**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *MSc in Data Analytics* |
| **Assessment Title:** | *MSC\_DA\_CA2* |
| **Lecturer Name:** | *Sam Weiss*  *Taufique Ahmed*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Leda Mercedes Gale Torre |
| **Student Number:** | 2023205 |
| **Assessment Due Date:** | 05/01/2024 |
| **Date of Submission:** | 05/01/2024 |

**Declaration**

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

## Abstract

The report focuses on a dataset obtained from The Central Statistics Office (CSO), Ireland's national statistical office, covering the population at each census from 1841 to 2016. The analysis is developed using Python programming, involving four key approaches: Programming, Statistics, Machine Learning, and Data Preparation and visualization.

Phyton has been used for data analytics projects, emphasizing clarity in code structures. The Data Preparation and visualization describes the initial steps of Exploratory Data Analysis (EDA), including data type identification, renaming of features for clarity, and the dropping of features not useful for the analysis. Visualizations, such as line graphs and density analysis, are employed to understand population trends and distributions across counties.

Descriptive statistics have been applied, including skewness, kurtosis, and central tendency measures. The data distribution is visualized through histograms and box plots. Inferential statistics involved testing the normality of the data using a Q-Q plot and the Shapiro-Wilk test, indicating a non-normal distribution. Probability distributions, specifically the binomial distribution, are explored for discrete variables.

The machine Learning section adopts the CRISP-DM framework, emphasizing the importance of understanding the business context before analysis. Supervised learning techniques are employed, including Linear Regression, Polynomial Regression, Ridge Regression, and Decision Tree Regression. The analysis focuses on predicting population trends based on census years. Evaluation metrics, such as R2 scores, are used to assess model performance.

## Introduction

The present CA report aims to detail the step-by-step work done in the chosen dataset, providing a description of each of the sections included in the analysis.

According to the guides of the project, the study has been based on data taken from The Central Statistics Office (CSO) which is Ireland’s national statistical office, their purpose is to impartially collect, analyze, and make available statistics about Ireland’s people, society, and economy.

The dataset analyzed is “Population at Each Census 1841 to 2016”, it was downloaded from the website as a CSV file named “E2001”.

It was loaded into a Jupiter Notebook where all the project was developed using libraries such as pandas, NumPy, seaborn, and some others that helped to achieve the goals expected on the project.

The programming language used is Python, its code is clear to read and write and facilitates getting results in fewer lines than other languages.

In this project 4 approaches were taken: Programming, Statistics, Machine Learning, and Data Preparation &Visualisation.

Approaches that are amplified in the following section.

## Approaches

### Programming

Python is one of the easiest programming languages nowadays. The code used is clear to read and write. Keeping a consistent and clear code is key to success in your projects.

Python counts with eight key pieces: data types, object references, logical operations, control flow statements, arithmetic operators, input/output, creation, and calling of functions.

All of the key pieces of Python mentioned before are one way or another implemented in a data analytics project. I developed the corresponding data type identification in order to understand my dataset (elaborated a dictionary with the most relevant information). Clear object references have been used in the project as it has been necessary to store information to give it proper use.

I developed functions such as “label\_graph”, “save\_results”, and “show\_results” to avoid the repetition of code and facilitate the job. I have also used control flow statements as the for loops with the purpose of extracting the data necessary to aim the objectives established.

*Programming Paradigms*

Programming paradigms interact facilitating data scientists and developers to combine different approaches. Imperative and procedural programming paradigms were used in the present project. A series of steps or instructions were created to extract from the dataset the points of interest as procedural programming at the moment of the creation of the definition to avoid unnecessary repetitive code.

### Data Preparation and Visualisation

The first step in any investigation is the detailed revision of the resource selected as a dataset, the clear understanding of the data determines the track the work done can take.

This initial stage is called Exploratory Data Analysis (EDA), which involves exploring and summarizing data to identify patterns and relationships.

A deeper understanding of the data and the detection of potential problems or errors in the analysis are some of the benefits of an effective EDA.

Firstly I imported the corresponding libraries such as pandas, seaborn, matplotlib.pyplot, and NumPy that will be used afterward, I found that the dataset originally had 6 Features which are: Statistic Label, County, Sex, CensusYear, UNIT, and VALUE with 2025 observations.

The datatypes of the features are the following:

* Statistic Label: Object (Str)
* County: Object (Str)
* Sex: Object (Str)
* CensusYear: int64
* UNIT: Object (Str)
* VALUE: int64

I proceeded to check if the dataset has null values, finding that the dataset seems to be complete as all the feature count matches the total observations. The same was done with the duplicated values which a 0 in this case. Having made the revisions mentioned before I realized that the feature named “UNIT” seemed to be of no use to my criteria, the same just had the label “Number” indicating that the next attribute named “VALUE” was a number, I considered it was not contributing on my analysis and dropped it.

For a better understanding of the data, I renamed the features and named them as shown next:

* Statistic Label as “Stat\_label”
* CensusYear as “Year”
* VALUE as “Population”

At this point, I believed it was the right time to do the recognition of what type of data and the levels of measurement the resource had in order to create the corresponding dictionary which will be helpful not only for other users using the Python notebook but also for my own organization of the project given that the correct identification of these two criteria provides an idea of what type of analysis we can develop.

The unique() and nunique() functions made me realize that in the case of the features “County” and “Sex” the dataset groups in the labels “State” and “Both sexes” respectively the totalized Population, otherwise stated using the dataset as it is would duplicate the values and lead to wrong conclusions. Hence, I decided to split the dataset into a general one and expanded to implement the necessary analysis for each.

Once I split the data, I plotted the Population feature to visualize how it has changed over the years included in the study.

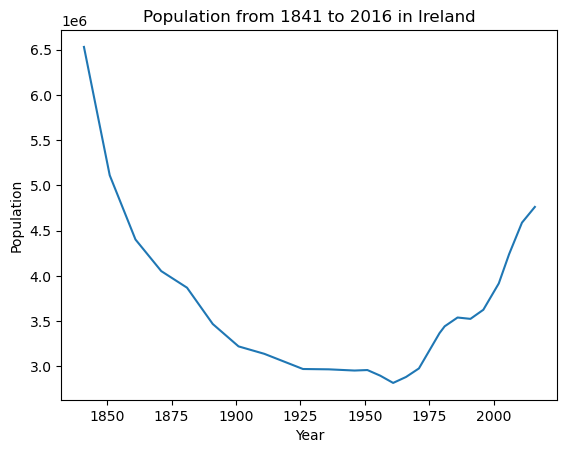


Figure 1: Population from 1841 to 2016 in Ireland from the general dataset of the project.

I considered that the line graph could be adequate as these ones are usually used to create a visual representation of values over time.

The graphic presents the Population between the years 1841 and 2016. As we can see the highest value is in the first year with a population of 6,528,799 having a big drop after it, this occurred at the same time as the Great Famine in the country, which is known as a period of starvation in Ireland in which a lot of people died.

The population kept decreasing and after Northern Ireland and Ireland split in two in 1921 this feature had its lowest point with 2,818,341 inhabitants in 1961. After the downfall, the value displays an increasing behavior recording 4,761,865 in the year 2016.

After visualizing how the population value has changed over the years, and having the data of population by County I thought it would be interesting to visualize how this population was distributed by counties. It is believed that Dublin being the capital has the largest amount of population, however I want to see graphically the difference.

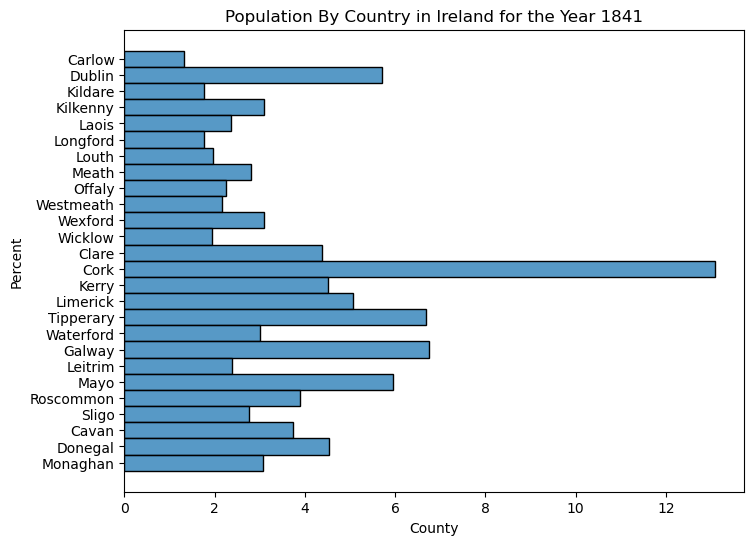


Figure 2: Population by Country in Ireland for the Year 1841.

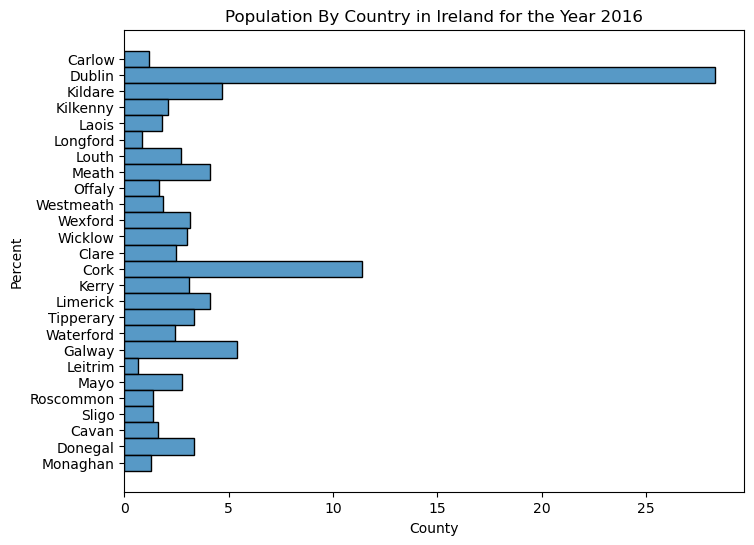


Figure 3: Population by Country in Ireland for the Year 2016.

It can be observed in Figure 3 for the year 2016 that Dublin occupies a relevant percentage of the population compared to the other counties having Cork as the second highest percentage, however, this has not been always the case as it can be observed in Figure 2 where Cork had the highest percentage.

I decided to do a density analysis, to do it I created a dictionary with the area of each county and proceeded to create a for loop to add this data to my original expanded dataset after calculated the density per area using the county area and population values.

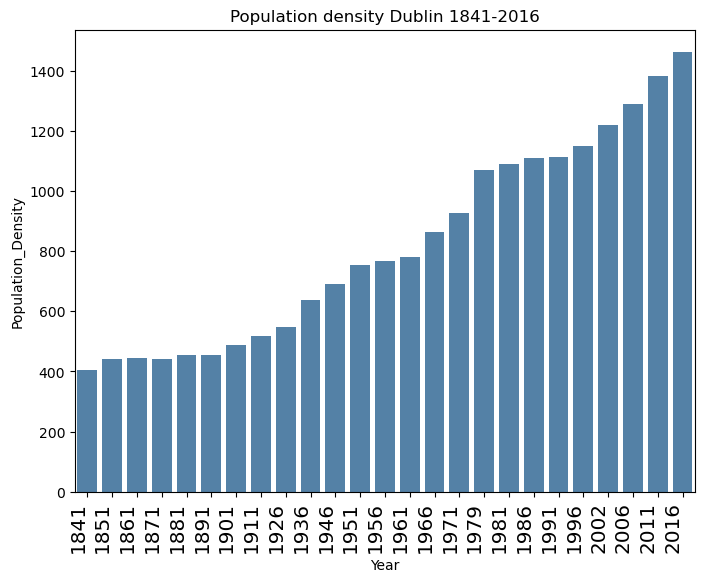


Figure 4: Population density Dublin 1841 – 2016.

Figure 4 shows how the density in the county of Dublin has increased which is consistent with the results observed before.

The data given allowed me to calculate as well the growth rate over the year in the country which is shown in the Figure 5 presented below.

The line plot presents a big decrease in the growth that can be identified for the first year included in the study which corresponds to the years a after the Great Famine, as stated earlier in this document.



Figure 5: Growth rate over the years in Ireland.

### Statistics

Statistics is the science of designing studies or experiments, collecting data, and modeling/analyzing data for the purpose of decision-making and scientific discovery when the available information is both limited and variable. That is, statistics is the science of Learning from Data. (Ott, L. and Longnecker, M. 2016)

There are two types of statistics:

* Descriptive statistics: describes basic features of data to provide an overview of the data, as it assists in summarizing, reviewing, and communicating in a meaningful way. There are two categories: Measure of Central Tendency and Measure of Variability.
* Inferential statistics: constructs predictions, and inferences and makes decisions from data.

*Descriptive Statistics*

The first step taken in this section was the visualization of the data distribution using histograms. In this case, the variable of interest is “Population”, consequently I used the graphic type mentioned before to observe the frequency of occurrences in the data.

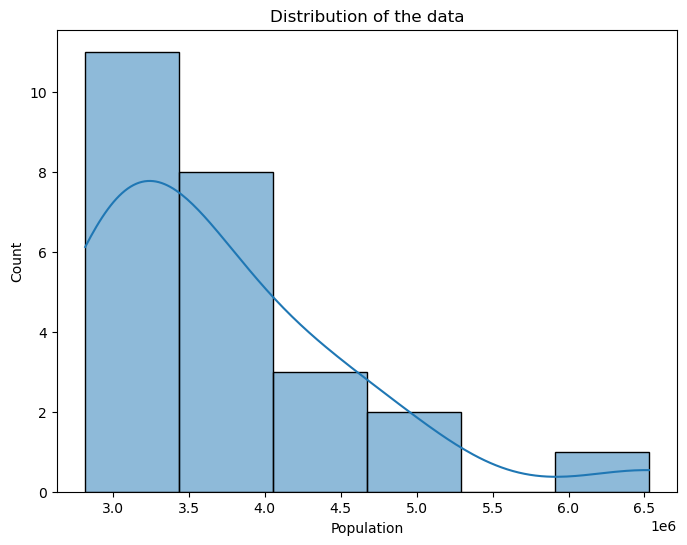


Figure 6: Distribution of the data for Population.

The values are grouped into bins on the x-axis and the height indicates how many values of the dataset fall into that bin. Having said that we can observe that the data has more years that have a population value of approximately three million inhabitants it decreases as the population number increases.

* *Shape of the date and its measures*

Skewness

Visualizing the histogram, I could identify the data studied has a Right-skewed distribution (Positively-skewed), the long tail is on the right side of the distribution, and there is an extreme value far from the peak. Having a right skew data means that the mean is greater than the media and the mean overestimates the most common values in a positively skewed distribution.

For the changes in the mean over or underestimating the most frequently occurring values in asymmetric distributions, analysts recommend using the median in these cases.

Numerically the coefficient of skewness “S” has to be greater than 0 so the distribution can be considered as positively skewed. The calculus of this coefficient reassures what is evident in the graph.

|  |  |
| --- | --- |
| Coefficient of skewness (S) | Value |
| S | 1.633 |

Modes

The distribution is unimodal due to the existence of one peak in the distribution.

Kurtosis

This statistical measure quantifies the shape of a probability distribution. It provides information about the tails and peakedness of the distribution compared to a normal distribution. Observing the histoplot and the data distribution considering the parameters given by the literature I concluded the kurtosis in this case was Leptokurtic.

This type of kurtosis has a very long tails, which means there are more changes of outliers.

Numerically the kurtosis has to be greater than 3 to be considered Leptokurtic.

The calculus of this coefficient reassures what is evident in the graph.

|  |  |
| --- | --- |
| Kurtosis (K) | Value |
| K | 3.279 |

* *Central tendency and its measure*

After analysing the shape of the data, I concluded as stated before that the best measure of tendency was the media as the mean seems to be affected by the skewness of data. I also included the min, max, and range of the population to enrich the knowledge about the data in the study.

|  |  |
| --- | --- |
| Parameter | Value |
| Median | 3,468,694.0 |
| Max Value | 6,528,799.00 |
| Min Value | 2,818,341.00 |
| Range | 3,710,458.00 |

* *Spread, variability and its measure*

Quantiles

Quantiles are values that split sorted data or a probability distribution into equal parts.

Quartiles are three values that split sorted data into four parts each with an equal number of observations.

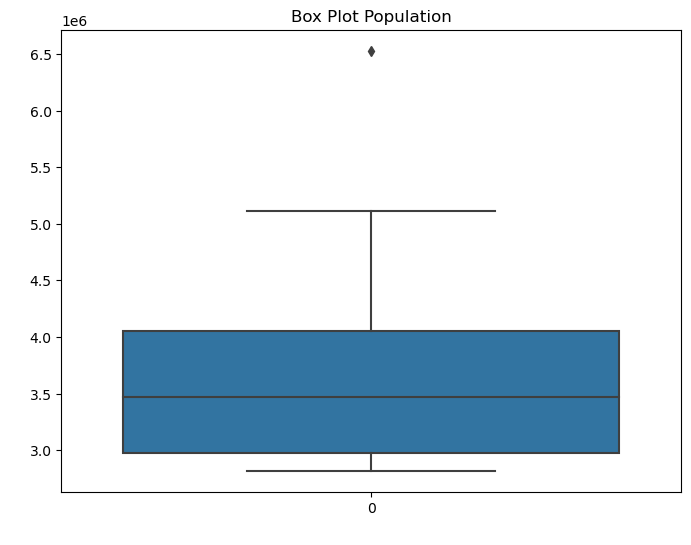


Figure 7: Box Plot Population.

|  |  |
| --- | --- |
| Quartile / Calculation value | Value |
| q1 | 2,971,992.0 |
| q2 | 3,468,694.0 |
| q3 | 4,053,187.0 |
| iqr | 1,081,195.0 |

A boxplot is a standardized way of displaying the distribution of the data based on the following numbers; minimum, first quartile (q1), median, third quartile (q3), and maximum. It also helps identify outliers. This shows graphically the distribution of the data confirming once again the skewness of the data to the right.

I calculated the values with the corresponding functions in order to calculate the Interquartile range (iqr) which represents the range between the 25th and 75th percentile, showing the spread of the middle half of the distribution.

Standard deviation

Measures of variability and spread summarise the data in a way that shows how scattered the values are and how much they differ from the median value in this case. It gives an idea of how well the mean or media represents the data. The larger the number the less representative the central tendency parameter chosen is.

As shown below standard deviation gives no useful information for skewed data. It is considered of better to use the first and third quartile since these give some sense of the asymmetry of the distribution.

|  |  |
| --- | --- |
| Standard deviation (Std) | Value |
| Std | 877,986.30 |

*Inferential Statistics*

*Testing normality*

One of the most common assumptions for statistical tests is that the data used are normally distributed.

Normal distribution can be tested either analytically (statistical tests) or graphically.

Graphical verification

Q-Q plot

This type of plot is a technique to analyze comparing two distributions (the normal one and the actual distribution of the data) by plotting their quantiles against each other.

It shows a straight line from each point corresponding to the normal distribution, if the plotted points of the actual dataset meet the straight line it can be assumed it corresponds to a normal distribution.

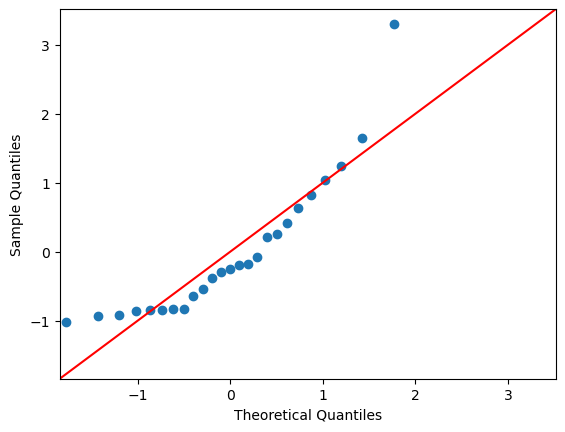


Figure 8: Q-Q Plot – Testing normality of the data

In the graph shown before it can clearly be seen that the points from the dataset do not follow a normal distribution. This plot provides even more information, as it can be observed a deviation on the right side which is consistent with the conclusion made before in relation to the Right skewness of the data.

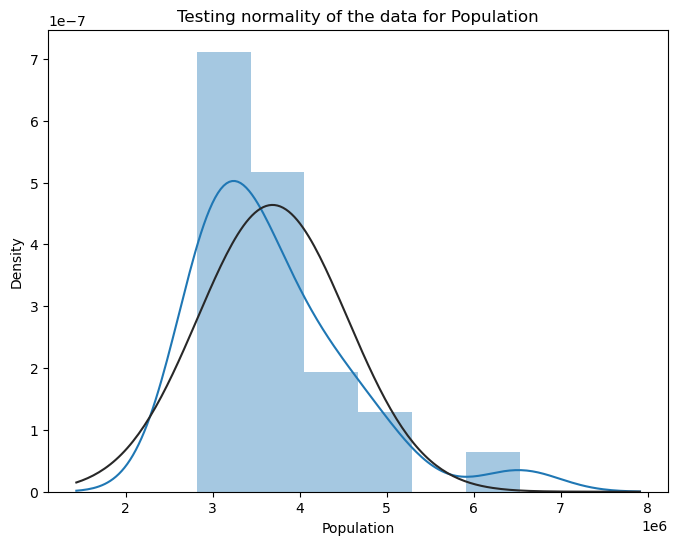


Figure 9: Testing normality of the data.

Figure 9 above compares the distribution of the data for the attribute of Population with the normal distribution of that which has the bell shape, it is clear that the data is right skewed, and the plot works to reassure what has been calculated and proved before.

Statistical Test

Shapiro – Wilk Test

This test of normality determines whether the given data is normally distributed or not. The null hypothesis of Shapiro’s test is that the population is distributed normally.

I elaborated on the following hypothesis:

H0 (Accepted): The data are normally distributed (Pvalue > 0.05)

Ha (Rejected): The data are not normally distributed

|  |  |
| --- | --- |
| Parameters | Value |
| Statistic | 0.842 |
| Pvalue | 0.001 |

The chart above resumes the results of using the shapiro() function, Pvalue is the parameter of interest and it gives a value of 0.001 which is not greater than 0.05 as a result the null hypothesis is rejected thus the data studied are not normally distributed.

*Probability distributions*

A probability distribution is a mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.

There are two types of probability distributions:

* Discrete probability distributions
* Continuous probability distributions

Having categorical and discrete variables guided me to conclude I could essay some discrete probability distributions.

A discrete distribution describes the probability of occurrence of each value of a discrete variable. Each possible value of the discrete variable can be associated with a non–zero probability in a discrete probability distribution.

Binomial Distribution

A binomial probability distribution is when there is only a probability of two outcomes. In this distribution, data are collected in one of two forms after repetitive trials and classified into either success or failure. In this scenario I decided to use my data expanded, in other words, the one that has the population divided in “County” and “Sex”, as the dataframe seemed to be a good fit for a test like this.

Problem Formulation

What is the probability of getting correctly identified a Male or Female (in Dublin) on the following scenarios of success (0 to 10) given that each scenario takes into account n (10) trials?

Considerations

Possible outcomes: Male or Female (0 or 1 respectively).

n: number of times the experiment runs (10 for the analysis).

p: the probability of one specific outcome (“y” individual classifies as Male or Female for the population of the county “Dublin”).

x: range of (0,10) success to run the experiment through.

Figures 10 and 11 the results obtained from the test described before. The interpretation is, for instance, that there’s a probability of 24% of 5 people being a man out of 10 people.

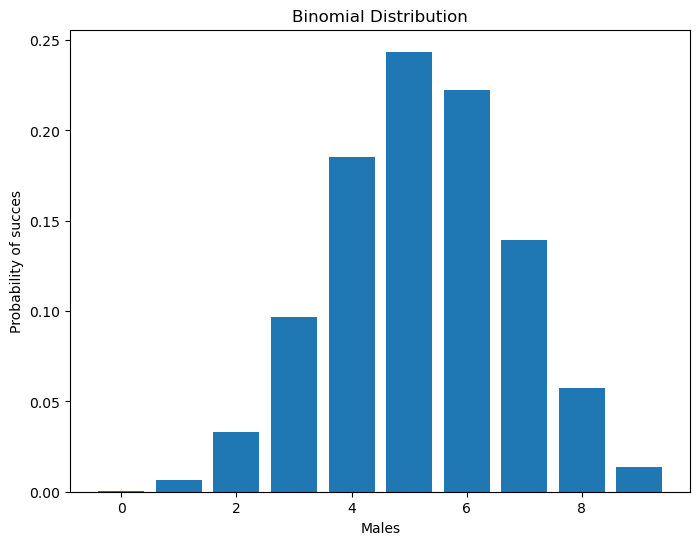


Figure 10: Testing normality of the data.

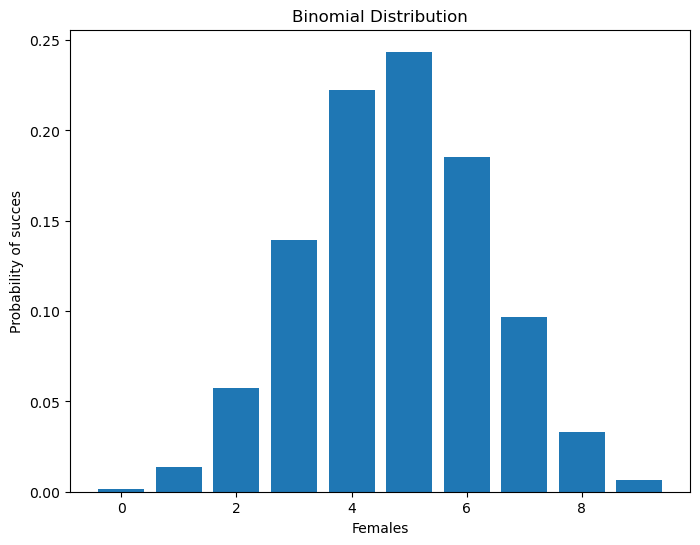


Figure 11: Testing normality of the data.

### Machine Learning

Machine learning is about extracting knowledge from data. It is a research field at the intersection of statistics, artificial intelligence, and computer science and is also known as predictive analytics or statistical learning. (Müller and Guido, 2017).

*Project Management Framework*

Project management frameworks ensure consistency in the process followed in the development of a project. Such frameworks facilitate to deliver of projects in a consistent manner, achieving faster results.

*CRISP-DM*

CRISP-DM, which stands for The Cross Industry Standard Process for Data Mining, is a comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project. (Shearer, 2000)

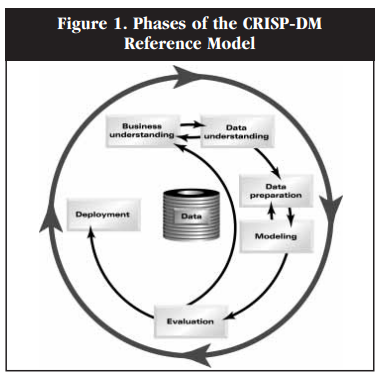


Figure 12: Phases of the CRISP-DM framework.

The image above shows the 6 phases CRISP-DM organizes the data mining process into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The arrows indicate the most important and usual dependency between the phases, while the outer circle symbolizes the cycle nature of data mining.

The highlighting aspect of the CRIPS-DM methodology is that emphasizes the importance of understanding the business context and objective before moving to the data analysis itself. Starting with a business understanding phase, the methodology ensures the analyst aligns the investigation with the needs of the organization.

*Machine Learning Techniques*

Supervised Machine Learning techniques were chosen in the study as the dataset had the data already categorized and labeled which makes it optimal to make classifications or predictions. Supervised Learning has the following characteristics: is less complex, predicts outcomes for new data, gives more accurate results, and the input and output variables are given.

Successful machine learning algorithms are those that automate decision-making processes by analyzing known examples. In these models, the analyst provides the algorithm with pairs of inputs and desired outputs, and the algorithm finds a way to produce the desired output given an input.

*Machine Learning Models*

Having recognized the type of data gives a direction of which kind of models or analysis. In this case, the appropriate model is about Classification or Regression. The image below shows a general idea of the type of machine learning.

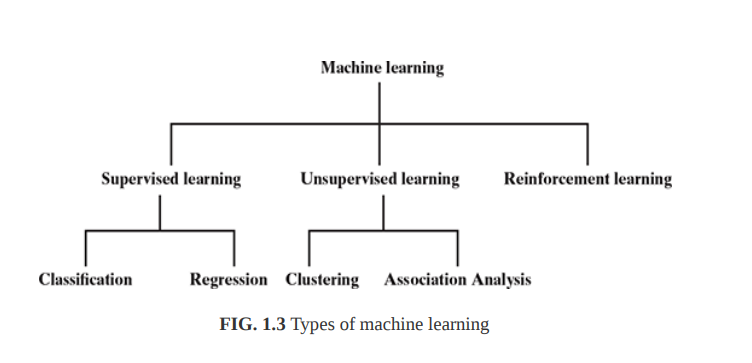


Figure 13: Types of machine learning (Dutt, Chandramouli and Kumar Das, 2019)

Linear Regression

Regression is a set of techniques for estimating relationships, once we have acquired data with multiple variables, we try to figure out how the variables are related. (Saleh, 2022)

In this particular case, I decided to perform a Linear Regression model to verify the relationship between my two variables (“Year” as the independent variable and “Population” as the dependent one) in order to get a model that could allow me to make predictions.

To begin I proceeded to split my dataset into train data and test data and I trained my model with LinearRegression() from the sklearn library. To verify the accuracy of the model I checked on the R2 score (the coefficient of determination) and I plotted the results.

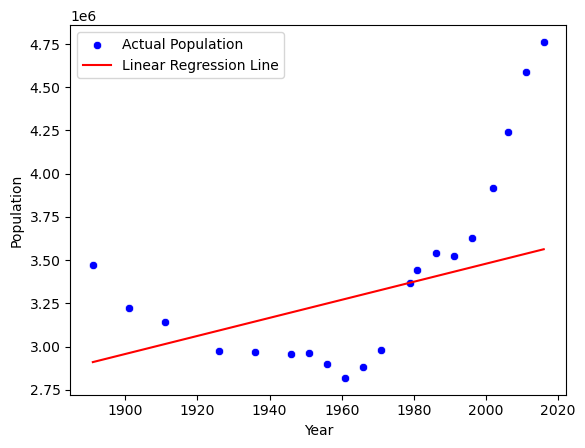


Figure 14: Linear regression model.

As the graph shows above there is not linear relationship between these two variables making the model not efficient for prediction. To try to make the model useful I tried to use a lagged variable of the population feature, adding 5 new columns with population values, giving the model more data to search for a relationship.

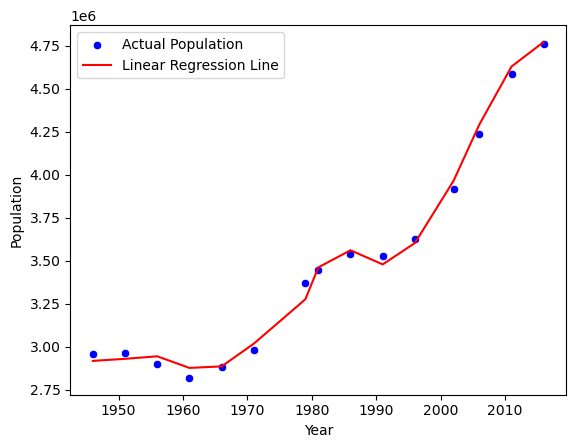


Figure 15: Lagged Linear regression model.

As shown above, the model improved by getting a training score of 89% and a test score of 54%, it seems that with the lagged values the model works as a proper linear regression. As the essays showed that linear regression does not make a complete fit with the dataset, I decided to analyze the polynomial regression.

Polynomial Regression

Polynomial regression is a technique based on a trick that allows the use of linear models even when the dataset has strong non-linearities. The idea is to add some extra variables computed from the existing ones and using (in this case) only a polynomial combination. (Saleh, 2022)

The polynomial regression model showed a better performance getting a training score of 97% and a test score of 96%, which led me to believe I could get good predictions using it.

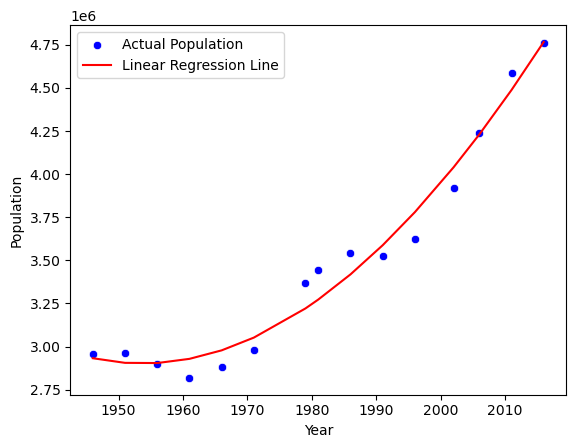


Figure 16: Polynomial regression model.

Ridge

Ridge regression is a statistical technique that is used when analyzing regression data or models that suffer from a condition known as multicollinearity or ill-conditioning. It is a method of estimating the coefficients of multiple regression models in scenarios where the independent variables are highly correlated.

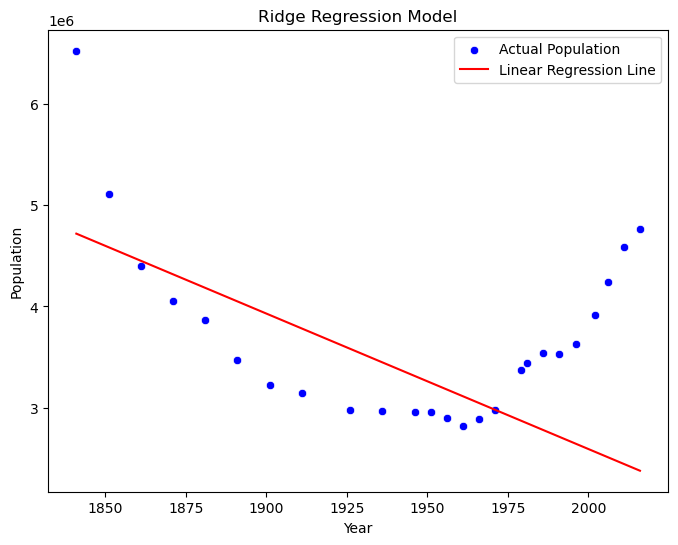


Figure 17: Ridge regression model.

The Ridge regression does not look accurate even graphically the training and test results throw numbers that reveal that the model can not obtain good results with the dataset provided to train.

Decision Tree regression

The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility.

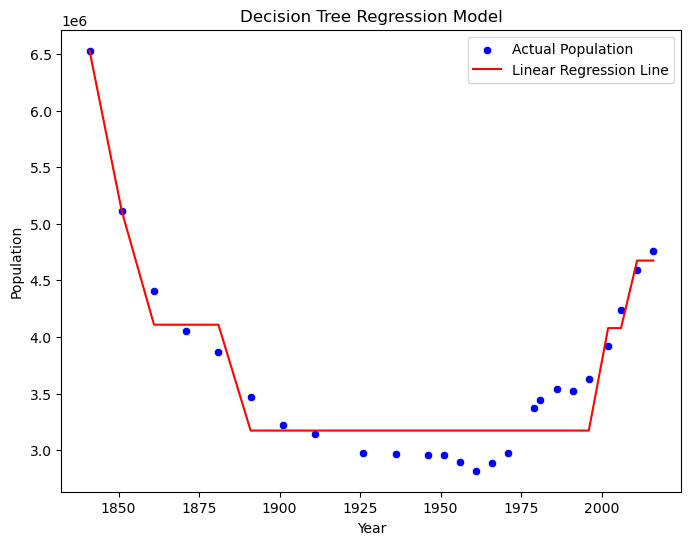


Figure 18: Decision tree regression model.

Decision tree regression shows good training and test scores however due to the spirit of the model and the type of dataset I would think more investigation will be needed to conclude that this model is accurate in this particular case.

Having performed the models shown before I collected all the training and test scores to compare the results all together.

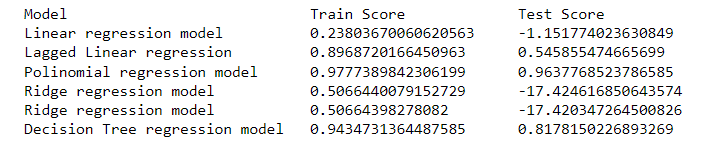


Figure 19: Train and Test results from the models/

## References

Ott, L. and Longnecker, M. (2016). An introduction to statistical methods & data analysis. Boston, Ma: Cengage Learning.

Müller, A.C. and Guido, S. (2017). Introduction to machine learning with Python: a guide for data scientists. Beijing: O’reilly.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4).

‌

Dutt, S., Chandramouli, S. and Kumar Das, A. (2019). Machine Learning. Pearson India Education Services Pvt. Ltd.

Saleh, H. (2022). Machine Learning - Regression.

‌

‌

‌