**Naan Mudhalvan Project**

**TN Marginal Workers Assessment**

**Phase 5**

**TEAM MEMBERS**

**Ranjith Kumar 2021115084**

**Rakshitha Senthil 2021115082**

**Ramkrishna 2021115083**

**Priyadarshini 2021115078**

**Saiganesh 2021115316**

**OVERVIEW :**

**1. Project Definition:**

The project involves analyzing the demographic characteristics of marginal workers in TamilNadu based on their age, industrial category, and sex. The objective is to perform a socioeconomic analysis and create visualizations to represent the distribution of marginal workers across different categories. This project includes defining objectives, designing the analysis approach, selecting appropriate visualization types, and performing the analysis using Python and data visualization libraries.

1. **Design Thinking for Analyzing Demographic Characteristics of Marginal Workers in TamilNadu:**

**2.1 Analysis Objectives:**

Before delving into the data analysis, it's essential to establish clear objectives for our project. In this phase, we define the following analysis objectives to provide a well defined direction for our analysis, focusing on key demographic factors related to marginal workers:

**\***Compare the age distribution of marginal workers within different industrial categories in Tamil Nadu.

**\***Examine the gender distribution among marginal workers across various age groups and industries in Tamil Nadu.

**2.2 Data Collection:**

To proceed with the analysis, we need to gather the necessary data. We will obtain the dataset containing information on demographic characteristics of marginal workers in Tamil Nadu. Data sources may include government reports, surveys, or other relevant sources. Data cleaning and preprocessing will be essential to handle any missing values, format inconsistencies, and outliers.

**Dataset Link:**  
[**https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil**](https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil)

**2.3 Visualization Strategy:**

Effective data visualization is crucial for conveying insights. In this phase, we plan how to visualize the demographic characteristics of marginal workers using Python and data visualization libraries. Our visualization strategy includes:

\* Selecting appropriate chart types (e.g., bar charts, pie charts, scatter plots) to represent age and gender distribution.Creating visualizations that facilitate comparisons across different industrial categories and age groups.

\*Ensuring visualizations are interactive and can be filtered by specific demographics or industrial categories.

\* Adding labels, legends, and tooltips for clarity and context.

Python and relevant data visualization libraries, such as Matplotlib or Seaborn, will be employed to create dynamic visualizations that aid in understanding the demographic data**.**

**2.4 Insights Generation:**

The primary objective of this project is to derive valuable insights from the analysis of demographic data pertaining to marginal workers in Tamil Nadu. Insights may include:

\* Identifying age groups with a higher concentration of marginal workers in specific industrial categories.

\* Revealing any gender disparities in the distribution of marginal workers within different age brackets and industries.

\* Recognizing any notable variations or patterns in the demographic composition of marginal workers.

\* Drawing connections between demographic trends and potential policy recommendations or interventions to improve the conditions of marginal workers in Tamil Nadu.

These insights will be derived through a combination of statistical analysis and visual examination of the demographic data. They will serve as valuable information for policymakers, government agencies, and organizations aiming to address the needs of marginal workers in Tamil Nadu.

**ANALYSIS TOOLS AND METHODS:**

**1. DATA COLLECTION:**

- Gather relevant demographic data related to marginal workers in Tamil Nadu, including age, industrial category, and sex, from reliable sources such as government surveys and databases.

**2. DESCRIPTIVE STATISTICS:**

- Utilize statistical measures and techniques to provide an initial overview of the data, including mean, median, and standard deviation for age.

- Calculate the frequency distribution of industrial categories and sex to identify the most common categories and gender distribution.

**3. DATA VISUALIZATION:**

- Create various visualizations to represent the data effectively, including:

- Bar charts and pie charts to show the distribution of marginal workers across different age groups and industrial categories.

- Stacked bar charts to display the intersection of age and industrial category.

- Gender-specific visualizations to highlight any gender disparities among marginal workers.

**4. SOCIOECONOMIC ANALYSIS:**

- Perform a deeper analysis of the data, considering factors such as income levels, education, and geographic distribution.

- Use clustering techniques or regression analysis to identify patterns and correlations within the data.

**5. MACHINE LEARNING:**

**K-Means Clustering:**

**Description:** K-Means is an unsupervised clustering algorithm used to group data points into clusters based on similarity.

**Application:** We will apply K-Means clustering to group marginal workers with similar demographic characteristics. This can help identify distinct groups within the population based on age, industrial category, and sex.

**Linear Regression:**

**Description:** Linear regression is a supervised learning algorithm used for regression tasks. It models the relationship between the dependent variable and one or more independent variables.

**Application:** We can employ linear regression to predict factors like income or education level based on demographic characteristics like age, allowing for a deeper understanding of the socioeconomic status of marginal workers.

**Random Forest:**

**Description:** Random Forest is an ensemble learning technique that combines multiple decision trees to improve accuracy and reduce overfitting.

**Application:** We can use Random Forest for both classification and regression tasks, providing a more robust analysis of the data. This algorithm can help analyze the socioeconomic status and distribution of marginal workers, considering a range of demographic characteristics.

**Support Vector Machines (SVM):**

**Description:** SVM is a supervised learning algorithm that is effective for classification tasks. It finds a hyperplane that best separates data into distinct classes.

**Application**: SVM can be applied to classify marginal workers into different industrial categories based on their demographic characteristics, helping to identify patterns and disparities in employment across different categories.

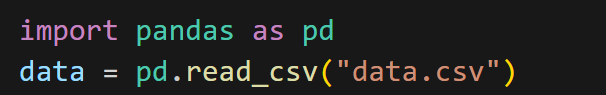
***Executive Summary:***

This report outlines the data preprocessing steps undertaken for the "TN Marginal Workers Assessment" project. The primary objective of this preprocessing is to prepare the dataset for subsequent analysis, ensuring data consistency and completeness.

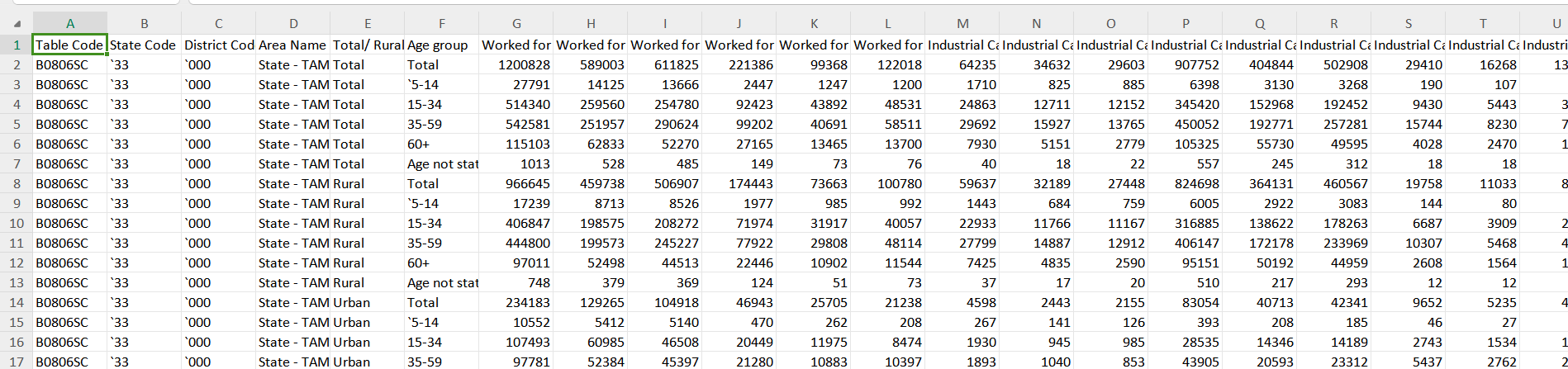
***Data Preprocessing Steps***

**Step 1:** Data Loading

- The dataset was loaded from the provided CSV file.

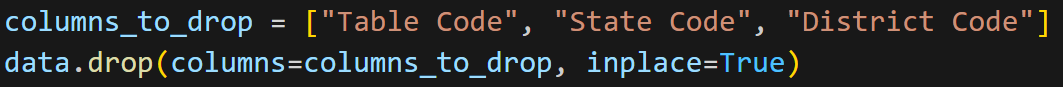


**INITIAL DATASET**

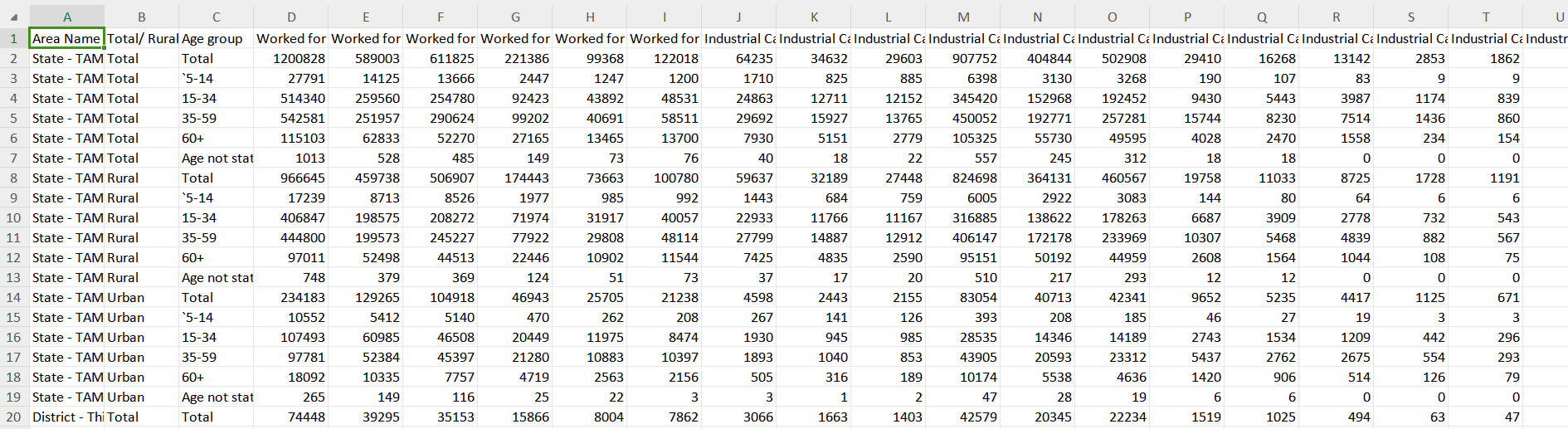


**Step 2:** Column Removal

- Unnecessary columns, namely "Table Code," "State Code," and "District Code," were removed to simplify the dataset and focus on relevant attributes.

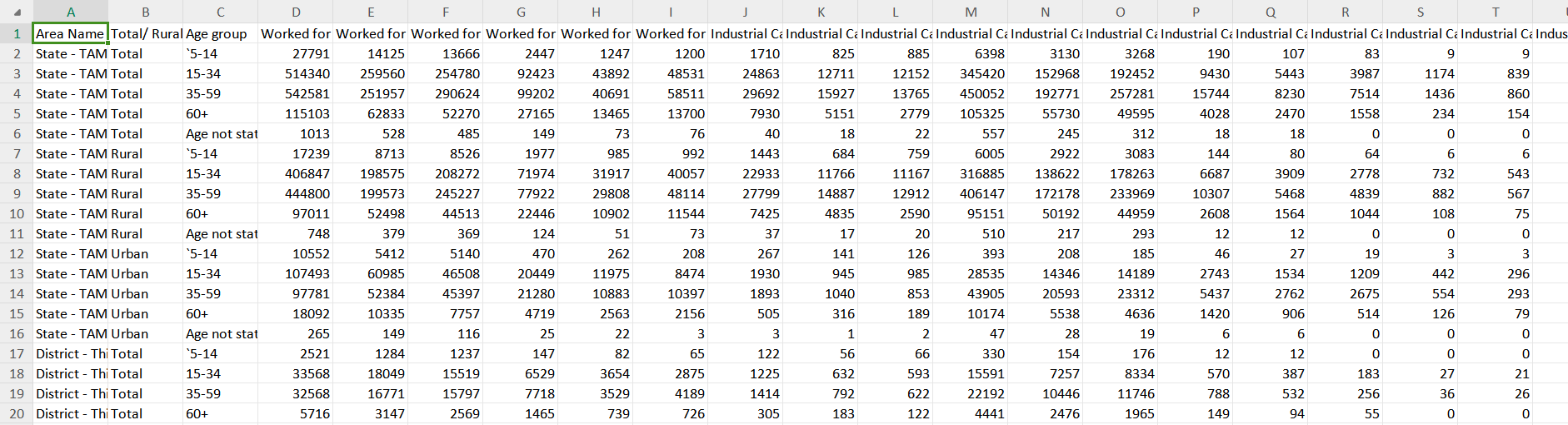


**OUTPUT:**



**Step 3:** Removal of Rows with "Total" in Age Group

- Rows with "Total" in the "Age Group" column were eliminated as they represent aggregated data and do not pertain to individual age groups.



**Step 4**: Age Group Mapping

- The "Age Group" column was standardized to ensure a consistent format.

- Specific mappings:

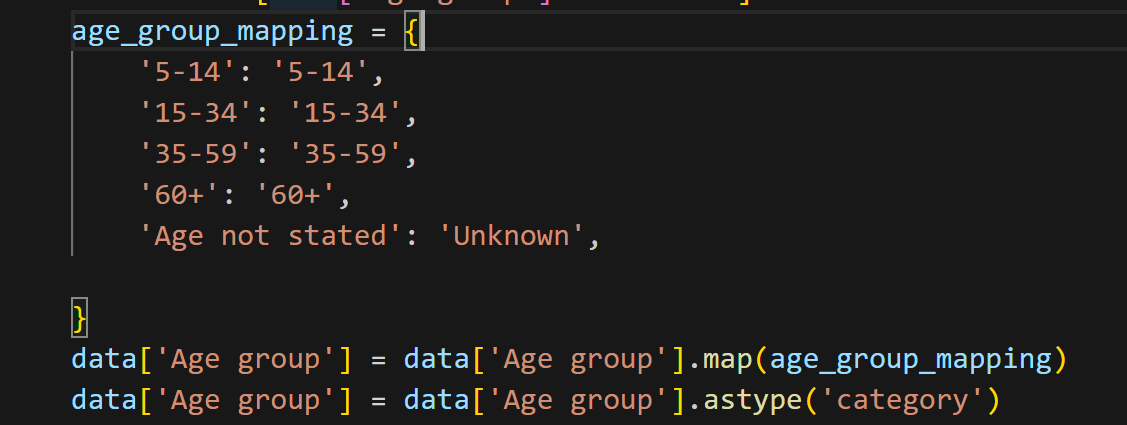
- "5-14" remains "5-14"

- "15-34" remains "15-34"

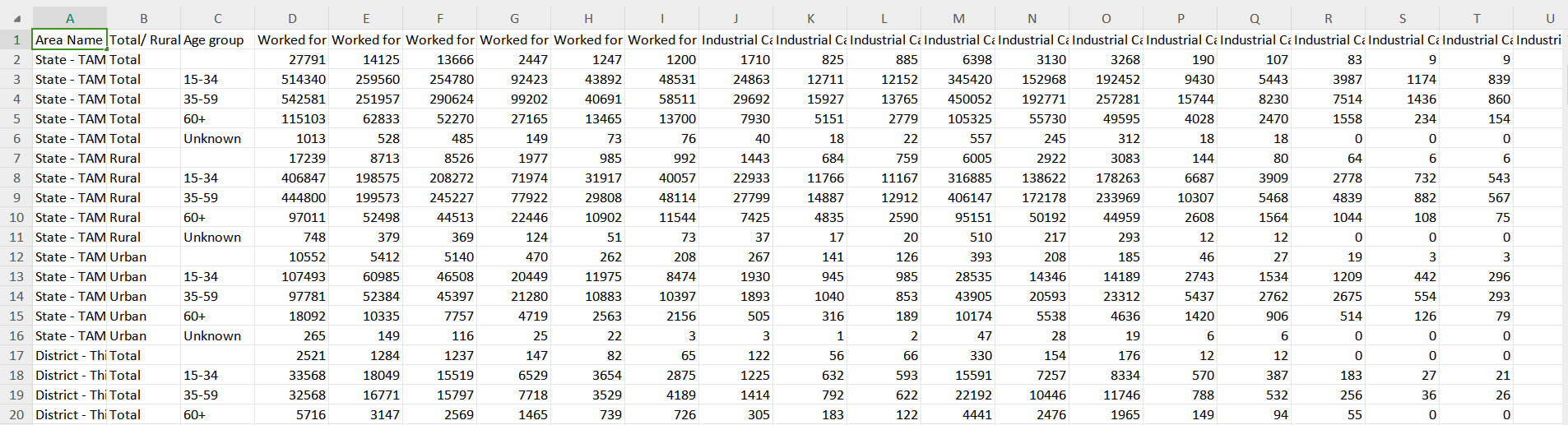
- "35-59" remains "35-59"

- "60+" remains "60+"

- "Age not stated" has been replaced with "Unknown" to address missing or unspecified age group entries.



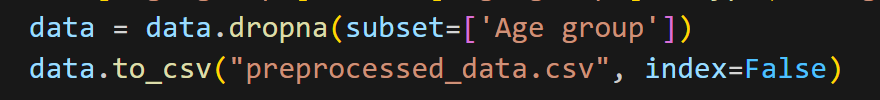
**OUTPUT:**

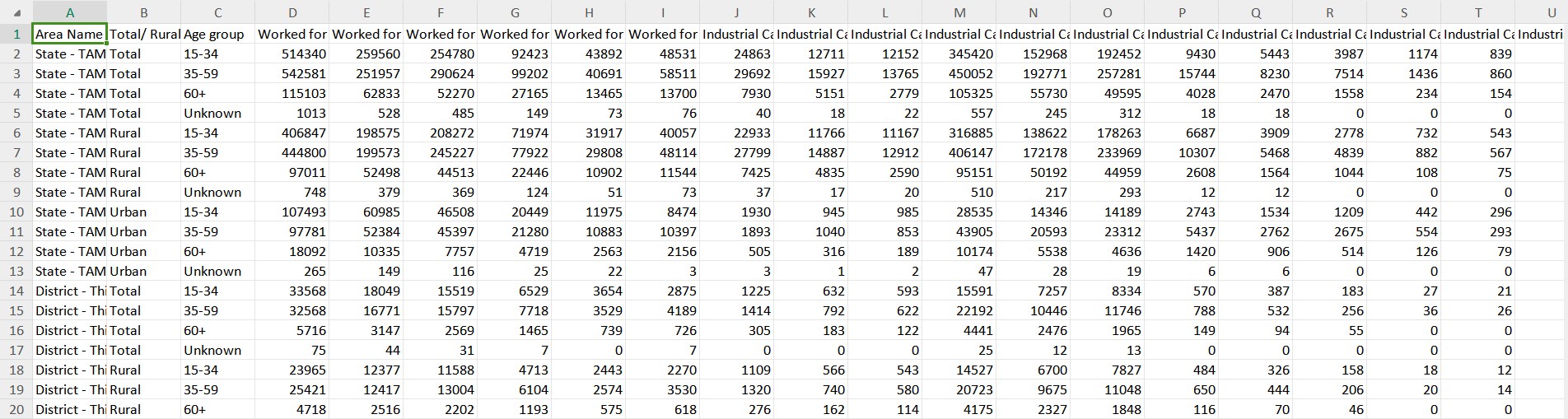


**Step 5:** Removal of Rows with Empty Entries in Age Group and Data Saving

- Rows containing empty or missing entries in the "Age Group" column were removed to maintain data consistency.

- The preprocessed data was saved to a new CSV file, "preprocessed\_data.csv," making it suitable for further analysis.





***Project Scope***

The scope of this project encompasses several key aspects:

**1. Data Collection:** Acquiring and accessing the relevant dataset containing information about marginal workers in Tamil Nadu.

**2. Data Preprocessing:** Preparing the dataset for analysis by cleaning and structuring the data. This involves removing irrelevant columns and handling missing or erroneous data points. Additionally, filtering out the 'Total' category from the 'Age group' column to focus on specific age groups.

**3. Data Analysis:** Delving into the dataset to extract insights about the demographic characteristics of marginal workers. This phase involves statistical calculations and exploration of patterns and trends.

**4. Data Visualization:** Creating visualizations that vividly represent the findings of the analysis. These visualizations are a critical component of the project, as they make complex data more understandable and help stakeholders grasp the key insights.

***Approach***

The project follows a systematic approach:

**1. Data Collection:**

- The dataset is obtained from [Specify the data source]. This dataset contains information about marginal workers in Tamil Nadu, covering a range of demographic variables and categories.

**2. Data Preprocessing:**

- Irrelevant columns such as 'Table Code,' 'State Code,' and 'District Code' are removed to focus on essential attributes.

- 'Total' rows in the 'Age group' column are filtered out, as the analysis is intended to be age-specific.

- Age group categories are mapped to more descriptive labels to improve clarity and understanding.

- Missing values are handled to ensure data integrity.

***Data Visualizations***

**1. Bar Plots:**

- Bar plots are generated to visualize the distribution of various characteristics of marginal workers segmented by age groups. These plots offer a clear and concise representation of data distribution.

**2. Histogram:**

- A histogram is created to illustrate the distribution of individuals who worked for 3-6 months. This histogram provides an overview of the distribution of this specific category.

**3. Box Plot:**

- A box plot is generated to show the distribution of individuals who worked for 3-6 months, segmented by age groups. It helps identify variations in different age groups.

**4. Scatter Plot:**

- A scatter plot is produced to reveal the relationship between the number of individuals who worked for 3-6 months and the number of males in that category. Scatter plots are useful for understanding correlations.

**5. Heatmap:**

- A heatmap is designed to visualize correlations between selected columns. This heatmap provides insights into the relationships between key variables.

***Code Explanation***

The project code begins by loading the dataset and performing data preprocessing. Irrelevant columns are removed, the 'Total' category is filtered out from the 'Age group' column, age group categories are mapped to descriptive labels, and missing values are handled to ensure data quality and accuracy.

The subsequent data visualization phase is critical for the project's success. The chosen visualization types help convey complex information in a clear and accessible manner. These visualizations, including bar plots, histograms, box plots, scatter plots, and a heatmap, offer unique insights into the demographic characteristics of marginal workers.

***Conclusion***

The data analysis and visualizations have unearthed valuable insights into the demographic characteristics of marginal workers in Tamil Nadu. These insights include:

- A clear concentration of marginal workers within specific age groups.

- Notable variations in the number of males engaged in 3-6 months of work across different age groups.

- The heatmap shows varying degrees of correlation between selected columns.

The findings from this analysis can serve as valuable inputs for policymaking and the development of initiatives aimed at supporting and improving the socioeconomic conditions of marginalized workers in Tamil Nadu.