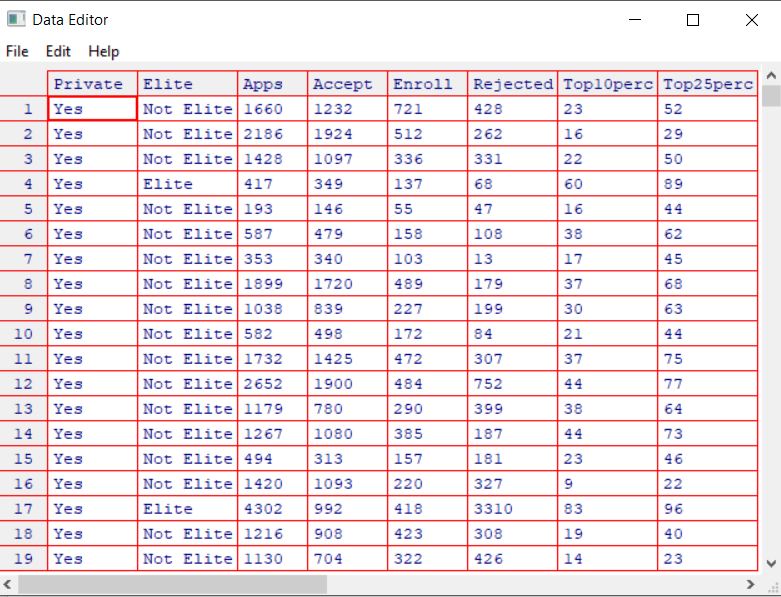
**8A)**

# Load the dataset

College\_HW = ISLR::College

**8B) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later.**

fix(College\_HW)



**Ci) Use the summary() function to produce a numerical summary of the variables in the data set.**

summary(College\_HW)

Private Apps Accept Enroll Top10perc Top25perc

No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00 Min. : 9.0

Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00 1st Qu.: 41.0

Median : 1558 Median : 1110 Median : 434 Median :23.00 Median : 54.0

Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56 Mean : 55.8

3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00 3rd Qu.: 69.0

Max. :48094 Max. :26330 Max. :6392 Max. :96.00 Max. :100.0

F.Undergrad P.Undergrad Outstate Room.Board Books Personal

Min. : 139 Min. : 1.0 Min. : 2340 Min. :1780 Min. : 96.0 Min. : 250

1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850

Median : 1707 Median : 353.0 Median : 9990 Median :4200 Median : 500.0 Median :1200

Mean : 3700 Mean : 855.3 Mean :10441 Mean :4358 Mean : 549.4 Mean :1341

3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700

Max. :31643 Max. :21836.0 Max. :21700 Max. :8124 Max. :2340.0 Max. :6800

PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate

Min. : 8.00 Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186 Min. : 10.00

1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751 1st Qu.: 53.00

Median : 75.00 Median : 82.0 Median :13.60 Median :21.00 Median : 8377 Median : 65.00

Mean : 72.66 Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660 Mean : 65.46

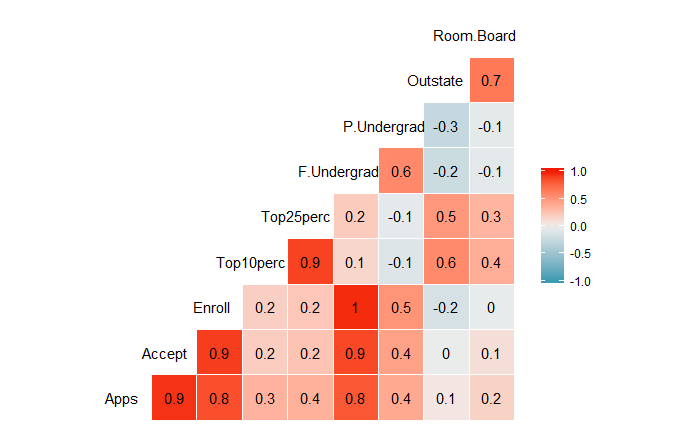
3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830 3rd Qu.: 78.00

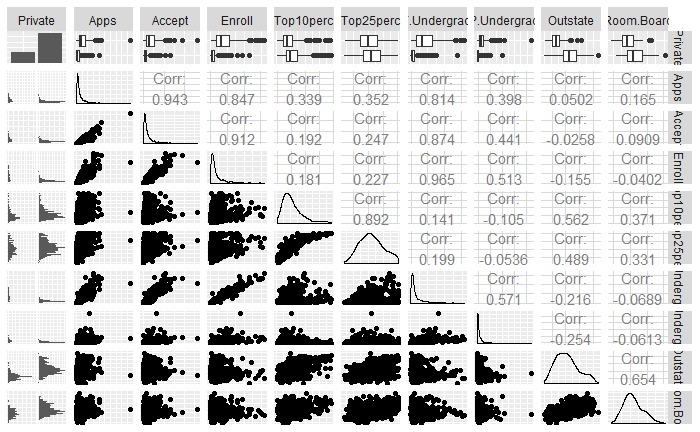
Max. :103.00 Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233 Max. :118.00

**Cii) Use the pairs() function to produce a scatterplot matrix the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].**

ggcorr(College\_HW[,1:10], palette = "RdBu", label = TRUE, hjust = 0.7)

ggpairs(College\_HW[,1:10], axisLabels = "none", progress = FALSE)

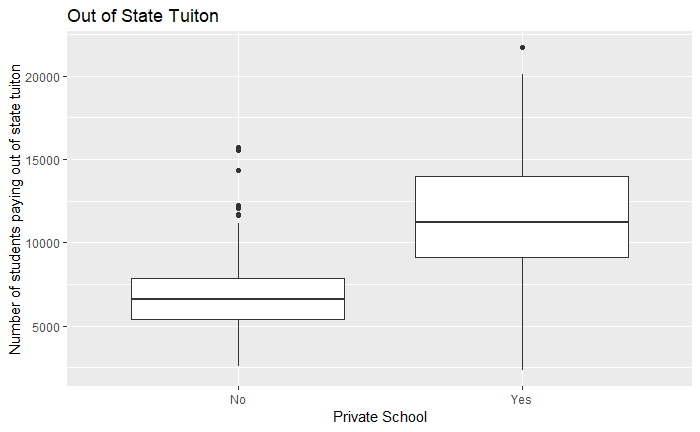




**C\_iii) Use the plot() function to produce side-by-side boxplots of Outstate versus Private.**

out\_vs\_private = ggplot(data = College\_HW, aes(Private, Outstate))

out\_vs\_private + geom\_boxplot() + ggtitle("Out of State Tuiton") + xlab("Private School") + ylab("Number of students paying out of state tuiton")



**C\_iv) Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.**

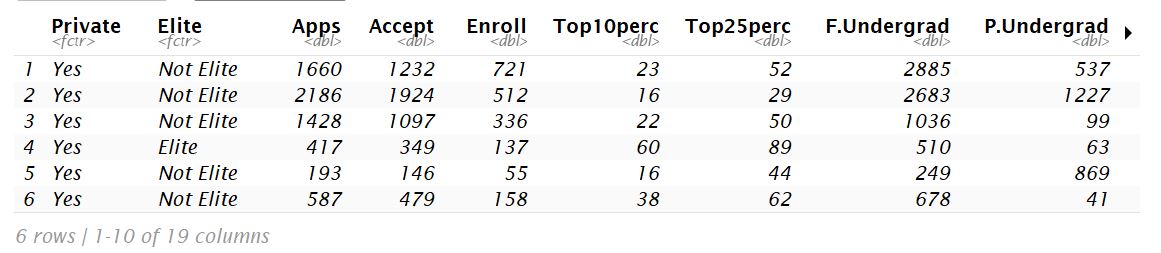
College\_HW = mutate(College\_HW, Elite = ifelse(College\_HW$Top10perc > 50, "Elite", "Not Elite"))

College\_HW$Elite = as.factor(College\_HW$Elite)

colnames(College\_HW)

College\_HW = College\_HW[, c(1, 19, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18)]

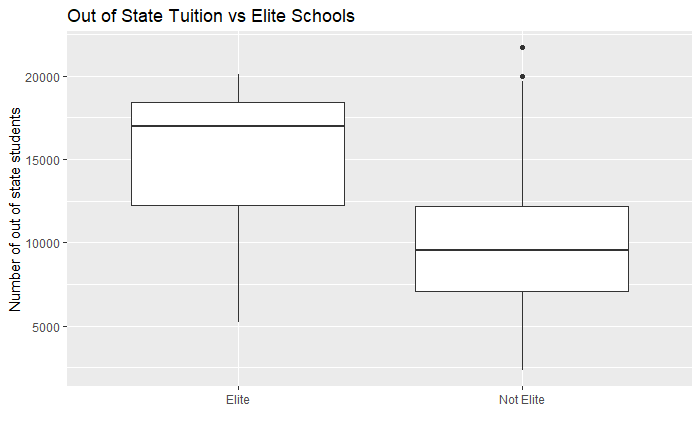
head(College\_HW)



summary(College\_HW)

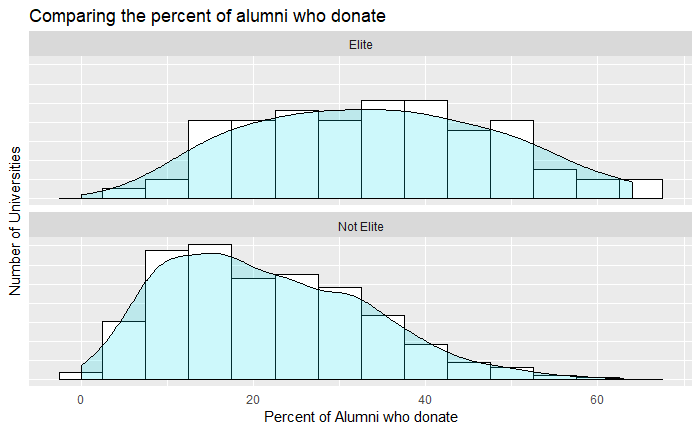
out\_vs\_elite = ggplot(data = College\_HW, aes(Elite, Outstate))

out\_vs\_elite + geom\_boxplot() + ggtitle("Out of State Tuition vs Elite Schools") + xlab("") + ylab("Number of out of state students")

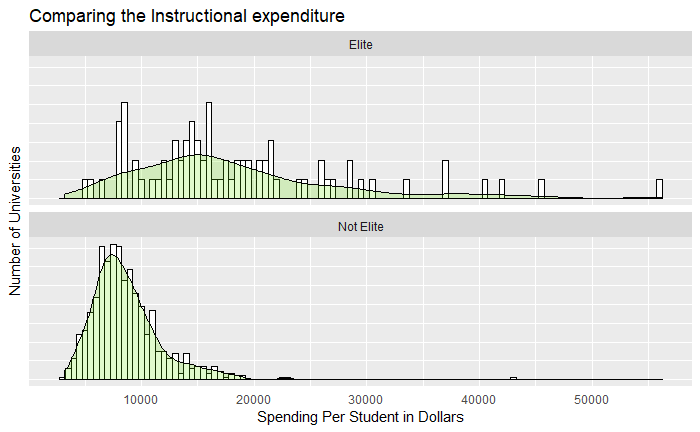


**D) Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.**

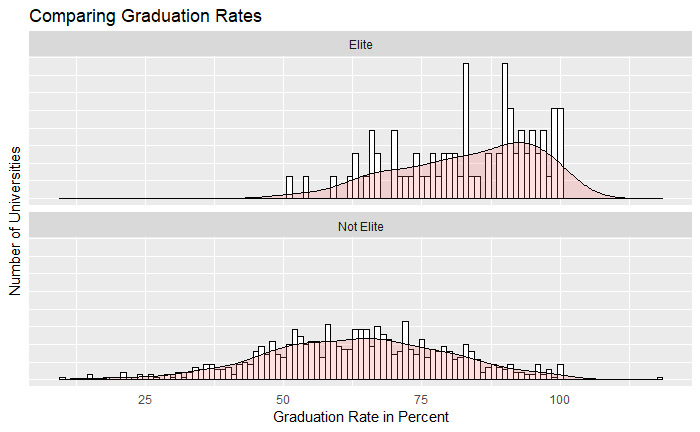
ggplot(data = College\_HW, aes(x = perc.alumni)) + geom\_histogram(aes(y=..density..), binwidth = 5, color = "black", fill = "white") + geom\_density(alpha = 0.2, fill = "#07DEED") + theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) + facet\_wrap(~Elite, ncol = 1) + ggtitle("Comparing the percent of alumni who donate") + xlab("Percent of Alumni who donate") + ylab("Number of Universities")



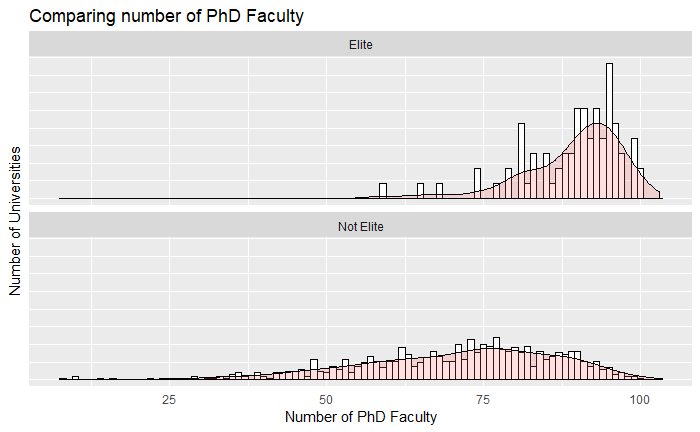
ggplot(data = College\_HW, aes(x=Expend)) + geom\_histogram(aes(y = ..density..), binwidth = 500, color = "black", fill = "white") + geom\_density(alpha = 0.2, fill = "#67ED00") + theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) + facet\_wrap(~Elite, ncol = 1) + ggtitle("Comparing the Instructional expenditure") + xlab("Spending Per Student in Dollars") + ylab("Number of Universities")



ggplot(data = College\_HW, aes(x=Grad.Rate)) + geom\_histogram(aes(y=..density..),binwidth = 1, color = "black", fill = "white") + geom\_density(alpha = 0.2, fill = "#FF6666") + theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) + facet\_wrap(~Elite, ncol = 1) + ggtitle("Comparing Graduation Rates") + xlab("Graduation Rate in Percent") + ylab("Number of Universities")



ggplot(data = College\_HW, aes(x=PhD)) + geom\_histogram(aes(y=..density..),binwidth = 1, color = "black", fill = "white") + geom\_density(alpha = 0.2, fill = "#FF6666") + theme(axis.ticks = element\_blank(), axis.text.y = element\_blank()) + facet\_wrap(~Elite, ncol = 1) + ggtitle("Comparing number of PhD Faculty") + xlab("Number of PhD Faculty") + ylab("Number of Universities")



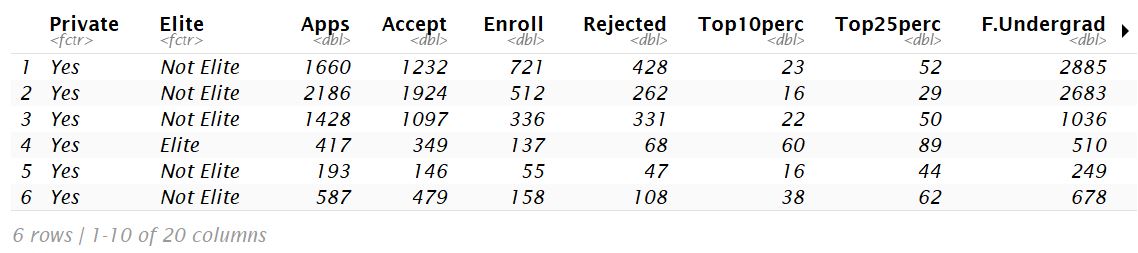
To help better analyze the data we want to create a new column titled "Rejected". This will be the difference between the number of application accepted (Accept) and the number of new students enrolled (Enroll). We will look at some descriptive statistics and basic graphs. We want to see of elite schools reject more students than not elite schools.

College\_HW = mutate(College\_HW, Rejected = Apps - Accept)

colnames(College\_HW)

College\_HW = College\_HW[, c(1, 2, 3, 4, 5, 20, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19)]

head(College\_HW)



summary(College\_HW)

rejected\_vs\_elite = ggplot(data = College\_HW, aes(Elite, Rejected))

apps\_vs\_elite = ggplot(data = College\_HW, aes(x = Elite, y = Apps))

rejected\_vs\_elite + geom\_boxplot() + ggtitle("Number of rejected students from both Elite School and not Elite schools", subtitle = "Elite School admits over 50% of the top 10% of High School Students") + xlab("") + ylab("Number of rejected students")

require(gridExtra)

plot1 = rejected\_vs\_elite + geom\_col() + scale\_y\_continuous(labels = comma

plot2 = apps\_vs\_elite + geom\_col() + scale\_y\_continuous(labels = comma)

grid.arrange(plot1, plot2, ncol = 2, top = "Comparing Rejection Numbers")

Private Elite Apps Accept Enroll Rejected

No :212 Elite : 78 Min. : 81 Min. : 72 Min. : 35 Min. : 0.0

Yes:565 Not Elite:699 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.: 131.0

Median : 1558 Median : 1110 Median : 434 Median : 326.0

Mean : 3002 Mean : 2019 Mean : 780 Mean : 982.8

3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.: 1066.0

Max. :48094 Max. :26330 Max. :6392 Max. :21764.0

Top10perc Top25perc F.Undergrad P.Undergrad Outstate Room.Board

Min. : 1.00 Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340 Min. :1780

1st Qu.:15.00 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320 1st Qu.:3597

Median :23.00 Median : 54.0 Median : 1707 Median : 353.0 Median : 9990 Median :4200

Mean :27.56 Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441 Mean :4358

3rd Qu.:35.00 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925 3rd Qu.:5050

Max. :96.00 Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700 Max. :8124

Books Personal PhD Terminal S.F.Ratio perc.alumni

Min. : 96.0 Min. : 250 Min. : 8.00 Min. : 24.0 Min. : 2.50 Min. : 0.00

1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00

Median : 500.0 Median :1200 Median : 75.00 Median : 82.0 Median :13.60 Median :21.00

Mean : 549.4 Mean :1341 Mean : 72.66 Mean : 79.7 Mean :14.09 Mean :22.74

3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00

Max. :2340.0 Max. :6800 Max. :103.00 Max. :100.0 Max. :39.80 Max. :64.00

Expend Grad.Rate

Min. : 3186 Min. : 10.00

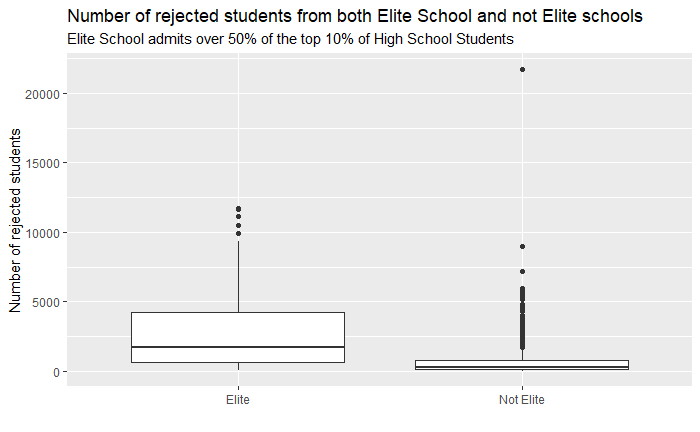
1st Qu.: 6751 1st Qu.: 53.00

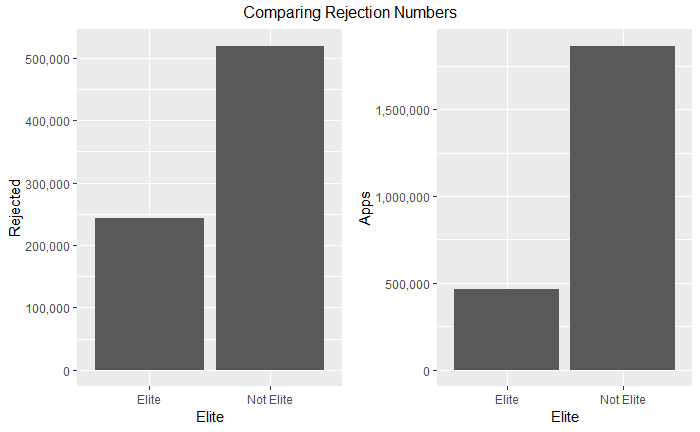
Median : 8377 Median : 65.00

Mean : 9660 Mean : 65.46

3rd Qu.:10830 3rd Qu.: 78.00

Max. :56233 Max. :118.00





**D) Continue exploring the data, and provide a brief summary of what you discover**

We can see a strong correlation between many of the variables. There is almost perfect positive linear correlation between Apps and Accept, Accept and Enroll, Apps and Enroll, Enroll and F.Undergrad, Accept and F.Undergrad, Apps and F.Undergrad, and Top10Perc and Top25Perc.

There is a weak negative correlation between undergraduate students and out of state tuition and room and board. We can interpret this as either when the number of undergraduate students increase the number of students paying room and board decreases or vice versa.

From the data exploration we can see that Elite schools admit more out of state students. Additionally, the Elite schools have a higher percentage of alumni who donate to their programs. Elite schools also spend more per student then non-elite schools and they have a higher graduation rate.

If we compare the number of students rejected from elite and non-elite schools the proportion of students rejected from elite schools is about 250,000/500,000 = 0.5 or 50% while non-elite schools are 550,000/2,500,000 = .22 or 22%.

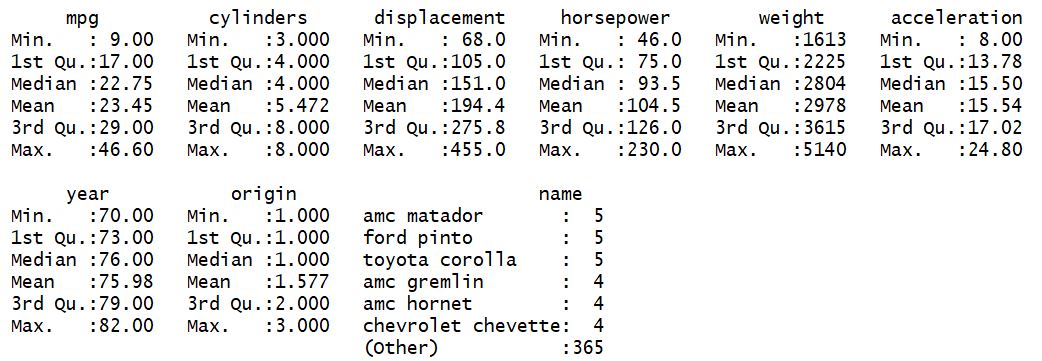
The next step would be a run a regression analysis on some of the variables to determine the nature of their relationship.

**Problem # 9**

**Import and clean data**

Auto\_Data = ISLR::Auto

summary(Auto\_Data)



# Check for missing values

sum(is.na(Auto\_Data))

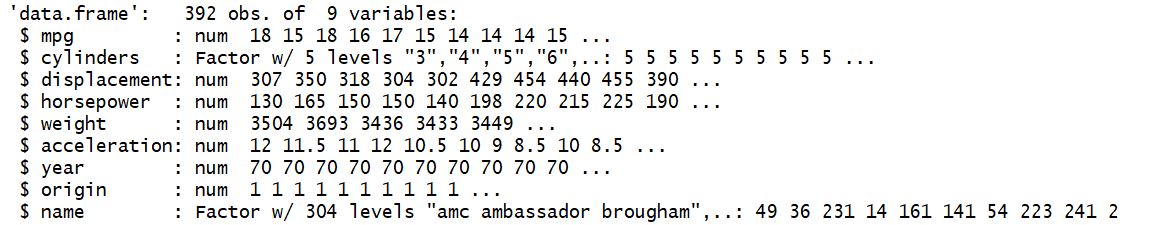


# Note: The orginal data contained 408 observations but 16 observations with missing values were removed

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

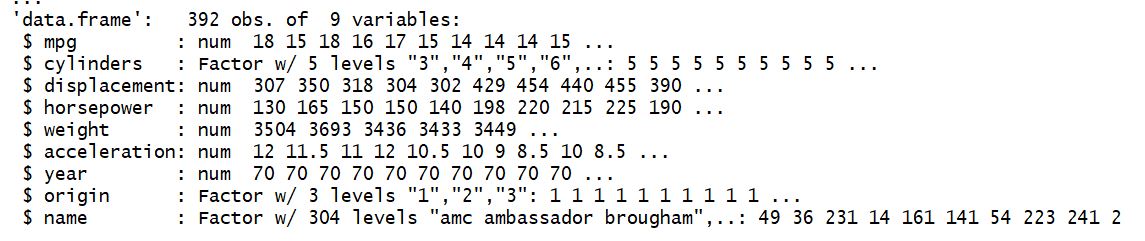
# See what type of data each variable is

str(Auto\_Data)

Auto\_Data$cylinders = as.factor(Auto\_Data$cylinders)

Auto\_Data$origin = as.factor(Auto\_Data$origin)

str(Auto\_Data)



**What is the range of each quantitative predictor?**

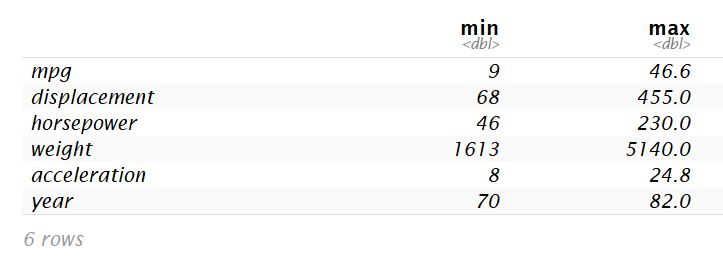
#Subset the data to remove non-quantitative

num\_subset = Auto\_Data%>%select(-name, -cylinders, -origin)

# Create dataframe of min and max

variable\_range = data.frame(min=sapply(num\_subset, min),max=sapply(num\_subset, max))

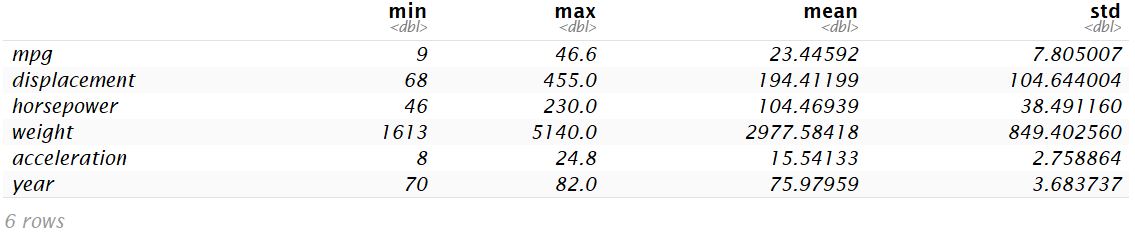
variable\_range



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

des\_num\_subset = data.frame(min=sapply(num\_subset, min),max=sapply(num\_subset, max), mean=sapply(num\_subset, mean), std=sapply(num\_subset,sd))

des\_num\_subset



**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?**

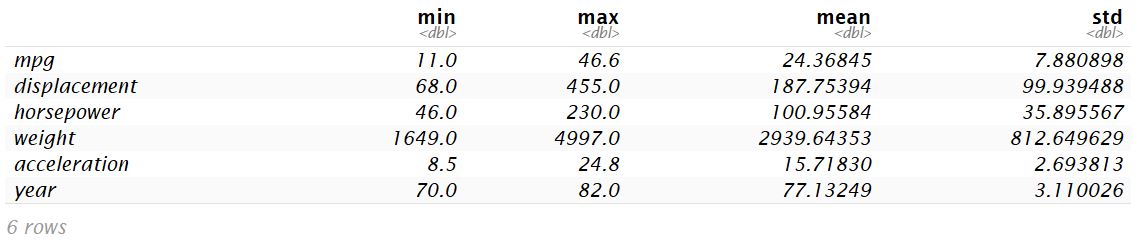
# Remove obs # 10-85

num\_not\_10\_85 = num\_subset[-(10:84),]

# Create dataframe of range, mean, and std

des\_num\_not\_10\_85 = data.frame(min=sapply(num\_not\_10\_85, min),max=sapply(num\_not\_10\_85, max), mean=sapply(num\_not\_10\_85, mean), std=sapply(num\_not\_10\_85,sd))

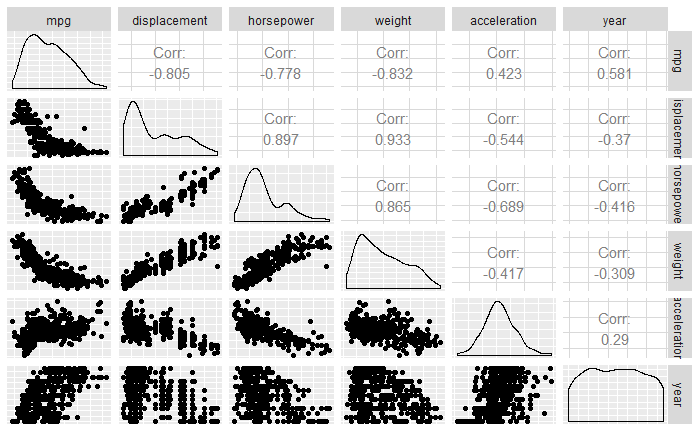
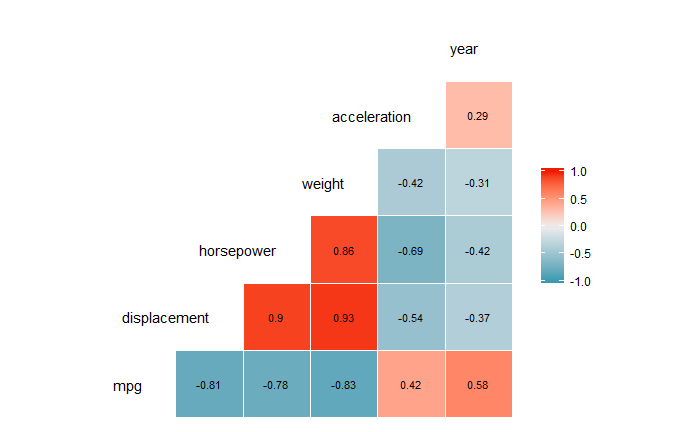
des\_num\_not\_10\_85



**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.**

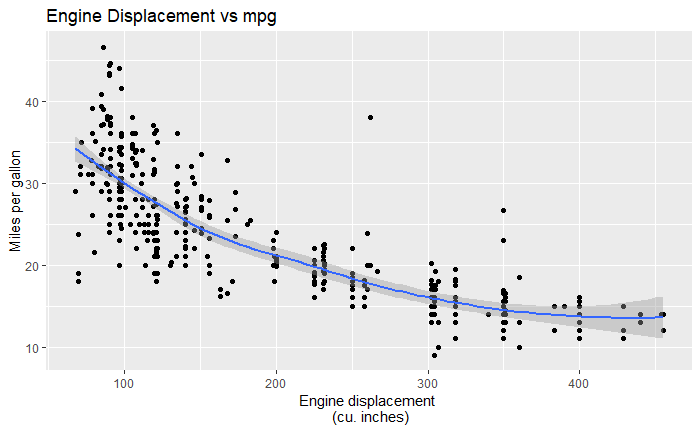
ggcorr(Auto\_Data, palette = "RdBu", label = TRUE, label\_size = 3, label\_round = 2, hjust = 1 )

ggpairs(num\_subset, axisLabels = "none", progress = FALSE)

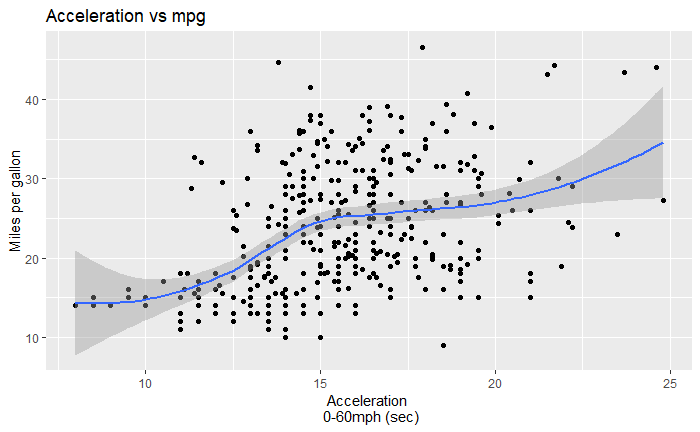


ggplot(data = Auto\_Data, mapping = aes(x = displacement, y = mpg)) +

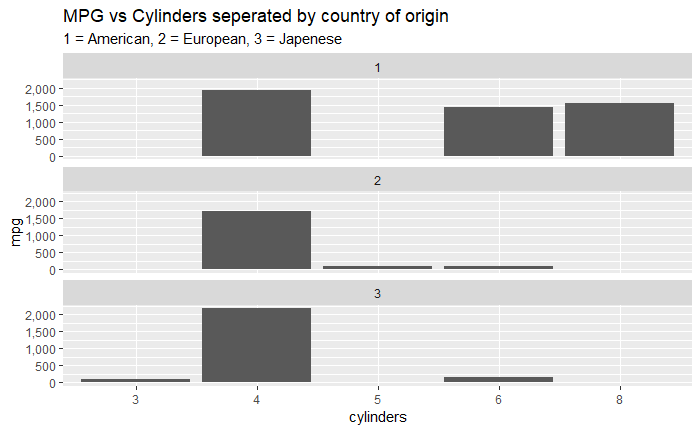
geom\_point() + geom\_smooth() + ggtitle("Engine Displacement vs mpg") + xlab("Engine displacement \n (cu. inches)") + ylab("Miles per gallon")



ggplot(data = Auto\_Data, mapping = aes(x = acceleration, y =mpg)) + geom\_point() + geom\_smooth() + ggtitle("Acceleration vs mpg") + xlab("Acceleration \n 0-60mph (sec)") + ylab("Miles per gallon")



ggplot(data = Auto\_Data, mapping = aes(x = cylinders, y = mpg)) + geom\_col() + facet\_wrap(~origin, ncol = 1) + ggtitle("MPG vs Cylinders seperated by country of origin", subtitle = "1 = American, 2 = European, 3 = Japenese") + scale\_y\_continuous(labels = comma)



**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.**

From the correlation analysis we can see that MPG is strongly correlated with displacement, horsepower, weight, and acceleration. However; weight and horsepower, and weight and displacement have almost perfect linear correlation. We should check for multicollinearity and maybe remove one of those three variables.

We could also use the country of origin as predictor as well. We can see that American made cars tend to favor higher cylinder engines while European and Japanese cars tend to have smaller cylinder engines.