Table of Contents

Table of Figures

Table of Tables

Introduction

Stocks trading has been going on for a long time now and has been a great medium for people as well as companies to grow their investment since its inception. Since, stocks are unpredictable and are highly volatile, that is where we implement Machine Learning and algorithms to identify patterns in data and to make better decision on the stocks (Pahwa & Agarwal, 2019). In this assignment, I will make use of various regression and clustering models on a stock exchange dataset and interpret the findings.

Main Objectives

The major objectives of the analysis that I have outlined analysing the dataset are:

- To perform exploratory data analysis on the stock dataset to view the trends of different exchanges.
- To view the relationship between the various attributes of data and how they affect the close price of the stocks.
- To identify the major variables that affect the stock close prices at the end of the day.
- To cluster the data based on the various attributes.
- To improvise the regression and clustering models by implementing other algorithms to get better findings.

I have prepared these research questions based on the attributes and structure of the stock dataset and how the provided attributes might be affecting one another. I want to view the trends of stocks in all these years and produce ML models to further predict or forecast the stock exchange trends in the coming years using the available data.

Methodology and Dataset

Dataset features and summary

I chose the Stock exchange data for my analysis, and it incorporates the stock values for various exchanges around the world starting from end of December of 1965. The main data frame in the dataset is the indexData which consists of 110253 rows and has 8 different attributes. The attributes are:

- Index: The index of the stock exchange
- Date: The day for the certain stock value.
- High: The highest price of the stock value during the day.
- Low: The lowest price of the stock during the day.
- Open: The opening price of the stock for the day.
- Volume: The total share of stocks that has been traded during the day.
- Adj Close: The closing value adjusted to reflect any corporate actions.

```
df_indexData
```

	Index	Date	Open	High	Low	Close	Adj Close	Volume
0	NYA	1965-12-31	528.690002	528.690002	528.690002	528.690002	528.690002	0.0
1	NYA	1966-01-03	527.210022	527.210022	527.210022	527.210022	527.210022	0.0
2	NYA	1966-01-04	527.840027	527.840027	527.840027	527.840027	527.840027	0.0
3	NYA	1966-01-05	531.119995	531.119995	531.119995	531.119995	531.119995	0.0
4	NYA	1966-01-06	532.070007	532.070007	532.070007	532.070007	532.070007	0.0
...	...	...	...	...	...	...	...	...
110248	N100	2021-05-27	1241.119995	1251.910034	1241.119995	1247.069946	1247.069946	379696400.0
110249	N100	2021-05-28	1249.469971	1259.209961	1249.030029	1256.599976	1256.599976	160773400.0
110250	N100	2021-05-31	1256.079956	1258.880005	1248.140015	1248.930054	1248.930054	91173700.0
110251	N100	2021-06-01	1254.609985	1265.660034	1254.609985	1258.579956	1258.579956	155179900.0
110252	N100	2021-06-02	1258.489990	1263.709961	1258.239990	1263.619995	1263.619995	148465000.0

110253 rows × 8 columns

Figure 1: IndexData Data frame

The other dataset provided along with the first data is the index info data, which consists of the information about the region of the exchange, the exchange name, the index of the exchange and the currency used by the certain exchange to trade their shares.

```
df_indexInfo
```

	Region	Exchange	Index	Currency
0	United States	New York Stock Exchange	NYA	USD
1	United States	NASDAQ	IXIC	USD
2	Hong Kong	Hong Kong Stock Exchange	HSI	HKD
3	China	Shanghai Stock Exchange	000001.SS	CNY
4	Japan	Tokyo Stock Exchange	N225	JPY
5	Europe	Euronext	N100	EUR
6	China	Shenzhen Stock Exchange	399001.SZ	CNY
7	Canada	Toronto Stock Exchange	GSPTSE	CAD
8	India	National Stock Exchange of India	NSEI	INR
9	Germany	Frankfurt Stock Exchange	GDAXI	EUR
10	Korea	Korea Exchange	KS11	KRW
11	Switzerland	SIX Swiss Exchange	SSMI	CHF
12	Taiwan	Taiwan Stock Exchange	TWII	TWD
13	South Africa	Johannesburg Stock Exchange	J203.JO	ZAR

Figure 2: Index Info

The final dataset is the processed index data which contains the same attributes as the index data, but an additional attribute named ‘CloseUSD’ which converts all the close values to USD.

```
df_indexProcessed = pd.read_csv("indexProcessed.csv")
df_indexProcessed.head()
```

	Index	Date	Open	High	Low	Close	Adj Close	Volume	CloseUSD
0	HSI	1986-12-31	2568.300049	2568.300049	2568.300049	2568.300049	2568.300049	0.0	333.879006
1	HSI	1987-01-02	2540.100098	2540.100098	2540.100098	2540.100098	2540.100098	0.0	330.213013
2	HSI	1987-01-05	2552.399902	2552.399902	2552.399902	2552.399902	2552.399902	0.0	331.811987
3	HSI	1987-01-06	2583.899902	2583.899902	2583.899902	2583.899902	2583.899902	0.0	335.906987
4	HSI	1987-01-07	2607.100098	2607.100098	2607.100098	2607.100098	2607.100098	0.0	338.923013

Figure 3: Processed Index Data

Methodology

After importing and viewing the dataset, I cleaned the data removing any null values from the indexData data frame and reset the index. I then changed the date attribute to a datetime format to select values based on data time index effectively. Exchange index is the major categorical variable in the data; thus, it is used to group data based on exchanges. I then viewed the general statistics of the attributes, which showed me that the mean, standard deviation, the percentiles of data were very close to each other, for the various variables which has presented me the idea that the relationship between the attributes might be very high.

Furthermore, I generated a new feature named 'Price Change' to analyse the difference between the close and open values for the day.

Price Change = Close – Open

I then generated line plots of the attributes against each other viewing the trends of data throughout the years, majorly my focus was identifying the trends in close values, and the volume of stocks being traded as it seemed to have a lesser correlation to other attributes based on the basic statistics.

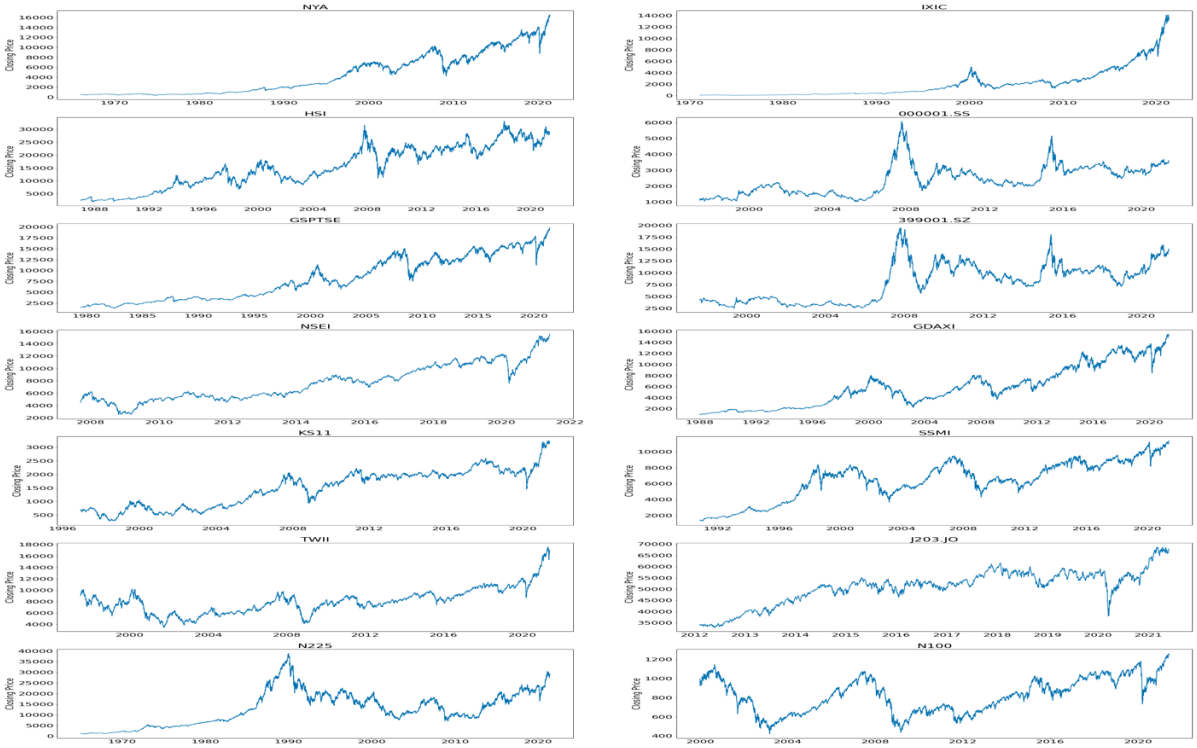

Figure 4: Closing price throughout the years for exchanges

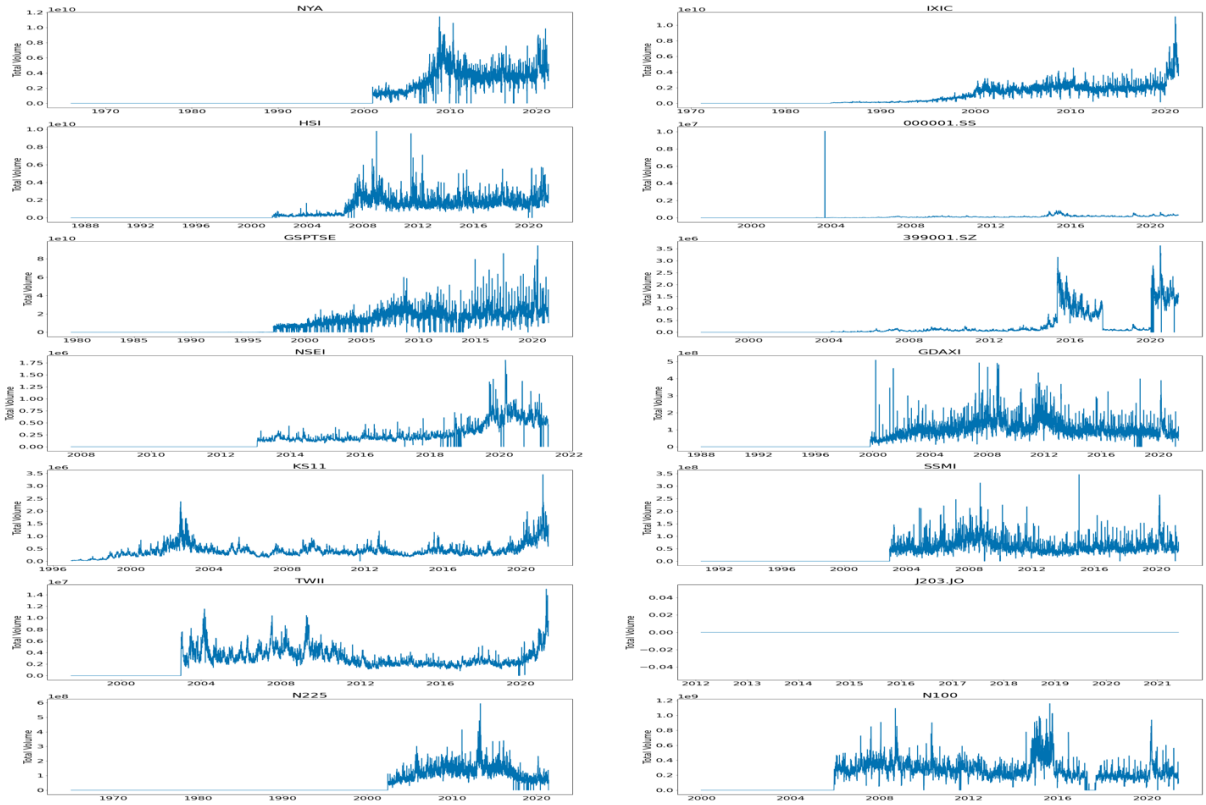

Figure 5: Volume of shares traded for different exchanges

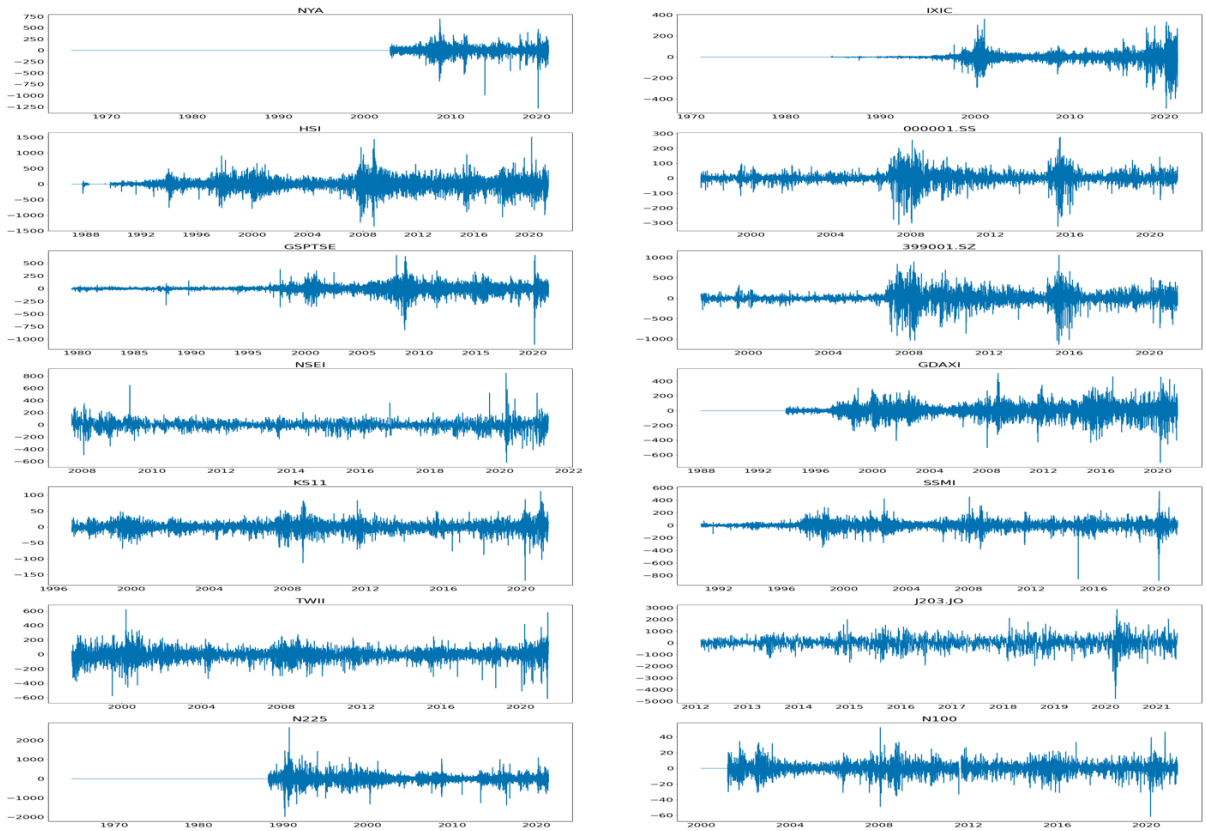

Figure 6: Price Change through the years for the exchanges

Afterwards, we moved on to the modelling and analysis of the data, where I implemented regression approaches as well as clustering approaches to view the relationship among the attributes and cluster of the variables. Despite the implementation of clustering approaches, I will majorly analyse the three regression techniques namely Multiple Linear Regression, Lasso Regression and Ridge Regression and their effectiveness in predicting the target variable which is the close price (Freund, et al., 2006). For the clustering, I will view the general cluster of the data generated by DBSCAN which is a density-based clustering method (Mahesh & Reddy, 2016).

Results

For the regression I make use of the categorical variable in the data as well to form a more robust regression measurement among all the attributes in the data frame. I implement the feature from pandas to generate dummy values for all the index of exchanges, so I can also pass the index values as parameters for the regression modelling and include the categorical difference as well. Before moving on to multiple linear regression and other regressions, I plotted a simple regression line of the data to identify the relationship between the high values and the close values. Figure 7 showcases a very strong relationship between the close and the high values as all the data points almost fall on the regression line, and the attributes display a linear relationship between them.

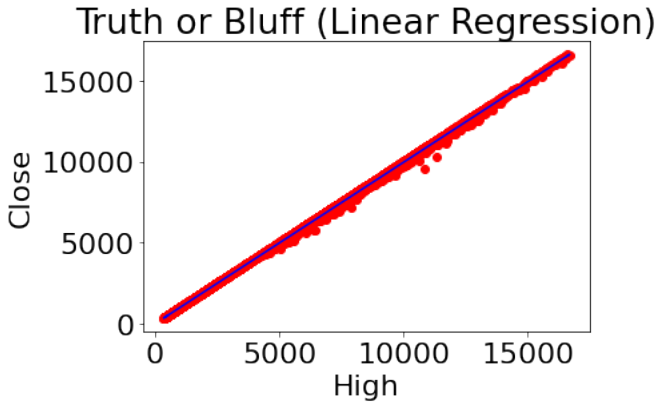

Figure 7: Linear Regression model

Regression Models

a. Multiple Linear Regression

The first model of my analysis is the multiple linear regression model which helped me to identify the robustness of the attributes and the strong relationship between them (Patel, et al., 2015). I fitted the model with my training data, and viewed the predictions on the test data, which provided me a high R^2 value of 0.99997 or 99.98% that denotes strong output variability.

b. Lasso Regression

The second model of my analysis is the Lasso Regression model where I implemented the model on the same data used for my previous model to keep the uniformity in interpreting the results (Ranstam & Cook, 2018). I firstly performed Lasso with 5-fold cross-validation to get the suitable alpha value which resulted me in getting an alpha value of 9.012. I then fitted the Lasso Regression model using this value and got an R^2 value of 0.9998.

c. Ridge Regression

The last regression model I implemented is Ridge Regression (Marquardt & Snee, 1975) where I also implemented cross-validation to get a suitable alpha value of 0.005 for the model. After implementing the Ridge regression model with alpha value as 0.005 I got a R^2 value of 0.9999 which is a 0.0001 increase from the previous model signifying the ridge regression to have the highest output variability among all the models.

Besides from measuring the output variability, I found it interesting that I was getting such high R^2 values, thus I performed a Permutation feature importance method to view which attributes were affecting the prediction of the close value the most. I found the attributes, High to be the most important, followed by the Low attribute and the Open attribute.

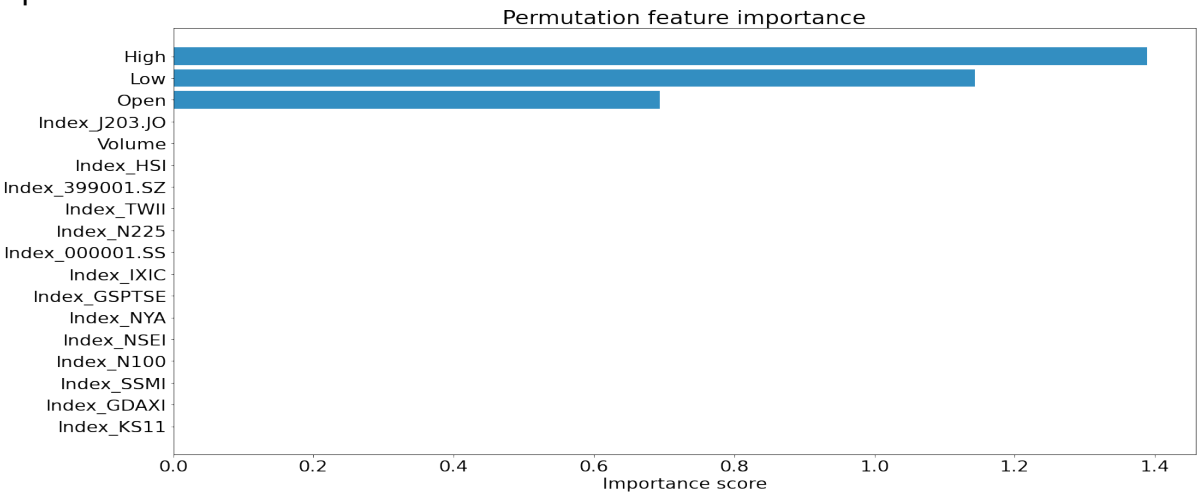

Figure 8: Permutation feature importance

Clustering Models

a. DBSCAN

Besides the regression model for predicting the close prices, I also implemented a DBSCAN model to separate the clusters high density to clusters of low density and to also visualize the outliers in the data (Ali, et al., 2010). I implemented Principal Component Analysis to identify the two main axes of variance in the data. I used the K-distances chart to find the suitable epsilon value of 0.3 and minimum samples of 7. DBSCAN presented me with 17 labels which are shown in Figure 9.

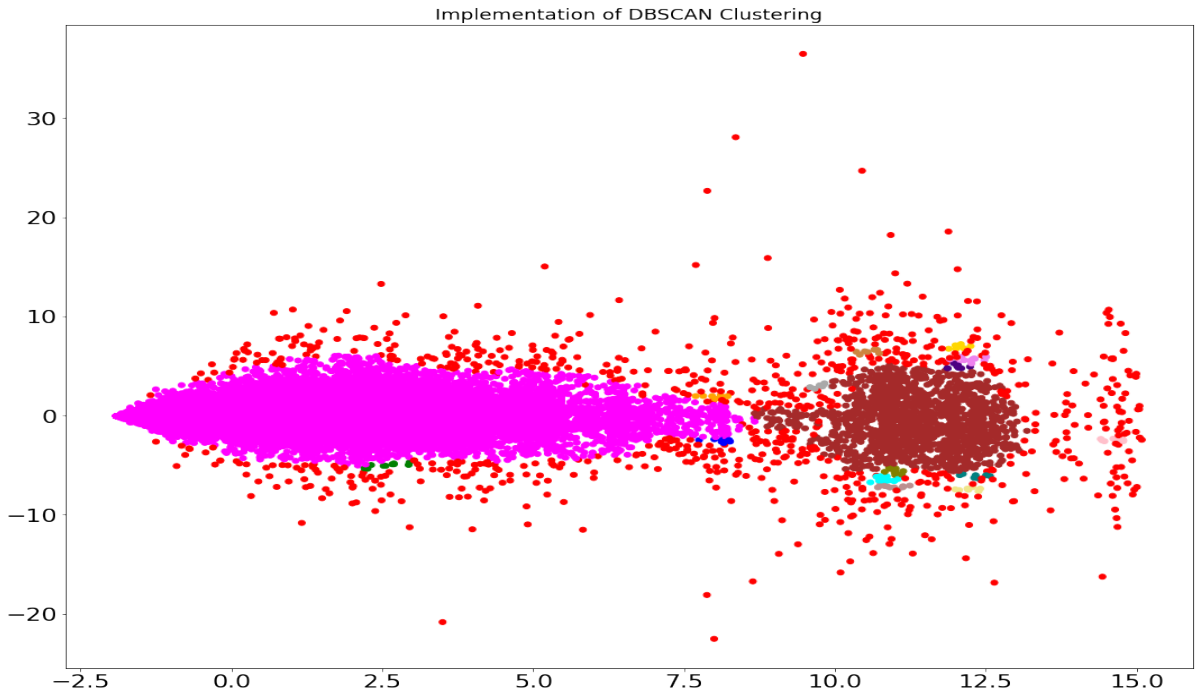

Figure 9: DBSCAN Clustering

Comparison of the Models

Regression Model	R^2 score	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Multiple Linear Regression	0.9997	2336.91495	48.34165
Lasso Regression	0.9998	12965.992	113.868
Ridge Regression	0.9998	7263.45	95.226

Table 1: Comparison of regression models using various metrics

Here, I have found that the Lasso regression models, and the Ridge regression model have high R^2 value and the dependent variables of the data fits better in these models compared to the Multiple Linear regression model. But despite the lower R^2 value, Multiple Linear regression model has a significantly better MSE and RMSE compared to the other two models. The low RMSE of Multiple linear models signifies the lower variance in the residuals or lower difference in the original and the predicted values in our data. Thus, we can conclude that Multiple Linear regression model is the best regression model due to the lower variance and standard deviation of the residuals amongst all the models (Botchkarev, 2018).

Discussion

Through the analysis I have found that the attributes ‘High’, ’Low’ and ‘Open’ affect the target variable the most, as they have the most importance in the regression models to help predict the closing price. I also noticed the relationship between ‘High’ and ‘Close’ was linear and helped produce robust prediction models. I identified Multiple Linear Regression model to be the best regression model to predict the close values, despite the Lasso Regression and Ridge Regression having a better variability score. While clustering I found three major clusters and identified some of the outliers in the data with the help of DBSCAN. The relationship between the independent variables and the dependent variables for the various exchanges has been quite linear through the years. To further improve on the analysis, we can make use of classification techniques and make use of other clustering techniques such as OPTICS and KMeans clustering which I have implemented in the python notebook. The analysis can be improved by incorporating missing values in some of the data for exchanges like J203.JO and 000001.SS. The data can also be improved by adding further new features such as adjusted prices for open, high and low values as well.

Works Cited

Ali, T., Asghar, S. & Sajid, N. A., 2010. *Critical analysis of DBSCAN variations.* s.l., IEEE.

Botchkarev, A., 2018. *Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology,* s.l.: arxiv.

Freund, R. J., Wilson, W. J. & Sa, P., 2006. *Regression Analysis.* 2nd Edition ed. s.l.:Elsevier.

Mahesh, K. K. & Reddy, A. R. M., 2016. *A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. Pattern Recognition.,* s.l.: Elsevier.

Marquardt, D. W. & Snee, R., 1975. Ridge regression in practice. *The American Statistician ,* 29(1), pp. 3-20.

Pahwa, K. & Agarwal, N., 2019. *Stock Market Analysis using Supervised Machine Learning.* s.l., s.n., pp. 197-200.

Patel, J., Shah, S., Thakkar, P. & Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications,* 42(1), pp. 259-268.

Ranstam, J. & Cook, J., 2018. LASSO regression. *Journal of British Surgery,* 105(10), pp. 1348-1348.