**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:** | *MSC\_DA\_CA1* |
| **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 Renting Prices**

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?*

*To answer these questions, I have collected and analysed data on population trends and recorded crime in Ireland.*

*The main source of these dataset is: Central Statistics Office (CSO),*

*Once you have collected the data, you can use a variety of statistical methods to analyse it. 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 renting prices.*

**Data preparation and Visualization**

**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:

MSC\_DA\_CA1\_Jose\_Mario.ipynb # Fase 01 - EDA:

**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.