Analysis and Visualization of S&P 500 Companies and Index through ETL Pipeline

This project document provides a comprehensive analysis of the S&P500 companies by combining sectoral performance and historical market behavior to uncover key trends and insights. It identifies high-growth sectors based on market capitalization, revenue growth, and profitability, analyzing attributes like EBITDA, employee count, and sector-wise averages. Simultaneously, it explores historical S&P 500 index data, examining daily fluctuations, trading volumes, and market volatility through exploratory data analysis (EDA). The dual perspective equips investors and financial strategists with data-driven insights to support informed decisions on investments, portfolio management, and risk assessment in the U.S. stock market.

Question

What sectors demonstrate the highest market capitalization and revenue growth in the S&P 500, and how do they compare in terms of EBITDA and employee count?

1st Data Source Description	**Data URL and Type**	**Source and License**	**About License**
The S&P500, a key financial benchmark, tracks 500 major U.S. companies. As of December 31, 2020, over $5.4 trillion was invested in assets linked to it (Source). Despite its name, the index includes 505 stocks due to multiple share classes, like Alphabet's Class A (GOOGL) and Class C (GOOG).	Metadata URL: Link Data URL: Link to Data Data Type: CSV	Data Source: The dataset is taken from Kaggle under the CC0: Public Domain License. Data License: CC0: Public Domain	The author of this work has waived all copyright and related rights worldwide, dedicating it to the public domain. You are free to copy, modify, distribute, and use it, even commercially, without seeking permission. More Information: Link Here
2nd Data Source Description	**Data URL and Type**	**Source and License**	About License
The S&P500 index is a stock market index that tracks the performance of 500 major U.S. companies across various industries. It is widely used as a benchmark for the overall health of the stock market and the economy. Managed by S&P Dow Jones Indices, the index includes leading companies like Apple, Microsoft, and Tesla, with its components selected by a committee.	Metadata URL: Link Data URL: Link to Data Data Type: CSV	Data Source: The dataset is taken from Kaggle under the CC0: Public Domain License. Data License: CC0: Public Domain	The author of this work has waived all copyright and related rights worldwide, dedicating it to the public domain. You are free to copy, modify, distribute, and use it, even commercially, without seeking permission. More Information: Link Here

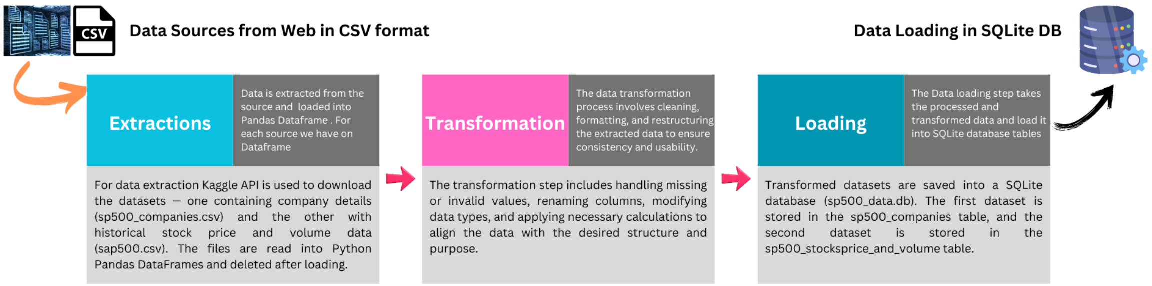
Why? These datasets I chose provide sector classifications, market capitalization, revenue, EBITDA, and employee count, enabling a comprehensive analysis of growth, profitability, and workforce metrics across S&P 500 Companies. Simultaneously, they explore historical S&P 500 index data, examining daily fluctuations, trading volumes, and market volatility.

Methodology Overview: The datasets, to address the main question, were extracted from Kaggle using the API, transformed for consistency, stored in SQLite, and analyzed through exploratory data analysis (EDA).

Importance: Equips investors and financial strategists with data-driven insights to support informed decisions on investments, portfolio management, and risk assessment in the U.S. stock market.

ETL Data Pipeline

Section	Description
High-Level Overview: The ETL process involves extracting the data, transforming it by cleaning and preprocess it for analysis, and loading it into a SQLite database for querying. For data extraction Kaggle API is used to download the datasets — one containing company details (sp500_companies.csv) and the other with historical stock price and volume data (sap500.csv). The files are read into Python Pandas DataFrames and deleted after loading.	The data pipeline is an ETL (Extract, Transform, Load) process for handling two datasets related to S&P 500 Companies and Index. **Technologies**: - **Python**: Programming language for scripting the ETL process. - **Pandas**: Library for data manipulation and transformations. - **SQLite**: Database for storing the processed data. - **Kaggle API**: For downloading datasets directly from Kaggle. - **Logging**: For tracking the ETL process and debugging issues. - **VS Code:** IDE for writing, debugging, and running Python code. - **Jupyter Notebook:** An interactive environment to write and test Python code for data analysis and visualization.
Transformation and Preprocessing: The data transformation process involves cleaning, formatting the extracted data to ensure consistency and readability. This includes handling missing or invalid values, renaming columns, modifying data types, and applying necessary calculations to align the data with the desired structure and purpose.	**First Dataset (S&P 500 Companies):** - Converted 'Fulltimeemployees' column from decimal to integers. - Transform 'Marketcap' and 'Ebitda' columns to billions and renamed them for better readability. - Rounded the 'Weight' column to two decimal places for precision. **Second Dataset (S&P 500 Volume and Prices):** - Converted 'Date' column to datetime format. - Dropped rows where Volume column has value 0. - Renamed all columns for consistency and contextual information for better readability.
Execution, Meta-Quality Measures, and Error Handling: The loaded data is analyzed to perform exploratory data analytics to address our main question and build visualization to help us identify patterns, segmentation and correlations. The entire ETL process, including extraction, transformation, and loading, is logged for error tracking and completion confirmation. The pipeline ensures high data quality through error handling, logging.	**Error Handling:** - Used try-except blocks and detailed logging. **Data Validation:** - Ensured consistent data types and column names. **Pipeline Resilience:** - Automated reprocessing of raw data, compatibility with SQLite.
Issues Encountered Data Access Issues Data Quality Issues Temporary File Management: Logging and Debugging Error Handling	**Corresponding Solution** - Checks file existence and logs download errors. - Dataset 1: Handles invalid 'Fulltimeemployees' column type, scales 'Marketcap' and 'Ebitda' to billions, and renames columns. - Dataset 2: Converts Date to datetime, removes rows with zero Volume, and renames columns for consistency. - Deletes files after loading into DataFrame. - Detailed logging at all stages. Exceptions are logged and propagated to prevent failures.



Data Sources from Web in CSV format

Data Loading in SQLite DB

Extractions

Data is extracted from the source and loaded into Pandas Dataframe . For each source we have on Dataframe

For data extraction Kaggle API is used to download the datasets — one containing company details (sp500_companies.csv) and the other with historical stock price and volume data (sap500.csv). The files are read into Python Pandas DataFrames and deleted after loading.

Transformation

The data transformation process involves cleaning, formatting, and restructuring the extracted data to ensure consistency and usability.

The transformation step includes handling missing or invalid values, renaming columns, modifying data types, and applying necessary calculations to align the data with the desired structure and purpose.

Loading

The Data loading step takes the processed and transformed data and load it into SQLite database tables

Transformed datasets are saved into a SQLite database (sp500_data.db). The first dataset is stored in the sp500_companies table, and the second dataset is stored in the sp500_stocksprice_and_volume table.

Output Results and Limitations

On execution the pipeline produces two tables in the SQLite database with S&500 Companies (sp500_companies) and Index (sp500_stocksprice_and_volume) as shown in the tables below (Few rows). Both the tables are processed and cleaned that can used for exploratory data analysis. The details analysis are performed using difference visuals/graphs such as Distribution of Companies Across Sectors, Top 5 Companies by Market Capitalisation, Revenue Growth by Sector, Relationship Between Market Cap and EBITDA, Top Sectors in the S&P 500 by Market Capitalisation and Revenue Growth and how do they compare in terms of EBITDA and employee count. The analytics will equip investors and financial strategists with data-driven insights to support informed decisions on investments, portfolio management, and risk assessment in the U.S. stock market. The detailed exploratory data analytics can be found at Link here.

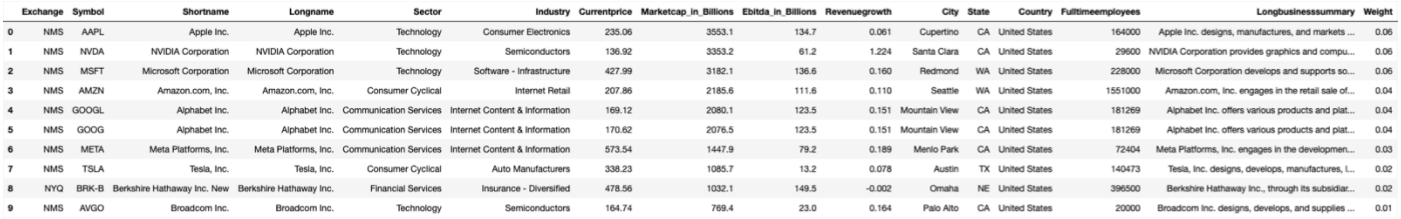
Output Data

The output data is loaded into tables in SQLite database. SQLite database was chosen because it is lightweight, easy to set up, and requires no complex server configuration, making it ideal for small-scale ETL processes. Its simplicity and compatibility with Python allow seamless data storage and retrieval during pipeline execution.

Potential Issues

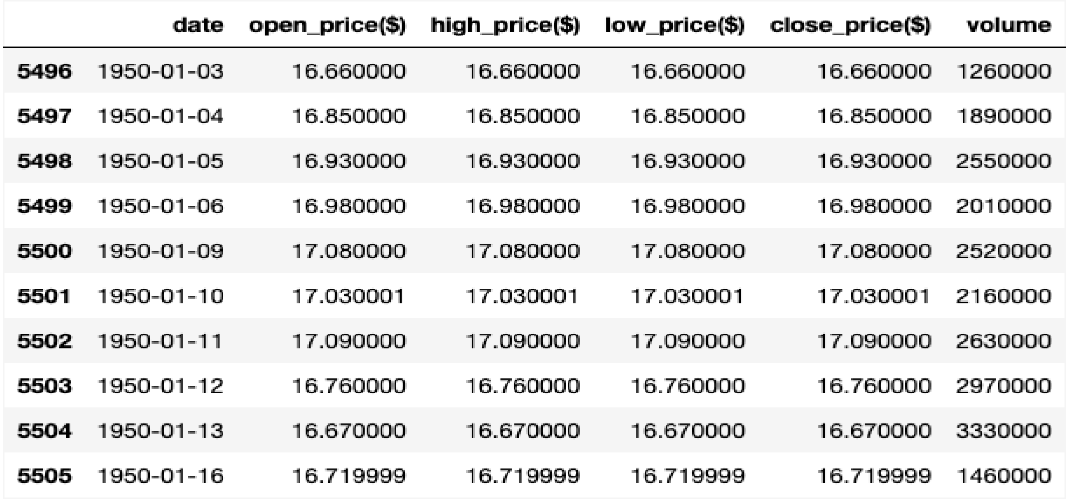
- The ETL process depends on the accuracy and reliability of the source data.
- Adaptations must be implemented to handle potential changes in data structure or schema over time.
- Challenges related to data consistency and completeness may arise due to variations in the data source.

Table1 (S&P500 Companies Data)



	Exchange	Symbol	Shortname	Longname	Sector	Industry	Currentprice	Marketcap_in_Billions	Ebitda_in_Billions	Revenuegrowth	City	State	Country	Fulltimeemployees	Longbusinesssummary	Weight
0	NMS	AAPL	Apple Inc.	Apple Inc.	Technology	Consumer Electronics	235.06	3553.1	134.7	0.061	Cupertino	CA	United States	164000	Apple Inc. designs, manufactures, and markets ...	0.06
1	NMS	NVDA	NVIDIA Corporation	NVIDIA Corporation	Technology	Semiconductors	136.92	3353.2	61.2	1.224	Santa Clara	CA	United States	29600	NVIDIA Corporation provides graphics and compu...	0.06
2	NMS	MSFT	Microsoft Corporation	Microsoft Corporation	Technology	Software - Infrastructure	427.99	3182.1	136.6	0.160	Redmond	WA	United States	228000	Microsoft Corporation develops and supports so...	0.06
3	NMS	AMZN	Amazon.com, Inc.	Amazon.com, Inc.	Consumer Cyclical	Internet Retail	207.86	2185.6	111.6	0.110	Seattle	WA	United States	1551000	Amazon.com, Inc. engages in the retail sale of...	0.04
4	NMS	GOOGL	Alphabet Inc.	Alphabet Inc.	Communication Services	Internet Content & Information	169.12	2080.1	123.5	0.151	Mountain View	CA	United States	181269	Alphabet Inc. offers various products and plat...	0.04
5	NMS	GOOG	Alphabet Inc.	Alphabet Inc.	Communication Services	Internet Content & Information	170.62	2076.5	123.5	0.151	Mountain View	CA	United States	181269	Alphabet Inc. offers various products and plat...	0.04
6	NMS	META	Meta Platforms, Inc.	Meta Platforms, Inc.	Communication Services	Internet Content & Information	573.54	1447.9	79.2	0.189	Menlo Park	CA	United States	72404	Meta Platforms, Inc. engages in the developmen...	0.03
7	NMS	TSLA	Tesla, Inc.	Tesla, Inc.	Consumer Cyclical	Auto Manufacturers	338.23	1085.7	13.2	0.078	Austin	TX	United States	140473	Tesla, Inc. designs, develops, manufactures, l...	0.02
8	NYQ	BRK-B	Berkshire Hathaway Inc. New	Berkshire Hathaway Inc.	Financial Services	Insurance - Diversified	478.56	1032.1	149.5	-0.002	Omaha	NE	United States	396500	Berkshire Hathaway Inc., through its subsidiar...	0.02
9	NMS	AVGO	Broadcom Inc.	Broadcom Inc.	Technology	Semiconductors	164.74	769.4	23.0	0.164	Palo Alto	CA	United States	20000	Broadcom Inc. designs, develops, and supplies ...	0.01

Table2 (S&P500 Index Data)



	date	open_price($)	high_price($)	low_price($)	close_price($)	volume
5496	1950-01-03	16.660000	16.660000	16.660000	16.660000	1260000
5497	1950-01-04	16.850000	16.850000	16.850000	16.850000	1890000
5498	1950-01-05	16.930000	16.930000	16.930000	16.930000	2550000
5499	1950-01-06	16.980000	16.980000	16.980000	16.980000	2010000
5500	1950-01-09	17.080000	17.080000	17.080000	17.080000	2520000
5501	1950-01-10	17.030001	17.030001	17.030001	17.030001	2160000
5502	1950-01-11	17.090000	17.090000	17.090000	17.090000	2630000
5503	1950-01-12	16.760000	16.760000	16.760000	16.760000	2970000
5504	1950-01-13	16.670000	16.670000	16.670000	16.670000	3330000
5505	1950-01-16	16.719999	16.719999	16.719999	16.719999	1460000