Duraid Syed, Merna Mostafa, Uthara Das
Intro to Data Science
Professor Chaturvedi
5/8/25

Introduction

The stock market is a complex and ever-changing system influenced by countless factors, making it both fascinating and difficult to analyze. Being able to understand its volatility is a challenge, especially when trying to understand how different stocks behave over time. As the use of data science and machine learning continues to grow, new methods have emerged that allow us to explore financial data in more insightful ways. This project focuses on uncovering patterns in stock performance using five years of historical data from companies in the S&P 500 index. Our goal is to group stocks based on shared characteristics such as volatility, returns, and trading activity. The core problem this project addresses is trying to extract usable insights from volatile market data. To solve this, we use unsupervised learning techniques along with data science knowledge to try to identify structure.

Initially, our project sought to examine patterns in traffic and weather data to identify their relationships. However, after extensive searching, we were unable to find reasonably sized, publicly available datasets that aligned with our goals. Due to these limitations, we revised our focus to a more manageable and well-documented dataset: historical stock data from companies in the S&P 500. This allowed us to conduct a meaningful analysis while still applying the same core data science techniques discussed in class.

Question

The question we are trying to answer is: Can we identify meaningful clusters of S&P 500 stocks based on their volatility, returns, and trading behavior over time?

We plan to answer this question through exploratory data analysis and statistical visualization. We will preprocess and clean the raw stock data by handling missing values, normalizing data, etc. Then we engineer features like daily returns, moving averages, and volatility indicators. Finally, we'll apply clustering to group similar stock behavior.

Course Relation

This approach corresponds with the techniques covered in our lectures on time series analysis, data preprocessing, and unsupervised learning. We used data normalization and transformation methods from our preprocessing lectures. We used clustering from unsupervised learning discussions.

Motivation

Stock market analysis is an evolving challenge in data science and also has benefits in finance and technology. Our group was drawn to this project because of its practical implications and relevance. It also required a variety of skills on our end including visualizing data and being able to evaluate statistics. We were especially interested in this project because of its practical approaches that could be used in a professional setting.

Existing Work

Some existing questions in the area are:

- How can we quantify and model market risk or uncertainty based on price fluctuations?
- Can stock prices be reliably predicted using historical data alone, or is external data, like earning reports, necessary for accurate forecasting?

There are many existing bodies of works that revolve around using technical indicators, and time series models which provided a helpful foundation and opened up several questions. Some studies focus on deep learning models like LSTMs, while others have explored more traditional approaches, including ARIMA forecasting and regression models. This project intends to take a more interpretable, transparent approach by working with clustering.

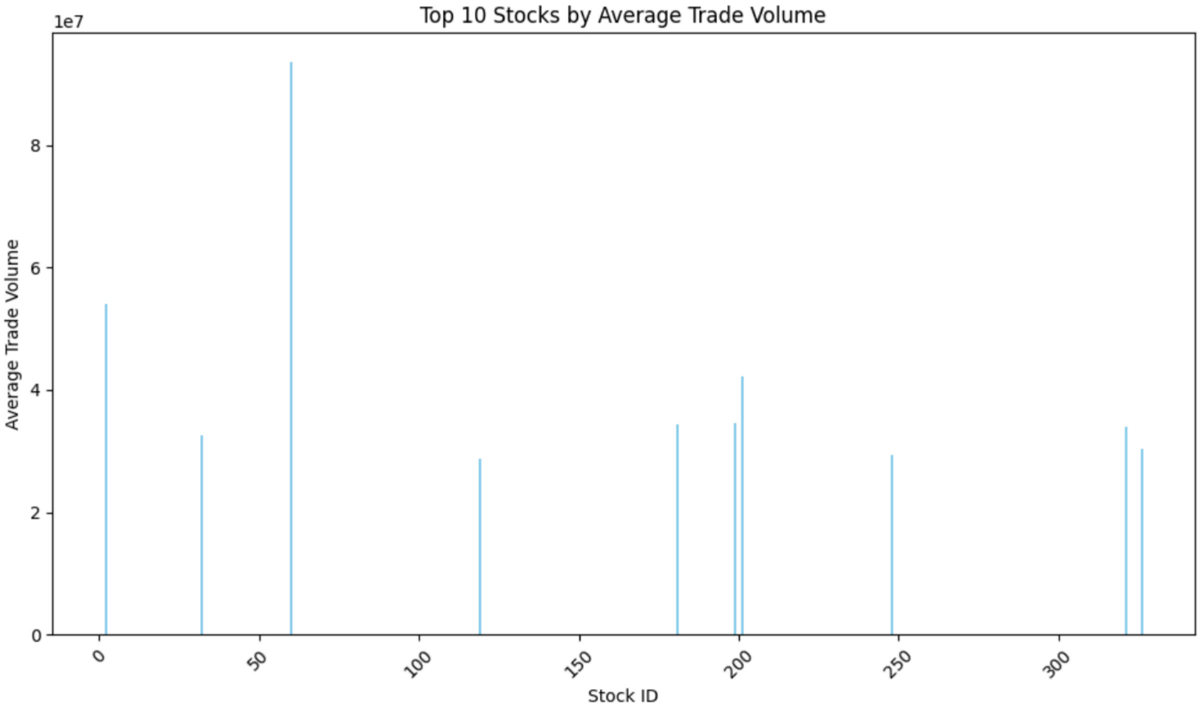
Method:

We used the *S&P 500 5-Year Stock Data* from Kaggle, which provides daily stock price information for all companies listed on the S&P 500 index over a five-year period. The dataset contains the columns date, open, high, low, close, volume, and name, which represents the stock ticker. Each row in the dataset corresponds to a single trading day for a particular stock, making it well-suited for time-series and performance analysis.

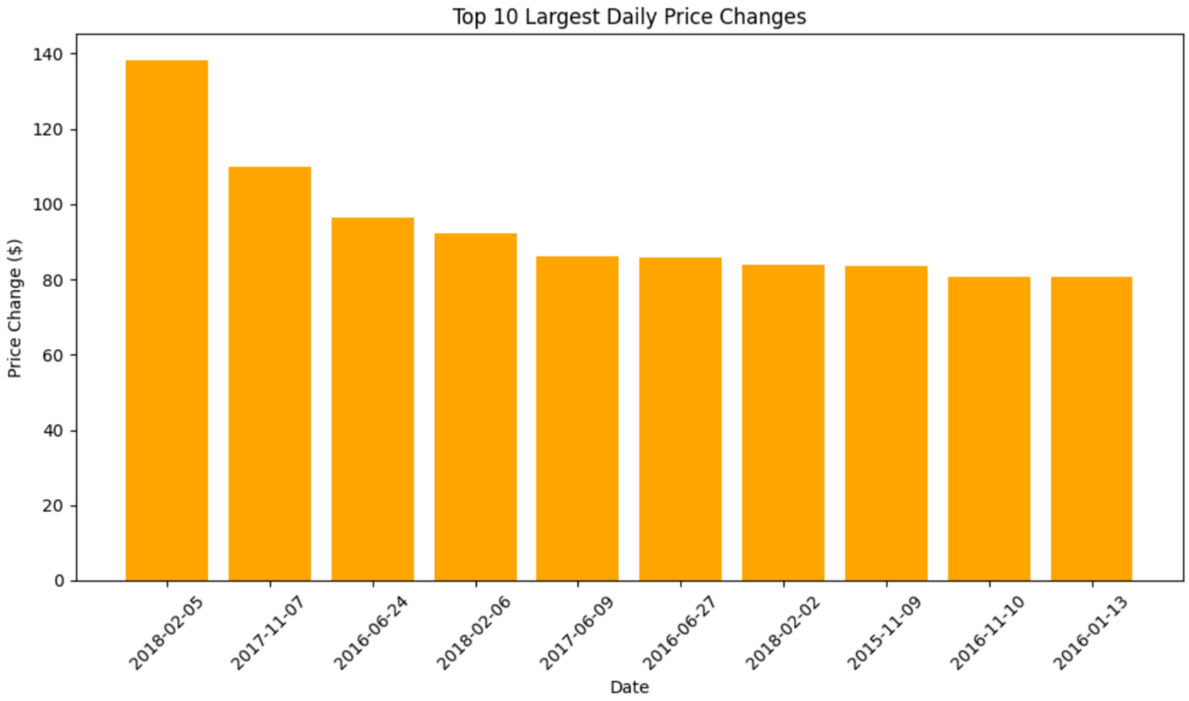
To prepare the data, we converted the date column to datetime format, added a unique stock_id for each ticker, and performed various preprocessing steps such as handling missing values and sorting by date. We then engineered several new features to help capture the behavior of each stock, including daily return (the percent change from open to close), rolling volatility (standard deviation of daily returns), 20-day and 50-day moving averages, and the Sharpe ratio (average return divided by volatility). These features provided a basis for deeper analysis.

Our analysis combined exploratory data analysis with unsupervised learning. We began by visualizing trading volume, daily price changes, and return trends to understand the spread and behavior of different stocks. Then we aggregated the data at the stock level using the engineered features and applied KMeans clustering to group stocks together. We visualized these clusters directly using average return and volatility. This allowed us to identify distinct groups of stocks with similar risk-returns, demonstrating how unsupervised learning can uncover meaningful patterns in financial data.

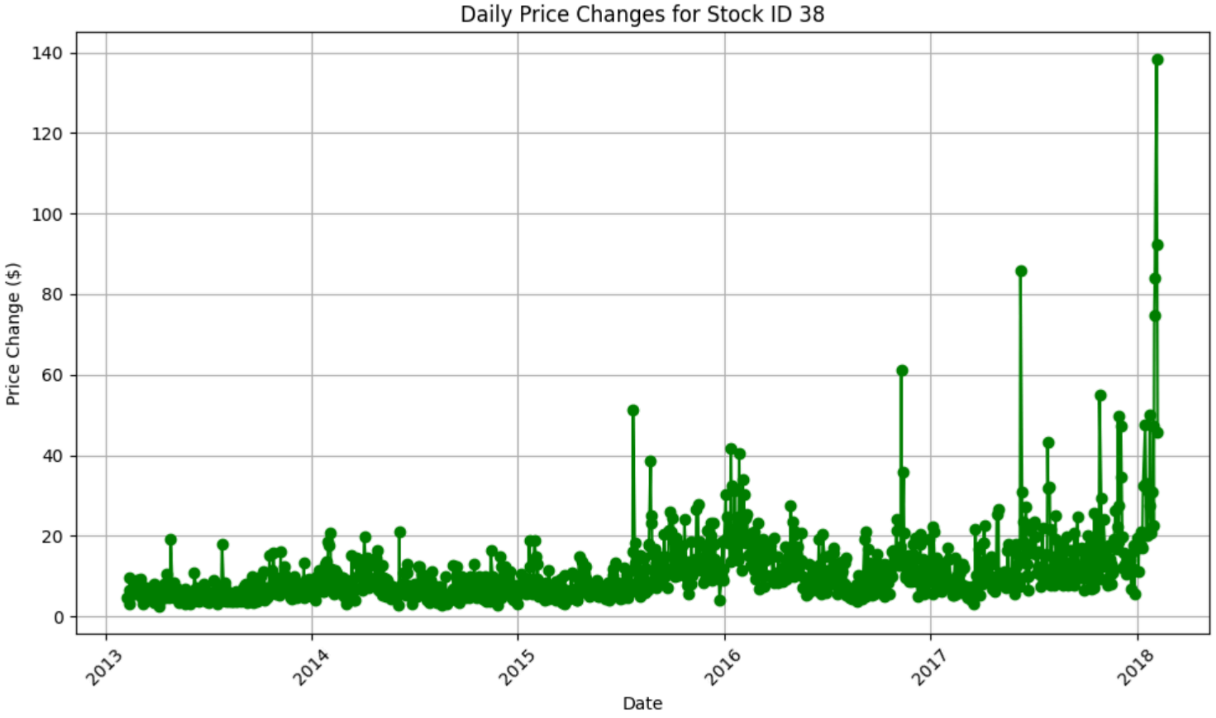
Results:



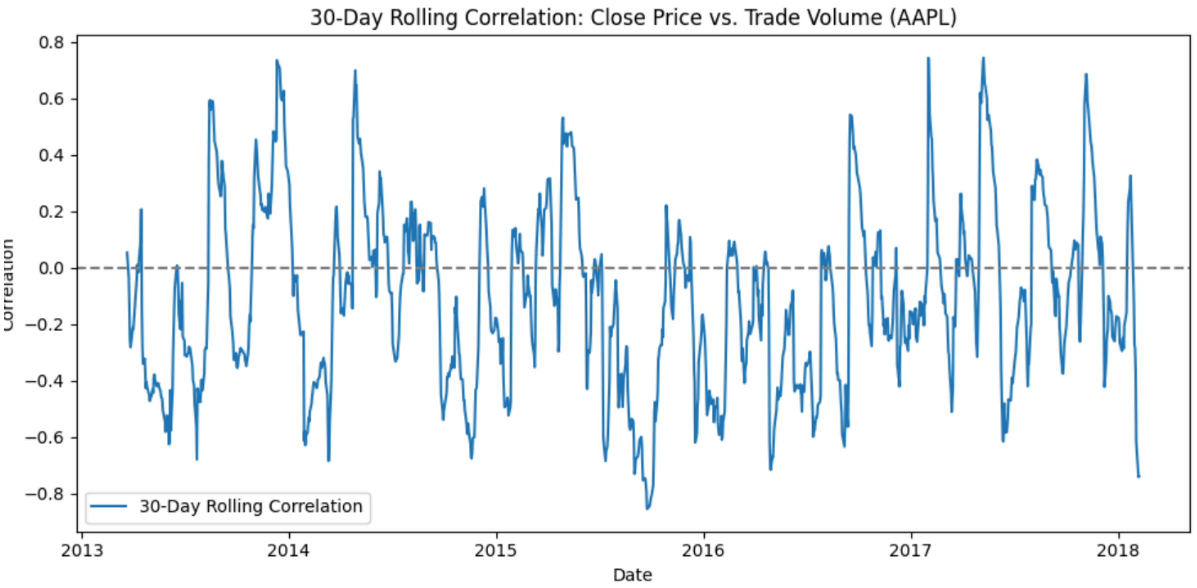
This bar chart shows the top 10 stocks with the highest average trade volume, with one stock (around ID 60) standing out with nearly 90 million average trades. High trade volume often indicates strong investor interest, institutional activity, or frequent news coverage. These stocks are likely more liquid and may experience different volatility patterns than lower-volume stocks. This could open discussions about how volume relates to price stability and whether high-volume stocks behave differently during market shifts.



This bar chart shows the top 10 largest daily price changes for Stock ID 38, with the highest spike occurring on February 5, 2018, at nearly $140. Most of these large changes cluster between 2016 and 2018, reinforcing the earlier observation that volatility increased sharply during this period. This trend suggests a shift in market behavior, possibly tied to external events or internal company changes. It raises further questions about what specifically triggered these surges and whether similar spikes occurred across other stocks in the index.



The analysis of daily price changes for Stock ID 38 shows a sharp increase in volatility starting around 2016, with significant spikes in 2017–2018 reaching up to $140. This suggests a major shift in the stock's behavior, possibly due to company changes, or market speculation. The earlier years show relatively stable, low-magnitude changes. These results raise questions about what triggered this volatility and whether similar patterns exist across other stocks.



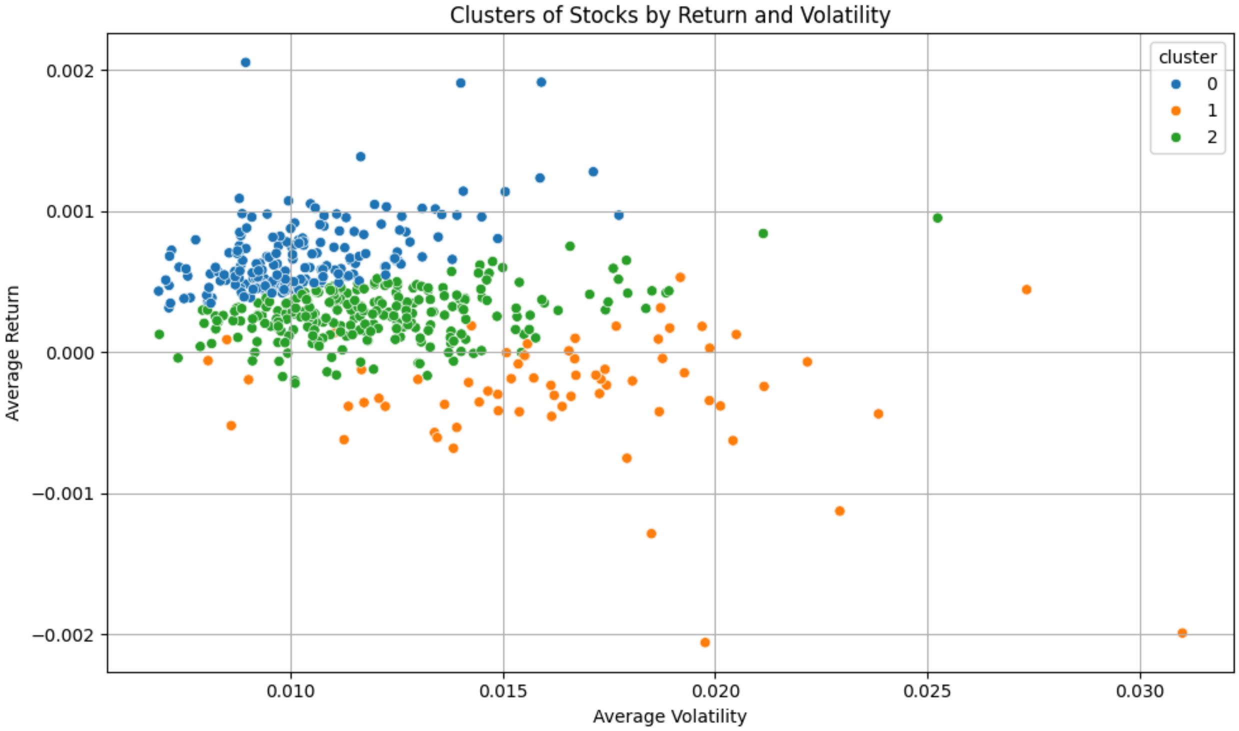
This plot shows the 30-day rolling correlation between Apple's closing price and trade volume. The correlation is highly variable over time, occasionally peaking above 0.6 and dipping below -0.7. This suggests that the relationship between price and volume is not stable and likely depends on broader market conditions or investor sentiment. Periods of negative correlation, like 2014 and 2016, may reflect panic selling, where volume surges as prices drop. Positive correlation spikes might indicate buying momentum, where increased volume pushes prices higher.

Expectations

We expected that applying unsupervised learning, specifically KMeans clustering to engineered stock features would help us uncover meaningful patterns within the S&P 500 companies. Our primary assumption was that certain financial indicators such as average daily return, rolling volatility, trading volume, and the Sharpe ratio would naturally group stocks with similar behavior into distinct clusters. These clusters could potentially represent categories like high-growth stocks, stable blue-chip companies, or high-risk, high-volatility assets. We believed that these clusters would reflect clear investment strategies. For example, one group might contain stocks with high returns and moderate volatility, which could be ideal for growth-focused investors. Another group might consist of stocks with low returns but also low volatility, appealing to conservative investors seeking stability. Because the features we used are common in financial analysis and frequently referenced, we expected the model to form clusters that aligned with real-world investment categories.This would not only support our assumptions about groupings but also allow us to interpret the results more effectively.

Actual Results

While the clustering was able to group the stocks based on their characteristics, the outcomes were not as clear as we had hoped. Some stocks appeared in clusters that didn't clearly align with our expectations. For example, there were some cases where high-volume stocks were grouped with stocks that had low returns, which made it difficult to label the clusters accordingly.



This tells us that even if our features may have been useful, they also may have not been able to properly separate the stocks by investment style/risk profile. It's also possible that some of the variation in stock behavior is influenced by factors outside the scope of our dataset such as sector information, recent earnings reports, or global news events that weren't included in our clustering inputs. Another limitation we encountered was that clustering, as an unsupervised learning technique, doesn't provide a straightforward way to evaluate performance. Since we didn't have ground truth labels for what a correct cluster would look like, we relied mostly on visualizations and summary statistics to assess whether the clusters made sense. This made it harder to know if the model had truly captured something meaningful or if the groupings were random.

Conclusion

Overall, our project showed that clustering can be a helpful tool in analyzing stock market behavior, but it also highlighted the complexity of financial data. While we were able to group stocks based on shared statistical features like volatility and returns, the resulting clusters were not always easy to interpret or clearly aligned with real-world investment categories. This suggests that our current feature set, while informative, may have missed key factors such as external events that influence stock performance. In the future, adding more context-specific features and exploring other unsupervised techniques could help produce more meaningful groupings and deeper insights into stock behavior.