**Ireland and the impact of CAP on Crops and Cereals**

Raúl Martín Sánchez

MSc in Data Analytics (Semester I), Integrated Assignment 2, CCT College Dublin

E-mail: [rmsryu@gmail.com](mailto:rmsryu@gmail.com) / [sbs22021@gmail.com](mailto:sbs22021@gmail.com)

Lectures: Marina Iantorno, Muhammad Iqbal, David McQuaid

Submission Date 24th April 2022

Word count: 2802

Abstract

The CAP [1] is a common policy for all European countries which manages the budget and resource allocation to the different member states, it aims to support farmers to improve agricultural productivity and ensure affordable food across all members. It’s also responsible to tackle climate change and ensuring sustainability of the natural resources. Another important area of development of the CAP is to keep the rural economy alive, promoting jobs and agricultural food industries that are associated with them. With a baseline set on Ireland, it will be detailed in this work how the different aspects of the CAP can be identified in Ireland among other member states on Crops and Cereals and how they can be used to determine a key indicator for the agriculture such as the Gross Value Added (GVA) of agriculture for the Irish economy.

Keywords: Ireland, Agriculture, Crops, Cereals, CAP, Inferential Analysis, Data Visualization, Regression, Sentiment Analysis

Introduction

The relationship between the GVA and the CAP core objectives will be studied by analyzing several country indicators and measurements for the years between 2004 and 2020. These 17 years include 3 main reforms of the CAP. With the last one coming to an end so the full period of study will be a good measure and base of study for the next reform starting on 1st January 2023. To gain a better understanding and perspective on how these reforms have been adopted by the Irish farmers and consumers, sentiment analysis will also be discussed with details gathered from the discussion of the users of the popular blog site boards.ie.

In terms of the different datasets required for the study, the following indicates the data source and details of them:

* ***Crop Mean Residues in Kg***. Average of crop residues by country and year in kilograms. Source: FAO Tier 1.
* ***Crop Production Index:*** Gross Production Index Number (2014-2016 = 100). *Source*: FAO Tier 1.
* ***Cereals Produce Price US$/tonne:*** Annual Producer Price (USD/tonne)
* ***% Employment Ratio In Rural Areas:*** Employment-to-population ratio by age, total (15+), rural areas. *Source*: FAO / Eurostat (\*)
* ***Average weekly working hours:*** Mean weekly hours actually worked per employed person in agriculture, forestry and fishing. *Source:* FAO / International reliable sources (\*)
* ***Crop Land Use in Thousands of ha:*** Reported land used for crops yearly questionnaire. Source: FAO / Official data reported on FAO Questionnaires from countries.
* ***Total Energy Use in Agriculture (Terajoules):*** Total Energy Use in Agriculture per year. *Source*: FAO (Aggregate, may include official, semi-official, estimated, or calculated data)
* ***Cereals Import Index:*** Trade Indices, Import Quantity Index (2014-2016 = 100). *Source*: FAO
* ***Cereals Export Index:*** Trade Indices, Export Quantity Index (2014-2016 = 100). *Source*: FAO
* ***Total Subsidies On Field Crops:*** All farm subsidies on crops, including compensatory payments/area payments, set-aside premiums, and aid under Art 68 for a given accounting year.*Source*: FADN
* ***Rented U.A.A (ha):*** Utilised agricultural area rented by the holder under a tenancy agreement for at least one year (remuneration in cash or in-kind). It is expressed in hectares. *Source*: FADN
* ***Rent paid (€):*** Rent paid for farmland and buildings and rental charges.
* ***Total Utilised Agricultural Area (ha):*** Total utilised agricultural area of holding. It is expressed in hectares. *Source*: FADN
* ***% Rented Land of UAA:*** Percentage rented land of the total utilised for agricultural area. *Source*: Calculated
* ***Gross Value Added:*** Gross value added and income by A\*10 industry breakdowns [nama\_10\_a10]. *Source*: Eurostat
* ***Compensation of Employees:*** Total compensation paid by employers in agriculture in Millions of euros. *Source*: Eurostat
* ***Wages and Salaries:*** Total wages and salaries in agriculture in Millions of euros. *Source*: Eurostat
* ***Production of Cereals Real Price Index:*** Production of cereals Real Price Index, n-1 = 100. *Source*: Eurostat

*(\*) Extra 2 columns added with a breakdown by gender (female/male)*

Data Dictionary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column** | **Name** | **Source(s)** | **Type** | | **Options** |
|  | **Qualitative / Quantitative** | **Categorical Discr / Contin** |
| country | Country Code ISO2 | ISO 3166-1 alpha-2 | Qualitative | Categorical | (AT) Austria,(BE) Belgium,(BG) Bulgaria,(CY) Cyprus,(CZ) Czechia,(DE) Germany,(DK) Denmark,(EE) Estonia,(EL) Greece,(ES) Spain,(FI) Finland,(FR) France,(HU) Hungary,(IE) Ireland,(IT) Italy,(LT) Lithuania,(LV) Latvia,(NL) Netherlands,(PL) Poland,(PT) Portugal,(RO) Romania,(SE) Sweden,(SI) Slovenia,(SK) Slovakia |
| year | Year | FAO, FADN, Eurostat | Qualitative | Discrete | 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020 |
| crop\_mean\_residues\_kg | Mean Residues Kg | FAO | Quantitative | Continues |  |
| crop\_production\_idx | Crop Production Index | FAO | Quantitative | Continues |  |
| cereals\_produce\_price\_usd\_tonne | Cereals Produce Price $/tonne | FAO | Quantitative | Continues |  |
| employment\_ratio\_rural\_areas\_pct | % Employment Ration Rural Areas | FAO | Quantitative | Continues |  |
| female\_employment\_ratio\_rural\_areas\_pct | % Female Employment Ration Rural Areas | FAO | Quantitative | Continues |  |
| male\_employment\_ratio\_rural\_areas\_pct | % Male Employment Ration Rural Areas | FAO | Quantitative | Continues |  |
| mean\_weekly\_working\_hours | Avg Weekly hours working | FAO | Quantitative | Continues |  |
| female\_mean\_weekly\_working\_hours | Avg Female Weekly hours working | FAO | Quantitative | Continues |  |
| male\_mean\_weekly\_working\_hours | Avg Female Weekly hours working | FAO | Quantitative | Continues |  |
| crop\_land\_use\_1000ha | Crop land used 1000ha | FAO | Quantitative | Continues |  |
| agri\_energy\_use\_tj | Energy Used in agriculture (tj) | FAO | Quantitative | Continues |  |
| avg\_import\_idx | Average Import Index Indicator | FAO | Quantitative | Continues |  |
| avg\_export\_idx | Average Export Index Indicator | FAO | Quantitative | Continues |  |
| total\_subsides\_on\_field\_crops | Total subsidies on crops (€) | FADN | Quantitative | Continues |  |
| rented\_land\_ha | Rented UAA (ha) | FADN | Quantitative | Continues |  |
| rent\_paid | Rent paid (€) | FADN | Quantitative | Continues |  |
| total\_uaa\_ha | Total Utilised Agricultural Area (ha) | FADN | Quantitative | Continues |  |
| pct\_rented\_land\_of\_uaa | % Rented land of UAA | Calculated | Quantitative | Continues |  |
| gross\_value\_added | Gross Value Added of Agriculture | Eurostat | Quantitative | Continues |  |
| compensation\_of\_employees | Compensation of employees | Eurostat | Quantitative | Continues |  |
| wages\_and\_salaries | Wages and Salaries | Eurostat | Quantitative | Continues |  |
| prod\_cereals\_real\_price | Prodcution Cereals Real Price Index | Eurostat | Quantitative | Continues |  |

Data Preparation

[code 1]

Original combined Datasets produced missing values, and the decision of merging data for the years 2004-2020 was also a limitation on the data available in Eurostat, compared to the historical records available in FAO and FADN sources. While historic data can be found available in different paid places [4], it was noted that the current EU structure or the Europe of the 28th was formed in 2003 with the inclusion of Croatia in the union. Hence this was set as the cut line for the data set.

However the following countries were excluded:

* *"HR" Croatia* didn’t meet the required 10 samples of data.
* *"LU" Luxembourg* is excluded from the analysis as it forms a very small percentage of the Duch's economy, thus most agricultural indicators and surveys from FAO are reported with low or no values.
* *"MT" Malta*, no data reported on fao datasets.

Handling missing data

Given the size of the data set, and that each aggregated value contains a full year of information, it was necessary to review and assess different techniques for treatment of missing at random (MAR) data several studies and methodologies were chosen for this from simple to complex [5], [6], [7]:

1. Adjustment of Cell Methods (bfill / bpad)
2. Use mean value in Series
3. Linear Regression
4. KNN-Cluster

From the 4 above the main reasons that we choose KNNI method to estimate missing data are:

* K nearest neighbor can deal with heterogeneous (i.e. mixed-attributes) data.
* K nearest neighbor is little affected by the missingness mechanism.
* K nearest neighbor can easily treat instances with multiple missing values (the occurrence of multiple missing values are more common in the process of decision-making)

To determine the appropriate technique to be applied for each series of data relevant to the missing value. It must be evaluated and carefully reviewed the conditions required for apply any technique:

Linear Regression:

***Linear Correlation Coefficient*** [8] to determine the linearity of the dependent variable:

where Sx and Sy denote the sample standard deviations of the x-values and y-values, respectively.

*Linearity: It states that the dependent variable Y should be linearly related to independent variable X.*

Nearest-Neighbour Imputation Methodology:

***Cluster tendency:*** It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed

* H0: Data is generated by a Poisson point process.
* H1: Data is NOT generated by a Poisson point process.

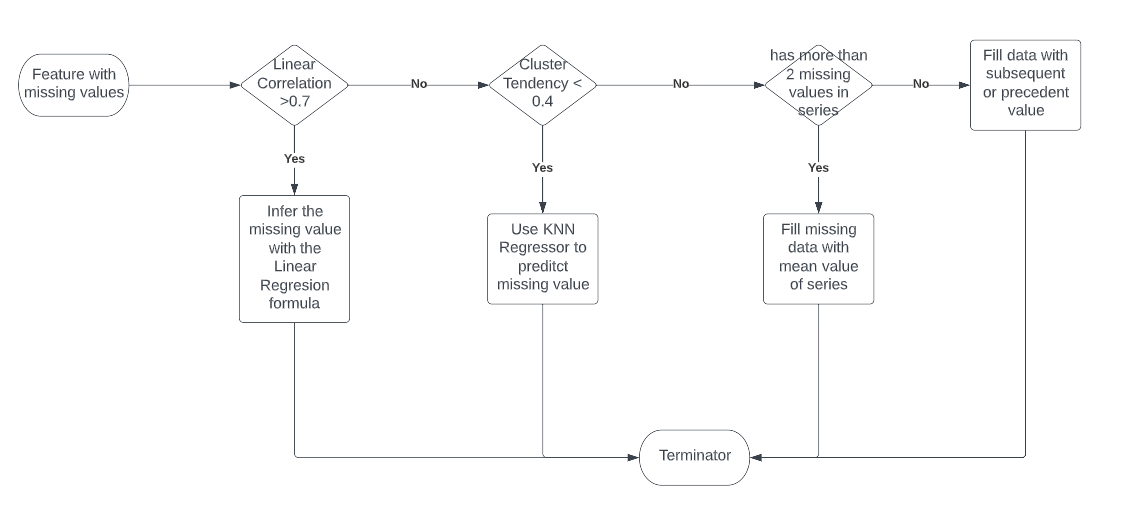
A value close to 0 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 1.

* If the value is between {0.07, ...,0.99}, the data is regularly spaced.
* If the value is around 0.5, it is random.
* If the value is between {0.1, ..., 0.33}, it has a high tendency to cluster.

Adjust of Cell Methods (bfill / bpad):

**maximum of 2 missing** values will be allowed in the full series. The rationale for this is that a higher number of missing values estimated with this method will intrude a big bias on the final series given each series consist of 17 results.

Impute missing values according to the following method.



*[Fig 1] Impute missing data flow*

383 missing values were generated making 4% of all the data points estimated values.

|  |  |
| --- | --- |
| ***impute\_estimator*** | ***total*** |
| bfill-pad | 113 |
| kNN | 43 |
| linear | 196 |
| mean | 31 |
|  | **383** |



1. *[Fig 2] impute values for Cereals Produce Price*

See full details [Table 1] impute values.

Data Visualization

[code 1]

To visualize and get a better understanding of the dataset, bar, line and a choropleth map has been put together into a dashboard to allow the interaction with the data, with filters to select different countries and features to render.

The open-source project Plotly Express [10] in combination with Dash [11] and plugin of the popular CSS library Boostrap allows rendering responsive sites. In terms of the color palette selected qualitative palette G10 has been choosing to cover the 8 final country candidates for study (*refer to inferential analytics section*)



The dashboard is available in the code, aswell by running the server with the distribution version of the dashboard, see app.exe attached to this submission (run.bat shortcut). [Fig 3], [Fig4], [Fig 5].

Descriptive and inferential Analysis

[code 2]

2 mission-critical decisions for the analysis such as “Feature Selection” and “Similar countries” definitions, were studied in this phase to provide an empirical and solid support for the decisions made.

In both areas, the objective is to identify which countries have similar means, variance, and distribution to Ireland, as well as which features can be compared.

Similar Countries:

Ireland has unique characteristics [Table 2] and [Table 3] of the descriptive analysis of the Irish variables. To determine which countries can be included in the analysis 2 selection criteria were defined:

1. Criteria 1: Countries with similar GVA and Total Used Agricultural Area (within percentile 25-75).
2. Criteria 2: Countries where Inferential analysis indicates means and variance of the variables are similar.

The first criteria identified 5 countries for inclusion, *Belgium, Denmark, Latvia, Lithuania, and Portugal*.

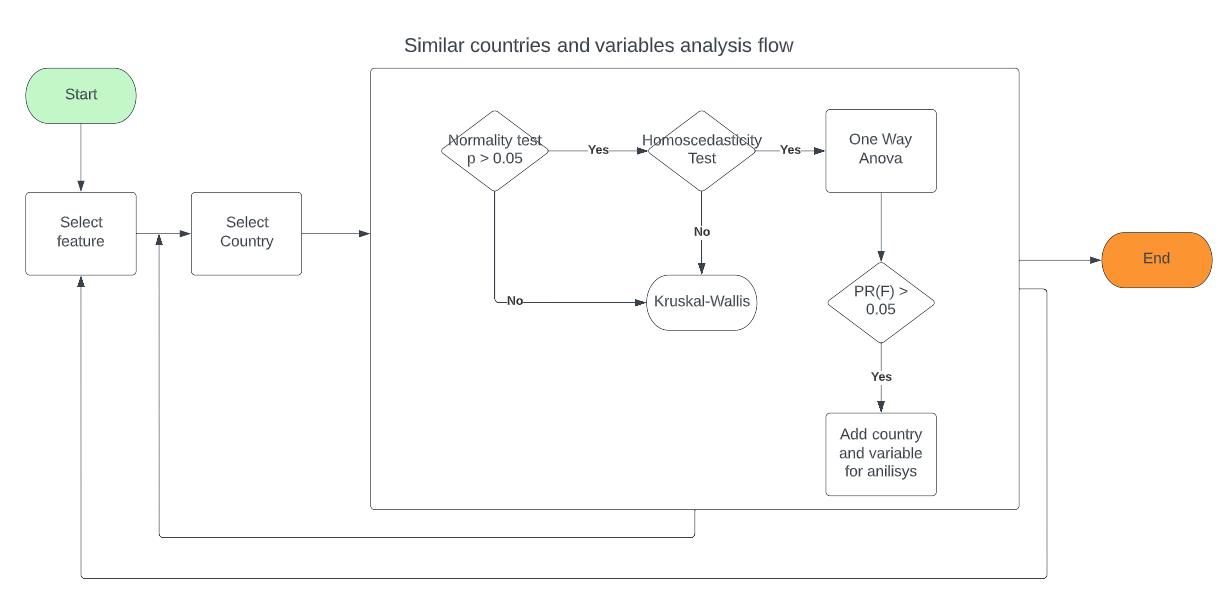
For the second criteria, one-way ANOVA and post hoc analysis within the group for multicomparison of individual countries were used, and pre-conditions were satisfied [12].

* The scale of measurement applied to the data collected follows a continuous or ordinal scale
* Data is collected from a representative, randomly selected portion of the total population.
* Data is normally distributed.
* Reasonable sample size is used.
* Homogeneity of variance.

If the conditions were not met, non-parametric test Kruskal–Wallis one-way analysis of variance.

The comparison runs also in 2 directions. Country and Ireland comparison (no specified variable), and variables (no specific country) with Ireland comparison.

[code 5] and [code 6] function runs the analysis and produces a detailed report variance comparison for ANOVA and Kruskal-Wallis respectively [Anexo 1], [Anexo 2] to identify countries and variables where means and variance are similar to Ireland.



*[Fig 6]: Inferential analysis similar countries flow*

Shapiro Wilk test is utilised for Normality test as well as histogram and probability plot for visualizations e.g. [Fig 7 – Fig 23]. Further post hoc analysis was completed [Fig 24 – Fig 26]. With results summarised below:

Variables for selection

=======================

avg\_export\_idx

avg\_import\_idx

cereals\_produce\_price\_usd\_tonne

compensation\_of\_employees

crop\_production\_idx

employment\_ratio\_rural\_areas\_pct

female\_mean\_weekly\_working\_hours

pct\_rented\_land\_of\_uaa

prod\_cereals\_real\_price

rent\_paid

total\_subsides\_on\_field\_crops

wages\_and\_salaries  
  
Countries for study

=======================

(PT) Portugal

(LT) Lithuania

(DK) Denmark

(LV) Latvia

(BE) Belgium

(BG) Bulgaria

(SK) Slovakia

(IE) Ireland

*On a side note, we can clearly see the effect of the CAP and how cereals value at real price results are the same across many countries [Table 4] and [Fig 26]*

Machine Learning

[code 3]

The aim on this exercise is to find the right selection of features and the adequate ML algorithms for the problem of predicting the Gross Value Added of Ireland in conjunction with other Member State which has shown similarity in different variables. *Countries selection as per Inferential analysis described above.*

Feature Selection and reduction of dimensionality  
  
The following are candidates for the study

* Same mean: Variables identified
* Correlation with GVA > 0.5 [Table 5]
* [13] Principal Component Analysis [Table 6]

ML Model selection

A list of the desired characteristic of the ML models was assessed [Table 7] to come up with the following list of models.

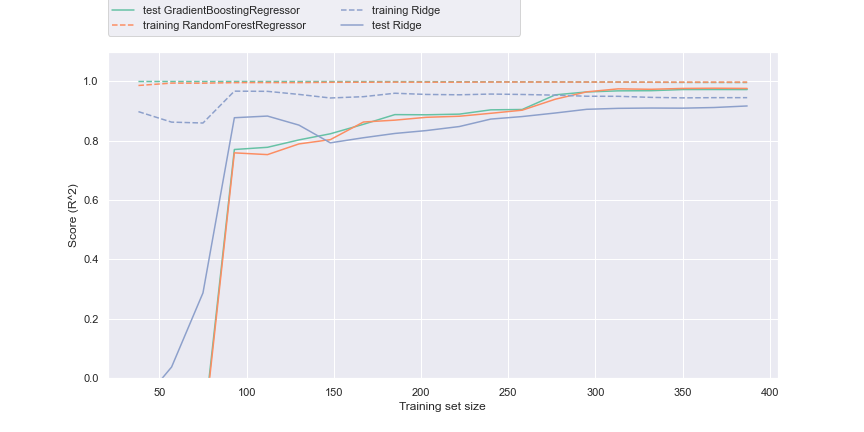
* Linear Regression
* Ridge
* Elastic Net
* RF Random Forest
* XGBoost

All models were executed first with the default parameters. And a second pass after GridSearchCV method was implemented to determine the ideal selection of parameters.

This exhaustive process takes some time to run, especially for the Gradient Boosting Regressor model, which is equivalent to the XGBoost but the version implemented for Sklearn (scikit-learn). The benefit of using it instead of XGBoost, even though the performance of the last is better, was that all previous models were avaiable within Sklearn, making the coding of the algorithms simplified.

The final result can be seen in the table below for the 10 ten best parameters/models/parameter selections:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **model** | **Hyper parameters** | **Feature selection** | **r2 train** | **r2 test** | **r2 ireland** |
| Gradient Boosting Regressor | default | Same Means | 0.997512 | 0.977237 | 0.790796 |
| Random Forest Regressor | {n\_estimators: 30,criterion: squared\_error,max... | Same Means | 0.997759 | 0.98231 | 0.787964 |
| Random Forest Regressor | default | Same Means | 0.997881 | 0.981224 | 0.783877 |
| Gradient Boosting Regressor | {n\_estimators: 30,criterion: squared\_error,max... | Same Means | 0.991945 | 0.973228 | 0.779155 |
| Random Forest Regressor | default | PCA | 0.997872 | 0.980417 | 0.77486 |
| Gradient Boosting Regressor | default | PCA | 0.997964 | 0.980141 | 0.729851 |
| Gradient Boosting Regressor | default | Correlated Variables | 0.997479 | 0.986018 | 0.687192 |
| Gradient Boosting Regressor | default | All | 0.999002 | 0.987238 | 0.666689 |
| Random Forest Regressor | default | All | 0.997972 | 0.986344 | 0.645591 |
| Ridge | default | Same Means | 0.946802 | 0.9387 | 0.638986 |



*[Fig 28] Training Learning best 2 models configuration.*

Conclusions

“Same mean” feature selections played an important role in the analysis, having 4 of the top best performance models at the top of the ranking. While models are overfitted for the test set, a secondary prediction for the full Ireland dataset was also passed achieving a significant good value of up to 79% on the R2 score, for the Gradient Boosting Regressor. The worst performance model was ElasticNet, however, the importance of selecting the best parameters can be seen pushing this algorithm to the middle of the ranking table for the selection of alpha, l1 ration, and random selection parameters.

Sentiment Analysis

[code 4]

Sentiment analysis was a challenging topic to address especially when it comes to finding data available on the topic for producers and consumers. Twitter was found to be quite a production approach, but it comes with a limitation on the developer portal to a maximum of 50 historic queries per month on the license obtained for students. Alternatives data sources were reviewed and finally, web scrappage method was used to pull out data from a different website. Scrapy open source project was used and 2 spiders were configured to retrieve data from 2 sources using CSS selectors:

1. Irish Farmers Associations: Market Updates
2. Boards.ie: CAP discussions

The 2 datasets obtained are disparities in their forms, one coming from official sources and to just publish market updates, and the other with personal options from consumers and producers with regards to the new reforms introduced during the years of this study.

These 2 disparages dataset allow us to set the basis for determining what is a “neutral” text programmatically and utilizing the pre-trained ML algorithm provided with TextBlob.

In both cases, the following steps were done.

* Remove Special Characters.
* All text in lower case
* Applied Stop words.
* Removed top 10 most frequent words
* Removed 10 less frequent words
* Emojis conversion to name (only applicable to boards.ie dataset)
* Calculate sentiment using TextBlob

For the boards.ie dataset and using as reference neutral results percentiles 25 to 75. The following thresholds obtained from IFA were applied:

positive: ( 0.1 , 1 ]

neutral: [-0.06 , 0.1 ]

negative: [-1 ,-0.06)

NLP

Count Vectorized and TF-IFD were executed on the boards.ie dataset for a max number of features of 3000.

The accuracy of both methods comes close to 52% - 57% for the F1 score, resulting in Count-Vectorized model performing better than TF-IFD.

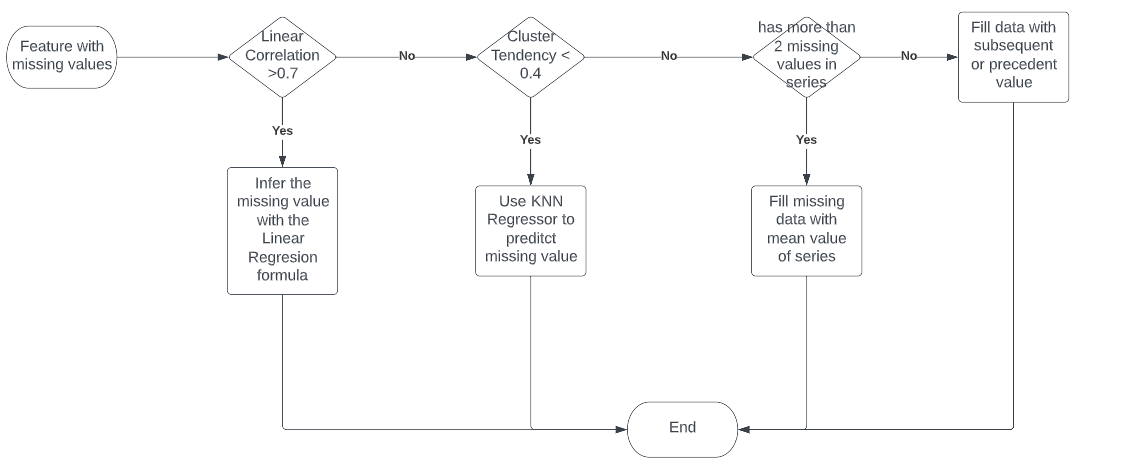
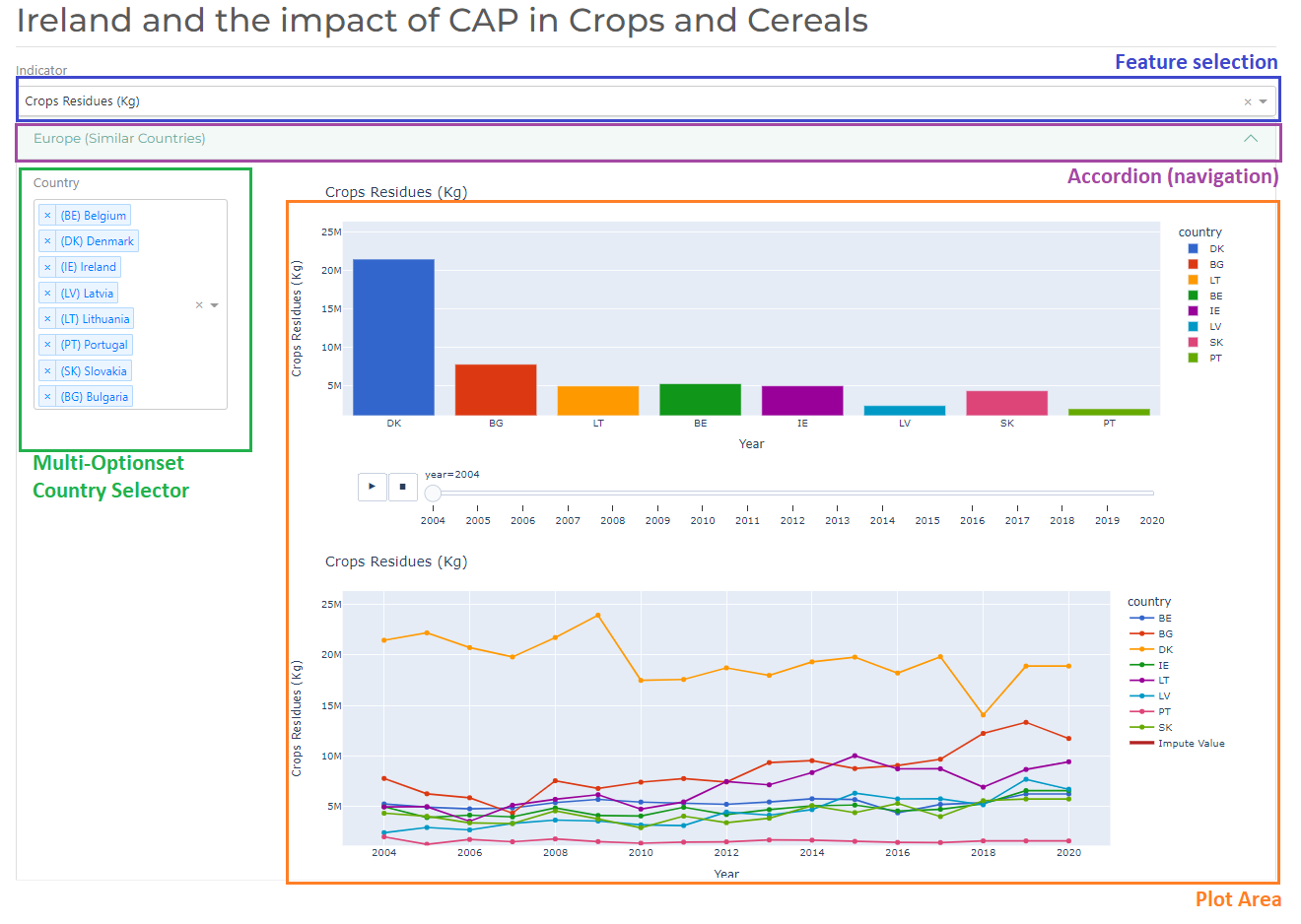
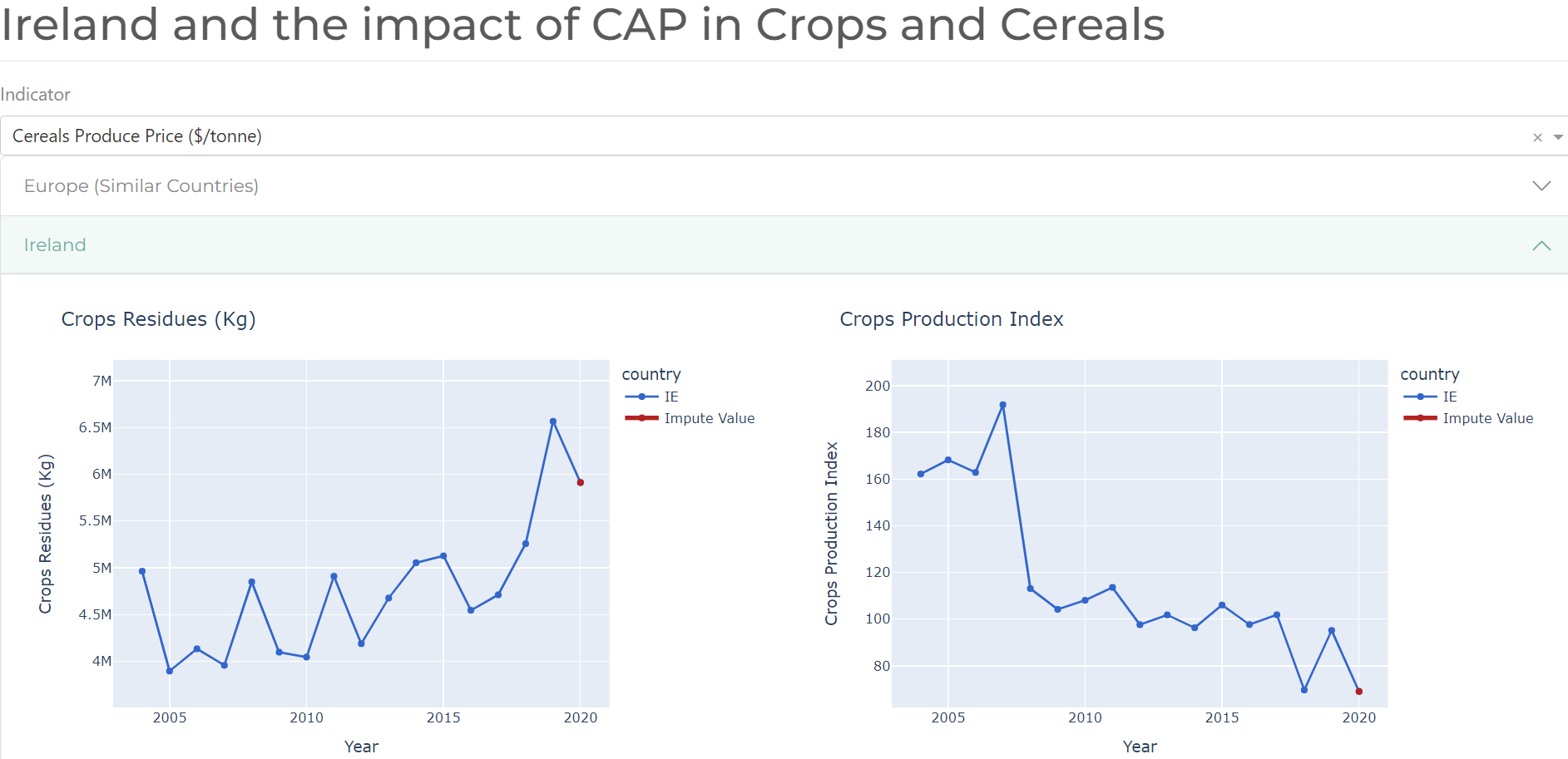
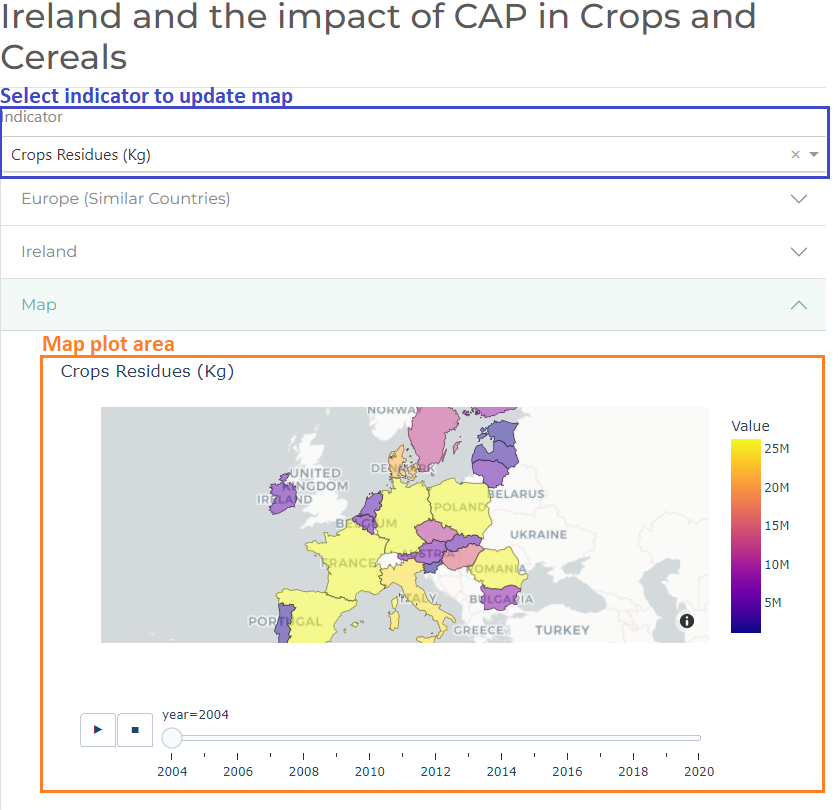
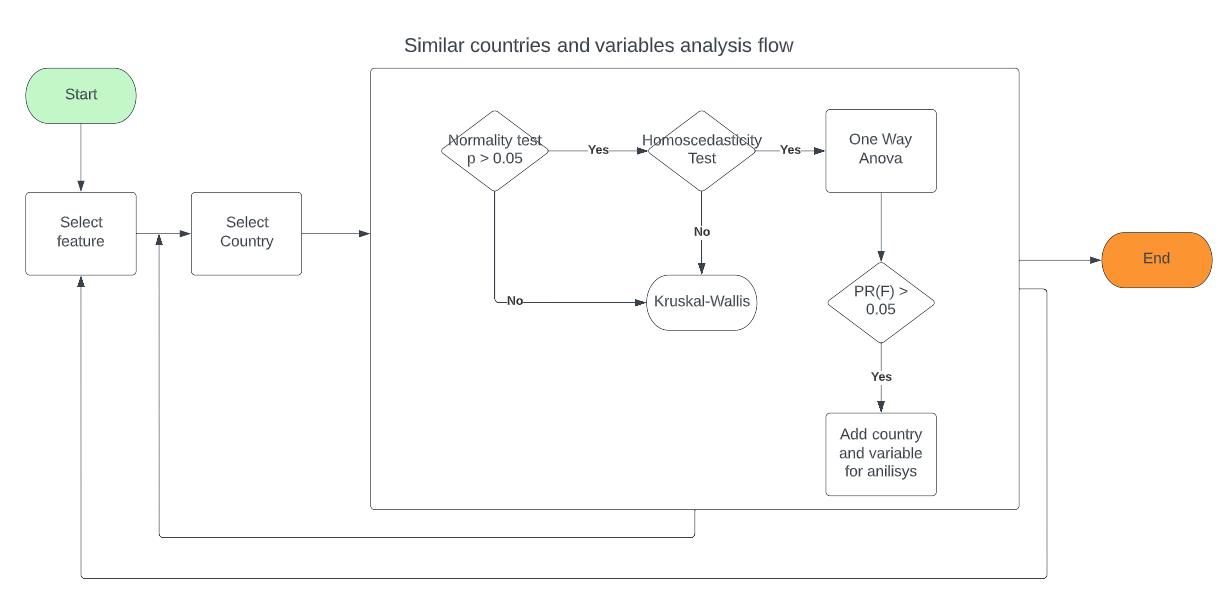
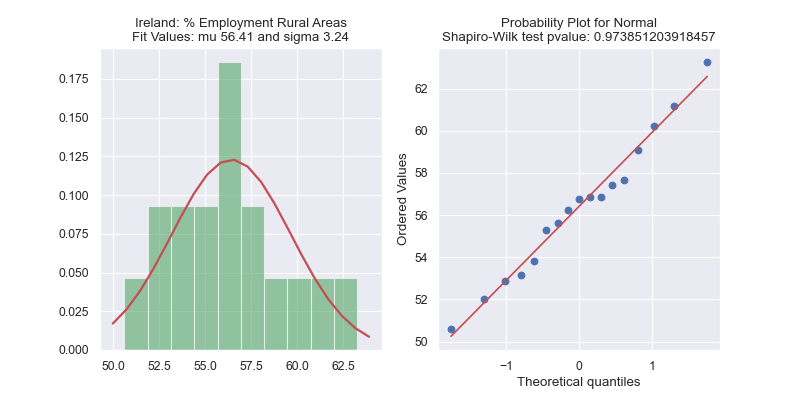
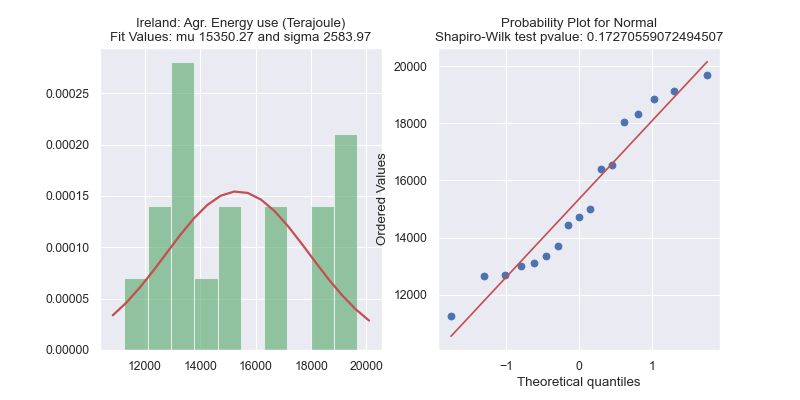
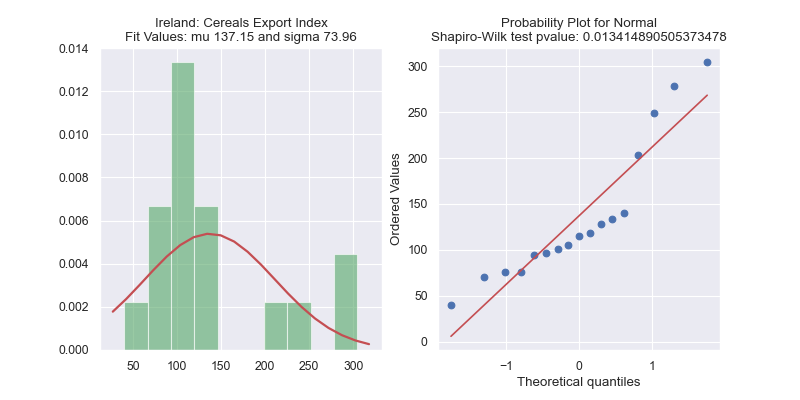
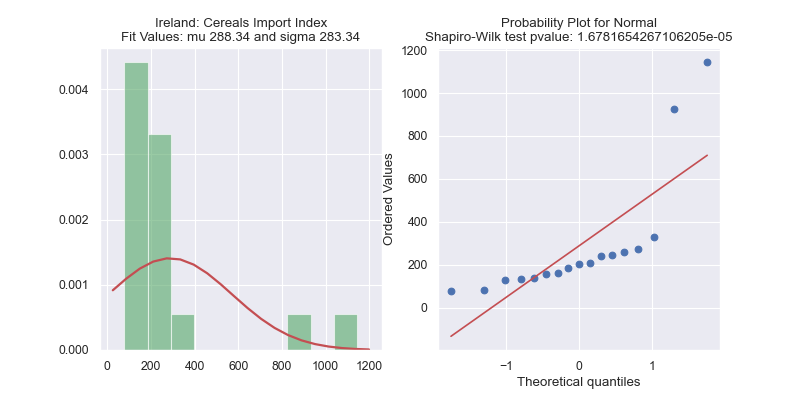
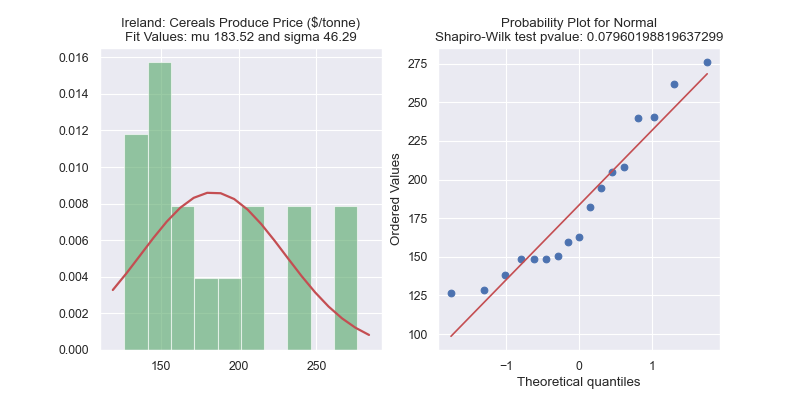
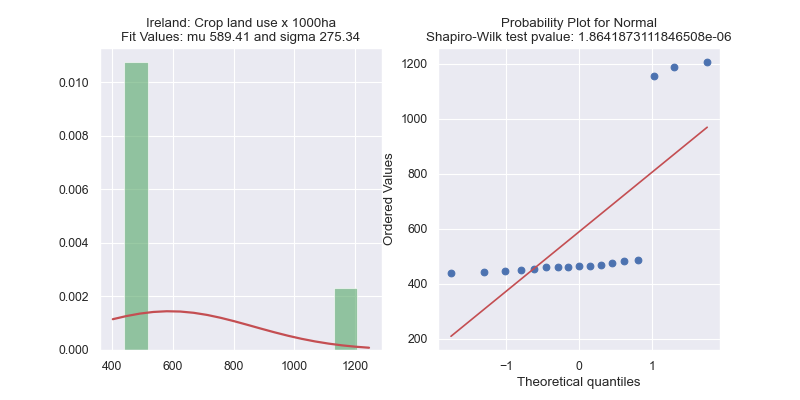
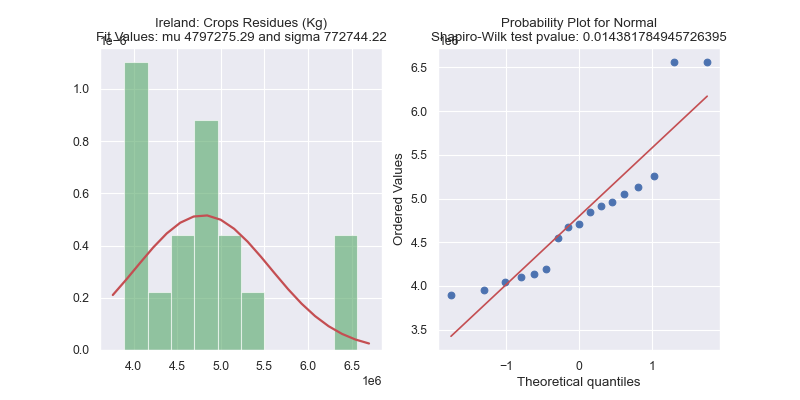
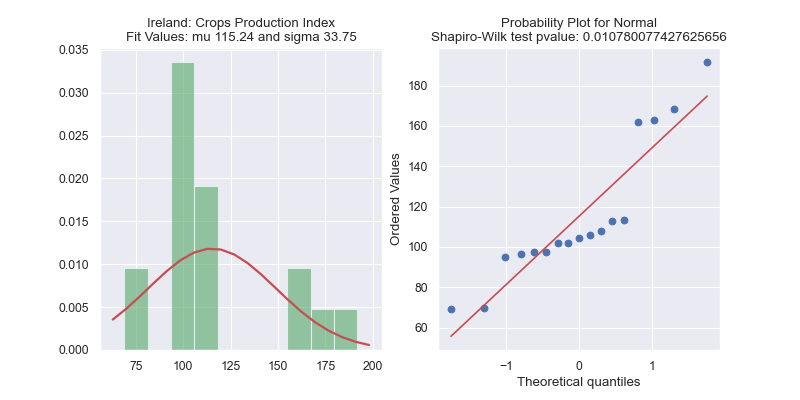
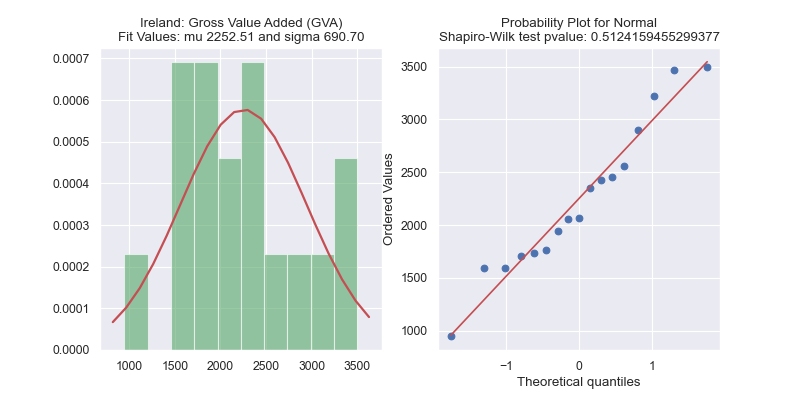
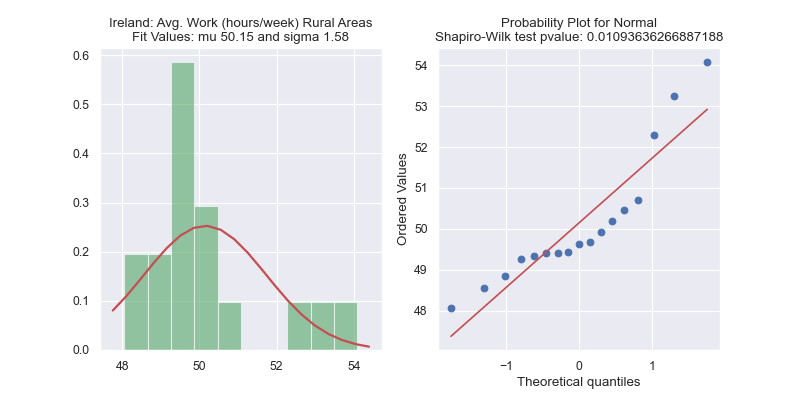
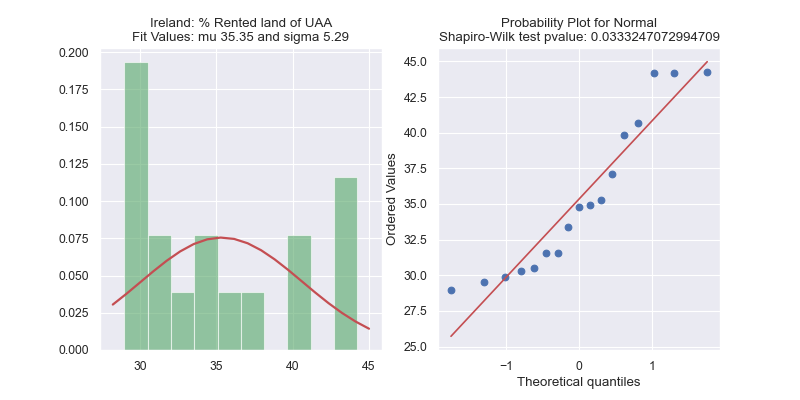
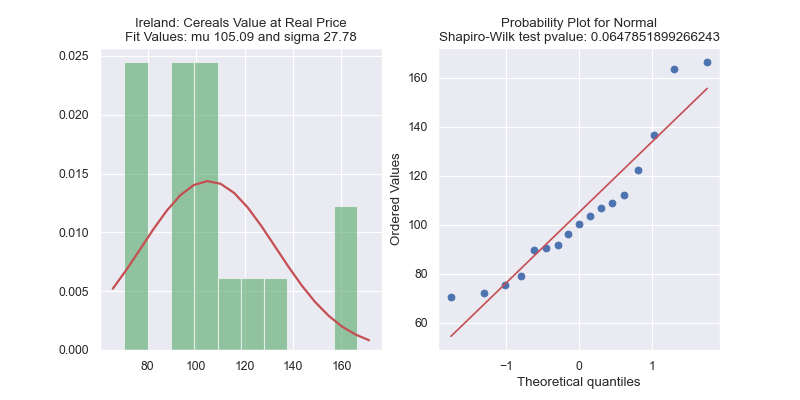
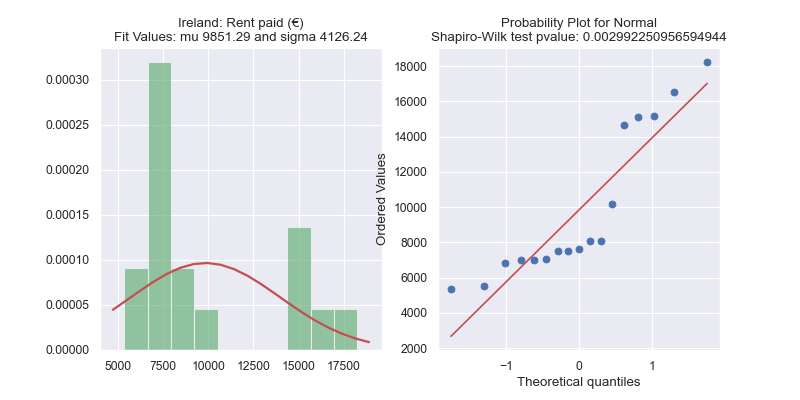
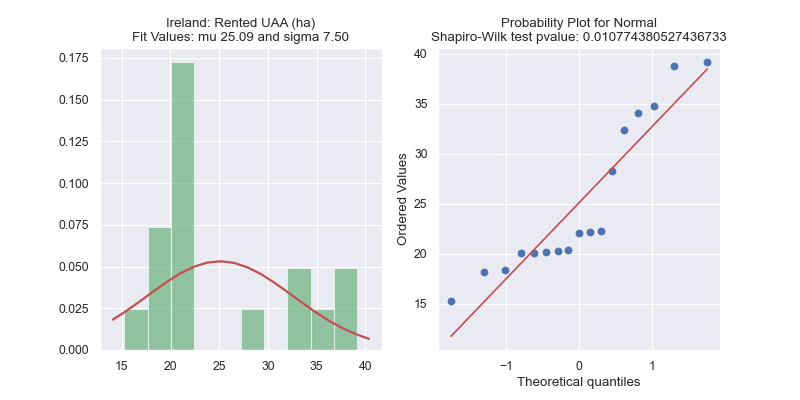
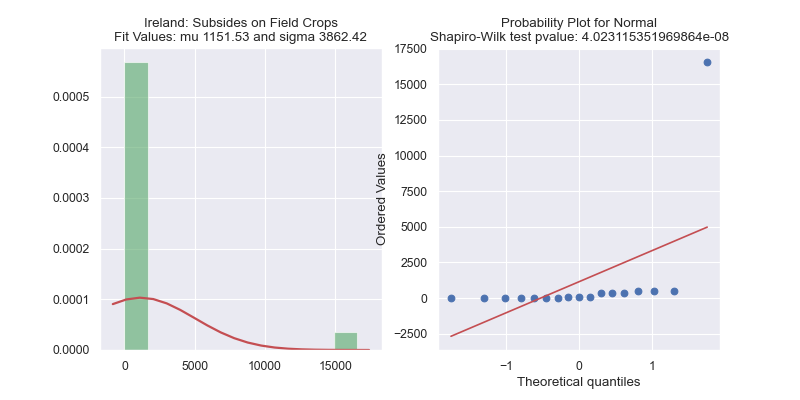
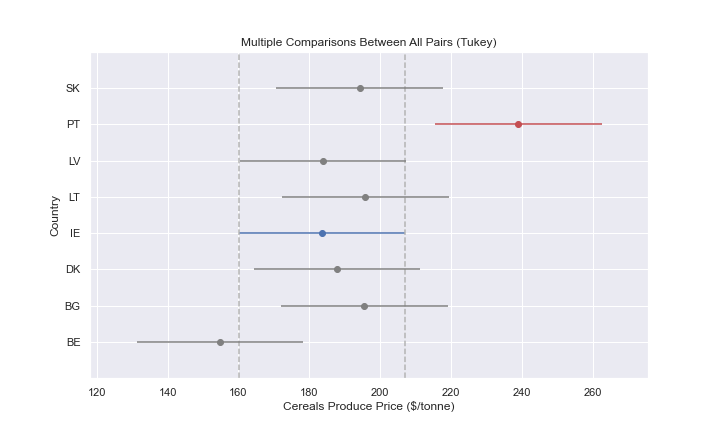
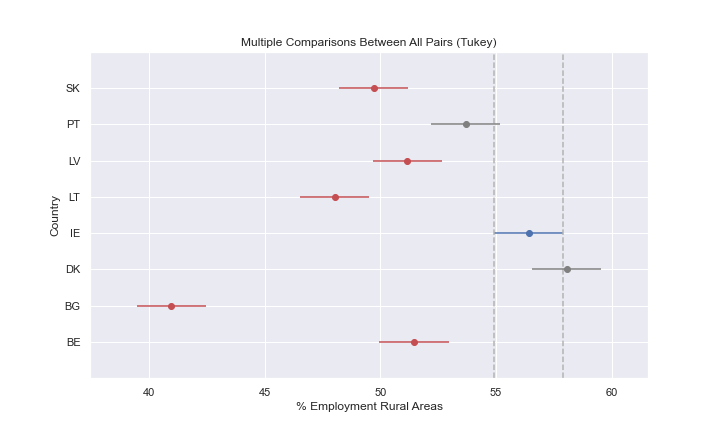
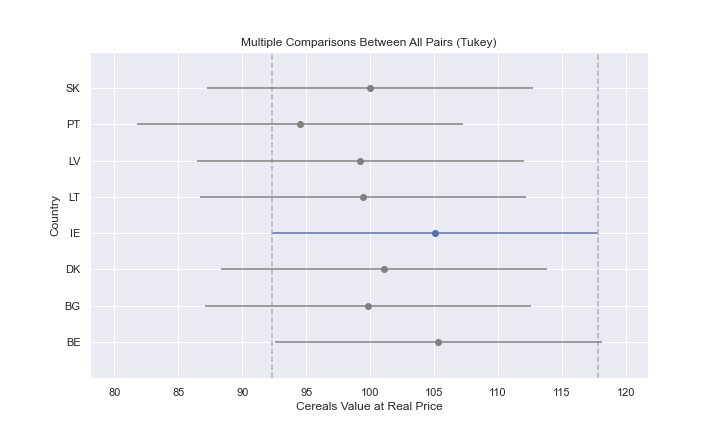
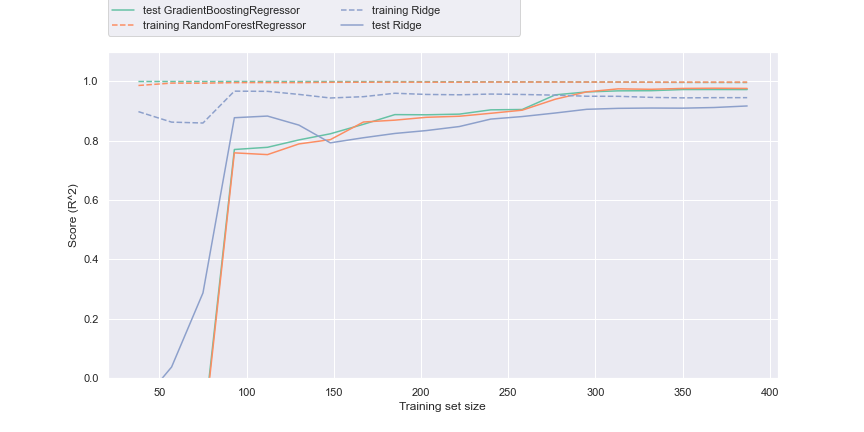
References

1. European Commission, CAP at a glance Available at: <https://ec.europa.eu/info/food-farming-fisheries/key-policies/common-agricultural-policy/cap-glance_en>
2. empty
3. empty
4. Statista Ltd. Agriculture in Ireland, Available at: <https://www.statista.com/statistics/377005/employment-by-economic-sector-in-ireland/>
5. ‘A Review of Hot Deck Imputation for Survey Non-response’ (2011). Pag 3-4 Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=edsoai&AN=edsoai.ocn894435509&site=eds-live> (Accessed: 14 May 2022).
6. Mohammad H. Nadimi-Shahraki et al. (2021) ‘A Hybrid Imputation Method for Multi-Pattern Missing Data: A Case Study on Type II Diabetes Diagnosis’, Electronics, 10(3167), p. 3167. doi: 10.3390/electronics10243167.
7. Ma, X. and Zhong, Q. (2016) ‘Missing value imputation method for disaster decision-making using K nearest neighbor’, Journal of Applied Statistics, 43(4), pp. 767–781. doi: 10.1080/02664763.2015.1077377.
8. Weiss, N. and Weiss, C., 2017. Introductory statistics. 10th ed. Pearson Education Limited 2017, p. 667
9. J. G. Skellam and Brian Hopkins (2017) ‘A New Method for determining the Type of Distribution of Plant Individuals’, Annals of Botany, 18, pp. 213–227. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=edsair&AN=edsair.doi...........6ef6ec0c85f3b83af04881659901aa52&site=eds-live> (Accessed: 14 May 2022).
10. Plotly: Open-source JavaScript charting library behind Plotly and Dash. Open Source. <https://github.com/plotly/plotly.js>
11. Dash: Analytical Web Apps for Python, R, Julia, and Jupyter. Open Source. <https://github.com/plotly/dash>
12. Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proc. R. Soc. Lond. A, 160(901), 268-282.
13. Dessureault, J.-S. and Massicotte, D. (2021) ‘Feature selection or extraction decision process for clustering using PCA and FRSD’. Available at: https://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.2111.10492&site=eds-live.

Code and Jupiter Notebook References

1. [code 1] Datapre and Dashboard: 00\_agriculture\_datasets.ipynb
2. [code 2] Statistical Analysis: 01\_agriculture\_stats.ipynb
3. [code 3] ML: 02\_agriculture\_machine\_learning.ipynb
4. [code 4] Sentiment Analysis: 2\_agriculture\_sentiment\_analysis.ipynb
5. [code 5] function anova\_paircomparison\_ireland\_oms used in 01\_agriculture\_stats.ipynb available in stats.py line 101
6. [code 6] function kruskal\_report used in 01\_agriculture\_stats.ipyn*b* available in *stats.py* line 181

Table of figures

1. [Fig 1]: Impute missing values  
   
2. [Fig 2] Impute values for Cereals Product Price  
   
3. [Fig 3] Dashboard – Similar countries view  
     
   
4. [Fig 3] Dashboard – Ireland  
     
   
5. [Fig 5] Dashboard – Europe Map  
      
    
6. [Fig 6] Inferential analysis. Similar countries and variables method  
   
7. [Fig 7] % Employment Rural Area:   
   
8. [Fig 8] Avg Energy Use.  
   
9. [Fig 9] Cereals Export Index .  
   
10. [Fig 10] Cereals Import Index.  
    
11. [Fig 11] Cereals Produce Price .  
    
12. [Fig 12] Salary Gross (pre-tax) Total compensation of employees.  
    
13. [Fig 13] Cropland used.  
    
14. [Fig 14] Crop Residues (Kg)  
    
15. [Fig 15] Crops Production Index  
    
16. [Fig 16] Gross Value Added  
      
    
17. [Fig 17] Avg. Working hours per week.  
    
18. [Fig 18] Rented Land of UAA.  
    
19. [Fig 19] Cereals Value at Real Price Index.  
    
20. [Fig 20] Rent Paid (€)  
    
21. [Fig 21] Rent UAA (ha)  
      
    
22. [Fig 21] Subsides on Field Crops  
    
23. [Fig 23] MC Mean / Std by Cereals Produce Price  
    
24. [Fig 24] MC Mean / Std by Cereals Produce Price
25. [Fig 25] MC Mean / Std by % Employment Rural Areas
26. [Fig 26] MC Mean / Std by % Cereals Value at Real Price  
    
27. [Fig 27] Best 2 models comparison Learning curve.  
    

Tables

1. [Table 1] Impute value estimators

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Country** | **Year** | **Variable** | **Original Value** | **Impute**  **estimator** | **Impute estimator notes** | **Impute Value** |
| IE | 2020 | agri\_energy\_use\_tj | NaN | linear | Person Correlation Coef: -0.9619341210320422 | 11265.79 |
| IE | 2014 | cereals\_produce\_price\_usd\_tonne | NaN | bfill-pad | Max 2 missing value in series | 148.50 |
| IE | 2015 | cereals\_produce\_price\_usd\_tonne | NaN | bfill-pad | Max 2 missing value in series | 148.50 |
| IE | 2013 | crop\_land\_use\_1000ha | NaN | kNN | hopkins: 0.07665447286475542 | 474.90 |
| IE | 2020 | crop\_land\_use\_1000ha | NaN | kNN | hopkins: 0.07665447286475542 | 448.00 |
| IE | 2020 | crop\_mean\_residues\_kg | NaN | bfill-pad | Max 2 missing value in series | 6566029.00 |
| IE | 2020 | crop\_production\_idx | NaN | linear | Person Correlation Coef: -0.8147547792553999 | 69.02 |
| IE | 2010 | employment\_ratio\_rural\_areas\_pct | NaN | kNN | hopkins: 0.32518482594005305 | 53.84 |
| IE | 2010 | female\_employment\_ratio\_rural\_areas\_pct | NaN | bfill-pad | Max 2 missing value in series | 47.21 |
| IE | 2010 | male\_employment\_ratio\_rural\_areas\_pct | NaN | bfill-pad | Max 2 missing value in series | 56.93 |
| IE | 2020 | pct\_rented\_land\_of\_uaa | NaN | bfill-pad | Max 2 missing value in series | 44.22 |
| IE | 2020 | rent\_paid | NaN | linear | Person Correlation Coef: 0.7570188058765643 | 15143.95 |
| IE | 2020 | rented\_land\_ha | NaN | linear | Person Correlation Coef: 0.7569522265317488 | 34.72 |
| IE | 2020 | total\_subsides\_on\_field\_crops | NaN | kNN | hopkins: 0.27036966710832966 | 374.00 |
| IE | 2020 | total\_uaa\_ha | NaN | linear | Person Correlation Coef: 0.871264143000627 | 86.55 |

1. [Table 2] Descriptive analysis 1 of 2

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **crop\_mean\_residues\_kg** | **crop\_production\_idx** | **cereals\_produce\_price\_usd\_tonne** | **employment\_ratio\_rural\_areas\_pct** | **female\_employment\_ratio\_rural\_areas\_pct** | **male\_employment\_ratio\_rural\_areas\_pct** | **mean\_weekly\_working\_hours** | **female\_mean\_weekly\_working\_hours** | **male\_mean\_weekly\_working\_hours** | **crop\_land\_use\_1000ha** | **agri\_energy\_use\_tj** |
| **count** | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 |
| **mean** | 4797275.00 | 115.24 | 183.52 | 56.41 | 48.94 | 63.79 | 50.15 | 34.87 | 52.39 | 589.42 | 15350.27 |
| **std** | 796526.50 | 34.78 | 47.71 | 3.34 | 2.26 | 5.03 | 1.63 | 1.99 | 1.36 | 283.82 | 2663.49 |
| **min** | 3896111.00 | 69.02 | 126.67 | 50.60 | 45.21 | 55.97 | 48.06 | 31.20 | 50.57 | 440.84 | 11265.79 |
| **25%** | 4134453.00 | 97.61 | 148.50 | 53.84 | 47.21 | 59.94 | 49.33 | 33.86 | 51.68 | 454.00 | 13123.16 |
| **50%** | 4711826.00 | 104.18 | 163.00 | 56.76 | 49.33 | 63.28 | 49.63 | 34.88 | 52.00 | 465.59 | 14731.88 |
| **75%** | 5054516.00 | 113.58 | 208.33 | 57.67 | 50.11 | 68.29 | 50.47 | 35.13 | 52.88 | 483.61 | 18051.64 |
| **max** | 6566029.00 | 191.84 | 276.33 | 63.27 | 53.12 | 73.32 | 54.09 | 38.47 | 55.74 | 1207.00 | 19671.02 |

1. [Table 3] Descriptive analysis 1 of 3

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **avg\_import\_idx** | **avg\_export\_idx** | **total\_subsides\_on\_field\_crops** | **rented\_land\_ha** | **rent\_paid** | **total\_uaa\_ha** | **pct\_rented\_land\_of\_uaa** | **gross\_value\_added** | **compensation\_of\_employees** | **wages\_and\_salaries** | **prod\_cereals\_real\_price** |
| **count** | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 | 17.00 |
| **mean** | 288.34 | 137.15 | 1151.53 | 25.09 | 9851.29 | 70.12 | 35.35 | 2252.51 | 638.64 | 524.44 | 105.09 |
| **std** | 292.06 | 76.24 | 3981.29 | 7.73 | 4253.23 | 11.66 | 5.45 | 711.96 | 76.51 | 63.28 | 28.63 |
| **min** | 79.00 | 40.20 | 0.00 | 15.29 | 5367.00 | 52.73 | 29.00 | 946.50 | 529.90 | 432.50 | 70.38 |
| **0.25** | 140.00 | 94.80 | 28.00 | 20.08 | 6990.00 | 60.95 | 30.49 | 1733.70 | 595.40 | 486.60 | 89.78 |
| **0.50** | 204.17 | 114.80 | 76.00 | 22.06 | 7603.00 | 67.87 | 34.78 | 2070.00 | 613.70 | 510.40 | 100.13 |
| **0.75** | 258.67 | 140.00 | 383.00 | 32.36 | 14683.00 | 81.20 | 39.85 | 2558.40 | 685.70 | 565.30 | 112.03 |
| **max** | 1145.67 | 304.80 | 16581.00 | 39.18 | 18255.00 | 88.58 | 44.23 | 3501.90 | 789.50 | 649.50 | 166.61 |

1. [Table 4] Post hoc Anova Multicomparison Cereals at Real Price results

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **group1** | **group2** | **meandiff** | **p-adj** | **lower** | **upper** | **reject** | **variable** | **country\_name** |
| BE | IE | -0.2441 | 1 | -25.7708 | 25.2825 | FALSE | prod\_cereals\_real\_price | (BE) Belgium |
| BG | IE | 5.2688 | 0.9983 | -20.2578 | 30.7955 | FALSE | prod\_cereals\_real\_price | (BG) Bulgaria |
| DK | IE | 4.0294 | 0.9997 | -21.4972 | 29.556 | FALSE | prod\_cereals\_real\_price | (DK) Denmark |
| IE | LT | -5.6288 | 0.9974 | -31.1555 | 19.8978 | FALSE | prod\_cereals\_real\_price | (LT) Lithuania |
| IE | LV | -5.8476 | 0.9967 | -31.3743 | 19.679 | FALSE | prod\_cereals\_real\_price | (LV) Latvia |
| IE | PT | -10.6071 | 0.9045 | -36.1337 | 14.9196 | FALSE | prod\_cereals\_real\_price | (PT) Portugal |
| IE | SK | -5.0918 | 0.9986 | -30.6184 | 20.4349 | FALSE | prod\_cereals\_real\_price | (SK) Slovakia |

1. [Table 5] Variables with Person Correlation > 0.7 with GVA

|  |  |
| --- | --- |
| **variable** | **value** |
| crop\_mean\_residues\_kg | 0.724757 |
| crop\_land\_use\_1000ha | 0.861387 |
| agri\_energy\_use\_tj | 0.806896 |
| compensation\_of\_employees | 0.960475 |
| wages\_and\_salaries | 0.963715 |

1. [Table 6] PCA Feature importance results

|  |  |  |
| --- | --- | --- |
| **PC** | **Variable** | **Importance** |
| PC0 | compensation\_of\_employees | 0.218245 |
| PC1 | total\_uaa\_ha | 0.165166 |
| PC2 | employment\_ratio\_rural\_areas\_pct | 0.154536 |
| PC3 | mean\_weekly\_working\_hours | 0.100192 |
| PC4 | prod\_cereals\_real\_price | 0.063979 |
| PC5 | avg\_import\_idx | 0.058312 |
| PC6 | avg\_export\_idx | 0.051678 |

1. [Table 7] ML models desired characteristics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DS** | **SOD** | **DOD** | **O** | **HO** | **MC** |
| **Desired features** | S/L | S/L | Y/N | Y | Y |
| Linear Regression | S | S | N | ***N*** | ***N*** |
| Ridge | L | L | Y | Y | Y |
| Elastic Net | L | L | Y | Y | Y |
| RF | L | L | Y | Y | ***n/a*** |
| XGBoost | L | L | Y | Y | ***n/a*** |

*Size of Data Set (SOD)*

*Dimension of Data Set (DOD)*

*Outliers (O)*

*Handling of Overfitting (HO)*

*Multicollinearity (MC)*

Anexos

1. Anexo 1: Anova execution log  
   
2. Anexo 2: Kruskal–Wallis execution log  
   