

MVP | ENGENHARIA DE DADOS

ALUNO: CLAYTON DOS SANTOS CARVALHO

Fonte de dados:

- **Diabetes:** raw.githubusercontent.com/ClaytonCarvalho/MVP.v1-Clayton_Carvalho/main/diabetes_binary_health_indicators_BRFSS2015.csv
- **Doenças cardíacas:** <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data>

Ferramenta: Google Cloud

PROJETO

Objetivo

Identificar a representatividade referente ao volume de pacientes com doenças cardíacas e diabetes, abrindo por idade e sexo.

Armazenamento

Utilizando o Google Cloud Storage, a primeira etapa é a criação do local de armazenamento dos dados a serem utilizados no estudo. Sendo assim, criei a pasta chamada clayton1987mvp (Buckets).

The screenshot displays the Google Cloud Storage interface. At the top, the 'Google Cloud' logo and the project name 'mvp1-tom' are visible. A search bar contains the text 'Pesquise (/) recursos, documentos, produtos e muito mais'. The left sidebar shows the 'Cloud Storage' menu with 'Buckets' selected. The main area displays the details for the bucket 'clayton1987mvp'. Below the bucket name, a table lists its properties:

Local	Classe de armazenamento	Acesso público	Proteção
us (várias regiões nos Estados Unidos)	Standard	Não público	Nenhum

At the bottom, a 'CLOUD SHELL' terminal window is open, showing the following text:

```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to neat-bongo-399917.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
```

Coleta

Download na pasta criada em nuvem, para futura manipulação

Print de parte do processo de download da base contendo os dados de diabetes. Essa base foi salva na minha conta no Github inicialmente e baixada no cloud Shell diretamente de lá.

```
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.108.133, 185.199.109.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)[185.199.110.133]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 22738154 (22M) [text/plain]
Saving to: 'diabetes_binary_health_indicators_BRFSS2015.csv'

diabetes_binary_health_indicators_BRFSS2015.csv      100%[=====]
2023-10-01 00:23:18 (127 MB/s) - 'diabetes_binary_health_indicators_BRFSS2015.csv' saved [22738154/22738154]

faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ls
diabetes_binary_health_indicators_BRFSS2015.csv  imb  imbd  imdb  README-cloudshell.txt
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ gcloud storage cp *.csv gs://clayton1987mvp/
WARNING: Omitting file://imb because it is a container, and recursion is not enabled.
WARNING: Omitting file://imbd because it is a container, and recursion is not enabled.
WARNING: Omitting file://imdb because it is a container, and recursion is not enabled.
Copying file://diabetes binary health indicators BRFSS2015.csv to gs://clayton1987mvp/diabetes_binary_health_indicators_BRFSS2015.csv
Copying file://README-cloudshell.txt to gs://clayton1987mvp/README-cloudshell.txt
Completed files 2 | 21.7MiB

Average throughput: 87.7MiB/s
ERROR: (gcloud.storage.cp) The following URLs matched no objects or files:
-imb
-imbd
-imdb
-csv
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ^C
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ gcloud storage cp diabetes binary health indicators BRFSS2015.csv gs://clayton1987mvp/
Copying file://diabetes binary health indicators BRFSS2015.csv to gs://clayton1987mvp/diabetes_binary_health_indicators_BRFSS2015.csv
Completed files 1/1 | 21.7MiB/21.7MiB

Average throughput: 66.1MiB/s
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ls
diabetes_binary_health_indicators_BRFSS2015.csv  imb  imbd  imdb  README-cloudshell.txt
faturamentocarv@cloudshell:~ (neat-bongo-399917)$
```

Print de parte do processo de download da base contendo os dados de doenças cardíacas. Essa base vem do repositório de aprendizado de máquina da UCI (Universidade da Califórnia, Irvine).

```
-.csv
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ^C
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ gcloud storage cp diabetes binary health indicators BRFSS2015.csv gs://clayton1987mvp/
Copying file://diabetes binary health indicators BRFSS2015.csv to gs://clayton1987mvp/diabetes_binary_health_indicators_BRFSS2015.csv
Completed files 1/1 | 21.7MiB/21.7MiB

Average throughput: 66.1MiB/s
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ls
diabetes_binary_health_indicators_BRFSS2015.csv  imb  imbd  imdb  README-cloudshell.txt
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ls
-bash: ls: command not found
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ls
diabetes_binary_health_indicators_BRFSS2015.csv  imb  imbd  imdb  README-cloudshell.txt
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ gcloud storag https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data
ERROR: (gcloud) Invalid choice: 'storag'.
Maybe you meant:
  gcloud storage

To search the help text of gcloud commands, run:
  gcloud help -- SEARCH TERMS
faturamentocarv@cloudshell:~ (neat-bongo-399917)$ wget https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data
--2023-10-01 00:57:48-- https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data
Resolving archive.ics.uci.edu (archive.ics.uci.edu)... 128.195.10.252
Connecting to archive.ics.uci.edu (archive.ics.uci.edu)[128.195.10.252]:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified
Saving to: 'processed.cleveland.data'

processed.cleveland.data      [ <-> ]
2023-10-01 00:57:49 (286 KB/s) - 'processed.cleveland.data' saved [18461]

faturamentocarv@cloudshell:~ (neat-bongo-399917)$ ls
diabetes_binary_health_indicators_BRFSS2015.csv  imb  imbd  imdb  processed.cleveland.data  README-cloudshell.txt
faturamentocarv@cloudshell:~ (neat-bongo-399917)$
```

Segue o print dos arquivos importados, possibilitando o início dos trabalhos no google Cloud.

		Pesquise (/) recursos, documentos, produtos e muito mais	
	Detalhes do bucket		
	clayton1987mvp		
	Local	Classe de armazenamento	Acesso público
	us (várias regiões nos Estados Unidos)	Standard	Não público
			Proteção
			Nenhum
	OBJETOS	CONFIGURAÇÃO	PERMISSÕES
		PROTEÇÃO	CICLO DE VIDA
	Intervalos > clayton1987mvp		
	FAZER UPLOAD DE ARQUIVOS		
	CARREGAR PASTA		
	CRIAR PASTA		
	TRANSFERIR DADOS		
	EXCLUIR		
	Filtrar apenas pelo prefixo do nome		
	Filtro Filtrar objetos e pastas		
	<input type="checkbox"/>	Nome	Tamanho
	<input type="checkbox"/>	README-cloudshell.txt	913 B
	<input type="checkbox"/>	diabetes_binary_health_indicators_BRFSS201...	21,7 MB
	<input type="checkbox"/>	processed.cleveland.csv	18 KB
	<input type="checkbox"/>	processed.cleveland.data	18 KB

Modelagem

Ambas as bases, foram tratadas anteriormente, de maneira que não há erros a serem corrigidos, como missings e etc.

Porém, foi necessário renomear as colunas de maneira a torná-las compreensíveis, bem como substituir 0.0 por Masculino e 1.0 por feminino.

Exemplo:

Output Schema

col_63_0	string
col_1_0	string
col_1_0_2	string
col_145_0	string
col_233_0	string
col_1_0_3	string
col_2_0	string
col_150_0	string
col_0_0	string
col_2_3	string
col_3_0	string
col_0_0_2	string
col_6_0	string
col_0	int

Idade
Sexo
cp
pressao_arterial_repouso
colesterol_serico
glicemia_jejum
frequencia_cardiaca
angin_induzida
exang
oldpeak
slope
ca
thal
num

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

Linha	idade_numerica	sexo_nom	quantidade
1	29.0	Feminino	111706
2	34.0	Masculino	141974
3	34.0	Feminino	111706
4	35.0	Masculino	141974
5	35.0	Feminino	335118
6	37.0	Masculino	141974
7	37.0	Feminino	111706
8	38.0	Feminino	223412
9	39.0	Masculino	283948
10	39.0	Feminino	223412
11	40.0	Feminino	335118
12	41.0	Masculino	567896
13	41.0	Feminino	670236
14	42.0	Masculino	283948
15	42.0	Feminino	670236

Carga

1. Utilização de duas bases diferentes em 2 GCS (Google Cloud Storage)

Armazenando duas bases de dados distintas no Google Cloud Storage, uma solução de armazenamento de objetos.

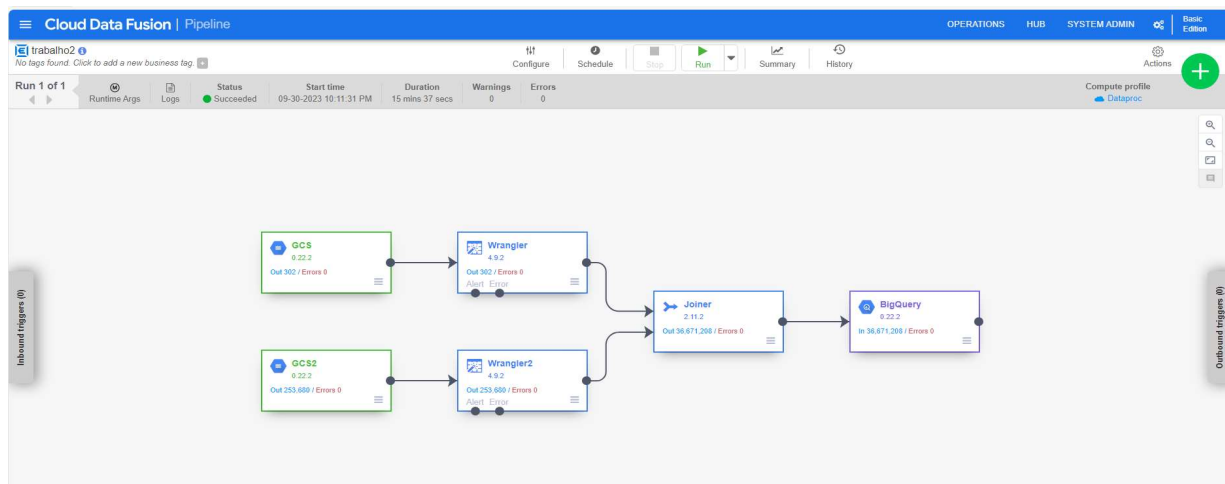
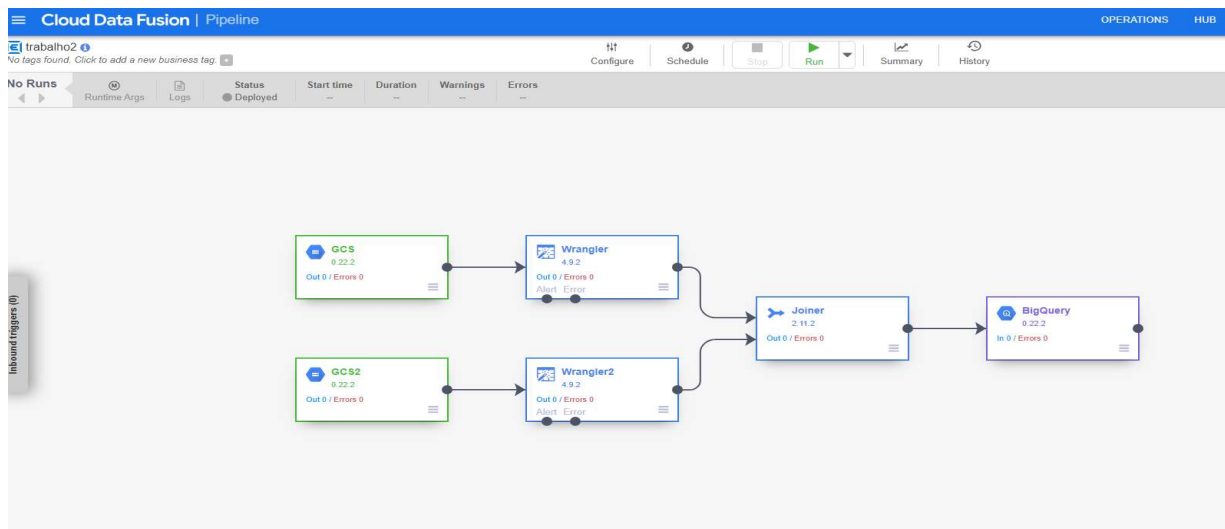
2. Vinculação de cada base a um Wrangler

Wrangler: Wrangler é uma etapa no Cloud Data Fusion que permite limpar e transformar os dados. Aqui, estou usando Wranglers para realizar algumas transformações ou limpezas iniciais nos dados.

3. Vinculação dos dois Wranglers a um único Joiner utilizando a coluna chamada "idade". Joiner: Esta é uma etapa que permite combinar dados de diferentes fontes ou tabelas baseando-se em uma chave comum, que neste caso, é a coluna "idade". Neste caso, estou realizando um "join" das duas bases de dados usando a coluna "idade", que é comum entre elas, para criar um conjunto de dados unificado.

4. Vinculação do Joiner ao BigQuery

Finalmente, estou enviando os dados resultantes para o BigQuery, um serviço de armazenamento de dados altamente escalonável e solução de análise. Aqui, posso realizar análises adicionais e consultas SQL ou usar os dados para treinar modelos de machine learning.



Nessa etapa, retiro todas as colunas que não serão utilizadas na minha análise e unifico as colunas pertinentes, que são as colunas que contêm a informação de idade e sexo.

Obs: Foi criado o bucket temporário “gs://clayton1987mvp2-temp”, para atender a premissa da plataforma.

ANÁLISE

a. Qualidade de dados

Existem problemas no conjunto de dados?

1)verificando nulos

```
1 SELECT
2   COUNT(*) AS total_linhas,
3   SUM(CASE WHEN quantidade IS NULL THEN 1 ELSE 0 END) AS quantidade_nulos
4 FROM (
5   SELECT
6     ROUND(CAST(idade AS FLOAT64)) AS idade_numerica,
7     sexo_nom,
8     COUNT(*) AS quantidade
9   FROM MVP30092023.analiseidade
10  GROUP BY idade_numerica, sexo_nom
11 )
12
```

Resultados da consulta

INFORMAÇÕES DO JOB		RESULTADOS	JSON	DETALHES DA EXECUÇÃO
id	total_linhas	quantidade_nulos		
1	73	0		

2)verificando mínimo e máximo

```
1 SELECT
2   MIN(idade_numerica) AS min_idade_numerica,
3   MAX(idade_numerica) AS max_idade_numerica,
4   MIN(quantidade) AS min_quantidade,
5   MAX(quantidade) AS max_quantidade
6 FROM (
7   SELECT
8     ROUND(CAST(idade AS FLOAT64)) AS idade_numerica,
9     sexo_nom,
10    COUNT(*) AS quantidade
11  FROM MVP30092023.analiseidade
12  GROUP BY idade_numerica, sexo_nom
13 )
14
```

Resultados da consulta

INFORMAÇÕES DO JOB		RESULTADOS	JSON	DETALHES DA EXECUÇÃO
id	min_idade_numerica	max_idade_numerica	min_quantidade	max_quantidade
1	29.0	77.0	111706	1452178

3)Encontrado 14 linhas com a mesma quantidade. Não se trata de um erro, devido a baixa quantidade encontrada.

```
1 SELECT quantidade, COUNT(*)
2 FROM (
3   SELECT
4     ROUND(CAST(idade AS FLOAT64)) AS idade_numerica,
5     sexo_nom,
6     COUNT(*) AS quantidade
7   FROM MVP30092023.analiseidade
8   GROUP BY idade_numerica, sexo_nom
9 )
10 GROUP BY quantidade
11 HAVING COUNT(*) > 1;
12
```

Resultados da consulta

INFORMAÇÕES DO JOB		RESULTADOS	JSON	DETALHE
id	quantidade	total		
1	1452178	3		
2	335118	5		
3	558530	3		
4	425922	9		
5	709870	3		
6	567896	5		
7	446824	7		
8	283948	6		
9	781942	2		
10	111706	4		
11	893648	2		
12	223412	3		
13	670236	6		
14	1452178	3		

4)verificando valores zerados

```
1 SELECT COUNT(*)
2 FROM (
3     SELECT
4         ROUND(CAST(idade AS FLOAT64)) AS idade_n
5         sexo_nom,
6         COUNT(*) as quantidade
7     FROM MVP30092023.analiseidade
8     GROUP BY idade_numerica, sexo_nom
9 )
10 WHERE quantidade = 0;
```

Resultados da consulta

INFORMAÇÕES DO JOB		RESULTADOS	JSON
inha	f0_		
1	0		

5)verificando os outliers (quartil)

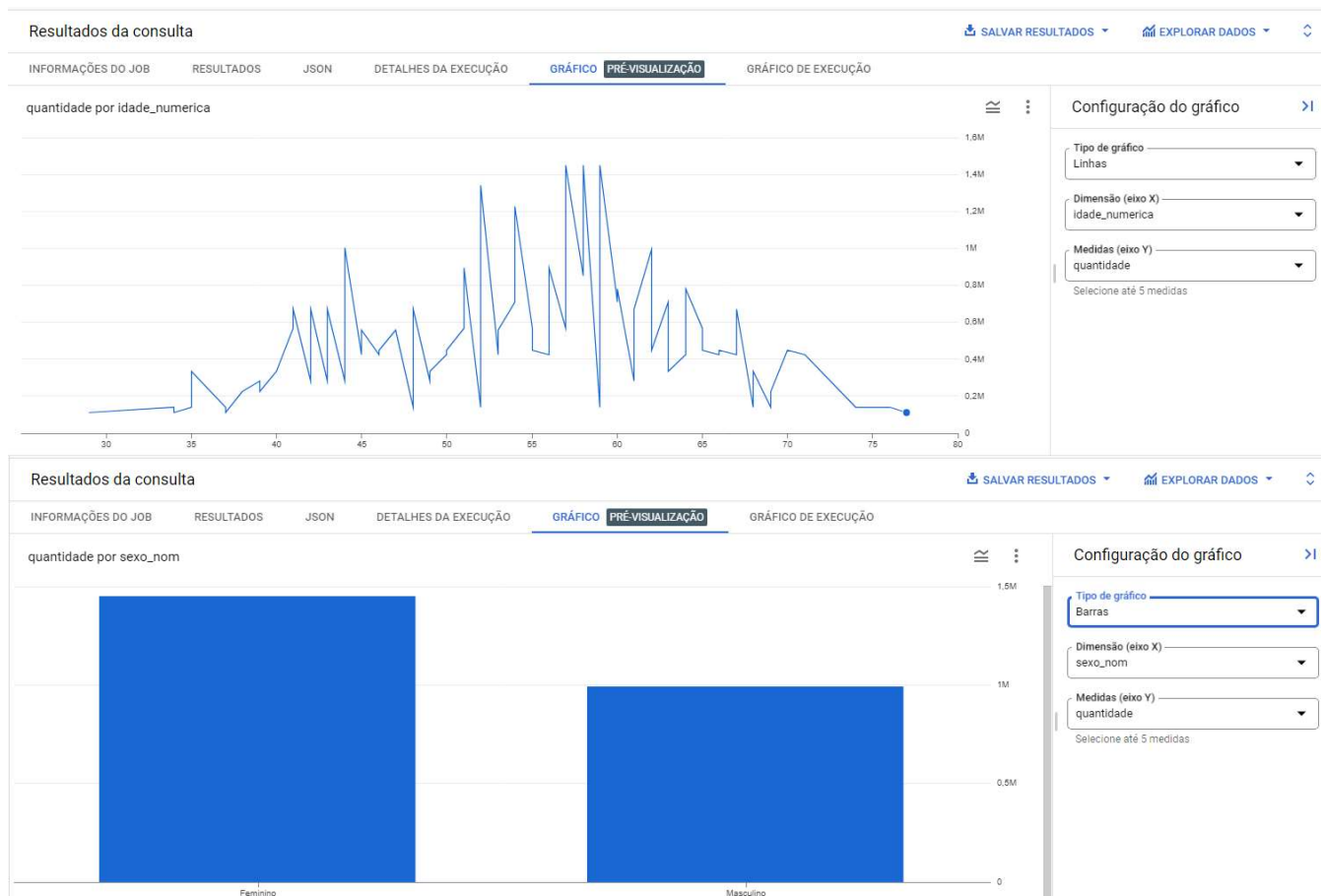
```
1 WITH agrupado AS (
2     SELECT
3         ROUND(CAST(idade AS FLOAT64)) AS idade_numerica,
4         sexo_nom,
5         COUNT(*) as quantidade
6     FROM MVP30092023.analiseidade
7     GROUP BY idade_numerica, sexo_nom
8 ),
9
10 quartis AS (
11     SELECT
12         MAX(CASE WHEN idx = 1 THEN idade_numerica END) AS Q1,
13         MAX(CASE WHEN idx = 3 THEN idade_numerica END) AS Q3
14     FROM (
15         SELECT
16             idade_numerica,
17             idx
18         FROM (
19             SELECT
20                 idade_numerica,
21                 ROW_NUMBER() OVER (ORDER BY idade_numerica) as idx
22             FROM (
23                 SELECT
24                     DISTINCT idade_numerica
25                 FROM agrupado
26             )
27         )
28         WHERE idx IN (1, 3)
29     )
30 )
31
32 SELECT
33     a.idade_numerica
34 FROM agrupado a, quartis q
35 WHERE a.idade_numerica < (q.Q1 - 1.5 * (q.Q3 - q.Q1))
36     OR a.idade_numerica > (q.Q3 + 1.5 * (q.Q3 - q.Q1));
```

6) Caso haja, como esses problemas podem ser resolvidos para que não afetem as respostas das perguntas que quer solucionar?

Resposta: Não houve problemas significativos na base, a ponto de comprometer o estudo.

b. Solução do problema

Utilizando os gráficos abaixo, podemos observar a representatividade por sexo e por idade, referente ao volume de pacientes com doenças cardíacas e diabetes. No primeiro gráfico percebemos que a maior amostragem está entre os 55 e os 60 anos e no segundo gráfico, vemos a predominância feminina entre os pacientes.



Autoavaliação

Como iniciante na linguagem python e SQL, tive muita dificuldade com a construção dos scripts de comando, bem como nas configurações dentro do google cloud.

Porém, sinto cada vez mais gosto pela área, e, também a vontade de desenvolver o conhecimento que já adquiri.

Sinto que de forma básica, consegui atender a proposta do exercício do MVP dessa sprint e pretendo evoluir cada etapa desse trabalho no futuro, afim de, melhorar o meu portfólio.

Agradeço muito a ajuda dos professores !