SUMMARY OF EDA –YC COMPANY

DATA COLLECTION

In this part,it reads data from a JSON file containing information about companies, flattens the nested structure of the data, and then saves the flattened data into a CSV file. Here's a summary of the code and its functionality:

- Reading JSON Data: The code begins by importing the necessary libraries (json and pandas) and then reads data from a JSON file named company_data.json using the json.load() function.

- Flattening Data: The nested structure of the JSON data is flattened to make it more suitable for analysis. Each company's information is extracted and stored in a flattened dictionary format.

- Iterating Through Companies: The code iterates through each company in the JSON data and checks if all required keys exist. If a company's dictionary is not empty and contains all the necessary keys, its information is extracted and stored in a flattened format.

- Founder Information: Information about the founders of each company is also extracted and stored in the flattened structure. It includes the founder's name, role, biography, and LinkedIn profile.

- Creating DataFrame: The flattened data is converted into a pandas DataFrame for easier manipulation and analysis.

- Writing to CSV: Finally, the DataFrame is written to a CSV file named company_data.csv using the to_csv() method, with the index set to False to exclude the index column.

INITIAL EXPLORATION SUMMARY:

The dataset, comprising 4270 entries and 29 columns, offers a substantial pool of company data for analysis. Most columns appear to be fully populated, with non-null counts matching the total number of entries. However, some columns, notably Twitter Handle and Founder roles, exhibit notable amounts of missing data. All columns are currently stored as objects, suggesting a need for potential data type conversion, particularly for numerical attributes like Year Founded and Team Size. Insights into company characteristics can be gleaned from categorical variables such as Name, Tagline, and Location, while analysis of founder information across multiple columns provides an opportunity to understand the composition and backgrounds of founding teams. With the dataset consuming approximately 967.6 KB of memory, there's moderate memory usage to handle. However, addressing missing data and

optimizing data types may be necessary to streamline memory usage and facilitate further analysis. Overall, this initial exploration sets the stage for deeper investigation into company profiles, trends, and patterns within the dataset.

DATA CLEANING

The data cleaning process involved several key steps aimed at enhancing the quality and usability of the dataset. Here's a summary of the process along with important insights:

1. Identifying Duplicates and Missing Values:Initial checks revealed no duplicate rows (df.duplicated().sum()) and highlighted missing values across various columns (df.isnull().sum()). Notably, columns like Twitter Handle and Founder information exhibited a significant number of missing values.
2. Removing Unnecessary Columns:Certain columns deemed unnecessary for analysis, such as 'Industry Tags' and detailed founder information, were dropped from the DataFrame (df.drop()). This step aimed to streamline the dataset and focus on relevant attributes.
3. Handling Missing Values:Missing values in the 'Twitter Handle' column were replaced with a default value ('Not Available') using the fillna() function. This ensured consistency and completeness in the data, especially for companies without a Twitter presence.
4. Cleaning Numeric Columns:The 'Year Founded' and 'Team Size' columns were cleaned to extract only numerical values (str.extract()). Additionally, 'Year Founded' was converted to datetime format (pd.to_datetime()) for easier manipulation and analysis.

Insights:

- The absence of duplicate rows indicates the dataset's integrity, minimizing the need for deduplication efforts.
- Addressing missing values in critical columns like 'Twitter Handle' enhances the dataset's completeness and usability for subsequent analysis.
- Cleaning numeric columns like 'Year Founded' and 'Team Size' enables numerical analysis and temporal comparisons, facilitating deeper insights into company characteristics and trends.

Overall, the data cleaning process focused on ensuring data quality, completeness, and consistency, laying a solid foundation for further exploration and analysis of the company dataset. These steps not only improved the dataset's usability but also set the stage for more robust and insightful analyses of company attributes and trends.

VISUALIZATIONS

1. Frequency of Companies in Batches (Countplot):The countplot visualizes the frequency of companies in different batches. I observe that Batch w22 AND S21 has

the highest number of companies, indicating a potential trend in the distribution of companies across batches.

2. Distribution of Company Types (Pie Chart):The pie chart illustrates the distribution of company types. Here,Active constitute the majority of companies, representing 71.7% of the dataset, while other company types like public, private and inactive make up the remaining percentage.
3. Distribution of Year Founded (Histogram):The histogram displays the distribution of the founding years of companies. Here, the dataset primarily comprises companies founded in recent years, with a peak around 2018-2022, suggesting a concentration of newer companies.
4. Scatter Plot for Year Founded vs Team Size: The scatter plot illustrates the relationship between the year founded and the team size of companies.This suggests that the founding year does not significantly influence the size of a company's team. The scattered distribution of data points across the plot confirms the lack of a meaningful correlation.
5. Box Plot for Company Type vs Team Size:The box plot compares the team sizes across different company types. Here it indicate that companies in the 'Public' category tend to have larger or more defined team sizes compared to others. This could imply differences in organizational structure, resources, or business models between publicly traded companies and other types of companies.
6. Correlation Analysis (Correlation Matrix and Heatmap): The correlation analysis reveals a very weak positive correlation coefficient of 0.038 between the year founded and team size. This indicates that there is almost no linear relationship between these two variables. In other words, the year a company was founded has minimal impact on the size of its team. This insight suggests that factors other than founding year may play a more significant role in determining the team size of companies. Further exploration or analysis may be warranted to uncover these additional factors influencing team size.