Avaliação dados MBA

Alunos:

Bruno Santos Wance de Souza

Lucas de Jesus Matias

Luiz Cesar Costa Raymund

Objetivo:

Predizer, a partir da manipulação dos dados disponíveis, o salário inicial de um ex-aluno do MBA em questão.

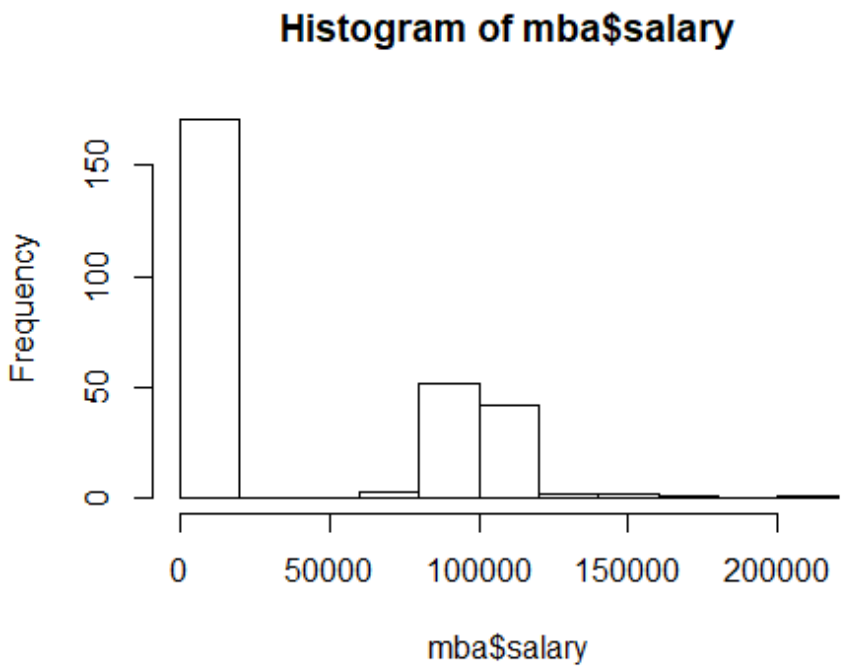
1. Leitura dos dados

```
library(readxl)

## Warning: package 'readxl' was built under R version 3.5.1

mba <-read_excel("C:/Users/cesar/Desktop/Analise Exploratória
MBA/mba.xlsx")
View(mba)
str(mba)

## Classes 'tbl_df', 'tbl' and 'data.frame':    274 obs. of  13
variables:
##  $ age     : num  23 24 24 24 24 24 25 25 25 25 ...
##  $ sex     : num  2 1 1 1 2 1 1 2 1 1 ...
##  $ gmat_tot: num  620 610 670 570 710 640 610 650 630 680 ...
##  $ gmat_qpc: num  77 90 99 56 93 82 89 88 79 99 ...
##  $ gmat_vpc: num  87 71 78 81 98 89 74 89 91 81 ...
##  $ gmat_tpc: num  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 : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ work_yrs: num  2 2 2 1 2 2 2 2 2 2 ...
##  $ frstlang: num  1 1 1 1 1 1 1 1 2 1 ...
##  $ salary  : num  0 0 0 0 999 0 0 0 999 998 ...
##  $ satis   : num  7 6 6 7 5 6 5 6 4 998 ...
```

2. Analise do comportamento da variável “salary”:

- Para avaliar o comportamento da variável “salary” optado por comparar os valores absolutos de média e mediana e, por se tratar de uma variável contínua, traçar um histograma. A partir dos resultados encontrados, percebemos que o comportamento da variável não corresponde a uma normal.

```
mean(mba$salary)
## [1] 39025.69
median(mba$salary)
## [1] 999
hist(mba$salary)
```

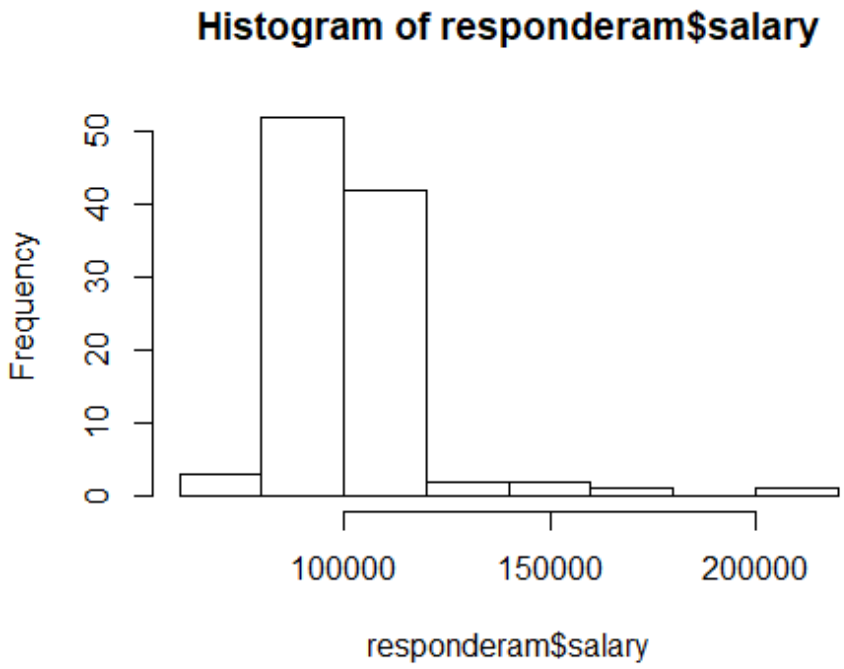


3. Limpando a variável “salary”

- A partir da observação de que valores inferiores a $1.000 foram computados para ex-alunos que não divulgaram seu salário, optado por retirar esses indivíduos da análise, considerando então apenas aqueles que responderam.

- Ao avaliar o histograma e o boxplot da “nova variável salary” percebemos que a variável possui curva que se assemelha a uma distribuição normal. Utilizaremos portanto essa nova base. (percebemos também no boxplot a existência de outliers; um indivíduo ganha especial destaque já que possui salário superior a $200.000), que retiramos da amostra.

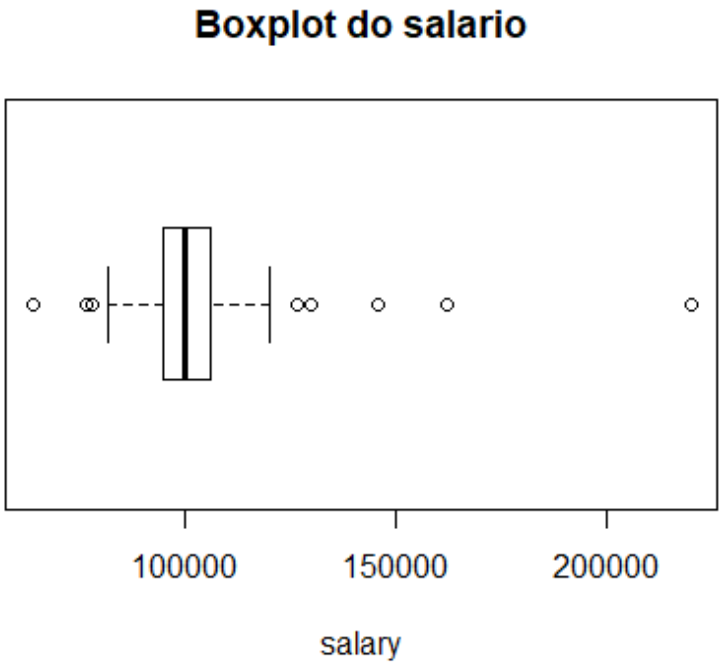
```
responderam <-mba[which (mba$salary >1000)  , ]
dim(responderam)

## [1] 103  13

View (responderam)
hist(responderam$salary)
```




```
describe(responderam)

##          vars   n      mean       sd   median   trimmed      mad
min
## age         1 103     26.78     3.27 2.60e+01     26.30     2.97
22.0
## sex         2 103      1.30     0.46 1.00e+00      1.25     0.00
1.0
## gmat_tot    3 103    616.02    50.69 6.20e+02    615.90    59.30
500.0
## gmat_qpc    4 103     79.73    13.39 8.20e+01     81.05    13.34
39.0
```

```
## gmat_vpc    5 103     78.56    16.14 8.10e+01     80.33   16.31
30.0
## gmat_tpc    6 103     84.52    11.01 8.70e+01     85.60   11.86
51.0
## s_avg       7 103      3.09     0.38 3.10e+00      3.10    0.44
2.2
## f_avg       8 103      3.09     0.49 3.25e+00      3.13    0.37
0.0
## quarter     9 103      2.26     1.12 2.00e+00      2.20    1.48
1.0
## work_yrs   10 103      3.68     3.01 3.00e+00      3.11    1.48
0.0
## frstlang   11 103      1.07     0.25 1.00e+00      1.00    0.00
1.0
## salary     12 103 103030.74 17868.80 1.00e+05 101065.06 7413.00
64000.0
## satis      13 103      5.88     0.78 6.00e+00      5.89    1.48
3.0
##              max    range  skew kurtosis      se
## age           40     18.0  1.92     4.90    0.32
## sex            2      1.0  0.86    -1.28    0.05
## gmat_tot     720    220.0  0.01    -0.69    4.99
## gmat_qpc      99     60.0 -0.81     0.17    1.32
## gmat_vpc      99     69.0 -0.87     0.21    1.59
## gmat_tpc      99     48.0 -0.84     0.19    1.08
## s_avg          4      1.8 -0.13    -0.61    0.04
## f_avg          4      4.0 -2.52    13.86    0.05
## quarter        4      3.0  0.27    -1.34    0.11
## work_yrs      16     16.0  2.48     6.83    0.30
## frstlang       2      1.0  3.38     9.54    0.02
## salary    220000 156000.0  3.18    17.16 1760.67
## satis          7      4.0 -0.40     0.44    0.08
```

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

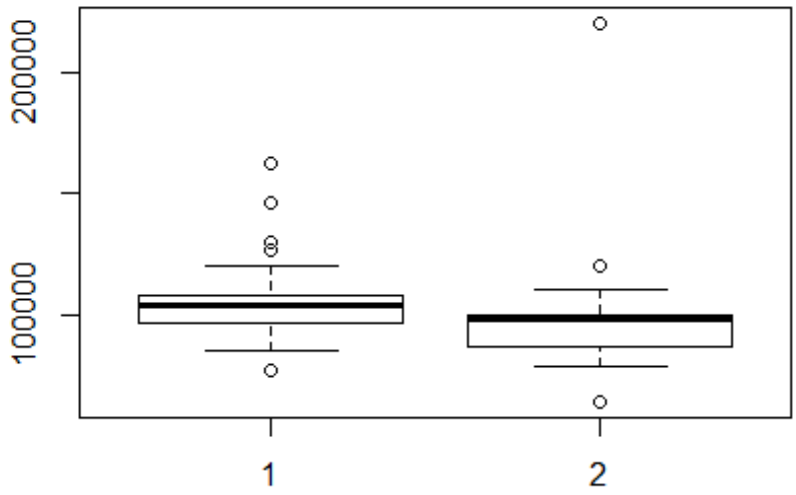


4. Testando a relação entre salário e gênero

- Por se tratar de um cruzamento entre variável quantitativa contínua e qualitativa nominal, optamos por utilizar o teste oneway para testar se existe relação entre “salary” e “sex”.

- Encontrado p valor elevado (0,1809) não sendo possível, portanto rejeitar com segurança a hipótese nula de que “não existe relação entre salário e gênero”

```
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),

Observamos que a média do salário do boxplot em questão, é a praticamente
a mesma, tanto para o Genero Masculino (1) e feminino (2, por isso
comprova a não rejeição da hipotese nula.
```

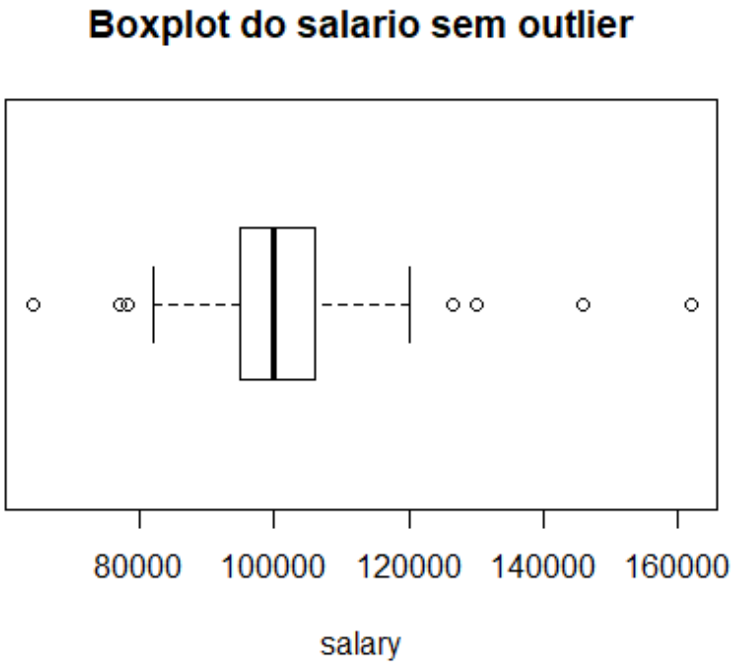


5. **Excluindo o outlier:**

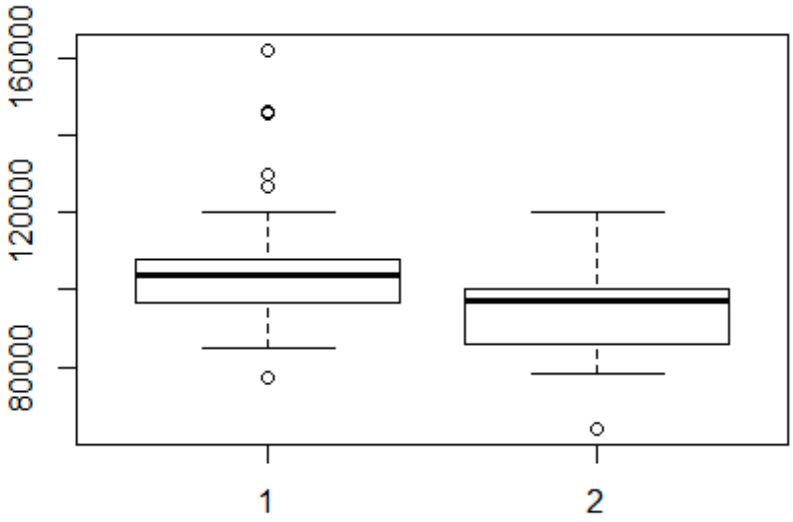
- Optado por excluir o outlier citado anteriormente no item 3 desse relatório (salário superior a $200.000) e refazer o teste de relação entre “salary” e “sex”.
- Encontrado então p valor baixo, o que significa que podemos aceitar a hipótese do investigador com segurança e afirmar que existe relação entre salário e gênero (homens costumam ter salários mais altos, o que pode ser observado de forma mais clara no boxplot “salary” x “sex)

(trabalharemos daqui em diante sem o outlier em questão)

```
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).

Podemos observar abaixo a diferença em media entre os salários do gênero
masculino (1) e feminino (2), assim poderemos rejeitar a hipótese nula.
```

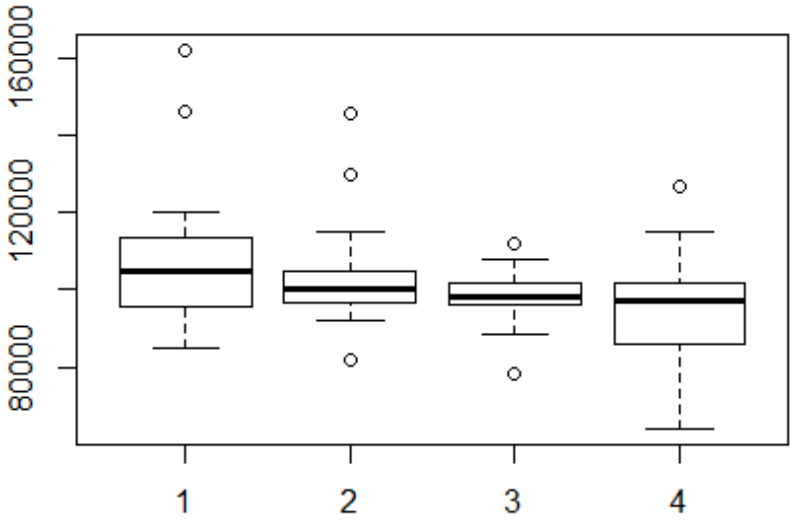


6. Testando a relação entre salário e quartil:

- Por se tratar de um cruzamento entre variável quantitativa contínua e qualitativa categórica, optamos por utilizar o teste oneway para testar se existe relação entre “salary” e “quarter”.

- Encontrado então p valor baixo, o que significa que podemos aceitar a hipótese do investigador com segurança e afirmar que existe relação entre salário e quartil (quanto mais baixo o quartil maior o salário, o que pode ser observado de forma mais clara no boxplot “salary” x “quarter”)

```
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
```

7. Testando a relação entre salário e nota gmat:

- Por se tratar de um cruzamento entre duas variáveis quantitativas contínuas, optamos por utilizar regressão linear para testar se existe relação entre “salary” e “gmat_tot”.

- Encontrado então p valor elevado, o que significa que não podemos rejeitar com segurança a hipótese nula de que não existe relação entre salário e nota gmat.

```
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
```

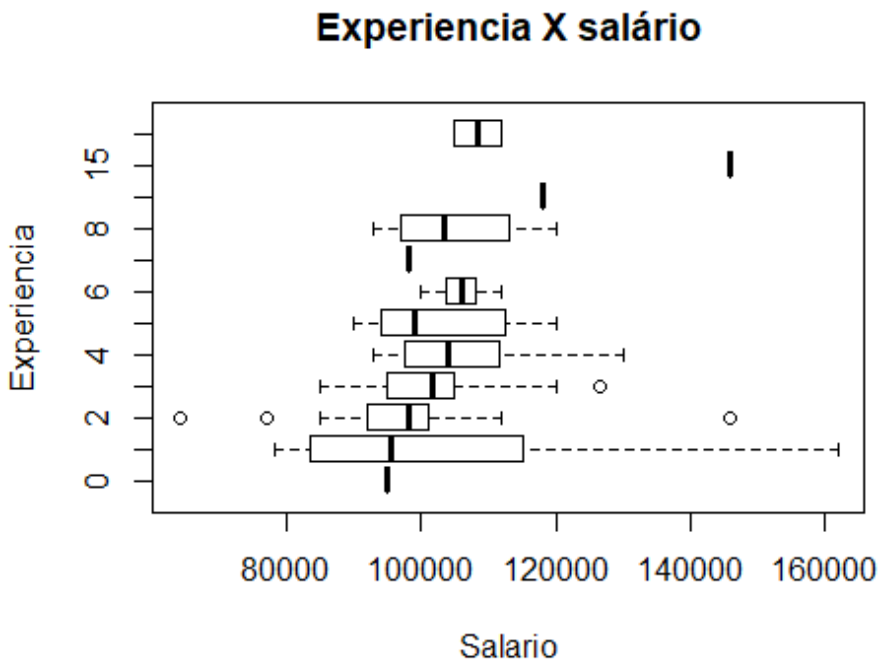
8. Testando a relação entre salário e primeira lingua:

- Realizada regressão linear e encontrado então p valor elevado, o que significa que não podemos rejeitar com segurança a hipótese nula de que não existe relação entre salário e primeira língua inglês.

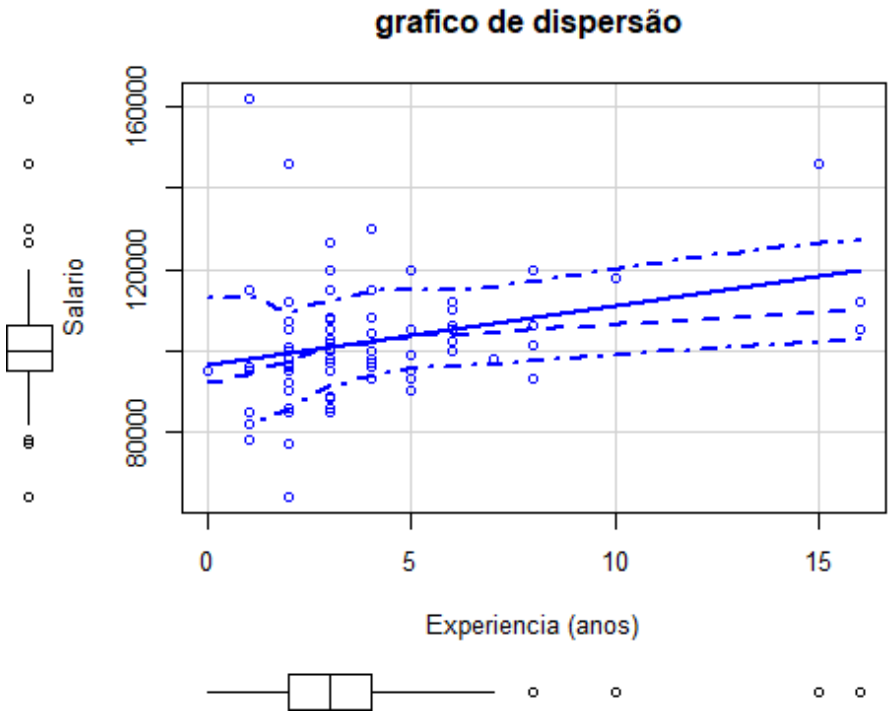
```
regressao3<-lm(responderamsoutlier$salary ~responderamsoutlier$frstlang)
summary (regressao3)

##
## 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
```

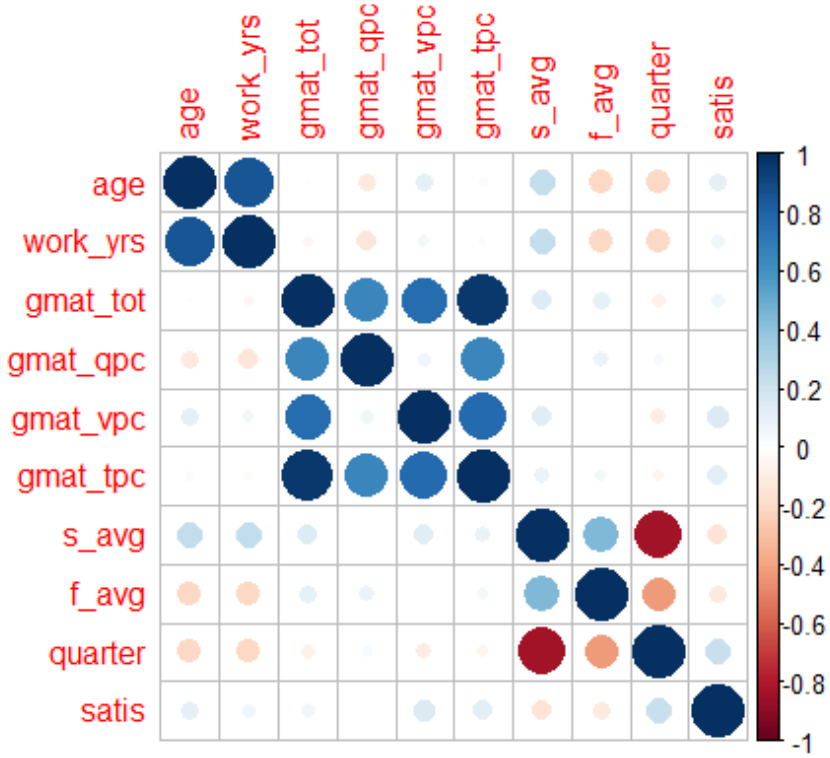
```
boxplot(salary ~work_yrs ,data=responderamsoutlier, main="Experiencia X salário", ylab="Experiencia", xlab="Salario", horizontal=TRUE)
```




```
scatterplot(salary ~work_yrs ,data=responderamsoutlier, main="grafico de dispersão ", xlab="Experiencia (anos)", ylab="Salario")
```




```
#Análise de correlação

Para complementação do trabalho e propostas de regressões futuras,
fizemos a análise de correlação dos dados do MBA, onde há muitas
variáveis correlacionadas entre si e que podemos exclui-las da regressão.
Assim as variaveis finais para fazermos as regressões são “work_yrs”,
“gmat_qpc”, “gmat_vpc”, quarter “, satis”.



library(corrplot)

## Warning: package 'corrplot' was built under R version 3.5.1

## 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")
```


age
work_yrs
gmat_tot
gmat_qpc
gmat_vpc
gmat_tpc
s_avg
f_avg
quarter
satis
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1