Population Growth Prediction model Using Demographic Measures

By Leon Pereira

1. Abstract :

Nowadays population and factors affecting population growth are very much in discussion. Some of the main reasons for this are the national bodies and government want to make good use of their resources and also want to plan the future infrastructure of the nation for which they need to pay more attention to the population change. This research project dives deep into the various factors and variables that affect population growth and makes use of them to develop potential solutions and machine learning models to forecast the population growth trends. The research explores various machine learning algorithms and techniques to measure accurate population growth using demographic variables which are more reliable and numeric based variables to make predictions that will help in environmental sustainability ,social services , urban development and resources allocation of the region. The prime technique involves creating, improving and applying machine learning models to determine the key demographic variables and gauge how they affect population increase. Once these models are trained ,it will help reduce the need for substantial historical data and help produce results using demographic indicators. The study advances our understanding of the application of machine learning in demographic research and provides insightful information on the creation of population growth prediction models.

2. Acknowledgements :

I would start by sincerely thanking my project supervisor Ms. Alaa Mohasseb who has helped me with their knowledge ,understanding and experience to work on this project and significantly enriching my graduation experience. Your guidance has been really helpful for me throughout this project and has helped me a lot in achieving my desired aims for the project.Additionally, I want to thank you for your insightful criticism, thought-provoking enquiries, and persistent support. I would also like to thank all the professors who have given me help throughout the academic year of studying at University of Portsmouth. To the MSc Data Analytics Program at University of Portsmouth , I would like to thank you for providing me with all the necessary resources , guidance , courses and proper environment to complete this project.

Table of Contents

3. Introduction

Regional Planning is one of the most important challenges faced by nations and their government bodies and it is heavily focused on the issue of human population growth. It involves several departments that act as a result of this regional planning like urban development and social services,resource management and environmental sustainability. A good and accurate forecasting of population rise can benefit policy makers, stakeholders and other planners to make developmental and informed decisions on factors like economic growth, social deliverings and infrastructural planning in a timely and efficient manner. However, when it comes to population growth rate estimates it is not feasible to depend on the historical data and traditional approaches.As a result, the goal of this research is to find a solution to this issue and offer substitute strategies for estimating the rate of increase in the human population without relying on past data.

Machine learning (ML) is used in various fields of study including demographic research. This provides us with a good understanding and a potential option for our research. We will use forecasting techniques , visualisations , correlations and regressions on complicated datasets and extract various patterns relevant to our study. The correlations of demographic characteristics and variability in the population growth rate (pgr) over a long period of time with a large dataset will be worked on in this project. Once determining the factors affecting the population growth by these machine learning techniques we can forecast future population growth rate with specific demographic variables and no longer rely on the historical data.

4. Problem Statement

Forecasting the growth of the human population is essential to efficient regional planning. However, in contexts where data is scarce, obtaining the demographic information required to predict population growth rates can be difficult. It is therefore extremely valuable to design a method that can forecast population growth independent of past data.The aim of this study is to forecast the population growth rate, or "pgr," in a given region using a variety of machine learning (ML) techniques. The study attempts to determine whether demographic indicators are correlated with the rate of population expansion by looking at a variety of data that were gathered over a long period of time. The objective is to use machine learning techniques to model these correlations, and then use the trained models to anticipate population increase based on the given demographic factors.

5. Literature Review

5.1 Global Population Projections: Insights into Future Trends and Demographic Structures

Human population growth and the factors that impact this factor have been predicted in many different ways through science and statistics.Numerous philosophers and intellectuals have made contributions to the historical discourse on population dynamics, but Thomas Malthus introduced quantitative analysis in the late 1700s. Malthus' seminal work made possible later discussions on striking a balance between population growth and resource availability.Techniques for examining the population increase of a mixed-gender population were covered in this journal by Goodman [1953] and Keyfitz and Murphy [1967].

The factors considered in this literature review for population projects include age and sex structure, fertility rates, death rates, and migration rates.The population is split up into groups based on age and sex.

Let's start understanding the mathematical equations and how it works taking into account different factors and variables.

Consider *P(a,t)* is the population of a region at age *a* at time *t* , then:

$$P\ (a + t\ ,\ t + 1) = P\ (a,t) - D\ (a,t) \quad For\ a = 0,1,2....$$

Where *D (a,t)* represents the number of deaths at time *t* of age *a*.

Now from this equation we say that the population at the next age group *a+1* in the next time period *t+1*.

For instance, if t is 2024 and age is 20 years old, then population that we will be looking at is 21 years old in 2025 and is represented by *P(21,2025) = P(21, 2025)*.

The likelihood of passing away at age a in the given time t is represented by the mortality rate. We denote it as *M (a,t)* and the overall number of deaths is *D(t)*,

$$D(t) = \sum_{a} P(a,t)M(a,t)$$

This formula, which tracks the evolution of an age cohort over time by taking just mortality into account, is a basic yet straightforward method of modelling population dynamics.

The above equation used death rates to calculate the population but now lets see how we can incorporate Birth rate in the equation.

Let *B(t)* be the number of births at time t.

Let *P(0,t)* be the population of newborns*(a=0)* at time *t*.

The formula that takes births into account is:

$$P\ (\ 0,\ t+1) = B(t)$$

The above equation will give us our newborn population using birth rates.

The quantity of births a woman of age a might have at time t is a representation of fertility rates *F (a, t)*. A total number of births is indicated by *B (t)*.

$$B(t = \sum_{a=0} P\,(a,t)\,F\,(a,t)$$

This equation helps us to explain the birth rate at any time interval by making use of the population at any given time and age.

To determine the population at time $t + 1$:

Determine the quantity of babies born:

$$P(0,t+1) = \sum_{a=0} k\, b(a,t) \cdot P(a,t)$$

where k is the maximum reproductive age.

Update the population for every age group taking mortality and ageing into account:

$$P\,(a + 1,t + 1) = P\,(a,t) - D\,(a,t)$$

We can estimate the population dynamics taking into account both the birth and death rates by utilising these equations.As it is very clear that how these variable rates can help us in finding out the population at any given time period *t* we can make use of this information for better predicting the population growth rate (pgr).

5.2 The Influence of Migration on Population Dynamics

Now let's move onto other factors that affect the trend in the population rise and fall of which migration is a very important one. The lack of job opportunities ,good education ,better health quality and defined infrastructure are some of the things that lead to migration of people.

This research paper by Minoru Tabata, Nobuoki Eshima and Ichiro Takagi speaks about migration and its effect on the population explosion or depletion. When there are large movements of human population in a specific geographical region then we say that there has been a Population explosion.

Human Population Explosion - Important for understanding the dynamics of the human population.

Research Needed -

- There aren't many mathematical studies on population explosions.
- The value of using a mathematical model.

Regional Overpopulation -

- Caused by migration as a result of income inequality
- Affected by rising or falling birth and/or mortality rates.
- Regional overcrowding is more affected by migration than by population growth.

Migration as a Focus :

- Migration contributes to larger and quicker regional population expansions.
- Economic differences lead to major urban migration in China and Southeast Asia.

Impact of Migration on the Economy -

- Migrants enhance economic potential and present commercial prospects.

To calculate net migration *N (t)*, the population change resulting from migration is simply added or removed from the total.When net migration, births, and deaths balance out, a population stabilises which is mathematically expressed as:

$$B(t) - D(t) + N(t) = 0$$

If the above equation value is positive that means the population is in an incline position meaning that the combined effect of birth,death and migration is leading to population growth but if this value is negative that means the population is declining meaning the population is decreasing. For the positive value of the above equation it will indicate that more births have taken place than deaths and net migration has also increased and people are moving in. For the negative value of the above equation it will indicate that fewer births have taken place and more deaths have happened and the net migration is negative meaning more people are moving out.

The advent of the cohort-component technique marked the beginning of the contemporary era of population projection, notwithstanding the numerous disputes and projections that have taken place over the years.The accuracy and amount of detail of population forecasts have significantly improved thanks to this method, which takes the age and sex makeup of the population into account.The first global projection utilising this technique was completed in 1945 by Frank Notestein of the Princeton Office of Population Research (Notestein 1945).

For our research project we need to understand how the effect of migrations can affect the population growth rate of a country or the world.Understanding the dynamics of migration is very important for any country or a region and its government . If there are more people entering the country then the government needs to plan for higher accommodation methods, more well structured infrastructure and good medical facilities.If there are many people leaving a specific nation then the government should find the reason for that as many of them leaving might be the youth which will indirectly affect the future growth of the nation. From this research paper we understand the fundamentals of population migration and population explosion and depletion. Our objective will be to better use this research for our project and implement the models in the right manner.

5.3 Functional Data Models for Mortality, Fertility, and Migration in Stochastic Population Forecasts

This research is by . It was published by the Department of Econometrics and Business Statistics, Monash University,VIC 3800, Australia.
The study focuses on the creation and use of stochastic population projections using functional data models for migration, fertility, and death rates.The functional data models that are used in this study consist of models using mortality and fertility rates with time series coefficients and net migration estimation using historical population data and modelled with a functional data model due to the lack of detailed migration data.The data that was used in this study is similar to the data that we have used consisting of birth and death numbers for each calendar year and population numbers at January 1 of each year.

Central Death Rates:

- $mt(x)=Dt(x) / Et(x)$ where $Dt(x)$ is the number of deaths and $Et(x)$ is the population at risk in year t.

Sex-Ratio at Birth:

- Births are divided by sex using ρ\rhoρ, the sex-ratio at birth.

Fertility Rates:

- $Ft(x) = Bt(x) / Et^{f(x)}$ where $Bt(x)$ is the number of births and $Et^{f(x)}$ is the female population at risk in the year t.

In this study the deaths are estimated using the standard life table approach where the standard life table is a statistical tool developed in actuarial science to summarise the death patterns of a population.

Functional data models that were used for stochastic population forecasts made use of transformation and modelling techniques.

5.3.1 Box-Cox Transformation :

This transformation is used to stabilise the variance in the fertility,mortality and net migration rates.

$$yt(x) = s_t(x) + \sigma_t(x)\epsilon_{t,x}$$

The above equation is the model assumption for this transformation where $yt(x)$ is the transformed quantity which will help us predict the population growth while $st(x)$ is the smooth function that represents age specific trends observed with errors.

5.3.2 Model Implementation:

- Estimation Steps:
    1. Smooth Function $st(x)$: Estimated using nonparametric regression.
    2. Mean Function $\mu(x)$: Estimated as the mean of $st(x)$ across years.
    3. Principal Components $\phi k(x)$: and time series coefficients $\beta t,k$ estimated.
    4. Time Series Models for $\beta t,k$: Use exponential smoothing state space models.

This model can be applied to mortality,fertility and migration data and it also includes room for adjusting forecast variance in prediction and error.

5.4 Comparative Analysis of 17 Machine Learning Algorithms for Forecasting Population Growth at the Country Level

The author of this study is Mohammad Mahmood Otoom. It appears in the January 2021 edition of the International Journal of Computer Science and Network Security. The study's primary focus is the pace of increase in the human population, which is essential for practical planning.The population, death rate, and fertility rate all of which are based on historical data are fixed factors used in conventional techniques to calculate the population growth rate. The aim of the research is to shed light on several machine learning models used in the process of estimating population growth rates. The research area is defined as nations, and the data used in the study comes from the United Nations database.

The research has used various demographic variables. These include concepts of mortality ,fertility and migration rates which have been explained in the above section. The demographic features used for the models are :

- Total human population
- Human population density

- Under-five mortality rate
- Infant mortality rate (IMR)
- Female life expectancy at birth
- Life expectancy at birth
- Total fertility rate

The main purpose of using these features is because somehow they directly associate with the factor of population growth and hence are used in the models.

The UN provided the data for this literature evaluation, which was then cleaned and processed for analysis. For the purpose of this study the researchers have transformed the population growth rate (pgr) into a categorical outcome. Three results have been identified: Negative to Low Growth Rate (NLGR), Medium Growth Rate (MGR), and High Growth Rate (HGR).

For categorization, K-means clustering is employed with Euclidean distance as the distance metric. K-means clustering is a popular method of categorization that creates a single cluster out of values that are comparable to each other.

Speaking of K-means clustering ,let's understand how it works. K-means clustering basically is used to minimise the sum of squares within the clusters.It groups points based on their similarities and creates clusters for easier modelling. Mathematically ,

$$\sum_{i=1}^{k} \sum_{j=1}^{ni} \left| xj^{(i)} - ci \right|^2$$

Where :

- K is the point or object size
- ni is the number of data points in cluster i
- $xj^{(i)}$ - ci implies the jth data point which is in the ith cluster
- Ci is known as the centroid of the cluster

One important principle that this algorithm works on is removing the highly correlated features from the analysis.Using Pearson's correlation coefficient to determine the high connection between the variables, the Infant Mortality Rate (IMR) and Female Life Expectancy at Birth are eliminated from the feature set in the literature review.

Pearson's correlation coefficient:

Consider two variables X and Y ,then the Pearson correlation coefficient is given as

$$\rho(X,Y) = cov(X,Y) / \sigma X \, \sigma Y$$

Where :

$cov(X,Y)$ is the covariance between the two variables X and Y

σx and σy represent the standard deviations of X and Y respectively.

By using correlation we can discard some features and keep only the ones which are most essential for the study thereby reducing time and reducing complexity in the models.

Further ahead in the literature review ,all the 17 machine learning techniques are trained to predict the pgr from the input features. Eight ensemble learners and nine base learners are educated to create models. The following nine basic learners are listed:

- Artificial Neural Network (ANN)
- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Linear Discriminant Analysis (LDA)
- Logistic Regression (LR)
- Localised Generalised Matrix Learning Vector Quantization (Lgmlvq)
- Naive Bayes (NB)
- Quadratic Discriminant Analysis (QDA)
- Support Vector Classifier (SVC)

The study normalises the data for modelling and splits it into 80:20 where 80 % is the training data and 20 % is the validation data. A three fold cross validation is used to optimise the hyperparameters .Grid search method is used in the research to identify the hyperparameters.

The obtained findings offer descriptive statistics of the dataset, emphasising the distribution of values across various attributes and their range. Many parameters are listed with their mean, standard deviation, median, minimum, and maximum values for the three categories of population growth rate (pgr): low, medium, and high. These parameters include population density, total fertility, under-five mortality, and population growth rate.

The Population Growth Rate (pgr) is negatively correlated with population, population density, and life expectancy, but directly correlated with mortality and fertility rate, according to the results. Using all of the demographic data that was available, the machine learning models' accuracy in forecasting the nation's population growth rate (pgr) ranged from 0.58 to 0.96, highlighting how important the machine learning model selection is when creating a predictive model for pgr. Random Forest performed the best in terms of forecasting pgr, whereas Naive Bayes performed the worst.The table 5.4.1 below shows all the models and their performance in different scenarios in test data shows us which model can be best used for our research.

#	Technique	Overall pgr Performance (Accuracy)					
		Scenario					
		All	NLEB	NPOP	NPD	NTFR	NUMR
1	ann	0.88	0.88	0.88	0.88	0.89	0.88
2	dt	0.94	0.94	0.94	0.94	0.94	0.94
3	dtADA	0.94	0.93	0.93	0.93	0.94	0.93
4	knn	0.93	0.93	0.93	0.93	0.93	0.93
5	lda	0.82	0.82	0.82	0.82	0.82	0.82
6	ldaBG	0.82	0.82	0.82	0.82	0.82	0.82
7	Lgmlvq	0.86	0.86	0.86	0.86	0.86	0.86
8	lr	0.84	0.84	0.84	0.84	0.84	0.84
9	lrADA	0.83	0.83	0.83	0.83	0.83	0.83
10	lrBG	0.84	0.84	0.84	0.84	0.84	0.84
11	nb	0.58	0.58	0.58	0.58	0.58	0.58
12	nbADA	0.78	0.78	0.78	0.78	0.78	0.78
13	nbBG	0.80	0.79	0.80	0.80	0.79	0.79
14	qda	0.61	0.61	0.61	0.61	0.61	0.61
15	qdaBG	0.80	0.80	0.80	0.80	0.80	0.80
16	rf	0.96	0.95	0.95	0.95	0.96	0.96
17	svc	0.86	0.86	0.86	0.86	0.86	0.86

Table 5.4.1

Main output from this Research

The study was successfully able to overcome the problem faced by historical data and used machine learning techniques to identify areas population growth rate. Random forest outperformed all other models with highest accuracy and good results.

Another important thing to notice from this study was that by using a Random forest model with specific features the model predicts the population growth rate really well. The features associated in population growth rate prediction are

- Total human population
- Human population density

- Under-five mortality rate
- Life expectancy at birth
- Total fertility rate

This study makes us aware of how different machine learning modes are used to predict the population growth rate and which models are best effective. Our objective will be to create some models different from the ones used in this research and to see how their performance is as compared to the ones used in this research. We also see how the research uses the standardisation in their data preprocessing since the different indicators are measured in different units. We would also want to make use of the correlation analysis technique and implement it in a more efficient way in our research to find out and get the perfect indicators for our models. The research uses 3 fold cross validation and for our study we would make it more advanced by using 10 fold cross validation.In all this research can serve as a good base to advance with our pre-processing, feature selection and modelling.

5.5 Cohort Component Technique in Population Prediction

This research paper by Nicholas Ormiston-Smith, Jonathan Smith and Alison Whitworth for the Office of National Statistics shows us how the Cohort component method is implemented in population prediction. This research focuses on Nations and shows us various aspects that get into play while applying this method. It also gives us various other methods that can be used. For our project we take key points from the research and use the cohort component method on our dataset.

Key components involved in Cohort component method for population estimation :

1. Starting Point : Indicates the month or the year you want to start the estimation from.
2. Population change estimation : Here we add births ,subtract deaths and adjust accordingly for migration. All of these are done for the exact same period as the estimation and none of it changes.
3. Population base and adjustments : A decennial census serves as the foundation for population estimations. Every year, estimates are revised and rebased using fresh census information.A correction known as the One Number Census (ONC) was implemented for the 2001 Census in order to account for undercounting.
4. Alternative Population bases: The population bases can be changed based on user demands such as workday or weekend populations.

Using these 4 points and availability of proper data and life tables one can estimate population prediction. Though there are various drawbacks as well in this method, it doesn't handle the uncertain events that happen in a population or sometimes doesn't take into account variability across the same attribute for multiple times.
Some other alternative methods were also suggested :

- Housing Unit Method
- Ratio Change Method
- Apportionment Method
- Regression Methods

This research tells us that the above methods offer different approaches to estimating populations with their own strengths and limitations and the choice really depends on the needs of the researcher.

6. Population Growth Rate Modelling

The cohort component technique for population projection (CCMPP) is the basis of our primary methodology for population growth modelling. It is now the projection technique that demographers use the most. An accounting framework is offered by CCMPP for the three main components contributing to population change that are fertility , mortality and net migrations.

In order to maintain the "demographic balancing equation" (Equation 1), the three demographic components as spoken above are applied to the population in question.

$$P(t+n) = P(t) + B(t \text{ to } t+n) - D(t \text{ to } t+n) + NM(t \text{ to } t+n) \qquad (1)$$

n = The length of projection interval

t = Initial time t

$t+n$ = The interval time plus the initial time

$P(t)$ = The population at the initial time t

$P(t+n)$ = The population at time $t + n$ (The projection period n could be any specific number of years for eg . n = 5 years or n = 10 years)

$B(t \text{ to } t+n)$ = The number of births between time t to $t+n$

$D(t \text{ to } t+n)$ = The number of deaths between time t to $t+n$

$NM(t \text{ to } t+n)$ = The net migration of people in time t to $t+n$

Explanation : The population at time $t+n$ is computed by first starting with initial population $P(t)$ and then we add the number of births $B(t \text{ to } t+n)$ in that ,subtract the number of deaths $D(t \text{ to } t+n)$ and finally add the net migration $NM(t \text{ to } t+n)$. This method provides a comprehensive way to project population changes by considering natural population growth (births and deaths) and migration effects.

This equation gives us clarity about various things and points that the population dynamics of a particular region depends on four main factors namely current population of the region , migrations ,births and deaths of the region. Keeping this point in time we will further carry out some visualisations to see if these factors do have an impact on the population growth rate variable in our dataset.

Based on these factors we make some assumptions which we will test further in the report.

6.1 Current Population Distribution : We assume that a higher value of population distribution will lead to a lower population growth of a region .

$$\text{Population growth rate (pgr)} \propto 1 / \text{Current population distribution (Po)}$$

6.2 Fertility Rate : The higher the fertility rate the higher will be the population growth rate stating a direct relationship between the both.

$$\text{Population growth rate (pgr)} \propto \text{Fertility rate (BR)}$$

6.3 Mortality Rate : The higher the mortality rate the lower will be the population growth rate stating an inversely proportional relationship.

$$Population\ growth\ rate\ (pgr) \propto \ 1\ /\ Mortality\ rate\ (DR)$$

6.4 Migration Rate : The higher net migration rate will result in higher population growth rate.

$$Population\ growth\ rate\ (pgr) \propto \ Net\ migration\ rate\ (NMR)$$

Since our focus is on Population growth rate we take a relation between pgr and migration rate which will result in growth of population.

Based on the above assumptions we can now model the population growth rate (pgr) equation as follows

$$PGR = xoPo + x1BR - x2DR + x3NMR \qquad (2)$$

Where xo ,x1,x2 and x3 can be considered as constants of proportionality. Our pgr prediction equation relates our assumptions with the four important factors that have a major contribution.
Our equation represents a linear relationship but in reality it may not necessary be a linear relation and hence we make a more generalised form of our equation

$$PGR = F(P, B, D, M) \qquad (3)$$

Where F denotes the non-linear relationship and P,B,D and M denote the set of demographic variables.
In our research our aim will be to estimate the function F using the related variables in relationship with our equation.

7. Methodology

In this study, our primary focus has been on utilising machine learning techniques to estimate the expansion of the human population based on several demographic factors. This section outlines a number of research issues, relevant data and resources that are available, how the available data is handled and preprocessed, and various algorithms or modelling techniques. Specifically, we describe the methodology for estimating the parameters of the function F(.) and identifying the demographic variables that comprise the sets P, B, D, and M. We first discuss the details of the original data , the cleaning procedure and key visualisations on the data. We then offer statistical analysis to determine the demographic characteristics that comprise the sets P, B, D, and M using cleaned data. At last, machine learning is demonstrated.

7.1 Project Objectives

- The main goal of this project is to estimate the population growth rate using demographic indicators and understand the importance of each indicator in the process. The original census survey method is very time consuming , needs a lot of manpower , needs a lot of resources and is not very reliable. While looking for the best indicators we look out for some indicators that very strongly affect the target population. Hence this study first focuses on finding the individual indicators that can be used for population growth rate prediction and then use them as a whole in our appropriate models.
- Data is the most essential component for developing any system in machine learning. How can the existing data be converted into a format that an algorithm can comprehend so that it can be used to develop models without sacrificing quantity and quality?
- Selecting which algorithms and data pre processing techniques can be best applied to the data. This will help us first understand our data and get useful inputs while removing the unnecessary noise and outliers from the data.
- How successful are the various models? With the limited conditioned data now available, what should the system's generalised structure be?
- In the end we have a look at the limitations and drawbacks in our models and what other scope we can have in improving the research.

7.2 Data Collection and Pre Processing

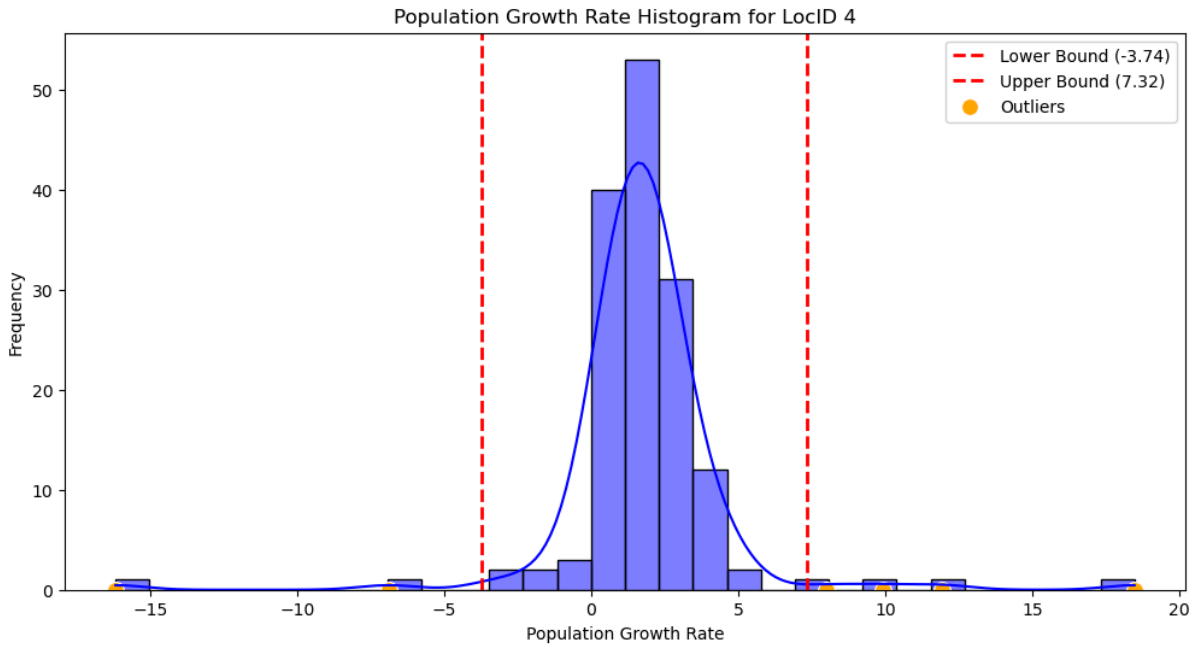
This section has covered the pre-processing procedures that were necessary as well as the dataset that was used.The Population Division of the United Nations Department of Economic and Social Affairs made the data accessible at https://population.un.org/wpp/Download/Common/CSV/. The aforementioned link has further information on the data.

Since the data was in csv format, we loaded it in a python environment in csv format and stored it in a dataframe called ‘data’. The data consists of a total 43472 rows and 67 columns which have various demographic indicators .Most of the data was in numerical format ,more specifically in floating and integer data type while there were only few categorical attributes in the data.
After this we first checked for any duplicate observations in the dataset. Doing this was necessary because if there are any duplicate observations then they can affect the process of modelling and can create errors in further processing. There were no duplicate values found and hence we moved to the next step of checking and handling any missing values in the dataset.
Handling missing values is very important because these empty rows can create error possibilities and can also sometimes give bias results for some values.
We used python code to check for missing values under each attribute and we found some missing values in the data which can be seen in table 7.2.1.

SortOrder	0
LocID	0
Notes	31920
ISO3_code	7448
ISO2_code	7600
SDMX_code	608
LocTypeID	0
LocTypeName	0
ParentID	0
Location	0
VarID	0
Variant	0
Time	0
TPopulation1Jan	0
TPopulation1July	286
TPopulationMale1July	286
TPopulationFemale1July	286
PopDensity	286
PopSexRatio	286
MedianAgePop	286
NatChange	286
NatChangeRT	286
PopChange	286
PopGrowthRate	286
DoublingTime	19166
Births	286
Births1519	286
CBR	286
TFR	286
NRR	286
MAC	286
SRB	286
Deaths	286
DeathsMale	286
DeathsFemale	286
CDR	286
LEx	286
LExMale	286
LExFemale	286
LE15	286
LE15Male	286
LE15Female	286
LE65	286
LE65Male	286
LE65Female	286
LE80	286
LE80Male	286
LE80Female	286
InfantDeaths	286
IMR	286
LBsurvivingAge1	286
Under5Deaths	286
Q5	286
Q0040	286
Q0040Male	286
Q0040Female	286
Q0060	286
Q0060Male	286
Q0060Female	286
Q1550	286
Q1550Male	286
Q1550Female	286
Q1560	286
Q1560Male	286
Q1560Female	286
NetMigrations	286
CNMR	286

Table 7.2.1

The handling of these missing values was first done on the basics of our target variable which is 'PopGrowtRate'. Since population growth rate is our main target variable we first check if there are any outliers present under this variable. For this we calculate the mean and standard deviation of the PopGrowthRate column and then we use Z score to find the outliers as using them for training the model can alter the training process. In this case we use 95% confidence interval i.e. mean +- 2*(standard deviation) as the threshold range. But these ranges were calculated for every single unique location ID in the dataset as the outliers for every location may vary and it may not be wise to use the overall threshold range so we create a threshold range for every location id based on population growth rate and remove any observations lying outside the range.
The graph 7.2.1 shows the bounds for a specific location ID (4) and we can see the observations lying outside the red lines are outliers and will be removed.



Graph 7.2.1

After implementing this step all of the missing values for demographic indicators were removed in the process of eliminating the outliers and we only had the ISO codes with missing values .These ISO code missing values were kept as it is because these features were not to be used further.
Furthermore we checked for any values that were infinity or negative infinity as it could be a mistake in the dataset and dropped them off.
We also take out any observation that has a Crude Net Migration Rate (CNMR) as 0 . The reason behind this is that a CNMR value of 0 means no net migration, that is the number of people entering and leaving the population is equal and this won't have any effect on the population increase or decrease and hence dropping it would reduce the dimensionality to some extent. We do not drop any columns now because in further analysis we select the desired features leaving the extra columns behind.
The rest of the data was clean without much discrepancy or inconsistency and the final dimensions after all the data cleaning and preprocessing was 39178 rows and 67 columns.
The table 7.2.2 below shows all the demographic indicators in the dataset alongside the data description.

Feature Index	Feature	Feature description	Unit
1	'TPopulation1July'	Total Population, as of 1 July	thousands
2	'TPopulationMale1July'	Male Population, as of 1 July	thousands
3	'TPopulationFemale1July'	Female Population, as of 1 July	thousands
4	'PopDensity'	Population Density, as of 1 July	persons per square km
5	'PopSexRatio'	Population Sex Ratio, as of 1 July	males per 100 females
6	'MedianAgePop'	Median Age, as of 1 July	years
7	'NatChange'	Natural Change, Births minus Deaths	thousands
8	'NatChangeRT'	Rate of Natural Change	per 1,000 population
9	'PopChange'	Population Change	thousands
10	'Births'	Births	thousands
11	'Births1519'	Births by women aged 15 to 19	thousands
12	'CBR'	Crude Birth Rate	births per 1,000 population
13	'TFR'	Total Fertility Rate	live births per woman
14	'NRR'	Net Reproduction Rate	surviving daughters per woman
15	'MAC'	Mean Age Childbearing	years
16	'Deaths'	Total Deaths	thousands
17	'DeathsMale'	Male Deaths	thousands
18	'DeathsFemale'	Female Deaths	thousands
19	'CDR'	Crude Death Rate	deaths per 1,000 population
20	'LEx'	Life Expectancy at Birth, both sexes	years
21	'LExMale'	Male Life Expectancy at Birth	years
22	'LExFemale'	Female Life Expectancy at Birth	years
23	'LE15'	Life Expectancy at Age 15, both sexes	years
24	'LE15Male'	Male Life Expectancy at Age 15	years
25	'LE15Female'	Female Life Expectancy at Age 15	years
26	'LE65'	Life Expectancy at Age 65, both sexes	years
27	'LE65Male'	Male Life Expectancy at Age 65	years
28	'LE65Female'	Female Life Expectancy at Age 65	years
29	'LE80'	Life Expectancy at Age 80, both sexes	years
30	'LE80Male'	Male Life Expectancy at Age 80	years
31	'LE80Female'	Female Life Expectancy at Age 80	years
32	'InfantDeaths'	Infant Deaths, under age 1	thousands
33	'IMR'	Infant Mortality Rate	infant deaths per 1,000 live births
34	'LBsurvivingAge1'	Live births Surviving to Age 1	thousands
35	'Under5Deaths'	Deaths under age 5	thousands
36	'Q5'	Under-five Mortality Rate	deaths under age 5 per 1,000 live births
37	'Q0040'	Mortality before Age 40, both sexes	deaths under age 40 per 1,000 live births
38	'Q0040Male'	Male mortality before Age 40	deaths under age 40 per 1,000 male live births
39	'Q0040Female'	Female mortality before Age 40	deaths under age 40 per 1,000 female live births
40	'Q0060'	Mortality before Age 60, both sexes	deaths under age 60 per 1,000 live births
41	'Q0060Male'	Male mortality before Age 60	deaths under age 60 per 1,000 male live births
42	'Q0060Female'	Female mortality before Age 60	deaths under age 60 per 1,000 female live births
43	'Q1550'	Mortality between Age 15 and 50, both sexes	deaths under age 50 per 1,000 alive at age 15
44	'Q1550Male'	Male mortality between Age 15 and 50	deaths under age 50 per 1,000 males alive at age 15
45	'Q1550Female'	Female mortality between Age 15 and 50	deaths under age 50 per 1,000 females alive at age 15
46	'Q1560'	Mortality between Age 15 and 60, both sexes	deaths under age 60 per 1,000 alive at age 15
47	'Q1560Male'	Male mortality between Age 15 and 60	deaths under age 60 per 1,000 males alive at age 15
48	'Q1560Female'	Female mortality between Age 15 and 60	deaths under age 60 per 1,000 females alive at age 15
49	'CNMR'	Net Migration Rate	per 1,000 population

Table 7.2.2

7.2.1 Z score in Outlier detection :

Z score is widely used in many outlier detection methods during data preprocessing and cleaning. This article tells us how Z-scores help in the outlier detection process.
Z -score basically calculates the number of standard deviations and observation is from the mean in the dataset that is normally distributed. A Z score value with its positive and negative signs implies various things like a 2 score means observation is 2 standard deviations higher than the mean and -2 score means it is 2 standard deviations below the mean.

Z-score = standard deviation (raw value − mean)
This is how the Z score is calculated for point estimation.
To get bounds or intervals we make use of something like this

CI = mean +- $Z*$ (Std)

Where we have standard Z values for confidence intervals like for 90 % confidence interval the X value is 1.645 , for 95 % confidence interval it is 1.96 ≅ 2.
This is what we use in our data cleaning to make a bound of 95% confidence interval and detect outliers.

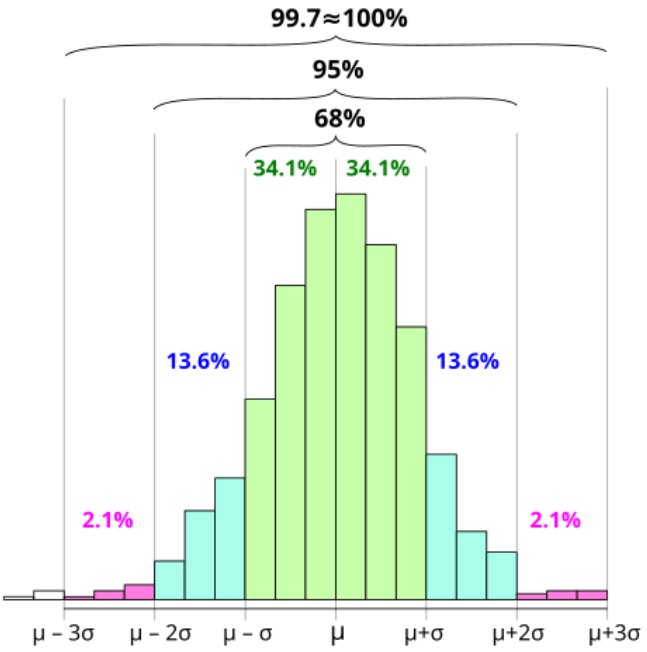


Figure 7.2.1

The figure 7.2.1 shows us the distribution of various confidence intervals and the spreading of data and the limits of the confidence interval. Using the same approach we can compare our graph 7.2.1 and see that we have made use of 95% confidence interval (u-2σ) for outlier detection.

7.3 Data Visualization

Visualising the data and getting important information out of it is the main objective of this section. As mentioned earlier in the modelling part, we have three main factors apart from population number itself that are births , deaths and migrations that are related to population growth rate. To see and

validate the claims of direct and inverse relationships between these variables and target variables we visualise them with the help of graphs.
Since our dataset is very large and the indicator units are also too many we take the approach of binning the samples. We first see the overall statistics of the attribute and then create 10 bins of the respective attributes. Then we calculate the simple mean of each bin to have an idea about what observations are there in the bins. Along with this we also calculate the mean population growth rate for each respective bin which then gives us the 10 bins with the average of each bin and average population growth rate for each bin. With help of this information we plot a line chart by keeping the mean population growth rate on Y axis and changing the X axis with respective attributes. We make use of logarithmic scale on the X axis as the interval gaps are very large between respective points and clarity of viewing the points is not possible if used on a normal scale.

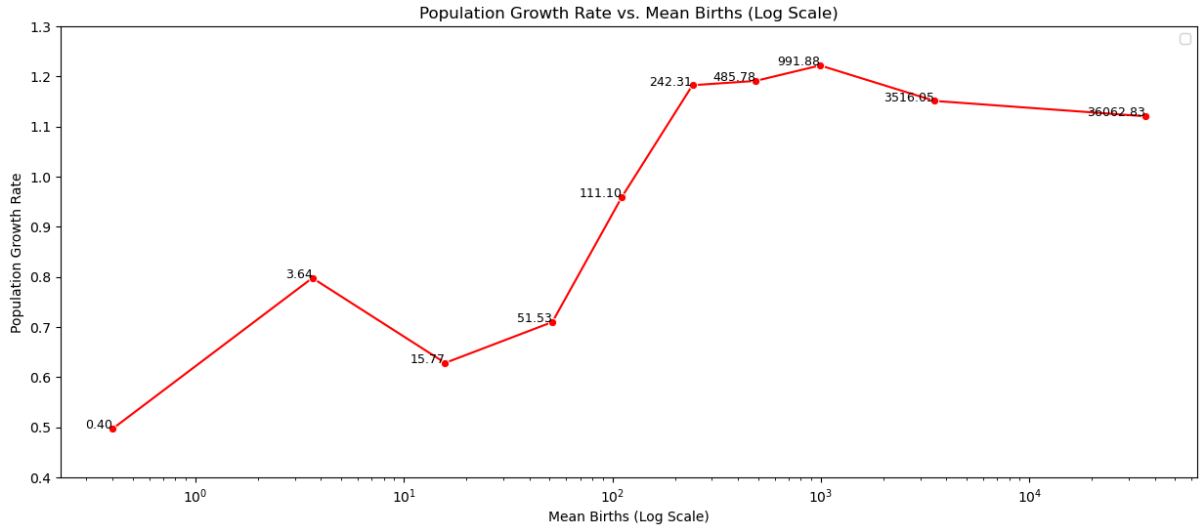
These graphs are mainly done for us to understand the relationship and patterns on how population growth and various factors associated with it go hand in hand.

7.3.1 Visualising Relation between Births and Population Growth rate

We have births in thousands in our data which we took which represents the overall birth count. The graph(1) below shows the relationship between births and population growth. It is clearly visible that as the mean birth rate increases the mean population growth rate also increases stating a direct relationship between both the factors.
Furthermore we also want to check this with another birth related attribute and hence we make use of crude birth rate.

	Births_bins	Mean_Births	Mean_PopGrowthRate
0	0	0.401062	0.497282
1	1	3.639682	0.798059
2	2	15.773058	0.628595
3	3	51.530818	0.710153
4	4	111.104414	0.959732
5	5	242.310801	1.182692
6	6	485.780847	1.191205
7	7	991.883790	1.222170
8	8	3516.045759	1.151328
9	9	36062.827977	1.120715

Table 7.3.1



Graph 7.3.1

Crude Birth Rate (CBR)

Crude birth rate represents the number of live births happening in a population per thousand individuals during a specific year.

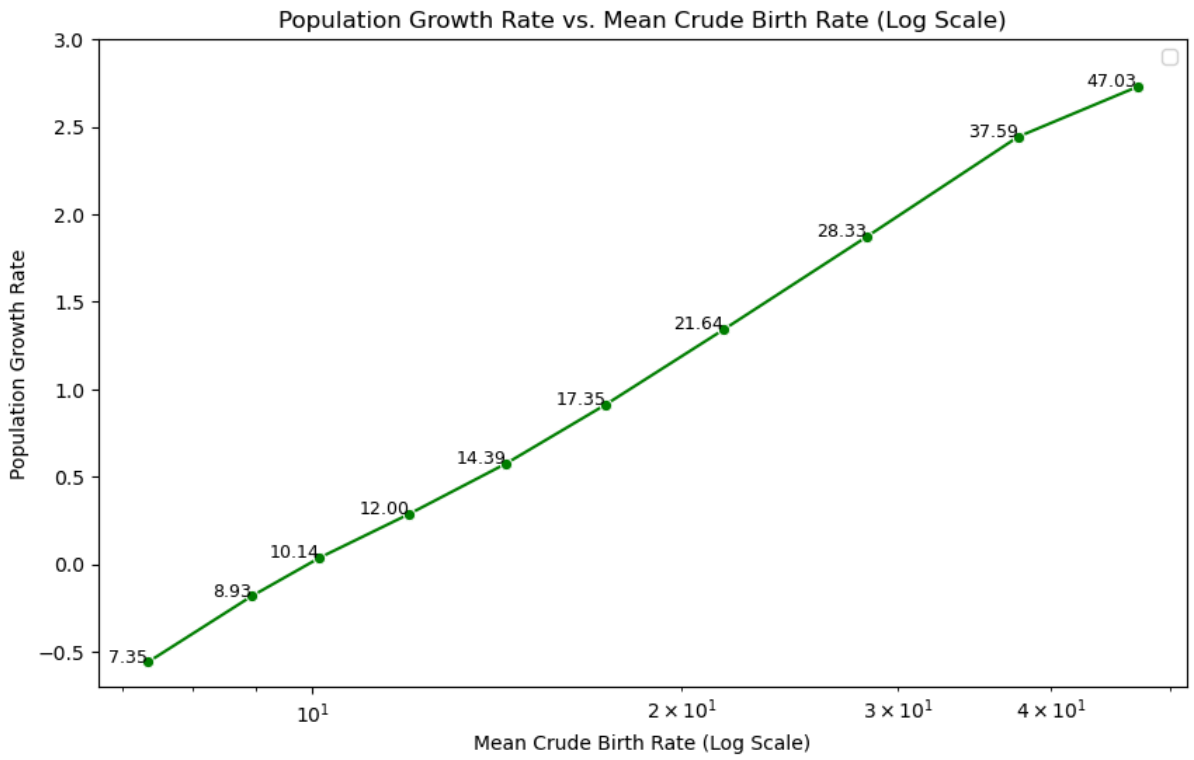
$$CBR = (B / N) \times 1000$$

Where B stands for number of births and N stands for midyear population.
The advantage of using crude rates here is that it standardised the number of births relative to population size allowing to compare it with other different factors.

Using the same method we plot a graph for Mean CBr and mean Population growth rate and get to see that there is a clear linear direct relationship between both the factors.

	CBR_bins	Mean_CBR	Mean_PopGrowthRate
0	0	7.349779	-0.558307
1	1	8.926760	-0.181752
2	2	10.139721	0.037646
3	3	11.996449	0.287861
4	4	14.392417	0.576625
5	5	17.349866	0.913562
6	6	21.643419	1.341940
7	7	28.325936	1.873668
8	8	37.591556	2.442710
9	9	47.025663	2.728525

Table 7.3.2



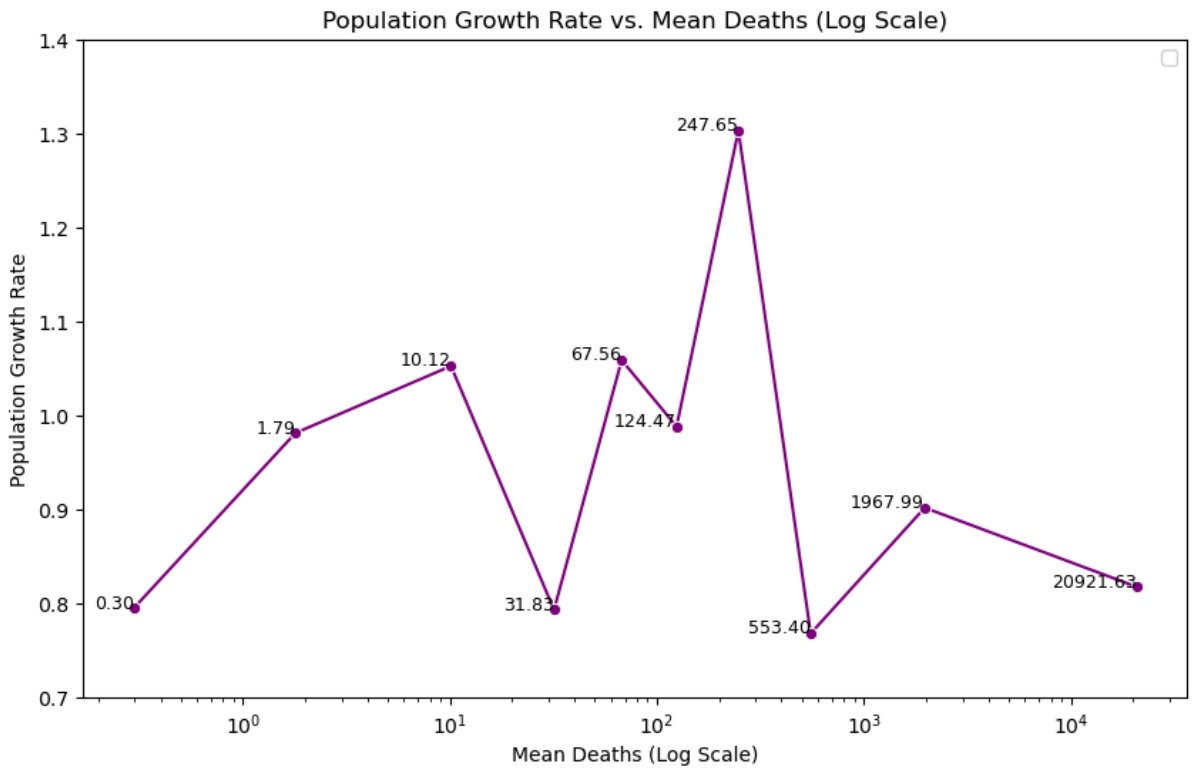
Graph 7.3.2

7.3.2 Visualising Relation between Deaths and Population Growth rate

We keep the same approach and plot the graph for mean deaths vs mean population growth rate for 10 specific bins which equally represent the whole data. From the graphs we can see that the deaths and growth rate of population show a varying pattern and hence we can't confirm a relationship between them and hence we need to make use of Crude Death rate (CDR) to obtain a relationship.

	Deaths_bins	Mean_Deaths	Mean_PopGrowthRate
0	0	0.296098	0.795016
1	1	1.791742	0.981557
2	2	10.115783	1.052903
3	3	31.832900	0.793397
4	4	67.563670	1.059193
5	5	124.472065	0.988427
6	6	247.652093	1.303459
7	7	553.395487	0.768191
8	8	1967.985504	0.902036
9	9	20921.633076	0.817702

Table 7.3.3



Graph 7.3.3

Crude Death Rate (CDR)

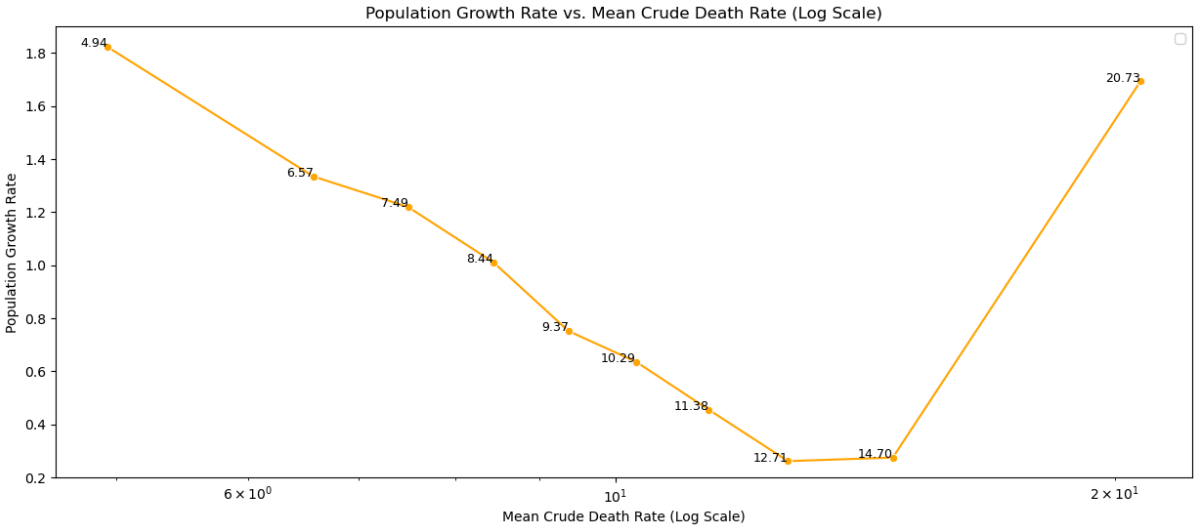
Crude death rate represents the number of deaths happening in a population per thousand individuals during a specific year.

$$CDR = (D / N) x 1000$$

Where D stands for number of deaths and N stands for midyear population.
Plotting the mean CDR against mean population growth rate we get to see that for most of the part the graph is declining stating an inverse relationship except for two points in the graph where it rises. But from overall both the graphs it is safe to conclude that deaths and population growth rate have an inverse relationship as more number of deaths will lead to decrease in the population.

	CDR_bins	Mean_CDR	Mean_PopGrowthRate
0	0	4.937629	1.822297
1	1	6.570400	1.334611
2	2	7.493827	1.219665
3	3	8.436262	1.011749
4	4	9.367810	0.751597
5	5	10.289675	0.635900
6	6	11.381839	0.455341
7	7	12.706110	0.261691
8	8	14.696864	0.275056
9	9	20.733257	1.693042

Table 7.3.4



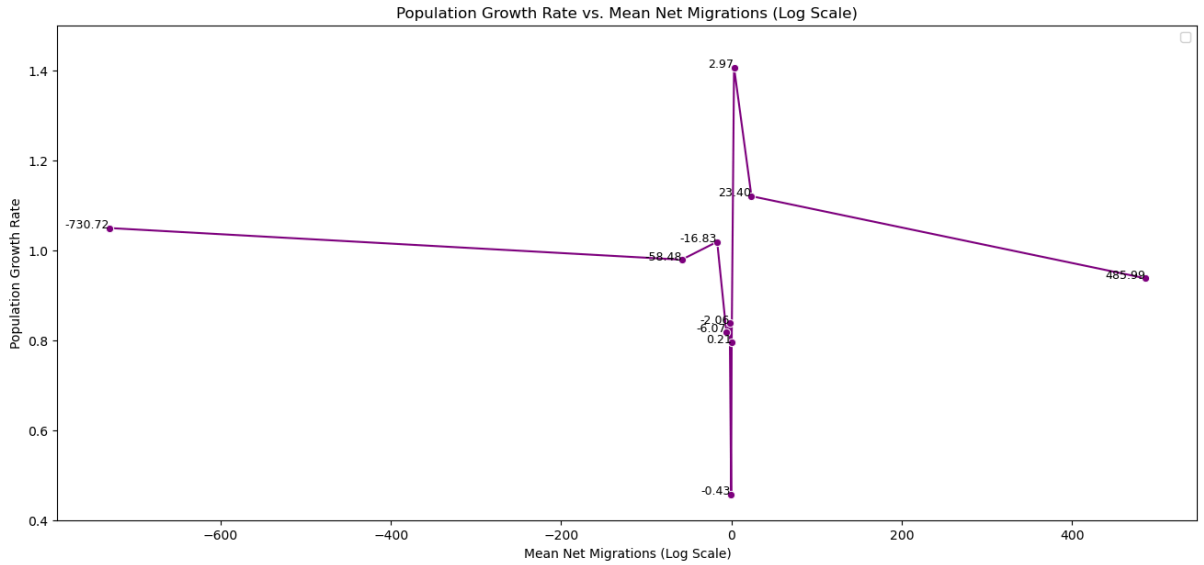
Graph 7.3.4

7.3.3 Visualising Relation between Migrations and Population Growth rate

Using the same approach as used for births and deaths we first plot the graph of net migrations against population growth rate to obtain a relationship and then we make use of crude net migration rate (CNMR). From the below graph of mean migrations against mean population growth we see a kind of constant trend where the migrations are not affecting the population growth rate as much and then we have some spikes up and down which are suggesting not so clear relationships between the two.

	NetMigrations_bins	Mean_NetMigrations	Mean_PopGrowthRate
0	0	-730.715942	1.049261
1	1	-58.479304	0.978901
2	2	-16.826987	1.018760
3	3	-6.066235	0.818604
4	4	-2.062255	0.837724
5	5	-0.434459	0.457533
6	6	0.207173	0.794335
7	7	2.970710	1.406520
8	8	23.404584	1.120087
9	9	485.986644	0.937569

Table 7.3.5



Graph 7.3.5

Crude Net Migration Rate (CNMR)

Crude net migration rate is the number of immigrants per 1,000 people in a population minus the number of emigrants per 1,000 people in a population.

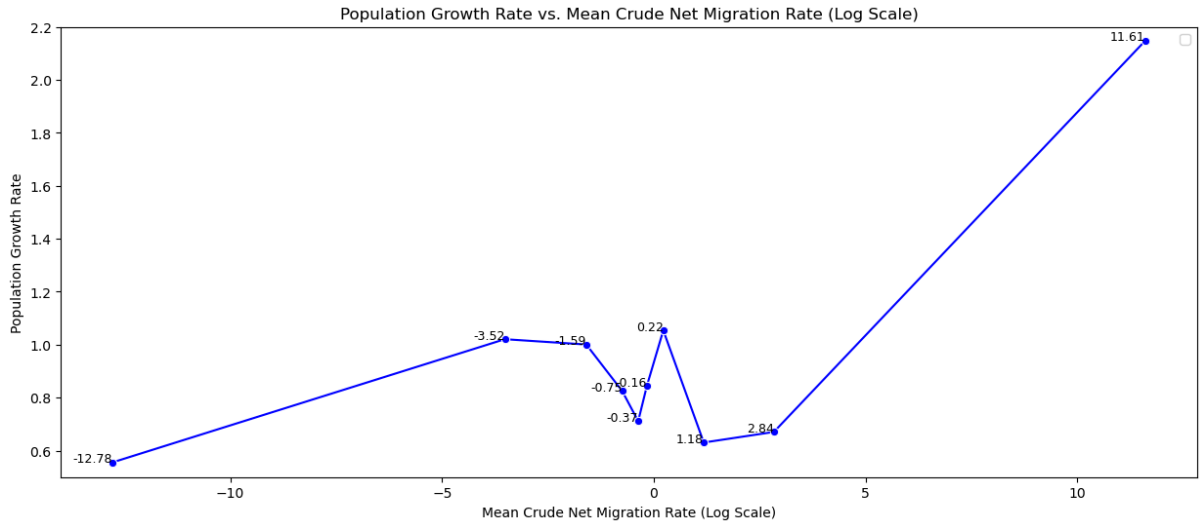
$$CNMR = (I - E)/ N \times 1000$$

Where I is immigrants entering the area , E is emigrants leaving the area and N is the midyear population.

From this graph we clearly see that the number of migrations and population growth rate have a direct relationship and the graph line keeps increasing as the crude net migration increases.

	CNMR_bins	Mean_CNMR	Mean_PopGrowthRate
0	0	-12.779520	0.555761
1	1	-3.517351	1.020758
2	2	-1.594371	1.000683
3	3	-0.745859	0.824845
4	4	-0.367910	0.711285
5	5	-0.161931	0.845018
6	6	0.224606	1.054013
7	7	1.180704	0.630534
8	8	2.836202	0.670506
9	9	11.611860	2.148991

Table 7.3.6



Graph 7.3.6

7.3.4 Analysing the Relationship Between Reduced Birth Rates and Net Reproduction Rate Across 10 Years

a. Net Reproduction Rate (NRR)

The net reproduction rate denoted by R0 is defined as the expected number of female newborns given birth by a woman during her entire life. The Mathematical formula for NRR is given as

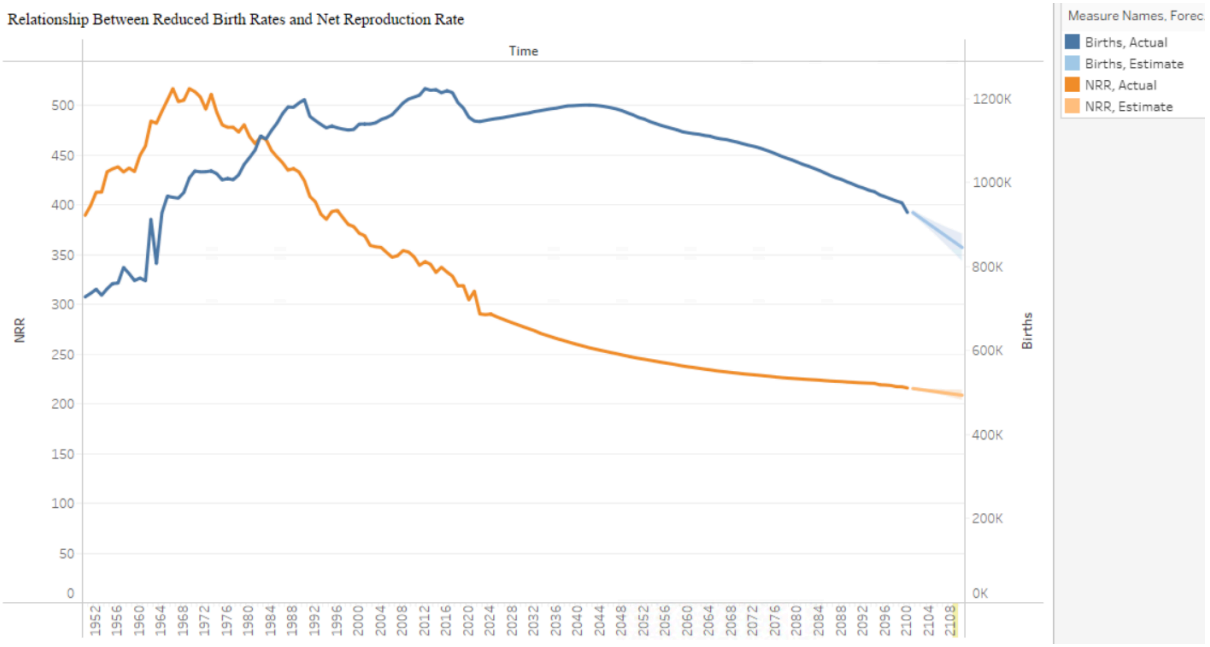
$$R_0 = \int_0^{inf} \beta(a) l(a) da$$

□*(a)* = maternity function (The specific age rate of having female birth)
l(a) = Female survival rate at age a
and da is the integration function ,where the integration is taken over from 0 to infinity.

The graph below shows the forecast for next 10 years for Net Reproduction Rates (NRR) and number of births. The purpose of this graph is to see how the effect of lower net reproductive rate can affect the number of births over the years. As we can clearly see that from 1950 to early 2000's the number of births kept on increasing but on the other side the net reproductive rate kept slowly decreasing. After the year 2020 the net reproductive rate is going only downwards which inturn will affect the number of births happening.
We then do a forecast at 95% confidence interval in Tableau and make a forecast for the next 10 years from 2101 to 2110. On plotting the graph we see that based on the past data and the trend of birth rates and net reproduction rates the graph for both the indicators is a decreasing one. Some of the main reasons for the reproductive rate to decrease are an increase in sexually transmitted diseases , various changes in the lifestyle of males and females , urban lifestyle and urbanisation etc. All these factors are highly affecting the fertility rates and causing a rise in the male and female infertility rates.

b. Solutions on Making Net Reproduction Rates better

Apart from the medical reasons there are several other socio economic factors that have led to womens not willing to bear a child or couples delaying to have children. Some of the factors involve no proper housing facilities for couples to set a family ,more flexible jobs in part time and full time which has led to a delay in conceiving a child. Aparat from this family planning also plays a vital role as couples and women decide to bear children at a later stage of their life after which the womens might face problems like decline in fertility due to ovarian ageing which all lead to a reduced chance of conception.
The government and the nations should provide proper family planning insights to the couples and everyone should try following it. The government should also fund reproductive health and social care facilities to achieve their targeted population with more births so that the nation can have a younger population for further development.



Graph 7.3.7

7.4 Correlation Analysis to identify key demographic indicators

7.4.1 Pearson's correlation coefficient

Similarity scores are based on viewing and comparing data objects of one attribute to another and then summing the squares of the differences in their magnitude .
Pearson's correlation score lies in the range of -1 to 1 where 1 or close to one score indicates high correlation among the attributes meaning one rises when the other rises ,while -1 score indicates a negative correlation meaning one rises when other falls. A Pearson's score of 0 means no correlation among the attributes.This also focuses more on linear relationships between the attributes. The Pearson correlation for two attributes is calculated as

$$\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$$

Where x and y are the two attributes and x bar and y bar are their means and $\sum(xi - x\square)^2$ *and* $\sum(yi - y\square)^2$ are the variances and the root of that is what our standard deviation is.

7.4.2 Spearman's Rank Correlation

Spearman's correlation examines monotonic correlations whether or not they are linear whereas Pearson's correlation studies linear relationships. The Pearson correlation between the rank values of two variables equals the Spearman correlation between those two variables. A perfect Spearman correlation of +1 or −1 happens when every variable is a perfect monotone function of every other variable, provided that no data values are repeated.

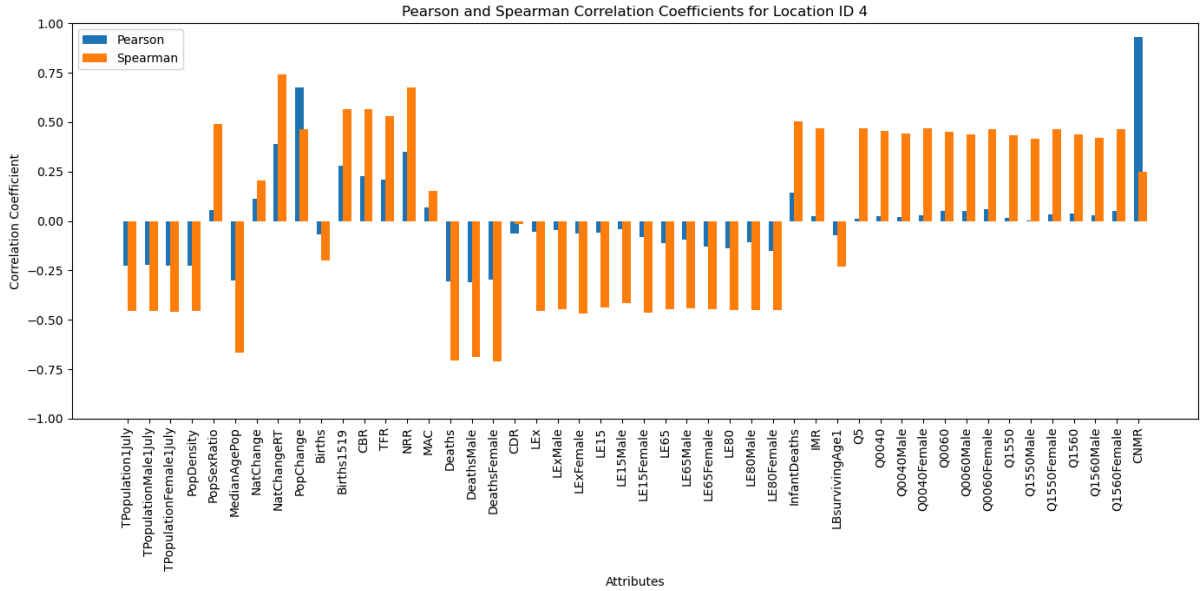
When two variables have a similar (or identical, for a correlation of 1) rank that is, when the observations are labelled as being in the same relative position within the variable the Spearman correlation between them should be high. Conversely, if the observations have a dissimilar (or fully opposed, for a correlation of −1) rank then the Spearman correlation between the two variables should be low. It is calculated as

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\operatorname{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}},$$

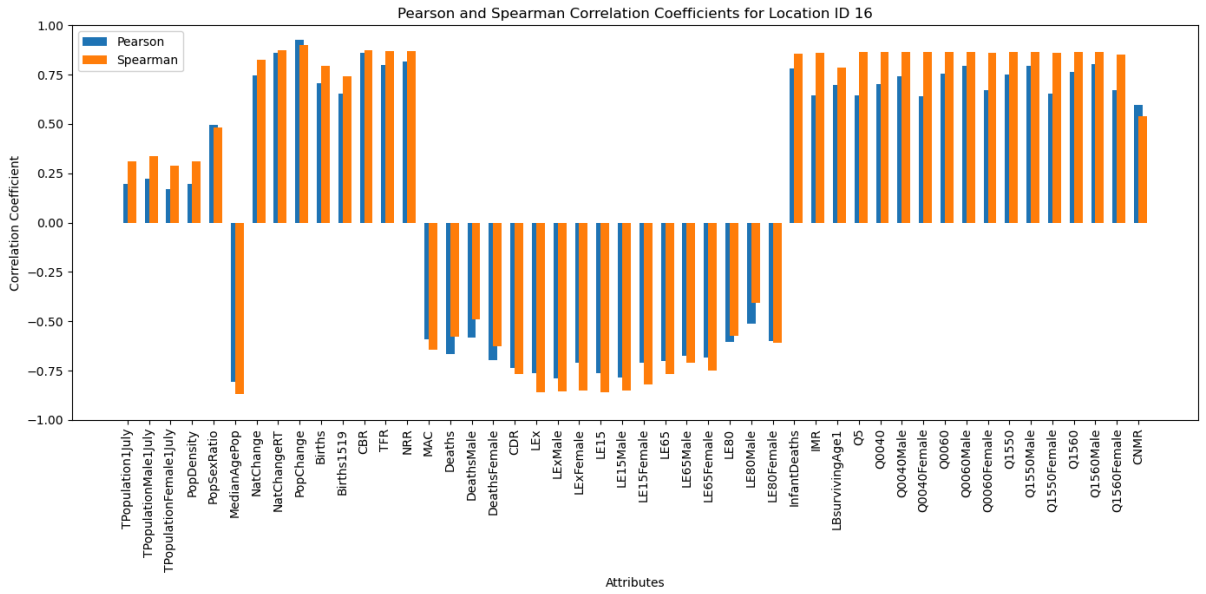
Where P*R(x),R(y)* is the pearson's correlation coefficient implied on rank variables , cov(*R(x),R(y)*) is the covariance of attributes x and y and $\sigma R(x)$ and $\sigma R(y)$ are the standard deviations respectively.

Also when the dataset is numerous it is suggested to calculate both the correlations as they go hand in hand and also help in finding the attributes that are inter correlated other than the target attribute.

In our project we use both the correlations to find our predictor attributes. For this we compute the correlations between target variable and the predictor variables. First to see how the correlations work we compute it for two random location ids. Here in Graph 7.4.1 and 7.4.2 we see that the correlations among the demographic variables are different for both locations and this will be the same case for all the other locations present. Hence to solve this problem we make use of the weighted average method and find out the weighted average of Pearsons and Spearman's Rank correlations for all the unique location ids to get a proper view as which of the indicators are best correlated with the target i.e. Population growth rate.

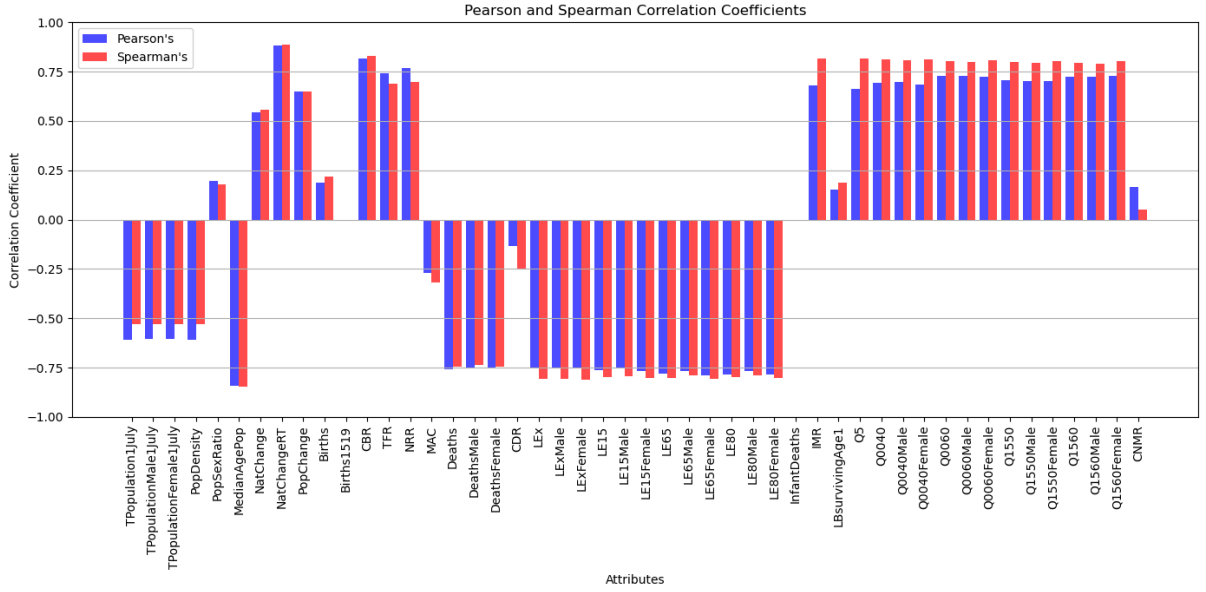


Graph 7.4.1



Graph 7.4.2

Figure 3 shows us the weighted average of correlations for all the demographic indicators. Now from this we see that various indicators have high correlation present with the target variable. Statistically we should also keep in mind that our models might overfit or underfit because of high inter correlation amongst the attributes.This helps in eliminating redundancy in feature space and helps reduce the dimensionality for modelling. Hence we choose representative indicators for our model which are based on the cohort component analysis and which show high correlation with the target variable.



Graph 7.4.3

Table : Demographic variables shortlisted for Modelling

Feature Set to represent	**Features Selected**	**Feature Explanation**
Population	TPopulation1July	Total Population, as of 1 July
Population	PopDensity	Population Density, as of 1 July
Fertility	TFR	live births per woman
Fertility	NRR	surviving daughters per woman
Mortality	LEx	Life Expectancy at Birth, both sexes
Mortality	Q1560	deaths under age 60 per 1,000 alive at age 15
Migration	CNMR	Net Migration Rate

Table 7.4.1

The above features were shortlisted from the list of demographic variables on which we performed correlation analysis. Based on our population growth rate modelling we ensured to have at least one feature from each of the demographic features (P , B , D , M) and also eliminated any redundant features. These features best represent their feature set and also have high correlation with the target variable. By selecting these features we best represent our population growth modelling equation based on cohort component technique.

7.5 Data Normalisation

We normalise our data and get it into standard format for several reasons. This is done because the demographic indicators are measured with different units and using them as it is may cause discrepancies.
Hence we use standard scalar fit where we get our observations in such a manner so that they have a mean of 0 and a standard deviation of 1. After this our data is ready to be used in the models.

7.6 Cross Validation

Cross validation is a largely used data resampling method for models and evaluation (Daniel Berrar Machine Learning Research Group School of Mathematics and Statistics The Open University, Milton Keynes, United Kingdom) A crucial technique for adjusting model hyperparameters, contrasting learning algorithms, and assessing model performance is cross-validation.According to research, the least biassed resampling techniques are the.632+ bootstrap method, 10-fold cross-validation, and LOOCV (Leave-One-Out Cross-Validation).Based on this as we have a huge dataset it would only make sense to use cross validation with 10 folds which will ensure us that all of the data has been used to its fullest.

7.6.1 Training Models with Cross validation

We make use of the k-fold cross validation technique where we divide the data into k equally sized subsets. We split the data into 10 folds so the cross validation will involve 10 iterations so each iteration will use a different fold as test sets. The shuffle value is set to 'True' meaning that the data will be shuffled before splitting into folds ensuring that the data is randomly divided and there is no chance of biasness. The random state set to 0 ensures that the shuffling is done in a reproducible way.

7.7 Machine Learning Models to Estimate Population Growth

In this part, we estimate the function F(.) that most accurately captures the relationship between the population growth rate and a subset of demographic factors using machine learning techniques.
The subsequent segment provides an in-depth explanation of every function F(.) formulation, together with its corresponding parameters and the machine learning techniques employed to extract these parameters from the dataset.

7.7.1 Polynomial Regression

Multiple Regression Model

The following represents a multiple regression model with a dependent variable (y) and a collection of x independent variables (predictors):

$$y1 = \square 0 + \square 1x11 + \square 2x12 + \square 3x13 + \ldots\ldots + \square nx1n + e1$$

$$y2 = \square 0 + \square 1x21 + \square 2x22 + \square 3x23 + \ldots\ldots + \square nx2n + e2$$

$$y3 = \square 0 + \square 1x31 + \square 2x32 + \square 3x33 + \ldots\ldots + \square nx3n + e3$$

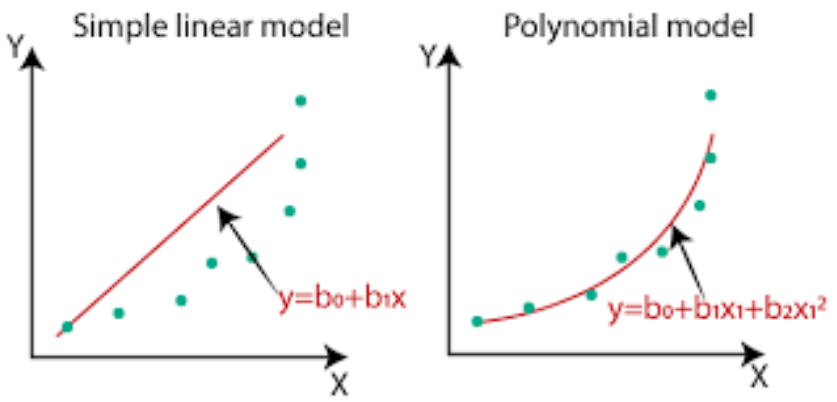
Where *y1,y2,y3* are the values of the dependent variable Y based on the independent variables *xij* and *e1,e2* and *e3* are error terms.

This is the base of multiple regression that is used by the polynomial regression. Now the factors our model will be evaluated are explained ahead.

Polynomial regression is a different case of multiple regression; power terms follow one another in a polynomial regression model. All lower order terms, whether significant or not, will be included in each model along with the highest order term , its mathematical equation is given as

$$yi = \beta 0 + \beta 1xi + \beta 2xi^{2} + \beta 3xi^{3} + \ldots\ldots + \beta kxi^{k} \quad , for\ i = 1,2 \ldots\ldots , n$$

where k is the polynomial degree, which also denotes the model's order.
This is equivalent to having many models with $X1 = X$, $X2 = X^{2}$, $X3 = X^{3}$, etc.



Graph 7.7.1

a. Mean Squared Error (MSE)

The mean squared error Equation defines MSE as an unbiased estimator of the variance σ^ 2 of the random error term.

$$MSE = \frac{SSE}{df_E} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}$$

The dependent variable Y's fitted values for the ith case are denoted by yi. An indicator of how well the regression fits the data is the mean squared error (MSE), which is calculated by dividing the total by the number of degrees of freedom.

b. R-squared

The multiple regression's R-squared R2 (coefficient of determination) is comparable to that of the simple regressions coefficient of determination. It measures the variation in response variable y based on the independent or predictor variables in term of percentage .Its mathematical formula is given as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

SST : total sum of squares
SSE : Sum squared of errors
Y bar : it is the arithmetic mean

R-squared value always lies between 0 and 1 ,value above 0.9 is always very good , anything above 0.8 is good and a value above 0.6 can be satisfactory based on the approach and other factors involved in modelling.

c. Adjusted R-squared

The number of variables in the regression equation is taken into account while adjusting R-square. When the value of adjusted R -square is significantly less than the value of R-squared, it suggests that our regression equation may have limited generalisability and may have been overfit to the sample.

$$R^{*2} = R^2 - \frac{(1-R^2)\,k}{n-(k+1)}.$$

d. Advantages of Polynomial Regression in Population Growth Rate modelling

The dataset in itself is very large and population data is generally very complex and it doesn't follow any one particular trend. Hence making use of polynomial regression would help predict the non linear relationship between the target and predictor variables.
By making use of higher degree polynomial terms ,the model can predict with better precision and can fit in various curve shapes helping with intricate population dynamics.
Population growth rates are usually affected by various interacting factors like population density and birth rates. Polynomial regression can capture these interactions and provide a more advanced understanding to predictions and help in real world application.

7.7.2 Decision Tree Regression

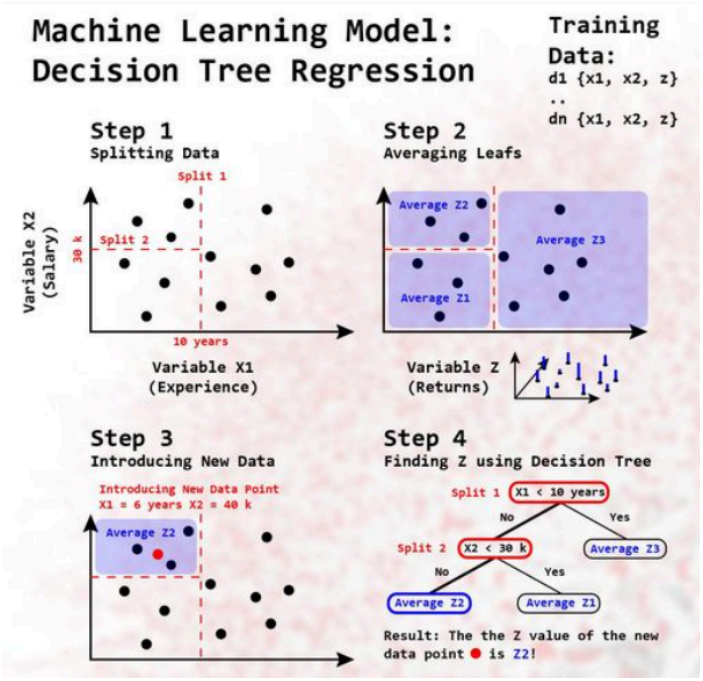
A kind of tree-based structure called decision tree regression is used to forecast the dependent variable's numerical results. An implementation of Quinlan's M5 algorithm, it is also referred to as the M5P algorithm .

a. Working of Decision Tree

First, a tree is constructed using a traditional decision-tree approach. A splitting criterion that reduces the intra-subset volatility in the class-values of instances that proceed down each branch is used in this decision tree. The root node is determined by selecting the property that maximises the projected reduction in error.

The standard deviation of reduction is given by the formula

$$SDR = sd(T) - \sum i\,|Ti\,/\,|T| \times sd(Ti)$$



Graph 7.7.2

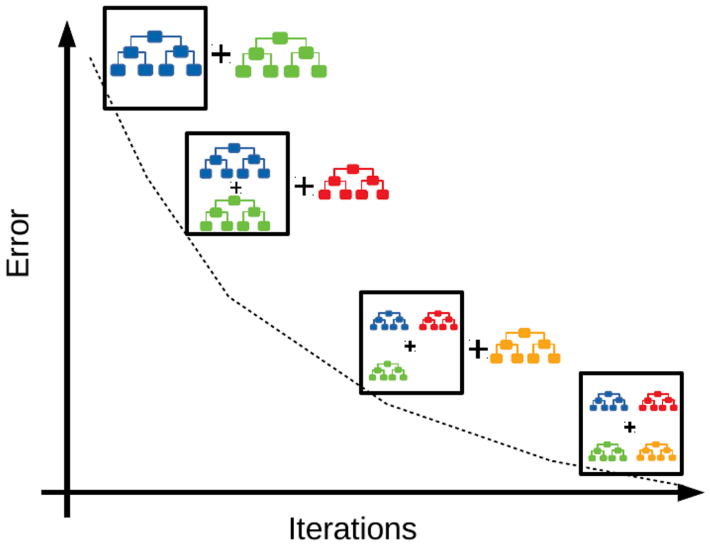
The tree is then trimmed back to just a few leaves. Lastly, severe discontinuities between consecutive linear models at the pruned tree's leaves are compensated for using a smoothing process.
In figure the steps are shown on how the splitting is performed in the dataset then how the leafs are taken into account and then further how the new data points are added and the final tree is formed for analysis.

b. Advantages of Decision Tree Regression in Population Growth Rate modelling

Choosing a decision tree for this project is a right choice because it is able to predict the numeric outcome of the dependent variable with more precision.
Another advantage of using decision tree regression is that it handles datasets with high dimensionality very well and helps in better decision making.
Decision Tree can work well with both categorical and numerical data which makes it adaptable to mix data types.

7.7.3 Gradient Boosting Regression

Unlike bagging algorithms the boosting algorithm works in a different manner where it generates base models in a sequential format. Multiple nodes are used to improve the prediction accuracy and these nodes put more emphasis on training cases that are tough to estimate. What the additional base models do is basically correct the mistakes made by the previous base models and in this manner the overall error is reduced and we get more accurate predictions.
The boosting algorithm is based on the logic that whether a set of weak learners is equivalent to a single strong learner? A strong base model is well correlated with the model and is more accurate in predictions while the weak learner is an algorithm that performs a bit better than random guessing. Now it is very easy to estimate a weak model instead of a string model and Schapire gives the solution to this by applying boosting algorithms where many weak models are combined into a single strong accurate model.



Graph 7.7.3

In figure we can see how the error reduces when several iterations are performed under a boosting algorithm where it learns from various weak models and gives accurate results by combining them into one good model.

a. Advantages of Gradient Boosting Regression in Population Growth Rate modelling

The Gradient boost algorithm is an ensemble method and its learning from its previous base models helps in reducing errors which may be caused due to the interacting factors in the population dynamics.
Like polynomial regression this algorithm also helps in predicting nonlinear relationships which is important because the population growth rates may not have simple linear relationships.
Though the ensemble methods are more prone to overfitting, the gradient boosting algorithm has the mechanisms to prevent it.

8. Results and Discussions

We provide and examine the findings from the methodologies discussed in the previous part in this section. Using cross validation where we trained various machine learning models and reported on their predictions. We start by discussing each model's evaluation metrics followed by some graphs that help in learning about the actual and predicted values. We then compare various models and identify the best model for our data.

8.1 Results on Polynomial Regression

The polynomial regression model that we use is of degree 2 ,what this does is that it generates new features that are a polynomial combination of the original features. This helps our model to fit more complex non linear regression models in the dataset. We fit the linear regression to the model which actually is applied to the polynomial transformed features.

```
Model: Polynomial Regression
Mean Squared Error (Cross-Validation): 0.1086
Mean Absolute Error (Cross-Validation): 0.2499
Root Mean Squared Error (Cross-Validation): 0.3295
R-squared (Cross-Validation): 0.9384
Adjusted R-squared (Cross-Validation): 0.9384
```

Figure 8.1

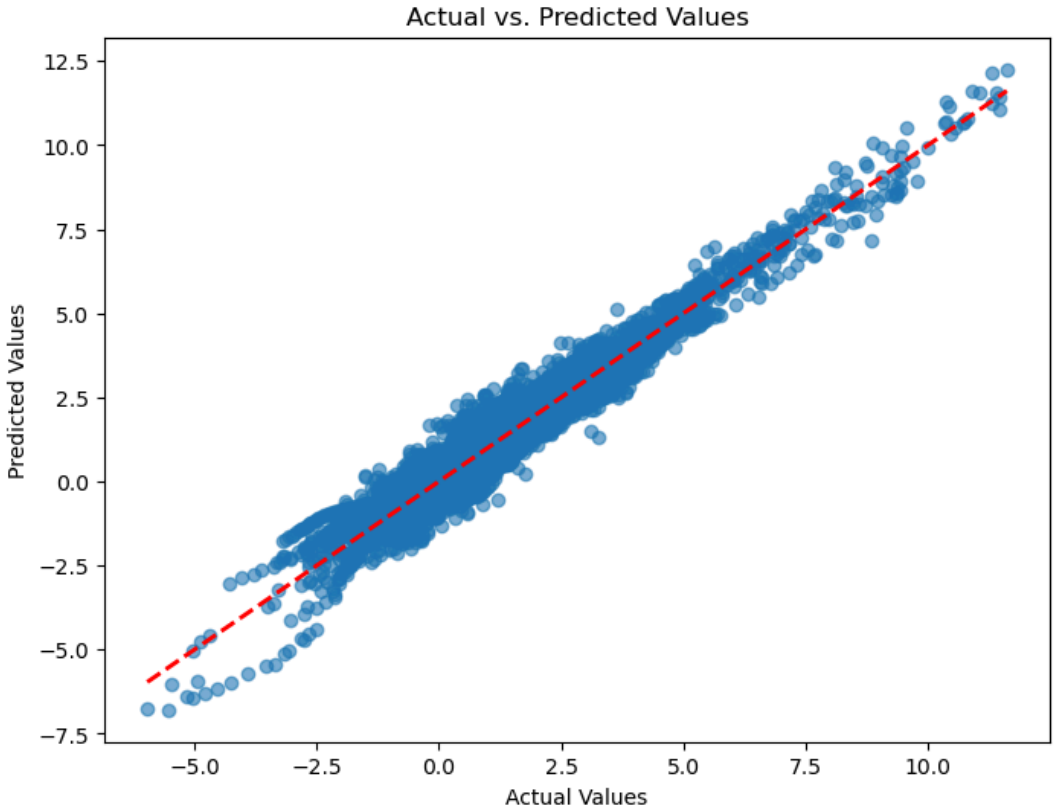
8.1.1 Model Performance

As we can see in the figure 8.1, the model gives a very good R-squared value of 93.84%. It means that approximately 93.84% of the variance in the target variable (PopGrowthRate) is explained by the polynomial regression model and the model is a strong fit.
Since the adjusted R-squared value and R-squared value are the same it explains that there has been no overfitting in the model and the model complexity is justified.

8.1.2 Error Metrics

A mean squared error (MSE) of 0.1086 means that the squared difference between the actual and predicted values is very less and the model performs well.
A mean absolute error (MAE) of 0.2499 means that on average the absolute error between the actual and predicted values is low and the model performs well.
The root mean squared error (RMSE) of 0.3295 implies that the average prediction error is about 0.3295 units.

8.1.3 Graphical Interpretation of Actual vs Predicted values



Graph 8.1

The graph 8.1 shows the actual values on the x axis and the predicted values on the y axis. The red dotted line passing through the middle is the line of best fit and the blue scatter dots are the predicted values. In this graph we can see that the predicted values lie in a clustered form on the line of best fit indicating that the predicted values are not lying too far from the actual values. Since most of our population growth rate observations are lying in the ranges -1.5 to 6.0 ,we see a more intense cluster of points in that particular region.

8.2 Results on Decision Tree Regression

For our decision tree regression we set the hyperparameter values where we keep the maximum depth of the tree as 5 . This is done so that to avoid complexities and overfitting of the training data. This will help to control the number of splits in the data . A shallow tree might not capture all complexities but with a deeper tree there are high chances of overfitting.

```
Model: Decision Tree Regression
Mean Squared Error (Cross-Validation): 0.2119
Mean Absolute Error (Cross-Validation): 0.3461
Root Mean Squared Error (Cross-Validation): 0.4603
R-squared (Cross-Validation): 0.8797
Adjusted R-squared (Cross-Validation): 0.8796
```

Figure 8.2

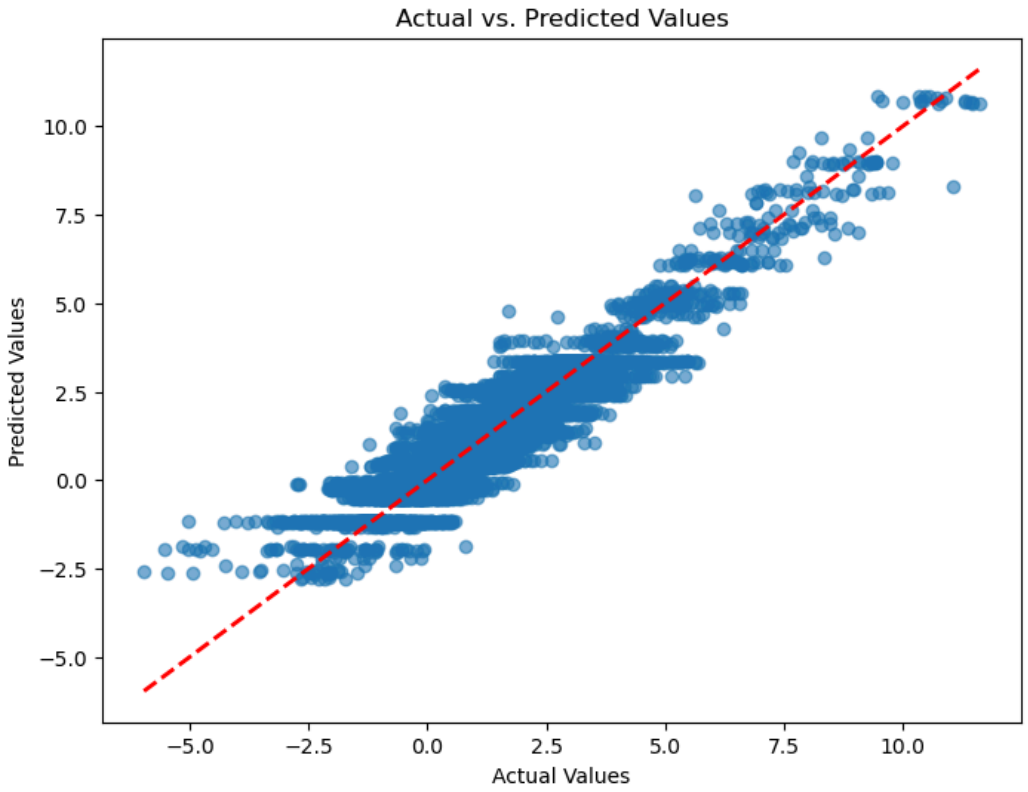
8.2.1 Model Performance

As we can see in the figure 8.2, the model gives a good R-squared value of 87.96%. It means that approximately 87.96% of the variance in the target variable (PopGrowthRate) is explained by the Decision Tree regression model and the model is a good fit.
We see that the adjusted R-squared value is lower than the R-squared value and there can be several reasons for this.
First of all the difference is not that much and can be ignored but some of the possible reasons for this can be using of cross validation as the metrics are being averaged over several folds and all the folds might have different characteristics then this could result in adjusted r-squared value being lower than the r-squared value.
Though there is a difference in R-squared and adjusted R-square values, the difference is very small which tells us that the model has avoided over complexities.

8.2.2 Error Metrics

A mean squared error (MSE) of 0.2119 means that the squared difference between the actual and predicted values is very less and the model performs well.
A mean absolute error (MAE) of 0.3461 means that on average the absolute error between the actual and predicted values is low and the model performs well.
The root mean squared error (RMSE) of 0.4603 implies that the average prediction error is about 0.4603 units.

8.2.3 Graphical Interpretation of Actual vs Predicted values



Graph 8.2

As over here in this graph 8.2 we can see that the predicted values are not as clustered toward the line of best fit implying that there is a difference between actual and predicted values. Some of the points in the lower region and upper region are way deviated from the best fit line and this explains the error terms being high and the adjusted r-squared value being slightly less than the R-squared value.

8.3 Results on Gradient Boosting Regression

The gradient boosting algorithm works very well on the dataset as it learns from several weak models to form one good model. We pass the transformed standard data through the 'Gradientboostingregressor'.

```
Model: Gradient Boost Regression
Mean Squared Error (Cross-Validation): 0.0789
Mean Absolute Error (Cross-Validation): 0.2092
Root Mean Squared Error (Cross-Validation): 0.2808
R-squared (Cross-Validation): 0.9552
Adjusted R-squared (Cross-Validation): 0.9552
```

Figure 8.3

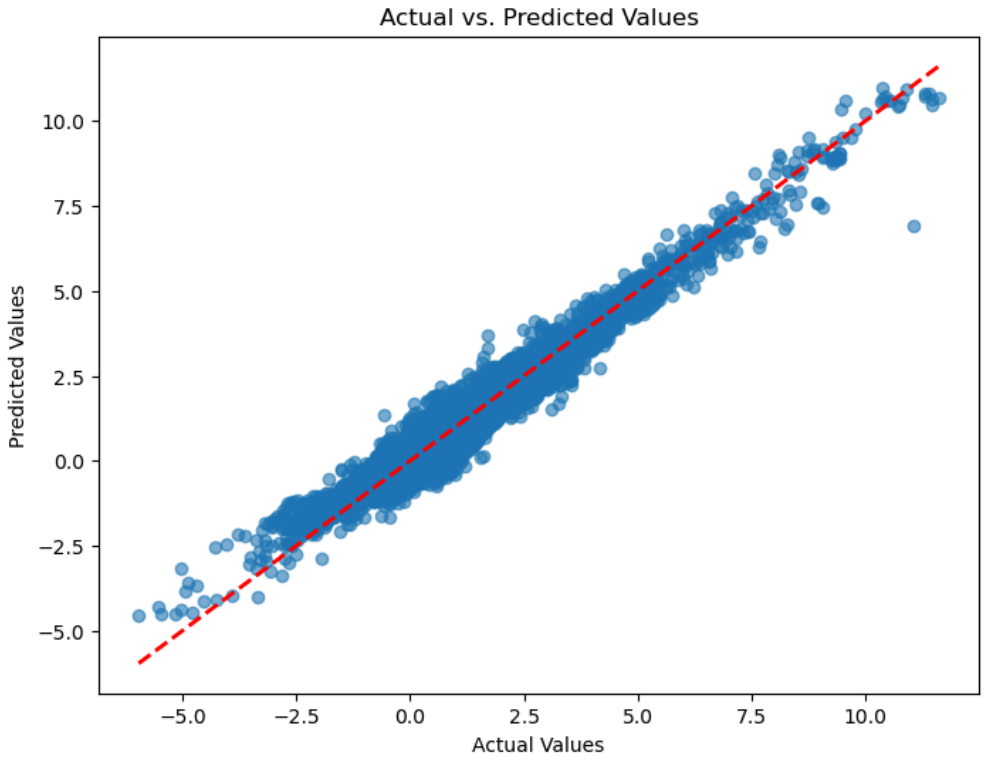
8.3.1 Model Performance

As we can see in the figure 8.3 , the model gives a good R-squared value of 95.52%. It means that approximately 95.52% of the variance in the target variable (PopGrowthRate) is explained by the Gradient Boost regression model and the model is a very strong fit.
Since the adjusted R-squared value and R-squared value are the same it explains that there has been no overfitting in the model and the model complexity is justified.

8.3.2 Error Metrics

A mean squared error (MSE) of 0.0789 means that the squared difference between the actual and predicted values is very less and the model performs well.
A mean absolute error (MAE) of 0.2092 means that on average the absolute error between the actual and predicted values is low and the model performs well.
The root mean squared error (RMSE) of 0.2808 implies that the average prediction error is about 0.2808 units.

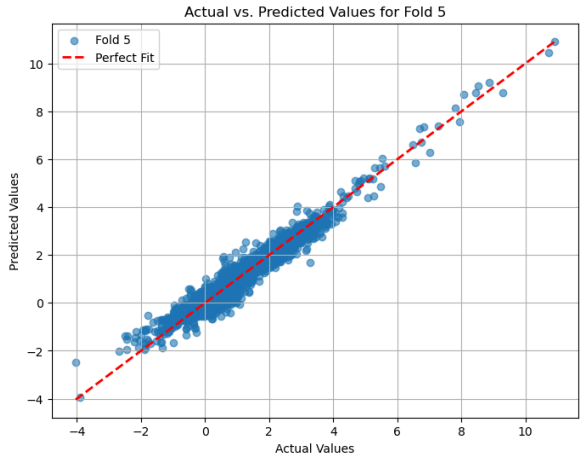
8.3.3 Graphical Interpretation of Actual vs Predicted values



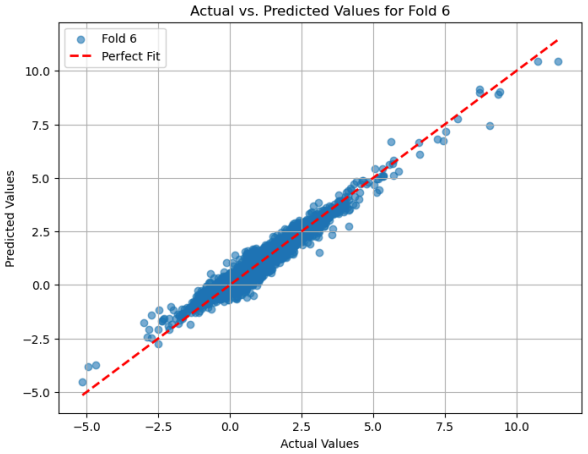
Graph 8.3

In this graph 8.3 we see that the scatter points of the predicted values are clustered almost perfectly around the line of best fit without much discrepancies. Only a few observations are deviating away

from the line of best fit indicating that the model has done a good job of predicting values. This explains the high value of R-square and low values of error terms.



Graph 8.4



Graph 8.5

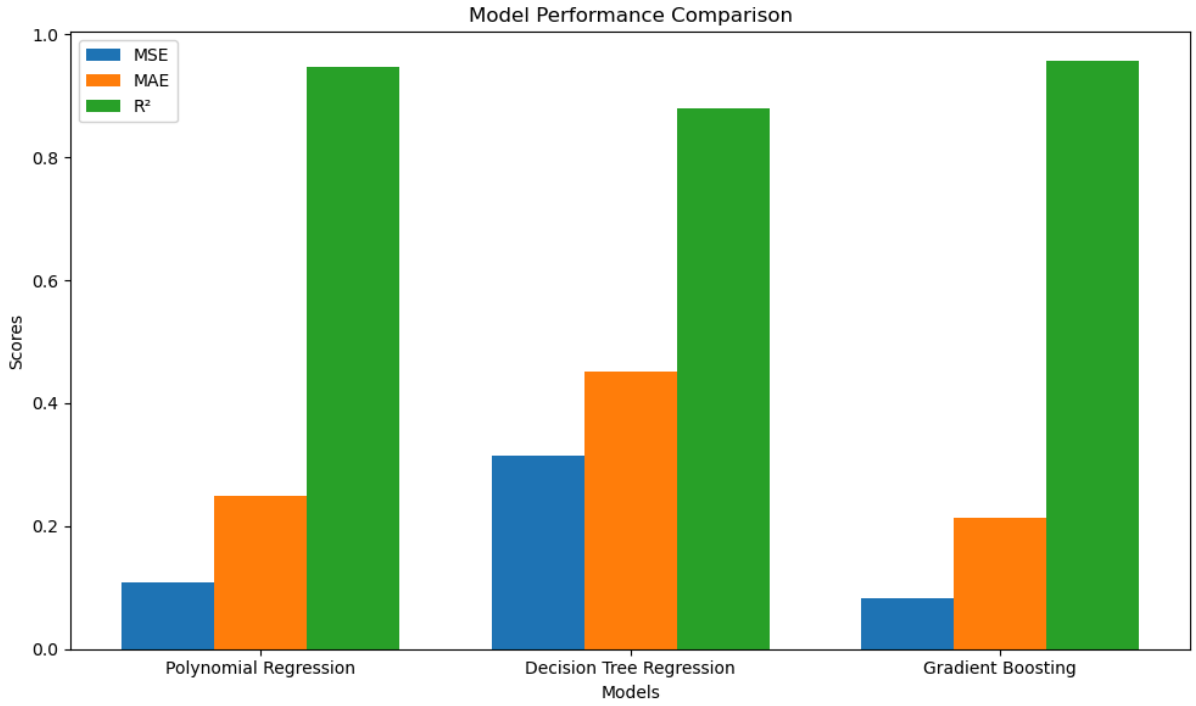
Graph 8.4 and 8.5 show us how the gradient boosting algorithm works in two different consecutive folds of cross validation. As we can see, in the 5th fold the scatter points at the higher end are more scattered and away from the line of best fit ,while in the 6th fold the scatter points are much more close to the line of best fit . This is because the predictions of the gradient boosting algorithm keep getting better after every iteration which is clearly visible from figures _ and _.

8.4 Comparison between the 3 Regression models

Model Name	**Polynomial Regression**	**Decision Tree Regression**	**Gradient Boosting Regression**
MSE	0.1086	0.2119	0.0789
MAE	0.2499	0.3461	0.2092
RMSE	0.3295	0.4603	0.2808
R-square	93.84%	87.97%	95.52%
Adjusted R-square	93.84%	87.96%	95.52%

After comparing all the three regression models ,Gradient Boosting Regression turns out to be the most robust and accurate model. The highest values of R-squared and adjusted R-square indicate a superior fit to the data ,while also maintaining lowest error metrics (MSE, MAE , RMSE). Overall Gradient Boosting Regression effectively captures the underlying patterns and trends in the data and at the same time avoids large errors.
Polynomial regression also performs well ,particularly in capturing large proportions of variance in the data with relatively low error rates. However the main difference between Gradient Boosting and Polynomial regression comes from the Root means square error (RMSE) where the higher value RMSE of polynomial regression indicates its proneness to large errors.
Decision Tree Regression which is one of the most simpleset regression and has wide scale

applications in different contexts does not perform that well as compared to the other two models. It captures less variance and has higher sensitivity towards error metrics indicating that the predictions are less accurate and more susceptible to outliers and noise in the data.



Graph 8.6

In predicting the population growth rate there are several factors that play a vital role like the uncertain fluctuations in demographic indicators which may be caused due to wars, pandemic or any government activity, and hence the model should be good enough to identify these fluctuations and work on it. Gradient Boosting model makes use of various low performance models and then adds up all to give the best one which is what fits the population growth rate prediction scenario well and hence it performs the best. As it is clearly visible from the Graph 8.6 that in all aspects the gradient boosting model performs well while the polynomial regression model also gets close by being the second best model.

8.5 Future Population Growth Rates Prediction based on the best fitting model

After performing regression analysis over the data we get Gradient Boosting Regression to be the best model for population growth rate predictions. Now further we try to use this model and make predictions for the next 10 years.
Our approach follows a simple method where we will be using time series prediction of data for different locations using where the predictions involve re-training the model iteratively using both historical data and newly predicted data to forecast future population growth rates.
For this process first we filter out our data until the year 2100 as historical data. After this we define the time period of future years for which we want to make our predictions on which is 2101 to 2111. In our dataset we have years and its population growth rate values for every unique location id, hence it would make sense to predict the values for every unique location id for the next 10 years. For this we filter out our data based on unique location ids and then the code iterates over each year from 2101 to 2111,updating ‘Time’ and ‘LocID’ fields in the current data for the specific year and location. After re-training, the model(Gradient Boosting Regression) predicts the population growth rate for the

current year using the most recent data. Once it has done the prediction then the newly predicted value is stored and used as a part of historical data for predicting the next year growth rate .
Time series data benefit from this continual updating procedure since the most recent trends the model has captured can affect future projections.The model adjusts to any new trends or patterns in the data by retraining itself at each stage, guaranteeing that forecasts stay precise and indicative of possible future events.

	LocID	Time	Predicted_PopGrowthRate
0	4	2101	0.239728
1	4	2102	0.247726
2	4	2103	0.249531
3	4	2104	0.250756
4	4	2105	0.255254
5	4	2106	0.254315
6	4	2107	0.255486
7	4	2108	0.257041
8	4	2109	0.258251
9	4	2110	0.257452
10	8	2101	-2.395604
11	8	2102	-2.395152
12	8	2103	-2.394045
13	8	2104	-2.393762
14	8	2105	-2.393521
15	8	2106	-2.392680
16	8	2107	-2.392898
17	8	2108	-2.392811
18	8	2109	-2.392580
19	8	2110	-2.392177
20	12	2101	-0.075260
21	12	2102	-0.073420
22	12	2103	-0.070694
23	12	2104	-0.070423
24	12	2105	-0.070961
25	12	2106	-0.069951
26	12	2107	-0.069402
27	12	2108	-0.069567
28	12	2109	-0.069422
29	12	2110	-0.068998

Figure 8.4

The above figure 8.4 shows us the predictions for the next 10 years for three locations. We see how the model has incorporated the historical data and also made the predicted values as historical data and given us the results. This approach of using the best fit model and time series prediction will help nations and government bodies to take crucial steps for the benefit of the population. By knowing whether there will be a rise or drop in the population a better infrastructure planning can be made and good development of the population can take place.

8.5.1 Detailed Analysis on Specific Locations

a. Location 4 :

Location 4 is on the path of modest but stable population growth, indicating potential scope in development and expansion. The population growth for this location will face only a minor dip in the final year 2110 and apart from that it is going to increase. This upward trend might indicate various things such as improved living conditions ,better medical facilities ,good economic opportunities and favourable demographic patterns for this location. The government and local bodies can plan for

infrastructure expansion , housing development and development in transport facilities to accommodate the growing population.

b. Location 8 :

Location 8 shows a consistent negative population growth rate, starting at -2.395 in 2101 and slightly improving by the year 2110. The continuous negative growth rate of this location might indicate that the location is facing various significant challenges like economic decline, high emigration rates ,decline in birth rates etc leading to a steady population decay over the decades. The governments and local bodies for this location can target economic revitalization , attracting new residents , some changes in tax systems and addressing the root cause of outmigration or low birth rates.

c. Location 12 :

For Location 12 the trend in the growth rate is more flat with minor fluctuations in negative form. The decline rate is slow and stabilising over the years .This could point to less severe underlying issues than those affecting Location 8 that are producing a slow population drop, like ageing populations or outmigration. The goal of population stabilisation could be achieved by enticing newcomers, enhancing living conditions, or offering incentives to families.

9. Model comparison with previous researches

In this section we would compare our results and findings of the model with some previous research papers and studies to better understand how our project has done in terms of model performance , results analysis and other important aspects. As in our literature review we saw the machine learning performance of 17 different machine learning models out of which we made use of one same machine learning technique that is decision tree regression and the other two were different as compared to the 17 regression model used. We would now like to compare and see where our findings and results stand with respect to previous studies.
The table 9.1 is based on the results from a previous research which is the work of Mohammad Mahmood Otoom and is taken from the research of machine learning for 17 different models based on various countries.

Models Used in This Project	**Polynomial Regression**	**Decision Tree Regression**	**Gradient Boosting Regression**
Overall Performance (R- squared)	94%	88%	96%
Top 3 models from previous research	**Random Forest Regression**	**Decision Tree Regression**	**K Nearest Neighbours Regression**
Overall Performance (R-squared)	96%	94%	92%

Table 9.1

From table 9.1 we see the comparison between our models and previous models. Here the decision tree algorithm in our project lags behind as compared to the decision tree in previous research. The other two models polynomial and gradient Boost perform equally well as compared to random forest and KNN models. Our aim was to use models different from the traditionally used models and by this comparison we can say that our 3 models can be used in real life applications.

10 . Conclusion

This Project sheds some light on how demographic data plays a vital role in population prediction and how it can be better utilised than the census data. We start by reviewing various important research papers that have been published on population growth prediction and use of machine learning in them. We then carefully took account of the information from the researchers and then created a plan for our research study. In the next step we created our model based on which we were gonna implement our models and that was Cohort Component Technique where we figured out our four main features that were important for the population growth prediction(Population at present time P , Fertility rate F, Mortality rate M and net migrations NM).

Our data exploration gave us various meaningful insights such as the importance of locations and time in the data. The various relationships between mortality ,fertility and migration were shown using graphs. We then pre processed and cleaned the data according to our project requirements and used various statistical techniques like Z-score evaluation for outlier detection , handling missing values and much more.

Next we laid a foundation to select the demographic indicators that would best fit our models and hence we used correlation analysis for this purpose. We took into account Pearsons and Spearman's rank correlation to figure out our best indicators. Since these results were different for different locations we decided to use the technique of weighted average of the correlations which in the end helped us get our final predictor variables.

For model selection we ventured through various articles and papers and made use of the algorithms that would suit our research and at the same time were unique to population prediction. This gave us our three algorithms that were Polynomial Regression , Decision tree Regression and Gradient Boosting Regression. Under polynomial regression we made use of degree 2 polynomials to avoid overfitting and get the best results. Decision tree was adjusted to a max depth of 5. Out of all the algorithms we used, Gradient Boost turned out to best fit into the data and gave good results and predictions. For the models we had to make use of a cross validation method. The main purpose of choosing cross validation over train-test split was that the data was segregated into years for all the different locations and to properly access the whole dataset while shuffling the data randomly at every fold ,cross validation proved to be the right choice. By evaluating the effectiveness of different machine learning algorithms further, we were able to identify several benefits and areas for improvement. Certain algorithms performed better at capturing linear relationships than others at managing non-linear complications.

To sum up, our study highlights the importance of methodological diversity in addition to the usefulness of machine learning in demographic research. The thorough understanding that is produced by the insights gained from each technique will aid future research in this field.

11. Scope of Improvement in models

While the three models in our project have shown successful results and predictions there is always a scope of further improvement. Here are several things that can be used to enhance our results.

1. Feature Engineering and selection:

   Use of more Predictor variables : In our project we made use of predictors that were based on high correlation with the target variable. While we eliminated some of them which still had high correlation with the target variable because of inter correlation and discrepancies. While we focused mainly on the indicators that were based on our population growth model we can focus on indicators referring to various other factors like economic indicators (like migration rates) and health based indicators like life expectancy and its several forms.

   Improvement in polynomial features: While we made use of degree 2 polynomial regression ,experimenting with higher degree polynomials or adding more interacting terms could capture more complex and meaningful results from the data.

   Dimensionality Reduction : We could make use of techniques like Principal component analysis (PCA) which could be implemented on the data for tree based modelling which inturn would help us in reducing the dimensionality and complexity of the data.

2. Cross Validation and evaluation :

   While we used cross validation with a specific amount of k-folds ,we can experiment by using a larger amount of folds and see how the results might change over the folds. This would help focus more on the unseen parts of the data and help in better understanding the data for the models.

3. Feature Important analysis :

   For the models decision tree and gradient boost a deeper analysis of feature importance can help in understanding which indicators drive the predictions and help in refining the models.

4. Implementation of other linear and non linear techniques :

   While the current research used Decision tree and Gradient Boost to simulate interactions, future studies may examine other alternative non-linear and linear modelling techniques. Experimenting with various non-linear and linear models can help better capture the subtleties of population dynamics and improve the prediction capability of the models.

5. Focusing on more Geographical based analysis :

   As we have seen while calculating the correlation(7.4) and predicting the future growth rate (8.5) that every location acts differently and hence making predictions based on location based information might lead us to better results. This can imply targeting a specific country or a continent and then comparing the results with other countries and continents.

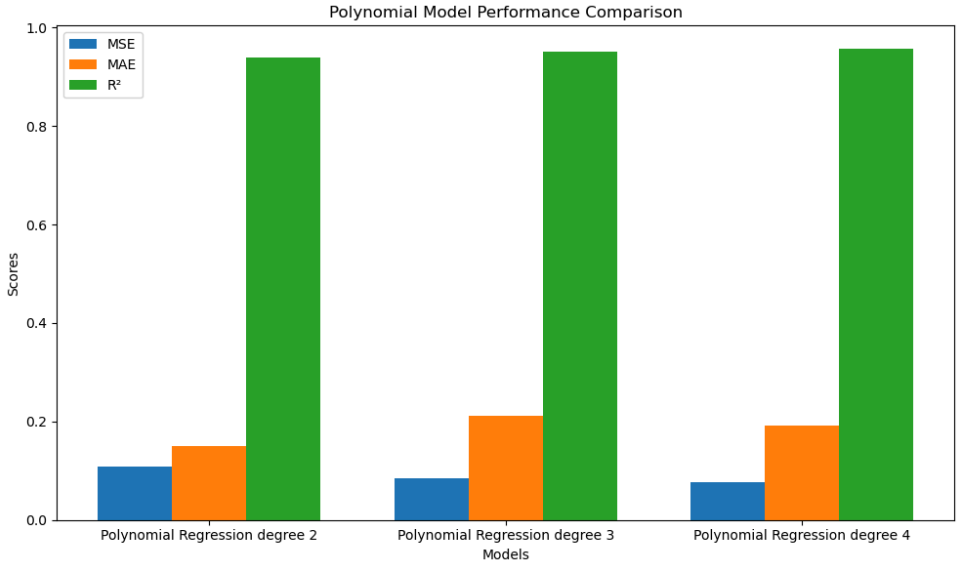
6. Real time Forecasting models :

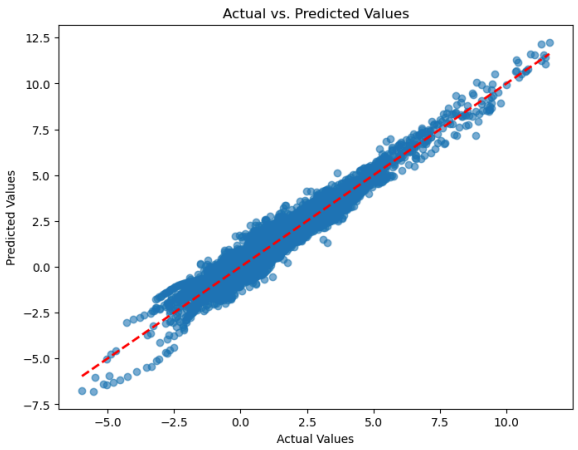
   An important scope for future work is predicting the population growth in real time. Real time analysis is important because there are continuous changes that take place like unexpected demographic patterns, unexpected situations like pandemic (Covid-19) or wars , rise and fall in immigrations and emigrations of specific regions etc. Policymakers and planners can

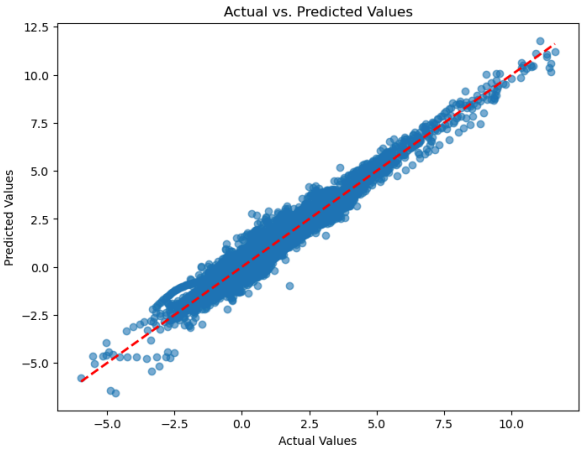
benefit from real-time forecasting models' timely insights that allow them to take proactive measures in response to emerging demographic trends and challenges.

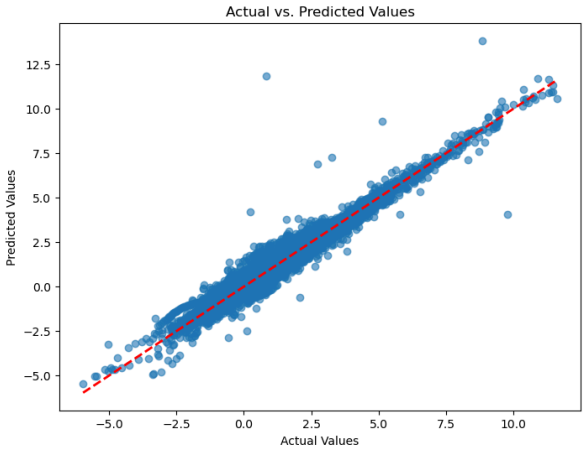
Future versions of the population growth rate prediction models may be even more accurate and resilient if these possible areas for development are investigated. This will improve the models' capacity for prediction while also offering greater understanding of the fundamental causes of population shifts, facilitating the formulation of more sensible policies and decisions.

12. Polynomial Regression Model Performance with 2,3,4 degrees of polynomial

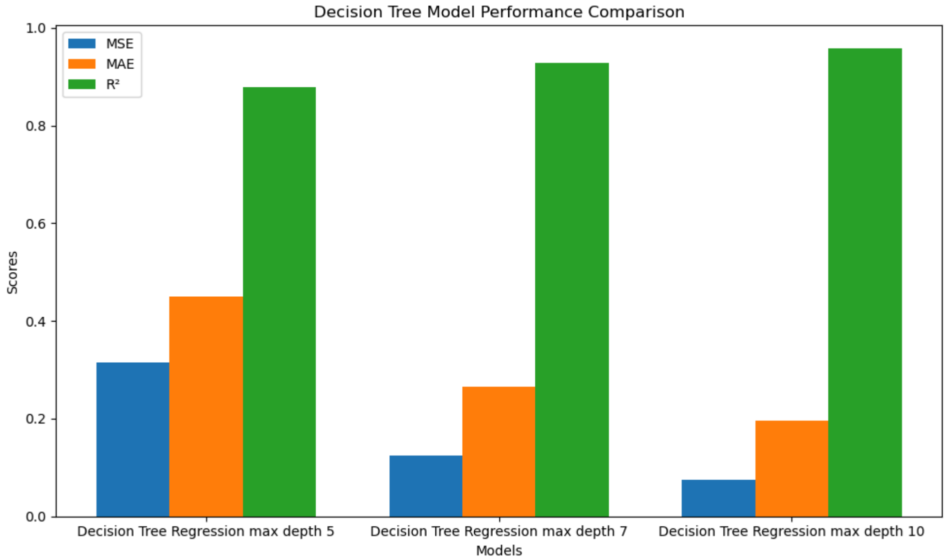


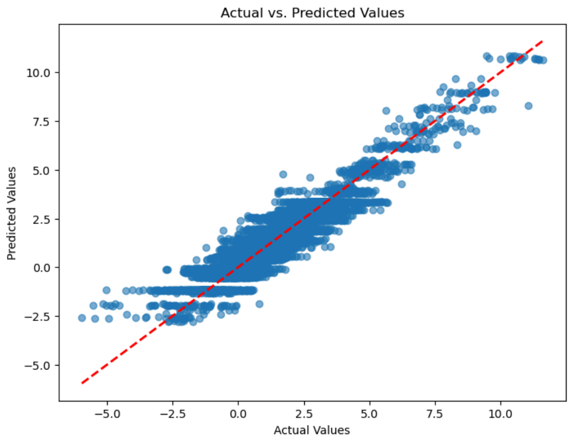


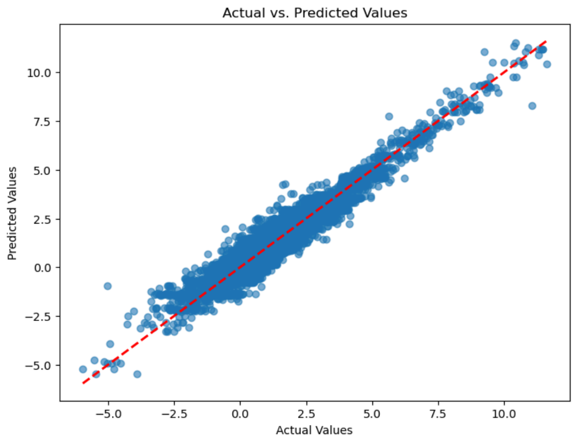


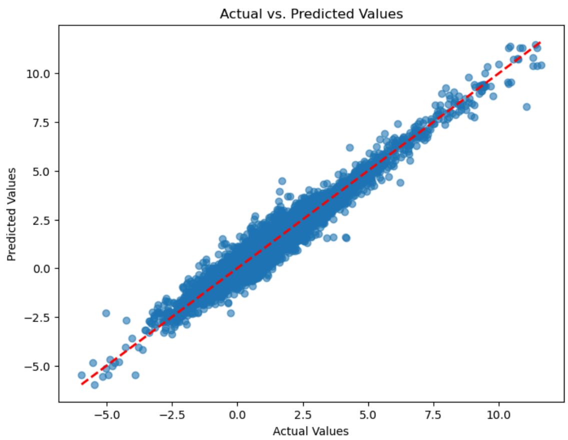


13. Decision Tree Regression model performance with 5,7,10 max depths









14. References :

Dataset :

https://population.un.org/wpp/Download/Files/1_Indicator%20(Standard)/CSV_FILES/WPP2024_Demographic_Indicators_Medium.csv.gz

Section 5.2

Tabata, M., Eshima, N., & Takagi, I. (2010). A mathematical-model approach to human population explosions caused by migration. *Nonlinear Analysis: Real World Applications*, 11(5), 4027-4042. https://doi.org/10.1016/j.nonrwa.2010.03.009

Section 5.3

Hyndman, R. J., & Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3), 323-342. https://doi.org/10.1016/j.ijforecast.2008.02.009

Section 5.4

Otoom, M. M., Jemmali, M., Qawqzeh, Y., Nazim S. A., K., & Al Fayez, F. (2019). Comparative analysis of different machine learning models for estimating the population growth rate in data-limited areas. *International Journal of Computer Science and Network Security (IJCSNS)*, 19(12).

Otoom, M. M. (2021). Comparing the Performance of 17 Machine Learning Models in Predicting Human Population Growth of Countries. *International Journal of Computer Science & Network Security*, *21*(1), 220-5.

S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, Jian, China, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.

Section 5.5

Ormiston-Smith, N., Smith, J., & Whitworth, A. (2006). An international comparative study on the use of the Cohort Component Method for estimating national populations. *Population Trends*, 125, 37.
Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308-324. https://doi.org/10.1016/j.trc.2015.02.019

Section 7.2

68–95–99.7 rule. (n.d.). In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

Image Source : https://en.wikipedia.org/wiki/File:Empirical_rule_histogram.svg
Shiffler, R. E. (1988). Maximum Z scores and outliers. *The American Statistician*, 42(1), 79-80. https://doi.org/10.1080/00031305.1988.10475530

Section 7.3

Inaba, H. (2009). The net reproduction rate and the type-reproduction number in multiregional demography. *Vienna Yearbook of Population Research*, 7, 197–215. http://www.jstor.org/stable/23025529

Section 7.3.1

Ünal, C., & Özel, G. (2023). A comparison of statistical distributions for the crude birth rate data. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 22(44), 281-291. https://doi.org/10.55071/ticaretfbd.1277633

Section 7.4

Duarte, F. S. L. G., Rios, R. A., Hruschka, E. R., & de Mello, R. F. (2019). Decomposing time series into deterministic and stochastic influences: A survey. *Digital Signal Processing*, 95, 102582. https://doi.org/10.1016/j.dsp.2019.102582

Rovetta, A. (2020). Raiders of the lost correlation: A guide on using Pearson and Spearman coefficients to detect hidden correlations in medical sciences. *Cureus*, 12(11), e11794. https://doi.org/10.7759/cureus.11794

Section 7.6

Berrar, D. (2019). Cross-validation. *In Encyclopedia of Bioinformatics and Computational Biology* (Vol. 1). Elsevier. Retrieved from http://berrar.com/resources/Berrar_EBCB_2nd_edition_Cross-validation_preprint.pdf

Section 7.7.1

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering*, 48, 500-506. https://doi.org/10.1016/j.proeng.2012.09.545

Image source : https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18

Section 7.7.2

(n.d.). Image source: *Artificial Intelligence Digest Facebook page*. Retrieved from https://www.facebook.com/artificialintelligence.digest/posts/233496684209361/

Section 7.7.3

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308-324. https://doi.org/10.1016/j.trc.2015.02.019

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5, 197–227. https://doi.org/10.1007/BF00116037