Data Engineering Bootcamp

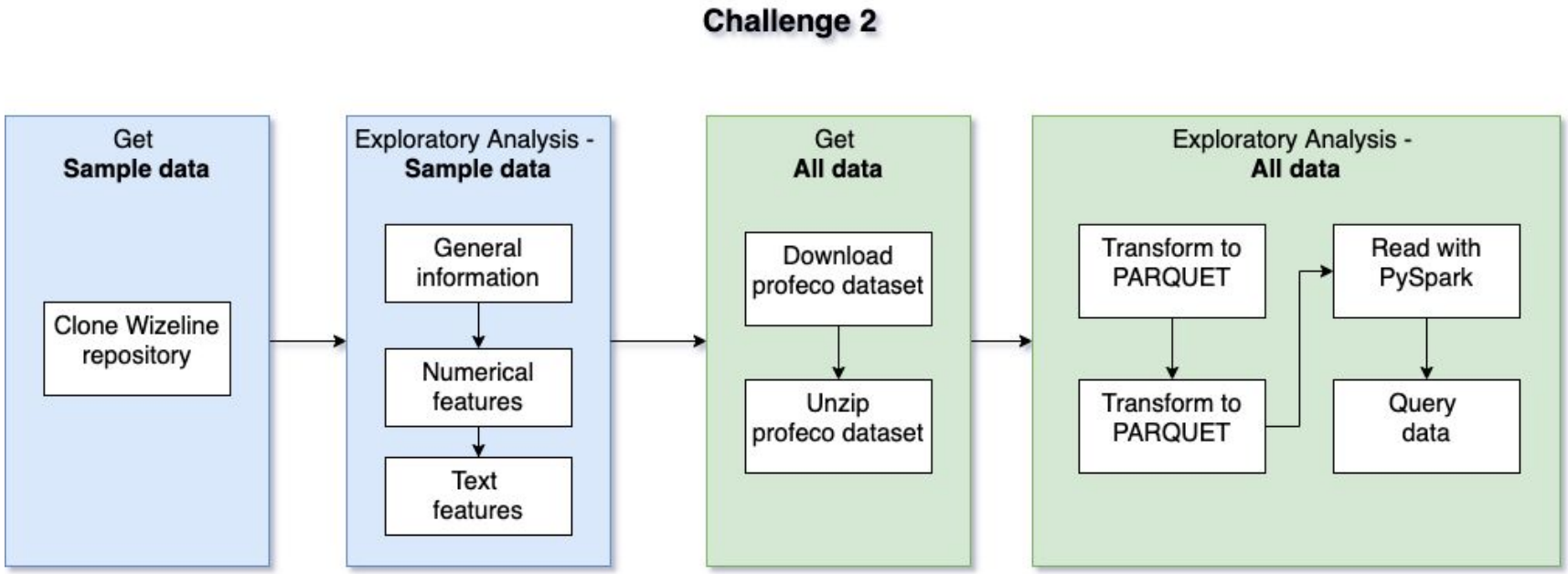
Technical Challenge Part 2 - Practical Applications

Carlos Lopez

Content

1. Data processing approach
2. Introduction - Exploratory Analysis
3. Questions

Data processing approach



Introduction - Exploratory Analysis

The Customer Service team at Profeco (Mexican Consumer Protection Agency) wants to analyze the monitored products in Mexico. The IT team downloaded the database into an Google Drive on a CSV file of about 20GB.

Your task as a Data Engineer is processing the data and creating an exploratory analysis with Python Pandas without using pure Python functions.

Questions

1. How many commercial chains are monitored, and therefore, included in this database?
2. What are the top 10 monitored products by State?
3. Which is the commercial chain with the highest number of monitored products?
4. Use the data to find an interesting fact.
5. What are the lessons learned from this exercise?
6. Can you identify other ways to approach this problem? Explain.

Question 1.

How many commercial chains are monitored, and therefore, included in this database?

There are **520** commercial chains monitored in the database

```
1 from pyspark.sql.functions import countDistinct
2 df_commercial_chains = parDF.select(countDistinct("cadenaComercial"))
3 df_commercial_chains.show()
```

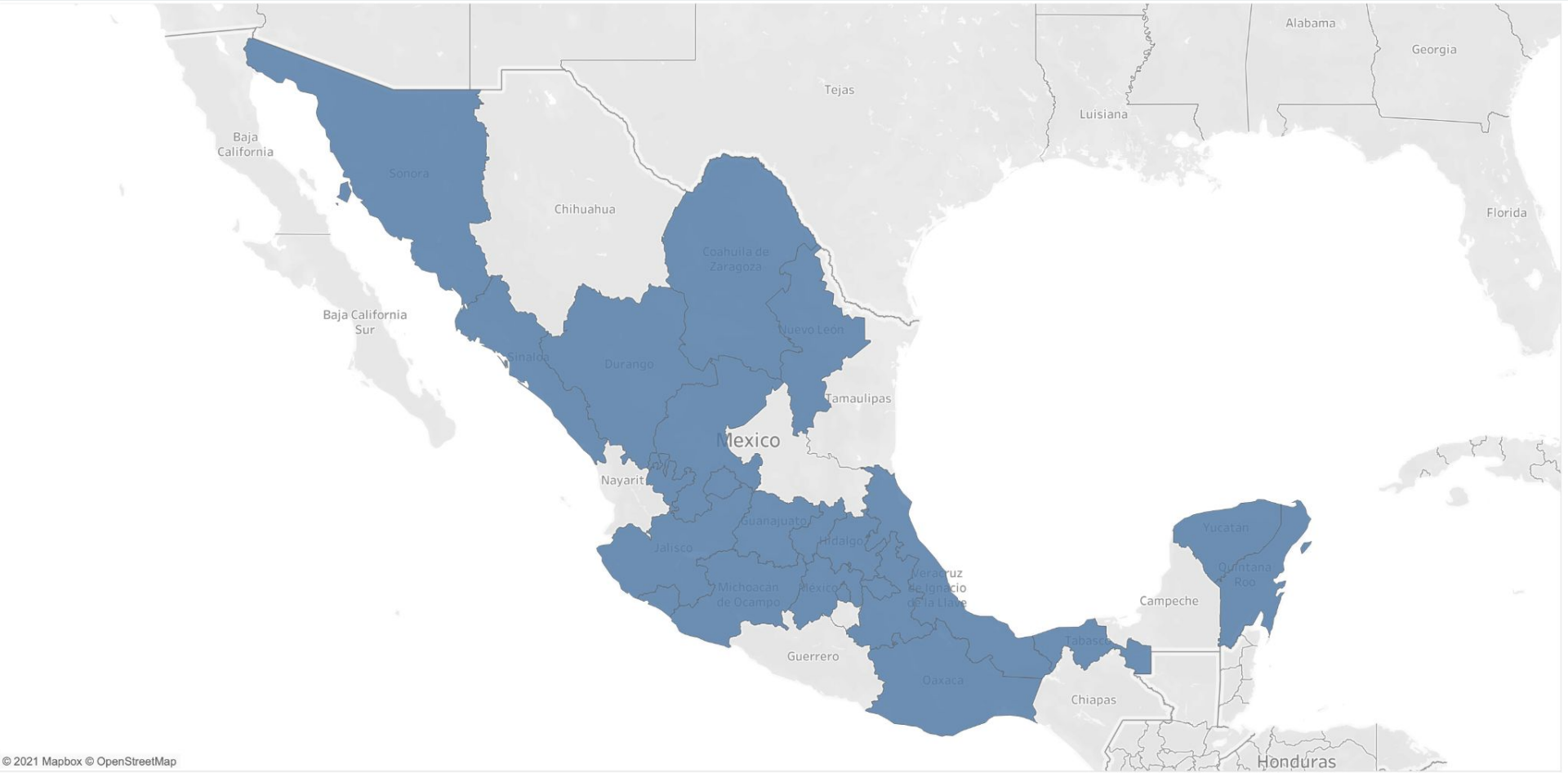
```
+------------------------------+
|count(DISTINCT cadenaComercial)|
+------------------------------+
|                           520|
+------------------------------+
```

Question 2.

What are the top 10 monitored products by State?

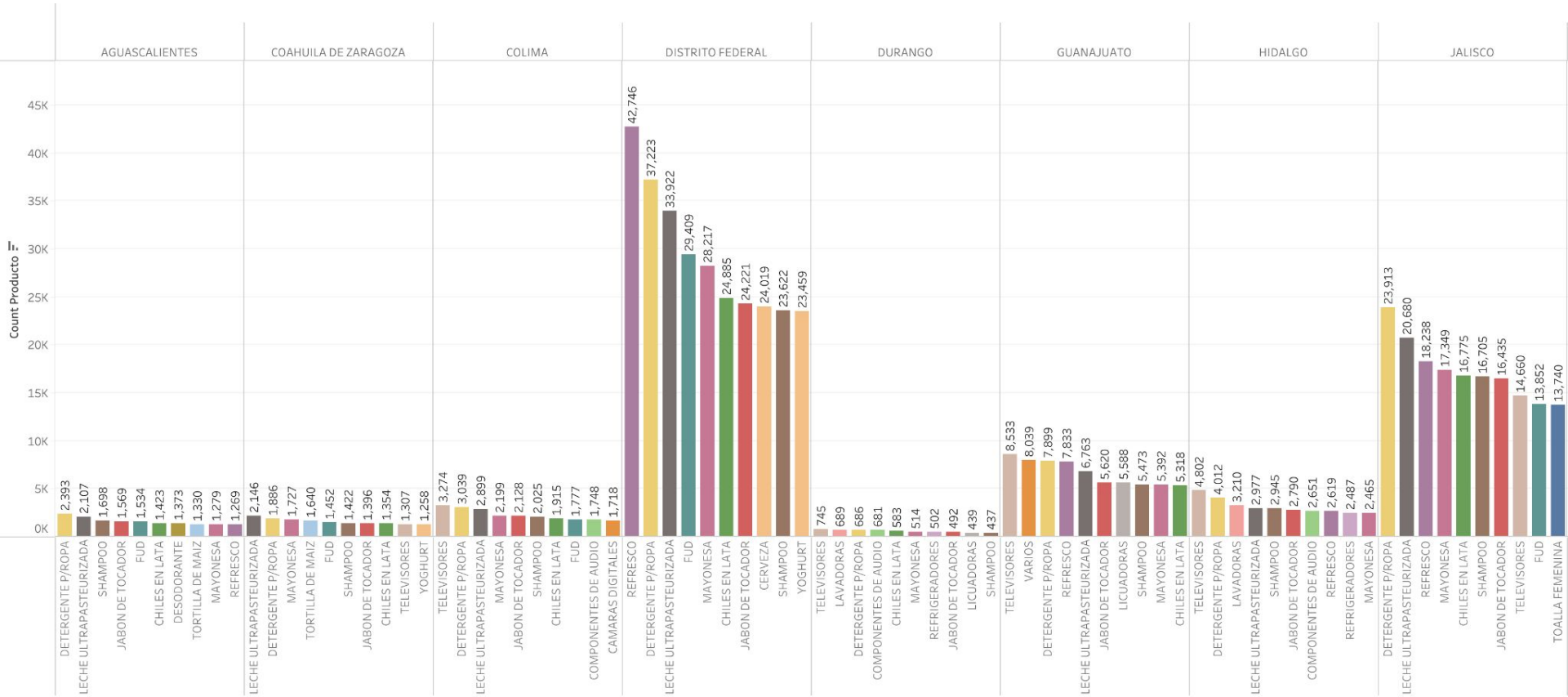
QUINTANA ROO
NUEVO LEÓN
SINALOA
TABASCO
TLAXCALA
COAHUILA DE ZARAGOZA
VERACRUZ DE IGNACIO DE LA LLAVE
SONORA
YUCATÁN
MICHOACÁN DE OCAMPO

DURANGO
DISTRITO FEDERAL
HIDALGO
ZACATECAS
GUANAJUATO
AGUASCALIENTES
OAXACA
PUEBLA
JALISCO
QUERÉTARO
COLIMA
MÉXICO

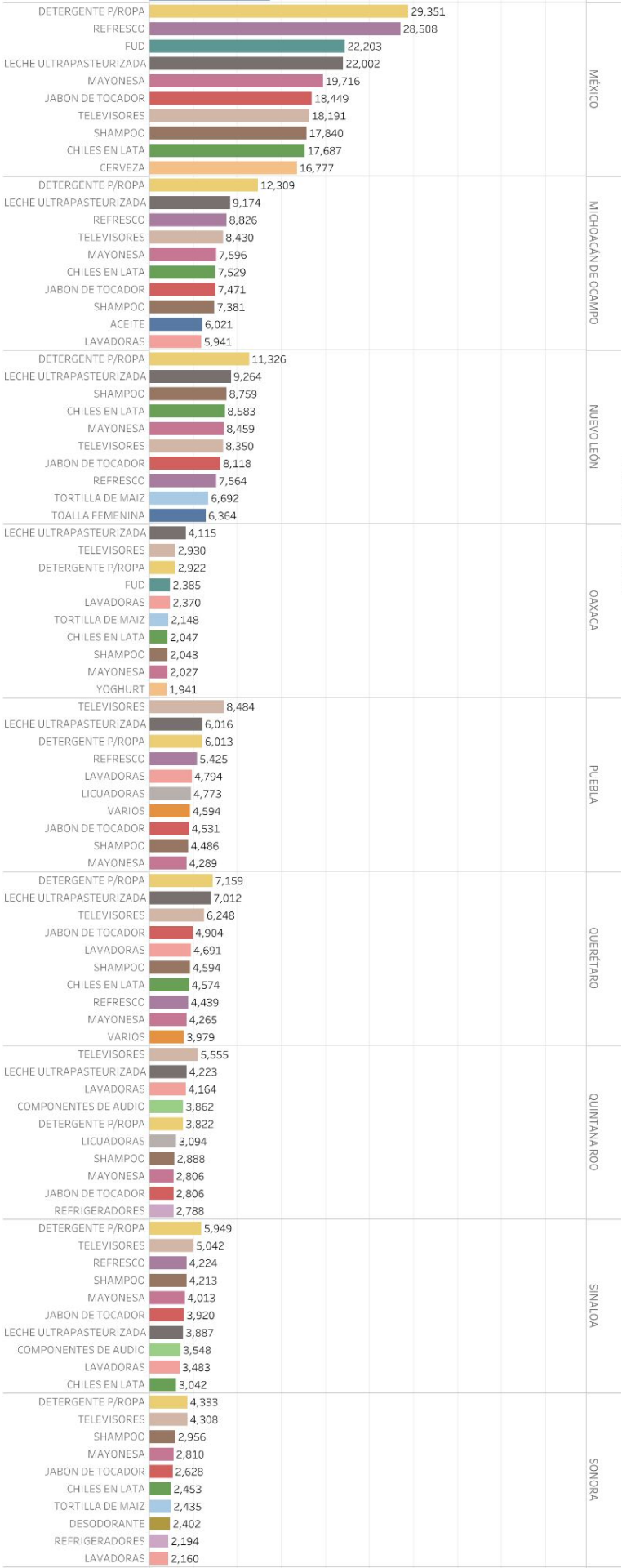


Mapa basado en Longitud (generado) y Latitud (generado). Se muestran detalles para PaisyEstado.

Top monitored products



Suma de Count Producto para cada Producto desglosado por Estado (grupo). El color muestra detalles acerca de Producto.


MÉXICO
DETERGENTE P/ROPA 29,351
REFRESCO 28,508
FUD 22,203
LECHE ULTRAPASTEURIZADA 22,002
MAYONESA 19,716
JABON DE TOCADOR 18,449
TELEVISORES 18,191
SHAMPOO 17,840
CHILES EN LATA 17,687
CERVEZA 16,777
MICHOACÁN DE OCAMPO
DETERGENTE P/ROPA 12,309
LECHE ULTRAPASTEURIZADA 9,174
REFRESCO 8,826
TELEVISORES 8,430
MAYONESA 7,596
CHILES EN LATA 7,529
JABON DE TOCADOR 7,471
SHAMPOO 7,381
ACEITE 6,021
LAVADORAS 5,941
NUEVO LEÓN
DETERGENTE P/ROPA 11,326
LECHE ULTRAPASTEURIZADA 9,264
SHAMPOO 8,759
CHILES EN LATA 8,583
MAYONESA 8,459
TELEVISORES 8,350
JABON DE TOCADOR 8,118
REFRESCO 7,564
TORTILLA DE MAIZ 6,692
TOALLA FEMENINA 6,364
OAXACA
LECHE ULTRAPASTEURIZADA 4,115
TELEVISORES 2,930
DETERGENTE P/ROPA 2,922
FUD 2,385
LAVADORAS 2,370
TORTILLA DE MAIZ 2,148
CHILES EN LATA 2,047
SHAMPOO 2,043
MAYONESA 2,027
YOGHURT 1,941
PUEBLA
TELEVISORES 8,484
LECHE ULTRAPASTEURIZADA 6,016
DETERGENTE P/ROPA 6,013
REFRESCO 5,425
LAVADORAS 4,794
LICUADORAS 4,773
VARIOS 4,594
JABON DE TOCADOR 4,531
SHAMPOO 4,486
MAYONESA 4,289
QUERÉTARO
DETERGENTE P/ROPA 7,159
LECHE ULTRAPASTEURIZADA 7,012
TELEVISORES 6,248
JABON DE TOCADOR 4,904
LAVADORAS 4,691
SHAMPOO 4,594
CHILES EN LATA 4,574
REFRESCO 4,439
MAYONESA 4,265
VARIOS 3,979
QUINTANA ROO
TELEVISORES 5,555
LECHE ULTRAPASTEURIZADA 4,223
LAVADORAS 4,164
COMPONENTES DE AUDIO 3,862
DETERGENTE P/ROPA 3,822
LICUADORAS 3,094
SHAMPOO 2,888
MAYONESA 2,806
JABON DE TOCADOR 2,806
REFRIGERADORES 2,788
SINALOA
DETERGENTE P/ROPA 5,949
TELEVISORES 5,042
REFRESCO 4,224
SHAMPOO 4,213
MAYONESA 4,013
JABON DE TOCADOR 3,920
LECHE ULTRAPASTEURIZADA 3,887
COMPONENTES DE AUDIO 3,548
LAVADORAS 3,483
CHILES EN LATA 3,042
SONORA
DETERGENTE P/ROPA 4,333
TELEVISORES 4,308
SHAMPOO 2,956
MAYONESA 2,810
JABON DE TOCADOR 2,628
CHILES EN LATA 2,453
TORTILLA DE MAIZ 2,435
DESODORANTE 2,402
REFRIGERADORES 2,194
LAVADORAS 2,160
Estado (grupo) / Producto

SONORA
TABASCO
TLAXCALA
VERACRUZ DE IGNACIO DE LA LLAVE
YUCATÁN
ZACATECAS
DETERGENTE P/ROPA 4,333
TELEVISORES 4,308
SHAMPOO 2,956
MAYONESA 2,810
JABON DE TOCADOR 2,628
CHILES EN LATA 2,453
TORTILLA DE MAIZ 2,435
DESODORANTE 2,402
REFRIGERADORES 2,194
LAVADORAS 2,160
TELEVISORES 6,837
DETERGENTE P/ROPA 5,408
LECHE ULTRAPASTEURIZADA 4,568
VARIOS 4,347
LAVADORAS 4,100
CHILES EN LATA 3,972
COMPONENTES DE AUDIO 3,883
LICUADORAS 3,850
REFRESCO 3,736
CAMARAS DIGITALES 3,713
DETERGENTE P/ROPA 9,282
LECHE ULTRAPASTEURIZADA 7,487
JABON DE TOCADOR 6,544
SHAMPOO 6,531
REFRESCO 6,505
MAYONESA 5,961
CHILES EN LATA 5,857
TOALLA FEMENINA 5,273
ACEITE 5,119
CAFE SOLUBLE 4,694
LECHE ULTRAPASTEURIZADA 4,444
DETERGENTE P/ROPA 3,161
REFRESCO 2,690
TELEVISORES 2,689
SHAMPOO 2,252
JABON DE TOCADOR 2,159
MAYONESA 2,037
HUEVO 1,884
LAVADORAS 1,879
YOGHURT 1,828
TELEVISORES 7,795
LECHE ULTRAPASTEURIZADA 7,400
DETERGENTE P/ROPA 7,153
FUD 5,034
MAYONESA 4,941
SHAMPOO 4,859
COMPONENTES DE AUDIO 4,849
CHILES EN LATA 4,725
LAVADORAS 4,552
PLANCHAS 4,447
DETERGENTE P/ROPA 5,009
LECHE ULTRAPASTEURIZADA 3,845
MAYONESA 3,556
JABON DE TOCADOR 3,472
SHAMPOO 3,470
CHILES EN LATA 3,058
TELEVISORES 2,954
LICUADORAS 2,865
REFRESCO 2,817
FUD 2,783
Producto
ACEITE
CAFE SOLUBLE
CAMARAS DIGITALES
CERVEZA
CHILES EN LATA
COMPONENTES DE AUDIO
DESODORANTE
DETERGENTE P/ROPA
FUD
HUEVO
JABON DE TOCADOR
LAVADORAS
LECHE ULTRAPASTEURIZADA
LICUADORAS
MAYONESA
PLANCHAS
REFRESCO
REFRIGERADORES
SHAMPOO
TELEVISORES
TOALLA FEMENINA
TORTILLA DE MAIZ
VARIOS
YOGHURT

Question 3.

Which is the commercial chain with the highest number of monitored products?

```
[43]  1 ɔl('cadenaComercial').isNotNull()).groupBy('producto', 'cadenaComercial').count().select('cadenaComercial','producto', f.col('count').alias('count'))
      2 ascending=False).show(10, False)
```

```
+-------------------------+----------------------+-----+
|cadenaComercial          |producto              |count|
+-------------------------+----------------------+-----+
|TORTILLERIAS TRADICIONALES|TORTILLA DE MAIZ      |47264|
|WAL-MART                 |DETERGENTE P/ROPA     |27030|
|SORIANA                  |DETERGENTE P/ROPA     |25132|
|WAL-MART                 |REFRESCO              |23732|
|BODEGA AURRERA           |DETERGENTE P/ROPA     |22156|
|BODEGA AURRERA           |LECHE ULTRAPASTEURIZADA|22150|
|WAL-MART                 |LECHE ULTRAPASTEURIZADA|21779|
|BODEGA AURRERA           |REFRESCO              |20920|
|SORIANA                  |REFRESCO              |20237|
|WAL-MART                 |MAYONESA              |20081|
+-------------------------+----------------------+-----+
only showing top 10 rows
```

Question 4.

Use the data to find an interesting fact.

- According to the data inspection, there are 62,530,716 of rows.
- 52,530,710 are null data, this represents **84% out of 100%**.
- The most monitored presentation is for **1 KG. GRANEL** - 175,580
- The most monitored by all states brand is **“La Costena” (La Costeña)** - 212,308
    - **CHILES EN LATA** is the most consumed product from “La Costeña”
    - Followed by MAYONESA, maybe used for the “elotes con harto chile del que si pica”

Question 5.

What are the lessons learned from this exercise?

- Process large csv files take a lot of time if I have limited resources, specially with RAM memory (that's why I used PySpark instead of Pandas).
- PARQUET is the best ally when compressing data.
- It would be interesting to know how to stream 20GB (or maybe terabytes) instead of batching.
- I need to know more Data Architectures
- Google Colab is a good tool for a proof of concept
- "The more I know, the more I realize I don't know"... and want to learn more :)

Question 6.

Can you identify other ways to approach this problem? Explain.

- First one, load all data in memory - FAILED
  - Didn't work, Google Colab session has limited resources
- Deploy a more sophisticated environment that has:
  - A Spark session
  - Bucket to storage data
  - Partition data by specific columns
  - Orchestrator like Airflow to control the tasks executions
  - Send data processed to Data Warehouse such as Redshift or Bigquery
  - Connect the data warehouse to Tableau or any other visualization tool to perform the analysis