Alignment of Amazon Stock Trends with Global Market Indices and Indices Changes During Key Economic Events

A H M RAIHAN - 23089000

1 Introduction

We will discuss the relationship between Amazon's stock price and analyze various external factors related to Amazon's stock price. Further explore how Amazon's stock align with global stock data indeses patterns over time. Aim is to identify trends, correlations, and key factors that affect Amazon's stock price with global stock price, find valuable insights, and try to find corelation trends of the stock price. To do this, collect Amazon's stock data and Global stock data from kaggle.

2 Question

Do Amazon stock trends align with global stock market indices during key economic events?

3 Data Sources

3.1 Global Stock Market (2008-2023) [1]

- Description: This dataset contains all the global stock data since 2008 to 2023.
- Data Structure: This dataset contains total 40550 row number of records in tabular format. There are main features such as open price, high price, low price and volume. The dataset is complete and there are only a few null values.
- License: This dataset is under CC0 public domain license [?], which is open to use.

3.2 US Stock Market Data(Amazon stock data 2010-2021)) [2]

- Description: This dataset contains 2010 to 2021 and. The stock's closing price adjusted for dividends, stock splits, and new stock offerings, providing a more accurate representation of the stock's performance over time.
- Data Structure:This dataset is also in tabular format, it contains total 2474 number of records with lot of potential features for our analysis.
- License: This dataset is under CC0: public domain license [4]license, which is open to use.

3.3 Data Quality

- Accuracy: The datasets contain real-world amazon stock price data which is collected form kaggle.
- Completeness: : Both datasets contain all necessary information to carry out this research.
- Consistency: The data format is consistent across all records and columns.
- Timeliness: Global stock dataset cover 2008 to 2023 and another one is from 2010 to 2021.
- Relevancy: The datasets focus on alignment trendency over the year that relevant to the research question.

4 Data Pipeline

In this research,We have been used python which has rich libraries to build a data pipeline. ETA(Extraction, Transformation and Loading) is widely used data pipeline architecture, that we are using to build the data pipeline. The data has been pulled from kaggle, then performed transformation step to clean the dataset and make the dataset usable. Finally, cleaned data have been stored to the data folder in SQLite database.

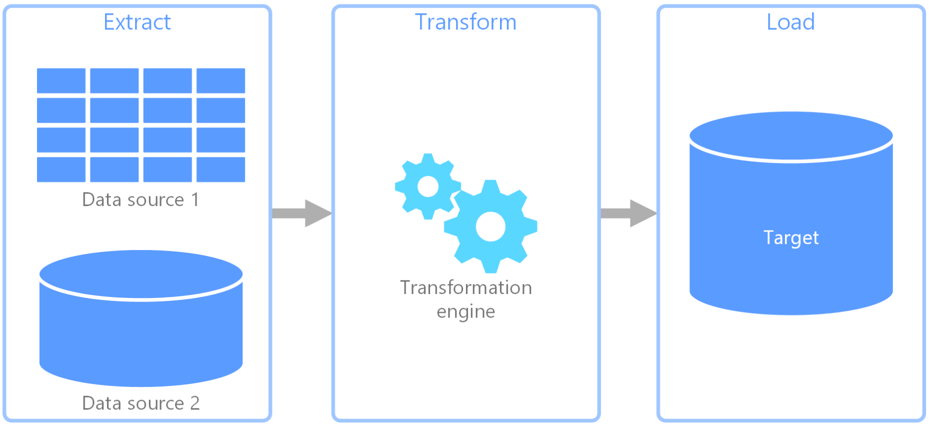

Figure 1: ETA Data Pipeline

4.1 Data Extraction

both dataset have been fetched from Kaggle using the KaggleApi library. This library is responsible for authenticating the valid Kaggle user and downloading the static data in zip format. The datasets have been downloaded to a temporary directory using the tempfile library. Finally, CSV files are read through the pandas library and stored in a dictionary as DataFrames for further processing.

4.2 Data Transformation

The (Global stock data(2008-2023)) datasets are loaded into a Python dictionary, and these DataFrames have been preprocessed and make the data usable. The second dataset (US Stock Market Data Technical Indicators) dataset is also fetched and stored in sqlite datafile. Already mentioned in the data sources section, filtered the data remove the unnecessary row and columns which is not needed in final analysis and also handle the null value. There are lots of feature in second dataFrame, and dropped unnecessary columns such as EMA100, EMA200, ATR, ADX etc. Besides, some features contained only a little amount of null values, which have been replaced by the mean or median for numerical and categorical features, respectively.

4.3 Data Loading

After transformation of data, transformed data into the SQLite database. There are a total of 2 DataFrames for this research from two different sources as mentioned earlier. Both DataFrames are stored as tables inside the database in the data repository.

4.4 Quality

This data pipeline can handle any errors related to update the dataset by the owner. For example, attention has been provided while dealing with features and file names. The file names and the number of datasets from each data source are dynamically handled based on the names retrieved from the sources. Additionally, while dropping the columns, attention has been provided to ensure that if any features are deleted or renamed, no errors occur while running the pipeline. To achieve this, the ‘errors‘ parameter of the ‘drop‘ function is set to ‘’ignore”‘.

4.5 Challenges and Solution

The main challenges during building this pipeline were, two dataset comes from kaggle and there are data format issues. However, both dataset align in same format and usable.Another challenges is not getting any direct dataset link for datasets that could be used to fetch the data. To solve the issue, research was done on how to take data from Kaggle. There are many techniques available to fetch data from Kaggle. The KaggleApi has been used in this pipeline which i mentioned in extraction section.

5 Conclusion

The outcome of this data pipeline is storing the usable data into the SQLite database. As there are two data sources for this analysis, the data has been stored in two tables as two different data frames inside the database in the data repository.

References

[1] Global stock(2008-2023). Global Stock Market (2008-2023). Retrieved from https://www.kaggle.com/datasets/pavankrishnanarne/global-stock-market-2008-present

[2] nikhilkohli. (2010–2020). US Stock Market Data Technical Indicators. Retrieved from https://www.kaggle.com/datasets/nikhilkohli/us-stock-market-data-60-extracted-features?select=AMZN.csv

[3] Creative Commons. CC0: Public Domain Dedication. Retrieved from https://creativecommons.org/publicdomain/zero/1.0/

[4] Creative Commons. CC0: Public Domain Dedication. Retrieved from https://creativecommons.org/publicdomain/zero/1.0/