Load the auto-mpg sample dataset from the UCI Machine Learning Repository
URL <-"https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data"
ndata <- read.table(URL,stringsAsFactors = FALSE)
colnames(ndata) <- c("mpg","cylinders","displacement","horsepower","weight","acceleration
","model_year","origin","car_name")

#Display number of rows
nrow(ndata)

## [1] 398

#Display number of rows:
ncol(ndata)
## [1] 9

#Diplay first 6 rows:
head(ndata)
```

Assignment 1

```
##   mpg cylinders displacement horsepower weight acceleration model_year
## 1  18         8          307      130.0   3504         12.0         70
## 2  15         8          350      165.0   3693         11.5         70
## 3  18         8          318      150.0   3436         11.0         70
## 4  16         8          304      150.0   3433         12.0         70
## 5  17         8          302      140.0   3449         10.5         70
## 6  15         8          429      198.0   4341         10.0         70
##   origin                  car_name
## 1      1 chevrolet chevelle malibu
## 2      1         buick skylark 320
## 3      1        plymouth satellite
## 4      1             amc rebel sst
## 5      1               ford torino
## 6      1          ford galaxie 500

#Display structure of variables:
str(ndata)

## 'data.frame':    398 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : chr  "130.0" "165.0" "150.0" "150.0" ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ model_year  : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ car_name    : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satell
ite" "amc rebel sst" ...

#Converting datatype of column horsepower from character to numeric:
ndata$horsepower<- as.numeric(ndata$horsepower)

## Warning: NAs introduced by coercion

#Checking datatype of horsepower column whether it is converted to numeric value or not:
is.numeric(ndata$horsepower)
## [1] TRUE

#Display structure of variables
str(ndata)

## 'data.frame':    398 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ model_year  : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ car_name    : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satell
ite" "amc rebel sst" ...
```

Assignment 1

```
#Calculate median of column horsepower:
median_horsepower<-median(ndata$horsepower,na.rm = TRUE)
median_horsepower
## [1] 93.5

#Calculate mean of column horsepower:
mean_horsepower<-mean(ndata$horsepower,na.rm = TRUE)
mean_horsepower
## [1] 104.4694

#Calculate number of rows having null value for column horsepower:
num<-sum(is.na(ndata$horsepower))
num
## [1] 6

#Replace NA values of horsepower column with its median value:
ndata$horsepower[is.na(ndata$horsepower)]<-median_horsepower

#Check number of rows having null value for column horsepower:
num<-sum(is.na(ndata$horsepower))
num
## [1] 0

#Check the new mean
mean(ndata$horsepower)
## [1] 104.304
```

From the above results we can conclude that **on replacing all the NA values of Horsepower attribute with median value, mean of the attribute will slightly change from 104.469 to 104.304.**

===