Assignment 2

Chiu-Yuan Wu

9/8/2020

1.

(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("College.csv")
```

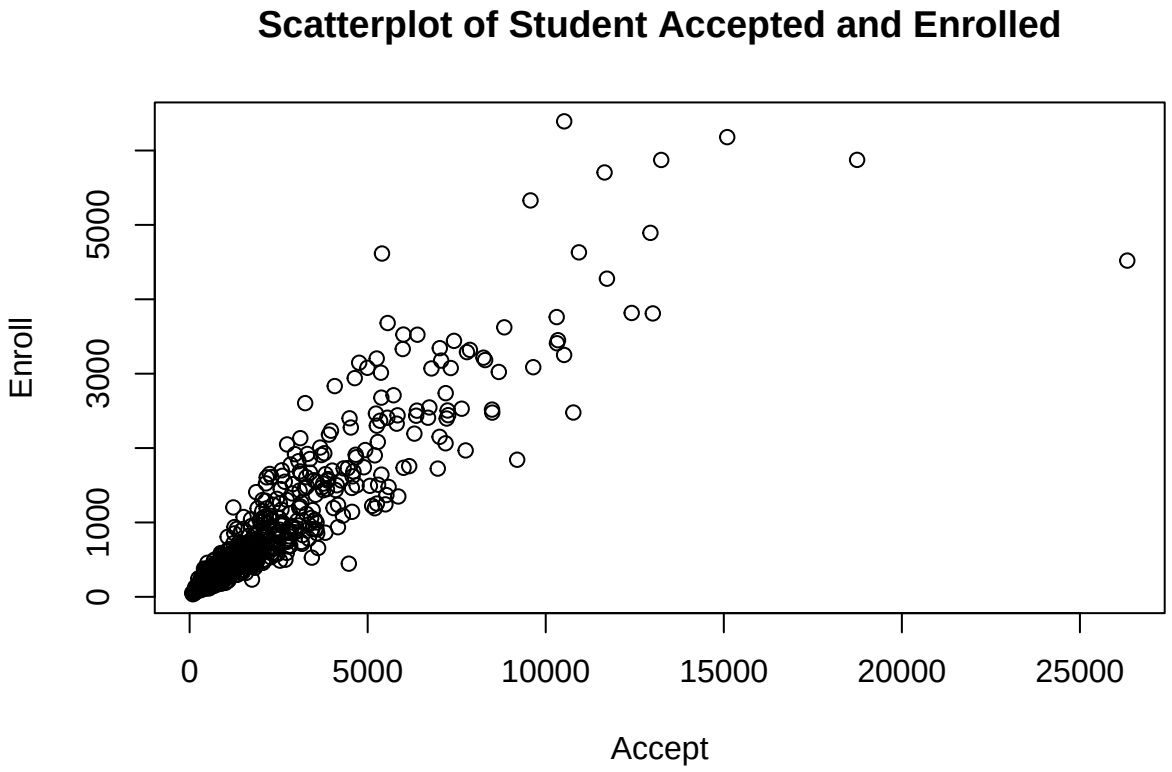
(b) Find the median cost of books for all schools in this dataset.

```
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.

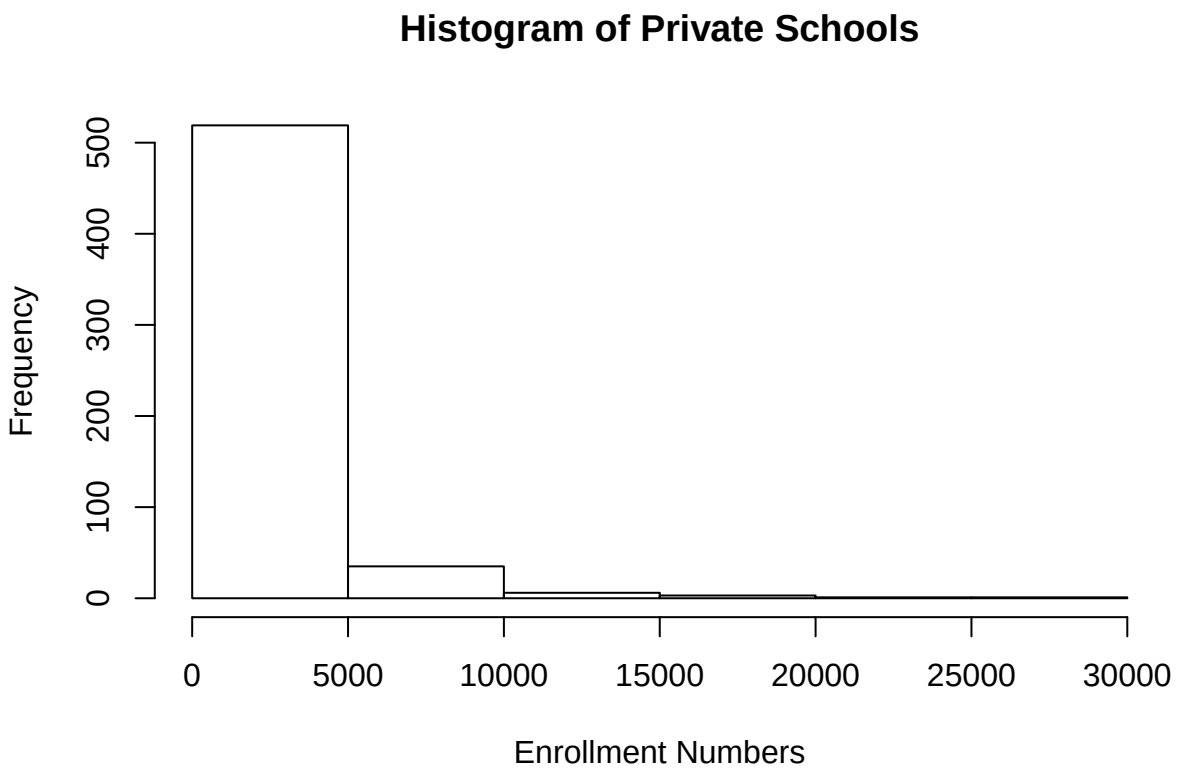
```
plot(college$Accept, college$Enroll, xlab = "Accept", ylab = "Enroll",
     main = "Scatterplot of Student Accepted and Enrolled")
```



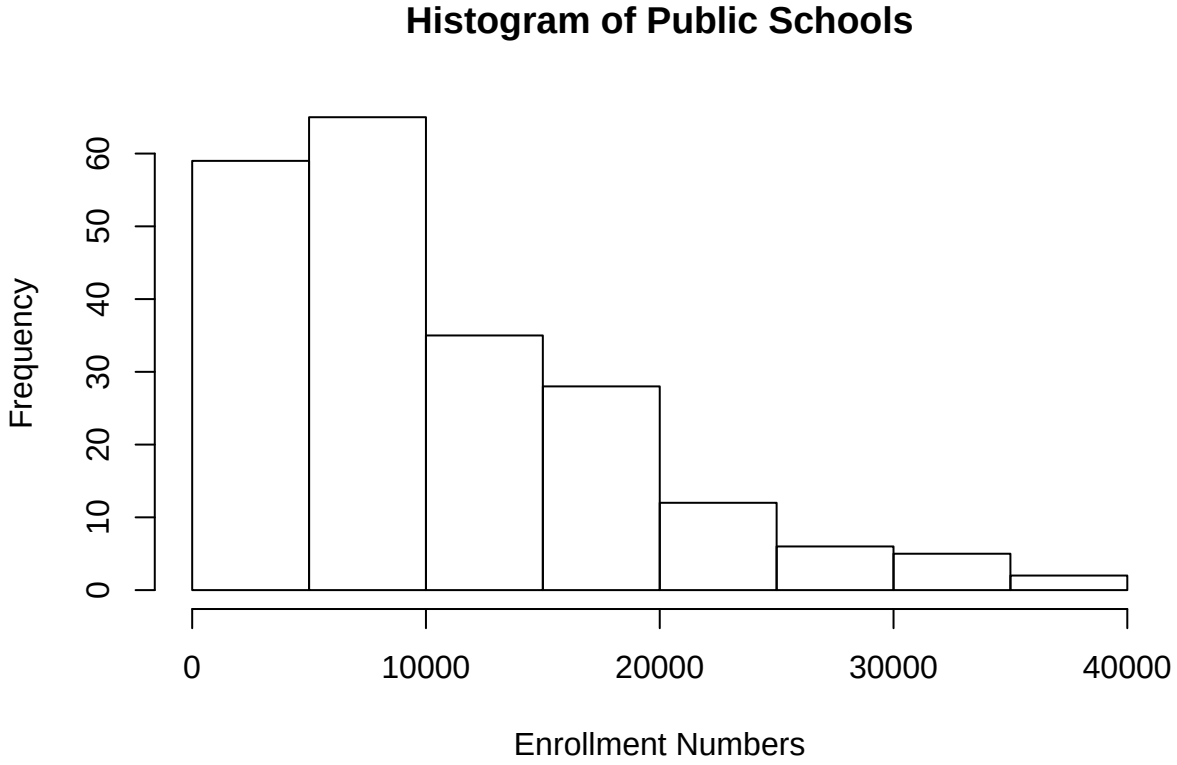
(d) Produce a histogram showing the overall enrollment numbers (P.Undergrad plus F.Undergrad) for both public and private (Private) schools. You may choose to show both on a single plot (using side by side bars) or produce one plot for public schools and one for private schools. Ensure whatever figures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

```
Private_schools <- college[which(college$Private=="Yes"),]
Public_schools <- college[which(college$Private=="No"),]

hist(Private_schools$P.Undergrad + Private_schools$F.Undergrad,
     breaks = 5, xlab = "Enrollment Numbers", main = "Histogram of Private Schools")
```




```r
hist(Public_schools$P.Undergrad + Public_schools$F.Undergrad,
     xlab = "Enrollment Numbers", main = "Histogram of Public Schools")
```



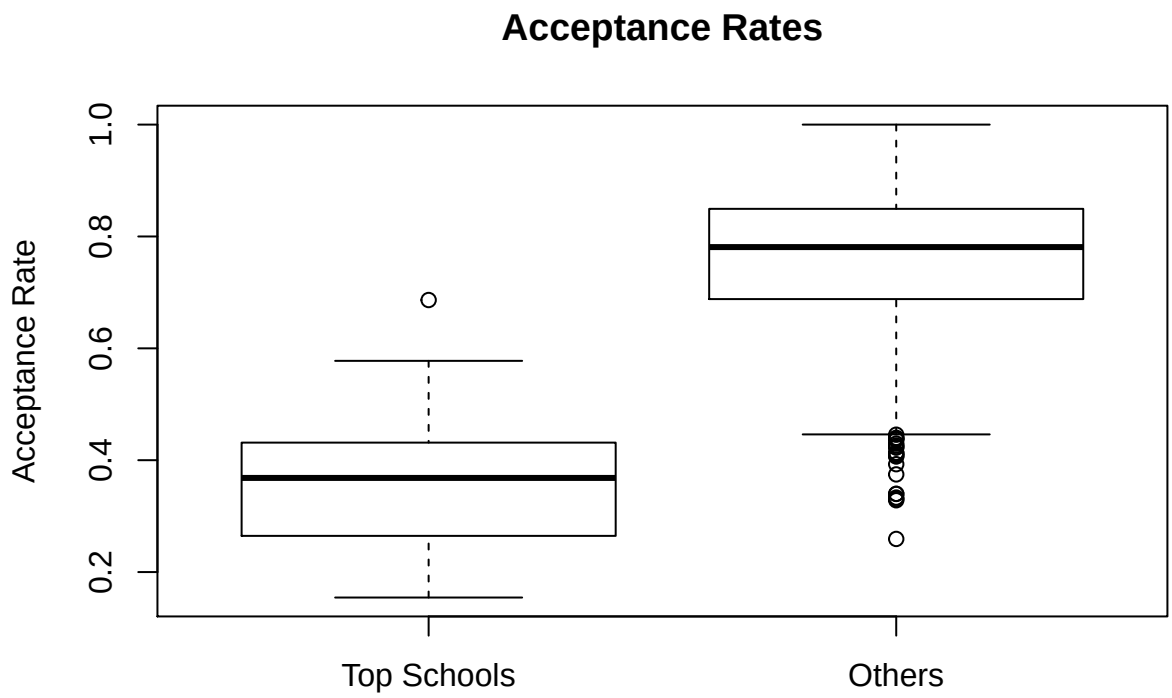
(e) Create a new qualitative variable, called Top, by binning the Top10perc variable into two categories

(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?

```
college$Top <- ifelse(college$Top10perc > 75, "Yes", "No")

Top_schools <- college[which(college$Top=="Yes"),]
Nottop_schools <- college[which(college$Top=="No"),]

boxplot(Top_schools$Accept / Top_schools$Apps,
        Nottop_schools$Accept / Nottop_schools$Apps, main = "Acceptance Rates",
        names = c("Top Schools", "Others"), ylab = "Acceptance Rate")
```



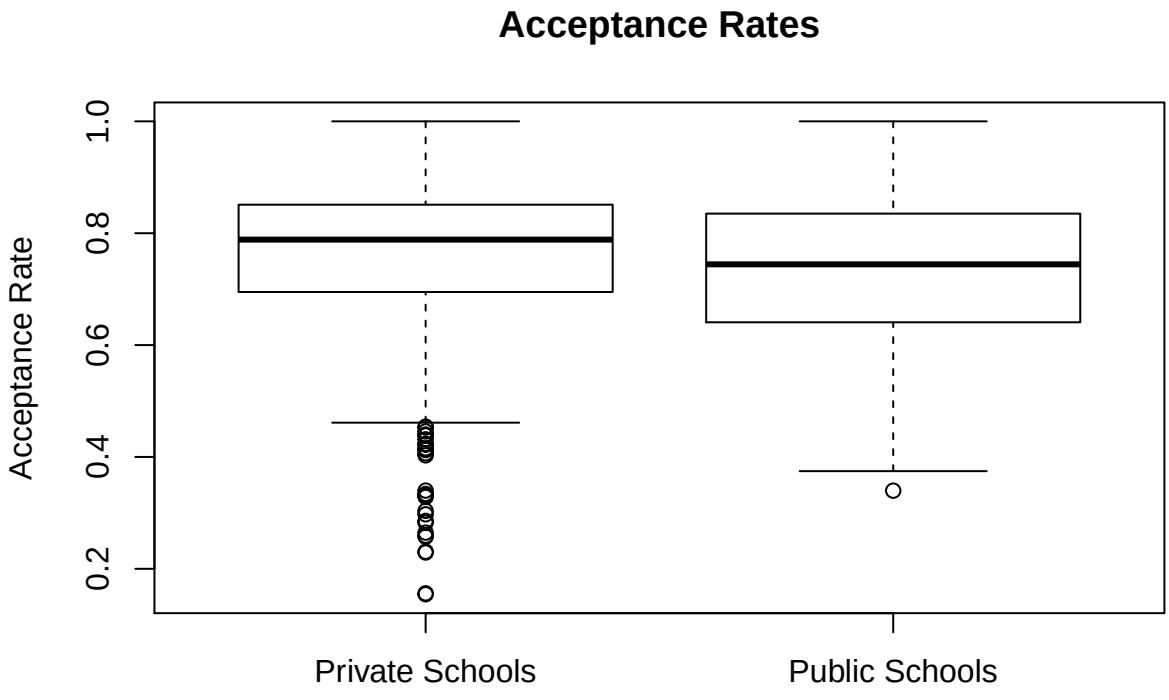

```
length(Top_schools$X)
```

```
## [1] 22
```

There are 22 top universities.

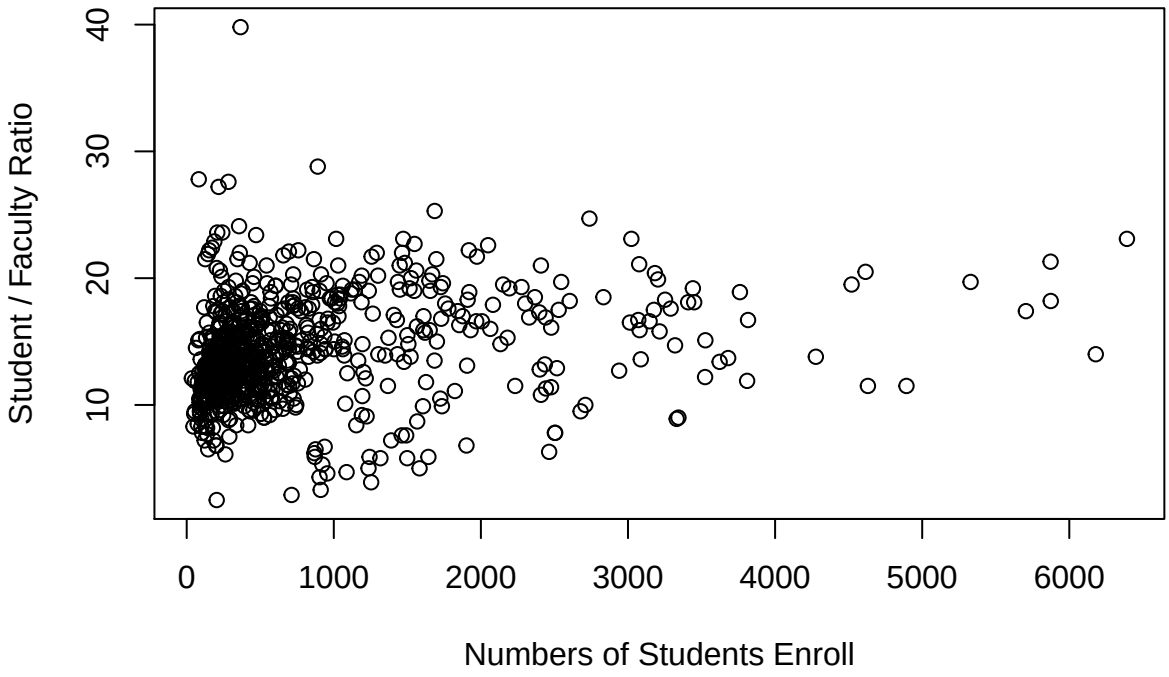
(f) Continue exploring the data, producing two new plots of any type, and provide a brief (one to two sentence) summary of your hypotheses and what you discover. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics 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.

```
boxplot(Private_schools$Accept / Private_schools$Apps,
        Public_schools$Accept / Public_schools$Apps, main = "Acceptance Rates",
        names = c("Private Schools", "Public Schools"), ylab = "Acceptance Rate")
```



Acceptance rate of private schools is higher than acceptance rate of public schools, but just a little bit.

```
plot(college$Enroll, college$S.F.Ratio,
     xlab = "Numbers of Students Enroll", ylab = "Student / Faculty Ratio")
```



There seems no relation between S.F. Ratio and the numbers of students enrolled, most of the school has a ratio in between 10-20%.

2.

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("Auto.csv")

horsepower <- as.double(auto$horsepower)
horsepower[is.na(horsepower)] <- mean(horsepower, na.rm = TRUE)
auto$horsepower <- horsepower
```

(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.
Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration.
Qualitative: year, origin, name

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

```
range(auto$mpg)
```

```
## [1]  9.0 46.6
```

```
range(auto$cylinders)
```

```
## [1] 3 8
```

```
range(auto$displacement)
```

```
## [1]  68 455
```

```
range(auto$horsepower)
```

```
## [1]  1 94
```

```
range(auto$weight)
```

```
## [1] 1613 5140
```

```
range(auto$acceleration)
```

```
## [1]  8.0 24.8
```

Mean:

```r
mean(auto$mpg)
```

```
## [1] 23.51587
```

```r
mean(auto$cylinders)
```

```
## [1] 5.458438
```

```r
mean(auto$displacement)
```

```
## [1] 193.5327
```

```r
mean(auto$horsepower)
```

```
## [1] 51.51637
```

```r
mean(auto$weight)
```

```
## [1] 2970.262
```

```r
mean(auto$acceleration)
```

```
## [1] 15.55567
```

Standard deviation:

```r
sd(auto$mpg)
```

```
## [1] 7.825804
```

```r
sd(auto$cylinders)
```

```
## [1] 1.701577
```

```r
sd(auto$displacement)
```

```
## [1] 104.3796
```

```r
sd(auto$horsepower)
```

```
## [1] 29.8627
```

```r
sd(auto$weight)
```

```
## [1] 847.9041
```

```r
sd(auto$acceleration)
```

```
## [1] 2.749995
```

(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?

```r
temp_auto <- auto[-c(40:80),-10]
```

Range:

```r
range(temp_auto$mpg)
```

```
## [1]  9.0 46.6
```

```r
range(temp_auto$cylinders)
```

```
## [1] 3 8
```

```r
range(temp_auto$displacement)
```

```
## [1]  68 455
```

```r
range(temp_auto$horsepower)
```

```
## [1]  1 94
```

```r
range(temp_auto$weight)
```

```
## [1] 1649 4997
```

```r
range(temp_auto$acceleration)
```

```
## [1]  8.0 24.8
```

Mean:

```r
mean(temp_auto$mpg)
```

```
## [1] 24.02472
```

```r
mean(temp_auto$cylinders)
```

```
## [1] 5.398876
```

```r
mean(temp_auto$displacement)
```

```
## [1] 189.2303
```

```r
mean(temp_auto$horsepower)
```

```
## [1] 51.66854
```

```r
mean(temp_auto$weight)
```

```
## [1] 2935.36
```

```r
mean(temp_auto$acceleration)
```

```
## [1] 15.60983
```

Standard deviation:

```r
sd(temp_auto$mpg)
```

```
## [1] 7.827748
```

```r
sd(temp_auto$cylinders)
```

```
## [1] 1.659254
```

```r
sd(temp_auto$displacement)
```

```
## [1] 100.8794
```

```r
sd(temp_auto$horsepower)
```

```
## [1] 30.36097
```

```r
sd(temp_auto$weight)
```

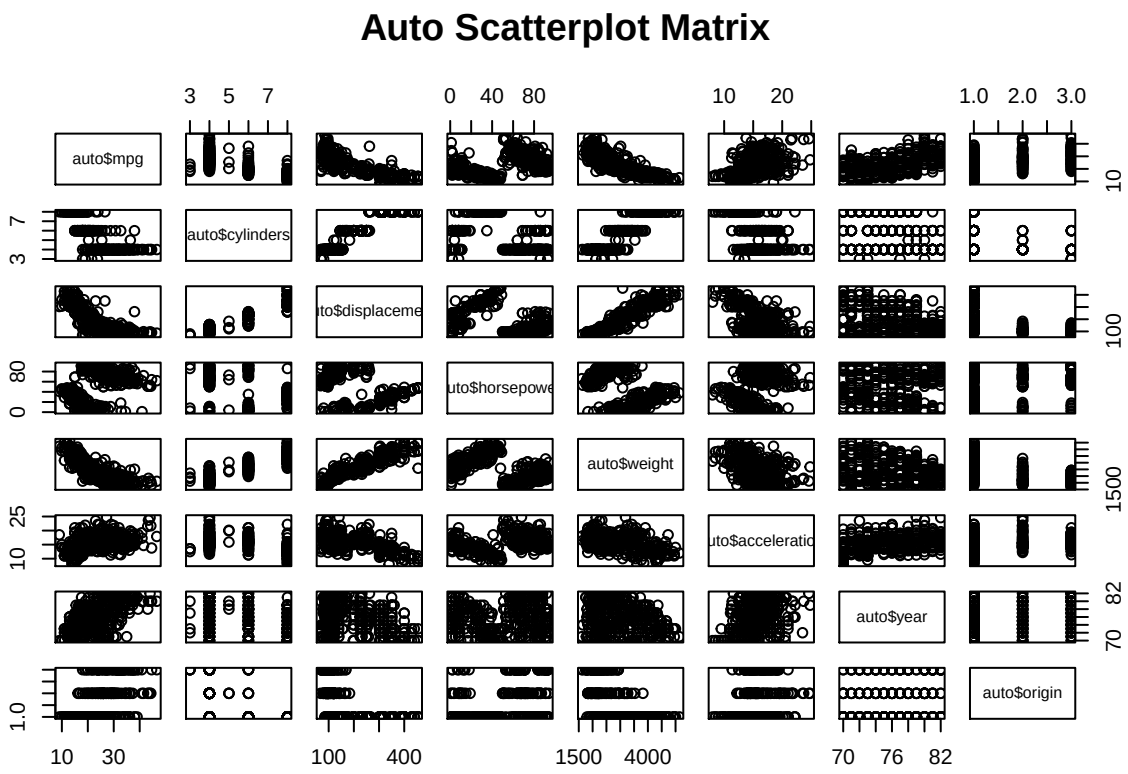
```
## [1] 810.8406
```

```r
sd(temp_auto$acceleration)
```

```
## [1] 2.712348
```

(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.

```
pairs(~auto$mpg+auto$cylinders+auto$displacement+auto$horsepower+
          auto$weight+auto$acceleration+auto$year+auto$origin,
      main="Auto Scatterplot Matrix")
```




```
auto_cor <- auto[c(1:7)]
res <- cor(auto_cor)
round(res, 2)
```

```
##                mpg cylinders displacement horsepower weight acceleration  year
## mpg           1.00     -0.78        -0.80       0.42  -0.83         0.42  0.58
## cylinders    -0.78      1.00         0.95      -0.55   0.90        -0.50 -0.35
## displacement -0.80      0.95         1.00      -0.48   0.93        -0.54 -0.37
## horsepower    0.42     -0.55        -0.48       1.00  -0.48         0.27  0.13
## weight       -0.83      0.90         0.93      -0.48   1.00        -0.42 -0.31
## acceleration  0.42     -0.50        -0.54       0.27  -0.42         1.00  0.28
## year          0.58     -0.35        -0.37       0.13  -0.31         0.28  1.00
```

(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.
Weight and Displacement, because the correlations are higher.