**D214 Capstone – Performance Assessment**

Steven Schindler

Regression and Time Series analysis of market research dataset

Sewell, William; PhD

College of I.T., Western Governors University

August 04, 2023

Table of Contents

[Research Question 3](#_Toc142726208)

[Data Collection 3](#_Toc142726209)

[Data Extraction and Preparation 4](#_Toc142726210)

[Analysis 5](#_Toc142726211)

[7](#_Toc142726212)

[8](#_Toc142726213)

[8](#_Toc142726214)

[9](#_Toc142726215)

[9](#_Toc142726216)

[10](#_Toc142726217)

[13](#_Toc142726218)

[13](#_Toc142726219)

[Data Summary and Implications 14](#_Toc142726220)

[Sources 14](#_Toc142726221)

# Research Question

The Coca-Cola company went public in 1919 with 600,000 shares at 40$ per share (Nasdaq. (2020, December 6)). Over 100 years later the company holds a market share of 42% of the global soft drink market and reported revenue of $37.3 billion in 2020( tutor2u Economics. (n.d.)). This analysis attempts to make a predictive model of Coca-Cola stock prices of up to a year in advance. This analysis could then assist a stock trader with the best times to buy or sell Coca-Cola stock up to a year in advance.

Can a Regression and Time Series analysis make a predictive model for market research dataset of the Coca-Cola stock price? The null hypothesis is a predictive model for Coca-Cola Stock prices cannot be constructed with an accuracy of 70%. The alternative hypothesis is a predictive model for Coca-Cola Stock prices can be constructed with an accuracy of at least 70%.

# Data Collection

The data includes 60 years of stock price history collected and shared from the website Kaggle. Kaggle is an online community platform for data scientists that publish open-source datasets (DataCamp. (n.d.))1. As stated above the data shared is 60 years of stock price history from Jan 1st, 1962, to Oct 25th, 2022. The main advantage of using Kaggle is that the data is already collected and presented in an easy CSV format. A disadvantage is that the data is not the most recent as it only goes to Oct 25th, 2022. The data is located [here](https://www.kaggle.com/datasets/kalilurrahman/coca-cola-stock-live-and-updated) (Rahman, K. (n.d.)).

The fields in the data are listed in the table below:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Data type** |
| Date | Date from 1962-2022 | datetime/continuous |
| Open | Open price at opening of the stock market | continuous |
| High | High price for that day | continuous |
| Low | Low price for that day | continuous |
| Close | Close price at closing of stock market | continuous |
| Volume | Volume Traded | continuous |
| Dividends | Dividends paid | continuous |
| Stock Splits | Stock Splits | continuous |

# **Data Extraction and Preparation**

The next step is to prepare the data for analysis using Python as the tool of choice. Python has many tools used for data science including the Pandas library. Pandas offers many advantages such as the ability to efficiently handle large datasets and provide the data in an easy representation (Patil, R. (n.d.)). A disadvantage of Pandas is the memory consumption as working with large data sets Pandas will use a large amount of memory (Patil, R. (n.d.)).

A screenshot of a computer

Description automatically generated The first step of the cleaning process is to read the CSV file into an easily manipulated DataFrame, below shows that step as well as information about the data set.

A screenshot of a computer code

Description automatically generatedThe info method shows the data type of the column and that there are 15,311 entries with the number of non-null values for each column. Using the isna() and the sum() methods which check for null values and counts the number respectively, then confirm that there are no null values.

**Next convert the date column to a datetime object to give the date consistency of how it is formatted.** Duplicate values are the next thing to check for using, the nunique() method which gives the number of unique values that are shown below.

A screenshot of a computer

Description automatically generated

Date is the only column that has completely unique values which is alright though as the other columns could have prices that are the same such as, the high price for the day being equal to the close price of the day. There could also be dates where the values are the same but are on different dates such as, two Open dates that have the same value but are weeks apart. Checking for null values and duplicates as well as removing unnecessary columns the cleaning is now complete and the data is ready for analysis.

Python was chosen for this analysis over R or SAS because of its advantages such as being free and open source and being faster than R for large data sets (GeeksforGeeks. (n.d.)1). Some disadvantages are that Python is slower than SAS when it comes to data manipulation (GeeksforGeeks. (n.d.)1). Python also consumes much more RAM than R and as such the process gets much slower the more objects that need to be accessed(InterviewBit Blog. (n.d.)).

# **Analysis**

The two analysis techniques used are Multiple Linear Regression (MLR) and Autoregressive integrated moving average or Arima. MLR is a statistical technique that attempts to model the relationship between two or more explanatory variables and a response variable y fitting the data to a linear equation (Yale University Department of Statistics. (n.d.)). Arima is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends (Investopedia. (n.d.)). One advantage of regression is that it is simple to implement, and the coefficients are easy to interpret while a disadvantage is that it is susceptible to overfitting (GeeksforGeeks. (n.d.)).MLR is in the form of Y = B0 +B1X1+B2X2… with B0 being the y-intercept, B1,B2… being the coefficients of the X variables which are the independent variables, while Y is the dependent variable. One advantage of an Arima model is that it can handle a wide range of time series data if it is univariate, while a disadvantage is that because it is univariate it cannot capture the interactions and dependencies between different variables (LinkedIn. (n.d.)).

The analysis begins with an exploratory look at some data visualizations.A line graph of different colored lines

Description automatically generated

We see from the picture above that Open, Close ,High and Low are very similar which may lead to multicollinearity. We see from the heatmap below as well as the VIF scores that Open, Close, High and Low have high multicollinearity.

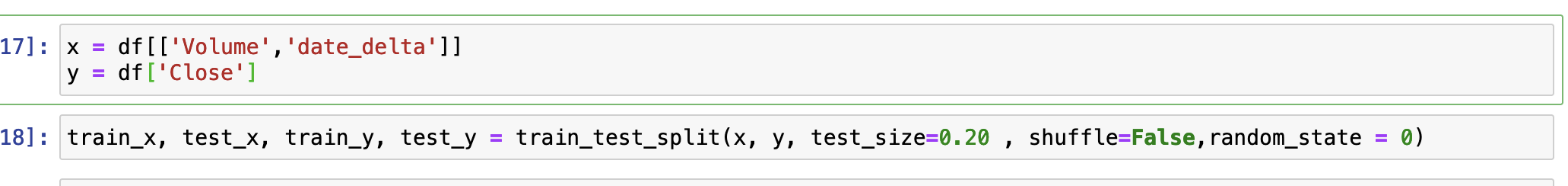
A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated

Because of the high multicollinearity the independent variables will be Date and Volume while the dependent variable will be Close for the Linear Regression analysis. Date has been converted to float as the number of days since the beginning of the data called delta\_date.

# A screenshot of a computer Description automatically generated

 **First the data is separated into a training data set and testing data set with the split at 80%:20% respectively.**

A screenshot of a computer code

Description automatically generated Next, import the regression library from sklearn and fit the data to the regression model to extract the coefficients and the y-intercept.

From the picture above we now know the equation is Y = -4.699 + -1.123X1 + 1.227X2. Next using the predict method on testing data and comparing it to actual prices of Coca-Cola this dataframe is created:

A screenshot of a computer program

Description automatically generated

**Then running mean absolute error(MAE), mean square error(MSE), and root mean square error(RMSE) on the dataframe these values are returned:**

# A screenshot of a computer code Description automatically generated

**Accuracy is calculated at 48.29:**

# A screenshot of a computer code Description automatically generated

**With this low accuracy and poor MAE, MSE, and RMSE the NULL hypothesis cannot be rejected.**

**Next, the Arima analysis is started by** a grouping **the data to monthly dates instead of daily dates using the mean. This is done to change the resolution of the data and reduce the number of data dimensions(**Copernicus Climate Change Service. (n.d.).**).**

A screenshot of a computer

Description automatically generated

Visualizing the monthly average Close stock price:

# A graph showing a line Description automatically generated

**Next visualize at the seasonal decompose with a period of 12 months:**

# A graph showing the time and time Description automatically generated with medium confidence

**We see that the Data is indeed seasonal so SARIMA will need to be used instead of ARIMA.** SARIMA (Seasonal Autoregressive Integrated Moving Average) is a **statistical technique used for forecasting time series data,** which is a series of observations recorded at **regular intervals over time (**GitConnected. (n.d.)**)**. Running the Dickey-Fuller test we get a p-value of greater than .05 further giving evidence of seasonality.

# **A screenshot of a computer program Description automatically generated**

Now the data is split into training and testing with an 80:20 split respectively.

A screenshot of a computer

Description automatically generated

Then auto Arima is ran so that the optimal P,D, and Q are found. P(AR) is the order of the autoregressive model. Q(MA) is the moving average model of order and D is degree of first differencing involved.( Hyndman, R. J., & Athanasopoulos, G. (n.d.))

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThis gives the optimal P, D, and Q as 2, 1, 0 respectively with a seasonal component of 12 and seasonal order of 2,0,0. Now Sarima can be run on the Monthly data of Close to start predictions and forecasting. 

A graph with numbers and lines

Description automatically generatedThe P-values are all less than .05, now predictions can be made.

**Above are predictions made up to one year in advance from Oct 27th, 2022, to Oct 27th, 2023. Then the data is split into a test set and compared with the mean predictions set gathered from the model.**

A graph with red and blue lines

Description automatically generatedA screen shot of a computer code

Description automatically generated

Finally, forecast graphs of up to one year are made.

# A screenshot of a computer code Description automatically generated

# A graph with lines and numbers Description automatically generated

**\*Note KO is the stock ticker symbol for Coca-Cola.**

A screenshot of a computer code

Description automatically generated**Then calculating the MAE, MSE and RMSE the values are 1.8, 4.9, 2.2 respectively.**

A screenshot of a computer

Description automatically generated The Accuracy is calculated at 99% accurate:

**With the high degree of accuracy, the Null Hypothesis can be rejected and a prediction model of at least 70% can be made for Coca-Cola stock prices.**

# **Data Summary and Implications**

With the low degree of accuracy for the linear regression analysis the Null Hypothesis cannot be rejected however, the time series analysis using Sarima has a degree of accuracy over 70% so the Null Hypothesis can be rejected. Therefore, a predictive model to forecast the stock price of a market data set can be made using time series analysis. A course of action is for companies or individuals can then use the analysis to invest accordingly in Coca-Cola. Another course of action is to modify the Sarima model to use a daily data set instead of the monthly data set that I used. The analysis is limited by its theoretical nature, in theory the model should predict market data sets with a high degree of accuracy in reality, many outside factors have an influence in stock market prices. All these outside factors are not included in the scope of this analysis. Further research could be conducted by trying to mitigate outside influences and reduce their impact on market data analysis. Another limitation of my analysis is that I used a monthly data set for my Sarima analysis a different approach would be to use Arima on a daily data set. Then forward testing can be applied to the Sarima model to see how well it does against real time daily market data (LinkedIn. (n.d.)).

As stated earlier the scope of this analysis is limited to theoretical analysis with no outside factors having an influence on the stock market such as inflation, interest rates, world events etc.(TIME. (n.d.).) Further study can be conducted by including current inflation rates or interest rates and incorporating them into the predictive model. Further study could also be used to expand the analysis’s time frame. The analysis conducted was for only 1 year in advance, future studies can expand 1 year to 1.5 years, 2 years, etc. Also mentioned earlier was a course of action to modify the Sarima model to use a daily data set instead of a monthly data set this can also be included in future studies.

# Sources

Nasdaq. (2020, December 6). A $120 Investment in Coca-Cola's IPO Would Be Worth This Much Money Now. Nasdaq. Retrieved August 05, 2023, from <https://www.nasdaq.com/articles/a-%24120-investment-in-coca-colas-ipo-would-be-worth-this-much-money-now-2020-12-06>

tutor2u Economics. (n.d.). Why Is Coca-Cola So Profitable? tutor2u Economics. Retrieved August 05, 2023, from <https://www.tutor2u.net/economics/reference/why-is-coca-cola-so-profitable>

DataCamp. (n.d.)1. What is Kaggle? Retrieved August 05, 2023, from <https://www.datacamp.com/blog/what-is-kaggle>

Rahman, K. (n.d.). Coca-Cola Stock - Live and Updated. Kaggle. Retrieved August 05, 2023, from <https://www.kaggle.com/datasets/kalilurrahman/coca-cola-stock-live-and-updated?select=Coca-Cola_stock_history.csv>

Patil, R. (n.d.). Exploring the Pros and Cons of Pandas. Medium. Retrieved August 06, 2023, from <https://medium.com/@rohanjpatil63/exploring-the-pros-and-cons-of-pandas-1675463971d4>

Yale University Department of Statistics. (n.d.). Linear Regression Analysis - Least Squares Estimation. Yale University Department of Statistics. Retrieved August 10, 2023, from <http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm>

Investopedia. (n.d.). Autoregressive Integrated Moving Average (ARIMA). Investopedia. Retrieved August 10, 2023, from <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp>

GeeksforGeeks. (n.d.). ML | Advantages and Disadvantages of Linear Regression. Retrieved August 10, 2023, from <https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/>

LinkedIn. (n.d.). What Are the Advantages and Disadvantages of ARIMA Models in Forecasting? Retrieved August 10, 2023, from <https://www.linkedin.com/advice/3/what-advantages-disadvantages-arima-models-forecasting>

Copernicus Climate Change Service. (n.d.). How to Resample and Aggregate Time Series Data. Retrieved August 11, 2023, from [https://cds.climate.copernicus.eu/toolbox/doc/how-to/14\_how\_to\_resample\_and\_aggregate/14\_how\_to\_resample\_and\_aggregate.html#](https://cds.climate.copernicus.eu/toolbox/doc/how-to/14_how_to_resample_and_aggregate/14_how_to_resample_and_aggregate.html)

GitConnected. (n.d.). Boost Your Sales Forecasting Accuracy with SARIMA: A Comprehensive Guide. Retrieved from <https://levelup.gitconnected.com/boost-your-sales-forecasting-accuracy-with-sarima-a-comprehensive-guide-553a8e8623d7>

Hyndman, R. J., & Athanasopoulos, G. (n.d.). Stationarity. Forecasting: Principles and Practice (2nd ed.). Retrieved August 11, 2023, from <https://otexts.com/fpp2/stationarity.html>

Hyndman, R. J., & Athanasopoulos, G. (n.d.). Autoregressive (AR) Models. Forecasting: Principles and Practice (2nd ed.). Retrieved August 12, 2023, from <https://otexts.com/fpp2/AR.html>

LinkedIn. (n.d.). What is Forward Testing? Hue Frame. Retrieved August 12, 2023, from <https://www.linkedin.com/pulse/what-forward-testing-hue-frame>

GeeksforGeeks. (n.d.)1. SAS vs R vs Python. Retrieved August 14, 2023, from GeeksforGeeks. <https://www.geeksforgeeks.org/sas-vs-r-vs-python/>

InterviewBit Blog. (n.d.). Python vs R. Retrieved August 14, 2023, from InterviewBit Blog. <https://www.interviewbit.com/blog/python-vs-r/>

TIME. (n.d.). How Are Stock Prices Determined? TIME. Retrieved August 14, 2023, from <https://time.com/personal-finance/article/how-are-stock-prices-determined/>