# Exercicio1.R

Maria Eduarda Betman, Rafael Nascimento e Rafael Ribeiro de Lima

2024-04-03

O primeiro passo para a elaboração deste trabalho foi a importação da tabela “dados_salarios.csv” para a variável “dados”.

```
# 1) Classifique todas as variáveis que estão na base de dados em relação a qualitativa (nominal/ordinal) e quantitativa (discreta/contínua);

dados = read.csv2("dados_salarios.csv", dec=".")
names(dados)

## [1] "ano"  "experiencia"   "emprego"      "cargo"        "salario_USD"
## [6] "pais_empreg"   "trab_remoto"   "pais_empresa"   "tam_empresa"
```

Após isso, utilizando função “names” foi possível conferir todas as colunas que compõem a tabela “dados” sendo listadas acima. Assim, as variáveis foram classificadas da seguinte maneira:

Ano: numérica discreta

Experiência: qualitativa ordinal

Emprego: qualitativa nominal

Cargo: qualitativa nominal

Salário: quantitativa contínua

País Emprego: qualitativa nominal

Trabalho remoto: qualitativa nominal

País Empresa: qualitativa nominal

Tamanho Empresa: qualitativa ordinal

```r
# 2) Recodifique a variável trab_remoto: 0=Não, 50=Parcial, 100=Sim


library(dplyr)

dados$trab_remoto = case_match(dados$trab_remoto, 0 ~ "Não", 50 ~ "Parcia
l", 100 ~ "Sim")
```

Utilizou-se a função “case_match” para converter os valores 0, 50 e 100 em “não”, “parcial” e “sim”, respectivamente.

```r
# 3) Faça uma tabela e um gráfico para a variável qualitativa experiencia
e tire conclusões


tabela_experiencia = table(dados$experiencia, useNA = "ifany")
tabela_media_experiencia = round(prop.table(tabela_experiencia)*100,1)
tabela_freq_media_experiencia = data.frame(tabela_experiencia,tabela_medi
a_experiencia)
tabela_freq_media_experiencia = tabela_freq_media_experiencia[,-3]
colnames(tabela_freq_media_experiencia) <- c("Experiência","Frequência","
Porcentagem")
```

Para criar uma tabela para a variável “experiência”, utilizou-se a função “table”, que gerou a tabela com as categorias de experiência e suas respectivas frequências. A partir disso, com a função “prop.table”, criou-se a coluna “porcentagem”, conforme visto na variável “tabela_media_experiencia”.

A tabela gerada pode ser vista a seguir.

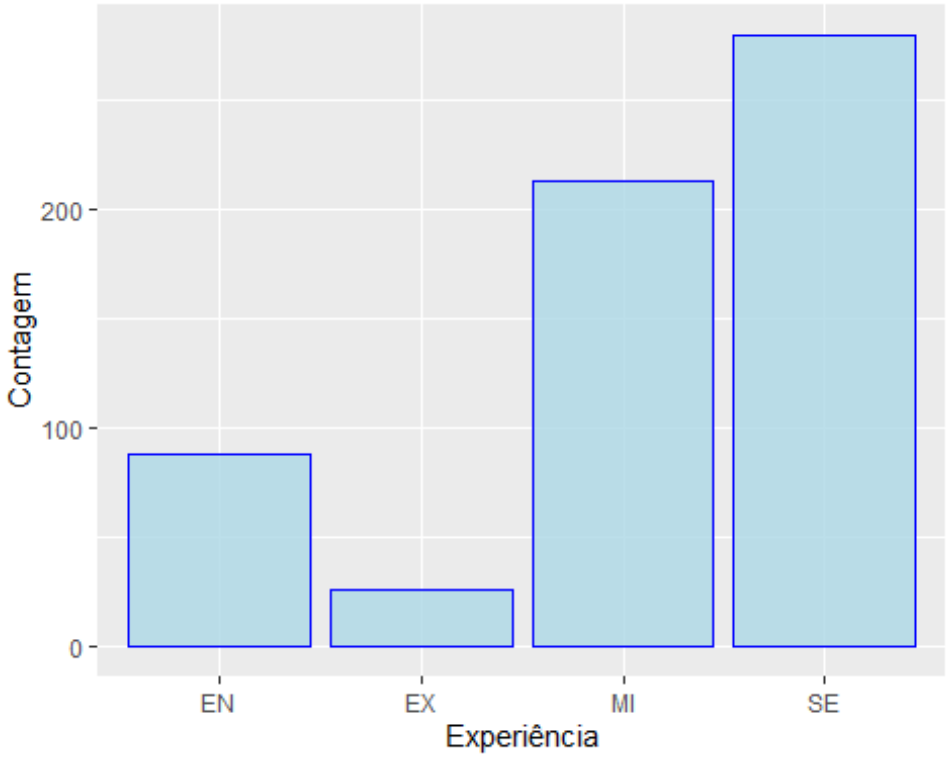
```r
tabela_freq_media_experiencia
```

```
##   Experiência Frequência Porcentagem
## 1          EN         88        14.5
## 2          EX         26         4.3
## 3          MI        213        35.1
## 4          SE        280        46.1
```

```r
write.table(tabela_freq_media_experiencia,"tabela_media_freq_experiencia.
csv", sep=";", dec=",", row.names=FALSE)
```

O gráfico gerado a partir dessas informações pode ser observado na imagem abaixo.

```
library(ggplot2)
ggplot(dados, aes(x=experiencia)) +
  geom_bar(fill="lightblue", color="blue", alpha=0.8) +
  labs(x="Experiência", y="Contagem")
```

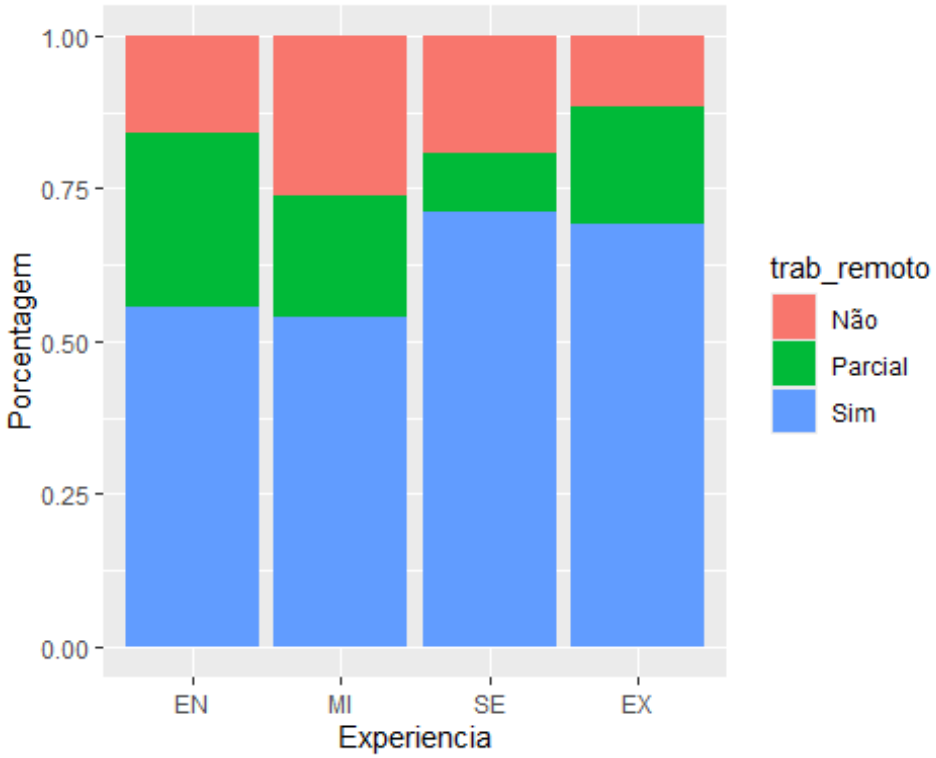



A partir dos valores observados na tabela "tabela_freq_media_experiencia", bem como no gráfico apresentado acima, pode-se afirmar que o grupo mais representativo da pesquisa é de empregados de nível sênior, com quase 50% das respostas. Em seguida tem-se nível médio, básico e, por fim, executivo.

Sabe-se que no mercado a quantidade de profissionais de nível básico é superior à quantidade de sêniors, portanto a pesquisa utilizada nesse estudo pode apresentar distorções em relação à realidade, principalmente no que diz respeito ao salário dos funcionários da área de dados. Isso será discutindo posteriormente neste relatório.

```r
# 4) Faça um gráfico para analisar a relação das duas variáveis qualitati
vas experiencia e trab_remoto e tire conclusões;


ggplot(dados, aes(x=experiencia, fill=trab_remoto)) +
  geom_bar(position="fill") +
  labs(x="Experiencia", y="Porcentagem") +
  scale_x_discrete(limits = c("EN", "MI", "SE", "EX"))
```




A partir do gráfico gerado observou-se que, para todos os níveis de experiência, o trabalho remoto é predominante, com mais de 50% de ocorrência. Há também uma tendência de maior adesão ao trabalho remoto para profissionais mais experientes.

Uma das hipóteses para esse comportamento é que profissionais de nível de experiência inferior precisam de treinamentos que, muitas vezes, ocorrem presencialmente na sede das empresas.

Além disso, profissionais com maior nível experiência possuem maior probabilidade de trabalharem para empresas de outros países de forma remota.

Para provar essas hipóteses, porém, seriam necessários estudos mais aprofundados e amostras mais representativas.

```r
# 5) Faça uma tabela relacionando as variáveis experiencia e salario_USD
e tire conclusões;


tabela_experiencia_salario <- aggregate(dados$salario_USD, by=list(dados$
experiencia), FUN="mean")
colnames(tabela_experiencia_salario) <- c("Salário","Média")
tabela_experiencia_salario_reorganizada = tabela_experiencia_salario[c(1,
3, 4, 2),]
tabela_experiencia_salario_reorganizada
```

```
##   Salário     Média
## 1      EN  61643.32
## 3      MI  87996.06
## 4      SE 138617.29
## 2      EX 199392.04
```

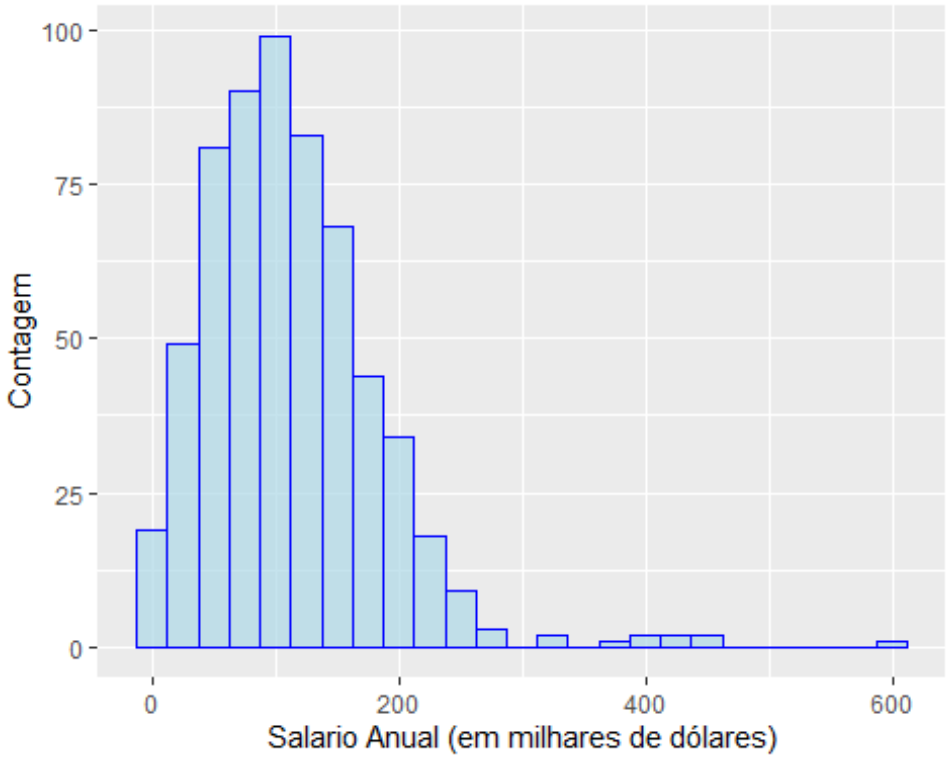
```r
write.table(tabela_experiencia_salario,"tabela_experiencia_salario.csv",
sep=";", dec=",", row.names=FALSE)
```

Gerou-se uma tabela relacionando o nível de experiência e a média salarial correspondente, utilizando a função “aggregate”.

É possível afirmar que cada aumento do nível experiência representa um ganho salarial médio considerável. Do nível básico para o nível médio há um aumento de 42,8%; do nível médio para o nível sênior há o maior salto: 57,5%; do nível sênior para o nível executivo o aumento foi de 43,8%.

```r
#6) Faça um gráfico para a variável salario_USD e tire conclusões.


ggplot(dados, aes(x=salario_USD / 1000)) +
geom_histogram(binwidth=25, alpha=0.7,  color="blue", fill="lightblue") +
labs(x="Salario Anual (em milhares de dólares)", y="Contagem")
```




A partir do histograma plotado, é possível observar a distribuição dos salários anuais dos profissionais que responderam à pesquisa. Cada barra possui uma largura equivalente a U$25.000.

A maior parte está na faixa até 200.000 dólares anuais, havendo maior concentração na barra em torno de 100.000 dólares, com quase 100 respondentes.

Os salários observados nesse gráfico podem ser considerados altos, uma vez que há maior participação de profissionais de nível sênior e médio em relação à quantidade de profissionais iniciantes esperada.