**MSC\_DA\_CA2**

**Ireland Milk Production Comparison with the EU Countries**

**Mahender Kumar**

**SBA22212**

**Table of Contents**

[**1. Report Description** 3](#_Toc123739025)

[**2. Abstract** 4](#_Toc123739026)

[**3. Introduction** 4](#_Toc123739027)

[**4. Exploratory data analysis (EDA)** 5](#_Toc123739028)

[4.1 Dealing with missing data and data cleaning 7](#_Toc123739029)

[4.2 Extract countries data where human population is between +/- 1 million 8](#_Toc123739030)

[**5. Handling outliers** 8](#_Toc123739031)

[**6. Descriptive statistics** 11](#_Toc123739032)

[6.1 Measures of Central Tendency 11](#_Toc123739033)

[6.2 Variation Measures 13](#_Toc123739034)

[**7. Poisson distribution** 15](#_Toc123739035)

[**8. Normal Distribution** 17](#_Toc123739036)

[**9. Hypothesis Test** 20](#_Toc123739037)

[**10. Data visualisation** 20](#_Toc123739038)

[**11. Feature Engineering** 20](#_Toc123739039)

[**12. Dashboard** 21](#_Toc123739040)

[**13. Project Management Framework** 21](#_Toc123739041)

[**12. Supervised Learning** 22](#_Toc123739042)

[**14. Decision Trees** 23](#_Toc123739043)

[**15. Linear Regression** 25](#_Toc123739044)

[**16. Random Forest** 27](#_Toc123739045)

[**17. Cross-Validation with Linear Regression** 28](#_Toc123739046)

[**18. Machine Learning outcomes** 31](#_Toc123739047)

[**19. Sentiment Analysis** 33](#_Toc123739048)

[**20. Conclusion** 33](#_Toc123739049)

[**21. References** 33](#_Toc123739050)

[**22. Appendix** 34](#_Toc123739051)

[22.1 Gantt Chart 34](#_Toc123739052)

[22.2 Population in EU countries in 2020 35](#_Toc123739053)

**Table of Figures**

[Figure 01 Display first five rows of the livestock products data frame 5](#_Toc123661416)

[Figure 02 Display first five rows of the world population data frame 5](#_Toc123661417)

[Figure 03 Common methods to get the insight of data 5](#_Toc123661418)

[Figure 04 Merging two datasets 6](#_Toc123661419)

[Figure 05 null, nan, zero value in data frame 7](#_Toc123661420)

[Figure 06 Histogram plot to find outliers in Milk production value 8](#_Toc123661421)

[Figure 07 Boxplot to find out outliers in milk production value 10](#_Toc123661422)

[Figure 08 Mean Value 11](#_Toc123661423)

[Figure 09 Mode value 11](#_Toc123661424)

[Figure 10 Median Value 12](#_Toc123661425)

[Figure 11 Variation and Standard Deviation 13](#_Toc123661426)

[Figure 12 Box Plot Human population in Ireland Configuration 13](#_Toc123661427)

[Figure 13 Histogram Human population configuration in Ireland 14](#_Toc123661428)

[Figure 14 Normal Distribution 17](#_Toc123661429)

[Figure 15 CDF calculation 18](#_Toc123661430)

[Figure 16 SF Calculation 18](#_Toc123661431)

[Figure 17 Milk production comparison 19](#_Toc123661432)

[Figure 18 test view 20](#_Toc123661433)

[Figure 19 CRISP-DM Model Overview 20](#_Toc123661434)

[Figure 20 Decision Tree 24](#_Toc123661435)

# **1. Report Description**

Git hub repository address  
<https://github.com/Mahender1007/MSC_DA_ContAssignment2>

With Git version control, version control is much smoother and easier to implement. Using an online platform like GitHub to store my files means that I have an online backup of my work, which is beneficial for both my collaborators and I.

**Project outcome**To create a datasheet of dairy products over time (1960 to 2020) for Ireland and other EU nations. The data frame will be created from milk production, yield value, quantity of animals, total human population and farmer population.

**Data Description**This data, which is in a csv file format, consists of Year, Ireland and other European countries. These columns contain information about the columns mentioned from 1960 to 2020.

**How Was This Data Collected?**Crop and Food data was obtained from the Food and Agricultural Organisation of the United Nations. FAO collects the data from all the countries in the world. FAO is an international organisation and it is a part of the United Nations. I am using raw data to form the base of my data frames. FAO is a website that provides reliable data in various fields.For more detailed information, please click the link below. Link:<https://www.fao.org/faostat/en/#data/QCL>

Total yearly population data is collected from EUROSTAT which is freely available and there is no licensing requirement to access this data. This population is then compared to the other 27 countries in the EU. Four countries are identified as having a similar (+/- 1m people) population in 2020. The population data for these countries is downloaded, which can only be downloaded separately.  
For more detailed information, please click the link below. Link:  
<https://data.worldbank.org/indicator/SP.POP.TOTL?end=2020&start=1960>

**Is there a License for the Data Used?**The FAO and EUROSTAT websites where the data was obtained, has a Creative Commons Attribution 4.0 International license. This license type makes shared data publicly available according to open data standards and license datasets. Data is freely available and no licence is required to access or download the data.

**Colour choice for graphs**Colour helps to highlight the most important aspects of message and simplify complex graphs. I chose a blue colour for graphs because the blue colour sets a more neutral tone and tends to focus on strengthening the message within the graph.

# **2. Abstract**

This report examines Ireland and EU milk production data from 1960-2020. Data cleaning and pre-processing tasks were performed on the data set and 4 European countries were identified based on the human population where the population difference is +/- 1 million in year 2020.

Hypothesis tests were performed and machine learning processes were used to apply the various models and these models will be discussed throughout the report. This report is divided into 17 sections and these sections consist of the following steps; data cleaning and preparation, statistics and machine learning.

# **3. Introduction**

*Data explanation*The data set Milk production consists of data from all the countries in the world. Furthermore, data is collected from 2 other datasheets. Data is gathered over a period of 60 years (1960-2020). All data files are in the form of csv file extension. The file dimensions for milk production data set are 60 rows and each row contains 1 year of milk related data.

*Why choose data related to milk production*

Milk and dairy products play an important role in a healthy and balanced diet. They are rich sources of calcium which is easily absorbed by the body. This mineral, along with other nutrients present in dairy foods, such as protein, magnesium and phosphorus, is essential to build and maintain strong bones.

Dairy farming (milk production) is one of the most important sectors in the agriculture economy in Ireland. During the last ten years farms involved in commercial milk production often had substantial modernisation.. (Fig 17)

Exploratory data analysis (EDA) was used to analyse and investigate data sets. It was necessary to summarise the main characteristics of data sets using visualisation methods. Descriptive Analysis were used to analyse the data and it aided to describe, show and summarise the data points in a constructive way such as patterns that may emerge that fulfil every condition of the data. Finally, machine learning techniques used to analyse and make data-driven recommendations and decisions based on the input data (Google, 2022).

# **4. Exploratory data analysis (EDA)**

EDA is the first step in the data analysis process, it is an approach to analysing data sets to summarise their main characteristics. It allows to us check and analyse the data before we make any conclusion or assumption. EDA also ensures that the results produced are valid and applicable to the business needs (Wes McKinney 2017).

Following the benefits, we begin using EDA on the data sets.

* It helps to understand the dataset variable and the relationship among the data;
* It provides better knowledge of the data set;
* Helps to identify if there are any errors in our dataset. I.e., Duplicates, missing data; and
* Helps detect outliers or anomalies

In my project, data is collected from 3 datasheets. These datasheets consist of milk production data from all around the world. Data is gathered over a period of 60 years (1960-2020).

The following are 3 types of datasets.

1. Livestock products;
2. World population; and
3. World population in farming

After loading the data from csv files into the data frame using Pandas library, simple tasks were performed to get an insight of data.

Initially, I worked on following the two data sets  
a. lp\_df (Livestock products)

b. wpt\_df (World population)

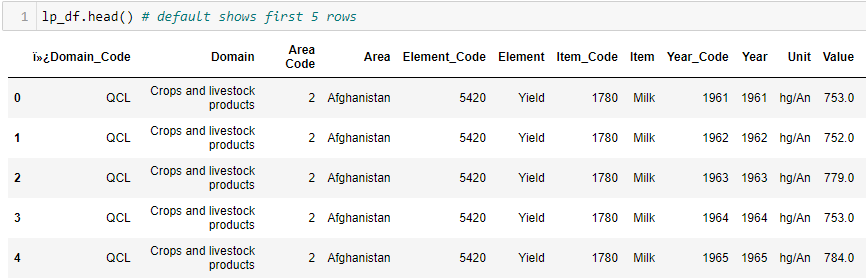


Figure 01 Display first five rows of the livestock products data frame

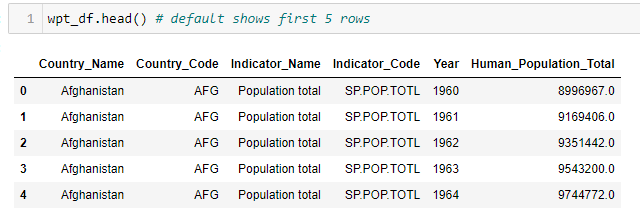


Figure 02 Display first five rows of the world population data frame

I also performed the following methods on the DataFrame to get more information

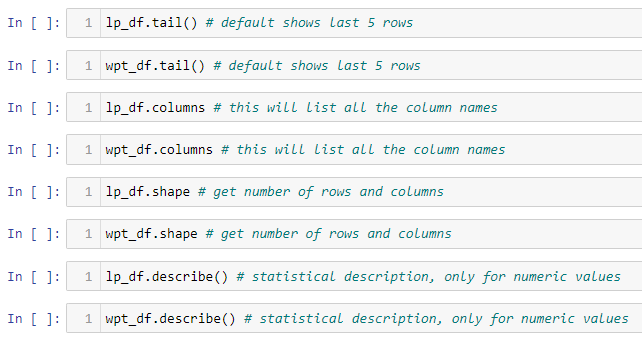


Figure 03 Common methods to get the insight of data

I also did some cleaning of data fin data frame lp\_df (Livestock products). I merged all unique “Element” rows into columns. [Please see section 2.0.3 in jupyter notebook]

*Merging the data frames*

Data frame lp\_df (Livestock products) and wpt\_df (World population) were merged into one dataframe (Please see section 2.0.4 in jupyter notebook)

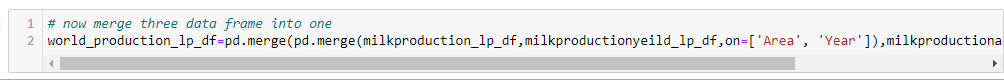


Figure 04 Merging two datasets

*Extraction of European country data from the main data frame*

This project evaluates milk production only in European countries so I need to extract European countries data from the main data frame world\_production\_inc\_lp\_df. Library countrygroups was used to get the list of countries in EU union. (Please see section 2.0.5 in jupyter notebook)

## 4.1 Dealing with missing data and data cleaning

One of the most common issues in dataset are missing values. Data can either be missing during data extraction or collection due to corruption or data not being recorded. I checked both datasheets for missing and empty(zero) data.

Below are some formats that could be in the missing data:

* n/a
* NA
* —
* na
* NaN

Why is it important to deal with missing data?

Missing data is important because, depending on the type, they can sometimes bias your results. This means results may not be generalisable outside of study because data comes from an unrepresentative sample.

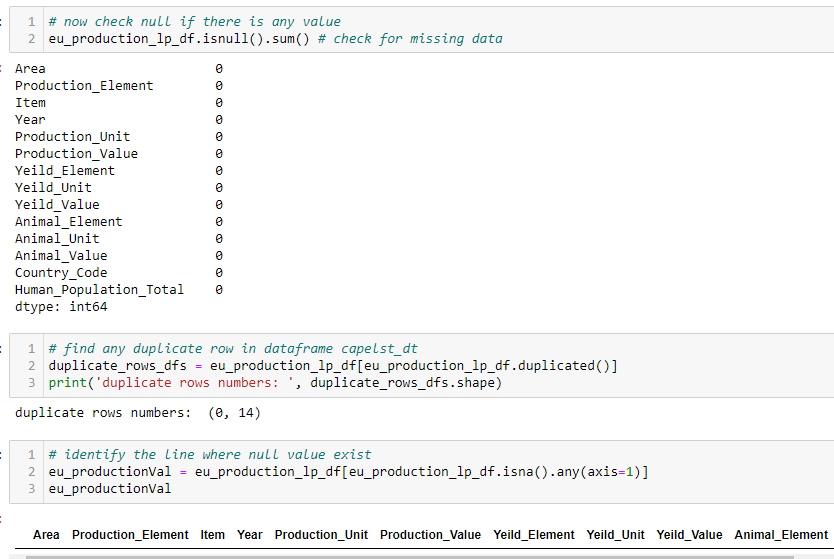


Figure 05 null, nan, zero value in data frame

From the above (see figure 5) observation I found that there are no missing or null value in the data frame. However, if there are missing or null values, they can be fixed by deleting the row or replacing it by mean, mode or with the other values. If the data base is small then it is best to replace the null value instead deleting it. Mean values can be used to replace the missing value columns if the data distribution is symmetric.

## 4.2 Extract countries data where human population is between +/- 1 million

I also analysed the dataset and extracted the data where countries population are +/- 1 million in year 2020. (Appendix number 22.2 to see the bar graph)

Upon researching, I identified the following 4 countries where the human population is similar to that of Ireland’s human population..

1. Slovakia
2. Denmark
3. Croatia
4. Finland

# **5. Handling outliers**

This is one of the most important parts of the data pre-processing and treating the outliers, as they can negatively affect the statistical analysis resulting in inaccuracy. The main reason they occur is due to variability in the data, or human error. It is good practice to detect and remove these outliers from the data as it could also give biased results. The simplest way to detect an outlier is by graphing the features or the data points. Visualisation is one of the best and easiest ways to have an inference about the overall data and the outliers. Scatter plots and box plots are the most preferred visualisation tools to detect outliers. The other method to detect outliers is using IQR (Interquartile Range) techniques (Chun-houh, Wolfgang, and Antony 2008).

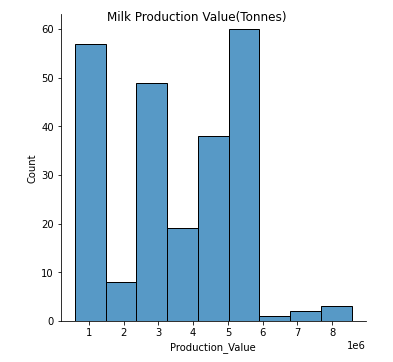
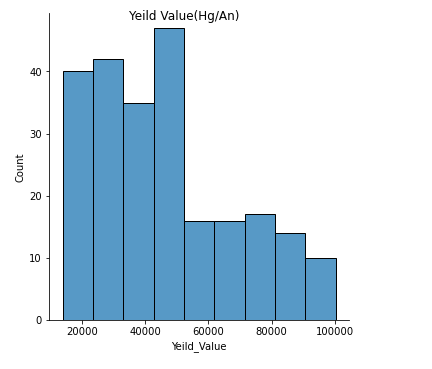
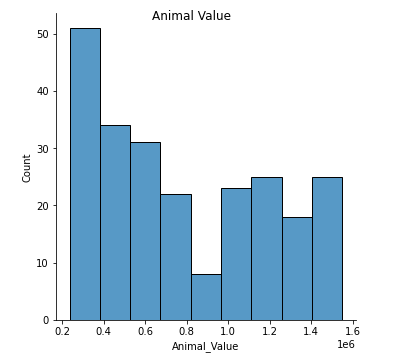
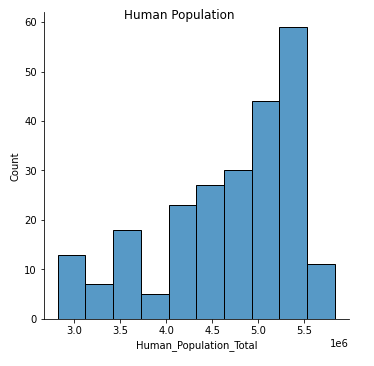


Figure 06 Histogram plot to find outliers in Milk production value







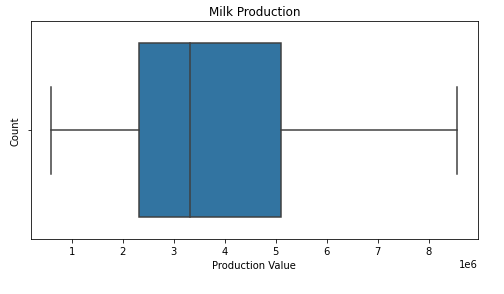


Figure 07 Boxplot to find out outliers in milk production value

After visualising and analysing the data frames, (See figure 6 and 7) I noticed there were no outliers in the data frame. Luckily, the data sets I am working with have no anomalies and I did not have to take any further steps. If there are outliers in the datasets, the following steps can be taken to handle them.

1. Remove or change outliers during post-test analysis;and
2. Change the value of outliers.

# 

# **6. Descriptive statistics**

Descriptive statistics were used to summarise and describe the characteristics of the data set. It was important to use these descriptive statistics as they can be useful for two purposes:   
a. to provide basic information about variables in a dataset; and   
b. to highlight potential relationships between variables

## 6.1 Measures of Central Tendency

*Mean*This is a typical average in the dataset column. The following formula is used to calculate the mean value;

\bar{x} = \dfrac{\sum x}{n}

Mean value is calculated in Ireland’s data frame Ireland\_production\_inc\_lp\_df on column Production\_Value

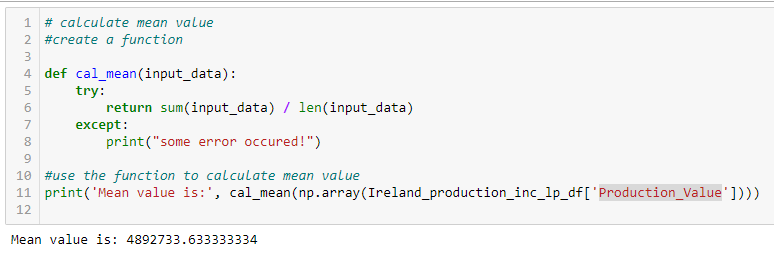


Figure 08 Mean Value

**Mean value can be used to replace null or empty columns in the dataset**

*Mode*The mode value of a data set is the most frequently occurring value. It visualises the most popular choice or the most common characteristic in the datasets.

Mode value is calculated in data frame Ireland\_production\_inc\_lp\_df on column Production\_Value

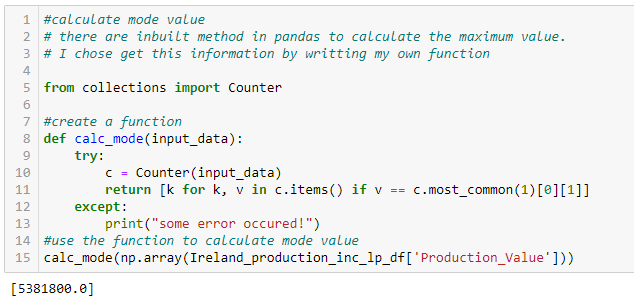


Figure 09 Mode value

**Mode value can be used to replace null or empty columns in dataset**

*Median*Median represents the middle value for the dataset. It is the point at which half of the data is more and half the data is less. Median helps to represent a large number of data points with a single data point.

Median value is calculated in the data frame Ireland\_production\_inc\_lp\_df on column Production\_Value



Figure 10 Median Value

**Median value can be used to replace null or empty columns in the dataset**

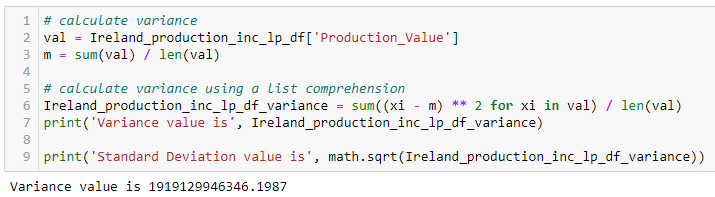
## 

## 6.2 Variation Measures

*Variance*The variance is a measure of variability. It is calculated by taking the average of squared deviations from the mean. Variance explains the degree of spread in the data set. The more spread the data is, the larger the variance is in relation to the mean.

s^2= \dfrac{\sum (X - \bar{x})^2}{n - 1}

Variance value is calculated in data frame Ireland\_production\_inc\_lp\_df on column Production\_Value



*Standard Deviation*Standard deviation is the square root of the variance. The Standard Deviation is a measure of how spread-out the numbers are.  
Standard Deviation value is calculated in the data frame Ireland\_production\_inc\_lp\_df on column Production\_Value

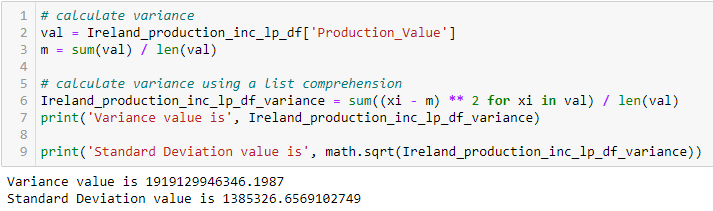


Figure 11 Variation and Standard Deviation

*Box Plot*I used Box and Whisker or Box Plot to statistically represent a column in the data set. A Box Plot is the visual representation of the statistical five number summary of a given data set.

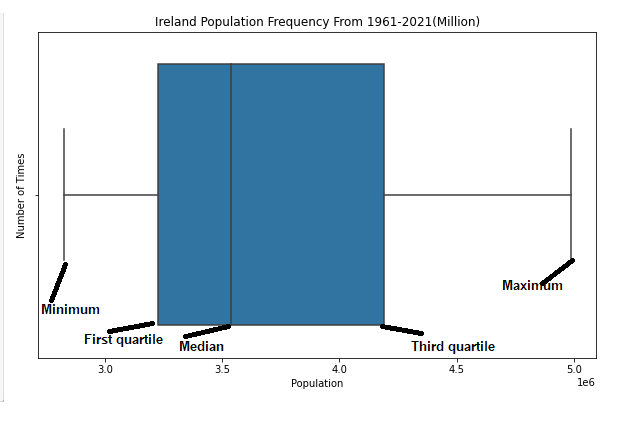


Figure 12 Box Plot Human population in Ireland Configuration

The following are the dimensions:

1. Minimum
2. First Quartile
3. Median
4. Third Quartile
5. Maximum

Box plots portray the distribution of data, outliers, and the median. The box within the chart displays where approximately 50 percent of the data points fall. As shown in fig 12 it summarises a data set in five marks. The mark with the greatest value is called the maximum, and it will likely fall far outside the box. The mark with the lowest value is called the minimum, and it will likely fall outside the box on the opposite side of the maximum.

The box itself contains the first or lower quartile, the third or upper quartile, and the median in the centre. The median is the value separating the higher half from the lower half from the data frame, where a total number of pedestrians crossed in a day throughout the year. The median value shown can also be deemed as "the middle" value in a set of numbers based on a count of column values rather than the middle based on a numeric value. The sections in a box plot help the viewer to see where the median falls within the distribution. The lower quartile is the 25th percentile, while the upper quartile is the 75th percentile. The median is the middle, but it helps give a better sense of what to expect from these measurements. The whiskers (the lines extending from the box on both sides) typically extend to 1.5\* the Interquartile Range (the box) to set a boundary beyond which would be considered outliers. The outliers are individual dots that occur outside the upper and lower extremes (Chun-houh, Wolfgang, and Antony 2008).

*Histogram Plot*A histogram is a graph that shows the frequency of numerical data using rectangles. The height of a rectangle (the vertical axis) represents the distribution frequency of a variable (how often that variable appears). The width of the rectangle (horizontal axis) represents the value of the variable (Total Human Population in Millions).

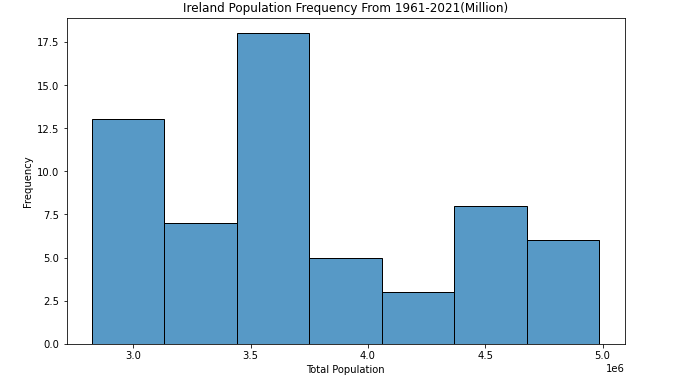


Figure 13 Histogram Human population configuration in Ireland

Histograms were useful to statistically represent the data as there was a need to display a comparison of the highest population in a year and the number of times it occurred in Ireland. In fig 13, it can be observed that there was a skewed distribution of data (Google 2022)

The anatomy of a histogram fig 13  
It was a visual representation of a data set which show how often each value in the data set Ireland\_production\_inc\_lp\_df on column Human\_Population\_Total occurs. The values are grouped into bins along the x-axis. The height of the bar indicates how many values of the data set fell into that bin. In above fig 13, a histogram illustrates the total of Ireland’s population values from a data set of Ireland’s milk productions. These 60 row values are sorted into 7 bins; the first bin includes the total population count that occurred which is 2.5 to 3.1 million; the second and third bin includes the total population count that occurred which are 3.1 to 3.4 million and so on and so forth. The height of the blue bar indicates the monotonal population crossed in a year, along in each bin. There were more than 17 times when 3.5 million or more human population observed in year.

# **7. Poisson distribution**

A Poisson distribution is a discrete probability distribution, it gives the probability of a discrete (i.e., countable) outcome. A Poisson distribution is a great tool that helps to predict the probability of certain events happening when you know how often the event has occurred. It gives us the probability of a given number of events happening in a fixed interval of time.

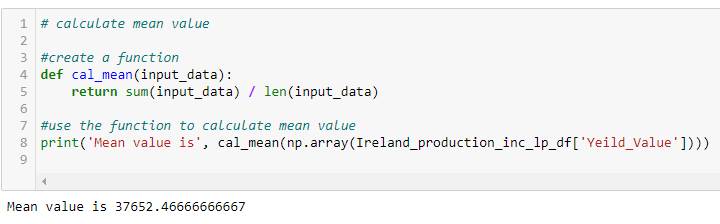
I used Poisson distribution to predict the yield value (hectogram/animal) in Ireland within a given interval of time. In my dataset, individual counts occurred at random and independently. That is, that the probability of one event was not affecting the probability of another event. (Please see jupyter notebook section for 3.0.2)

Furthermore, I was aware of the mean number of events occurring within 60 years and this number is called λ (lambda), and it was a constant number.

The following is the formula to calculate the Poisson distribution.



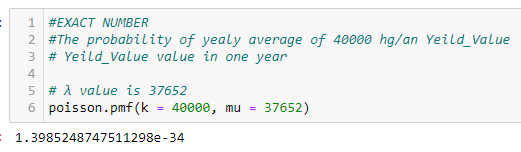
*Calculating mean value*

In my data sets column λ value is 37652

*The below criteria were tested*

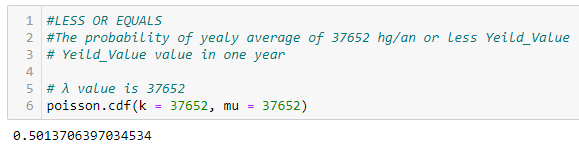
1. The probability of the yearly average of 40000 hg/an yield value

*Answer: 1.39*



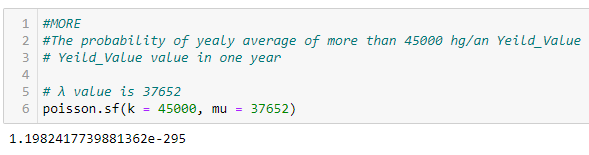
1. The probability of the yearly average of 40000 hg/an or less yield value

*Answer: 0.50*



1. The probability of the yearly average of more than 40000 hg/an yield value

*Answer: 1.19*



# **8. Normal Distribution**

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean isoccurring more frequently than data that is further from the mean. The normal distribution is easy to work with mathematically.

I implemented normal distribution on data frame Ireland\_production\_inc\_lp\_df on column Production\_Value. Production value is calculated in tonnes. This data refers to milk related values from 1961-2021. (Jupyter notebook section 3.0.3)

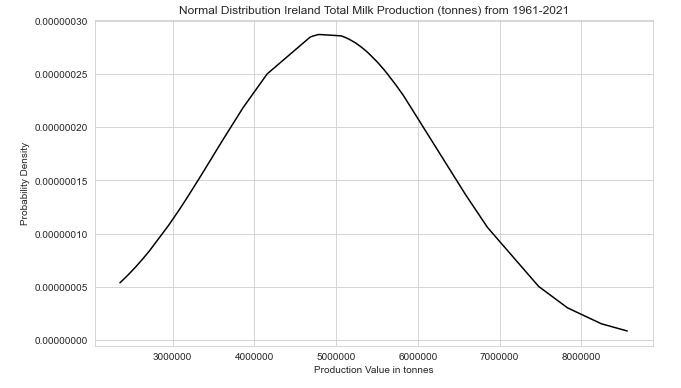
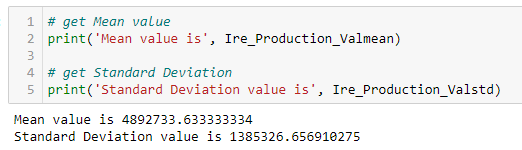


Figure 14 Normal Distribution



Mean value is 4892733 and the Standard deviation is 1385326.

The normal distribution density function simply accepts a data point along with a mean value and a standard deviation, and throws a value which we call probability density. The shape can be altered by changing the mean and the standard deviation. Changing the mean will shift the curve towards that mean value, this means we can change the position of the curve by altering the mean value while the shape of the curve remains intact. The shape of the curve can be controlled by the value of Standard deviation. A smaller standard deviation will result in a closely bounded curve while a high value will result in a more spread-out curve.

*Some properties of a normal distribution:*

1. The mean, mode, and median are all equal.
2. The curve is symmetric around the mean.

To find the probability of a value occurring within a range in a normal distribution, I needed to find the area under the curve in that range. i.e., I need to integrate the density function. Since the normal distribution is a continuous distribution, the area under the curve represents the probabilities. A standard normal distribution is just similar to a normal distribution with mean = 0 and standard deviation = 1.

Following is the formula to calculate the z value;



The z value above is also known as a z-score. A z-score gives you an idea of how far from the mean a data point is.

1. To find the probability of a total number of people that visited the streets that has a value of less than or equal to 4000000. (Using Cumulative Density Function)

X = 4000000  
μ = 4892733  
σ = 1385326

*Answer:*To calculate this probability I firstly needed to follow standardisation process. This process will calculate the requested data points.

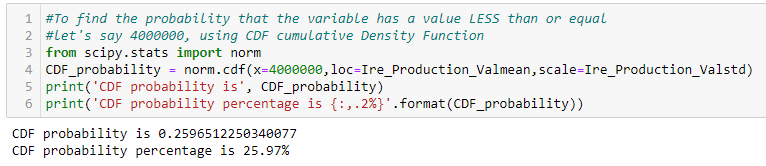


Figure 15 CDF calculation

As shown in fig 15, the probability of the milk production in one year that has a value of less than or equal to 4000000 is 25.97%.

1. To find the probability of the total number of people visited the streets that has a value of greater than or equal to 500000. (Using Survival Function)

X = 5000000  
μ = 4892733  
σ = 1385326

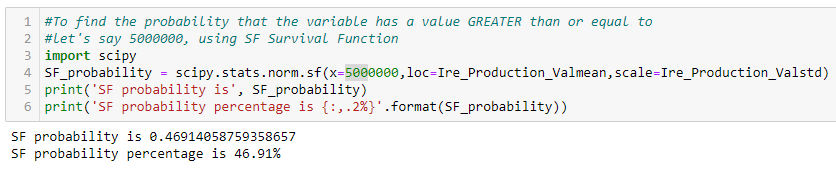
*Answer:*  


Figure 16 SF Calculation

As shown in fig 16, the probability of the milk production in one year that has a value of greater than or equal to 5000000 is 46.91%.

# **9. Hypothesis Test**

Add stuff here

# **10. Data visualisation**

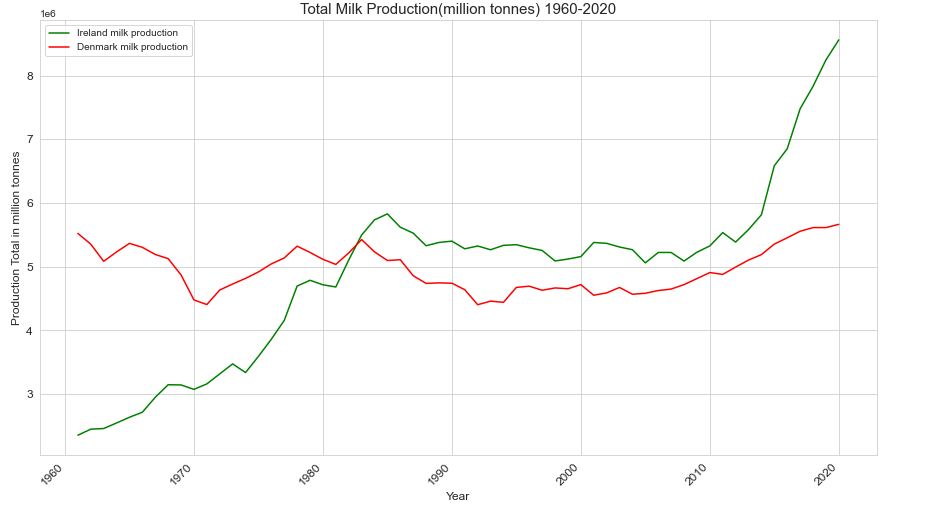


Figure 17 Milk production comparison

I chose line graphs to visualise data for the milk production comparison between Ireland and Denmark. Line graphs are used to compare different sets of data between different groups or to track changes over time. However, when trying to measure change over time, line graphs are the best when the changes are larger. (Jupyter notebook section 4.0.1)

Milk production data is processed from the data frame selected\_eu\_production\_lp\_df, this data frame consists of data from EU countries over 60 years. Upon observation it is noticeable that milk production in Ireland was very low in 1960. In fig 17, it is shown that there is a continuous increase from 1960 to around 1985 and then production declined due to the introduction of milk quota in Europe (<https://ec.europa.eu/commission/presscorner/detail/en/MEMO_15_4697>).

Ireland surpassed Denmark in production in 1982 and since then Ireland has produced more milk when it is compared with Denmark. It is also noticeable in fig 17 that the production of milk in Ireland increased very fast since 2012 due to abolition of milk quota in Europe (<https://ec.europa.eu/commission/presscorner/detail/en/MEMO_15_4697>).

# **11. Feature Engineering**

To add stuff here

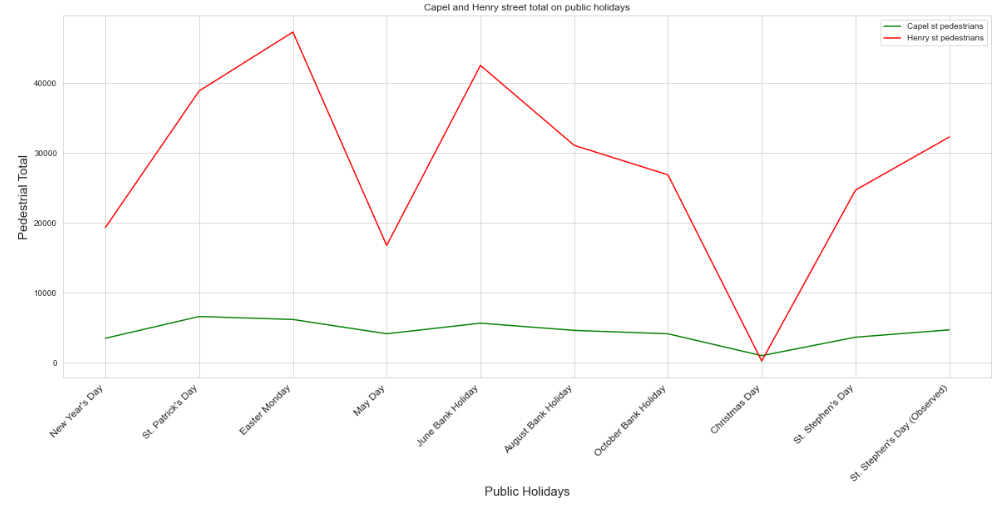


Figure 18 test view

# **12. Dashboard**

Add stuff here

# **13. Project Management Framework**

There are several project management frameworks that can be applied to modern Data Science projects. Some of the examples are CRISP-DM, SEMMA and KDD. I have used CRISP-DM methodology in my project.

CRISP-DM

CRISP-DM stands for Cross-Industry Standard Process for Data mining. This methodology was developed in IBM for Data Mining tasks.

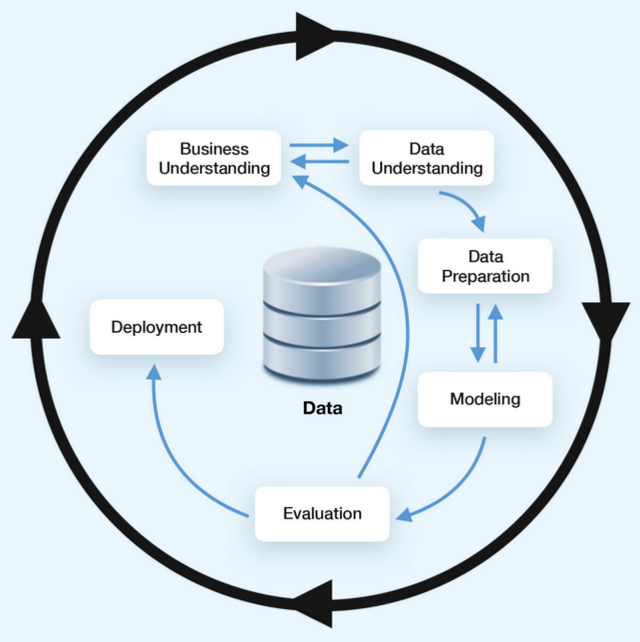


Figure 19 CRISP-DM Model Overview

CRISP-DM is divided into 6 phases:

a. Business Understanding

b. Data Understanding

c. Data Preparation

d. Modeling

e. Evaluation

f. Deployment

The main advantage of choosing CRISP-DM is due to it being a cross-industry standard. It means that this methodology can be implemented in any data science project notwithstanding its domain or destination.

**Flexibility**This is another feature CRISP-DM offers to its users. At the beginning of a project, it can suffer pitfalls and mistakes which are crucial, and it needs to be rectified at the beginning. When starting a project, there could be an issue of the lack of domain knowledge or ineffective models of data evaluation. Thus, a project can become successful only if project model has an option to reconfigure its strategy and is able to improve technical processes it applies. Another advantage of CRISP-DM approach is its flexibility. This makes it possible for models and processes to be imperfect at the very beginning. It provides a high level of flexibility that helps improve hypotheses and data analysis methods in a regular manner during further iterations.

**Functional Templates**Functional templates are another benefit of using a CRISP-DM framework. There is a possibility to develop functional templates for data science management processes. In order to take as many benefits as possible from CRISP-DM implementation is to create strict checklists for all phases of the work.

**Long-term Strategy**CRISP-DM frameworks allows for creation of a long-term strategy based on short iterations at the beginning of the project development. During first iterations, this model helps to create a basic and simple model cycle that can easily be improved in further iterations. This principle allows to perfect a preliminarily developed strategy after obtaining additional information and insights. (Perez 2021)

# **12. Supervised Learning**

Supervised learning is a machine learning method in which models are trained using labelled data. In supervised learning, models need to find the mapping function to map the input variable (X) with the output variable (Y).

Supervised Machine learning

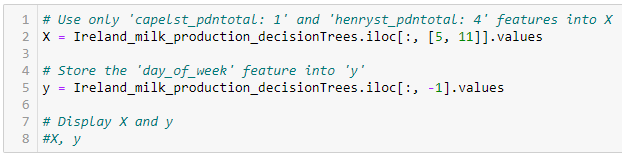
The Following findings were the reason for selecting the supervised learning model for my project;

1. Data frame selected\_eu\_production\_lp\_df had labelled data which I used to train supervised learning algorithms.
2. I was able to predict the learning outcomes.
3. Input data were available to provide the model along with the output.
4. My goal of using supervised learning is to train the model so that it can predict the output when it is given new data.

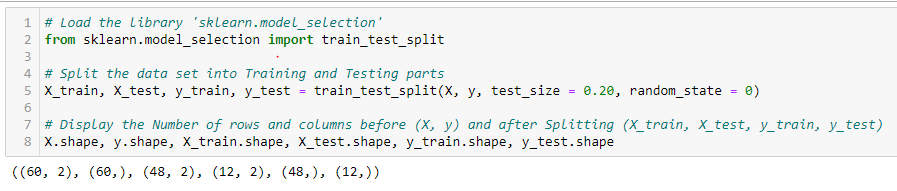
# **14. Decision Trees**

Decision tree is a popular supervised learning algorithms in machine learning. I used decision tree on my data frame to analyse how increased yield value effected milk production. This algorithm was performed on Ireland’s data from 1960 to 2020. I made a new copy of the data frame from selected\_eu\_production\_lp\_df to Ireland\_milk\_production\_decisionTrees. Some data preparation tasks performed on the copied data set. I chose 2 columns (Milk Production\_Value (tonnes) and Animal\_Value(head count)) from the data frame and these columns were X axis. (Jupyter notebook section 5.0.1)

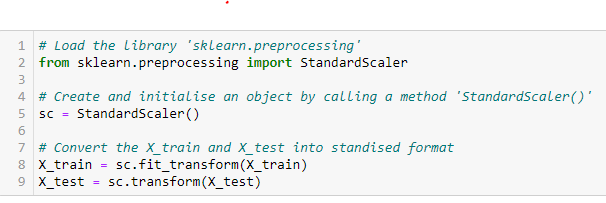
*A. Separate the independent and dependent variables using the slicing method*



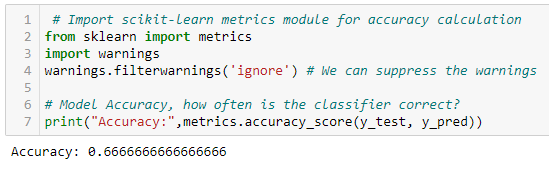
*B. Split the data into training and testing sets*



*C. Training the Decision Tree Classification model on the Training set*

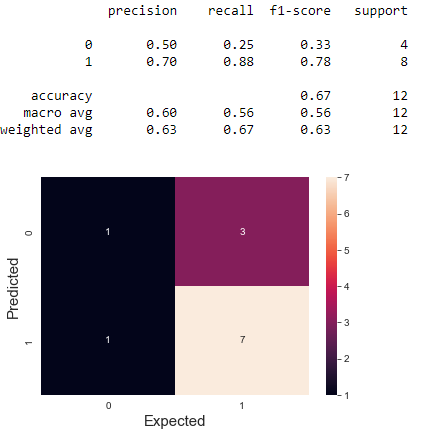


*D. Check the model score*



Decision tree got classification rate of 66.66% considered as good accuracy.

*E. Confusion Matrix*



*F. Visualising Decision Trees*

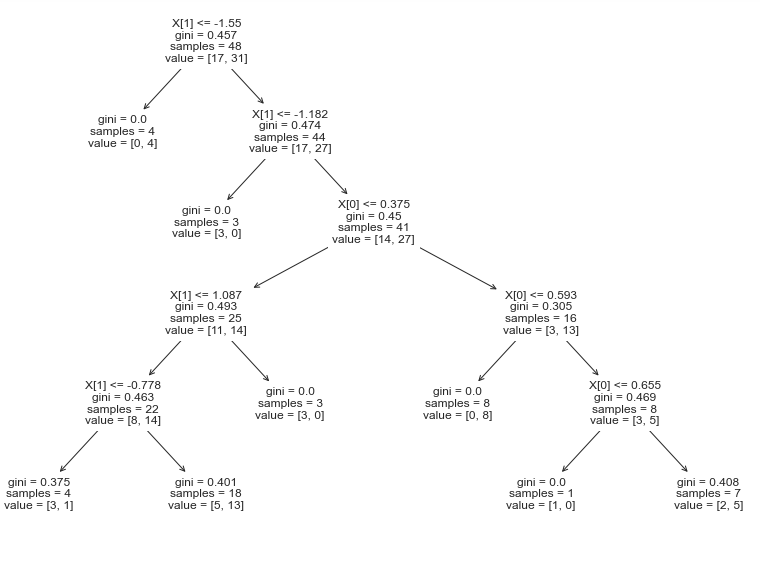


Figure 20 Decision Tree

In fig 20, each internal node has a decision rule that splits the data. Gini referred to as the Gini ratio, which measures the impurity of the node.

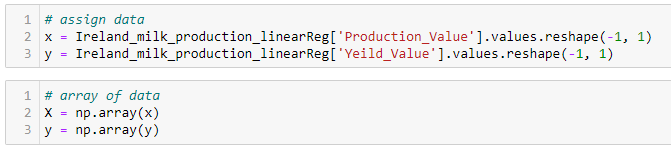
# **15. Linear Regression**

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. I used the linear regression to analyse whether the milk production value and yield value had a correlation. This analysis was performed on the data from Ireland milk production from 1960 to 2020. (Jupyter notebook section 6.0.1)

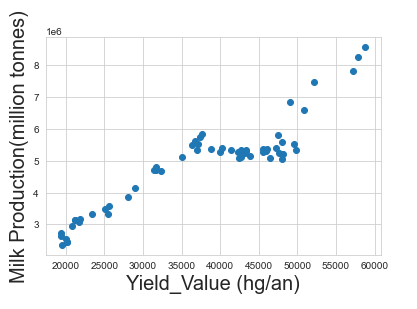
*A. Data preparation*

A copy was made from the original data frame. Data preparation and pre-processing tasks were performed on the copied data set.

*B. identify X and Y variables*

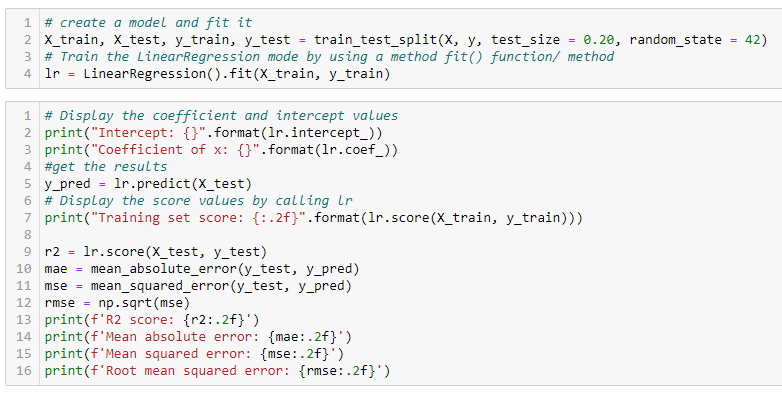


*C. Visualising the data*



The results show that when a yield value increased, the milk production also increased.There is a fairly high positive correlation here.

*D. Train and test the model*



Data was split into train and test size and the 80/20 train-test split was used to split the

Prediction. The prediction was made on the test data and after evaluating the model the following scores were achieved:

**Intercept: [709606.84976969]**

**Coefficient of x: [[110.20913804]]**

**Training set score: 0.82**

**R2 score: 0.90**

**Mean absolute error: 379799.73**

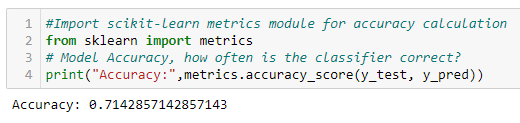
**Mean squared error: 239295536082.24**

**Root mean squared error: 489178.43**

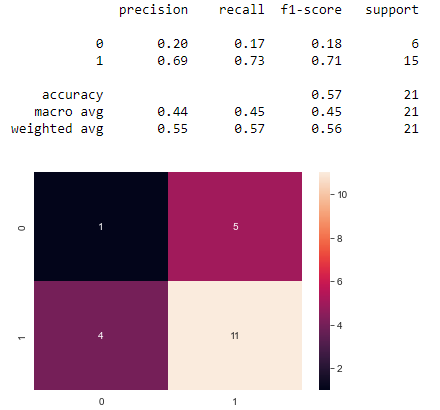
# **16. Random Forest**

Random forest was performed on the Ireland milk production data frame to analyse how increased yield value effected milk production.

Model accuracy



Model confusion matrix



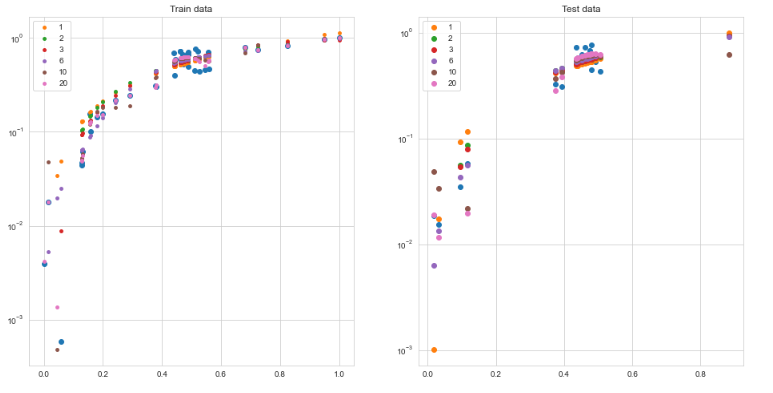
# **17. Cross-Validation with Linear Regression**

Cross validation is a useful tool when the size of the data is limited. Cross validation iteratively splits the data set into two portions; a test and trailing set. The prediction error from each of the test sets are then averaged to determine the expected prediction error for the whole model.

*A. Experiment to understand overfitting*

There were some experiments preformed on the data set to understand the overfilling. I used polynomial regression for this experiment.

*Visualise the train and test predictions*



*B. Compare R2 for Train and Test sets*

**R-squared values:**

**Polynomial degree 1: train score=0.86, test score=0.81**

**Polynomial degree 2: train score=0.88, test score=0.83**

**Polynomial degree 3: train score=0.88, test score=0.83**

**Polynomial degree 6: train score=0.89, test score=0.85**

**Polynomial degree 10: train score=0.92, test score=0.79**

**Polynomial degree 20: train score=0.93, test score=-15172029.45**

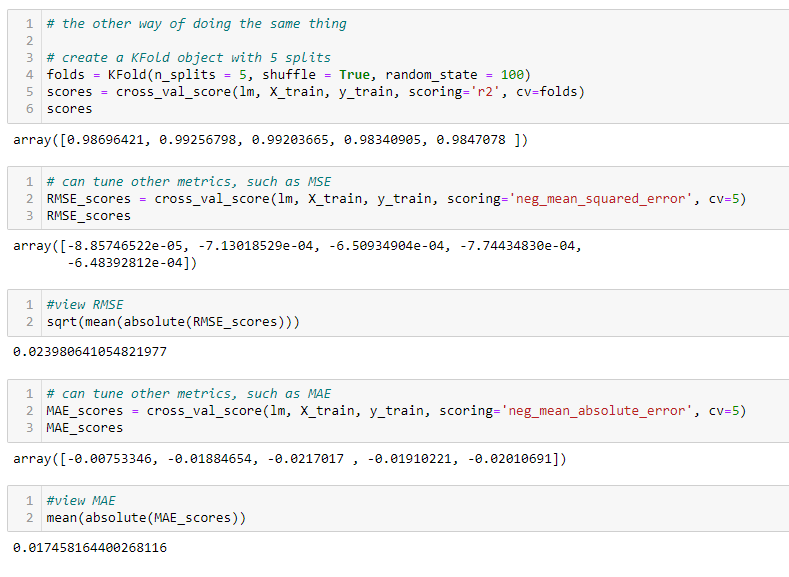
After examining the above values, it is concluded that the model is fitting well and the R2 score is also very good.

*C. Cross validation methos*

Firstly, I built a model without cross validation, then the following validation methods were performed on the model.

1. **K-Fold Cross Validation**

This validation randomly divides the dataset into k groups or folds, of roughly equal size. Data were prepared from the main data frame.



From the output I could see that the mean absolute error was 0.017. The result is that the average absolute error between the model prediction and the actual observed data. The lower MAE means that the model is able to predict the actual observations very well. Root mean squared error (RMSE) is another metric that was calculated to evaluate the model. RMSE output is 0.023. Lower RMSE also means that the model can predict the actual observations very well.

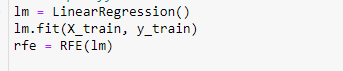
1. **Grid Search Cross Validation**

Grid Search assesses the performance for each possible combination of the hyperparameters and their values. It also chooses the combination with the best performance.

Step 1. Create a cross validation scheme

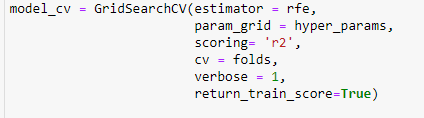


Step 2. Specify model



Step 3. Call GridSearchCV().

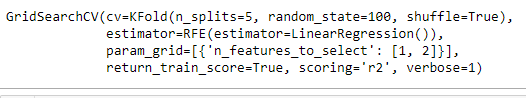
This process will construct and evaluate the model for each combination of parameters.



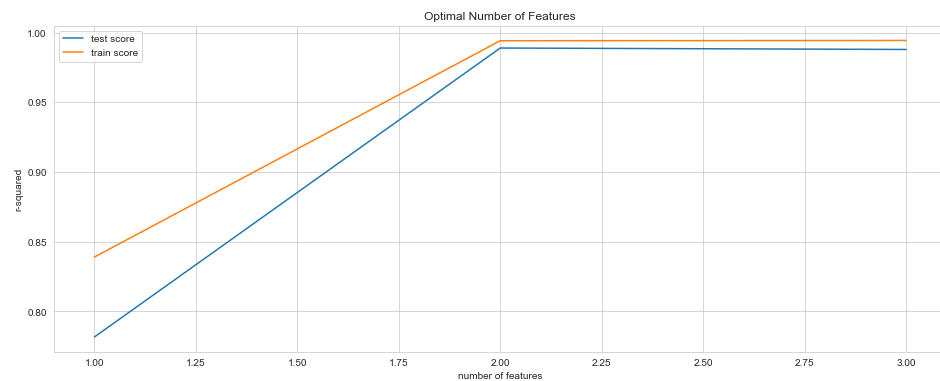
Step 4. Fit the model



Results are



Step 5. Plotting the Cross-validation result

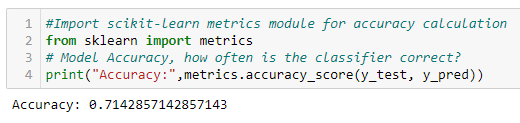


Step 6. R2 test score

**0.99218**

# **18. Machine Learning outcomes**

***Random forest***



***Linear regression score***

Intercept: [709606.84976969]

Coefficient of x: [[110.20913804]]

Training set score: 0.82

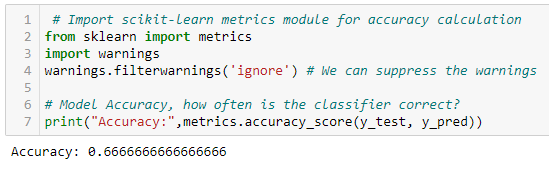
R2 score: 0.90

Mean absolute error: 379799.73

Mean squared error: 239295536082.24

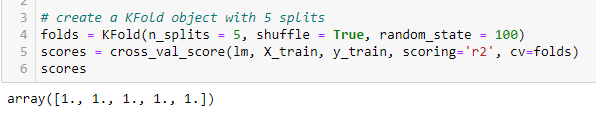
Root mean squared error: 489178.43

***Decision tree accuracy***

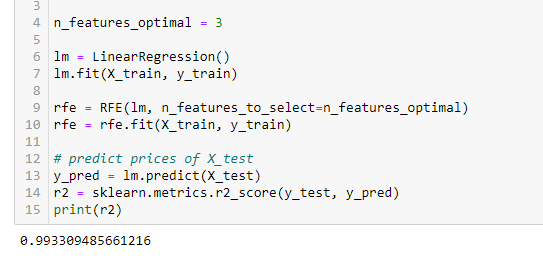


***Cross validation scores***

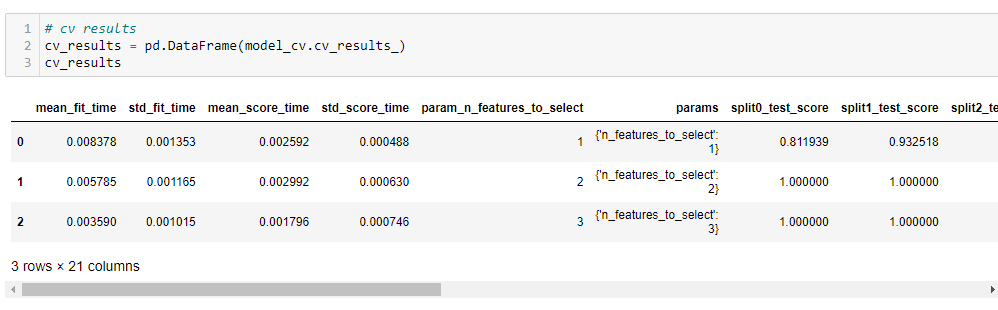
***K-Fold CV score***



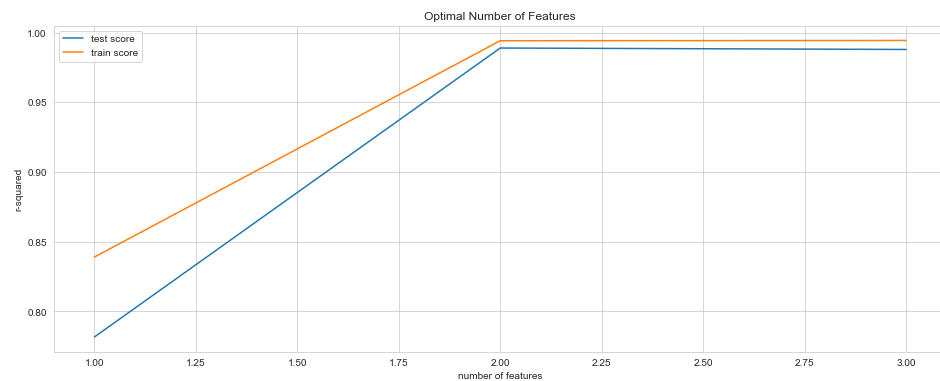
***Grid search CV score***



Grid search cross validation result table



Grid search cross validation train and test graph



Both validation scores are very close to each other, This means that the weekdays and weekend has a positive impact on the pedestrian count in Capel Street.

# **19. Sentiment Analysis**

Add stuff here

# **20. Conclusion**

Add stuff here

# **21. References**

***Books***Andreas, C. M. and Sarah, G. and O'Reilly (2016) Introduction to Machine Learning with Python, Media, Inc.

Hands on Machine Learning with Scikit Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, September 2019, ISBN: 9781492032649.

Python Machine Learning Third Edition, Sebastian Raschka , Vahid Mirjalili , Copyright © 2017 Packt Publishing.

Discovering Knowledge In Data: An Introduction To Data Exploration, Second Edition, By Daniel Larose And Chantal Larose, John Wiley And Sons, Inc., 2014.

Chun-houh, C. and Wolfgang, H. and Antony, U. (2008) Handbook of Data Visualization, Springer publishing.

Wes, M. (2017). Python for Data Analysis, 2nd Edition.

Perez (2021). Data Mining. The CRISP-DM Methodology. The CLEM language and IBM SPSS MODELER.

Ojala and Garriga (2010) Permutation Tests for Studying Classifier Performance.

***Websites***

Google (2022) Machine Leaning. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values. (Accessed: 17 Dec 2022).

Google (2022) Exploratory data analysis. Available at: https://businessanalyst.techcanvass.com/objective-of-exploratory-data-analysis/ (Accessed: 11 Dec 2022).

Google (2022) Handling outliers. Available at: https://towardsdatascience.com/exploratory-data-analysis-topic-that-is-neglected-in-data-science-projects-9962ae078a56. (Accessed: 12 Dec 2022).

Google (2022) Normal Distribution. Available at: https://www.askpython.com/python/normal-distribution (Accessed: 12 Dec 2022).

Google (2022) Pedestrian footfall. Available at: https://www.retailsensing.com/people-counting/count-pedestrian-footfall-town-centre/ (Accessed: 22 Dec 2022).

Google (2022) Linear regression. Available at: https://realpython.com/linear-regression-in-python/(Accessed: 20 Dec 2022).

Google (2022) Supervised Leaning. Available at: https://www.sciencedirect.com/topics/computer-science/supervised-learning (Accessed: 15 Dec 2022).

Google (2022) Data Science project management methodologies. Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb (Accessed: 13 Dec 2022).

Google (2022) Descriptive analysis. Available at: https://www.analyticssteps.com/blogs/overview-descriptive-analysis (Accessed: 18 Dec 2022).

Google (2022) Descriptive analysis. Available at: https://www.labxchange.org/library/items/lb:LabXchange:10d3270e:html:1 (Accessed: 22 Dec 2022).

# **22. Appendix**

## 22.1 Gantt Chart

## 22.2 Population in EU countries in 2020

