Project Documentation: Marginal Workers Data Analysis and Clustering

Table of Contents

1. Project Overview

The goal of this project is to analyze and cluster data related to marginal workers in Tamil Nadu. This dataset includes demographic information, as well as details about different industrial categories. The primary objectives are:

- Conduct demographic analysis to understand the age and gender distribution of marginal workers.

- Perform K-Means clustering to identify patterns among various industrial categories and age groups.

The code presented here accomplishes these objectives by leveraging Python, Pandas, Matplotlib, Seaborn, and scikit-learn's K-Means clustering algorithm.

2. Code Overview

2.1. Demographic Analysis

2.1.1. Data Loading

The project starts by loading the dataset using Pandas. Missing values are removed to ensure data cleanliness and consistency.

2.1.2. Age Distribution

The age distribution of marginal workers is visualized using a histogram with 20 bins and a kernel density estimate (KDE). This visualization provides insights into the distribution of workers across different age groups.

2.1.3. Gender Distribution

The gender distribution of marginal workers is visualized with a pie chart, displaying the percentage of total workers, male workers, and female workers.

2.1.4. Industrial Category Distribution

A bar plot is generated to visualize the distribution of marginal workers across various industrial categories. This plot offers insights into the count of workers in different industrial sectors.

2.2. K-Means Clustering

2.2.1. Data Preprocessing

Before clustering, missing values and non-numeric entries (e.g., 'Total') are removed from the dataset. Additionally, the 'Age group' column is one-hot encoded to represent age groups in a suitable numerical format.

2.2.2. Feature Selection

Industrial categories are chosen as the features for K-Means clustering. These features will be used to identify patterns among different industrial categories and age groups.

2.2.3. Choosing the Number of Clusters (k)

The optimal number of clusters (k) is determined using methods like the elbow method or silhouette score. The value of `chosen_k` can be adjusted based on the results of the analysis.

2.2.4. K-Means Clustering

The K-Means clustering algorithm is applied to the selected features with the chosen number of clusters. Cluster assignments are added to the dataset in the 'Cluster' column.

2.2.5. Visualization

The code includes an example of creating a scatter plot to visualize the clustering results. While not the primary focus of the code, additional visualization can be added for in-depth analysis.

3. Project Flow

The project flow involves data loading, demographic analysis, data preprocessing, and K-Means clustering. The resulting clusters can be further analyzed for insights into marginal workers' characteristics and distribution across industrial categories and age groups.

5. Conclusion

This project offers valuable insights into the demographic and industrial distribution of marginal workers in Tamil Nadu. The K-Means clustering analysis is a powerful tool to identify patterns and clusters within the data. The code serves as a foundation for further analysis and interpretation.