Introduction

For the second phase of the project, we had to work on the Preprocessing and Transformation phase of a KDD work. For the preprocessing phase, we based our decision on the analysis we did in the first step.

Statements enhance

The first statement remains the same, it's a quite simple statement, that doesn't require a lot of work, and the data we need is contained in the first dataset from DataAgri. We can improve it by trying to Categorize the products more affected by the price change.
To do this we can use a clustering algorithm.

- **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. We can determine some groups of products that are more affected than others.

In the second statement, we had to remove the relationship between consumption and the rise of daily cases because we didn't find daily consumption datasets. The rest of the statement remains the same.

- **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".

The third statement remains the same.

- **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.

Data Enrichment

For the data enrichment part, we agreed to use the Dataset mentioned in the previous work. We did the cleaning part on the dataset regarding the covid, but we confirmed that we couldn't use it because it didn't have enough data.
We used the Ourworldindata dataset of COVID-19. The link to the source is here.
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. An important part is that this data 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, and Policy responses
We are more interested in the Case data

Preprocess and Transformation

Here we tried to prepare the data for analysis. The code aims to prepare and clean the data, ensuring that unnecessary information is removed, missing values are handled, and the content is more focused and ready for analysis or further processing.

Dataset 1 - DatosConsumoAlimentario

- We modify the data DataFrame by converting month names in Spanish to numerical equivalents in the 'Mes' column and sort the data by year, month, and region without assigning it to a variable. After June the dataset shafts to presenting only the total national data, resulting in varying frequencies for specific months due to the absence of regional-level information, so the last 120 rows reflect aggregated national data.

- To compare the months better, we changed them as numbers. For example "Enero:1, Febrero: 2" etc.

- We compared Dataset 1 and Dataset 2 and we added labels to Dataset 1. To create the data card; we had to change some names of the products to match them. Down below you can see we changed the product name 'MELON to 'MELONES'.

```
df_1['Producto'] = df_1['Producto'].replace('MELON', 'MELONES')
```

- For all the CCAAs in Spain, we have separated the data and the category which is called 'national total'. We need to check if the total"s summarization or the average is the same as the other CCAAs. We made a code that compares the data between the total Nacionales and the other CCAAs.

- After everything was corrected and checked on Total Nacional data since June of 2020 we dropped the other CCAAs and saved them.

Dataset 2 - PreciosSemanales

- From the loaded data, we choose certain columns ('INICIO', 'FIN', 'SECTOR', 'SUBSECTOR', 'PRODUCTO', 'PRECIO') for further processing.

- We handle the missing values so we removed rows from the DataFrame where the 'PRODUCTO' column has no data ('NaN' - Not a Number). This ensures that only rows with valid product information remain.

- Our data's date was weekly and we had to make them monthly to merge them with the other data. (You can see the uncleared data down below)

```
INICIO|FIN|GRUPO|SECTOR|SUBSECTOR|PRODUCTO|TIPO|SUBTIPO|POSICION|CATEGORIA|FORMATO|PRECIO|UNIDAD
01/01/2018|07/01/2018|Agrícola|Frutales|Citricos|LIMON|||Mercas|||0,92|Euros/kg
01/01/2018|07/01/2018|Agrícola|Frutales|Citricos|MANDARINA|||Mercas|||0,84|Euros/kg
01/01/2018|07/01/2018|Agrícola|Frutales|Citricos|NARANJA|||Mercas|||0,78|Euros/kg
01/01/2018|07/01/2018|Agrícola|Frutales|Citricos|POMELO|||Mercas|||1,01|Euros/kg
```

- Also, we had to clean the product information because the 'PRODUCTO' column had a lot of unnecessary data. We removed the extra details enclosed in parentheses leaving only the main product name.

Dataset 3A & 3B - DatosMercaMadrid & DatosMercaBarna

- This dataset, was a little bit different than the others because we had to merge them. Firstly, we removed the column named ***Unidad***. Because it had a constant value of Euro/kg.

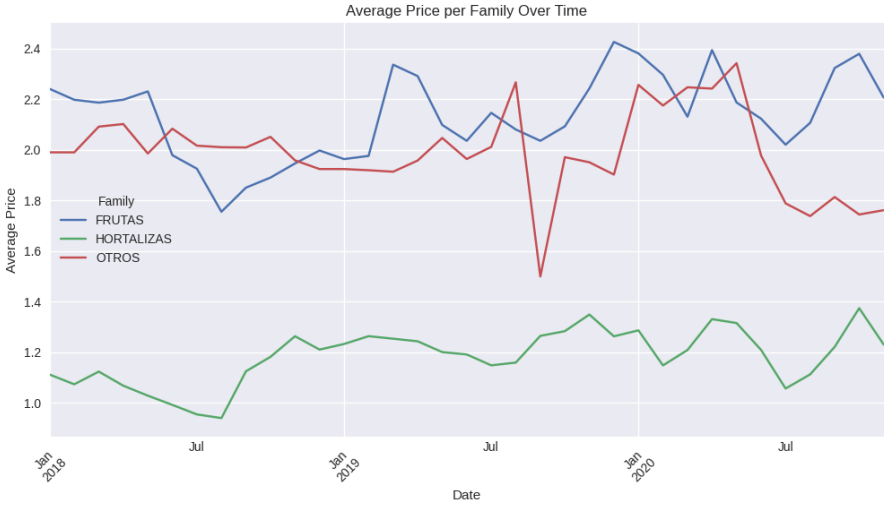
- Because we didn't need a specific family category we decided to reduce the values of the column ***Familia***. Instead, we tried to achieve 2 big categories which are ***Fruits*** and ***Vegetables***.

- We kept the ***Other*** and ***Seta*** categories. Even if our statements are not related to them.

- We replaced the commas with the dots that are between the numbers to get the numeric conversion properly.

```
product|variedad|origen|Unidad|familia|YEAR|MONTH|price_mean|price_min|price_max|Volumen
ACEITUNAS|ACEITUNAS|ALMERIA|kg|FRUTAS|2018|2|3,46|3,31|3,61|6700
ACEITUNAS|ACEITUNAS|ALMERIA|kg|FRUTAS|2018|4|3,46|3,31|3,61|400
ACEITUNAS|ACEITUNAS|ALMERIA|kg|FRUTAS|2018|5|3,46|3,31|3,61|260
ACEITUNAS|ACEITUNAS|BADAJOZ|kg|FRUTAS|2018|4|3,46|3,31|3,61|24060
ACEITUNAS|ACEITUNAS|BARCELONA|kg|ULTRAMARINO|2019|2|0|0|0|1000
```

- We merged Dataset1 and Dataset2. We grouped common columns and calculated their average(mean).
- Lastly, we define the file paths for the CSV file's outputs. And saved the DataFrames to CSV files.
- We combined the ***YEAR*** and ***MONTH*** into a single date for columns.
- We plotted the mean price for each family over time.



Dataset 4 - ComercioExterior

- Some of the rows containing missing or duplicate values across any column were dropped. we ensured that the dataset contained only complete information.

- We removed the column 'PARTNER' from the DataFrame as it's not necessary for further analysis.

- Some rows have an invalid 'PERIOD', they have December and January merged together in the same row. For our purpose, we need to separate them. We don't know how to separate the value of the price and the quantity of those products so we decided to split them equally.

```
Dec. 2018|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|EXPORT|VALUE_IN_EUROS|:
Dec. 2018|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|EXPORT|QUANTITY_IN_100KG|:
Jan.-Dec. 2018|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|IMPORT|VALUE_IN_EUROS|1058
Jan.-Dec. 2018|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|IMPORT|QUANTITY_IN_100KG|4
Jan.-Dec. 2018|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|EXPORT|VALUE_IN_EUROS|:
Jan.-Dec. 2018|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|EXPORT|QUANTITY_IN_100KG|:
Jan. 2019|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|IMPORT|VALUE_IN_EUROS|:
Jan. 2019|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|IMPORT|QUANTITY_IN_100KG|:
Jan. 2019|Austria|ES|Bananas, fresh (excl. plantains)(2012-2500)|EXPORT|VALUE_IN_EUROS|56680
```

- We created new columns ('MONTH' and 'YEAR') to separate the store of month and year extracted from the 'PERIOD' column.

- Removed the null values, which in this dataset were signed with ":".
- The DataFrame is sorted primarily by 'YEAR',' MONTH', and 'PRODUCT' columns, arranging the data in ascending order.

Dataset 5 - CoronavirusCases

- In this dataset we processed both datasets and we decided to use dataset 6 ***owid-covid-data.csv*** which has bigger data than dataset 5.

- We created a new data frame for Spain, removing the data from all the other countries. We summed every case of the same month. By doing this, we cleaned the data and we had values per month, like in all the other datasets.

- We replaced the ***NaN*** values with ***0*** in this line:

```
df_spain.fillna(0, inplace=True)
```

Dataset Merge

In order to achieve our goal we had to merge the datasets. We decided to merge Food Consumption Data, Monthly Trade Data, and COVID-19. Data from MercaMadrid and MercaBerna were used to label the products, but we decided to not merge them because the data was the same in the Food consumption dataset. We haven't used the Weekly Pricing of the products dataset because it had data on the price in the Consumption Dataset, and also because our statements are monthly based.
Once we finished the preprocessing and transformation phase we decided to merge the datasets in one big table, with the columns we will use later.
To merge the datasets we tried to unify the naming of the products, renaming them all in plural, and adding the types labeling to the first dataset.
We merge the datasets based on the Month and Year, adding the data mapped on these columns. This allowed us to have a well-distributed dataset with a clear time frame of the values.

The final Dataset has the following columns:

Column	Description	Example of data
Year	The year of the data	2018
Month	The month of the data transformed to integers	1 (as January)
Producto	Specific product	CHIRIMOYAS
Grupo	Broad category or group the product belongs to	FRUTAS

Volumen (miles de kg)	Volume of the product, in thousands of kilograms	79445.66
Valor (miles de €)	Value of the product consumed, in thousands of euros	84640.08
Precio medio kg	Average price per kilogram (kg)	0.78
Consumo per capita	Monthly consumption per person	2.38
Gasto per capita	Money spent by a person during the month specified	1.85
Reporter	Country reporting the data	Belgium
Flow	Type of trade	0(import),1(export)
Indicators	Type of data	QUANTITY_IN_100KG
Value	Value for the given indicator	11188
total_cases	Number of new cases reported	345

Work Proposal

For the next part of the work, we will choose the right algorithms for our statements.
For the first Statement, we think that a clustering algorithm would suit us the most for our purpose. We still need to decide which one. The first part of the second statement is quite easy to do, with some correlation detection we can prove it. The second part will be the hardest. We think that we can define the patterns of the consumption of specific products with some Clustering algorithms. The third statement is, like the previous one, a matter of demonstrating the correlation between four factors. Covid-19, Import export, Price, and Consumption of fruits and vegetables.

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/12l24iQ38PANwFkDyinXbRSj_h6l820KS#scrollTo=0umvwTUoxJPu

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