**A Simple Arima Model**

**What is Arima model?**

ARIMA is the abbreviation for Auto Regressive Integrated Moving Average. Auto Regressive (AR) terms

refer to the lags of the differenced series, Moving Average (MA) terms refer to the lags of errors and I is the number of difference used to make the time series stationary.

**What are the assumptions of time series modelling?**

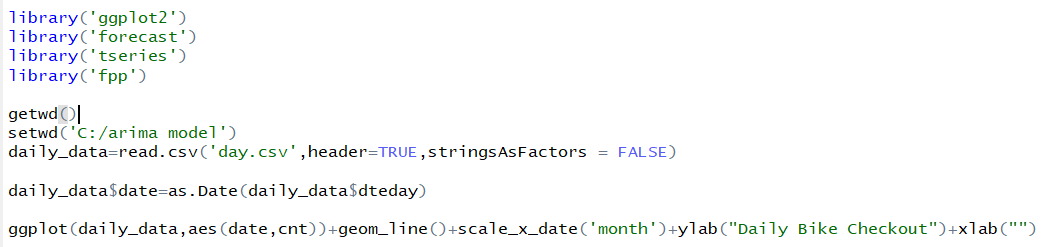
Assumptions of ARIMA model

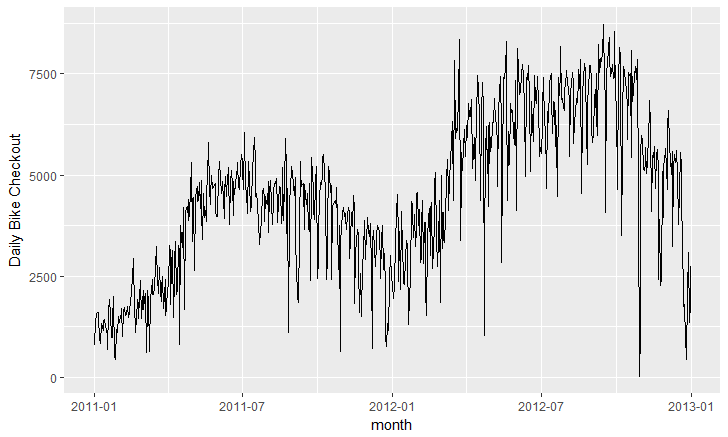
1. Data should be stationary – by stationary it means that the properties of the series don’t depend on the time when it is captured. A white noise series and series with cyclic behavior can also be considered as stationary series.

2. Data should be univariate – ARIMA works on a single variable. Auto-regression is all about regression with the past values.

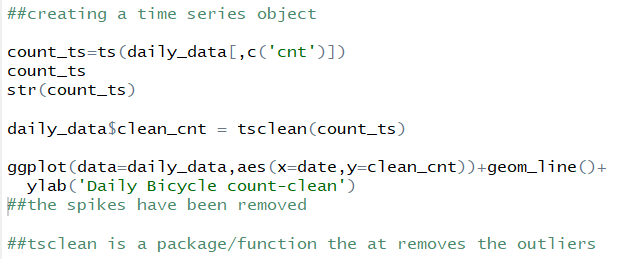
Code:

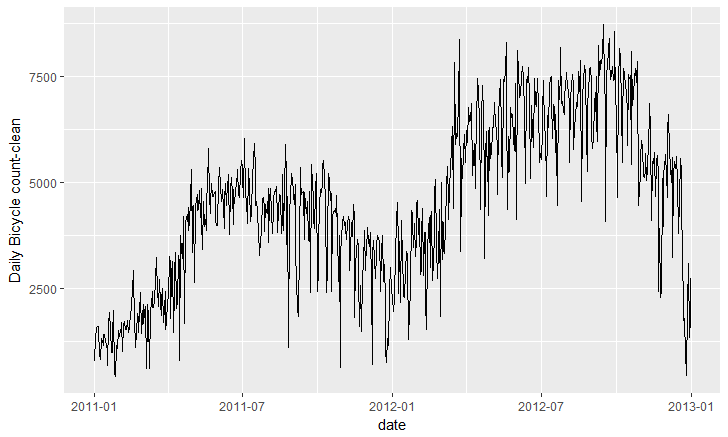
##importing the necessary libraries and plotting data:



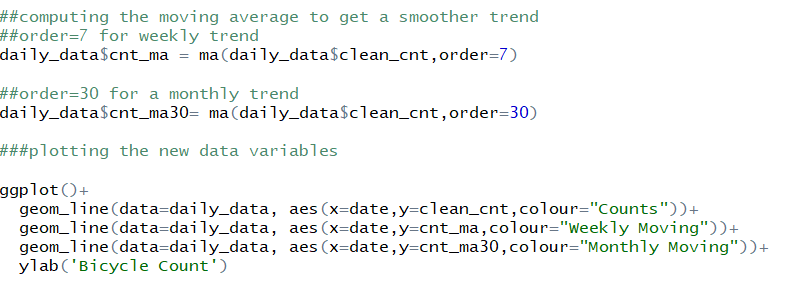


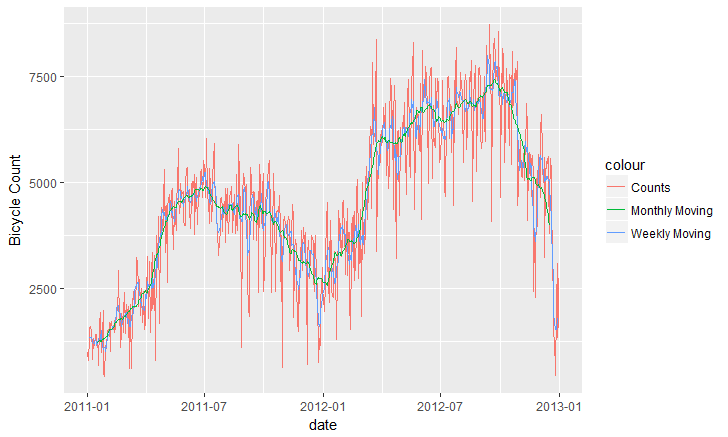
In some cases, the number of bicycles checked out dropped below 100 on day and rose to over 4,000 the next day. These are suspected outliers that could bias the model by skewing statistical summaries. R provides a convenient method for removing time series outliers: tsclean() as part of its forecast package. tsclean() identifies and replaces outliers using series smoothing and decomposition.



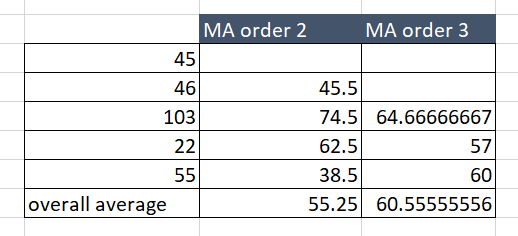


We observe that the trend is still not very smooth. In order to create a smoother trend, we compute the moving average of the time series object. Please remember this moving average is the numerical moving average and not the MA component that we will use to analyze the ARIMA model.





Simple example to compute moving average



We then try to decompose our time series data

**What is decomposition of time series object?**

Time series data essentially consist of four components

1)Seasonal Component

2)Cyclic Component

3)Trend

4) Random Residual

Seasonal component refers to fluctuations in the data related to calendar cycles.

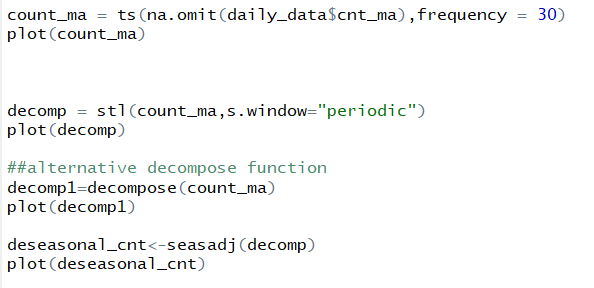
For example, more people might be riding bikes in the summer and during warm weather, and less during colder months. Usually, seasonality is fixed at some number; for instance, quarter or month of the year.

Trend component is the overall pattern of the series: Is the number of bikes rented increasing or decreasing over time?

Cycle component consists of decreasing or increasing patterns that are not seasonal. Usually, trend and cycle components are grouped together. Trend-cycle component is estimated using moving averages.

Residual component-Finally, part of the series that can't be attributed to seasonal, cycle, or trend components is referred to as residual or error.

The process of extracting these components is referred to as decomposition.

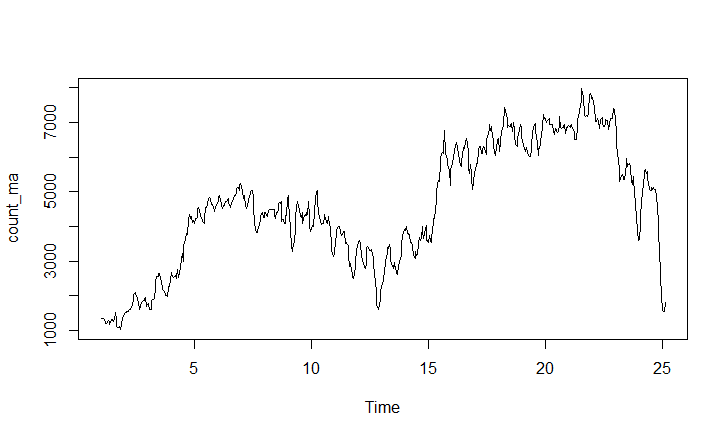


**What is the ts function here?**

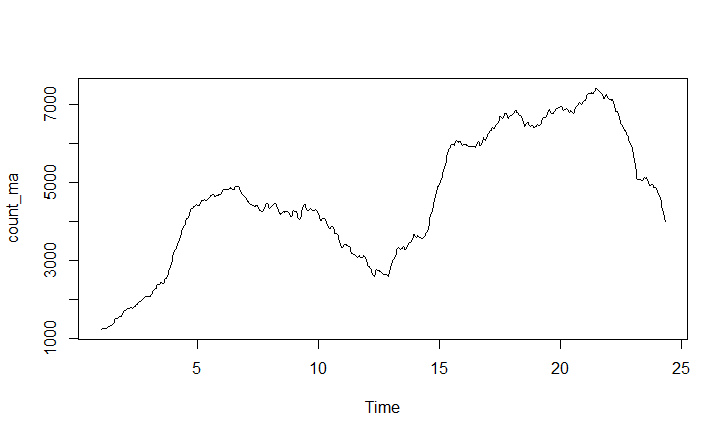
The **ts()** function will convert a numeric vector into an R time series object. The format is **ts(**vector,**start=, end=, frequency=)** where start and end are the times of the first and last observation and frequency is the number of observations per unit time (1=annual, 4=quartly, 12=monthly, etc.).

First, we calculate seasonal component of the data using stl(). STL is a flexible function for decomposing and forecasting the series. It calculates the seasonal component of the series using smoothing and adjusts the original series by subtracting seasonality in two simple lines.

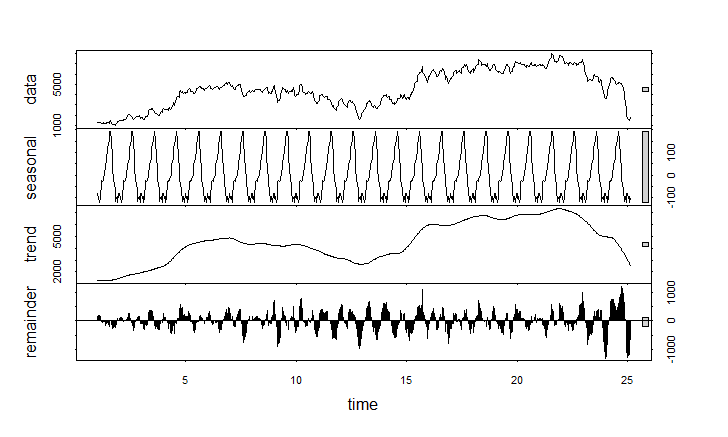
Weekly trend plot



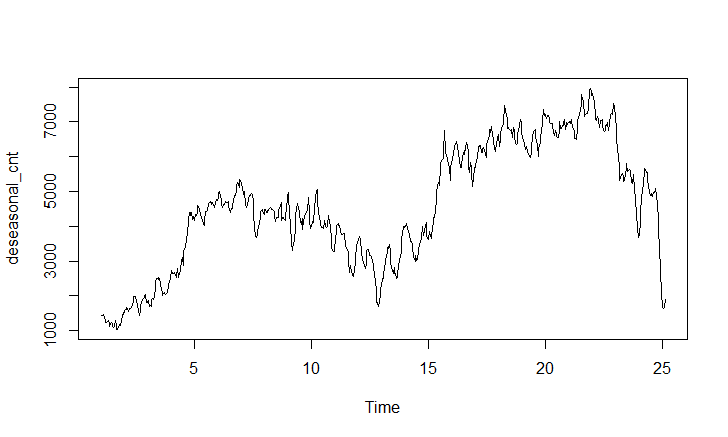
Monthly trend plot



Decomposed components of the time series dataset



Time series data adjusted for the seasonal component



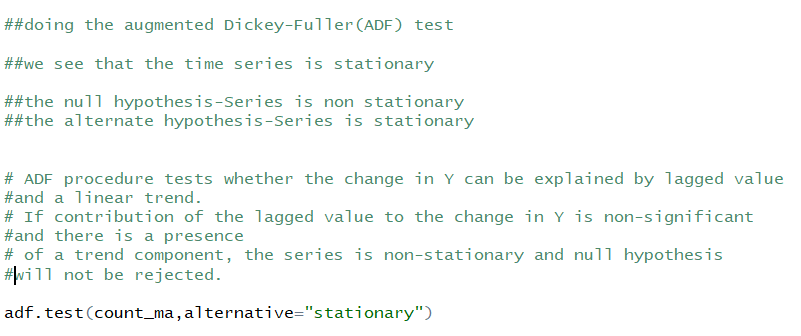
We don’t observe any seasonality or trend in the dataNote that stl() by default assumes additive model structure.Use allow.multiplicative.trend=TRUE to incorporate the multiplicative model.

**Is the time series dataset stationary?**

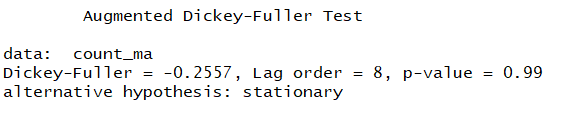
Fitting an ARIMA model requires the series to be stationary.

A series is said to be stationary when its mean, variance, and autocovariance are time invariant. This assumption makes intuitive sense: Since ARIMA uses previous lags of series to model its behavior, modeling stable series with consistent properties involves less uncertainty.

An example of a stationary series, where data values oscillate with a steady variance around the mean of 1. For a non-stationary series; mean of this series will differ across different time windows.

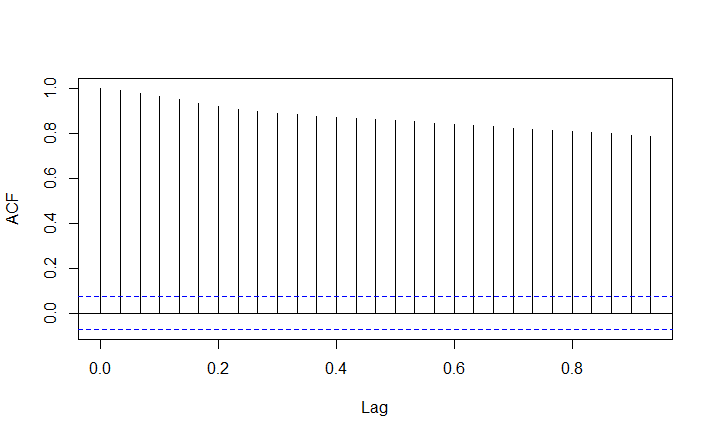


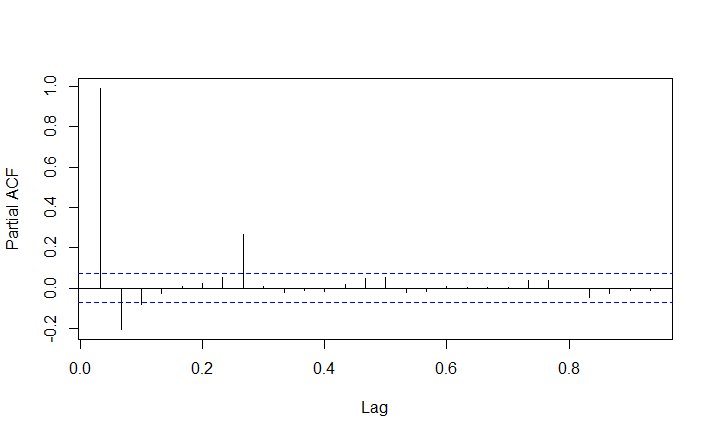
Test results



The test shows that the series in non-stationary.

To treat this series, we take a differencing factor. Checking the Auto Correlation and Partial Auto Correlation Plots.





**What are auto correlations in Time Series Data?**

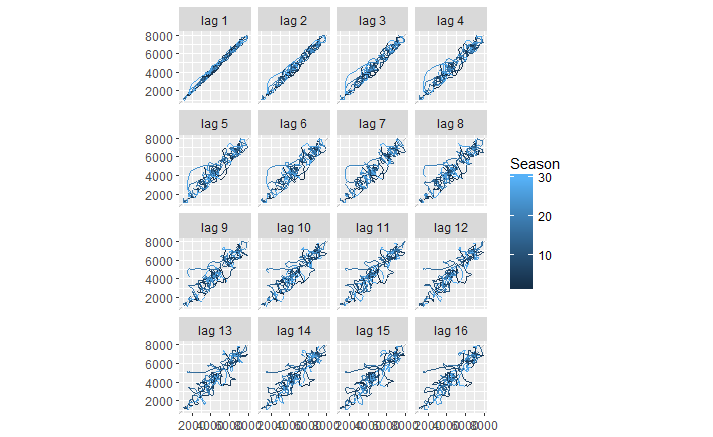
Autocorrelation of Time Series

Another way to look at time series data is to plot each observation against another observation that occurred some time previously. For example, you could plot ytyt against yt−1yt−1. This is called a lag plot because you are plotting the time series against lags of itself. The gglagplot() function produces various types of lag plots.

The correlations associated with the lag plots form what is called the “autocorrelation function”. Autocorrelation is nearly the same as correlation, which you can learn about in the Assessing Correlations tutorial. However, autocorrelation is the correlation of a time series with a delayed copy of itself. Autocorrelation between ytyt and yt−kyt−k for different values of *k* can be written as:

where *T* is the length of the time series. And similar to correlation, autocorrelation will always between +1 and -1.

Gglagplot

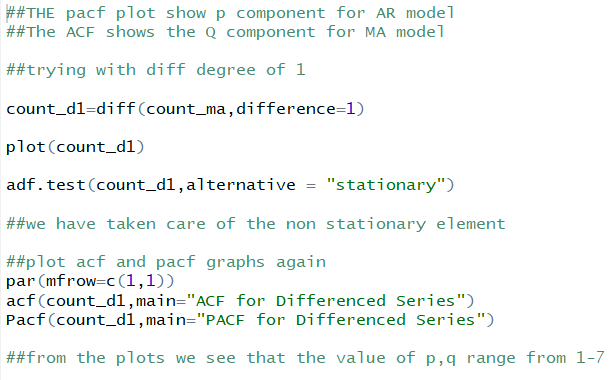


From these plots we see a high auto correlation occurring at lag 1. From the PACF plots also, we notice spike at order1.

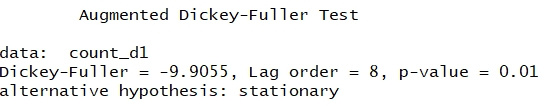
The PACF plots show the p component of an AR model

The ACF show the q component of an MA model

Assuming the order of differentiating factor is one:

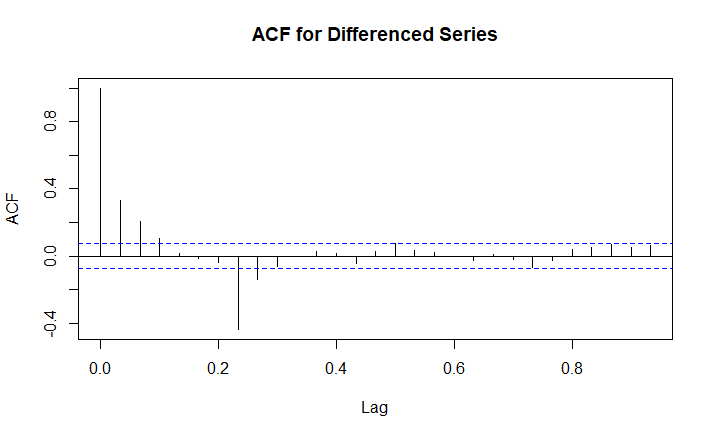


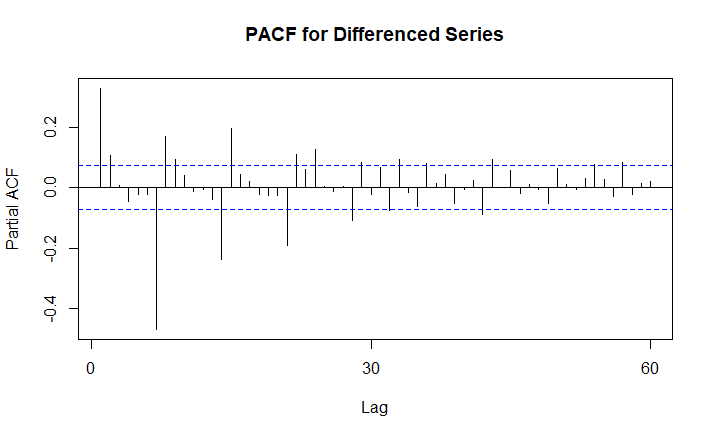
Testing the series again to see if it stationary or not



The test shows that the series is stationary now.

Plotting the treated time series

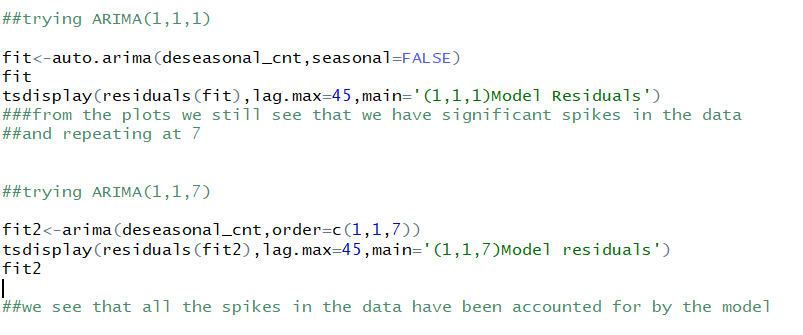




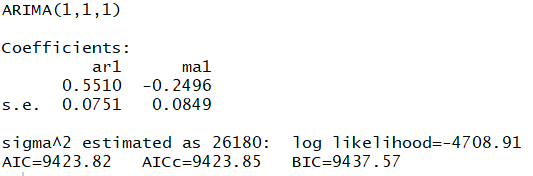
From the ACF and PACF plots we decide that the values of (p,q) range from 1,7

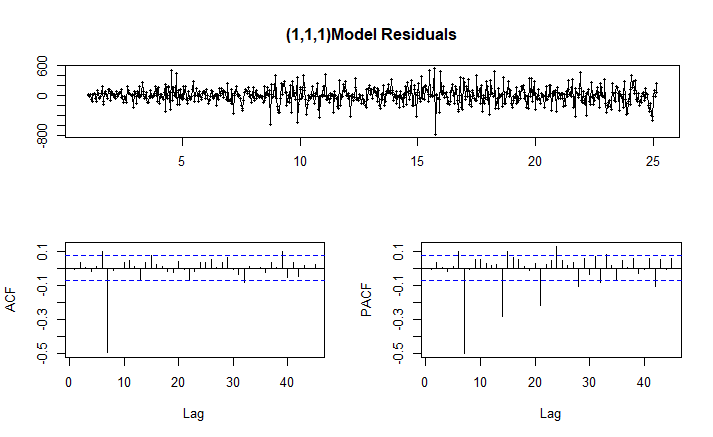
**FITTING ARIMA MODELS**

Two different models were tried to understand the results.



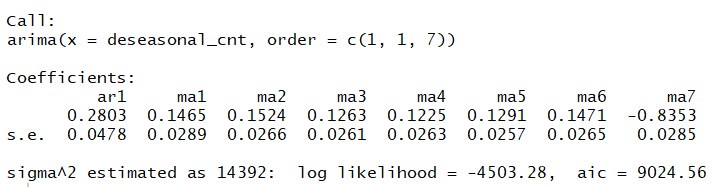
Results from ARIMA model (1,1,1)

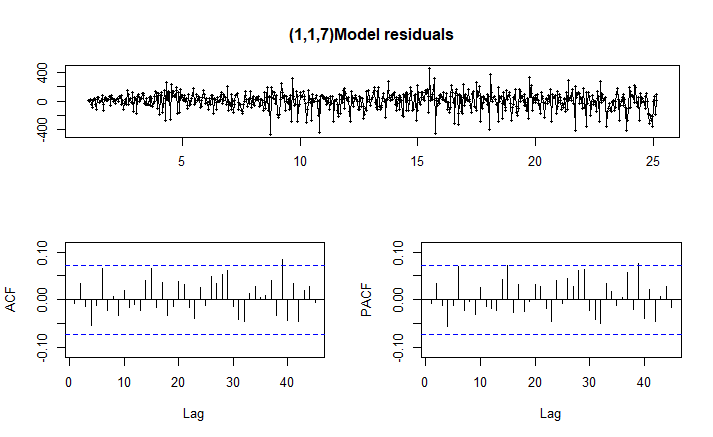




From the plots we can observe that most of the spikes have been treated for except the one at 7.

Results of ARIMA (1,1,7)

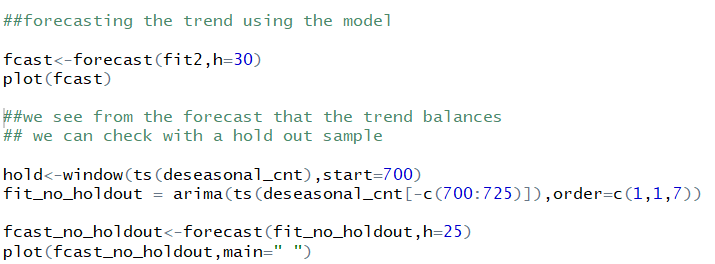




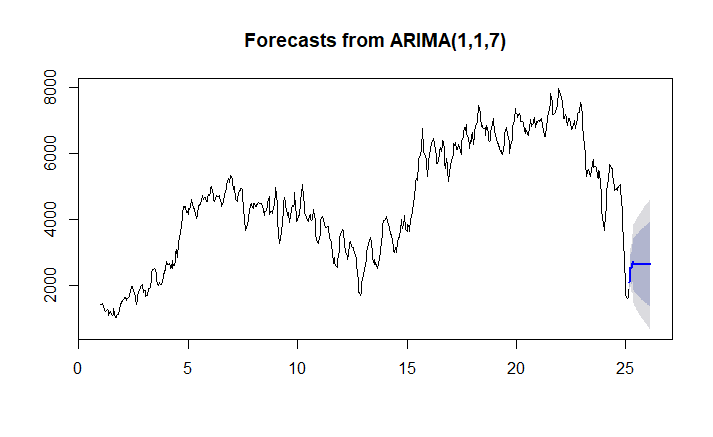
In the ACF and PACF plots we can see that all the lag variables are well within the 95% confidence interval. All the decay component is accounted for, by the model.

This is the final model that we select for our analysis.

**Forecasting data using the ARIMA (1,1,7) model**



**Without a hold out sample**



**With a holdout sample**

