Case - Time de Dados e Analytics

Thaís Souza Godoy

1.1. Solução com SQL:

```
-- Verificação da quantidade de linhas da tabela e suas diferenças
```

SELECT

```
*
,TB_LOCAL - TB_GCP AS DIFERENCA

FROM( SELECT

(SELECT COUNT(*) FROM application_record_local) AS TB_LOCAL, (SELECT COUNT(*) FROM application_record_gcp) AS TB_GCP
);
```

Diferença absoluta de linhas entre as tabelas local e GCP:



-- Verificação da quantidade de ID distintos da tabela e suas diferenças

SELECT

```
*
,TB_LOCAL - TB_GCP AS DIFERENCA_ID_DISTINTOS

FROM( SELECT

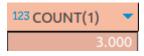
(SELECT COUNT(DISTINCT ID) FROM application_record_local) AS TB_LOCAL,
(SELECT COUNT(DISTINCT ID) FROM application_record_gcp) AS TB_GCP
);
```

Diferença da quantidade de IDs distintos entre as tabelas:

```
123 TB_LOCAL ▼ 123 TB_GCP ▼ 123 DIFERENCA_ID_DISTINTOS ▼ 438.510 434.459 4.051
```

```
É possível concluir que existe uma diferença absoluta de 1051 linhas de uma tabela para outra. No entanto, há 3000 linhas repetidas na tabela do GCP. Há duplicidade de IDs na tabela GCP, se considerarmos apenas os ID distintos.
```

```
-- Lista do ID que tem na tabela local e não tem no GCP
SELECT tb local.ID
FROM application record local tb local
LEFT JOIN application record gcp tb gcp
ON tb local.ID = tb gcp.ID
WHERE tb gcp.ID IS NULL;
Retorna os IDs que tem na tabela local e não na do GCP.
-- Lista de ID que tem no GCP e não tem no local
SELECT tb local.ID
FROM application record gcp tb gcp
LEFT JOIN application record local tb local
ON tb local.ID = tb gcp.ID
WHERE tb local.ID IS NULL;
Com o retorno desta consulta, conclui-se que todos os IDs que estão no GCP
estão presentes também na tabela local.
-- Contagem de IDs duplicados no GCP
SELECT COUNT (1) FROM (
SELECT ID, COUNT (1)
FROM application record gcp tb gcp
GROUP BY ID
HAVING COUNT (1) > 1
)
```



Número de IDs duplicados na tabela do GCP.

```
-- IDs duplicados no GCP
SELECT ID, COUNT (1)
FROM application_record_gcp tb_gcp
GROUP BY ID
HAVING COUNT (1) > 1
ORDER BY COUNT (1) DESC
Retorna quais IDs estão duplicados e quantas vezes aparecem na tabela.
-- Análise de padrão das colunas das duas tabelas
SELECT tb local.*
FROM application_record_local tb_local
LEFT JOIN application_record_gcp tb_gcp
ON tb local.ID = tb gcp.ID
WHERE tb gcp.ID IS NULL;
O retorno desta consulta mostra que dos IDs que não estão presentes na tabela
GCP, todos possuem o campo FLAG MOBIL como 1. Não há mais informações sobre
esse campo para que uma análise mais profunda seja feita.
SELECT COUNT (1)
FROM application record local tb local
LEFT JOIN (SELECT DISTINCT * FROM application record gcp) tb gcp
ON tb local.ID = tb gcp.ID
AND tb local.CODE GENDER = tb gcp.CODE GENDER
AND tb local.FLAG OWN CAR = tb gcp.FLAG OWN CAR
AND tb_local.FLAG_OWN_REALTY = tb_gcp.FLAG_OWN_REALTY
AND tb local.CNT CHILDREN = tb gcp.CNT CHILDREN
AND tb local.AMT INCOME TOTAL = tb gcp.AMT INCOME TOTAL
```

```
AND tb_local.NAME_INCOME_TYPE = tb_gcp.NAME_INCOME_TYPE

AND tb_local.NAME_EDUCATION_TYPE = tb_gcp.NAME_EDUCATION_TYPE

AND tb_local.NAME_FAMILY_STATUS = tb_gcp.NAME_FAMILY_STATUS

AND tb_local.NAME_HOUSING_TYPE = tb_gcp.NAME_HOUSING_TYPE

AND tb_local.DAYS_BIRTH = tb_gcp.DAYS_BIRTH

AND tb_local.DAYS_EMPLOYED = tb_gcp.DAYS_EMPLOYED

AND tb_local.FLAG_MOBIL = tb_gcp.FLAG_MOBIL

AND tb_local.FLAG_WORK_PHONE = tb_gcp.FLAG_WORK_PHONE

AND tb_local.FLAG_PHONE = tb_gcp.FLAG_PHONE

AND tb_local.FLAG_EMAIL = tb_gcp.FLAG_EMAIL

AND tb_local.OCCUPATION_TYPE = tb_gcp.OCCUPATION_TYPE

AND tb_local.CNT_FAM_MEMBERS = tb_gcp.CNT_FAM_MEMBERS
```

WHERE tb gcp.ID IS NOT NULL;

Por meio da consulta ON + AND para cada campo foi possível identificar quais colunas apresentam diferenças entre as tabelas. As colunas CODE GENDER, AMT INCOME TOTAL, DAYS BIRTH e OCCUPATION TYPE tem diferenças de formatação da tabela local para a GCP. A coluna FLAG WORK PHONE apresenta inconsistências na tabela GCP, parte dos dados não foram migrados e não há um padrão que permita algum tipo de tratamento.

Exemplo de inconsistências encontradas nas tabelas:

| ABC CODE_GENDER | ▼ ABC CODE_GENDER | • | 123 AMT_INCOME_TOTAL | • | 123 AMT_INCOME_TOTAL | • |
|-----------------|-------------------|---|----------------------|-----|----------------------|-----|
| М | Male | | 427. | 500 | 42.750.0 | 000 |
| M | Male | | 427. | 500 | 42.750.0 | 000 |
| M | Male | | 112. | 500 | 11.250.0 | 000 |

-- Tratando a tabela do GCP

SELECT DISTINCT

ID,

CASE WHEN CODE_GENDER = 'Male' THEN 'M'
WHEN CODE GENDER = 'Female' THEN 'F'

```
ELSE NULL END AS CODE GENDER,
FLAG OWN CAR,
FLAG_OWN_REALTY,
CNT CHILDREN,
AMT INCOME TOTAL/100 AS AMT INCOME TOTAL,
NAME INCOME TYPE,
NAME EDUCATION TYPE,
NAME FAMILY STATUS,
NAME HOUSING TYPE,
DAYS BIRTH*(-1) AS DAYS BIRTH,
DAYS EMPLOYED,
FLAG MOBIL,
FLAG WORK PHONE,
FLAG_PHONE,
FLAG EMAIL,
CASE WHEN OCCUPATION TYPE = 'Without Occupation' THEN ''
ELSE OCCUPATION TYPE END AS OCCUPATION TYPE,
CNT FAM MEMBERS
FROM application_record_gcp;
```

As diferenças foram tratadas com essa query, para que o dado fosse apresentado da mesma forma que na tabela origem local.

1.2. Solução com Python

*Os desenvolvimentos em Python foram feitos usando Pyspark por conta dos meus estudos atuais na pós-graduação, mas poderiam ser feitos em Pandas.

O código Python para solucionar o problema foi desenvolvido em um notebook e apresenta os mesmos resultados obtidos pela consulta SQL. Para executar o arquivo `.ipynb`, é necessário ter o PySpark instalado. Além disso, comentários explicativos foram incluídos ao longo do código para facilitar a compreensão.

- verificando a quantidade de linha e diferenças entre as tabelas:

DIFERENCA: 1051 TB_LOCAL: 438510 TB_GCP: 437459

- verificando a quantidade de IDs distintos da tabela e suas diferenças:

TB_GCP_DISTINCT: 434459
TB_LOCAL_DISTINCT: 438510
DIFERENCA_ID_DISTINTOS: 4051

- lista de IDs que aparecem na tabela local e não na GCP:

| ID C | | | | | | | | | | | | | |
|------------|-------------------|---------------|-------------------|-------------|---------------------|-----------------|--------------|------------|---------------------|-----------|--------------|---------------|-----------------|
| | ODE_GENDER FLAG_0 | WN_CAR FLAG_0 | WN_REALTY CNT_CHI | LDREN AMT | _INCOME_TOTAL NAME | _INCOME_TYPE NA | ME_EDUCATION | ON_TYPE | NAME_FAMILY_STATUS | NAME_HO | USING_TYPE D | AYS_BIRTH DAY | S_EMPLOYED FLAG |
| ++- | | | | | ******* | | | | | + | | | 2052421 |
| 5033986 | <u> </u> | N I | N | 91 | 144000.0 | Pensioner | Lower se | | | | apartment | -21675 | 365243 |
| 5045921 | <u> </u> | N I | | 9 | 40500.0 | Pensioner | Lower se | | | | apartment | -23933 | 365243 |
| 5052997 | <u> </u> | N I | | 9 | 94500.0 | Pensioner | Lower se | | Separated | | apartment | -22926 | 365243 |
| 5087743 | | N I | | 0 | 112500.0 | Pensioner | Lower se | | Married | | apartment | -24821 | 365243 |
| 5126743 | | N | N | 0 | 135000.0 | Working | Lower se | | Married | | apartment | -19019 | |
| 5260593 | F | N | N | 0 | 225000.0 | Working | Lower se | | Widow | | apartment | -21811 | -2563 |
| 5354898 | M | | | | 247500.0 | Working | Lower se | | Married | House / | apartment | -20950 | -196 |
| 5427531 | | | | | 135000.0 | Pensioner | Lower se | condary | Married | House / | apartment | -24210 | 365243 |
| 5467022 | | | | | 112500.0 Commercia | al associate | Lower se | condary S: | ingle 🖊 not married | Municipal | apartment | -12824 | |
| 5501357 | | | | | 135000.0 | Working | Lower se | condary | Civil marriage | House / | apartment | -14550 | -516 |
| 5585006 | | | | | 112500.0 Commercia | al associate | Lower se | condary | Married | House / | apartment | -19742 | -4230 |
| 5702391 | | | | | 112500.0 | Working | Lower se | condary | Married | House / | apartment | -10123 | -190 |
| 5742776 | F | | | | 135000.0 | Pensioner | Lower se | condary | Married | House / | apartment | -22486 | 365243 |
| 5872792 | F | | | | 157500.0 | Pensioner | Lower se | condary | Married | Municipal | apartment | -22941 | 365243 |
| 5970743 | F | | | | 126000.0 | Pensioner | Lower se | condary | Widow | House / | apartment | -21595 | 365243 |
| 6009146 | FΪ | | | | 60750.0 | Pensioner | Lower se | condary | Widow | House / | apartment | -22653 | 365243 |
| 6060382 | | | | | 270000.0 | Working | Lower se | condary | Married | House / | apartmenti | -14200 | -467 |
| 6077889 | | | | | 202500.0 | Working | Lower se | condary | Married | House / | apartmenti | -14489 | -2377 |
| 6088956 | FΪ | Nİ | | Θİ | 202500.0 | Pensioner | Lower se | condary S: | ingle ∠ not married | House / | apartmenti | -23232 | 365243 |
| 6103539 | | | | | 112500.0 S | tate servant | Lower se | | Married | | apartment | -19238 | -3509 |
| | | | | | | | | | | | | | |
| only showi | ng top 20 rows | | | | | | | | | | | | |

- Contagem de IDs duplicados no GCP e ordenação:

```
Número de IDs com mais de uma ocorrência: 3000
| gcp_ID|ID_Count|
|5008957|
5009141
|5009198|
|5009628|
                2
|5010058|
|5010338|
|5010568|
                2|2|2|
|5010623|
|5010674|
|5010781|
|5010801|
                2
|5010820|
|5018477|
|5021662|
|5021682|
|5021738|
|5021818|
|5021849|
                2
[5021878]
|5021947|
only showing top 20 rows
```

- filtro para identificar as colunas diferentes:

| | _CODE_GENDER tb_gc | p_CODE_GENDER tb_local | AMT_INCOME_TOTAL tb_gcp | _AMT_INCOME_TOTAL tb_lo | cal_DAYS_BIRTH tb_gcp | _DAYS_BIRTH tb_local_ | FLAG_WORK_PHONE tb_gcp_FLA | G_WORK_PHONE |
|--------------------|--------------------|------------------------|-------------------------|-------------------------|-----------------------|-----------------------|----------------------------|--------------|
| | | | | | | | | |
| 5008895 | | Female | 297000.0 | 29700000.0 | -15519 | 15519 | 0 | |
| 5009246 | F | Female | 135000.0 | 13500000.0 | -14201 | 14201 | 0 | 0.0 |
| 5009335 | F | Female | 139500.0 | 13950000.0 | -16203 | 16203 | 0 | null |
| 5009766 | | Male | 135000.0 | 13500000.0 | -14118 | 14118 | | null |
| 5009999 | | Female | 157500.0 | 15750000.0 | -12330 | 12330 | | 0.0 |
| 5010368 | | Female | 157500.0 | 15750000.0 | -9800 | 9800 | | null |
| 5010949 | | Female | 270000.0 | 27000000.0 | -13768 | 13768 | | |
| 5021309 | | Male | 270000.0 | 27000000.0 | -16896 | 16896 | | |
| 5021967 | | Female | 103500.0 | 10350000.0 | -11720 | 11720 | | |
| 5022410 | | Female | 153000.0 | 15300000.0 | -21899 | 21899 | | 0.0 |
| 5022555 | F) | Female | 112500.0 | 11250000.0 | -11234 | 11234 | | 0.0 |
| 5022720 | | Male | 157500.0 | 15750000.0 | - 15202 | 15202 | | |
| [5022759] | | Female | 180000.0 | 18000000.0 | -9791 | 9791 | | |
| [5022800] | F | Female | 202500.0 | 20250000.0 | -11998 | 11998 | | |
| [5022836] | | Male | 202500.0 | 20250000.0 | -17262 | 17262 | | |
| [5022988] | F | Female | 76500.0 | 7650000.0 | -24311 | 24311 | | 0.0 |
| [5023261] | | Male | 180000.0 | 18000000.0 | -21202 | 21202 | | |
| [5023441] | F | Female | 180000.0 | 18000000.0 | -22730 | 22730 | | 1.0 |
| [5023454] | F | Female | 180000.0 | 18000000.0 | -20560 | 20560 | | |
| 5023617 | F | Female | 157500.0 | 15750000.0 | -10806 | 10806 | | |
| | | | | | | | | |
| only showing top 2 | 20 rows | | | | | | | |

- transformação da tabela GCP:

2. Código SQL

Na primeira questão desenvolvi o código no ambiente Dbeaver + Sqlite, nesta precisei de um ambiente mais robusto para executar a procedure, portanto, desenvolvi a consulta no ambiente do Bigquery.

Para a solução 2 foi necessária a criação de uma procedure, conjunto de instruções que funcionam como uma função encapsulando várias operações. Como a data referência para a análise precisaria ser uma variável, ela foi definida como o parâmetro para a execução da procedure.

Uma tabela temporária foi criada para facilitar a operação das consultas:

```
CREATE TEMP TABLE tb_dias_atraso AS (
SELECT

re.nm_revendedor, IF(dt_pagamento IS
NULL, DATE_DIFF(data_parametro, dt_vencimento, DAY
), MAX(DATE_DIFF(dt_pagamento, dt_vencimento, DAY ))) AS dias_atraso

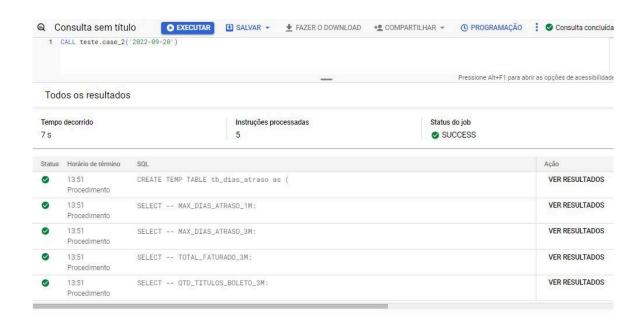
FROM teste.tb_revendedor re

LEFT JOIN teste.tb_titulos ti

ON re.id_revendedor = ti.id_revendedor

GROUP BY nm_revendedor, dt_pagamento, dt_vencimento);
```

Como resultado para a data parâmetro 2022-09-20, temos:



- tabela temporária:



- resultado para a consulta máximo dias de atraso em 1 mês:

| JSON | GRÁFICO | MAÇÕES DO JOB RESULTADOS GRÁF | | INFORMAÇÕES DO JOB | |
|------|-------------|-------------------------------|-----------------|--------------------|--|
| | DIAS_ATRASO | MAX_D | nm_revendedor ▼ | Linha / | |
| | 30 | | Ariel | 1 | |
| | 23 | | Lucas | 2 | |
| | 26 | | Mateus | 3 | |
| | 30 | | Renata | 4 | |
| | 25 | | Adriane | 5 | |
| | 30 | | Beatriz | 6 | |
| | 21 | | Gabriel | 7 | |
| | 29 | | Jessica | 8 | |
| | 28 | | Mariana | 9 | |
| | 28 | | Fernando | 10 | |

- resultado para a consulta máximo dias de atraso para 3 meses:

| INFOR | MAÇÕES DO JOB | RESULTADOS | GRÁFICO |
|----------|-----------------|------------|------------|
| Linha // | nm_revendedor ▼ | // MAX_D | IAS_ATRASO |
| 1 | Ariel | | 55 |
| 2 | Lucas | | 44 |
| 3 | Mateus | | 58 |
| 4 | Renata | | 54 |
| 5 | Adriane | | 59 |
| 6 | Beatriz | | 47 |
| 7 | Gabriel | | 57 |
| 8 | Jessica | | 53 |
| 9 | Mariana | | 58 |
| 10 | Fernando | | 54 |

- total faturado em 3 meses:

| INFOR | MAÇÕES DO JOB | RESULTADOS | GRÁFICO |
|---------|-----------------|----------------|---------|
| Linha / | nm_revendedor ▼ | // vlr_total ▼ | |
| 1 | Lucas | | 6890.6 |
| 2 | Fernando | | 5493.4 |
| 3 | Renata | | 7959.1 |
| 4 | Mariana | | 4797.0 |
| 5 | Beatriz | | 9889.7 |
| 6 | Mateus | | 4834.1 |
| 7 | Gabriel | | 8744.4 |
| 8 | Ariel | | 6248.0 |
| 9 | Adriane | | 8867.5 |
| 10 | Jessica | | 6955.2 |

- quantidade de boletos a prazo em 3 meses:

← Resultados da consulta

| INFOR | MAÇÕES DO JOB | RESULTADOS | GRÁFICO |
|---------|-----------------|------------|-----------|
| Linha / | nm_revendedor ▼ | / qtde_ti | tulo 🕶 // |
| 1 | Lucas | | 2 |
| 2 | Fernando | | 4 |
| 3 | Renata | | 4 |
| 4 | Mariana | | 4 |
| 5 | Beatriz | | 2 |
| 6 | Mateus | | 2 |
| 7 | Ariel | | 2 |
| 8 | Adriane | | 6 |
| 9 | Jessica | | 2 |

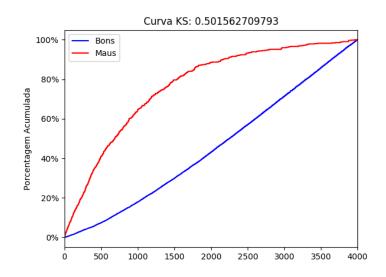
3. Código Python

No problema 3 foi implementada uma função que calcula a métrica KS (Kolmogorov-Smirnov). Medida usada para verificar a eficácia de um modelo de classificação, comparando as distribuições acumuladas de duas classes, a fim de identificar o maior desvio entre elas.

O código foi feito para ler um arquivo, passado como parâmetro para a função. Alguns outros parâmetros como nome das colunas também poderiam ser parametrizados, mas considerei que todos os arquivos possuem a mesma formatação.

É feita a contagem de 'BOM' e 'MAU'. Pelos valores do 'SCORE' foi aberta uma janela de ordenação para o cálculo da porcentagem acumulada da classificação binária de bons e maus.

O valor de KS foi calculado pela diferença absoluta entre os valores acumulados de 'BOM' e 'MAU'.



4. Código Python

No problema 4 foi desenvolvido um script 100% automatizado que retorna a tabela filtrada para o ano de 2023 e 2024:

| + | | | | | | | | + | + | | + | | |
|-----|---------|------|-----------|-----------|------|-----------|--------|---------|----------|------------|---------|-------------|---------|
| C0 | _AN0 C0 | _MES | CO_NCM CO | _UNID CO_ | PAIS | SG_UF_NCM | CO_VIA | CO_URF | QT_ESTAT | KG_LIQUIDO | VL_F0B | VL_FRETE VL | _SEGUR0 |
| + | | | | | | | | + | ++ | 4 | + | | |
| | 2023 | 01 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 333 | 333 | 7799.0 | 469 | 22 |
| | 2023 | 11 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 807 | 807 | 23778.0 | 107 | 34 |
| i i | 2023 | 09 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 24 | 24 | 986.0 | 4 | 1 |
| | 2023 | 06 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 1024 | 1024 | 29167.0 | 236 | 42 |
| | 2023 | 03 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 544 | 544 | 14606.0 | 172 | 37 |
| | 2023 | 05 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 164 | 164 | 4229.0 | 55 | 10 |
| | 2023 | 04 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 500 | 500 | 12619.0 | 209 | 31 |
| | 2023 | 10 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 76 | 76 | 3319.0 | 11 | 5 |
| | 2023 | 12 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 393 | 393 | 9909.0 | 55 | 14 |
| | 2023 | 08 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 499 | 499 | 9114.0 | 96 | 13 |
| Ī | 2023 | 07 3 | 3030010 | 10 | 275 | SP | 1 | 0817800 | 626 | 626 | 29725.0 | 136 | 43 |
| + | | +- | | | | | | + | + | | + | | |

| + | | | + | | | | | | | + | | |
|------|------------|--------|----------|-----------|-----------|------------|-------------|--------------|----------|---------|--------------|---------|
| C0 | _AN0 C0_ | MES | CO_NCM | CO_UNID C | O_PAIS SG | _UF_NCM CO | VIA CO_URF | QT_ESTAT K | _LIQUIDO | VL_F0B | VL_FRETE VL | _SEGUR0 |
| + | | | + | | | | | | | + | | |
| | 2024 | 02 3 | 33030010 | 10 | 275 | SP | 1 0817800 | 155 | 155 | 5076.0 | 33 | 7 |
| i i | 2024 | 07 3 | 33030010 | 10 | 275 | SP | 1 0817800 | 1076 | 1076 | 24415.0 | 120 | 35 |
| l i | 2024 | 06 3 | 33030010 | 10 | 275 | SP | 1 0817800 | 126 | 126 | 4542.0 | 24 | 7 |
| l i | 2024 | 04 | 33030010 | 10 | 275 | SPİ | 1 0817800 | 23 | 23 | 225.0 | 5 į | Θİ |
| l i | 2024 | 05 | 33030010 | 10 | 275 | SP | 1 0817800 | 1014 | 1014 | 33592.0 | 105 | 48 |
| l i | 2024 | 03 i 3 | 33030010 | 10 İ | 275 İ | SPİ | 1 0817800 | 58 | 58 İ | 2494.0 | 8 | 4 ј |
| l i | 2024 | | 33030010 | | 275 | SPİ | 1 0817800 | 282 | 282 | 6311.0 | 83 | 9 i |
| Ti . | 2024 | | 33030010 | | 275 | SP | 1 0817800 | 938 | | 41757.0 | | 60 |
| ļ | | | | | | | | | + | | ₁ | |

O script pode ser reexecutado a cada mês já que ainda não finalizamos o ano de 2024. O código faz o download da tabela direto do site do Governo, foi necessário um tratamento para essa primeira parte, porque o site está apresentando um problema de SSL. A solução temporária foi desabilitar a verificação do SSL, não é algo recomendado, mas impossibilitaria o script ser totalmente automático.

Após baixar o arquivo, o script faz o tratamento de algumas colunas, transformando-as em tipo int ou float, filtra o dataframe de acordo com os requisitos: país França, via navio, produto 33030010 e negociação no estado de SP.

As colunas **KG_LIQUIDO** e **VL_FOB** foram somadas, e a média anual do valor da negociação foi calculada. Mantive os valores em dólar para preservar a moeda de referência utilizada nas transações e evitar as flutuações cambiais do real. Isso garante uma análise mais precisa e consistente do valor negociado, evitando distorções que poderiam ocorrer, como no caso de variações no câmbio em 2023, por exemplo.

Média em 2023: U\$29,11

Média em 2024: U\$32,23