

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

Contents

[Time Series Analysis 1](#_Toc44148290)

[INTRODUCTION 3](#_Toc44148298)

[ABSTRACT 3](#_Toc44148299)

[KEYWORDS 5](#_Toc44148300)

[OVERVIEW 6](#_Toc44148301)

[ASSUMPTION 6](#_Toc44148302)

[SUMMARY 6](#_Toc44148303)

[DATA EXPLORATION 8](#_Toc44148304)

[Group S01 11](#_Toc44148305)

[Group S02 30](#_Toc44148307)

[Group S03 37](#_Toc44148308)

[Group S04 45](#_Toc44148309)

[Group S05 58](#_Toc44148310)

[Group S06 67](#_Toc44148311)

[DISCUSSION AND CONCLUSIONS 74](#_Toc44148312)

[REFERENCES 75](#_Toc44148313)

### INTRODUCTION

Given six variables separated into six different groups that convey information over a series of time, our task is to create models that forecast futures values for these groups and variables. The variables might represent daily sales or prices for different segments of a business or industry. We’re not sure. Based on the gaps in the data in a column called “SeriesInd,” or Series Indicator, our assumption is that this data captures weekday values while skipping weekends. There are some instances of missing values for what we presume to be weekdays. Values for absent days will be handled using a method called value imputation, which approximates missing values based on present neighboring ones.

As consumers of this report will have both technical and non-technical backgrounds, we hope to not overwhelm readers with forecasting jargon. At a high level, we’ve sought to examine how the individual variables within groups have changed over time and use that analysis to predict future values for the next 140 periods, which may represent six months of weekdays.

The clearest path to communicating technical information successfully to non-technical staff is to provide real-world examples using workflows and business processes that end users understand. In this case, as we are not sure of what exactly we’re forecasting, that is not possible. Please bear with us and do not become overwhelmed by technical details. At a high level, we explore the data we have, attempting to understand both trends and summary values such as averages and medians. We find and handle the aforementioned missing values. Outlier values that differ significantly from neighboring values from the same variable are reviewed and possibly excluded, as they might represent data-entry errors or other invalid values.

Next, we examine the series data for aspects like seasonality. Do we find significant increases at regular intervals that reflect a certain day of the week or part of the year – like you might see with retails sales at Christmas? Understanding patterns in the data helps us predict where it might go in the future. Finally, we attempt to construct models that capture how the data change over time. After finding the best model for the variables in the groups, we then project that model forward for 140 periods.

Finding a model that captures a data set’s change over time can be difficult. We use three different methods that are explained slightly more later in this document. At a high level, we considered time series decomposition, exponential smoothing, and ARIMA models, each of which can sometimes be the best model for a dataset based on its characteristics.

We found that ARIMA models most often best captured our series data over time followed by exponential smoothing models. Sometimes our data had to be transformed using methods such as logs to create the most accurate models.

Why is forecasting important? Having an accurate idea of future activity like sales allows for an organization to optimize planning for functions such raw materials purchases, sales staffing, and marketing expenditures. Note that the best applications of data science are paired with a deep understanding of the business process being analyzed, which was not feasible in this case.

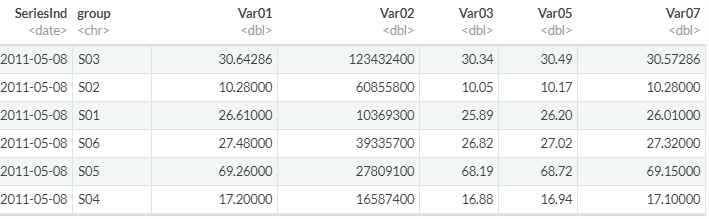
### ABSTRACT

* In this project work we have data set which is grouped in six different groups, each of which includes six variables. Our objective it to predict two of the variables for the next 140 periods.
* We analyzed the data using Time Series analytical techniques.
* We will identify if these times series meet the 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 by comparing their AIC and MAPE scores
* Finally, we have stored the predicated value in the Excel file for next 140 points.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| KEYWORDS  * **Time Series:** 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 summary of data: Data has groups from S01 to S06, and Variables Var01, Var02, Var03, Var05, Var07. * We will predict following variables from 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.  SUMMARY Here we have included all the models used and their summary for quick view:   |  |  |  |  |  | | --- | --- | --- | --- | --- | | Group /Var | Series /Model | RMSE | AIC | MAPE | | S01 – VAR02 | Series: dt\_s01\_v2\_xts | 2709242 | 76361.69 | 0.2020669 | | ARIMA(1,1,2) | |  | | S01 – VAR02 | Series: log(dt\_s01\_v2\_xts) | 0.01001024 | 229.29 | 0.1812568 | | ARIMA(1,1,1) | |  | | S02 – VAR02 | ARIMA model | 14704115 | 58113.31 | 0.247144897 | | S02 – VAR03 | ARIMA model | 7789411 | 3795.87 | 0.021378604 | | S03 – VAR05 | ARIMA(0,1,1) with drift | 1.496811 | 5914.84 | 0.013225681 | | | S03 – VAR05 | log(tsclean(dt\_s03\_v5\_xts)) | 0.01375072 | -13480.84 | 0.009118772 | | ARIMA(0,1,2) with drift | |  | | S03 – VAR07 | ARIMA(1,1,0) with drift | 1.342269 | 5561.52 | 0.012237645 | | | S03 – VAR07 | log(tsclean(dt\_s06\_v7\_xts)) | 0.01278172 | -13824.59 | 0.008305979 | | ARIMA(2,1,0) with drift | |  | | S04 – VAR01 | ARIMA(1,1,1) | 0.3892143 | 2242.88 | 0.008355055 | | S04 – VAR02 | Series: log(dt\_s04\_v2\_xts) | 0.3076678 | 1142.09 | 0.214089174 | | ## ARIMA(2,1,2) | | S05 – VAR02 | ARIMA(1,1,2) | 302903.7 | -5242.21 | 0.1721 | | Box Cox transformation: lambda= -0.09436074 | | S05 – VAR02 | ETS(A,N,N) | 263534.8 | 2219.804 | 0.1755 | | Box-Cox transformation: lambda= -0.0944 | | S05 – VAR03 | ARIMA(0,1,3) | 0.8849026 | 17683.02 | 0.0076 | | ## Box Cox transformation: lambda= 1.944948 | | S05 – VAR03 | ETS(A,N,N) | 0.8929011 | 25109.3 | 0.0077 | | Box-Cox transformation: lambda= 1.9449 | | S06 – VAR05 | Series: Var05\_ets | 0.565 | 5824.118 | 0.01130514 | | ETS(A,N,N) | | S06 – VAR05 | Series: log(tsclean(dt\_s06\_v5\_xts)) | 0.01202235 | -14106.61 | 0.008784268 | | ARIMA(3,1,3) | |  | | S06 – VAR07 | Series: Var07\_ets | 0.559 | 6669.095 | 0.011383246 | | ETS(A,N,N) | | S06 – VAR07 | Series: log(tsclean(dt\_s06\_v7\_xts)) | 0.01227619 | -14010.39 | 0.008840895 | | ARIMA(2,1,3) | |  | |  |  |

### DATA EXPLORATION

Actual data adjusted by date for better visualization can be seen below in figure 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.

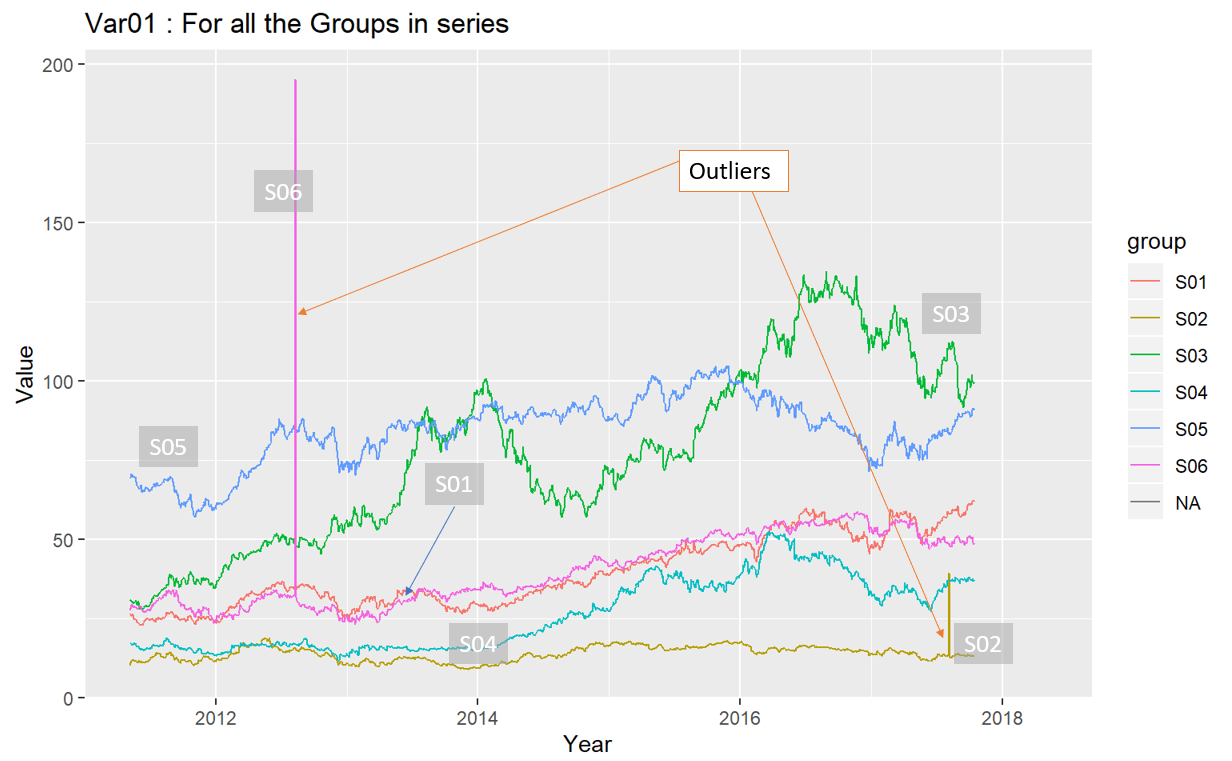
As you can see that this data set is very complex to do any time series analysis in the given form, lets break the data in the readable from for each group, and then by each variable along with date.

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 |

### Group S01

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.

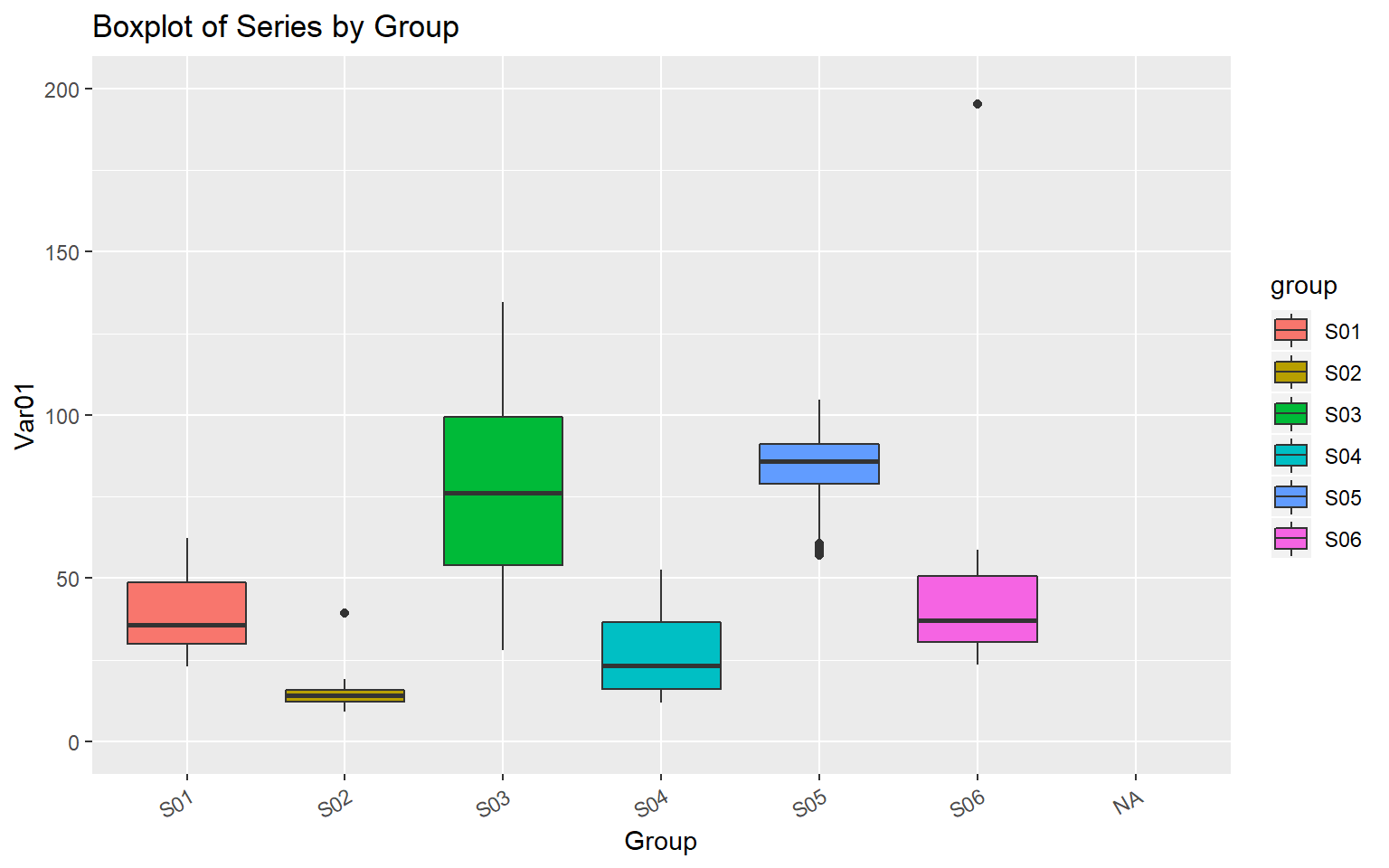


Figure 7 : Boxplot for var01 across groups

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.

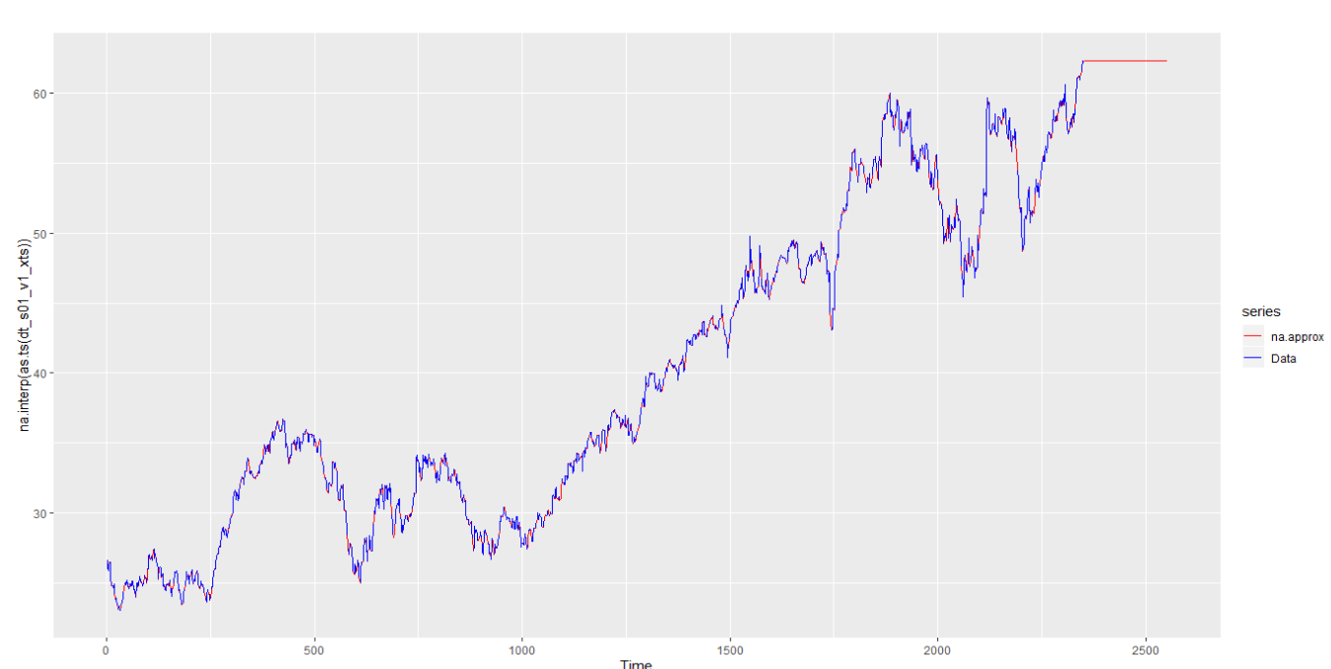


Figure 8 a. Missing value and actual data of Var01 from S01 group

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 |

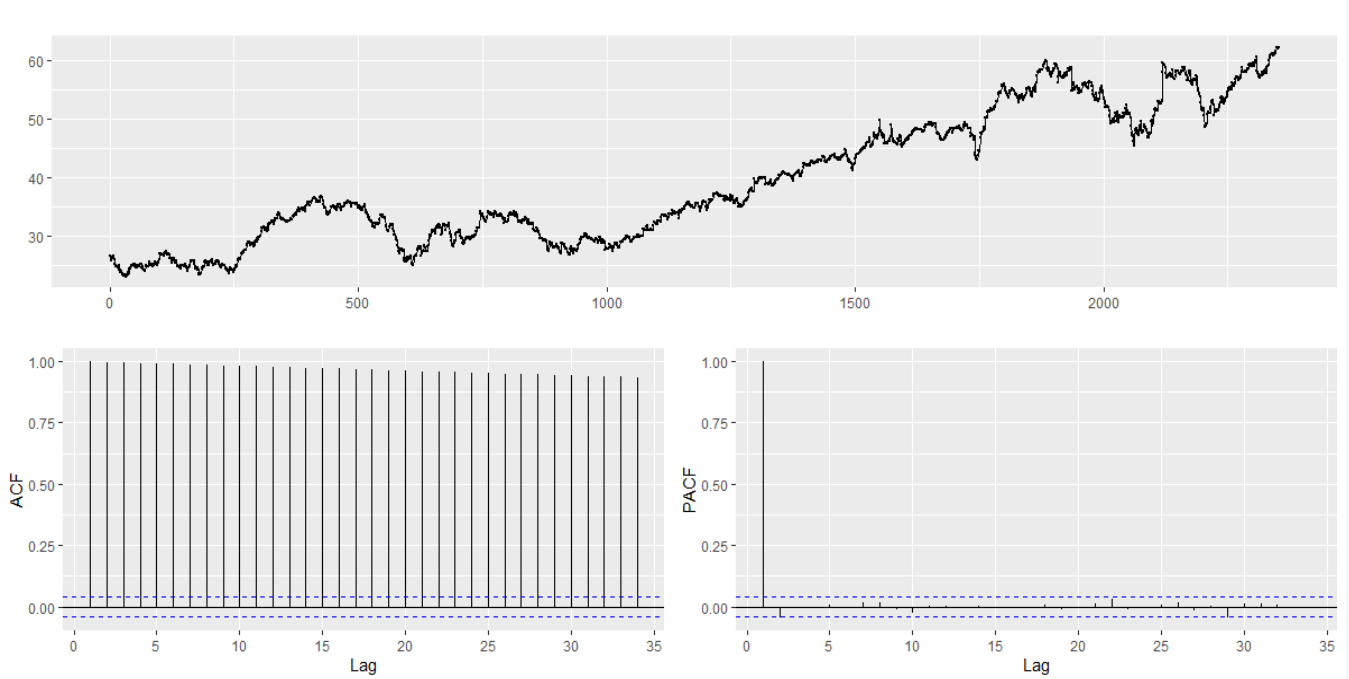
Seasonality: We created the lag of Var01 for 365 data points, to see yearly pattern. Below graph shows how the trend and Seasonally adjusted after removing trend are going moving.

Since there is no major data deviation between Seasonally Adjusted data and actual data, we think Seasonality is not present in the data.



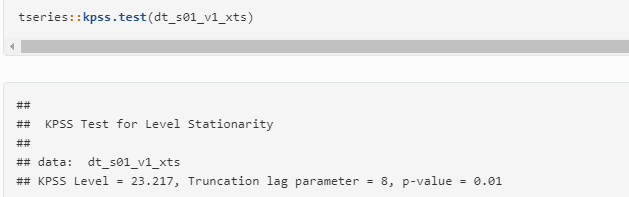
ACF and PACF:

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.



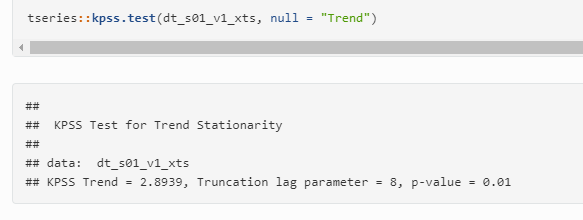
Plot of Var01 from S01 group

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

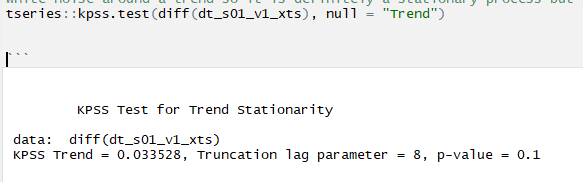


Above test suggest that p value is less than .05 and hence we reject the null and say that data is not stationary around the Level/trend/Mean line.

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.

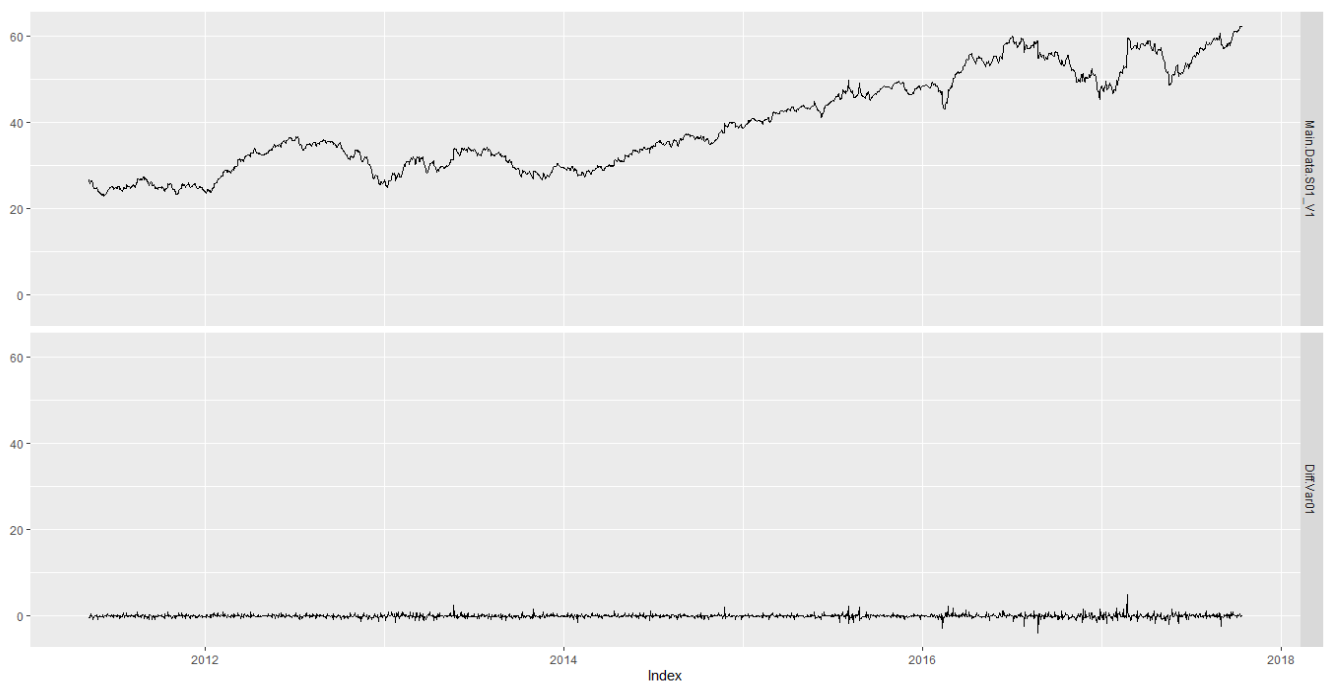
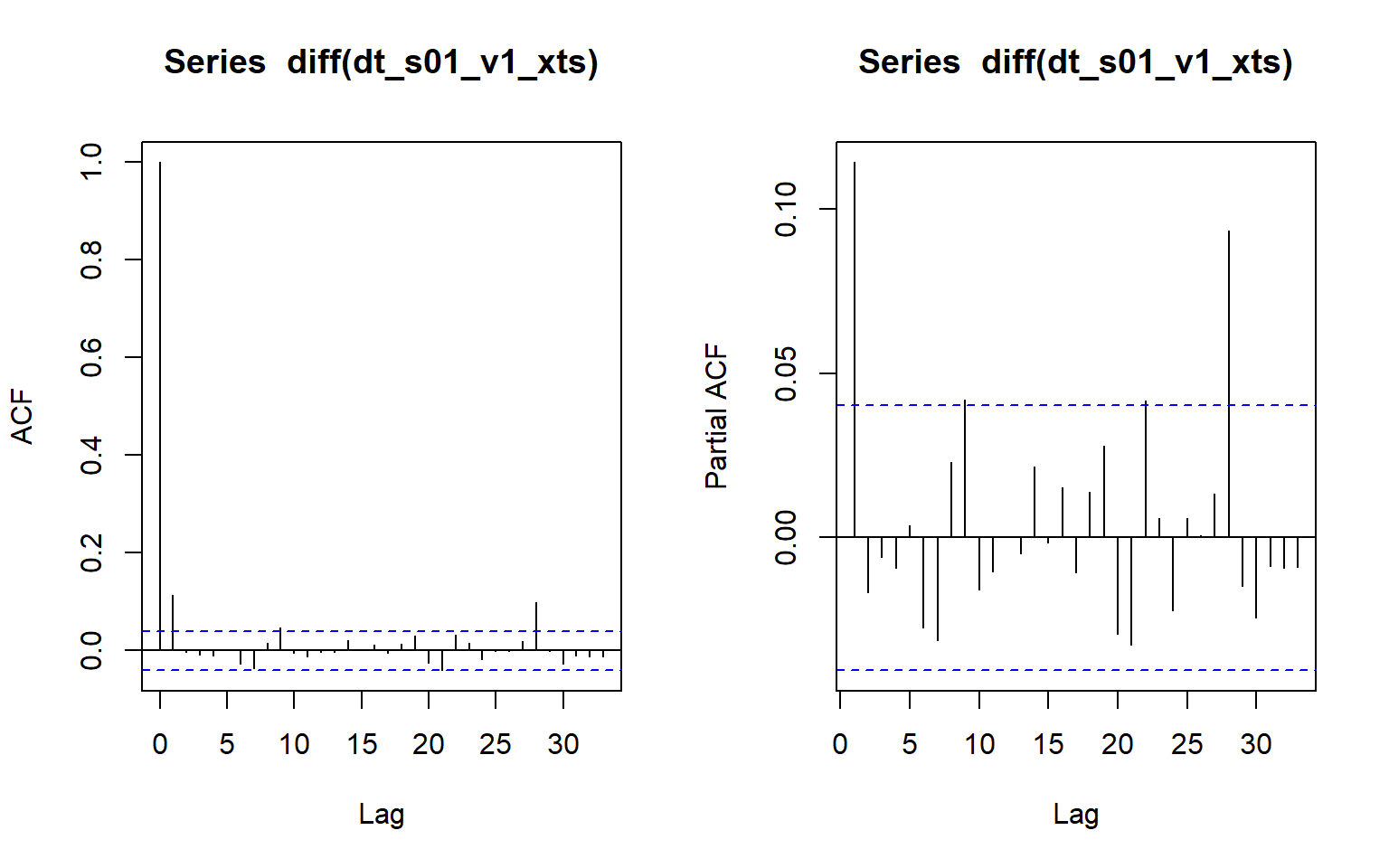


Figure : Var01 from s01 group

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

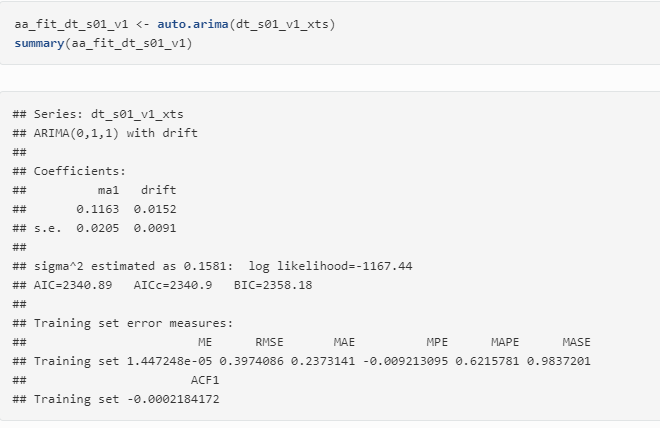
### 



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.

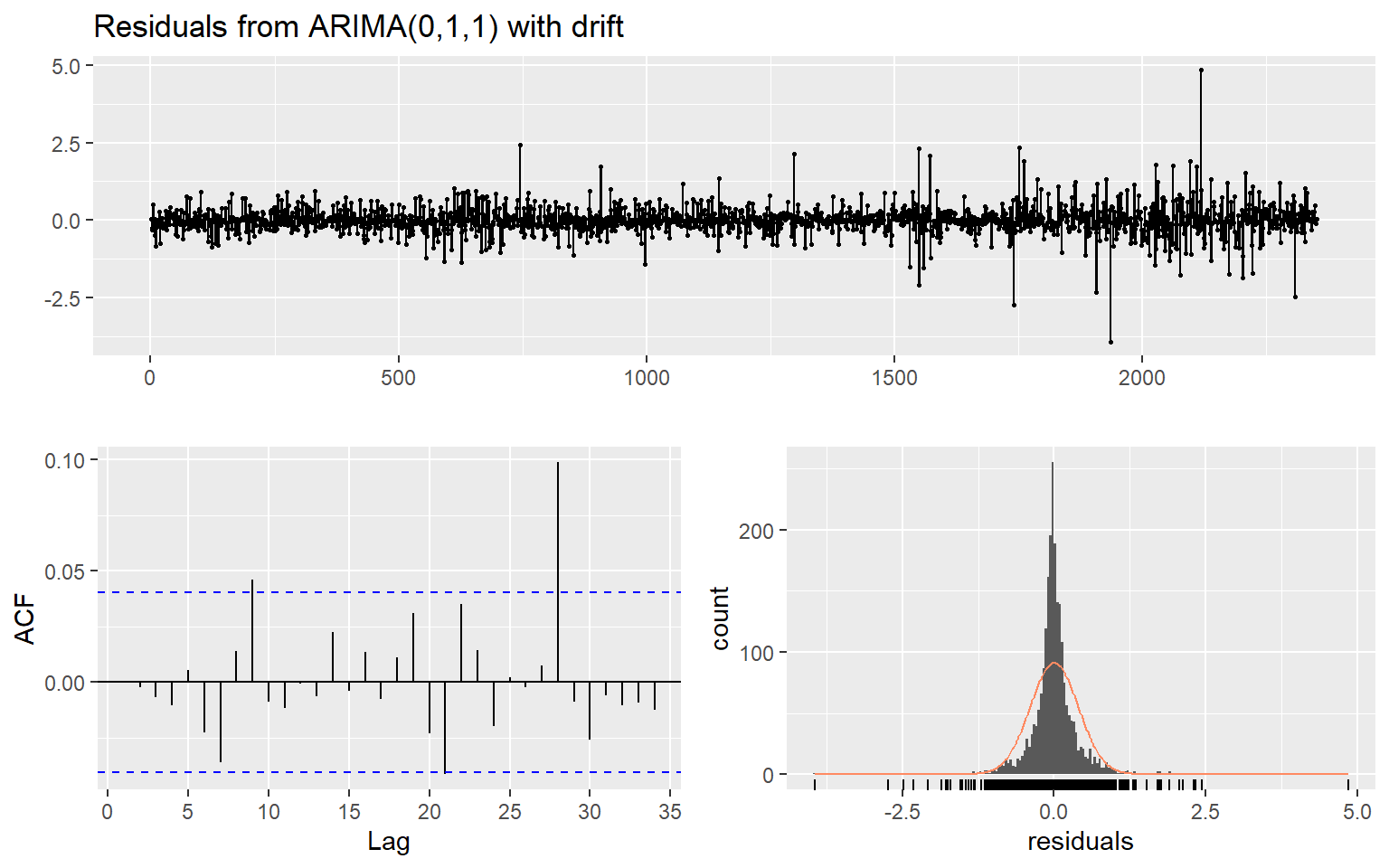


Auto.arima result for Var01

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 :

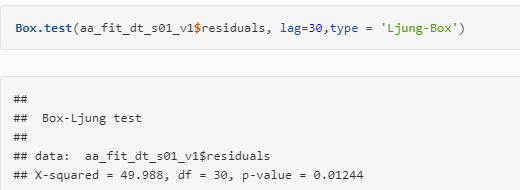
## ma1 drift   
## 0.11626944 0.01517711

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.



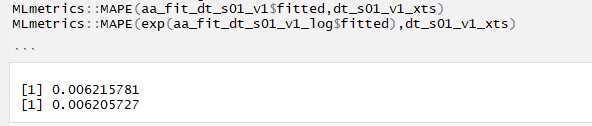
ACF plot of the residuals are white noise, as no prominent patterns can be seen here.

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



Lets us this model to forecast next 140 data points, we have also done log transformation of the Var01, before applying ARIMA on the data. Below plots shows the output from these two models.

|  |  |
| --- | --- |
|  | Series: dt\_s01\_v1\_xts  ARIMA(0,1,1) with drift  Coefficients:  ma1 drift  0.1163 0.0152  s.e. 0.0205 0.0091  sigma^2 estimated as 0.1581: log likelihood=-1167.44  AIC=2340.89 AICc=2340.9 BIC=2358.18 |
|  | Series: log(dt\_s01\_v1\_xts)  ARIMA(1,1,0) with drift  Coefficients:  ar1 drift  0.1117 4e-04  s.e. 0.0205 2e-04  sigma^2 estimated as 0.0001003: log likelihood=7491.16  AIC=-14976.31 AICc=-14976.3 BIC=-14959.02 |

Calculating MAPE of the Fitted value : 

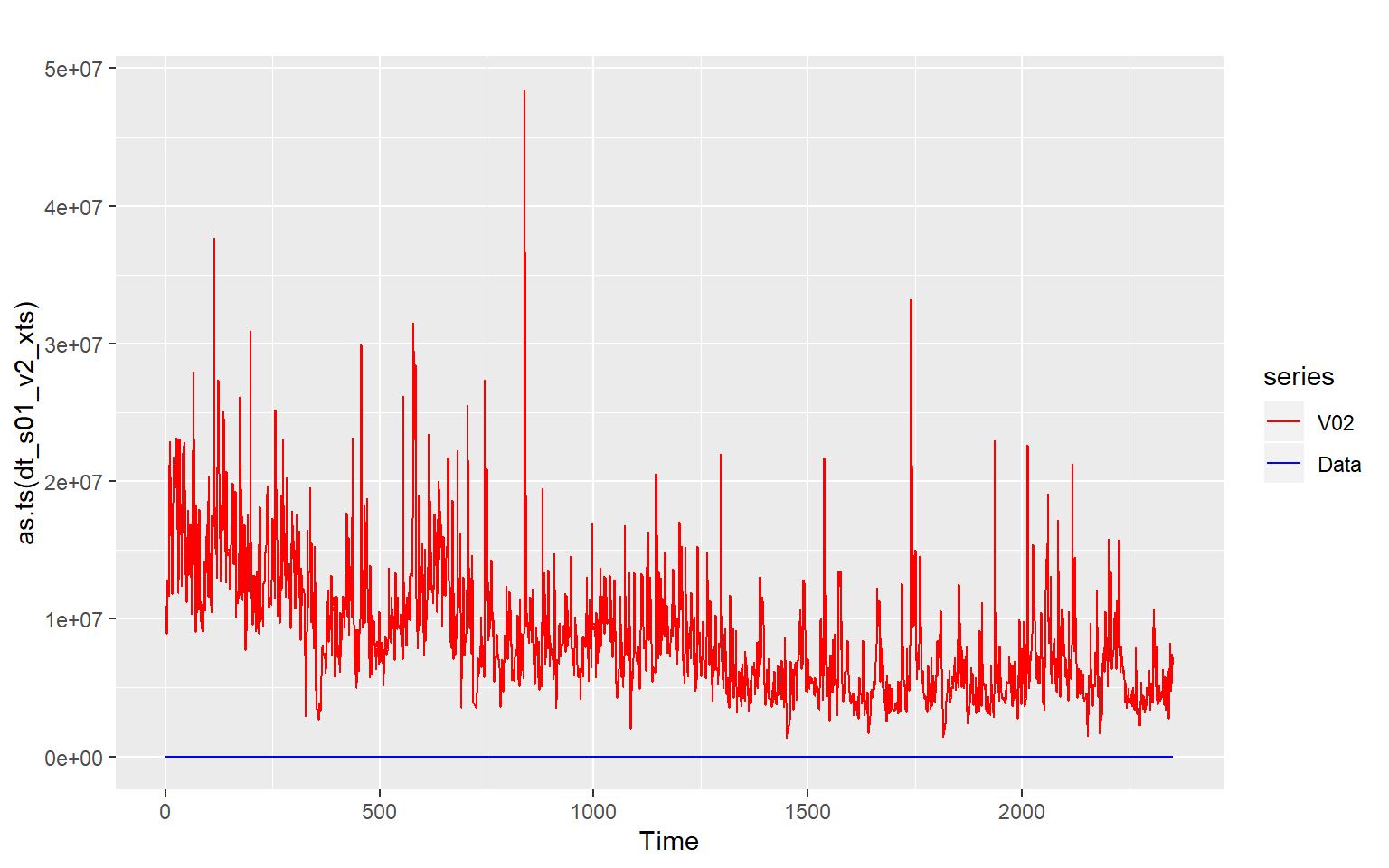
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group /Var | Series /Model | RMSE | AIC | MAPE |
| S01 – VAR01 | ARIMA(0,1,1) with drift | 0.3974086 | 2340.89 | 0.006215781 |
| S01 – VAR01 | Series: log(dt\_s01\_v1\_xts)  ARIMA(1,1,0) with drift | 0.01001024 | -14976.31 | 0.006205727 |

We don’t see much difference in the MAPE score of ARIMA(0,1,1) and ARIMA(1,1,0) with drift, AIC score seem to favor log transformation of the data before applying ARIMA. We will forecast next 140 for submission using ARIMA(1,1,0) with drift.

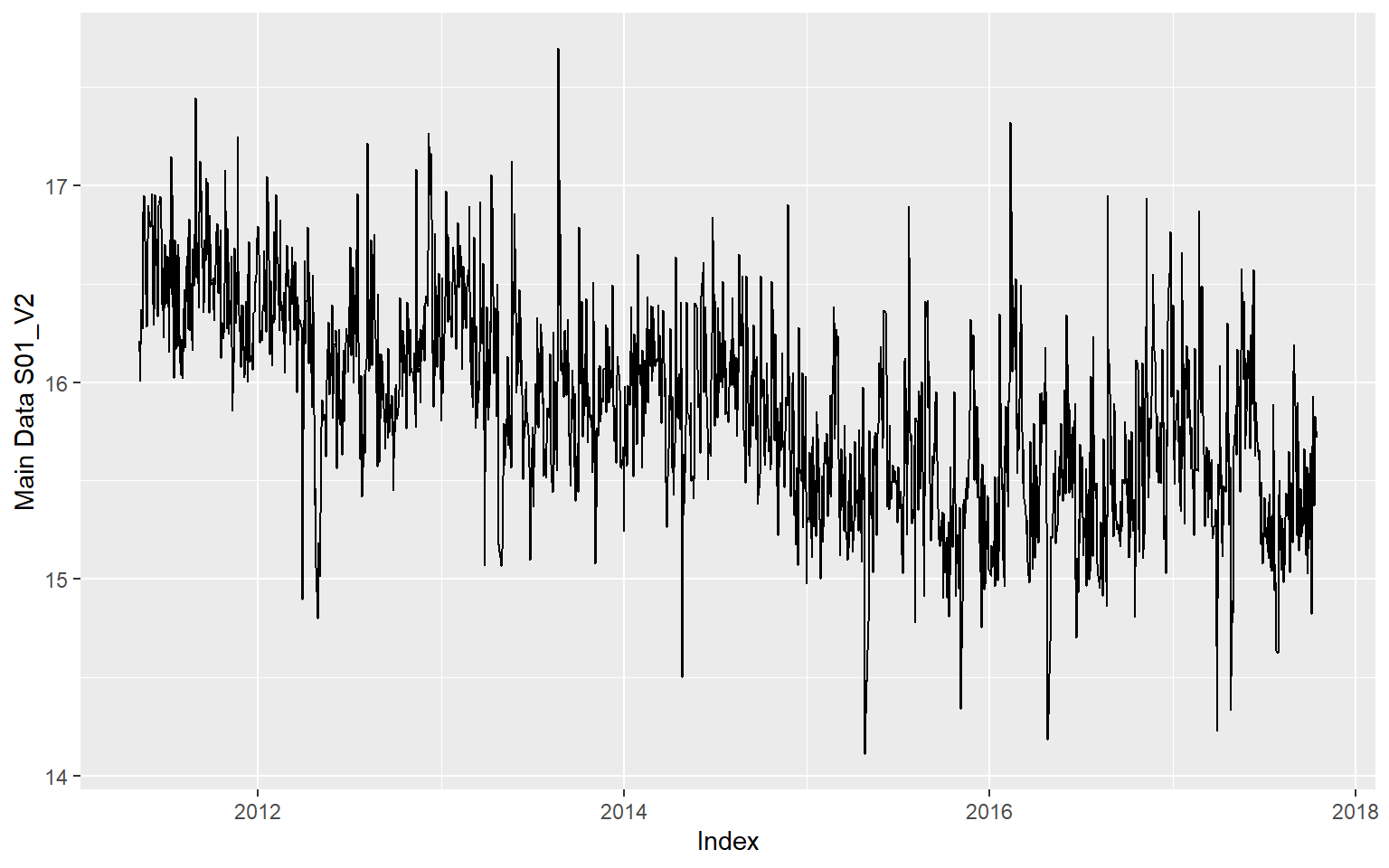
##### Group S01 - Var02:

As we have noted above the Var02 is very noisy a and is not following any trend. It’s mostly a straight line or always coming back to its mean. We plotted the log and actual data of he var02 in the below plot:

**autoplot**(**as.ts**(dt\_s01\_v2\_xts) , series="V02") **+** **autolayer**(**log**(**as.ts**(dt\_s01\_v2\_xts)),series="Data")**+** **scale\_colour\_manual**(values=**c**("blue","red"), breaks=**c**("V02","Data"))

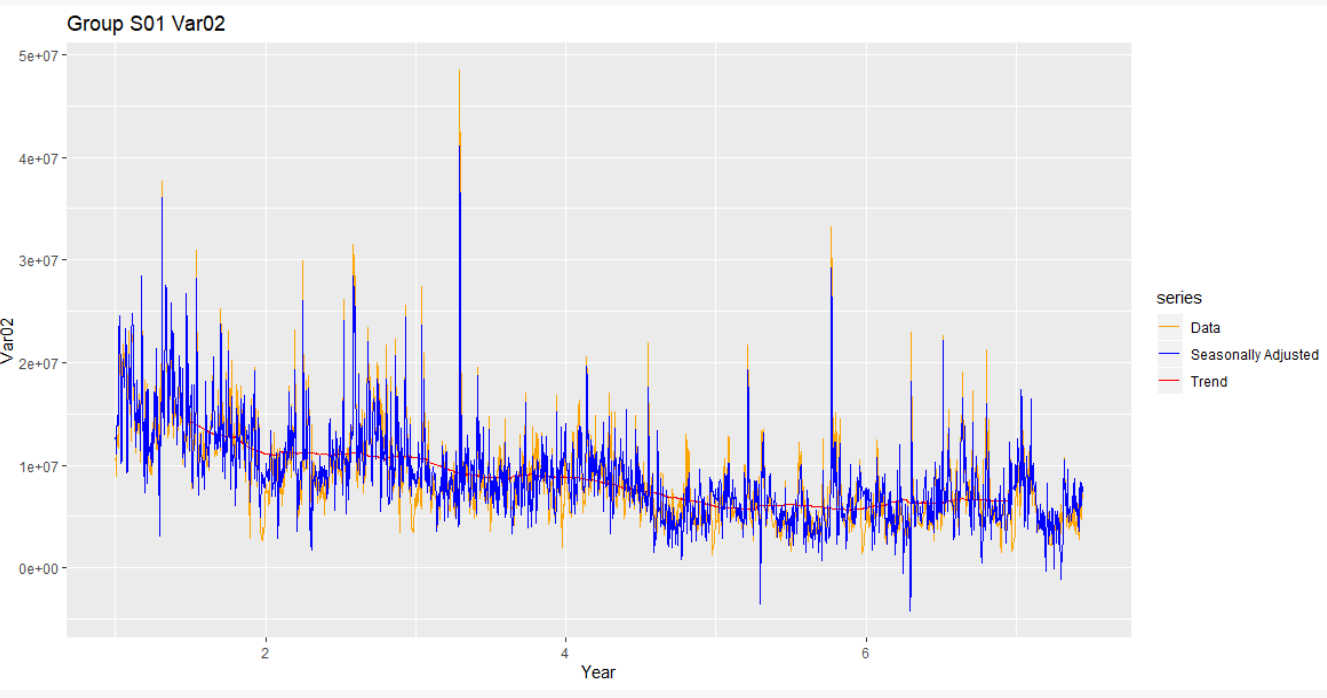


We think log transformation of the data before we apply time series model would be produce a better result. Below is the log transformed data for Var02.



Let’s check for seasonality of the data.

**Seasonality**: Looking at the yearly (lag 365 ) data of the var02, we hardly find any trend or seasonality.



Stationary Test:

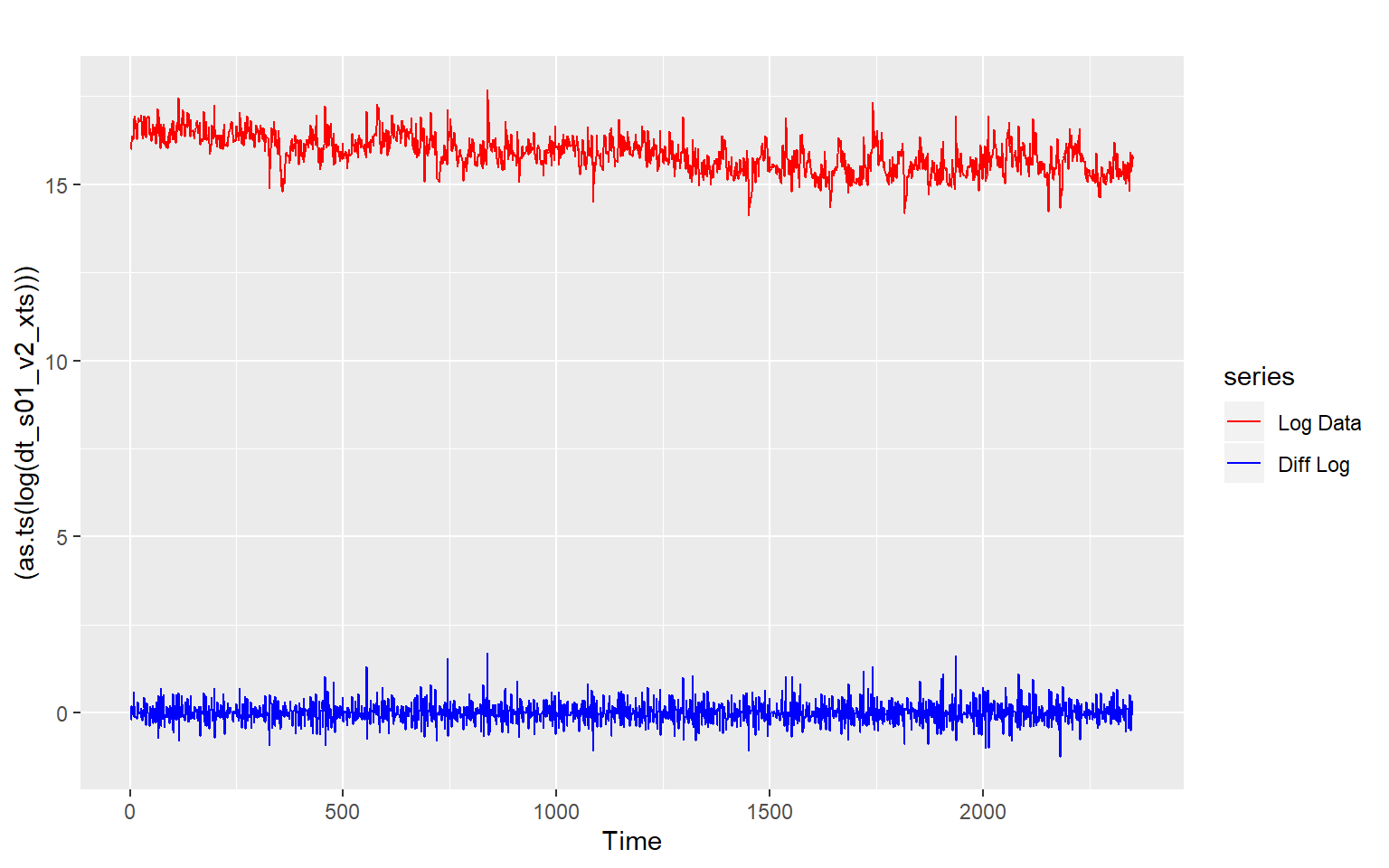
The null hypothesis for the test is that the data is stationary. The alternate hypothesis for the test is that the data is not stationary. The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test figures out if a time series is stationary around a mean or linear trend or is non-stationary due to a unit root. A stationary time series is one where statistical properties - like the mean and variance - are constant over time.

|  |  |
| --- | --- |
|  | The null hypothesis for the test is that the data is stationary.  Here you should reject the null hypothesis of stationarity as the value of the test statistic is more extreme than the 10%, 5% and 1% critical values (13.4497 > 0.347 , 13.4497 > 0.463, 13.4497 > 0.739).  Hence, we will reject the null and accept the alternate i.e. data is not stationary. |
| After doing Differencing: | We applied some one differencing to the data and then checked it for KPSS test.  Since the value of the test statistic = 0.003 is less than the 10%, 5% and 1% critical values (0.003 > 0.347 , 0.003 > 0.463, 0.003 > 0.739). We can say that data is now stationary. |
|  | As discussed above we think log transformed data can give better model. Lets test log differenced data for KPSS test.  We see no issue and our data is still stationary, as noted above without log difference model. |

This suggest that our data would be differenced once before we can apply our model. i.e. I(1).

Below plots shows how the log transformed data looks before and after differencing. One can see that data is very much close to its mean with log transformed.

**autoplot**((**as.ts**(**log**(dt\_s01\_v2\_xts))) , series="Log Data") **+** **autolayer**(**diff**(**as.ts**(**log**(dt\_s01\_v2\_xts))),series="Diff Log")**+** **scale\_colour\_manual**(values=**c**("blue","red"),breaks=**c**("Log Data","Diff Log"))



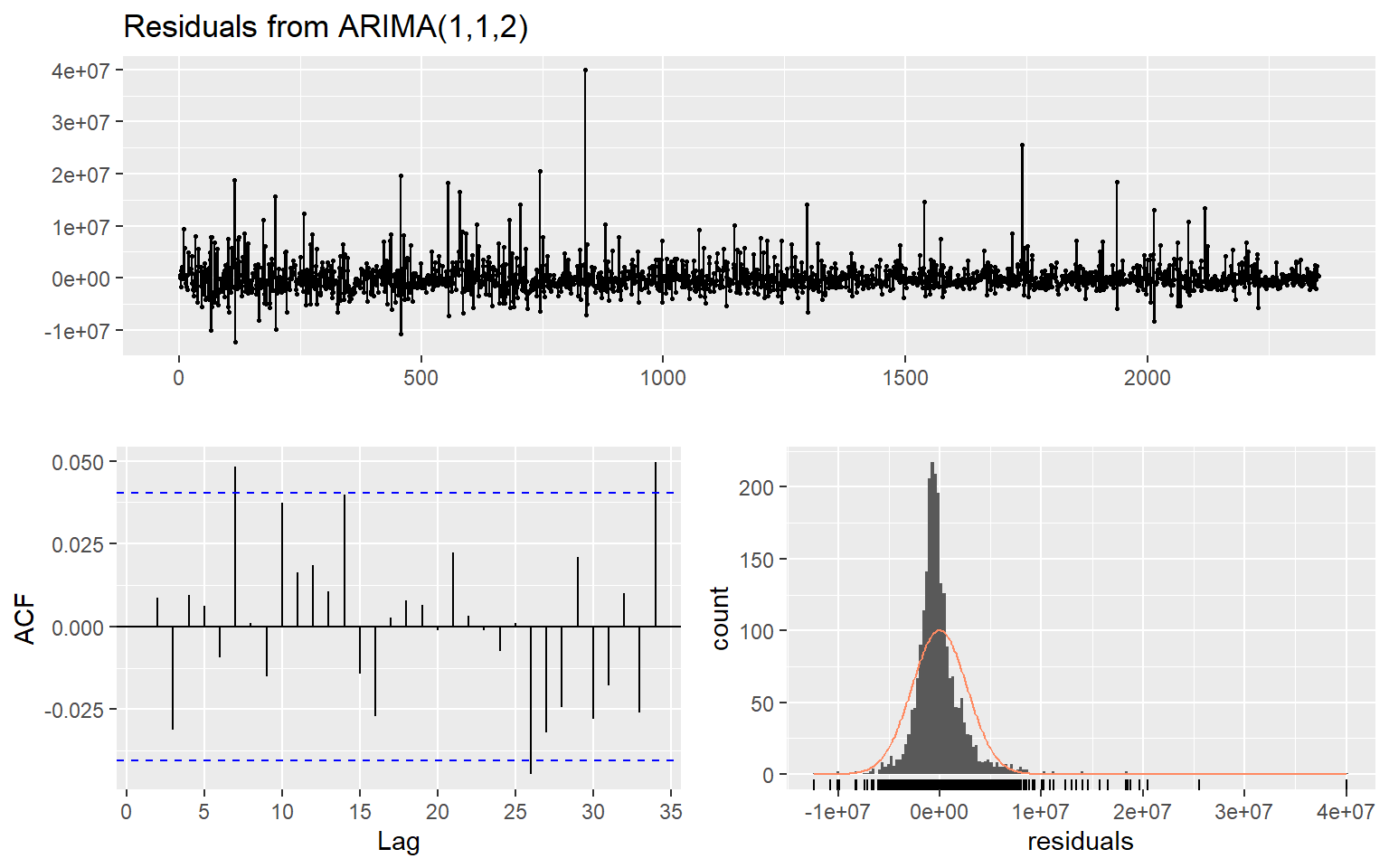
**ACF and PACF**

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\A186762D.tmp |  |
| PACF plot suggest too much variations which are not close to zero, Let’s try doing some log transaction and check. PACF plot also suggest Nonzero autocorrelations in the data from lag 1 to 35.  ACF plots shows that plots has some strong correlation in the first few lags then it is very much of random walk model, but slowly it exhibits Moderate to none Autocorrelation among other lags . |  |
| We will apply one difference on the logged data for var02. Below ACF and PACF plots suggest an mix model where we see AR(3) as ACF plot shows correlation till lag 3 before its goes down to zero. PACF model is little high in number of lags , as we don’t see it going down to zero before lag 18.  As you may note ACF plot suggest that after 3 lag data is mostly negatively auto correlated to zero mean. This suggest we can apply AR model to forecast some part of such data.  **Negative ACF** means that a positive value return for one observation increases the probability of having a negative value return for another observation (depending on the lag) and vice-versa. Or you can say (for a stationary time series) if one observation is above the average the other one (depending on the lag) is below average and vice-versa.  A negative autocorrelation changes the direction of the influence. A negative autocorrelation implies that if a particular value is above average the next value (or for that matter the previous value) is more likely to be below average. If a particular value is below average, the next value is likely to be above average. |  |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\DADEE525.tmp |  |

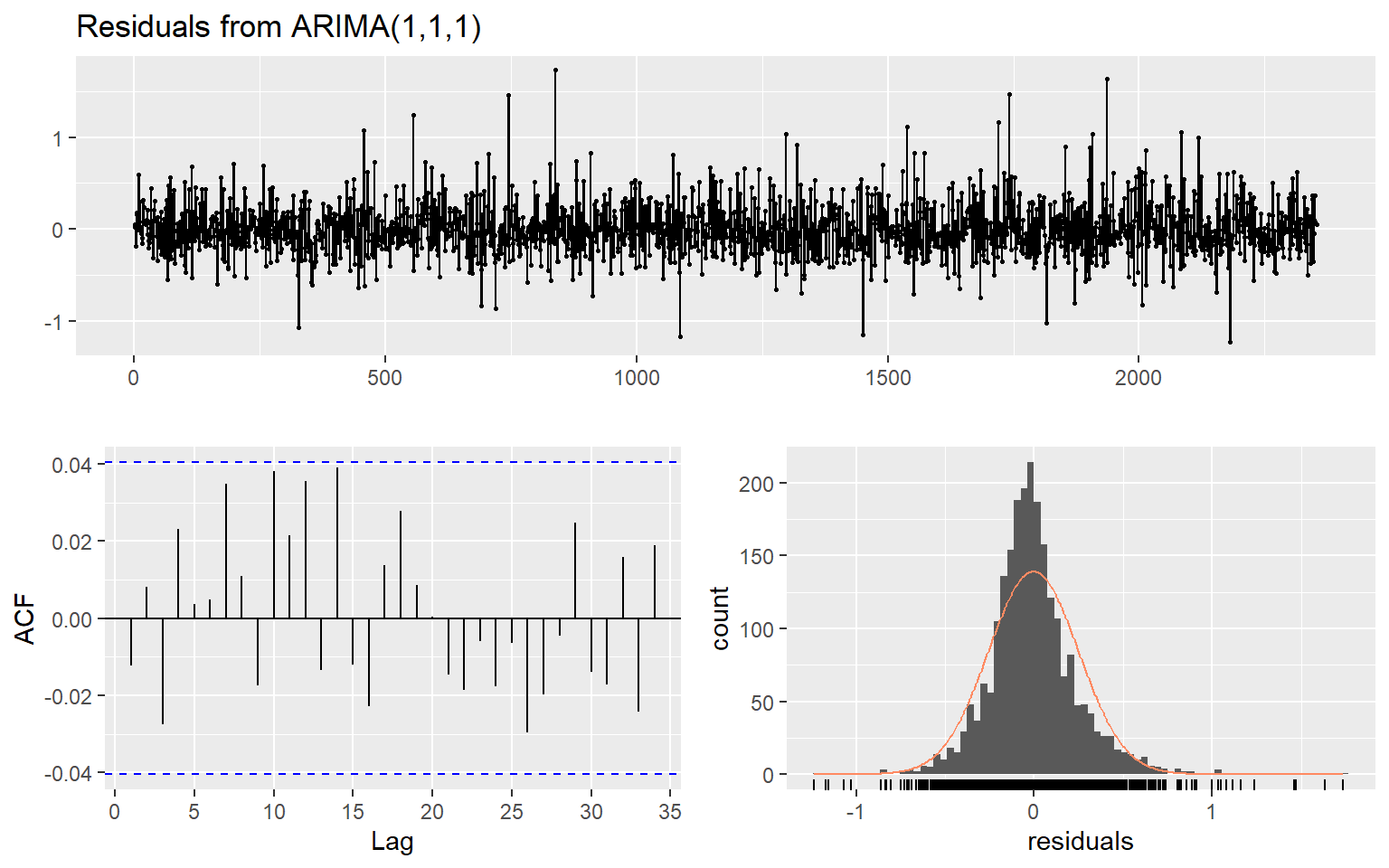
Based on our analysis we think ARIMA(3,1,0) may be best representative of this data. Lets test it with auto.arima.

|  |  |
| --- | --- |
| Here ARIMA(1,1,2) was fitted on the real data before we applied log to it .  AIC Score = 76361.69 |  |
| Here ARIMA(1,1,1) was fitted on the logged transformed data.  AIC Score = 229.29 |  |

AIC Score of Log transformed model is better so we will use it. Lets check the ressidul of both the model.



Residuals are giving very distributed errors and mostly close to zero. Let’s see the residual plot for Logged data:



Let’s do Ljung-Box test to see if data is white Nosie. Null hypothesis for this test is data is white noise.

|  |  |
| --- | --- |
|  | The Ljung-Box test also returned High p-value indicating that we can't reject the null hypothesis for model ARIMA(1,1,2) , and residuals data is white Nosie. |
|  | The Ljung-Box test also returned High p-value for logged data model ARIMA(1,1,1)indicating that we can't reject the null hypothesis, and residuals data is white Nosie. |

Now will forecast the result for next 140 data points, using both model.

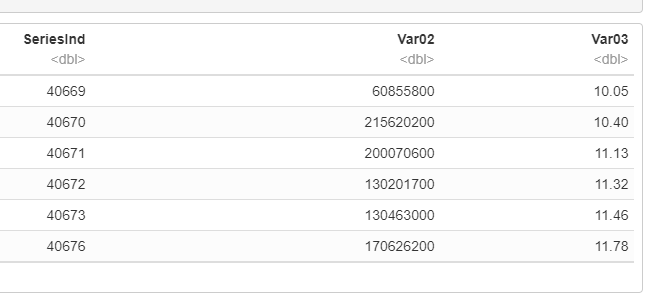
|  |  |
| --- | --- |
|  | Series: dt\_s01\_v2\_xts  ARIMA(1,1,2)  Coefficients:  ar1 ma1 ma2  0.6594 -1.0199 0.0448  s.e. 0.0300 0.0366 0.0329  sigma^2 estimated as 7.352e+12: log likelihood=-38176.85  AIC=76361.69 AICc=76361.71 BIC=76384.75 |
|  | Series: log(dt\_s01\_v2\_xts)  ARIMA(1,1,1)  Coefficients:  ar1 ma1  0.7196 -0.9763  s.e. 0.0188 0.0076  sigma^2 estimated as 0.0644: log likelihood=-111.65  AIC=229.29 AICc=229.3 BIC=246.58 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group /Var | Series /Model | RMSE | AIC | MAPE |
| S01 – VAR02 | Series: dt\_s01\_v2\_xts  ARIMA(1,1,2) | 2709242 | 76361.69 | 0.2020669 |
| S01 – VAR02 | Series: log(dt\_s01\_v2\_xts)  ARIMA(1,1,1) | 0.01001024 | 229.29 | 0.1812568 |

Since AIC score of logged models is lower we choose logged model for submitting the predicted score.

### Group S02

Lets see the S02 data and explore its components :



Statistical Analysis of the variables suggest Var02 has no missing values and Var03 has 4 missing values.

|  |  |
| --- | --- |
|  |  |

Exploring Data for S02 - var02:

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\8E57158.tmp | |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\846E6824.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\78CD926.tmp |

Looking at above plots fixed distribution assumption do not hold true, as histogram is not be bell-shaped, and the normal probability plot is not linear. it is right skewed. Box plot show many outliers. May be after suppressing outlier’s distribution plot will improve and series smoothens.

Analysis for S02 Var03:

|  |  |
| --- | --- |
|  | |
|  |  |

Looking at above plots fixed distribution assumption holds good for V03, as histogram is bell-shaped, and the normal probability plot is linear. Box Plot has one outlier.

ACF plots:

|  |  |
| --- | --- |
|  |  |

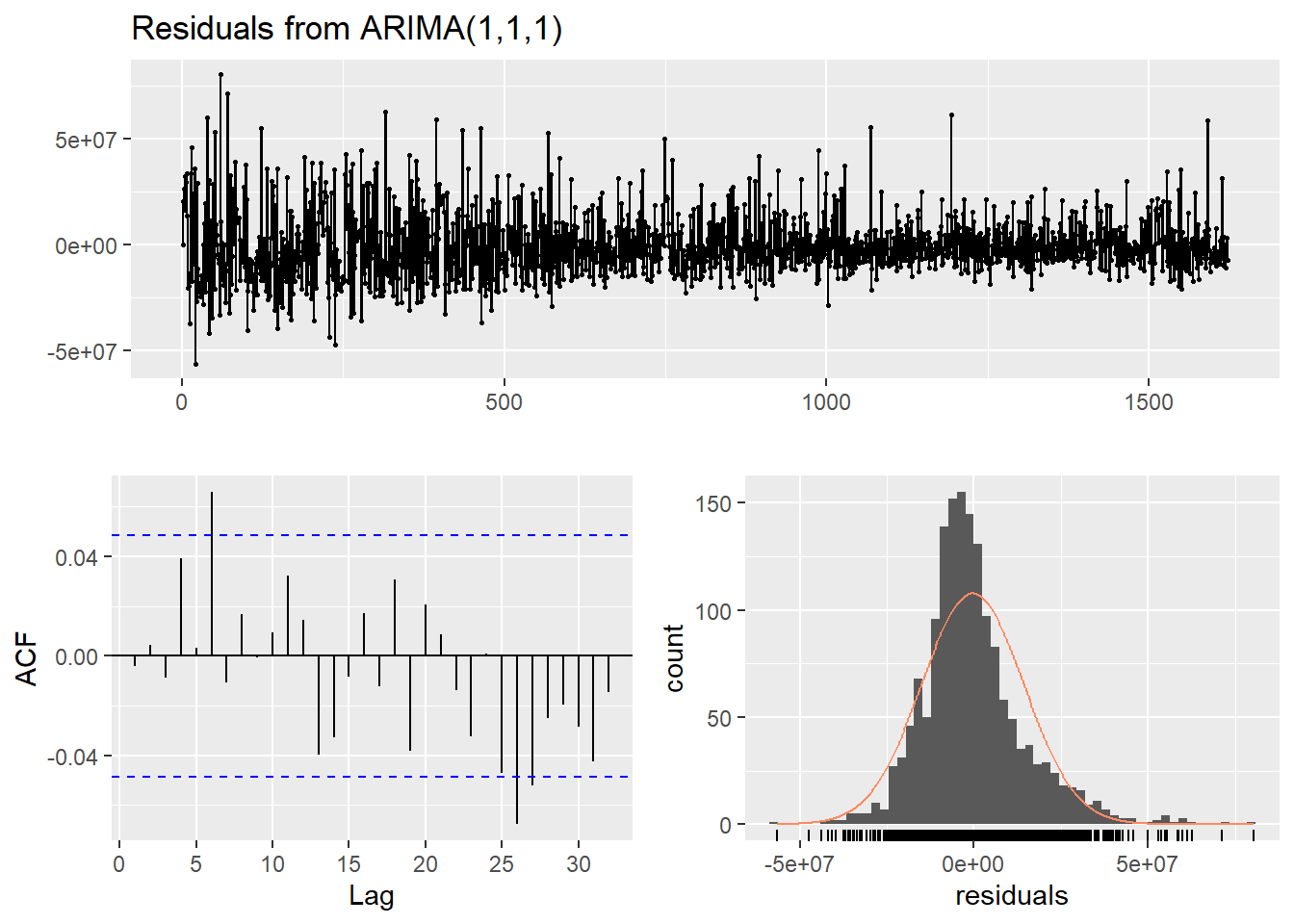
ACF plots shows that plots has some strong correlation in the first few lags then it is very much of random walk model, but slowly it exhibits Moderate to none Autocorrelation among other lags

##### Group S02 Var02

Applying Model Arima to see what is the best possible model that can predicate this data:

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\1E4915D2.tmp | ## Series: S02v2C  ## ARIMA(1,1,1)  ##  ## Coefficients:  ## ar1 ma1  ## 0.5075 -0.9502  ## s.e. 0.0283 0.0128  ##  ## sigma^2 estimated as 2.166e+14: log likelihood=-29053.65  ## AIC=58113.31 AICc=58113.32 BIC=58129.48 |

Residual Check:



|  |  |
| --- | --- |
|  | Ljung test suggest that data is white noise. |

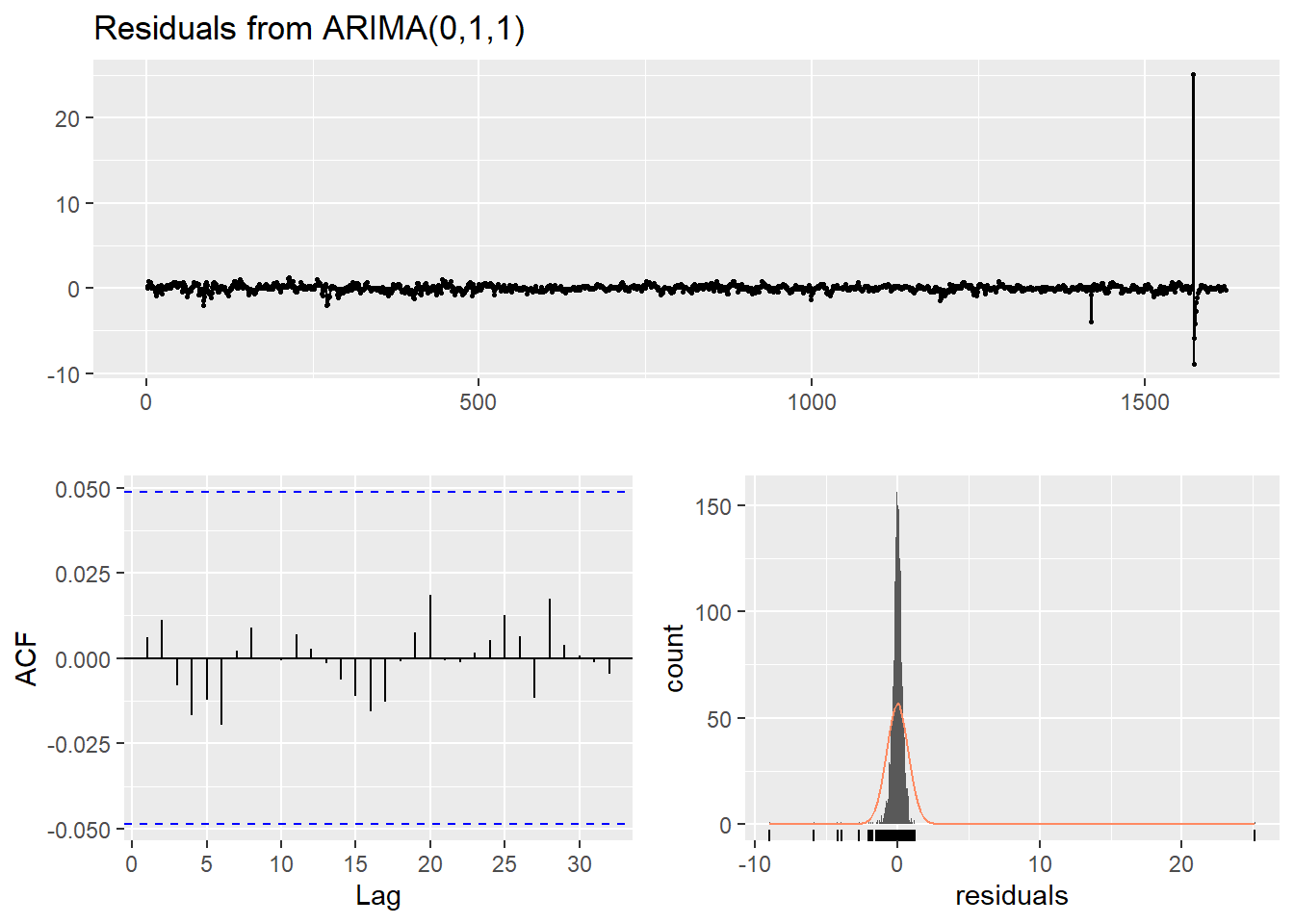
The auto.arima results with ARIMA(1,0,1) model with no drift .

##### Group S02 Var03

Lets apply auto.airma to var03 and analyze the result :

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\92F2A73E.tmp | ## Series: S02v3C  ## ARIMA(0,1,1)  ##  ## Coefficients:  ## ma1  ## -0.6687  ## s.e. 0.0191  ##  ## sigma^2 estimated as 0.6075: log likelihood=-1895.94  ## AIC=3795.87 AICc=3795.88 BIC=3806.66 |

Lets check the residual :



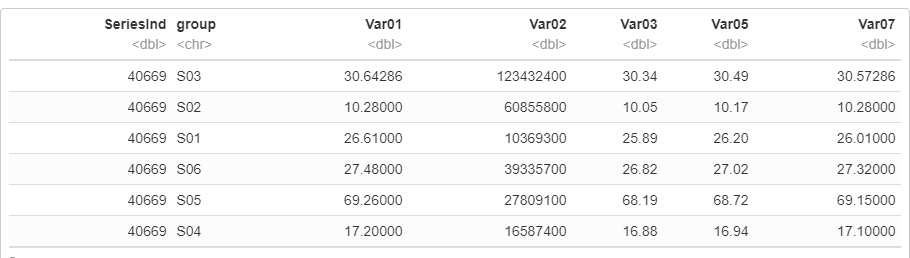
|  |  |
| --- | --- |
|  | A portmanteau test returns a large p-value 0.9938, also suggesting that the residuals are white noise. |

The ACF plot of the residuals from the ARIMA(0,1,1) model shows that all autocorrelations are within the threshold limits, indicating that the residuals are behaving like white noise

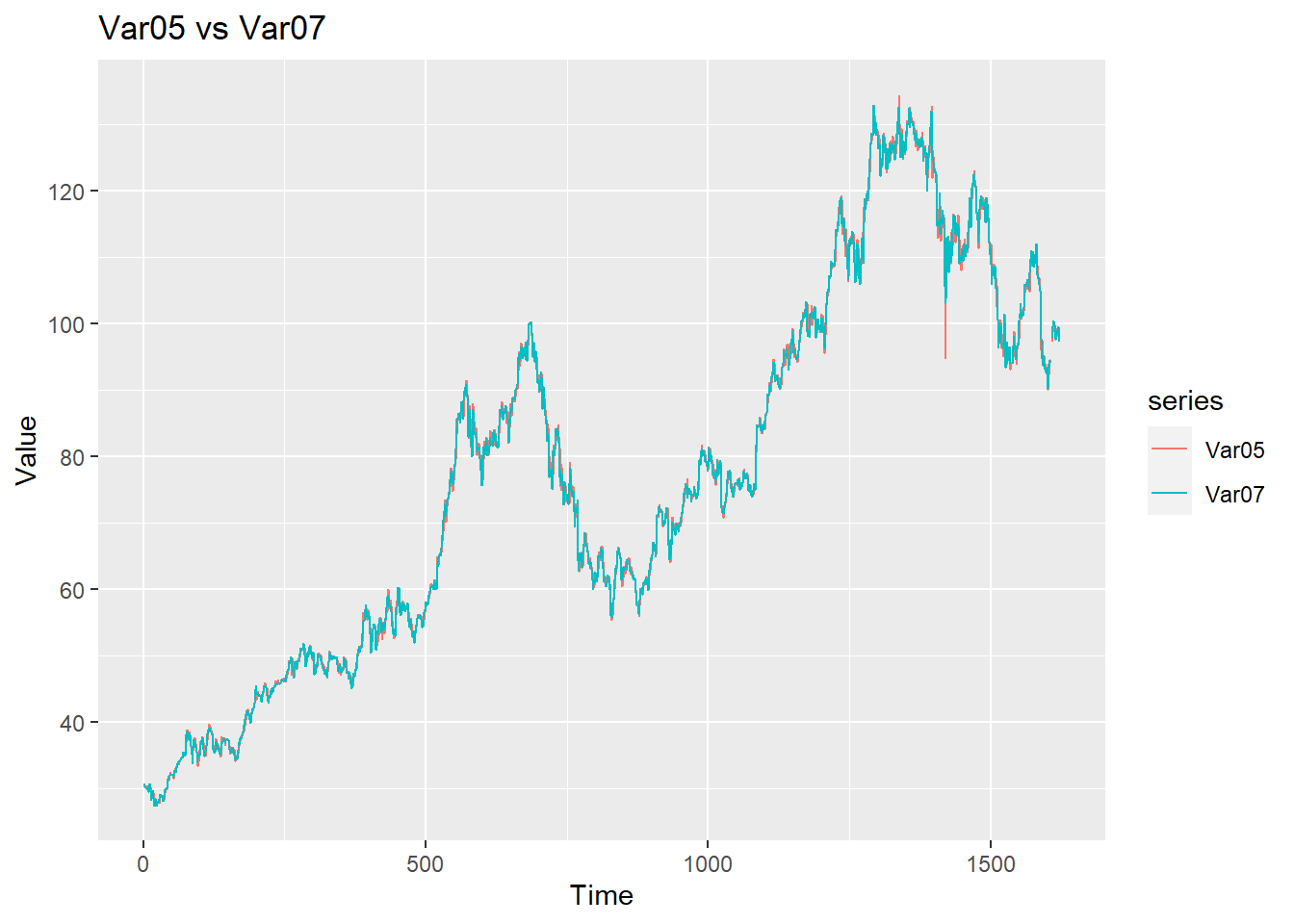
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group /Var | Series /Model | RMSE | AIC | MAPE |
| S02 – VAR02 | ARIMA model | 14704115 | 58113.31 | 0.247144896628619 |
| S02 – VAR03 | ARIMA model | 7789411 | 3795.87 | 0.0213786043674216 |

### 

### Group S03

Let’s see how data of group S03 look, we have created subset of Var05 and var07 for further analysis.   


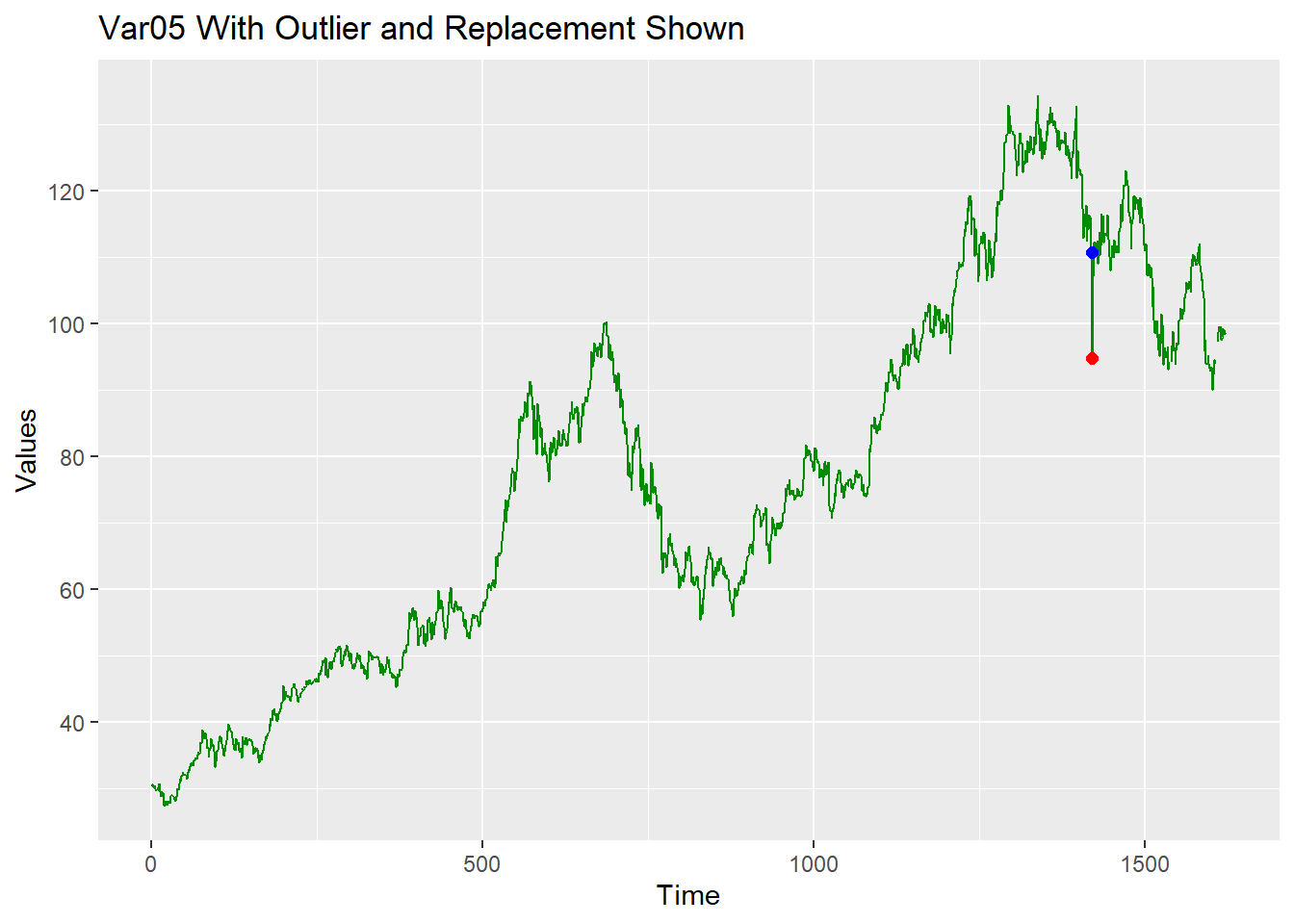
Visualization of variables Var07, and var05, suggest that most probably the forecasts for both series will be almost identical. The data shows strong trend but no evident seasonal pattern, which will be bases of our modelling.



|  |  |
| --- | --- |
| Var05 has 4 missing value, we have imputed the NA data , and let’s check the statistic after imputation : | Var07 has 4 missing values, we have imputed the NA data , and let’s check the statistic after imputation : |

##### Group S03 Var05

**Outliers:** The series show only one outlier between them which is shown in the above timesries plot. As, series are so similar and this is one of the defining differences of the series, We are not removing outliers.



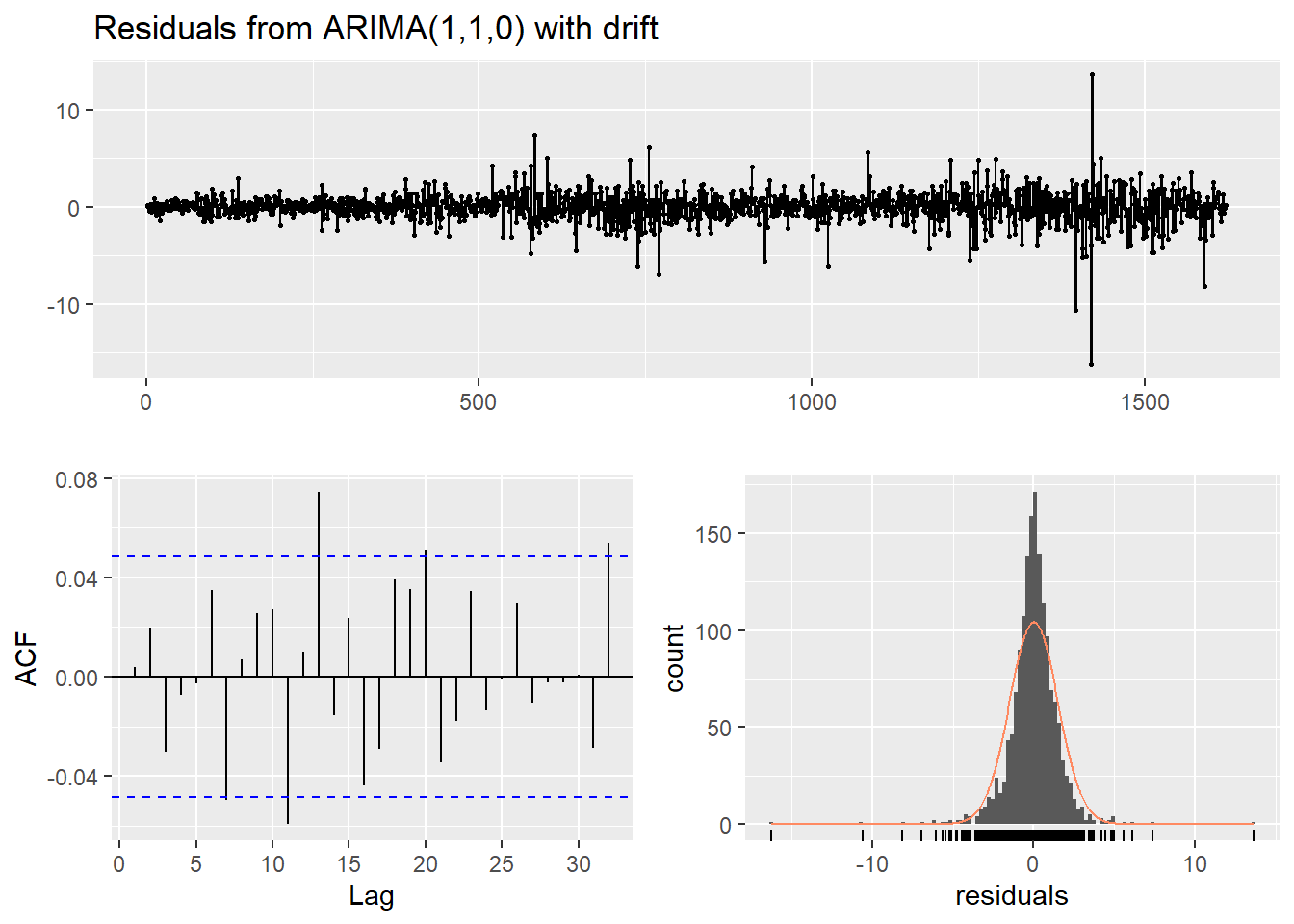
ACF for Var05 : Below graph show that data is clearly non stationary.

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\53988551.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\706A8487.tmp |

Applying ARIMA model: ARIMA(1,1,0) choice of model parameters is confirmed by data visualization.

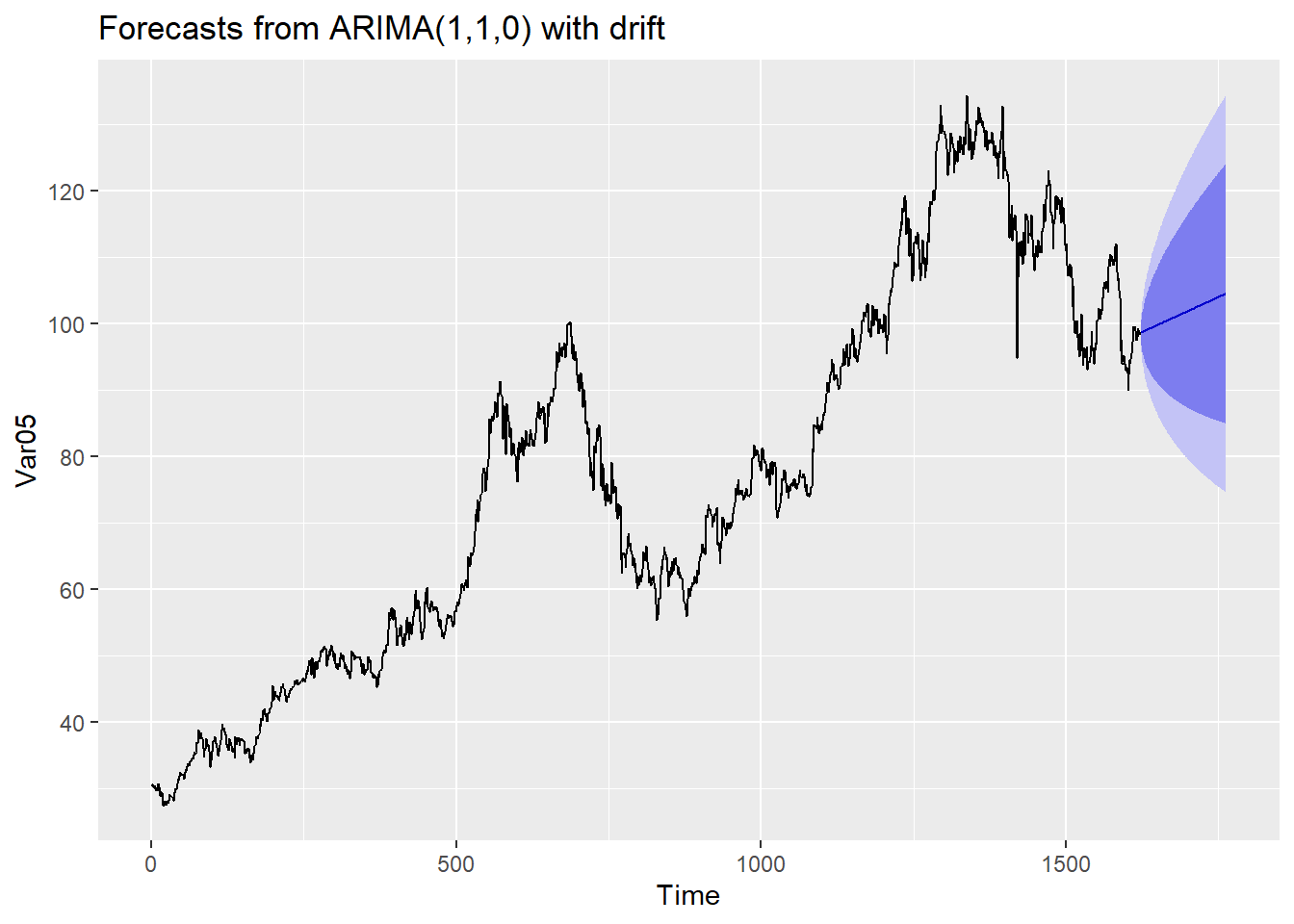
|  |  |
| --- | --- |
| Arima(1,1,0) with drift is used for the forecast. Summary of the mode as shown to the right of this para. | Series: s03v5  ARIMA(1,1,0) with drift  Coefficients:  ar1 drift  -0.1640 0.0421  s.e. 0.0245 0.0320  sigma^2 estimated as 2.245: log likelihood=-2954.42  AIC=5914.84 AICc=5914.86 BIC=5931.02  Training set error measures:  ME RMSE MAE MPE MAPE MASE ACF1  Training set 2.6104e-05 1.496811 1.001878 -0.009345759 1.322568 0.9899828 0.003934698 |

Residual analysis of the model:



|  |  |
| --- | --- |
|  | Pvalue 0.22 is greater than .05 and hence we can’t reject the null hypnosis here , which say that residual data is white noise.  The residuals are independent. |

Forecasted data:



##### Group S03 Var07

Data of Var07, show increasing trend but not confirmed seasonality.



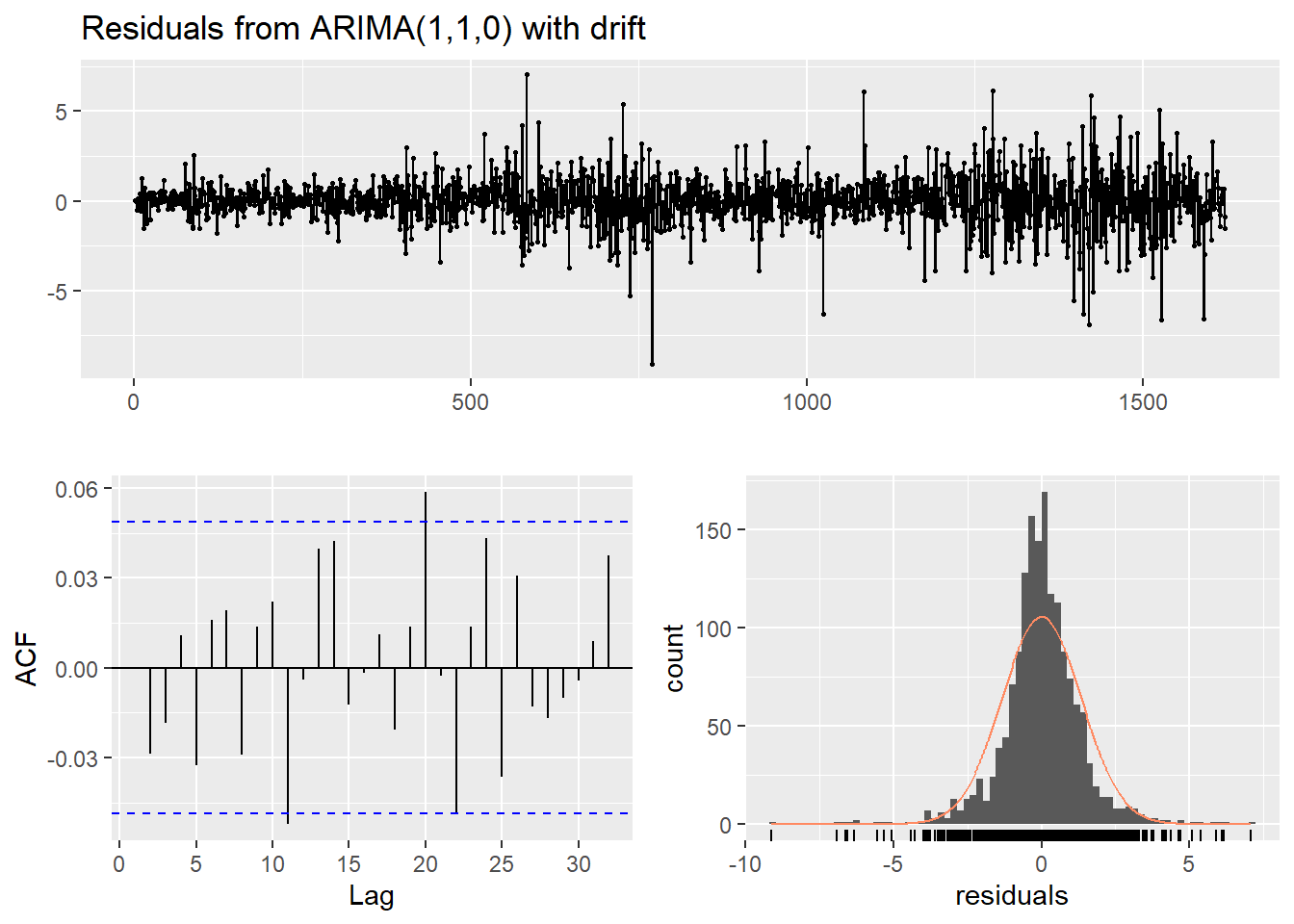
Lets check the ACF plots and difference data for var07:

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\9C93672D.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\F354AEC3.tmp |

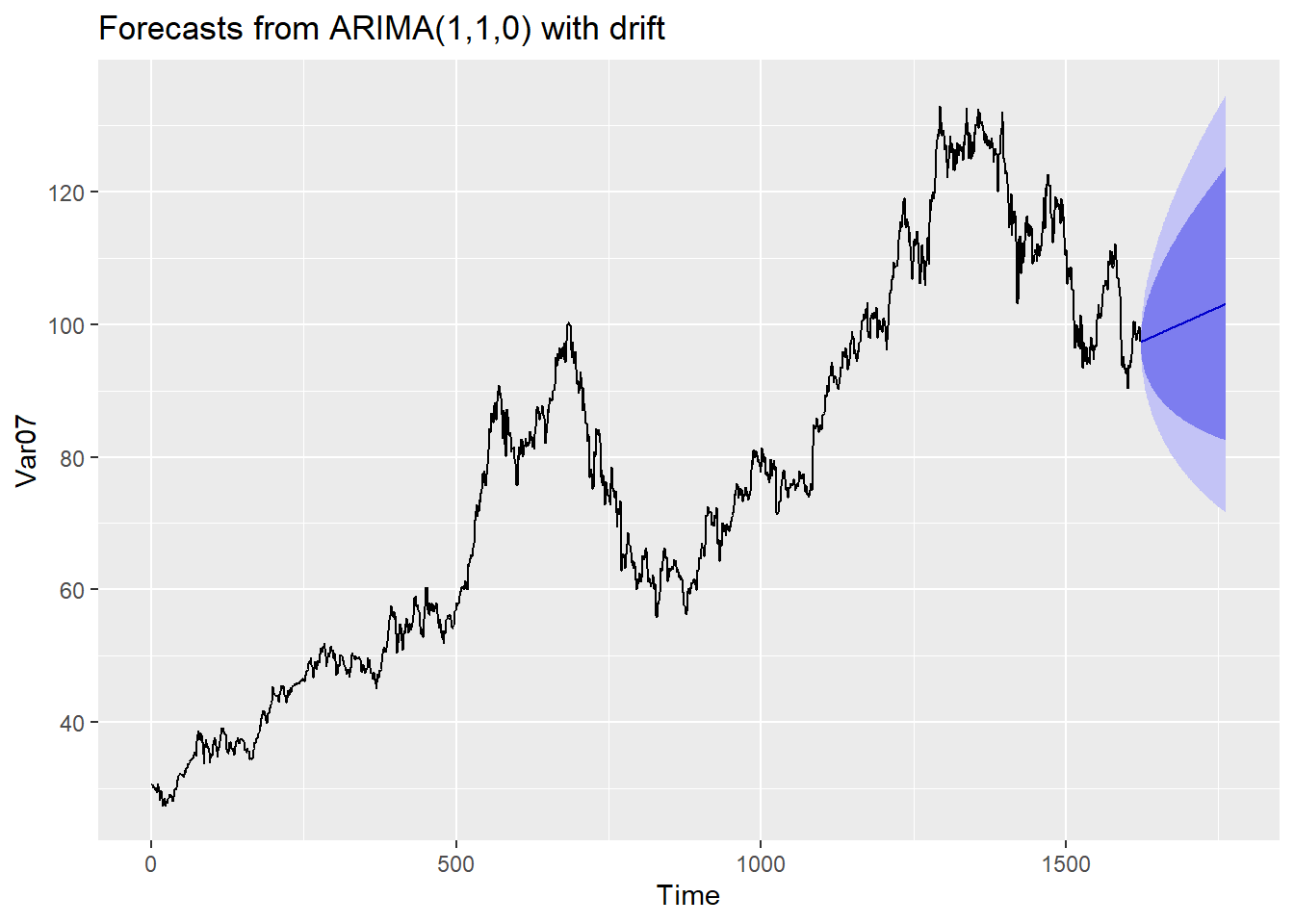
The new differenced data is stationary and ACF plots show its close to zero mean correlation .

|  |  |
| --- | --- |
| The new differenced data is stationary and therefore appropriate for ARIMA  modeling. | ARIMA(1,1,0) with drift is applied for forecast.  Series: s03v7  ARIMA(1,1,0) with drift  Coefficients:  ar1 drift  0.0109 0.0412  s.e. 0.0248 0.0337  sigma^2 estimated as 1.805: log likelihood=-2777.76  AIC=5561.52 AICc=5561.53 BIC=5577.69  Training set error measures:  ME RMSE MAE MPE MAPE MASE ACF1  Training set 1.074076e-05 1.342269 0.9292027 -0.005290655 1.223764 0.9990103 0.0003231799 |

Residual analysis of the model:



|  |  |
| --- | --- |
|  | Pvalue 0.5006 is greater than .05 and hence we can’t reject the null hypnosis here , which say that residual data is white noise.  The residuals are independent. |

Forecasted data using model:  


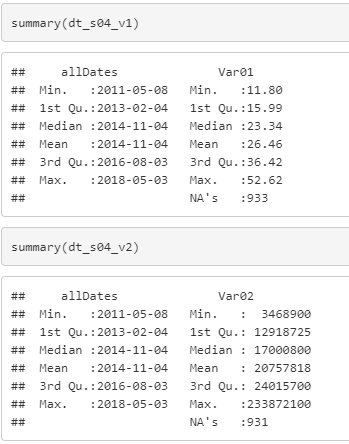
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group /Var | Series /Model | RMSE | AIC | MAPE |
| S01 – VAR05 | ARIMA(0,1,1) with drift | 1.496811 | 5914.84 | 0.0132256809974202 |
| S01 – VAR05 | log(tsclean(dt\_s06\_v5\_xts))  ARIMA(0,1,2) with drift | 0.01375072 | -13480.84 | 0.009118772 |
| S03 – VAR07 | ARIMA(1,1,0) with drift | 1.342269 | 5561.52 | 0.0122376449235163 |
| S03 – VAR07 | log(tsclean(dt\_s06\_v7\_xts))  ARIMA(2,1,0) with drift | 0.01278172 | -13824.59 | 0.008305979 |

### Group S04

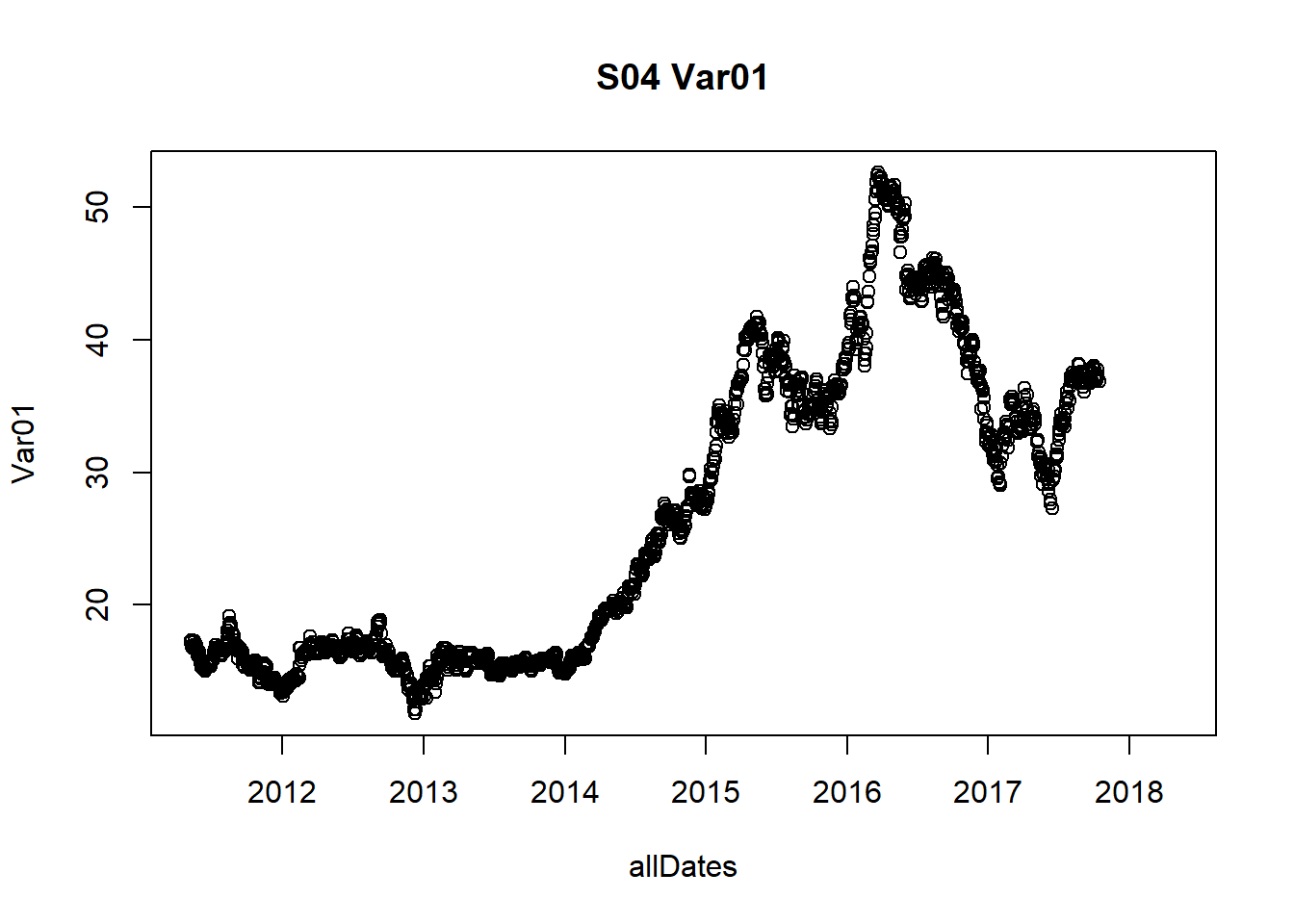
**S04 - Forecast Var01, Var02:**

We usually see consecutive sequences of five ‘SeriesInd’ entries followed by a two-integer (-days, based on our assumption) break, which might be indicative of weekdays. There are some additional breaks in the data beyond the five days on, two days off pattern.

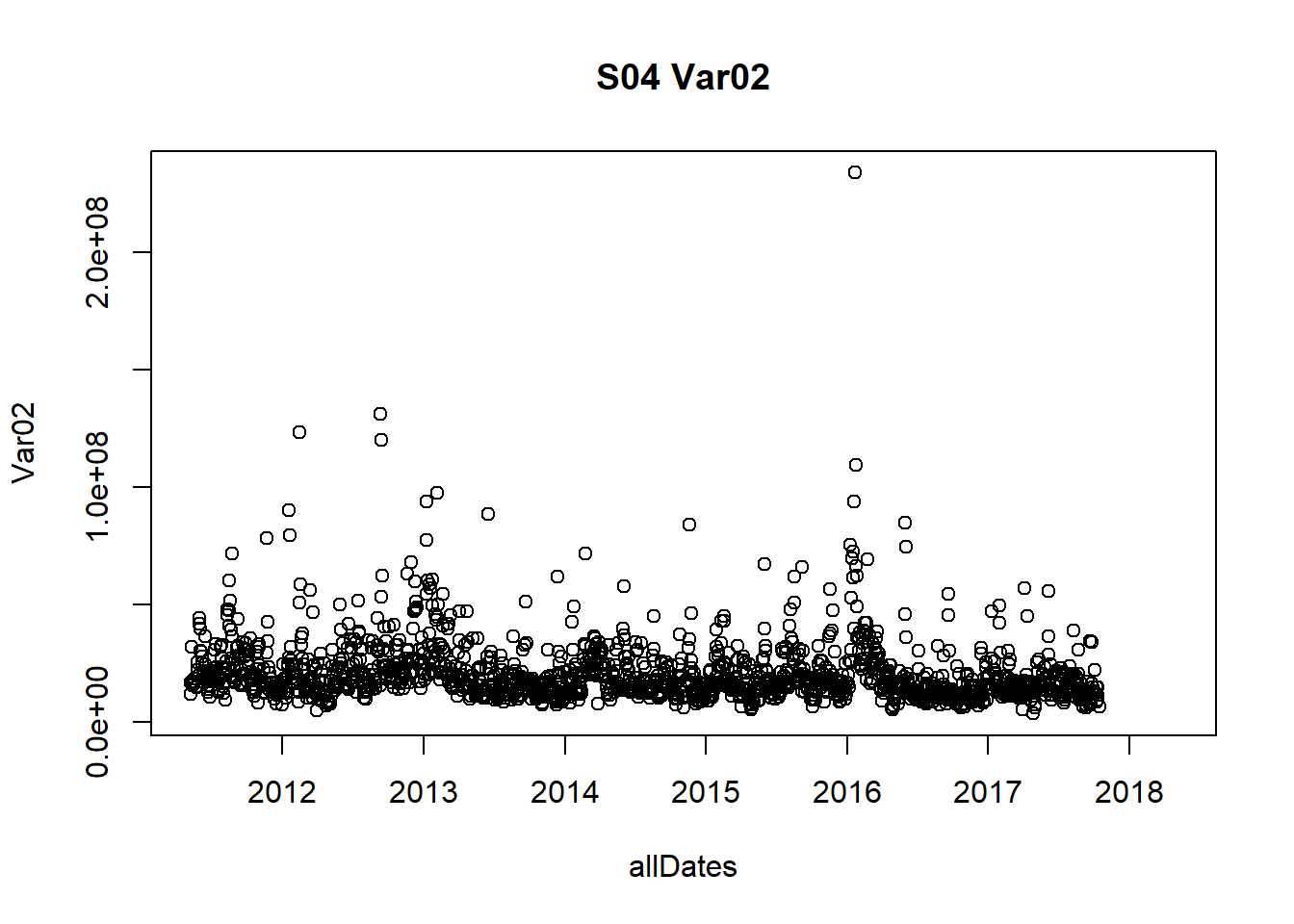
Note that if it were confirmed that the ‘SeriesInd’ values represented weekdays, it might be prudent to remove the weekend days from the sequence for forecasting purposes. From the exploratory Data Analysis below we see NA’s as expected. The maximum value of Var02 looks like it could be a substantial outlier.



**S04 Var01:** Below plot shows There might be a seasonal drop near the end of each year, but it’s not definitive. Business cycles could be present.

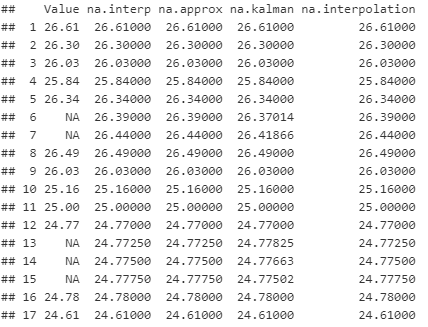
****

**S04 Var02:** There is no obvious trend or seasonality, but there are some outlier values - one clear outlier value is apparent in late 2015. It appears relatively stationary.

****

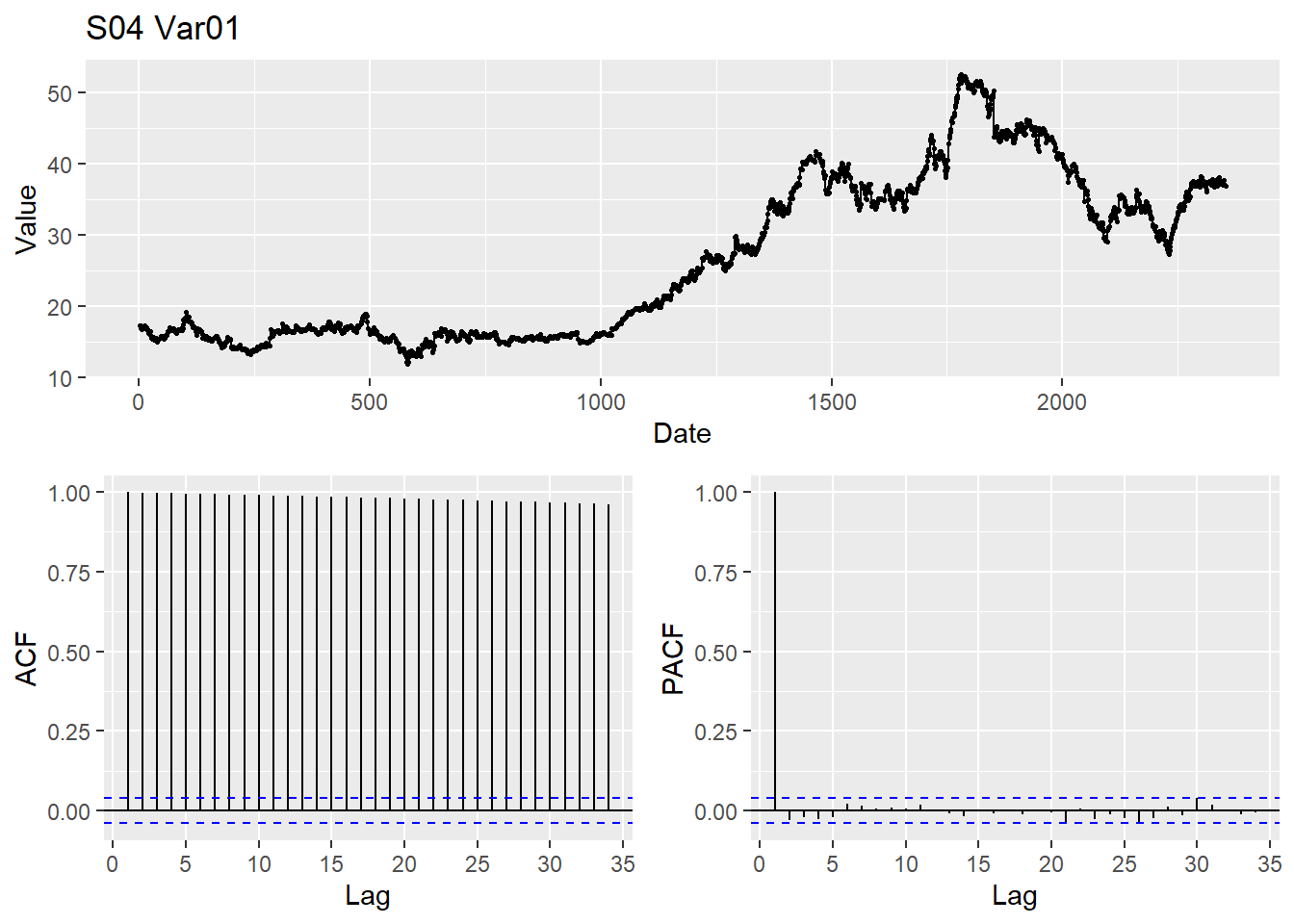
Because we haven’t confirmed the weekend hypothesis, we will go ahead and fill the missing SeriesInd values for both Var01 and Var02 with “N/A” entries for now. We won’t remove the presumed SeriesInd weekend values from the dataset.

Let’s review some different methods of handling NA values in the data. We will use the approx. function. Below is the View of some different imputation methods and their values.



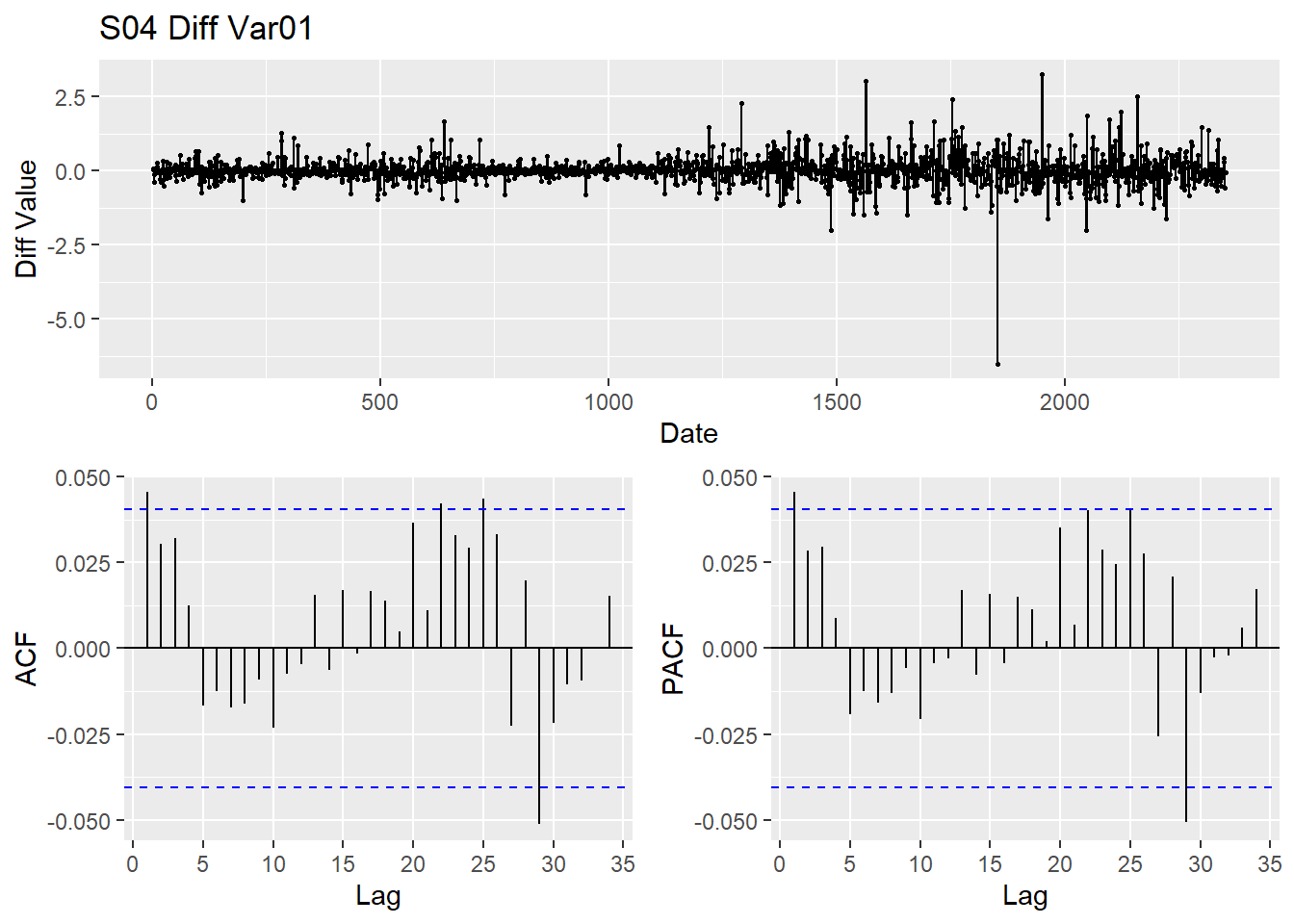
We appear to have two disparate datasets, so we will address them separately. Let’s find the right model to forecast S04 Var01.

##### Group S04 Var01



|  |  |
| --- | --- |
|  | KPSS test doesn’t support stationarity hypothesis. Let check do some differencing on the data. |

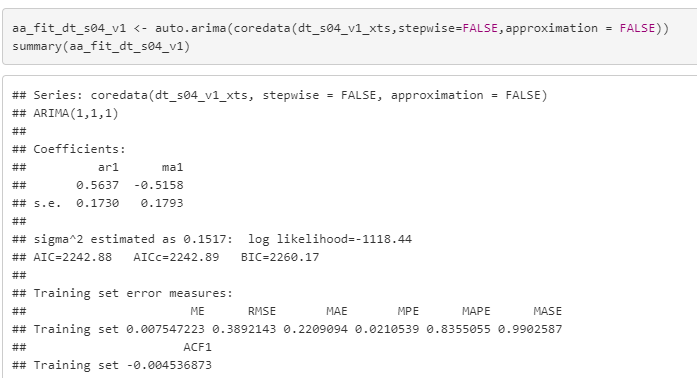
We see clear autocorrelation between variables in the ACF. Let’s take the first-order difference to see if our data can be made stationary for modeling.



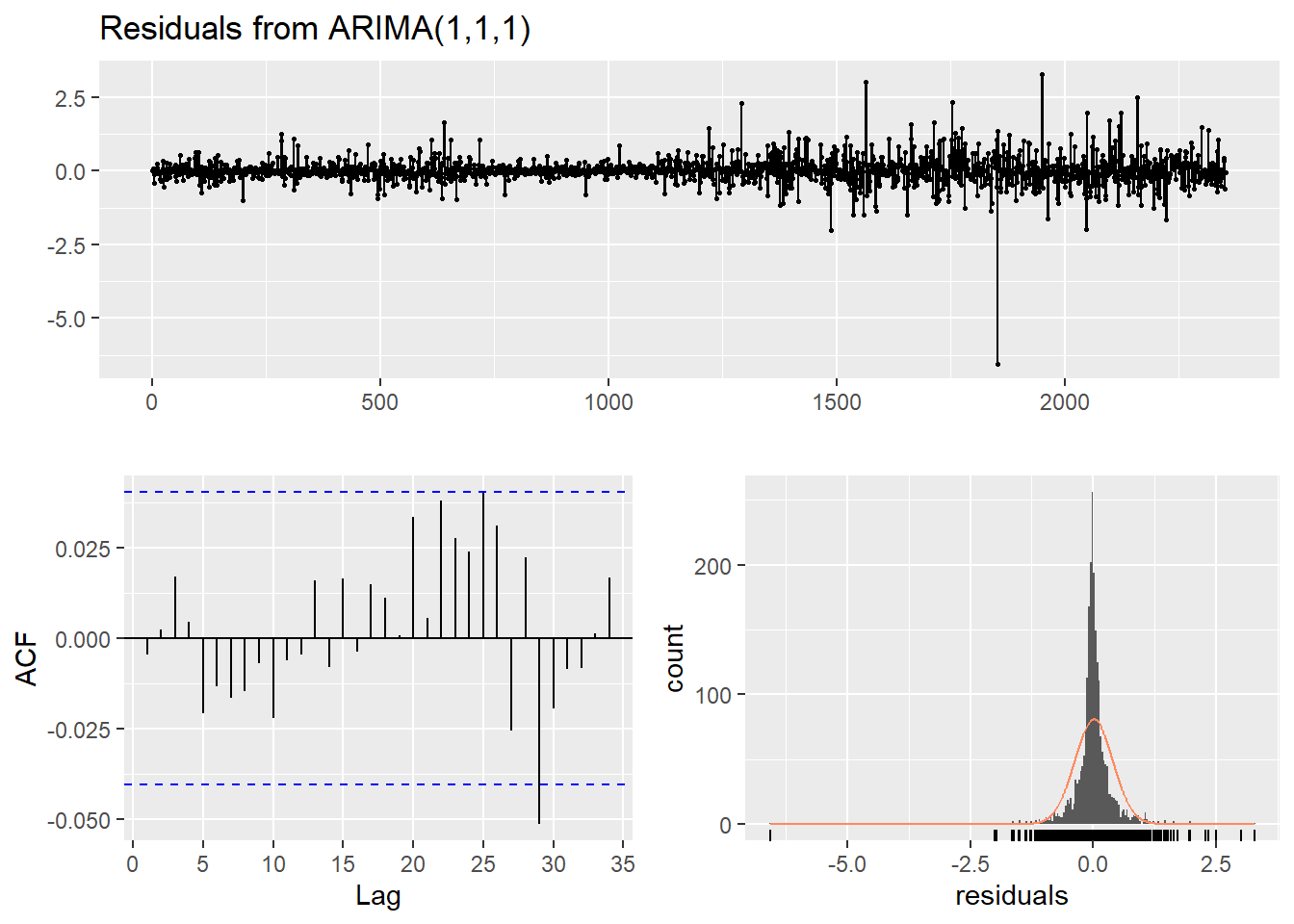
KPSS This test statistic looks better, so we conclude that the differences data are stationary. We verify with ndiffs and it also supports our finding.

|  |  |
| --- | --- |
|  |  |

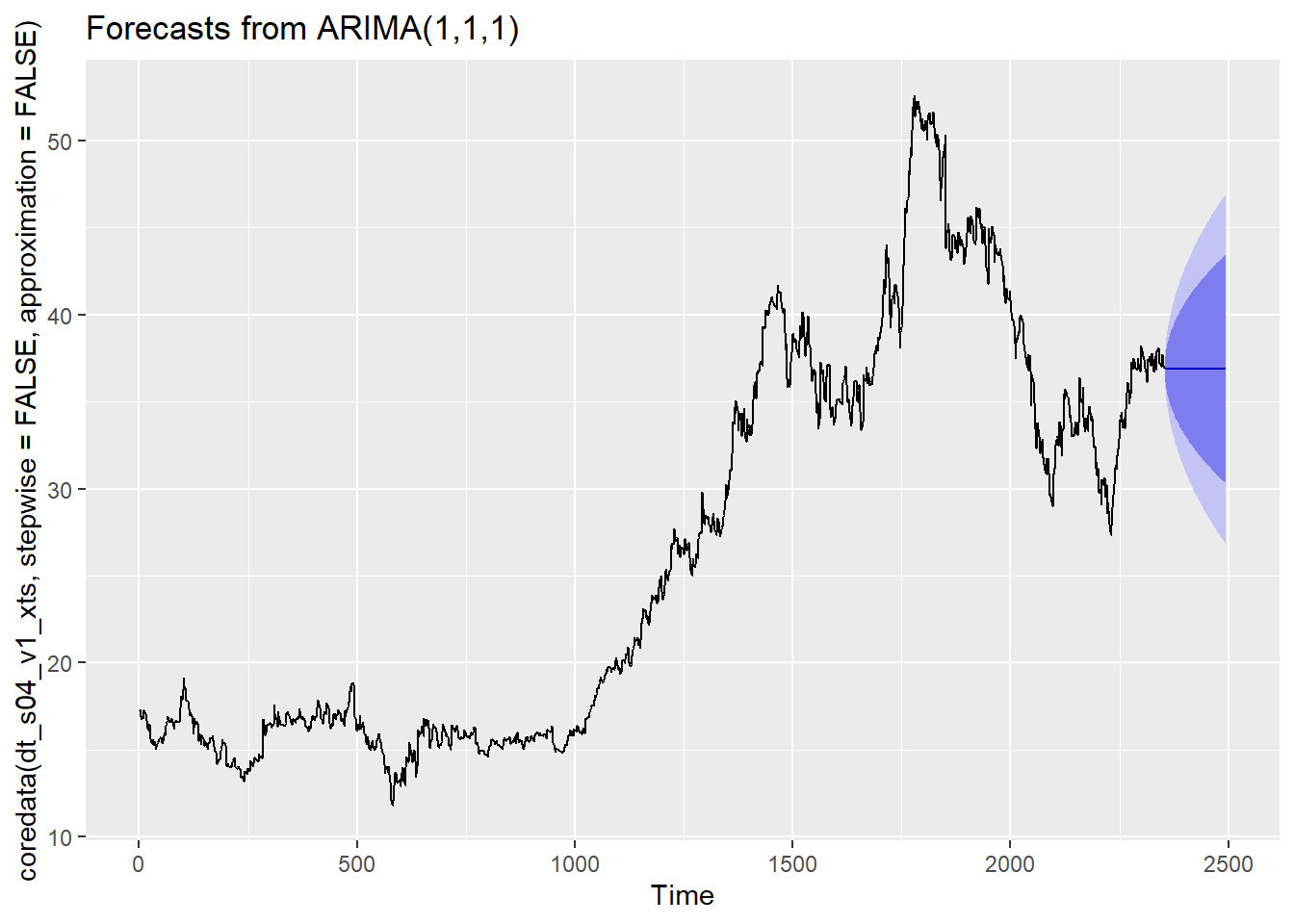
The S04 Var01 plots seem to indicate that ARIMA would be a good fit for this model.



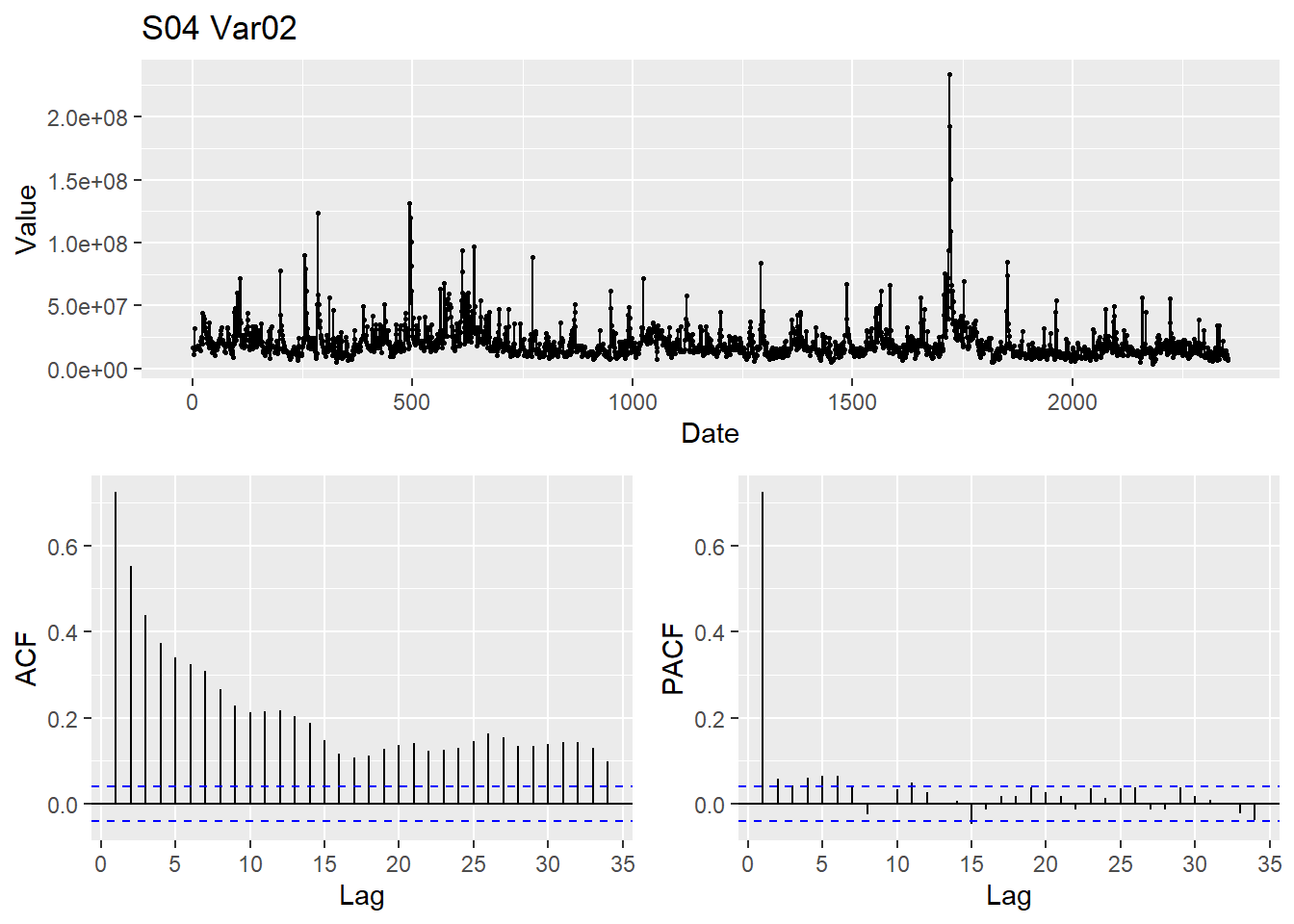
Let’s check the residual of this ARIMA(1,1,1) model.



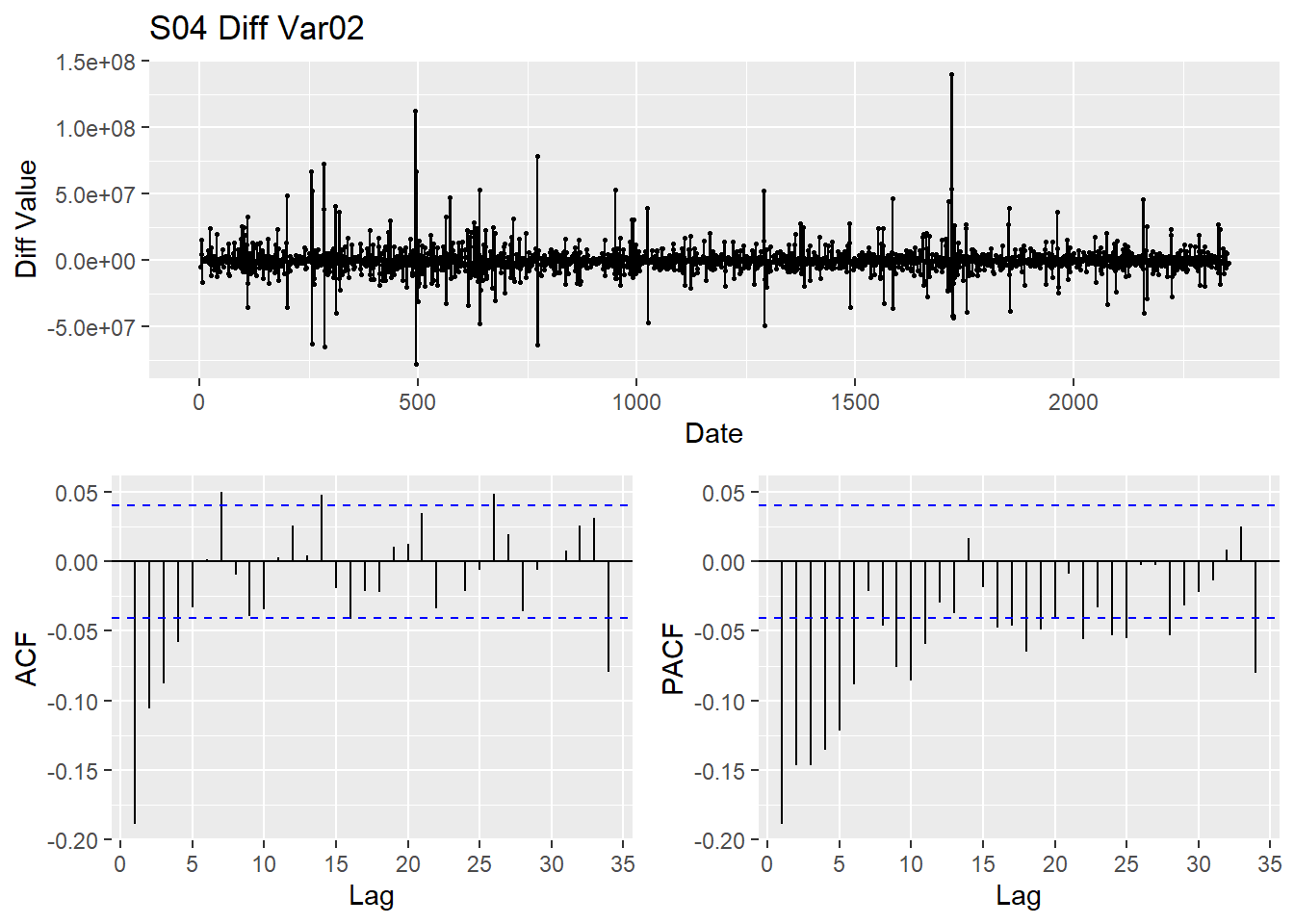
|  |  |
| --- | --- |
|  | Pvalue 0.79 is greater than .05 and hence we can’t reject the null hypnosis here , which say that residual data is white noise.  The residuals are independent, which is what we want. |

****

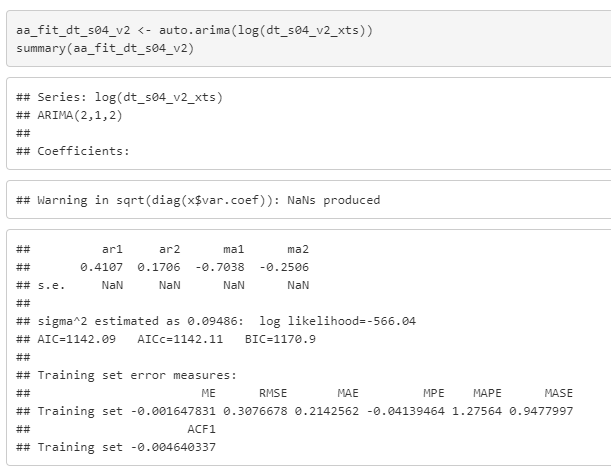
##### Group S04 Var02

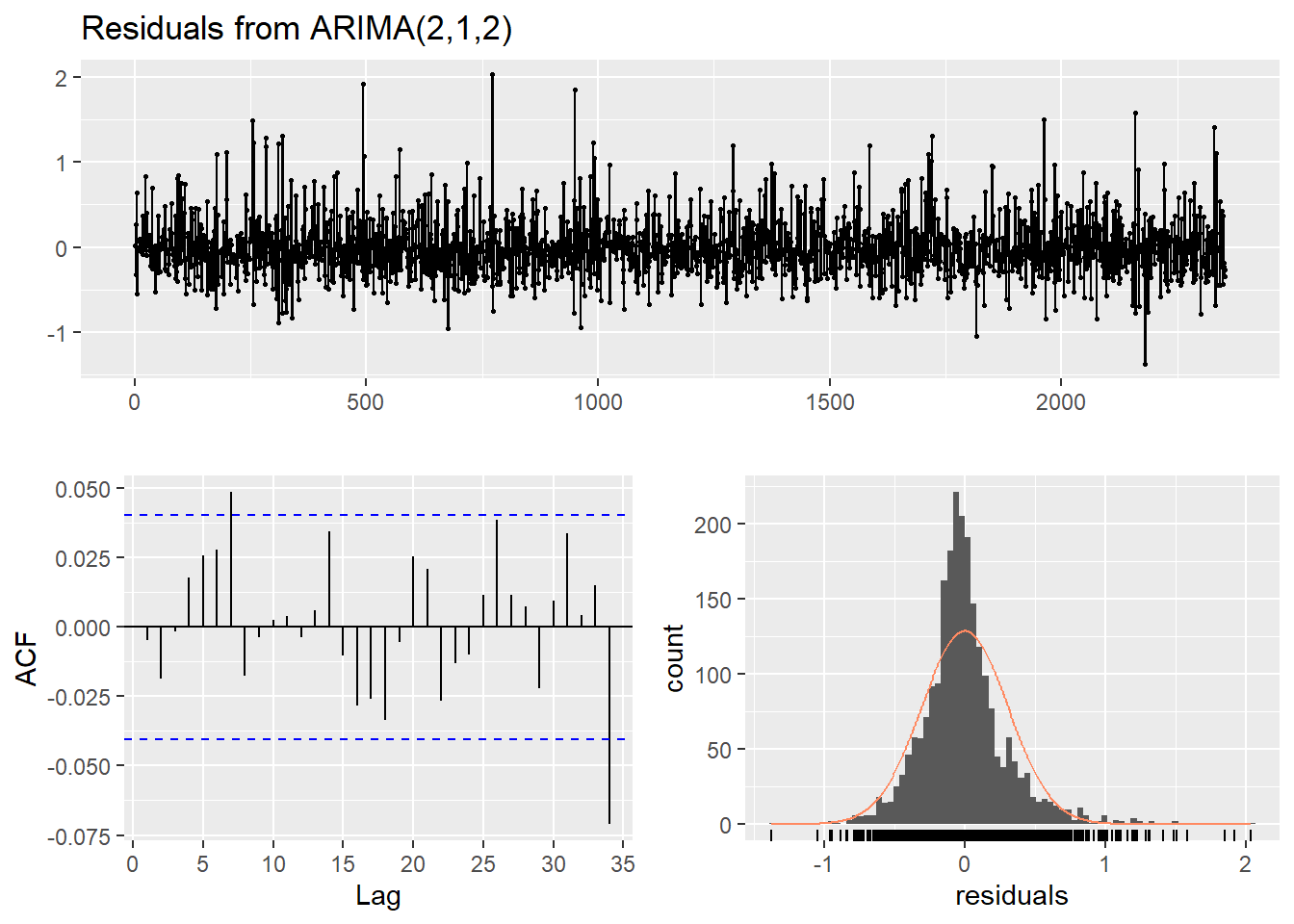
Now, let’s find the right model to forecast S04 Var02. 

|  |  |
| --- | --- |
| We see clear autocorrelation between variables in the ACF. Let’s take the first-order difference to see if our data can be made stationary for modeling. |  |



|  |  |
| --- | --- |
| This test statistic looks better, so we conclude that the differences data are stationary. We verify with ndiffs. |  |

The S04 Var02 plots seem to indicate that ARIMA(3,1,1) would be a good fit for this model. I am going to use log transformed data before we pass it to ARIMA Let run this with auto.arima. 



|  |  |
| --- | --- |
|  | The Ljung test shows pvalue greater than 0.05 so we can reject the null , hence. data is white noise. The residuals are independent, which is what we want. |

Let’s forecast the next 140 data point :



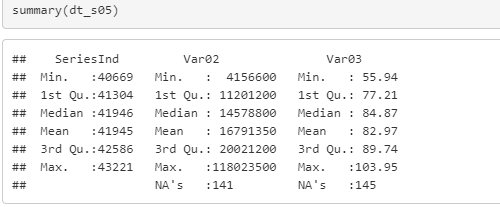
We will see some unit roots are present in this model for some data point , it is little disappointing to see that, as we may not able giving close predication in few case as unit data points is close to unit circle.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group /Var | Series /Model | RMSE | AIC | MAPE |
| S04 – VAR01 | ARIMA(1,1,1) | 0.3892143 | 2242.88 | 0.00835505464453924 |
| S04 – VAR02 | Series: log(dt\_s04\_v2\_xts)  ## ARIMA(2,1,2) | 0.3076678 | 1142.09 | 0.21408917391016 |

### Group S05

This group S05 is constituted of variables Var02 and Var03. Our goal is to find the best forecast for the variables Var02 and Var03 in S05. For that, we are going to process the dataset to change missing values and outliers. After some statistical analysis, we can apply several models and check for the accuracy of those models.

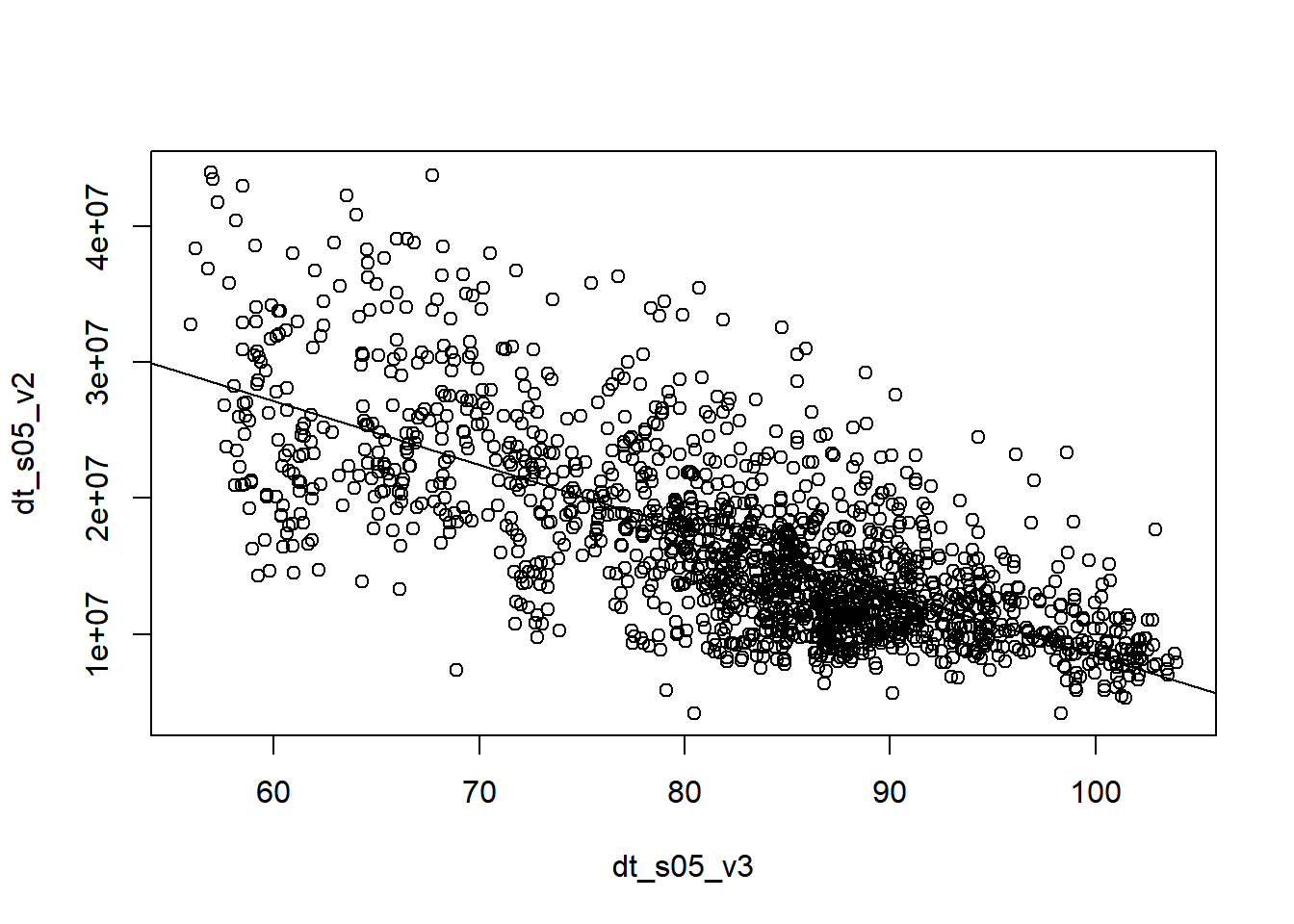
Statistical Analysis of Var02, and Var03 from group S05.



|  |  |
| --- | --- |
| Var02 has 1 missing value | Var03 has 5 missing values Median and mean are in the same order. |

**Imputing missing values:**

|  |  |
| --- | --- |
|  |  |



**Correlation between variables Var02 and Var03** : Above plots shows that there exists a linear correlation between V02 and V03 This is prove by the correlation test below, Var02 can explain Var03 by 50.75% and vice-versa.

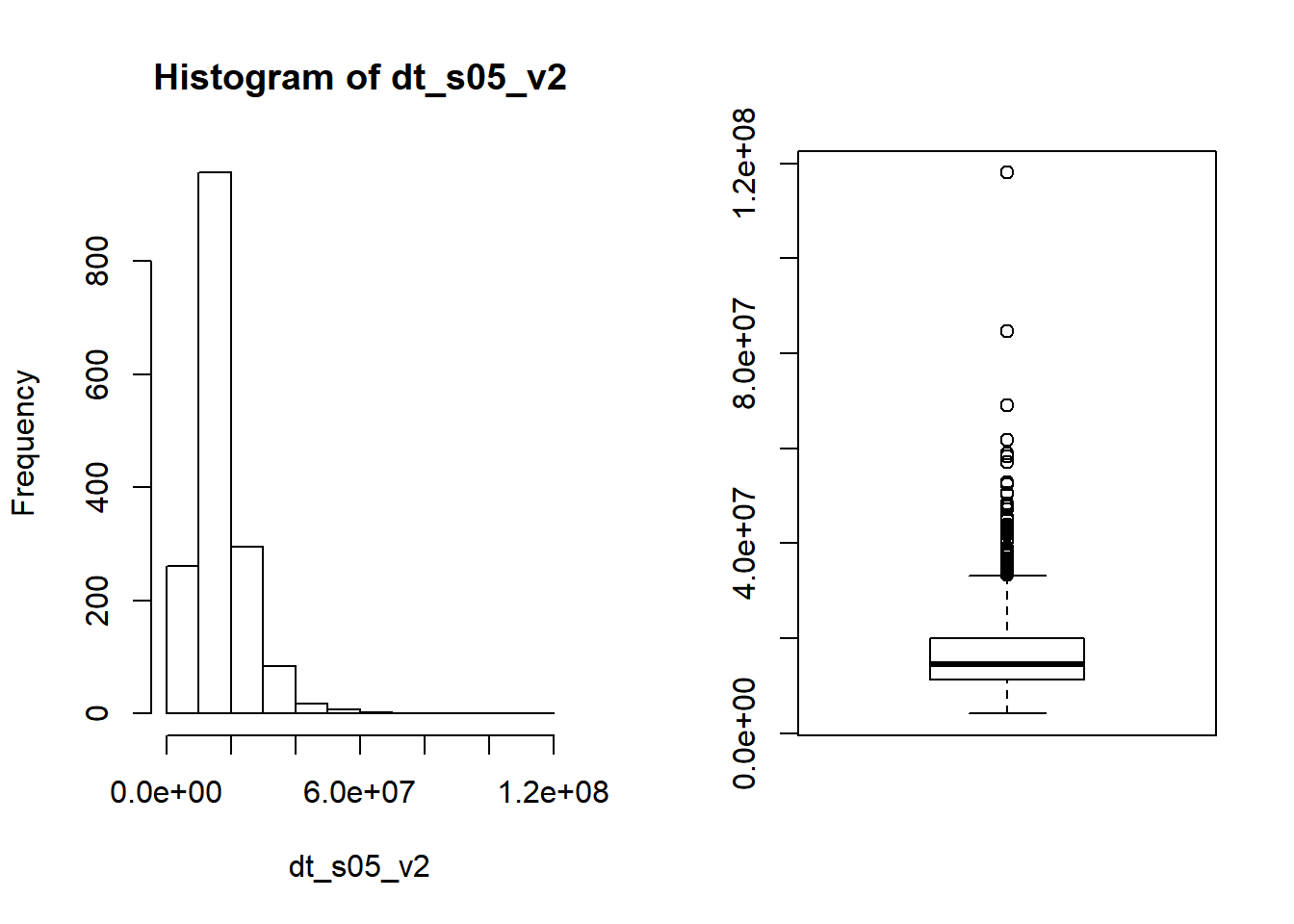
##### Group S05 Var02

Lets see the flow of variable Var02 over time :

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\F37BFA08.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\29B7E460.tmp |
| Time series S05\_V02 before cleaning | Time series S05\_V02 after cleaning |

Trend non seasonal time series

The distribution of data Var02 below shows that it is right skewed. We can suppress the ultimate outliers that skew the distribution then transform the data.



We have 31 extreme value in the time series V02 and only 1 in the time series V03. We use the tsclean function to replace outlier

Cleaning the time series.

Since the histogram and boxplot show that S05\_V02 has a skewed distribution, we can attempt to transform the dataset using Box-Cox transformation is necessary turn the time series to a normal distribution. More general than log transformation, the Box-Cox can also automatically chose the parameter lambda for efficient transformation.

**Forecast and Accuracy check**

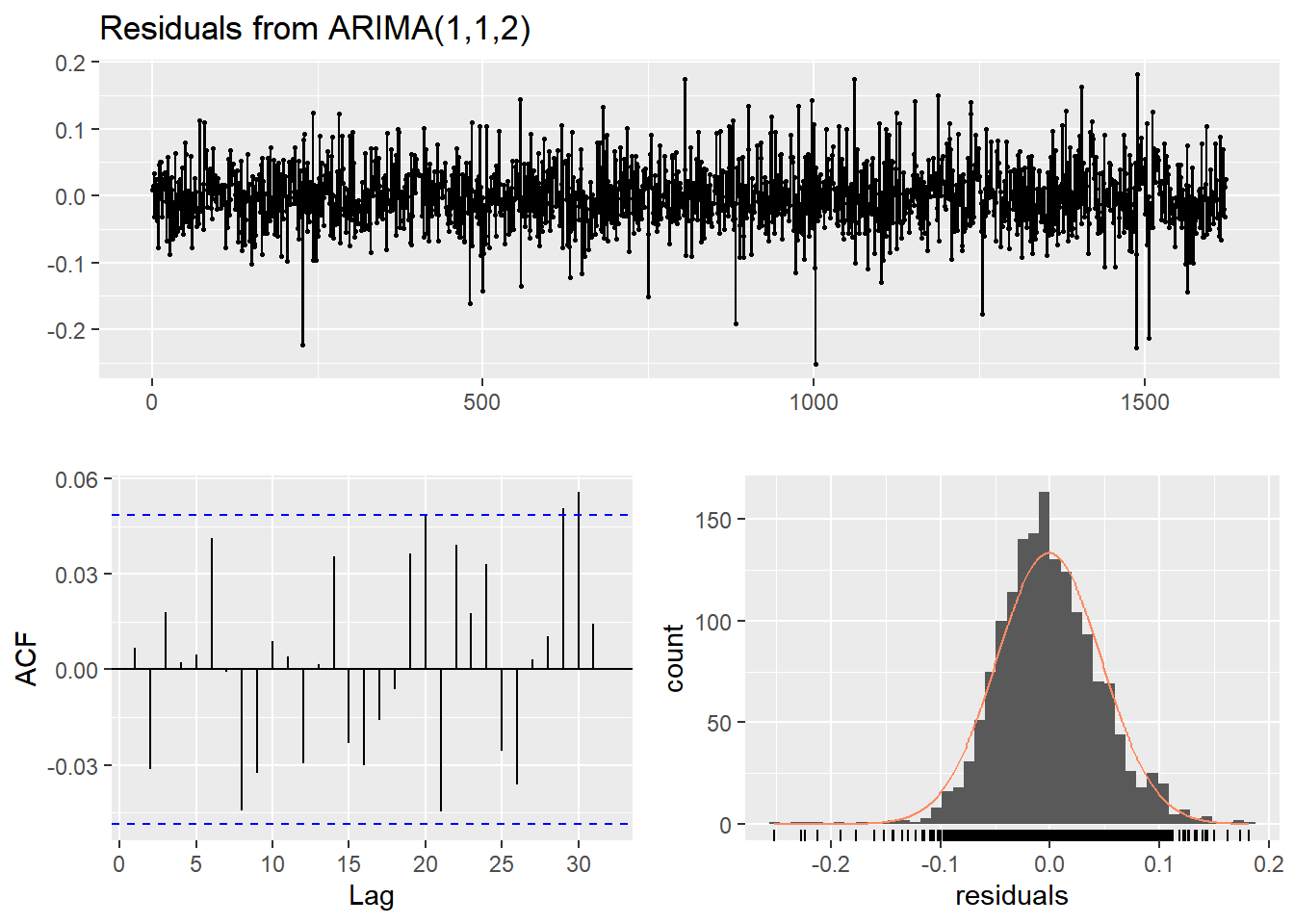
To forecast our time series, we selected two models. The Arima and the ETS which stand for Error Trends and Seasonality. The Arima model as well as the ETS can use the lambda parameter from the Box-Cox transformation to find the best model for each time series base on the metric called AIC or Akaike Information Criterion. The AIC is base on the information the model lost. So, the less a model loses information, the better quality that model is.

We use the train and test set to first build the model and then find the accuracy of the model thru different metrics such as RMSE(Root Mean Square Error), MSE(Mean Square Error), MAPE(Mean Absolute Percent Error)… The best model will generate the least value of each metric. We check for the best model that it represent the data by checking that the residuals are white noise. The function ‘checkresiduals’ give the parameters a the p-value to estimate the veracity of the residuals as white noise.

In this forecasting, the 80% confidence interval (light blue) on the test forecast cover the entire fluctuation of the data. The 95% confidence interval (dark blue) almost fully cover the amplitude of oscillation of the test data.

Check the residuals if the model is valid:

|  |  |
| --- | --- |
| C:\Users\alain\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\1FD2FAB6.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\80FC069A.tmp |
| Forecast with ETS model on V05\_S02 | Forecast with Arima model on V05\_S02 |



|  |  |
| --- | --- |
|  | with p-value greater than 0.05, there is convincing evidence that residuals for Var02 are white noise. On ACF, the residuals are uncorrelated. The histogram shows that the residuals are normal distributed. |

ACF of Var02 difference:

The time series is the sum of Trends, seasonality and residuals. The differencing remove the trend and seasonality so that we can explore the residuals. We can then check visually if the time series is stationary or not.

We trying to confirm that S05\_V02 is not stationary by plotting ACF of Var02 difference

The ACF of S05\_V02 shows that the time series V02 has trends. This means that V02 is not stationary.

|  |  |
| --- | --- |
| C:\Users\alain\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B8DBEE1F.tmp | C:\Users\alain\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\DAA0711D.tmp |
| Time series Var02 | ACF of time series ACF |

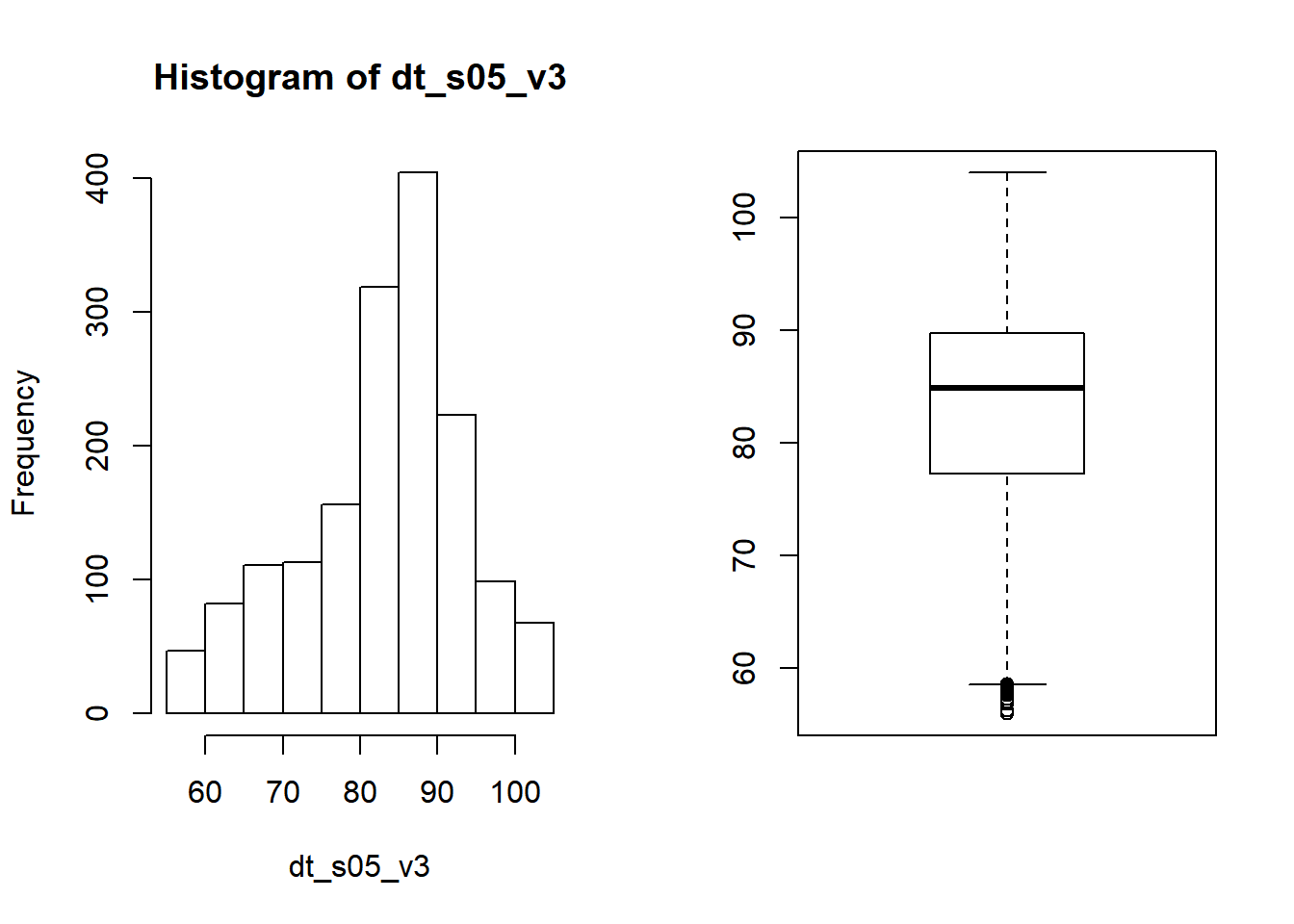
.

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\AD98FA10.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B7743A9E.tmp |
| The difference has removed the trend and the residuals look like white noise;. | This ACF of residuals on V02 has one order autocorrelation. Even thought there is no seasonality, there still information that we can get from this residuals. |

##### Group S05 Var03

Lets see the flow of variable Var03 on time : Trend non seasonal time series. It can also be cyclic time series

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\B48C5556.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\3591A96E.tmp |
| Time series Var03 before cleaning | Time series Var03 after cleaning |

The distribution of data Var03 is nearly normal distributed and has outliers at the left.

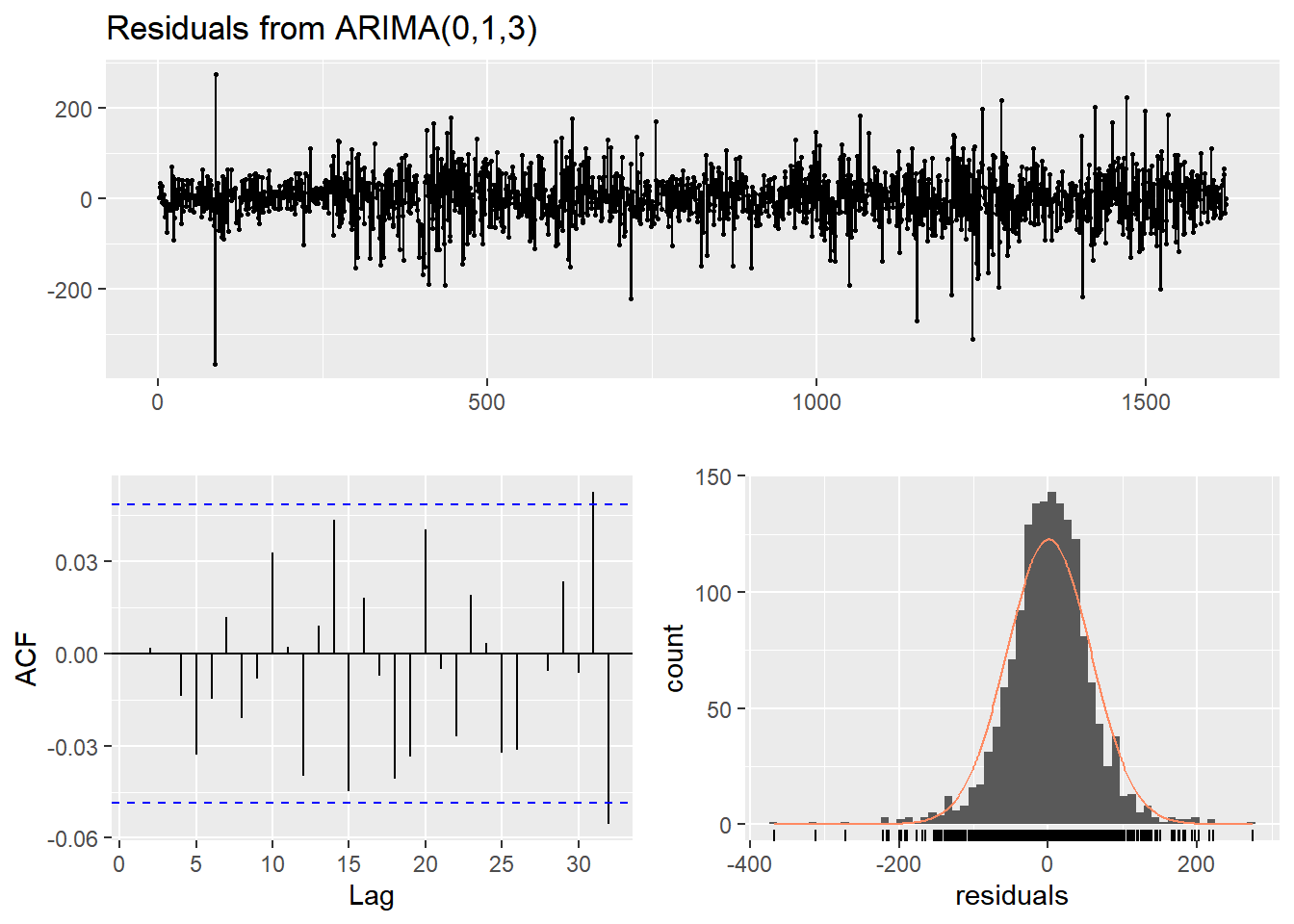
We have 31 extreme value in the time series V02 and only 1 in the time series V03. We used tsclean to replace outlier Cleaning the time series

ACF of Var03 difference:

|  |  |
| --- | --- |
| C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\6A6F6684.tmp | C:\Users\951250\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\955B8332.tmp |
| This residuals present extreme outliers. | Even though the ACF of these residuals presents one lag, it still white noise. Most of the information of the dataset are on the Trends. |

Var03 residuals for Model ARIMA (0,1,3):

|  |  |
| --- | --- |
| C:\Users\alain\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\3392692.tmp | C:\Users\alain\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\31BBE743.tmp |
| Forecast with Arima model on S05\_V03 | Forecast with ETS model on S05\_V03 |



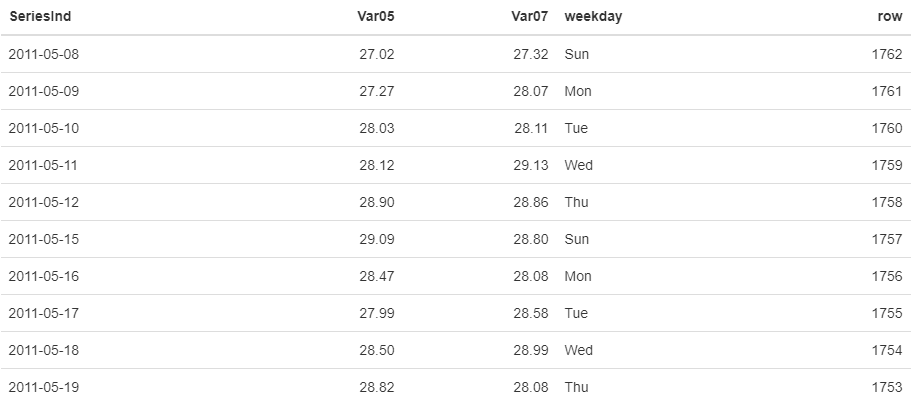
|  |  |
| --- | --- |
| with p-value greater than 0.05, there is convincing evidence that residuals for Var03 are white noise. On ACF, the residuals are uncorrelated. The histogram shows that the residuals are normal distributed. |  |

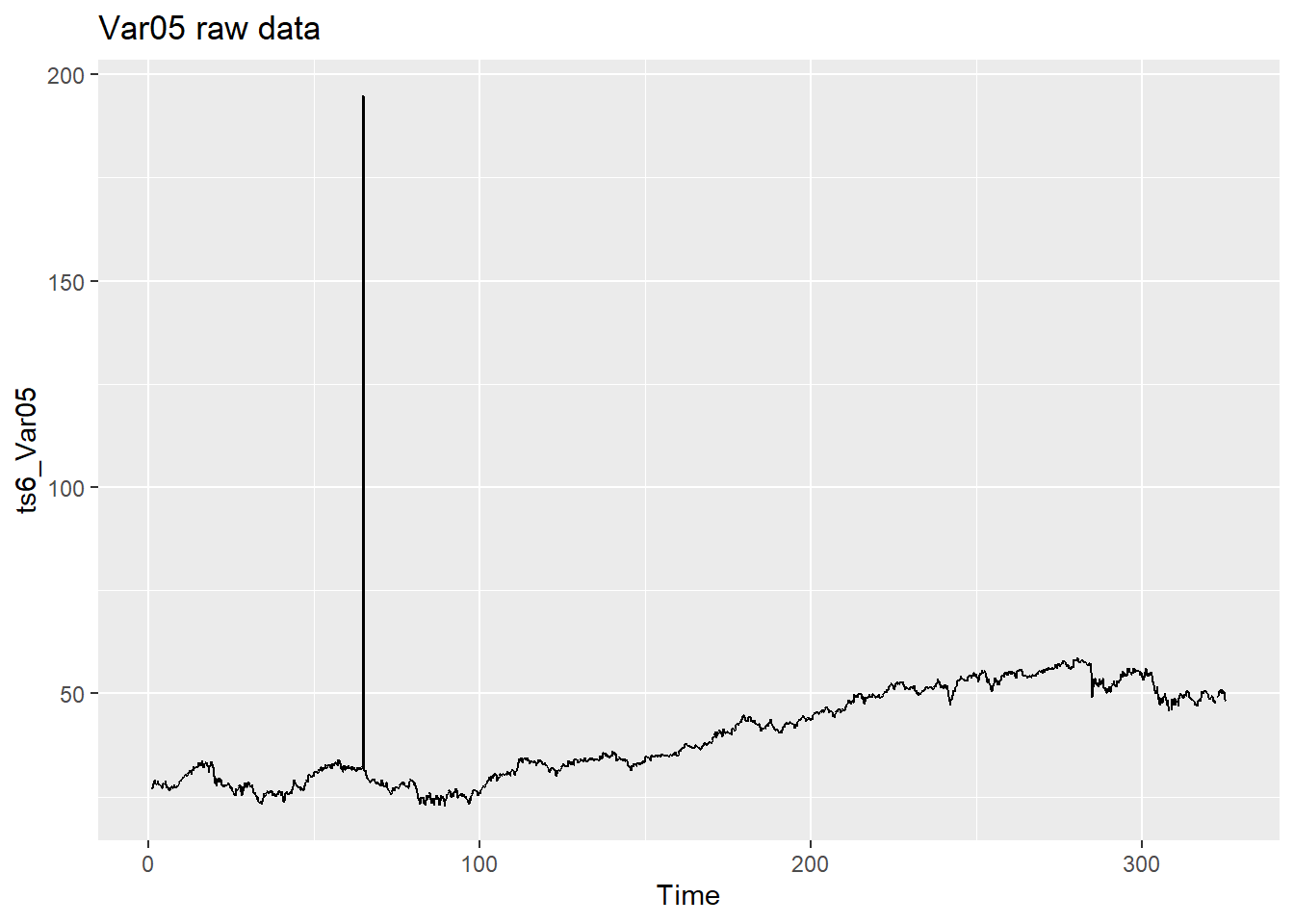
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group /Var | Series /Model | RMSE | AIC | MAPE |
| S05 – VAR02 | ARIMA(1,1,2)  Box Cox transformation: lambda= -0.09436074 | 302903.7 | -5242.21 | 0.1721 |
| S05 – VAR02 | ETS(A,N,N) Box-Cox transformation: lambda= -0.0944 | 263534.8 | 2219.804 | 0.1755 |
| S05 – VAR03 | ARIMA(0,1,3)  ## Box Cox transformation: lambda= 1.944948 | 0.8849026 | 17683.02 | 0.0076 |
| S05 – VAR03 | ETS(A,N,N)  Box-Cox transformation: lambda= 1.9449 | 0.8929011 | 25109.30 | 0.0077 |

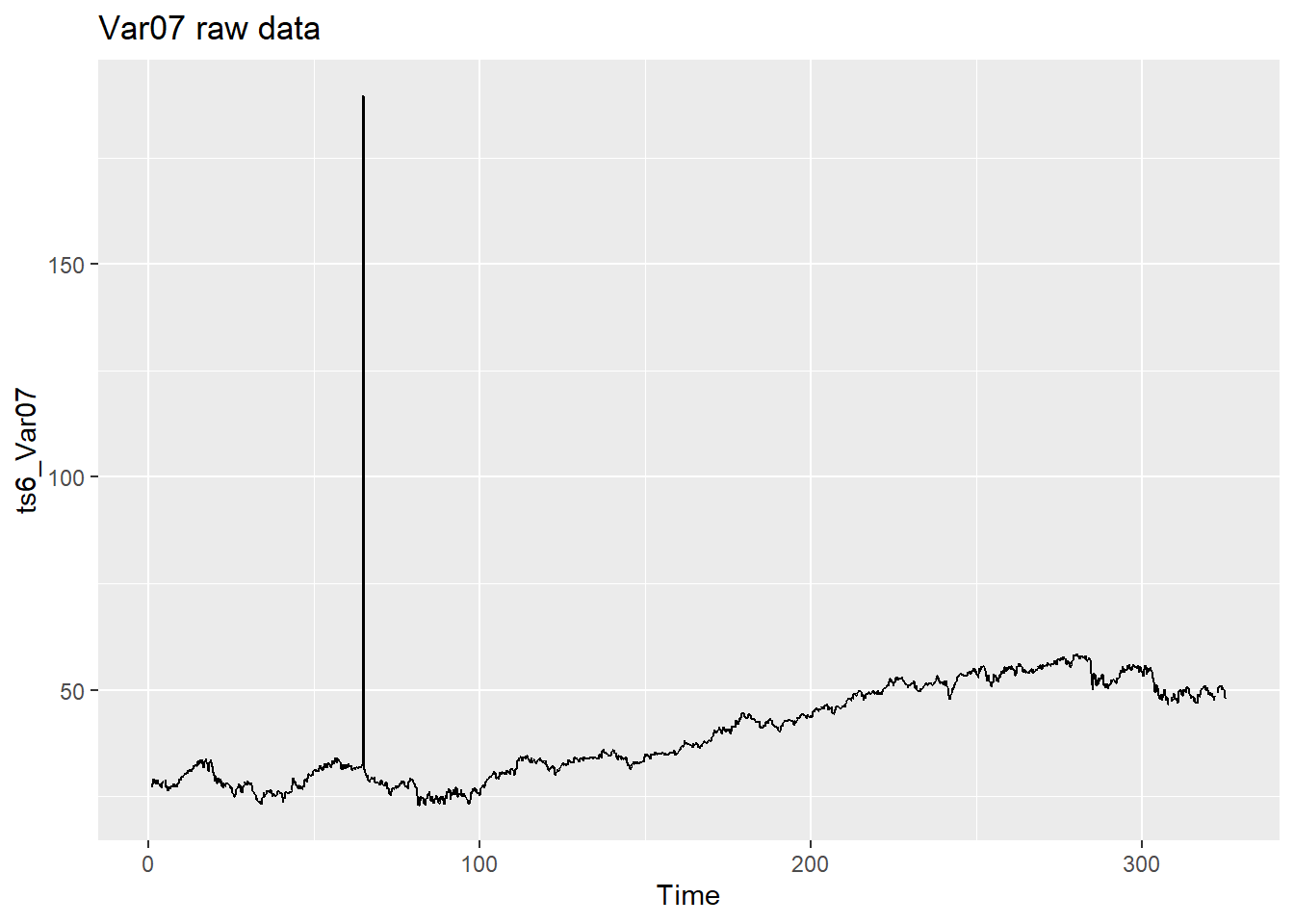
### Group S06

For group S06, we need to build forecasts for Var05 and Var07 respectively. We look at the data and find that it has a frequency of 5 (cycling every 5 days). We count the number of missing dates, and we also plot the raw data to visualize any extreme value. We realize that there’s one extreme and 5 missing values in each variable.

# extract group S06, Var05, Var07

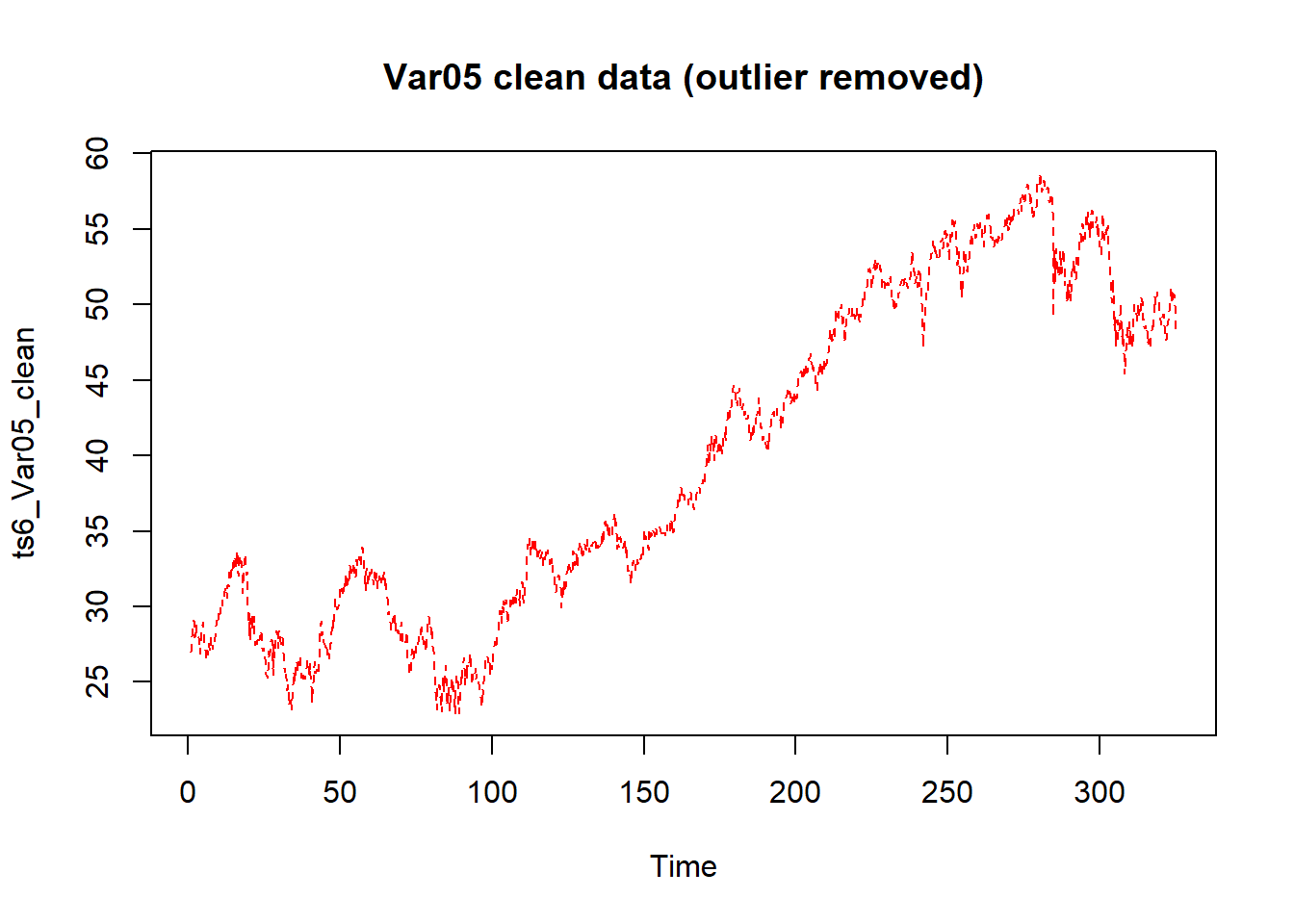


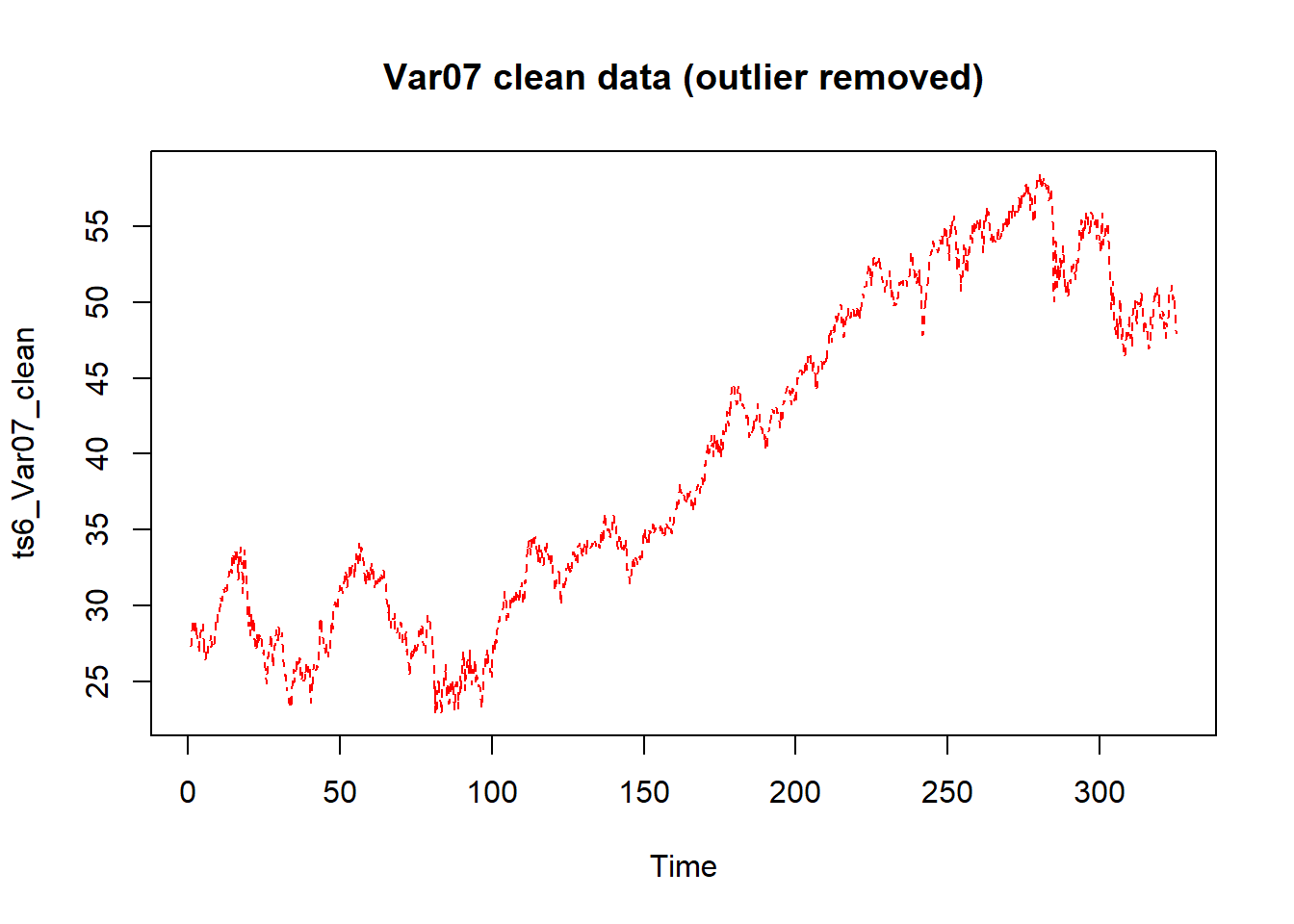




**Remove outlier(s), impute missing value:**

Next step, we need to remove outliers and impute missing values before building our forecast model. We can locate the extreme outlier(s) and check for replacement(s) using the **forecast::tsoutliers** function. Furthermore, we can apply the **forecast::tsclean** to remove the outlier(s) and replace any missing value. Subsequently, we examine the time series and plot the clean data. Now, it’s ready for building forecast.

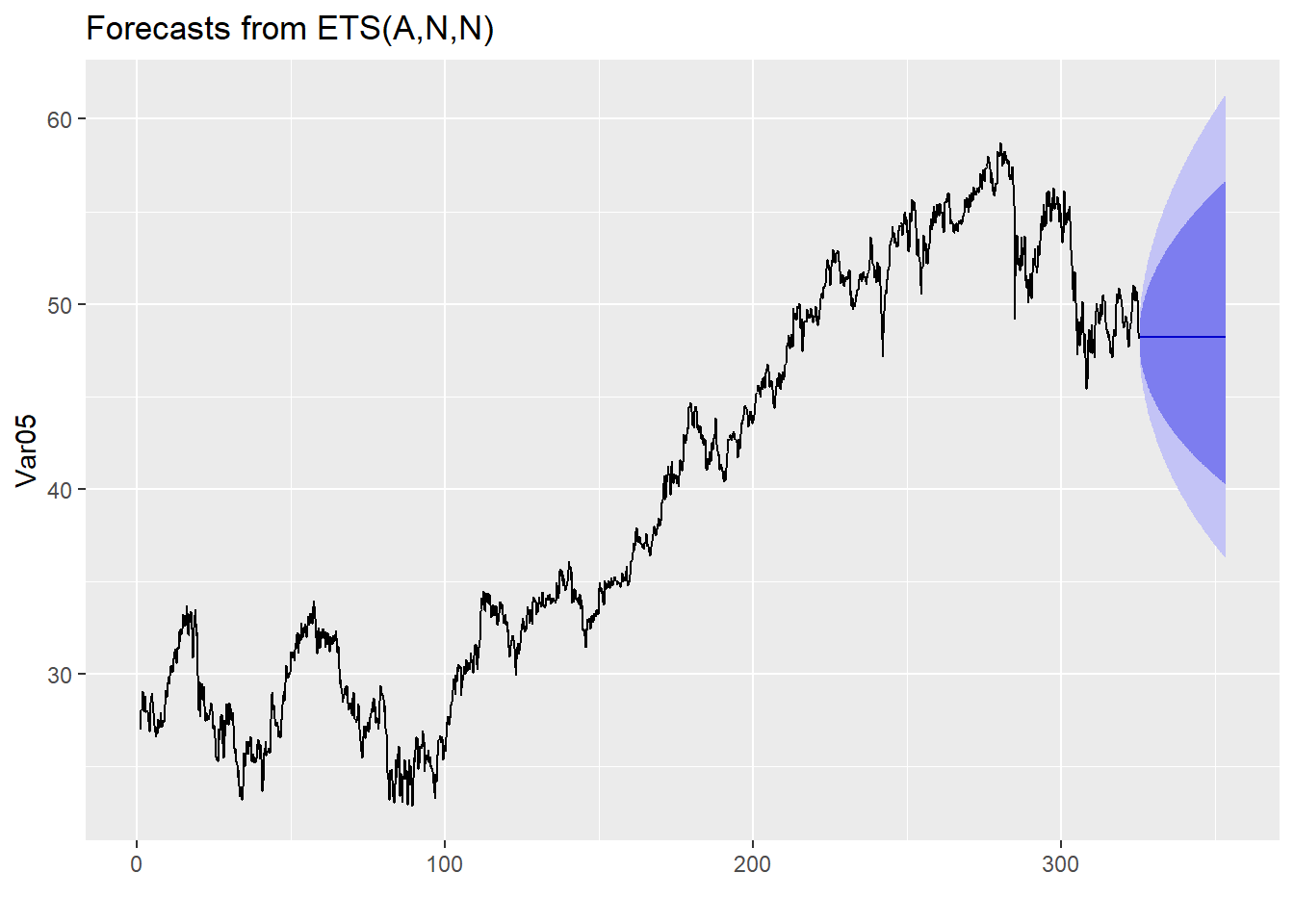
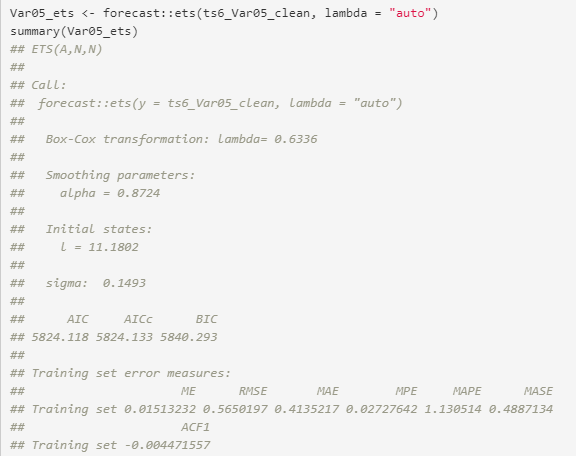
****

****

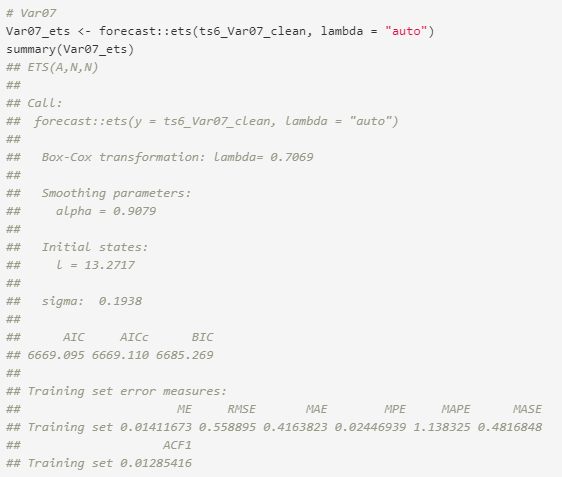
**Model – ETS:**

**Var05 :**

We decide to automate the process by relying on ETS model. We use the **forecast::ets** to look for the most appropriate forecasting model for each variable. In both cases, the ets recommends [A,N,N], i.e. simple exponential smoothing with additive errors. We fit the data and visualize the forecast with prediction interval (80, 95). In addition, we evaluate our models using **forecast::accuracy**. The MAPE is 1.13% and 1.14% for Var05 and Var07 respectively.

****

**Var07**

****

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group /Var | Series /Model | RMSE | AIC | MAPE |
| S06 – VAR05 | Series: Var05\_ets  ETS(A,N,N) | 0.565 | 5824.118 | 0.0113051402177851 |
| S06 – VAR05 | Series: log(tsclean(dt\_s06\_v5\_xts))  ARIMA(3,1,3) | 0.01202235 | -14106.61 | 0.008784268 |
| S06 – VAR07 | Series: Var07\_ets  ETS(A,N,N) | 0.559 | 6669.095 | 0.0113832455506511 |
| S06 – VAR07 | Series: log(tsclean(dt\_s06\_v7\_xts))  ARIMA(2,1,3) | 0.01227619 | -14010.39 | 0.008840895 |

We have used logged model with Arima after removing the outlier to generate the prediction for submitting.

### 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.)