Homework1

Bethany Smith

January 19, 2018

# Problem 8

### (A)

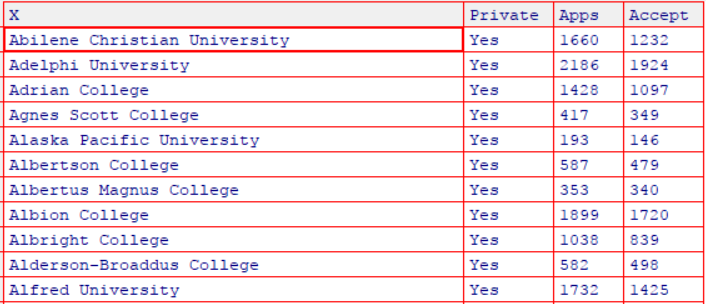
#### Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

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

### (B)

#### 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)



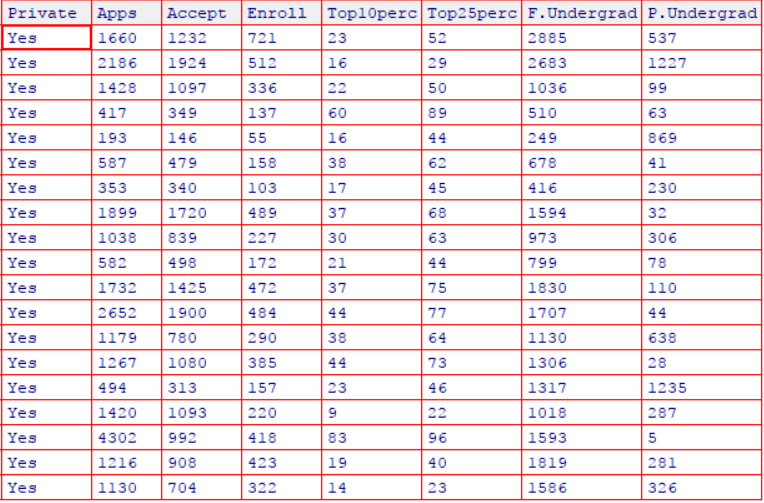
#### 

#### 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.

rownames <- college[, 1]  
college <- college[, -1]  
head(college[, 1:5])

## Private Apps Accept Enroll Top10perc  
## 1 Yes 1660 1232 721 23  
## 2 Yes 2186 1924 512 16  
## 3 Yes 1428 1097 336 22  
## 4 Yes 417 349 137 60  
## 5 Yes 193 146 55 16  
## 6 Yes 587 479 158 38

fix(college)



### 

### (C) i. Use the summary() function to produce a numerical summary of the variables in the data set

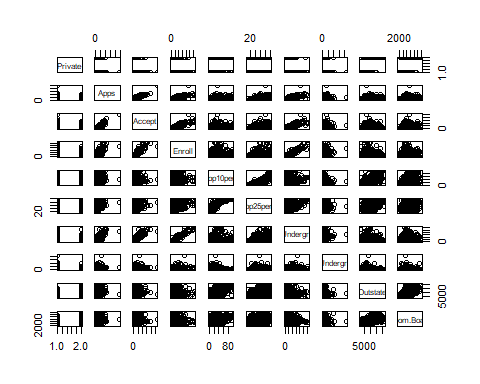
***Answer****: We see from the numerical summary outputed below the Minimum, Maximum, Mean, Median etc. for each variable found in the college data set. From initially looking at the data it appears as if all of the variables are numeric continuous while the Private variable is categorical. We see that there are wide ranges of values that our variables fall into, some have quite large minimums and maximum while others have rather small minimum and maximums. If we were doing any sort of predictive modeling with the college data set we may want to consider standardizing the variables to account for these large deviations in scale across all variables.*

summary(college)

## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00   
## Max. :48094 Max. :26330 Max. :6392 Max. :96.00   
## Top25perc F.Undergrad P.Undergrad Outstate   
## Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340   
## 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320   
## Median : 54.0 Median : 1707 Median : 353.0 Median : 9990   
## Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441   
## 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925   
## Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700   
## Room.Board Books Personal PhD   
## Min. :1780 Min. : 96.0 Min. : 250 Min. : 8.00   
## 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00   
## Median :4200 Median : 500.0 Median :1200 Median : 75.00   
## Mean :4358 Mean : 549.4 Mean :1341 Mean : 72.66   
## 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00   
## Max. :8124 Max. :2340.0 Max. :6800 Max. :103.00   
## Terminal S.F.Ratio perc.alumni Expend   
## Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186   
## 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751   
## Median : 82.0 Median :13.60 Median :21.00 Median : 8377   
## Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660   
## 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830   
## Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233   
## Grad.Rate   
## Min. : 10.00   
## 1st Qu.: 53.00   
## Median : 65.00   
## Mean : 65.46   
## 3rd Qu.: 78.00   
## Max. :118.00

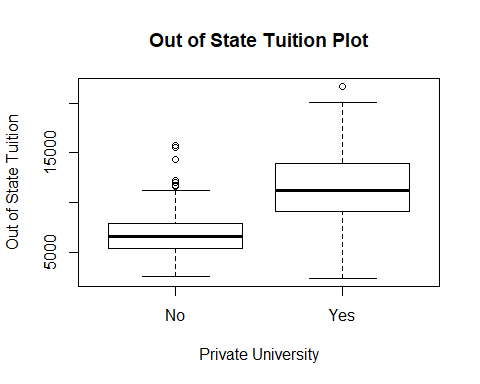
#### ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

pairs(college[,1:10])



#### iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

plot(college$Private, college$Outstate, xlab="Private University", ylab= "Out of State Tuition", main = "Out of State Tuition Plot")



***Interpretation:*** *We see from the above boxplot that Private Universities tend to have a higher average out of state tuition than public universities.*

#### 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 %.

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

#### Use the summary() funcation to see how many elite universities there are.

summary(college$Elite)

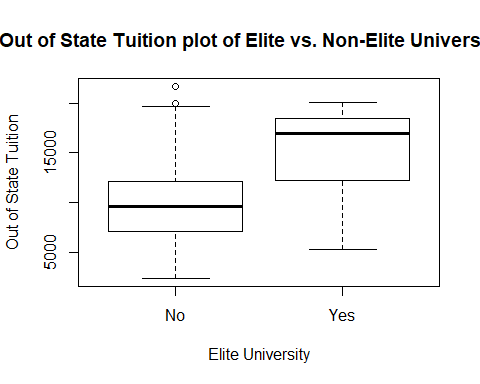
## No Yes   
## 699 78

***Answer****: There are a total of 78 elite universities and 699 non elite universities within the college data set.*

#### 

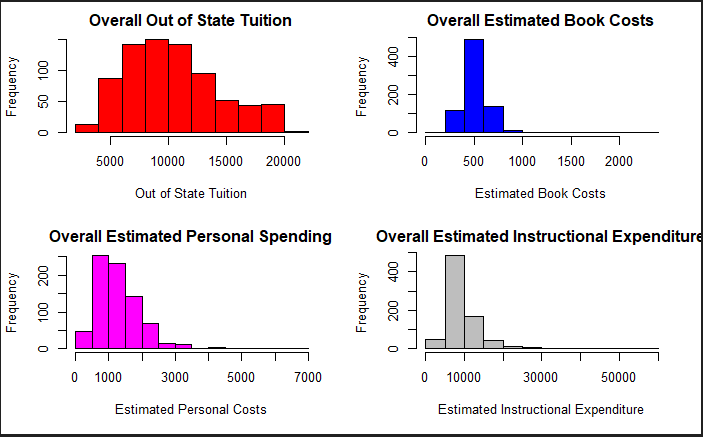
#### Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite

plot(college$Elite, college$Outstate, xlab= "Elite University", ylab= "Out of State Tuition", main = "Out of State Tuition plot of Elite vs. Non-Elite Universities")

   
***Interpretation:*** *Based on the box plot above: the average of out of state tuition is substantially higher for those universities considered “Elite”. Those students who are out of state and are concerned about tuition would be wise to explore there options at universities other than those considered “Elite”.*

#### v. 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.

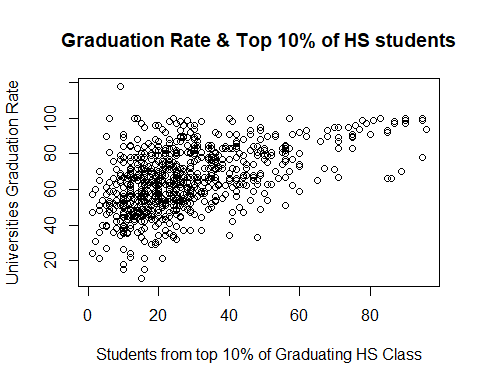
par(mfrow=c(2,2))  
hist(college$Outstate,col=2, xlab="Out of State Tuition", ylab = "Frequency", main="Overall Out of State Tuition")  
hist(college$Books,col=4, xlab="Estimated Book Costs", ylab= "Frequency", main = "Overall Estimated Book Costs")  
hist(college$Personal, col=6, xlab= "Estimated Personal Costs", ylab="Frequency", main="Overall Estimated Personal Spending")  
  
hist(college$Expend, col=8, xlab= "Estimated Instructional Expenditure", ylab="Frequency", main="Overall Estimated Instructional Expenditure")



***Interpretation:*** *We see from the histograms above that overall there is a large amount of variation between the frequencies of estimates collected for the overall out of state tuition costs. Estimated instructional expenditure finds the largest frequency (over 400 total estimated) of estimates falling between $5,000 and $10,000.*

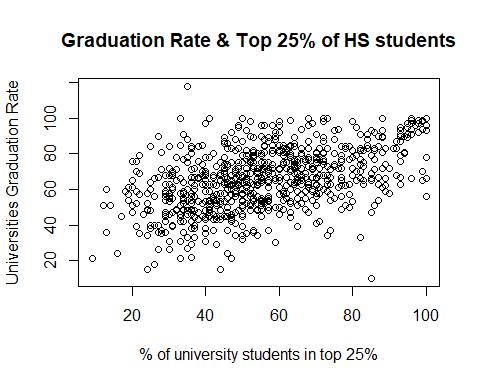
#### vi. Continue exploring the data, and provide a brief summary of what you discover.

plot(college$Top10perc, college$Grad.Rate, xlab="Students from top 10% of Graduating HS Class", ylab = "Universities Graduation Rate", main = "Graduation Rate & Top 10% of HS students")

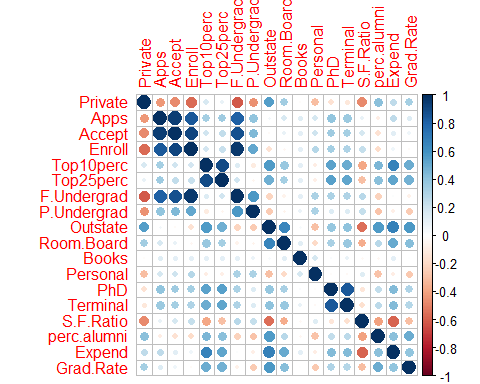


***Answer:*** *We can see from the above scatterplot that those universities whose student bodies have a larger proportion of students who were among the top 10% of students in their high school graduating class are not necessarily associated with higher graduation rates. The same is true for those universities whose student bodies have a larger proportion of students who were among the top 25% of students in their high school graduating class.*

plot(college$Top25perc, college$Grad.Rate, xlab="% of university students in top 25%", ylab = "Universities Graduation Rate", main = "Graduation Rate & Top 25% of HS students")



college2 = select(college, Private, Apps, Accept, Enroll, Top10perc, Top25perc, F.Undergrad, P.Undergrad, Outstate, Room.Board, Books, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend, Grad.Rate)  
college2$Private = as.numeric(college2$Private)  
correlation1 = cor(college2)  
corrplot(correlation1)



***Answer****: We investigate the association and relationships amongst the variables within the college data set by first visually inspecting the correlation plot above. We see that the variables Applications, Accept, and Enroll are highly correlated, top 10 percent is correlated with out of state, PhD, Expenditure, and graduation rate.Furthermore, terminal and PhD are highly correlated. To further investigate the correlation amongst these variables we next explore the pearson correlation coefficients to assess whether there is further evidence to support the conclusions made from the visual inspection of association amongst the variables.*

A <- cor(college2)  
A <- round(A,digits = 2)  
A <- as.table(A, nrow(19), ncol(19))  
  
A

## Private Apps Accept Enroll Top10perc Top25perc F.Undergrad  
## Private 1.00 -0.43 -0.48 -0.57 0.16 0.10 -0.62  
## Apps -0.43 1.00 0.94 0.85 0.34 0.35 0.81  
## Accept -0.48 0.94 1.00 0.91 0.19 0.25 0.87  
## Enroll -0.57 0.85 0.91 1.00 0.18 0.23 0.96  
## Top10perc 0.16 0.34 0.19 0.18 1.00 0.89 0.14  
## Top25perc 0.10 0.35 0.25 0.23 0.89 1.00 0.20  
## F.Undergrad -0.62 0.81 0.87 0.96 0.14 0.20 1.00  
## P.Undergrad -0.45 0.40 0.44 0.51 -0.11 -0.05 0.57  
## Outstate 0.55 0.05 -0.03 -0.16 0.56 0.49 -0.22  
## Room.Board 0.34 0.16 0.09 -0.04 0.37 0.33 -0.07  
## Books -0.02 0.13 0.11 0.11 0.12 0.12 0.12  
## Personal -0.30 0.18 0.20 0.28 -0.09 -0.08 0.32  
## PhD -0.16 0.39 0.36 0.33 0.53 0.55 0.32  
## Terminal -0.13 0.37 0.34 0.31 0.49 0.52 0.30  
## S.F.Ratio -0.47 0.10 0.18 0.24 -0.38 -0.29 0.28  
## perc.alumni 0.41 -0.09 -0.16 -0.18 0.46 0.42 -0.23  
## Expend 0.26 0.26 0.12 0.06 0.66 0.53 0.02  
## Grad.Rate 0.34 0.15 0.07 -0.02 0.49 0.48 -0.08  
## P.Undergrad Outstate Room.Board Books Personal PhD Terminal  
## Private -0.45 0.55 0.34 -0.02 -0.30 -0.16 -0.13  
## Apps 0.40 0.05 0.16 0.13 0.18 0.39 0.37  
## Accept 0.44 -0.03 0.09 0.11 0.20 0.36 0.34  
## Enroll 0.51 -0.16 -0.04 0.11 0.28 0.33 0.31  
## Top10perc -0.11 0.56 0.37 0.12 -0.09 0.53 0.49  
## Top25perc -0.05 0.49 0.33 0.12 -0.08 0.55 0.52  
## F.Undergrad 0.57 -0.22 -0.07 0.12 0.32 0.32 0.30  
## P.Undergrad 1.00 -0.25 -0.06 0.08 0.32 0.15 0.14  
## Outstate -0.25 1.00 0.65 0.04 -0.30 0.38 0.41  
## Room.Board -0.06 0.65 1.00 0.13 -0.20 0.33 0.37  
## Books 0.08 0.04 0.13 1.00 0.18 0.03 0.10  
## Personal 0.32 -0.30 -0.20 0.18 1.00 -0.01 -0.03  
## PhD 0.15 0.38 0.33 0.03 -0.01 1.00 0.85  
## Terminal 0.14 0.41 0.37 0.10 -0.03 0.85 1.00  
## S.F.Ratio 0.23 -0.55 -0.36 -0.03 0.14 -0.13 -0.16  
## perc.alumni -0.28 0.57 0.27 -0.04 -0.29 0.25 0.27  
## Expend -0.08 0.67 0.50 0.11 -0.10 0.43 0.44  
## Grad.Rate -0.26 0.57 0.42 0.00 -0.27 0.31 0.29  
## S.F.Ratio perc.alumni Expend Grad.Rate  
## Private -0.47 0.41 0.26 0.34  
## Apps 0.10 -0.09 0.26 0.15  
## Accept 0.18 -0.16 0.12 0.07  
## Enroll 0.24 -0.18 0.06 -0.02  
## Top10perc -0.38 0.46 0.66 0.49  
## Top25perc -0.29 0.42 0.53 0.48  
## F.Undergrad 0.28 -0.23 0.02 -0.08  
## P.Undergrad 0.23 -0.28 -0.08 -0.26  
## Outstate -0.55 0.57 0.67 0.57  
## Room.Board -0.36 0.27 0.50 0.42  
## Books -0.03 -0.04 0.11 0.00  
## Personal 0.14 -0.29 -0.10 -0.27  
## PhD -0.13 0.25 0.43 0.31  
## Terminal -0.16 0.27 0.44 0.29  
## S.F.Ratio 1.00 -0.40 -0.58 -0.31  
## perc.alumni -0.40 1.00 0.42 0.49  
## Expend -0.58 0.42 1.00 0.39  
## Grad.Rate -0.31 0.49 0.39 1.00

***Answer:*** *The table above displays the pearson correlation coefficients for the variables found within the college dataset. Values above .5 that are either positive or negative, tend to be good indications of correlation, the closer a value is to positive or negative one the stronger the association between those two variables. We conclude that our visual inspection and the conclusions made above are valid. Below is a summary of the insight gathered about the association and relationships amongst the college data set variables :*

*Applications are highly correlated with Accept (.94), Enroll (.85), Full Time Undergraduates (.814) Furthermore, there is a lack of* *correlation between applications and Student/Faculty ratio (.10), Book Expenditure (.13), and Graduation Rate (.15).*

*Top 10 percent is highly correlated with out of state tuition (.56), percent of facult with PhD’s (.53), stimated instructional expenditure* *per student (.66) and is marginally associated with Graduation Rate (.49). This conclusion about top 10 percent and graduation rate provides further evidence in support of the conclusions drawn based off of scatterplot above*  
**Problem 9**

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

Auto <- read.csv("Auto.csv", header=T, na.strings = "?")  
fix(Auto)  
dim(Auto)

## [1] 397 9

Auto= na.omit(Auto)  
dim(Auto)

## [1] 392 9

names(Auto)

## [1] "mpg" "cylinders" "displacement" "horsepower"   
## [5] "weight" "acceleration" "year" "origin"   
## [9] "name"

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

***Answer:*** *We see that the variable “name” is a qualitative variable. Due to the fact that the variable origin only has three unique values, we can conclude that this is a categorical variable and should be treated as qualitative rather than quantitative.Furthermore, Cylinders only has 8 unique values, we could treat it as a quantitative or qualitative variable, for this analysis I choose to treat it as a qualitative variable.*

summary(Auto)

## mpg cylinders displacement horsepower   
## Min. : 9.00 Min. :3.000 Min. : 68.0 Min. : 46.0   
## 1st Qu.:17.00 1st Qu.:4.000 1st Qu.:105.0 1st Qu.: 75.0   
## Median :22.75 Median :4.000 Median :151.0 Median : 93.5   
## Mean :23.45 Mean :5.472 Mean :194.4 Mean :104.5   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:275.8 3rd Qu.:126.0   
## Max. :46.60 Max. :8.000 Max. :455.0 Max. :230.0   
##   
## weight acceleration year origin   
## Min. :1613 Min. : 8.00 Min. :70.00 Min. :1.000   
## 1st Qu.:2225 1st Qu.:13.78 1st Qu.:73.00 1st Qu.:1.000   
## Median :2804 Median :15.50 Median :76.00 Median :1.000   
## Mean :2978 Mean :15.54 Mean :75.98 Mean :1.577   
## 3rd Qu.:3615 3rd Qu.:17.02 3rd Qu.:79.00 3rd Qu.:2.000   
## Max. :5140 Max. :24.80 Max. :82.00 Max. :3.000   
##   
## name   
## amc matador : 5   
## ford pinto : 5   
## toyota corolla : 5   
## amc gremlin : 4   
## amc hornet : 4   
## chevrolet chevette: 4   
## (Other) :365

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

*We see that origin only has three categories, the minimum value being one and the maximum value being three.Below is the code used to try and determine what the levels of origin stand for:*

head(unique(Auto$name[Auto$origin==1]),11)

## [1] chevrolet chevelle malibu buick skylark 320   
## [3] plymouth satellite amc rebel sst   
## [5] ford torino ford galaxie 500   
## [7] chevrolet impala plymouth fury iii   
## [9] pontiac catalina amc ambassador dpl   
## [11] dodge challenger se   
## 304 Levels: amc ambassador brougham ... vw rabbit custom

head(unique(Auto$name[Auto$origin==2]),11)

## [1] volkswagen 1131 deluxe sedan peugeot 504   
## [3] audi 100 ls saab 99e   
## [5] bmw 2002 opel 1900   
## [7] peugeot 304 fiat 124b   
## [9] volkswagen model 111 volkswagen type 3   
## [11] volvo 145e (sw)   
## 304 Levels: amc ambassador brougham ... vw rabbit custom

head(unique(Auto$name[Auto$origin==3]),11)

## [1] toyota corona mark ii datsun pl510   
## [3] toyota corona toyota corolla 1200   
## [5] datsun 1200 toyota corona hardtop   
## [7] mazda rx2 coupe datsun 510 (sw)   
## [9] toyouta corona mark ii (sw) toyota corolla 1600 (sw)   
## [11] toyota carina   
## 304 Levels: amc ambassador brougham ... vw rabbit custom

***Answer:*** *From looking at the origin variable alongside the name of the manufacture at each level of origin we can conclude that each level of origin correlates to manufacturers found in specific countries/regions as so: 1=United States 2=European 3= Japenese. We fix the origin variable into a factor with the following code:*

Auto$origin <- factor(Auto$origin, levels=1:3, labels= c("United States", "Europe", "Japan"))  
sapply(Auto,class)

## mpg cylinders displacement horsepower weight   
## "numeric" "factor" "numeric" "numeric" "numeric"   
## acceleration year origin name   
## "numeric" "numeric" "factor" "factor"

***Answer:*** *In sum the auto datasets variables can be summarized as follows: Quantitative: mpg, displacement, horsepower, weight, acceleration, year Qualitative : cylinders, origin, name*

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

quantitative <- sapply(Auto, is.numeric)  
B <- sapply(Auto[,quantitative], function(x) c(range(x)))  
row.names(B) <- c("Minimum", "Maximum")

***Answer****: The range of each quantitative predictor is shown in the table below:*

b <- as.table(B,nrow(3), ncol(7))  
b

## mpg displacement horsepower weight acceleration year  
## Minimum 9.0 68.0 46.0 1613.0 8.0 70.0  
## Maximum 46.6 455.0 230.0 5140.0 24.8 82.0

### (C) What is the mean and standard deviation of each quantitative predictor?

***Answer:*** *The mean and standard deviation of each quantitative predictor is shown in the table below:*

quantitative <- sapply(Auto, is.numeric)  
S <- sapply(Auto[,quantitative], function(x) c(mean(x), sd(x)))  
row.names(S) <- c("Mean", "Standard Deviation")  
S

## mpg displacement horsepower weight  
## Mean 23.445918 194.412 104.46939 2977.5842  
## Standard Deviation 7.805007 104.644 38.49116 849.4026  
## acceleration year  
## Mean 15.541327 75.979592  
## Standard Deviation 2.758864 3.683737

### (D) 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?

***Answer:*** *After removing the 10th through 85th observations, the range, mean, and standard deviation of each quantitative variable is displayed in the output below:*

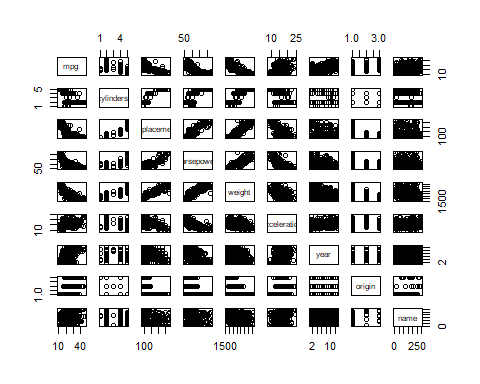
Auto\_new <- sapply(Auto[-10:-85,quantitative], function(x) c(range(x), mean(x), sd(x)))  
row.names(Auto\_new) <- c("Minimum", "Maximum", "Mean", "Standard Deviation")  
Auto\_new

## mpg displacement horsepower weight  
## Minimum 11.000000 68.00000 46.00000 1649.0000  
## Maximum 46.600000 455.00000 230.00000 4997.0000  
## Mean 24.404430 187.24051 100.72152 2935.9715  
## Standard Deviation 7.867283 99.67837 35.70885 811.3002  
## acceleration year  
## Minimum 8.500000 70.000000  
## Maximum 24.800000 82.000000  
## Mean 15.726899 77.145570  
## Standard Deviation 2.693721 3.106217

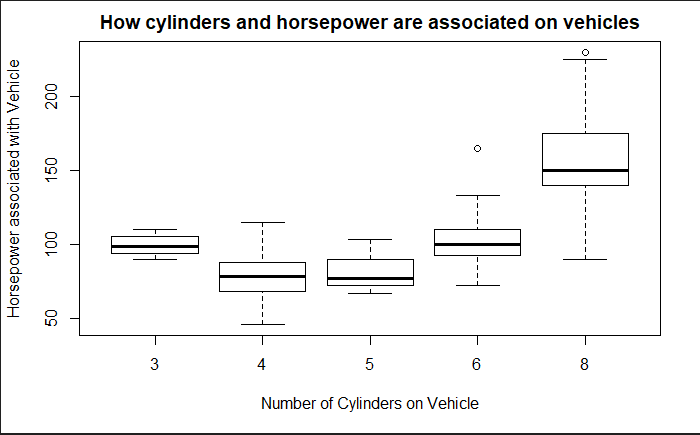
### (E)

#### 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.

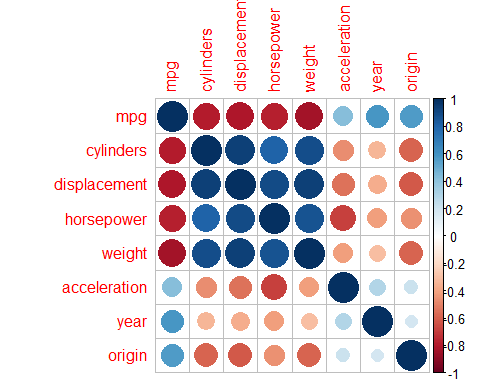
Auto$year <- as.factor(Auto$year)  
Auto$origin <- as.factor(Auto$origin)  
pairs(Auto)



plot(Auto$cylinders,Auto$horsepower, xlab= "Number of Cylinders on Vehicle", ylab="Horsepower associated with Vehicle", main=" A look at how cylinders and horsepower are associated on vehicles")

  
  
  
  
***Answer:*** *We see from the plot above that the amount of horsepower a vehicle has corresponds to vehicles with a greater number of cylinders.*

auto = select(Auto, mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin)  
auto$cylinders = as.numeric(auto$cylinders)  
auto$year = as.numeric(auto$year)  
auto$origin = as.numeric(auto$origin)  
correlation = cor(auto)  
corrplot(correlation)



***Answer:*** *We see a that year is most correlated with miles per gallon on a vehicle, that is newer vehicles are associated with better gas mileage. Furthermore, cylinders is most correlated with displacement, horsepower, and weight.*

### (F) 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.

***Answer:*** *We would expect from the plots above, that weight, displacement, horsepower, and year to be good predictors of mpg. From calculating the correlations amongst the quantitative variables we get the Pearson correlation coefficients for each variable. We see that cylinders(-.78), displacement(-0.81), horespower(-0.78) and weight(-0.83) are negatively correlated with mpg and year (0.58) is somewhat possitively correlated with mpg. The correlation coefficient for each variable with mpg is displayed below and is mentioned within parentheses in the previous sentence.*

*However we see that cylinders, displacement,weight, and horsepower are highly correlated, thus it would be best to use one of these four variables in a predictive model, if we are concerned with multicollinearity. Out of the four, weight should be used due to the fact that it has the highest correlation with mpg of the four. Year and origin should also be used due to the fact that they have correlation coefficients with mpg greater than (.5).*

correlation <- round(correlation,digits = 2)  
Q <- as.table(correlation, nrow(9), ncol(9))  
  
Q

## mpg cylinders displacement horsepower weight acceleration  
## mpg 1.00 -0.78 -0.81 -0.78 -0.83 0.42  
## cylinders -0.78 1.00 0.93 0.80 0.89 -0.46  
## displacement -0.81 0.93 1.00 0.90 0.93 -0.54  
## horsepower -0.78 0.80 0.90 1.00 0.86 -0.69  
## weight -0.83 0.89 0.93 0.86 1.00 -0.42  
## acceleration 0.42 -0.46 -0.54 -0.69 -0.42 1.00  
## year 0.58 -0.33 -0.37 -0.42 -0.31 0.29  
## origin 0.57 -0.59 -0.61 -0.46 -0.59 0.21  
## year origin  
## mpg 0.58 0.57  
## cylinders -0.33 -0.59  
## displacement -0.37 -0.61  
## horsepower -0.42 -0.46  
## weight -0.31 -0.59  
## acceleration 0.29 0.21  
## year 1.00 0.18  
## origin 0.18 1.00