DATASET AND ANALYSIS OVERVIEW

I. Description and Source

Dataset: Forbes 2000 Global Companies

Link:
https://www.kaggle.com/datasets/rakkesharv/forbes-2000-global-companies/data

Source: This dataset was sourced from kaggle.com. According to Kaggle, this dataset was created by Rakkesh Aravind G via web scraping around 1 year ago. The author claims the data is scraped from Forbes.

Description: This dataset contains information of the top 2000 companies internationally, ranked by Forbes using four metrics: sales, profits, assets, and market value. These top 2000 companies account for $47.6 trillion in revenues, $5.0 trillion in profits, $233.7 trillion in assets and $76.5 trillion in market cap. This data is retrieved from the Forbes' article titled: The Global 2000 of 2022 (https://www.forbes.com/sites/forbesstaff/2022/05/12/forbes-global-2000-list-2022-the-top-200/). This dataset contains information about the location and industries of these companies, as well as other details about their profitability and yearly revenue.

II. Why We Chose This Dataset

We chose this dataset because it can help us analyze the global business landscape, identify trends and patterns in the corporate world. This is directly applicable to our current situation, as we will soon be joining the corporate world. This dataset is also fairly recent, so it can be used to approximate the current standings of these companies. From a technical point of view, there are nearly 2000 rows in this dataset, with minimal entries missing data or information. This ensures that during data processing, samples can be taken with assurance of their randomness, to present an unbiased analysis.

Overall, this dataset is robust and provides a comprehensive amount of information in a well-summarized and descriptive manner, regarding international businesses and their financial standings.

III. Dataset Statistics

This data set has 1999 rows. This dataset contains 11 columns:

- *2022 Ranking*: int
- *Organization Name*: string
- *Industry*: string.

- *Country*: string
- *Year* Founded: int
- *CEO*: string
- *Revenue (Billions):* float
- *Profits (Billions):* float
- *Assets (Billions):* float
- *Market Value (Billions):* float
- *Total Employees*: float (though the number of employees can only be a whole number, this is designated as a float in the dataset because whole number values are stored as 125.0)

According to the description provided, this dataset contains no columns with missing or corrupted values, the exception being the 'Total Employees' row which contains 0s for values not made publicly available by the companies or Forbes. Around 1% of the rows do not have any CEO data. These fields are then labeled "No Data." Approximately, 96 rows do not have a year in the 'Year Founded' row. In place of a year, there is a 0. Overall, this dataset contains minimal entries that have default values.

IV. Analysis Goals and Hypothesis

Using this dataset, we hope to develop a thorough financial analysis of the top 2000 companies in the year 2022. We hope to analyze market trends in different industries and companies, as well as trends in annual revenue, profit, and much more. Our primary hypotheses are centered around determining the leading industries and leading companies through the creation of our own metric and point allocating system (Our primary hypothesis can be found explicitly listed in 'CS210 Final Project.ipynb'). To support this analysis we will perform various types of general and advanced analysis. Overall, we hope to paint a picture of the financial condition of the economy in the year 2022, with the help of this dataset, representing the performance of the world's top 2000 companies.