**CCT College Dublin**

**Assessment Cover Page**

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

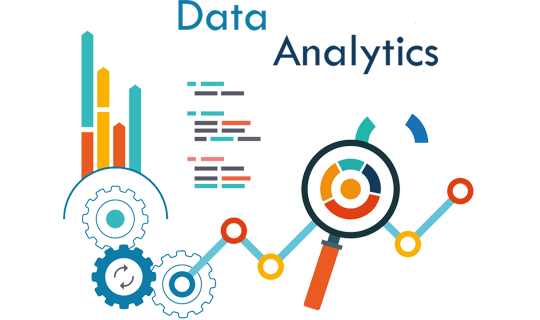
|  |  |
| --- | --- |
| **Module Title:** | Programming, Stadistics, Machine learning for Data Analytics and Data preparation and Visualization |
| **Assessment Title:** | MSC\_DA\_CA1 |
| **Lecturer Name:** | David McQuaid, Muhammad Iqba, Taufique Ahmed and Sam Weiss |
| **Student Full Name:** | Jesus Rodrigo Colina Nunez |
| **Student Number:** | 2017156 |
| **Assessment Due Date:** | 12/11/2023 |
| **Date of Submission:** | 12/11/2023 |

**Jesus Rodrigo Colina Nunez**

**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. |

**MSc in Data Analytics**



MSC\_DA\_CA1

Jesus Rodrigo Colina Nunez

Student Number: 2017156

Contents

[Abstract 4](#_Toc150728169)

[Data preparation and Visualization 5](#_Toc150728170)

[Machine learning for Data Analytics 9](#_Toc150728171)

[Statistics 13](#_Toc150728172)

[Programming 19](#_Toc150728173)

[GitHub Link 20](#_Toc150728174)

[Bibliography 21](#_Toc150728175)

# Abstract

I have downloaded several data sets from the website: <https://data.cso.ie/product/pme>, I have performed EDA on various data sets, and I have decided to only use two for my assignment.

The ones that I have chosen are: “Annual\_Population\_Change.csv”, where I noticed this data set it is focused on the population over the years.

The second data set I decided to use is: “Estimated\_Population\_from\_2006.csv”, as the previous data set it also focuses on the population but this one from 2006, it also gives us insight information into the genders and citizenships.

After reviewing the different sets, I decided that my area of interest will be Annual population change and Population forecast. As well I will only focus from the year 2006 in Ireland. Like that I am going to be able to use both data sets from the 2006 onwards.

# Data preparation and Visualization

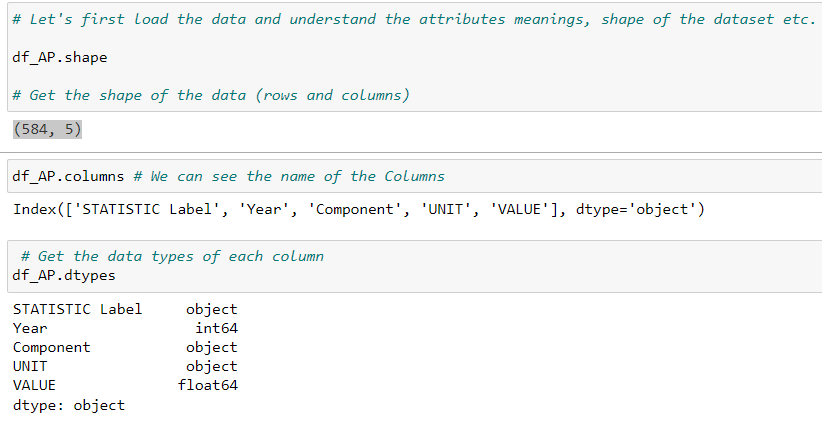
Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions (IBM, 2023).

EDA is like digging into data to find out interesting things, not just sticking to formal models or testing specific ideas. It helps you understand the different parts of the data and how they relate to each other. Also, it is useful to check if the statistical methods you plan to use for analysing the data are the right ones.

I will not put the code that I already use in my Jupyter Notebook file, in this section I will explain generally what I did but I will use snapshots in order to explain it.)

**1.- Load and Inspect the Data**

We need to load the data and get a basic understanding of its structure and contents.



For out Annual\_Population\_Change.csv file we noticed that there are 584 columns and 5 rows.

Then we proceed to know the name of the columns. We also get the data type of each column, this is important. For example, if we want to use the column Year and we want to manipulate it, it has to be an Integer and not an object.

**2.- Data cleaning and preprocessing**

Before diving our EDA, it is important to clean and preprocess the data.

In here we would have to handle missing values, so we check for missing values. For our df\_AP we noticed that there is 144 missing values in our column Value.   
Want we can do is replace it with mean or drop them, in this case we decide to replace them with mean. But in the end, it will not matter because I will use the data set form 2006 and the values that are missing are before that year.

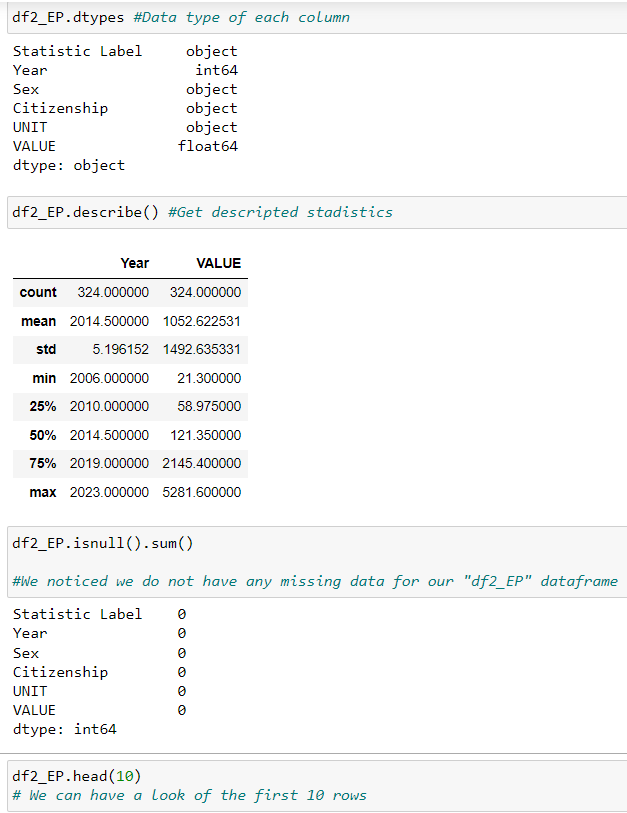
And if it is necessary, we would have to convert data types if necessary (e.g., dates to datetime format). For our both data sets it is not necessary this step.

**3.- Exploratory Data Analysis**

This step involves visualizing and analysing the data to gain insights.

Here we will analyse individual variables using histograms, box plots, etc.  
Explore relationships between pairs of variables using scatter plots, correlation matrices.  
Understand interactions between multiple variables.

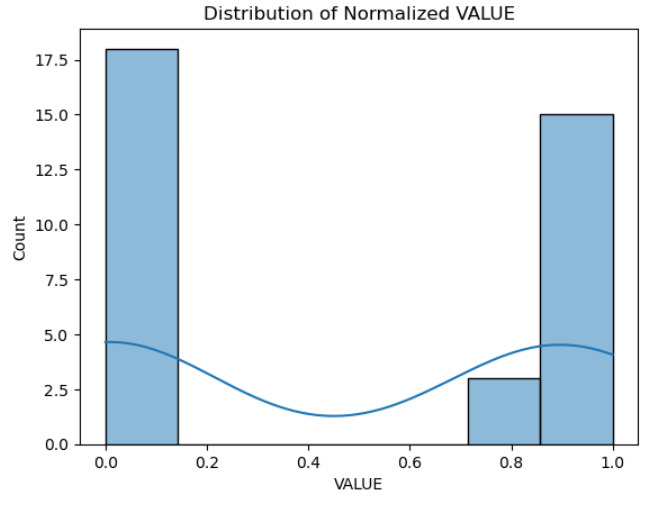
Here we also do the same for second Data Population same Estimated\_Population\_from\_2006.csv



**4. Preparing Data for Machine Learning**

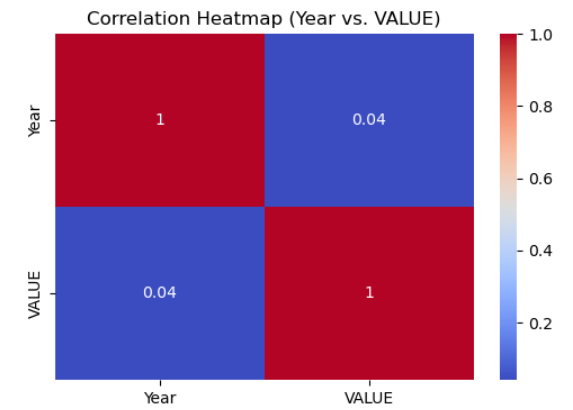
In here we are going to choose relevant features for the model.  
We would have to normalize if it is necessary

Here we could convert categorical variables to a format that can be used by ML models.



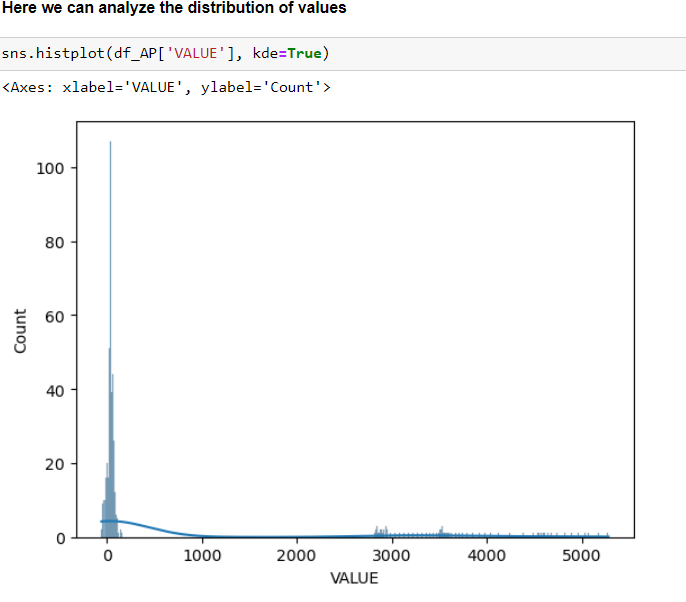
I mostly used bar plots for my graphs. Bar plots are straightforward and easy to interpret, making them accessible to a wide audience. They provide a clear visual representation of data where the length of the bar is proportional to the value it represents. They are also easy to compare data.

**5.- Visualization for Insights**

****

This heatmap shows the correlation coefficients between pairs of variables in your dataset.

The colors indicate the strength and direction of the correlation, with warmer colors (red) indicating strong positive correlations and cooler colors (blue) indicating strong negative correlations.  
A correlation heatmap is essential for understanding the linear relationships between variables.   
Identifying highly correlated variables can inform feature selection in ML models, helping to avoid issues and improve model performance.



The histogram for the VALUE column visualizes the distribution of population changes. This is crucial for understanding the spread and skewness of the data.

This approach provides a comprehensive understanding of the dataset and prepares it effectively for ML modeling, with each step designed to enhance the overall analysis and model performance.

# Machine learning for Data Analytics

The management framework that I have chosen is CRISP-DM and I used a Supervised machine learning technique for my data modelling.

CRISP-DM. It stands for Cross-Industry Standard Process for Data Mining. This methodology was originally developed in IBM for Data Mining tasks. This model has the same cyclic nature as both KDD and SEMMA. The key difference in the structure is that the transitions between stages can be reversed. If during the modeling stage the specialist found the data not sufficient to resolve the goal of the project, they can return to the data preparation stage and select different target variables, generate features, etc (Quantum, 2023).

One of the reasons I decided to use this framework is because it is widely used and flexible.

CRISP-DM approach for my population dataset.

1.- **Business Understanding:** Our objective is to forecast population growth and analyse population during the last years. This involves examining the data structure and identifying the relevant variables in order to understand the data.

2.- **Data Understanding:** In this step I am going to Explore and familiarize myself with the population data set. In order to do that I have to examine the data structure.

3.- **Data Preparation:** Here we are going to process the data, checking if there is any missing values, here we can handle outliers as well and If it is necessary we will transform the variables.

4.- **Modeling:** Here we apply statistical or machine learning models to analyse the population data. For this example, I decided to use Lenear Model, Lasso Model, KNN model and Random Forest Model. In here we can predict future population trends based on demographic similarities and other relevant analyses.

5.- **Evaluation:** Here we are going to assess the performance of the models and analyses. We will determinate if the results meet the initial objectives, he we can also make adjustments or improvements needed.

6.- **Deployment:** In this last section we are going to implement the findings from the analysis, here we create visualizations. In my case I created a visualisation where I can forecast the next 5 years of population growth using linear modelling.

As I have a specific target variable and I want to predict population growth, my data is well labelled I have decided to use **Supervised learning**.

Supervised learning is based on training. During its training phase, the system is fed with labeled data sets, which instruct the system what output variable is related to each specific input value. The trained model is then presented with test data. This is data that has been labeled, but the labels have not been revealed to the algorithm. The aim of the testing data is to measure how accurately the algorithm performs on unlabeled data (TechTarget, 2023) .

I used GridSeachCV for my Random Forest

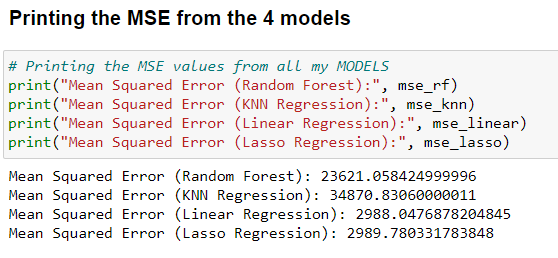
(Grid Search Cross-Validation) is a technique used in machine learning to search and find the optimal combination of hyperparameters for a given model. It systematically explores a predefined set of hyperparameter values, creating a “grid” of possible combinations. It then evaluates each combination using cross-validation and selects the one that produces the best performance. GridSearchCV helps in automating the process of hyperparameter tuning, enhancing model performance, and avoiding manual trial-and-error (Analytics Vidhya, 2021).

For data modelling I used:

Lineal Regression, Lasso Regression, KNN Regression and Random Forest.

The absolute value of MSE can be somewhat abstract and depends on the scale of my target variable. For instance, an MSE of 100 might be very small if your target variable ranges in the thousands, but quite large if your target variable ranges around 10.

Here is my display for the 4 Models

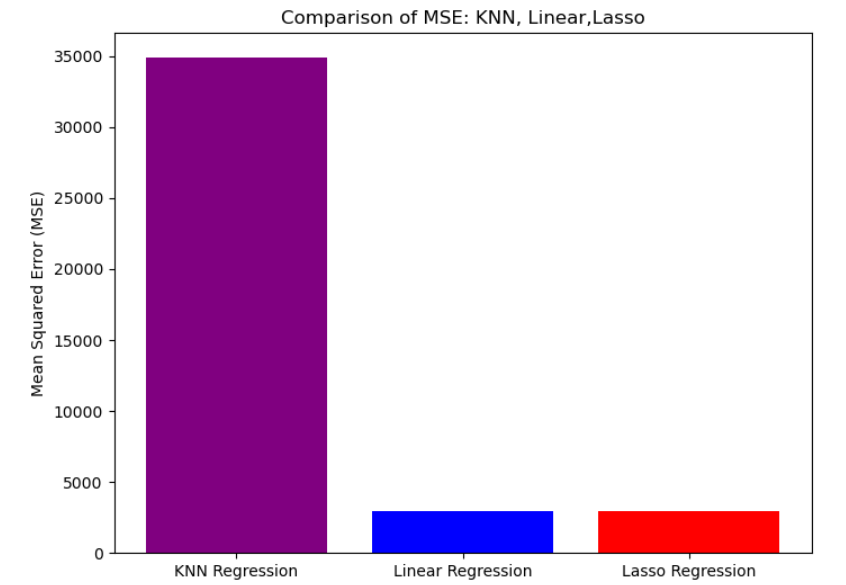


MSE can be heavily influenced by outliers. A few large errors can dominate the MSE, potentially leading to a misleading interpretation of the model’s performance. That is why EDA is important in order to see the outliers.

A lower MSE for one model suggests that it is better at predicting the target variable in your specific context.

* MSE gives an idea of how close a regression model's predictions are to the actual data points.
* A lower MSE indicates a better fit of the model to the data.
* It is always non-negative, and values closer to zero are better.

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD) (Frost, 2023).



KKN Regression and Random Forest overfit my model and those are not useful to forecast the population.

We also noticed that Linear Regression and Lasso Regression the MSE are close to each other.

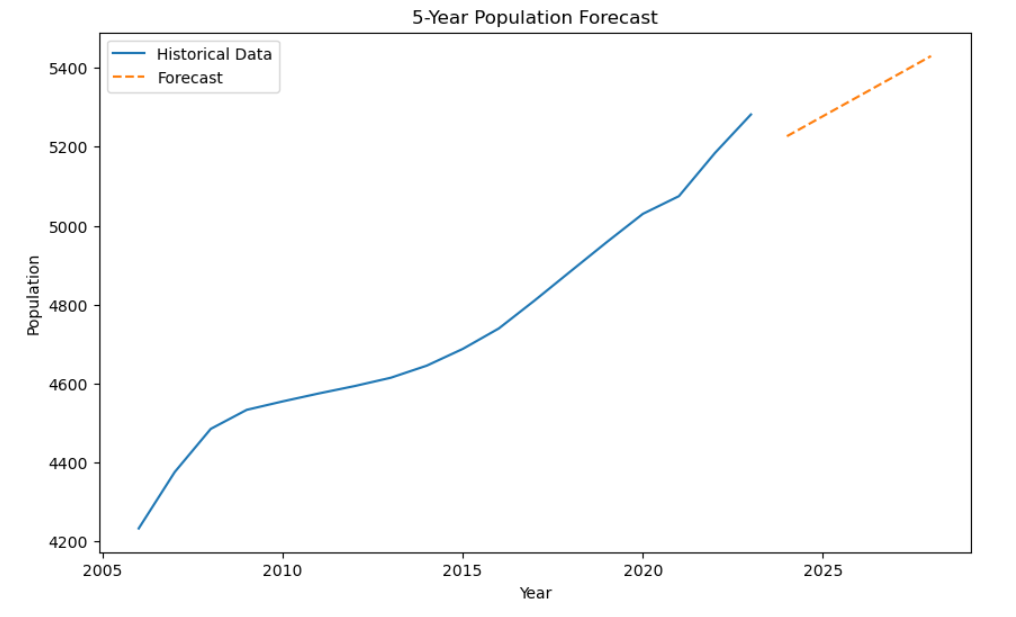
**Linear Regression** provided the best performance among the models tested, as indicated by its lower MSE. This suggests that the features in your dataset likely have a linear relationship with the target variable.

**Random Forest** is the one with the highest MSE this could imply that the model is either overfitting the training data or that the data's relationship is not complex enough to benefit from this model capabilities. Random Forest is typically more suited for complex, non-linear data. This could be the reason it did not work for my dataset as it is more lineal.

**Lasso Regression** would also be a good choice, Lineal and Lasso Regression had pretty similar results as we can see in our chart. But the one with lower MSE is linear Regression.

**KNN** applicability and performance would not work to forecast population because it had hight MSE. It is best suited for scenarios where proximity-based similarity is a strong indicator of the target variable.

Having all these in consideration we see that Linear Regression is the one that best fits to forecast the next 5 years for the population from 2023 to 2028. Here an example:



A screenshot of a graph

Description automatically generated

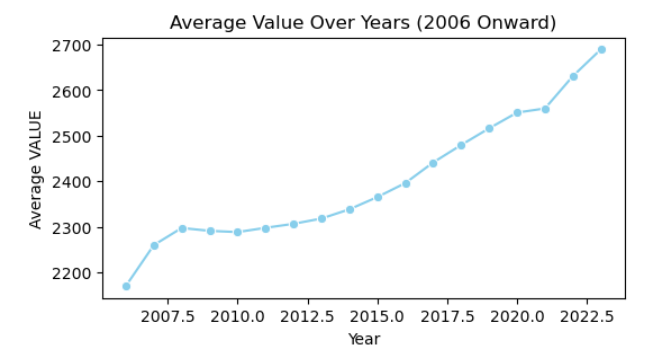
Here we can Cleary see the prediction of Population for the next 5 years and its increase over the years. It is really easy to visualize using this plot.

We also have the table indicating the amount of population. I think in the end I was able to archive my goal that was predict population for the next 5 year. The prediction is quite accurate, and I am happy with the result of my data modelling. If I had more time, I would have use more models to see if I had found a better suitable than the Linear one.

# Statistics

Here I will display 3 graphs and analyse my findings. (More graphs can be found on my python file)

Average Value Over Years (2006 Onward)

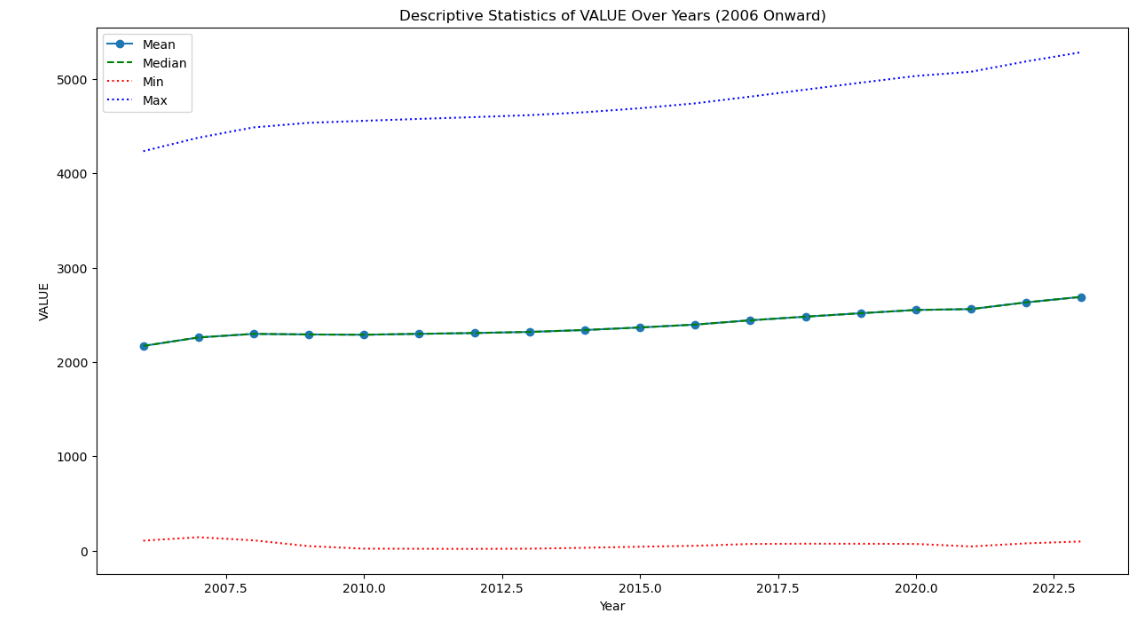


The line plot provides a visual representation of how the 'VALUE' metric has changed over the years since 2006. The shape of the line indicates the trend increasing over time.

The line plot uses the mean 'average' of 'VALUE' for each year to represent the data. This approach is useful for understanding the overall trend but does not show the distribution or variability of values within each year.

Overall, the plot is a clear and straightforward way to understand how the 'VALUE' has evolved annually from 2006 onwards, but it does not provide insights into the distribution or variability of the data within each year.

Descriptive Statistics of Value Over the Years from 2006



The chart displays various descriptive statistics of the 'VALUE' metric from 2006 onward, including the mean, median, minimum, and maximum. Here are the findings:

Mean (Average) Trend: The line marked with circles represents the mean value for each year. This line shows how the average 'VALUE' has changed over time, indicating the central tendency of the data.

Median Trend: The dashed green line shows the median value each year. The median is particularly useful in understanding the distribution's center, especially if the data is skewed.

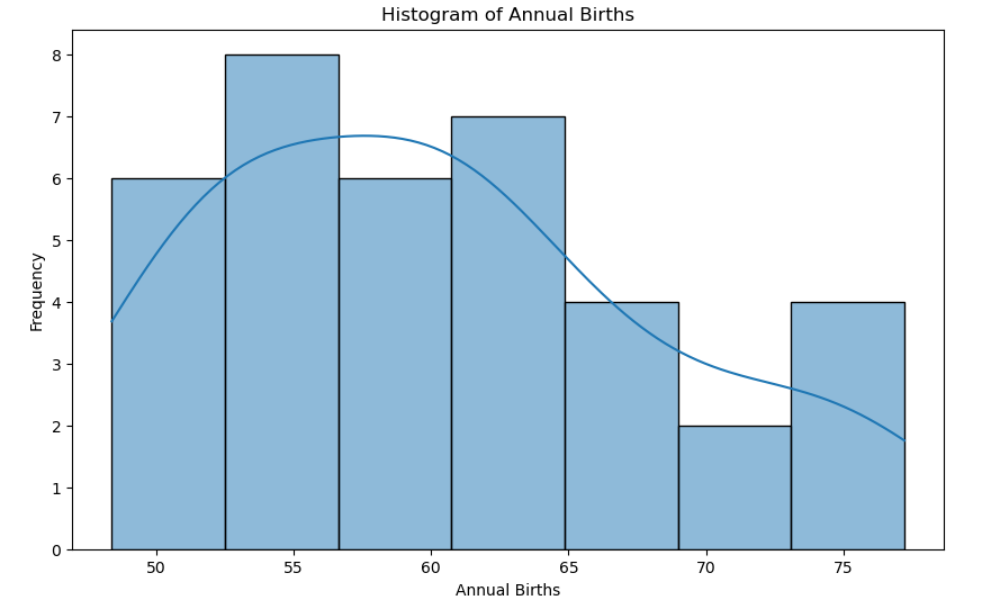
Minimum and Maximum Trends: The dotted red and blue lines represent the minimum and maximum values for each year, respectively. These lines provide insights into the range of the data, showing how extreme the values can get.

Overall, this chart provides a comprehensive overview of the data's distribution characteristics over time. It highlights not only the central tendency (mean, median) but also range (min, max) of the 'VALUE' metric, offering a more detailed understanding than a simple average trend line.​​

Min and max: Shows you the lowest (minimum) and highest (maximum) values in your column.

Mean: Also called the average. The sum of all the values in your column divided by the total number of values. Median: The number that would be in the middle of an ordered list of your values (academy.datawrapper, 2023).

The histogram of annual births provides a visual representation of the frequency distribution of the number of annual births recorded in my dataset “Annual\_Population\_Change”

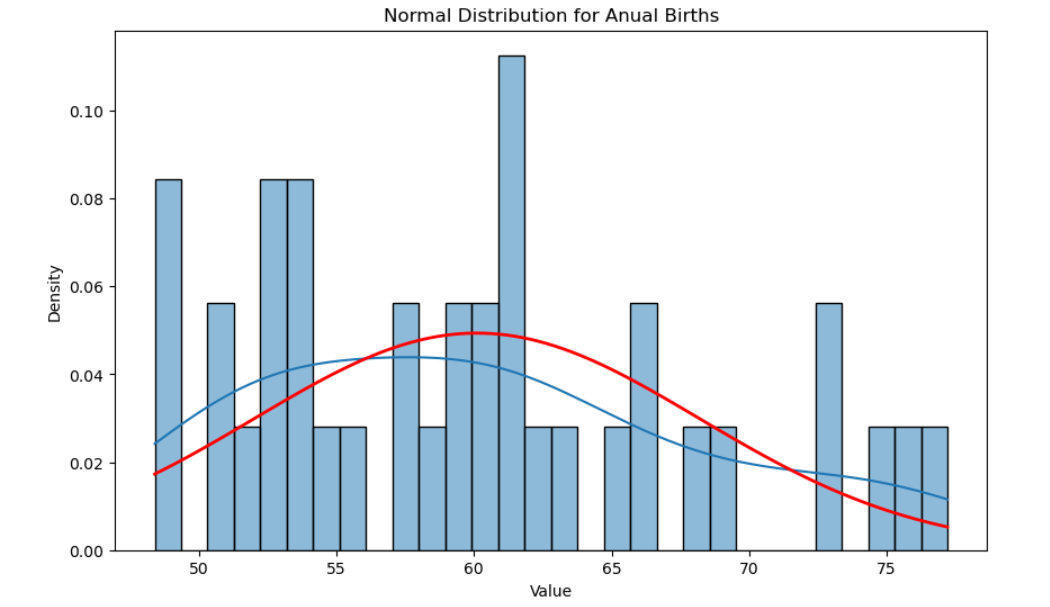


By analyzing this histogram, we can gain insights into the typical range of annual births, identify any unusual patterns or outliers, and understand the general trend in birth rates over the years covered in the dataset.

**Normal Distribution**

Here we will analyze the normal distribution of a specific component of data, in this case “Annual Births”, we also going to visualize its distribution, comparing it to a 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 are more frequent in occurrence than data far from the mean (Investopedia, 2023).

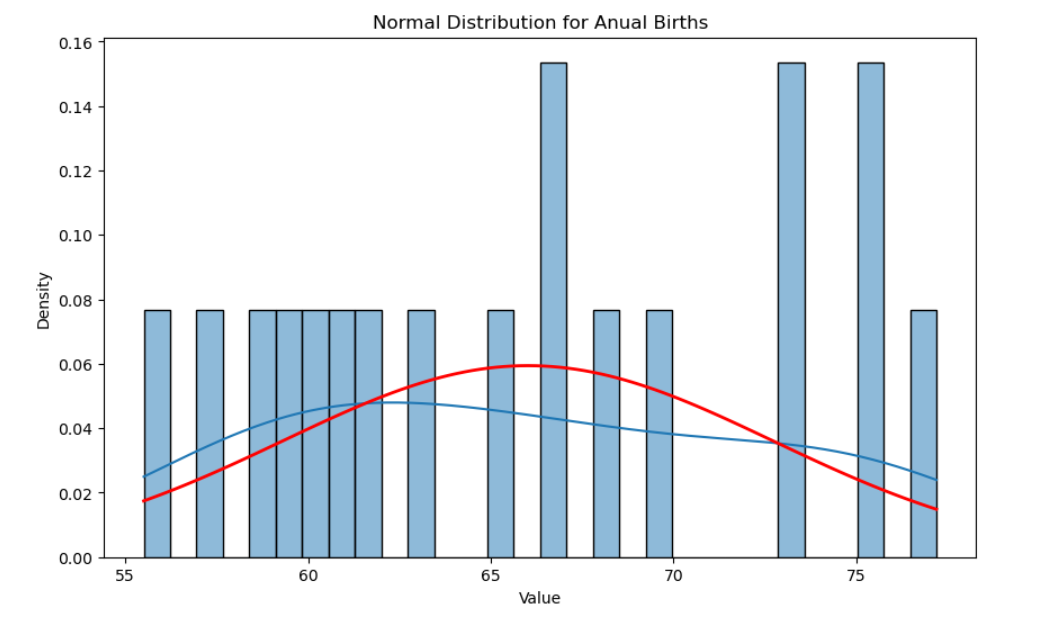
****

The result of this script is a graph that shows how closely the distribution of annual births follows a normal distribution that means that the red line closely matches the histogram. The histogram represents the actual distribution of the data, while the red line shows what a perfect normal distribution with the same mean and standard deviation would look like.

In this example we notice how the data is a bit big on the sides, this can also indicate skewness.

If I take the same example but I apply it to the data set that I have from 2006 we notice that the data is well distributed.

This graph shows the normal distribution for annual Births from 2006

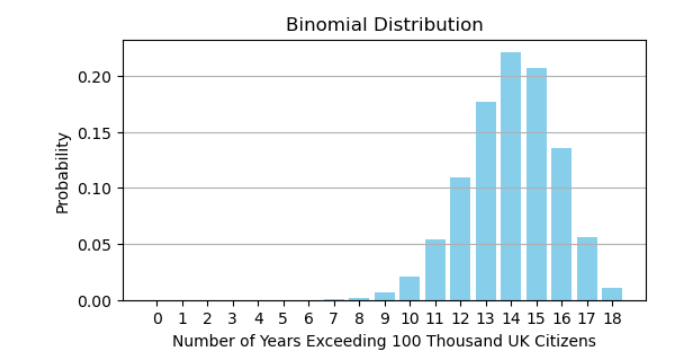


**Binomial Distribution**

Binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as normal distribution. This is because binomial distribution only counts two states, typically represented as 1 (for a success) or 0 (for a failure), given a number of trials in the data. Binomial distribution thus represents the probability for x successes in n trials, given a success probability p for each trial (Investopedia, 2023).

Here we are going to calculate the probability of a specific event happening a certain number of times over a given period, using the Binomial distribution. The event in this case is the population of the United Kingdom for both sexes exceeding 100,000 in a given year.

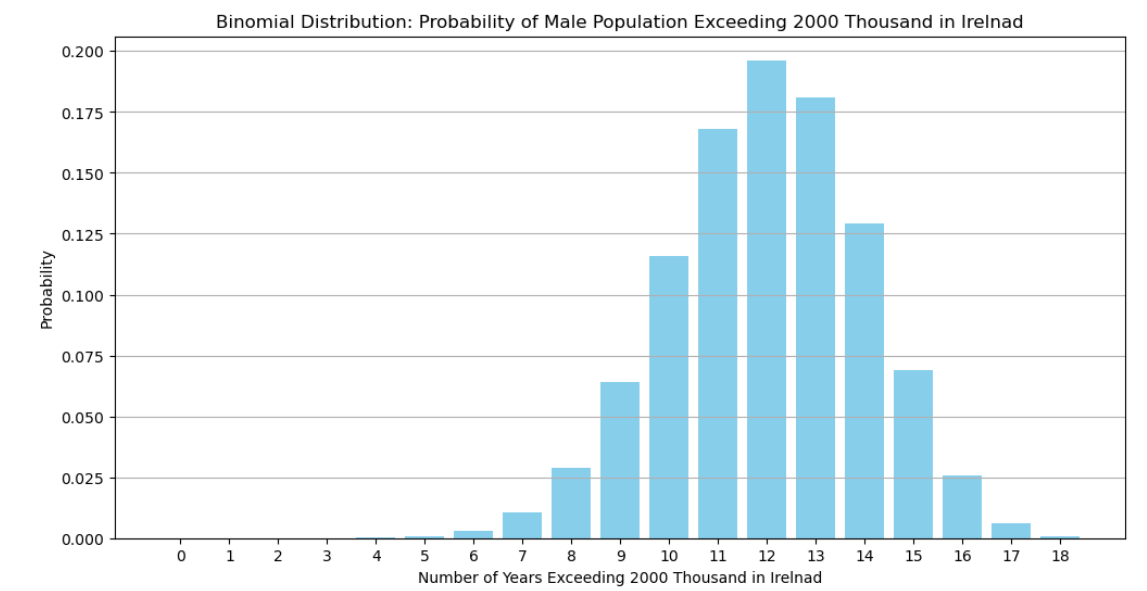
So we are going to see the probability and percentage of times that it can occur.



We noticed that there is a probability of 77% that means that there is the probability that we the population of 100 thousand UK citizens exceeds. In this example we can observe that 14 out of 18 would happen. That means that there is like hood that happened.

**For our second giving example let’s make a second event:**

The event in question is the population of male citizens from Ireland exceeding 2,000 (thousands) in a given year. This event it is similar to the first one.



We calculated the probability of 66% for the event to occur exactly 12 times in 18 years. That means that there is the probability of exceeding 2000 thousand male Irish citizens in Ireland.

In summary this binomial distribution shows that it is likelihood of a repeated event happening a specific number of times over a set period.

This can be useful in demographic studies, resource planning where understanding the probability of such events are crucial.

**Poisson distribution**

Unfortunately did not work and that is why I decided to use two Binomial Distribution, maybe my data was not suited for my Poisson distribution example.

A Poisson distribution is a discrete probability distribution, meaning that it gives the probability of a discrete (i.e., countable) outcome. For Poisson distributions, the discrete outcome is the number of times an event occurs, represented by k. You can use a Poisson distribution to predict or explain the number of events occurring within a given interval of time or space. “Events” could be anything from disease cases to customer purchases to meteor strikes. The interval can be any specific amount of time or space, such as 10 days or 5 square inches (Turney, 2023).

# Programming

Now I am going to briefly discuss my use of aspects of various programming paradigms in the development of my project.

**1. Procedural Programming:**

Commonly used in data preprocessing and analysis. The step-by-step instructions and functions in Python, such as loading data, cleaning, and transforming data, are typical examples.

It guides the project's structure from data loading, preprocessing, to modeling and evaluation.

It helped me in breaking down the analysis process into manageable and logical steps, ensuring a clear and understandable code flow.

**2. Object-Oriented Programming (OOP):**

Predominantly observed in structuring machine learning models and handling data. Python libraries like scikit-learn extensively use classes and objects. For example, model instances (LinearRegression(), RandomForestRegressor() are objects with attributes and methods.

How it helped me: It enables code reusability and modularity. Designing custom functions or classes for repetitive tasks (like evaluation metrics) and makes the code more organized.

The advantages of OOP include reusability, modularity, encapsulation, and inheritance. These features make it easier to develop, test, and maintain code (Mandal, 2023).

**3. Functional Programming:**

Functional programming is a programming paradigm in which code is structured primarily in the form of functions. The origins of this programming style arise from a branch of mathematics known as lambda calculus, which is the study of functions and their mathematical properties (codecademy., 2023).

I used it in operations that involve data manipulation and functional transformations for example lambda functions.

Promotes writing more concise and potentially more readable code. I used it in my data transformation and analysis steps.

It helped me in simplifying code, which is crucial in data processing tasks. For example, applying a function to transform a column in a DataFrame without altering the original data.

**4. Declarative Paradigm:**

For example, pandas query methods and regular expressions for data extraction and cleaning.

It allows for a more intuitive specification of what the program should accomplish rather than detailing the control flow to achieve it.

Declarative programming is a method to abstract away the control flow for logic required for software to perform an action, and instead involves stating what the task or desired outcome is. Declarative programming is a high-level programming concept, which is the opposite of imperative programming. It is typically found in databases (Beltram, 2023).

**My conclusion:**

Each of these paradigms contributed uniquely to different stages of my project. Procedural programming establishes a clear workflow, OOP encapsulates data and functionalities, functional programming streamlines data processing, and the declarative paradigm simplifies data querying and manipulation. The combined strengths of these paradigms enhance my project's efficiency, maintainability, and scalability.

If I had to do it all over again, I would have started first with the programming section because, I did small research of each paradigm, in order to understand them better. I left this section at the end and then I had to search what paradigms I used, it would have been better if I had done it the other way around. I also learned that in the future assignments, it is better to know the theorical part then apply it to the practical part.

# GitHub Link

[**https://github.com/CCT2017156/MSC\_DA\_\_CA1**](https://github.com/CCT2017156/MSC_DA__CA1)

# Bibliography

academy.datawrapper, 2023. *academy.datawrapper.de.* [Online]   
Available at: https://academy.datawrapper.de/article/136-histogram-min-max-median-mean#:~:text=Min%20and%20max%3A%20Shows%20you,ordered%20list%20of%20your%20values.  
[Accessed 09 11 2023].

Analytics Vidhya, 2021. *analyticsvidhya.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/  
[Accessed 10 11 2023].

Ashly, 2023. *medium.com.* [Online]   
Available at: https://medium.com/@ashiq0ashi/the-process-involved-in-data-analysis-involves-several-different-steps-54f543421e93  
[Accessed 10 11 2023].

Beltram, A., 2023. *techtarget.com.* [Online]   
Available at: https://www.techtarget.com/searchitoperations/definition/declarative-programming  
[Accessed 10 11 2023].

codecademy., 2023. *codecademy.* [Online]   
Available at: https://www.codecademy.com/article/functional-programming-in-python  
[Accessed 10 11 2023].

Frost, J., 2023. *statisticsbyjim.* [Online]   
Available at: https://statisticsbyjim.com/regression/mean-squared-error-mse/  
[Accessed 09 11 2023].

IBM, 2023. *www.ibm.com.* [Online]   
Available at: https://www.ibm.com/topics/exploratory-data-analysis#:~:text=Exploratory%20data%20analysis%20(EDA)%20is,often%20employing%20data%20visualization%20methods.  
[Accessed 05 11 2023].

Investopedia, 2023. *investopedia.com.* [Online]   
Available at: https://www.investopedia.com/terms/n/normaldistribution.asp#:~:text=Error%20Code%3A%20100013)-,What%20Is%20a%20Normal%20Distribution%3F,as%20a%20%22bell%20curve%22.  
[Accessed 11 11 2023].

Investopedia, 2023. *investopedia.com.* [Online]   
Available at: https://www.investopedia.com/terms/b/binomialdistribution.asp#:~:text=without%20the%20other.-,Binomial%20distribution%20is%20a%20common%20discrete%20distribution%20used%20in%20statistics,of%20trials%20in%20the%20data.  
[Accessed 11 11 2023].

Mandal, D., 2023. *krayonnz.com.* [Online]   
Available at: https://www.krayonnz.com/user/doubts/detail/61cf0bf6d873e800403a059d/what-are-the-advantages-and-disadvantages-of-OOP  
[Accessed 10 11 2023].

Quantum, 2023. *quantumobile.* [Online]   
Available at: https://quantumobile.com/blog/data-science-project-management-methodologies/  
[Accessed 12 11 2023].

TechTarget, 2023. *techtarget.com.* [Online]   
Available at: https://www.techtarget.com/searchenterpriseai/definition/supervised-learning  
[Accessed 12 11 2023].

Turney, S., 2023. *scribbr.com.* [Online]   
Available at: https://www.scribbr.com/statistics/poisson-distribution/#:~:text=A%20Poisson%20distribution%20is%20a,the%20mean%20number%20of%20events.  
[Accessed 08 11 2023].