The Aging Irish Population

# Abstract

This project explores the Irish Population by Region and Age Group, with particular focus in parts on the change in ratio of Working, Retired and Dependent Populations over time.

# Introduction

Data gathered from the CSO regarding Irish Population growth over a range of variable including Age, Region and Sex. With inspiration from the works of Harper (2016) regarding aging populations and the potential issues that future generations will face as a result.

*Accompanying Jupyter Notebook can be access through Github*:

<https://github.com/Whyvonne/CA1_MSc_DA_Yvonne_Smyth.git>

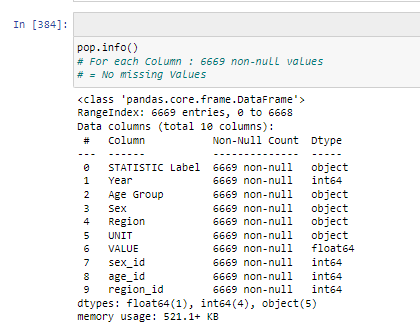
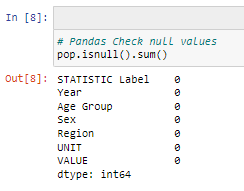
# Dataset 1: D1: 2011-2023 Irish Population, by Age, Sex & Region

## Exploring D1

The dimensions of D1 are 10 Columns with 6669 rows/observations. From reviewing the head and tail (top and bottom), it is apparent that the data is split into several categories including Age Groups, All Ages, Individual Sex, Both Sex, individual Regions and State.

It important to review the shape of the data to become aware of any aspects that may affect plotting and subsequent interpretation. It also allows for review of any missing or null values, which should be dealt with on a case-by-case basis; with the correct solution based on the level of direct influence it will have on original research motives / questions and potential results.

As all columns state 0 “null” i.e. missing values, there is no investigation or corrective action required in this regard.



*pop.info* indicates that there are a variety of data types present. Integers and float are Numerical Variables, and objects are Categorical Variables.

## Numerical Variables

Python function *pop.describe* shows the dimensions of the data that python recognises as numerical.

There are 2 Numerical Variables:

A screenshot of a data

Description automatically generated

## Categorical Variables

Python function *pop.describe(include=object)* shows the variables that are assigned the "object" data types by python upon import of the data. The observations within these columns typically contain strings. Reviewing unique values will assist with identifying number and type of Categories.

There are 5 categorical Variables;

A screenshot of a computer

Description automatically generated

## Feature Engineering

To facilitate analysis and ease of manipulation of the categorical variable in Python, they are converted/assigned unique numerical equivalents.

It is important that during this conversion, the categorical variable is not changed in any way, other than the label that represents it i.e. the numerical equivalent should not change the "meaning" of the variable.

### Convert Categorical Variables into Numerical (DPREP)

A screenshot of a computer

Description automatically generated

#### Convert Sex

A screenshot of a computer

Description automatically generated

#### Convert Age Group

A screenshot of a computer

Description automatically generated

#### Convert Region

A screenshot of a computer

Description automatically generated

### Subsets

Multiple subsets of the original dataset were created to facilitate exploratory analysis. This included the summary variables; All Ages, Boths Sexes and State and combinations of the other grouped variables.

Various python data manipulation techniques were employed including, but not limited to those listed below.

Many of the resulting subsets are catalogued in *df\_index\_table.*

#### Python Methods Implemented

|  |  |
| --- | --- |
| **Method** | **Objective** |
| .loc | Reduce the dimensionality of the dataframe by filtering by a desired criteria |
| .pivot  (Pandas) | Reshape / restructure a dataframe based on desired column while aggregated values those desired columns |
| .groupby  (Pandas) | Split dataframe into groups, or aggregate based on a value from a desired column |
| .drop | Drop unnecessary columns from a dataframe |
|  |  |
|  |  |

## Data Exploration & Visualisations

### Plotting & Tufte's Principles

Tufte (2007) 6 principles of graphical integrity (summarised):

1. Numbers presented should be directly proportional to the physical representation on the visualisation
2. Labelling should be Clear, detailed and thorough, with zero ambiguity. Explanation and flagging of important data should be present, but not distort the visual
3. Focus should remain on data variation, not design variation.
4. Use standardised unit of monetary measurements in time-series displays. Avoid nominal units where appropriate
5. Quantity dimension of information variable never to exceed number of dimensions displayed.

All visualisations in this project seek to follow these principles and adhere to their fundamental intent. An explanation of how these principles, and other guidance from Tufte, were considered during choice and design of visualisation is included throughout, with some showing the adjustment and evolution of design to illustrate application.

### Irish Population 2011-2023 Visualisations

#### Region Population Over Time

A graph of different colored lines

Description automatically generated

**Motives & Design**

To show the change in population across the Regions in Ireland over 2011-2023

A Lineplot is a simple and effective way of comparing multiple numerical values across a time period. This plot is skewed by the largest of the Regions, Dublin. This region will be removed and plotted separately so the other can be compared in a meaningful way.

**A graph of a number of people

Description automatically generated**

**Design:**

The reduced axis on this plot means it is much easier to review and compare the changes in population for each region individually, and to compare regions. An improvement would be to order the plots in the same as the population values highest to lowest, which would decrease the processing require for the reader to match the line colour to the corresponding legend line.

**Interpretation:**

All regions show and increase in populations at a steady rate. This was to be expected considering the general increase across the state.

The lineplot below of Population in Dublin 2011-2023 shows an increase over the time period, matching the other regions.

The Dublin and All Other Regions plot should be use for comparison as they have different axes. The reader is likely to interpret that Dublin has had an accelerated increase compared with Other Regions over the same time period, when they have all experienced approx. 10-15% increase.

A graph showing the growth of the number of people in the world

Description automatically generated

#### Boxplots of Population by Region

A graph showing a box plot

Description automatically generated

#### Male vs Female by State Population Over Time

**A graph with a green line and orange line

Description automatically generated**

**Motives & Design**

To review the change in population over time for the state, and split by Male / Female

As with the above plot, the theme and colours used are clear and simple. A new colour is introduced for the stat population; again it is a contrasting yet complimentary colour. Colours representing Male and Females are kept consistent to maintain reader processing time at a minimal amount.

**Interpretation:**

As expected, this plot also shows a steady increase in state population from 2011-2023. The overall increase is approximately 700. (700,000 actual population)

**A graph showing the number of people in the irish population

Description automatically generated**

**Motives & Design**

To review the change in population over time split by Male / Female

A line plot clearly shows the change of numerical data over time. As per Tufte's principles and general guidance (2018), a focus is maintained on the data-ink ratio and reducing all unnecessary "noise" from the plot. This also allows the reader to focus solely on the data, and not on the design. Contrasting Colours are used for Male and Female, but they are complimentary so as not to create strain for the reader while they review the data.

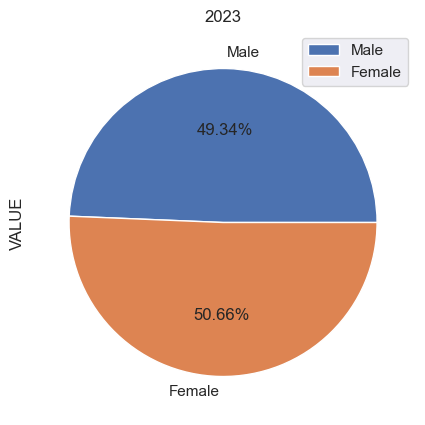
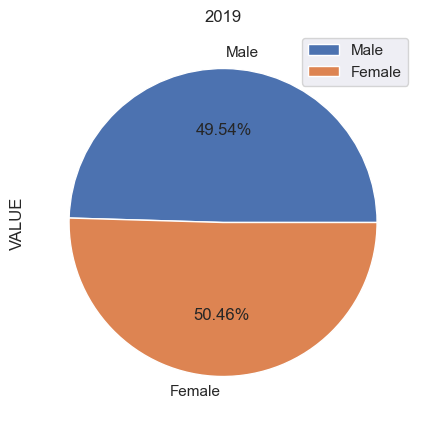
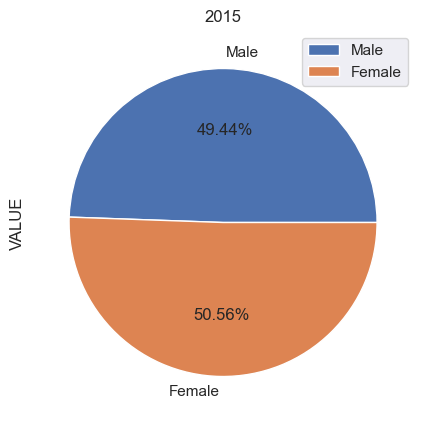
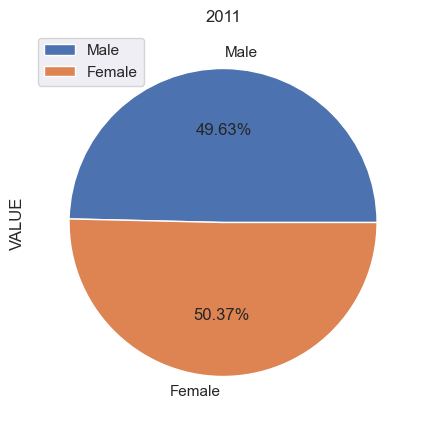
**Interpretation:**

There is a steady increase in population across the State for both Male and Female over the 12-year period. Female maintains a higher population by approx. 50 (50,000 actual population) throughout the time period. There is a small blip in the steady increase for both sexes in 2021. One reason for this may be an increase in death rate as a result of the Covid-19 Pandemic, but more in-depth investigation analysis would be required to confirm this, or find the correct cause.

#### Male vs Female by State Population: Pie Charts by Year

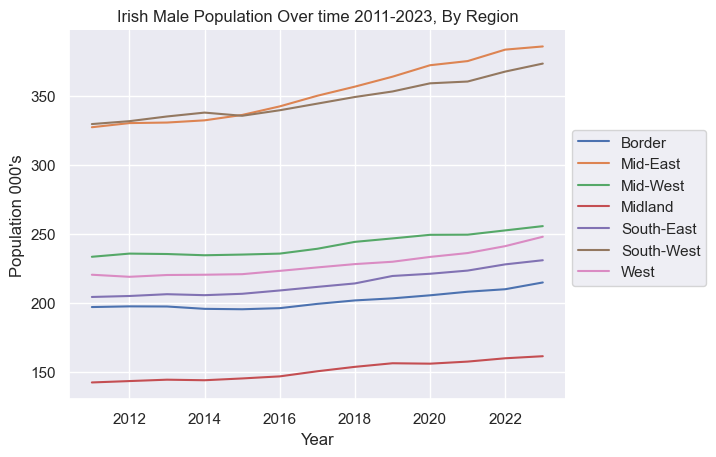
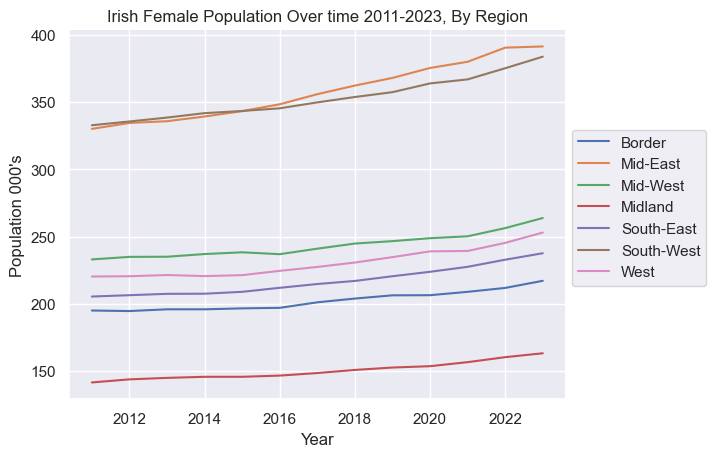
Pie Charts are a method of clearly exhibiting proportionality between a small number of variables. Here they show that there is a consistent split in the Male / Female Population in Ireland over 2011-2023. While Female maintains a 1% higher level throughout the time period, the population level of females is also trending upwards.

This trend is also evident in the line chart above, with female population beginning to divert from the parallel growth after 2021.



#### Male vs Female by Region Population

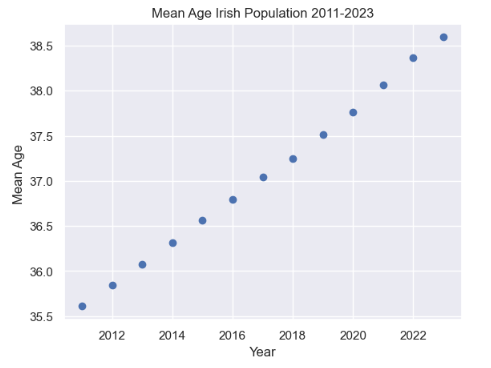
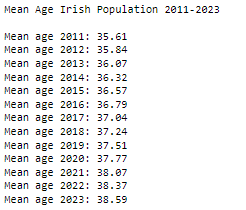
The Male and Female Population split by region supports the conclusions drawn from the previous plots. Both Sexes in each region are trending upward in line with the general population growth. For both sexes, the Mid-East Region appear to reach a plateau in 2022, especially evident in Female Population. Female population in each other region also appears to be trending up at a high rate than Male. The gradient of each line is increasing compared to that of the Male population for the same region.



#### Mean Age of Population

The Mean age of the population is trending upward, with approx. rise from 35.61 in 2011 to 38.59 in 2022, an increase of 3 years over 12 years.

This supports further research into the spread of age across the populations, and investigation into growth and relative ratios of the population at different ages.



## Working Age Population vs Retired Population

Organisation for Economic Co-operation and Development (OECD) reviews this data worldwide and provides the below % Population comparative information on their website:

*Source:* [*https://data.oecd.org/pop/elderly-population.htm#indicator-chart*](https://data.oecd.org/pop/elderly-population.htm#indicator-chart)

## Working Age Population = Age 15-64

A screenshot of a graph

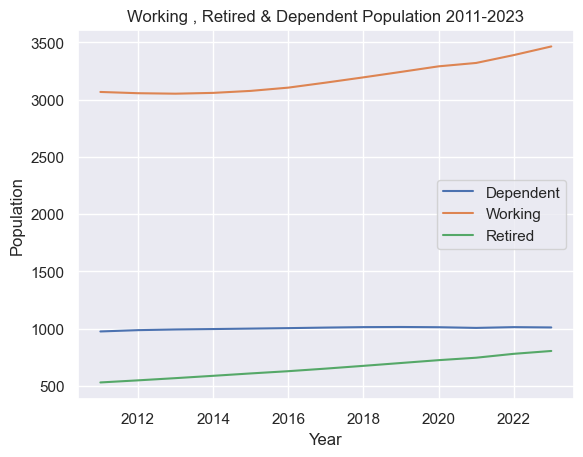
Description automatically generated

Elderly (Retired) = Over 65A screenshot of a computer

Description automatically generated

A table with numbers and numbers

Description automatically generated



# Dataset 2 : D2: Ireland Population 1950-2023, by Age & Sex

## Exploring D2

The format and content of D2 is similar to that of D1 with an extended number of years, but not split by region. D2 will be used to review and explore the change in Dependent, Working and Retired / Elderly Population over an extended period of time (70+ years).

A screenshot of a white page

Description automatically generated

There are additional age groups in D2 that overlap with the standard groups which will be removed. There are null values in D2 that should be investigated before analysis can take place.

A screenshot of a computer

Description automatically generated

The below groups will be kept facilitating exploration of the Irish population across all ages groups.

A screenshot of a computer screen

Description automatically generated

All null values appear in Age Group 0-4 years, so with that age group removed, no null values remain. This age range is served within *Under 1 year* and *1-4 years.*

A screenshot of a computer

Description automatically generated

These age groups were then segregated into their larger age sectors of Dependent, Working and Retired

A table with numbers and numbers

Description automatically generated

## Working, Retired & Dependent Population

### A graph showing the growth of population Description automatically generatedWorking Population

A Steady increase in working population from approx. 1600 in 1960, to 3500 in 2023.

A graph showing the growth of population

Description automatically generated

### Retired Population

The retired population has grown by approx. 110 from 1950 to 2000, and then experienced exponential growth of over 350 in the next 20 years.

A graph showing the growth of population

Description automatically generated

### Dependent Population

The dependent population has grown to a peak of 1040 and declined again to under 850. It has remaining at approx. 1000 since 2015.

### Combined Dependent, Working & Retired Population

A graph showing the number of retirement age

Description automatically generated

### Working Population Vs Retired Population

A graph showing a line graph

Description automatically generated with medium confidenceThere is a positive correlation between Working vs Retired Population but, as the scatter plot below shows, this is not a steady or consistent relationship.

As Working Population increases from 2000 to 3000 (50% increase), the corresponding Retired Population increases from 370 to 470 (27% increase)

Using these approximate figures, the ratio of Retired to Working population decreased from 18.5% to 15.6%.

# Choropleth Maps of Ireland

## CSO Classifications of regions:

The CSO provide additional information regarding the counties that make up the regions. These will be used to plot a map of Ireland for some statistics.

<https://www.cso.ie/en/methods/classifications/standardcountiesandnutsregions/>

## Plotting the Choropleth

**Motives & Design**

For those readers who may be unfamiliar with the geography of Ireland, or for those who wish to clarify which counties are included the Regions in he data, context should be provided. The below Choropleth / Map visualisation uses the Region to apply a different colour to each region, with definition of Region provided in the legend.

This contextual information is best visualised on a map as this is the format the reader will be familiar with. It is more informative at a glance, than a table showing counties and their associated Regions. This follow Tufte's intent of making the visualisation highly accessible to the reader, and reducing the processing time to as little as possible.

The colour map used for this plot is an alternative to that of the below plots, as this plot is for informational / contextual

A map of ireland with different colored states

Description automatically generated

## Choropleth: Population by Region

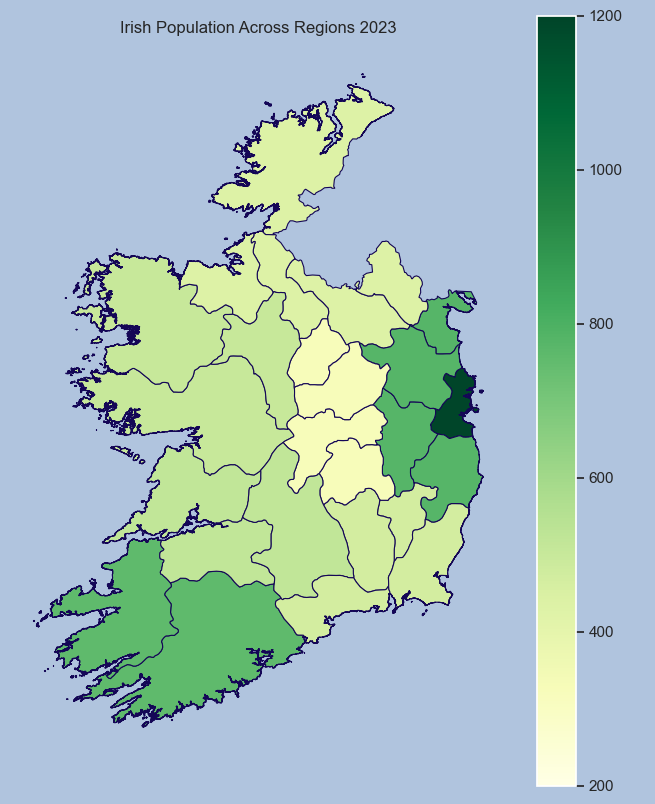
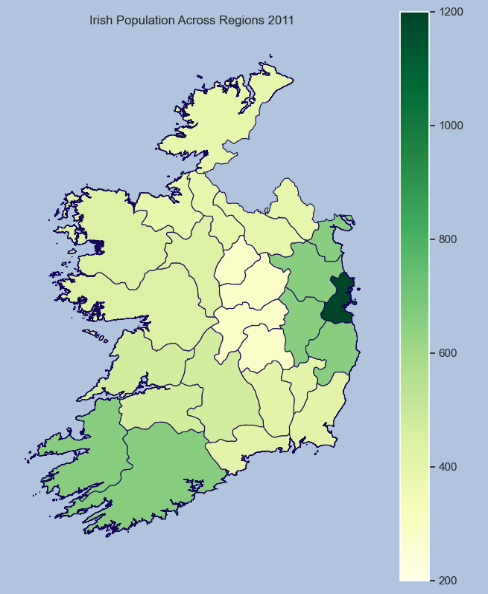
## Population by Region

## Plotting the Choropleth

**Motives & Design**

The goal with this visualisation is to allow the reader to review the change in population across the regions over the given time period.

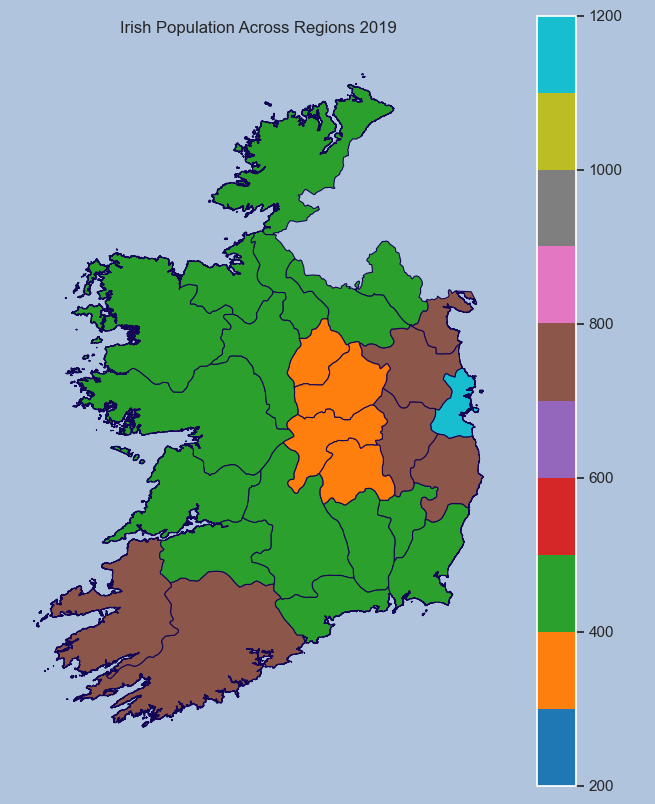
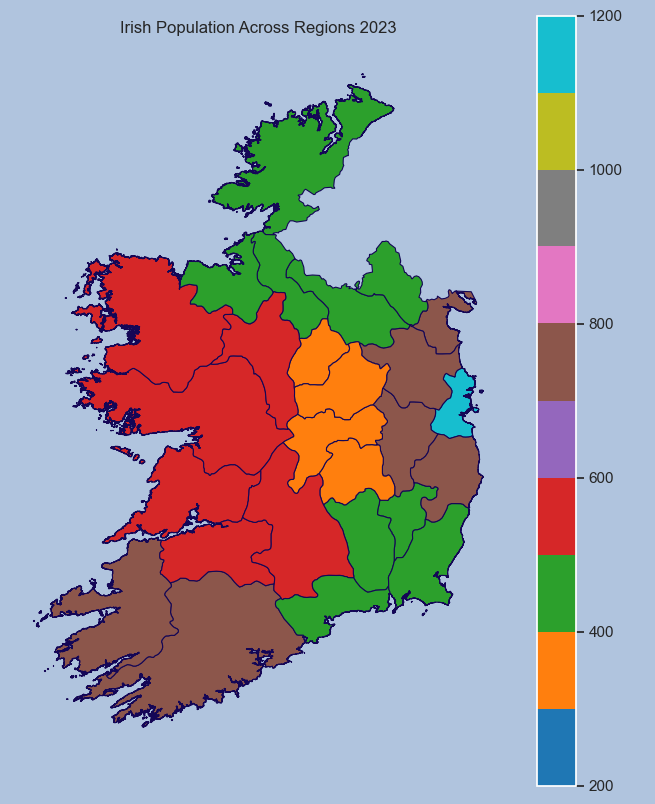
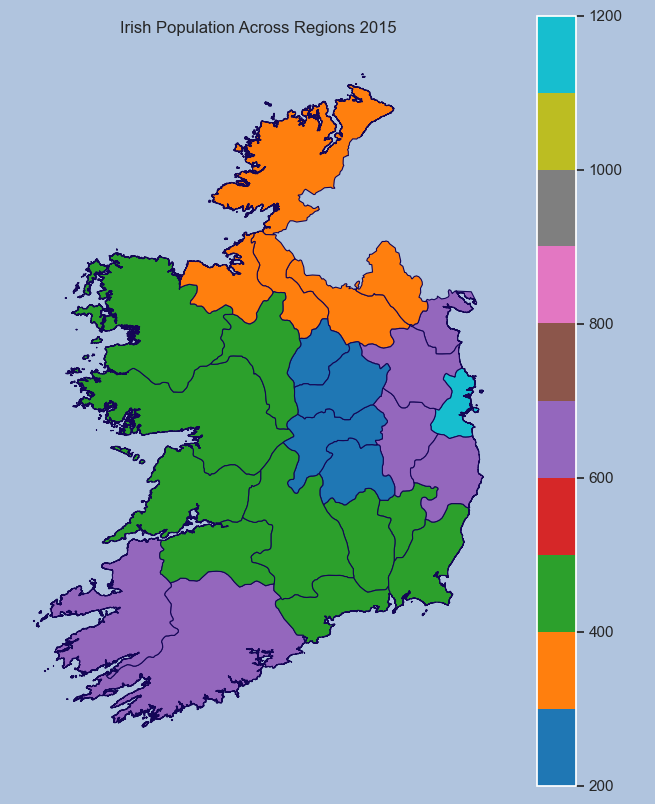
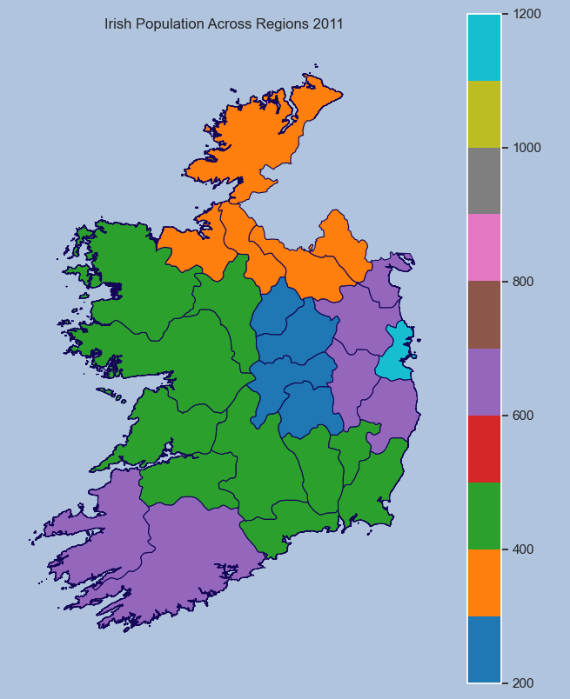
Readers are likely familiar with Ireland plots in shades of green and this would be the preferred colour to show the varying population levels across the regions in the country. However, as seen below, there are not a significant enough difference in the sliding scale of shade Yellow-Green to adequately communicate the population level differences.



**Design Evolution:**

Adjusting the Matplotlib Colormap to tab10 allows for better indication of both difference across regions, and also the actual figures. From reviewing the min and max of the populations levels it is seen that the range is approx. 300 - 1500. As the colour shades are split into 10 sections, this range of population is applied at 1,000. This means each colour represents a range of 100, with the intent of increasing the meaning of the colour to the viewer. It can be used to review the approx. (within 100) population level of the region, compare with others, and compare with the same region in other year plots. Consistency in set up across of these yearly plots is important so the reader can review and compare applying one set of definitions i.e red = 500-600 across all plots.

From previous exercises completed, it is know that there is significant jump between Dublin Region and the next highest population, so the gap between 1000 and the highest population of 1500 will not cause any issues with colouring here.



**Final Design:**

The plots clearly show the growth, decline and general evolution of the population across the regions over the time period 2011-2023. As this is a 12 year period, the plots are separated evenly into 4 year intervals. This allows the reader to follow growth / decline of individual regions, or compare with other regions over the given intervals. An possible improvement, or next iteration of this plot would be to create an animation showing each year so the reader could view the changes over time and pause at will. The reader would have to be viewing the plot in an interactive environment

## Choropleth: of Working / Retired Population Across Regions 2011/2015/2019/2023

A map of ireland with different colored states

Description automatically generated A map of ireland with different colored states

Description automatically generated

A map of ireland with different colored states

Description automatically generated A map of ireland with different colored states

Description automatically generated

**Final Design:**

The plots clearly show the growth, decline and general evolution of the population that are at working age across the regions over the time period 2011-2023. Again these plots are separated into 4 year intervals. Most regions are in a state of growth of working population.

# Statistics

## Scenario 1:

Choose 10 people at random from population in 2023, what is probability that less than 4 of them are from Mid-East

### Binomial

This meets the criteria for Binomial Distribution as we have a number of independent Bernoulli Trials. These are trials in which there are only two outcomes, success or failure. The criteria require to use Binomial Distribution are as follows:

1. Set Number of Trials

In this scenario the set number of trials is 10

2. Trials are independent

In this scenario, the trials will not affect one another and so they are considered independent. If the sample was taken from the same household or family group, then is would not be independent as the respondents have a relationship with each other.

3. Probability of success is constant

In this scenario the probability of success for trial remains the same for each trial. Regardless of the order in which we choose the 10 people, the probability that there is a success i.e. they are from Mid-East, remains the same.

4. Bernoulli Trials i.e a Binary Outcome

In this scenario there are only two possible results - Success or failure i.e. the chose person is either from Mid-East or not. There are no other possible options within the results gathered

5. The Data is Discrete

In this scenario the data only has values that are specific i.e. 1 person, 2 people, 3 people. It cannot take on any values in between these given discrete values.

### Workings

**A screenshot of a computer

Description automatically generated**

A screenshot of a graph

Description automatically generated*Source:* [*https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html*](https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html)



### Result:

The probability that less than 4 of the 10 people chosen at random from the population are from the Mid-East Area, is 95.30%

## Scenario 2 :

**Choose 25 people at random from Dublin Population in 2023, what is probability that exactly 15 of them are Male**

### Binomial

This Scenario also meets the criteria for Binomial Distribution as set out above.

### Workings

**A screenshot of a computer

Description automatically generated**

A graph of a function

Description automatically generated

*Source:* [*https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html*](https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html)



### Result:

The probability that exactly 15 of the 25 people chosen at random from the Dublin population are Male, is 8.78%

## Exploring Increasing Trials and Normality

By plotting Binomial Distribution for multiple trials of this event, we can explore how the distribution changes as the number of trials increases.

Generate results based on this trial being completed t times.

**Plot**

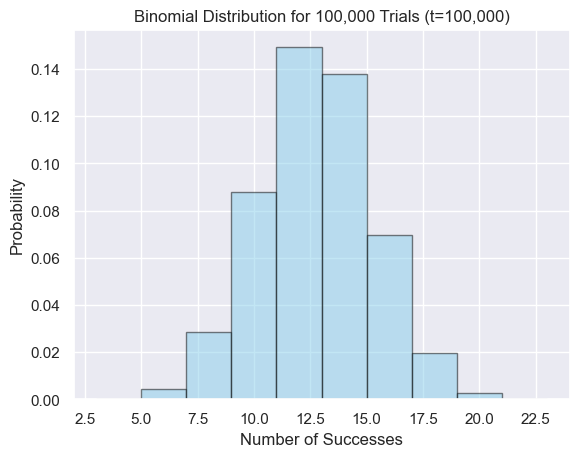
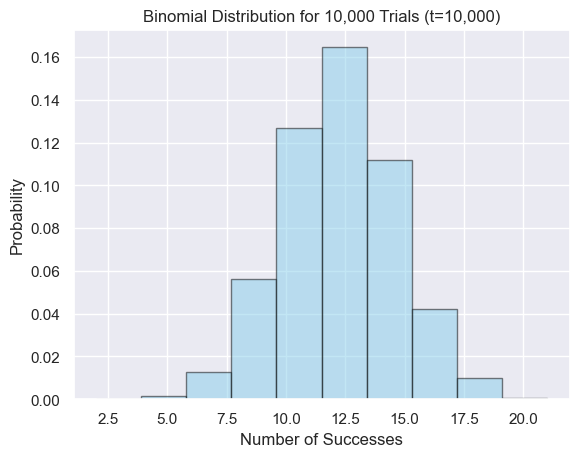
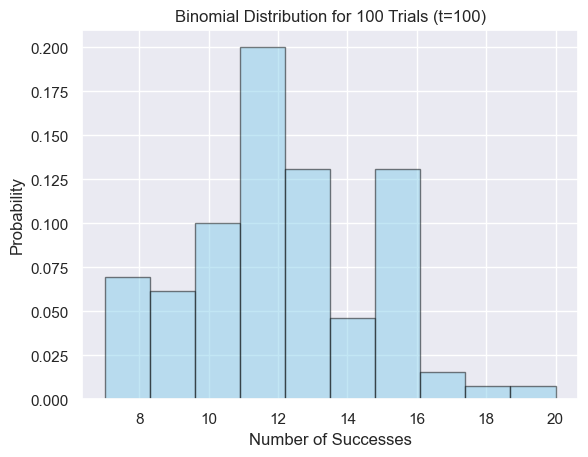
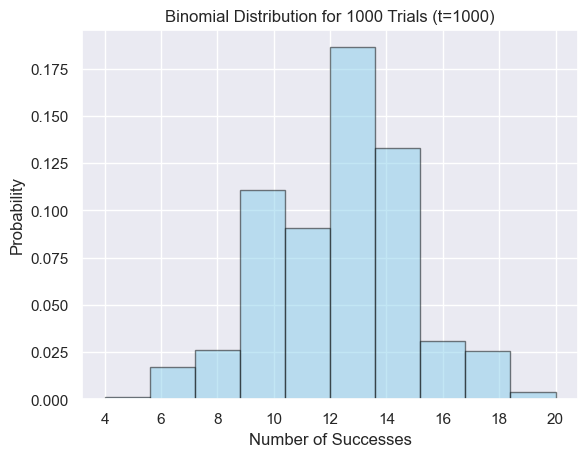
* t=10
* t=100
* t=1000
* t=10,000
* t=100,000

A graph of a number of individuals

Description automatically generated

Review of the first Binomial Plot above for t=10 trials reveals that the distribution is scattered and does not show "Normal Distribution". From Weiss (2017) a Normal distribution, also referred to as Gaussian Distribution, is symmetric and centered around the mean of the variable.

As the number of trials (t) increases throughout the Plots, the distribution of successes becomes increasingly Normally distributed. The distributed begins to resemble a bell curve more as the number of trials is increased.



### Central Limit Theorem & Law of Large numbers

Weiss (2017) explores the concept of the law of large numbers, and concludes that with a “large number of independent observations of a random variable X”, it is likely the average of the value will be close to the mean of the random variable X. This is the effect seen above.

## Exploring Normality

Normality can be assessed with regards to the age data in this dataset because although age appears to be measured in a discrete fashion, it is often considered to be continuous.

The below review of normality extends this consideration to the age group (and age\_id) as they are representative of age. The original categorical Age Groups have also been transformed into Ordered Numerical Variables, therefore their continuous attributes are considered to have been transformed also.

A graph of a number of people

Description automatically generated

The above plot showing Age and relevant % of Total Population appears to be roughly normal.

The plot for Normal distribution present as the shape of a bell curve, centered around the mean; which is roughly the case here. Based on the Age Midpoints / Mean Age exercise completed, the approx. mean age is known to be approx. 39 (age\_id = 9). The median is at VALUE=2641, which also lies within age\_id = 9. These matching measures of centrality would indicate that a normal distribution is present.

As the plot has 2 peaks and appears slightly right-skewed, normality can't be assumed from this plot, despite the supporting Mean / Median results. Other tests to confirm normality include Sharpio\_Wilk Test as seen below.

### Shapiro-Wilk Test:

A black text on a white background

Description automatically generated

The p-value > 0.05, therefore not in the rejection region. H0 is not rejected, and data can be assumed to be normally distributed. In the context of this dataset, this confirms that that Age does have a Normal Distribution across the population in 2023.

# Project Management Frameworks

Swamynathan (2017) outlines the steps in various project management frameworks, which can be seen in the image below. CRISP-DM (Cross-Industry Standard Process for Data Mining) present an iterative flow, which allows for evaluation and revisiting previous stages to adjust of the model through out its development. Developed just over 25 years ago, Wirth (2000) explored and evaluated the, then relatively new, concept and concluded that it was an answer to the common process model requirement the data community had been looking for. Martinez-Plumed (2021) explored the framework after 20 years and found that, although the industry has grown considerably, the CRISP\_DM framework is still appropriate for use, especially in goal-driven and process-orientated projects. Real life applications vary from Segmenting Customer in the Retail industry, to Fraud Detection in financial institutions and Government Tax agencies.

A diagram of a process

Description automatically generated

# Machine Learning Models

## Preparing the Data

### OneHotEncoder

Categorical Variables require transformation to be use in machine learning models. Burkov (2019) explains this is because many machine learning algorithms are based on mathematical equations and therefore cannot use categorical labels.

Some categorical variables have been transformed to numerical for plotting and ease of manipulation throughout the data exploration phase of the project. However they have been given ordered numerical Values, which will affect the machine learning model. Burkov (2019) reviews this matter and advises that non-ordered data should be encoded use the OneHotEncoder tool available on Python.

The Region and Sex will need to be transformed via the encoder, but Age can be ordered numerical data as the order of the Age groups forms an important part of any relationship identified. The main goal of machine learning models is to identify relationships and create a reliable model around strong relationships, therefore if the age variable maintains it's true definition during transformation, the model results that is created around it can easily be related back to the original categorical variables; in this case, the age groups.

### Scaling the Data

This is the first step; This helps to reduce the distance between any variable with large values and variable with small values. It brings the distance to a smaller, more manageable size while maintaining the proportion of the differences.

## 2011-2023 Population Dataset

Encoded values for Sex and Region are numerical, as is age\_id:

A screenshot of a computer

Description automatically generated

### Correlation between featuresA screenshot of a computer Description automatically generated

**From the Correlation Matrix:**

- Region\_Dublin and VALUE have a moderately positive correlation which means that more of the population is represented within the Dublin Region

- age\_ID and VALUE have a weak negative correlation which means that slightly less of the population is represented within the higher age groups

Neither of these correlations are strong enough to affect the Machine learning model to the point that they would cause over fitting, so they will not be removed from train / test variables.

### Supervised Model

A Supervised model approach will be taken as the data is labelled and is therefore suitable. The goal will be to test models that can accurately predict future Population Values based on the Region and Sex. This is also a suitable goal for a Supervised Machine Learning Model.

Dataset: popc\_ec\_r\_num

- Sex encoded value

- Region encoded values

- age\_id

A screenshot of a computer

Description automatically generated

There are no missing values in the dataset so no adjustment is necessary before beginning the Model creation and testing.

### Regression Models

### Metrics for Evaluation

For each of the below regression model, the below metrics will be requested and used to compare against the other models.

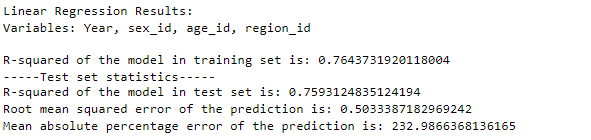
**R-squared (R2) :** Indicates the proportion of the variance in the dependent variable (target) that is predictable from the independent variable(s).

* Train vs Test score = Similar values = good fit
* Low Train and/or Low Test score = Underfitting
* R2 Train > R2 Test = Overfitting.

**Root Mean Squared Error (RMSE) :** Indicates the magnitude of error, in the context of the dataset in question.It is the square root of the MSE i.e. square root of the average of the squared differences between the actual value vs predicted values.

**Mean Absolute Error (MAE) :** Measures the average absolute error between predicted and actual values.

### Linear Regression



R2 score indicates 76% of the variance of the dependent variable within the data is captured by the model. The closeness in the train vs test R2 indicates that model is well fitted. RMSE of 0.5 in the context of this data indicates a large magnitude of error, as the scaling of the significantly reduces the range. This is supported by MAE of 232, indicating that, on average, the predictions are 232% away from actual values.

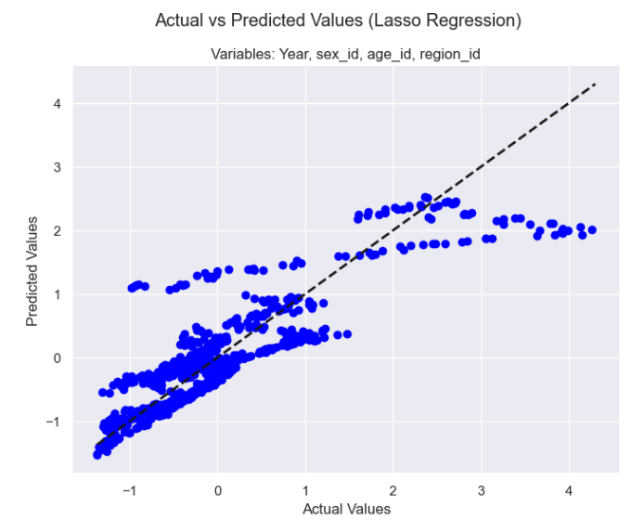
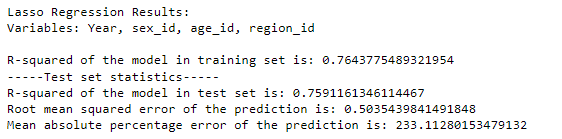
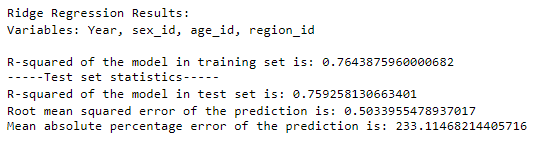
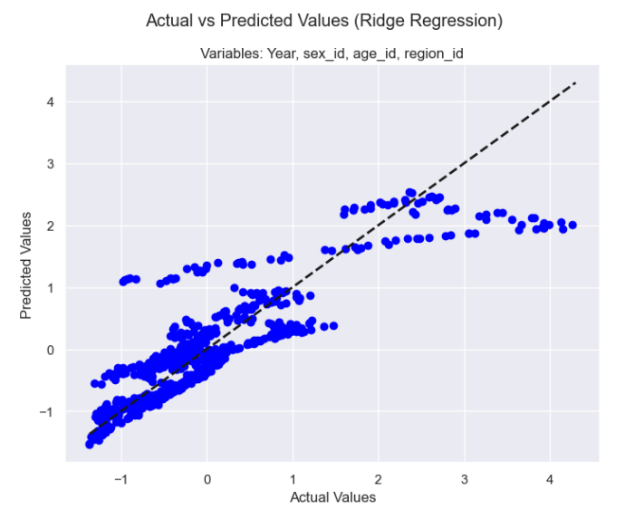
Plotting of predicted vs actual values, shows these errors as points that stray far from the line of prediction created by the model.

A graph with blue dots

Description automatically generated

### Ridge & Lasso Regression

The Lasso and Ridge model produce similar results to Linear model above and can be interpreted in the same way. From plotting, many points distant from the prediction line indicate a significant error margin.

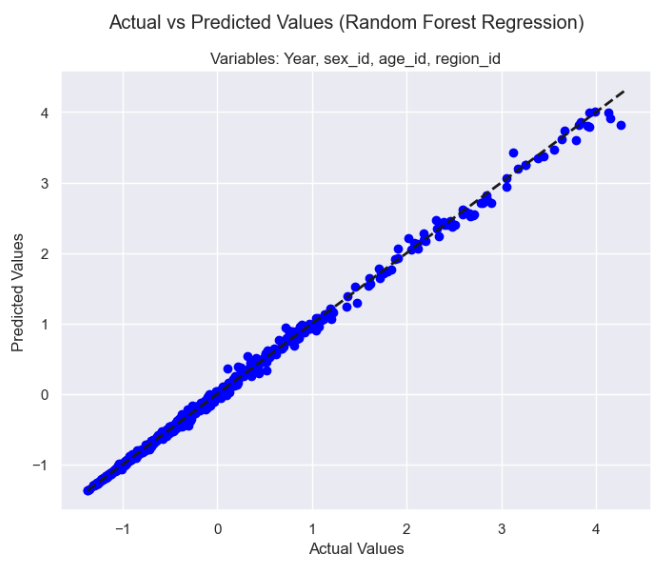
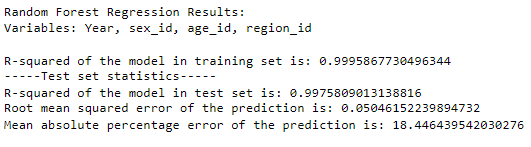
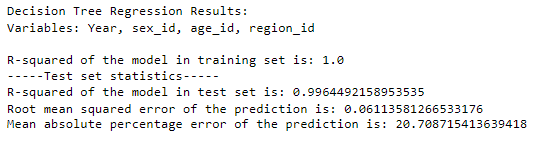
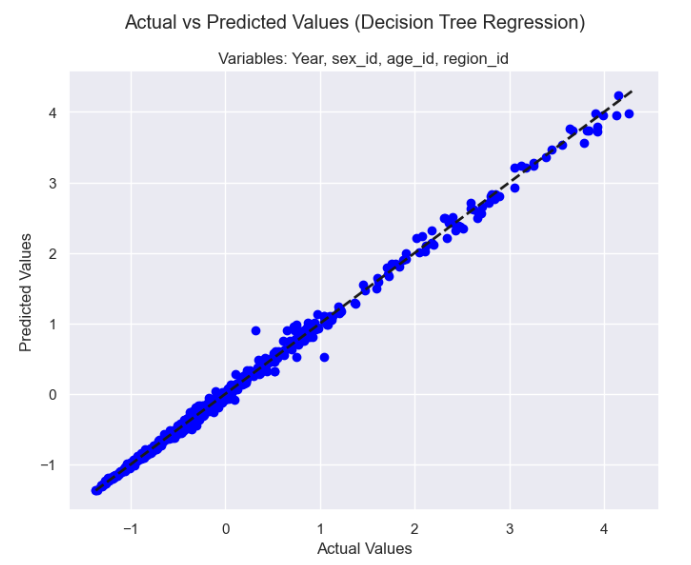


### Decision Tree & Random Forest

The Decision Tree and Random Forest Results are substantially improved compared with the previous models. High R2 values indicate 99% of the variance of the dependent variable is explained within the model.

Low RMSE scores 0.06/0.05 indicate low magnitude of error, even considering the scaling data. This is supported by the vastly reduced MAE of 20/18 respectively. On average the model are18%/20% away from the actual values.

This is reflected in the plots showing all points significantly closer to the prediction line.



### Hyperparameter Tuning - GridSearchCV

#### Lasso Before:

Testing Lasso with Alpha = 1 and 0.1.

From Results below, the R2 values are increasing and getting better but to find the best value of Alpha, GridSearchCV hyperparameter tuning should be implemented.

A white background with black text

Description automatically generatedA white text with black text

Description automatically generated with medium confidence

#### Lasso After:

A white box with black text

Description automatically generatedAs seen below, the best alpha parameter for Lasso Regression in this instance is indicated to be 0.001.

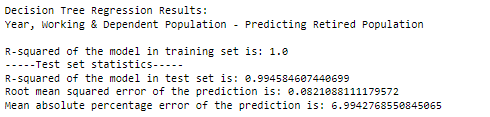
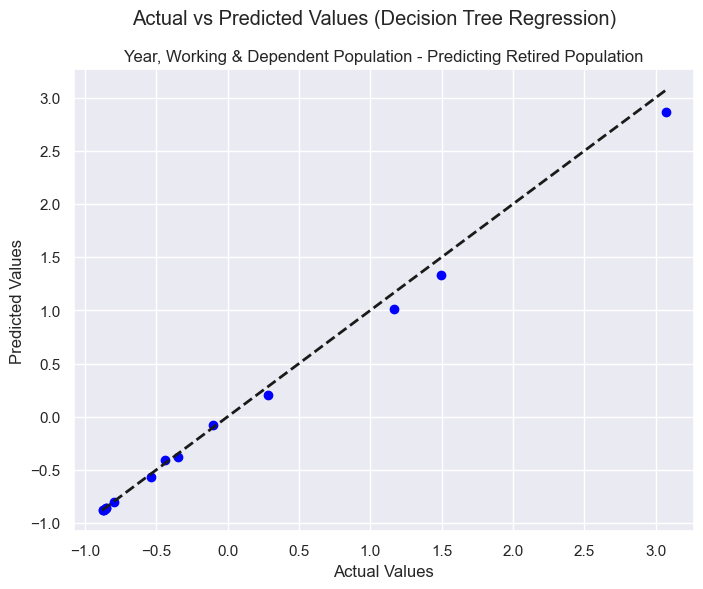
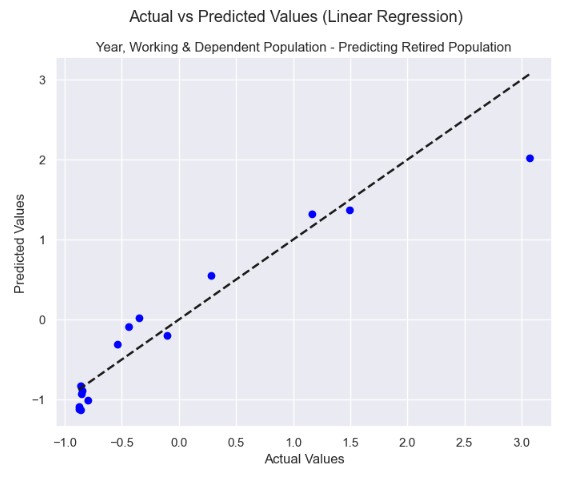
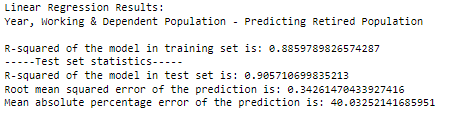
A black text on a white background

Description automatically generatedThis produces a higher R2 Value with similar Train/Test scores, indicating a good fit. The error values are quite high, indicating that may not be the best model.

## Regression: Working, Retired & Dependent Population 1950-2023

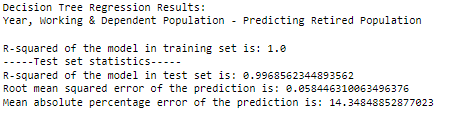
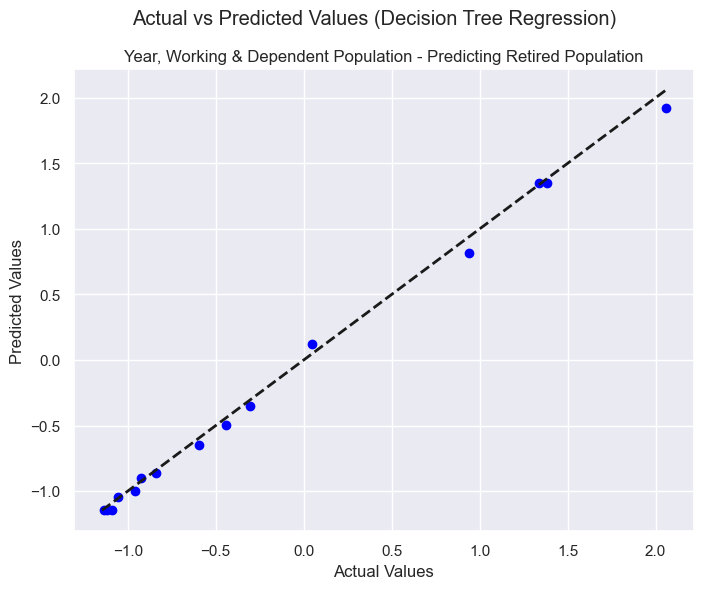
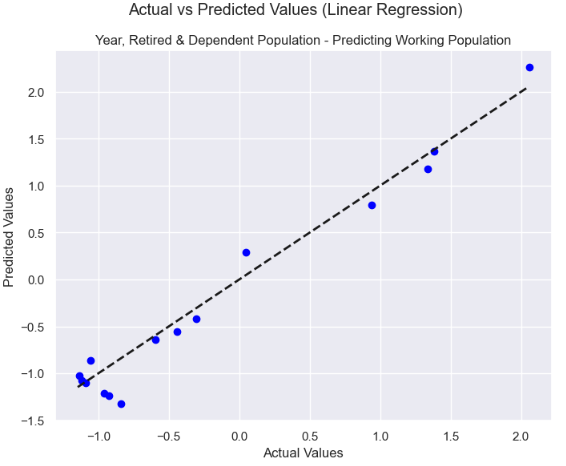
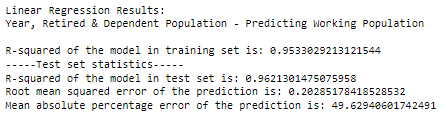
### Predicting Retired Population (Year, Working & Dependent Population)

#### Linear Decision Tree Regression



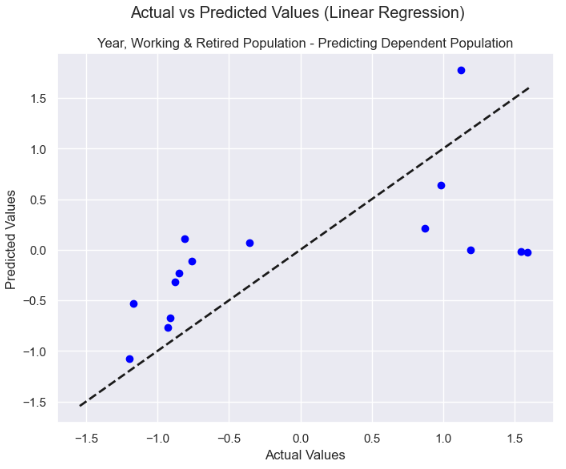
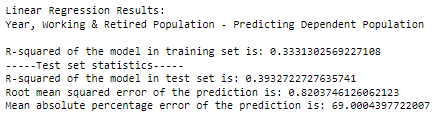
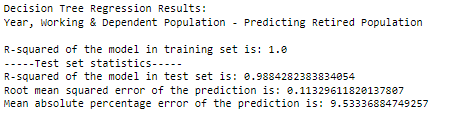
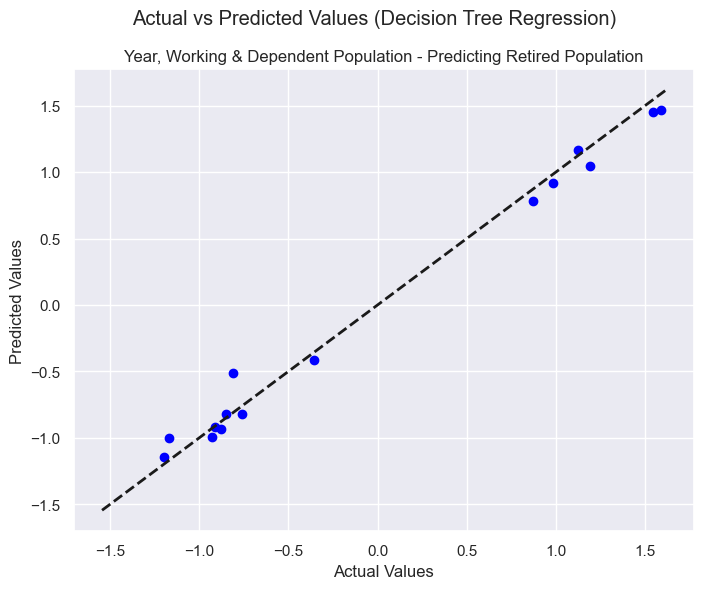
### Predicting Working Population (Year, Retired & Dependent Population)

#### Linear & Decision Tree Regression



### Predicting Dependent Population (Year, Working & Retired Population)

#### Linear & Decision Tree Regression



## K Mean Clustering

### Sum of Squared Distances - Elbow Method

A graph with a line

Description automatically generated

* Elbow here looks to be at 5.
* Check silhouette method to confirm.

### Silhouette Method

A graph showing the growth of a number of people

Description automatically generated

### Confirmed k=5

Create Clusters based on this:

A bar chart with different colored bars

Description automatically generated

From the above K means Cluster Bar Chart, it can be interpreted that most of the VALUE data is grouped into ClusterID 2

A bar chart with different colored bars

Description automatically generated

From the above K means Cluster Bar Chart, it can be interpreted that most of the age\_id data is grouped into Cluster ID 1 and 2, with a smaller amount in Cluster ID 4 and even less again in Cluster ID 3.

# Programming & Jupyter Notebook Design

## Extensions

Install & Enable Jupyter Notebook Extensions

**Enabled Spellchecker**

help with Spellchecking markdown entries as using exported Jupyter Notebook Markdown cells to create some report sections.

**Collapsible Headings**

Help with ongoing structure of Notebook and tracking of Sections & Sections Goals, including success failure of attempted plots and models.

## Print a list of all created dataframes

The below code creates a *df\_index\_table* to catalogue and describe all dataframes created throughout.

A screenshot of a computer code

Description automatically generated

This is placed at the at start of Notebook dataframe names and descriptions are added to this dataframe throughout using the below template.

A screenshot of a computer program

Description automatically generated

The datafames and subsets became extensive in number during the project. To prevent the duplication of work this cataloguing convention was implemented. As a later addition to improve ongoing project management, not all dataframes are entered here.

A potential improvement would be to upgrade and insert this method into a function and call the function to complete the addition to the DF Index Table each time.

*df\_index\_table* output:

A table of text with numbers and letters

Description automatically generated with medium confidence

# Conclusions

The Irish population is increasing year on year.

While working population is increasing also, the elderly /retired populations is increasing at a faster rate.

# References

*Data.cso.ie*. Available at: https://data.cso.ie/ (Accessed: 10 November 2023).

Wirth, R. and Hipp, J., 2000*, April. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (Vol. 1, pp. 29-39).* [Preprint].

Burkov, A. (2019) *The hundred-page machine learning book*. S.L.: s.n.

*Demography - working age population - OECD data* (no date) *theOECD*. Available at: https://data.oecd.org/pop/working-age-population.htm#indicator-chart (Accessed: 10 November 2023).

Harper, S. (2016) *How population change will transform our world*. Oxford University Press.

Martinez-Plumed, F. *et al.* (2021) ‘CRISP-DM twenty years later: From data mining processes to data science trajectories’, *IEEE Transactions on Knowledge and Data Engineering*, 33(8), pp. 3048–3061. doi:10.1109/tkde.2019.2962680.

Swamynathan, M. (2017) *Mastering machine learning with python in six steps a practical implementation guide to Predictive Data Analytics using python*. New York, NY: Apress, Springer Science+Business Media.

Tufte, E.R. (2018) *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Weiss, N.A. (2017) *Introductory statistics*. Boston etc.: Pearson.