Avaliação dados MBA

Objetivo:

Análise do salário inicial de recem formados em MBA

#1. Leitura dos dados

read.csv2("./dados/mba.csv", stringsAsFactors = FALSE) -> mba  
  
str(mba)

## '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 ...

#2. Analise do Salários e limpeza do banco

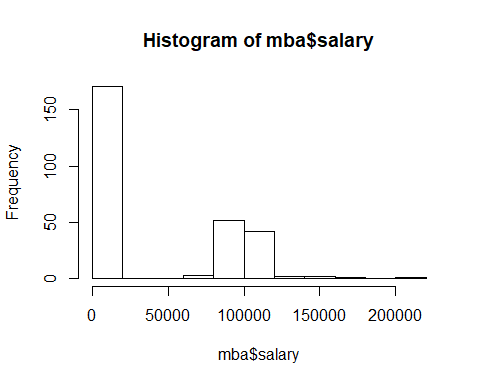
mean(mba$salary)

## [1] 39025.69

median(mba$salary)

## [1] 999

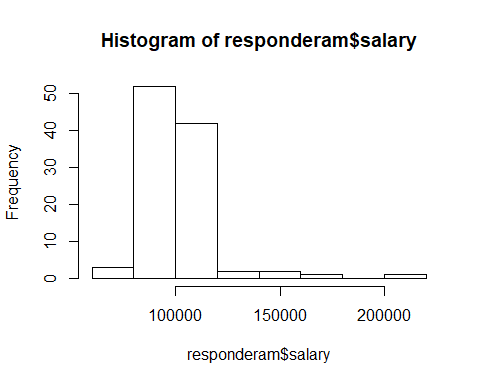
hist(mba$salary)



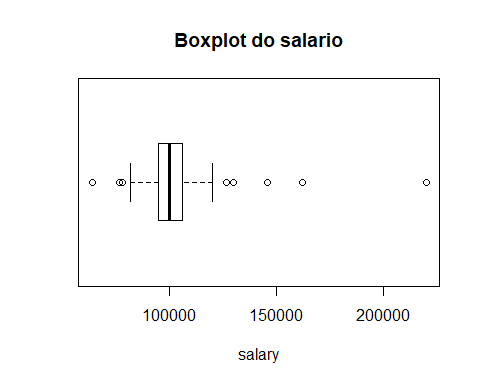
# Estudantes que revelaram o seu salario  
responderam <- mba[which (mba$salary > 1000) , ]  
dim(responderam)

## [1] 103 13

hist(responderam$salary)



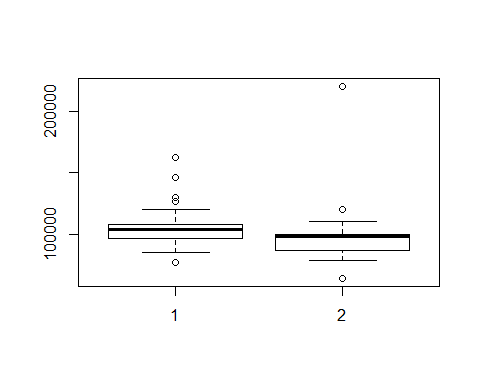
#describe(responderam)  
  
boxplot(responderam$salary,  
 main= "Boxplot do salario",  
 horizontal=TRUE,  
 xlab="salary")



oneway.test(responderam, formula=salary~sex)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: salary and sex  
## F = 1.8573, num df = 1.000, denom df = 38.115, p-value = 0.1809

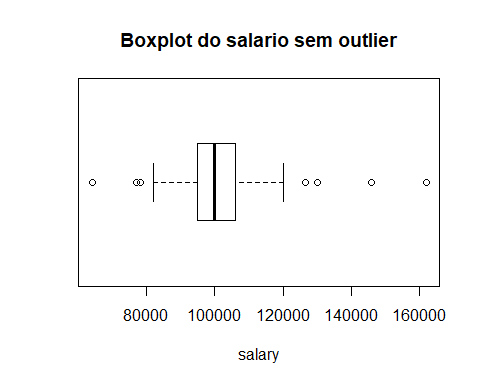
boxplot(responderam$salary ~ responderam$sex)



#Valor p alto, aceita a Hipotese Nula, os salários de homens e mulheres em média são iguais

#Retirada do Outlier

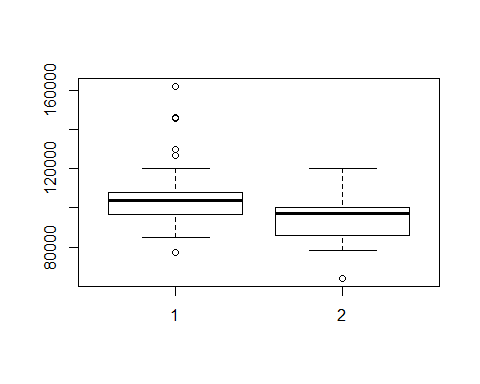
responderamsoutlier<- responderam[which (responderam$salary < 200000) , ]  
  
boxplot(responderamsoutlier$salary,  
 main= "Boxplot do salario sem outlier",  
 horizontal=TRUE,  
 xlab="salary")



oneway.test(responderamsoutlier, formula=salary~sex)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: salary and sex  
## F = 17.729, num df = 1.000, denom df = 70.693, p-value = 7.384e-05

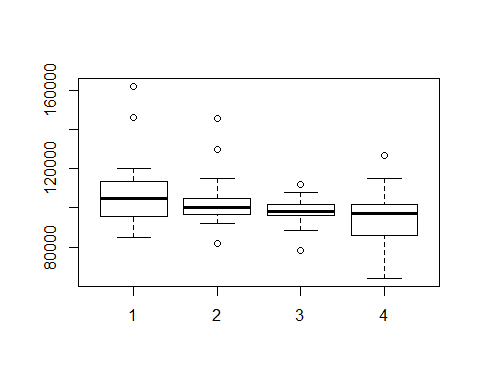
boxplot(responderamsoutlier$salary ~ responderamsoutlier$sex)



oneway.test(responderamsoutlier$salary ~ responderamsoutlier$quarter)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: responderamsoutlier$salary and responderamsoutlier$quarter  
## F = 3.4424, num df = 3.000, denom df = 47.963, p-value = 0.02389

boxplot(responderamsoutlier$salary ~ responderamsoutlier$quarter)



regressao1<-lm(responderamsoutlier$salary ~ responderamsoutlier$quarter)  
regressao1

##   
## Call:  
## lm(formula = responderamsoutlier$salary ~ responderamsoutlier$quarter)  
##   
## Coefficients:  
## (Intercept) responderamsoutlier$quarter   
## 110290 -3744

#Quem está no primeiro quartil tem salário em média mais alto  
  
#redução do salário anual em -3744 por diminuição do quartil

regressao2<-lm(responderamsoutlier$salary ~ responderamsoutlier$gmat\_tot)  
summary (regressao2)

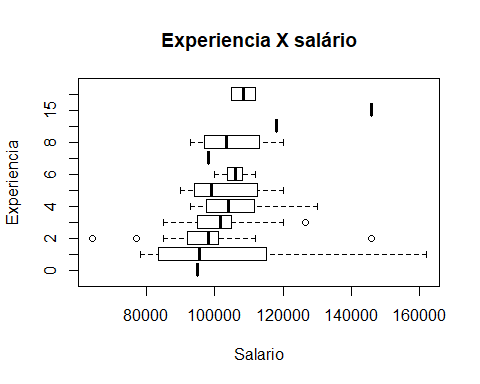
##   
## Call:  
## lm(formula = responderamsoutlier$salary ~ responderamsoutlier$gmat\_tot)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36659 -6410 -1745 4405 58340   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 88652.78 16956.49 5.228 9.4e-07 \*\*\*  
## responderamsoutlier$gmat\_tot 21.44 27.39 0.783 0.436   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13650 on 100 degrees of freedom  
## Multiple R-squared: 0.00609, Adjusted R-squared: -0.003849   
## F-statistic: 0.6128 on 1 and 100 DF, p-value: 0.4356

regressao3<-lm(responderamsoutlier$salary ~ responderamsoutlier$frstlang)  
summary (regressao3)

##   
## Call:  
## lm(formula = responderamsoutlier$salary ~ responderamsoutlier$frstlang)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37749 -6749 -1749 4001 60251   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 99447 6245 15.92 <2e-16 \*\*\*  
## responderamsoutlier$frstlang 2301 5758 0.40 0.69   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13680 on 100 degrees of freedom  
## Multiple R-squared: 0.001595, Adjusted R-squared: -0.008389   
## F-statistic: 0.1598 on 1 and 100 DF, p-value: 0.6902

#R2 baixo

boxplot(salary ~ work\_yrs ,data=responderamsoutlier, main="Experiencia X salário", ylab="Experiencia", xlab="Salario", horizontal=TRUE)



library(car)

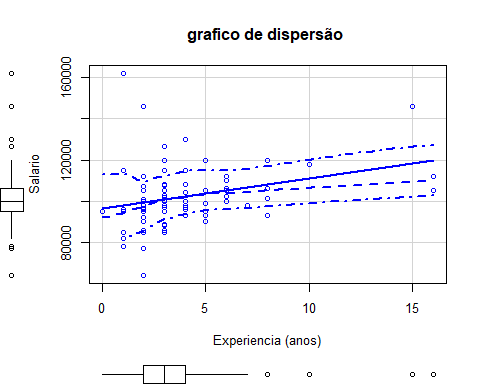
## Loading required package: carData

##   
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':  
##   
## some

scatterplot(salary ~ work\_yrs,   
 data=responderamsoutlier,  
 main="grafico de dispersão ",  
 xlab="Experiencia (anos)",  
 ylab="Salario")



#Análise de correlação ( achei bem legal)  
library(corrplot)

## corrplot 0.84 loaded

C <-cor(responderamsoutlier [,   
 c("age",  
 "work\_yrs",  
 "gmat\_tot",  
 "gmat\_qpc",  
 "gmat\_vpc",  
 "gmat\_tpc",  
 "s\_avg",  
 "f\_avg",  
 "quarter",  
 "satis")])   
corrplot(C, method="circle")

