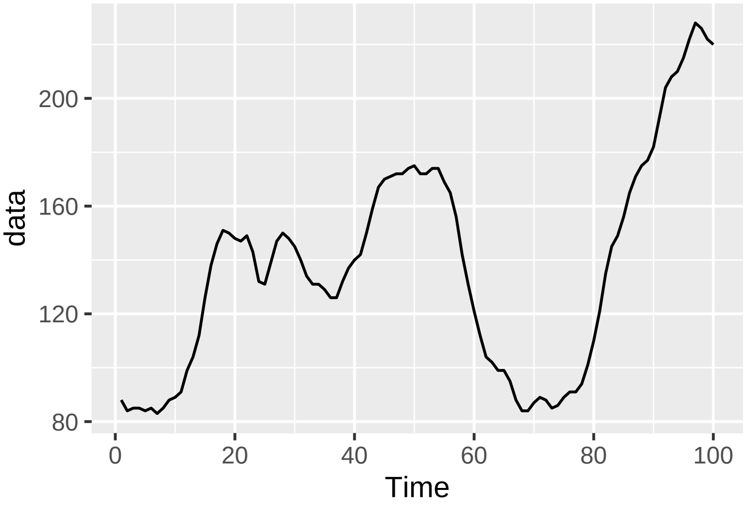
ARIMA Modelling on WWWusage Dataset

2024-02-29

```
library(fpp2)
```

```
data = read.csv("tsa1_data.csv")
data = ts(data$x, frequency = 1)
```

```
autoplot(data)
```

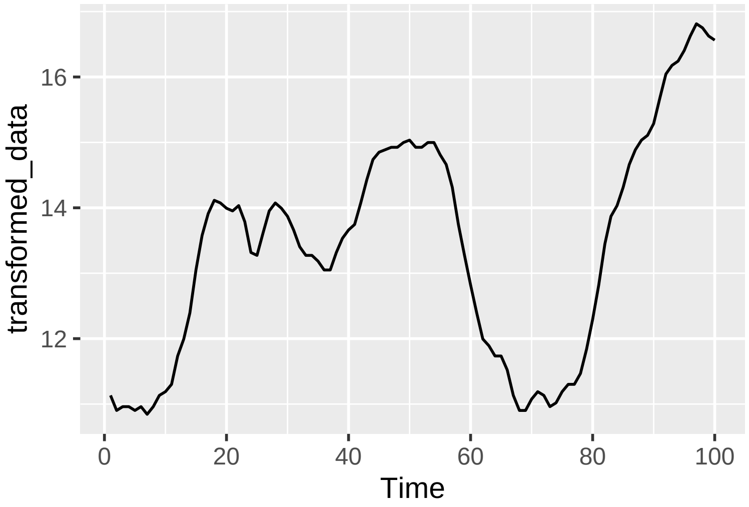


Remove Deterministic Variation

Since the time series exhibits non-constant variance (as seen in the plot above), and all values are positive, I will apply a Box-Cox transformation to stabilise the variance.

```
lambda = BoxCox.lambda(data)
transformed_data = BoxCox(data, lambda)
```

```
autoplot(transformed_data)
```



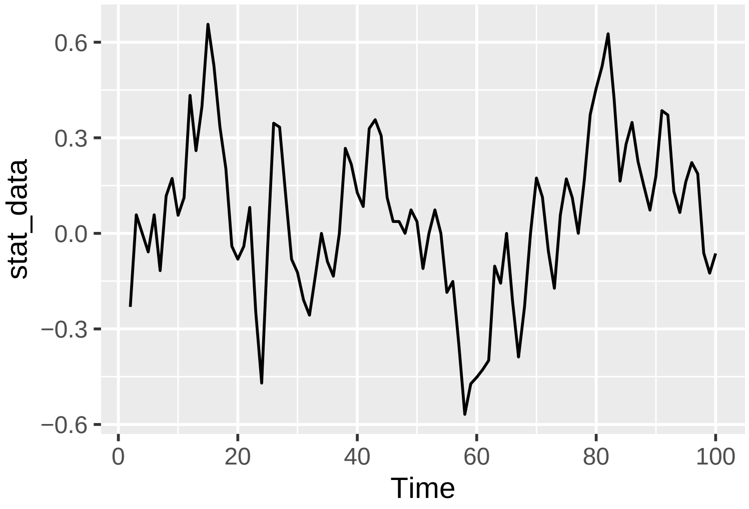
Stationarity

Looking at the plot of the data, it is not stationary. I will use differencing to make it stationary.

```
#nsdiffs(transformed_data) # 0, because it is non-seasonal data
ndiffs(transformed_data) # 1
```

```
## [1] 1
```

```
stat_data = diff(transformed_data)
autoplot(stat_data)
```



After differencing, I will use the KPSS test to confirm that the once-differenced data is stationary and does not have any unit roots present in the data.

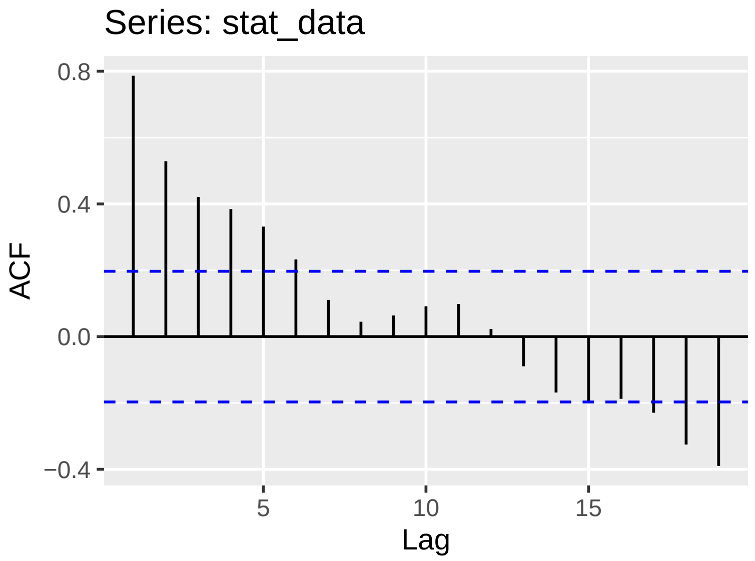
```r
library(urca)
summary(ur.kpss(stat_data))
```

```
##
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 3 lags.
##
## Value of test-statistic is: 0.1943
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

At all significance levels, the test-statistic = 0.1943 is smaller than all critical values. Hence, we do not reject the null hypothesis (H0) that the data are stationary.

Finally, I will look at the ACF plot to confirm that the once-differenced data is, in fact, stationary.

```r
ggAcf(stat_data)
```

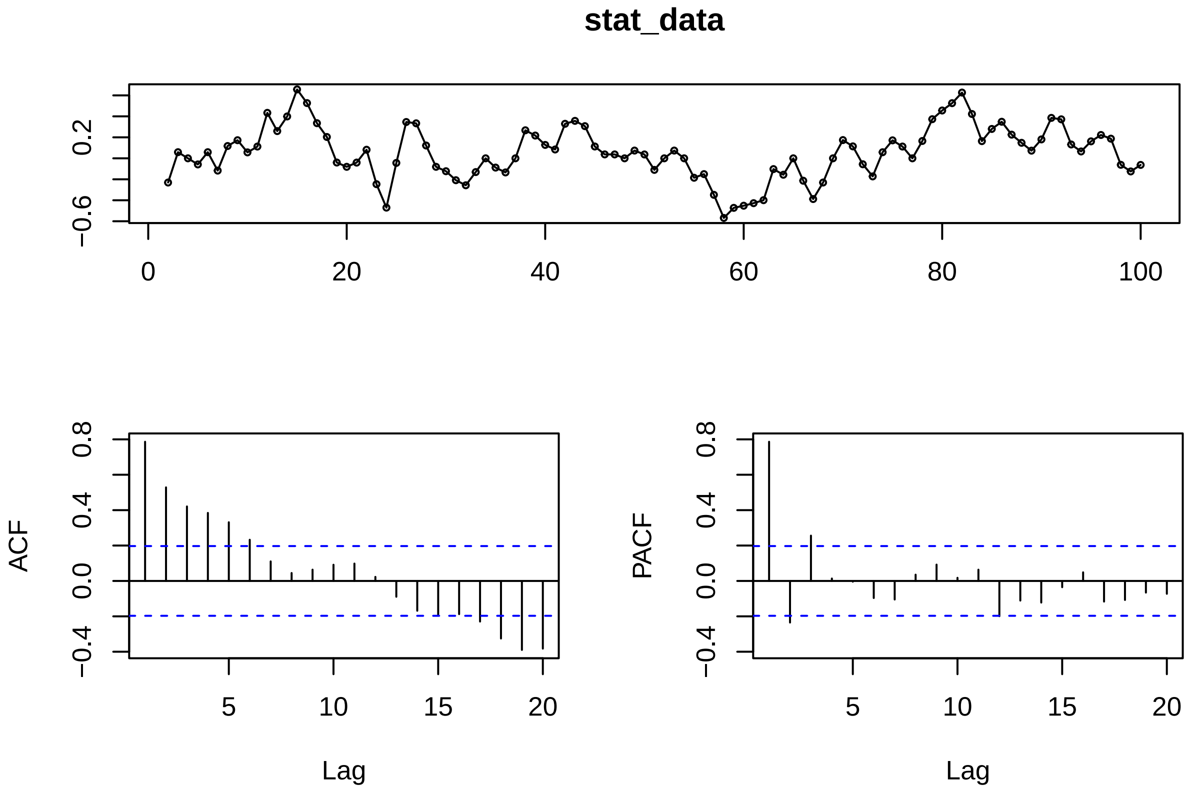


The autocorrelations drop off after the 6th lag, and most of the lags are within the confidence bands. Hence, based on the KPSS test and the ACF plot, I can conclude that there is no clear indication of non-stationarity, suggesting that the once-differencing has been effective in stabilising the mean of the series.

ARIMA Model

I will now fit the ARIMA(p,d,q) model. Firstly, d=1 because I have differenced the data once. Next, I will look at the ACF and PACF plots to determine the MA and AR orders respectively.

```r
tsdisplay(stat_data)
```



1. Firstly, looking at the ACF plot, there seems to be a gradual decay. This suggests that there is an AR component, but the order (p) cannot be determined from the ACF plot, since all AR(p) processes show similar decay patterns of ACF. Turning to the PACF, the lags drop off after lag 3, suggesting p=3. Hence, I will first try fitting an ARIMA(3,1,0) model.

2. Secondly, looking at the ACF plot, the lags drop off after the 6th lag, suggesting an MA order of q=6. Hence, I will also try fitting an ARIMA(0,1,6).

3. Lastly, I hypothesise that this could also be a mixed model with both AR and MA components. This is because of the gradual decay in the first 8 lags of the ACF plot, which suggests the presence of an AR component, but at the same time, makes it difficult to determine the exact order of the MA component. I will use multiple few steps to determine the AR and MA orders: first, I will start with an initial model of ARIMA(1,1,0), since the PACF shows a very significant spike at lag 1, compared to lags 2 & 3 which are much shorter. Based on the ARIMA(1,1,0) model, I will analyse the ACF plot of the residuals to determine the possible MA(q) order, which will give me the final ARIMA model with both AR and MA components.

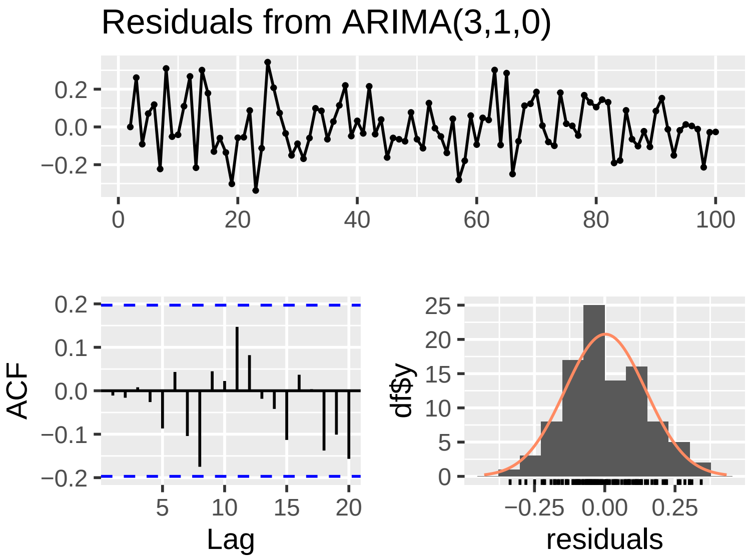
I will then compare the performance of all the models using the AIC from diagnostic checking.

Model 1: ARIMA(3,1,0)

```
fit1 = Arima(stat_data, order=c(3,1,0))
summary(fit1)
```

```
## Series: stat_data
## ARIMA(3,1,0)
##
## Coefficients:
##          ar1      ar2      ar3
##       0.1836  -0.3954  -0.0480
## s.e.  0.1027   0.0953   0.1027
##
## sigma^2 = 0.02132:  log likelihood = 50.84
## AIC=-93.67   AICc=-93.24   BIC=-83.33
##
## Training set error measures:
##                       ME      RMSE       MAE MPE MAPE     MASE         ACF1
## Training set 0.002332629 0.1430267 0.1146693 NaN  Inf 0.870001 -0.01103637
```

```
checkresiduals(fit1)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,0)
## Q* = 5.9399, df = 7, p-value = 0.5468
##
## Model df: 3.   Total lags used: 10
```

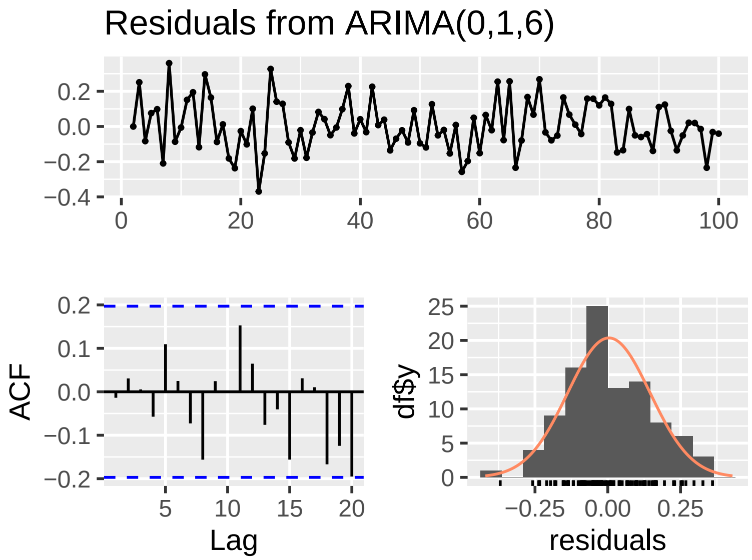
For Model 1, the AIC = -93.67. From the results of the Ljung-Box test, it is clear that we fail to reject the null hypothesis (H0) that there is no autocorrelation/ information in the residuals, as the p-value = 0.5468 > 0.05 at 5% significance level. This means that the model adequately captures all time series information in the data and there is no autocorrelation in the residuals. This can also be seen in the ACF plot of the residuals, where all the lags are within the confidence bands.

Model 2: ARIMA(0,1,6)

```
fit2 = Arima(stat_data, order=c(0,1,6))
summary(fit2)
```

```
## Series: stat_data
## ARIMA(0,1,6)
##
## Coefficients:
##          ma1      ma2      ma3     ma4      ma5      ma6
##       0.1999  -0.4296  -0.2477  0.1752  -0.0727  -0.0504
## s.e.  0.1054   0.1305   0.1207  0.1712   0.1058   0.1432
##
## sigma^2 = 0.02135:  log likelihood = 52.15
## AIC=-90.3   AICc=-89.06   BIC=-72.21
##
## Training set error measures:
##                       ME      RMSE       MAE MPE MAPE      MASE        ACF1
## Training set 0.003953705 0.1408405 0.1126783 NaN  Inf 0.8548952 -0.01376106
```

```
checkresiduals(fit2)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,6)
## Q* = 5.1328, df = 4, p-value = 0.2739
##
## Model df: 6.   Total lags used: 10
```

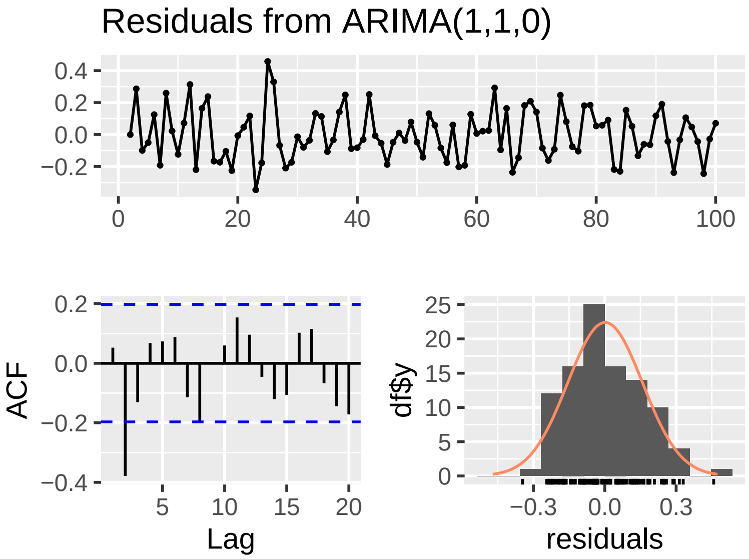
Model 2 performs slightly worse than Model 1 as Model 2's AIC = -90.3 is greater than Model 1. However, Model 2 is still adequate as the p-value = 0.2739 > 0.05 at 5% significance level, which means that we fail to reject the null hypothesis (H0). Furthermore, the ACF plot of the residuals also shows that all the lags fall within the confidence bands, meaning that there is no significant time series information left in the model.

Model 3: mixed model
Step 1: ARIMA(1,1,0)

```
fit3_1 = Arima(stat_data, order=c(1,1,0))
summary(fit3_1)
```

```
## Series: stat_data
## ARIMA(1,1,0)
##
## Coefficients:
##          ar1
##       0.1412
## s.e.  0.1013
##
## sigma^2 = 0.02504:  log likelihood = 42.12
## AIC=-80.24   AICc=-80.11   BIC=-75.07
##
## Training set error measures:
##                       ME      RMSE       MAE MPE MAPE      MASE       ACF1
## Training set 0.001512218 0.1566255 0.1289157 NaN  Inf 0.9780887 0.05250421
```

```
checkresiduals(fit3_1)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,0)
## Q* = 24.803, df = 9, p-value = 0.003197
##
## Model df: 1.   Total lags used: 10
```

For the initial ARIMA(1,1,0) model, the AIC = -80.24. Based on the results of the Ljung-Box test, it is clear that there is still time series information left in the residuals which are not captured by the model in

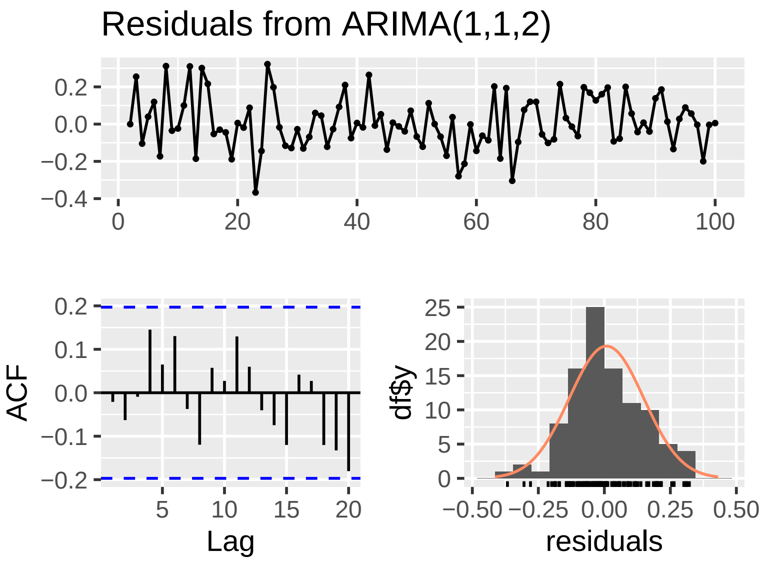
one or more of the first 10 lags. The p-value = 0.003197 < 0.05 at 5% significance level. Hence, we reject the null hypothesis (H0) that there is no autocorrelation/ information in the residuals. This model is therefore inadequate.
The ACF plot of the residuals shows that lag 2 falls outside the confidence bands. This not only suggests that there is time series information not fully captured by the model, but also indicates that there might be an MA component present in the data, since the ACF plot is an indicator of MA terms. Hence, I will try adding an MA component with both q=1 and q=2.

Step 2: adding q=2
ARIMA(1,1,2)

```
fit3_2 = Arima(stat_data, order=c(1,1,2))
summary(fit3_2)
```

```
## Series: stat_data
## ARIMA(1,1,2)
##
## Coefficients:
##          ar1      ma1     ma2
##       0.6536  -0.5180  -0.482
## s.e.  0.0912   0.1051   0.100
##
## sigma^2 = 0.02056:  log likelihood = 51.26
## AIC=-94.52   AICc=-94.09   BIC=-84.18
##
## Training set error measures:
##                       ME      RMSE       MAE MPE MAPE      MASE        ACF1
## Training set 0.008206427 0.1404543 0.1092966 NaN  Inf 0.8292375 -0.02088172
```

```
checkresiduals(fit3_2)
```



```
##
##  Ljung-Box test
##
```

```
## data:  Residuals from ARIMA(1,1,2)
## Q* = 7.1316, df = 7, p-value = 0.4153
##
## Model df: 3.   Total lags used: 10
```

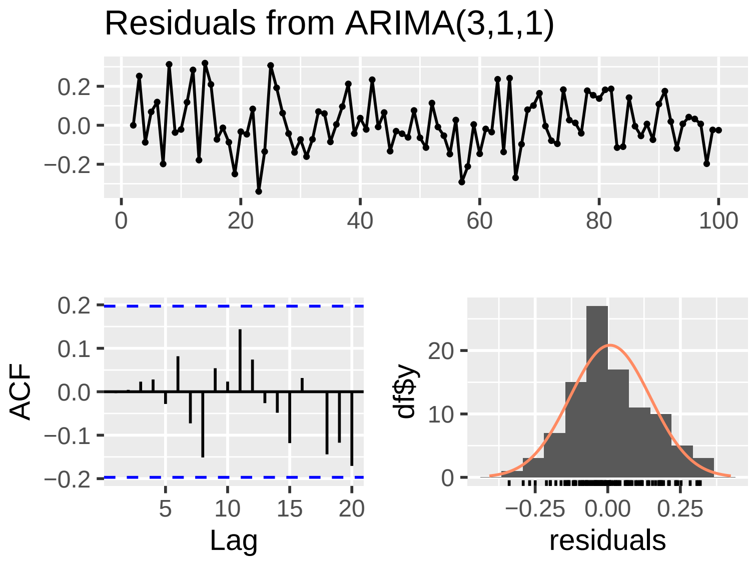
The ARIMA(1,1,2) model is a much better fit, as it has a lower AIC = -94.52. From the Ljung-Box test, since the p-value = 0.4153, at 5% significance level, we do not reject the null hypothesis (H0) that there is no autocorrelation/ information in the residuals. Looking at the ACF plot of the residuals, all of the lags are within the confidence bands. The ARIMA(1,1,2) adequately captures all time series information in the data.

Step 3: adding q=1. For this model, I will try p=3 instead, since the AR(3) model produced fairly good results, and the PACF of the data also suggests that the AR component has an order of 3.
ARIMA(3,1,1)

```
fit3_3 = Arima(stat_data, order=c(3,1,1))
summary(fit3_3)
```

```
## Series: stat_data
## ARIMA(3,1,1)
##
## Coefficients:
##          ar1      ar2     ar3      ma1
##       1.1050  -0.5802  0.3135  -1.0000
## s.e.  0.0989   0.1372  0.0980   0.0288
##
## sigma^2 = 0.0201:  log likelihood = 53.13
## AIC=-96.25   AICc=-95.6   BIC=-83.33
##
## Training set error measures:
##                      ME      RMSE       MAE MPE MAPE      MASE         ACF1
## Training set 0.00804099 0.1381633 0.1081012 NaN  Inf 0.8201681 -0.003164626
```

```
checkresiduals(fit3_3)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,1)
## Q* = 4.4187, df = 6, p-value = 0.6202
##
## Model df: 4.   Total lags used: 10
```

The ARIMA(3,1,1) model performed the best out of all the models, with an AIC = -96.25. Based on the results of the Ljung-Box test, we fail to reject the null hypothesis (H0) as the p-value = 0.6202 > 0.05 at 5% significance level. The ACF plot of the residuals also lies within the confidence bands for all the lags.

In conclusion, the final model I will use is ARIMA(3,1,1).