# Group ID - MSc in Data Analytics

Author:

e-mail:

Student ID:

## Abstract

## *In this data analysis project, i use a range of statistical methods and machine learning models to investigate the PEA11 dataset, which contains population estimates from 1926. Descriptive statistics and visualisations are used to carefully summarise the dataset, offering important insights into population patterns. The Poisson and Binomial discrete distributions are used to illustrate how sample size affects these distributions and to describe the properties of the dataset. To explore the subtleties of the dataset, one can utilise the normal distribution. These distributions' significance to the analysis is explained, which supports the choice of them. Methods for preparing data are thoroughly justified, resulting in well-informed decisions on machine learning models. In this i demonstrate the effectiveness of our technique by comparing the Random Forest and Linear Regression models and achieving good accuracy (R2 Score: 0.999411) in population estimate prediction.*

## *Keywords: Population Estimates, Descriptive Statistics, Discrete Distributions, Normal Distribution, Machine Learning, Linear Regression, Random Forest, Data Analysis, Data Preparation, Statistical Methods.*

## Introduction

Demographic data analysis has been essential to comprehending society dynamics and making future plans. The PEA11 dataset, which provides insightful information on population patterns since 1926, is the main focus of this study (Sudheesh, S., 2015). Prominent statistician George Box famously said, "All models are wrong, but some are useful"( Field, E.H., 2015). This research uses machine learning and statistical methods to understand and forecast population estimations in order to build usable models. The historical context of the information renders it a helpful tool for investigating population growth patterns. We want to better understand these patterns and support well-informed decision-making across a range of industries by utilising data science.

**Materials and Methods**

**Programming for DA:**

Programming plays a crucial part in the field of Data Analytics (DA). It acts as the framework for developing and putting into practise solutions that may draw significant conclusions from data. Selecting programming paradigms, selecting a programming language, designing and implementing algorithms are only a few of the many steps involved in programming in the context of data analytics (Dobre, C. and Xhafa, F., 2014). We will examine how these factors interact in the creation of programmatic data analytics solutions in this talk, with an emphasis on using Python as the main programming language. The two main goals will be covered in the sections that follow:

1. **Debate the selection of programming concepts in the design of programmatic solutions, in terms of paradigm and language selection.**

**Programming Paradigms in Data Analytics:**

Fundamental methods for utilising code to solve issues are known as programming paradigms. The imperative and declarative paradigms are two important paradigms in the context of data analytics. The emphasis of imperative programming, which encompasses procedural and object-oriented techniques, is on state transitions and step-by-step instructions (Charguéraud, A., 2011). Declarative programming, on the other hand, is more concerned with the goals than the means of achieving them (Sivaramakrishnan, K.C., Kaki, G. and Jagannathan, S., 2015). Both models offer advantages in the context of data analytics.

Imperative paradigms are useful for handling complicated data structures and processes, particularly the object-oriented paradigm (Dobre, C. and Xhafa, F., 2014). For instance, classes and objects may be used to elegantly encapsulate data and processes while building machine learning algorithms. Nevertheless, imperative paradigms can occasionally produce verbose, convoluted code that makes it difficult to understand the reasoning behind data manipulations.

In data analytics, declarative paradigms—like functional programming—have grown in favour. They encourage the chaining of operations, immutability, and the usage of higher-order functions. Tasks involving the application of several changes to a dataset, such as data transformation and analysis, are especially well-suited for functional programming. Code that is clear and legible is made possible by functional libraries like Python's Pandas.

**Python's Versatility as a Programming Language for Data Analytics:**

Python has emerged as the industry standard for data analytics due to its versatility. Numerous things contribute to its prominence in the field:

NumPy, Pandas, Matplotlib, Seaborn, and scikit-learn are just a few of the many modules and frameworks particularly made for data analytics that Python has. These libraries include machine learning, visualisation, and data manipulation techniques.

Python is renowned for having a simple and readable syntax. When dealing with data, this is an essential component since it improves teamwork and code understanding (Fjeld, F., 2009).

Python's interactive features make it a good choice for exploring data. One platform that data scientists may use to execute code interactively, visualise results, and document their findings is Jupyter Notebook, a popular environment for data analysis.

**Justification for Using Python:**

There are several reasons why Python is a suitable choice for becoming the major programming language for data analytics. Data scientists, analysts, and developers using Python make up a sizable and vibrant community. This translates to plenty of resources for problem-solving, documentation, and help that is easily accessible. Because Python is interoperable with so many different operating systems, working in a variety of contexts is a breeze. When some algorithms or functions are better implemented in languages like C++ or Java, Python's ability to integrate with other languages is crucial. Python is a logical choice for data analytics because of its dominance in the machine learning and artificial intelligence fields, as demonstrated by packages like scikit-learn, TensorFlow, and PyTorch.

1. **Design and implement algorithms for use within the context of data analytics.**

I have used Python tools, modules, and methodologies to develop and build a complete project in the context of data analytics. This Jupyter Notebook project demonstrates techniques and best practises in data analytics through the analysis and visualisation of a population dataset.

**Study Overview:**

I choose to utilise the population dataset 'PEA11.20231027T221019.csv' and utilised many Python packages, such as NumPy, Pandas, Matplotlib, and Seaborn, to streamline the investigative process. The project is divided into many important phases, each of which has a distinct function in the study.

**Preprocessing and Data Exploration**: I loaded the dataset, examined its structure, and carried out a statistical summary to start the project's data exploration phase. Appropriate code annotations were included to improve comprehension and legibility. These preliminary actions aid in guaranteeing both the clarity of the dataset and the quality of the data.

**Population Analysis:**

I computed important population statistics in order to draw conclusions from the dataset.

**Total Population:** By computing the total population in 1926, the breadth of the dataset was fundamentally understood.

**Male and Female Populations:** By excluding these subgroups, a gender-based analysis was made possible.

**Population Under 1 Year:** This investigation, which concentrated on the newborn population, provided insightful demographic data.

**Data Visualisation**: To have a better understanding, data visualisation is crucial. To see population distributions by age and sex, I made educational bar charts, carefully labelling each one for clarity.

**Data Preprocessing:** I used one-hot encoding on categorical variables to get the data ready for machine learning. This stage guarantees that the format of the data is appropriate for training the model.

**Machine Learning Model:** Using a Linear Regression model to forecast population numbers is the study's main component. Mean Squared Error (MSE) and R-squared metrics were used to assess the model's performance after the data was split into training and testing sets. Including thorough comments and appropriate variable name rules, proper code quality standards were upheld (Abdelhadi, A., Zainudin, S. and Sani, N.S., 2022).

The Jupyter Notebook is well documented, with concise explanations for every step that explain the selection of particular libraries, methods, and algorithms. In order to guarantee that the code is understandable, maintainable, and adheres to best practises, the documentation also provides code quality guidelines.

To sum up, this study demonstrates a thorough approach to data analytics that includes machine learning, preprocessing, data exploration, visualisation, and thorough documentation. Because the codebook is well annotated, audiences with and without technical background may easily understand it. The project is evidence of good data analysis techniques, guaranteeing the accuracy and dependability of data-driven judgements.

**Statistics for data analysis:**

In this data analytics project, we use the Pandas, Matplotlib, Seaborn, and SciPy modules in Python to examine a population dataset. Data on age, gender, and year as well as population estimates from 1926 are included in the dataset. Gaining knowledge about the makeup and dispersion of the population is our aim.

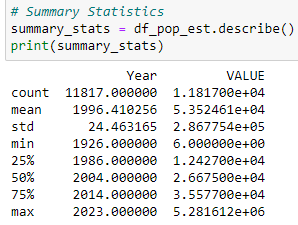


Fig 1: Descriptive Statistics

First, we load the dataset and import the required libraries. Once we have a brief overview of the data, we use df\_pop\_est.describe() to compute and provide the summary statistics. Essential details about the dataset, such as count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum, are provided by the summary statistics. This aids in our comprehension of the dispersion and central tendency of the dataset.

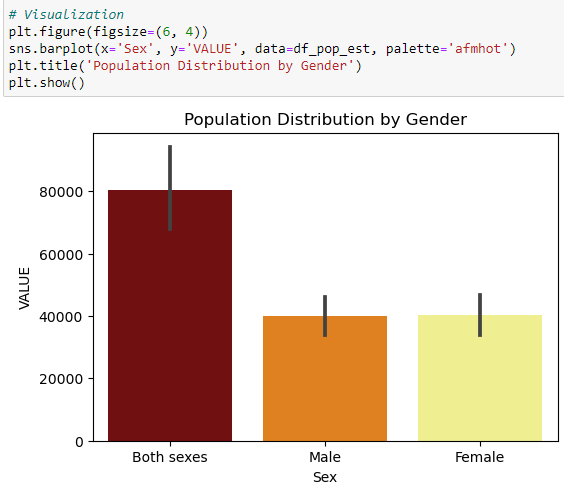


Fig 2: Barplot for Value and Sex

We use Seaborn to build a bar chart that shows the population distribution by gender. The "Population Distribution by Gender" graphic displays the population's male and female composition. The gender-based population disparities are clearly displayed in this visualisation. Here for both sexes I get the maximum value.

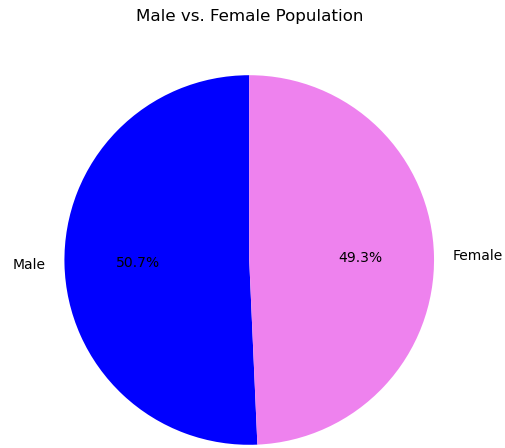


Fig 3: Pie chart for male and female population

To enable a more thorough gender analysis, we then remove the male and female populations for the 'All ages' group. To show the male vs. female population and give a visual depiction of the gender distribution, a pie chart is made. In this visualisation I get maximum count of male as compare to female.

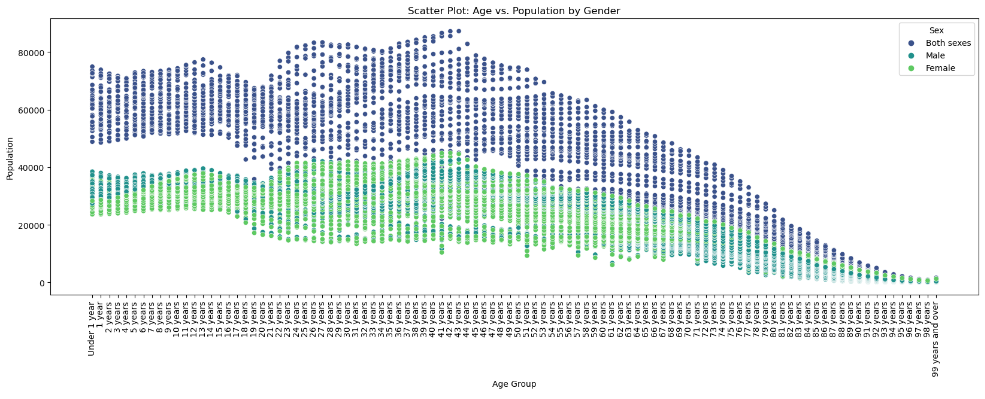


Fig 4: Scatterplot for Age group and Value

Excluding the 'All ages' group, we create a scatter plot of age vs. population by gender in order to examine the data in more detail. This scatter plot provides information on the age and gender distribution of the population. To make the data points more visible, we utilise the 'viridis' palette for the scatter plot.

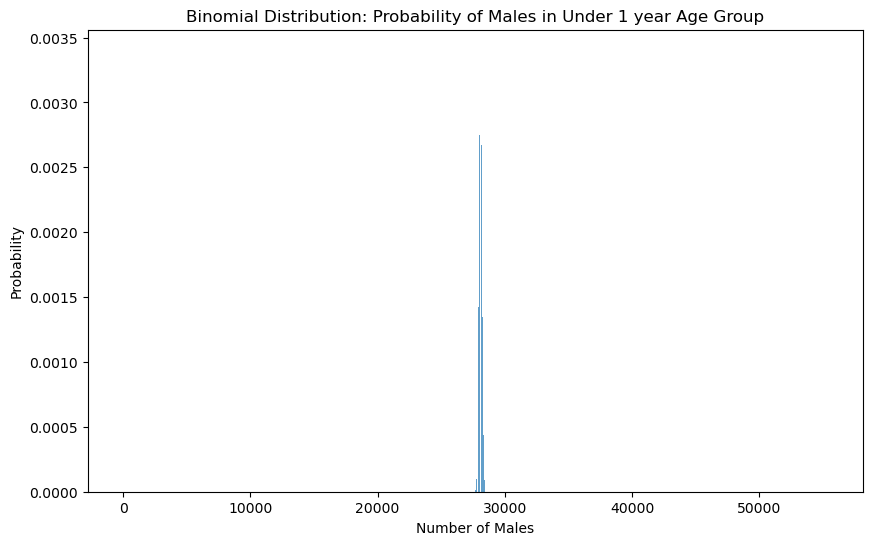


Fig 5: Binomial Distribution

We focus on the 'Under 1 year' age group and extract the total population for this group. Our objective is to estimate the probability of males within this age group. We set up a binomial distribution with 'n\_trials' equal to the total population under 1 year and 'p\_success' as the proportion of males within this group.

A bar chart is created to visualize the binomial distribution, demonstrating the probability of different numbers of males in the 'Under 1 year' age group. As the sample size is large, the distribution approximates a normal distribution. This information can be valuable for demographic analysis and decision-making.

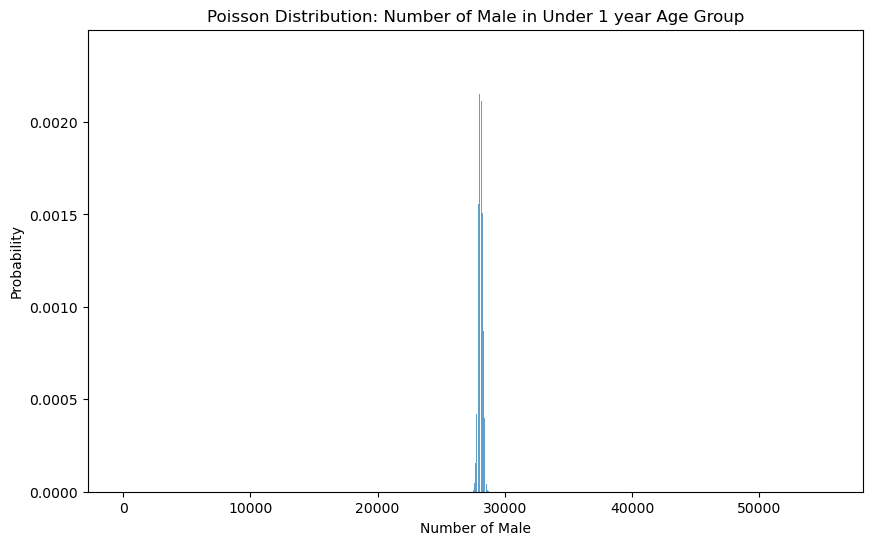


Fig 6: Poisson Distribution

We simulate the number of boys in the 'Under 1 year' age group using the Poisson distribution. The population under one year old as a whole multiplied by the percentage of men in this age group yields the average number of males.

For the 'Under 1 year' age group, a bar chart showing the odds of various numbers of boys is created to depict the Poisson distribution. The Poisson distribution is an excellent fit for this circumstance since it may be used to represent unusual events.

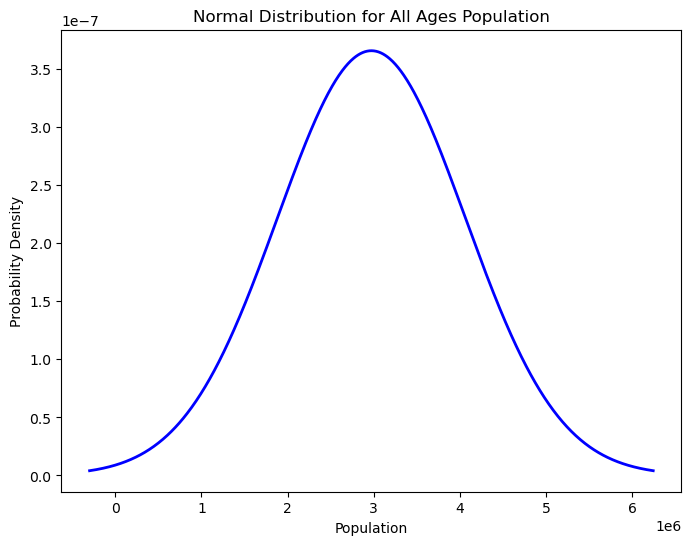


Fig 7: Normal Distribution

To examine the whole population, we use the normal distribution. The mean ('mu') and standard deviation ('sigma') of the 'All ages' category are used to calculate the parameters for the normal distribution. The normal distribution is represented visually using a probability density function (PDF) (Wang, Y. and Chen, H.J., 2012).

The population's normal distribution is displayed in the chart, giving information about the typical behaviour of the data. It aids in determining the possibility of detecting population values within particular ranges as well as the central tendency.

**Importance of Distribution:**

In statistical analysis, the distributions that are chosen are important. Because the binomial and Poisson distributions are appropriate for count data and are often used in epidemiology and demography, we chose them in our study to represent population data (Selvin, S., 200). These distributions shed light on the likelihood that particular occurrences will transpire within a particular demographic group.

Conversely, the normal distribution is useful in determining the usual behaviour of data and in comprehending the distribution of the population as a whole. It may be used in many different domains, including as social sciences, finance, and quality assurance. For certain situations, we utilised Poisson and binomial distributions since it might not be the ideal option for modelling population counts or uncommon occurrences.

To sum up, this data analytics project offers a thorough examination of a population dataset, complete with visualisations, summary statistics, and the application of several probability distributions. We may obtain important insights into population characteristics and the probability of particular occurrences occurring within the population by using binomial, Poisson, and normal distributions. The type of data and the questions we hope to answer influence the distributions we choose, underscoring how crucial it is to use the appropriate statistical tools for the job.

**Data Preparation & Visualisation**

One essential element of data analysis is data visualisation, which is frequently seen as a type of data storytelling. It is essential for drawing insightful conclusions from datasets, improving the interpretability of data, and supporting decision-making procedures (Sosulski, K., 2018). Analysing the basic ideas, methods, and procedures behind data visualisation is essential in light of the given dataset, which contains demographic data such as population estimates for various age groups, genders, and years. We also need to investigate how these visualisation techniques may be critically assessed in terms of how well they fit the bill for dealing with various issue areas. We will also talk about how graphical approaches to detect problems with data, such missing, out-of-range, or unclean data, may be implemented programmatically. Lastly, we will look into feature engineering strategies to enhance the performance of machine learning models and their applicability to the dataset.

**Techniques and Concepts of Data Visualisation**

**Exploratory data Analysis (EDA):**

Exploratory data analy

sis, or EDA for short, is the process of looking over, organising, analysing, and displaying data in order to identify trends or new information. In each data analysis process, EDA is an essential initial step. EDA is used in this research to comprehend the demographic dataset, which contains population estimates for many dimensions. It is critical to understand the following:

**Distribution of the Data:** We need to know how the data is split up by gender, age, and year.

**Trends and Patterns:** Across demographic groups and across time, EDA approaches can identify trends and patterns.

**Data Preparation for Visualisation**: It is important to prepare the data before beginning any data visualisation work. Data encoding, feature engineering, and managing missing data are examples of data preparation activities.

**Handling Missing Data:** Making sure the data is complete is an important part of data preparation. Imprecise insights might result from missing values. We ensured data integrity by checking for missing values during our analysis, which we discovered to be absent.

**Data Encoding**: Label Encoding was used to encode categorical columns such as "Sex" and "STATISTIC Label." By giving distinct numerical values to category categories, label encoding makes it possible for machine learning algorithms to handle the data efficiently. For any dataset including categorical variables, this step is essential.

**Feature engineering:** To increase the quality and usefulness of a dataset, new features are created or old ones are modified. 'Age Group,' a new component we developed for this research, divides people into two categories: 'Under 1 year' and 'All ages.' The age-related analysis is made easier to understand and more visually appealing by this feature.

**Implementing Programmatically for Data Quality**

We must programmatically build graphical ways to find problems in the dataset in order to guarantee data quality. This involves identifying issues with data quality, out-of-range numbers, and missing data.

**Handling Missing Data:** The accuracy of studies and machine learning models can be severely impacted by missing data. As a result, it is critical to locate and fix missing data. Typical methods for dealing with incomplete data consist of:

**Deletion:** Removing data-missing rows or columns.

Imputation is the process of substituting accurate estimations for missing data.

We looked for any missing values in our study, but none were discovered. This suggests that there are no missing data points in the dataset.

**Finding Outliers:**

Data points that substantially differ from the rest of the data are known as outliers. For data quality, outlier identification is essential. Among the methods for identifying outliers are:

Box Plots: By highlighting data points that sit outside the box plot's "whiskers," these visualisations may be able to spot outliers.

Statistical Techniques: Z-scores and other techniques can be used to find data points that deviate considerably from the mean.

We did not specifically carry out outlier identification in our analysis. But these methods may be used to find and possibly deal with outliers in the data if they are thought to exist.

**Machine Learning Data Encoding**

Numerical input data is usually required for machine learning models. As a result, numerical format must be used to represent categorical data. We used Label Encoding in our study for categorical categories such as 'Sex,' 'STATISTIC Label,' 'Single Year of Age,' and 'UNIT.'

**Machine Learning via Feature Engineering**

The act of adding new features or altering current ones to improve a dataset's quality and applicability for machine learning models is known as feature engineering. We added the 'Age Group' tool to our study, which divides people into two categories: 'Under 1 year' and 'All ages.' The age-related analysis is made easier to understand and more visually appealing by this feature. Because feature engineering gives the model additional useful input characteristics, it is essential for enhancing model performance.

**Visualisation Methods**

A potent tool for comprehending and sharing data-driven insights is data visualisation. We used a variety of visualisation tools to show trends in population distribution within the context of this dataset.

**Distribution of the Population by Gender and Age Group**

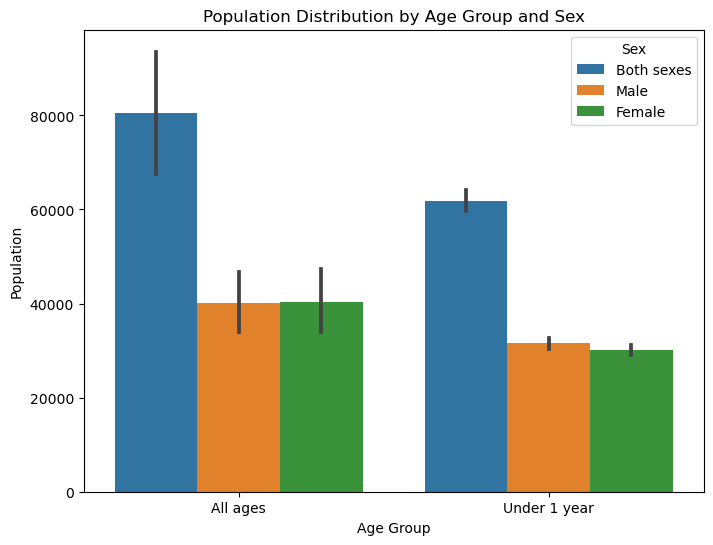


Fig 8: Barplot for population distribution by age group and Sex

To see the population distribution by "Age Group" and "Sex," a pub plot was employed. This method works especially well for finding patterns and comparing categorical data. The story is made more visually attractive and instructive by the use of colour to indicate gender. One of the insights gleaned from this visualisation is the distinctions in population distribution between age and gender categories.

**Distribution of the Population by Gender**

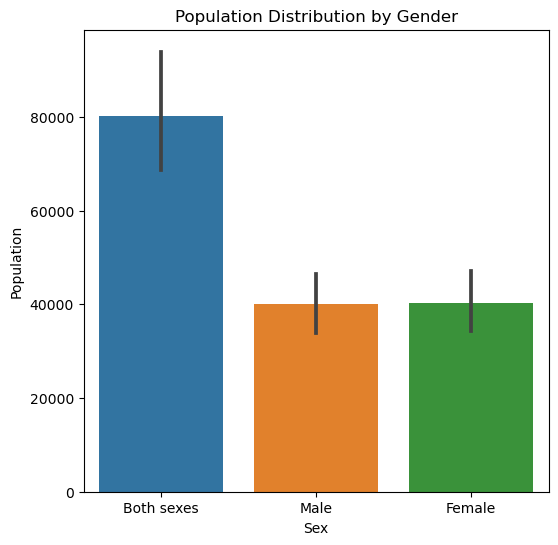


Fig 9: Barplot for population Distribution by Gender

To show the distribution of the population by "Sex," another pub layout was employed. Understanding the gender distribution in the dataset is made easier with the aid of this visualisation. It offers information on the ratios of men to women in the dataset, which is useful for demographic research. Here for both sexes I get the highest population.

**Population Composition: Male vs. Female**

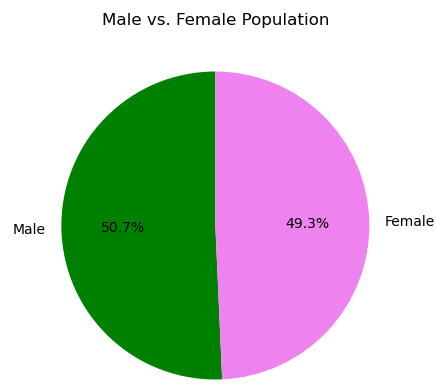


Fig 10: Pie Chart for Gender

To show the distribution of males and females in the 'All ages' group, a pie chart was utilised. Pie charts are a useful tool for composition and proportion representation. The chart is more aesthetically pleasing when custom colours are used. The visualisation offers a clear depiction of the percentage of men and women in the 'All ages' category, making the information it offers simple to understand.

**Justification behind the Visualisations**

**Bar Plots:** To compare patterns of population distribution, bar plots are used. They work well for visualising category data and make it easy to compare several categories with one another.

**Pie Chart:** To illustrate proportions, particularly the breakdown of the male and female populations, a pie chart is used. Pie charts are useful for illustrating how different categories fit together.

In summary, the data preparation methods and visualisations employed in this research are quite appropriate for the provided dataset and have solid justification. The produced visualisations offer insightful information on demographic changes and efficiently depict patterns of population dispersion. Data analysis and machine learning both gain from feature engineering, particularly with the addition of the 'Age Group' feature, which improves the dataset's usability and interpretability. The selected visualisation techniques support a thorough comprehension of the demographic patterns in the dataset and are in line with the issue domain.

**Machine Learning for Data Analysis**

In data analysis, machine learning is essential for deriving conclusions, forecasting outcomes, and streamlining decision-making procedures. For the provided demographic dataset, we will create a machine learning technique in this part and share the insights we discover from the process. We will investigate several machine learning methodologies, assess their efficacy critically, and suggest optimisation approaches.

**Structure of Project Management**

Selecting a project management framework for a data science project is essential to efficiently planning and carrying out activities. Knowledge Discovery in Databases (KDD), CRISP-DM (Cross-Industry Standard Process for Data Mining), and SEMMA (Sample, Explore, Modify, Model, and Assess) are three frequently used frameworks. The project's scale and nature will determine which framework is best.

**Justification for CRISP-DM:** The CRISP-DM framework is suitable for this study of the demographic dataset. CRISP-DM offers data mining projects an organised methodology with clearly defined phases. The following is a summary of how it relates to the project:

**Business Understanding:** It's critical to comprehend the significance of the data and the company's objectives. Here, our goal is to learn more about the demographic changes as time goes on.

**Understanding the Data**: We have examined the dataset, looked for any missing information, and created a population trend visualisation. This is consistent with the stage of data comprehension.

**Data Preparation:** To get ready for modelling, data encoding and feature engineering were done.

**Modelling:** We have applied hyperparameter tweaking and machine learning techniques such as Random Forest and Linear Regression.

**Evaluation and Deployment:** Metrics like RMSE and R2 Score have been used to assess the models' performance.

**Selecting a Machine Learning Approach**

The purpose of the study and the type of dataset determine which machine learning approach is best. Here, our goal is to forecast population values using a dataset that contains both numerical and category information. Because we wish to generate predictions based on labelled data (past population estimates), we have opted to use supervised machine learning algorithms.

**Linear Regression**

When predicting numerical quantities (population in this example) based on input data, linear regression is a good option. It offers a straightforward yet understandable paradigm. Because it is the most straightforward to understand and can be used as a benchmark for more intricate models, we decided to choose Linear Regression as the baseline model.

**Random Forest**

Regression and classification challenges can be handled using the potent ensemble method known as Random Forest. Its ability to capture intricate correlations between characteristics and the target variable makes it a good fit for this dataset. After one-hot encoding, Random Forest likewise manages categorical data well.

**Hyperparameters Tuning**

We used GridSearchCV, a method that methodically investigates a variety of hyperparameters to discover the optimal combination, to do hyperparameter tweaking in order to optimise the performance of machine learning models. We tweaked the hyperparameters of the Random Forest and Linear Regression models.

Hyperparameter tuning for Linear Regression: We investigated if data normalisation enhanced the model's functionality. We found the optimal model parameters with the aid of GridSearchCV.

Hyperparameter tuning for the Random Forest model: We changed the number of estimators, the maximum depth, and the minimum split of samples. We were able to determine the ideal set of these parameters by using GridSearchCV.

**Performance Comparison of Models**

Two essential metrics—RMSE, or root mean squared error, and R2 score—were used to evaluate the performance of the Random Forest and Linear Regression models. These measures shed light on the quality of fit and accuracy of the models.

**Results of Comparing Model Performance:**

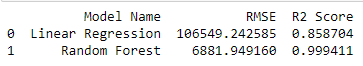


Fig 11: Comparison of Results

Linear Regression: R2 Score: 0.859; RMSE: 106549.24

Random Forest: R2 Score: 0.999; RMSE: 6881.95

In terms of both RMSE and R2 Score, the Random Forest model performs much better than the Linear Regression model. The population values may be predicted with a high degree of accuracy using the Random Forest model.

Visualisation of Model Performance Comparison

Bar charts for RMSE and R2 Score are used to display the outcomes of the model performance comparison. The performance disparities between the two models are easily seen thanks to these visuals.

**Comparison of RMSE:**

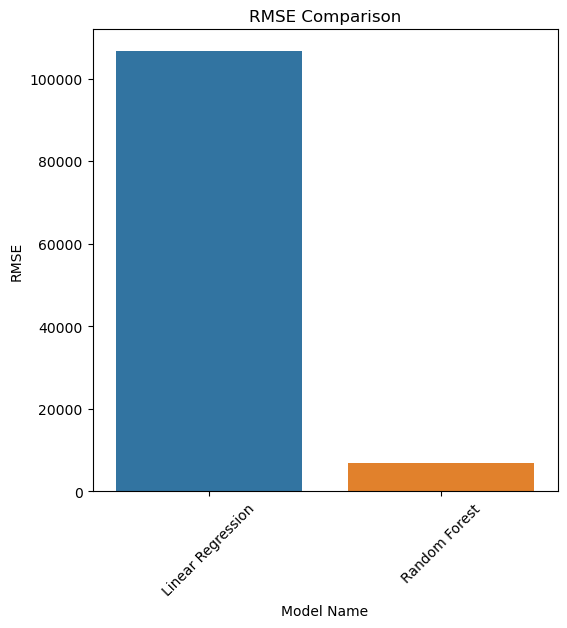


Fig 11: Barplot for RMSE

The bar plot demonstrates the greater accuracy of the Random Forest model over Linear Regression, with the Random Forest model having a substantially lower RMSE.

**Comparison of R2 Scores:**



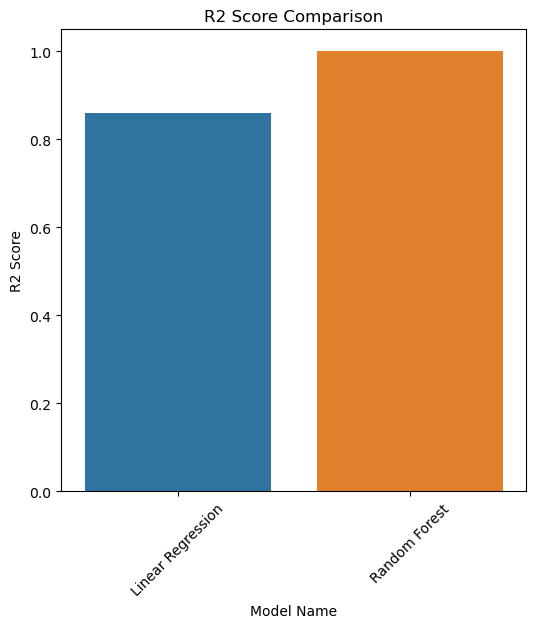


Fig 12: R2 score

While Linear Regression yields decent but less precise results, the Random Forest model offers a near-perfect fit, as seen by the R2 Score bar plot.

**Similarities and Differences in the Model's Outcomes**

**Similarities:**

Metrics like RMSE and R2 Score were used to assess both models.

GridSearchCV was used to refine both models in order to maximise their performance.

**Differences:**

In terms of accuracy (lower RMSE) and quality of fit (higher R2 Score), the Random Forest model performed better than the Linear Regression model.

The reason for Random Forest's improved performance was its capacity to grasp complicated correlations in the data.

**Explaination and Interpretation of the Data**

The study that made use of machine learning models—more especially, Random Forest and Linear Regression—offered insightful information on demographic changes. Predicting population numbers was remarkably accurate because to the Random Forest model's complex ability to identify patterns in the data. The Random Forest model outperformed Linear Regression in this situation, as seen by the models' comparison.

**Interpretation of Results:** These models' insights may direct decision-making in a number of industries, including urban planning, healthcare, and education. Policymakers, for example, might utilise these forecasts to better allocate healthcare resources, prepare for future infrastructure needs, and create educational initiatives that are demographically targeted.

To sum up, this data science project was made more organised by the CRISP-DM framework, which made it possible to conduct methodical investigation, model creation, and assessment. The best possible model performance was obtained by hyperparameter adjustment in conjunction with the suitable machine learning approach selection. Stakeholders are given practical insights by comparing the models and then interpreting the results, which highlights the effectiveness of machine learning influencing data-driven, strategic decision-making.

**Conclusion:** To sum up, our CRISP-DM framework-based data analysis research demonstrated the predictive capacity of machine learning. When it came to providing extremely precise demographic trend estimates, the Random Forest model outperformed Linear Regression. Policymakers in fields like healthcare and urban planning will be able to confidently make data-driven decisions thanks to these discoveries, which have important ramifications. In influencing the future of strategic resource allocation and policy development, the initiative highlights the revolutionary potential of advanced analytics.

**References:**

1. Sudheesh, S., 2015. *Genetics and genomics of key agronomic traits in field pea (Pisum sativum L.)* (Doctoral dissertation, La Trobe).
2. Field, E.H., 2015. All models are wrong, but some are useful. *Seismological Research Letters*, *86*(2A), pp.291-293.
3. Dobre, C. and Xhafa, F., 2014. Parallel programming paradigms and frameworks in big data era. *International Journal of Parallel Programming*, *42*(5), pp.710-738.
4. Charguéraud, A., 2011, September. Characteristic formulae for the verification of imperative programs. In *Proceedings of the 16th ACM SIGPLAN international conference on Functional programming* (pp. 418-430).
5. Sivaramakrishnan, K.C., Kaki, G. and Jagannathan, S., 2015. Declarative programming over eventually consistent data stores. *ACM SIGPLAN Notices*, *50*(6), pp.413-424.
6. Dobre, C. and Xhafa, F., 2014. Parallel programming paradigms and frameworks in big data era. *International Journal of Parallel Programming*, *42*(5), pp.710-738.
7. Fjeld, F., 2009. *Extending DataPool: A Tool for Handling Input Data in Scientific Computing, Using Python Web Frameworks* (Master's thesis).
8. Abdelhadi, A., Zainudin, S. and Sani, N.S., 2022. A Regression Model to Predict Key Performance Indicators in Higher Education Enrollments. *International Journal of Advanced Computer Science and Applications*, *13*(1).
9. Selvin, S., 2004. Statistical analysis of epidemiologic data (Vol. 35). Oxford University Press.
10. Wang, Y. and Chen, H.J., 2012. Use of percentiles and z-scores in anthropometry. In *Handbook of anthropometry: Physical measures of human form in health and disease* (pp. 29-48). New York, NY: Springer New York.
11. Sosulski, K., 2018. Data visualization made simple: insights into becoming visual. Routledge.