A picture containing icon

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for Data Analytics*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *Exploring the Link Between Population Trends and Crime Rates in Ireland* |
| **Lecturer Name:** | *Marina Iantorno/Taufique Ahmed*  *Sam Weiss/David Gonzalez*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Jose Mario da Cruz Costa |
| **Student Number:** | 2023422 |
| **Assessment Due Date:** | *5th Jan 2024* |
| **Date of Submission:** |  |

**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. |

**Analysing Transportation System in Ireland, Focused on Bike Sharing in Dublin and Comparing to equivalent service in Barcelona - Spain.**

# Abstract

*This project is available on github:*

*link:* https://github.com/jmdtanalyst/MSC\_DA\_CA2\_Transport\_Ireland

***KEYWORDS:*** *population growth, crime rates, Ireland, crime prevention, policy*

Word count: 3.232

Table of Contents

Abstract 2

Introduction 4

1.0 Data Preparation and Visualization 5

1.1 Data Wrangling 5

1.2 Exploratory Data Analysis (EDA) Method and Insights 5

1.2.1 Performing the EDA: 6

1.2.2 Exploring Crime Rate dataset: 10

2.0 Statistics for Data Analytics 11

2.1 PFM and CDF 12

2.2 Crime Rate Statistics 14

3.0 Machine Learning for Data Analysis 15

3.1 Handling Outliers and Skewness: 15

3.2 Principal Component Analysis (PCA): 19

3.3 Predicting Population growth: 22

4.0 Programing 23

4.1 Programming Paradigms 23

4.2 Benefits of using different programming paradigms. 25

5.0 Bibliography 25

# Introduction

DUBLIN

As of today, December 5, 2023, there are three bikeshare services operating in Dublin:

Dublinbikes: This is the oldest and most well-established bikeshare service in Dublin, having been launched in 2009. It is operated by TFI Cycles, a subsidiary of the National Transport Authority (NTA). Dublinbikes has over 3,500 bikes and 450 docking stations located throughout the city center and inner suburbs.

Bleeper: This is a stationless bikeshare service that was launched in Dublin in 2018. It is operated by Bleeper Bike, a private company. Bleeper bikes can be unlocked and locked using a smartphone app. There are currently over 800 Bleeper bikes available in Dublin.

Moby: This is an electric bikeshare service that was launched in Dublin in 2023. It is operated by Moby Bikes, a private company. Moby bikes can be unlocked and locked using a smartphone app. There are currently over 300 Moby bikes available in Dublin.

All three bikeshare services are convenient and affordable ways to get around Dublin. Dublinbikes is the most extensive service, with the largest number of bikes and docking stations. Bleeper is a good option for shorter trips, as there are no docking stations to worry about. Moby is a good option for longer trips or for those who want the assistance of an electric bike.

BARCELONA

Bicing is the older of the two services, having been launched in 2007. It is operated by Clear Channel and has over 6,000 bikes and 400 docking stations. Bicing is a public service and is intended for short-term journeys of up to 30 minutes.

AMBici is a newer service that was launched in 2023. It is operated by Transports Metropolitans de Barcelona (TMB) and has over 2,600 bikes and 236 docking stations. AMBici is an electric bike-sharing service and is also intended for short-term journeys of up to 30 minutes.

Both Bicing and AMbici are convenient and affordable ways to get around Barcelona. However, there are some key differences between the two services. Bicing is a more established service with a wider network of docking stations. AMbici is a newer service with a smaller network of docking stations, but it offers electric bikes.

Ultimately, the best bike-sharing service for you will depend on your individual needs. If you are looking for a convenient and affordable way to get around Barcelona for short-term journeys, then either Bicing or AMbici would be a good option.

Identify the most popular stations based on the number of available bikes: This can be done by calculating the average number of available bikes per station and identifying the stations with the highest and lowest average.

The most popular station based on the number of available bikes is: Smithfield

Analyze the trend of the available bikes over time: This can be done by plotting the number of available bikes against the time of day. This can help you to identify patterns in bike usage, such as peak hours and periods of low demand.

The average number of available bikes per hour:

Identify the stations with the highest and lowest number of available bikes: This can be done by creating a table or chart that shows the number of available bikes at each station for a given date or time period.

The station with the highest number of available bikes:

The station with the lowest number of available bikes:

In addition to these basic analyses, you can also use this dataset to answer more specific questions about bike usage in Dublin, such as:

How does bike usage vary by day of the week?

How does bike usage vary by season?

What factors are associated with higher bike usage?

By analyzing this data, you can gain valuable insights into the patterns of bike usage in Dublin and use this information to improve the bike-sharing system and make it more efficient and user-friendly.

# Data Preparation and Visualization

## 1.1 Data Wrangling

Data Wrangling is the first step to be performed with the data. This process consists of collecting, cleaning, transforming, and manipulating raw data to make it usable for analysis.

For any data analysis project, the data collection must be performed in a trustable source.

The datasets for DublinBikes was collected on data.gov.ie, an Irish government trustable source. The data was downloaded in CSV file, under the “Creative Commons Attribution 4.0 (CC BY 4.0)” Licence.

The advantage of this format is that the data is easy to be collected and this licence allows to copy and use the data for free, one disadvantage is that the data isn’t automatically updated, requiring manual download and analysis.

for the station information in Barcelona, the data was collected from the website: opendata-ajuntament.barcelona.cat, also a government website, and it was collected through an API, in JSON format. The advantage of using API for data collection is that the data is regularly updated, ensuring access to the latest data. The advantage is that API access and technical expertise is required.

This dataset is also under “Creative Commons Attribution 4.0 (CC BY 4.0)” Licence.

For reviews of bikestation(Dublin) and bicing (Barcelona), a TripAdvisor API was used. The advantage of using an API for data collection is that the data is regularly updated, ensuring access to the latest data. However, API\_KEY must be managed and stored properly.

Licence: License Creative Commons Attribution 4.0

https://creativecommons.org/licenses/by/4.0/

## 1.2 Exploratory Data Analysis (EDA) Method and Insights

* Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.**[0-25]**

For the historical dataset, a multivariate analysis was applied. It is used to identify patterns and trends in the data. Also, PCA - principal component analysis was applied reduce the dimensionality of the dataset and identify the most important variables.

Data Cleaning and Preparation

Duplicate rows from the dataset were removed, there were no missing values in the dataset, no additional data cleaning was needed.

* Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.**[0-30**]

For machine learning purposes, The status column was changed to numerical variables, where 1 to open and 0 to closed. I also scaled the numerical variables BIKES AVAILABLE and BIKE STANDS AVAILABLE to a standard normal distribution.

* Modern Transport planning has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern Transport planning, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development. **Note you may not use Powerbi, rapidminer, tableau or other such tools to accomplish this (at this stage).[0-30]**

Considering Tufte's principles of data visualization, a scatter Map was ploted, making easy to understand the pattern of bike station around the city.

The size of the bubble is proportional to the number of bikes available, it will provide a clear bike availability at each station, also, the color of the bubble was chosen to represent the open and close station.

A map of a city

Description automatically generated

Insights from the map:

In the figure, it is observed that the concentration of bike stations is in the city center.

Almost all stations are operational (green bubbles), only two (red bubbles) are closed

## 1.2.1 Performing the EDA:

This process will be performed on Jupyter notebook: **MSC\_DA\_CA1\_Jose\_Mario.ipynb** and it was called **“**Phase 02 - EDA ”

**Skewness:**

Skewness is a measure of how much the distribution of a random variable deviates from symmetry.

It is important to consider the skewness when performing statistical tests. In some statistical tests, if the data is skewed, these tests may not be valid.

Skewness is classified as follows:

Highly skewed: Less than -1 or greater than 1

Moderately skewed: Between -1 and -0.5 or between 0.5 and 1

Approximately symmetrical: Between -0.5 and 0.5

Histograms or distplots can be used to visualize skewness. However, distplots are generally considered to be a better option because they provide more information about the shape of the data distribution.

A group of blue and white graphs

Description automatically generated*Figure 01: distplot - Skewness by region*

**Insights**

From this figure, we can observe that aside from the Border region, which exhibits a highly positive skew, the skewness of the data is relatively mild. Therefore, it is not necessary to perform any transformations on the data for the purpose of data visualization.

Below, the print of the skew value for each region:

A screenshot of a computer

Description automatically generated

**BoxPlot:**

We can get some insights from the boxplot and answer a variety of questions about the data, for example: Identify the median value of the data, determine whether the data is skewed, identify outliers and compare the distributions of two or more groups of data.

A chart with different colored boxes

Description automatically generated  
*Figure 02: Boxplot - population by region*

**Insights:**

From this boxplot we can get some Insights:

The median population of the "Border" region is the highest, also the boxplot shows that there is a significant variation in population between the different regions.

The Border and West regions have the highest populations, while the South-West region has the lowest population.

It is possible to detect some outliers on the Border region, what require further investigation.

**Outliers:**

Outliers are data points that significantly differ from the majority of observations in a dataset. They can be caused by various factors, including measurement errors, data entry errors, or natural variation in the population.

To identify potential outliers, statistical methods like the Z-score or the Interquartile Range (IQR) can be employed. The Z-Score Method is particularly useful when dealing with data that follows a normal distribution. In contrast, the IQR method is robust and less sensitive to extreme values, making it well-suited for data that is skewed or non-normally distributed.

Considering that the 'Border' column in this dataset is highly skewed, the IQR method is the more appropriate choice for identifying potential outliers.

A screenshot of a computer

Description automatically generated

Although the values are potential outliers, observing the histological population growth in the “Border” region, it appears to be a natural growth of the population.

**Line Chart:**

Line charts are a useful tool for detecting trends, relationships, and comparisons.

To visualize the grown of the population, a line chart has been created.

A graph of different colored lines

Description automatically generated  
*Figure 03 – Population by region over time*

**Insights:**

One of the most obvious insights is the overall growth of population over time, The chart shows the regional variation in population growth. Some regions, such as Dublin, have experienced more rapid growth than others.

Also, we can observe some anomalies in the regions "Border" and "South-East", which show a decrease in population.

Considering the trend line, and furthermore using Machine Learning, we can get insights into the future of population.

**Pie Chart:**

Pie charts are also useful for detecting trends, relationships, and comparisons, but it is important to limit the number of items in the chart to avoid creating a cluttered image.

Pie charts use a 1D array as input, so we need to convert the data variable to a 1D array using the NumPy ravel() function.

In this case, we will plot two pie charts: one with data from 2013 and one with data from 2023, so that we can observe any changes over the 10-year period.

A close-up of a pie chart

Description automatically generated*Figure 04 – Population by region 2013 and 2023*

**Insights:**

From the pie chart, we can observe that the biggest region in population Is the Dublin region, and the smallest is Midland.

Comparing both pie charts, we can see the following changes:

The Dublin region has become even more populous over the past decade, increasing its population share from 32.1% to 33.2%.

The Border region has also experienced a slight increase in population share, from 16.5% to 16.9%.

Overall, the population of Ireland is becoming more concentrated in the Dublin region. This may be due to a number of factors, such as job opportunities, educational opportunities, and cultural amenities.

## 1.2.2 Exploring Crime Rate dataset:

From the crime dataset, we can get some information.

Which Garda Station has registered more incidents, and which one has registered less incidents:

|  |  |
| --- | --- |
| **More Incidents:**  A screenshot of a phone  Description automatically generated | **Less Incidents:** |

Also, we can have an overview of the Type of Offences

A graph with text and numbers

Description automatically generated  
*Figure 05 – Overview of type of offences*

**Insights:**

From this plot, we can observe that the most prevalent offenses in Ireland are theft and related offenses, public order and other social code offenses, and damage to property and the environment. Conversely, kidnapping and related offenses are the least common type of offense.

# 2.0 Statistics for Data Analytics

**Notes:**

Demonstrate the confidence interval for each column

Explain about the correlation between the variables

Increase the size of the text in markdown

Use inferencial statistics

Define a hypothesis and apply the hypothesis.

* Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. **[0-20]**
* Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with public transport, you could find a confidence interval for the population proportion of users commuting to Dublin by train). **[0-20]**
* Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests. **[0-40]**
* Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process. **[0-20]**

*Note: All your calculations and reasoning behind your models must be documented in the report and/or the appendix.*

We can use a set of tools and techniques for collecting, organizing, summarizing, analysing, and interpreting data.

Statistical techniques were employed in this project within the Jupyter notebook named “MSC\_DA\_CA1\_Jose\_Mario.ipynb”, encompassing both “Phase 3, which focused on statistics, and Phase 4, which centred on machine learning”.

By using the method describe, we can get important information from the result.

**Mean**: The average of crime over these types.  
**Max**: The largest value in the type of crime.  
**Median**: (middle value) of recorded crime incidents.  
**Minimum**: and maximum values in the dataset.  
**Standard deviation** :to measure the spread of the data.

Applyiing the .describe() method on crime\_df:

A screenshot of a computer code

Description automatically generated

From this “.describe()” method, we can get some insights:

* The average number of crimes in the dataset is 34.78.
* The standard deviation is 143.66, which means that the number of crimes varies widely.
* There are a few outliers in the dataset, with the maximum number of crimes being 6523.
* The median number of crimes is 4, which means that half of the crimes in the dataset are less than or equal to 4, and half are greater than or equal to 4.

Histograms provide a visual representation of the distribution of data, it is useful for understanding the characteristics of the data.

A graph with numbers and lines

Description automatically generated

*Figure 06 – histogram of crime\_df*

**Insights:**

The distribution of the VALUE column is skewed to the right. This means that there are more values on the right side of the distribution than on the left side.

The median value of the VALUE column is between 10,000 and 15,000.

## 2.1 PFM and CDF

Pmf and Cdf are classes that represent probability mass functions (PMFs) and cumulative distribution functions (CDFs), respectively.

* A PMF is a function that gives the probability of each possible value of a discrete random variable.
* A CDF is a function that gives the probability that a random variable will take on a value less than or equal to a given value.

The best approach to visualize the PMF is plotting a wide bar chart.

A graph with blue and black text

Description automatically generated *Figure 06 – pmi of Type of Offence*

To get the exact PMI values, we can also explore the PMI itself by printing the values.

The highest probability:

A computer screen shot of a computer code

Description automatically generated

The lowest probability:

A close-up of a computer screen

Description automatically generated

**Insights:**

Based on the PMF chart we can make the following analyses:

The most common type of offence is theft and related offences with 30%. This is followed by public order and other social code offences

The least common types of offences are kidnapping and related offences.

The distribution of offences is skewed, with a few types of offences being much more common than others.

**Exploring CDF:**

A CDF is a function that gives the probability that a random variable will take on a value less than or equal to a given value.

A graph with numbers and lines

Description automatically generated*Figure 07 – CDF of crime\_df*

**Insights from the CDF:**

The CDF chart shows that the VALUE column is skewed to the right. This means that there is a relatively small number of crimes with very high values, and a large number of crimes with lower values.

Most crimes have relatively low values.

There is a small number of crimes with very high values.

## 2.2 Crime Rate Statistics

**Evolution of top 5 type of offence from 2003 to 2022:**

A graph with different colored lines

Description automatically generated

*Figure 08 –Evolution of top 5 type of Offence 2003 to 2023 Q2*

**Insights:**

The chart shows that all five types of offences have increased over time. Theft and related offences has shown the smallest increase, followed by public order and other social code offences. The other three types of offence have all shown larger increases. After 2021, all 5 types of offences started to increase, with theft and related offences leading the increase.

**Population Growth vs Crime Rate**

To compare population growth and crime rates, we will filter the population dataframe to extract total population figures between 2003 and 2022, excluding regional data. We must then multiply the population by 1000 to account for the fact that the population is shown in thousands and crimes are shown in units.

To compare the evolution of the population and crime rate, a line chart has been plotted.

A graph of a line and a line

Description automatically generated with medium confidence  
*Figure 09 – Line Chart comparing total crime and population growth*

**Insight**

From the figure above, we can see that the crime rate in Ireland has declined significantly over time, while the population has grown steadily.

This suggests that the crime rate in Ireland has declined even faster than the population has grown. This is a positive development.

**Correlation**

To get a better understanding of the relationship between population growth and crime rate, we can calculate the correlation coefficient between the two variables.

A screenshot of a graph

Description automatically generated  
*Figure 10 – Heatmap correlation coefficient population and crime*

**Insight from Correlation Heatmap**

The correlation coefficient for the population growth and crime rate dataset is --0.582652. This indicates a moderate negative correlation between the two variables. This means that as the population increases, the crime rate tends to decrease.

# 3.0 Machine Learning for Data Analysis

**Machine Learning Tasks**

Use of multiple models (at least two) to compare and contrast results and insights gained.

* Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.

**[0 - 30]**

* Collect and develop a dataset based on the transport topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate transport topic (e.g., public transport, freight movement etc…) for producers and consumers point of view in Ireland.

**[0 - 25]**

* You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.

**[0 - 30]**

* A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.

**[0 - 15]**

For the purpose of developing a machine learning model, I will utilize the new generated dataset, Pop\_and\_crime\_df.csv. To prepare the data for Machine Learning, I will standardize the crime rate providing the crime rate per 100,000 people, which is a common way to standardize crime rates for comparison.

I will use CRISP-DM project management framework, applying multiple machine learning models and examine the performance of each model.

These tasks will be performed in the jupyter notebook: MSC\_DA\_CA1\_Jose\_Mario.ipynb, Phase 4 - Machine Learning.

## 3.1 Handling Outliers and Skewness:

In the Machine Learning phase, it is crucial to address skewness, handle outliers, and standardize the data.

To identify potential outliers, statistical methods like the Z-score or the Interquartile Range (IQR) can be employed.

Considering that this dataset is highly skewed, the IQR method is the more appropriate choice for identifying potential outliers.

After applying the IQR method, we have a table with the rows with outliers:

A screenshot of a computer

Description automatically generated

For skewness, as we have more columns, I've implemented a “For” loop to identify and print the skewness only when it exceeds the threshold.

**Skewness vs outliers:**

A screenshot of a computer code

Description automatically generated

As a result, we have 4 skewed columns, with some right skewness and others left skewness. Also, we can observe that there are 5 rows with outliers.

Considering that the dataframe contains 20 rows, removing outliers would eliminate 5 rows. In this scenario, the preferred approach is to transform the data to mitigate skewness rather than removing outliers.

For data transformation, we need to understand the skewness, for right skewness we can apply some transformation methods, whereas for left skewness, we need to apply another transformation method.

For right skewness, I’ve applied the following transformation methods:

**Square Root Transformation:**

Square root transformation is suitable for right-skewed (positively skewed) data, it has better performance in mildly right-skewed.

A screenshot of a cell phone

Description automatically generated

After applying square root, the result was this:

As the result, them column with mild skewness was adjusted, but the columns with high skewness was not.

**Cube Root Transformation:**

Cube rooting can help reduce the influence of extreme values or outliers in the data, making the data more robust to the presence of outliers.

After the cube root transformation, the outcome was as follows: The cube root proved to be more effective than the square root in reducing skewness, but it was still not sufficient to fully adjust the highly skewed column.

A screenshot of a computer

Description automatically generated

**Log Transformation:**

Log transformation is a common way to handle right-skewed data. It can be used to improve the normality of the data, which is a requirement for regression models.

A screenshot of a computer

Description automatically generated

Log transformation proved to be more effective than the square root and cube root in reducing skewness, but it was still not sufficient to fully adjust the highly skewed column.

**Reciprocal Transformation:**

A screenshot of a computer

Description automatically generatedReciprocal transformation is effective in reducing the impact of outliers, especially for data with right-skewness.

Even after handling the skew with reciprocal transformation, the data was reduced to zeros, which is not useful for regression.

The best approach was applying both Log transformation and Reciprocal transformation consecutively, which handled the skewness successful, and it is still suitable for regression.

A screenshot of a computer

Description automatically generated

**Left Skewness:**

Now we have only the columns with negative skewness (left skew). For left skewness, we have some transformation methods:

* Square Transformation;
* Cubes Transformation;

**Square Transformation:** The square transformation is effective in reducing left skewness. It's the counterpart of the square root transformation and can be used when the data is left-skewed.

A screenshot of a computer

Description automatically generated

**Cubes Transformation:** Exponential transformations can be applied to data with negative skewness to expands small values and compresses large values, which will make the distribution more symmetric.

A screenshot of a computer

Description automatically generated

Applying square or cube transformation was effective at reducing skewness, but it was not enough to fully adjust the highly skewed columns. The best option is to apply both transformations.

A screenshot of a computer

Description automatically generated

Now we have our data cleaned and more symmetric, we can apply some predictions and Machine learning models.

## 3.2 Principal Component Analysis (PCA):

As we have a dataset with 16 columns, it is a good approach to reduce the dimension, for this purpose, Principal component analysis (PCA) technique will be applied. PCA simplifies the complexity of the dataset while preserving its essential structure.

Applying PCA with 10 components to determine the optimal number of components for dimensionality reduction:

A graph with a red line

Description automatically generated  
*Figure 11 – Cumulative Explained Variance*

Based on this chart, we can assume that the best number of components is 5, because from 1 to 5 they explain over 90% of the variance in the data. After 5 components, the variance explained by each additional component drops off sharply, indicating that these components are not capturing much important information.

**Predicting Crime Rate after applying PCA.**

For predicting the crime rate, Its necessary to use Machine Learning to predict the number of crime in a given year.

As we know the features and the target of our data, we will apply supervised learning to predict the crime rate:

**Linear regression:**

Linear regression is a powerful model that can be used to predict continuous variables, such as population growth or crime rate, that is because it is used to model the relationship between a dependent variable and one or more independent variable.

For this purpose it’s the best option because the variables and the target are known.

After creating the linear regression model and applying, the residual plot can provide the accuracy of the model.

A graph with a line and dots

Description automatically generated with medium confidence  
*Figure 12 –Linear regression residual plot*

The residual plot shows that the residuals for linear regression model are randomly scattered around the zero line. This indicates that the model is a good fit for the data.

A close-up of a computer screen

Description automatically generated

**R-squared score (R^2)** is a measure of how well the model explains the variation in the data. A higher R^2 score indicates a better fit. The R^2 score on the training set is 1.000 and the R^2 score on the test set is 0.998.

**Mean squared error (MSE)**

The mean squared error (MSE) is a measure of how close the predicted values are to the actual values. It is calculated by taking the average of the squared differences between the predicted values and the actual values.

Based on the MSE and R^2 results, we can assume that the model is performing well, and can be applied in a new dataset.

**Lasso Regression:**

Appling Lasso Regression, we had also a good result, which means that this model can be applied in a new dataset:

A graph with a red line

Description automatically generated  
*Figure 13 – Lasso Regression Residual Plot*

A computer code with text

Description automatically generated with medium confidence

Similar to Linear Regression, Lasso regression also had a very good result. The training set score and test set score are both 1.00, which means that the model is perfectly fitting the data.

**Random Forest Regressor:**

To compare the performance of different models. I also applied Random Forest Regressor.

Applying the Random Forest Regressor model, this was the result:

A screenshot of a computer program

Description automatically generated

With a MSE very low (0.00563426) and an R-squared very high (0.99359531), this model also performs very well.

**Comparing the results:**

|  |  |  |
| --- | --- | --- |
| Linear Regression:  MSE train: 0.000, test: 0.002  R^2 train: 1.000, test: 0.997 | Lasso Regression:  Training set score: 1.00  Test set score: 1.00 | Random Forest Regressor:  MSE: 0.00563426  R-squared: 0.99359531 |

**Conclusion:**

All three models have performed well, Linear Regression performed the best. It perfectly fit the training data and performed well on unseen data. Lasso Regression and Random Forest Regressor also achieved good results, but they were slightly less accurate than Linear Regression.

After Understanding PCA e Linear Regression and Lasso Regression and Random Forest Regressor, we will now use only the principal features of our datasets to predict population growth and crime rates.

First, we will predict the population growth for the next 8 years using simple linear regression, then, we will use this result as an input to a lasso regression model to predict crime rates for the next 8 years.

## 3.3 Predicting Population growth:

For this purpose, I’ve created a new model, “lr\_pop” that utilize simple regression, since we have only the Year and Population.

After, I’ve created a new dataframe with a range of years from 2023 to 2030, and applied the model “lr\_pop” to predict the population.

A screenshot of a computer

Description automatically generated

Subsequently, I employed the outcome in a new model, "lasso\_crime," which estimates crime rates based on population demographics.

A screenshot of a computer screen

Description automatically generated

**Result**:

As a result, using linear regression and lasso regression, we could predict well that crime rates would decline over the next seven years. This is a positive trend, and it suggests that the model is able to identify factors that are contributing to the current crime rate and that these factors are likely to change in the future.

# 4.0 Programing

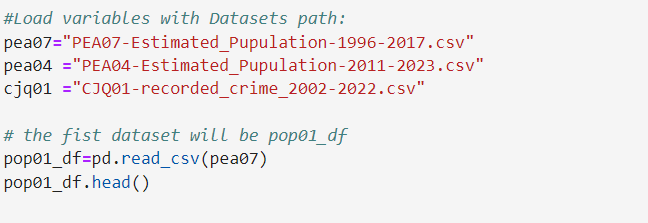
1. **Programming:** The project must be explored programmatically: this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. **[0-20]**
2. **Data structures:** You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format. **[0-20]**
3. **Documentation:** The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied. **[0-20]**
4. **Testing & Optimisation:** You are required to document and evaluate a testing and optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). Note any trade-offs that you've made in these areas. **[0-20]**
5. **Data manipulation:** For each of the different data sources, compare and contrast at least two relevant libraries and techniques for a) processing and b) aggregating the respective data, in order to justify your chosen libraries/techniques. **[0-20]**

## 4.1 Programming Paradigms

Through the development of this project, Imperative programming, functional programming and object oriented programming has been aplied.

Imperative programming was applied in tasks such: reading and writing data to and from CSV files, Performing data cleaning and preprocessing operations.

This is an example of imperative programming, it tells the computer how to perform the task step by step.

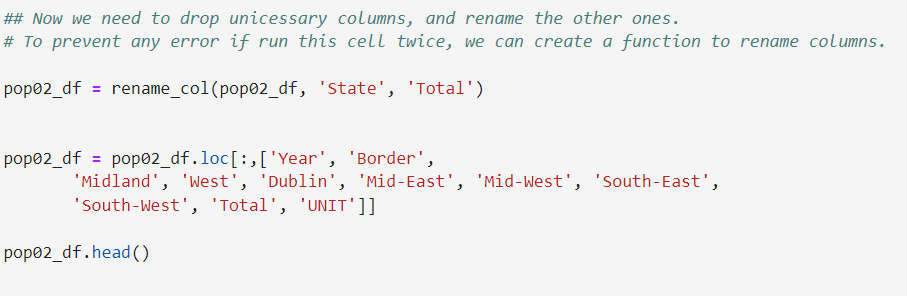


Functional programming was applied to tasks such: filtering and manipulating data frames using pandas and creating and applying custom functions, for example, rename\_col().

A screenshot of a computer code

Description automatically generated

Object-oriented programming: I used object-oriented programming to organize my code into reusable and maintainable classes and objects. For example, I reused the function to rename columns.



Programming paradigms was used to merging the population datasets, as the result of the merge was different columns, I used functional programming to create a function to rename the columns, this was one specific problem that I solved using programming knowledge.

A screenshot of a computer

Description automatically generated

Moreover, I created a function to calculate outliers and applied a for loop to execute this function in each column.

A screenshot of a computer program

Description automatically generated

## 4.2 Benefits of using different programming paradigms.

By using different programming paradigms will increase the performance of the code, as well as, the reusability.

Functional programming and object-oriented programming allow to create code more reusable and maintainable. For example, create functions and classes that can be used to perform common tasks, such as data cleaning and data visualization.

Furthermore, applying different programing paradigms and commenting the code, improve the readability, it makes the code more readable and easier to understand.

A screenshot of a computer program

Description automatically generated

# 5.0 Bibliography

References

--- data collection

Creswell, J. W., & Plano Clark, V. L. (2018). Introduction to research methods: A global perspective (5th ed.). Sage Publications.

Flicker, S. (2007). The human research subject: The ethics of participant observation. SAGE Publications.

Sieber, J. E. (1992). Planning ethically responsible research: A guide for students and faculty. Jossey-Bass.

**Tufte, E. R.** (2001). The Visual Display of Quantitative Information (2nd ed.). Cheshire, CT: Graphics Press.