**ASSESSMENT OF MARGINAL WORKERS IN TAMIL NADU-A SOCIOECONOMIC ANALYSIS**

**PHASE 4: DEVELOPMENT PART 4**

**FEATURE ENGINEERING, MODEL TRAINING & EVALUATION**

**OVERVIEW OF THE PROCESS:**

***DATA COLLECTION:***

* Collect relevant data on water quality. This can come from various sources, including government agencies, sensors, or even crowdsourced data.
* Ensure the data is well-documented, clean, and in a format suitable for analysis.

***DATA EXPLORATION AND PREPROCESSING:***

* Explore the dataset to understand its structure and identify any missing or erroneous data.
* Preprocess the data, which may involve handling missing values, data imputation, and outlier detection.
* Conduct feature selection and engineering to create relevant input variables for your models. This can involve transforming, scaling, or creating new features from the raw data.

***DATA VISUALISATION:***

* Create visualizations to gain insights into the data. Plots, charts, and graphs can help you identify patterns, correlations, and anomalies.

***MODEL EVALUATION:***

* Evaluate your model(s) using appropriate evaluation metrics. For regression tasks, this might include Mean Absolute Error (MAE) or Root Mean Square Error (RMSE). For classification, you may use metrics like accuracy, precision, recall, and F1 score.
* Perform cross-validation to ensure your model's robustness.

***FEATURE ENGINEERING:***

* Feature engineering involves creating new features or transforming existing ones to improve the performance of your model.
* common techniques include one-hot encoding for categorical variables, scaling or normalizing numerical features, and creating interaction terms or polynomial features.

***MONITORING AND MAINTAINANCE:***

* Continuously monitor the model's performance in real-world settings and update it as needed.

***DOCUMENTATION:***

* Maintain comprehensive documentation of your work, including data sources, preprocessing steps, model details, and results.

***DEMOGRAPHIC ANALYSIS AND VISUALISATION:***

import pandas as pd

import matplotlib.pyplot as plt

df= pd.read\_csv(‘Marginal Workers.csv’)

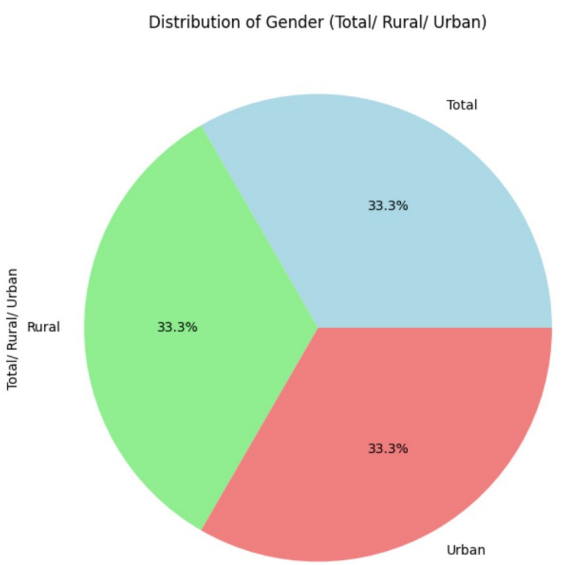
**# Demographic Analysis: Gender**

gender-counts=df[‘Total/Rural/Urban’].value\_counts()

gender\_counts.plot(kind=’pie’, autopct=’%1.11%%’,colors=[‘lightblue’, ‘lightgreen’, ‘lightcoral’],figsize=(8,8))

plt.title(‘Distribution of Gender (Total/Rural/Urban)’)

plt.show()

******

***#*Demographic Analysis: Age Groups**

age\_group\_counts = df [‘Age group’]. value\_counts()

age\_group\_counts.plot(kind=’bar’,figsize=(8,6),color=’skyblue’)

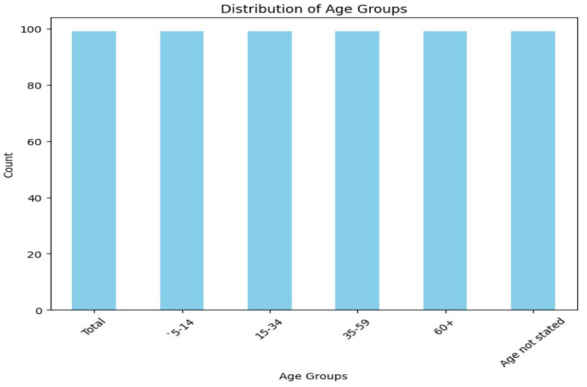
plt.title(‘Distribution of Age Groups’)

plt.xlabel(‘Age Groups’)

plt.ylabel(‘Count’)

plt.xticks(rotation=45)

plt.show()

******

**#Demographic Analysis: Total/Rural/Urban Distribution**

urban\_rural\_counts=df[‘Total/Rural/Urban’].value\_counts()

urban\_rural\_counts.plot(kind=’bar’, color=[‘lightgreen’, ‘lightblue’],figsize=(6,6))

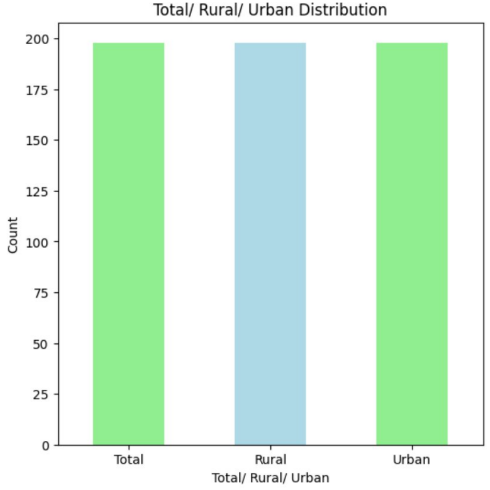
plt.title(“Total/Rural/Urban Distribution’)

plt.xlabel(“Total/Rural/Urban’)

plt.ylaabel(‘Count’)

plt.xsticks(rotation=0)

plt.show()

**

***DATA AGGREGATION***

Import pandas as pd

Df = pd.read\_csv(‘Marginal Workers.csv’)

# Group by age group , industrial categories – a - cultivators – persons ,and worked for 3 months or more but less than 6 months – persons then aggregate the counts

marginal\_workers\_distribution = df.groupby([‘Age group’, ‘industrial Category – A -Cultivators – persons’ , ‘Worked for 3 months or more but less than 6 months – Persons’])[‘Industrial Category - A – Cultivators – Males’].sum().reset\_index()

# Print the resulting DataFrame (optional)

Print(marginal\_workers\_distribution)

#To save the aggregated data to a new CSV file (optional)

#marginal\_workers\_distribution.to\_csv(‘marginal\_workers\_distribution.csv’. index=false)

Age group … Industrial Category - A – Cultivators – Males

0 15-34 … 0

1 15-34 … 3

2 15-34 … 6

3 15-34 … 4

4 15-35 … 8 .. … … …

540 ‘5-14 … 133

541 ‘5-14 … 137

542 ‘5-14 … 141

543 ‘5-14 … 664

544 ‘5-14 … 825

[545 rows x 4 columns]

***Data Visualization***

***Countplot:***

**#1**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("Marginal Workers.csv")

sns.set(style="whitegrid")

plt.figure(figsize=(10,6))

sns.countplot(data=df, x="Area Name")

plt.title("Count of Marginal Workers by Area Name")

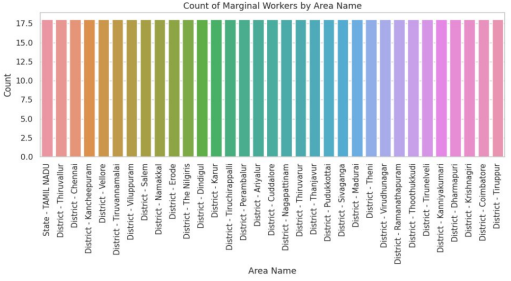
plt.xlabel("Area Name")

plt.ylabel("Count")

plt.xticks(rotation=90)

plt.tight\_layout()

plt.show()

**

**#2**

import seaborn as sns import

matplotlib.pyplot as plt

import pandas as pd

#Assuming ’Age group’ is the column representing age groups in your dataset

df = nd pd.read\_csv("Marginal Workers.csv")

sns.set(style="whitegrid")

plt.figure(figsize = (10, 6)

sns.countplot x=^ prime Aqe groupdata=dfpalette='viridis)

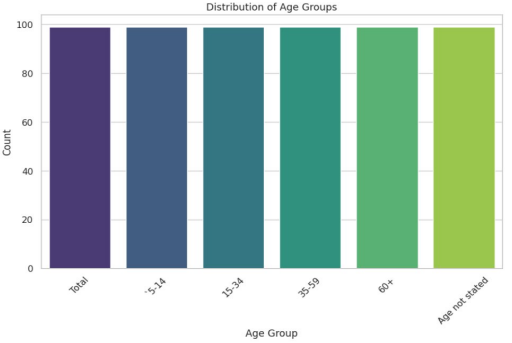
plt.title('Distribution of Age Groups')

plt.xlabel('Age Group')

plt.ylabel('Count)

plt.xticks (rotation = 45)

plt.show()

**

***Line Graph:***

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("Marginal Workers.csv")

x\_values =df['Age group']

y\_values =df['Area Name']

plt.figure(figsize (10, 6))

plt.plot(x\_values, y\_values, marker='o', linestyle='-')

plt.title("Line Graph for Marginal Workers Dataset")

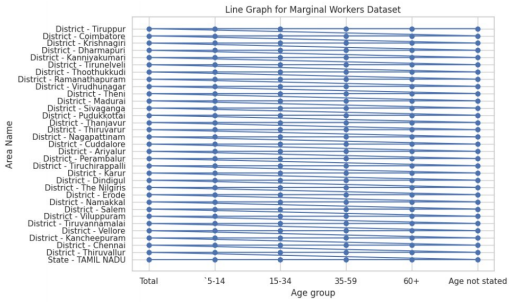
plt.xlabel("Age group")

plt.ylabel("Area Name")

plt.grid(True)

plt.tight\_layout()

plt.show()



***Histogram:***

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("Marginal Workers.csv")

data=df['Age group']

plt.figure(figsize-(10, 6))

plt.hist(data, bins=20, color='skyblue', edgecolor='black')

plt.title("Histogram of Age in Marginal Workers Dataset")

plt.xlabel("Age group")

plt.ylabel("Frequency")

plt.tight\_layout()

plt.show()



***Scatter Plot:***

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("Marginal Workers.csv")

x\_values =df['Worked for 3 months or more but less than 6 months - Persons']

y\_values =df['Worked for 3 months or more but less than 6 months - Males']

plt.figure(figsize-(10, 6))

plt.scatter(x\_values, y\_values, c='blue', alpha=0.5, edgecolors='k')

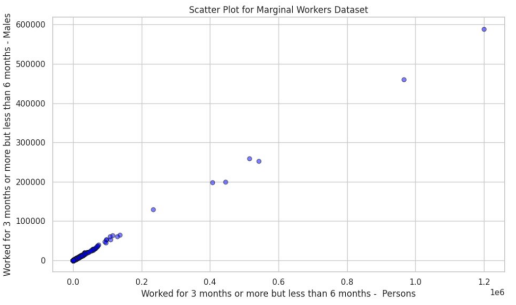
plt.title("Scatter Plot for Marginal Workers Dataset")

plt.xlabel("Worked for 3 months or more but less than 6 months - Persons")

plt.ylabel("Worked for 3 months or more but less than 6 months - Males")

pit tight\_layout()

plt.show()



***Heatmap:***

import seaborn as sns

import pandas as pd

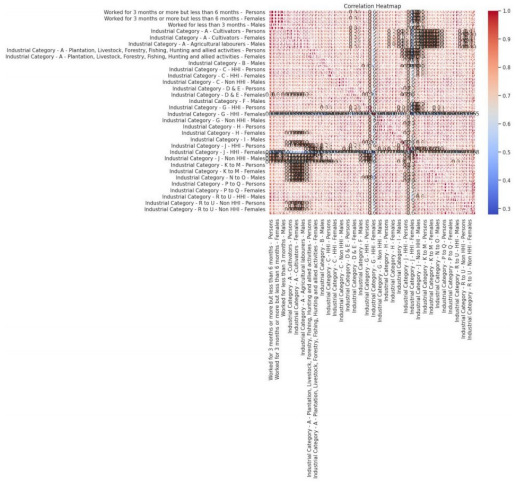
import matplotlib.pyplot as plt

df = pd.read\_csv('Marginal Workers.csv')

plt.figure(figsize=(10,8))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt=".2f") plt.title("Correlation Heatmap')

plt.show()



***Scatterplot:***

import seaborn as sns

import pandas as pd

import matplotlib.pyplot as plt

df=pd.read\_csv('Marginal Workers.csv')

sns.set(style="whitegrid")

plt.figure(figsize=(10,6))

sns scatterplot (x='Area Name', y='Total/ Rural/Urban',

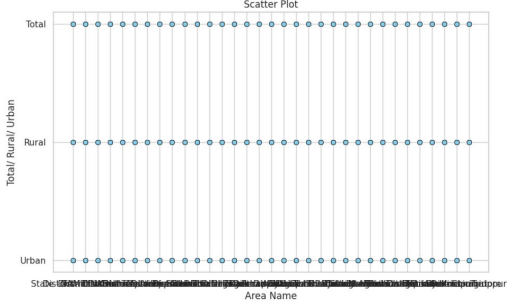
data=df, color='skyblue', marker='o', edgecolor="black")

plt.title('Scatter Plot')

plt.xlabel('Area Name')

plt.ylabel('Total/ Rural/Urban')

plt.show()



***Histplot:***

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df-pd.read\_csv("Marginal Workers.csv')

plt.figure(figsize=(8.6))

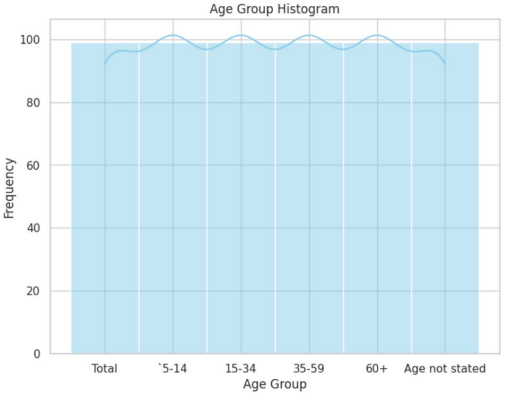
sns.histplot(df['Age group'], bins=20, kde=True, color='skyblue')

plt.title('Age Group Histogram')

plt.xlabel('Age Group')

plt.ylabel('Frequency')

plt.show()



***Pie Chart:***

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("Marginal Workers.csv")

category\_counts =df['Worked for 3 months or more but less than 6 months Persons').value\_counts()

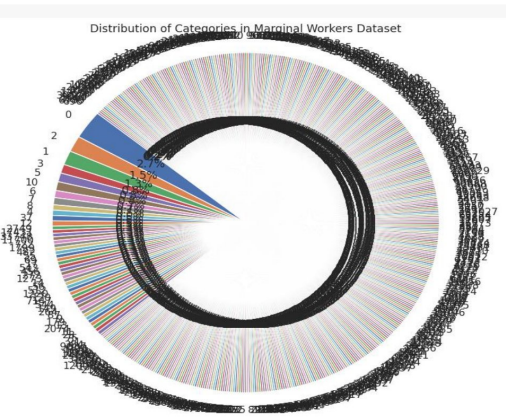
plt.figure(figsize (8, 8))

plt.pie(category\_counts, labels=category\_counts.index, autopct='%1.11%%', startangle=140)

plt.title("Distribution of Categories in Marginal Workers Dataset")

plt.axis('equal")

plt.show()



import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv('Marginal Workers.csv')

print(df.head())

urban\_counts =df['Total/ Rural/

Urban'].value\_counts()

plt.figure(figsize=(10, 6))

urban\_counts.plot(kind='bar', color='skyblue')

plt.title('Distribution of Total/ Rural/Urban')

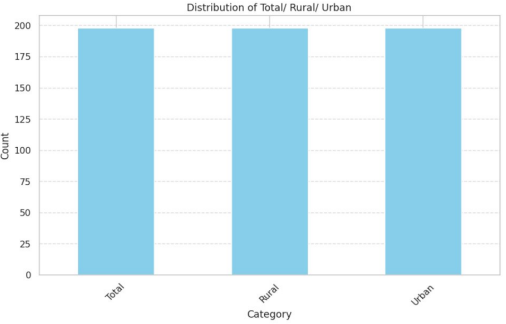
plt.xlabel('Category')

plt.ylabel('Count')

plt.xticks(rotation=45)

plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()



***CONCLUSION:***

* A reflection on the significance of your socioeconomic analysis and its potential impact.
* Express your commitment to continued efforts to contribute to the improvement of economic and social well-being in the area you've studied.