**Comparing Stocks of Storage Device Manufacturers**

Emilio Ramos Monzalvo

IST 652 Scripting for Data Analysis

Professor Deborah Landowski

Syracuse University // March 2021

# Table of Contents

[Table of Contents 1](#_Toc67145253)

[Introduction 3](#_Toc67145254)

[Data Selection 4](#_Toc67145255)

[Data Preprocessing 6](#_Toc67145256)

[Time Aggregation 6](#_Toc67145257)

[Scaling Data 6](#_Toc67145258)

[Additional Columns 8](#_Toc67145259)

[Analysis 8](#_Toc67145260)

[Correlation Study 8](#_Toc67145261)

[Auto Correlation Study 9](#_Toc67145262)

[Modeling 10](#_Toc67145263)

[Modeling Results 11](#_Toc67145264)

[Conclusion 13](#_Toc67145265)

[References: 14](#_Toc67145266)

# Introduction

Since the start of the computer age, data storage has been a key element in the growth potential of the functionalities of computers. Without a place to store your software and data, the uses of a computer will not expand. As an employee of a Data Storage Manufacturing Company, I have heard stories from the 1980s where our company was called crazy for inventing the first Hard Drive Disk (HDD) the size of a bookcase that could hold 1GB worth of data. Back then, 1GB was a needless amount of data. On the other hand, they are now considered visionaries as their efforts to change the way we view data storage led to personal computers having the option to hold two terabytes, or 2000GB, worth of data.

With the growth of data storage technologies, the growth of data also increased exponentially in recent years. According to IDC Global Data-Sphere, the amount of data generated in the world is going to get to 175 zettabytes, or 112GB, in 2025. Then, it should also be expected to see a significant growth in data storage device manufacturers, especially in their stock price.

**Figure 1: Data Sphere Growth over Time**

Chart, bar chart

Description automatically generated

In the world of data storage devices, there are two main technologies: Solid State Drives (SSD) and Hard Drive Disks (HDD). HDDs were invented back in 50s and are effective in storing data for a long period of time. On the other hand, SSDs are newer, and can have quadruple the read speed of an HDD. Consequently, SSDs are also significantly costlier. Both technologies have gone through an inexpressible number of improvements; therefore, there is only a handful of companies that control most of the market. According to a 2015 market report, less than ten companies share 99% of the market. Hence, the question arises, does the handful of companies follow a similar pattern of growth to keep up with the demand of data storage? If so, can we understand and predict the company’s stock price using the other company’s stock prices?

**Figure 2: Storage Market Share**

Chart, pie chart

Description automatically generated

# Data Selection

In the HDD world, there only exist three key companies. In the SSD world, there exists nine companies, but only five hold a significant amount of the market. Therefore, seven companies’ stock was selected for the study which are outlined below. Now it is important to note that SanDisk was absorbed by Western Digital in 2016.

**Table 1: Companies and Their Stock Information**

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Storage Tech Type** | **Stock Identifier** | **Stock Market Index** |
| Western Digital | SSD and HDD | WDC | NASDAQ |
| Seagate | HDD | STX | NASDAQ |
| Toshiba | HDD | TOSYY | OTC Markets |
| Intel | SSD | INTC | NASDAQ |
| Samsung | SSD | 005930.KS | KRW |
| Micron | SSD | MU | NASDAQ |
| Kingston | SSD | KINS | NASDAQ |

To bring the data into our local environment, we used a free package called Yahoo Finance Package called ‘yfinance’ which offers a straightforward API to pull data for our stock of interest. It first asks to provide the stock identifier, and a time frame. To avoid external factors, we excluded the 2008 Housing Market Crash, and the 2020 COVID Market Crash, so we were left with stock market data from January 2010 to December 2019. The API then returns a pandas Data Frame.

**Code 1: Pulling Data Using Yahoo Finance Python Package**

**Table 2: Finance Python Package Resulting Data Frame Columns**

|  |  |
| --- | --- |
| **Column** | **Meaning** |
| **High** | Highest stock price in time interval unit. |
| **Low** | Lowest stock price in time interval unit. |
| **Open** | Starting stock price in time interval unit. |
| **Close** | Final stock price in time interval unit. |
| **Volume** | Number of trades in time interval unit. |
| **Adjusted Close** | Close stock price adjusted by accounting actions by the company. |
| **Date** | Starting date of interval unit. |

# Data Preprocessing

## Time Aggregation

The volatility of stock is high if we view it by day. On the other hand, the volatility of the stock can be reduced by zooming out and viewing it by a larger time frame. Therefore, the stock prices were preprocessed and aggregated by week, month, quarter, and year. Overall, the daily view shows a high volatility while the yearly view results in time series much different from the rest. Therefore, we will focus on the weekly and monthly view of the data. In the weekly view we are left with 521 data points per company while in the monthly we are left with 120 data points per company. The number of rows is greatly reduced and will lead to a much smaller dataset.

**Figure 3: WDC Stock Price by Different Time Frames**

Chart, histogram

Description automatically generated

## Scaling Data

As previously shown in Table 1, the stock price for most companies come from the NASDAQ stock index except Samsung and Toshiba. This also mean that the stock price is in a different currency. We could pass the foreign currencies back to US Dollars (USD), but we would still be left with a different USD baseline for each stock as seen in Figure 4. For example, Kingston shows an average adjusted close stock price of 7.48USD while Western Digital’s is 52.17USD. Therefore, to make the stocks comparable, we used a Min Max Scaler, and you can see the results in Figure 5.

**Figure 4: US Companies’ Stock Price Distribution**

Chart, line chart

Description automatically generated

**Figure 5: Stock Price Before and After Scaling**

Chart, histogram

Description automatically generated

## Additional Columns

In order to create a model, we need a standard target value for all stocks. In this case, we can clearly see that the stock market price for each company can be different from each at different time periods. For example, Western Digital is closer to .2 on 2019 while Samsung is closer to .8. On the other hand, the percent change for a stock price is highly volatile and it does not follow a trend.

**Table 3: New Columns for Model Input**

|  |  |
| --- | --- |
| **Column** | **Meaning** |
| **Adj. Close Lag 1** | Adj. Close shifted up by one time interval. |
| **Adj. Close Percent Change Lag 1** | Adj. Close Percent Change shifted up by one time interval. |
| **Volume Lag 1** | Volume shifted up by one time interval. |
| **Volume Change Lag 1** | Change in the Volume by one time interval |

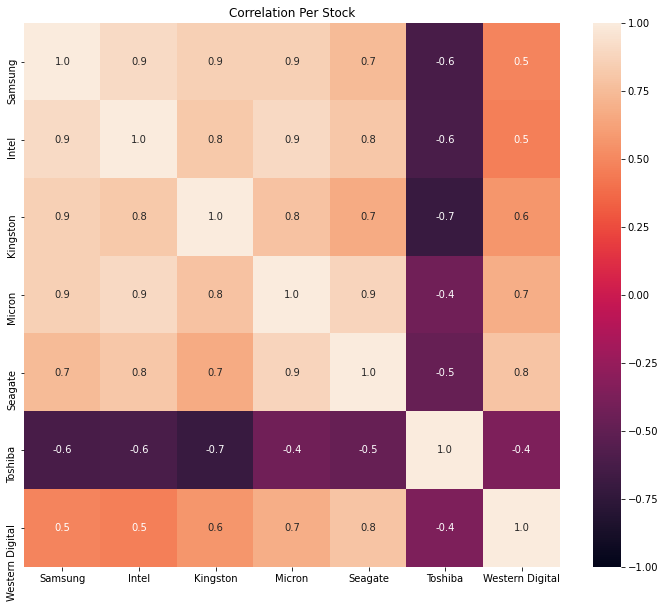
# Analysis

## Correlation Study

Since our goal is to predict the stock price changes by using other company’s stock price, we first find the correlation for one company to the other. In Figure 6, we can clearly see that Toshiba follows a very different trend than the rest of the rest companies. It is negatively correlated with all companies and if we look at the stock price over time, it shows a clear negative trend in its stock. Therefore, we should exclude it from the model building analysis.

Another highlight from the study is the fact that Western Digital’s stock shows a weak positive correlation to the other companies except Seagate. This is expected as Seagate it’s Western Digital’s main competitor in the HDD world. Since Western Digital shows a weak correlation, but a correlation nonetheless, we will still use it for the model.

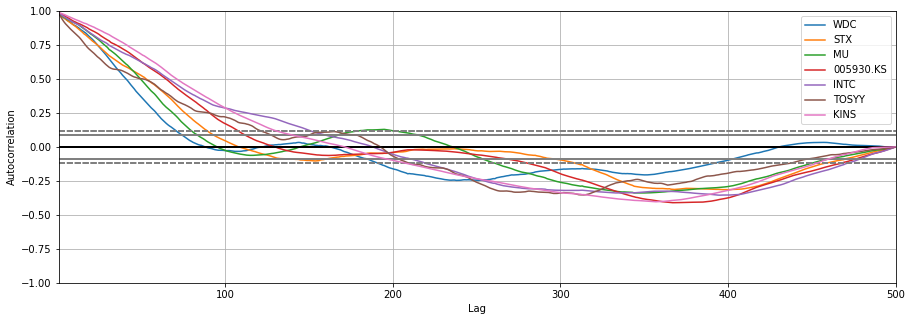
**Figure 6: Correlation Study between Companies**

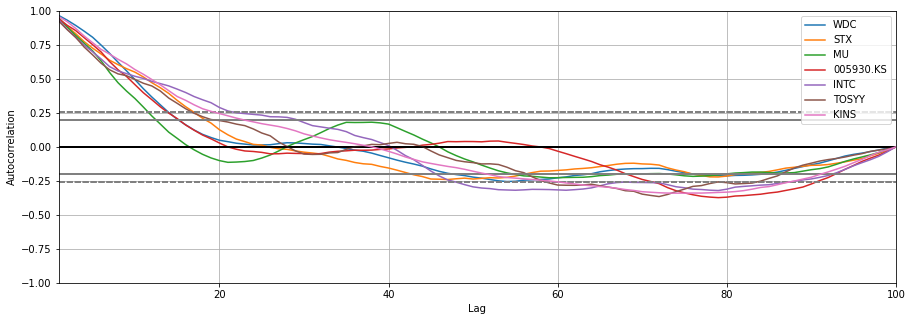


## Auto Correlation Study

In order to see how much lag should be used in our model, we plot the autocorrelation function for each stock’s time series. The autocorrelation plots by week show how about one-hundred time-interval before the present carry important information on the current stock market value. This means that if we want to predict the stock price, we need the data for the previous 100 weeks. If we look at it by month, we get about 20 months, by day 500 days. Therefore, it shows a similar pattern of about 1.5years to 2years worth of data needed to understand the time series trend. Only problem with using 100 weeks’ worth of data is that our models are not capable to use take many lag columns. Therefore, we only tried using one lag which resulted in 4 input columns for the models.

**Figure 7: Auto Correlation Study between Companies by Week**



**Figure 8: Auto Correlation Study between Companies by Month** ****

# Modeling

In order to Model Time Series data, we used a set of five different Regression Models from the python Package SKLearn which are highlighted in Table 4.

**Table 4: Model Description**

|  |  |
| --- | --- |
| **Model** | **Model Type** |
| **Linear Regression** | Linear |
| **MLP Regressor** | Neural Network |
| **KNN Regressor** | Neighbor |
| **Random Forest Regressor** | Tree |
| **SVM Regressor** | Support Vector |

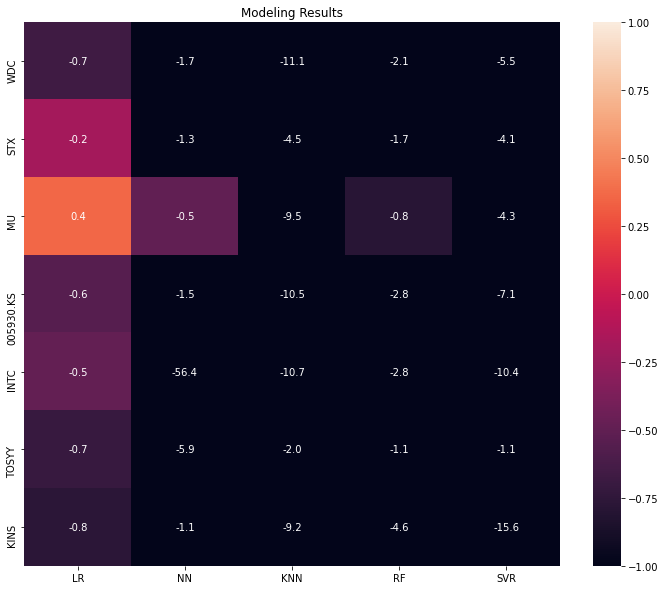
The data was splitted into a training and testing data where the training set included 80% of the data which included data before December 2017. Then, we sequentially trained each model using a cross validation approach with 10 time series splits. The score of each model was recorded. This was done for each stock and for the week and month time interval.

# Modeling Results

In order to test the accuracy of the results, the cross validation score function from SKLearn returns a score value which should be close to one if the model shows a good prediction ability, close to zero if it’s poor, and negative if it shows no predictability. Almost all models show a negative score except two.Linear Regression and Neural Network show the best predictibility when we aggreagted the data by week. The aggregation by month showed no good results. Overall, the best model was the Linear Regression, but it does not show a good enough result to be able to use it to predict each other’s stock prices. We are looking for an accuracy score of at least .8 in all categories.

**Figure 9: Modeling Results by Week** Graphical user interface, application

Description automatically generated

**Figure 10: Modeling Results by Month** 

# Conclusion

Our attempts to accurately predict the changed in individual stock market prices, we were unsuccsseful with the columns used . While the predictive analysis shows an extremely poor predictability score in most cases, there does seem to be a strong correlation among most Data Storage Device Manufacturers when we aggreagate the data by week. The Autocorrelation study shows a significant amount of lag data needed to accurately describe the variance in the data which would make our column and row ratio close to 1:1 which can lead to overfitting. Next steps would include finding a more scalable approach for the limited data points that we have. One where we can add more lag columns into the input variable set. Overall, with more resources, the preliminary results do show correlation in the stock price for the companes selected, and a promising predictability of individual stock price.

# References:

1. Model Building:
   1. <https://towardsdatascience.com/time-series-modeling-using-scikit-pandas-and-numpy-682e3b8db8d1>
2. Datasphere Growth:
   1. <https://www.datanami.com/2018/11/27/global-datasphere-to-hit-175-zettabytes-by-2025-idc-says/>
3. Storage Device Market:
   1. <https://www.businesswire.com/news/home/20151130006114/en/HDDs-SSDs-Grow-Seasonally-to-143-Exabytes-TRENDFOCUS-3rd-Calendar-Quarter-Analysis>
4. Yahoo Finance Stock Python Package:
   1. <https://pypi.org/project/yahoofinancials/>