Ireland’s Agriculture Analysis

Abstract

This paper provides a detailed analysis of agricultural inputs and outputs products in Ireland, focusing on understanding the specific patterns and dynamics within the sector. By examining the types of inputs required for agricultural production, such as seeds, fertilizers, machinery, and labor, alongside the outputs produced, including crops and dairy products, this study aims to elucidate the underlying mechanisms driving agricultural productivity in Ireland.

Furthermore, comparative analyses will be conducted to assess how the behaviour of inputs and outputs in Ireland differs from that of other European Union countries. By identifying potential variations in agricultural practices, this research seeks to uncover unique challenges and opportunities specific to Irish agriculture.

Ultimately, the primary objective of this study is to generate insights and establish behavioural patterns regarding agricultural inputs and outputs in Ireland, with a particular emphasis on discerning any significant differences in behaviour compared to other European Union countries. Through this analysis, policymakers, researchers, and stakeholders can gain a deeper understanding of the factors shaping agricultural sustainability and competitiveness in Ireland and the broader European context.

Introduction

The agricultural sector is a cornerstone of Ireland's economy, involving a complex interplay of practices, resources, and outputs. This paper explores the dynamics between agricultural inputs and outputs to understand productivity and sustainability. Recent transformations in Ireland's agriculture, driven by technology, consumer preferences, and environmental changes, highlight the importance of inputs like seeds, fertilizers, machinery, and labor. The analysis of outputs, including crops and dairy, provides insights into the sector's economic and environmental impact.

Additionally, the study compares Ireland's agricultural practices with those of other EU countries to identify patterns, disparities, and opportunities for collaboration. This comparative analysis aims to enhance understanding of Ireland's agricultural sector and foster cross-border innovation.

Overall, this research examines the intricacies, challenges, and opportunities within Ireland's agricultural sector, aiming to inform policy, promote sustainability, and strengthen Ireland's global agricultural position.

DATABASES

The study utilized databases containing Price Indices of Agricultural Production Means in European Union countries. The first study concentrates on Input Price Indices for Agricultural Production, whereas the second study pertains to Agricultural Product Price Indices. The data were acquired from the Eurostat website, namely from the agriculture section. In addition, a third database was obtained straight from the IBAN website. This database contains a reference table that includes national information such as the country name, Alpha-2 code, and Alpha-3 code. This reference table will be used to merge the major datasets. Further information regarding the databases is provided below.

Price Indices for Means of Agricultural Input Production

The **Price Indices of the Means of Agricultural Input Production** database contains statistics on the mean prices of agricultural inputs utilized by European Union nations from 2000 to 2017. It includes information on the products, countries, years, and quarters in which they were recorded.

	freq	p_adj	unit	product	geo_TIME_PERIOD	2000-Q1	2000-Q2	2000-Q3	2000-Q4	2001-Q1	...	2015-Q3	2015-Q4	2016-Q1	2016-Q2	2016-Q3	2016-Q4	2017-Q1	2017-Q2	2017-Q3	2017-Q4
0	Q	NI	I10	200000	AT	:	:	:	:	:	...	113.1	111.7	111.1	112.7	112.3	111.8	113.8	114.6	113.2	113.3
1	Q	NI	I10	200000	BE	:	:	:	:	83.3	...	107.7	107	104.7	105.9	107.1	107.7	111.1	109.7	108.8	109.8
2	Q	NI	I10	200000	BG	:	:	:	:	:	...	108.1	105.6	102.9	101.8	100.8	102.2	104.2	103.8	102	104.1
3	Q	NI	I10	200000	CY	:	:	:	:	:	...	118.4	107	108	109.9	113.3	108	113	115	110	107
4	Q	NI	I10	200000	CZ	83.7	85.1	87.7	91.1	89.4	...	111	109.4	107.9	107.3	105.9	106.2	108.4	108.6	107.8	108.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5419	Q	RI	PCH_SM	220000	RO	:	:	:	:	:	...	-5.1	-3.4	-2.3	-2.6	-4.2	-1.3	1.8	0.4	3.2	1.9
5420	Q	RI	PCH_SM	220000	SE	:	:	:	:	:	...	-1	-0.8	-2	-3.1	-2.6	-1.7	0	0.7	0.9	1.2
5421	Q	RI	PCH_SM	220000	SI	:	:	:	:	5.8	...	-2.3	-1.2	-1.1	-2.1	-1.5	-2.2	-0.8	-1.1	-1.2	0.7
5422	Q	RI	PCH_SM	220000	SK	:	:	:	:	8.5	...	-5.8	-3.4	-3.4	-3.9	-4.2	-5.1	-2	-2.1	-0.4	1.4
5423	Q	RI	PCH_SM	220000	UK	:	:	:	:	2	...	-5.4	-3.1	-5.5	-4	-2.9	0.6	3.7	2.4	2.1	0.8

Data Dictionary

- **Product:** Input Used in Agricultural Activities
- **geo_TIME_PERIOD:** Country code of location
- **[2000-Q2 … 2017-Q4]:** Year and Quarter Columns with Average Indices of Agricultural Inputs Products.

Price Indices of Agricultural Output Products

The **Price Indices of Agricultural Output Products** database contains statistics on the average prices of agricultural products produced by European Union countries from 2000 to 2017. The database has the same structure as the previous one, with information on products, countries, years, and quarters in which they were recorded.

	product	subclass	geo_TIME_PERIOD	2000-Q1	2000-Q2	2000-Q3	2000-Q4	2001-Q1	2001-Q2	2001-Q3	...	2015-Q3	2015-Q4	2016-Q1	2016-Q2	2016-Q3	2016-Q4	2017-Q1	2017-Q2	2017-Q3	2017-Q4
0	010000	CEREALS	AT	:	:	:	:	:	:	:	...	75.8	82.1	82.6	83.5	62.6	69.5	79.2	81.7	77.6	79.8
1	010000	CEREALS	BE	:	:	:	:	67.2	68.4	67.9	...	93.1	97.7	86.7	79	69.5	92.5	92.1	90.1	88.3	89.1
2	010000	CEREALS	BG	:	:	:	:	:	:	:	...	123.1	108.8	115.9	117.5	108.9	105.9	122.2	119.6	112.6	102.3
3	010000	CEREALS	CY	:	:	:	:	:	:	:	...	190	:	:	170.3	216.8	:	:	167.3	178.4	:
4	010000	CEREALS	CZ	91.1	96	102.7	113.5	120.4	126.5	117.1	...	129.4	131.2	128.2	118.3	113.5	113.2	116	121.4	120.7	121.9
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6427	141000	TOTAL AGRICULTURAL GOODS	RO	:	:	:	:	38.6	29.9	5.6	...	0.6	5	4.8	2.3	-4.9	-3	-1.8	0.4	-1.5	-0.4
6428	141000	TOTAL AGRICULTURAL GOODS	SE	:	:	:	:	:	:	:	...	-1	0.9	0.3	-0.4	0.5	2.8	5.5	4.3	5.9	5.9
6429	141000	TOTAL AGRICULTURAL GOODS	SI	:	:	:	:	-5.5	0.2	-2.8	...	-5	-0.8	-2	-5.6	-5.7	-1	-0.5	6.5	9.1	5.4
6430	141000	TOTAL AGRICULTURAL GOODS	SK	:	:	:	:	2.9	2.6	-1.1	...	-0.8	4.8	0.5	-2.7	-5.6	-10.9	-6.1	1.1	5.8	10.2
6431	141000	TOTAL AGRICULTURAL GOODS	UK	:	:	:	:	6.2	9.8	5.6	...	-8	-6.2	-7.7	-5.7	-0.3	5.9	9.8	12.8	9.4	4.4

Data Dictionary

- **Product:** Output Produced in Agricultural Activities
- **geo_TIME_PERIOD:** Country code of location
- **[2000-Q2 … 2017-Q4]:** Year and Quarter Columns with Price Indices of Agricultural Output Products.

Table of Country Codes

The **Country Codes - ALPHA-2 & ALPHA-3** data table is a complete list of all ISO country codes as described in the international standard ISO 3166. These codes are used throughout the IT industry by computer systems and software to facilitate the identification of country names. The alpha-2 and alpha-3 codes are part of the ISO 3166-1 standard, which defines codes for representing the names of countries and their subdivisions. These codes are widely used in international contexts, such as in internet country identifiers, official documents, and data classification systems.

Country	Alpha-2 code	Alpha-3 code	Numeric
Afghanistan	AF	AFG	004
Albania	AL	ALB	008
Algeria	DZ	DZA	012
American Samoa	AS	ASM	016
Andorra	AD	AND	020

Data Dictionary

- **Country:** Output Produced in Agricultural Activities
- **Alpha-2 Code:** It consists of two letters and is used to represent countries in an abbreviated form.
- **Alpha-3 Code:** It consists of three letters and provides a more detailed representation of country names.

DATA PREPARATION

Importing Dataset
Price Indices for Means of Production of Agricultural Inputs Products

1. The dataset was imported in a standard way with the read_csv function from the Pandas library.

Importing Dataset
Price Indices of Agricultural Output Products

The Price Indices of Agricultural Output Products dataset was imported by connecting to the MySQL database. The steps for loading the database into MySQL Workbench software until importing the data into Jupyter Notebook are detailed below:

1. The MySQL Workbench software was opened;
2. The connection to the MySQL database was made;
3. A schema called 'agriculture' was created;
4. By pressing the right mouse button on the "agriculture" schema, the Table Data Import Wizard option was selected to import the dataset;
5. The directory where the dataset was saved was selected;
6. The 'Create a new table' option was selected and a name was defined for the table;
7. On the next screen, the format of each of the dataset variables and the encoding used were selected;
8. The process of importing the dataset into the 'agriculture' schema in the database was carried out.
9. The database was exported into a *.sql* file and then loaded into the Jupyter notebook for manipulation.
10. Environment variables were configured with the connection data to the database.
11. After this, the connection to the MySQL database was made.
12. Query was run on the database with the aim of selecting variables necessary for the study.
13. Finally, the data returned through the query was saved in a dataframe and the MySQL connection with the database was closed.

Data Preparation
Price Indices for Means of Production of Agricultural Inputs Products

Below are the steps taken to prepare the **Price Indices for Means of Production of Agricultural INPUTS** database:

1. Removal of variables *freq*, *p_adj*, and *unit* that will not be used in the study.

	product	geo_TIME_PERIOD	2000-Q1	2000-Q2	2000-Q3	2000-Q4	2001-Q1	2001-Q2	2001-Q3	2001-Q4	...	2015-Q3	2015-Q4	2016-Q1	2016-Q2	2016-Q3	2016-Q4	2017-Q1	2017-Q2	2017-Q3	2017-Q4
0	200000	AT	:	:	:	:	:	:	:	:	...	113.1	111.7	111.1	112.7	112.3	111.8	113.8	114.6	113.2	113.3
1	200000	BE	:	:	:	:	83.3	83.2	83.3	82.8	...	107.7	107	104.7	105.9	107.1	107.7	111.1	109.7	108.8	109.8
2	200000	BG	:	:	:	:	:	:	:	:	...	108.1	105.6	102.9	101.8	100.8	102.2	104.2	103.8	102	104.1
3	200000	CY	:	:	:	:	:	:	:	:	...	118.4	107	108	109.9	113.3	108	113	115	110	107
4	200000	CZ	83.7	85.1	87.7	91.1	89.4	91.1	91	89.6	...	111	109.4	107.9	107.3	105.9	106.2	108.4	108.6	107.8	108.7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5419	220000	RO	:	:	:	:	:	:	:	:	...	-5.1	-3.4	-2.3	-2.6	-4.2	-1.3	1.8	0.4	3.2	1.9
5420	220000	SE	:	:	:	:	:	:	:	:	...	-1	-0.8	-2	-3.1	-2.6	-1.7	0	0.7	0.9	1.2
5421	220000	SI	:	:	:	:	5.8	4	2.5	-2.2	...	-2.3	-1.2	-1.1	-2.1	-1.5	-2.2	-0.8	-1.1	-1.2	0.7
5422	220000	SK	:	:	:	:	8.5	5.8	-1.9	-3.4	...	-5.8	-3.4	-3.4	-3.9	-4.2	-5.1	-2	-2.1	-0.4	1.4
5423	220000	UK	:	:	:	:	2	0.9	2.4	1.4	...	-5.4	-3.1	-5.5	-4	-2.9	0.6	3.7	2.4	2.1	0.8

2. Pivoting the dataset and placing year and quarter information in rows.

	product	geo_TIME_PERIOD	date	value
0	200000	AT	2000-Q1	:
1	200000	BE	2000-Q1	:
2	200000	BG	2000-Q1	:
3	200000	CY	2000-Q1	:
4	200000	CZ	2000-Q1	83.7
...	...	...	...	...
390523	220000	RO	2017-Q4	1.9
390524	220000	SE	2017-Q4	1.2
390525	220000	SI	2017-Q4	0.7
390526	220000	SK	2017-Q4	1.4
390527	220000	UK	2017-Q4	0.8

3. The variable *class* was included to distinguish which information comes from Agricultural Inputs or Outputs Products.

4. The **Price Indices for Means of Production of Agricultural Inputs Products** dataset was merged with the Country Codes - ALPHA-2 & ALPHA-3 dataset to provide information about the name of the country.

	key_0	product	geo_TIME_PERIOD	date	value	type	Item	class	subclass	Description
0	200000	200000	AT	2000-Q1	:	inputs	200000	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
1	200000	200000	BE	2000-Q1	:	inputs	200000	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
2	200000	200000	BG	2000-Q1	:	inputs	200000	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
3	200000	200000	CY	2000-Q1	:	inputs	200000	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
4	200000	200000	CZ	2000-Q1	83.7	inputs	200000	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
...	...	...	...	...	...	...	...	...	...	...
390523	220000	220000	RO	2017-Q4	1.9	inputs	220000	INPUT TOTAL	INPUT TOTAL	INPUT TOTAL (INPUT 1 + INPUT 2)
390524	220000	220000	SE	2017-Q4	1.2	inputs	220000	INPUT TOTAL	INPUT TOTAL	INPUT TOTAL (INPUT 1 + INPUT 2)
390525	220000	220000	SI	2017-Q4	0.7	inputs	220000	INPUT TOTAL	INPUT TOTAL	INPUT TOTAL (INPUT 1 + INPUT 2)
390526	220000	220000	SK	2017-Q4	1.4	inputs	220000	INPUT TOTAL	INPUT TOTAL	INPUT TOTAL (INPUT 1 + INPUT 2)
390527	220000	220000	UK	2017-Q4	0.8	inputs	220000	INPUT TOTAL	INPUT TOTAL	INPUT TOTAL (INPUT 1 + INPUT 2)

5. Removal of columns *key_0*, *class_y* and *Description*

Data Preparation - Price Indices of Agricultural Outputs Products

Below are the steps taken to prepare the **Price Indices of Agricultural Outputs Products** database:

1. Pivoting the dataset and placing year and quarter information in rows.

	product	subclass	geo_TIME_PERIOD	date	value
0	010000	CEREALS	AT	2000-Q1	:
1	010000	CEREALS	BE	2000-Q1	:
2	010000	CEREALS	BG	2000-Q1	:
3	010000	CEREALS	CY	2000-Q1	:
4	010000	CEREALS	CZ	2000-Q1	91.1
...	...	...	...	...	...
463099	141000	TOTAL AGRICULTURAL GOODS	RO	2017-Q4	-0.4
463100	141000	TOTAL AGRICULTURAL GOODS	SE	2017-Q4	5.9
463101	141000	TOTAL AGRICULTURAL GOODS	SI	2017-Q4	5.4
463102	141000	TOTAL AGRICULTURAL GOODS	SK	2017-Q4	10.2
463103	141000	TOTAL AGRICULTURAL GOODS	UK	2017-Q4	4.4

2. Including the variable *class* in the dataframe with the value *output*.

Merging the Datasets

After preparing the **Price Indices of Agricultural Outputs Products** data, the dataset was merged with **Price Indices for Means of Production of Agricultural Inputs Products**.

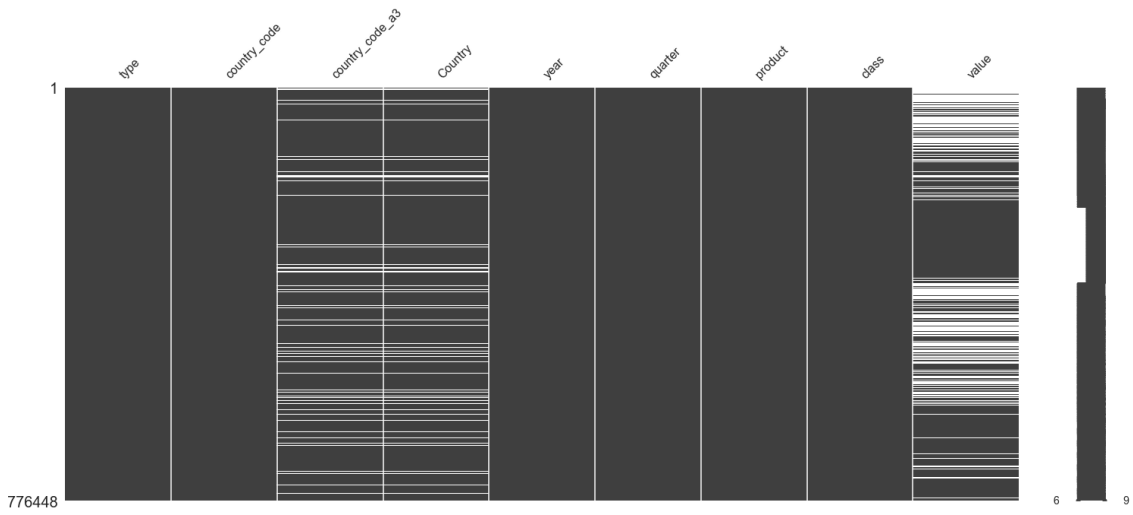
	product	geo_TIME_PERIOD	date	value	type	subclass
0	200000	AT	2000-Q1	:	inputs	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
1	200000	BE	2000-Q1	:	inputs	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
2	200000	BG	2000-Q1	:	inputs	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
3	200000	CY	2000-Q1	:	inputs	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
4	200000	CZ	2000-Q1	83.7	inputs	GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...
...	...	...	...	...	...	...
463099	141000	RO	2017-Q4	-0.4	outputs	TOTAL AGRICULTURAL GOODS
463100	141000	SE	2017-Q4	5.9	outputs	TOTAL AGRICULTURAL GOODS
463101	141000	SI	2017-Q4	5.4	outputs	TOTAL AGRICULTURAL GOODS
463102	141000	SK	2017-Q4	10.2	outputs	TOTAL AGRICULTURAL GOODS
463103	141000	UK	2017-Q4	4.4	outputs	TOTAL AGRICULTURAL GOODS

Removing Columns and Treating Missing Values

After merging the data, some columns needed to be removed due to duplication or simply because they were not part of the analysis. Are they:

- Date
- Key_0
- Alpha-2 code
- Numeric

With the removal of columns, we need to check the presence of null or missing values. To do this, the graph below was plotted:

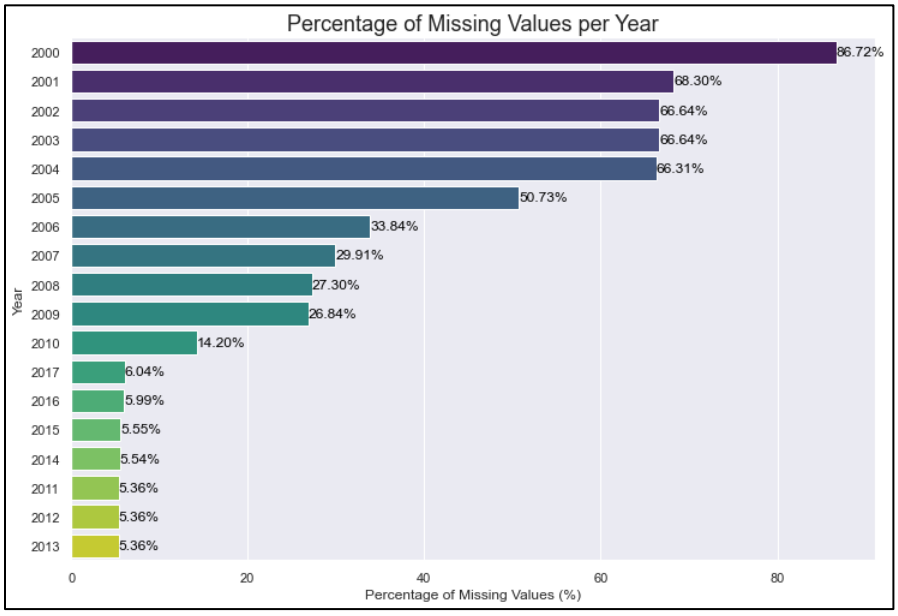


We can observe the presence of several missing values in the dataframe, especially in the Country and Value variables. Therefore, we need to study in detail the absence and impact of these values on the variables.

Analysing the Missing Values in 'values' according to other variables

1. **Missing Values of 'values' per 'year'**

	year	total_values	missing_values	Missing_Values (%)
0	2000	43136	37406	86.716432
1	2001	43136	29460	68.295623
2	2002	43136	28744	66.635757
3	2003	43136	28744	66.635757
4	2004	43136	28604	66.311202
5	2005	43136	21882	50.727930
6	2006	43136	14596	33.837166
7	2007	43136	12900	29.905415
8	2008	43136	11778	27.304340
9	2009	43136	11576	26.836053
10	2010	43136	6126	14.201595
11	2017	43136	2604	6.036721
12	2016	43136	2584	5.990356
13	2015	43136	2396	5.554525
14	2014	43136	2390	5.540616
15	2011	43136	2312	5.359792
16	2012	43136	2312	5.359792
17	2013	43136	2312	5.359792

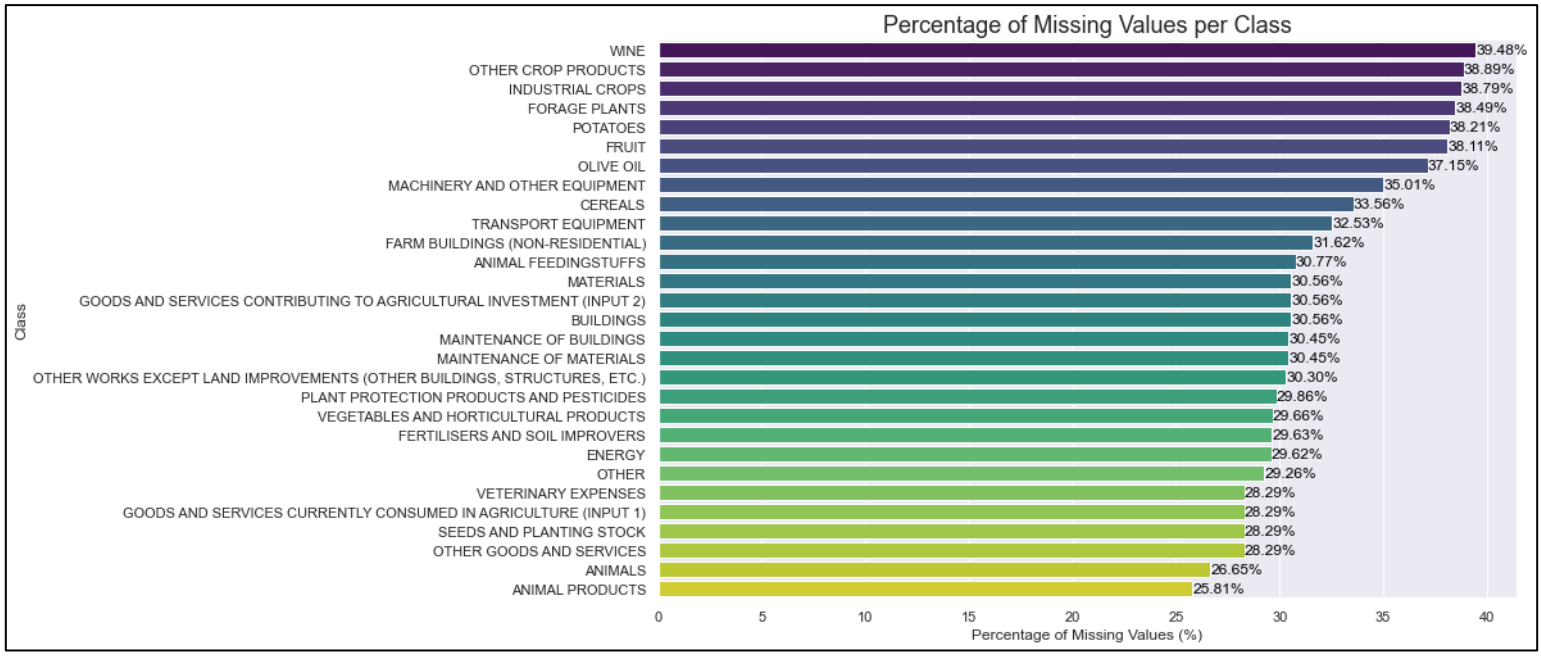


The data displayed illustrates the proportion of missing values in the 'value' variable from 2000 to 2017. The study uncovers three clearly defined periods:

- During the period from 2000 to 2004, there was a significant number of missing values, with the proportion starting at 86.72% and gradually reducing to approximately 66.31%.
- During the period from 2005 to 2010, there was a notable decrease in the proportion of missing data, declining from 50.73% in 2005 to 14.20% in 2010.
- From 2011 to 2017, there was a consistent level of stability with regards to missing values, which remained at a low rate of approximately 5-6%.

2. Missing Values of 'values' per 'class'

class	total_values	missing_values	Missing_Values (%)
WINE	11520	4548	39.479167
OTHER CROP PRODUCTS	10656	4144	38.888889
INDUSTRIAL CROPS	46944	18208	38.786639
FORAGE PLANTS	14400	5542	38.486111
POTATOES	37440	14304	38.205128
FRUIT	52704	20084	38.107165
OLIVE OIL	2304	856	37.152778
MACHINERY AND OTHER EQUIPMENT	46368	16232	35.006901
CEREALS	70560	23678	33.557256
TRANSPORT EQUIPMENT	21024	6840	32.534247
FARM BUILDINGS (NON-RESIDENTIAL)	7488	2368	31.623932
ANIMAL FEEDINGSTUFFS	80928	24904	30.773033
MATERIALS	7776	2376	30.555556
GOODS AND SERVICES CONTRIBUTING TO AGRICULTURA...	7776	2376	30.555556
BUILDINGS	7776	2376	30.555556
MAINTENANCE OF BUILDINGS	7776	2368	30.452675
MAINTENANCE OF MATERIALS	7488	2280	30.448718
OTHER WORKS EXCEPT LAND IMPROVEMENTS (OTHER BU...	3168	960	30.303030
PLANT PROTECTION PRODUCTS AND PESTICIDES	34560	10320	29.861111
VEGETABLES AND HORTICULTURAL PRODUCTS	44064	13068	29.656863
FERTILISERS AND SOIL IMPROVERS	64224	19032	29.633782
ENERGY	36288	10748	29.618607
OTHER	4320	1264	29.259259
VETERINARY EXPENSES	7776	2200	28.292181
GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	7776	2200	28.292181
SEEDS AND PLANTING STOCK	7776	2200	28.292181
OTHER GOODS AND SERVICES	7776	2200	28.292181
ANIMALS	76896	20496	26.654182
ANIMAL PRODUCTS	40896	10554	25.806925

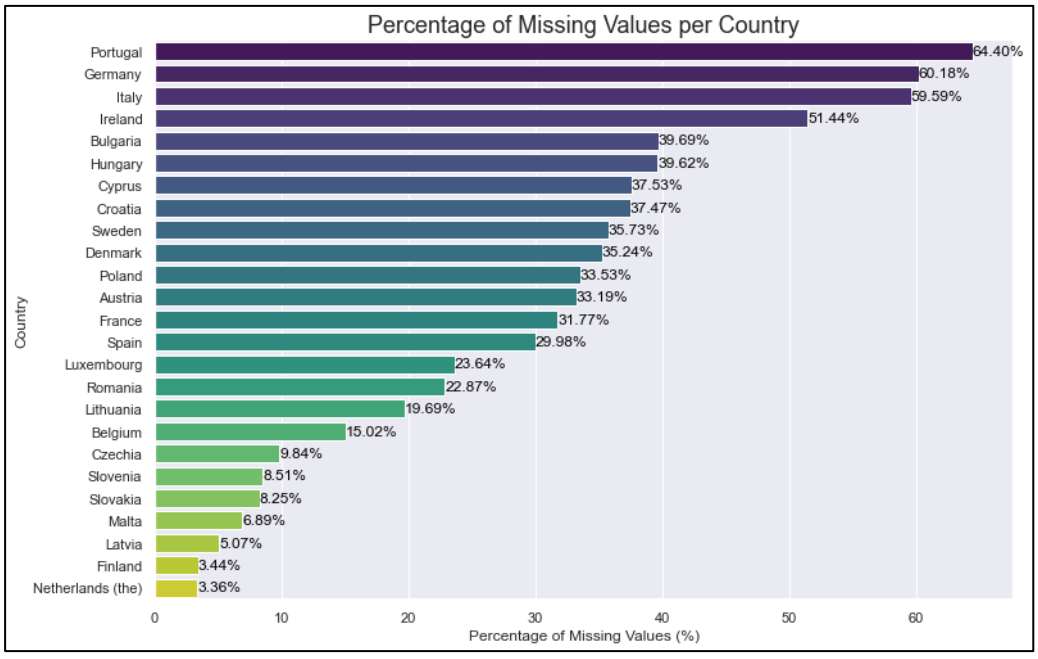


Through data analysis, it is evident that various product categories exhibit varying rates of missing values. For instance, several classes, like 'WINE', 'OTHER CROP PRODUCTS', and 'INDUSTRIAL CROPS', exhibit missing value rates exceeding 35%, whilst other classes like 'ANIMAL PRODUCTS', 'ANIMALS', and 'SEEDS AND PLANTING STOCK' have rates below 30%. This implies that certain categories of products are more prone to having incomplete data compared to others.

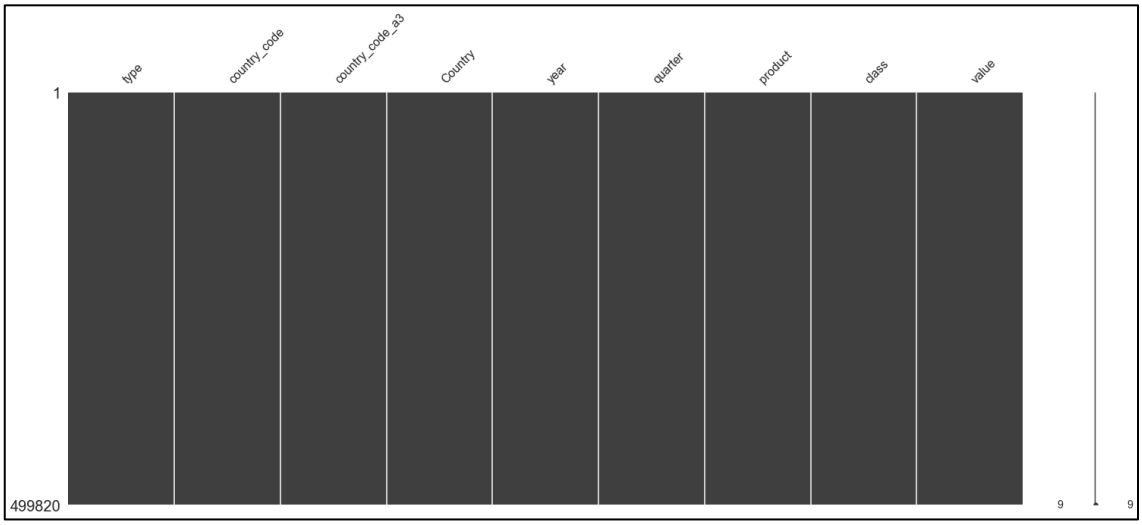
3. Missing Values of 'values' per 'Country'

	Country	country_code	total_values	missing_values	Missing_Values (%)
0	Portugal	PT	33408	21516	64.403736
1	Germany	DE	22752	13692	60.179325
2	Italy	IT	31104	18534	59.587191
3	Ireland	IE	23328	12000	51.440329
4	Bulgaria	BG	28800	11430	39.687500
5	Hungary	HU	32544	12894	39.620206
6	Cyprus	CY	29376	11026	37.534041
7	Croatia	HR	32544	12194	37.469272
8	Sweden	SE	27072	9672	35.726950
9	Denmark	DK	28800	10150	35.243056
10	Poland	PL	30240	10140	33.531746
11	Austria	AT	32256	10706	33.190724
12	France	FR	32832	10432	31.773879
13	Spain	ES	32832	9842	29.976852
14	Luxembourg	LU	25344	5992	23.642677
15	Romania	RO	31104	7112	22.865226
16	Lithuania	LT	26496	5216	19.685990
17	Belgium	BE	27648	4152	15.017361
18	Czechia	CZ	26496	2606	9.835447
19	Slovenia	SI	27072	2304	8.510638
20	Slovakia	SK	27936	2304	8.247423
21	Malta	MT	20448	1408	6.885759
22	Latvia	LV	24480	1240	5.065359
23	Finland	FI	24192	832	3.439153
24	Netherlands (the)	NL	29088	978	3.362211

The analysis indicates that there are varying levels of missing data in the records of different countries. For instance, Portugal, Germany, and Italy exhibit missing value rates exceeding 50%, whereas the Netherlands, Finland, and Latvia display rates below 10%.



Treating the Missing Values



Upon eliminating all null values, it becomes evident that the dataframe is now devoid of any missing values. However, previously we noticed that the year variable has a significant amount of missing data, with some years having about 100% of their information missing. By examining the variable country, it becomes apparent from the graphic provided that some countries lack data for certain years. Hence, in order to prevent any influence on subsequent studies, we will exclude years in which there is missing data for at least one country. Therefore, the years 2000 to 2010 will be excluded from the analysis.

country_code	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
AT	NaN	NaN	NaN	NaN	NaN	75280.1	82529.7	99954.6	108674.2	84446.0	85042.9	107316.7	98987.3	101711.0	86229.2	90396.1	87776.4	89408.8
BE	NaN	70421.1	63960.1	71536.7	66392.2	64857.2	77668.0	80768.2	91013.9	59367.6	77350.4	87434.1	90858.7	86455.3	66767.5	79159.5	77438.1	80498.9
BG	NaN	NaN	NaN	NaN	NaN	NaN	64103.9	79785.9	84822.4	76059.9	75764.8	94958.1	92010.9	92799.6	82275.6	86912.7	82969.6	82132.9
CY	NaN	NaN	NaN	NaN	NaN	81813.4	84337.9	91402.4	100759.0	78345.6	66913.8	79906.5	83704.3	90209.0	84864.8	85812.1	94585.9	87077.6
CZ	69746.3	77410.9	68953.6	68994.2	73271.4	62863.6	71748.9	83132.8	84189.1	58670.7	73327.8	87311.9	81125.0	87230.9	77165.5	75612.3	72554.3	75337.7
DE	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	59747.9	77621.0	73230.8	74206.0	62044.0	67033.0	69223.0	71703.3
DK	NaN	NaN	NaN	NaN	NaN	66189.4	70268.7	82524.5	100638.7	65294.3	77055.3	94265.9	91421.8	94474.1	89760.0	87401.8	85464.7	89889.7
ES	51523.7	52781.0	50347.8	53424.9	51883.4	51057.1	47539.5	59540.7	59207.7	42306.1	92411.3	102818.9	107211.5	105757.0	88187.0	101850.3	96169.1	96554.3
FI	58576.1	60232.3	59536.4	60208.8	63203.1	60482.8	65704.5	73792.9	85230.8	60556.0	69524.6	87714.7	83370.8	85681.8	71473.0	74527.7	71326.8	74967.3
FR	NaN	NaN	NaN	NaN	NaN	78435.4	88051.1	102395.3	111026.7	79754.4	94313.0	111236.2	108228.0	108471.9	93682.5	96988.9	99609.3	99941.0
HR	NaN	NaN	NaN	NaN	NaN	71495.4	73259.4	86410.2	95652.9	66325.8	90171.2	108211.4	103008.8	98760.7	84571.0	92512.6	80190.5	84917.7
HU	NaN	NaN	NaN	NaN	NaN	34133.1	36748.1	88345.9	106198.8	78826.9	90870.6	110973.0	109899.1	107871.9	96114.3	102906.3	98538.7	100197.2
IE	NaN	NaN	NaN	NaN	NaN	NaN	NaN	53094.9	65580.8	49862.6	55996.6	66354.4	66021.0	67009.7	55567.7	58363.8	57583.7	59170.7
IT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	83809.4	98784.6	97812.7	99271.7	89559.6	90390.7	87368.2	92503.2
LT	30011.6	33159.4	38263.4	30292.3	30845.1	73824.9	91765.2	105385.2	112983.1	60960.3	78934.6	106737.2	92102.1	92497.6	82926.3	85390.7	78856.7	83405.3
LU	39060.0	42217.4	40693.3	42790.4	50687.9	53136.1	56640.3	65666.5	70820.1	51076.3	62935.6	69607.0	71574.5	63080.6	59759.0	59525.6	57891.7	59836.1
LV	56323.7	61344.0	60384.7	58674.5	65208.6	66508.0	75046.1	87377.8	85119.8	58551.2	71519.0	89069.9	80756.9	74314.1	73475.1	74726.1	72215.5	79394.5
MT	52385.5	57113.6	52397.6	55070.9	47327.9	49677.8	49358.2	55912.3	58414.9	57437.4	55328.6	56765.3	66225.8	59508.6	52607.7	60337.9	59324.5	57308.6
NL	71412.0	79643.4	71647.4	73674.1	71691.3	70678.6	81761.3	89668.6	96949.8	72082.0	86062.0	97466.8	97687.3	96830.0	82512.2	86390.5	82495.3	92690.6
PL	NaN	NaN	NaN	NaN	NaN	69866.1	77554.4	95868.9	89094.5	75458.6	85813.8	102214.9	92940.6	96209.4	87859.7	88089.4	90075.0	96778.9
PT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	80381.5	85989.1	89166.2	91701.8	76475.9	82758.4	83708.1	80900.7
RO	25884.5	39732.3	31329.9	45772.4	43604.6	54091.9	66279.0	75609.8	94401.9	82726.5	85592.0	104610.6	111784.0	112394.9	93238.9	89891.7	89053.3	101330.4
SE	NaN	NaN	NaN	NaN	NaN	57986.2	65081.8	76126.4	87608.3	64050.1	70974.4	80083.5	76499.0	74945.5	70299.9	74416.3	70917.8	76041.2
SI	62985.3	68412.0	67158.4	67830.0	64990.8	60808.5	71114.5	78984.0	97633.2	66993.7	70853.8	89046.0	84573.3	92890.4	76533.3	81950.1	79583.8	79477.5
SK	NaN	97086.6	87262.9	83717.1	82926.0	77198.5	77483.8	81589.4	94619.0	66543.9	75258.4	98255.9	85033.8	87076.0	77922.9	75704.7	74824.6	77145.6

EXPLORATORY DATA ANALYSIS (EDA)

Descriptive Analysis

Starting the exploratory data analysis stage, it is first necessary to describe the variables to understand how they behave, the data range and some basic statistical metrics.

	value
count	259730.000000
mean	57.177159
std	59.819649
min	-90.300000
25%	0.500000
50%	68.500000
75%	107.900000
max	1231.100000

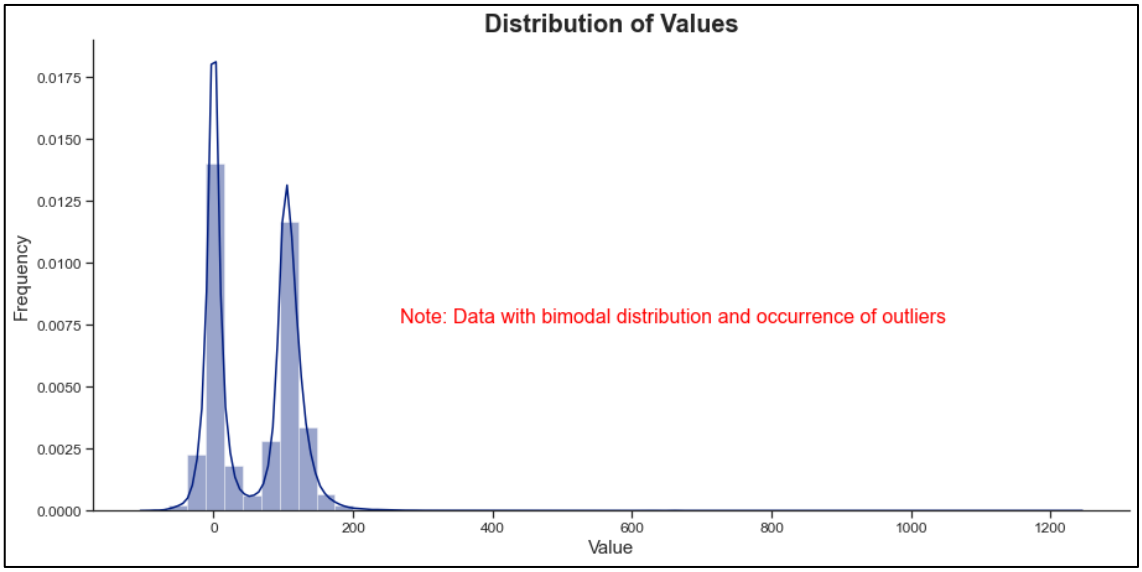
The data provides a concise overview of the 'value' variable's activity, revealing a total of 259,730 records. The average value is roughly 57.18, with a standard deviation of approximately 59.82, indicating a significant amount of variation from the average. The values go from -90.30 to 1231.10, covering an extensive range of values. Quartiles indicate that 25% of the values are equal to or lower than 0.50, 50% are equal to or lower than 68.50 (the median), and 75% are equal to or lower than 107.90. This information helps to understand the distribution of the data.

	type	country_code	country_code_a3	Country	year	quarter	product	class
count	259730	259730	259730	259730	259730	259730	259730	259730
unique	2	25	25	25	7	8	129	29
top	outputs	FR	FRA	France	2011	3	200000	ANIMAL FEEDINGSTUFFS
freq	130022	12544	12544	12544	37208	34484	2800	28664

Each category contains a total of 259,730 records. The categories 'type' and 'country_code_a3' have 2 and 25 distinct values, respectively. The most common values in these categories are 'outputs' and 'FRA'. France is the most prevalent country, with 12,544 records representing it. The years exhibit 7 distinct values, with 2011 being the most common. The third quarter being the most frequent. Additionally, there are 29 distinct classifications, with 'ANIMAL FEEDINGSTUFFS' being the most commonly occurring.

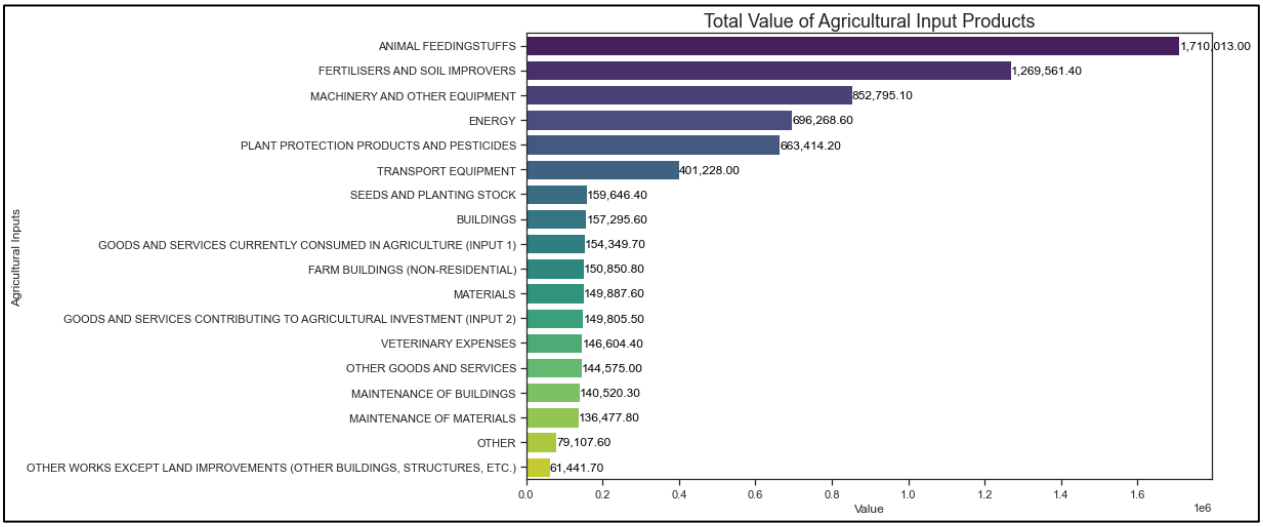
Graphical Analysis

Data Overview



The graph illustrates the distribution of the variable 'value' and exhibits a bimodal distribution, suggesting the existence of two different peaks in the frequency of values. This implies that the data may consist of two primary subsets with distinct features. Moreover, there are outliers present, which can be identified by values that are very far from the major peaks, especially at higher values.

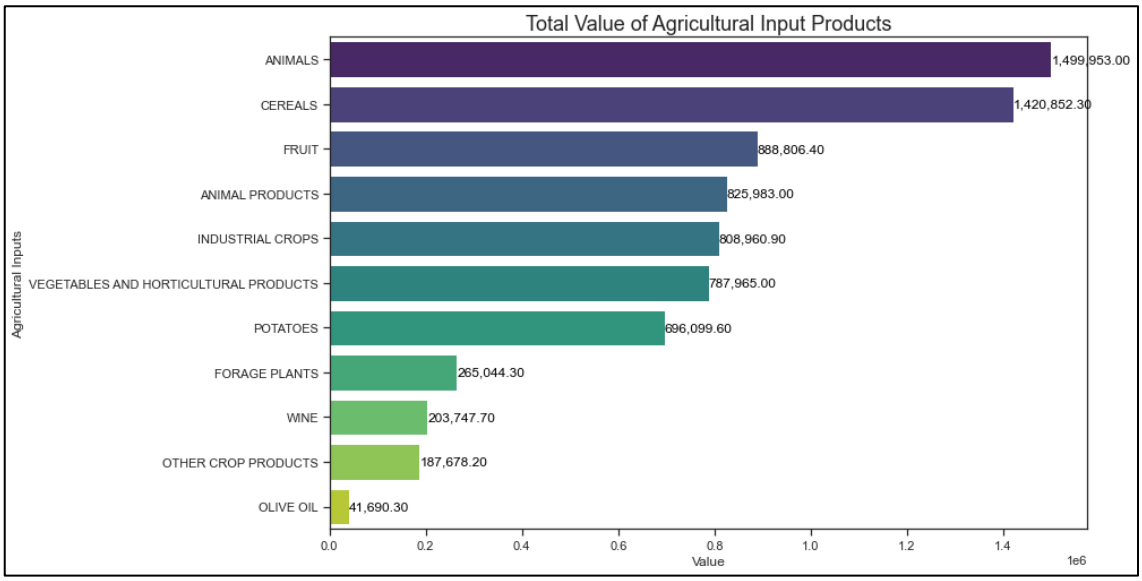
Plotting the histogram of Values - Agricultural Inputs



The graph displays the total amount of agricultural input values for each product. The category 'ANIMAL FEEDINGSTUFFS' has the highest total value of around 1,710,013.00. It is followed by 'FERTILISERS AND SOIL IMPROVERS' and

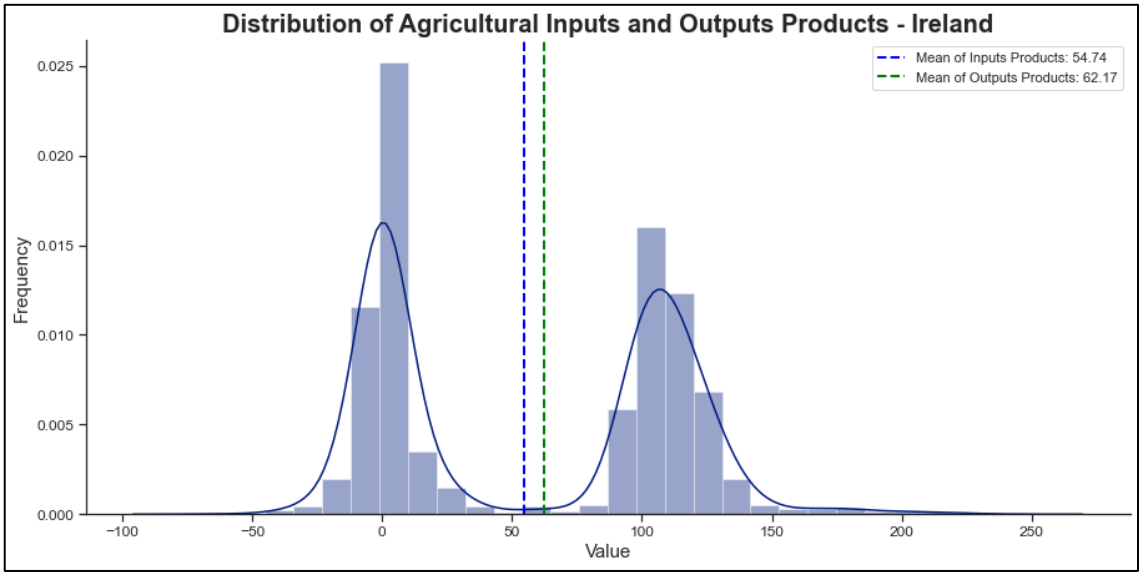
'MACHINERY AND OTHER EQUIPMENT', with values of approximately 1,269,561.40 and 852,795.10, respectively. Additional products such as 'ENERGY', 'PLANT PROTECTION PRODUCTS AND PESTICIDES', and 'TRANSPORT EQUIPMENT' also hold considerable values, whilst products like 'OTHER WORKS EXCEPT LAND IMPROVEMENTS' have the lowest values.

Plotting the histogram of Values - Agricultural Outputs

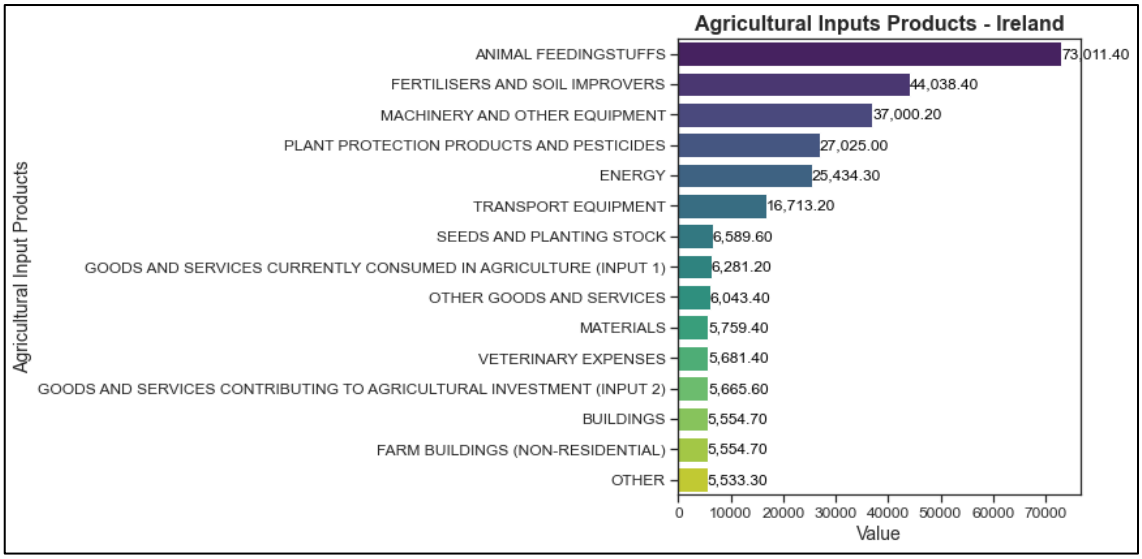


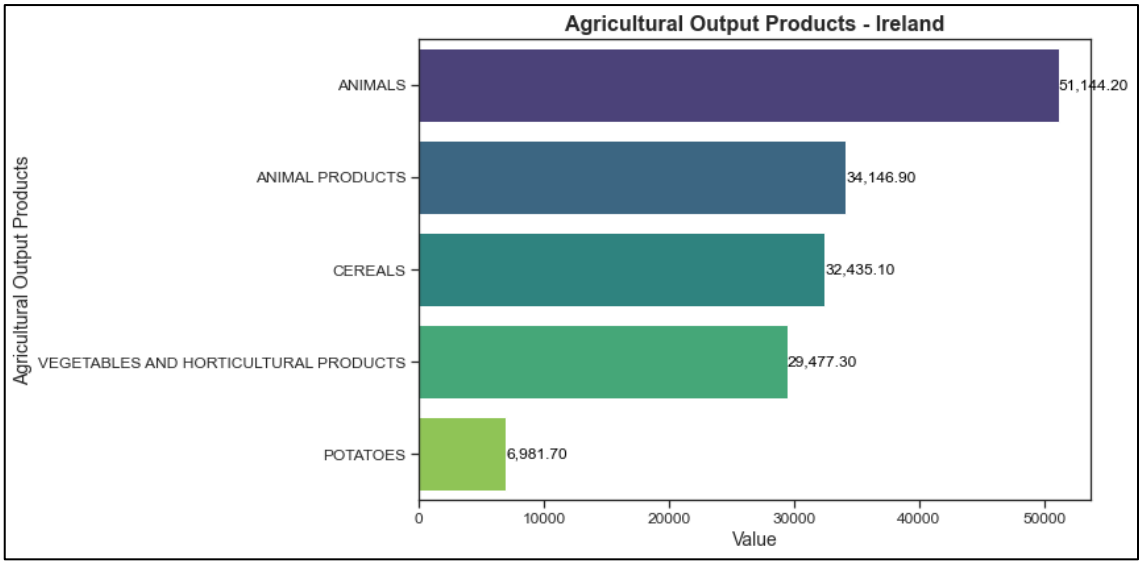
The graph illustrates the cumulative sum of the values of agricultural products generated. The category 'ANIMALS' has the highest total value of around 1,499,953.00, followed by 'CEREALS' with approximately 1,420,852.30, and 'FRUIT' with approximately 888,806.40. Additional commodities, including 'ANIMAL PRODUCTS', 'INDUSTRIAL CROPS', and 'VEGETABLES AND HORTICULTURAL PRODUCTS', exhibit substantial worth. Conversely, 'OTHER CROP PRODUCTS' and 'OLIVE OIL' possess the least value.

Ireland Overview

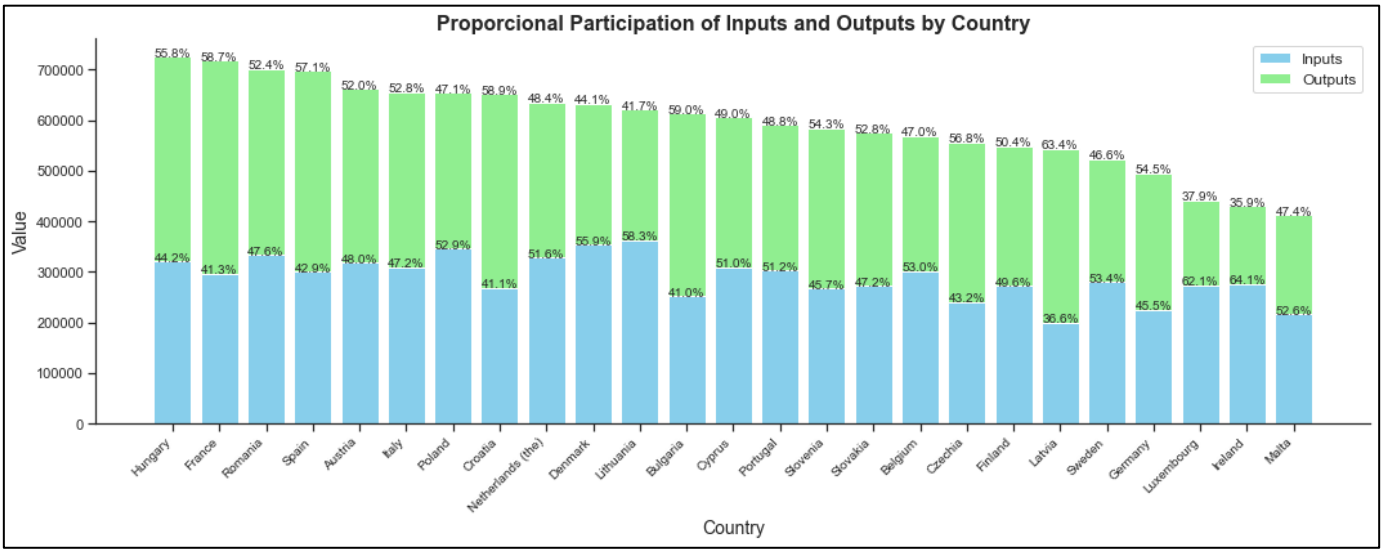


The graph illustrates the distribution of the 'value' variable for Ireland, encompassing both agricultural inputs and products. The distribution has a bimodal pattern, signifying the existence of two different peaks in the frequency of values. The blue dashed line indicates the mean value of agricultural inputs, which is 54.74, whereas the green dashed line reflects the mean value of agricultural goods produced, which is 62.17.





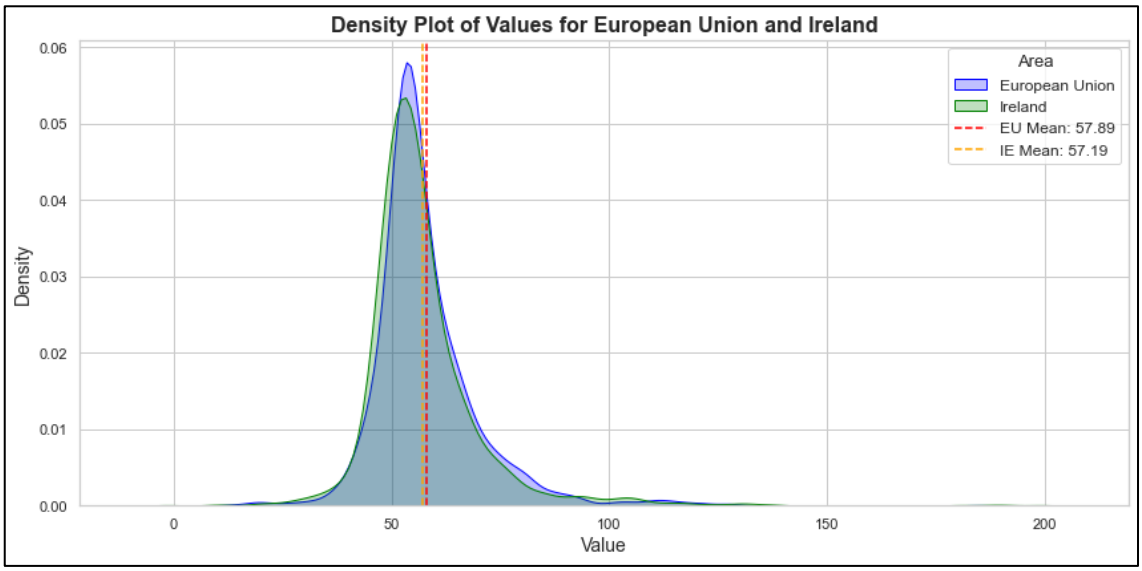
Comparison of Ireland with Other European Union Countries



The graph illustrates the distribution of agricultural inputs and products among European Union countries. It reveals that Ireland is responsible for 64.1% of inputs and 35.9% of goods, indicating a significant investment in inputs in relation to final products. Luxembourg shows a similar pattern, in which 62.1% of its trade involves inputs and 37.9% involves outputs. Also with similar behavior, Sweden has 53.4% inputs and 43.6% outputs, while Hungary, which appears in first position, has around 44.2% inputs and 55.8% outputs.

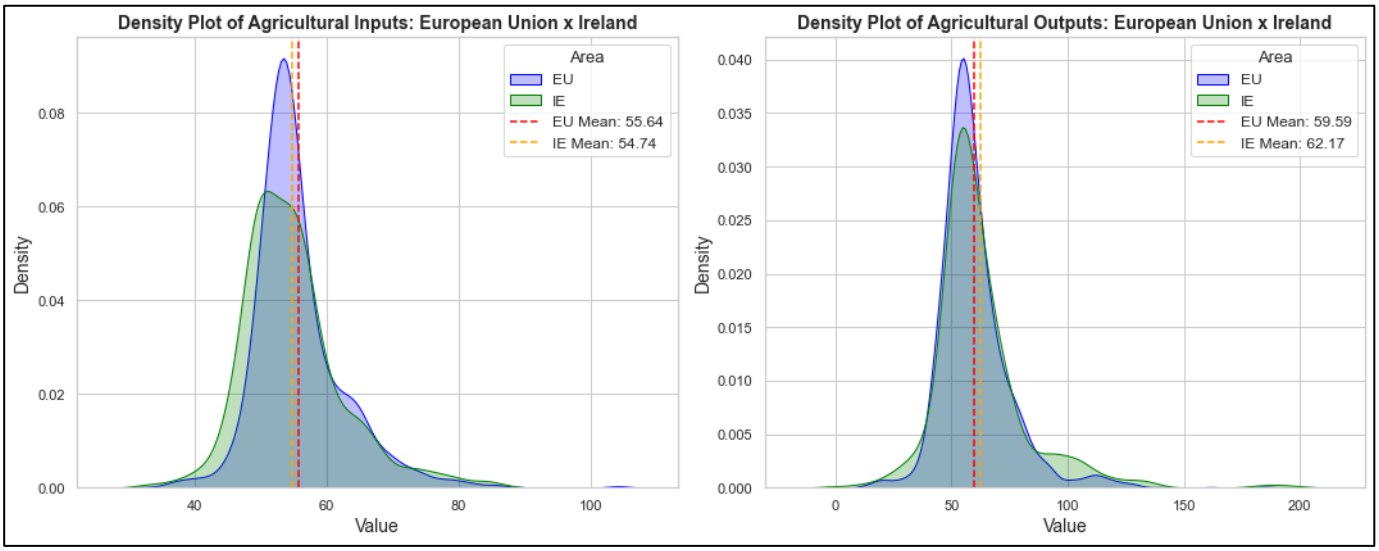
Country	Inputs	Outputs	Total
Hungary	321,022.90	405,477.60	726,500.50
France	296,889.60	421,268.20	718,157.80
Romania	333,954.50	368,349.30	702,303.80
Spain	299,935.20	398,612.90	698,548.10
Austria	317,842.80	343,982.70	661,825.50
Italy	309,491.20	346,199.50	655,690.70
Poland	345,842.50	308,325.40	654,167.90
Croatia	268,033.40	384,139.30	652,172.70
Netherlands (the)	327,917.00	308,155.70	636,072.70
Denmark	353,924.60	278,753.40	632,678.00
Lithuania	362,428.50	259,487.40	621,915.90
Bulgaria	251,975.70	362,083.70	614,059.40
Cyprus	308,996.80	297,163.40	606,160.20
Portugal	302,313.70	288,386.50	590,700.20
Slovenia	267,008.40	317,046.00	584,054.40
Slovakia	271,926.00	304,037.50	575,963.50
Belgium	301,344.20	267,267.90	568,612.10
Czechia	240,168.80	316,168.80	556,337.60
Finland	272,546.60	276,515.50	549,062.10
Latvia	199,297.70	344,654.40	543,952.10
Sweden	279,383.00	243,820.20	523,203.20
Germany	225,187.10	269,874.00	495,061.10
Luxembourg	273,913.90	167,360.60	441,274.50
Ireland	275,885.80	154,185.20	430,071.00
Malta	216,612.80	195,465.60	412,078.40

Comparing Agricultural Input and Output Products: Mean of Ireland with Mean of European Union



The density graph illustrates the comparative distribution of agricultural input and production values between Ireland and the European Union. The green curve corresponds to Ireland, whilst the blue curve corresponds to the European Union. The European Union average, indicated by the red line, stands at 57.89, while the Irish average, represented by the yellow line, is slightly lower at 57.19. Both distributions exhibit a high degree of similarity, suggesting that Ireland's agricultural inputs and

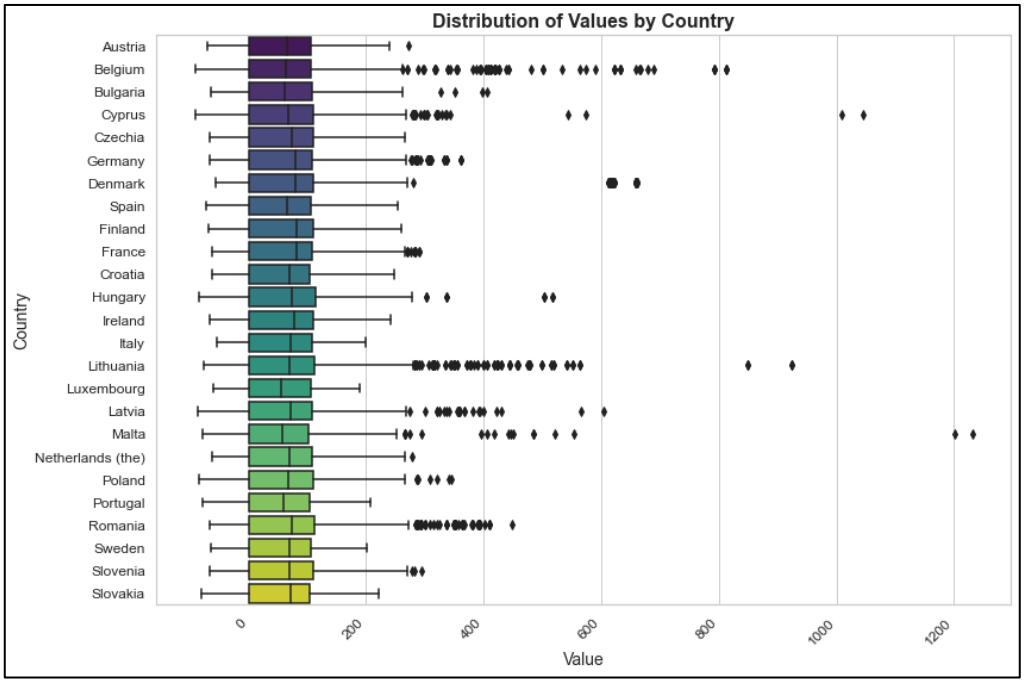
products adhere to a distribution pattern that closely resembles the average distribution of the European Union.



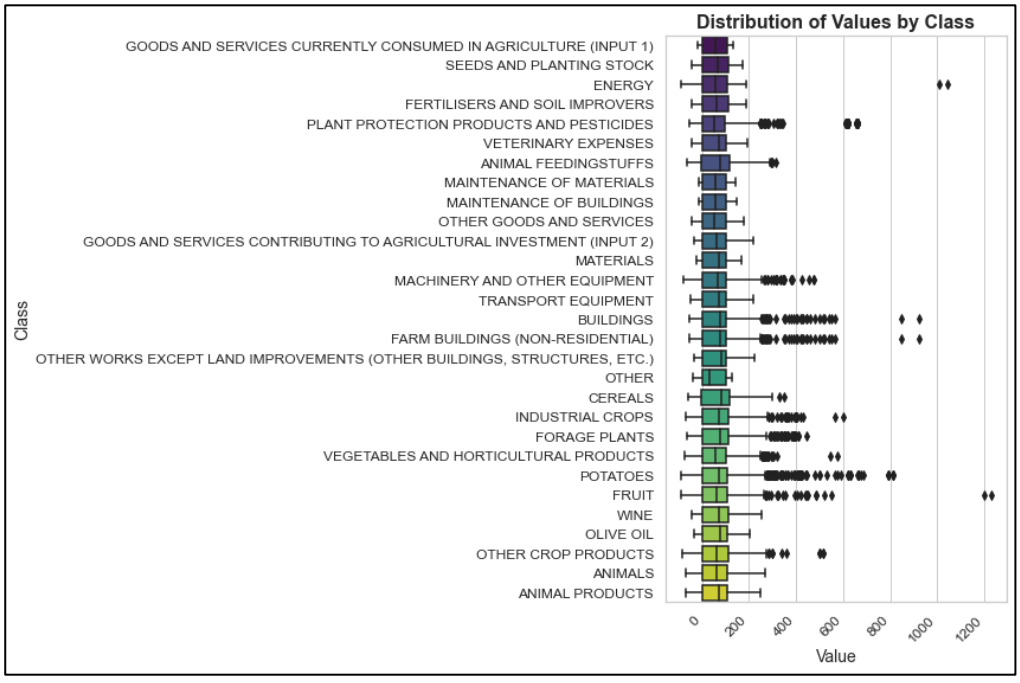
Density plots analyse the distribution of agricultural input and production values in Ireland and the European Union, allowing for a comparison between the two. The left graph illustrates the inputs, with the green curve representing Ireland and the blue curve representing the European Union. The mean input for the European Union is 55.64 (shown by the red line), whereas for Ireland it is 54.74 (represented by the yellow line). The graph on the right depicts the agricultural products generated, using the same colour scheme. The mean number of items manufactured is 59.59 for the European Union (shown by the red line) and 62.17 for Ireland (represented by the yellow line).

Outliers Analysis

Checking the dataset outliers considering variables Class and Country



The research demonstrates a substantial range of values and the notable occurrence of outliers in many nations, including Cyprus, Lithuania, and Malta, suggesting a diversity in agricultural data. Certain countries, such as Ireland and France, exhibit lower levels of dispersion and a reduced number of outliers. Outliers in the data may suggest the presence of extremely diverse values that deviate greatly from the rest of the data. This could be attributed to large variances in agricultural inputs or production. This trend has the potential to impact data analysis and interpretation, emphasizing the importance of taking these outliers into account in any subsequent statistical study.



For the variable ‘class’, the study demonstrates that categories such as 'BUILDINGS', 'FARM BUILDINGS', and 'FRUIT' exhibit a larger spread of values and a notable number of outliers, suggesting a high degree of variability in the data. The classes 'CEREALS' and 'FORAGE PLANTS' exhibit lower dispersion and a reduced number of outliers, indicating a higher level of data consistency.

Removing Outliers considering the variables Class and Country

Country	country_code	mean_previous	sd_previous	mean_after	sd_after
Austria	AT	54.841357	55.948945	54.605638	55.534529
Belgium	BE	57.043750	68.475006	54.958083	61.248362
Bulgaria	BG	56.668457	58.319654	56.265038	57.726402
Croatia	HR	54.474833	56.172743	53.860232	55.514390
Cyprus	CY	58.987953	61.869322	57.640911	58.311971
Czechia	CZ	57.711369	58.298079	57.494354	57.973992
Denmark	DK	60.578131	68.933177	58.196323	58.969608
Finland	FI	58.361193	58.857083	58.202298	58.628802
France	FR	57.251100	58.284517	56.416143	57.156098
Germany	DE	58.545542	60.345841	57.348846	58.199986
Hungary	HU	60.481227	61.722378	59.905925	60.478110
Ireland	IE	57.190293	57.392117	56.592204	56.686691
Italy	IT	55.889081	55.883206	55.706530	55.716522
Latvia	LV	58.514641	61.422174	56.749261	58.173144
Lithuania	LT	61.698006	70.839270	59.448195	65.914972
Luxembourg	LU	53.971930	54.990771	53.922829	54.852180
Malta	MT	54.106933	60.854829	53.026559	55.426486
Netherlands (the)	NL	56.499618	57.861544	55.862846	57.135547
Poland	PL	58.117262	58.994997	57.549799	58.066453
Portugal	PT	53.273828	54.445013	53.145939	54.243825
Romania	RO	60.879317	63.524465	59.542567	61.366399
Slovakia	SK	54.562666	56.636122	54.079618	56.104420
Slovenia	SI	58.995394	60.231792	58.034285	58.940904
Spain	ES	56.571761	57.673167	55.825175	56.794390
Sweden	SE	53.694910	53.896581	53.662437	53.720535

An examination of the disparities in averages and variations before and after eliminating anomalies for multiple European Union nations reveals a minor decrease in averages and variations in the majority of instances. Overall, the averages declined marginally after excluding extreme values, as evidenced in Belgium (from 57.04 to 54.96) and Denmark (from 60.58 to 58.20). Similarly, the standard deviations decreased as well, indicating a decrease in the range of values in the data set after removing outliers. This can be seen in the examples of Lithuania (from 70.84 to 65.91) and Germany (from 60.35 to 58.20).

class	mean_previous	sd_previous	mean_after	sd_after
ANIMAL FEEDINGSTUFFS	59.657166	59.669849	59.243788	59.141746
ANIMAL PRODUCTS	57.280374	57.204542	56.579670	56.449255
ANIMALS	55.570280	55.593071	55.277305	55.195593
BUILDINGS	56.177000	67.524398	55.401752	63.066137
CEREALS	62.992210	65.008152	62.565167	64.567521
ENERGY	54.720890	57.181040	54.491726	55.804449
FARM BUILDINGS (NON-RESIDENTIAL)	56.120089	68.000417	55.312439	63.383825
FERTILISERS AND SOIL IMPROVERS	56.235002	57.392348	56.229059	57.386672
FORAGE PLANTS	62.687867	68.846740	60.989833	65.414564
FRUIT	57.714701	61.456333	55.931975	57.228752
GOODS AND SERVICES CONTRIBUTING TO AGRICULTURA...	53.501964	52.971573	53.443873	52.891758
GOODS AND SERVICES CURRENTLY CONSUMED IN AGRIC...	55.124893	54.740260	55.124893	54.740260
INDUSTRIAL CROPS	59.976342	62.154598	58.572397	59.730377
MACHINERY AND OTHER EQUIPMENT	54.387443	55.199366	53.675368	53.698976
MAINTENANCE OF BUILDINGS	52.276897	51.704385	52.276897	51.704385
MAINTENANCE OF MATERIALS	52.980512	52.220413	52.980512	52.220413
MATERIALS	53.531286	52.738302	53.531286	52.738302
OLIVE OIL	62.039137	60.752747	62.039137	60.752747
OTHER	50.451276	51.232576	50.451276	51.232576
OTHER CROP PRODUCTS	62.227520	66.644837	60.011926	62.490727
OTHER GOODS AND SERVICES	51.633929	51.466451	51.550393	51.389489
OTHER WORKS EXCEPT LAND IMPROVEMENTS (OTHER BU...	54.858661	56.844338	54.858661	56.844338
PLANT PROTECTION PRODUCTS AND PESTICIDES	53.848555	63.142190	51.478729	52.871207
POTATOES	63.247283	76.004670	59.981674	68.076070
SEEDS AND PLANTING STOCK	57.016571	56.154254	57.016571	56.154254
TRANSPORT EQUIPMENT	53.468550	53.001695	53.355561	52.868373
VEGETABLES AND HORTICULTURAL PRODUCTS	52.741968	55.928986	51.046397	53.346085
VETERINARY EXPENSES	52.358714	52.102598	52.358714	52.102598
WINE	61.666979	62.883560	60.347407	61.330936

Upon removing outliers, the majority of classes had a marginal decline in both means and standard deviations, indicating a decrease in the variability of the data. As an illustration, the average value of the 'BUILDINGS' category decreased from 56.18 to 55.40, and the measure of variability, known as the standard deviation, decreased from 67.52 to 63.07.

STATISTICAL ANALYSIS

Measures of Central Tendency

An examination of the major patterns in agricultural goods in Ireland and the European Union demonstrates that the average value of agricultural inputs in Ireland is 54.74, whereas the average value of agricultural products produced is 60.11. This indicates that the products generated have a greater average worth than the inputs. The median value for inputs in Ireland is 53.53, whereas for outputs it is 56.50, thereby confirming this pattern. The mean value for agricultural inputs in the European Union is 55.12, while the mean value for agricultural products is 58.07. The median value for agricultural inputs is 53.95, and for agricultural products, it is 55.99.

Variability Measures

Ireland's agricultural inputs have a coefficient of variation of 14.06% and a standard deviation of 7.70, indicating substantial dispersion around the mean, according to an analysis of measures of variation for agricultural outputs in Ireland and the European Union. Significantly higher standard deviation (19.86) and coefficient of variation (23.59%) of agricultural products produced in Ireland point to higher variability. Comparatively speaking, agricultural inputs in the European Union show less variability than in Ireland with a coefficient of variation of 11.82% and a standard deviation of 6.51. Ireland and the European Union have comparable relative variability of final products despite variations in absolute standard deviations; agricultural products produced in the EU have a standard deviation of 13.70 and the same coefficient of variation of 23.59%. This implies that although the absolute values of agricultural products in Ireland vary more, the relative fluctuation is comparable to that of the European environment.

Analysis of Statistical Inferences for Agricultural Products

- **Agricultural INPUTS Products:**

  Ireland:

  - **Population Mean:** 54.74
  - **Estimated Sample Mean:** 55.97
  - **Population Standard Deviation:** 7.70
  - **Estimated Sample Standard Deviation:** 7.74
  - **Confidence Interval of the Population Mean:** (54.44, 57.50)

  European Union:

  - **Population Mean:** 55.12
  - **Estimated Sample Mean: 55**.21
  - **Population Standard Deviation:** 6.51
  - **Estimated Sample Standard Deviation:** 6.80
  - **Confidence Interval of the Population Mean:** (53.99, 56.43)

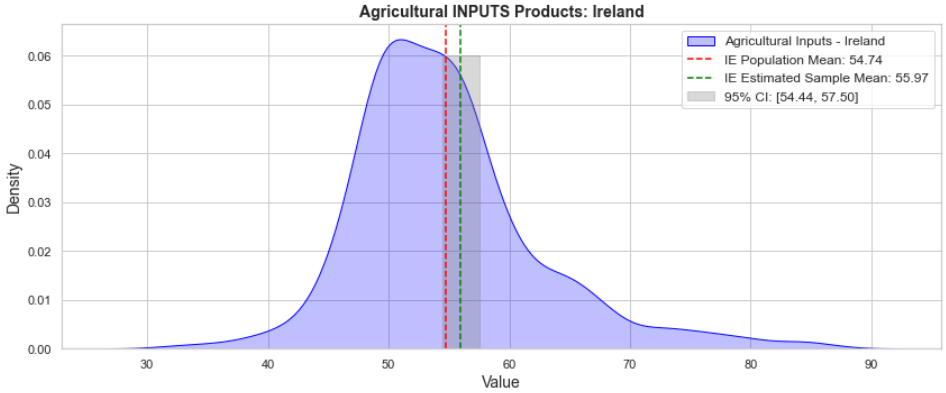
- **Agricultural OUTPUTS Products:**

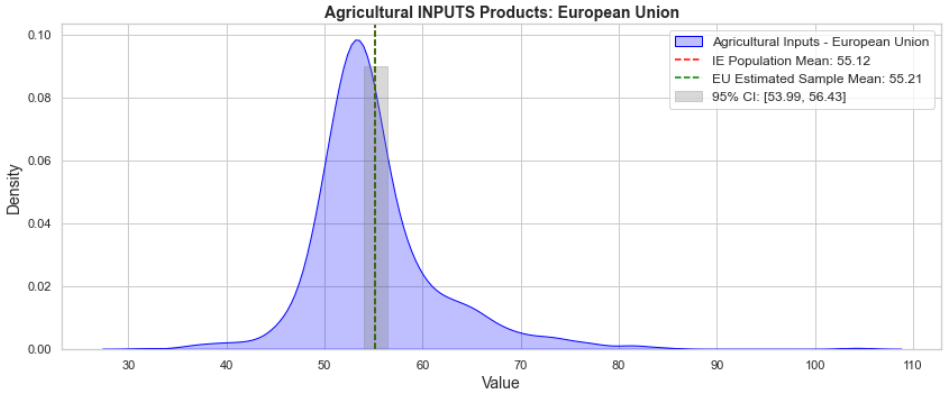
  Ireland:

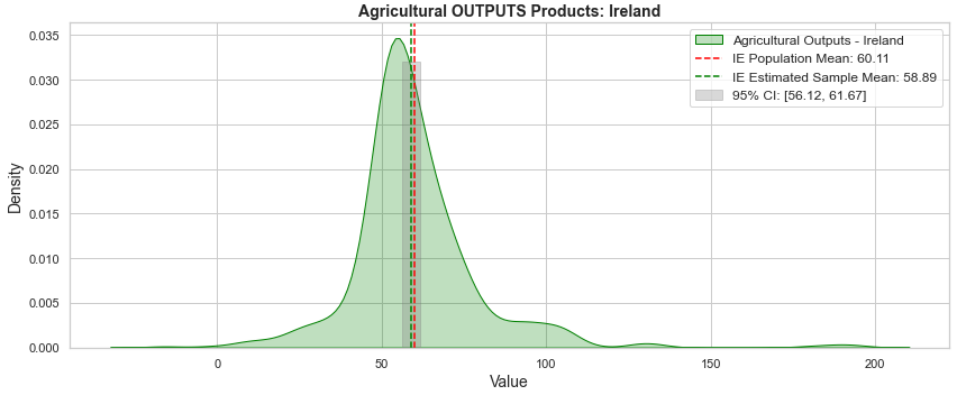
  - **Population Mean:** 60.11
  - **Estimated Sample Mean:** 58.89
  - **Population Standard Deviation:** 19.86
  - **Estimated Sample Standard Deviation:** 10.97
  - **Confidence Interval of the Population Mean:** (56.12, 61.67)

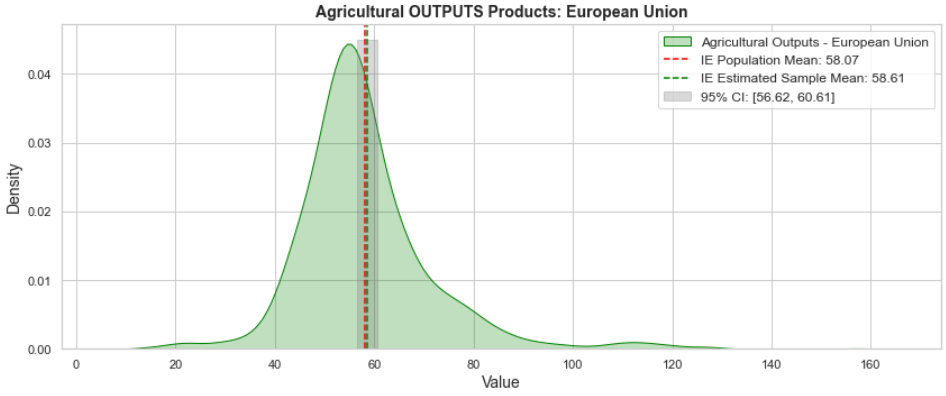
European Union:

- **Population Mean:** 58.07
- **Estimated Sample Mean:** 58.61
- **Population Standard Deviation:** 13.70
- **Estimated Sample Standard Deviation:** 14.25
- **Confidence Interval of the Population Mean:** (56.62, 60.61)









- **General Observations:**

1. **Comparison of Means:**

- In Ireland, the mean of agricultural inputs is slightly lower than the mean of agricultural outputs.

- In the European Union, the mean of agricultural inputs is also lower than the mean of agricultural outputs.

2. **Standard Deviations:**

- The population standard deviations for both inputs and outputs in Ireland are higher than in the European Union, indicating greater variability in the Irish data.

- The small difference between sample and population standard deviations suggests that the samples are representative.

3. Confidence Intervals:

- The confidence intervals for the population means show that the estimates are precise, with sample means falling within the confidence intervals.

- The confidence intervals are slightly wider in Ireland for agricultural outputs, reflecting the observed higher variability.

These statistical inferences provide a detailed and accurate view of the central tendencies and variability of agricultural products in Ireland and the European Union, allowing for a robust comparison between the two regions.

Statistical Tests - Parametric and Non-Parametric

The study of the tests will be carried out in two ways. Characteristics of Agricultural Inputs and Outputs Products will be studied among:

- Ireland x European Union
- Ireland x Countries with Similarities

Shapiro-Wilk Test (Parametric Test)

Apply some tests that assume data with normal distribution, we will first check this assumption using the Shapiro-Wilk test.

- **Comparison**: Agricultural Inputs and Outputs Products - Normal Distribution or not.

- **Test Proposal**: Check if the Agricultural Inputs and Outputs Products values are normally distributed.

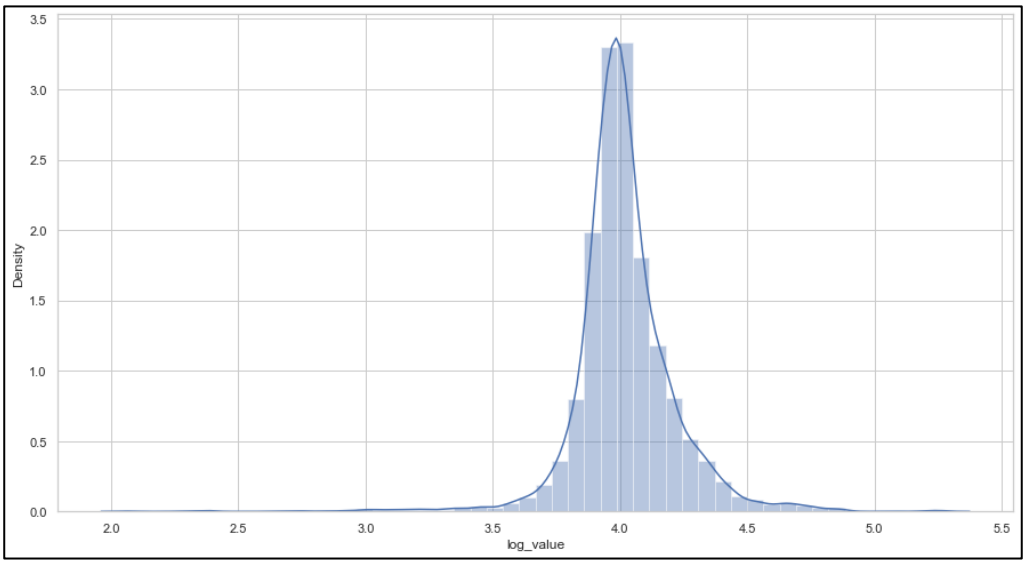
- **- Hypotheses:**
  - Null Hypothesis (H0): The data follows a normal distribution.
  - Alternative Hypothesis (H1): The data does not follow a normal distribution.

1. Applying with the Original Data

- Shapiro-Wilk Test Statistic: 0.8170347213745117
- P-value: 0.0

- The data does not follow a normal distribution (we reject H0).

2. Applying the Log-transformation in the Data



- Shapiro-Wilk Test Statistic with Log-transformation: 0.17034
- P-value with Log-transformation: 1.0

- The data follows a normal distribution (we do not reject H0).

Independent Two-Sample t-Test (Parametric Test)

- **Comparison**: Agricultural Inputs and Outputs Products - Mean of Ireland x Mean of European Union.

- **Test Proposal:** Compare Ireland's Agricultural Inputs and Outputs Products averages with the average for European Union countries (excluding Ireland) and determine whether the averages differ significantly.

- **Hypothesis**:
  - Null Hypothesis (H0): There is no difference in the mean of Agriculture Inputs and Outputs Products between Ireland and European Union countries (without Ireland).
  - Alternative Hypothesis (H1): There is a difference in the mean between Ireland and European Union.

- **Results**

➢ Parametric t-test p-value: 0.09878009481915333

➢ Parametric t-test: Fail to reject null hypothesis. There is no significant difference in the means between Ireland and European Union.

Mann-Whitney U Test (Non-Parametric Test)

- **Comparison**: Agricultural Inputs and Outputs Products - Mean of Ireland x Mean of European Union.

- **Test Proposal**: Compare Ireland's Agricultural Inputs and Outputs Products averages with the average for European Union countries (excluding Ireland) and determine whether the averages differ significantly.

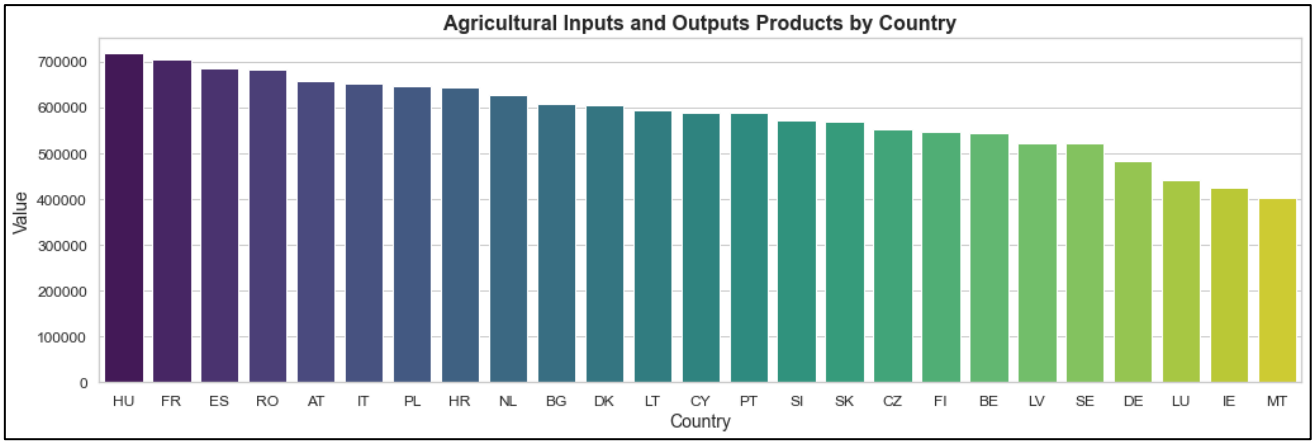
- **Hypothesis**:
  - Null Hypothesis (H0): There is no difference in the mean of Agriculture Inputs and Outputs Products between Ireland and European Union countries (without Ireland).
  - Alternative Hypothesis (H1): There is a difference in the mean between Ireland and European Union.

➢ **Results**

➢ Non-parametric Mann-Whitney U test p-value: 0.012400667208821284

- Non-parametric Mann-Whitney U test: Reject null hypothesis. There is a significant difference in the means between Ireland and European Union.

Checking the values by country



The countries Malta (MT), Portugal (PT) and Luxembourg (LU) have Agricultural Inputs and Outputs Products total values similar to Ireland. We will use the following tests to find out if these similarities are really significant.

One-Way ANOVA (Parametric Test)¶

The ANOVA test assumes data normality. This way, we will use log-transformation to make our data Normally distributed and add a constant to avoid NA and Inf values.

- **Comparison**: Agricultural Inputs and Outputs Products - Means comparison between (Ireland x Malta x Portugal x Luxembourg)

- **Test Proposal:** Define if the Mean of Agricultural Inputs and Outputs Products for the countries Ireland (IE), Malta (MT), Portugal (PT) and Luxembourg (LU) has a significant difference or if the countries have the same behaviour.

- **Hypotheses**:
  - Null Hypothesis (H0): All group means are equal.
  - Alternative Hypothesis (H1): At least one group mean is different.

- **Results**

- Parametric One-Way ANOVA p-value: 0.0005404903695815851

- Parametric One-Way ANOVA: Reject null hypothesis. There is a significant difference between the means of the countries Ireland, Malta, Portugal and Luxembourg.

Kruskal-Wallis Test (Non-Parametric Test)

- **Comparison**: Agricultural Inputs and Outputs Products - Distributuion Population comparison between (Ireland x Malta x Portugal x Luxembourg)

- **Test Proposal:** Comparison of the medians of Agricultural Inputs and Outputs Products for the countries Ireland (IE), Malta (MT), Portugal (PT) and Luxembourg (LU).

- **Hypotheses:**

  - Null Hypothesis (H0): Population distributions are equal.

  - Alternative Hypothesis (H1): At least one of the population distributions is different.

- **Results**

- Non-parametric Kruskal-Wallis Test p-value: 2.309473471900063e-16

- Non-parametric Kruskal-Wallis test: Reject the null hypothesis. There is a significant difference between the population distributions of the countries Ireland, Malta, Portugal and Luxembourg.

MACHINE LEARNING

Sentiment Analysis on Agriculture Using Reddit

A sentiment analysis on agriculture was conducted using data collected from Reddit. The process involved configuring connection variables and environment settings to access Reddit's API. Posts and comments related to agriculture were collected, and sentiment analysis was performed on the gathered data.

The analysis included the following steps:

- **Configuration**: Connection variables and environment settings were configured to ensure seamless access to Reddit's API, including necessary authentication keys and tokens.

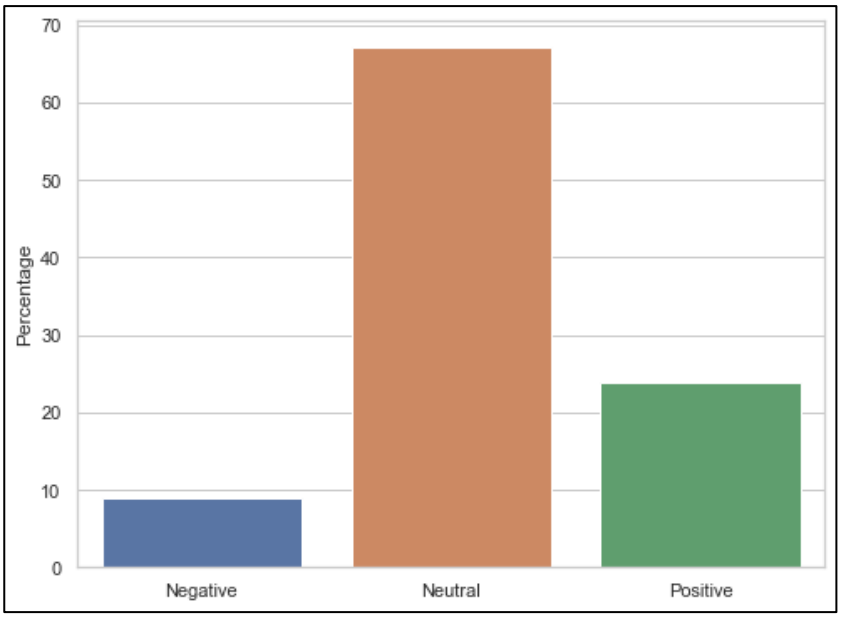
- **Data Collection**: Posts and comments about agriculture were collected from various subreddits. This involved querying relevant keywords and topics to gather a comprehensive dataset.


```
[{'neg': 0.0,
  'neu': 1.0,
  'pos': 0.0,
  'compound': 0.0,
  'headline': 'Paramérica: Semillas mejoradas en la agricultura moderna'},
 {'neg': 0.0,
  'neu': 1.0,
  'pos': 0.0,
  'compound': 0.0,
  'headline': 'Anyone have tips for a young western Canadian wanting to learn more '},
 {'neg': 0.0,
  'neu': 1.0,
  'pos': 0.0,
  'compound': 0.0,
  'headline': 'What are the operating modes of large-scale agriculture in plain areas?'}]
```


- **Sentiment Analysis**: The collected data was analyzed to determine the overall sentiment (positive, negative, or neutral) expressed in the posts and comments. Sentiment analysis algorithms were used to classify the sentiments.

	neg	neu	pos	compound	headline	class
0	0.000	1.000	0.000	0.0000	Paramérica: Semillas mejoradas en la agricultu...	0
1	0.000	1.000	0.000	0.0000	Anyone have tips for a young western Canadian ...	0
2	0.000	1.000	0.000	0.0000	What are the operating modes of large-scale ag...	0
3	0.000	1.000	0.000	0.0000	Small scale Biofuel education	0
4	0.196	0.804	0.000	-0.2960	Cash corn is signaling basis may finally feel ...	-1
...	...	...	...	...	...	...
885	0.000	0.860	0.140	0.0772	I am a aviation technician that want to get in...	0
886	0.099	0.901	0.000	-0.6062	Hi guys! Does anyone know what this is called?...	-1
887	0.000	1.000	0.000	0.0000	Looking for feedback on a concept for on-farm ...	0
888	0.000	1.000	0.000	0.0000	What to do with used farm plastics?	0
889	0.000	0.588	0.412	0.6369	Best source for raw material prices regionally?	1

- **Graphical Representation**: The results of the sentiment analysis were presented graphically using visualizations included sentiment distribution charts.



The graphical analysis provided insights into the general public's perception of agriculture on Reddit, highlighting prevailing sentiments and key topics of discussion.

Random Forest & Support Vector Machine (SVM) Algorithms

- **Proposal**:

Utilizing the Random Forest and SVM algorithms to categorize agricultural data entries according to the "Country" response variable. Employing collected and processed data to forecast the nation of origin for each entry.

- **Application of Random Forest and SVM Models for Classification of the "Country" Variable**

1. Data Loading and Preparation:

Data was loaded and pre-processed to handle missing values by replacing them with the mean of their respective columns. Categorical variables were encoded using `get_dummies`.

2. Separation of Features and Target:

The dataset was divided into features (`X`) and the target variable (`y`), with "Country" being the target for classification.

3. Data Splitting:

The data was split into training and testing sets with an 80/20 ratio to ensure a representative sample.

4. Data Normalization:

To ensure all features contribute equally to the models, the data was normalized using `StandardScaler`.

5. Model Definition and Training:

Random Forest:

A grid search (`GridSearchCV`) was configured to optimize hyperparameters, testing parameters such as the number of estimators, maximum depth, and the minimum number of samples per leaf.

SVM (Support Vector Machine):

A grid search (`GridSearchCV`) was configured to optimize hyperparameters, testing parameters such as different kernels (linear, polynomial, RBF), regularization parameter (`C`), and kernel parameters (`gamma`).

6. Model Evaluation and Selection:

Both models were trained using `GridSearchCV`, and the best sets of hyperparameters were selected. The models' accuracy was evaluated, resulting in classification reports that include metrics such as precision, recall, and F1-score.

7. Confusion Matrix:

For both models, a confusion matrix was generated to visualize the models' performance in correctly classifying countries. The matrices were plotted for better interpretation of the results.

8. Learning Curve:

A learning curve was generated for both models to assess their performance with different training sample sizes, showing the evolution of training and cross-validation scores.

High accuracy scores of the Random Forest and SVM models indicate that they were both quite successful in forecasting the "Country" variable. Both models' ideal

hyperparameters were effectively found by GridSearchCV, guaranteeing peak performance. Each model's confusion matrices showed how well they classified nations. The learning curves also showed that, with SVM exhibiting strong generalization capabilities and Random Forest exhibiting low variance and low bias, both models adapted well to new data.

Comparative Analysis

Comparative Performance Table: Random Forest vs. SVM

Metric	Random Forest	SVM
Best Hyperparameters	{'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200}	{'C': 10, 'gamma': 0.1, 'kernel': 'linear'}
Accuracy	0.8073	0.7091
Confusion Matrix	[[126, 33], [20, 96]]	[[108, 51], [29, 87]]
Cross-Validation Scores	[0.7727, 0.8318, 0.7900, 0.7626, 0.8311]	[0.7091, 0.7227, 0.7078, 0.6986, 0.7032]
Mean CV Score	0.7976	0.7083

Analysis - Random Forest

Best Hyperparameters: GridSearchCV identified the best hyperparameters for the Random Forest model, including unlimited tree depth, square root of the total number of features, a minimum of 2 samples per leaf, a minimum of 5 samples per split, and 200 estimators.

Accuracy: The model achieved an accuracy of 0.8073.

Confusion Matrix: The model correctly classified 126 instances of class 0 and 96 instances of class 1, with some misclassifications (33 and 20, respectively).

Cross-Validation: Cross-validation scores ranged from 0.7626 to 0.8318, with a mean of 0.7976, indicating consistent performance.

Analysis – Support Vector Machine (SVM)

Best Hyperparameters: For the SVM model, the best hyperparameters included C of 10, gamma of 0.1, and a linear kernel.

Accuracy: The model achieved an accuracy of 0.7091.

Confusion Matrix: The model correctly classified 108 instances of class 0 and 87 instances of class 1, with some misclassifications (51 and 29, respectively).

Cross-Validation: Cross-validation scores ranged from 0.6986 to 0.7227, with a mean of 0.7083, indicating more varied performance compared to Random Forest.

Analysis – Conclusions

Overall Performance: The Random Forest model outperformed the SVM model in terms of accuracy.

Consistency: Random Forest demonstrated slightly better consistency in cross-validation scores compared to SVM.

Feature Importance: Random Forest also allowed for the analysis of feature importance, providing additional insights into the most influential variables, which is not directly possible with SVM.

In summary, Random Forest demonstrated superior and more consistent performance in classifying the "Country" variable compared to SVM, making it the preferred choice for this dataset.

CONCLUSION

The agricultural sector is a fundamental part of Ireland's economy, with complex interactions between inputs and outputs driving productivity and sustainability. This study examines these dynamics, focusing on critical inputs and diverse outputs.

Early data (2000-2004) showed high missing values, which decreased over time, indicating improved data collection. Sentiment analysis of Reddit posts revealed public perceptions of agricultural practices.

Comparing agricultural practices in Ireland and other European countries, the study used Random Forest and SVM models. Random Forest achieved higher accuracy (80.73%) compared to SVM (70.91%), highlighting important features in the dataset.

Statistical analysis showed Ireland's mean for agricultural inputs was 54.74, with a mean output of 60.11, reflecting variability. Similar trends were observed in EU data, indicating diverse practices across member states.

In conclusion, this study provides detailed insights into Ireland's agricultural sector, supporting evidence-based policy, sustainable practices, and cross-border collaboration to enhance productivity and sustainability.