**Contents**

[Introduction 2](#_Toc150608328)

[A- Statistics for Data Analytics 2](#_Toc150608329)

[A.1 Data Cleaning 2](#_Toc150608330)

[A.2 Data Aggregation Check 2](#_Toc150608331)

[A.3 Exploration and Evaluation of Dataset 3](#_Toc150608332)

[A.3.1 Descriptive Statistics 3](#_Toc150608333)

[A.3.2 Data Visualization: 3](#_Toc150608334)

[A.4. Binomial Distribution 7](#_Toc150608335)

[A.5. Normal Distribution 9](#_Toc150608336)

[B. Data Preparation and Visualization 10](#_Toc150608337)

[B.1 Exploratory Data Analysis (EDA) 10](#_Toc150608338)

[B.2. Descriptive Statistics: 12](#_Toc150608339)

[B.3. Distribution Plots: 12](#_Toc150608340)

[C. Preparing Data for ML 15](#_Toc150608341)

[C.1. Identify Categorical Variables: 15](#_Toc150608342)

[C.3. Machine Learning for Data Analytics 16](#_Toc150608343)

[C.3.1 Project Management Framework 16](#_Toc150608344)

[C.3.2 CRISP-DM: 16](#_Toc150608345)

[C.3.3. KDD: 16](#_Toc150608346)

[C.3.4. SEMMA: 17](#_Toc150608347)

[C.4. Machine Learning Technique: 17](#_Toc150608348)

[C.5. Model Training and Evaluation 17](#_Toc150608349)

[C.6. Model Selection and Training: 18](#_Toc150608350)

[C.7. Model Evaluation: 18](#_Toc150608351)

[C.8.1 Comparison Table: 19](#_Toc150608352)

[C.8.2 Linear Regression: 19](#_Toc150608353)

[C.8.3 Random Forest (Initial & Optimized): 19](#_Toc150608354)

[C.8.4 Similarities and Differences: 20](#_Toc150608355)

[C.8.5 Graphical Comparison: 20](#_Toc150608356)

[C.8.6 Interpretation 20](#_Toc150608357)

**Predictive Analysis for Population**

# **Introduction**

The primary objective of this study is to conduct an analysis of a dataset including estimated population figures. The dataset encompasses demographic information, such as population figures disaggregated by geographic regions, age cohorts, and gender categories. The primary objective of this analysis is to enhance comprehension of population trends and utilize this comprehension to forecast future population magnitudes.

The dataset, named "Estimated Population (Persons in April)", is organized and ready for detailed study. In this report, we start by exploring the data to see what it looks like and what it tells us about the population.

# **Statistics for Data Analytics**

Before starting Statistics for Data Analytics, the steps performed are:

* Load the dataset to understand its structure.
* Identify and remove any unnecessary columns.
* Check each column for aggregated data that may need to be excluded.
* Locate and remove the column named 'unit' as it appears constant and therefore not useful for analysis.

## **A.1 Data Cleaning**

The *STATISTIC, TLIST(A1),* and *UNIT* columns have been successfully removed. Also, columns like *C02076V02508, C02199V02655,* and *C02196V04140* are also removed.

* Removed the *STATISTIC* column as it contains a constant value that doesn't vary across rows and doesn't contribute to any analysis.
* The *TLIST(A1)* column seems redundant due to the presence of the Year column, and thus, it is removed.
* The *UNIT* column was also removed since it contains the same value for all entries and does not contribute to differentiation in data.
* Columns with unclear codes (C02076V02508, C02199V02655, C02196V04140) also removed for not giving any information.

## **A.2 Data Aggregation Check**

The check for aggregated data within the Year, Age Group, and Sex columns indicates multiple counts for each unique combination of these three categories. Since the count exceeds one for all combinations, this suggests that the dataset likely contains multiple entries for each demographic segment within different regions.

* Removed rows where the Age Group is "All ages" to exclude age-related aggregations.
* Removed rows where the Sex is "Both sexes" to exclude gender-related aggregations.
* Removed rows where the Region is "State" to exclude region-related aggregations.

The refined dataset now exclusively contains data points that are differentiated by specific age groups, distinct sex categories, and individual regions.

## **A.3 Exploration and Evaluation of Dataset**

### **A.3.1 Descriptive Statistics**

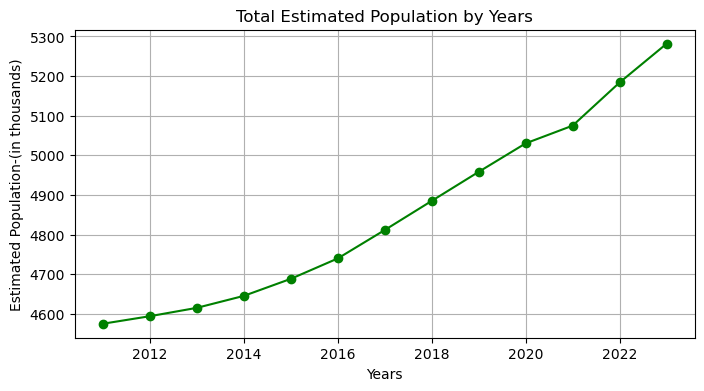
The dataset, now refined to exclude aggregates, describes the estimated population (in thousands) for various age groups, sexes, and regions across different years. The descriptive statistics for the population values (VALUE) are as follows:

* **Count**: There are 3,744 non-aggregated entries in the dataset.
* **Mean**: The average population size in the dataset is approximately 16.85 thousand.
* **Standard** **Deviation**: The population sizes have a standard deviation of around 11.43 thousand, indicating variability in the population sizes across different categories.
* **Minimum**: The smallest population size recorded is 1.2 thousand.
* **25th** **Percentile** (Q1): 25% of the population size is below 10.2 thousand.
* **Median** (**Q2**): The median population size is 14.3 thousand.
* **75th** **Percentile** (**Q3**): 75% of the population sizes are below 20.0 thousand.
* **Maximum**: The largest population size recorded is 66.0 thousand.

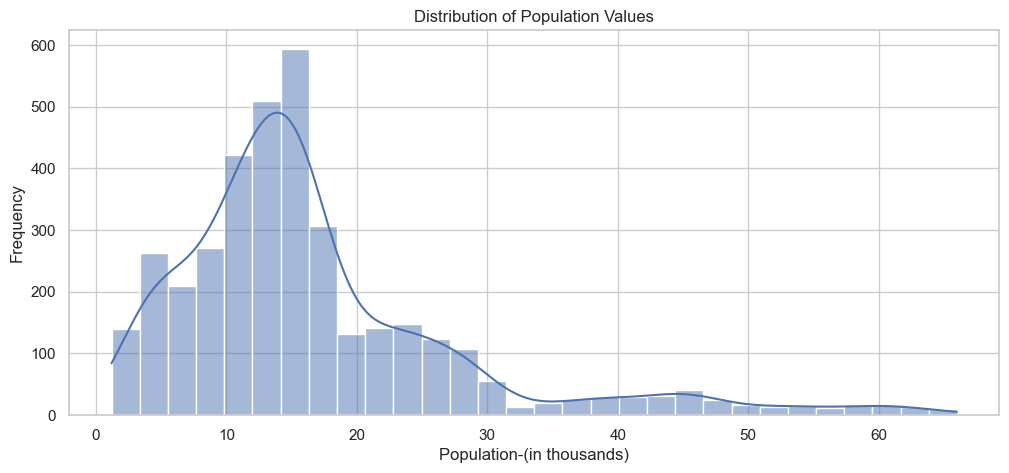
### **A.3.2 Data Visualization:**

**Temporal Analysis of Population Dynamics:** The presented visual representation depicts the longitudinal trajectory of the aggregate population over a span of many years. This phenomenon serves as an indicator of fluctuations in the size of a population and can be utilized to deduce patterns of population expansion or contraction.

The population has been steadily increasing over the years. Trend shifts a bit during 2020 and 2021 but still population keeps rising up.



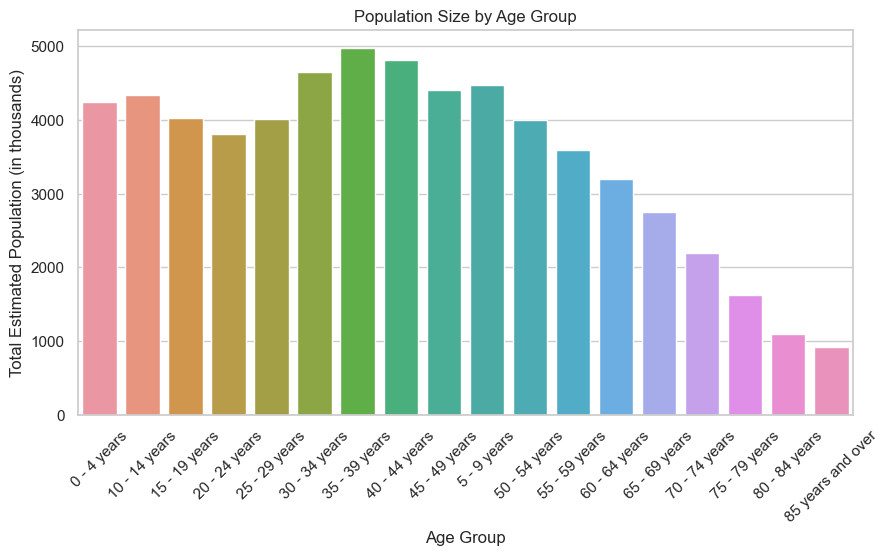
**Histogram of Population Values**: The histogram presents the distribution of population sizes, showing how often each population size occurs.



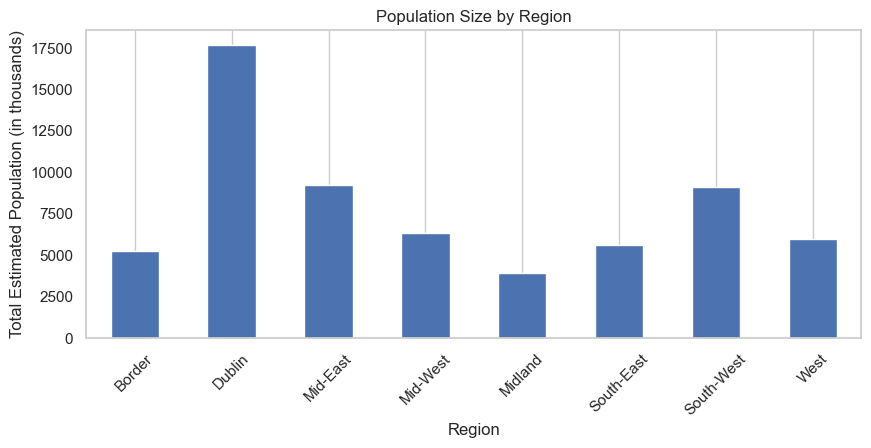
* The histogram exhibits a positively skewed distribution, indicating the prevalence of multiple locations with relatively small populations and a smaller number of regions with significantly larger populations.
* The majority of the population is comprised of approximately 10,000 to 20,000 individuals.
* The absence of bars on the far-right side of the histogram suggests a scarcity or absence of occurrences with exceptionally high population values, which could potentially be outliers, within the dataset.

**Population by Age Group**: The provided figure visually represents the aggregate population across various age groups. Comparing population sizes can provide valuable insights into the distribution of age groups, allowing for the identification of age cohorts with the largest or lowest population sizes.

* The age cohort of individuals aged 30-34 years exhibits the largest population count, suggesting a notable concentration of individuals within this specific age range.
* The age groupings of "25-29 years" and "35-39 years" exhibit substantial population counts, indicating a significant presence of individuals ranging from their late twenties to late thirties.
* The analysis reveals that the oldest age group(s) exhibit the lowest population, suggesting a diminished proportion of elderly adults.
* The population sizes of younger age groups, specifically those aged "0-4 years" and "5-9 years", are of intermediate magnitude, falling neither within the smallest nor the largest categories.
* A decline in population size becomes evident as individuals surpass the age range of 30 to 34 years, indicating a gradual reduction in population figures as age advances.



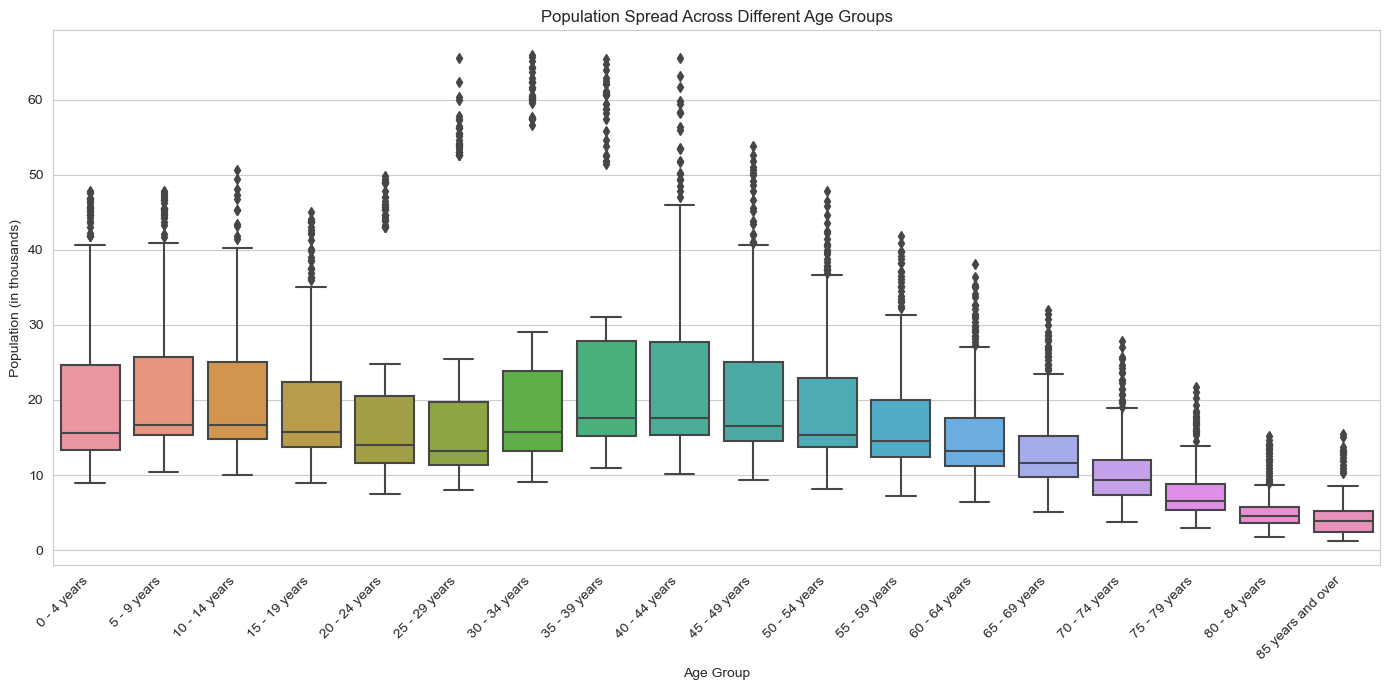
**Population by Region**: This visualizes the total population by region. It's useful for identifying which regions have larger or smaller populations.



* One region, "South-West" significantly outnumbers the others in terms of population.
* The regions "Mid-West" and "South-East" also have substantial populations after “South-West”.
* The "Border" region and another region "West" have the lowest population counts among all the regions.
* This variation in regional populations can indicate differing levels of urbanization, economic activity, or geographic size across these regions.

**Population Values by Age Group**: The boxplot provides a visual summary of the distribution of population sizes across different age groups.

* The age groups of "0-4 years" and "5-9 years" exhibit a relatively limited span of population values and maintain a stable population size.
* The age group of individuals between 30 and 34 years old displays a broader range and a higher median, suggesting that this particular age group tends to have a larger and more diverse population size.
* There exist significant deviations from the norm among age cohorts, namely among individuals aged between 20 and 24 years, as well as those between 25 and 29 years. The outliers in question denote instances where the population for the specified age groups had an unusually elevated level.
* The age groups at the higher end of the spectrum, such as "65-69 years" and "70-74 years," exhibit a comparatively lower median population, indicating a lesser representation of elderly individuals within the sample.



## **A.4. Binomial Distribution**

The Binomial Distribution is:

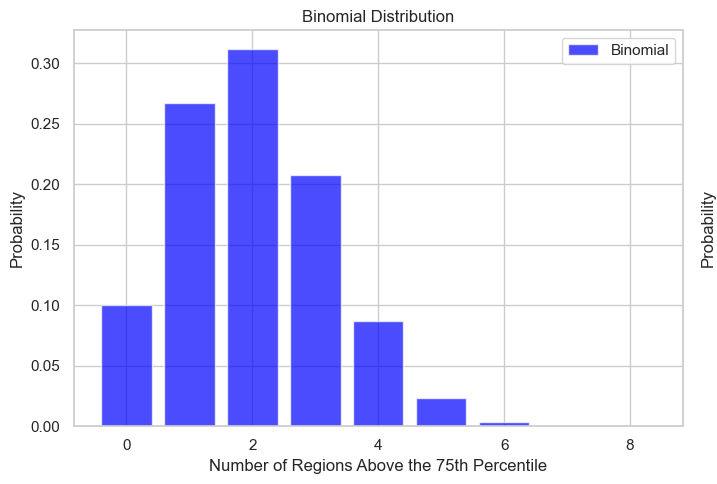
**n (Trials):** The number of regions in each year.

**p (Probability of success):** The probability that a region’s population for the 20-24 age group is above the 75th percentile for that age group.

**X (Number of successes):** The number of regions with a population above the 75th percentile in a given year.

Probability of success for a single trial. The binomial distribution to model the probability of k successes across all trials for a single year.

The binomial distribution was utilized to model the probability of a certain number of regions having a population for the age group of 20-24 years. The 75th percentile threshold was calculated to be approximately 20.525 thousand individuals. With each region considered a trial, the probability of a 'success' was determined to be 0.25. This means that in any given year, there is a 25% chance that a randomly selected region will have a population size in this age group above the 75th percentile. The graph illustrates these probabilities for 0 through 8 successes.



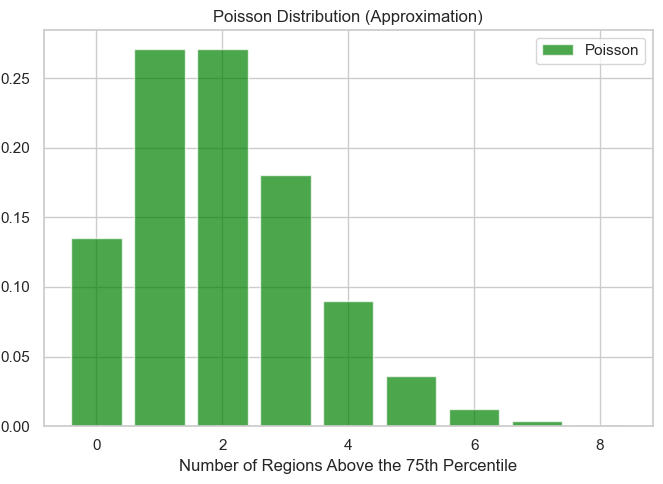
**Poisson Distribution**

For the Poisson distribution, we usually deal with the number of events in a fixed interval of time or space. We'll use each region year as an independent observation:

**Lambda (λ)**: The average number of regions in a year that exceed the 75th percentile population size for the 20-24 age group.

Calculate the 75th percentile for the population size of the 20-24 age group and the probability of a region being above this percentile.

The Poisson distribution is used as an approximation to the binomial distribution, assuming a substantial number of trials and a low probability of success. The average rate of 2.0 indicates the anticipated annual count of regions in which the population size for the age group of 20-24 years is above the 75th percentile.



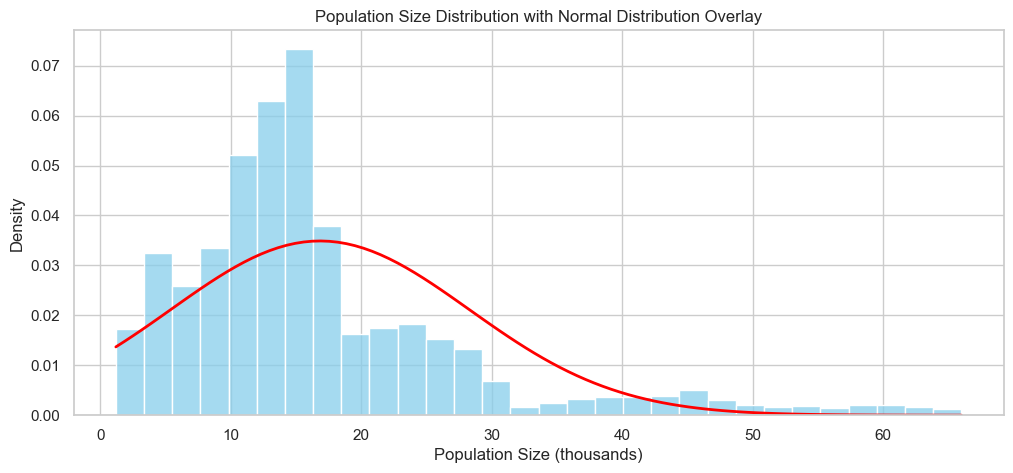
**Large Sample Behavior:**

For large samples, we would expect the binomial distribution to resemble a normal distribution due to the Central Limit Theorem. In the case of the Poisson distribution, the average rate λ becomes large, it too begins to resemble a normal distribution. However, with only 8 regions, we do not have a large enough sample to observe these asymptotic behaviors directly in our dataset.

The plots visualize the discrete probability distributions based on the data from the age group 20-24, highlighting the probability of a region exceeding the population size threshold, as well as the expected frequency of such events across regions

## **A.5. Normal Distribution**

* The overall mean (μ) across all age groups is approximately 16.85 thousand individuals.
* The overall standard deviation (σ) is around 11.43 thousand, indicating the variation in population sizes across all regions and age groups.



The red-plotted curve shows the calculated overall mean and standard deviation values.

**Observations:**

* The histogram displays the entire spread of population sizes across all age groups, and it appears to be roughly symmetric around the mean. However, there's an indication of slight right-skewness, which is often present in population data.
* The Normal distribution curve provides an approximation across the entire range of population sizes, but it may not capture all features of the distribution, especially since the data comprises multiple age groups.

**Large Sample Behavior:**

* With a large enough sample size, we could use the Normal distribution to estimate probabilities and conduct hypothesis tests about the population mean, even if individual data points are not normally distributed.
* The empirical rule (68-95-99.7 rule) for Normal distributions can be applied to interpret the spread of the data relative to the mean and standard deviations if the overall distribution were Normal.

# **B. Data Preparation and Visualization**

## **B.1 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is a critical step in the data analysis process because it allows us to understand the underlying structures, detect anomalies, test assumptions, and check for underlying patterns in the data.

**Descriptive Statistics**:  
**Rationale**: To get a basic understanding of the central tendencies and dispersions within the data.  
**Method**: Calculate mean, median, mode, range, variance, and standard deviation for the population sizes.  
**Insight**: Understanding the average population size, the spread, and the commonality of data points, which are fundamental for any subsequent analysis.

**Distribution Plots:  
Rationale**: To visually assess the distribution of the population data and check for normality, skewness, or outliers. **Method**: Histograms for the overall population and for specific age groups, and boxplots for comparing distributions across age groups or regions. **Insight**: Identification of any skewness in the data, which could influence the use of certain statistical tests or models that assume normality.

**Trend Analysis**:   
**Rationale**: To investigate any patterns or trends over time.  
**Method**: Line plots showing the population over the years.  
**Insight**: Understanding of growth or decline in the population, which could be significant for forecasting and planning purposes.

**Comparative Analysis**:  
**Rationale**: To compare population sizes across different groups.  
**Method**: Bar plots to compare the population sizes across different age groups and regions.  
**Insight**: Recognizing which age groups or regions have larger or smaller populations, which is vital for targeted policy-making and resource allocation.

**Missing Values and Data Quality Check:  
Rationale**: To ensure the integrity of the analysis. **Method**: Checking for null or missing values, and inconsistencies in the dataset. **Insight**: Ensuring that the dataset is complete and accurate before proceeding with any further analysis.

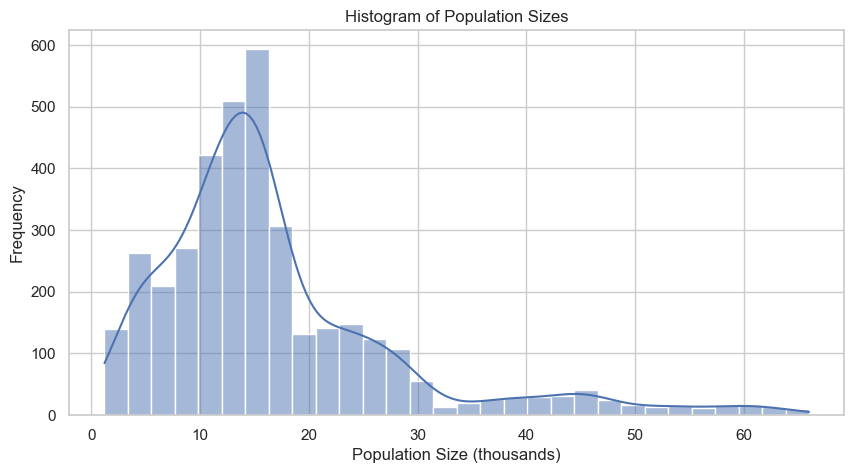
## **B.2. Descriptive Statistics:**

The descriptive statistics for the population values (VALUE) provide an initial understanding of the dataset:

* There are 3,744 data points (count).
* The average population size (mean) is approximately 16.85 thousand.
* The median (50% percentile) is 14.3 thousand, which is less than the mean, indicating a slight right skew.
* The population sizes vary with a standard deviation of about 11.43 thousand.
* The minimum population size is 1.2 thousand, and the maximum is 66 thousand.

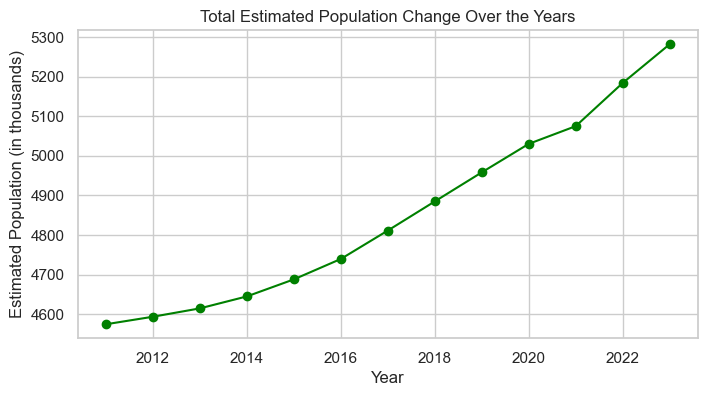
## **B.3. Distribution Plots:**

The histogram for the overall population values shows a distribution that is approximately normal but with a slight right skew (skewness of 1.77), as indicated by the mean being larger than the median. A kernel density estimate (KDE) overlay provides a smooth estimate of the distribution.

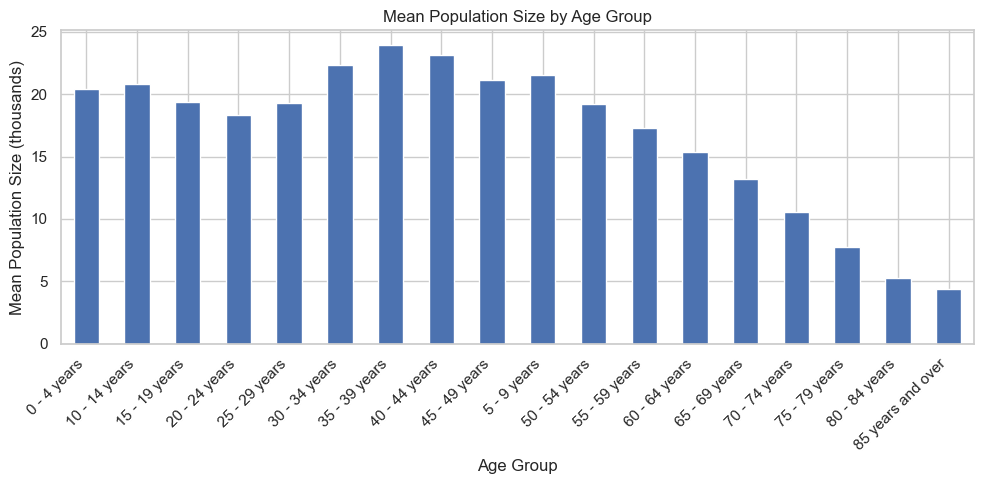


**Trend Analysis:**

The line plot of the population over the years illustrates the total population size for each year, showing fluctuations over time. This can indicate economic, social, or policy changes affecting population growth.



**Comparative Analysis:**



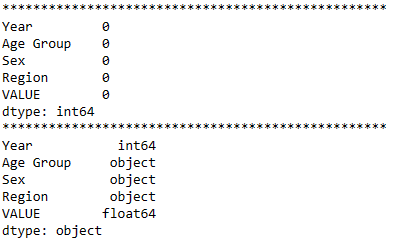
The bar plots for all the age groups and regions reveal that certain age groups and regions have significantly higher populations. This information could be critical for targeted interventions or resource allocations.

**Distribution of Population among different Regions across Age-groups**



* "30-34 years" for a region like "South-West" is significantly darker than other columns, which indicates that this age group is particularly populous in the South-West region.
* "65-69 years" consistently shows lighter shades across most regions, which suggests that this age group generally has a lower population compared to other age groups.
* Some regions show an overall darker hue across all age groups, indicating they are generally more populated than other regions like Dublin.
* A specific region has a unique distribution, such as being predominantly darker in younger age groups and lighter in older age groups, indicating a younger population demographic like Dublin, Middle East, and South-West.
* A region like the South-East shows uniformity because it has the same color across all ages.

**Missing Values and Data Quality Check:**



There are no missing values in the dataset, suggesting good data quality and completeness.

# **C. Preparing Data for ML**

## **C.1. Identify Categorical Variables:**

There are 3 Categorical Variables. E.g., 'Sex', 'Age Group', and 'Region' in our dataset.

After identifying Categorical Variables, we need to do proper encoding for them.

**C.2. Encoding Categorical Variables**:

We used Label Encoding to convert the binary categorical variable 'Sex' into a numeric format that can be understood by ML algorithms. 0 stands for Male and 1 stands for Female.

We applied OneHot Encoding to the 'Region' variable, which has multiple categories. This process created a new binary (0 or 1) column for each region category. The region has 6 unique values in it so we can safely encode it without getting worries for column expansions.

For 'Age Group', which is an ordinal variable, we used Ordinal Encoding. This method assigns a numeric value to each category based on order. Since 'Age Group' has a natural ordering (younger to older), this encoding captures that relationship accordingly.

**Data Splitting**: The dataset was split into features (X) and the target variable (y), where 'X' contains the encoded categorical variables and other features, and 'y' is the population size we want to predict. We then split these into training and test sets, with 80% of the data reserved for training and 20% for testing.

Now as the data is fully ready to apply ML model. The next step that will be a part are:

**Model Training**: Train regression models on the training set. Considering the nature of the data, we could start with simple linear regression and then explore more complex models if necessary.

**Model Selection**: The model suited for this dataset are Linear Regression and Random Forest Regressor.

**Model Evaluation**: Evaluate the final model on the test set using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared to assess its predictive performance.

**Interpretation of Results**: Interpret the model coefficients to understand the influence of each feature on the population size.

## **C.3. Machine Learning for Data Analytics**

### **C.3.1 Project Management Framework**

When managing a data science project, selecting an appropriate project management framework is vital to guide the process from conception to deployment. Three well-known frameworks are CRISP-DM (Cross-Industry Standard Process for Data Mining), KDD (Knowledge Discovery in Databases), and SEMMA (Sample, Explore, Modify, Model, Assess). Each has its strengths and is suited to different types of projects.

### **C.3.2 CRISP-DM:**

**Description**: CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It's a robust and well-established framework that outlines a structured approach to data mining projects. It consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

**Justification with Scenario**: CRISP-DM is particularly suited for projects that require a deep alignment with business objectives and involve iterative explorations of data and models. For instance, a retail company looking to enhance customer segmentation for targeted marketing campaigns would benefit from CRISP-DM. It allows for a cyclical process where business understanding guides data preparation and modeling, and results can be directly tied back to business implications.

### **C.3.3. KDD:**

**Description**: KDD, or Knowledge Discovery in Databases, is a framework that focuses on the extraction of useful knowledge from large volumes of data. It involves selection, preprocessing, transformation, data mining, and interpretation/evaluation.

**Justification with Scenario**: KDD is ideal for projects where the primary goal is the discovery of patterns and knowledge from large datasets, often without a predefined hypothesis. An example could be a data-driven research institution analyzing satellite data to identify environmental changes. The emphasis is on uncovering new patterns that may not have been previously considered.

### **C.3.4. SEMMA:**

**Description**: SEMMA stands for Sample, Explore, Modify, Model, Assess. It is a sequence of steps that guide the process of data mining. It is a framework developed by SAS Institute Inc.

**Justification with Scenario**: SEMMA is well-suited for scenarios where data is vast and sampling is essential before any analysis, such as in pharmaceutical research where initial experiments are conducted on a small scale before broader application. The framework is highly structured, which is beneficial in tightly regulated environments that require rigorous documentation and validation of each step.

## **C.4. Machine Learning Technique:**

The choice between supervised, unsupervised, or semi-supervised learning depends on the nature of the problem and the data available:

**Supervised Learning**: This approach is chosen when the dataset includes both input features and known output labels. It's suitable when the task is to predict or classify based on these labels. In our dataset, we used supervised learning techniques such as Linear Regression and Random Forest Regression because we have a clear target variable (population size) and a set of input features. The goal is to predict the population size, making supervised learning the appropriate choice.

**Unsupervised Learning**: This would be selected if we only had input data without labeled responses and wanted to uncover patterns or structures within the data, such as clustering similar data points or reducing dimensionality.

**Semi-Supervised Learning**: This is a blend of the previous two, often used when a large amount of input data is unlabeled, but some labeled examples are available. It's not applicable in our scenario as we have labeled data for all our observations.

## **C.5. Model Training and Evaluation**

The aim of this analysis was to develop models to predict population values (in thousands) using a dataset containing features like 'Year', 'Age Group', 'Sex', and various regional indicators.

**Linear Regression**

**Data Splitting**: The dataset was divided into two parts: a training set (80%) for building the models, and a testing set (20%) for evaluating their performance. We use the Sklearn function train\_test\_split () to perform splitting of dataset.

## **C.6. Model Selection and Training:**

Two types of models were chosen for this analysis:

**Linear Regression**: A basic model for establishing a baseline performance. Work better

**Random Forest**: A more complex model known for its high accuracy in various scenarios.

These models were trained using the training dataset.

## **C.7. Model Evaluation:**

The models were evaluated on the test dataset using two metrics:

**Mean Squared Error (MSE):** A lower MSE indicates a model's predictions are closer to the actual values.

**R-squared (R²):** This represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Closer to 1 is better.

**Parameter Optimization**: Using **GridSearchCV**, an advanced technique for hyperparameter tuning, the Random Forest model is further optimized.

**Linear Regression Model:**

**MSE**: 33.10

**R²**: 0.759

**Interpretation**: The model had moderate accuracy, with approximately 75.9% of the variability in the target variable explained by the model.

**Initial Random Forest Model**:

**MSE**: 0.338

**R²**: 0.998

**Interpretation**: This model showed high accuracy and excellent fit to the data, significantly outperforming the Linear Regression model.

**Optimized Random Forest Model**:

**MSE**: 1.484

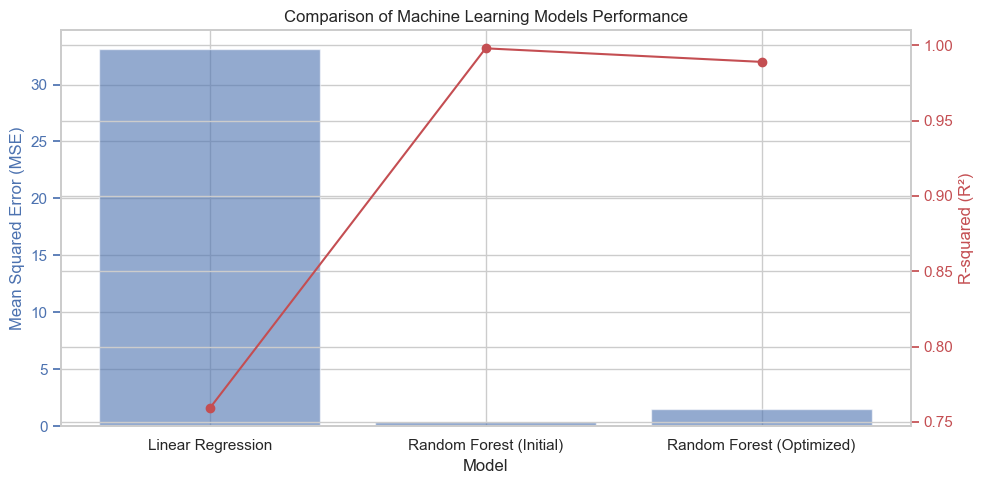
**R²:** 0.989

**Interpretation**: Although the MSE slightly increased compared to the initial Random Forest model, the optimized model still exhibited high accuracy and a strong fit to the data.

**C.8. Comparison**:

### **C.8.1 Comparison Table:**

|  |  |  |
| --- | --- | --- |
| Model | MSE | R2 |
| Linear Regression | 33.10 | 0.759 |
| Random Forest Initial | 0.338 | 0.998 |
| Random Forest Optimized | 1.484 | 0.989 |



### **C.8.2 Linear Regression:**

* It is a simple and interpretable model, often used as a baseline in machine learning tasks.
* The moderate R² value indicates that while the model captures some of the variance in the data, it's not highly precise.
* The high MSE compared to the Random Forest models suggests it is less accurate in predicting population values.

### **C.8.3 Random Forest (Initial & Optimized):**

* Both versions of the Random Forest model significantly outperform the Linear Regression model.
* The high R² values indicate these models are very effective in capturing the variance in the data.
* The difference in MSE between the initial and optimized Random Forest models (0.338 vs. 1.484) suggests a potential trade-off between bias and variance, where the optimized model may be slightly overfitting.

### **C.8.4 Similarities and Differences:**

**Similarity**: Both Random Forest models exhibit a high degree of accuracy and strong predictive capabilities, as evidenced by their high R² scores.

**Difference**: The initial Random Forest model has a lower MSE, indicating better accuracy in predictions. However, the optimized model, despite a slightly higher MSE, maybe capturing more complex patterns in the data.

**Relevance and Effectiveness**: The Random Forest models are highly effective for this specific dataset, especially for capturing complex relationships between features and the target variable. The Linear Regression model, while less accurate, offers a simpler and more interpretable alternative.

### **C.8.5 Graphical Comparison:**

The bar chart above provides a clear visual comparison of the performance metrics (Mean Squared Error and R-squared) for the three models: Linear Regression, Random Forest (Initial), and Random Forest (Optimized).

Blue Bars (MSE): Represent the Mean Squared Error for each model. Lower values are better, indicating that the model's predictions are closer to the actual values.

Red Line (R²): Shows the R-squared values, where higher values (closer to 1) suggest a better fit of the model to the data.

From the chart, it's evident that while the Linear Regression model has a significantly higher MSE, indicating less accuracy, the Random Forest models (both initial and optimized) demonstrate superior performance with much lower MSE values and higher R² scores.

### **C.8.6 Interpretation**

The Random Forest models, particularly the optimized version, are highly effective in capturing the complex relationships in the dataset. This makes them well-suited for predictive tasks where accuracy is paramount.

The Linear Regression model, despite its lower performance metrics, provides a simpler and more interpretable option. This could be valuable in scenarios where model interpretability is as important as predictive accuracy.