NUS FINTECH SOCIETY

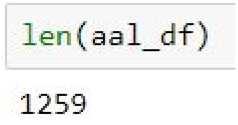
S&P 500 Stock Prediction with Time Series
AY18/19 Semester 2

Project Report

Prepared by:
Valary Lim (A0190343L)
Nicole Png (A0188425W)
Zheng Yi (A0190010A)
Ernest Chng (A0189158M)
Ivan Ho (A0167519U)

1. Descriptive Statistics from EDA

Total number of records (specific to “AAL”)



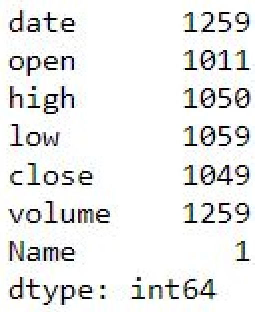
```
len(aal_df)
```

```
1259
```

# There are 1259 rows of data.

Unique values

Checking for distinct values



```
date      1259
open      1011
high      1050
low       1059
close     1049
volume    1259
Name         1
dtype: int64
```

# Date: There are no duplicated records (every value in the date column is unique).
# Open: There are 248 duplicated records.
# High: There are 209 duplicated records.
# Low: There are 200 duplicated records.
# Close: There are 210 duplicated records.
# Volume: There are no duplicated records (every value in the volume column is unique).
# Name: Only one unique record “AAL”

2. Process of Data Pre-processing and Analysis

Before we carry out our analysis on the dataset, we have to clean the data by removing any missing values present in the dataset. Instead of omitting data with missing values, we replaced these missing values with the mean values of each individual stock. This process is repeated for the “High” and “Open” columns. We also checked for duplications in the various stock data since duplications may cause biases in the analysis of the results.

For simplicity, we carried out the analysis on one of the 500 stocks provided. We chose to use “AAL” for our analysis. We plotted the line plot, as shown in the figure below, for ‘High’, ‘Open’, ‘Low’ and ‘Close’ prices across the years after aggregating the data.

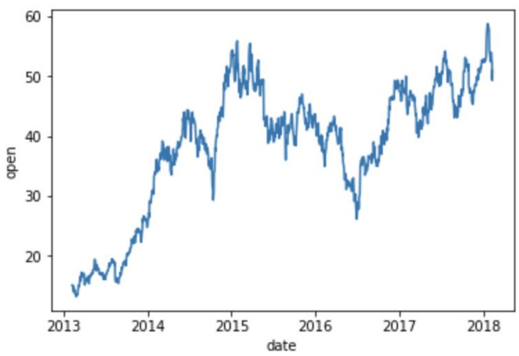


Fig: Plot of ‘Open’ values over 6 years

We also created a heatmap to find out whether each of the columns have a correlation with each other. From the heatmap below, since all the correlation values are 1, we conclude that they are highly correlated pairwise.

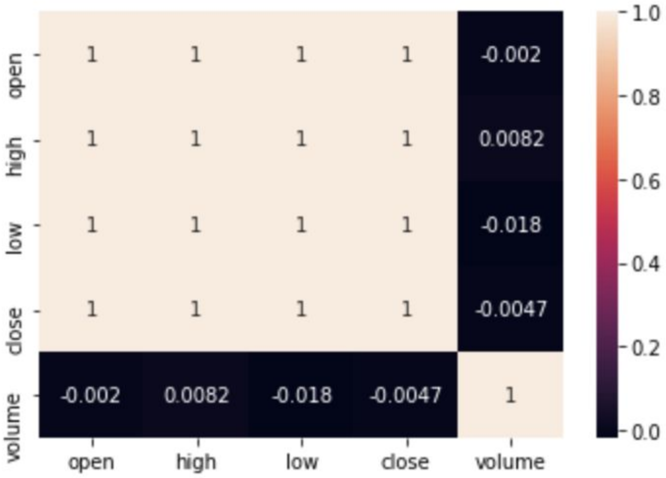


Fig: Heatmap of ‘Open’, ‘Close’, ‘Low’ and ‘High’ and ‘Volume’ of Stocks

This is further proven by the pairplot as shown below. Since the variables indicate high collinearity, it suggests that we could possibly make use of the non-stationary time series model in our time series analysis.

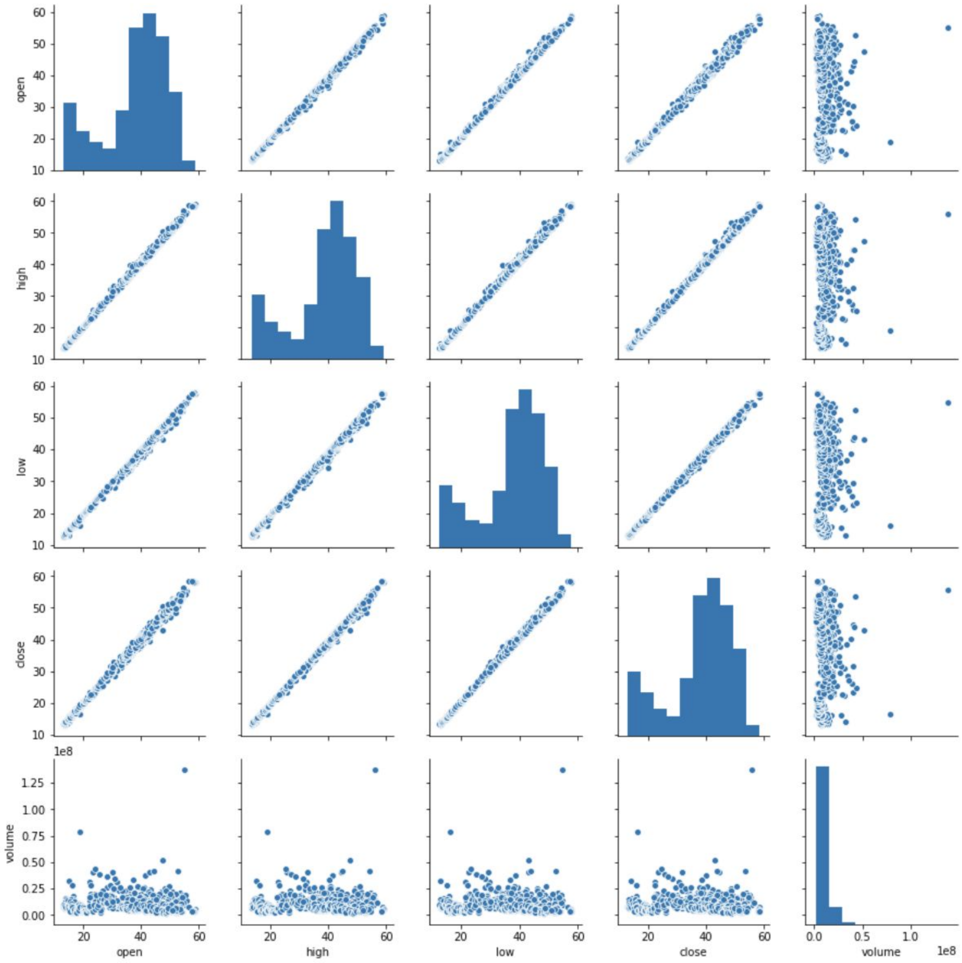


Fig: Pairplot of ‘Open’, ‘Close’, ‘Low’ and ‘High’ and ‘Volume’ of Stocks

3. Model Results

Before we can figure out how to forecast our time series, we have to figure out 3 things.

1. Identifying whether the current trend is stationary
2. Whether there is a need to stationarise the data
3. What model to fit and forecast the stock data

Firstly, in order to identify if the data stock is stationary, we will use 2 test.

1. Augmented Dicker Fuller(ADF) Test
2. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

Augmented Dickey Fuller (ADF) Test ¶

The ADF Test can be used to determine the presence of unit root in the series, and hence help us understand if the series is stationary or not. The null and alternate hypothesis of this test are:

- Null Hypothesis: The series has a unit root (value of a =1)
- Alternate Hypothesis: The series has no unit root.

If we fail to reject the null hypothesis, we can say that the series is non-stationary.

Refer to Notebook regarding outcomes of ADF Test.

Results of ADF Test

For **all** the data series, the test statistic > critical values, we do not reject the null hypothesis. We conclude that the series is non-stationary (or more specifically, non-difference stationary). Hence, we will need to stationarise the data to counter the changing variance before we can develop our time series model.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

The null and alternate hypothesis of this test are:

- Null Hypothesis: The process is trend stationary
- Alternate Hypothesis: The series has a unit root (series is not stationary).

Do note that the null and alternative hypothesis of the KPSS are the opposites of ADF. Here, we can say that the series is non-trend stationary if we reject the null hypothesis.

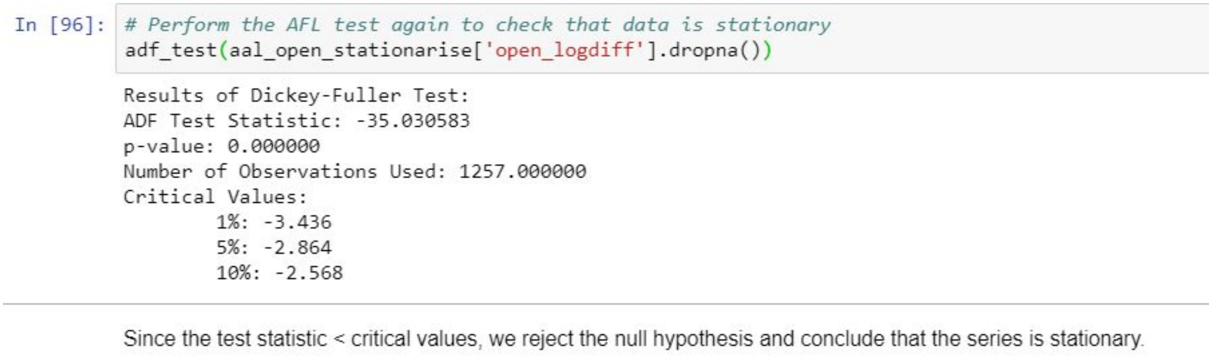
Refer to Notebook regarding outcomes of KPSS Test.

Results of KPSS Test

For **all** the data series, the test statistic > critical values, we do not reject the null hypothesis. We conclude that the series is trend-stationary.

Thus, since series data follows KPSS = stationary and ADF = not stationary. We conclude that the data is trend stationary, and we need to stationarise the series to make the series strict stationary.

Stationarising the series can be done through either taking the logarithm and square rooting the series to stabilise the variance. For our case, we have chosen the log the series.



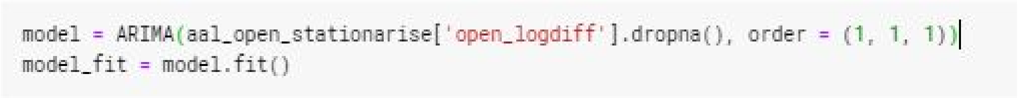
```
In [96]: # Perform the AFL test again to check that data is stationary
         adf_test(aal_open_stationarise['open_logdiff'].dropna())

         Results of Dickey-Fuller Test:
         ADF Test Statistic: -35.030583
         p-value: 0.000000
         Number of Observations Used: 1257.000000
         Critical Values:
                 1%: -3.436
                 5%: -2.864
                 10%: -2.568
```

Since the test statistic < critical values, we reject the null hypothesis and conclude that the series is stationary.

An ADF test on the “log-ed” data has resulted in our series showing a stationary trend.

When a time series is stationary, it can be easier to model. Statistical modeling methods assume or require the time series to be stationary to be effective.

Since the trend of the data is now considered stationary. We are now able to run an ARIMA model on it as ARIMA expects it data to be stationary.

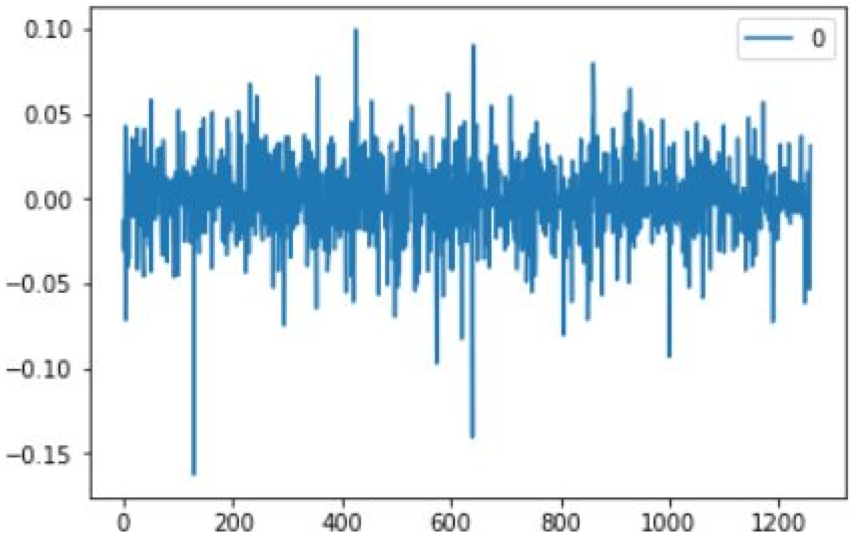


```
model = ARIMA(aal_open_stationarise['open_logdiff'].dropna(), order = (1, 1, 1))
model_fit = model.fit()
```

ARIMA model; specifically:

- **p**: Trend autoregression order.
- **d**: Trend difference order.
- **q**: Trend moving average order.

We have chosen (1, 1, 1) as one of the p-values of one variable is 0 which is less than 0.05, indicating that it is statistically significant. On the other hand, if the model was running on the parameters of (1,0,0), none of the variables have a p-value of less than 0.05, indicating that a model with parameters (1,1,1) would be a better fit.



Following which, we used the arima model to forecast our stock data.

4. Limitations / Other Considerations

Incompleteness of individual tests

A disadvantage of using solely the KPSS test is that it has a high rate of Type I errors, meaning it tends to reject the null hypothesis fairly often. If attempts are made to control these errors by having larger p-values, it will negatively affect the test's effectiveness. To deal with this high likelihood of Type I errors, we chose to complement it with the ADF test. If results from both tests suggest the the time series is stationary, it with very high possibility that it actually is.