

Case - CIVITAS

Análise Exploratória dos Dados (EDA)

Para análise exploratória, normalmente meu processo envolve ter um primeiro contato com os dados da tabela, executando o seguinte comando.

1.1. Visualizar as primeiras linhas da tabela

```
SELECT *
FROM rj-cetrio.desafio.readings_2024_06
LIMIT 1000;
```

CASE-CIVITAS

Pesquise (/) recursos, documentos, produtos e muito mais

Pesquisa

📁

📄

🔔

?

⋮

G

SANDBOX

Configure o faturamento para fazer upgrade para a experiência completa do BigQuery.

Saiba mais

DISPENSAR

FAZER UPGRADE

Explorer

🔍 Digite para pesquisar

Você está visualizando os recursos.

MOSTRAR APENAS COM ESTRELA

▶ case-civitas

▶ datario

▼ rj-cetrio

▶ ➔ Conexões externas

▼ 📁 desafio

📄 readings_2024_06

Consulta sem título

EXECUTAR

SALVAR

FAZER O DOWNLOAD

1 SELECT *

2 FROM rj-cetrio.desafio.readings_2024_06

3 LIMIT 1000;

4

Resultados da consulta

SALVAR RESULTADOS

EXPLORAR DADOS

<

INFORMAÇÕES DO JOB

RESULTADOS

GRÁFICO

JSON

DETALHES DA EXECUÇÃO

GRÁFICO DE E

>

Linha	datahora	datahora_captura	placa	empresa	tipoveiculo
1	2024-06-09 10:55:23 UTC	2024-06-09 10:56:04 UTC	/27nwwSfnx/sg35AmBkV...	CJGWe0E/pA==	AxzAA36BbQ=
2	2024-06-11 13:34:02 UTC	2024-06-11 13:35:02 UTC	Pcu1gmN4V7Rnu2r49Nv...	CJGWe0E/pA==	AxzAA36BbQ=
3	2024-06-11 11:55:36 UTC	2024-06-11 11:56:03 UTC	LbGYQjw3NSbTAFK8qY...	CJGWe0E/pA==	AxzAA36BbQ=
4	2024-06-10 12:54:28 UTC	2024-06-10 12:56:03 UTC	gln2brzL9keudnYNr3Cbr...	CJGWe0E/pA==	AxzAA36BbQ=
5	2024-06-11 16:23:51 UTC	2024-06-11 16:25:03 UTC	k52ZvGbjZoF8wt52iGND...	CJGWe0E/pA==	AxzAA36BbQ=
6	2024-06-09 12:19:31 UTC	2024-06-09 12:20:04 UTC	WDYk3QaDgbM1XUFmP...	CJGWe0E/pA==	AxzAA36BbQ=

Resultados por página: 50 1 - 50 de 1000

< < > >

RESUMO

Nada foi selecionado no momento

Histórico de jobs

ATUALIZAR

Aqui eu já consigo ver a formatação de cada coluna e como os dados estão apresentados. Depois, sigo para entender a dimensão da tabela com a seguinte query.

1.2. Contagem total de registros

```
SELECT COUNT(*) AS total_registros
FROM rj-cetrio.desafio.readings_2024_06;
```

The screenshot shows the Google Cloud BigQuery interface. On the left, the 'Explorer' pane displays the project hierarchy: 'case-civitas' > 'desafio' > 'readings_2024_06'. The main editor shows a SQL query: `SELECT COUNT(*) AS total_registros FROM rj-cetrio.desafio.readings_2024_06;`. The 'Results' pane at the bottom displays the query output as a table with one row and one column, 'total_registros', with the value 363585336.

Linha	total_registros
1	363585336

Aqui podemos continuar explorando os dados por coluna. Por exemplo, fazer uma contagem por tipo de veículo para termos um melhor entendimento dos dados apresentados.

1.3. Contagem de registros por tipo de veículo

```
SELECT tipoveiculo, COUNT(*) AS total
FROM rj-cetrio.desafio.readings_2024_06
GROUP BY tipoveiculo
ORDER BY total DESC;
```

The screenshot shows the Google Cloud BigQuery interface. On the left, the 'Explorer' pane displays the project hierarchy: 'case-civitas' > 'desafio' > 'readings_2024_06'. The main editor shows a SQL query titled 'Consulta sem título'. Below the query editor, the 'Resultados da consulta' (Query Results) pane is active, displaying a table with 4 rows and 2 columns: 'tipoveiculo' and 'total'.

Linha	tipoveiculo	total
1	4uACn8DT5Q==	34386894
2	emN29HypFQ==	1157096
3	AxzAA36BbQ==	482850
4	uIZSERCZ7Q==	331696

At the bottom of the results pane, there is a 'Histórico de jobs' (Jobs History) section with an 'ATUALIZAR' (Refresh) button.

Outra possibilidade de análise para identificar inconsistências nos dados é comparar a `data_hora` com a data de captura. Através da query abaixo, podemos identificar o número de casos a serem tratados nesta situação. Neste caso específico, por se tratar de 495 mil casos, não recomendo a exclusão do modelo. Possivelmente exista um cenário onde a `data_hora` possa ser maior que a data de captura, que eu não esteja considerando.

1.4. Verificação de inconsistências nas datas

```
SELECT
  COUNT(*) AS inconsistencias_data
FROM rj-cetrio.desafio.readings_2024_06
WHERE datahora > datahora_captura;
```

The screenshot shows the Google Cloud BigQuery interface. On the left, the 'Explorer' pane displays a project named 'case-civitas' with a dataset 'desafio' containing a table 'readings_2024_06'. The main editor shows a SQL query titled 'Consulta sem título' with the following code:

```
1 SELECT
2   COUNT(*) AS inconsistencias_data
3 FROM rj-cetrio.desafio.readings_2024_06
4 WHERE datahora > datahora_captura;
5
```

Below the query editor, the 'Resultados da consulta' (Query Results) section is visible. It includes tabs for 'INFORMAÇÕES DO JOB', 'RESULTADOS' (selected), 'GRÁFICO', 'JSON', 'DETALHES DA EXECUÇÃO', and 'GRÁFICO DE ERROS'. The 'RESULTADOS' tab shows a table with one row of data:

Linha	inconsistencias_data
1	495797

At the bottom of the results section, there is a 'RESUMO' (Summary) tab which currently shows 'Nada foi selecionado no momento' (Nothing was selected at the moment).

Além disso, podemos buscar variáveis numéricas e executar análises de distribuição para identificar padrões. Pessoalmente, esse tipo de análise eu costumo fazer em Python para aproveitar as bibliotecas de plotagem de dados, mas em SQL se faz da seguinte forma.

1.5. Distribuição da velocidade dos veículos

```
SELECT velocidade, COUNT(*) AS total
FROM rj-cetrio.desafio.readings_2024_06
GROUP BY velocidade
ORDER BY velocidade DESC;
```

Case-CIVITAS

Pesquise (/) recursos, documentos, produtos e muito mais

Pesquisa

SANDBOX Configure o faturamento para fazer upgrade para a experiência completa do BigQuery. Saiba mais

DISPENSAR FAZER UPGRADE

Explorer

Q Digite para pesquisar

Você está visualizando os recursos.

MOSTRAR APENAS COM ESTRELA

- case-civitas
- datario
- rj-cetrio
 - Conexões externas
 - desafio
 - readings_2024_06

Consulta sem título

EXECUTAR SALVAR FAZER O DOWNLOAD Consulta concluída

```

1 SELECT velocidade, COUNT(*) AS total
2 FROM rj-cetrio.desafio.readings_2024_06
3 GROUP BY velocidade
4 ORDER BY velocidade DESC;

```

Resultados da consulta

SALVAR RESULTADOS EXPLORAR DADOS

INFORMAÇÕES DO JOB RESULTADOS GRÁFICO JSON DETALHES DA EXECUÇÃO GRÁFICO DE E

Linha	velocidade	total
1	255	32
2	254	50
3	253	22
4	252	27
5	251	42
6	250	26

Resultados por página: 50 1 - 50 de 256

Histórico de jobs ATUALIZAR

Identificação de Placas Clonadas

Para identificar possíveis placas clonadas, devemos procurar por situações onde a mesma placa foi registrada em locais diferentes (latitude e longitude) com um intervalo de tempo muito curto, o que seria impossível fisicamente para o mesmo veículo.

```

WITH veiculos AS (
  SELECT
    placa,
    datahora,
    camera_latitude,
    camera_longitude,
    LEAD(datahora) OVER (PARTITION BY placa ORDER BY datahor

```

```

a) AS prox_datahora,
    LEAD(camera_latitude) OVER (PARTITION BY placa ORDER BY d
atahora) AS prox_camera_latitude,
    LEAD(camera_longitude) OVER (PARTITION BY placa ORDER BY
datahora) AS prox_camera_longitude
    FROM rj-cetrio.desafio.readings_2024_06
),
distancias AS (
    SELECT
        placa,
        datahora,
        prox_datahora,
        ST_DISTANCE(ST_GEOGPOINT(camera_longitude, camera_latitud
e), ST_GEOGPOINT(prox_camera_longitude, prox_camera_latitud
e)) AS distancia
    FROM veiculos
    WHERE prox_datahora IS NOT NULL
)
SELECT
    placa,
    COUNT(*) AS vezes_clonada
FROM distancias
WHERE distancia > 1000 AND TIMESTAMP_DIFF(prox_datahora, data
hora, SECOND) < 600
GROUP BY placa
ORDER BY vezes_clonada DESC;

```

