

|  |  |
| --- | --- |
| Time Series Analysis | |
| **Group3\_Project1\_Summer624 DATA624 Project One** |  |
| DATE : Jun 23nd 2020 | TEAM:Kuiete Tchoupou AlainLittlejohn JeffreyMalhotra SamritiMishra RajwantNg Jimmy |

Contents

[ABSTRACT 2](#_Toc41047762)

[KEYWORDS 2](#_Toc41047763)

[INRTODUCTION 3](#_Toc41047764)

[LITERATURE REVIEW 3](#_Toc41047765)

[METHODOLOGY 4](#_Toc41047766)

[OVERVIEW 4](#_Toc41047767)

[DATA EXPLORATION 5](#_Toc41047768)

[Build Data 7](#_Toc41047769)

[Time Series Forecasting 8](#_Toc41047770)

[BUILDING MODEL 11](#_Toc41047771)

[EXPERIMENTATION AND RESULTS 12](#_Toc41047774)

[DISCUSSION AND CONCLUSIONS 13](#_Toc41047775)

[REFERENCES 14](#_Toc41047776)

[APPENDIX 14](#_Toc41047784)

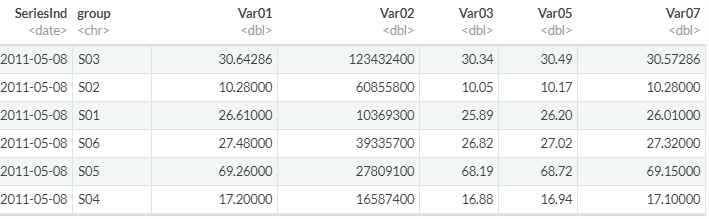
### ABSTRACT

* In this project work we have data set which is grouped in Six different groups and each group is having Six variables. Our objective it to Predict the variables for next 140 140 periods.
* Given data must be analyzed using Time Series analytical techniques.
* We will identify how this time series fits on Stationarity assumption.
* We have evaluated the data and residuals using ACF and PACF, to check for autocorrelation among different lags of data.
* We are discussing which model can best represent our data, AR(Auto Regression) and MA(Moving Average) , or ARIMA .
* We have evaluated all data models comparing it’s AIC score
* Lastly, we have stored the predicated value in the Excel file for next 140 points.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| KEYWORDS  * **Time Series:** It’s 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. * **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). For example, yt on yt−1: yt=β0+β1y(t−1)+ϵt. * **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. * **Trend**: When we see data follows a certain trend as it moves along the time, it can be going upward or downward. Each time series may have Level, Trend, Seasonality and Noise. * **Trend Stationary:** When data follows a stationary pattern over the trend line, (a stochastic process is trend stationary), then such series is called trend stationary. To simplify, removing of trend would make this series stationary. * **Drift:** It is a constant which is mostly related to the average change in the value.  OVERVIEW  * High Level View of data: Data has groups from S01 to S06, and Variables Var01, Var02, Var03, Var05, Var07. * We are expected to Predict following variables, form each group.  |  | | --- | | **GROUPS** **Variables**   * S01 - Forecast Var01, Var02 * S02 - Forecast Var02, Var03 * S03 - Forecast Var05, Var07 * S04 - Forecast Var01, Var02 * S05 - Forecast Var02, Var03 * S06 - Forecast Var05, Var07 |  ASSUMPTION  1. For simplicity we have converted “SeriesInd” to date. by setting Origin of date to 1900 Jan 1st. 2. Below is table where we have replaced missing value with some approximate value  |  |  | | --- | --- | | Group | Approximated Missing value | | S01 | Yes | | S02 |  | | S03 |  | | S04 |  | | S05 |  | | S06 |  | |  |  |

### DATA EXPLORATION

Actual data adjusted by date for better visualization can be seen in fig 2.

  
(figure 2: Adjusted `**SeriesInd**` by adding Series to date 1900 Jan 1st.)

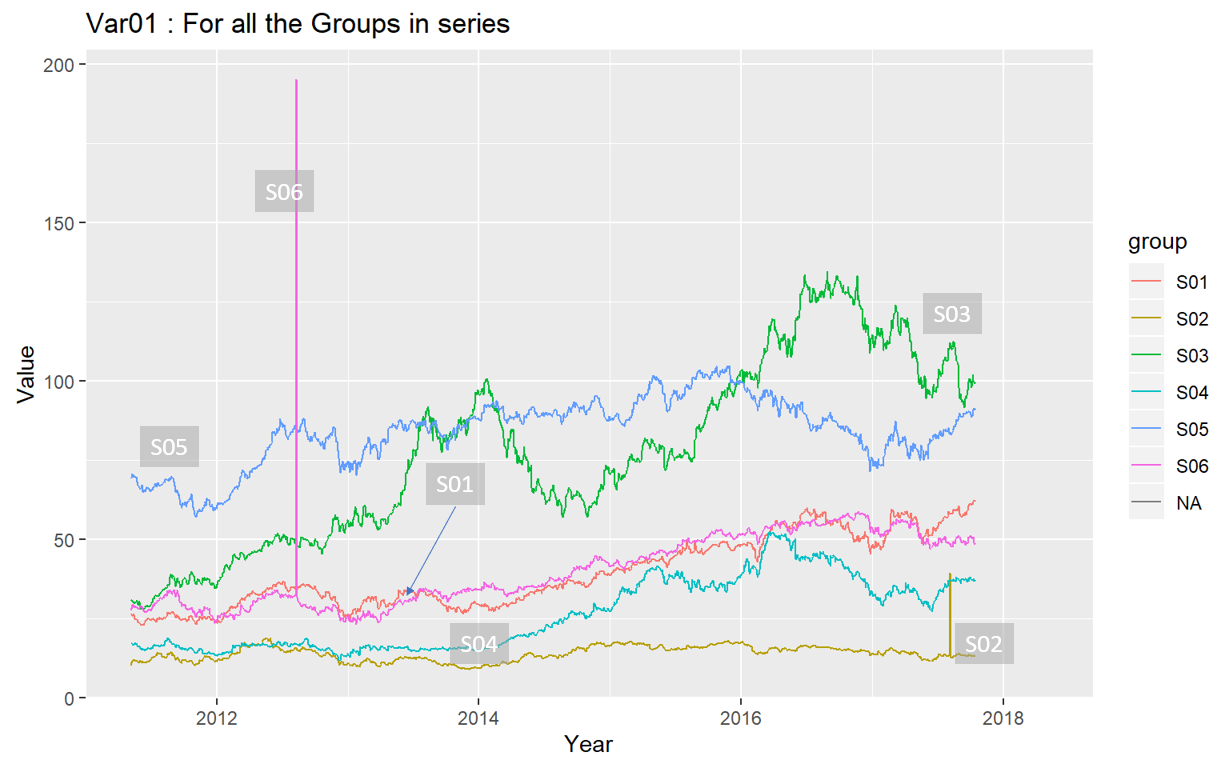
The above data has “Start Of Data: 2011-05-08” and “End Of Data: 2018-05-03”, expected to Forecast data from : 2018-05-04 to 2018-09-20.

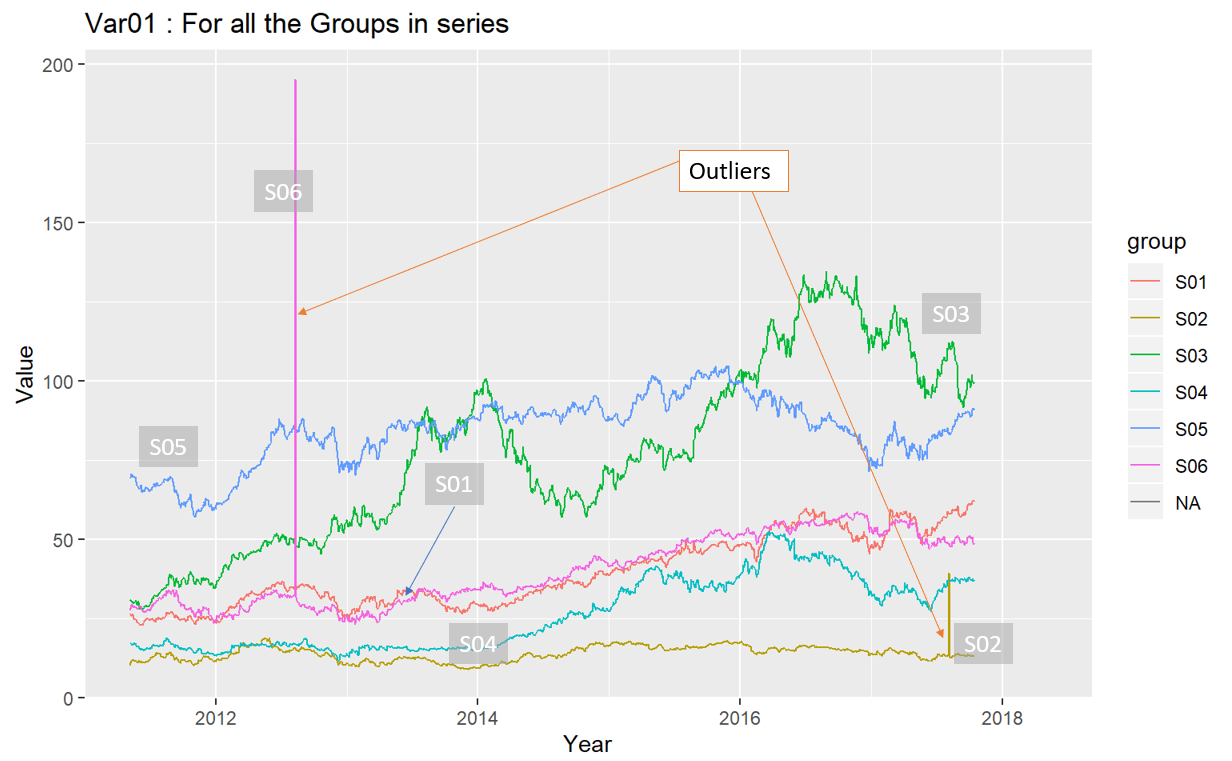
As you can see that this data set is very complex to do any time series analysis in the given form, lets break the data in the readable from for each group, and then by each variable along with date.

Code 1. Subset Data for Group and Variable

dt\_s01\_v1 = **subset**(full\_data[,**c**(1,2,3)], group **==** "S01")**%>%** .[,**c**(1,3)]**%>%** **left\_join**(**as.data.frame**(allDates),.,by=**c**("allDates"= "SeriesInd"))

When checking the flow of Var01 across all the groups, we can that some groups has data that shows some trend and seasonality for example group S03, and group S05 , S06, S01 and S04 show some upward trend. *Technical Note*: This is very important to identify the trend from S05,S06, S01 and S04 , and check if it’s a real trend or just drift with Random walk. We will try to do some statistical test to check these in latter part our report.

  
(figure 3: Var01 across all Groups)

Outliers: Data points that don’t follow pattern of whole data set or stand out with in the whole dataset are called outliers. They have serious impact on the data and its forecast. Figure 4, suggest how group S06 and S02 for Var02 have outliers and the prediction of S06 may not be much impacted by the outlier as it’s in the beginning of series , whereas the outliers from group S02 is very close to the end of the series, hence it would have impact on the prediction if we don’t drop this outlier from the series before building out model.  
  
(Figure 4: Outliers in Var01 for each group)

Below we have listed all the Variables across each group, we see same pattern as we noted in the from **figure 3** and **figure 4**, for var01, var03, var05, var07. But for Var02 we see very no trend across each group or downward drift over the time as shown in figure 5a.

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\81937748.tmp  (Figure 5 a: Var02 for all Groups) | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\19110496.tmp  (Figure 5 b: Var03 for all Groups) |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\24CB5514.tmp  (Figure 5 c: Var05 for all Groups) | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\C547C042.tmp (Figure 5 d: Var07 for all Groups) |

Figure 6a and 6b, shows how datapoints from group S01 are flowing over time. From figure 6a, the Var02, and is the only prominent data points , so when we dropped it in figure 6b, we were able to see how mostly all the variables (Var01, Var03, Var05, Var07) are mostly following the same trend / drift over time.

|  |  |
| --- | --- |
| Figure 6a | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\73EC40AE.tmp  Figure 6b |

Let’s see how data is centered around its mean by plotting the box plot. We can see that similar to figure 3 (line plot of the var01 across group ), we can see that all the data points are with the range , except group S02, S05 and S06.

We are wondering how we couldn’t catch S05 in our earlier graph, possible reason is these outliers are very close minimum value in the dataset and they would not show any spike in the dataset rathe a fall, which we missed to locate.

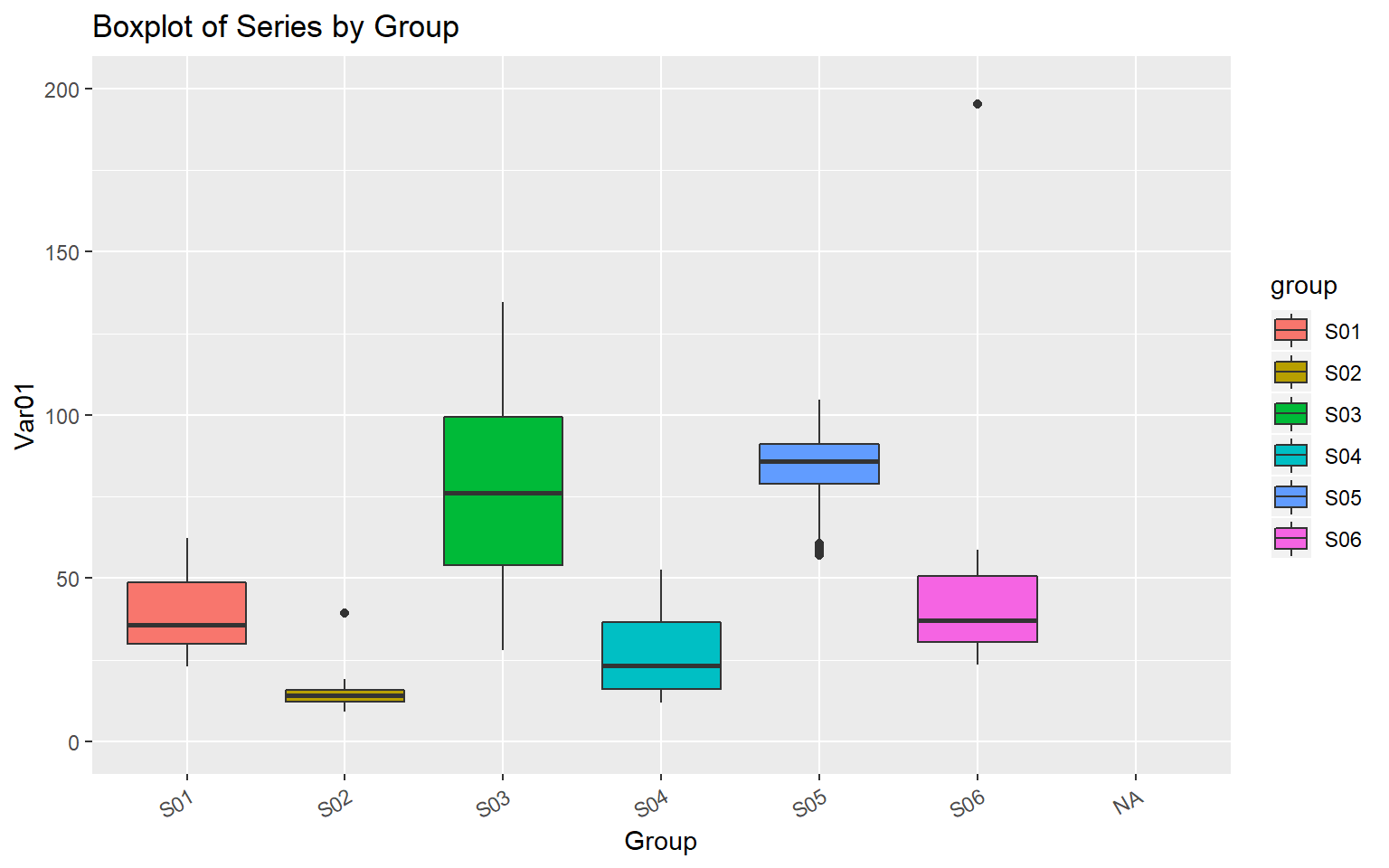


Figure 7 : Boxplot for var01 across groups

Missing Value:

The red spots in the below plots of Var01 from group S01, we can see res spots quite often in the data. This is the indicator of missing data points. We are approximating this value for the full series before we can work on data. This process is called imputing of the data.

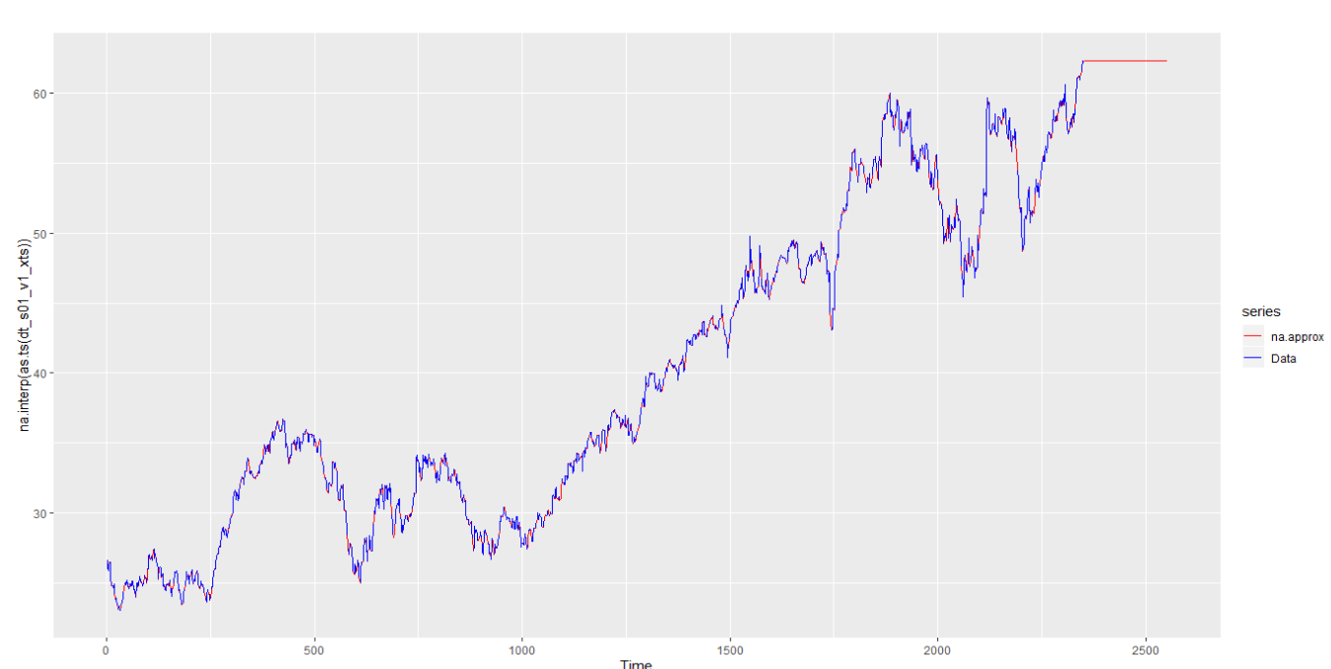


Figure 8 a. Missing value and actual data of Var01 from S01 group

Below we are trying to show how our data imputation has approximated some of the data points of Var01.

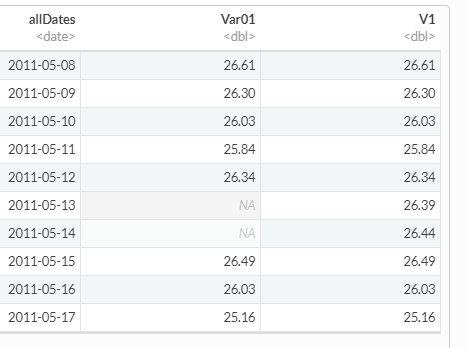


Figure 8b: Imputed data

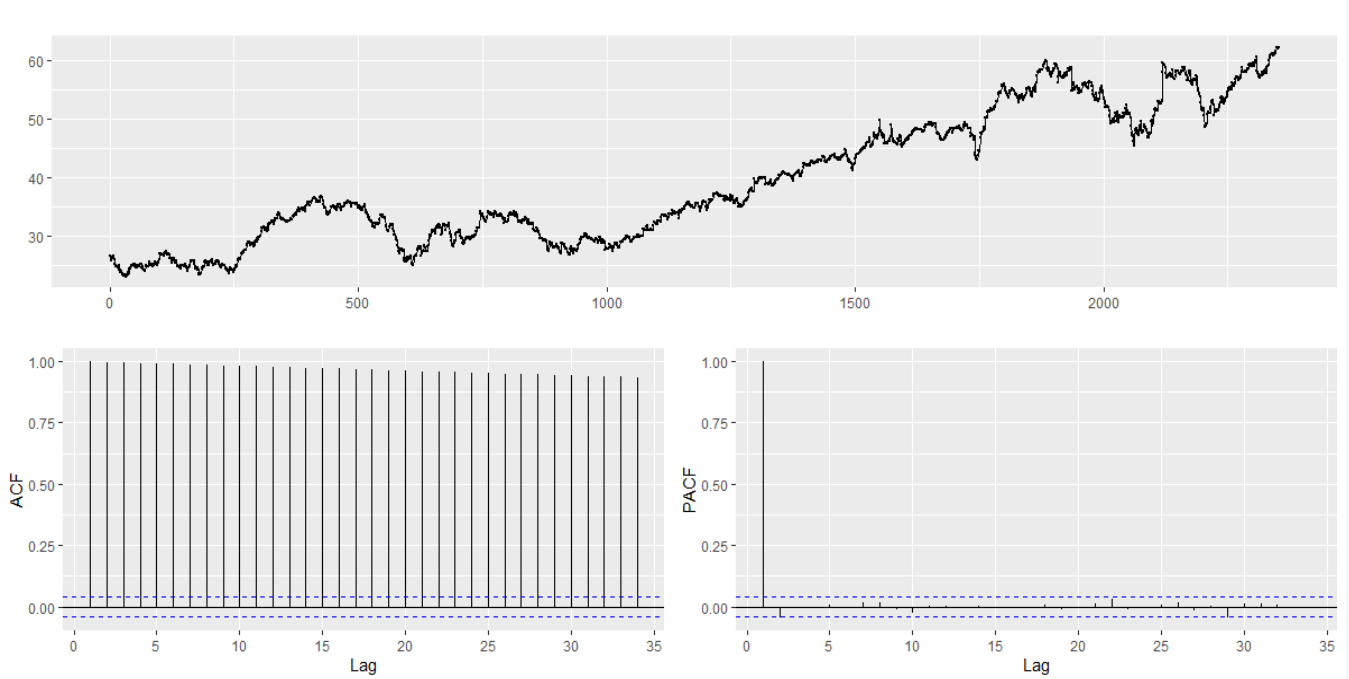
|  |  |
| --- | --- |
| We noted that we have around 1645 values missing, which we have replaced with “NA” in above data preparation. Here we have assumed that data should be present for each day and if its missing due to some reason, so we have accounted for that.  Var01- have 1645 NAs, Var02 has 1633 NA. | Figure 8 c : Stats of the data. |

For better analysis we have created dataset for each variable with date, lets evaluate the statistical summary of each data point :

|  |  |
| --- | --- |
| Figure 9a: Var01 from Group S01 | Figure 9b: Var02 from Group S01 |
| Figure 9c: Var03 from Group S01 | Figure 9d: Var05 from Group S01 |
|  | Figure 9e: Var07 from Group S01 |

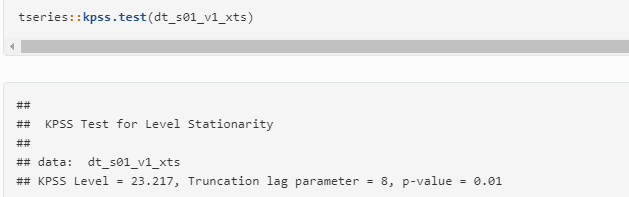
ACF and PACF:

We will now plot the ACF and PACF of Var01 and Var02 from group S01. Below plots of Var01 from S01 group, Since ACF plots very slowly moving towards ZERO for all the variables, we can say that our data series is non-stationary.



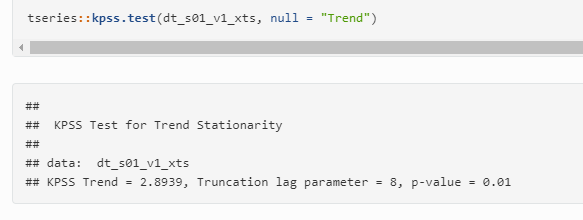
Plot of Var01 from S01 group

We will user KPSS test to check Null Hypothesis : Data is stationary around the ~~Level/trend/~~Mean line.

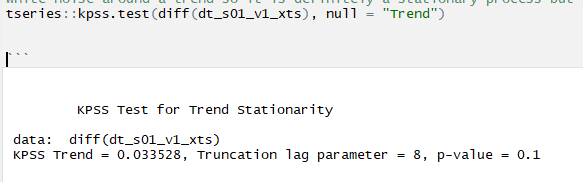


Above test suggest that p value is less than .05 and hence we reject the null and say that data is not stationary around the Level/trend/Mean line.

Technical Info: We noted that this method by default tests for stationarity around a 'mean' only. Let’s test this with Trend parameter. So our null Hypothesis would be: Data is stationary around the Trend line.



The p-value is less than 0.05. The null hypothesis of stationarity around a Trend is rejected. Lets apply some diff function in the data, and then see the impact of KPSS test on Trend line.



This is white noise around a trend so it is definitely a stationary process but has a trend. Here its very tricky to say that if this trend is a real trend or drift from random walk model. More test can be done to clearly find this info.

So far, we noted that our data needs 1 level of differencing, then we can see data is stationary. Let’s plot ACF and PACF plot after differencing and evaluate the data. Below data shows how we have removed some trend/drift from the data after differencing.

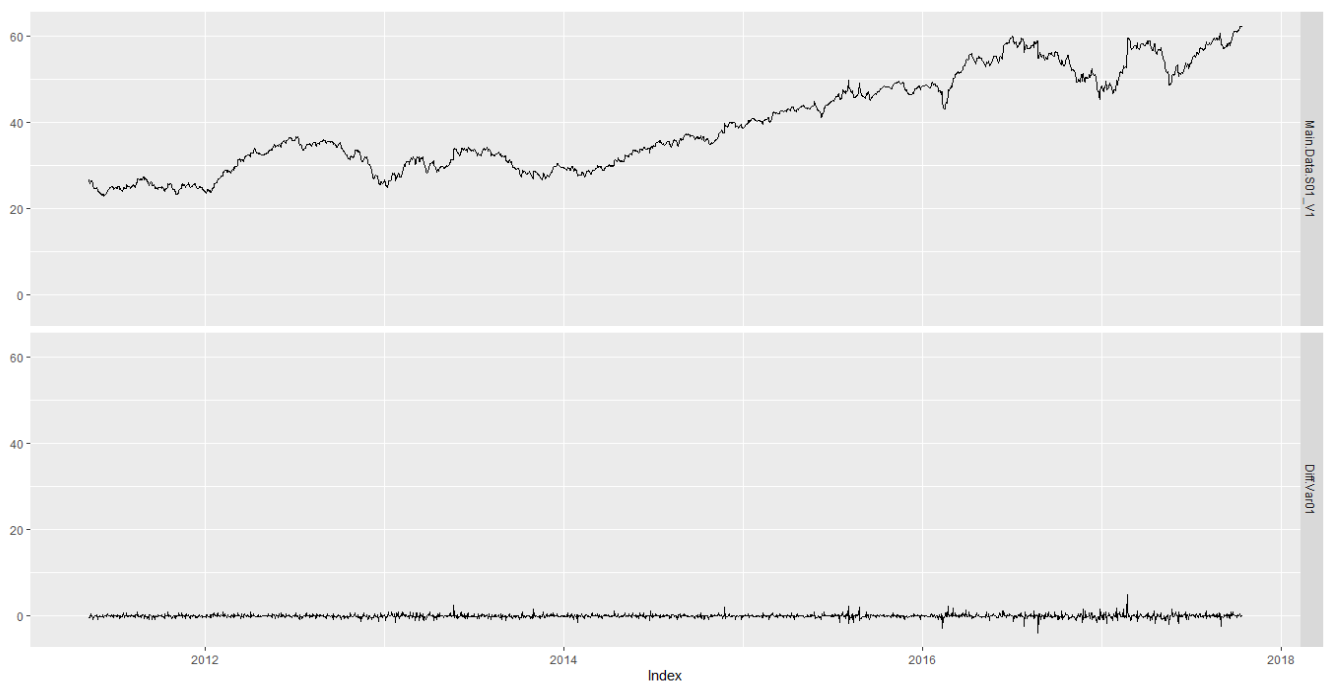
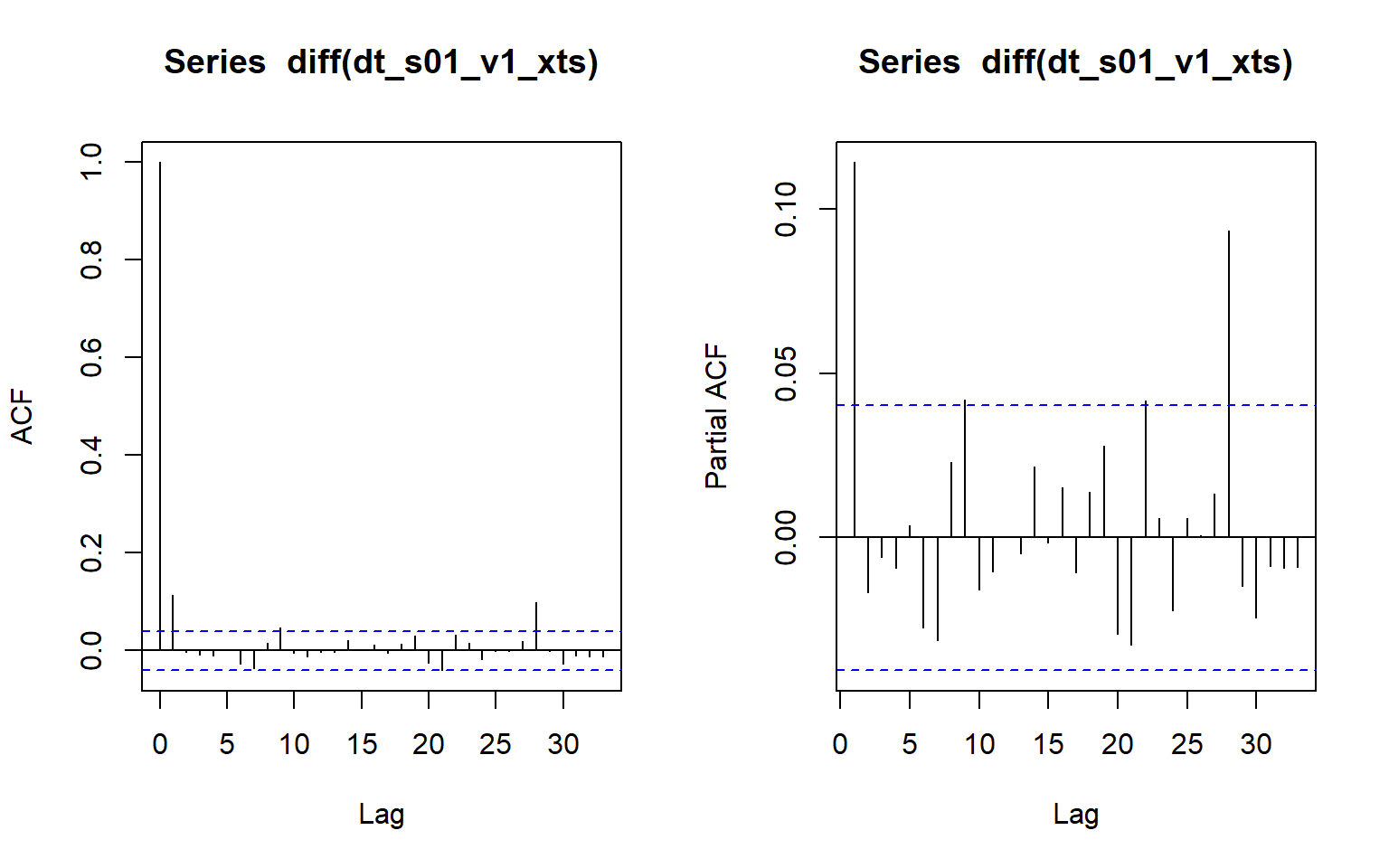


Figure : Var01 from s01 group

Difference data looks very much white noise in the bottom panel of the above graph.

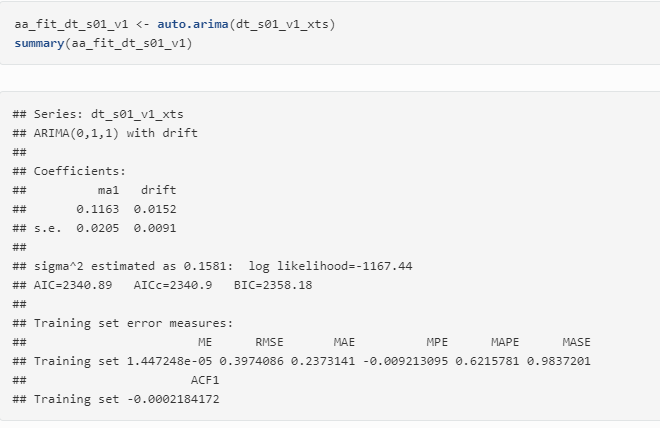
### 



ACF plot after Diff(var01) from S01 group

Above ACF plots suggest that data for variable 1 is more correlated with 1st lag, i.e. first order MA model can be used to define such data after applying difference on the data. PACF plots shows so many significant PACF values, so we would have to consider too many variables for AR model, which would make it complicated. Hence, we are going to use MA first order model.

We think variable Var01 , can be best predicted by I(1) and MA(1) model,  Let’s use auto.arima to validate our model and fir it.

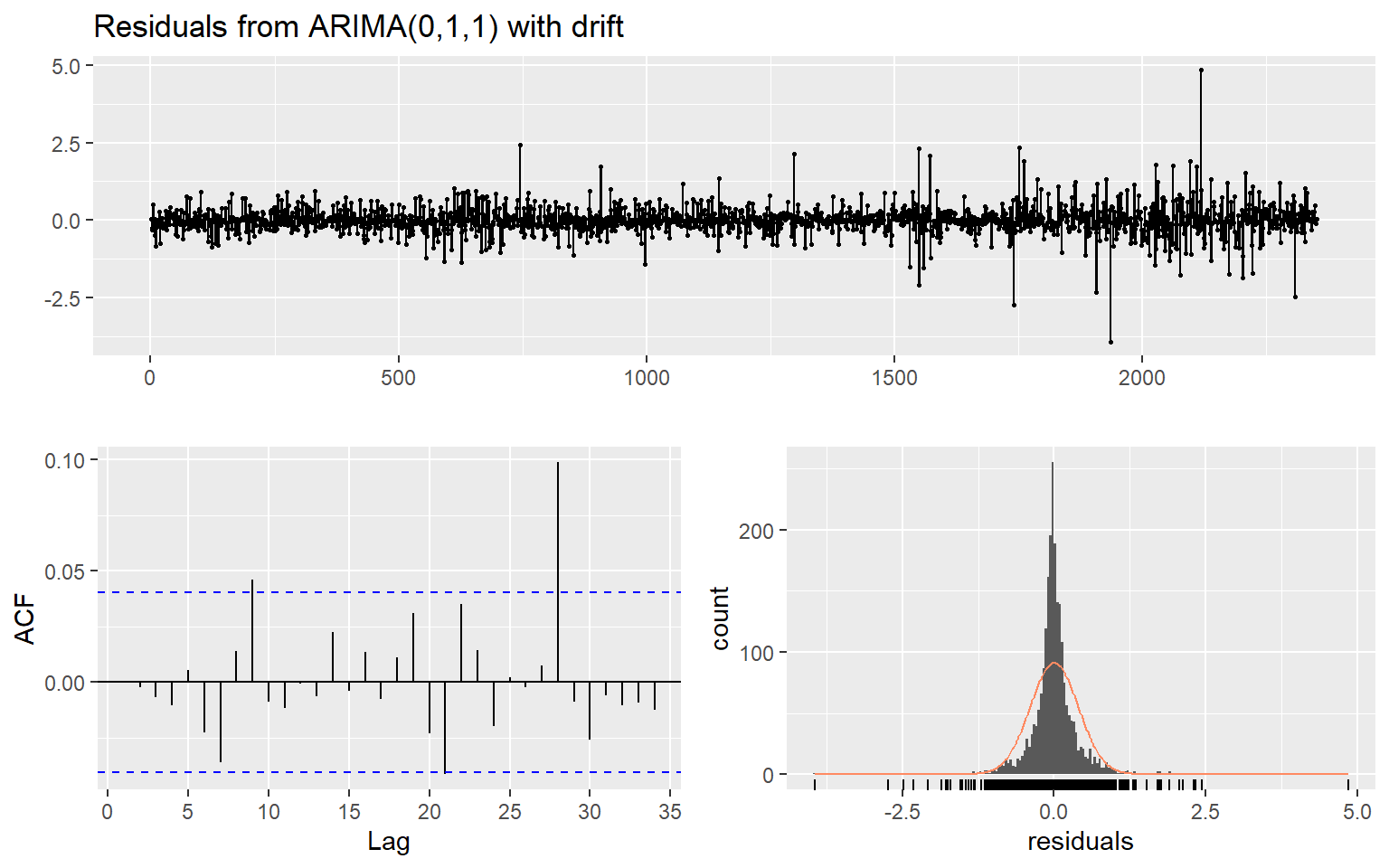


Auto.arima result for Var01

The auto.arima results with ARIMA(0,1,1) model with drift , which means that our data is not having any trend in it, the increase in the data over the year is better explained by a constant known as drift and not a trend which is function of time. Below are Coefficients of the model :

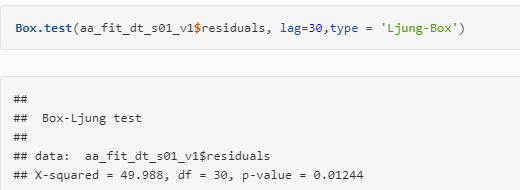
## ma1 drift   
## 0.11626944 0.01517711

Lets check the ACF plot of residuals, as we want to have white noise or very random residual in order to have better performing model.

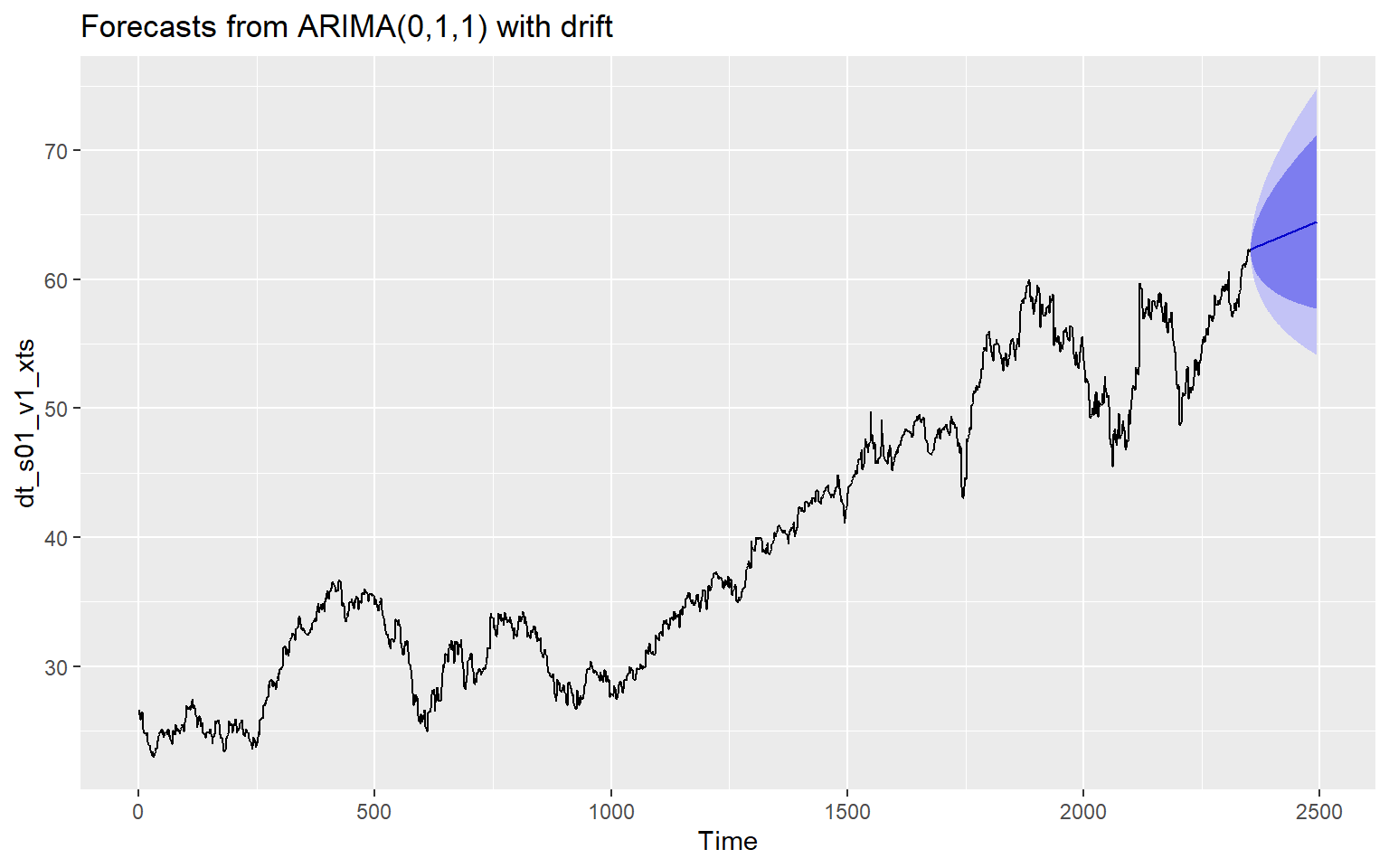


ACF plot of the residuals are white noise, as no prominent patterns can be seen here.

The Ljung-Box test also returned High p-vlaue indicating that we can't reject the null hypothesis, and data is white Nosie.



Forecasting :



### EXPERIMENTATION AND RESULTS

### 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](https://www.datacamp.com/community/blog/r-xts-cheat-sheet)
* [Introduction to Stock Analysis](https://lamfo-unb.github.io/2017/07/22/intro-stock-analysis-1/)
* [R for Data Science cheat-sheet](https://s3.amazonaws.com/assets.datacamp.com/blog_assets/xts_Cheat_Sheet_R.pdf)
* [A little book of R for Time Series](https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html)
* [Applied Time Series Analysis for Fisheries and Environmental Sciences](https://nwfsc-timeseries.github.io/atsa-labs/sec-tslab-correlation-within-and-among-time-series.html)
* [Autoregressive Models](https://online.stat.psu.edu/stat501/lesson/14/14.1)
* [Moving-average model](https://en.wikipedia.org/wiki/Moving-average_model%23Definition)
* <https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/#:~:text=A%20given%20time%20series%20is,non%2Dsystematic%20component%20called%20noise.&text=Trend%3A%20The%20increasing%20or%20decreasing,random%20variation%20in%20the%20series.>

### [Trend stationary - Wikipedia](https://en.wikipedia.org/wiki/Trend_stationary" \l ":~:text=In%20the%20statistical%20analysis%20of,not%20have%20to%20be%20linear.)