AI_Phase5

Serial Number	Team Member Name	Registration Number
1	K.Navinraj	310821104064
2	M.P.Praveen Raja	310821104070
3	Sachin A	310821104081
4	K.P.Tharun	310821104100

Project Title:

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

Problem Definition:

In today's rapidly evolving business landscape, data is an invaluable asset. The Registrar of Companies (RoC) maintains an extensive repository of data on registered companies, encompassing a wide array of attributes, and paid-up capital. Leveraging this vast dataset, our project endeavours to address a multifaceted challenge: conducting AI-driven exploration and predictive analysis.

The overarching goal of our project is threefold:

1. Uncovering Hidden Patterns.
2. Gaining Deep Insights.
3. Predictive Analysis.

By providing predictive capabilities, our project equips stakeholders with the ability to anticipate market shifts, make proactive investments, and enact forward-thinking policies.

Design Thinking

In the principles of Design Thinking, a human-centred methodology that fosters innovation. The following step-by-step process outlines our approach, which incorporates technology and data-driven techniques:

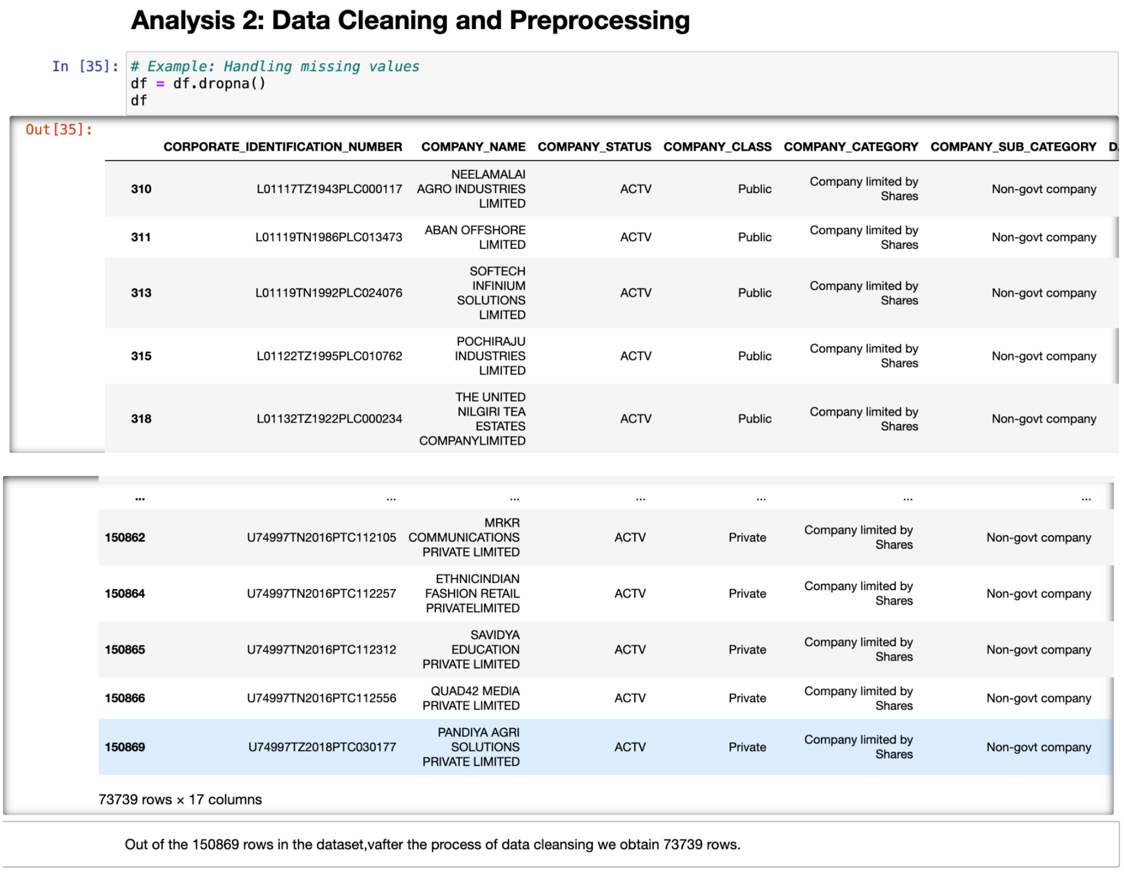
Step 1: Data Source

We begin by tapping into the dataset containing information about registered companies from the Registrar of Companies (RoC). It encompasses a multitude of attributes. This wealth of data is a treasure trove of information waiting to be explored and analysed.

We used the "Company Master Data of Tamil Nadu up to 28th February 2019" dataset obtained from the Open Government Data (OGD) Platform India.
Dataset link: Company Master Data of Tamil Nadu upto 28th February 2019 | Open Government Data (OGD) Platform India

Step 2: Data Preprocessing

To ensure the reliability and accuracy of our analysis, we embark on a comprehensive data preprocessing journey. And converting categorical features into numerical representations through methods like one-hot encoding or label encoding. This step lays the foundation for robust and reliable analysis.
Data cleaning and preprocessing are foundational to the success of any AI-driven exploration and prediction project.Data cleaning and preprocessing is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset while preparing it for analysis or machine learning.



Analysis 2: Data Cleaning and Preprocessing

```
In [35]: # Example: Handling missing values
         df = df.dropna()
         df
```

Out[35]:

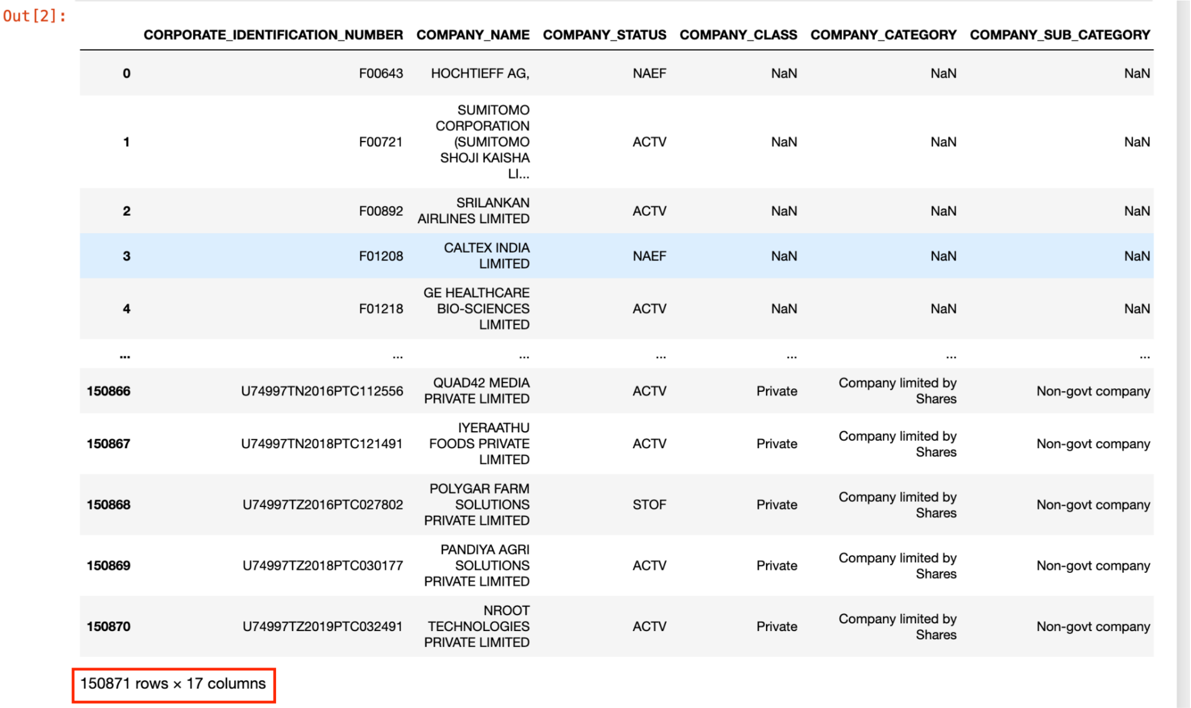
	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	D
310	L01117TZ1943PLC000117	NEELAMALAI AGRO INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
311	L01119TN1986PLC013473	ABAN OFFSHORE LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
313	L01119TN1992PLC024076	SOFTECH INFINIUM SOLUTIONS LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
315	L01122TZ1995PLC010762	POCHIRAJU INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
318	L01132TZ1922PLC000234	THE UNITED NILGIRI TEA ESTATES COMPANYLIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
...	...	...	...	...	...	...	
150862	U74997TN2016PTC112105	MRKR COMMUNICATIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150864	U74997TN2016PTC112257	ETHNICINDIAN FASHION RETAIL PRIVATELIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150865	U74997TN2016PTC112312	SAVIDYA EDUCATION PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150866	U74997TN2016PTC112556	QUAD42 MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	
150869	U74997TZ2018PTC030177	PANDIYA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company	

73739 rows × 17 columns

Out of the 150869 rows in the dataset,vafter the process of data cleansing we obtain 73739 rows.

Dataset Overview:

Basic Information about the Dataset is to explore and predict company registration trends using Registrar of Companies (RoC) data with AI-driven methods, below is basic information about the dataset required for the project:



Out[2]:

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY
0	F00643	HOCHTIEFF AG,	NAEF	NaN	NaN	NaN
1	F00721	SUMITOMO CORPORATION (SUMITOMO SHOJI KAISHA LI...	ACTV	NaN	NaN	NaN
2	F00892	SRILANKAN AIRLINES LIMITED	ACTV	NaN	NaN	NaN
3	F01208	CALTEX INDIA LIMITED	NAEF	NaN	NaN	NaN
4	F01218	GE HEALTHCARE BIO-SCIENCES LIMITED	ACTV	NaN	NaN	NaN
...	...	...	...	...	...	...
150866	U74997TN2016PTC112556	QUAD42 MEDIA PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company
150867	U74997TN2018PTC121491	IYERAATHU FOODS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company
150868	U74997TZ2016PTC027802	POLYGAR FARM SOLUTIONS PRIVATE LIMITED	STOF	Private	Company limited by Shares	Non-govt company
150869	U74997TZ2018PTC030177	PANDIYA AGRI SOLUTIONS PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company
150870	U74997TZ2019PTC032491	NROOT TECHNOLOGIES PRIVATE LIMITED	ACTV	Private	Company limited by Shares	Non-govt company

150871 rows × 17 columns

Company Registration Data: This dataset should include details about the companies registered with the Registrar of Companies. Key attributes may include:

- ➔ Company Name
- ➔ Registration Number
- ➔ Date of Registration

Data cleaning:

Data cleaning is foundational to the success of any AI-driven exploration and prediction project. It is the process of identifying and correcting errors, inconsistencies, and inaccuracies in a dataset while preparing it for analysis or machine learning

Step 3: Exploratory Data Analysis (EDA)

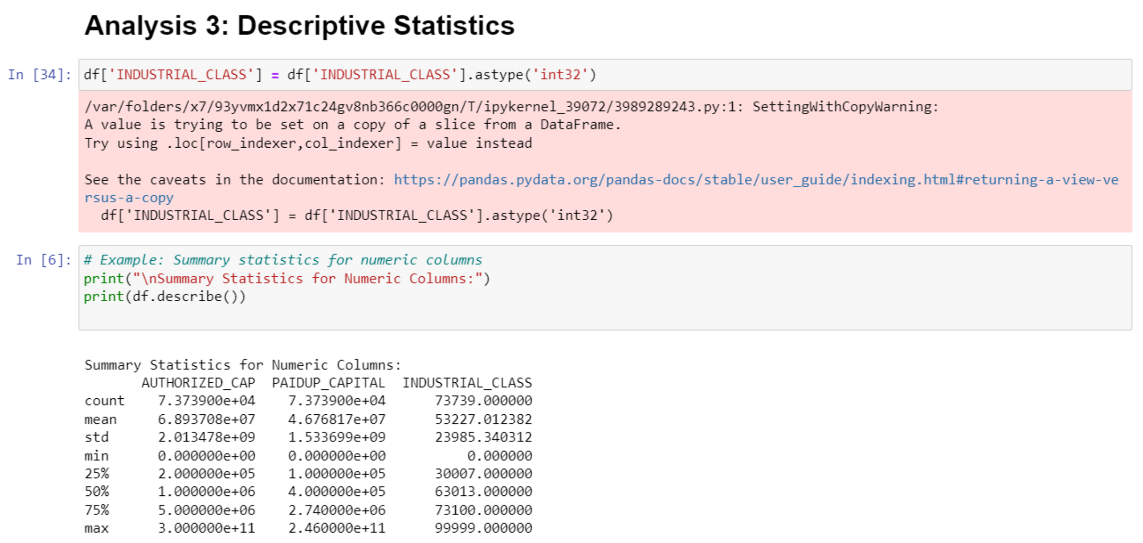
We explore data distributions, identify correlations between variables, detect anomalies, and unveil unique insights. Visualisation techniques such as histograms, scatter plots, and heatmaps are employed to make sense of the data's intricacies.

- EDA was crucial for gaining insights into the dataset.
- Descriptive statistics provided a preliminary understanding of the dataset.
- Effective visualisations, such as histograms, scatter plots, and heatmaps, were employed to convey insights, patterns, and trends in the data.
- Correlation analysis helped identify relationships between variables, and Principal Component Analysis (PCA) was applied for dimensionality reduction

Descriptive Statistics

Gains a preliminary understanding of your data when conducting AI-driven exploration and prediction of company registration trends with Registrar of Companies (RoC) data.

Descriptive statistics provide valuable insights into the characteristics of your data, enabling you to make informed decisions about data preprocessing, feature selection, and the choice of AI-driven modelling techniques.



Analysis 3: Descriptive Statistics

```
In [34]: df['INDUSTRIAL_CLASS'] = df['INDUSTRIAL_CLASS'].astype('int32')
```

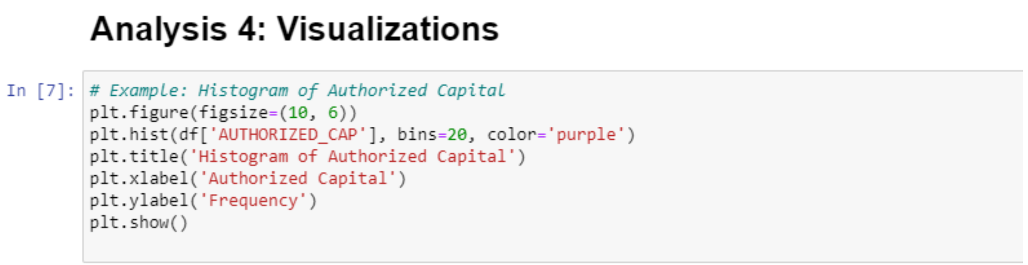
```
/var/folders/x7/93yvmx1d2x71c24gv8nb366c0000gn/T/ipykernel_39072/3989289243.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy
  df['INDUSTRIAL_CLASS'] = df['INDUSTRIAL_CLASS'].astype('int32')
```

```
In [6]: # Example: Summary statistics for numeric columns
print("\nSummary Statistics for Numeric Columns:")
print(df.describe())
```

```
Summary Statistics for Numeric Columns:
       AUTHORIZED_CAP  PAIDUP_CAPITAL  INDUSTRIAL_CLASS
count    7.373900e+04    7.373900e+04      73739.000000
mean     6.893708e+07    4.676817e+07      53227.012382
std      2.013478e+09    1.533699e+09      23985.340312
min      0.000000e+00    0.000000e+00          0.000000
25%      2.000000e+05    1.000000e+05      30007.000000
50%      1.000000e+06    4.000000e+05      63013.000000
75%      5.000000e+06    2.740000e+06      73100.000000
max      3.000000e+11    2.460000e+11      99999.000000
```

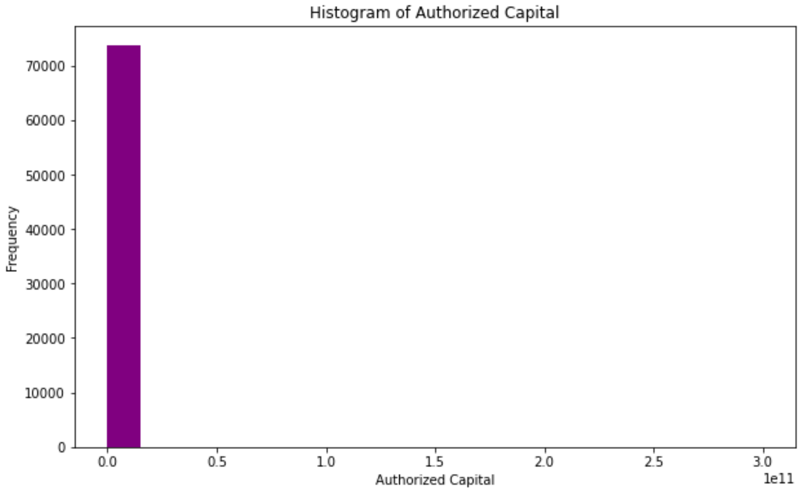
Visualisations

Visualisations are a powerful tool for conveying insights, patterns, and trends in your data when conducting AI-driven exploration and prediction of company registration trends with Registrar of Companies (RoC) data. Effective visualisations play a crucial role in conveying insights, facilitating data exploration, and aiding in decision-making in AI-driven projects focused on company registration trends



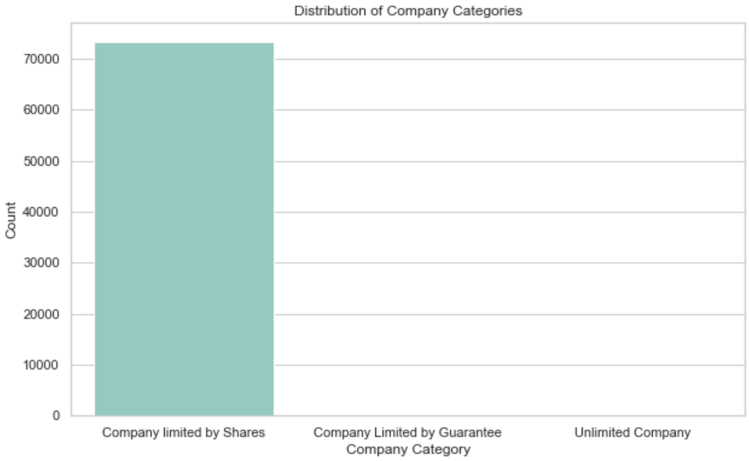
Analysis 4: Visualizations

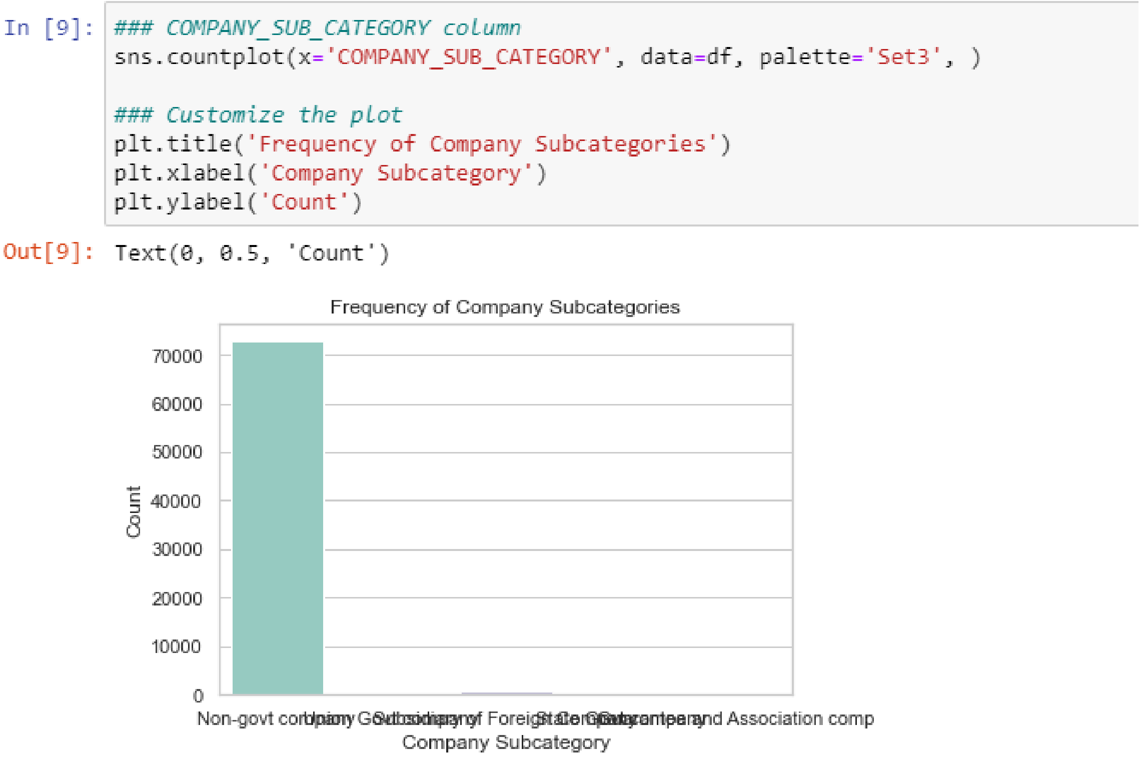
```
In [7]: # Example: Histogram of Authorized Capital
plt.figure(figsize=(10, 6))
plt.hist(df['AUTHORIZED_CAP'], bins=20, color='purple')
plt.title('Histogram of Authorized Capital')
plt.xlabel('Authorized Capital')
plt.ylabel('Frequency')
plt.show()
```



```
In [8]: #Distribution of Company Categories
        sns.set(style='whitegrid')
        plt.figure(figsize=(10, 6))

        # Create the count plot for the COMPANY_CATEGORY column
        sns.countplot(x='COMPANY_CATEGORY', data=df, palette='Set3', color = 'skyblue')

        # Customize the plot (optional)
        plt.title('Distribution of Company Categories')
        plt.xlabel('Company Category')
        plt.ylabel('Count')
```

```
Out[8]: Text(0, 0.5, 'Count')
```

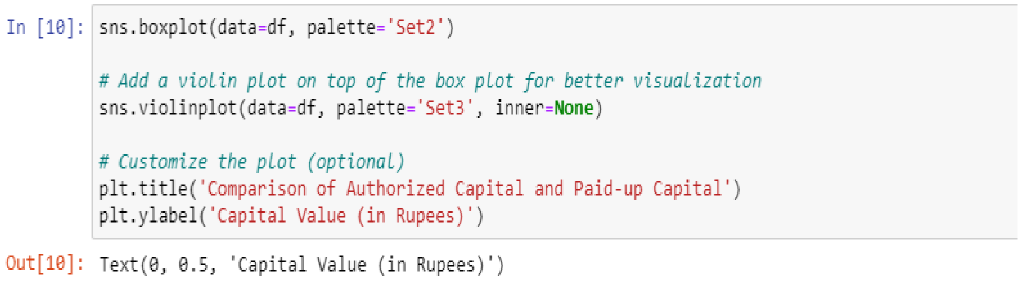




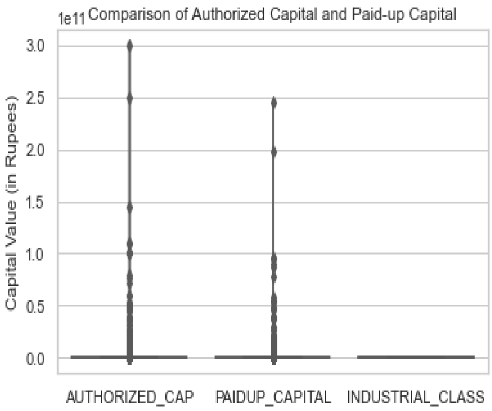
```
In [9]: ### COMPANY_SUB_CATEGORY column
        sns.countplot(x='COMPANY_SUB_CATEGORY', data=df, palette='Set3', )

        ### Customize the plot
        plt.title('Frequency of Company Subcategories')
        plt.xlabel('Company Subcategory')
        plt.ylabel('Count')
```

```
Out[9]: Text(0, 0.5, 'Count')
```

Frequency of Company Subcategories
Count
70000
60000
50000
40000
30000
20000
10000
0
Non-govt company Union Govt company Subsidiary of Foreign company State Govt company Guarantee and Association comp
Company Subcategory



```
In [10]: sns.boxplot(data=df, palette='Set2')

         # Add a violin plot on top of the box plot for better visualization
         sns.violinplot(data=df, palette='Set3', inner=None)

         # Customize the plot (optional)
         plt.title('Comparison of Authorized Capital and Paid-up Capital')
         plt.ylabel('Capital Value (in Rupees)')

Out[10]: Text(0, 0.5, 'Capital Value (in Rupees)')
```



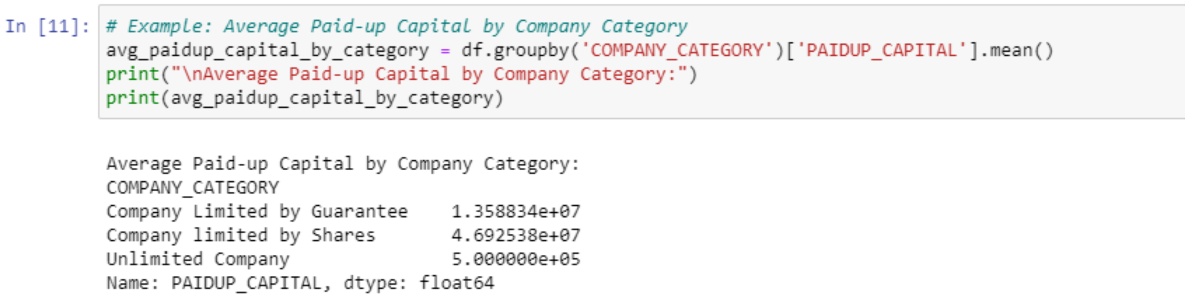
Step 4: Feature Engineering

This creative process involves crafting new features or transforming existing ones to maximise their relevance and predictive power. The data expertise intersect to extract valuable information that may be hidden within the data. Decide which variables or features to include in the PCA analysis. Carefully select the features that are most relevant for exploring and predicting company registration trends.

Grouping and Aggregation

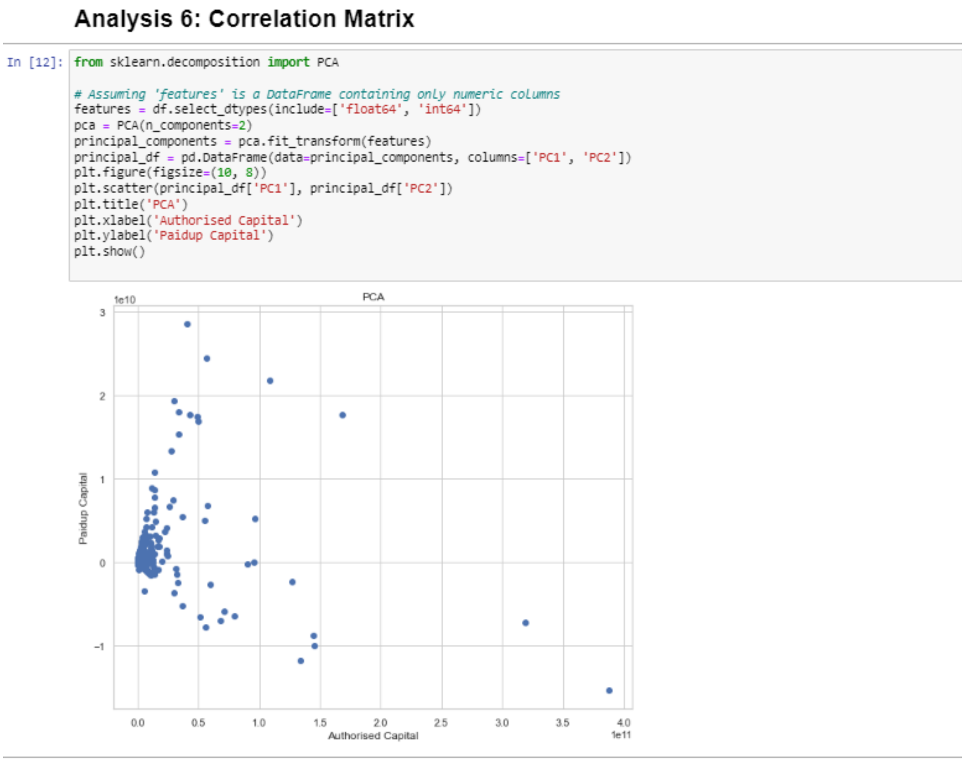
Descriptive statistics are essential for gaining a preliminary understanding of data when conducting AI-driven exploration and prediction of company registration trends with Registrar of Companies (RoC) data.

Analysis 5: Grouping and Aggregation



```
In [11]: # Example: Average Paid-up Capital by Company Category
         avg_paidup_capital_by_category = df.groupby('COMPANY_CATEGORY')['PAIDUP_CAPITAL'].mean()
         print("\nAverage Paid-up Capital by Company Category:")
         print(avg_paidup_capital_by_category)

         Average Paid-up Capital by Company Category:
         COMPANY_CATEGORY
         Company Limited by Guarantee    1.358834e+07
         Company limited by Shares       4.692538e+07
         Unlimited Company               5.000000e+05
         Name: PAIDUP_CAPITAL, dtype: float64
```

Correlation and Matrix



Analysis 6: Correlation Matrix

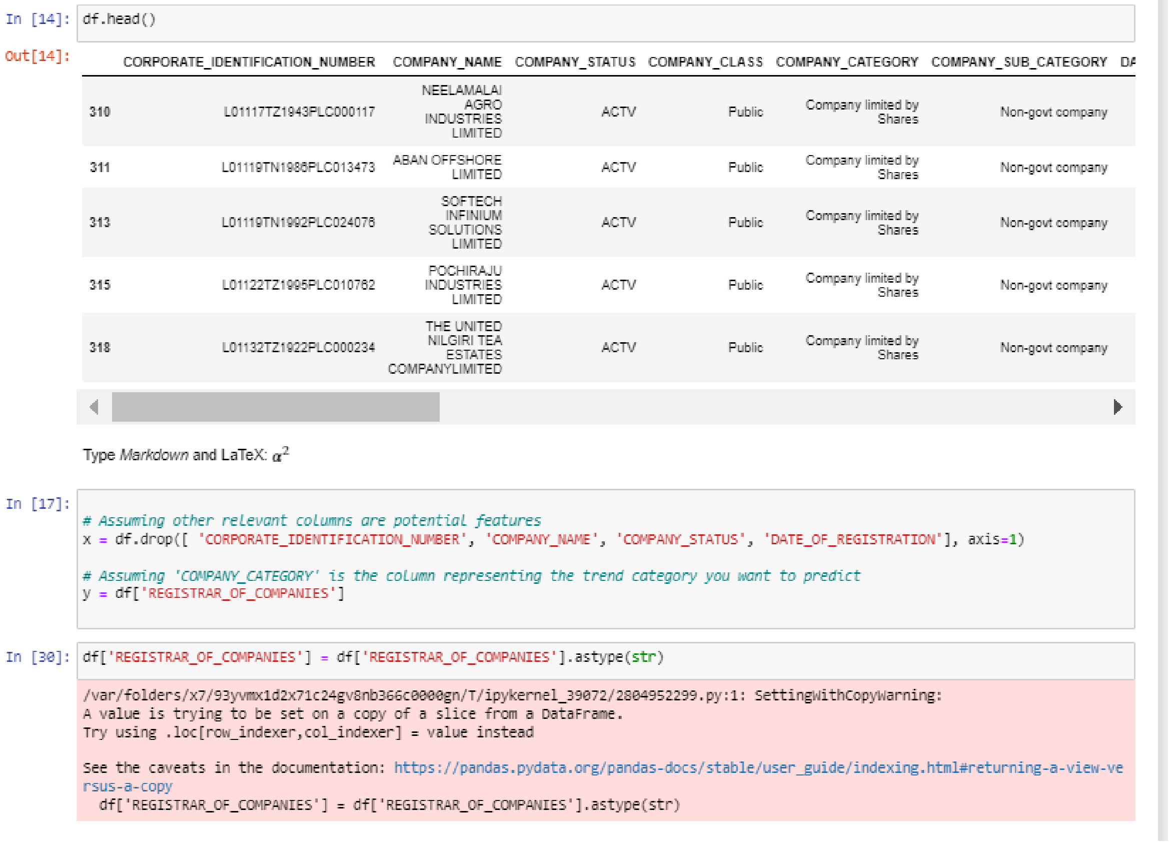
In [12]:
```
from sklearn.decomposition import PCA

# Assuming 'features' is a DataFrame containing only numeric columns
features = df.select_dtypes(include=['float64', 'int64'])
pca = PCA(n_components=2)
principal_components = pca.fit_transform(features)
principal_df = pd.DataFrame(data=principal_components, columns=['PC1', 'PC2'])
plt.figure(figsize=(10, 8))
plt.scatter(principal_df['PC1'], principal_df['PC2'])
plt.title('PCA')
plt.xlabel('Authorised Capital')
plt.ylabel('Paidup Capital')
plt.show()
```

Step 5: Predictive Modelling

- We apply cutting-edge AI algorithms, such as machine learning and deep learning techniques, to develop models that forecast future company registrations.
- The choice of algorithms, hyperparameter tuning, and model selection are crucial in this phase.
- We applied a machine learning model, specifically the Random Forest algorithm, to develop models for forecasting future company registrations.
- Hyperparameter tuning was performed to optimise model performance.

Principal Component Analysis (PCA)

- Dimensionality reduction technique commonly used in data analysis and machine learning to reduce the complexity of datasets while retaining as much valuable information as possible.
- When applying PCA to AI-driven exploration and prediction of company registration trends with Registrar of Companies (RoC) data.
- Data quality is essential for the success of PCA.
- Selection of features that are most relevant for exploring and predicting company registration trends play a vital role.



```
In [14]: df.head()
```

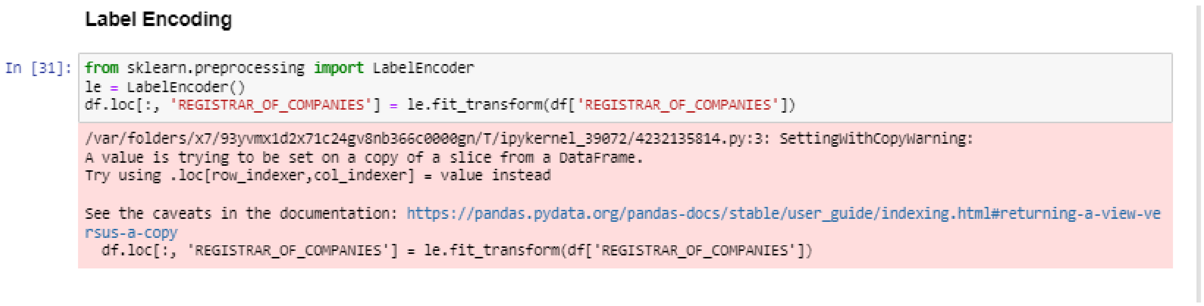
Out[14]:

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DA
310	L01117TZ1943PLC000117	NEELAMALAI AGRO INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
311	L01119TN1986PLC013473	ABAN OFFSHORE LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
313	L01119TN1992PLC024076	SOFTECH INFINIUM SOLUTIONS LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
315	L01122TZ1995PLC010762	POCHIRAJU INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
318	L01132TZ1922PLC000234	THE UNITED NILGIRI TEA ESTATES COMPANYLIMITED	ACTV	Public	Company limited by Shares	Non-govt company	

Type *Markdown* and LaTeX: α^2

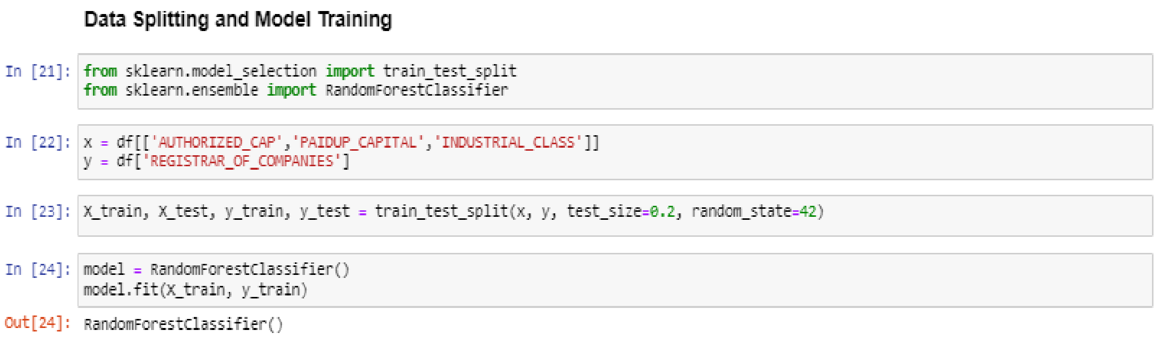
```
In [17]: # Assuming other relevant columns are potential features
         x = df.drop([ 'CORPORATE_IDENTIFICATION_NUMBER', 'COMPANY_NAME', 'COMPANY_STATUS', 'DATE_OF_REGISTRATION'], axis=1)

         # Assuming 'COMPANY_CATEGORY' is the column representing the trend category you want to predict
         y = df['REGISTRAR_OF_COMPANIES']
```

```
In [30]: df['REGISTRAR_OF_COMPANIES'] = df['REGISTRAR_OF_COMPANIES'].astype(str)

/var/folders/x7/93yvmx1d2x71c24gv8nb366c0000gn/T/ipykernel_39072/2804952299.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy
  df['REGISTRAR_OF_COMPANIES'] = df['REGISTRAR_OF_COMPANIES'].astype(str)
```



Label Encoding

```
In [31]: from sklearn.preprocessing import LabelEncoder
         le = LabelEncoder()
         df.loc[:, 'REGISTRAR_OF_COMPANIES'] = le.fit_transform(df['REGISTRAR_OF_COMPANIES'])

/var/folders/x7/93yvmx1d2x71c24gv8nb366c0000gn/T/ipykernel_39072/4232135814.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy
  df.loc[:, 'REGISTRAR_OF_COMPANIES'] = le.fit_transform(df['REGISTRAR_OF_COMPANIES'])
```

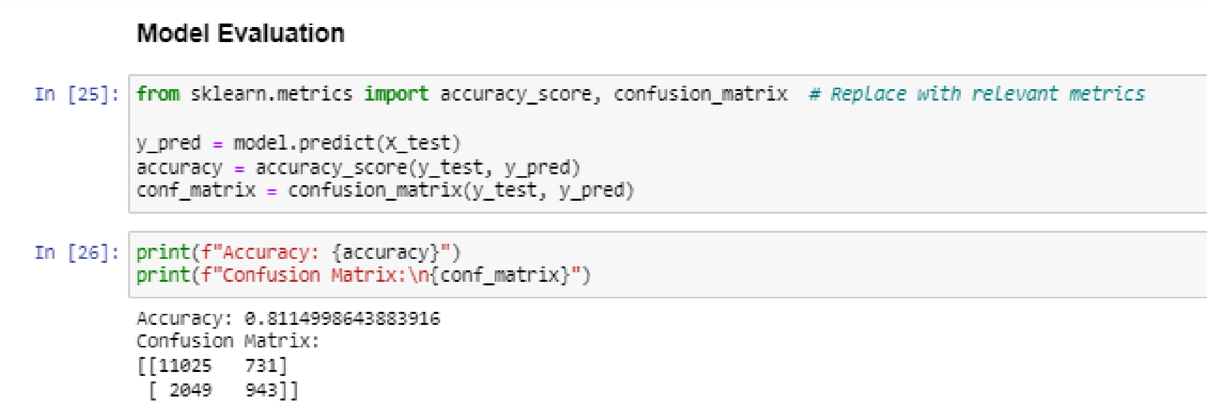


Data Splitting and Model Training

```
In [21]: from sklearn.model_selection import train_test_split
         from sklearn.ensemble import RandomForestClassifier

In [22]: x = df[['AUTHORIZED_CAP','PAIDUP_CAPITAL','INDUSTRIAL_CLASS']]
         y = df['REGISTRAR_OF_COMPANIES']

In [23]: X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

In [24]: model = RandomForestClassifier()
         model.fit(X_train, y_train)

Out[24]: RandomForestClassifier()
```

Step 6: Model Evaluation

To ensure the reliability and accuracy, We employ appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score, depending on the problem's nature. Hence, we fine-tune the models to deliver actionable predictions.

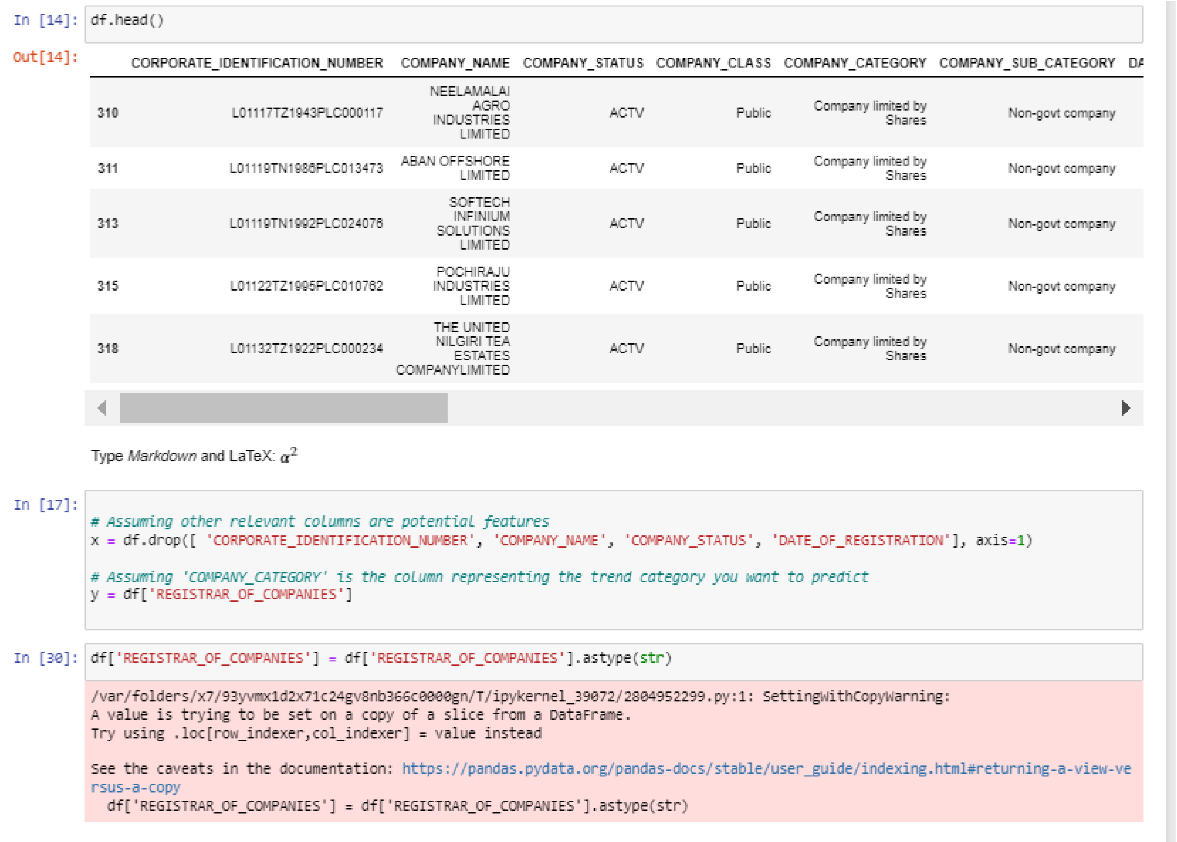


Model Evaluation

```
In [25]: from sklearn.metrics import accuracy_score, confusion_matrix  # Replace with relevant metrics

         y_pred = model.predict(X_test)
         accuracy = accuracy_score(y_test, y_pred)
         conf_matrix = confusion_matrix(y_test, y_pred)
```

```
In [26]: print(f"Accuracy: {accuracy}")
         print(f"Confusion Matrix:\n{conf_matrix}")
```

```
Accuracy: 0.8114998643883916
Confusion Matrix:
[[11025   731]
 [ 2049   943]]
```



```
In [14]: df.head()
```

Out[14]:

	CORPORATE_IDENTIFICATION_NUMBER	COMPANY_NAME	COMPANY_STATUS	COMPANY_CLASS	COMPANY_CATEGORY	COMPANY_SUB_CATEGORY	DA
310	L01117TZ1943PLC000117	NEELAMALAI AGRO INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
311	L01119TN1986PLC013473	ABAN OFFSHORE LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
313	L01119TN1992PLC024076	SOFTECH INFINIUM SOLUTIONS LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
315	L01122TZ1995PLC010762	POCHIRAJU INDUSTRIES LIMITED	ACTV	Public	Company limited by Shares	Non-govt company	
318	L01132TZ1922PLC000234	THE UNITED NILGIRI TEA ESTATES COMPANYLIMITED	ACTV	Public	Company limited by Shares	Non-govt company	

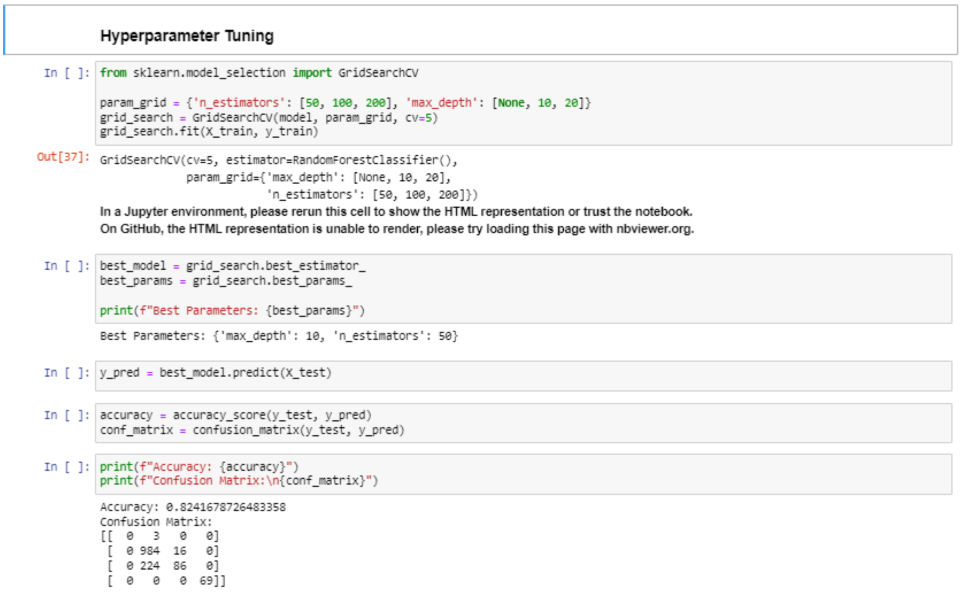
Type *Markdown* and LaTeX: α^2

```
In [17]: 
         # Assuming other relevant columns are potential features
         X = df.drop([ 'CORPORATE_IDENTIFICATION_NUMBER', 'COMPANY_NAME', 'COMPANY_STATUS', 'DATE_OF_REGISTRATION'], axis=1)

         # Assuming 'COMPANY_CATEGORY' is the column representing the trend category you want to predict
         y = df['REGISTRAR_OF_COMPANIES']
```

```
In [30]: df['REGISTRAR_OF_COMPANIES'] = df['REGISTRAR_OF_COMPANIES'].astype(str)
```

```
/var/folders/x7/93yvmx1d2x71c24gv8nb366c0000gn/T/ipykernel_39072/2804952299.py:1: SettingWithCopyWarning: 
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-ve
rsus-a-copy
  df['REGISTRAR_OF_COMPANIES'] = df['REGISTRAR_OF_COMPANIES'].astype(str)
```

Hyperparameter Tuning



Hyperparameter Tuning

```
In [ ]: from sklearn.model_selection import GridSearchCV

        param_grid = {'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20]}
        grid_search = GridSearchCV(model, param_grid, cv=5)
        grid_search.fit(X_train, y_train)

Out[37]: GridSearchCV(cv=5, estimator=RandomForestClassifier(),
                      param_grid={'max_depth': [None, 10, 20],
                                  'n_estimators': [50, 100, 200]})
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

```
In [ ]: best_model = grid_search.best_estimator_
        best_params = grid_search.best_params_

        print(f"Best Parameters: {best_params}")

        Best Parameters: {'max_depth': 10, 'n_estimators': 50}

In [ ]: y_pred = best_model.predict(X_test)

In [ ]: accuracy = accuracy_score(y_test, y_pred)
        conf_matrix = confusion_matrix(y_test, y_pred)

In [ ]: print(f"Accuracy: {accuracy}")
        print(f"Confusion Matrix:\n{conf_matrix}")

        Accuracy: 0.8241678726483358
        Confusion Matrix:
        [[  0   3   0   0]
         [  0 984  16   0]
         [  0 224  86   0]
         [  0   0   0  69]]
```

Conclusion:

This project successfully utilised AI-driven methods, with the Random Forest model, to explore and predict company registration trends using Registrar of Companies (RoC) data. The insights gained from EDA and the performance of predictive models provide valuable information for stakeholders to make informed decisions in the dynamic business landscape.

The project has the potential to empower businesses and policymakers to anticipate market shifts and take proactive measures, leading to informed investments and forward-thinking policies.
