

**MSc data analytics ca1**



ireoluwa olajitan

SBA23029@STUDENT.CCT.IE

Contents

[INTRODUCTION 2](#_Toc150620294)

[PROJECT FRAMEWORK SELECTION 2](#_Toc150620295)

[1. BUSINESS UNDERSTANDING 5](#_Toc150620296)

[2. Data Understanding and Data Preparation 5](#_Toc150620297)

[3. Modelling and Evaluation 22](#_Toc150620298)

[4. Deployment 25](#_Toc150620299)

[RESULTS AND CONCLUSION 26](#_Toc150620300)

[REFERENCES 26](#_Toc150620301)

# 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 process 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

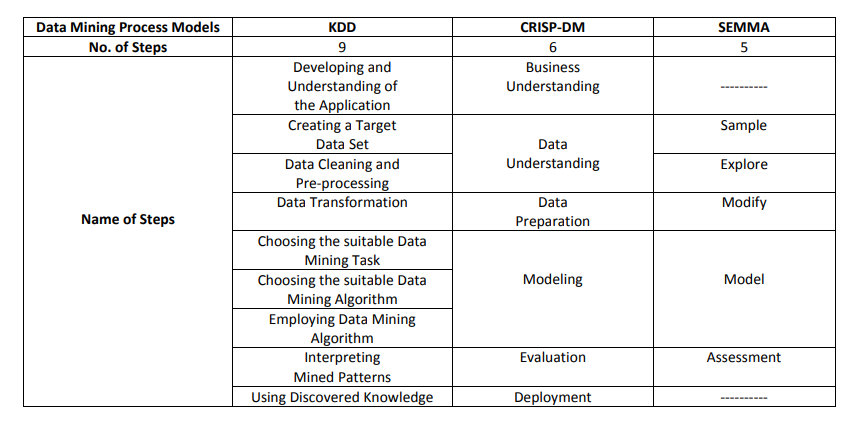
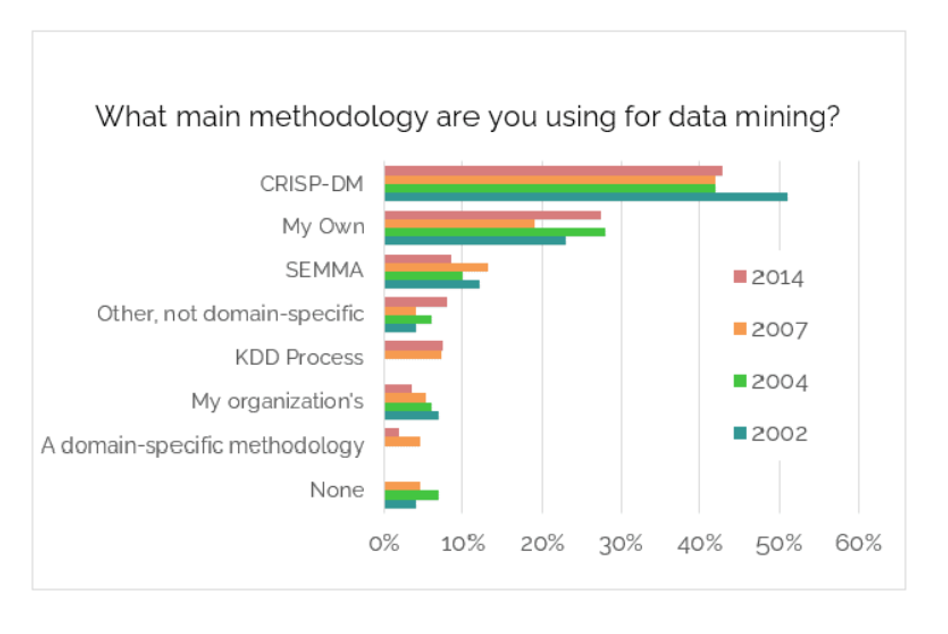


Figure xx: Comparison of KDD, CRISP-DM and SEMMA

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)

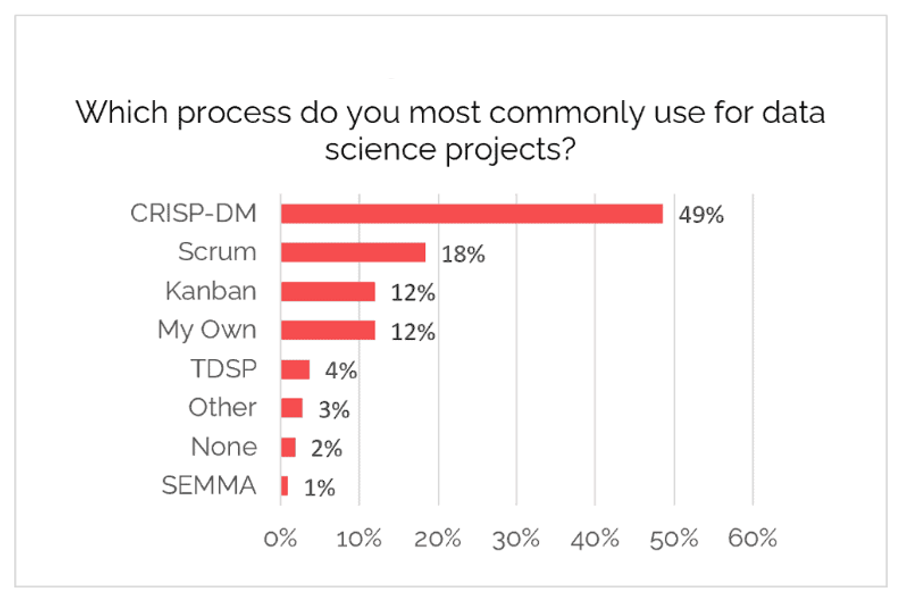


Figure xx: Project Framework Poll (2020)

The methodology of CRISP-DM is illustrated below

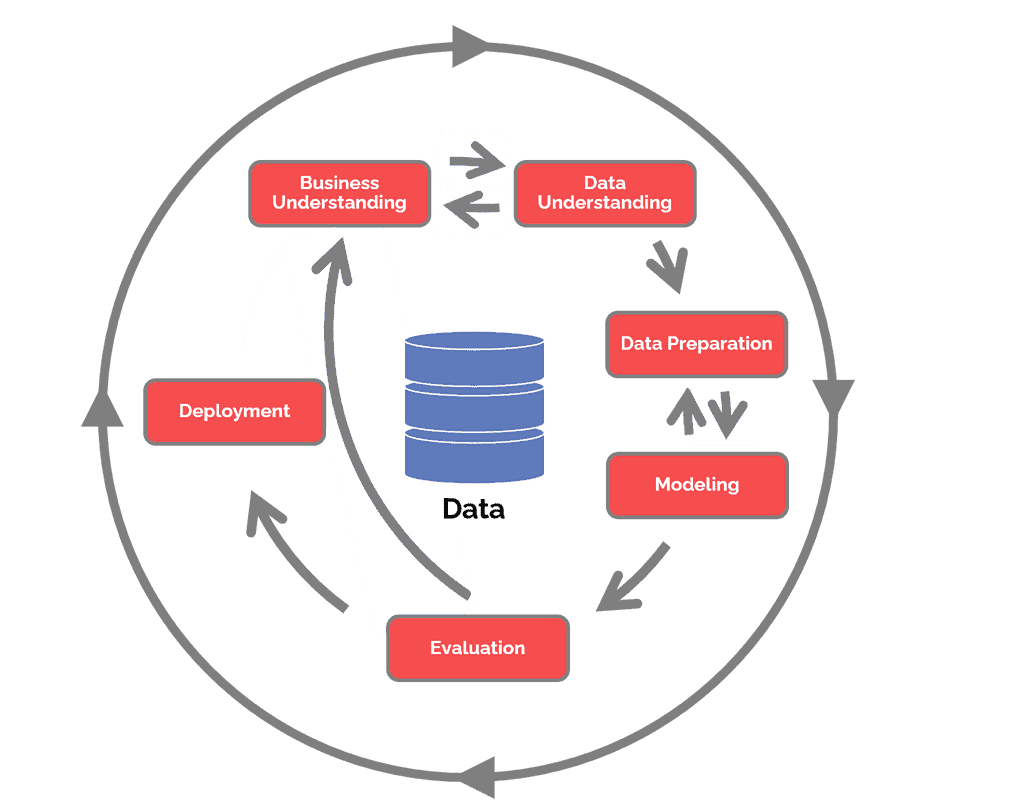


Figure xx: CRISP-DM Methodology (Hotz, 2023)

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: dataset showing the income distribution of the country across the different age groups

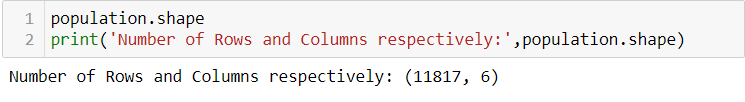
All datasets can also be grouped by Sex. These datasets were gotten from the Central Statistics Office (CSO), so in terms of data credibility and quality, the data collected should be sufficient, but this will be ensured through the EDA process

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

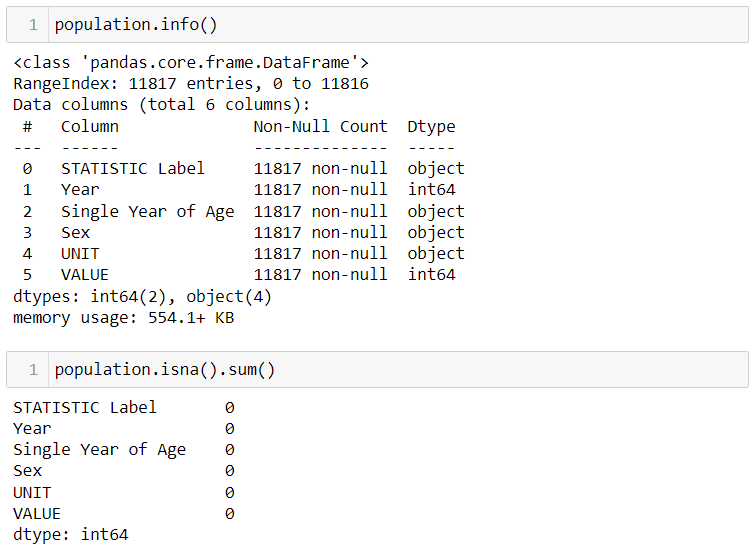
The first step in the EDA process was reading the csv files and converting them into dataframe for the analysis and preparation.



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:

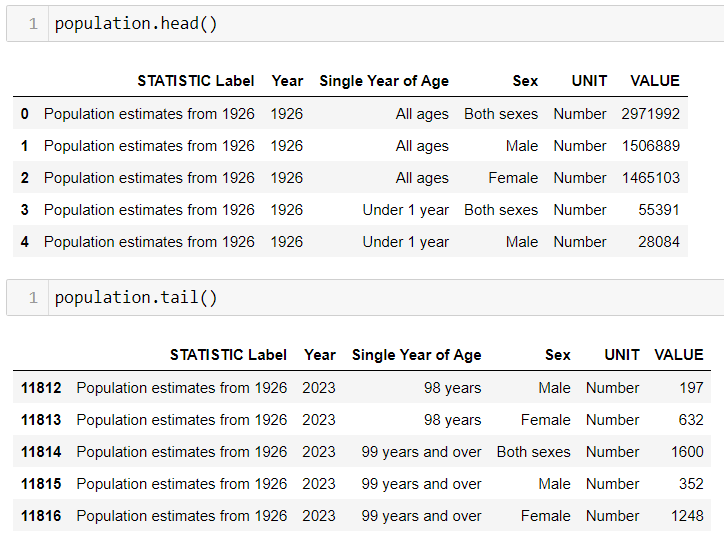


The quality of the data was checked by assessing the number of missing rows:



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 no rows 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 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:

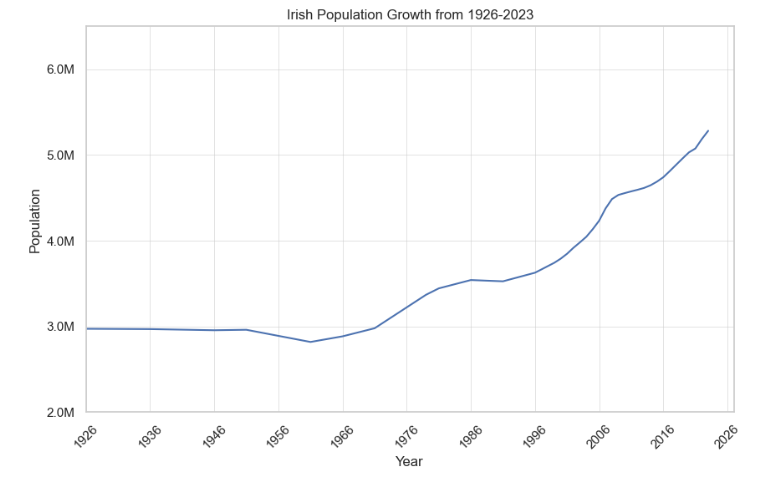


From above, the mean, median (50%), first quartile (25%), third quartile (75%) among many other statistical illustrations were deduced.

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 to 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:

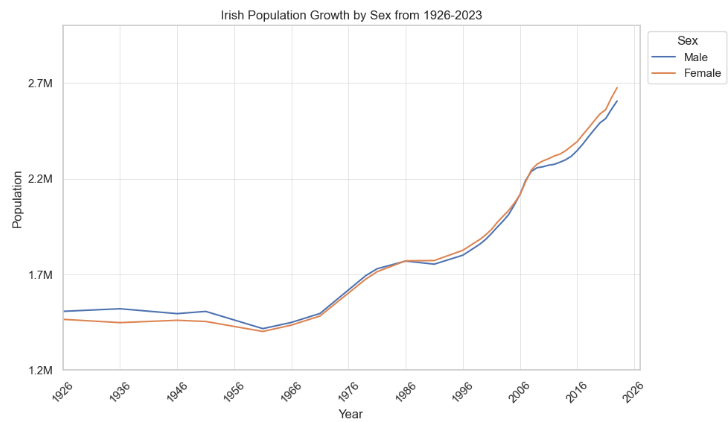


Some stylistic choices that were applied to this graph and successive graphs, where possible:

* 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 was also modified for visual and practical purposes. For example, where the axis should’ve stated 2700000, it was formatted to 2.7M

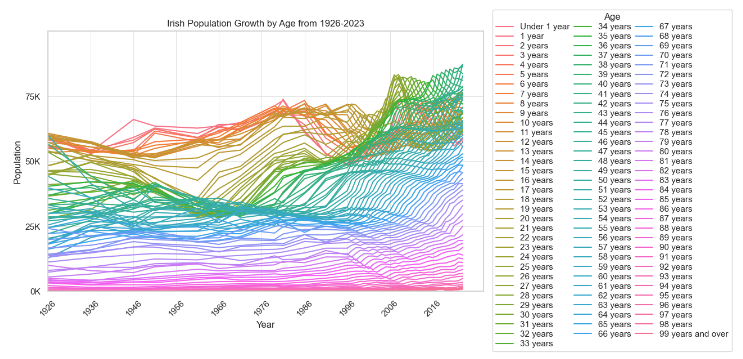
As seen from the graph, Ireland has experienced an upward trend of population since the 60’s. Through research, there has been no conclusive data that suggests the reason for a relatively stable population as of the writing of this report.

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

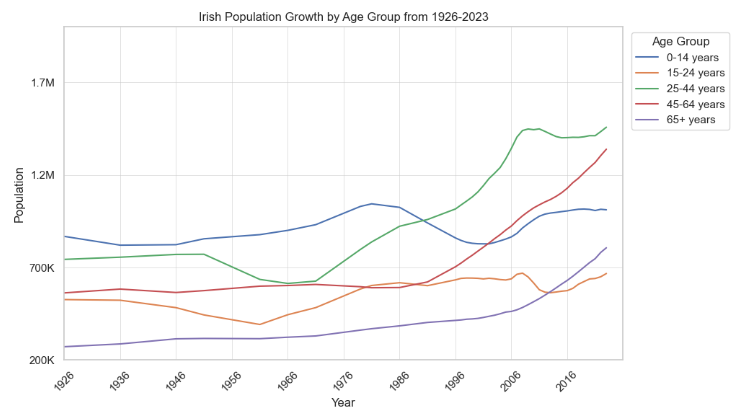


In reference to more methodologies from Tufte, a legend is added into this graph to differentiate the two lines.

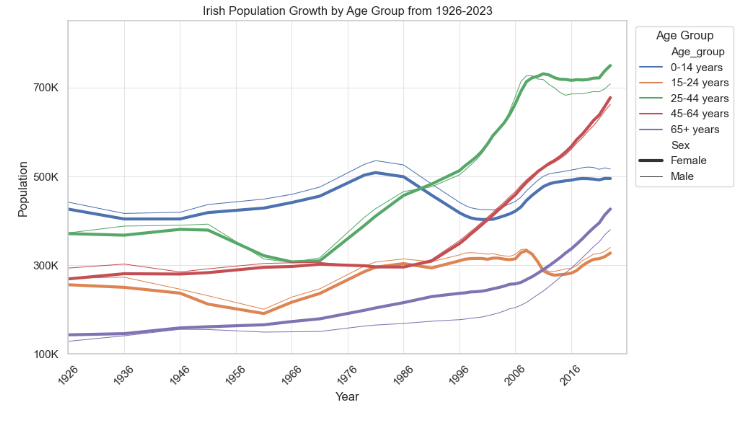
The next demographic to be assessed was the age distribution:



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 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 scenario 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.



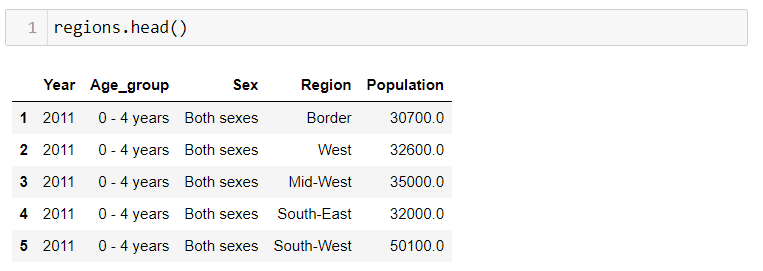
The age group hypothesised to overtake another was the 25-44 group overtaking the 0-14 group. With both gender and age demographics individually assessed, the analyses were combined:



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.

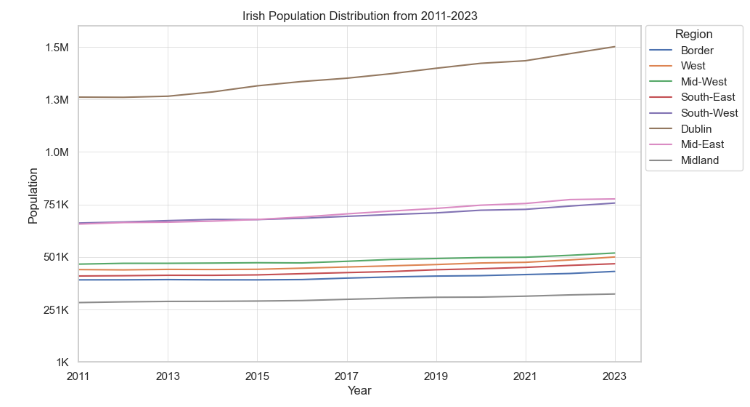
After the analysis of the population as a whole, the second dataset was used to get a deeper understanding on how this population was distributed across the country.



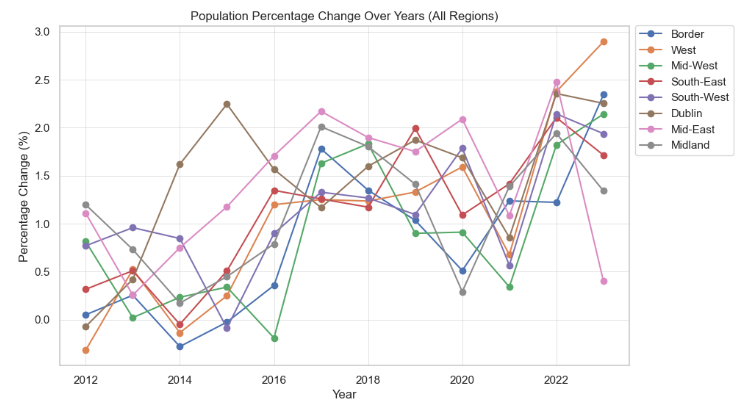
The dataset gotten from the CSO (2021) split the country into the 8 NUTS3 regions:

* 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.



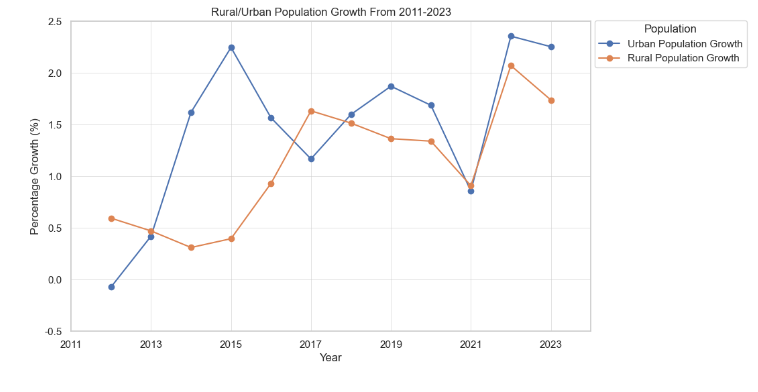
Dublin as expected, is the most populous and looks on the face of it to be the constant highest grower, which is why a percentage growth plot was made to assess this, below.



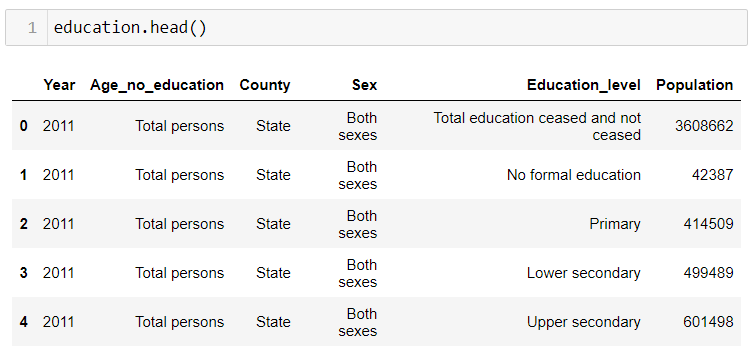
The population trend is fairly similar for all regions. Key findings:

* From 2013-2016, there could’ve been potential urbanisation as the 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

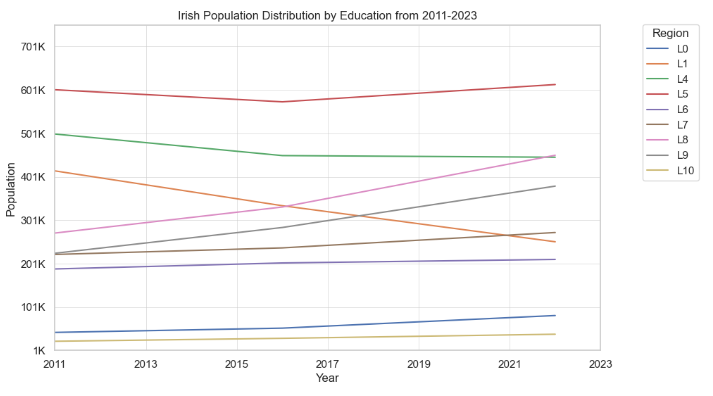
The urbanisation theory is also assessed by consolidation all the other regions as 1 rural population and Dublin as the urban population, shown below:



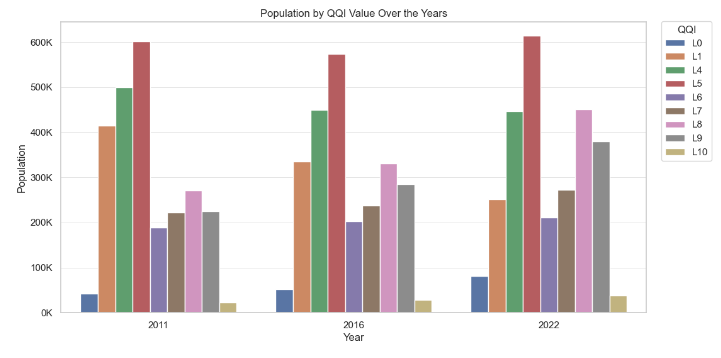
Next, the education dataset is explored:



There were a number of varying mentioned to education levels in this dataset, so to aid understanding, these were classified according to the National Framework of Qualifications (QQI, 2021) ,from Level 0 signifying no formal education to Level 10.

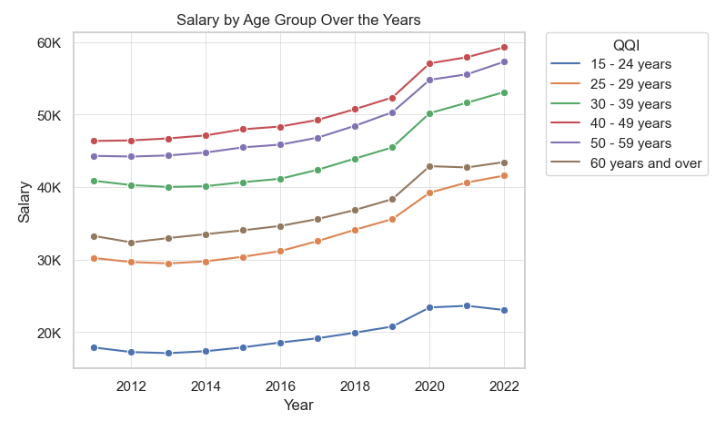


Up until this stage, the line plot was utilised as the visualisation of choice as it is able to show a trend in time relative to previous data, however, there doesn’t seem to be a lot of information deduced from above graph, so a bar plot was utilised in this instance.

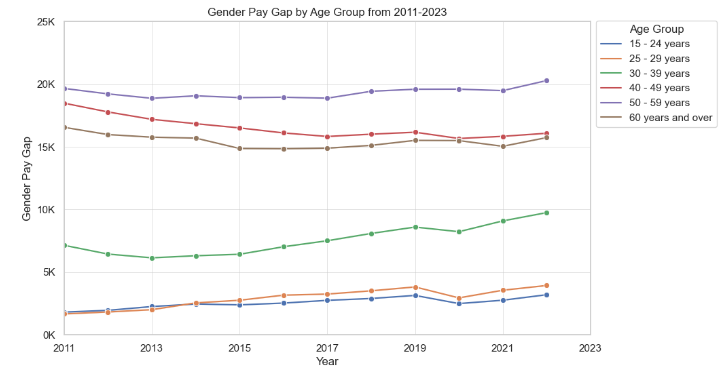


The plot above shows the population who have a level 5 as their highest educational degree attainment to be the highest since 2011 and has remained relatively constant. It’s also interesting to see that while the numbers of the uneducated population rose over the years and the Level 1 population declined, it seemed like more and more of the population were looking to further their education.

The income datasets also gave insight as to how that was grouped in the country.



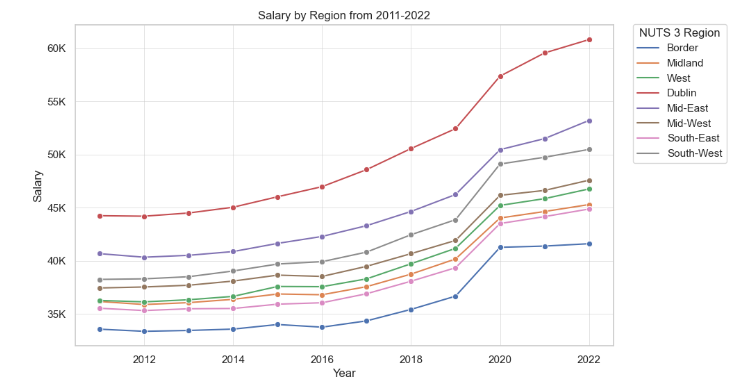
As expected, the 40-49 group and the 50-59 group were the highest earners through the years. This makes sense as they would typically have the most experience and be established in their respective carers. It is however good to see that there has been a relative increase in the average incomes across the age groups. When the issue of income is usually discussed in a countries demographic profile the gender pay gap is often discussed along. The gender pay gap was assessed in the country below. This is the non-absolute difference subtracting the female salary from the male salary



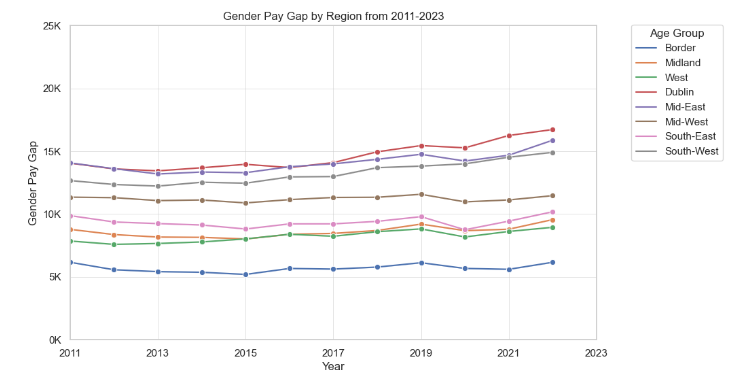
As seen above there is a significant disparity in the incomes as the population ages. Wilson (2021) corroborates this stating that for all occupations across Ireland, males are paid more than females, especially during the middle working years. She attributes this to the major factor of childbirth.

This is a significant difference; however, Ireland looks to slightly better than average when compared to other countries in the EU with an average of 13% compared to Ireland’s 9%.

The earnings were also assessed by regions.

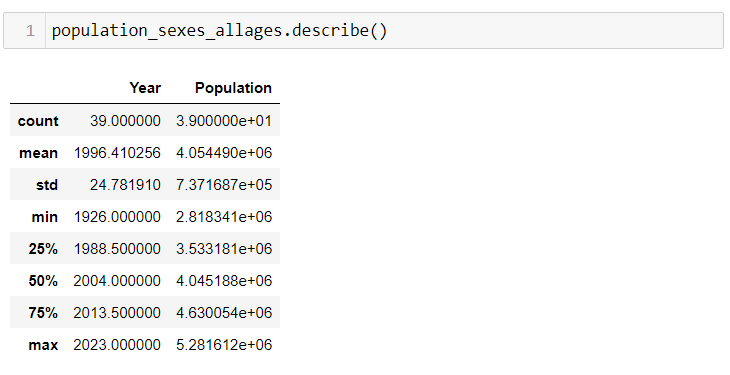


The same gender pay gap analysis was done for the regions to assess if this might vary by regions.

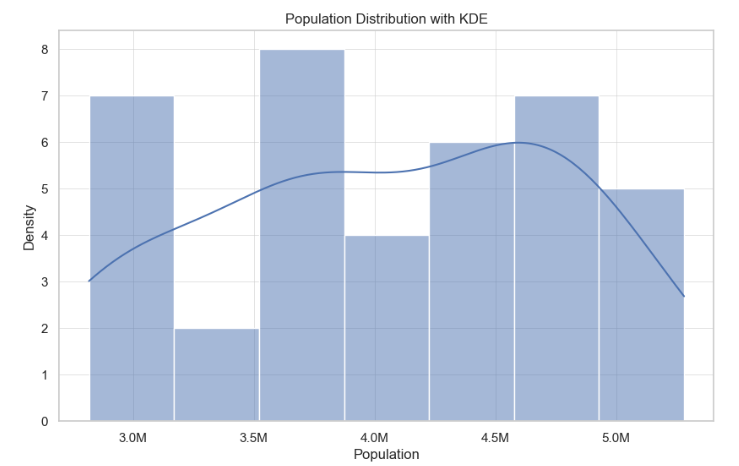


There were no variations, it looked to follow the trend where as the earnings got higher, the gap increased. It is, however, not ideal from a diverse perspective to see that in both cases, the gap looks to be constant and, in some cases, even rising.

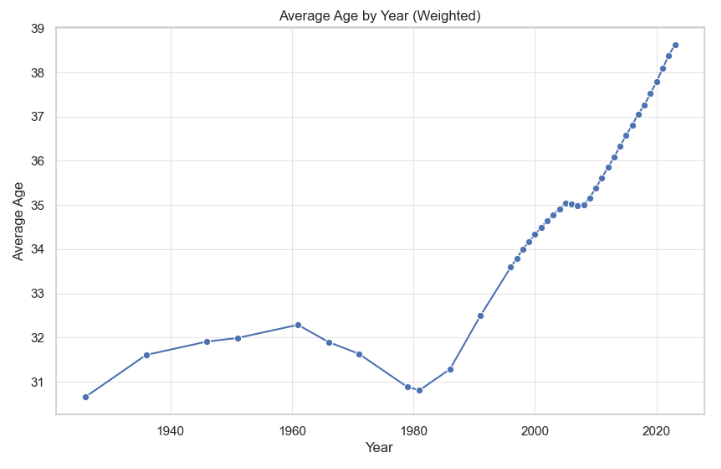
As part of the data understanding phase, it is also important to get information on the statistical distribution of the different datasets. Python has some basic inbuilt statistical tools to carry this out.



For the population, analysis, from above we see the average age over the years (mean) to be slightly above 4million people with the median being roughly the same. In a distribution, this often implies the absence of a skew, however when the distribution was plotted and curve fitted as per below, the data distribution looked to be bimodal with a slight skew to the left where majority of the population over the years look to be concentrated.

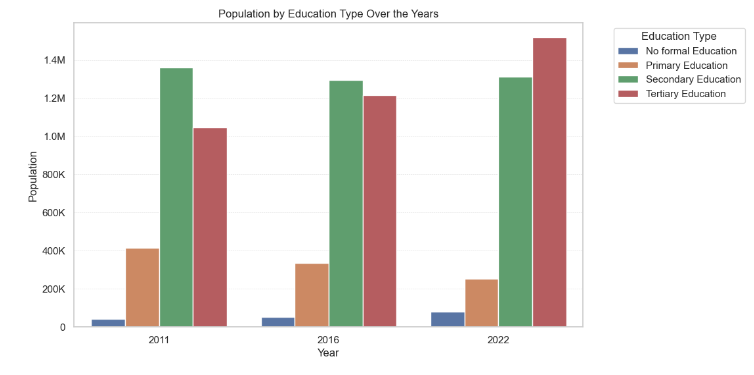


To visualise the mean age referred to in the pandas output, the average was plotted



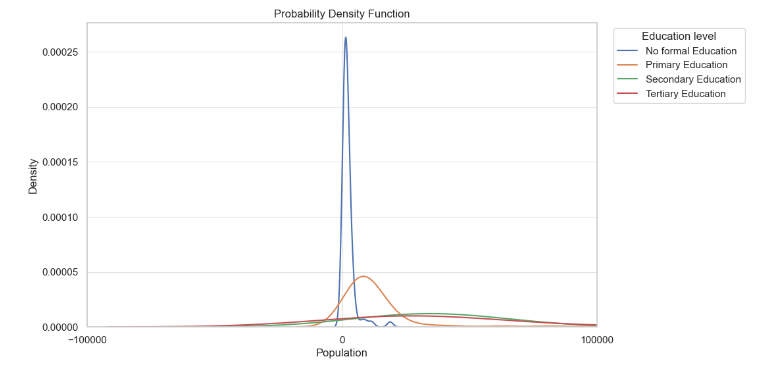
This would signify a growing population in the older ages, indicating a growing life expectancy, also reported by Malone (2020), where the expectancy has risen up an average of 2.5% in the last 3 years

Looking at the education distribution, when grouped further into 4, it gives more of a statistical insight to the previous education data.

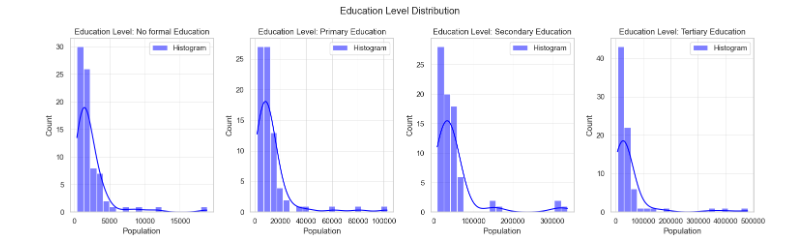


Looking at the distribution, it looks like the population is also left-skewed as the population is concentrated on the right-side alluding to the fact that more people are looking to further their education.

A typical way to infer from a dataset using binomial distribution is a probability density function, shown below.



As seen above, the population is highly concentrated for the uneducated population at the lower population numbers. At the population, the chances where someone doesn't have any formal education whatsoever is significantly high, practically towering over the other education levels, so it would be good to split them up to dig deeper.

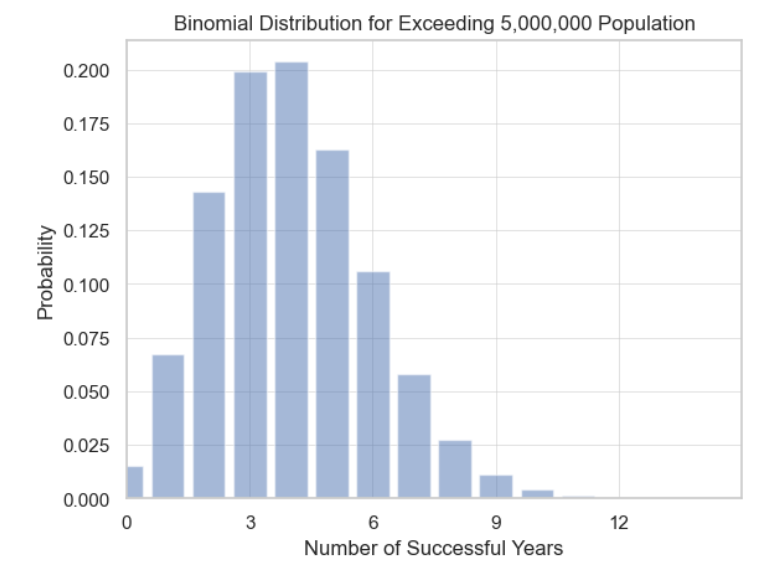


From above, at all education types the data looks to be right skewed, which means in order of rank, the mode is the highest, followed by the median and then the mean, this can also be referred to as a positively skewed distribution. It can also be deduced that the distribution would not be normal as it lacks symmetricity. In terms of the binomial analysis and information gotten, Binomial analysis can be utilised to model the probability of picking a specific number of people where a certain degree is their highest educational attainment in the country.

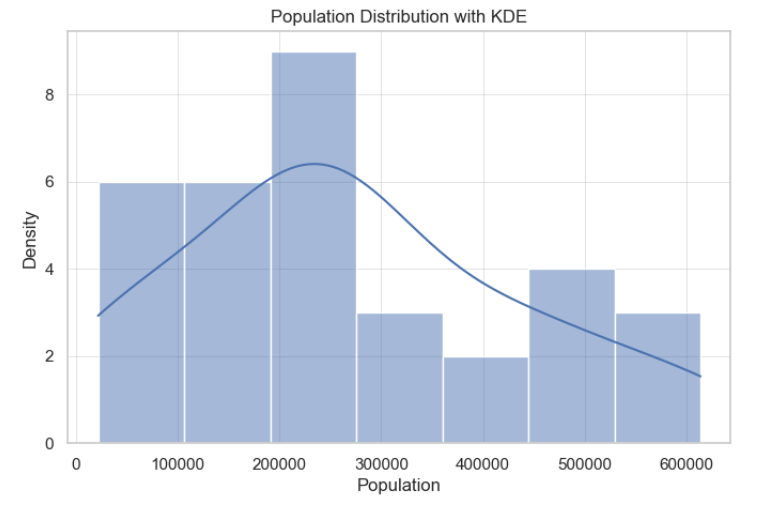
Key findings:

* The binomial distribution can estimate how likely it is to get a specific number of people with a certain degree in the country.
* You can see the populations identified where there are significantly high or low numbers for the different education levels.
* The probability of successfully finding someone who has no formal distribution in the country is really low reflected by the relatively low scale of the X axes.
* Given a random sample from the population of Ireland, the chances you will pick someone with a tertiary education as their highest degree is highest due to the curve distribution and reflected values.
* The Probability density function (PDF) shows the Binomial distribution of educated individuals across the country. We observe the probability of achieving a specific count of educated individual by population number.

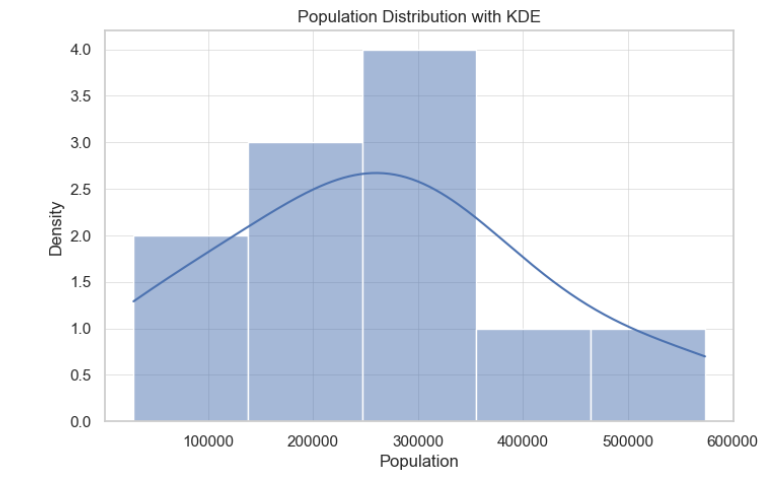
Binomial distribution was also used in terms of more practicality. The population was used to find the probability that an event (a population exceeding a threshold) occurs in a number of years withing a limit (50 years).



In terms of normal distribution analysis, this was done with the population distribution over the years.



Total Population Histogram



Sample Population Histogram

From the 2 histograms above this is an example of analysing the data with more focus on normal distribution. Theoretically, when a sample is taken and its distribution plotted, when the total of the population is plotted and its distribution plotted, the latter will naturally tend to be closer to a normal distribution. This is however not the case in the population distribution when a sample is taken versus the whole thing. In this case the sample looks to exhibit more of a symmetry in its normal distribution than when viewed as a whole. When viewed as a whole the distribution looks to be right-skewed signifying the mean is to the right of the median.

In terms of the rationales used for the binomial and normal distribution. Binomial distributions were employed in this situation as they are usually used for discrete random variable with a set number of runs where the probability of an event comes out equal with others. The binomial distribution was selected to see the probabilities that the population exceeded a number in the past. This can be used for future planning for multiple cases in a country. The variable selected, being the population threshold was appropriate as that is a key way to assess the country’s past profile but also from that it can be divided further into counties, gender, education income etc for further analysis, hence why the number of trials in this instance is synonymous for the number of years. It’s also assumed that the periods between the years are constant and set.

However, since this is also real-world data, the normal distribution also suits as it’s for a continuous distribution. As mentioned, the more samples assessed the more normal the distribution, however it seems like more samples are required to make the distribution selected more normal, signifying the sample size isn’t large enough for the Central Limit Theorem to be applicable. The variable chose for the distribution was also population, as it is numeric continuous data, however it doesn’t assume full normality.

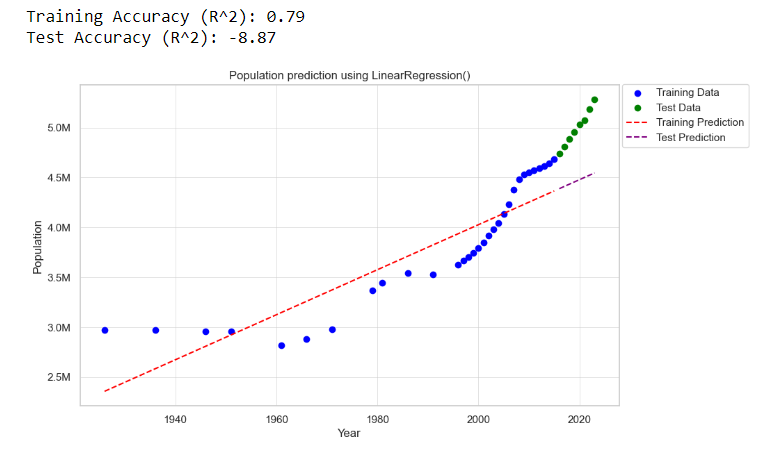
It is very important to choose the appropriate distribution for an analysis as it directly impacts the accuracy and validity of results.

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]**

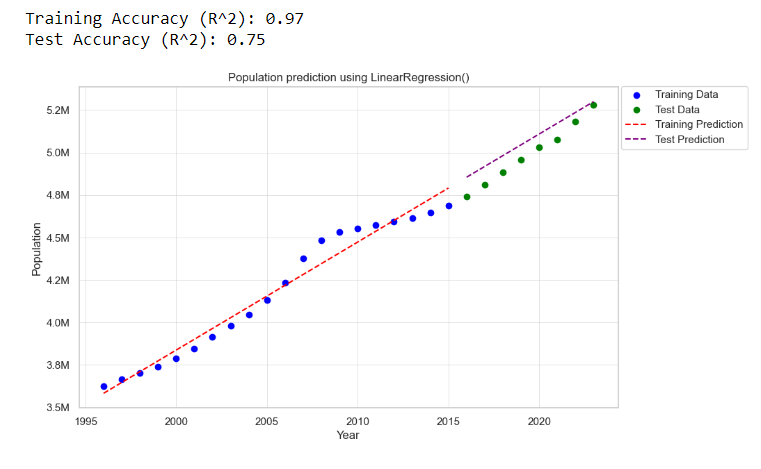
# Modelling and Evaluation

This is the phase of the CRISP-DM where the models are created based on the preparation and cleaning of the data. The first model developed was to predict the population of Ireland till 2040. This was inspired by Ireland project 2040 , which according to the Department of Public Expenditure, NDP Delivery and Reform, aims to plan and develop various socio-economic areas within the country, so 2040 will be where the predictions will go to for the analyses

For all models developed, the training, test split was 80/20%. Gholamy, Kreinovich and Kosheleva (2018) site 70/30 and 80/20 as the most common splits with a preference for the 80/20. So as a result, all models will use the 80/20 split. For repeatability and reproducibility, the random seed was set to 42 for all models.

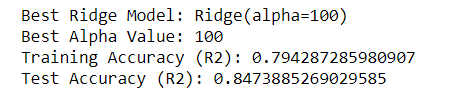


The first model developed was a linear regression model. It had a reasonable training accuracy but negative testing accuracy, so some additional data preparation was done to what was mentioned as a relatively constant population from 1926 onward. As a result, this part of the data was ignored for an alternate model.

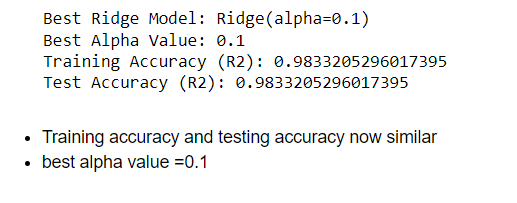


For the model, the training and test splits were also done in a way where the first 80% were used for the training and the rest for the 20% to give some continuity to the model prediction.

The next model developed was the ridge regression. Gridsearch CV was used to find the best alpha

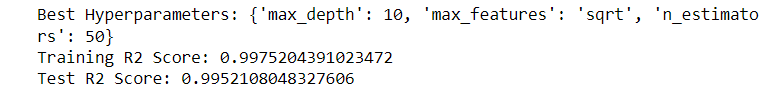


The testing accuracy came back higher than the test accuracy. Typically, if the difference was significant, it would indicate underfitting of the model, however this is not the case but is not ideal for modelling purposes. The same data preparation is done by ignoring the stable population.



The training and test accuracy are high and similar, which signifies the data was fitted well.

The last regressor used was random forest where Gridsearch was used again to find the best combination of hyperparameters. Results below

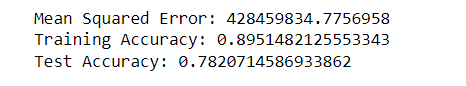


Comparing all 3 regressors used

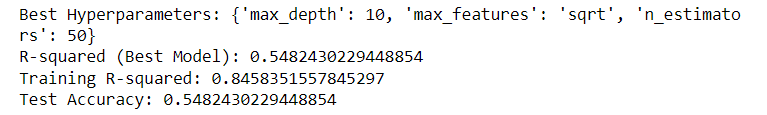
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Gridsearch use | Hyperparameters | Training Accuracy | Test Accuracy |
| Linear Regression | No | Default | 97.00% | 75.00% |
| Ridge Regression | Yes | alpha=0.1 | 98.33 | 98.33 |
| Random Forest Regression | Yes | max\_depth=10, max\_features=’sqrt’, n\_estimators=50 | 99.8% | 99.5% |

Random forest regression looks to be the standout model

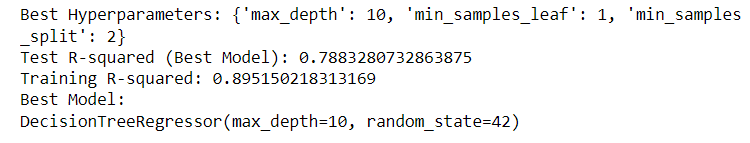
The next modelling was done for the educated population. The first model used was random forest



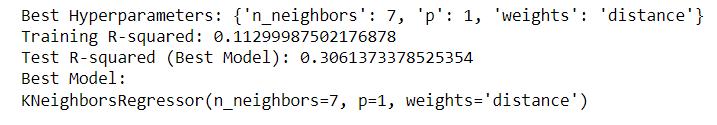
The training and test accuracies are good, but GridSearchCV was used to try and optimise



The accuracies were lower so as a result, the default hyperparameters were selected instead. The next model was the decision tree regressor, optimised with Gridsearch.

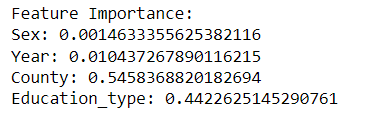


The final model developed was K-Nearest Neighbours which had the lowest parameters even with Gridsearch.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Gridsearch use | Hyperparameters | Training Accuracy | Test Accuracy |
| Random Forest Regression | No | Default | 89.50% | 75.2% |
| Random Forest Regression | Yes | Max\_depth=10, max\_features=’sqrt’, n\_estimators=50 | 84.60% | 54.82% |
| Decision Tree Regression | Yes | max\_depth=10, min\_samples\_leaf=1, min\_samples\_split=2 | 89.5% | 78.8% |
| K-Nearest Neighbours Regression | Yes | n\_neighbours=7, p=1, weights=’distance | 99.8% | 99.5% |

The optimised decision tree looks to be the best regression model in the second case. As there are 4 input parameters into the model, it was insightful to run the feature importance to see how much the data inputted in affected the model.

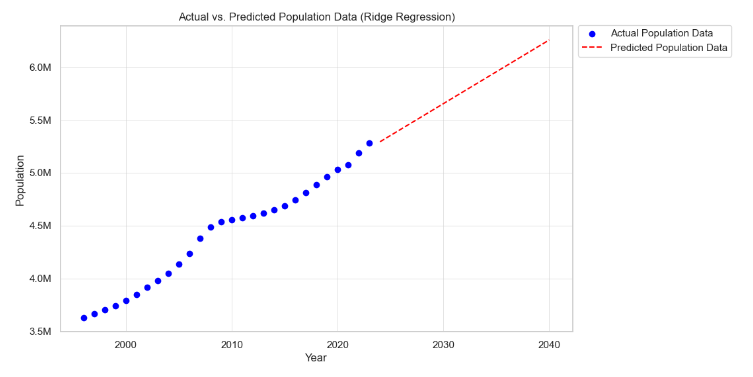


The county value affected the model the most.

# Deployment

Now that the optimum models were selected , they were used to predict the demographic population

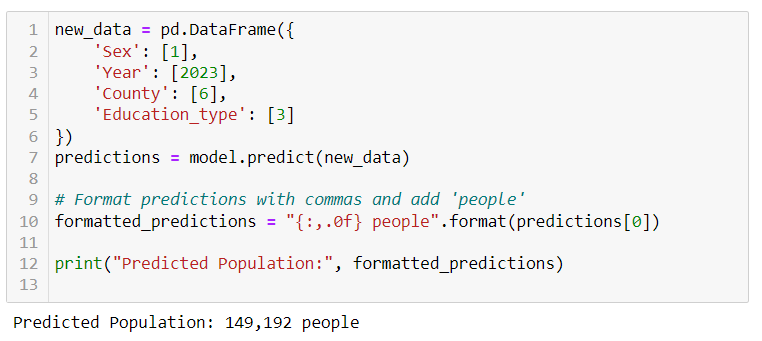
For the population prediction, the random forest prediction was initially chosen as it had the highest testing and training accuracy, however through the iterative method of this phase of the CRISP-DM methodology, it was proven that the predictions flatlined and as a result the ridge regression was taken as that was identical in accuracies. So, to get the predicted population in Ireland in 2040:



The population was predicted to be as per below.



The second prediction model was to estimate the population of people in a certain county, in a certain year, being either male or female who would’ve attained a certain level of education as their highest. As part of the deployment, the input values were mapped for best practices.



# RESULTS AND CONCLUSION

This project implemented comprehensive data analysis of the population of Ireland, its demographic and its future. The overarching management methodology was CRISP-DM due to its iterative nature and commonness. In terms of key findings, Ireland has witnessed steady population since the 60’s and expected to keep rising till 2040. The models developed that proved superior to the key aim were Ridge Regression and Random Forest due to their high accuracies and similarities in the training and test accuracies, signifying, perfectly fitting the model.

In terms of practicality, this project can be used as part of future planning , in this specific case, if the population predicted in a county and their educational background is know this could influence an array of things such as more/less schools in the area, increase in companies due to increase of skilled labour and many more focal topics in a country’s future planning.

# 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

Gholamy, A., Kreinovich, V. and Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (CS)*. [online] Available at: https://digitalcommons.utep.edu/cs\_techrep/1209 [Accessed 4 Nov. 2023].

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].

KH (2023). *Gender Pay Gap at 13% in EU, Luxembourg at Top with 0.7% Gap*. [online] Chronicle.lu. Available at: https://chronicle.lu/category/surveys-reports/43308-gender-pay-gap-at-13-in-eu-luxembourg-at-top-with-0-7-gap [Accessed 11 Nov. 2023].

Malone, P. (2020). *Irish Population Health: life expectancy and mortality*. [online] Available at: https://publicpolicy.ie/wp-content/uploads/2020/11/Irish\_Population\_Health\_Life\_Expectancy\_and\_Mortality.pdf.

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.