MSc in Data Analytics

CA1

Author: Maria Koleva

E-mail: sba23020@student.cct.ie

Student ID: 23020

**Abstract**

*The study probes demographic trends in Ireland, employing the Linear Regression algorithm to forecast population growth. Data analysis confirms that the overall population has shown a steady increase over the past few decades, a trend anticipated to persist in the foreseeable future. A distinct correlation emerges between net migration and population shifts during this period. Periods of heightened net migration significantly influenced these growth rates. These results not only emphasize the crucial role of migration in Ireland's demographic evolution but also underscore the utility of machine learning tools in predictive population studies. In an extension of my analysis, I explore the predictive power of machine learning by incorporating the K-Nearest Neighbors (KNN) and decision tree algorithms to classify annual migration as either positive (inflow) or negative (outflow).*

**Introduction**

Ireland's demographics have long been a subject of study and interest, revealing patterns and trends that reflect both historical events and evolving societal dynamics. The interplay between natural population growth and the effect of net migration offers a complex narrative, one that is crucial for understanding the nation's past and anticipating its future. Based on features like year and country, this research seeks to classify if the migration will be positive (inflow) or negative (outflow) for a given year using KNN and decision tree algorithms. Delving deeper into Ireland's population changes, the study places a spotlight on the influence of net migration. It aims to uncover the degree to which immigration and emigration have shaped Ireland's population over the decades. With the large amount of data available, ensuring its quality and relevance becomes paramount. As such, thorough data preparation techniques are employed, paving the way for accurate and insightful analysis. Incorporating modern analytical tools, the study harnesses the power of the Linear Regression Machine Learning algorithm. This approach aims to provide a predictive framework, offering projections of Ireland's population trajectory based on historical data. Through a combination of data visualization, meticulous preparation, and advanced analytical techniques, the research sets out to offer a comprehensive perspective on Ireland's demographic journey.

**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.

Firstly, I aimed to figure out migration trends and estimate Ireland's future population. I then got familiar with our data, checked its quality, and made sure it was ready for analysis.

For the actual predictions, I used supervised learning, a method that learns from past data to make future predictions. I picked Linear Regression to guess Ireland's population since it's good at seeing growth patterns. To understand migration trends, I used K-Nearest Neighbors (KNN) and Decision Trees. KNN looks at data groupings, while Decision Trees make decisions based on the data's features.

**Data Preparation and Visualization**

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 two datasets from the Central Statistics Office (CSO) to examine Ireland’s population and migration trends.

The two datasets used in this project are as follow:

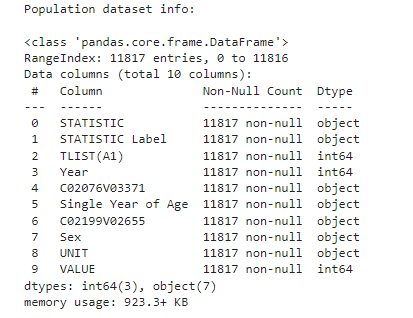
1. PEA11.20231013T111046.csv - Population estimates from 1926
2. PEA18.20231016T141007.csv - Estimated Migration (Persons in April)

The first dataset tracks the shifts in Ireland's population over the years and includes details about the age distribution. The second dataset, on the other hand, tells us about migration patterns, detailing with both incoming and outgoing movements. Notably, this dataset also breaks down the countries of origin, 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 make sense and are reliable. This preliminary examination was conducted using specific pandas methods, as detailed in the “2. Examine the structure and shape of the datasets” section of my Jupyter notebook.

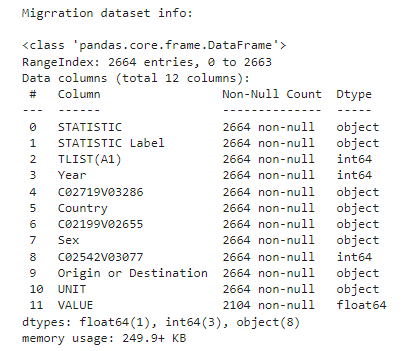
Here are some of the most relevant information gathered for the first dataset:

* There are 11817 rows and 10 columns
* Present data types are: int and object (which later on 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

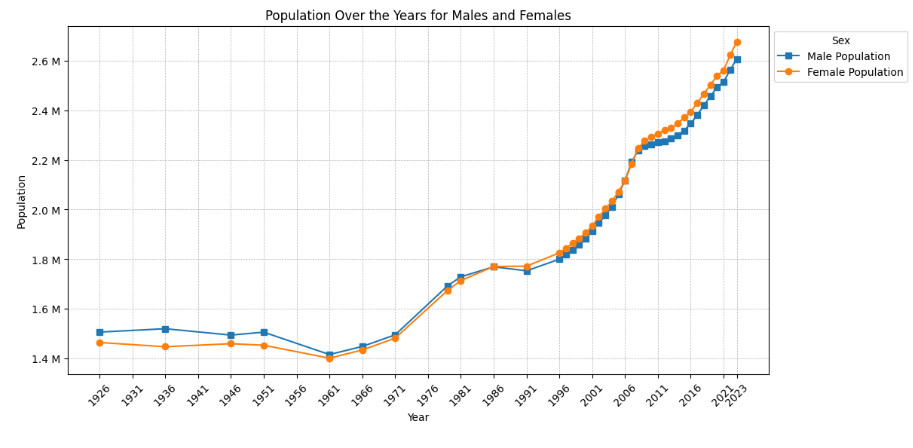


Checking for missing data is crucial because gaps in data can affect our results. Incomplete data can greatly influence the outcomes of data analysis. There are two main approaches to address this: deletion and imputation. Taking a further look at the dataframe, I was able to confirm that those missing 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. I also checked both dataframes 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.

Patterns and Trends

As later on we will be applying Linear Regression to predict Ireland’s population, I took a further look on how the population has been changing all over the years, using a line chart to represent the trend.



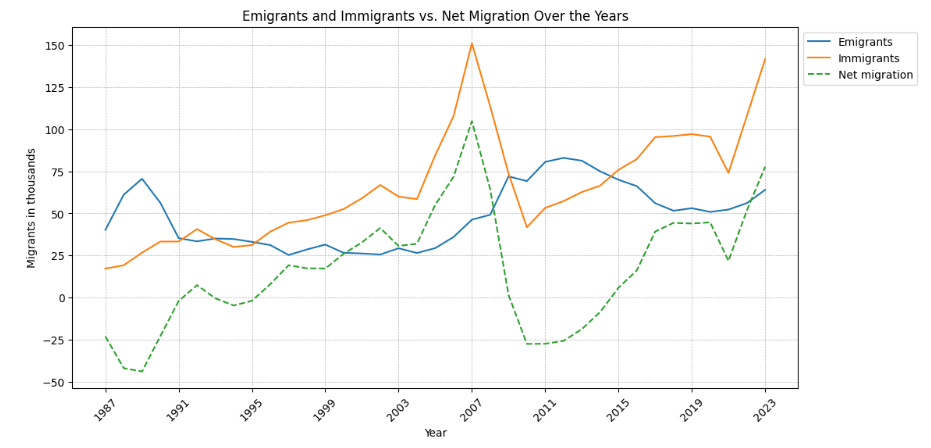
We can see a steady ascent in both male and female populations in Ireland, with the female count slightly surpassing the male in recent years, highlighting evolving demographic patterns over the decades.

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 solely 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 colors 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.

To have I adjusted the population data to focus on specific age groups. 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 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.

#TODO – format the value on the left.

I did a similar visualization for the migration dataset, 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.

Reference:

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. pivot\_table. 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.