CptS 575 Data Science: Assignment 2

Md Muhtasim Billah

9/11/2020

Qestion 1

1 (a)

Reading in the data in R as a data frame.

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

(b)

The median cost of books for all schools in this dataset.

```
median(college$Books)
```

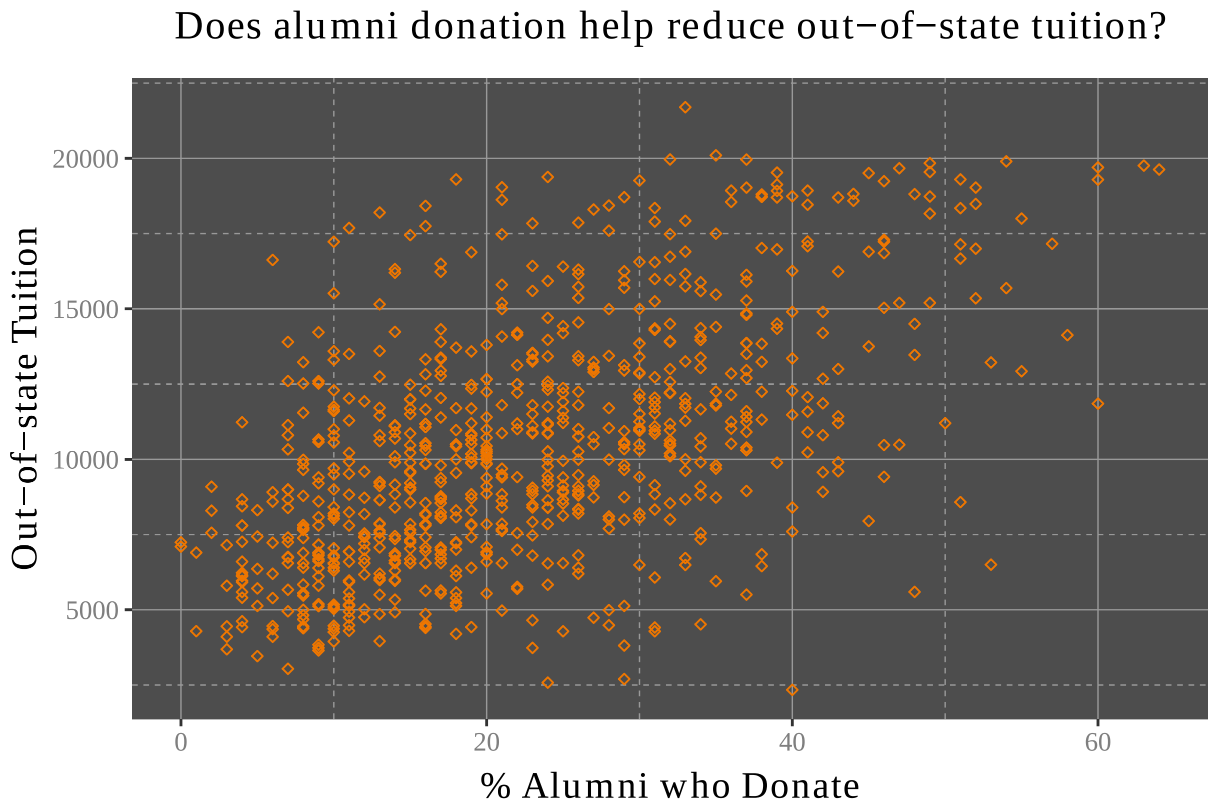
```
## [1] 500
```

(c)

Scatter plot of Outsate vs perc.alumni. Surprisingly, the out-of-state tuition shows a linearly increasing relationship with the percentage of alumni of the school who donate.

```
library(ggplot2)
plot1 = ggplot(data=college, aes(x=perc.alumni, y=Outstate)) +
  geom_point(color="darkorange2", shape=5, size=1) +
  ggtitle("Does alumni donation help reduce out-of-state tuition?") + ylab("Out-of-state Tuition") + xla
  theme(
    panel.background = element_rect(fill = "grey30"),
    panel.grid.major = element_line(colour = "grey60", size=0.25),
    panel.grid.minor = element_line(colour = "grey60", linetype = "dashed"),
    plot.title = element_text(size=15, hjust=0.5, vjust = 3.5, family = "Palatino", colour = "Black",
                              margin = margin(10, 0, 0, 0)),
    axis.title.x = element_text(size=14, vjust = -0.3, family = "Palatino", colour = "Black",
                                margin = margin(0, 0, 20, 0)),
    axis.text.x = element_text(size = 10, family = "Palatino", colour = "grey50"),
    axis.title.y = element_text(size=14, vjust = 1.5, family = "Palatino", colour = "Black",
                                margin = margin(10, 0, 10, 10)),
    axis.text.y = element_text(size = 10, family = "Palatino", colour = "grey50"),
  )

plot1
```



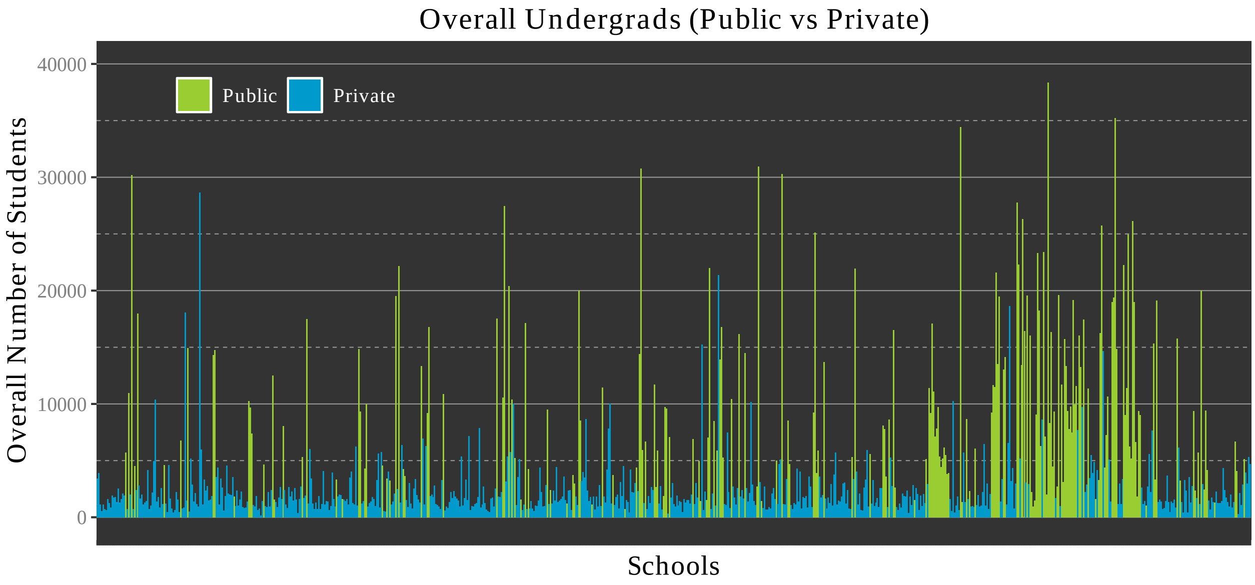
(d)

The histogram showing the overall enrollment numbers (P.Undergrad plus F.Undergrad) for both public and private (Private) schools.

```
#create column for overall undergrads
college$O.undergrad = college$P.Undergrad + college$F.Undergrad
#head(college)

# Use position=position_dodge()
plot2 = ggplot(data=college, aes(x=X, y=O.undergrad, fill=Private)) +
  geom_bar(stat="identity", position=position_dodge()) +
  #geom_text(aes(label=X), position = position_dodge(0.7), vjust=-0.3, size=3.5, color = "White", famil
  coord_cartesian(ylim=c(0,40000)) + scale_fill_manual(labels = c("Public","Private"),values = c("olive
  ggtitle("Overall Undergrads (Public vs Private)") + ylab("Overall Number of Students") +   xlab("Scho
  theme(
    #plot.background = element_rect(fill = "grey20"),
    panel.background = element_rect(fill = "grey20"),
    panel.grid.major = element_line(colour = "grey60", size=0.25),
    panel.grid.minor = element_line(colour = "grey60", linetype = "dashed"),
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank(),
    plot.title = element_text(size=15, hjust=0.5, vjust = 1.5, family = "Palatino", colour = "Black", ma
    axis.title.x = element_text(size=14, vjust = -0.3, family = "Palatino", colour = "Black", margin = m
    axis.text.x = element_blank(),
    axis.title.y = element_text(size=14, vjust = 1.5, family = "Palatino", colour = "Black", margin = ma
    axis.text.y = element_text(size = 10, family = "Palatino", colour = "grey50"),
    # Change legend background color
    legend.background = element_rect(fill = "transparent"),
    #legend.title= element_text(color = "White",family = "Palatino", size=10),
    legend.title=element_blank(),
    #legend.key = element_rect(fill = "lightblue", color = NA),
    legend.text = element_text(color = "White",family = "Palatino", size=10),
    # Change legend key size and key width
    #legend.key.size = unit(1.5, "cm"),
    #legend.key.width = unit(0.5,"cm")
    legend.position = c(.05, .95),
    legend.justification = c("left", "top"),
    legend.box.just = "left",
    legend.margin = margin(6, 6, 6, 6),
    legend.direction = "horizontal"
  )

plot2
```



(e)

```
Top = ifelse(college$Top10perc >75, "Yes", "No")
table(Top)
```

```
## Top
##  No Yes
## 755  22
```

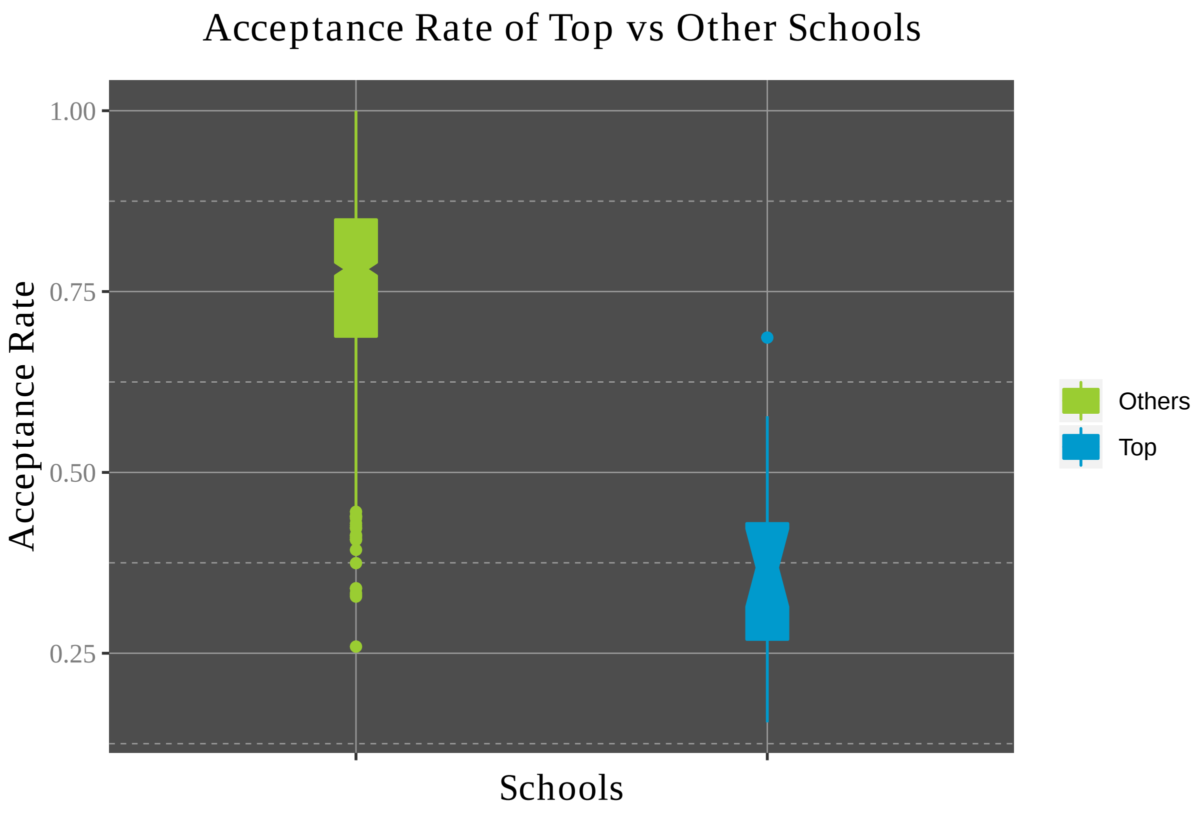
So, there are 22 top universities.

The side-by-side boxplots of the schools’ acceptance rates.

```r
college$Acceptance.Rate = college$Accept / college$Apps

plot3 = ggplot(data=college, aes(x=Top, y=Acceptance.Rate, color=Top, fill=Top)) +
  geom_boxplot(notch=TRUE, width=0.1) +
  scale_color_manual(labels = c("Others","Top"), values = c("olivedrab3", "deepskyblue3")) +
  scale_fill_manual(labels = c("Others","Top"), values = c("olivedrab3", "deepskyblue3")) +
  ggtitle("Acceptance Rate of Top vs Other Schools") + ylab("Acceptance Rate") + xlab("Schools") +
  theme(
    panel.background = element_rect(fill = "grey30"),
    panel.grid.major = element_line(colour = "grey60", size=0.25),
    panel.grid.minor = element_line(colour = "grey60", linetype = "dashed"),
    plot.title = element_text(size=15, hjust=0.5, vjust = 3.5, family = "Palatino", colour = "Black",
                              margin = margin(10, 0, 0, 0)),
    axis.title.x = element_text(size=14, vjust = -0.3, family = "Palatino", colour = "Black",
                                margin = margin(0, 0, 20, 0)),
    axis.text.x = element_blank(),
    axis.title.y = element_text(size=14, vjust = 1.5, family = "Palatino", colour = "Black",
                                margin = margin(10, 0, 10, 10)),
    axis.text.y = element_text(size = 10, family = "Palatino", colour = "grey50"),
    legend.title = element_blank(),
    legend.position="right",
  )

plot3
```



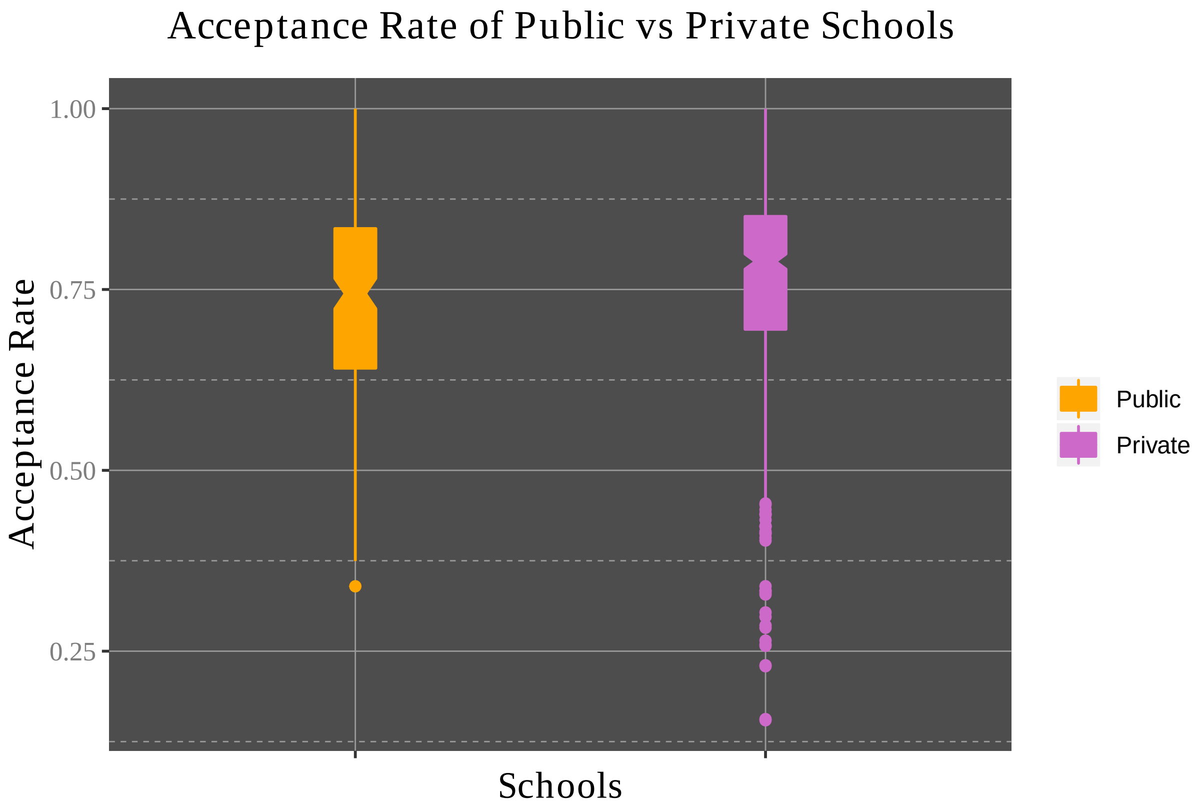
(f)

First Plot:

The bar plot below demonstrates the comparison of the acceptance rates between the public and private schools. It is apparent from the plot that the private schools have a slightly higher acceptane rate. However, for private schools, the value is verh widely spread indicating to a higher standard deviation, unlike the public schools.

```r
plot4 = ggplot(data=college, aes(x=Private, y=Acceptance.Rate, color=Private, fill=Private)) +
  geom_boxplot(notch=TRUE, width=0.1) +
  scale_color_manual(labels = c("Public", "Private"), values = c("orange", "orchid3")) +
  scale_fill_manual(labels = c("Public", "Private"), values = c("orange", "orchid3")) +
  ggtitle("Acceptance Rate of Public vs Private Schools") + ylab("Acceptance Rate") + xlab("Schools") +
  theme(
    panel.background = element_rect(fill = "grey30"),
    panel.grid.major = element_line(colour = "grey60", size=0.25),
    panel.grid.minor = element_line(colour = "grey60", linetype = "dashed"),
    plot.title = element_text(size=15, hjust=0.5, vjust = 3.5, family = "Palatino", colour = "Black",
                              margin = margin(10, 0, 0, 0)),
    axis.title.x = element_text(size=14, vjust = -0.3, family = "Palatino", colour = "Black",
                                margin = margin(0, 0, 20, 0)),
    axis.text.x = element_blank(),
    axis.title.y = element_text(size=14, vjust = 1.5, family = "Palatino", colour = "Black",
                                margin = margin(10, 0, 10, 10)),
    axis.text.y = element_text(size = 10, family = "Palatino", colour = "grey50"),
    legend.title = element_blank(),
    legend.position="right",
  )

plot4
```

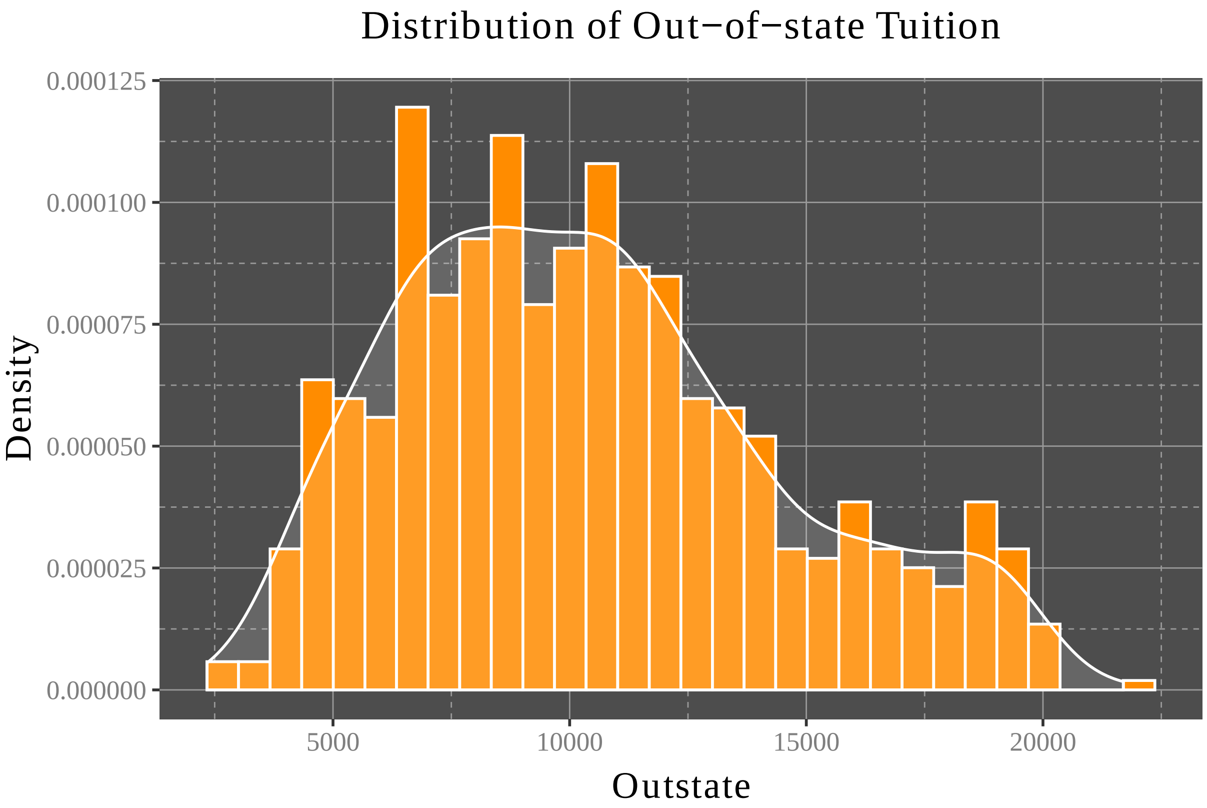


Second Plot:

The following plot shows the Kernel density of the variable Outstate which is the out-of-state tuition for the schools in the dataset. The distribution seems to be a bit skewed to the left.

```
# Histogram overlaid with kernel density curve
plot5 =ggplot(college, aes(x=Outstate)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead of count on y-axis

                   colour="white", fill="darkorange") +
  geom_density(alpha = .15, color = "white", fill="white") +
    ggtitle("Distribution of Out-of-state Tuition") + ylab("Density") + xlab("Outstate") +
  theme(
    panel.background = element_rect(fill = "grey30"),
    panel.grid.major = element_line(colour = "grey60", size=0.25),
    panel.grid.minor = element_line(colour = "grey60", linetype = "dashed"),
    plot.title = element_text(size=15, hjust=0.5, vjust = 3.5, family = "Palatino", colour = "Black",
                              margin = margin(10, 0, 0, 0)),
    axis.title.x = element_text(size=14, vjust = -0.3, family = "Palatino", colour = "Black",
                                margin = margin(0, 0, 20, 0)),
    axis.text.x = element_text(size = 10, family = "Palatino", colour = "grey50"),
    axis.title.y = element_text(size=14, vjust = 1.5, family = "Palatino", colour = "Black",
                                margin = margin(10, 0, 10, 10)),
    axis.text.y = element_text(size = 10, family = "Palatino", colour = "grey50"),
  )
plot5
```


Distribution of Out−of−state Tuition
Density
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
5000
10000
15000
20000
Outstate

Question 2

```
autodata = read.csv(url("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/Auto.csv"),
                   header = TRUE)
attach(autodata)
#summary(autodata)
```

No missing values were found in the dataset.

2 (a)

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

2 (b)

The range, mean and standard deviation of each quantitative predictor is given below.

```
horsepower = as.numeric(autodata$horsepower)
quant.var = data.frame(mpg,cylinders,displacement,horsepower,weight,acceleration)
summary(quant.var)
```

```
##       mpg          cylinders      displacement     horsepower        weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 1.00   Min.   :1613
##  1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.0   1st Qu.:26.00   1st Qu.:2223
##  Median :23.00   Median :4.000   Median :146.0   Median :61.00   Median :2800
##  Mean   :23.52   Mean   :5.458   Mean   :193.5   Mean   :51.52   Mean   :2970
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0   3rd Qu.:79.00   3rd Qu.:3609
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :94.00   Max.   :5140
##   acceleration
##  Min.   : 8.00
##  1st Qu.:13.80
##  Median :15.50
##  Mean   :15.56
##  3rd Qu.:17.10
##  Max.   :24.80
```

```
sd(quant.var$mpg)
```

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

```
sd(quant.var$cylinders)
```

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

```
sd(quant.var$displacement)
```

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

```r
sd(quant.var$horsepower)
```

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

```r
sd(quant.var$weight)
```

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

```r
sd(quant.var$acceleration)
```

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

Variables	Range	Mean	SD
mpg	37.6	23.52	7.82
cylinders	5	5.46	1.70
displacement	387	193.53	104.38
horsepower	93	51.12	29.86
weight	3509	146	33.29
acceleration	16.8	15.56	2.74

Table 1: Statistical parameters for quantitative variables (whole dataset)

2 (c)

After removing the 40th through 80th (inclusive) observations from the dataset, the range, mean, and standard deviation of each predictor in the subset of the data are given below.

```r
autodata2 = autodata[-c(40:80),] # Data after discarding rows 40 to 80 inclusive
horsepower = as.numeric(autodata2$horsepower)
summary(autodata2)
```

```
##       mpg          cylinders      displacement     horsepower      weight    
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   88     : 18   Min.   :1649  
##  1st Qu.:18.00   1st Qu.:4.000   1st Qu.:103.2   90     : 18   1st Qu.:2222  
##  Median :23.65   Median :4.000   Median :146.0   110    : 17   Median :2782  
##  Mean   :24.02   Mean   :5.399   Mean   :189.2   150    : 17   Mean   :2935  
##  3rd Qu.:29.82   3rd Qu.:6.000   3rd Qu.:252.0   100    : 16   3rd Qu.:3508  
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   75     : 14   Max.   :4997  
##                                                  (Other):256                 
##   acceleration        year           origin                      name    
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   ford pinto    :  6  
##  1st Qu.:14.00   1st Qu.:74.00   1st Qu.:1.000   amc matador   :  5  
##  Median :15.50   Median :77.00   Median :1.000   ford maverick :  5  
##  Mean   :15.61   Mean   :76.51   Mean   :1.593   toyota corolla:  5  
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc gremlin   :  4  
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   amc hornet    :  4  
##                                                  (Other)       :327  
```

Variable	**Range**	**Mean**	**SD**
mpg	37.6	24.02	7.83
cylinder	5	5.399	1.65
displacement	387	189.2	100.88
horsepower	93	51.67	30.36
weight	3348	2935	810.84
acceleration	16.8	15.61	2.712

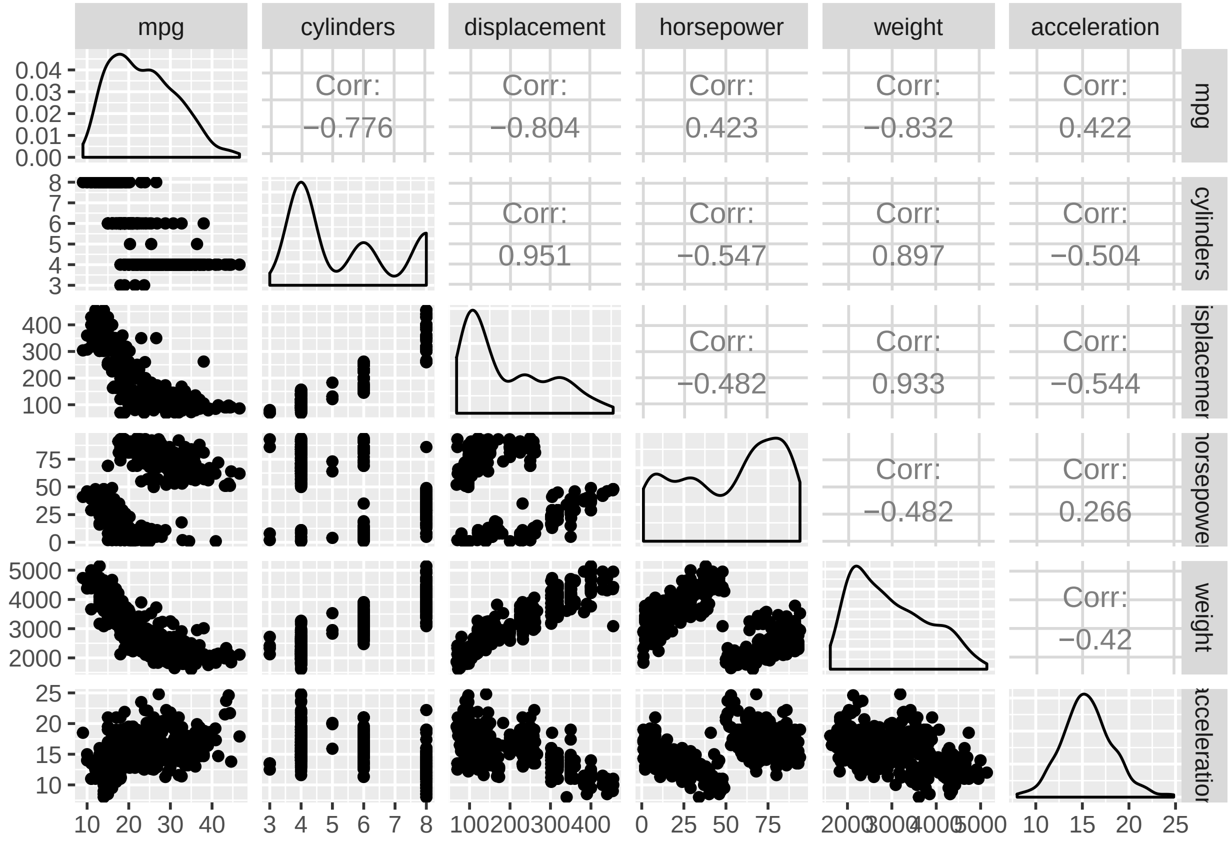
Table 2: Statistical parameters for quantitative variables (data subset)

2 (d)

The scatter plots and the pairwise correlations among the quantitative predictors of are generated below. The plot below provides all of these in the same graph.

The positive correlation coefficient value indicates that there lies a linearly positive relationship between the variables where the negative sign means the opposite. Higher the value, higher the correlation. For example, it is apparent that the "displacement" variable is highly correlated with the "cylinder" variable which is expected.

```
library(GGally)
ggpairs(quant.var)
```



2 (e)

From the above correlation plot, it is evident that the “mpg” variable is highly correlated with “weight”, “displacement” and “cylinders” and they have inversely proportional relationship with the gas mileage (“mpg”). Thus, these three variables will be useful in predicting “mpg”. However, the “horsepower” and “acceleration” variables are moderately correlated and thus won’t be helpful as much towards predicting the gas mileage.