CCT College Dublin

Assessment Cover Page

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

---

Module Title:	Data Analytics Project | MSc in Data Analytics Programming for DA Statistics for Data Analytics Machine Learning for Data Analysis Data Preparation & Visualisation
Assessment Title:	MSC_DA_CA2
Lecturer Name:	David McQuaid Dr. Muhammad Iqbal Sam Weiss Marina Iantorno
Student Full Name:	Alexandru Ciaica
Student Number:	**sba22251**
Assessment Due Date:	06/01/2023
Date of Submission:	06/01/2023

Alexandru Ciaica

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.

CA 2

Prepared by: Alexandru Ciaica

MSC Data Analytics

Student Number: **sba22251**

Date: 06/01/2023

Abstract

I. Introduction

- Overview of CRISP DM
- Why CRISP DM is important in data science projects

II. Business Understanding

- Defining the project objectives and goals
- Identifying stakeholders and their needs
- Defining the project scope

III. Data Understanding

- Collecting and preparing data for analysis
- Exploring and visualizing data
- Verifying data quality

IV. Data Preparation

- Cleaning and transforming data
- Selecting relevant data for analysis
- Creating derived variables

V. Modeling

- Selecting and training a model
- Evaluating model performance
- Fine-tuning the model

VI. Evaluation

- Assessing model performance against business objectives
- Identifying potential improvements
- Communicating results to stakeholders

VIII. Conclusion

- Summary of key points from the project
- Implications and next steps
- Reflection on the use of CRISP DM in the project.

Abstract:

This project is based on the Food and **Agriculture Organization of the United Nations (FAO)**. Data from FAO was selected to represent Producer Prices and **Producer Price Index for Agriculture**.

Agri-Producer Prices refer to the prices farmers receive at the point of sale for their primary crops, live animals, and livestock. Data for the 27 EU member countries is provided by the FAO dataset beginning in 1991 and continuing through 2020.

An agricultural producer price index measures changes in the average selling prices received by farmers over time (prices at the farm gate or at the first point of sale).

This study examines and identifies changes in the agricultural producer price index over time in terms of the average selling prices received by farmers. Data for EU countries have been downloaded, and on the basis of this data, a separate dataset is generated for the Republic of Ireland. This will be compared with other countries in the European Union at a later stage.

- (FAO - Agriculture Organization of the United Nations - https://www.fao.org)

I. Introduction

Based on the information obtained from FAO, it seems that the business problem being addressed in this project is to examine changes in agricultural producer prices and producer price indices over time in the European Union, with a focus on the Republic of Ireland. The goal is to identify trends and patterns in these prices and indices, and potentially to use this information to inform policy or decision-making related to agriculture in the EU.

One potential approach to solving this problem using the CRISP-DM framework would be to:

There are several reasons why CRISP-DM has been chosen over other PM Data science frameworks:

- CRISP-DM is a widely used and well-established framework that has been around for over two decades. It has been tested and refined through years of practical use and is widely recognized as a reliable and effective approach to data mining.
- CRISP-DM is a flexible framework that can be adapted to fit the needs of a wide range of projects. It is not tied to any specific industry or technology and can be applied to a variety of business contexts.
- CRISP-DM is a comprehensive framework that covers all the key steps in the data mining process, from business understanding and data preparation to modeling, evaluation, and deployment. This makes it well-suited for projects that require a structured and systematic approach.

- CRISP-DM is easy to understand and use, with clear guidance and best practices for each step in the process. This makes it accessible to practitioners who are new to data mining, as well as those who are more experienced.
- Overall, the CRISP-DM framework is a solid choice for data mining projects due to its widespread adoption, flexibility, comprehensiveness, and ease of use.

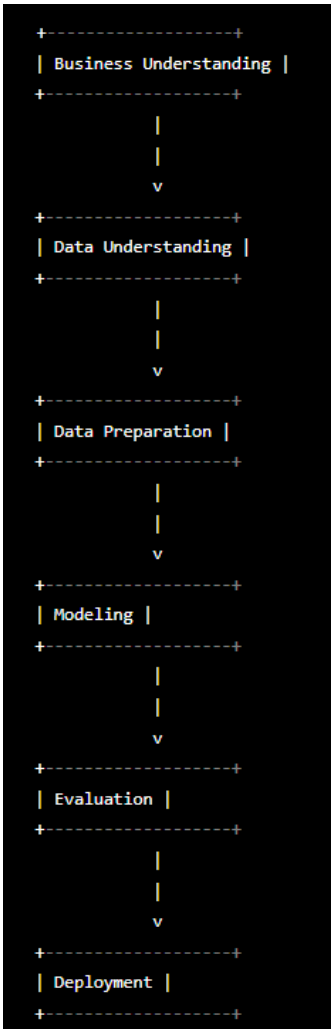
- **Overview of CRISP DM process**

The steps in the CRISP-DM process are as follows:

- **Business Understanding:** Define the project goals and objectives, as well as the stakeholders who will be affected by the project.
- **Data Understanding:** Collect and summarize data and assess the quality of the data.
- **Data Preparation:** Clean and transform the data to make it ready for analysis.
- **Modeling:** Select and apply the appropriate data mining techniques to build the model.
- **Evaluation:** Evaluate the results of the model and determine whether it meets the project goals and objectives.
- **Deployment:** Plan for the deployment of the model and prepare for the ongoing maintenance of the model.

Here is a flowchart that illustrates the CRISP-DM process in the CA-2:

Fig 1:



Business Understanding

In this project, the business problem being addressed is the examination of changes in agricultural producer prices and producer price indices over time in the European Union, with a focus on the Republic of Ireland. The goal of the project is to identify trends and patterns in these prices and indices, and potentially use this information to inform policy or decision-making related to agriculture in the EU.

Initial Steps

This step involves setting up the GitHub environment that I will be using as part of my development process. It has been connected to my git repository, and the assignment has been sent to my git hub account, which has been pushed to my git repository. My GitHub repository can be found at the following link, if you would like to check it out:

https://github.com/sba22251/CA2/blob/main/CA_2-Alex_Ciaica-SBA22251.ipynb

Data Understanding

For this study, I have downloaded the following datasets from the FAO website:

Load data

```
df1 = pd.read_csv('Producer Prices.csv')
df2 = pd.read_csv('Consumer Price Indices.csv')
df3 = pd.read_csv('Temperature_Change.csv')
df4 = pd.read_csv('Value_Added_Agri_Frs_Fishing.csv')
ire_df1 = df1.loc[df1['Area'] == 'Ireland']  # based on df1 ('Producer Prices.csv')
```

As soon as the data has been imported, I begin exploring the dataset to identify specific trends. During the course of the process of building the machine learning model, I checked the dataset for null values, duplicate values, and I checked each feature for the type of data it contained. That was because we needed to know which features to use when building the machine learning model.

According to my analysis in my Jupyter notebook, there were no null values in the data and there were no duplicates in the data.

In the next step, I'm going to use the "idxmax" function in order to find the maximum value in the series for the European countries. Based on the results, I discovered that the highest value of game meat, fresh, chilled or frozen, was produced in Cyprus in 2013 at a value of 29243,2 million dollars.

In a similar manner, I perform the same function to determine the minimum value for European countries.

Maximum Value per Europe:

Domain Code	PP
Domain	Producer Prices
Area Code (M49)	196
Area	Cyprus
Element Code	5532
Element	Producer Price (USD/tonne)
Item Code (CPC)	21170.02

Minimum Value per Europe:

Domain Code	PP
Domain	Producer Prices
Area Code (M49)	642
Area	Romania
Element Code	5532
Element	Producer Price (USD/tonne)
Item Code (CPC)	1229.0

Then, I proceed to perform the same report for the Republic of Ireland. Screenshots attached:

Maximum Value in Ireland:

```
Domain Code                             PP
Domain                      Producer Prices
Area Code (M49)                         372
Area                                Ireland
Element Code                           5532
Element          Producer Price (USD/tonne)
Item Code (CPC)                      1270.0
Item                 Mushrooms and truffles
Year Code                              2007
Year                                   2007
Months Code                            7021
Months                          Annual value
Unit                                    USD
Value                                3642.2
Flag                                      A
Flag Description            Official figure
Name: 19533, dtype: object
```

Minimum Value in Ireland:

```
Domain Code                             PP
Domain                      Producer Prices
Area Code (M49)                         372
Area                                Ireland
Element Code                           5532
Element          Producer Price (USD/tonne)
Item Code (CPC)                      1801.0
Item                              Sugar beet
Year Code                              1999
Year                                   1999
Months Code                            7021
Months                          Annual value
Unit                                    USD
Value                                  51.4
Flag                                      A
Flag Description            Official figure
Name: 19626, dtype: object
```

According to our discoveries, Ireland received the maximum value per ton of mushrooms and truffles in 2007 at the price of $3642.2 per ton.

In comparison, Sugar Beet was valued at $51.4 per ton at the time of the minimum value in 1999.

Why this information could be important?

This information could be used to understand the relative value of different types of produce, how the value of different types of produce has changed over time, or how the value of a particular type of produce compares to the value of other types of produce. This information could also be useful for people who are involved in the production, trade, or sale of these types of produce, or for researchers who are studying the economic or market trends related to these types of produce.

Data Visualization

The plot includes three horizontal lines in red, which represent the "mean" and the standard deviations above and below the mean of the 'Value' column.

The horizontal lines added to the plot represent statistical measures of the 'Value' column.

The red line at the center of the plot represents the mean of the 'Value' column, which is the average value of all the data points.

The red lines above and below the mean represent the standard deviations above and below the mean. This represents the measure of the spread or dispersion of a set of data. It can be used to identify data points that are significantly different from the mean.

Fig 1: Mean & Standart Deviation Value of trends in data over time

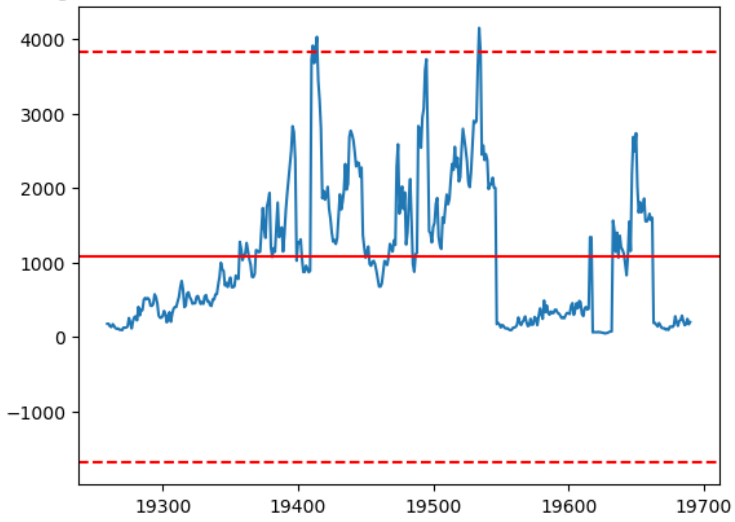
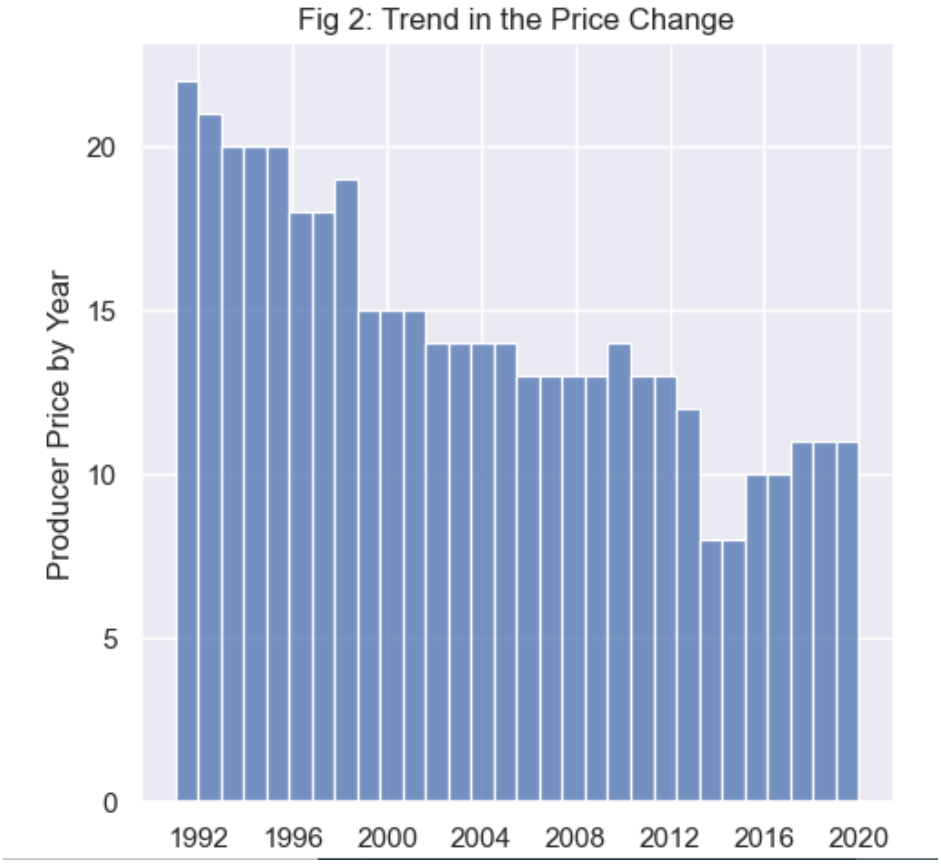


Fig 2: Represents the distribution of the Items respective to the 'Year' column and identifies patterns in trends of the data over time.



The data being plotted is from a dataframe called df1, and the x-axis of the plot is the 'Year' column of the dataframe. The plot is using 30 bins, which means that the data is divided into 30 intervals. The x-axis label is 'Year', the y-axis label is 'Producer Price by Year', and the title of the plot is 'Fig 2: Trend in the Price Change'. It is also setting the theme of the plot to 'darkgrid' using the set_theme function.

As a result, this plot can be used to identify relationships between variables as well as to detect anomalies or unexpected results that might occur.

Fig 3: This plot represents the mean producer value in Europe over each of the years in which the calculation was made.

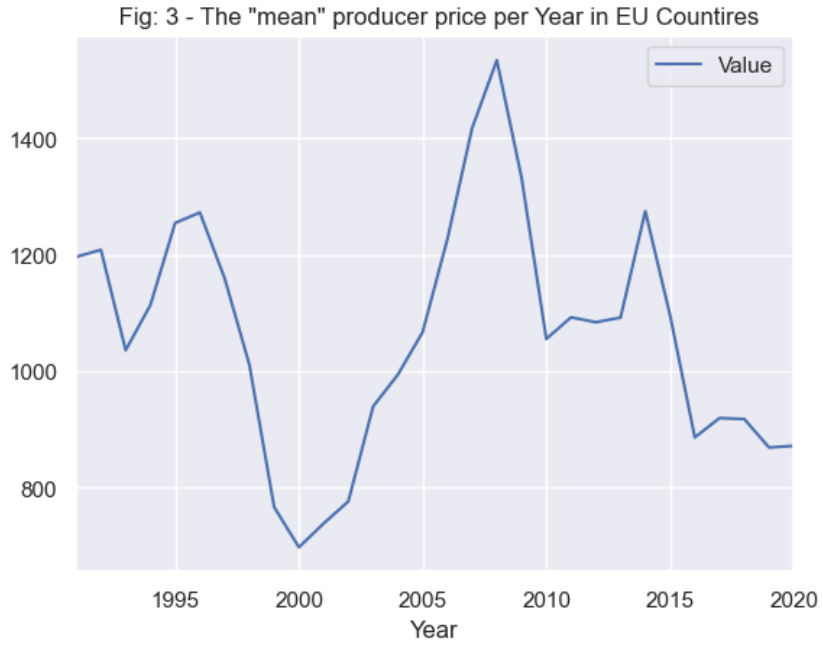
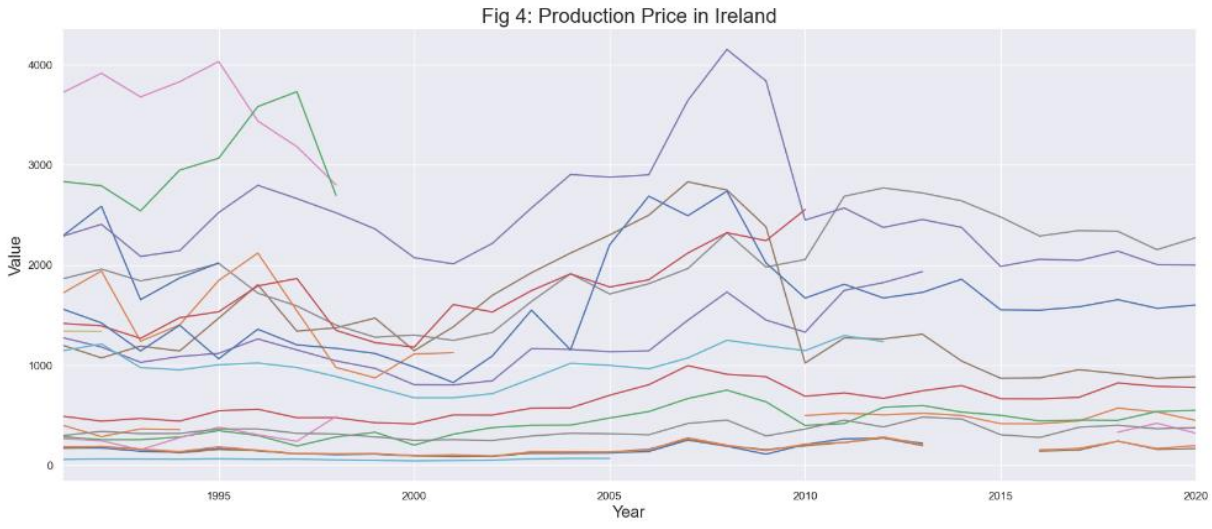


Fig4. Production Price in Ireland



This information in these graphs could be useful for understanding the average price that producers in European countries are receiving for their products, which could be useful for people who are involved in the production, trade, or sale of these products. It could also be useful for understanding the economic conditions or trends in European countries, or for comparing the prices of different products in different countries. Additionally, this information could be useful for researchers who are studying economic or market trends related to these products, or for policymakers who are making decisions related to the production or trade of these products.

Furthermore, there are a few more potential reasons why it could be important to know the mean value of the producer price per product in a given year for European countries:

- To inform business decisions: If you are a producer of a particular product, knowing the average price that producers in European countries are receiving for that product could help you make decisions about pricing, production, or marketing strategies.
- To track economic trends: Knowing the average price of a product over time can help you understand how the demand for that product is changing, and how it is being affected by economic trends or other factors.
- To compare prices across countries: By comparing the average price of a product in different countries, you can get a sense of how the price of the product compares across different markets.
- To inform policy decisions: Policymakers who are responsible for making decisions related to the production or trade of a particular product may find it useful to have information about the average price of the product in different countries. This could help them understand the economic implications of different policy options.

Statistical Analysis

Calculation the mean, standard deviation, and variance

The next step is to calculate the **mean, standard deviation, and variance**. What is the purpose of such a calculation?

In order to understand and interpret data, identify patterns and trends, and make statistical inferences, it is useful to calculate the mean, standard deviation, and variance of the data.

From the result I obtained the following results:

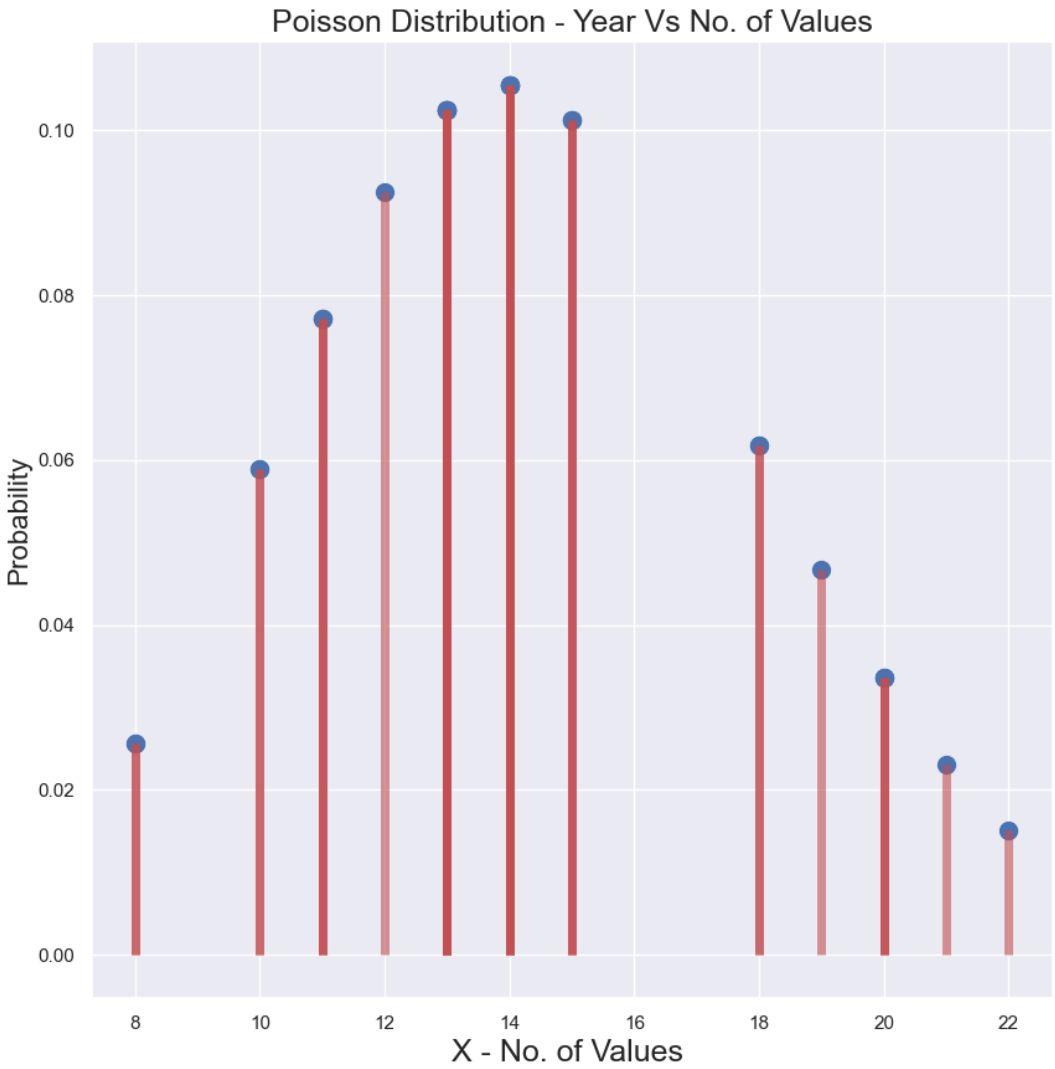
```
Total number of Values :
36840
Average number of Values each year :
1228.0
The standard deviations of the Value is :
114.56514484690841
The Varience of the Value is :
13125.172413793105
```

Calculation of the contingency table using chi2 method:

Next I perform the Calculation of the contingency table using chi2 method. This method can be used to calculate the contingency table by calculating the chi-square statistic, the p-value, the degrees of freedom, and the expected frequencies. Based on these values, it can be determined whether the two categorical variables are significant. You can conclude that there is a significant association if the p-value is lower than the significance level (usually 0.05).

In our case the (p) value is 1 therefore we fail to reject the Null Hypothesis and Null Hypothesis stays.

Calculation of Poisson Distribution



Calculating the Poisson distribution can be useful for predicting the likelihood of future events, understanding the distribution of events over time, and modeling real-world phenomena.

It is important to perform this calculation in order to predict the likelihood of future events occurring: By calculating the Poisson distribution, you can determine the probability that a certain number of events will occur in the future.

Performing Anova test

Performing an Anova test can be important for comparing the means of different groups, determining the relationship between variables, and testing hypotheses. Here we get our result for Anova test.

```
In [55]: oneway.anova_oneway((df1['Year'],df1['Value'],df1['Item Code']), use_var='equal')

Out[55]: <class 'statsmodels.stats.base.HolderTuple'>
         statistic = 59.60909623711433
         pvalue = 1.72556542882542e-25
         df = (2.0, 1293.0)
         df_num = 2.0
         df_denom = 1293.0
         nobs_t = 1296.0
         n_groups = 3
         means = array([ 2003.42592593,  1071.30347222, 24467.10602083])
         nobs = array([432., 432., 432.])
         vars_ = array([7.46719945e+01, 8.50253224e+05, 3.81426304e+09])
         use_var = 'equal'
         welch_correction = True
         tuple = (59.60909623711433, 1.72556542882542e-25)
```

Forecasting with ARIMA

After performing Arima test we Reject the Null Hypothesis and we confirm that our data is Stationary.

Forecasting with ARIMA can be extremely important for making predictions about the future, understanding patterns and trends in data, and informing decision-making processes. Here are the result of our test.

```
adfuller_test(df1['Value'])
```

```
ADF Test statistics : -3.6266230264038213
p-value : 0.005271982397588646
lags_used : 2
nobs : 429
Reject the Ho, data is stationary
```

Calculation of the IQR value

The next step in our Statistical analysis is the Calculation of the IQR values.

Calculation of the IQR value by subtracting the 25th percentile value from the 75th percentile value.

We need to perform the IQR calculation in order to identify and remove outliers from a dataset. First we mark the outlier as Nan then we remove the Nan by using the function dropNa.

Feature Encoding

I then performed a feature encoding process. Machine learning algorithms utilize feature encoding to convert categorical variables (variables with a limited number of values) into numerical form. This is an important step in preparing data for machine learning, and it is often required since most machine learning algorithms are designed to work with numerical data rather than categorical data.

Feature encoding is important for several reasons:

- It is difficult for many machine learning algorithms to process categorical data directly. In order to use these algorithms, categorical data is encoded into numerical form.
- It is possible for some machine learning algorithms to perform better when numerical data is provided rather than categorical data. Random forests and decision trees are often capable of handling categorical data directly, but may perform better with numerical data.
- The feature encoding method can also be used to reduce the dimensionality of the data by encoding multiple categorical variables into a single numerical feature. Data can be made more manageable by reducing its complexity.
- In general, feature encoding is important because it facilitates the use of categorical data in machine learning algorithms, enhances the performance of the algorithms, and reduces the dimensionality of the data.

The Feature Encoding gives me the following features that I can use to perform machine learning:

```
array([[nan, 429.99999999999994],
       [nan, 429.99999999999994],
       [nan, 429.00000000000017],
       [nan, 429.00000000000017],
       ['Country Code', 427.0000000000001],
       ['Element', 427.0000000000001],
       ['Item', 427.0000000000001],
       ['Months', 427.0000000000001],
       [nan, 427.0000000000001],
       [nan, 427.0000000000001]], dtype=object)
```

Cross-Validation Model

- Our results show that we have **286** entries that we can TRAIN these predictions
- Our results show that we have 143 entries that we can Test

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=0)
```

```
print(X_train.shape)
```

```
(289, 9)
```

Our results show that we have 286 entries that we can TRAIN these predictions

```
print(X_test.shape)
```

```
(143, 9)
```

Our results show that we have 143 entries that we can Test

Lazypredict Lybrary

In this step I decide to use a library called lazypredict . This library can be helpful for machine learning practitioners because it can save time and effort by automating certain tasks and allowing them to focus on other aspects of the modeling process. However, it's always a good idea to carefully evaluate whether a particular library is suitable for your needs and to consider the trade-offs of using a library versus implementing certain functionality yourself.

Although it’s a good library I haven’t received great results.

```
clf = LazyClassifier(verbose=0,
                     ignore_warnings=True,
                     custom_metric=None)
models, predictions = clf.fit(xtrain, xtest, ytrain, ytest)
models
```

```
100%|██████████████████████████████████████████████████████████| 29/29 [07:42<00:00, 15.96s/it]
```

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
BaggingClassifier	0.09	0.08	None	0.09	1.29
NearestCentroid	0.05	0.08	None	0.03	0.11
RandomForestClassifier	0.09	0.08	None	0.09	10.70
ExtraTreesClassifier	0.08	0.08	None	0.08	10.50
DecisionTreeClassifier	0.09	0.07	None	0.09	0.25
ExtraTreeClassifier	0.08	0.07	None	0.08	0.13
LGBMClassifier	0.08	0.07	None	0.07	9.75

Random Forest

Since we shall use a random forest regressor during our random search implementation, it is of value to introduce random forests. Random forests refer to an ensemble of untrained decision trees capable of both regression and classification tasks.

These methods involve the use of bagging, which combines many models in order to provide a generalized result.

As a first step, samples of training observations are taken at random when building trees. In addition, random subsets of features are used when splitting nodes.

There are many parameters that are used in a random forest. I have listed the most important ones below.

- **n_estimators.** In an ensemble/forest, this parameter indicates the maximum number of trees.
- **max_features.** When splitting a node, this represents the maximum number of features taken into consideration.
- **max_depth. max_depth** represents the maximum number of levels that are allowed in each decision tree.
- **min_samples_split.** There must be a minimum number of samples in a node for it to split. This minimum number of data points is what is represented by to as min_samples_split.
- **min_samples_leaf.** Data points that can be stored in a leaf node at a minimum.
- **bootstrap**

Bootstrap sampling is used to sample data points. Sampling may be carried out with or without replacement. Sampling with replacement can be described as when a sample is selected from a random population, then returned to the population. If bootstrap = True, sampling is carried out randomly with replacement. If bootstrap = False, sampling is without replacement

The end

REFERENCES:

- DUBLIN CITY COUNCIL. (2022) DCC_GULLYREPAIRSPENDING2004-11ALL: [ONLINE] AVAILABLE FROM: https://data.gov.ie/dataset/drainage-gully-cleaning-programme/resource/a7dcf386-55b8-4809-9487-c846b6fef145 [LAST UPDATED JANUARY 22, 2016].
- MEIDIUM.COM:(https://medium.com/fintechexplained/ever-wondered-why-normal-distribution-is-so-important-110a482abee3)
- DATACAMP.COM (RANDOM FORESTS (RF) https://campus.datacamp.com/courses/machine-learning-with-tree-based-models-in-python/bagging-and-random-forests?ex=7

- Udemy.com: (Created by: Jose Portilla)Python for Data Science and Machine Learning Bootcamp https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/learn/lecture/5733492#overview [Last updated 5/2020]
- medium.com: (Farhad Malik) Ever Wondered Why Normal Distribution Is So Important?: https://medium.com/fintechexplained/ever-wondered-why-normal-distribution-is-so-important-110a482abee3 [Jun 20, 2019]