

analise_salarios.r

Matheus Silvano

2025-04-08

```
# analise_salarios.R

# Pacotes necessários
library(dplyr)

##
## Anexando pacote: 'dplyr'

## Os seguintes objetos são mascarados por 'package:stats':
##
##     filter, lag

## Os seguintes objetos são mascarados por 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)

# Leitura da base de dados
base <- read.csv("dados_salarios.csv", sep = ";")

# 1) Amostra
set.seed(19092005) # Dia, mês e ano de nascimento
base1 <- base[sample(nrow(base), 300), ]

# 2) Estatísticas descritivas da variável salario_USD
estatisticas_salario <- base1 %>%
  summarise(
    media = mean(salario_USD, na.rm = TRUE),
    mediana = median(salario_USD, na.rm = TRUE),
    percentil_5 = quantile(salario_USD, 0.05, na.rm = TRUE),
    percentil_25 = quantile(salario_USD, 0.25, na.rm = TRUE),
    percentil_75 = quantile(salario_USD, 0.75, na.rm = TRUE),
    percentil_95 = quantile(salario_USD, 0.95, na.rm = TRUE),
    minimo = min(salario_USD, na.rm = TRUE),
    maximo = max(salario_USD, na.rm = TRUE)
  )

print("Estatísticas descritivas de salario_USD:")
## [1] "Estatísticas descritivas de salario_USD:"
print(estatisticas_salario)
```

```
##      media mediana percentil_5 percentil_25 percentil_75
percentil_95 minimo
## 1 112015.2 1e+05 24324.9 62707.25 150018.8
220005.5 4000
##      maximo
## 1 600000
```

3) Média e mediana por ano

```
salario_por_ano <- base1 %>%
  group_by(ano) %>%
  summarise(
    media = mean(salario_USD, na.rm = TRUE),
    mediana = median(salario_USD, na.rm = TRUE)
  )
```

```
print("Média e mediana de salario_USD por ano:")
```

```
## [1] "Média e mediana de salario_USD por ano:"
```

```
print(salario_por_ano)
```

```
## # A tibble: 3 × 3
##   ano  media mediana
##   <int> <dbl> <dbl>
## 1  2020  87303.  74130
## 2  2021 102797.  85000
## 3  2022 122210. 116075
```

Conclusão:

Há um crescimento consistente tanto na média quanto na mediana ao longo dos anos, indicando valorização salarial.

A mediana é mais apropriada para essa análise porque é menos sensível a valores extremos que são comuns em salários. A diferença entre média e mediana em 2022 (122.210 vs 116.075) sugere presença de valores elevados que puxam a média para cima.

4) Análise por experiência (agrupando SE + EX)

```
base1 <- base1 %>%
  mutate(experiencia_mod = ifelse(experiencia %in% c("SE", "EX"),
    "SE_EX", experiencia))
```

```
salario_por_experiencia <- base1 %>%
  group_by(experiencia_mod) %>%
  summarise(
    media = mean(salario_USD, na.rm = TRUE),
    mediana = median(salario_USD, na.rm = TRUE)
  )
```

```
print("Média e mediana de salario_USD por experiência (com SE + EX  
juntos):")
```

```
## [1] "Média e mediana de salario_USD por experiência (com SE + EX  
juntos):"
```

```
print(salario_por_experiencia)
```

```
## # A tibble: 3 × 3  
##   experiencia_mod   media mediana  
##   <chr>          <dbl>   <dbl>  
## 1 EN             67409.   61363  
## 2 MI             82202    73565  
## 3 SE_EX          145970.  140000
```

Existe um aumento significativo nos salários conforme a experiência, especialmente no grupo SE/EX.

A mediana é novamente mais apropriada porque a distribuição salarial, especialmente no grupo SE/EX, pode conter salários muito altos que distorcem a média.

5) Gráfico com as medianas por experiência

```
ggplot(salario_por_experiencia, aes(x = experiencia_mod, y = mediana))  
+  
  geom_col(fill = "steelblue") +  
  labs(  
    title = "Mediana de salário por nível de experiência",  
    x = "Nível de Experiência",  
    y = "Mediana do Salário (USD)"  
  ) +  
  theme_minimal()
```

Mediana de salário por nível de experiência

