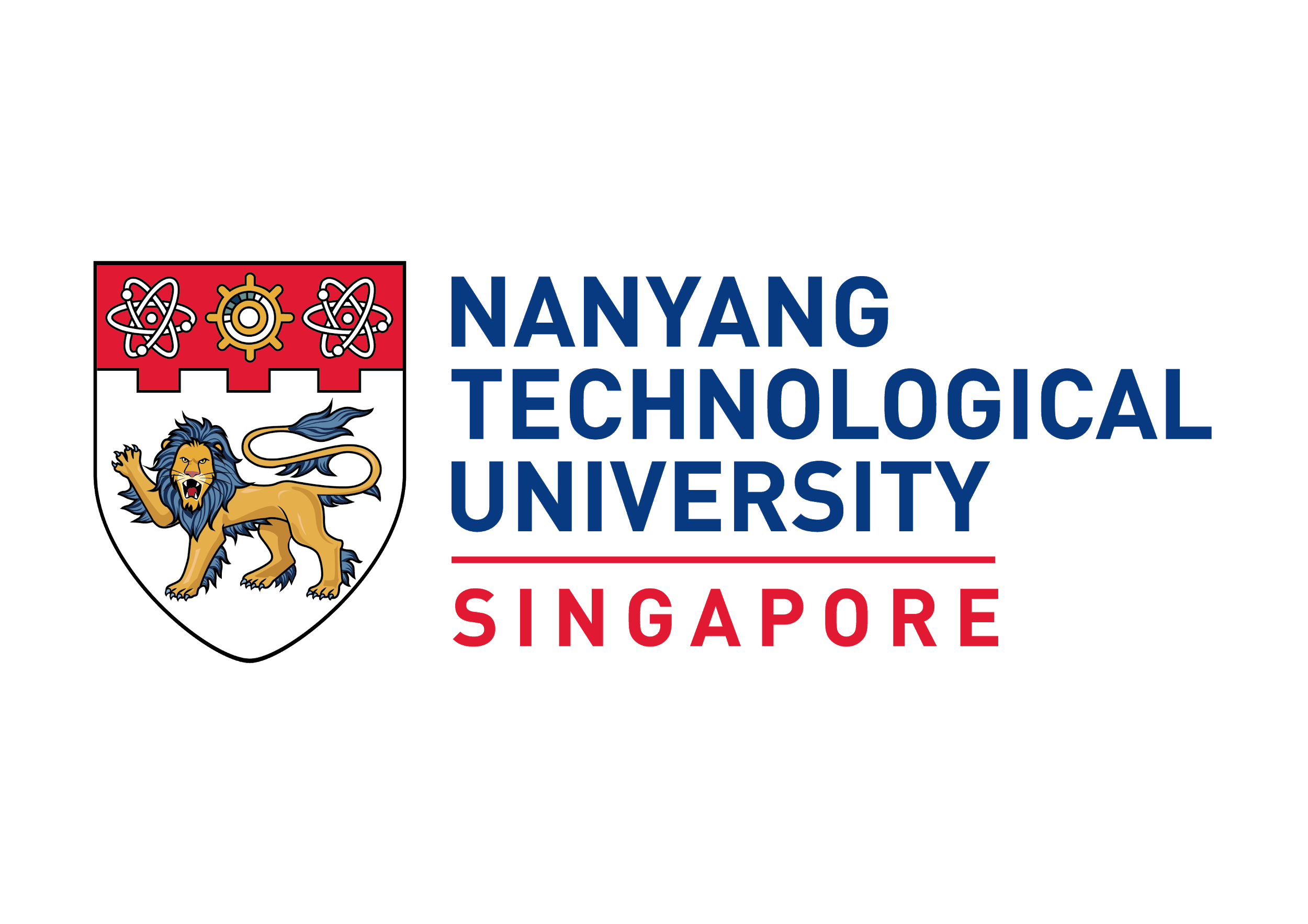
****

**BC2406 Analytics I**

**Project Report**

|  |  |
| --- | --- |
| **Team Members:** | |
| Tan Jun Hong | U1920048B |
| Chin Min Ray | U1910595L |
| Ryan Tan Yu Xiang | U1922774F |
| Ian Pang Yi En | U1921704F |
| Sherwin Lui Jia Xing | U1910064H |
| Shao Yakun | U1920578C |

**Table of Contents**

[Executive Summary 4](#_Toc55162699)

[1 Introduction 5](#_Toc55162700)

[1.1 Background/ Context 5](#_Toc55162701)

[1.2 Investment decision-making process 5](#_Toc55162702)

[a) Fundamental Analysis 5](#_Toc55162703)

[b) Technical Analysis 5](#_Toc55162704)

[1.3 Problem Identification for White Rock 5](#_Toc55162705)

[1.4 Our approach 7](#_Toc55162706)

[2 Solution to predict from the company’s internal financial performance 7](#_Toc55162707)

[2.1 Overview of Dataset & Choice of Variables 7](#_Toc55162708)

[2.2 Data Cleaning 9](#_Toc55162709)

[2.3 Data Exploration 10](#_Toc55162710)

[2.4 CART & Logistic Regression Model 11](#_Toc55162711)

[A. CART Model 11](#_Toc55162712)

[B. Logistic Regression Model 14](#_Toc55162713)

[2.5 Comparison & selecting the better model 15](#_Toc55162714)

[2.6 Prediction results & conclusion 16](#_Toc55162715)

[3 Solution to predict from external market patterns 17](#_Toc55162716)

[3.1 Overview of Dataset 17](#_Toc55162717)

[3.2 Data Cleaning & Choice of Variables 17](#_Toc55162718)

[3.3 Data Exploration 17](#_Toc55162719)

[a) Exploring stock price indicators 17](#_Toc55162720)

[b) Exploring trade volumes 18](#_Toc55162721)

[3.4 Time Series Model 18](#_Toc55162722)

[Step 1: Perform rolling mean on the dataset 18](#_Toc55162723)

[Step 2: Training the ARIMA model and checking its accuracy 19](#_Toc55162724)

[3.5 Prediction results & evaluations 19](#_Toc55162725)

[4 Insights & Analysis of both solutions 19](#_Toc55162726)

[4.1 Insights from Logistic Regression & CART models 19](#_Toc55162727)

[Key factor Influencing Investment Decision: 19](#_Toc55162728)

[Factor 1: Free Cash Flow Yield 19](#_Toc55162729)

[Factor 2: Operating Cash Flow Sales Ratio 20](#_Toc55162730)

[Factor 3: Return on Equity 20](#_Toc55162731)

[4.2 Analysis of the ARIMA Time Series Model 20](#_Toc55162732)

[a) Strengths 20](#_Toc55162733)

[b) Limitations 21](#_Toc55162734)

[5 Business Applications 21](#_Toc55162735)

[Step 1: Using ARIMA model to quickly filter out a group of good potential stocks 22](#_Toc55162736)

[Step 2: Using LR and CART to predict the stock price based on the stock’s own financial performance 22](#_Toc55162737)

[Step 3: Predictions 22](#_Toc55162738)

[6 Future Considerations 22](#_Toc55162739)

[7 Conclusion 23](#_Toc55162740)

[References 24](#_Toc55162741)

[Appendix 27](#_Toc55162742)

[Appendix A Justification for Variables 27](#_Toc55162743)

[Appendix B data cleaning for solution 1 29](#_Toc55162744)

[B.1 Bar plot of number of ‘NA’ values in each row 29](#_Toc55162745)

[B.2 Bar plot of number of ‘NA’ values in each column 30](#_Toc55162746)

[B.3 Boxplot for all variables to identify extreme data points 31](#_Toc55162747)

[Appendix C Solution 1 - Data Exploration 33](#_Toc55162748)

[C.1 Barplot for Distribution of Class 33](#_Toc55162749)

[C.2 Barplot for Distribution of Market Capitalization 33](#_Toc55162750)

[C.3 Barplot for Distribution of Class based on Market Capitalization 34](#_Toc55162751)

[C.4 Violin plot for Distribution of Price Variation based on Market Capitalization 34](#_Toc55162752)

[C.5 Barplot for Analyst Rating based on Market Capitalization 35](#_Toc55162753)

[Appendix D CART Models 36](#_Toc55162754)

[D.1 Variable importance of all 17 variables from initial CART 36](#_Toc55162755)

[D.2 Model 1 36](#_Toc55162756)

[D.3 Model 2 38](#_Toc55162757)

[D.4 Model 3 40](#_Toc55162758)

[Appendix E Logistic Regression Model 42](#_Toc55162759)

[E.1 VIF of logistic regression model 2 42](#_Toc55162760)

[E.2 VIF of logistic regression model 3 42](#_Toc55162761)

[E.3 Logistic regression model train set accuracy 42](#_Toc55162762)

[E.4 Logistic regression model train set accuracy 42](#_Toc55162763)

[Appendix F Solution 2 - Data Exploration 43](#_Toc55162764)

[F.1 Graphs of individual metrics against time 43](#_Toc55162765)

[F.2 Financial barchart (which incorporates high, low, open, close) against time 43](#_Toc55162766)

[F.3 Graph of average stock price against time 44](#_Toc55162767)

[F.4 Graphs of trade volume against time 44](#_Toc55162768)

[F.5 Linear Regression graphs of average stock price against trade volume 45](#_Toc55162769)

[Appendix G ARIMA Time Series Model 47](#_Toc55162770)

[G.1 Graphs of rolling mean of individual metrics against time 47](#_Toc55162771)

[G.2 Results of ARIMA Time Series model 48](#_Toc55162772)

[Appendix H Insights analysis 52](#_Toc55162773)

[H.1 Distribution of Free Cash Flow Yield by Class 52](#_Toc55162774)

[H.2 Distribution of Operating Cash Flow Sales Ratio by Class 52](#_Toc55162775)

[H.3 Distribution of Return On Equity by Class 53](#_Toc55162776)

# 

# Executive Summary

This report aims to recommend strategies to enhance White Rock’s decision making process and reduce risk in technology stock investment using accurate and timely predictions of technology stock price movements through data-driven solutions.

Our team outlined an overview of the current investment decision-making approaches used by investors, namely fundamental and technical analysis. Based on the assumption that White Rock mainly utilises fundamental analysis, we identified areas for improvement using technical analysis. We identified 2 issues with over-reliance on fundamental analysis, namely the inaccuracies and inefficiencies in predictions due to the high volatility of technology stocks and the need to consider external stock market patterns.

Our first strategy involves utilising classification models to predict changes in stock prices based on the selected financial indicators. Notable discoveries were made on the variable ‘Class’ and its relationship with ‘Market Capitalization’, as well as ‘Price Variation’ and its relationship with ‘Market Capitalization’. We used both CART and Logistic Regression to develop models in identifying the financial indicators that are most significant in affecting stock prices. Subsequently, we verified the accuracy of our findings using the test set and obtained an accuracy of 69.7% and 63.2% for the CART and Logistic Regression models respectively. Due to a higher prediction accuracy, our team concluded to employ CART and its corresponding variables - ‘Free Cash Flow Yield, ‘Operating Cash Flow Sales Ratio’ and ‘Return on Equity’ - to identify profitable investment opportunities in technology stocks.

Our second strategy involves utilising Auto Regressive Integrated Moving Average (ARIMA) to perform time series forecasting on short-term stock price movements based on the selected stock’s historical price data. We employed the ARIMA model on the test set to predict future stock prices and plotted the forecasted prices against the actual prices. The forecasted data closely matched and captured the actual data within the error range of the model which implies that the ARIMA model is capable of accurately forecasting stock prices.

The 2 above-mentioned strategies are to be used in conjunction; the ARIMA model is used first to forecast the future price trends of the identified stock then using the CART model to validate the results obtained from applying the ARIMA model. By combining the 2 strategies, White Rock can integrate the benefits of CART with the ARIMA model to overcome the 2 issues identified as mentioned above.

# 1 Introduction

## 1.1 Background/ Context

Technology stocks are one of the most popular sectors that asset managers and individual investors look into due to their high growth potential (Osterland, 2015). However, they are notorious for being highly volatile as compared to other sectors (CapitalGroup, 2020). Therefore, managing investment risk and making the right investment decisions are particularly crucial for investors looking to invest in technology stocks.

This paper dives deep into the **decision-making process of investing in technology stocks** by first looking at the commonly deployed strategies by investors, and seeks to recommend solutions that White Rock could utilise to make faster and more informed decisions when investing in technology stocks with minimal risks.

## 1.2 Investment decision-making process

There are two different approaches when it comes to the investment of equities, mainly ***Fundamental Analysis*** and ***Technical Analysis****.*

### Fundamental Analysis

The most commonly adopted strategy when investing in stocks is to conduct *Fundamental Analysis* of the potential investment opportunities. Fundamental Analysis (FA) is a stock trading strategy that attempts to determine the value of a stock by analysing its revenue, expenses, growth prospects and the competitive landscape (Bonga 2015). Through conducting an FA, one is able to obtain a better understanding of the company's business model, its current business performance, as well as future prospects (Euroinvestor 2012).

### Technical Analysis

On the other end of the spectrum is technical analysis, where more emphasis is placed on statistics and data analysis rather than understanding the company’s performance and the market. Investors taking such an approach will analyse the stock from a purely statistical standpoint, where various mathematical or statistical models are used to analyze the patterns of stock prices and volume traded.

## 1.3 Problem Identification for White Rock

As the technical analysis approach typically requires strong statistical and data analytics skills and the fact that equity investors mostly consist of finance experts and business development professionals who are not professionals in data, fundamental analysis is more commonly adopted by traditional asset management firms such as White Rock[[1]](#footnote-1).

While the fundamental analysis does provide a comprehensive understanding of a potential stock, considering the investment target to be technology stocks, our team has recognized **two** major issues with this existing approach:

1. **Inaccuracy and inefficiency in predictions due to the high-volatility nature of technology stocks**

A research done by Dimensional Fund Advisors suggests that any attempt to forecast short-term price movement in a volatile market is unlikely to be successful, due to the very nature of stocks being highly volatile and unpredictable. (Dimensional, 2016)

Considering the high volatility and frequent fluctuations of technology stocks, coupled with the fast evolution of technology companies over the years, traditional methods involving manual intervention are becoming increasingly inefficient in terms of time and resources. This is due to the large amount of data that needs to be monitored and processed among numerous technology companies with investment potential.

1. **The need to consider external stock market patterns**

Investment decisions on stocks do not only depend on its price and value obtained from fundamental analysis, but also the external buying and selling patterns within the stock market. The prices of stock may drop due to lesser buys and more sells and vice versa. Also, the buy/sell behaviours of investors may not be directly related to the stock’s price and value, but more about the expectations on future stock price patterns.

For example, one may buy more stocks now if he/she expects the price will increase in the near future, or wait if he/she expects the price will drop further before it rises again. This is also known as ***speculations*** - an external factor which affects stock prices (Sornette, D., 2000), but is not considered in fundamental analysis.

Coupled with the fact that technology stocks fluctuate more and thus have more speculation than other stocks (Brunetti, C., Büyükşahin, B., & Harris, J., 2016), there is a greater need to conduct technical analysis to study the statistical patterns of the buying and selling patterns and enable the asset managers to invest at the most suitable time.

**Problem statement**

*How can we enable asset managers to make more timely and more accurate decisions when investing in technology stocks (so as to minimize risk)?*

## 1.4 Our approach

According to the Asset Management report, 2020 by PwC, big data analysis is becoming critical for asset managers to gain a competitive edge in making the right investment decisions (PwC, 2020).

Therefore, our team developed data-driven solutions to address both aspects of the problem statement (as stated in *section 1.3*):

1. To derive an accurate and timely conclusion on the change in stock prices based on the stock company’s key financial metrics (under section “**2 Solution 1**”)

**Data models used:**

Logistic Regression (LR) model

Classification and Regression Tree (CART) model

1. To predict the future trends of stock prices and trade volumes for a specific stock statistically (under section “**3 Solution 2**”)

**Data model used:**

Auto-Regressive Integrated Moving Average (ARIMA) time series model (a subset of the linear regression model)

By focusing our efforts on developing such data-driven solutions, our team utilised various data analytics tools and methods to identify and minimise investment risk on technology-related investments for White Rock’s Clients and recommend appropriate investment opportunities based on our findings.

# 2 Solution to predict from the company’s internal financial performance

## 2.1 Overview of Dataset & Choice of Variables

The objective of the first solution is to derive the change in the stock price of a specific stock based on its past financial performance. This will require a dataset that contains both the financial data of the company as well as the change in its stock price over the same period.

Therefore, our team has selected the [200+ Financial Indicators of US stocks (2014-2018)](https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018) as the dataset for our first solution, as it covers over 4000 US stocks with more than 200 financial indicators commonly found in companies’ annual reports. In addition, it also contains a “PRICE VAR[%]” column, which indicates the difference in the stock’s price between the start and the end of the year.

Correspondingly, the “Class” column is a binary (or categorical) indicator for the value of “PRICE VAR[%]”, which our team will be using as the variable for predicting outcomes in the Logistic Regression and CART models.

Among the 200+ financial indicators within the dataset, our group has decided to narrow them down to 21 different variables based on 5 main categories as seen in the table below, which shows a brief summary of the variables and the justifications for the choices. A more detailed explanation can be found in [**Appendix A.**](#_Appendix_A_Justification)Additionally, we mainly selected to use ratios instead of absolute values as it allows for better comparison between companies of varying market capitalizations.

|  |  |  |
| --- | --- | --- |
| **Category** | **Variable(s)** | **Definition/Justification** |
| **Profitability Ratio** | Profit Margin  EBITDA Margin  Enterprise Value over EBITDA  Return on Equity(ROE) | Profitability ratios are ratios which ***assess a company’s ability to generate earnings relative to revenue, operating cost.*** Ratios such as ROE are proven to be effective in predicting the future value of stock price.(Arkan,2016) |
| **Leverage Ratios** | Debt to Equity Ratio  (D/E Ratio)  Interest Coverage | Leverage ratios are ratios which ***assess the ability of a company to meet its financial obligations,*** and they are closely related to stock prices. For example, a research report states that there is a significant positive relationship between D/E ratio and stock price. (Arkan,2016) |
| **Liquidity Ratio** | Free Cash Flow Yield  Current Ratio  Operating Cash Flow to Sales Ratio | Liquidity ratio measures a company’s ability to pay debt obligations and its margin of safety. Investors consider if a company has the ability to convert assets into cash. |
| **Efficiency Ratio** | Asset Turnover Ratio | Asset Turnover ratio shows how likely a company is able to effectively generate revenue from their assets. |
| **Market Prospect Ratios** | P/E ratio  P/B ratio  Dividend Yield  Book Value per share  EPS Diluted Growth  Revenue growth | Market prospect ratios provide investors with an estimate of the value they would receive from their investment. For example, ratios such as P/E ratio do in fact have a strong relationship with price of stocks and can be used to forecast market returns. (Mozes, 2017) |
| **Others** | Market Capitalisation (Market Cap) | Market capitalisation is the total dollar value of outstanding shares of a company. It reflects the value investors are willing to pay for its stock (wan, 2018). |
| PRICE VAR[%] | Indicate the % change in the stock’s price between the start and the end of the calendar year. |
| Class | A binary metric for “PRICE VAR[%]”:   * **class = 0:** negative PRICE VAR[%] value   A ***NOT BUY***decision for a rational investor   * **class = 1:** positive PRICE VAR[%] value   A ***BUY***decision for a rational investor |
| Ticker | A unique sequence of characters used to represent a particular stock listed on an exchange. |

**Table 1: Definition/Justification of variables used**

## 2.2 Data Cleaning

Firstly, we renamed the column “PRICE VAR [%]” to “Price Variation” and included a column “Ticker” to have a clearer representation of the two columns such that it is easier to interpret.

Secondly, we converted the “Class” column to a categorical variable for our model to be able to distinguish between the 2 values.

Next, as our focus is on technology stocks and according to the variables discussed in the previous subsection and in [Appendix B](#_Appendix_B_data), we filtered out the data set accordingly. After this step, we have 3126 rows and 18 variables in our initial dataset, excluding our response variables (“Price Variation” and “Class”) and “Ticker”.

We assumed that time is not a factor in this analysis, therefore, we combined data across 5 years into a single dataset, as there are enough financial indicators to leave out the time factor. This allows us to have more data to train our model, which results in a model with better accuracy.

1. **Missing values for each row**

Within the data set, there are multiple rows with many NA values ([Appendix B.1](#_B.1_Bar_plot)). Due to insufficient data to predict the “Class” variable for these rows, we decided to remove those rows with more than 50% NA values. In this cleaning step, we removed 81 rows.

1. **Missing values for each column**

Within the data set, there are multiple columns with many NA values. ([Appendix B.2](#_B.2_Bar_plot)) Due to insufficient data to predict the “Class” variable for these rows, we decided to remove those rows with more than 20% NA values. In this cleaning step, we removed 1 column: “priceEarningsToGrowthRatio”.

1. **Illogical Data**

Within the data set, there are 3 rows with a market capitalization of 0 but they have an EPS of > 40. All the 3 rows have the same ticker which implies that there might be some error when the data was being keyed in. In order to preserve the integrity of the data, we chose not to estimate a market cap for the rows and decided to remove these 3 rows.

1. **Removal of duplicates**

We ran code to remove duplicate data entries from the entire data set but no rows were affected in this cleaning step.

1. **Removal of extreme values**

To ensure that extreme values do not skew the accuracy of our models, we prepared a boxplot for each of the 17 remaining variables ([Appendix B.3](#_B.3_Boxplot_for)) to identify and remove such points from our dataset. In total, we have identified 31 rows to be removed from the dataset.

## 2.3 Data Exploration

After cleaning the data, data visualization tools from ggplot2 package were used to derive insights from the data.

1. **Distribution of Class**

As seen in [Appendix C.1](#_C.1_Barplot_for), we have 55% of stocks with class “0” and 45% of stocks with class “1”, which means that the dataset is well-balanced and not very skewed. Furthermore, there are no null values for that column.

1. **Distribution of Class Based on Market Capitalization**

Since a stock’s price and its market capitalization are closely related, our team has categorised the stocks by market capitalization into small (Less than US$2 billion), medium (US$ 2 - 10 Billion) and large (More than US$10 Billion) (Investopedia, 2020) to observe the distribution of “Class” ([Appendix C.2 & C.3](#_C.2_Barplot_for)). Generally, we can see that companies with higher market capitalization will see an increase in stock price.

In terms of actual price variation, we see that the distribution of price variation is similarly low for medium and large stocks whereas small stocks have the highest variation ([Appendix C.4](#_C.4_Violin_plot)). This is inline with our analysis that we have gathered from Rolf W. BANZ which suggests that the size of a company might not be correlated to the returns of a particular company’s stock. (Banz, 1979)

1. **Distribution of Price Variation based on Market Capitalization**

Instead of measuring price variation in dollars, our team created a variable to categorize 5 levels of ratings ranging from strong sells to strong buys based on the percentile of price variation ([Appendix C.5](#_C.5_Barplot_for)). According to research, market analysts have identified market capitalization as an important indicator in determining how strong the market rating of the stock is (Investopedia, 2018). The results further support our findings regarding the positive relationship between market capitalization and stock price.

## 2.4 CART & Logistic Regression Model

Using the above discussed dataset, we adopted 2 methods - Classification and Regression Tree (CART) and Logistic Regression - to predict whether stock prices will increase or decrease based on the selected financial indicators.

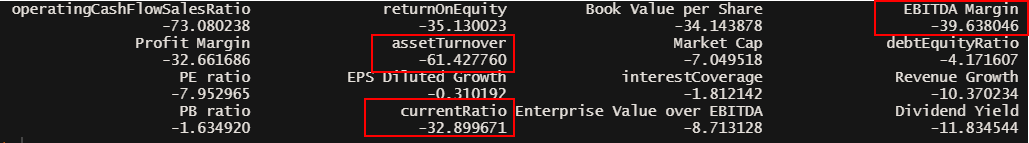
Decision trees are useful for its explicability as it makes the results simple to explain via rules established by the model, which is efficient from a business perspective. Furthermore, a decision tree will allow White Rock to correctly identify the most important financial indicators of a company to base investment decisions on. To ensure a fair comparison between models, we used set.seed(30) before running each of our models.

We ran both models separately to optimise each model, followed by analysing the variables generated by both the models and comparing which variables are the most significant in affecting stock prices.

Before employing either models, we performed a train-test split of 70% - 30% using the CaTools package. Splitting the dataset into a train set and test set will allow us to verify the accuracy of our model after developing it. This is supported by Ramesh Medar whose research suggests that the accuracy of predictions made depend on the training data used and the predictions are closer to the actual values when the model is well trained. (Medar, 2017)

### A. CART Model

Firstly, an initial CART model containing all 17 variables is developed using the entire dataset to identify the most important variables via variable importance. The variables and their respective variable importance is as shown in [Appendix D.1](#_D.1_Variable_importance). To develop our model, the input variables are selected based on the severity of drop in importance between consecutive variables. For instance, the first cut-off point is “operatingCashFlowSalesRatio” as there is a 35.1 variable importance drop from “operatingCashFlowSalesRatio” to “returnOnEquity”. In this case, we do not cut-off at “Free Cash Flow Yield” as it will only leave us with 1 variable in our CART model. The red boxes indicate all the cut-off points that will be used for the subsequent CART models.



At the first cut-off point, we will use the 4 variables: “Free Cash Flow Yield”, “operatingCashFlowSalesRatio”, “returnOnEquity” and “Book Value per Share” to develop our first model ([Appendix D.1](#_D.1_Variable_importance))

For the CART model, the tree is grown to the maximum before it is pruned based on the 1 Standard Error Rule (Hastie, Tibshirani and Friedman, 2009). The smallest tree, with error estimate within one standard error from the minimum cross validation error, is chosen. This would improve stability of the tree selection and reduce tree complexity, with a statistically insignificant decrease in model accuracy. A confusion matrix is then constructed by applying model predictions to the train set and test set.

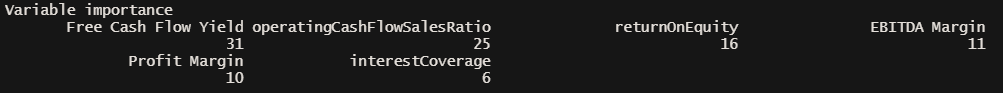
|  |  |
| --- | --- |
| Confusion Matrix for train set | Confusion Matrix for test Set |

The results of the train set are shown for reference while the results for the test set is taken to test for accuracy. The train set accuracy for the CART model is 71.1% while the test set accuracy for the CART model is 68.8%. The similarity of accuracies between the train set and test set ensures that there is no overfitting of the model.

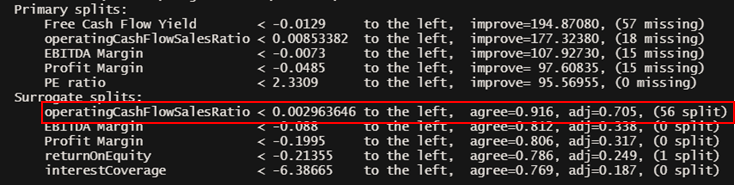
As the first model built may not necessarily be the best, additional models, with different input variables are being built based on other cut-off points in variable importance. The pruned tree and the confusion matrix for the other 2 models are shown below:

|  |  |  |
| --- | --- | --- |
| **Models** | **Input Variables** | **Test Set Accuracy (3dp)** |
| Model.cart.pruned  **(**[**Appendix D.2**](#_D.2_Model_1) **)** | `Free Cash Flow Yield`+ `operatingCashFlowSalesRatio` +`returnOnEquity` + `Book Value per Share` | 68.771% |
| Model.cart.pruned2  **(**[**Appendix D.3**](#_D.3_Model_2)**)** | `Free Cash Flow Yield`+ `operatingCashFlowSalesRatio` +`returnOnEquity` + `Book Value per Share` + `EBITDA Margin` + `Profit Margin` | 68.771% |
| Model.cart.pruned3  **(**[**Appendix D.4**](#_D.4_Model_3) **)** | `Free Cash Flow Yield`+ `operatingCashFlowSalesRatio` +`returnOnEquity` + `Book Value per Share` + `EBITDA Margin` + `Profit Margin` + `assetTurnover` + `Market Cap` + `debtEquityRatio` + `PE ratio` + `EPS Diluted Growth` + `interestCoverage` + `Revenue Growth` + `PB ratio` | **69.657%** |

From the table, we can see that the 3rd model developed has the highest test set accuracy. To establish the variables that are crucial in determining “Class” of a stock, we used variable importance.



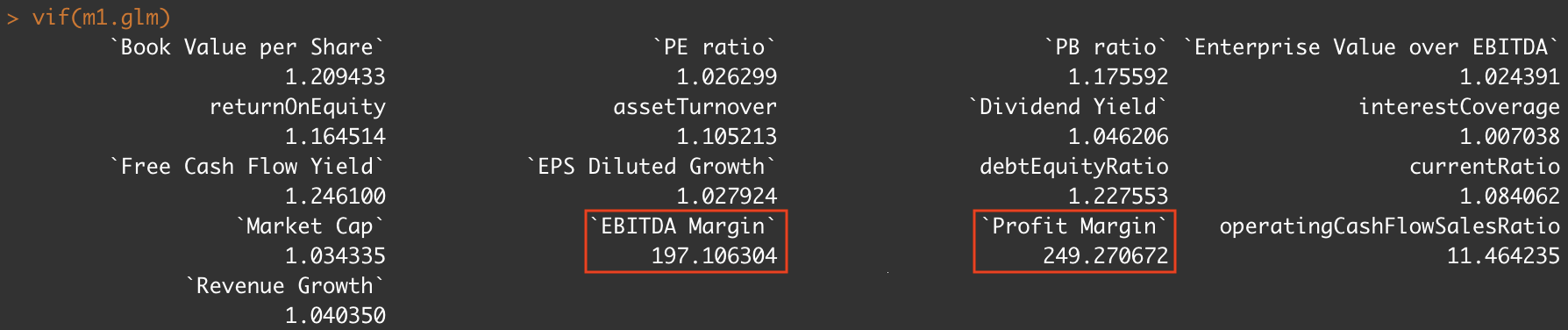
Using variable importance, “Free Cash Flow Yield”, “Operating Cash Flow Sales Ratio” and “Return on Equity” are the most important factors to decide if a technology stock is worth investing in. This is further supported that free cash flow yield has the explanatory power in predicting subsequent returns. (David, 1994) In reality, it might be difficult to get perfect information regarding a particular stock due to the lack of publicly available information. In this case, CART provides an alternative solution: To use surrogate factors which are almost similar in terms of importance to determine the split within the decision tree.



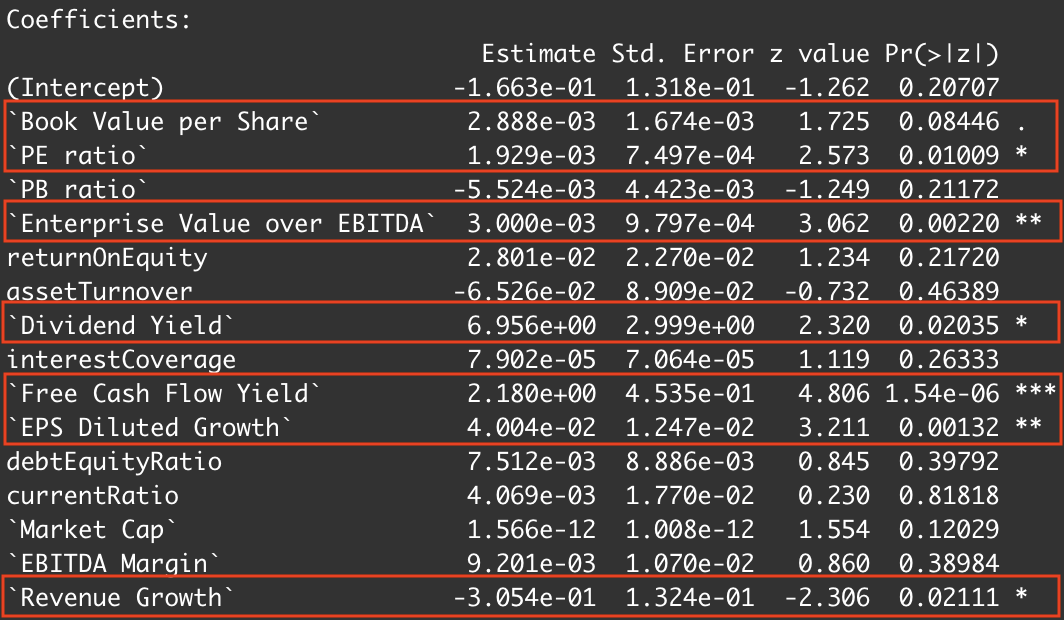
Taking the first split at node 1 as an example, we can see that “operatingCashFlowSalesRatio” can be used as a substitute as it is 91.9% similar to “Free Cash Flow Yield”. This surrogate split is also 70.5% better than doing using the majority rule for identifying a stock’s Class. This surrogate split is also used in training or predicting the model and the number of rows affected by the split is also displayed, which is 56 rows. Knowing these surrogate splits can help White Rock managers be flexible in their decision making and when evaluating technology stocks.

### B. Logistic Regression Model

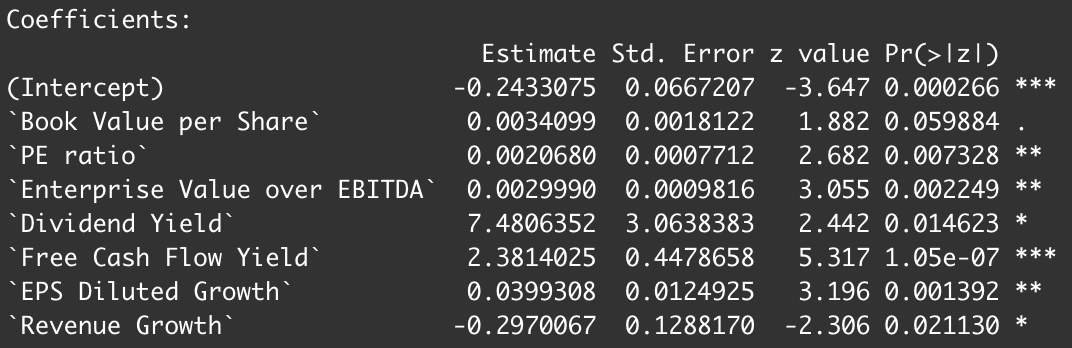
We conducted another round of cleaning of the dataset by removing 568 rows that contained any NA values. The rationale is due to the limitation of logistic regression not being able to automatically deal with NA values. We chose not to replace or estimate the NA values as the financial indicators of a company are specific to that particular company and we are unable to obtain rough estimates from similar sized companies. After conducting a train test split on the new dataset, we ran a logistic regression model using “Class” as the dependent variable and the remaining 17 variables as the independent variables. Subsequently, we checked for multicollinearity between the variables by running the “vif()” function under the “car” package. For our logistic regression model, we decided that variables with VIF values greater than 10 indicate a presence of multicollinearity.



We noticed that “EBITDA margin” and “Profit Margin” have very high Variance Inflation Factor(VIF) values. We decided to remove “Profit Margin” as it had the highest VIF value. After training another logistic regression model and checking VIF values again, we noticed that “EBITDA Margin” and “operatingCashFlowSalesRatio” have VIF values that are greater than 10 ([Appendix E.1](#_E.1_VIF_of)). Therefore, we removed “operatingCashFlowRatio”. The variables of the new model have VIF values that are close to 1 ([Appendix E.2](#_E.2_VIF_of)), indicating low presence of multicollinearity. The below image shows the summary of the new logistic regression model after removing “Profit Margin” and “operatingCashFlowSalesRatio”.



The p-values for each variable are listed under the column “Pr(>|z|)”. For this model, we decided to treat statistically significant variables as those whose p-values are less than 0.10, which are variables indicated by stars(\*) or a dot(.) next to them (highlighted above). Running another logistic regression model with only these as the independent variables, we obtained the final logistic regression model as seen below. These variables are determined to be statistically significant in determining whether “Class” is 0 or 1.



A confusion matrix was then constructed by applying model predictions each to the train and test sets. Similar to the CART model, the results of the train set are shown for reference while the results for the test set is taken to test for accuracy. The values shown are recorded as percentage values.

|  |  |
| --- | --- |
| **Confusion matrix for train set** | **Confusion matrix for test set** |

The similarity of accuracies between the train set and test set ensures that there is no overfitting of the model ([Appendix E.3 & E.4](#_E.3_Logistic_regression)) . The final test set accuracy for the logistic regression model is 63.2%.

## 2.5 Comparison & selecting the better model

By improving and modifying both models separately and repeatedly, we are able to optimise the best model to obtain its respective accuracy on the test set. This will enable us to identify the most appropriate model and factors to decide if a technology stock is worth investing. Both models have identified Free Cash Flow Yield as the most statistically significant variable in logistic regression and the most important variable in CART. In the final model, the accuracy of CART is 6.5% more accurate on the testset than that of logistic regression. In this case, this accuracy difference can be costly to White Rock. A wrong prediction in the Class variable could potentially cost White Rock a lot of money. For instance, Knight Capital Group had a computer glitch in 2012 which caused them to make trades they did not intend to, which resulted in a total loss of US$440 Million within 10 minutes (Harford, 2012). This illustrates the importance of making proper and accurate judgement calls in investing stocks as minute errors can be magnified. Hence, we decided to employ CART and use the corresponding variables defined as important in the final model.

## 2.6 Prediction results & conclusion

Using the final classification tree built, it is possible to identify technology stocks that have the potential to rise in price. After obtaining the required financial metrics regarding the stocks that they are interested in, asset managers can follow the branches of the classification tree to evaluate if the stock is worth to invest in. This will greatly reduce the time invested in analysing stocks in great detail even though they do not have the potential. The indicators of technology stocks that are likely to increase in prices are as follows :

1. Free Cash Flow Yield greater than -0.0013
2. Return on Equity greater than -0.44
3. Operating Cash Flow Sales Ratio greater than -0.02

|  |
| --- |
|  |

# 3 Solution to predict from external market patterns

## 3.1 Overview of Dataset

The second dataset was chosen consisting of the [Big Five Stocks](https://www.kaggle.com/abdullahmu/big-five-stocks), which are NASDAQ stock data for the big 5 tech companies - Apple, Google, Amazon, Facebook and Microsoft. These stocks’ data are included in the NASDAQ composite index, which is an index that includes the majority of stocks listed on NASDAQ and is the 3rd most-followed stock market indices in the United States. This dataset ranges from 1972 to 2019 and has 5 main “data” columns - high, low, open, close, volume, which are the main stock price indicators.

We chose this dataset because it is comprehensive and it contains data that is fundamental to stock analysis as discussed above in Section 1.3. The data is also heavily used in other research and contains daily values of the stocks, which allows us to have a more in-depth analysis.

In this dataset, we have NASDAQ composite stock index data from 1972, and the big 5 companies data span from the day they are publicly listed to 2019.

## 3.2 Data Cleaning & Choice of Variables

For data cleaning for the second data, we first converted the data types from strings to factors in order to deploy them in our data exploration and model more easily.

As our data set does not consist of any NULL values and that we will require all data in the dataset in order to see the patterns across the years, there is no need to clear any outliers or delete any specific rows of data.

Hence, the variables included for the second solution are Stock Ticker, Dates, High, Low, Open, Close, and Trade Volume.

## 3.3 Data Exploration

### a) Exploring stock price indicators

Firstly, we plotted the time series graph for the 4 stock price indicators (i.e. the high, low, open, close) of the 5 stocks and the NASDAQ composite index (i.e. “^IXIC” index) in order to have a general overview of how the stock price has changed over time ([**Appendix F.1**](#_F.1_Graphs_of)**)**.

From the first exploration, we found out that the 4 different stock price indicators have similar patterns over the years. Therefore, we compiled the 4 indicators into a financial barchart, and plotted the time series graph for the financial barchart[[2]](#footnote-2) ([**Appendix F.2**](#_F.2_Financial_barchart)). The results from this graph have proven that the 4 stock price indicators indeed are highly similar in their patterns, as they are basically the variations of daily stock prices which do not fluctuate significantly.

Therefore, in order to analyse the stock price patterns across the year, there is no strong need to analyse the 4 stock price indicators separately. We then created an additional column called “average” which is the mean stock price of the 4 indicators (high, low, open, close) and used it in our model subsequently. A time-series graph of “average” stock price for all 5 stocks and the composite index can be found in [**Appendix F.3**](#_F.3_Graph_of)**.**

### b) Exploring trade volumes

To explore the “volume” column[[3]](#footnote-3), similarly, we plotted a time series graph to see how the trade volume changed over time ([**Appendix F.4**](#_F.4_Graphs_of)).

Next, as the objective of this solution is to predict the change in stock price, we wanted to see if there is any correlation between the stock prices and the trade volumes. Therefore, we plotted Linear Regression graphs of stock price against volume. ([**Appendix F.5**](#_F.5_Linear_Regression)). However, the results of this data exploration show that there is no clear correlation between stock price and trade volumes. Hence, we will not include trade volumes in our model.

## 3.4 Time Series Model

Our team decided to use the ARIMA model. ARIMA is a class of models, similar to a linear regression model, that tries to create a model of time series based on its own past values which are then used to forecast and predict future values. ARIMA model is used mainly due to it being robust and efficient in short-term prediction of stock prices. (Adebiyi, 2014) Our team decided to perform a time series analysis, in order to predict future values of stock prices.

### Step 1: Perform rolling mean on the dataset

Due to the actual raw data being too unpredictable and containing too much “white noise”, meaning that the stock fluctuated too often in a small time period, we have performed a rolling mean of the dataset. Relying solely on the raw data might disrupt the model and ARIMA would not be able to predict future stock prices. By incorporating rolling means, it smooths out the data so that any sudden spike in stock prices does not disrupt the model that much.

We then created an additional column to store the moving average, and used the zoo package to create a rolling-mean of 121 days, which approximates to 4 months. 4 months provides a good rough guide of the trend’s stock in the near future, and smooths out day to day fluctuations of the stock. We then plot comparison of the actual data and the rolling-mean data for all 5 stocks. ([**Appendix G.1**](#_G.1_Graphs_of))

### Step 2: Training the ARIMA model and checking its accuracy

To use the ARIMA model, we first converted the dataset into a time series, followed by the splitting of the data into train and test sets. We allocated the last 2 years of data as the test set and the rest as the train set. The rationale is to be able to predict short term data given the stock price of the previous years. We then used ARIMA via the “forecast” R package to predict the future stock price value, and then plot it against the actual data. Refer to the graphs in [**Appendix G.2**](#_G.2_Results_of).

## 3.5 Prediction results & evaluations

From the graphs ([**Appendix G.2**](#_G.2_Results_of)), the results are promising. The predicted data models closely match the actual data, and the actual data are within the error range of the model. This implies that ARIMA has very good confidence that the actual value falls in the error range, and is unlikely to be due to random chance since all stocks modelled to have fairly good accuracy. As such, we conclude that this model is fairly accurate at predicting stock prices given enough data.

# 4 Insights & Analysis of both solutions

## 4.1 Insights from Logistic Regression & CART models

Our findings show that the top 3 factors that White Rock asset managers should take into account when deciding how to invest are:

1. Free Cash Flow Yield
2. Return on Equity
3. Operating Cash Flow Sales Ratio

### Key factor Influencing Investment Decision:

#### Factor 1: Free Cash Flow Yield

The median Free Cash Flow Yield for stocks that are classified as “1” is 0.0425 while that for “0” is -0.0042. ([**Appendix H.1**](#_H.1_Distribution_of)). More than 75% of technology stocks that are classified as “1” have a free cash flow yield above 0 while those that are classified as 0 have a wider distribution ranging from - 0.5 to 0.5 with no obvious patterns. From this plot, we can conclude that technology stocks that have negative free cash flow yield are not likely to see an increase in stock prices. This could be explained by the fact that negative free cash flow yields imply that the business is unable to generate sufficient cash to sustain its own business. As a result, it is also likely that the business is unable to pay off its debts and other obligations or not able to offer high dividends. Moreover, a low free cash flow yield implies that investors are not receiving good returns on their investments which might deter investors from investing in the company or even choose to sell off those stocks they hold which might result in a decrease in the share price for the company.

#### Factor 2: Operating Cash Flow Sales Ratio

The median Operating Cash Flow Sales Ratio for stocks that are classified as “1” is 0.139 while that for “0” is 0.0338. ([**Appendix H.2**](#_H.2_Distribution_of)). 75% of technology stocks that are classified as “1” have an Operating Cash Flow Sales Ratio free yield above 0.0678, which is twice as high as the median of stocks classified as “0”. While there are a handful of outliers within both classes, there is a clear indicator that the Operating Cash Flow Sales Ratio of a stock is crucial in deciding its worthiness to invest in. Our findings are consistent as operating cash flow sales ratio reflects a company’s ability to generate cash from its sales. A negative or low operating cash flow sales ratio indicates that a company is not generating sufficient revenue from its core revenue and implies that they require to generate additional forms of cash flow from financing or investment activities.

#### Factor 3: Return on Equity

With reference to [**Appendix H.3**](#_H.3_Distribution_of), stocks that are of Class 1 are very concentrated as seen by the narrow box-plot while the stocks of Class 0 are more spread out and vary across a large range of values. Interestingly, 50% of Class 1 stocks are concentrated slightly above the 0 marks. Return on equity reflects the profitability of a company in relation to the shareholder’s equity. It is also a measure of how much an investor can potentially get in returns from his investment. These results suggest that ROE does indeed have a correlation in the increase of stock price and therefore be used to predict the future trend of the stock.

## 4.2 Analysis of the ARIMA Time Series Model

As stated in Section 3.5, our ARIMA model is proven to have relatively good accuracy in its predictions. This section will discuss its strength and limitations:

### a) Strengths

Firstly, as we perform a rolling mean of the dataset which reduces the overall fluctuation, we are able to give a short term prediction of the general trend of the stocks. Since companies are likely to keep the stock in the near future and not sell the stock on a weekly or even a monthly basis, this model will certainly be beneficial to White Rock in understanding which stocks to invest in.

Furthermore, the ARIMA model is designed in such a way that as more rows are added to the dataset, the model will adjust itself and do a modified prediction accordingly. This will allow White Rock to not only apply the model purely for technology stocks, but also to any other stocks that have frequent access to the market.

### b) Limitations

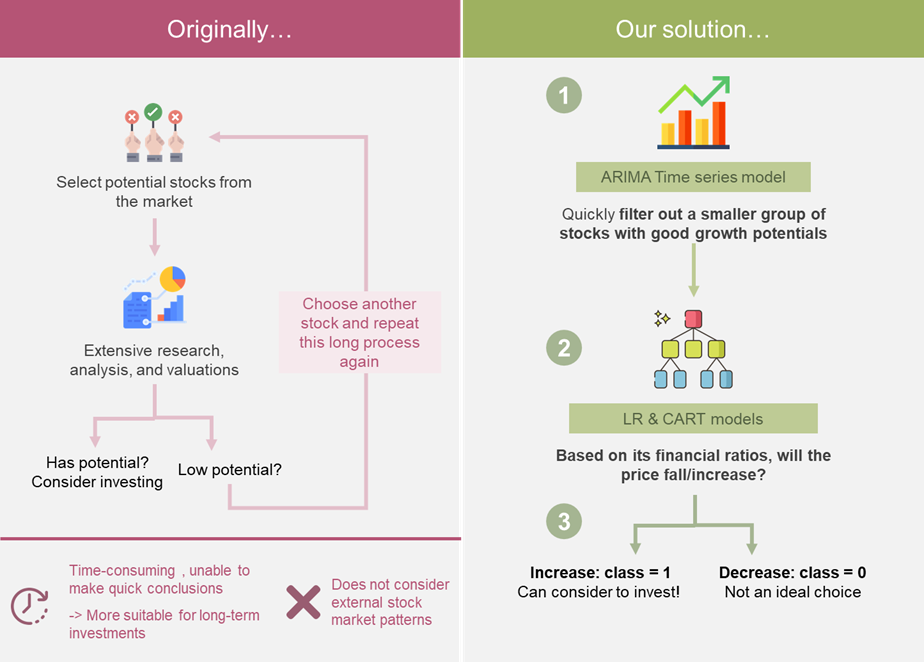
While the ARIMA model has proven to be fairly accurate in predicting the future trend of a certain stock, we do recognise that there are limitations.

Unexpected events may cause a sudden drop or rise in stock price which may distort the past stock price patterns.For instance, Facebook’s stock price plummeted tremendously between 2018-2019 due to an unexpected scandal. Such unexpected events will throw even the most advanced models off-guard (Forbes, 2020). As our model is trained to conduct predictions based on historical patterns, it is therefore limited in handling such unforeseeable circumstances.

# 5 Business Applications

Based on our findings and analysis, our team suggests a guide for White Rock to implement our solutions to improve their current investment decision-making process, so as to make more accurate and timely decisions when investing in technology stocks.

The below diagram illustrates how the new process will work:

****

### Step 1: Using ARIMA model to quickly filter out a group of good potential stocks

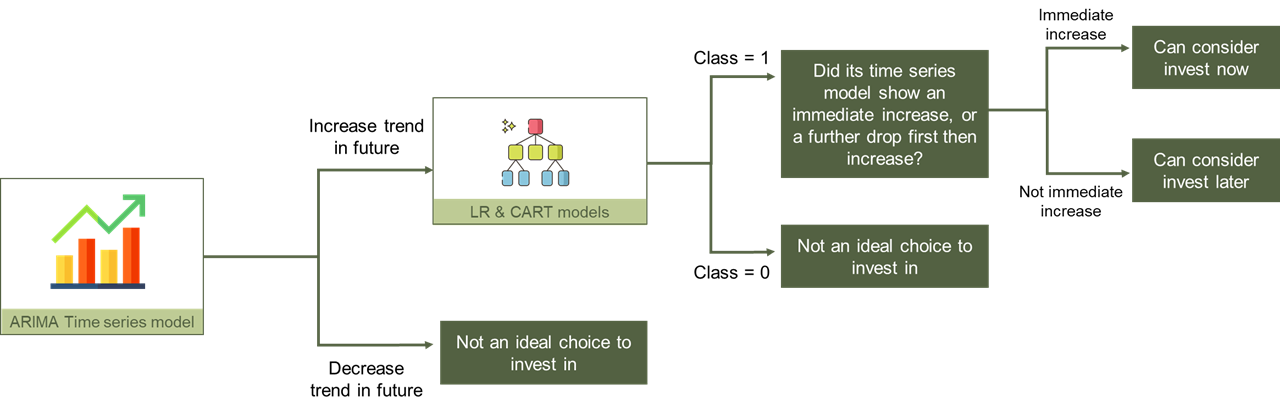
Prior to investing in a stock, White Rock could first analyse using the ARIMA model provided to forecast the future price trend of a particular stock they are interested in investing. This is because stock prices are readily accessible and can be easily obtained.

### Step 2: Using LR and CART to predict the stock price based on the stock’s own financial performance

Upon forecasting a positive trend using ARIMA, White Rock will have to further analyse the stock using our team’s CART model. This is because stock prices are known to be highly volatile and purely analysing based on historical trends would be insufficient in ensuring that risk is minimized. Therefore, as a safety net, White Rock would have to calculate the respective equity’s free cash flow yield, return on equity, operating cash flow sales ratio in order for the CART to derive the results.

### Step 3: Predictions

The CART model will predict a result on the “class” value, where 1 represents an increase in the stock price in the next year, and 0 represents otherwise. The graphic below summaries the decision making process that managers will need to take when evaluating the potential growth of a stock.



# 6 Future Considerations

While our solution has enabled asset management firms such as White Rock to improve their investment decision-making process to be more accurate and timely, we do recognize that there are areas that can be improved on. This section will provide a few potential areas that can be explored as future considerations to further improve our solutions.

The stock price of technology stocks also tends to be heavily affected by external social, economic, and political factors. Furthermore, according to research by Gbenga Ibukunle from the University of Edinburgh regarding the prediction of highly volatile markets, the findings show that a stock’s trend is dependent on the price trends of other stocks. (Gbenga, 2020).

To address these possible issues, the team has devised a few approaches to mitigate some of the above-mentioned factors:

1. **Text Mining**

Changes in stock prices are highly reliant on news articles and discussion forums where investors may reveal some of their sentiments regarding the current market situation or certain stock options. Thus, text mining technologies can be employed to do web scraping of websites with large volumes of news or opinions to derive an overall sentiment of how a specific sector is faring. Common approaches to derive sentiments in the market are to use the Lexicon-based approach or Machine Learning approach.

1. ***Analysis of other sectors that can possibly influence the identified sector (technology)***

Besides the technology sector, other sectors or markets could potentially influence the stock price of technology stocks. Supply and demand changes in other sectors can affect stock prices as well. Since the correlation between other sectors and technology may not be prominent, more advanced machine learning techniques will be required to identify and analyse such trends as the relationship may not be obvious to the naked eye.

1. ***Algorithmic Trading***

Once sufficient knowledge regarding the market has been obtained, White Rock could potentially explore the option of algorithmic trading by employing various methods. The advantages of algorithmic trading are that a computer is able to digest and analyse large amounts of information simultaneously and make decisions in a split second. This could potentially be a new source of revenue for White Rock if they are able to devise an effective algorithm which is able to make rule-based trades, using price, the volume of stocks. However, this will require a thorough and in-depth understanding of the market pricing and many factors to be included in the algorithm.

# 7 Conclusion

In conclusion, our team provides White Rock with a concrete way of how to effectively and efficiently predict technology stock prices for the foreseeable future. By combining solutions 1 and 2, we are able to integrate the benefits of CART with the ARIMA model to come up with an elegant way of deciding which technology stock prices will increase or decrease. At the same time, White Rock can overcome the inaccuracies and inefficiencies in predictions due to high volatility of technology stocks and account for external stock market patterns. With this model, we are able to help White Rock minimize risk and make better technology investments.

# References

16th International Conference on Computer Modelling and Simulation (pp. 106-112). IEEE.

Adebiyi, Ayodele & Adewumi, Aderemi & Ayo, Charles. (2014). Stock price prediction using the ARIMA model. Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014. 10.1109/UKSim.2014.67.

Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).

Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS

Arkan, T. (2016). The importance of financial ratios in predicting stock price trends: A case study in emerging markets. Finanse, Rynki Finansowe, Ubezpieczenia, (79), 13-26.

Banz, R. W. (1981). The relationship between return and market value of common stocks. Journal of financial economics, 9(1), 3-18.

Benjamin, B. (2017). Asset & Wealth Management Insights–Asset Management 2020: Taking Stock. PWC Lond. UK.

Brunetti, C., Büyükşahin, B., & Harris, J. (2016). Speculators, Prices, and Market Volatility. Journal of Financial and Quantitative Analysis, 51(5), 1545-1574. doi:10.1017/S0022109016000569

Chew C.H. (2020). Artificial Intelligence, Analytics and Data Science, Volume 1: Core Concepts and Models. Cengage Asia.

Corporate Finance Institute. (2020, January 28). Operating cash flow ratio - Formula, guide for financial analysts. <https://corporatefinanceinstitute.com/resources/knowledge/finance/operating-cash-flow-ratio/>

Davis, J. (1994). The Cross-Section of Realized Stock Returns: The Pre-COMPUSTAT Evidence. The Journal of Finance, 49(5), 1579-1593. doi:10.2307/2329263

Dimensional. (2016, July). Can Volatility Predict Returns? <https://www.sparrowwealth.com/wp-content/uploads/2016/07/Can-Volatility-Predict-Returns.pdf>

Harford, T. (2012, August 10). High-frequency trading and the $440m mistake. Retrieved from https://www.bbc.com/news/magazine-19214294

Hastie, T, R. Tibshirani and J. Friedman (2009) The Elements of Statistical Learning 2nd ed. Springer.

Investopedia. (2018, August). “How Can I Use Market Capitalization to Evaluate a Stock?” https://www.investopedia.com/ask/answers/042415/how-can-i-use-market-capitalization-evaluate-stock.asp.

Klebnikov, S. (2020). Big Tech Stocks Are Soaring—Here’s Why That Could Spell Trouble For The Market. Forbes. <https://www.forbes.com/sites/sergeiklebnikov/2020/07/11/big-tech-stocks-are-soaring-heres-why-that-could-spell-trouble-for-the-market/?sh=7ccb64e1576b>

KRAMER.J.MICHAEL (2018). Facebook Stock Seen Dropping to New Low.Investopedia. <https://www.investopedia.com/news/facebook-stock-seen-dropping-new-low/>

Medar, R., Rajpurohit, V. S., & Rashmi, B. (2017, August). Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning. In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA) (pp. 1-6). IEEE

Moews, B., & Ibikunle, G. (2020). Predictive intraday correlations in stable and volatile market environments: Evidence from deep learning. Physica A: Statistical Mechanics and Its Applications, 547, 124392. doi:10.1016/j.physa.2020.124392

Moews, B., & Ibikunle, G. (2020). Predictive intraday correlations in stable and volatile market environments: Evidence from deep learning. Physica A: Statistical Mechanics and its Applications, 124392.

Mozes, H. A., & Rozen, H. (2017). When do PE ratios matter? Journal of Investing, 26(3), 10-20. doi:<http://dx.doi.org/10.3905/joi.2017.26.3.010>

O’Neal, D., Buchbinder, C., & Casey, M. (2018). Putting technology stocks’ volatility in perspective | Capital group Canada | Insights. Capital Group. <https://www.capitalgroup.com/advisor/ca/en/insights/content/articles/putting-technology-stocks-volatility-in-perspective.html>

Ortis, A., Farinella, G. M., & Battiato, S. (2019). An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges. In ICETE (1) (pp. 296-306).

Press, T. (2020, May 14). Brookfield Asset Management reports US$157-million first-quarter loss - BNN Bloomberg. Retrieved October 31, 2020, from https://www.bnnbloomberg.ca/brookfield-asset-management-reports-us-157-million-first-quarter-loss-1.1436303

PwC. (2014). Asset management 2020: A brave new world.

Sornette, D. (2000). Stock market speculation: Spontaneous symmetry breaking of economic valuation. Physica A: Statistical Mechanics and its Applications, 284(1-4), 355-375

Wan, Guna & Hendrawaty, Ernie. (2018). Stock Liquidity and Market Capitalization: Investment Challenges in Indonesian Capital Markets. International Journal of Economics and Management Studies. 5. 11-20. 10.14445/23939125/IJEMS-V5I11P103.

# Appendix

## Appendix A Justification for Variables

|  |  |  |
| --- | --- | --- |
| **Category** | **Variable(s)** | **Meaning/Justification** |
| **Profitability Ratio** | Profit Margin  EBITDA Margin  Enterprise Value over EBITDA  Return on Equity(ROE) | Profitability ratios are ratios which assess a company’s ability to generate earnings relative to revenue, operating cost.  In general, investors would determine the profitability ratio and compare them within the industry prior to investments.  Ratios such as ROE has proven to be a most effective variable in predicting the future value of stock price in the investment sectors.(Arkan,2016) |
| **Leverage Ratios** | Debt to Equity Ratio  Interest Coverage | Leverage ratios are ratios which assess the ability of a company to meet its financial obligations. Investors would consider whether companies have enough cash flow to pay for interest and equity.  A research conducted showed a significant positive relationship between debt to equity ratio with stock price where the significance of 0.074 is above the confidence degree of 0.05. (Arkan, 2016) |
| **Liquidity Ratio** | Free Cash Flow Yield  Current Ratio  Operating Cash Flow to Sales Ratio | Liquidity ratio measures a company’s ability to pay debt obligations and its margin of safety. Investors would consider if a company has the ability to convert assets into cash. |
| **Efficiency Ratio** | Asset Turnover Ratio | Asset Turnover ratio is the ratio of sale or revenue to total assets which shows how likely a company is able to effectively generate revenue from their assets. Investors are more inclined to invest in companies with better Asset Turnover. |
| **Market Prospect Ratios** | PE ratio  PB ratio  Dividend Yield  Book Value per share  EPS Diluted Growth  Revenue growth | Market prospect ratios provide investors with an estimate of the value they would receive from their investment. Ratios such as PE ratio do in fact have a strong relationship with price of stocks and can be used to forecast market returns. (Mozes, 2017) |
| **Others** | Market Capitalisation (Market Cap) | Market capitalisation is the total dollar value of outstanding shares of a company. The greater the market capitalization, the more investors are interested in buying or to retain shares that have been owned because of the large estimated profits to be gained (wan, 2018). |
| PRICE VAR[%] | Indicate the % change in the stock’s price between the start and the end of the calendar year. |
| Class | A binary metric for “PRICE VAR[%]”:   * **class = 0:** negative PRICE VAR[%] value   *A hypothetical trader* ***should NOT BUY*** *at the start of the year and sell at the end of the year for a profit*   * **class = 1:** positive PRICE VAR[%] value   *A hypothetical trader* ***should BUY*** *at the start of the year and sell at the end of the year for a profit* |

[*Click here to go back to the main report content*](#_2_Solution_to)

## Appendix B data cleaning for solution 1

### B.1 Bar plot of number of ‘NA’ values in each row

|  |
| --- |
|  |

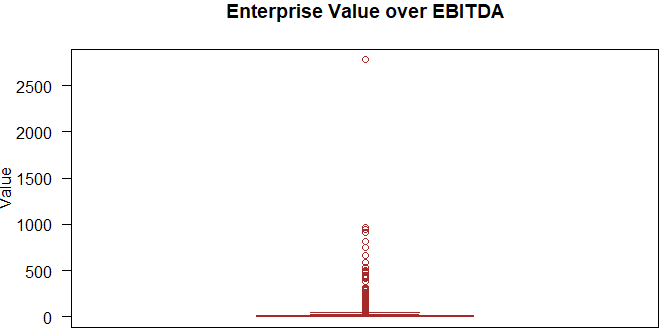
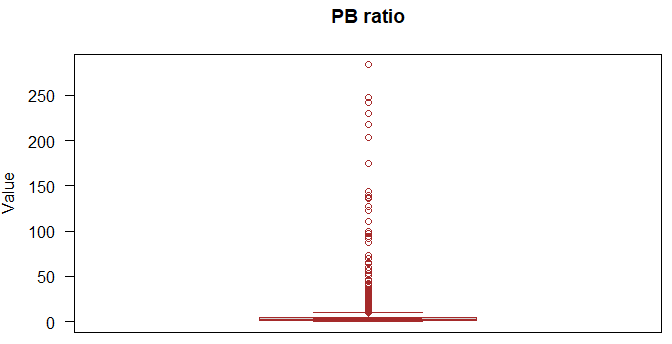
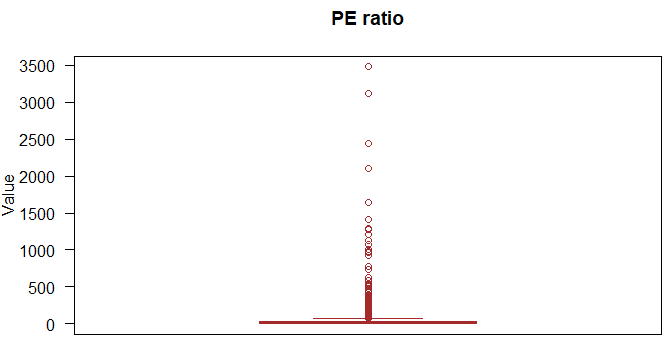
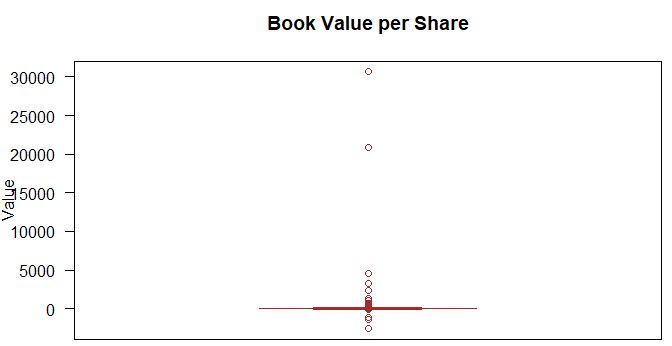
[*Click here to go back to the main report content*](#_2.2_Data_Cleaning)

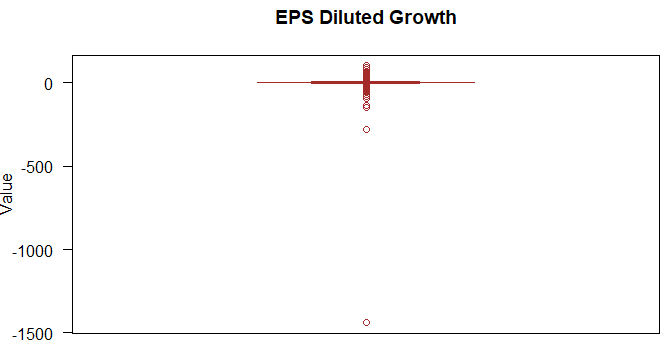
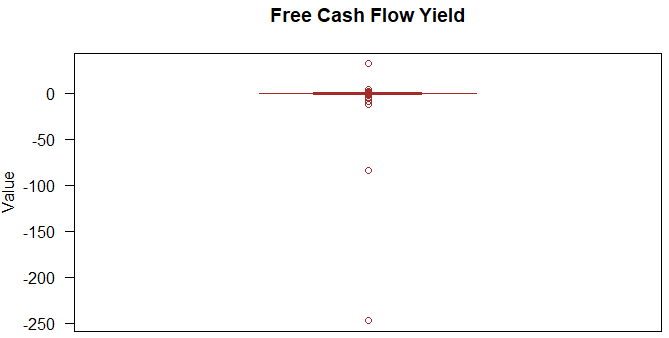
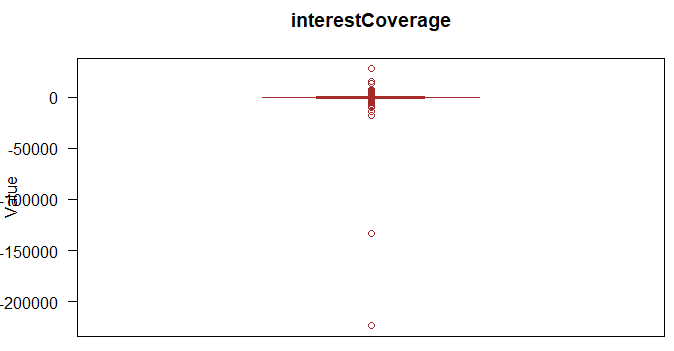
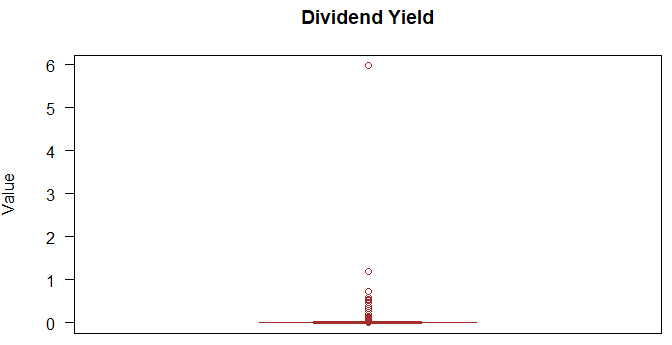
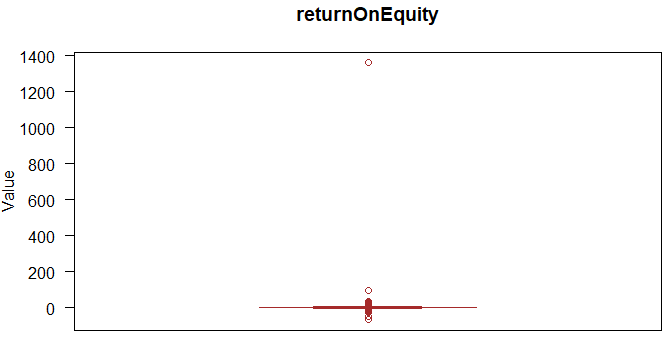
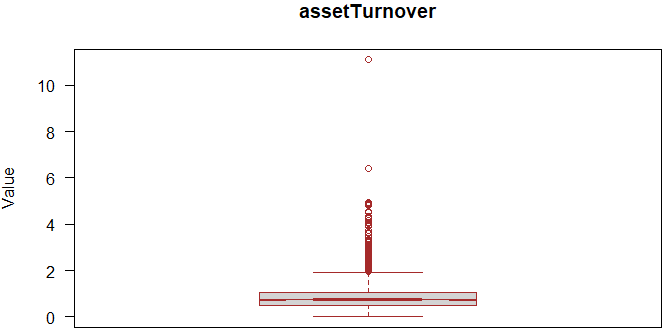
### B.2 Bar plot of number of ‘NA’ values in each column

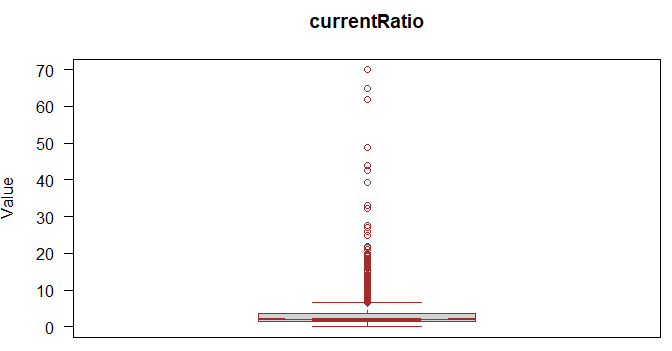
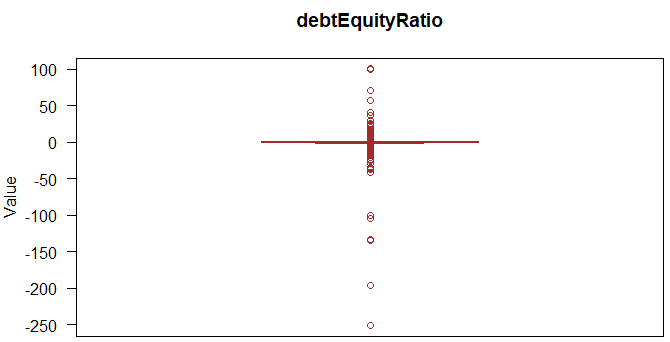
|  |
| --- |
|  |

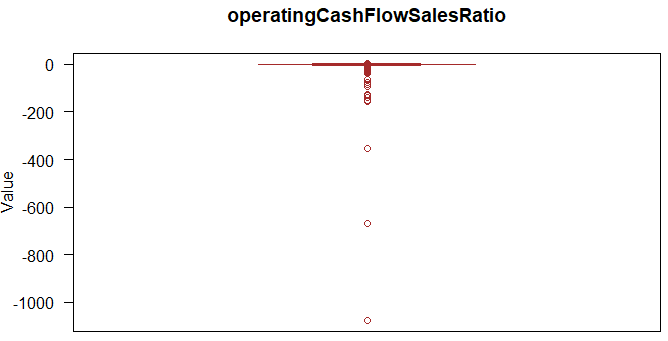
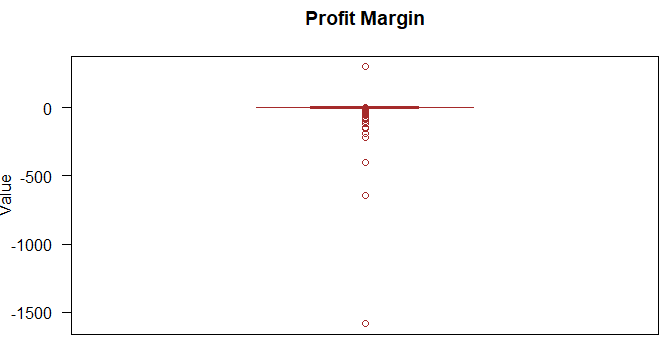
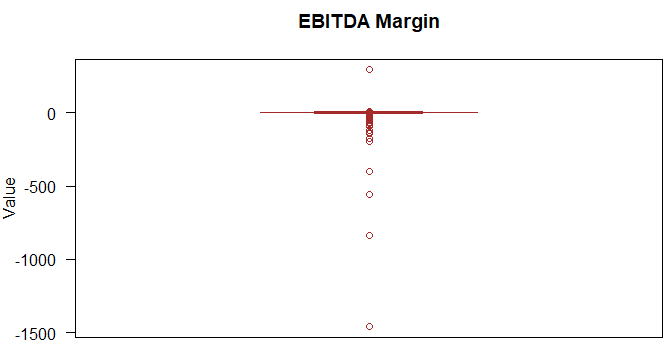
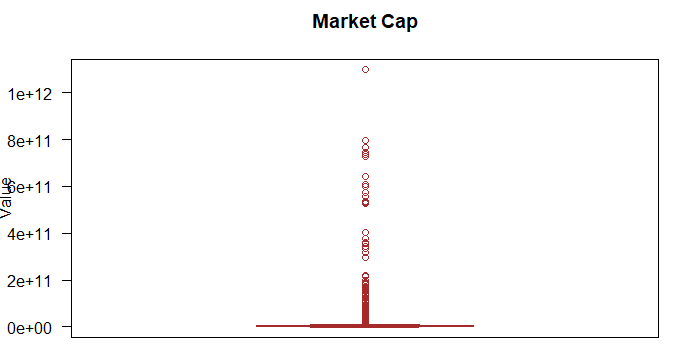
[*Click here to go back to the main report content*](#_2.2_Data_Cleaning)

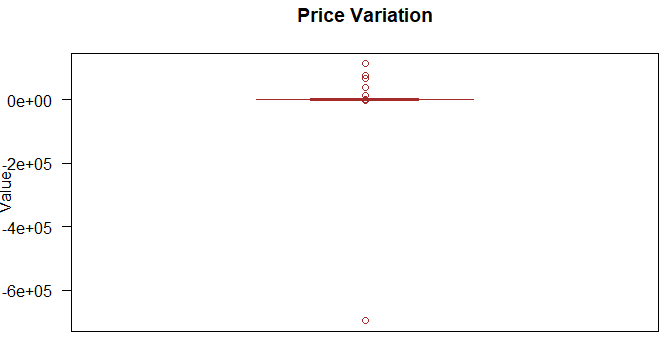
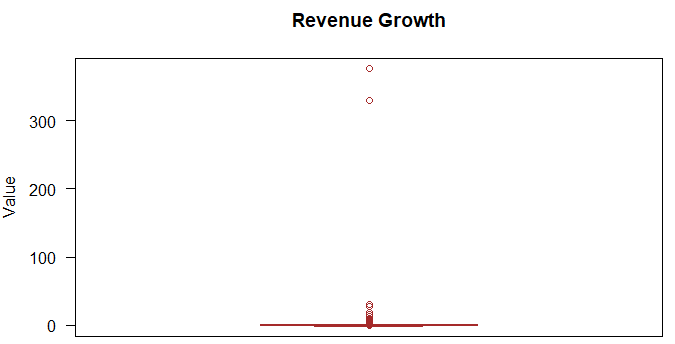
### B.3 Boxplot for all variables to identify extreme data points

****

****

****

****

****

[*Click here to go back to the main report content*](#_2.2_Data_Cleaning)

## Appendix C Solution 1 - Data Exploration

### C.1 Barplot for Distribution of Class

|  |
| --- |
|  |

[*Click here to go back to the main report content*](#_2.3_Data_Exploration)

### C.2 Barplot for Distribution of Market Capitalization

|  |
| --- |
|  |

[*Click here to go back to the main report content*](#_2.3_Data_Exploration)

### C.3 Barplot for Distribution of Class based on Market Capitalization

|  |
| --- |
|  |

*[Click here to go back to the main report content](#_2.3_Data_Exploration)*

### C.4 Violin plot for Distribution of Price Variation based on Market Capitalization

|  |
| --- |
|  |

[*Click here to go back to the main report content*](#_2.3_Data_Exploration)

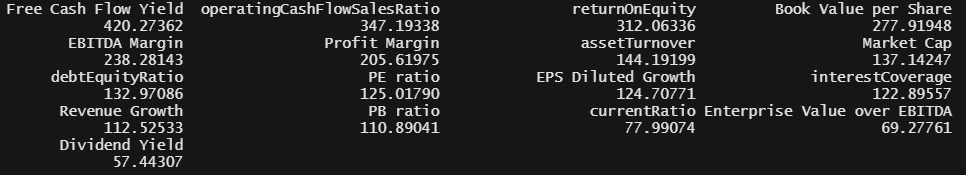
### C.5 Barplot for Analyst Rating based on Market Capitalization

|  |
| --- |
|  |

[*Click here to go back to the main report content*](#_2.3_Data_Exploration)

## Appendix D CART Models

### D.1 Variable importance of all 17 variables from initial CART



*[Click here to go back to the main report](#_2.4_CART_&)*

### D.2 Model 1

**R code to develop model 1**

****

*[Click here to go back to the main report](#_2.4_CART_&)*

**Pruned tree with optimal cp for model 1**

|  |
| --- |
|  |

*[Click here to go back to the main report](#_2.4_CART_&)*

|  |
| --- |
| **Confusion Matrix on test set for model 1** |

*[Click here to go back to the main report](#_2.4_CART_&)*

|  |
| --- |
| **Confusion Matrix on train set for model 1** |
| **Test Set Accuracy**    **Train Set Accuracy** |

*[Click here to go back to the main report](#_2.4_CART_&)*

### D.3 Model 2

**R code to develop model 2**

****

*[Click here to go back to the main report](#_2.4_CART_&)*

**Pruned tree with optimal cp for model 2**

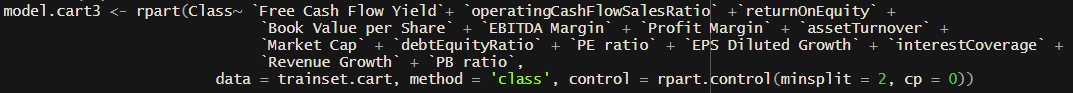
|  |
| --- |
|  |
| **Confusion Matrix on test set for model 2** |

|  |
| --- |
| **Confusion Matrix on train set for model 2** |
| **Test Set Accuracy**    **Train Set Accuracy** |

*[Click here to go back to the main report](#_2.4_CART_&)*

### D.4 Model 3

**R code to develop model 3**

****

**Pruned tree with optimal cp for model 3**

|  |
| --- |
|  |

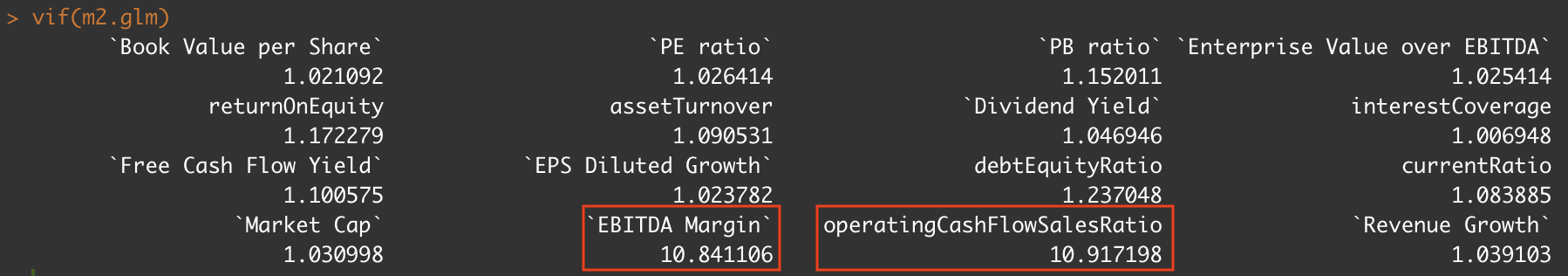
|  |
| --- |
| **Confusion Matrix on test set for model 3** |

|  |
| --- |
| **Confusion Matrix on train set for model 3** |
| **Test Set Accuracy**    **Train Set Accuracy** |

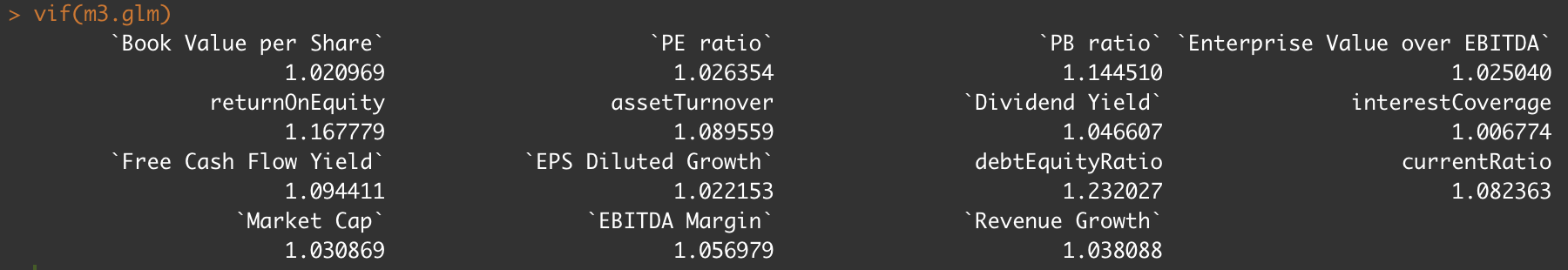
*[Click here to go back to the main report](#_2.4_CART_&)*

## Appendix E Logistic Regression Model

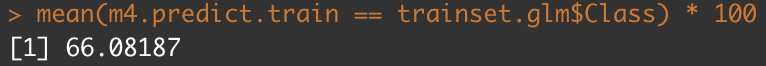
### E.1 VIF of logistic regression model 2

****

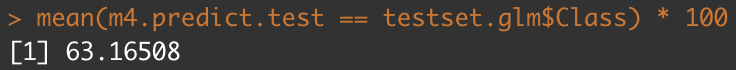
### E.2 VIF of logistic regression model 3

****

### E.3 Logistic regression model train set accuracy



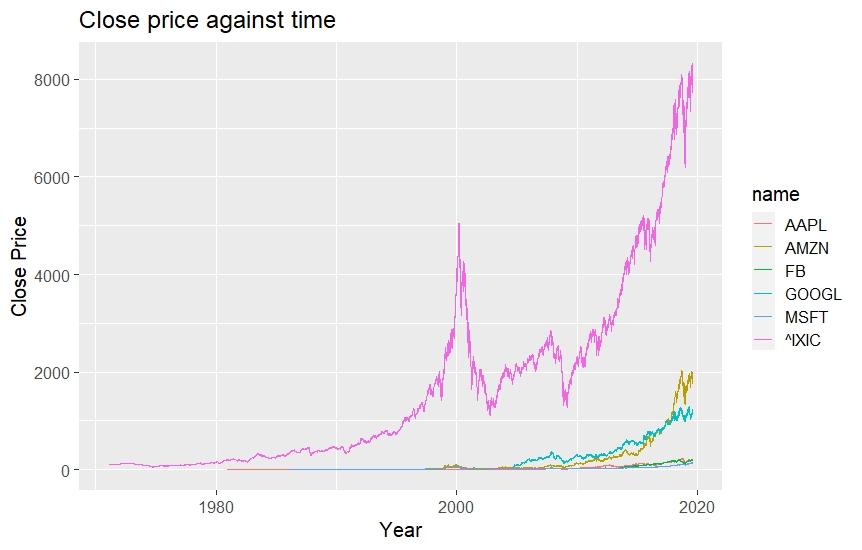
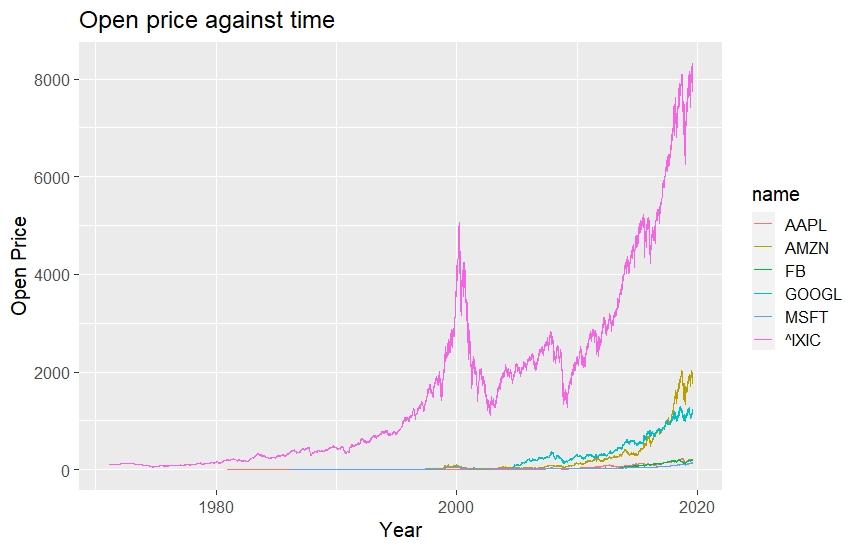
### E.4 Logistic regression model train set accuracy

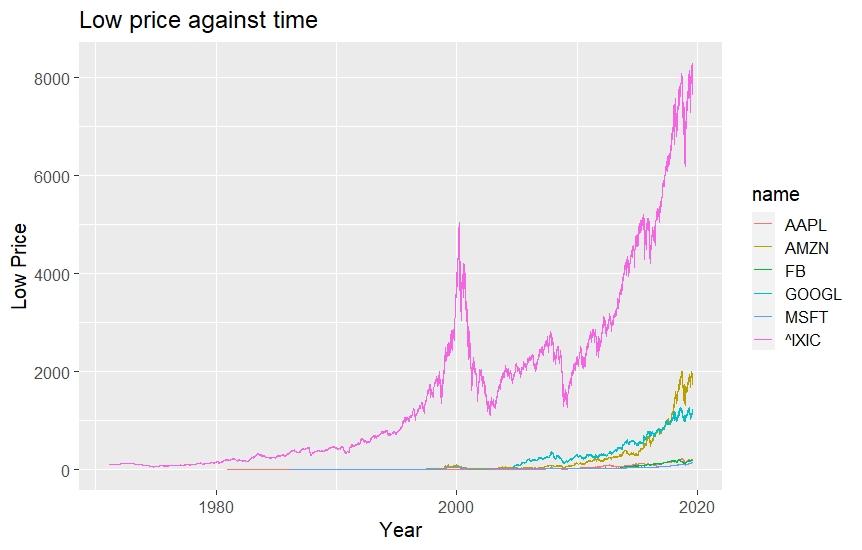
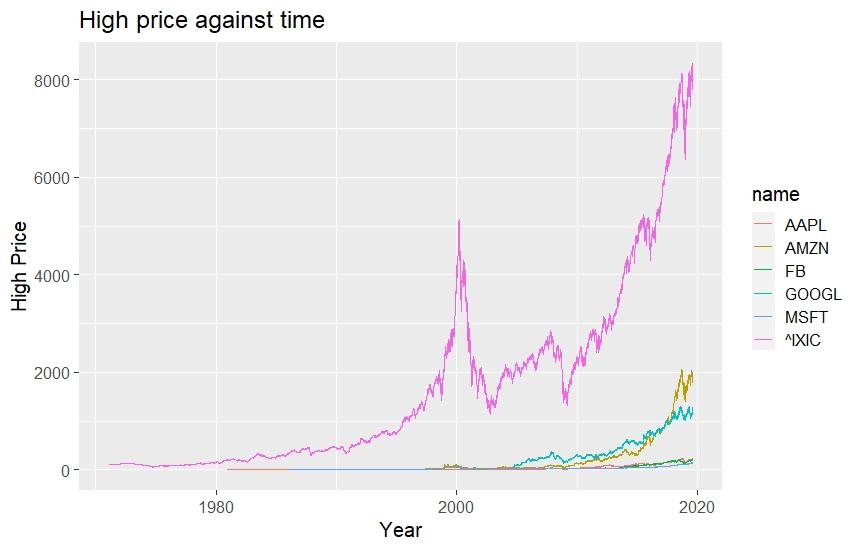


[*Click here to go back to the main report*](#_heading=h.7mdxd9ci059n)

## Appendix F Solution 2 - Data Exploration

### F.1 Graphs of individual metrics against time





[*Click here to go back to the main report*](#_heading=h.c2ei7fokffn4)

### F.2 Financial barchart (which incorporates high, low, open, close) against time

|  |
| --- |
|  |

[Click here to go back to the main report](#_heading=h.c2ei7fokffn4)

### F.3 Graph of average stock price[[4]](#footnote-4) against time

|  |
| --- |
|  |

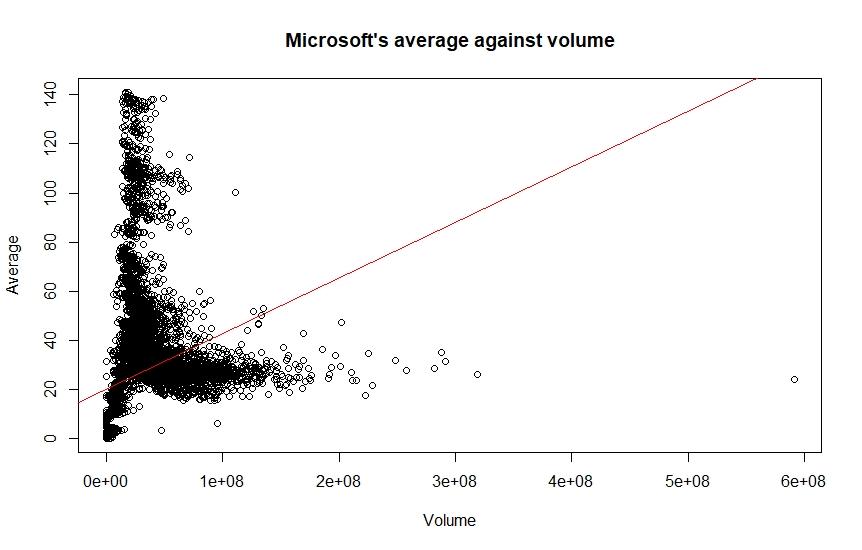
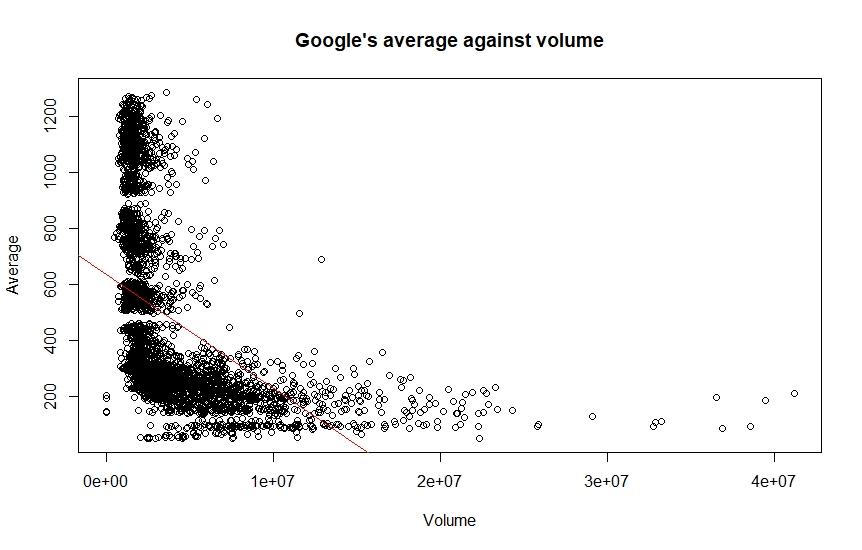
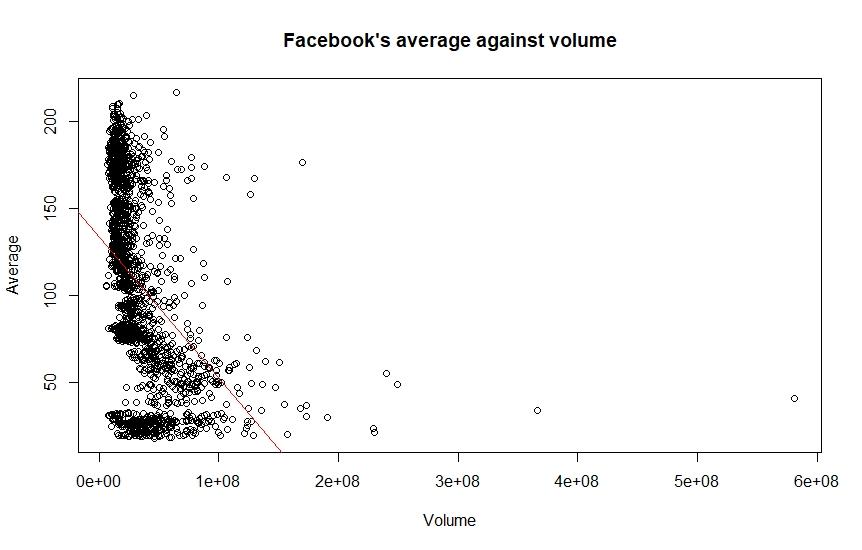
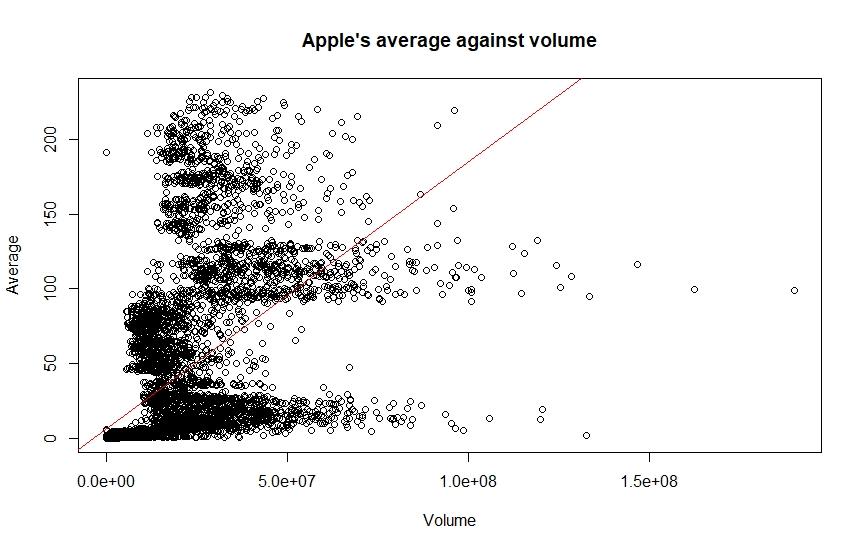
[Click here to go back to the main report](#_heading=h.c2ei7fokffn4)

### F.4 Graphs of trade volume against time

|  |
| --- |
|  |

[Click here to go back to the main report](#_heading=h.c2ei7fokffn4)

### F.5 Linear Regression graphs of average stock price against trade volume

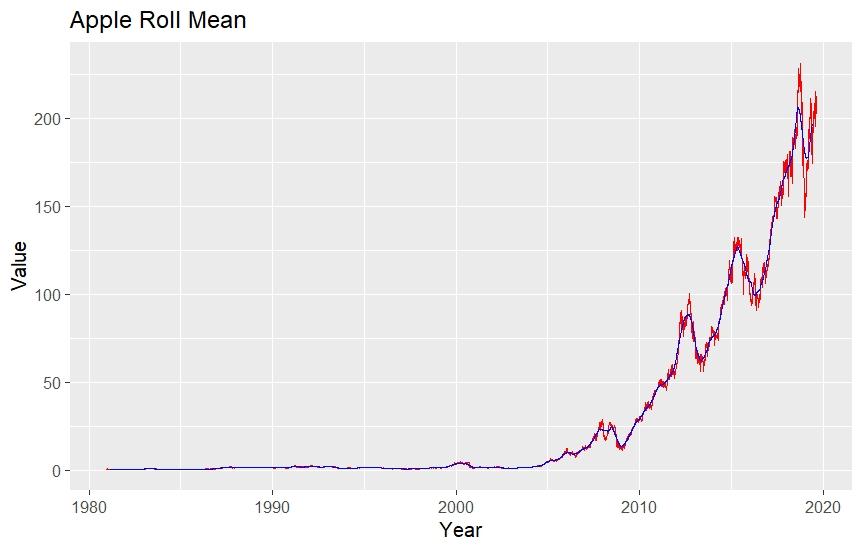
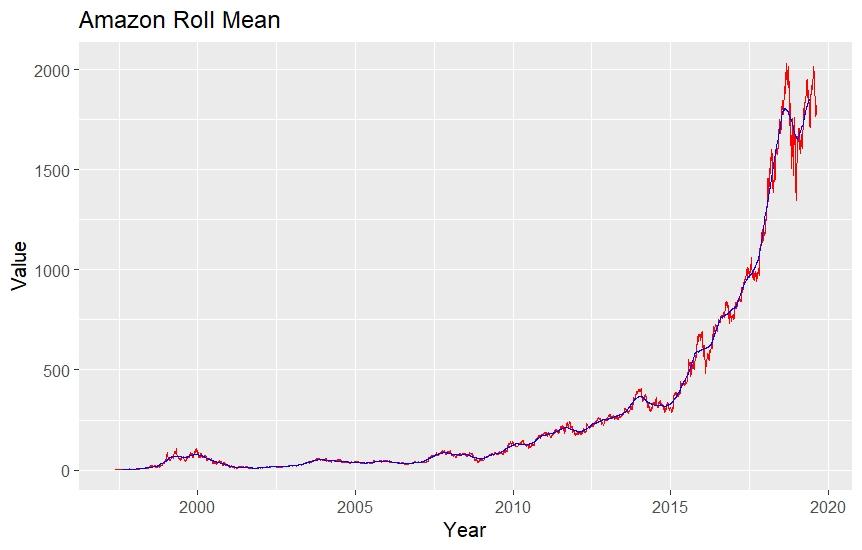


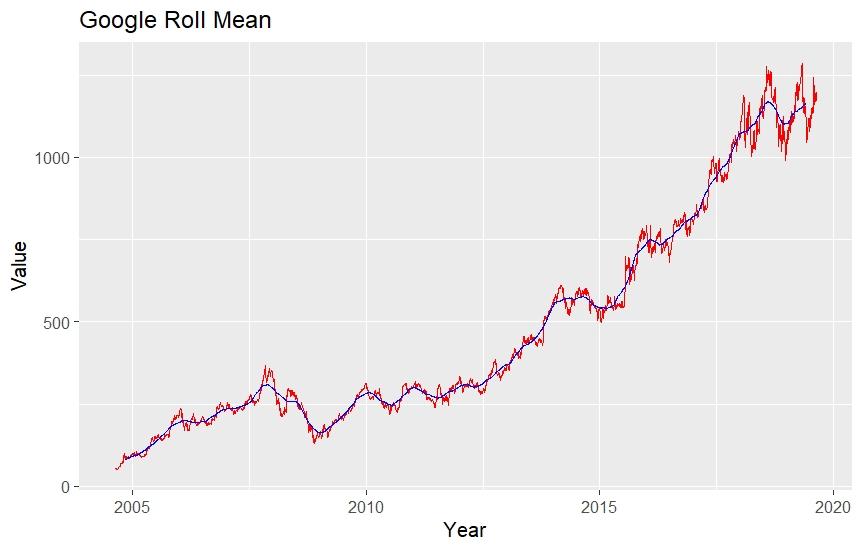
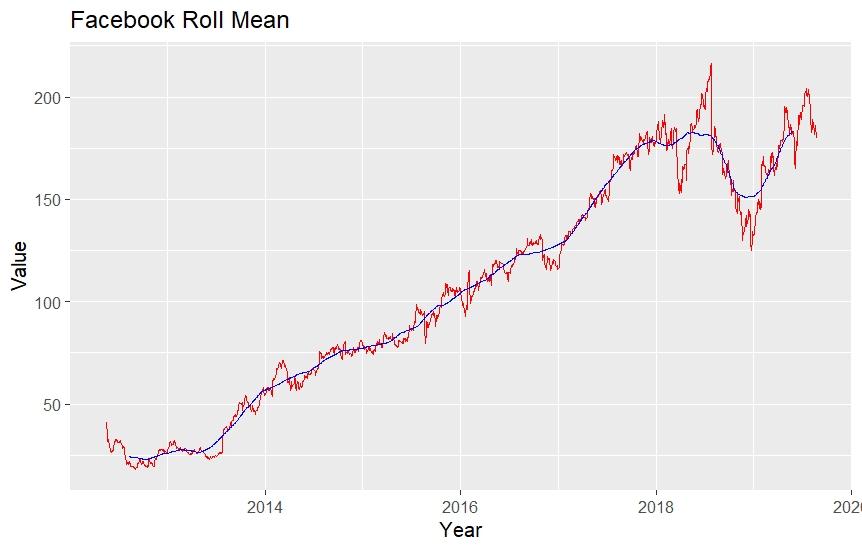
[Click here to go back to the main report](#_heading=h.c2ei7fokffn4)

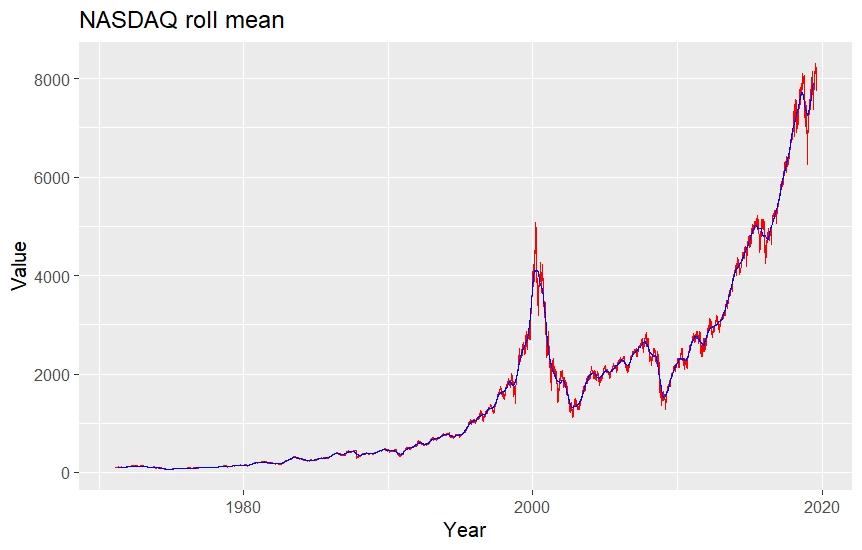
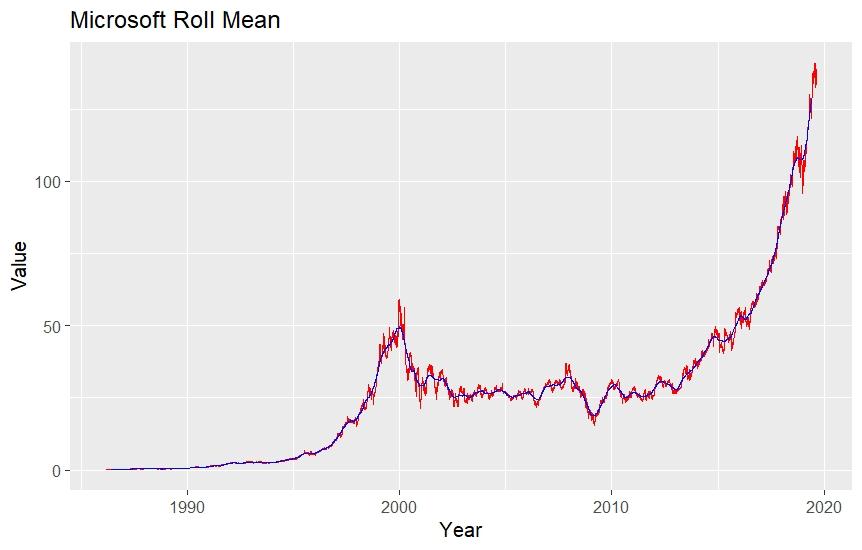
## 

## Appendix G ARIMA Time Series Model

### G.1 Graphs of rolling mean of individual metrics against time



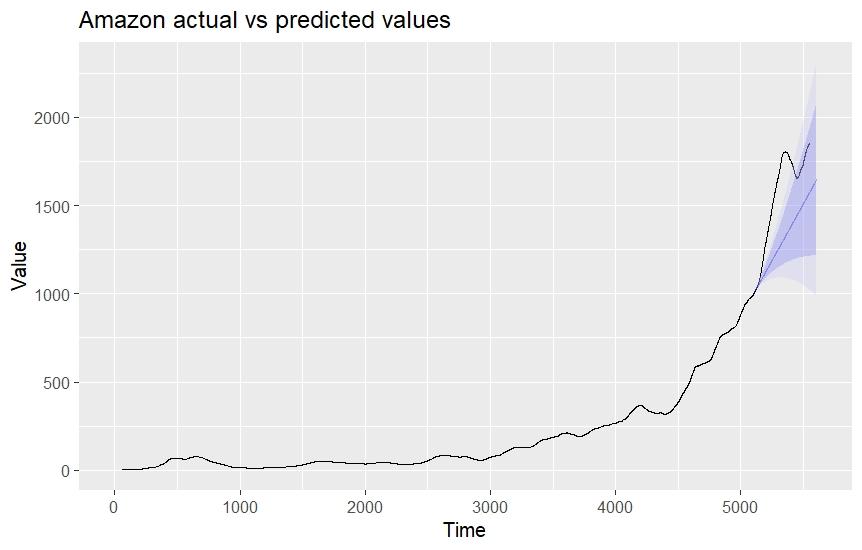


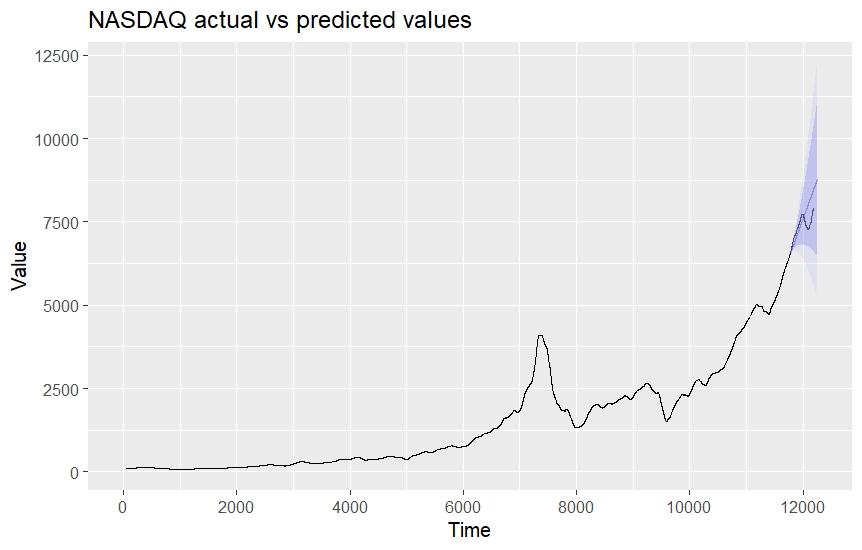
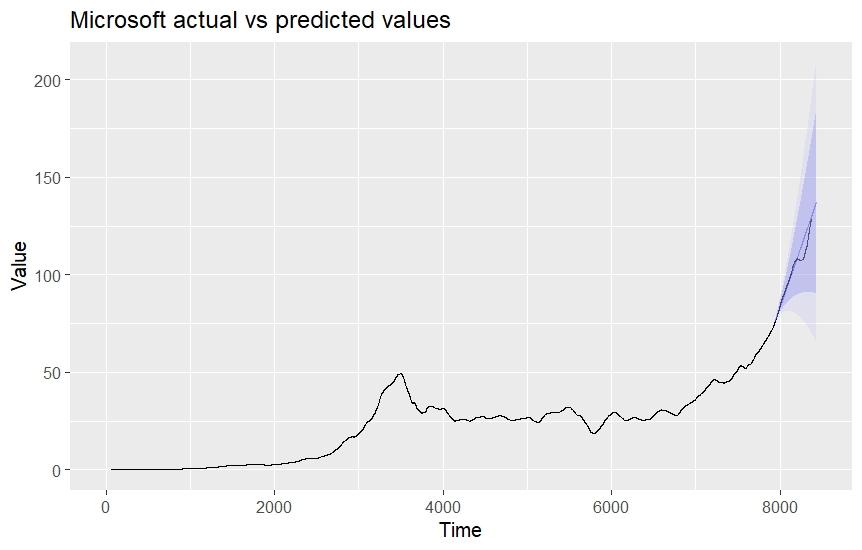
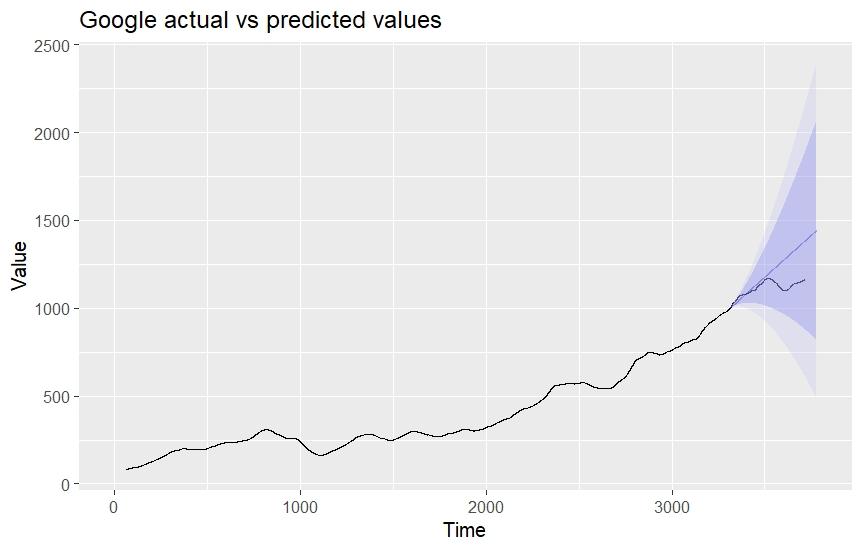
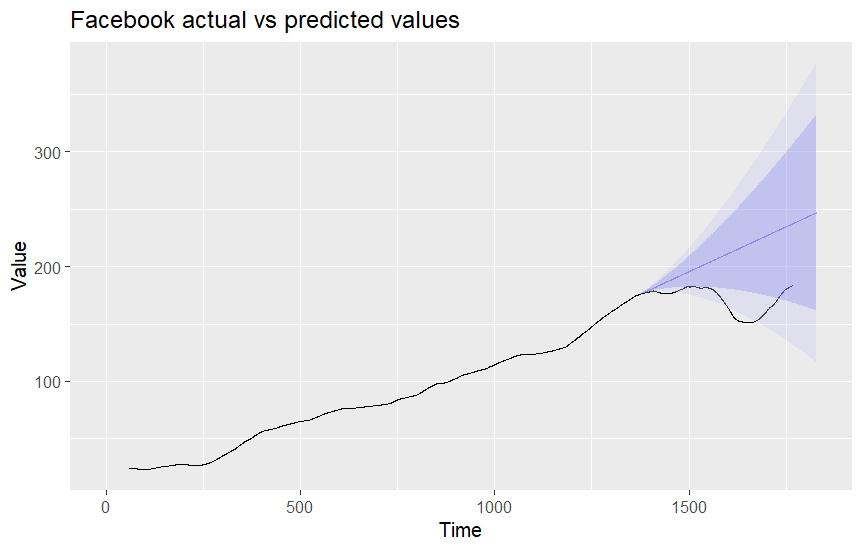
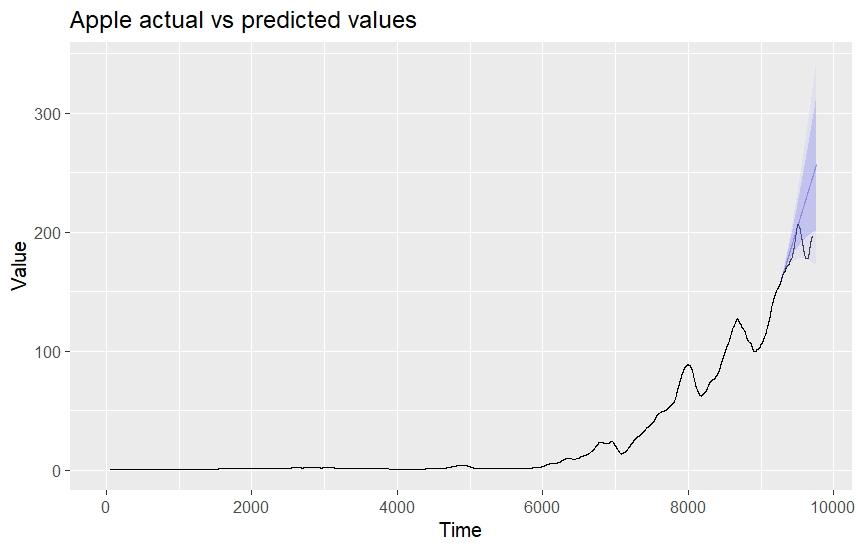


[Click here to go back to the main report](#_heading=h.ddwsnggdxwlh)

### G.2 Results of ARIMA Time Series model

|  |
| --- |
| Guide to interpret the graph |

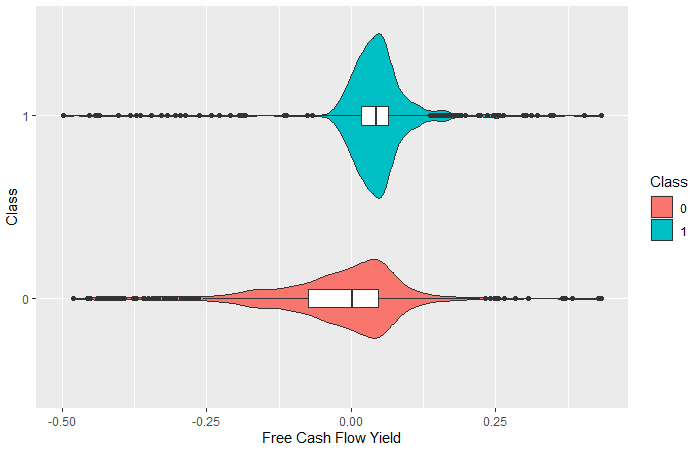




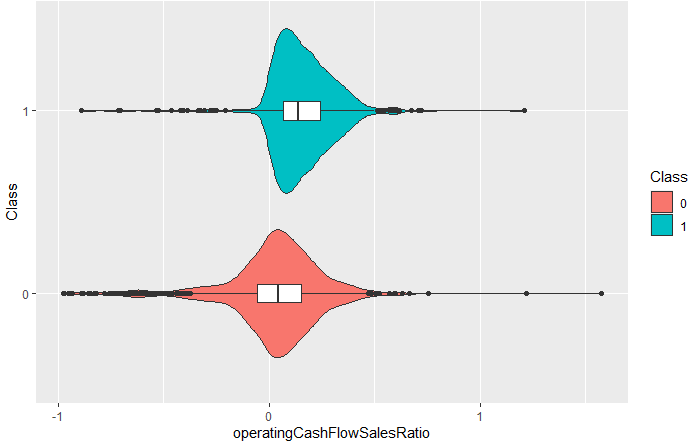
[Click here to go back to the main report](#_heading=h.y4bq21xb5piv)

## Appendix H Insights analysis

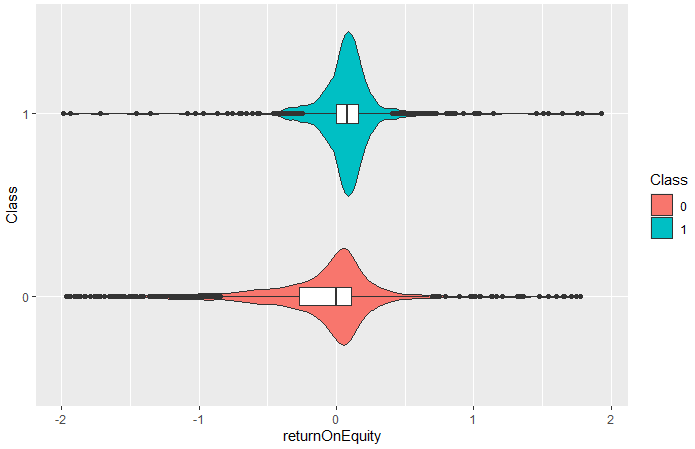
### H.1 Distribution of Free Cash Flow Yield by Class

****

### H.2 Distribution of Operating Cash Flow Sales Ratio by Class

****

### H.3 Distribution of Return On Equity by Class

****

[Click here to go back to the main report](#_heading=h.y4bq21xb5piv)

1. As no specific details about White Rock’s business model and investment approach were given, we reasonably assume White Rock to be a traditional asset management firm (in contrast to non-traditional asset management firms such as Citadel, Two Sigma, and Bridgewater, which adopted a purely technical approach for investment) [↑](#footnote-ref-1)
2. Financial barchart: A line graph with a barchart (consisting of open, close, high and low) at each point of time. [↑](#footnote-ref-2)
3. “Volume” column refers the trade volume of the stock [↑](#footnote-ref-3)
4. Average stock price = (high+low+open+close)/4 [↑](#footnote-ref-4)