**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 performed with the data, this process consists of cleaning, transforming, and manipulating data to make it more usable for analysis.

This process will be performed on Jupyter notebook: **MSC\_DA\_CA1\_Jose\_Mario.ipynb** and It was called “**Phase 01 - Data Wrangling** “

The project will use 3 datasets:

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

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

\* Duplicates / NaN

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

EDA allows us to gain an overall understanding of the dataframes, detecting 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 best approach is the multivariate analysis, what is exploring the relationships between three or more variables, so I will look at the relationship between population, and crime rates.

I chose these methods because they will allow me to gain insights over the overall trends in population, as well as the overall situation of the crime rate and relation with the population trend.

After the first process, “Data Wrangling”, the population\_df dataset is shown bellow:

A table with numbers and lines

Description automatically generated

The crime\_df dataset is shown bellow.

A screenshot of a computer

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 – Visualization”

**Skewness and Outliers:**

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

A group of blue and white graphs

Description automatically generated*Figure 01: 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.
* 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.

Based in this insight, que can answer the following question:

Which region has the highest median population?

*The Border region has the highest median population.*

**Line Chart:**

From the Line charts we can get the following insights:

* Trends: we can identify trends, such as whether the population is increasing, decreasing, or remaining stable.
* Relationships: We can identify relationship between two variables.
* Comparisons: We can compare two or more variables over time. This can be useful for identifying differences between regions.

A graph of different colored lines

Description automatically generated  
*Figure 02 – 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 further more using Machine Learning, we can get insights into the future of population.

**Pie Chart:**

From Pie charts we can get following insights:

* Composition: From Pie charts we can get how a whole is divided into its constituent parts.
* Comparisons: We can also compare the population of each region. This can be useful for identifying similarities and differences between different regions.
* Trends: We can also plot more than one Pie chart and observe how it changes over time.

Pie chart uses a 1D array as input. so, we need to convert the data variable to a 1D array using numpy.ravel() function we can flattens a NumPy array into a 1D array.

In this case, we will plot two Pie Carts one with data from 2013 and other with data from 2023, so that, we can observe any change in 10 years.

A close-up of a pie chart

Description automatically generated*Figure 03 – 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**

Statistics is essential for data analytics because it enables data scientists and analysts to:

Summarize and describe the data

Identify patterns, trends, and relationships in data

Make predictions and draw conclusions about populations based on a sample of data

Test hypotheses and determine the statistical significance of relationships between variables

Communicate insights and findings to stakeholders in a clear and concise manner

Using the method describe, we can get important information from the result

Ex:

\* 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.

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.

**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 importante 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

To get the exact values, we can also explore the pmi itself, here are a way to explore the pmi.

A screenshot of a computer program

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

**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.

# 3.0 Sample Report – Methodologies

John utilized a widely adopted approach to performing penetration testing that is effective in testing

## 3.1 Sample Report – Information Gathering

## 3.2 Sample Report – Service Enumeration

The service enumeration portion of a penetration test focuses on gathering information about what

## 3.3 Sample Report – Penetration

The penetration testing portions of the assessment focus heavily on gaining access to a variety of

|  |
| --- |
| Vulnerability Exploited: Ability Server 2.34 FTP STOR Buffer Overflow |
|  | |

## 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 Additional Items Not Mentioned in the Report

This section is placed in the overall report.

## Abstract

*The aim of this project is to explore the relationship between population trends and recorded crimes in Ireland. To achieve this goal, I will be investigating the impact of population growth on crime rates.*

*For this purpose, we need to answer the following 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?*

1 - Data preparation and Visualization

* 1. **Early Data Analysis/Exploratory Data Analysis (EDA)**

Before getting any result from the data, it is necessary perform the correct EDA process, this process provide insights into its characteristics, relationships, and patterns.

**EDA method for the dataset**

The project will use 4 datasets:

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

As we have three different datasets, the best approach is the multivariate analysis, what is exploring the relationships between three or more variables, so I will look at the relationship between immigration, population, and renting prices.

I chose these methods because they will allow me to gain insights over the overall trends in population, as well as the overall situation of the crime rate and relation with the population trend.

Performing the EDA:

This process will be performed on Jupyter notebook:

File: MSC\_DA\_CA1\_Jose\_Mario.ipynb

Phase 01 - EDA:

In the phase 01, I will organize and clean the data, loading the dataset and checking the header. After I will drop unnecessary columns and rename some to be more understandable.

After, I will check if is there any null value .

**4. Insights gained from EDA**

The insights gained from EDA will help me to choose the appropriate prediction method for the dataset and to develop a more accurate and reliable model. For example, if I find that immigration is a strong predictor of population growth, then I can include immigration in my prediction model.

**Here are some specific examples of insights that I might gain from EDA:**

* I might find that immigration has been increasing steadily over time, while population growth has been slowing down. This could suggest that immigration is becoming a more important driver of population growth in Ireland.
* I might find that there is a strong positive correlation between immigration and renting prices. This could suggest that an increase in immigration leads to an increase in demand for housing, which drives up renting prices.
* I might find that there is a negative correlation between population growth and renting prices. This could suggest that an increase in population leads to an increase in the supply of housing, which drives down renting prices.

By identifying these insights, I can develop a better understanding of the data and choose the appropriate prediction method.

1. The first goal is using Data Visualization, analize the population growth in each region and in total 1996.

As we have 2 datasets with these datas: PEA07-Estimated\_Pupulation-1996-2017.csv and PEA04-Estimated\_Pupulation-2011-2023.csv. We need to visualize, clean, verify if is there any “bad data” (null or bad formatted data), and after merge these datasets.

Analysing the dataset, we can see that the population is split into regions, each region represents a group of conties, as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

**Machine Learning**

**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.