320 Final Project

Amy Vuong, Katheryn Fox, Shawdi Hootkany

5/22/2019

# Introduction

In the US, attending college is a true financial investment. Students take out hundreds of thousands of dollars in student loans, to attend school and earn their degree. According to [studentloanhero.com](https://studentloanhero.com/student-loan-debt-statistics/), 69% of the graduating class of 2018 left school with student loans, with an average debt of $29,800. Altogether, American citizens “owe over $1.56 trillion in student loan debt, spread out among about 45 million borrowers” ([studentloanhero.com](https://studentloanhero.com/student-loan-debt-statistics/)).

This crisis centers around money: paying for college as well as the ability to pay off student debt after graduation.

To better understand the statistics and relationships between factors of these costs and salaries, we found a dataset that could provide financial information about students and costs as well as extensive information about the schools themselves. This provides greater insight into the possible driving factors behind published statistics like median starting salary and school tuition prices.

### The Dataset

Utilizing the publicly available and widely trusted source, [U.S. News 2019 Best National Univeristy](https://www.usnews.com/best-colleges/rankings/national-universities) ranking list, we were able to access information about top schools that was organized in a digestable and scrapable manner.

We obtained information regarding 12 different variables (using the URL and School Name to uniquely identify the Univerisities):

* Tuition Cost (in USD - out of state tuition used if university is a state school)
* Room & Board (in USD)
* Total Enrollment
* School Type (Private/Public)
* Year Founded
* Setting
* Endowment 2017
* Median Starting Salary Of Alumni (3 years postgraduate)
* Selectivity
* Fall 2017 Acceptance Rate
* Male Percentage
* Four Year Graduation Rate

We chose to use the top 110, removing one row, because of data inconsistency, as it did not provide a Median Starting Salary for Alumni. This was the Univerity of California - Santa Barbara.

### Research Questions

* What variables are statistically significant in predicting college tuition prices?
  + Which factors seem to drive the biggest increases?
  + Does a higher endowment mean that students have to spend less on tuition? (i.e. does it appear that schools use their endowments to supplement tuition prices at all?)
* What variables are statistically significant in predicting median starting salary for alumni (with 3 years of experience post-grad)?
  + Is the cost of undergraduate tuition a key factor? (i.e. spending more on education correlated with getting more money later?)
  + Is median starting salary correlated with the male to female ratio at the institution in any kind of way?

### Required Tools

You will use R and the following libraries:

* ggplot2
* rvest
* tidyverse
* stringr

# Part 1: Data Scraping

The main url we will be using contains very limited information about each of the schools, such as ranking and tuition, therefore the first step that needs to be taken to be able to gain all the information we need to be able to analyze the data and make predictions is to parse the information into readable data. Detailed information about each school is spread across multiple websites so we will need to retrieve the proper url for each university from the US News website containing the ranking and then parse important information into tables that could be used for data analysis.

We are scraping the data of 100 schools from <https://www.usnews.com/best-colleges/rankings/national-universities>. The data we have is stored in a text file since it loads on the page in increments. We parse the data to find the URL for each college’s informational page.

Note: The information for the University of California–Davis was removed from the dataset because it didn’t contain median alumni salary, which plays a large role in our analysis.

Note: Room and Board, Tuition and Fees, and Median Alumni Salary are all in thousands of dollars.

library(rvest)  
library(tidyverse)  
url <-"html\_top100.txt"  
college\_urls <- url %>%  
 read\_html() %>%  
 html\_node("body") %>% html\_nodes("ol[class~=bEyEue]") %>% html\_nodes("li[id]")%>% html\_nodes("h3") %>%   
 html\_nodes("a[href]") %>%  
 html\_attr("href")   
head(college\_urls)

## [1] "/best-colleges/princeton-university-2627"   
## [2] "/best-colleges/harvard-university-2155"   
## [3] "/best-colleges/columbia-university-2707"   
## [4] "/best-colleges/massachusetts-institute-of-technology-2178"  
## [5] "/best-colleges/university-of-chicago-1774"   
## [6] "/best-colleges/yale-university-1426"

A data frame is created to store the information of each college in rows. Columns are initialized with their corresponding default values.

index\_num <- 0  
  
college\_tab\_1 <- data.frame("URL" = gsub(" ", "", paste("https://www.usnews.com",college\_urls, sep = "")),   
"CollegeName"= "", "TuitionFeesThousands" = 0, "RoomBoardThousands" = 0, "TotalEnrollment" = 0, "SchoolType" = "", "YearFounded" = 0, "Setting" = "", "Endowment2017Millions" = 0, "MedianStartingSalaryOfAlumniThousands" = 0, "Selectivity" = "", "Fall2017AcceptanceRate" = 0, "MalePercentage" = 0, "FourYearGraduationRate" = 0, stringsAsFactors = FALSE)   
  
#removing one college that doesn't have a median starting salary, for data uniformity  
college\_tab\_1 <- college\_tab\_1[-c(40),]

Below are functions used to obtain data from the website and parse it.

#retrieves of vector of size three containing the Tuition&Fees, Room&Board, and total enrollment  
get\_info <- function(url\_html){  
 attr <- url\_html %>% html\_node("body") %>% html\_nodes("div[id~=content-main]") %>%   
 html\_nodes("section[class~=hero-stats-widget-stats]") %>%  
 html\_nodes("ul") %>% html\_nodes("li") %>% html\_nodes("strong")  
}  
  
#takes in a vector and index, and parses that information to a double  
#ex: $47,263 -> 47263.0  
get\_tuition\_rm <- function(info, num){  
 a\_1 <- info[num] %>% html\_text()  
 tuition\_rm <-   
 as.double(paste(substring(a\_1, 2, str\_locate(a\_1, ",")[1] - 1), substring(a\_1, str\_locate(a\_1, ",")[1] + 1, str\_locate(a\_1, " ")[1] - 1), sep=""))  
 tuition\_rm / 1000.0  
}  
  
#takes in a vector and parses the total enrollment information to a double  
get\_enrollment <- function(info){  
 a\_1 <- info[3] %>% html\_text()  
 as.double(paste(substring(a\_1, 1, str\_locate(a\_1, ",")[1] - 1), substring(a\_1, str\_locate(a\_1, ",")[1] + 1), sep=""))  
}  
  
#gets the percentage of the majority gender at a certain university  
get\_percent <- function(url\_html){  
 attr <- url\_html %>% html\_node("body") %>% html\_nodes("div[id~=content-main]") %>%   
 html\_nodes("div[class~=block-normal]") %>% html\_nodes("span[class~=distribution-breakdown\_\_percentage]") %>% html\_text()  
 as.double(substring(attr, 1, str\_locate(attr, "%")[1] - 1)) / 100.0  
}  
  
#retrieves the gender of the majority sex and parses the percentage to be in terms of males  
get\_gender\_ratio <- function(url\_html){  
 attr <- url\_html %>% html\_node("body") %>% html\_nodes("div[id~=content-main]") %>%   
 html\_nodes("div[class~=block-normal]") %>% html\_nodes("span[class~=distribution-breakdown\_\_percentage-copy]") %>% html\_text()  
 attr <- sub("\n ","",attr)  
 attr <- sub("\n ","",attr)  
 if (attr == "Female"){  
 1 - get\_percent(url\_html)  
 }else{  
 get\_percent(url\_html)  
 }  
}

Here, we use both the functions above and the html\_node function to fill out the table.

college\_tab <- college\_tab\_1  
  
for (i in 1:nrow(college\_tab)){  
 url\_html <- college\_tab[i,1] %>%read\_html()  
 college\_tab[i,]$CollegeName <- url\_html %>% html\_node("body") %>% html\_nodes("h1[class~=hero-heading]") %>% html\_text()  
 priv\_tuition <- url\_html %>% html\_node("body") %>% html\_nodes("span[data-test-id~=v\_private\_tuition]") %>% html\_text()  
 college\_tab[i,]$TuitionFeesThousands <- ifelse(length(priv\_tuition) > 0, priv\_tuition,   
 url\_html %>% html\_node("body") %>% html\_node("span[data-test-id~=v\_out\_state\_tuition]") %>% html\_text())  
 college\_tab[i,]$RoomBoardThousands <- url\_html %>% html\_node("body") %>% html\_node("span[data-test-id~=w\_room\_board]") %>% html\_text()  
 college\_tab[i,]$TotalEnrollment <- url\_html %>% html\_node("body") %>% html\_node("span[data-test-id~=total\_all\_students]") %>% html\_text()  
 college\_tab[i,]$MalePercentage <- get\_gender\_ratio(url\_html)  
 college\_tab[i,]$Fall2017AcceptanceRate <- url\_html %>% html\_node("span[data-test-id~=r\_c\_accept\_rate]") %>% html\_text()  
 college\_tab[i,]$Selectivity <- url\_html %>% html\_node("span[data-test-id~=c\_select\_class]") %>% html\_text()  
 college\_tab[i,]$FourYearGraduationRate <- url\_html %>% html\_node("span[data-test-id~=grad\_rate\_4\_year]") %>% html\_text()  
 college\_tab[i,]$MedianStartingSalaryOfAlumniThousands <- url\_html %>% html\_nodes("div[data-field-id=averageStartSalary]") %>%html\_node("span[data-test-id]") %>% html\_text()  
 temp\_vector <- url\_html %>% html\_node("body") %>% html\_nodes("div[id~=content-main]") %>%html\_nodes("div[class~=flex-row]") %>% html\_nodes("span[class~=heading-small]") %>% html\_text()  
 college\_tab[i,]$SchoolType <- temp\_vector[1]  
 college\_tab[i,]$YearFounded <- temp\_vector[2]  
 college\_tab[i,]$Setting <- temp\_vector[5]  
 college\_tab[i,]$Endowment2017Millions <- temp\_vector[6]  
}  
  
head(college\_tab)

## URL  
## 1 https://www.usnews.com/best-colleges/princeton-university-2627  
## 2 https://www.usnews.com/best-colleges/harvard-university-2155  
## 3 https://www.usnews.com/best-colleges/columbia-university-2707  
## 4 https://www.usnews.com/best-colleges/massachusetts-institute-of-technology-2178  
## 5 https://www.usnews.com/best-colleges/university-of-chicago-1774  
## 6 https://www.usnews.com/best-colleges/yale-university-1426  
## CollegeName  
## 1 \n Princeton University\n   
## 2 \n Harvard University\n   
## 3 \n Columbia University\n   
## 4 \n Massachusetts Institute of Technology\n   
## 5 \n University of Chicago\n   
## 6 \n Yale University\n   
## TuitionFeesThousands RoomBoardThousands  
## 1 \n $47,140 (2018-19) \n $15,610 (2018-19)  
## 2 \n $50,420 (2018-19) \n $17,160 (2018-19)  
## 3 \n $59,430 (2018-19) \n $14,016 (2018-19)  
## 4 \n $51,832 (2018-19) \n $15,510 (2018-19)  
## 5 \n $57,006 (2018-19) \n $16,350 (2018-19)  
## 6 \n $53,430 (2018-19) \n $16,000 (2018-19)  
## TotalEnrollment SchoolType YearFounded Setting  
## 1 \n 8,273 Private, Coed 1746 Suburban  
## 2 \n 20,604 Private, Coed 1636 Urban  
## 3 \n 25,968 Private, Coed 1754 Urban  
## 4 \n 11,466 Private, Coed 1861 Urban  
## 5 \n 13,736 Private, Coed 1890 Urban  
## 6 \n 12,974 Private, Coed 1701 City  
## Endowment2017Millions MedianStartingSalaryOfAlumniThousands  
## 1 $23.4 billion \n $68,400\*  
## 2 $37.1 billion \n $66,500\*  
## 3 $10.0 billion \n $64,900\*  
## 4 $14.8 billion + \n $79,800\*  
## 5 $6.6 billion + \n $57,700\*  
## 6 $27.2 billion + \n $63,200\*  
## Selectivity Fall2017AcceptanceRate MalePercentage  
## 1 \n Most selective \n 6% 0.51  
## 2 \n Most selective \n 5% 0.52  
## 3 \n Most selective \n 6% 0.52  
## 4 \n Most selective \n 7% 0.54  
## 5 \n Most selective \n 9% 0.51  
## 6 \n Most selective \n 7% 0.50  
## FourYearGraduationRate  
## 1 \n 89%  
## 2 \n 84%  
## 3 \n 88%  
## 4 \n 85%  
## 5 \n 88%  
## 6 \n 87%

Below, we reformat many of the columns to get usable data. Each column is categorized into the appropriate type of data.

formatted\_college\_tab <- college\_tab  
#fix type of School Type, Setting, Year Founded  
formatted\_college\_tab$SchoolType <- as.factor(formatted\_college\_tab$SchoolType)  
formatted\_college\_tab$Setting <- as.factor(formatted\_college\_tab$Setting)  
formatted\_college\_tab$YearFounded <- as.integer(formatted\_college\_tab$YearFounded)  
#fix Endowment2017 formatting  
formatted\_college\_tab$Endowment2017Millions <- ifelse(grepl("billion", formatted\_college\_tab$Endowment2017Millions ), sub("\\.","",formatted\_college\_tab$Endowment2017Millions ),formatted\_college\_tab$Endowment2017Millions )  
formatted\_college\_tab$Endowment2017Millions <-sub(" billion","00",formatted\_college\_tab$Endowment2017Millions )  
formatted\_college\_tab$Endowment2017Millions <-sub(" million","",formatted\_college\_tab$Endowment2017Millions )  
formatted\_college\_tab$Endowment2017Millions <-sub("[[:punct:]]", "",formatted\_college\_tab$Endowment2017Millions )  
formatted\_college\_tab$Endowment2017Millions <-sub("\\$", "",formatted\_college\_tab$Endowment2017Millions )  
formatted\_college\_tab$Endowment2017Millions <-sub(" \\+", "",formatted\_college\_tab$Endowment2017Millions )  
formatted\_college\_tab$Endowment2017Millions <- as.double(formatted\_college\_tab$Endowment2017Millions)  
#fix College Name formatting  
formatted\_college\_tab$CollegeName <- sub("^\n ","",formatted\_college\_tab$CollegeName)  
formatted\_college\_tab$CollegeName <-sub("\n ","",formatted\_college\_tab$CollegeName)  
#fixing Acceptance Rate formatting  
formatted\_college\_tab$Fall2017AcceptanceRate <- sub("\n ","",formatted\_college\_tab$Fall2017AcceptanceRate)  
formatted\_college\_tab$Fall2017AcceptanceRate <- sub("%","",formatted\_college\_tab$Fall2017AcceptanceRate)  
formatted\_college\_tab$Fall2017AcceptanceRate <- as.double(formatted\_college\_tab$Fall2017AcceptanceRate)  
formatted\_college\_tab$Fall2017AcceptanceRate <- formatted\_college\_tab$Fall2017AcceptanceRate/100  
#fixing Grad Rate formatting  
formatted\_college\_tab$FourYearGraduationRate <- sub("\n ","",formatted\_college\_tab$FourYearGraduationRate)  
formatted\_college\_tab$FourYearGraduationRate <- sub("%","",formatted\_college\_tab$FourYearGraduationRate)  
formatted\_college\_tab$FourYearGraduationRate <- as.double(formatted\_college\_tab$FourYearGraduationRate)  
formatted\_college\_tab$FourYearGraduationRate <- formatted\_college\_tab$FourYearGraduationRate/100  
#fixing Salary formatting  
formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands <-   
 sub("\n ","",formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands)  
formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands <- gsub("\\\*","",formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands)  
formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands <- gsub("\\$","",formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands)  
formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands <- gsub("\\,","",formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands)  
formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands <- as.double(formatted\_college\_tab$MedianStartingSalaryOfAlumniThousands)/1000  
#fixing Selectivity formatting  
formatted\_college\_tab$Selectivity <- sub("\n ","",formatted\_college\_tab$Selectivity)  
formatted\_college\_tab$Selectivity <- as.factor(formatted\_college\_tab$Selectivity)  
#fixing Tuition formatting  
formatted\_college\_tab$TuitionFeesThousands <- sub("\n ", "",formatted\_college\_tab$TuitionFeesThousands )  
formatted\_college\_tab$TuitionFeesThousands <- sub(" \\(2018-19\\)", "",formatted\_college\_tab$TuitionFeesThousands )  
formatted\_college\_tab$TuitionFeesThousands <-sub("\\,", "",formatted\_college\_tab$TuitionFeesThousands )  
formatted\_college\_tab$TuitionFeesThousands <-sub("\\$", "",formatted\_college\_tab$TuitionFeesThousands )  
formatted\_college\_tab$TuitionFeesThousands <- as.double(formatted\_college\_tab$TuitionFeesThousands)/1000

## Warning: NAs introduced by coercion

#fixing RoomBoard formatting  
formatted\_college\_tab$RoomBoardThousands <- sub("\n ", "",formatted\_college\_tab$RoomBoardThousands )  
formatted\_college\_tab$RoomBoardThousands <- sub(" \\(2018-19\\)", "",formatted\_college\_tab$RoomBoardThousands )  
formatted\_college\_tab$RoomBoardThousands <-sub("\\,", "",formatted\_college\_tab$RoomBoardThousands )  
formatted\_college\_tab$RoomBoardThousands <-sub("\\$", "",formatted\_college\_tab$RoomBoardThousands )  
formatted\_college\_tab$RoomBoardThousands <- as.double(formatted\_college\_tab$RoomBoardThousands)/1000

## Warning: NAs introduced by coercion

#fixing Enrollment formatting  
formatted\_college\_tab$TotalEnrollment <- sub("\n ", "",formatted\_college\_tab$TotalEnrollment )  
formatted\_college\_tab$TotalEnrollment <-sub("\\,", "",formatted\_college\_tab$TotalEnrollment )  
formatted\_college\_tab$TotalEnrollment <- as.double(formatted\_college\_tab$TotalEnrollment)  
  
  
formatted\_college\_tab <- formatted\_college\_tab %>% mutate(TotalCostThousands =TuitionFeesThousands + RoomBoardThousands )  
  
formatted\_college\_tab <- na.omit(formatted\_college\_tab)  
nrow(formatted\_college\_tab)

## [1] 107

as.tibble(formatted\_college\_tab)

## Warning: `as.tibble()` is deprecated, use `as\_tibble()` (but mind the new semantics).  
## This warning is displayed once per session.

## # A tibble: 107 x 15  
## URL CollegeName TuitionFeesThou… RoomBoardThousa… TotalEnrollment  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 http… Princeton … 47.1 15.6 8273  
## 2 http… Harvard Un… 50.4 17.2 20604  
## 3 http… Columbia U… 59.4 14.0 25968  
## 4 http… Massachuse… 51.8 15.5 11466  
## 5 http… University… 57.0 16.4 13736  
## 6 http… Yale Unive… 53.4 16 12974  
## 7 http… Stanford U… 51.4 15.8 17178  
## 8 http… Duke Unive… 56.0 15.9 16294  
## 9 http… University… 55.6 15.6 21907  
## 10 http… Johns Hopk… 53.7 15.8 25151  
## # … with 97 more rows, and 10 more variables: SchoolType <fct>,  
## # YearFounded <int>, Setting <fct>, Endowment2017Millions <dbl>,  
## # MedianStartingSalaryOfAlumniThousands <dbl>, Selectivity <fct>,  
## # Fall2017AcceptanceRate <dbl>, MalePercentage <dbl>,  
## # FourYearGraduationRate <dbl>, TotalCostThousands <dbl>

#to save as csv to easily work on it without having to reload  
write.csv(formatted\_college\_tab, file = "college\_info.csv")

formatted\_college\_tab <- read.csv("college\_info.csv")  
formatted\_college\_tab <- formatted\_college\_tab[,-c(1)]  
as.tibble(formatted\_college\_tab)

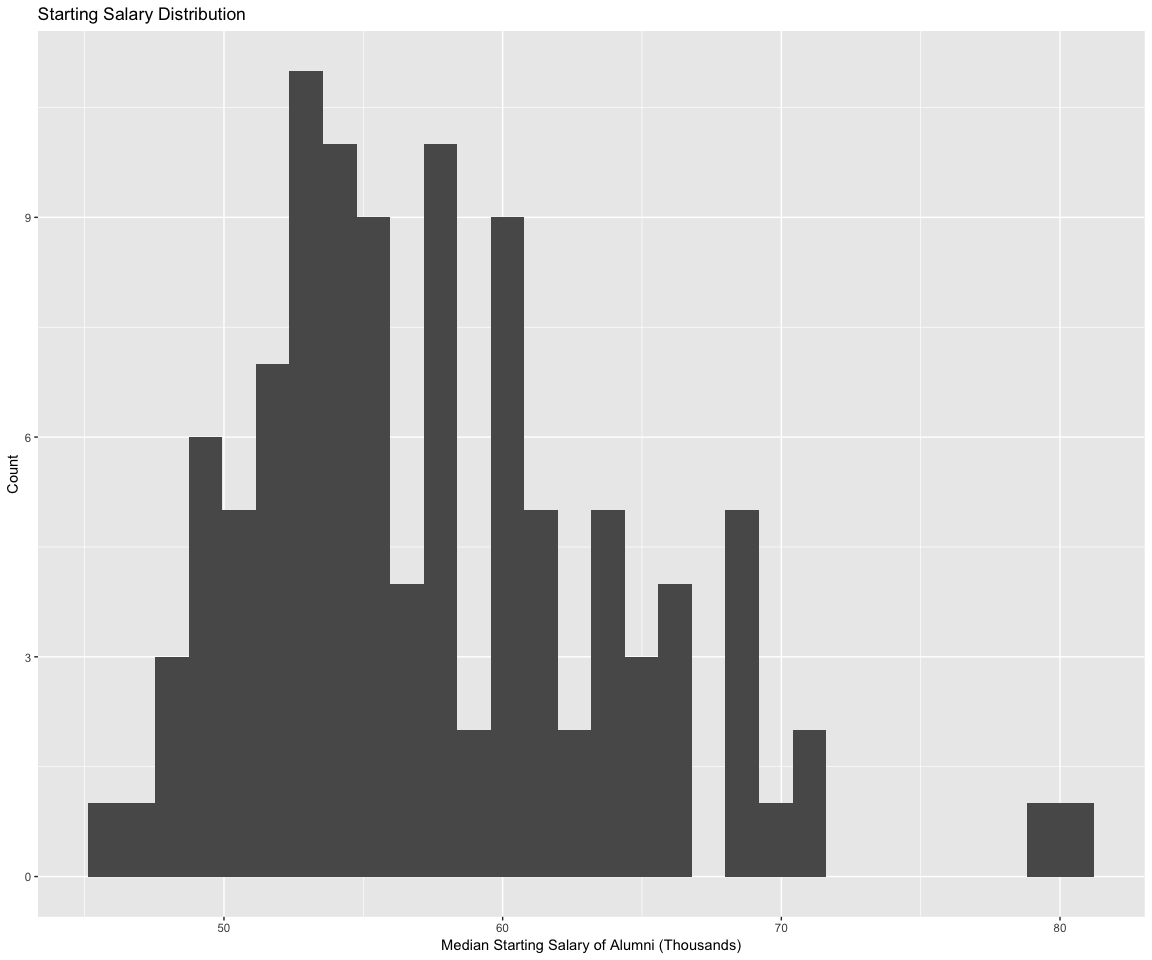
## # A tibble: 107 x 15  
## URL CollegeName TuitionFeesThou… RoomBoardThousa… TotalEnrollment  
## <fct> <fct> <dbl> <dbl> <int>  
## 1 http… Princeton … 47.1 15.6 8273  
## 2 http… Harvard Un… 50.4 17.2 20604  
## 3 http… Columbia U… 59.4 14.0 25968  
## 4 http… Massachuse… 51.8 15.5 11466  
## 5 http… University… 57.0 16.4 13736  
## 6 http… Yale Unive… 53.4 16 12974  
## 7 http… Stanford U… 51.4 15.8 17178  
## 8 http… Duke Unive… 56.0 15.9 16294  
## 9 http… University… 55.6 15.6 21907  
## 10 http… Johns Hopk… 53.7 15.8 25151  
## # … with 97 more rows, and 10 more variables: SchoolType <fct>,  
## # YearFounded <int>, Setting <fct>, Endowment2017Millions <dbl>,  
## # MedianStartingSalaryOfAlumniThousands <dbl>, Selectivity <fct>,  
## # Fall2017AcceptanceRate <dbl>, MalePercentage <dbl>,  
## # FourYearGraduationRate <dbl>, TotalCostThousands <dbl>

# Part 2: Data Visualization

We plot the data in order to visualize relationships among the attributes.

#Starting Salary  
#-histograms  
library(ggplot2)  
plot\_1 <- formatted\_college\_tab %>%  
 ggplot(aes(MedianStartingSalaryOfAlumniThousands)) +  
 geom\_histogram()+   
 labs(title="Starting Salary Distribution", x="Median Starting Salary of Alumni (Thousands)", y="Count")  
plot\_1

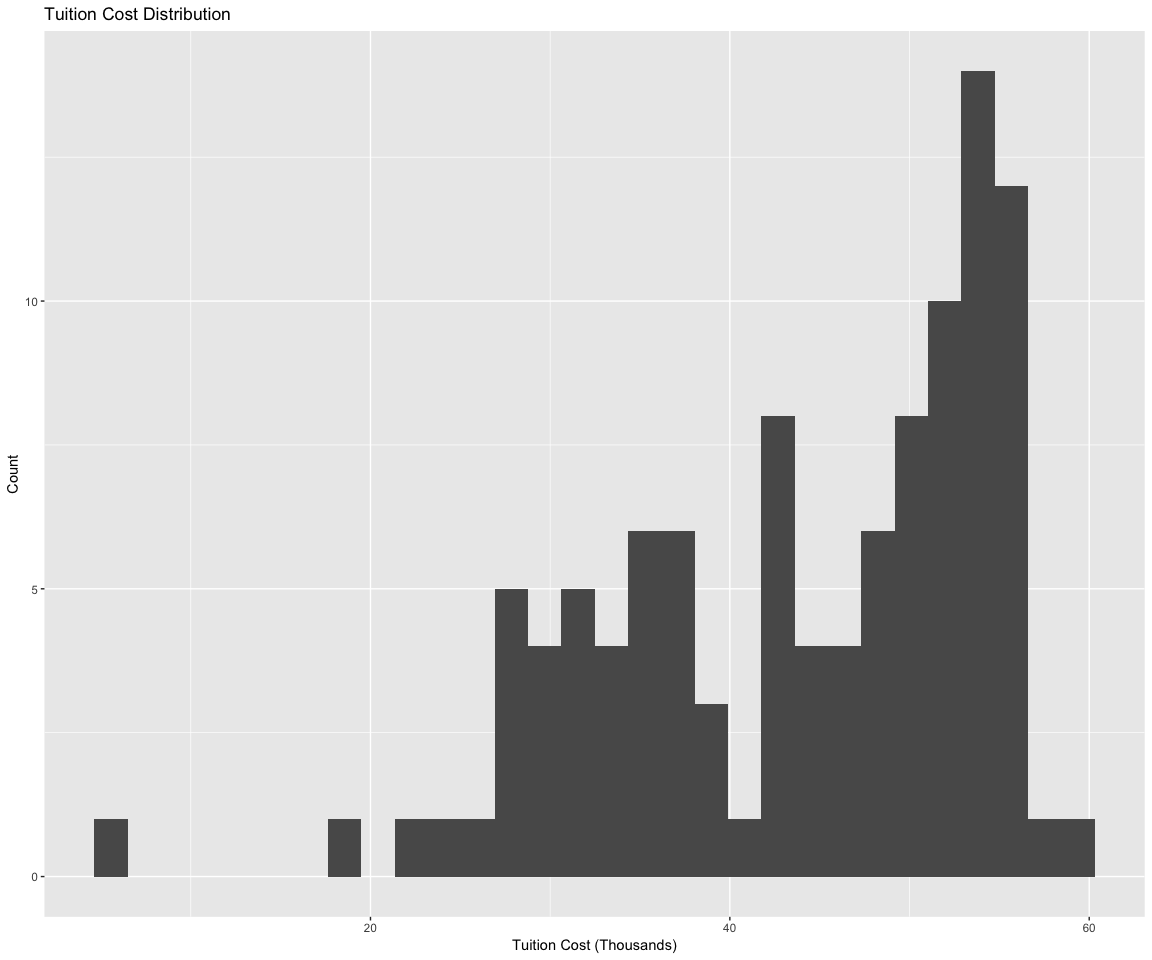
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The distribution of the median starting salary of alumni from all the school seems to be a bell-shaped curve (a little skewed right), centering around $55,000.

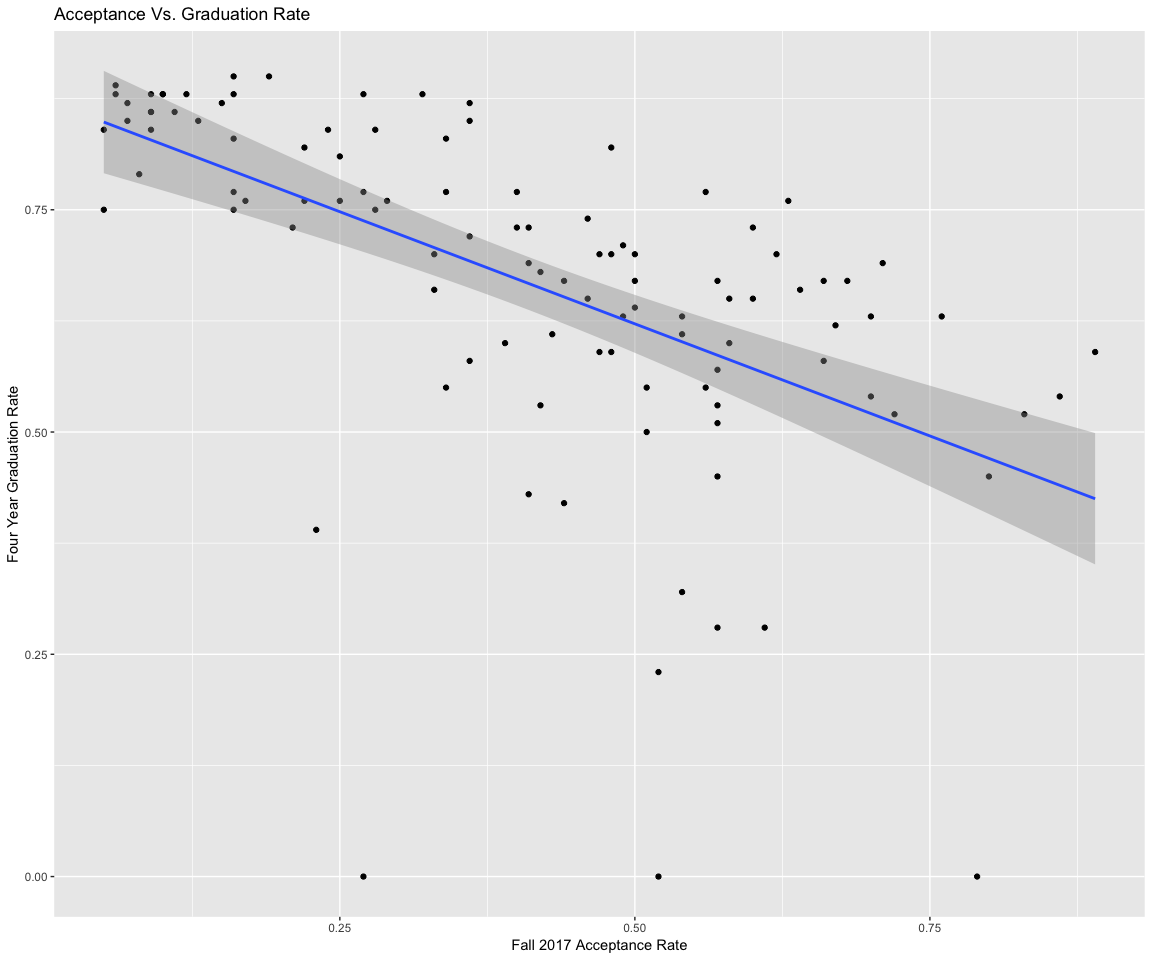
#Tuition Cost  
#-histograms  
library(ggplot2)  
plot\_2 <- formatted\_college\_tab %>%  
 ggplot(aes(TuitionFeesThousands)) +  
 geom\_histogram()+   
 labs(title="Tuition Cost Distribution", x="Tuition Cost (Thousands)", y="Count")  
plot\_2

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

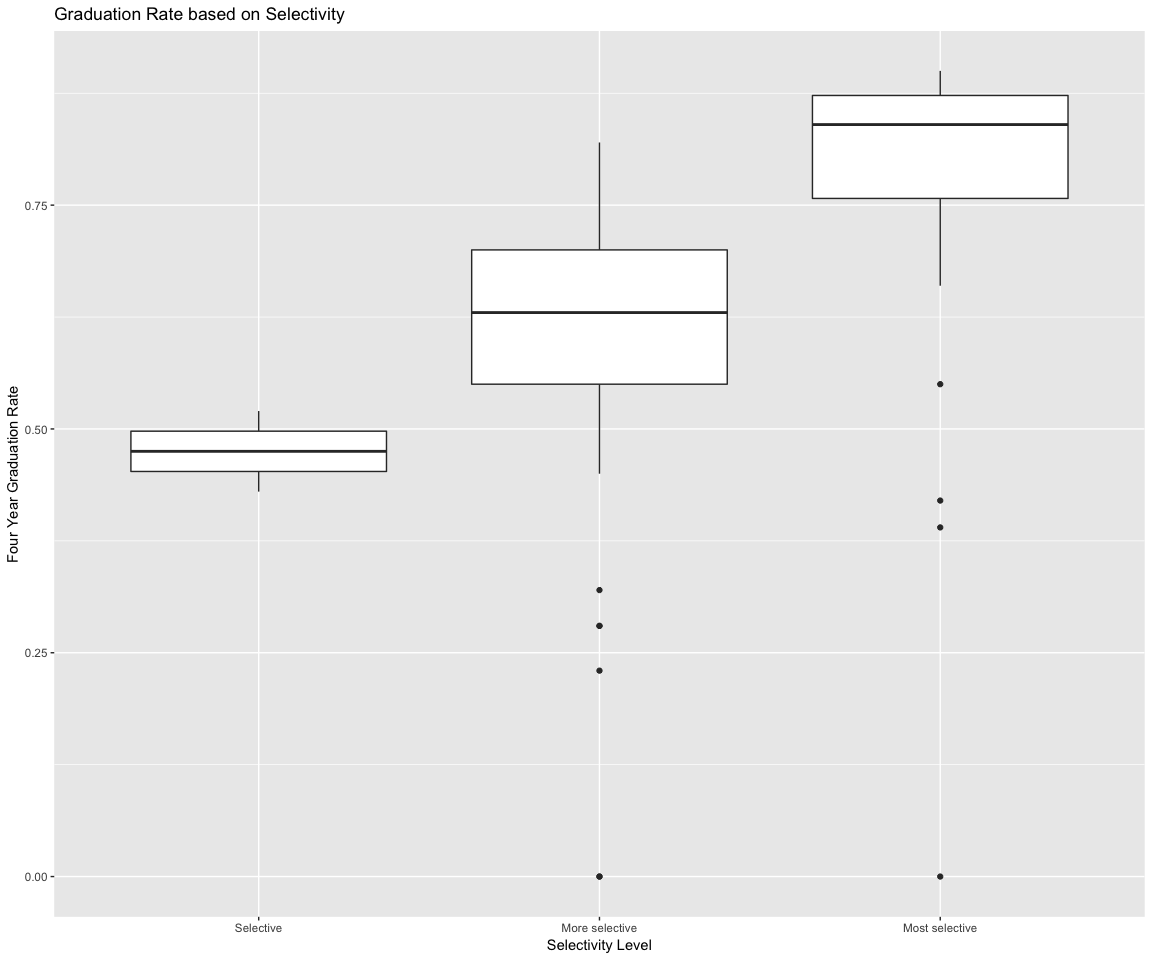


The distribution of tution costs of all the schools is skewed left, with a range of $60,000.

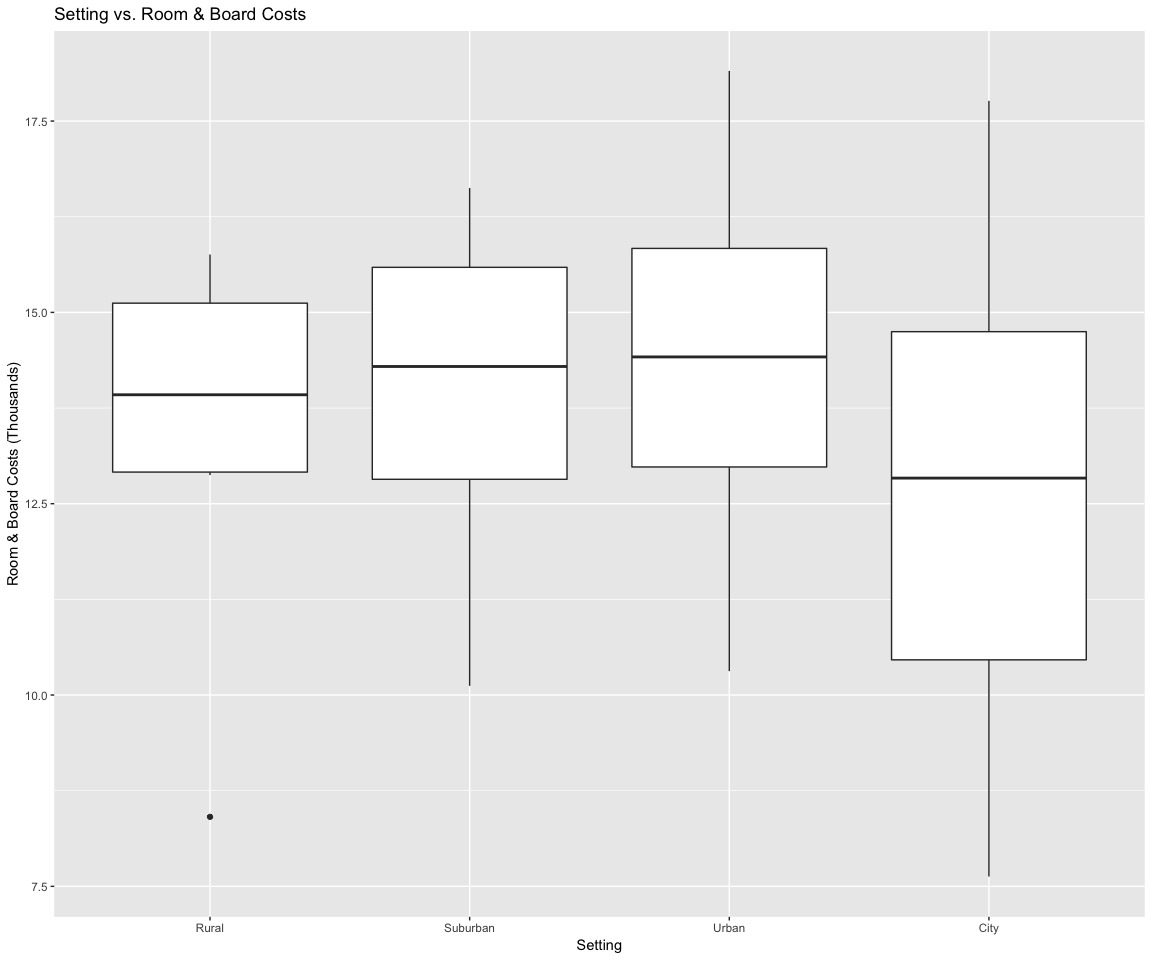
#Acceptance rate vs graduation rate  
  
library(ggplot2)  
plot\_3 <- formatted\_college\_tab %>%  
 ggplot(aes(x=Fall2017AcceptanceRate, y=FourYearGraduationRate)) +  
 geom\_point()+   
 geom\_smooth(method=lm)+  
 labs(title="Acceptance Vs. Graduation Rate", x="Fall 2017 Acceptance Rate", y="Four Year Graduation Rate")  
plot\_3

 There is a linear relationship between acceptance rate (Fall 2017) and the four year graduation rate. It is an overall negative relationship. The higher the acceptance rate, the lower the rate of graduation.

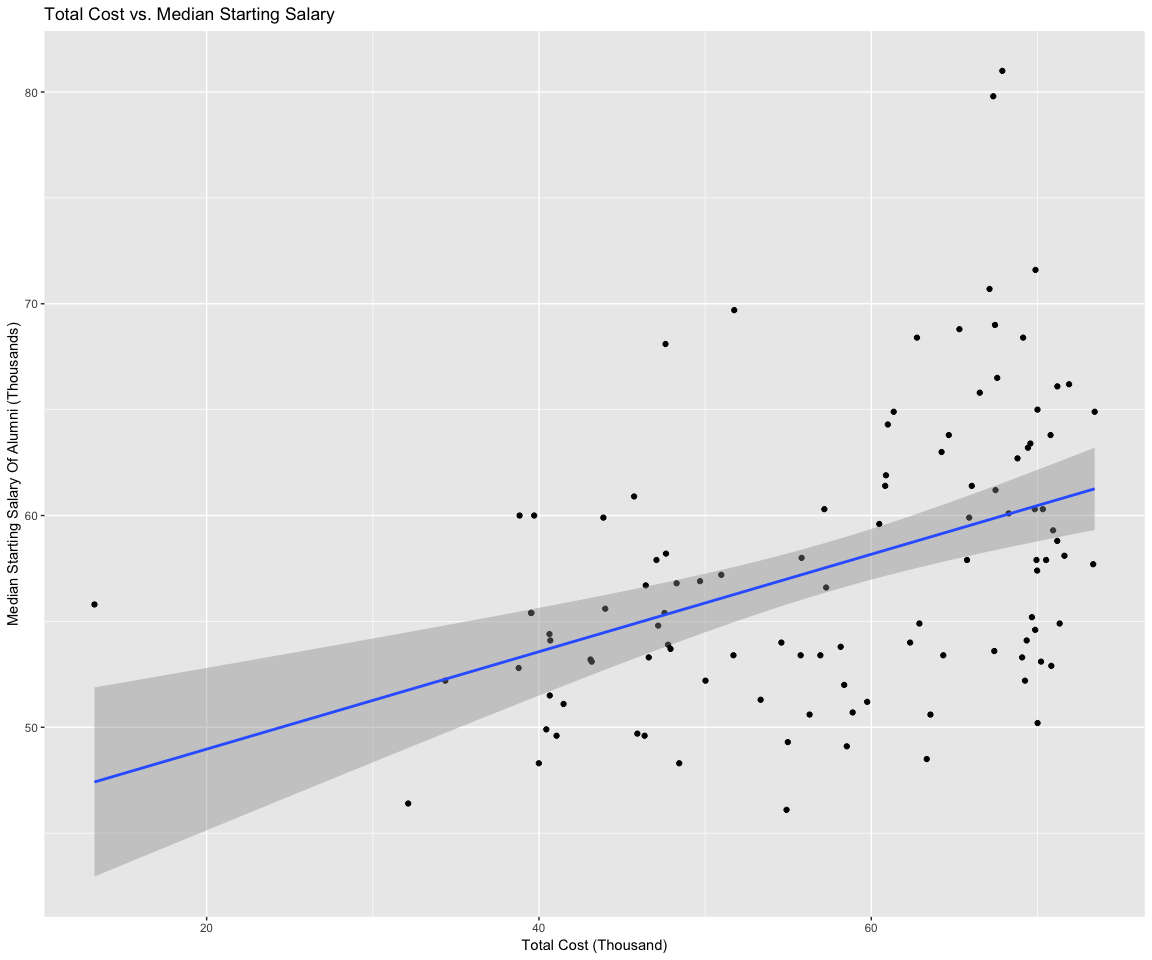
#Boxplots of (1) gradruattion rate & (2) admission rate by selectivity   
library(ggplot2)  
formatted\_college\_tab$Selectivity <- factor(formatted\_college\_tab$Selectivity, c("Selective","More selective","Most selective"))  
plot\_4 <- formatted\_college\_tab %>%  
 ggplot(aes(x=Selectivity, y=FourYearGraduationRate)) +  
 geom\_boxplot()+  
 labs(title="Graduation Rate based on Selectivity", x="Selectivity Level", y="Four Year Graduation Rate")  
plot\_4

 This is significant difference in four year graduation rates based on their Selectivity Level of accepting students. These boxplots show that each 3 selectivity level vary significantly on range and central tendency. The more selective a college is, the greater their graduation rates seem to be.

#Setting vs. room board  
  
library(ggplot2)  
formatted\_college\_tab$Setting <- factor(formatted\_college\_tab$Setting, c("Rural","Suburban","Urban", "City"))  
plot\_5 <- formatted\_college\_tab %>%  
 ggplot(aes(x=Setting, y=RoomBoardThousands)) +  
 geom\_boxplot()+  
 labs(title="Setting vs. Room & Board Costs", x="Setting", y="Room & Board Costs (Thousands)")  
plot\_5

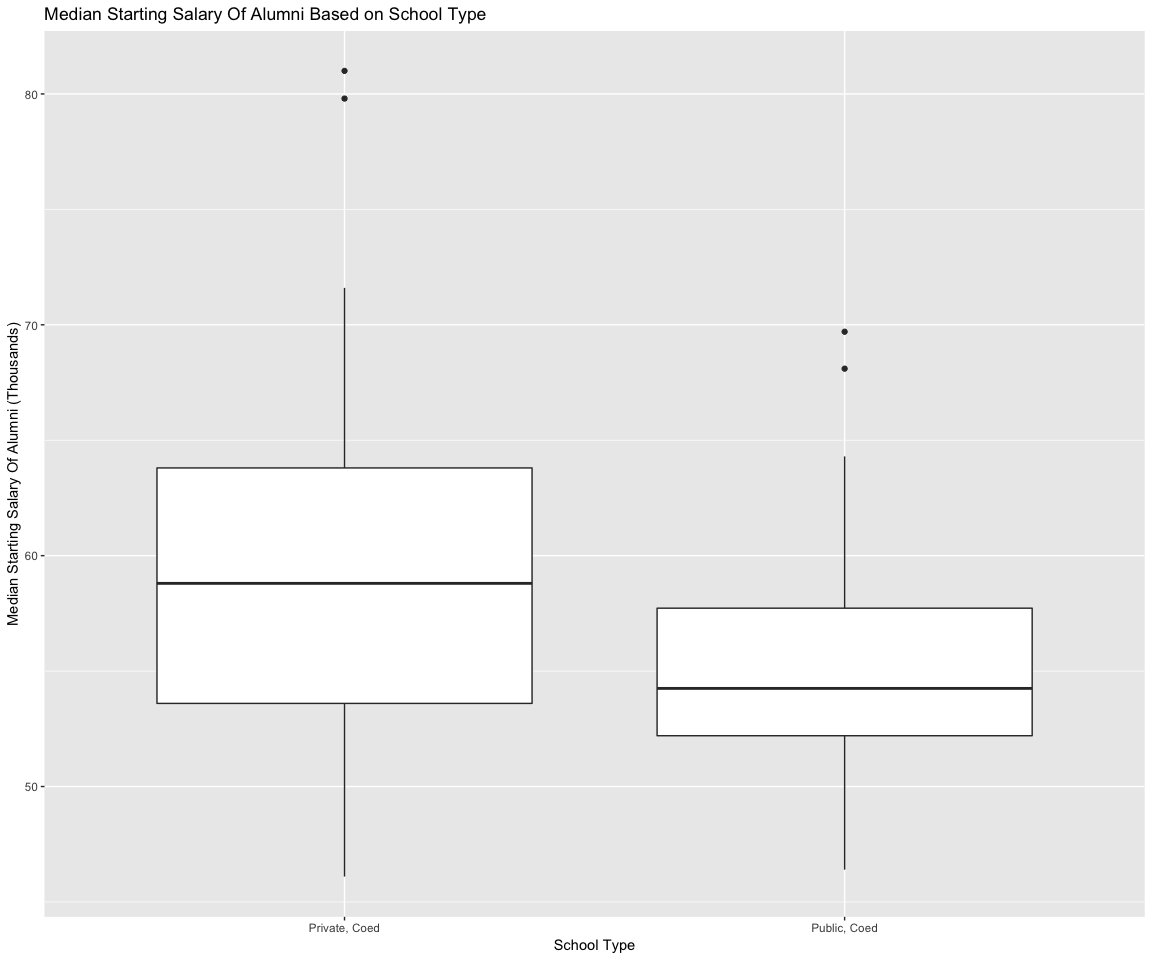
 The boxplots of room & board costs based on setting shows that the setting of the college has some influence the room and board costs for the students. The median room and board costs of the City settingvary from that of the others. The spread is also greater for the City setting while it is much smaller for the rural setting.

plot\_6 <- formatted\_college\_tab %>%  
   
 ggplot(aes(x=TotalCostThousands, y=MedianStartingSalaryOfAlumniThousands)) +  
 geom\_point()+   
 geom\_smooth(method=lm)+  
 labs(title="Total Cost vs. Median Starting Salary", x="Total Cost (Thousand)", y="Median Starting Salary Of Alumni (Thousands)")  
plot\_6



There appears to be a positive linear relationship between median starting salary and total cost of colleges. The general trends shows that the more students spend on tution, room, and board, the more likely that their starting salary is higher.

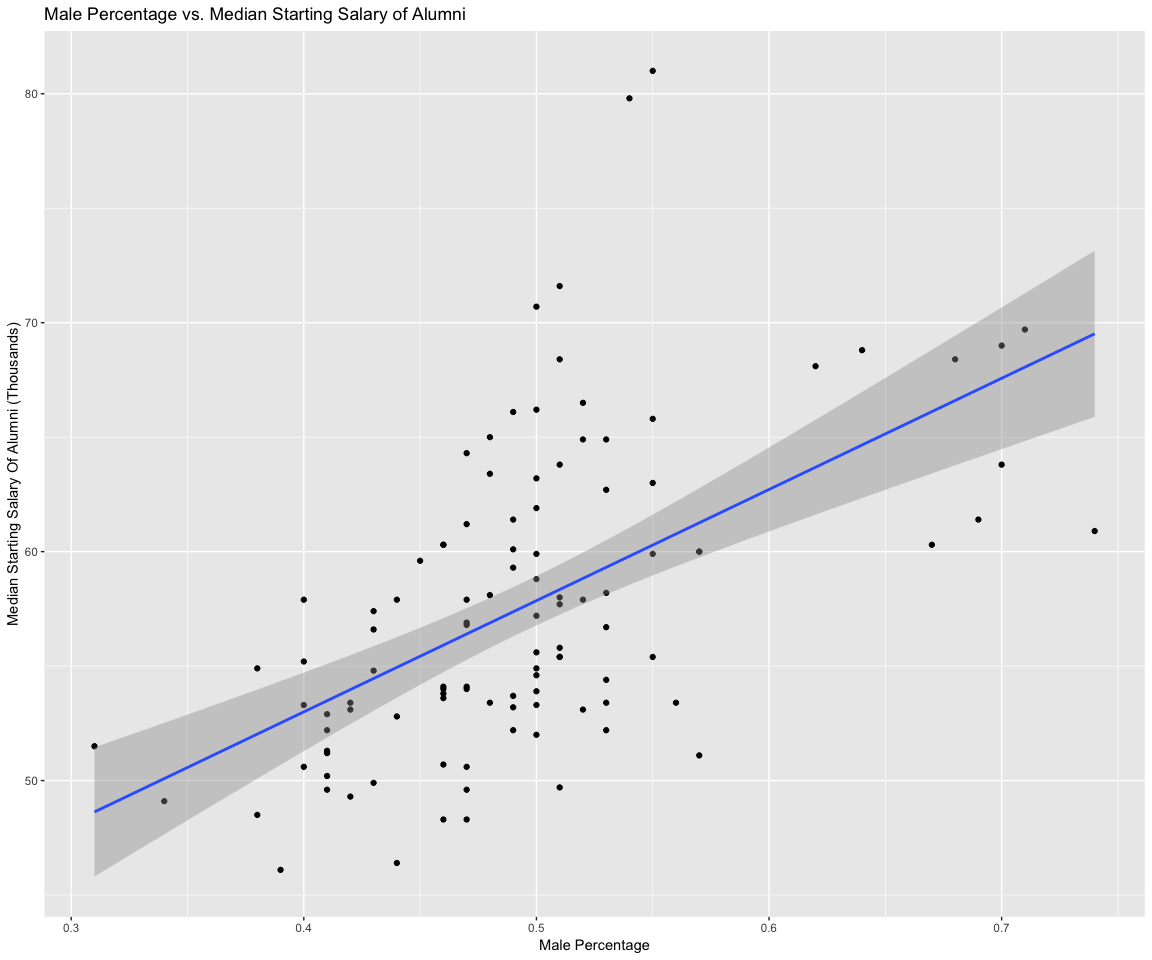
plot\_7 <- formatted\_college\_tab %>%  
 ggplot(aes(x=SchoolType, y=MedianStartingSalaryOfAlumniThousands   
)) +  
 geom\_boxplot()+  
 labs(title="Median Starting Salary Of Alumni Based on School Type ", x="School Type", y="Median Starting Salary Of Alumni (Thousands)")  
plot\_7

 Between school types, private colleges seem to have greater starting salaries than public schools, based on the medians of these boxplots.

formatted\_college\_tab %>% group\_by(Selectivity) %>%  
 summarise(n())

## # A tibble: 3 x 2  
## Selectivity `n()`  
## <fct> <int>  
## 1 Selective 2  
## 2 More selective 61  
## 3 Most selective 44

plot\_8 <- formatted\_college\_tab %>%  
 ggplot(aes(x=MalePercentage, y=MedianStartingSalaryOfAlumniThousands   
)) +  
 geom\_point()+  
 geom\_smooth(method=lm)+  
 labs(title="Male Percentage vs. Median Starting Salary of Alumni ", x="Male Percentage", y="Median Starting Salary Of Alumni (Thousands)")  
plot\_8

 Although the points are scattered with some variation, there is a general positive correlation between median starting salary of alumni and the male percentage of the student body of colleges.

# Part 3: Model Fitting and Selection

## Fitting model for tuition prices

#adjusting dataset to remove variables not able to be used in model fitting  
college\_info <- formatted\_college\_tab[,-c(1,2)]  
head(college\_info)

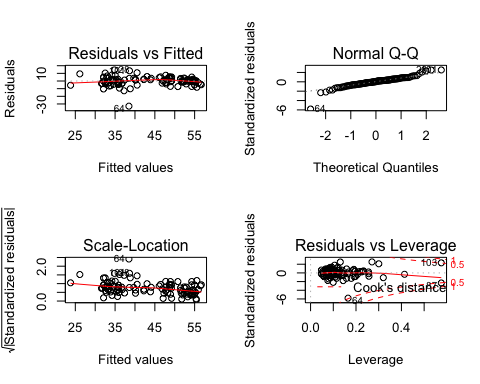
## TuitionFeesThousands RoomBoardThousands TotalEnrollment SchoolType  
## 1 47.140 15.610 8273 Private, Coed  
## 2 50.420 17.160 20604 Private, Coed  
## 3 59.430 14.016 25968 Private, Coed  
## 4 51.832 15.510 11466 Private, Coed  
## 5 57.006 16.350 13736 Private, Coed  
## 6 53.430 16.000 12974 Private, Coed  
## YearFounded Setting Endowment2017Millions  
## 1 1746 Suburban 23400  
## 2 1636 Urban 37100  
## 3 1754 Urban 10000  
## 4 1861 Urban 14800  
## 5 1890 Urban 6600  
## 6 1701 City 27200  
## MedianStartingSalaryOfAlumniThousands Selectivity  
## 1 68.4 Most selective  
## 2 66.5 Most selective  
## 3 64.9 Most selective  
## 4 79.8 Most selective  
## 5 57.7 Most selective  
## 6 63.2 Most selective  
## Fall2017AcceptanceRate MalePercentage FourYearGraduationRate  
## 1 0.06 0.51 0.89  
## 2 0.05 0.52 0.84  
## 3 0.06 0.52 0.88  
## 4 0.07 0.54 0.85  
## 5 0.09 0.51 0.88  
## 6 0.07 0.50 0.87  
## TotalCostThousands  
## 1 62.750  
## 2 67.580  
## 3 73.446  
## 4 67.342  
## 5 73.356  
## 6 69.430

college\_info$FourYearGraduationRate <- college\_info$FourYearGraduationRate\*100  
college\_info$MalePercentage <- college\_info$MalePercentage\*100  
college\_info$Fall2017AcceptanceRate <- college\_info$Fall2017AcceptanceRate\*100

#linear model fitting   
tuition\_lm\_1 <- lm(TuitionFeesThousands~.-RoomBoardThousands-TotalCostThousands, data = college\_info)  
summary(tuition\_lm\_1)

##   
## Call:  
## lm(formula = TuitionFeesThousands ~ . - RoomBoardThousands -   
## TotalCostThousands, data = college\_info)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.860 -2.743 0.146 3.237 14.500   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 8.311e+00 2.691e+01 0.309  
## TotalEnrollment -3.283e-05 5.921e-05 -0.555  
## SchoolTypePublic, Coed -1.150e+01 1.909e+00 -6.021  
## YearFounded 4.876e-03 1.305e-02 0.374  
## SettingSuburban 9.797e-01 2.825e+00 0.347  
## SettingUrban 1.483e+00 2.899e+00 0.512  
## SettingCity -2.112e-01 2.878e+00 -0.073  
## Endowment2017Millions -5.211e-05 1.483e-04 -0.351  
## MedianStartingSalaryOfAlumniThousands 1.396e-01 1.705e-01 0.819  
## SelectivityMore selective 6.810e+00 4.740e+00 1.437  
## SelectivityMost selective 1.136e+01 5.175e+00 2.194  
## Fall2017AcceptanceRate 5.185e-02 5.897e-02 0.879  
## MalePercentage 2.208e-02 1.319e-01 0.167  
## FourYearGraduationRate 1.754e-01 4.529e-02 3.873  
## Pr(>|t|)   
## (Intercept) 0.7581   
## TotalEnrollment 0.5806   
## SchoolTypePublic, Coed 3.42e-08 \*\*\*  
## YearFounded 0.7095   
## SettingSuburban 0.7295   
## SettingUrban 0.6100   
## SettingCity 0.9417   
## Endowment2017Millions 0.7262   
## MedianStartingSalaryOfAlumniThousands 0.4151   
## SelectivityMore selective 0.1541   
## SelectivityMost selective 0.0307 \*   
## Fall2017AcceptanceRate 0.3815   
## MalePercentage 0.8674   
## FourYearGraduationRate 0.0002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.182 on 93 degrees of freedom  
## Multiple R-squared: 0.7003, Adjusted R-squared: 0.6584   
## F-statistic: 16.72 on 13 and 93 DF, p-value: < 2.2e-16

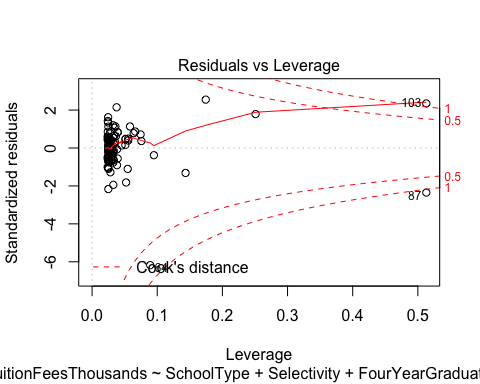
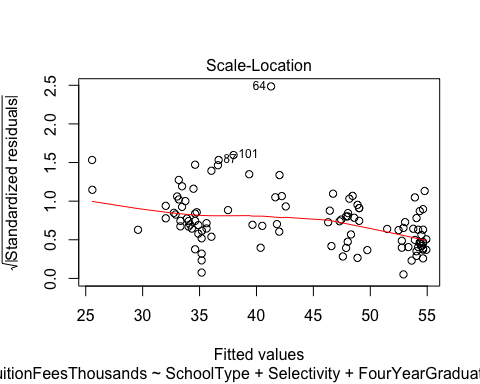
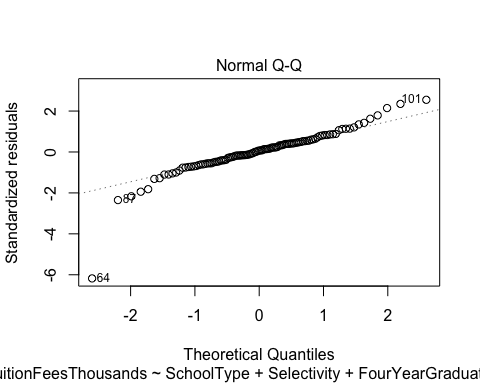
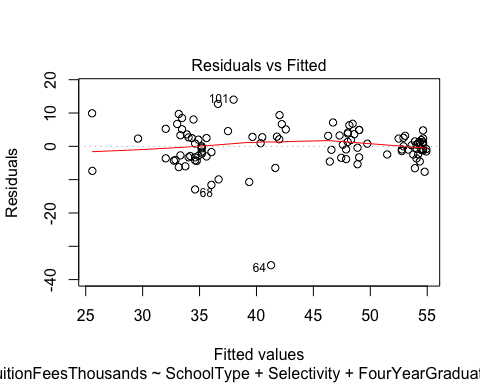
par(mfrow=c(2,2))  
plot(tuition\_lm\_1)



tuition\_lm\_2 <- step(tuition\_lm\_1, direction = "both", steps = 1000, trace = F)  
summary(tuition\_lm\_2)

##   
## Call:  
## lm(formula = TuitionFeesThousands ~ SchoolType + Selectivity +   
## FourYearGraduationRate, data = college\_info)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -35.663 -2.870 0.402 3.025 14.015   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.51347 4.69313 6.502 2.98e-09 \*\*\*  
## SchoolTypePublic, Coed -12.40552 1.29191 -9.602 6.18e-16 \*\*\*  
## SelectivityMore selective 7.47358 4.37232 1.709 0.090437 .   
## SelectivityMost selective 11.51048 4.51421 2.550 0.012265 \*   
## FourYearGraduationRate 0.14329 0.03637 3.940 0.000149 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.049 on 102 degrees of freedom  
## Multiple R-squared: 0.6854, Adjusted R-squared: 0.6731   
## F-statistic: 55.55 on 4 and 102 DF, p-value: < 2.2e-16

plot(tuition\_lm\_2)



anova(tuition\_lm\_2,tuition\_lm\_1, test="Chisq")

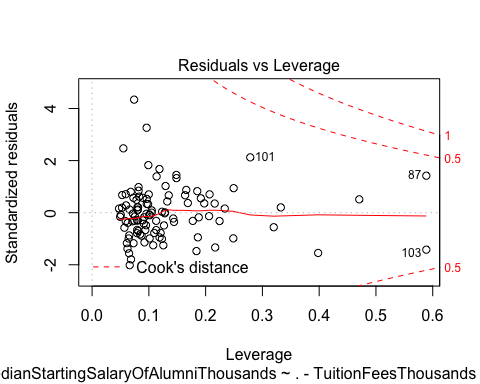
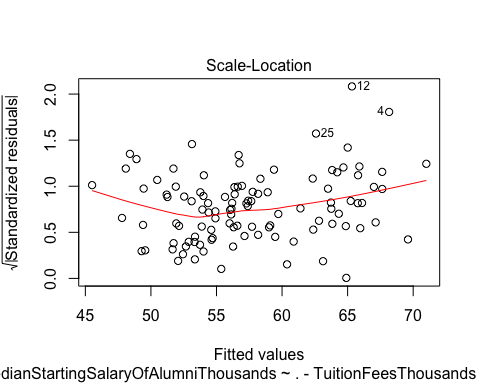
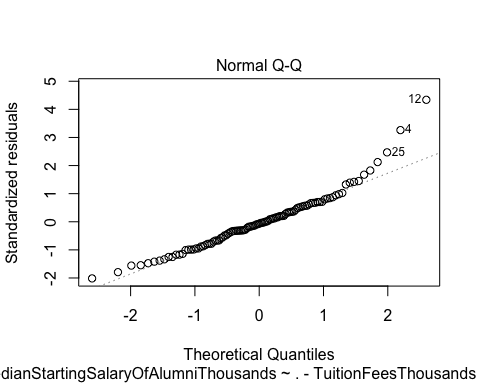
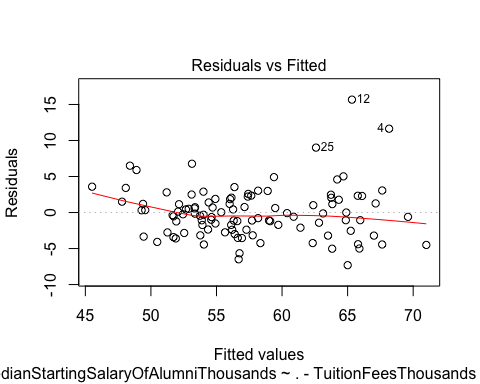
## Analysis of Variance Table  
##   
## Model 1: TuitionFeesThousands ~ SchoolType + Selectivity + FourYearGraduationRate  
## Model 2: TuitionFeesThousands ~ (RoomBoardThousands + TotalEnrollment +   
## SchoolType + YearFounded + Setting + Endowment2017Millions +   
## MedianStartingSalaryOfAlumniThousands + Selectivity + Fall2017AcceptanceRate +   
## MalePercentage + FourYearGraduationRate + TotalCostThousands) -   
## RoomBoardThousands - TotalCostThousands  
## Res.Df RSS Df Sum of Sq Pr(>Chi)  
## 1 102 3731.7   
## 2 93 3554.7 9 176.99 0.8653

## Fitting model for graduation rate

#linear model fitting   
gradrate\_lm\_1 <- lm(MedianStartingSalaryOfAlumniThousands~.-TuitionFeesThousands-RoomBoardThousands, data = na.omit(college\_info))  
summary(gradrate\_lm\_1)

##   
## Call:  
## lm(formula = MedianStartingSalaryOfAlumniThousands ~ . - TuitionFeesThousands -   
## RoomBoardThousands, data = na.omit(college\_info))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.3062 -2.3889 -0.2445 1.8739 15.6698   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.371e+01 1.616e+01 1.468 0.1455   
## TotalEnrollment 8.453e-06 3.614e-05 0.234 0.8156   
## SchoolTypePublic, Coed -1.430e+00 1.325e+00 -1.079 0.2833   
## YearFounded 5.017e-03 7.928e-03 0.633 0.5284   
## SettingSuburban -1.682e+00 1.707e+00 -0.986 0.3269   
## SettingUrban -1.043e+00 1.760e+00 -0.593 0.5547   
## SettingCity -1.384e+00 1.741e+00 -0.795 0.4286   
## Endowment2017Millions 2.166e-04 8.718e-05 2.485 0.0147 \*   
## SelectivityMore selective -2.567e+00 2.885e+00 -0.890 0.3759   
## SelectivityMost selective -8.552e-02 3.207e+00 -0.027 0.9788   
## Fall2017AcceptanceRate -8.235e-02 3.478e-02 -2.368 0.0200 \*   
## MalePercentage 5.486e-01 5.637e-02 9.732 7.55e-16 \*\*\*  
## FourYearGraduationRate 1.843e-02 2.907e-02 0.634 0.5276   
## TotalCostThousands 3.458e-02 5.496e-02 0.629 0.5307   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.752 on 93 degrees of freedom  
## Multiple R-squared: 0.7248, Adjusted R-squared: 0.6864   
## F-statistic: 18.85 on 13 and 93 DF, p-value: < 2.2e-16

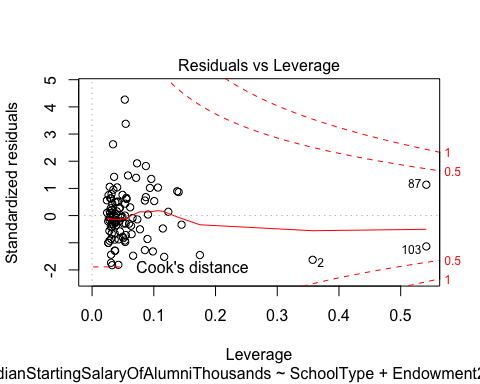
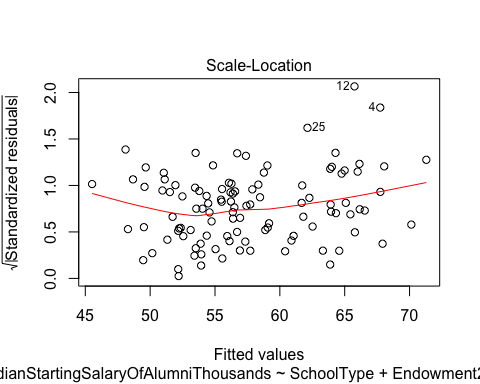
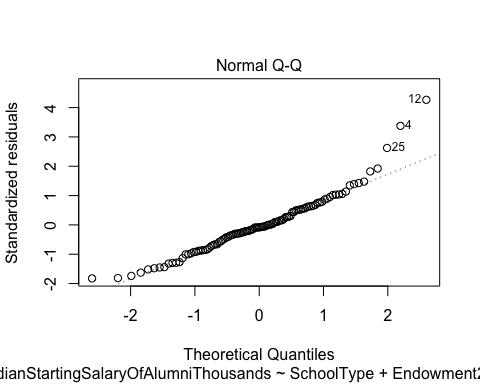
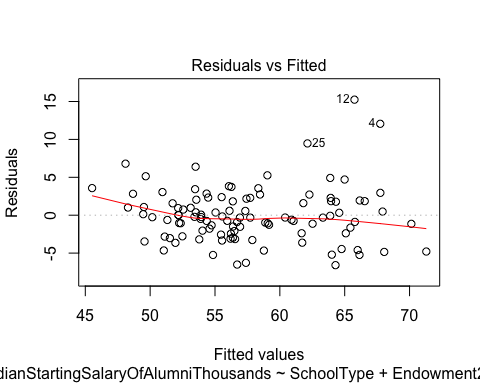
plot(gradrate\_lm\_1)



gradrate\_lm\_2 <- step(gradrate\_lm\_1, direction = "both", steps = 1000, trace = F)  
summary(gradrate\_lm\_2)

##   
## Call:  
## lm(formula = MedianStartingSalaryOfAlumniThousands ~ SchoolType +   
## Endowment2017Millions + Selectivity + Fall2017AcceptanceRate +   
## MalePercentage, data = na.omit(college\_info))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.5908 -2.3830 -0.3075 1.8991 15.2461   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.545e+01 3.638e+00 9.744 3.6e-16 \*\*\*  
## SchoolTypePublic, Coed -1.912e+00 8.010e-01 -2.386 0.01889 \*   
## Endowment2017Millions 1.998e-04 7.306e-05 2.735 0.00738 \*\*   
## SelectivityMore selective -2.042e+00 2.713e+00 -0.752 0.45357   
## SelectivityMost selective 6.553e-01 2.981e+00 0.220 0.82643   
## Fall2017AcceptanceRate -9.100e-02 3.173e-02 -2.868 0.00504 \*\*   
## MalePercentage 5.428e-01 4.828e-02 11.243 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.672 on 100 degrees of freedom  
## Multiple R-squared: 0.7166, Adjusted R-squared: 0.6996   
## F-statistic: 42.15 on 6 and 100 DF, p-value: < 2.2e-16

plot(gradrate\_lm\_2)



anova(gradrate\_lm\_2,gradrate\_lm\_1, test="Chisq")

## Analysis of Variance Table  
##   
## Model 1: MedianStartingSalaryOfAlumniThousands ~ SchoolType + Endowment2017Millions +   
## Selectivity + Fall2017AcceptanceRate + MalePercentage  
## Model 2: MedianStartingSalaryOfAlumniThousands ~ (TuitionFeesThousands +   
## RoomBoardThousands + TotalEnrollment + SchoolType + YearFounded +   
## Setting + Endowment2017Millions + Selectivity + Fall2017AcceptanceRate +   
## MalePercentage + FourYearGraduationRate + TotalCostThousands) -   
## TuitionFeesThousands - RoomBoardThousands  
## Res.Df RSS Df Sum of Sq Pr(>Chi)  
## 1 100 1348.5   
## 2 93 1309.3 7 39.15 0.9045

# Conclusion

Being aware of all these factors in succeeding in college is very important when deciding where to go.

References: -College Ranking Data: <https://www.usnews.com/best-colleges/rankings/national-universities>