Time Series Analysis

Critical Thinking
Group 4:
DATA621 Final Project

DATE : May 22nd 2020

TEAM:
Rajwant Mishra
Priya Shaji
Debabrata Kabiraj
Isabel Ramesar
Sin Ying Wong
Fan Xu

Contents

ABSTRACT

- In this project, 5 years of live historical dataset (2015 to 2020) of stocks has been explored and analyzed with time series models.
- Time series models used are AR(Auto Regression) and MA(Moving Average) .
- Time series forecasting process has been executed using ACF() function and ACF plots.
- Lastly, we have evaluated all data models comparing its prediction score to analyze which model has performed better

KEYWORDS

- Lag : A “lag” is a fixed amount of passing time; One set of observations in a time series is plotted (lagged) against a second, later set of data. The kth lag is the time period that happened “k” time points before time i. The most commonly used lag is 1, called a first-order lag plot.
- Seasonality : In time series data, seasonality is the presence of variations that occur at specific regular intervals less than a year, such as weekly, monthly, or quarterly.

- Stationary : Stationary graphs are relevant to time series analysis, where we seek to understand the changes of a graph over time. With time series analysis, it is expected for data to vary over time, however, it is difficult to figure out the exact pattern by which a graph will change over time.
- Random Walk : A random walk, on the other hand, does not have this same tendency to centralize towards the mean due to the individual points along the walk being dependent on the previous points. This adds variance the more points are included in the walk, which can cause the path of the walk to deviate very far away from the mean.
- White Noise : With a white noise graph, we know that the distribution of the points will be normal and centered around zero with the same variance because the points are independent, so the tendency over time will be towards the mean
- AR (Auto regressive) : In this regression model, the response variable in the previous time period has become the predictor and the errors have our usual assumptions about errors in a simple linear regression model. The order of an autoregression is the number of immediately preceding values in the series that are used to predict the value at the present time. So, the preceding model is a first-order autoregression, written as AR(1).
- MA (Moving Average) :Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting. Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series.
- In time series analysis, the moving-average model (MA model), also known as moving-average process, is a common approach for modeling univariate time series. The moving-average model specifies that the output variable depends linearly on the current and various past values of a stochastic (imperfectly predictable) term.

INRTODUCTION

- We collected data from NYSE from Lat Five year using the API in R
- Also scrapped data of Sectors and Stock so that we can understand the trend by Sector from our data.
- Analyzed 5 year data by Sector and choose one of the stocks from Healthcare sector.
- We partitioned our data in data before 2020 and after 2020.
- Build AR and MA model on data before 2020 and predicted stock value for Year 2020
- We were able to check accuracy of the model by Model comparison and graph .

LITERATURE REVIEW

- One of the interesting works in stocks analysis is “using data mining with time series data in short-term stocks prediction” which explores methodologies similar to our project.
- Their approach uses data mining with time series data using examples related with short-term stocks prediction which is proved to be important to a better understanding of the field
- Specific challenges: developers focus on the issue of representing time series data in order to effectively and efficiently apply data mining.
- Another interesting issue was to find out if different time series or parts of time series have similar behavior.

- This issue can be approached through the use of similarity measures or indexing techniques.
- Over-fitting is a common problem across data mining applications.
- Achievements: A new concept, named as “median strings” is presented as a simple and at the same time powerful representation of time series data.
- Our work/investigation on our project is different from their’s by our approach used to solve a similar issue
- Time series models used by us are AR and MA model, as both these models perform with better predictions specifically when market is not stable, which hold true for current covid-19 scenario.
- Link to book we used to learn how other researchers have solved similar issue: Data mining with time series data

METHODOLOGY

- We have used Auto Regressive Integrated Moving Average Model with AR and MA model.
- Together with the autoregressive (AR) model, the moving-average model is a special case and key component of the more general ARMA and ARIMA models of time series, which have a more complicated stochastic structure.
- A time series is a sequence of measurements of the same variable(s) made over time. Usually the measurements are made at evenly spaced times - for example, monthly or yearly. Let us first consider the problem in which we have a y-variable measured as a time series. As an example, we might have y a measure of global temperature, with measurements observed each year. To emphasize that we have measured values over time, we use "t" as a subscript rather than the usual "i," i.e., yt means y measured in time period t.
- An autoregressive model is when a value from a time series is regressed on previous values from that same time series. for example, yt on yt−1: $y_t=\beta 0+\beta 1 y_{t-1}+\epsilon t$.
- In this regression model, the response variable in the previous time period has become the predictor and the errors have our usual assumptions about errors in a simple linear regression model. The order of an autoregression is the number of immediately preceding values in the series that are used to predict the value at the present time. So, the preceding model is a first-order autoregression, written as AR(1).

OVERVIEW

In this analysis we will analyze Stocks data and check how Covid-19 had impacted it from the beginning of the year 2020. We have collected data from exchange for last Five year starting from year 2015 to Apr 27, 2020. We would be using Moving Average Model and Auto Regressive model to analyze the time series data.

Column Name	Type	Sample Data	Description
X1	num	0	Sequence

begins_at	POSIXct,	4/28/2015	Date
open_price	num	24.9	Open Price
close_price	num	24.9	Close Price
high_price	num	24.9	High Price of stock on the given date.
low_price	num	24.9	Lowest price of the stock on the given date
volume	num	0	Volume of the shares trade on the give date
session	chr	reg	Type of the data session Regular /Extended
interpolated:	TRUE	TRUE	Not used
sname	chr	AA	Stock Name

Below is a short snippet of the data in the data set:

X1 <dbl>	begins_at <S3: POSIXct>	open_price <dbl>	close_price <dbl>	high_price <dbl>	low_price <dbl>	volume <dbl>	session <chr>	interpolated <lgl>	s <c
0	2015-04-28	24.8871	24.8871	24.8871	24.8871	0	reg	TRUE	A
1	2015-04-29	24.8871	24.8871	24.8871	24.8871	0	reg	TRUE	A
2	2015-04-30	24.8871	24.8871	24.8871	24.8871	0	reg	TRUE	A
3	2015-05-01	24.8871	24.8871	24.8871	24.8871	0	reg	TRUE	A
4	2015-05-04	24.8871	24.8871	24.8871	24.8871	0	reg	TRUE	A
5	2015-05-05	24.8871	24.8871	24.8871	24.8871	0	reg	TRUE	A

Sector information for the stocks:

```
 [1] "Basic Materials Sector"        "Communication Services Sector" "Consumer Cyclical Sector"
 [4] "Consumer Defensive Sector"     "Energy Sector"                 "Financial Services Sector"
 [7] "Healthcare Sector"             "Industrials Sector"            "Technology Sector"
[10] "Utilities Sector"
```

DATA EXPLORATION

We will be exploring the 500 stocks data by sectors and then we will choose one stocks to do further analysis.

Top 3 Stocks In Each Sector:

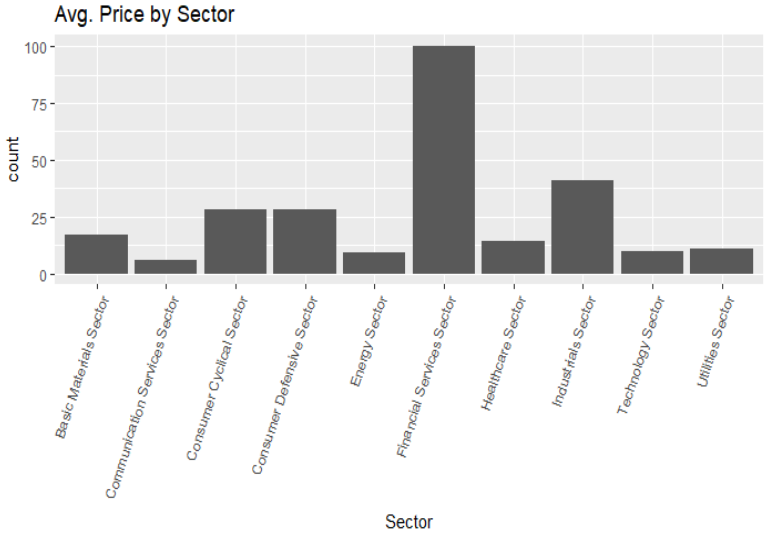
Index	Price	Sector	Firm Name
AZO	800.9067	Consumer Cyclical Sector	AutoZone, Inc
AZO	800.9067	Consumer Defensive Sector	AutoZone, Inc
BLK	419.6630	Financial Services Sector	BlackRock, Inc
BH	291.6055	Consumer Cyclical Sector	Biglari Holdings Inc
BH	291.6055	Consumer Defensive Sector	Biglari Holdings Inc
BA	245.1263	Industrials Sector	The Boeing Company
BIO	233.8155	Healthcare Sector	Bio-Rad Laboratories, Inc
AGN	211.6256	Healthcare Sector	Allergan plc
ADS	208.0295	Financial Services Sector	Alliance Data Systems Corporation
ANTM	206.0807	Healthcare Sector	Anthem, Inc

We will study the flow on some of the stocks from Health and Tech Sectors like:

ANTM Anthem, Inc
ANET Arista Networks, Inc
BA The Boeing Company

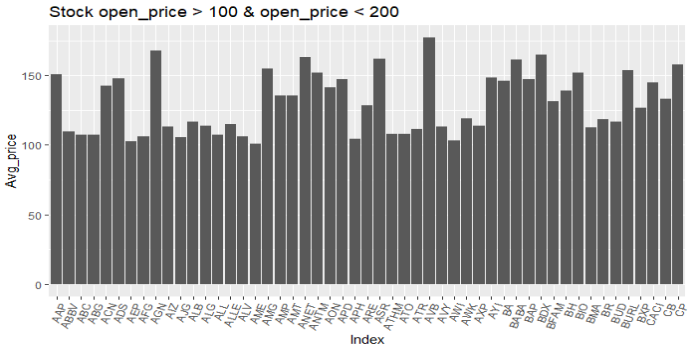
We have collected all the stocks from NYSE and corresponding sector info:

- Basic Materials Sector
- Communication Services Sector
- Consumer Cyclical Sector
- Consumer Defensive Sector
- Energy Sector
- Financial Services Sector
- Healthcare Sector
- Industrials Sector
- Technology Sector
- Utilities Sector

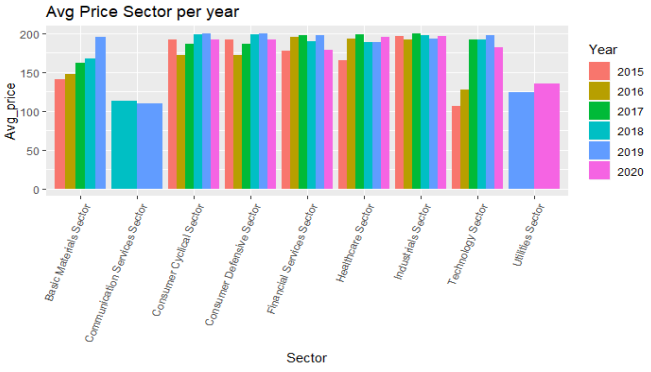


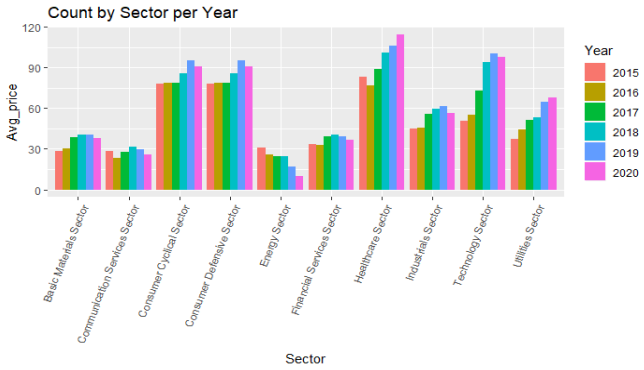
Let's looks at our data structure of the data :

Observations: 370,440
Variables: 10
$ X1 <*dbl*> 0, 1, 2, 3, 4, 5,
$ begins_at <*dttm*> 2015-04-28
$ open_price <*dbl*> 24.8871, 24.8871,
$ close_price <*dbl*> 24.8871, 24.8871,
$ high_price <*dbl*> 24.8871, 24.8871,
$ low_price <*dbl*> 24.8871, 24.8871,
$ volume <*dbl*> 0, 0, 0, 0, 0, 0, 0, 0
$ session <*chr*> "reg", "reg", "reg
$ interpolated <*lgl*> TRUE, TRUE
$ sname <*chr*> "AA", "AA",

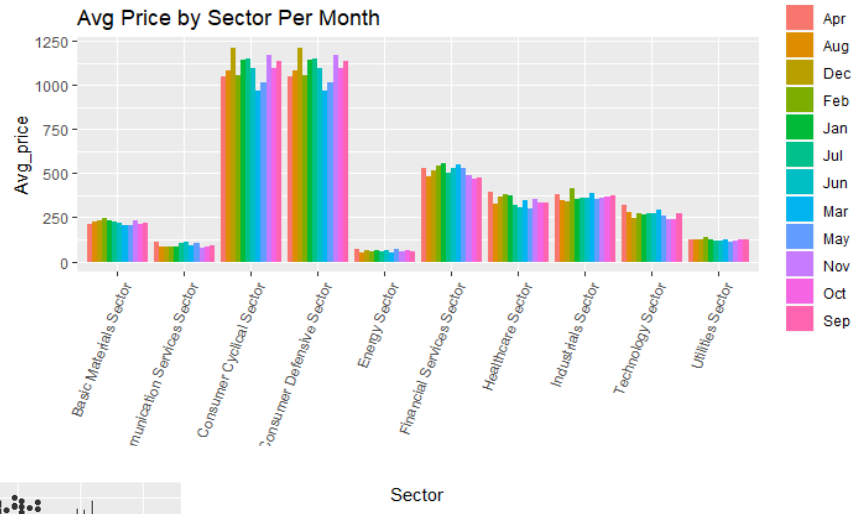


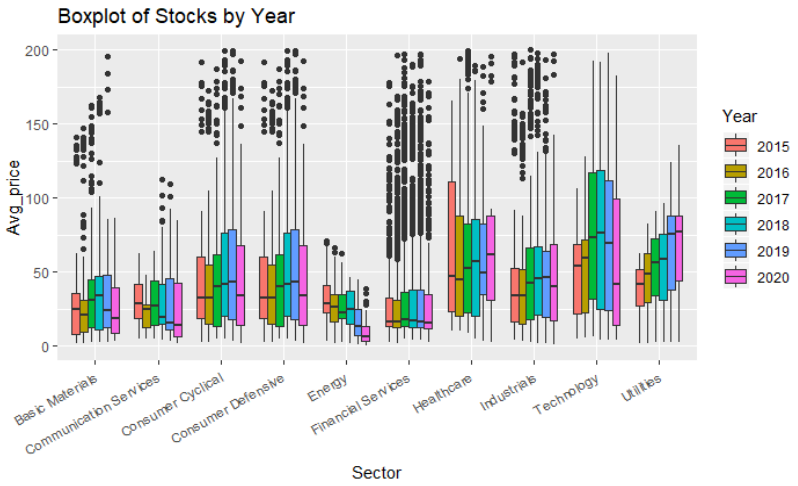
Summary The graph above tells us about yearly increase avg. Price in each sector.





Graph of the sector by Month and year, which shows some pattern. We will do some analysis to see how stocks from few of these industries fit with `AR(Auto Regression)` and `MA(Moving Average)` model. Analyzing average price of stocks yearly for each sector in the dataset.





From the box plot above we can analyze that mostly all the sectors in our dataset have some outliers throughout 5 years, except two sectors that are: `Technology` and `Utilities`.

Converting the data of stocks in wide format:

begins_at <S3: POSIXct>	AAP <dbl>	ADS <dbl>	AGN <dbl>	ANET <dbl>	ANTM <dbl>	APD <dbl>	ASR <dbl>	AYI <dbl>	AZO <dbl>
2015-04-28	145.00	303.00	281.90	64.82	150.15	139.4615	151.5279	166.54	691.68
2015-04-29	144.50	301.27	286.35	65.67	155.96	140.3583	150.7279	164.61	690.00
2015-04-30	144.43	299.54	287.27	64.56	151.83	141.7728	149.6180	168.95	682.63
2015-05-01	143.09	300.06	285.17	64.08	151.92	132.8975	145.1780	167.22	675.06
2015-05-04	145.25	299.77	290.19	64.24	153.45	136.7619	146.6580	170.07	683.00
2015-05-05	145.45	300.00	290.45	64.66	154.67	137.0023	143.9980	172.83	682.00

6 rows | 1-10 of 18 columns

Build Data

We will Fit an AR model to the following stocks, we would be using xts package in R to convert our data set to required time series data format.:

ANTM Anthem, Inc

ANET Arista Networks, Inc

BA The Boeing Company

Creating Time Series Object

```
# wide_data_Main$begins_at  <- as_datetime(wide_data_Main$begins_at)
stocks_ANTM <-xts(wide_data_Main$ANTM,  order.by=as.Date(wide_data_Main$begins_at))
stocks_ANET <- xts(wide_data_Main$ANET, order.by=as.Date(wide_data_Main$begins_at))
stocks_BA <- xts(wide_data_Main$BA,  order.by=as.Date(wide_data_Main$begins_at))
```

Checking the index of xts object:

With the commands head() and tail() we can see the first and last 6 lines of the base. There are 6 columns with: opening price, maximum and minimum prices, closing price, volume of transactions and adjusted price. Using the command summary() we verify the descriptive statistics of each price series and volume. The command str() returns the object structure. In this case, it's a xts object, a time series.

```
str(stocks_ANTM_MY)
## An 'xts' object on 2020-01-02/2020-04-27 containing:
##   Data: num [1:80, 1] 303 294 296 299 301 ...
##   Indexed by objects of class: [Date] TZ: UTC
##   xts Attributes:
##  NULL
```

Time Series Forecasting

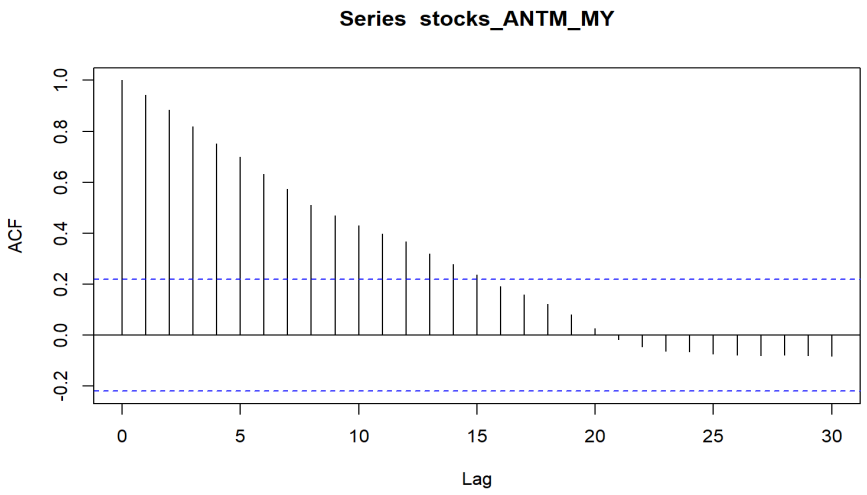
'stocks_ANTM' hold data of the Anthem stocks from April 28 2015 to April 27, 2020. Head and tail for the this stock can be seen below:

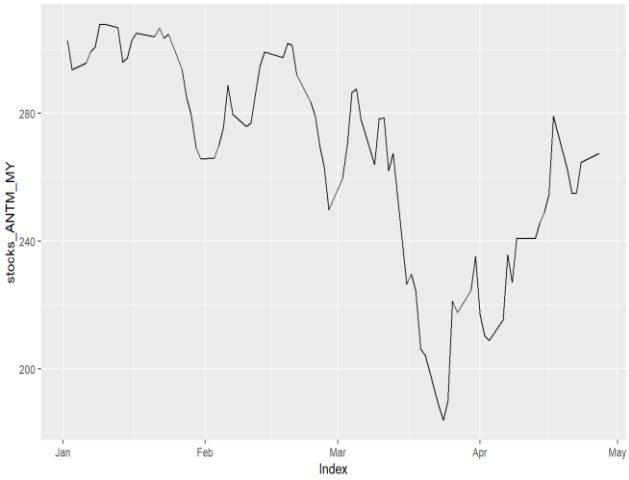
```
tail(stocks_ANTM)
##            [,1]
## 2020-04-20 262.48
## 2020-04-21 255.00
## 2020-04-22 255.09
## 2020-04-23 264.71
## 2020-04-24 265.48
## 2020-04-27 267.56
```

```
head(stocks_ANTM)
##            [,1]
## 2015-04-28 150.15
## 2015-04-29 155.96
## 2015-04-30 151.83
## 2015-05-01 151.92
## 2015-05-04 153.45
## 2015-05-05 154.67
```

```
summary(stocks_ANTM)
##      Index              stocks_ANTM
##  Min.   :2015-04-28   Min.   :117.0
##  1st Qu.:2016-07-26   1st Qu.:145.2
##  Median :2017-10-24   Median :195.9
##  Mean   :2017-10-25   Mean   :206.5
##  3rd Qu.:2019-01-26   3rd Qu.:262.8
##  Max.   :2020-04-27   Max.   :317.6
```

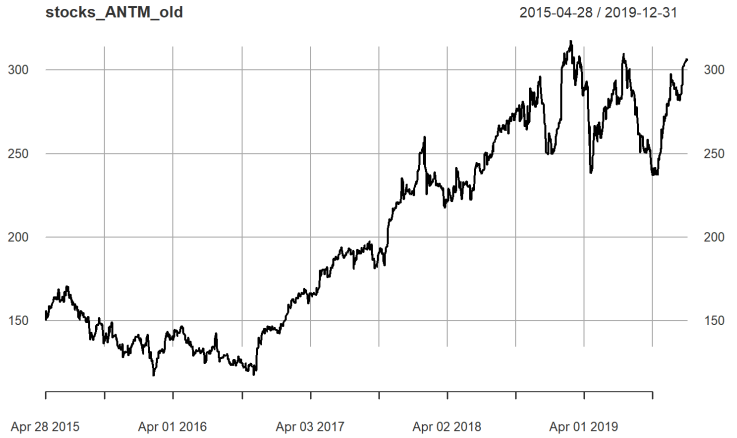
ACF plots below shows it's not a stationary, as it shows some seasonality in the data in latter part of data. This seasonality is due to COVID-19 impact on the stocks. If you compare the same plot for full ANTHEM data , its create that this stock prices follows some study random walk model and is non-stationary data.





Graph shows how ANTHEM Stock moved in the year 2020. From the trend, its very easy to see sharp decline in March and little recovery in April 2020.

White Noise

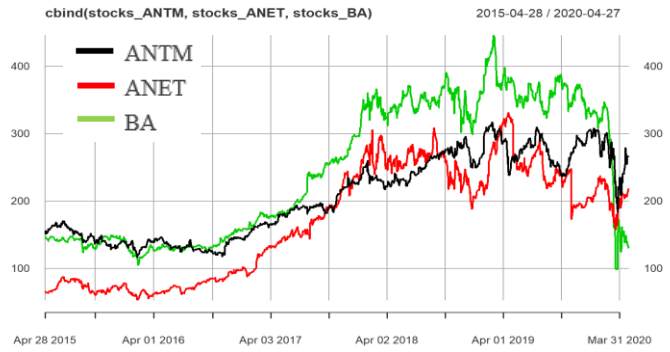


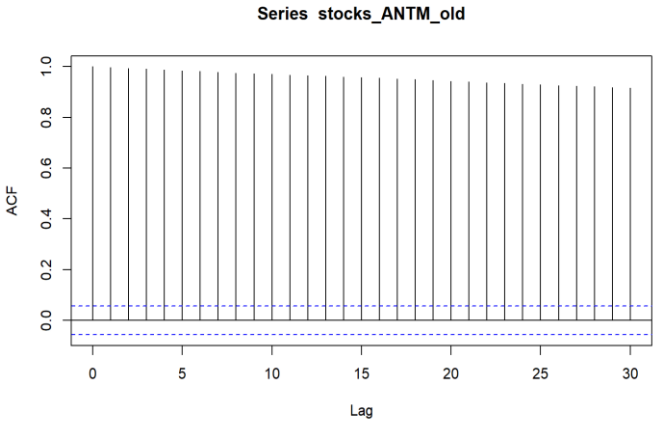
All the above plots doesn't show any white noise, The white noise (WN) and random walk (RW) models are very closely related. However, only the RW is always non-stationary. And we do see that our dataset is supporting random walk and its not stationary.
The ACF plots test if an individual lag autocorrelation is different than zero. An alternative approach is to use the Ljung-Box test, which tests whether any of a group of autocorrelations of a time series are different from zero.

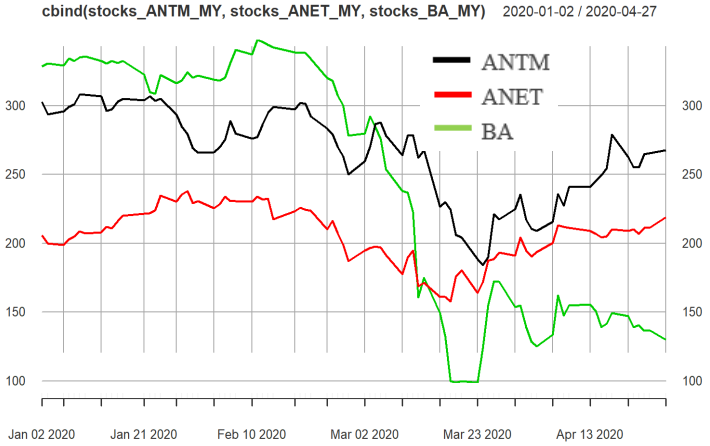
```
Box.test(wide_data_Main_20$ANTM, lag = 30, fitdf = 0, type = "Lj")
##
##  Box-Ljung test
##
## data:  wide_data_Main_20$ANTM
## X-squared = 480.01, df = 30, p-value < 2.2e-16

Box.test(wide_data_Main$ANTM, lag = 4, fitdf = 0, type = "Lj")
##
##  Box-Ljung test
##
## data:  wide_data_Main$ANTM
## X-squared = 4976, df = 4, p-value < 2.2e-16
```

It tests the "overall randomness" based on a number of lags. If the result is a small p-value than it indicates the data are probably not white noise. For 2020 Data we will see if it's while noise or not:
Here, we perform a Ljung-Box test on the first 24 lag autocorrelations. The resulting p-value is significant at $p < .001$, so this supports our ACF plot consideration above where we stated it's likely this is not purely white noise and that some time series information exists in this data.

Autocorrelation



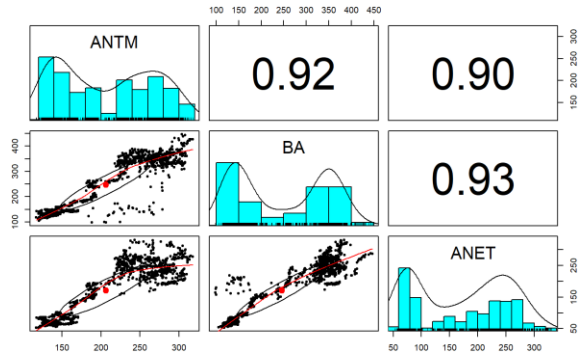




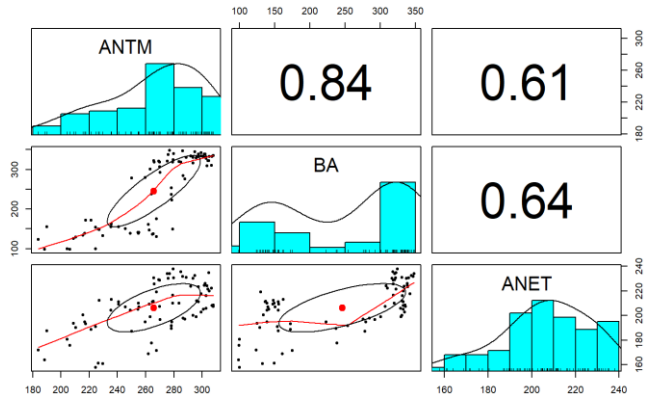
These plots suggest that these slots the stocks improved from their position from mid of 2016 though 2018, and then it remained constant in progress until Late 2019 and early 2020.
The trend is the long-term increase or decrease in the data. There is an increasing trend in the cement data. the seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. The daily data of the stocks_ANTM doesn't show any seasonality in the graph.
the cycle occurs when the data exhibit rises and falls that are not of a fixed period. These fluctuations are usually due to economic conditions and are often related to the "business cycle". We can see a few cycles in our in stocks_ANTM data from 2015 to 2018 and then in 2020 we have sudden drop due to covid 19.

Correlation in the stocks is very high and COR function also shows the same.

Correlation on Full Stock data from 2015- 2020.



Correlation only for 2020 data:



BUILDING MODEL

Fitting Data

For a given time series x we can fit the autoregressive (AR) model using the arima() command and setting order equal to c(1, 0, 0). Note for reference that an AR model is an ARIMA(1, 0, 0) model.

```
# Fit with Full Data

AR_ANTM <- arima(stocks_ANTM, order  = c(1,0,0))
MA_ANTM <- arima(stocks_ANTM, order  = c(0,0,1))
AR_ANTM_fit <- as.ts(stocks_ANTM) - resid(AR_ANTM)
MA_ANTM_fit <- as.ts(stocks_ANTM) - resid(MA_ANTM)
```

Fitting Full data :

Fitting Anthem full year data from 2015 to 2020 using AR Model, and MA Model.

AR Model

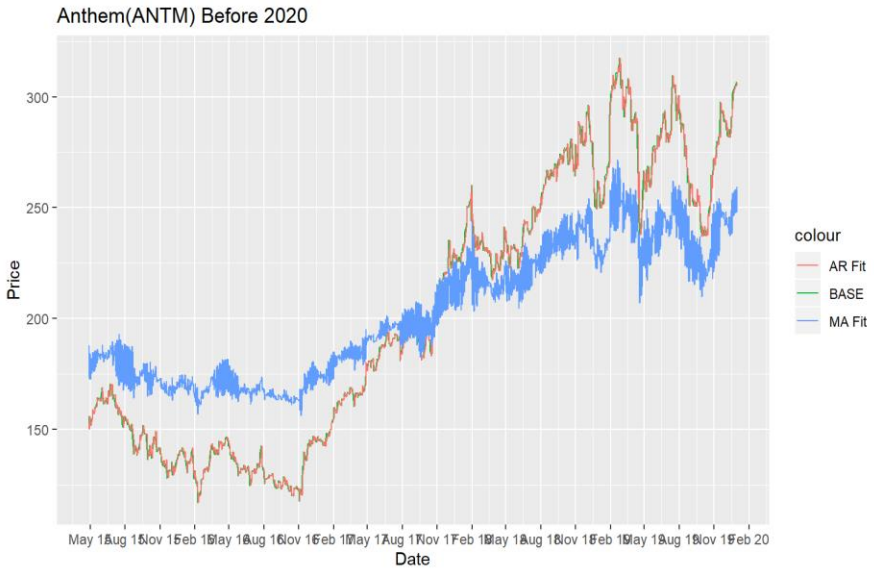
```
summary(AR_ANTM)

##
## Call:
## arima(x = stocks_ANTM, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.9978   222.5894
## s.e.  0.0018    45.1308
##
## sigma^2 estimated as 16.88:  log likelihood = -3568.12,
aic = 7142.25
##
## Training set error measures:
##                     ME     RMSE      MAE        MPE    MAPE
MASE
## Training set 0.0542719 4.108244 2.529343 -0.01210249
1.19899 1.002039
##                    ACF1
## Training set -0.0209907
```

MA Model

```
summary(MA_ANTM)
##
## Call:
## arima(x = stocks_ANTM, order = c(0, 0, 1))
##
## Coefficients:
##          ma1  intercept
##       0.9678   206.4969
## s.e.  0.0058     1.7300
##
## sigma^2 estimated as 973.9:  log likelihood = -
6119.59,  aic = 12245.19
##
## Training set error measures:
##                     ME     RMSE      MAE       MPE
MAPE     MASE      ACF1
## Training set 0.01566924 31.20706 27.96412 -
4.741534 15.01558 11.07842 0.9108069
```

Fit with Data Before 2020

Now we will use the data Until December 2019 and see how that fits with the data off AR and MA model. We can very clearly see that AR model is doing better and is very close to the base line, whereas MA model is not staying close to the actual data.

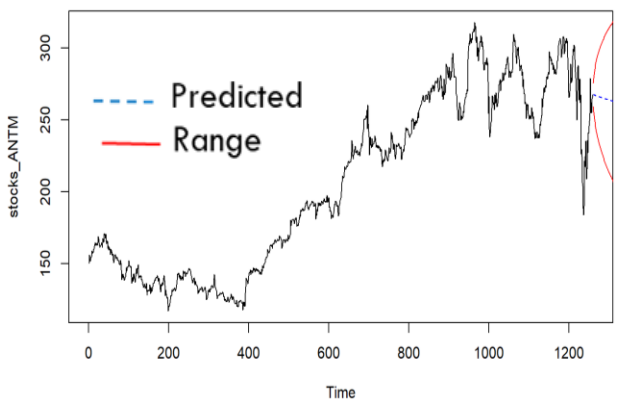


Predicting Time Series data

We will evaluate all the data models and see its prediction using both the models with Current Years data.

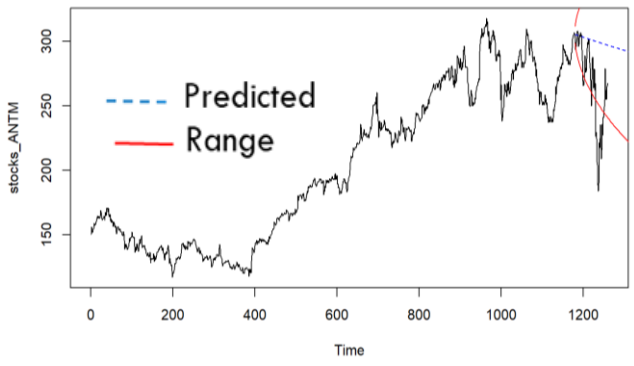
We can use predict command to predict the future days price by using n.ahead command. Here AR(1) model is the so-called "random walk" model (without drift): it assumes that, from one period to the next, the original time series merely takes a random "step" away from its last recorded

Using Full Data till -Apr 2020



Below graph shows how the price would be if we use AR(1) model to predict. We are using full data set to predict future price after April 27 2020.

Using Data Fit Till Dec 2019



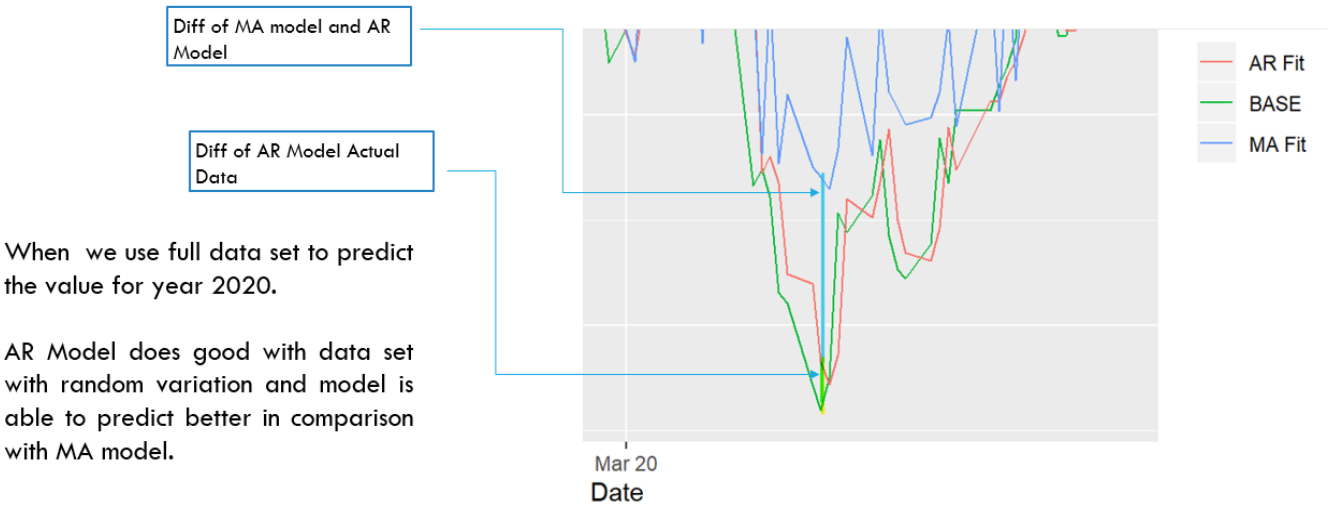
Now let use AR(1) model fit from data of only dec 2019, and try to predict 2020 stock price. Predicting next 300 days stock prices for the Anthem stock , based on AR(1) and MA(1) model.

EXPERIMENTATION AND RESULTS

Check Model

If we compare the data and results from AR and MA Model , we noted that:

- MA model seems to be doing better with predication when we used data till Dec 2019.
- For instance from the figure in right we can see very clearly a sharp drop in price, due to COVID-19 Pandemic. Lets see how these two Model predicted if we go by full data:
- When we use full data set to predict the value for year 2020.
- AR Model does good with data set with random variation and model can predict better in comparison with MA model.



Compare the Model

Our Model comparison shows that AR(1) model is predicting the price correctly compare with MA model.

Data Group	Model	df	AIC	BIC
Ful Data	AR_ANTM	3	7142.246	7157.66
	MA_ANTM	3	12245.19	12260.6
Till Jan- Apr 2020	AR_ANTM_MY	3	608.5171	615.6632
	MA_ANTM_MY	3	707.9615	715.1076
Til Dec 2019	AR_ANTM_old	3	6151.07	6166.287
	MA_ANTM_old	3	11435.55	11450.77

DISCUSSION AND CONCLUSIONS

Based upon our underacting of this time series analysis we noted that:

- Different model can be used to better predict same set of time series
- AR model is would always perform better for few predictions if market is not stable

- MA model may give better predication when market is very unstable
- Training and testing in Time series data depends on portioning data by date, Random selection of such data may not be accurate choice to better check the efficiency of the model.

REFERENCES

- Data Camp R cheat-sheet
- Introduction to Stock Analysis
- R for Data Science cheat-sheet
- A little book of R for Time Series
- Applied Time Series Analysis for Fisheries and Environmental Sciences
- Autoregressive Models
- Moving-average model

APPENDIX

Detail Code Base
Github Link