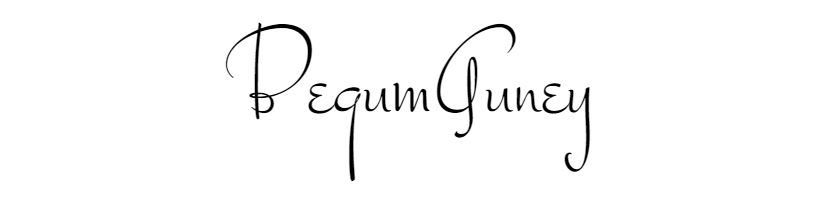
**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title(s):** | Statistics for Data Analytics  Programming for Data Analytics  Data Preparation & Visualization  Machine Learning for Data Analytics |
| **Assessment Title:** | MSC\_DA\_CA2 |
| **Lecturer Name(s):** | John O’Sullivan  Sam Weiss  David McQuaid  Muhammad Iqbal |
| **Student Full Name:** | Begum Guney |
| **Student Number:** | 2022317 |
| **Assessment Due Date:** | 06.01.2023 |
| **Date of Submission:** | 06.01.2023 |

****

**Declaration**

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

# *Abstract*

*Food security is a pressing social and economic problem. One of the most popular ways of combating food insecurity is the production and preservation of more food. However, food production is affected by various factors including geo-political wars, policies, arable land, etc. The current study sought to understand the cereal production situation in Ireland, Portugal, and the rest of Europe. Besides, the study examined the factors that influence the production of cereals in Ireland, using linear regression, decision trees, and random forest. Using a trend test, Ireland was noted to have an increasing cereal production trend while Portugal’s rate of cereal production is decreasing over time. The linear regression model had the best performance (RMSE = 467.662, R2 = 0.9406), compared to the decision tree (RMSE = 836.071, R2 = 0.8103), and random forest (RMSE = 620.517, R2 = 0.8955). Following feature selection, the modified linear regression (RMSE = 467.662, R2 = 0.9406) indicated that cereal production in Ireland is influenced by Rural Population, Crop Production Index, Land under cereal production (hectares), Arable land (% of land area), and Fertilizer consumption (kilograms per hectare of arable land).*

***KEYWORDS:*** *Cereal Production, EDA, Normality Test, Homogeneity of Variance Test, Wilcoxon Test, Kruskal-Walls Test, Trend Test, Clustering, Random Forest Classifier, Linear Regression, Decision Tree, Sentiment Analysis*

***Word Count****: 3236 words (not including code, code comments, titles, references or citations)*

*5414 in total*

# TABLE OF CONTENTS

[*Abstract* ii](#_Toc123881175)

[TABLE OF CONTENTS iii](#_Toc123881176)

[LIST OF FIGURES iv](#_Toc123881177)

[LIST OF TABLES iv](#_Toc123881178)

[1 INTRODUCTION 1](#_Toc123881179)

[1.1 Role of Cereals in Food Security 1](#_Toc123881180)

[1.2 Europe and Cereal Production 2](#_Toc123881181)

[1.3 Policies 2](#_Toc123881182)

[1.4 Objective 2](#_Toc123881183)

[2 PROJECT APPROACH 3](#_Toc123881184)

[2.1 Analytical Framework 3](#_Toc123881185)

[2.1.1 Problem Identification 3](#_Toc123881186)

[2.1.2 Data Wrangling 3](#_Toc123881187)

[2.1.3 Exploratory Data Analytics (EDA) 3](#_Toc123881188)

[2.1.4 Detecting Outliers 4](#_Toc123881189)

[2.1.5 Pre-processing and Training Data Development 5](#_Toc123881190)

[2.1.6 Checking missing observations 6](#_Toc123881191)

[2.1.7 Variable Selection 7](#_Toc123881192)

[2.1.8 Modelling 8](#_Toc123881193)

[2.1.9 Presentation of Findings 8](#_Toc123881194)

[2.2 Project Planning 9](#_Toc123881195)

[3 Statistics for Data Analytics 10](#_Toc123881196)

[3.1 Data Wrangling 10](#_Toc123881197)

[3.1.1 Descriptive Statistics 11](#_Toc123881198)

[3.1.1.1 Ireland 11](#_Toc123881199)

[3.1.1.2 Portugal 12](#_Toc123881200)

[3.1.1.3 The rest of the countries in Europe 12](#_Toc123881201)

[3.1.2 Visualization (Food production index vs. cereal yield) 13](#_Toc123881202)

[3.1.3 Population Estimates 16](#_Toc123881203)

[3.2 Statistical Tests 16](#_Toc123881204)

[3.2.1 Distributional tests 16](#_Toc123881205)

[3.2.1.1 Test for normality 16](#_Toc123881206)

[3.2.1.2 Test for Homogeneity of Variance 17](#_Toc123881207)

[3.2.2 Test for Differences in Cereal Production 17](#_Toc123881208)

[3.2.2.1 Wilcoxon signed-rank test 18](#_Toc123881209)

[3.2.2.2 Kruskal-Wallis test 18](#_Toc123881210)

[3.2.3 Trend Test 19](#_Toc123881211)

[4 Machine Learning for Data Analytics 21](#_Toc123881212)

[4.1 Machine Learning Problem 21](#_Toc123881213)

[4.1.1 Clustering 21](#_Toc123881214)

[4.1.1.1 K-Means Clustering 21](#_Toc123881215)

[4.1.1.2 Average Cereal Production and food production index per Cluster 23](#_Toc123881216)

[4.2 Sentiment Analysis 27](#_Toc123881217)

[5 Discussion and Conclusion 29](#_Toc123881218)

[5.1 Study Limitations and Difficulties 29](#_Toc123881219)

[5.1.1 Availability of data 29](#_Toc123881220)

[5.1.2 Licensing 29](#_Toc123881221)

[5.2 Dashboard Overview 30](#_Toc123881222)

[REFERENCES 32](#_Toc123881223)

[GitHub Reference 34](#_Toc123881224)

# LIST OF FIGURES

[Figure 1: Box plot of cereal production in Ireland, Portugal, and the rest of Europe 4](#_Toc123881225)

[Figure 2: Violin plot of the rural population 4](#_Toc123881226)

[Figure 3: Geo plot of average cereal production in Europe since 2010 5](#_Toc123881227)

[Figure 4: Heat map of missing observations in each of the data frames 6](#_Toc123881228)

[Figure 5: Heatmap of correlation matrix of Ireland 7](#_Toc123881229)

[Figure 6: Project progress 9](#_Toc123881230)

[Figure 7: Relationship between agricultural production indices and cereal production in Ireland 13](#_Toc123881231)

[Figure 8: Relationship between agricultural production indices and cereal production in Portugal 14](#_Toc123881232)

[Figure 9: Relationship between agricultural production indices and cereal production (the rest of Europe) 15](#_Toc123881233)

[Figure 10: Trend test for Ireland 19](#_Toc123881234)

[Figure 11: Trend test for Portugal 19](#_Toc123881235)

[Figure 12: Trend test for the rest of Europe 20](#_Toc123881236)

[Figure 13: K-means clustering 22](#_Toc123881237)

[Figure 14: Silhouette Score per model 22](#_Toc123881238)

[Figure 15: RMSE scores 25](#_Toc123881239)

[Figure 16: R-Squared scores 26](#_Toc123881240)

[Figure 17: Common words associated with cereal products from Ireland 27](#_Toc123881241)

[Figure 18: Distribution of sentiments 28](#_Toc123881242)

[Figure 19: Cluster = 3 30](#_Toc123881243)

[Figure 20: Cluster = 0 31](#_Toc123881244)

# LIST OF TABLES

[Table 1: Project schedule and task progress 9](#_Toc123881245)

[Table 2: Data Characteristics 10](#_Toc123881246)

[Table 3: Indicator information 11](#_Toc123881247)

[Table 4: Agricultural produce indices in Ireland 11](#_Toc123881248)

[Table 5: Agricultural produce indices in Portugal 12](#_Toc123881249)

[Table 6: Agricultural produce indices in rest of the Europe 12](#_Toc123881250)

[Table 7: Population estimates in region or countries 16](#_Toc123881251)

[Table 8: Test for Normality 17](#_Toc123881252)

[Table 9: Test for Homogeneity of Variance 17](#_Toc123881253)

[Table 10: Trend test evaluation 20](#_Toc123881254)

[Table 11: Average Cereal Production and food production index per cluster 23](#_Toc123881255)

[Table 12: Parameter space 24](#_Toc123881256)

[Table 13: Significant coefficients 25](#_Toc123881257)

# INTRODUCTION

As reflected in the United Nations Sustainable Development Goals, the significance of food security and nutrition cannot be overemphasized (Raheem, et al., 2021). Conceptually, food security comes about when there is physical and economic access to sufficient, safe, and nutritious food for everyone at all times with which to fulfil their dietary needs and food desires for an active and healthy life (Hassen & Bilali, 2021). Primarily, food security is parameterized by four standard dimensions including availability (access to a sufficient amount of food); access (possessing adequate resources to obtain the necessary food); utilization (having a reasonable food use for basic nutrition requirements based on learned knowledge); and the stability of the availability, access, and utilization of food (United Nations System High-Level Task Force on Global Food Security, 2011).

To date, food and agriculture systems have been the key drivers of the well-being of humanity. As such, regarding food security, the aforementioned systems need to not only provide safe and healthy food for the well-being of humanity but also act as a means of livelihood and income to actors in the field of food production (Kelly, 2019) e.g., farmers, food processing industries, etcetera. Besides, the systems also play a significant role in rural and economic growth and development of many economies around the globe making them integral components of human advancement (Kelly, 2019).

## Role of Cereals in Food Security

According to Kelly (2019), the fundamental objective of food security is to produce sufficient cereals globally, to meet the rising demand for food, animal feeds, and biofuels. Therefore, the questions one might be interested in range from, what is the state of cereals production? To the role of cereals in ensuring food security.

Pioneering research in improving the key cereals- maize, rice, and wheat has greatly improved global food security for over half a century (Kropff & Morell, 2019). Regardless of the feats achieved, there are over 800 million people who are still faced with chronic hunger and malnutrition, leaving a margin for more work.

## Europe and Cereal Production

Cereals are grown on approximately 50% of the European Union’s (EU) farms, which is a third of the region’s agricultural area and accounts for a quarter of the EU’s crop production value (EU, 2014). Primarily, Europe’s cereal production constitutes 20% of the global cereal production and is a net exporter of 15% of the resultant cereal (Schils, et al., 2018) making cereal farming a viable social and economic prospect.

## Policies

Various regulations that govern agriculture in Europe have been put in place including the EU crop rotation rules which were suspended temporarily in 2018 to facilitate an increase in cereal production and help offset the global food security crisis following the impact of the Russia-Ukraine war (Fortuna & Foote, 2022). The Common Agricultural Policy (CAP) which generally governs agricultural activities in Europe has nine main objectives including enhancing market competitiveness and market orientation (Grodzicki & Jankiewicz, 2022).

## Objective

The current study seeks to examine the state of cereal production in Europe with Ireland as the baseline. In particular, the current study aims to examine differences in cereal production between Ireland and Portugal as well as Ireland and the rest of the European countries.

To address the preceding research objective, the following question will be answered:

1. How does Ireland compare with Portugal and the rest of Europe’s cereal production?
2. What are the factors that influence cereal production in Ireland?
3. How is cereal production changing in Ireland, Portugal, and the rest of Europe?

# PROJECT APPROACH

The project is comprised of three tasks including statistics for data analytics, data preparation, and visualization, as well as machine learning tasks.

## Analytical Framework

To this end, a modified Data Science Method (DSM) framework (Johnson, 2018; Martinez, et al., 2021) was adopted which follows processes such as:

### Problem Identification

The problem identification step involves specifying and establishing the research problem that the project seeks to address. This step allows the formation of the problem and the establishment of a solution scope. For instance, the current study seeks to propose a machine-learning model for predicting the amount of cereal produced in Ireland as well as detecting sentiments regarding Cereals from Ireland.

This step also includes the specification of data sources.

### Data Wrangling

During data wrangling, various processes are undertaken including the collection of data, organization, and cleaning.

### Exploratory Data Analytics (EDA)

Exploratory data analytics allows the building of data profile tables and plots to determine issues that are present in the data. This step also enables the detection of underlying relationships between various attributes as well as the distribution of the variables of interest. With respect to the current study, both the statistics for the data analytics and visualization tasks will be conducted during the Exploratory data analytics step in DSM.

During EDA, a dashboard was built to facilitate the visualization of the relationships and distributions of various attributes.

### Detecting Outliers

Box plots were used to detect outliers in the data. Figures 1 and 2 below show the distribution of cereal production in Ireland relative to Portugal and Europe. No outliers were detected in the region data hence no treatment was applied.

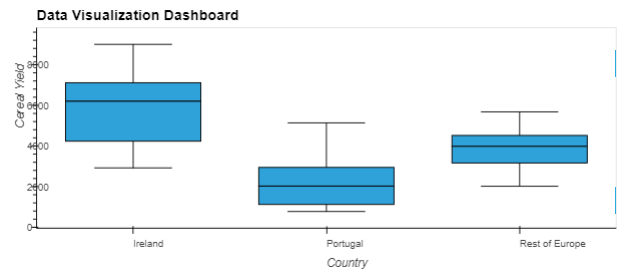


Figure 1: Box plot of cereal production in Ireland, Portugal, and the rest of Europe

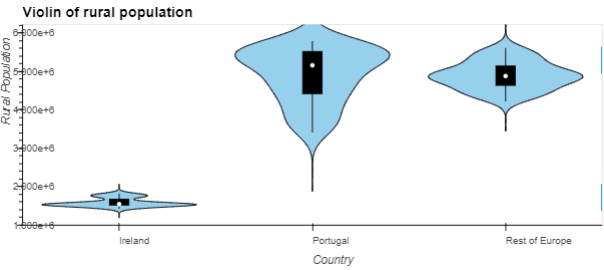


Figure 2: Violin plot of the rural population

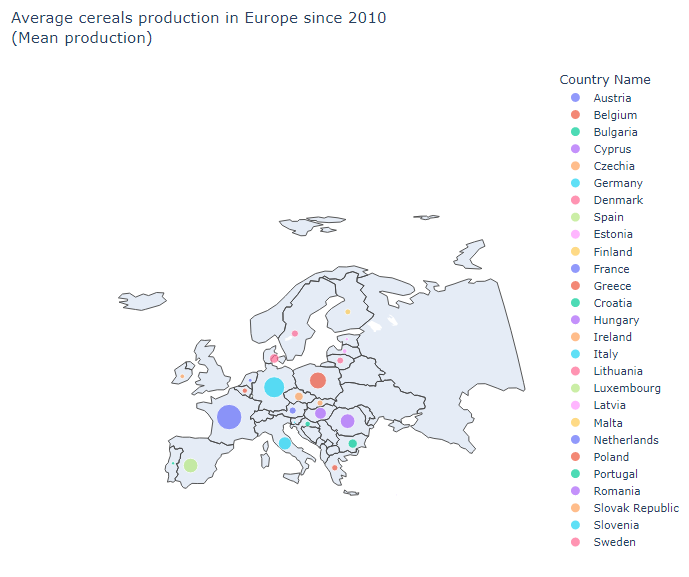


Figure 3: Geo plot of average cereal production in Europe since 2010

Since 2010, it is noted that France, Germany, and Poland have the highest cereal production in Europe respectively (*see figure 4*).

### Pre-processing and Training Data Development

Depending on the preliminary relationships that are learned above and using the resulting data following the previous steps, this step involves the transformation of existing features and the generation of new features, i.e., feature engineering. Moreover, this step involves partitioning the data into train and test sets for use during the training and testing of the machine learning models. Ideally, this step is mainly composed of the *data preparation* activities which address the problems identified during the exploratory data analytics process.

### Checking missing observations

Missing observations were observed in each of the data frames (*see figure 4 and table 3*). To handle the missing observations, each column was imputed using the respective attribute mean.

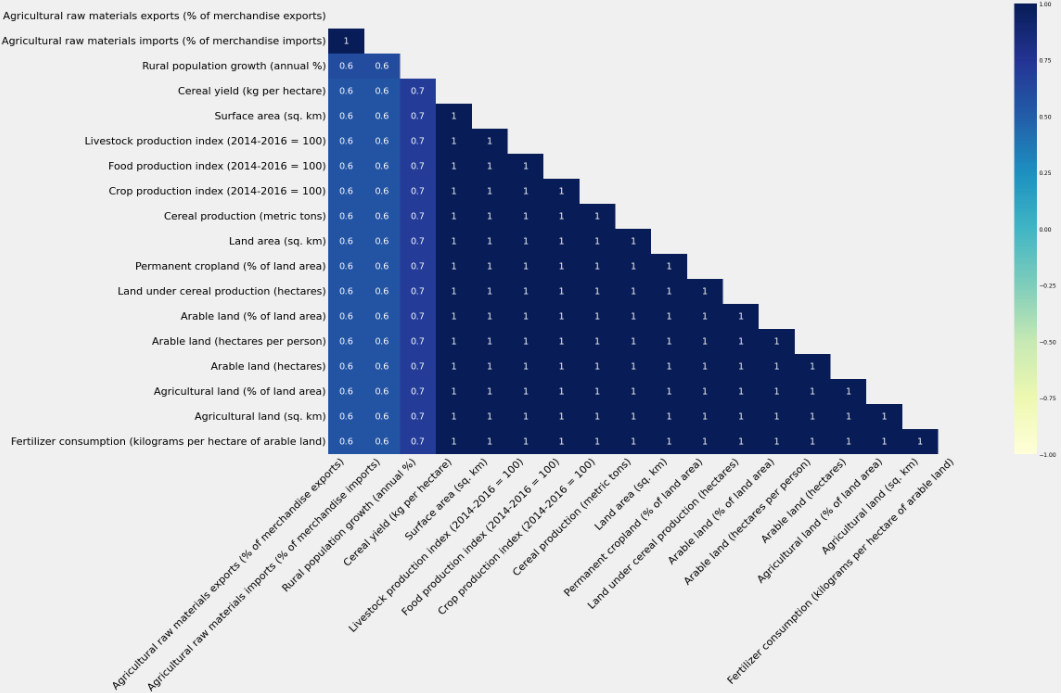


Figure 4: Heat map of missing observations in each of the data frames

### Variable Selection

Models such as linear regression are sensitive to multicollinearity. For each of the predictor attributes i.e., all attributes listed in table 3 except the Cereal production (metric tons) were checked for multicollinearity (*see figure 5*). Each pair that had a correlation score greater than 0.9 were excluded from the data.

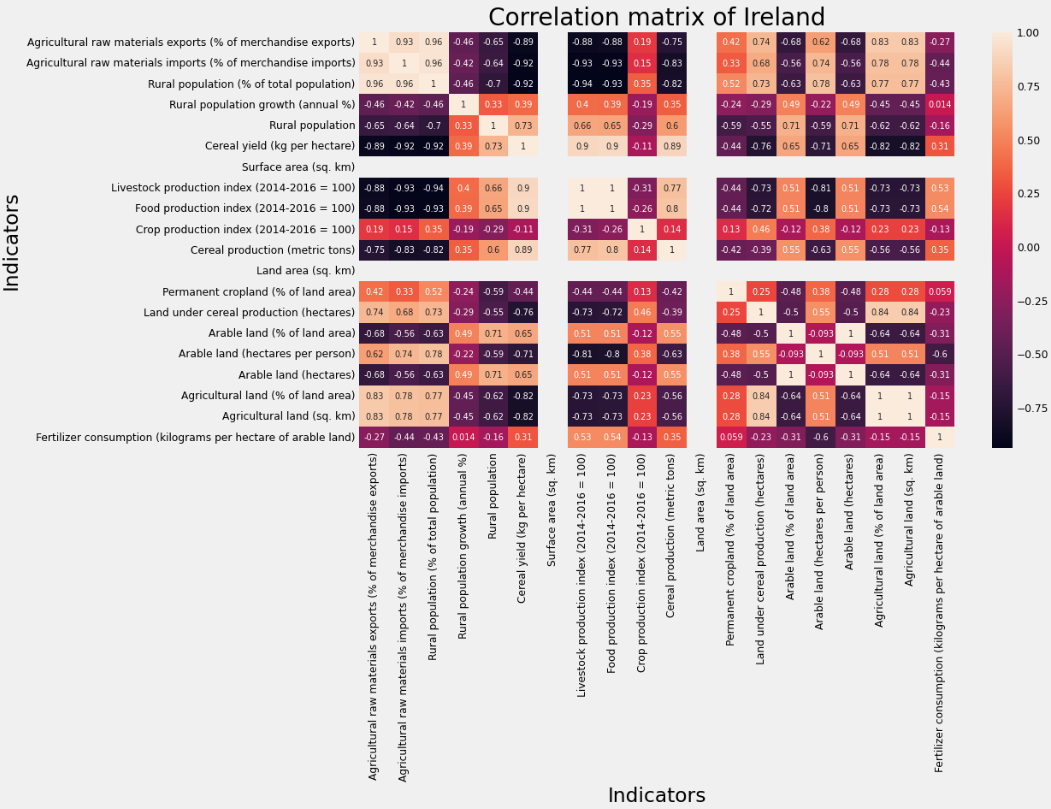
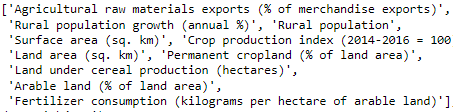


Figure 5: Heatmap of correlation matrix of Ireland

The following list shows the attributes that were retained following the correlation dimensionality reduction.

**

### Modelling

During modelling, various models are proposed and trained depending on the corresponding modelling problem. Subsequently, the performance of each of the predictive models will be reviewed using the amount of variance explained (R-Squared) by each model as well as the error score (root mean squared error, i.e., RMSE).

The model with the best performance is then proposed for addressing the prediction research problem.

### Presentation of Findings

Akin to deployment, this step involves presenting the outcome of the modelling process in an interactive dashboard. That is, this step involves creating an interactive dashboard to show the various aspects of cereal production in Ireland based on the predictive model. This includes visualization of the various clusters and the predicted cereal production.

## Project Planning

The entire project is expected to be conducted between 29th November and 22nd December (*see figure 6*).



Figure 6: Project progress

Table 1 below shows the project timeline for each of the proposed DSM activities.

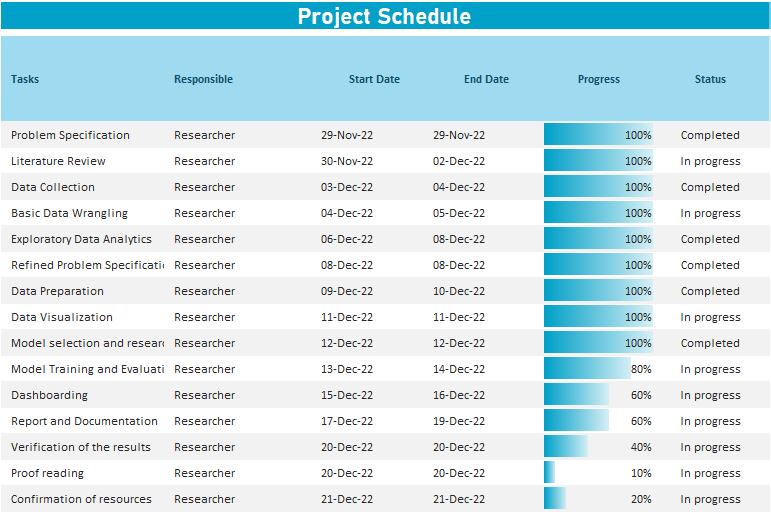


Table 1: Project schedule and task progress

# Statistics for Data Analytics

## Data Wrangling

To address the research objective, data related to agricultural produce in Europe were collected. The data included information on agricultural indicators for the 27 European Union member states. Data was collected from <https://data.worldbank.org/topic/agriculture-and-rural-development>. Before further analysis, the data was filtered to only include information on EU countries and restructured to convert the indicators into column names and the yearly observations as row instances for the respective indicators. Table 1 below shows the characteristics of the raw data.

|  |  |
| --- | --- |
| Data Characteristic | Observation |
| Number of variables (indicators) | 42 |
| Number of observations | 62 |
| Period of observations | 1960 to 2021 |
| Aggregation | Annually |

Table 2: Data Characteristics

To ensure the research data had sufficient information to inform findings regarding the objective, indicators that had more than 20% missing observations were excluded from the data. The subsequent data contained 20 indicators as listed in table 3 below including the number of missing observations in each indicator.

|  |  |  |
| --- | --- | --- |
| Indicator | Portugal | Ireland |
| Agricultural raw materials exports (% of merchandise exports) | 2 | 4 |
| Agricultural raw materials imports (% of merchandise imports) | 2 | 4 |
| Rural population (% of the total population) | 0 | 0 |
| Rural population growth (annual %) | 1 | 1 |
| Rural population | 0 | 0 |
| Cereal yield (kg per hectare) | 2 | 2 |
| Surface area (sq. km) | 2 | 2 |
| Livestock production index (2014-2016 = 100) | 2 | 2 |
| Food production index (2014-2016 = 100) | 2 | 2 |
| Crop production index (2014-2016 = 100) | 2 | 2 |
| Cereal production (metric tons) | 2 | 2 |
| Land area (sq. km) | 2 | 2 |
| Permanent cropland (% of land area) | 2 | 2 |
| Land under cereal production (hectares) | 2 | 2 |
| Arable land (% of land area) | 2 | 2 |
| Arable land (hectares per person) | 2 | 2 |
| Arable land (hectares) | 2 | 2 |
| Agricultural land (% of land area) | 2 | 2 |
| Agricultural land (sq. km) | 2 | 2 |
| Fertilizer consumption (kilograms per hectare of arable land) | 2 | 2 |

Table 3: Indicator information

### Descriptive Statistics

Tables 4, 5, and 6 below provide an overview of the mean, standard deviation, and median of the livestock, food, and crop production index indicators in Ireland, Portugal, and the rest of Europe respectively.

#### Ireland

|  |  |  |  |
| --- | --- | --- | --- |
|  | mean | std | median |
| Livestock production index (2014-2016 = 100) | 79.0490 | 19.491166 | 85.210 |
| Food production index (2014-2016 = 100) | 80.9345 | 17.104684 | 85.905 |
| Crop production index (2014-2016 = 100) | 96.2450 | 8.795854 | 96.140 |

Table 4: Agricultural produce indices in Ireland

#### Portugal

|  |  |  |  |
| --- | --- | --- | --- |
|  | mean | std | median |
| Livestock production index (2014-2016 = 100) | 76.035500 | 25.548746 | 82.02 |
| Food production index (2014-2016 = 100) | 87.701667 | 10.752323 | 89.72 |
| Crop production index (2014-2016 = 100) | 97.732667 | 12.018663 | 96.04 |

Table 5: Agricultural produce indices in Portugal

#### The rest of the countries in Europe

|  |  |  |  |
| --- | --- | --- | --- |
|  | mean | std | median |
| Livestock production index (2014-2016 = 100) | 101.667516 | 12.968301 | 101.782600 |
| Food production index (2014-2016 = 100) | 95.487047 | 8.764715 | 96.739000 |
| Crop production index (2014-2016 = 100) | 95.851598 | 7.102554 | 95.851598 |

Table 6: Agricultural produce indices in rest of the Europe

As noted in tables 4, 5, and 6 above, Portugal had the highest crop production index compared to Ireland and the rest of Europe indicating that on average, Portugal and Ireland have high crop production compared to the rest of European countries. However, Ireland (*M* = 80.93, *SD* = 17.10) had the least food production index relative to Portugal (*M* = 87.70, *SD* = 10.75) and the rest of Europe (*M* = 95.49, *SD* = 8.76).

### Visualization (Food production index vs. cereal yield)



Figure 7: Relationship between agricultural production indices and cereal production in Ireland

As noted in figure 7 above, the food production index (β = 18403.58) has the strongest positive effect on cereal yield in Ireland indicating an increase in the food production index by 1 unit corresponds to an increase in cereal yield by approximately 18403.58 (metric tons). In contrast, it is observed in figure 8 below that the food production index has a negative relationship with cereal yield in Portugal.

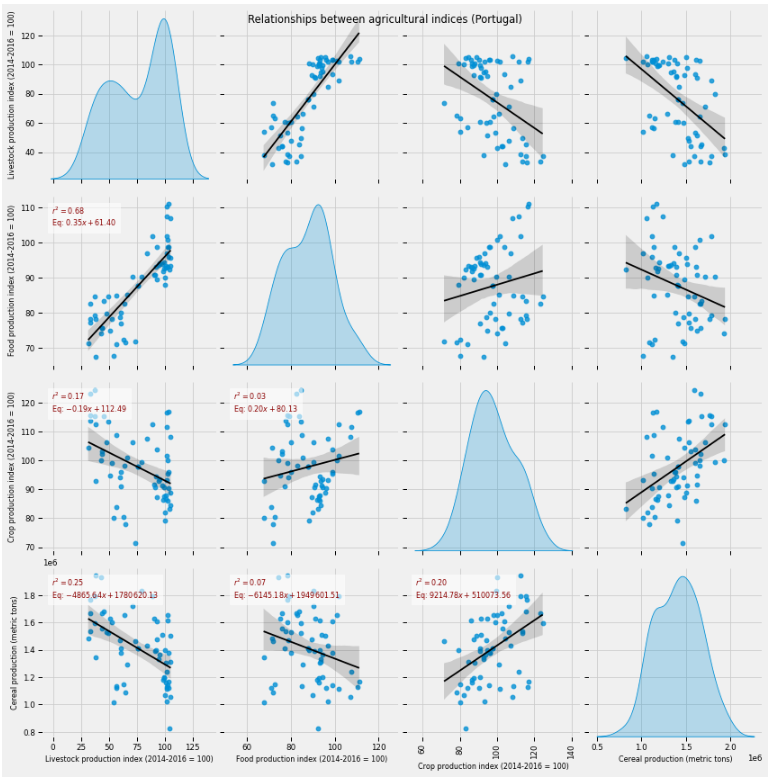


Figure 8: Relationship between agricultural production indices and cereal production in Portugal

While in the rest of Europe, the Crop production index has the strongest effect on cereal yield (*see figure 9*).

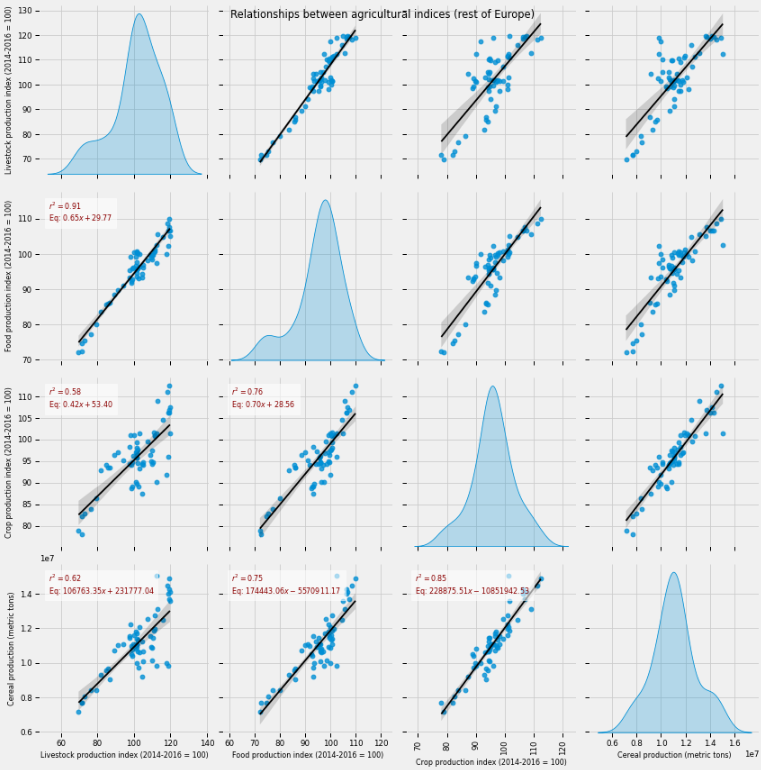


Figure 9: Relationship between agricultural production indices and cereal production (the rest of Europe)

### Population Estimates

For each of the three regions, a confidence interval was computed for the region’s cereal production. Table 7 below shows the population estimates at a 95% confidence interval.

|  |  |  |  |
| --- | --- | --- | --- |
| Region/Country | Upper limit (00s) | Point Estimate (00s) | Lower limit (00s) |
| Ireland | 14742.02 | 14106.59 | 13471.15 |
| Portugal | 19899.22 | 18894.7 | 17890.17 |
| Rest of Europe | 115330.8 | 110861.4 | 106392 |

Table 7: Population estimates in region or countries

Based on table 7 above, Ireland is estimated to have the lowest cereal yield compared to Portugal and the rest of Europe indicating that on average, Ireland produces relatively low cereals per year.

## Statistical Tests

The primary objective of the statistical tests is to compare the cereal yields of Ireland relative to Portugal and Europe.

### Distributional tests

To determine whether to conduct parametric or non-parametric tests, a normality test was conducted and the assumption was that the selected attributes are independent since they come from different samples.

#### Test for normality

A Shapiro-Wilk test was used to determine whether the samples follow a normal distribution (King & Eckersley, 2019). To this end, the hypothesis H0: **the population is normally distributed** vs H1: **the population deviates from a normal distribution was tested**.

The results for each region are given in table 8 below.

|  |  |  |
| --- | --- | --- |
| Region | W | P-value |
| Ireland | 0.9561 | 0.0266 |
| Portugal | 0.9819 | 0.4909 |
| Rest of Europe | 0.9736 | 0.2003 |

Table 8: Test for Normality

At a 0.05 level of significance, it is observed that Ireland does not follow a normal distribution (p < 0.05) as opposed to samples from Portugal and the rest of Europe with p-values greater than 0.05.

#### Test for Homogeneity of Variance

The homogeneity of variance test seeks to determine whether the distribution of two samples has the same population variance. Using a Levene’s homogeneity test which tests the hypothesis, H0: **the population variances are equal** vs H1: **the population variances of at least one pair of groups have unequal variance**, it was noted that at 0.05, none of the pairs (Europe – Ireland, Europe – Portugal, and Portugal – Ireland) had equal variances (*see table 9*).

|  |  |  |
| --- | --- | --- |
| Region | W | P-value |
| Ireland - Europe | 41.4291 | < 0.0001 |
| Portugal - Ireland | 7.6972 | 0.0064 |
| Rest of Europe - Portugal | 52.1439 | < 0.0001 |

Table 9: Test for Homogeneity of Variance

### Test for Differences in Cereal Production

From the results of distributional tests, it is established that the three groups do not meet the assumption of homogeneity of variance which is a requirement for ANOVA whereas both *Europe* and *Portugal* data follow a normal distribution. As such non-parametric tests were used i.e., a Wilcoxon signed-rank test (King & Eckersley, 2019) to compare cereal production in Ireland and Portugal and a Kruskal-Wallis test (Hoffman, 2019) to compare the cereal output of the three regions.

#### Wilcoxon signed-rank test

**Statistical claim**

H0: The median of the population of differences between the cereal production in Ireland and Portugal data is zero.

H1: There is a significant difference in cereal production in Ireland and Portugal

At 0.05, it was noted that with *W* = 227, *p* < 0.05, the null hypothesis of no difference in cereal production between Ireland (*M* = 1889470.0), and Portugal (*M* = 1410659.0) is rejected indicating that cereal production in the two countries is different.

#### Kruskal-Wallis test

Kruskal-Wallis test is used as an alternative to the ANOVA test since it does not assume normality and the data is not sensitive to outliers (Hoffman, 2019).

**Statistical claim**

H0: The mean ranks of the groups (Ireland, Portugal, and the rest of Europe) are the same.

H1: The mean ranks of at least one of the groups (Ireland, Portugal, and the rest of Europe) differs.

At 0.05, the p-value < 0.05 indicates that there is sufficient evidence to suggest that the cereal yield differs across Ireland, Portugal, and the rest of Europe, *K* = 141.34, *p* < 0.05.

### Trend Test

Primarily, the objective of a trend test was to determine the change in cereal production in Portugal, Ireland, and Europe. To this end, a Mann-Kendall statistical test (Fan, et al., 2020) was used to assess whether the cereal yield (tons) is increasing over time or decreasing over time. The Mann-Kendal test assumes there is no serial correlation in the data. However, as shown in the autocorrelation (ACF) plots in figures 10, 11, and 12 below, there is autocorrelation in the first lag for each series. Therefore, a modified Mann-Kendall Trend Test is used. The following hypothesis is tested at a 0.05 level of confidence:

**Statistical claim**

H0: There is no trend present in the data.

H1: A trend is present in the data. (This could be a positive or negative trend)

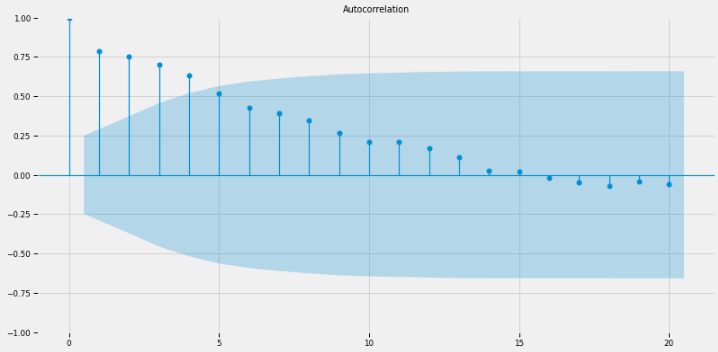


Figure 10: Trend test for Ireland

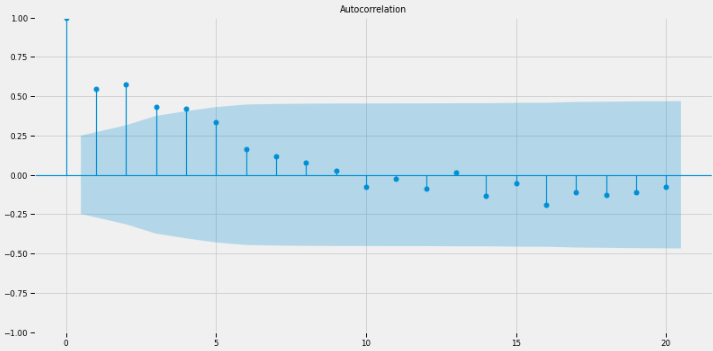


Figure 11: Trend test for Portugal

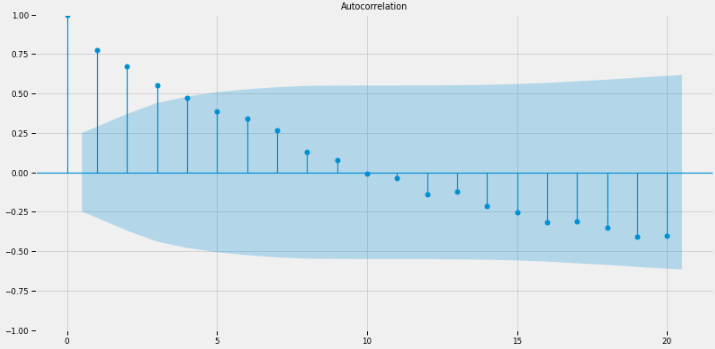


Figure 12: Trend test for the rest of Europe

Table 10 below shows the results of the trend test for each series.

|  |  |  |  |
| --- | --- | --- | --- |
| Series | Tau-Statistic | P-Value | Trend |
| Ireland | 0.5172 | 0.0007 | Increasing |
| Portugal | -0.3871 | 0.0005 | Decreasing |
| Europe | 0.1978 | 0.2657 | No trend |

Table 10: Trend test evaluation

Overall, it is noted that Ireland’s cereal production is increasing while Portugal’s production is decreasing and no change is observed in cereal production in the rest of Europe.

# Machine Learning for Data Analytics

## Machine Learning Problem

The primary objective of the current study was to develop an understanding of cereal production in Ireland relative to the rest of the world and in particular Portugal and Europe. Two machine-learning problems were proposed to extract patterns regarding cereal production in Ireland:

1. Group yearly cereal production in Ireland using clustering
2. Developing an optimal predictive regression model

### Clustering

Ideally, clustering is used to detect structures in a data (Alashwal, et al., 2019). Since clustering is unsupervised, it works on data with no outcome (target) variable as well as that which no relationship between the observations is known. Clustering is useful in creating generalizations about groups in the data. In particular, both DBSCAN (Mund, 2019) and *K-means* algorithm were fitted with the data. During model selection, a silhouette score (Bhardwaj, 2020) was used to determine the model with the optimal separation degree among the clusters.

#### K-Means Clustering

Figure 13 below shows the *sum of squared errors* (SSE) for the K-means algorithm across different values of *k*.

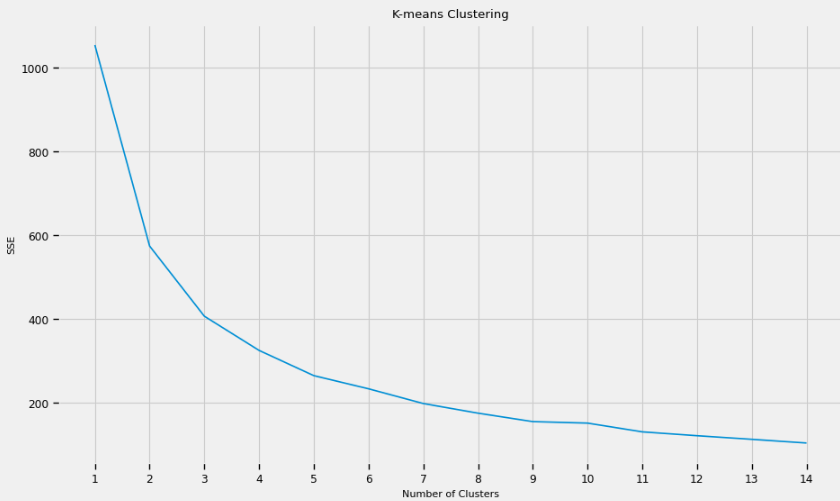


Figure 13: K-means clustering

Using the elbow method above, it is noted that the *SSE* of the model begins to flatten out at around *k* = 4. Therefore, the optimal number of groups of cereal production in Ireland was selected to be 4 using the k-means algorithm. The performance of the model was compared against a DBSCAN model as shown in figure 14.

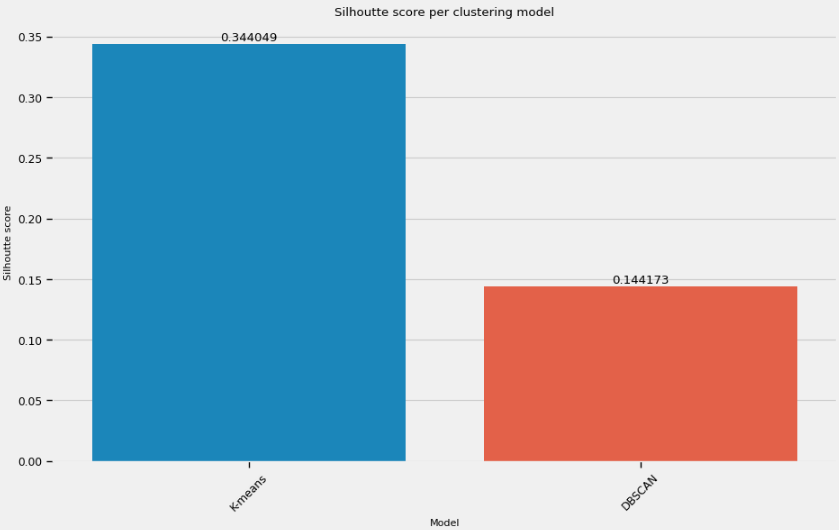


Figure 14: Silhouette Score per model

As demonstrated, the *K-means* algorithm has better degrees of separation among the clusters (0.3440) compared to the DBSCAN algorithm.

#### Average Cereal Production and food production index per Cluster

Cluster 0 had the highest average cereal production (2217298.20) as well as the highest food production index (95.40) while cluster 1 had the lowest cereal production (1332010.73) and lowest food production index (55.44)

|  |  |  |
| --- | --- | --- |
| Cluster Group | Cereal production (metric tons) | Food production index (2014-2016 = 100) |
| 0 | 2217298.20 | 95.40 |
| 1 | 1332010.73 | 55.44 |
| 2 | 1987421.33 | 80.94 |
| 3 | 1923074.30 | 90.23 |

Table 11: Average Cereal Production and food production index per cluster

Three regression models including linear regression, decision tree, and random forests were used to predict cereal production in Ireland. During implementation, the data was partitioned into train and test using a 70:30 ratio. The train data was used to train the model using RandomSearchCV which randomly selects hyperparameter combinations thus correcting the shortcomings of GridSearchCV where the combinations are defined manually (Satheesh, 2020).

The parameter space for each model is given in table 12 below.

|  |  |
| --- | --- |
| Model | Parameter Space |
| Linear Regression |  |
| Random Forest |  |
| Decision Tree |  |

Table 12: Parameter space

The linear regression model with no selection was noted to have the lowest prediction error (*see figure 15*) and hence selected as the optimal model for predicting cereal production in Ireland.

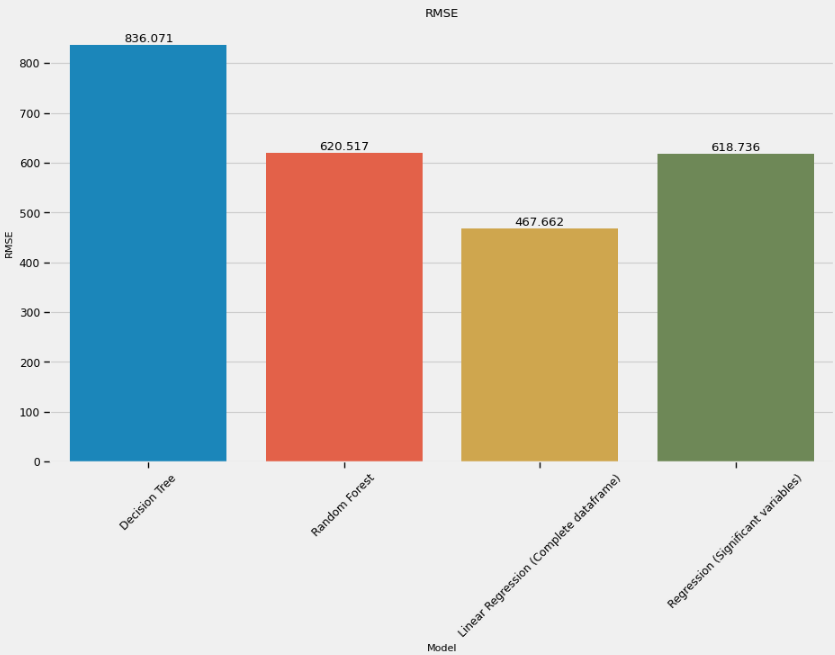


Figure 15: RMSE scores

Subsequently, model selection was conducted for the regression model which included defining a new model using only significant values from the original regression model. Table 13 below shows the coefficients of the regression model.

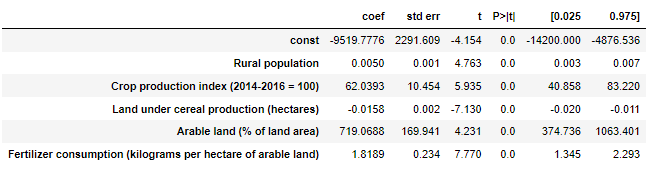


Table 13: Significant coefficients

It is observed that at 0.05 level of significance, *Arable land (% of land area), β* = 719.0688, and *Crop production index* (2014-2016 = 100), *β* = 62.0393, had the strongest positive effect on cereal production in Ireland. However, interestingly the *land under cereal production (hectares*) had a negative but small effect on cereal production indicating that an increase in land under cereal production slightly reduces the cereal production of Ireland.

Overall, the regression model explained the most variance before and after the variable selection was conducted (*see figure 16*) indicating that using the significant predictors of cereal production in Ireland, the linear regression model explains up to 89.61% of the variability in the data.

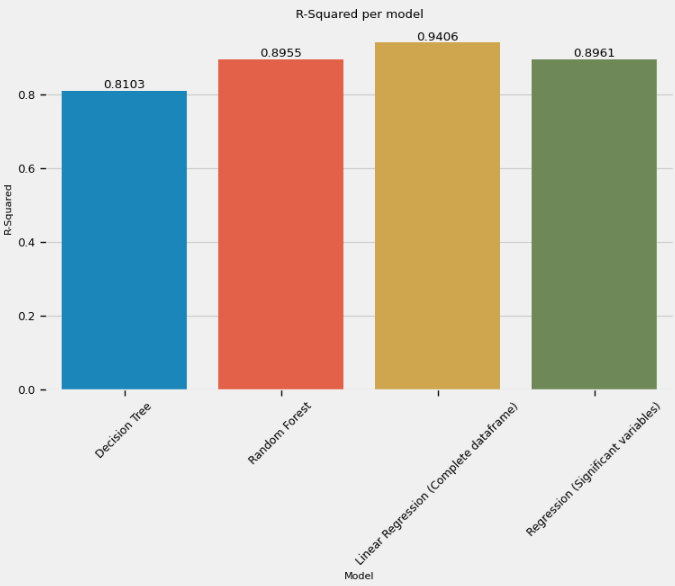


Figure 16: R-Squared scores

## Sentiment Analysis

Due to data availability, sentiment analysis was conducted on reviews made on cereal products produced in Ireland using the VADER (Valence Aware Dictionary and Sentiment Reasoner) algorithm which is a lexicon and rule-based sentiment analysis (Chauhan, et al., 2018). VADER is particularly useful since it is sensitive to sentiment expressions which is a core objective of the study’s sentiment analysis. The customer review data was scraped using *Open Web Scraper* (Web Scraper, 2022) from [https://www.dailyedge.ie](https://www.dailyedge.ie/breakfast-cereal-ranking-1473022-May2014/) which contains customer reviews for Irish cereal brands after which it was pre-processed to remove stop words and punctuations. Figure 17 below shows the top 10 most commonly used words in product reviews.

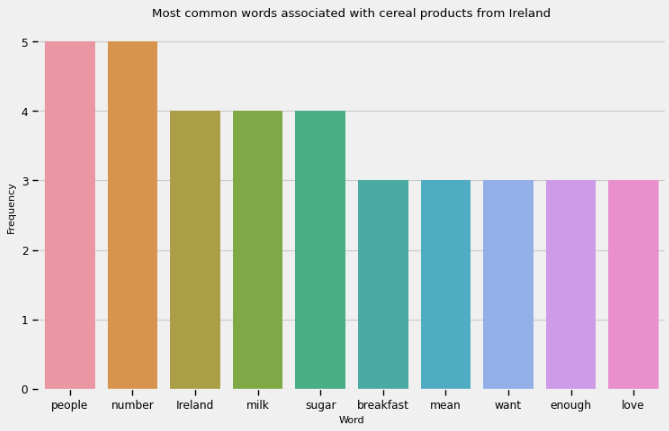


Figure 17: Common words associated with cereal products from Ireland

As shown in figure 18 below, most of the customers provided positive reviews about the Irish cereal products. However, 25% of the users had a negative overview of the products.

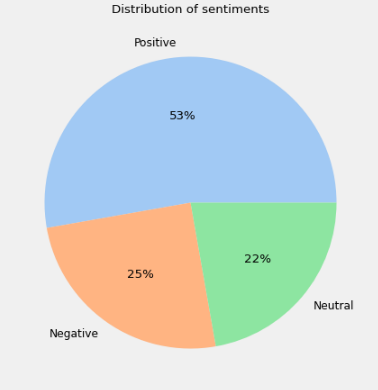


Figure 18: Distribution of sentiments

# Discussion and Conclusion

There is a wide range of cereal products ranging from wheat and corn to sorghum and barley, etcetera. Therefore, one could argue that cereal production is a key component in fighting food insecurity both in Europe and the rest of the world. The current study sought to explore the phenomenon of food production in Ireland and compare the country’s production with the rest of the world including Portugal and Europe in particular.

Based on the current study’s findings, it was noted that:

1. There is a significant difference in the amount of cereal produced in Ireland, Portugal, and the rest of Europe.
2. Cereal production in Ireland is increasing over time while Portugal’s production is decreasing whereas no change is observed in the rest of Europe, combined.
3. Factors that influence cereal production in Ireland include *Rural Population*, *Crop Production*, *Land under cereal production (hectares*), *Arable land (% of land area*), and *Fertilizer consumption (kilograms per hectare of arable land*).
4. Most consumers (53%) have a positive sentiment regarding Ireland’s cereal products.

## Study Limitations and Difficulties

### Availability of data

Availability of relevant data was a key difficulty since most of the data either included observations for a single country or had few observations. Besides, there are no available repositories with consumer or producer reviews for cereal production over time.

### Licensing

Collection of the agricultural production data related to various countries around the globe necessitated that an account should be created before downloading. The data was collected under the Creative Commons Attribution 4.0 license which allows for further modification and use of the collected data.

## Dashboard Overview

An interactive dashboard was developed to visualize the results of clustering and the cereal prediction from the linear regression model. Figures 19 and 20 provide an overview of the dashboard when the cluster is set to 3 and 0 respectively. The dashboard provides an overview of the change in the results depending on the cluster selected by the user (Farmer).

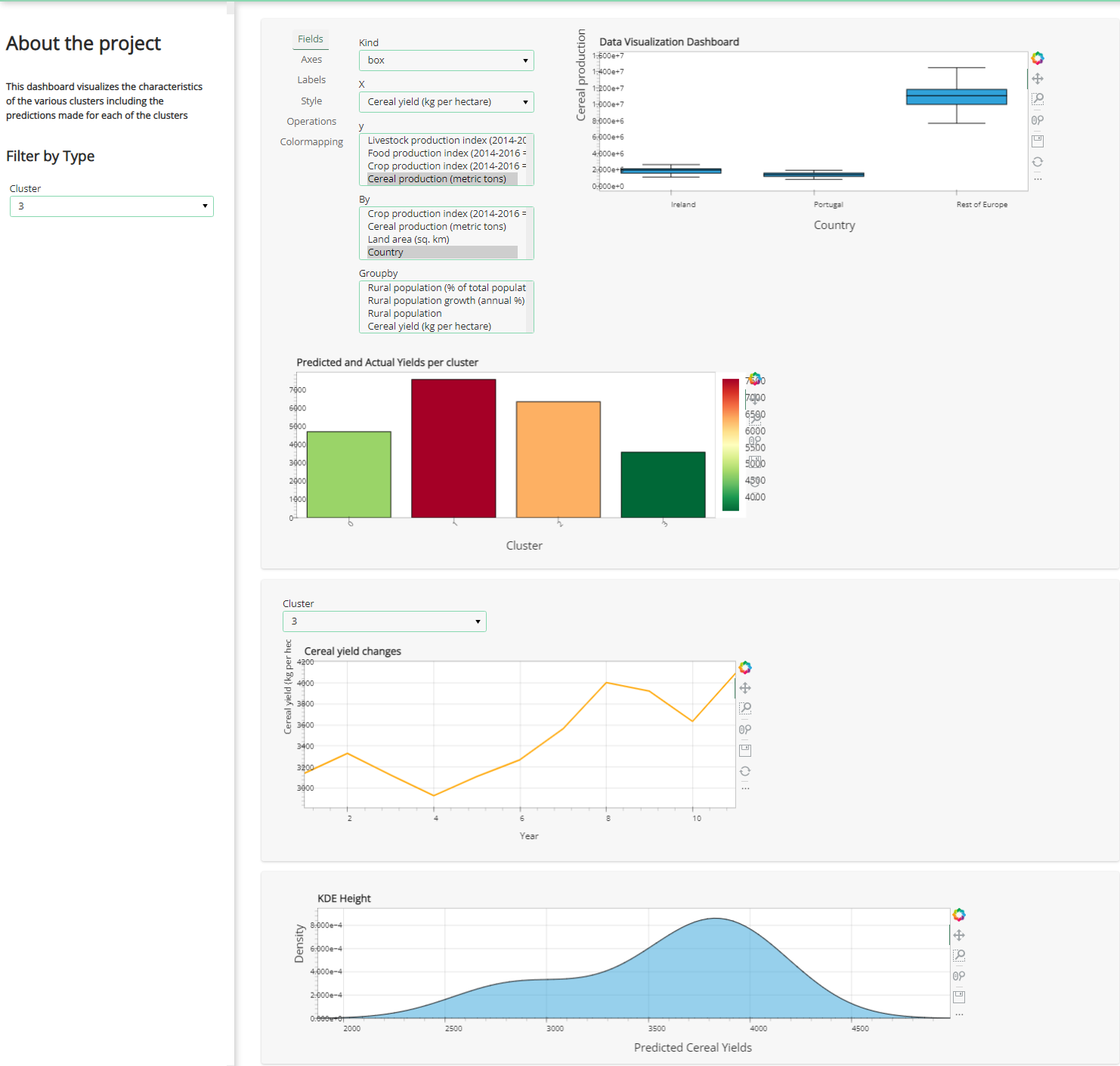


Figure 19: Cluster = 3

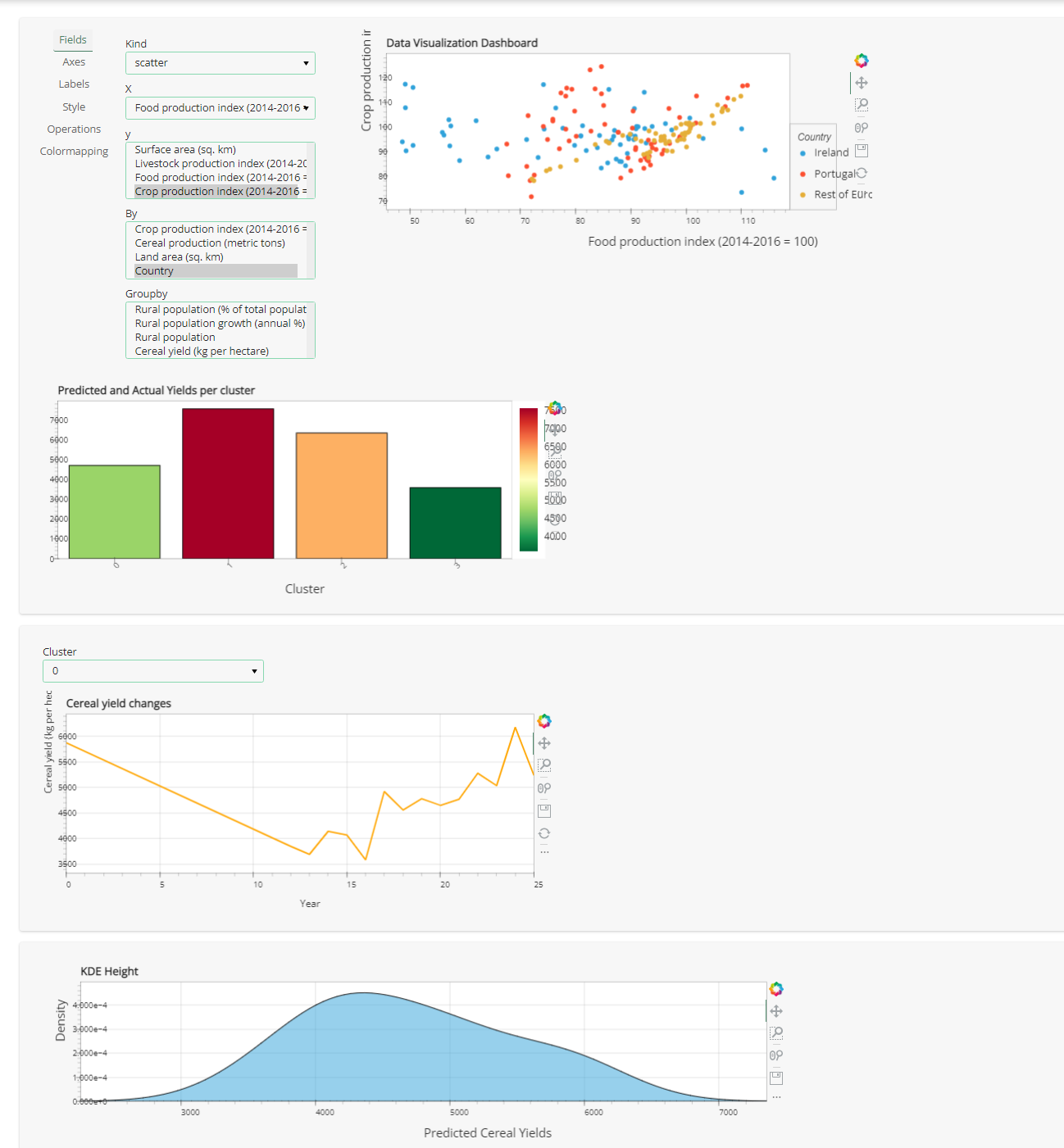


Figure 20: Cluster = 0

# REFERENCES

1. Alashwal, H. et al., 2019. The Application of Unsupervised Clustering Methods to Alzheimer's Disease. *Front Comput Neurosci,* 13(31).
2. Bhardwaj, A., 2020. *Silhouette Coefficient.* [Online]   
   Available at: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c [Accessed 20 December 2022].
3. Chauhan, V. K., Bansal, A. & Goel, D. A., 2018. Twitter Sentiment Analysis Using Vader. *International Journal of Advance Research, Ideas and Innovations in Technology,* 4(1), pp. 485-489.
4. EU, 2014. *EU CEREAL FARMS REPORT 2013,* Brussels: European Union.
5. Fan, W. et al., 2020. Re-evaluation of the Power of the Mann-Kendall Test for Detecting Monotonic Trends in Hydrometeorological Time Series. *Frontiers in Earth Science,* 8(2020).
6. Fortuna, G. & Foote, N., 2022. *EU adopts further relaxation of environmental measures to increase cereal production.* [Online]   
   Available at: https://www.euractiv.com/section/agriculture-food/news/eu-adopts-further-relaxation-of-environmental-measures-to-increase-cereal-production/ [Accessed 19 December 2022].
7. Grodzicki, T. & Jankiewicz, M., 2022. The role of the common agricultural policy in contributing to jobs and growth in EU’s rural areas and the impact of employment on shaping rural development: Evidence from the Baltic States. *Plos ONE,* 17(2).
8. Hassen, T. B. & Bilali, H. E., 2021. Impacts of the Russia-Ukraine War on Global Food Security: Towards More Sustainable and Resilient Food Systems?. *Foods,* 11(15).
9. Hoffman, J. I., 2019. Analysis of Variance. I. One-Way. In: J. I. Hoffman, ed. *Basic Biostatistics for Medical and Biomedical Practitioners.* 2nd ed. s.l.:Academic Press, pp. 391-47.
10. Johnson, A. V., 2018. *The Data Science Method (DSM) — A framework on how to take your data science projects to the next level..* [Online]   
    Available at: https://aiden-dataminer.medium.com/the-data-science-method-dsm-a-framework-on-how-to-take-your-data-science-projects-to-the-next-91f9fd81e5d1 [Accessed 20 December 2022].
11. Kelly, P., 2019. *The EU cereals sector: Main features, challenges and prospects,* s.l.: European Parliamentary Research Service.
12. King, A. P. & Eckersley, R. J., 2019. Inferential Statistics III: Nonparametric Hypothesis Testing. In: A. P. King & R. J. Eckersley, eds. *Statistics for Biomedical Engineers and Scientists.* s.l.:Academic Press, pp. 119-145.
13. King, A. P. & Eckersley, R. J., 2019. Inferential Statistics IV: Choosing a Hypothesis Test. In: A. P. King & R. J. Eckersley, eds. *Statistics for Biomedical Engineers and Scientists.* s.l.:Academic Press, pp. 147-171.
14. Kropff, M. & Morell, M., 2019. *The cereals imperative of future food systems,* Carretera México-Veracruz: International Maize and Wheat Improvement Center (CIMMYT).
15. Martinez, I., Viles, E. & Olaizola, I. G., 2021. Data Science Methodologies: Current Challenges and Future Approaches. *Big Data Research ,* 24(3), p. 100183.
16. Mund, S. K., 2019. *How does DBSCAN clustering algorithm work?.* [Online]   
    Available at: https://shritam.medium.com/how-dbscan-algorithm-works-2b5bef80fb3#:~:text=A%20silhouette%20score%20ranges%20from,of%200%20suggest%20overlapping%20clusters. [Accessed 20 December 2022].
17. Raheem, D., Dayoub, M., Birech, R. & Nakiyemba, A., 2021. The Contribution of Cereal Grains to Food Security andSustainability in Africa: Potential Application of UAV inGhana, Nigeria, Uganda, and Namibia. *Urban Science,* 5(8).
18. Satheesh, V., 2020. *Hyper Parameter Tuning (GridSearchCV Vs RandomizedSearchCV).* [Online] Available at: https://medium.com/analytics-vidhya/hyper-parameter-tuning-gridsearchcv-vs-randomizedsearchcv-499862e3ca5 [Accessed 20 December 2022].
19. Schils, R. et al., 2018. Cereal yield gaps across Europe. *European Journal of Agronomy,* 101(2018), pp. 109-120.
20. United Nations System High-Level Task Force on Global Food Security, 2011. *Food and Nutrition Security: Comprehensive Framework for Action. Summary of the Updated Comprehensive Framework for Action (UCFA),* Rome, Italy: United Nations System High Level Task Force on Global Food Security.
21. Web Scraper, 2022. *Open Web Scraper: Documentation.* [Online]   
    Available at: https://webscraper.io/documentation/open-web-scraper [Accessed 21 December 2022].

## GitHub Reference