Sai\_Gopi\_Krishna\_Govindarajula\_MBASalaries\_Code.R

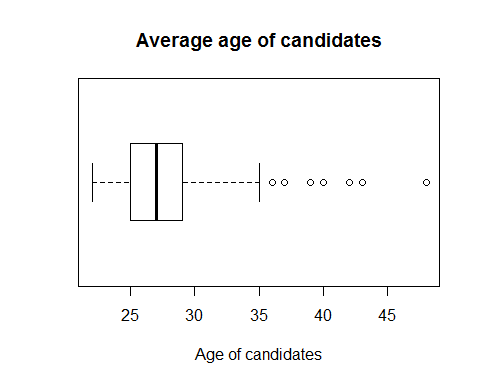
HansZimmer

Thu Jul 06 08:54:10 2017

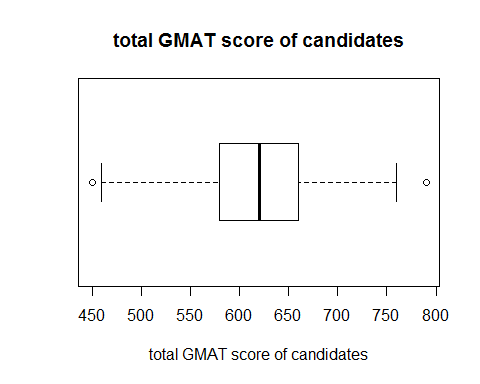
# Analysis of MBA SALARIES  
# NAME: Sai Gopi Krishna Govindarajula  
# EMAIL: saigopikrishna.g@gmail.com  
# COLLEGE / COMPANY: BITS PILANI/Capgemini   
  
#To read the data  
MBA.df <- read.csv(paste("E:/Studies/Sai Gopi Krishna Govindarajula/Udemy/Data sets/MBA Starting Salaries Data.csv", sep=""))  
View(MBA.df)  
  
#Summary  
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

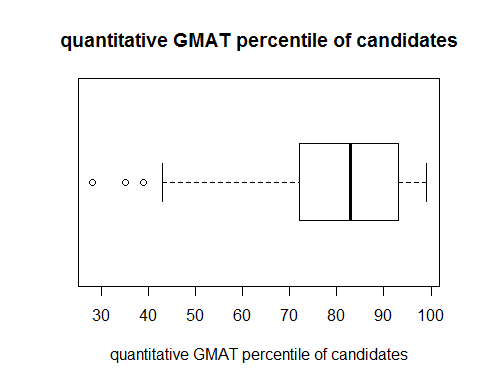
#Boxplot  
#Boxplot to see Average age of candidates  
boxplot(MBA.df$age,horizontal=TRUE,xlab="Age of candidates",main ="Average age of candidates")



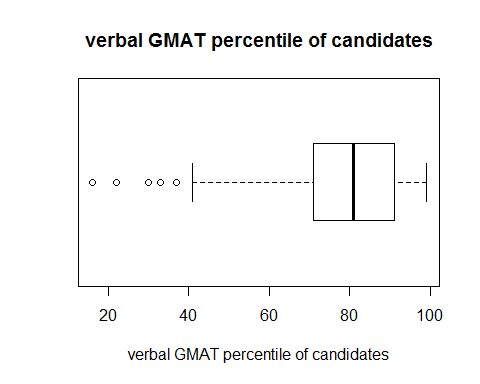
# Boxplot to see total GMAT score of candidates  
boxplot(MBA.df$gmat\_tot,horizontal=TRUE,xlab="total GMAT score of candidates",main ="total GMAT score of candidates")



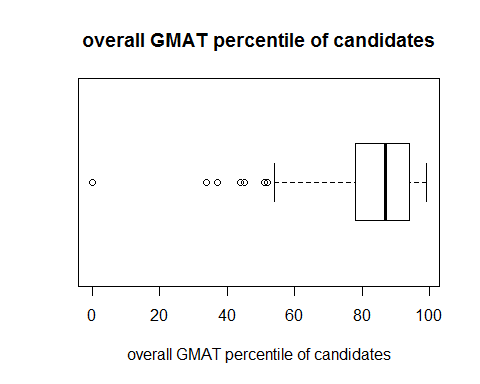
#Boxplot to see quantitative GMAT percentile of candidates  
boxplot(MBA.df$gmat\_qpc,horizontal=TRUE,xlab="quantitative GMAT percentile of candidates",main ="quantitative GMAT percentile of candidates")



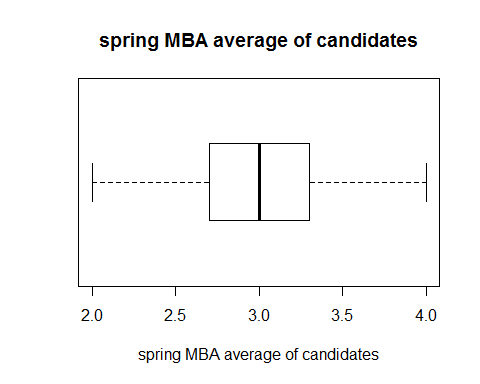
#Boxplot to see verbal GMAT percentile of candidates  
boxplot(MBA.df$gmat\_vpc,horizontal=TRUE,xlab="verbal GMAT percentile of candidates",main ="verbal GMAT percentile of candidates")



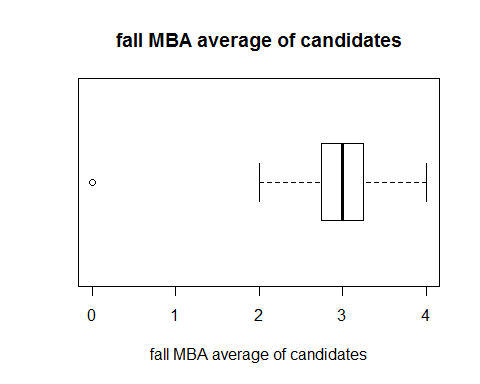
# Boxplot to see overall GMAT percentile  
boxplot(MBA.df$gmat\_tpc,horizontal=TRUE,xlab="overall GMAT percentile of candidates",main ="overall GMAT percentile of candidates")



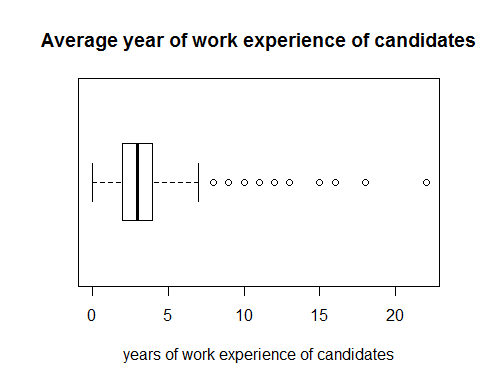
# Boxplot to see spring MBA average pf candidates  
boxplot(MBA.df$s\_avg,horizontal=TRUE,xlab="spring MBA average of candidates",main ="spring MBA average of candidates")



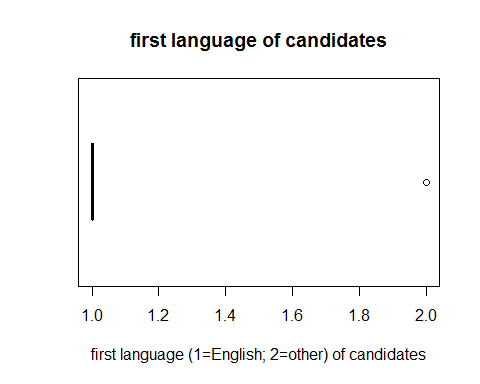
#Boxplot to see fall MBA average of candidates  
boxplot(MBA.df$f\_avg,horizontal=TRUE,xlab="fall MBA average of candidates",main ="fall MBA average of candidates")



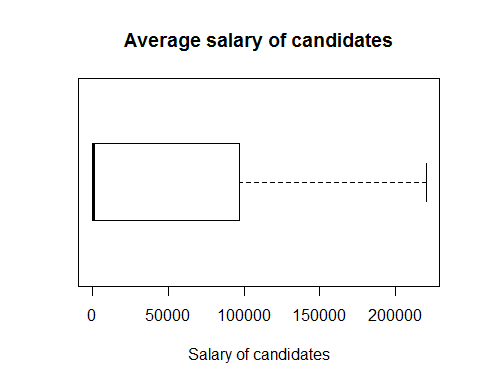
#Boxplot to see years of work experience  
boxplot(MBA.df$work\_yrs,horizontal=TRUE,xlab="years of work experience of candidates",main ="Average year of work experience of candidates")



# Boxplot to see first language (1=English; 2=other)   
boxplot(MBA.df$frstlang,horizontal=TRUE,xlab="first language (1=English; 2=other) of candidates",main ="first language of candidates")

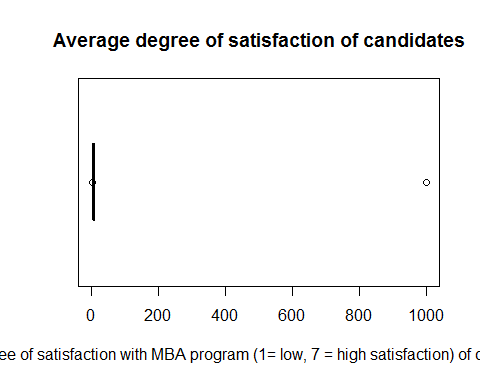


#Boxplot to see Average salary  
boxplot(MBA.df$salary,horizontal=TRUE,xlab="Salary of candidates",main ="Average salary of candidates")

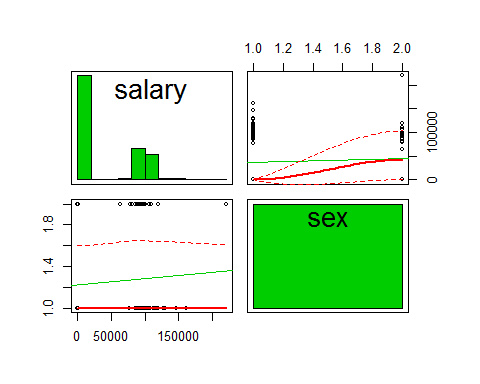


# Boxplot to see degree of satisfaction with MBA program (1= low, 7 = high satisfaction)  
boxplot(MBA.df$satis,horizontal=TRUE,xlab="degree of satisfaction with MBA program (1= low, 7 = high satisfaction) of candidates",main ="Average degree of satisfaction of candidates")  
  
#Scatterplot matrix  
library(car)

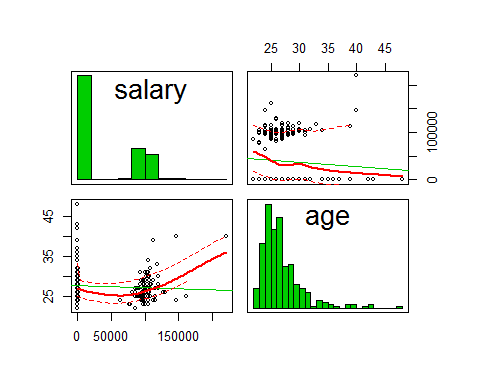
## Warning: package 'car' was built under R version 3.3.3



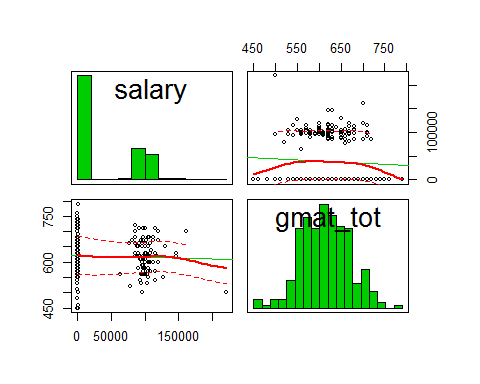
#scatterplot matrix for salary and sex  
scatterplotMatrix(formula=~salary+sex,cex=0.6,data=MBA.df,diagonal="histogram")



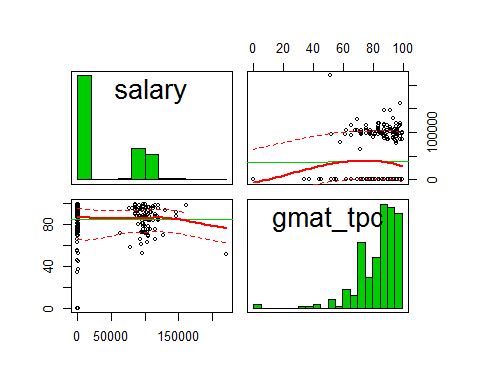
#Scatterplot matrix for salary and age  
scatterplotMatrix(formula=~salary+age,cex=0.6,data=MBA.df,diagonal="histogram")



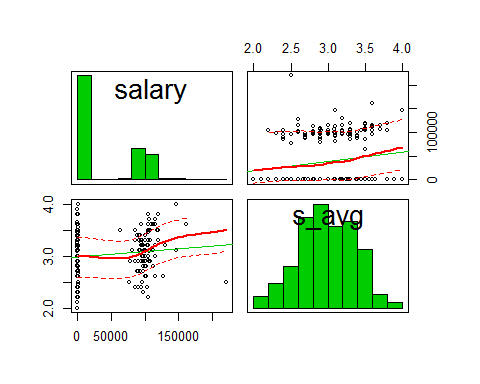
#Scatterplot matrix for salary and gmat total  
scatterplotMatrix(formula=~salary+gmat\_tot,cex=0.6,data=MBA.df,diagonal="histogram")



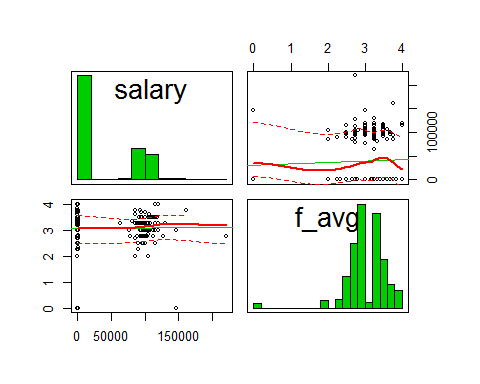
#Scatterplot matrix for salary and gmat overall  
scatterplotMatrix(formula=~salary+gmat\_tpc,cex=0.6,data=MBA.df,diagonal="histogram")



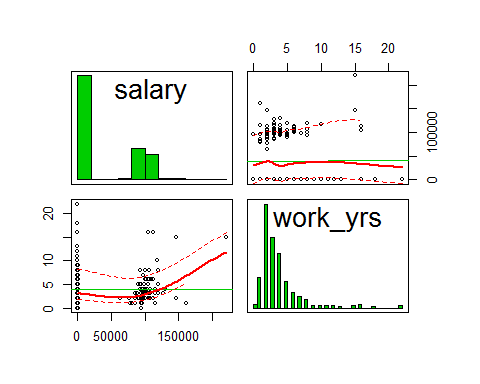
#Scatterplot matrix for salary and spring avg  
scatterplotMatrix(formula=~salary+s\_avg,cex=0.6,data=MBA.df,diagonal="histogram")



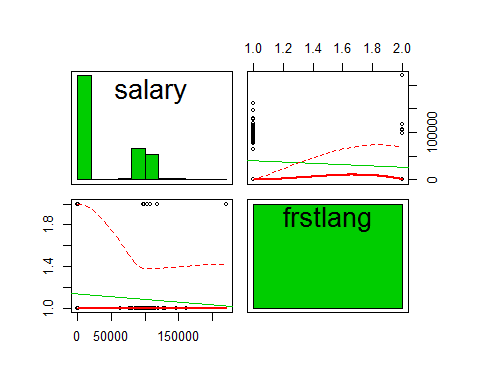
#Scatterplot matrix for salary and fall avg  
scatterplotMatrix(formula=~salary+f\_avg,cex=0.6,data=MBA.df,diagonal="histogram")



#Scatterplot matrix for salary and work experience  
scatterplotMatrix(formula=~salary+work\_yrs,cex=0.6,data=MBA.df,diagonal="histogram")



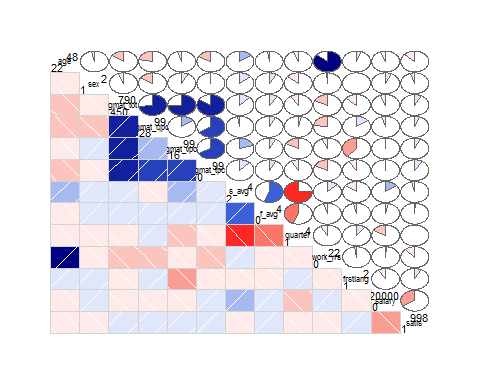
#Scatterplot matrix for salary and first language  
scatterplotMatrix(formula=~salary+frstlang,cex=0.6,data=MBA.df,diagonal="histogram")



#corrgram  
library(corrgram)

## Warning: package 'corrgram' was built under R version 3.3.3

corrgram(MBA.df, lower.panel=panel.shade, upper.panel=panel.pie, diag.panel=panel.minmax, text.panel=panel.txt)



#Subset of candidates who got a job fith factors that affect it  
job <- subset(MBA.df, salary>1000, select = sex:salary)  
#Subset of candidates who are not employed  
unemployed.sub <- subset(MBA.df, salary<1, select = sex:salary)  
  
#Regression model  
#Regression model through origin to estimate the salary (Y) from predictors sex(X1), overall GMAT(X2),spring avg(X3), Fall avg(X4), work exp(X5),First lan(X6)  
#Y=B1\*X1+B2\*X2+B3\*X3+B4\*X4+B5\*X5+B6\*X6+B7\*X7  
#Not taking intercept and running the regression through origin as without experience and degree, no one can get a job of high salary  
reg1<- lm(salary~sex+gmat\_tpc+s\_avg+f\_avg+work\_yrs+frstlang+quarter-1, data = job)  
summary(reg1)

##   
## Call:  
## lm(formula = salary ~ sex + gmat\_tpc + s\_avg + f\_avg + work\_yrs +   
## frstlang + quarter - 1, data = job)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45154 -8487 -402 6770 82372   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## sex -5719.51 3651.42 -1.566 0.120550   
## gmat\_tpc 48.74 144.80 0.337 0.737164   
## s\_avg 19950.67 5324.42 3.747 0.000305 \*\*\*  
## f\_avg 1192.11 3931.03 0.303 0.762350   
## work\_yrs 2245.94 607.90 3.695 0.000366 \*\*\*  
## frstlang 20797.73 6506.96 3.196 0.001885 \*\*   
## quarter 4535.44 1668.25 2.719 0.007779 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16510 on 96 degrees of freedom  
## Multiple R-squared: 0.9768, Adjusted R-squared: 0.9751   
## F-statistic: 576.5 on 7 and 96 DF, p-value: < 2.2e-16

#removing the insignificant variables  
reg2<- lm(salary~s\_avg+work\_yrs+frstlang+sex-1, data = job)  
summary(reg2)

##   
## Call:  
## lm(formula = salary ~ s\_avg + work\_yrs + frstlang + sex - 1,   
## data = job)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -39107 -8995 -487 7998 80935   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## s\_avg 21327.3 2354.9 9.057 1.25e-14 \*\*\*  
## work\_yrs 1993.7 591.6 3.370 0.00107 \*\*   
## frstlang 31366.5 5964.9 5.259 8.39e-07 \*\*\*  
## sex -3446.1 3721.8 -0.926 0.35674   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17260 on 99 degrees of freedom  
## Multiple R-squared: 0.9738, Adjusted R-squared: 0.9728   
## F-statistic: 920.3 on 4 and 99 DF, p-value: < 2.2e-16

#Model 2 gives beeter r square so considering it  
# Positive t value for s\_avg, first language English and work experience(Higher the positive value, lesser the likelihood of value being 0 by chance)  
# The model's, p-value:2.2e-16 is also lower than the statistical significance level of 0.05, this indicates that we can safely reject the null hypothesis that the value for the coefficient is zero (or in other words, the predictor variable has no explanatory relationship with the response variable)  
# The model has a F Statistic of 920.3 on 4 and 99 DF which is moderately high  
#t.test  
t.test(MBA.df$salary~MBA.df$sex)

##   
## Welch Two Sample t-test  
##   
## data: MBA.df$salary by MBA.df$sex  
## t = -1.123, df = 111.89, p-value = 0.2638  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -22412.277 6197.373  
## sample estimates:  
## mean in group 1 mean in group 2   
## 37013.62 45121.07

t.test(MBA.df$salary~MBA.df$frstlang)

##   
## Welch Two Sample t-test  
##   
## data: MBA.df$salary by MBA.df$frstlang  
## t = 1.3595, df = 38.488, p-value = 0.1819  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6698.237 34122.797  
## sample estimates:  
## mean in group 1 mean in group 2   
## 40627.12 26914.84

#Chi-square test  
mytable <- xtabs(~salary+gmat\_tpc, data = job)  
chisq.test(mytable)

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

##   
## Pearson's Chi-squared test  
##   
## data: mytable  
## X-squared = 1422.2, df = 1230, p-value = 0.0001065

mytable <- xtabs(~salary+s\_avg, data = job)  
chisq.test(mytable)

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

##   
## Pearson's Chi-squared test  
##   
## data: mytable  
## X-squared = 792.97, df = 861, p-value = 0.9524

mytable <- xtabs(~salary+f\_avg, data = job)  
chisq.test(mytable)

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

##   
## Pearson's Chi-squared test  
##   
## data: mytable  
## X-squared = 596.28, df = 574, p-value = 0.2518

mytable <- xtabs(~salary+quarter, data = job)  
chisq.test(mytable)

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

##   
## Pearson's Chi-squared test  
##   
## data: mytable  
## X-squared = 129.85, df = 123, p-value = 0.3186

mytable <- xtabs(~salary+work\_yrs, data = job)  
chisq.test(mytable)

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

##   
## Pearson's Chi-squared test  
##   
## data: mytable  
## X-squared = 535.23, df = 451, p-value = 0.003809

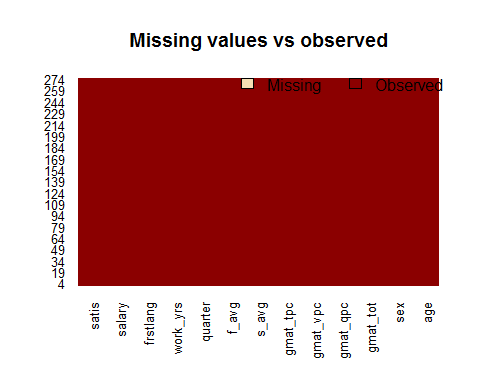
#Logistic Regression  
library(Amelia)

## Warning: package 'Amelia' was built under R version 3.3.3

## Loading required package: Rcpp

## ##   
## ## Amelia II: Multiple Imputation  
## ## (Version 1.7.4, built: 2015-12-05)  
## ## Copyright (C) 2005-2017 James Honaker, Gary King and Matthew Blackwell  
## ## Refer to http://gking.harvard.edu/amelia/ for more information  
## ##

missmap(MBA.df, main = "Missing values vs observed")



#No missing value so will consider whole data