**CCT College Dublin**

**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:** | *Sam Weiss/* *David Gonzalez*  *Bharathi Chakravarthi/* *Marina Iantorno*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Jose Mario da Cruz Costa |
| **Student Number:** | 2023422 |
| **Assessment Due Date:** | 10/11/2023 |
| **Date of Submission:** | 10/11/2023 |

**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. |

**Exploring the Link Between Population Trends and Crime Rates in Ireland**

**Abstract**

*Population growth can have positive and negative impacts in a country. On the positive side, population growth can lead to increased economic growth and innovation, in other hand, population growth can also lead to increased competition for resources and increased crime rates.*

*This project will explore the link between population trends and crime rates in Ireland. It will use a variety of data sources including population data from the Central Statistics Office of Ireland and recorded crime data from the Garda Síochána also found at Central Statistics Office of Ireland.*

*The study will focus on the following key questions:*

* *How has the population of Ireland changed in recent years?*
* *Which region has changed most?*
* *Is there a relationship between population growth and recorded crimes in Ireland in the last 20 years?*
* *Population trends can be used to predict future crime rate Ireland.*

*To choose the best way to predict something, we are exploring and comparing different regression models. The models we are considering are multiple linear regression, lasso and ridge regularization, and decision tree regression.*

*As different regression models make different assumptions about the data. Comparing different models, I will find the one that produces the most accurate predictions for the dataset.*

*I also will use regression analysis to model the relationship between population growth and crime rate.*

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

Table of Contents

Abstract ……………………………………………………………………………………………………2

Table of Contents ………………………………………………………………………………...………3

Introduction ………………………………………………………………………………………………..4

1.0 Data Preparation and Visualization 4

1.1 Data Wrangling 6

1.2 Exploratory Data Analysis (EDA) Method and Insights 6

2.0 *Statistics for Data Analytics* 9

2.1 PFM and CDF 11

3.0 *Programming for Data Analytics* 11

3.1 Sample Report – Information Gathering 11

3.2 Sample Report – Service Enumeration 12

3.3 Sample Report – Penetration 13

3.4 Sample Report – Maintaining Access 17

3.5 Sample Report – House Cleaning 18

4.0 *Machine Learning for Data Analysis* 18

**Introduction**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Region and county** | | | | | | | |
| **Border** | **Midland** | **West** | **Dublin** | **Mid-East** | **Mid-West** | **South-East** | **South-West** |
| Cavan | Laois | Galway City | Dublin City | Kildare | Clare | Carlow | Cork City |
| Donegal | Longford | Galway County | Dún Laoghaire-Rathdown | Louth | Limerick | Kilkenny | Cork County |
| Leitrim | Offaly | Mayo | Fingal | Meath | Tipperary | Waterford | Kerry |
| Monaghan | Westmeath | Roscommon | South Dublin | Wicklow |  | Wexford |  |
| Sligo |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Source: https://www.cso.ie/en/releasesandpublications/ep/p-rsdgi/regionalsdgsireland2017/nt/

**1.0 Data Preparation and Visualization**

**1.1 Data Wrangling**

Data Wrangling is the first step to be performed with the data. This process consists of cleaning, transforming, and manipulating the data to make it more usable for analysis. This process will be performed in a Jupyter Notebook:

* File name: MSC\_DA\_CA1\_Jose\_Mario.ipynb
* Process: 'Phase 01 - Data Wrangling'

The project will use three datasets:

* CJQ01-recorded\_crime\_2002-2023.csv
* PEA04-Estimated\_Population-2011-2023.csv
* PEA07-Estimated\_Population-1996-2017.csv

It will be necessary to clean the data, organize, and rename some columns, as well as merge the datasets.

During the cleaning process, on “crime\_df” dataframe, one ‘NaN’ value was found in one row, filling it with the mean wouldn't be a suitable approach. Using the mean would distort the statistics because the mean would consider the other types of offenses in its calculations. The best approach is to drop the row, ensuring that the dataset contains only complete and reliable information.

The method .dropna() has been used in this dataframe.

**1.2 1 Exploratory Data Analysis (EDA) Method and Insights**

EDA allows us to gain an overall understanding of the dataframes, detect relationships between variables, and examine the distribution of the variables of interest. In this study, the exploratory data analysis (EDA) step will be used to perform both statistical analysis and visualization tasks.

The preferred approach is multivariate analysis, which explores the relationships between three or more variables. Specifically, I will examine the relationship between population and crime rates.

I chose these methods because they will enable me to gain insights into the overall trends in population, as well as the overall situation of the crime rate and its relation to population trends.

After the first process, “Data Wrangling”, the population\_df and crime\_df datasets are shown below:"

A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

**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.

**2.0 Statistics for Data Analytics**

We can use a set of tools and techniques for collecting, organizing, summarizing, analyzing, 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 centered 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:

* There is a wide range of 'VALUE' in the dataset, with values ranging from 0 to 20,707.
* The high standard deviation suggests that there is a significant dispersion of values.
* The median value of 109 indicates the central tendency, but the mean is higher at 1,560.19, suggesting potential skewness in the data.

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

A screenshot of a graph

Description automatically generated

*Figure 05 – histogram of crime\_df*

**Insights:**

There are some insights that you can get from the bin plot:

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.

Before getting the PMF and CDF information, its important to prepare the dataset.

To improve the clarity and visualization of the data, it’s important to add a new reference column that identifies each Type of Offence.

A screenshot of a computer code

Description automatically generated

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

A graph of a number of blue bars

Description automatically generated with medium confidence  
*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 code on a white background

Description automatically generated

The lowest probability:

A screenshot of a computer code

Description automatically generated

**Insights from this PFM**

The most common crime type is 67 "Theft and related offences", with a probability over 0.14.

The least common crime type is 34, "Infanticide", with probability of 2.08e-07 or 0.000000208

The distribution of crime types is skewed, with a few crime types accounting for a large proportion of the total number of crimes.

**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 of data on a white background

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 are more values in the dataset that are greater than the median value than there are values that are less than the median value. The median value of the VALUE column is approximately 10,000.

The CDF chart also shows that there are a small number of outliers in the dataset. Outliers are values that are much larger or smaller than the rest of the values in the dataset. The outliers in the VALUE column are greater than 15,000.

**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 2023, 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 graph and a graph of a graph

Description automatically generated  
*Figure 08 – Line Chart comparing crime rate and population growth*

**Insight**

The overall trend in the dataset shows that both population and crime have increased over time. However, the crime rate has increased at a slower rate than the population. This means that the crime rate per capita has decreased over time.

**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 screen shot of a computer

Description automatically generated

**Heatmap**

A black and white squares with white text

Description automatically generated*Figure 08 – Heatmap correlation coefficient popularion and crime*

**Insight**

The correlation coefficient for the population growth and crime rate dataset is -0.602567. 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*

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.

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

**Handling with 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 screen

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:**

As a result, we have 19 skewed columns, with some right skewness and others left skewness. Also, we can observe that there are 15 rowns with outliers.

Considering that the dataframe contains 21 rows, removing outliers would eliminate 15 rows. In this scenario, the preferred approach is transform the data to mitigate skewness rather than removing outliers.

A screenshot of a data

Description automatically generated

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 computer

Description automatically generated

After applying square root, the result was this:

As the result, some columns with mild skewness where adjusted, but the columns with high skewness were 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.

A screenshot of a computer

Description automatically generated

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 columns.

**Reciprocal Transformation:**

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

A screenshot of a computer

Description automatically generated

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 columns.

Applying Square Root and Reciprocal Tramsformation

A screenshot of a computer

Description automatically generated

The best approach was applying both the Square Root and Reciprocal transformations consecutively.

**Left Skewness:**

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

**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.

**Exponential Transformation:** Exponential transformations can be applied to data with negative skewness to stretch the lower values and compress the higher values.

**Box-Cox Transformation:** The Box-Cox transformation is a family of power transformations that can be used to stabilize variance and make data more normal.

|  |  |  |
| --- | --- | --- |
| Square Transformation result: | Exponential Transformation result: | Box-Cox Transformation result: |
| A screenshot of a computer code  Description automatically generated | A screenshot of a computer  Description automatically generated | A screenshot of a computer  Description automatically generated |

Applying the Square Transformation, we can observe that one column still exhibits skewness. In other hand, using exponential or Box-Cox transformations, both were effective approaches for mitigating and completely eliminating skewness in the data.

Considering that using Exponential transformation, will change the column to a infinite number, the best approach is using the Box-Cox transformation to handle left skew.

Result of applying exponential transformation:

A screenshot of a computer screen

Description automatically generated

Result of applying Box-Cox transformation:

A screenshot of a computer screen

Description automatically generated

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

**Forecasting**

For forecasting the quantity of crimes and 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 simple but powerful model that can be used to predict continuous variables, such as population growth. It works by finding a linear relationship between the population growth and one or more independent variables, such as fertility rates, mortality rates, and migration rates.

## 3.1 Sample Report – Information Gathering

## 3.2 Sample Report – Service Enumeration

|  |
| --- |
| The |

## 3.4 Sample Report – Maintaining Access

Maintaining access to a system is important to us as attackers, ensuring that we can get back into a system after

## 3.5 Sample Report – House Cleaning

The

After the trophies on both the lab network and exam network were completed, John removed all user accounts and passwords as well as the Meterpreter services installed on the system. Offensive Security should not have to remove any user accounts or services from the system.

# 4.0 *Programing*

A screenshot of a computer program

Description automatically generated