

Data Engineer Bootcamp

Exploratory Analysis of monitored products in Mexico

Engineer: José Carlos Gallardo Aceves

12/10/2021

To solve this Project i created subsets of the main dataframe to make the program lighter tan if i used the complete dataframe. In each question i created a specific dataframe just with the columns i needed to succed.

Imagen que contiene Escala de tiempo

Descripción generada automáticamente

Figure 1: Diagram of proccess to   
solve the excersice.

The procedure to create a specific dataframe was to define a chunksize, in my case the chunsize was of 10,000,000 and then for every chunk in the csv the program is making the neccesary operations to add the solutions of every chunk into the past chunk data frame and make it lighter tan it woul be if the program proccess everything at the end of the for cyccle.

**Results**

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

The answer obtained was **706** commercial chains were included in the database.

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

The answer of this question is on the same folder as this document with the name “**answer2**”

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

The answer obtained for this question is “**TORTILLERIAS TRADICIONALES**”.

1. Use the data to find an interesting fact.

The fact that i wanted to show was to find the top 10 monitored commercial chains by their number of products and then calculate the percentage of products that this 10 commercial chains have in respect to all the data base.

The top 10 comercial chains and their number of products were:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **cadenaComercial** | **count** | | WAL-MART | 8643133 | | BODEGA AURRERA | 6751030 | | SORIANA | 6546211 | | MEGA COMERCIAL MEXICANA | 4899509 | | CHEDRAUI | 4221625 | | COMERCIAL MEXICANA | 2598272 | | SUPERAMA | 2590459 | | HIPERMERCADO SORIANA | 1706956 | | MERCADO SORIANA | 1533080 | | SORIANA SUPER | 1389901 | |  |

And the percentage of products in those 10 commercial chains were **%65.38** aproximately.

1. What are the lessons learned from this exercise?

One of the best leasson i learned is that pandas is not the best way to analyze bigdata, it can be slow and hard to proccess all the data. But in the other hand, i learned how to properly manage huge data frames by dividing in chunks and at the same time by only choosing the data that is going to help you solve the excercise and not waist extra memory in data that it’s not going to be used.

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

If I could choose another way to solve the problem I would go with spark, that could improve by far the processing velocity and even take better chances to analyze the complete data from the csv so I could get more relevant information and more completed analysis.