# Analysis of MBA SALARIES

# NAME: Sonal Somani

# EMAIL: [sonalm300@gmail.com](mailto:sonalm300@gmail.com)

# COLLEGE / COMPANY: Antuit India Pvt. Ltd.

#Load packages  
  
library(statsr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(gplots)

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

#set working directory  
  
setwd("C:/Users/Sonal Somani/Desktop/IIMInternship/R\_code")  
  
#load dataset into R  
  
salary <- read.csv(paste("MBA\_Salary\_Data.csv",sep=""))  
  
#View dataset  
View(salary)

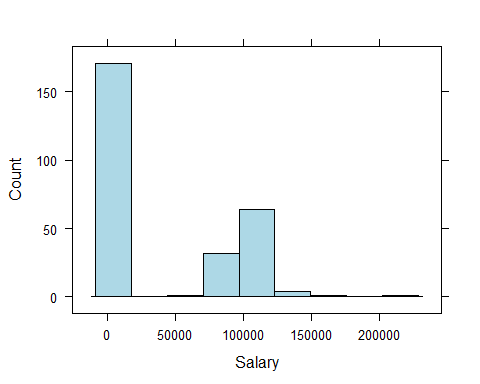
#Check how your data looks like with their datatypes  
str(salary)

## 'data.frame': 274 obs. of 13 variables:  
## $ age : int 23 24 24 24 24 24 25 25 25 25 ...  
## $ sex : int 2 1 1 1 2 1 1 2 1 1 ...  
## $ gmat\_tot: int 620 610 670 570 710 640 610 650 630 680 ...  
## $ gmat\_qpc: int 77 90 99 56 93 82 89 88 79 99 ...  
## $ gmat\_vpc: int 87 71 78 81 98 89 74 89 91 81 ...  
## $ gmat\_tpc: int 87 87 95 75 98 91 87 92 89 96 ...  
## $ s\_avg : num 3.4 3.5 3.3 3.3 3.6 3.9 3.4 3.3 3.3 3.45 ...  
## $ f\_avg : num 3 4 3.25 2.67 3.75 3.75 3.5 3.75 3.25 3.67 ...  
## $ quarter : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ work\_yrs: int 2 2 2 1 2 2 2 2 2 2 ...  
## $ frstlang: int 1 1 1 1 1 1 1 1 2 1 ...  
## $ salary : int 0 0 0 0 999 0 0 0 999 998 ...  
## $ satis : int 7 6 6 7 5 6 5 6 4 998 ...

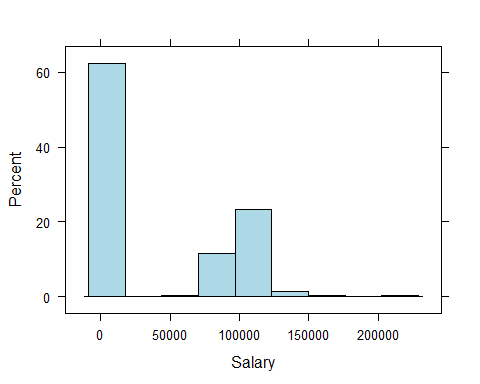
#To get summary statistics of your data, like mean, median, mode for numeric data etc.  
  
summary(salary)

## age sex gmat\_tot gmat\_qpc   
## Min. :22.00 Min. :1.000 Min. :450.0 Min. :28.00   
## 1st Qu.:25.00 1st Qu.:1.000 1st Qu.:580.0 1st Qu.:72.00   
## Median :27.00 Median :1.000 Median :620.0 Median :83.00   
## Mean :27.36 Mean :1.248 Mean :619.5 Mean :80.64   
## 3rd Qu.:29.00 3rd Qu.:1.000 3rd Qu.:660.0 3rd Qu.:93.00   
## Max. :48.00 Max. :2.000 Max. :790.0 Max. :99.00   
## gmat\_vpc gmat\_tpc s\_avg f\_avg   
## Min. :16.00 Min. : 0.0 Min. :2.000 Min. :0.000   
## 1st Qu.:71.00 1st Qu.:78.0 1st Qu.:2.708 1st Qu.:2.750   
## Median :81.00 Median :87.0 Median :3.000 Median :3.000   
## Mean :78.32 Mean :84.2 Mean :3.025 Mean :3.062   
## 3rd Qu.:91.00 3rd Qu.:94.0 3rd Qu.:3.300 3rd Qu.:3.250   
## Max. :99.00 Max. :99.0 Max. :4.000 Max. :4.000   
## quarter work\_yrs frstlang salary   
## Min. :1.000 Min. : 0.000 Min. :1.000 Min. : 0   
## 1st Qu.:1.250 1st Qu.: 2.000 1st Qu.:1.000 1st Qu.: 0   
## Median :2.000 Median : 3.000 Median :1.000 Median : 999   
## Mean :2.478 Mean : 3.872 Mean :1.117 Mean : 39026   
## 3rd Qu.:3.000 3rd Qu.: 4.000 3rd Qu.:1.000 3rd Qu.: 97000   
## Max. :4.000 Max. :22.000 Max. :2.000 Max. :220000   
## satis   
## Min. : 1.0   
## 1st Qu.: 5.0   
## Median : 6.0   
## Mean :172.2   
## 3rd Qu.: 7.0   
## Max. :998.0

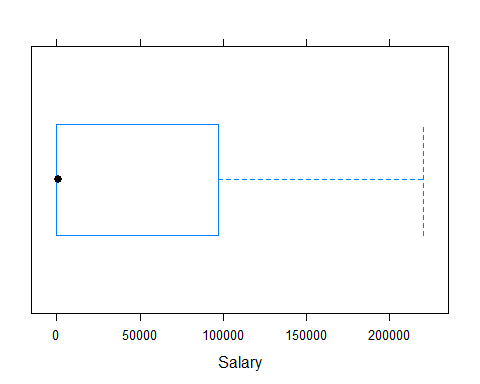
#Single varibale visualizations - Explorartory Data Analysis   
  
#Histogram of salary  
library(lattice)  
histogram(salary$salary, type="count", xlab="Salary",ylab = "Count",col=c("lightblue"))



#Histogram of salary with percentages  
  
histogram(salary$salary, xlab="Salary",ylab = "Percent",col=c("lightblue"))

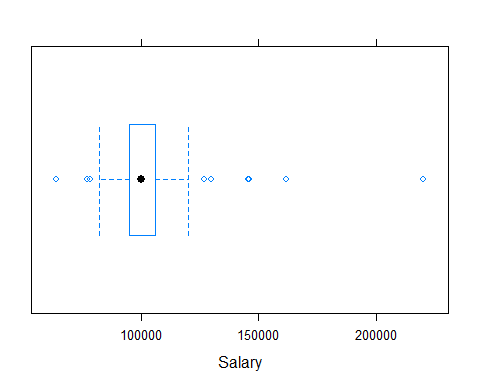


#Box plot of salary  
bwplot(salary$salary,   
 horizontal=TRUE, xlab="Salary")



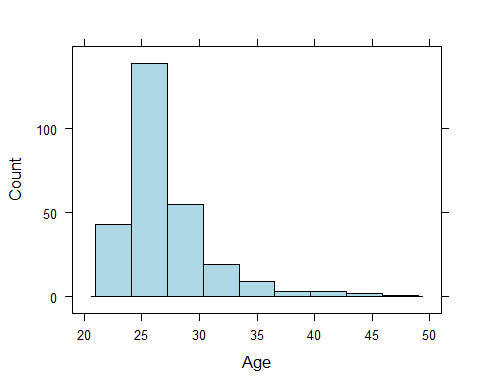
#Here, our results are right skewed as there are more than 60% people who are getting salary as 0 as they are not placed.   
  
#Filter people not placed i.e. people with salary less than 999.  
placed <- subset(salary, salary > 999)

#Box plot of salary with data frame placed  
bwplot(placed$salary,   
 horizontal=TRUE, xlab="Salary")



#Here we see, that the median of salary for the placed people is 1 lakhs.

#Histogram of age  
histogram(salary$age, type="count", xlab="Age",ylab = "Count",col=c("lightblue"))

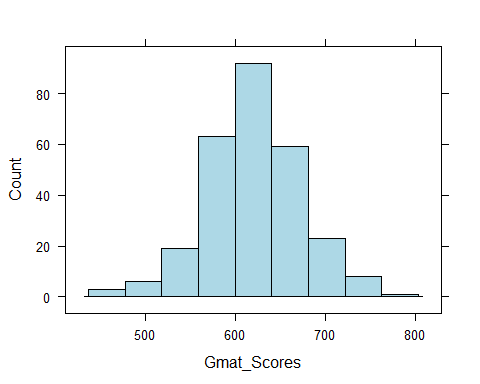


#It can be seen the distribution is unimodal and right skewed which suggests that there are more people in their 20s who are enrolled than who are 30 +.

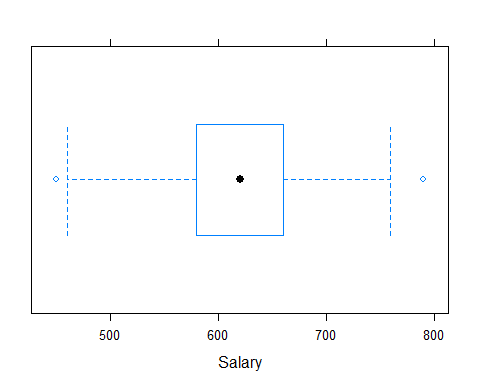
summary(salary$gmat\_tot)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 450.0 580.0 620.0 619.5 660.0 790.0

#Histogram of gmat scores  
  
histogram(salary$gmat\_tot, type="count", xlab="Gmat\_Scores",ylab = "Count",col=c("lightblue"))



#Box plot of gmat total score of all students  
  
bwplot(salary$gmat\_tot,   
 horizontal=TRUE, xlab="Salary")

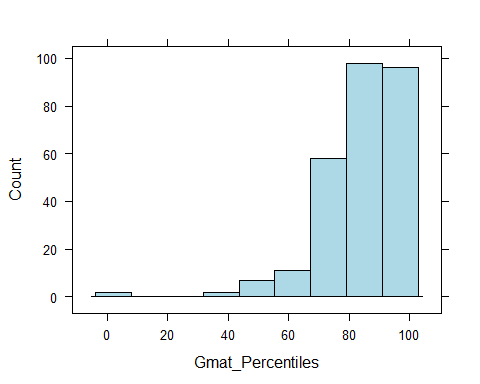


#It clearly follows a normal distrbution with 620 as the median score.

summary(salary$gmat\_tpc)

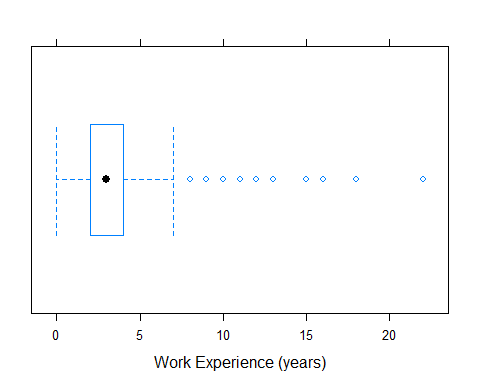
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 78.0 87.0 84.2 94.0 99.0

#Histogram of gmat total percentiles  
  
histogram(salary$gmat\_tpc, type="count", xlab="Gmat\_Percentiles",ylab = "Count",col=c("lightblue"))



#This distribution is left skewed suggesting there are far more people with 80 + percentiles than lower.

#Box plot of work experience (years)  
  
bwplot(salary$work\_yrs,   
 horizontal=TRUE, xlab="Work Experience (years)")

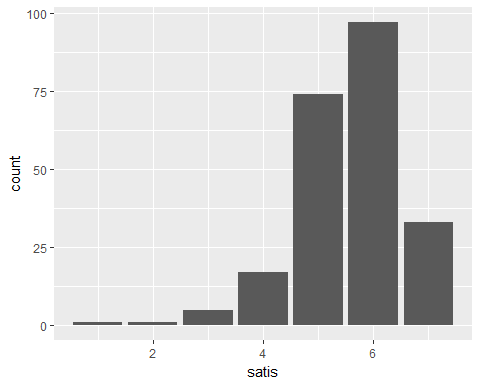


#The distribution has a lot of outliers suggesting generally people with 2-4 years join MBA programs while some have a lot more experience (over 20 years too).

table(salary$satis)

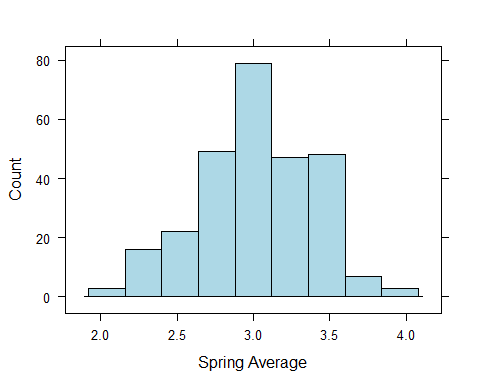
##   
## 1 2 3 4 5 6 7 998   
## 1 1 5 17 74 97 33 46

#Since many didn't fill out the survey for satisfaction levels, we will filter them out and see the distribution again  
  
survey <- subset(salary, satis < 8)  
  
#Bar chart of satisfaction level with the MBA Program  
  
ggplot(survey, aes(x = satis, fill = satis)) + geom\_bar()



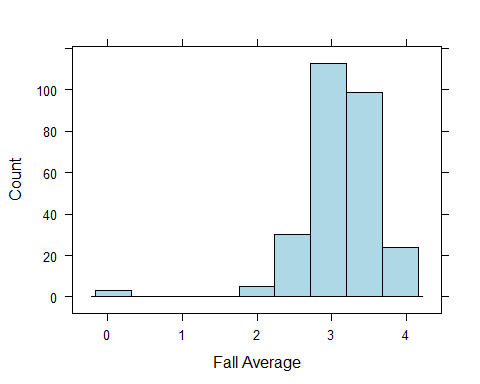
#The chart shows that of those who fill out responses for satisfaction levels of MBA Program gave the course a 6 or had high satisfaction levels.

# Histogram for Spring Average (s\_avg)  
  
histogram(salary$s\_avg, type="count", xlab="Spring Average",ylab = "Count",col=c("lightblue"))



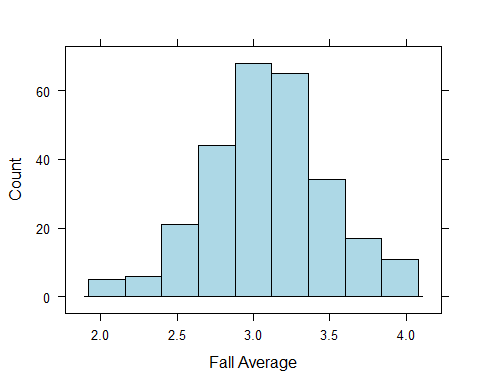
#The histogram shows unimodal and normal distribution for the spring averages with median being 3.

# Histogram for Fall Average (f\_avg)  
  
histogram(salary$f\_avg, type="count", xlab="Fall Average",ylab = "Count",col=c("lightblue"))



#Fall average for a couple of students is 0, which may be beacuse those people dropped out of the course or some other reason. If we exclude those and re run our analysis, let's see whta do we get -  
  
passed <- subset(salary, f\_avg > 0 )

histogram(passed$f\_avg, type="count", xlab="Fall Average",ylab = "Count",col=c("lightblue"))

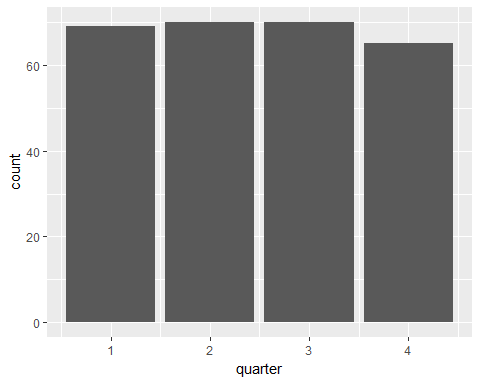


#The fall average distribution too follows a normal curve with the median being around 3.25.

table(salary$quarter)

##   
## 1 2 3 4   
## 69 70 70 65

#Bar chart of the quartile ranking of students  
ggplot(salary, aes(x = quarter)) + geom\_bar()

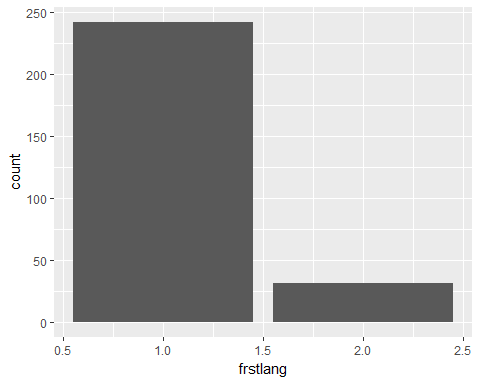


#The distribution looks pretty uniform with students falling in each of the quartiles pretty uniformly.

table(salary$frstlang)

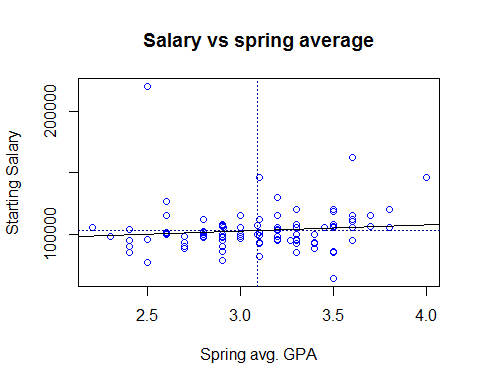
##   
## 1 2   
## 242 32

#Bar chart of the first language of people enrolled in the course  
ggplot(salary, aes(x = frstlang)) + geom\_bar()



#The chart shows that most students enrolled have English as their first language.

#Time for bivariate analysis - scatterplots  
  
library(lattice)  
#Let's do this only for students who were placed.  
  
#Scatter plot for Salary and spring average (Since both are numeric in nature)  
  
plot(placed$s\_avg,placed$salary,  
 col="blue",  
 main="Salary vs spring average",  
 xlab="Spring avg. GPA", ylab="Starting Salary")  
  
# Add the sample means to the Scatterplot  
  
abline(h=mean(placed$salary), col="dark blue", lty="dotted")  
abline(v=mean(placed$s\_avg), col="dark blue", lty="dotted")  
  
  
# Add a regression line  
  
abline(lm(placed$salary ~ placed$s\_avg))



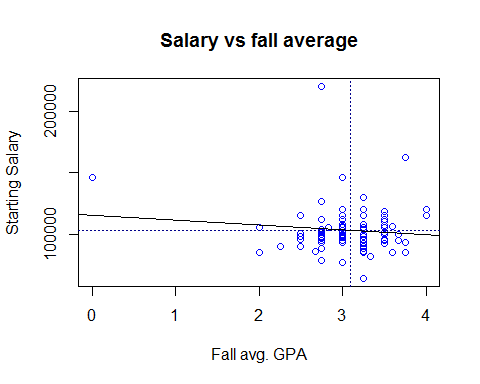
# With this information, it is hard to say more spring GPA avg. is correlated with more starting salary.

#Let's do a correlation test to confirm this observation of ours.  
  
cor.test(placed$salary,placed$s\_avg)

##   
## Pearson's product-moment correlation  
##   
## data: placed$salary and placed$s\_avg  
## t = 1.0277, df = 101, p-value = 0.3065  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.09363639 0.28955576  
## sample estimates:  
## cor   
## 0.1017317

#The results confirm our observation showing a weak correlation between the two variables. Also, the p-value is 0.3065 which is above 0.05 and thus we fail to reject the null hypothesis that salary and spring average affect each other.

#Scatter plot for Salary and fall average (Since both are numeric in nature)  
  
plot(placed$f\_avg,placed$salary,  
 col="blue",  
 main="Salary vs fall average",  
 xlab="Fall avg. GPA", ylab="Starting Salary")  
  
# Add the sample means to the Scatterplot  
  
abline(h=mean(placed$salary), col="dark blue", lty="dotted")  
abline(v=mean(placed$f\_avg), col="dark blue", lty="dotted")  
  
  
# Add a regression line  
  
abline(lm(placed$salary ~ placed$f\_avg))



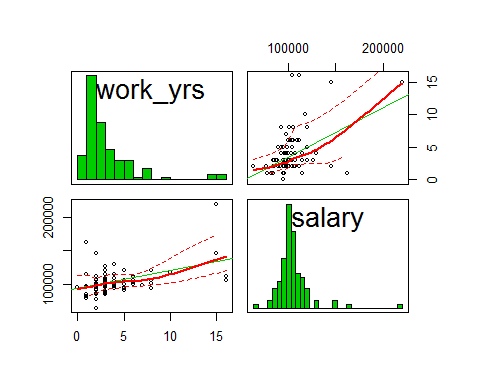
# With this information, it is hard to say more fall GPA avg. is correlated with more starting salary.

#Scatter plot matrix between work experience and salary  
  
library(car)

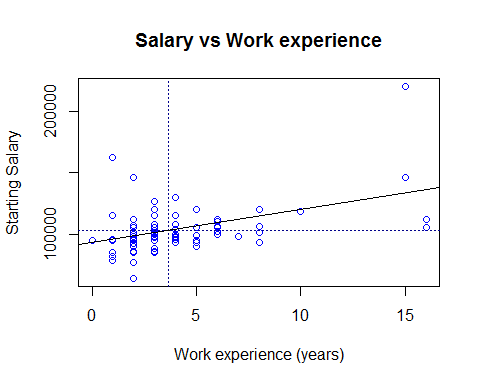
##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

scatterplotMatrix(formula = ~work\_yrs + salary, cex = 0.6, data = placed, diagonal = "histogram")



#Scatter plot for Salary and work experience (Since both are numeric in nature)  
  
plot(placed$work\_yrs,placed$salary,  
 col="blue",  
 main="Salary vs Work experience",  
 xlab="Work experience (years)", ylab="Starting Salary")  
  
# Add the sample means to the Scatterplot  
  
abline(h=mean(placed$salary), col="dark blue", lty="dotted")  
abline(v=mean(placed$work\_yrs), col="dark blue", lty="dotted")  
  
  
# Add a regression line  
  
abline(lm(placed$salary ~ placed$work\_yrs))

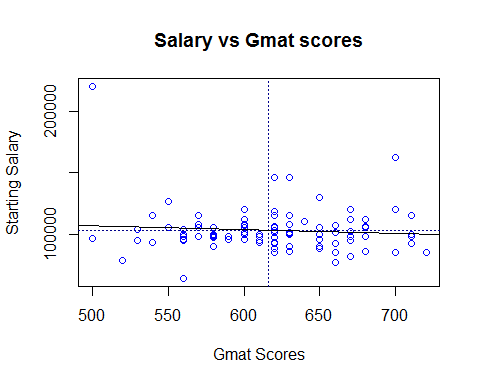


# The regression line shows an upward trend in the starting salaries and work experience (yrs), we can confirm this by doing a correlation test.  
  
cor.test(placed$salary, placed$work\_yrs)

##   
## Pearson's product-moment correlation  
##   
## data: placed$salary and placed$work\_yrs  
## t = 5.1303, df = 101, p-value = 1.403e-06  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2863362 0.5957697  
## sample estimates:  
## cor   
## 0.4546663

# We can see that the p-value = 1.403e-06 which falls much below 0.05 and hence we can say that salary and work experience are correlated and the correlation value being 0.45 shows that they both are positively correlated and their relationship has moderate strength.

#Scatter plot for Salary and Gmat Score (Since both are numeric in nature)  
  
plot(placed$gmat\_tot,placed$salary,  
 col="blue",  
 main="Salary vs Gmat scores",  
 xlab="Gmat Scores", ylab="Starting Salary")  
  
# Add the sample means to the Scatterplot  
  
abline(h=mean(placed$salary), col="dark blue", lty="dotted")  
abline(v=mean(placed$gmat\_tot), col="dark blue", lty="dotted")  
  
  
# Add a regression line  
  
abline(lm(placed$salary ~ placed$gmat\_tot))

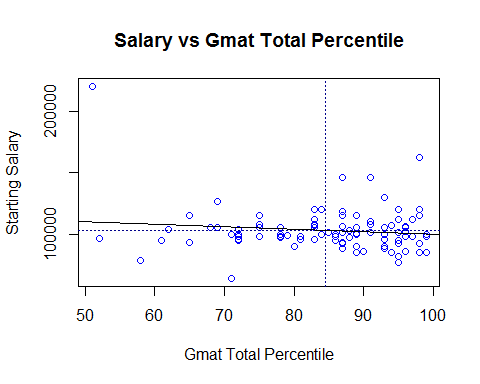


# With the above information, it is hard to say if more GMAT scores are correlated with starting salary, we can confirm this by doing a correlation test.  
  
cor.test(placed$salary,placed$gmat\_tot)

##   
## Pearson's product-moment correlation  
##   
## data: placed$salary and placed$gmat\_tot  
## t = -0.91501, df = 101, p-value = 0.3624  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.2792952 0.1046903  
## sample estimates:  
## cor   
## -0.09067141

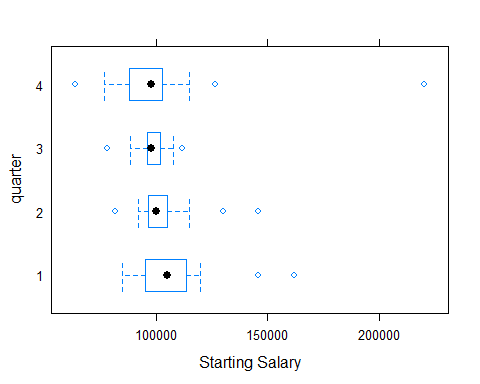
# We can see that the p-value = 0.3624 which is above 0.05 and thus we fail to reject the null hypothesis that salary and gmat scores affect each other.

#Scatter plot for Salary and Gmat Total Percentile (Since both are numeric in nature)  
  
plot(placed$gmat\_tpc,placed$salary,  
 col="blue",  
 main="Salary vs Gmat Total Percentile",  
 xlab="Gmat Total Percentile", ylab="Starting Salary")  
  
# Add the sample means to the Scatterplot  
  
abline(h=mean(placed$salary), col="dark blue", lty="dotted")  
abline(v=mean(placed$gmat\_tpc), col="dark blue", lty="dotted")  
  
  
# Add a regression line  
  
abline(lm(placed$salary ~ placed$gmat\_tpc))



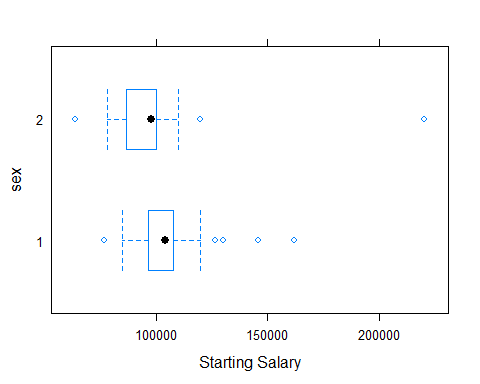
# With the given plot, it is hard to say if more GMAT scores are correlated with starting salary.

#Box plots for salaries by quartiles  
  
library(lattice)  
  
bwplot(quarter ~ salary, data=placed, horizontal=TRUE,   
 xlab = "Starting Salary")



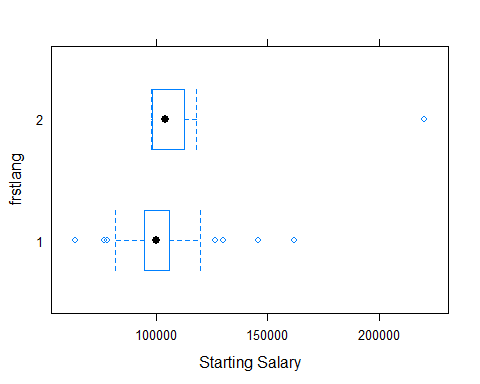
#The plot shows that quartiles does not affect much the distribution of salaries.

# Box plot for salary by sex  
  
bwplot(sex ~ salary, data=placed, horizontal=TRUE,   
 xlab = "Starting Salary")



# The plot shows that there seems to be some gender disparity in place when comparing salaries by gender. Median salary for females seems to be comparitively lower than that of males.

#Box plot for salary by first language  
  
bwplot(frstlang ~ salary, data=placed, horizontal=TRUE,   
 xlab = "Starting Salary")

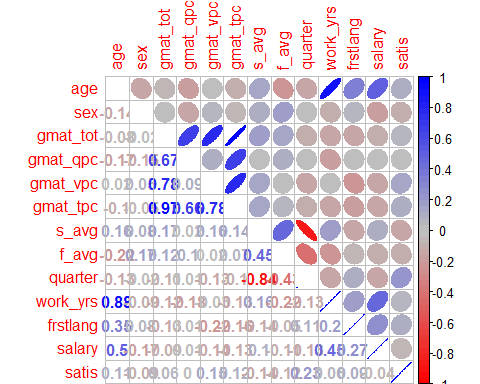


#While english speakers salary follows a normal distribution with some outliers on both end, none of the students whose first language is not english and are placed got a salary below 90000.  
  
#Let's do a correlation test -  
cor.test(placed$salary, placed$frstlang)

##   
## Pearson's product-moment correlation  
##   
## data: placed$salary and placed$frstlang  
## t = 2.7846, df = 101, p-value = 0.0064  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.07749965 0.43791500  
## sample estimates:  
## cor   
## 0.2670195

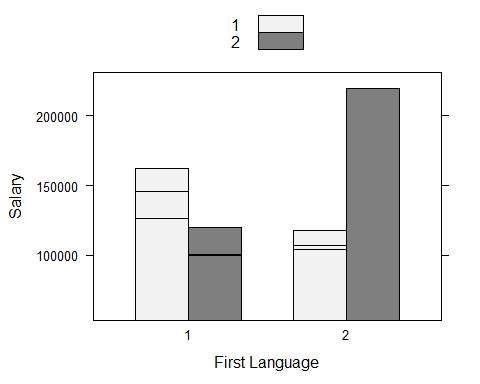
#The p-value suggests that first language and salaries are correlated to each other as it lies below 0.05 but the correlation coefficent is 0.27 which although shows a positive correlation but the strength of the relation is weak in nature.

#Corrgram  
  
library(corrplot)  
  
par(mfrow=c(1, 1))  
corrplot.mixed(corr=cor(placed[ , c(1:13)], use="complete.obs"),  
 upper="ellipse", tl.pos="lt",  
 col = colorpanel(50, "red", "gray", "blue"))



#This corrgram clearly shows that work experience and age has a strong positive correlation with salaries while first language and spring average also are positively correlated with moderate strength to salary.

#Contingency table for salary,sex and first language  
  
#Since this would only work on factors, I would be converting variables like sex and firstlang into factors.  
  
placed <- placed %>% mutate(sex1=factor(sex),firstlang1=factor(frstlang),quarter1=factor(quarter))  
  
barchart(salary ~ firstlang1 , data=placed,   
 groups=sex1, auto.key=TRUE,  
 par.settings = simpleTheme(col=c("gray95", "gray50")),xlab="First Language",ylab="Salary" )



#This plot shows that females whose first language is English tend to get lower salaries than their male counterparts while its the opposite for the females whose first language is not English.

#Contingency tables between sex,first language and quartiles  
  
prop.table(table(placed$sex1))

##   
## 1 2   
## 0.6990291 0.3009709

prop.table(table(placed$sex1,placed$firstlang1))

##   
## 1 2  
## 1 0.66019417 0.03883495  
## 2 0.27184466 0.02912621

mytable <- table(placed$sex1,placed$quarter1)  
prop.table(table(placed$sex1,placed$quarter1))

##   
## 1 2 3 4  
## 1 0.22330097 0.18446602 0.16504854 0.12621359  
## 2 0.11650485 0.05825243 0.06796117 0.05825243

#Chi sq test of independence between quartiles and sex  
  
chisq.test(mytable)

##   
## Pearson's Chi-squared test  
##   
## data: mytable  
## X-squared = 0.76332, df = 3, p-value = 0.8582

#Since the p-value is not below 0.05, we fail to reject our null hypothesis that quartiles and sex are independent of each other.  
  
mytable1 <- table(placed$quarter1,placed$firstlang1)  
prop.table(table(placed$quarter1,placed$firstlang1))

##   
## 1 2  
## 1 0.330097087 0.009708738  
## 2 0.223300971 0.019417476  
## 3 0.213592233 0.019417476  
## 4 0.165048544 0.019417476

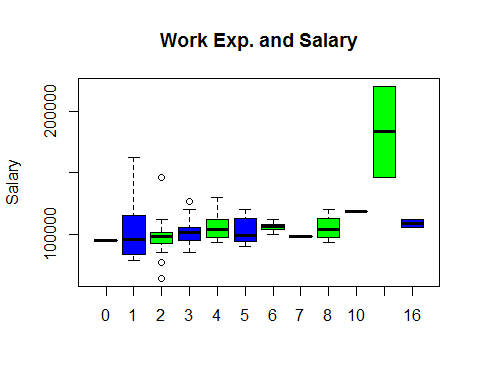
#Chi sq test of independence between quartiles and first language  
  
chisq.test(mytable1)

## Warning in chisq.test(mytable1): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: mytable1  
## X-squared = 1.4214, df = 3, p-value = 0.7005

#Since the p-value is not below 0.05, our null hypothesis that quartiles and first language are independent of each other is not rejected.

#t-tests  
  
#Articulating hypothesis as -  
#H1 = Work experience does have an effect on salary  
#Running t-test to test our hypothesis (H0,H1) -  
#Here, the null hypothesis (H0) is that work experience does not have an effect on salary.  
#Let's see -  
  
boxplot(placed$salary~placed$work\_yrs,main = "Work Exp. and Salary",col = (c("green","blue")), ylab = "Salary")

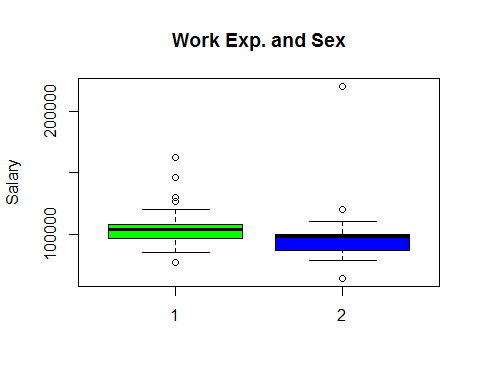


t.test(placed$salary,placed$work\_yrs)

##   
## Welch Two Sample t-test  
##   
## data: placed$salary and placed$work\_yrs  
## t = 58.516, df = 102, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 99534.79 106519.33  
## sample estimates:  
## mean of x mean of y   
## 1.030307e+05 3.679612e+00

#Since the results show that the p-value is < 2.2e-16, we can reject our null hypothesis in favour of our alternate hypothesis (H1) i.e. work experience does have an effect on starting salary.

#Articulating hypothesis as -  
#H1 = Males have a higher mean starting salary than females  
#Running t-test to test our hypothesis (H0,H1) -  
#Here, the null hypothesis (H0) is that Males and females have equal mean starting salaries.  
  
#Let's see -  
  
boxplot(placed$salary~placed$sex1,main = "Work Exp. and Sex",col = (c("green","blue")), ylab = "Salary")



t.test(placed$salary~placed$sex1,alternative="greater")

##   
## Welch Two Sample t-test  
##   
## data: placed$salary by placed$sex1  
## t = 1.3628, df = 38.115, p-value = 0.09047  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -1527.96 Inf  
## sample estimates:  
## mean in group 1 mean in group 2   
## 104970.97 98524.39

#Since the p-value is not below 0.05, our null hypothesis that Males and females have equal mean starting salaries is not rejected.

#Linear Regression Models  
  
#Model\_1  
model\_1 <- lm(salary ~ work\_yrs + s\_avg + sex, data = placed)  
summary(model\_1)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + s\_avg + sex, data = placed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31968 -8252 -1577 4715 92322   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 94047.3 13467.7 6.983 3.35e-10 \*\*\*  
## work\_yrs 2588.1 535.8 4.831 4.97e-06 \*\*\*  
## s\_avg 1935.0 4257.2 0.455 0.650   
## sex -5014.2 3463.7 -1.448 0.151   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15980 on 99 degrees of freedom  
## Multiple R-squared: 0.2239, Adjusted R-squared: 0.2004   
## F-statistic: 9.522 on 3 and 99 DF, p-value: 1.388e-05

#Multiple R-squared: 0.2239, Adjusted R-squared: 0.2004  
#p-value: 1.388e-05  
  
model\_2 <- lm(salary ~ work\_yrs + s\_avg + sex + gmat\_tot, data = placed)  
summary(model\_2)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + s\_avg + sex + gmat\_tot, data = placed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -33148 -8069 -1216 4586 91181   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 103625.22 22277.81 4.651 1.03e-05 \*\*\*  
## work\_yrs 2540.97 544.70 4.665 9.78e-06 \*\*\*  
## s\_avg 2407.29 4360.83 0.552 0.582   
## sex -5111.47 3480.77 -1.468 0.145   
## gmat\_tot -17.43 32.23 -0.541 0.590   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16040 on 98 degrees of freedom  
## Multiple R-squared: 0.2262, Adjusted R-squared: 0.1947   
## F-statistic: 7.163 on 4 and 98 DF, p-value: 4.208e-05

#Multiple R-squared: 0.2262, Adjusted R-squared: 0.1947  
#p-value: 4.208e-05  
  
model\_3 <- lm(salary ~ work\_yrs + s\_avg + sex + gmat\_tpc, data = placed)  
summary(model\_3)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + s\_avg + sex + gmat\_tpc, data = placed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34221 -8178 -1706 5162 89040   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 104867.5 17297.6 6.063 2.51e-08 \*\*\*  
## work\_yrs 2497.9 543.4 4.597 1.28e-05 \*\*\*  
## s\_avg 2674.4 4321.5 0.619 0.537   
## sex -5282.2 3474.2 -1.520 0.132   
## gmat\_tpc -147.0 147.5 -0.997 0.321   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15980 on 98 degrees of freedom  
## Multiple R-squared: 0.2317, Adjusted R-squared: 0.2004   
## F-statistic: 7.389 on 4 and 98 DF, p-value: 3.036e-05

#Multiple R-squared: 0.2317, Adjusted R-squared: 0.2004   
#p-value: 3.036e-05  
  
model\_4 <- lm(salary ~ work\_yrs + f\_avg + sex , data = placed)  
summary(model\_4)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + f\_avg + sex, data = placed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32046 -8090 -1827 4890 90627   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 98423.5 11423.9 8.616 1.13e-13 \*\*\*  
## work\_yrs 2643.7 539.7 4.898 3.77e-06 \*\*\*  
## f\_avg 416.5 3363.8 0.124 0.902   
## sex -4925.4 3489.9 -1.411 0.161   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15990 on 99 degrees of freedom  
## Multiple R-squared: 0.2224, Adjusted R-squared: 0.1989   
## F-statistic: 9.44 on 3 and 99 DF, p-value: 1.523e-05

#Multiple R-squared: 0.2224, Adjusted R-squared: 0.1989   
#p-value: 1.523e-05  
  
  
model\_5 <- lm(salary ~ work\_yrs + f\_avg + sex + quarter, data = placed)  
summary(model\_5)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + f\_avg + sex + quarter, data = placed)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31003 -7786 -1585 4085 93897   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 106974 14936 7.162 1.48e-10 \*\*\*  
## work\_yrs 2520 558 4.515 1.76e-05 \*\*\*  
## f\_avg -1201 3827 -0.314 0.754   
## sex -4792 3497 -1.370 0.174   
## quarter -1444 1623 -0.890 0.376   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16010 on 98 degrees of freedom  
## Multiple R-squared: 0.2287, Adjusted R-squared: 0.1972   
## F-statistic: 7.263 on 4 and 98 DF, p-value: 3.644e-05

#Multiple R-squared: 0.2287, Adjusted R-squared: 0.1972   
#p-value: 3.644e-05  
  
#Let's now compare the above regressions for picking the best one -  
  
#Among all the above models, model\_3 has higher Multiple R-square as well as higher adjusted R sqaure as compared to other models, hence I would pick model\_3 with explanatory variables as work\_yrs,s\_avg,sex,gmat\_tpc as the best fit model.  
  
#According to model\_3, explanatory variable work\_yrs is statistically significant with its p-value as 1.28e-05 which is way below the level of significance (alpha) which is 0.05.  
#Then, the overall model's p-value is 3.036e-05 which is also way below alpha 0.05, hence we can accept this model to predict the variations in salary as it means that our model is doing better than the intercept model.  
  
# Our model's equation is as follows : y = b0 + b1\*x1 + b2\*x2 + b3\*x3 + b4\*x4 + e where y is salary, x1 is work\_yrs and b1 is its coefficient, similary for the rest of the explanatory variables added in this model. e denotes the rror term or the residuals and b0 is the intercept (when no regressors are added).  
  
# Our final equation for the model becomes :   
  
# salary = 104867.5 + 2497.9\*work\_yrs + 2674.4\*s\_avg -5282.2\*sex - 147.0\*gmat\_tpc + e  
  
#Thus, for every year increase in years of experience, the salary will go up by 2497.9 and similarly, we can conclude for other variables too.

#Comparing those who didn't get a job whith those who did  
  
#Removing not answered survey responses records   
salary1 <- subset(salary, salary!=998 & salary!=999)  
View(salary1)  
  
#Subset of salary dataframe for people who weren't placed  
notplaced <- subset(salary, salary == 0)  
View(notplaced)  
summary(notplaced)

## age sex gmat\_tot gmat\_qpc   
## Min. :22.00 Min. :1.000 Min. :450.0 Min. :28.00   
## 1st Qu.:25.00 1st Qu.:1.000 1st Qu.:570.0 1st Qu.:68.25   
## Median :27.00 Median :1.000 Median :610.0 Median :82.00   
## Mean :28.51 Mean :1.256 Mean :614.3 Mean :78.91   
## 3rd Qu.:29.75 3rd Qu.:1.750 3rd Qu.:650.0 3rd Qu.:93.00   
## Max. :48.00 Max. :2.000 Max. :760.0 Max. :99.00   
## gmat\_vpc gmat\_tpc s\_avg f\_avg   
## Min. :22.00 Min. : 0.00 Min. :2.000 Min. :0.000   
## 1st Qu.:70.25 1st Qu.:73.50 1st Qu.:2.800 1st Qu.:2.750   
## Median :81.00 Median :86.00 Median :3.000 Median :3.000   
## Mean :77.63 Mean :82.29 Mean :3.031 Mean :3.062   
## 3rd Qu.:89.00 3rd Qu.:93.00 3rd Qu.:3.300 3rd Qu.:3.250   
## Max. :99.00 Max. :99.00 Max. :3.900 Max. :4.000   
## quarter work\_yrs frstlang salary   
## Min. :1.000 Min. : 0.000 Min. :1.000 Min. :0   
## 1st Qu.:2.000 1st Qu.: 2.000 1st Qu.:1.000 1st Qu.:0   
## Median :2.500 Median : 3.000 Median :1.000 Median :0   
## Mean :2.544 Mean : 4.589 Mean :1.089 Mean :0   
## 3rd Qu.:3.000 3rd Qu.: 5.000 3rd Qu.:1.000 3rd Qu.:0   
## Max. :4.000 Max. :22.000 Max. :2.000 Max. :0   
## satis   
## Min. :4.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.622   
## 3rd Qu.:6.000   
## Max. :7.000

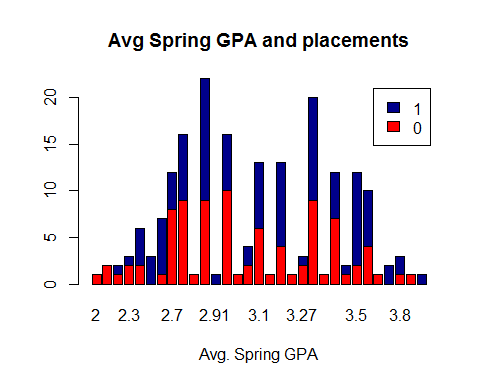
#Let's keep 1 in place of salaries which are greater than 1, as we are only concerned with whether the person got placed or not.  
  
salary1$salary = ifelse(salary1$salary > 1, 1,0)  
View(salary1)  
summary(salary1)

## age sex gmat\_tot gmat\_qpc   
## Min. :22.00 Min. :1.00 Min. :450.0 Min. :28.00   
## 1st Qu.:25.00 1st Qu.:1.00 1st Qu.:570.0 1st Qu.:72.00   
## Median :27.00 Median :1.00 Median :610.0 Median :82.00   
## Mean :27.59 Mean :1.28 Mean :615.2 Mean :79.35   
## 3rd Qu.:29.00 3rd Qu.:2.00 3rd Qu.:650.0 3rd Qu.:91.00   
## Max. :48.00 Max. :2.00 Max. :760.0 Max. :99.00   
## gmat\_vpc gmat\_tpc s\_avg f\_avg   
## Min. :22.00 Min. : 0.00 Min. :2.000 Min. :0.000   
## 1st Qu.:71.00 1st Qu.:75.00 1st Qu.:2.800 1st Qu.:2.750   
## Median :81.00 Median :87.00 Median :3.090 Median :3.000   
## Mean :78.13 Mean :83.48 Mean :3.064 Mean :3.078   
## 3rd Qu.:91.00 3rd Qu.:93.00 3rd Qu.:3.300 3rd Qu.:3.330   
## Max. :99.00 Max. :99.00 Max. :4.000 Max. :4.000   
## quarter work\_yrs frstlang salary   
## Min. :1.000 Min. : 0.000 Min. :1.000 Min. :0.0000   
## 1st Qu.:1.000 1st Qu.: 2.000 1st Qu.:1.000 1st Qu.:0.0000   
## Median :2.000 Median : 3.000 Median :1.000 Median :1.0000   
## Mean :2.394 Mean : 4.104 Mean :1.078 Mean :0.5337   
## 3rd Qu.:3.000 3rd Qu.: 5.000 3rd Qu.:1.000 3rd Qu.:1.0000   
## Max. :4.000 Max. :22.000 Max. :2.000 Max. :1.0000   
## satis   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.762   
## 3rd Qu.:6.000   
## Max. :7.000

#Subset of salary1 dataframe for people who were placed  
placed1 <- subset(salary1, salary == 1)  
View(placed1)  
summary(placed1)

## age sex gmat\_tot gmat\_qpc   
## Min. :22.00 Min. :1.000 Min. :500 Min. :39.00   
## 1st Qu.:25.00 1st Qu.:1.000 1st Qu.:580 1st Qu.:72.00   
## Median :26.00 Median :1.000 Median :620 Median :82.00   
## Mean :26.78 Mean :1.301 Mean :616 Mean :79.73   
## 3rd Qu.:28.00 3rd Qu.:2.000 3rd Qu.:655 3rd Qu.:89.00   
## Max. :40.00 Max. :2.000 Max. :720 Max. :99.00   
## gmat\_vpc gmat\_tpc s\_avg f\_avg   
## Min. :30.00 Min. :51.00 Min. :2.200 Min. :0.000   
## 1st Qu.:71.00 1st Qu.:78.00 1st Qu.:2.850 1st Qu.:2.915   
## Median :81.00 Median :87.00 Median :3.100 Median :3.250   
## Mean :78.56 Mean :84.52 Mean :3.092 Mean :3.091   
## 3rd Qu.:92.00 3rd Qu.:93.50 3rd Qu.:3.400 3rd Qu.:3.415   
## Max. :99.00 Max. :99.00 Max. :4.000 Max. :4.000   
## quarter work\_yrs frstlang salary   
## Min. :1.000 Min. : 0.00 Min. :1.000 Min. :1   
## 1st Qu.:1.000 1st Qu.: 2.00 1st Qu.:1.000 1st Qu.:1   
## Median :2.000 Median : 3.00 Median :1.000 Median :1   
## Mean :2.262 Mean : 3.68 Mean :1.068 Mean :1   
## 3rd Qu.:3.000 3rd Qu.: 4.00 3rd Qu.:1.000 3rd Qu.:1   
## Max. :4.000 Max. :16.00 Max. :2.000 Max. :1   
## satis   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.883   
## 3rd Qu.:6.000   
## Max. :7.000

#Contingency table between salary and s\_avg  
  
mytable2 <- table(salary1$salary,salary1$s\_avg)  
  
#Stacked barplot for placed or not depending on s\_avg  
  
barplot(mytable2, main="Avg Spring GPA and placements",  
 xlab="Avg. Spring GPA", col=c("red","darkblue"),  
 legend = rownames(mytable2))

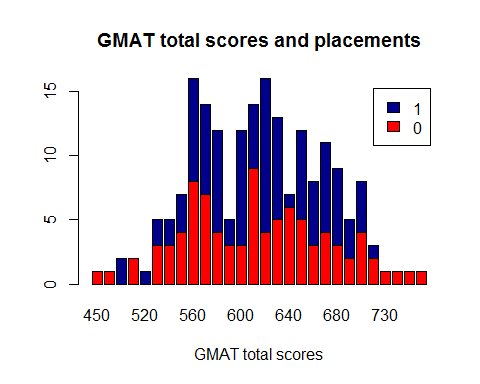


chisq.test(salary1$salary,salary1$s\_avg)

## Warning in chisq.test(salary1$salary, salary1$s\_avg): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: salary1$salary and salary1$s\_avg  
## X-squared = 33.09, df = 30, p-value = 0.3187

#Contingency tables between salary and gmat\_tot  
  
mytable3 <- table(salary1$salary,salary1$gmat\_tot)  
  
#Stacked barplot for placed or not depending on gmat total scores  
  
barplot(mytable3, main="GMAT total scores and placements",  
 xlab="GMAT total scores", col=c("red","darkblue"),  
 legend = rownames(mytable3))

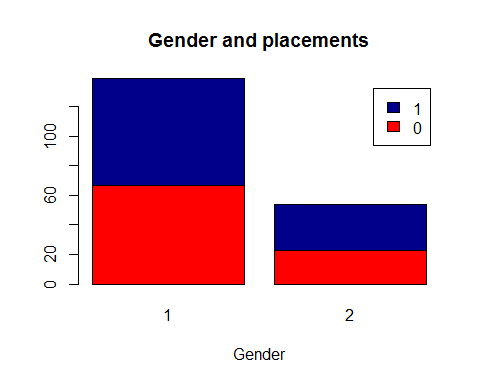


chisq.test(salary1$salary,salary1$gmat\_tot)

## Warning in chisq.test(salary1$salary, salary1$gmat\_tot): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: salary1$salary and salary1$gmat\_tot  
## X-squared = 27.919, df = 27, p-value = 0.4152

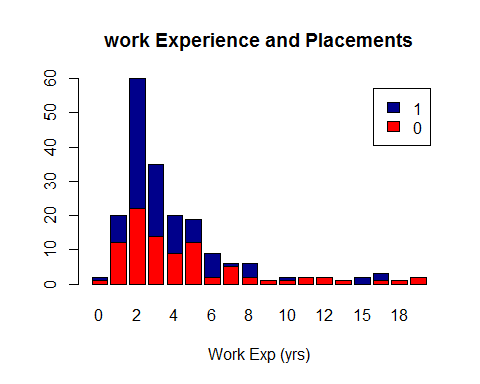
#Contingency tables between salary and gmat\_tot  
  
mytable4 <- table(salary1$salary,salary1$sex)  
  
#Stacked barplot for placed or not depending on gmat total scores  
  
barplot(mytable4, main="Gender and placements",  
 xlab="Gender", col=c("red","darkblue"),  
 legend = rownames(mytable4))



chisq.test(salary1$salary,salary1$sex)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: salary1$salary and salary1$sex  
## X-squared = 0.29208, df = 1, p-value = 0.5889

#Contingency tables between salary and work experience  
  
mytable5 <- table(salary1$salary,salary1$work\_yrs)  
  
#Stacked barplot for placed or not depending on gmat total scores  
  
barplot(mytable5, main="work Experience and Placements",  
 xlab="Work Exp (yrs)", col=c("red","darkblue"),  
 legend = rownames(mytable5))

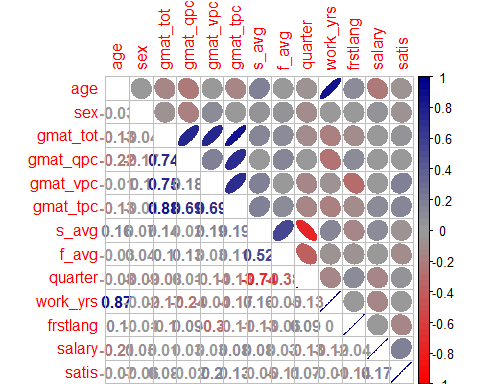


chisq.test(salary1$salary,salary1$work\_yrs)

## Warning in chisq.test(salary1$salary, salary1$work\_yrs): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: salary1$salary and salary1$work\_yrs  
## X-squared = 24.663, df = 17, p-value = 0.1025

#We ran a couple of chi sq tests above but none of them suggests that we reject the null hypothesis as for all the tests, the obtained p-value is not statistically significant.  
  
#Corrgram  
corrplot.mixed(corr=cor(salary1[ , c(1:13)], use="complete.obs"),  
 upper="ellipse", tl.pos="lt",  
 col = colorpanel(50, "red", "gray60", "blue4"))



#No strong correlations are found between the salary varibale and other variables.

#Logistic Regression  
  
#Here, we would run this regression to predict if a student will get placed or not.  
  
#Before proceeding ahead, I would want to convert the categorical variables from int to factors.  
  
salary1 <- salary1 %>% mutate(sex1=factor(sex),firstlang1=factor(frstlang),quarter1=factor(quarter),satis1=factor(satis))  
  
str(salary1)

## 'data.frame': 193 obs. of 17 variables:  
## $ age : int 23 24 24 24 24 25 25 27 27 28 ...  
## $ sex : int 2 1 1 1 1 1 2 1 1 2 ...  
## $ gmat\_tot : int 620 610 670 570 640 610 650 740 750 540 ...  
## $ gmat\_qpc : int 77 90 99 56 82 89 88 99 99 75 ...  
## $ gmat\_vpc : int 87 71 78 81 89 74 89 96 98 50 ...  
## $ gmat\_tpc : int 87 87 95 75 91 87 92 99 99 65 ...  
## $ s\_avg : num 3.4 3.5 3.3 3.3 3.9 3.4 3.3 3.5 3.4 3.6 ...  
## $ f\_avg : num 3 4 3.25 2.67 3.75 3.5 3.75 3.5 3.5 4 ...  
## $ quarter : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ work\_yrs : int 2 2 2 1 2 2 2 3 1 5 ...  
## $ frstlang : int 1 1 1 1 1 1 1 1 2 1 ...  
## $ salary : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ satis : int 7 6 6 7 6 5 6 6 5 5 ...  
## $ sex1 : Factor w/ 2 levels "1","2": 2 1 1 1 1 1 2 1 1 2 ...  
## $ firstlang1: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 2 1 ...  
## $ quarter1 : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...  
## $ satis1 : Factor w/ 5 levels "3","4","5","6",..: 5 4 4 5 4 3 4 4 3 3 ...

salary2 <- subset(salary1,select=c(1,3,4,5,6,7,8,10,12,14,15,16,17))  
  
str(salary2)

## 'data.frame': 193 obs. of 13 variables:  
## $ age : int 23 24 24 24 24 25 25 27 27 28 ...  
## $ gmat\_tot : int 620 610 670 570 640 610 650 740 750 540 ...  
## $ gmat\_qpc : int 77 90 99 56 82 89 88 99 99 75 ...  
## $ gmat\_vpc : int 87 71 78 81 89 74 89 96 98 50 ...  
## $ gmat\_tpc : int 87 87 95 75 91 87 92 99 99 65 ...  
## $ s\_avg : num 3.4 3.5 3.3 3.3 3.9 3.4 3.3 3.5 3.4 3.6 ...  
## $ f\_avg : num 3 4 3.25 2.67 3.75 3.5 3.75 3.5 3.5 4 ...  
## $ work\_yrs : int 2 2 2 1 2 2 2 3 1 5 ...  
## $ salary : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ sex1 : Factor w/ 2 levels "1","2": 2 1 1 1 1 1 2 1 1 2 ...  
## $ firstlang1: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 2 1 ...  
## $ quarter1 : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...  
## $ satis1 : Factor w/ 5 levels "3","4","5","6",..: 5 4 4 5 4 3 4 4 3 3 ...

#Splitting the observations into training and test datasets  
  
train <- salary2[1:154,]  
test <- salary2[155:193,]  
  
names(train)

## [1] "age" "gmat\_tot" "gmat\_qpc" "gmat\_vpc" "gmat\_tpc"   
## [6] "s\_avg" "f\_avg" "work\_yrs" "salary" "sex1"   
## [11] "firstlang1" "quarter1" "satis1"

#Logistic regression model implemented  
  
logit\_model <- glm(salary ~.,family=binomial(link='logit'),data=train)  
summary(logit\_model)

##   
## Call:  
## glm(formula = salary ~ ., family = binomial(link = "logit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2721 -1.1228 0.6802 1.0007 1.7604   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 1.298e+01 1.455e+03 0.009 0.993  
## age -1.538e-01 1.069e-01 -1.439 0.150  
## gmat\_tot -1.144e-02 1.551e-02 -0.737 0.461  
## gmat\_qpc -4.050e-02 5.516e-02 -0.734 0.463  
## gmat\_vpc -3.310e-02 5.275e-02 -0.627 0.530  
## gmat\_tpc 1.322e-01 8.183e-02 1.615 0.106  
## s\_avg 2.367e+00 1.544e+00 1.533 0.125  
## f\_avg 4.792e-03 4.018e-01 0.012 0.990  
## work\_yrs 3.078e-02 1.156e-01 0.266 0.790  
## sex12 -1.221e-01 4.065e-01 -0.300 0.764  
## firstlang12 -3.741e-02 7.158e-01 -0.052 0.958  
## quarter12 2.035e-02 7.235e-01 0.028 0.978  
## quarter13 6.943e-01 1.100e+00 0.631 0.528  
## quarter14 -1.342e+01 1.028e+03 -0.013 0.990  
## satis14 -1.607e+01 1.455e+03 -0.011 0.991  
## satis15 -1.468e+01 1.455e+03 -0.010 0.992  
## satis16 -1.454e+01 1.455e+03 -0.010 0.992  
## satis17 -1.421e+01 1.455e+03 -0.010 0.992  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 212.21 on 153 degrees of freedom  
## Residual deviance: 188.31 on 136 degrees of freedom  
## AIC: 224.31  
##   
## Number of Fisher Scoring iterations: 14

#Taking all variables may not have helped much as none of them are significant acc. to the model. AIC: 224.31

#In the logit model the response variable is log odds: ln(odds) = ln(p/(1-p)) = a\*x1 + b\*x2 + . + z\*xn.  
  
logit\_model\_1 <- glm(salary ~ sex1 + work\_yrs + gmat\_tot + s\_avg + firstlang1,family=binomial(link='logit'),data=train)  
summary(logit\_model\_1)

##   
## Call:  
## glm(formula = salary ~ sex1 + work\_yrs + gmat\_tot + s\_avg + firstlang1,   
## family = binomial(link = "logit"), data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8720 -1.1808 0.7977 1.0387 1.6573   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.5382175 2.5984827 -1.362 0.17331   
## sex12 -0.0264893 0.3693174 -0.072 0.94282   
## work\_yrs -0.0983141 0.0469670 -2.093 0.03633 \*   
## gmat\_tot -0.0007684 0.0031297 -0.246 0.80606   
## s\_avg 1.4842058 0.5720918 2.594 0.00948 \*\*  
## firstlang12 -0.3498053 0.6404454 -0.546 0.58493   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 212.21 on 153 degrees of freedom  
## Residual deviance: 201.53 on 148 degrees of freedom  
## AIC: 213.53  
##   
## Number of Fisher Scoring iterations: 4

#AIC:213.53  
  
#We found here that work\_yrs and s\_svg are statistically significant explanatory variables with p values less than 0.05. As for the statistically significant variables, s\_avg has the lowest p-value suggesting a strong association of the spring average GPA of the student with the probability of getting placed.   
  
#We can interpret this model as : a male would reduce the log odds by 0.0264893 while a unit increase in s\_avg would increase the log odds by 1.4842058.  
  
#Of the two models, the second one has has got lesser AIC, which is desirable and thus we choose that for further analysis.  
  
#Also, we can do ANOVA test on the models, the greater the deviation between null and residuals, the better.  
  
anova(logit\_model, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: salary  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 153 212.22   
## age 1 4.6860 152 207.53 0.030409 \*   
## gmat\_tot 1 0.0167 151 207.51 0.897246   
## gmat\_qpc 1 0.2276 150 207.28 0.633297   
## gmat\_vpc 1 0.8235 149 206.46 0.364148   
## gmat\_tpc 1 3.4272 148 203.03 0.064131 .   
## s\_avg 1 8.4643 147 194.57 0.003622 \*\*  
## f\_avg 1 0.0155 146 194.55 0.900816   
## work\_yrs 1 0.2267 145 194.33 0.633983   
## sex1 1 0.1778 144 194.15 0.673290   
## firstlang1 1 0.0160 143 194.13 0.899267   
## quarter1 3 2.5475 140 191.59 0.466766   
## satis1 4 3.2711 136 188.31 0.513528   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(logit\_model\_1, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: salary  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 153 212.22   
## sex1 1 0.0262 152 212.19 0.871451   
## work\_yrs 1 2.8494 151 209.34 0.091408 .   
## gmat\_tot 1 0.0211 150 209.32 0.884405   
## s\_avg 1 7.4848 149 201.83 0.006222 \*\*  
## firstlang1 1 0.2997 148 201.53 0.584051   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#The drop in the devaince upon dropping a couple variables suggests that logit\_model\_1 is better among the two.

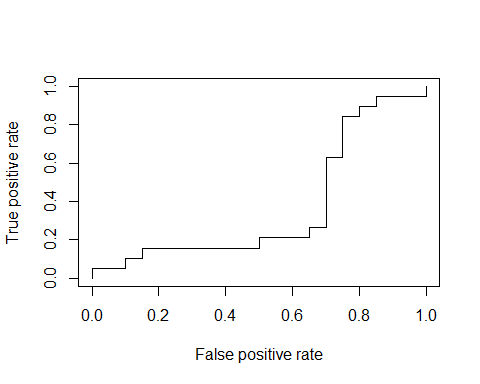
#Assessing predictabilty of the model  
  
fitted.results <- predict(logit\_model\_1,newdata=subset(test,select=c(1,2,3,4,5,6,7,8,10,11,12,13)),type='response')  
fitted.results <- ifelse(fitted.results > 0.5,1,0)  
misClasificError <- mean(fitted.results != test$salary)  
print(paste('Accuracy',1-misClasificError))

## [1] "Accuracy 0.333333333333333"

#ROC and AUC curve for performance measures : As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5.  
  
library(ROCR)

## Warning: package 'ROCR' was built under R version 3.4.1

p <- predict(logit\_model\_1, newdata=subset(test,select=c(1,2,3,4,5,6,7,8,10,11,12,13)), type="response")  
pr <- prediction(p, test$salary)  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf)



auc <- performance(pr, measure = "auc")  
auc <- auc@y.values[[1]]  
auc

## [1] 0.3710526