Name: Chiayu Tu (Louis)
Stat 387 Assignment One

Conceptual:

n : Sample Size
p : Predictors
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.

- 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.

    - The scenario is Regression
    - Most interested in Inference
    - n = 500 firms in the US
    - p = 3 (profit, number of employees, industry)

- 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.

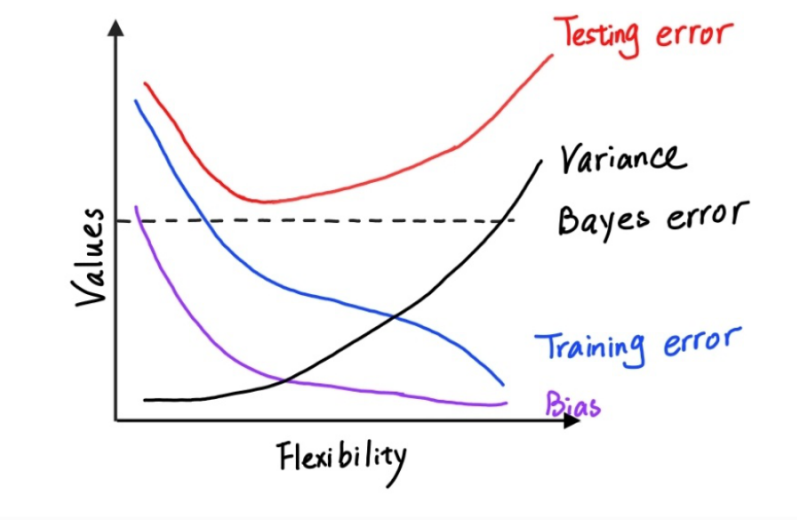
    - The scenario is Classification
    - Most interested in Prediction
    - n = 20 similar products
    - p = 13 (price charged for the product, marketing budget, competition price, and ten other variables)

- We are interested in predicting the % change in the USD/Euro exchange rate 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 USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

    - The scenario is Regression
    - Most interested in Prediction
    - n = 52
    - p = 4 (the % charge in the USD/Euro, the % charge in the US market, the % charge in the British market, and the % charge in the German market.

We now revisit the bias-variance decomposition

- Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.



- Explain why each of the five curves has the shape displayed in part (a).
  - Bayes error is the lowest possible prediction error that can be achieved and is same as irreducible error. If One would know exactly what process generates the data, then errors will still be made if the process is random.
  - Training error is a flexibility increases, the error decreases as the model is fitted to the training data set.
  - Test error will always be higher than the variance, because the variance can not be predicted by the model. This is where the flexibility.
  - Bias is inversely proportional to flexibility. Bias decreases as the model is fitted to the training data set.
  - variance is a feasible increase in the curre as flexibility increases from its initial value. With the increase in flexibility, the system function becomes less robust, thereby increase in the variance.

Applied:

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

- Which of the predictors are quantitative, and which are qualitative?

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
print("mpg          is quantitative.")
```

```
## [1] "mpg          is quantitative."
```

```
print("Cylinders    is quantitative.")
```

```
## [1] "Cylinders    is quantitative."
```

```
print("Displacement is quantitative.")
```

```
## [1] "Displacement is quantitative."
```

```
print("Horsepower   is quantitative.")
```

```
## [1] "Horsepower   is quantitative."
```

```
print("Weight       is quantitative.")
```

```
## [1] "Weight       is quantitative."
```

```
print("Acceleration is quantitative.")
```

```
## [1] "Acceleration is quantitative."
```

```
print("Year         is quantitative.")
```

```
## [1] "Year         is quantitative."
```

```
print("Origin       is quantitative.")
```

```
## [1] "Origin       is quantitative."
```

```
print("Name         is qualitative.")
```

```
## [1] "Name         is qualitative."
```

- What is the range of each quantitative predictor? You can answer this using the range() function.

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,]  9.0         3           68         46   1613          8.0   70
## [2,] 46.6         8          455        230   5140         24.8   82
```

- What is the mean and standard deviation of each quantitative predictor?

```
print("The mean of each quantitative predictor:")
```

```
## [1] "The mean of each quantitative predictor:"
```

```
apply(Auto[, 1:7], 2, mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year
##    75.979592
```

```
print("The standard deviation of each quantitative predictor:")
```

```
## [1] "The standard deviation of each quantitative predictor:"
```

```
apply(Auto[, 1:7], 2, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

- Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
print("The range of data removed the 10th through 85th:")
```

```
## [1] "The range of data removed the 10th through 85th:"
```

```
apply(Auto[-c(10:85), 1:7], 2, range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,] 11.0         3           68         46   1649          8.5   70
## [2,] 46.6         8          455        230   4997         24.8   82
```

```
print("The mean of data removed the 10th through 85th:")
```

```
## [1] "The mean of data removed the 10th through 85th:"
```

```
apply(Auto[-c(10:85), 1:7], 2, mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899
##         year
##    77.145570
```

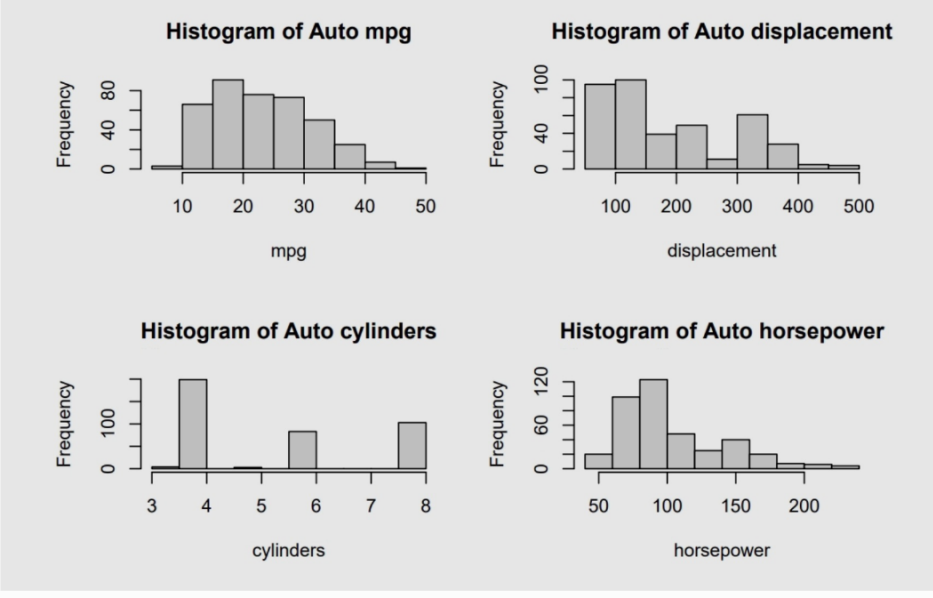
```
print("The standard deviation of data removed the 10th through 85th:")
```

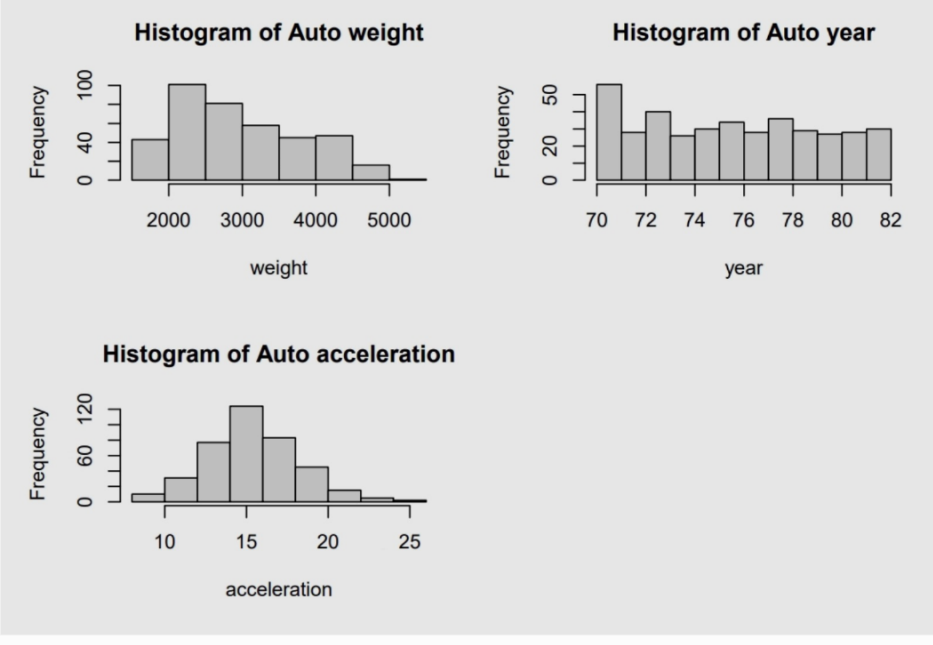
```
## [1] "The standard deviation of data removed the 10th through 85th:"
```

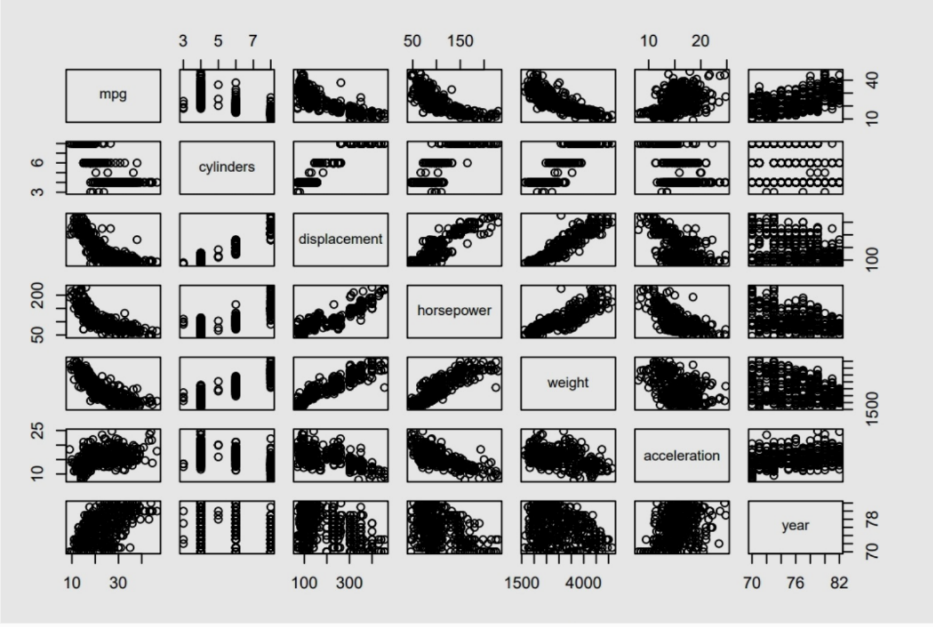
```
apply(Auto[-c(10:85), 1:7], 2, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721
##         year
##     3.106217
```

- Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.







- Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    1.0000000   -0.7776175   -0.8051269   -0.7784268   -0.8322442    0.4233285
##         year
##    0.5805410
```

```
##       Variable    Min    Max    Mean Standard Deviation
## 1          mpg   11.0   46.6   24.40               7.87
## 2    cylinders    3.0    8.0    5.37               1.65
## 3 displacement   68.0  455.0  187.24              99.68
## 4   horsepower   46.0  230.0  100.72              35.71
## 5       weight 1649.0 4997.0 2935.97             811.30
## 6 acceleration    8.5   24.8   15.73               2.69
## 7         year   70.0   82.0   77.15               3.11
```

R code:

```
library(ISLR)
library(ggplot2)
library(GGally)
Auto = na.omit(Auto)
head(Auto)

# Question One
print("mpg          is quantitative.")
print("Cylinders    is quantitative.")
print("Displacement is quantitative.")
print("Horsepower   is quantitative.")
print("Weight       is quantitative.")
print("Acceleration is quantitative.")
print("Year         is quantitative.")
print("Origin       is quantitative.")
print("Name         is qualitative.")

# Question Two
apply(Auto[, 1:7], 2, range)

# Question Three
print("The mean of each quantitative predictor:")
apply(Auto[, 1:7], 2, mean)
print("The standard deviation of each quantitative predictor:")
apply(Auto[, 1:7], 2, sd)

# Question Four
print("The range of data removed the 10th through 85th:")
apply(Auto[-c(10:85), 1:7], 2, range)

print("The mean of data removed the 10th through 85th:")
apply(Auto[-c(10:85), 1:7], 2, mean)

print("The standard deviation of data removed the 10th through 85th:")
```

```
apply(Auto[-c(10:85), 1:7], 2, sd)

# Question Five
#auto
auto = Auto[, 1:7]

# Histogram
par(mfcol = c(2, 2))
    for(i in 1:ncol(auto)){
        hist(auto[, i],
        main = paste("Histogram of Auto", names(auto)[i], sep = " "),
        xlab = names(auto)[i])
    }

# Pair
pairs(auto)

# Question Six
sapply(auto[, !sapply(auto, is.factor)], function(x) cor(auto$mpg, x))

auto.rm <- auto[-c(10:85), ]

temp <- NULL

for (i in 1:ncol(auto)) {
    if(is.factor(auto[, i]) == F) {
        temp=rbind(temp, data.frame(colnames(auto.rm[i]),
                round(min(auto.rm[, i]), digits = 2),
                round(max(auto.rm[, i]), digits = 2),
                round(mean(auto.rm[, i]), digits = 2),
                round(sd(auto.rm[, i]), digits = 2)))
        }
}

colnames(temp)=c("Variable", "Min", "Max", "Mean", "Standard
Deviation")
```

```
temp
rm(temp)
```