CPT\_S 575 Data Science: Assignment 2

Reet Barik

September 9, 2019

## Excercise 1

1. 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. Call the loaded data college. Ensure that your column headers are not treated as a row of data.

college = read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/College.csv")  
head(college)

## X Private Apps Accept Enroll Top10perc  
## 1 Abilene Christian University Yes 1660 1232 721 23  
## 2 Adelphi University Yes 2186 1924 512 16  
## 3 Adrian College Yes 1428 1097 336 22  
## 4 Agnes Scott College Yes 417 349 137 60  
## 5 Alaska Pacific University Yes 193 146 55 16  
## 6 Albertson College Yes 587 479 158 38  
## Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD  
## 1 52 2885 537 7440 3300 450 2200 70  
## 2 29 2683 1227 12280 6450 750 1500 29  
## 3 50 1036 99 11250 3750 400 1165 53  
## 4 89 510 63 12960 5450 450 875 92  
## 5 44 249 869 7560 4120 800 1500 76  
## 6 62 678 41 13500 3335 500 675 67  
## Terminal S.F.Ratio perc.alumni Expend Grad.Rate  
## 1 78 18.1 12 7041 60  
## 2 30 12.2 16 10527 56  
## 3 66 12.9 30 8735 54  
## 4 97 7.7 37 19016 59  
## 5 72 11.9 2 10922 15  
## 6 73 9.4 11 9727 55

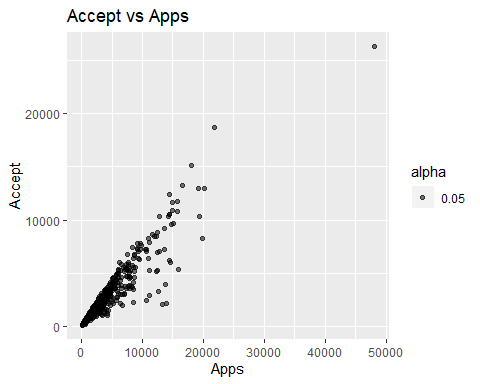
1. Find the median cost of books for all schools in this dataset.

books = summary(college$Books)  
books['Median']

## Median   
## 500

1. Produce a scatterplot that shows a relationship between two features of your choice in the dataset. Ensure it has appropriate axis labels and a title.

Relationship between Acceptence and Applications :



1. Produce a histogram showing the overall enrollment numbers (P.Undergrad plus F.Undergrad) for both public and private (Private) schools. Ensure it has appropriate axis labels and a title.

Adding both the fields P.Undergrad and F.undergrad gives us the overall enrollment

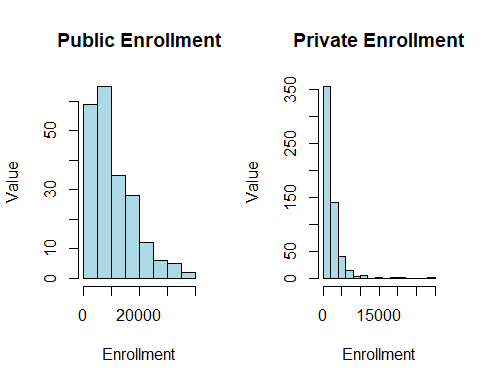
enrollTotal = college$P.Undergrad+college$F.Undergrad

Splitting public and private colleges

pub = which(college$Private=="No")  
pri = which(college$Private =="Yes")

Overall Enrollment plots

par(mfcol = c(1,2))  
hist(enrollTotal[pub], col="light Blue", main="Public Enrollment", xlab="Enrollment", ylab="Value")  
hist(enrollTotal[pri], col="light Blue", main="Private Enrollment", xlab="Enrollment", ylab="Value")



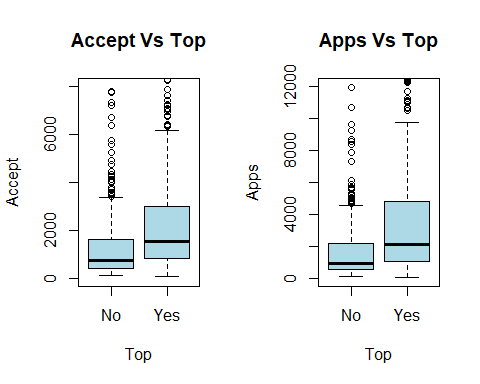
1. Create a new qualitative variable, called Top, by binning the Top25perc variable into two categories. Specifically, divide the schools into two groups based on whether or not the proportion of students coming from the top 25% of their high school classes exceeds 50%. Now produce side-by-side boxplots of acceptance rate (based on Accept and Apps) with respect to the two Top categories (Yes and No). How many top universities are there?

top = rep("No",nrow(college))  
top[college$Top25perc > 50] = "Yes"  
top= as.factor(top)  
college = data.frame(college, top)  
summary(college$top)

## No Yes   
## 328 449

Acceptance and Applications for Top

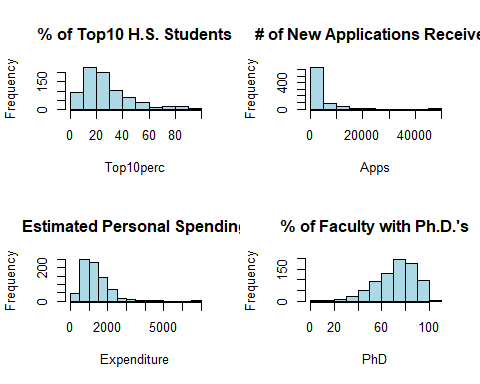
par(mfcol = c(1,2))  
  
topUni = boxplot(college$Accept ~ college$top, col = "light blue", main = "Accept Vs Top", xlab = "Top", ylab = "Accept", ylim = c(0, 8000))  
boxplot(college$Apps ~ college$top, col = "light blue", main = "Apps Vs Top", xlab = "Top", ylab = "Apps", ylim = c(0, 12000))



From the above, it is observed that number of top universities are 449.

1. Continue exploring the data, producing two or more new plots of any type, and provide a brief summary of your hypotheses and what you discover. You may use additional plots or numerical descriptors as needed. 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.

The following 4 histograms show us the frequency distrbution over the variables ‘Top10perc’, ‘Apps’, ‘Personal’ and ‘PhD’. This gives us some idea of the demographic of the total college population in terms of the mentioned features.



## Excercise 2

Handling missing values using na.strings parameter and na.omit function

auto = read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/Auto.csv",  
na.strings = "?")  
auto <- na.omit(auto)  
head(auto)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

1. Specify which of the predictors are quantitative, and which are qualitative? 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 varibles are numeric while qulitative variables are descriptions, which categorizes the data

* Quantitative variables
  + mpg
  + Cylinders
  + Displacement
  + Horsepower
  + Weight
  + Acceleration
  + Year
  + Origin
* Qualitative variables
  + Name

1. What is the range, mean and standard deviation of each quantitative predictor?

Range:

sapply(auto[, -9], range)

## mpg cylinders displacement horsepower weight acceleration year  
## [1,] 9.0 3 68 46 1613 8.0 70  
## [2,] 46.6 8 455 230 5140 24.8 82  
## origin  
## [1,] 1  
## [2,] 3

Mean:

sapply(auto[, -9], mean)

## mpg cylinders displacement horsepower weight   
## 23.445918 5.471939 194.411990 104.469388 2977.584184   
## acceleration year origin   
## 15.541327 75.979592 1.576531

Standard Deviation:

sapply(auto[, -9], sd)

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

1. Now remove the 45th through 85th (inclusive) observations from the dataset. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

auto\_d = auto[-c(45:85), -9]

Range:

sapply(auto\_d, range)

## mpg cylinders displacement horsepower weight acceleration year  
## [1,] 9.0 3 68 46 1649 8.0 70  
## [2,] 46.6 8 455 230 5140 24.8 82  
## origin  
## [1,] 1  
## [2,] 3

Mean:

sapply(auto\_d, mean)

## mpg cylinders displacement horsepower weight   
## 23.780057 5.470085 194.048433 103.863248 2977.233618   
## acceleration year origin   
## 15.541880 76.475783 1.578348

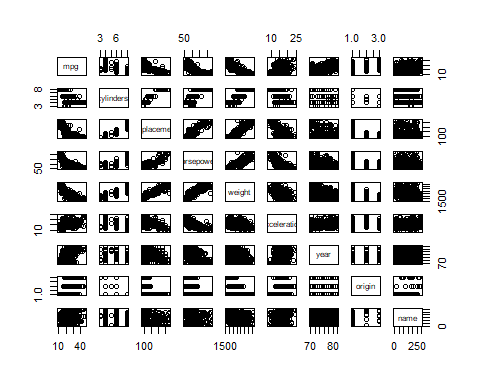
Standard Deviation:

sapply(auto\_d, sd)

## mpg cylinders displacement horsepower weight   
## 7.9008789 1.6830550 103.2050688 38.2367600 835.3627353   
## acceleration year origin   
## 2.7525751 3.5735313 0.8099302

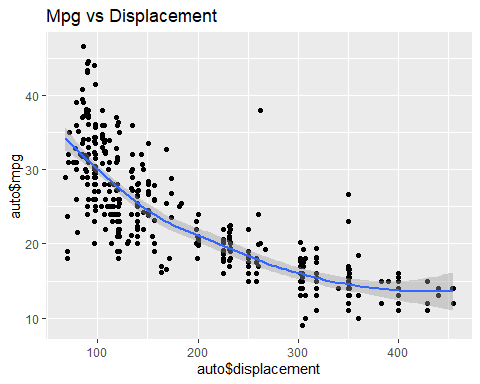
1. Using the full data set, investigate the predictors graphically, using scatterplots, correlation scores or other tools of your choice. Create some plots highlighting the relationships you find among the predictors. Explain briefly what the relationships between variables are, and what they mean.

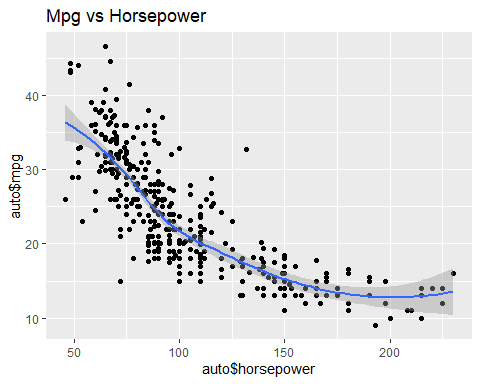
pairs(auto)

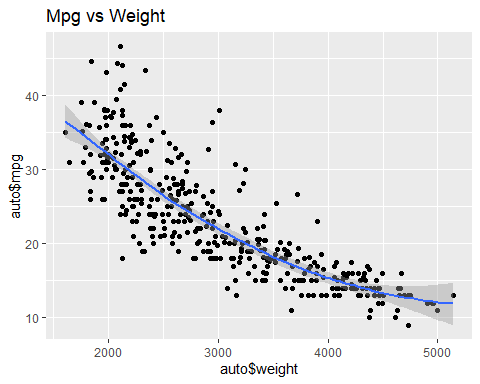


From the scatterplots, we find that there might be a relationship between the following features: \* mpg vs displacement \* mpg vs horsepower \* mpg vs weight \* weight vs horsepower \* weight vs displacement \* horsepower vs displacement \* acceleration vs horsepower

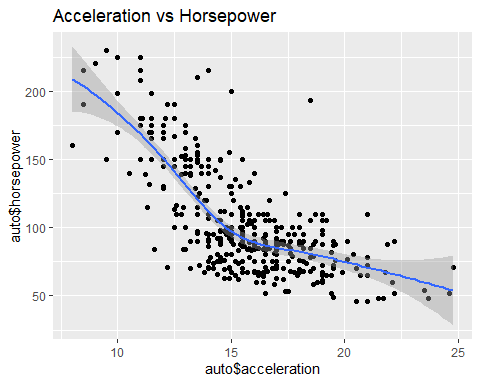
Taking a closer look at the dependence between ‘mpg’ and other features:







From the following plot, we also see that acceleration and horsepower are inversely proportional:



This seems to follow basic physics which says that at lower gears, where horsepower is more, acceleration is less.

1. 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.

Horsepower, cylinders, year and origin can be used as predictors for mpg. Displacement and Weight can not be used as they are highly correlated to each other and to horsepower as seen from the scatterplot.