**AI-Driven Exploration and Prediction of Company**

**Registration Trends with the Registrar of Companies (ROC)**

**NAME:KARTHIKA A**

**REGISTER NUMBER: 61772131018**

# PHASE 2:

# Import the dataset and perform data cleaning &data analysis:

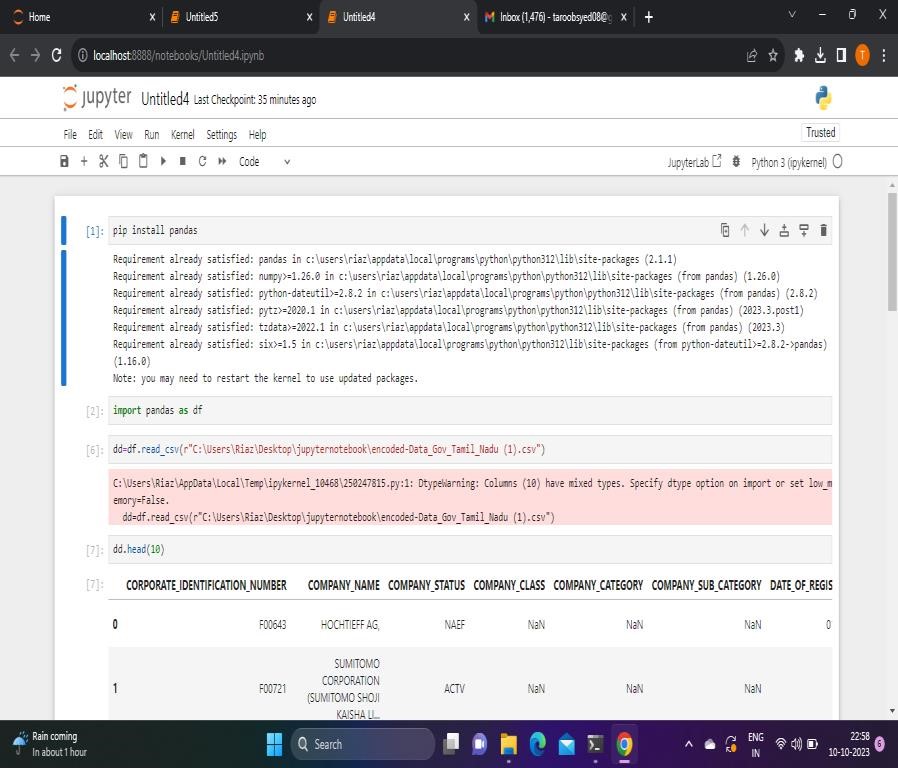
* The given dataset is imported into jupyter notebook. The required modules are imported to perform the cleaning operation.
* The necessary libraries are imported by the following commands:

pip install pandas import pandas as df

* The missing data and the duplicate data are handled by the following commands:

dd.drop\_duplicates() dd.dropna(inplace=True)

* Clean the pre-processed data which includes the renaming of columns.
* The cleaned dataset file is saved.



**DATA CLEANING:**

Data cleaning means fixing bad data in the dataset. The bad data could be empty cells, data in wrong format and wrong data.

Data cleaning is a critical step in preparing data for the prediction of company registration using AI-driven exploration. Clean and well-structured data is essential for building accurate and reliable predictive models. Here are the steps to clean the data:

# Handling Missing Data:

* Missing data in the dataset is identified and handled. Missing data can significantly impact the quality of predictions.
* Options for handling missing data include:
* Rows with missing values are handled.
* Imputing missing values with the mean, median, or mode of the respective column.

# Handling Duplicates:

* Duplicate records are checked and removed, as duplicate entries can skew the analysis and modeling results.
* drop\_duplicates method in pandas is used to remove duplicate rows.

# Data Transformation:

* Categorical variables are converted into numerical format using one-hot encoding or label encoding. This is necessary for many machine learning algorithms.
* Normalize or scale numerical features to ensure that they are on a common scale, especially if you plan to use algorithms sensitive to feature scaling.

# Data Splitting:

 Dataset is split into training and testing sets for model evaluation. Typically, the training set is used to train the model, and the testing set is used to evaluate its performance.

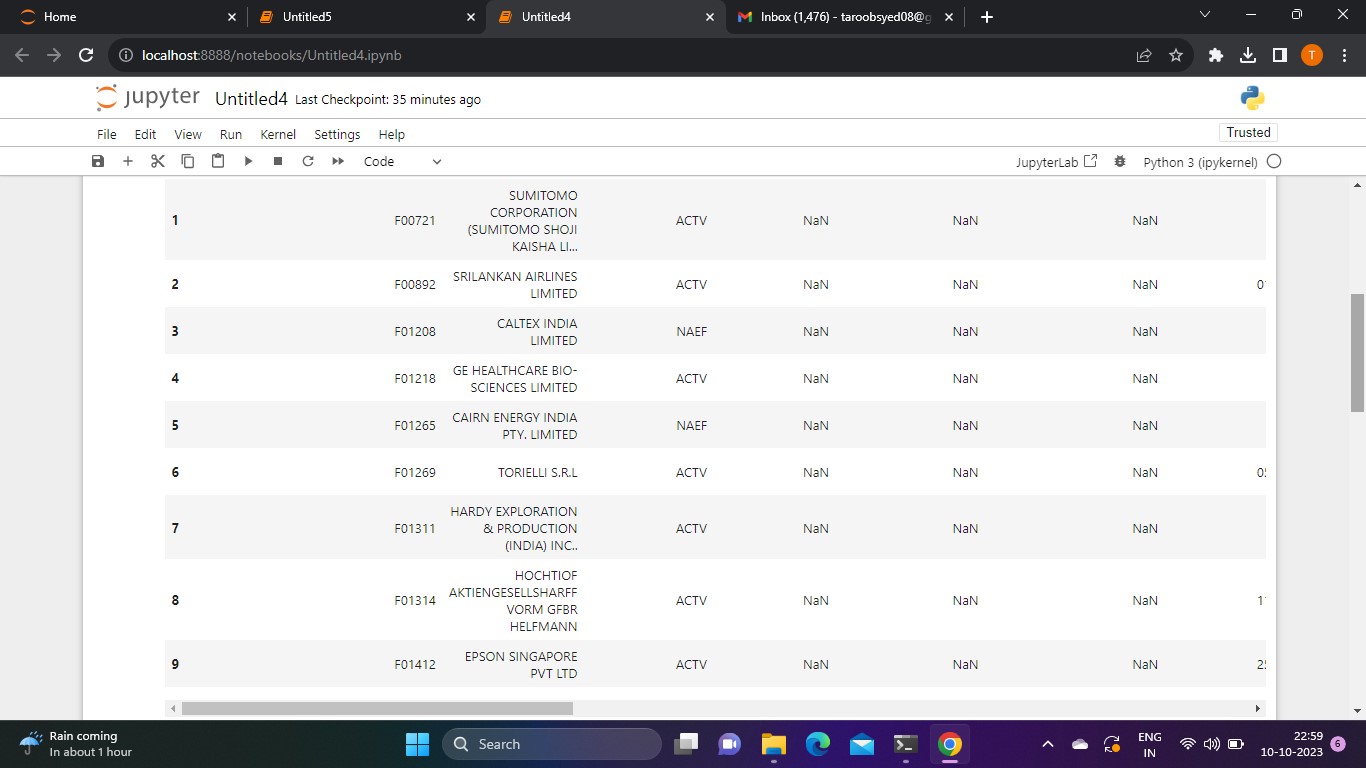
# Data Quality Checks:

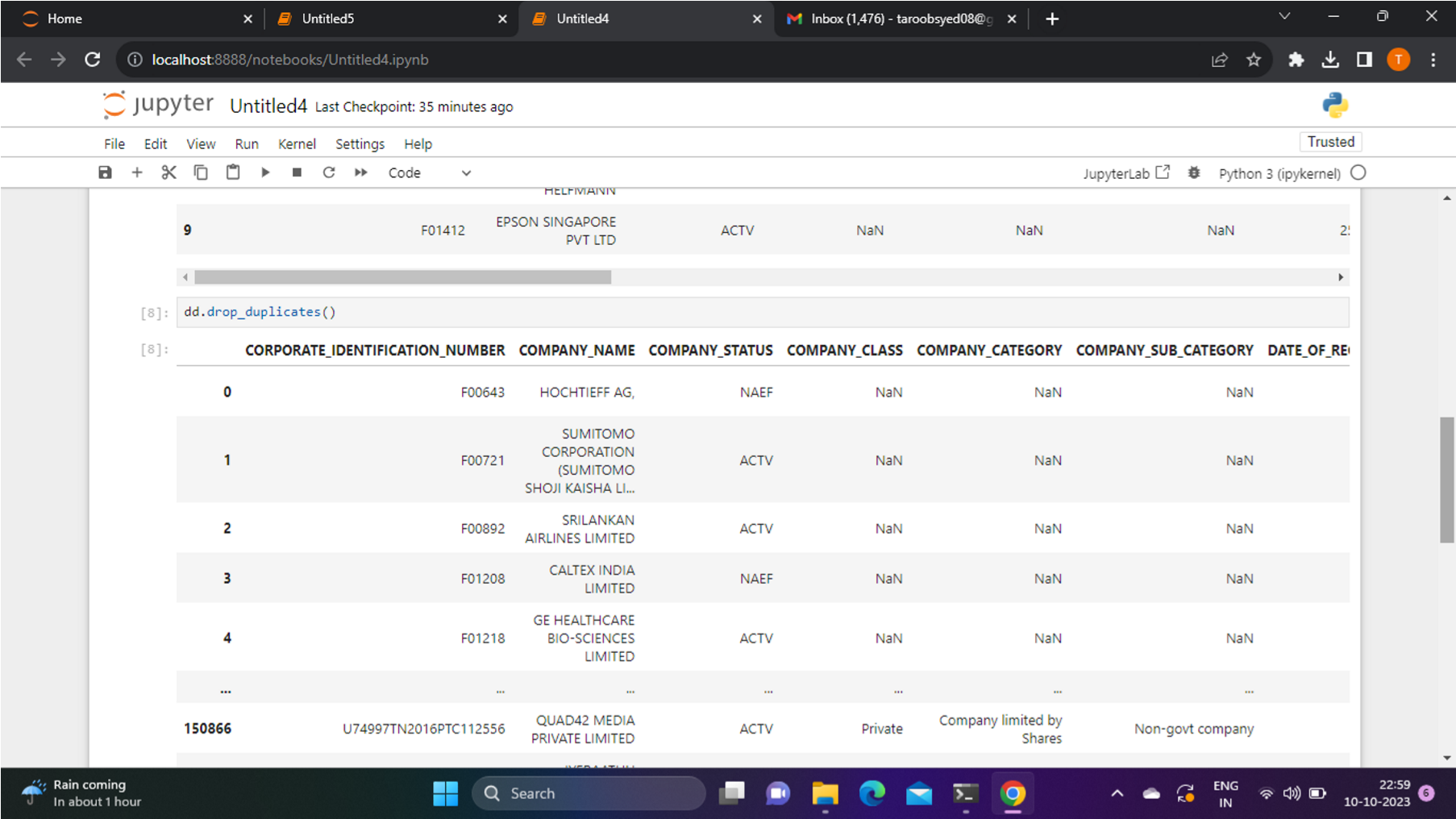
 Any anomalies or data quality issues is checked. This includes verifying that data is within expected ranges and adheres to domain-specific rules.

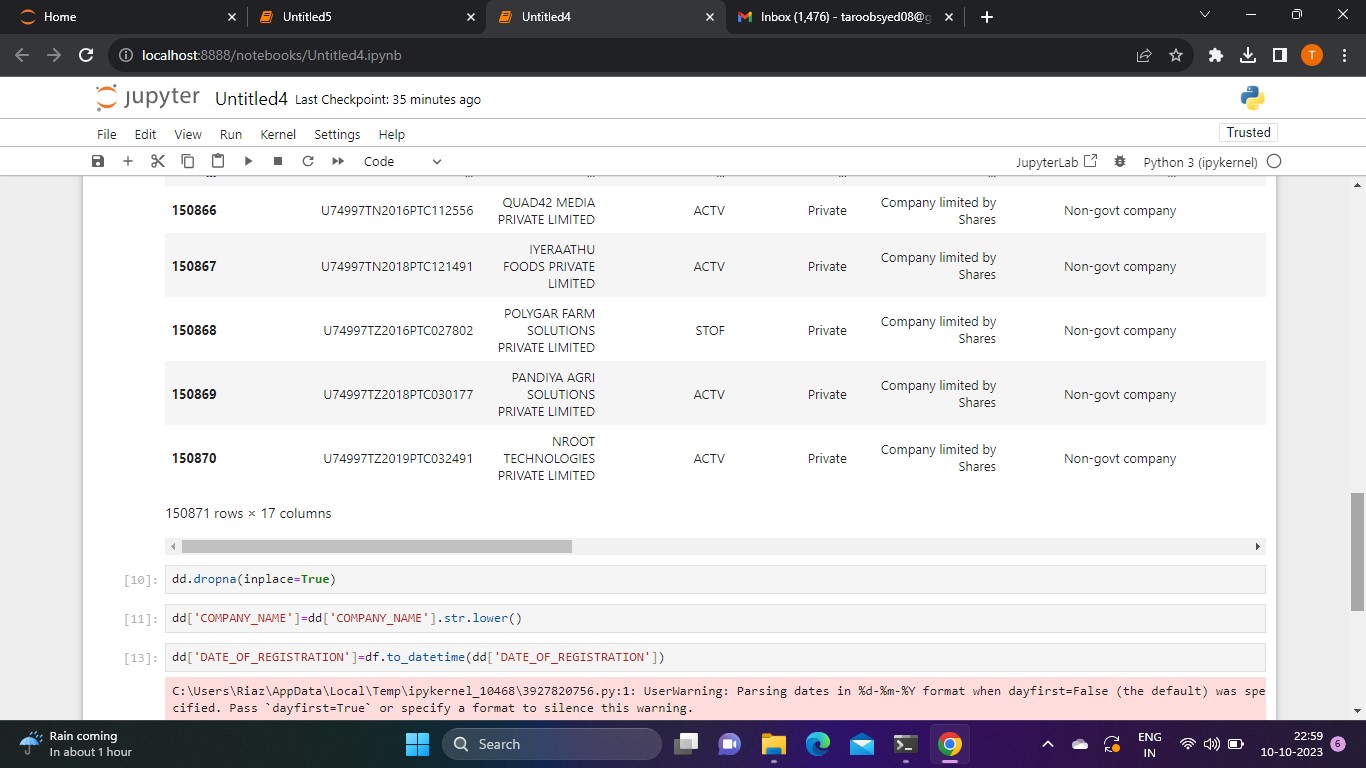
# Data Visualization:

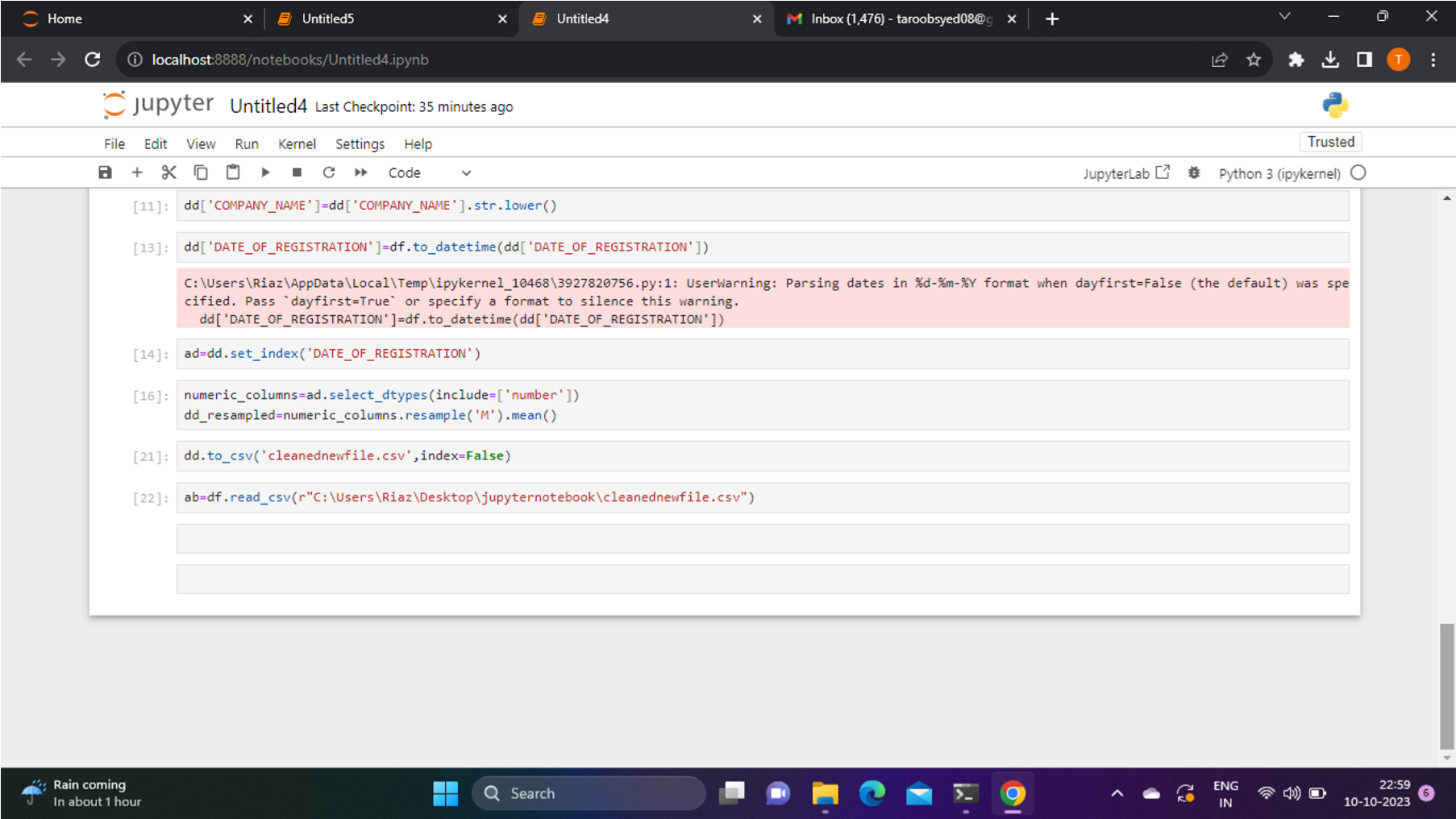
 Data is visualized to gain insights into its distribution and relationships between variables. Data visualization can help identify patterns and trends.

PYTHON CODE:



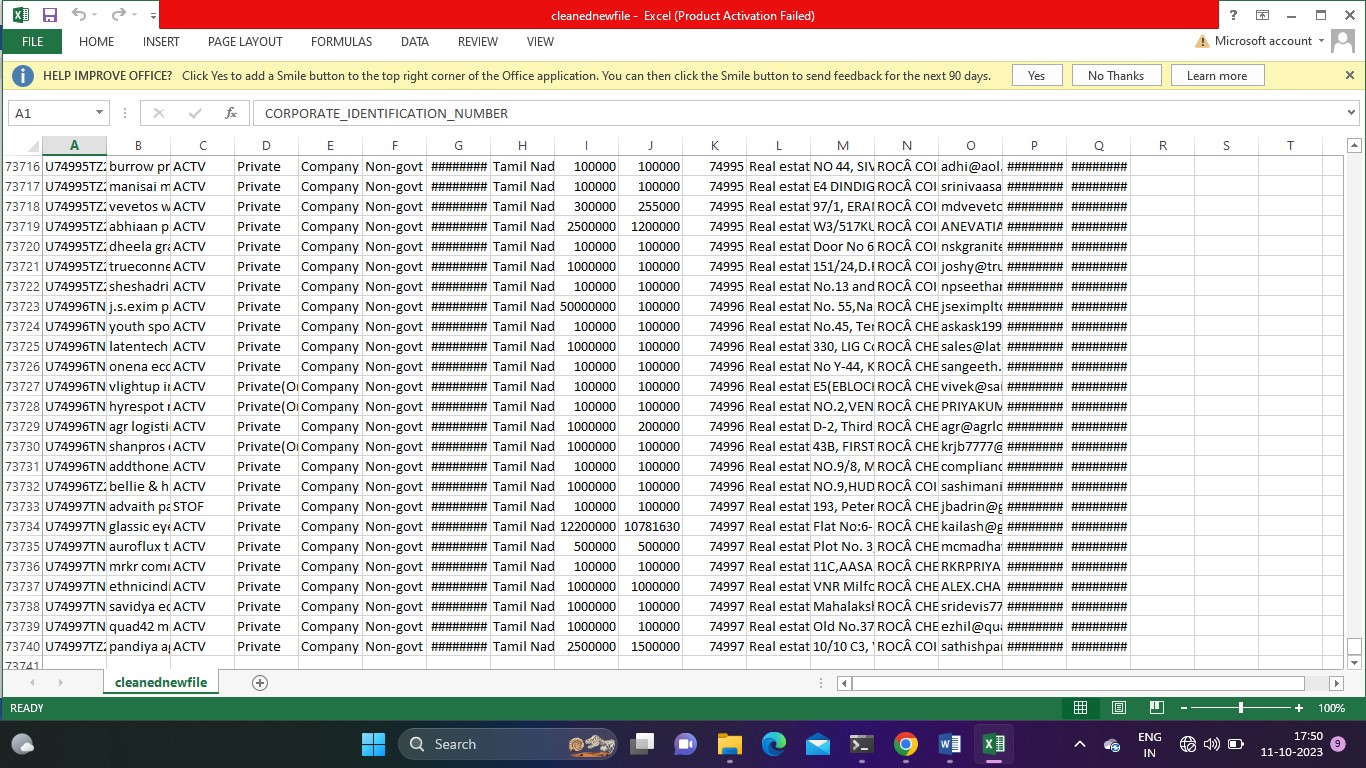






# OUTPUT AFTER CLEANING:

After the data cleaning is done, the number of cells in the Excel sheet in the final dataset is reduced to 73741 by removing the duplicate data, wrong data.



# DATA ANALYSIS:

Data analysis is a systematic approach which follows the process of inspecting, cleaning, transforming and interpreting data to extract valuable insights.

Data analysis is done using Python- jupyter notebook.

It includes various various methods:

* Data Collection and Pre-processing.
* Conduct EDA to gain insights into the dataset.
* Feature and Model Selection.
* Data Splitting
* Model training and evaluation.
* Deployment and Monitoring.

