GRUPO 1

- ★ CAIO RODRIGUES GOMES *11208012*
- ★ RODRIGO DORNELES FERREIRA DE SOUZA *11295831*
- ★ VITOR CAETANO DA SILVA *9276999*

ROTEIRO DE ATIVIDADES 2

I.Identificação do dataset escolhido

- Quem fez o levantamento dos dados (empresa, instituição de pesquisa, universidade, endereço)?
    - Todas as informações contidas neste dataset foram produzidas a partir do ai-jobs.net Salaries: https://salaries.ai-jobs.net/
- Quando o levantamento foi feito?
    - 3 meses atrás
- Quantas linhas e quantas colunas tem o dataset?
    - 607 linhas e 12 colunas

II.Procedimentos de amostragem

Amostra total, não foi realizado extração de amostra para utilização dos dados, utilizaremos o dataset original.

III.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; tamanho da companhia (pequeno, médio e grande) com base na quantidade de número de empregados; nível de experiência do empregado (iniciante, júnior, sênior, nível executivo);

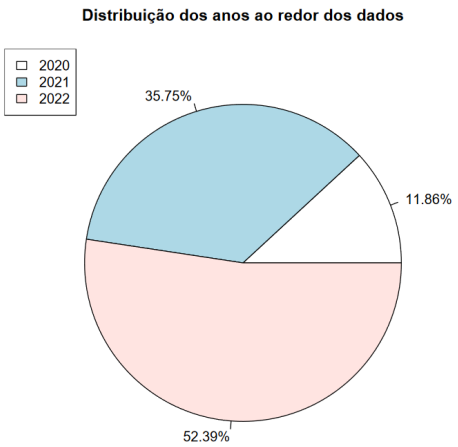
- Qualitativa ordinal discreta: ano de trabalho;
- Quantitativa contínua: salário; salário em dólares; taxa de trabalho feito de modo remoto;

IV.Observações, casos ou instâncias

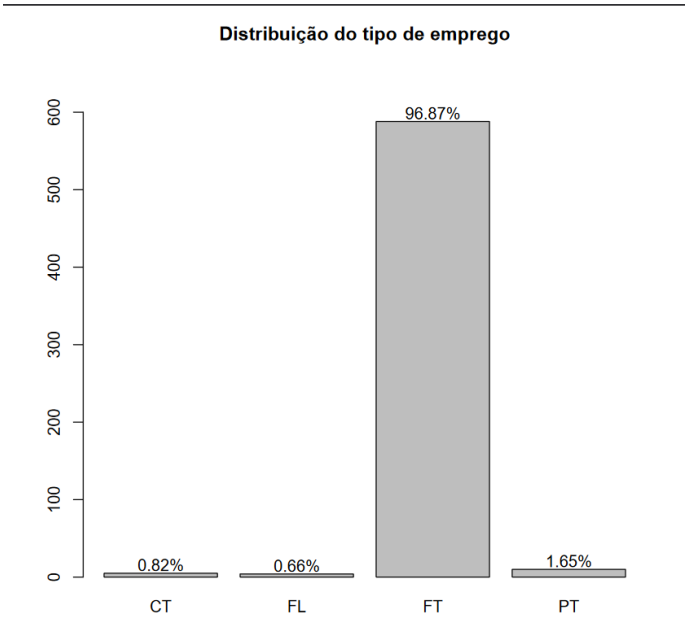
- Com esse dataset, é possível verificar os salários de profissionais que trabalham com Ciência de Dados, tendo informações ainda sobre o nível de experiência, a localização da empresa e do empregado, o tamanho da empresa, além também do ano em que aquele salário foi pago.

V. Estatística descritiva (EM DESENVOLVIMENTO)

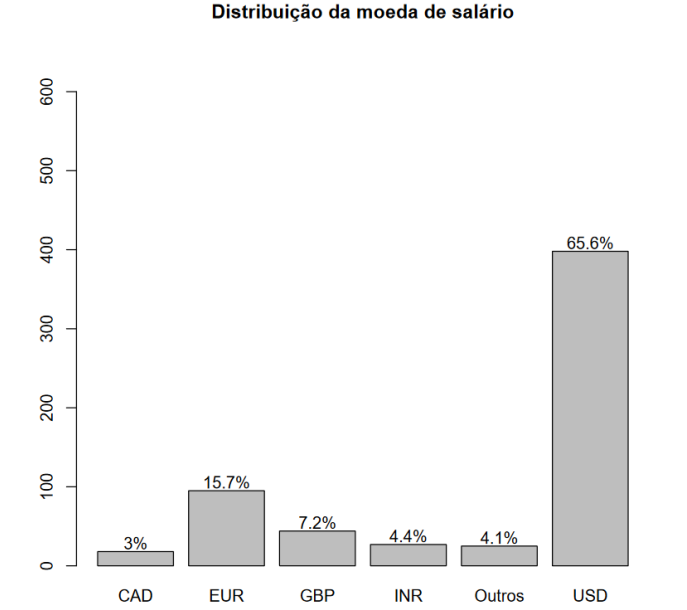
- **Todas** as variáveis deverão ser analisadas no R e apresentadas sob a forma de:



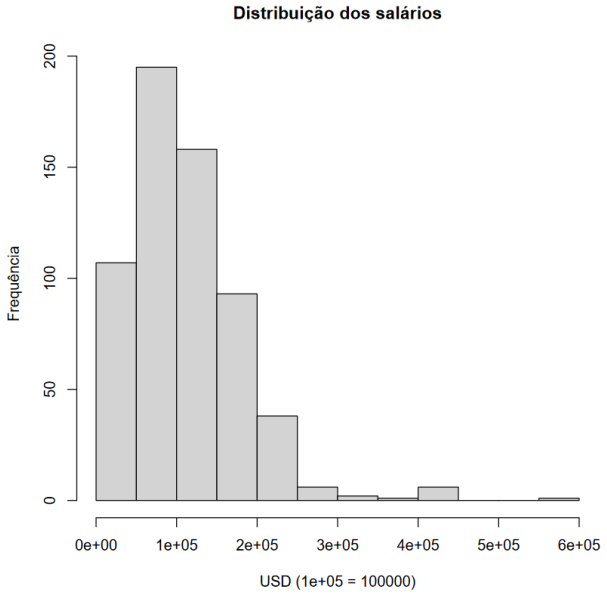

```
#grafico de pizza da distribuição por ano
dados1 <- table(mydata$work_year)
div1 <- sum(dados1)
pie(dados1, labels=paste0(round(dados1/div1*100,2), "%"), main="Distribuição dos anos ao redor
dos dados")

legend("topleft", legend = c("2020", "2021", "2022"),
       fill =  c("white", "lightblue", "mistyrose"))
#-------------------------------------------------------------------------------------------
```



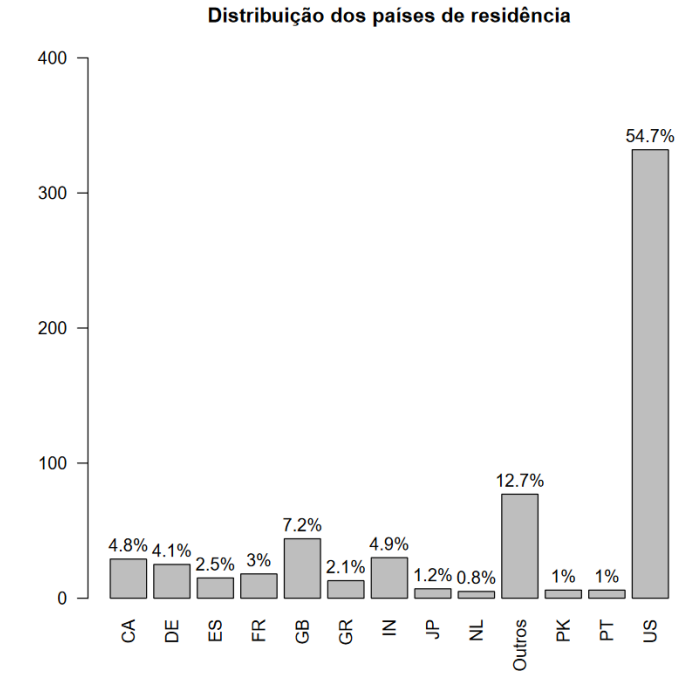

```
#-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------
#gráfico de barras do tipo de emprego

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, "%"))
#-------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------
```



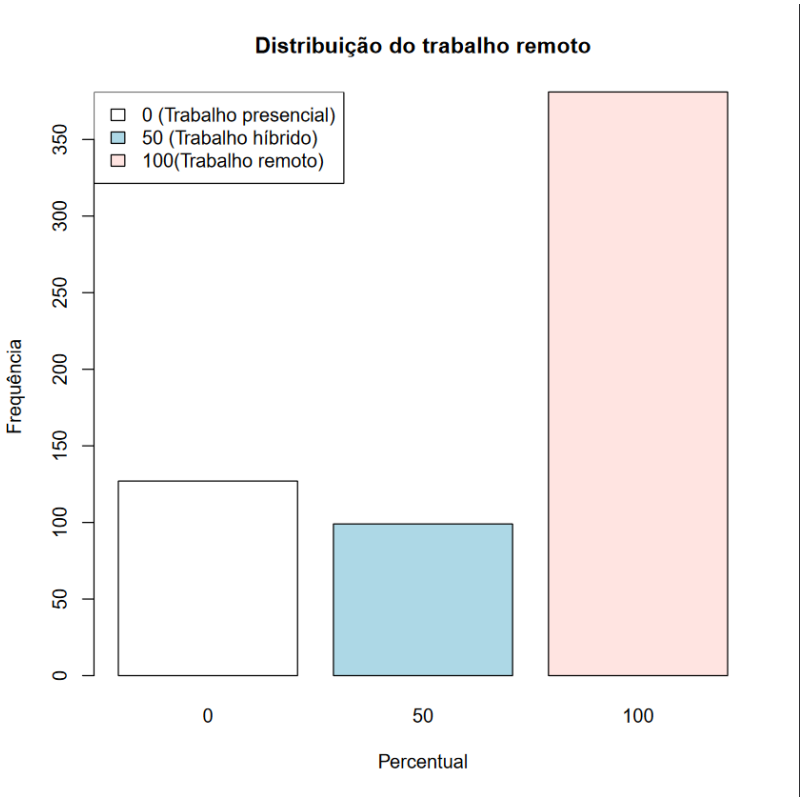
```
#-------------------------------------------------------------------------------
#gráfico de barras da moeda do salário
dados3 <- table(mydata$salary_currency)

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, "%"))
```



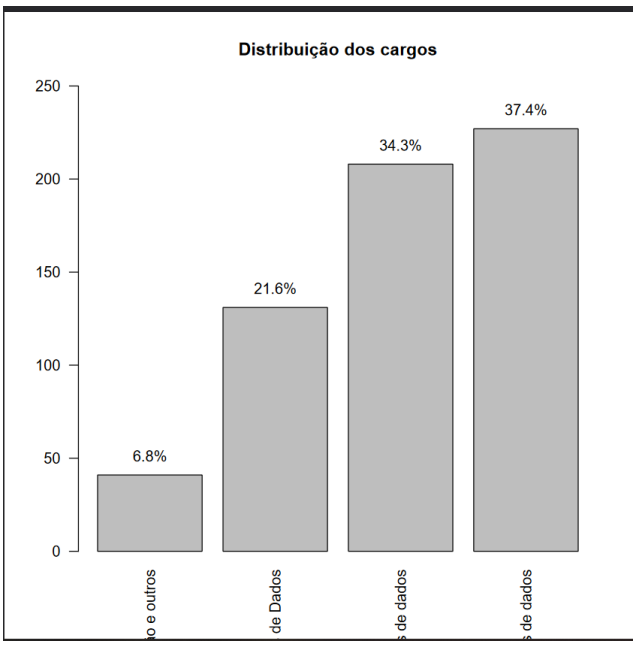

```
#-------------------------------------------------------------------------------
#plot dos salários

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

#-------------------------------------------------------------------------------
```




```r
#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 do trabalho
```




```r
#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, "%"))
```

Como medidas, temos para as variáveis:

Salário:

- Média: 112297.9
- Mediana: 101570
- Moda: 100000
- Desvio Padrão: 70957.26
- Q1:62726 , Q2: 101570 , Q3:150000

VI.Que tipo de pesquisa/pergunta você pretende fazer com este dataset? (EM DESENVOLVIMENTO)

Salarios de desenvolvedores de ciência de dados, quais são e como estão distribuidos