The primary aim for this project is to investigate population demographics specifically within each region of the Republic of Ireland. I will then use this data to continue my investigation into a broader subject, analysing the number of “displaced persons” in Ireland.

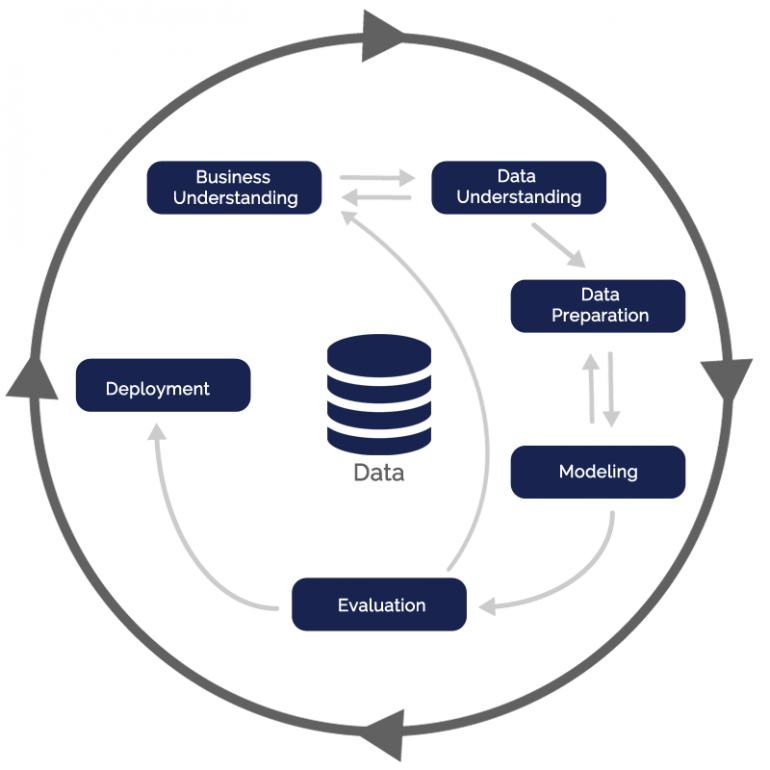
I would like to investigate key statistics such as Asylum application acceptance rates, Refugees numbers and potential influences on the overall level of displaced persons within Ireland.

Before I begin, I think it is necessary to understand a few key definitions before continuing with my report. I would first like to outline a number of key definitions which will be critical in enabling a comprehensive understanding the analysis.

Not every asylum seeker will ultimately be recognised as a refugee, but every refugee is initially an asylum seeker. Therefore, asylum-seekers and refugees are mutually exclusive. Meaning that at any given point in time, a person cannot be labelled both a refugee and an asylum-seeker. When I speak about this group collectively, I will refer to them as “Displaced Persons” for convenience.

CRISP-DM (“Cross-Industry Standard Process for Data Mining”) is a staple of a successful data science projects, outlining the typical phases that are involved in a project and the relationships between each of these phases. I have used a real-world scenario to outline each phase below.

Over the course of this investigation, I have used multiple datasets and implemented various machine learning models. As I have completed this, I have attempted to mimic the basic concepts outlined by CRISP-DM.



I am an Actuary by trade, I will discuss each step of the CRISP-DM framework regarding a common actuarial matter, focusing on insurance premium modelling.

1. **Business Understanding:** An insurer aims to set appropriate premiums for different customers based on their individual risk profiles. The objective is to accurately predict the expected claims for each policy.
2. **Data Understanding:** Historical claims data and customer demographics are often explored to understand patterns and correlations affecting claim occurrences.
3. **Data Preparation:** Data cleaning involves handling missing values and outliers. Features like policyholder age, coverage types, and claim history are prepared for modelling.
4. **Modelling:** Linear regression or generalized linear models (GLMs), are used to predict the expected claims amount.
5. **Evaluation:** The model's performance is evaluated using metrics like Mean Squared Error to measure the accuracy of predicted claim amounts.
6. **Deployment:** The actuarial model is integrated into the insurance company's systems to determine premiums based on predicted claims.

To begin my analysis I will explore population in each region of Ireland. An early decision was taken surrounding how to divide the Republic of Ireland into respective regions. After some independent research, I concluded that subdividing Ireland into the 7 NUTS3 regions would be most appropriate. I came to this conclusion as I felt that the 26 county divisions were a step to far regarding the level of detail and the NUTS3 regions were more directly comparable to each other.

To counter this point, I felt that using the NUTS2 regions did not give me enough granularity to discuss explore particular avenues.

Firstly, I read in each of the CSV files, taking an initial look at the dataset I saw that there are a number of columns which would be unnecessary for the investigation, so edited each dataframe to only include necessary columns.

After some simple data manipulation I was able to see that the “PEA04” dataset ranges from the 2011-2023 whereas the “PEA07” dataset ranges from 1996 – 2017. This meant that I had a decision to make surrounding how to join the datasets. Ultimately, I decided to remove the year of 2011 – 2017 from the PEA07 dataset.

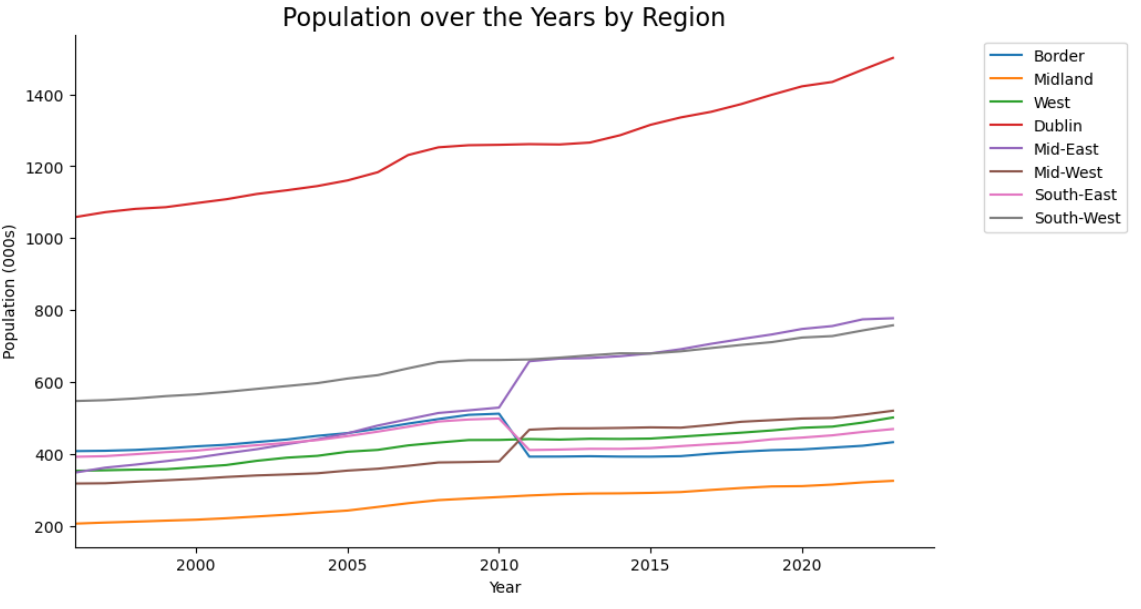
I then used Panda’s concat function to join the two datasets into a larger dataset to use for my investigation. However, I soon realised that there was a problem with this newly constructed dataset, which I will discuss below.

I used some common functions to analyse this new data such as head, describe and info in order to give myself the opportunity to understand and visualise the data using these key statistics.

I also removed some unnecessary data such as the data relating to gender, this would perhaps be useful to keep if wanted to extend the investigation but is unnecessary for what I want to achieve for now. I also removed particular rows which overlapped with our individual data subsets, I didn’t want these rows disrupting future calculations.

I created a ‘dictionary’ which would allow me to pull data from particular years. I created this dictionary using a loop function which created a subset of the original dataframe for each year and assigns this dataframe to the dictionary.

As we are looking at population growth by region it would be most appropriate to use line charts to display my results. I constructed a line chart showing the population of each region for each year from 1996 – 2023.



The chart produced is correct despite the visual suggesting an error occurring between 2010 and 2011. This is, as mentioned previously, caused by the changes that were made to the NUT3 regions in 2011, detail of this can be found in the Jupyter notebook and I will not repeat here.

The chart is correct as it details the populations of each NUT3 regions for each year as per the geographical definition of these regions at the time of the data being recorded. I do acknowledge that I will have to be careful with how I will use this dataset in the future given this breach of geographical consistency.

I then thought that it would be useful to visualise our dataset using a practical real-world visual and decided to plot a choropleth map detailing the population of each region for each year.

I was able to source a GeoJSON file online which details the geographical data on the Republic of Ireland. I explored this GeoJSon file to find the locations of the NUT3 Region and proceeded to manipulating the dataset to assign colours and labels to each region.

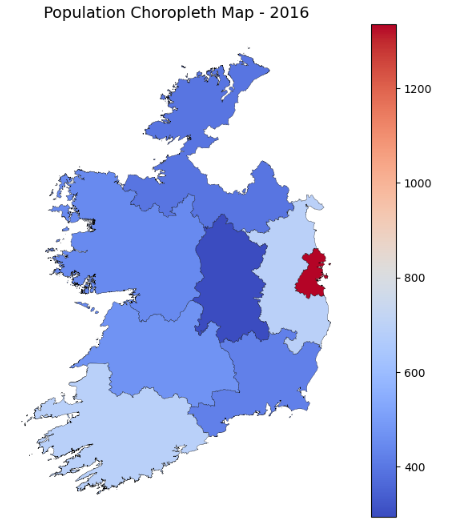
I was able to add labels to each of the regions using their respective centroid co-ordinates. I included a number of if and elif (else-if) statements to ensure the labels did not overlap each other.



My aim was to then combine the dictionary I created earlier along with my newly created map to visualise our data in the form of a choropleth map.

Before combining the two datasets I took a look at each dataset separately to see if I noticed any inconsistencies and I soon spotted one. The GeoJSON file had a slightly different naming convention as it is using ‘Midlands’ instead of ‘Midland’ which was used in the CSO data. I corrected this before proceeding with my calculations.

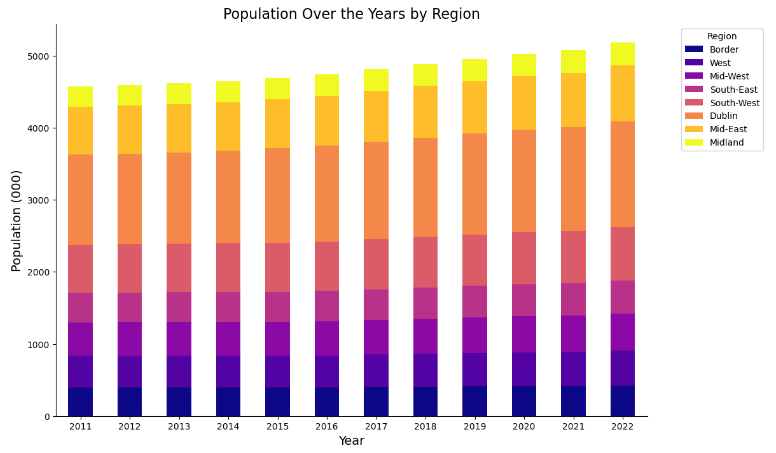
I merged the two datasets using a loop function which cycles through each year in the dictionary and assigns the respective regional populations to the GeoJSON dataset. Once I have that combined dataset then it is quite straightforward to plot my results in a choropleth map. I started by plotting this for only one year, 2016. I also performed a quick exercise to remove the region of Dublin, purely out of curiosity as the Irish population is highly concentrated around Dublin.



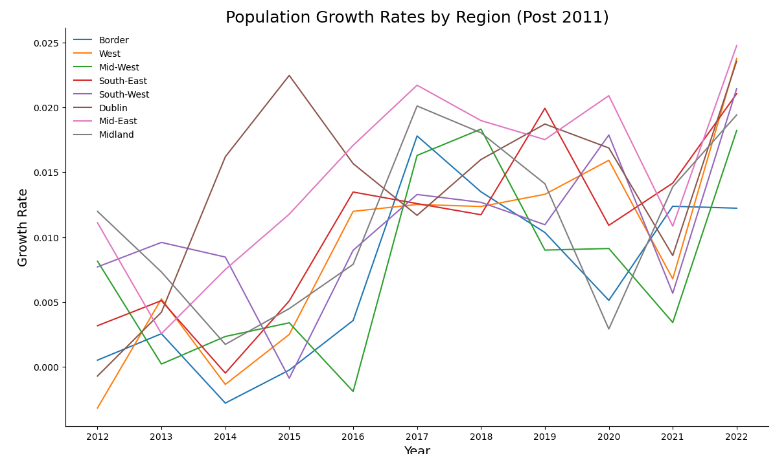
I also wanted to display this data for each year using an interactive display. I was able to find a

function online which allowed me to do this. I was able to implement this supplemented by my earlier code. I made use of the vmin and vmax variable to keep the legend consistent even when the year is updated.

I was conscious that the regions plotted on the map were aligned with the latest NUT3 geographical borders so I made sure not to plot any data prior to 2011 as this simply wouldn’t be accurate with respect to the current regional borders.

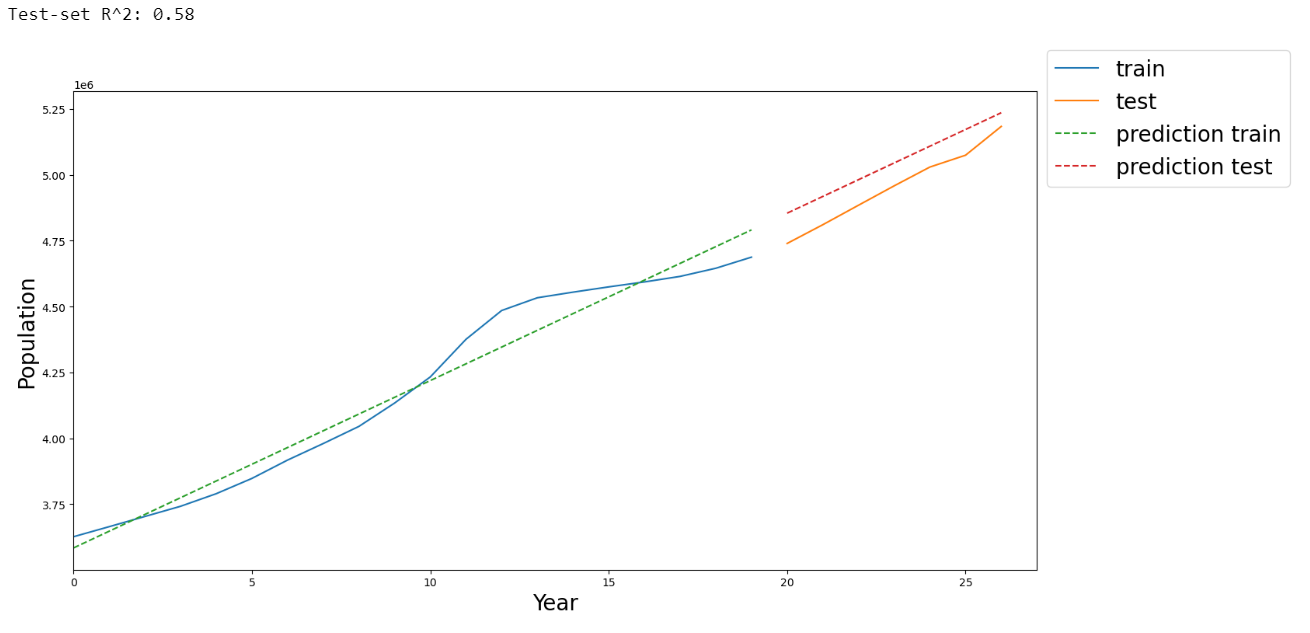
Alternatively, I can display this data in the form of a stacked bar chart, like the one below:

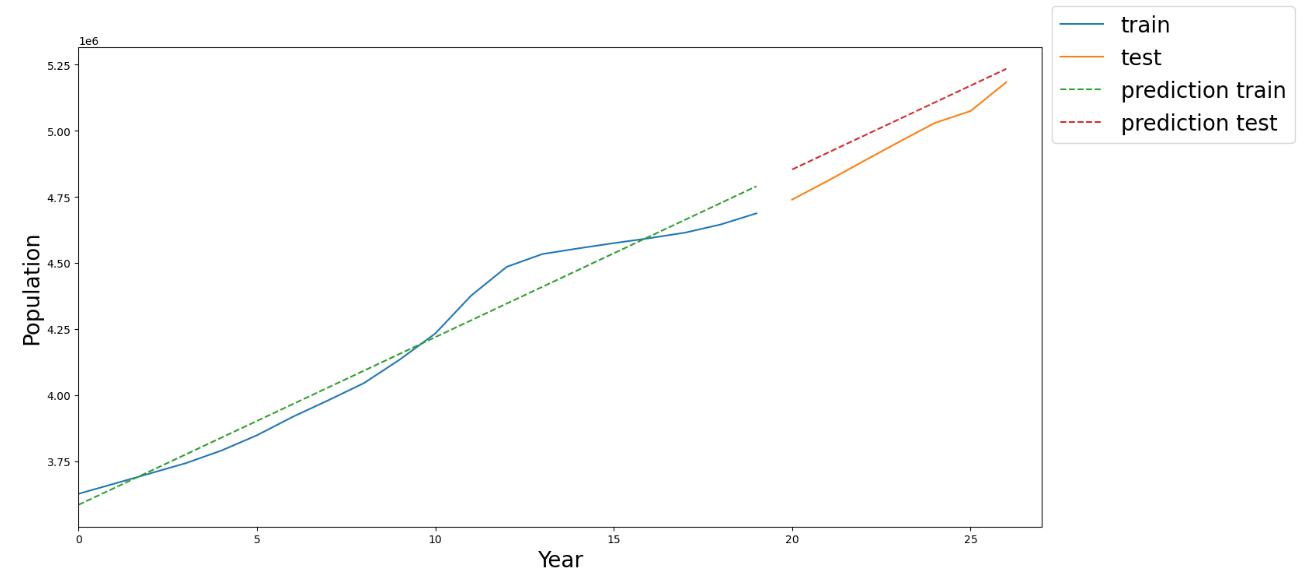
Next, I was keen to investigate the population growth in each region in terms of growth rates and not just looking at the absolute population figures. I was able to take the merged dataset that was used in the choropleth exercise and convert this into a standardised growth rates table using a reasonably straightforward loop function.

Acknowledging the lack of continuity in the geographical details of my dataset I made the decision to split my dataset in pre-2011 and post-2011 figures for growth rates and display these individually.

To further the investigation I wanted to apply a regression model to our population dataset. After loading in the necessary libraries and functions, we adjust the dataframe to suit the desired format. I have created the data in such a way that the model can be easily adjusted by the user. The user is able to select which regions to model or they can choose to use the aggregated population. For the purposes of the report I will only discuss the aggregated figures.

We then outlined the number of datapoints to be used in the model, I have chosen to use the first 20 as our training set. I then implemented a custom function which is taken from one of my Machine Learning lectures. This function is great as it allows the user to specify the type of regression for the model.

I began using linear regression with results below:

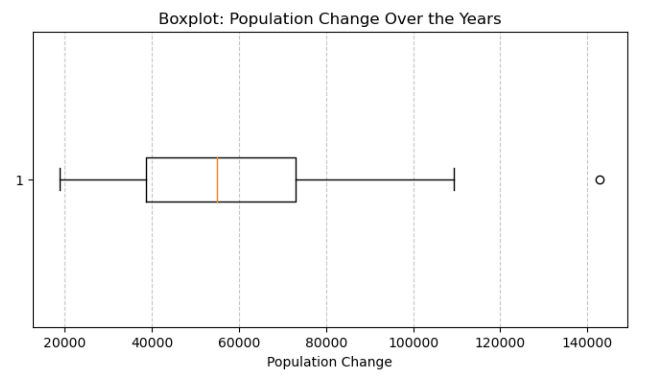
Results look reasonable but I decided to try different regression types to develop the model. Next, I used ridge regression.

The ridge regression results are slightly improved compared to the linear model. I also applied a Random Forrest regression model, but the results were poor.

I decided not to delve too deep into this subject as I felt my dataset was small and the validity of any further analysis would be limited.

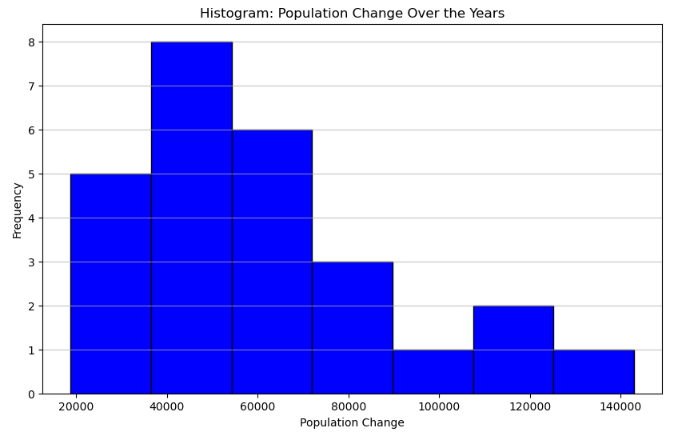
As per the assignment criteria, I am required to apply a normal distribution to explain some aspect of my data. I theorised that the annual population change could potentially follow a normal distribution, so I decided to put this to the test.

Firstly, I plotted the population change data in a boxplot, I visualised the probability distribution if we assumed the data was normally distributed.



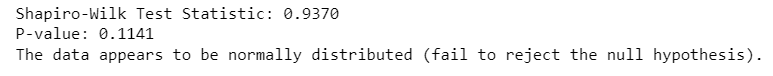
If the dataset follows a normal distribution, the boxplot will show a symmetrical box with the median line in the centre, so this was promising.

I then plotted my data into a frequency histogram to visual the distribution of the data.



The data does appear to have that characteristic bell-shape that is often associated with normal distribution, but there is clearly a number of outliers and small sample size is also evident.

To remove my doubt I decided to perform a Shapiro-Wilk test and have printed the results below.



It’s important to note that the test might not be perfect, and examination of the histogram is also valuable.

As we can see the p-value (0.1141) is larger than the significance level of 0.05, and so we fail to reject the null hypothesis. Therefore, there is not enough evidence to conclude that the population

change is not normally distributed.

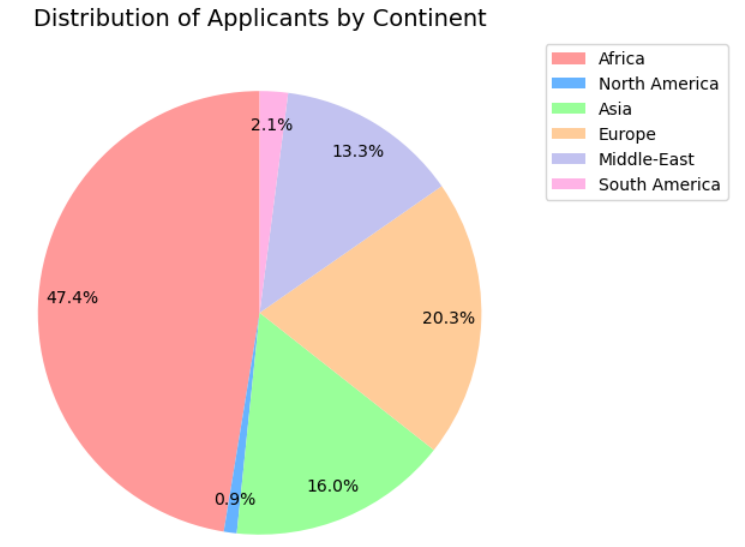
Before I began investigating the population statistics in terms of proportion of displaced persons, I wanted to explore the statistics underlying the individuals who are applying to enter Ireland as asylum-seekers, in particular applicant’s success rate and the factors influencing a successful applicant.

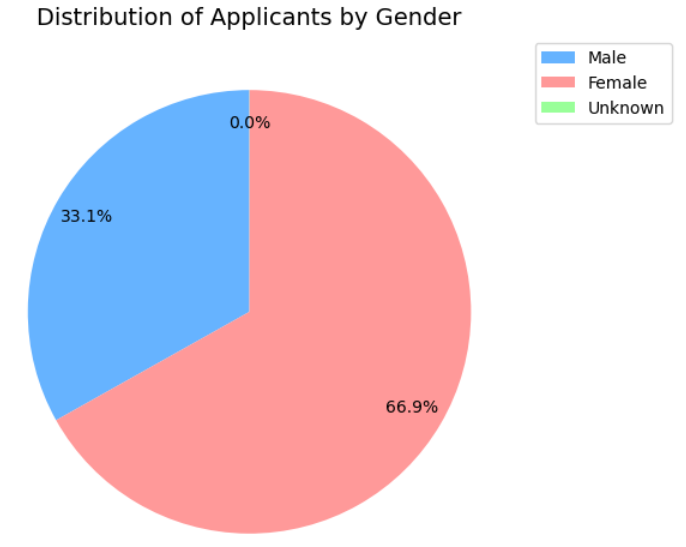
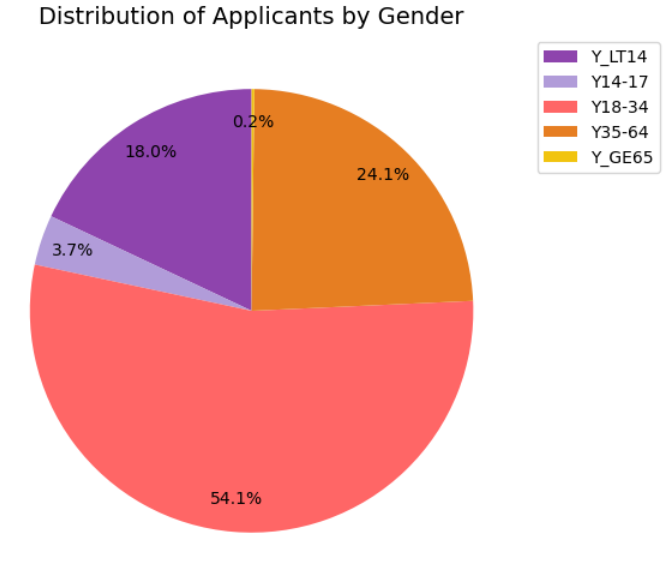
I begin by reading in my dataset, “Dataset 3) Eurostat - Application Demographics”, details of this can be found in the Appendix. I performed some basic data manipulation by trimming the dataframe to exclude the columns and rows that will be unnecessary for our investigation.

I then wanted to flesh out our dataset by adding extra columns which will be useful further into our investigation. I did this by reading in a custom dataset that I myself have created called “Dataset 4) Additional Country Info”. Again, you can read details of this dataset in the Appendix.

After joining our 2 datasets using Panda’s merge function, I then proceeded to create some visualisations in order to get a clearer understanding of our dataset.

I wanted to get a sense of the demographics of the dataset, so I created 3 pie charts each displaying a different feature of these individuals who are applying for asylum in Ireland.

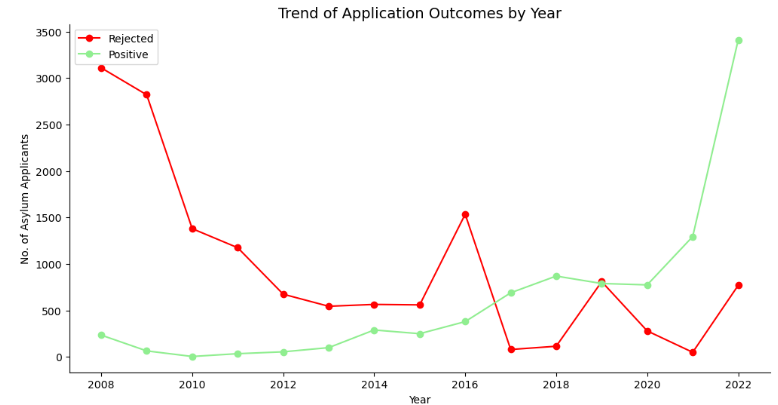




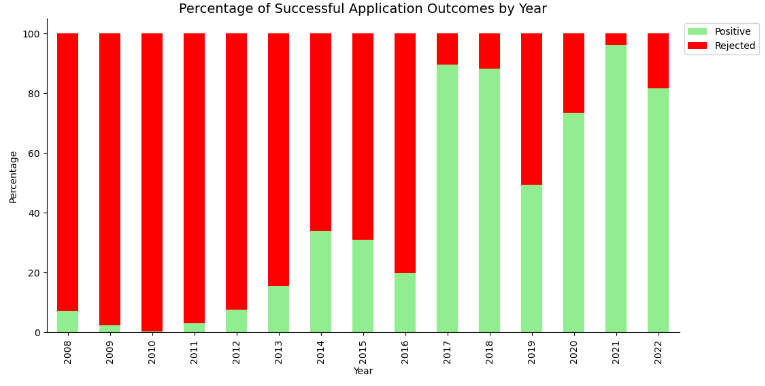
From these charts we are able to see that the majority of applicants are African, female, and they are generally between the ages of 18-34.

I then moved on to exploring application success rates but unfortunately my prior dataset does not include this information. Therefore I began by reading in another dataset (“Dataset 5) Eurostat - Application Outcomes”) that now includes the decisions made on each application by year.

I then proceeded to refine this dataset in a similar way to how I edited the initial dataset above. However, I then added an additional step where I split the dataset in 2 datasets, one for rejected applicants and one for successful applicants. I then used Panda’s merge function to re-join these two datasets side-by-side so that we now have a column for rejected applicants and successful applicants.

After grouping my data by year, I began by plotting a line chart depicting the number of successful and unsuccessful applicants by year.

Alternatively I can display the same data as percentages by year, and so I plotted a bar chart showing the success percentages for each year from 2008-2022.

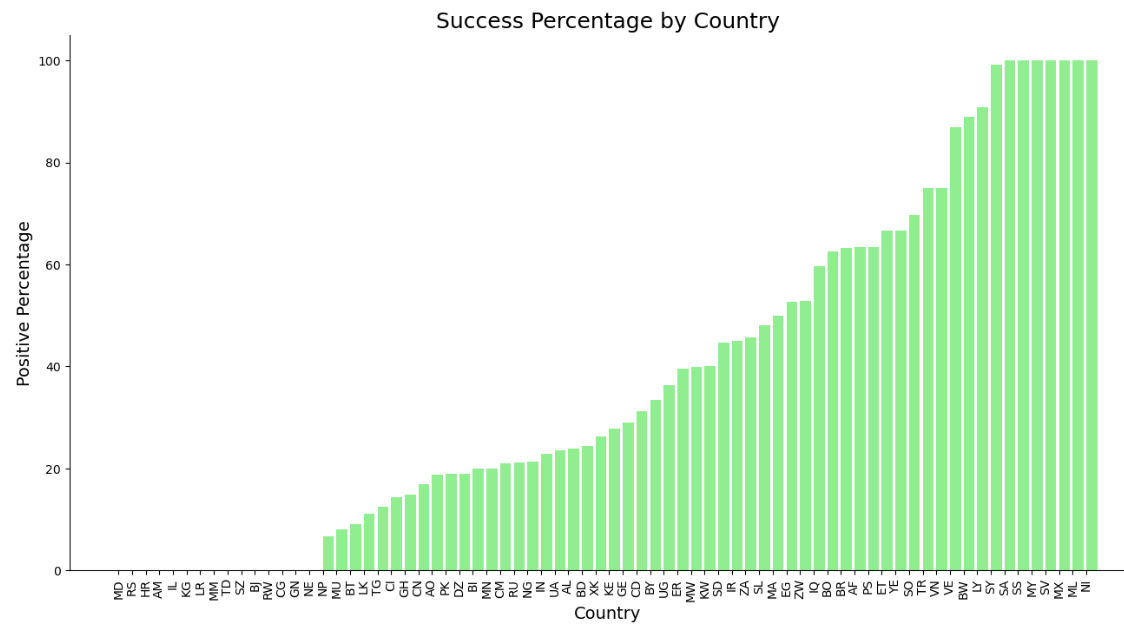


This shows a clear trend that recent applicants are experiencing a higher success rate compared to applicants from prior years. There appears to be a significant jump in 2017 in particular.

I have used a green/red colour scheme for both of these charts which is quite intuitive given the common associated of green with success and red with failure.

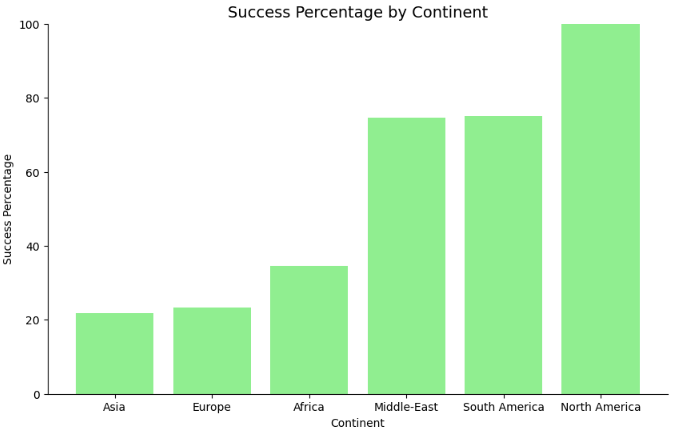
Next, I wanted to look into the various success rates that are experienced by applicants and so I decided to investigate where citizenship (or ‘Country of Origin’) has any influence on success rates.

I began by grouping my data by citizenship using the groupby function and then adding a column showing the percentage of successful decisions by each country. I also reordered the data and plotted this data using a simple histogram.

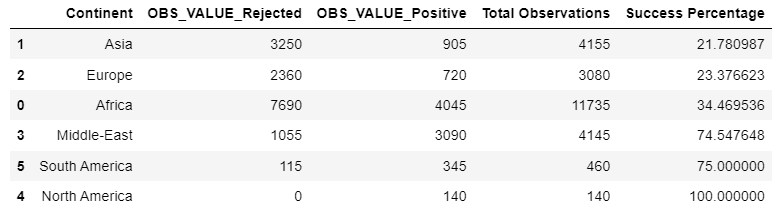


After taking an initial look at this data I thought that the data could potentially be normally distributed, so I decided to perform a couple of high-level checks in an attempt to confirm my theory. These checks both concluded that my data was indeed not normally distributed. I will not discuss in detail here but have left these tests in my Jupyter notebook for completeness’ sake.

Ultimately, I decided that using country as a variable in my investigation was much too granular. I decided to manipulate my data to instead group the applicants by continent (or continental citizenship to be exact). I then merged this with my “Additional Country Info” dataset to expand the data to now include information like Continent, GDP per Capita and Global Peace Index.



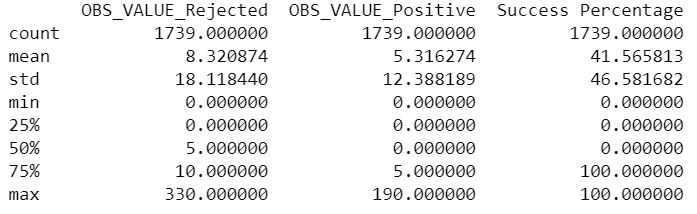
From the data above we can see the apparent discrepancy between the various success rates between the 6 observable continents.



It is important to note that although North and South America have the highest success rates, they also have a significantly lower number of applicants in general. This could suggest that perhaps applicants from these countries are prioritised in some way because of their rarity. Alternatively, applicants from the American continents may submit a higher quality of application given their distance from Ireland. It is interesting to note there are 0 applicants from Oceania, potentially signifying that distance from Ireland plays a crucial role in number of applications.

Now that we have concluded which variables are appropriate, we can now perform one final data manipulation step to prepare our dataset to be used as a model input. I was concerned that each row of these previous datasets did not follow the convention that one row would be one observation. So I decided that converting my data to follow this convention.

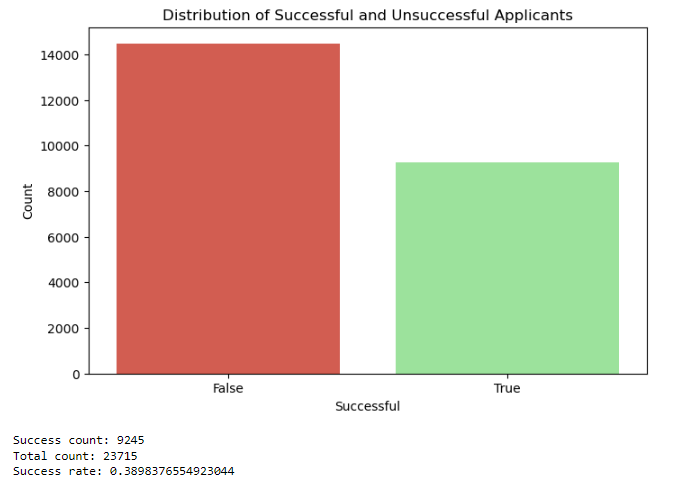
I began by starting with my Filtered\_df and merging this with my “Additional Country Info”. I included a new column to look at success percentage, before taking a look at some key summary statistics.



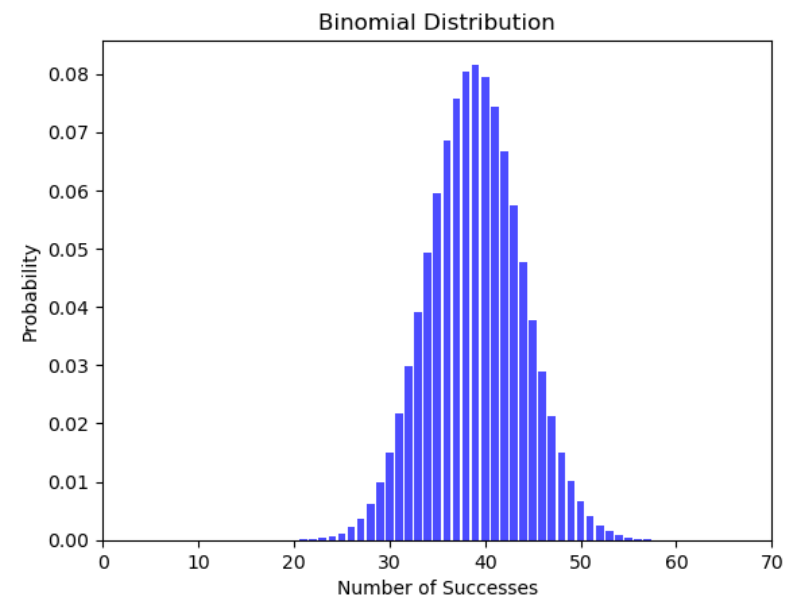
The ultimate aim from this section is to convert the data to have one individual per row. So I decided to do this by creating a function, this function code has been heavily annotated in the notebook to avoid any confusion.

I then performed a number of quick checks on the data to ensure the function had worked appropriately and the numbers matched that of the original data.

Whilst completing section 2.9, I realised that the data is in a format in where I could now apply a binomial distribution. I began by plotting the results along with some key statistics.



I then took a sample of 100 applicants to visualise the probability of successes within this sample.



As you can see the mean centres around 39 successful applicants which we would expect. Also, the distribution’s resemblance to that of a normal distribution is notable. We will discuss this in greater detail with a later example.

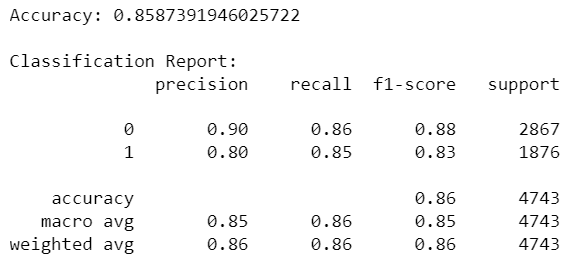
As the dataset includes a clear target variable (success), a supervised learning approach is most suitable. Supervised learning models are trained on historical data with known outcomes, allowing them to make predictions on new, unseen data. This is the perfect choice to apply to our current scenario.

Firstly, I encoded the dataset using a label encoder, this is a simple tool used to standardise the inputs into the model. It is not necessary to do this for a decision tree model but will be necessary for future models.

I then defined my target and independent variables, with the ‘Successful’ columns being my target variable and the remaining columns being my independent variables.

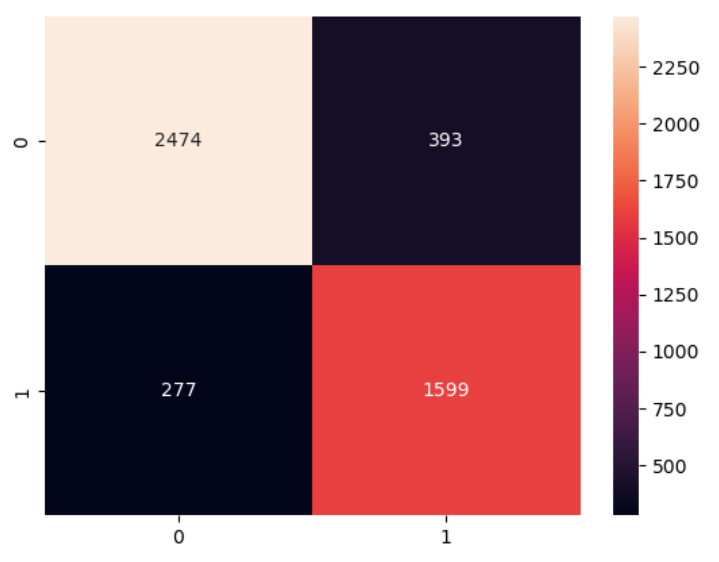
I then split my dataset into a training set and a testing set, with 80% of the model being used as the training set and 20% being used as a test set. I then initialised the decision tree model, defining the max depth and random state, I trained the model and made predictions using their respective functions.

I then summarised my model outputs with an accuracy statistic and classification report.

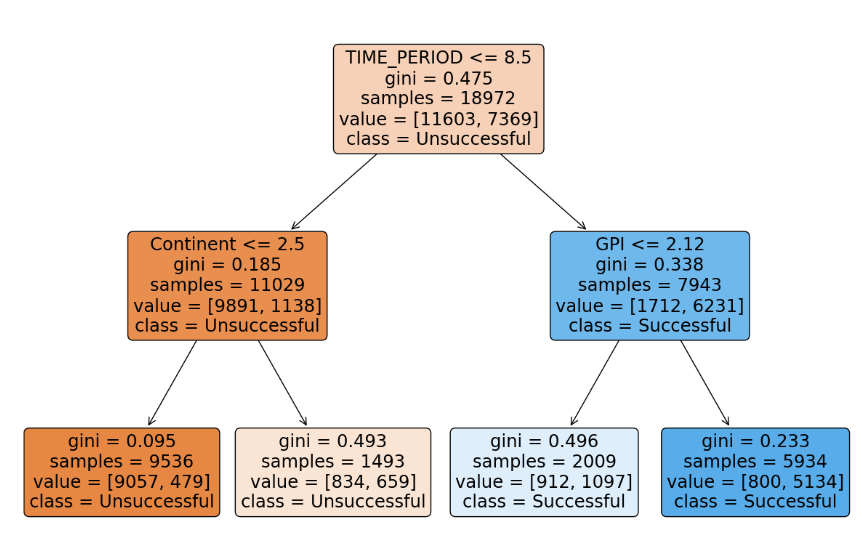


In short, the model appears to be performing well. The precision regarding class 0 tells us that out of all instances predicted as class 0, 90% were actually class 0. Similarly, out of all instances predicted as class 1, 80% were actually class 1. The model is precise but slightly less so for true positives.

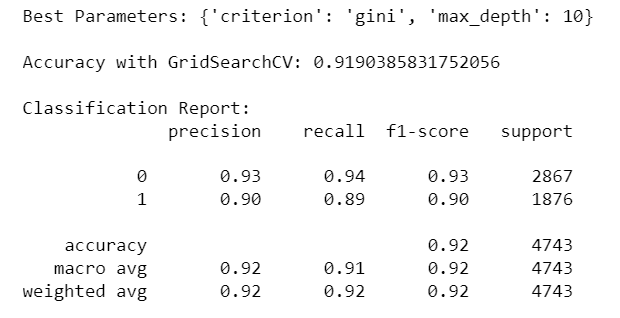
The overall model accuracy (86%) provides us with a general measure of model performance.

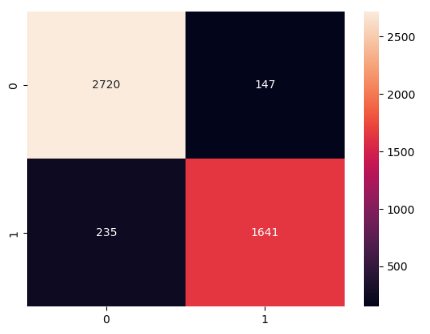


* 2474: This is the count of instances where the model correctly identified applicants who were accepted.
* 1599: This is the count of instances where the model correctly identified applicants who were rejected.
* 393: These are instances where the model predicted success, but the applicants were actually rejected.
* 277: These are instances where the model predicted rejection, but the applicants were actually successful.

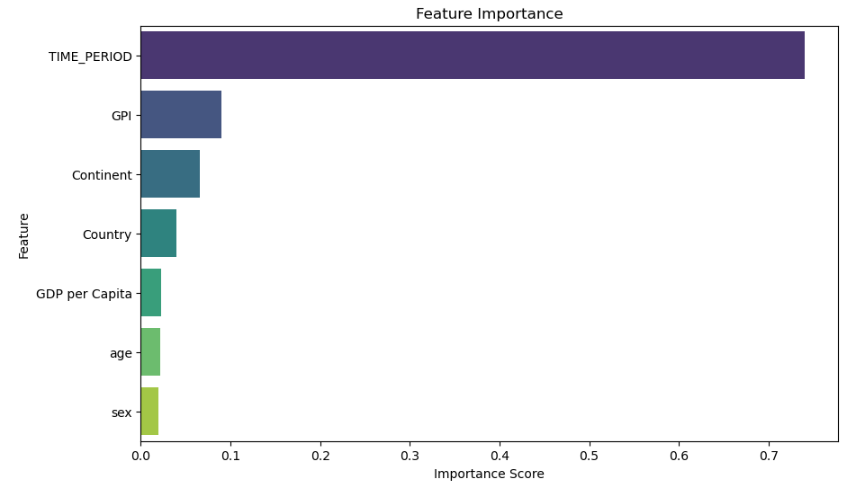


I then performed a search of the best hyperparameters using “GridSearchCV” function, which retrains the model using a range of different parameters to achieve the best results. The results of this model are displayed below:





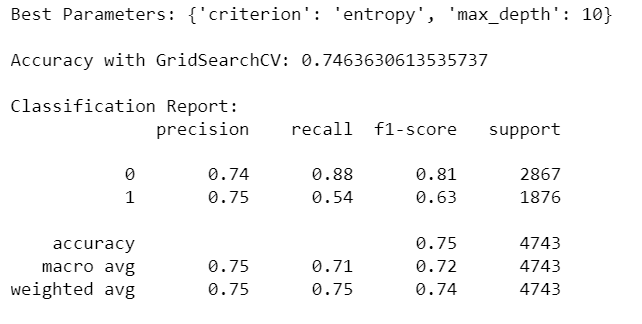
I also displayed the importance of each feature within the model below:



From this we can clearly see the significant influence that ‘TIME\_PERIOD’ has on the model outcomes.

I am conscious of the models practically, and so I decided experiment by removing the ‘TIME\_PERIOD’ column. I came to this realisation as the model could potentially be used by an aspiring asylum applicant to predict their future chances of success. In this scenario they obviously are unable to apply in prior years.

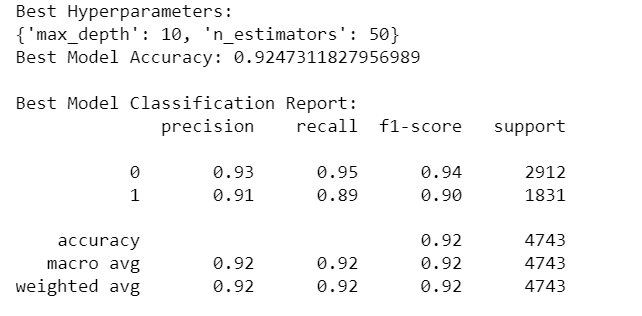
The results of this are printed below:

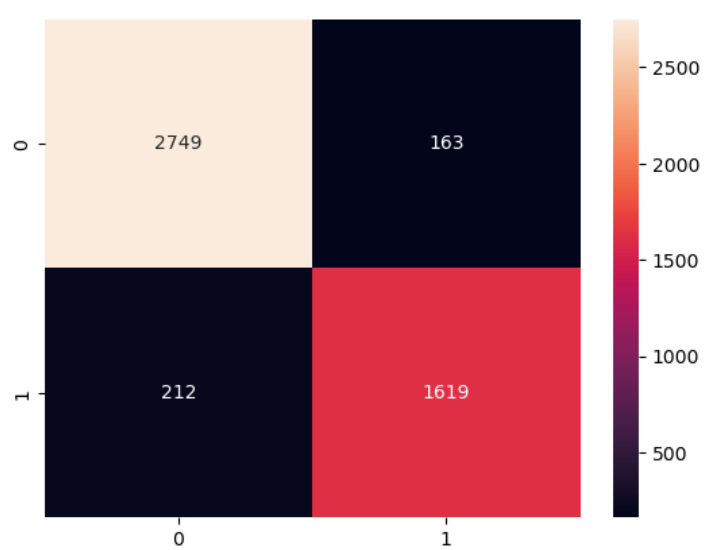


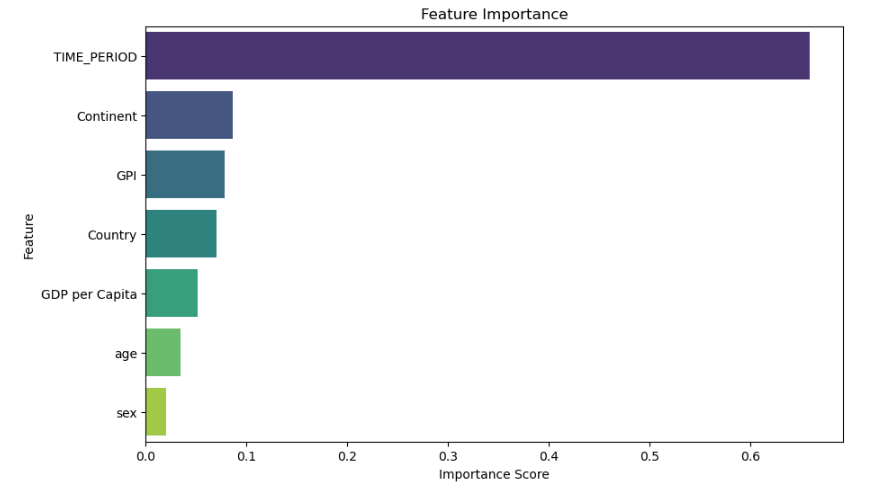
As you can see the model’s overall accuracy has reduced greatly from removing the ‘TIME\_PERIOD’ column, proving just how influential this factor is on applicant success. I will not speak about this in great detail here, but I have allowed the user to explore this option as the code can be easily adjusted.

After applying the decision tree model to the dataset I then shifted my focus on applying a random forest model.

Once again, I completed similar steps to how I created the decision tree model, splitting the dataset into a training and testing set, initialising the model, training the model, and making predictions. I then used the GridSearchCV function to tune my model to achieve the most optimal result (within the specified bounds). The results of this model are printed below:



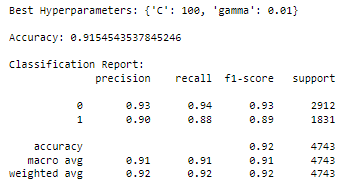


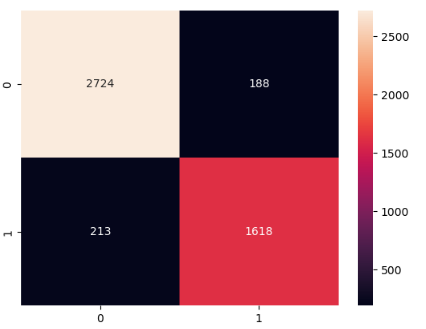


From these results, we can see that the model is performed similarly when compared to the decision tree model.

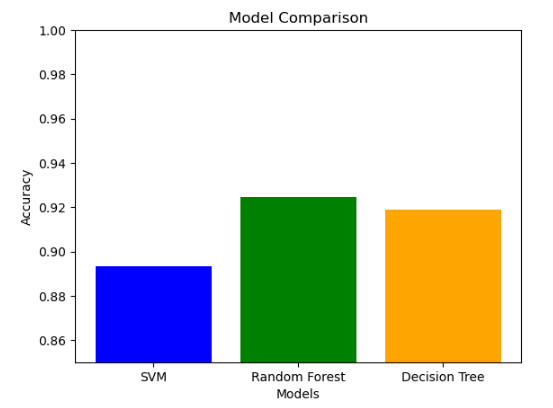
For my final model, I decided to implement a SVM (Support Vector Model) in my investigation. As SVM models are sensitive to the scale of the input I was required to standardise my inputs. To do this I used the ‘Standard Scalar’ approach.

I then repeated the exact same steps as outlined above for Decision Tree and Random Forrest models to achieve the following results for my SVM model.



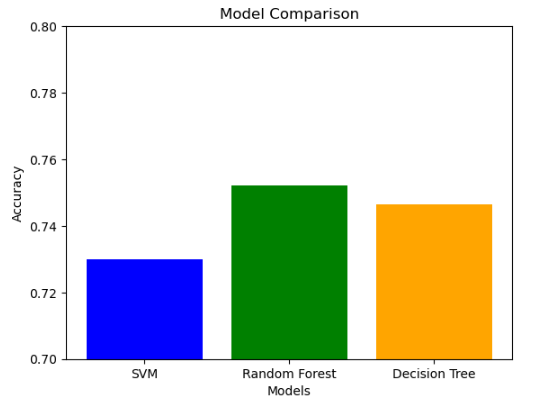


This section simply collates the accuracy score of the 3 models for direct comparison in the below bar chart:



From this we can see that the Random Forrest model is most accurate with an accuracy score of 92.5%. However, it is important to note that accuracy is not the sole measure of a good model.

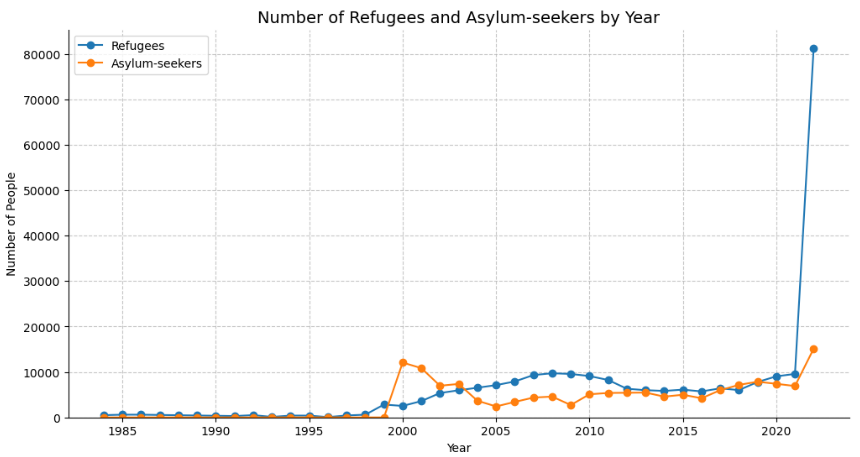
I have also included the results of model if were to remove the time period component. This significantly reduces the accuracy of the models. However, the Random Forrest model remains as the most accurate, as we might expect.



This notebook has been set up to explore the number of ‘displaced persons’ within the Irish population throughout the years and the demographics behind our current ‘displaced persons’ population.

In this section I read in the data and performed some basics trimming functions to only include rows and columns that are necessary in our investigation.

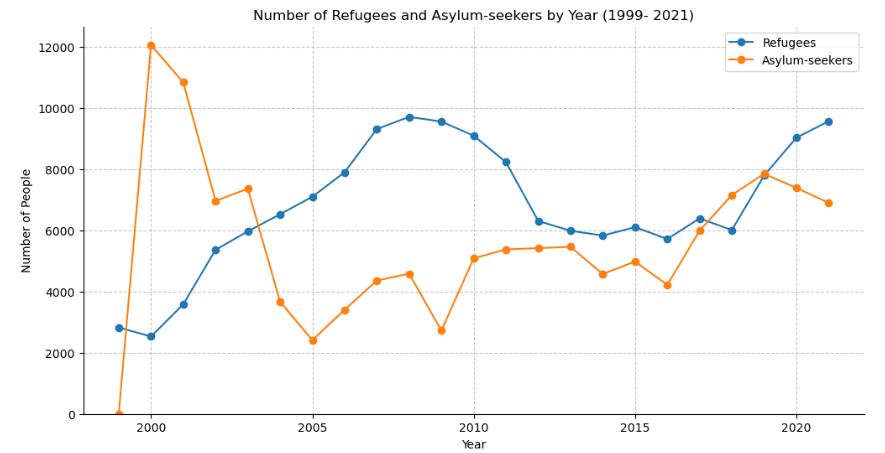
I wanted to visualise the dataset in terms of trends over the years. To do this I used the groupby function to group the dataset by each year. I then plotted the data using a line graph:



After plotting the data, two things became apparent. Firstly, the lack of ‘Displaced persons’ in the year prior to 1998 is significantly lower than what we would expect to see in more recent years, I concluded it is not appropriate to compare data from this time period with the data of more recent times.

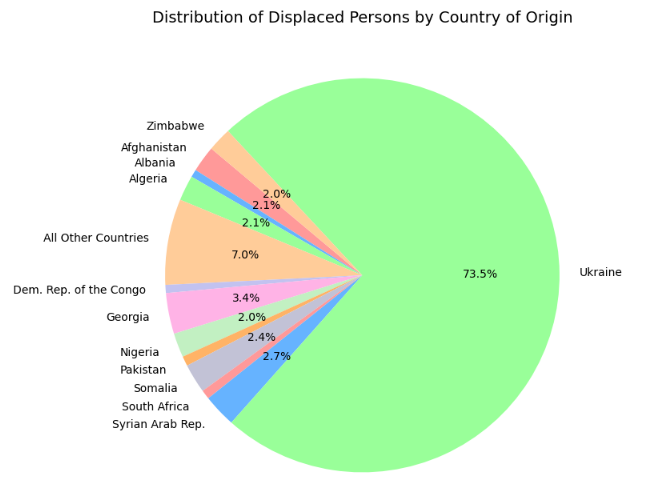
Secondly, the sheer number of Ukrainian refugees given refuge in 2022 is skewing the graph significantly and taking our attention away from the true underlying trends.

As a result of these phenomenon, I decided to replot my graph with a filtered dataset, excluding all years prior to 1998 and 2022.

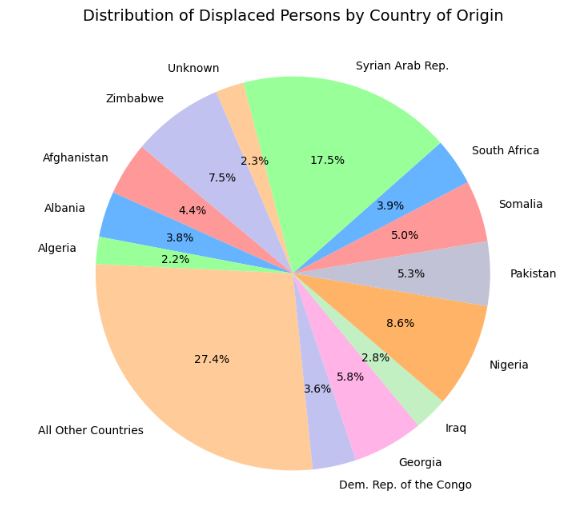


The graph above shows more coherent view of recent trends associated with the intake of displaced persons.

Next, I began investigating the demographics of the displaced persons in Ireland, mainly focusing on their country of origin. The data currently shows the number of displaced persons in Ireland at a given point in time, i.e. it is a stock dataset, so we cannot group this data across multiple years. Instead, we must select one specific year to analyse, so I began by taking a subset of the data for the year 2022.



From this pie chart it is apparent that the presence of the Ukrainian refugees is overwhelming our data for 2022. As a result of this, I decided to replot my chart for displaced persons in 2021.

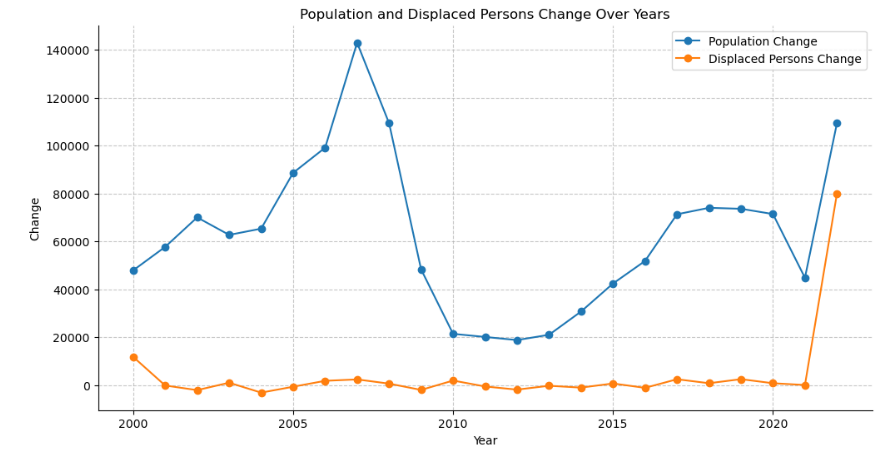


This section was created purely as I was curious to explore the trends of incoming displaced persons from particular areas of conflict. I will not go into great detail here but have left this section in my Jupyter notebook for the readers curiosity.

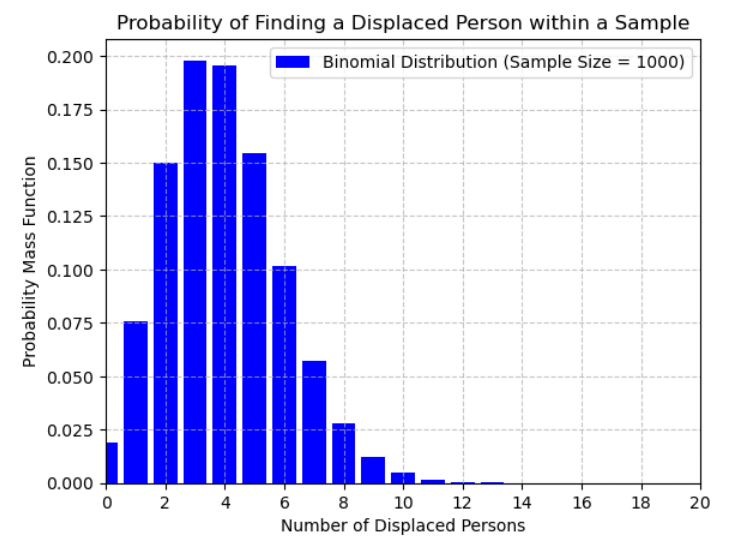
### 

This section focuses on the number of Displaced Persons per Capita. I firstly had to read in some population data (“Dataset 6) PEA01.csv”) to supplement my pre-existing dataset. I edited this data before merging it with the displaced persons dataset.

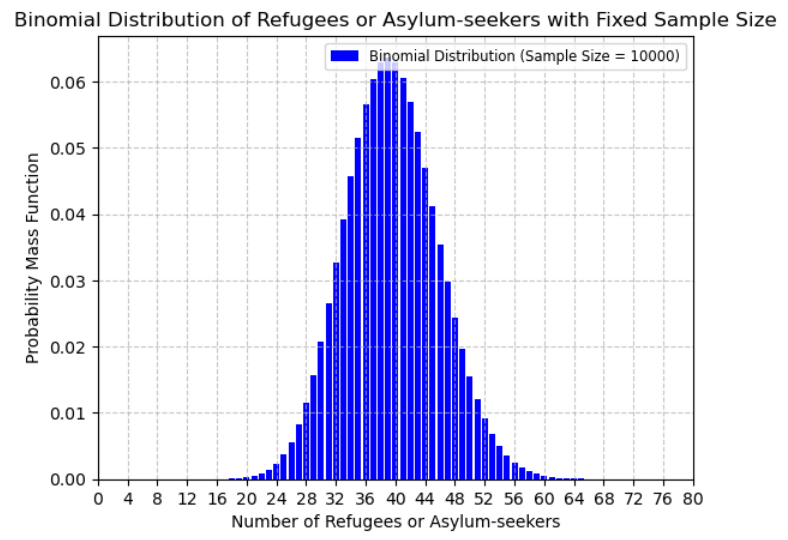
I then created 3 additional columns, outlining the population change per year, displaced persons population change per year and the displaced persons per capita. I then plotted a line chart to visualise these columns.

In this section I applied a binomial distribution to our dataset by creating a hypothetical scenario. The scenario I have in mind is that if we took a random sample of the Irish population in any given year (from 2010 onwards), how many of these individuals would be displaced persons?

So I began by trimming my dataset and then calculating the mean of the ‘Displaced Persons per Capita’ (0.00395). I then applied this to a sample size of 1000 individuals and plotted my result in a histogram.



I then performed my calculations again with a larger sample size of 10000.



We can see from the above histograms that as we increase sample size, the probability distribution of the larger sample significantly resembles that of a normal distribution. This phenomenon is known as the Central Limit Theorem (CLT).

The CLT dictates that, as the sample size increases, the shape of the binomial distribution becomes more symmetric and bell-shaped, resembling a normal distribution. This happens because, with a larger sample size, there are more possible combinations of outcomes, and the distribution becomes smoother.

In conclusion, this investigation has been extremely valuable to me personally and has given me the freedom to explore and analyse avenues that would not have been possible without the independence that was granted to me by my lecturers.

Throughout this report I have often had to refrain myself from delving too deep into certain topics as I understand that the examination criterion is a priority, and it is necessary to devote my attention to satisfying each individual point.

Ultimately, I hope that my report has given the reader some interesting insights into these topics whilst remaining specific enough to satisfy each criterion that was required of me.