**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA,  Statistics for Data Analytics,  Machine Learning for Data Analysis,  Data Preparation & Visualisation |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Rita Raher |
| **Student Number:** | sbs22115 |
| **Assessment Due Date:** | 22/05/2022 |
| **Date of Submission:** | 22/05/2022 |

Icon

Description automatically generated with medium confidence

**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.

Table of Contents

[Introduction 3](#_Toc104153132)

[Methodology 3](#_Toc104153133)

[Data Collection 4](#_Toc104153134)

[Data understanding 5](#_Toc104153135)

[Inferential statistics 6](#_Toc104153136)

[T-Test 6](#_Toc104153137)

[Anova 6](#_Toc104153138)

[Comparing countries against Ireland 7](#_Toc104153139)

[T-Test 7](#_Toc104153140)

[ANOVA 7](#_Toc104153141)

[Non-parametric 7](#_Toc104153142)

[Wilcoxon signed-rank test 7](#_Toc104153143)

[Kruskal-Wallis H Test 8](#_Toc104153144)

[Mann-Whitney Test 8](#_Toc104153145)

[Data preparation 8](#_Toc104153146)

[I. Handling missing values 8](#_Toc104153147)

[II. Handle duplicates 9](#_Toc104153148)

[III. Remove outliers 9](#_Toc104153149)

[IV. Encoding values 9](#_Toc104153150)

[V. Apply feature scaling 9](#_Toc104153151)

[Modelling 9](#_Toc104153152)

[Linear regression 10](#_Toc104153153)

[Ridge regression 10](#_Toc104153154)

[Lasso regression 10](#_Toc104153155)

[ElasticNet 11](#_Toc104153156)

[Sentiment analysis 11](#_Toc104153157)

[Topic modelling 11](#_Toc104153158)

[Evaluation 12](#_Toc104153159)

[Dashboard 12](#_Toc104153160)

[Conclusion 12](#_Toc104153161)

[Bibliography 14](#_Toc104153162)

# Introduction

There has been increased pressure on Irish farmers due to rising input costs (Teagasc, 2022). One of the main input costs on the rise is fertiliser, which has seen a huge jump this year (Blaney, 2022) (Murphy, 2022). There is also rising concern that these soaring costs will have a knock-on effect on Irish consumers in the near future (O'Brien, 2022). Many attribute the rising costs to the war in Ukraine as Russia is a big producer of the ingredient that make up fertilisers (REUTERS, 2022). This report will analyse fertiliser price, understand what input variables have an effect on the price and also build a model to predict fertiliser price based on those inputs.

Please find links to GitHub repo and dashboard:

<https://github.com/RitRa/Msc_CA2>

<https://share.streamlit.io/ritra/msc_ca2/notebooks/dashboard.py>

Report word count: 3066

# Methodology

The Cross-Industry Standard Process for Data Mining, as known as CRISP-DM, the methodology was developed with DiamlerChrysler, SPSS and NCR in 1996 (Santos, 2008). CRISP-DM, KDD, and SEMMA have all been compared, and CRISP-DM was found to be the most robust process for data mining projects (Qaiser, 2014). It consists of a cycle of 6 steps, as seen in Figure 1 (Chapman, 2000).

Diagram

Description automatically generated

Figure 1: Phases of the CRISP-DM reference model

1. Business understanding: Assess the data project from a business perspective and formulate a plan to achieve those goals.
2. Data understanding: Collect and understand the data to help gain insight and formulate the hypothesis.
3. Data preparation: begin activities for preparing the data for modelling, including pre-processing, transformation and cleaning.
4. Modelling: apply various modelling techniques to the data
5. Evaluation: evaluate and review the modelling techniques used during the modelling step and analyses whether the business goals were achieved.
6. Deployment: Create a real-world webpage for the model

# Data Collection

Using the [Central Statistics Office](https://www.cso.ie/en/statistics/), CSO, an Irish dataset on fertiliser prices was sourced from 1980 to 2022. According to multiple sources, including a leading fertilisers manufacturer in Norway, fertilisers are made up of potassium, nitrogen and phosphate. During production, it uses a large amount of natural gas (Yara, 2017) . Due to this research, additional features were added to the datasets for potassium, Urea (as known as Urea), phosphate and Natural gas. These datasets are readily available on [www.indexmundi.com](http://www.indexmundi.com) and will be used as part of the analysis and models, see Table 1 .

According to Teagasc, there are multiple factors for the increase in fertiliser prices; they include inflation due to covid recovery, sanctions on Russia due to the War in Ukraine and farmers delaying purchasing fertilisers with hopes that the price will drop (Teagasc, 2022; Liboreiro, 2022). Due to this evidence, a feature for inflation is added to the dataset to hopefully try and capture these effects on the price of fertilisers. The dataset is sourced from <https://db.nomics.world/> and looks at housing, water, electricity, gas and other fuels.

Table : Data sources

|  |  |
| --- | --- |
| Datasets | Link |
| Fertilisers | [source](https://data.cso.ie/table/AJM05) |
| Potassium | [source](https://www.indexmundi.com/commodities/?commodity=potassium-chloride&months=240&currency=eur) |
| Urea (Nitrogen) | [source](https://www.indexmundi.com/commodities/?commodity=urea&currency=eur) |
| Phosphate | [source](https://www.indexmundi.com/commodities/?commodity=rock-phosphate&months=360&currency=eur) |
| Natural Gas | [source](https://www.indexmundi.com/commodities/?commodity=natural-gas&months=120&currency=eur) |
| Consumer price index ireland | [source](https://db.nomics.world/OECD/MEI/IRL.CP040500.CTGY.M) |

Part of this project is to analyse sentiment related to agriculture. One of the best places to find out what the general population are thinking is Twitter. One way to gain access to tweets is by using a developer account; this will give you an API key, API key secret, access token and access token secret. To get elevated access to Twitter, you need to explain what you are doing with the tweets. The developer account gives you access to 2M Tweets per month / Project. A python library called Tweepy allows for easy access to the Twitter API. Unfortunately, Twitter only provides API access for up to 5-7 days of tweets or 1,500 tweets in total, so the total number of tweets returned was only 9, not enough for this project. Snscrape is a python web scraper for social networking services which allows you to add search criteria and gives back tweets without the restrictions of others like Tweepy. Specific keywords were targeted to focus on the topic of farmer costs. They included: ‘farmers prices’, ‘fertilisers prices’, ‘farming prices’, ‘agriculture spend, ‘farm cost’, ‘fertilisers’, ‘farm spend’ and each search criteria included “near: Ireland” to try and collect tweets in Ireland only. After merging all the datasets, it resulted in 46 tweets ready for sentiment analysis. (Jupyter: tweets)

# Data understanding

Descriptive statistics is an effective way of summarising data and identifying patterns.

They are measures that show where the centre of the data line is and are called measures of central tendency, which is the measure of the centre (Weiss, 2017). Central tendency includes: mean, median and mode.

* The mean is the sum of observations and dividing by the total
* The median finds the middle of the data
* The mode finds the most frequently reoccurring value

A boxplot shows the range and spread of data, with the middle of the data being represented by the line in the middle of the box. In this case, the median is 245. Therefore, 50% of the data is represented to the right and left. The first and second quartiles have a smaller spread and are much closer to the median than the third and fourth quartiles. This means that the number of observations condensed into the lower quartiles is greater than the upper quartile. The boxplot also identifies a large number of extreme outliers ranging from 620 to 890 (Jupyter: CA2\_Agri [22-23]).

|  |  |
| --- | --- |
| Figure 2: Boxplot of fertiliser price | Figure 3: Histogram of fertiliser price |

In this dataset, they are 24 types of fertilisers. In the 1980s and the 1990s, there were only 11 to 12 types, and in the last decade, the number of fertilisers available has exploded to over 20 types (Jupyter: CA2\_Agri [56]). When focusing on the last few years, it is clear that fertiliser Urea prices have been increasing steadily since mid-2021, well before the war in Ukraine began, see Figure 4.

Chart

Description automatically generated

Figure 4: Price increases

## Inferential statistics

Inferential statistics allows you to make inferences or estimations about the population based on the sample data (B. S. Everitt, 2010) . Several sampling methods can be used; some examples are random sampling and stratified sampling.

Stratified sampling is a way of obtaining samples that best represent the population as it divides the data into subgroups and takes a sample from each subgroup (Weiss, 2017).

Using this stratified sampling approach, the dataset was grouped by fertiliser\_type to sample each subgroup of fertiliser type, taking five from each (Jupyter: CA2\_Agri[27]). A probability plot was graphed, and a Shapiro-Wilk test was performed to determine whether the sample data were normally distributed. The Shapiro-Wilk test revealed that the sample data was not normally distributed as it had a p-value of 5.32.

### T-Test

For this project, it will be assumed that it is normally distributed. A t-test is performed on the sample dataset to check whether the sample mean price of fertilisers is equal to 288 yielding a p-value of 0.00000668. The null hypothesis is rejected, and the alternative hypothesis is accepted. The difference between the mean of the sample data and 288 is statistically significant (Jupyter: CA2\_Agri[33]).

### Anova

After creating sample data, the data for selected fertiliser types are checked to see if it is normally distributed using a Shapiro test (Jupyter: CA2\_Agri[35-40]) .Two out of the five samples were found not to be normally distributed, so continuing with three normally distributed, a ANOVA test was performed. A p-value of 0.06 fails to reject the null hypothesis (Jupyter: CA2\_Agri[43]).

## Comparing countries against Ireland

Here are the following tests which were used to compare countries

|  |  |
| --- | --- |
| Parametric | Non-Parametric |
| T-Test | Wilcoxon signed-rank test |
| Anova | Kruskal-Wallis H Test |
|  | Mann-Whitney U Test |

### T-Test

A t-test is a parametric statistical test used to compare the means of two groups (Brownlee, 2021). The Independent variable is geo, IE and PL (Ireland and Poland), and a t-test can be used to see if there is a difference between the fertiliser consumption volume of each group or if it is the same.

A sample of 15 is randomly selected from each dataset (Jupyter: CA2\_Agri[81]), and then a shapiro\_wilk test is performed to check if the data is normal for each one. It is found not to be normal, but in this project, we shall proceed with a t-test. The test output yields a t-statistic of 3.843 and a p-value is 0.001, as the p-value is less than alpha (0.05), and the null hypothesis is rejected. There is a difference between the means, which is statistically significant (Jupyter: CA2\_Agri[85].

### ANOVA

Analysis of variance test can be used to compare more than two groups. The Independent variable is geo, IE, PL and DE (Ireland, Poland, Germany). A random sample of 15 from each country was selected for the test. The ANOVA test produced a p-statistics of less than 0.05; the null hypothesis is rejected. The difference between the three countries is statistically significant (Jupyter: CA2\_Agri[95]).

### Non-parametric

Non-parametric tests do not assume a normal distribution.

### Wilcoxon signed-rank test

Statistical non-parametric tests are used for paired or dependent samples to discover if there is a difference between the samples (O'Loughlin, 2021). It is used when the data does not follow a normal distribution (pythonfordatascienceorg, 2018).

Firstly the data for the test was prepared, focusing on the years 2019 and 2020 and ‘Nitrogen’. Pandas pivot was used to create a dataframe for the analysis with geo, 2019 and 2020 as the columns (Jupyter: CA2\_Agri[99]). . A Shapiro test was performed to ensure that the data was suitable for a non-parametric test; it found that the data was not from a normal distribution and a good candidate for a Wilcoxon test. Using the scipy library, the Wilcoxon test was performed, and a p-value of 0.00000210 was returned (Jupyter: CA2\_Agri[102]).Therefore the null hypothesis can be rejected. There is a significant difference in fertiliser consumption from 2019 to 2020.

### Kruskal-Wallis H Test

Test of hypothesis to determine whether there is a difference in rank totals between independent groups (O'Loughlin, 2021). This test is an alternative to the ANOVA test; an ANOVA can only be used if the data is normally distributed (Brownlee, 2018).

Using the scipy library, the Kruskal test was performed on df\_de, df\_ie and df\_pol (Jupyter: CA2\_Agri[103]). It returned a p-value of 0.0002879, which means that the ranks of the groups were not the same and rejected the null hypothesis.

### Mann-Whitney Test

A non-parametric test, the Mann-Whitney test compares two sample means from the same population (Weiss, 2017). It is an alternative to the T-test and can be used when the data does not follow a Normal distribution. Using the scipy library, the Mann-Whitney U test was performed using sample data created previously, ie\_sampled and pol\_sampled. The p-value was found to be 0.0464, which is less than alpha (0.05) (Jupyter: CA2\_Agri[104]). Therefore the sample means are different, and the null hypothesis can be rejected.

# Data preparation

To prepare the data for machine learning algorithms, here are the steps followed:

1. Handling missing values
2. Handle duplicates
3. Remove outliers
4. Encoding values
5. Apply feature scaling

### Handling missing values

After merging new features, the dataset was now missing values from multiple columns, such as consumer price index and milk price. Some fertiliser types were missing pricing data because they were new to the market. From 1980 to 2013, there were only 10 to 11 fertiliser types; however, this grew in 2014 to over 20 types, see Figure 5. After dropping rows earlier than 2015, 14% of data were missing from fertiliser types, which is still too large to impute without introducing bias into the dataset. The fertiliser types with the most significant missing values were identified and dropped from the dataset (Jupyter: CA2\_Agri[116]). This left 150 missing values, 8.7%, which means the rest of the values can be imputed. The missing values are replaced using the KKNimputer from sklearn (Jupyter: CA2\_Agri[120]).

Chart, bar chart

Description automatically generated

Figure 5: Fertilisers type growth

### Handle duplicates

Duplicates can create bias in the model; it is essential to identify and remove them. Luckily, no duplicates were found in the dataset (Jupyter: CA2\_Agri[122]).

### Remove outliers

An outlier is an anomaly outside of the lower and upper quartiles of the data, generally representing either high or low extremes (GRUBBS, 1969). After removing outliers from the data, the model performance dropped 5%, so outliers were left in the dataset.

### Encoding values

Machine learning models require all input and output variables to be numeric to allow for them to perform mathematical computations and statistical analysis (Casari, 2018). Firstly, categorical data are transformed into numeric values using the category encoders library (Jupyter: CA2\_models[12]). To ensure that these values are not given any statistical significance over each other, OneHotEncoding is used to create new columns for each value, and the rows are then filled with 1 or 0, representing true or false if that value exists (Jupyter: CA2\_models[13]). Before completing this step of the process, the model reported a precision of 40%, and after it yielded a precision of 85%.

### Apply feature scaling

Machine learning algorithms will give more weight to features with larger numeric values (Roy, 2020). Milk price, gas price and consumer price index have small values and fertilisers price, phosphate price, urea price, and potassium price. Using StandardScaler from sklearn will bring everything to the same magnitude (Jupyter: CA2\_models[15]).

# Modelling

The goal of this project is to predict the price of fertilisers. Therefore, this is a regression problem rather than a classification problem. The heatmap, Figure 6, uses three distinct colors, so it would be easy to distinguish between high, medium and low correlation. The independent variables that have mid to high correlation with fertiliser price are: gas price (0.41), consumer\_price\_index(0.51), phosphate price(0.65) and urea price(0.71). Urea has the strongest correlation with fertilisers price. It seems likely that a regression task will solve this problem easily, especially with features showing such a high correlation. The regression algorithms utilised in this project are Linear, Ridge, Lasso and ElasticNet.

**Chart

Description automatically generated**

Figure 6: Correlation heatmap of variables

## Linear regression

Linear regression or ordinary least squares (OLS), is one of the most well-known machine learning algorithms (Brownlee, 2020; Guido, 2017). It is a supervised learning algorithm which tries to find a linear relationship between the independent and dependent variables. Linear regression is represented by ***Y = a + bX***, where ***X*** is the independent variable and ***Y*** is the value that we wish to predict. To measure how good the model is explained using the R2 value. The model resulted in a high R2 score of 0.89, meaning 89% of the variance in observed data is explained by the model, and 11% is unexplained (Jupyter: CA2\_models[16-22]).

## Ridge regression

Ridge regression is another supervised learning algorithm that uses a linear model for regression (Guido, 2017). It includes adding penalties, called L2 regularisation, to coefficients which do not contribute to the overall model (Trevor Hastie, 2004). Ridge results in a high R2 score of 0.89 (Jupyter: CA2\_models[30-34]).

## Lasso regression

Lasso Regression is another regularising linear regression algorithm which uses a shrinkage method similar to Ridge. It also restricts the coefficients, however, this algorithm used an L1 regularisation method which means that some coefficients are exactly 0 and are ignore from the model completely. (Guido, 2017). Lasso resulted in a high R2 score of 0.88 (Jupyter: CA2\_models[24-29]).

## ElasticNet

ElasticNet is a regularising linear regression algorithm that uses L1 and L2 penalties from Ridge and Lasso methods (Guido, 2017). ElasticNet yielded the worst score of 0.85 (Jupyter: CA2\_models[35-39])

## Sentiment analysis

Appropriate cleaning techniques were used to prepare the tweets for natural language processing: text lowercased, Twitter handles removed, square brackets removed, punctuation removed, and words containing numbers removed. Stop words were identified using the NLTK library and removed. Later, extra stop words were included for “farm” and “farms”. Text blob, another python library for text analysis, can be used to extract the sentiment of a tweet. It can identify the polarity and subjectivity of the tweet. Polarity represents whether the tweet is positive or negative from a range of -1 to 1.

Subjectivity represents whether the tweet is fact or an opinion, it is presented between 0 and 1 (Jain, 2018). Each dot on the plot, Figure 3, represents a tweet, and most of them are in the bottom left quadrant of the plot, which means the tweets being interpreted

as negative and factual apart from two outliers in the top right quadrant: positive and opinions (Jupyter: CA2\_models[80-84]).

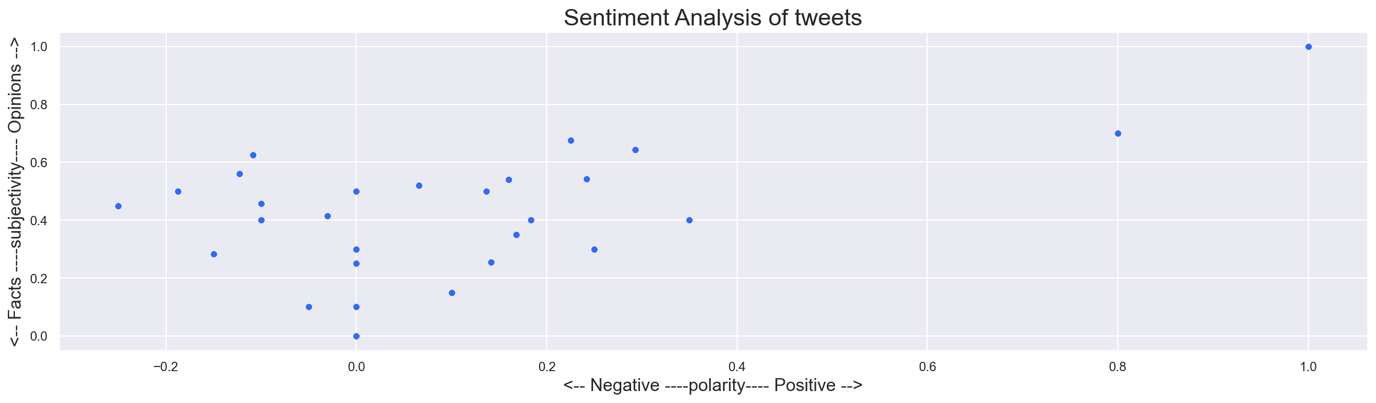


Figure : Sentiment Analysis of tweets

### Topic modelling

Using Genism, which is a library from Natural Language Processing, the tweets will be analyseds and grouped into topics using LDA, latent [Dirichlet](https://en.wikipedia.org/wiki/Dirichlet_distribution) allocation. This is an example of topic modelling and an unsupervised technique (Guido, 2017).

Firstly, the cleaned tweets are turned into a matrix of tokens by using CountVectorizer from sklearn (Jupyter: CA2\_models[85]) and then converted into a bag of words format using genism, and then the model is trained on the corpus. After multiple iterations and multiple keyword changes, after starting too broad at first, the 2 main topics were identified: Price increase and War. (Jupyter: CA2\_models[87-109])

[(0,

'0.029\*"prices" + 0.027\*"farmers" + 0.015\*"fertiliser" + 0.013\*"year" + 0.011\*"beef" + 0.009\*"record" + 0.008\*"meat" + 0.008\*"live" + 0.006\*"milk" + 0.005\*"increase"'),

(1,

'0.043\*"fertiliser" + 0.038\*"prices" + 0.014\*"war" + 0.014\*"global" + 0.012\*"fuel" + 0.012\*"day" + 0.010\*"food" + 0.010\*"climate" + 0.010\*"oils" + 0.010\*"grains"')]

# Evaluation

Metrics used to measure the performance of regression models are Mean Absolute Error (MAE), Mean Squared Error(MSE), Mean Squared Error (RMSE) and Coefficient of determination (R2 score)

|  |  |  |
| --- | --- | --- |
|  | Precision | Precision after outliers removed |
| Linear Regression | 0.90 | 0.86 |
| Lasso Regression | 0.89 | 0.84 |
| Ridge Regression | 0.90 | 0.86 |
| ElasticNet | 0.86 | 0.79 |

Table 2: ML Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Precision | MAE | MSE | RMSE | R2 Score | Gridsearch Best score |
| Linear Regression | 0.90 | 14.03 | 667.93 | 25.84 | 0.89 | 0.88 |
| Lasso Regression | 0.89 | 15.79 | 723 | 26.89 | 0.88 | 0.88 |
| Ridge Regression | 0.90 | 14.01 | 667.93 | 25.84 | 0.89 | 0.88 |
| ElasticNet | 0.86 | 19.38 | 950.03 | 30.82 | 0.85 | 0.87 |

All models performed pretty well, linear regression and ridge regression performed the best; they have the lowest MAE and highest R2 score. ElasticNet performed the worst.

Using GridSearchCV all the hyperparameters were passed in for each of the above models to find the best score. Elasticsearch is the only algorithm which saw an improvement.

(Jupyter: CA2\_models[40-44])

## Dashboard

Streamlit(<https://streamlit.io/> ) is an open-source framework for building apps, and it can be used alongside plotly to build very nice interactive dashboards or apps. Getting up and running was very easy, it links to the GitHub project and checks for a requirements.txt to know what to install for the application to work and allows you to host your site for free. Here is a link to the dashboard: <https://share.streamlit.io/ritra/msc_ca2/notebooks/dashboard.py>

The dashboard includes some descriptive statistics of the fertilisers dataset with an animation of fertiliser price for each fertiliser type over the years and a simple boxplot with all the fertiliser types by price. It also allows the user to select from the dropdown the fertiliser type they want and shows the price of the fertiliser and histogram of the price.

There is also another page for models, which allows the user to pick from a dropdown the type of model, and it draws a plot, table and metrics based on the selection.

# Conclusion

After analysing the data, it is clear that the War in Ukraine was not the main factor causing an increase in prices, fertiliser prices were rising from mid-2021 due to inflation from the covid recovery. Fertiliser prices are also affected by the price of materials used to make fertiliser. However, most of the data is until March 2022 and it may tell a different story in the next few month as the war in the Ukraine continues and further sanctions are placed on Russia. Interestingly, when removing outliers from the dataset the models performed poorer. The biggest increase in the model's precision was after applying the onehotencoder, prior to that model precision was 40%. All regression models performed very well.

# Bibliography

B. S. Everitt, A. S., 2010. *The Cambridge Dictionary of Statistics.* 4th edition ed. s.l.:Emerald Group Publishing Limited.

Blaney, A., 2022. *Price of fertiliser doubles for farmers with knock on anticipated in food prices.* [Online]   
Available at: https://www.irishtimes.com/news/ireland/irish-news/price-of-fertiliser-doubles-for-farmers-with-knock-on-anticipated-in-food-prices-1.4827498  
[Accessed 02 05 2022].

Brownlee, J., 2018. *How to Calculate Nonparametric Statistical Hypothesis Tests in Python.* [Online]   
Available at: https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/#:~:text=The%20Kruskal%2DWallis%20H%2Dtest,p%2Dvalue%20as%20the%20result  
[Accessed 11 05 2022].

Brownlee, J., 2020. *Linear Regression for Machine Learning.* [Online]   
Available at: https://machinelearningmastery.com/linear-regression-for-machine-learning/  
[Accessed 22 05 2022].

Brownlee, J., 2021. *17 Statistical Hypothesis Tests in Python (Cheat Sheet).* [Online]   
Available at: https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/  
[Accessed 01 05 2022].

Casari, A. Z. a. A., 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* First Edition ed. Beijing: O’Reilly Media, Inc.

Chapman, P., 2000. *CRISP-DM 1.0,* s.l.: SPSS.

GRUBBS, F. E., 1969. *Procedures for Detecting Outlying Observations in Samples,* s.l.: Technometrics.

Guido, A. C. M. a. S., 2017. *Introduction to Machine Learning with Python.* First Edition ed. Beijing: O’Reilly Media.

Jain, S., 2018. *Natural Language Processing for Beginners: Using TextBlob.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/#:~:text=Polarity%20is%20float%20which%20lies,of%20%5B0%2C1%5D.  
[Accessed 22 05 2022].

Liboreiro, J., 2022. *Ukraine war pushes eurozone's inflation to a record 7.5%.* [Online]   
Available at: https://www.euronews.com/my-europe/2022/04/01/ukraine-war-pushes-eurozone-s-inflation-to-a-record-7-5#:~:text=Russia's%20invasion%20of%20Ukraine%20has,of%20the%20post%2Dpandemic%20recovery.  
[Accessed 21 05 2022].

Murphy, B., 2022. *Official figures confirm the soaring price of fertiliser Costs more than doubled compared to 12 months ago.* [Online]   
Available at: https://www.farmersjournal.ie/fertiliser-prices-up-127-in-year-to-january-2022-686075  
[Accessed 02 05 2022].

O'Brien, A., 2022. *Fertiliser prices impact farmers and consumers – McGuinness.* [Online]   
Available at: https://www.agriland.ie/farming-news/fertiliser-prices-impact-farmers-and-consumers-mcguinness/#:~:text=The%20spiraling%20cost%20of%20fertiliser,(CAN)%20and%20urea%20fertilisers  
[Accessed 14 05 2022].

O'Loughlin, E., 2021. *How To... Perform a Kruskal-Wallis Test in R #97.* [Online]   
Available at: https://www.youtube.com/watch?v=NVlctlg\_sIA  
[Accessed 11 05 2022].

O'Loughlin, E., 2021. *How To... Perform a Wilcoxon Signed Rank Test in R #96.* [Online]   
Available at: https://www.youtube.com/watch?v=zE4Os7JCg34&t=5s  
[Accessed 11 05 2022].

pythonfordatascienceorg, 2018. *Wilcoxon Sign-Ranked Test.* [Online]   
Available at: https://pythonfordatascienceorg.wordpress.com/wilcoxon-sign-ranked-test-python/  
[Accessed 11 05 2022].

Qaiser, U. S. a. H., 2014. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research,* 12(1), pp. 217-222.

Roy, B., 2020. *All about Feature Scaling.* [Online]   
Available at: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35  
[Accessed 26 03 2022].

Santos, A. A. a. M., 2008. *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW,* Amsterdam: IADIS European Conference on Data Mining 2008.

Teagasc, 2022. *Situation and Outlook for Irish Agriculture April 2022,* Galway: Teagasc.

Trevor Hastie, R. T. F., 2004. *The Elements of Statistical Learning.* Second Edition ed. New York: Springer.

Weiss, N. A., 2017. *Introductory Statistics.* 10th edition ed. Essex: Pearson Education Limited.

Yara, 2017. *How we make our fertiliser.* [Online]   
Available at: https://www.yara.com/crop-nutrition/why-fertilizer/production-of-fertillizer/  
[Accessed 14 05 2022].