Data Cleaning

We started with the idea of analyzing companies using ranking magazines such as Fortune and Forbes. We found Forbes Global 2000 dataset for years from 2017 to 2022. The first task was to import each dataset and merge them into a single csv file. The problem during merging was that in some files exact figures were reported for assets or market value whereas in some files they were reported in millions or billions. For example, assets were reported as 20,000,000 in some files and 20 M in others. We had to detect this number, remove the commas, and convert the figures to a common unit. We accomplished this using a set of three functions. Thus, all the figures were converted to billions for the sake of consistency and ease of analysis and visualization.

We also thought of analyzing the stock prices of these companies to help in making better investment decisions. Since these companies were all listed on different stock exchanges and getting data from all these stock exchanges would have been difficult, we decided to just consider the companies listed on NYSE, the largest stock exchange in the world. We found a dataset having stock prices and fundamentals of different companies listed on NYSE. But later, found that the timeline of this dataset was from 2010 to 2016 whereas our Forbes data was from 2017 to 2022. Using these two datasets for the same analysis would have been wrong. Therefore, we found another dataset which had stock prices of each company from the 90s to the present. This dataset had a separate csv for each company. We had to merge these csv files into a single file for ease of analysis. The problem was that these files were named using the ticker symbols of the companies and our Forbes dataset had no column for tickers. We had to figure out a way to merge these datasets.

This is where the previous dataset having stock prices up to 2016 came to the rescue. This dataset had a file named securities.csv which had the company name as well as the tickers. Therefore, the securities file was first merged with the consolidated Forbes dataset. The problem was there were differences in the way names were present in these two files. For example, Walmart was reported as Walmart Inc. We had to figure out a way beyond a simple merger which only compared string for exact matches. For this, we compared each company name with every other company and calculated the Levenshtein distance which is the number of changes required to be made to the first string to get to the second string. After this we kept the company pairing with the least Levenshtein distance considering these companies are similar. Still there was further cleaning required to ensure consistency of names and ease of merging in Tableau. We hard coded a list of common words in names of companies which are not important such as corp., corporation, inc., investments, bank, etc. After removing these words, we got around 350 companies in the dataset.

But we also had to merge these companies with the NYSE stock prices dataset. We did that using the ticker symbol column. It was not very difficult except for some hiccups here and there. At the end we got around 150 companies for our visualization in Tableau. All the files were exported to csv. At the end we had 4 files to be imported in Tableau. First was the Forbes Global 2000 dataset from 2017-2022 which had the company names and their assets, profits, sales, and market value for every year. The second dataset was the securities dataset which had the company name and ticker. This dataset was not going to be used but it had some valuable information in the form of company headquarters which has been used for creating a geomap visualization. The third file was the NYSE stock prices for all the ~150 companies from 2017 to first few months of 2023. The last file had the company name from the Forbes dataset, its representative name in the securities file and the respective ticker symbol. This file is used to merge all the other files.

Stock Analysis Dashboard

The first dashboard can be used for analysis the fundamentals and historical stock performance of a company. The fundamental analysis can be done using the top left visualization which shows assets, profits, sales, and market value of the company for all the years for which the data is available. The top right visualization is a line chart of the closing price for each day with the moving average. The bottom visualization can be used for technical analysis. This worksheet displays the open, high, low, and close of the stock prices for each company for each day in the form of candlesticks. A filter is also provided using which can be used to only analyze the required timeframe. This is a relative date filter such that we can see the data from a time in the past till today. All the visualizations have a common filter of ticker symbol. If you know the ticker symbol of a company, you can search for it and see if the company will be a good investment.