# Analysis of Immigration and Emigration in Ireland for the Period \_\_\_\_ - 2023.

Author: Emma McCready

Student ID: sba23001

MSc in Data Analytics

Postal address: Meenagowan, Lettermacaward, Co. Donegal

E-mail: [sba23001@student.cct.ie](mailto:sba23001@student.cct.ie)

Github link:

# **Abstract**

Every submission should begin with an abstract of about 100 words in the normal text style but italicized. The abstract should be a concise statement of the problem, approach, findings, and conclusions of the work described.

## 2. Introduction

## Sections

Breaking your report into sections can make it much easier to read. Main sections (Introduction, Materials and Methods, Results and Discussion/Conclusions should generally be in Arial 12-point bold title-case or follow specific instructions given with your real assignments.The use of screen shots or diagrams to enhance your explanations is encouraged.

### Subsections

Arial 12-point title-case.

#### Sub-subsection headings

Arial 12-point italic.

### 2.1 Central Statistics Office

The purpose of the Central Statistics Office (CSO) is to collect, analyse and make available statistics about Ireland’s people and society, across w range of areas including construction, health, welfare, the environment and economy. maintain a record of population stocks and migration flows on an annual basis.

Could possibly expand a bit more on the CSO? [Who We Are - CSO - Central Statistics Office](https://www.cso.ie/en/aboutus/whoweare/)

### 2.2 Migration to and from Ireland

With a particular interest arising from the volume of young people emigrating, this analysis focuses on the migration to and from Ireland. The population migration flows are collected from mid-April over a 12-month period (REF CSO1). The CSO have a

Ref news articles citing increased

### 2.3 Third heading

## 3. Methodology

### 3.1 Dataset

Data was obtained from the CSO (Ref website?) and a dataset containing information on the migration count to and from Ireland for the published for 1987 to 2023 was selected for this analysis. In its raw form, the dataset contains 7 columns (Fig X). The migration data is collected from mid-April to mid-April of the following year, and the year listed in the dataset refers to the latter year when the data is published. For example, the 2023 rows refer to data collected from mid-April 2022 to mid-April 2023. in the data refers to the April of the year the migration. This dataset contains information on the total migration split by sex for each country of origin or destination country.

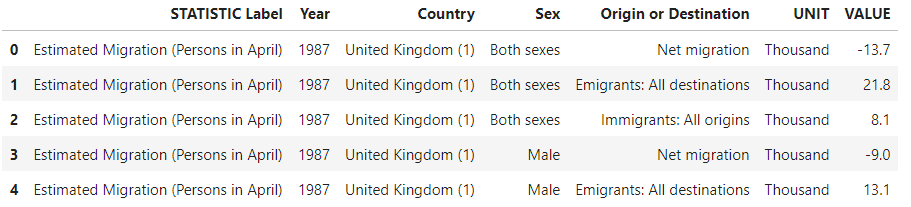


Fig X. The first 5 rows of the dataset.

A table with text on it

Description automatically generated

Fig Y. The first 5 rows of the dataset post-data prep

3.2 Exploratory Data Analysis

A standard method for Exploratory Data Analysis was followed. The data was loaded onto the variable pop\_data using the function .read\_csv() from the Pandas library. The first five rows and the last five rows were assessed and compared using pop\_data.head() and pop\_data.tail() respectively. The function .info() was used to gain insight into the number of rows, as well as the number of missing values in each row. The total number of missing values was obtained by calling pop\_data.isnull().sum(). To look at the statistical information of the dataset, .describe() was used.

Boolean masks were used to filter the data. Specific areas of the dataset were explored to investigate areas which may not directly form part of the overall analysis. In this case, line plots were used to look at migration over time, boxplots were used to assess the distribution of migration values and identify outliers. Histograms and bar charts were also used to visualise the data. The Seaborn library was used for all plotting.

3.3 Data Preparation

The columns that were deemed redundant were dropped ('STATISTIC Label', 'UNIT'). Column headings with .rename(columns = {‘old\_heading’:’new\_heading’,…}). Similarly, the values under the ‘Country’ and ‘Origin or Destination’ columns were shortened and/or acronymised using .replace(). The missing values were explored to garner additional information into the type of missing data and where the values were missing by moving them into a separate dataset and exploring it.

Univariate Analysis:

Not interested in univariate analysis, except for histogram of total\_migration to communicate visually information about minimum and maximum values, central location, and spread, as well as skew. Not interested in looking closer at the other variables, as they have the same number of values in it, so there’s no useful information to be extracted.

Bivariate anlaysis

4. Results and Discussion

4.1 Dataset

4.1.1 Exploratory Data Analysis

EDA steps and why I did each.. try to ref some papers

The EDA process was cyclical and was adjusted throughout the analysis as more information was discovered about the dataset.

4.1.2 Data Preparation

For ease of exploration and to aid in analysis, some data tidying and preparation was performed in Section 2 of the codebook. Firstly, redundant columns that provided no useful information for the analysis were dropped (i.e., 'STATISTIC Label', 'UNIT'). The UNIT column specified that the units for the VALUE column were in thousands (See Fig. X). Thus, it was considered to multiply the VALUE column by 1,000 to reflect this. However, later, when visualising the data, it was realised this was additional data ink which goes against Tufte’s 4th principle of graphical integrity (See Section 4.1.3 of this report on Data Visualisation).

Furthermore, the column headings were renamed to make them easier to call with functions, and to better reflect the values they contained. For example, the column title ‘Origin or Destination’ implies that the data it contains is binary (i.e. either ‘Origin’ or ‘Destination’), but it actually lists the type of migration that the data row refers to (i.e. whether it was emigration, immigration or net migration). Therefore, the column was renamed to “migration\_type”.

The countries included in this dataset were quite long and contained additional text in some strings which wasn’t deemed necessary to this analysis, and so they were shortened. In some cases it was deemed appropriate to acronymise them with widely accepted terms (i.e., for the UK and USA). However, it was not considered appropriate to shorten all countries, as in some cases it may not have been clear to the reader (e.g. “Aus” for Australia may refer be misinterpreted as Austria).

The dataset spanned from the years 1987 to 2023. Exploring the data revealed that there were multiple missing values, particularly for before 2008. As the dataset is of a moderate size with only a few features, it was possible to conclude through exploring the dataset that the values were Missing Not at Random (MNAR). Prior to 2008, migration data wasn’t categorised into the list of countries it is now, namely for migration to and from Australia, Canada, and EU 15 to EU 27. Prior to this, the migration data for these countries were classified under “Other countries”. Thus, the category for ‘Other countries’ pre- and post- 2008 are not comparable, and so it was deemed appropriate to split the dataset here, and to focus all analysis on migration data from 2008 onwards.

A function was defined to filter data, which when called, used Boolean masks to create a subset of data based on the desired country, sex and migration type. This helped improve the efficiency of the code. The function was validated by creating the same subset of data manually, and after this was used in the creation of all subsets of data.

4.1.3 Data Visualisation

When visualising this data, Tufte’s Principles were followed,,,

While Tufte frowns on horizontal reference lines, a horizontal line was added in graphs for net migration at y=0 to highlight when there was a negative migration total. However, this line was made slimmer so it wouldn’t be a distraction or use too much data ink.

Made sure to use 3:2 aspect ratio by default

In choosing a colour scheme for visualisations, the documentation for the Seaborn package for assistance in choosing the colours. The default colour scheme in Seaborn was used as this is in line with what viewers are most accustomed to, and it’s optimised for representing categorical data. Moreover, the default ordering of the colours are distinct, which will aid in the viewer’s interpretability.

Avoided use of gendered colours when representing male vs female. Despite blue being associated with males, it is recognised as a generic colour on graphs. Orange is not usually associated with females.

Histograms are used to visualise distributions bc…

Bar charts used to .. bc..

When plotting multiple graphs on the one output, care was taken to ensure there was adequate spacing and that they were on the same axes for ease of interpretation. If

|  |  |
| --- | --- |
| **Graphical Integrity** | **Graphical Excellence** |
| 1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured. | 1. Above all else show the data |
| 2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data. | 2. Maximise the data-ink ratio |
| 3. Show data variation, not design variation. | 3. Erase non-data ink |
| 4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units. | 4. Erase redundant data-ink |
| 5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data. | 5. Revise and edit |
| 6. Graphics must not quote data out of context. |  |

For plotting the histogram, I initially plotted in matplot lib and noted how changing the bin size impacted the distribution of the feature. Then I decided to use seaborn to plot as it automatically calculates an appropriate bin size based on the statistical distribution of the dataset.

* I didn't include the number of bins, but I wanted to experiment with different numbers to see how the plot changed, particularly to unearth any multimodality. I then did some research to find the best number of bins, and noted the "square root rule" `n\_bins = int(np.sqrt(len(overall\_net\_migration))) `

4.2 Net migration 1987 to 2023

Put in two figures

Ad discussed in Section 4.1.2 of this report, a decision was made to split the dataset into two, and focus the analysis on 2008 to 2023. However, the nean net migration for each country in each dataset was

Discuss separately. It’s interesting that there is more person immigrating to Ireland from most countries, except for the USA, Australia and Canada.

Statistics section

The sample contained 117 which is

Univariate analysis would be migration total for diff countries

Bi variate would be migration total for diff countries split by gender

Noticed 2007 was the year for the highest immigration – could discuss?>

[Article: Ireland: From Rapid Immigration to Recession | migrationpolicy.org](https://www.migrationpolicy.org/article/ireland-rapid-immigration-recession)

[Population And Migration Estimates, April 2007, Foreign Nationals: PPSN Allocations And Employment, 2002-2006 - CSO - Central Statistics Office](https://www.cso.ie/en/csolatestnews/pressreleases/2007pressreleases/populationandmigrationestimatesapril2007foreignnationalsppsnallocationsandemployment2002-2006/)

[WP69\_The\_changing\_face\_of\_Irish\_migration\_2000\_2012\_0.pdf (maynoothuniversity.ie)](https://www.maynoothuniversity.ie/sites/default/files/assets/document/WP69_The_changing_face_of_Irish_migration_2000_2012_0.pdf)

**ML** Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios. Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modeling. **[0 - 20]**

Show the results of two or more ML modeling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. **[0 - 30]**

**Programming** 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]**

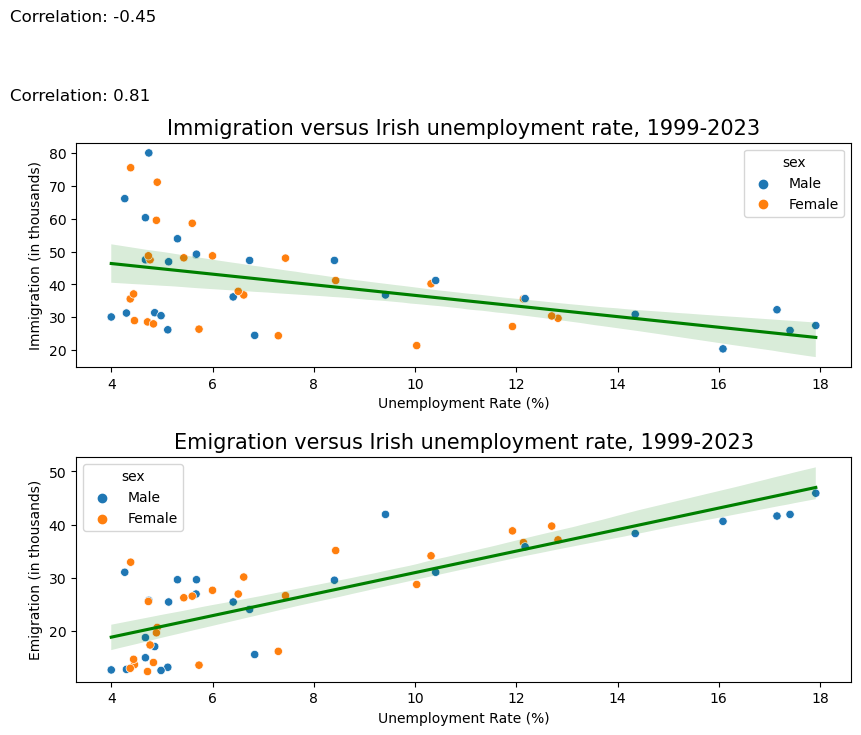
**Machine learning**

A solid project management framework is required for all data science projects to ensure efficiency in terms of time and resources. They are beneficial for providing a structured and systematic approach for data scientists to follow. One such project management framework is that of Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a popular framework consisting of only six phases (business understanding, data understanding, data preparation, modelling, evaluation, and deployment), which are represented as a cycle and so are iterative. Because CRISP-DM is flexible and easily adapted, it can be used for data science projects of all size, but it is renowned for designing and implementing data mining projects. The CRISP-DM framework was implemented in this analysis, as first the project’s objectives and requirements were defined, then EDA was performed, the data was prepared for modelling, models were built and then evaluated, and now with this report the framework is complete. Throughout this analysis, the project moved backwards and forwards in the steps of the CRISP-DM framework.

Supervised machine learning methods were applied in order to predict the migration, as the dataset used is labelled which the machine uses to earn. Further data from the CSO was used to supplement the dataset with the rate of unemployment. The data was filtered to obtain the average annual rate from April of a particular year to March of the following year, in accordance with the period covered by the main dataset. However, data was only available to cover 1999 to 2023, which limited the dataset size used in the machine learning methods.

Initially, the goal was to predict migration based on the Irish unemployment rate, and a basic linear regression analysis was performed to assess the relationship of these two variables (Fig Z). As expected, emigration exhibited a linear relationship with the unemployment rate. As the unemployment rate rose, emigration also increased increased, with an R2 (correlation) value of 0.81. It was expected that immigration would decrease with the unemployment rate, but the linear regression didn’t reflect this, with an R2 value of -0.45 and 0.81 for immigration.

Fig Z. Scatterplots showing the relationship between the two types of migration (emigration and immigration) and unemployment rate. The line of best fit and associated confidence bands are represented in green.



Owing to the disappointing results, it was decided to include other parameters into an actual machine learning model. As the target value is numerical, regression models were used. One-hot encoding was used to add binary values to the ‘sex’ and ‘migration\_type’ columns so that these could be included in the analysis. As the data is skewed due to outliers and extreme values, the data was pre-processed using min-max scaling. The fit and performance of five different models were evaluated (See Table X for the results). It was concluded that the Random Forest Regressor fit the model the strongest, and so the model underwent hyperparameter tuning using GridSearchCV() to find the optimal number of estimators and the maximum depth of trees. This was important, as adding more estimators to the ensemble can improve the model's predictive performance, but can be computationally-demanding. The maximum depth of trees refers to how long each decision tree grow, but this needs to be controlled to prevent overfitting. For this analysis, a maximum depth of 100 and 100 random estimators were found to be optimal.

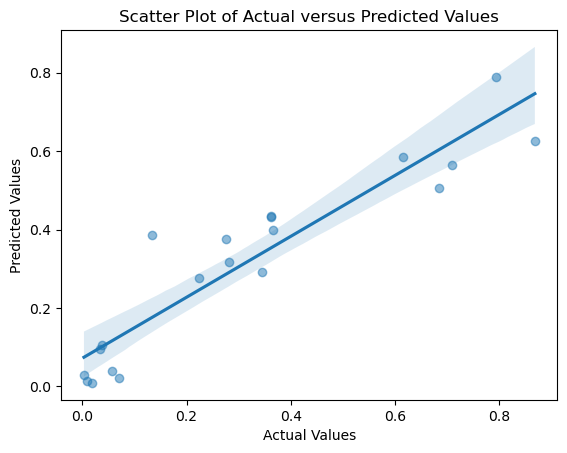
Table X. Evaluation of the fit of machine learning methods applied to the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Machine Learning Model** | **R2 (training set)** | **R2 (test set)** | **Root Mean Squared Error** | **Mean Absolute Percentage Error** |
| Linear regression | 0.3921044 | 0.4891022 | 0.1977052 | 360.7682124 |
| Ridge Regression | 0.3891232 | 0.4743532 | 0.2005386 | 403.0211287 |
| Lasso Regression | 0.3408335 | 0.3993978 | 0.2143604 | 560.2011447 |
| Decision Tree Regressor | 1.0000000 | 0.7692657 | 0.1328639 | 85.1876140 |
| Random Forest Regressor (pre-hyperparameter tuning) | 0.9723276 | 0.8884682 | 0.0923741 | 92.4583043 |
| Random Forest Regressor (post-hyperparameter tuning) | 0.9605161 | 0.8600728 | 0.1034671 | 91.9288365 |

To evaluate fit of the models to the both the training and the test sets, the R2 was calculated, which is measured on a scale of 0 to 1, where 1 indicates a perfect match. The linear regression models (including Ridge and Lasso) tended to perform poorly, whereas the decision tree and random forest regressors fit very well. The Root Mean Squared Error (RMSE) for the test set was used to measure the average prediction error of the model, and the lower the RMSE value the better. The models all performed similarly, with the random forest regressor showing the most potential. Lastly, the Mean Absolute Percentage Error (MAPE) was used to assess the accuracy of predictions. The lower the MAPE the better, with decision tree and random forest regressors again proving to be better ML methods for this analysis. The random forest regressor was identified as the optimal model on account of the higher R2 value for the test set and the lower RMSE.

A scatter plot of the actual versus predicted values was generated to further assess the performance of the chosen ML model which can be seen in Figure A. While the model appears to be performing reasonably well, there are a number of outliers outside of the confidence intervals of the line of best fit, which indicates variability in the data. The ML analysis in this study was limited by a small dataset and the analysts inexperience. It’s unknown if the “year” feature should’ve been removed from this ML model.

Fig A. A visualisation of the performance of the random forest regressor model.



## References

CSO1 [Population and Migration Estimates - CSO - Central Statistics Office](https://www.cso.ie/en/methods/surveybackgroundnotes/populationandmigrationestimates/) [Accessed 30-10-23]