Autor: Gustavo Coimbra de Souza Teixeira

Relatório - Desafio 03

Neste Relatório abordarei a estratégia usada para ler um banco de dados no formato **PARQUET** e outro no formato **JSON**.

PARQUET

Para ler o formato PARQUET no R-studio, usamos do pacote “arrow” que nos permite ler este formato, e também usei o “tidyverse”, para isso usamos os comandos:

```
#install.packages("arrow")
library(arrow)
```

Feito isso, baixamos o banco de dados que iremos utilizar.

O banco de dados baixamos do kaggle pelo link “https://www.kaggle.com/datasets/aadyasingh55/twitter-emotion-classification-dataset” que se chamava “train-00000-of-00001.parquet” que consiste em frases/textos que as pessoas colocam no twitter e a emoção que a frase que elas escreveram que revela o'que elas estavam sentindo (sendo as opções #0: sadness, #1: joy, #2: love, #3: anger, #4: fear, #5: surprise.

Agora que baixamos nosso banco de dados, eu o li e o nomeei, e fiz isso usando caminhos relativos:

```
#Lendo o banco de dados em parquet
d_parquet = read_parquet("train-00000-of-00001.parquet")
```

A tabela vinda, estava classificada com os números das emoções, então adicionei uma coluna com a legenda mostrando cada uma, usando o seguinte comando:

```
#colocando a legenda no banco de dados
d_parquet_leg = d_parquet %>%
  mutate(emotion = recode(label,
                          `0` = "sadness",
                          `1` = "joy",
                          `2` = "love",
                          `3` = "anger",
                          `4` = "fear",
                          `5` = "surprise"))
```

Feito, usei o “head()” que nos mostra uma prévia da tabela para usar neste relatório

```
head(d_parquet_leg)
```

e tivemos uma amostra da tabela como saída.

```
# A tibble: 6 × 3
  text                                                          label emotion
  <chr>                                                         <int> <chr>
1 i feel awful about it too because it s my job to get him in a p…     0 sadness
2 im alone i feel awful                                           0 sadness
3 ive probably mentioned this before but i really do feel proud o…     1 joy
4 i was feeling a little low few days back                        0 sadness
5 i beleive that i am much more sensitive to other peoples feelin…     2 love
6 i find myself frustrated with christians because i feel that th…     2 love
```

JSON

Para ler o formato JSON no R-studio, usamos do pacote “” que nos permite ler este formato usando os seguintes comandos

```
install.packages("jsonlite")
library(jsonlite)
```

Feito isso, baixamos o banco de dados que iremos utilizar.

O banco de dados baixamos do kaggle pelo link “https://www.kaggle.com/datasets/pratyushpuri/heart-disease-dataset-3k-rows-python-code-2025” que se chamava “heart_disease_dataset.json”, que consistia em dados medicos de diversos pacientes, que buscava ver quais tinham tido ataque cardiaco ou não pra estabelecer relações. Porém este banco, ao contrário do outro estava zipado, então antes tivemos que extrair o arquivo do zip usando o seguinte comando:

```
unzip("jonson.zip")
```

*nota: jonson.zip foi como renomeei o arquivo zipado assim que eu o baixei

E afeito isso, nós o lemos e o renomeamos:

```
d_json = fromJSON("heart_disease_dataset.json")
```

Feito, usei o “head" que nos mostra uma prévia da tabela para usar neste relatório que ficou na próxima página.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
1	67	1	2	111	536	0	2	88	0	1.3	3	2	3
2	57	1	3	109	107	0	2	119	0	5.4	2	0	3
3	43	1	4	171	508	0	1	113	0	3.7	3	0	7
4	71	0	4	90	523	0	2	152	0	4.7	2	1	3
5	36	1	2	119	131	0	2	128	0	5.9	3	1	3
6	49	1	1	186	571	0	0	176	0	4.0	3	0	3

	smoking	diabetes	bmi	heart_disease
1	1	0	23.4	1
2	0	1	35.4	0
3	1	1	29.9	0
4	1	0	15.2	1
5	1	0	16.7	1
6	1	0	33.8	0