Assignment 2

Tayyab Munir - 11716089

9/11/2020

Question 1

This exercise relates to the College data set, which can be found in the file College.csv on the course's public webpage (https://scads.eecs.wsu.edu/index.php/datasets/). The dataset contains a number of variables for 777 different universities and colleges in the US.

(a) Use the read.csv() function to read the data into R, or the csv library to read in the data with python. In R you will load the data into a dataframe. In python you may store it as a list of lists or use the pandas dataframe to store your data. Call the loaded data college. Ensure that your column headers are not treated as a row of data.

```
college = read.csv("https://scads.eecs.wsu.edu/wp-
content/uploads/2017/09/College.csv")
```

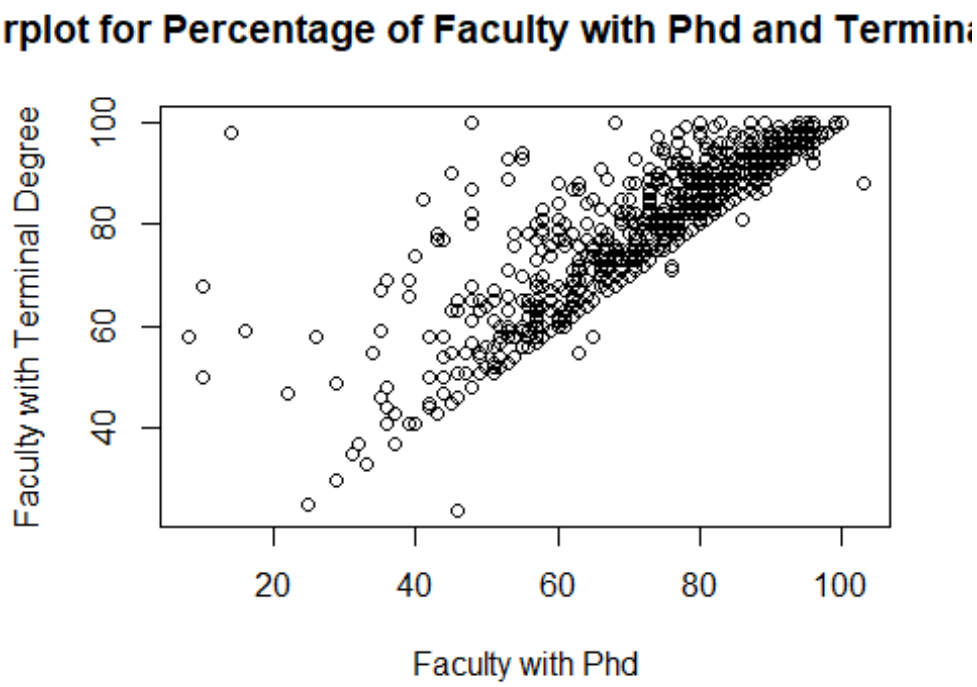
(b) Find the median cost of books for all schools in this dataset.

```
median(college$Books)

## [1] 500
```

(c) Produce a scatterplot that shows a relationship between two numeric (not factor or boolean) features of your choice in the dataset. Ensure it has appropriate axis labels and a title.

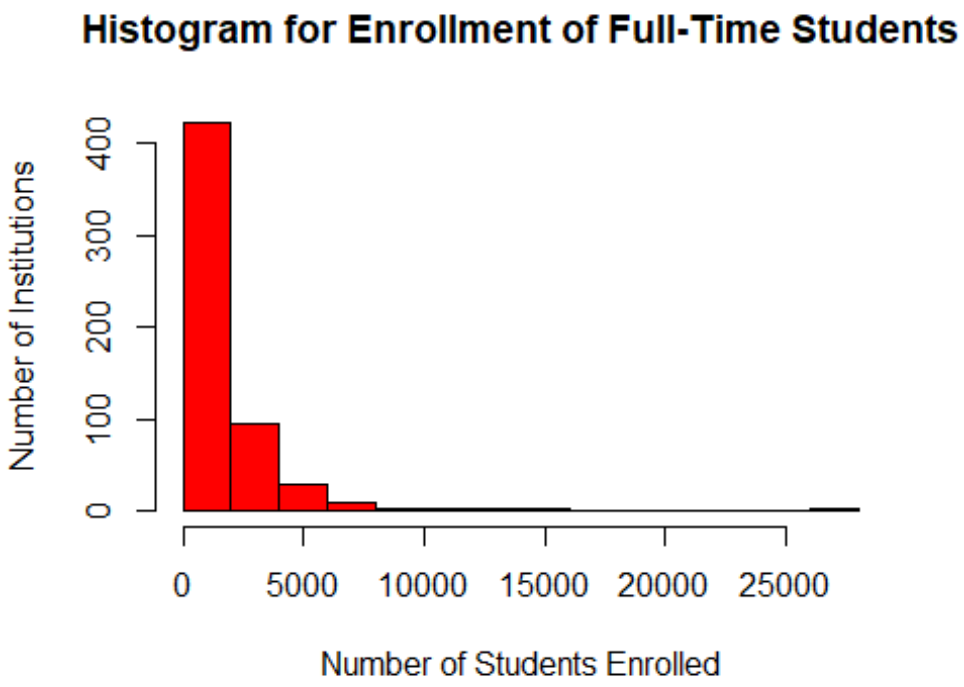
```
plot(college$PhD,college$Terminal, main = "Scatterplot for Percentage of
Faculty with Phd and Terminal Degrees", xlab = "Faculty with Phd", ylab =
"Faculty with Terminal Degree")
```



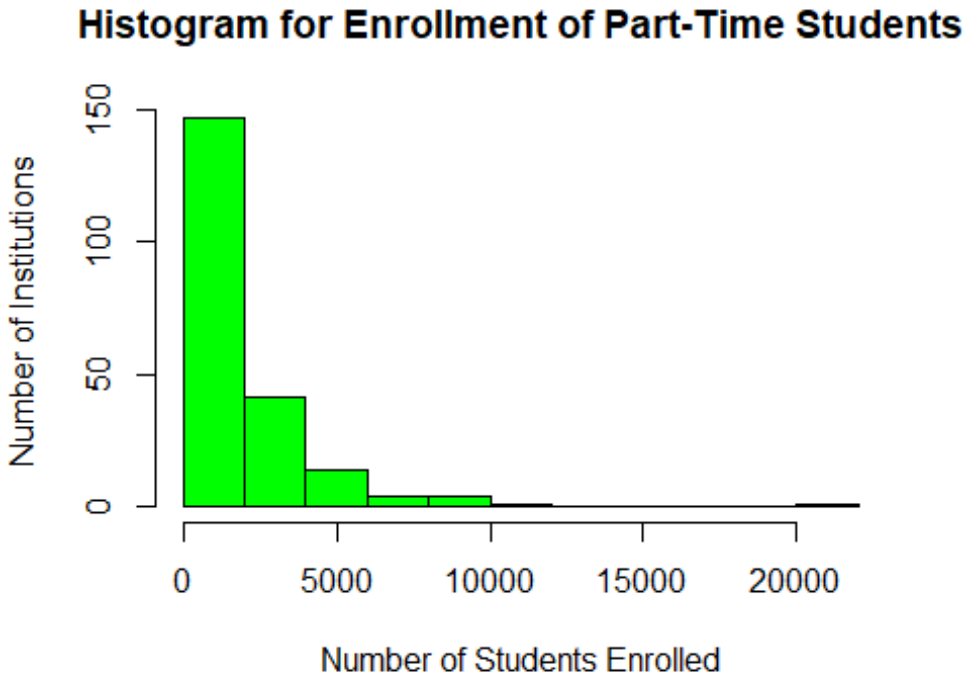
(d) Produce a histogram showing the overall enrollment numbers (P.Undergrad plus F.Undergrad) for both public and private (Private) schools. You may choose to show both on a single plot (using side by side bars) or produce one plot for public schools and one for private schools. Ensure whatever figures you produce have appropriate axis labels and a title.

```
private_college = college[college$Private=="Yes",]
public_college = college[college$Private=="No",]

hist(private_college$F.Undergrad, main = "Histogram for Enrollment of Full-
Time Students", col = "red", xlab = "Number of Students Enrolled", ylab =
"Number of Institutions")
```



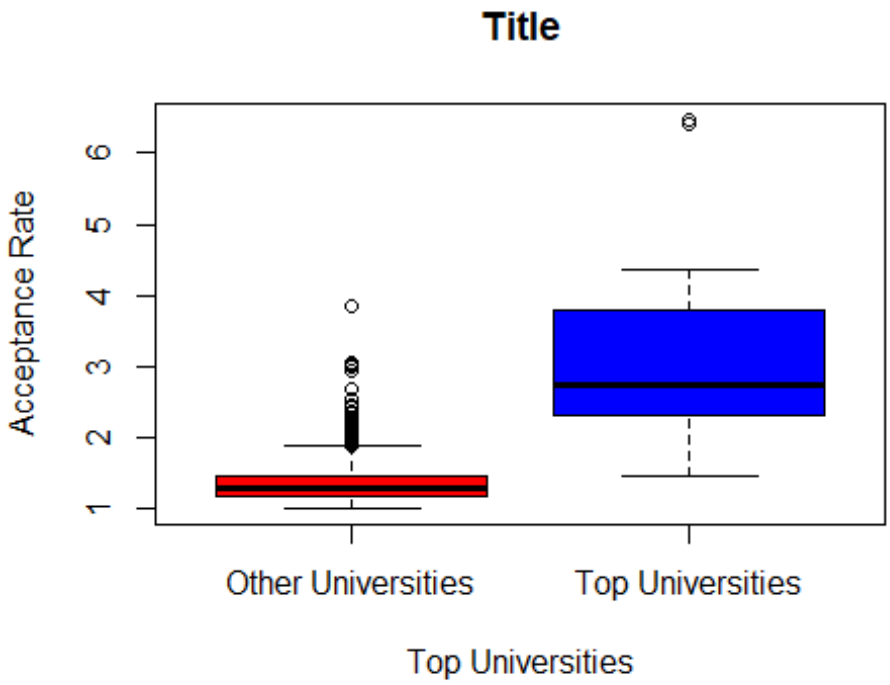

```r
hist(public_college$P.Undergrad, main = "Histogram for Enrollment of Part-Time Students", col = "green", xlab = "Number of Students Enrolled", ylab = "Number of Institutions")
```



(e) Create a new qualitative variable, called Top, by binning the Top10perc variable into two categories (Yes and No). Specifically, divide the schools into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 75%. Now produce side-by-side boxplots of the schools' acceptance rates (based on Accept and Apps) for each of the two Top categories. There should be two boxes on your figure, one for top schools and one for others. How many top universities are there?

```
Top=rep("Other Universities", nrow(college))
Top[college$Top10perc>75]="Top Universities"
Top=as.factor(Top)
college=data.frame(college, Top)
college$acceptance <- (college$Apps / college$Accept)

boxplot(college$acceptance ~ college$Top, col = c("red", "blue"), main =
"Title", xlab = "Top Universities", ylab = "Acceptance Rate")
```

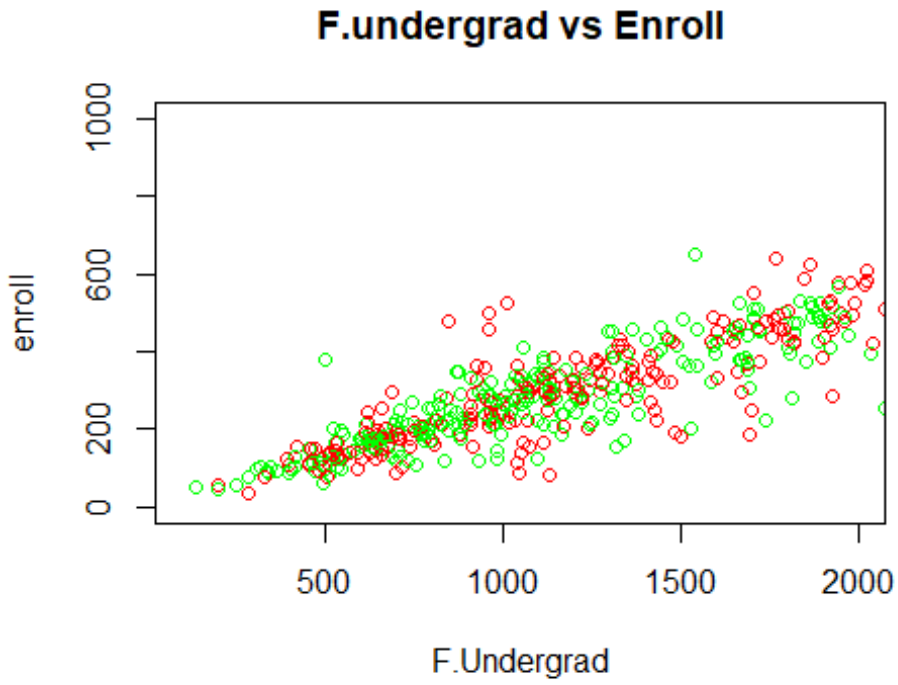


(f) Continue exploring the data, producing two new plots of any type, and provide a brief (one to two sentence) summary of your hypotheses and what you discover. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge.

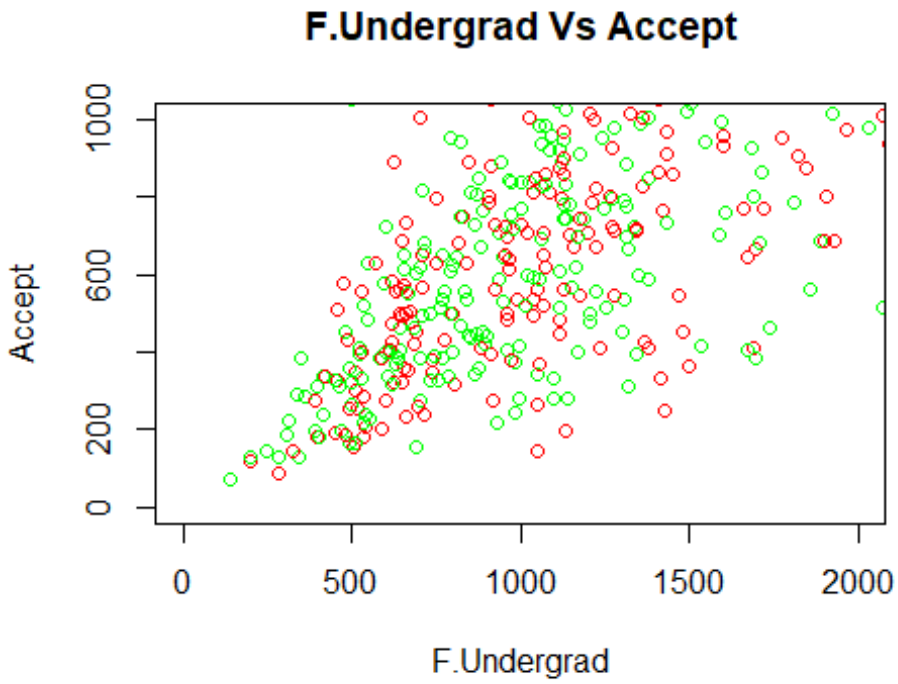
After plotting different possible combination of graphs. we can relate some variable such as mentioned below: (1.) The variables F.Undergrad and Enroll has almost a linear

relationship. i.e if the variable enroll increases there will also be a increase in F.Undergrad. (2.)The variables F.Undergrad and Accept has almost a linear relationship. i.e if the variable Accept increases there will also be a increase in F.Undergrad. (3.) Variables Apps and Accept have almost a linear relationship.In other words they both are directly proportional in nature.

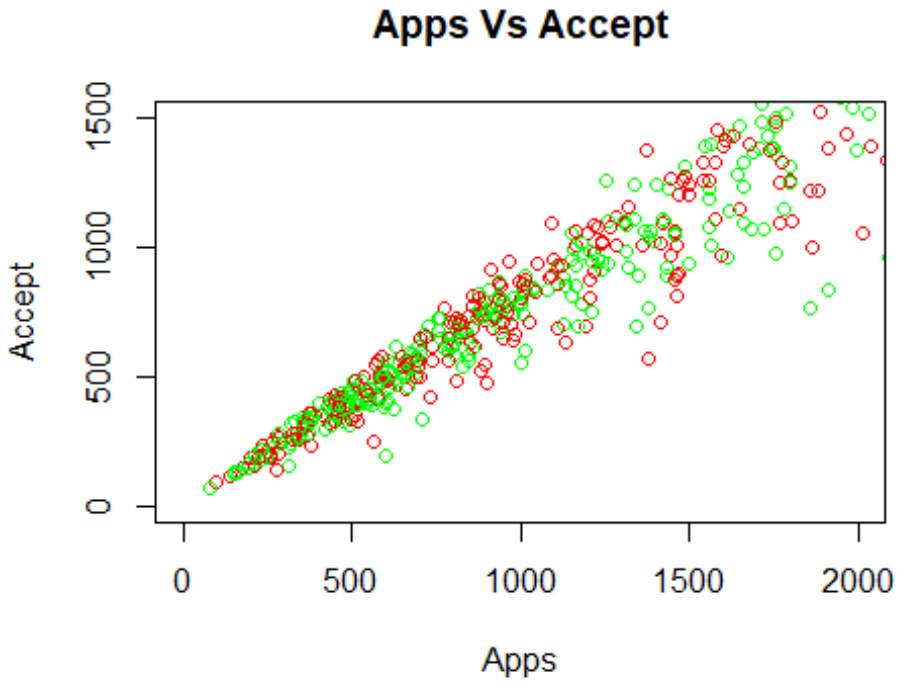
```
plot(x = college$F.Undergrad, y = college$Enroll, xlim= c(100, 2000), ylim=
c(00,1000), xlab = "F.Undergrad", ylab = "enroll", main ="F.undergrad vs
Enroll", col = c("green","red"))
```




```
plot(x = college$F.Undergrad, y = college$Accept, xlim=c(0,2000),
ylim=c(0,1000), xlab ="F.Undergrad", ylab = "Accept", main = "F.Undergrad Vs
Accept", col = c("green","red"))
```




```
plot(x = college$Apps, y = college$Accept, xlim=c(0,2000), ylim=c(0,1500),
xlab ="Apps", ylab = "Accept", main = "Apps Vs Accept", col =
c("green","red"))
```



Question 2

This exercise involves the Auto.csv data set found on the course website. The features of the dataset are as follows:

• mpg: miles per gallon

• cylinders: number of cylinders

• displacement: volume of air displaced by cylinders

• horsepower: power of the car (rate of work)

• weight: how much the car weighs in lb

• acceleration: rate at which car accelerates

• year: when the car was made

• origin: where the car comes from (1=USA, 2=Germany, 3=Japan)

• name: the make and model of the car

Make sure that rows with missing values have been removed from the data. For part, show both the code you used and any relevant outputs.

```
auto = read.csv("C:/Users/tayya/OneDrive/Desktop/Fall 2020/Data
Science/Assignment 2/Auto.csv")
auto = na.omit(auto)
```

(a) Specify which of the predictors are quantitative (measuring numeric properties such as size, or quantity), and which are qualitative (measuring non-numeric properties such as color, appearance, type etc.)? Keep in mind that a qualitative variable may be represented as a quantitative type in the dataset, or the reverse. You may wish to adjust the types of your variables based on your findings.

```
sapply(auto, class)
```

```
##          mpg    cylinders displacement   horsepower       weight
acceleration
##    "numeric"    "integer"    "numeric"    "integer"    "integer"
"numeric"
##         year       origin         name
##    "integer"    "integer"  "character"
```

Following list shows the variables of the Auto data and their classification as quantitative and qualitative data.

1. mpg -> Quantitative (consist data of numeric value)
2. cylinders -> Quantitative (consist data of integer value)
3. displacement -> Quantitative (consist data of numeric value)

4. horsepower -> Quantitative (consist data of integer value)
5. weight -> Quantitative (consist data of integer value)
6. acceleration -> Quantitative (consist data of numeric value)
7. year -> Qualitative (consist data of integer value but is qualitative)
8. origin -> Qualitative (consist data of integer value but is qualitative)
9. name -> Qualitative (consist of name of the car and is treated as factor)

(b) What is the range, mean and standard deviation of each quantitative predictor?

```
sapply(auto[,c(1:6),],range)
```

```
##       mpg cylinders displacement horsepower weight acceleration
## [1,]  9.0         3           68         46   1613          8.0
## [2,] 46.6         8          455        230   5140         24.8
```

```
sapply(auto[,c(1:6),],mean)
```

```
##          mpg    cylinders displacement   horsepower       weight
acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184
15.541327
```

```
sapply(auto[,c(1:6),],sd)
```

```
##          mpg    cylinders displacement   horsepower       weight
acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560
2.758864
```

(c) Now remove the 40th through 80th (inclusive) observations from the dataset. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
new.auto = subset(auto[-c(40:80),])

sapply(new.auto[,c(1:6),],range)
```

```
##       mpg cylinders displacement horsepower weight acceleration
## [1,]  9.0         3           68         46   1649          8.0
## [2,] 46.6         8          455        230   4997         24.8
```

```
sapply(new.auto[,c(1:6),],mean)
```

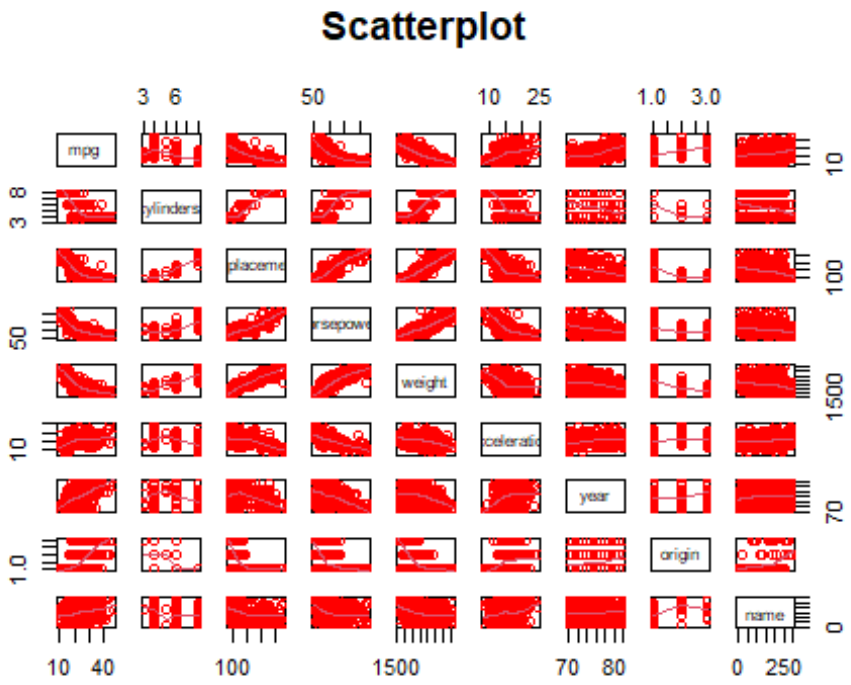
```
##          mpg    cylinders displacement   horsepower       weight
acceleration
##    23.931054     5.424501   190.943020   103.019943  2948.934473
15.581766
```

```
sapply(new.auto[,c(1:6),],sd)
```

```
##          mpg    cylinders displacement   horsepower       weight
acceleration
##     7.826817     1.667975   101.726508    37.711797   815.903085
2.730831
```

(d) Using the full data set, investigate the predictors graphically, using scatterplots, correlation scores or other tools of your choice. Create a correlation matrix for the relevant variables.

```
values = auto[, !sapply(auto, is.factor)]
plot(values, panel = panel.smooth, main = "Scatterplot", col = "red")
```

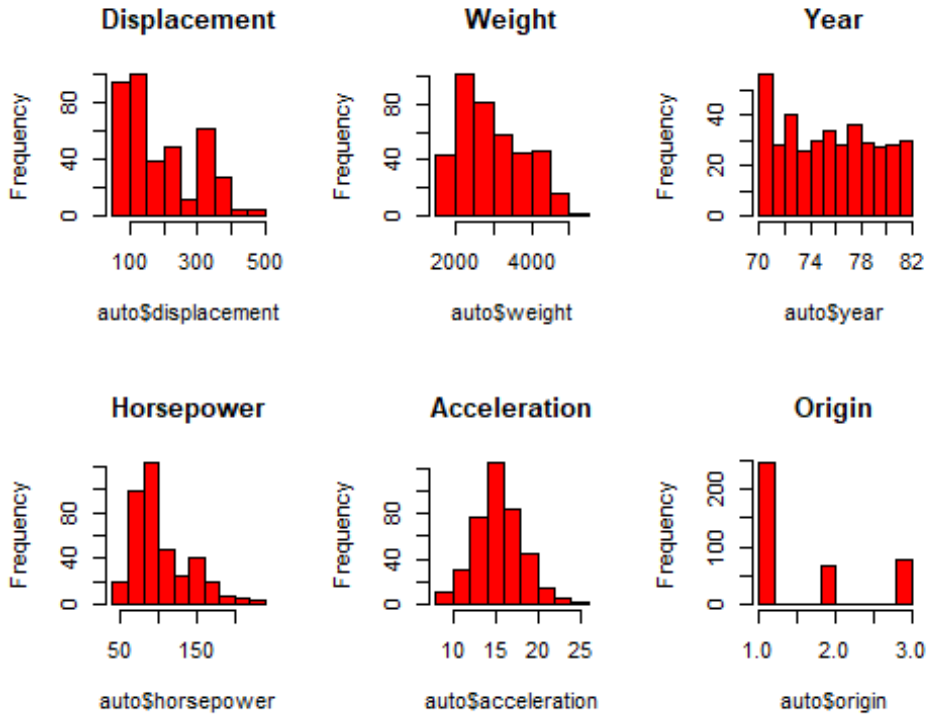


The above shown graph shows that the Miles per galon (mpg) of a car is continously decreasing with respect to each attribute that is increasing. For example if we look over the relation of cylinder of a car and mileage, till 4 cylinder the mileage increases but after that it decreases gradually. It also depict that weight, cylinder, displacement, horse power, and acceleration are directly proportional to each other, with increase in weight of a car it increases the number of cylinders in it, its horse power, acceleration, and displacement.

Following histograms are of displacement, horsepower, weight, acceleration, year and origin.

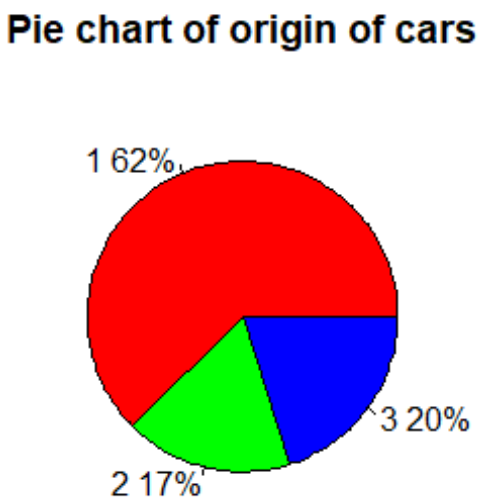
```
par(mfcol = c(2, 3))
hist(auto$displacement, col = "red", main = "Displacement")
hist(auto$horsepower, col = "red", main = "Horsepower")
hist(auto$weight, col = "red", main = "Weight")
hist(auto$acceleration, col = "red", main = "Acceleration")
```

```
hist(auto$year, col = "red", main = "Year")
hist(auto$origin, col = "red", main = "Origin")
```



Following is the pie graph of origin of cars. It shows that 62% of cars are designed from the origin no 1, 17% from origin no. 2, and 20% from thee origin no. 3. These 1, 2, and 3 represent any place, or makers of the cars.

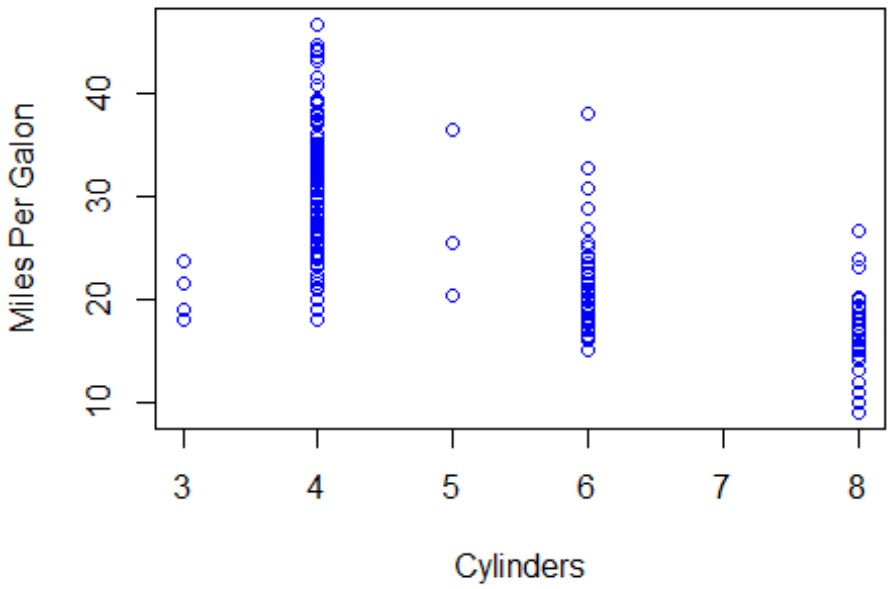
```
chart = table(auto$origin)
perc = round(chart/sum(chart)*100)
name = paste(names(chart),perc)
lbl = paste(name, "%", sep = "")
pie(chart, labels = lbl, col = rainbow(length(lbl)), main = "Pie chart of
origin of cars")
```



(e) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Which, if any, of the other variables might be useful in predicting mpg? Justify your answer based on the prior correlations.

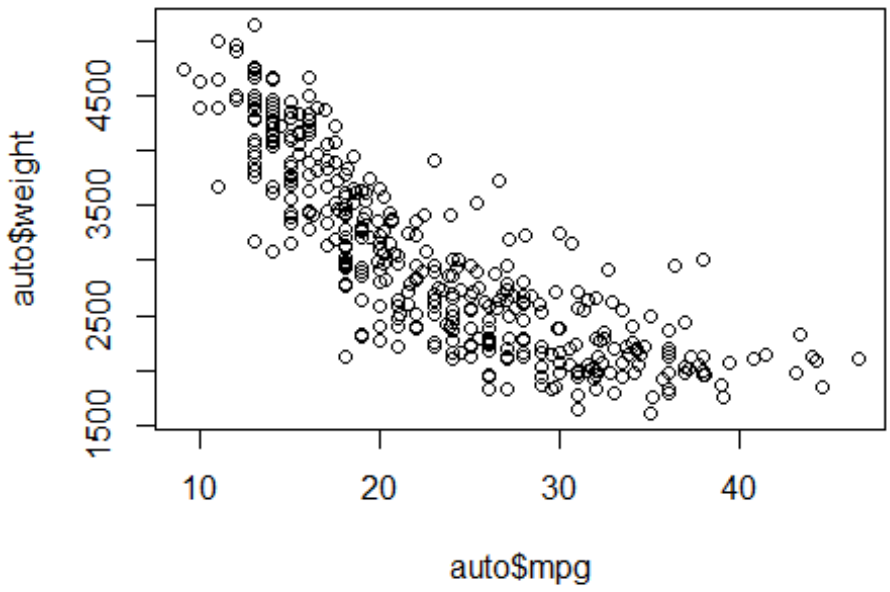
Following plots shows the behaviour of predicting gas mileage(mpg) on the basis of other variables.

```
plot(auto$cylinders, auto$mpg, xlab ="Cylinders", ylab = "Miles Per
Galon",col="blue")
```



The above drawn plot depicts that when we increas the number of cylinders to 4, the gas mileage increases too, but increasing it more than 4 cylinders drops down the mileage. So it shows that increasing the number of cylinders is not efficient to increase gas mileage.

```
plot(auto$mpg, auto$weight)
```



The plot drawn above shows the realation of car weight and its gas mileage. It clearly shows that increasing the weight of the car decreases the car mileage, that is mpg is inversely proportional to the weight of the car.
