**Milestone 5: Stock market Data trends prediction, Merging and Visualization**

**Summary:**

Out project is to use various stock market dataset and find the Stock market performance and various trends across the various sectors to identify the sector wise patterns and performance metrics and its key measures. We have used the dataset from Kaggle , Yahoo Finance Web scraping, Yahoo Finance Query API and the Finance Modeling pre API. We have merged these dataset and created 5 different visualizations. Our cleaned and merged datasets are stored in the SQLite database which can be useful for the future analysis.

We have followed the following approach:

1. **Data Extraction:** Our dataset was sourced from the multiple platforms like Kaggle, Yahoo Finance, Financing modeling prep API
2. **Data cleansing**: I have extracted the dataset and cleaned it by removing the duplicate dataset and the missing datasets are handled properly. I have also formatted the data into a readable format and standardized the column name, headers
3. **Data Merging:** I have merged the dataset extracted from Kaggle, Yahoo Finacle Web scrapping, Yahoo Finance query API, Finance modeling prep API. As we know that the sector information was missing from Kaggle but we mapped the sector from other dataset and then merged the dataset to identify the key performance information for the sector.
4. **Data Storage:** Our final cleaned merged data set is stored into the SQLite database for the future analysis and implementation.
5. **Data Visualization**: As per the requirement in milestone five, I have included the five different visualizations.

With all these steps followed we have successfully determined the sector wise performance of the stock market trends based on the historical stock price with the available datasets.

**Changes made to the dataset:**

We have made several changes to the dataset to clean and standardizing the data.

1. Renamed column with the standard naming to follow the pattern. For example open is renamed to Open
2. I have formatted the data column with the datetime format
3. Replaced the missing value with N/A wherever it is applicable
4. To retain the consistency I have removed the duplicate values
5. Ticker name is standardized by renaming it all to uppercase letter.
6. Web scraping using yahoo website is done to get the sector wise data from yahoo finance.
7. I have merged the additional valuable column that was provided by the yahoo finance website such as Year to data, daily return
8. Converted the percentage-based values into the consistent format.
9. Formatted the dataset into a readable format.

**Legal and regulatory considerations:**

1. The stock market in US follows strict protocol with the regulatory guidelines and any misrepresentation of the data could lead to legal consequences.
2. I had faced an issue with the API limitation when used the free version of the API keys for yahoo finance and Finance modeling prep API so due to this limitation I had the limit the number of data to process my analysis. So this could cause a biased dataset.
3. Regulatory standards such as data privacy, API data sharing has to be followed
4. Yahoo Finance terms of service must be followed before doing the web scraping and properly follow the ethical guidelines and standards.
5. We have to ensure accuracy and transparency in the financial data to maintain our credibility and also for the legal compliance.

**Risks from data transformations:**

1. As we are restricted to 50 records only due to the API limits our analysis will not be 100 % perfect and it could be a biased dataset and it will not completely represent the market trends
2. Removing duplicate data or missing values without the proper verification could lead to incorrect trend analysis.
3. Web scraping inconsistencies are possible as there could be some variations in how yahoo finance displays their web information and it could be changed anytime in the future without we know what format they updated to. So this would require a new code changes
4. Filling the missing values with N/A instead of estimations could introduce gaps in our analysis.

So it is very important to document our transformations and verify it with multiple data sources.

**Assumption made during data cleaning:**

1. Ticker names are ensured that they are sensitive for consistency.
2. I assumed all the duplicate records has to be removed.
3. Replaced the missing values with N/A without estimating the impact

**Data sourcing and verification of credibility:**

1. I used Kaggle which is one of the best reliable data sources.
2. I used Yahoo Finance and web scraping which came from the official yahoo finance website.
3. I also used financial model prep API which I cross checked with the stock market performance reports.

**As this data source is widely used for data analysis, this is considered the reliable website and dataset.**

**Ethical considerations in Data acquisition:**

1. Dataset of Kaggle is used after reviewing the terms of service from Kaggle website.
2. API data used by acquiring new API keys for this project by properly following their standard.
3. As we extracted key sector performance information from web scraping, we have to review the terms of the service of yahoo finance.

**Mitigation strategy for ethical concerns:**

1. Transparency in the documentation is very important, so we have to document all the transformations with clear explanations
2. Adding users the disclaimer information about the data limitations due to the API constraints.
3. Validation of data source information from alternative reliable datasets.
4. Avoiding misleading information by ensuring the stock market price predictions are not overstated or biased.

**Conclusion:**

With the milestone 5, I have successfully merged, stored, and visualized the stock market data using the advanced data wrangling techniques. Despite various issues and API limitations our project effectively highlighted the sector performance insights. Our future improvements could be including the dataset with the premium API key and various machine learning model which could help us in analyzing stock trends.