PHASE 3 ASSIGNMENT

PROJECT TITLE: PREPROCESSING THE DATASET

PROBLEM STATEMENT:

AI-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC).

PROBLEM DEFINITION:

The problem is to perform an AI-driven exploration and predictive analysis on the master details of companies registered with the Registrar of Companies (RoC). The objective is to uncover hidden patterns, gain insights into the company landscape, and forecast future registration trends. This project aims to develop predictive models using advanced Artificial Intelligence techniques to anticipate future company registrations and support informed decision-making for businesses, investors, and policymakers.

GITHUB LINK: https://github.com/20Santhiya03/RoC.git

https://github.com/20Santhiya03/innovation.git

DOCUMENT:

Building the project by preprocessing the data.

DATASET LINK: https://tn.data.gov.in/resource/company-master-data-tamil-nadu-upto-28th-february-2019
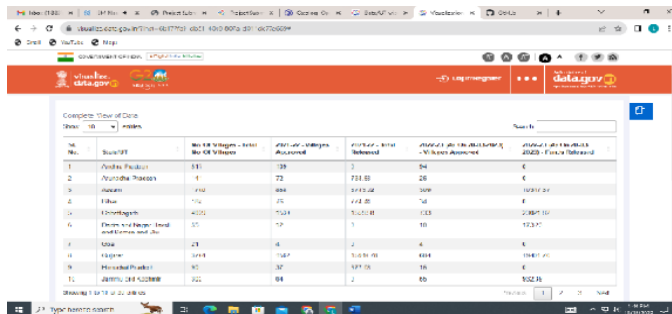
Preprocessing a dataset is a crucial step in preparing data for machine learning models. The specific steps can vary depending on the nature of your data and the problem you're trying to solve. However, here's a general set of steps you might follow:

1. **Import Libraries:**

   - Import the necessary libraries for data manipulation and analysis such as Pandas, NumPy, and others.

   ```python
   import pandas as pd

   import numpy as np
   ```
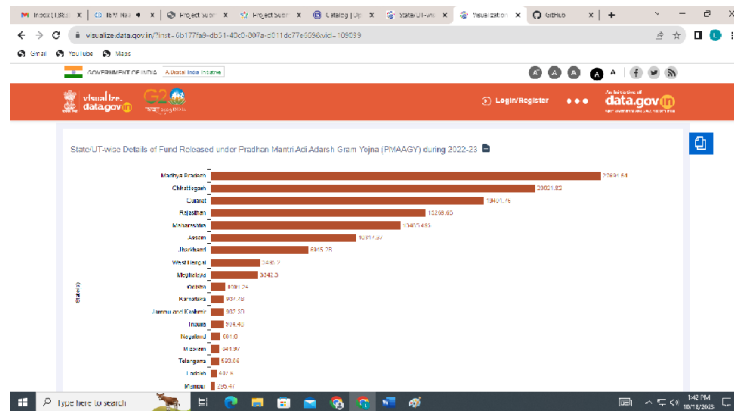

2. **Load the Dataset:**

   - Read the dataset into a Pandas DataFrame.

   ```python
   data = pd.read_csv('your_dataset.csv')
   ```

3. **Explore the Data:**

- Check for missing values, understand the structure of the data, and explore basic statistics.

```python
# Check for missing values
print(data.isnull().sum())
# Basic statistics
print(data.describe())
```

4. **Handle Missing Values:**

- Decide on a strategy for handling missing data. Options include dropping missing values, filling them with mean or median, or using more advanced imputation techniques.

```python
# Drop rows with missing values

data = data.dropna()


# Fill missing values with mean

data = data.fillna(data.mean())
```

5. **Remove Duplicates:**
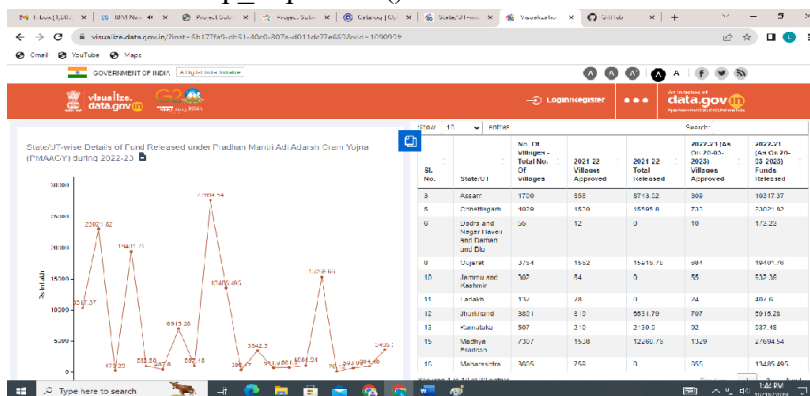
- Check for and remove duplicate rows.

```python
data = data.drop_duplicates()
```

6. **Handle Categorical Data:**

   - Convert categorical variables into numerical format, using techniques like one-hot encoding or label encoding.

   ```python
   # One-hot encoding
   data = pd.get_dummies(data, columns=['categorical_column'])
   ```
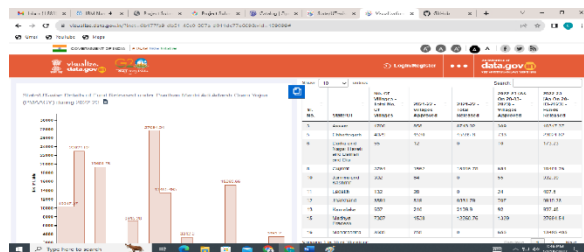

7. **Feature Scaling:**

   - Standardize or normalize numerical features to ensure they are on similar scales.

   ```python
   from sklearn.preprocessing import StandardScaler

   scaler = StandardScaler()
   data[['numerical_column']] = scaler.fit_transform(data[['numerical_column']])
   ```
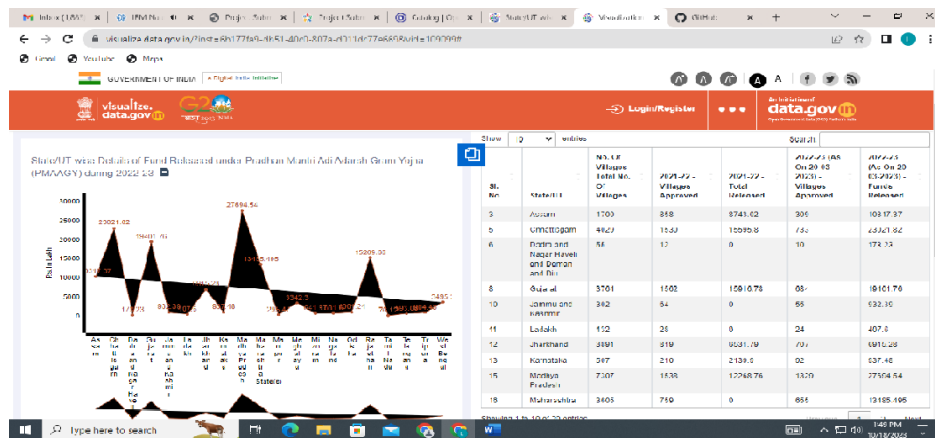
8. **Feature Engineering:**

- Create new features or transform existing ones to better represent the underlying patterns in the data.

```python
# Example: Create a new feature
data['new_feature'] = data['feature1'] * data['feature2']
```
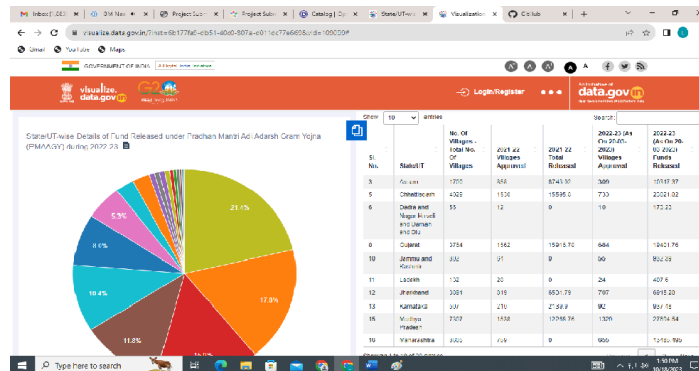
9. **Split the Dataset:**

   - Split the dataset into training and testing sets.

   ```python
   from sklearn.model_selection import train_test_split

   X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
   ```


10. **Save Preprocessed Data (Optional):**

   - Save the preprocessed data to a new file for future use.

   ```python
   data.to_csv('preprocessed_data.csv', index=False)
   ```

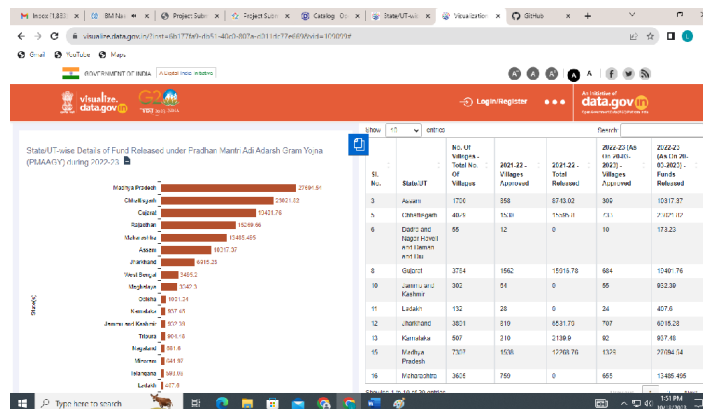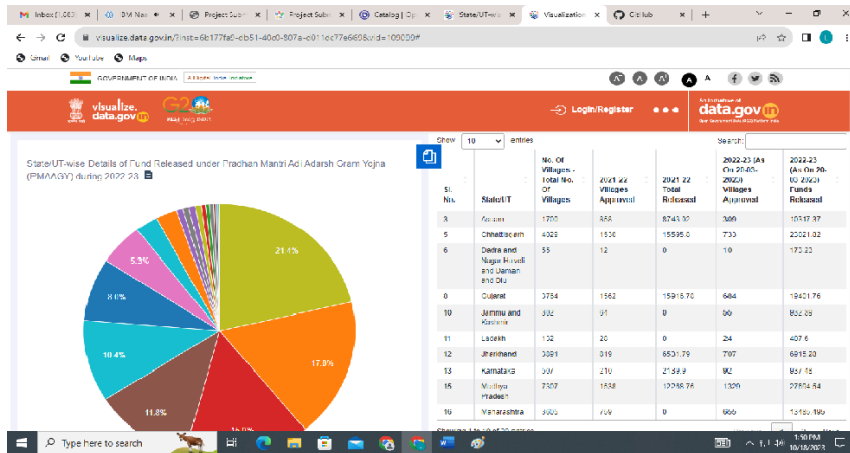State/UT-wise Details of Fund Released under Pradhan Mantri Adi Adarsh Gram Yojna (PMAAGY) during 2022-23

Sl. No.	State/UT	No. Of Villages - Total No. Of Villages	2021-22 Villages Approval	2021-22 Total Released	2022-23 (As On 20-05-2020) Villages Approval	2022-23 (As On 20-05-2020) Funds Released
3	Assam	1700	868	8749.30	389	10917.37
5	Chhattisgarh	4026	1638	15536.0	733	23021.32
6	Dadra and Nagar Haveli and Daman and Diu	55	12	0	16	173.23
9	Gujarat	3754	1562	15916.70	684	19401.76
10	Jammu and Kashmir	392	91	0	96	852.89
11	Ladakh	132	20	0	24	407.6
12	Jharkhand	5991	819	6501.79	797	6915.20
13	Karnataka	997	210	2159.9	92	957.16
15	Madhya Pradesh	7307	1638	10248.76	1329	27604.54
16	Maharashtra	3605	769	0	666	13160.196

SUBMITTED BY,

STUDENT REG NO: 711221104047

NAAN MUDHALVAN: au711221104047