SENTIMENT & CAPITAL: THE DATA BEHIND FINANCIAL MOVEMENTS

Varun Tirthani
New York University
vyt2006

Rahul Raman
New York University
rr4549

Srivats Poddar
New York University
sp7811

Satyanarayna Chillale
New York University
sc9960

January 23, 2025

ABSTRACT

In today's age where financial indicators are worth their weight in gold, we are experimenting with legacy data in the form of bitcoin and personal loan data as well as tweets involving publicly-traded companies to test and understand their impact on the market prices of stocks and their level of interdependence with each other.

We are leveraging distributed tools and platforms like Apache Spark and Apache Zeppelin to preprocess, clean, derive insights and test our hypotheses.

1 Introduction

We have used bitcoin, loan, tweet and stock price data to analyze the magnitude and directional dependence with each other and derived crucial insights. We have performed sentiment analysis on tweets, enhanced the financial data on hand to calculate additional metrics and derived some astonishing results and insights that further cement tried and tested market hypotheses. We have leveraged concepts taught in class and have built all our analysis in Apache Spark and using Apache Zeppelin.

Financial markets (bitcoin, stock prices) are considered as leading indicators of an economy whereas loans give a lagging picture of an economy's health. Through this analysis we are trying to confirm whether the leading and lagging indicators are actually leading and lagging respectively.

We are also experimenting to analyze the impact related tweets have on stock prices and bitcoin. Apart from this, we have gone a step ahead and attempted to categorize stocks into sectors and come up with their cross-sectoral impact in the form of shock propagation analysis. Correlating stock prices to bitcoin, stocks amongst their previous lagged prices and future stock prices to loan approvals and rejections are some of the further analysis we have conducted given the data we have.

We have performed our analysis and derived insights over a time interval common across datasets, which is from 2015 to 2018.

Apart from the financial aspect of our work, we have ensured to implement and encourage the use of technical concepts taught to us while navigating complex challenges in code

This analysis is useful and intended specially for Hedge Funds, Portfolio Managers and Equity Researchers and basically anyone enthusiastic about capital markets and curious to know which factors can actually significantly move markets.

This analysis is important as those vaguely familiar with finance would be slightly aware of the factors that could potentially affect market prices but might not be aware of their magnitude and direction and hence assigning estimate numerical values to this solidifies our understanding of the impact

2 Data

2.1 Stock Market Dataset

The stock dataset [1] is 700 MB in size and contains multiple files, each corresponding to different stock names. In total, it includes approximately 14 million entries. The dataset provides stock information for U.S. companies spanning the period from 1962 to 2017. To enrich this dataset, we integrated it with the Nasdaq Screener to incorporate sector(12) and industry(142) information. The schema of the dataset is shown in Table 1.

2.2 Tweets Dataset

The tweet dataset [2] is approximately 800 MB in size and contains the uncleaned tweet body which is the main

Table 1: Stock Dataset Schema Representation

Column Name	Data Type
date	Date
stockName	String
open	Double
close	Double
low	Double
high	Double
volume	Double
Symbol	String
Name	String
Country	String
Sector	String
Industry	String

focus of the preprocessing and cleaning. The schema of the dataset is shown in Table 2:

Table 2: Tweet Data Schema Representation

Column Name	Data Type
tweet_id	String
writer	String
post_date	String
body	String
comment_num	String
retweet_num	String
like_num	String

2.3 Data: Bitcoin Dataset

The bitcoin data [3] is approximately 350MB in size. This dataset contains BTC-USD price at 1-minute intervals from 2012-01-01 to Present (currently 2024-11-22). This dataset has $\sim$6.72 million rows. Schema of dataset is shown in Table 3.

Table 3: Description of columns in the dataset.

Column Name	Data Type
Timestamp	Double
Open	Double
Close	Double
Low	Double
High	Double
Volume	Double

2.4 Personal Loan Dataset

The personal loan dataset [4] utilized in this study is a comprehensive collection of loan application data, totaling approximately **648 MB** in size. It is divided into two CSV files:

- **Accepted Dataset:** This file comprises **2.2 million rows** and **151 columns**. Following a thorough process of data cleaning and feature selection, we refined the dataset to focus on **30 key columns** relevant to our analysis.
- **Rejected Dataset:** This file is significantly larger, with **27.6 million rows** and **9 columns**. Our study is focused on the rejected dataset, as it provides insights into loan rejection trends. Schema as shown in Table 4.

Column Name	Type
Amount Requested	double
Application Date	date
Loan Title	string
Risk_Score	double
Debt-To-Income Ratio	string
Zip Code	string
State	string
Employment Length	string
Policy Code	double

Table 4: Preprocessing Summary of Rejected Loans

3 Architecture Overview

As illustrated in our design diagram (refer to Figure 1), the datasets were sourced from Kaggle and ingested into NYU Google Dataproc's HDFS. We performed all cleaning, profiling, and analysis using **Scala Spark** within Zeppelin notebooks, with some plots generated using Python.

Our analysis is categorized based on the tools utilized:

- **SQL (DataFrames and Datasets):**
  - Examining the impact of tweets on Bitcoin and the stock market.
  - Analyzing the effect of stock market trends on Bitcoin.
  - Studying how general market trends influence the rejection rate of personal loans.
- **MLlib:**
  - Linear regression for stock prediction.
- **GraphX:**
  - Modeling shock propagation between companies across sectors.

4 Data Ingestion

4.1 Stocks Dataset

The data cleaning process began with **extracting stock names** from file names, which were formatted as *<stockName>.us.txt*. A new column named "stockName" was

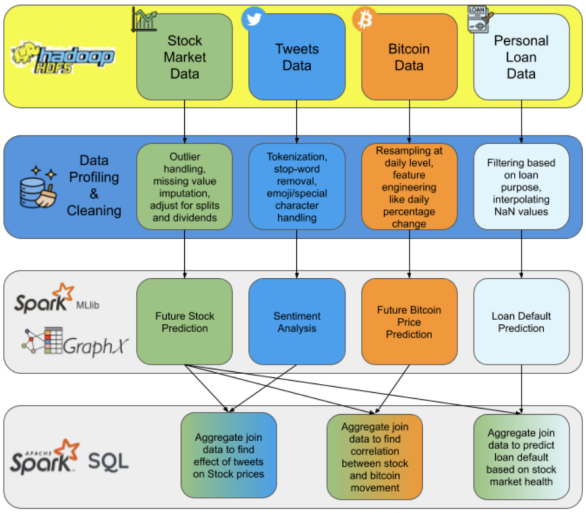

Figure 1: Project Design Diagram

created by extracting the stock name using regex. To **handle missing data**, we first checked for rows with missing stock names or date values and found no such rows. All numerical columns were converted to *double* type, and missing values were imputed using the median strategy after grouping the data by "stockName".

Next, **deduplication** was performed by identifying and removing duplicate rows. Rows with the same date for the same stock were also eliminated to ensure data integrity. The dataset was then **limited** to include only rows where the date is after 2011, maintaining consistency with other group data. Stocks with less than one year of data were excluded, and rows with volume values below 1000 were filtered out.

The **validation process** ensured that the data met specific quality criteria. It was verified that the "low" values were greater than zero and less than both "open" and "close" prices. Additionally, an upper bound of 100,000 was applied to "high" values to limit anomalies, and it was confirmed that the "high" values were greater than both "open" and "close" prices.

Outliers were handled by grouping data based on stock names and calculating the mean, standard deviation, and z-scores. Rows with values more than two standard deviations away from the mean were filtered out. This step was applied to the "open", "close", and "volume" columns to ensure consistency and remove extreme anomalies.

4.2 Tweets Dataset

For the tweet data, the numerical data was first converted from string to double. The dates were provided in Unix timestamps so to ensure consistency across all financial data, they were converted to the EST format. The statistics in the data set were later profiled using the **describe** method.

The next step was to do a sanity check for missing/null values of the tweet body. It does not make sense to perform any analysis for an entry that has a missing tweet body. Thankfully, there were no such rows in the dataset.

The next step was to count the unique number of tweet writers to understand the distribution of the sample size on which the data set was based. The figure was in line with expectations (around **140,000** unique writers for **3,700,000** tweets) and hence no further cleaning with respect to the writers column was needed.

Within the tweet body, we then tried to find the count of the mentions of the companies so that it later becomes easier to link a tweet to a particular company. Observing the results, it was pretty obvious that data needed to be cleaned as for example, **$tsla** and **$TSLA** were being categorized separately and so was **$TSLA.** and **$TSLA** and hence the need for data normalization and punctuation cleaning as part of the data cleaning process was highly needed.

Now that we have pre-processed the data and identified the pain points in the data, mainly to do with the tweet body, we focused on cleaning the same and leveraged Spark ML features for the cleaning. We started by cleaning it of any hyperlinks and special characters (except $) and then tokenized and removed stopwords using Spark ML.

4.3 Bitcoin Dataset

The Bitcoin dataset was the cleanest among the selected datasets, containing only one `NaN` value in a single column. To preserve data integrity, the affected row was removed. Duplicate entries were checked using the `Timestamp` column as the primary identifier, and no duplicates were found. The dataset was further validated against the following conditions to ensure correctness:

$$\text{Low} \leq \text{Open},\ \text{Low} \leq \text{Close},\ \text{Low} \leq \text{High},$$
$$\text{High} \geq \text{Open},\ \text{High} \geq \text{Close},\ \text{Volume} \geq 0$$

No violations were detected, confirming the dataset's consistency with the defined rules and enabling data enhancement.

4.4 Personal Loan Dataset

Here are the following steps done to pick the columns required in accepted loan dataset:

1. Drop columns that have more than 30% NaN / NULL values.
2. Took reference from the official Kaggle dataset and understand the importance of each column and its relevance to our analysis.

In rejected loan dataset, the *Risk Score* is similar to `fico_score`.

To clean our dataset, the following steps were performed:

1. Standard cleaning based on minimum and maximum values.
2. Imputation for numerical columns were imputed using the median strategy.
3. All categorical string columns were imputed using the mode.
4. Specific cleaning methods for individual columns:
   (a) For **Accepted Dataset**:
      i. `Term`: Converted values (36 months, 60 months) to integers.
      ii. `emp_length`: Converted values like `< x years` or `x+ years` into `x year` and transformed them into integers.
      iii. `fico_range_low` and `fico_range_high`: Averaged these two values and stored the result in `fico_score`.
   (b) For **Rejected Dataset**:
      i. `Debt-To-Income Ratio`: Removed the % symbol and converted values to float.

5 Data Enhancement

5.1 Tweets Dataset

Once the tweet data was cleaned, the focus was now on assigning a sentiment to a tweet. Since the data was not previously labeled, we had to use a renowned lexicon called `affin` to assign high-impact words in the tweet text with the intention such that the sum of all such scores of words within a text will reveal the overall results. This allows us to calculate the column `sentiment` with values `Positive`, `Negative` or `Neutral`

The lexicon needed to be shared to all partitions of our tweet data since they all needed to be joined with the lexicon scores but this would have meant very high communication traffic and less efficient in terms of execution.

To improve efficiency, we used the concept of broadcast values. Broadcasting the lexicon that was being used extensively for the join resulted in significant reduction in execution times and gave us an efficient way to compute the tweet sentiment.

5.2 Stocks Dataset

The Stocks dataset was enriched by joining it with the Nasdaq Screener to add sector and industry information. Rows with null values for sector or industry were removed to facilitate sector and industry-based analysis. This comprehensive cleaning and validation process resulted in a high-quality dataset suitable for further analysis.

Next, we joined the stocks dataset with Twitter sentiment data. We mapped the tweet sentiment values to **1 for positive**, **0 for neutral**, and **-1 for negative**. The data was then grouped by **stock** and **date** to calculate the sum of positive and negative sentiment scores. The difference between the positive and negative sums was taken as the net sentiment score for each stock on a given day.

For the Linear Regression models predicting stocks and the LR models for GraphX edges, we compute the daily returns for each sector. The returns are calculated using the formula:

$$\text{Return} = \log(\text{close}_T) - \log(\text{close}_{T-1})$$

This method helps smooth the graph by minimizing extreme fluctuations.

5.3 Bitcoin Dataset

To align with the granularity of other datasets, which operate at the daily level, the 1-minute interval data in the Bitcoin dataset was aggregated to a daily level. The `Timestamp` column, originally in UNIX format, was converted to the `date` column in the EST time zone to ensure consistency across datasets. Aggregation was performed as follows:

1. `Open`: The first `Open` value within each day.
2. `Close`: The last `Close` value within each day.
3. `High`: The maximum `High` value within each day.
4. `Low`: The minimum `Low` value within each day.
5. `Volume`: The sum of `Volume` within each day.

The average daily Bitcoin price was calculated using:

$$\text{Average Price} = \frac{\text{High} + \text{Low} + \text{Close}}{3}$$

To capture short- and long-term trends, two additional features, `5Day_SMA` and `28Day_SMA`, were computed as the 5-day and 28-day simple moving averages (SMA) of the `AveragePrice`. These calculations utilized a rolling window ordered by `date`, with ranges of `rowsBetween(-4, 0)` for the 5-day SMA and `rowsBetween(-27, 0)` for the 28-day SMA.

To analyze the impact of Twitter sentiment, four additional features (`1Day_Future_Returns`, `3Day_Future_Returns`, `7Day_Future_Returns`, `14Day_Future_Returns`) were created by calculating future returns using a `Window` ordered by `date` and `lead(period)` for periods of 1, 3, 7, and 14 days.

$$\text{p-day Future Return} = \frac{\text{Close}_{\text{p-day future}} - \text{Close}_{\text{today}}}{\text{Close}_{\text{today}}} \times 100$$

Figure 2: Tweet Sentiment Distribution Pie Chart From [2]

Similarly, to correlate Bitcoin data with stock movements, four features for past returns (`1Day_Past_Returns`, `3Day_Past_Returns`, `7Day_Past_Returns`, `14Day_Past_Returns`) were created in both Bitcoin and Stock dataset using `lag(period)` operations.

$$\text{p-day Past Return} = \frac{\text{Close}_{\text{p-day past}} - \text{Close}_{\text{today}}}{\text{Close}_{\text{today}}} \times 100$$

6 Data Analysis

6.1 Sentiment Analyis

We assigned a positive sentiment for those tweets that had a total score greater than 0, negative sentiment for those tweets that had a score less than 0 and a neutral sentiment for those tweets that had a score of 0. The distribution (shown in Figure 1) was such that half of the tweets were neutral while 40% were positive and the rest were negative.

This data was shared across for further analysis and impact calculation on stocks and bitcoin data.

6.2 Twitter Sentiment and Stocks Correlation

For the data set of stocks, we calculated the **daily change percentage** using the formula:

$$\text{Daily Change Percentage} = \frac{\text{close} - \text{open}}{\text{open}} \times 100$$

We used the Spark `corr` function to calculate the correlation between the net sentiment score and the daily change percentage for each stock across different years. We limited data to sentiment score greater than 50 and had more than 50 days of data per year. Additionally, we correlated the **total engagement** (positive, negative, neutral) with the **trading volume** to analyze the relationship between social media activity and market behavior.

To examine the impact of sentiment on future price movements, we calculated the **next-day change percentage** using the Spark `lead` function. Non-consecutive trading days were filtered out to ensure data continuity. Finally, we computed the correlation between the Twitter sentiment score and the next-day change percentage to identify potential predictive relationships.

6.3 Top Gaining and Losing Stocks Analysis

In addition to the sentiment and correlation analysis, we identified the **top three stocks** in each **sector** for each **year** that showed the most significant gains and losses.

To achieve this, we calculated the **annual percentage change** (APC) for each stock using the formula:

$$APC = \frac{P_{\text{end}} - P_{\text{start}}}{P_{\text{start}}} \times 100$$

where P_{end} is the closing price at year-end and P_{start} is the closing price at year-start.

We utilized Spark functions such as **window partitioning** to efficiently perform this calculation. Specifically:

- The data was partitioned by **sector**, **stock**, and **year**.
- Window functions were applied to compute P_{start} and P_{end} for each stock within its respective sector and year.

Once the annual percentage change was calculated, the dataset was grouped by **sector** and **year**. Within each group, we ranked the stocks using the `row_number()` function based on their annual percentage change to determine the following:

- **Top 3 gaining stocks**: Stocks with the highest positive percentage change within each sector for the given year.
- **Top 3 losing stocks**: Stocks with the most significant negative percentage change (largest losses) within each sector for the given year.

This ranking provided insights into sector-level stock performance, highlighting top-performing and underperforming companies, and helping identify trends and outliers within sectors.

6.4 Twitter Sentiment and Bitcoin Correlation

To identify potential correlations, the Spark `corr` function was applied to multiple comparisons, including the following:

1. Bitcoin `AveragePrice` vs. the count of positive, negative, and neutral tweets for each date.
2. Bitcoin trading volume vs. the total Bitcoin-related tweet count for each date.
3. `1Day_Future_Returns`, `3Day_Future_Returns`, `7Day_Future_Returns`, and `14Day_Future_Returns` vs. the count of positive, negative, and neutral tweets for each date.

6.5 Bitcoin and Stocks Correlation

To analyze the relationship between Bitcoin and stocks, the Spark `corr` function was utilized to compute correlations between features of Bitcoin and stocks. The features included `1Day_Future_Returns`, `3Day_Future_Returns`, `7Day_Future_Returns`, `14Day_Future_Returns`, `Volume`, and `AveragePrice` for both Bitcoin and individual stocks in the dataset.

The stocks were ranked based on their correlation coefficients with each feature. For each feature, the top 10 most positively correlated and the bottom 10 most negatively correlated stocks were identified and reported.

7 Insights and Results

7.1 Sector Shock Propogation

In financial markets, a sudden change or shock from one company can have a larger impact due to the interconnectedness between companies. These shocks, or ripple effects, can spread across different sectors, industries, or even geographies, causing market volatility, shifts in investor behavior, disruptions in supply chains, or changes in consumer demand. The propagation of these effects describes how they move through the market and interact with other factors, leading to broader and sometimes unpredictable market reactions.

Given the importance of understanding shock propagation in financial markets, we built a relationship model (represented as in Figure 3) between various sectors to study the effects of shock propagation and predict how one sector's shock impacts others. This model is represented using graphs, where the nodes represent sectors and the edges show the relationships between them. To measure the correlation between sectors, we created 132 Linear Regression models (since there are 12 sectors) to predict the average return of a sector based on the past 2 days' average returns of other sectors. The node values are struct of data of initialized with stock market data at timestep T, along with its sector name and the past 2 days AvgReturn for this sector; and the edge values are represented as tuples of coefficients and bias from the Linear Regression models.

Once the model is initialized Table 5 – For this experiment, we set T = 2014-01-10, we simulate a ripple effect by changing one sector's value – For this experiment we changed Finance at timestep $T + 1$. The shock then propagates to nearby sectors through graph message passing, where the shock's effect is predicted using the Linear Regression coefficients stored on the edges. Each sector's return is updated based on the average of all incoming messages Table 6, which predict the returns at timestep $T + 1$. This graph message passing is implemented using the Pregel API, and the model essentially consists of multiple chained Linear Regression models represented as a graph. To analyze the shock propagation in our model, we amplified the effect of returns in the Finance sector by a factor of 10. This adjustment led to the following observed impacts across different sectors:

- The Real Estate sector experienced massive returns, suggesting that it is highly sensitive to changes in the Finance sector. This could indicate a strong correlation between these two sectors, where financial market fluctuations heavily influence real estate investments.
- Consumer Staples and Energy sectors showed slight improvements in their average returns. This implies that these sectors may be somewhat insulated from financial shocks, or that their behavior is less volatile compared to more sensitive sectors like Real Estate.
- Investing in Technology shifted from a loss to massive profits, demonstrating the sector's ability to rebound from negative shocks. This reflects the growth potential in the technology market, which can benefit significantly from positive changes in the broader financial landscape.
- The Health Care sector experienced a slight negative impact, indicating that it may be more vulnerable to financial market fluctuations than other sectors. This suggests that healthcare investments could be affected by shifts in investor sentiment driven by changes in the finance sector.

These results highlight how interconnected the financial market sectors are, with some sectors more sensitive to shocks than others. The amplification of returns in the Finance sector provides insights into how these ripple effects can lead to broad market movements, affecting various sectors in diverse ways. This analysis also underscores the importance of understanding shock propagation for investors, as different sectors exhibit varying levels of resilience or vulnerability to financial shocks.

Another analysis that was performed was to try and use a Linear Regression model to predict a day's return based on its return the previous day, the volume of the stock traded on the day and open interest. The model returned a R^2 score of **0.004** for the split of data that it was tested with. This further reiterates our assumption that linear regression models are not the most ideal for estimating time-series data

7.2 Twitter sentiment and stock data correlation

- The **top 10 stocks each year** that gained and lost the most based on correlation with positive and negative Twitter sentiment respectively. This highlights how sentiment aligns with the best and worst performing stocks annually.
- Top 10 stock correlation with total Twitter engagement. This explores the relationship between social media activity and trading volume.

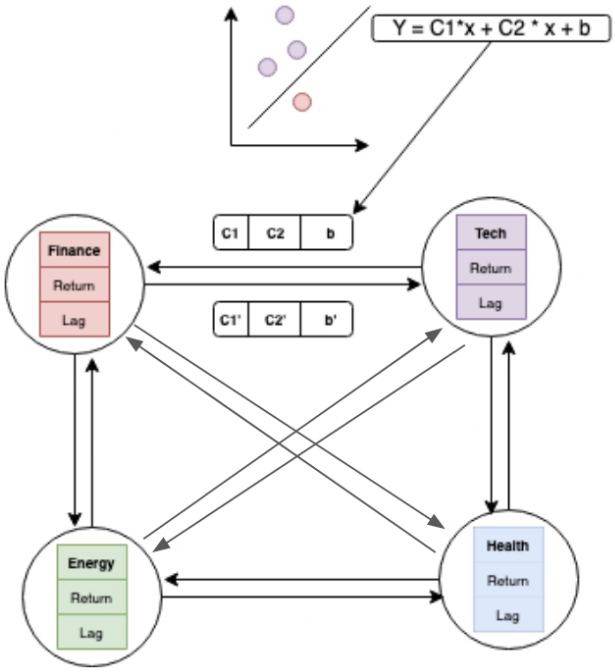

Figure 3: Representing the financial market relations as a Graph. Each Node consists of the sector details and the edges represents the relationship between the sectors.

Table 5: Sector-wise Average Returns: The graph represents the average returns across various sectors. The values are relatively low because they are calculated as simple averages across multiple small and large companies, without accounting for the weight of each company in the financial market.

Node (sector)	AvgReturn
Miscellaneous	-0.0510
Finance	0.0032
Real Estate	-0.0007
Health Care	0.0135
Consumer Staples	0.0009
Industrials	-0.0006
Technology	-0.0057
Utilities	-0.0015
Basic Materials	-0.0187
Telecommunications	-0.0010
Energy	-0.0043
Consumer Discretionary	-0.0022

- The **top 10 stocks each year** that gained and lost the most based on correlation with previous day positive and negative Twitter sentiment respectively. This highlights how Twitter sentiment impacts next day stock price. Some of these graphs are shown in the Figure 4. From the graphs we see that there is minimal correlation which suggests that stock prices are not only dependent on the twitter sentiment but various other factors.

Table 6: Sector-wise Average Returns after shock propagation (i.e., message passing) in the graph.

Node (sector)	AvgReturn
Miscellaneous	-0.0007
Finance	0.0033
Real Estate	0.0039
Health Care	-0.0030
Consumer Staples	0.0015
Industrials	0.0004
Technology	0.0018
Utilities	0.0036
Basic Materials	-0.0032
Telecommunications	-0.0002
Energy	-0.0055
Consumer Discretionary	-0.0001

These visualizations provide valuable information on how Twitter sentiment and engagement influence stock performance trends, both annually and in predicting short-term price movements.

7.3 Top performing stocks each sector annually

- The **top stock that dipped the most** in each sector for year 2016 and 2017. This graph identifies the worst-performing stocks annually within each sector. [Figure 5]

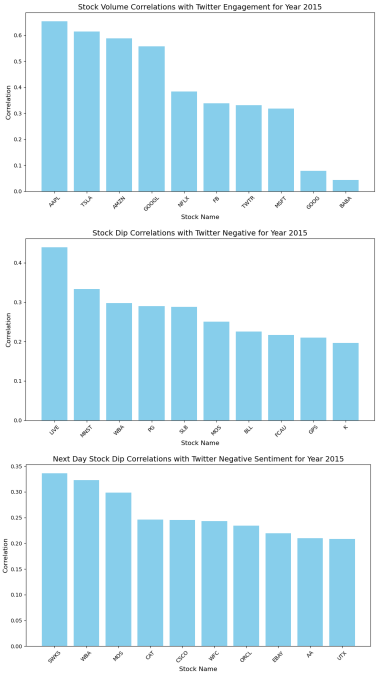

Figure 4: Twitter Sentiment vs Stock Correlation

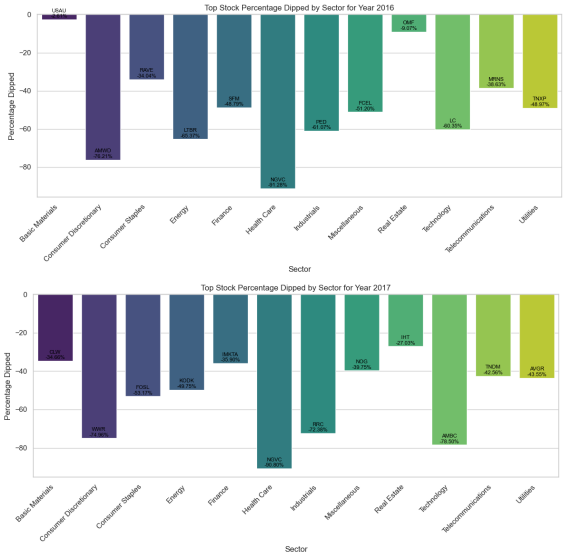

Figure 5: Top stock dipped per sector

- The **top stock that gained the most** in each sector for year 2016 and 2017. This graph highlights the best-performing stocks annually across different sectors. [Figure 6]

These visualizations provide a sector-level perspective on stock performance over time, enabling comparisons between sectors and identifying trends among the best and worst performing stocks.

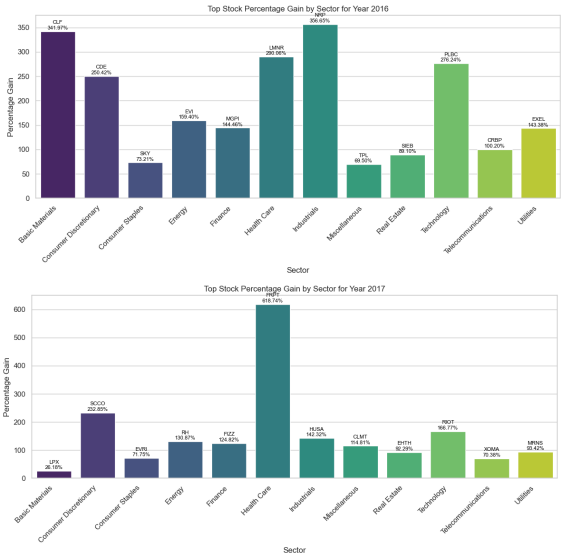

Figure 6: Top stock gain per sector

7.4 Twitter Sentiment and Bitcoin Price

The analysis shows minimal correlation between Bitcoin prices or returns and Twitter sentiment. As seen in Table 7, the correlations of `AveragePrice` with positive, neutral, and negative tweets are 0.49, 0.55, and 0.48, respectively, indicating no strong relationship between sentiment and price. Additionally, as shown in Table 8, return correlations for all sentiment categories are very low, with values close to zero. For instance, the 1-day return correlations with positive, neutral, and negative tweets are -0.02, 0.01, and 0.01, respectively, and similar minimal values are observed for 3-day, 7-day, and 14-day returns. The correlation of Bitcoin trading volume with the total number of tweets is 0.03, as seen in Table 9.

The data reveals that Twitter sentiment data is available for 1013 days (55%) out of 1826 days in the analyzed period (2015-01-01 to 2019-12-31), as presented in Table 10. Furthermore, the number of negative tweets (1668) is significantly lower compared to positive (6436) and neutral tweets (5254), which skews the dataset and may contribute to inconclusive trends.

Table 7: Correlation of Average Bitcoin Price with Tweet Sentiments

Metric	Positive	Neutral	Negative
AveragePrice	0.49	0.55	0.48

7.5 Additional Trends

In this section, we analyze the Lending Club loan dataset in conjunction with the state of the stock market and

Table 8: Correlation of Bitcoin Returns with Tweet Sentiments

Metric	Positive	Neutral	Negative
1-day	-0.02	0.01	0.01
3-day	-0.05	0.01	-0.01
7-day	-0.08	0.01	-0.03
14-day	-0.11	0.02	0.04

Table 9: Correlation of Bitcoin Volume with Total Tweet Count

Metric	Correlation
Volume with Total Tweets	0.03

Bitcoin trends to identify indirect effects of the global market on loan rejection rates and criteria over time.

First, we examined the counts of rejected and accepted loan applications over time, as shown in Figure 7. Our analysis reveals that the acceptance of loan proposals has plateaued despite a steady increase in the number of applicants. A notable trend was observed between September 2013 and November 2013, during which there was a sharp spike in the number of applicants, accompanied by a corresponding increase in the rejection rate.

This trend is further reflected in the Average FICO scores of applicants, as illustrated in Figure 8. Specifically, there was a sudden 58-point increase in the FICO scores of rejected applicants during the same period. This observation suggests either an influx of higher-credit applicants or a possible shift in market dynamics affecting loan approval criteria.

To investigate this further, we plotted a time-series line chart (Figure 9) showing the average returns of each sector. These returns were calculated based on the top

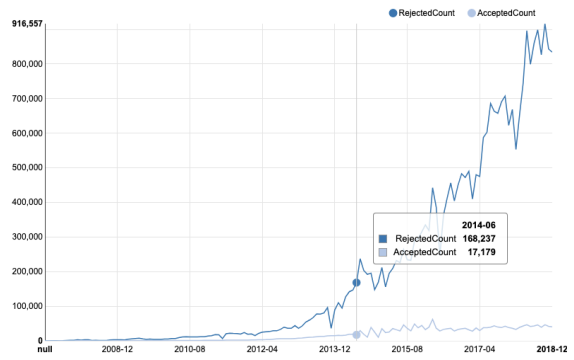

Figure 7: Trend of accepted and rejected loan applications over time. The plateau in loan acceptance rates despite an increase in applicants highlights a potential change in funding dynamics, particularly during the period of September 2013 to November 2013.

Table 10: Tweet Counts and Coverage in Analyzed Period

Metric	Count
Total Tweets	13358
Positive Tweets	6436
Neutral Tweets	5254
Negative Tweets	1668
Dates with Tweets	1013 out of 1826 (55%)

5 companies in each sector, and Bitcoin stock data was also incorporated. The analysis revealed a significant increase in Bitcoin returns during the same period (September 2013 to November 2013).

One plausible hypothesis is that, since Lending Club's funding primarily comes from investors, the spike in Bitcoin returns might have attracted investors to shift their focus from funding personal loans to investing in Bitcoin. This redirection of investments could have indirectly impacted loan rejection rates by limiting the available pool of funding for loan approvals.

On further investigation, we analyzed potential indicators for loan rejection, focusing on the influence of external factors such as the stock market and geographical distribution.

First, we examined whether the geographical location of loan applications could affect loan approval outcomes. Figure 10 illustrates the distribution of loan requests across different states. While certain states consistently request more loans than others, the rejection and acceptance trends over time show inconclusive results when correlated with market conditions or the performance of companies in those states. This indicates that geographical area alone might not serve as a significant factor in determining loan outcomes, particularly with respect to external market influences.

Next, we explored the role of the debt-to-income (DTI) ratio as a potential indicator for loan approval, while also considering its relationship with market dynamics. From Figure 11, it is evident that an ideal debt-to-income ratio lies within the range of 10–20%. However, the connection between DTI ratios and fluctuations in the stock market remains unclear. Although we hypothesize that rising market investments could influence borrowing behavior and DTI ratios, further analysis is needed to draw meaningful conclusions.

8 Conclusion

As part of the final findings of our experiments, we have correctly proved that capital markets are indeed leading indicators whereas loan disbursal is a lagging indicator of an economy. This was by comparing the future returns of the stock prices with the rejection rates of loans and they were in tandem with each other which

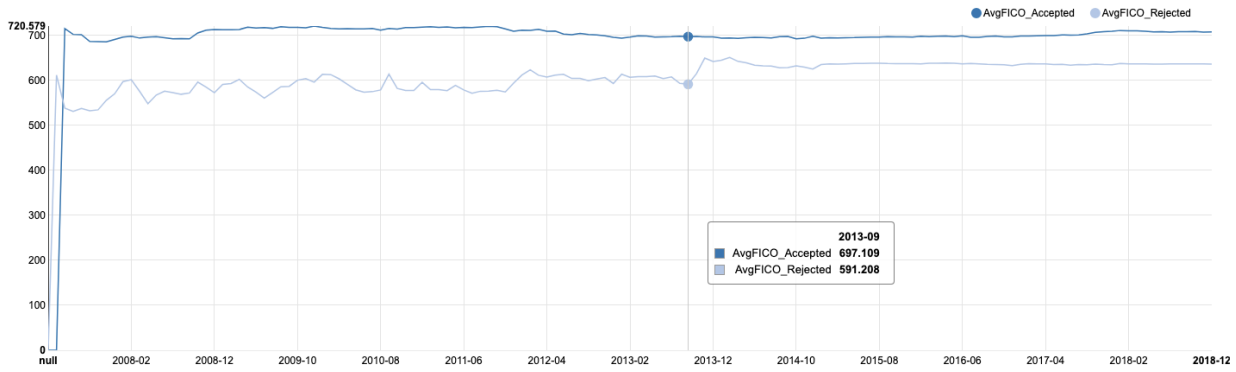

Figure 8: The FICO score requirement for loan approval has been steadily rising. Notably, during Q4 of 2011 and 2013, there was a sudden spike in the required FICO score.

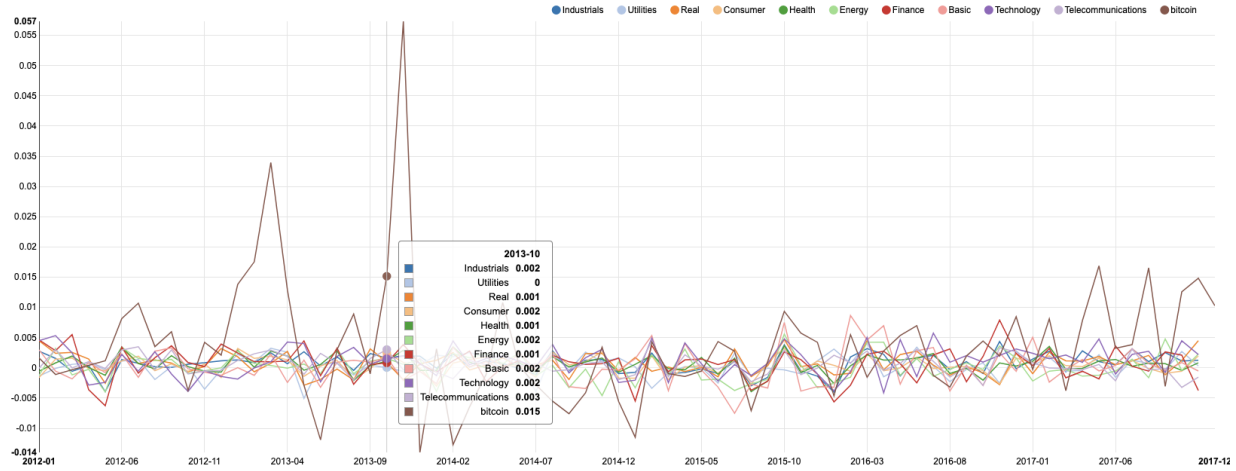

Figure 9: Time-series analysis of average returns across sectors and Bitcoin from September 2013 to November 2013. The significant increase in Bitcoin returns during this period suggests a potential shift in investor focus, influencing funding availability for personal loans.

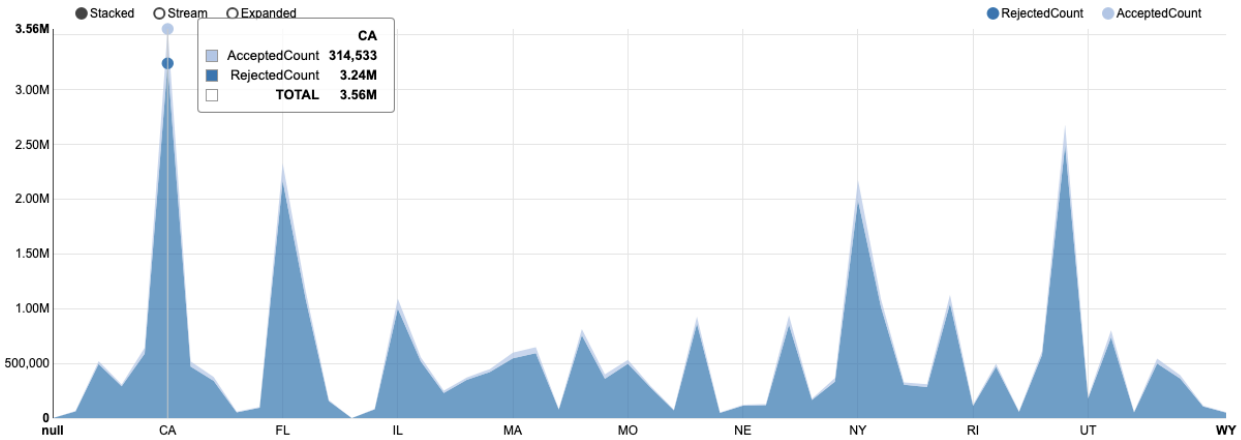

Figure 10: Geographical distribution of loan requests across states.

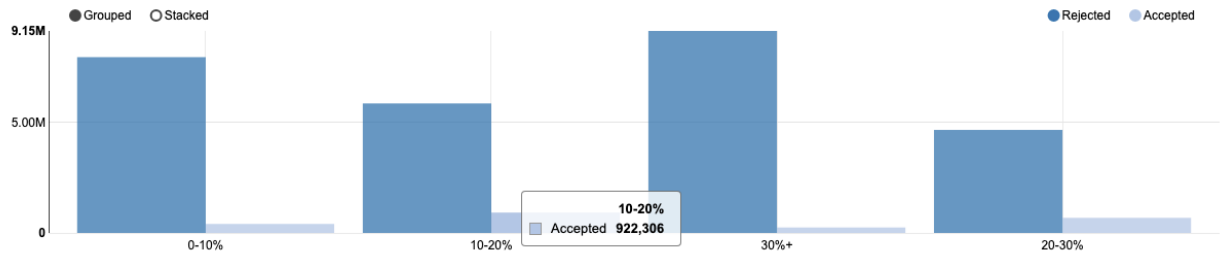

Figure 11: Analysis of debt-to-income ratios for loan applicants.

led us to believe our hypothesis was correct. Bitcoin and stock prices are highly positive correlated in cases where the particular stock has some direct or indirect investment/interest in cryptocurrencies. This further reaffirms our belief that stock prices go hand in hand with bitcoin data in cases where there is a related interest and when there is none, then there is not necessarily a strong correlation as such. Twitter sentiment data has a positive correlation with the volume of the stock traded. This might be surprising given we anticipated that prices would be affected but the data reveals that it does not significantly affect price, it does affect the volume of the stock traded. This translates to the fact that tweets and hashtags and trends contribute to the activity of a stock in the market but not necessarily the price of the stock itself.

9 Obstacles

- Linear Regression (LR) models are not ideal for time-series data analysis. Spark Scala ML does not support time-series models such as ARIMA. The existing Spark time-series model is deprecated.
- There is limited correlation statistics available, as Spark Scala Stats does not provide hypothesis tests like Granger causality or models like VAR.
- In Zeppelin, creating multiple plots (e.g., line or scatter plots) is difficult.

10 Future Improvements

- PySpark can simplify the integration of time-series models, offering a viable alternative to Linear Regression (LR) models for stock dataset analysis.
- The stock market dataset contains several gaps; therefore, scraping data from Yahoo Finance could be a useful method to fill these gaps more accurately.
- A deeper analysis of the rejected loan dataset suggests that while individual factors such as geographical location and Debt-To-Income (DTI) ratio are important for loan approvals, their direct correlation with market trends requires further investigation. Future work should focus on integrating more granular data and advanced analytical models to uncover hidden relationships between these factors and broader market dynamics.

11 Acknowledgements

Thank you to the NYU HPC for making great guides for how to use the HPC on their Google Sites webpage, and for providing the Spark cluster to conduct these analytics. Thank you to the Kaggle Collaborators to open source their data and share on Kaggle. Lastly, thank you to Professor Yang for all the support that you have provided us throughout the semester, and analyzing and approving our project idea!

References

[1] Boris Marjanovic. Price volume data for all u.s. stocks and etfs, 2021. Accessed: December 13, 2024.

[2] Omer Mertin and Mustafa Dogan. Tweets about the top companies from 2015 to 2020, 2020.

[3] Zielak. Bitcoin historical data, 2024.

[4] Nathan George. All lending club loan data, 2018.