# MSc in Data Analytics - CA1

Author: Stephen Burke

e-mail: sba23027[@student.cct.ie](about:blank)

Student ID: 23027

Github: https://github.com/step-hen-burke/ca1

## Abstract

*It is a useful and common practice to put the abstract in Times New Roman 12-point italics. Throughout this document the styles used reflect the styles we suggest you use in your scientific report*.

## **Introduction**

Choice of programming style

The principals outlined in Clean Code (Martin, 20XX) were largely followed; names are descriptive, repetition is minimised within reason,

Choice of Machine Learning Framework

Choice of packages – pandas, sklearn

The seaborn plotting library was used extensively to produce the graphics in this report and accompanying notebook. Its style was set to “darkgrid”, which is similar to default themes used in graphics libraries such as R’s ggplot2, and conforms to style directives outlined by Tufte (p. 112, 116); namely that the grid background should be a muted grey so as to not be distracting, and that gridlines should be use minimal ink. In this case, gridlines are given as negative space, so do not use ink at all. Tufte also argues in favour of using a font family with serifs in order to create more “friendly” graphics, which was also set globally.

Additionally, the default linewidth was set to slightly thinner than seaborn’s default. Again, Tufte argues that thinner lines, and therefore less data ink, result in more aesthetically pleasing graphics (p. 184-185).

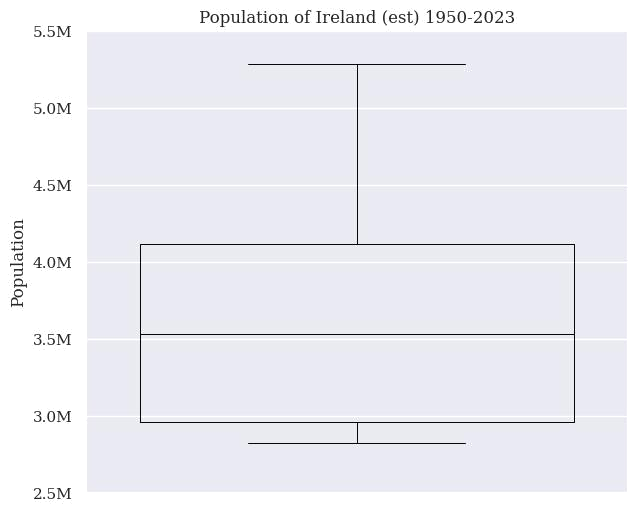
## **Population Data Exploration**

The first dataset explored concerned population estimates in April from 1950 to 2023. The dataframe was loaded from csv and the column names were changed to snake\_case by converting to lowercase and replacing spaces with underscores, this is done as a stylistic choice so that the column names are in a consistent case with our variable names. The same column transformation was applied to multiple csvs throughout the notebook, so a function was created to encapsulate this operation in order to minimise code-reuse.

The data was summarised and missing values were identified - however, only in the 0-4 years age groups, not every year, and only in years until 1995. The data was pivoted in order to make working with multiple factor levels easier, and an attempt was made to fill in the missing data using a combination of other columns; “Under 1 year” was a value within age\_group, so adding this to the “1 - 4 years” value to get “0 - 4 years” was attempted. However, when this method was checked against the completed entries post-1995, it was discovered that this doesn’t work, as the candidate values did not equal the reported values in 47/108 cases.

Further investigation revealed that in certain cases, the discrepancies between the additions of age groups and their supposed combined groups were abnormally large. Census data was imported and compared against the population estimates, and it was determined that the 1-4 years data for Females was reported erroneously in the population estimates table.

The CSO was contacted about this, and it was confirmed that the data for 15-24 year olds had mistakenly overwritten the data for 1-4 year olds when it was originally ingested.

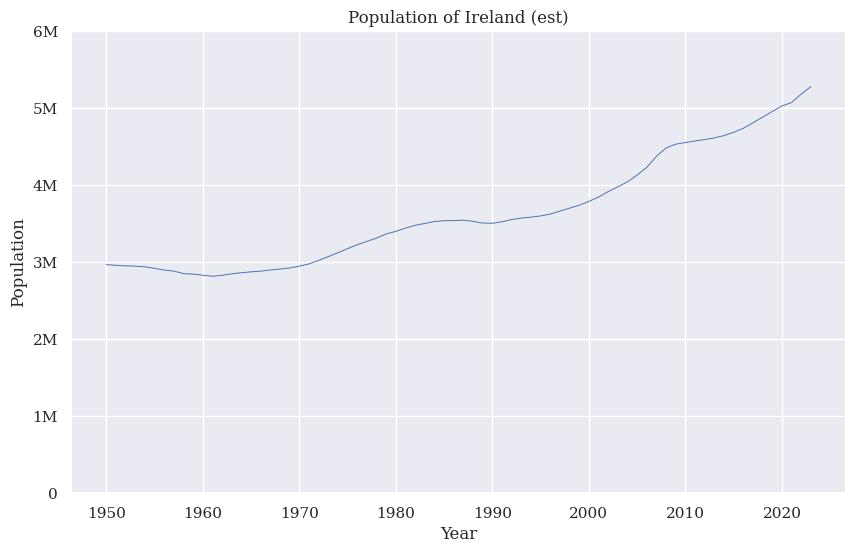
As these values are erroneous, they were also treated as missing values and imputed. As the Female figures could be thought of as a univariate time series, interpolation or other time-series-appropriate imputation methods could be used, such as those outlined by Moritz et al in 2015. However, due the structure of this particular dataset, the opposite procedure to how the missing 1-4 years data was inferred could be used; subtracting the under 1 year value from the 0-4 year value gives a candidate imputation for the 1-4 years variable. This was performed, and the resulting data was found to have no remaining missing values.

Wickham (2011) argues that when data is collected at multiple levels, as was the case here with both the individual figures per sex and their sum being stored on separate rows, the data should be split into tables containing individual observational units. This was done by grouping on the condition that sex = ‘Both sexes’, resulting in two clean datasets – one split by sex, and one containing the data summed across both sexes. The data was then summarised visually – focusing initially on the Both Sexes case in order to give a in idea of high level trends in population over time.

It was seen that, across all 74 years within the data, the mean population count was about 3.6M, with a standard deviation of 718K. We can also immediately see that the data is right-skewed, as the mean is greater than the median, and the upper tail of the variable’s boxplot is longer.

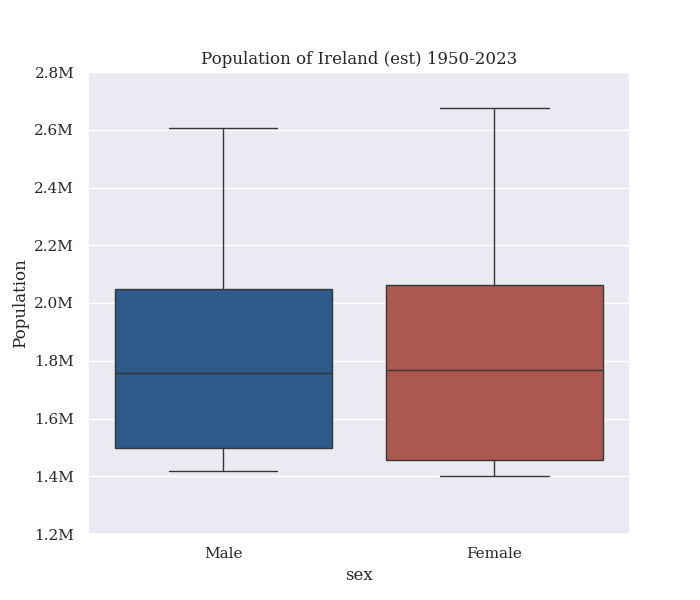
|  |  |  |
| --- | --- | --- |
|  | year | value |
| count | 74 | 74 |
| mean | 1986.5 | 3641.32 |
| std | 21.51 | 718.94 |
| min | 1950 | 2818.3 |
| 25% | 1968.25 | 2954.83 |
| 50% | 1986.5 | 3527.35 |
| 75% | 2004.75 | 4111.65 |
| max | 2023 | 5281.6 |

The year and population variables are both continuous and form a univariate time series, so a line chart is a natural choice of visualisation to highlight the time-dependencies between each successive data point. We see that the overall population trend has been increasing, however periods of decline/stability are seen in the 50s - early 70s, the late 80s - mid-90s, and the late 2000s - mid 2010s. The Y axis was formatted with M suffixes denoting millions to aid readability.



The same trend split by gender was produced. A colour palette that has high contrast to differentiate between the two lines was required, which the seaborn default does give. However, the chosen palette is also colourblind friendly as it was retrieved from David Nichol’s “Coloring for Colorblindness” tool. Tufte advocates for colourblind-conscious choices of palette in aid of creating more “friendly” graphics (p. 183).

It can be seen that the mean population when split by gender is roughly equal, however the distribution of the female population is slightly more right-skewed than the male equivalent.





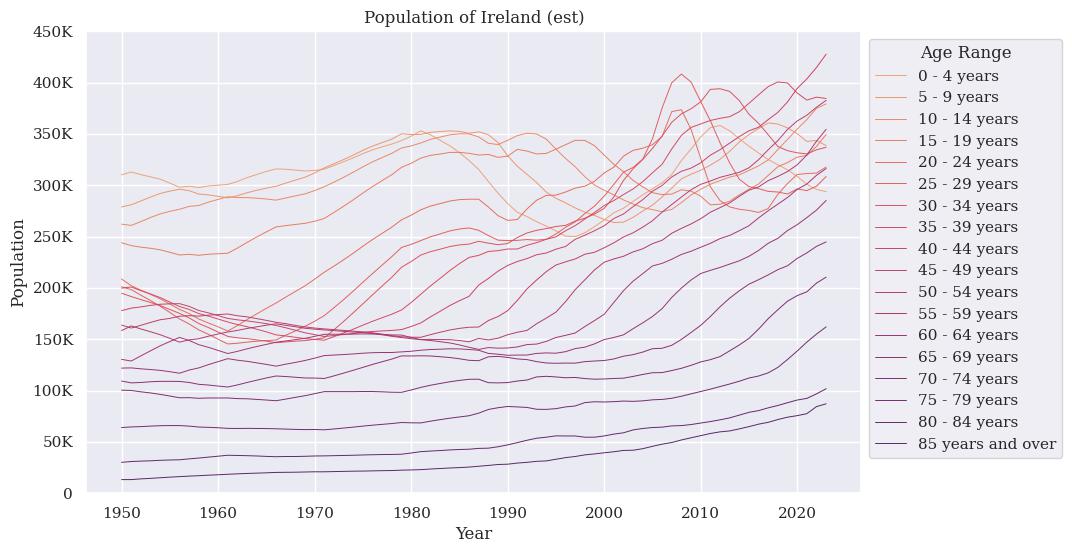
It can again be seen that for most of the time series the gender balance was roughly equal - however in the 1950s there were slightly more males than females, and since the late 2000s there appear to have been more females than males.

Much like the sex variable, there are overlapping age bounds within the data’s age\_group variable. Were the populations per year to be added up, double counting could occur in multiple instances. This can be avoided by using a covering of all ages that is comprised of non-overlapping age ranges.

A covering can be defined using the levels within the age\_group factor in multiple ways; The covering which gives the highest level of granularity / the most levels within the age factor (i.e. - choosing under 1 year and 1-4 years, rather than 0-4 years), choosing to have each of the age ranges (except 85+) be equal width. In either of these cases the 1-4 year or 0-4 year age range which was previously cleaned will be used.

Alternatively, a higher covering with fewer age groups could be chosen to simplify visual summary.

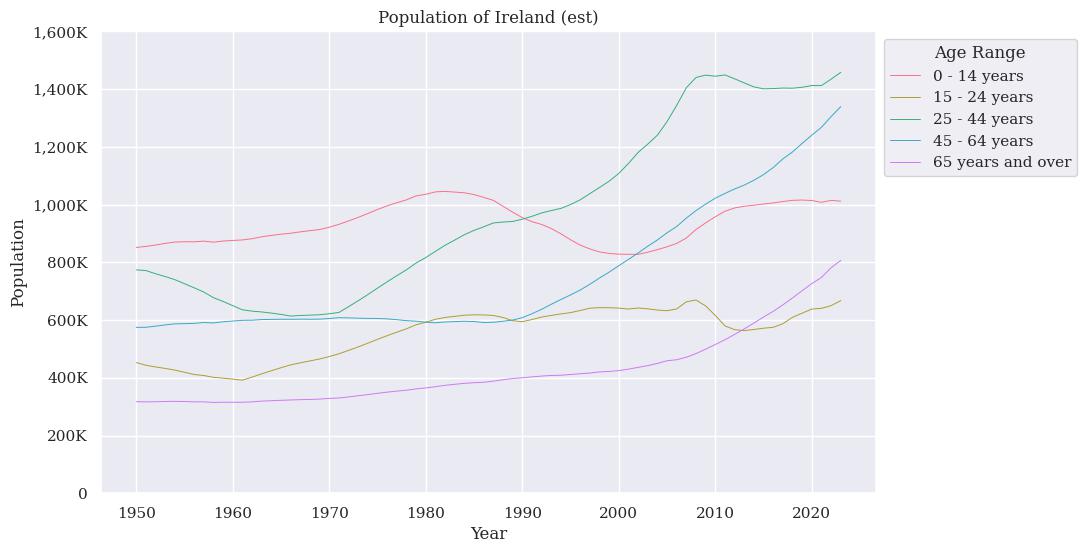
Choosing the most granular covering, population per age group can be visualised as a line chart as before, splitting on age range. However, colour palette previously used does not have enough levels to represent this many age groups, so a sequential colour palette is instead chosen. This darkens in hue across the age ranges. The age levels are ordered, so it makes sense to use a sequential palette as opposed to a categorical one.



The resulting graphic is very busy, or something that Tufte might call “a duck” or “a puzzle”, and is in general diﬀicult to follow – therefore a coarser age covering should be chosen in order to simplify this visualisation. One observation that can be noted from this chart however, is that we can see the same patterns occurring on each line lagged by 5 years, as we have the same population ageing and continuously falling into the next category across time, with the slopes lessening towards the higher age categories indicative of mortality at higher ages.

It can also be seen that the youngest age categories transition from being the most populous in the 1950s, to being overtaken by working ages in the 2000s. This is indicative of a shift in demographic makeup, although it is unclear from this chart alone what the driving force behind this change is.

An alternate covering of the span of ages can be used to make the above visualization less cluttered. With this visualisation, we can clearly see the number of 0-14 year olds declining in the 90s until the mid 2000s, and the number of 25-44 year olds increases dramatically over this timeframe.

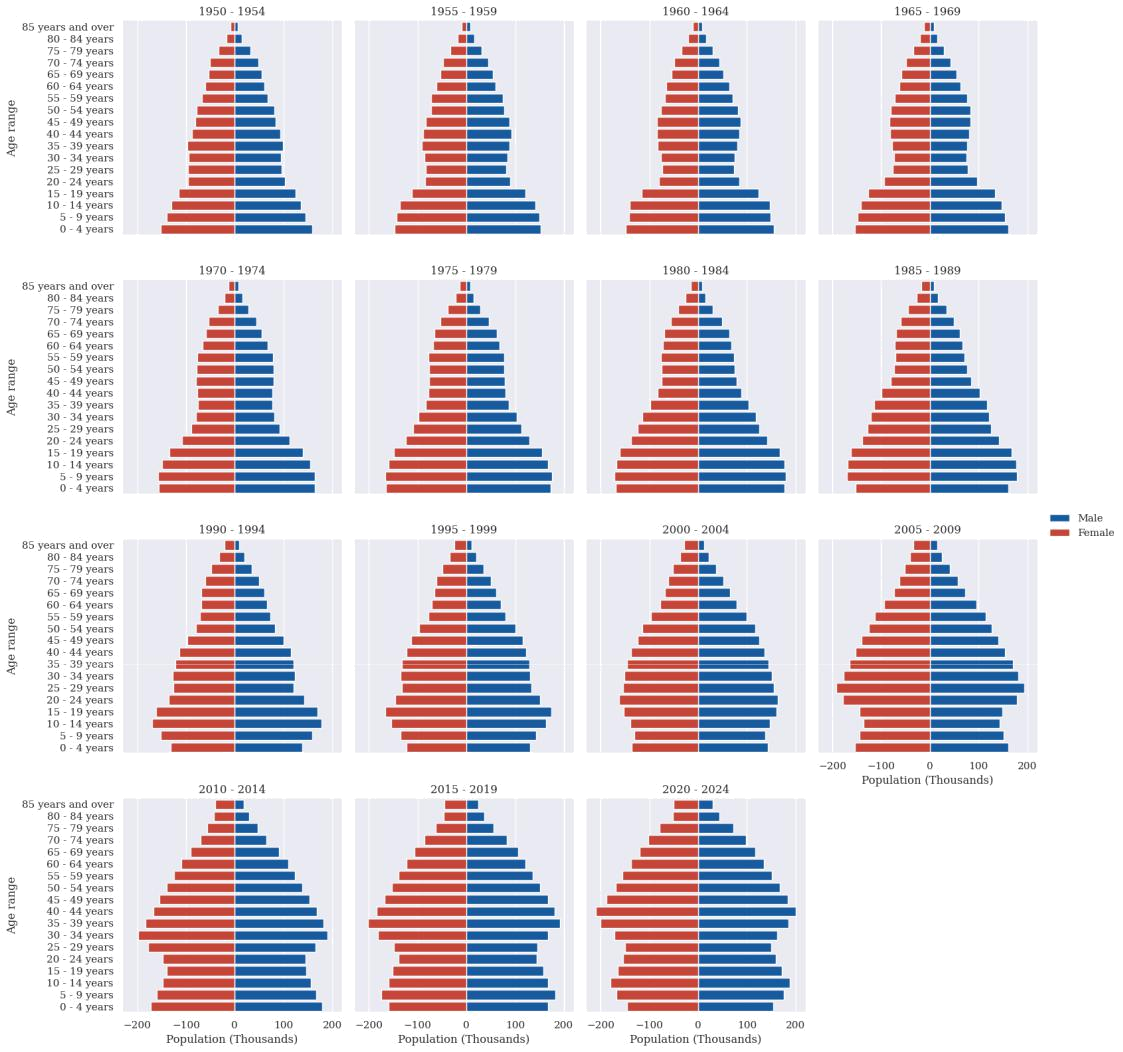


As we are only working with 5 levels in this chart, we elect to switch the palette to one with more variation, so that individual trends can be followed more easily, rather than having the hue change continuously with age as before.

One thing to note here though, is that the age ranges should not necessarily be compared to one another, as the number of ages in each bucket is not equal.

We would also like to look at population age composition separated by year, however 70+ charts would be excessive, so we will bucket the years together into a range, similar to the age\_range variable. Grouping by decade would be the obvious choice, however we can see from the above chart that some of the rapid changes in demographic occur over a shorter period than that, so we will choose buckets of 5 years.

As we will be looking at multiple individual charts, we can also reintroduce gender at this point without creating too much visual clutter.



We can visualise the changing population makeup across genders and ages using multiple population pyramids (or age structure diagrams), which are a type of paired bar plots typically used for this purpose (Wilson 2016). Tufte argues (p. 170-175) in favour of these “small multiple” style visualizations in order to observe a change over time.

This shows that the population in Ireland has shifted from having a relatively wide base - indicative of many births and young people - to having a wide middle and comparatively narrower base, which indicates a lower birth rate. A particularly notable change in the width of the middle ages can be seen in 2000-2010 - which we expected to see given the above line charts.

This could also explain the higher percentage of females in recent years we saw previously, as women have a longer life expectancy than men (Gryclewska, 2016), therefore an aging population would naturally be comprised of more women - the gender difference in the older cohorts can be seen clearly in the plots from 2000 onwards.

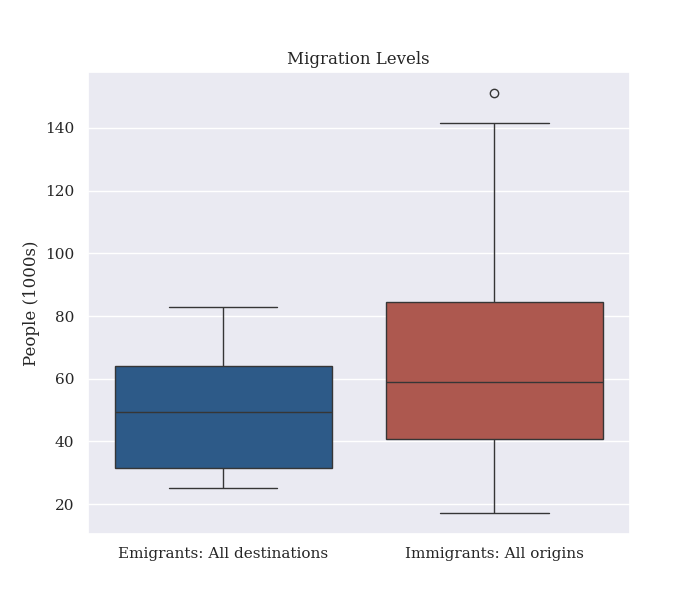
## Migration Data Exploration

We would like to try and gain some understanding of what is driving some of the population changes seen above, and bringing in migration data will help us to do this.

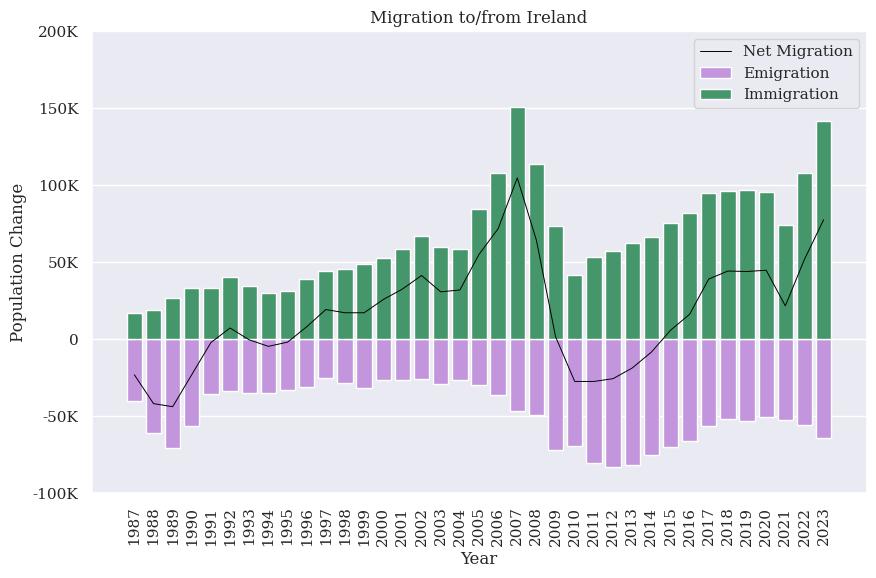
On loading the estimated migration data we see that again, we have missing values in the value column. These primarily impact the net migration and All ages datapoints, so can be repaired using combinations of other values as above.

The net migration values are not missing in more recent data, so it’s possible that this was not originally calculated when records began. We can easily fill in the NaNs in net migration by subtracting emigrants from immigrants with the data in a pivoted form, however we first need to address the other missing values, or else we would be left with missing values corresponding to All ages. We first pivot with age group across the columns, repair All ages by adding the figures for the other ages together, melt and repivot with inward\_or\_outward\_flow across the columns, then finally repair Net Migration. After this procedure we have once again obtained a cleaned dataset.

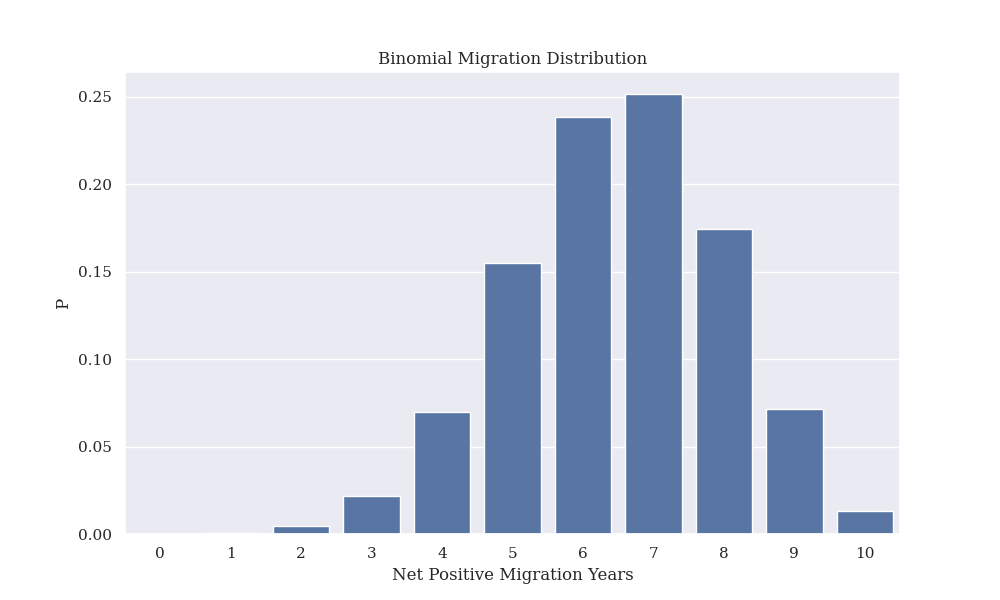
We can observe the distribution of Emigration and Immigration levels across the years using another pair of boxplots – this shows that the mean immigration level is higher than the mean emigration level, and the distribution of immigration levels is right-skewed, with one outlier point.



In order to plot this data further, we continue working with the pivoted form. We can display emigration and immigration as bars, with an accompanying line denoting net migration. Although emigration results in a reduction in the population and immigration results in an increase in the population, I have elected not to use typical increasing / decreasing colour choices (blue for increasing, red for decreasing, etc.) as these may also be interpreted as ascribing goodness or badness to either phenomenon, which I want to avoid, while still choosing colours that are visually distinct.



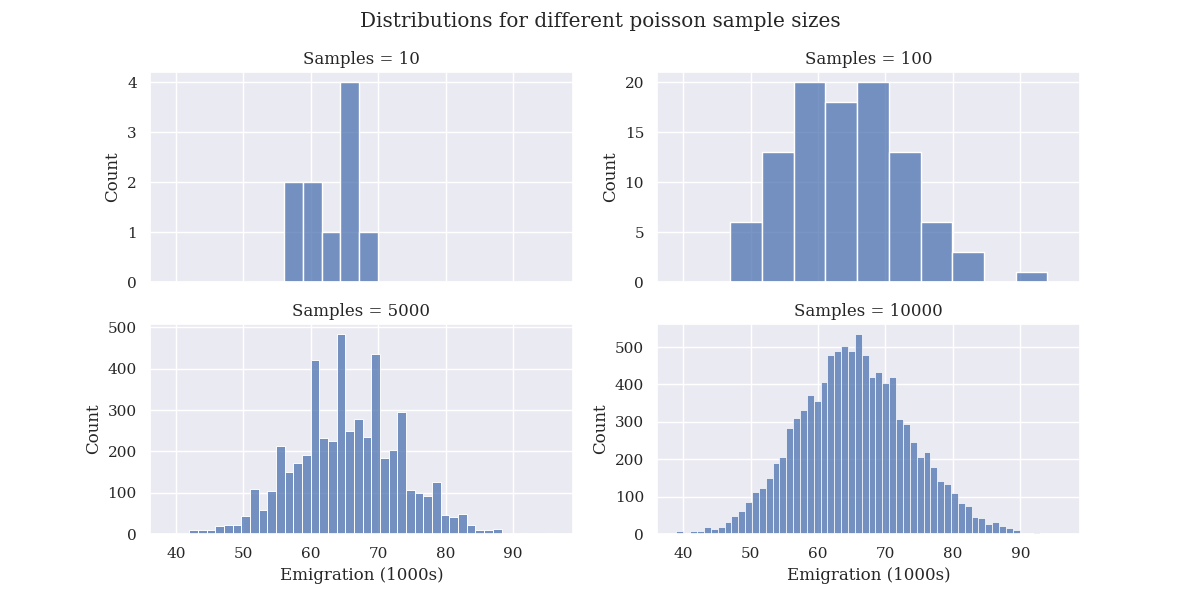
We can model the relationship between emigration and immigration using a binomial distribution; where each year is an independent Bernoulli trial with success being defined as immigration > emigration, or that net migration contributes to population growth. If we denote 1 as a success and 0 as a failure, we can estimate the success probability over all years by taking the mean of our success column.

We can then draw up the binomial distribution for the number of successes in a decade using the binom.pmf() function by setting our number of trials to 10. As our estimated P(X=x) is 0.648 (> 0.5) our distribution is left-skewed (Weiss, 2017).

Looking at the decades before and after the millenium, we can see that the 90s had 5 successes, which according to our model had a probability of 0.155 of occurring, wheras the 2000s had 10 successes, which had only a probability of 0.013 of occurring. This indicates that the distribution of successes may have shifted over time, as the probability of the post 2000s data occurring according to our model is low.

We can also model migration using a poisson distribution - in this case, by treating one immigration event as a success, we can calculate lambda as the mean number of immigrants per year. Using this, we can estimate the probability that immigration is greater than the mean emigration level, which over the wholde dataset comes out to have a probability of 0.487

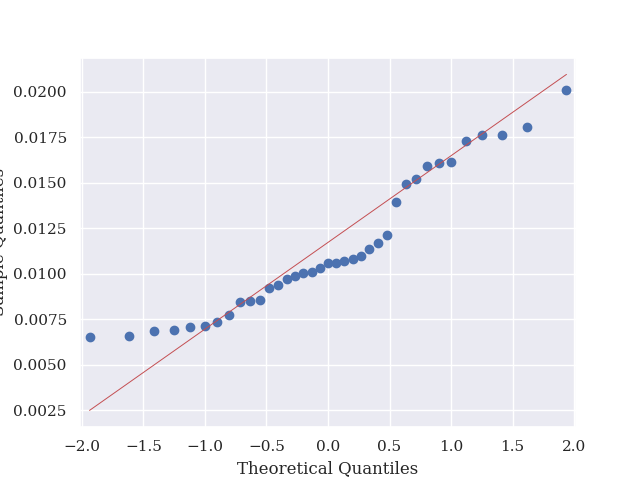
Additionally, by drawing successively larger random samples from this distribution we can see how it converges to a normal distribution in the limit.



### Comparing pre & post 2008 Emigration

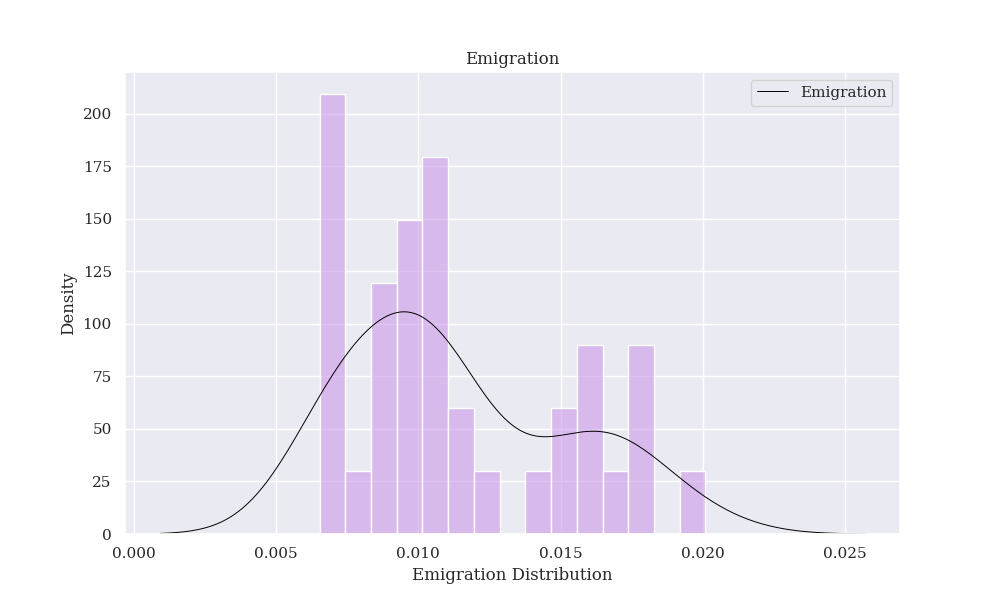
There was a period of net emigration after the 2008 crash. Looking at the above line and bar graph there was an inflection point in 2008, and it looks like before and after this our emigration figures can be thought of as coming from different distributions.

We can use a qqplot to check the emigration variable’s distribution agains a theoretical normal distribution. Although, we would expect emigration to increase as the overall population increases, so in this case we can look at emigration as a percentage of total population to control for this.

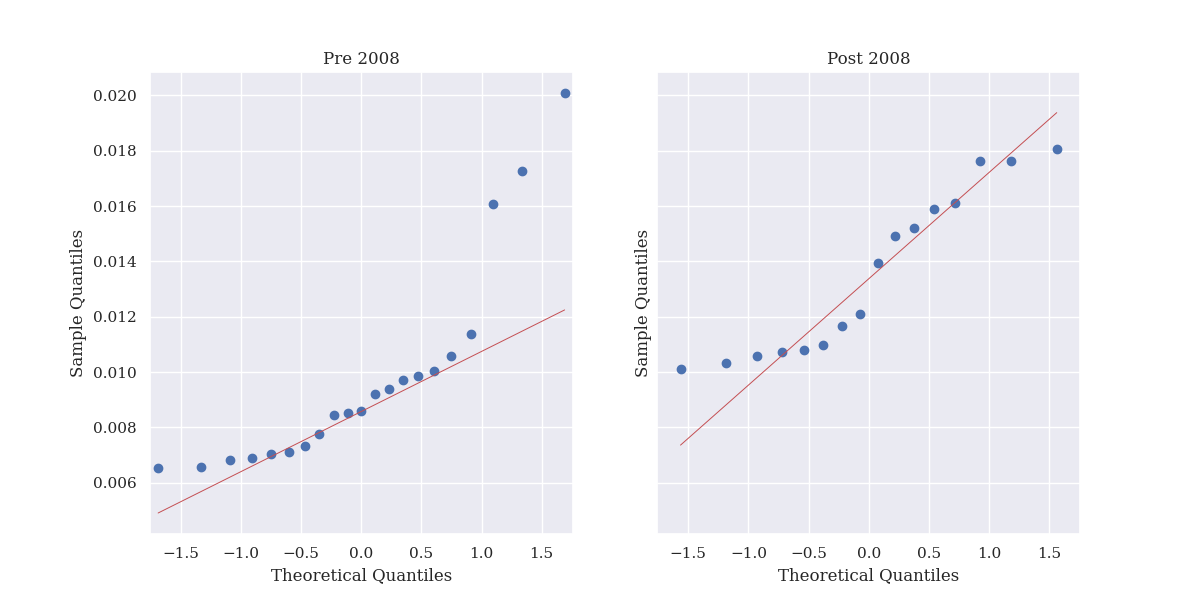


This is a poor fit - implying a nonnormal distribution for the overall series.

The inflection point at 0.5 indicates bimodality - something we can again see with a histogram.



So we look at separate qq plots pre and post 2008 to see if these partitoned series are distributed normally.



Neither of these are convincing fits, so this temporal partition does not cleanly split our data into two normally distributed subsets, and our emigration variable is not drawn from a Gaussian mixture delineated by 2008.

### Bootstrap Comparison

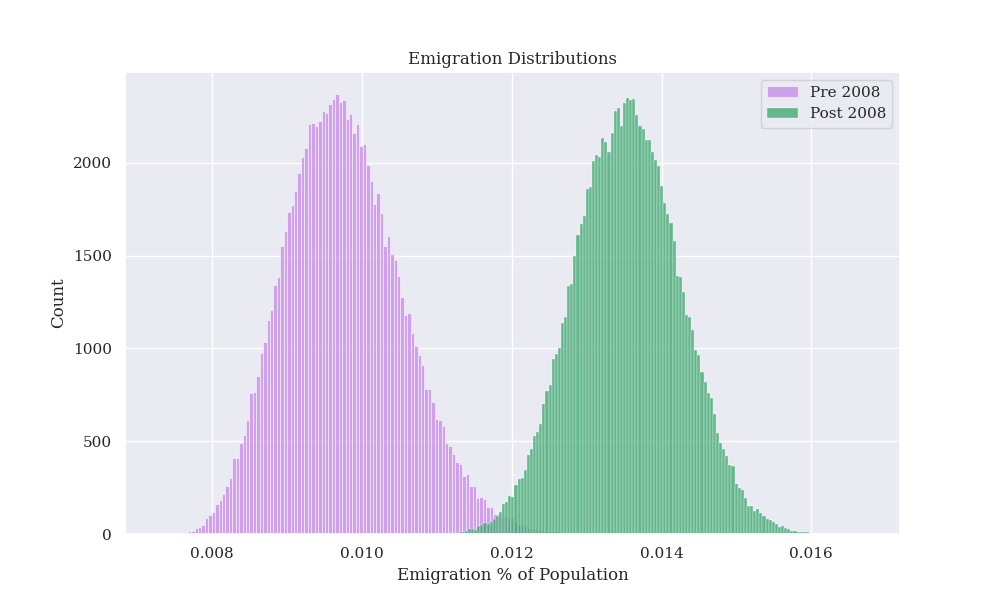
Even without knowing the exact distributions of these subsets, we can investigate whether they are drawn from the same distribution numerically using a non-parametric bootstrap (Efron, 1994). This allows us to construct confidence intervals for any given statistic by the following procedure.

- A bootstrap sample is chosen by drawing from the data with replacement up to the size of the original data.

- The chosen statistic (typically a measure of central tendency, in this case the mean) is calculated on each bootstrap sample

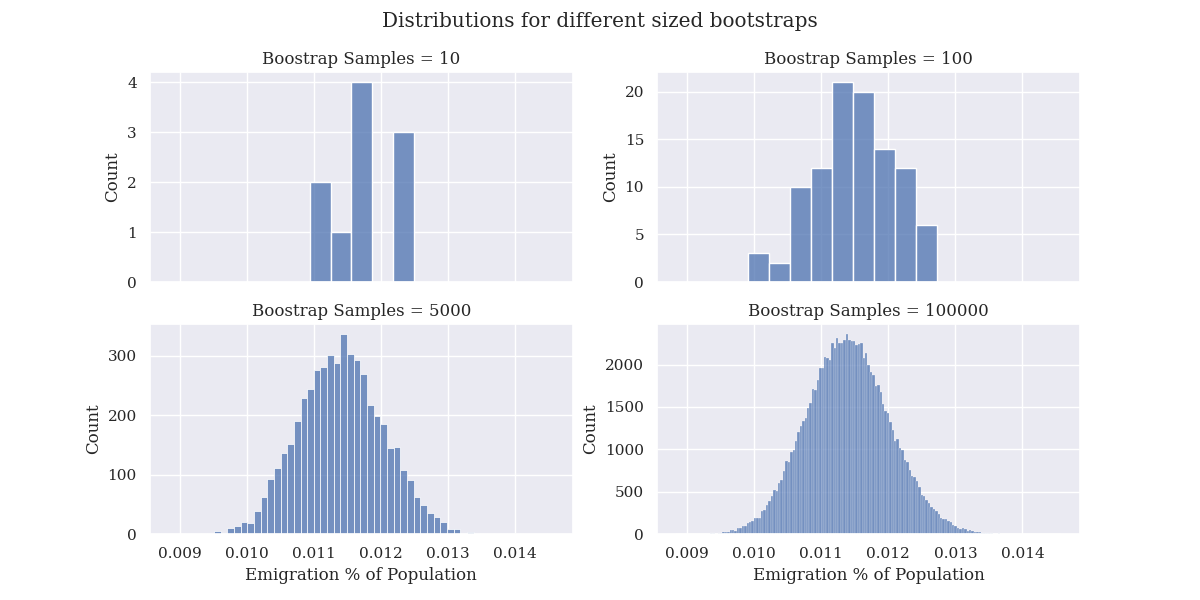
- Quantiles are derived for the computed statistics, and we can obtain a confidence interval from this (95% by convention)

Different distributions have different moments, so if two datasets result in non-overlapping bootstrapped confidence intervals, it follows that they are drawn from different distributions, as illustrated below.



The 95% confidence intervals for the means of the pre and post 2008 distributions do not overlap, so we can reject the hypothesis that the data were drawn from the same distribution, and conclude that the mean emigration rate was significantly greater post-2008.

Additionally, we can observe that the distribution of means of our bootstrap samples are normally distributed thanks to the central limit theorem, with the distribution converging to normality as the number of samples increases. This occurs even when starting with a nonnormal distribution, such as the combined pre and post '08 emigration\_pop\_pct variable below.



## Predicting future Emigration Levels

We would like to be able to forecast emigration levels into the future based on past data. We can frame this as a supervised learning application, as we know the value of our target variable - the emigration level. In this case, as we have few features to work with, we can use the lags of our variables in addition to the variables themselves. This can be thought of as an autoregressive model, which are commonly seen in time series analysis (Chatfield, 2003).

It is also important that when we split our data into train and test sets, we set a temporal cutoff, and don't randomly sample data points as would be typical when generating a train/test split. If we were to randomly draw data points from the full history, data leakage could occur and the model could become informed of values in the test set based on information in the training set. This procedure of withholding the tail of a time series is commonly used in traditional forecasting applications (Bergmeir & Benitez, 2012).

We will compare multiple regression algorithms - OLS regression, linear regression with L1 and L2 regularization (Lasso & Ridge), ElasticNet, and Random Forest Regression, all made available through the scikit-learn package.

Finally, we will compare the performance of these algorithms to that of a method formulated specifically for time series data - ARIMA (Autoregressive Integrated Moving Average models), which can conveniently be tuned using the pmdarima package, which ports R's auto.arima() functionality to python.

### CRISP-DM

We are following the CRISP-DM process outlined by Wirth & Hipp (2000), which gives an iterative process for the progression of data mining projects. We have already gained an understanding of our data and done some preliminary data preparation. We will now do additional data preparation in order to frame our problem in a way compatible with the machine learning method we wish to apply, before actually applying those methods, evaluating them, and finally either deploying (or in this case accepting) the finalised model, or iterating through the previous steps in order to gain a deeper understanding and improve performance.

### **Data Preparation**

Some additional data processing is required in order to get our dataframe into a format where machine learning algorithms can be applied. We first want to get our independent variables into their own dataframe; the continuous variables of year, and population, migration, and emigration levels, and the categorical variables of sex and age group. We will generate lags for each of population, emigration and migration, and also generate 1 lead term for emigration to serve as our target variable. The number of lags is parameterised and will inform our train/test split criteria later, in this case 3 lag terms was selected.

As the first few years of the data don't have data in their lag terms, these come out to be null. We will ignore these rows and just deal with the rows with complete data - we could alternatively back-fill the earliest seen value, but in order to not bias the models towards older values this was not done and the rows were removed instead.

Additionally, the last year has no lead variable for the target, as that is in the future, so these rows are discarded as well.

Migration Exploration & imputation

**Forecasting & Machine Learning**

ML data prep

Regression

Hyperparameter tuning

Iteration

Binary Classification

**References**

* Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences 191, 192–213. <https://doi.org/10.1016/j.ins.2011.12.028>
* Burkov, A., 2019. The hundred-page machine learning book. Andriy Burkov, Polen.
* Burnham, K.P., Anderson, D.R., Burnham, K.P., 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. ed. Springer, New York.
* Chatfield, C., 2003. The Analysis of Time Series, 0 ed. Chapman and Hall/CRC. <https://doi.org/10.4324/9780203491683>
* CSO, n.d. 2016 Census.
* Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap, 0 ed. Chapman and Hall/CRC. <https://doi.org/10.1201/9780429246593>
* Frost, Jim, 2020. Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models.
* Gryclewska, B., 2016. Why women live longer than men? Review of biological and non-biological factors. Przeglad lekarski 73, 392–394.
* Hastie, T., Tibshirani, R., Friedman, J.H., 2017. The elements of statistical learning: data mining, inference, and prediction, Second edition, corrected at 12th printing 2017. ed, Springer series in statistics. Springer, New York, NY. <https://doi.org/10.1007/b94608>
* James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer Texts in Statistics. Springer New York, New York, NY. <https://doi.org/10.1007/978-1-4614-7138-7>
* Martin, R.C., 2012. Clean code: a handbook of agile software craftsmanship, Repr. ed, Robert C. Martin series. Prentice Hall, Upper Saddle River, NJ Munich.
* Mckinney, W., 2010. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference.
* Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., Stork, J., 2015. Comparison of different Methods for Univariate Time Series Imputation in R.
* Nichol, D., n.d. Coloring for Colorblindness.
* Tufte, E., 2001. The visual display of quantitative information / E.R. tufte. American Journal of Physics 31. <https://doi.org/10.1109/MPER.1988.587534>
* Weiss, N.A., 2017. Introductory statistics, 10th edition, global edition. ed. Pearson, Boston; Columbus; Indianapolis New York.
* Wickham, H., 2014. Tidy Data. J. Stat. Soft. 59. <https://doi.org/10.18637/jss.v059.i10>
* Wilson, T., 2016. Visualising the demographic factors which shape population age structure. Demographic Research 35, 867–890.
* Wirth, R., Hipp, J., 2000. Crisp-dm: towards a standard process modell for data mining.
* Yee, J.L., Niemeier, D.A., 1996. Advantages and disadvantages : longitudinal vs. repeated cross-section surveys.