Ankit\_Kashyap\_MBA\_Salary\_Code.r

Ankit

Thu Jul 06 03:15:49 2017

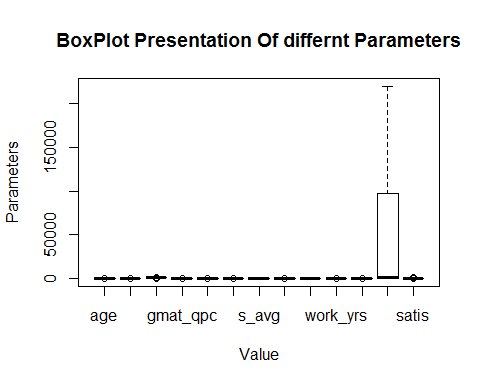
# Analysis of MBA Salaries  
# NAME: ANKIT KASHYAP  
# EMAIL: ankitkashyap.te15@rvce.edu.in  
# COLLEGE: R.V COLLEGE OF ENGINEERING   
####### MBA Starting Salaries #######  
##setting the directory and assigning a variabel to the data frame  
setwd("C:/Users/Ankit/Desktop/harvard")  
  
#Reading the dataset and creating a data frame  
mbasal.df<-read.csv(paste("MBA Starting Salaries Data.csv",sep = ""))  
  
#Viewing the data frame  
View(mbasal.df)  
  
##Analyzing the summary of the data and describing the variables  
library(psych)  
describe(mbasal.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

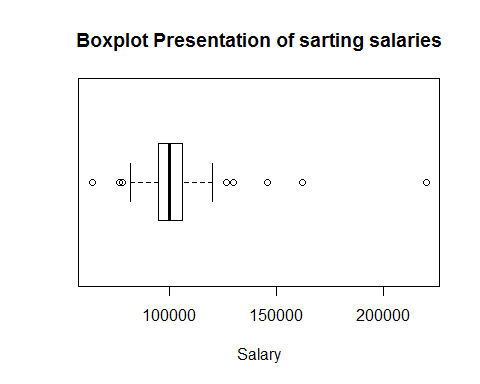
summary(mbasal.df)

## 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

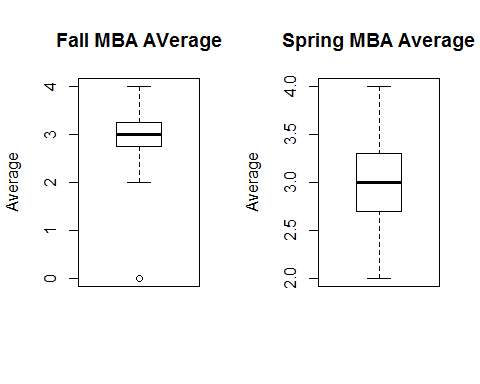
##Creating a new data frame for the students who had a starting salary  
startsalary.df<-mbasal.df$salary[mbasal.df$salary>999]  
  
##Drawing boxplots among comparable parameters  
  
boxplot(mbasal.df, xlab="Value", ylab="Parameters", main="BoxPlot Presentation Of differnt Parameters")



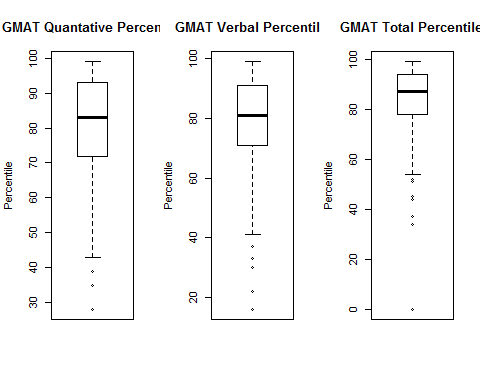
##Individual boxplots for comparable parameter  
boxplot(startsalary.df,horizontal = TRUE,xlab="Salary",main="Boxplot Presentation of sarting salaries")



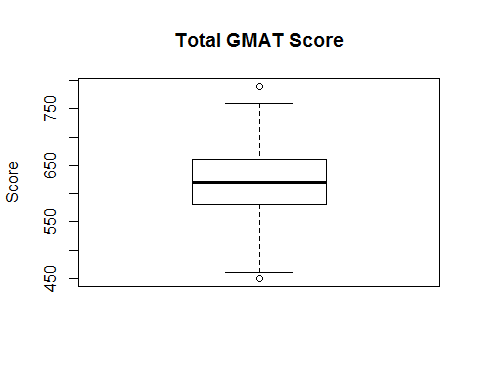
par(mfrow=c(1,2))  
with(mbasal.df, boxplot(mbasal.df$f\_avg,main="Fall MBA AVerage",ylab="Average"))  
with(mbasal.df, boxplot(mbasal.df$s\_avg,main="Spring MBA Average",ylab="Average"))



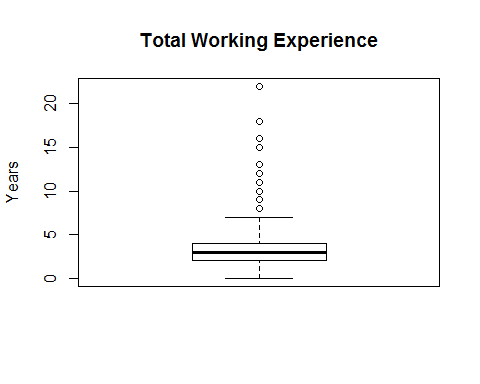
par(mfrow=c(1,1))  
  
par(mfrow=c(1,3))  
with(mbasal.df, boxplot(mbasal.df$gmat\_qpc,main="GMAT Quantative Percentile",ylab="Percentile"))  
with(mbasal.df, boxplot(mbasal.df$gmat\_vpc,main="GMAT Verbal Percentile",ylab="Percentile"))  
with(mbasal.df, boxplot(mbasal.df$gmat\_tpc,main="GMAT Total Percentile",ylab="Percentile"))



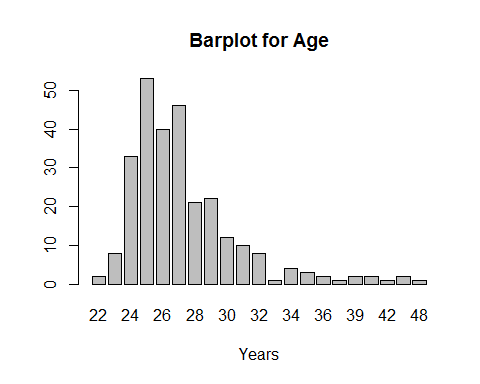
par(mfrow=c(1,1))  
  
boxplot(mbasal.df$gmat\_tot,main="Total GMAT Score",ylab="Score")



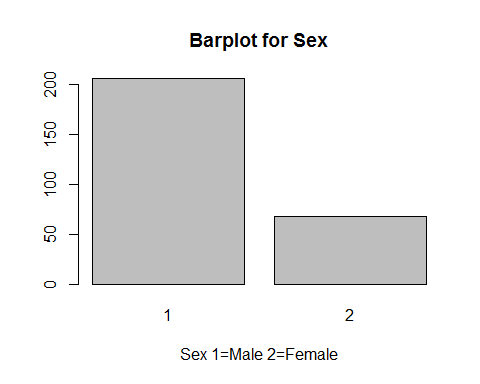
boxplot(mbasal.df$work\_yrs,main="Total Working Experience",ylab="Years")



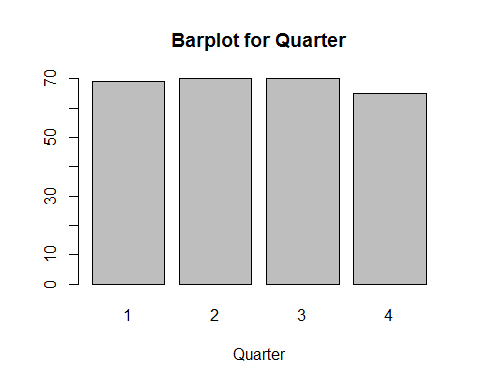
## Bar Plots to visualize the distribution of each variable independently  
  
count1<-table(mbasal.df$age)  
barplot(count1,main="Barplot for Age",xlab="Years")



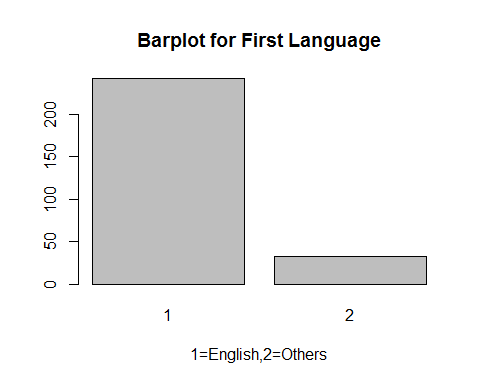
count2<-table(mbasal.df$sex)  
barplot(count2,main="Barplot for Sex",xlab="Sex 1=Male 2=Female")



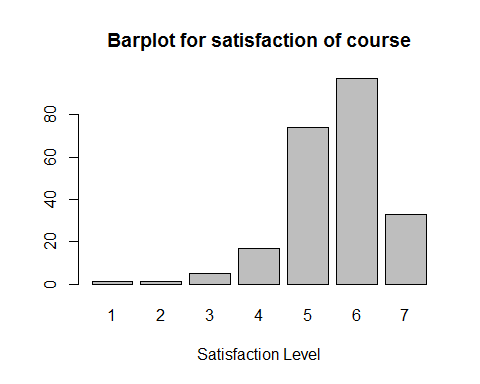
count3<-table(mbasal.df$quarter)  
barplot(count3,main="Barplot for Quarter",xlab="Quarter")



count4<-table(mbasal.df$frstlang)  
barplot(count4,main="Barplot for First Language",xlab="1=English,2=Others")



count5<-table(mbasal.df$satis[mbasal.df$satis<998])  
barplot(count5,main="Barplot for satisfaction of course",xlab="Satisfaction Level")

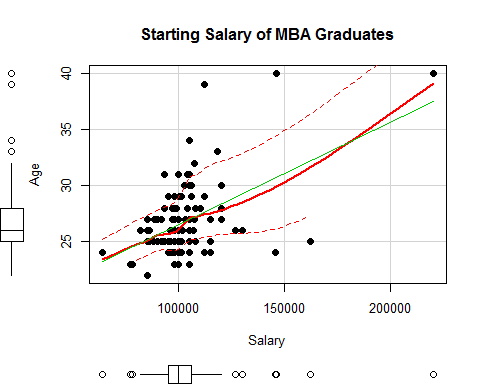


##Scatter Plots/ Plots to understand how are the variables correlated pair-wise  
  
salary1.df<-mbasal.df[which(mbasal.df$salary>999),]  
View(salary1.df)  
library(car)

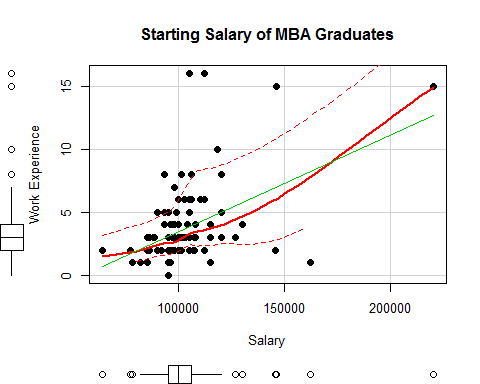
##   
## Attaching package: 'car'

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

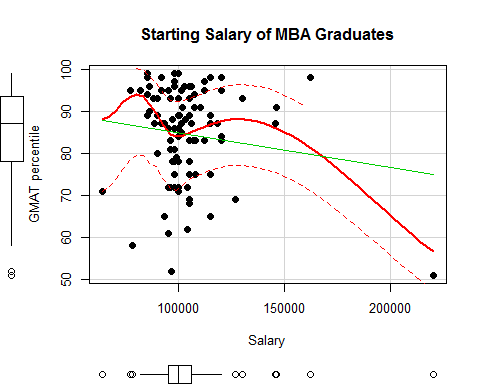
scatterplot(salary1.df$salary,salary1.df$age,main="Starting Salary of MBA Graduates",ylab = "Age", xlab="Salary",cex=1.1,pch=19)



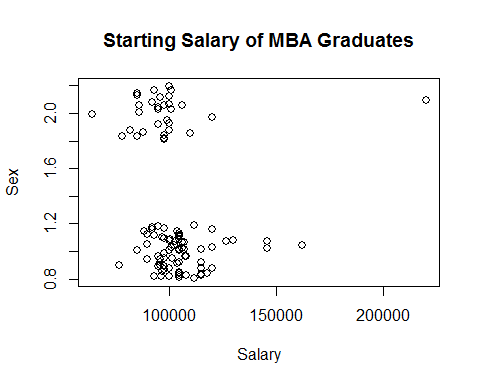
scatterplot(salary1.df$salary,salary1.df$work\_yrs,main="Starting Salary of MBA Graduates",ylab = "Work Experience", xlab="Salary",cex=1.1,pch=19)



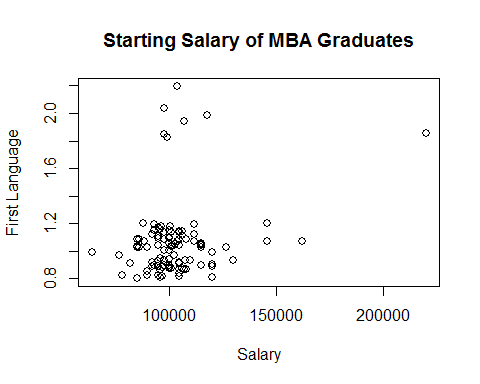
scatterplot(salary1.df$salary,salary1.df$gmat\_tpc,main="Starting Salary of MBA Graduates",ylab = "GMAT percentile", xlab="Salary",cex=1.1,pch=19)



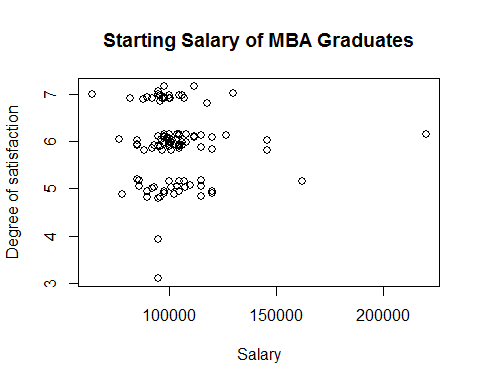
##Plots for binary categorical data with starting salaries  
plot(jitter(salary1.df$salary),jitter(salary1.df$sex),main="Starting Salary of MBA Graduates",ylab = "Sex", xlab="Salary",cex=1.1)



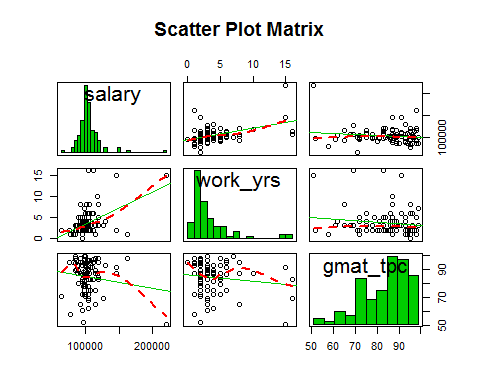
plot(jitter(salary1.df$salary),jitter(salary1.df$frstlang),main="Starting Salary of MBA Graduates",ylab = "First Language", xlab="Salary",cex=1.1)



plot(jitter(salary1.df$salary),jitter(salary1.df$satis),main="Starting Salary of MBA Graduates",ylab = "Degree of satisfaction", xlab="Salary",cex=1.1)



library(car)  
scatterplotMatrix(  
 salary1.df[  
 ,c("salary","work\_yrs","gmat\_tpc")],  
 spread=FALSE, smoother.args=list(lty=2),  
 main="Scatter Plot Matrix", diagonal = "histogram")



##Correlation tests to find relationship between different parameters  
# Correlation matrix,covariance matrix, Corrgram  
  
x<-mbasal.df[,c("age","sex","gmat\_tot","gmat\_qpc", "gmat\_vpc","gmat\_tpc","s\_avg", "f\_avg","quarter", "work\_yrs", "frstlang", "salary","satis")]  
y<-mbasal.df[,c("salary","gmat\_tpc","work\_yrs","satis","age")]  
cor(x,y)

## salary gmat\_tpc work\_yrs satis age  
## age -0.062573547 -0.169903066 0.858298096 -1.278882e-01 1.00000000  
## sex 0.068858628 -0.008090213 -0.011296374 -5.460222e-02 -0.02810644  
## gmat\_tot -0.054971880 0.847799647 -0.182354339 8.255770e-02 -0.14593840  
## gmat\_qpc -0.044032933 0.651377538 -0.236608270 6.060004e-02 -0.21616985  
## gmat\_vpc -0.006139340 0.666216035 -0.066390490 6.262375e-02 -0.04417547  
## gmat\_tpc 0.004930901 1.000000000 -0.173361859 9.293427e-02 -0.16990307  
## s\_avg 0.145836062 0.117362449 0.129292714 -3.268664e-02 0.14970402  
## f\_avg 0.029443027 0.079732099 -0.039056921 1.089273e-02 -0.01744806  
## quarter -0.164369865 -0.083035351 -0.086026406 -1.267198e-05 -0.04967221  
## work\_yrs 0.009023407 -0.173361859 1.000000000 -1.092553e-01 0.85829810  
## frstlang -0.086592096 -0.103362747 -0.027866747 7.932264e-02 0.05692649  
## salary 1.000000000 0.004930901 0.009023407 -3.352171e-01 -0.06257355  
## satis -0.335217114 0.092934266 -0.109255286 1.000000e+00 -0.12788825

cov(x,y)

## salary gmat\_tpc work\_yrs satis  
## age -1.183042e+04 -8.8399775 10.29493864 -1.763499e+02  
## sex 1.518264e+03 -0.0490896 -0.01580172 -8.780808e+00  
## gmat\_tot -1.611600e+05 683.9910698 -33.91633914 1.765263e+03  
## gmat\_qpc -3.335823e+04 135.7996845 -11.37186171 3.348371e+02  
## gmat\_vpc -5.273852e+03 157.4932488 -3.61816529 3.923563e+02  
## gmat\_tpc 3.522750e+03 196.6057057 -7.85751718 4.842467e+02  
## s\_avg 2.831601e+03 0.6271001 0.15926392 -4.628845e+00  
## f\_avg 7.876560e+02 0.5869862 -0.06628700 2.125329e+00  
## quarter -9.296214e+03 -1.2923719 -0.30866822 -5.227133e-03  
## work\_yrs 1.486147e+03 -7.8575172 10.44882490 -1.312408e+02  
## frstlang -1.419586e+03 -0.4663244 -0.02898318 9.484532e+00  
## salary 2.596062e+09 3522.7500067 1486.14704152 -6.347115e+06  
## satis -6.347115e+06 484.2466779 -131.24080907 1.380974e+05  
## age  
## age 1.376904e+01  
## sex -4.513248e-02  
## gmat\_tot -3.115879e+01  
## gmat\_qpc -1.192655e+01  
## gmat\_vpc -2.763643e+00  
## gmat\_tpc -8.839978e+00  
## s\_avg 2.116874e-01  
## f\_avg -3.399348e-02  
## quarter -2.045935e-01  
## work\_yrs 1.029494e+01  
## frstlang 6.796610e-02  
## salary -1.183042e+04  
## satis -1.763499e+02

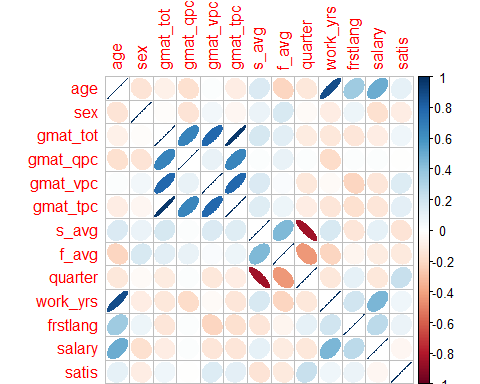
var(x,y)

## salary gmat\_tpc work\_yrs satis  
## age -1.183042e+04 -8.8399775 10.29493864 -1.763499e+02  
## sex 1.518264e+03 -0.0490896 -0.01580172 -8.780808e+00  
## gmat\_tot -1.611600e+05 683.9910698 -33.91633914 1.765263e+03  
## gmat\_qpc -3.335823e+04 135.7996845 -11.37186171 3.348371e+02  
## gmat\_vpc -5.273852e+03 157.4932488 -3.61816529 3.923563e+02  
## gmat\_tpc 3.522750e+03 196.6057057 -7.85751718 4.842467e+02  
## s\_avg 2.831601e+03 0.6271001 0.15926392 -4.628845e+00  
## f\_avg 7.876560e+02 0.5869862 -0.06628700 2.125329e+00  
## quarter -9.296214e+03 -1.2923719 -0.30866822 -5.227133e-03  
## work\_yrs 1.486147e+03 -7.8575172 10.44882490 -1.312408e+02  
## frstlang -1.419586e+03 -0.4663244 -0.02898318 9.484532e+00  
## salary 2.596062e+09 3522.7500067 1486.14704152 -6.347115e+06  
## satis -6.347115e+06 484.2466779 -131.24080907 1.380974e+05  
## age  
## age 1.376904e+01  
## sex -4.513248e-02  
## gmat\_tot -3.115879e+01  
## gmat\_qpc -1.192655e+01  
## gmat\_vpc -2.763643e+00  
## gmat\_tpc -8.839978e+00  
## s\_avg 2.116874e-01  
## f\_avg -3.399348e-02  
## quarter -2.045935e-01  
## work\_yrs 1.029494e+01  
## frstlang 6.796610e-02  
## salary -1.183042e+04  
## satis -1.763499e+02

#Visualizing relation through corrplots  
  
library(corrplot)

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

corrplot(corr=cor(salary1.df[,c(1:13)],use = "complete.obs"), method = "ellipse")

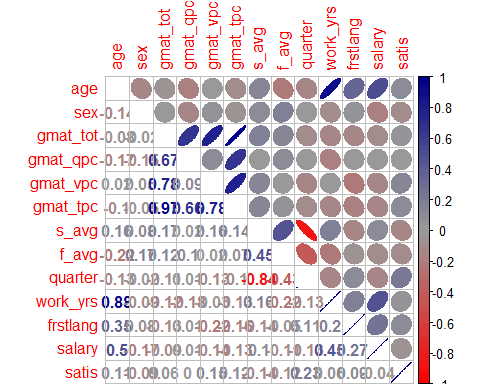


library(gplots)

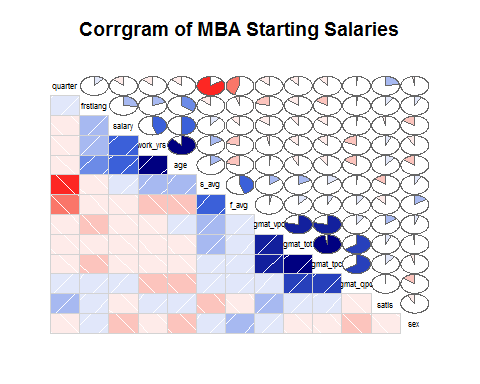
##   
## Attaching package: 'gplots'

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

corrplot.mixed(corr=cor(salary1.df[,c(1:13)],use = "complete.obs"), upper = "ellipse", tl.pos = "lt", col = colorpanel(50, "red", "gray60", "blue4"))



#VIsualizing by corrgram  
   
library(corrgram)  
  
corrgram(salary1.df, order=TRUE, lower.panel=panel.shade,  
 upper.panel=panel.pie, text.panel=panel.txt,  
 main="Corrgram of MBA Starting Salaries")



##Chi Square Test   
mytable<-xtabs(~sex+work\_yrs,data=salary1.df)  
chisq.test(mytable)

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

##   
## Pearson's Chi-squared test  
##   
## data: mytable  
## X-squared = 8.1579, df = 11, p-value = 0.6991

addmargins(mytable)

## work\_yrs  
## sex 0 1 2 3 4 5 6 7 8 10 15 16 Sum  
## 1 1 4 24 16 10 4 5 1 3 1 1 2 72  
## 2 0 4 14 5 1 3 2 0 1 0 1 0 31  
## Sum 1 8 38 21 11 7 7 1 4 1 2 2 103

##Because p value is more than 0.05 we cannot reject the null hpothesis and  
##the parameter sex and work experience are independant  
  
mytable1<-xtabs(~satis+work\_yrs,data=salary1.df)  
chisq.test(mytable1)

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

##   
## Pearson's Chi-squared test  
##   
## data: mytable1  
## X-squared = 131.13, df = 44, p-value = 1.35e-10

addmargins(mytable1)

## work\_yrs  
## satis 0 1 2 3 4 5 6 7 8 10 15 16 Sum  
## 3 0 0 0 1 0 0 0 0 0 0 0 0 1  
## 4 1 0 0 0 0 0 0 0 0 0 0 0 1  
## 5 0 5 8 6 3 3 2 1 0 0 0 1 29  
## 6 0 1 19 12 5 3 5 0 3 0 2 0 50  
## 7 0 2 11 2 3 1 0 0 1 1 0 1 22  
## Sum 1 8 38 21 11 7 7 1 4 1 2 2 103

##Since te p value is less than 0.05 we can't reject the null hypothesis and the  
##parameters of work experience and level of satisfaction are not independant  
  
mytable2<-xtabs(~sex+frstlang,data=salary1.df)  
chisq.test(mytable2)

## Warning in chisq.test(mytable2): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: mytable2  
## X-squared = 0.11264, df = 1, p-value = 0.7372

addmargins(mytable2)

## frstlang  
## sex 1 2 Sum  
## 1 68 4 72  
## 2 28 3 31  
## Sum 96 7 103

##Because p value is more than 0.05 we cannot reject the null hpothesis and  
##the parameter sex and first language are independant  
  
mytable3<-xtabs(~work\_yrs+frstlang,data=salary1.df)  
chisq.test(mytable3)

## Warning in chisq.test(mytable3): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: mytable3  
## X-squared = 22.274, df = 11, p-value = 0.02233

addmargins(mytable3)

## frstlang  
## work\_yrs 1 2 Sum  
## 0 1 0 1  
## 1 8 0 8  
## 2 36 2 38  
## 3 20 1 21  
## 4 10 1 11  
## 5 6 1 7  
## 6 7 0 7  
## 7 1 0 1  
## 8 4 0 4  
## 10 0 1 1  
## 15 1 1 2  
## 16 2 0 2  
## Sum 96 7 103

##Since te p value is less than 0.05 we can't reject the null hypothesis and the  
##parameters of work experience and level of satisfaction are not independant   
  
  
##Average Salary of male is greater than average salary of females.   
t.test(salary ~ sex,alternative= "greater",data=salary1.df)

##   
## Welch Two Sample t-test  
##   
## data: salary by sex  
## 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 more than 0.05 we can't reject the null hypothesis  
  
##Average salary of people whose first language is English is greater than average  
##salary of other language speakers  
   
t.test(salary ~ frstlang,alternative= "greater",data=salary1.df)

##   
## Welch Two Sample t-test  
##   
## data: salary by frstlang  
## t = -1.1202, df = 6.0863, p-value = 0.8476  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -51508.45 Inf  
## sample estimates:  
## mean in group 1 mean in group 2   
## 101748.6 120614.3

##Since the p value is more than 0.05 we can't reject the null hypothesis  
  
  
##Average GMAT percentile of male is greater than that of female  
t.test(gmat\_tpc ~ sex,alternative= "greater",data=salary1.df)

##   
## Welch Two Sample t-test  
##   
## data: gmat\_tpc by sex  
## t = 0.43873, df = 48.83, p-value = 0.3314  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -3.157889 Inf  
## sample estimates:  
## mean in group 1 mean in group 2   
## 84.86111 83.74194

##Since the p value is more than 0.05 we can't reject the null hypothesis  
  
  
##Generating A Multi Variable Linear Regressional Model for MBA Starting Salary  
##1.  
  
linear1.mod<- lm(salary~ work\_yrs + gmat\_tot -1, data = salary1.df)  
summary(linear1.mod)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + gmat\_tot - 1, data = salary1.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30428 -9691 -624 8110 97678   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## work\_yrs 3264.289 579.553 5.632 1.61e-07 \*\*\*  
## gmat\_tot 146.716 4.449 32.976 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17920 on 101 degrees of freedom  
## Multiple R-squared: 0.9712, Adjusted R-squared: 0.9706   
## F-statistic: 1702 on 2 and 101 DF, p-value: < 2.2e-16

#Coefficients of the model  
coefficients(linear1.mod)

## work\_yrs gmat\_tot   
## 3264.2887 146.7158

#Residuals of the model  
residuals(linear1.mod)

## 35 36 37 38 39 40   
## -15096.6876 -24229.6066 -20295.2915 -17158.1075 -18696.7642 -14285.2123   
## 41 42 43 44 45 46   
## -3299.5565 9574.8879 7447.7832 -2629.5301 -3619.4509 -1893.8188   
## 47 48 49 50 51 52   
## 11310.5992 -9082.3435 3106.1812 14513.4681 11579.1529 13376.2841   
## 53 54 55 56 57 58   
## 649.1028 1786.2868 -11078.0785 -13352.4464 5276.2076 10177.6803   
## 59 60 61 62 63 64   
## -3483.8162 -28258.0732 18314.8642 7567.5246 -5606.6559 22177.6803   
## 65 66 67 68 69 115   
## 977.5272 15649.1028 -4413.8663 4604.7429 56034.6821 -19563.8452   
## 116 117 118 119 120 121   
## -5492.3461 -12340.8180 -9828.1339 -4289.4773 7310.5992 16613.5446   
## 122 123 124 125 126 127   
## 4909.1265 7843.4417 4579.1529 -2416.5821 8376.2841 4782.0218   
## 128 129 130 131 132 133   
## 3507.6539 2243.3651 8782.0218 -4559.5802 4243.3651 -27922.2854   
## 134 135 136 137 138 139   
## 12441.9690 -5885.2888 20771.9426 9511.9189 21577.6037 48307.6539   
## 186 187 188 189 190 191   
## -1300.4818 -12256.6349 -1623.7159 -8959.5037 -9553.7660 -3482.2670   
## 192 193 194 195 196 197   
## -7020.9236 2111.9954 -6828.1339 9310.5992 6376.2841 -8295.2915   
## 198 199 200 201 202 203   
## -4553.7660 -29018.2078 -13961.0529 1040.4963 6441.9690 -21846.7087   
## 204 205 206 207 208 209   
## -15385.2888 13376.2841 -13352.4464 3939.0236 11314.8642 5704.7085   
## 256 257 258 259 260 261   
## -24689.4008 -26360.9764 -30428.2105 -12492.3461 -16223.7924 -21686.6850   
## 262 263 264 265 266 267   
## -11360.9764 -1025.1886 1441.9690 7843.4417 5441.9690 -1893.8188   
## 268 269 270 271 272 273   
## -1823.7924 -623.7924 13183.4944 13376.2841 25980.6256 36223.4681   
## 274   
## 97677.7912

#Fitting the model  
fitted(linear1.mod)

## 35 36 37 38 39 40 41   
## 100096.69 109229.61 106295.29 105158.11 110696.76 107285.21 98299.56   
## 42 43 44 45 46 47 48   
## 85425.11 87552.22 98629.53 99619.45 101893.82 88689.40 109082.34   
## 49 50 51 52 53 54 55   
## 101893.82 90486.53 93420.85 91623.72 104350.90 103213.71 117078.08   
## 56 57 58 59 60 61 62   
## 119352.45 102223.79 97822.32 113483.82 140258.07 96685.14 107432.48   
## 63 64 65 66 67 68 69   
## 123606.66 97822.32 119022.47 104350.90 124413.87 141395.26 105965.32   
## 115 116 117 118 119 120 121   
## 101563.85 97492.35 105340.82 104828.13 99289.48 88689.40 79886.46   
## 122 123 124 125 126 127 128   
## 93090.87 90156.56 93420.85 101416.58 91623.72 95217.98 97492.35   
## 129 130 131 132 133 134 135   
## 100756.63 95217.98 109559.58 100756.63 132922.29 94558.03 117885.29   
## 136 137 138 139 186 187 188   
## 94228.06 105488.08 108422.40 97492.35 79556.48 100756.63 91623.72   
## 189 190 191 192 193 194 195   
## 98959.50 102553.77 98482.27 104020.92 94888.00 104828.13 88689.40   
## 196 197 198 199 200 201 202   
## 91623.72 106295.29 102553.77 127018.21 113961.05 98959.50 94558.03   
## 203 204 205 206 207 208 209   
## 122946.71 117885.29 91623.72 119352.45 103360.98 96685.14 106295.29   
## 256 257 258 259 260 261 262   
## 88689.40 103360.98 115428.21 97492.35 102223.79 111686.68 103360.98   
## 263 264 265 266 267 268 269   
## 96025.19 94558.03 90156.56 94558.03 101893.82 102223.79 102223.79   
## 270 271 272 273 274   
## 90816.51 91623.72 89019.37 90486.53 122322.21

## Model: salary = b0 + b1\*work\_yrs + b2\*gmat\_tot  
## b0 = -1(assumption), b1 = 3264.2887,b2= 146.7158  
## Model: salary = -1 + 3264.2887\*work\_yrs+146.7158\*gmat\_tot  
  
  
##2.  
linear2.mod<-lm(salary~work\_yrs+age\*frstlang+gmat\_tot+sex-1,data= salary1.df)  
summary(linear2.mod)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + age \* frstlang + gmat\_tot +   
## sex - 1, data = salary1.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28406 -9496 -820 6174 69521   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## work\_yrs -958.87 1091.08 -0.879 0.381667   
## age 2905.80 830.35 3.499 0.000706 \*\*\*  
## frstlang -15290.15 25367.28 -0.603 0.548081   
## gmat\_tot 40.36 31.33 1.288 0.200705   
## sex -2260.92 3407.03 -0.664 0.508518   
## age:frstlang 794.41 797.85 0.996 0.321878   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15790 on 97 degrees of freedom  
## Multiple R-squared: 0.9785, Adjusted R-squared: 0.9772   
## F-statistic: 736.6 on 6 and 97 DF, p-value: < 2.2e-16

#Coefficients of the model  
linear2.mod$coefficients

## work\_yrs age frstlang gmat\_tot sex   
## -958.86681 2905.80137 -15290.15225 40.35853 -2260.92128   
## age:frstlang   
## 794.41173

#Residuals of the model  
residuals(linear2.mod)

## 35 36 37 38 39 40   
## -2270.4554 -21426.9952 -12219.3985 -8049.7758 -17091.5018 -11021.9256   
## 41 42 43 44 45 46   
## -3294.2550 664.7582 3793.2476 -4228.4307 -369.6104 10391.7836   
## 47 48 49 50 51 52   
## 8062.9168 -1220.1817 5730.4361 10725.1558 6217.7721 4855.3200   
## 53 54 55 56 57 58   
## -7876.1025 -9406.6929 -4806.6727 -6884.9967 2596.0473 11707.2293   
## 59 60 61 62 63 64   
## 6129.7706 -23630.4855 13476.4258 4949.6314 -12915.0130 16306.8032   
## 65 66 67 68 69 115   
## 10488.4708 7123.8975 9436.3222 4499.6787 59753.6429 -20474.8931   
## 116 117 118 119 120 121   
## -3797.8868 -18278.2035 -5076.7345 -1696.3559 4062.9168 6984.4285   
## 122 123 124 125 126 127   
## 8552.3740 5659.3315 -3885.8592 -10360.9605 9516.6675 -15920.8413   
## 128 129 130 131 132 133   
## 5202.1132 -1500.3674 -820.2020 1778.3340 499.6326 -10111.4936   
## 134 135 136 137 138 139   
## 6048.1494 957.8804 19682.5383 11054.9141 28947.9566 51441.4051   
## 186 187 188 189 190 191   
## -7064.4745 -16000.3674 -6444.4669 -8462.3933 -2737.4891 -10300.6270   
## 192 193 194 195 196 197   
## -6541.5006 -9586.2394 3884.3999 4623.6250 3816.4544 -16039.7103   
## 198 199 200 201 202 203   
## -8838.1284 -13697.5940 2968.0043 3798.5280 9709.4969 -7620.8007   
## 204 205 206 207 208 209   
## -12242.3327 12255.7462 -5445.7049 -28405.6638 2776.2127 15219.8933   
## 256 257 258 259 260 261   
## -25676.1619 -15272.7230 -13874.8728 -16759.0211 -12942.8183 -13793.3897   
## 262 263 264 265 266 267   
## -7673.1492 -2655.2227 -4951.8506 5659.3315 8709.4969 -708.8557   
## 268 269 270 271 272 273   
## -11904.3789 396.2604 -20046.6967 15955.9593 21128.7411 28734.9427   
## 274   
## 69520.8920

#Fitting the model  
fitted(linear2.mod)

## 35 36 37 38 39 40 41   
## 87270.46 106427.00 98219.40 96049.78 109091.50 104021.93 98294.25   
## 42 43 44 45 46 47 48   
## 94335.24 91206.75 100228.43 96369.61 89608.22 91937.08 101220.18   
## 49 50 51 52 53 54 55   
## 99269.56 94274.84 98782.23 100144.68 112876.10 114406.69 110806.67   
## 56 57 58 59 60 61 62   
## 112885.00 104903.95 96292.77 103870.23 135630.49 101523.57 110050.37   
## 63 64 65 66 67 68 69   
## 130915.01 103693.20 109511.53 112876.10 110563.68 141500.32 102246.36   
## 115 116 117 118 119 120 121   
## 102474.89 95797.89 111278.20 100076.73 96696.36 91937.08 89515.57   
## 122 123 124 125 126 127 128   
## 89447.63 92340.67 101885.86 109360.96 90483.33 115920.84 95797.89   
## 129 130 131 132 133 134 135   
## 104500.37 104820.20 103221.67 104500.37 115111.49 100951.85 111042.12   
## 136 137 138 139 186 187 188   
## 95317.46 103945.09 101052.04 94358.59 85320.47 104500.37 96444.47   
## 189 190 191 192 193 194 195   
## 98462.39 95737.49 105300.63 103541.50 106586.24 94115.60 93376.38   
## 196 197 198 199 200 201 202   
## 94183.55 114039.71 106838.13 111697.59 97032.00 96201.47 91290.50   
## 203 204 205 206 207 208 209   
## 108720.80 114742.33 92744.25 111445.70 135705.66 105223.79 96780.11   
## 256 257 258 259 260 261 262   
## 89676.16 92272.72 98874.87 101759.02 98942.82 103793.39 99673.15   
## 263 264 265 266 267 268 269   
## 97655.22 100951.85 92340.67 91290.50 100708.86 112304.38 101203.74   
## 270 271 272 273 274   
## 124046.70 89044.04 93871.26 97975.06 150479.11

## Model: salary = b0 + b1\*work\_yrs + b2\*age +b3\*frstlang +b4\*gmat\_tot +b5\*sex +b6\*sex\*frstlang  
## b0 = -1(assumption), b1 = -958.86681,b2=2905.80137,b3=-15290.15225,b4=40/35853,b5=-2260.92128,b6=794.41173  
## Model: salary = -1 + -958.86681\*work\_yrs + 2905.80137\*age +-1290.15225\*frstlang +40.35853\*gmat\_tot + -2260.92128\*sex + 794.41173\*sex\*frstlang  
  
  
  
##3.  
linear3.mod<-lm(salary~work\_yrs+age,data=salary1.df)  
summary(linear3.mod)

##   
## Call:  
## lm(formula = salary ~ work\_yrs + age, data = salary1.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -31675 -8099 -2108 4411 80650   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36967.5 23323.8 1.585 0.1161   
## work\_yrs 388.8 1084.0 0.359 0.7206   
## age 2413.8 997.4 2.420 0.0173 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15620 on 100 degrees of freedom  
## Multiple R-squared: 0.2506, Adjusted R-squared: 0.2356   
## F-statistic: 16.72 on 2 and 100 DF, p-value: 5.438e-07

#Coefficients of the model  
linear3.mod$coefficients

## (Intercept) work\_yrs age   
## 36967.4546 388.8347 2413.7599

#Residuals of the model  
residuals(linear3.mod)

## 35 36 37 38 39   
## -5459.00732 -17916.64155 -12089.12173 -10477.95640 -10916.64155   
## 40 41 42 43 44   
## -13496.90548 102.30753 -2700.28706 -3477.95640 -1700.28706   
## 45 46 47 48 49   
## -5280.55098 6738.39810 4324.63818 -4471.98024 6910.87827   
## 50 51 52 53 54   
## 6522.04360 4108.28369 4497.11836 -6324.42530 -9127.01989   
## 55 56 57 58 59   
## -6490.92932 -5713.25997 4194.52378 9522.04360 3114.25985   
## 60 61 62 63 64   
## -25325.44590 11305.68911 12472.19312 -2509.87840 16694.52378   
## 65 66 67 68 69   
## 13503.09452 8675.57470 7509.07068 6649.62886 64299.71294   
## 115 116 117 118 119   
## -18114.04697 -6089.12173 -21904.68923 -3089.12173 -3477.95640   
## 120 121 122 123 124   
## 324.63818 824.63818 4738.39810 2324.63818 -2891.71631   
## 125 126 127 128 129   
## -7496.90548 4324.63818 -13349.35054 2910.87827 -305.47622   
## 130 131 132 133 134   
## -2108.07081 4108.28369 1694.52378 -20256.64634 6497.11836   
## 135 136 137 138 139   
## 2700.49994 19713.47286 11305.68911 28719.44902 50124.63818   
## 186 187 188 189 190   
## -14616.76723 -14805.47622 -8089.12173 -8089.12173 -5866.79107   
## 191 192 193 194 195   
## -13910.66539 -6694.31089 -8719.23613 2324.63818 -89.12173   
## 196 197 198 199 200   
## -89.12173 -4916.64155 -8108.07081 -11688.33473 3935.80351   
## 201 202 203 204 205   
## 1910.87827 5324.63818 -8977.16941 -9213.25997 9324.63818   
## 206 207 208 209 256   
## -3299.50006 -7685.44111 1891.92919 16324.63818 -31675.36182   
## 257 258 259 260 261   
## -16261.60190 -13477.95640 -15502.88164 -14891.71631 -14083.14557   
## 262 263 264 265 266   
## -6089.12173 -3089.12173 -4502.88164 2324.63818 4324.63818   
## 267 268 269 270 271   
## -502.88164 -7732.99605 708.28369 -9349.35054 11738.39810   
## 272 273 274   
## 16522.04360 25818.28369 80649.62886

#Fitting the model  
fitted(linear3.mod)

## 35 36 37 38 39 40 41   
## 90459.01 102916.64 98089.12 98477.96 102916.64 106496.91 94897.69   
## 42 43 44 45 46 47 48   
## 97700.29 98477.96 97700.29 101280.55 93261.60 95675.36 104471.98   
## 49 50 51 52 53 54 55   
## 98089.12 98477.96 100891.72 100502.88 111324.43 114127.02 112490.93   
## 56 57 58 59 60 61 62   
## 111713.26 103305.48 98477.96 106885.74 137325.45 103694.31 102527.81   
## 63 64 65 66 67 68 69   
## 120509.88 103305.48 106496.91 111324.43 112490.93 139350.37 97700.29   
## 115 116 117 118 119 120 121   
## 100114.05 98089.12 114904.69 98089.12 98477.96 95675.36 95675.36   
## 122 123 124 125 126 127 128   
## 93261.60 95675.36 100891.72 106496.91 95675.36 113349.35 98089.12   
## 129 130 131 132 133 134 135   
## 103305.48 106108.07 100891.72 103305.48 125256.65 100502.88 109299.50   
## 136 137 138 139 186 187 188   
## 95286.53 103694.31 101280.55 95675.36 92872.77 103305.48 98089.12   
## 189 190 191 192 193 194 195   
## 98089.12 98866.79 108910.67 103694.31 105719.24 95675.36 98089.12   
## 196 197 198 199 200 201 202   
## 98089.12 102916.64 106108.07 109688.33 96064.20 98089.12 95675.36   
## 203 204 205 206 207 208 209   
## 110077.17 111713.26 95675.36 109299.50 114985.44 106108.07 95675.36   
## 256 257 258 259 260 261 262   
## 95675.36 93261.60 98477.96 100502.88 100891.72 104083.15 98089.12   
## 263 264 265 266 267 268 269   
## 98089.12 100502.88 95675.36 95675.36 100502.88 108133.00 100891.72   
## 270 271 272 273 274   
## 113349.35 93261.60 98477.96 100891.72 139350.37

## Model: salary = b0 + b1\*work\_yrs + b2\*age  
## b0 = 36967.4546, b1 = 388.8347,b2= 2413.7599  
## Model: salary = 36967.4546 + 388.8347\*work\_yrs+2413.7599\*age  
  
##Creating a subset for those who did not get a job  
salary2.df<-mbasal.df[which(mbasal.df$salary<998),]  
  
##Chi square test  
mytable4<-xtabs(~sex+work\_yrs,data=salary2.df)  
chisq.test(mytable4)

## Warning in chisq.test(mytable4): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: mytable4  
## X-squared = 21.229, df = 16, p-value = 0.1699

addmargins(mytable4)

## work\_yrs  
## sex 0 1 2 3 4 5 6 7 8 9 10 11 12 13 16 18 22 Sum  
## 1 1 12 16 9 8 7 2 3 2 0 0 1 2 0 1 1 2 67  
## 2 0 0 6 5 1 5 0 2 0 1 1 1 0 1 0 0 0 23  
## Sum 1 12 22 14 9 12 2 5 2 1 1 2 2 1 1 1 2 90

##Since the p value is more than 0.05 we can't reject the null hypothesis and  
##the parameter sex and work experience are independant  
  
mytable5<-xtabs(~sex+frstlang,data=salary2.df)  
chisq.test(mytable5)

## Warning in chisq.test(mytable5): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: mytable5  
## X-squared = 0.21376, df = 1, p-value = 0.6438

addmargins(mytable5)

## frstlang  
## sex 1 2 Sum  
## 1 60 7 67  
## 2 22 1 23  
## Sum 82 8 90

##Since the p value is more than 0.05 we can't reject the null hypothesis and  
##the parameter sex and first language are independant  
  
  
  
##################################CHALLENGE#########################  
#Logistic Regression Analysis for students who got a job  
salary1.df$sex<-factor(salary1.df$sex)  
is.factor(salary1.df$sex)

## [1] TRUE

logic1.mod<-glm(sex~.,family = binomial(link = 'logit'),data=salary1.df)  
summary(logic1.mod)

##   
## Call:  
## glm(formula = sex ~ ., family = binomial(link = "logit"), data = salary1.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4863 -0.7894 -0.5805 0.7626 2.3292   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.1064455 8.2384884 0.013 0.9897   
## age -0.3643742 0.1889782 -1.928 0.0538 .  
## gmat\_tot 0.0162521 0.0269925 0.602 0.5471   
## gmat\_qpc -0.0435054 0.0770321 -0.565 0.5722   
## gmat\_vpc 0.0084836 0.0780797 0.109 0.9135   
## gmat\_tpc -0.0561304 0.1181993 -0.475 0.6349   
## s\_avg 0.1751868 1.5508906 0.113 0.9101   
## f\_avg 1.5943945 1.0429927 1.529 0.1263   
## quarter 0.2901630 0.4253040 0.682 0.4951   
## work\_yrs 0.2410914 0.1783851 1.352 0.1765   
## frstlang 2.4111026 1.0665299 2.261 0.0238 \*  
## salary -0.0000184 0.0000191 -0.963 0.3353   
## satis -0.2638553 0.3332759 -0.792 0.4285   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 126.01 on 102 degrees of freedom  
## Residual deviance: 107.49 on 90 degrees of freedom  
## AIC: 133.49  
##   
## Number of Fisher Scoring iterations: 5

anova(logic1.mod,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: sex  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 102 126.01   
## age 1 2.3856 101 123.62 0.12245   
## gmat\_tot 1 0.0744 100 123.55 0.78507   
## gmat\_qpc 1 4.1847 99 119.36 0.04079 \*  
## gmat\_vpc 1 1.8543 98 117.51 0.17329   
## gmat\_tpc 1 0.0823 97 117.43 0.77423   
## s\_avg 1 0.4155 96 117.01 0.51919   
## f\_avg 1 2.1057 95 114.90 0.14675   
## quarter 1 0.4742 94 114.43 0.49107   
## work\_yrs 1 0.5956 93 113.83 0.44026   
## frstlang 1 4.6687 92 109.17 0.03072 \*  
## salary 1 1.0389 91 108.13 0.30808   
## satis 1 0.6359 90 107.49 0.42521   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

fitted.results<-predict(logic1.mod,data=salary1.df,type = 'response')  
fitted.results<-ifelse(fitted.results>0.5,1,0)  
misClassificError<-mean(fitted.results != salary1.df$sex)  
print(paste('Accuracy',1-misClassificError))

## [1] "Accuracy 0.0485436893203883"

#Logistic Regression Analysis for students who did not got a job  
salary2.df$sex<-factor(salary2.df$sex)  
is.factor(salary2.df$sex)

## [1] TRUE

logic2.mod<-glm(sex~.,family = binomial(link = 'logit'),data=salary2.df)  
summary(logic2.mod)

##   
## Call:  
## glm(formula = sex ~ ., family = binomial(link = "logit"), data = salary2.df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5451 -0.7582 -0.4838 0.6019 2.1976   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 13.39699 7.84772 1.707 0.0878 .  
## age 0.05353 0.12071 0.443 0.6574   
## gmat\_tot -0.03439 0.02183 -1.576 0.1151   
## gmat\_qpc 0.02944 0.06260 0.470 0.6381   
## gmat\_vpc 0.10328 0.06711 1.539 0.1238   
## gmat\_tpc 0.03205 0.06128 0.523 0.6010   
## s\_avg -0.47864 1.17187 -0.408 0.6830   
## f\_avg -0.58170 0.57645 -1.009 0.3129   
## quarter -0.49321 0.36673 -1.345 0.1787   
## work\_yrs -0.08643 0.14181 -0.609 0.5422   
## frstlang -0.31776 1.29059 -0.246 0.8055   
## salary NA NA NA NA   
## satis -0.51118 0.40913 -1.249 0.2115   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 102.304 on 89 degrees of freedom  
## Residual deviance: 86.742 on 78 degrees of freedom  
## AIC: 110.74  
##   
## Number of Fisher Scoring iterations: 5

anova(logic2.mod,test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: sex  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 89 102.304   
## age 1 0.4712 88 101.833 0.49244   
## gmat\_tot 1 0.3130 87 101.520 0.57585   
## gmat\_qpc 1 4.3705 86 97.150 0.03657 \*  
## gmat\_vpc 1 5.0395 85 92.110 0.02478 \*  
## gmat\_tpc 1 0.6560 84 91.454 0.41798   
## s\_avg 1 0.0490 83 91.405 0.82487   
## f\_avg 1 0.5497 82 90.855 0.45844   
## quarter 1 1.6354 81 89.220 0.20096   
## work\_yrs 1 0.8609 80 88.359 0.35348   
## frstlang 1 0.0078 79 88.351 0.92960   
## salary 0 0.0000 79 88.351   
## satis 1 1.6093 78 86.742 0.20459   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

fitted.results<-predict(logic2.mod,data=salary2.df,type = 'response')  
fitted.results<-ifelse(fitted.results>0.5,1,0)  
misClassificError<-mean(fitted.results != salary2.df$sex)  
print(paste('Accuracy',1-misClassificError))

## [1] "Accuracy 0.0444444444444444"