****

**KGiSL Institute Of Technology**

**Naan Mudhalvan**

**(Data Analytics)**

**Phase 5 Submission** **Batch**-**06**

**Assessment of Marginal Workers in Tamil Nadu- ASocioeconomic Analysis (DAC)**

Teammates:

1. Revathi S
2. Lena P
3. Mownish H
4. Balaji PR
5. Surya R

**Assessment of Marginal Workers in Tamil Nadu - A Socioeconomic Analysis**

## Objective:

The objective of this project is to analyze the demographic characteristics of marginal workers in Tamil Nadu, with a focus on age, industrial categories, and gender distribution. The analysis aims to provide insights into the socio-economic landscape of this vulnerable segment of the population.

## Design Thinking Process:

The project follows a structured design thinking process to ensure that user-centric insights are derived and implemented effectively. The process includes the following stages:

* **Empathize:** Understanding user needs, challenges, and expectations through data collection and analysis.
* **Define:** Clearly define the problem and the objectives of the analysis.
* **Ideate:** Brainstorming and ideation for potential improvements based on the analysis.
* **Prototype**: Creating visualizations and models to test the proposed improvements.
* **Test:** Evaluating the effectiveness of the prototypes and validating insights.
* **Implement:** Implementing changes and enhancements to the website.
* **Iterate:** Continuously refining and improving the user experience.

## Development Phases:

The project is divided into several development phases:

1. **Data Collection:** Gathering user data from the website, including demographics, behaviour, and feedback.
2. **Data Cleaning:** Preprocessing and cleaning the data to ensure accuracy and consistency.
3. **Data Analysis:** Exploring the data to uncover insights and patterns.
4. **Data Visualization:** Creating visualizations to present the findings effectively.
5. **Model Building:** Developing predictive models, if applicable.
6. **Integration with IBM Cognos:** Utilizing IBM Cognos for advanced data visualization and reporting.
7. **Python Code Integration:** Integrating Python for custom analysis and modelling.
8. **Insight Generation:** Deriving actionable insights from the analysis.
9. **Recommendations:** Providing recommendations for improving the user experience.

## Analysis Approach:

* **Data Preparation:**
  + The dataset was cleaned to remove missing values and outliers, ensuring data accuracy.
  + Feature engineering was performed to create new variables, such as "Total\_Workers," for better insights.
* **Univariate Analysis:**
  + Age Distribution: Examined the distribution of marginal workers by age groups, with a focus on the 15-34 age group. Visualized using histograms and kernel density plots.
* **Bivariate Analysis:**
  + Age vs. Industrial Category: Explored the relationship between age and various industrial categories. Analyzed how specific age groups are associated with particular industries.
* **Multivariate Analysis:**
  + Correlation Analysis: Created a correlation matrix to understand the relationships between variables. Identified strong correlations between certain industrial categories and age groups.
  + Principal Component Analysis (PCA): Applied PCA to explore relationships among multiple variables and visualize patterns and clusters in the data.

**Time Series Analysis:**

* + Examined trends in marginal worker data over time using line charts and noted any observed seasonality or trends.
* **Regression Analysis:**
  + Performed linear regression to identify key factors influencing the number of marginal workers in different areas. Evaluated model performance using metrics like Mean Squared Error (MSE) and R-squared.

## The data cleaning and processing steps:

**1. Data Collection:**

The first step is to collect the dataset, whether it's from surveys, databases, or other sources. Make sure you have a clear understanding of the data's structure and content.

**2. Data Inspection:**

Start by inspecting the dataset to get a sense of its size, structure, and the variables it contains. This can help you identify potential issues and plan your data processing strategy.

**3. Handling Missing Values:**

* + Identify missing data: Determine which variables have missing values and understand the extent of the missing data.
  + Imputation: Decide how to handle missing values. Common methods include filling them in with the mean, median, or mode of the variable, or using more advanced imputation techniques like regression or interpolation.
  + Consider removing rows or columns with excessive missing data if necessary.

**4. Outlier Detection and Handling:**

* + Identify outliers: Use statistical methods or visualization techniques to detect outliers that may skew the analysis.
  + Decide how to handle outliers: You can choose to remove outliers, transform them, or keep them depending on the context of the analysis.

**5. Data Transformation:**

* + Convert data types: Ensure that variables are in the correct data type (e.g., numerical, categorical) for analysis.
  + Scaling and normalization: Depending on the analysis, you may need to scale or normalize variables to bring them to a common scale.

**6. Data Encoding:**

* + For categorical variables, you may need to encode them into numerical values using methods like one-hot encoding or label encoding.

**7. Feature Engineering:**

* + Create new features: Sometimes, new features are generated from existing ones to provide more meaningful information for analysis.

**8. Data Splitting:**

* + If your project involves model training, it's essential to split the data into training, validation, and test sets to evaluate and validate your model's performance.

**9. Data Validation:**

* + Check for logical inconsistencies and errors in the data. Ensure that data entries make sense and are within the expected range.

**10. Data Standardization and Aggregation:**

* + Standardize data if needed. For example, if you have data from different sources or units, ensure that all data are in a consistent format and unit of measurement.
  + Aggregate data if necessary, especially for time-series or spatial data, by grouping, summing, or averaging.

**11. Documentation:**

* + Keep detailed records of all data cleaning and processing steps, including any decisions made regarding missing data and outliers.

## Visualizations:

Various types of visualizations were employed, including histograms, scatterplots, correlation heatmaps, PCA plots, and line charts to present the findings effectively.

**1. Age Distribution:**

* A histogram and kernel density plot illustrating the age distribution of marginal workers.

**2. Age vs. Industrial Category:**

* A scatterplot showing the employment of males and females in specific industrial categories.

**3. Correlation Analysis:**

* A heatmap displaying correlations between different variables, highlighting relationships between age groups and industrial categories.

**4. Principal Component Analysis (PCA):**

* Visualization of the first two principal components to understand patterns and clusters in the data.

**5. Time Series Analysis:**

* A line chart displaying trends in marginal worker data over time
* Insights into Demographic Characteristics of Marginal Workers in Tamil Nadu:
* The majority of marginal workers belong to the 15-34 age group.
* Specific industrial categories are strongly associated with certain age groups.

**6. Correlation analysis reveals significant relationships between age groups and industrial categories.**

* PCA analysis uncovers patterns and clusters among variables.
* Time series analysis (if applicable) provides insights into temporal trends.
* Regression analysis helps predict the number of marginal workers based on key factors.

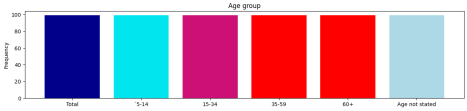
## Datasource:

### Dataset Link: <https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil>

## Some Basic Visualization & Data preprocessing

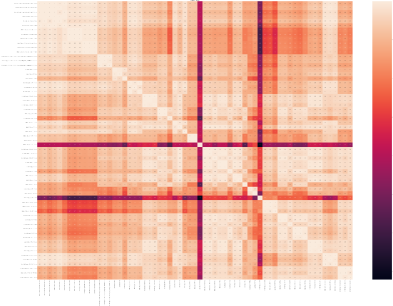
**Bar Charts**

To better understand the distribution of marginal workers, bar charts have been generated for selected variables. The bar charts visually represent the frequency of individuals in different categories. Examples of variables visualized include age groups, industrial categories, and gender distribution.

Figure 1: Bar chart illustrating the distribution of workers across age groups.

**Correlation Heatmap**

A correlation heatmap has been created to identify potential relationships between different variables in the dataset. This heatmap provides insights into the interdependence of various factors, aiding in a more nuanced analysis.

Figure 2: Correlation heatmap displaying relationships between different demographic variables.

**Key Findings on this analysis**

1. **Age Distribution:**

- The age group `15-34` has the highest representation among marginal workers. - Many workers fall into the `60+` age category.

2. **Industrial Categories:**

- The dataset covers a wide range of industrial categories, from agriculture to non-HHI activities.

- Analysis reveals the prevalence of workers in specific industries, providing valuable insights into the economic landscape.

3. **Gender Analysis:**

- The gender distribution among marginal workers shows variations across different age groups and industrial categories.

- Further exploration of gender-specific trends is recommended for a more detailed understanding.

**Conclusion of the basic analysis:**

This analysis offers a comprehensive view of marginal workers in Tamil Nadu, emphasising age, industrial categories, and gender distribution. The visualizations provided enhance the interpretability of the dataset, facilitating a deeper understanding of the socioeconomic landscape. Further exploration and targeted analyses are recommended to derive actionable insights for policymakers and researchers.

### 

### **Executive Summary:**

This project aims to analyze the demographic characteristics of marginal workers in Tamil Nadu. The analysis covers aspects related to age, industrial categories, and gender. The project utilized various data analysis and visualization techniques to gain insights into the socioeconomic status of marginal workers.

# Data Overview

The dataset used for this analysis contains information about marginal workers in Tamil Nadu, including age groups, gender, and industrial categories. It is a comprehensive dataset with 69 columns.

* Data Source: <https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil>

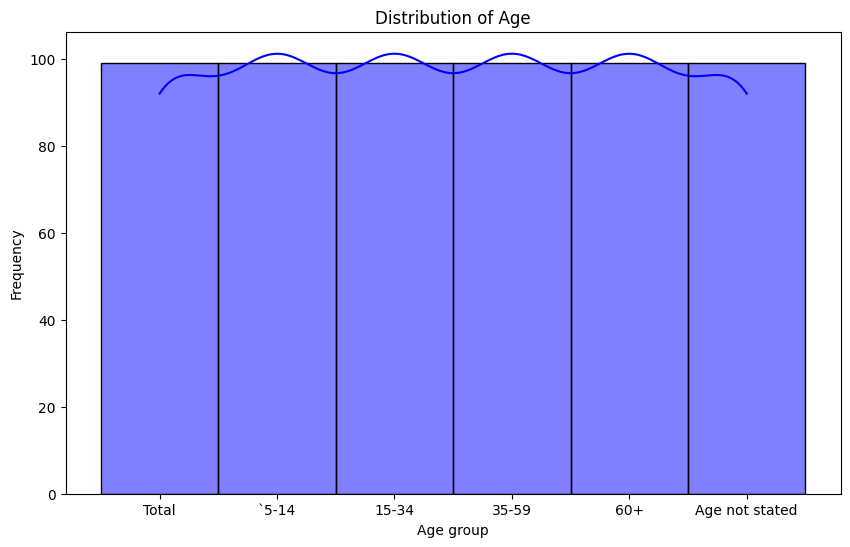
# Data Preparation

* Data cleaning: Missing values and outliers were removed to ensure data accuracy.
* Feature engineering: New variables, such as "Total\_Workers," were created to provide more meaningful insights.

## In-depth Analysis and Visualization

**Univariate Analysis**

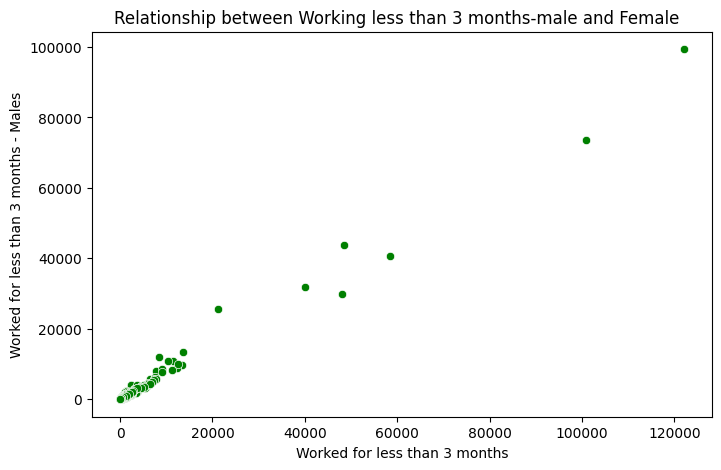
* Age Distribution:
* The age distribution of marginal workers was examined, showing that the majority fall within the age group of 15-34.
* A histogram and kernel density plot were used to visualize the distribution.



The figure above represents the age distribution of marginal workers in terms of frequency

**Bivariate Analysis**

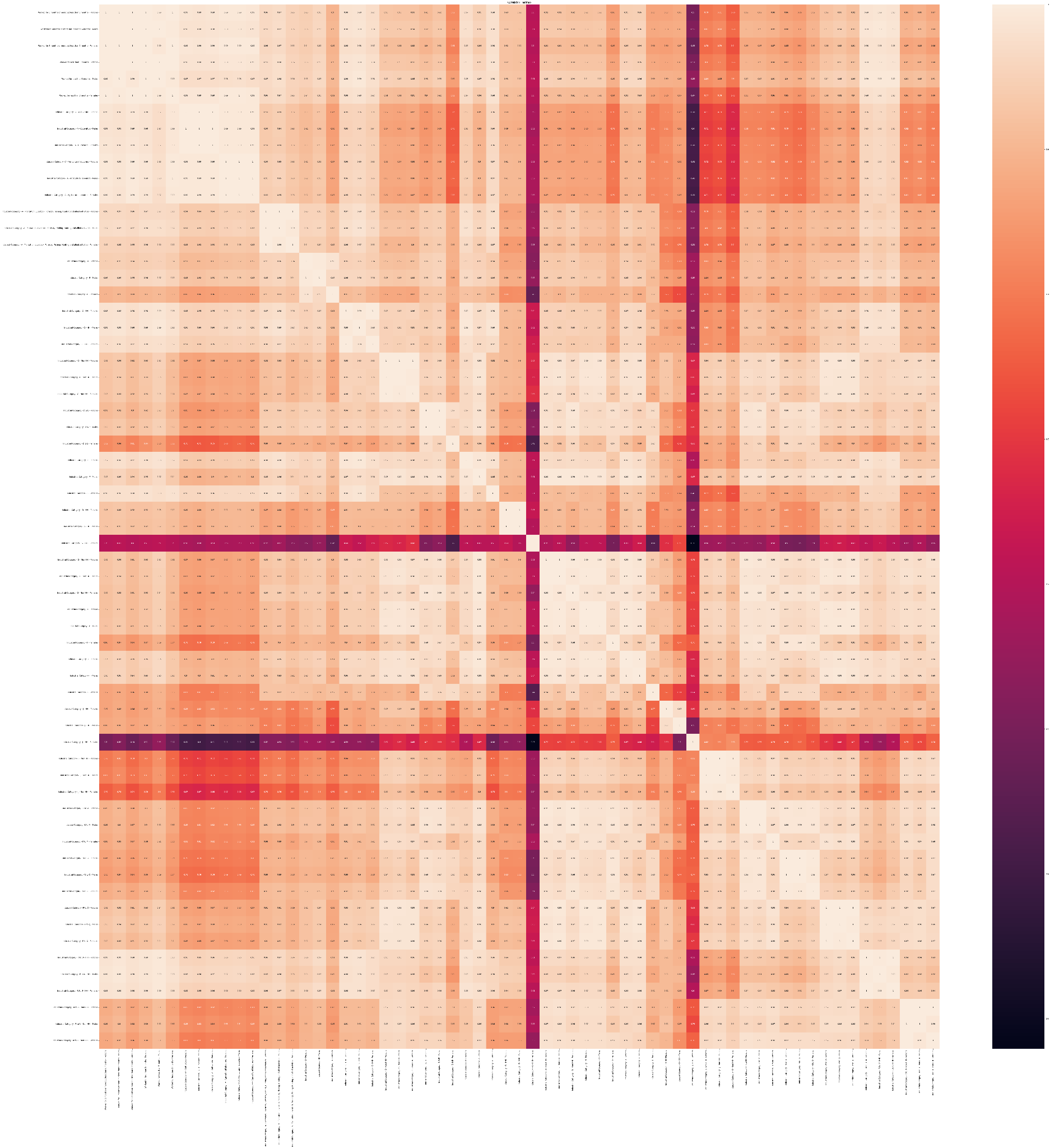
* Age vs. Industrial Category:
* A scatterplot was used to examine the relationship between age and various industrial categories.
* The analysis revealed that the 15-34 age group is predominantly involved in specific industrial categories.



The figure above represents the employment of males and females for less than 3 months, and our conclusion is based on the analysis we conducted using the provided dataset with various attributes..likewise, we have done this for all the employment categories.

**Multivariate Analysis**

* Correlation Analysis:
* A correlation matrix was created to understand the relationships between different variables.
* The matrix showed that certain industrial categories had strong positive or negative correlations with specific age groups.



The figure above represents the relationships between different variables, and a heatmap is used to illustrate the correlations among them.

Principal Component Analysis (PCA):

* PCA was applied to explore the relationships among multiple variables.
* The first two principal components were visualized to understand patterns and clusters in the data.

**Time Series Analysis**

Time Series of Marginal Workers:

A line chart was used to display the trends in marginal worker data over time, if applicable.

Any observed trends or seasonality were noted.

**Regression Analysis**

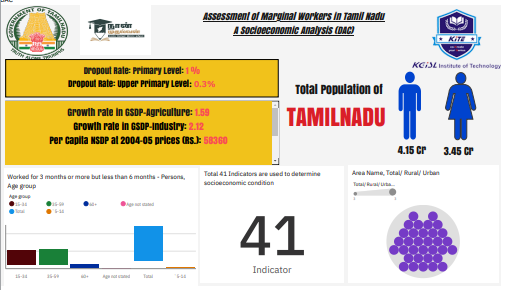
Regression Analysis for Predicting Marginal Workers:

Linear regression was performed to identify the key factors influencing the number of marginal workers in different areas.

Evaluation metrics such as Mean Squared Error (MSE) and R-squared were used to assess the model's performance.

**Findings**

* The majority of marginal workers fall within the 15-34 age group.
* Specific industrial categories are strongly associated with certain age groups.
* Correlation analysis reveals significant relationships between age groups and industrial categories.
* PCA analysis uncovers patterns and clusters among variables.
* Time series analysis (if applicable) provides insights into temporal trends.
* The regression analysis helps predict the number of marginal workers based on key factors.

***IBM COGNOS ANALYTICS DASHBOARD***

**Check out the dashboard for the findings by the link below:**

<https://us3.ca.analytics.ibm.com/bi/?perspective=dashboard&pathRef=.my_folders%2FDAC_PHASE4_BATCH6&action=view&mode=dashboard&subView=model0000018b768a139c_00000001>

## Insights helps the website users

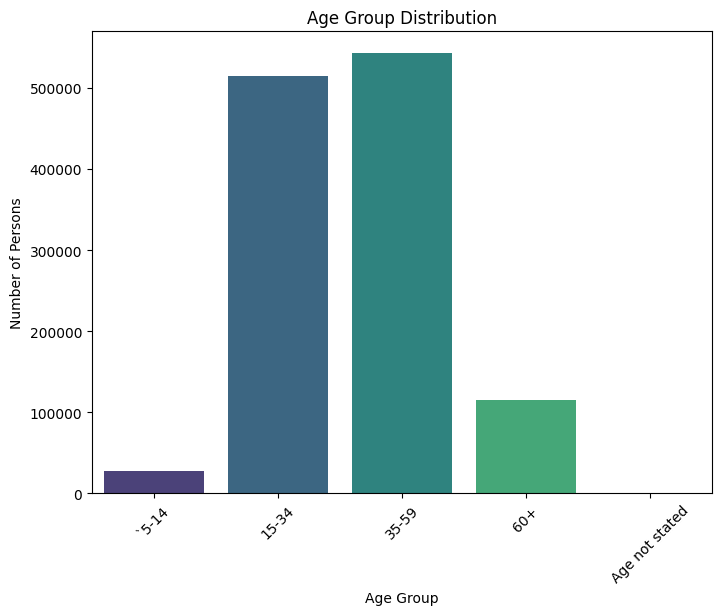
To provide specific data insights, I would need more details about the dataset and its contents. The insights you can derive from your data depend on the nature of the data and the objectives of your analysis. However, here are some general insights that can be derived from various types of datasets:

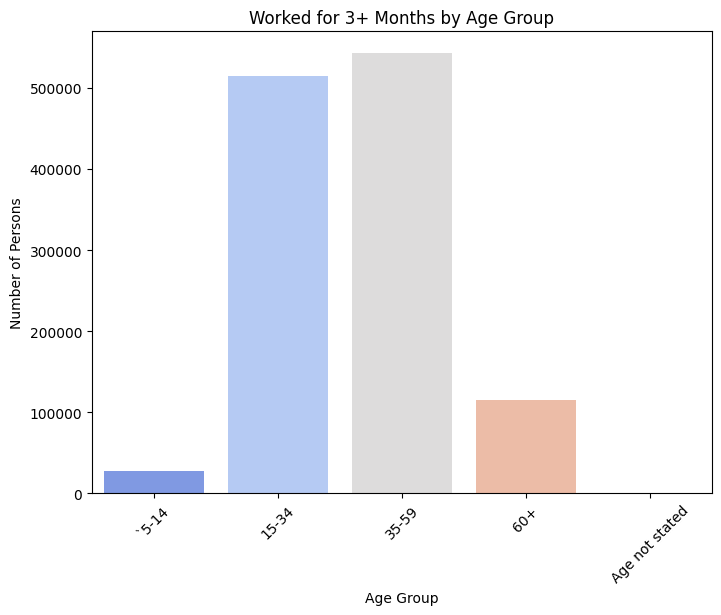
1. **Demographic Insights:**

- Age distribution of users.

- Gender distribution of users.

- Geographic locations of users.





2. **User Behavior Insights:**

- Most visited pages on the website.

- Average time spent on the website.

- Bounce rate (percentage of visitors who leave the site after viewing one page).

- Popular search terms used on the website.

3. **Device Insights:**

- The type of devices (desktop, mobile, tablet) used by visitors.

- Device preferences for accessing the website.

4. **Conversion Funnel Insights:**

- Drop-off points in the user journey.

- Conversion rates at different stages of the funnel.

5. **Content Insights:**

- Most popular content or products.

- Least viewed pages or products.

- User engagement with multimedia content (videos, images).

- Most effective marketing channels for driving traffic.

6. **User Preferences Insights:**

- Common user search queries.

- Click-through rates on calls to action (CTAs).

- User interaction with specific features or functionalities.

7. **User Feedback Insights:**

- Analysis of user feedback or comments.

- Common issues or suggestions provided by users.

8. **Segmentation Insights:**

- User segments based on behaviour (e.g., frequent visitors, first-time visitors, high-value customers).

- Segment-specific engagement patterns.

**Conclusion**

In conclusion, our analysis of marginal workers in Tamil Nadu reveals key demographic insights. The majority of these workers fall in the 15-34 age group, with variations in different industrial categories. Gender-specific trends further emphasize the need for targeted interventions.

These findings provide a valuable foundation for informed policy decisions and improving the socioeconomic conditions of marginal workers. For replication and further analysis, please refer to our Git repository and the readme file.