Forecasting the Stock Performance of Top 10 Pharmaceutical Companies in the U.S. by Utilizing Quantitative and Qualitative tools in the realm of Data Science for Long Term Investme...

Preprint · September 2022

DOI: 10.13140/RG.2.2.13728.66562

Ronen Tarnow-Fine
City University of New York - Bernard M. Baruch College

Forecasting the Stock Performance of Top 10 Pharmaceutical Companies in the U.S. by Utilizing Quantitative and Qualitative tools in the realm of Data Science for Long Term Investment

CIS3920 Term project Spring 2022
Prof. Mohammad Chowdhury
Team Members: Yifan Zang (Project Lead), Ronen Tarnow-Fine, Jiyoon Lim

Goal

In this term project, we would like to conduct an integrated approach to predict the stock performance for major pharmaceutical companies with a focus on price level, profitability potential and market movement direction. We will evaluate the forecasting quality of the mainstream data mining including simple linear regression and KNN. In the end, text mining will be applied as a qualitative method to discover human-friendly insights including price-boosting events and stock popularity based on term frequency.

Abstract

According to the time-series data from MarketWatch.com, the S&P pharmaceutical industry index exceeds the S&P 500 by 4.11% in terms of investment returns based on the observation window from 01.01.2020 to 05.04.2022. Although the COVID-19 quarantine, and recession of the economy caused the S&P500 slump by 29.41% on 03.15.2020; the S&P500 pharmaceutical industry index demonstrated a stronger defensive value as it fell by only 20.15%. The compounded annual growth rate of pharmaceutical companies was 13.21%, which outperformed the S&P 500 index by 1.83%. These growth ratios serve as numerical evidence to justify the real-world investment opinion that pharmaceutical stocks worth investment during the COVID-19 period. Therefore, we would like to research how mainstream the data mining algorithms simple linear regression and KNN perform for predicting stock performance.

Index Terms

Simple Linear Regression, K Nearest Neighbor, Text Mining, Pharmaceutical Stock Investment

Research Introduction

The stock market is a dynamic and challenging environment where daily price levels and market movement directions constantly change. The stock performances of public companies also vary from time to time due to competitive efforts and business cycles within their industry. Due to volatility in the stock market, people may feel discouraged from investing in stocks, however, it is feasible to gain profits with the help of data mining algorithms. The semi-strong efficiency market hypothesis (EMH) believes that the price movement reflects all past and current information available to the public, which means that stock market information carries analytical value, and the semi-strong efficiency market will lose effectiveness when future information causes stock-price mismatches. This is when a profitable opportunity can be identified as price mismatching leads to self-adjustment from deviated values to market-consensus values or mean. Because of this, data mining can be expected to be able to analyze stock price and potential profitability, as well as discover the patterns of movement. Hopefully, the data mining algorithms can allow people without a business background to gain insights related to the patterns and assist people in trading for profit in various means including long selling and short selling.

This research project focuses on exploring and evaluating the predictive quality of data mining algorithms: should users rely on price and linear regression to predict stock? Is KNN accurate enough to predict market movement?

The project starts with data preprocessing with the goal of recognizing the price vs time patterns and the volatility of selected pharmaceutical stocks based on scatterplots. This is because volatility is necessary for seeking profits when liquidity continues in either long or short selling. After which, we will examine how simple linear regression performs in predicting stock price.

We will then examine several targeted variables that are measurements of a stock's profitability and risk. These targeted variables are stock-adjusted close price, time (daily), the return in percentage change (return_Perc), and the logarithm of the daily price range (LogRange). Various visualizations including single variable scatterplots, histograms, and bivariate scatterplots will be utilized to discover the underlying probability distributions. Then we will move on to conducting KNN for predicting the market movement direction ("Up" or "Down") using a new set of features that existing research ignores. KNN would allow investors to decide the trading direction "long" (bull) or "short" (bear) as well as which action to take: "sell" or "hold" based on predicted market movements. Lastly, we will utilize a news dataset gathered from major investment news platforms including Barron's, Yahoo Finance, Bloomberg, Motley Fool, and Seeking Alpha to explore stock popularity and price boosting events from 03.01.2022 to 05.04.2022. The text mining task is implemented with the purpose of offering qualitative insights that the quantitative approach may ignore.

Literature Survey

[1] The first research related to our project mentions that EMH (efficient market hypothesis) is a theory that bridges the gap between stock markets and stock information which allows stock price or movement prediction to be conducted based on stock data. The research focuses on KNN and nonlinear regression for price prediction. In the feature engineering processes, they emphasized that the stock price has various characteristics, and it is important to consider the closing price as the next day's price indicator. They trained their KNN model based on: Price, Opening Price, Closing Price, Daily High, and Daily Low. Their research illuminates our thought process for feature engineering and feature selection**.** [2] The second research conducted by Ulster University in the UK offers several technical insights related to KNN when predicting stock performances. First, it showed that the KNN label design can affect model performance. KNN obtains better predictive accuracy when labels are designed into 2 groups "Up" or "Down" instead of 3 groups of labels ("Up", "No Movement" and "Down"). This is because it is rare to detect "No Movement" in the stock market and as such, a No-Movement label would be noise that would deteriorate model performance. This research also proposed a set of 10 selected features that can be trained when implementing KNN; the features are used to measure the price

level, risk level, and transactional trend. Their research suggests that the input window length would affect the predictive quality and the best predictive quality is achieved when the input window length equals the predictive window length. These findings shed a light on the importance of trail design for ensuring higher predictive quality.[3] The third research is conducted by Claremont College, which is utilizing K-Means clustering analysis based on historical price movements and financial ratios for predicting stock performance. The research shows that K-means clustering performs poorly for predicting stock and this finding excludes K-means from our data mining task for stock performance prediction. [4] The fourth related research was conducted by Universitas Sumatera Utara, and it experimented with feature selection, particularly the predictive performance and number of features. The finding shows that the KNN tasks with fewer features (2 or 3) outperform KNN tasks with 5 or 6 features in terms of predictive accuracy. This reinforces the data science commonsense that KNN performs well with fewer features. [5] The fourth related research is conducted by Ali and Neagu. Their research goal is to evaluate the predictive quality of KNN based on different K-Values and various training-to-test ratios for heterogeneous datasets. It indicates that predictive quality would vary with K-values and training-to-test ratios and therefore, we consider designing different trails based on K-values and training-to-test set splitting ratios in our research. [6] The sixth research involved text analysis in the stock market. The researchers gathered news from financial platforms, social media, and corporate disclosures. The research also separated words based on sentiment and conducted feature engineering based on both qualitative words and quantitative technical indicators. The data science task conducted includes deep learning models.

Methodology

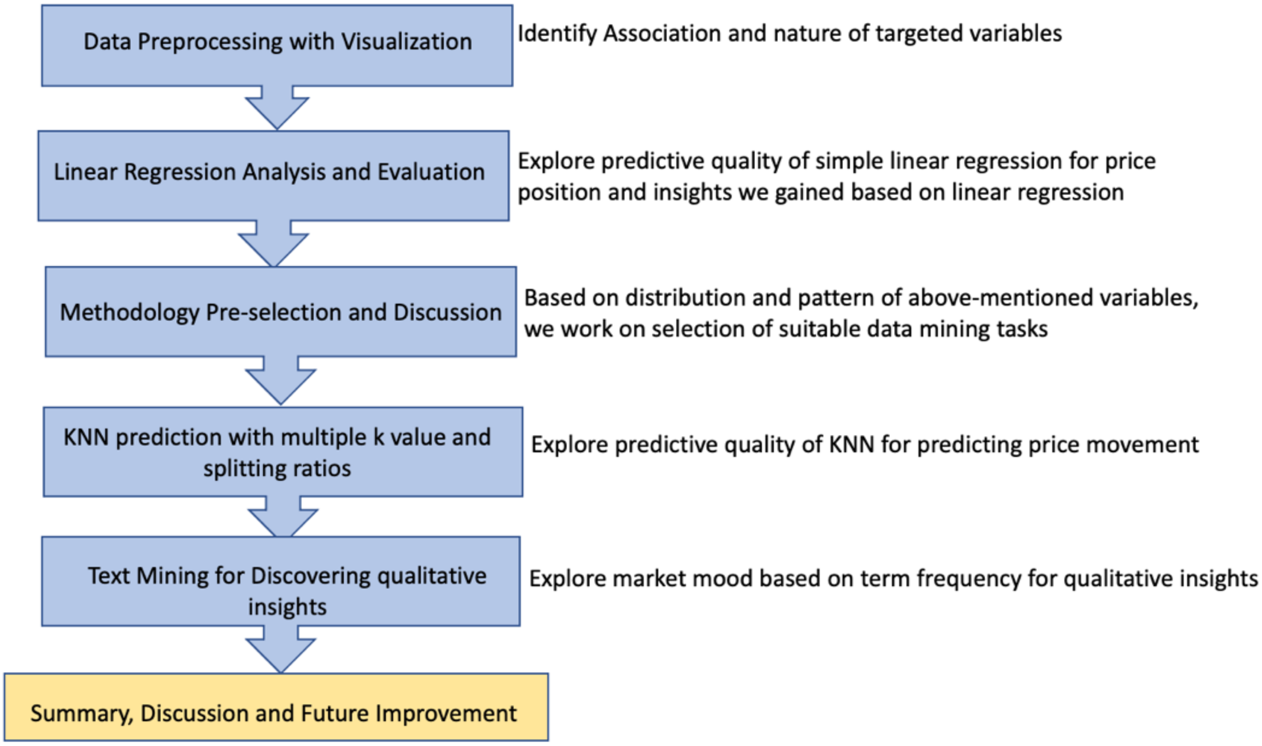
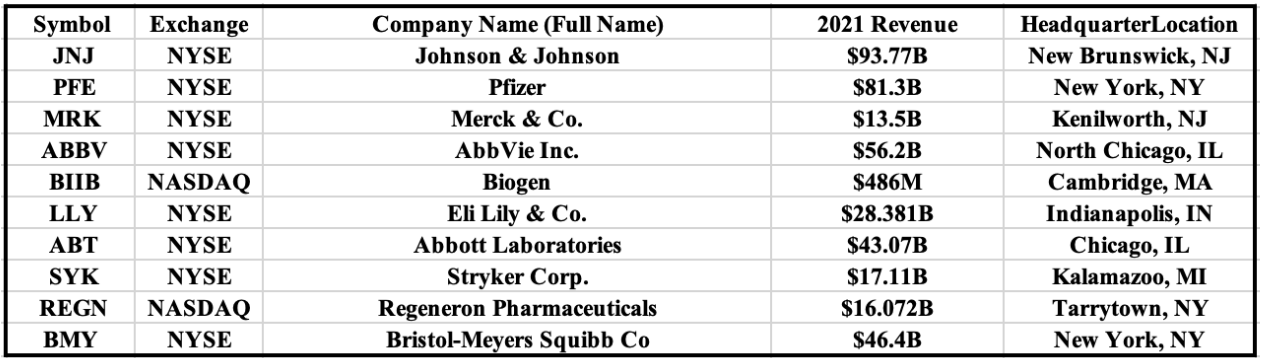

Data Preprocessing with Visualization
Identify Association and nature of targeted variables
Linear Regression Analysis and Evaluation
Explore predictive quality of simple linear regression for price position and insights we gained based on linear regression
Methodology Pre-selection and Discussion
Based on distribution and pattern of above-mentioned variables, we work on selection of suitable data mining tasks
KNN prediction with multiple k value and splitting ratios
Explore predictive quality of KNN for predicting price movement
Text Mining for Discovering qualitative insights
Explore market mood based on term frequency for qualitative insights
Summary, Discussion and Future Improvement

Table 1-2 Research Workflow

Dataset Introduction

Targeted Pharmaceutical Company Stock List Table



Symbol	Exchange	Company Name (Full Name)	2021 Revenue	HeadquarterLocation
JNJ	NYSE	Johnson & Johnson	$93.77B	New Brunswick, NJ
PFE	NYSE	Pfizer	$81.3B	New York, NY
MRK	NYSE	Merck & Co.	$13.5B	Kenilworth, NJ
ABBV	NYSE	AbbVie Inc.	$56.2B	North Chicago, IL
BIIB	NASDAQ	Biogen	$486M	Cambridge, MA
LLY	NYSE	Eli Lily & Co.	$28.381B	Indianapolis, IN
ABT	NYSE	Abbott Laboratories	$43.07B	Chicago, IL
SYK	NYSE	Stryker Corp.	$17.11B	Kalamazoo, MI
REGN	NASDAQ	Regeneron Pharmaceuticals	$16.072B	Tarrytown, NY
BMY	NYSE	Bristol-Meyers Squibb Co	$46.4B	New York, NY

Table 1-1 Top10 Pharmaceutical Companies in terms of 2021 annual Revenue

Dataset for Simple Linear Regression

The stock price dataset was downloaded from Yahoo finance based on publicly available data. The time window selected is from 01.01.2020 to 05.04.2022, which is the early beginning and end of COVID-19. We prepared 10 datasets and each dataset is associated with a targeted company. A total of 588 data points were included in each data set. Data preprocessing is conducted to delete irrelevant data like price and volume. Because we only focused on time (day) and ADJclose price in the simple linear regression task; ADJclose price is the stock price that factors out stock split and dividend payment.

Dataset for KNN

We conducted variable engineering to use LogRange as a risk indicator and daily percentage return as a market movement indicator as well as a profitability indicator.

LogRange= Log(Daily High – Daily Low)

Daily change in Percentage Return = (Today ADJclose – Prior Day ADJclose)/ Prior Day ADJclose

A total of 10 datasets were prepared based on the Yahoo Finance stock price and each dataset contains 588 data points. The time window selected is from 01.01.2020 to 05.04.2022.

Dataset for Text Mining

100 pharmaceutical-related investment news are collected from Motley Fool, Barron's, Bloomberg, MarketWatch, and Yahoo Finance. These websites are mainstream platforms that investors would use to read the news and discuss investments. News is compiled and stored in txt format, and then converted to csv file for reading into R. We hope to extract the most frequent words to discover the stock popularity and price-driving events that people care about the most.

Preliminary Stage: Data Preprocessing

The stock price datasets of targeted companies were downloaded from Yahoo Finance based on an observation window from 01.01.2020 to 05.04.2022; the time frame selected can be considered as the emerging stage of COVID-19 and the post-stage of COVID-19. The selected dataset is sampled independently from the targeted company's whole available stock price dataset. All missing values are removed from the sample.

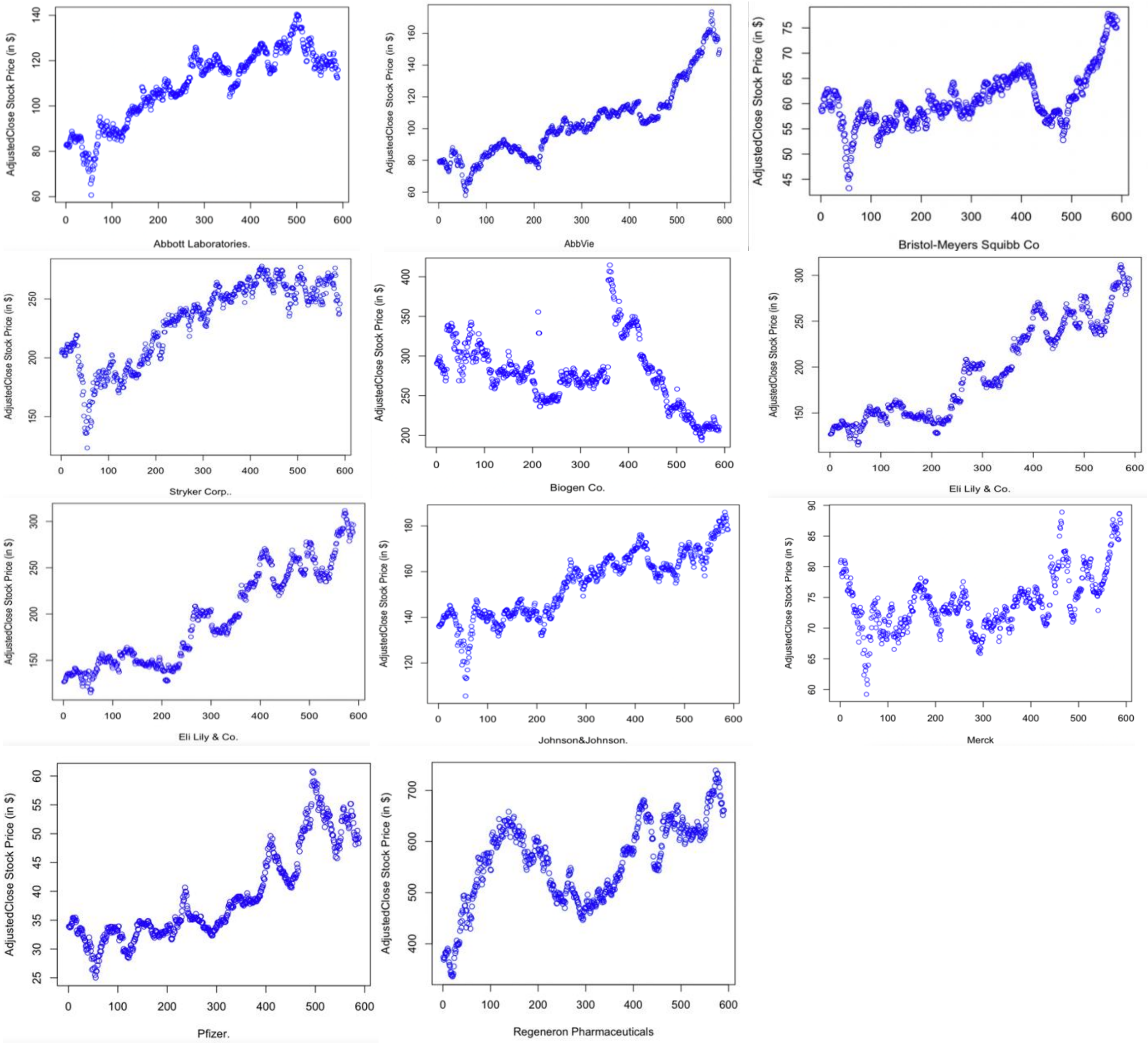
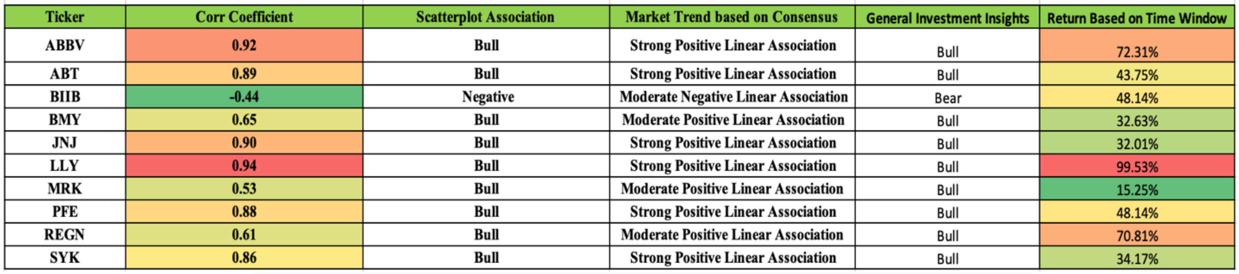


Table1-3 Time Series Scatterplot of ADJclose in $ vs Time on Daily Basis ADJclose price refers to stock closing price factoring out dividend payment and stock split.

Based on the time-series scatterplot of ADJclose price of selected pharmaceutical companies, we identified the association and volatility of selected stocks. Based on the scatterplot output in R, we found that the daily adjusted close price movement and time generally follow a linear association, but some patterns are strong whereas others are moderate. Outliers do exist as we discovered two sharp slumps in day range 40-60 and day range 400 to 450 respectively. The first slump was caused by the outbreak of COVID-19 in late Feb 2020, and it negatively impacted the market until April 20th, 2020. After, the stock flash crash happened on 02.20.2020 and the recession occurred following a short period of bearish markets. The second price tank occurred during days 400 to 450 and this time window was associated with the outbreak of the delta variant in the U.S. Apart from the general trend, we spotted that Biogen's stock tumbled after it reached its peak in late December 2020 due to the failure of its Alzheimer's Drug.

The association pattern and volatility indicate that the selected companies provide an opportunity for profit in either long or short selling. This preliminary step seems to not be directly related to a major data mining task, but it serves as a precondition for evaluating an investment's worthiness: volatility is necessary for stock market profit gains. Data preprocessing dismay the concerns that pharmaceutical stocks are too stable to trade. Based on the correlation coefficient ratio and the stock trend of association, the trading direction can be summarized as follows for reference purposes only.



Ticker	Corr Coefficient	Scatterplot Association	Market Trend based on Consensus	General Investment Insights	Return Based on Time Window
ABBV	0.92	Bull	Strong Positive Linear Association	Bull	72.31%
ABT	0.89	Bull	Strong Positive Linear Association	Bull	43.75%
BIIB	-0.44	Negative	Moderate Negative Linear Association	Bear	48.14%
BMY	0.65	Bull	Moderate Positive Linear Association	Bull	32.63%
JNJ	0.90	Bull	Strong Positive Linear Association	Bull	32.01%
LLY	0.94	Bull	Strong Positive Linear Association	Bull	99.53%
MRK	0.53	Bull	Moderate Positive Linear Association	Bull	15.25%
PFE	0.88	Bull	Strong Positive Linear Association	Bull	48.14%
REGN	0.61	Bull	Moderate Positive Linear Association	Bull	70.81%
SYK	0.86	Bull	Strong Positive Linear Association	Bull	34.17%

Table 1-4 Association Trend and General Investment Insight

Data Mining Task 1: Simple Linear Regression based on ADJclose Price and Time

The rationale behind such selection is attributed to easy deployment and convenient interpretation based on the cost-effectiveness principle. Both independent variables and dependent variables are quantitative, which satisfies the precondition for applying simple linear regression. We chose to include outliers in our simple linear regression to reflect the real-world stock market fluctuation during the COVID-19 period. The outbreak of the COVID-19 variant, drug failure, and vaccine accident are all relevant events to drug stock performance. Within the world of statistics, it is acceptable to report linear regression results with outliers.

The independent variable selected is time ("Day" or "Date"). The dependent variable observed is ADJclose price, which is the stock closing price factoring out stock dividend and stock splits. On each day of the observation window, the observed day and corresponding ADJclose price can be treated as an independent observation.

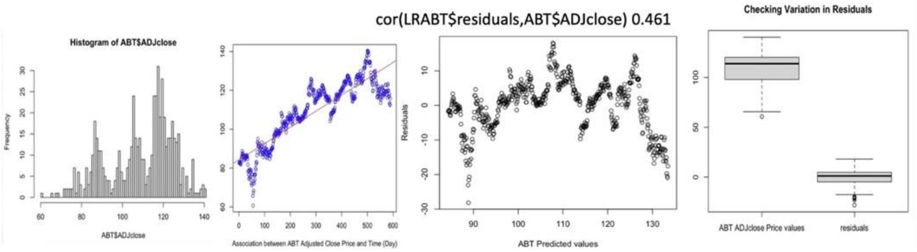
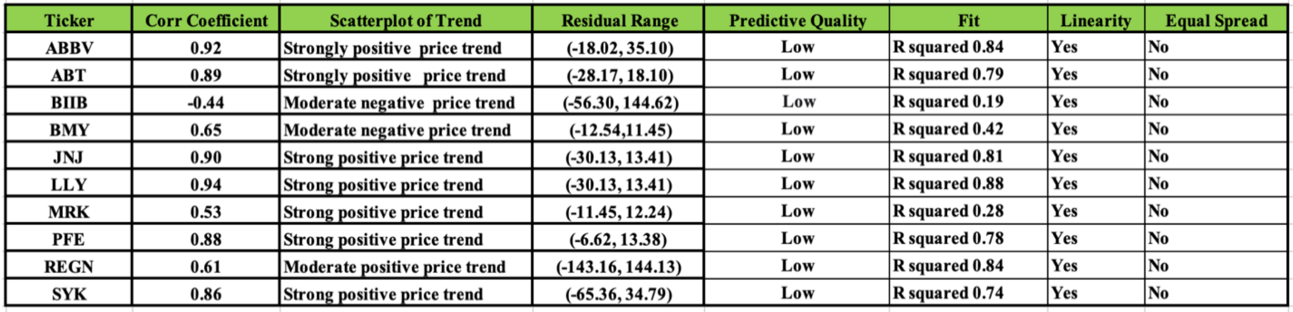


Table 1-5 Example: AbbVie Stock Data Visualization with Linear Regression Assumption Check Visualization Tools: Histogram of Stock Price, Linear Regression line vs Actual Value, Residual Plot, and Variation Boxplot.

However, simple linear regression performs poorly for predicting the stock market price. For example, the visualization and R summary(lm) of AbbVie shows that linear regression carries high residual ranging from -18.02 to 35.10, which indicates a low predictive value.

Following the same visualization approach, the table below summarizes all visualization output and assesses the fitness and predictive quality for all selected pharmaceutical companies:



Ticker	Corr Coefficient	Scatterplot of Trend	Residual Range	Predictive Quality	Fit	Linearity	Equal Spread
ABBV	0.92	Strongly positive price trend	(-18.02, 35.10)	Low	R squared 0.84	Yes	No
ABT	0.89	Strongly positive price trend	(-28.17, 18.10)	Low	R squared 0.79	Yes	No
BIIB	-0.44	Moderate negative price trend	(-56.30, 144.62)	Low	R squared 0.19	Yes	No
BMY	0.65	Moderate negative price trend	(-12.54,11.45)	Low	R squared 0.42	Yes	No
JNJ	0.90	Strong positive price trend	(-30.13, 13.41)	Low	R squared 0.81	Yes	No
LLY	0.94	Strong positive price trend	(-30.13, 13.41)	Low	R squared 0.88	Yes	No
MRK	0.53	Strong positive price trend	(-11.45, 12.24)	Low	R squared 0.28	Yes	No
PFE	0.88	Strong positive price trend	(-6.62, 13.38)	Low	R squared 0.78	Yes	No
REGN	0.61	Moderate positive price trend	(-143.16, 144.13)	Low	R squared 0.84	Yes	No
SYK	0.86	Strong positive price trend	(-65.36, 34.79)	Low	R squared 0.74	Yes	No

Table 1-6: Evaluation of Linear Regression

Based on the histogram of ADJclose stock price, we found that the ADJstock price of each selected company follows a stochastic pattern instead of a bell-shaped curve based on histogram visualization. This means that ADJclose price is random, and its distribution does not have a pattern, which is an underlying reason why linear regression performs poorly for predicting stock price.

In summary, simple linear regression is not suitable for predicting stock market price because of its high residual, less than ideal fitness, and violation of equal spread condition. The histogram plot of stock ADJclose price does not always show a bell-shaped curve, which means that stock price distribution is random or stochastic. Due to its random nature, predicting stock ADJclose price is not a task that can be achieved with high accuracy using simple linear regression.

Our evaluation of the performance linear regression contradicts the real-world popular opinion stating that stock price fluctuates around a mean and linear regression is useful for traders to see where stock is overbought or underbought. Ideas or opinion like this on Investopedia and TowardDatascience are not true.

Our research then shifts focus to exploring other characteristics of stocks with the hope of finding certain variables that follow an underlying probabilistic distribution, if there are any. If certain variables follow a pattern, we hopefully can gain actionable insights based on the underlying distribution.

What we explored are

(1) Risk indicator:

$$\textbf{LogRange} \quad = \textbf{Log(Daily High – Daily Low)}$$

(2) Profitability Indicator:

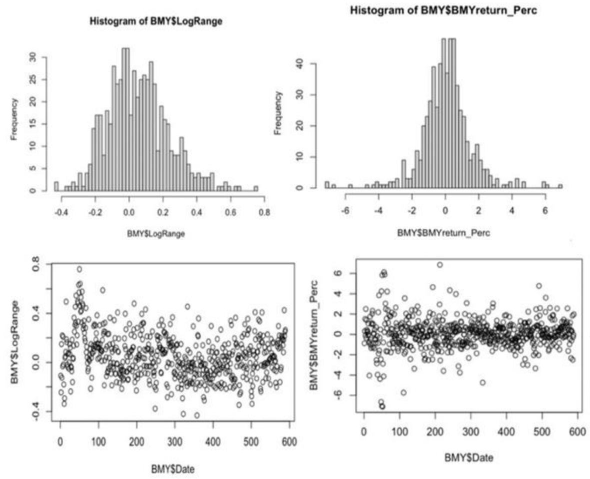
$$\textbf{return\_Perc} = \textbf{Rt\%} = \frac{\textbf{(}P_{t+1}\textbf{ -}P_t\textbf{)*100\%}}{P_t}$$

Interpretation: LogRange value has a range from (-∞， +∞). A higher positive value of LogRange means the stock is more volatile. If LogRange is close to 0, it means the stock has low volatility. The reason to take logarithm for the range is to scale both the large and small numbers so that when we conduct a visualization of selected data points, the graph output would be easily viewed and judged. For return_Perc, we take percentage (%) as measurement. Large positive value means strong profitability whereas negative percentage means loss, and 0% indicates no gain or loss.

We conducted such variable engineering because researchers seldom try to visualize and explore the underlying patterns of risk and return % change or visualize these 2 variables together. We expected to discover some meaningful patterns of these 2 features using a single variable scatterplot, histogram, and bivariate scatterplot. The scatterplot of LogRange and return_Perc cluster around a central horizontal line. Based on statistical knowledge, the stability of the scatterplot is graphical evidence that there is an underlying probability distribution of the return in percentage change, and this is a relatively stable pattern across time because the histogram shows that investors can expect a central value, μ of LogRange or return in percentage.

Right skewed bell curve distribution of Return_Perc is spotted. Mean value of percentage return is near to 0.05 and risk indicator LogRange fluctuates from -0.2 to 0.4 most of the time.

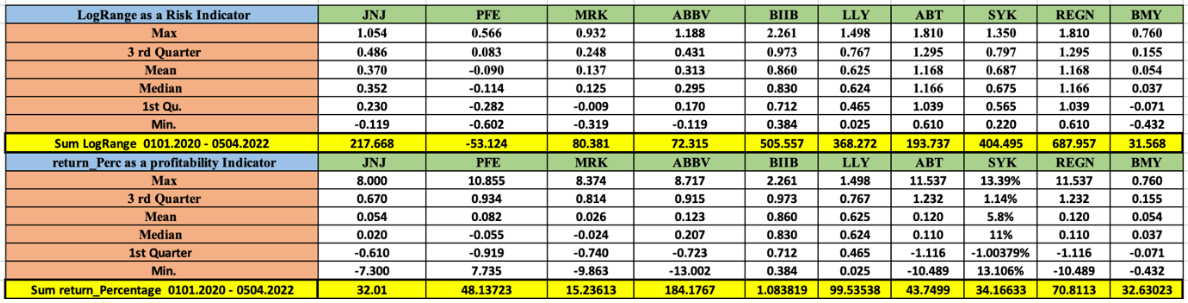
The scatterplot looks relatively stable, which means underlying pattern may exist in terms of distribution



Bell curve distribution of Return_Perc is spotted. Mean value of percentage return is near to 0.05 % and range from -2% to 2% most of the time.

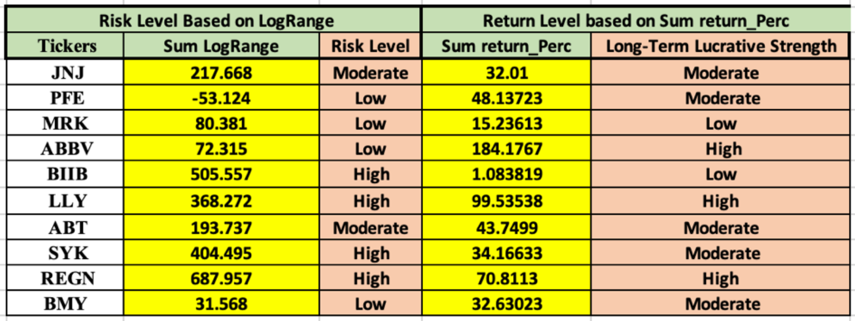
The scatterplot looks stable, which means underlying pattern may exist in terms of distribution

Table 1-7: Visualization Example of Bristol-Meyers Squibb Co: LogRange Scatterplot,LogRange histogram, return_Perc Scatterplot, Return_Percentage Histogram



LogRange as a Risk Indicator	JNJ	PFE	MRK	ABBV	BIIB	LLY	ABT	SYK	REGN	BMY
Max	1.054	0.566	0.932	1.188	2.261	1.498	1.810	1.350	1.810	0.760
3 rd Quarter	0.486	0.083	0.248	0.431	0.973	0.767	1.295	0.797	1.295	0.155
Mean	0.370	-0.090	0.137	0.313	0.860	0.625	1.168	0.687	1.168	0.054
Median	0.352	-0.114	0.125	0.295	0.830	0.624	1.166	0.675	1.166	0.037
1st Qu.	0.230	-0.282	-0.009	0.170	0.712	0.465	1.039	0.565	1.039	-0.071
Min.	-0.119	-0.602	-0.319	-0.119	0.384	0.025	0.610	0.220	0.610	-0.432
Sum LogRange 0101.2020 - 0504.2022	217.668	-53.124	80.381	72.315	505.557	368.272	193.737	404.495	687.957	31.568
return_Perc as a profitability Indicator	JNJ	PFE	MRK	ABBV	BIIB	LLY	ABT	SYK	REGN	BMY
Max	8.000	10.855	8.374	8.717	2.261	1.498	11.537	13.39%	11.537	0.760
3 rd Quarter	0.670	0.934	0.814	0.915	0.973	0.767	1.232	1.14%	1.232	0.155
Mean	0.054	0.082	0.026	0.123	0.860	0.625	0.120	5.8%	0.120	0.054
Median	0.020	-0.055	-0.024	0.207	0.830	0.624	0.110	11%	0.110	0.037
1st Quarter	-0.610	-0.919	-0.740	-0.723	0.712	0.465	-1.116	-1.00379%	-1.116	-0.071
Min.	-7.300	7.735	-9.863	-13.002	0.384	0.025	-10.489	13.106%	-10.489	-0.432
Sum return_Percentage 0101.2020 - 0504.2022	32.01	48.13723	15.23613	184.1767	1.083819	99.53538	43.7499	34.16633	70.8113	32.63023

Table 1-8: Table of Threshold Value for LogRange and return_Perc for Selected Companies



Risk Level Based on LogRange			Return Level based on Sum return_Perc	
Tickers	Sum LogRange	Risk Level	Sum return_Perc	Long-Term Lucrative Strength
JNJ	217.668	Moderate	32.01	Moderate
PFE	-53.124	Low	48.13723	Moderate
MRK	80.381	Low	15.23613	Low
ABBV	72.315	Low	184.1767	High
BIIB	505.557	High	1.083819	Low
LLY	368.272	High	99.53538	High
ABT	193.737	Moderate	43.7499	Moderate
SYK	404.495	High	34.16633	Moderate
REGN	687.957	High	70.8113	High
BMY	31.568	Low	32.63023	Moderate

Table 1-9 Assessment of Accumulative Market Strength based on Risk and LogRange

Key Findings based on Preliminary Data Preprocessing and Simple Linear Regression:

[Bullet Point Summary]:

1. Linear regression is helpful for obtaining an overview of recent market trends so that investors can plan the trading direction: long selling or short selling for long-term investment purposes. Linear regression graphical outputs can inform investors of the ballpark of the stock price range.
2. Linear regression is not proper for predicting stock prices due to the randomness of stock price distributions and high residuals between the predicted price and actual price. Linear regression is not a proper model for fitting stock prices due to the violation of homoscedasticity. This contradicts the real-world opinion that linear regression is helpful for predicting stock prices.
3. What follows a distribution pattern is the risk indicator LogRange and profit indicator Return in Percentage. The underlying bell-curved distribution allows investors to identify the deviation of the return or risk level compared to its mean and frequent value based on the histogram. Therefore, it is more realistic to set a percentage return target and risk level threshold instead of a price level target. By identifying the deviation of LogRange and return in percentage change, profit can be identified.

[Summary] Linear regression alone has low predictive quality. Within a certain period of the observation window, the integrated interpretation of ADJclose price and its associated risk and return% will provide better information for investors in terms of current profit potential, risk level, and market mood (general bull or bear, strong bull or weak bull).

Second Data Mining Task: KNN for Predicting Stock Market Movement

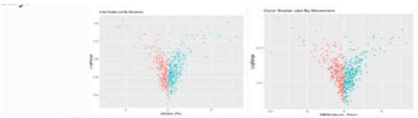
After identifying a market trend and profitable opportunity based on target value derived from underlying distribution, it would be beneficial to predict the stock market movement direction. Within the realm of data science, KNN as a supervised learning method for classification can be adopted to assess the quality of movement prediction.

The rationale behind the selection of KNN:

1. K-NN is a non-parametric algorithm, which means it does not make any assumptions about underlying data. As stock-related data are random and theories behind the stock market can be complex, it is safe to make no assumption for classification.
2. The KNN algorithm assumes that similar things exist in proximity for pattern recognition.

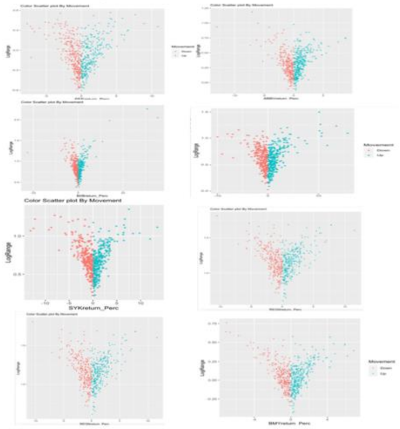
A bivariate scatterplot demonstrates such a pattern, which enhances our confidence to further conduct KNN to assess the possibility of movement prediction and then evaluate the quality of prediction.

When we designed a label for the KNN dataset, we created a categorical variable called “movement”. Two values can be assigned to the movement variable: “Up” or “Down”. The “Up” label would be assigned to the datapoints with positive return_Perc and the “Down” label would be assigned to datapoints with negative return_Perc. Based on the bivariate data scatterplot, KNN is expected to perform well based on proximity because LogRange and return_Perc show a near symmetric pattern with a vertical boundary.

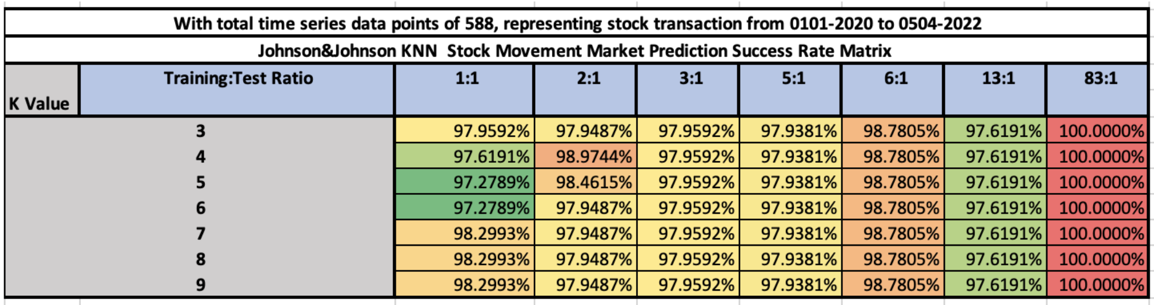


Graph 2-1 Bivariate Scatterplot of LogRange and Return_Perc

Theoretical KNN function: $\hat{f}$: $\text{Yhat}_{\text{movement}} = f(\text{LogRange},\text{return\_Perc}, \text{k value})$

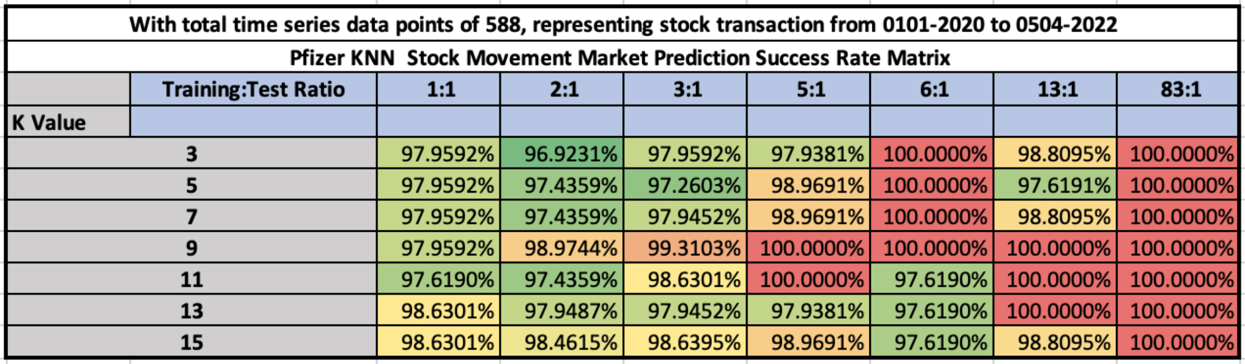


We conducted 10 KNN tasks and designed 49 trials for each task. The trial is designed based on various k values and train-to-test dataset ratios with the purpose of exploring the best k values and splitting ratio combinations for optimizing predictive performance. The predictive accuracy based on the confusion matrices is summarized as below:



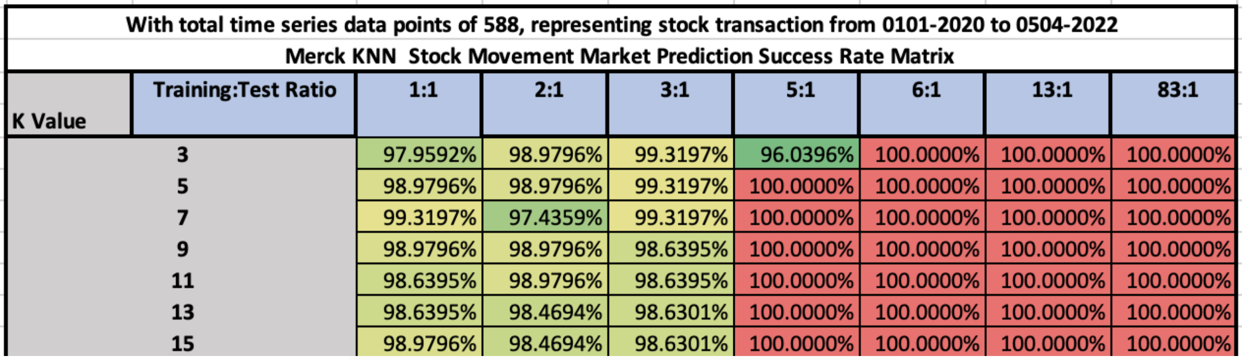
With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022							
Johnson&Johnson KNN Stock Movement Market Prediction Success Rate Matrix							
K Value / Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3	97.9592%	97.9487%	97.9592%	97.9381%	98.7805%	97.6191%	100.0000%
4	97.6191%	98.9744%	97.9592%	97.9381%	98.7805%	97.6191%	100.0000%
5	97.2789%	98.4615%	97.9592%	97.9381%	98.7805%	97.6191%	100.0000%
6	97.2789%	97.9487%	97.9592%	97.9381%	98.7805%	97.6191%	100.0000%
7	98.2993%	97.9487%	97.9592%	97.9381%	98.7805%	97.6191%	100.0000%
8	98.2993%	97.9487%	97.9592%	97.9381%	98.7805%	97.6191%	100.0000%
9	98.2993%	97.9487%	97.9592%	97.9381%	98.7805%	97.6191%	100.0000%

Table 2-1 Johnson & Johnson Accuracy Heatmap



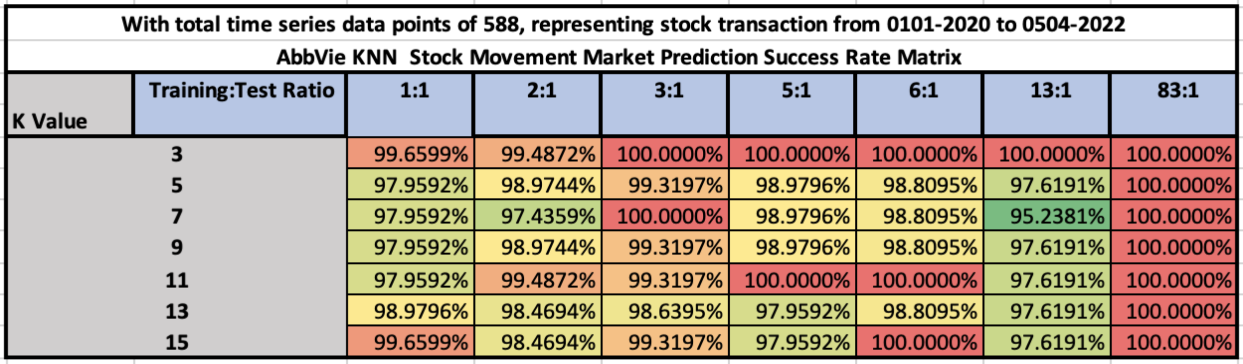
With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022								
Pfizer KNN Stock Movement Market Prediction Success Rate Matrix								
	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
K Value								
3		97.9592%	96.9231%	97.9592%	97.9381%	100.0000%	98.8095%	100.0000%
5		97.9592%	97.4359%	97.2603%	98.9691%	100.0000%	97.6191%	100.0000%
7		97.9592%	97.4359%	97.9452%	98.9691%	100.0000%	98.8095%	100.0000%
9		97.9592%	98.9744%	99.3103%	100.0000%	100.0000%	100.0000%	100.0000%
11		97.6190%	97.4359%	98.6301%	100.0000%	97.6190%	100.0000%	100.0000%
13		98.6301%	97.9487%	97.9452%	97.9381%	97.6190%	100.0000%	100.0000%
15		98.6301%	98.4615%	98.6395%	98.9691%	97.6190%	98.8095%	100.0000%

Table 2-2 Pfizer Accuracy Table Heatmap



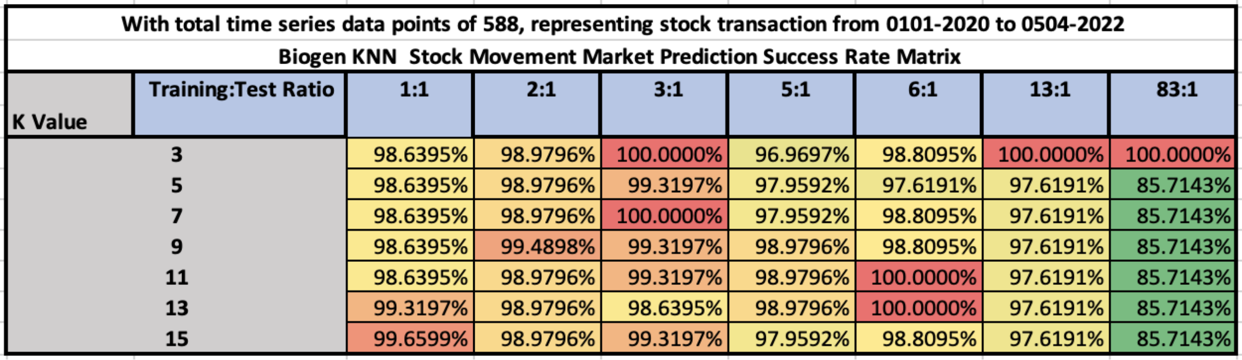
With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022								
Merck KNN Stock Movement Market Prediction Success Rate Matrix								
K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3		97.9592%	98.9796%	99.3197%	96.0396%	100.0000%	100.0000%	100.0000%
5		98.9796%	98.9796%	99.3197%	100.0000%	100.0000%	100.0000%	100.0000%
7		99.3197%	97.4359%	99.3197%	100.0000%	100.0000%	100.0000%	100.0000%
9		98.9796%	98.9796%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%
11		98.6395%	98.9796%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%
13		98.6395%	98.4694%	98.6301%	100.0000%	100.0000%	100.0000%	100.0000%
15		98.9796%	98.4694%	98.6301%	100.0000%	100.0000%	100.0000%	100.0000%

Table 2-3 Merck Accuracy Table Heatmap



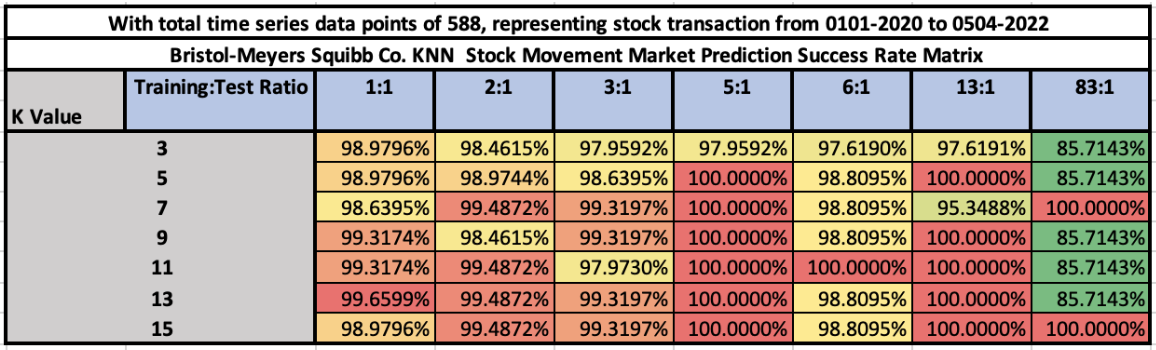
With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022								
AbbVie KNN Stock Movement Market Prediction Success Rate Matrix								
K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3		99.6599%	99.4872%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
5		97.9592%	98.9744%	99.3197%	98.9796%	98.8095%	97.6191%	100.0000%
7		97.9592%	97.4359%	100.0000%	98.9796%	98.8095%	95.2381%	100.0000%
9		97.9592%	98.9744%	99.3197%	98.9796%	98.8095%	97.6191%	100.0000%
11		97.9592%	99.4872%	99.3197%	100.0000%	100.0000%	97.6191%	100.0000%
13		98.9796%	98.4694%	98.6395%	97.9592%	98.8095%	97.6191%	100.0000%
15		99.6599%	98.4694%	99.3197%	97.9592%	100.0000%	97.6191%	100.0000%

Table 2-4 AbbVie Accuracy Table Heatmap



With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022								
Biogen KNN Stock Movement Market Prediction Success Rate Matrix								
K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3		98.6395%	98.9796%	100.0000%	96.9697%	98.8095%	100.0000%	100.0000%
5		98.6395%	98.9796%	99.3197%	97.9592%	97.6191%	97.6191%	85.7143%
7		98.6395%	98.9796%	100.0000%	97.9592%	98.8095%	97.6191%	85.7143%
9		98.6395%	99.4898%	99.3197%	98.9796%	98.8095%	97.6191%	85.7143%
11		98.6395%	98.9796%	99.3197%	98.9796%	100.0000%	97.6191%	85.7143%
13		99.3197%	98.9796%	98.6395%	98.9796%	100.0000%	97.6191%	85.7143%
15		99.6599%	98.9796%	99.3197%	97.9592%	98.8095%	97.6191%	85.7143%

Table 2-5 Biogen Accuracy Table Heatmap

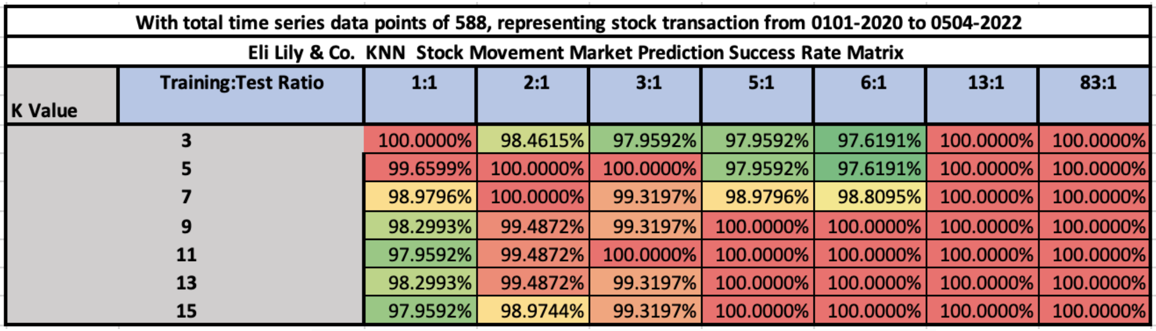


With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022

Bristol-Meyers Squibb Co. KNN Stock Movement Market Prediction Success Rate Matrix

K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3		98.9796%	98.4615%	97.9592%	97.9592%	97.6190%	97.6191%	85.7143%
5		98.9796%	98.9744%	98.6395%	100.0000%	98.8095%	100.0000%	85.7143%
7		98.6395%	99.4872%	99.3197%	100.0000%	98.8095%	95.3488%	100.0000%
9		99.3174%	98.4615%	99.3197%	100.0000%	98.8095%	100.0000%	85.7143%
11		99.3174%	99.4872%	97.9730%	100.0000%	100.0000%	100.0000%	85.7143%
13		99.6599%	99.4872%	99.3197%	100.0000%	98.8095%	100.0000%	85.7143%
15		98.9796%	99.4872%	99.3197%	100.0000%	98.8095%	100.0000%	100.0000%

Table 2-6 Bristol-Meyers Squibb Co. Accuracy Table Heatmap

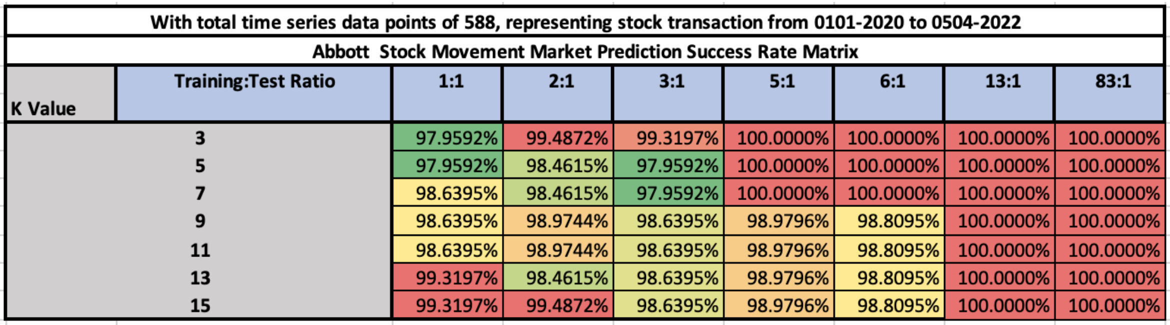


With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022

Eli Lily & Co. KNN Stock Movement Market Prediction Success Rate Matrix

K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3		100.0000%	98.4615%	97.9592%	97.9592%	97.6191%	100.0000%	100.0000%
5		99.6599%	100.0000%	100.0000%	97.9592%	97.6191%	100.0000%	100.0000%
7		98.9796%	100.0000%	99.3197%	98.9796%	98.8095%	100.0000%	100.0000%
9		98.2993%	99.4872%	99.3197%	100.0000%	100.0000%	100.0000%	100.0000%
11		97.9592%	99.4872%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
13		98.2993%	99.4872%	99.3197%	100.0000%	100.0000%	100.0000%	100.0000%
15		97.9592%	98.9744%	99.3197%	100.0000%	100.0000%	100.0000%	100.0000%

Table 2-7 Eli Lily&Co Accuracy Table Heatmap

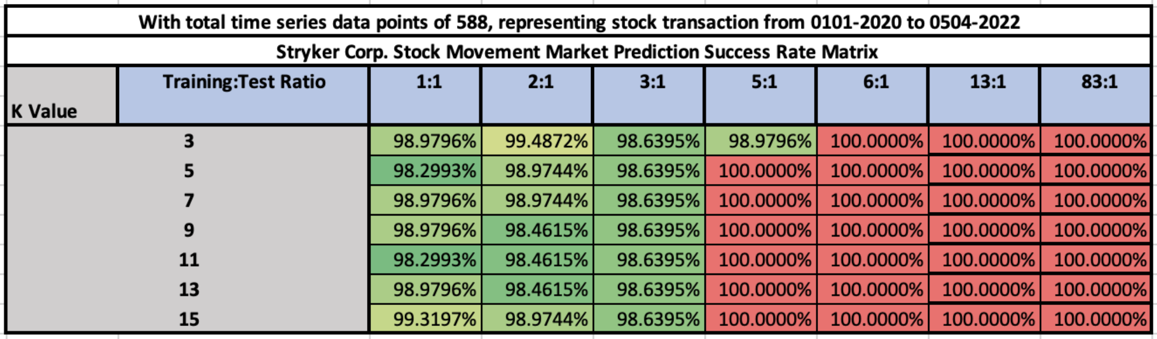


With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022

Abbott Stock Movement Market Prediction Success Rate Matrix

K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3		97.9592%	99.4872%	99.3197%	100.0000%	100.0000%	100.0000%	100.0000%
5		97.9592%	98.4615%	97.9592%	100.0000%	100.0000%	100.0000%	100.0000%
7		98.6395%	98.4615%	97.9592%	100.0000%	100.0000%	100.0000%	100.0000%
9		98.6395%	98.9744%	98.6395%	98.9796%	98.8095%	100.0000%	100.0000%
11		98.6395%	98.9744%	98.6395%	98.9796%	98.8095%	100.0000%	100.0000%
13		99.3197%	98.4615%	98.6395%	98.9796%	98.8095%	100.0000%	100.0000%
15		99.3197%	99.4872%	98.6395%	98.9796%	98.8095%	100.0000%	100.0000%

Table 2-8 Abbott Laboratories Accuracy Table Heatmap

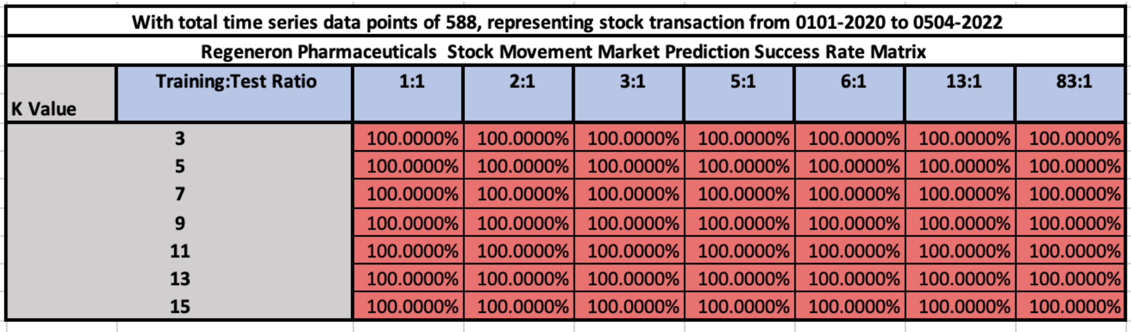


With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022

Stryker Corp. Stock Movement Market Prediction Success Rate Matrix

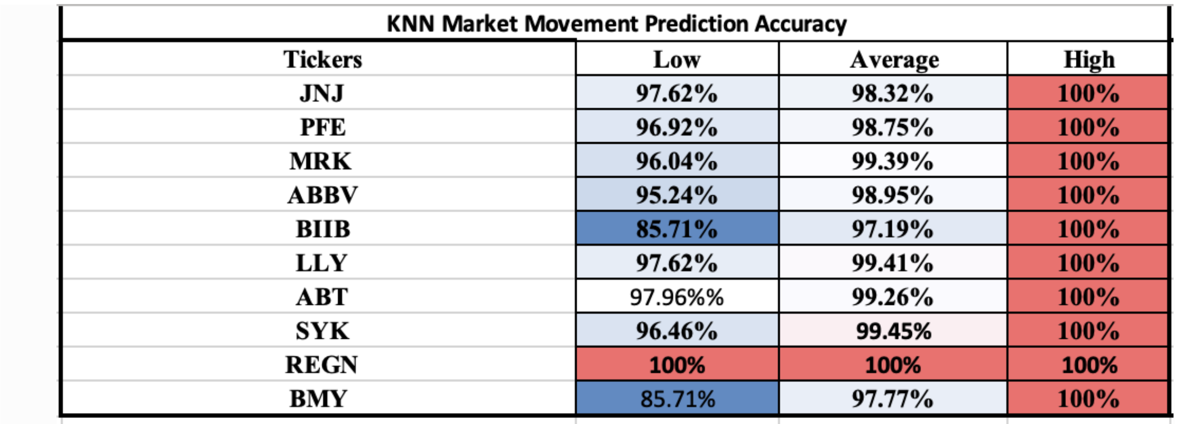
K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
3		98.9796%	99.4872%	98.6395%	98.9796%	100.0000%	100.0000%	100.0000%
5		98.2993%	98.9744%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%
7		98.9796%	98.9744%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%
9		98.9796%	98.4615%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%
11		98.2993%	98.4615%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%
13		98.9796%	98.4615%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%
15		99.3197%	98.9744%	98.6395%	100.0000%	100.0000%	100.0000%	100.0000%

Table 2-9 Stryker Corp Accuracy Table Heatmap



With total time series data points of 588, representing stock transaction from 0101-2020 to 0504-2022								
Regeneron Pharmaceuticals Stock Movement Market Prediction Success Rate Matrix								
K Value	Training:Test Ratio	1:1	2:1	3:1	5:1	6:1	13:1	83:1
	3	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
	5	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
	7	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
	9	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
	11	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
	13	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%
	15	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%	100.0000%

Table 2-10 Regeneron Pharmaceuticals Accuracy Table Heatmap



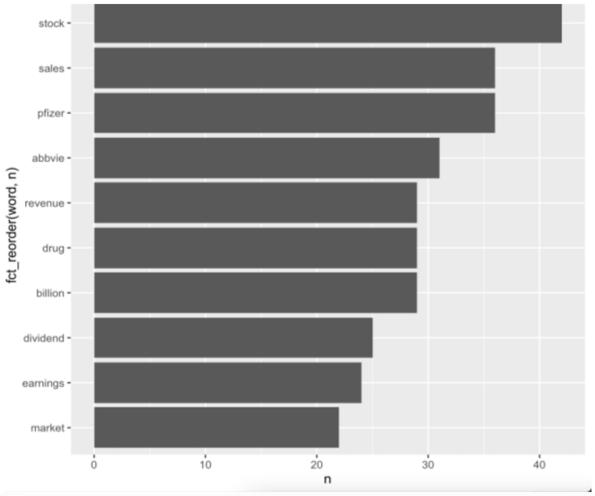
KNN Market Movement Prediction Accuracy			
Tickers	Low	Average	High
JNJ	97.62%	98.32%	100%
PFE	96.92%	98.75%	100%
MRK	96.04%	99.39%	100%
ABBV	95.24%	98.95%	100%
BIIB	85.71%	97.19%	100%
LLY	97.62%	99.41%	100%
ABT	97.96%%	99.26%	100%
SYK	96.46%	99.45%	100%
REGN	100%	100%	100%
BMY	85.71%	97.77%	100%

Table 2-11 KNN Market Movement Prediction Accuracy per Ticker

Based on KNN prediction and trials with different K values and splitting ratios, we found that KNN achieves high accuracy for predicting stock market movements based on the features LogRange and return_Perc. The lowest accuracy is 85.71% whereas the highest is 100%. The more data points are contained in the dataset, the more accurate the test results will be. Such a finding allows potential investors to use LogRange and return_Perc to detect market movement and build positions based on their trading direction.

Data Mining Task III: Text Mining to Gain Qualitative Insights based on Term Frequency

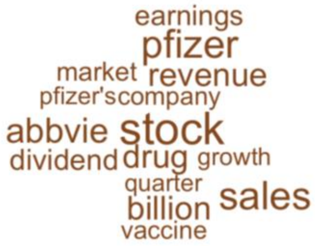
The dataset we utilized in this task is an unstructured news dataset in txt format. The news was collected from Motley Fool, Barron's, Bloomberg, Reuters, Yahoo Finance, and MarketWatch. All the news is pharmaceutical-related stock news. The news collection window is between 02.28.2022 and 05.04.2022. The purpose of text mining is to gain qualitative insights based on term frequency with the goal of identifying price-boosting events and stock popularity. Several word clouds are shown below after excluding stop words:



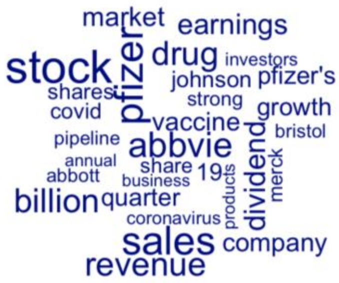
Graph 3-1 Barplot of top 10 most frequent words in News Dataset

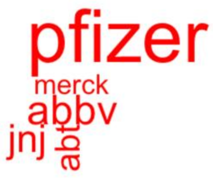
The top 2 companies mentioned in the news are Pfizer and AbbVie, which contradicted the popular opinion that Johnson & Johnson might be the most popular pharmaceutical stock even though J&J ranked No.1 in terms of revenue and market share within the pharmaceutical industry. Other frequent words mentioned are revenue, drug, dividend, billion, earnings, vaccine, etc. These indicated that investors care about dividend payment, revenue performance, and the drug product that the company associates with.

market
dividend
revenue
drug
billion
abbvie
pfizer
earnings
stock
sales



Graph 3-2 Top 10 frequent word WordCloud Top 15 frequent word WordCloud





Top30 frequent word WordCloud Top 5 Companies Frequently Mentioned

By increasing the word cloud coverage from 10 to 15 and then 30 words, we can see more qualitative details that quantitative data mining tasks fail to communicate. However, the interpretation of such qualitative information varies from individual to individual. Text mining can be utilized to extract meaningful business information of qualitative nature, and they direct people to search for information based on these frequent words. For example, a pharmaceutical investor might further research the drug product, vaccine offering, and production pipeline related to a certain company.

Conclusion, Discussion, and Future Improvement

Conclusion and Contribution

1. We contradicted the popular opinion that simple linear regression can be applied to predict stock price, and we found that what simple linear can do is provide an overview of a general market trend in the long term: Bull or Bear. To predict price, it is practical to set a target return and risk indicator based on "LogRange" and "return_Percentage", and then utilize a histogram plot to explore when a stock's return% or risk levels are deviated from its distribution mean. This will allow investors to earn extraordinary profits by detecting rare deviations. Therefore, we contribute a new mechanism, which is utilizing integrated efforts of linear regression and histogram plot of LogRange and return_Perc.
2. Out selection of features achieved high classification successful rate for predicting stock market movement. By including percentage_return and LogRange as features in the KNN training dataset, the test result performed well with accuracy averaged at 97.52% for "Up" or "Down" classification. KNN performs well for the stock market classification problem. It can be used to predict next-day movements if investors can come up with an expected value of LogRange and percentage_return; then let KNN return a class label based on the new LogRange and percentage_Return
3. Unlike other quantitative-intensive research, we incorporate text mining for extracting human-friendly qualitative insights for stock investment.

Discussion: The Rationale of Selection of Certain Data Mining Tasks and Why not Others

1. Why not use multilinear regression: We did conduct trials of multilinear regression. But multilinear regression had a similar issue to simple linear regression: Multiple regression assumes that the residuals are normally distributed; but in our cases, a high residual is found in multilinear regression and the homoscedasticity condition is violated. The model will perform poorly.
2. Why not use decision trees: decision trees can achieve good accuracy for training data, but poor accuracy for unseen samples. During COVID-19, pharmaceutical company stock prices fluctuated with the rollout of vaccines and the delta variant outbreak. Within the selected period, there are 2 price slumps and 2 price rebounds that were out of pattern due to the COVID-19 official outbreak and the delta variant outbreak. Based on the volatile trend of stock prices, we assumed the decision tree will perform poorly especially when we try to split datasets when the test dataset contains datasets that acted out of the pattern.
3. Why not SVM: the boundary is tight between the LogRange vs return_Perc bivariate scatterplot, where the hypothetical line might get across both "Up" class and "Down" class points, so the boundary will lose effectiveness for dividing 2 classes.

4. Why not Binary Logistic Regression: Logistic regression is sensitive to outliers. During the COVID-19 period, stock price frequently acts out of pattern with significant events that occurred (outbreak of delta variant, Russian-Ukraine War).
5. Why not K-means: Based on a literature survey, K-means perform poorly for predicting stock performance in terms of return and it is time-consuming to select proper variables for the K-means clustering, which violates the cost-effective principle. Although k-means can be conducted using EPS, PE, dividend, etc; these variables are not released frequently, which may fail to provide enough data for analysis. Besides, it is controversial to choose proper variables for the k-means clustering within the realm of finance.

Future Improvement

If the project continues, we would like to apply simple linear regression and KNN to a wide variety of companies and industries outside of pharmaceutical companies to explore the predictive quality of the two approaches. Secondly, we would like to explore how the LSTM model can be applied to examine the performance for predicting stock price and movement.

Reference

[1] Khalid,H.,Ismail,M.&Ali,S. “Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm,” 2013 International Journal of Business, Humanities and Technology Vol. 3 No. 3, https://www.ijbhtnet.com/journals/Vol_3_No_3_March_2013/4.pdf.

[2] Shynkevich, Yauheniya et al. “Forecasting Price Movements using Technical Indicators: Investigating Window Size Effects.” (2018).

[3]Shu Bin, “K-Means Stock Clustering Analysis Based on Historical Price Movements and Financial Ratios.”(2020).

[4] R. Puspadini, H. Mawengkang and S. Efendi, "Feature Selection on K-Nearest Neighbor Algorithm Using Similarity Measure," *2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*, 2020, pp. 226-231, doi: 10.1109/MECnIT48290.2020.9166612.

[5]Ali, N., Neagu, D. & Trundle, P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Appl. Sci.***1,** 1559 (2019). https://doi.org/10.1007/s42452-019-1356-9

[6] Kamaladdin,Wei Liu et al. “Text-based Stock Market Analysis: A Review.”(2021).https://arxiv.org/pdf/2106.12985.pdf