­CCT College Dublin

|  |  |
| --- | --- |
| Modules Titles: | Data Preparation & Visualisation, Machine Learning for Data Analytics, Programming for Data Analytics, Statistics for Data Analytics |
| Assessment Title: | Ireland Agri-Business |
| Lecturers: | David McQuaid, Dr. Muhammad Iqbal, Sam Weiss, Taufique Ahmed |
| Student: | Ricardo Alves de Souza |
| Student Number: | 2018499 |
| Assessment Due Date: | 18/04/2024 |
| Date of Submission: | 26/05/2024 |

Declaration

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Title: Ireland Agri-Business

Author: Ricardo Alves de Souza

E-mail: [2018499@student.cct.ie](mailto:2018499@student.cct.ie?subject=Integrated%20CA1)

GitHub Address: https://github.com/ricardoasouz/2024---MSc-in-Data-Analytics---Feb---FT/tree/36c758bf7493b93b5199709afc8300c4e80051d3/%20MSC\_DA\_CA2

Table of Contents

[Scenario 4](#_Toc168090455)

[METHODOLOGY 5](#_Toc168090456)

[DATA SOURCES 5](#_Toc168090457)

[- Scraping by “Beautiful Soup” 5](#_Toc168090458)

[- Scraping by “API” 5](#_Toc168090459)

[SENTIMENT ANALYSIS 6](#_Toc168090460)

[Beef Production on Europe, Slaughterings and Price’s 7](#_Toc168090461)

[Obtaining the dataset’s 7](#_Toc168090462)

[Preparing the dataset for analyses. 9](#_Toc168090463)

[INTERACTIVE DASH BOARD 10](#_Toc168090464)

[European Beef Analyses (slaughterings and prices) 10](#_Toc168090465)

[Forecast Predictions 13](#_Toc168090466)

[References 14](#_Toc168090467)

­Table of Figures

[Figure 1: Sentiment Analysis, Polarity and Subjectivity Distributions. 8](#_Toc166751733)

[Figure 2: Heatmap of OBS\_VALUE Over TIME\_PERIOD for Each Country 11](#_Toc166751734)

­­

# Scenario

*“Today, big data is ubiquitous, machine learning applications are thriving, artificial intelligence appears in everyday conversations, and the internet of things is present even in household appliances. Businesses and organizations are increasingly managed through cloud computing and high-performance computing is progressively accessible as a service…More effective operations, reduced uncertainties, and real-time decision support could revolutionize agriculture to a great extent. Food could be produced more efficiently, of higher nutritional quality, in more stable supplies, with less environmental damage, and likely with additional economic, social, and ecological benefits.”*(Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis (2022))

# METHODOLOGY

The methodology is “CRISP-DM”, a robust and well-established framework that outlines a comprehensive process of understanding the business context, data understanding, data preparation, modelling, evaluation, and deployment. It's widely adopted due to its industry-agnostic and flexible nature, facilitating the study and development.

# DATA SOURCES

This study has used the computer programming language “Python” in the format of “Jupyter Notebook” with dedicated libraries for different tasks. It has acquired data to analyse from different sources, including tree files provided together with this assessment:

* Agriculture\_\_crops\_\_livestock\_and\_land\_use\_by\_general\_farm\_type\_\_region\_18042024\_071227.csv
* EuropeanAgriculture\_FarmStructureIndicators\_Eurostat2016.csv
* Irish-agri-food-exports-208-2022\_21032023.csv

https://ec.europa.eu/eurostat/databrowser/view/tag00044/default/table?lang=en&category=t\_agr.t\_apro.t\_apro\_mt

tag00044\_linear.csv

The segment chosen for analysis was:

“Ireland Agri-Business on CAP”

## Scraping by “Beautiful Soup”

A discussion content was obtained from, “Forum4Farming - Britain & Ireland Farming Forum, Agricultural Discussion Boards. IE & UK”. On the following address:

(“https://www.forum4farming.com/forum/index.php?threads/cap-2023-2027.20587/”)

And saved in the format “JSON”, an acronym for, “JavaScript Object Notation” on the file:

(“extracted\_forum4farming.json”)

This format is chosen to make it easy to read by humans and upload to a NoSQL Database without any more conversion.

The library utilised to obtain this data part is popular among Data Analytics by your use facility and good performance and is called “BeatfulSoup”.

The code for this extraction can be viewed on file, “StractingForumsBSandAPI.ipynb” and the libraries “request”, “BeautifulSoup”, “re” and “JSON” is utilised to connect to the web address and manage the requisition to realise the download from six pages of the contents mentioned before.

## Scraping by “API”

Another discussion content was obtained from Reddit. On the following address:

(“https://www.reddit.com/r/ireland/comments/1aghrfn/irish\_farmers\_protest\_in\_solidarity\_with\_eu/”)

And also saved in the format “JSON”, on the file:

(“extracted\_redditfarming.json”)

The code for this extraction is also on file, “StractingForumsBSandAPI.ipynb” and the environment variables are hidden on the, “.env” file to not exposure the credentials utilised to access the Reddit API also the, “.env” file is cited inside the file, “.gitignore” making this environment credentials invisible on Open Source platforms like, “GitHub”.

“PRAW”, is the library to access the Reddit API, she can hold the credentials to open access for obtain content submitted from address specified on variable “URL”.

After that, both files have been merged into only one, type, “JSON” and also in format, “CSV” to make it easy to analyse from Machine Learning Models.

(“extracted\_combined.json”)

(“extracted\_combined.csv”).

# SENTIMENT ANALYSIS

With the advance of techniques in machine learning and improvement on code’s dedicatees to analyse what feelings bring each specific word typed by users on the internet, now are possible to detect and or find out what the users want manifest, if this is positive or negative on the question about.

For this case let’s make use of a library called “TextBlob”, with her is possible to process the file previously scrapped and prepared in a formatted dataset, and obtain two variables: polarity and subjectivity.

Polarity measures the positivity or negativity of the text with a range between -1 to 1.

Subjectivity measures the objectivity of the text with a range between 0 to 1.

author statement

0 Franz\_Werfel the famers protests on the continent were abou...

1 lamahorses Over 30% of the European budget is on CAP etc

2 ConnolysMoustache What are they even protesting? The IFA repr...

3 bintags These people are brainwashed.

4 goodforyourself21 Stop bitching at each other and get out and su...

textblob\_sentiment polarity subjectivity

0 (0.0, 0.5) 0.00000 0.500000

1 (0.0, 0.0) 0.00000 0.000000

2 (0.04375, 0.685417) 0.04375 0.685417

3 (0.0, 0.0) 0.00000 0.000000

4 (-0.125, 0.375) -0.12500 0.375000

Now with all phrases analysed by the “TextBlob” library and converted into measured numeric values, is possible to perform descriptive statistics and also plot a graphic to visualize the distribution of these results.

Polarity Subjectivity

count 129.000000 129.000000

mean 0.071990 0.409139

std 0.156366 0.207902

min -0.375000 0.000000

25% 0.000000 0.310606

50% 0.052000 0.419643

75% 0.141667 0.512352

Max 0.800000 1.000000

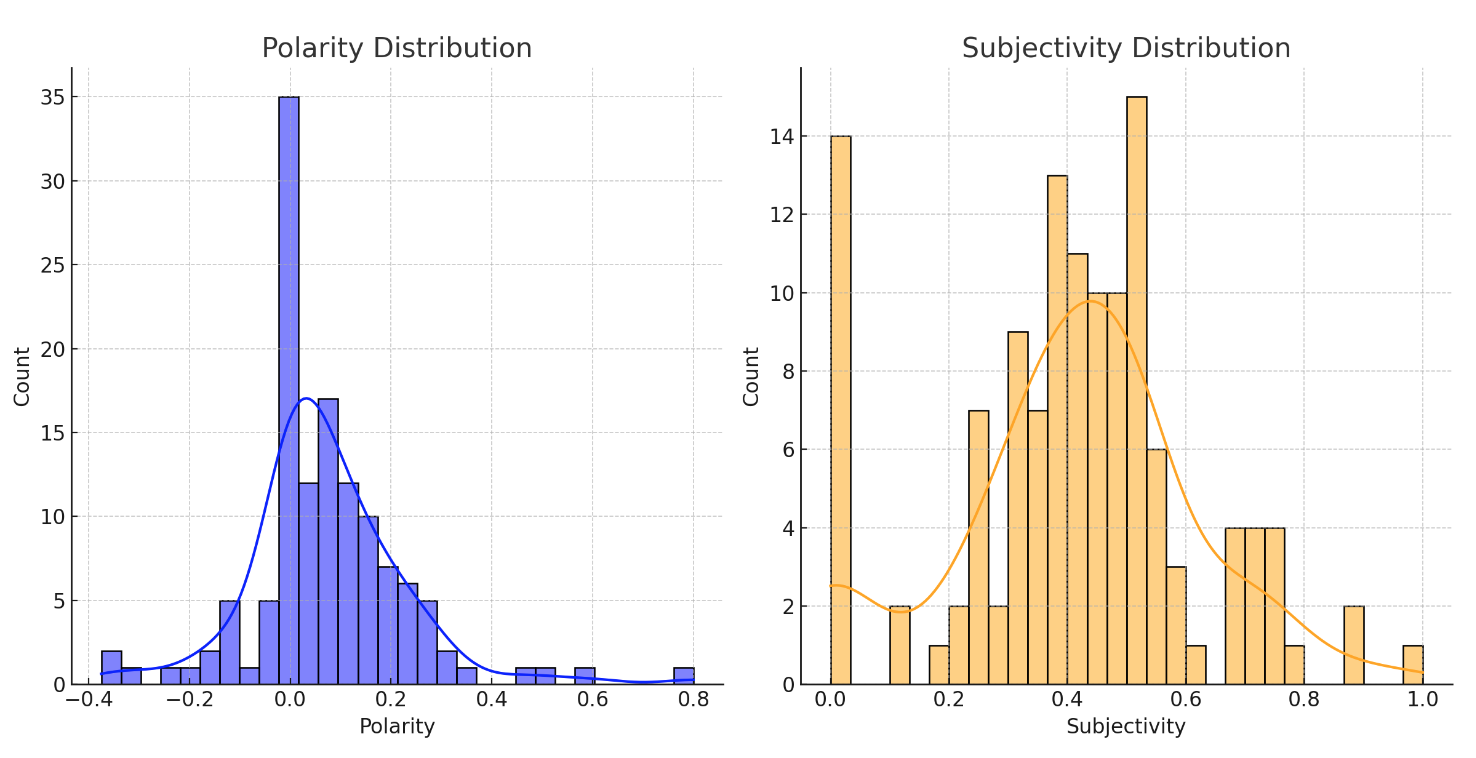


Figure 1: Sentiment Analysis, Polarity and Subjectivity Distributions.

Now with statistics measured and graphically exposed it is easy to see the major part of people manifesting positivity to the subject analysed in this case the influence of CAP on the farmers in Ireland.

# Beef Production on Europe, Slaughterings and Price’s

## Obtaining the dataset’s

For this case study, let us analyse meat production: Slaughtering between the countries in the European Union, from 2012 to 2023. For this, a dataset from the website “Eurostat” can provide us with accurate statistics numbers about countries and the subject cited.

The datasets were obtained from the following address in the format “SDMX-CSV 1.0”,

“https://ec.europa.eu/eurostat/databrowser/view/apro\_mt\_pheadm\_\_custom\_11602618/default/table?lang=en”.

After making a selection of countries to reduce the number of “null” values and get the maximum available data as possible, as downloaded and saved with the file name:

“apro\_mt\_pheadm\_page\_linear(Slaughterings).csv”. I added “(Slautherings)” to help easily recognise. This one is about meat production in European countries, it also downloaded three more datasets corresponding to prices observed over the same period (2017~2022). This period was chosen because it is more consistent, avoiding missing data.

And about the prices, the dataset as obtained from “https://ec.europa.eu/eurostat/databrowser/view/prc\_fsc\_idx\_\_custom\_11611346/default/table?lang=en”, where I downloaded three more datasets of prices about Commodity price, Consumer Price and Import Price.

*Production of meat: Slaughtering*

*The Dataset page description:*

This indicator covers the slaughtering of bovine animals (calves, bullocks, bulls, heifers and cows) slaughtered in slaughterhouses and on the farm, whose meat is declared fit for human consumption.

Online data code:

Online data code:

prc\_fsc\_idx

Source of data:

Eurostat

Last data update:

31/05/2024 10:00

Last structure update:

31/05/2024 10:00

Overall data coverage:

2005-01 - 2024-04

Number of values:

974 943

prc\_fsc\_idx\_\_custom\_11611346

FULL DATASET (prc\_fsc\_idx)

Time-frequency: Month

Unit of measure: Thousand tonnes

Item of meat: Slaughtering

Meat product: Bovine meat

*Prices of meat: Consumer, Commodity, Import*

*The Dataset page description:*

 Food price monitoring tool

Online data code:

prc\_fsc\_idx

Source of data:

Eurostat

Last data update:

31/05/2024 10:00

Last structure update:

31/05/2024

Overall data coverage:

2005-01 - 2024-04

Number of values:

974 943

Data navigation tree location

The source datasets for prc\_fsc\_idx are:

prc\_hicp\_manr

<https://ec.europa.eu/eurostat/databrowser/product/view/prc_hicp_manr>

prc\_hicp\_midx

<https://ec.europa.eu/eurostat/databrowser/product/view/prc_hicp_midx>

sts\_inppd\_m

<https://ec.europa.eu/eurostat/databrowser/product/view/sts_inppd_m>

prc\_fsc\_idx\_\_custom\_11611346

FULL DATASET (prc\_fsc\_idx)

## Preparing the dataset for analyses.

Opening the file about production, “Slaughtering”, in a “Pandas Data frame”, it is possible to observe the presence of columns with unique values representing redundant information that can be deleted. The structure of data can be rearranged and organised to eliminate the presence of categorical data, and “object” to make then possible and easy to perform statistics analyses and run machine learning models.

Following the original structure of the data frame:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1872 entries, 0 to 1871

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 DATAFLOW 1872 non-null object

1 LAST UPDATE 1872 non-null object

2 freq 1872 non-null object

3 meat 1872 non-null object

4 meatitem 1872 non-null object

5 unit 1872 non-null object

6 geo 1872 non-null object

7 TIME\_PERIOD 1872 non-null object

8 OBS\_VALUE 1872 non-null float64

9 OBS\_FLAG 181 non-null object

dtypes: float64(1), object(9)

memory usage: 146.4+ KB

To clean and organise this data frame, let’s delete the columns ( DATAFLOW, LAST UPDATE, freq, unit, meatitem, meat, OBS\_FLAG ), keeping only the columns, “geo, TIME\_PERIOD, OBS\_VALUE”.

To add the values from Commodity, Import, and Consumer, let’s merge the OBS\_VALUE of each one on this data frame, renaming the columns for your respective metrics.

The column “geo” represents the country where the value is observed and comes with only two letters, let’s convert these codes to a full country name with the help of the library, “pycountry”.

Unfortunately at this moment the problem of Greece with your abbreviation “EL”, it’s not recognised from the library “pycountry”, is easily solved by doing a custom\_country\_mapping, as is possibly observed on the Jupiter notebook code.

Now let’s pivot this table making the column, “TIME\_PERIOD” the index of the table and countries columns with your respective values observed with your respective metrics.

df.info()

<class 'pandas.core.frame.DataFrame'>

Index: 72 entries, 2017-01 to 2022-12

Columns: 104 entries, Austria\_commodity to Sweden\_prod

dtypes: float64(104)

memory usage: 61.1+ KB

The result is a data frame with 72 rows and 104 columns, all in numeric format helping us perform the statistical analysis and machine learning models in each country and in four parameters collected, (prod, commodity, consumer, import).

For this let’s implement an interactive dashboard with the capability of choosing the countries to be analysed and performing summary statistics and specific machine learning models.

# INTERACTIVE DASH BOARD

## European Beef Analyses (slaughterings and prices)

With the data frame now right obtained and organized about almost all countries in the European Union, let’s create a way to make a comparison between each one, of course not losing the target to see how Ireland is beside the others.

For this let’s create a way to choose multiple countries at the same time and a dashboard with selectors about what country and what metric is to be calculated and visualized in the way implemented.

To create the dashboard, let’s import and use the libraries “dash, dcc, html, Input, Output, State, dash\_table”

With “dash” library is easy to create a canvas and start inserting components on the dashboard.

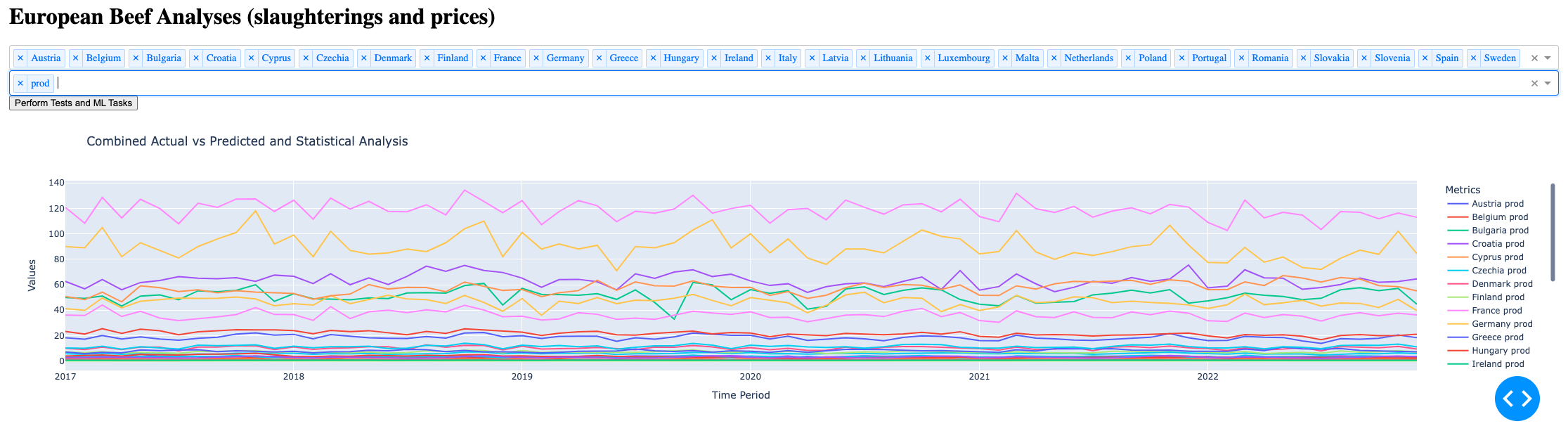
First of all, let’s insert a country selector a metric selector, and a line graph to visualize the values of selections over time.

With all countries selected at the same time can be hard to understand or visually polluted to see anything.



How is possible to see the picture above for the first time can be hard to understand what’s going on with the beef in the countries in Europe, but this capacity to filter the data in real time can clarify some information.

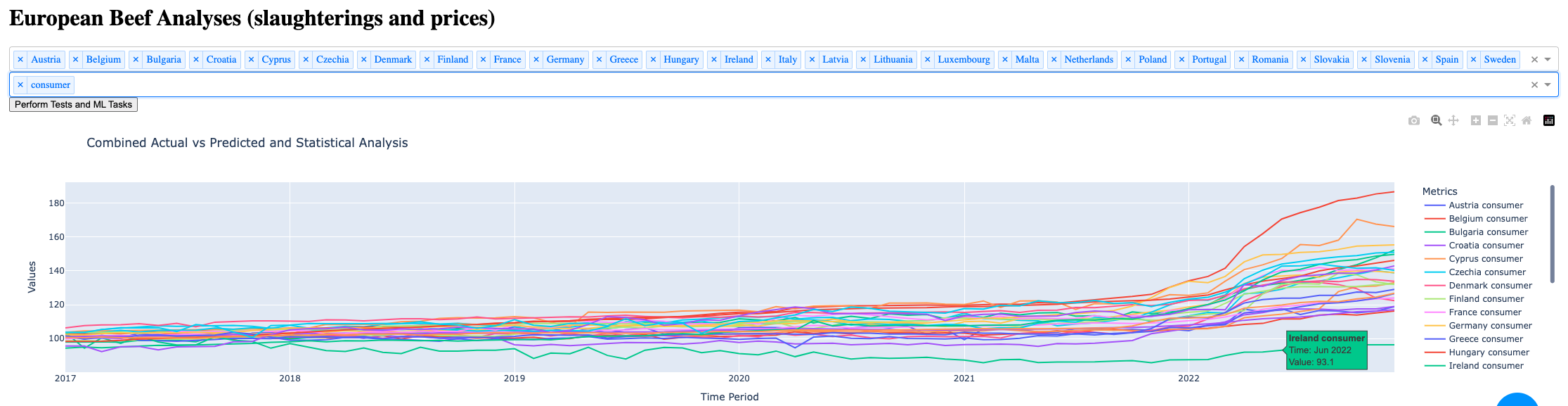
The focus for us is to try to understand how is Ireland on Agribusiness, for this first let’s filter the parameters and keep them just one by one to check if is possible to visualize something.



How is possible to see in the picture above, keeping only the metric, “prod” relative to the production of each country and keeping all countries selected, the graph starts a be less mazy and is possible to see Ireland is the fifth biggest producer of beef on Europe almost over all years analysed, literally swapping position with Italy, Spain, Poland and Netherlands.

Another information that can be obtained from this visualization is possible can classify the beef producers in Europe into three distinct groups by the quantity produced.

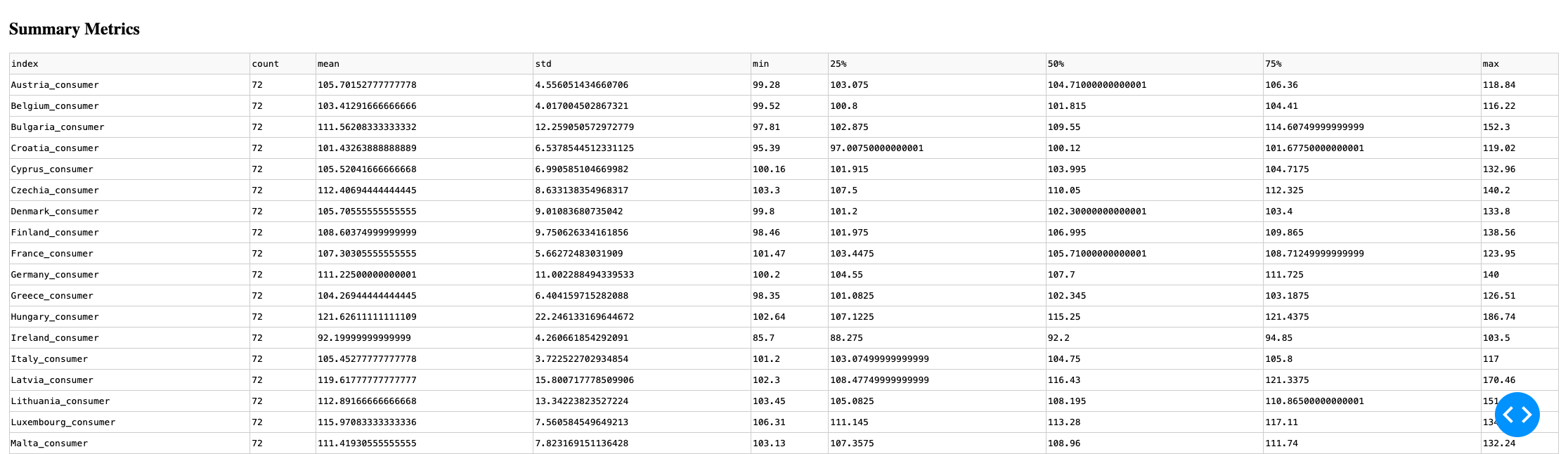
Now let’s try to keep just the metrics about consumer price to see what is possible to visualize.

******

In the middle of this maze picture, Ireland stands out from others keeping almost all years analysed with the small price of beef to consumers compared with other countries in the European Union, showing to us a strong characteristic about Ireland.

Another trend that can be visualised is all other countries after the year 2022 start increasing beef prices for consumers specifically Hungary leader on increase with a big jump followed by Latvia and Poland.

To try to understand better the behaviour of this data, let’s get some statistical information from the selected countries and for this let’s create a summary table with statistics values on the dashboard, keeping the characteristics of selected countries to visualize the numbers.

******

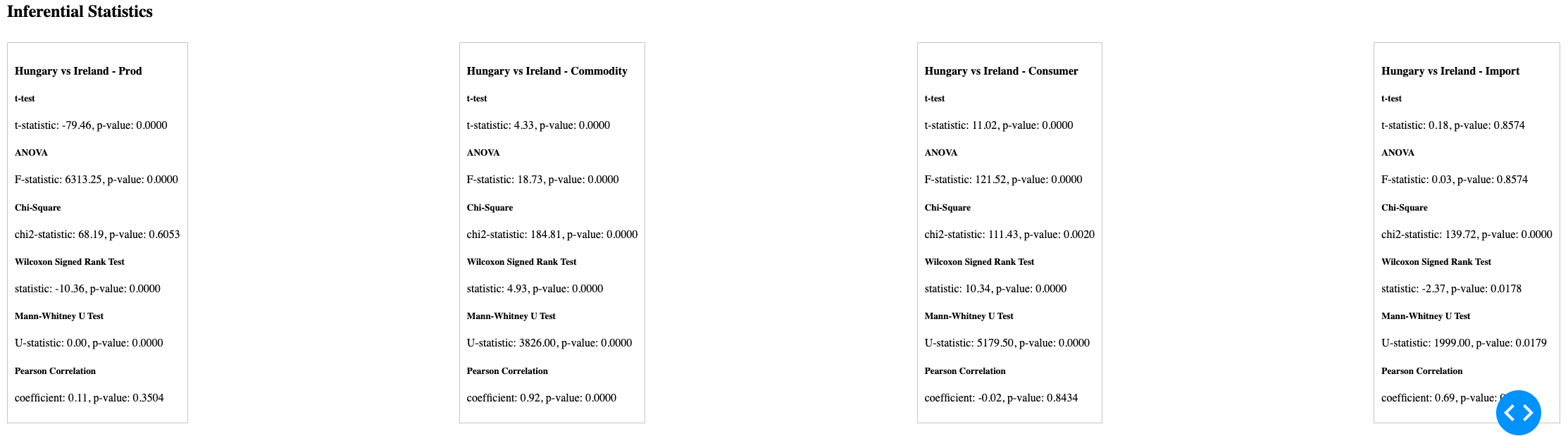
Sorting alphabetically Ireland comes just after Hungary, how can see that is the maximum opposite of Ireland in consumer price and with a bigger jump in the consumer price after the year 2021.

The summary metrics can show us the biggest differences between Ireland and Hungary in standard deviation.

To try to understand better the differences between the countries let’s try to perform some inferential statistics upside the data, for this let’s implement the calculations and the dashboard visualization to also visualize the results side by side.

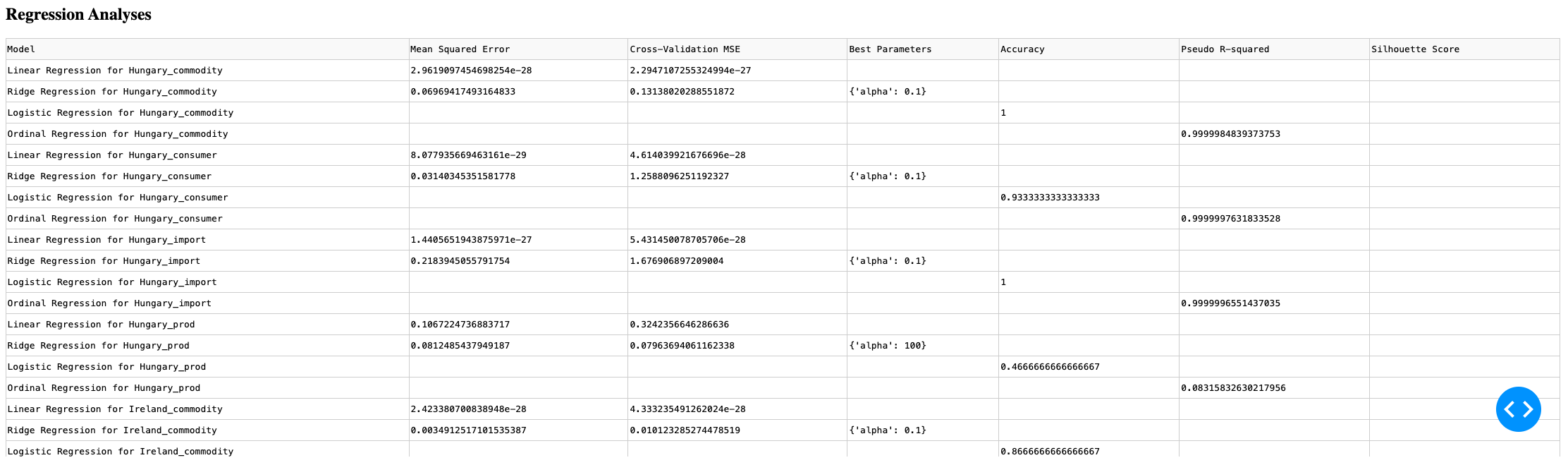
Let’s try to implement this on the dash board.

For this let’s make use of “Scipy.stats” library, with some features like (ttest\_ind, f\_oneway, chi2\_contingency, ranksums, mannwhitneyu, pearsonr), and try to organize the results on way can be better visualized.



Also, let’s perform Regression Analysis trying to get some insights about the data, performing “Linear, Ridge, Logistic and Ordinal Regressions”, and trying to compare the results

Of the similarity of needing steps to perform Machine Learning Algorithms, like splitting the data and training models, let’s implement and perform the Machine Learning algorithms together and for this try to apply supervised and unsupervised algorithms also making use of GridSearchCV to tune the parameters.



Now let’s try to interpret these values:

Linear Regression for “Hungary\_commodity”:

Mean Squared Error: 2.961997454698254e-28

Cross-Validation MSE: 2.2947107255324994e-27

This indicates that the linear regression model for predicting Hungary's commodity metric has a very low MSE, suggesting a good fit.

The Cross-Validation MSE is slightly higher but still low, indicating consistent performance across different subsets of the data.

Ridge Regression for “Hungary\_commodity”:

Mean Squared Error: 0.06969417493164833

Cross-Validation MSE: 0.13138020288551872

Best Parameters: {'alpha': 0.1}

This shows that the ridge regression model for Hungary's commodity metric has a higher MSE compared to the linear regression model but benefits from regularization, which can help in cases where overfitting might be an issue. The best alpha value found is 0.1.

Logistic Regression for Hungary\_commodity:

Accuracy: 1

The logistic regression model achieves perfect accuracy in classifying Hungary's commodity metric into binary classes, suggesting that the model fits the data extremely well for the given target.

Ordinal Regression for Hungary\_commodity:

Pseudo R-squared: 0.9999984839373753

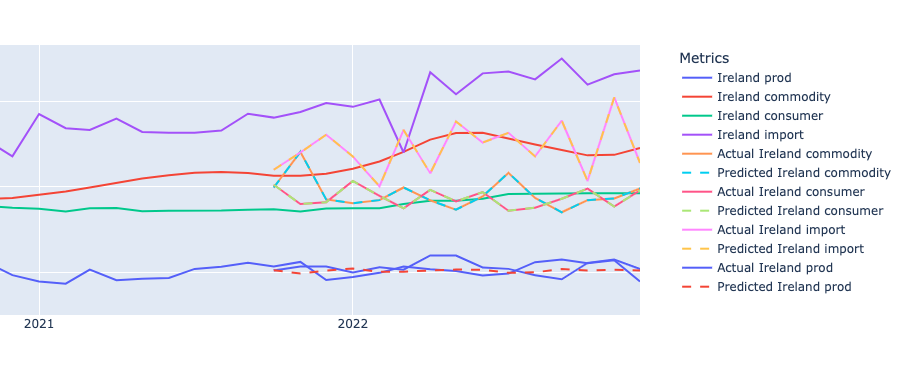
This high Pseudo R-squared value indicates that the ordinal regression model provides an excellent fit for the ordinal target variable “Hungary\_commodity”.

KMeans Clustering:

The silhouette score of approximately 0.44 suggests that the KMeans clustering algorithm has successfully grouped the data into clusters with a reasonable level of cohesion and separation.

## Forecast Predictions

How can it possible to see in picture, that the trained models got good accuracy on predictions with a minimal marge of error in some cases.



# References

Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis (2022): “Big data in agriculture: Between opportunity and solution”

<https://doi.org/10.1016/j.agsy.2021.103298>,(https://www.sciencedirect.com/science/article/pii/S0308521X21002511)