S&P 500 STOCK DATA ANALYSIS

Presented by: -
Hiren Patel
Hitesh Saai Mananchery Panneerselvam
Nikhil Rajendra Mutha
Preetkumar Patel

Under the Guidance of
Prof. Ji Meng Loh

INTRODUCTION:

The Standard & Poor's 500, often abbreviated as the S&P 500, or just the S&P, is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ. The S&P 500 index components and their weightings are determined by S&P Dow Jones Indices. It differs from other U.S. stock market indices, such as the Dow Jones Industrial Average or the Nasdaq Composite index, because of its diverse constituency and weighting methodology. It is one of the most commonly followed equity indices, and many consider it one of the best representations of the U.S. stock market, and a bellwether for the U.S. economy. The National Bureau of Economic Research has classified common stocks as a leading indicator of business cycles.

In this report we are going to analyze stock data of S&P 500 from 2012 to 2018, Jan. There are 500 companies in the dataset containing different values like high, low, close, open and volume of stocks for the last five years. We are narrowing down the analysis just for the top five companies from banking & finance sector and trying to answer questions arising as we move along with the data analysis. We are also going to perform time-series analysis and ARIMA prediction method as we move forward with the analysis to predict the high stock value for Bank of America for next five years.

DATA ANALYSIS:

We found dataset containing stock price of S&P 500 with various factors like high, low, open, close, volume.
Before we start analyzing data, we are checking if the dataset has any missing value.

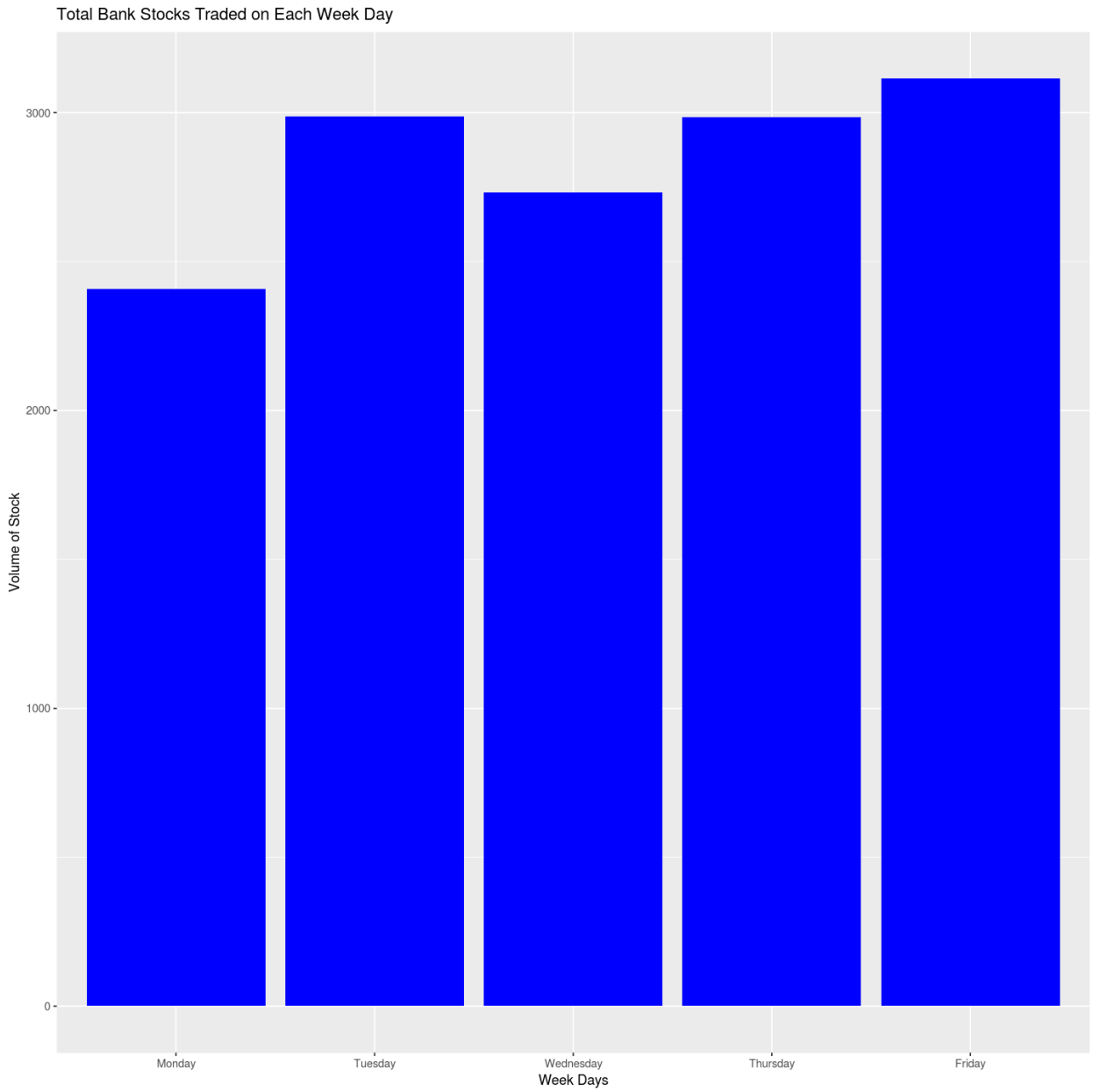
Date		Name			Open	High	Low	Close	Volume
2017-12-05	505	A	1259	Min.	1.62	1.69	1.50	1.59	0
2017-12-06	505	AAL	1259	1st Qut.	40.22	40.62	39.83	40.24	1070320
2017-12-07	505	AAP	1259	Median	62.59	63.15	62.02	62.62	2082094
2017-12-08	505	AAPL	1259	Mean	83.02	83.78	82.26	83.04	4321823
2017-12-11	505	ABBV	1259	3rd Qut.	94.37	95.18	93.54	94.41	4284509
2017-12-12	505	ABC	1259	Max.	2044	2067.99	2035.11	2049	618237630
Other	616010	611486		NA's	11	8	8	-	-

Now as we can see from above summary, there are some missing value in open, high and low column. So, we will try to remove them by replacing with zero to do an efficient analysis.

For further analysis we had data for 500 companies belonging to different segments of business, we decided to narrow down our analysis just for the banking and finance sector considering top 20 banks, by volume traded in five years, i.e. BAC, BK, BBT, COF, SEHW, C, CFG, CMA, FITB, GS, HBAN, JPM, KEY, MTB, MS, PBCT, PNC, RF, STT, STI.
Summary analysis for these banks are as follows:

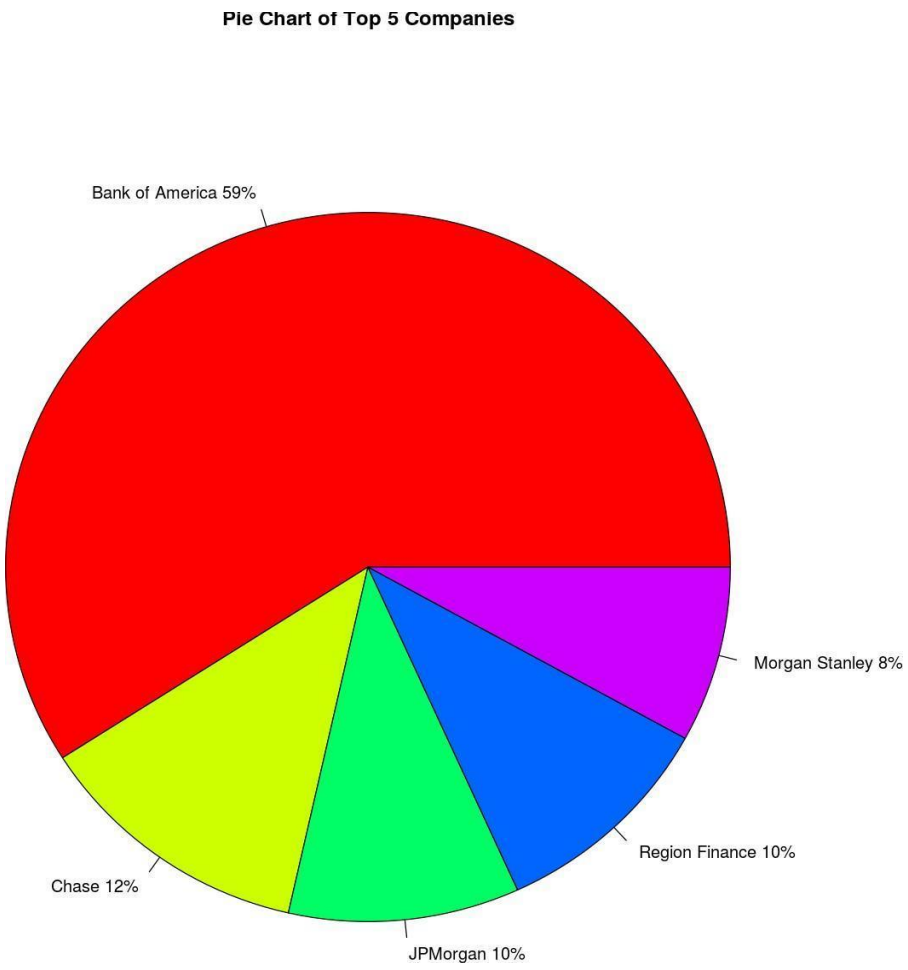
Date		Name			Open	High	Low	Close	Volume
2013-02-14	2	BAC	63	Min.	7.07	7.17	7.05	7.12	0.3574
2013-02-26	2	BBT	63	1st Qut.	18.36	18.50	18.23	18.39	2.5010
2013-02-27	2	BK	63	Median	40.55	40.88	40.19	40.44	4.9470
2013-03-04	2	C	63	Mean	53.72	54.16	53.25	53.72	12.0972
2013-03-06	2	CMA	63	3rd Qut.	70.81	71.74	70.34	71.09	11.8393
2013-03-07	2	CDF	63	Max.	269.04	273.79	268.81	272.48	231.4992
Other	1164	Other	798	NA's	-	-	-	-	-

Anyone with a little knowledge of stock market could say that people trade more stock volume on first day of week i.e. Monday and last day i.e. Friday. To confirm or say to check this assumption, we plot the volume of stock traded per day for the last five years of data. To do that analysis we first added the day for the corresponding dates in data file and we found result as follows.



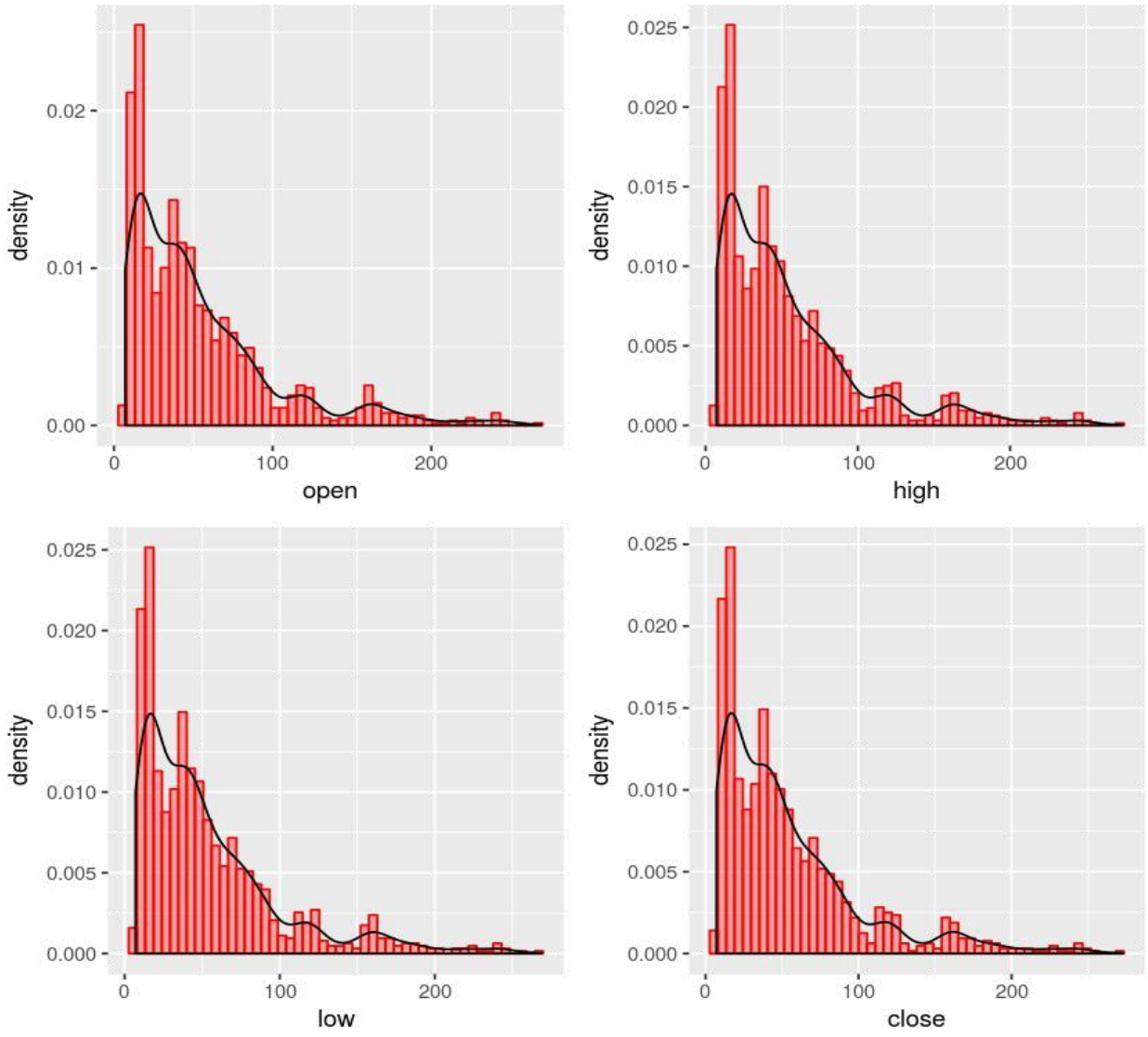
Well, the result isn't as we expected. It is literally the opposite for Monday, in-real volume of stocks traded on Monday is least of all the days. And the assumption stands true for Friday as it records highest volume traded during week for banking sector. Also, Volume traded on second day from start and end of the week i.e. on Tuesday and Thursday is almost same, which is surprising. And on middle of the week i.e. on Wednesday, it is low comparatively but higher than Monday.

We further check which banks top the list for highest volume traded in last five years and the percentage of it, we created a pie chart showing share of top five banks as follows. We only considered top five banks in this analysis.



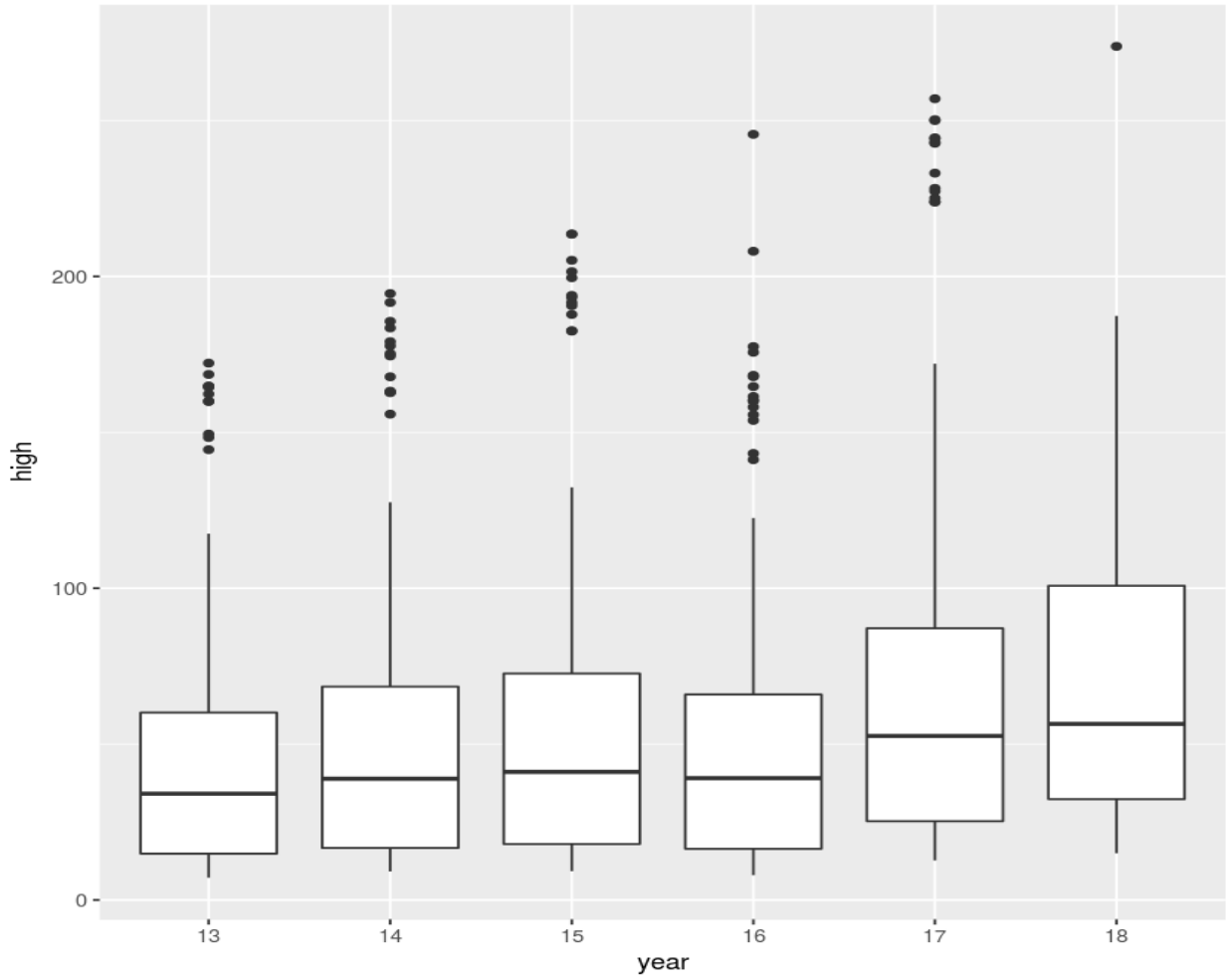
From the above pie chart, we can see that Bank of America tops the list by 59% of stock share among top five banks by very large margin. Other top four banks are Chase, JP Morgan, Region Finance and Morgan Stanley with 40% share combined.

Now, to just get the idea of variation of stock prices for open, close, high and low factors, we are plotting univariate distribution for the values of each factors as a factor on X-axis and density on Y-axis.
And as we can see below, the values of stock prices vary somewhere between $0 to $500 while density stay the same for all.



The most frequently used analysis in stock market is Time-Series analysis, to predict the future value of stock for investment planning. We try to do time series analysis for Bank of America by Autoregressive Integrated Moving Average (ARIMA) method.

Before we work on ARIMA model, we plot the boxplot for high value vs year to see how the performance of the Bank of America over the year is. The box plot below shows that high value of stock for BAC increased suddenly in year 2017 and 2018 that too with all time high recorded because of the company's policy and its steps to increase profit by cost reduction, continuous improvements in risk management, massive capital return to shareholders, it also not only recovered from the meltdown, as being the hardest-hit major US banks during the financial crisis, but transformed its business into a well-run, sustainable banking operation that continues to improve.

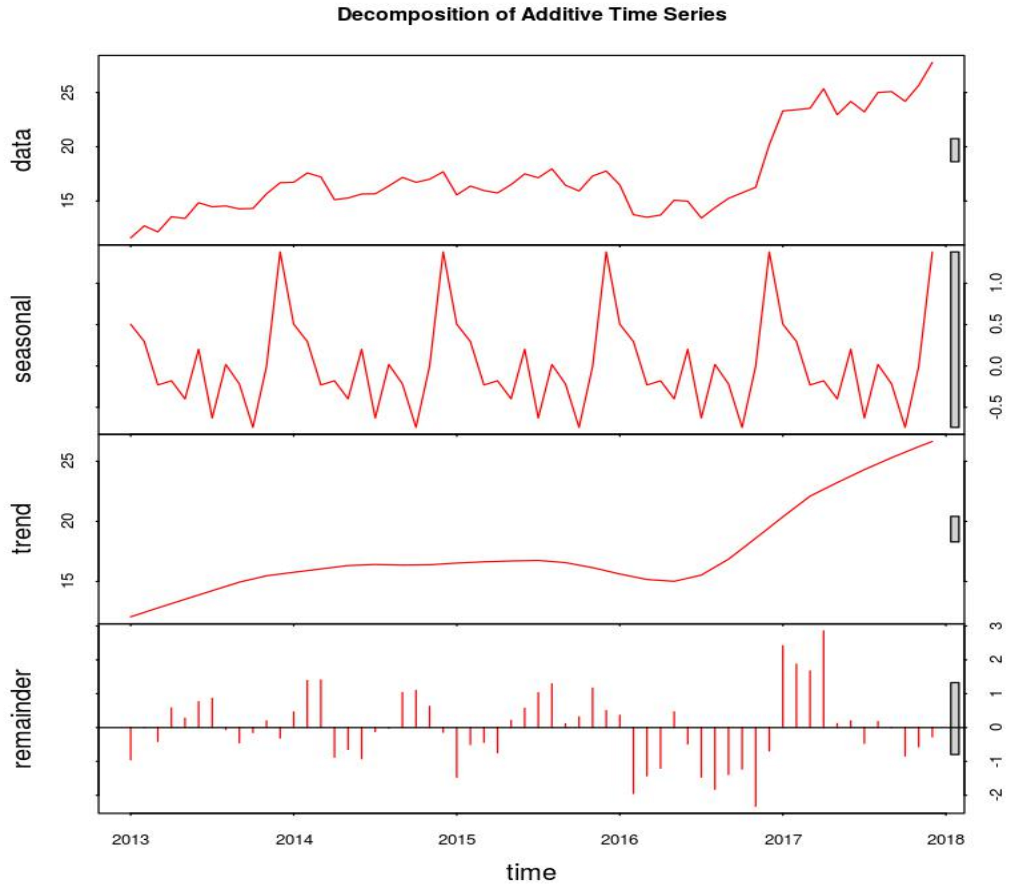


The high, low and each quadrant value can be seen from the summary below.

Date		Name			Open	High	Low	Close	Volume	Day	
2013-02-14	2	BAC	63	Min.	7.07	7.17	7.05	7.12	0.357	Mon	228
2013-02-26	2	BBT	63	1st Qut.	18.36	18.50	18.23	18.39	2.501	Tue	239
2013-02-27	2	BK	63	Median	40.55	40.88	40.19	40.44	4.947	Wed	237
2013-03-04	2	C	63	Mean	53.72	54.16	53.25	53.72	12.097	Thru	240
2013-03-06	2	CMA	63	3rd Qut.	70.81	71.74	70.34	71.09	11.839	Fri	232
2013-03-07	2	CDF	63	Max.	269.04	273.79	268.81	272.48	231.499	-	-
Other	1164	Other	798	NA's	-	-	-	-	-	-	-

Now we are plotting decomposition graph for high value of Bank of America. By decomposition, we mean breaking it down into trend, seasonal and irregular (noise) components.

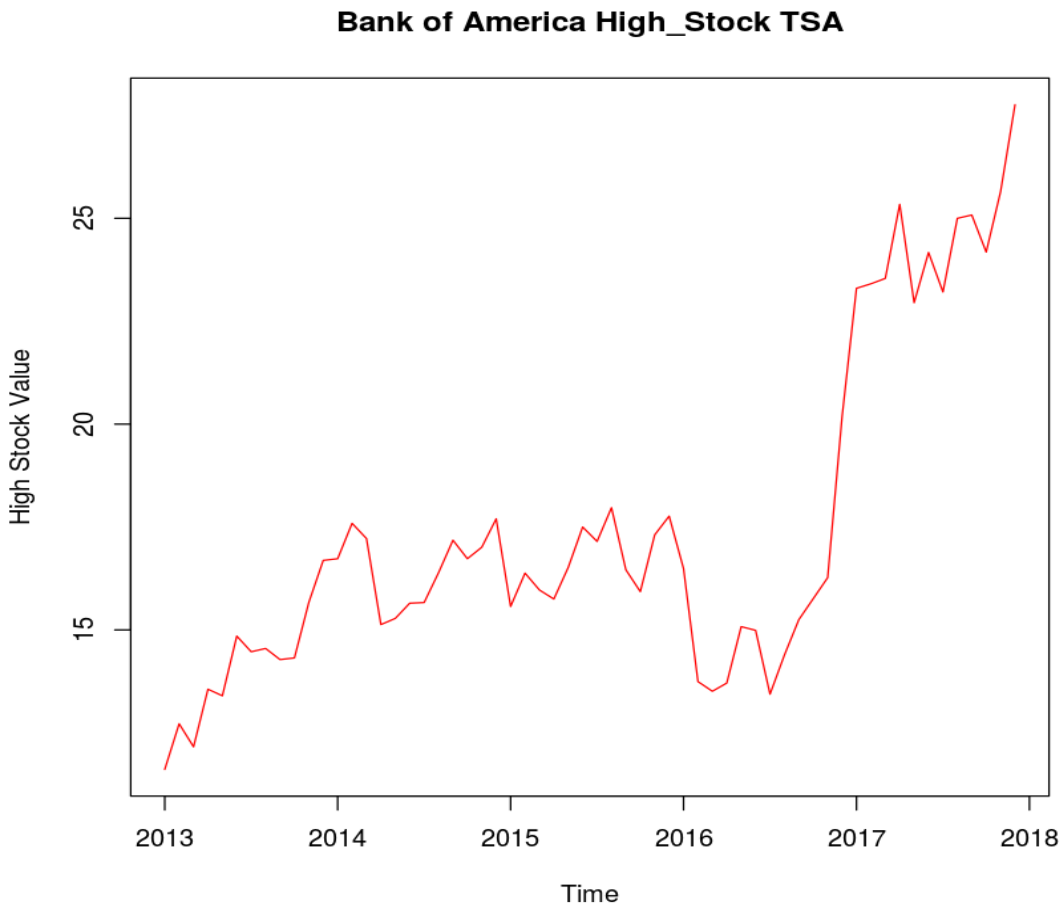
In the graph below, we can see seasonal, trend and observed (data) pattern for the high value over the last five years. We can see that the seasonal changes happen all over the year and to see the clear pattern in high value over the years, we plot high value vs time graph as follows.



Following is the time-series data for BAC over the years. We didn't take 2018 data into consideration because we don't have the data for whole year which could affect our prediction accuracy.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2013	11.61	12.72	12.16	13.56	13.40	14.85	14.47	14.55	14.28	14.32	15.67	16.69
2014	16.73	17.59	17.22	15.13	15.28	15.65	15.67	16.39	17.18	16.73	17.01	17.70
2015	15.57	16.38	15.97	15.75	16.52	17.50	17.15	17.97	16.64	15.93	17.31	17.77
2016	16.49	13.75	13.51	13.71	15.08	14.99	13.44	14.39	15.25	15.76	16.27	20.18
2017	23.30	23.41	23.54	25.34	22.95	24.17	23.21	25.00	25.08	24.18	25.65	27.76

The following is the graph of how high value of BAC fluctuated over the years.



Now we are going to do the predictive analysis for time series using “ARIMA” (Autoregressive Integrated Moving Average) model.

ARIMA Definition:

ARIMA is an acronym that stands for Autoregressive Integrated Moving Average. A popular and widely used statistical method for time series forecasting is the ARIMA model. It is a class of model that captures a suite of different standard temporal structures in time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.

- **AR**: ***Autoregression***. A model that uses the dependent relationship between an observation and some number of lagged observations.
- **I**: ***Integrated***. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) to make the time series stationary.
- **MA**: ***Moving Average***. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Parameters required to calculate the ARIMA model is defined as follows:

- **p**: The number of lag observations included in the model, also called the lag order.
- **d**: The number of times that the raw observations are differed, also called the degree of differencing.
- **q**: The size of the moving average window, also called the order of moving average.

To implement the ARIMA model to the time-series data, we need the dataset to be stationary. We are using the dickey-fuller test to see if the dataset is stationary or not. If the p-value is close to zero, then it means the dataset is stationary. Below is the obtained after performing the test.

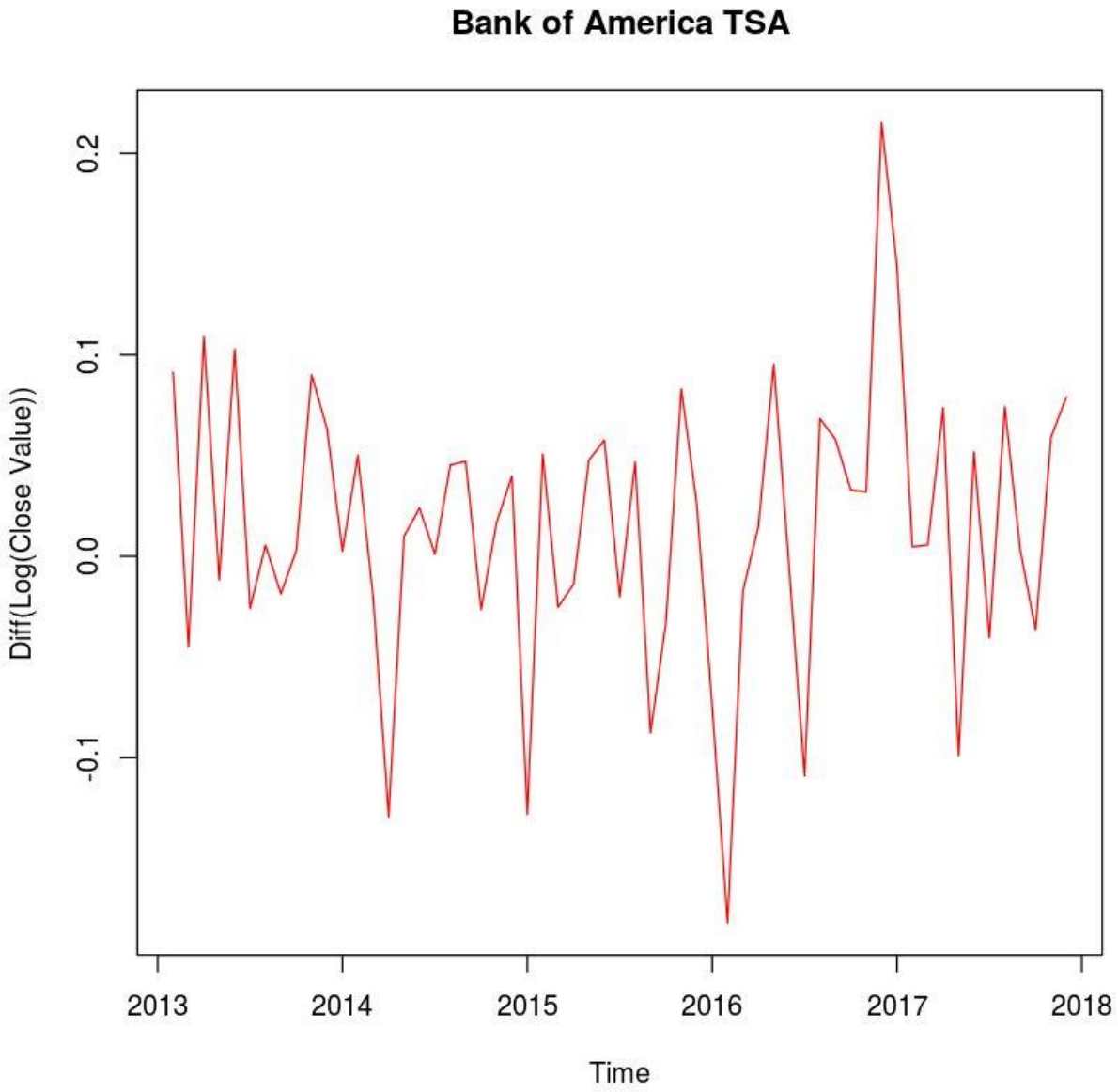
Augmented Dickey-Fuller Test:

Data: BAC
Dickey-Fuller = -1.1402, Lag order = 0, p-value = 0.9086
alternative hypothesis: stationary

From the above result we can see the p-value is far from zero, hence the dataset is not stationary. So, we must do the log transformation, so that the variance becomes equal over the period and differencing is a common solution used to stationaries the mean of the variable, therefore we will perform differencing using R function “diff”. Again, now using the Dicky-Fuller test to check whether the data is stationary or not.

Augmented Dickey-Fuller Test

Data: stat_c_ts
Dickey-Fuller = -7.2621, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary



From the above we can see that the time-series is stationary by having the mean and variance are constant.

Now that the p-value is closer to zero after log transform and differencing, so with one differencing, we get dataset to be stationary. Now the value of **d=1**, to proceed further with the ARIMA prediction method we need **p and q value.**

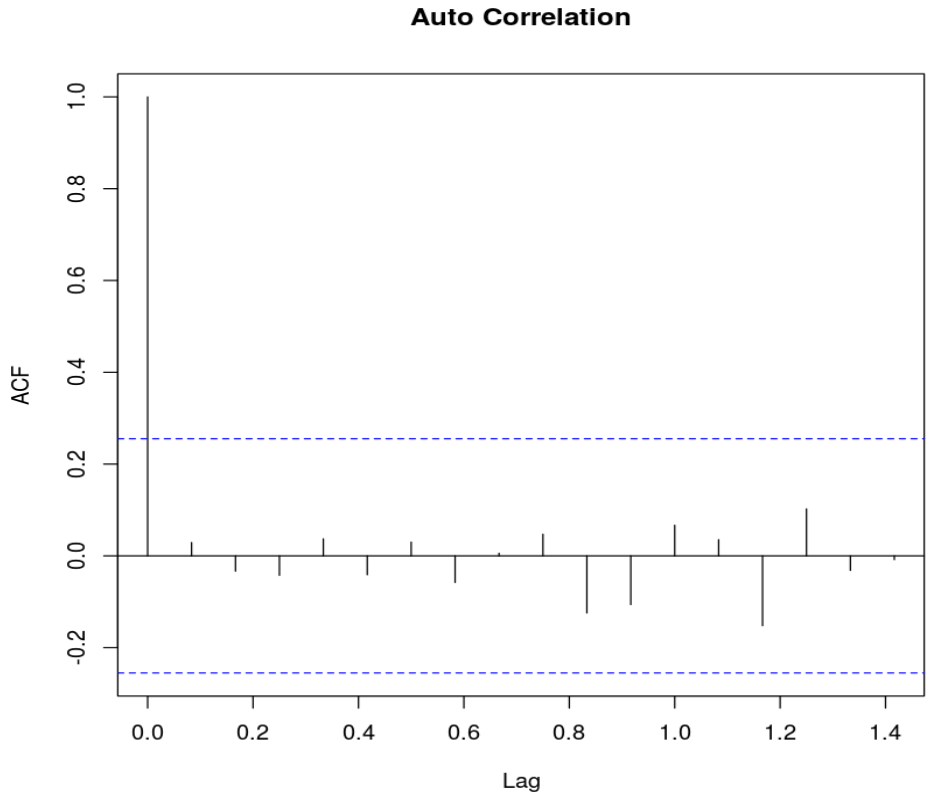
To obtain the **p** and **q** value we need to plot two analyses as follows:

Autocorrelation Function (q): The autocorrelation function is one of the tools used to find patterns in the data. Specifically, the autocorrelation function tells you the correlation between

points separated by various time lags. The notation is ACF (n=number of time periods between points) = correlation between points separated by n time periods.

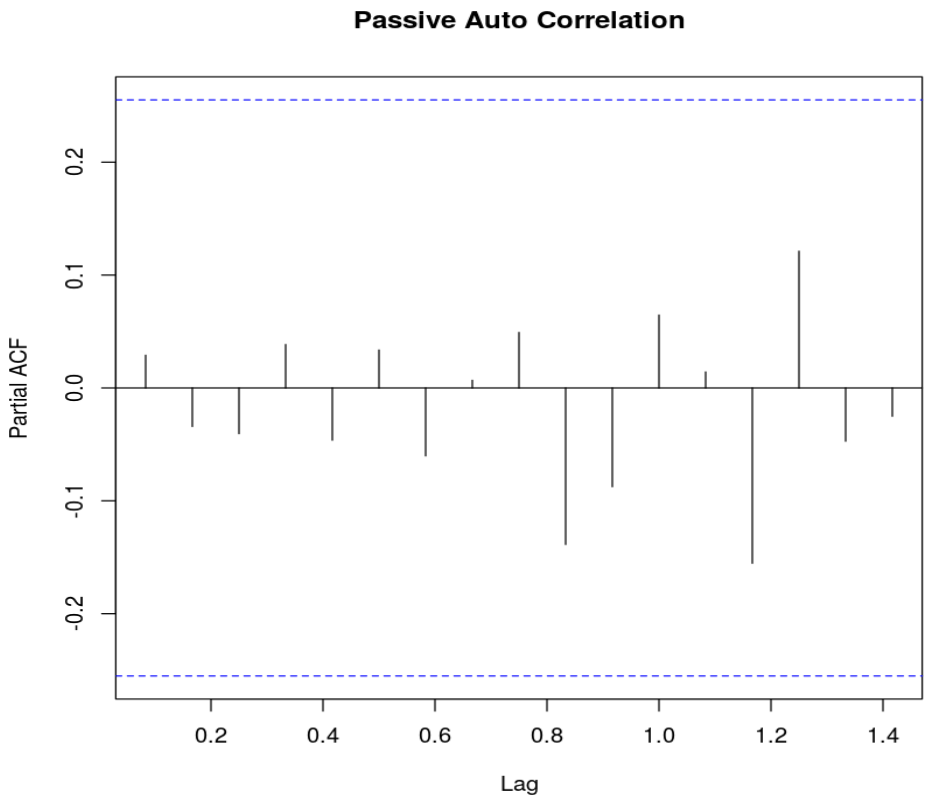
Passive Autocorrelation Function (p): In time-series analysis, the partial autocorrelation function (PACF) gives the partial correlation of a time-series with its own lagged values, controlling for the values of the time-series at all shorter lags. It contrasts with the autocorrelation function, which does not control for other lags.

First, we'll find the value of the q from the autocorrelation function graph:



From the above graph we can interpret the q-value based on considering the line before the first inverted line's count from zero. Hence, here the **q-value = 1**

Now to obtain the p-value we are going to plot the passive correlation graph as follows,



Similar process is used to find the p-value as how we discussed above for interpreting q-value from the graph. Hence, **p-value = 0.**

Now we have all the required parameters to fit and predict using the **ARIMA model**.

Below is the code used to fit the ARIMA model and to make the predictions for the next five years

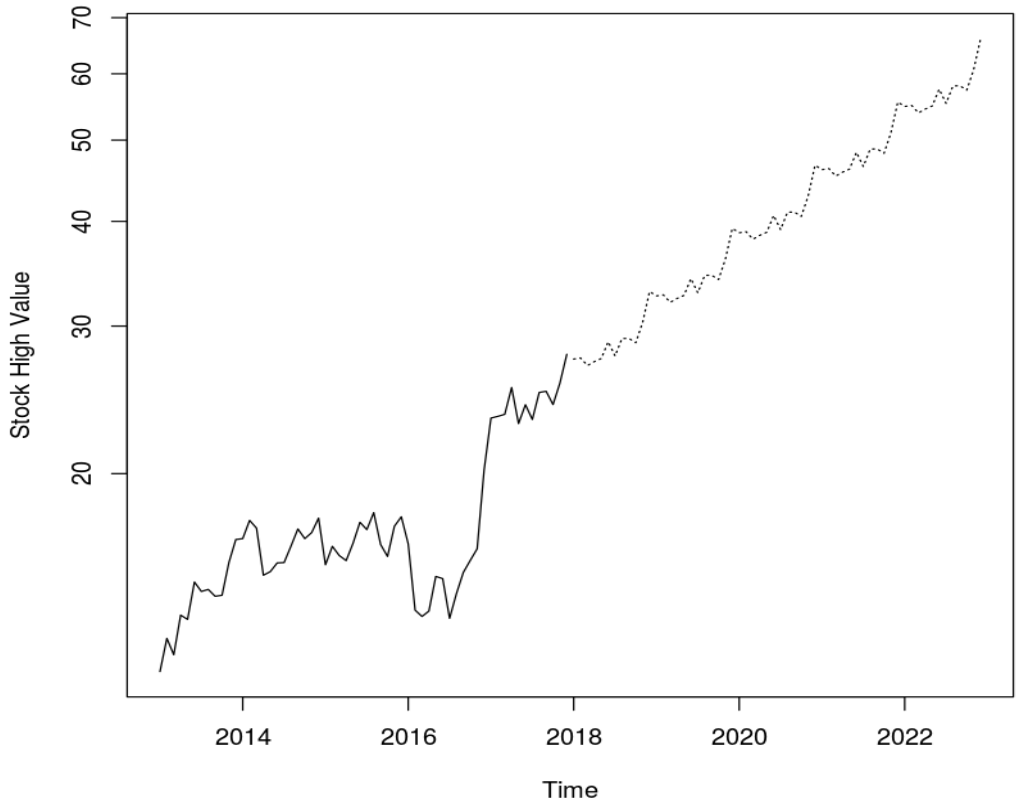
```
fit <- arima(log(c_ts), c(0,1,1), seasonal = list(order = c(0,1,1), period = 12))

pred <- predict(fit, n.ahead = 5*12)

pred1 <- 2.718^pred$pred

ts.plot(c_ts,pred1,log = "y",lty = c(1,3), ylab = "Stock High Value")
```

Resulting Prediction Graph for Bank of America High Stock Value for 5 Years:



Below are the predictions for Bank of America from 2018 to 2022.

year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018	27.40	27.48	26.92	27.21	27.44	28.73	27.63	28.99	29.00	28.66	30.31	32.99
2019	32.59	32.69	32.02	32.37	32.64	34.18	32.87	34.49	34.50	34.09	36.06	39.24
2020	38.76	38.88	38.08	38.50	38.82	40.65	39.10	41.02	41.04	40.55	42.89	46.67
2021	46.11	46.24	45.30	45.79	46.18	48.35	46.50	48.79	48.81	48.23	51.01	55.52
2022	54.84	55.01	53.88	54.47	54.93	57.51	55.32	58.04	58.06	57.36	60.67	66.03

ACCURACY TESTING:

In our original time-series dataset, we have data from 2013 to 2017, from this data we trained the dataset for the year duration from 2013 to 2016 using the ARIMA model and we predicted for the year 2017, and we did the cross validation to check for the accuracy.

Training Data:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2013	11.61	12.72	12.16	13.56	13.40	14.85	14.47	14.55	14.28	14.32	15.67	16.69
2014	16.73	17.59	17.22	15.13	15.28	15.65	15.67	16.39	17.18	16.73	17.01	17.70
2015	15.57	16.38	15.97	15.75	16.52	17.50	17.15	17.97	16.64	15.93	17.31	17.77
2016	16.49	13.75	13.51	13.71	15.08	14.99	13.44	14.39	15.25	15.76	16.27	20.18

Predicted Values:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2017	18.73	15.61	15.34	15.57	17.12	17.02	15.26	16.34	17.32	17.90	18.48	22.92

Original Values:

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2017	23.30	23.41	23.54	25.34	22.95	24.17	23.21	25.00	25.08	24.18	25.65	27.76

Now, comparing the predicted and the original value using the cross-validation method and we obtained an accuracy of 70.75%.

CONCLUSION:

We were trying to do the predictive analysis for time-series for the monthly high stock value for Bank of America for next five years i.e. from 2018 to 2022 which is one of the banking and finance organization listed in S&P 500. To do the predictions we split the data into two sets i.e. training data and testing data and fitted the training data in ARIMA model and checked our exactness by performing cross-validation. Hence, we were able to anticipate the month to month high stock value of Bank of America for the following five years with the accuracy of 70.75%. Therefore, from the prediction we were able to see that the high stock value for Bank of America increases in next five years. Due to some limitation of the dataset we were only able to pull 70.75% accuracy at its best.

Reference:

https://www.kaggle.com/camnugent/sandp500
https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
http://www.statosphere.com.au/check-time-series-stationary-r/