Technical Challenge Solution

Oscar Garcia

Q1. How Many commercial chains are monitored?

I calculated the number of unique commercial chains using the nunique function. There are 705 commercial chains.

```
df.cadenaComercial.nunique(dropna = True)
✓ 2.2s
705
```

Q2. What are the top 10 monitored products by State?

First I make a group by states and then I use a lambda function to pass the value_counts() method to each of the groups, then the nlargest method will select only the top 10 products for each of the groups. The result is too big to be displayed here but the information could, for example, be saved to a .csv file so it can be read.

```
df.groupby("estado").apply(lambda x: x.producto.value_counts(dropna = False).nlargest(10))
```

```
estado
AGUASCALIENTES  FUD                       12005
                DETERGENTE P/ROPA         10188
                LECHE ULTRAPASTEURIZADA    9824
                SHAMPOO                    9654
                REFRESCO                   9481
                                           ...
ZACATECAS       SHAMPOO                   15012
                CHILES EN LATA            14866
                COMPONENTES DE AUDIO      14799
                REFRESCO                  13925
estado          producto                     20
Name: producto, Length: 321, dtype: int64
```

Q3. Which is the commercial chain with the highest number of monitored products?

Similar to the last question, first I group by commercial chains and then I count the number of unique products, then I sort them in descending order and the method head() will select the top value which in this case is Soriana.

```
df.groupby("cadenaComercial").producto.nunique().sort_values(ascending=False).head(1)
```

```
cadenaComercial
SORIANA    1059
Name: producto, dtype: int64
```

Q4. Use the data to find an interesting fact

t first I was trying to analyze the distribution of expensive products in cities when I found that there are repeated municipio names in some states, it can be seen that "ZACATECAS" appear twice, there is probably some extra space in the data entry.
Fixing this error would imply going state by state to find repeated names and fixing them manually.

```
#All products
df.groupby("estado").municipio.value_counts(normalize = True)
```

```
estado           municipio
AGUASCALIENTES   AGUASCALIENTES     0.608401
                 AGUASCALIENTES     0.391599
BAJA CALIFORNIA  MEXICALI           0.351807
                 ENSENADA           0.217337
                 TIJUANA            0.213059
                                       ...
ZACATECAS        ZACATECAS          0.563561
                 ZACATECAS          0.263421
                 GUADALUPE          0.135417
                 GUADALUPE          0.037601
estado           municipio          1.000000
Name: municipio, Length: 210, dtype: float64
```

Q5. What are the lessons learned from this exercise?

Must ensure the data quality of the database before performing analysis. Failing to do this might invalidate an analysis and we might not catch this error before its too late.

Q6. Can you identify other ways to approach this problem? Explain.

If the file would be larger it would be impossible to run this program, I suspect that many are not able to load the complete file, to solve this we would need to process it in a distributed manner. Downloading the file to databricks and then running the analysis using PySpark APIs would allows to handle the whole file.

We can even use the Koalas library to reuse the pandas code on spark although it would run slower than using only Spark APIs.