MBA Starting Salaries

Aman Gupta

February 20, 2018

### Read the data

mba.df <- read.csv(paste("MBA Starting Salaries Data.csv", sep=""))  
View(mba.df)

### Attach the dataframe

attach(mba.df)

### Summarize the data

summary(mba.df)

## 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

library(psych)  
describe(mba.df)

## vars n mean sd median trimmed mad min max  
## age 1 274 27.36 3.71 27 26.76 2.97 22 48  
## sex 2 274 1.25 0.43 1 1.19 0.00 1 2  
## gmat\_tot 3 274 619.45 57.54 620 618.86 59.30 450 790  
## gmat\_qpc 4 274 80.64 14.87 83 82.31 14.83 28 99  
## gmat\_vpc 5 274 78.32 16.86 81 80.33 14.83 16 99  
## gmat\_tpc 6 274 84.20 14.02 87 86.12 11.86 0 99  
## s\_avg 7 274 3.03 0.38 3 3.03 0.44 2 4  
## f\_avg 8 274 3.06 0.53 3 3.09 0.37 0 4  
## quarter 9 274 2.48 1.11 2 2.47 1.48 1 4  
## work\_yrs 10 274 3.87 3.23 3 3.29 1.48 0 22  
## frstlang 11 274 1.12 0.32 1 1.02 0.00 1 2  
## salary 12 274 39025.69 50951.56 999 33607.86 1481.12 0 220000  
## satis 13 274 172.18 371.61 6 91.50 1.48 1 998  
## range skew kurtosis se  
## age 26 2.16 6.45 0.22  
## sex 1 1.16 -0.66 0.03  
## gmat\_tot 340 -0.01 0.06 3.48  
## gmat\_qpc 71 -0.92 0.30 0.90  
## gmat\_vpc 83 -1.04 0.74 1.02  
## gmat\_tpc 99 -2.28 9.02 0.85  
## s\_avg 2 -0.06 -0.38 0.02  
## f\_avg 4 -2.08 10.85 0.03  
## quarter 3 0.02 -1.35 0.07  
## work\_yrs 22 2.78 9.80 0.20  
## frstlang 1 2.37 3.65 0.02  
## salary 220000 0.70 -1.05 3078.10  
## satis 997 1.77 1.13 22.45

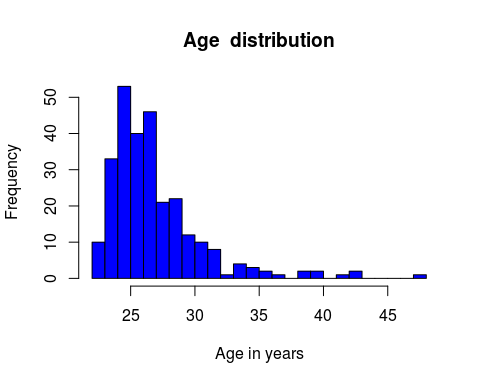
### Data Types

str(mba.df)

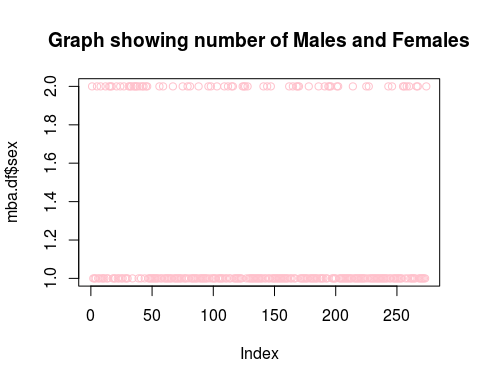
## 'data.frame': 274 obs. of 13 variables:  
## $ age : int 23 24 24 24 24 24 25 25 25 25 ...  
## $ sex : int 2 1 1 1 2 1 1 2 1 1 ...  
## $ gmat\_tot: int 620 610 670 570 710 640 610 650 630 680 ...  
## $ gmat\_qpc: int 77 90 99 56 93 82 89 88 79 99 ...  
## $ gmat\_vpc: int 87 71 78 81 98 89 74 89 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 ...

# Visualization

hist(mba.df$age, breaks=20,col="blue",xlab="Age in years", main="Age distribution")



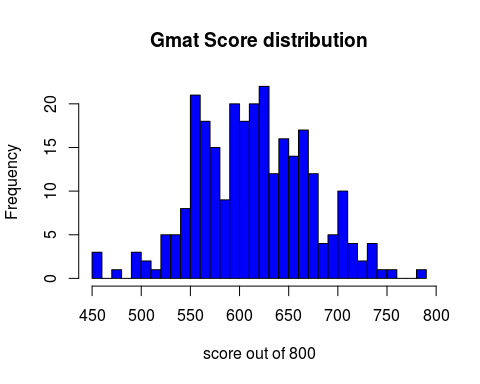
plot(mba.df$sex,main = "Graph showing number of Males and Females",col="pink")



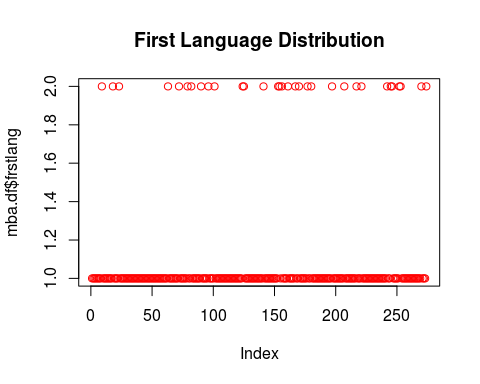
hist(mba.df$work\_yrs, breaks=20,col="blue",xlab="Work Experience in years", main="Work experience distribution")



hist(mba.df$gmat\_tot, breaks=40,col="blue",xlab="score out of 800", main="Gmat Score distribution")



plot(mba.df$frstlang,main = "First Language Distribution",col="red")



newdata <- mba.df[ which(mba.df$satis<='7'), ]  
hist(newdata$satis, breaks=5,col="yellow",xlab="Degree of Satisfaction,1=low 7=high", main="Satisfaction distribution")



newdata1 <- mba.df[ which(mba.df$salary !="998" & mba.df$salary !="999"), ]  
hist(newdata1$salary, breaks=10,col="yellow",xlab="starting salary", main="Salary distribution")



aggregate(cbind(salary, work\_yrs, age) ~ sex, data = mba.df, mean) # Effect of gender on salary

## sex salary work\_yrs age  
## 1 1 37013.62 3.893204 27.41748  
## 2 2 45121.07 3.808824 27.17647

boxplot(salary ~ sex ,data=mba.df,col = c("magenta","green"), main="Effect of Gender on Salary", ylab="Gender", xlab="Starting Salary")



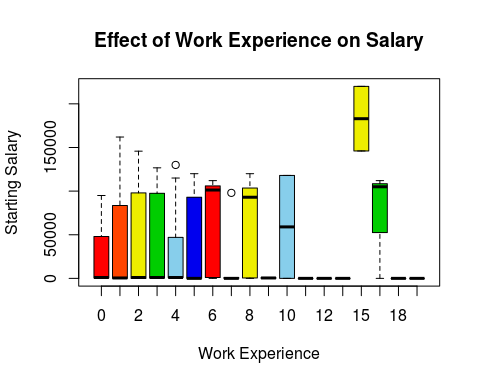
aggregate(cbind(salary, work\_yrs) ~ age, data = mba.df, mean) # Effect of age on Salary

## age salary work\_yrs  
## 1 22 42500.00 1.000000  
## 2 23 57282.00 1.750000  
## 3 24 49342.24 1.727273  
## 4 25 43395.55 2.264151  
## 5 26 35982.07 2.875000  
## 6 27 31499.37 3.130435  
## 7 28 39809.00 4.666667  
## 8 29 28067.95 4.500000  
## 9 30 55291.25 5.583333  
## 10 31 40599.40 5.800000  
## 11 32 13662.25 5.625000  
## 12 33 118000.00 10.000000  
## 13 34 26250.00 11.500000  
## 14 35 0.00 9.333333  
## 15 36 0.00 12.500000  
## 16 37 0.00 9.000000  
## 17 39 56000.00 10.500000  
## 18 40 183000.00 15.000000  
## 19 42 0.00 13.000000  
## 20 43 0.00 19.000000  
## 21 48 0.00 22.000000

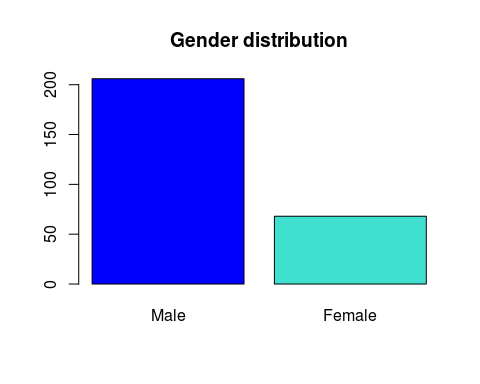
aggregate(cbind(salary, work\_yrs) ~ satis , data = mba.df, mean) # Effect of Salary on the Satisfaction level

## satis salary work\_yrs  
## 1 1 999.000 3.000000  
## 2 2 999.000 2.000000  
## 3 3 19799.200 4.200000  
## 4 4 6293.412 2.941176  
## 5 5 40476.311 4.243243  
## 6 6 54383.536 4.185567  
## 7 7 65718.152 3.727273  
## 8 998 998.000 3.086957

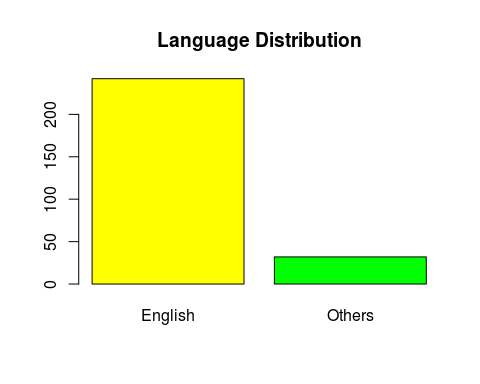
boxplot(salary ~ work\_yrs ,data=mba.df, main="Effect of Work Experience on Salary", xlab="Work Experience", ylab="Starting Salary",col=c("red","orangered","yellow2","green3","skyblue","blue2"))



mba.df$sex=factor(mba.df$sex, levels=c(1,2), labels=c("Male","Female"))  
plot(mba.df$sex,col = c("blue","turquoise"),main = "Gender distribution")



mba.df$frstlang = factor(mba.df$frstlang, levels=c(1,2), labels=c("English","Others"))  
plot(mba.df$frstlang,col=c("yellow","green"),main = "Language Distribution")



# Scatter Plots

library(car)

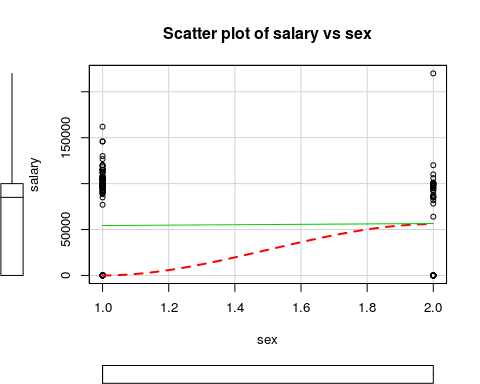
##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

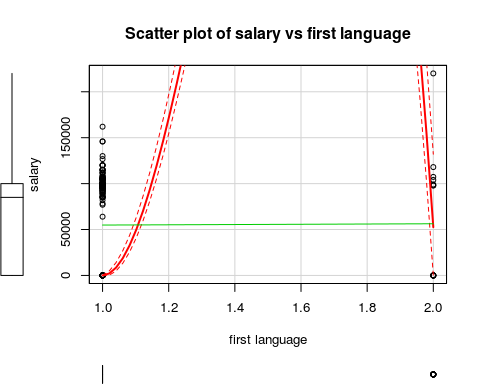
scatterplot(salary ~age, data=newdata1,  
 spread=FALSE, smoother.args=list(lty=2),  
 main="Scatter plot of salary vs age",  
 xlab="age",  
 ylab="salary")



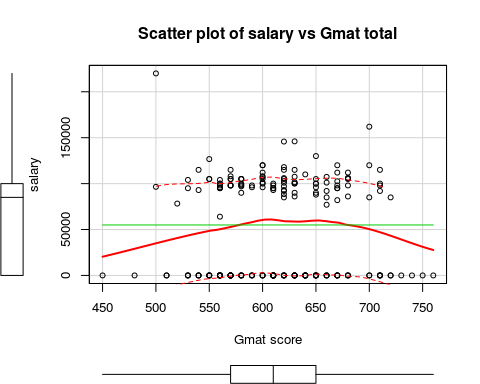
scatterplot(salary ~sex, data=newdata1,  
 spread=FALSE, smoother.args=list(lty=2),  
 main="Scatter plot of salary vs sex",  
 xlab="sex",  
 ylab="salary")



scatterplot(salary ~frstlang, data=newdata1,  
 main="Scatter plot of salary vs first language",  
 xlab="first language",  
 ylab="salary")



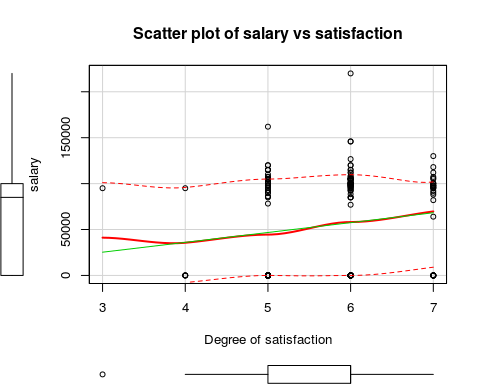
scatterplot(salary ~gmat\_tot, data=newdata1,  
 main="Scatter plot of salary vs Gmat total",  
 xlab="Gmat score",  
 ylab="salary")



scatterplot(salary ~work\_yrs, data=newdata1,  
 main="Scatter plot of salary vs Work exp.",  
 xlab="Work experience in years",  
 ylab="salary")

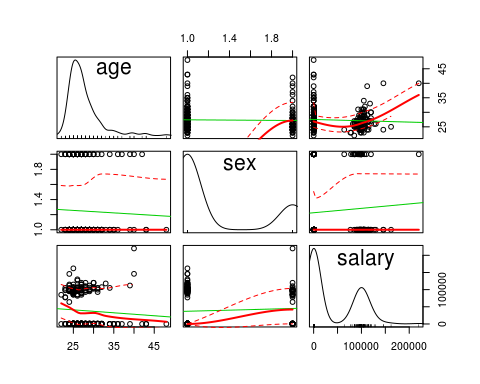


scatterplot(salary ~satis, data=newdata1,  
 main="Scatter plot of salary vs satisfaction",  
 xlab="Degree of satisfaction",  
 ylab="salary")

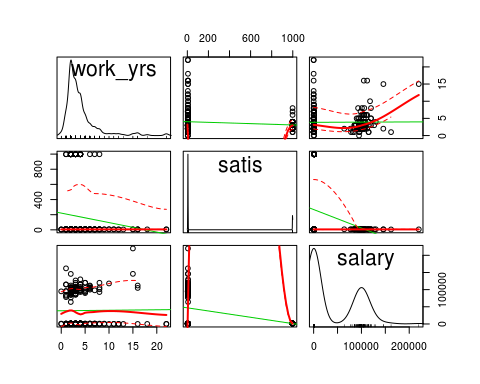


# Scatterplot Matrix

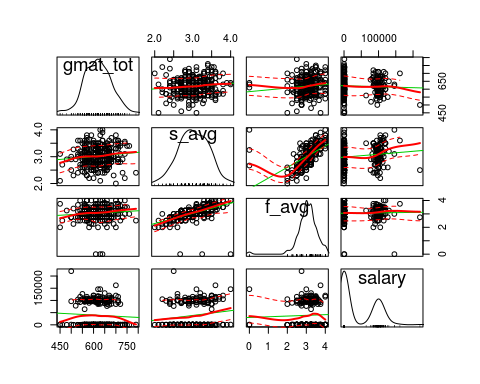
scatterplotMatrix(~age+sex+salary, data=mba.df)



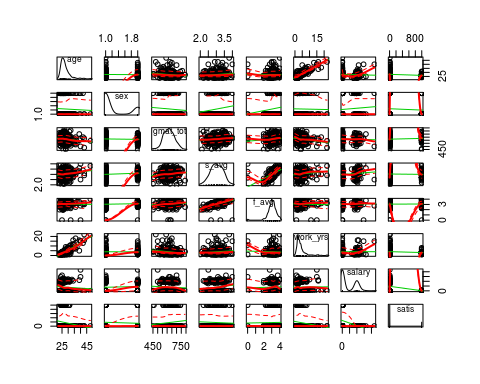
scatterplotMatrix(~work\_yrs+satis+salary, data=mba.df)



scatterplotMatrix(~gmat\_tot+s\_avg+f\_avg+salary, data=mba.df)

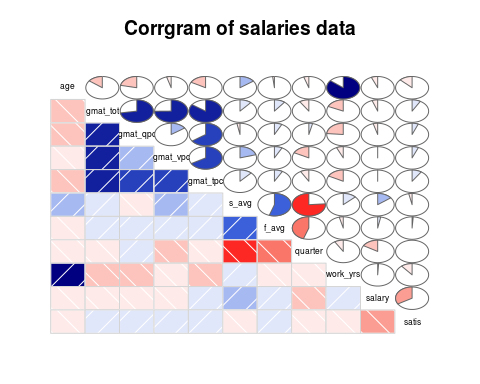


scatterplotMatrix(~age+sex+gmat\_tot+s\_avg+f\_avg+work\_yrs+salary+satis, data=mba.df)



# Corrgram

library(corrgram)  
corrgram(mba.df, order=FALSE,   
 lower.panel=panel.shade,  
 upper.panel=panel.pie,   
 text.panel=panel.txt,  
 main="Corrgram of salaries data")



# Correlation Matrix

correlationmatrix <- cor(mba.df[,c(3:10,12,13)])  
round(correlationmatrix,digits = 2)

## gmat\_tot gmat\_qpc gmat\_vpc gmat\_tpc s\_avg f\_avg quarter work\_yrs  
## gmat\_tot 1.00 0.72 0.75 0.85 0.11 0.10 -0.09 -0.18  
## gmat\_qpc 0.72 1.00 0.15 0.65 -0.03 0.07 0.04 -0.24  
## gmat\_vpc 0.75 0.15 1.00 0.67 0.20 0.08 -0.17 -0.07  
## gmat\_tpc 0.85 0.65 0.67 1.00 0.12 0.08 -0.08 -0.17  
## s\_avg 0.11 -0.03 0.20 0.12 1.00 0.55 -0.76 0.13  
## f\_avg 0.10 0.07 0.08 0.08 0.55 1.00 -0.45 -0.04  
## quarter -0.09 0.04 -0.17 -0.08 -0.76 -0.45 1.00 -0.09  
## work\_yrs -0.18 -0.24 -0.07 -0.17 0.13 -0.04 -0.09 1.00  
## salary -0.05 -0.04 -0.01 0.00 0.15 0.03 -0.16 0.01  
## satis 0.08 0.06 0.06 0.09 -0.03 0.01 0.00 -0.11  
## salary satis  
## gmat\_tot -0.05 0.08  
## gmat\_qpc -0.04 0.06  
## gmat\_vpc -0.01 0.06  
## gmat\_tpc 0.00 0.09  
## s\_avg 0.15 -0.03  
## f\_avg 0.03 0.01  
## quarter -0.16 0.00  
## work\_yrs 0.01 -0.11  
## salary 1.00 -0.34  
## satis -0.34 1.00

# Variance- Covariance Matrix

VarianceCovariancematrix <- var(mba.df[,1:13])

## Warning in var(mba.df[, 1:13]): NAs introduced by coercion

round(VarianceCovariancematrix, 2)

## age sex gmat\_tot gmat\_qpc gmat\_vpc gmat\_tpc s\_avg  
## age 13.77 NA -31.16 -11.93 -2.76 -8.84 0.21  
## sex NA NA NA NA NA NA NA  
## gmat\_tot -31.16 NA 3310.69 620.02 726.00 683.99 2.48  
## gmat\_qpc -11.93 NA 620.02 221.07 38.15 135.80 -0.17  
## gmat\_vpc -2.76 NA 726.00 38.15 284.25 157.49 1.31  
## gmat\_tpc -8.84 NA 683.99 135.80 157.49 196.61 0.63  
## s\_avg 0.21 NA 2.48 -0.17 1.31 0.63 0.15  
## f\_avg -0.03 NA 3.15 0.58 0.67 0.59 0.11  
## quarter -0.20 NA -5.89 0.60 -3.27 -1.29 -0.32  
## work\_yrs 10.29 NA -33.92 -11.37 -3.62 -7.86 0.16  
## frstlang NA NA NA NA NA NA NA  
## salary -11830.42 NA -161159.99 -33358.23 -5273.85 3522.75 2831.60  
## satis -176.35 NA 1765.26 334.84 392.36 484.25 -4.63  
## f\_avg quarter work\_yrs frstlang salary satis  
## age -0.03 -0.20 10.29 NA -11830.42 -176.35  
## sex NA NA NA NA NA NA  
## gmat\_tot 3.15 -5.89 -33.92 NA -161159.99 1765.26  
## gmat\_qpc 0.58 0.60 -11.37 NA -33358.23 334.84  
## gmat\_vpc 0.67 -3.27 -3.62 NA -5273.85 392.36  
## gmat\_tpc 0.59 -1.29 -7.86 NA 3522.75 484.25  
## s\_avg 0.11 -0.32 0.16 NA 2831.60 -4.63  
## f\_avg 0.28 -0.26 -0.07 NA 787.66 2.13  
## quarter -0.26 1.23 -0.31 NA -9296.21 -0.01  
## work\_yrs -0.07 -0.31 10.45 NA 1486.15 -131.24  
## frstlang NA NA NA NA NA NA  
## salary 787.66 -9296.21 1486.15 NA 2596061571.52 -6347115.38  
## satis 2.13 -0.01 -131.24 NA -6347115.38 138097.38

# Dataframe of those who were placed

placed.df <- mba.df[ which(mba.df$salary !="998" & mba.df$salary !="999" & mba.df$salary!="0"), ]  
head(placed.df)

## age sex gmat\_tot gmat\_qpc gmat\_vpc gmat\_tpc s\_avg f\_avg quarter  
## 35 22 Female 660 90 92 94 3.5 3.75 1  
## 36 27 Female 700 94 98 98 3.3 3.25 1  
## 37 25 Female 680 87 96 96 3.5 2.67 1  
## 38 25 Female 650 82 91 93 3.4 3.25 1  
## 39 27 Male 710 96 96 98 3.3 3.50 1  
## 40 28 Female 620 52 98 87 3.4 3.75 1  
## work\_yrs frstlang salary satis  
## 35 1 English 85000 5  
## 36 2 English 85000 6  
## 37 2 English 86000 5  
## 38 3 English 88000 7  
## 39 2 English 92000 6  
## 40 5 English 93000 5

# Contingency tables showing the affect of various factors on the starting salary

t1 <- xtabs(~salary+sex,data=placed.df)  
t1

## sex  
## salary Male Female  
## 64000 0 1  
## 77000 1 0  
## 78256 0 1  
## 82000 0 1  
## 85000 1 3  
## 86000 0 2  
## 88000 0 1  
## 88500 1 0  
## 90000 3 0  
## 92000 2 1  
## 93000 2 1  
## 95000 4 3  
## 96000 3 1  
## 96500 1 0  
## 97000 2 0  
## 98000 6 4  
## 99000 0 1  
## 100000 4 5  
## 100400 1 0  
## 101000 0 2  
## 101100 1 0  
## 101600 1 0  
## 102500 1 0  
## 103000 1 0  
## 104000 2 0  
## 105000 11 0  
## 106000 2 1  
## 107000 1 0  
## 107300 1 0  
## 107500 1 0  
## 108000 2 0  
## 110000 0 1  
## 112000 3 0  
## 115000 5 0  
## 118000 1 0  
## 120000 3 1  
## 126710 1 0  
## 130000 1 0  
## 145800 1 0  
## 146000 1 0  
## 162000 1 0  
## 220000 0 1

From this table it is evident that mostly men have higher starting salaries compared to women.

t2 <- xtabs(~salary+work\_yrs,data=placed.df)  
t2

## work\_yrs  
## salary 0 1 2 3 4 5 6 7 8 10 15 16  
## 64000 0 0 1 0 0 0 0 0 0 0 0 0  
## 77000 0 0 1 0 0 0 0 0 0 0 0 0  
## 78256 0 1 0 0 0 0 0 0 0 0 0 0  
## 82000 0 1 0 0 0 0 0 0 0 0 0 0  
## 85000 0 1 2 1 0 0 0 0 0 0 0 0  
## 86000 0 0 1 1 0 0 0 0 0 0 0 0  
## 88000 0 0 0 1 0 0 0 0 0 0 0 0  
## 88500 0 0 0 1 0 0 0 0 0 0 0 0  
## 90000 0 0 2 0 0 1 0 0 0 0 0 0  
## 92000 0 0 3 0 0 0 0 0 0 0 0 0  
## 93000 0 0 0 0 1 1 0 0 1 0 0 0  
## 95000 1 1 2 2 0 1 0 0 0 0 0 0  
## 96000 0 1 2 0 1 0 0 0 0 0 0 0  
## 96500 0 0 1 0 0 0 0 0 0 0 0 0  
## 97000 0 0 0 1 1 0 0 0 0 0 0 0  
## 98000 0 0 7 1 1 0 0 1 0 0 0 0  
## 99000 0 0 0 0 0 1 0 0 0 0 0 0  
## 100000 0 0 6 1 1 0 1 0 0 0 0 0  
## 100400 0 0 0 1 0 0 0 0 0 0 0 0  
## 101000 0 0 2 0 0 0 0 0 0 0 0 0  
## 101100 0 0 0 0 0 0 0 0 1 0 0 0  
## 101600 0 0 0 1 0 0 0 0 0 0 0 0  
## 102500 0 0 0 0 0 0 1 0 0 0 0 0  
## 103000 0 0 0 1 0 0 0 0 0 0 0 0  
## 104000 0 0 0 0 2 0 0 0 0 0 0 0  
## 105000 0 0 4 4 0 1 1 0 0 0 0 1  
## 106000 0 0 0 0 0 0 2 0 1 0 0 0  
## 107000 0 0 1 0 0 0 0 0 0 0 0 0  
## 107300 0 0 1 0 0 0 0 0 0 0 0 0  
## 107500 0 0 0 1 0 0 0 0 0 0 0 0  
## 108000 0 0 0 1 1 0 0 0 0 0 0 0  
## 110000 0 0 0 0 0 0 1 0 0 0 0 0  
## 112000 0 0 1 0 0 0 1 0 0 0 0 1  
## 115000 0 2 0 1 2 0 0 0 0 0 0 0  
## 118000 0 0 0 0 0 0 0 0 0 1 0 0  
## 120000 0 0 0 1 0 2 0 0 1 0 0 0  
## 126710 0 0 0 1 0 0 0 0 0 0 0 0  
## 130000 0 0 0 0 1 0 0 0 0 0 0 0  
## 145800 0 0 1 0 0 0 0 0 0 0 0 0  
## 146000 0 0 0 0 0 0 0 0 0 0 1 0  
## 162000 0 1 0 0 0 0 0 0 0 0 0 0  
## 220000 0 0 0 0 0 0 0 0 0 0 1 0

From the above table it is evident that a minimum of 2 years of work experience is necessary for a good salary.

t3 <- xtabs(~salary+gmat\_tot,data=placed.df)  
t3

## gmat\_tot  
## salary 500 520 530 540 550 560 570 580 590 600 610 620 630 640 650 660  
## 64000 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0  
## 77000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1  
## 78256 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 82000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 85000 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1  
## 86000 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0  
## 88000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0  
## 88500 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0  
## 90000 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0  
## 92000 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1  
## 93000 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0  
## 95000 0 0 1 0 0 2 0 0 0 0 2 0 0 0 0 0  
## 96000 0 0 0 0 0 1 0 0 1 1 0 0 0 0 1 0  
## 96500 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 97000 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0  
## 98000 0 0 0 0 0 1 3 1 1 0 1 0 0 0 0 0  
## 99000 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0  
## 100000 0 0 0 0 0 2 0 1 0 1 1 0 1 0 2 0  
## 100400 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0  
## 101000 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0  
## 101100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1  
## 101600 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0  
## 102500 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 103000 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0  
## 104000 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0  
## 105000 0 0 0 0 2 0 2 3 0 1 0 1 0 0 1 0  
## 106000 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0  
## 107000 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  
## 107300 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1  
## 107500 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0  
## 108000 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0  
## 110000 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0  
## 112000 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0  
## 115000 0 0 0 1 0 0 1 0 0 0 0 1 1 0 0 0  
## 118000 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0  
## 120000 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0  
## 126710 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0  
## 130000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0  
## 145800 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0  
## 146000 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0  
## 162000 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## 220000 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  
## gmat\_tot  
## salary 670 680 700 710 720  
## 64000 0 0 0 0 0  
## 77000 0 0 0 0 0  
## 78256 0 0 0 0 0  
## 82000 1 0 0 0 0  
## 85000 0 0 1 0 1  
## 86000 0 1 0 0 0  
## 88000 0 0 0 0 0  
## 88500 0 0 0 0 0  
## 90000 0 0 0 0 0  
## 92000 0 0 0 1 0  
## 93000 0 0 0 0 0  
## 95000 2 0 0 0 0  
## 96000 0 0 0 0 0  
## 96500 0 0 0 0 0  
## 97000 0 0 0 0 0  
## 98000 1 1 0 1 0  
## 99000 0 0 0 0 0  
## 100000 0 0 0 1 0  
## 100400 0 0 0 0 0  
## 101000 0 0 0 0 0  
## 101100 0 0 0 0 0  
## 101600 0 0 0 0 0  
## 102500 1 0 0 0 0  
## 103000 0 0 0 0 0  
## 104000 0 0 0 0 0  
## 105000 0 1 0 0 0  
## 106000 0 2 0 0 0  
## 107000 0 0 0 0 0  
## 107300 0 0 0 0 0  
## 107500 0 0 0 0 0  
## 108000 0 0 0 0 0  
## 110000 0 0 0 0 0  
## 112000 1 1 0 0 0  
## 115000 0 0 0 1 0  
## 118000 0 0 0 0 0  
## 120000 1 0 1 0 0  
## 126710 0 0 0 0 0  
## 130000 0 0 0 0 0  
## 145800 0 0 0 0 0  
## 146000 0 0 0 0 0  
## 162000 0 0 1 0 0  
## 220000 0 0 0 0 0

Generally, people with high Gmat Score also have high salaries.

t4 <-xtabs(~salary+frstlang,data=placed.df)  
t4

## frstlang  
## salary English Others  
## 64000 1 0  
## 77000 1 0  
## 78256 1 0  
## 82000 1 0  
## 85000 4 0  
## 86000 2 0  
## 88000 1 0  
## 88500 1 0  
## 90000 3 0  
## 92000 3 0  
## 93000 3 0  
## 95000 7 0  
## 96000 4 0  
## 96500 1 0  
## 97000 2 0  
## 98000 8 2  
## 99000 0 1  
## 100000 9 0  
## 100400 1 0  
## 101000 2 0  
## 101100 1 0  
## 101600 1 0  
## 102500 1 0  
## 103000 1 0  
## 104000 1 1  
## 105000 11 0  
## 106000 3 0  
## 107000 1 0  
## 107300 0 1  
## 107500 1 0  
## 108000 2 0  
## 110000 1 0  
## 112000 3 0  
## 115000 5 0  
## 118000 0 1  
## 120000 4 0  
## 126710 1 0  
## 130000 1 0  
## 145800 1 0  
## 146000 1 0  
## 162000 1 0  
## 220000 0 1

Employees with English as first language are mostly preferred and are given higher salaries compared to those who don’t have English as their first language.

### Chi-squared test

chisq.test(placed.df$age,placed.df$salary)

## Warning in chisq.test(placed.df$age, placed.df$salary): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: placed.df$age and placed.df$salary  
## X-squared = 717.62, df = 574, p-value = 3.929e-05

chisq.test(placed.df$sex,placed.df$salary)

## Warning in chisq.test(placed.df$sex, placed.df$salary): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: placed.df$sex and placed.df$salary  
## X-squared = 52.681, df = 41, p-value = 0.1045

chisq.test(placed.df$gmat\_tot,placed.df$salary)

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

##   
## Pearson's Chi-squared test  
##   
## data: placed.df$gmat\_tot and placed.df$salary  
## X-squared = 927.24, df = 820, p-value = 0.005279

chisq.test(placed.df$s\_avg,placed.df$salary)

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

##   
## Pearson's Chi-squared test  
##   
## data: placed.df$s\_avg and placed.df$salary  
## X-squared = 792.97, df = 861, p-value = 0.9524

chisq.test(placed.df$f\_avg,placed.df$salary)

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

##   
## Pearson's Chi-squared test  
##   
## data: placed.df$f\_avg and placed.df$salary  
## X-squared = 596.28, df = 574, p-value = 0.2518

chisq.test(placed.df$work\_yrs,placed.df$salary)

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

##   
## Pearson's Chi-squared test  
##   
## data: placed.df$work\_yrs and placed.df$salary  
## X-squared = 535.23, df = 451, p-value = 0.003809

chisq.test(placed.df$frstlang,placed.df$salary)

## Warning in chisq.test(placed.df$frstlang, placed.df$salary): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: placed.df$frstlang and placed.df$salary  
## X-squared = 69.847, df = 41, p-value = 0.003296

The results of the Chi-Squared tests tell us that age, GMAT percentiles, work experience and first language are factors that are statistically significant for starting salary (p < 0.05), whereas gender, average GPA for Spring and Fall semesters and quartile ranking with degree are not statistically significant for salary (p > 0.05).

##### T-test

log.transformed.salary=log(placed.df$salary)  
t.test(log.transformed.salary~ placed.df$sex, var.equal = TRUE)

##   
## Two Sample t-test  
##   
## data: log.transformed.salary by placed.df$sex  
## t = 2.4552, df = 101, p-value = 0.01579  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.01470674 0.13847594  
## sample estimates:  
## mean in group Male mean in group Female   
## 11.55390 11.47731

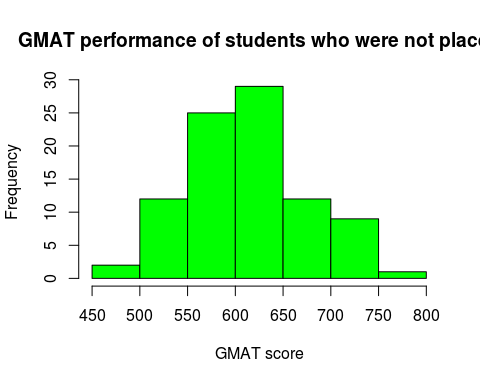
This T-test shows that there is a significant difference in salaries of men and women.

# Dataset consisting of people who were not placed

notPlaced.df <- mba.df[ which(mba.df$salary !="998" & mba.df$salary !="999" & mba.df$salary==0), ]  
head(notPlaced.df)

## age sex gmat\_tot gmat\_qpc gmat\_vpc gmat\_tpc s\_avg f\_avg quarter  
## 1 23 Female 620 77 87 87 3.4 3.00 1  
## 2 24 Male 610 90 71 87 3.5 4.00 1  
## 3 24 Male 670 99 78 95 3.3 3.25 1  
## 4 24 Male 570 56 81 75 3.3 2.67 1  
## 6 24 Male 640 82 89 91 3.9 3.75 1  
## 7 25 Male 610 89 74 87 3.4 3.50 1  
## work\_yrs frstlang salary satis  
## 1 2 English 0 7  
## 2 2 English 0 6  
## 3 2 English 0 6  
## 4 1 English 0 7  
## 6 2 English 0 6  
## 7 2 English 0 5

hist(notPlaced.df$gmat\_tot,  
 main = "GMAT performance of students who were not placed",  
 xlab="GMAT score",  
 breaks=10,  
 col = "green")

 GMAT score is distributed between 550-650 for unplaced students while it is more scattered amongst those who do have a job.

chisq.test(notPlaced.df$work\_yrs,notPlaced.df$satis)

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

##   
## Pearson's Chi-squared test  
##   
## data: notPlaced.df$work\_yrs and notPlaced.df$satis  
## X-squared = 44.974, df = 48, p-value = 0.5976

This shows that the unplaced students with work experience are satisfied with the MBA program.

# Regression Analysis

### Preparing for regression analysis

mba.df$sex[mba.df$sex == 1] <- 'Male'   
mba.df$sex[mba.df$sex == 2] <- 'Female'  
mba.df$sex <- factor(mba.df$sex)   
mba.df$frstlang[mba.df$frstlang == 1] <- 'English'   
mba.df$frstlang[mba.df$frstlang == 2] <- 'Other'

## Warning in `[<-.factor`(`\*tmp\*`, mba.df$frstlang == 2, value =  
## structure(c(1L, : invalid factor level, NA generated

mba.df$frstlang <- factor(mba.df$frstlang)

### Model 1

fit1 <- lm(salary ~ gmat\_tot + gmat\_vpc + gmat\_qpc + gmat\_tpc , data=placed.df)  
summary(fit1)

##   
## Call:  
## lm(formula = salary ~ gmat\_tot + gmat\_vpc + gmat\_qpc + gmat\_tpc,   
## data = placed.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40370 -8250 -2164 5253 100097   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 109539.54 48054.24 2.279 0.0248 \*  
## gmat\_tot 55.01 181.71 0.303 0.7627   
## gmat\_vpc 546.10 543.85 1.004 0.3178   
## gmat\_qpc 718.40 541.90 1.326 0.1880   
## gmat\_tpc -1663.16 801.57 -2.075 0.0406 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17670 on 98 degrees of freedom  
## Multiple R-squared: 0.06089, Adjusted R-squared: 0.02256   
## F-statistic: 1.589 on 4 and 98 DF, p-value: 0.1834

Gmat\_tpc is a significant variable in model 1 The multiple R squared value indicates that the model accounts for 6% of the variance in the variables The residual error (17670) can be thought of as the average error in predicting salary using the various gmat data available.

### Model 2

fit2 <- lm(salary ~ frstlang + satis + work\_yrs , data=placed.df)  
summary(fit2)

##   
## Call:  
## lm(formula = salary ~ frstlang + satis + work\_yrs, data = placed.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31764 -9640 -604 4816 76193   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 104142.2 11899.4 8.752 5.73e-14 \*\*\*  
## frstlangOthers 13541.5 6305.7 2.147 0.0342 \*   
## satis -1913.1 2000.0 -0.957 0.3411   
## work\_yrs 2506.8 528.6 4.742 7.11e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15740 on 99 degrees of freedom  
## Multiple R-squared: 0.2466, Adjusted R-squared: 0.2237   
## F-statistic: 10.8 on 3 and 99 DF, p-value: 3.354e-06

work\_yrs and frstlang are significant variables in model 2 The multiple R squared value indicates that the model accounts for 24.66% of the variance in the variables The residual error(15740) can be thought of as the average error in predicting salary using work experience, job satisfaction and first language.

### Model 3

fit3 <- lm(salary ~ s\_avg + f\_avg , data=placed.df)  
summary(fit3)

##   
## Call:  
## lm(formula = salary ~ s\_avg + f\_avg, data = placed.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41509 -7388 -1723 3119 119810   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 97277 15352 6.336 6.8e-09 \*\*\*  
## s\_avg 8781 5171 1.698 0.0926 .   
## f\_avg -6924 4013 -1.725 0.0875 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17690 on 100 degrees of freedom  
## Multiple R-squared: 0.03896, Adjusted R-squared: 0.01974   
## F-statistic: 2.027 on 2 and 100 DF, p-value: 0.1371

##### We can see that model 2 is better than model 1 and model 3, with a higher R-squared value.

# beta coefficients  
fit2$coefficients

## (Intercept) frstlangOthers satis work\_yrs   
## 104142.167 13541.466 -1913.088 2506.764

# confidence intervals  
confint(fit2)

## 2.5 % 97.5 %  
## (Intercept) 80531.137 127753.197  
## frstlangOthers 1029.606 26053.326  
## satis -5881.593 2055.418  
## work\_yrs 1457.812 3555.716

### Visualizing the beta coefficients

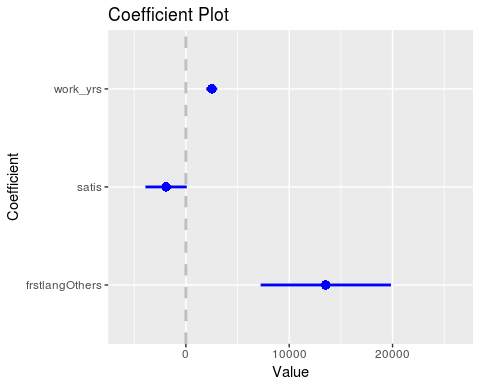
library(coefplot)

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

library(ggplot2)  
coefplot(fit2, predictors=c("work\_yrs", "frstlang", "satis"))



### Executive Summary

* The starting salary of the Mba program of any individual student depends critically on the first language of the student and the degree of satisfaction estimated through various boxplots and the scatterplots.
* Even from the corrogram and the correlation matrices , it is quite clear that the starting salaries are strongly correlated with the first language.
* From the chi- squared tests and the t-tests between the people who got a job and those who did not get a job , it can be analysed that there is a significant relationship between the starting salaries , degree of satisfaction of the MBA program and the first language of the people.
* The Regression model ,i.e. the best fit model , here the second model helps us in concluding that the salary has more or less a significant effect from work years experience, first language and satisfaction degree.