**Phase 5: Project Documentation & Submission**

**Project Overview:**

**Objectives:** The primary objectives of this project are to analyze the demographic characteristics of marginal workers in Tamil Nadu, specifically based on age, industrial category, and gender. The project aims to provide socioeconomic insights into this population.

**Analysis Approach:**

**1.Data Loading and Preprocessing**: The project starts with loading a dataset containing information about marginal workers in Tamil Nadu. The data is preprocessed to handle missing values, encode categorical variables, and prepare it for analysis.

**2.Demographic Analysis** : A demographic analysis is conducted to understand the distribution of marginal workers in terms of age, industrial category, and gender. This analysis includes basic statistics, visualizations, and insights into the population's characteristics.

**3.Clustering Analysis:** To identify patterns among different industrial categories and age groups, a clustering analysis (e.g., K-Means) is performed. This helps in categorizing marginal workers into groups based on similar characteristics.

**Code:**

df['Age\_Midpoint'] = df['Age group'].apply(lambda x: sum(map(int, x.split('-'))) / 2)

# Standardize the data

scaler = StandardScaler()

scaled\_data = scaler.fit\_transform(df[['Age\_Midpoint','Industrial Category - A - Cultivators - Persons', 'Industrial Category - A - Cultivators - Males', 'Industrial Category - A - Cultivators - Females']])

# Determine the optimal number of clusters (K) using the elbow method

inertia = []

for k in range(1, 11):

kmeans = KMeans(n\_clusters=k, random\_state=42)

kmeans.fit(scaled\_data)

inertia.append(kmeans.inertia\_)

# Plot the elbow curve

plt.figure(figsize=(8, 6))

plt.plot(range(1, 11), inertia, marker='o')

plt.title('Elbow Method for Optimal K')

plt.xlabel('Number of Clusters (K)')

plt.ylabel('Inertia')

plt.show()

# Based on the elbow method, choose an appropriate value for K

k = 3

# Apply K-Means clustering

kmeans = KMeans(n\_clusters=k, random\_state=42)

df['cluster'] = kmeans.fit\_predict(scaled\_data)

# Visualize clusters

plt.figure(figsize=(12, 8))

sns.scatterplot(x='Age\_Midpoint', y='cluster', hue='Industrial Category - A - Cultivators - Persons', data=df, palette='viridis')

plt.title('Clustering of Marginal Workers by Age group and Industrial Category')

plt.xlabel('Age group')

plt.ylabel('Cluster')

plt.show()

**Visualization Types:**

**1.Demographic Analysis Visualizations:** The following visualizations are used to understand the demographic characteristics:

- Histograms to show the distribution of age.

- Bar plots to display the distribution of marginal workers in different industrial categories.

- Countplots to visualize the gender distribution.

- Summary statistics to describe the age distribution.

**2.Clustering Analysis Visualizations:** For clustering analysis:

- Scatter plots to visualize clusters based on age and industrial category.

- Elbow method plot to determine the optimal number of clusters (K) in K-Means.

**Code Implementation:**

I have provided code snippets for data preprocessing, demographic analysis, and clustering analysis in earlier responses. Here is an example output of a demographic analysis:

* Histogram of Age Distribution
* Bar Plot of Industrial Category Distribution
* Gender Countplot

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.linear\_model import LinearRegression

from sklearn.linear\_model import Lasso

from sklearn.ensemble import RandomForestRegressor

from sklearn.svm import SVR

dataset= pd.read\_csv("/content/DDW\_B06\_3300\_State\_TAMIL\_NADU-2011.csv")

dataset.info()

dataset.hist(figsize=(155,100))

These visualizations help to understand the distribution of marginal workers in terms of age, industrial category, and gender.

Insights:

- The histogram of age distribution reveals that most marginal workers are in the age range

- The bar plot shows the distribution of marginal workers across different industrial categories, which is essential for understanding their economic activities.

- The gender countplot provides insights into the gender distribution among marginal workers.

By performing clustering analysis, you can further uncover hidden patterns and group marginal workers based on shared characteristics. The insights gained from this analysis help policymakers, researchers, and organizations understand and address the specific needs and challenges of this demographic group in Tamil Nadu.

**Dataset Link:**[**https://tn.data.gov.in/catalog/marginal-workers-classified-age-industrial-category-and-sex-census-2011-india-and-states**](https://tn.data.gov.in/catalog/marginal-workers-classified-age-industrial-category-and-sex-census-2011-india-and-states)