WPI

Analysis of U.S. Job Market and Reddit Sentiment

Project Team:
Luke Gebler
Zhifei Ma
Andrew Nicklas

Project Advisor
Professor Robert Sarnie
Department of Business

Project Advisor
Professor Wilson Wong
Department of Computer Science

Project Advisor
Professor Marcel Blais
Department of Mathematical Science

Abstract

In this project, data was utilized to make more informed investment decisions as well as mitigate stock market investment risk. The initial project was to analyze a series of job search databases in an effort to make predictions about future migration trends and make more informed real estate investment decisions. Next, we worked to improve the company's existing stock portfolio risk analysis by scraping comments from the subreddit WallStreetBets and analyzing stock purchase sentiment and comparing it to historical market data using natural language processing and time series analysis.

Acknowledgements

We would like to thank everyone who made our project possible. We would like to thank our sponsor for their continued support throughout the duration of the project. They gave us incredibly valuable experience and insight into the FinTech and data science worlds. We would also like to thank our advisors: Professor Wong, Professor Sarnie, and Professor Blais. They provided us with the guidance and direction that made this project possible.

Executive Summary

Over the course of seven weeks, this project team attempted to explore two major themes using analytics and data science techniques. For the first half of our project, we explored Google Trends and employment data in an attempt to draw conclusions and make predictions about employment data. During the second half of our project, our team worked to analyze sentiment data of comments within the WallStreetBets subreddit in an attempt to draw useful conclusions about future trends in the stock market.

When exploring employment data, we focused specifically on those searching for jobs, figuring that this would begin to allow us to predict migration patterns. To do this, we utilized both Google Trends and Indeed historical data as ways to track people that were searching for jobs in different geographical areas. While we were not able to make predictions or strong correlations with this data, we did analyze a series of different employment-related searches through Google Trends, which is detailed in this report.

We also delved into sentiment analysis of comments made in the WallStreetBets forum. After comparing different options, we settled on using BERT as our language processing tool in order to obtain the perceived sentiment of different comments. After obtaining sentiment, we compared past comments to historical stock trends in an attempt to predict stock market trends and mitigate stock investment risk.

In order to ensure accountability and to most effectively use our time throughout the seven week project, we utilized the Agile Scrum methodology. After comparison with other options, we chose this methodology for its flexible structure and ability to adapt quickly with the changing goals and tasks throughout the duration of our project.

Authorship

Section	Primary Author	Primary Editor
Abstract	Andrew Nicklas	Luke Gebler
Acknowledgements		
Executive Summary		
1. Introduction	Andrew Nicklas, Luke Gebler	Zhifei Ma
1.1 Background		
1.2 Problem Being Addressed		
1.3 Goals and Scope		
2. Research	Andrew Nicklas	Luke Gebler
2.1 Domain Research		
2.2 Background Research		
2.3 Related MQP Research Projects	Zhifei Ma	Andrew Nicklas
2.4 Business and Project Risk Management	Andrew Nicklas	Luke Gebler
2.5 Math Background	Zhifei Ma	Luke Gebler
2.6 Machine Learning		
3. Software Development Methodology	Andrew Nicklas, Luke Gebler	Zhifei Ma
3.1 Agile Scrum Management		
3.2 Alternative Software Development Methodology		
4. Software Development Environment	Zhifei Ma, Luke Gebler	Andrew Nicklas
4.1 Project Management Software		
4.2 Programming Environment		
4.3 Software Tools		
4.4 Data Sources		
4.5 Summary		
5. Software Requirements	Andrew Nicklas	Luke Gebler
5.1 Software Requirements Gathering Strategy		
5.2 Functional and Nonfunctional Requirements		
5.3 User Stories and Epics		

6. Design		Zhifei Ma	Andrew Nicklas
	6.1 Software Frameworks and Architecture		
	6.2 Models		
	6.3 Data schema		
7. Development		Andrew Nicklas, Luke Gebler	Zhifei Ma
	7.1 Sprint 0: Acclimation		
	7.2 Sprint 1: Exploration of Possible Data Sources		
	7.3 Sprint 2: Looking for a correlation with Google Trends		
	7.4 Sprint 3: Google Trends Denormalization		
	7.5 Sprint 4: Monitoring/Predicting stock market using WallStreetBets		
	7.6 Sprint 5: Continued Data Collection		
	7.7 Sprint 6: Analysis and Deliverables		
8. Discussion		Zhifei Ma	Andrew Nicklas
	8.1 Computer Programs and Math Models Used		
	8.2 Data Processing and bias		
	8.3 Data observation and analysis		
	8.4 Data visualization and mapping		
	8.5 Accuracy of datasets and outliers		
9. Assessment		Andrew Nicklas	Luke Gebler
	9.1 Business Learnings		
	9.2 Technical Learnings	Luke Gebler	Zhifei Ma
	9.3 Accomplishments	Zhifei Ma	Andrew Nicklas
	9.4 Limitations	Luke Gebler	Andrew Nicklas
10. Future Work		All	Zhifei Ma
	10.1 Google Trends		
	10.2 WallStreetBets		
	10.3 Machine Learning		
11. Conclusion		Andrew Nicklas	Zhifei Ma

Table of Content

Table of Figures

Table of Tables

1. Introduction

The FinTech industry continues to cement itself as a growing and important field within the financial sector. As more of the daily processes of many financial institutions become automated, the reliance on technology and data-driven decisions continues to grow. Companies within these sectors have had to rely on more and more data in order to make informed decisions about the future and maintain a competitive advantage.

This reliance on data is especially prevalent within the private investment industry. When it comes to real estate, it is crucial for firms within this industry to be able to identify and accurately predict trends by analyzing large amounts of data. Within the real estate industry, it is important to find ways to use data in order to predict possible future migration trends.

One strategy of predicting these trends would be through the analysis of employment data. If there was a clear and accurate way to see both where there were open employment opportunities and where the most candidates for certain jobs were, it was believed that accurate migration predictions would be made. We gathered data from a number of different sources that people would utilize when searching for a job. Despite thorough investigation of this data and an attempt to draw correlations and trends between the employment data or job search-related data that we were able to capture and migration patterns, we were unable to make any significant correlations.

This led us to begin investigating if we could discover trends in the stock market, and keep the investment firm away from potentially riskier investments through analyzing public sentiment around a set group of stock holdings. The increasing popularity of investment forums and discussion boards has provided a large database of comments that may be able to provide insight

into the public sentiment surrounding certain stock tickers. This insight surrounding sentiment, we believe, could provide private investors with increased risk mitigation strategies, as well as an improved investment portfolio. These sentiments were analyzed through the training and deployment of a natural language processing tool to find a correlation between these sentiments and historical stock data.

1.1 Background

Our group worked with an alternative investment firm with assets spanning across multiple continents. While maturing as a company over the past decades, they have invested in a multitude of areas including corporations, foundations, pension funds, endowments and more. To acquire these billions of dollars in assets, they use data driven analytics to determine currently undervalued items as well as predict future market trends to determine the best candidates to invest in.

Within the company, a range of financial technology is utilized in order to make wise investment decisions. This includes predictive models, historical analysis, and risk management, with much of their success coming from noticing new trends in real estate before others have a chance to invest.

1.2 Problem Being Addressed

Our project worked to analyze trends and answer two separate problems. At its onset, our project attempted to predict the migration of job patterns using denormalized Google Trends data. After we were able to denormalize a series of different job-related search terms through Google

Trends, we used machine learning techniques in order to predict the migration of different jobs and compared this against the firm's given job databases.

We also used comments throughout Reddit's WallStreetBets forum as predictors for the changes and possible future trends of stock prices to give the firm better return of future investments.

1.3 Goals and Scope

Our project adapted to a series of different goals and deliverables throughout the course of the process. As we began to explore different databases and sources of information, we were able to change the scope of our project in order to make more informative predictions for the investment firm.

Using tools given we analyzed and visualized trends within multiple large datasets with hundreds of thousands of entries in order to gain insights into existing and potential investments. Furthermore, it was important to be able to verify that all the alternative data provided by the company is accurate and large enough of a sample in order to make real estate investment decisions.

In order to perform this analysis, we used a multitude of different software and data science techniques. Data science techniques including linear regression and machine learning were utilized to make predictions.

2. Research

2.1 Domain Research

There are many clear differences between alternative and traditional investments. Alternative investments are commonly rather illiquid and less regulated than their counterparts. Traditional investments (stocks, bonds, cash) can be easily converted into cash and are therefore considered liquid investments. Alternative investments are illiquid, meaning they cannot be easily converted into cash (Chladek, 2020). Due to the inherent nature of these investments having fewer buyers than traditional investments, a few issues can arise. Namely, investments can have a lack of performance data. This not only makes it harder to get investors involved, but can also cause issues with the valuation of the investment itself.

While real estate is still considered an alternative investment, it still clearly stands out from other forms of alternative investment due to its massive market. With its overwhelming popularity, it can create cash flows that may make it more similar to a bond or a more liquid asset, but it is still considered an alternative investment (Chladek, 2020). Around half of the company's assets are based in real estate, making it the biggest revenue generator for them. Given the large amount of risk coming with real estate, it is important to be extra vigilant in tracking trends in human activity and being able to make predictions on what will be more valuable one, five, and even ten years down the line. This is where the majority of our research went into and it was critical for us to be able to make accurate predictions in order to help real estate investment strategies.

2.2 Background Research

2.2.1 Google Trends

Google Trends is a dataset aggregated by Google to show the popularity of different search requests made to Google. It is largely unfiltered and can show both real-time data and historical data from as far back as 2004 (“FAQ about…”, 2021). The data comes from a representative sample of Google searches and is normalized in an effort to make comparisons between different search terms simpler. While this caused some frustration throughout our project, this normalization does make search terms easier to compare.

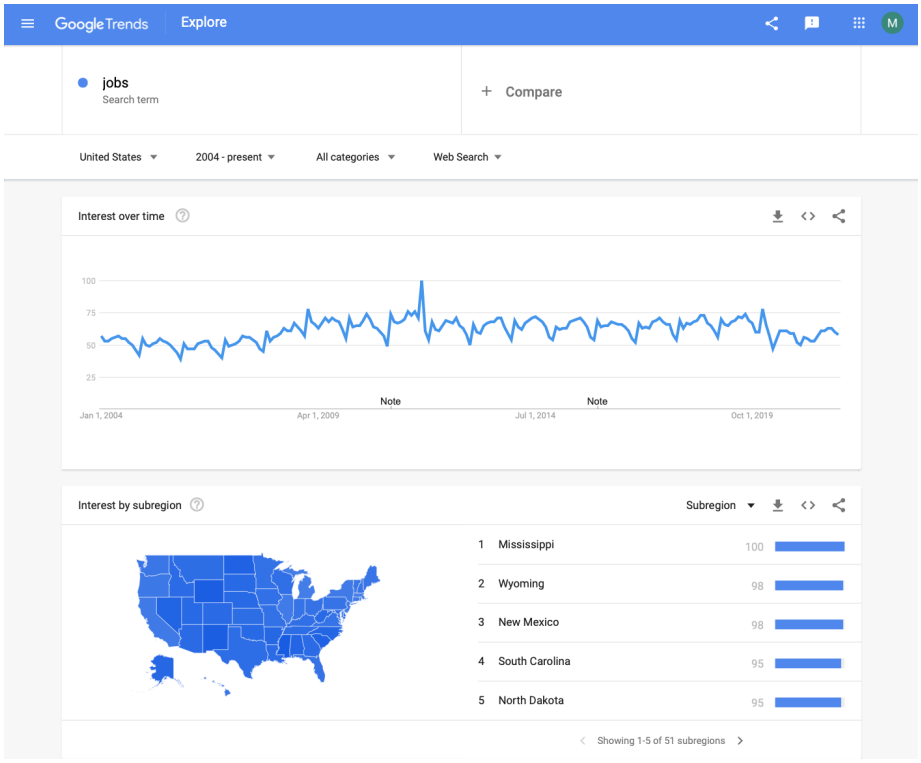

Figure 2.1. Google Trends results with search term “jobs” over time across different states

As detailed in the figure above, Google Trends normalizes its data by first dividing each data point by the aggregate number of searches within either the geographical or chronological

ranges that are specified. This is intended to normalize it so that it can be visualized by its proportional share of the searches, rather than always just giving preference to the places with the highest search volume. The resulting proportions are then scaled on a range of 0 to 100, with the region having the highest proportion of searches being indexed to 100. While this made comparing searches within matching geographical areas incredibly convenient, it made comparisons across different searches or comparisons to any outside datasets near impossible.

2.2.2 WallStreetBets

WallStreetBets is a forum in which people can discuss investments, trading and stocks. It is self-described as “like 4chan found a Bloomberg terminal”(“wallstreetbets”, 2012). Founded in 2012, it recently gained noticeable traction throughout the COVID-19 pandemic as people looked to gain more knowledge and familiarity with investing and the stock market. This following allowed it to become a fairly strong player within the market, and something that firms had to be wary of when monitoring their investment portfolios. A good portion of the forum’s popularity came from its ability to cause short squeezes with meme stocks, as detailed in the next section.

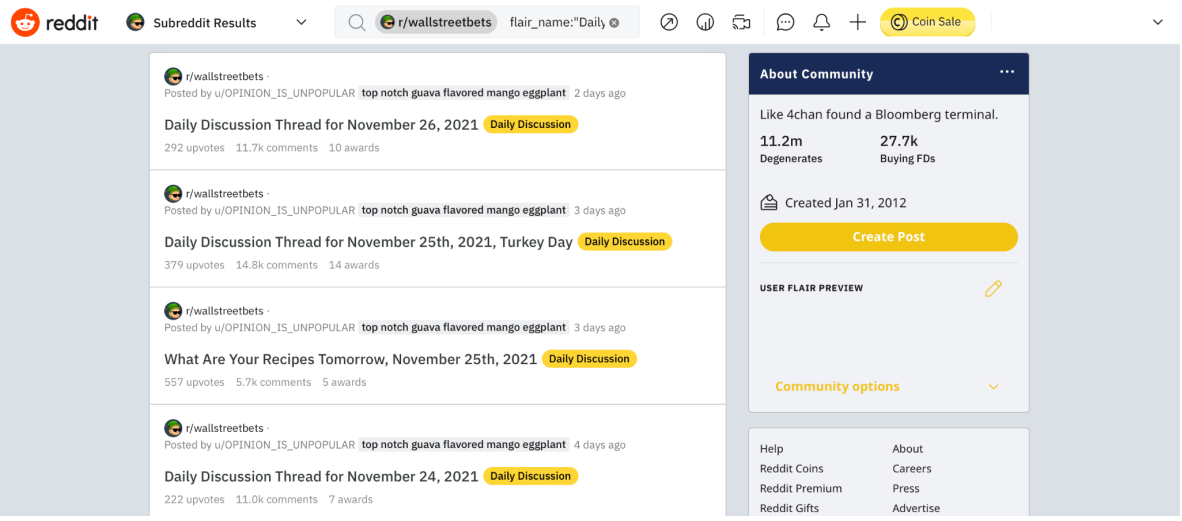

Figure 2.2. WallStreetBets Daily Discussion

2.2.3 Meme stocks

Meme stocks are a select group of stocks that have their stock price greatly affected by sudden internet popularity (Gravier, 2021). This high trading volume usually causes their stock price to skyrocket, allowing some traders to make a lot of money and also results in a short squeeze. A short squeeze is when a security or stock that has an unusually large number of short sellers is targeted in an effort to increase its prices dramatically (Mitchell, 2021). Short sellers are people that hold short positions on the stock, meaning that they are predicting a decline in the stock's price and will lose money if the stock's price increases (Hayes, 2021). The increase caused by the target of a particular stock results in heavy losses for those holding the short positions, especially firms with large amounts of holdings. The most well-known example of this event is WallStreetBets with AMC and GameStop in early 2021, where GameStop saw its stock jump 134% in a single day and more than 2,000% over a period in January due to concerted efforts by online investors (Schneider, 2021). This dramatic increase and stock price volatility is demonstrated in Figure 2.3.

Figure 2.3. Changes in Stock Price of GameStop Corp. (GME)

2.2.4 Time Series Analysis

Time series analysis is a type of analysis which looks to examine data over particular intervals and perform trend analysis for this data (Statistics Solutions, 2021). This data examines a sequence of data points over a period of time (Hayes, 2021). This analysis is especially useful when examining financial or migrational data, as it can lead to the discovery of trends over a certain period of time. This can also help with predictive analysis, as the past periods of time can be used as a tool to predict trends for future time periods. It can be used for analysis in different ways, including both to understand the underlying trends or forces which affect the examined model or to fit a model and make forecasts (Natrella, 2003). For our project, we used time series data within both the Google Trends and WallStreetBets of our project for both of these uses.

2.2.5 Sentiment Analysis

Sentiment analysis is a text classification tool used to quantify the emotion behind a given blurb of text. Through the use of machine learning and natural language processing techniques, the model is able to output a score from 0 to 1 where 0 represents a negative comment while 1 signifies a very positive comment. While this can add a new depth to your data, it comes at the price of computational power. Sentiment analysis is often built on the back of hefty resource consumption and is impossible to complete without the proper amount of time and hardware due to the computationally intensive algorithms required for the analysis. It works to dissect the text, while identifying and extracting information from text and helping a business to understand public sentiment (Gupta, 2018). While this can be useful for a business in numerous ways, there is some disagreement as to whether or not it can be useful in determining long-run market trends (Brown & Cliff, 2001). For our project, we used sentiment analysis in an effort to draw conclusions and make predictions about potential stock market trends.

2.3 Related MQP Research Projects

2.3.1 Summary of Previous MQP Paper: Analyzing Migration Trends through Credit Card and Foot Traffic Data Using Machine Learning Model

The MQP group trained a machine learning model for predicting human migration patterns across the United States. They analyzed and validated different data sets using Azure Databricks Notebook and the programming language Python. They used the population of each of the top two hundred counties in the United States from 2018 to 2019 from US Census and the number of bars and full service restaurants from 2018 to 2019 from SafeGraph as inputs. The output was a prediction for the population of each county based on the number of bars and restaurants recorded in SafeGraph in 2020.

The group first focused on understanding, visualizing, normalizing, and combining data from SafeGraph and credit card transactions data from Yodlee. Then, the project team learned to document and categorize notebooks for reproducibility and started using the Pyspark Plotly framework for visualizing a correlation they found between the population of Texas and offline credit card transactions at a specific supermarket in Texas called HEB. After that, the MQP group kept focusing on HEB and found a strong correlation between the number of unique HEB locations in SafeGraph and the population both by county from 2018 to 2019. Next, they created a machine learning regression model with SkLearn to predict the population of states by county by utilizing the existing linear correlation between populations from the U.S. Census and the unique number of grocery stores stored in SafeGraph with a confidence interval above 92%. Finally, they applied the model to the top 200 counties in the US with a reasonable mean squared error demonstrating the high accuracy of their Random Forest Regression model.

2.3.2 Summary of Previous MQP Paper: Tracking U.S. Migration Patterns of People in Result of the COVID-19 Pandemic

The MQP group developed a Jupyter Notebook and Power Business Intelligence dashboard to provide visuals of migration patterns from March 2020 to November 2020 in the United States, it included the most popular areas where people are moving into and out of. They utilized filters and animation to show key demographics like age and median income of the population movement. The displayed migration patterns helped sponsors to strategize where to invest. By using census block group data, they predicted foot traffic at different home improvement and appliance stores within the U.S.

The group started by viewing each Home Depot within the United States to measure foot traffic in each census block group and using median income to understand money movement patterns. Then, they expanded to a variety of commercial areas that people are likely to visit when moving to new locations and considered greater distances from residential locations to points of interest, like grocery stores, all presented on a geometric map. After that, the project team made their deliverables more interactive and dynamic by using a heat map with darker color to indicate greater movement densities and adding arrows for showing movement patterns in the US, or even within specific cities. Next, additional data was added to create a multidimensional map for analyzing movements based on demographics like age, income, and mile range. At the end, the MQP group drew insightful conclusions from the filtered data on the map and refined their deliverables.

2.4 Business and Project Risk Management

Our project aimed to increase the bottom line for the company by providing key insights into potential trends in a number of different databases that were provided to us by the private investment firm. There were significant risks to consider when undertaking a project with this amount of data. If the project were to be unsuccessful, it could present a significant loss in both efficiency and possible profits for the firm. If the cleaning of the data is not completed thoroughly and correctly, there could be significant trends in the data and company decisions made based on these trends that could be either slightly or even majorly incorrect. This could have a significant effect on the company's bottom line.

If done properly, the project aimed to both directly and indirectly increase revenue for the company. First, properly cleaning and providing beginning analysis on these large databases can directly reduce costs for the company. Also, proper data and accurate analysis can lead to decisions by the firm that will indirectly lead to increases in the company's bottom line.

Through this project, we hoped to further the firm's knowledge in possible migration trends and migration data, through analyzing both Google Trends and GDelt data. We also aimed to assist them in their analysis of job listing and merger data in order to fully and accurately work to predict possible future migration of people.

As we experienced throughout the duration of our project, the firm had a series of different groups and processes that were implemented in order to ensure proper risk mitigation. Some risk is inherent within any series of investments, but the important thing was that the firm wanted to maintain that they were taking the right amount of risk. One of the analogies that was stressed to us was the analogy that investment was similar to driving a racecar: take too much risk and you

crash, too little and you have no chance of winning. To ensure they have proper risk analysis techniques, the firm implemented a series of different strategies.

Risk mitigation existed throughout the firm at both the company level, and throughout different sectors. Within individual sectors, there was constant reporting and predictions to ensure they were never exposing themselves to unnecessary risks. They also utilized a multitude of different alternative data sources in order to ensure they are taking the proper amount of risk. Acknowledging the importance of having risk mitigation at the sector level within the company, the importance of having a separate group at the company level that is able to address risk was also stressed. One of the main reasons for this was to manage the concentration risk throughout the company. Concentration risk is the exposure to an excessive amount of risk by becoming too dependent on a single investment or area (Miratech Holdings, 2020). With a group managing company-wide risk, they were able to ensure the mitigation of the risk by making sure the firm is never investing too much in one company, sector or country.

The existence of this risk mitigation group also allowed the company to ensure that they would be able to sustain the impact of a multitude of different potential shocks. Simulating macroeconomic shocks such as inflation, impacts on the supply chain, and price increases allowed the risk management group to ensure that they are not overleveraged in any particular department and do not expose themselves to any unnecessary risk.

2.5 Math Background

To find correlations between different data sets, we looked into linear regression. We also utilized machine learning, which is the field of predictive modeling that is concerned with minimizing the error of a model and making the most accurate predictions possible. Linear regression was developed in statistics for understanding the relationship between input and output numerical variables (Brownlee, 2016).

Assume we want to compare two data sets each with n number of data points, we evaluate R-squared, a statistical measure of how close the data are to the fitted regression line, which indicates how strong two data sets are correlated with each other. It's a number between 0 to 1(100%). In general, the higher the R-squared, the better the model fits the data (Minitab Blog Editor, 2013). Practically, 0.3 correlation is decent and meaningful enough for economical data compared to 0.9 in the field of physics and engineering. R-squared is calculated by the following equation (1) where $SSregression$ represents sum of squared regression error and $SStotal$ represents sum of squared total error, $\bar{x}$ represents sample mean and $x_{regression}$ is the regression value correspond to each data point x_i :

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - x_{regression})} \qquad (1)$$

Besides R-squared in linear regression, Z-score also helps to interpret how interest scores from Google Trends can be compared with other data sets. Z-score interprets how far each data point is from the mean value of the entire data set. It measures how many standard deviations below or above the mean a data point is. A Z-score can be placed on a normal distribution curve. Z-score ranges from negative 3 standard deviations up to positive 3 standard deviations (Glen, n.d.). We computed Z-score using equation (2) where σ represents the standard deviation, which measures the dispersion of a dataset relative to its mean:

$$z_i = \frac{x_i - \bar{x}}{\sigma} = \frac{x_i - \bar{x}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}} \tag{2}$$

One difficulty we encountered when comparing potential redundant data sets is collinearity, which describes the problem when two or more predictor variables are highly correlated with each other, then it is difficult to separate their effects on the response (Ruppert, 2004). The Variance Inflation Factor (VIF) is a common tool to detect collinearity. A VIF of 4 indicates that the standard error is increased by a factor of 2. Calculating VIF using equation (3) is also related to the calculation of R-squared in equation (1):

$$VIF = \frac{1}{1 - R^2} \tag{3}$$

The last important math concept that will be used in this project is the confidence interval (CI), representing a range of values that is likely to include a population value with a certain degree of confidence, usually expressed as a percentage where a population mean lies within a lower and upper interval (McLeod, 2019).

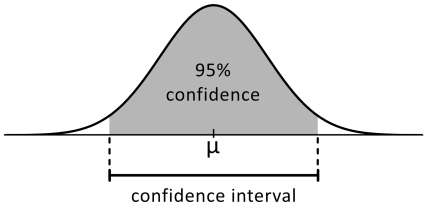

Figure 2.4. Distribution of sample mean around population mean with 95% CI (Szyk et al., 2021)

A 95 percent confidence interval corresponds to a Z-score of about 1.96. The actual interval can be calculated using equation (4):

$$CI = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} \tag{4}$$

2.6 Machine Learning

Machine Learning is defined as the capability of a machine to perform complex tasks which emulate intelligent human behaviors (Brown, 2021). Specifically for this project, we utilized machine learning for estimating a number and understanding a text written in natural language.

Types of machine learning techniques that the project team applied are Sklearn linear regression framework and BERT natural language processing model. They both belong to supervised learning algorithms, in which the machine analyzes a training dataset, applies what has been learned in the past, and produces an inferred function or measurement to make predictions as output (Expert.ai Team, 2021).

2.6.1 Sklearn Linear Regression Framework

We employed one of the supervised learning algorithms from the Sklearn Framework, particularly the ordinary least squares linear regression method. It fitted a linear model to minimize the residual sum of squares between the observed data and the predicted data generated by the linear approximation.

2.6.2 BERT Natural Language Processing Model

BERT stands for Bidirectional Encoder Representations from Transformers. As described by the name, instead of reading text input sequentially, either left-to-right or right-to-left, a BERT model reads in both directions at once. It helps computers understand the meaning of an ambiguous text by analyzing surrounding text to establish the most precise context (Lutkevich, 2020).

3. Software Development Methodology

3.1 Agile Scrum Methodology

Agile scrum methodology is a project management tool used to ensure accountability and effectively manage the different tasks assigned throughout a project. Scrum is a sprint-based approach where tasks are created each week and then distributed out to each group member for them to finish within the sprint. It is especially useful for projects that continuously adapt by dividing many of the agile development methods into smaller tasks ("Agile - Characteristics", n.d.). This adaptability works well with data science driven projects, as the results gathered during the project can have large impacts on how the rest of the project will be completed.

User stories are the main tool used in the agile methodology to convey what needs to be worked on. These consist of a theoretical user who is asking for a specific feature or fix to the product. These stories both provide the team members with tasks to complete as well as give the reason behind implementing that task. These user stories usually include tasks. The user story details the needs of the user and talks about what needs to be done. It does so with the form: As a <type of user>, I want to <some goal> so that <reason for goal>. The task discusses how the particular user story will be implemented. These tasks are given an estimation of time, usually in hours. For user stories, time is tracked by story points. These points define how much time should be committed to a certain story. Stories are defined as done when all tasks are completed, no defect is open and the product owner has accepted the user story (Rehkopf, n.d.).

This process also includes a number of different roles that team members must fulfill throughout the agile process, including the Scrum Master and the Product Owner. The Scrum Master is tasked with ensuring cooperation between all roles and functions and removing any

blocks or disturbances that may occur through the agile process. They also coordinate with the organization and ensure that all meetings and stand-ups are leveraged properly. The Product Owner works to drive the business perspective of the project. They are tasked with defining and prioritizing the requirements. They also want to make sure to be the voice of the customer and take an active role in iteration and release planning (Cprime, n.d.). The product owner is also responsible for maintaining and creating the sprint backlog, where all the future user stories are stored. This serves to both indicate the future direction of the project, and ensure that group members always know what the future stories to complete will be.

Daily stand-ups are daily meetings among all members of the team ("Agile - Daily Stand-up", n.d.). These meetings ensure accountability by reviewing what the team had accomplished in the previous day, as well as detailing what they will be completing throughout the upcoming day. These meetings also served to ensure that groups are able to maintain focused progress by reviewing any impediments they might have been facing and discussing ways to address them. Finally, these meetings serve to ensure that groups are all on the same page when it comes to the direction and the progress of the project.

Along with daily meetings throughout the sprint, the agile methodology also requires three different meetings for every sprint. The sprint planning meeting, at the beginning of each sprint, as well as the sprint review and retrospective meetings at the end of each sprint. These meetings work to ensure that the group is fully utilizing the scrum methodology, as well as constantly adjusting and adapting within the project.

The sprint planning meeting sets the expectations for the sprint. In it, the group works to define the what, the how, the who, and the inputs and outputs of the upcoming sprint. The what is defined mainly by the product owner and details exactly what the group will work to accomplish

throughout the week. The how falls on the development team to define the work necessary to accomplish the goals of the sprint. The inputs and outputs clearly outline what the sprint will need to be successful and what exactly the group aims to produce or accomplish throughout the sprint (West, n.d.). If this meeting is done properly, it allows the team to have a sprint that is well-organized, efficient and successful.

The second type of meeting that agile utilizes consistently is the retrospective. Retrospectives serve to satisfy the final principle of the agile manifesto: "the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly" (Beedle et al., 2001). This is done through meetings after every sprint, where you discuss the sprint, how it went, and what could be improved upon for the next sprint. These serve to ensure that the group is constantly communicating and improving upon their processes, something stressed throughout the duration of the agile process (Atlassian, n.d.).

Agile also contains a sprint review meeting at the end of each sprint, which differs from a retrospective meeting. Sprint reviews focus more on the work of the team throughout the sprint. These are a time for team members to gather, describe and show the work they have been able to complete throughout the week (Radigan, n.d.). Although sprint reviews are much more casual than the more formal retrospective, they are as necessary to the Agile Scrum process.

3.2 Alternative Software Development Methodology

SDLC or Software Development Life Cycle is another approach to creating software. This methodology focuses on a structured approach to development with five main stages: planning, designing, developing, testing and deploying. SDLC is commonly used in large application building for companies around the world and is a proven methodology for developing high

quality software efficiently (Pedamkar, n.d.).

While SDLC is a proven methodology to developing software, it does have drawbacks as well. SDLC focuses on a much more structured approach to development which would not bode well for our project. Agile allows for us to constantly adapt to new information and implement within the next week, or sprint, cycle (Altvater, 2021).

The earliest approach that was designed for SDLC in software development was the waterfall model (also known as the linear-sequential life cycle model). In this approach, the process is divided into phases as detailed in Figure 3.2. Each phase must each be fully completed before moving to the next phase, cascading into each other as the process flows steadily downward.

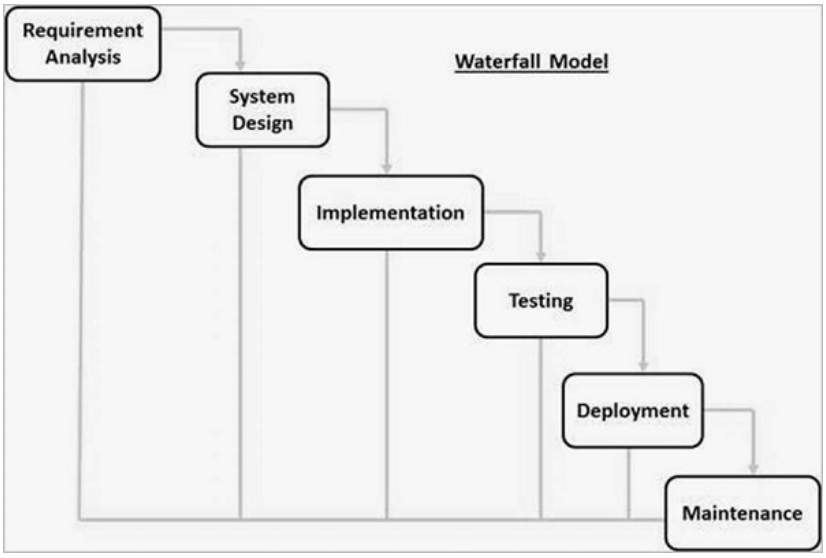

Figure 3.1. Example of the phases of the Waterfall model (Tutorials Point, n.d.)

This approach has several advantages, including its allowance of departmentalization and control. It is also fairly simple to understand and has clearly defined stages with well understood milestones that allow for easy management. The first stage is requirement analysis, the backbone of the project. You first gather information from the user on what needs to be implemented and

then confer as a team as to whether or not it is feasible, making sure the end product is what the user wants it to be. This is followed by system design, a zero code approach to creating the initial mockup of what the product will look like. This stage includes the design of user interfaces, databases, the network and more without any implementation. After system design, the system is divided into small modules which are used in the next phases. Implementation focuses on turning the agreed upon system design into working code. After implementation is complete, one moves on to testing to make sure each requirement has been met. Once this is completed, the system is deployed. After it has been deployed, the system must go through regular maintenance to ensure it continues to function properly ("SDLC - Waterfall Model", n.d.).

With these advantages, however, come several disadvantages. The main disadvantage of the waterfall approach is that it does not make it easy to have reflection or revision. This introduces several large disadvantages, including a high amount of risk and uncertainty. This also makes it a very poor model for long, ongoing projects or for complex projects with moderate to high risk of changing as it cannot accommodate changing requirements. It is also extremely inflexible. Before the group can continue to the next phase, all previous parts of the phase before must be complete; if the group ever wants to go back to a previous phase, they must go back to that phase of the waterfall process and start over.

4. Software Development Environment

The sponsor gave us the freedom to "choose our own weapons", so we selected software packages and tools that we were already familiar with, like Trello and Github, along with the platforms that are provided by the private investment firm, including Azure Databricks and Power BI, together with what the past MQP groups have used, such as Jupyter Notebooks and the Sklearn framework.

4.1 Project Management Software

4.1.1 Trello

We kept track of all the project's tasks on Trello. It is a web-based, Kanban-style, list-making application similar to an actual Agile development scrum board. Trello is also a subsidiary of Atlassian (Wiggers, 2019). This project management tool is designed to help visualize work, prioritize tasks, and maximize efficiency (Rehkopf, n.d.).

4.1.2 Github

For version control, we used GitHub repo to keep track of collaborative works, including scripts, graphs, and CSV files. Since each file on GitHub has a history, it was easy to view the changes that occurred at different time points (Daskalova, n.d.). We were able to assign issues to different programs which clearly identified each person's responsibility and also helped to achieve a balanced workload for everyone. Besides navigating among different versions of the files, we also maintained online backups for the project.

4.2 Programming Environment Including IDE

4.2.1 Microsoft Azure Databricks 8.2

Microsoft Azure Databricks is a SQL database based on Apache Spark that is designed for pivoting over big data. It is a managed platform where data scientists and programmers focus on generating insights from datasets and query tables, without worrying about other tasks like managing Databricks clusters, libraries, dependencies, and upgrades (Kennedy, 2021). We utilized the online user interface to access data stored in Microsoft Azure DataLake. We also learned PySpark, an API framework developed to work with Apache Spark for programmers to easily manipulate large data sets (Desmond, 2021).

4.2.2 VSCode 1.62 and Anaconda 2021.05 Individual Edition for Python3

For programming tasks outside of PySpark Notebook, we used Visual Studio Code and Anaconda as IDEs. Our team's programmers have had extensive experience using this development environment. VSCode provides user-friendly extensions including Microsoft's Python extension, Azure Data Lake Tools, and Azure Machine Learning, while Anaconda offers easy access to Jupyter Notebook.

4.2.3 Pandas Python Framework 1.3.3

Besides Apache Spark's DataFrame object, we also used Pandas DataFrames for more efficient data visualization tools including Plotly Express API and Matplotlib library. Furthermore, there are more Pandas DataFrames examples and tutorials online, since it is the most widely used open source package for data analysis and machine learning tasks. It was built on top of the Numpy package (ActiveState, n.d.).

4.2.4 SkLearn Python 0.24

We utilized the SkLearn framework for building the linear regression model. We retrieved job data through Bright Data from Indeed We gathered public interest scores by state using Google Trends. Then, we interpreted R-square results given by the linear regression model to decide whether we should keep investigating the relationship between different datasets.

4.2.5 pyTrends 4.7.3

Instead of manually downloading csv files from Google Trends Interface for every word and for every specific time range, we looked into pyTrends, an unofficial API for Google Trends, which allows automating downloading of interest scores from Google Trends (General Mills Inc., 2016). It provided hourly, daily, and weekly data, along with options for geological locations: country, state, or metro level interest scores.

4.2.6 Praw 7.5.0

Praw is the official Python Reddit API Wrapper for obtaining a very large number of comments from Daily Discussion in subreddit WallStreetBets. It requires a Reddit account to create Client ID, Client Secret, and User Agent, which is a unique identifier that helps Reddit determine the source of network requests (Boe, n.d.). PRAW supports Python with version 3.6 and above. After grabbing data using PRAW, we put the comments into JSON files and used natural language processors to discover sentiment and trends, and eventually recorded aggregated results in a dataframe (Minot, 2021).

4.2.7 spaCy 3.2.0 vs. NLTK 3.6.4 vs. BERT

We compared several natural language processing (NLP) tools. First, spaCy is an open-source software python library used in advanced natural language processing and machine learning, mainly for information extraction, natural language understanding systems, and preprocessing text for deep learning (Singh, 2020). NLTK is also one of the most popular NLP

tools available in Python. NLTK was built by scholars and researchers as a tool to create complex NLP functions; it is essentially a string processing library for education purposes. In comparison, spaCy takes an object-oriented approach that each function returns objects instead of strings or arrays, and it runs faster since it is written in Cython, thus favored by app developers (Kakarla, 2019). We ended up training a BERT model for sentiment analysis due to its fast fine-tuning characteristic. As opposed to directional models, which read the text input sequentially from left-to-right or right-to-left, the BERT model is bidirectional because it reads the entire sequence of words at once, allowing it to learn the context of a word based on all of its surroundings, thus generating more accurate results (Horev, 2018).

4.3 Software Tools

4.3.1 Power BI 2.97.801.0 vs. Kepler GL 0.3.2 vs. Excel

When plotting a data set on a US map, for example, Google Trends interest scores for the search term “software engineer” by state, we mainly use Power BI, a collection of software services which turns data into a form of coherent, immersive, and interactive visualization (Hart et al., 2021). The other tool we utilized is Kepler.gl, a web-based application for visual exploration of large-scale geolocation data sets. It can render millions of data points representing thousands of locations and perform spatial aggregations while in progress, which is favored by companies like Uber and Airbnb (He, 2018). Power BI became the main deliverable because the private investment firm preferred using it for business analysis and visualization.

4.3.2 Jupyter Notebooks 6.4.4 vs. Databricks Notebook with Apache Spark 3.1.2 Scala 2.12

We performed analysis and sql queries using Databricks notebook, a web-based interface that can combine runnable code, visualizations, and narrative text all together (Perla et al., 2021).

Similarly, when working on our local desktop, we used Jupyter Notebooks, a free and open-source web tool, which contains software code, computational output, explanatory text and multimedia resources in a single document (Perkel, 2018). Both notebooks are helpful for presenting analysis results to non-programmers.

4.3.3 BrightData

BrightData is an automated online data collection platform that helps companies to access up-to-date data and develop business insights by data analysis. We focused on grabbing data from career websites including Indeed, LinkedIn, and Monster.

4.3.4 Zoom 5.8.3 and iMovie 10.3

Zoom recording was the easiest way to create our user guide videos, since we were already familiar with the software from remote meetings. All post-processing was done in iMovie, a built-in video editing application on Apple devices.

4.3.5 Webex 41.10.0.20395, Microsoft Teams 1.4.00.26376, and Discord

We completed daily standups using Webex with the sponsors and asked pop-up questions in group chat on Microsoft Teams. For internal real-time communication between group members, Discord provided a casual setting for interacting and reaching out to each other instantly. Discord voice channel allowed us to have convenient and instant daily meetings.

4.4 Data Sources

4.4.1 LEHD

LEHD is a data source provided to us which lists the total number of jobs by zip code within the US. This data covered from 2008 until 2017 and was used to cross reference with other historical data.

4.4.2 Google Trends

Google Trends is an online resource provided by Google where the user can query up to five search terms within a specified region of the world over a given time period. The output is a normalized dataset where the period with the maximum search volume across all terms is given the value of 100 and the rest is scaled accordingly. This data source was the main focus of our project for the first five weeks. The search terms we used were: jobs, jobs hiring, jobs near me, hire, indeed jobs.

4.4.3 GDelt

GDelt is a periodically updated data source provided to us which contains articles from many different media outlets along with a theme and tone analysis on each. This data allowed us to track the media's tone on a given topic over a certain time period which was used to cross reference with Google Trends data.

4.4.4 BLS

The BLS or Bureau of Labor Statistics is a public US government database with a plethora of labor data. Our project mainly focused on the employment and unemployment data provided which lists the rate of unemployment by month since 2001.

4.4.5 Indeed

Indeed is an employment website for job listings which has been running since 2004. Indeed recently decided to start releasing data to the public which listed the change in job postings since February 1st of 2019 until present.

4.4.6 Reddit

Reddit is a social media platform which covers all topics on the internet. For our project we focused on the subreddit WallStreetBets, a public forum which mainly discussed the stock

market. This forum rised in popularity during the COVID-19 pandemic, growing tenfold with its current user base at 11 million members. Our project focused on their daily discussion posts where each user would talk about their plans for that day of trading. We were able to aggregate that data and run sentiment analysis on it in order to track current day trading.

4.5 Summary

We chose the Python Pandas dataframe(df) as the standardized data structure. We also developed a script for filtering out useful data from BrightData, but the official data published by Indeed seemed more promising so we ended up abandoning BrightData as an option. We started to focus more on comparing other data sets with Google Trends. Instead of downloading a single CSV file each time we searched a term and decided to only look at metro or state-level areas, we utilized pyTrends for acquiring large scale Google Trends data.

We used linear regression for mapping between two datasets to show how strongly they are correlated with each other. There were no scaling coefficients needed for the linear regression. We were simply trying to validate whether one dataset was an indicator of another, for instance, whether Google Trends was an indicator for U.S. job market and hence correlated to migration pattern. We created a denormalization algorithm to undo the effect of Google Trends normalization in order to conduct spatio-temporal analysis using search behavior by applying the equation in figure 4.1 (Memon, Razak, and Weber, 2020).

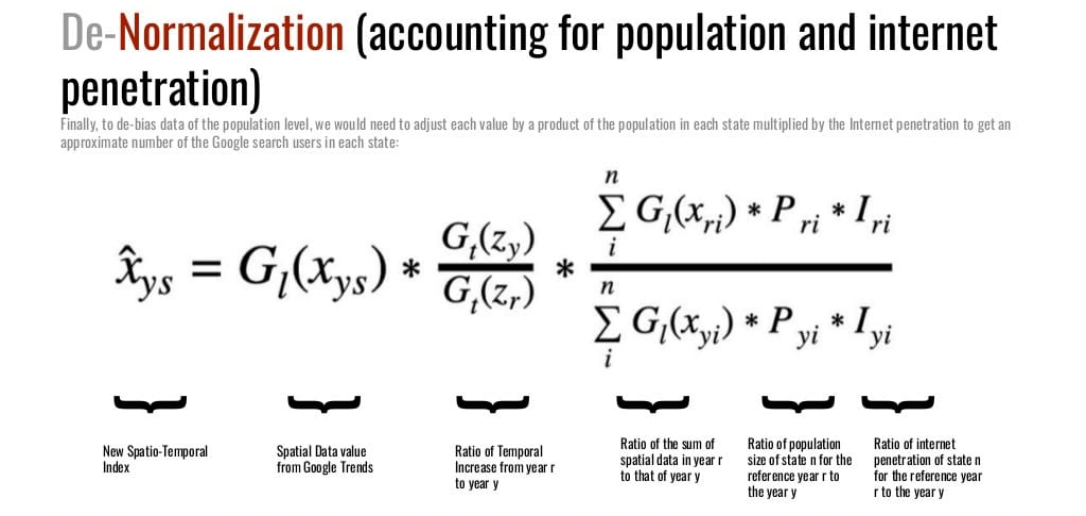

Figure 4.1. Equation for Google Trends Denormalization (Memon, Razak, and Weber, 2020)

We first denormalized Google Trends interest scores across temporal data. For example, Mississippi with 100 interest score in 2020 is not equivalent to Mexico with 100 in 2021 because the scores are calculated with respect to the unique population size and time period. To remedy this, we calculated the factor based on temporal data and the general interest in the US over time, then multiplied these factors to each column, while making sure the interest scores are scaled over multiple years. We then denormalized the processed interest score using geological data by state. Google normalized raw search numbers according to local populations, so that states with lower populations would not be underrepresented. Therefore, we calculated the factor based on the region with the lowest population: Wyoming with 576,851. For comparison, the population of California is about 66 times that of Wyoming, so we applied a factor of 66 to Google Trends interest scores. These population factors, unfortunately, held greater influence on the interest score that the data calculated. The correlation with other data collected, therefore, was determined to be more strongly correlated with actual population and diminished the significance of the Google Trends data. All denormalization tasks were completed on Databricks Notebook.

5. Software requirements

5.1 Software Requirement Gathering Strategy

Our project requirements changed constantly, and were mainly determined by our sponsor, who we met with daily. In this daily meeting, we met with our sponsor to determine what tasks we wanted to complete that day and the direction of our project. This allowed us to form our sprint backlogs, as well as make changes to our existing ones. We also met as a group without our sponsor throughout the day to discuss what we had accomplished and any trouble we had been having. This included a weekly retrospective meeting to overview what we had accomplished and what changes we could be making to be more effective for the upcoming sprint.

5.2 Functional and Non-functional Requirements

The first of our two main functional requirements from our sponsor was to provide analysis and show comparisons between Indeed data and Google Trends data around employment searches. The second functional requirement was to provide sentiment analysis for comments made on the WallStreetBets forum. We did this through an analysis of a large amount of historical Indeed data and historical Google Trends data.

5.3 User Stories and Epics

Sprint	User Story	Points
Epic: Familiarization with Programs and Tools		
0	As a student, I would benefit from getting familiar with Microsoft Teams, BrightData, PySpark, Databricks so that I can seamlessly work on the rest of the project	4

0	As a student, I want to set up a Trello in order to easily track our entire agile process through user stories	1
1	As a data scientist, I want to learn how to use the GDelt dataset and what information it was to offer so I can correctly utilize it in the future	2
1	As a student, it would benefit me to familiarize with PowerBI visualizations early on so down the line it will not impede our work	2
Epic: Gathering Employment Search Data		
1	As a student, I want to familiarize myself with Google Trends (GT) to understand the limitations of the project	3
1	As a developer, I want to be able to grab Indeed Job Postings from their site in order to aggregate job posting data over time	1
1	As a developer, I utilize Plotly express for plotting line graph and scatter plot so that our findings are comprehensible to non-developers	1
2	As a developer, I try to find the best correlation between Indeed and GT data nationally and state wise in order to perform further analysis	5
2	As an investor, I would like to know if GT is an indicator for number of new jobs so that it can help to predict future job migration pattern	5
2	As a data scientist, I create linear regression script in order to easily identify correlation between data sources	3
2	As an investor, I would like to see a comparison between normalized GT data and other data sources so that	8
2	As an investor, I would like to see whether there is correlation between unemployment rates and GT (or Indeed data) in order to determine GT may be an indicator for job market	5
2	As a developer, I would like to look into denormalizing GT data so that we can compare it with actual job data	5
Epic: Analysis of Employment Search Data		
3	As a data scientist, I want to figure out if GT is correlating to LEHD after they are scaled by population, so that we can use one dataset to predict the other	3
3	As a developer, I need to find alternate ways to utilize GT data or GT normalization in order to get more accurate data	4

3	As a developer, I look into using GTab to denormalize Google Trends data so that data gathering become more efficient	3
3	As a data scientist, I search for GT related research papers to understand how other researchers utilize GT data for their project	1
3	As a data scientist, I review the calculation for Z-scores so that we are prepared to calculate Z-score of GT data for job searches in US	2
3	As a developer, I perform Z-scores calculation on Databricks using GT data of 10 metros in order to understand current job market	2
3	As a developer, I develop pipeline for quickly processing csv files and plot Z-scores using Databricks so that Notebooks can be shared and used by other data scientists	3
3	As a data scientist, I normalize job data and compare it with GT to validate that GT is not the best indicator for job market	5
Epic: Gathering Stock and Comment Data		
3	As a data scientist, I want to find more correlations in GT by analyzing over different time periods	4
4	As a developer, I want to modify our given reddit scraping script in order to gather comment data to be processed.	1
5	As a data scientist, I would like to provide a Power BI deliverable comparing Google Trends Z-score with Indeed on 10 metro areas so that we can showcase analysis performed with GT data and provide insight into our results.	4
5	As a developer, I would like to write a Python script to summarize sentiment results to be able to analyze the sentiment of different gathered comments.	2
5	As a developer, I would like to gather as much data as possible from WallStreetBets comments so that I am able to analyze them fully later.	6
Epic: Analysis/Machine Learning for Reddit Data		
4	As a student, I want to draw Entity-Relationship Diagram for section 6	2
Epic: Tracking and Reporting our Findings		
1	As an academic, I want to see the background of the company working with the group in order to gain a better understanding of the context of the project	1

1	As a student, I would like to write outlines for the background and intro so that later on we can fill in the missing information.	1
1	As a professor, I want to see the outline include both business and math so that I can understand all aspects of the project	1
1	As a student, I need to write a brief and simple abstract on employment in order to provide more necessary background information	1
2	As a developer, I would like to know which software tools were used in the research and analysis portion of this project	2
2	As a academic, I want to understand why one development methodology was preferred over another so I can understand the process behind which one they chose	1
4	As a data scientist, I would like to see an explanation of each of the data sources used in the project to gain a better understanding of why each source was used and what it brought to the project	1
4	As a data scientist, I want to create and train BERT model for NLP sentiment analysis on reddit comments in order to gain insight into why each stock was mentioned	2
4	As a professor, I want to see the business section of the paper complete so I can track the progress of the project	2
4	As a student, I want to create sprint tables so I can easily track each sprint within the paper	1
4	As a professor, I want to see a complete summary of Agile Scrum methodology with comparisons to other methodologies so that I know the students are correctly implementing agile methodology.	4
4	As an investor, I would like a research section on meme stocks mentioned in the reddit comments in order to get a better context on why the data is being gathered	2
4	As an investor, I want to see a findings section so I know what conclusions have been drawn from the analysis talked about in the paper	2
4	As a professor, I want to see sprint overviews and risk analysis for previous sprints so I can track the entire process of the project and what was being done by the end of each sprint	1
4	As a student, I want to rewrite the introduction and abstract to include the new reddit portion of the project in order to avoid conflicting information later in the paper	3

4	As an investor, I want to see a fully fleshed out background in order to gain a better understand of the context of the project	2
4	As a professor, I would like to see a section on what specifically has been learned by the students over the course of the project in order to understand and validate the work done by the students	1
5	As a data scientist, I would like to recalculate Z-scores to match Indeed data time range and metros so that we are able to easily compare data gathered between the two sources	2
5	As a data scientist, I would want to get GT using PyTrends so that it can be easily exported and visualized.	2
5	As a student, I would like to continue to document our sprints and retrospectives in order to showcase the work we have completed.	1
5	As a student, I would like to provide a summary of our project in order to document the work we have completed.	1
5	As a student, I would like to review linear regression concepts and incorporate collinearity in paper so that I can fully understand them and to provide context into processes discussed in future sections.	2
5	As a project group, we wanted to make an improved ERD and make two new context diagrams to better explain and visualize our processes.	2
5	As an academic, I want to see a review of linear regression concepts and the incorporation of collinearity in paper so that I know what formulas are being used to perform each validation	2
All	As a group, we wanted to look into different ways of achieving deliverables so that we are able to create the most effective one for our project.	2
L. Gebler	As a data scientist, we wanted to recreate a timelapse of tickers mentioned over time in Power BI so that we are able to best display our collected data.	2
All	As a project group, we wanted to display our final project through a presentation to communicate our findings to the sponsor.	2
All	As a project group, we wanted to finalize the report detailing our findings.	4
L. Gebler	As a data scientist, I wanted to create a data visualization dashboard to display our findings.	2

Table 5.1 Epics and User Stories

6. Design

6.1 Existing Software Frameworks and Architectures

Our final deliverables were:

1. Python scripts with a trained Bert model which measures the sentiment of each comment ranging from 0 to 1.
2. SQL queries and Plotly linear regression plots demonstrating Google Trends is not the best indicator for job markets.
3. Power BI showing the z-scores of 10 metropolitan areas that have the greatest growth rates from 2010 to 2010 with populations greater than 2 millions.
4. Other python scripts and the software packages they used: Praw for grabbing comments from Reddit; BERT for natural language processing; SkLearn for finding correlation between datasets.

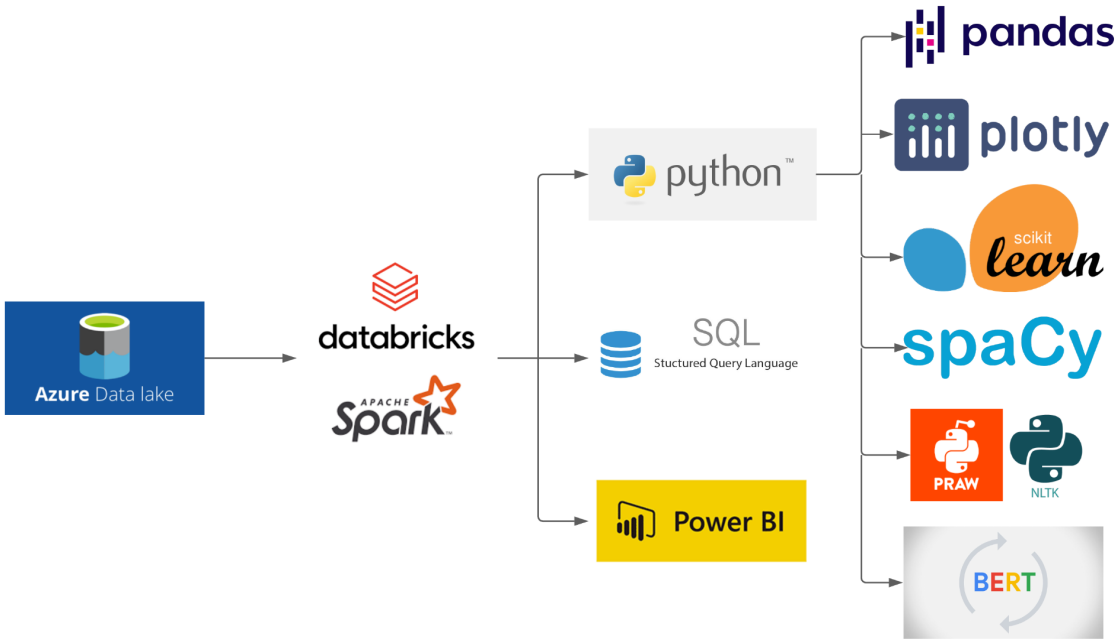

Figure 6.1. Project Team's Architectural Design Flowchart

6.2 Models

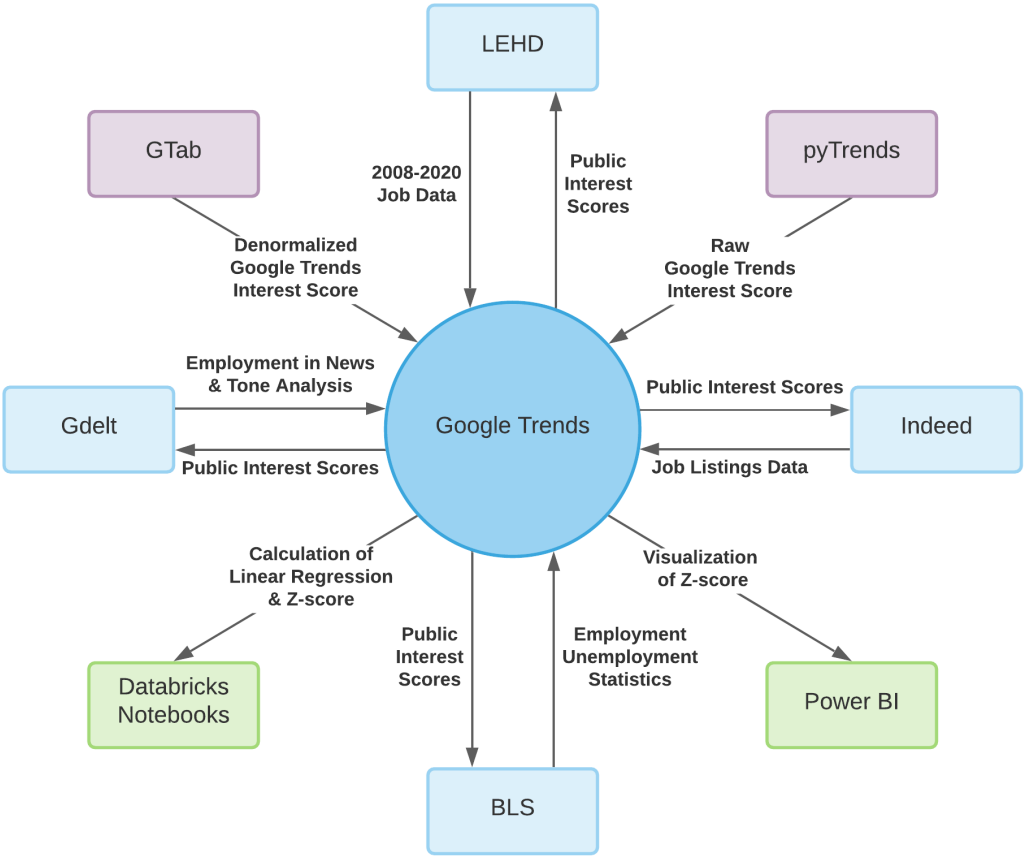

Figure 6.2 Project Team's Context Diagram for analyzing Google Trends

Figure 6.2 showed the data flow of Google Trends interest scores processing: we first developed a pipeline grabbing Google Trends for the search term "job" from GTab and pyTrends, then compared them with 4 different database: LEHD, Indeed, Gdelt, and BLS, after that, we calculated R-squared and Z-score values on Databricks Notebooks, finally presented visualization using Power BI.

Figure 6.3 illustrated the data flow of analyzing Wallstreetbets comments: we first utilized PRAW to request and scrape comments, then cleaned and filtered necessary information using

Jupyter Notebooks and Python script, after that, we applied the BERT model to measure the average sentiment comments that corresponded with a stock ticker, at the end, we produced animated bar chart in Power BI as the final deliverable.

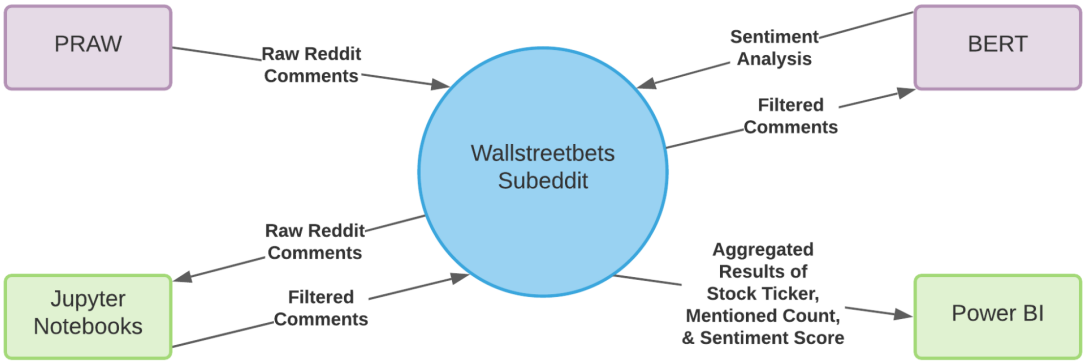

Figure 6.3. Project Team's Context Diagram for analyzing WallStreetBets

6.3 Data Schema

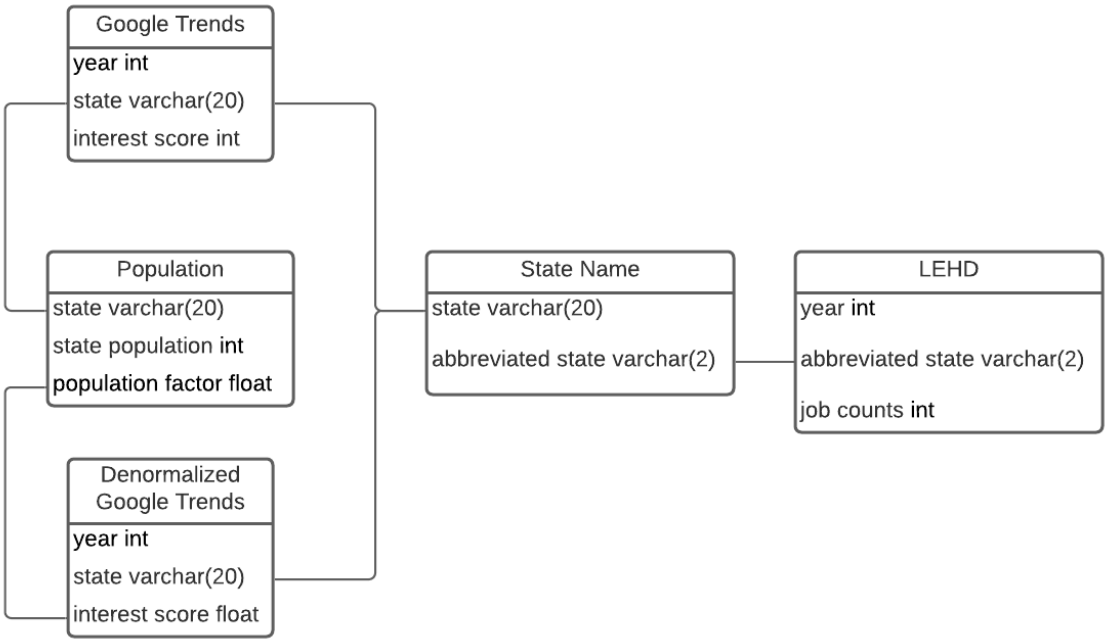

Figure 6.4. Project Team's data schema for comparing Google Trends with other databases

7. Software Development

For our project, we decided to utilize the Agile Scrum methodology. We felt that agile's adaptability and flexibility worked best with our project. Our schedule consisted of weekly sprints with daily standup meetings, each sprint starting on Monday and ending that same Friday. At the start of every sprint, a team creates user stories and assigns them to each group member based on their role. These stories dictate what will be done during the upcoming week and can also be modified throughout the week. For daily standup, each member will declare what they have been working on, if they have anything blocking them from working, and what they plan on working on before the next standup. It's important to be meeting every work day as agile focused projects are constantly changing and adapting to new information or requirements relating to the project. This also includes weekly retrospective meetings that detail what was accomplished in the previous week and what challenges the group faced throughout the week.

We utilized Trello's kanban feature in order to complete this, as shown in Figure 7.1.

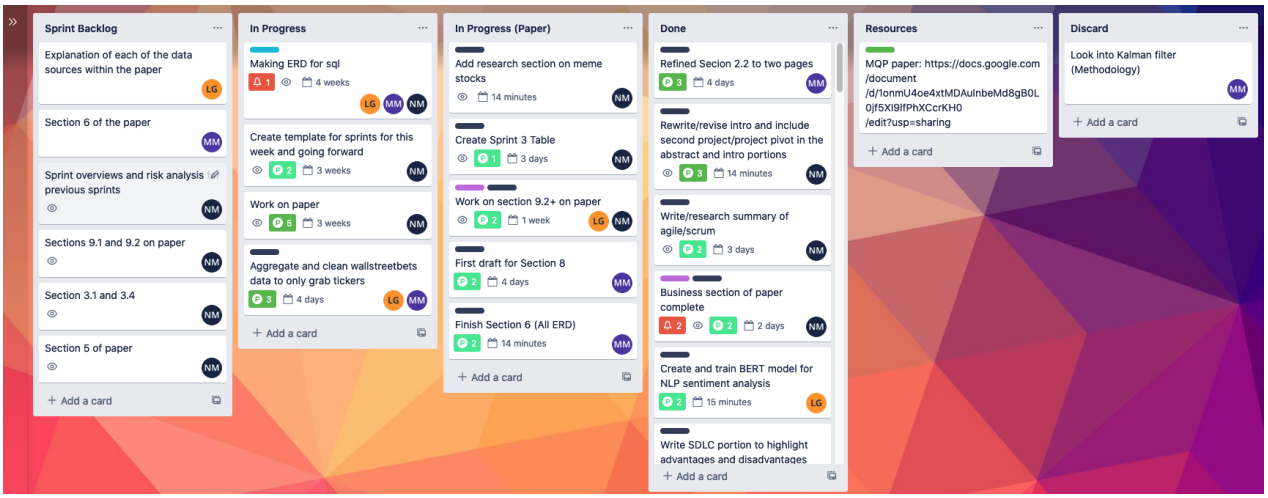

Figure 7.1. Agile Scrum Trello Board

We completed a series of five main sprints over the course of our project, with a starting sprint used to get acclimated to our project, working environment, and different languages and tools that we would be utilizing throughout the duration of our project. For our sprint tables, we felt it important to include our user stories, along with their corresponding epic, point values and status. We equated one point to four hours of estimated work. The status could either be completed during that sprint (green), carried over to the next sprint (yellow) or discontinued (red). We also included risk analysis with possible mitigation strategies, as well as burndown charts for every sprint. For risks, high probability or high status risks are highlighted in red, with medium in yellow and low status or low probability risks denoted with green.

7.1 Sprint 0: Acclimation

7.1.1 Documentation

Status	Story Owner	User Story	Points
Epic: Continuing familiarity with programs and tools			
Completed	All	As a group, we wanted to continue to gain familiarity with PySpark, BrightData, DataBricks, and Microsoft Teams in order to ensure we are able to utilize these programs and tools fully throughout the duration of our project	5
Completed	A. Nicklas	As a student, it would benefit me to familiarize with PowerBI visualizations early on so down the line it will not impede our work	2
Completed	L. Gebler, Z. Ma	As programmers, we wanted to get familiar with Google Trends and how to best utilize the data given to us in order to prepare for analysis and drawing conclusions and making predictions from the data	3
Epic: Continue to document progress and work on paper			
Completed	Z. Ma	Write software environment part of the paper	1

Completed	L. Gebler	Add company info into paper	1
Completed	A. Nicklas	Finish intro and write outlines for the paper	1
Total points competed			**13 out of 13**

Table 7.1.1 Sprint 0 User Stories

7.1.2 Risk Analysis

Description	Risk Category	Probability	Risk Status	Mitigation
Poor organization on paper could lead to a disjointed vision on direction of project/paper	Operational Risk	Low	Medium	Have explicit and agreed upon project goals and paper outline
Lack of completion on paper could lead to difficulties later	Operational Risk	High	Medium	Assigned multiple user stories for next week concerning paper completion
Lack of knowledge/ competency with databases and programs could lead to mistakes or wasted time later in project	Training risk	Low	High	Spent week familiarizing all group members in environment/databases
Not enough completed/ assigned tasks for this sprint could lead to too many tasks later in project	Strategy risk	Medium	Low	Readjusted point value estimation for next week

Table 7.1.2 Sprint 0 Risks

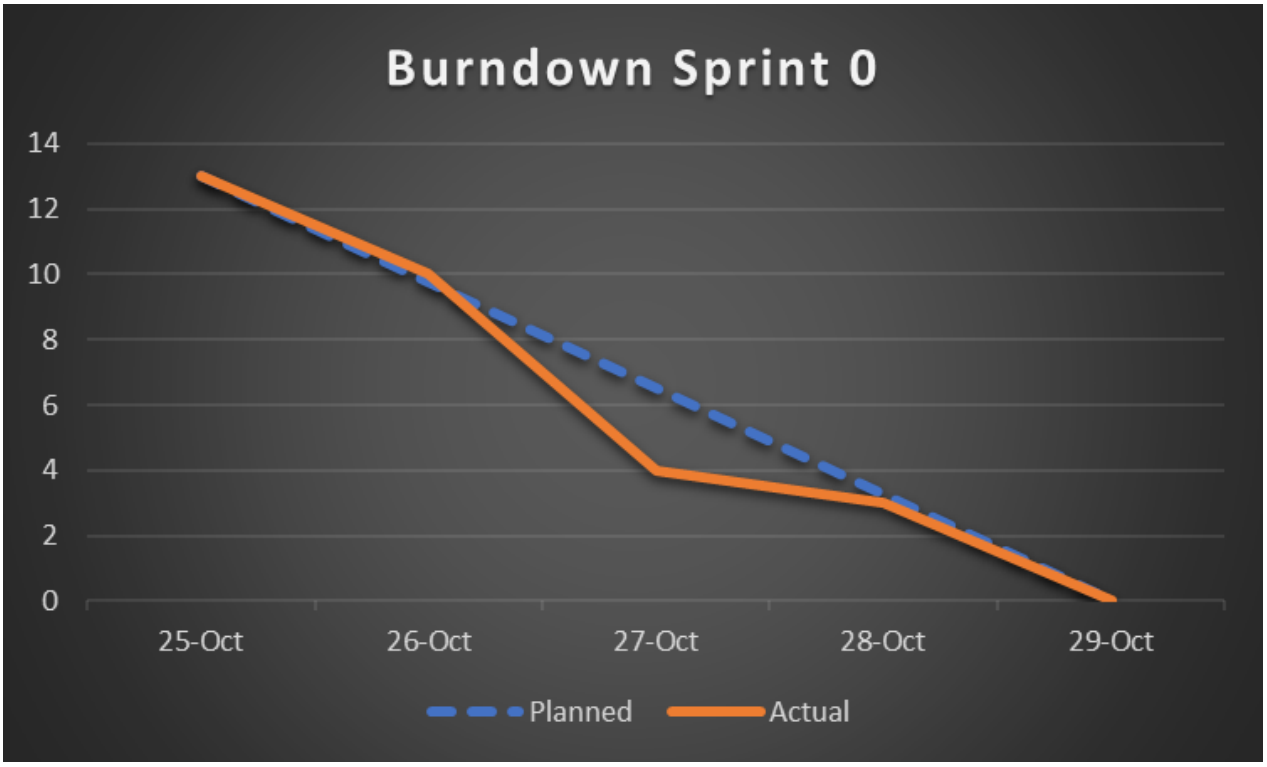

Figure 7.1.3 Sprint 0 Burndown Chart

7.1.3 Overview

For our first sprint, our focus was mainly to acclimate ourselves with the tools that we would be using throughout the duration of our project. We spent the majority of our time acclimating ourselves to the programs we would be using, as well as the remote environment. Throughout this week, we also established contact with the firm and set up daily meetings to discuss our progress throughout the project.

7.1.4 Retrospective

This sprint was our acclimation period with the firm. We spent most of it familiarizing ourselves with the softwares we would be using, including Python and BrightData. Several of these were new to us as a group, which led to some difficulties in getting acclimated. BrightData

proved to be an especially challenging platform to utilize properly. We were able to get set up with the firm on Wednesday in the middle of this sprint. While there were some challenges in getting to this point, once we were set up with the firm, it was a fairly seamless process. We detailed a few challenges we had this week, including finding or creating a basic overview for our project. There were also several improvements we could have made from our progress this week, including doing a better job preparing for the start of our project and finishing more sections in the report, which would have given us a better idea of the structure and progress we were making throughout the project and ensured that we were all on the same page when it came to the requirements of our project.

7.2 Sprint 1: Exploration of Possible Data Sources

7.2.1 Documentation

Status	Story Owner	User Story	Points
Epic: Analysis of Available Data			
Completed	L. Gebler, Z. Ma	As a group, we wanted to continue to analyze and experiment with GDelt data to familiarize ourselves and look for potential trends within the data.	2
Completed	L. Gebler	As a developer, I want to be able to grab Indeed Job Postings from their site in order to aggregate job posting data over time	1
Discontinued	L. Gebler, Z. Ma	Continue to analyze different job listing services (Monster, Indeed, etc.) by scraping through BrightData in an effort to collect data that will help in the analysis of job market data	2
Epic: Continue to work through reporting/paper			
Completed	A. Nicklas, Z. Ma	Expand outline to include business and math sections	2
Completed	A. Nicklas	Brief simple abstract on employment	1
Total points completed			**6 out of 6**

Table 7.2.1 Sprint 1 User Stories

7.2.2 Risk Analysis

Description	Risk Category	Probability	Risk Status	Mitigation
Wasted time by allocating time to tasks that will not end up contributing to our final deliverable	Operational Risk	High	Medium	Make sure that we are constantly meeting with sponsor to realign goals to match theirs
Over-reliance on a single dataset/way of collecting data	Concentration Risk	Medium	High	Explored an assortment of different data sources and databases
Group members/business lead getting COVID	Management Risk	Low	Medium	All remote meetings, made sure every group member is informed of everything going on
Over-variation when it comes to data sources; too spread out to get anything meaningful	Operational Risk	Low	Low	Additional focus on sticking with possible data sources to fullest extent possible

Table 7.2.2 Sprint 1 Risks

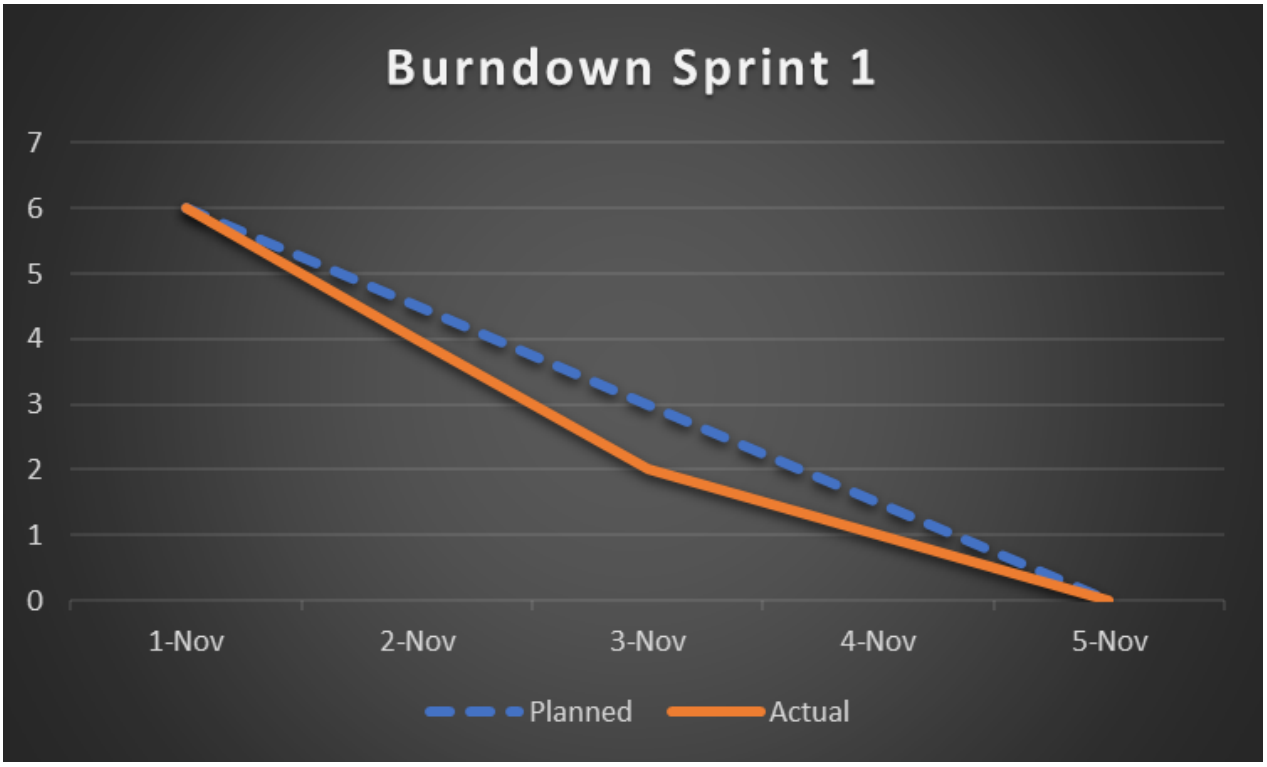

Figure 7.2.3 Sprint 1 Burndown Chart

7.2.3 Overview

This sprint was mainly focused on digging deeper into all of our data sources and beginning to develop a better idea of the direction that our analysis will go. We started the week scraping Indeed data from BrightData, but discontinued that when better data sources were found. We began to collect and analyze Indeed job data.

7.2.4 Retrospective

A large portion of this week was dedicated to exploring BrightData as a source of information for our project. It was thought that if we were able to use this source to scrape current information on a series of different job posting websites, including Monster, Indeed and others, we would then be able to compare this data in some capacity to the data given to us through

Google Trends. This comparison, we thought, would provide us insight into certain trends or predictions that could be useful to the firm.

Despite our continued attempts to utilize this data, we were unable to get BrightData to work for us in any useful capacity. At the end of this sprint, we abandoned BrightData in favor of using Indeed data, for the ability to collect both recent and historical data, and the ease of scraping needed data directly from the site. We also spent much of this week getting familiar with our data sources and choosing the best ones, which we ended up settling on Indeed and GoogleTrends for future analysis. We also looked through a series of public employment and economic databases, including ones given by the Bureau of Labor Statistics (BLS) and Bureau of Economic Analysis (BEA).

7.3 Sprint 2: Looking for a correlation with Google Trends

7.3.1 Documentation

Status	Story Owner	User Story	Points
Epic: Project Organization			
Completed	A. Nicklas	As a project group, we want to continue to work on completing a full initial draft of the paper in order to better understand our project and ensure we complete the project on time.	4
Completed	A. Nicklas	As a group, we want to further understand the business operations of our sponsor and ensure that we are aligned with their goals.	4
Epic: Data Organization/Analysis			
Completed	L. Gebler	As a developer, I want to collect as much useful data as possible around Google Trends job searches and Indeed data in order to begin to start looking into possible correlations within the data	2

Carried over	Z. Ma	As a developer, I want to attempt to de-normalize the Google Trends data in order to allow for more accurate comparisons between different searches, as well as between searches and other datasets.	4
Completed	Z. Ma	As an investor/associate of the firm, I want to test whether Google Trends is an indicator for the number of new jobs that are created to gain insight into where possible future jobs may be.	5
Carried over	L. Gebler	As an investor/associate of the firm, I want to explore correlations between Indeed and Google trends data to validate both datasets and begin to predict future trends based on collected data.	4
Completed	L. Gebler	As a data scientist, I create linear regression script in order to easily identify correlation between data sources	3
Total points completed			**18 out of 26**

Table 7.3.1 Sprint 2 User Stories

7.3.2 Risk Analysis

Description	Risk Category	Probability	Risk Status	Mitigation
Risk of business lead/any group member getting COVID	Management Risk	Low	Medium	Made sure all members are informed, remote meetings
Seven week deadline quickly approaching	Organizational Risk	High	Medium	Continue to add more to each sprint, increase number of points per sprint

Table 7.3.2 Sprint 2 Risks

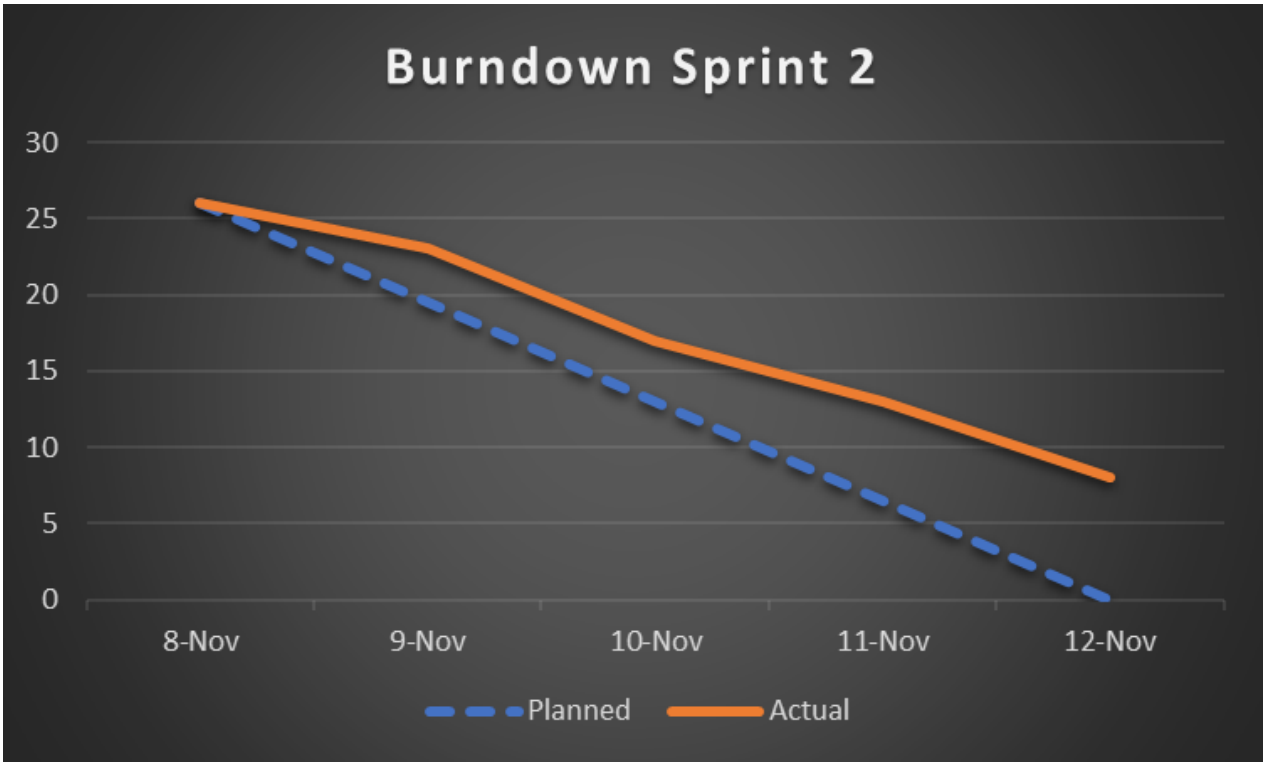

Figure 7.3.3 Sprint 2 Burndown Chart

7.3.3 Overview

For this week, we focused on attempting to denormalize and find ways of drawing meaningful conclusions from the Google Trends data we collected. We continued to collect as much data as possible while beginning our analysis on the data we had already collected.

7.3.4 Retrospective

Throughout this week, we continued our search for a reasonable correlation between the data presented by Google Trends and the numerous other data sources that we had begun to accumulate. We used linear regressions, different Python visualization techniques and Power BI in an attempt to find correlations between these sources and visualize this data.

One struggle we had this week was the lack of any strong correlation apparent between the Google Trends data and the other data sources. Due to the obscuring of the data that Google does,

we struggled to find a good way to compare this data to other sources, which leads us to believe that some strategy of denormalization of this data will lead to correlation for the next sprint.

7.4 Sprint 3: Google Trends Denormalization

7.4.1 Documentation

Status	Story Owner	User Story	Points
Epic: Analyzing Correlation Between Google Trends and Other Data Sources			
Completed	L. Gebler	As a developer, we wanted to analyze correlations between Google Trends and job source data and figure out if GT is correlating to LEHD Data	2
Completed	L. Gebler	As a developer, I need to find alternate ways to utilize GT data or GT normalization in order to get more accurate data	3
Completed	L. Gebler, Z. Ma	As a group, we wanted to look into Google Trends related research papers in order to give us more perspective and information and figure out if we could find a better way to analyze or denormalize the Google Trends data.	1
Epic: Analyzing Trends in GT by Comparing Z-Scores of Different Metro Areas			
Completed	Z. Ma	As a data scientist, I review the calculation for Z-scores so that we are prepared to calculate Z-score of GT data for job searches in US	1
Completed	Z. Ma	As a developer, I perform Z-scores calculation on Databricks using GT data of 10 metros to understand current job market	2
Completed	Z. Ma	As a developer, I develop pipeline for quickly processing csv files and plot Z-scores using Databricks so that Notebooks can be shared and used by other data scientists	2
Completed	Z. Ma	As a data scientist, I normalize job data and compare it with GT to validate that GT is not the best indicator for job market	2

Carried over	Z. Ma	As a data scientist, I want to find more correlations in GT by analyzing over different time periods	2
Epic: Data Organization/Analysis			
Completed	Z. Ma	As a developer, I want to attempt to de-normalize the Google Trends data in order to allow for more accurate comparisons between different searches, as well as between searches and other datasets.	4
Completed	L. Gebler	As an investor/associate of the firm, I want to explore correlations between Indeed and Google trends data to validate both datasets and begin to predict future trends based on collected data.	4
Total points completed			**17 out of 19**

Table 7.4.1 Sprint 3 User Stories

7.4.2 Risk Analysis

Description	Risk Category	Probability	Risk Status	Mitigation
Spending too much time working on a deliverable that will not be useful for us at the end of the project	Operational RIsk	Low	Medium	Align and confirm vision and progress with the sponsor, advisors and group
Seven week deadline: almost halfway through our project and need to make sure we're able to have meaningful deliverables	Deadline Risk	High	Medium	Continue to update Trello and sprints to maintain accountability and effectiveness
Getting too far down the road with Google Trends to be able to change our course of the project	Organizational Risk	Low	Medium	Begin to explore the possibility of using other tools/data sources
Not having clarity when it comes to our project goals, deadlines, or focus	Organizational Risk	Medium	High	Schedule meeting with sponsors, possible with lead sponsor to ensure project clarity

Table 7.4.2 Sprint 3 Risks

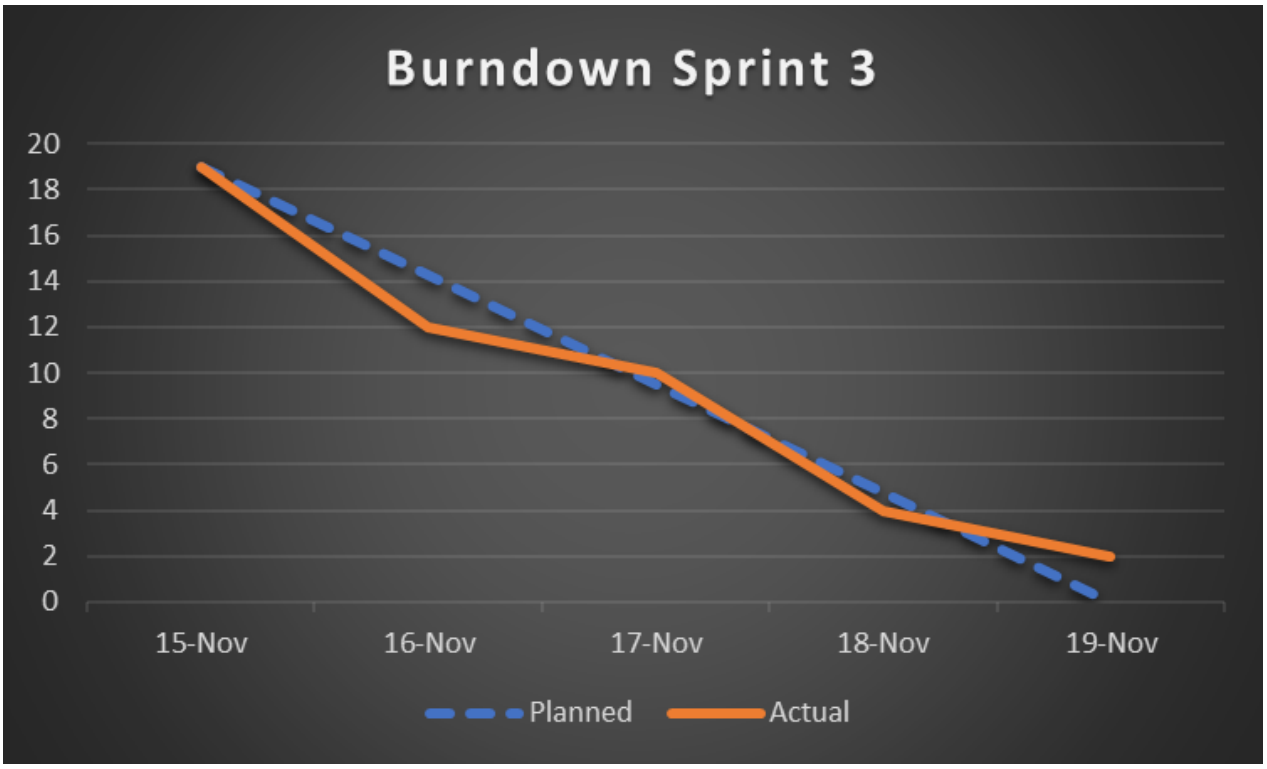

Figure 7.4.3 Sprint 3 Burndown Chart

7.4.3 Overview

This week was spent trying to find and utilize new ways of analyzing Google Trends data in our continued attempt to draw meaningful conclusions from the data we had acquired. We did this through the implementation of two main strategies: using Z-score calculations to normalize and compare tendencies in Google Trends data with growth between years and the use of an anchor search or a group of anchor searches to standardize our Google Trends data.

7.4.4 Retrospective

Throughout this sprint, we ran into a series of problems that we had to overcome in order to continue with our project. In order to solve the problems with normalization of Google Trends data that we had from last week, we attempted to implement a series of different denormalization

techniques in order to better utilize the Google Trends data. Our first idea was to calculate the z-scores for given metro areas and then compare these to each other. This would allow us to compare the rate of year-over-year change in the Google Trends score, given us some metric of easier comparison. Both the data and the visualization for this yielded no useful results.

We also attempted to denormalize the data through the use of an anchor search given to us by GTab. This uses a group of different searches that should be uniform in the different areas of comparison in order to 'anchor' the search we are comparing. This too struggled to produce meaningful results as we ran into problems with query limits and useful correlation.

After feeling like we had exhausted our possible use of Google Trends data throughout the course of this week, we have decided to pivot our work for the next sprint and focus on comments with Reddit's Wall Street Bets forum and their usefulness in monitoring stock holdings of the given firm.

7.5 Sprint 4: Monitoring/Predicting stock market using WallStreetBets

7.5.1 Documentation

Status	Story Owner	User Story	Points
Epic: Make major improvements to reporting of paper			
Completed	Z. Ma, L. Gebler	As a data scientist, I would like to see an explanation of each of the data sources used in the project to gain a better understanding of why each source was used and what it brought to the project.	1
Carried over	L. Gebler	As a data scientist, I want to create and train BERT model for NLP sentiment analysis on reddit comments in order to gain insight into why each stock was mentioned	2

Completed	A. Nicklas	As a professor, I want to see the business section of the paper complete so I can track the progress of the project	2
Completed	A. Nicklas	As a student, I want to create sprint tables so I can easily track each sprint within the paper.	1
Completed	A. Nicklas	As a professor, I want to see a complete summary of Agile Scrum methodology with comparisons to other methodologies so that I know the students are correctly implementing agile methodology.	4
Carried over	A. Nicklas	As an investor, I would like a research section on meme stocks mentioned in the reddit comments in order to get a better context on why the data is being gathered	2
Carried over	L. Gebler, Z. Ma	As an investor, I want to see a findings section so I know what conclusions have been drawn from the analysis talked about in the paper	2
Completed	A. Nicklas	As a professor, I want to see sprint overviews and risk analysis for previous sprints so I can track the entire process of the project and what was being done by the end of each sprint	1
Completed	A. Nicklas	As a student, I want to rewrite the introduction and abstract to include the new reddit portion of the project in order to avoid conflicting information later in the paper	3
Completed	A. Nicklas	As an investor, I want to see a fully fleshed out background in order to gain a better understand of the context of the project	2
Completed	All	As a professor, I would like to see a section on what specifically has been learned by the students over the course of the project in order to understand and validate the work done by the students	1
Epic: Machine Learning			
Completed	Z. Ma	As a data scientist, I want to find more correlations	2

		in GT by analyzing over different time periods	
Completed	L. Gebler, Z. Ma	As a developer, I want to modify our given reddit scraping script in order to gather comment data to be processed.	1
Total points completed			**18 out of 24**

Table 7.5.1 Sprint 4 User Stories

7.5.2 Risk Analysis

Description	Risk Category	Probability	Risk Status	Mitigation
Not being able to present a concrete deliverable with project pivot	Scheduling/Organizational Risk	Medium	Medium	Concurrently working on GT deliverable in case we are unable to finish
Running out of time, not being able to develop a working sentiment analysis model in time	Organizational Risk	Medium	High	Diversifying models programmers are working on, once we get one working we will focus efforts onto that one

Table 7.5.2 Sprint 4 Risks

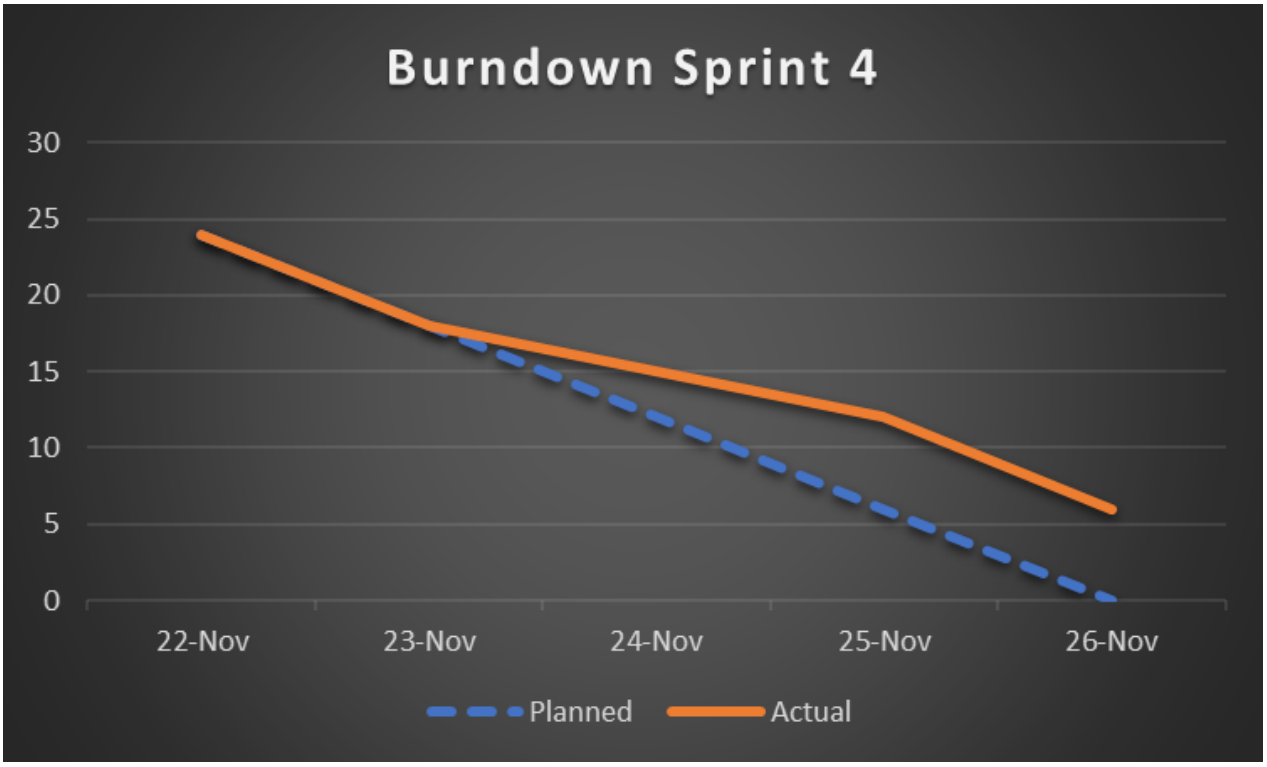

Figure 7.5. Sprint 4 Burndown Chart

7.5.3 Overview

This sprint marked the official beginning of our project's pivot. In this sprint, we got away from our exploration of Google Trends data and began to explore WallStreetBets data. We settled on a language processor and began to look for strategies to gather and analyze this data.

7.5.4 Retrospective

One of the major challenges we discussed for this week was the fact that it had been shortened to a three day sprint with the holiday at the end of the week. For this reason, we adjusted our project sprint points in the sprint planning meeting, and were able to accomplish almost all of the points that we had planned for the week. We found for this week that we continue to make improvements with our ability to work as a group, specifically when it came to both gathering data and reporting it within the paper. We also felt like we were spending

significantly less time throughout this week disorganized. We felt like we really started to hit a stride in terms of the focus of the project and the final goals that we had. We did also talk about the transition from Google Trends to WallStreetBets and ways we could make it a smoother and easier transition, while maintaining focus on Google Trends enough to provide the firm with a meaningful deliverable.

7.6 Sprint 5: Continued Data Collection

7.6.1 Documentation

Status	Story Owner	User Story	Points
Epic: Tracking and Reporting our Findings			
Completed	Z. Ma	As a data scientist, I would like to recalculate Z-scores to match Indeed data time range and metros so that we are able to easily compare data gathered between the two sources	2
Completed	Z. Ma	As a data scientist, I would want to get GT using PyTrends so that it can be easily exported and visualized.	2
Completed	A. Nicklas	As a student, I would like to continue to document our sprints and retrospectives in order to showcase the work we have completed.	1
Completed	A. Nicklas	As a student, I would like to provide a summary of our project in order to document the work we have completed.	1
Completed	Z. Ma	As a student, I would like to review linear regression concepts and incorporate collinearity in paper so that I can fully understand them and to provide context into processes discussed in future sections.	2
Completed	A. Nicklas	As an investor, I would like a research section on meme stocks mentioned in the reddit comments in order to get a better context on why the data is being gathered	2
Completed	L. Gebler, Z. Ma	As an investor, I want to see a findings section so	2

		I know what conclusions have been drawn from the analysis talked about in the paper	
Carried over	Z. Ma	As a project group, we wanted to make an improved ERD and make two new context diagrams to be able to better explain and visualize our processes.	2
Carried over	A. Nicklas	As a project group, we wanted to create an authorship table to showcase who created each section of the project.	1
Carried over	Z. Ma	As an academic, I want to see a review of linear regression concepts and the incorporation of collinearity in paper so that I know what formulas are being used to perform each validation	2
Epic: Gather Reddit and Comment Data			
Carried over	Z. Ma, A. Nicklas	As a data scientist, I would like to provide a Power BI deliverable comparing Google Trends Z-score with Indeed on 10 metro areas so that we can showcase analysis performed with GT data and provide insight into our results.	4
Carried over	L. Gebler	As a developer, I would like to write a Python script to summarize sentiment results to be able to analyze the sentiment of different gathered comments.	2
Completed	L. Gebler	As a data scientist, I want to create and train BERT model for NLP sentiment analysis on reddit comments in order to gain insight into why each stock was mentioned	2
Carried over	L. Gebler, Z. Ma	As a developer, I would like to gather as much data as possible from WallStreetBets comments so that I am able to analyze them fully later.	6
Total Points Completed			**14 out of 31**

Table 7.6.1 Sprint 5 User Stories

7.6.2 Risk Analysis

Description	Risk Category	Probability	Risk Status	Mitigation
Losing sight of our first deliverable and not being able to	Organizational Risk	Medium	High	Splitting programmers between efforts on past deliverable and

have a meaningful conclusion				future one
Wasted time due to acclimation to new data source	Training Risk	Medium	Low	Collaboration with sponsor and group member to talk through any problems we may be having and possible solutions
Members of team/sponsor getting COVID	Operational Risk	Low	High	Completely remote meetings, making sure all members of the group are up to date

Table 7.6.2 Sprint 5 Risks

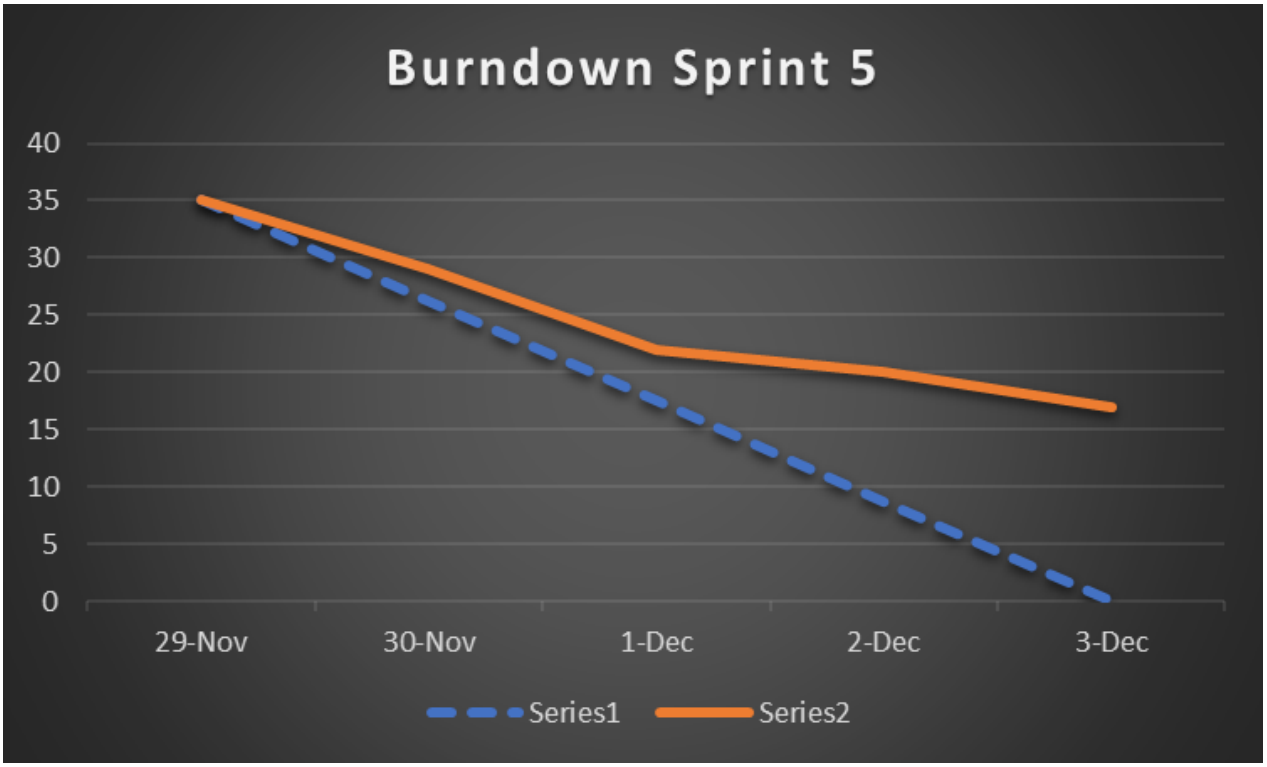

Figure 7.6. Sprint 5 Burndown Chart

7.6.3 Overview

For this sprint, we continued to gather and analyze WallStreetBets data that we had started to collect last week. We looked to gather the past three months of data and begin to analyze using the BERT model that we had created and trained. We also continued our attempts to analyze the

Indeed and Google Trends data that we had collected using Power BI to visualize and present the data in an understandable format. Finally, we continued our work to research and report the different topics surrounding our project and the deliverables we were creating.

7.6.4 Retrospective

Our retrospective meeting for this week was considerably smoother than we have had in past weeks. We noted that we have felt that everything is running much more smoothly than it was in previous weeks and we have grown much more accustomed to the new project. This has led us to have a much clearer goal and vision when it comes to the end of our project. One thing we did mention as a struggle that we had throughout the week was the processing power that it took to scrape, gather, train and analyze the comments and the model. While this was limited by the computers and resources we had, we were able to somewhat circumvent the problem by running the programs consistently in the background.

7.7 Sprint 6: Analysis and Deliverables

7.6.1 Documentation

Status	Story Owner	User Story	Points
Epic: Tracking and Reporting our Findings			
Completed	All	As a group, we wanted to look into different ways of achieving deliverables so that we are able to create the most effective one for our project.	2
Completed	L. Gebler	As a data scientist, we wanted to recreate a timelapse of tickers mentioned over time in Power BI to best display our collected data.	2
Completed	All	As a project group, we wanted to	2

		display our final project through a presentation to communicate our findings to the sponsor.	
Completed	All	As a project group, we wanted to finalize the report detailing our findings.	4
Completed	L. Gebler	As a data scientist, I wanted to create a data visualization dashboard to display our findings.	2
Completed	Z. Ma	As a project group, we wanted to make an improved ERD and make two new context diagrams to be able to better explain and visualize our processes.	2
Completed	A. Nicklas	As a project group, we wanted to create an authorship table to showcase who created each section of the project.	1
Completed	Z. Ma	As an academic, I want to see a review of linear regression concepts and the incorporation of collinearity in paper so that I know what formulas are being used to perform each validation	2
Epic: Gather Reddit and Comment Data			
Completed	Z. Ma, A. Nicklas	As a data scientist, I would like to provide a Power BI deliverable comparing Google Trends Z-score with Indeed on 10 metro areas so that we can showcase analysis performed with GT data and provide insight into our results.	4
Completed	L. Gebler	As a developer, I would like to write a Python script to summarize sentiment results to be able to analyze the sentiment of different gathered comments.	2
Completed	L. Gebler, Z. Ma	As a developer, I would like to gather as much data as possible from WallStreetBets comments so	6

		that I am able to analyze them fully later.	
Total Points Completed			**29 out of 29**

Table 7.7.1 Sprint 6 User Stories

7.6.2 Risk Analysis

Description	Risk Category	Probability	Risk Status	Mitigation
Focus on scraping/analysis prevents proper focus on presentation	Organizational Risk	Medium	High	Delegation of work throughout members of the group so that we are able to maintain focus on both analysis and presentation
Group member/Leader gets COVID/is unable to meet in person	Operational Risk	Low	Low	All remote meetings and maintained communication
Lack of conclusive data when constructing final data presentation	Organization Risk	Low	Medium	Continued work on a myriad of different datasets and analysis techniques

Table 7.7.2 Sprint 6 Risks

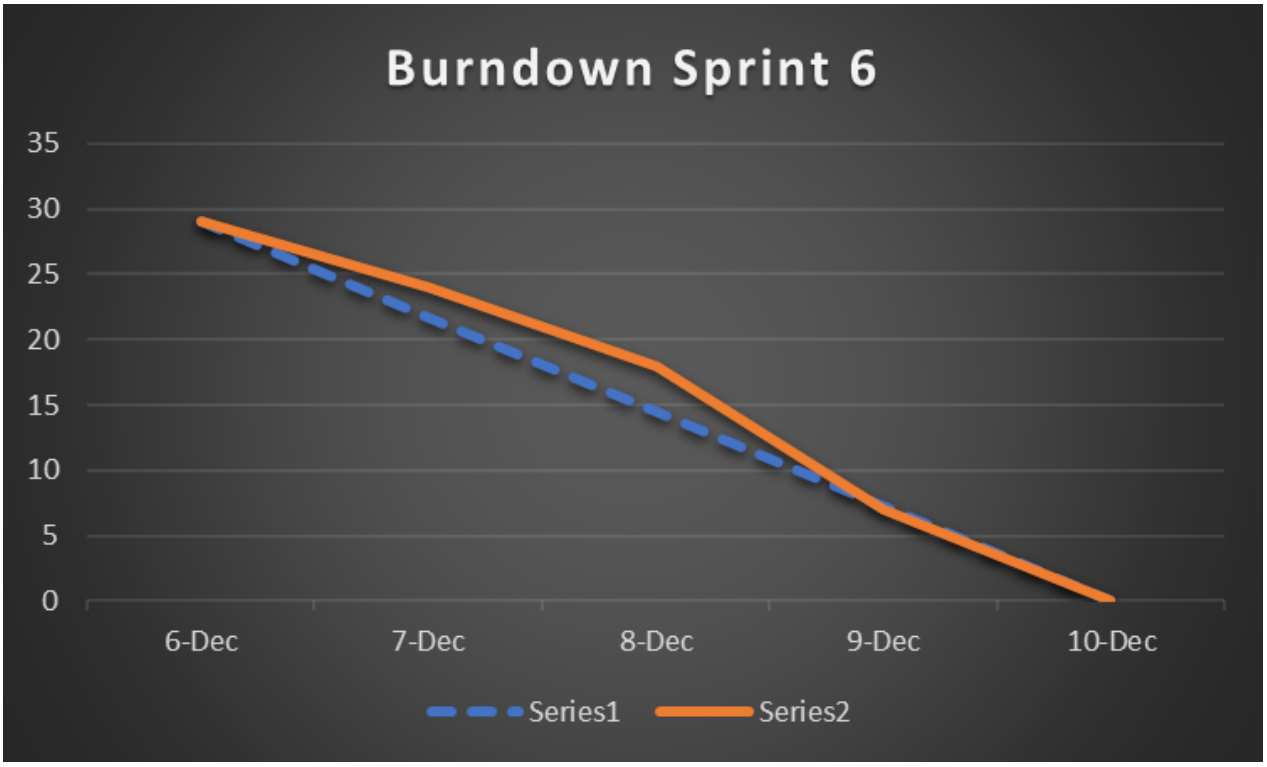

Figure 7.7. Sprint 6 Burndown Chart

7.6.3 Overview

This week marked the last sprint of our project. Throughout it, we were able to work through many of the final deliverables that we were beginning to create for the duration of our project. We spent the week creating different methods of displaying the data and presenting it to most effectively display the analysis we had created.

7.6.4 Retrospective

The point values for this sprint benefited greatly from those that we were unable to finish for last week. Sprint 5 suffered, by only having 8 total points completed, mostly due to the fact that we spent a good portion of the time working towards user stories that we were unable to complete within the sprint. We worked through the final deliverables within this last sprint, and we were able to complete these carried over points as well.

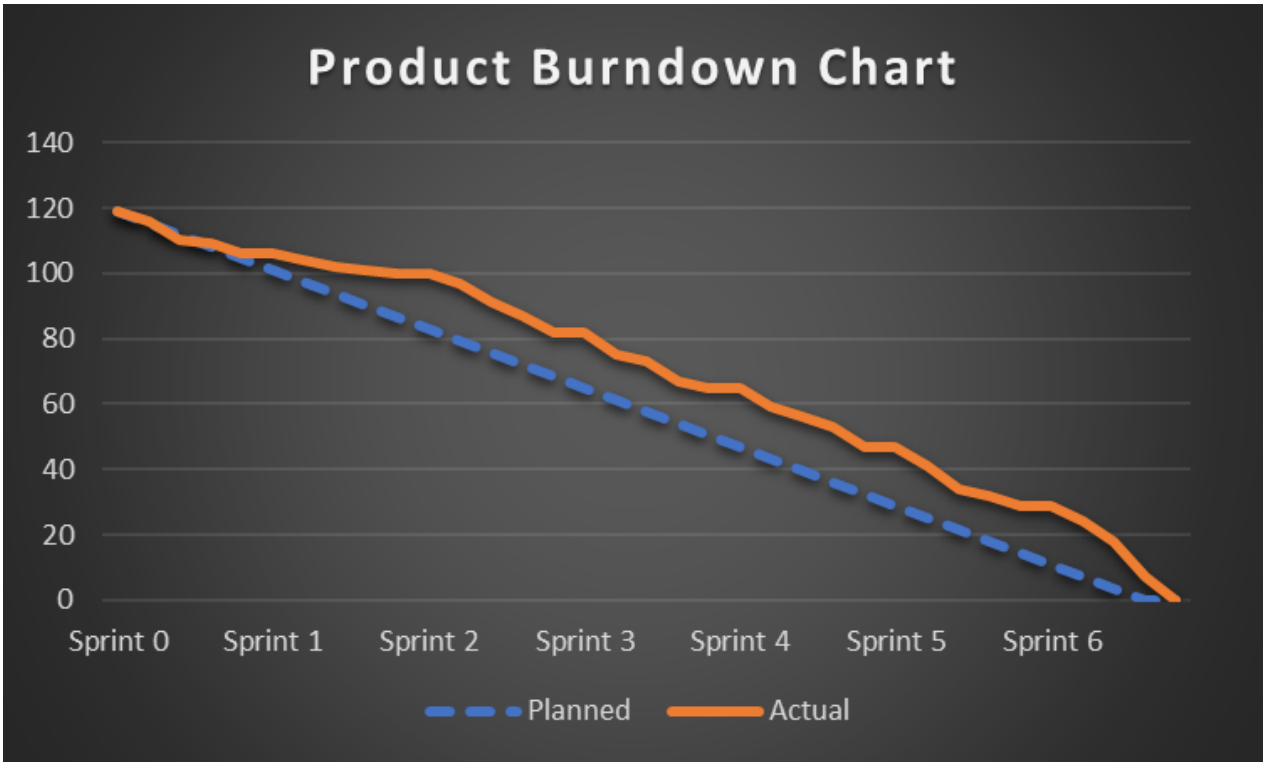

Figure 7.8. Product Burndown Chart

8. Findings/Discussion

We ascertained that Google Trends is not the most valid and reliable data source for estimating job markets and migration patterns due to a lack of correlations between de-normalized Google Trends and historical job data. The other reason is Google Trends compares the relative popularity that interest scores are generated by dividing the total searches of the geography and time range to avoid the situation that areas with small populations are underrepresented, while places with the most search volume would always be ranked the highest. The interest scores become insignificant once we tried to reverse the normalization process by multiplying population factors.

We had a major pivot in the project goals by switching from using Google Trends as job market indicator, to analyzing sentiments from WallStreetBets daily discussions for stock market comparisons. We summarized the most mentioned top 15 stocks every day by grabbing comments from Daily Discussion Thread and analyzing the sentiment of each comment, then calculating the average sentiment score for each ticker. We compared them with the actual stock market behaviors as our second deliverable. An example of aggregated results can be found in figure 8.1.

	stock	count	score
2	SKLZ	81	0.40435539
3	GME	59	0.31065232
4	BABA	53	0.34404567
5	CRSR	42	0.29505163
6	BB	41	0.32171306
7	WKHS	36	0.34975137
8	AMC	30	0.4114217
9	AMZN	27	0.3749781
10	SOFI	25	0.42236574
11	AAPL	25	0.34051043
12	CHWY	24	0.62230973
13	SWBI	20	0.31632267
14	AMD	18	0.39968404
15	ABBV	18	0.34100436
16	TTCF	17	0.23702489

Figure 8.1. Aggregated Results for Wallstreetbets comments on September 1st, 2021

8.1 CS Programs & Math Model

When downloading data from Google Trends, the development team used pyTrends, an unofficial API for processing massive Google Trends search interests, in order to grab data more efficiently in Databricks Notebook. PyTrends returned results in Pandas dataframe, so we transformed it to pySpark dataframe and SQL temporary table for storing and calculating Z-scores. When creating visualizations, we used the Plotly Python library for generating bar graphs and compared the general patterns with job data published by Indeed Hiring Lab for 10 metro areas. We were successful in both maintaining 10 separate Databricks Notebooks that are sustainable to use for the private investment firm and presenting final deliverables on Power BI.

To complete data scraping from Wallstreetbets, we utilized the Universal Reddit Scraper with PRAW, the official Reddit API for obtaining information. We trained a BERT model using movie reviews and fed it with comments to produce sentiment scores as output. The analysis was displayed using Power BI with a timeline from September to November in 2021. We constructed a final presentation showing the alternative investment firm the significant findings in our research, the methodology used, and the future directions for continuing to analyze Google Trends and Wallstreetbets.

8.2 Data Processing & Bias

Since Indeed Hiring Lab only provided job posting data starting from February 1, 2020, we had to limit our search for Google Trends to match up the timeline. We referred to the data of the 2020 Census in comparison to the 2010 Census in order to determine the 10 metropolitan areas we wanted to focus on. Among all the metropolitan statistical areas that have numbers of

population greater than 2 million, we decided to examine those with the fastest growth rate over the past 10 years.

Rank	Metro (metro code)	Census10	Census20	Growth
1	Dallas-Fort Worth-Arlington, TX (623)	6,366,542	7,637,387	19.96%
2	Houston-The Woodlands-Sugar Land, TX (618)	5,920,416	7,122,240	20.30%
3	Seattle-Tacoma-Bellevue, WA (819)	3,439,809	4,018,762	16.83%
4	Denver-Aurora-Lakewood, CO	2,543,482	2,963,821	16.53%
5	Orlando-Kissimmee-Sanford, FL (534)	2,134,411	2,673,376	25.25%
6	Charlotte-Concord-Gastonia, NC-SC (517)	2,243,960	2,660,329	18.56%
7	San Antonio-New Braunfels, TX (641)	2,142,508	2,558,143	19.40%
8	Austin-Round Rock-Georgetown, TX (635)	1,716,289	2,283,371	33.04%
9	Las Vegas-Henderson-Paradise, NV (839)	1,951,269	2,265,461	16.10%
10	Nashville-Davidson–Murfreesboro–Franklin, TN (659)	1,646,200	1,989,519	20.86%
11	Jacksonville, FL (561)	1,345,596	1,605,848	19.34%

Table 8.1 Metro Areas

Due to Indeed currently not sharing postings data for Colorado and Colorado metropolitan areas, including Denver, we switched to investigate the Google Trends search pattern of the 11th metro area, Jacksonville, FL.

We found that all 10 metro areas in the Indeed job posting data have a similar pattern, shown in figure 8.5. In comparison to figure 8.4, when grabbing the corresponding Google Trends data, each metro varied drastically per day. However, in the U.S., the Z-score of the weekly Google Trends searches in figure 8.3 appeared to be more correlated than daily data. Due to time constraint, we suggested that as future work, as described in section 10.1.

8.3 Data Observation & Analysis

8.3.1 Denormalization of Google Trends

We concluded that Google Trends is not a good measure for variation in job markets. First, due to a lack of correlation between raw Google Trends and job data, we agreed to denormalize the interest scores. Second, because the denormalized Google Trends have strong correlation with population data, we encountered the problem of collinearity that their effects on job data are interdependent.

8.3.2 Data Redundancy

We noticed that denormalized Google Trend data and the actual population provided redundant information because of their high correlation with r^2 = 0.983. Collinearity increases standard errors when we tried to compare these data sets with job counts.

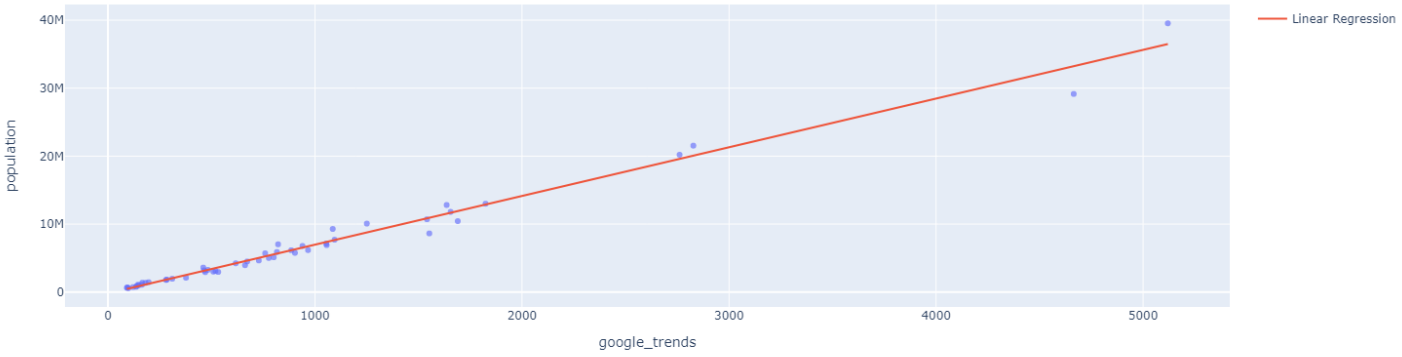

Figure 8.2. Linear Regression between Denormalized Google Trends and State Population

In this case, using denormalized Google Trends data gave us a VIF around 57, which means the standard error increased 7.6 times. Therefore, we reported to the private investment firm that denormalized Google Trends did not work out as we expected. Accordingly, we made a major pivot to study sentiment and stock market analysis.

8.3.3 Stock Tickers and Sentiments

We observed that the more times that a stock sticker gets mentioned, people tend to have a more negative sentiment toward it. The top 5 most mentioned stock tickers per day rarely receive an optimistic judgment from the Reddit users as they are all usually below 0.5. This might be due to some bias in the trained Bert model. There was an exception that we noticed with respect to the comment “BB carry me to the moon”. This sounds positive for a human reader, but seems negative in the analysis of the BERT model. This exception received a sentiment score of 0.204, but there appeared to be no other inaccuracy.

8.4 Data Visualization & Mapping

When Google Trends Denormalization ceased to be promising, we started to compute Z-scores using raw Google Trends data. The comparison ran from the end of 2019 to 2021. For the US in general, the trend is below 0 from week 0 to 17 (November to February), meaning people are less motivated to search for jobs. After that, Z-scores stayed positive and reached the first peak on week 21, indicating people were more willing to start job seeking around April. Then, the scores dropped but the general public remained optimistic toward the labor market. At the end, Z-score reached its maximum in week 51, indicating a flourishing job market.

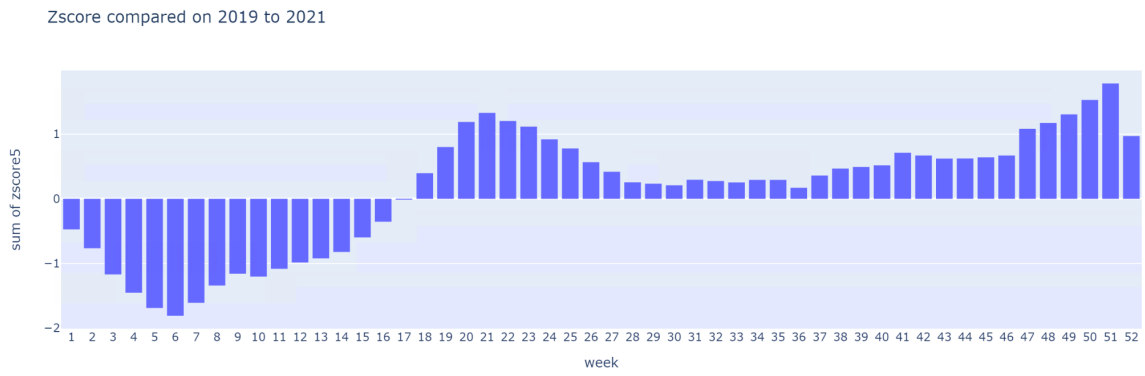

Figure 8.3. Z-Scores Computed Using Weekly Google Trends from 2019 to 2021 in the US

We used the search data of Dallas, TX to explain the daily Z-scores. From figure 8.5, the graph almost looks like a sine wave showing the inconsistency of the data sets. Although a rolling average could give us more steady data, this result is already averaged across 5 consecutive days.

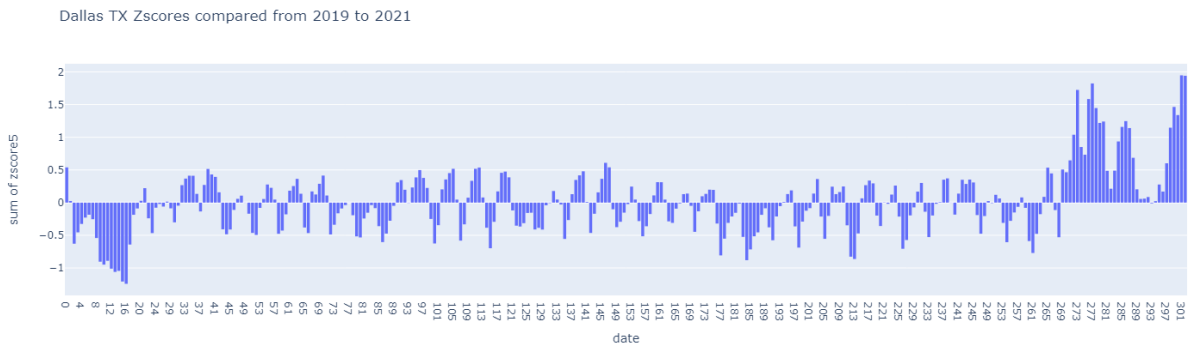

Figure 8.4. Z-Scores Computed Using Daily Google Trends from 2019 to 2021 in Dallas, TX

The way that Indeed summarized their job posting trends is by scaling all data using the number of job postings on Feb 1st, 2020 as a benchmark. It was comparable to our Google Trends Z-scores calculation because they both measure how daily data deviate from a standard or a mean number. Figure 8.3 and 8.5 seem to have a strong connection; it might generate interesting results if we have more time to explore the R-squared value between the two data sets.

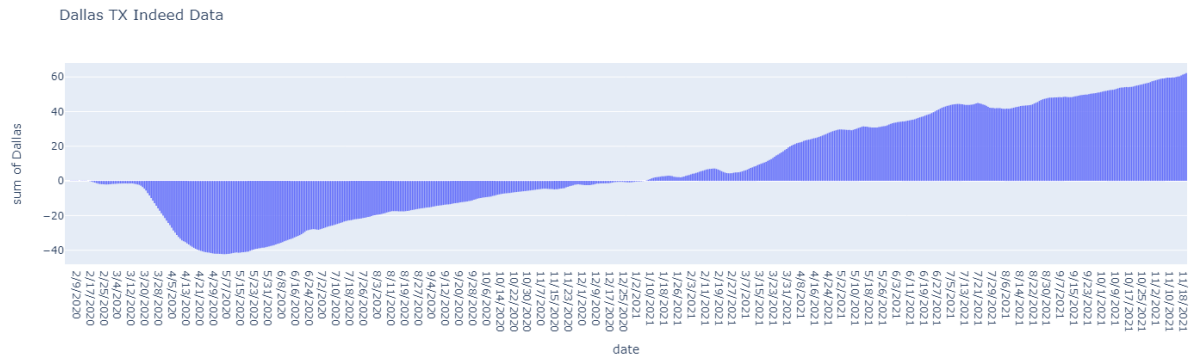

Figure 8.5. Indeed Job Data from Feb 1st, 2020 to 2021 in Dallas, TX

8.5 Accuracy of Datasets & Outliers

Both Indeed and Google Trends reflected a general pattern based on large-scale data sets, like the number of jobs and number of times that people search about jobs in certain areas. They are good measures for showing public intention on job seeking. Percentages and Z-scores did not produce many outliers because both estimated how values deviate from a standard. In rare cases, a Google Trends score of 0 in 2020 caused the Z-score to be null, this exception constitutes the only outlier in our analysis. It was uncommon for a region to receive a Google Trends of 0, showing during a time period and in that area, people are uninterested in employment opportunities. Therefore, We focused on the 10 metro areas in Table 8.1 to avoid such subtle outliers.

There are usually 10,000 to 15,000 comments on Wallstreetbets Daily Discussion per day. The sample is large enough to be considered statistically significant, but we could not detect whether there was bias when using online comments as the major data source. A professor of globalization and development at Oxford university brought up the idea of herd mentality, known as herd instinct or herd behavior, and concluded that WallStreetBets was able to channel this social contagion into financial decision-making, which means Reddit users bought or sold a stock not because of any pattern or news, but because other users also bought or sold the stock (Goldin, 2021). We recognized there might be herd bias and tried to rule out potential outliers by only evaluating the first 15 most-mentioned stock tickers every day.

9. Assessment

9.1 Business Learnings

9.1.1 Leadership and Risk Mitigation

Throughout the duration of our project, there were a number of key takeaways that helped us improve over the course of the project and that we feel could be beneficial for future groups completing similar projects. The first takeaway we had is that strong leadership and a clear vision are vitally important to a successful project. This was something that we found to be an incredibly challenging part of our project at its onset, and something that we caused us to struggle. For the later weeks, our project greatly benefited from the inclusion of an explicit timeline and clear project objectives. This was accomplished through a kanban board updated daily and daily stand-up meetings, both with the group and with the firm.

This also helped to assist us with risk mitigation and risk management throughout the duration of our project. Properly identifying the possible risks that we could face while completing each portion of our project allowed us to develop mitigation plans and strategies to implement when we were faced with such risks. This enabled the team to all be on the same page when it came to our risk strategies and we were able to stay organized throughout what could have been a difficult or turbulent time. This also paid dividends when it came to our efficiency and effectiveness as a group, and ensured that we were not wasting time on improper or ineffective risk management strategies.

9.1.2 Adaptability and Pace of Data Science

Another takeaway that was impressed upon throughout the duration of the project was

how quickly projects can transform and adapt, especially within the Data Science and FinTech industries. Our project constantly adapted as we found new data sources and new questions to investigate were passed down to us. This led to a culture of fast-moving and constantly changing projects, with a heavy focus on being able to quickly abandon a project that was not proving fruitful in favor of a potentially more telling endeavor. From a business perspective, we found it imperative to make sure that we were exploring many different strategies at the same time in order to accomplish a single goal. While we may have had a single business objective or question we wanted to investigate, we had to implore many simultaneous ways of accomplishing it, so as not to get too attached to a singular strategy and be unable to adapt.

Figure 1: Hype Cycle for Data Science and Machine Learning, 2021

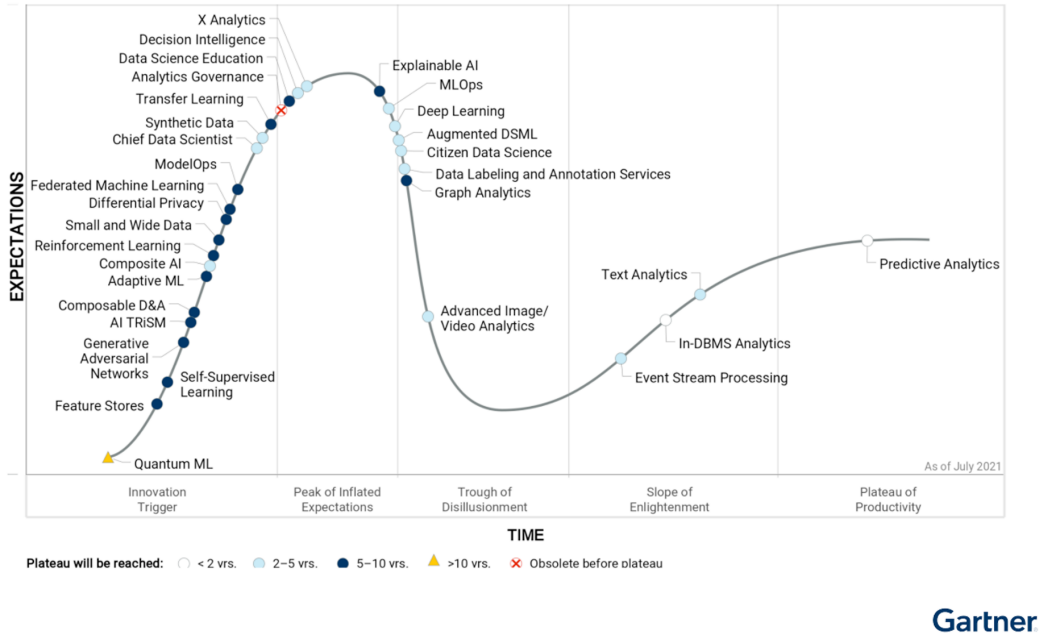

Figure 9.1 Data Science Hype Cycle (Vashisth et al., 2020)

We also found striking similarities between our projects and the Hype Cycle for Data Science, as shown in Figure 9.1. Both of our projects followed this curve where we would start

out with what we felt was an innovative idea passed down from the firm. This spurred us to begin investigating, gathering data and analyzing the data through a myriad of different techniques. Through this, we were often faced with a series of challenges, some of which we would overcome. We would then choose either to pivot or continue to persevere with what we were focusing on, or sometimes a combination of both. This fast-paced cycle was not something we were used to, but something that was incredibly effective once we were able to adapt.

9.1.3 Remote Work

The final part of our project that we had to consider was the benefits and limitations of working on our project remotely throughout the duration of the term. While it allowed us some convenience when it came to having daily meetings, it did cause some difficulty in fully communicating our possible goals and project objectives. Working remotely also made it more difficult for us to work on any portion of the project with multiple people concurrently. This was something that had to factor into our project and time management strategies, and also something we took into account when designing our weekly sprints and project timeline.

9.2 Technical Learnings

During the project, our technical team was able to gain a plethora of real world technical experience with a leading alternative investment firm. Through a multitude of different tasks such as data scraping, denormalization, regression models and more, our team was able to experience the full spectrum of being a data scientist.

The diversity in data sources being used allowed our team to gain insight on the many different ways to gather data from the internet. We used BrightData in order to create bots which scrape job posting websites such as Indeed. Furthermore, we also learned how to leverage public

government data sources such as the Bureau of Labor Statistics in order to gather large amounts of public employment data, a valuable skill in the modern world as a new data source can be as easy as a google search away. The culmination of all this data gathering led us to learn an unexpected new skill, being able to identify which data sources will be valuable and accurate towards our project. This was a unique issue for all of us as we had never before faced the problem of having too many data sources.

After gathering data, the next largest portion of our project required the application of many data wrangling techniques in order to accurately analyze our datasets. These techniques include cleaning, denormalizing, normalizing, and querying datasets in order to modify our data. To do so, we learned how to use Azure Databricks within a company environment, where data sources are shared amongst the entire data science team. This in turn led us to learn how to use PySpark, a Python API which allows you to more easily query and modify very large datasets. To further clean our Google Trends data, we learned the process of denormalizing data over temporal and geological factors which were used by Google in order to normalize the given results.

After sourcing and preparing the data, our technical team was then able to perform an analysis on it. The majority of analysis was done through Python in either Databricks or Visual Studio Code. In order to determine correlation between multiple data sources, we studied and applied a linear regression technique through PySpark. To display the analysis before drawing conclusions, we also learned the vast amount of ways to display data through Plotly.

To create the final data visualizations, we utilized Power BI, a Microsoft data analytics tool with a plethora of functionality. This piece of software was at first very difficult to work with but after learning the basics, it was easy to make complicated visualizations. Power BI allowed us to effectively present our data and is a tool all of us will be utilizing in the future.

This project has allowed our team to experience the life of a data scientist within a leading FinTech company, gaining skills and experiences we would have never otherwise had. This project allowed us to break out of the limitations of the academic environment and get a taste of what real world problems can look like and how these large companies leverage data science teams to solve them. The project team successfully implemented the Agile Scrum Methodology for software development and project management.

9.3 Accomplishments

The major achievements were summarized in Power BI based on the preference of the private investment firm. Wallstreetbets sentiment analysis and Google Trends Z-score were presented in separate Power BI Dashboards.

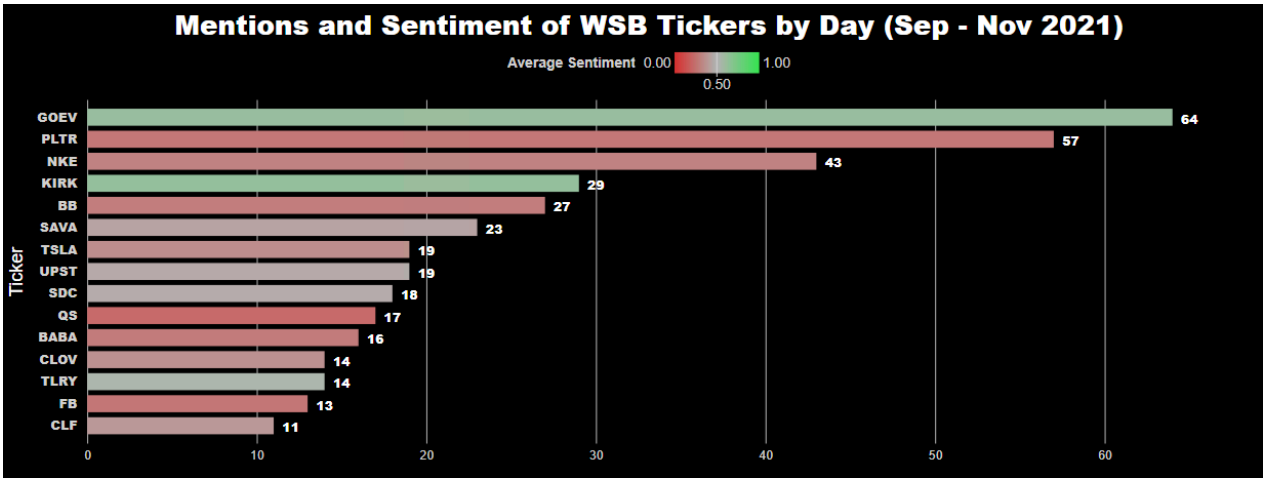

Figure 9.2. Measurements on Ticker Sentiments per Day

We trained a BERT Model to evaluate sentiments. Given a sentiment in a range from 0 to 1, 0.5 represented a neutral opinion, red indicated negative attitude and green suggested positive feeling toward a stock, shown in figure 9.2. We chose the color schema to match with the increases and decreases of the stock market.

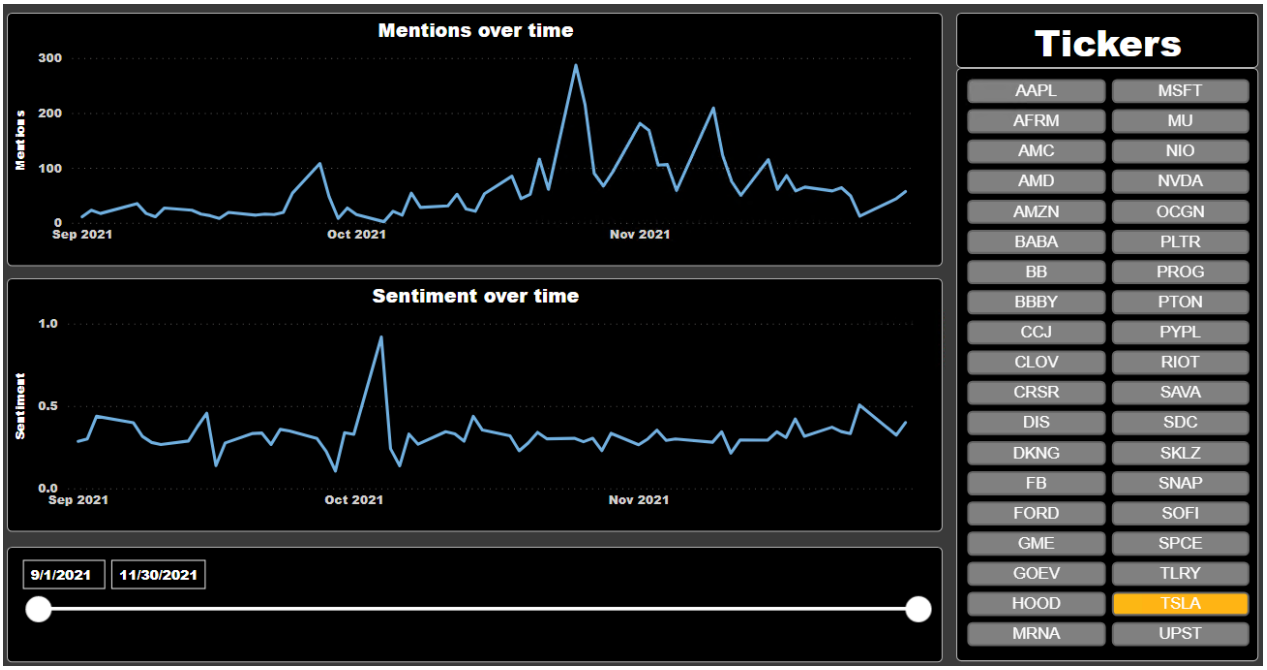

Figure 9.3. Changes in Sentiments for Specific Stock over Time

Figure 9.4. Stock Price of TSLA from June to December 2021

Given sufficient data from September to November, we built a timeline for visualizing the number of times that a stock ticker is mentioned, along with its average sentiments over time. It would be interesting to compare both trends with the actual changes in stock price. Taking TSLA

(Tesla Inc.) in figure 9.3 and 9.4 as an example, mentions over time seemed to agree with the steady performance of the stock, while sentiment over time showed an extreme positive attitude right about the actual stock price started to go up. Although we did not draw any formal conclusion connecting sentiment with stock volatility, it might be promising to monitor an overshoot in sentiments and analyze how closely it correlated to the growth or decline in stock price.

We summed up the number of times that each ticker was mentioned during the past three months in figure 9.5 with an animated bar chart showing how stock tickers were racing with each other comparing their total mentions.

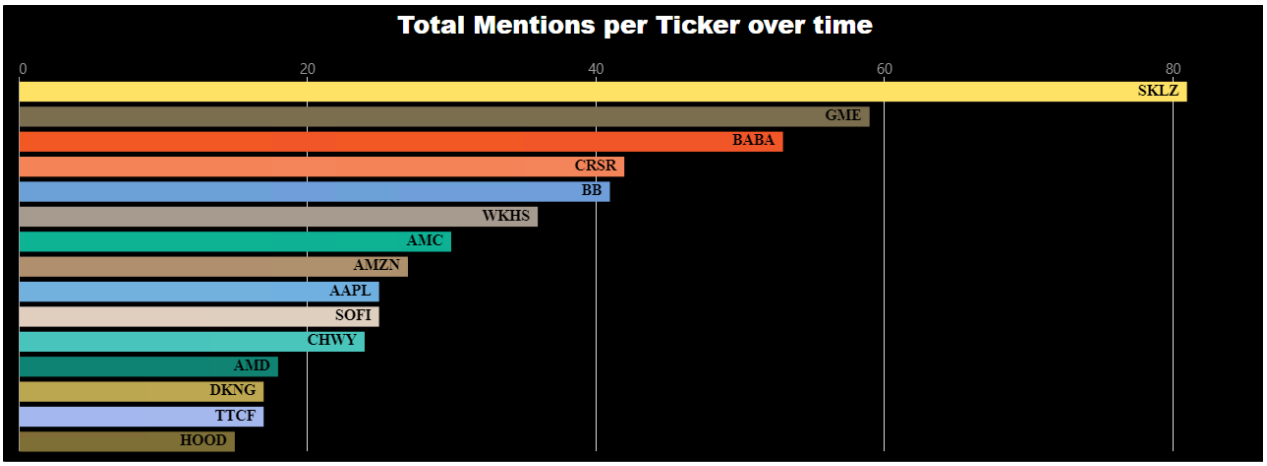

Figure 9.5. Accumulated Results of Tickers Mentioned over Time

The price of a stock is the long-term expectancy of investors to the future profit margin of that company. A research paper suggested that instead of investigating the relation between stock prices one day and investor sentiment several days ago, it believed stock prices always react to the real-time investor sentiment (Ni et al., 2019). We observed that many comments actually reflected previous changes in the stock market. It would be exciting to study the relationship between the comments and actual stock market behavior: whether comments are only a post-product of

volatility of a stock, or there are hidden significance behind these Reddit comments that they are possibly reflecting or even manipulating the stock price.

Besides all the fascinating results we generated from Wallstreetbets, Google Trends and Indeed data also produced patterns that were worth investigating.

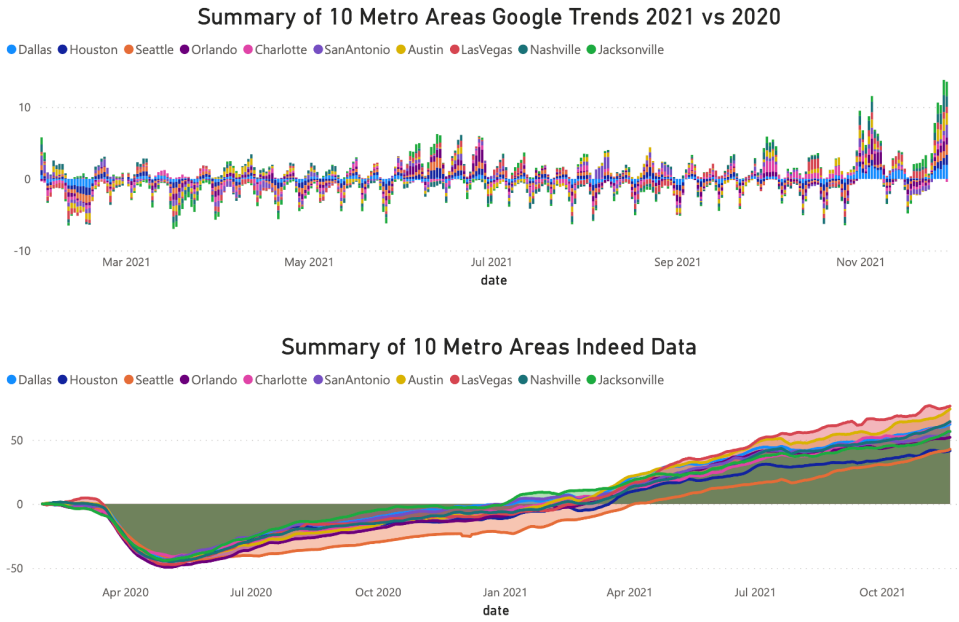

Figure 9.6. Google Trends and Indeed Data Comparison for 10 Metro Areas

We used a stacked column chart to plot Google Trends data which amplified the fluctuation, usually the local max is Sunday while local min corresponded with Tuesday. We chose an area chart to present Indeed job postings data. All metro areas followed a similar trend, which implied the booming of the job market starting around March 2021.

9.4 Limitations

9.4.1 Data sources

Many times throughout the length of this project, we would run into issues with sourcing the data we wanted. This is one of the most common problems in the field of data science and this project was no exception. Certain sources such as Google Trends would either normalize or in some way modify the data which would mask the raw data that was gathered, making it much harder for us to work with it.

Another common issue we faced was the lack of historical data from many of our sources. Indeed and Reddit were two sources we used, both of which had a barrier to entry on their historical data. Indeed has a public github with job posting data since February of 2019 with no mention of data beforehand. Reddit has all its historical data available but you will be rate limited by the API based on the activity on your Reddit account. These types of data sources often led us to either abandon the source all together or spend a copious amount of time just gathering the data.

9.4.2 Time constraint

The largest issue by far in this project was the time constraint given to us at the beginning. Having seven weeks to complete a project on this scale can quickly limit your options on what is feasible. While gathering and sourcing data can take a lengthy period of time, by far the most time consuming and computationally intensive task we completed was training and using a BERT model for sentiment analysis.

When deciding how much data we believed we should gather from Reddit, we had to choose a small enough period of time in order to allow us to run the sentiment analysis on all the

comments. On average, it would take an hour and a half to run a Reddit thread through our BERT model. Each thread would have about 10,000 comments which would first be pruned to only include those with tickers. Given this metric and the fact that each month would have about 23 threads, it would take over a full day of computational power to run a sentiment analysis on a month's worth of data.

10. Future Work

10.1 Google Trends

Throughout the duration of our project, there were a number of times in which we stumbled across data sources or research papers that piqued our interest, but that we did not have the time or the resources to investigate more. When it came to Google Trends, there were a few different ways of utilizing Google Trends that we felt could have been interesting to explore further. One method of comparison we found was through the use of an anchor term. This provides a method of comparison between two or more different search terms that otherwise may have not been comparable (Schmitz, 2019). This could also be done using the Google Trends Anchor Base (GTAB). While this is something we did look into during the end of our exploration with Google Trends, we feel that further investigation could be useful.

Although we generated our final deliverable by juxtaposing Google Trends and Indeed data, their x-axis did not match due to different measurements: Indeed data was scaled based on the number of jobs on Feb.1st, 2020, while Google Trends compared the same day from 2020 with 2021. We could improve by two approaches:

1. Scale Google Trends to match Indeed, which would eventually abandon Z-score calculation but use the interest score on Feb.1st, 2020 as a benchmark.

2. Trivial details are not needed for daily data. It might be more reasonable to shift to weekly data and instead compute the Z-score of Indeed data to match with Google Trends.

We did find that with many of the comparisons and analyses we were trying to perform, Google Trends ended up being a dead end. Most of the future work that we found promising came from our analysis of the Reddit comments.

10.2 WallStreetBets

As we were investigating the natural language processing and WallStreetsBets comments, there were numerous sources that we stumbled upon that we felt would merit further study. In our project, we opted toward the use of the BERT model for its accuracy, ease to understand and the firm's preference, but we feel investigation into other models could be interesting. The first was a distilled spaCy model, which we determined from the idea that "it is possible to train spaCy's convolutional neural network to rival much more complex model architectures such as BERT's" (Peirsman, 2019). This article goes on to detail that using the distilled spaCy model, it was able to improve spaCy's accuracy in their testing by 7.3% and reduce the error of the model by 39%. This could be an interesting model to continue to explore given more time and resources.

The other ways that we talked about possible future exploration would be the use of additional data to better train the BERT model. This could help us with our analysis of the data we ended up getting for our project. The further back we could go in time would also help us to create more accurate and informative results. One of the biggest problems throughout the course of our project was the processing power that we were limited to for both the Reddit scraping and the analysis through BERT. With more processing power, it would be easier to go back further in time to train and process more data.

We also considered the idea of investigating other factors that may contribute to becoming a meme stock. With the GameStop example, it was shown that it was not chosen at random. The company had gone through a period of poor management and failed acquisitions, but they were still valued by some well-known Wall Street names (Lynch, 2021). This made it a prime target for

the WallStreetBets forum. We believe it may be an interesting investigation to create or train a machine learning model that could identify these factors and better predict future meme stocks.

10.3 Machine Learning

We did not have time within the duration of our project to look into the Ensemble Learning Method. If there were causation or correlation between the comments and actual stock market fluctuation, we recommended future MQP teams to apply the Random Forest Regression Model for making scientific predictions. Ensemble learning refers to a group or ensemble of base models working collectively to achieve a better final prediction. A single model, a base or weak learner, may not perform well individually due to high variance or bias. However, when weak learners are aggregated, they can form a strong learner, as their combination reduces bias or variance, yielding better model performance. The idea of the "wisdom of crowds" suggests that the decision-making of a larger group of people is typically better than an individual expert (IBM Could Education, 2021). Random Forest Regression uses ensemble learning methods for regression. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees (Bakshi, 2020).

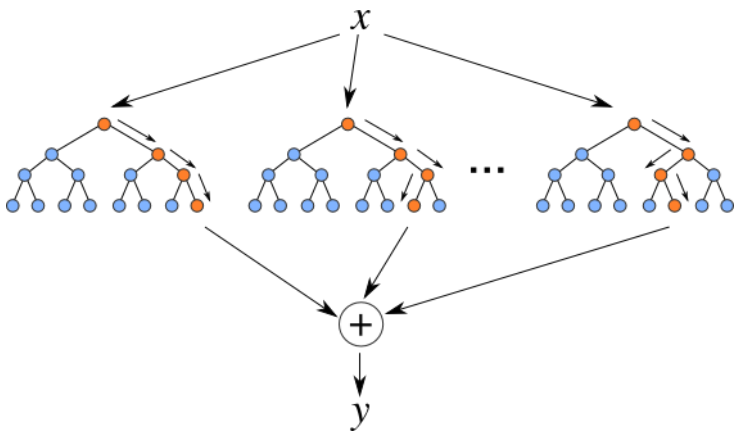

Figure 10.1. Decision Trees and Random Forest Regression Model (Bakshi, 2020)

11. Conclusion

This project worked to provide exploration and analysis of a series of different datasets. We combined databases from a private investment firm, as well as data that we worked to scrape from various different online sources. We first conducted time-series analysis using data sources surrounding employment and job search data, looking for correlations between this employment data and possible trends in migration throughout metro areas in the United States from 2019 to 2021. While we were unable to formally provide a predictive model due to the normalization of Google Trends interest scores, we completed Z-score calculation and data visualization dashboards for the investment firm to continue on detecting potential future correlation. In the latter half of this project, we worked to scrape and analyze comments made about different stock tickers on an online social media forum. With this, we performed time-series analysis and sentiment analysis on both the frequency that comments were made and the corresponding attitudes they had towards various stocks from September to November in 2021. While no conclusions were drawn from the effect that these frequencies or sentiments had towards the actual stock prices, we initiated this sentiment analysis which can be both informative and useful for long-term study for all stock tickers.

12. References

ActiveState. (n.d.). *What Is Pandas in Python? Everything You Need to Know*. https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/

Agile - Daily Stand-up. (n.d.). Tutorial's Point. https://www.tutorialspoint.com/agile/agile_daily_standup.htm

Atlassian. (n.d.). *Retrospective*. https://www.atlassian.com/agile/scrum/retrospectives

Altvater, A. (2021, April 8). *What Is SDLC? Understand the Software Development Life Cycle*. Stackify. https://stackify.com/what-is-sdlc/

Barnier, B. (2021, September 5). *What Is a Hedge Fund?* Investopedia. https://www.investopedia.com/terms/h/hedgefund.asp

Bakshi, C. (2020, June 8). *Random Forest Regression*. Medium. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Boe, B. (n.d.). *Quick Start — PRAW 7.5.0 documentation*. PRAW: The Python Reddit API Wrapper. https://praw.readthedocs.io/en/stable/getting_started/quick_start.html

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brown, G. W., & Cliff, M. T. (2001). Investor Sentiment and Asset Valuation. *SSRN Electronic Journal*, 405–410. https://doi.org/10.2139/ssrn.292139

Brown, S. (2021, April 21). *Machine learning, explained*. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Brownlee, J. (2016, March 25). *Linear Regression for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/linear-regression-for-machine-learning/

Cammenga, J. (2021, December 9). *Where Did Americans Move in 2020?* Tax Foundation. https://taxfoundation.org/state-migration-trends/

Charles, A. (2017, December). *Kalman Filtering: A Bayesian Approach*. The John Hopkins University.

Chladek, N. (2020, May 7). *7 Types of Alternative Investments Everyone Should Know | HBS Online*. Business Insights - Blog. https://online.hbs.edu/blog/post/types-of-alternative-investments

Cprime. (n.d.). *What is AGILE? - What is SCRUM? - Agile FAQ's*. Cprime. https://www.cprime.com/resources/what-is-agile-what-is-scrum/

Crawford, D. (n.d.). *Bright Data Review*. ProPrivacy. https://proprivacy.com/privacy-service/review/brightdata

Daskalova, G. (n.d.). *Intro to Github for version control*. Coding Club.

https://ourcodingclub.github.io/tutorials/git/

Desmond, K. (2021, June 3). *Why Learn Python? 6 Reasons Why it's So Hot Right Now.* CodingNomads. https://codingnomads.co/why-learn-python/

Expert.Ai Team, E. A. (2021, May 6). *What is the Definition of Machine Learning?* Expert.Ai. https://www.expert.ai/blog/machine-learning-definition/

FAQ about Google Trends data. (n.d.). Trends Help. https://support.google.com/trends/answer/4365533?hl=en

FRED Economic Data. (2021, December 3). FRED. https://fred.stlouisfed.org/series/UNRATE#

Fronckova, K., & Slaby, A. (2020, July). Kalman Filter Employment in Image Processing. In *International Conference on Computational Science and Its Applications* (pp. 833-844). Springer, Cham.

General Mills Inc. (2016). *GitHub - GeneralMills/pytrends: Pseudo API for Google Trends*. GitHub. https://github.com/GeneralMills/pytrends#readme

Glen, S. (n.d.). *Z-Score: Definition, Formula and Calculation*. Statistics How To. https://www.statisticshowto.com/probability-and-statistics/z-score/

Goldin, I. (2021, February 10). *How herd behaviour drives action on r/WallStreetBets*. Financial Times. https://www.ft.com/content/971df303-726a-4bdf-93eb-9a9e848f7109

Gravier, E. (2021, October 18). *Meme stocks: What are they and why you should be careful buying them*. CNBC. https://www.cnbc.com/select/what-is-a-meme-stock/

Gupta, S. (2018, June 17). *Sentiment Analysis: Concept, Analysis and Applications*. Medium. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Hart, M., Buck, A., Berdugo, M., Sharabi, K., Sparkman, M., Sharkey, K., & Blythe, M. (2021, September 21). *What is Power BI? - Power BI*. Microsoft Docs. https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

Hayes, A. (2021, April 24). *Understanding Time Series*. Investopedia. https://www.investopedia.com/terms/t/timeseries.asp

Hayes, A. (2021, September 23). *Short Selling*. Investopedia. https://www.investopedia.com/terms/s/shortselling.asp

He, S. (2018, May 29). *From Beautiful Maps to Actionable Insights: Introducing kepler.gl, Uber's Open Source Geospatial Toolbox*. Uber Engineering Blog. https://eng.uber.com/keplergl/

Horev, R. (2018, November 10). *BERT Explained: State of the art language model for NLP*. Medium.https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

IBM Cloud Education. (2021, May 26). *Boosting*. IBM Cloud Learn Hub. https://www.ibm.com/cloud/learn/boosting

Kagan, J. (2020, August 28). *Financial Technology – FintechDefinition*. Investopedia. https://www.investopedia.com/terms/f/fintech.asp

Kakarla, S. (2019, October 17). *Natural Language Processing: NLTK vs spaCy*. ActiveState. https://www.activestate.com/blog/natural-language-processing-nltk-vs-spacy/

Kennedy, R. (2021, March 9). *What is Azure Databricks?* MSSQLTips. https://www.mssqltips.com/sqlservertip/6779/azure-databricks/

Lutkevich, B. (2020, January 27). *BERT language model*. SearchEnterpriseAI. https://searchenterpriseai.techtarget.com/definition/BERT-language-model

Lynch, D. J. (2021, February 2). *The GameStop stock craze is about a populist uprising against Wall Street. But it's more complicated than that.* Washington Post. https://www.washingtonpost.com/business/2021/02/01/gamestop-origins/

McLeod, A. (2019, June 10). What are confidence intervals in statistics? Simply psychology: https://www.simplypsychology.org/confidence-interval.html

Memon, S., Razak, S., & Weber, I. (2020). Lifestyle disease surveillance using population search behavior: Feasibility study. *Journal of medical Internet research, 22*(1), e13347.

Minitab Blog Editor. (2013, May 30). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* Minitab Blog. https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Minot, J. (2021, January 21). *How to use PRAW and crawl Reddit for subreddit post data?* Honchō.https://www.honchosearch.com/blog/seo/how-to-use-praw-and-crawl-reddit-for-subreddit-post-data/

Miratech Holdings, Inc. (2020, August 20). *What is Concentration Risk and How to Reduce It?* JD Supra. https://www.jdsupra.com/legalnews/what-is-concentration-risk-and-how-to-50651/

Mitchell, C. (2021, October 17). *Short Squeeze*. Investopedia. https://www.investopedia.com/terms/s/shortsqueeze.asp

Natrella, M. (2003, June 1). *6.4.1. Definitions, Applications and Techniques*. NIST SemaTech. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc41.htm

Ni, Y., Su, Z., Wang, W., & Ying, Y. (2019). A novel stock evaluation index based on public opinion analysis. *Procedia computer science, 147*, 581-587.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial*

intelligence research, *11*, 169-198.

Pedamkar, P. (n.d.). *SDLC vs Agile*. EDUCBA. https://www.educba.com/sdlc-vs-agile/

Peirsman, Y. (2019, August 26). *Distilling BERT models with spaCy - Towards Data Science*. Medium. https://towardsdatascience.com/distilling-bert-models-with-spacy-277c7edc426c

Perkel, J. (2018, October 30). *Why Jupyter is data scientists' computational notebook of choice*. Nature.https://www.nature.com/articles/d41586-018-07196-1?error=cookies_not_supported&code=60ab3326-e1b7-45da-ae51-2b0c4284fd93

Perla, S., Brown, L., & McCready, M. (2021, July 2). *Notebooks - Azure Databricks*. Microsoft Docs. https://docs.microsoft.com/en-us/azure/databricks/notebooks/

Radigan, D. (n.d.). *Sprint Review Meeting*. Atlassian. https://www.atlassian.com/agile/scrum/sprint-reviews

Rehkopf, M. (n.d.). *What is a kanban board?* Atlassian. https://www.atlassian.com/agile/kanban/boards

Rehkopf, M. (n.d.). *User Stories | Examples and Template*. Atlassian. https://www.atlassian.com/agile/project-management/user-stories

Ruppert, D. (2004). *Statistics and finance: An introduction* (Vol. 27). New York: Springer.

Schmitz, M. (2019, December 6). *Using Google Trends data to leverage your predictive model*. Medium.https://towardsdatascience.com/using-google-trends-data-to-leverage-your-predictive-model-a56635355e3d

Schneider, A. (2021, January 28). *GameStop Stock Mania: Why Everyone Is Talking About It And Many Are Worried*. NPR. https://www.npr.org/2021/01/28/961349400/gamestop-how-reddit-traders-occupied-wall-streets-turf

Singh, A. (2020, December 16). *Hands-On Guide To Natural language Processing Using Spacy*. Analytics India Magazine. https://analyticsindiamag.com/nlp-deep-learning-nlp-framework-nlp-model/

Statistics Solutions. (2021, September 16). *Time Series Analysis - Understand Terms and Concepts*.https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/time-series-analysis/

Szyk, B., Mah, J., & Pal, T. (n.d.). *Confidence Interval Calculator*. OMNI Calculator. https://www.omnicalculator.com/statistics/confidence-interval

Tutorials Point. (n.d.). *SDLC - Waterfall Model*. Tutorials Point SDLC Tutorials. https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm

Vashisth, S., Linden, A., Hare, J., & Hamer, P. (2020, July 28). *Hype Cycle for Data Science and Machine Learning, 2020.* Gartner.

https://www.gartner.com/en/documents/3988118/hype-cycle-for-data-science-and-machine-learning-2020

wallstreetbets • r/wallstreetbets. (2012, January 31). Reddit. https://www.reddit.com/r/wallstreetbets/

West, D. (n.d.). *Sprint planning*. Atlassian. https://www.atlassian.com/agile/scrum/sprint-planning

Wiggers, K. (2019, March 19). *Trello limits teams on free tier to 10 boards, rolls out Enterprise automations and admin controls*. VentureBeat. https://venturebeat.com/2019/03/19/trello-limits-free-users-to-10-boards-rolls-out-enterprise-automations-and-admin-controls/

Willems, K. (2019, January 14). *Pandas Tutorial: DataFrames in Python*. DataCamp Community. https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python