MSDS-694-01: Group Project Task 1 - Select Data Sets

Group name: “the best group”
Group Members:

- Boston Bautista
- Chris Dhong
- Ignacio Gomez
- Paul Ng
- Tianyi Luo
- Niki Naderzad

Dataset: SEC Financial Statement Data Sets from 2020
Link: https://www.sec.gov/data-research/sec-markets-data/financial-statement-data-sets
Example file: https://www.sec.gov/files/dera/data/financial-statement-data-sets/2020q1.zip
Data catalog: https://www.sec.gov/files/financial-statement-data-sets.pdf

We decided to use the SEC Financial Statement Data Sets from 2020 for our project. The dataset contains financial statements from public companies, which can be used to understand the impact of the COVID-19 pandemic on the financial market.

The SEC offers a comprehensive collection of financial data, allowing us to extend the analysis to multiple quarters and years. This dataset is ideal for our project as it provides a rich source of information to analyze the financial performance of companies during the pandemic. Thanks to the nature of the data, we can validate our results against major financial news sources (Bloomberg) and trends during that period. This dataset also allows us to familiarize ourselves with officially reported financial data, the same data used by professionals in the finance industry.

Finally, the SEC maintains a well documented data catalog (linked above) that we can use to better understand each feature from the data.
The SEC releases the data per quarter in 4 separate files (documented in data catalog):

- Submissions
- Tags
- Numbers
- Presentation

We can create 4 RDDs in spark to answer the questions in the next assignment, or denormalize the data to have all this data as a single object.

We will merge 4 quarters (Q1, Q2, Q3, Q4) of the year 2020 to create a comprehensive dataset for our analysis. The data is available as a tab separated file that can be converted to a csv file for easier manipulation and analysis.

As a team, we can propose the following questions:

Member	**Questions**	**EDA**
Boston Bautista	Which industries show the most consistent profitability, and which experience the highest volatility?	Use NetIncomeLoss, Revenues, and Assets, as financial indicators to measure profitability Calculate net profit margin and return on assets, using the financial indicators above, for each company per quarter from 2015 to 2024 Use the SIC code in sub.txt to group companies into industries Calculate average profitability and standard deviation for each industry
Chris Dhong	How many companies had IPOs per quarter?	Compare all reporting companies per quarter and identify first reports or new companies.
Ignacio Gomez	What are the top 5 industries with the best return over assets per quarter	Create column return over assets per quarter.
Paul Ng	Which quarter had the best earnings, in USD and in number of profitable companies	Create column is_profitable for each combination of company/quarter as 0 or 1 group by quarter and count.
Tianyi Luo	Which industries rely most on accruals, and do those industries show more volatile earnings?	For every company, find the accrual ratio, create the column Accruals = NetIncome – OperatingCashFlow) / TotalAssets. Find the average accrual ratio per industry.
Niki Naderzad	How has the average corporate leverage ratio (total liabilities / total assets) changed since 2010 across different industries? Do companies with strong operating cashflow also show higher profitability, and does	Compute leverage ratio for each company and observe trends using spark. Group data by industry and year to identify which sector maintains the highest debt ratio. Summarize average and variance of leverage across time to infer risk exposure at the industry level

	this change for different industries?	Compute correlation between operating cash flow and net income. Examine how this relationship differs in various industries to see if there are some of them that don't have correlation between profitability and cash flow. Identify outliers in both profitability and cash flow.

Datasets proposed by team members:

Member	**Dataset**	**Size**	**Why**
Boston Bautista	Yelp User Open Dataset	3.13 GB 1,987,897 rows	It provides a detailed view into real online communities and patterns of user engagement.
Chris Dhong	Motor Vehicle Collision Data	400 MB 2,216,649	Explains crash details. It can be used by insurance companies. We can identify the most dangerous intersections.
Ignacio Gomez	SEC data for year 2020	1.9 GB 12,000,000	We can study how the financial markets reacted during the early months of the COVID-19 pandemic.
Paul Ng	U.S. Geological Survey API for Rivers and Groundwater Levels	<scraped from API>	Water plays an important role in our daily lives and to better understand this resource, USGS has a significant number of stations (monitoring sites) throughout the country — usually on a river, creek, canal, reservoir, groundwater well — where instruments record

			water data at a set interval (~15 minutes/hourly/daily). The data can be used to help identify real-time flooding risks and predict drought throughout the U.S.
Tianyi Luo	Amazon Reviews'23	1.2 GB 5,000,000 rows	True ecommerce dataset with rich user text and ratings. Good for business insight on customer satisfaction, price, and reviews.
Niki Naderzad	Financial News and Stock Price	21 GB 36,000,000 rows	We can understand the impact of news in financial markets.