Eduardo Cisneros, Diego Pretelt

Professor Williamson

DS 2002

December 4, 2024

Gamestop’s Relationship With Reddit

Our project aimed to explore the relationship between GameStop's stock performance and its activity on Reddit, focusing on analyzing data trends to uncover potential insights. To achieve this, we utilized three datasets from Kaggle, implemented an ETL pipeline, and performed data analysis to derive actionable conclusions. The experience offered an opportunity to work with various data engineering and visualization tools, providing valuable technical and collaborative learning. However, it also presented significant challenges, particularly in data integration, transformation, and team coordination.

The first hurdle was identifying suitable datasets that aligned with our research goals. We used datasets containing GameStop stock data and Reddit mentions of gamestop, which contained varying structures and formats. Ensuring their usability required extensive preprocessing. While the datasets were comprehensive, they required verification for completeness and relevance. Aligning Reddit mentions data with stock performance data demanded consistency in date format. One of the important tasks was converting Unix timestamps into a standard YYYY/MM/DD format. This refactor, although simple in hindsight, required some research on python methods that could achieve this. Our merging of three datasets into a unified data frame was challenging due to differences in column names that needed to be normalized as well as the correct use of join operations.

The ETL pipeline involved extracting data from the gathered CSV files, transforming it through cleaning and joining, and loading it into storage systems. Using pandas, we loaded the datasets into data frames, created new data frames for Reddit posts and comments, and calculated daily mentions on trading days (excluding weekends). This transformed data was then combined with GameStop stock data to create a single dataset with all the information necessary for analysis. Storing pandas data frames as MongoDB documents posed challenges, particularly with format compatibility and indexing. This required extensive research on how to refactor from pandas dataframes into the JSON format that MongoDB expects. This also stands for the refactor when fetching from storage. We experienced some service outages with MongoDB so as a contingency, we stored a copy of the cleaned data in Google Cloud Storage. Managing access keys for each team member ensured security but occasionally slowed collaboration when keys were needed for any access or manipulation of data.

Once the cleaned data was ready, we faced some difficulties in choosing which tools to use for analyzing and visualizing it effectively. Creating a time series graphs to compare Reddit mentions and stock performance metrics was more complex than we expected due to the different Y-axis scales. Although with some effective data manipulation we were able to effectively portray our ideas through the visualized data. Upon finishing our visualization techniques we came to the conclusion that the statistical analysis demonstrated a moderate to strong relationship between Reddit mentions of Gamestop and its stock trading volume.

This project made us take into consideration the need and effectiveness of data standardization when it comes to working with different data sets. The transformation of data from different formats in order to effectively store them in cloud solutions also posed a challenge of refactoring whenever uploading or downloading data, especially with MongoDB JSON document format. The constant trial and error of using different merge and join operations on the three datasets until we were satisfied with the resulting data set also consumed a majority of our time, however this was crucial for the adequate data analysis steps that we needed to take. Collaboration and effective communication were also key aspects of this project, principally during the span of time where we were relying on google cloud storage. We were able to maintain clean documentation of our code, and through VCS (github) were able to work independently of each other with a minimized risk of corrupting data or wrongly refactoring code. In future projects of this magnitude and datasets we would like to implement a machine learning algorithm which could be used to predict volume on a given day based on mentions of the stock.

Choosing and analyzing these data sets was difficult, but at the same time rewarding. It definitely made the project more fun as we researched more and occasionally came across a dubious comment. The moderate to strong correlation demonstrated that there are ties between both datasets, however of course not totally influenced by Reddit activity. This project served to expand on our technical skills in ETL processes and data analysis in an interesting and challenging manner.