

**MSc data analytics ca1**



ireoluwa olajitan

SBA23029@STUDENT.CCT.IE

Contents

[ABSTRACT 2](#_Toc148809910)

[1. INTRODUCTION 3](#_Toc148809911)

# ABSTRACT

# INTRODUCTION

Over the course of this report, there were 4 focus areas in which the selected data was analysed- Data Preparation and Visualization, Machine Learning, Statistics and Programming.

# PROJECT FRAMEWORK SELECTION

Most data projects involve a large amount of data and multiple stages of cleaning and preparation and showing the data, so it is ideal for a project to have an efficient way to structure and manage the process. The assessment for the project management framework for this report was narrowed down to 3 frameworks, CRISP-DM, KDD or SEMMA. A quick overview of each framework is assessed. CRIPS-DM (Cross-Industry Standard Process for Data Mining) is a 6-phase framework that works based on a ranked order which starts off the project at a broad level but then towards the end of the lifecycle specialises the activities required for completion (Chapman et al 2000). KDD (Knowledge Discovery in Databases) as described by Fayyad, et al. (2000) is a 9-phase data management framework where throughout the project lifecycle, knowledge required is extracted and used to structure the data. Finally, SEMMA where the acronym stands for its 5 phases, manages a project from end-to-end using the proceses of Sample, Explore, Modify, Model and Assess.

According to Shafique and Qaiser (2014), there is a notable similarity in all 3 frameworks in terms of the ideology of the phases. Figure x below shows the similarity in the data mining processes

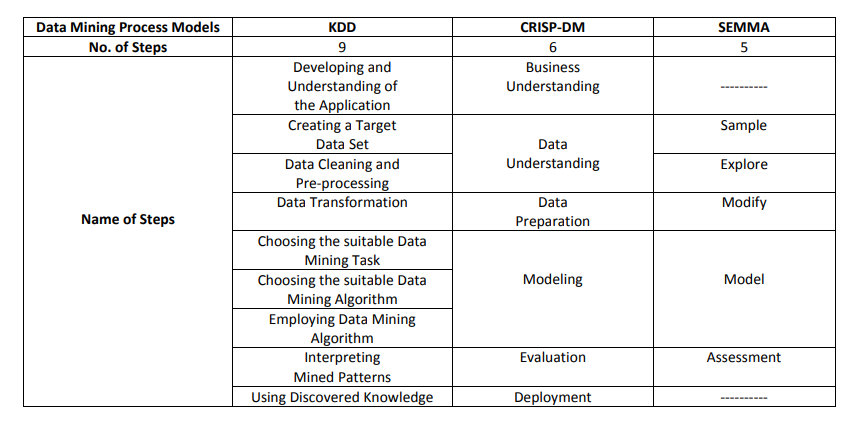
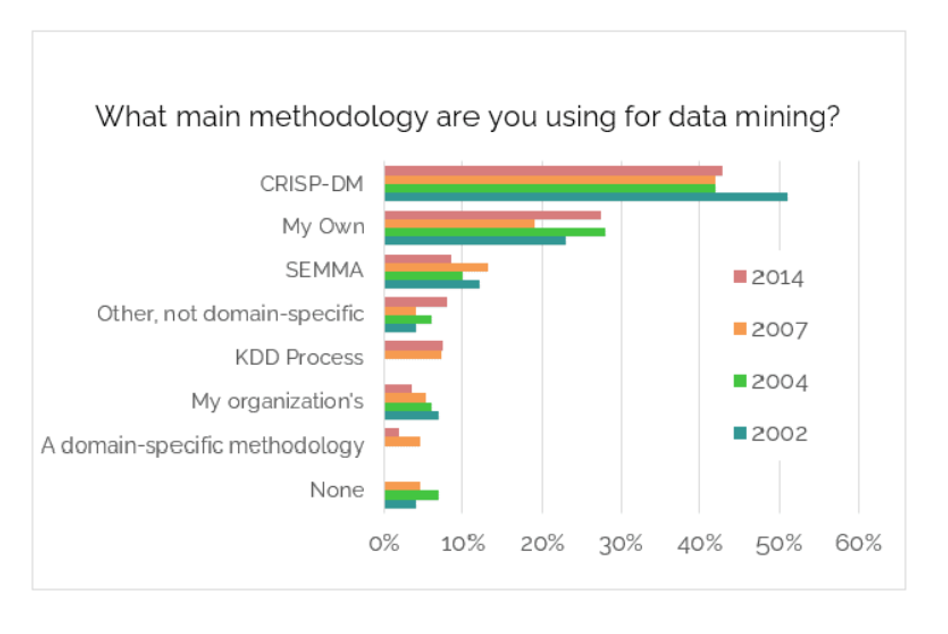


Figure xx: Comparison of KDD, CRISP-DM and SEMMA

From above it is seen that all 3 frameworks’ stages can be linked to the other. This indicates the best type of framework to use for a project is entirely project and data dependent.

The type of data that will be used and analysed will be population data in the country of Ireland, primarily from the census. The characteristics of the datasets to be used will be it being a time-series, having geospatial data, both numerical and categorical data. As there will be a significant amount of data being analysed with multiple features, there will be a level of iteration required and according to Wirth and Hipp (2000), the combination of this instances make CRISP-DM the best framework for the proposed dataset. The justification in selection of CRISP-DM was corroborated by Hotz (2023) who ran a poll collating how the usage of different frameworks changed over an 18 year period. As seen below in figure xx, CRISP-DM remains the preferred project framework by Data analysts

Figure xx: Project Framework Poll (2002-2014)

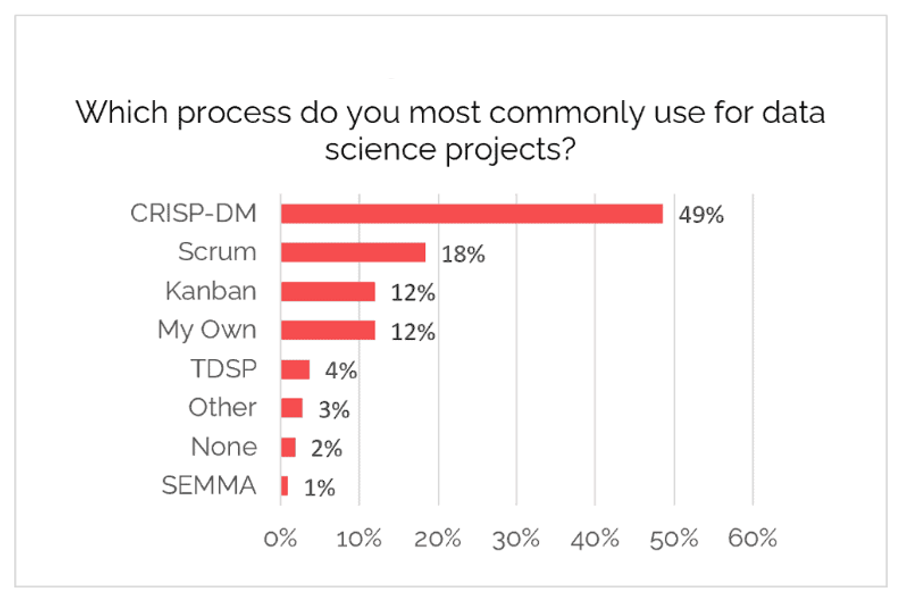


Figure xx: Project Framework Poll (2020)

The methodology of CRISP-DM is illustrated below

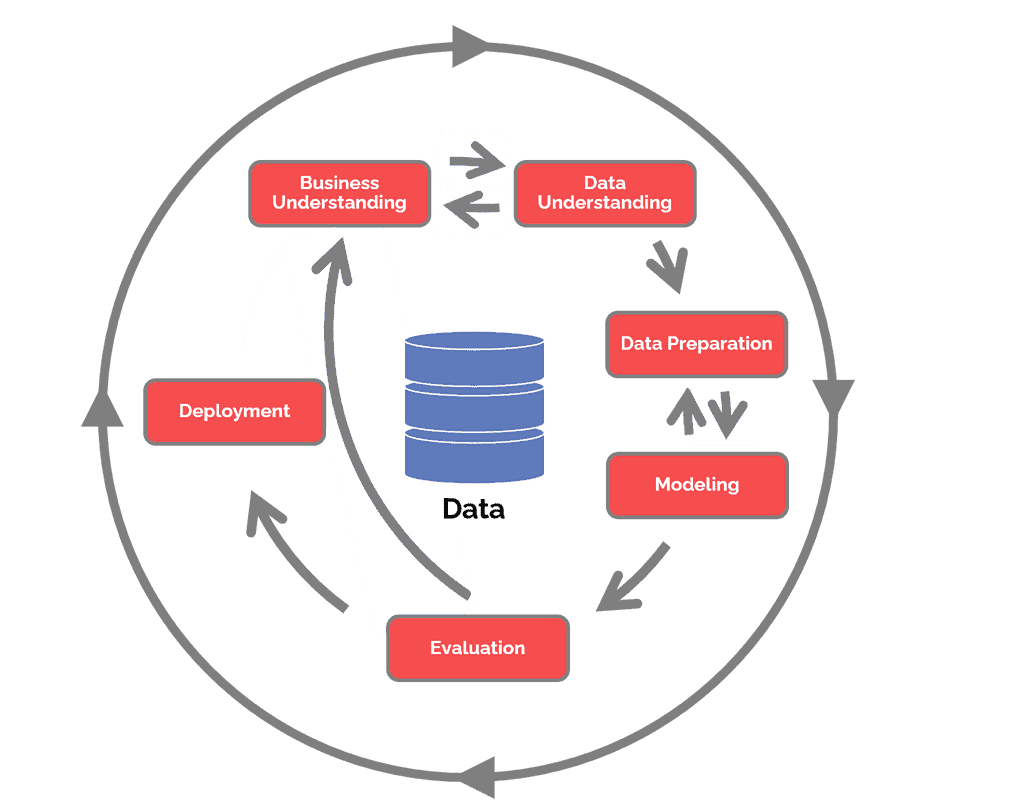


Figure xx: CRISP-DM Methodology (Hotz, 2023)

Over the course of this project, the various phases of how CRISP-DM was utilised will be expanded on. XX real life scenarios

# BUSINESS UNDERSTANDING

In the first phase of the CRISP-DM methodology, the aim and objectives of the project is established. Overall, this project aims to analyse the demographic nature of Ireland as a country now and then predict how this could look in the near future. The main questions it aims to answer is ‘How does the educational distribution in Ireland look like ’. Sub-areas that will be focused on will include current demographic structure in Ireland (age, gender, population distribution, urbanisation, income distribution etc) and effectively looking to predict the demographic structure in the future.

To get to the stage of analysing and modelling the data that will be collected will stem from the overall population in the country over a number of years. This will then dive deeper into specifics such as counties, age and gender distribution, economic status etc. Finally, a model will be developed to analyse the current status, highlight any potential issues and then predict the future in the country

# Data Understanding and Data Preparation

This second phase of the framework methodology aims to collect the data required for the modelling phase. During this phase, the data will be collected and then explored using appropriate EDA (Exploratory Data Analysis), this will be done to visually analyse the data. This part of the process will also look to ensure the data acquired is of good quality ie, no missing, duplicated or invalid data. The original dataset used included data on population in Ireland, population across counties, income distribution and education level. EDA will now be carried out on these 4 datasets. There were 5 datasets used in this project:

* population: dataset detailing population numbers
* regions: dataset dividing the Irish population into the NUTS3 regions
* education: dataset showing the age where the population’s highest educational degree was attained
* income\_region: dataset showing the income distribution across the different counties in ireland
* income\_age: datset showing the income distribution of the country across the different age groups

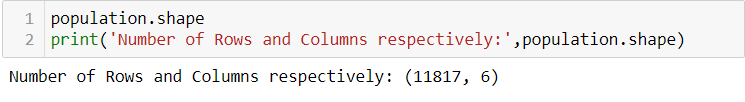
All datasets can also be grouped by Sex. These datasets were gotten from the Central Statistics Office (CSO), so in terms of data credibility and quality, the data collected should be sufficient but this will be ensured through the EDA process

The choice of library for data manipulation used for this projects was pandas. DataChef (2023) cite pandas as one of the most common and well-documented python libraries for data manipulation.

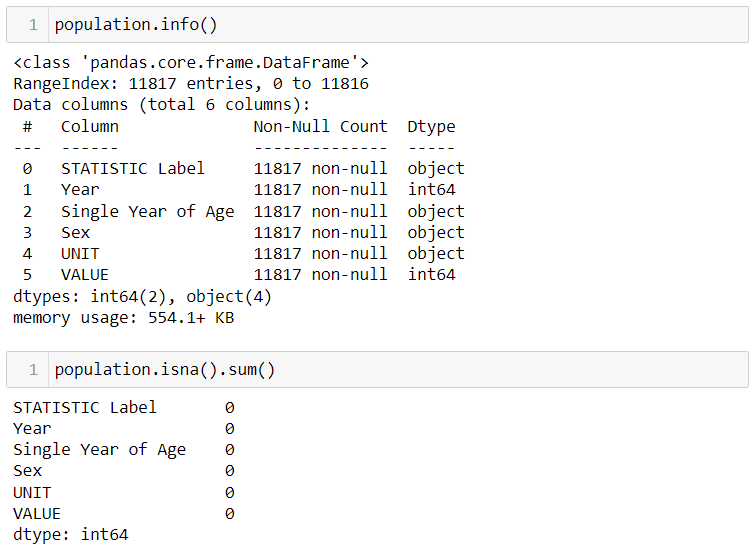
The first step in the EDA process was reading the csv files and converting them into dataframe for the analysis and preparation.



For the purposes of the report, all steps in the EDA that were repeated through all datasets will be referred to as being done for the other datasets to be seen in the accompanying Jupyter notebook for this project. The number of rows and columns are first checked:

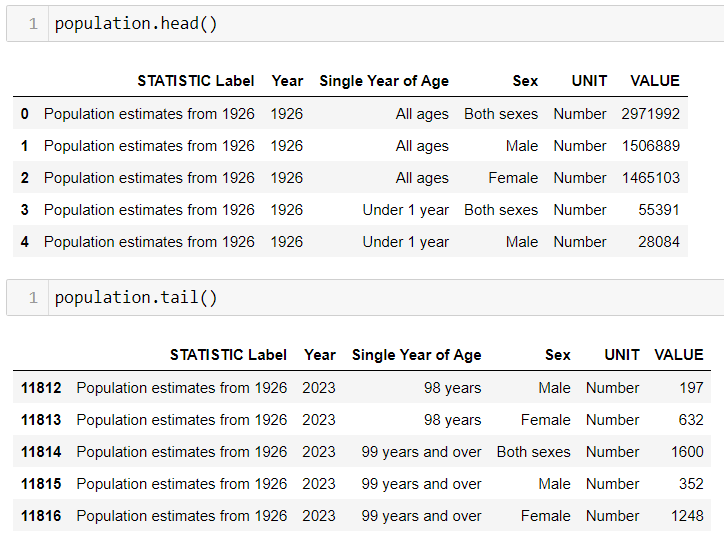


The quality of the data was checked by assessing the number of missing rows:



As seen above, the number or non-null values matched the number of rows previously displayed for the dataset. In typical large datasets with multiple entries, some values could be deemed ‘valid’ but could contain data-irrelevant data such as n.a', '-','--', 'NA', 'Not Applicable', 'n/a'. These were also checked for in the whole dataset. The quality check of all datasets also yielded norows with duplicates. After the quality of the datasets were confirmed, basic pandas commands were used to get a deeper understanding of the data. Some key findings across all datasets will be summarised below.

The first and last couple of entries were displayed to have a sample look at the dataset such as types of data, column names etc.



The columns were later renamed and some dropped for for easier data manipulation. From above taking a preview into the data showed the types of data in each column, the same can be gotten using ‘.info()’ above. Year and VALUE (later renamed to ‘Population’), were the only numerical columns and as a result, descriptive statistics could be used:

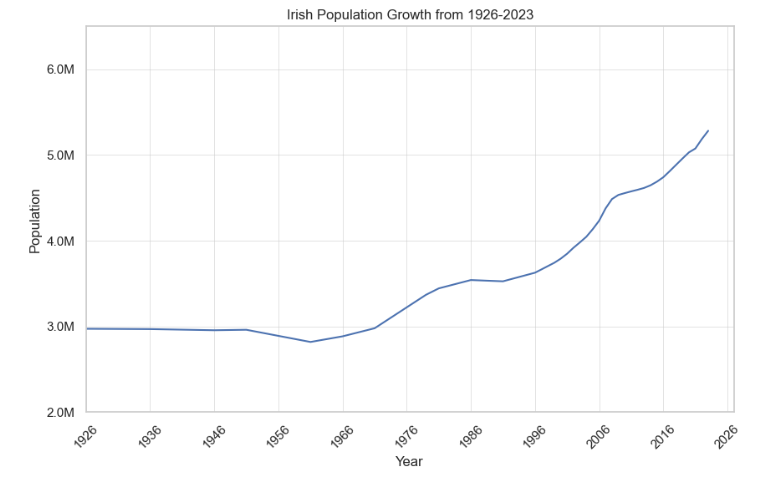


From above, the mean, median (50%), first quartile (25%), third quartile (75%) among many other statistical illustrations were deduced.

After the initial exploration, the datasets will now be visualised to aid in the understanding phase. The bulk of the visualization styles and effects will be done based on methodology as per Tufte’s principles. The main libraries used for visualisation were seaborn and matplotlib.

The firs dataset to which, all further information in this project could be said to be derived from was the population of Ireland over the years, in this case from 1926 to 1923. A recurring theme for all the datasets was the requirement to further split them to avoid aggregation errors and double-counting data. All the datasets referred to either Male or Female sexes but also had rows referencing the aggregation of both variables, this could easily skew and produce inaccurate visualisations and machine learning models. However, one of the reasons why CRISP-DM was chosen as the framework for this project was due t its degree of interaction between the stages. So as a result, there will be portions of this project where the Data understanding and preparation phase will be intertwined.

After the required data filtering of the main dataset to account for the aggregated rows of sex, the population trend for Ireland was plotted:

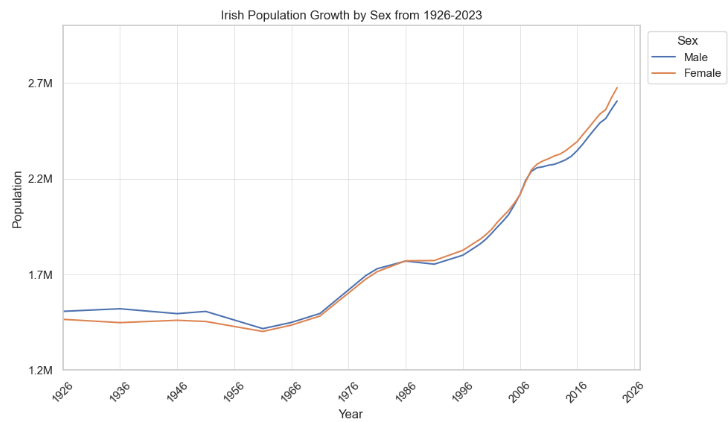


Some stylistic choices that were applied to this graph and successive graphs, where possible:

* Visibility of the gridlines were reduced to minimise Ink-data ratio and give more focus on the data.
* The absence of a legend box to eliminate what Tufte (1983) refers to as chart junk, which are parts of the visualisation that add no meaningful info. As it is just the 1 trendline in this graph, a legend was ignored.
* The default blue colour for the trendlines were selected as they gave a noticeable but not too distracting contrast with the black ink on the graph. This colour and some other colours in successive graphs were also assessed for colourblind audience where applicable. This was done through pilestone.com.
* Labels are present and minimalistic for the title and axes.
* The format of the axes were also modified for visual and practical purposes. For example, where the axis should’ve stated 2700000, it was formatted to 2.7M

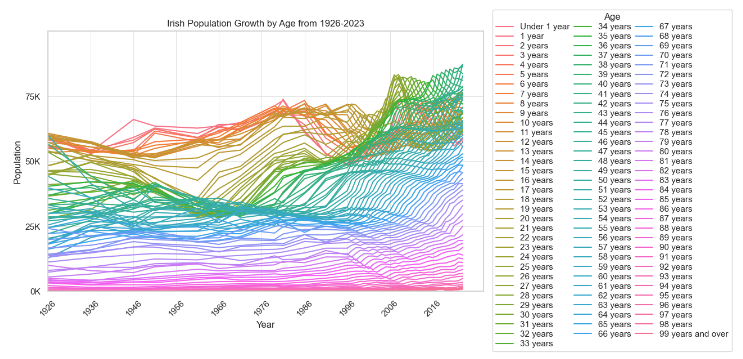
As seen from the graph, Ireland has experienced an upward trend of population since the 60’s. Through research, there has been no conclusive data that suggests the reason for a relatively stable population as of the writing of this report.

Nex the gender balance across the years is shown, where fromabout 1986, the female population is the more populous gender in Ireland.

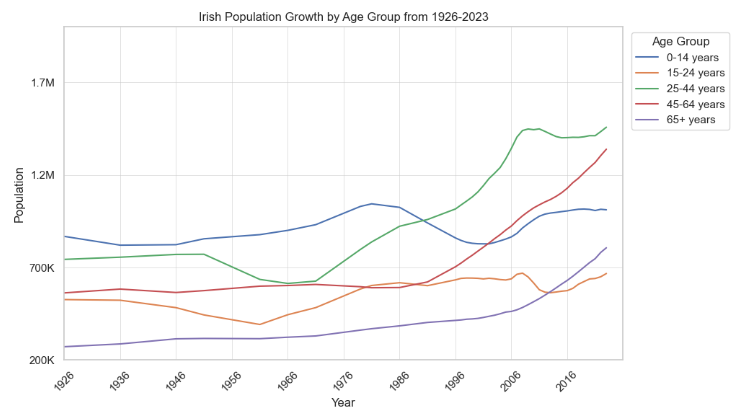


In reference to more methodologies from Tufte, a legend is added into this graph to differentiate the two lines.

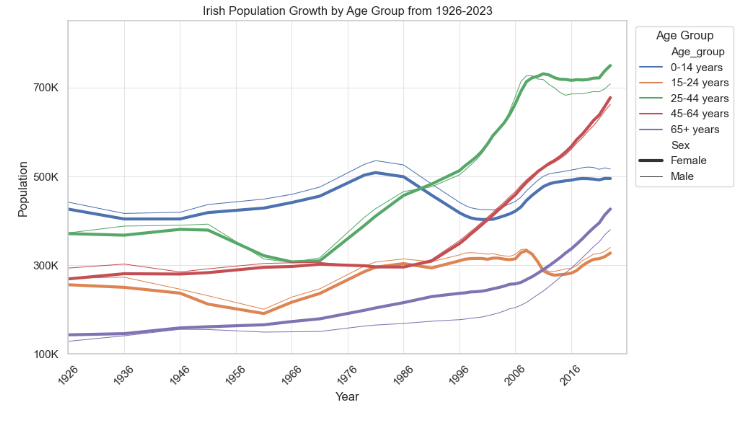
The next demographic to be assessed was the age distribution:



On first glance of the graph, it is extremely busy, meaningful information is hard to get, there is a lot of ink with diminishing returns, etc. Tufte (1983). However, it can be be seen that from the 80’s a certain age group looked to surpass another in terms of population, however the busy nature of the graph prevents one from seeing which. This is alluding to the previously mentioned scanrio where the iterative process of data preparation, visualisation and understanding are facilitated by the CRISP-DM method. The dataset was filtered and the individual ages were put into age groups as below.



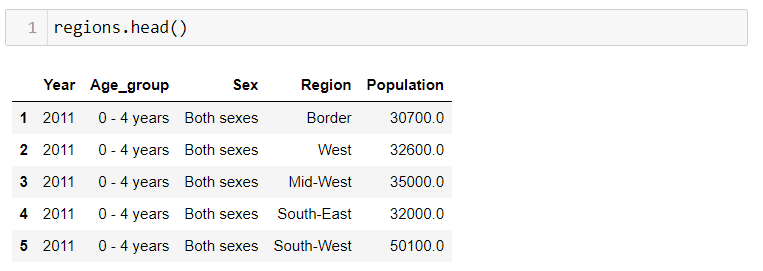
The age group hypothesised to overtake another was the 25-44 group overtaking the 0-14 group. With both gender and age demographics individually assessed, the analyses were combined:



From the dataset above, it is seen that in terms of current landscape (2023), as the population ages, the female population outnumbers the male population, signifying that the irish male population are potentially dying at earlier ages. It is however interesting to see that the older population (65+) have had the female population as the major population for majority of the dataset.

Kelly (2017) corroborated this from a medical standpoint and also gave some reasons and indicators reflecting this visualisation. She sighted factors such as higher risk of heart disease and stroke, higher likelihood to commit suicide, lifestyle choices such as smoking, high cholesterol diets, alcohol, all which men have been statistically shown to indulge more in.

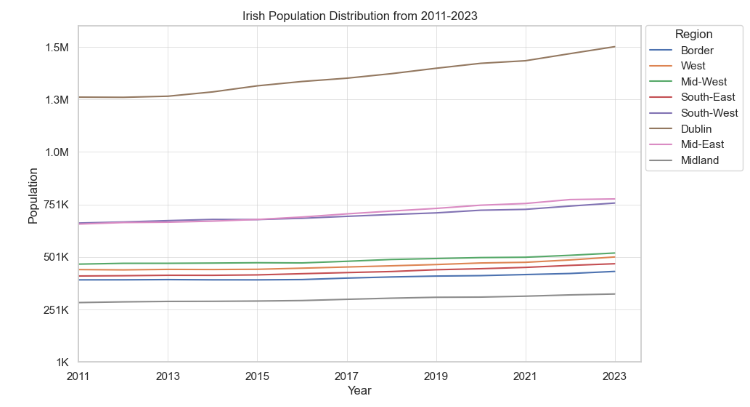
After the analysis of the population as a whole, the second dataset was used to get a deeper understanding on how this population was distributed across the country.



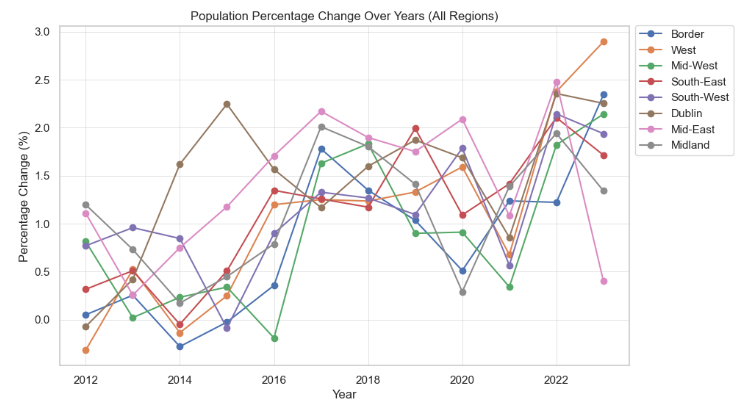
The dataset gotten from the CSO (2021) split the country into the 8 NUTS3 regions:

* Border: Cavan, Donegal, Leitrim, Monaghan, Sligo
* West: Galway, Mayo, Roscommon
* Mid-West: Claire, Limerick, Tipperary
* South-East: Carlow, Kilkenny, Waterford, Wexford
* South-West: Cork, Kerry
* Dublin
* Mid-East: Kildare, Louth, Meath, Wicklow
* Midland: Laois, Longford, Offaly, Westmeath

The graph below shows the irish population split into its regions.



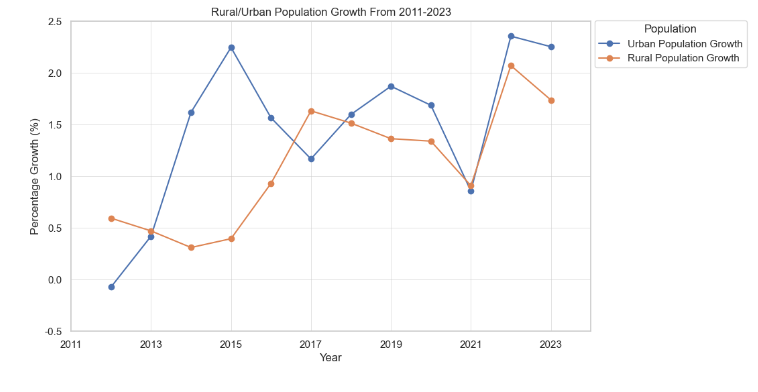
Dublin as expected, is the most populous and looks on the face of it to be the constant highest grower, which is why a percentage growth plot was made to assess this, below.



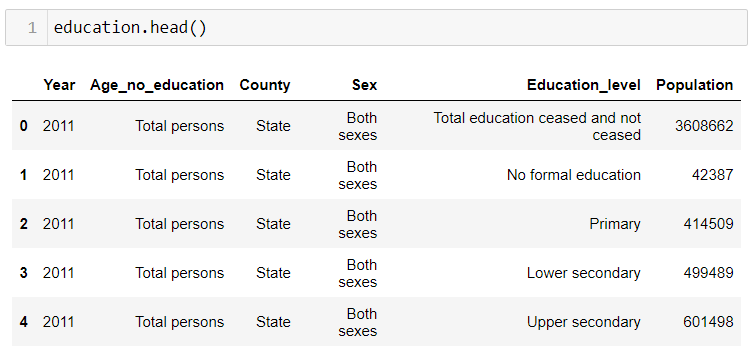
The population trend is fairly similar for all regions. Key findings:

* From 2013-2016, there could’ve been potential urbanisation as th increased in Dublin where it reduced across the country
* There was a downward trend across the country in in 2021 and then an upward one the following year, possibly post-COVID travels or the younger population going to explore other countries briefly

The urbanisation theory is also assessed by consolidation all the other regions as 1 rural population and Dublin as the urban population, shown below:



Next, the education dataset is explored:



There were a number of varying mentiond to education levels in this dataset, so to aid understanding, these were classified according to the National Framewrok of Qualifications (QQI, 2021)

You need to analyse the chosen dataset using statistical logic and statistical techniques. Note: ALL Statistical work MUST be carried out using Python.

You are required to:

Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at least three graphs**. [0-35]**

Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. **[0-30]**

Use Normal distribution to explain or identify some information about your dataset. [0-20]

Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case. [0-15]

# PROGRAMMING

The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should also be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. Note that marks may not be awarded if the discussion does not involve your specific project. **[0-50]**

# RESULTS AND CONCLUSION

# REFERENCES

Chapman at al., 2000. CRISP-DM 1.0: Step-by-step data mining guide. SPSS.

CSO (2021) Standard counties and nuts regions - CSO - central statistics office, CSO. Available at: https://www.cso.ie/en/methods/classifications/standardcountiesandnutsregions/ (Accessed: 30 October 2023)

DataChef (2023). *Choosing the Best Data Manipulation Package in Python: A Comprehensive Comparison*. [online] datachef.co. Available at: https://datachef.co/blog/data-manipulation/ [Accessed 11 Nov. 2023].

Fayyad, U. M., 1996. Data mining and knowledge discovery: making sense out of data. IEEE Expert, Vol. 11 No. 5, pp. 20-25

Hotz, 2023. CRISP-DM*.* [image online] Available at: https://www.datascience-pm.com/crisp-dm-2/ [Accessed 03 Nov. 2023].

Hotz, N. 2023. CRISP-DM. [online] Data Science Project Management. Available at: https://www.datascience-pm.com/crisp-dm-2/ [Accessed 03 Nov. 2023].

Kelly, C. (2017). *Dr Ciara Kelly: 15 reasons why men die younger*. [online] Independent.ie. Available at: https://www.independent.ie/life/health-wellbeing/health-features/dr-ciara-kelly-15-reasons-why-men-die-younger/35384958.html#:~:text=Irish%20men%20have%20a%20life [Accessed 11 Nov. 2023].

Molin, S. and Jee, K., 2021. *Hands-On Data Analysis with Pandas: A Python data science handbook for data collection, wrangling, analysis, and visualization*. Packt Publishing Ltd.

QQI (2021). *National Framework of Qualifications*. [online] Quality and Qualifications Ireland. Available at: https://www.qqi.ie/what-we-do/the-qualifications-system/national-framework-of-qualifications [Accessed 4 Nov. 2023].

SAS Institute, 2003. *Data mining using SAS Enterprise Miner : a case study approach.* Cary, Nc: Sas Pub.

Shafique, U., and Qaiser, H., 2014. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), pp.217-222.

Tufte, E.R. (1983). *The Visual display of quantitative information*. Cheshire, Conn.: Graphics Press.

Wirth, R. and Hipp, J. 2000. *CRISP-DM: Towards a Standard Process Model for Data Mining*. [online] Available at: https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.