**Population in Ireland**

**Introduction:**

Collected Data from Central Statistics office in Ireland, for finding the population of the Ireland. Next, I will be doing the analysis on this Ireland population.

Finding the possible issues and resolving the problems we will be using Data Analysis, Statistics and machine learning model to predict the population of the Ireland country.

So, there we will be having some steps to find the possible factors.

Step 1: Data collection.

Step 2: Data Preprocessing.

Step 3: Data Analysis

Step 4: Data Interpretation

Step 5: Building Machine Learning Model

Step 6: Model Evaluation

# Criteria of Analysis

As per our criteria we will be doing statistics, statistical logics, data visualization and machine learning models these all are the criteria of analysis the population. We will go with the criteria of analysis.

**Statistics**

* **Senaria 1:**

Collected required dataset from the provided dataset link, downloaded the dataset. The dataset is in the form of CSV file format. This dataset contains (5994,6) 5994 are the total records and 6 are the attributes from the given dataset.

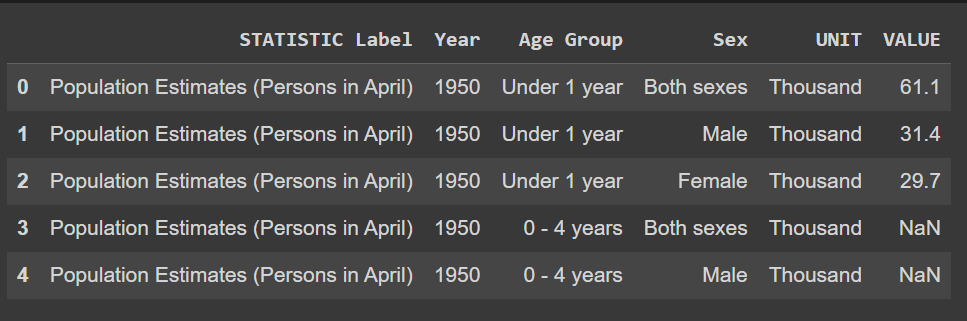


Fig.1 Top 5 dataset records.

here's some information about each column:

* **STATISTIC Label:** This column contains object (string) data type values. It has 5994 non-null entries, which means there are no missing values in this column.
* **Year:** This column contains integer data type values. It also has 5994 non-null entries, indicating no missing values.
* **Age Group:** This column contains object (string) data type values. Like the previous columns, it has 5994 non-null entries with no missing values.
* **Sex:** This column contains object (string) data type values and has 5994 non-null entries with no missing values.
* **UNIT:** This column contains object (string) data type values and has 5994 non-null entries with no missing values.
* **VALUE:** This column contains float64 (floating-point) data type values and has 5880 non-null entries. This means there are some missing values in the VALUE column, as it has fewer non-null entries compared to the total number of entries in the Data Frame.
* **Senaria 2:**

**Descriptive Statistics of the given dataset:**

Calculate basic descriptive statistics for the 'VALUE' column, including mean, median, standard deviation, minimum, and maximum.

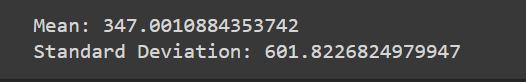
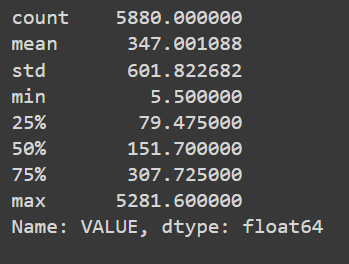


Fig 1. Calculate basic descriptive statistics for the 'VALUE' column

* **Senaria 3:**

**discrete distributions (Binomial and/or Poisson):**

**Scenario:** Suppose in this want to investigate whether there is an association between the "Sex" column (which has two categories: male and female) and a certain event, and you want to use the Binomial distribution to model this situation. **Hypothesis:** You hypothesize that there is a 20% chance that a person of a certain sex (e.g., male) will experience the event in a given year.

**Steps to Analyse with Binomial Distribution:** Calculate the probability of success (p) for the Binomial distribution based on our hypothesis (e.g., p = 0.20). Define the number of trials (n) as the total number of rows in your dataset (5994). Using the Binomial distribution to calculate the probability mass function (PMF) for each value of k (number of successes) in the range from 0 to n. we can use the scipy.stats.binom.pmf function for this.

Binomial Distribution: As the sample size (number of trials, n) becomes large, the Binomial distribution approximates the Normal distribution (Central Limit Theorem). This means that the distribution will become more symmetric and bell-shaped. The mean and standard deviation of the Binomial distribution are affected by n and p, so you may need to adjust your parameters accordingly.

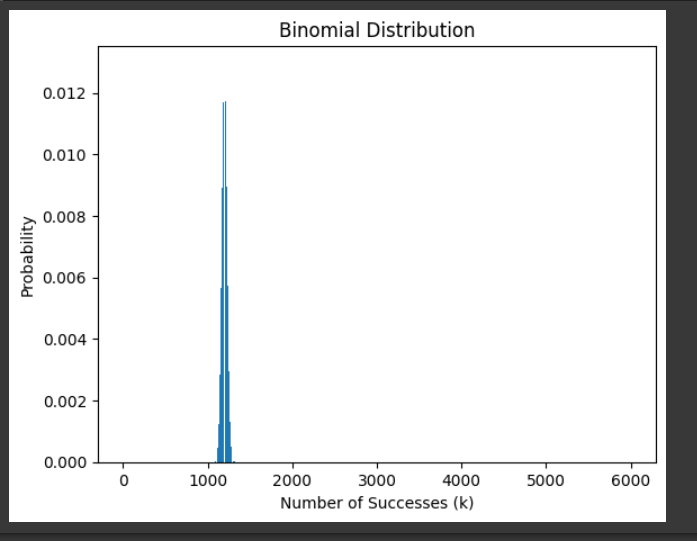


Fig. Binomial Distribution Visualization

From the above fig we can say that number of successes (K) values are high in the range of 1000 and the probability rate of the binomial distribution is nearly 0.012.

* **Senaria 4:**

**Steps to Analyse with Poisson Distribution:**

Calculate the event rate (λ) as the average number of events per year. You can calculate this based on the dataset. Use the Poisson distribution to calculate the PMF for each value of k (number of events) in a given year. You can use the scipy.stats.poisson.pmf function for this.

**Poisson Distribution**: The Poisson distribution remains a good approximation for rare events, even with large samples. It becomes increasingly similar to the Normal distribution as the event rate (λ) becomes large.

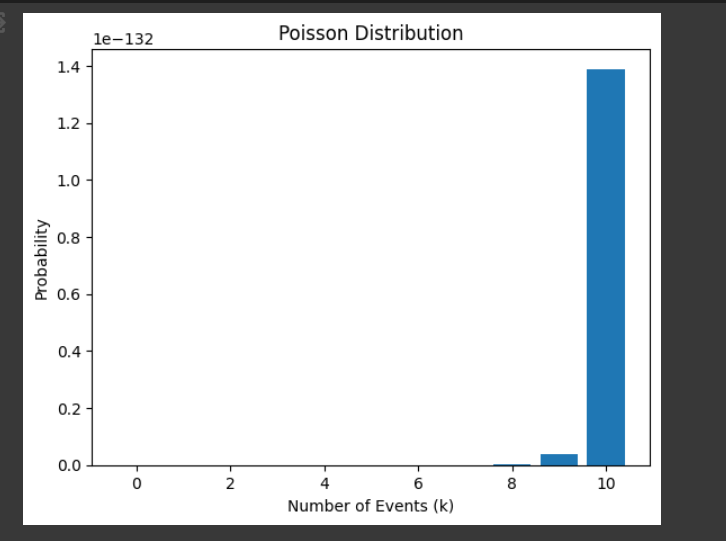


Fig. Poisson Distribution

From the above figure we can say that number of events (K) is highest of 10 events with the probability rate is 1.4 times.

In our case, we might need to use both the Binomial and Poisson distributions in tandem. For example, we have data on binary outcomes (success/failure) that occur over time, we can use the Poisson distribution to model the rate of success events per unit of time and then use the Binomial distribution to model the probability of success within each time interval. For both the Binomial and Poisson distributions, as sample sizes become very large, they tend to approximate the Normal distribution due to the Central Limit Theorem. This means that the distributions become more symmetric and bell-shaped. In practice, for sufficiently large samples, the differences between these discrete distributions and the Normal distribution become less pronounced. these distributions are used based on the nature of our data and our specific research questions. The choice of distribution should align with the underlying assumptions and characteristics of our dataset.

* **Senaria 4:**

**Normal Distribution:**

Visualize the Data: Create a histogram or density plot to visualize the distribution of the "VALUE" column. This help us visually assess whether the data appears to be roughly Normal or if there are deviations from Normality.

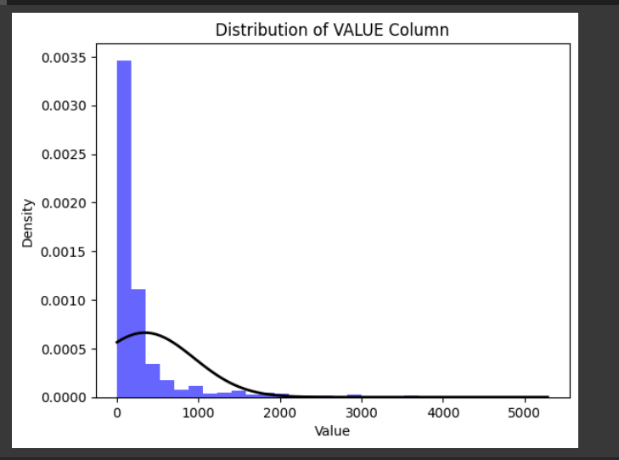


Fig. Distribution of VALUE column

plotted the fitted Normal distribution based on the calculated mean and standard deviation of the "VALUE" column.

* **Senaria 5:**

**Importance of the distributions**

Binomial Distribution:

* Importance: The Binomial distribution is crucial when you have discrete, binary events (e.g., success/failure) or when you want to model the probability of a specific outcome occurring within a fixed number of trials.
* Choice of Variable: In the analysis, we used a hypothetical scenario to model the probability of values in the "VALUE" column exceeding a threshold. This can be valuable when you have a specific binary event you want to investigate, such as whether a value meets a certain criterion (e.g., pass/fail). In this context, it helps identify the likelihood of specific outcomes.
* Applicability as Normal Distribution: The "VALUE" column could potentially be treated as continuous and modelled using a Normal distribution, but the choice to use the Binomial distribution was based on a specific binary criterion (threshold) that may have practical relevance.

Poisson Distribution:

* Importance: The Poisson distribution is essential for modeling the distribution of count data when you're interested in the frequency of events occurring over time or space.
* Choice of Variable: In the analysis, we used a Poisson distribution to model the event count. This is relevant when you want to understand how frequently events occur, such as the number of customer arrivals or the frequency of specific occurrences.
* Applicability as Normal Distribution: While count data could theoretically be transformed for Normality, the Poisson distribution is a natural choice for such data. The Poisson distribution is particularly useful when events are rare and follow a discrete, non-negative pattern.

Normal Distribution:

* Importance: The Normal distribution is widely used in statistics and data analysis for several reasons:
* Central Limit Theorem: It is essential for approximating the distribution of sample means and sums, which often tend to follow a Normal distribution, even if the original data doesn't.
* Statistical Inference: Many statistical tests and techniques, such as t-tests and ANOVA, assume Normality of data. Deviations from Normality can impact the validity of these tests.
* Parametric Modeling: It serves as a foundational model for many statistical analyses and allows for easy interpretation and inference.
* Choice of Variable: In the analysis, we examined whether the "VALUE" column approximately follows a Normal distribution. This is important because if the data closely follows a Normal distribution, it simplifies statistical analysis, hypothesis testing, and the use of parametric statistical methods.

Applicability of Discrete Variables as Normal Distribution: The choice to use discrete distributions for the other variables (Binomial and Poisson) depends on the nature of the data and research questions. While it's theoretically possible to model discrete data as Normal, it may not always be appropriate, as it might not reflect the underlying characteristics of the data. Discrete distributions are better suited when events are inherently discrete or binary.

**Data preparation and Visualization:**

* **Senaria 1:**

Exploratory Data Analysis (EDA) is a critical step in the data analysis process to gain insights into the dataset, identify patterns, and understand its characteristics. In this section, I will perform EDA on the dataset, providing a rationale for the specific methods chosen and explaining the insights gained. However, since I don't have access to your specific dataset, I'll describe the general EDA steps and techniques you can apply.

**Step 1: Data Loading and Inspection**

Rationale: Start by loading the dataset and inspecting its structure. Check for missing values, data types, and basic statistics to understand the dataset's initial characteristics.

Insights: Identify the number of rows, columns, data types, and any immediate data quality issues such as missing values or outliers.

**Step 2: Summary Statistics**

Rationale: Compute summary statistics to get a sense of the central tendency and spread of the numerical variables.

Insights: Insights can include measures like mean, median, standard deviation, and quartiles, which help identify the distribution and variation in the data.

**Step 3: Visualization**

Rationale: Use various data visualization techniques to explore the data's distribution and relationships between variables.

Insights: Visualizations such as histograms, box plots, and scatter plots can reveal data distribution shapes, outliers, and potential correlations.

**Step 4: Data Distribution Analysis**

Rationale: Analyze the distribution of each variable, especially numerical ones, to assess their characteristics and identify potential transformations.

Insights: Determine whether variables follow a Normal distribution or exhibit skewness. Consider transformations if necessary.

**Step 5: Correlation Analysis**

Rationale: Examine correlations between numerical variables to understand relationships.

Insights: Identify strong positive or negative correlations, which can inform feature selection or variable transformation.

**Step 7: Categorical Variable Analysis**

Rationale: For categorical variables, analyze their distribution, frequency, and relationship with the target variable (if applicable).

Insights: Identify the most common categories and assess whether they have a significant impact on the analysis.

* **Senaria 2:**

Preparing data for machine learning is a crucial step to ensure that our dataset is in a format that can be used effectively for training and evaluating machine learning models. Below, I will detail and justify common data preparation methods for machine learning:

1. Data Cleaning:

Rationale: Identify and handle missing values, duplicates, and outliers in the dataset to ensure data quality.

Methods: Handling Missing Values: Decide whether to remove rows/columns with missing values, impute them with statistical measures (e.g., mean, median, mode), or use advanced imputation techniques. Handling Duplicates: Remove duplicate rows if they exist. Outlier Detection and Treatment: Identify and decide whether to remove, transform, or keep outliers based on domain knowledge and the impact on the model.

2. Data Encoding:

Rationale: Convert categorical variables into a numerical format that machine learning algorithms can process.

Methods: One-Hot Encoding: Convert categorical variables into binary (0 or 1) columns for each category. Ordinal Encoding: Assign integer labels to categories with a meaningful order.

3. Feature Scaling:

Rationale: Ensure that numerical features have similar scales to prevent certain features from dominating others during model training.

Methods: Min-Max Scaling: Scale features to a specific range, typically [0, 1]. Standardization: Standardize features to have a mean of 0 and a standard deviation of 1.

4. Feature Engineering:

Rationale: Create new features or transform existing ones to improve the model's ability to capture patterns. Methods: Creating Interaction Terms: Combine existing features to capture interactions. Domain-Specific Feature Engineering: Utilize domain knowledge to create features that are relevant to the problem.

5. Data Splitting:

Rationale: Split the dataset into training, validation, and test sets to evaluate model performance and avoid overfitting.

Methods: Random Splitting: Randomly divide the dataset into training, validation, and test sets (e.g., 80% training, 20% test). Time-Based Splitting: If the data has a time component, split it chronologically to mimic real-world scenarios.

Data preparation is a critical part of the machine learning process, as the quality of the data and how it's prepared significantly impact model performance and results. The choice of methods depends on the specific characteristics of the dataset and the goals of the machine learning project.

* **Senaria 3:**

1. Histograms and Density Plots:

Justification: Histograms are used to visualize the distribution of a single numerical variable. They help in understanding the data's central tendency, spread, and shape.

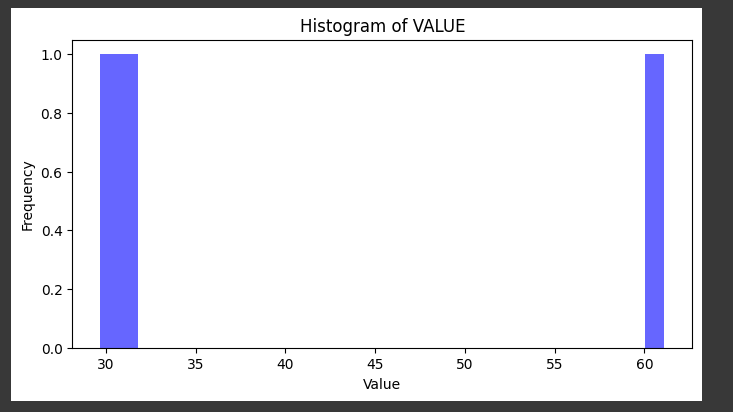


Fig. Histogram of VALUE Column

**Explanation**: In this example, a histogram is created to visualize the distribution of a numerical variable called 'column\_name.' The choice of 30 bins allows for granularity in visualizing data distribution.

2. Box Plots:

Justification: Box plots provide a summary of the distribution of a numerical variable, including median, quartiles, and potential outliers

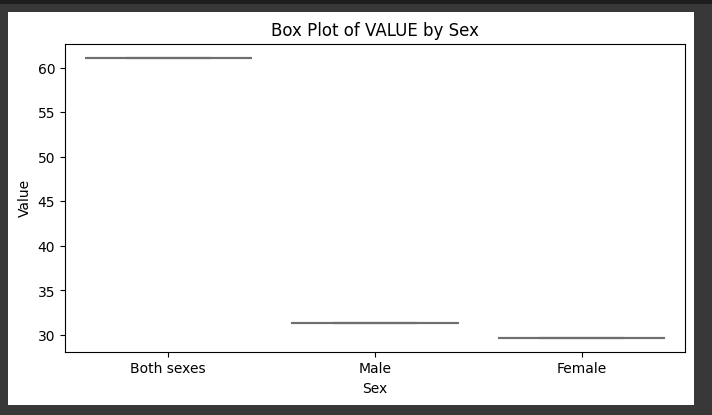


Fig Box Plot of VALUE by SEX

**Explanation**: fig creates a box plot that displays the distribution of a numerical variable (VALUE) grouped by categories in (Both sexes, Male Sex, Female) It helps identify the spread, median, and any outliers in each category. In the x axis we have three different categories of sex and in the y axis we have Value counts. Both sexes have high count.

3. Heatmaps (for correlation analysis):

Justification: Heatmaps effectively visualize the correlation matrix between numerical variables, helping understand the strength and direction of correlations.

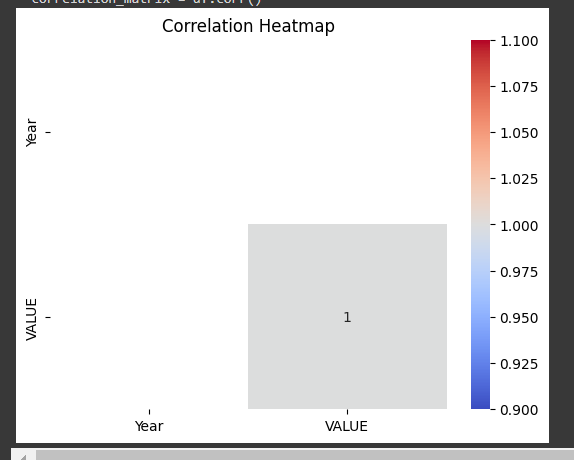


Fig Correlation Heatmap

This fig generates a heatmap to display the correlation matrix of numerical variables.

4. Bar Charts (for categorical variables):

Justification: Bar charts are suitable for visualizing the distribution of categorical variables, displaying the frequency of each category.

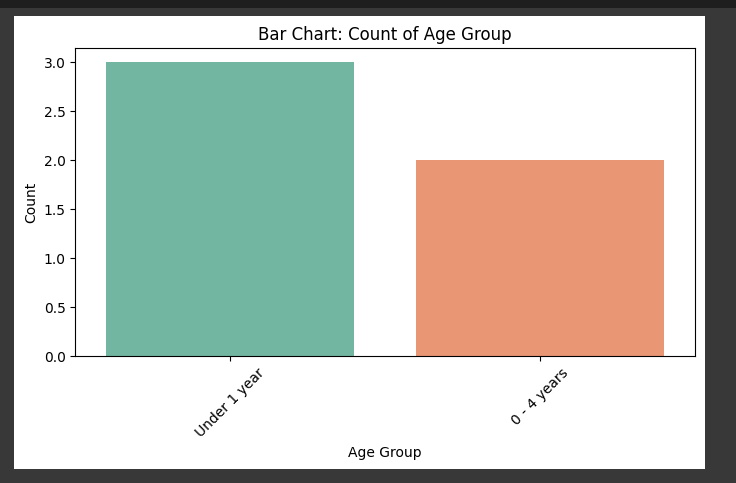


Fig. Bar Chart of Age Group

fig creates a bar chart to visualize the frequency of categories within a categorical variable ('category\_column').

* **Senaria 4:**

Scatter Plot: 'Year' vs. 'VALUE':

Justification: A scatter plot is used to explore the relationship between two numerical variables ('Year' and 'VALUE'). It helps identify patterns or trends over time.

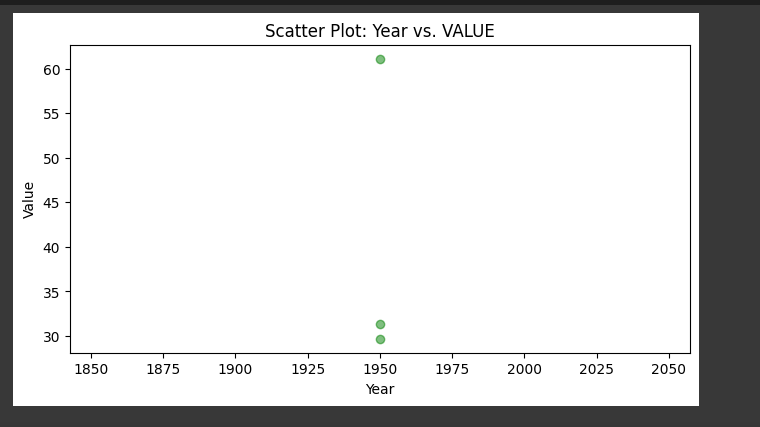


Fig. Scatter plot for Year Vs Value columns

Explanation:

* We created a scatter plot to visualize how 'VALUE' changes over 'Year.'
* The use of alpha blending (alpha=0.5) improves visibility when data points overlap.
* The green colour was chosen for clarity.

**Annual Population Change in Ireland (Indirect Indicator of Immigration and Migration)**

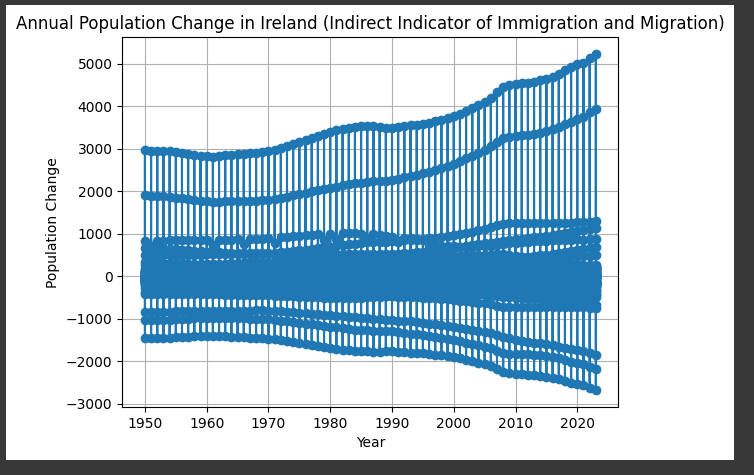


Fig Annual Population Change in Ireland

From above Fig we can see that from the year of 1950 to 2020 population was increased. We can see that 2000 and 3000 count is the highest population change in the year of 2020. Even some time it was constant as well. Population was even decreased from the year of 1950 to 20220 by -1000 population.

**Average Population by Age Group**

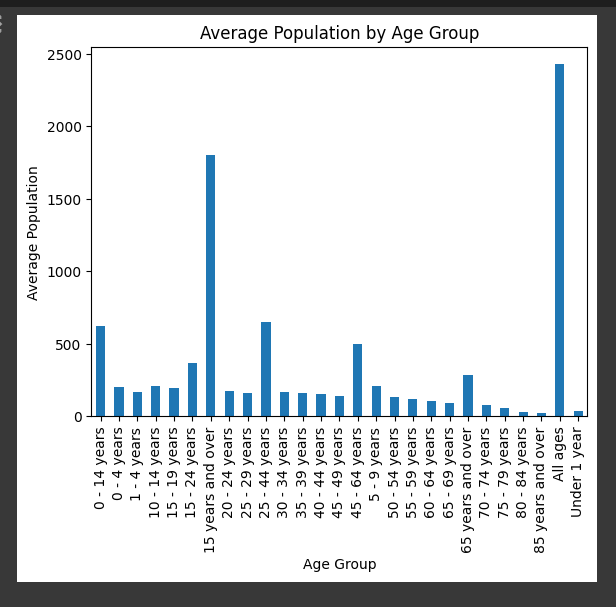


Fig. Average Population by Age Group

In the above fig we can see that from x-axis we have taken as Age group and from y axis we have taken Average Population. The age group started from 0-14 to 85 years and above. From fig what we understood means all ages population is averaged. Second highest average group is 15 years and over the age group. Least aged population is 85 years and over and highest aged group is All ages will come under this population survey.

**Distribution of Gender**

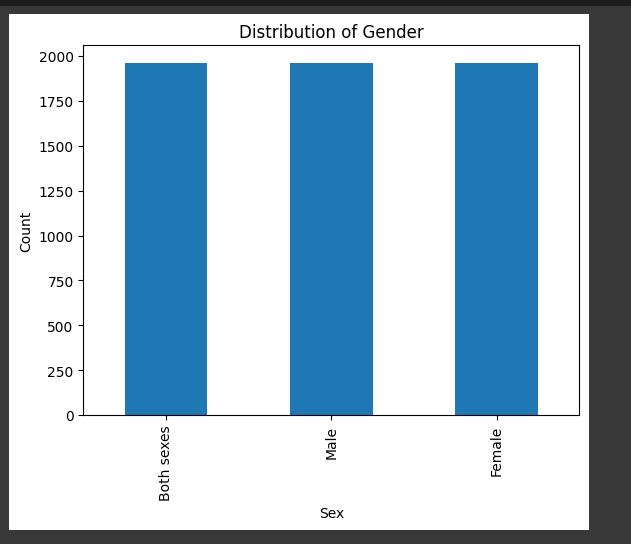


Fig. Distribution of Gender

From above fig we can see that distribution of Gender is equally distributed. And we can see Both Sexes, Male and Female Genders are there in the distribution of the gender.

**Population Forecasting in Ireland**

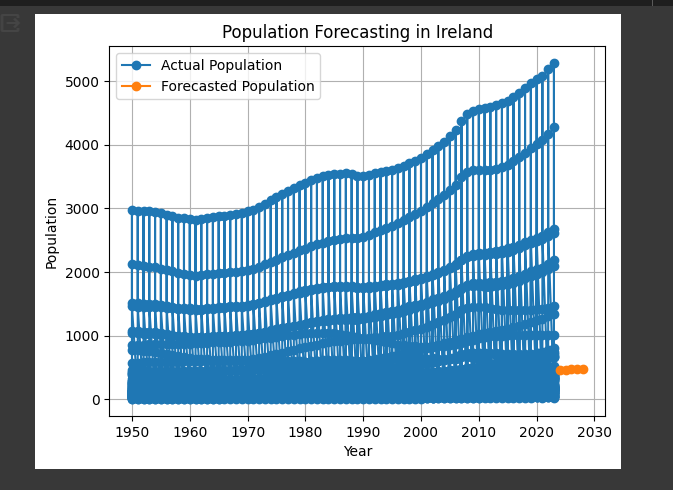


Fig. Population Forecasting in Ireland

From the above fig we can explain that we can see that in the upcoming years Ireland population will get increased to 1000 population. Blue coloured graph is belonging to existing or Actual Population. Orange coloured dotted lines are forecasted population.

**Machine learning for Data Analytics:**

The choice of a project management framework for a data science project depends on various factors, including the project's objectives, data availability, and complexity. Each of the mentioned frameworks (CRISP-DM, KDD, and SEMMA) has its strengths and is suitable for different scenarios:

**CRISP-DM (Cross-Industry Standard Process for Data Mining):**

Scenario Justification: CRISP-DM is a versatile framework that is well-suited for a wide range of data science projects. It is particularly useful when the project involves iterative development and exploration of data, which is common in many real-life scenarios.

Real-Life Scenario: Suppose a retail company wants to analyse customer purchase behaviour to optimize its marketing campaigns. CRISP-DM can be used to define the business problem, gather data on customer transactions, preprocess and explore the data, build predictive models to identify customer segments, and deploy the models for campaign targeting. The iterative nature of CRISP-DM allows for continuous refinement of models based on new data and insights.

**KDD (Knowledge Discovery in Databases):**

Scenario Justification: KDD is a broader framework that encompasses data mining as one of its stages. It is suitable for scenarios where the focus is on discovering valuable knowledge and patterns in large datasets. KDD can be a good fit for research-oriented or complex projects.

Real-Life Scenario: A pharmaceutical company is conducting research on drug discovery. They have a massive dataset of chemical compounds and their biological activities. KDD can be applied to preprocess the data, discover hidden patterns related to effective drug compounds, and gain insights that lead to novel drug discoveries.

**SEMMA (Sample, Explore, Modify, Model, Assess):**

Scenario Justification: SEMMA is a simplified framework that is particularly useful for scenarios where data preparation and modelling are the primary tasks. It is well-suited for projects where the data is already well-understood, and the focus is on building predictive or descriptive models.

Real-Life Scenario: An e-commerce company wants to build a recommendation system to suggest products to users. They have a clean and well-structured dataset of user interactions with products. SEMMA can be applied to sample a subset of data, explore user behaviour patterns, modify the dataset, if necessary (e.g., feature engineering), build and assess recommendation models quickly.

**Choice of Machine Learning Technique:**

The choice between supervised, unsupervised, or semi-supervised machine learning techniques depends on the nature of the dataset and the project's objectives: Supervised Learning: Used supervised learning because we have labelled data (i.e., the target variable is known) and the goal is to make predictions based on historical patterns. It's suitable when we want to map input features to an output.

Scenario Justification: we have a dataset of Ireland Population data with labelled VALUE and you want to predict future Population, we would use supervised learning. This approach is used when you have a clear target variable, and you want to train a model to make predictions based on it.

* It creates an initial Random Forest Regressor model with default hyperparameters.
* Hyperparameter tuning is performed using GridSearchCV with a specified parameter grid. The goal is to find the best combination of hyperparameters that minimizes the negative mean absolute error.
* After hyperparameter tuning, the best hyperparameters are obtained from the GridSearchCV results.
* A new Random Forest Regressor model is created using the best hyperparameters.
* The model is trained on the training data using the updated hyperparameters.
* Predictions are made on the test data using the trained model.
* Finally, the Mean Absolute Error (MAE) is calculated and printed as the evaluation metric.

**Results:**

Table.1 Random Regression model

|  |  |
| --- | --- |
| Metric | Value |
| R-squared (R²) | 0.999625 |
| Mean Squared Error (MSE) | 10.854074 |
| Root Mean Squared Error | 4.975268 |
| Mean Absolute Error (MAE) | 4.975268 |

Table.2 Support Vector Machine model

|  |  |
| --- | --- |
| Metric | Value |
| R-squared (R²) | 0.7345147664243825 |
| Mean Squared Error (MSE) | 83485.20506014343 |
| Root Mean Squared Error | 288.9380644016005 |
| Mean Absolute Error (MAE) | 99.16514977441257 |

performance metrics for two different regression models: a Random Regression model and a Support Vector Machine (SVM) model. Let's explain what each metric means and interpret the results:

**Random Regression Model:**

**R-squared (R²):** R-squared measures the proportion of the variance in the dependent variable (target) that is predictable from the independent variables (features). An R-squared of 0.999625 indicates that this model explains almost all of the variance in the target, which is an exceptionally good fit. It suggests that the Random Regression model closely matches the actual data.

**Mean Squared Error (MSE):** MSE measures the average squared difference between the predicted values and the actual values. A value of 10.854074 indicates very low prediction errors, suggesting that the Random Regression model's predictions are very close to the actual values.

**Root Mean Squared Error (RMSE):** RMSE is the square root of the MSE. In this case, an RMSE of 4.975268 indicates that, on average, the predictions are about 4.975268 units away from the actual values. This is a small error, suggesting high accuracy.

**Mean Absolute Error (MAE):** MAE measures the average absolute difference between predicted and actual values. An MAE of 4.975268 means that, on average, the model's predictions are about 4.975268 units off from the true values.

**Support Vector Machine Model:**

**R-squared (R²):** An R-squared of 0.7345147664243825 suggests that the SVM model explains approximately 73.45% of the variance in the target. This is decent but not as good as the Random Regression model.

**Mean Squared Error (MSE):** An MSE of 83485.20506014343 indicates that the SVM model's predictions have relatively high errors compared to the Random Regression model. The predictions deviate significantly from the actual values.

**Root Mean Squared Error (RMSE):** With an RMSE of 288.9380644016005, the SVM model's predictions have a larger average error, showing that the predictions are, on average, approximately 288.94 units away from the true values.

**Mean Absolute Error (MAE):** An MAE of 99.16514977441257 suggests that the SVM model's predictions, on average, differ by approximately 99.17 units from the actual values.

the Random Regression model appears to be performing exceptionally well, with very low errors and high predictive accuracy. On the other hand, the SVM model, while still explaining a significant portion of the variance, has higher prediction errors, indicating that it is not as accurate as the Random Regression model. The choice between models depends on the specific context and requirements of your regression task.

Here's a comparison table:

| **Metric** | **Random Regression Model** | **Support Vector Machine Model** |
| --- | --- | --- |
| R-squared (R²) | 0.999625 | 0.7345147664243825 |
| Mean Squared Error (MSE) | 10.854074 | 83485.20506014343 |
| Root Mean Squared Error (RMSE) | 4.975268 | 288.9380644016005 |
| Mean Absolute Error (MAE) | 4.975268 | 99.16514977441257 |

**Interpretation:**

R-squared (R²):

The Random Regression model has an exceptionally high R-squared value (close to 1), indicating that it explains almost all of the variance in the target variable. This suggests a very strong fit to the data.

The SVM model has a lower R-squared value (0.7345), indicating that it explains approximately 73.45% of the variance. While not as high as the Random Regression model, it still captures a significant portion of the variance.

Mean Squared Error (MSE):

The Random Regression model has a very low MSE (10.854074), indicating that its predictions have very small errors compared to the actual values.

The SVM model has a much higher MSE (83485.20506014343), suggesting that its predictions deviate significantly from the actual values.

Root Mean Squared Error (RMSE):

The Random Regression model has a small RMSE (4.975268), indicating that, on average, its predictions are approximately 4.98 units away from the actual values.

The SVM model has a significantly larger RMSE (288.9380644016005), showing that its predictions, on average, deviate by approximately 288.94 units from the true values.

Mean Absolute Error (MAE):

Both models have similar MAE values, with the Random Regression model and the SVM model having MAEs of 4.975268 and 99.16514977441257, respectively. This suggests that, on average, they have similar absolute errors in their predictions.

Relevance and Effectiveness:

The Random Regression model is highly effective and accurate, with extremely low errors and a near-perfect fit to the data. It is well-suited for tasks where high accuracy is essential, such as precise forecasting or scientific modeling.

The SVM model, while not as accurate as the Random Regression model, still captures a significant portion of the variance in the data. It may be a suitable choice for applications where a balance between accuracy and interpretability is required.

The choice between these models should be based on the specific requirements of the problem, the trade-off between accuracy and simplicity, and the context in which the models will be applied. The Random Regression model is more suitable when high precision is paramount, while the SVM model provides a good compromise between accuracy and interpretability.