Project Milestone 1
DSC540-T301
Joshua Greenert
9/24/2022

3 Data Sources

- **Website - NIR_Datasets_2021**
  - https://netimpactreport.com/datasets/largest-500#data
  - This dataset contains data over the impact produced by major companies all over the world
- **Flat File - Fortune 1000 Companies by Revenue**
  - https://www.kaggle.com/datasets/surajjha101/fortune-top-1000-companies-by-revenue-2022
  - This dataset contains information over the top 1000 companies based on revenue for 2022
- **API - The Muse**
  - https://www.themuse.com/developers/api/v2?ref=apilist.fun
  - This API contains information regarding companies that are hiring along with job listings for positions from entry to senior level

The Relationship

The relationship between all datasets will be the names of each respective company. Some minor data cleaning/preparation will be required to convert all of the strings amongst each dataset into one solitary group of data. The website data will contain company names that can be compared after pulling in the tables available on the data page. Likewise, the Muse API can be used to pull in data that matches the currently established list of names that would be formed amongst the data contained within the website (NIR Datasets) and the flat file (fortune 1000 companies by revenue).

What Is Required?

In order to properly connect the data between each dataset, some transformations are going to be essential. There will need to be a connection created to the API that will pull data through based on the company name; in this way, we can segment the data produced by the API using the appropriate endpoint. The company names aforementioned will be the product of the data scraped from the Net Profit Report (NIR) data matched against the Fortune 1000 data to ensure a comprehensive list. Anytime that web scraping is involved, there will be immense cleaning that will need to be performed for the data to fully match up. Likewise, the API data will need to be cleaned as well to remove any HTML elements included through the use of stripping libraries like Beautiful Soup.

Data from each dataset will be stored in a Postgresql database to be extracted and manipulated for visualizations. Therefore, PostgreSQL will need to be installed and set up to properly handle these processes that will be required at a later date. Python libraries including Matplotlib, numpy, and pandas will be used to properly establish the connections within the data; moreover, the psycopg2 library appears to have the required functions to connect to the database — once established — and collect the data needed to be joined and compared for visualizations. The use of these libraries in accordance with Jupyter Notebooks will be sufficient to provide clear and impactful graphs and charts that should clearly represent the connections made within the data.