**Group Number: Title of your Projects**

For example: Group 5: Build a recommender systems using collaborative filtering

|  |  |  |  |
| --- | --- | --- | --- |
| First Name | Last Name | Monday or Tuesday class | Share project with ITMD 525? (Y or N) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

[**1. Introduction** 2](#_Toc498879580)

[**2. Data** 2](#_Toc498879581)

[**3. Problems to be Solved 2**](#_Toc498879582)

[**4. Data Processing** 2](#_Toc498879583)

[**5. Methods and Process** 2](#_Toc498879584)

[**6. Evaluations and Results** 2](#_Toc498879585)

[6.1. Evaluation Methods 2](#_Toc498879586)

[6.2. Results and Findings 2](#_Toc498879587)

[**7. Conclusions and Future Work** 2](#_Toc498879588)

[7.1. Conclusions 2](#_Toc498879589)

[7.2. Limitations 2](#_Toc498879590)

[7.3. Potential Improvements or Future Work 2](#_Toc498879591)

# **1. Introduction**

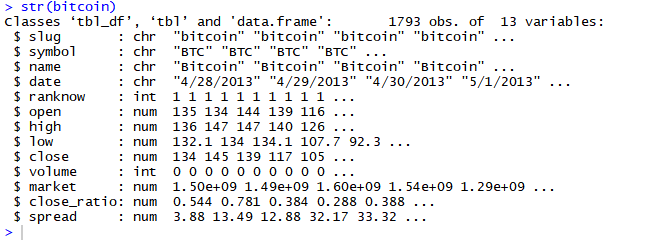
Blockchain is a new phenomenon which is going to change many aspects of modern societies. Most popular application of blockchain is digital crypto currencies. Digital cryptocurrencies are one of the hottest topics in data science, economics, finance, and technology. Currently, Bitcoin is the most famous digital cryptocurrency. So, exploring and discovering the relations and finding the patterns between data which are produced by selling and buying the Bitcoin can be valuable. In this project, we are going to find a prediction model for the Bitcoin price. We have the previous data of Bitcoin market and try to use Time-Series technique to predict the future price in the market.

# **2. Data**

Data Set which is used for this project is downloaded from Kaggle website.

However, it originally is collected from https://coinmarketcap.com/ and is accessible for Kaggle website’s users. As mentioned earlier, the goal for this project is to create a Time-Series model to forecast Bitcoin’s price in future.

There are eight variables in the dataset which are:



The following are descriptions of each variables:

$ **open**: The $ amount in US Dollars that the day started at

$ **high**: The highest $ amount it got to in US dollars that day

$ **low**: The lowest $ amount it got to in US dollars that day

$ **close**: The $ amount in US dollars that the day finished at

$ **volume**: The $ value in US dollars of how many were exchanged that day

$ **market**: The total amount of market capital (combined worth) in US dollars

$ **Close**\_ratio = The daily close rate, min-maxed with the high and low values for the day.

$ **Spread** = The $USD difference between the high and low values for the day.

Moreover, data is collected from 28/04/2013 to 25/03/2018 and it contains 1793 records.

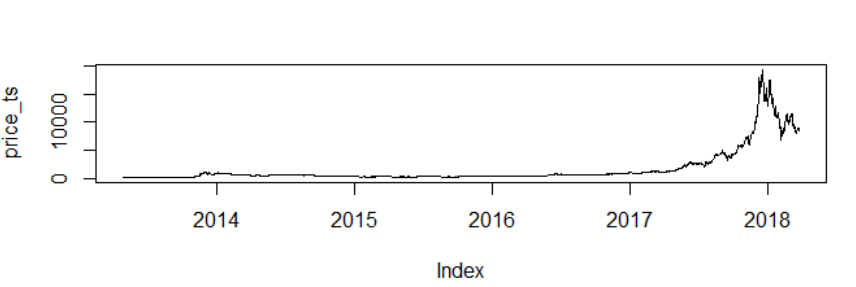
Among all available variable, close price is chosen as independent variable.

# **3. Problems to be solved**

Before making any time-series model we have to make sure that time-series data is stationary and serial dependence.

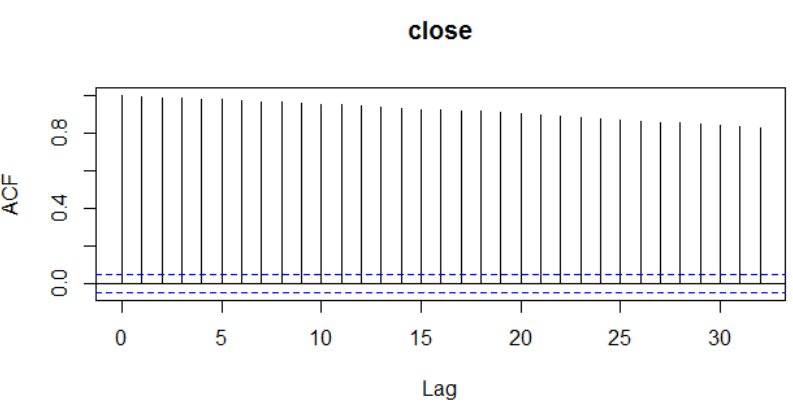
**Stationary:**

To do so, for stationary test linear time series is plotted.



It is cleared that the plot is not stationary. Both mean and variance is changed over time.

Also we must check serial dependency. ACF plot is needed to figure out whether data is stationary or not.



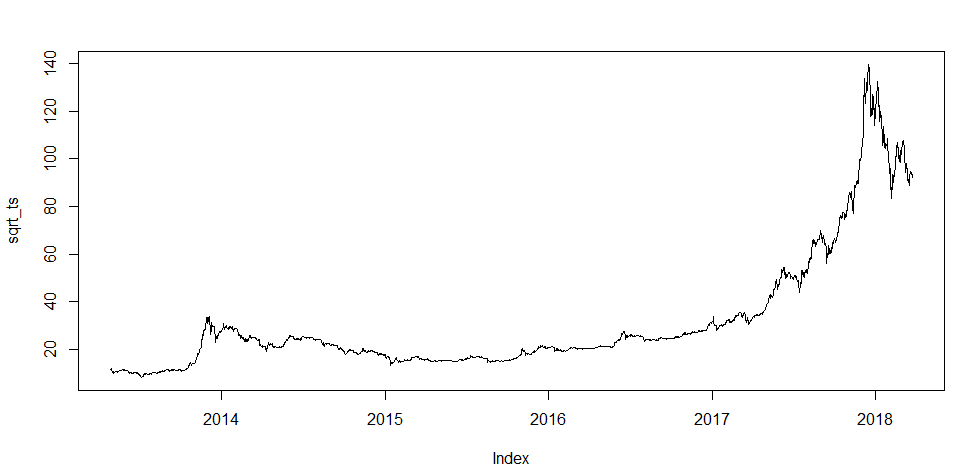
As it is clear from the plot, it is not decayed quickly. So, we need to apply data processing.

# **4. Data Processing**

# To achieve to a reliable data different data transformation is applied.

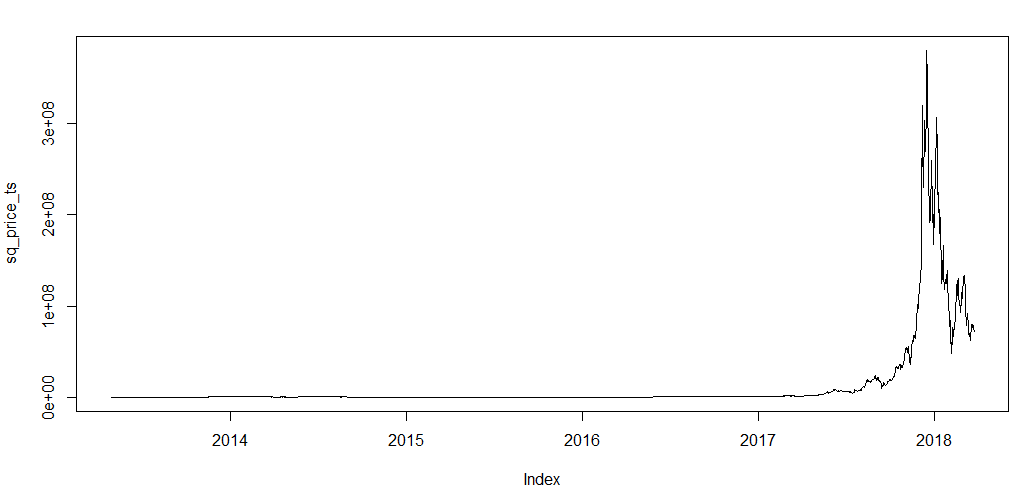
To see whether each transformation technique makes the data reliable or not the time-series plot is creayed for each one.

* **SQRT transformation**



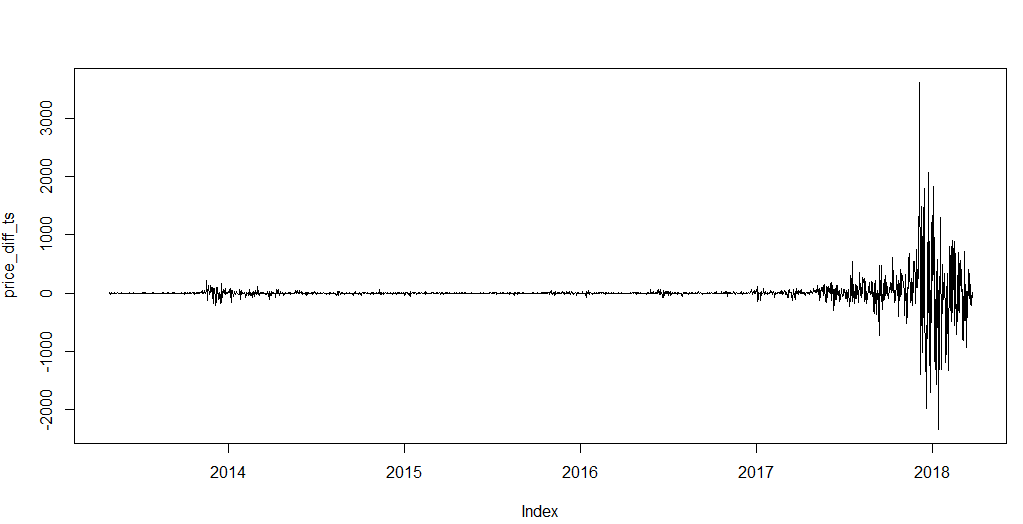
# As it is clear, there is not a significant change on the data by applying sqrt transformation.

* **Square transformation**



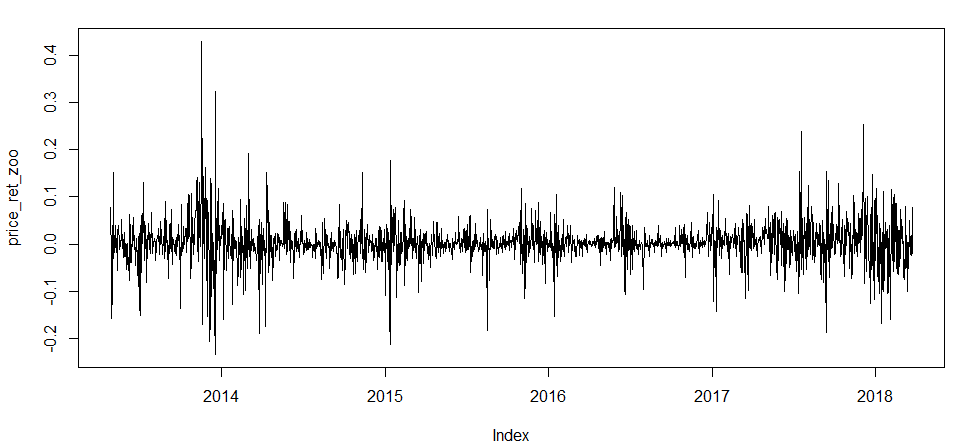
# Also the square transformation does not make a difference on the result.

# **Differencing**



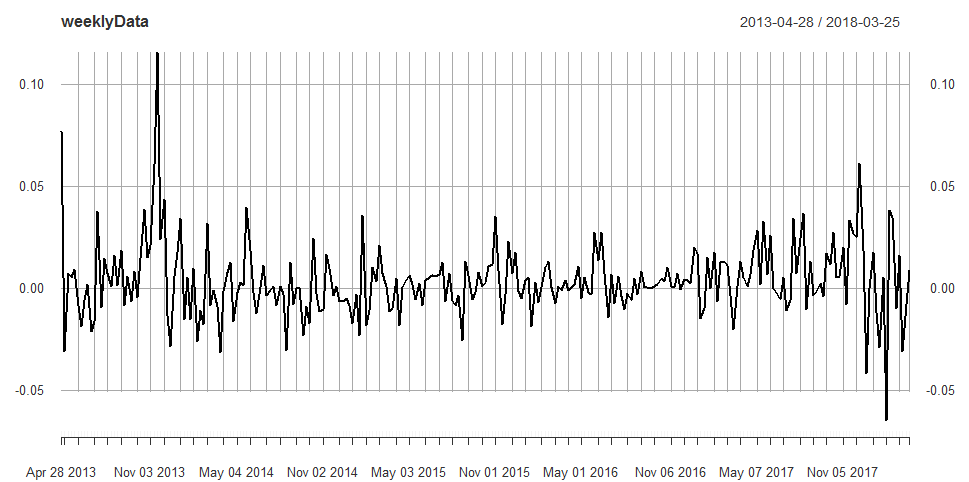
# After applying differencing, although mean does not change over time, variance does. So, data still is not stationary.

* **Return Transformation:**



Return transformation has made the data more reliable. The mean does not change over time and variance changes are smoother. To make sure if there is other transformation with better result, return transformation is applied on weekly data set. I mean daily data is converted to weekly data and each week has average value of the seven days of the week.

* **Return Transformation on weekly data:**

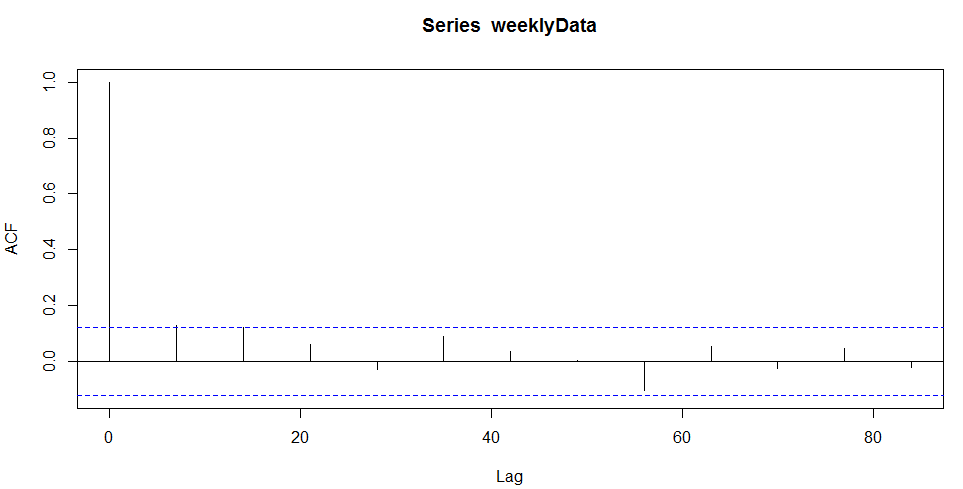


Return transformation on weekly data has made the data smoother.

As seen on the plot, mean does not change over time and variance changes are reliable to make sure the data is stationary.

**Serial dependency:**

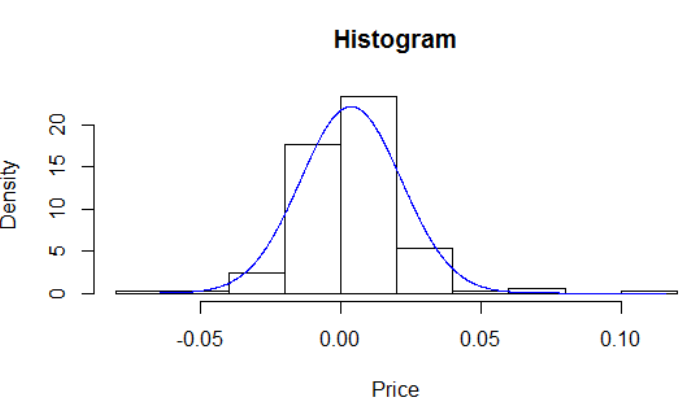
To figure whether the data is serially dependent or not, ACF plot is created.

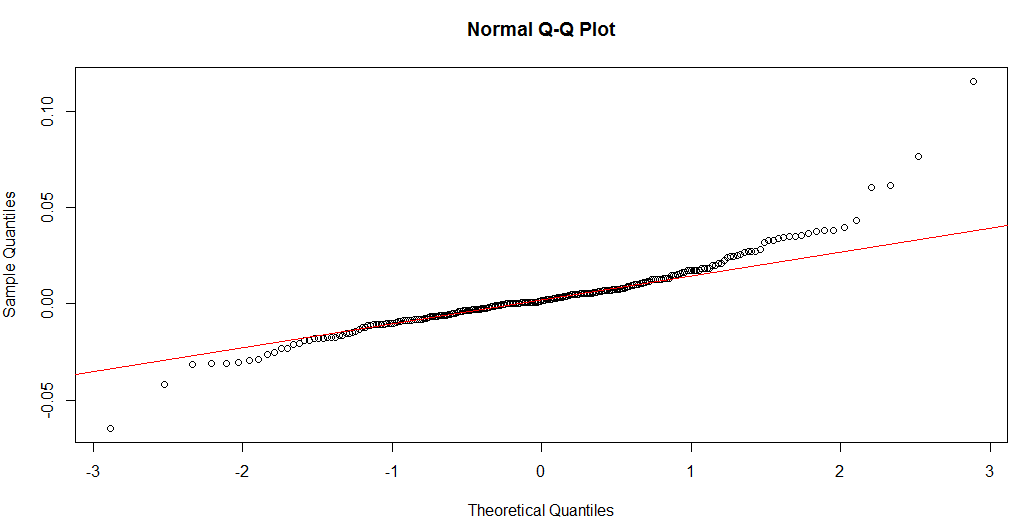


According to the ACF plot, there are some non-zero auto correlation values. Thus, it is serially correlated.

**Normality Test:**

To figure out the data is normally distributed or not, Histogram is created.





As it is clear, the data follows normal distribution.

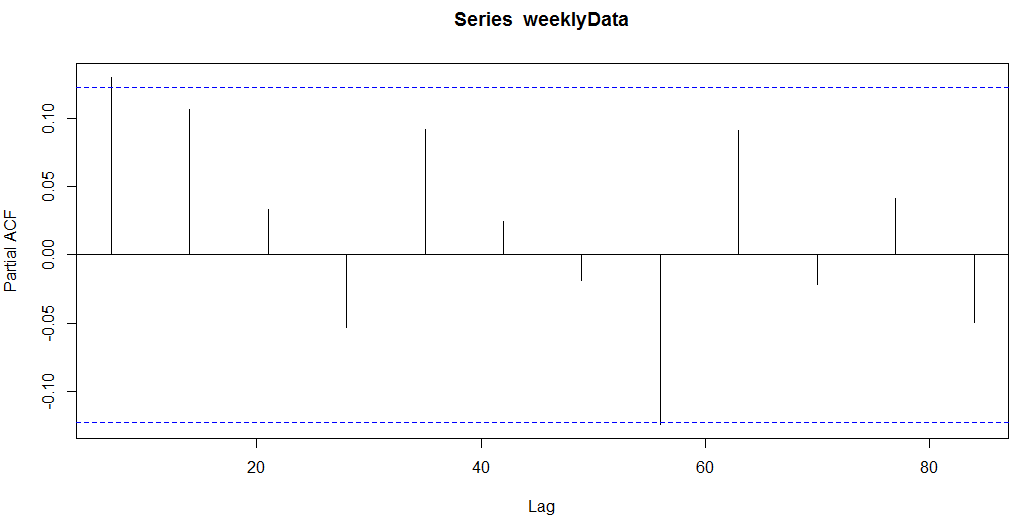
# **5. Methods and Process**

To find best model for the price in future, different models are built such as AR, MA, ARMA and ARIMA model.

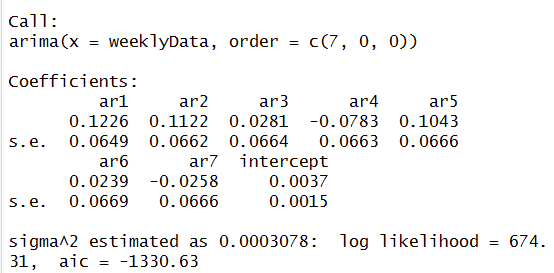
**AR:**

**Model creation:**

To build AR model, determination of value of p is required. To find the value, PACF plot is created.



According to the plot, PACF tails off at lag 7. So, the AR model with the p value of 7 is built.

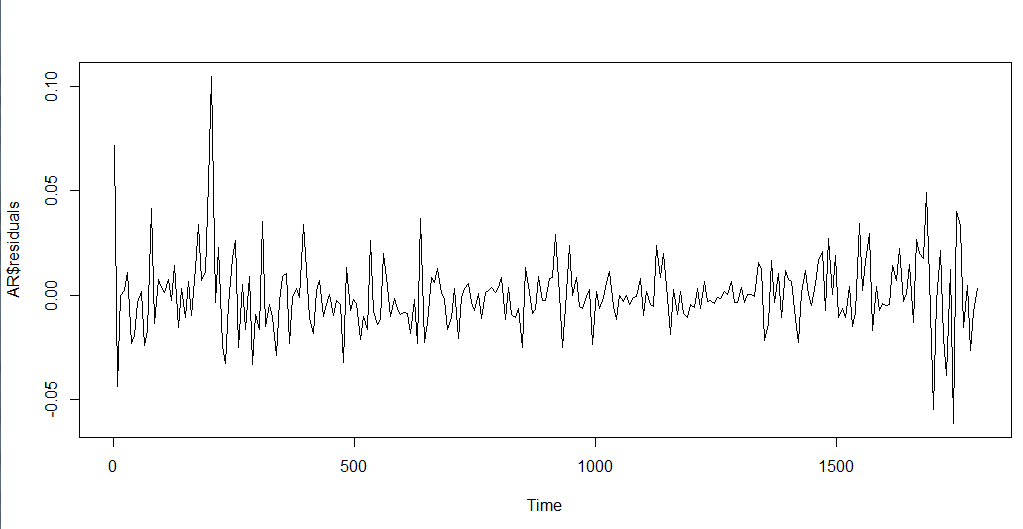


**AR model:**

**Xt = 2.6381 \* 10-3 + 0.1226 \* Xt-1 + 0.1122 \* Xt-2 + 0.0281 \* Xt-3 – 0.0783 \* Xt-4 + 0.1043 \* Xt-5 + 0.0239 \* Xt-6 – 0.0258 \* Xt-7 + at**

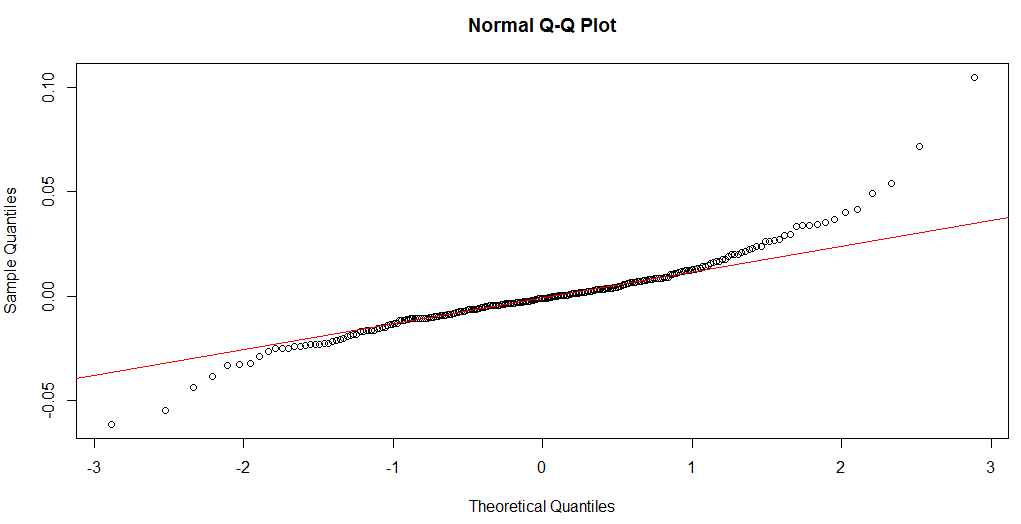
**Residual analysis:**

**Residual time plot is created.**



**Normality test:**

Normality for the model is validated by QQ plot.

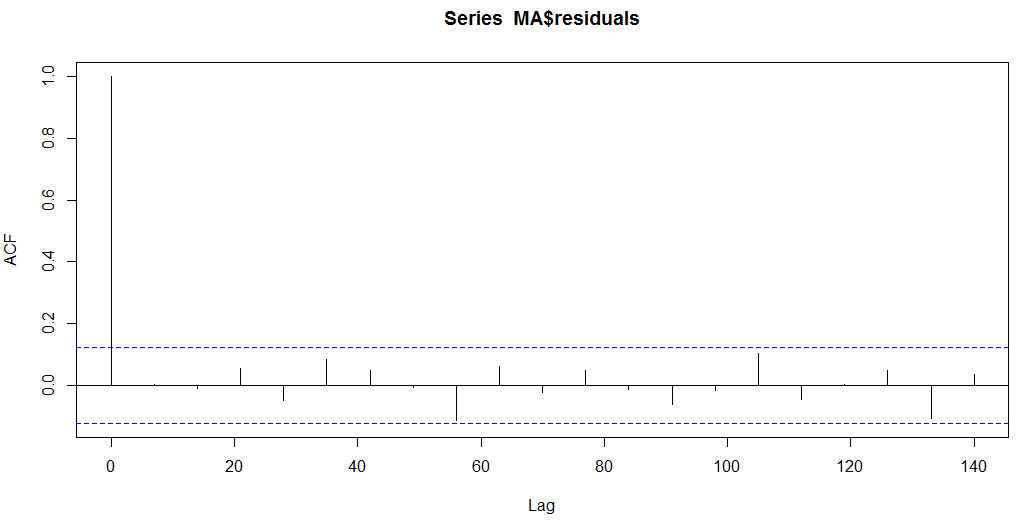


According to the plot, most of the spots are fallen on the line. So, it validates normally distribution.

# 

# Result of Ljung normality test (p-value>0.05) confirms the normally distribution of the residuals.

To check whether residuals are white noise or not ACF plot is created.

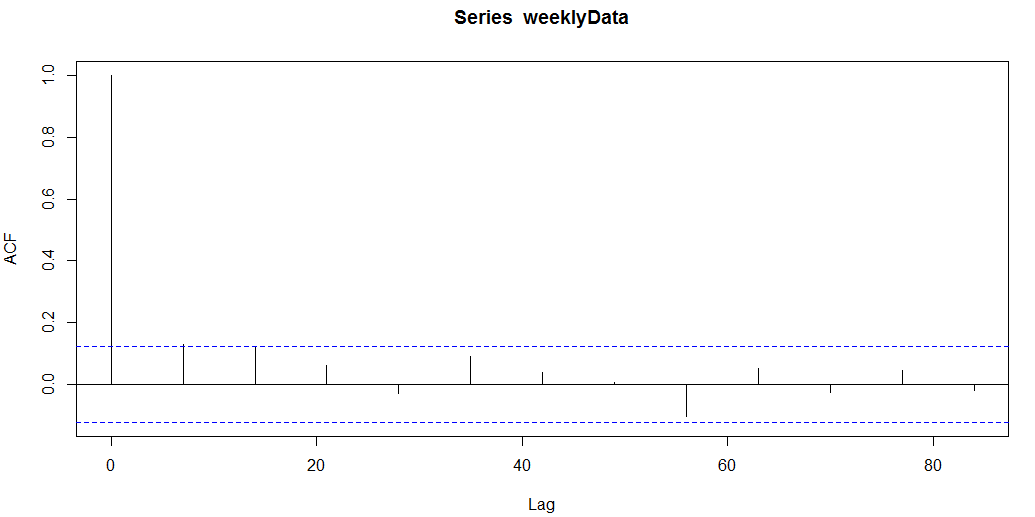


# Since all the values of the residual ACF plot are zero, it validates that residuals are white noise which meets the assumption on residual analysis.

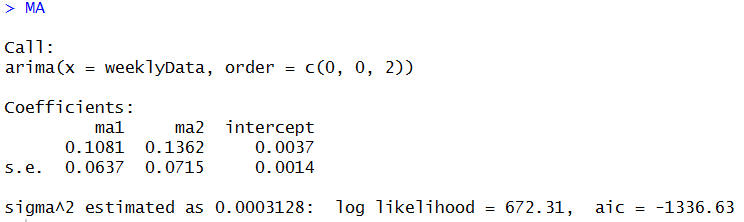
**MA:**

**Model creation:**

To build AR model, determination of value of q is required. To find the value, ACF plot is created.



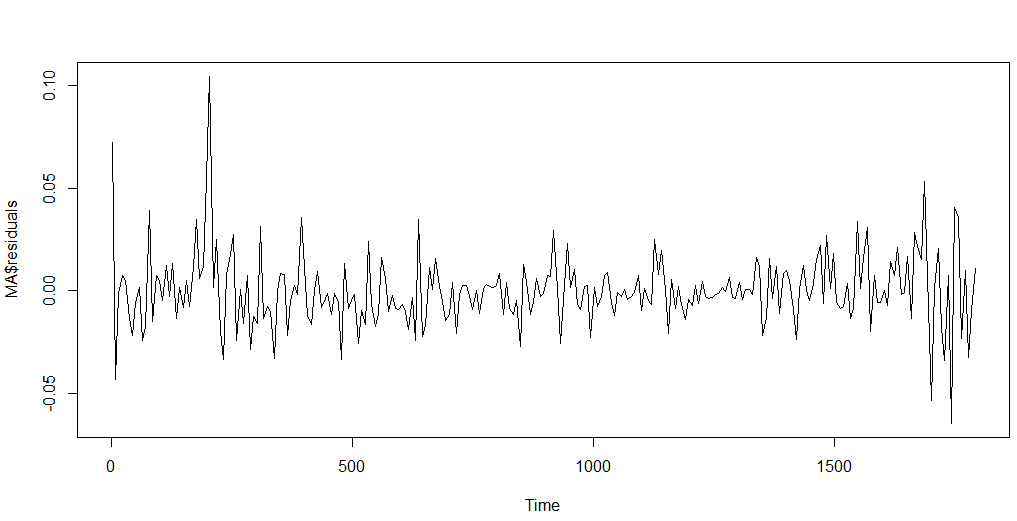
**According to the plot, ACF cuts off at lag 2. So, MA(2) is created.**



# MA model:

# rt = 3.62 \* 10-3 + at – 0.1081 \* at-1 - 0.1362 \* at-2

**Residual analysis:**



# **Normality test:**

Normality for the model is validated by QQ plot.

# 

According to the plot, most of the spots are fallen on the line. So, it validates normally distribution.

# 

# Result of Ljung normality test (p-value>0.05) confirms the normally distribution of the residuals.

To check whether residuals are white noise or not ACF plot is created.

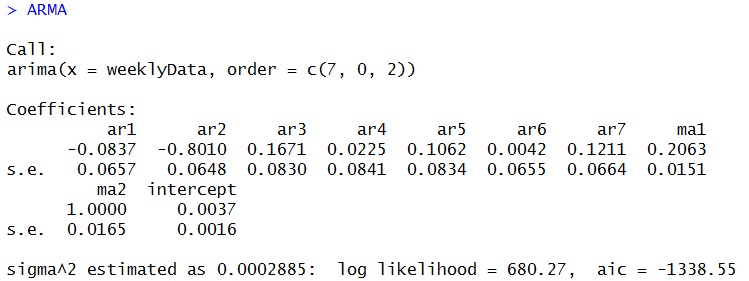
# 

# Since all the values of the residual ACF plot are zero, it validates that residuals are white noise which meets the assumption on residual analysis.

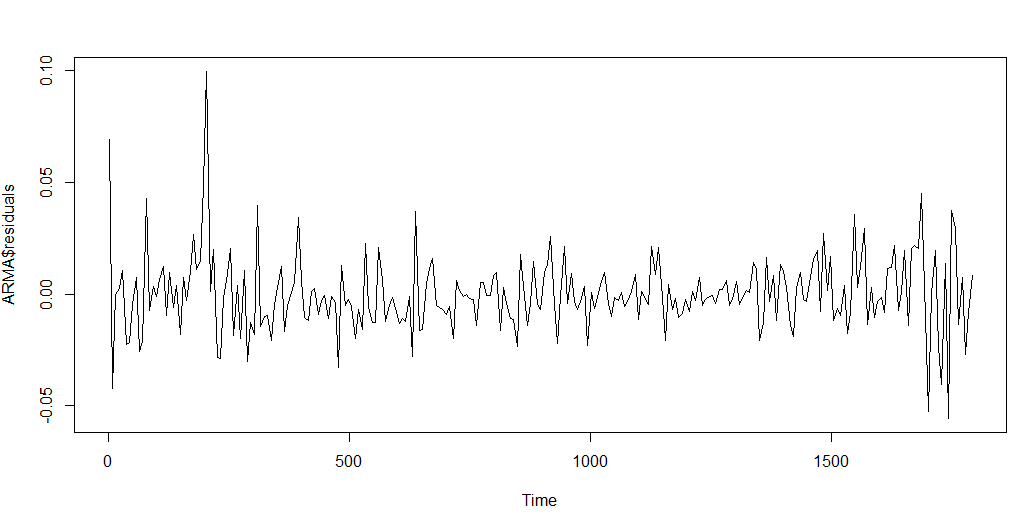
# **ARMA:**

**Model creation:**

As seen on MA and AR models, ACF plot tails of at lag 7 and PACF plot tails off at lag 2. So, p and q for the ARMA model are 7 and 2, respectively.

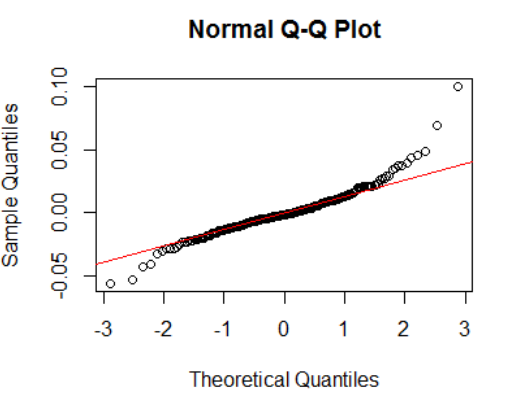


# **Residual Analysis:**



# **Normality test:**

Normality for the model is validated by QQ plot.

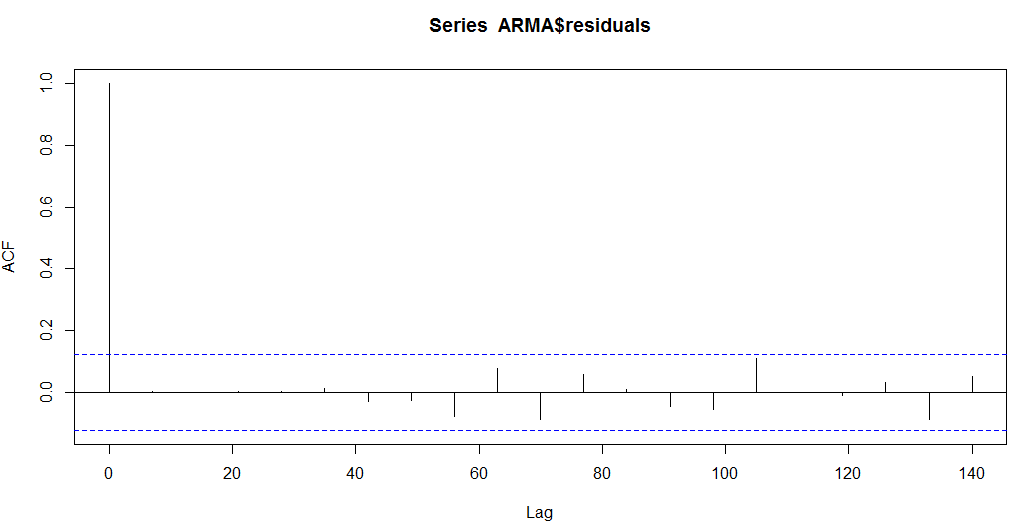


According to the plot, most of the spots are fallen on the line. So, it validates normally distribution.

# 

# Result of Ljung normality test (p-value>0.05) confirms the normally distribution of the residuals.

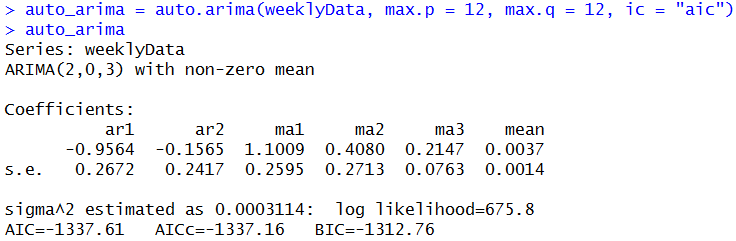
To check whether residuals are white noise or not ACF plot is created.



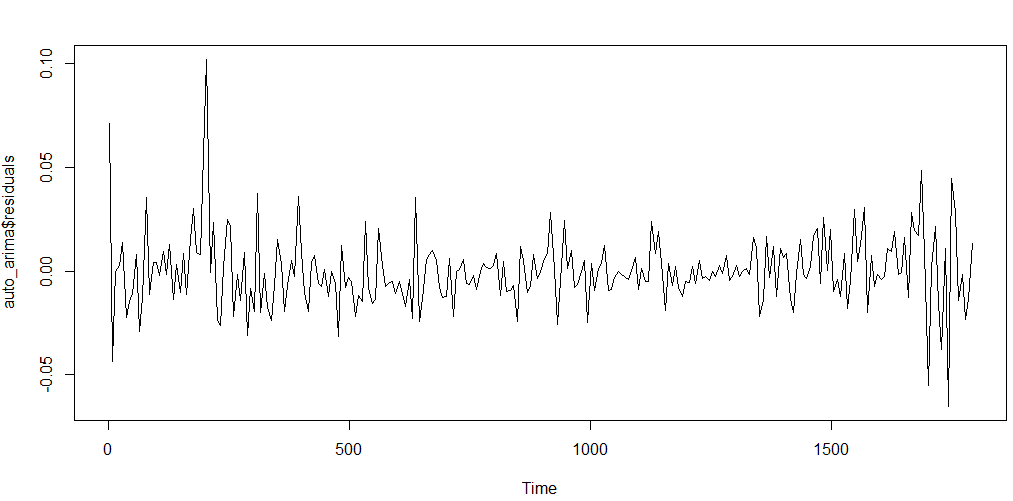
# Since all the values of the residual ACF plot are zero, it validates that residuals are white noise which meets the assumption on residual analysis.

**ARMA using auto.arima function:**

**Model Creation:**



# **Residual Analysis:**



Normality for the model is validated by QQ plot.

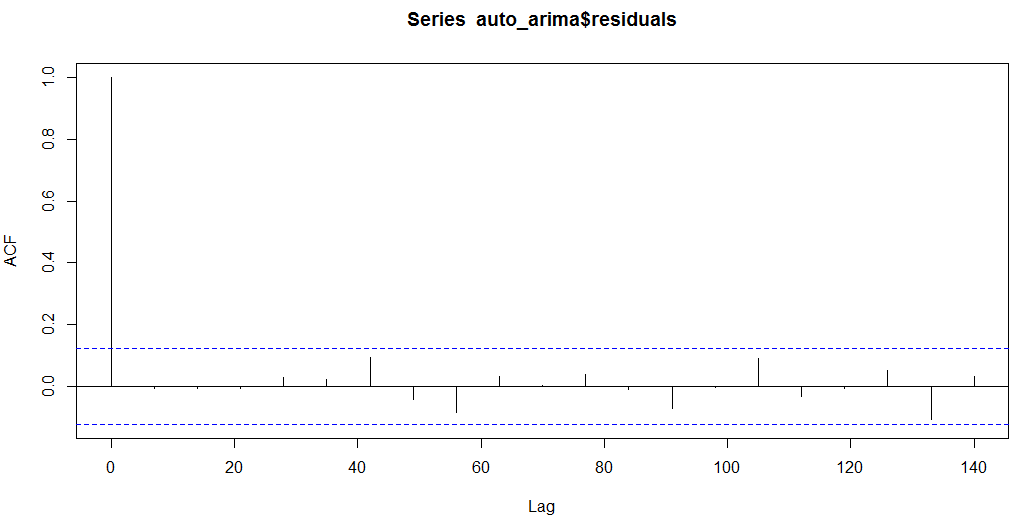
# 

According to the plot, most of the spots are fallen on the line. So, it validates normally distribution.

# 

# Result of Ljung normality test (p-value>0.05) confirms the normally distribution of the residuals.

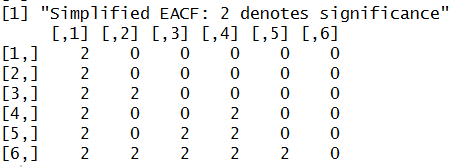
To check whether residuals are white noise or not ACF plot is created.

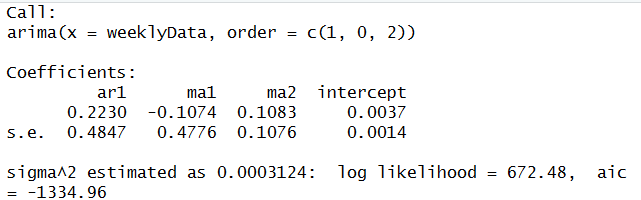


# Since all the values of the residual ACF plot are zero, it validates that residuals are white noise which meets the assumption on residual analysis.

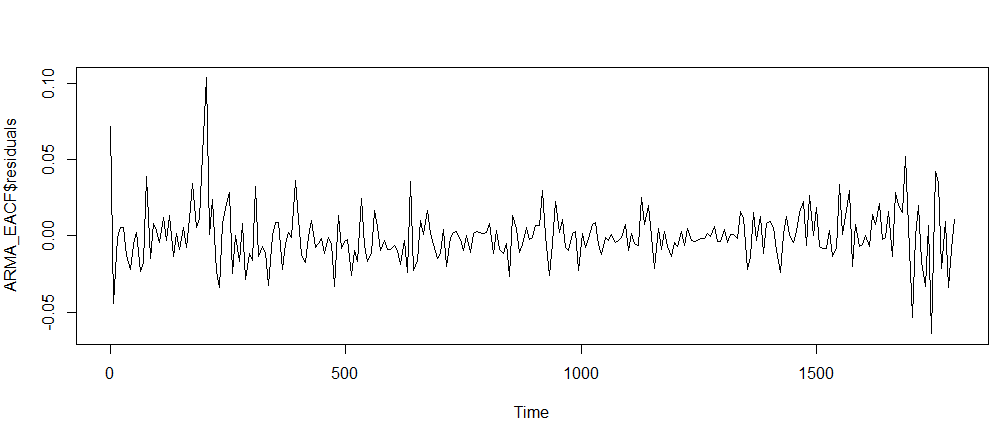
**ARMA using EACF Method:**

For this model creation, p and q are determined by EACF method. The following table show that p and q are 1 and 2, respectively.





# **Residual Analysis:**



Normality for the model is validated by QQ plot.

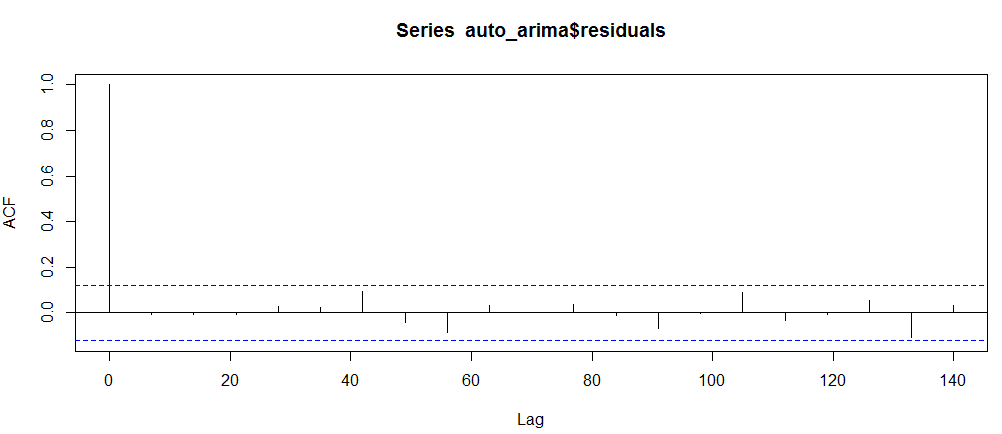
# 

According to the plot, most of the spots are fallen on the line. So, it validates normally distribution.

# 

# Result of Ljung normality test (p-value>0.05) confirms the normally distribution of the residuals.

To check whether residuals are white noise or not ACF plot is created.



# Since all the values of the residual ACF plot are zero, it validates that residuals are white noise which meets the assumption on residual analysis.

**Model selection:**

To select which model is the best fit for the data, AIC is chosen as a metric.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **AR(7)** | **MA(2)** | **ARMA(7,2)** | **ARMA(2,3)** | **ARMA(1,2)** |
| **AIC** | -1330.63 | -1336.63 | -1338.55 | -1337.61 | -1334.96 |

AIC values for different models are so close together. So, at this time we cannot claim which one is the best fit. We must apply evaluation on all the models to figure out the best one.

# **6. Evaluations and Results**

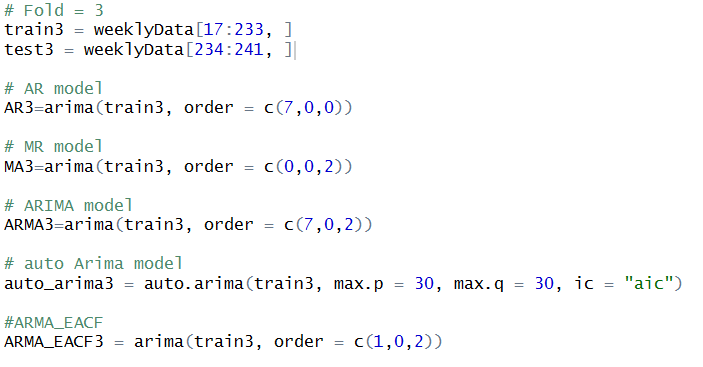
## 6.1. Five-fold cross validation

## Because number of the data are too small, validation method must be N-fold cross validation. To do so, data is separated to 5 fold. Each fold contains of 217 training data and 8 test data. For example, first fold’s training data set starts from first data and ends to 217th data and its test data set start from 218th data and ends to 225th data.

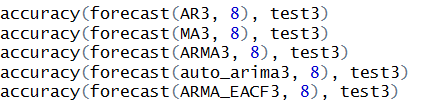
## In a same way, second fold’s training data set starts from 9th data and ends to 225th data and its test data set start from 226th data and ends to 233rd data. Other folds follow the same pattern.

The validation method is applied for all the models.

A sample of cross folding is shown:



Then accuracy is applied for each fold with eight predictions ahead to compare with the actual data of the test data set.



## To figure out which model fits the best, MAE is considered as a metric. Since there are five folds for each model, mean value of five related MAE is considered as each model’s MAE.

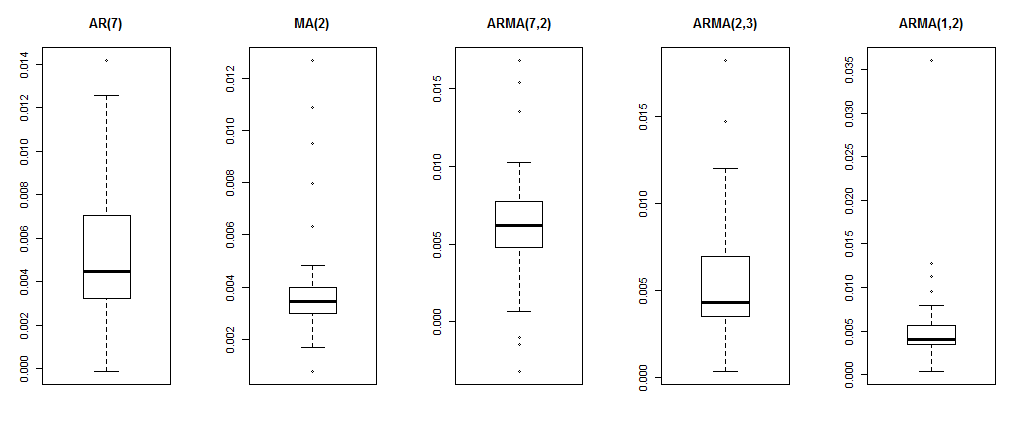
## 6.2. Results and Findings

As mentioned earlier, mean MAE value of five folds of each model is considered as the model’s MAE. The result is shown in the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **AR** | **MA** | **ARMA(7,2)** | **ARMA(2,3)** | **ARMA(1,2)** |
| **MAE** | 0.02670786 | 0.026707 | 0.02649215 | 0.02662835 | 0.02670202 |

As seen in the table, MAE values for all models are too close. However, ARMA(7,2) is maintaining the least value of MAE. So, we need to apply two-paired hypothesis testing on the ARMA(7,2) with all other models.

At first we need to create box plot of the predicted values of different models to see is it possible to compare the different models by their median values or not.



Because group variance is large in the models’ box plot it is not reliable to compare the groups based on their medians. In other words we cannot use box plot to make solid conclusion.

As I mentioned earlier, we need to apply hypothesis testing on the group of data.

1. ARMA(7,2) with AR(7)

# **7. Conclusions and Future Work**

## 7.1. Conclusions

xxx

## 7.2. Limitations

xxx

## 7.3. Potential Improvements or Future Work

Xxx