CMP.5.0.2 - Técnicas Estatísticas de Predição - Avaliação Final

Diego Leonardo Urban

09 de Junho de 2020

```
# Carregando a biblioteca fBasics
library(fBasics)
```

```
## Loading required package: timeDate
```

```
## Loading required package: timeSeries
```

Cenário

Este documento contém a análise dos dados de um levantamento feito com 1289 pessoas dos Estados Unidos da América.

Os dados estão organizados com as seguintes variáveis:

Rótulo	O que mede
obs	Rótulo das Observações.
salario	Salário dos indivíduos entrevistados – em milhares de reais por ano.
sexo	Se for do sexo feminino recebe valor 1, se for do seco masculino recebe valor 0.
cor	Recebe 1 se a pessoa for não-branca, 0 caso contrário.
est_civil	Recebe 1 caso a pessoa for casada, 0 caso contrário.
instrucao	Anos de educação formal que a pessoa recebeu.
experiencia	Anos de experiência que a possui na área em que trabalha.
idade	Anos de vida que a pessoa entrevistada possui.

```
# Carregando os dados
df <- read.csv("Dados.csv", header = T, sep = ";", dec = ",");
```

Questão 1

Calcule as medidas de posição (Média, Mediana, Máximo, Mínimo, 1o Quartil e 3o quartil) para as variáveis "salario", "instrucao", "experiência" e "idade". Apresente os cálculos e faça uma interpretação dos resultados.

Salário

```
summary(df$salario)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.84    6.92   10.08   12.37   15.63   64.08
```

Instrução

```
summary(df$instrucao)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   13.15   16.00   20.00
```

Experiência

```
summary(df$experiencia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   18.00   18.79   27.00   56.00
```

Idade

```
summary(df$idade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   29.00   37.00   37.93   47.00   65.00
```

O salário aparenta ter uma assimetria a direita pois o primeiro quartil, média, mediana e terceiro quartil estão entre R$ 6.000 e R$ 16.000 por ano, sendo que o salario máximo é de R$ 63.000 por ano.

Há pessoas sem instrução e com no máximo 20 anos de instrução. Média e Mediana entre 12 e 13 anos.

Há pessoas sem experiência e com no máximo 56 anos de experiência. Média e Mediana entre 18 e 19 anos. Aparenta ter assimetria a direita.

A variável Idade tem idade mínima de 18 e máxima de 65 anos. Média e Mediana entre 37 e 38 anos. Aparenta ser uma distribuição normal.

Questão 2

Calcule as medidas de dispersão (Amplitude, Desvio-Padrão, Variância, Coeficiente de Variação, Assimetria e Curtose) para as variáveis “salario”, “instrucao”, “experiência” e “idade”.

Salário

```
cat("Amplitude =", max(df$salario)-min(df$salario), "\n")
```

```
## Amplitude = 63.24
```

```
cat("Desvio Padrão =", sd(df$salario), "\n")
```

```
## Desvio Padrão = 7.89635
```

```r
cat("Variância =", var(df$salario), "\n")
```

```
## Variância = 62.35235
```

```r
cat("Coeficiente de Variação =", sd(df$salario) / mean(df$salario), "\n")
```

```
## Coeficiente de Variação = 0.6385611
```

```r
cat("Assimetria =", skewness(df$salario), "\n")
```

```
## Assimetria = 1.845964
```

```r
cat("Curtose =", kurtosis(df$salario), "\n")
```

```
## Curtose = 4.824411
```

Instrução

```r
cat("Amplitude =", max(df$instrucao)-min(df$instrucao), "\n")
```

```
## Amplitude = 20
```

```r
cat("Desvio Padrão =", sd(df$instrucao), "\n")
```

```
## Desvio Padrão = 2.813823
```

```r
cat("Variância =", var(df$instrucao), "\n")
```

```
## Variância = 7.917602
```

```r
cat("Coeficiente de Variação =", sd(df$instrucao) / mean(df$instrucao), "\n")
```

```
## Coeficiente de Variação = 0.2140592
```

```r
cat("Assimetria =", skewness(df$instrucao), "\n")
```

```
## Assimetria = -0.2900436
```

```r
cat("Curtose =", kurtosis(df$instrucao), "\n")
```

```
## Curtose = 2.968193
```

Experiência

```r
cat("Amplitude =", max(df$experiencia)-min(df$experiencia), "\n")
```

```
## Amplitude = 56
```

```r
cat("Desvio Padrão =", sd(df$experiencia), "\n")
```

```
## Desvio Padrão = 11.66284
```

```r
cat("Variância =", var(df$experiencia), "\n")
```

```
## Variância = 136.0218
```

```r
cat("Coeficiente de Variação =", sd(df$experiencia) / mean(df$experiencia), "\n")
```

```
## Coeficiente de Variação = 0.6207018
```

```r
cat("Assimetria =", skewness(df$experiencia), "\n")
```

```
## Assimetria = 0.3752323
```

```r
cat("Curtose =", kurtosis(df$experiencia), "\n")
```

```
## Curtose = -0.675665
```

Idade

```r
cat("Amplitude =", max(df$idade)-min(df$idade), "\n")
```

```
## Amplitude = 47
```

```r
cat("Desvio Padrão =", sd(df$idade), "\n")
```

```
## Desvio Padrão = 11.49428
```

```r
cat("Variância =", var(df$idade), "\n")
```

```
## Variância = 132.1184
```

```r
cat("Coeficiente de Variação =", sd(df$idade) / mean(df$idade), "\n")
```

```
## Coeficiente de Variação = 0.3030006
```

```r
cat("Assimetria =", skewness(df$idade), "\n")
```

```
## Assimetria = 0.2692705
```

```r
cat("Curtose =", kurtosis(df$idade), "\n")
```

```
## Curtose = -0.772023
```

2.1 Com relação ao Coeficiente de Variação, qual é a variável que possui maior discrepância em seus valores? E a com menor discrepância?

A variável Instrução possui menor discrepância com 0.2140592 enquanto que a variável de maior discrepância é o Salário com 0.6385611.

2.2 Qual deve ser a interpretação dada ao Coeficiente de Variação?

Quanto menor o Coeficiente de Variação mais homogêneo é o conjunto de dados.

2.3 Considerando que as medidas de Assimetria e Curtose qualificam a média como boa medida de tendência central, existe alguma das variáveis que possua problemas de assimetria e/ou curtose? Justifique.

As variáveis Experiência e Idade aparentam ter medidas de assimetria e curtose mais leves enquanto que o Salário e Instrução aparentam serem mais acentuadas.

Questão 3

Considere uma análise que possa ser realizada sobre a variável salario. Faça os procedimentos destacados a seguir:

3.1 Calcule a média e a mediana do "salario" para mulheres e homens separadamente. Qual é a tendência apresentada para média e para mediana?

```r
df_homens <- subset(df, (sexo == 0))
df_mulheres <- subset(df, (sexo == 1))

media_salario_homens <- mean(df_homens$salario)
mediana_salario_homens <- median(df_homens$salario)
cat("Média do Salário dos Homens =", media_salario_homens, "\n")
```

```
## Média do Salário dos Homens = 14.11889
```

```r
cat("Mediana do Salário dos Homens =", mediana_salario_homens, "\n\n")
```

```
## Mediana do Salário dos Homens = 12
```

```r
media_salario_mulheres <- mean(df_mulheres$salario)
mediana_salario_mulheres <- median(df_mulheres$salario)
cat("Média do Salário das Mulheres =", media_salario_mulheres, "\n")
```

```
## Média do Salário das Mulheres = 10.59367
```

```r
cat("Mediana do Salário das Mulheres =", mediana_salario_mulheres, "\n")
```

```
## Mediana do Salário das Mulheres = 8.89
```

Há uma tendência de homens ganharem mais do que mulheres.

3.2 Calcule a média do "salario" para brancos e não brancos. Qual é a tendência apresentada para média e para mediana?

```r
df_brancos <- subset(df, (cor == 0))
df_nao_brancos <- subset(df, (cor == 1))

media_salario_brancos <- mean(df_brancos$salario)
mediana_salario_brancos <- median(df_brancos$salario)
cat("Média do Salário de Brancos =", media_salario_brancos, "\n")
```

```
## Média do Salário de Brancos = 12.79442
```

```r
cat("Mediana do Salário de Brancos =", mediana_salario_brancos, "\n\n")
```

```
## Mediana do Salário de Brancos = 11
```

```r
media_salario_nao_brancos <- mean(df_nao_brancos$salario)
mediana_salario_nao_brancos <- median(df_nao_brancos$salario)
cat("Média do Salário de Não Brancos =", media_salario_nao_brancos, "\n")
```

```
## Média do Salário de Não Brancos = 9.990203
```

```r
cat("Mediana do Salário de Não Brancos =", mediana_salario_nao_brancos, "\n")
```

```
## Mediana do Salário de Não Brancos = 8
```

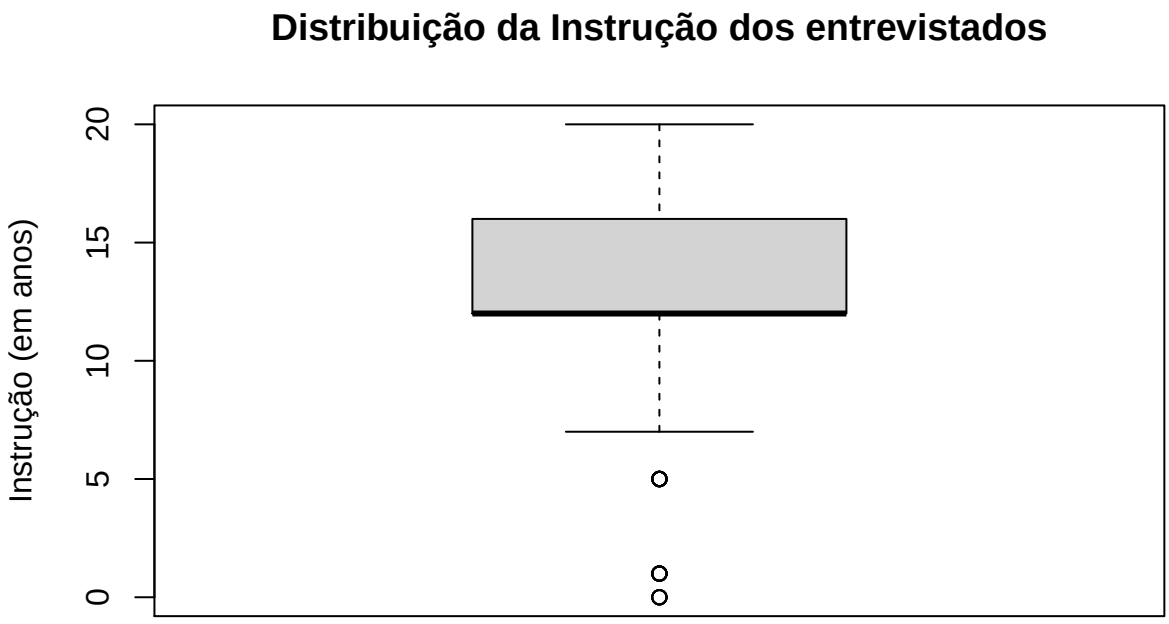
Há uma tendência de pessoas Brancas em média, ganharem mais do que pessoas Não Brancas.

Questão 4

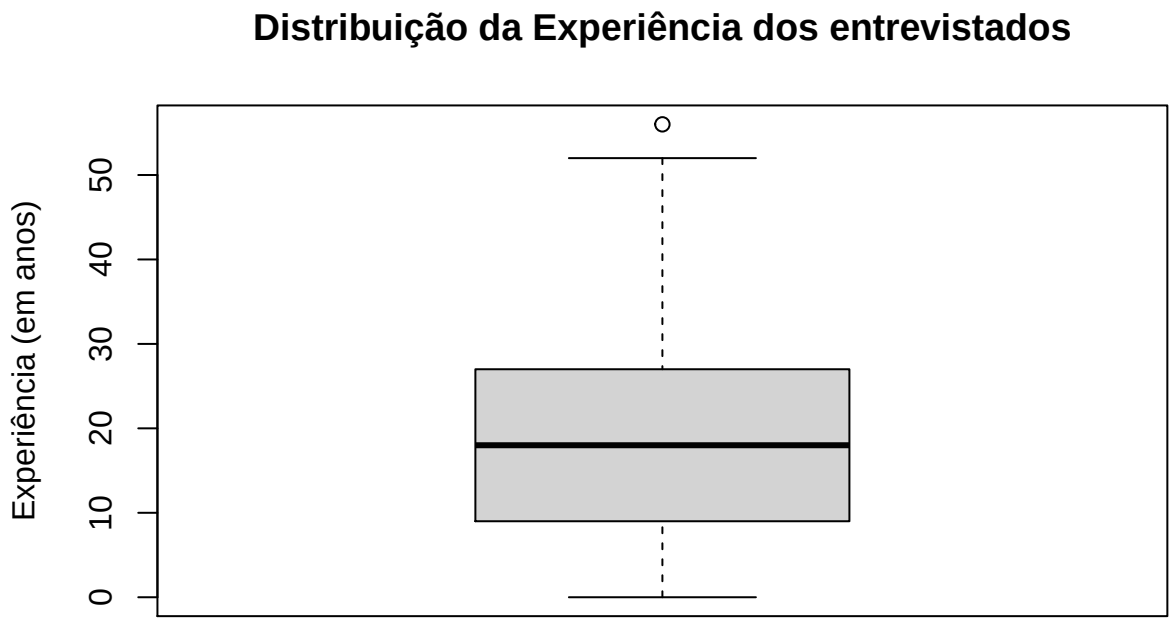
Faça um gráfico Box-Plot para as variáveis salario, instrucao, experiencia e idade e identifique se existem outliers. Quantas observações deveriam ser excluídas em cada variável por serem prováveis outliers?

```r
boxplot(df$salario, ylab="Salário (em milhares por ano)", main="Distribuição do Salário dos entrevistado
```

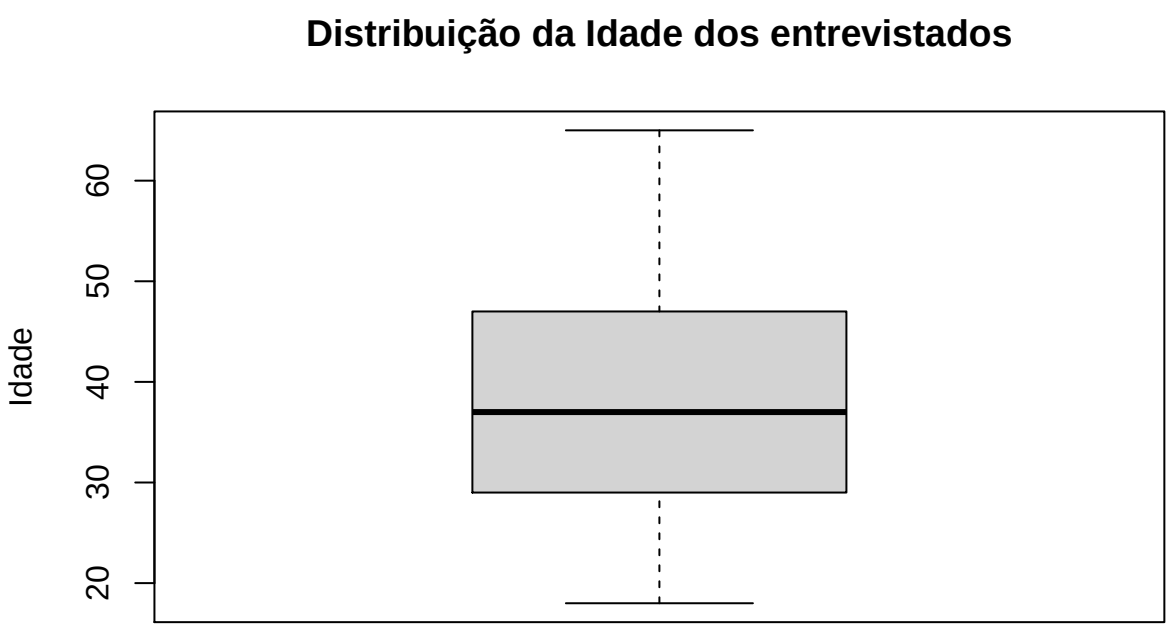

```
boxplot(df$instrucao, ylab="Instrução (em anos)", main="Distribuição da Instrução dos entrevistados")
```




```
boxplot(df$experiencia, ylab="Experiência (em anos)", main="Distribuição da Experiência dos entrevistado
```




```
boxplot(df$idade, ylab="Idade", main="Distribuição da Idade dos entrevistados")
```



A variável Idade não apresenta outliers. Experiência e Instrução tem 1 e 3 outliers respectivamente, enquanto que Salário há dezenas de outliers.

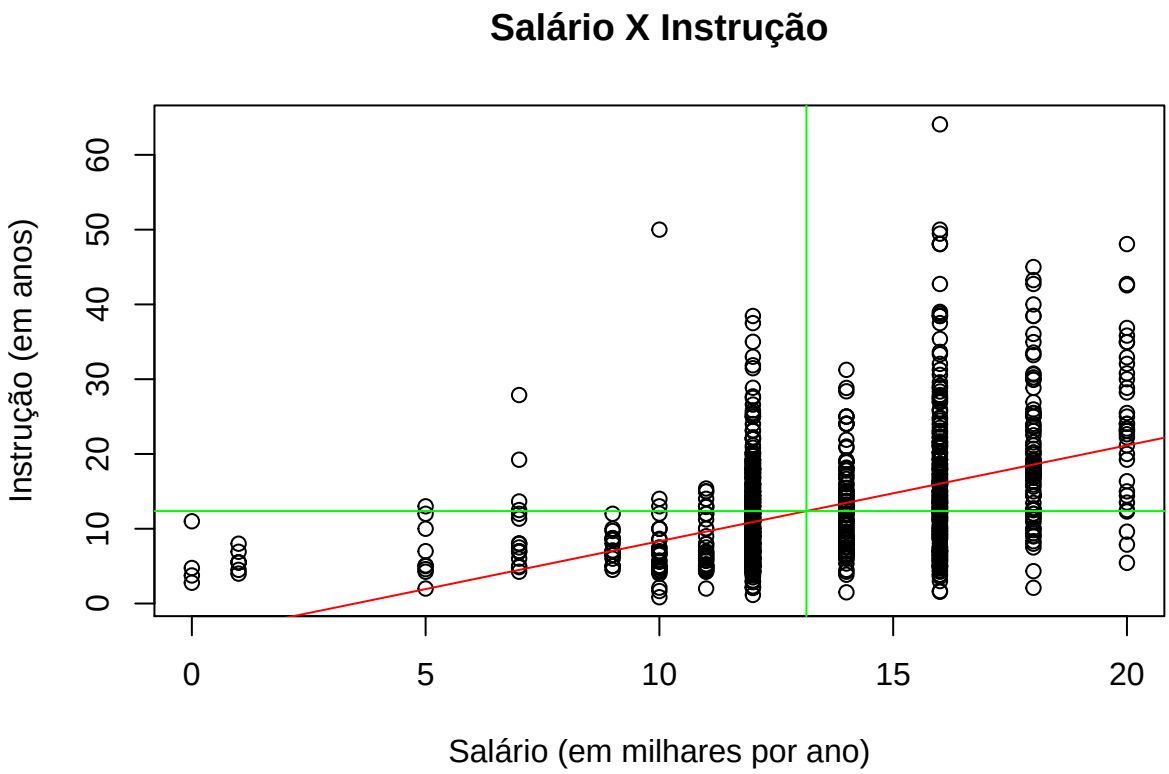
Questão 5

Considerando os gráficos de dispersão, construa-os conforme pedido a seguir:

```
salario <- df$salario
instrucao <- df$instrucao
experiencia <- df$experiencia
idade <- df$idade
```

5.1 Faça um gráfico que relacione o "salario" com o tempo de "instrucao". Analise uma eventual tendência dos dados.

```
plot(salario~instrucao, xlab="Salário (em milhares por ano)", ylab="Instrução (em anos)", main="Salário
model <- lm(salario~instrucao)
abline(model, col="red")
abline(h=mean(salario), col="green")
abline(v=mean(instrucao), col="green")
```



Há uma tendência de que quanto mais anos de instrução o entrevistado tem, maior é o seu salário.

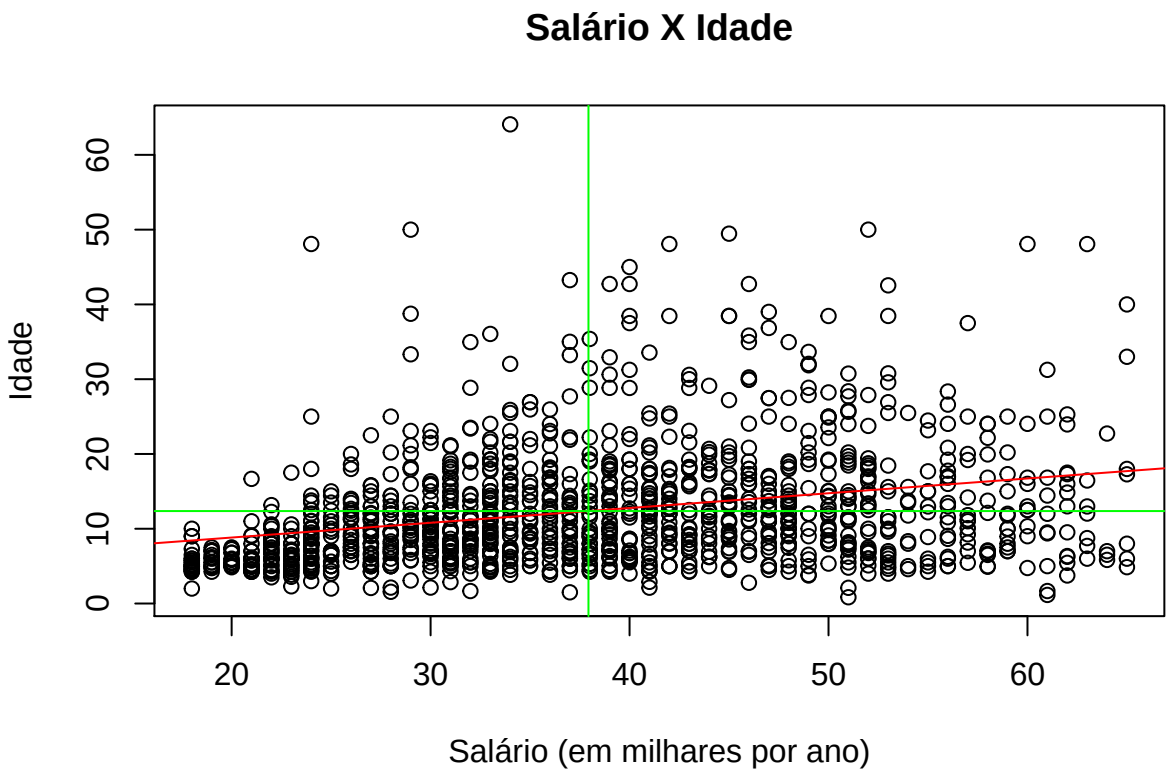
5.2 Faça um gráfico que relacione o "salario" com o tempo de "experiência". Analise uma eventual tendência dos dados.

```
plot(salario~experiencia, xlab="Salário (em milhares por ano)", ylab="Experiência (em anos)", main="Sala
model <- lm(salario~experiencia)
abline(model, col="red")
abline(h=mean(salario), col="green")
abline(v=mean(experiencia), col="green")
```

Há uma leve tendência de que quanto mais anos de experiência o entrevistado tem, maior é o seu salário.

5.3 Faça um gráfico que relacione o "salario" com a "idade". Analise uma eventual tendência dos dados.

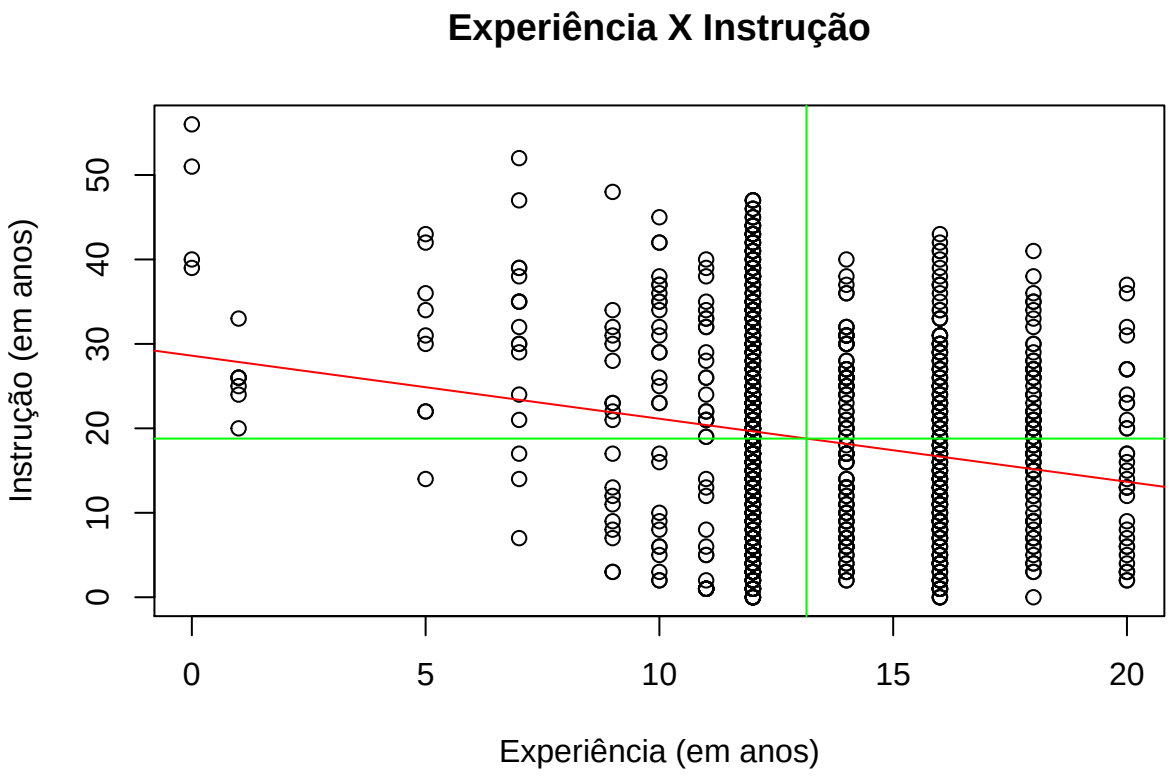
```
plot(salario~idade, xlab="Salário (em milhares por ano)", ylab="Idade", main="Salário X Idade")
model <- lm(salario~idade)
abline(model, col="red")
abline(h=mean(salario), col="green")
abline(v=mean(idade), col="green")
```



Há uma leve tendência de que quanto maior a indade do entrevistado, maior é o seu salário.

5.4 Faça um gráfico que relacione a "experiencia" com o tempo de "instrucao". Analise uma eventual tendência dos dados.

```
plot(experiencia~instrucao, xlab="Experiência (em anos)", ylab="Instrução (em anos)", main="Experiência
model <- lm(experiencia~instrucao)
abline(model, col="red")
abline(h=mean(experiencia), col="green")
abline(v=mean(instrucao), col="green")
```



Há uma tendência de que entrevistados com mais anos de experiência tem menos anos de instrução.

Questão 6

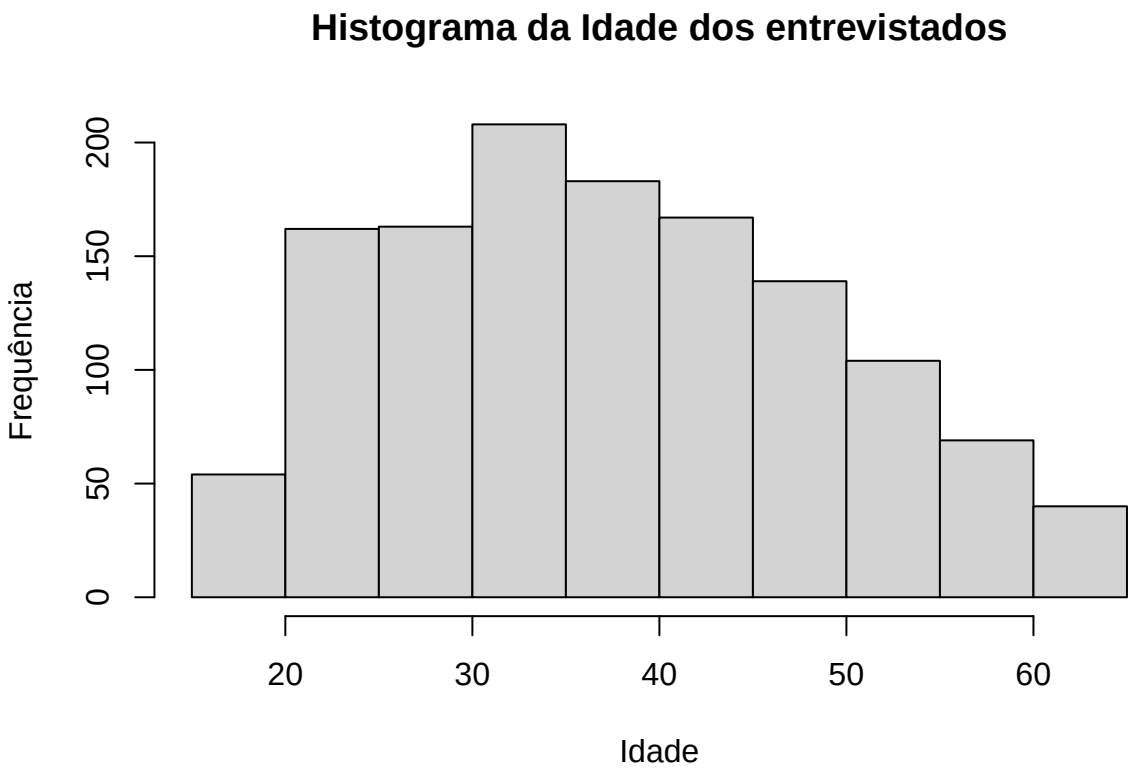
Considerando as variáveis estritamente quantitativas. Construa um Histograma e identifique a variável com melhor ajuste percebido para a distribuição normal de probabilidade.

```
hist(salario,
     main="Histograma do Salário dos entrevistados",
     xlab="Salário (em milhares por ano)",
     ylab="Frequência")
```


```
hist(instrucao,
     main="Histograma da Instrução dos entrevistados",
     xlab="Instrução (em anos)",
     ylab="Frequência")
```


```
hist(experiencia,
     main="Histograma da Experiência dos entrevistados",
     xlab="Experiência (em anos)",
     ylab="Frequência")
```


```r
hist(idade,
     main="Histograma da Idade dos entrevistados",
     xlab="Idade",
     ylab="Frequência")
```



A variável Idade aparenta ser a variável com melhor ajuste para a distribuição normal.

Questão 7

Considere que a variável "salario" segue uma distribuição normal de probabilidade. A média e o desvio-padrão já foram calculados. Assim determine o que se pede:

7.1 Qual a probabilidade estimada de uma pessoa ganhar mais do que o 3o quartil?

```
media_salario <- mean(df$salario)
mediana_salario <- median(df$salario)
dp_salario <- sd(df$salario)

q3 <- quantile(df$salario, probs=0.75)
prob_salario_maior_q3 <- 1 - pnorm(q3, mean = media_salario, sd = dp_salario)
cat("Probabilidade estimada de uma pessoa ganhar mais do que o 3º quartil =", prob_salario_maior_q3, "\n
```

```
## Probabilidade estimada de uma pessoa ganhar mais do que o 3º quartil = 0.3396661
```

7.2 Qual a probabilidade estimada de uma pessoa ganhar menos do que o 1o quartil?

```
q1 <- quantile(df$salario, probs=0.25)
prob_salario_menor_q1 <- pnorm(q1, mean = media_salario, sd = dp_salario)
cat("Probabilidade estimada de uma pessoa ganhar menos do que o 1º quartil =", prob_salario_menor_q1, "
```

```
## Probabilidade estimada de uma pessoa ganhar menos do que o 1º quartil = 0.2452019
```

7.3 O que é mais provável, considerando a probabilidade estimada, a pessoa ganhar menos do que a média ou a pessoa ganhar menos do que a mediana?

```
prob_salario_menor_media <- pnorm(media_salario, mean = media_salario, sd = dp_salario)
prob_salario_menor_mediana <- pnorm(mediana_salario, mean = media_salario, sd = dp_salario)

cat("Probabilidade estimada de uma pessoa ganhar menos que a média =", prob_salario_menor_media, "\n")
```

```
## Probabilidade estimada de uma pessoa ganhar menos que a média = 0.5
```

```
cat("Probabilidade estimada de uma pessoa ganhar menos que a mediana =", prob_salario_menor_mediana, "\n
```

```
## Probabilidade estimada de uma pessoa ganhar menos que a mediana = 0.3861064
```

É mais provável uma pessoa ganhar menos que a média.