Introduction

The pandemic we have experienced due to COVID-19 has changed our habits. Not only affected our health but also left an impact on the global economy.
One of the most impacted areas during the pandemic was the food industry.
This situation caused a first phase of urgency to buy food and household products leading to a compulsive purchase of products such as fruits, vegetables, and hygiene products.

Goals

Our main goal is to examine how COVID-19 has affected the Spanish fruit and vegetable market. To do this we've used a variety of agricultural datasets to support what we discover. We've tried to achieve the following goals:

1. **Consumer Behavior Analysis:** Examine the changes in consumer behavior during the pandemic, especially in fruit and vegetable consumption.
2. **Market Response:** Explore how the Spanish fruit and vegetable market managed the COVID-19 crisis.

Statements

- **Relationship between Consumption and Price**
    - The consumption of fruits and vegetables is directly related to the price of those products. With an increase in the price, there is a consumption decrease. This indicates an inverse relationship between price and consumption.

- **Consumption during the pandemic**
    - In the first months of the pandemic, people consumed more fruits and vegetables compared to the same period of the previous years
    - There is a relationship between the consumption of these products and the rise of the daily cases of COVID-19 in Spain. During high spikes of cases of COVID-19, there are high spikes of consumption
    - There are patterns of consumption of specific categories of fruits and vegetables during the pandemic like the citrus fruits were bought more because people thought it could help with the "influence".

- **Relationship between Import and Price during the pandemic**
    - The import of fruits and vegetables in Spain during COVID-19 decreased compared to the previous years. These fewer imports lead to an increase in the price
    - The products that had a high dependency on imports from specific countries were more affected by the rise of the price compared with the products with less dependency.

Considerations over statements

The first statements can have different problems. The first could be that not all the goods respond to the price changes in the same way. The elasticity of demand can affect our statement a lot. The consumption of some goods, for example, those considered necessities, like fruit and vegetables might not be affected a lot by the price.
The second statement relationship could not be linear or directly connected. Some panic buying could be possible, but at the same time consumption during that period of uncertainty could be affected by other factors such as job losses, or reduced incomes. Also, the supply chain could be affected impacting the availability of the products
The hypothesis of the price increase and a lower import of the products could be mitigated by internal production, which could reduce the impact of those problems.

Dataset description

There are five datasets. Our datasets are .txt files with a CSV format.

Datos Consumo Alimentario → Food Consumption Data		
This dataset represents fruit and vegetable consumption in Spain from the Ministry of Agriculture, Fisheries and Food, with data on volumes, prices, and penetration		
Column	**Description**	**Example of data**
Año	The year of the data	2018
Mes	The month of the data	Enero
CCAA	Spain Autonomous Community or Region	Aragon
Producto	Specific product	HORTALIZAS
Volumen	Volume of the product, in thousands of kilograms	209957,24
Valor	Value of the product consumed, in thousands of euros	376688,56
Precio medio	Average price per kilogram (kg)	1,79
Penetración	Penetration of the product in the market (%)	97,27
Consumo per capita	Consumption per person	4,6
Gasto per capita	Money spent by a person in the period taken into consideration	8,25

Dataset 1Wikipedia's definition of Market penetration: “*Market penetration refers to the **successful selling** of a good or service in a specific market. It is measured by the **amount of sales** volume of an existing good or service **compared to the total target market** for that product or service.”*

Dataset 2

PreciosSemanales → Weekly Pricing of the products		
It presents the average weekly prices both for farmers and markets of different fruits and vegetables. This dataset investigates the pricing dynamics of these products over the years.		
Column	**Description**	**Example of data**
INICIO	Start date of the week	01/01/2018
FIN	End date of the week.	01/01/2018
GRUPO	Category of the product	Agrícola
SECTOR	Sector of the product	Frutales
SUBSECTOR	Another classification of the product	Citricos
PRODUCTO	Specific product	LIMON
TIPO/SUBTIPO	Another classification of the product	SIN ESPECIFICAR
POSICION	Position of the sale	Agricultor, Mercas, Subasta
FORMATO	Format of the product	II
PRECIO	Price of the product	0,92
UNIDAD	Unit for the price	Euros/kg

Dataset 3a and 3b

DatosMercaMadrid →Data from MercaMadrid		
Provides information on the prices and volumes of the products sold weekly in the market.		
Column	**Description**	**Example of data**
product	Specific product	*ACEITUNAS*
variedad	Type of the product	ACELGAS
origen	Origin of the product	TOLEDO
Unidad	Unit of the volume	kg
familia	Broad category or family the product belongs to	FRUTAS
YEAR	Year of the data	2020
MONTH	Month of the data	11
price_mean price_min price_max	Mean, minimum, and maximum price of the product	3,46
Volumen	Volume of the product sold	1500

Monthly Product Price and Volume Data by Origin		
It has similar columns to dataset 3a but with fewer columns.		
Column	**Description**	**Example of data**
product	Specific product	*ACEITUNAS*
origen	Origin of the product	TOLEDO, ITALIA
Unidad	Unit of the volume	kg
familia	Broad category or family the product belongs to	FRUTAS
YEAR	Year of the data	2020
MONTH	Month of the data	11
price_mean	Mean, minimum, and maximum price of the product	3,46
Volumen	Volume of the product sold	1500

Dataset 4

Comercio Exterior de España → Monthly Trade Data		
It contains the data from the import and export of fruits and vegetables with the rest of the European countries from 2018.		
Column	**Description**	**Example of data**
PERIOD	Month and year of the data	Jan. 2018
REPORTER	Country reporting the data	Austria
PARTNER	Trading partner country	ES
PRODUCT	Specific product or category of product	Brussels sprouts
FLOW	Type of trade	IMPORT, EXPORT
INDICATORS	Type of data	VALUE_IN_EUROS
Value	Value for the given indicator	29, 0, :

Dataset 5

Daily COVID-19 Cases and Deaths Data		
It includes daily international COVID-19 statistics by country, from Jan 2020 to Nov 2020.		
Column	**Description**	**Example of data**
dateRep	Date of the report	16/11/2020
day/month/year	Day, month, and year of the report	16, 11, 2020
cases	Number of new cases reported	228
deaths	Number of new deaths reported	11
countriesAndTerritories	Specific country or territory	Afghanistan
geoId/countryterritoryCode	Geographical codes for the country	AF,AFG
popData2019	Population data for 2019	38041757
continentExp	Continent of the country or territory	Asia
Cumulative_number_for_14_days_of_cases_per_100000	Cumulative number of cases in the past 14 days per 100,000 population	6,39560365

Possibilities for Data Enrichment

In addition to the data types already mentioned, there are several potential sources of data enrichment useful for our analysis:

We Found a Covid-19 Dataset with all the Countries inside. It is a very big dataset but it is very helpful for us because it has the daily report. Having such a big granularity helps us to manipulate the data for our purposes.
It has a lot of null values also, so the preprocessing part of this dataset is going to be very important. The link to the source is here, it was made by Our World in Data.
An important part is that this data that is completely open access under the Creative Commons BY license. This allows us to use reproduce, and distribute it in any medium, provided the source and authors are credited.
It has a lot of columns, the main ones are the following Deaths, Cases, Tests, Hospitalizations, Vaccinations, Mortality risk, Excess mortality, Policy responses
We are more interested in the Cases data.

Background or similar work

Here are some winners of the CAJAMAR Agri Data competition and their approaches:
Datacrop:
They searched for the effects of COVID-19 on agri-food.

- Their main analysis was “How has the consumption and production of fruits and vegetables been affected during the pandemic compared to previous years?”.
- To achieve their goal they used a methodology such as processing the data, reading different articles, loading/enriching/cleaning their datasets, analyzing their datasets and displaying them on the website they created, and also using an auto machine learning service to improve their project.

TheDataMasters:
They created an interactive application to see the impacts of COVID-19 on the Spanish fruit and vegetable market.Their main objectives:

- What is the state of the Spanish fruit and vegetable market?
- How much has the virus affected this market?
- Has the diet been healthier during the pandemic period?

Their methodology was doing the project in 3 phases.

- First phase consisted of preprocessing the data with cleaning and classifications, making the data ready before the other phases.
- Second phase consisted of processing the data with aggregation of the data by averages and sums, creation of new variables, and unification of all textual fields
- Third phase was the implementation of interactive visual application. They created an application that has a menu for the interfaces, filters to select the variables to inspect, and visual objects to see the area and conclusions of that area.
- The interface they created has 3 prospects: agri analysis, the impact of Covid-19, and health impact.

To sum it up, we thought their approaches were helpful to us to understand what we needed to do and plan our methodology.

Data Selection

Understanding data and assessing the data quality

The first thing we did was to do a first analysis of the datasets. To do this we used GoogleColab and the Pandas Profile Report. The Code we used is here.
This report saved us some time for the analysis and helped us to understand the data more.

DatosConsumoAlimentario

- The time distribution of the data is mainly in 2019 with 41% of the data, then we have 2018 with 37% of the data and 2020 with 21%.

Common Values

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

- The first value starts from 2018 in January to 2020 in November. This means December 2020 is missing. We don't think this could be a problem.

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%

- We have a higher frequency in some months, with a 3.3% difference. This means we have more data in some months than others. We'll investigate that in the next phase.

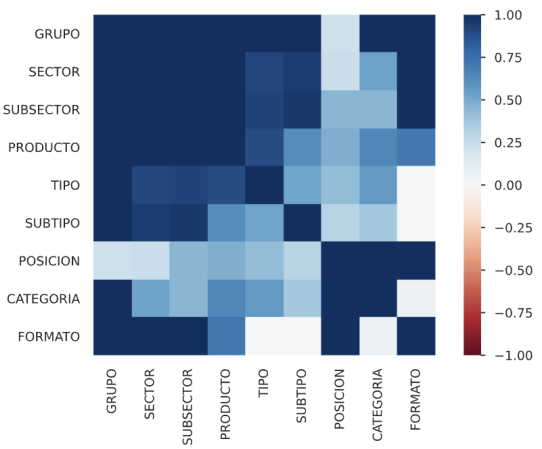
- We saw that we have separate data for all CCAA of Spain, and 1 more category which is Total. We think it's the sum/average of all the CCAA data, but we need to

query the data to ensure that. We don't have statements related only to the CCAA so we might use only the National Data, removing the other rows.

- We saw that we don't have missing data, but we will see if we have some zeros in the price, which could be useless to analyze.

PreciosSemanales

- The first thing we noticed is that the column "Unidad" has a constant value which is Euros/kg. We will surely delete this column because it is useless.
- The Time distribution here goes from 1/01/2018 to 10/12/2020.
- The Group, Sector, and Subsector are used to categorize the products.
- The column Type has 18% of the data missing, and 16% of them are NO DESIGNADO.
- Type columns are not useful with all this missing data.
- Subtype has the same problem, with 20% data missing, and more than 50% SIN ESPECIFICAR.
- The POSITION has only 3 Values Agricultor, Mercas, Subasta.
- The CATEGORY and FORMAT columns are not understandable and there are a lot of missing values. We Might delete them.
- We have a lot of correlation inside those columns, probably we can simplify the values between the columns, maybe with some clustering algorithms.



Comercio Exterior de España

- The “partner” column is constant and is ES, so we’ll delete it.
- The “flow” is only IMPORT and EXPORT, so we can change these values into 0 and 1. The “indication” is the same, VALUE_IN_EUROS and QUANTITY_IN_100KG.
- In the “reporter” column we have 28 European countries.

DatosMercaMadrid and DatosMercaBerna

- The column “Unidad” is only kg. We’ll delete it. The products are 125 distinct products with a high percentage of tomatoes, potatoes, lettuce, and apples.
- The Familia column is distributed like this.

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%

- The problem will be to merge their products because they are classified with different labels. We will decide later how to do it.

Coronavirus Cases

Overview | Alerts 6 | Reproduction

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

- The first thing we notice is the high correlation between cases and deaths.
- Another thing is that the zeros here are not a problem, are a valid value.

- The problem is that is a small dataset, with cases only for 2020. Not enough for our study cases.

Selection of Data

Relationship between Consumption and Price

- We need to gather the consumption and price data through the years. We can do this by splitting the data by years and collecting the columns that are valuable for us (product, ccaa and month). So then we can compare the consumptions and prices.

Consumption during the pandemic

- We need to check the consumptions in the first months of 2020.

- We can check the product, ccaa and month columns and compare the columns of first 3 months.

Relationship between Import and Price during the pandemic

- We'll check "comercio exterior de España" dataset for this. We'll use import, period and product columns.
- We also need their prices.

Preprocess and Transformation

Statement 3

remove outliers 68-95-...

For the first statement, we can use the dataset 1.
For the 2.1 ST, we need to label the product of dataset 1 “Fruits and vegetables” using dataset 1,2,3

For the 2.2 we need to use the weekly covid, and MAYBE…. search for another dataset with daily consumption of the products

For the 2.3 we need to have the labeling part mentioned earlier.

For the 3 ST, we can use the import-export dataset, adding also the COVID report for the countries, in order to understand how the price, covid cases, and volume change.

1) Try to unify the list of products along the first three datasets, changing the names

2) Try to label all the products with at least “fruit” and “vegetable” categories, it would be useful to have for example a higher granularity distinction like “berries” or “citrus”

3) Unify the dataset 1 (with the labeling), the Covid, and The Import/Export with the Month granularity

4) Once we have the big table we can start doing the documentation part, in which we have these points
    a) Enrich the datasets (covid, and maybe the daily consumption of products)
    b) Enhance the hypotesis (maybe change the 2.2 statement if we don't have the weekly prices)
    c) Pick the variables (explain what columns and datasets we used/added/deleted)
    d) Preprocesamento y trasformacion (describe the things we did on the datasets)
    e) Work proposal (High level description of what algorithms we want to use)
    f) Anexe (Google colab and Github repo)