CCT College Dublin

Assessment Cover Page

Module Title:	HDip in Science in Data Analytics for Business/ AI
Assessment Title:	Individual / Practical CA2 Project
Lecturer Name:	James Garza (james@cct.ie)
Student Full Name:	Oluwatimileyin Oladipo Ayomide
Student Number:	2023383
Assessment Due Date:	17 th December 2023
Date of Submission:	16th December 2023

Oladipo Ayomide. O

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Population in Ireland: Where are we heading to?

By

Oluwatimileyin Oladipo Ayomide (Student ID: 2023383)

Higher Diploma in Science in Data Analytics for Business

Strategic Thinking

James Garza

CCT College

Dublin, Ireland

Table of Contents

GitHub link: https://github.com/Ayomide-ola/Stategic_thinking-CA2

Introduction

Data are plain facts that have been gathered within a certain context. When data is summarised, it gives us information about the context within which it was collected, and with the use of analytical tools we can produce evidence from the information we have. The evidence we have today can be used to train machine learning model to predict future possibilities.

The first task in this project required me and my team mate to produce a capstone project proposal on an interesting and relevant subject. As a team we decided to research 'Population' in Ireland. In this aspect of the task, I would be conducting individual research on the subject using CRISP- DM project management framework. I intend to report on my findings through in-depth analysis of population trend in Ireland as many of today's largest economies are reportedly experiencing decline in population growth rate (World101, 2022).

(Honohan, 2021) reported Ireland as one of the largest GDP in Europe, as a result, it's economic stability is imperative, if the European Union (EU) will retain its economic viability on the world stage.

The Organisation for Economic Co-operation and Development (OECD, 2023) and the United Nations (UN, 2023) reported that a country requires a fertility rate of 2.1 children per woman to ensure a stable population. For several decades, the EU are reportedly below this fertility rate and for most case, this decline has been below replacement level (UN, 2023). Many factors influence a countries population, with birth rate, death rate and migration been key determining factors of population growth or decline (World101, 2022).

Through exploratory data analysis, I aim to find if there is any correlation between this 3 major factors and population decline / growth in Ireland. This project also aim to predict based on the decline or growth trend found in this analysis when the birth rate in Ireland will not compensate for the rate of deaths.

Objective

- Observe through exploratory analysis of data, the population growth/decline trend in Ireland from 1960 to 2021.
- Visualise correlation between the number of births, deaths, and marriages, been few of many determining factors of population growth/decline and the population growth/decline trend in Ireland.
- Visualize current trend between birth and death rate/1000 individuals in Ireland.

- Find an optimal machine learning model that predicts Ireland's future population estimate.
- Evaluate performance of model using appropriate performance evaluation metrics depending on the model used.
- Evaluate results and draw conclusions concerning Ireland's population stability based on evidence produced.

Could-Haves objective:

- Plot a predictive graph of birth and death rate of Individuals in Ireland.

Using CRISP-DM Methodology (Matsumoto and Carrinho, 2023)

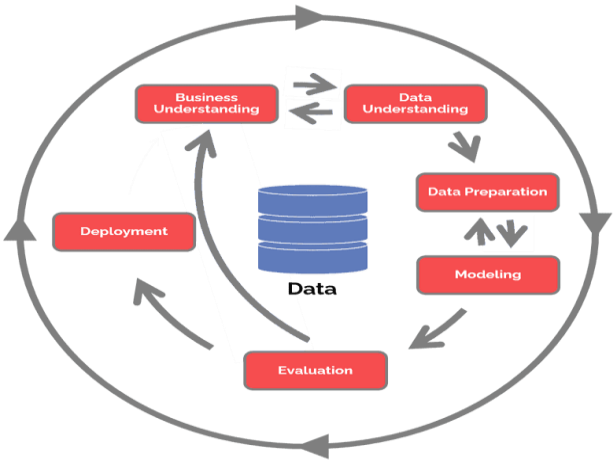


Figure 1: CRISP-DM Life Cycle

Problem Understanding

Human capital is an important bed rock of any stable economy, a large population means more workers and customers which ultimately boosts a country's GDP (World101, 2022). As stated by (Wilmoth, Menozzi and Bassarsky, 2022) world population is expected to peak around 2100 at a level of almost 11 billion. (OECD and UN, 2023) both reported a country requires a birth rate of 2.1 children per woman to maintain a healthy population growth. However as shown in fig.2, Ireland has recorded a progressive decline in its birth rate since 1970 but were still within a healthy birth rate.

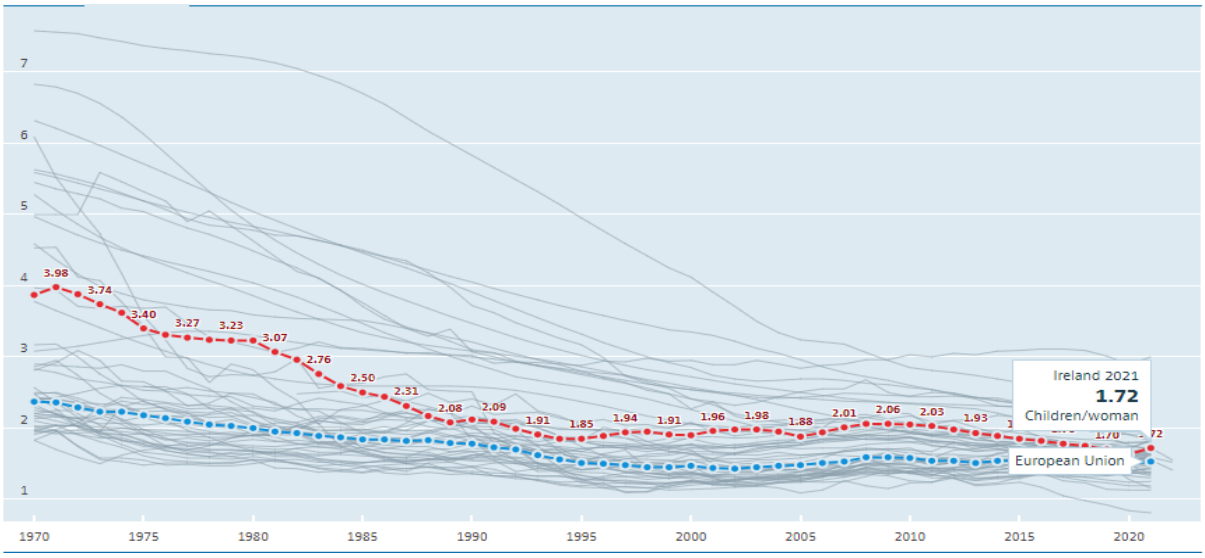


Figure 2: Fertility rate, total children per woman 1970-2022 (OECD 2023)

Since 1991, Ireland has declined below a healthy population birth rate, and trend suggests that this decline may continue unless objective measures are implemented to prevent it. We believe Ireland is facing a significant challenge in its birth rate which is one of major drivers of population growth or decline. This challenge based on scientific research could potentially impact a country's (Ireland's) economic and societal stability.

Scope/ Data Understanding

To educate myself on domain knowledge from experts, I reached out to the CSO office, my questions and their responses are detailed in *fig 3, fig 4, fig5 & fig 6* below but in summary, I sort their advice on the key data required to conduct this research. Considering time constraint, I have drawn a sub-scope from their advice to complete in this project phase.

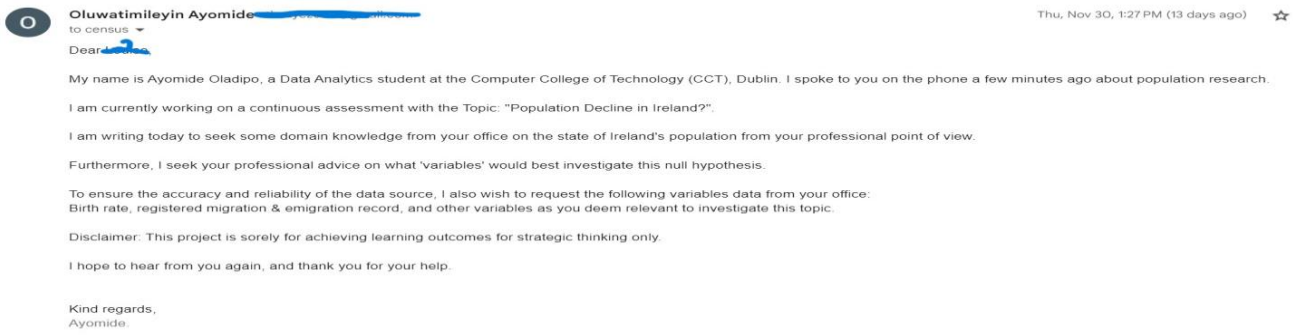
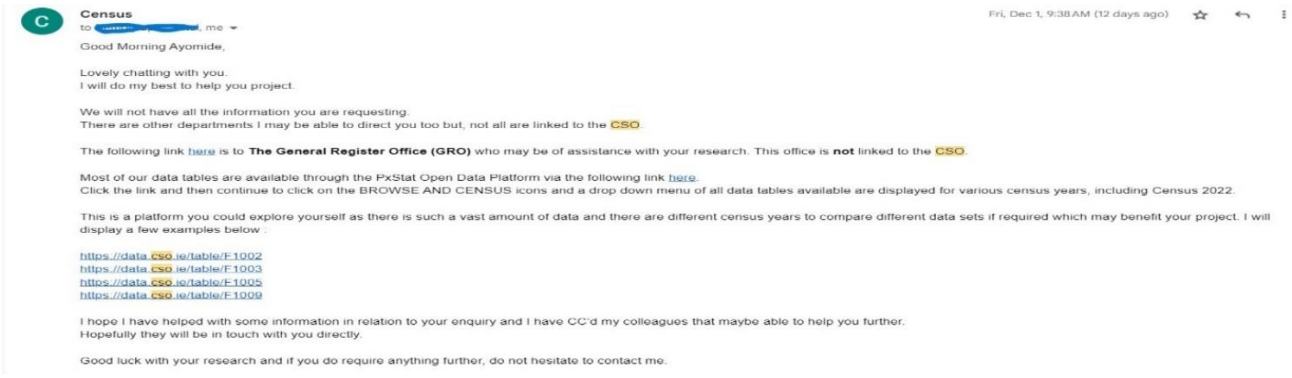


Figure 3:Investigative question from me to CSO expacts

The dataset mined from CSO direct links, contains data on number of Birth, death, and (birth and death rate)/1000 individuals. In other to predict population estimate, I need actual data on

population estimate/ year to train my model. I got population estimate dataset from UN open-source dataset available here.

I will explore and analyse the current trend of population estimate of Ireland by UN from 1960 to 2021. I will explore the correlation between the number of births, deaths, and marriages in Ireland to provide a better understanding of its population trend. I aim to predict when the number of deaths will be greater than the number of births per year.



Census Fri, Dec 1, 9:38 AM (12 days ago)

Good Morning Ayomide,

Lovely chatting with you.
I will do my best to help you project.

We will not have all the information you are requesting.
There are other departments I may be able to direct you too but, not all are linked to the CSO.

The following link here is to **The General Register Office (GRO)** who may be of assistance with your research. This office is **not** linked to the CSO.

Most of our data tables are available through the PxStat Open Data Platform via the following link here.
Click the link and then continue to click on the BROWSE AND CENSUS icons and a drop down menu of all data tables available are displayed for various census years, including Census 2022.

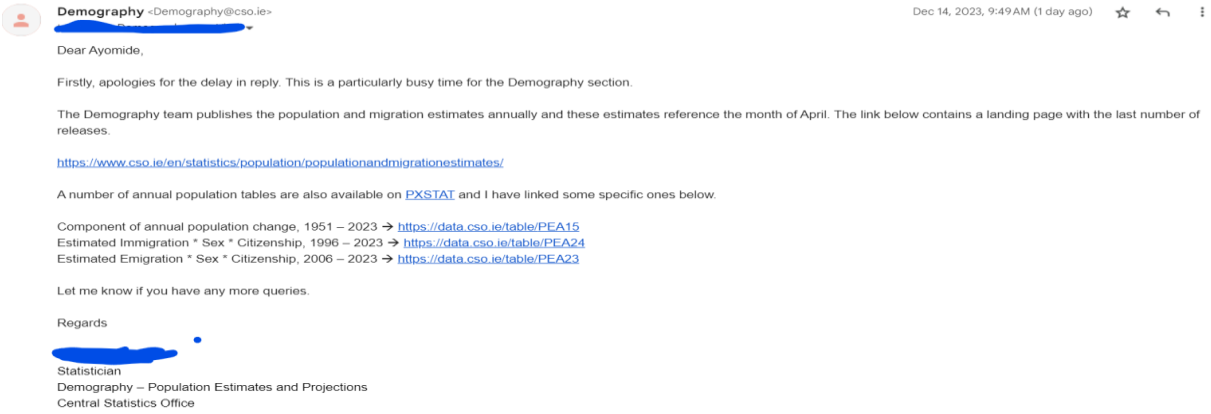
This is a platform you could explore yourself as there is such a vast amount of data and there are different census years to compare different data sets if required which may benefit your project. I will display a few examples below :

https://data.cso.ie/table/F1002
https://data.cso.ie/table/F1003
https://data.cso.ie/table/F1005
https://data.cso.ie/table/F1009

I hope I have helped with some information in relation to your enquiry and I have CC'd my colleagues that maybe able to help you further.
Hopefully they will be in touch with you directly.

Good luck with your research and if you do require anything further, do not hesitate to contact me.

Figure 4: Response from the statistics office, population department



Demography <Demography@cso.ie> Dec 14, 2023, 9:49 AM (1 day ago)

Dear Ayomide,

Firstly, apologies for the delay in reply. This is a particularly busy time for the Demography section.

The Demography team publishes the population and migration estimates annually and these estimates reference the month of April. The link below contains a landing page with the last number of releases.

https://www.cso.ie/en/statistics/population/populationandmigrationestimates/

A number of annual population tables are also available on PXSTAT and I have linked some specific ones below.

Component of annual population change, 1951 – 2023 → https://data.cso.ie/table/PEA15
Estimated Immigration * Sex * Citizenship, 1996 – 2023 → https://data.cso.ie/table/PEA24
Estimated Emigration * Sex * Citizenship, 2006 – 2023 → https://data.cso.ie/table/PEA23

Let me know if you have any more queries.

Regards

Statistician
Demography – Population Estimates and Projections
Central Statistics Office

Figure 5: Demography advice, and migration information

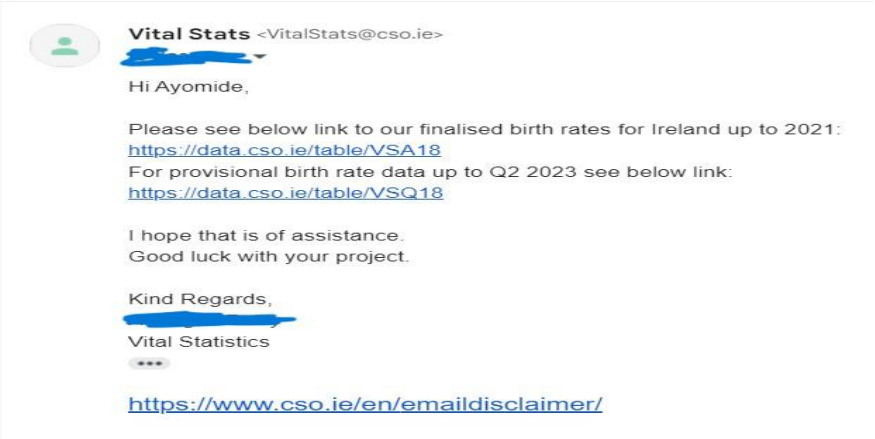


Figure 6:Birthrate information from Vital stat department in CSO

Ethical considerations

The socioeconomic importance of this project topic was considered, as this can influence public action or inactions, for example people deciding to have more children because of fear of decline, but we established after consideration that we do not seek to make any conclusions on the population stability of Ireland but just to analyse and report findings for capstone research purpose only.

Also, to prevent breach of anonymity of population from which the data was gathered the CSO anonymised all the dataset.

CSO directly provided the data I used for this project and CSO confirmed on phone there is no permission required to use data available on open source as shown in *fig.7* and here.

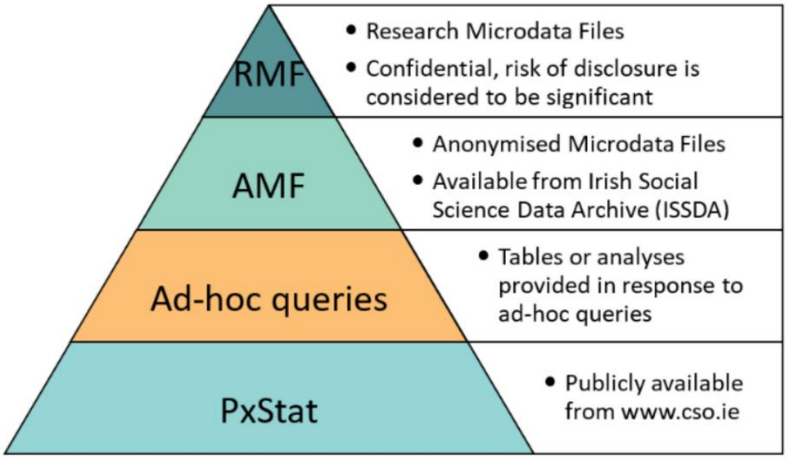


Figure 7: CSO policy on data usage

Data Preparation

As earlier stated, 2 dataset is required for the projects.

Birth, death, marriage and rates dataset

To be able to manipulate data using python in Jupyternotebook interface, I imported pandas, NumPy and visualisation libraries and load csv file into dataframe(df) using pandas and 'read_csv' method as shown in fig 8.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns

In [2]: pop= pd.read_csv("Total Births, Deaths and Marriages Registered.csv")
        pop.head()
```

Figure 8: Importing Libraries.

Understanding data

Initial exploration of df size shows there are 4554 rows and 5 columns in the df. No null or duplicated values where present in initial df. The df contains record of birth death and rates and other labels for 63 years (1960-2022). The 'statistical' label column has 20 unique labels in it, and domain knowledge enquiry and project scope definition enhanced the decision to extract records that are relevant to this project phase only.

Extracting needed 'labels' and their corresponding value and years.

Because the data where not stored in vertical column format, they were store in horizontal row format, the traditional column dropping was not usable. To extract, I stored the values of the labels I want into a list using their unique labels as filter and assigned the list to proposed column name as shown in fig 9, same code syntax was repeated to extract the values of birth and marriage and their rates.

```
pop['Reg_Death'] = pop['VALUE'][pop['Statistic Label']=='Deaths Registered']
pop['Rate_Death/1000'] = pop['VALUE'][pop['Statistic Label']=='Death Rate Registered per 1000 Estimated Population']
pop['Reg_inf_Death/1000'] = pop['VALUE'][pop['Statistic Label']=='Deaths of Infants under 1 Year Registered per 1000 Births']
pop['Reg_Birth'] = pop['VALUE'][pop['Statistic Label']=='Births Registered']
pop['B_outsideMarriage'] = pop['VALUE'][pop['Statistic Label']=='Percentage of births registered outside marriage']
pop['Mothers_age'] = pop['VALUE'][pop['Statistic Label']== 'Average age of mothers giving birth']
pop['CV_P_M'] = pop['VALUE'][pop['Statistic Label']== 'Civil partnership marriage rate']
pop['Rate_marriage/1000'] = pop['VALUE'][pop['Statistic Label']== 'Marriage Rate Registered per 1000 Estimated Population']
pop.head()
```

Figure 9: Extracting values by labels

This operation indeed extracted the values relevant to project objective, however a new problem was created. Now there are 4301 NAN in each of the newly created columns. This Is because pandas will automatically fill empty rolls with NAN, and lots of NAN have been created after extraction of unique rows.

Data Cleaning

Fixing Null-values:

NAN can be fixed using two broad methods, imputation and dropping. Because the NAN in df were artificially created, I decided to drop them all using 'dropna' method. This method worked, but created another problem, output fig 10 shows that all the rows in all columns were dropped.

```
# Now droping the the NaN
pop_dro= pop.dropna()
pop_dro.head()
```

Quarter	UNIT	Married	Reg_Death	Rate_Death/1000	Reg_inf_Death/1000	Reg_Birth	Rate_marriage/1000

Figure 10: Empty df after dropping Nan.

This is because dropna will automatically drop all the rows where Nan is found, and considering the proportion of Nan created the output in fig.10 was justifiable. I then decided to perform EDA while leaving the Nan in the data because they are too much to fill without introducing massive bias.

After several experimentation and rationale detailed in the Jupyternotebook, I was able to fix the null vales by individually storing the labels I intend to use and their values in a list, I then converted all the individual list to a new dataframe. The new dataframe has 234 rows and 8 columns.

Descriptive analysis:

Statistical analysis shows overall majority the variables a slightly skewed except for [birth outside marriage, mothers age, rate of marriage]. All variables where right skewed as the mean was slightly more than the median. All variables have very large standard deviation, which suggests a tightly grouped variance relative to the mean. The mean shifting to the right suggest there may be outlier on the right-side of the bell curve.

Visualising Descriptive analysis:

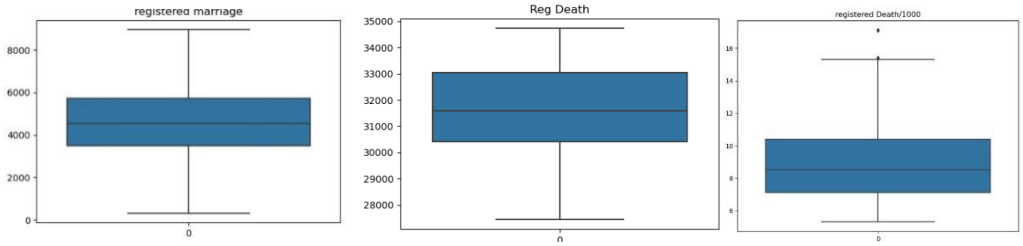


Figure 11: Registered marriage

Figure 12: registered death

Figure 13: Rate of death

The registered marriage showed a normal distribution, statistics table suggest slight skewness, the death variable is also slightly skewed but there are no outliers in them as shown in Fig 11 and 12. Rate of death however has outliers and it is highly left skewed.

Year column:

The observations were recorded quarterly with their corresponding values, this means each year had four independent values recorded, however the population dataset is yearly record, in other to get the data into the same shape to allow merging, I Sumed up the 4 values in each quarter year, and converted them to one value per year. The ‘Q1’ in each year was removed and a new yearly column was created using ‘range (1960-2023)’ which generated a list of years from 1960- 2022, and I joined this list to main ’df’ and the old year column was then dropped.

Population Dataset

The dataset contained population estimate of all the countries in the world, using ‘IRL’ as key, I was able to extract the population estimate of Ireland from the main loaded world population dataframe. This contain population estimate of Ireland from 1960-2021, this led to the decision to drop the 2022 row from the ‘death, birth’ df.

After joining, now I have cleaned data with 9rows and 62columns shape, with no null or duplicated values. All floats were converted to integers and the ‘rate’ columns were left as floats.

Data Visualisation:

Using pair plot grid to visualise the entire dataframe to see correlation, I saw a positive correlation between the population estimate and year, correlation coefficient of both variables also returned ≈1. As year increases the population estimate increased too fig 12.

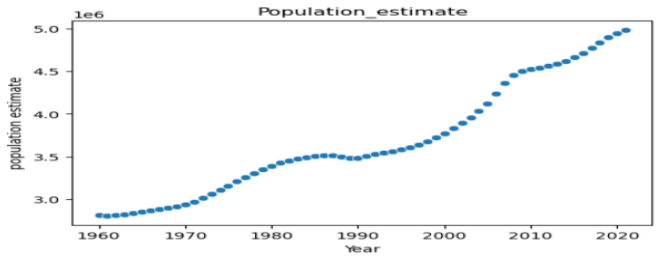


Figure 12: Visual relationship between the year and population estimate.

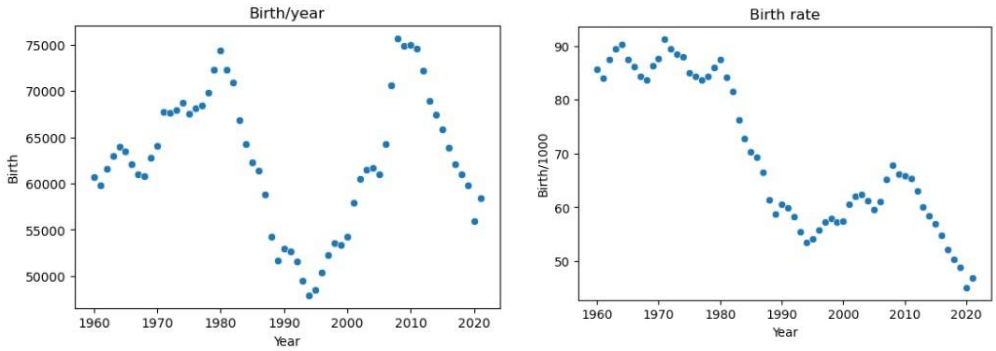


Figure 13: Birth rate/ 1000 people and birth rate/year

As shown in the plot, the number of births progressively increased every year from 1960, and reach its peak ≈1980, with over 70,000 births recorded in Ireland in ≈1980. From year 1980, birth in Ireland fall from ≈1980 - 1994 with the lowest birth recorded at under 50000 since 1960. ≈1995, births increased progressively to record high in ≈2010 higher than 1980, a record birth of over 75000 in 2010. From ≈2011, birth/year progressively declined.

Visual exploration of the birth rate/1000 individuals in Ireland suggest a record high between year 1960 - ≈1985 with over 90% birth rate, birth rate showed a downward surge from its highest in ≈1985 to less than 60% in ≈1995. From 1995 to ≈2010 there was a progressive increase in the birthrate from <60% to over 65% less than 70%. From 2010, Ireland's birth rate had been decreasing and data show it was at its lowest around 2019, and a little increase in

birth rate from 2019 – 2021, this result is consistent with report of OECD detailed in introduction.

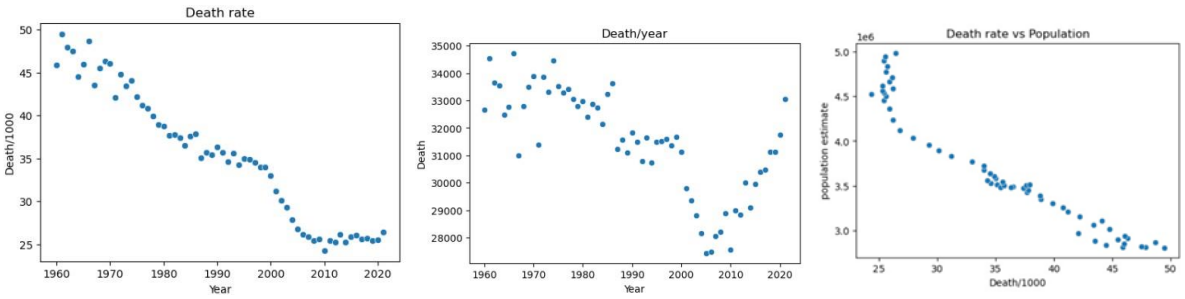


Figure 14: (Death/ year and death rate/ year) trend and Impact of death rate on population

According to data, in ≈1961 Ireland recorded the highest death registered and rate, over 34,000 actual deaths and ≈48% of 1000 individuals. Death rate shown a downward surge from 1960 to its lowest in 2010, with the death rate less than 25% of 1000 individuals. Acutual number of recorded deaths continually increased from 2010 significantly, but more significantly, the recorded death in Ireland exponentially increased between 2019 ≈ to 2021 compared to other years. Population was estimated at its highest when the death was low, it decreased as the rate of death increased.

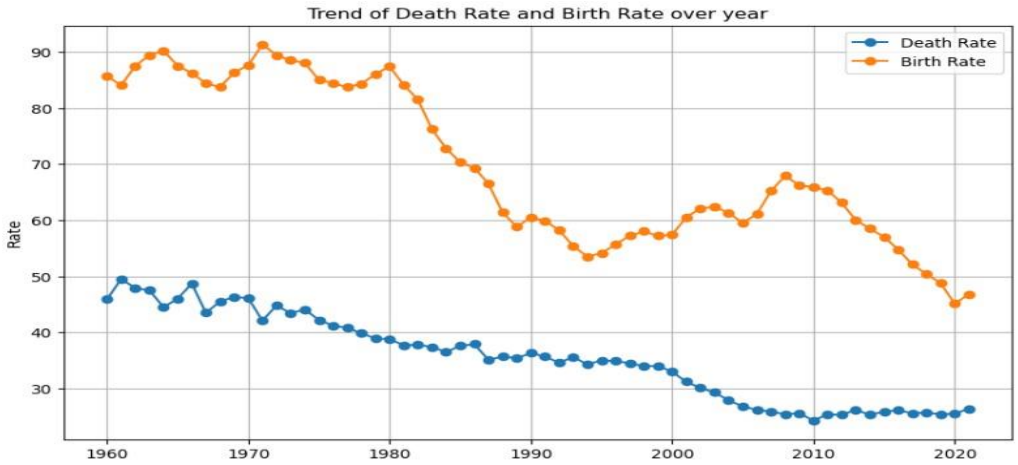


Figure 15: Rate of Death and birth trend

Overall, since 1960 to 2021, the rate of birth has been significantly higher than the death rate.

Modelling

I have a regression problem, and I am trying to predict continuous data which is why I chose Random forests regressor. They are considered powerful and robust because work well with high-dimensional data, missing values, and outliers (BHAT, 2023).

Population estimate was defined as dependent variable and remaining variables are defined as X (independent variables.

Model Evaluation

Metrics:

To evaluate the performance of the model, I used [Mean absolute error, Mean square error, R2, root square error] performance metrics.

MAE: (Chugh, 2020) “it is the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset after prediction. Model 1 had a lower MAE compared to model 2, it performed better at predicting y given x.

MSE: MSE is average of the squared difference between the original and predicted values in the dataset. It quantifies the variance of the residuals after prediction (Chugh, 2020). model 1 has better precision compared to model 2.

RMSE: RMSE is the square root of MSE, it measures the standard deviation of the error. Model 1 has better precision at prediction of estimated population because it shows lesser RMSE.

(R2): R2 represents the proportion of the variance in the dependent variable (Chugh, 2020). R2 in model 1 is slightly larger than that of model 2, this means it explains the target variable variance better than the model 2.

Conclusion

In conclusion, using the randomforestregressor without hyperameters, at 80% training and 20% testing, the model performed better at predicting estimated population of Ireland given 'X' variable. The results of the performance metrics show model 1 performed better, and based on trend from this data, there are no prove to support that “Ireland is in a population decline”.

Nevertheless, the subject of population Is broad, further research is imperative for other population determining factor for example migration, emigration, proportionality between age groups and gender. I did not consider these factors in this phase on my analysis, because of time and resource constraint.

References

BHAT, S. (2023). *A Comprehensive Guide to Random Forest Regression*. [online] Medium. Available at: https://medium.com/@bhatshrinath41/a-comprehensive-guide-to-random-forest-regression-43da559342bf [Accessed 14 Dec. 2023].

Central Statistics Office (n.d.). *Copyright Policy - CSO - Central Statistics Office*. [online] www.cso.ie. Available at: https://www.cso.ie/en/aboutus/whoweare/copyrightpolicy/ [Accessed 27 Oct. 2023].

Central Statistics Office, Ireland (2023). *Total Births, Deaths and Marriages Registered*. [online] Data.cso.ie. Available at: https://data.cso.ie/table/VSQ04 [Accessed 12 Oct. 2023].

Chugh, A. (2020). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?* [online] Medium. Available at: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e [Accessed 15 Dec. 2023].

Honohan, P. (2021). Is Ireland really the most prosperous country in Europe? *Central Bank of Ireland*, [online] 2021(1). Available at: https://www.centralbank.ie/docs/default-source/publications/economic-letters/vol-2021-no-1-is-ireland-really-the-most-prosperous-country-in-europe.pdf [Accessed 24 Oct. 2023].

HYDE (2017). *Population*. [online] Our World in Data. Available at: https://ourworldindata.org/grapher/population-long-run-with-projections [Accessed 13 Dec. 2023].

Matsumoto, R. and Carrinho, S. (2023). *Predicting the effects of Climate Change on Irish Agriculture Predicting the effects of Climate Change on Irish Agriculture*. [online] Available at: https://arc.cct.ie/cgi/viewcontent.cgi?article=1039&context=ict. [Accessed 13 Oct. 2023].

OECD (2023). *Demography - fertility rates - OECD data*. [online] OECD. Available at: https://data.oecd.org/pop/fertility-rates.htm [Accessed 13 Oct. 2023].

Organisation for Economic Co-operation and Development (2023). *Fertility rates - OECD data*. [online] OECD. Available at: https://data.oecd.org/pop/fertility-rates.htm [Accessed 13 Oct. 2023].

United Nations (2023). *Population*. [online] United Nations. Available at: https://www.un.org/en/global-issues/population [Accessed 24 Oct. 2023].

Wilmoth, J., Menozzi, C. and Bassarsky, L. (2022). *Why population growth matters for sustainable development POLICY BRIEF NO 130 Key messages*. [online] Available at: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/undesa_pd_2022_policy_brief_population_growth.pdf [Accessed 27 Oct. 2023].

World101 from The Council on Foreign Relations (2022). *Global Population Growth Is Slowing Down*. [online] World101 from the Council on Foreign Relations. Available at: https://world101.cfr.org/global-era-issues/development/global-population-growth-slowing-down#:~:text=Studies%20differ%20about%20the%20exact [Accessed 24 Oct. 2023].

Word count: 2110