MSc in Data Analytics

CA1

Author: Maria Koleva

E-mail: sba23020@student.cct.ie

Student ID: 23020

Github: https://github.com/mkoleva0/DataAnalytics\_CA1

**Abstract**

*The study looks into population dynamics in Ireland, through the application of different machine-learning algorithms, with a specific focus on Regression and Classification techniques, to predict future population and its relationship with migration. The analysis confirmed a consistent growth in the population over the past several decades, a path that is projected to continue. A correlation between net migration and demographic fluctuations was identified within the selected timeframe, with periods of increased migration influencing the growth of the population.*

**Introduction**

I’ve completed the analysis using Python, and executed the code in a Jupyter notebook. This is a handy way of easily processing the code, and visualizing the trends, along with using powerful libraries that Python offers for data manipulation, visualizations, statistical analysis, and machine learning algorithms, such as pandas, seaborn, numpy, scikit-learn etc. (McKinney, 2022)

The pandas library was used for data manipulation, which is a key for its high-performance, easy-to-use data structures, and data analysis tools. This library's functionality was essential for cleaning, transforming, and aggregating data necessary for this demographic study.

When it comes to visualizing trends, seaborn was used, which is built on top of matplotlib and provides a high-level interface for drawing informative statistical graphics. This library simplifies the creation of complex visualizations.

NumPy, another library in the Python data science stack, is fundamental for scientific computing in Python. It provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

https://numpy.org/doc/stable/user/whatisnumpy.html

In addition to these, scikit-learn was included for its machine learning

I expanded our analytical methods by incorporating statsmodels.api, a library offering numerous classes and functions for estimating a variety of statistical models, conducting tests, and exploring statistical data. For instance, in my analysis, I utilized statsmodels.tsa.arima.model to implement ARIMA models, which proved essential for forecasting time series and predicting demographic changes over time. Python's itertools module played a key role in our process, allowing us to iterate effectively over data and model parameters. This module offers a suite of tools for creating and managing iterators, enabling the efficient generation of complex iterators while conserving memory. This feature was especially helpful for dynamically adjusting parameters during model optimization. By strategically employing Python's comprehensive programming features and its potent analytical tools, we established a robust framework for in-depth data analysis. This approach ensures our study of demographic and migration patterns in Ireland is executed with both accuracy and precision.

I’ve followed the good practices of Python, such as naming variables in a clear and understandable way, and defining functions where possible in order to avoid repetition of code. However, rather than following the Object-oriented programming principles, as encapsulation and inheritance, my code is characterized by a step-by-step approach where data and functions are separate, and the code is organized into procedures and functions that are called as needed

**Methodology**

My research into Ireland's population and migration trends used a structured approach called CRISP-DM. This is like a roadmap for my study. The six phases of CRISP-DM guided me from understanding the business problem to deploying the model which will be followed in this report as well.

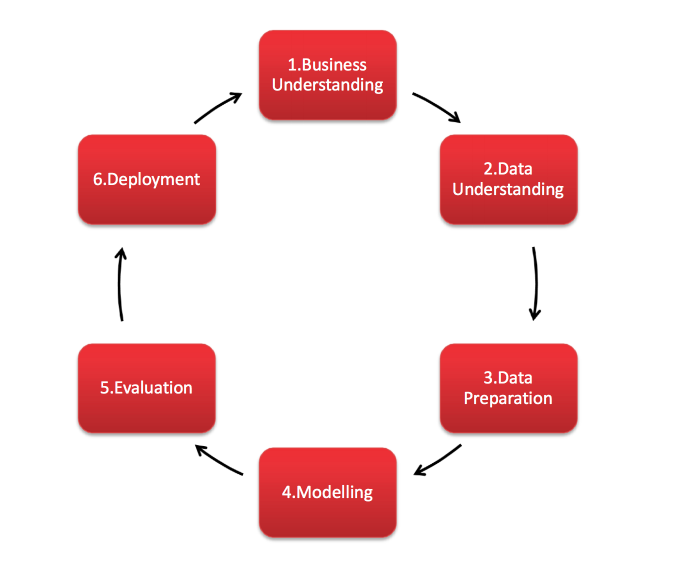


Figure 1

**Business Understanding**

The purpose of this analysis is to study Ireland’s changing population figures and migration flows over a period spanning from 1996 to 2023. My focus will be on the following areas:

* Investigating the patterns and trends in net migration, including inflows and outflows and their impact on population size and structure
* Examining shifts in the population’s age groups, noting trends in both the younger and older segments

Understanding these demographic changes is important because:

* It helps predict the need for public services like healthcare and education.
* It offers valuable information for the labor market to guide employment policies and economic plans.
* It supports urban planners and social service organizations in planning for housing and community needs in the future.

The evaluation of these elements provide a comprehensive report that enhances understanding of Ireland’s demographic changes over the specified period.

**Data Understanding**

In a 2018 Harvard Business Review article, Redman states, "Poor data quality is enemy number one to the widespread, profitable use of machine learning." He also notes the challenges bad data can pose, especially during the training of models and their real-world applications. Understanding the importance of good data, this study uses three datasets from the Central Statistics Office (CSO) to examine Ireland’s population and migration trends.

The three datasets used in this project are as follow:

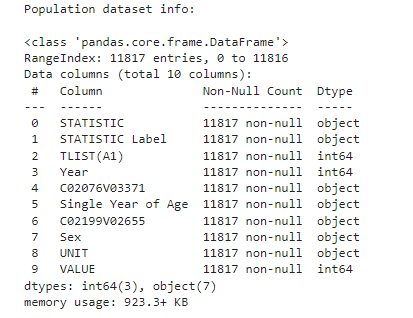
1. PEA11.20231013T111046.csv - Population estimates from 1926 with Age and Sex breakdown
2. PEA18.20231016T141007.csv - Estimated Migration (Persons in April) with Country and Sex breakdown
3. PEA03.20231022T221016.csv - Estimated Migration (Persons in April) with Age group and Sex breakdown

The first dataset tracks the shifts in Ireland's population over the years and includes details about the age distribution. The second and third datasets, on the other hand, tell us about migration patterns, detailing both incoming and outgoing movements. Notably, those datasets also break down by the countries of origin, as well as the age, allowing us to identify the primary sources of immigrants to Ireland.

Before analysing the data, it's important to first take a good look at it. This means seeing how much data there is, what kind of information it has, and if there are any gaps or odd patterns. By doing this first, I can make sure our later analyses are reliable. This preliminary examination was conducted using the necessary pandas methods, as detailed in the “Examine the structure and shape of the datasets” section of my Jupyter notebook.

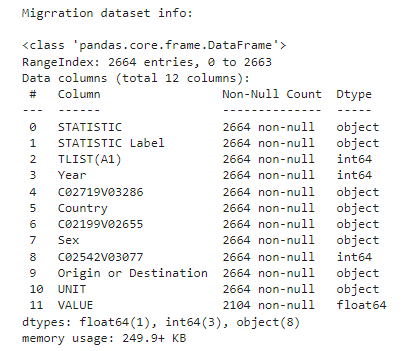
Here is a short summary of the structure of the first dataset:

* There are 11817 rows and 10 columns
* Present data types are: int and object (which later will be converted to categories)
* No null values



The same was applied for the second dataset and here is the information gathered:

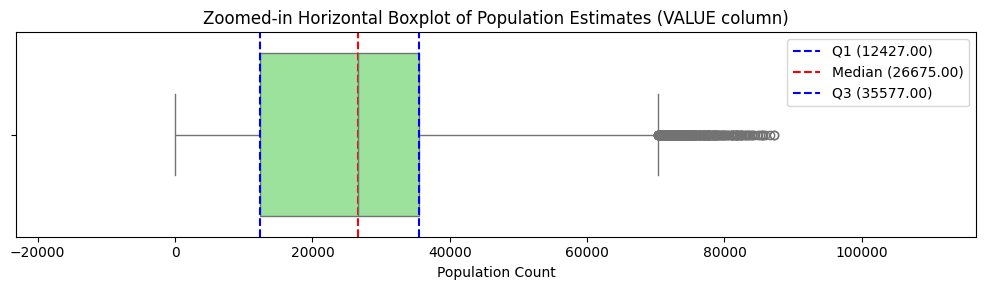
* There are 2664 rows and 12 columns
* Present data types are: int, object and float
* There are 560 missing values for the ‘VALUE’ column



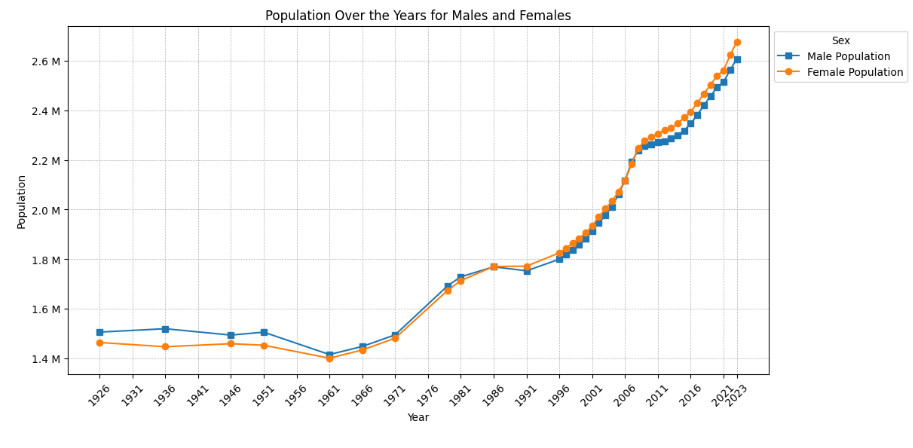
As the third dataset was added only for the optimization of the machine learning model and applying the normal distribution, a brief analysis was done, such as imputation of the null values and preparing the data for the application of the models. This has been performed in the CA1\_2\_final Jupyter notebook in the section ‘Migration dataset with age group data exploring and preparation’.

Checking for missing data is crucial because gaps in data can affect our results. There are two main approaches to address this: deletion and imputation. Taking a further look at the dataframe with the country breakdown, I was able to confirm that those missing values are for different countries and different years, where I do not have any data for the Immigrants, Emigrants and respectfully for the Net Migration. Imputing values where none exist can introduce bias, especially since the missing data spans different countries and years. Having said that, as of now I will be removing those rows. Imputation was performed for the third data set, where I was able to calculate the Net Migration (the missing values), as I had the values for Immigrants and Emigrants. I also checked all three data frames for duplicates to ensure accurate analysis and found none.

Part of my work was checking for unusual data, or outliers, as detailed in the "Dealing with Outliers" section of my Jupyter notebook. I found many in both datasets, but realized they represent totals, like the yearly sum for all ages and both genders in the population data, or the sum for all countries and both sexes in the migration data. So, I'll use two versions of each dataset: one for total values and another for detailed breakdowns. A visual representation of the outliers for the population data set can be seen below:



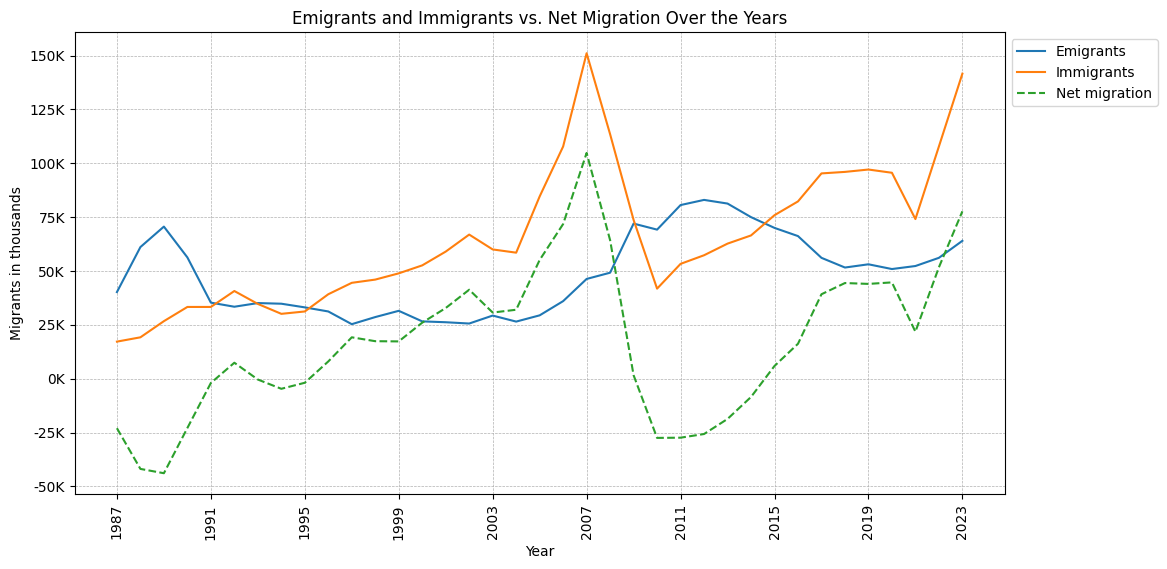
A couple of visualizations were performed for each dataset, representing the whole data available in order to get an idea of the different points of data.



We can see a steady growth in both male and female populations in Ireland, with the female count slightly surpassing the male in recent years, highlighting evolving demographic patterns over the decades. A simple linear regression analysis was applied and it shows Ireland's population is growing, a trend that will be examined more closely with machine learning later in this report. Early results match up with the CSO's population projections (CSO, 2023).

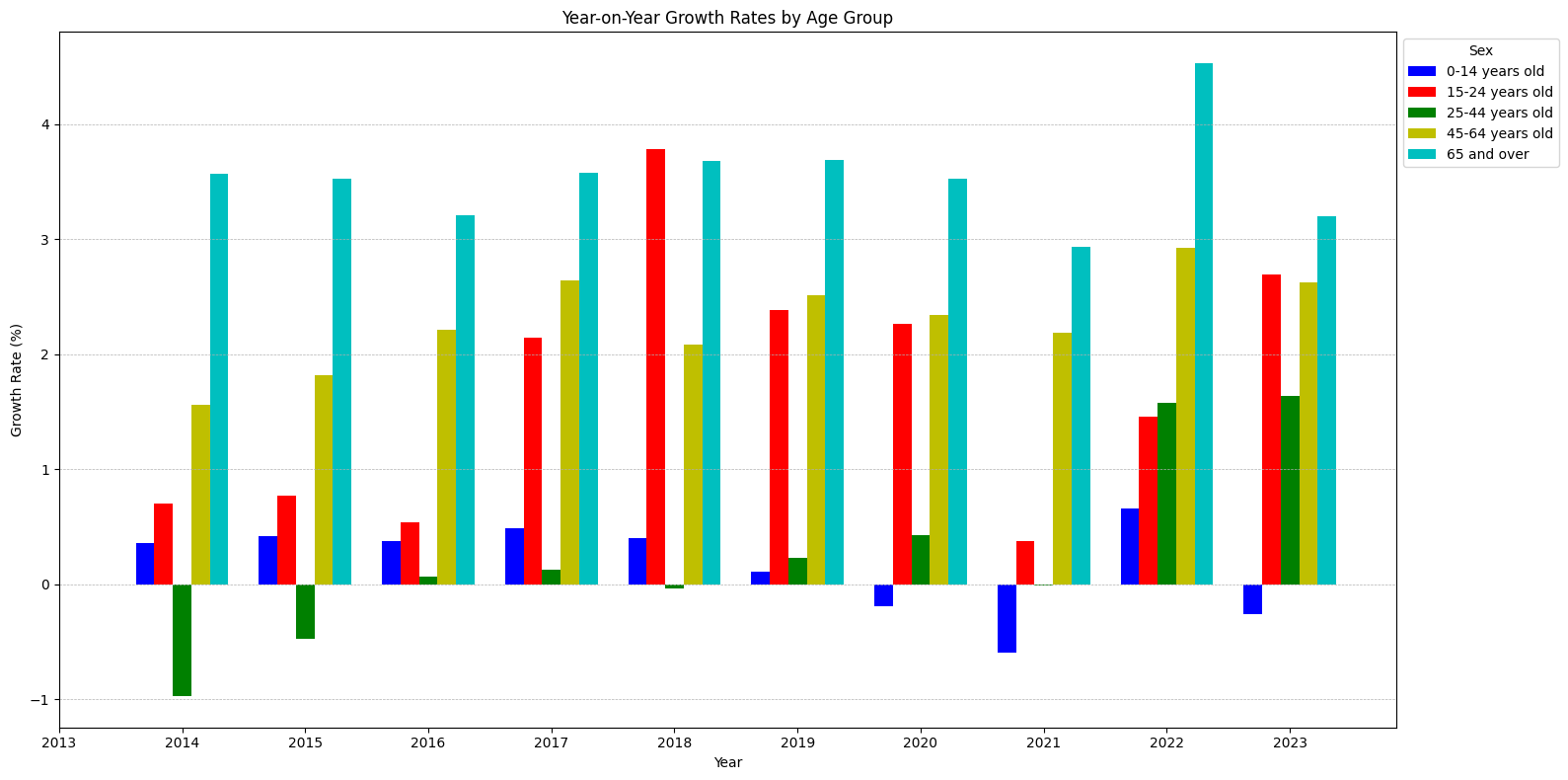
Tufte states, "The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data" (Tufte, 2007, p.77). In line with this, the visualization focuses on 'Sex' and the total population value, ensuring I don't overcomplicate the display. In alignment with Tufte’s principles, using a line chart for time-series data allows for clear visualization of trends over the years. I used blue for males and orange for females to clearly differentiate the two, making it easier for viewers to compare trends without relying on colours that might have cultural biases. I also decided to format the y-axis values with ‘M’ for millions, simplifying the numbers and making the chart cleaner.

A similar visualization for the migration dataset was done, including all present years within the dataset.



The chart reveals a 15-year high for immigration in 2023, marking the most significant inflow since the 2007 peak.

I adjusted the population data to focus on specific age groups and it's change during the years. I only looked at the past 10 years because recent changes, like tech advancements and job market shifts, make this period especially relevant. For better accuracy, I turned the 'Age' column into an integer, so that later on I can apply the descriptive statistics to the numerical values. This meant changing entries like 'Under 1 year' to '0', using extraction of the integer value using regex. With these changes, I added a new column to sort these ages into groups, making it easier to see the age distribution. This gave me a lot better outcome when it comes to the visualization of the data, as we only have 5 age groups now, instead of having each single year of age, which was going to be a very heavy representation.



The youngest age group (0-14 years) experiences a relatively stable but modest growth, indicating a steady birth rate. In contrast, the 15-24 and 25-44 age brackets show more volatility, reflecting the impact of migration trends and economic factors on these more mobile demographics. Significantly, the 45-64 and 65 and over age groups consistently exhibit higher growth rates. This trend suggests an aging population, with a possible increase in life expectancy and a growing number of individuals entering the retirement bracket.

For finding the standard deviation and mean of the Age I focused on the year 2022. Based on the calculations, where I had to use the weighted mean due to the structure of the data, I was able to calculate the average age in Ireland for 2022 being 38.37, which aligns closely with the report from the CSO (Central Statistics Office, 2023).

For the analysis, a Binomial distribution was employed to model the probability of different outcomes in trials that can result in either success or failure. The Binomial distribution is defined by two parameters: the number of trials *n* and the probability of success *p* in a single trial. It provides a way to calculate the probability of obtaining a certain number of successes in a fixed number of trials. (National Institute of Standards and Technology, no date). Those calculations can be found under the Binomial distribution module in the CA1\_2\_final Jupyter notebook.

In assessing the continuous data related to Ireland’s population and migration, I will evaluate whether it conforms to a normal distribution. A bell-shaped density curve is typical for the normal distribution. It is characterized by its mean and standard deviation. In such a distribution, the mean is not influenced by the data's outliers. When the data adheres to a normal distribution, approximately 68.2% of observations fall within one standard deviation of the mean, 95.4% within two standard deviations, and 99.7% within three standard deviations (Campbell, 2007).

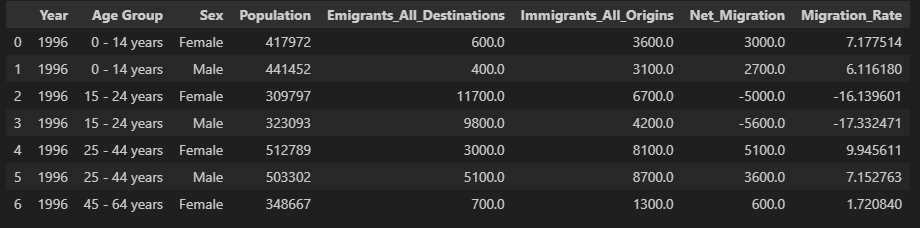
A test for normal distribution was performed on this data for different age groups, using the Shapiro-Wilk test, giving us a p-value of 0.328 for the age group 0-14, which does not provide us with enough evidence to reject the null hypothesis that the data is normally distributed

So, I will be assuming that this data is following the normal distribution and calculations will be performed. The whole analysis can be found in CA1\_2\_final, under Normal distribution.

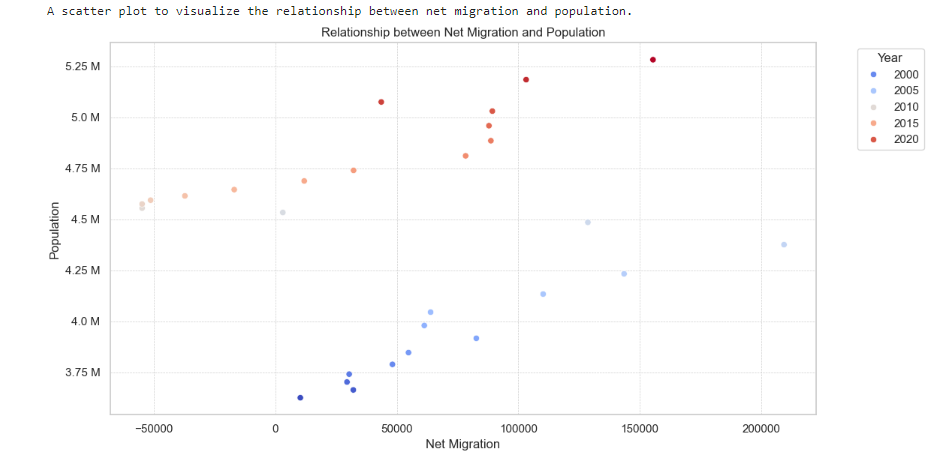
**Data Preparation**

Data will be segmented for detailed analysis. Several machine learning algorithms will be applied to model Ireland's population growth. The outcomes of these models will then be compared to assess their predictive accuracy. With the initial two datasets I had, I did not have the age breakdown in the migration data, so I only had two features – Year and Net Migration, and the population as the target variable. The data was split so that the first 80% is used for training the model and the last 20% for testing it. This way, the model learns from earlier years and is tested on the most recent ones, following a standard approach for predicting over time (Hyndman & Athanasopoulos, 2018).

In order to get more data for the machine learning algorithms and improve the accuracy of the models, a third dataset was introduced containing the Age Group breakdown, so that we can later merge the data with the population dataset that also has the Single Year of Age column. To do the merge, the same age groups that were created earlier were used, with some small changes in the syntax, as I noticed that they did not match exactly with the migration data. Once the converting was done, a right join was performed for both tables on the Year, Age Group and Sex category, giving us the population, net migration, immigration and emigration in one single table. The representation of the numbers was also adjusted, so that we can have the same unit for both values. Final outcome of the merging of the datasets:



A visual representation of the correlation between the migration and population has been created.



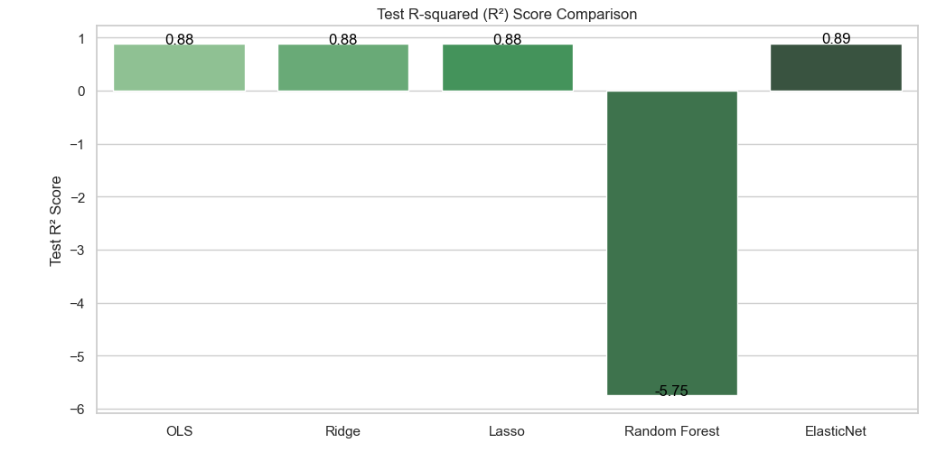
The scatter plot provided shows a clear visual correlation between net migration and population figures in Ireland from the year 2000 to 2020. Each point on the graph represents the interplay between the population of Ireland and the net migration for that year, with the colour coding corresponding to distinct years across two decades.

Observing the scatter plot, it is apparent that there is a general trend where an increase in net migration correlates with an increase in population. This trend is consistent with the basic demographic principle that migration contributes to population changes, alongside other factors such as birth and death rates.

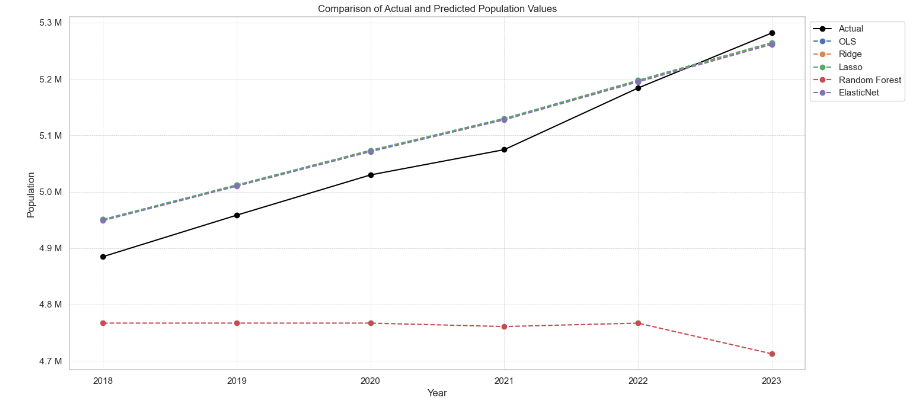
**Modelling**

For continuous data, regression analysis has been employed due to its suitability for predicting numerical outcomes. Regression is ideal for understanding and quantifying the relationship between a continuous dependent variable, such as Ireland's population size, and one or more independent variables. This phase involves selecting the appropriate regression techniques, adjusting model parameters, and assessing model fit to ensure accurate predictions and valuable insights into the factors influencing population growth.

The training set will be utilized to train a suite of models, including Linear Regression, Ridge, Lasso, ElasticNet, and Random Forest, with the latter four undergoing hyperparameter tuning for optimization. Post-training, the models will be evaluated on the test set to compare predictive accuracies, using metrics like MAE, RMSE, and R².



The bar chart indicates that ElasticNet Regression achieved the highest R-squared (R²) score of 0.89 among the tested models, suggesting it was the most effective at capturing the variance in Ireland's population data. Conversely, the Random Forest model significantly underperformed with a negative R² score of -5.75, indicating that it did not predict the test data accurately and was worse than a model that would simply predict the mean of the target variable. The other models — OLS, Ridge, and Lasso — show comparable performance, each with an R² score of approximately 0.88, reflecting a good fit to the test data. The actual results of the models, with the population data itself, can be visualised on a time series plot.



The chart effectively uses Tufte's principles by clearly contrasting actual population data with predictions from multiple regression models, using distinct markers and a time series layout for ease of comparison. The design is simple and focused, avoiding unnecessary embellishments to facilitate direct, unambiguous interpretation of the data trends.

Similar analysis was performed on the merged dataset and will be discussed below. To prepare the dataset for analysis, the 'Age Group' and 'Sex' columns were one-hot encoded to convert categorical data into a format that could be provided to machine learning algorithms. This transformation was carried out using the pd.get\_dummies() function from Pandas, producing an updated DataFrame named population\_migration\_df\_encoded with the categorical attributes encoded. We can see an improvement of around 0.04 in the Linear Regression, Ridge, Lasso, ElasticNet, meaning that in fact adding more data, the models are performing better. However, there is a very big improvement when it comes to the Random Forest, which from an R² score of -5.75 goes up to 0.99, which tells us that it is not performing good and it’s overfitting.

OLS (Ordinary Least Squares) Regression:

R²: 0.9301

Lasso Regression (with hyperparameter tuning):

R²: 0.9301

Ridge Regression (with hyperparameter tuning):

R²: 0.9288

ElasticNet Regression (with hyperparameter tuning):

R²: 0.9286

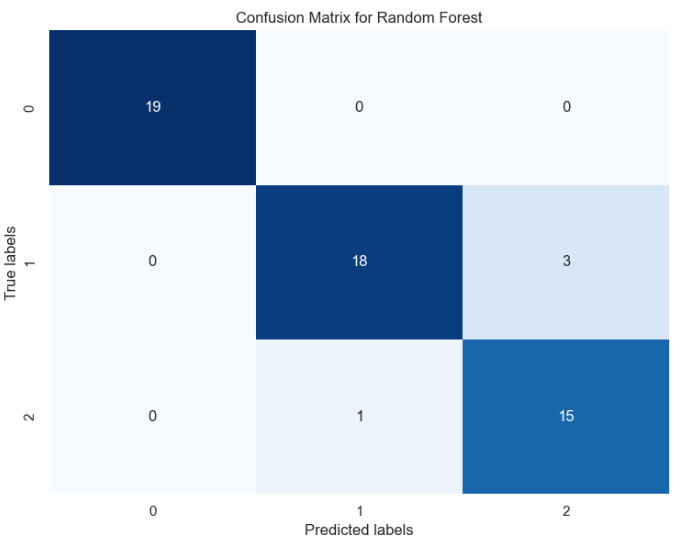
Random Forest Regression (with hyperparameter tuning):

R²: 0.9909



The scatter plot shows the relationship between actual and predicted population figures using Ordinary Least Squares (OLS) Regression. The close alignment of the data points along the red line of perfect fit suggests the model's predictions are highly accurate. Similar visualizations for all other models were created and can be found in the Jupyter notebook.

For the classification analysis, I will focus on the net migration and will create three labels referring to low, medium and high migration. I will keep the splitting of the data in the same ratio, being 80% for the testing and 20% for the training. I applied four different algorithms – Logistic Regression, Support Vector Machine, Random Forest and Gradient Boosting. The performance of all of them is quite good, however the Random Forest is the one performing the best with an accuracy of 0.92%. To understand how many of those labels were accurately predicted, I created a confusion matrix for each one of them which can be found in the Jupyter notebook. Let’s take a closer look at the confusion matrix for the Random Forest and analyze the results.



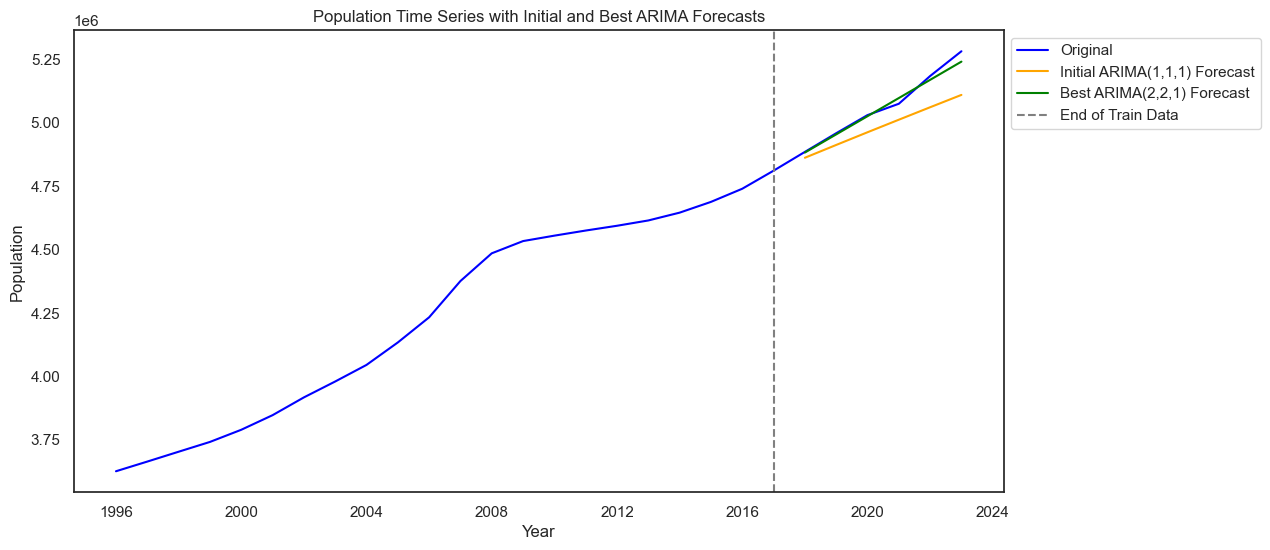
For the Label ‘0’, which in our case is the ‘low’, all 19 instances were predicted correctly with no mistakes. For the Label ‘1’, being the ‘medium’, the model predicted correctly 18 instances, but there are three errors as well – for three instances the model classified as ‘1’, while they should have been labelled as ‘2’. For the Label ‘2’ being the ‘high’, 15 instances were predicted correctly, with only one instance labelled as ‘2’, while it was actually ‘1’. Overall, the model is performing good, and also for one of the instances no errors were made. Getting more data, should improve the model, because as of now we do not have enough instances that can be tested – the model is performing on less than 20 instances per each Label.

Arima also known as autoregressive integrated moving average have been used in my analysis for statistical measure of the population data for the time period I’m working on (1996 – 2023). As we can get familiar from the **mastersindatascience.org** the ARIMA model is used to understand past data or/and predict future population data distribution. I assume that the model could be a good fit for my analysis because the data is recorded in regular (years) interval and it’s not a any nonseasonal series of numbers that exhibits patterns and is not a series of random events. An ARIMA model has three component functions:

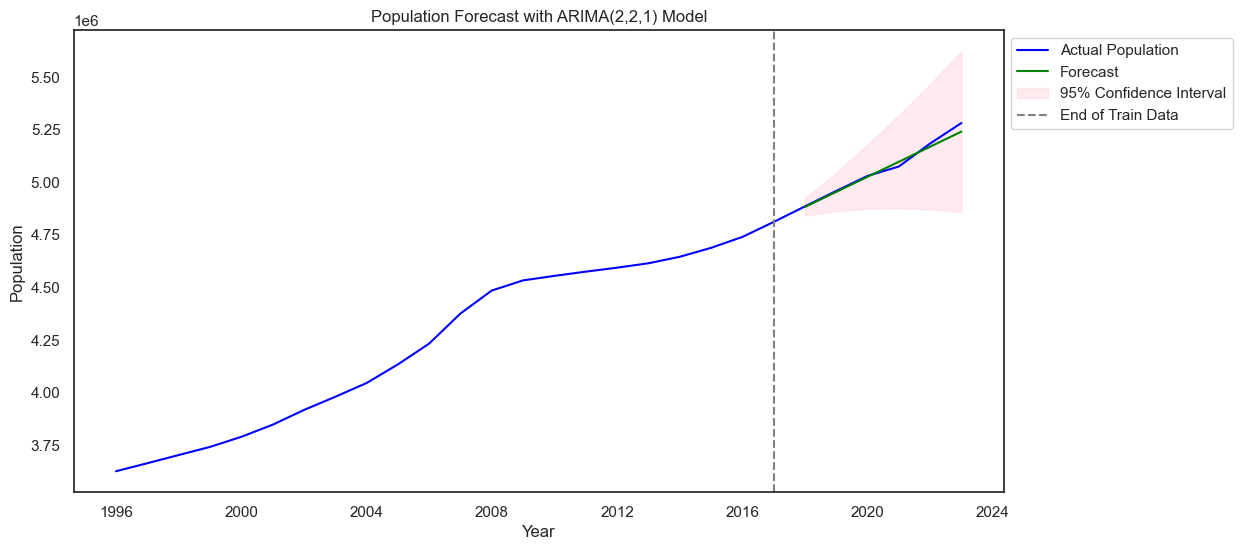
* AR (**p**), the number of lag observations or autoregressive terms in the model;
* I (**d**), the difference in the nonseasonal observations.
* MA (**q**), the size of the moving average window.

An ARIMA model order is depicted as (p, d, q) with values for the order or number of times the function occurs in running the model. Values of zero are acceptable.

After running a simplified grid search to tune the parameters of the ARIMA model, the best-performing model was ARIMA(2,2,1). ARIMA(2,2,1) model achieved a significantly lower Mean Squared Error with value of 409046390.49283 and Root Mean Squared Error with value of 20224.895314755773 indicating a better fit, compared to the initial ARIMA(1,1,1) model where MSE is 9326242855.819143 and RMSE is 96572.47462822491. These results suggest that the time series data likely has a second-order differencing pattern and a first-order moving average process, which the ARIMA(2,2,1) is able to capture. In terms of forecasting capability, ARIMA(2,2,1) model would be expected to perform better than the initial ARIMA(1,1,1).



* The blue line represents the actual population data over time.
* The orange line shows the forecast from the initial ARIMA(1,1,1) model.
* The green line indicates the forecast from the best ARIMA(2,2,1) model after parameter tuning.
* The grey dashed line marks the end of the training data and the beginning of the testing period where the models' forecasts are compared against the actual data.

****

* The pink shaded area represents the 95% confidence interval for the forecast, providing a range where future points are expected to fall with 95% certainty.

**Evaluation**

References:

Redman, T.C., 2018. If Your Data Is Bad, Your Machine Learning Tools Are Useless. Harvard Business Review. [Online] Available at: <https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>

Pandas documentation, n.d. Available at: https://pandas.pydata.org/docs/reference/api/pandas.pivot\_table.html

Tufte, E. (2007) The Visual Display of Quantitative Information (2nd ed.). Cheshire, Connecticut: Graphics Press.

Campbell, M. (2007) Medical statistics: A textbook for the Health Sciences. Wiley.

# CSO, n.d. Press Statement Census of Population 2022 - Summary Results.

SV-Europe. (no date) CRISP-DM Methodology. Available at: <https://www.sv-europe.com/crisp-dm-methodology> (Accessed: 5 November 2023).

CSO. (2023) Press Statement Census of Population 2022 - Summary Results. Available at: https://www.cso.ie/en/csolatestnews/pressreleases/2023pressreleases/pressstatementcensusofpop

Koleva, M. (2023) CA1\_1\_final [Jupyter notebook]. Available at: <https://github.com/mkoleva0/DataAnalytics_CA1/blob/main/CA1_1_final.ipynb> (Accessed: 10 November 2023). ulation2022-summaryresults/ (Accessed: 20 October 2023).

National Institute of Standards and Technology. (no date) Binomial Distribution. Available at: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda366i.htm> (Accessed: 10 November 2023).

Central Statistics Office. (2023) Population and Labour Force Projections 2017-2051. Available at: <https://www.cso.ie/en/releasesandpublications/ep/p-plfp/populationandlabourforceprojections2017-2051/populationprojectionsresults/> (Accessed: 10 November 2023).

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). Monash University. Available at: <https://otexts.com/fpp2/accuracy.html> (Accessed: 15 October 2023).

SV-Europe. (2023). What is the CRISP-DM methodology?. [online] Available at: https://www.sv-europe.com/crisp-dm-methodology/ (Accessed: 1 November 2023)

McKinney, W. (2022). 5. Getting Started with Pandas. In: Python for Data Analysis. [online] Available at: https://wesmckinney.com/book/pandas-basics (Accessed: 10 November 2023).