U.S. Stock Analysis

Group 1

ABSTRACT

In this project, the patterns within U.S. stock data between 2000 and 2019 are explored through two major data mining techniques: clustering and classification. K-Means Algorithm is performed on the data based on adjusted close price. As a result, four clusters separated by stock prices are found. The algorithm is performed again based on the volatility of stocks, five clusters are found. During association rule mining, six association rules are generated. Furthermore, four attributes are visualized, and two attribute pairs are pairwise compared to examine any linear association between them. The relation between economics and stock transactions is explored during this . The result shows that there is no linear relation between economic data and stock transaction.

1. INTRODUCTION & MOTIVATION

The stock market is very volatile and goes up and down based on various reasons. Our goal for this project was to analyze the data from the year 2000 onwards to highlight some hidden patterns in the data so that this information could be used to make better trading decisions, and thus maximise the profit gained from investing. In Section 2, we provide a detailed explanation of our dataset, including the project set up and attribute description. Section 3 has four subsections; each introduces an approach or model we use for analyzing the target dataset. In Section 4, we discuss the observations of our analysis result of each approach or model. Section 5 talks about what we have learned from this project. Moreover, in Section 6, we discuss our current status in this project and what can be done in the future.

2. DESIGN

2.1 Project Setup

The stock data is pulled from yfinance[2], a Yahoo Finance API, and Quandl[1]. Moreover, we also pulled general economic data, the yearly GDP and monthly IDP of the U.S., from the World Bank and the Federal Reserve Bank of St Louis [3]. All the data have a period from 2000 to 2019 and have been cleaned and stored on a MySQL database. We have realized the existence of dividends and splits. Only using open, high, low, and close price of a stock on a specific date, which are directly pulled from our sources, would not give an accurate result in this project. Therefore, we introduce four extra features: adjusted open, adjusted high, adjusted low, and adjusted close price for each instance. These are prices adjusted after dividends and splits happened.

2.2 Dataset Description

This data uses the following elements.

1. `Date:` The date a specific stock trading took place.
2. `Open:` The start price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. It is used to calculate Adj. Open, which will be used in our analysis.
3. `High:` The highest price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. It is used to calculate Adj. High, which will be used in our analysis.
4. `Low:` The lowest price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. It is used to calculate Adj. Low, which will be used in our analysis.
5. `Close:` The final price of a specific stock on a specific date. The value of this attribute is pulled from our sources, but it is not used in any of our analysis directly. The actual attribute we will use is Adj. Close.
6. `Volume:` Total number of stocks traded over the course of a day.
7. `Dividends:` The reward for the stockholders given by the corporation; usually rewarded quarterly or twice a year or monthly, some companies offer a one time dividend across the year. It is usually done when the company is booming with profits. The rewards are offered either in form of cash or some extra stocks(also known as splits).
8. `Company:` The company name of this specific stock. Each instance should have its unique Date and Company value. That is, there is only one stock record per day for each company.
9. `Adjusted closing price:` The adjusted close price of this specific stock on this specific date after a dividend and split happened.These values are calculated using the splits in the stocks.Using these values other adjusted values can be calculated.

10. `Adjusted Open price:` The adjusted open price of this specific stock on this specific date after a dividend and split happened. The data pulled from yfinance does not contain this attribute. Therefore, any instance which has null value on this attribute will be filled up a value through the equation:

$$Adj.\ Open = \frac{Adj.\ Close * Open}{Close} \quad (1)$$

11. `Adjusted High price:` The adjusted high price of this specific stock on this specific date after a dividend and split happened. The data pulled from yfinance does not contain this attribute. Therefore, any instance which has null value on this attribute will be filled up a value through the equation:

$$Adj.\ High = \frac{Adj.\ Close * High}{Close} \quad (2)$$

12. `Adjusted Low price:` The adjusted low price of this specific stock on this specific date after a dividend and split happened. The data pulled from yfinance does not contain this attribute. Therefore, any instance which has null value on this attribute will be filled up a value through the equation:

$$Adj.\ Low = \frac{Adj.\ Close * Low}{Close} \quad (3)$$

13. `Splits:` A ratio indicates how many more stock a stockholder now has after a split happens. It is a rare occurrence and 1 is a placeholder of this attribute indicates there is no split happened on that date. In order to combat over inflated stock prices a company will sometimes perform a split. This means that it will (usually) decrease the price of the stock in exchange for adding more of it into circulation.

14. `GDP:` Gross domestic product: the total value of goods produced and services provided in a country during one year in US Dollars. Simple way of evaluating if this particular trade occurred during a net positive or net negative year, can be used alongside aggregated stock data to show how reliant stock price is on the economy at large.

15. `Industrial Output Index:` A percent production of Industrial production is a measure of output of manufacturing based industries, including those producing goods for consumers and businesses based on the output of the same for 2012.Correlation (or lack thereof) can show how reliant a given company is on demand for US manufactured goods. Because the US has changed to being a service industry over time this is an important statistic as it can indicate the viability of this business into the going future.

3. IMPLEMENTATION

In this study, two data mining techniques: clustering and classification, are using to explore the patterns of the dataset. Furthermore, four attributes are visualized, and two attribute pairs are pairwise compared to examine any linear association between them.

3.1 Classification

For the classification, since we are not having a nominal target attribute in the data, we tried to focus on prediction of the stock based on the historic data. This is done by choosing the prior 30 day adjusted closing price as input and 31st adjusted closing price as output. So we get our training data using a sliding window approach. Currently we have chosen linear regression as our machine learning model. The model seems to work good with stable stocks, such as Apple Inc. Unfortunately the model fails predicting highly volatile stocks like Amazon and Alphabet .inc. Therefore, we are still developing the model, trying to generate a general model that works under all circumstances.

3.2 Clustering

To cluster the data, we plan to work on the Adj. Close attribute with own developed program, as this gives us the closing price of the stock everyday. For each of the 49 companies, we have computed mean, standard deviation, maximum value and minimum value for the adjusted close attribute. Then, with these values, we plan on clustering the data with two goals :

1. Establish clusters of companies based on their prices (Low, Medium, High and very High);

2. Establish clusters of companies based on how volatile and how fluctuating the stock is. This can be done by using the standard deviation attribute as it shows how much the closing values vary from the mean of the closing values. We intend to use K-Means to perform the clustering, and determine the value of k by using the elbow plot.

3.3 Association Rule Mining

The goal of this approach was to derive rules that described which stocks rise together and how they are related. First the data had to be transformed. For stock i,For Day j, if $Adjusted\_close[i][j] >= Adjusted\_close[i][j-1]$, then we append the name of the stock to the result. This data trasnformation was done on all 49 companies over the year 2008. Total number of rows in the transformed data were 252 records for 252 days of the stock market being open in the year 2008. Few entries of the transformed data are shown below:

- Day 2: 'Apple Inc', 'Microsoft Corp', 'Alphabet Inc', 'Cisco Systems Inc', 'Adobe Inc', 'PepsiCo Inc', 'Gilead Sciences Inc', 'Biogen Inc', 'Exelon Corp'
- Day 3: 'NVIDIA Corp', 'PepsiCo Inc', 'Amgen Inc', 'T-Mobile US Inc', 'Texas Instruments Inc', 'Starbucks Corp', 'Vertex Pharmaceuticals Inc', 'Exelon Corp', 'Monster Beverage Corp'
- Day 4: 'T-Mobile US Inc', 'Intuit Inc', 'Vertex Pharmaceuticals Inc', 'Applied Materials Inc'

This data is then fed into the Apriori algorithm with support = 0.3 and confidence = 0.9.

3.4 Visualizations of Attributes

Since all of the attributes in the dataset are numerical values and separated by companies, it is easy to visualize them. We choose to visualize four attributes using Microsoft

Excel: Open, Close, High, and Low, based on four randomly chosen companies: Illumina Inc, Biogen Inc, Exelon Corp, and Ross Stores Inc. We expect to see similar patterns each company may have with its stock price path among these four attributes.

3.5 Pairwise Comparison

We pairwise compared two attribute pairs using Weka and Microsoft Excel. The method in used is linear regression with a confidence interval of 95%. First is the adjust open and the adjust close price pair. In common, if the open price of a stock is high, the close price of that stock should also be high. Thus, we expected to see a strong positive linear relationship between them. Second, we wanted to examine the association between the economy and the stock market transactions. A new attribute is used in this comparison– the average annual stock transfer volume. It is calculated as:

$$avgVol = \frac{sum\ of\ volume\ of\ all\ stocks\ in\ a\ specific\ year}{number\ of\ stock\ record\ in\ that\ year} \quad (4)$$

The annual GDP and the average annual stock transfer volume is pairwise compared. We expected to see a positive linear relationship in this comparison, that is, when the economy went well, there were more stock transactions in the market.

4. RESULTS

4.1 Classification

Here we are using the linear regression to predict the stock prices. There are some known techniques to predict the stocks based on their historical values like moving-average, adjusted moving average, however we use linear regression to get the weights for each historical entry which will be different for each stock.

The RMSE achieved using this technique for all 50 companies is shown using a par graph below.

The better illustration Of stock prediction can be shown using the graph of actual closing prices and the predicted closing prices for a particular stock.

4.2 Clustering

To determine the number of clusters for the classification on adjusted closing price, the SSE vs K cluster graph as seen in Figure 3 was calculated and from the graph, k was selected to be 4.

K-Means clustering was then performed on the data and 4 meaningful clusters were found with no overlap. These centroid values are displayed in Table 1.

Cluster	Mean Adj. Price	StdDev Adj. Price	Minimum Adj. Price	Maximum Adj.Price
0	108.874	101.262	9.264	440.913
1	2149.142	37.526	2077.56	2206.09
2	37.077	24.398	9.706	115.979
3	354.751	285.693	28.064	1392.975

Table 1: Centroid Values of Each Cluster

The table shows that the stocks can be divided into low priced stocks, medium priced stocks, high prices stocks and very high priced stocks. This data can be used to choose the correct investment strategy and in what price range of stocks one can invest.

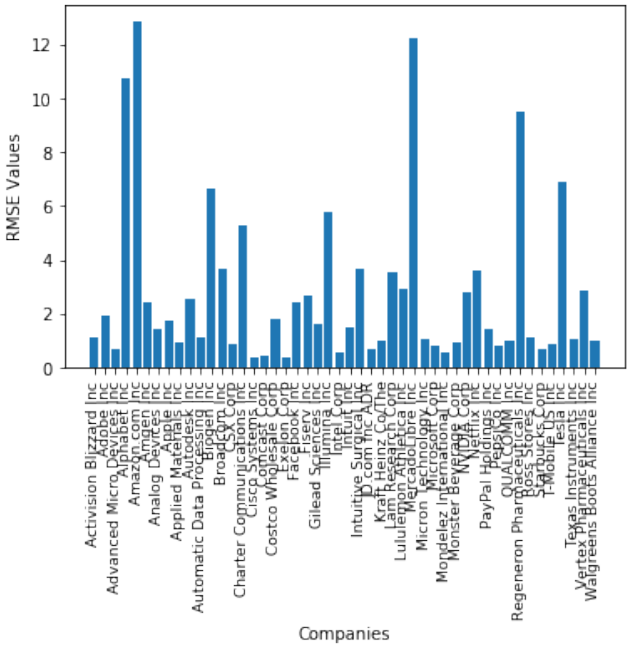

Figure 1: Stock vs RMSE

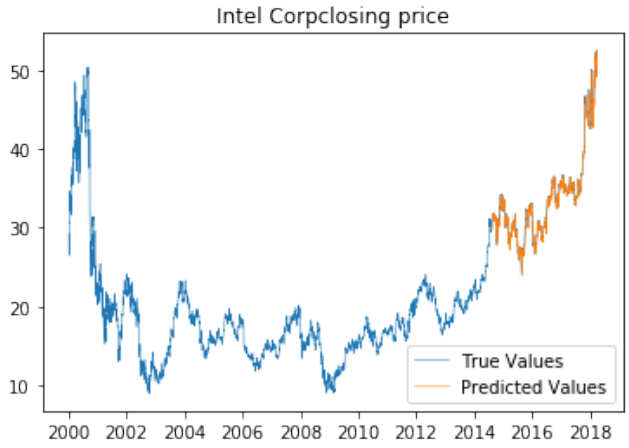

Figure 2: Time vs Adjusted Closing Price

To determine the number of clusters for the classification on the volatility of stocks, the SSE vs K cluster graph as seen in Figure 4 was calculated and from the graph, k was selected to be 5.

K-Means clustering was then perfomed on the data and 4 meaningful clusters were found with no overlap. There centroid values are displayed in Table 2.

The clusters are all unique, and they show that stocks can be divided into 5 different categories of volatility. These clusters can be used to decide where to invest in and what clusters can provide most returns.

4.3 Association Rule Mining

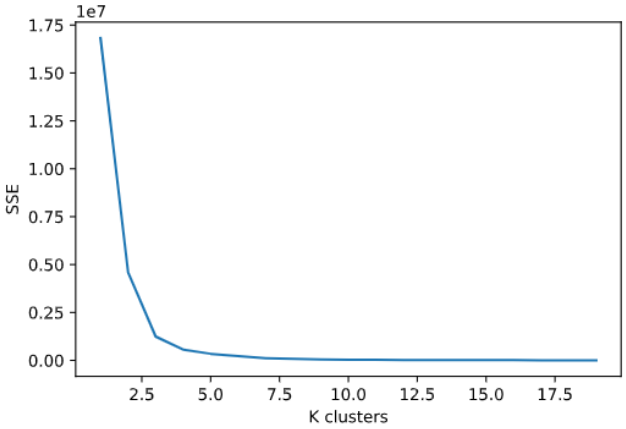

Figure 3: SSE vs K

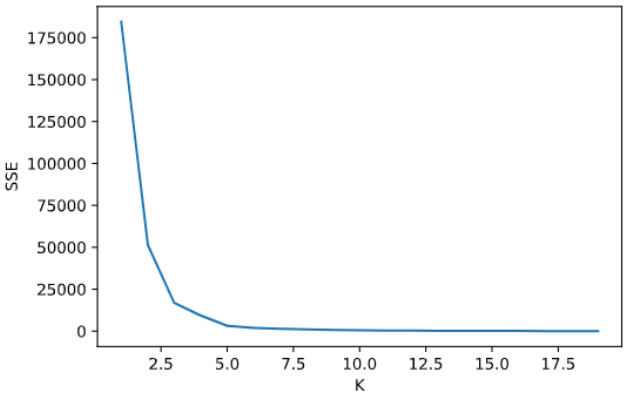

Figure 4: SSE vs K

Cluster	Mean Adj. Price	StdDev Adj. Price
0	182.576	43.112
1	148.894	168.933
2	121.969	100.465
3	354.751	285.693
4	30.158	16.299

Table 2: Mean and Std Dev of Adj. Price of Each Cluster

For data from the year 2008, and support is 0.3 and confidence is 0.9, we derived 6 rules in total:

- {Cisco Systems Inc, Comcast Corp} -> {Automatic Data Processing Inc}
- {Comcast Corp, Costco Wholesale Corp} -> {Automatic Data Processing Inc}
- {Comcast Corp, Fiserv Inc} -> {Automatic Data Processing Inc}
- {Intel Corp, Lam Research Corp} -> {Adobe Inc}
- {Analog Devices Inc, Intuit Inc} -> {Intel Corp}
- {Analog Devices Inc, Lam Research Corp} -> {Intel Corp}

With these rules we can now establish that certain stocks rise and fall in price together and this can help gain more profits from investment.

4.4 Visualizations of Attributes

The visualizations of Open, High, Low, and Close prices among the four randomly chosen companies (Illumina Inc, Biogen Inc, Exelcon Corp, and Ross Stores Inc) are shown in Figure 5, 6, 7, and 8.

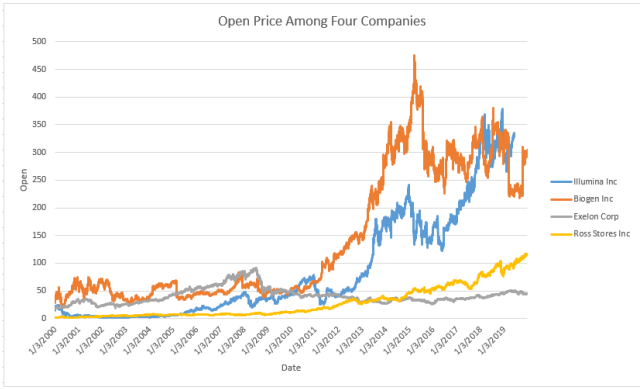

Figure 5: Visualization of Open Price Among Four Companies

Figure 6: Visualization of High Price Among Four Companies

As we expected, all four attributes have shown a same pattern for each company, which is not surprising because these four attributes are strongly positively correlated with each other, as the plot matrix generated by Weka shown in Figure 9.

4.5 Pairwise Comparison

4.5.1 *Pairwise Comparison Between Adj. Open and Adj. Close*

The plot graph of Adjusted Open vs. Adjusted Close is shown in Figure 10 and the linear regression statistics is shown in Figure 11.

The regression equation of the linear relation between adjusted open and adjusted close is y=0.9996x+0.035, with

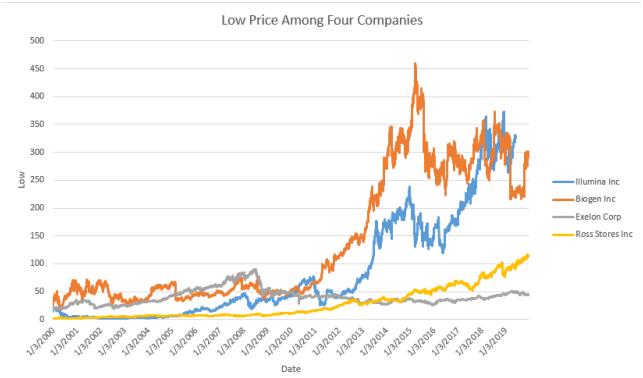

Figure 7: Visualization of Low Price Among Four Companies

Figure 8: Visualization of Close Price Among Four Companies

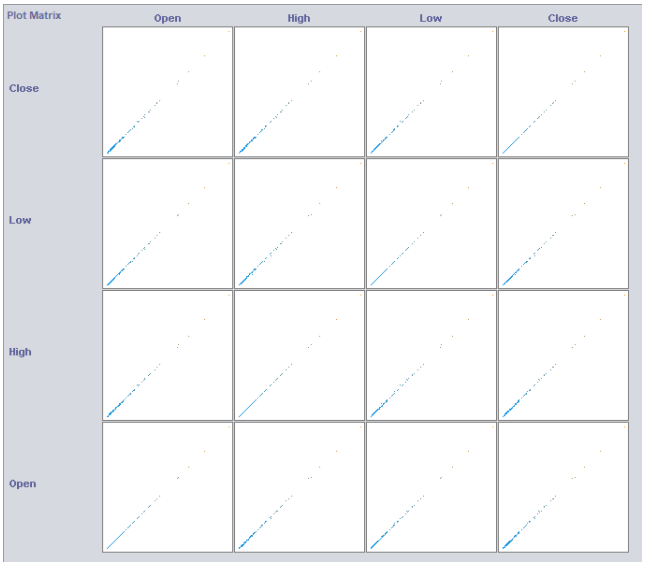

Figure 9: Plot Matrix of Close, Low, High, and Open Attributes

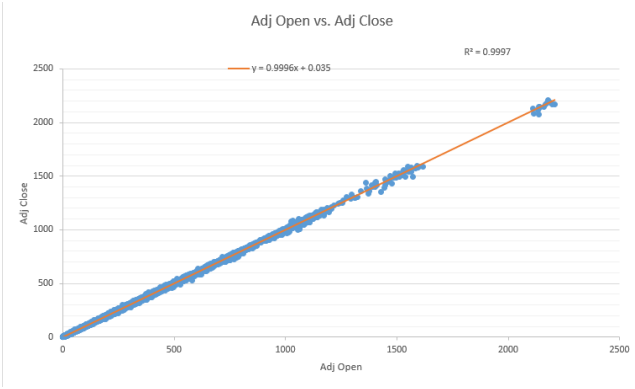

Figure 10: Plot Graph of Adj. Open vs. Adj. Close

Regression Statistics	
Multiple R	0.999873612
R^2	0.999747239
Adjusted R^2	0.999747237
Standard Error	2.130268867
Observations	119935

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2152721249	2152721249	474371901.7	0
Residual	119933	544261.4047	4.538045448		
Total	119934	2153265511			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.034960817	0.007021505	4.979106213	6.3967E-07	0.021198782	0.048722852
x Variable	0.999559099	4.58933E-05	21780.08039	0	0.999469149	0.999649049

Figure 11: Regression Statistics of Adj. Open vs. Adj. Close

a correlation coefficient of 0.9997. The standard error is 2.1303. With a 95% confidence interval, the lower intercept is only 0.0212, and the upper intercept is only 0.04872. Since the slope of this equation is very close to 1, and the correlation coefficient is very close to 1, we can say that there is a perfect positive linear relation between them. As the adjusted open price of a particular stock increased, the adjusted close price also increased. We have predicted this outcome because in general, if a stock has a high open price on a particular day, its close price should also be high. This comparison has provided this argument.

4.5.2 Pairwise Comparison Between GDP and Avg. Vol

The plot graph of annual GDP vs. average annual stock transfer volume is shown in Figure 12 and the linear regression statistics is shown in Figure 13.

The regression equation of the linear relation between annual GDP and average annual stock transfer volume is y=-0.0076x+6E+07, with a correlation coefficient of 0.1487. The standard error is 15777036. With a 95% confidence interval, the lower intercept is 5951484, and the upper intercept is 1.09E+08. Since the slope of this equation is very close to 0, and the correlation coefficient is very close to 0.1, we can say that there no linear relationship between them, which has surprised us. We were expecting to see if the economic indicators are potential factors affecting the stock during the design phase. We assumed that higher economic

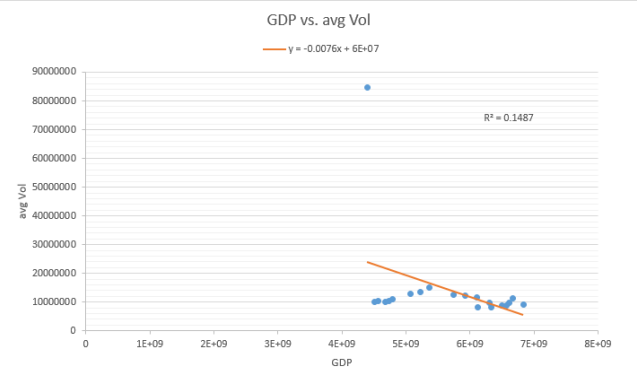

Figure 12: Plot Graph of GDP vs. Avg. Vol

Regression Statistics	
Multiple R	0.385658
R^2	0.148732
Adjusted R^2	0.101439
Standard Error	15777036
Observations	20

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7.82819E+14	7.82819E+14	3.144924805	0.093083706
Residual	18	4.48047E+15	2.48915E+14		
Total	19	5.26329E+15			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	57425691	24500769	2.343832	0.030765	5951484	1.09E+08
x Variable	-0.00761	0.004289	-1.77339	0.093084	-0.01662	0.001405

Figure 13: Regression Statistics of GDP vs. Avg. Vol

values, i.e., a better social economy, will lead to a higher stock transaction. However, the outcome of this comparison has shown that there is no such relation, at least not linear relation, between GDP and stock transaction volume.

5. LESSONS LEARNED

Although clustering does not give us any concrete information about which stocks to invest in, it does group stocks into clusters that can help us decide what cluster of stocks we can invest in, and it can also show which category a stock can belong to.

The results from association rule mining show that a pattern can be established to show stocks that rise together. This can help keep track of rising stocks and can help make informed decisions on which stocks to invest in.

The results of linear regression are promising however the predictions are not exact but close to the actual predictions. Certainly these results can be used in real time trading. The same technique can be used to predict the stock behaviour of future month or even a year.

From our pairwise comparison analysis, there is no linear relation between economic indicators and stock transactions. Thus, different from what we expected, economics is not a factor affects stock market.

6. CURRENT STATUS & FUTURE WORK

Currently the association rules generated are only for one year and only for the rise in stocks. Future work could include our entire dataset for more concrete rules, and also include rules that describe stocks that fall together in price. This would then give further information of which stocks to invest in so that one could make maximum profit. Moreover, in this project, only the U.S. stock records between 2000 and 2019 are analyzed. To do more reliable and accurate analyses on stock marketing, analyzing stock data from all over the world with a longer time period should be considered in the future.

7. REFERENCES

[1] Quandl Inc. *Quandl*, 2020 (accessed May 27, 2020).
[2] ranaroussi. *yfinance*, 2009 (accessed May 29, 2020).
[3] World Bank Group. *GDP/IDP*, 2020 (accessed May 30, 2020).