

PROJETO FINAL

ACIDENTES TERRESTRES

GRUPO 03, TURMA BCW6



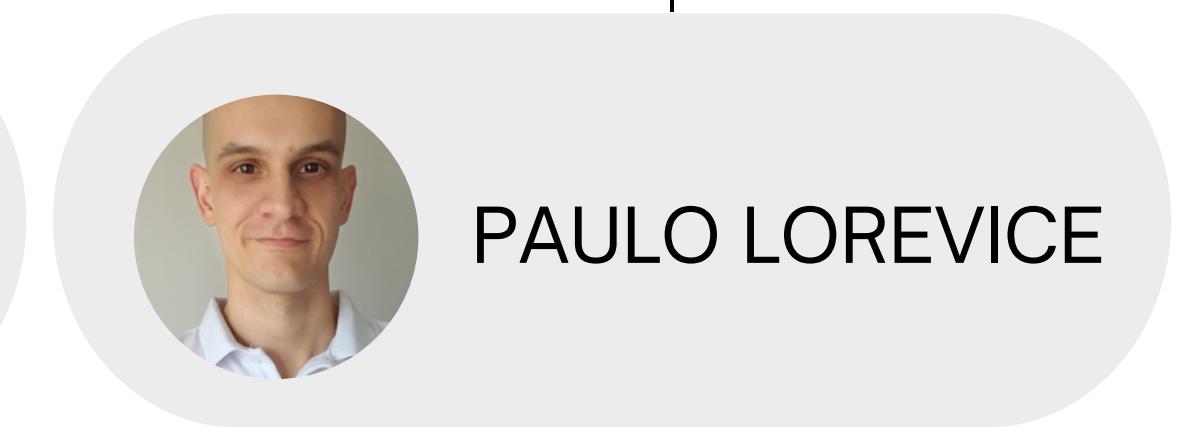
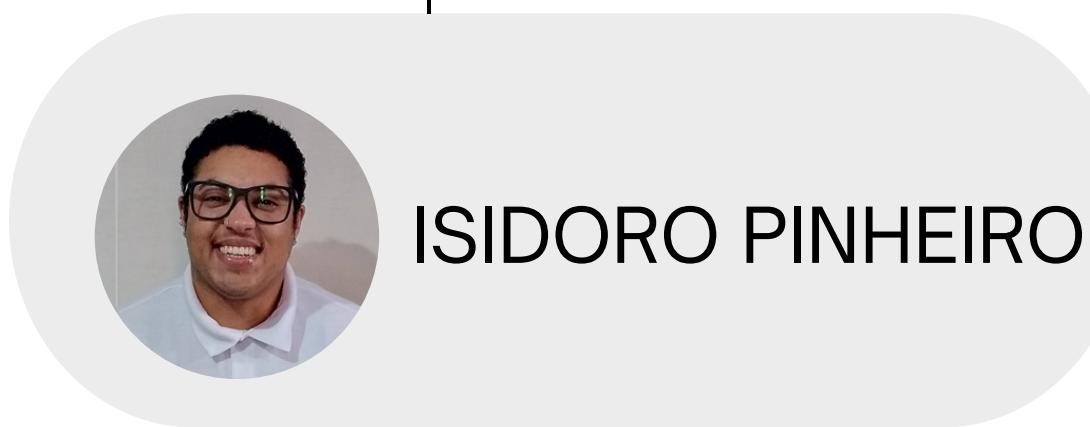


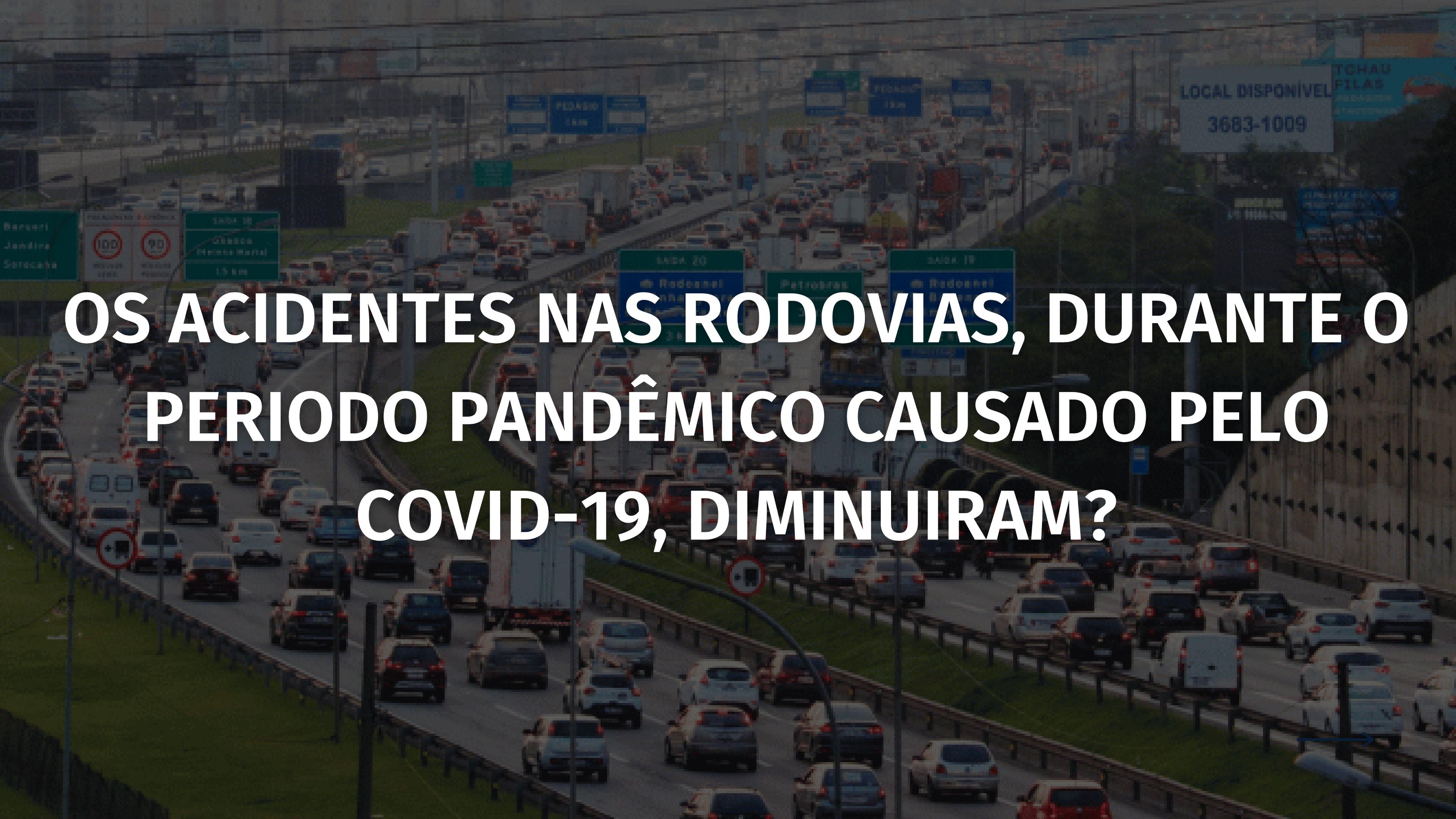
ESCOPO GERAL DO PROJETO

ITENS OBRIGATÓRIOS DO PROJETO

-
- Datasets em formatos diferentes;
 - Operações Pandas;
 - Operações PySpark;
 - Operações SparkSQL;
 - Datasets em armazenamento Cloud na plataforma GCP;
 - Armazenamento dos dados tratados em Datalake (GStorage) e DW(BigQuery);
 - Análises no BigQuery;
 - Criar Dashboard no Data Studio;
 - Demonstração em Workflow das etapas de ETL.
-

INTEGRANTES DO GRUPO





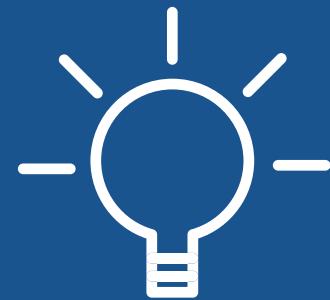
OS ACIDENTES NAS RODOVIAS, DURANTE O
PERÍODO PANDÉMICO CAUSADO PELO
COVID-19, DIMINUIRAM?

LOCAL DISPONÍVEL
3683-1009

TCHAU
FILAS
PRAZO
de vacinação

PARA RESPONDER A ESSA
PERGUNTA NOSSO GRUPO SE
ESTRUTUROU DA SEGUINTE
FORMA





1 - A PERGUNTA

- O que o tema Acidentes Terrestres podia nos trazer?
- Para onde podíamos seguir dentro do tema?
- Quais assuntos podíamos trazer para conversar com o tema principal?
- E dessas conversas, quais teriam os insights mais interessantes?
- Qual história nós queremos contar dentro do nosso tema?



2 - A BUSCA DOS DADOS

Depois de decidirmos o que queríamos contar, foi a vez de ir atrás dos dados que pudessem nos ajudar a trazer essa "história" à tona.

Algumas palavras chaves que ajudaram no processo foram:

- Trânsito;
- Tráfego;
- Pedágio;
- Acidentes;
- Pandemia;
- Covid-19;
- Óbitos;
- Isolamento;
- Pessoa na rua;
- Aumento, ou queda, de tráfego de carros nas rodovias.



3 - A AÇÃO

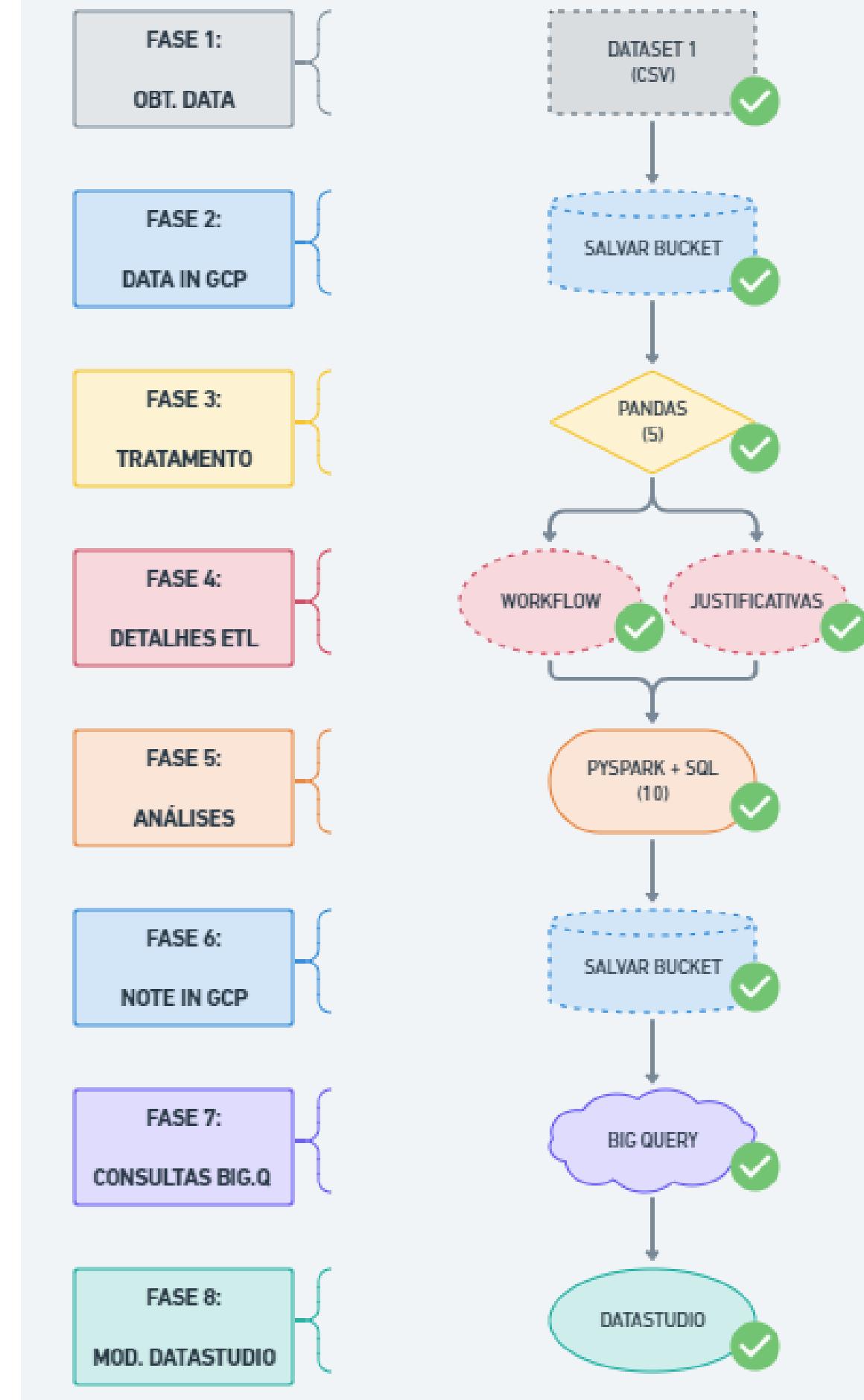
Depois de muito procurar eis que encontramos 3 Bases de Dados para trabalharmos:

- Covid, encontrado no site Ministério da Saúde;
- Acidentes, encontrado no site da Policia Rodoviaria Federal;
- Pedágio, encontrado no site ANTT - Agência Nacional de Transportes Terrestres

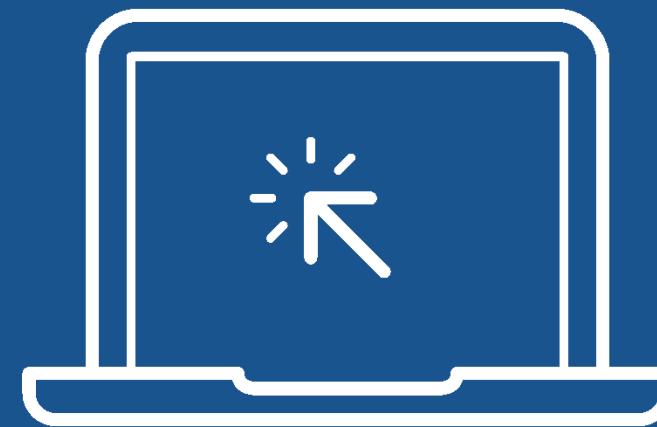
Com os dados nas mãos, chegou a hora de trabalhar para lapidar e extrair as melhores informações para ilustrar o que queríamos contar.

A CONSTRUÇÃO DO PLANEJAMENTO FOI ESSENCIAL PARA PROSSEGUIRMOS PROSPERANDO

WORKFLOW DE PLANEJAMENTO



E NO PROCESSO USAMOS DIVERSAS TECNOLOGIAS E FERRAMENTAS, CONHEÇA ALGUMAS DELAS:



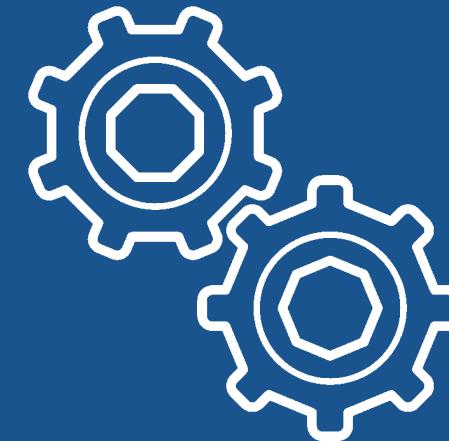
FERRAMENTAS & TECNOLOGIAS

Python;
Pandas;
PySpark;
SparkSQL;
BigQuery;
Colaboratory;
Google Cloud Platform;
Google Cloud Storage;
Canva;
DataStudio.



COMUNICAÇÃO

Google Meet;
WhatsApp;
E-mail.



METODOLOGIA ÁGIL

Trello;
Whimsical.



GOOGLE CLOUD STORAGE

The screenshot shows the Google Cloud Platform interface for Cloud Storage. The top navigation bar includes the 'Google Cloud Platform' logo, the project name 'Projeto-Final-Grupo03', and a search bar. The main area has a sidebar with 'Cloud Storage' selected, and a central 'Browser' view showing a list of buckets. The 'Browser' tab is active, along with 'CREATE BUCKET', 'DELETE', and 'REFRESH' buttons. A 'Filter' section allows filtering by bucket name. The list of buckets is as follows:

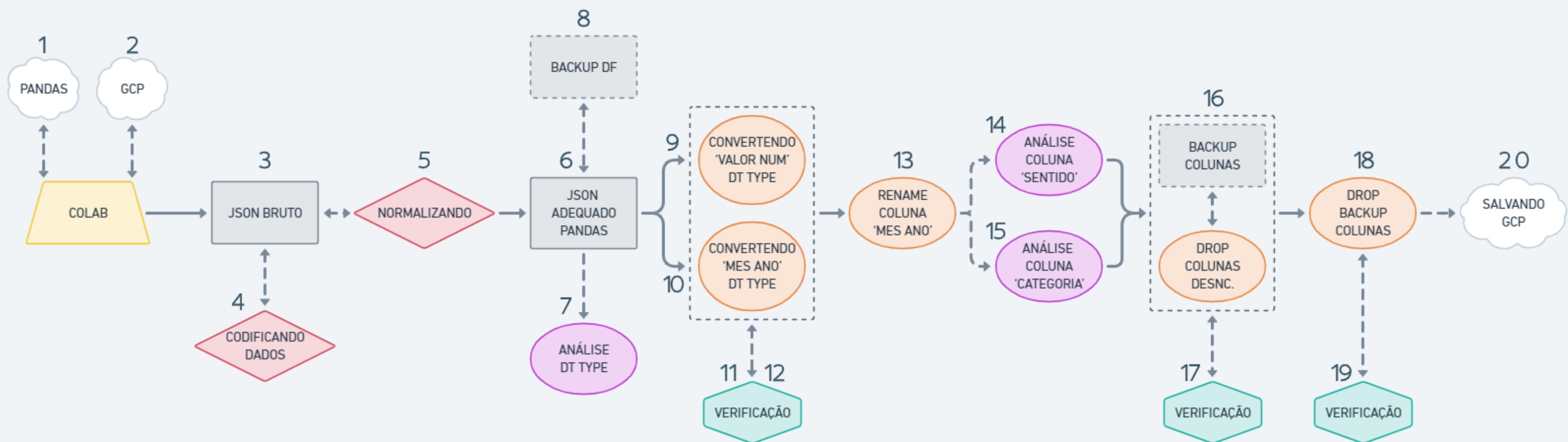
Name	Created
datasets_gp03	Nov 18, 2021, 2:48:46 PM
detalhamento_etl_gp03	Nov 17, 2021, 12:27:02 PM
notebooks_pandas_gp03	Nov 17, 2021, 12:19:38 PM
notebooks_pyspark_sql_gp03	Nov 17, 2021, 12:21:23 PM
sql_bigquery_queries_gp03	Nov 26, 2021, 3:14:31 PM

CONHEÇA UM POUCO DE >> PANDAS

DATASET DE PEDÁGIO



WORKFLOW DATAFRAME PEDÁGIO



NORMALIZAÇÃO E CONCATENAÇÃO DE DATASETS EM PANDAS

Quando lemos datasets em .json eles ficam desta forma.

```
#4 - IMPORTANDO O DOCUMENTO E CODIFICANDO DADOS
```

```
df1 = pd.read_json("/tmp/VolTrafego2019.json")
df2 = pd.read_json("/tmp/VolTrafego2020.json")
```

```
#5 - CONSULTANDO DATAFRAME EM JSON PARA VER SE OS DADOS CARREGARAM DE FORMA CORRETA
```

```
df1
```

```
          empresas_habilitadas_regular
0      {'Concessionaria': 'NOVADUTRA', 'mes_ano': '01...
1      {'Concessionaria': 'NOVADUTRA', 'mes_ano': '01...
2      {'Concessionaria': 'NOVADUTRA', 'mes_ano': '01...
3      {'Concessionaria': 'NOVADUTRA', 'mes_ano': '01...
4      {'Concessionaria': 'NOVADUTRA', 'mes_ano': '01...
...
33631  {'Concessionaria': 'AUTOPISTA FERNÃO DIAS', 'm...
33632  {'Concessionaria': 'AUTOPISTA FERNÃO DIAS', 'm...
33633  {'Concessionaria': 'AUTOPISTA FERNÃO DIAS', 'm...
33634  {'Concessionaria': 'AUTOPISTA FERNÃO DIAS', 'm...
33635  {'Concessionaria': 'AUTOPISTA FERNÃO DIAS', 'm...
33636 rows × 1 columns
```

CÓDIGO PARA EDIÇÃO

Para normalizar as estruturas e concatenar os dois datasets, usamos os seguintes códigos:

```
#6 - NORMALIZANDO DADOS CONTIDOS NOS ARQUIVOS JSON  
  
df1 = pd.json_normalize(df1["empresas_habilitadas_regular"])  
df2 = pd.json_normalize(df2["empresas_habilitadas_regular"])  
  
#8 - NORMALIZANDO TIPO DATE NO DATAFRAME1  
  
df1["mes_ano"] = pd.to_datetime(df1["mes_ano"], dayfirst=True)  
  
#9 - NORMALIZANDO TIPO DATE NO DATAFRAME2  
  
df2["mes_ano"] = pd.to_datetime(df2["mes_ano"], dayfirst=True)  
  
#10 - CONCATENANDO DATAFRAMES  
  
frames = [df1, df2]  
df = pd.concat(frames)
```

O RESULTADO FOI

#11 - VISUALIZANDO DATAFRAME

```
df
```

	Concessionaria	mes_ano	Sentido	Praça	Categoria	Tipo_de_veiculo	Volume_total
0	NOVADUTRA	2019-01-01	Sul	Praça 01 BR-116/SP km 204,50	Categoria 1	Passeio	640429
1	NOVADUTRA	2019-01-01	Norte	Praça 01 BR-116/SP km 204,50	Categoria 1	Passeio	787419
2	NOVADUTRA	2019-02-01	Sul	Praça 01 BR-116/SP km 204,50	Categoria 1	Passeio	511718
3	NOVADUTRA	2019-02-01	Norte	Praça 01 BR-116/SP km 204,50	Categoria 1	Passeio	638796
4	NOVADUTRA	2019-03-01	Sul	Praça 01 BR-116/SP km 204,50	Categoria 1	Passeio	579422
...
35538	AUTOPISTA FERNÃO DIAS	2020-12-01	Sul	Praça 06 BR-381/MG km 659	Outros	Comercial	16137
35539	AUTOPISTA FERNÃO DIAS	2020-12-01	Norte	Praça 07 BR-381/MG km 596,935	Outros	Comercial	15141
35540	AUTOPISTA FERNÃO DIAS	2020-12-01	Sul	Praça 07 BR-381/MG km 596,935	Outros	Comercial	15863
35541	AUTOPISTA FERNÃO DIAS	2020-12-01	Norte	Praça 08 BR-381/MG km 546	Outros	Comercial	14362
35542	AUTOPISTA FERNÃO DIAS	2020-12-01	Sul	Praça 08 BR-381/MG km 546	Outros	Comercial	14988

69179 rows × 7 columns

EXTRAÇÃO DE DADOS DE UMA DETERMINADA COLUNA

Precisamos extrair da coluna "Praca" do nosso dataset os nomes dos Estados e das Rodovias.

Data	Sentido	Praca
2019-01-01	Sul	Praça 01 BR-116/SP km 204,50
2019-01-01	Norte	Praça 01 BR-116/SP km 204,50
2019-02-01	Sul	Praça 01 BR-116/SP km 204,50
2019-02-01	Norte	Praça 01 BR-116/SP km 204,50
2019-03-01	Sul	Praça 01 BR-116/SP km 204,50
...
2020-12-01	Sul	Praça 06 BR-381/MG km 659
2020-12-01	Norte	Praça 07 BR-381/MG km 596,935
2020-12-01	Sul	Praça 07 BR-381/MG km 596,935
2020-12-01	Norte	Praça 08 BR-381/MG km 546
2020-12-01	Sul	Praça 08 BR-381/MG km 546

CÓDIGO PARA EDIÇÃO

Para a extração dos dados e a criação das novas colunas com os dados necessários, usamos os seguintes códigos:

```
#27 - DIVIDINDO DADOS DA COLUNA DE PRAÇAS PARA VISUALIZAR COMO EXTRAIR OS DADOS  
df['Praca'].str.split(" ", expand=True)  
  
#28 - DIVIDINDO DADOS DA COLUNA DE PRAÇAS E SALVANDO ELAS EM COLUNAS  
df[['1', '2', 'BRUF', '3', '4', '5', '6', '7', '8', '9', '10', '11']] = df["Praca"].str.split(" ", expand=True)
```

```
#31 - DIVIDINDO DADOS DA COLUNA "BRUF" E SALVANDO EM COLUNAS  
df[['BR', 'UF']] = df["BRUF"].str.split("/", expand=True)  
  
#32 - DROPANDO COLUNAS QUE FICARAM REPETIDAS  
df.drop(["BRUF"], axis=1, inplace=True)  
df.drop(["Praca"], axis=1, inplace=True)
```

O RESULTADO FOI

#33 - VISUALIZANDO DATAFRAME

df

	Concessionaria	Data	Sentido	Tipo_de_veiculo	Volume_total	BR	UF
0	NOVADUTRA	2019-01-01	Sul	Passeio	640429.0	BR-116	SP
1	NOVADUTRA	2019-01-01	Norte	Passeio	787419.0	BR-116	SP
2	NOVADUTRA	2019-02-01	Sul	Passeio	511718.0	BR-116	SP
3	NOVADUTRA	2019-02-01	Norte	Passeio	638796.0	BR-116	SP
4	NOVADUTRA	2019-03-01	Sul	Passeio	579422.0	BR-116	SP
...
35538	AUTOPISTA FERNÃO DIAS	2020-12-01	Sul	Comercial	16137.0	BR-381	MG
35539	AUTOPISTA FERNÃO DIAS	2020-12-01	Norte	Comercial	15141.0	BR-381	MG
35540	AUTOPISTA FERNÃO DIAS	2020-12-01	Sul	Comercial	15863.0	BR-381	MG
35541	AUTOPISTA FERNÃO DIAS	2020-12-01	Norte	Comercial	14362.0	BR-381	MG
35542	AUTOPISTA FERNÃO DIAS	2020-12-01	Sul	Comercial	14988.0	BR-381	MG

69179 rows × 7 columns

CONHEÇA UM POUCO DE >> PYSPARK

DATASET DE ACIDENTES



CRIANDO UMA COLUNA DE CLASSIFICAÇÃO DE REGIÕES

```
#CRIA NOVA COLUNA COM CLASSIFICAÇÃO DE REGIÃO DO BRASIL
```

```
df1 = (df.withColumn('regiao',  
                      F.when((df.uf == 'PR') | (df.uf == 'RS') | (df.uf == 'SC'), 'Sul')  
                        .when((df.uf == 'SP') | (df.uf == 'RJ') | (df.uf == 'MG') | (df.uf == 'ES'), 'Sudeste')  
                        .when((df.uf == 'MT') | (df.uf == 'DF') | (df.uf == 'GO') | (df.uf == 'MS'), 'Centro-Oeste')  
                        .when((df.uf == 'RR') | (df.uf == 'AP') | (df.uf == 'AM') | (df.uf == 'PA') | (df.uf == 'AC')  
                            | (df.uf == 'RO') | (df.uf == 'TO'), 'Norte').otherwise('Nordeste')))
```

Precisavamos criar uma coluna onde os estados fossem classificados em Sul, Sudeste, Leste, e assim por diante, para isso usamos o código acima.

O RESULTADO FOI

id_acidente	id_pessoa	data	dia_semana	uf	causa_acidente	estado_fisico	idade	sexo	regiao
182256	403856	2019-01-01	terça-feira	CE	Animais na Pista	Ileso	35	Masculino	Nordeste
182263	402859	2019-01-01	terça-feira	MT	Defeito Mecânico ...	Ileso	30	Masculino	Centro-Oeste
182277	402850	2019-01-01	terça-feira	PA	Velocidade Incomp...	Ileso	54	Masculino	Norte
182289	402431	2019-01-01	terça-feira	BA	Ingestão de Álcoo...	Ileso	43	Masculino	Nordeste
182307	402642	2019-01-01	terça-feira	BA	Ingestão de Álcoo...	Lesões Graves	37	Masculino	Nordeste
182307	402638	2019-01-01	terça-feira	BA	Ingestão de Álcoo...	Lesões Leves	57	Masculino	Nordeste
182316	402361	2019-01-01	terça-feira	RS	Falta de Atenção ...	Ileso	28	Masculino	Sul
182334	402401	2019-01-01	terça-feira	SE	Falta de Atenção ...	Não Informado	null	Não Informado	Nordeste
182361	402521	2019-01-01	terça-feira	MG	Ultrapassagem Ind...	Lesões Leves	52	Masculino	Sudeste
182362	404322	2019-01-01	terça-feira	PE	Ingestão de Álcoo...	Ileso	59	Masculino	Nordeste
182410	404399	2019-01-01	terça-feira	MG	Pista Escorregadia	Ileso	48	Masculino	Sudeste
182433	404380	2019-01-01	terça-feira	RJ	Ingestão de Álcoo...	Ileso	49	Masculino	Sudeste
182443	404484	2019-01-02	quarta-feira	RJ	Carga Excessiva e...	Lesões Leves	42	Masculino	Sudeste
182451	402669	2019-01-02	quarta-feira	RS	Condutor Dormindo	Ileso	40	Masculino	Sul
182459	402929	2019-01-02	quarta-feira	MG	Defeito Mecânico ...	Ileso	36	Masculino	Sudeste
182459	403072	2019-01-02	quarta-feira	MG	Defeito Mecânico ...	Ileso	30	Masculino	Sudeste
182472	405393	2019-01-02	quarta-feira	CE	Falta de Atenção ...	Ileso	43	Masculino	Nordeste
182486	403142	2019-01-02	quarta-feira	SP	Condutor Dormindo	Lesões Leves	48	Feminino	Sudeste
182486	402972	2019-01-02	quarta-feira	SP	Condutor Dormindo	Ileso	39	Masculino	Sudeste
182486	403140	2019-01-02	quarta-feira	SP	Condutor Dormindo	Ileso	26	Masculino	Sudeste

CRIANDO UMA COLUNA DE CLASSIFICAÇÃO DE MÊS E ANO

```
#CRIA NOVA COLUNA COM MÊS  
  
df2 = df1.withColumn('mes', F.month(df1.data))
```

```
#CRIA NOVA COLUNA COM ANO  
  
df2 = df2.withColumn('ano', F.year(df2.data))
```

Precisavamos criar uma coluna onde aparecesse Mês e Ano no dataframe, para isso usamos o código acima.

O RESULTADO FOI

id_acidente	id_pessoa	data	dia_semana	uf	causa_acidente	estado_fisico	idade	sexo	regiao	mes	ano
182256	403856	2019-01-01	terça-feira	CE	Animais na Pista	Ileso	35	Masculino	Nordeste	1	2019
182263	402859	2019-01-01	terça-feira	MT	Defeito Mecânico ...	Ileso	30	Masculino	Centro-Oeste	1	2019
182277	402850	2019-01-01	terça-feira	PA	Velocidade Incomp...	Ileso	54	Masculino	Norte	1	2019
182289	402431	2019-01-01	terça-feira	BA	Ingestão de Álcoo...	Ileso	43	Masculino	Nordeste	1	2019
182307	402642	2019-01-01	terça-feira	BA	Ingestão de Álcoo...	Lesões Graves	37	Masculino	Nordeste	1	2019
182307	402638	2019-01-01	terça-feira	BA	Ingestão de Álcoo...	Lesões Leves	57	Masculino	Nordeste	1	2019
182316	402361	2019-01-01	terça-feira	RS	Falta de Atenção ...	Ileso	28	Masculino	Sul	1	2019
182334	402401	2019-01-01	terça-feira	SE	Falta de Atenção ...	Não Informado	null	Não Informado	Nordeste	1	2019
182361	402521	2019-01-01	terça-feira	MG	Ultrapassagem Ind...	Lesões Leves	52	Masculino	Sudeste	1	2019
182362	404322	2019-01-01	terça-feira	PE	Ingestão de Álcoo...	Ileso	59	Masculino	Nordeste	1	2019
182410	404399	2019-01-01	terça-feira	MG	Pista Escorregadia	Ileso	48	Masculino	Sudeste	1	2019
182433	404380	2019-01-01	terça-feira	RJ	Ingestão de Álcoo...	Ileso	49	Masculino	Sudeste	1	2019
182443	404484	2019-01-02	quarta-feira	RJ	Carga Excessiva e...	Lesões Leves	42	Masculino	Sudeste	1	2019
182451	402669	2019-01-02	quarta-feira	RS	Condutor Dormindo	Ileso	40	Masculino	Sul	1	2019
182459	402929	2019-01-02	quarta-feira	MG	Defeito Mecânico ...	Ileso	36	Masculino	Sudeste	1	2019
182459	403072	2019-01-02	quarta-feira	MG	Defeito Mecânico ...	Ileso	30	Masculino	Sudeste	1	2019
182472	405393	2019-01-02	quarta-feira	CE	Falta de Atenção ...	Ileso	43	Masculino	Nordeste	1	2019
182486	403142	2019-01-02	quarta-feira	SP	Condutor Dormindo	Lesões Leves	48	Feminino	Sudeste	1	2019
182486	402972	2019-01-02	quarta-feira	SP	Condutor Dormindo	Ileso	39	Masculino	Sudeste	1	2019
182486	403140	2019-01-02	quarta-feira	SP	Condutor Dormindo	Ileso	26	Masculino	Sudeste	1	2019

CONHEÇA UM POUCO DE >> SPARKSQL

ACIDENTES & COVID



CONSULTA SQL

```
1 # TOTAL DE ACIDENTES POR ANO  
2  
3 spark.sql("SELECT COUNT(DISTINCT id_acidente) AS total_2019 FROM temp WHERE ano = 2019").show()  
4 spark.sql("SELECT COUNT(DISTINCT id_acidente) AS total_2020 FROM temp WHERE ano = 2020").show()
```

```
+-----+  
|total_2019|  
+-----+  
|      67351|  
+-----+
```

```
+-----+  
|total_2020|  
+-----+  
|      63372|  
+-----+
```

CONSULTA SQL

```
#26 SOMA NÚMERO DE NOVOS CASOS POR MÊS DE JANEIRO A SETEMBRO EM 2020
```

```
spark.sql("SELECT Mes, SUM(Novos_Casos) AS totalNovosCasos_2020 FROM temp WHERE ano = 2020 AND Mes BETWEEN 2 AND 9 GROUP BY Mes ORDER BY Mes ASC").show()
```

Mes	totalNovosCasos_2020
2	2.0
3	5822.0
4	81302.0
5	429011.0
6	896532.0
7	1257782.0
8	1244378.0
9	902536.0

CONHEÇA UM POUCO DE

>> BIGQUERY E DATA STUDIO

CONSULTAS E DASHBOARD



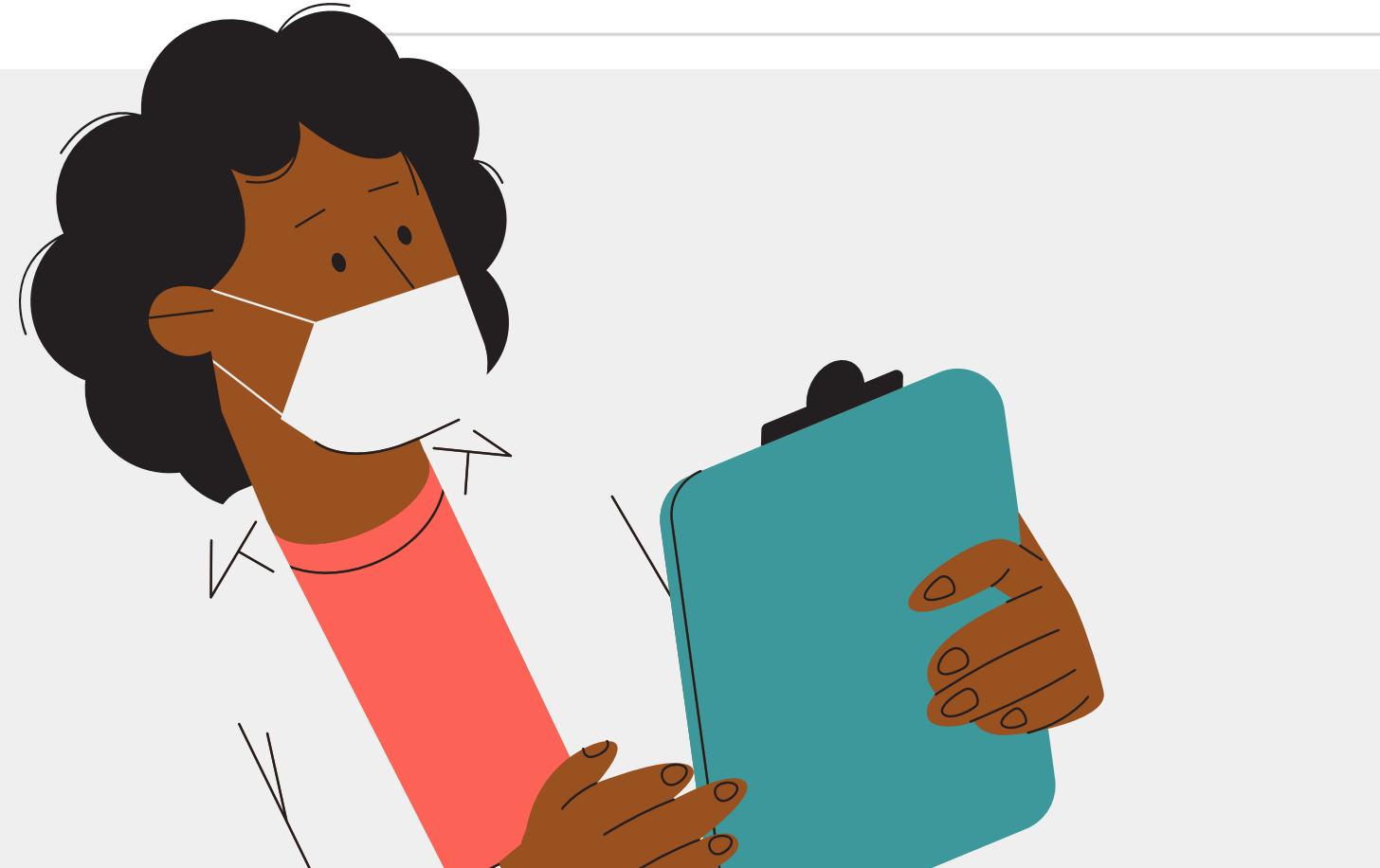
E NÃO FOI SÓ O NOSSO GRÁFICO QUE FALOU ISSO

Jovem Pan > Programas > Jornal da Manhã > Pandemia reduz o número de mortes no trânsito no Brasil em 2020

Pandemia reduz o número de mortes no trânsito no Brasil em 2020

Especialistas alertam que diminuição dos números deve ser comemorada, já que muitos veículos deixaram de circular por causa da pandemia

Por Jovem Pan 06/03/2021 12h28



Movimento de veículos na Rodovia dos Imigrantes, que dá acesso ao litoral de São Paulo, em julho

Imagem: Mister Shadow/Estadão Conteúdo



ESTADÃO conteúdo

São Paulo

01/08/2020 11h21

🕒 Esse conteúdo é antigo

Isolamento reduziu em 11% o número de mortes por acidente de trânsito em SP

PUBLICIDADE



REFERÊNCIAS BIBLIOGRÁFICAS

MINISTÉRIO DA SAÚDE

<https://opendatasus.saude.gov.br/>

POLÍCIA RODOVIÁRIA FEDERAL

<https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-acidentes>

ANTT - AGÊNCIA NACIONAL DE TRANSPORTES TERRESTRES

<https://dados.gov.br/dataset/volume-trafego-praca-pedagio>

DOCUMENTAÇÃO DE PANDAS

<https://pandas.pydata.org/docs/>

DOCUMENTAÇÃO DE PYSPARK E SPARKSQL

<https://spark.apache.org/docs/latest/index.html>

DOCUMENTAÇÃO GOOGLE CLOUD PLATFORM

<https://cloud.google.com/docs?hl=pt-br>

A TODOS,
MUITO OBRIGADO!

CONTATO DOS INTEGRANTES DO GRUPO

ISIDORO PINHEIRO

LinkedIn @isiumlord

JOÃO PEDRO GOTTI

LinkedIn @joaopedrogotti

PAULO LOREVICE

LinkedIn @paulo-loreviçe

