DESAFIO DIA 05 - FÁBRICA DE SOFTWARE

```
setwd("C:\\Users\\elain\\OneDrive\\Documents\\UNIPE\\P2\\fabrica\\R")
df = read.csv("ds_salaries.csv", sep=",", encoding = "UTF-8")
```

Inicialmente vamos setar a pasta de origem e o arquivo de dataframe a ser utilizado.

```
View(df)
```

Visualização do dataframe. Este dataframe se chama “Data Science Job Salaries”.

```
freq_abs = table(df$salary_in_usd)
View(freq_abs)
```

Criação e visualização da tabela de frequência absoluta, apenas com a coluna “salary_in_usd”. A frequência absoluta irá contabilizar quantas vezes determinado salário aparece no dataframe, na coluna “salary_in_usd”. Foi possível notar que o salário que mais aparece é de U$100000, com 15 ocorrências. Em seguida, os salários de U$120000 e U$150000 apareceram 12 vezes, cada um.

```
freq_rel = prop.table(freq_abs)
View(freq_rel)
```

Criação e visualização da tabela de frequência relativa, referente à tabela de frequência absoluta. Essa tabela apresenta quanto cada ocorrência representa em relação a todas as ocorrências.

```
p_freq_rel = 100* prop.table(freq_rel)
View(p_freq_rel)
```

Criação e visualização da tabela de frequência relativa percentual. Essa tabela apresenta quanto cada ocorrência representa em relação a todas as ocorrências, em porcentagem.

```
freq_abs = c(freq_abs, sum(freq_abs))
View(freq_abs)
names(freq_abs)[[370]] = "Total"
View(freq_abs)
```

Atualização da tabela frequência absoluta, com a inclusão dos valores cumulativos. Aqui, foi possível notar que o dataframe tem 370 valores distintos de salário, onde foi colocado o name “Total”.

```
freq_rel = c(freq_rel, sum(freq_rel))
p_freq_rel = c(p_freq_rel, sum(p_freq_rel))
```

Adicionando as somas das frequências relativas e percentuais.

```
tabela_final = cbind(freq_abs,
                     freq_rel = round(freq_rel, digits = 5),
                     p_freq_rel = round(p_freq_rel, digits = 2))
View(tabela_final)
```

Criação e visualização da tabela final, com a junção das tabelas de frequência absoluta, frequência relativa e porcentagem de frequência relativa.

```
k = 1+3.3*log(607)
A = (600000 - 2859)/23
intervalo_de_classes = seq(2859, 600000, 25962.65)
```
Através da análise das observações encontradas no dataframe, foi possível identificar os valores necessários para calcular o intervalo de classes (mínimo valor, máximo valor e total de observações). Com isso, foi identificado o valor do intervalo das classes.

```
tabela_de_classes = table(cut(df$salary_in_usd, breaks = intervalo_de_classes, right=FALSE))
View(tabela_de_classes)
```
Por fim, construímos nossa tabela de classes, mostrando as classes e suas frequências.

```
summary(df)
```
O summary mostra informações importantes sobre o dataframe, como mínimos e máximos valores, quartis, mediana e outras informações.

```
hist(df$salary_in_usd, col="purple")
```
Para uma melhor visualização, utilizei o comando acima para criar um histograma. Através deste histograma é possível identificar o tipo de distribuição dos dados salary_in_usd através da observação da sua distribuição de frequências.

```
barplot(tabela_de_classes, col = "purple")
```
Ainda, construi um gráfico de barras, que se assemelha com o histograma, sendo que desta vez utilizando a tabela de classes, para representar os dados que estão categorizados.

```
df1 = df
excluir1 = c("salary", "salary_currency")
novo_df1 = df1[ , !(names(df1) %in% excluir1)]
View(novo_df1)
```
Como eu desejava trabalhar apenas com os dados de salário em dólar, criei um novo dataframe para fazer as alterações que queria (df1). Assim, exclui as colunas “salary” e “salary_currency”, pois a primeira dava o valor na moeda corrente e a segunda dizia que moeda era esta.
Visualizei o novo dataframe e identifiquei que as colunas realmente não estavam mais presentes.

```
write.table(novo_df1, file = "salaries_in_usd.csv", sep = ",")
```
Realizei este comando para salvar o novo dataframe em formato csv.