**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | MSc Data Analysis |
| **Assessment Title:** | Integrated Continuous Assessment 1 |
| **Lecturer Name:** | David McQuaid |
| **Student Full Name:** | Ruairi Mc Hugh |
| **Student Number:** | SBA23300 |
| **Assessment Due Date:** | 12TH November 2023 |
| **Date of Submission:** | 12TH November 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. |

MSc Data Analytics - Integrated Continuous Assessment

Ruair´ı Mc Hugh SBA23300 12th November 2023

**Abstract**

This work investigates the change in the historical population data in Ireland and provides insights to population forecasting

# Introduction

## Background

Ireland has undergone a massive population change since 1950, this report will analyse this change based on the data sets provided by the Central Statistics office, the period of time which will be investigated spans seventy three years which include some seminal events in Irish history, the data in the data set starts one year after the foundation of the republic, it also includes the time period that saw sectarian violence known as the troubles, all the way up to more recent events that may influence Ireland’s population such as Brexit and the Covid-19 Pandemic

Other factors that we must take into account when analysing population change in Ireland over the last seventy three years must include economic swings. Ireland until recently performed poorly relative to it’s European neighbours the first major swing recent years of Ireland’s economic fortunes was in the early 2000’s when Ireland experienced unprecedented growth, this was followed by the 2008 financial crash.

## Aims

The aim of this report is to investigate and analyse the influential factors that affect population changes in Ireland over a period of the most recent seventy three years

## Technology

This project is based on the use of Python programming language as the main technology for data analysis, additional technologies used in this report include:

* numpy - Fundamental Python Library
* pandas - Data Manipulation and analysis Python library
* sklearn - Machine Learning Python library
* matplotlib - M2D Plotting Python library
* seaborn - Statistical data Visualization Python library
* scipy - Computation Python library
* statsmodels - Statistical Models Python Library
* Anaconda Navigator - UI for Launching Python environments
* Jupyter Notebook - Python environment
* Overleaf - LaTeX editor
* GitHub - Version control sofware
  1. **Structure**

# Theory

## Statistics

Statistics can be defined as facts or data, either numerical or nonnumerical, organized and summarized so as to provide useful and accessible information about a particular subject. The field of statistics can be defined as the science of organizing and summarizing numerical or nonnumerical information.(W[eiss,](#_bookmark4) [2017)](#_bookmark4) In this report statistics are used as a tool to analyse data from the CSO Annual Population Estimates data sets [(C.S.O.,](#_bookmark1) [2019)](#_bookmark1) by summarising the data to make it more understandable, statistics are also used when implementing descriptive statistics and analysing the data through the use of appropriate plots.

Modern statistics is based largely on the work of Ronald A. Fisher, Fisher pioneered likelihood estimations, distributions and diffusion equations.(Ow[en,](#_bookmark2) [1962)](#_bookmark2) A challenged faced in this report is that while we can measure and analyse the changes in population in Ireland we have no way of directly measuring why these changes are taking place, this means we rely on proxies or a measure that is related to what we want to measure but is not exactly what we want to measure.

### Descriptive and inferential Statistics

Descriptive statistics is means of analysing data which focuses on where the middle of the data is, measures of Central tendency and how spread out the data is another way of explaining descriptive statistics is that it is a method that is used to analyse data of a data set using functions such as the mean, median and mode.

Inferential Statistics allow us to analyse large data through the process of sampling a small portion of the data and inferring conclusions based on the analysis of that smaller sample, the conclusions we are able to draw from this method are based on probability scores.

### Central Tendencies and Distribution

Measures of Central tendency, the most used measures of Central tendency are the mean, median, and mode. Briefly the mean is the sum of all numbers in a data set divided by the number of data points, the mean is one of the first steps in analysing numerical data it is usually the best measure of central tendencies as it uses all data points however can be skewed by outliers. The Median is the middle data point, there are as many data points above it as there are below it, the medium is not affected by outliers however it ignores all but the middle data point of the distribution. The Mode is the value that appears the most in our data set, the mode is most useful when the data set is large and can also be used with numerical and non numerical data.

Data which has mostly the same values of data either side of the median can be referred to as normal distribution, normal, other types of distribution models that will be considered in this report are binomial and Poisson distributions.

## Data Preparation and Visualization

While producing this report a concerted effort was made to follow Tufte’s guiding principles for Data Visualisation(T[ufte,](#_bookmark5) [2001).](#_bookmark5)

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured.
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
3. Show data variation, not design variation
4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units
5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data
6. Graphics must not quote data out of context

In following Tufte’s guiding principles in data visualization this report aims to efficiently communicate the process and conclusions of the data analysis and research which the resulting visualisations are included in this report

## Machine Learning

### Project Management

Project management is crucial for ensuring the successful execution of Machine learning and data analysis projects. It helps define project objectives, allocate resources effectively, manage timelines, and mitigate risks. By providing structure and oversight, project management enhances the likelihood of delivering data analysis results that meet business needs and objectives.

CRISP-DM, as a systematic and adaptable framework, plays a key role in guiding the data analysis process. It ensures a clear understanding of business goals, supports iterative and flexible analysis, promotes transparency, and encourages collaboration. CRISP-DM helps organizations derive meaningful insights from their data, leading to informed decision-making and actionable results.CRISP-DM is an industry-independent process model for data mining. It consists of six iterative phases from business understanding to deployment.

Phase Short description

Business Understanding

The business situation should be assessed to get an overview of the available and required resources. The determination of the data mining goal is one of the most important aspect in this phase.

Data Understanding Collecting data from data sources, exploring and describing it and check- ing the data quality are essential tasks in this phase.

Data Preparation Data selection should be conducted by defining inclusion and exclusion criteria. Bad data quality can be handled by cleaning data.

Modeling The data modelling phase consists of selecting the modeling technique, building the test case and the model.

Evaluation

In the evaluation phase the results are checked against the defined business objectives. Therefore,the results have to be interpreted and further actions have to be defined.

Deployment The deployment phase is described generally in the user guide. It could be a final report or a software component.

Table 1: Description of the CRISP-DM project management framework. [(Wirth and Hipp,](#_bookmark6) [2000),(Chapman](#_bookmark6) [et al.,](#_bookmark0) [2000),](#_bookmark0) [(Schr¨oer et al.,](#_bookmark3) [2021)](#_bookmark3)

# Method

The environment used in this analysis was Jupyter Notebook and the specific notebook used was CA1.ipynb

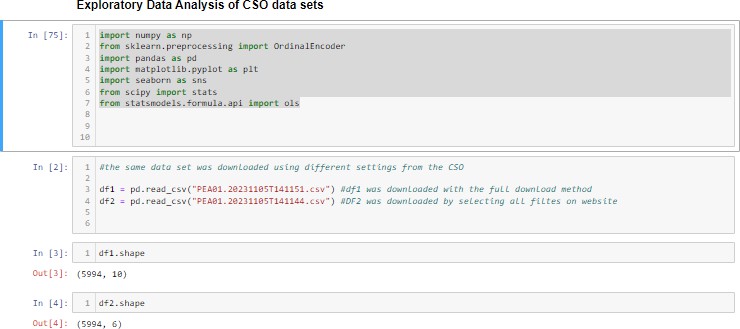


Figure 1

The data set used for this analysis was obtained by downloading it from the CSO website, the

first challenge was that there were two methods of downloading the data set which lead to data sets with different shapes.



Figure 2

df1 was downloaded using the download all technique while df2 was downloaded using the filtered download technique. After investigation df1 contained redundant information with columns like STATISTIC, TLIST(A1), C02076V02508, C02199V02655 these columns gave the same information of already present columns. df2 was chosen as the prefared data set and therefore the prefared technique for downloading the data set was using the filtered technique.

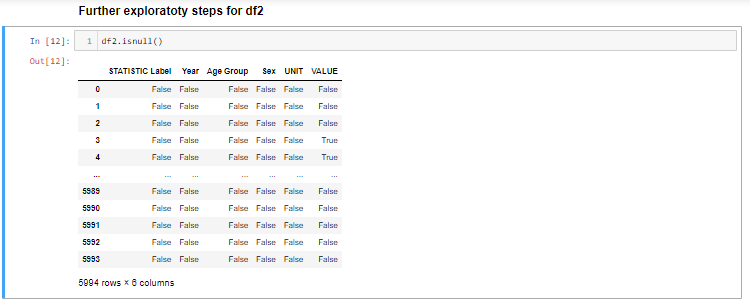


Figure 3

investigating null values in data set.

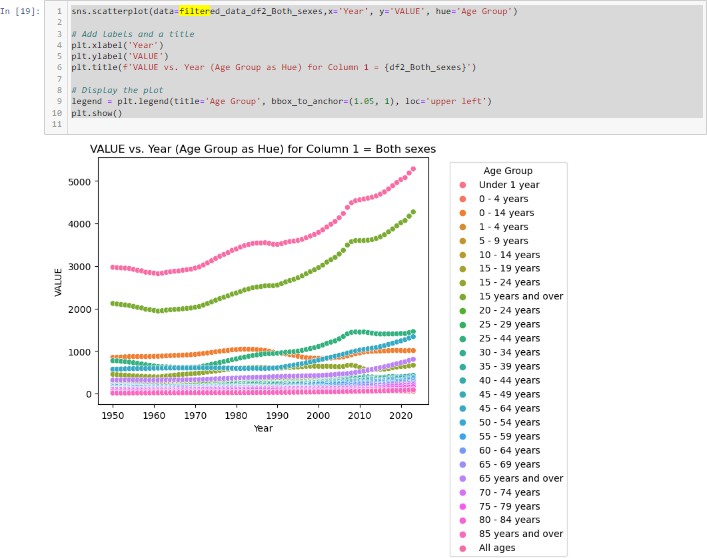


Figure 4

Initial scatter plot of filtered data.

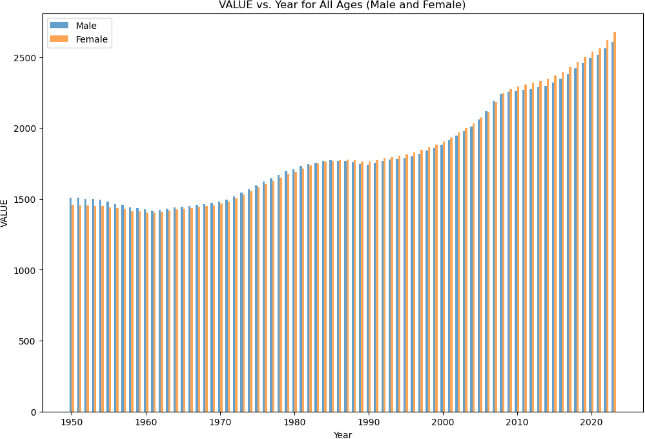


Figure 5

Male and Female populations displayed in number of thousands over the last seventy three years in Ireland.

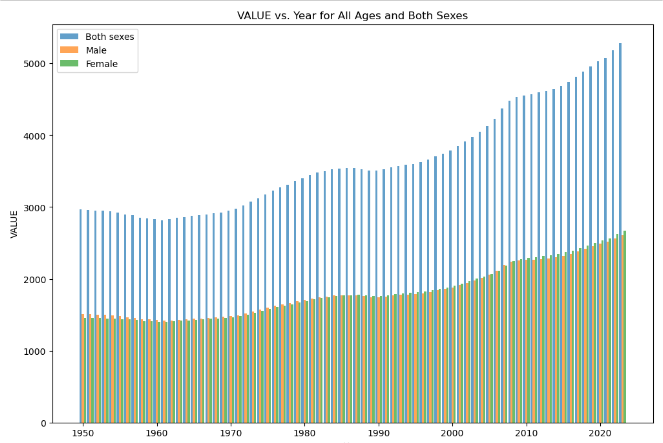


Figure 6

Male and Female populations displayed in number of thousands over the last seventy three years in Ireland along side the total population for comparason.

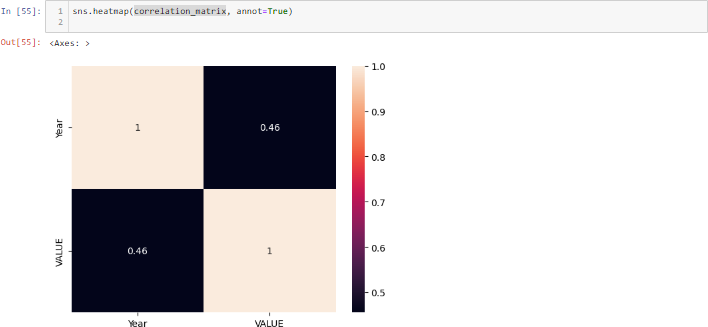


Figure 7

HeatMap

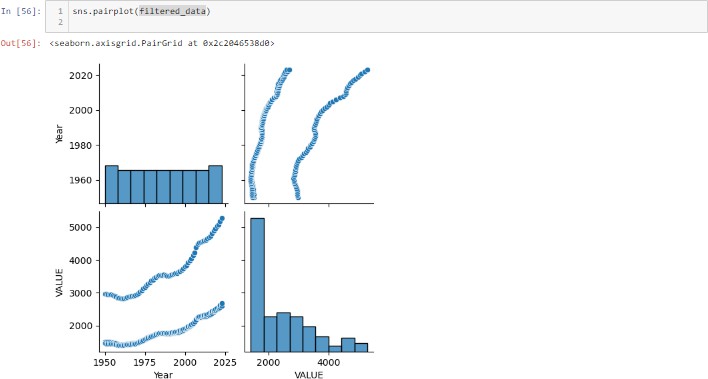


Figure 8

# Results and Discussion

Based on the data set analysed it is clear that Ireland insights can be gained from analysing Ireland’s population, these insights could inform policy-making, housing, healthcare, resource allocation.

# Conclusion

In conclusion the analysis of Ireland’s population change in the previously stated data set from the CSO Ireland has gone through a period of trans-formative change.

# References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide, *SPSS inc* **9**: 13.

C.S.O. (2019). Pme - annual population estimates, An Phr´ıomh-Oifig Staidrimh, The Central Statistics Office.

**URL:** [*https: // www. cso. ie/ /product/ PME*](https://www.cso.ie//product/PME)

Owen, A. R. G. (1962). An appreciation of the life and work of sir ronald aylmer fisher, *Journal of the Royal Statistical Society* **12**: 313.

Schr¨oer, C., Kruse, F. and G´omez, J. M. (2021). A systematic literature review on applying crisp-dm process model., *Procedia Computer Science* **181**: 526 – 534.

**URL:** *https://search.ebscohost.com/login.aspx?direct=truedb=edoAN=148883799site=eds-live* Tufte, E. (2001). *The visual display of quantitative information*, Cheshire: Graphic Press.–2001.–213 p. Weiss, N. (2017). *Introductory Statistics*, Pearson.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Vol. 1, Springer-Verlag London, UK.