Phase 5

PROJECT DOCUMENTATION & SUBMISSION

Date	**31-10-2023**
Team ID	**953**
Project Name	**Al-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC)**

Project Title: Al-Driven Exploration and Prediction of Company Registration Trends with Registrar of Companies (RoC)

Problem Statement

Objective: The objective is to utilize AI and data analytics to explore historical company registartion trends, develop predictive models, and facilitate informed decision-making for regulatory authorities and businesses using Registrar of Companies (RoC) data.

Problem identified:

AI-driven analysis of Registrar of Companies (RoC) data faces challenges in data quality, potentially inaccurate predictions due to biases, and ensuring compliance with privacy regulations. Identifying relevant features impacting registration trends is complex, while dynamic business environments require adaptable models. Model interpretability is crucial, yet complex AI models might lack transparency. Scalable infrastructure and expertise are needed to handle vast, complex datasets for meaningful insights and accurate trend predictions.

Introduction:

The integration of Artificial Intelligence (AI) in scrutinizing and forecasting company registration trends via Registrar of Companies (RoC) data stands at the forefront of modern data analytics. This innovation involves the application of advanced AI algorithms to navigate and interpret the complexities inherent in the RoC database, aiming to unearth valuable insights pivotal for comprehending the ever-evolving landscape of business registrations.

Nonetheless, this ambitious initiative encounters several challenges. Data quality issues, such as inconsistencies and inaccuracies within RoC records, pose a significant barrier to accurate predictions. Moreover, ensuring the compliance of AI methodologies with stringent regulatory frameworks, particularly in handling sensitive company information, remains a crucial concern.

The essence of this endeavor lies in its capacity to discern and analyze vast volumes of data. Extracting meaningful information from this wealth of data is fundamental in illuminating the intricate trends and patterns within company registration, fostering a deeper understanding of economic fluctuations and regulatory impacts. However, this journey is also marked by the complexity of identifying and incorporating the most influential variables into predictive models. Ensuring these models are interpretable is another critical facet. While AI algorithms can offer predictions, their opaque nature might challenge stakeholders' ability to understand the rationale behind these forecasts, necessitating transparent and interpretable models.

Furthermore, the dynamic nature of the business environment poses a perpetual challenge. Businesses are subject to an array of ever-changing factors—economic, social, and regulatory—that can swiftly influence registration trends. Adapting AI models to these dynamic shifts is crucial for relevance and accuracy. The potential for AI-driven analysis to uncover trends and anomalies within RoC data is immense, holding the promise of informing policymakers, investors, and businesses in making strategic, well-informed decisions. The amalgamation of AI's prowess with comprehensive domain expertise is pivotal in overcoming these challenges and realizing the transformative potential of predicting and understanding company registration trends with the Registrar of Companies.

Data: The project relies on Registrar of Companies (RoC) datasets, including historical company registration records, financial data, and relevant economic indicators.

LITERATURE SURVEY

1. "Implementation of Data Mining on a Secure Cloud Computing via Web API Using Supervised Machine Learning Algorithm", Tosin Ige [2022]

This research paper focuses on the implementation of data mining in a secure cloud computing environment through a combination of decision tree and Random Forest algorithms, accessible via a Restful Application Programming Interface (API). It addresses the challenge of efficiently and securely mining large volumes of data available through cloud computing for pattern detection. The study bypasses direct interaction with data warehouses to ensure security and scalability, using a combination of IBM Cloud storage, a web service, an API, and a decision tree/Random Forest algorithm. The achieved model exhibits a high accuracy rate of 94%.

2. “Machine Learning and IBM Cloud for Critical Patient Care: A Promising Solution”, Asif Ahmed Neloy [2019]

This research project, conducted by a team from North South University in Dhaka, Bangladesh, aims to revolutionize critical patient care in healthcare facilities. The team proposes the development of a comprehensive system that harnesses the power of machine learning (ML) and the IBM Cloud platform. The key objective is to enable real-time monitoring and prediction of critical patients' health conditions, allowing doctors and nurses to provide timely care. The project involves the use of various ML algorithms, ensemble methods, and the creation of a mobile application named "Critical Patient Management System - CPMS" for seamless data access. By integrating advanced technology, this project seeks to address the critical patient care challenges prevalent in developing countries like Bangladesh.

3. “The Development and Deployment of Machine Learning Models”, James A. Pruneski [2022]

The application of artificial intelligence, particularly machine learning, is gaining traction in Orthopaedic Surgery and the field of medicine at large. This growing interest is shared by data scientists and physicians, although there is often a gap in understanding the developmental process and potential applications of machine learning. Given the anticipated impact of new technology on clinical practice in the coming years, it is crucial for physicians to grasp the workings of these processes. This paper aims to provide clarity and a general framework for building and evaluating machine learning models.

4. "Explainable Machine Learning in Deployment" ,Shubham Sharma [2020]

This study, conducted by Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley, delves into the practical deployment of explainable machine learning (ML). While explainability is crucial for building trust in ML models, the research aims to understand how organizations use explainability techniques in practice. The study reveals that most deployments primarily serve machine learning engineers for model debugging, rather than end users affected by the models, creating a gap between explainability in practice and transparency goals. It highlights the limitations of current explainability techniques and proposes a framework for setting clear goals for explainability to facilitate end-user interaction.

5. "Accelerating the Machine Learning Model Deployment using MLOps", Mandepudi Nobel Chowdary[2022]

This research, led by Mandepudi Nobel Chowdary, Bussa Sankeerth, Chennupati Kumar Chowdary, and Manu Gupta, delves into the realm of Machine Learning Operations (MLOps) to expedite the deployment of machine learning models. Deploying machine learning models can be a complex task, involving multiple factors such as continuous builds for efficiency and different libraries for predictions. As datasets grow to enhance predictive accuracy, certain parameters must be dynamically adjusted for performance tuning. Implementing these changes manually can be time-consuming and labor-intensive for developers. This study introduces an end-to-end automation cycle designed to streamline the deployment of machine learning models with enhanced performance.

DESIGN THINKING

1. Empathize

- Understand Users: Gain an empathic understanding of the problem by engaging with and observing the people you're designing for.
- Empathy Tools: Conduct interviews, observations, and surveys to delve into the users' experiences, needs, and challenges.

2. Define

- Frame the Problem: Define the problem based on the insights gathered during the empathize stage.
- User-Centric Problem Statement: Reframe the issue in a human-centric manner to guide the design process.
- 3. Ideate
- Generate Ideas: Encourage the generation of a wide range of ideas. Use brainstorming or other creative techniques to foster innovative thinking.
- Divergent Thinking: Create a space where all ideas are welcome without judgment.

4. Prototype

- Create Solutions: Build scaled-down versions or prototypes of the best ideas generated during the ideation phase.
- Iterative Prototyping: Develop multiple prototypes to test different aspects and functionalities of the solution.

5. Test

- Gather Feedback: Test the prototypes with users and gather feedback.
- Iterate Based on Feedback: Refine the solutions based on the feedback received, which often involves going back to the previous stages.

6. Implement

- Execute the Solution: Once the final solution is refined and validated, implement it.
- Deploy the Solution: Launch the solution and observe its real-world impact.
- Actions:

Design Thinking Approach

1.Functionality:

- This project aims to explore historical company registration data, build
- predictive models, and deliver actionable insights, supporting informed
- decision-making for regulatory authorities and businesses based on Registrar of Companies (RoC) data.

2.UserInterface:

- The user interface will provide an intuitive platform for stakeholders to interact
- with and visualize the AI-driven insights and predictions derived from Registrar of Companies (RoC) data.

3.Natural language processing:

- Integrating NLP techniques enables the system to process unstructured text
- data, enhancing the comprehensiveness of insights from Registrar of Companies (RoC) documents and reports.

4.Responses:

- The system will provide automated responses in natural language to user
- queries, making it user-friendly and accessible for stakeholders seeking insights from Registrar of Companies (RoC) data.

5.Integration:

The system will seamlessly integrate with RoC databases and external data sources to access and analyze comprehensive data for trend exploration and prediction.

6.Testing and Improvement:

Continuous testing and refinement will be carried out to enhance the accuracy and reliability of predictions and insights generated from Registrar of Companies (RoC) data.

phases of development

1. Initiation or Conception:

- Idea Generation: The initial concept or idea is conceived.
- Feasibility Analysis: Evaluate the idea's feasibility, considering resources, market demand, or other relevant factors.

2. Planning:

- Setting Objectives: Define goals, outcomes, and success criteria.
- Creating a Roadmap: Develop a plan outlining tasks, timelines, resource allocation, and responsibilities**.**

3. Design or Conceptualization:

- Blueprint Creation: Create a detailed design or plan based on the requirements and objectives set during the planning phase.
- Prototyping: Develop initial models or prototypes for testing and feedback**.**

4. Development or Execution:

- Building or Producing: Implement the plan or design created during the previous phases.
- Continuous Improvement: Iteratively refine the work based on feedback and new insights.

5. Testing or Evaluation:

- Quality Assurance: Test the product, service, or development against pre-defined criteria to ensure it meets the required standards.
- Feedback Collection: Gather feedback from users or stakeholders to identify areas for improvement.

6. Deployment or Implementation:

- Launch or Rollout: Introduce the finalized product, service, or development to its intended audience or market.
- Installation or Implementation: Put the solution into action or use.

7. Monitoring and Maintenance:

- Continuous Assessment: Monitor the performance, collect data, and make adjustments as necessary.
- Regular Maintenance: Perform updates, fixes, or improvements to ensure sustainability and relevance.

8. Reflection and Iteration:

- Review and Learn: Reflect on the entire development process to identify successes, challenges, and areas for improvement.
- Iterate and Improve: Use the insights gained to refine and optimize future iterations or developments.

Dataset Used

1. CORPORATE_IDENTIFICATION_NUMBER

A unique identifier assigned to each registered company.

2. COMPANY_NAME

The name of the company as registered.

3. COMPANY_STATUS

Status indicating if the company is active, inactive, or any other defined status.

4. COMPANY_CLASS

Categorization of the company based on its structure or type (e.g., private, public, etc.).

5. COMPANY_CATEGORY

The classification of the company according to certain predefined categories.

6. COMPANY_SUB_CATEGORY

Further sub-categorization of the company within the broader category.

7. DATE_OF_REGISTRATION

The date when the company was officially registered.

8. REGISTERED_STATE

The state or region where the company is registered.

9. AUTHORIZED_CAP

The maximum amount of capital that a company is authorized to have through shares.

10. PAIDUP_CAPITAL

The amount of authorized capital that has been paid by shareholders.

11. INDUSTRIAL_CLASS

Classification based on the industry in which the company operates.

12. PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN

Description of the primary business activity undertaken by the company based on the Corporate Identification Number (CIN).

13. REGISTERED_OFFICE_ADDRESS

Address of the company's registered office.

14. REGISTRAR_OF_COMPANIES

The authority responsible for registering companies within a specific jurisdiction.

15. EMAIL_ADDR

Contact email address associated with the company.

16. LATEST_YEAR_ANNUAL_RETURN

The most recent year for which the annual return of the company is available.

17. LATEST_YEAR_FINANCIAL_STATEMENT

The most recent year for which the financial statement of the company is available.

Data Preprocessing Steps

1. Handling Missing Values:

Identification: Check for missing or null values in each column.
Strategies: Decide whether to remove rows with missing values, fill them with mean/median/mode, or use more complex imputation techniques.

2. Data Cleaning:

Outlier Detection: Identify and handle outliers that might skew the analysis.
Inconsistencies: Check for inconsistencies or errors in data entry (e.g., typographical errors in company names, inconsistent date formats) and correct them.

3. Feature Encoding and Transformation:

Categorical Data: Convert categorical variables (like COMPANY_CLASS, COMPANY_CATEGORY) into numerical format for machine learning algorithms. This can be done through one-hot encoding or label encoding.
Date Formatting: Ensure consistency in the date formats for uniform analysis.

4. Feature Scaling and Normalization:

Scaling Numerical Data: Normalize numerical columns like AUTHORIZED_CAP, PAIDUP_CAPITAL to a similar scale to prevent bias in models that are sensitive to the magnitude of values.

5. Text Processing (if needed):

Feature Extraction: Extract relevant information from text-based columns like PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN or REGISTERED_OFFICE_ADDRESS.
Text Cleaning: Remove unnecessary characters, lowercase text, handle special characters, and possibly perform stemming or lemmatization for natural language data.

6. Handling CIN (Corporate Identification Number):

Feature Engineering: Extract meaningful information from the CIN if it contains specific codes or identifiers that could be informative (e.g., region codes, industry codes).

7. Data Splitting:

Training and Testing Sets: Divide the dataset into training and testing sets for model evaluation.
Stratification: If needed, ensure representative distribution of classes across training and testing datasets.

8. Addressing Imbalanced Data (if applicable):

Resampling Techniques: Apply techniques if the dataset has imbalanced classes (e.g., oversampling minority class, undersampling majority class, SMOTE).

9. Data Integration and Transformation:

Merging Datasets: If additional data is available, merge it with the existing dataset to enhance analysis.

Transformation: Create new features or transform existing ones if needed for better model performance.

10. Checking Data Consistency:

Validate Data: Ensure the processed data remains consistent, adhering to the defined business rules and integrity.

AI Algorithms Applied

1. Logistic Regression:

Often used for binary classification tasks, such as predicting whether a company will register or not within a specific time frame.

2. Decision Trees and Random Forest:

Decision trees can be used to predict registration trends based on different parameters. Random Forest, an ensemble of decision trees, can enhance accuracy and handle large datasets.

3. Support Vector Machines (SVM):

SVM can predict registration trends by finding a hyperplane that best separates different classes in the dataset.

4. Gradient Boosting Models:

Algorithms like XGBoost, LightGBM, or CatBoost can be utilized to boost accuracy by sequentially adding models that correct the errors of the previous ones.

5. Neural Networks:

Artificial neural networks, including feedforward networks or more complex architectures, can capture intricate patterns in the data to predict registration trends**.**

6. Time Series Forecasting Techniques:

Methods such as ARIMA (AutoRegressive Integrated Moving Average) or Exponential Smoothing can be useful if you're dealing with time-dependent data to forecast registration trends over time.

7. Ensemble Methods:

Techniques that combine multiple algorithms, such as bagging or stacking, to improve prediction accuracy and robustness.

8. Naive Bayes:

Particularly useful in cases where probabilities of different classes are required based on given features.

Insights Gained from EDA:

Data Quality Assessment:

EDA reveals missing values, anomalies, and inconsistencies that might affect the predictive model's performance. Cleaning and handling these issues appropriately can enhance model accuracy.

Feature Engineering:

EDA helps identify the most relevant features that strongly correlate with the target variable. Selecting and transforming these features can improve model performance.

Understanding Relationships:

Insights into relationships and correlations between variables obtained from EDA guide the selection of appropriate algorithms, particularly for models sensitive to multicollinearity or dependencies between features.

Identification of Outliers:

Outliers might adversely impact model training. EDA helps decide whether to remove, transform, or handle outliers to improve model robustness.

Data Distribution Understanding:

Knowledge of data distributions aids in selecting suitable models, particularly when assumptions about data distribution are crucial (e.g., normality assumptions in some statistical models).

The Performance Of Predictive Models

Improved Feature Selection:

EDA informs better feature selection, ensuring that the model is trained on the most relevant and meaningful predictors, consequently enhancing its predictive power.

Enhanced Model Robustness:

Addressing outliers, missing values, and inconsistencies identified during EDA improves the model's robustness and stability.

Optimized Model Choice:

Understanding the nature of the data helps in choosing the most appropriate algorithm that suits the dataset's characteristics, resulting in better model performance.

Reduced Overfitting:

EDA assists in identifying data patterns, preventing overfitting by guiding the modeling process and feature selection, resulting in more generalizable models.

Refinement of Hyperparameters:

Insights from EDA might guide the hyperparameter tuning process, fine-tuning the model to achieve optimal performance.

Improved Interpretability:

EDA insights lead to a deeper understanding of the data, making the model's predictions more interpretable and trustworthy.

Technology Architecture

1. Data Collection and Integration:

- Data Sources: Collect data from Registrar of Companies databases, government registries, and other relevant sources.
- Data Integration: Utilize ETL (Extract, Transform, Load) processes to integrate and preprocess data into a suitable format for analysis**.**

2. Data Storage:

- Database or Data Warehouse: Store structured company registration data in a secure, scalable, and accessible repository.

3. Data Preprocessing:

- Data Cleaning: Handle missing values, outliers, and inconsistencies identified during preprocessing.
- Feature Engineering: Extract relevant features, encode categorical variables, and transform data for analysis.

4. Exploratory Data Analysis (EDA):

- Visualization Tools: Employ data visualization tools to gain insights into data distributions, correlations, and anomalies.
- Statistical Analysis: Use statistical methods to understand patterns and relationships in the data.

5. Machine Learning Models:

- Algorithm Selection: Choose appropriate algorithms for predictive analysis. Examples might include decision trees, regression models, neural networks, or time series models.
- Model Training: Train models to predict company registration trends based on historical data.

6. Model Evaluation and Validation:

- Performance Metrics: Assess model performance using metrics like accuracy, precision, recall, ROC-AUC, etc.
- Validation Techniques: Employ cross-validation or holdout sets to validate model robustness.

7. Deployment and Integration:

- Scalable Infrastructure: Deploy models on scalable infrastructure, such as cloud-based platforms or on-premises servers.
- Integration with Registrar of Companies Systems: Establish interfaces or APIs to fetch and update data from the Registrar of Companies databases.

8. Monitoring and Maintenance:

- Continuous Monitoring: Monitor model performance and data quality regularly.
- Model Maintenance: Update and retrain models with new data periodically to maintain accuracy.

9. User Interface and Reporting:

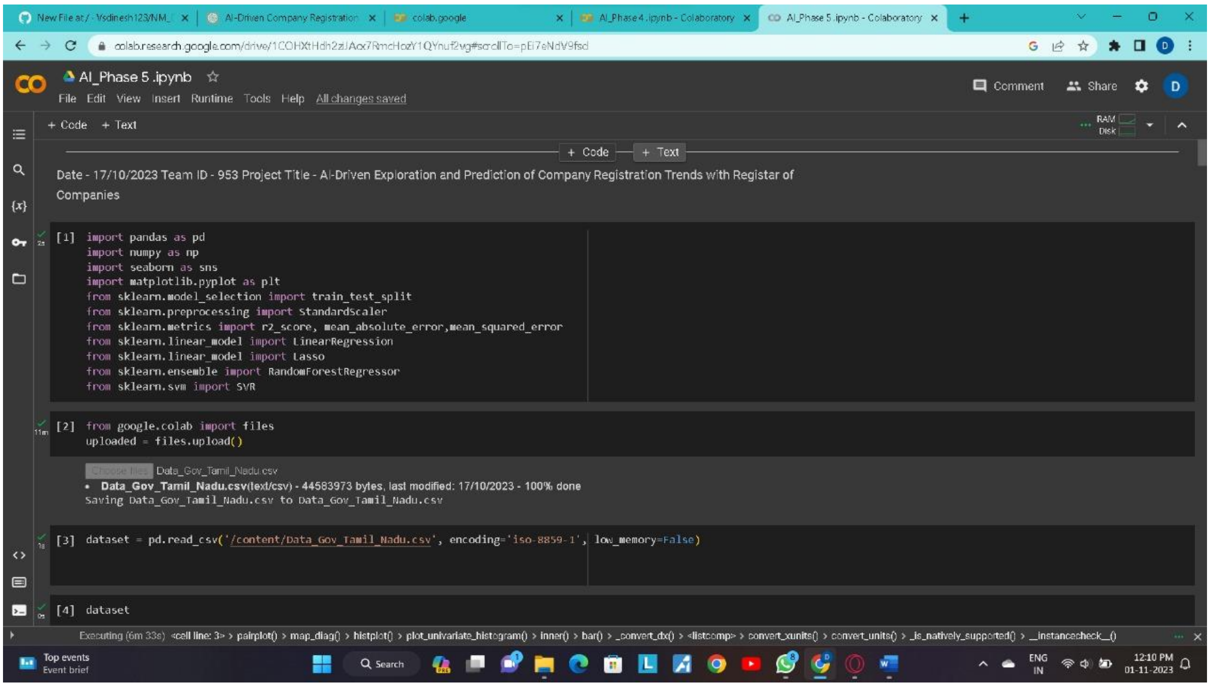
- Dashboard or Reporting Tools: Develop user interfaces or dashboards to present insights and predictions in a user-friendly manner.
- Automated Reporting: Generate regular reports on company registration trends and predictions.

10. Security and Compliance:

- Data Security Measures: Implement robust security measures to protect sensitive data.
- Compliance: Ensure compliance with data privacy and regulatory standards.

PROJECT DEVELOPMENT STEPS AND SCREENSHOT

Step 1: Import dependencies and loading the Dataset



Date - 17/10/2023 Team ID - 953 Project Title - AI-Driven Exploration and Prediction of Company Registration Trends with Registar of Companies

```
[1] 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.metrics import r2_score, mean_absolute_error,mean_squared_error
    from sklearn.linear_model import LinearRegression
    from sklearn.linear_model import Lasso
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.svm import SVR
```

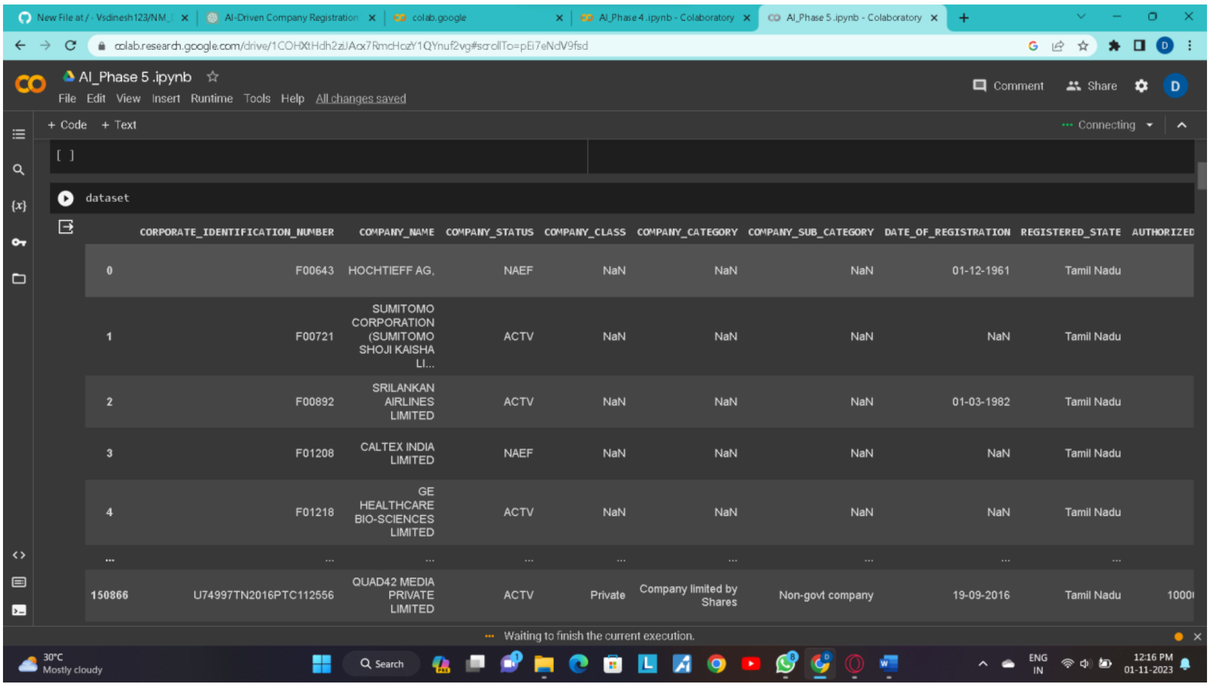
```
[2] from google.colab import files
    uploaded = files.upload()
```

```
Choose files Data_Gov_Tamil_Nadu.csv
• Data_Gov_Tamil_Nadu.csv(text/csv) - 44583973 bytes, last modified: 17/10/2023 - 100% done
Saving Data_Gov_Tamil_Nadu.csv to Data_Gov_Tamil_Nadu.csv
```

```
[3] dataset = pd.read_csv('/content/Data_Gov_Tamil_Nadu.csv', encoding='iso-8859-1', low_memory=False)
```

```
[4] dataset
```

```
Executing (6m 33s) <cell line: 3> > pairplot() > map_diag() > histplot() > plot_univariate_histogram() > inner() > bar() > _convert_dx() > <listcomp> > convert_xunits() > convert_units() > _is_natively_supported() > __instancecheck__()
```

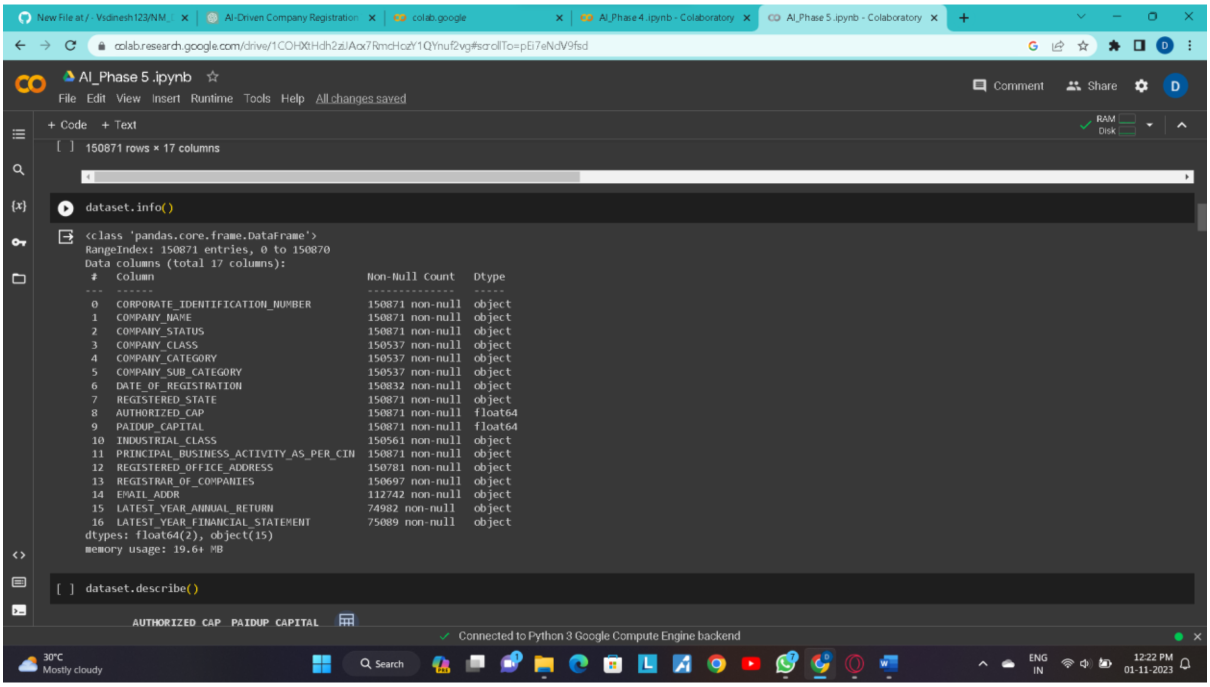


```
dataset
```

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DATE_OF_REGISTRATION	REGISTERED_STATE	AUTHORIZED
0	F00643	HOCHTIEFF AG,	NAEF	NaN	NaN	NaN	01-12-1961	Tamil Nadu	
1	F00721	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...	ACTV	NaN	NaN	NaN	NaN	Tamil Nadu	
2	F00892	SRILANKAN AIRLINES LIMITED	ACTV	NaN	NaN	NaN	01-03-1982	Tamil Nadu	
3	F01208	CALTEX INDIA LIMITED	NAEF	NaN	NaN	NaN	NaN	Tamil Nadu	
4	F01218	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	NaN	NaN	NaN	NaN	Tamil Nadu	
...	...	...	...	...	...	...	...	...	
150866	U74997TN2016PTC112556	QUAD42 MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	19-09-2016	Tamil Nadu	1000

Waiting to finish the current execution.

Step 2 : Dataset Information and describe the Dataset


New File at / - Vsdinesh123/NM_
AI-Driven Company Registration
colab.google
AI_Phase 4 .ipynb - Colaboratory
AI_Phase 5 .ipynb - Colaboratory
colab.research.google.com/drive/1COHXtHdh2ziJAcx7RmcHozY1QYnuf2vg#scrollTo=pEi7eNdV9fsd
AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Connected to Python 3 Google Compute Engine backend
30°C
Mostly cloudy
Search
ENG
IN
12:22 PM
01-11-2023

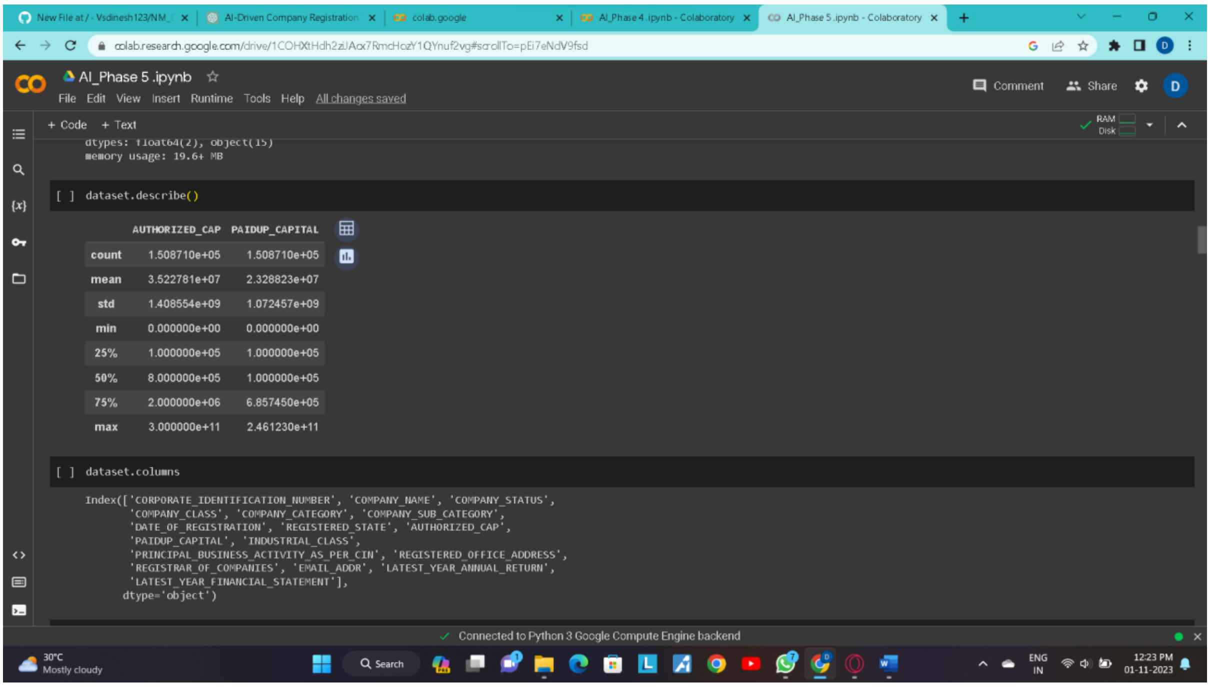
```
[ ] 150871 rows × 17 columns
```

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150871 entries, 0 to 150870
Data columns (total 17 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   CORPORATE_IDENTIFICATION_NUMBER       150871 non-null  object
 1   COMPANY_NAME                          150871 non-null  object
 2   COMPANY_STATUS                        150871 non-null  object
 3   COMPANY_CLASS                         150537 non-null  object
 4   COMPANY_CATEGORY                      150537 non-null  object
 5   COMPANY_SUB_CATEGORY                  150537 non-null  object
 6   DATE_OF_REGISTRATION                  150832 non-null  object
 7   REGISTERED_STATE                      150871 non-null  object
 8   AUTHORIZED_CAP                        150871 non-null  float64
 9   PAIDUP_CAPITAL                        150871 non-null  float64
 10  INDUSTRIAL_CLASS                      150561 non-null  object
 11  PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN  150871 non-null  object
 12  REGISTERED_OFFICE_ADDRESS             150781 non-null  object
 13  REGISTRAR_OF_COMPANIES                150697 non-null  object
 14  EMAIL_ADDR                            112742 non-null  object
 15  LATEST_YEAR_ANNUAL_RETURN             74982 non-null   object
 16  LATEST_YEAR_FINANCIAL_STATEMENT       75089 non-null   object
dtypes: float64(2), object(15)
memory usage: 19.6+ MB
```

```
[ ] dataset.describe()
```

```
AUTHORIZED_CAP  PAIDUP_CAPITAL
```


New File at / - Vsdinesh123/NM_
AI-Driven Company Registration
colab.google
AI_Phase 4 .ipynb - Colaboratory
AI_Phase 5 .ipynb - Colaboratory
colab.research.google.com/drive/1COHXtHdh2ziJAcx7RmcHozY1QYnuf2vg#scrollTo=pEi7eNdV9fsd
AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Connected to Python 3 Google Compute Engine backend
30°C
Mostly cloudy
Search
ENG
IN
12:23 PM
01-11-2023

```
dtypes: float64(2), object(15)
memory usage: 19.6+ MB
```

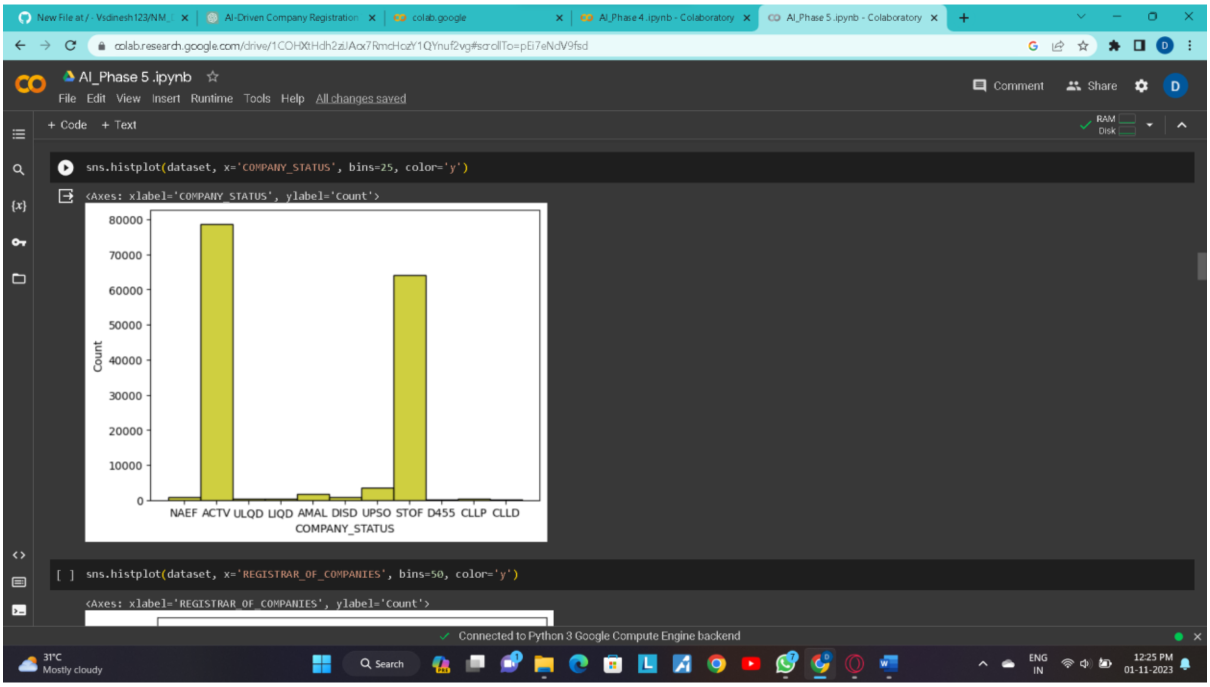
```
[ ] dataset.describe()
```

	AUTHORIZED_CAP	PAIDUP_CAPITAL
count	1.508710e+05	1.508710e+05
mean	3.522781e+07	2.328823e+07
std	1.408554e+09	1.072457e+09
min	0.000000e+00	0.000000e+00
25%	1.000000e+05	1.000000e+05
50%	8.000000e+05	1.000000e+05
75%	2.000000e+06	6.857450e+05
max	3.000000e+11	2.461230e+11

```
[ ] dataset.columns
```

```
Index(['CORPORATE_IDENTIFICATION_NUMBER', 'COMPANY_NAME', 'COMPANY_STATUS',
       'COMPANY_CLASS', 'COMPANY_CATEGORY', 'COMPANY_SUB_CATEGORY',
       'DATE_OF_REGISTRATION', 'REGISTERED_STATE', 'AUTHORIZED_CAP',
       'PAIDUP_CAPITAL', 'INDUSTRIAL_CLASS',
       'PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN', 'REGISTERED_OFFICE_ADDRESS',
       'REGISTRAR_OF_COMPANIES', 'EMAIL_ADDR', 'LATEST_YEAR_ANNUAL_RETURN',
       'LATEST_YEAR_FINANCIAL_STATEMENT'],
      dtype='object')
```

Step 3: Pre-processing and visualisation

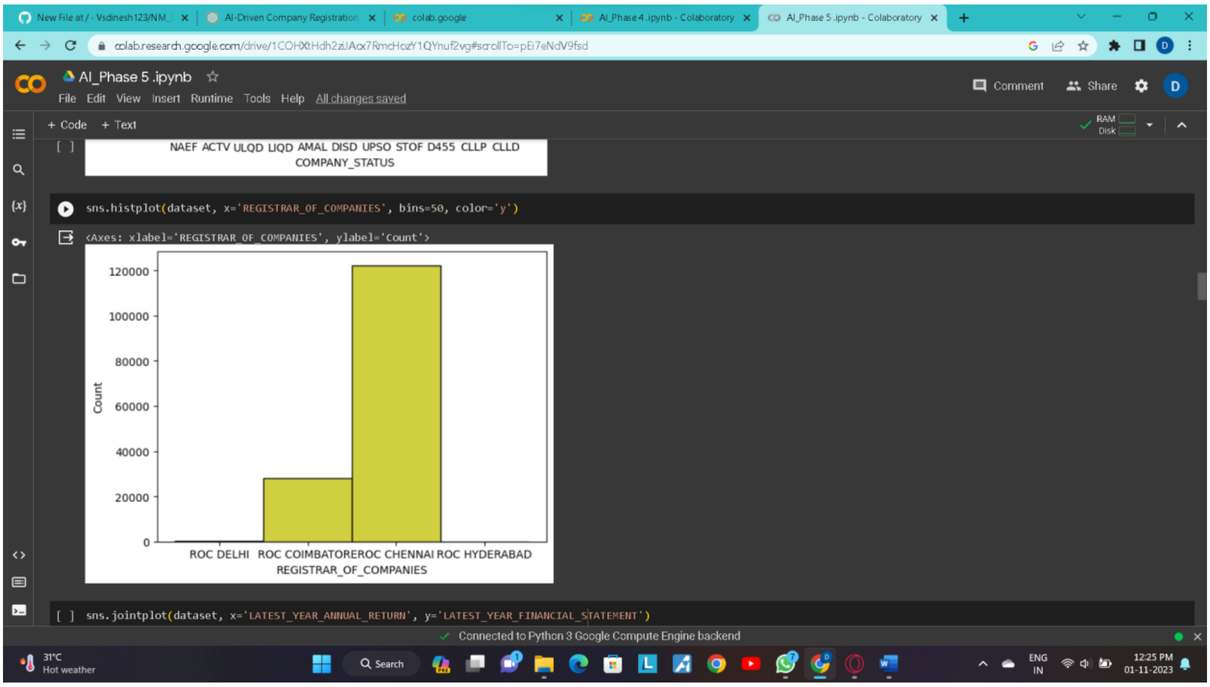


```
sns.histplot(dataset, x='COMPANY_STATUS', bins=25, color='y')
<Axes: xlabel='COMPANY_STATUS', ylabel='Count'>
```

```
sns.histplot(dataset, x='REGISTRAR_OF_COMPANIES', bins=50, color='y')
<Axes: xlabel='REGISTRAR_OF_COMPANIES', ylabel='Count'>
```

Connected to Python 3 Google Compute Engine backend

AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
colab.research.google.com/drive/1COHXtHdh2zJAcx7RmcHozY1QYnuf2vg#scrollTo=pEi7eNdV9fsd
Comment
Share
+ Code
+ Text
80000
70000
60000
50000
40000
30000
20000
10000
0
Count
NAEF ACTV ULQD LIQD AMAL DISD UPSO STOF D455 CLLP CLLD
COMPANY_STATUS
31°C
Mostly cloudy
Search
ENG IN
12:25 PM
01-11-2023

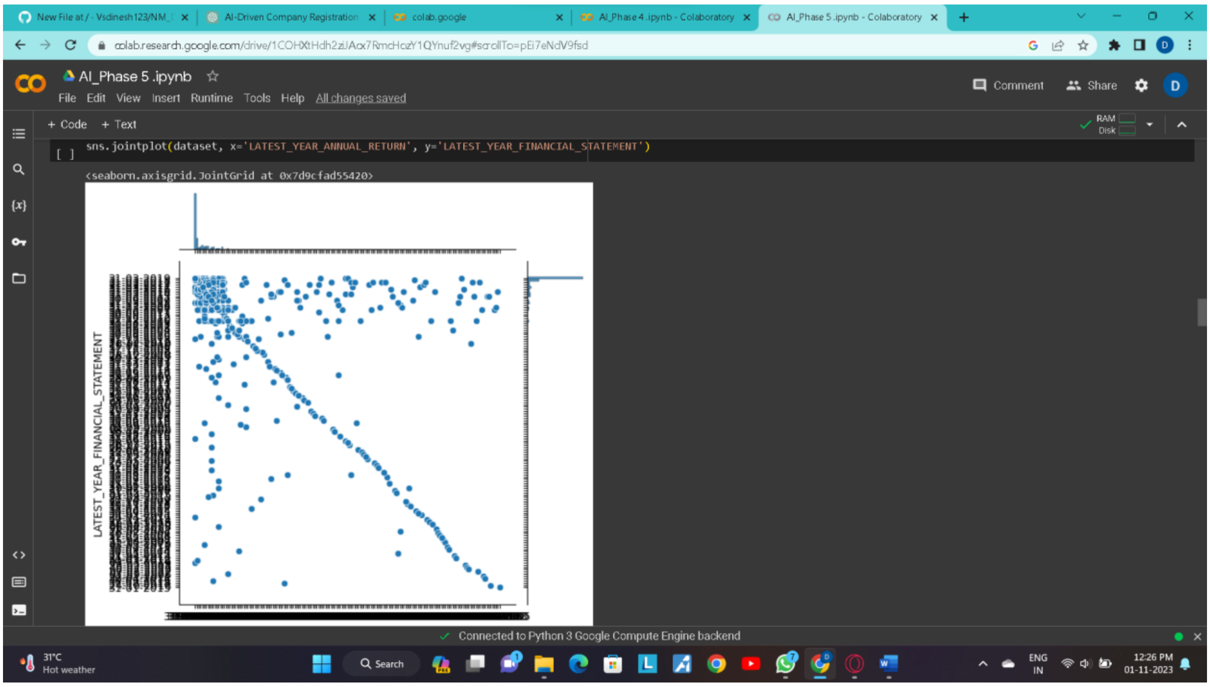


```
sns.histplot(dataset, x='REGISTRAR_OF_COMPANIES', bins=50, color='y')
<Axes: xlabel='REGISTRAR_OF_COMPANIES', ylabel='Count'>
```

```
sns.jointplot(dataset, x='LATEST_YEAR_ANNUAL_RETURN', y='LATEST_YEAR_FINANCIAL_STATEMENT')
```

Connected to Python 3 Google Compute Engine backend

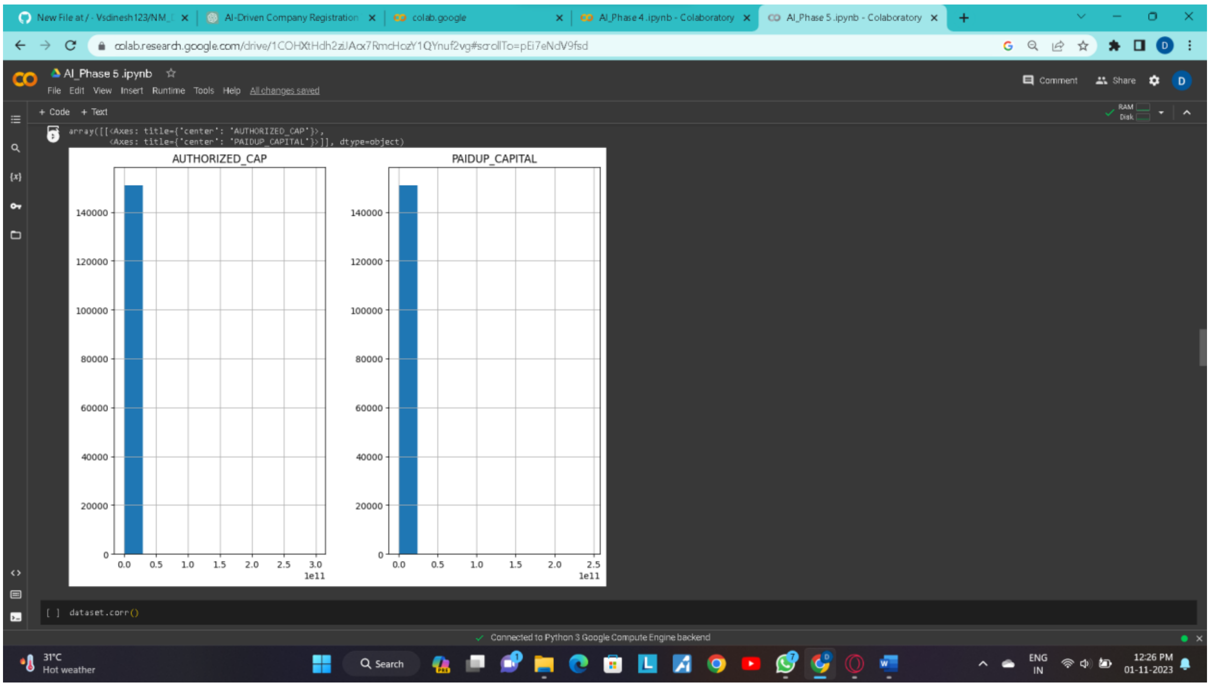
AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
NAEF ACTV ULQD LIQD AMAL DISD UPSO STOF D455 CLLP CLLD
COMPANY_STATUS
120000
100000
80000
60000
40000
20000
0
Count
ROC DELHI ROC COIMBATORE ROC CHENNAI ROC HYDERABAD
REGISTRAR_OF_COMPANIES
31°C
Hot weather
Search
ENG IN
12:25 PM
01-11-2023



```
sns.jointplot(dataset, x='LATEST_YEAR_ANNUAL_RETURN', y='LATEST_YEAR_FINANCIAL_STATEMENT')
```

```
<seaborn.axisgrid.JointGrid at 0x7d9cfad55420>
```

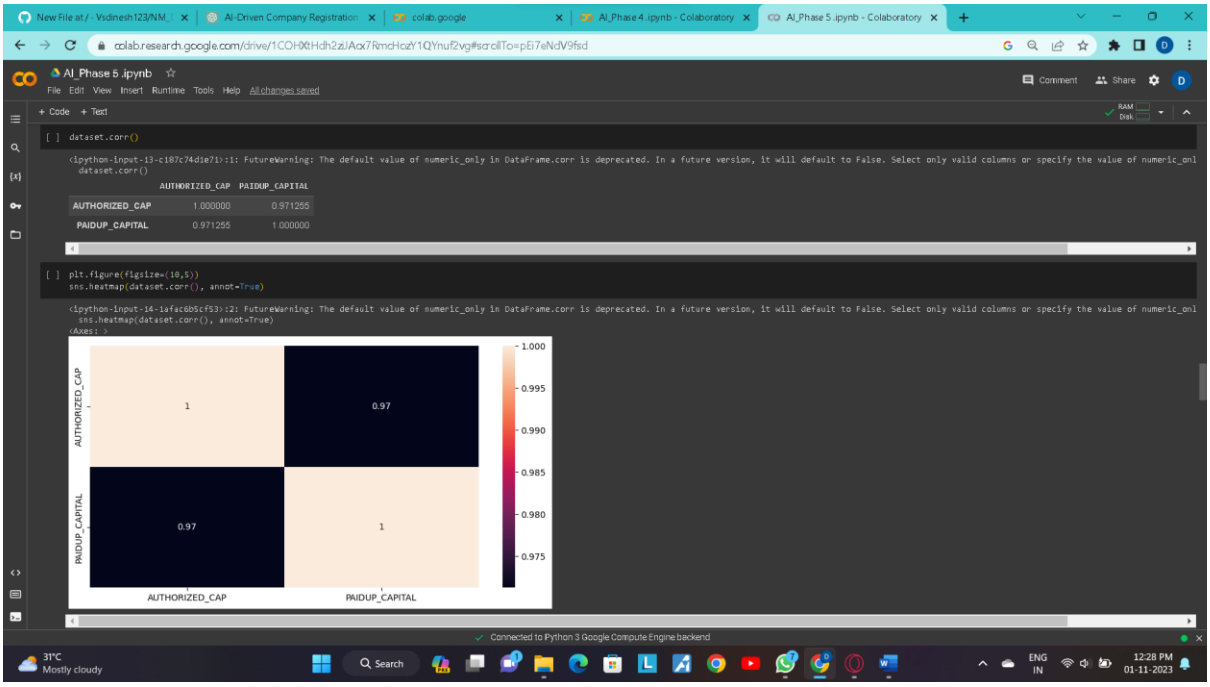
LATEST_YEAR_FINANCIAL_STATEMENT
31-03-2019
Connected to Python 3 Google Compute Engine backend



```
array([[<Axes: title={'center': 'AUTHORIZED_CAP'}>,
        <Axes: title={'center': 'PAIDUP_CAPITAL'}>]], dtype=object)
```

```
dataset.corr()
```

Step 4 : Dataset Correlation



AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
AUTHORIZED_CAP
PAIDUP_CAPITAL
1
0.97
0.97
1
1.000
0.995
0.990
0.985
0.980
0.975
Connected to Python 3 Google Compute Engine backend

```
dataset.corr()
```

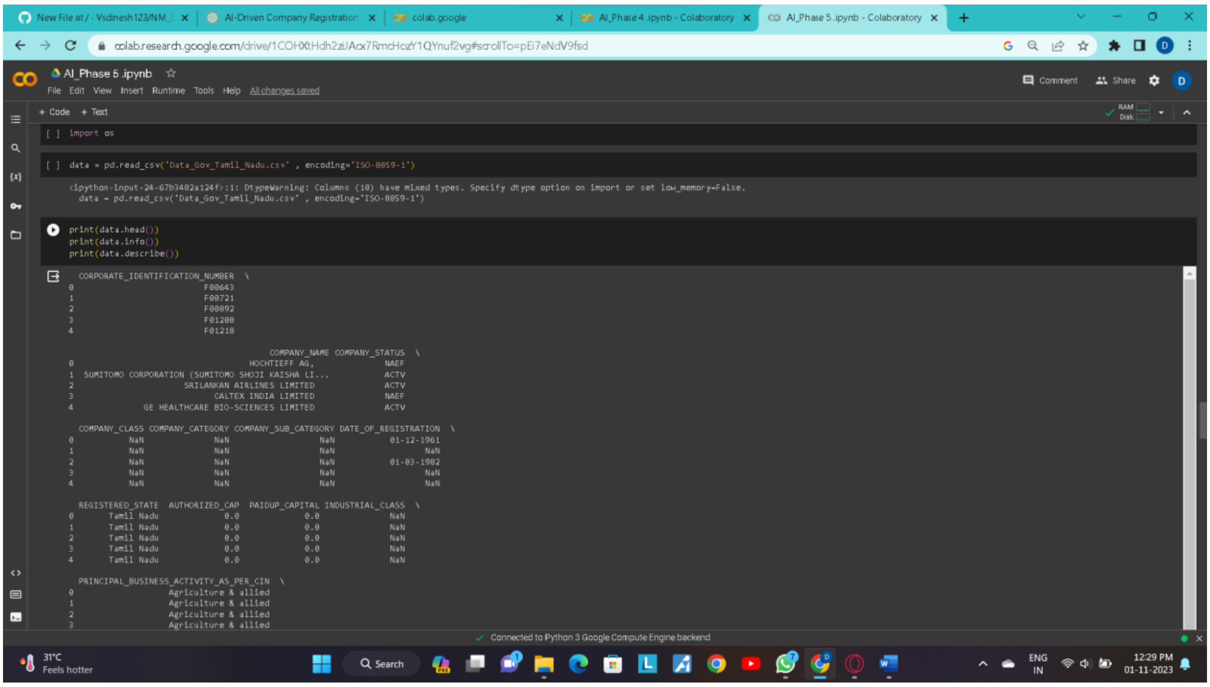
```
<ipython-input-13-c187c74d1e71>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_onl
  dataset.corr()
```

	AUTHORIZED_CAP	PAIDUP_CAPITAL
AUTHORIZED_CAP	1.000000	0.971255
PAIDUP_CAPITAL	0.971255	1.000000

```
plt.figure(figsize=(10,5))
sns.heatmap(dataset.corr(), annot=True)
```

```
<ipython-input-14-1afac6b5cf53>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_onl
  sns.heatmap(dataset.corr(), annot=True)
<Axes: >
```

Step 5: Import dependencies and print the Dataset



AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Connected to Python 3 Google Compute Engine backend

```
import os
```

```
data = pd.read_csv('Data_Gov_Tamil_Nadu.csv' , encoding='ISO-8859-1')
```

```
<ipython-input-24-67b3482a124f>:1: DtypeWarning: Columns (10) have mixed types. Specify dtype option on import or set low_memory=False.
  data = pd.read_csv('Data_Gov_Tamil_Nadu.csv' , encoding='ISO-8859-1')
```

```
print(data.head())
print(data.info())
print(data.describe())
```

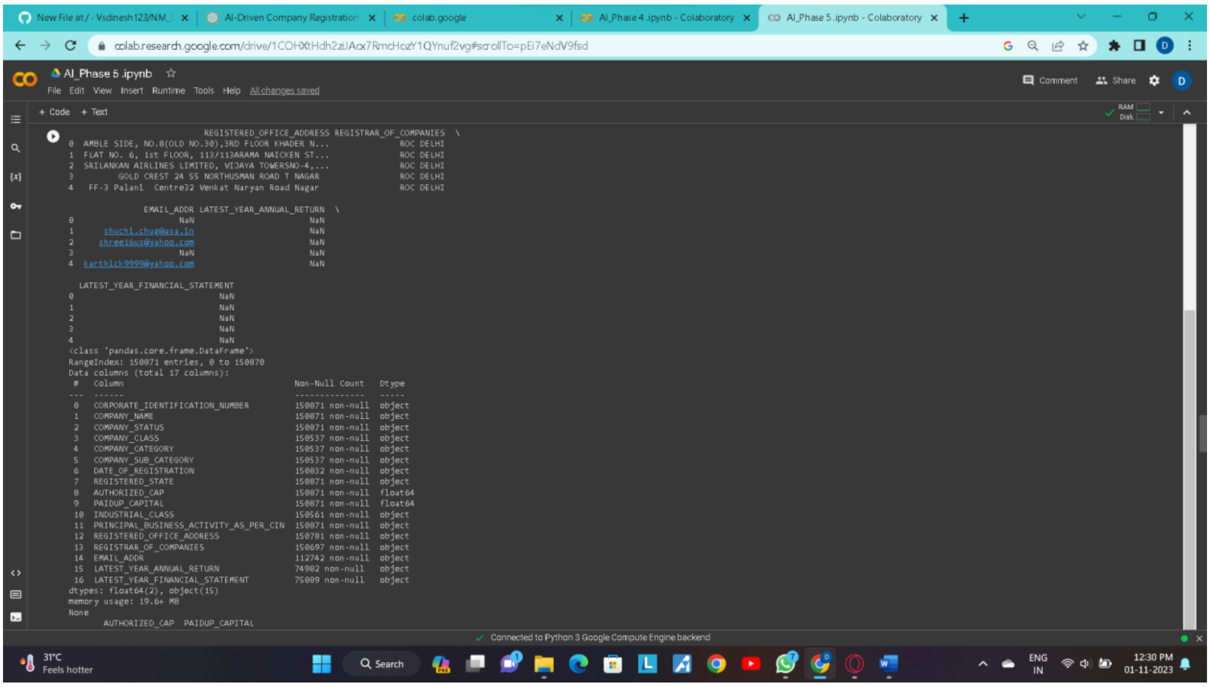
```
  CORPORATE_IDENTIFICATION_NUMBER  \
0                          F00643
1                          F00721
2                          F00892
3                          F01208
4                          F01218

                                    COMPANY_NAME COMPANY_STATUS  \
0                                HOCHTIEFF AG,             NAEF
1  SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...           ACTV
2                     SRILANKAN AIRLINES LIMITED           ACTV
3                           CALTEX INDIA LIMITED           NAEF
4               GE HEALTHCARE BIO-SCIENCES LIMITED           ACTV

  COMPANY_CLASS COMPANY_CATEGORY COMPANY_SUB_CATEGORY DATE_OF_REGISTRATION  \
0           NaN              NaN                  NaN           01-12-1961
1           NaN              NaN                  NaN                  NaN
2           NaN              NaN                  NaN           01-03-1982
3           NaN              NaN                  NaN                  NaN
4           NaN              NaN                  NaN                  NaN

  REGISTERED_STATE  AUTHORIZED_CAP  PAIDUP_CAPITAL INDUSTRIAL_CLASS  \
0       Tamil Nadu             0.0             0.0              NaN
1       Tamil Nadu             0.0             0.0              NaN
2       Tamil Nadu             0.0             0.0              NaN
3       Tamil Nadu             0.0             0.0              NaN
4       Tamil Nadu             0.0             0.0              NaN

  PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN  \
0                Agriculture & allied
1                Agriculture & allied
2                Agriculture & allied
3                Agriculture & allied
```

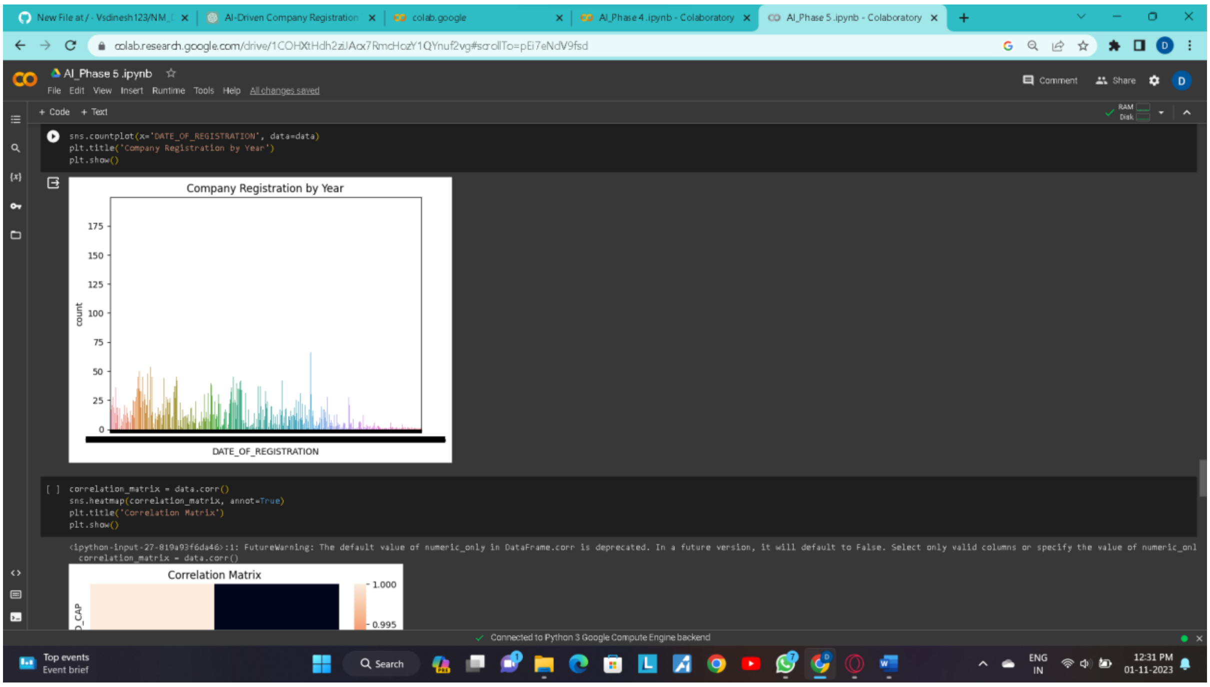

AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Connected to Python 3 Google Compute Engine backend

```
                         REGISTERED_OFFICE_ADDRESS REGISTRAR_OF_COMPANIES  \
0  AMBLE SIDE, NO.8(OLD NO.30),3RD FLOOR KHADER N...             ROC DELHI
1  FLAT NO. 6, 1st FLOOR, 113/113ARAMA NAICKEN ST...             ROC DELHI
2  SRILANKAN AIRLINES LIMITED, VIJAYA TOWERSNO-4,...             ROC DELHI
3         GOLD CREST 24 55 NORTHUSMAN ROAD T NAGAR               ROC DELHI
4   FF-3 Palani  Centre32 Venkat Naryan Road Nagar               ROC DELHI

               EMAIL_ADDR LATEST_YEAR_ANNUAL_RETURN  \
0                     NaN                       NaN
1       shuchi.chug@asa.in                       NaN
2     shree16us@yahoo.com                       NaN
3                     NaN                       NaN
4  karthick9999@yahoo.com                       NaN

  LATEST_YEAR_FINANCIAL_STATEMENT
0                             NaN
1                             NaN
2                             NaN
3                             NaN
4                             NaN
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150871 entries, 0 to 150870
Data columns (total 17 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   CORPORATE_IDENTIFICATION_NUMBER       150871 non-null  object
 1   COMPANY_NAME                          150871 non-null  object
 2   COMPANY_STATUS                        150871 non-null  object
 3   COMPANY_CLASS                         150537 non-null  object
 4   COMPANY_CATEGORY                      150537 non-null  object
 5   COMPANY_SUB_CATEGORY                  150537 non-null  object
 6   DATE_OF_REGISTRATION                  150832 non-null  object
 7   REGISTERED_STATE                      150871 non-null  object
 8   AUTHORIZED_CAP                        150871 non-null  float64
 9   PAIDUP_CAPITAL                        150871 non-null  float64
 10  INDUSTRIAL_CLASS                      150561 non-null  object
 11  PRINCIPAL_BUSINESS_ACTIVITY_AS_PER_CIN  150871 non-null  object
 12  REGISTERED_OFFICE_ADDRESS             150781 non-null  object
 13  REGISTRAR_OF_COMPANIES                150697 non-null  object
 14  EMAIL_ADDR                            112742 non-null  object
 15  LATEST_YEAR_ANNUAL_RETURN             74982 non-null   object
 16  LATEST_YEAR_FINANCIAL_STATEMENT       75089 non-null   object
dtypes: float64(2), object(15)
memory usage: 19.6+ MB
None
       AUTHORIZED_CAP  PAIDUP_CAPITAL
```

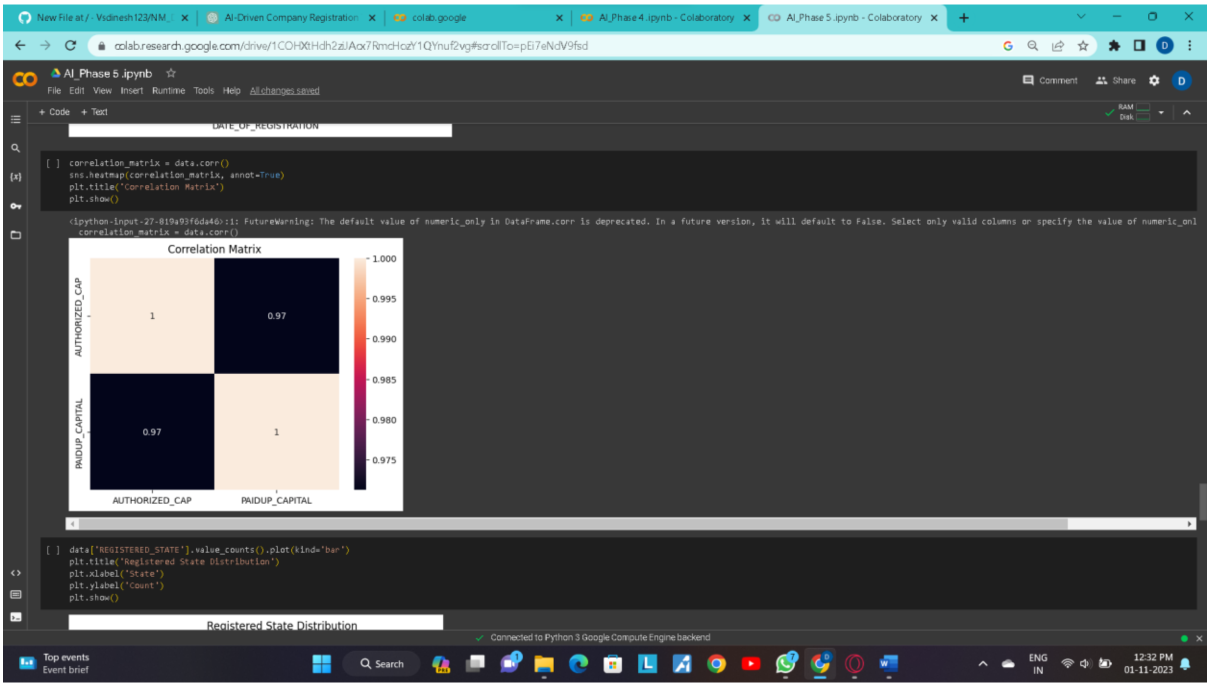
Step 6: Countplot and visualization of correlation matrix


AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Company Registration by Year
175
150
125
100
75
50
25
0
count
DATE_OF_REGISTRATION
Correlation Matrix
1.000
0.995
D_CAP
Connected to Python 3 Google Compute Engine backend

```
sns.countplot(x='DATE_OF_REGISTRATION', data=data)
plt.title('Company Registration by Year')
plt.show()
```

```
correlation_matrix = data.corr()
sns.heatmap(correlation_matrix, annot=True)
plt.title('Correlation Matrix')
plt.show()
```

```
<ipython-input-27-819a93f6da46>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_onl
  correlation_matrix = data.corr()
```



New File at / · Vsdinesh123/NM_
AI-Driven Company Registration
colab.google
AI_Phase 4 .ipynb - Colaboratory
AI_Phase 5 .ipynb - Colaboratory
colab.research.google.com/drive/1COHXtHdh2ziJAcx7RmcHozY1QYnuf2vg#scrollTo=pEi7eNdV9fsd
AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
DATE_OF_REGISTRATION

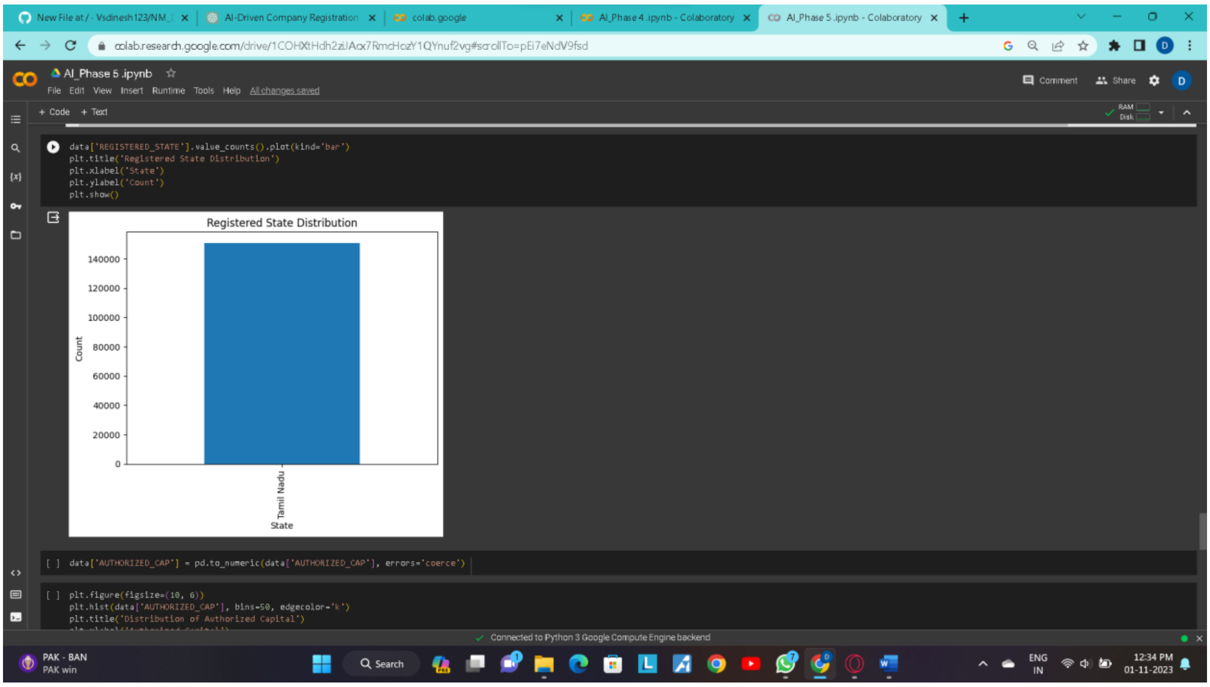
```
correlation_matrix = data.corr()
sns.heatmap(correlation_matrix, annot=True)
plt.title('Correlation Matrix')
plt.show()
```

```
<ipython-input-27-819a93f6da46>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_onl
  correlation_matrix = data.corr()
```

Correlation Matrix
AUTHORIZED_CAP
PAIDUP_CAPITAL
1
0.97
0.97
1
AUTHORIZED_CAP
PAIDUP_CAPITAL
1.000
0.995
0.990
0.985
0.980
0.975

```
data['REGISTERED_STATE'].value_counts().plot(kind='bar')
plt.title('Registered State Distribution')
plt.xlabel('State')
plt.ylabel('Count')
plt.show()
```

Registered State Distribution
Connected to Python 3 Google Compute Engine backend
Top events
Event brief
Search
ENG
IN
12:32 PM
01-11-2023



New File at / · Vsdinesh123/NM_
AI-Driven Company Registration
colab.google
AI_Phase 4 .ipynb - Colaboratory
AI_Phase 5 .ipynb - Colaboratory
colab.research.google.com/drive/1COHXtHdh2ziJAcx7RmcHozY1QYnuf2vg#scrollTo=pEi7eNdV9fsd
AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk

```
data['REGISTERED_STATE'].value_counts().plot(kind='bar')
plt.title('Registered State Distribution')
plt.xlabel('State')
plt.ylabel('Count')
plt.show()
```

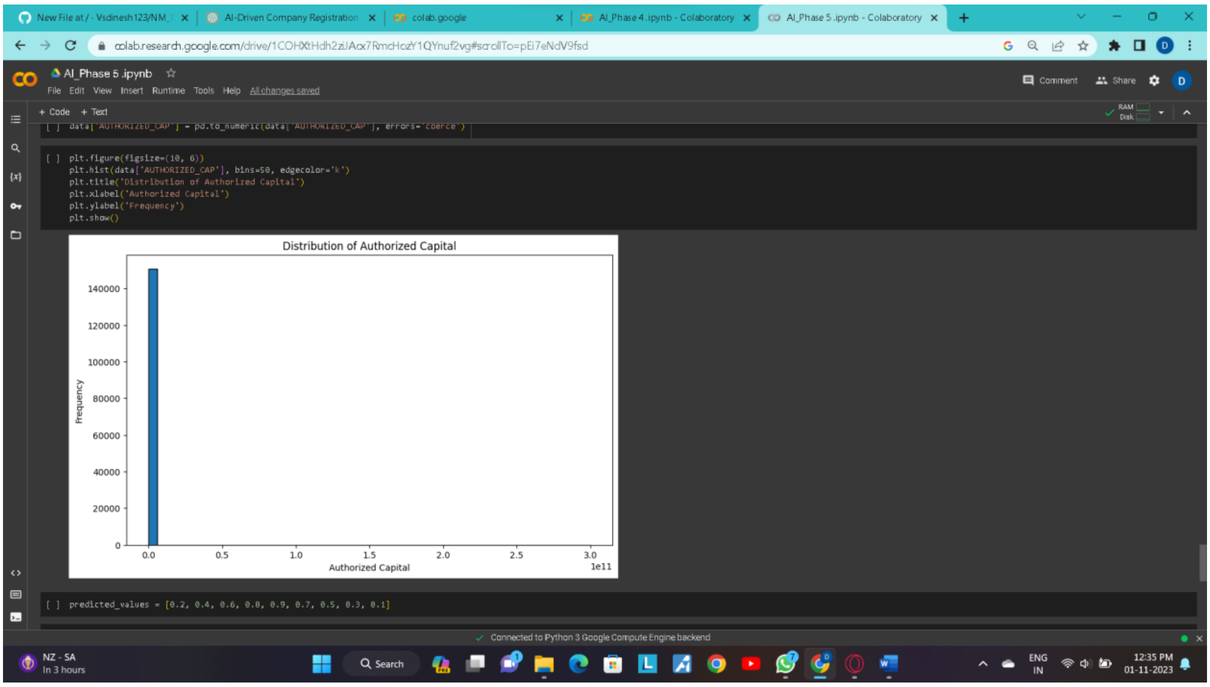
Registered State Distribution
Count
140000
120000
100000
80000
60000
40000
20000
0
Tamil Nadu
State

```
data['AUTHORIZED_CAP'] = pd.to_numeric(data['AUTHORIZED_CAP'], errors='coerce')
```

```
plt.figure(figsize=(10, 6))
plt.hist(data['AUTHORIZED_CAP'], bins=50, edgecolor='k')
plt.title('Distribution of Authorized Capital')
```

Connected to Python 3 Google Compute Engine backend
PAK - BAN
PAK win
Search
ENG
IN
12:34 PM
01-11-2023

Step 7: Distribution of capital and predicted values distribution

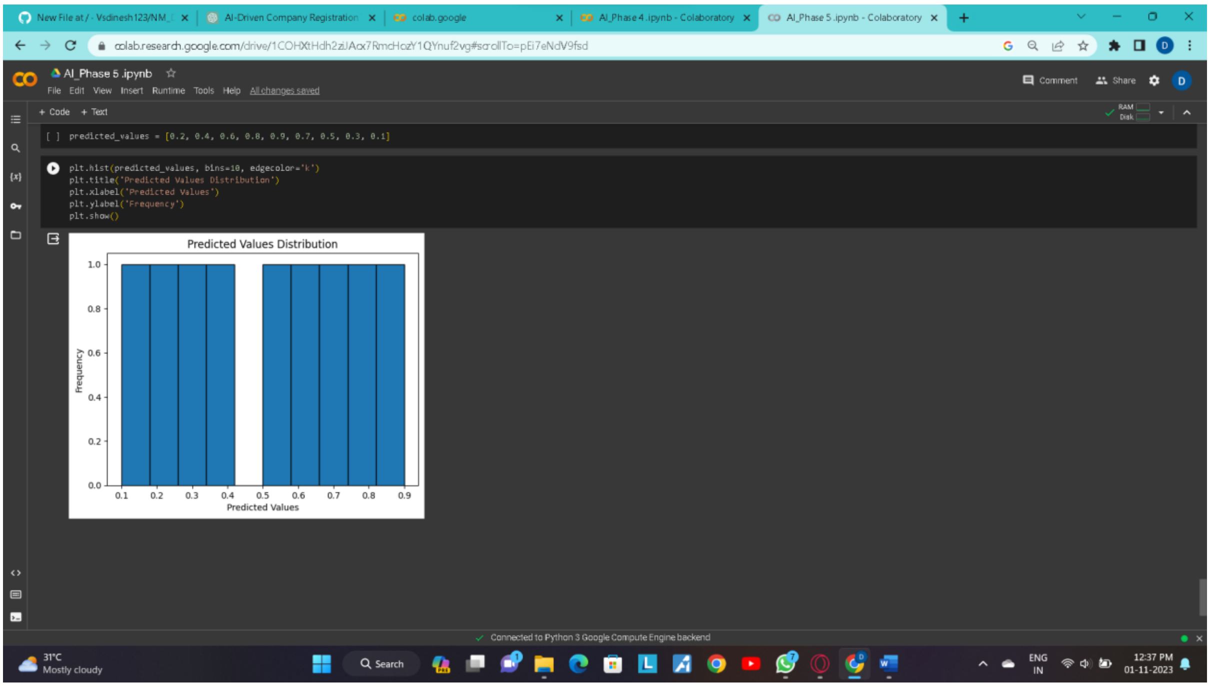


```
data['AUTHORIZED_CAP'] = pd.to_numeric(data['AUTHORIZED_CAP'], errors='coerce')
```

```
plt.figure(figsize=(10, 6))
plt.hist(data['AUTHORIZED_CAP'], bins=50, edgecolor='k')
plt.title('Distribution of Authorized Capital')
plt.xlabel('Authorized Capital')
plt.ylabel('Frequency')
plt.show()
```

AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Distribution of Authorized Capital
Frequency
140000
120000
100000
80000
60000
40000
20000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e11
Authorized Capital
Connected to Python 3 Google Compute Engine backend

```
predicted_values = [0.2, 0.4, 0.6, 0.8, 0.9, 0.7, 0.5, 0.3, 0.1]
```



```
predicted_values = [0.2, 0.4, 0.6, 0.8, 0.9, 0.7, 0.5, 0.3, 0.1]
```

```
plt.hist(predicted_values, bins=10, edgecolor='k')
plt.title('Predicted Values Distribution')
plt.xlabel('Predicted Values')
plt.ylabel('Frequency')
plt.show()
```

AI_Phase 5 .ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Predicted Values Distribution
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Predicted Values
Connected to Python 3 Google Compute Engine backend

CONCLUSION

In conclusion, the successful creation and deployment of a phishing detection machine learning model in IBM Watson Studio exemplify the transformative potential of technology in enhancing online security. Leveraging AutoAI, the model development process is streamlined, and with a dataset from Kaggle, the system is primed to swiftly recognize and mitigate evolving phishing threats. This proactive approach to safeguarding online interactions not only empowers users and organizations but also underscores the value of leveraging advanced machine learning solutions in an era where cybersecurity is paramount.