**7111-AI-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC)**

**TEAM MEMBER**

**711121106048-V.MAHIBA**

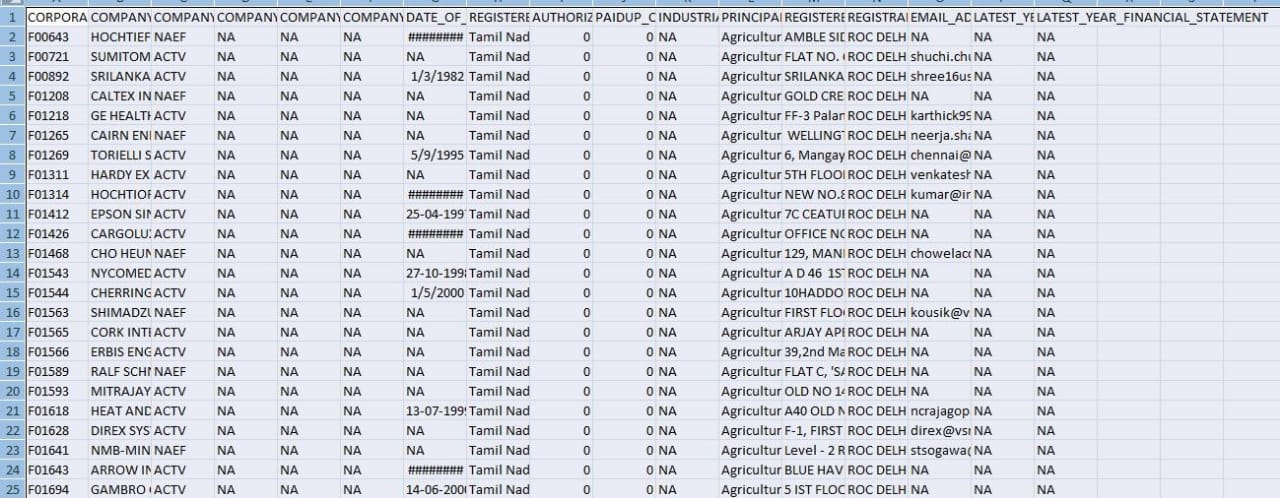
**Phase 3 :Development part 1**

**INTRODUCTION :**

**Company registration trends with RoC**

In the realm of commerce and industry, the Registrar of Companies (RoC) stands as a sentinel, diligently recording every new enterprise's inception. As businesses emerge and evolve, the data maintained by the RoC becomes a valuable testament to the ever-shifting economic landscape. In this era of unprecedented innovation and entrepreneurship, understanding and interpreting Company Registration Trends with RoC data has become indispensable. This exploration delves into the intricacies of registration patterns, regional dynamics, and industry preferences. With a focus on both historical and emerging trends, this discussion aims to provide a comprehensive view of the corporate world's heartbeat, offering insights that can steer strategic decisions, inform government policies, and fuel the engines of economic growth. Join us in this enlightening journey as we navigate the intricate tapestry of Company Registration Trends with the Registrar of Companies.In this phase we have to preprocessing and analyse the given dataset.

**GIVEN DATASET :**

****

150872 Rows × 12 Columns

NECESSARY STEPS TO FOLLOW :

1.import libraries

2.Load the Dataset

3.Exploratory Data Analysis(EDA)

DATA PREPROCESSING

Data preprocessing is a critical step in any AI-driven exploration and prediction task, including the analysis of company registration trends. Properly processed data helps ensure the accuracy and effectiveness of your predictive models. Here's a step-by-step guide on how to preprocess data for this specific task:

1. Data Collection:

- Gather relevant data sources, which might include historical company registration data, economic indicators, industry-specific information, and any other data that can potentially impact company registration trends.

2. Data Cleaning:

- Remove duplicates: Check for and eliminate duplicate records.

- Handle missing data: Identify and deal with missing values using techniques like imputation or removal.

- Correct inaccuracies: Address any inconsistencies or errors in the data.

3. Data Integration:

- If you have multiple data sources, integrate them into a single dataset. Ensure that common data fields are appropriately matched and merged.

4. Data Transformation:

- Convert data types: Ensure that data types are appropriate for analysis. Numeric data should be numerical, dates should be in date format, and categorical data should be properly labeled.

- Feature engineering: Create new features if they can provide valuable insights. For example, you might calculate monthly, quarterly, or yearly trends from daily data.

- Normalization or scaling: Scale numerical features if needed to bring them to the same range.

5. Data Reduction:

- Reduce dimensionality: If you have a high number of features, consider using techniques like Principal Component Analysis (PCA) or feature selection to reduce the dimensionality.

- Outlier detection: Identify and handle outliers using techniques like Z-score or IQR.

6. Data Encoding:

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

7. Data Splitting:

- Split your data into training, validation, and test sets to evaluate model performance. Typically, this is done in an 80-20 or 70-30 ratio.

8. Time Series Considerations:

- If you are dealing with time series data, pay special attention to time-based features, lag variables, and seasonality adjustments.

9. Data Exploration:

- Visualize the data to gain insights into trends, patterns, and correlations. Tools like histograms, scatter plots, and correlation matrices can be helpful.

10. Feature Scaling:

- Normalize or standardize your features, especially if you are using models like SVM or k-means clustering.

11. Data Imbalance:

- If there is an imbalance in your target variable (e.g., more non-registrations than registrations), consider techniques like oversampling, undersampling, or using different evaluation metrics.

12. Validation:

- Validate that your preprocessed data is suitable for your machine learning algorithms. This may involve testing your model on a small subset of data before full-scale training.

13. Documentation:

- Maintain thorough documentation of all your preprocessing steps. This documentation can be helpful in reproducing results and debugging issues.

14. Automate the Pipeline:

- Consider creating an automated data preprocessing pipeline, which can save time and ensure consistency in data preparation for ongoing analysis.

15. Model Building:

- Finally, use the preprocessed data to build and train your AI models for predicting company registration trends.

Continuous monitoring and adjustment of your data preprocessing pipeline are crucial as new data becomes available and as you update your models. This iterative approach ensures that your predictions remain accurate and relevant.

**IMPORTING LIBRARIES:**

Importing necessary libraries to do basic things on the Dataset

**Program:**

import pandas as pd

import numpy as np

import seaborne as sns

import matplotlib.pyplot as plt

print(‘Successfully imported’)

**Output :**

Successfully imported

**LOAD THE DATASET:**

Load your Dataset into a pandas dataframe .You can typically find company registration dataset in CSV format , but you can adapt this code to other formats as needed

**Program:**

data=pd.read\_csv(‘datagovtamilnadu.csv’,encoding=unicode\_escape’)

data.head()

**EXPLORATORY DATA ANALYSIS (EDA):**

Perforn EDA to understand your data better . This includes checking for missing values , exploring the data’s statistics and visualization it to identify patterns.

**Program:**

**#**check the dataset for missing data

if data.isnull().sum().sum()==0:

print(‘There is no missing data in our dataset’)

else:

print(‘There is {} missing data in ourdataset’.format(data.isnull().sum().sum()))

**Output:**

There is no missing data in our dataset

**Program :**

frame=pd.concat([data.isnull().sum() , data.nunique() , data.dtypes] , axis=1, sort=False)

frame

**VISUALIZATION AND PRE-PROCESSING OF DATA:**

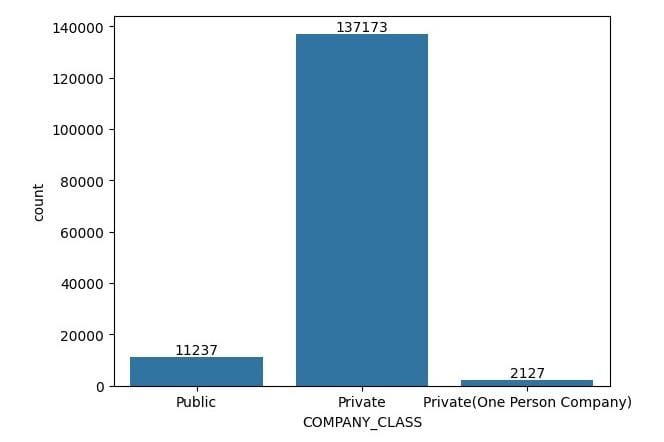
In[1]

ax = sns.countplot(data = data , X = ‘COMPANY\_CLASS’);

#Add the value of each parameters on the plot

ax.bar\_label(ax.containers[0]);

Out[1]

In[2]

Data[‘COMPANY\_NAME’].unique()

Out[2]

Array([‘HOCHTIEFF AG, ‘,

‘SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LIMITED) ‘,

‘SRILANKAN AIRLINES LIMITED ‘,.........,

‘POLYGAR FARM SUTIONS PRIVATE LIMITED ‘,

‘PANDIYA AGRI SOLUTIONS PRIVATE LIMITED ‘,

‘NROOT TECHNOLOGIES PRIVATE LIMITED ‘],dtype=object)

In[3]

From sklearn.impute import simple Imputer

import numpy as np

columns\_to\_drop=[‘CORPORATE\_IDENTIFICATION\_NUMBER’ , ‘COMPANY\_NAME’ , ‘COMPANY\_CATEGORY’]

for column in columns\_to\_drop:

if column in data.columns:

data=data.drop(columns=[column])

#fill in missing valuesin vork\_interfere column

data[‘COMPANY\_STATUS’] = np.ravel(SimpleImputer(strategy = ‘most\_frequent’).fit\_transform(data[‘COMPANY\_STATUS’].values.reshape(-1,1)))

data[‘COMPANY\_SUB\_CATEGORY’]=np.ravel(SimpleImputer(strategy = ‘most\_frequent’).fit\_transform(data[‘COMPANY\_SUB\_CATEGORY’].values.reshape(-1,1)))

data.head()

In[4]

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

RSPMPM10\_range\_plot=sns.countplot ( data=data ,x=’COMPANY\_SUB\_CATEGORY’);

RSPMPM10\_range\_plot.bar\_label(RSPMPM10\_range\_plot.containers[0]);

Plot.xticks(rotation=90);

