CAIO RODRIGUES GOMES
VITOR CAETANO DA SILVA
RODRIGO DORNELES FERREIRA DE SOUZA

I. IDENTIFICAÇÃO

O nome do dataset é Data Science Job Salaries, retirado do Kaggle, na qual a ai-jobs.net Salaries foi responsável por agregar todos os dados contidos nele. A última vez em que o dataset foi atualizado foi em julho de 2022. O dataset possui 12 colunas e 607 linhas.

II. VARIÁVEIS

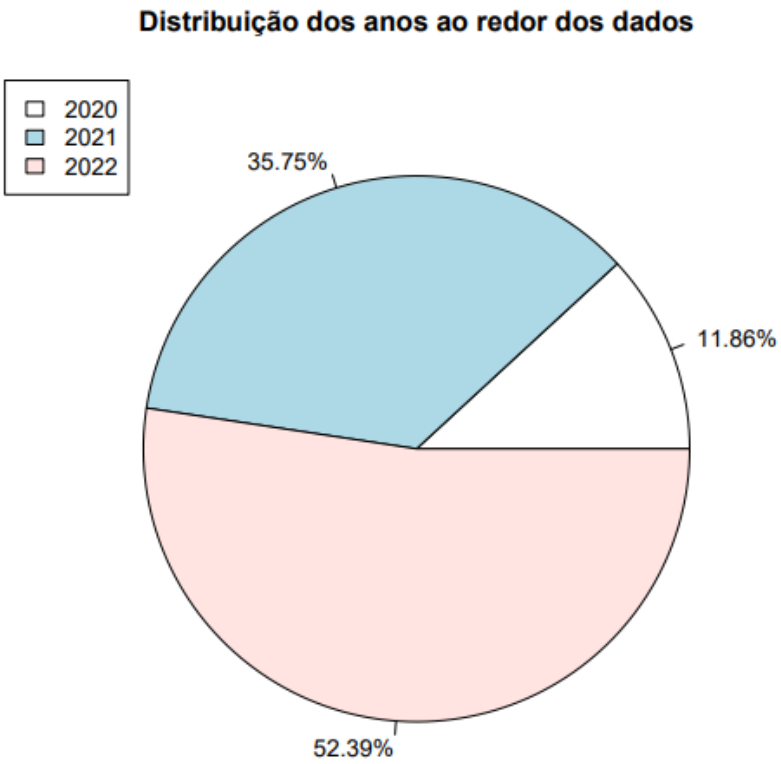
Neste dataset, estão contidas variáveis qualitativas e quantitativas, sendo elas divididas da seguinte forma:

- Qualitativa nominal: tipo de trabalho (meio período, tempo integral e freelancer); título do trabalho; moeda do trabalho; residência do empregado; localização da companhia que contratou;
- Qualitativa ordinal: ano de trabalho; nível de experiência do empregado (iniciante, júnior, sênior, nível executivo); tamanho da companhia (pequeno, médio e grande) com base na quantidade de número de empregados
- Quantitativa contínua: salário; salário em dólares; taxa de trabalho feito de modo remoto;

III. OBSERVAÇÕES, CASOS OU INSTÂNCIAS

As observações correspondem a pessoas, na qual para cada pessoa indica o seu salário, o seu cargo na área de ciência de dados, localização da empresa pela qual foi contratado, etc.

IV. ESTATÍSTICA DESCRITIVA



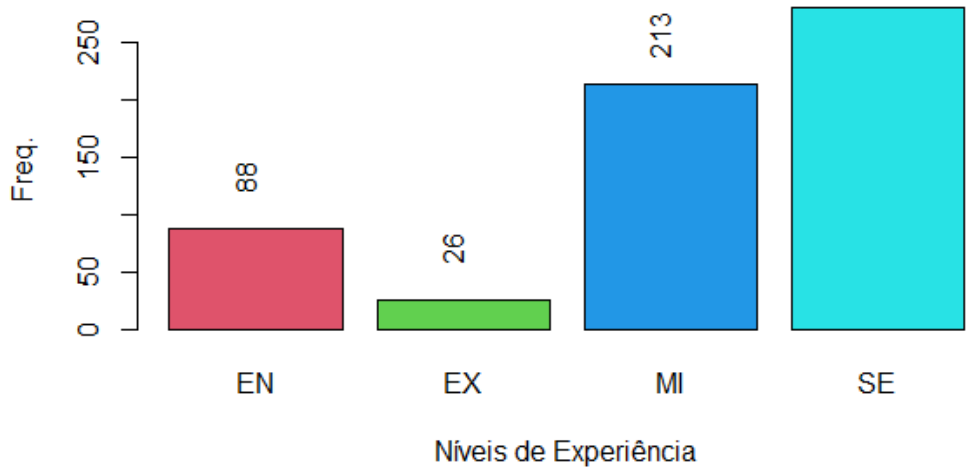
DISTRIBUIÇÃO DE FREQUÊNCIA DE NÍVEL DE EXPERIÊNCIA DOS EMPREGADOS

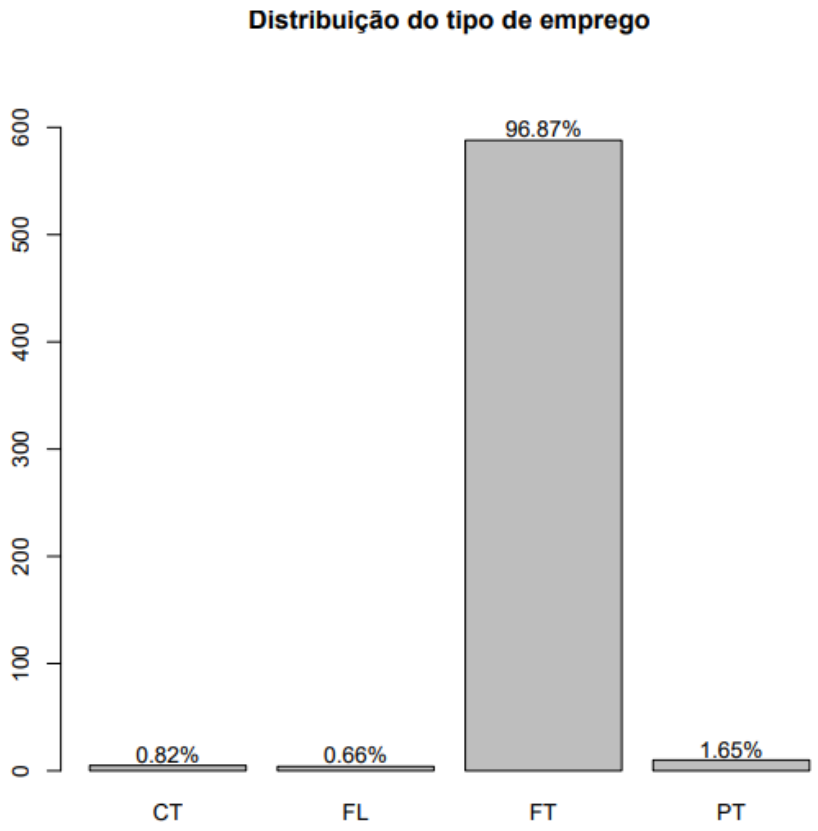
EN, se refere a **Entry-level / Júnior**

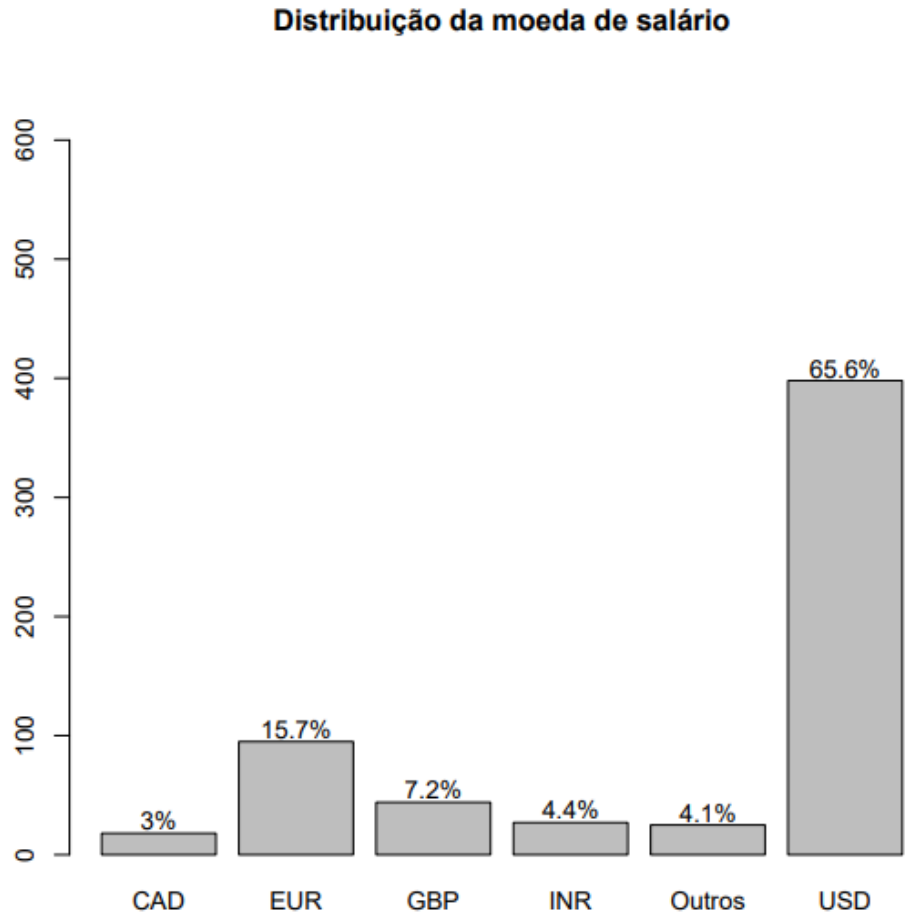
MI, se refere a **Mid-level / Intermediário**

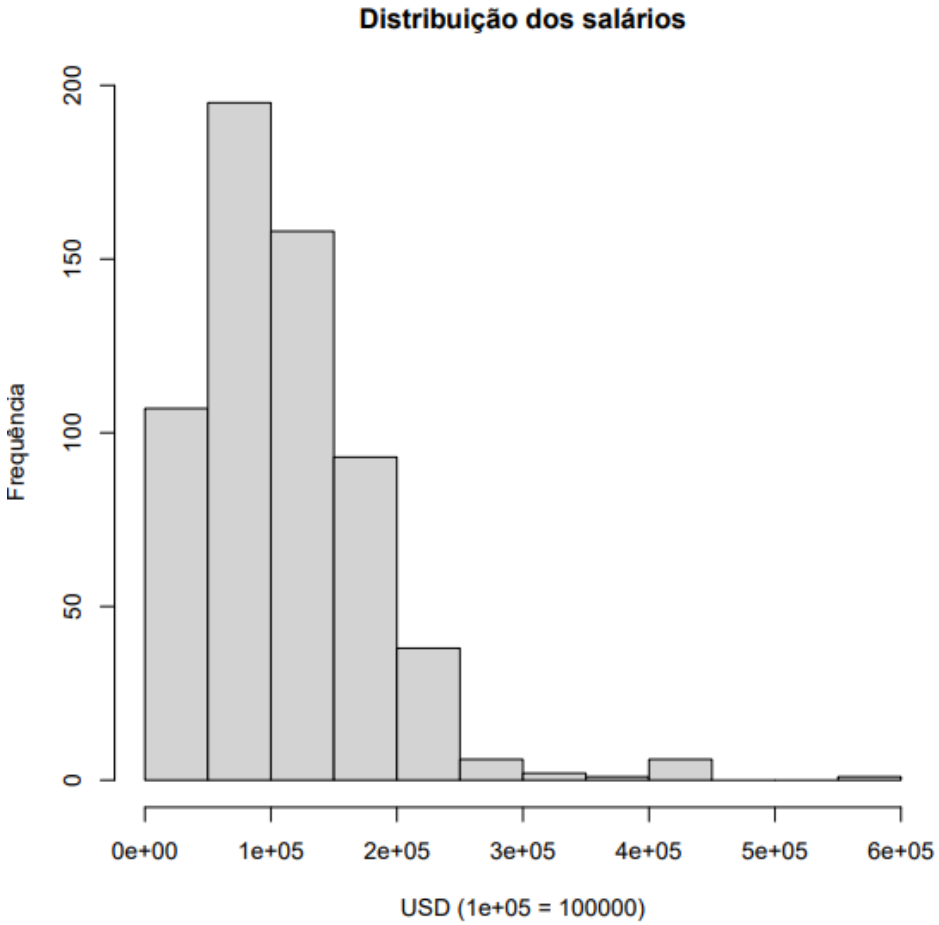
SE, se refere a **Senior-level / Expert**

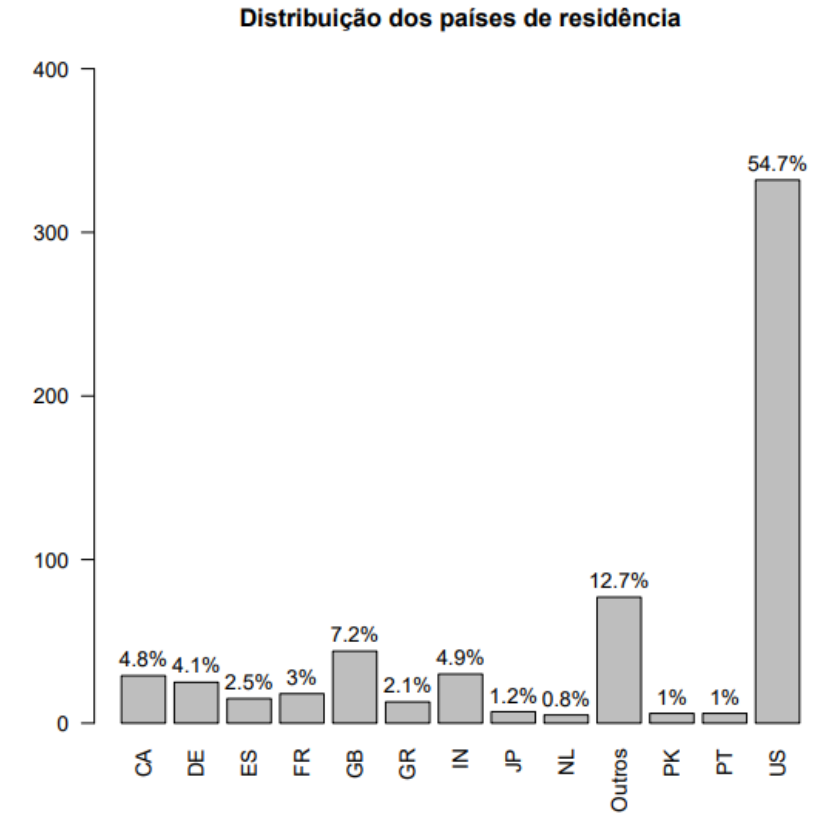
EX, se refere a **Executive-level / Diretor**

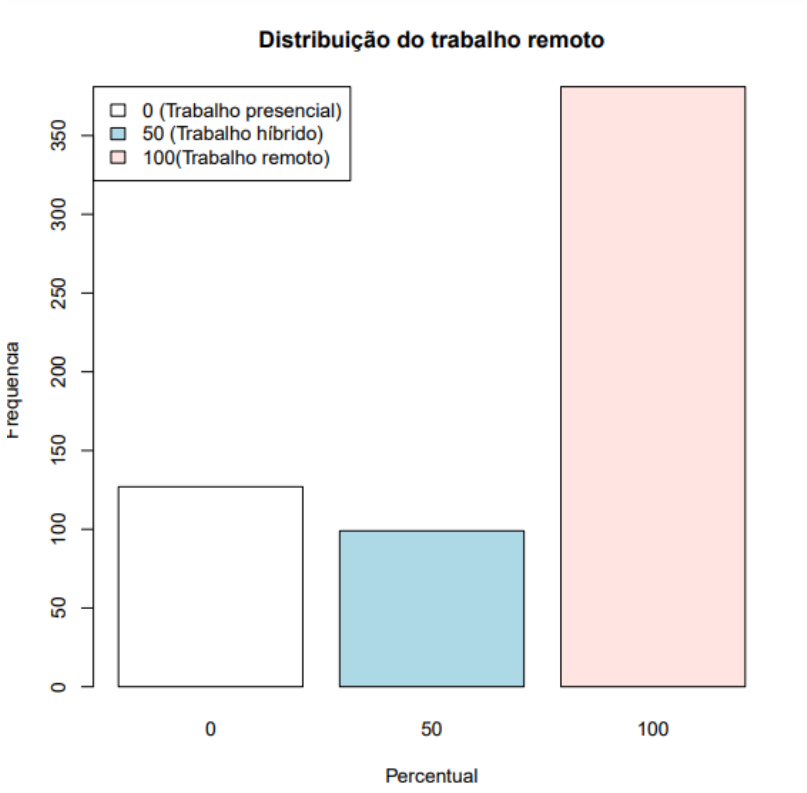


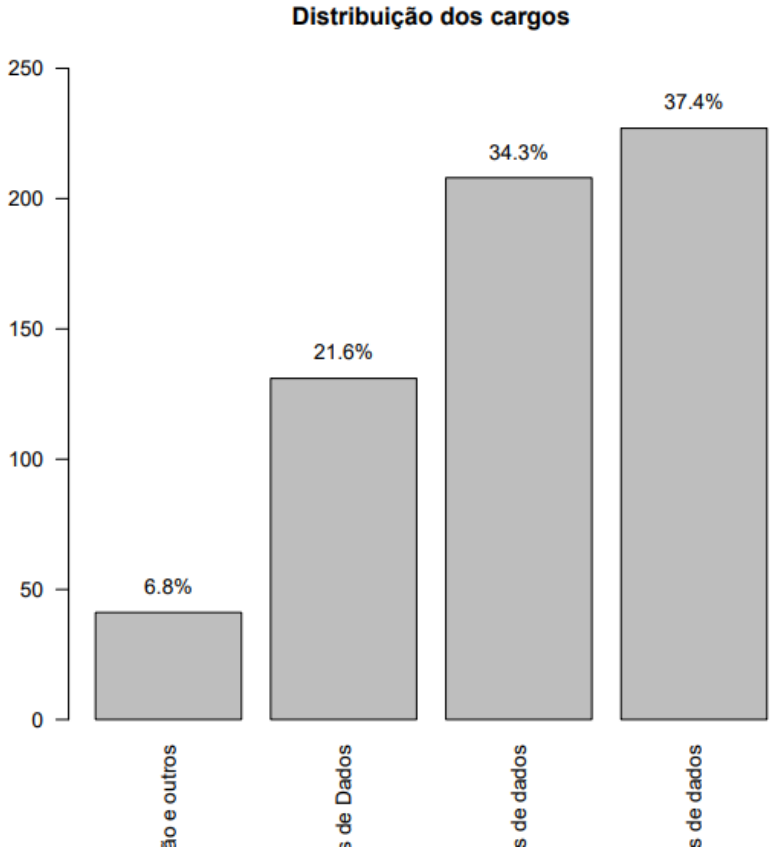


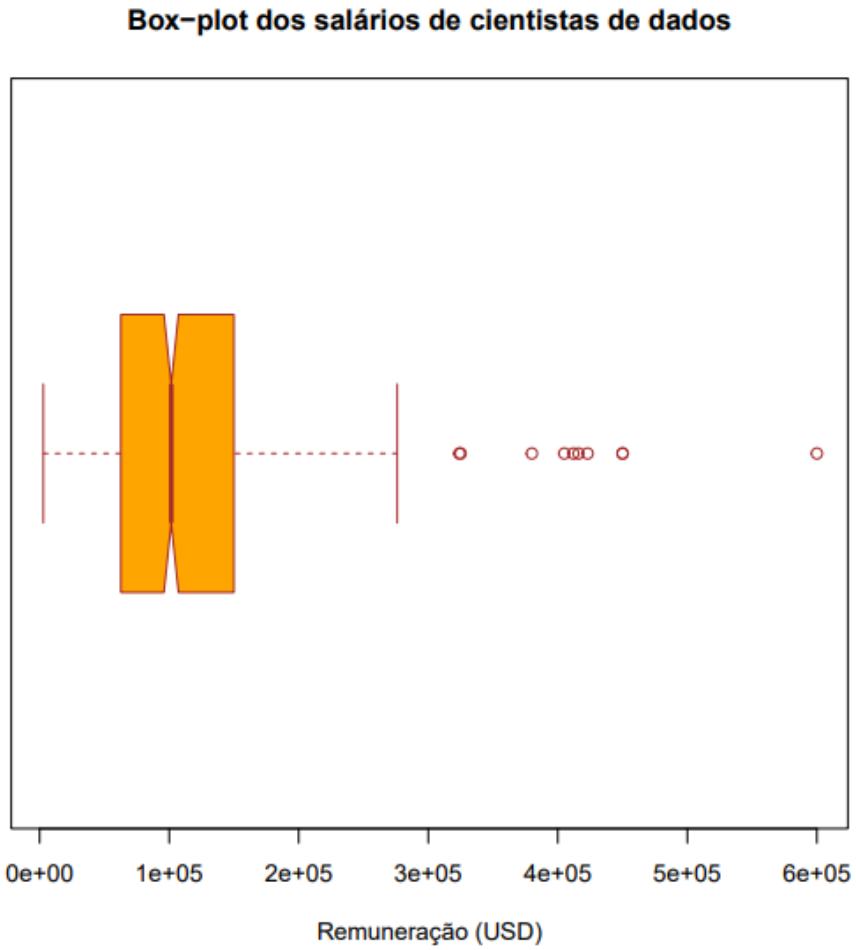

Distribuição da moeda de salário
600
500
400
300
200
100
0
3%
15.7%
7.2%
4.4%
4.1%
65.6%
CAD
EUR
GBP
INR
Outros
USD


Distribuição dos salários
Frequência
0
50
100
150
200
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
USD (1e+05 = 100000)

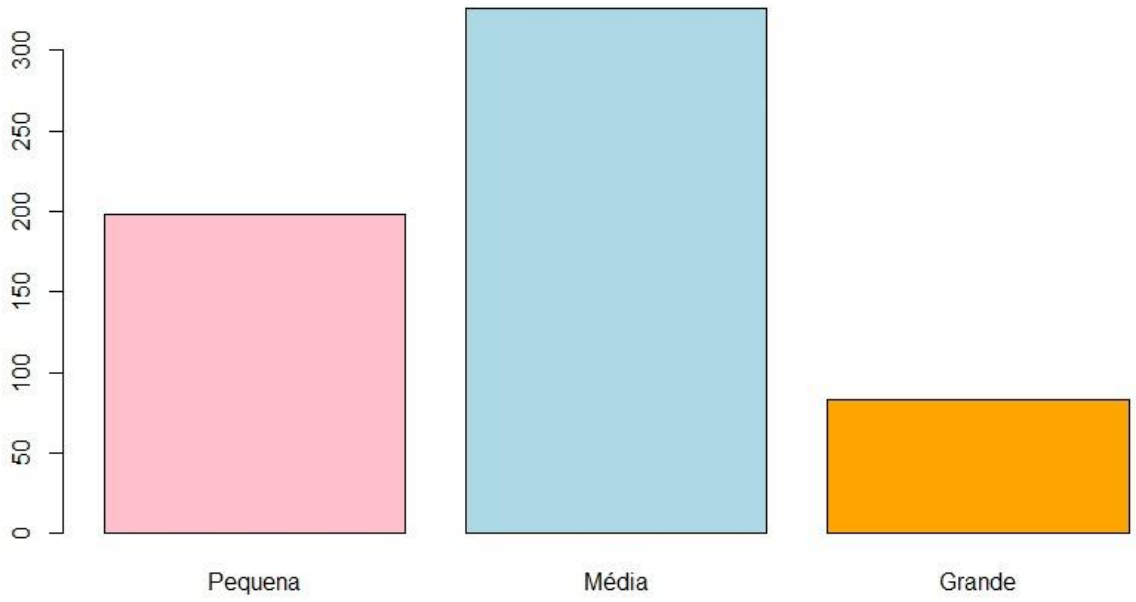

Distribuição dos países de residência
400
300
200
100
0
4.8%
4.1%
2.5%
3%
7.2%
2.1%
4.9%
1.2%
0.8%
12.7%
1%
1%
54.7%
CA
DE
ES
FR
GB
GR
IN
JP
NL
Outros
PK
PT
US


Distribuição do trabalho remoto
0 (Trabalho presencial)
50 (Trabalho híbrido)
100(Trabalho remoto)
Frequencia
0
50
100
150
200
250
300
350
0
50
100
Percentual


Distribuição dos cargos
250
200
150
100
50
0
6.8%
21.6%
34.3%
37.4%
ão e outros
s de Dados
s de dados
s de dados



DISTRIBUIÇÃO DE FREQUÊNCIA DE TAMANHO DAS EMPRESAS



SALÁRIO EM DÓLARES:

- Média: 112297.9
- Mediana: 101570
- Moda: 100000
- Desvio padrão: 70957.26
- Quartis: 2859 (0%), 62727 (25%), 101570 (50%), 15000 (75%), 600000 (100%)

TAXA EM PORCENTAGEM DE TRABALHO REMOTO:

- Média: 70.92
- Mediana: 100
- Moda: 100
- Desvio padrão: 40.70
- Quartis: 0 (0%), 50 (25%), 100 (50%), 100 (75%), 100 (100%)

Código usando a linguagem R para a construção dos gráficos:

```
mydata <- read.table("./ds_salaries.csv", header=TRUE,
  sep=",")

names(mydata)

#função de moda
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
#contagem das vagas de trabalho
table(mydata$job_title)

#contagem da experiencia de trabalho
table(mydata$experience_level)

#-----------------------------------------------------------------------------------------------------------
#média dos salários
mean(mydata$salary_in_usd)

#mediana dos salários
median(mydata$salary_in_usd)

#moda dos salários
result <- getmode(mydata$salary_in_usd)
print(result)
#desvio padrão dos salários
sd(mydata$salary_in_usd)

#percentis dos salários
```

```
quantile(mydata$salary_in_usd)

print('---------------------------------')
#-----------------------------------------------------------------------------------------------------------

#média do quão remoto a carreira trabalha
mean(mydata$remote_ratio)

#mediana do quão remoto a carreira trabalha
median(mydata$remote_ratio)

# Calculate the mode using the user function.
result <- getmode(mydata$remote_ratio)
print(result)

#desvio padrão do remote_Ratio
sd(mydata$remote_ratio)

#percentis do remote_ratio
quantile(mydata$remote_ratio)

#-----------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------
#grafico de pizza da distribuição por ano
dados1 <- table(mydata$work_year)
div1 <- sum(dados1)
pie(dados1, labels=paste0(round(dados1/div1*100,2), "%"), main="Distribuição dos anos ao
redor dos dados")

legend("topleft", legend = c("2020", "2021", "2022"),
       fill =  c("white", "lightblue", "mistyrose"))
#-----------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------
#gráfico de barras do tipo de emprego

dados2 <- table(mydata$employment_type)

b1 <- barplot(dados2, ylim=c(0, 650),main = "Distribuição do tipo de emprego")
text(x=b1, y=dados2 + 10, labels=paste0(round(proportions(dados2), 4)*100, "%"))
#-----------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------
#gráfico de barras da moeda do salário
dados3 <- table(mydata$salary_currency)

val_repl <- c("AUD","BRL","CHF","CLP", "CNY", "DKK", "HUF", "JPY", "MXN","PLN",
"SGD","TRY")

dados3new <- sapply(mydata$salary_currency, function(x) replace(x, x %in% val_repl,
"Outros"))
dados3 <- table(dados3new)
```

```
b2 <- barplot(dados3, ylim=c(0, 650),main = "Distribuição da moeda de salário")
text(x=b2, y=dados3 + 10, labels=paste0(round(proportions(dados3), 3)*100, "%"))
#-----------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------
#plot dos salários

hist(mydata$salary_in_usd, main = "Distribuição dos salários", xlab="USD (1e+05 =
100000)", ylab = "Frequência")

#-----------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------
#plot dos países de residência

val_repl2 <- c("AE","AR","AT","AU", "BE", "BG", "BO", "BR", "CH","CL", "CN","CO", "CZ",
"DK", "DZ", "EE", "HK", "HN", "HR", "HU", "IE", "IQ", "IR", "IT", "JE", "KE", "LU", "MD", "MT",
"MX", "MY", "NG", "NZ", "PH", "PL", "PR", "RO", "RS", "RU", "SG", "SI", "TN", "TR", "UA",
"VN" )

dados4new <- sapply(mydata$employee_residence, function(x) replace(x, x %in% val_repl2,
"Outros"))
dados4 <- table(dados4new)

b3 <- barplot(dados4, ylim=c(0, 400),main = "Distribuição dos países de residência", las=2)
text(x=b3, y=dados4 + 10, labels=paste0(round(proportions(dados4), 3)*100, "%"))
#-----------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------
#plot do remote ratio

barplot(table(mydata$remote_ratio), main = "Distribuição do trabalho remoto",
xlab="Percentual", ylab = "Frequência", col =  c("white", "lightblue", "mistyrose"))

legend("topleft", legend = c("0 (Trabalho presencial)", "50 (Trabalho híbrido)", "100(Trabalho
remoto)"),
       fill =  c("white", "lightblue", "mistyrose"))
#-----------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------
#plot dos cargos de trabalho

dados5 <- table(mydata$job_title)

analistas <- c("Product Data Analyst",  "Principal Data Analyst", "Marketing Data Analyst",
"Lead Data Analyst", "Finance Data Analyst", "Financial Data Analyst", "Data Analytics
Manager", "Data Analytics Lead", "Data Analytics Engineer", "Data Analyst", "Business Data
Analyst", "BI Data Analyst")

cientistas <- c("3D Computer Vision Researcher", "AI Scientist", "Applied Data Scientist",
"Applied Machine Learning Scientist", "Data Science Consultant", "Data Science Engineer",
```

```
"Data Scientist", "Lead Data Scientist", "Machine Learning Developer", "Machine Learning
Scientist", "Principal Data Scientist", "Research Scientist", "Staff Data Scientist")

engenheiros <- c("Principal Data Engineer", "NLP Engineer", "ML Engineer", "Machine
Learning Infrastructure Engineer", "Machine Learning Engineer", "Lead Machine Learning
Engineer", "Lead Data Engineer", "Data Engineer", "Data Architect", "Data Analytics
Engineer", "Computer Vision Software Engineer", "Computer Vision Engineer", "Cloud Data
Engineer", "Big Data Engineer", "Analytics Engineer")

admeoutros <- c("Data Analytics Manager", "Data Engineering Manager", "Data Science
Manager", "Director of Data Engineering", "Director of Data Science", "ETL Developer", "Head
of Data", "Head of Data Science", "Head of Machine Learning", "Machine Learning Manager",
"Data Specialist", "Big Data Architect")

dados5new <- sapply(mydata$job_title, function(x) replace(x, x %in% analistas, "Analistas
de Dados"))
dados5new <- sapply(dados5new, function(x) replace(x, x %in% cientistas, "Cientistas de
dados"))
dados5new <- sapply(dados5new, function(x) replace(x, x %in% engenheiros, "Engenheiros
de dados"))
dados5new <- sapply(dados5new, function(x) replace(x, x %in% admeoutros,
"Administração e outros"))
dados5 <- table(dados5new)

b4 <- barplot(dados5, ylim=c(0, 250),main = "Distribuição dos cargos", las = 2)
text(x=b4, y=dados5 + 10, labels=paste0(round(proportions(dados5), 3)*100, "%"))
#-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------

#boxplot dos salários
boxplot(mydata$salary_in_usd, main = "Box-plot dos salários de cientistas de dados",
xlab = "Remuneração (USD)",
ylim=c(2000,600000),
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE)
```

V. TIPO DE PESQUISA

Com esse dataset podemos descobrir a diferença de salários que ocorre entre os países nas profissões relacionadas a ciência de dados. Além disso, com base nas variáveis do dataset, como a localização da empresa, o nível de experiência do empregado e o cargo ocupado por ele podemos fazer um estudo para inferir a tendência salarial daquela pessoa.