**STATS 202 | HW:1 | Sagar Ganapaneni |** SUID# 06167633

**Problem: 1**

**Background:**

Error: The difference between actual or true value and predicted value derived from model.

Error term has two components:

1. Reducible error
   1. consists of variance and bias, need to strike a balance between these components to produce a high performance model
   2. Variance and bias depends on the Model flexibility (or) complexity and actual data nature whether it is more linear or non linear
2. Irreducible error
   1. due to unknown factors, uncaptured data, unpredictable factors
   2. nothing we can do much to reduce this component

Our main objective should be reducing the MSE of test data rather than train data as overfitting training data leads to include noise from train data which is not present in test data set.

1. **Large Sample size *n* &Small number of Predictors *p*:**

More Flexible model fits well with large sample size hence flexible model performance better than inflexible model

1. **Small Sample size *n* &Large number of Predictors *p*:**

With small sample size, flexible model over fits the training data which leads to poor performance with test data sets. Hence flexible model performs worse than inflexible model

1. **Relationship between predictors and response is highly non-linear:**

With highly non-linear data more flexible model will account for non-linearity in the data whereas inflexible model might lead to poor fit to the data hence Flexible model performance better than Inflexible model

1. ***σ*2 = Var(ɛ), is extremely high:**

This means high noise in the train data, a flexible method could over fit the data and include this noise in the model which is not present in the test data. So Inflexible method performs better then inflexible method

**Problem: 2**

1. **Prediction:** predict an event or outcome value (Y) based on the data in hand (X) by computing = (X).

(X) could be a black box, we are most interested in the outcome variable than understanding f.

1. **Inference**: some times our goal may not be necessarily to make prediction instead we want to understand the relationship between X and Y or more specifically how Y varies with changes in X. with this we can find out the exact relationship between response and each predictor.

**a)**

1. Data set of 500 forms 🡺 sample size n= 500
2. Predictors: profit, number of employees, industry 🡺 p = 3
3. Output variable, CEO’s Salary, a quantitative variable 🡺 Scenario: Regression
4. We are most interested the relationship between predictors and output 🡺 Inference

**b)**

1. Data of 20 similar products launched previously 🡺 sample size n= 20
2. Predictors: price charged for the product, marketing budget, competition price, and ten other variables 🡺

p = 13

1. Output variable, whether new product will be a success or a failure, a categorical variable

🡺 Scenario: Classification

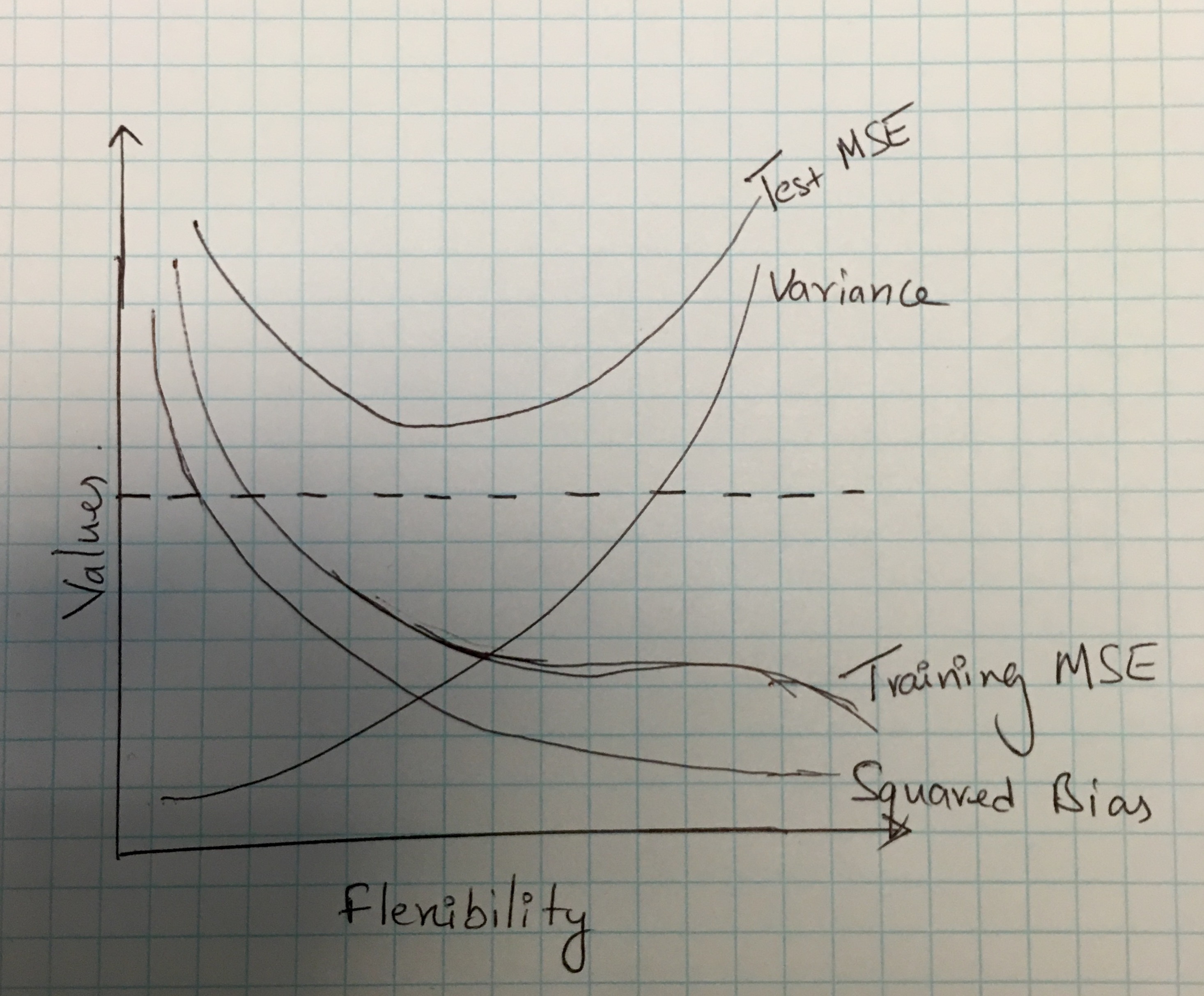
1. We are only interested in output 🡺 Prediction

**c)**

1. Weekly data for all of 2012: 53 weeks🡺 sample size n= 53
2. Predictors: % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.🡺 p = 3
3. Response variable, % change in the US dollar, a quantitative variable 🡺 Scenario: Regression
4. We are only interested in predicting output 🡺 Prediction

**Problem: 3**

**a)**



**b)**

**Variance:** measures how much changes as we change the training data set

**Bias:** measure how far the estimated value of (x) from the actual or true f(x)

**Test MSE = Var () + [Bias2 + Var(ϵ0)**

**Variance:** increases monotonically as flexibility increases. more flexible fit contains noise from the train data. As the training data changes changes hence higher variance.

**Squared bias:** declines monotonically as flexibility increases. Inflexible models approximate the relationship between the variable to greater extent.

**Var(ϵ), the irreducible error** is constant unless we add new data or come up with new predictors

**Test MSE** declines at first, because as flexibility increases the bias decreases. However, increased flexibility leads

to increased variance, so at some point the benefits of decreasing bias are outweighed by the variance, which

**Problem 4:**

Bias Variance Decomposition Equation:

E = Var((*x*0)) + +Var(*ɛ0*)

**Variance:** measures how much changes as we change the training data set

**Bias:** measure how far the estimated value of (x) from the actual or true f(x)

b) We cannot estimate the ***bias*** component as we don’t have the true f(x).

c) We can estimate the ***variance*** component by using already simulated models, find out corresponding estimated y0 and measure the variance.

d) we cannot estimate the variance in *ɛ0*  as y0 = *ɛ0* and we cannot simulate y0

a) We cannot compute TEST MSE as well as we don’t know true

To summarize, with unknown , we can only estimate the variance component.

**Problem 5:**

*## Read data*

college <-read.csv('College.csv', header=TRUE)

*## view data*

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

*## include college names in to the datafarame*

rownames (college )<- college [,1]

college <- college [,-1]

*## Summary*

summary(college)

## X Private Apps

## Abilene Christian University: 1 No :212 Min. : 81

## Adelphi University : 1 Yes:565 1st Qu.: 776

## Adrian College : 1 Median : 1558

## Agnes Scott College : 1 Mean : 3002

## Alaska Pacific University : 1 3rd Qu.: 3624

## Albertson College : 1 Max. :48094

## (Other) :771

## Accept Enroll Top10perc Top25perc

## Min. : 72 Min. : 35 Min. : 1.00 Min. : 9.0

## 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00 1st Qu.: 41.0

## Median : 1110 Median : 434 Median :23.00 Median : 54.0

## Mean : 2019 Mean : 780 Mean :27.56 Mean : 55.8

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

## Max. :26330 Max. :6392 Max. :96.00 Max. :100.0

##

## F.Undergrad P.Undergrad Outstate Room.Board

## Min. : 139 Min. : 1.0 Min. : 2340 Min. :1780

## 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320 1st Qu.:3597

## Median : 1707 Median : 353.0 Median : 9990 Median :4200

## Mean : 3700 Mean : 855.3 Mean :10441 Mean :4358

## 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925 3rd Qu.:5050

## Max. :31643 Max. :21836.0 Max. :21700 Max. :8124

##

## Books Personal PhD Terminal

## Min. : 96.0 Min. : 250 Min. : 8.00 Min. : 24.0

## 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00 1st Qu.: 71.0

## Median : 500.0 Median :1200 Median : 75.00 Median : 82.0

## Mean : 549.4 Mean :1341 Mean : 72.66 Mean : 79.7

## 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00 3rd Qu.: 92.0

## Max. :2340.0 Max. :6800 Max. :103.00 Max. :100.0

##

## S.F.Ratio perc.alumni Expend Grad.Rate

## Min. : 2.50 Min. : 0.00 Min. : 3186 Min. : 10.00

## 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751 1st Qu.: 53.00

## Median :13.60 Median :21.00 Median : 8377 Median : 65.00

## Mean :14.09 Mean :22.74 Mean : 9660 Mean : 65.46

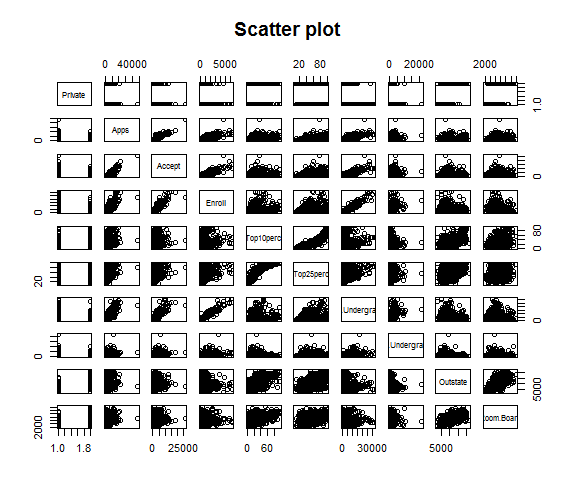
## 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830 3rd Qu.: 78.00

## Max. :39.80 Max. :64.00 Max. :56233 Max. :118.00

##

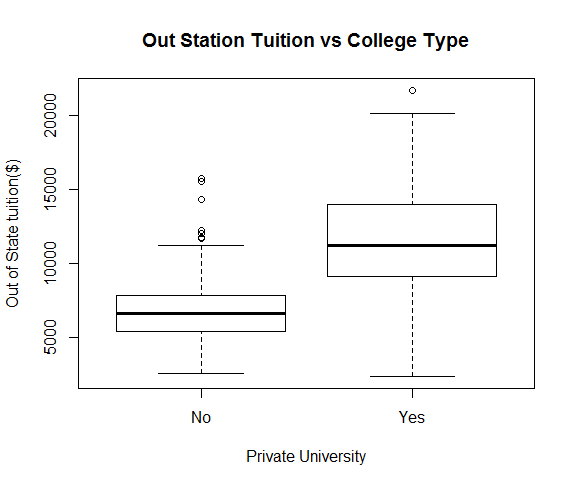
*## Use the pairs () function to produce a scatterplot matrix of the first ten columns or variables of the data.*

pairs(college[, 1:10], main='Scatter plot')



*## Use the plot() function to produce side-by-side boxplots of ‘Outstate’ vs. ‘Private’*

plot(college$Private, college$Outstate, xlab = "Private University", ylab ="Out of State tuition($)", main = "Out Station Tuition vs College Type")



* *Median Out of station Tuition fee is higher for Private colleges compared to Public colleges*

*## 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 exceed 50%.*

college$Elite<-'No'

college[college$Top10perc > 50,]$Elite<-'Yes'

college$Elite<- as.factor(college$Elite)

summary(college$Elite)

## No Yes

## 699 78

*##Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables*

par(mfrow = c(2,2))

hist(college$Expend, col = 'red', xlab = "Expenditure", ylab = "Count",

main='Histogram for Expenditure')

hist(college$PhD, col = 'blue', xlab = "PhD", ylab = "Count",

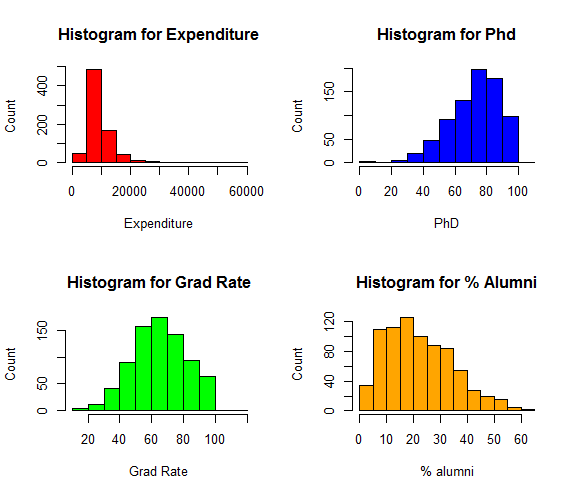
main='Histogram for Phd')

hist(college$Grad.Rate, col = 'green', xlab = "Grad Rate", ylab = "Count",

main='Histogram for Grad Rate')

hist(college$perc.alumni, col = 'orange', xlab = "% alumni", ylab = "Count",

main='Histogram for % Alumni')

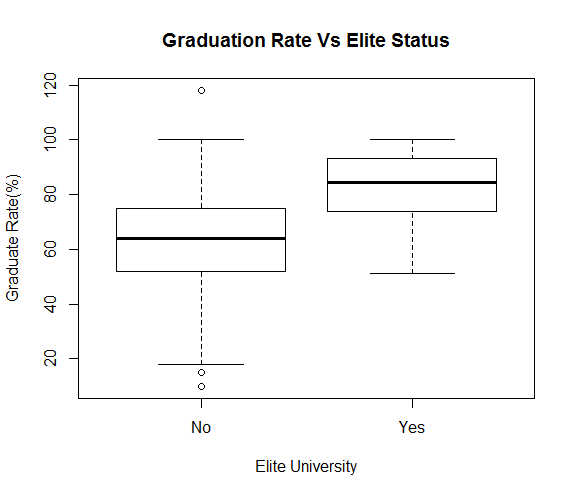


* % Phd Professors metric is negatively skewed
* % Alumni metric is positively skewed

*## further data exploration*

1. If we look at summary of Grad. Rate & PhD columns the max value is greater than 100, this could be an issue with data entry
2. Median Graduation rate is higher in Elite Colleges compared to non- Elite colleges and there is an outlier in the non-Elite college data

plot(college$Elite, college$Grad.Rate, xlab = "Elite University", ylab ="Graduate Rate(%)", main = "Graduation Rate Vs Elite Status")

.

1. Median Acceptance rate is lower Elite Colleges compared to non- Elite colleges

college$acceptance\_rate <- (college$Accept/college$Apps)\*100

plot(college$Elite, college$acceptance\_rate, xlab = "Elite University", ylab ="Acceptance Rate(%)", main = "Acceptance Rate Vs Elite Status")

