Analyzing the Impact of COVID-19 on the Spanish Fruit and Vegetable Market

INTRODUCTION

In this project we looked at the food industry when pandemic hit the world globally. COVID-19 had effects on the whole business world and their economy. It not only affected the business world, it also changed people's daily eating habits. Because of the pandemic, human beings' need has increased and their first instinct was to buy food and household products.

1. Aim

Our main goal of the project is to get a deeper understanding of how COVID-19 has affected the food industry and markets in Spain. We considered the markets who are named like Mercadona (and also MercaMadrid, MercaBarca).

a. **Analysis of the Consumer Behavior:** Investigate replacements in consumer conduct between the pandemic, particularly in the consumption of fruits and vegetables.
b. **Market Reaction:** To get a better view of how the Spanish market for fruits and vegetables navigated through the challenges posed by the COVID-19 crisis.

2. Statements and Hypotheses

Relationship between Consumption and Price: The consumption of fruits and vegetables is anticipated to have an inverse relationship with pricing changes.

Consumption during the Pandemic: It is hypothesized that during the initial months of the pandemic, there was an increase in fruit and vegetable consumption.

Relationship between Import and Price during the Pandemic: Import variations are expected to influence price fluctuations in the market.

In our exploration, we're delving into various aspects. In case of investigating whether higher prices lead to reduced purchases of fruits and vegetables. First, if fruits and veggies cost more, do people buy less of them? Second, did folks buy more fruits and veggies when COVID-19 started? Finally, did prices go up because Spain imported fewer fruits and veggies?

To figure this out, we're looking at data about what people bought, how prices changed every week, what got sold in markets, Spain's trade with other countries, and how many COVID-19 cases and deaths happened in Spain. All this info helps us understand how people shop, how markets work, and how COVID-19 affected things like fruit and veggie prices in Spain.

PHASES OF THE KDD PROCESS

The KDD process involves several parts. We can follow these several statements like down below:

- **Understanding the Business & Data**
- Trying to gain insights from the challenges and dynamics of the Spanish markets.
- Understanding how the pandemic affected the behavior of the consumers, market trends or supply chains.

- **Data Cleaning and Preprocessing**
- For this phase of the project we handled missing values and transformed the data into a better form.
- Cleaning the data that are related to sales, customer behavior.
- Trying to handle the dataset by looking if there’s missing values or inconsistency in case of dataset’s quality.

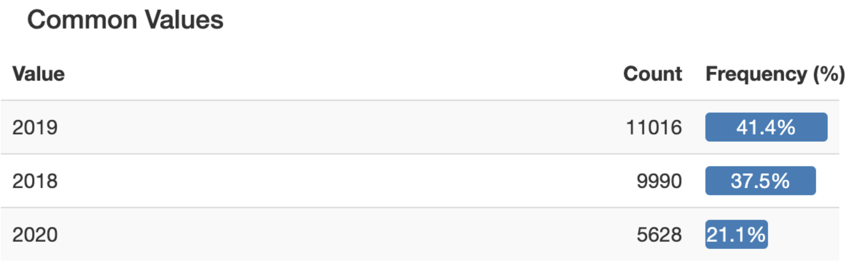
a. **Dataset 1 - DatosConsumoAlimentario**

Cleaning Steps:

- Convert month names in Spanish to numerical equivalents in the ‘Mes’
- Sort the data by year, month, and region.

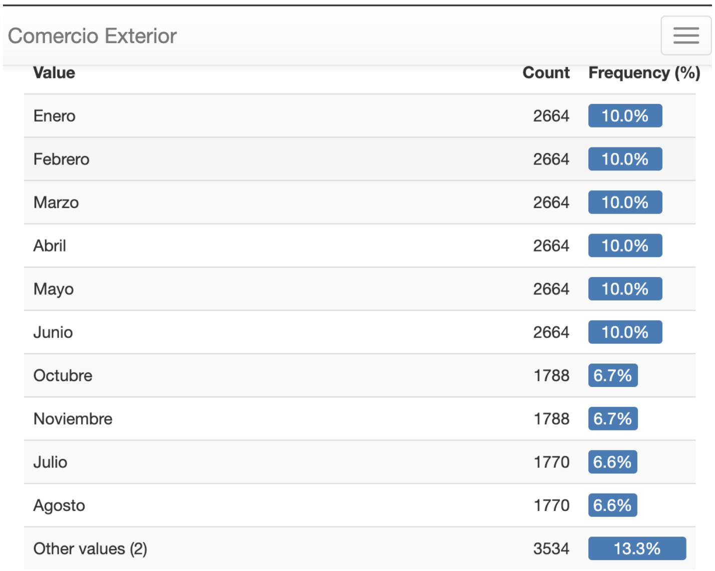
Preprocessing Steps:

- Convert to month names into number for a better understanding
- Added labels to Dataset 1 for a better comparison with Dataset 2.
- Seperated and saved the national data.0



Common Values

Value	Count	Frequency (%)
2019	11016	41.4%
2018	9990	37.5%
2020	5628	21.1%



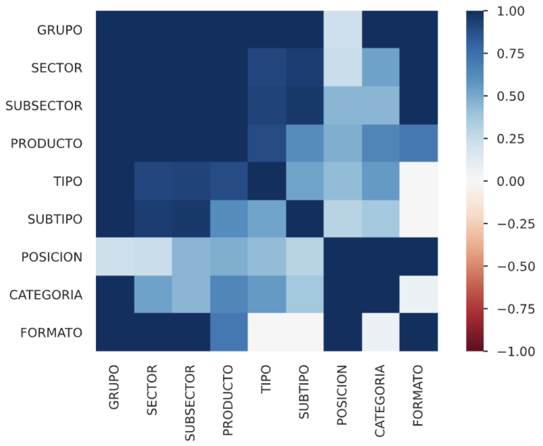
Comercio Exterior

Value	Count	Frequency (%)
Enero	2664	10.0%
Febrero	2664	10.0%
Marzo	2664	10.0%
Abril	2664	10.0%
Mayo	2664	10.0%
Junio	2664	10.0%
Octubre	1788	6.7%
Noviembre	1788	6.7%
Julio	1770	6.6%
Agosto	1770	6.6%
Other values (2)	3534	13.3%

b. Dataset 2 - PreciosSemanales

Cleaning Steps:

- Handled the missing values by removing the data where the ‘PRODUCTO’ column has no data
- Convert the data from weekly to monthly to merge them with the other data



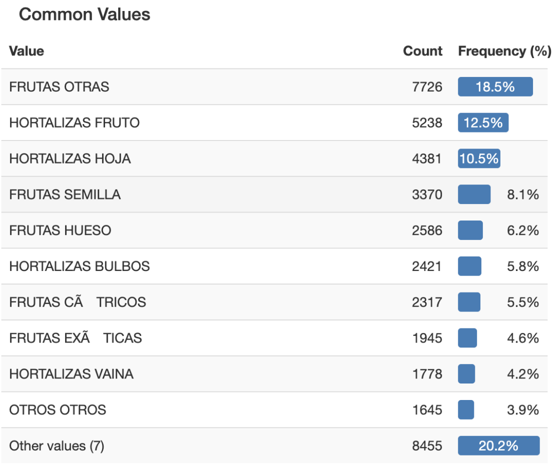
c. Dataset 3A & 3B - DatosMercaMadrid & DatosMercaBarna

Cleaning Steps:

- Removed the Unidad column which has a constant value
- Kept the categories like Other and Seta
- Commas have been replaced by dots

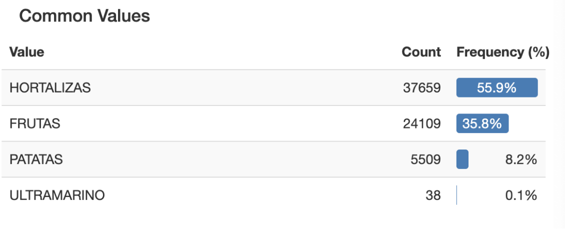
Preprocessing Steps:

- YEAR and MONTH combined into a single date column
- Defined the file paths for CSV outputs and saved the DataFrames to CSV files.



Common Values

Value	Count	Frequency (%)
FRUTAS OTRAS	7726	18.5%
HORTALIZAS FRUTO	5238	12.5%
HORTALIZAS HOJA	4381	10.5%
FRUTAS SEMILLA	3370	8.1%
FRUTAS HUESO	2586	6.2%
HORTALIZAS BULBOS	2421	5.8%
FRUTAS CÃ TRICOS	2317	5.5%
FRUTAS EXÃ TICAS	1945	4.6%
HORTALIZAS VAINA	1778	4.2%
OTROS OTROS	1645	3.9%
Other values (7)	8455	20.2%



Common Values

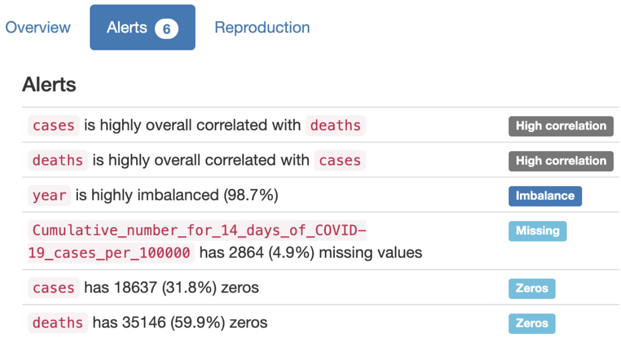
Value	Count	Frequency (%)
HORTALIZAS	37659	55.9%
FRUTAS	24109	35.8%
PATATAS	5509	8.2%
ULTRAMARINO	38	0.1%

d. Monthly Trade Data (Dataset 4 - Comercio Exterior de España)

Cleaning Steps:

- Rows with missing or duplicate values have been dropped
- Removed the 'PARTNER' column as it's unnecessary.
- Removed the null values, which in this dataset were signed with **“:”**.

e. Coronavirus Cases Data (Dataset 5):



Alerts

`cases` is highly overall correlated with `deaths`	High correlation
`deaths` is highly overall correlated with `cases`	High correlation
`year` is highly imbalanced (98.7%)	Imbalance
`Cumulative_number_for_14_days_of_COVID-19_cases_per_100000` has 2864 (4.9%) missing values	Missing
`cases` has 18637 (31.8%) zeros	Zeros
`deaths` has 35146 (59.9%) zeros	Zeros

- Replaced NaN values with 0.

- **Data İntegration**
- Coordinating information from different sources, including deals records, market reports, and client input, to make a complete dataset for examination.

Translation and Assessment of Results

Decipher the found examples and survey their importance with regards to the Spain markets.

Translation and Assessment of Results:

- Evaluate the discovered examples to understand their relevance within the Spanish markets.
- Assess the significance of identified patterns and trends concerning consumer behavior and market dynamics.

Strategic Analysis of Mercadona and Market Responses:

- Examine the adaptive strategies employed by Mercadona, Merca Madrid, and Merca Barca in response to pandemic-induced changes.
- Investigate alterations in their operational methods or offerings to address shifting market demands.

Visualization and Data Representation:

- Utilize visualization tools to present key findings effectively.
- Develop visual representations that depict consumer preferences, market dynamics, and performance insights of Mercadona, Merca Madrid, and Merca Barca during the pandemic.

Application of Knowledge in Business Strategies:

Translate acquired insights into actionable business strategies.
Suggest approaches such as strengthening supply chains, optimizing inventory management, and refining marketing tactics based on pandemic-related learnings.

1. Unify Product Naming:

- Renamed all products in plural form for a better view
- Add product type labels to the first dataset for clarity.

2. Select Relevant Datasets:

- Chose datasets for merging, including Food Consumption Data, Monthly Trade Data, and COVID-19.
- Excluded Data from MercaMadrid and MercaBerna datasets as they were used for labeling and contain similar data as Food Consumption Data.
- Excluded Weekly Pricing of the Products dataset due to redundancy in price information within the Consumption dataset.

3. Merged Datasets Based on Time Columns:

- Use the 'Month' and 'Year' columns as the key columns for merging.

- **Data Mining and Pattern Identification**
- Applying the data mining techniques to determine patterns and its trends in case of market demand, consumer's behavior and supply chain breakdowns during the pandemic.

- Understanding and exploring its patterns of how these Spanish markets reacted to it and their unique challenges.

4. Evaluation of the Final Results

Relationship between Consumption and Price (Statement 1):

Analysis:

- Two line plots created, one for fruits and one for vegetables. For Volume sold it is green and for the average price it is blue.

Results:

- Calculated the correlation coefficient between product categories and prices.

Consumption during the pandemic (Statement 2):

Analysis:

- We analyzed the months from March to November for the 2018, 2019, 2020
- We assumed that covid started in March 2020
- Then we calculated the medium of the prices from 2018 and 2019, grouping by product
- We used KMean algorithm with 5 clusters

Results:

- Noticed an increase in fruit and vegetable consumption during the initial months of the pandemic.
- We created the list of product that are impacted by covid

Slight Positive Impact: ['AGUACATE','AJOS','ALBARICOQUES','ALCACHOFAS','APIO','BERENJENAS','BROCOLI','CEREZAS','CHAMPIÑONES','CHIRIMOYAS','CIRUELAS','COLES','COLIFLOR','ESPARRAGOS','FRESONES','JUDIAS VERDES','MANGOS','NECTARINAS','PATATAS CONGELADAS','PATATAS FRITAS','PATATAS PROCESADAS','PEPINOS','PIÑAS','POMELO','PUERROS','UVAS','VERDURAS DE HOJA']

Slight Negative Impact: ['CALABACINES','FRUTAS IV GAMA','KIWI','LECHUGAS','LIMONES','MANDARINAS','MELOCOTONES','OTRAS FRUTAS FRESCAS','PERAS','PIMIENTOS','VERD./HORT. IV GAMA','ZANAHORIAS']

Significant Negative Impact: ['NARANJAS','PATATAS FRESCAS','TOMATES','TOTAL PATATAS']

Significant Positive Impact: ['T.FRUTAS FRESCAS', 'T.HORTALIZAS FRESCAS']

Relationship between import quantity and price during covid (Statement 3):

Analysis:

- Firstly categorized the products by fruits and vegetables for our statement
- Then made the quantity-price graphs for fruit and vegetables
- Also made the graph for covid cases
- In both of the graphs for fruits and vegetables we saw that the peak values were from December 2019.
- We used k-means clustering to see the relationship between the price and quantity

Results:

- We can see the imports increase in winter months and the increases were higher from the previous years.

- In the first months of 2020, first both of them decreased and then increased even so they have a correlation but that was when covid cases weren't too much.

- With covid cases rising up from 16 to 100k, the imports have decreased but their prices increased so we can say that our statement is valid for this month.

- For vegetables, in the first months the relation between quantity and values are more stable. They both increase or decrease together.

Annex

Here are our links:

Github link: https://github.com/orgs/UCLM-ESI-NECULA/projects/2

Google Colab link where we cleaned our data:
https://colab.research.google.com/drive/1TDcs3742BFN3uEN3wtIkcONiGihD9Lo1

Google Colab link where we worked on the data card:
https://colab.research.google.com/drive/14M2BPIyEtpZrWvssb8ZedWmJ_MbLwICh#scrollTo=iDy1GxEH4zMO

Google Colab link for our pandas profiling:
https://colab.research.google.com/drive/12I24iQ38PANwFkDyinXbRSj_h6I820KS#scrollTo=0umvwTUoxJPu

Drive link with the datasets and files we’ve worked on:
https://drive.google.com/drive/folders/1tDgUndXM9VgKSCjr6a5LsVTfbtemm3FZ?usp=sharing

Thank you!

1. Valentin Necula
2. Aymina Yılık
3. Cenk Karakaş
4. Öykü Sedef Öztürk
5. Anatoli Zournatzi
6. Dilan Kubay