**AI-DRIVEN EXPLORATION AND PREDICTION OF COMPANY REGISTRATION TRENDS WITH REGISTRAR OF COMPANIES (RoC)**

**EXPLORATORY DATA ANALYSIS (EDA)**

## Distribution of company registrations by industry, location, and size

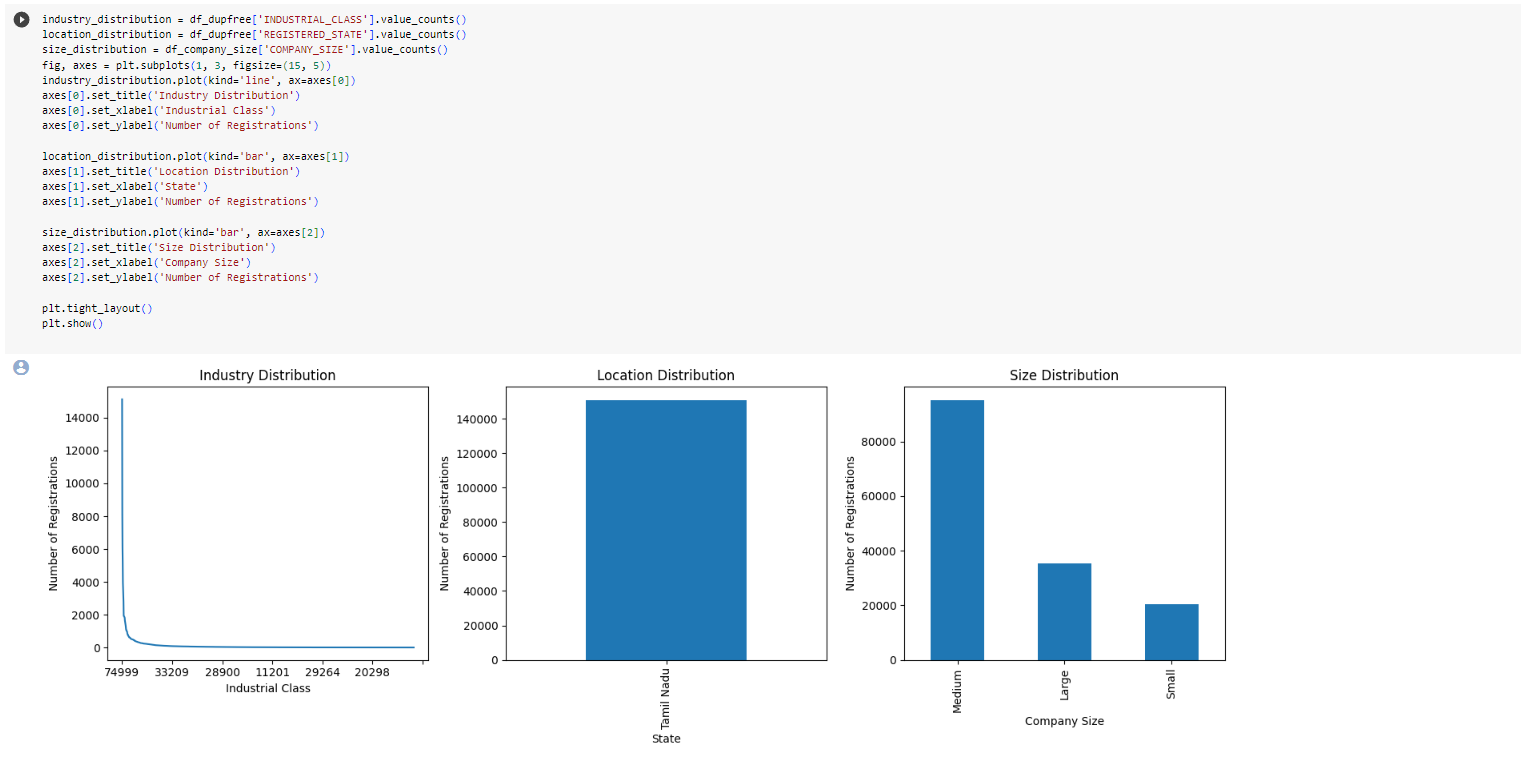


This segment of the code creates a new DataFrame df\_company\_size by extracting the first 150,870 entries for three columns: COMPANY\_NAME, AUTHORIZED\_CAP, and PAIDUP\_CAPITAL from the df\_dupfree DataFrame.

Companies are categorized into 'Small', 'Medium', and 'Large' based on their authorized capital and paid-up capital. Criteria for categorizing are defined in the size\_categories dictionary.

conditions contain boolean expressions for categorizing companies as 'Small' or 'Medium'.

Using np.select(), companies satisfying the conditions are labeled 'Small' or 'Medium'. Any company that doesn't match these conditions defaults to 'Large'.



The distribution (counts) for industry classes, registered states (locations), and company sizes are calculated.

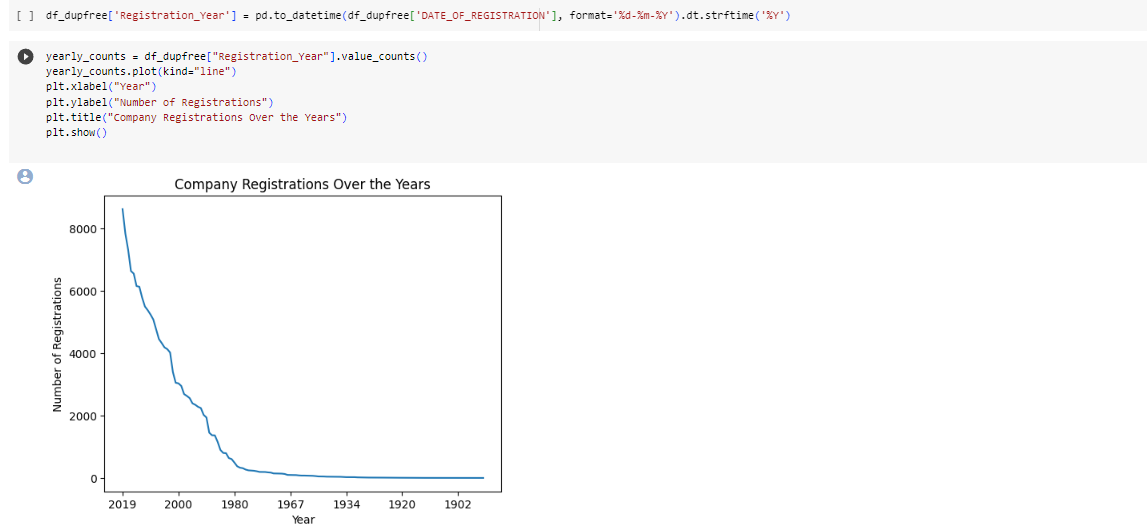
The visualization of the distributions is done using matplotlib.

plt.subplots(1, 3, figsize=(15, 5)): Initializes a single row of three subplots (charts).

plt.tight\_layout() adjusts the spacing between the plots for better aesthetics.

plt.show() displays the visualizations.

## Company Registrations Trends Over Time



pd.to\_datetime(...): The 'DATE\_OF\_REGISTRATION' column, which might be in a string format, is converted to a datetime object.

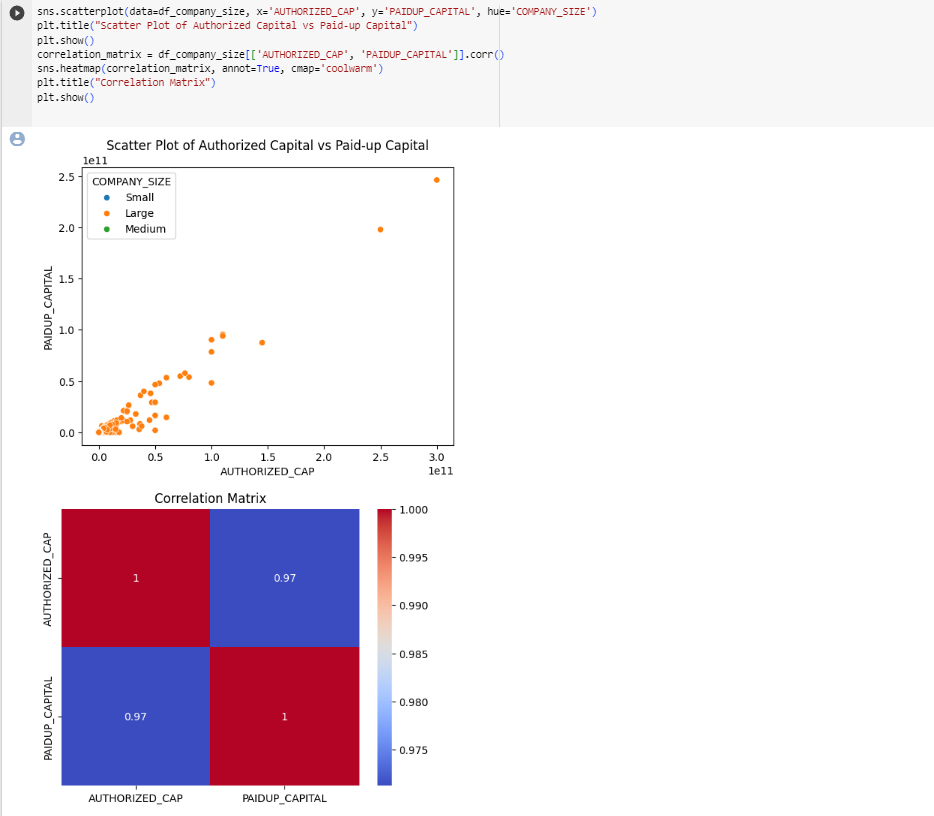
.dt.strftime('%Y'): Extracts only the year from the datetime object.

value\_counts(): Calculates the number of registrations for each year.

The plot displays the yearly registration trend using a line chart.

The x-axis represents the year, while the y-axis shows the number of registrations for that year.

Relationships between different variables, such as company size, authorized capital, and paid-up

****

This scatterplot illustrates the relationship between authorized capital and paid-up capital, while the colors (determined by the hue parameter) represent the size of the company.

Correlation matrix is also obtained which helps in investigating relationships between company financial metrics (authorized vs. paid-up capital) and how company size might affect this relationship.

The dataset used is df\_company\_size DataFrame, which is derived from df\_dupfree but only contains the first 150,870 rows and specifically the columns 'COMPANY\_NAME', 'AUTHORIZED\_CAP', and 'PAIDUP\_CAPITAL'. The reason for using this subset may be to focus on a specific segment of the data for more detailed analysis or due to computational efficiency.