# 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:

In general, use a 12-point Times New Roman font, or other Roman font with serifs.

The title (Arial 14-point bold) runs across the full width of the page and is centred. Authors' names (Arial 12 point not-bold) and affiliations (Arial 12-point not-bold) are entered into the table at the top. We also recommend you add your postal address and e-mail address using the same style as for authors.

Acronyms

|  |  |
| --- | --- |
| CSO | Central Statistics Office |
| EDA | Exploratory Data Analysis |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# **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 period 1987 to 2023 was selected for this analysis. This dataset contains information on the total migration split by sex for each country of origin or destination country.

3.2 Exploratory Data Analysis

A standard method for Exploratory Data Analysis was followed.

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

Machine learning

If I use linear regression, ref that Stanford book

4. Results and Discussion

EDA steps and why I did each..

Including prep steps(e.g. why I acronymised some and not all, UK and US are accepted acronyms but not Aus or Australia as could be Austria)

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)

Data Visualisation

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

* Need to explain EVERYTHING including details like what colour I used.

While Tufte frowns on horizontal reference lines, I felt it important to include one in certain graphs at y=0 to highlight when there was a negative migration total. So I altered the width of the line tomake it slimmer

Made sure to use 3:2 aspect ratio by default

I used the default colour scheme as this is in line with what viewers are most accustomed to (?). I referred to the documentation for the Seaborn package for assistance in choosing the colours, and decided to stick to the default as it’s optimised for representing categorical data, and the default ordering of them are distinct, which will aid in interpretability.

Histograms are used to visualise distributions bc…

Bar charts used to .. bc..

|  |  |
| --- | --- |
| **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))) `

**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]**

## References

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