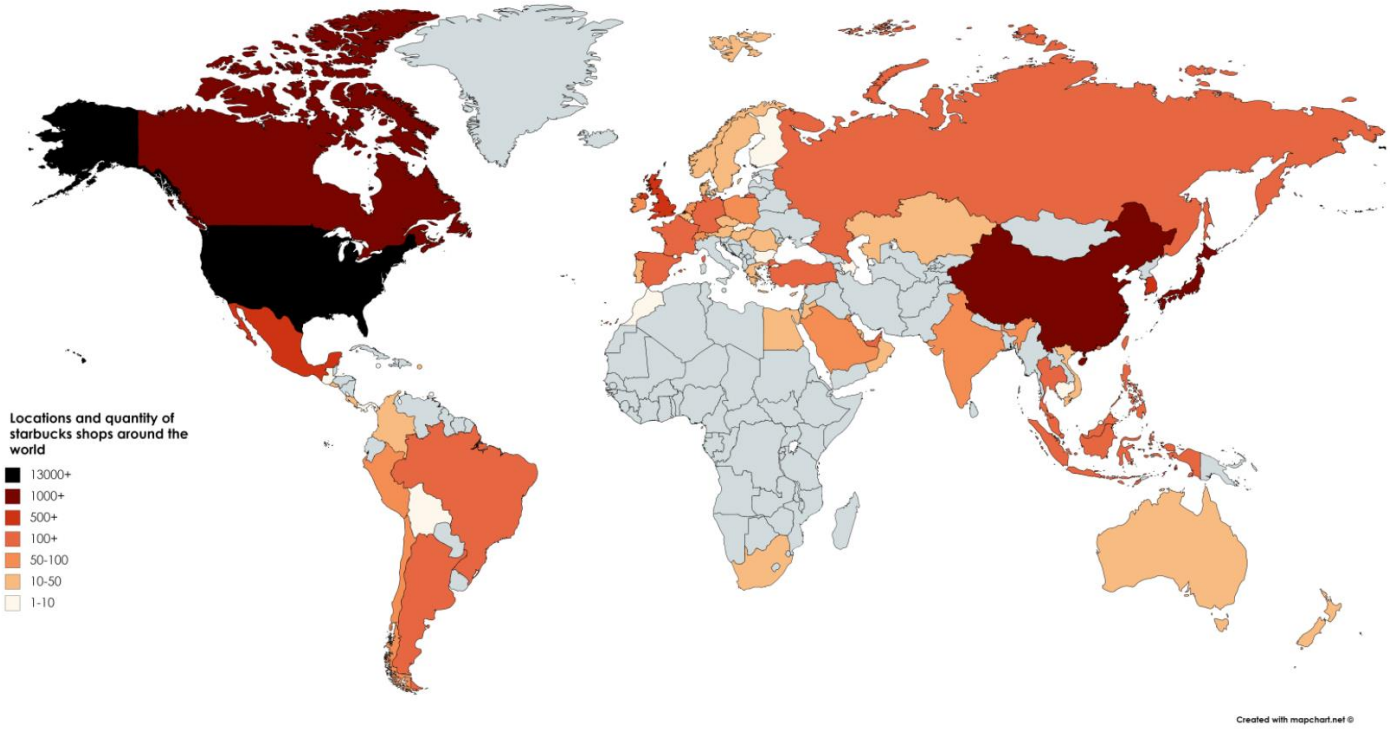
Starbucks time series model

Presented by: Ghliah maher.
Dr. Mejdal alqahtani

Content

Introduction: Starbucks Locations



Project Statement

This project studies the stock market prices for Starbucks company, then predict the closing price in the next year, using the ARIMA time series model.

Data and EDA

The data was collected from MarketWatch.
The collected data started from 1st Jan 2016 until 6th Dec 2021.

	0	1	2	3	4	5
0	Date	Open	High	Low	Close	Volume
1	12/30/2016	56.28	56.45	55.40	55.52	8,344,508
2	12/29/2016	56.35	56.47	56.14	56.32	3,781,721
3	12/28/2016	56.80	56.90	56.25	56.35	5,548,726
4	12/27/2016	56.99	57.39	56.81	56.86	4,186,157

- **Data cleaning:**

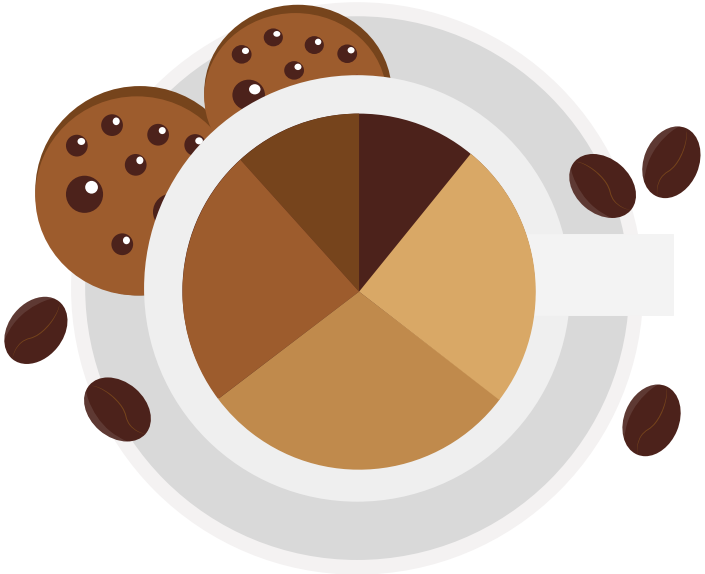
1. Rename the columns.
2. Drop unused columns, rows which included names of the columns, NA.
3. Set the index to the date.
4. Converting the date into date time, and the close column into float.
5. Sort the data frame.
6. Split the data frame into: train, and from the training set taking the validation set ,and lastly is the testing set.

DATE	CLOSE
2016-01-04	58.26
2016-01-05	58.65
2016-01-06	58.13
2016-01-07	56.69
2016-01-08	56.63
...	...
2021-11-30	109.64
2021-12-01	108.66
2021-12-02	111.42
2021-12-03	111.24
2021-12-06	113.36

Algorithm

- **ARIMA model:**

1. Checking the stationarity, using the dicky fuller and KPSS test.
2. Determine the best fit of the ARIMA model, by trying different orders of the p and q values, based on the lowest root mean of squared errors.
3. There is two approaches, first is to take all the dates from 2016 until 2021. the second is excluding the period between 2020- 2021.



Results: Two Approaches

First approach: from 2016-2021

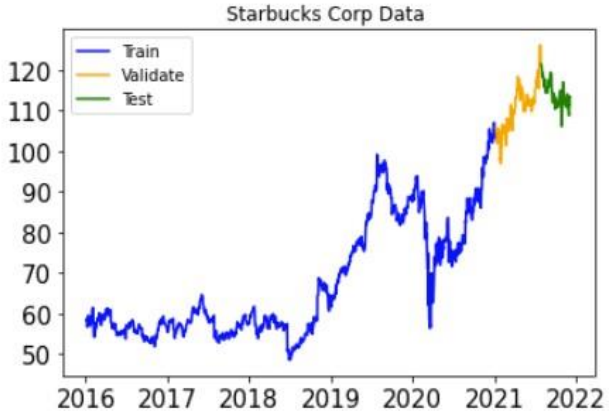
ARIMA (4,1,2)

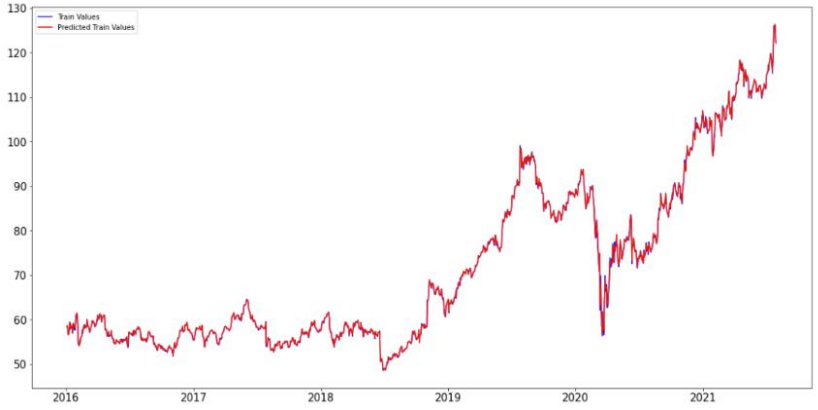
RMSE for the training set: 0.23596924955878643.

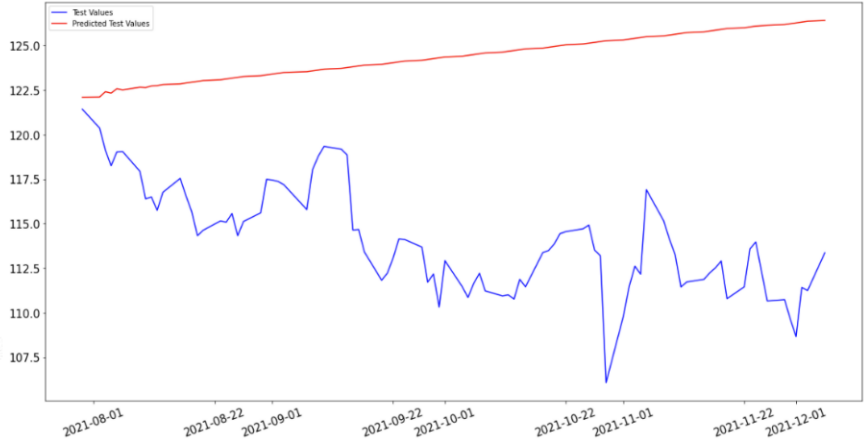
RMSE for the testing set: 10.993793405368965.

And the mean of the testing set is: 114.085778.

Value	Best order based on the RSME
(1,1,0)	0.18177178481712283 6.650977012275106
(2,1,1)	0.2288112984080849 6.559330841408596
(1,1,3)	0.2381034677522918 6.534312738053918
(4,1,2)	0.2583877561807341 6.487219747280541
(0,1,5)	0.19787184703825342 6.571224612712827



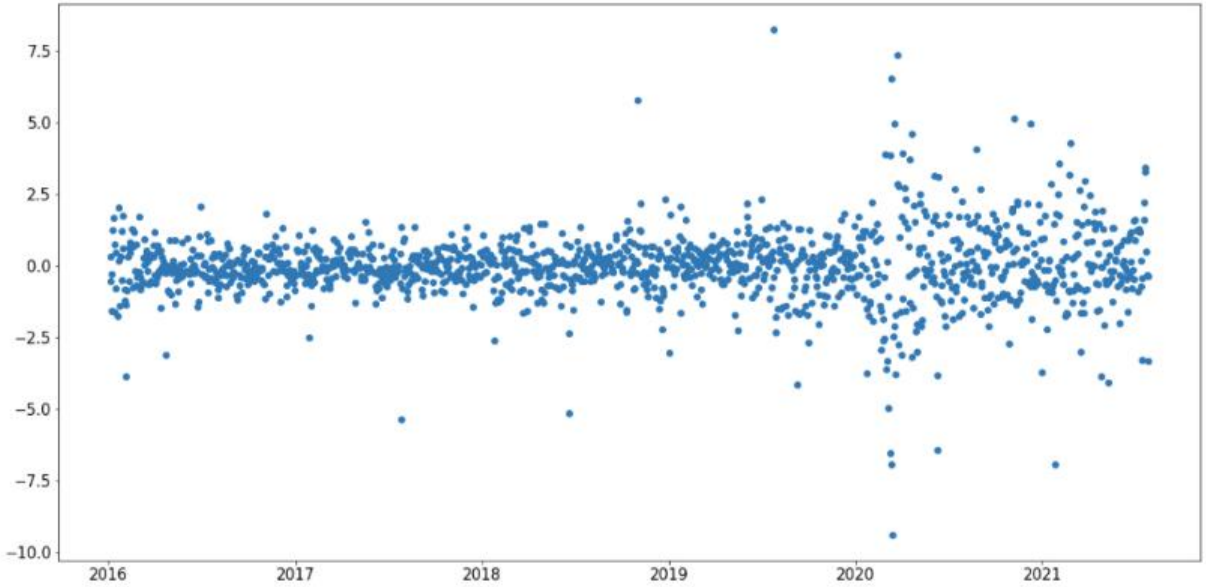




Results: Two Approaches

First approach: from 2016-2021

The equation of this model: $\hat{y}_t = c + \varphi_1 \hat{y}_{t-1} + \varphi_2 \hat{y}_{t-2} + \varphi_3 \hat{y}_{t-3} + \varphi_4 \hat{y}_{t-4} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$



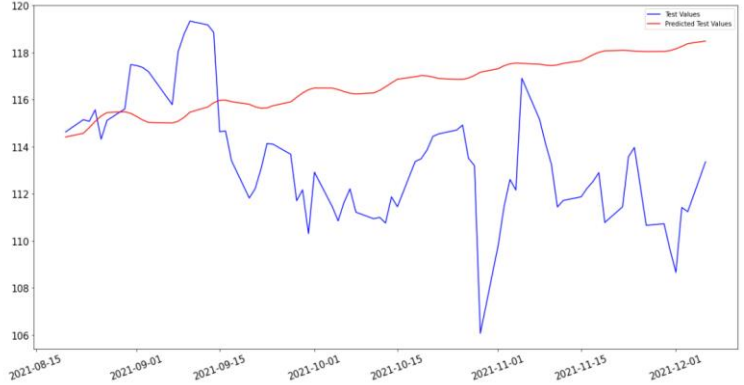
Results: Two Approaches

Second approach: removing the period from 2020-2021

ARIMA (2,1,3)

RMSE for the training set: 0.11975507325705646.
RMSE for the testing set: 4.6000970301995014.
And the mean of the testing set is: 113.373467.

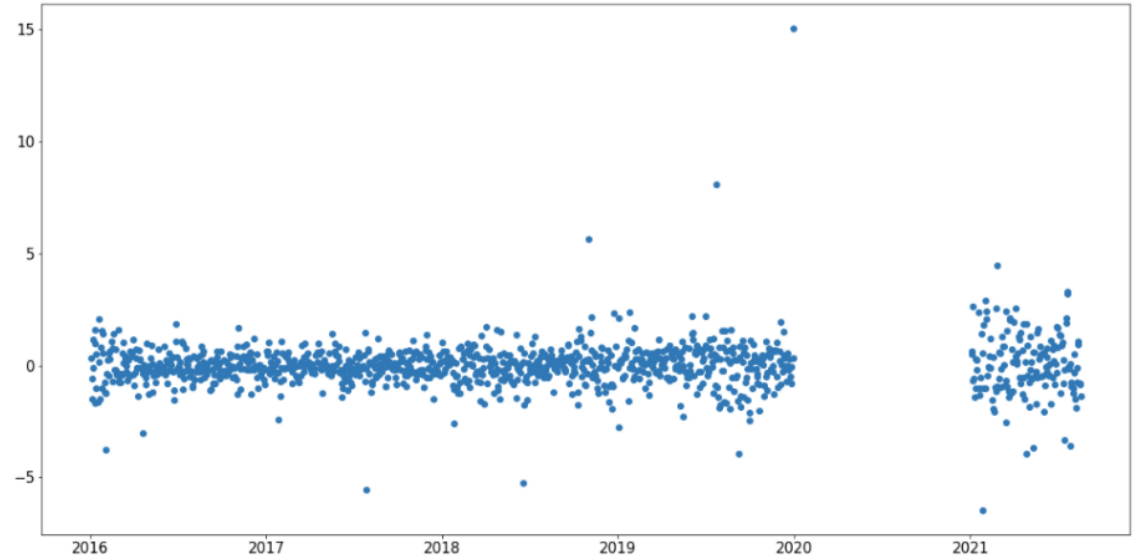
Value	Best order based on the RSME
(1,1,1)	0.05600379482858372 7.043124768971345
(2,1,2)	0.09343940612725077 7.041751447910158
(2,1,3)	0.11438360582233079 6.385317308497374
(2,1,4)	0.096605459925839 6.941709780979756
(5,1,5)	0.15172715699321743 6.884117701137903



Results: Two Approaches

Second approach: removing the period from 2020-2021

The equation of this model: $\hat{y}_t = c + \varphi_1 \hat{y}_{t-1} + \varphi_2 \hat{y}_{t-2} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t$



Conclusion

- The results are inconclusive
- The first approach, which included all the data from 2016 - 2021, included higher RMSE which is equal to 10.99.
- Second approach which excluded the data from 2020 -2021, included lower RMSE which is equal to 4.60.
- The second approach is better.

Thank you for listening
Any questions?