Homework #1

Sunny Lee

2021-02-03

1.

a) Regression, and we are interested in infering. $n = 500, p = 4$.

b) Classification, we are trying to predict if our product will fail or succeed, interested in prediction, $n = 20, p = 13$

c) Regression as we are trying to predict the % change in the US dollar, interested in prediction. $n = 52, p = 4$.

2. Parametric methods cut down on the possible solutions by assuming the parametric form of f. This makes it much easier to train a model, as we are really only trying to estimate the β_p not some arbitrary function. However, since we are assuming the form of the function, our model might not fit our data and taking more flexible parametric forms may result in overfitting. Non-parametric methods do not make assumptions about the parametric form of f and thus make them much more versatile than parametric methods. However, because non-parametric methods are so versatile, there needs to be a very large amount of data in order to actually find the form of f.

3.

a)
b) 3

ii) 2
iii) $\sqrt{10}$
iv) $\sqrt{5}$
v) $\sqrt{2}$
vi) $\sqrt{5}$

b) For $K = 1$, we predict green, as it is closest to observation 5, which is green with a distance of $\sqrt{2}$.

c) When $K = 3$, we find that our new observation point is closest to one red and one green, however there is a tie between one red and one green, so depending on which one of the tied we take as our third neightbor, our new prediction can be considered green or red.

4.

a)

```
college <- read.csv("College.csv")
```

b)

```
rownames(college) <- college[,1]

college <- college [,-1]
```

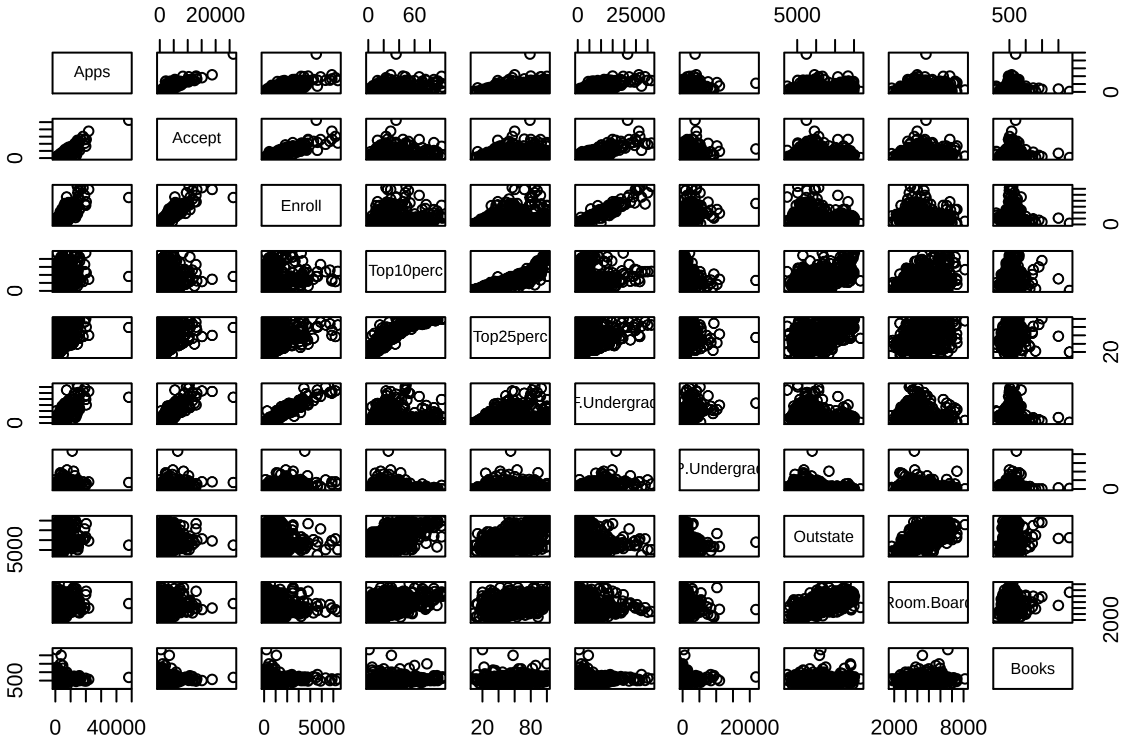
c)
d)

```
summary(college)
```

```
##    Private               Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 3002   Mean   : 2019   Mean   : 780
##                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                     Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board       Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD            Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend        Grad.Rate
##  Min.   : 3186   Min.   : 10.00
##  1st Qu.: 6751   1st Qu.: 53.00
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

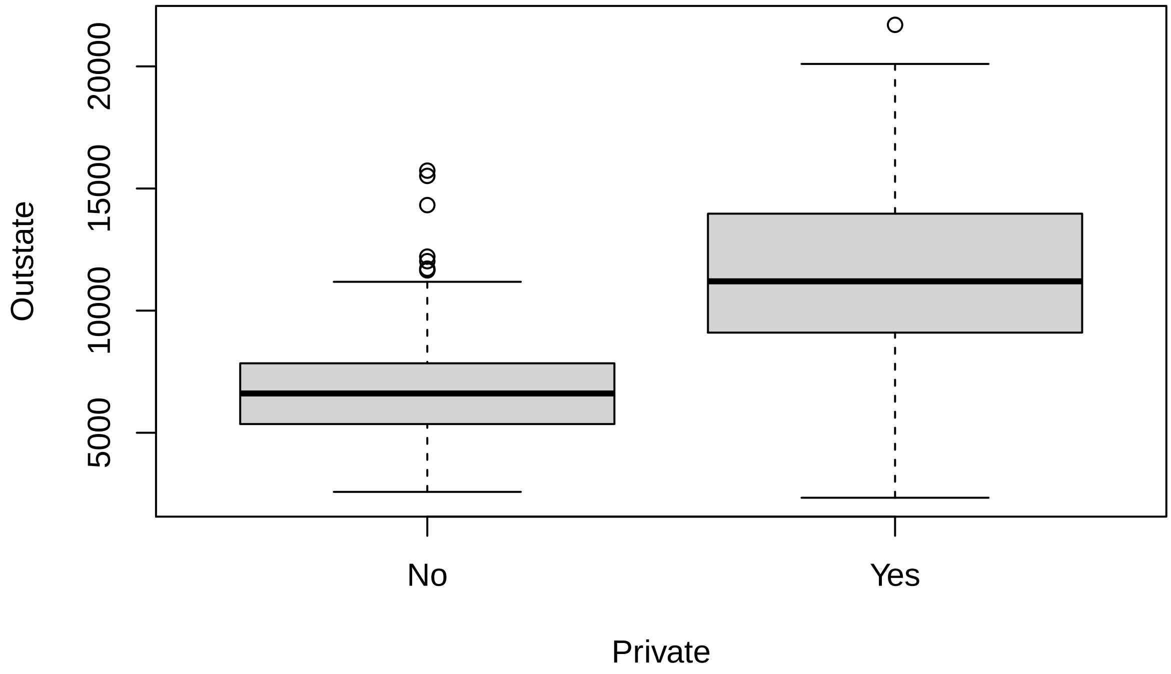
ii)

```
pairs(college[,2:11])
```



iii)

```r
boxplot(Outstate~Private, data = college)
```



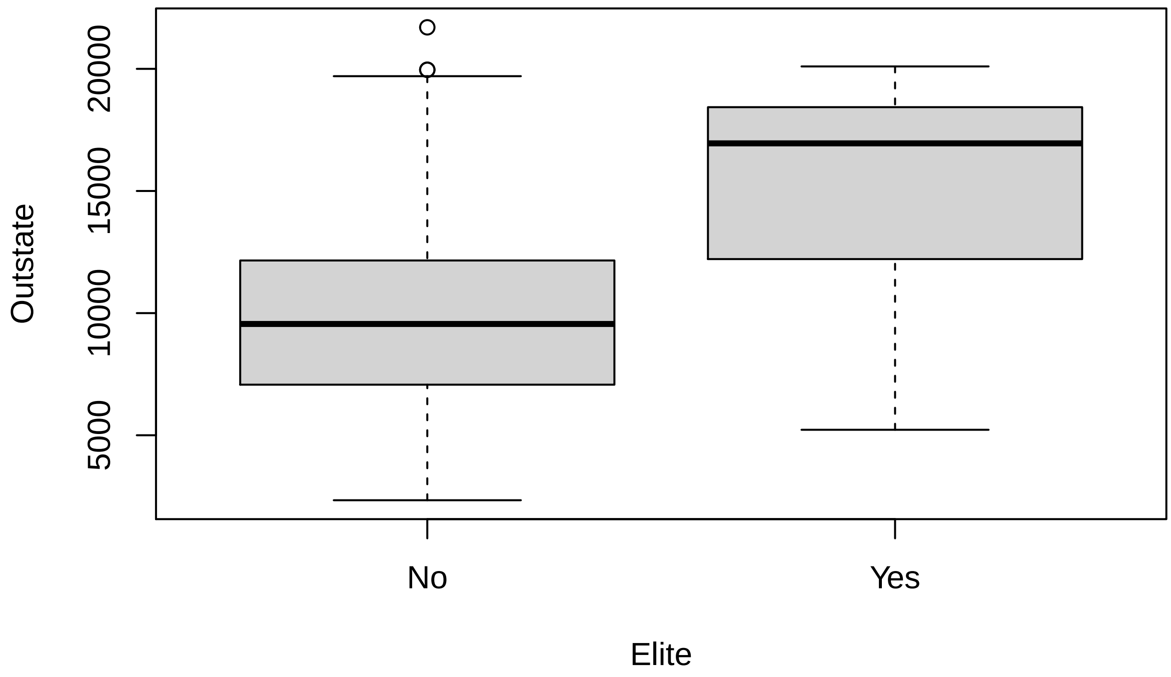
iv)

```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] = "Yes"
Elite = as.factor(Elite)
college = data.frame(college, Elite)

summary(college)
```

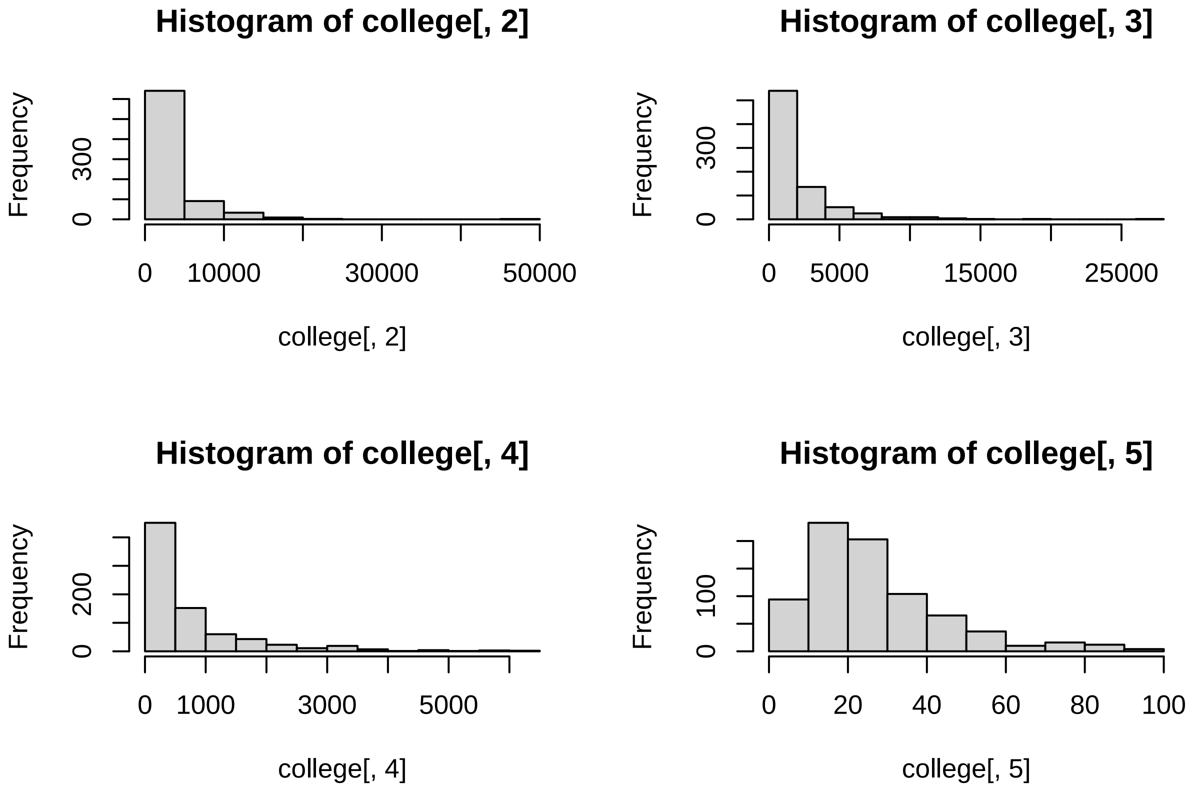
```
##    Private               Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 3002   Mean   : 2019   Mean   : 780
##                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                     Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board        Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341  
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700  
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800  
##       PhD            Terminal       S.F.Ratio      perc.alumni 
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00  
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00  
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00  
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74  
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00  
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00  
##      Expend        Grad.Rate      Elite    
##  Min.   : 3186   Min.   : 10.00   No :699  
##  1st Qu.: 6751   1st Qu.: 53.00   Yes: 78  
##  Median : 8377   Median : 65.00            
##  Mean   : 9660   Mean   : 65.46            
##  3rd Qu.:10830   3rd Qu.: 78.00            
##  Max.   :56233   Max.   :118.00            
```

```
plot(Outstate~Elite, data = college)
```



v)

```
par(mfrow= c(2,2))
hist(college[, 2])
hist(college[, 3])
hist(college[, 4])
hist(college[, 5])
```



vi)

```
#fix(college)

private = subset(college, college$Private == "Yes")
public = subset(college, college$Private == "No")

summary(private$Top10perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   17.00   25.00   29.33   36.00   96.00
```

```
summary(private$Top25perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   42.00   55.00   56.96   70.00  100.00
```

```
summary(public$Top10perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   12.00   19.00   22.83   27.50   95.00
```

```
summary(public$Top25perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    12.0    37.0    51.0    52.7    65.0   100.0
```

From the summary of the Top10perc and Top25perc data, we find that private and public institutions have a similar amount of new students from the 10 and 25 percent of thier high school classes.

5.

a) They are all numerical values except for name:

```
auto <- read.csv("Auto.csv",header=T,na.strings ="?")
auto <- na.omit(auto)
summary(auto)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration        year           origin          name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

so, everything is quantitative except for name which is qualitative. b) From the summary, we can find the range as it gives a min and max of each quantitative column.

```
summary(auto)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration        year           origin          name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :24.80   Max.   :82.00   Max.   :3.000
```

c) We can see the mean from the summary, but not the standard deviation. We can find the standard deviation from the sd command and apply it to each column:

```
summary(auto)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Median :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mean   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration        year           origin          name
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   Length:392
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   Class :character
##  Median :15.50   Median :76.00   Median :1.000   Mode  :character
##  Mean   :15.54   Mean   :75.98   Mean   :1.577  
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000  
##  Max.   :24.80   Max.   :82.00   Max.   :3.000  
```

```
stddev <- apply(auto[, 1:8], 2, sd)
stddev
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##    7.8050075    1.7057832  104.6440039   38.4911599  849.4025600    2.7588641 
##         year       origin 
##    3.6837365    0.8055182 
```

d) We can see the mean from the summary, but not the standard deviation. We can find the standard deviation from the sd command and apply it to each column:

```
auto1 <- auto[-c(10:85), ]
summary(auto1)
```

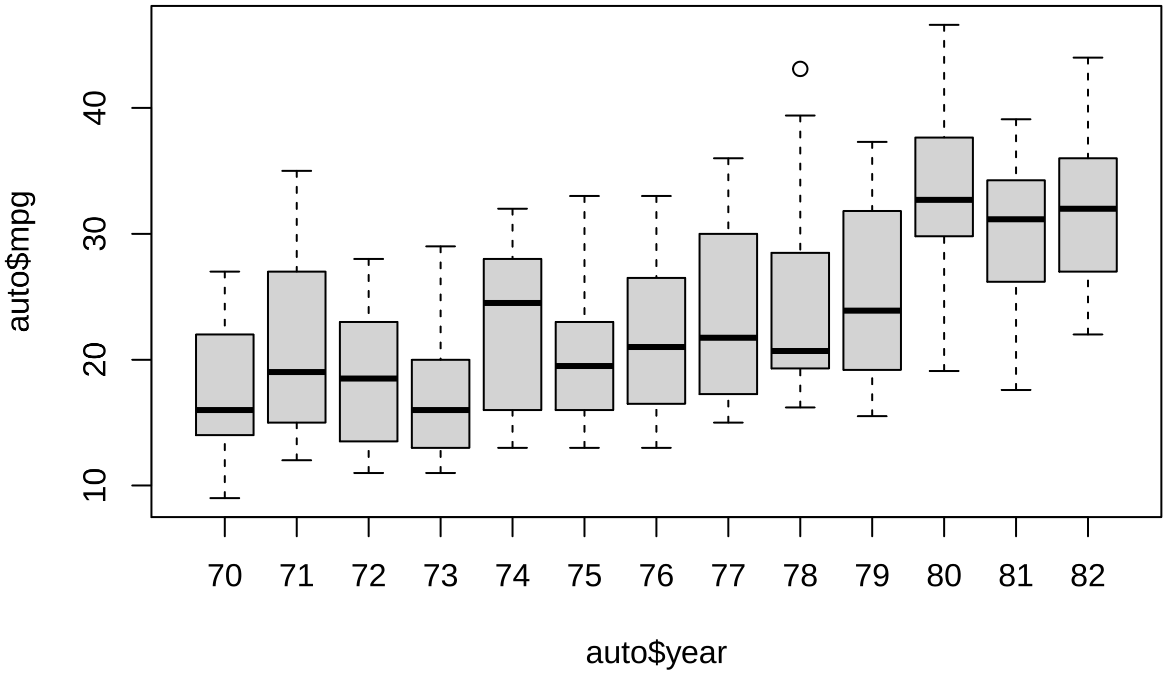
```
##       mpg          cylinders      displacement     horsepower        weight    
##  Min.   :11.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.   :1649  
##  1st Qu.:18.00   1st Qu.:4.000   1st Qu.:100.2   1st Qu.: 75.0   1st Qu.:2214  
##  Median :23.95   Median :4.000   Median :145.5   Median : 90.0   Median :2792  
##  Mean   :24.40   Mean   :5.373   Mean   :187.2   Mean   :100.7   Mean   :2936  
##  3rd Qu.:30.55   3rd Qu.:6.000   3rd Qu.:250.0   3rd Qu.:115.0   3rd Qu.:3508  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.   :4997  
##   acceleration        year           origin          name          
##  Min.   : 8.50   Min.   :70.00   Min.   :1.000   Length:316        
##  1st Qu.:14.00   1st Qu.:75.00   1st Qu.:1.000   Class :character  
##  Median :15.50   Median :77.00   Median :1.000   Mode  :character  
##  Mean   :15.73   Mean   :77.15   Mean   :1.601                     
##  3rd Qu.:17.30   3rd Qu.:80.00   3rd Qu.:2.000                     
##  Max.   :24.80   Max.   :82.00   Max.   :3.000                     
```

```
stddev1 <- apply(auto1[, 1:8], 2, sd)
stddev1
```

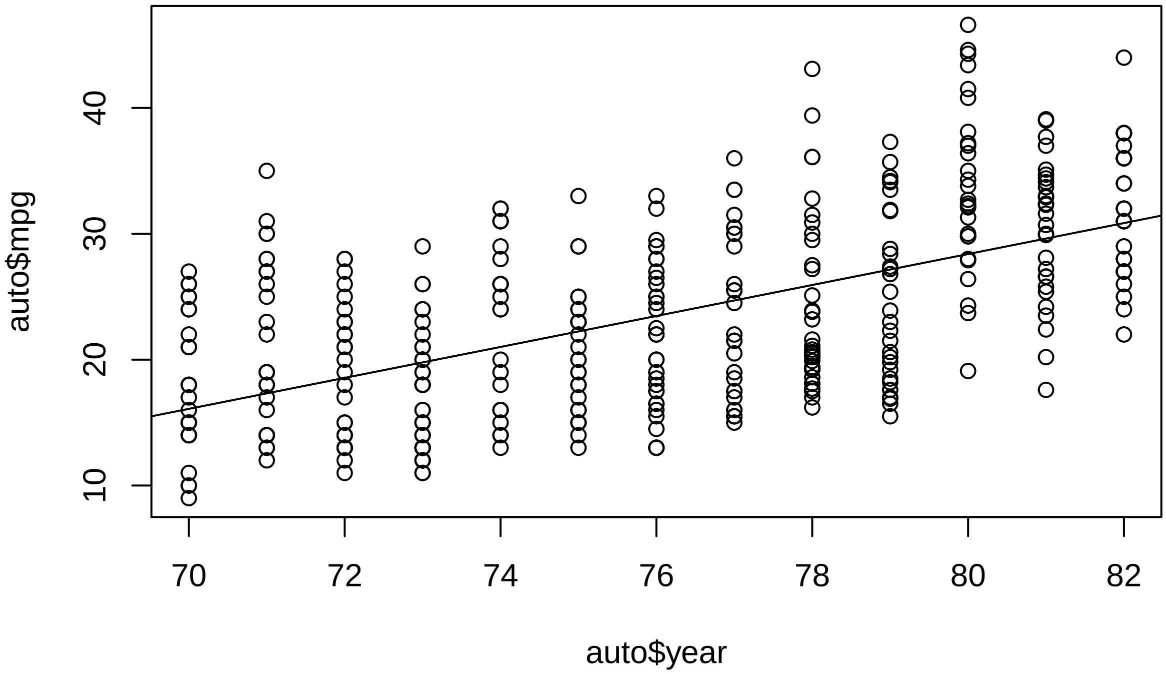
```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721 
##         year       origin 
##     3.106217     0.819910 
```

e) Using a boxplot, we find for the most part, that the average mpg for the cars in our dataset goes up the later it came out. We can also fit a linear model to the scatter plot of the mpg vs year in order to see this correlation.

```
boxplot(auto$mpg~auto$year)
```

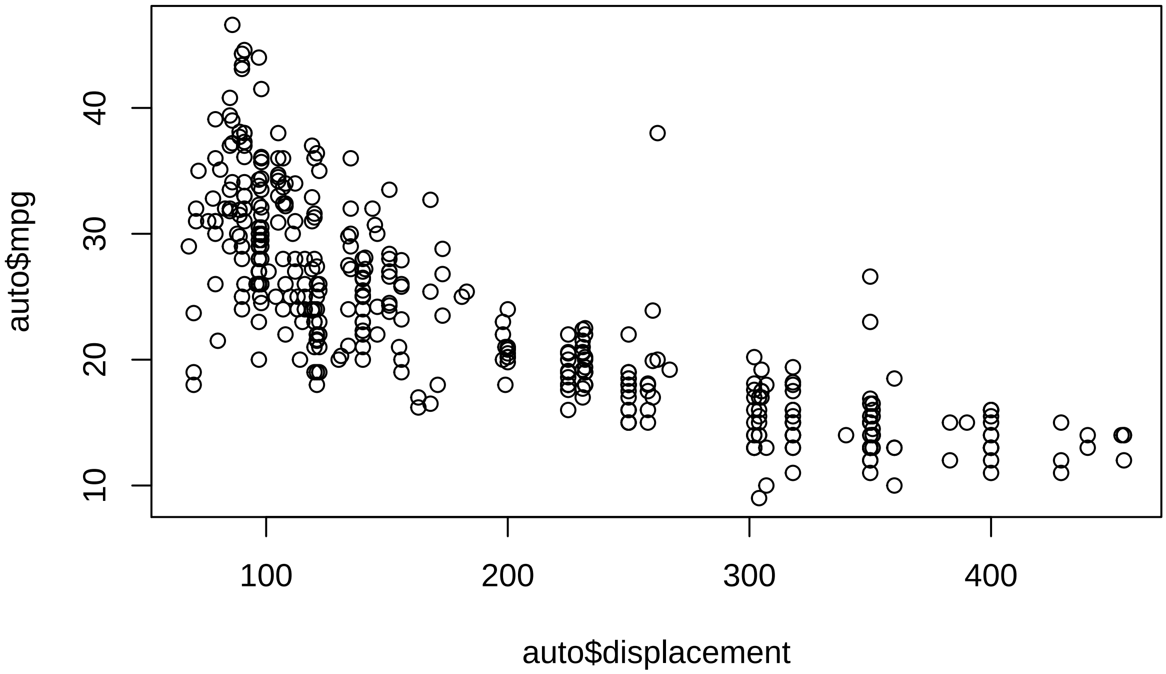


```
fit <- lm(auto$mpg~auto$year)
plot(auto$mpg~auto$year)
abline(fit)
```

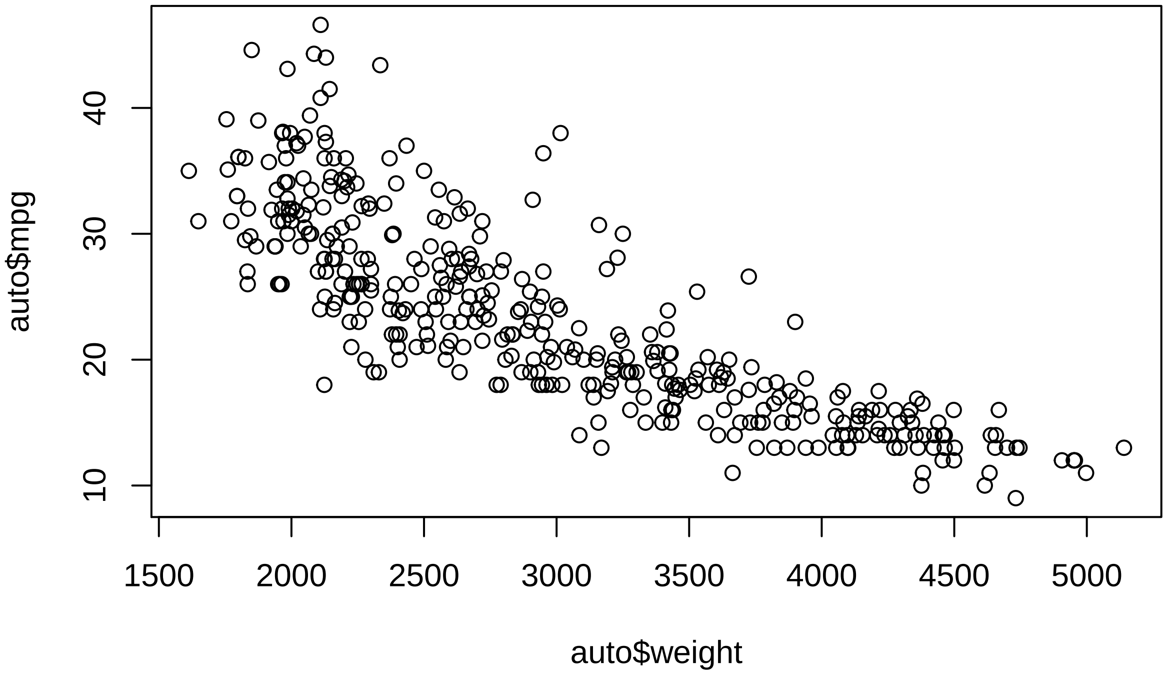


f) From the above, we find that there is a correlation between the year and mpg, but other factors also seem to correlate as well.

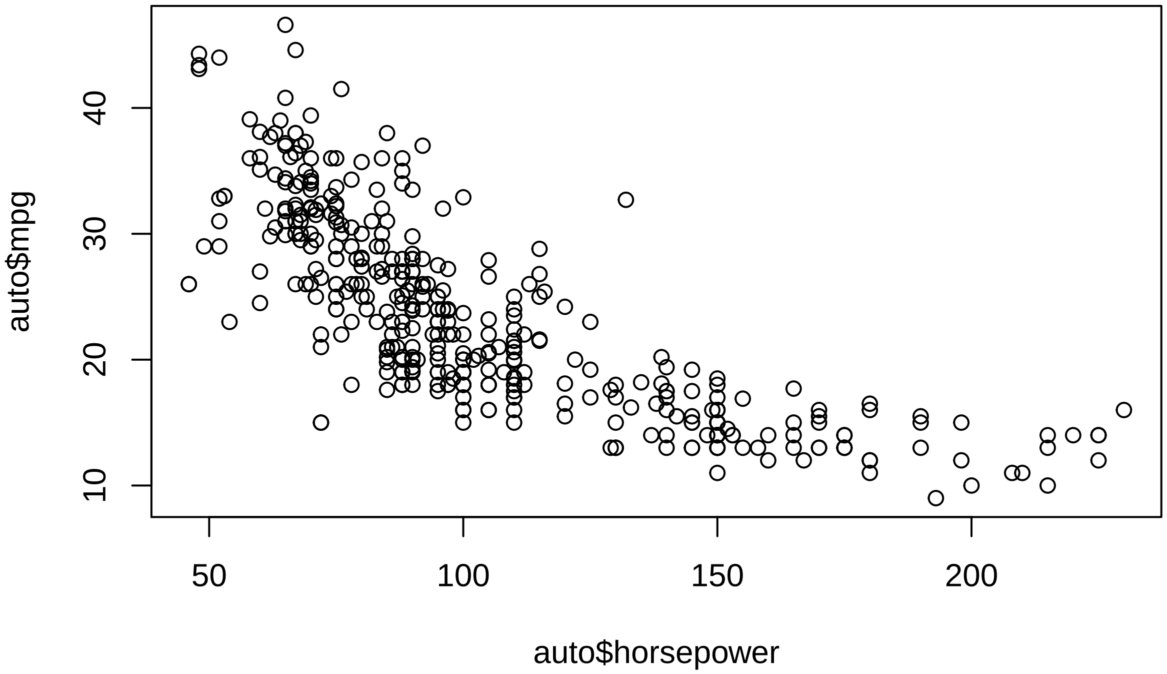
```
plot(auto$mpg~auto$displacement)
```



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



```
plot(auto$mpg~auto$horsepower)
```



From the above 3 graphs, there is a general trend downward in mpg with cars with higher horsepower, displacement and weight.