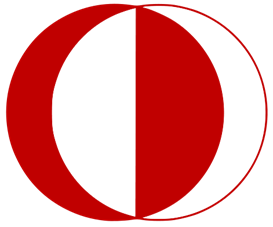
****

**MIDDLE EAST TECHNICAL UNIVERSITY**

**Department of Statistics**

**STAT 497 APPLIED TIME SERIES ANALYSIS**

**PROJECT**

**Contributors**

Gizem Cemile ÇELİK

İrem EKMEKÇİ

**TABLE OF CONTENT**

**Introduction…………………………………………………………………………………..3**

**Analysis………………………………………………………………………………………..3**

**Conclusion…………………………………………………………………………………...16**

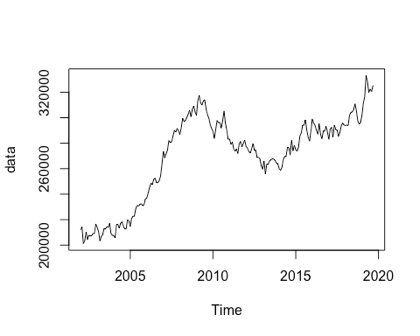
**Appendix…………………………………………………………………………………….18**

**Introduction**

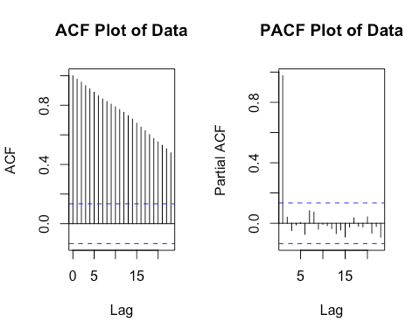
The project dataset is downloaded from [fred.stlouisfed.org](http://fred.stlouisfed.org/). This data provides information about monthly estimates of the total dollar value of construction in the U.S from 2002 to 2019. As cited in Census.gov, data estimates include the cost of labor and materials, cost of architectural and engineering work, overhead costs, interest and taxes paid during construction, and contractor’s profits. A million dollars define units of the value of spending. The project aims that develop a time series analysis using total public construction spendings information and offer an efficient model to use for generating future expenses.

**Analysis**

Before starting any analysis, it is better to look at the plot of the data.



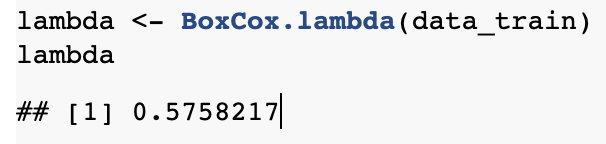
By looking at this plot, it can be said that the data has an increasing trend until 2010. However, between 2010-2015 there are factors that cause decline in trend. Later, there is also an increasing trend. It is also better to look at ACF and PACF plots.



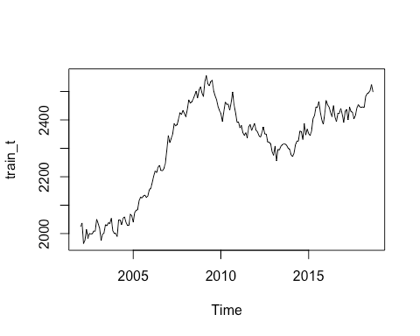
ACF plot shows slow decay that is an indication of being non-stationarity. Non-stationarity means that the mean and the variance of the observations change over time. Also, it is known that the series is seasonal because the data set is monthly.

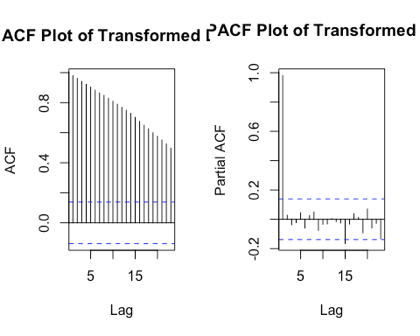
Firstly, it is necessary to divide data into two as train and test to measure the forecast accuracy of the models.

Secondly, Box-Cox transformation should be done to stabilize variance.



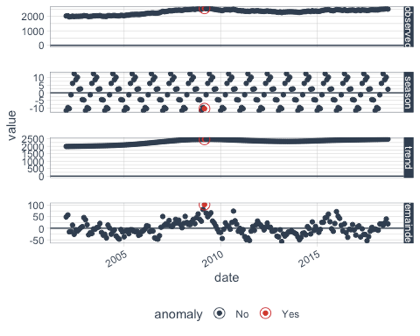
Since lambda value is not 1, Box Cox transformation is applicable. After transformation, the trend of the data still looks the same.

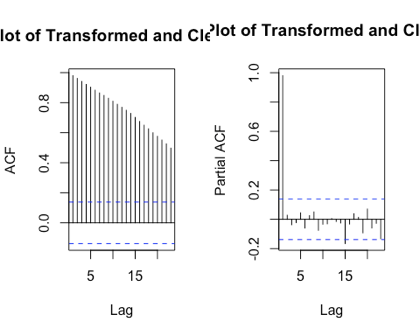




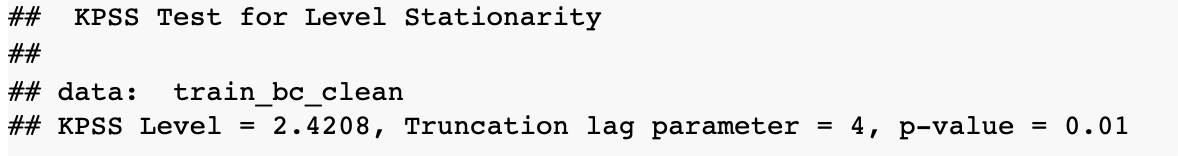
ACF shows slow decay that is an indication of being non-stationarity.

In order to detect unusual observations, anomaly detection is used. Furthermore, one observation is detected as an anomaly. After that this anomaly observation is cleaned from data.



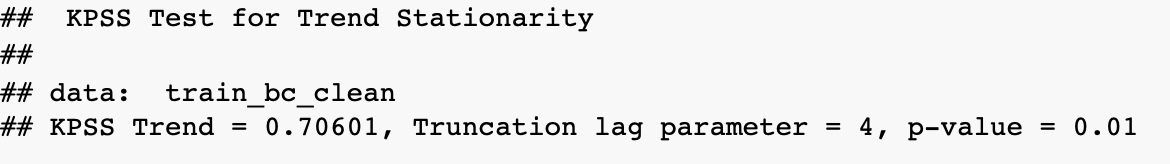


After anomaly detection and cleaning data process, the system still looks non-stationary. For this reason, it is better to look at KPSS test.



*H0: The system is stationary. H1: The system is non-stationary.*

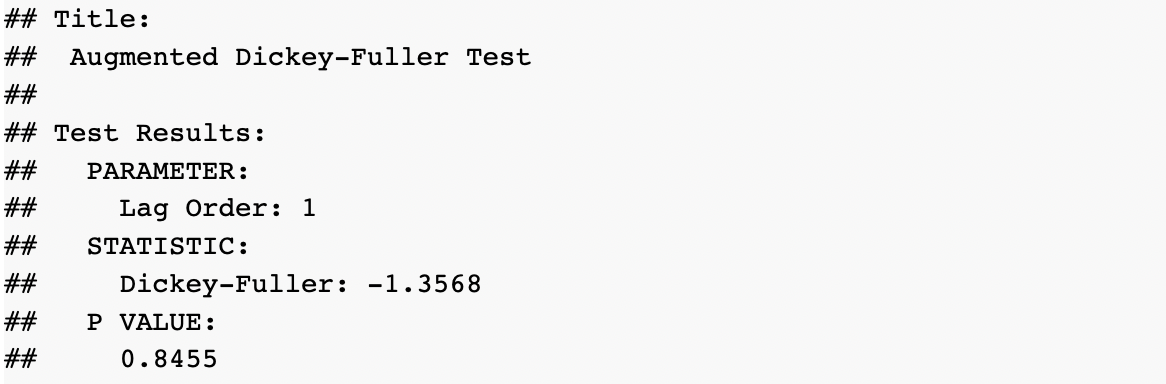
It can said that the system is non-stationary since p-value is less 0.05. Now, it can be looked at trend of the data.



*H0: There is a deterministic trend. H1: There is a stochastic trend.*

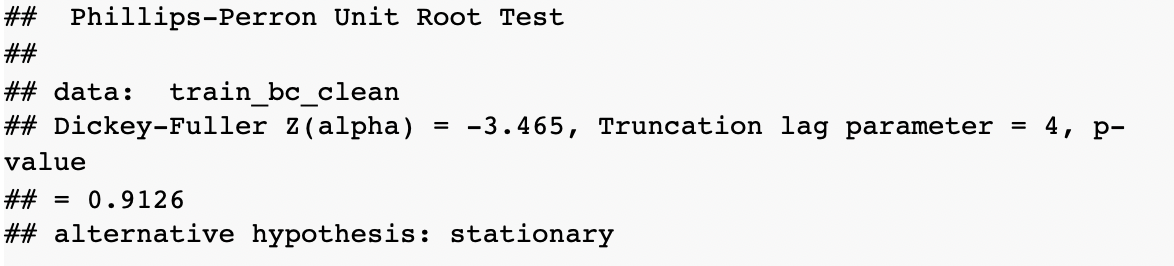
Since p-value is less than 0.05, we reject H0, and we can say that there is a stochastic trend.

ADF test and PP test are used to check unit root problems.



*Ho: The process has unit root (non-stationary). H1: The process is stationary.*

Since p value is greater than 0.05, we fail to reject H0. It means that we don’t have enough evidence to claim that we have a stationary system.

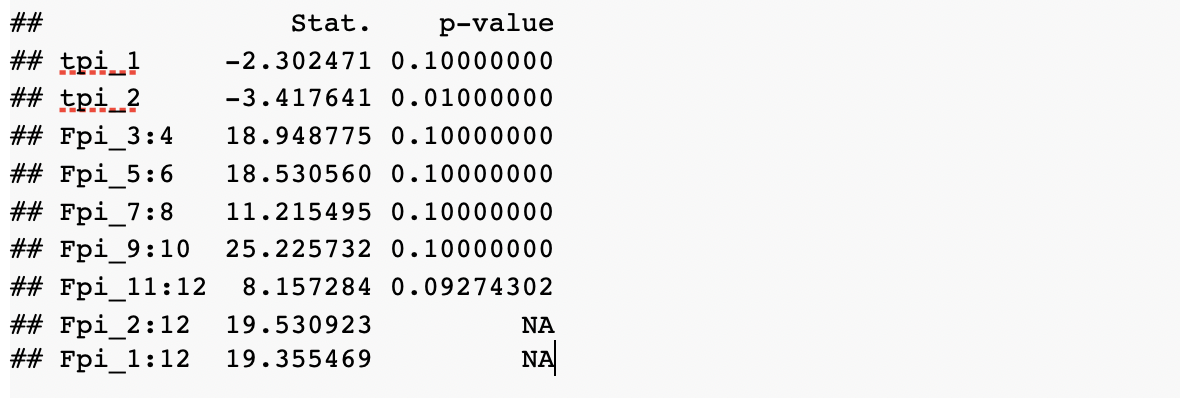


*Ho: The process has unit root (non-stationary). H1: The process is stationary.*

Since p value is greater than 0.05, we fail to reject H0. It means that we don’t have enough evidence to claim that we have a stationary system.

KPSS, ADF and PP Tests show that there is a unit root problem in the system.

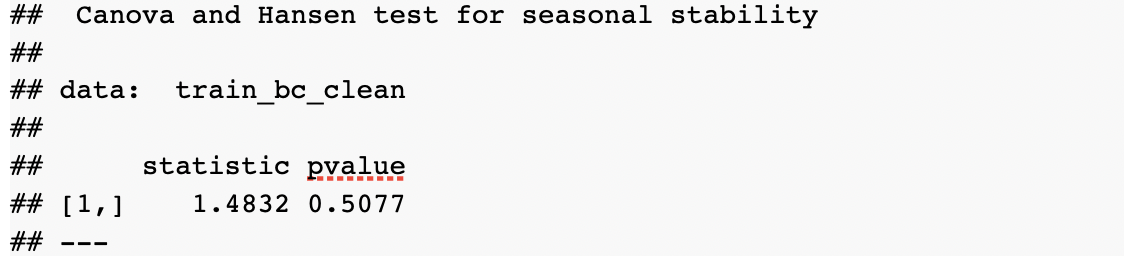
For seasonal unit root, HEGY and Canova-Hansen tests are used. The result below is from HEGY test.



In this output, p value of tpi\_1 gives the result for regular unit root and the p value of Fpi\_11:12 gives the result for testing seasonal unit root.

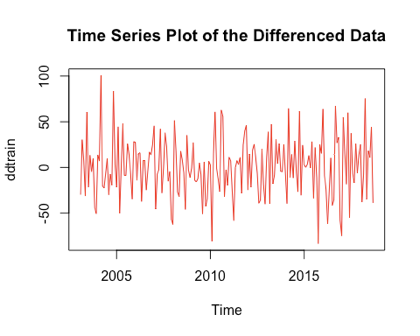
The output shows that the system has both regular and seasonal unit roots because both of the p-values of tpi\_1 and Fpi\_11:12 are greater than 0.05.

Canova and Hansen (CH) test statistic for the null hypothesis of a stable seasonal pattern.

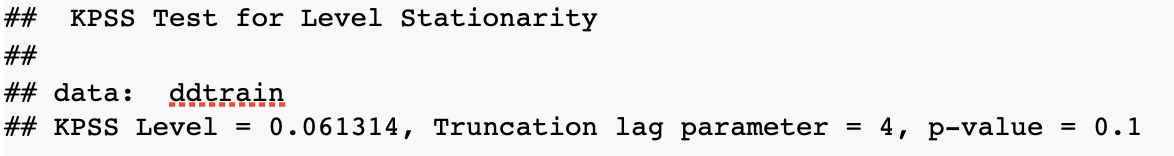


The output above shows that according to Canova-Hansen test, the seasonal pattern is stable.

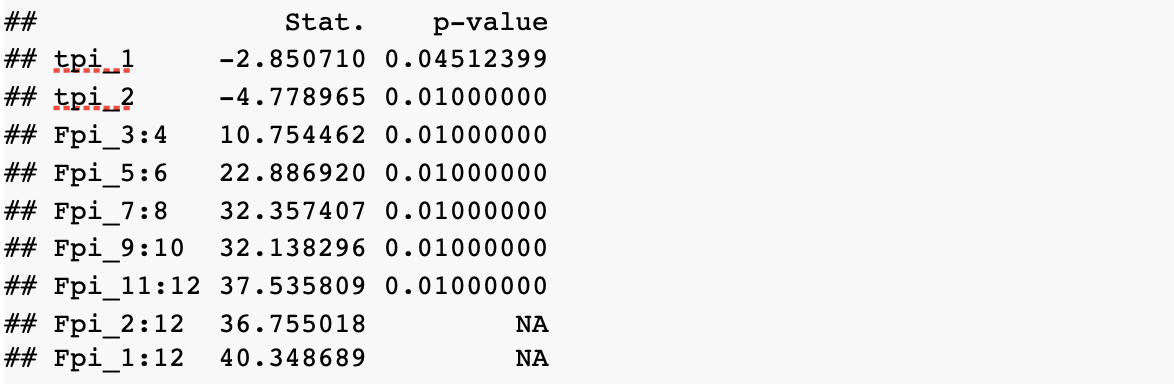
In order to remove the trend, differencing is necessary. Then, regular differencing is taken. Taking regular difference did not solved the regular unit root. (p-value of tpi\_1 is greater than 0.05) Therefore, seasonal differencing is also taken.



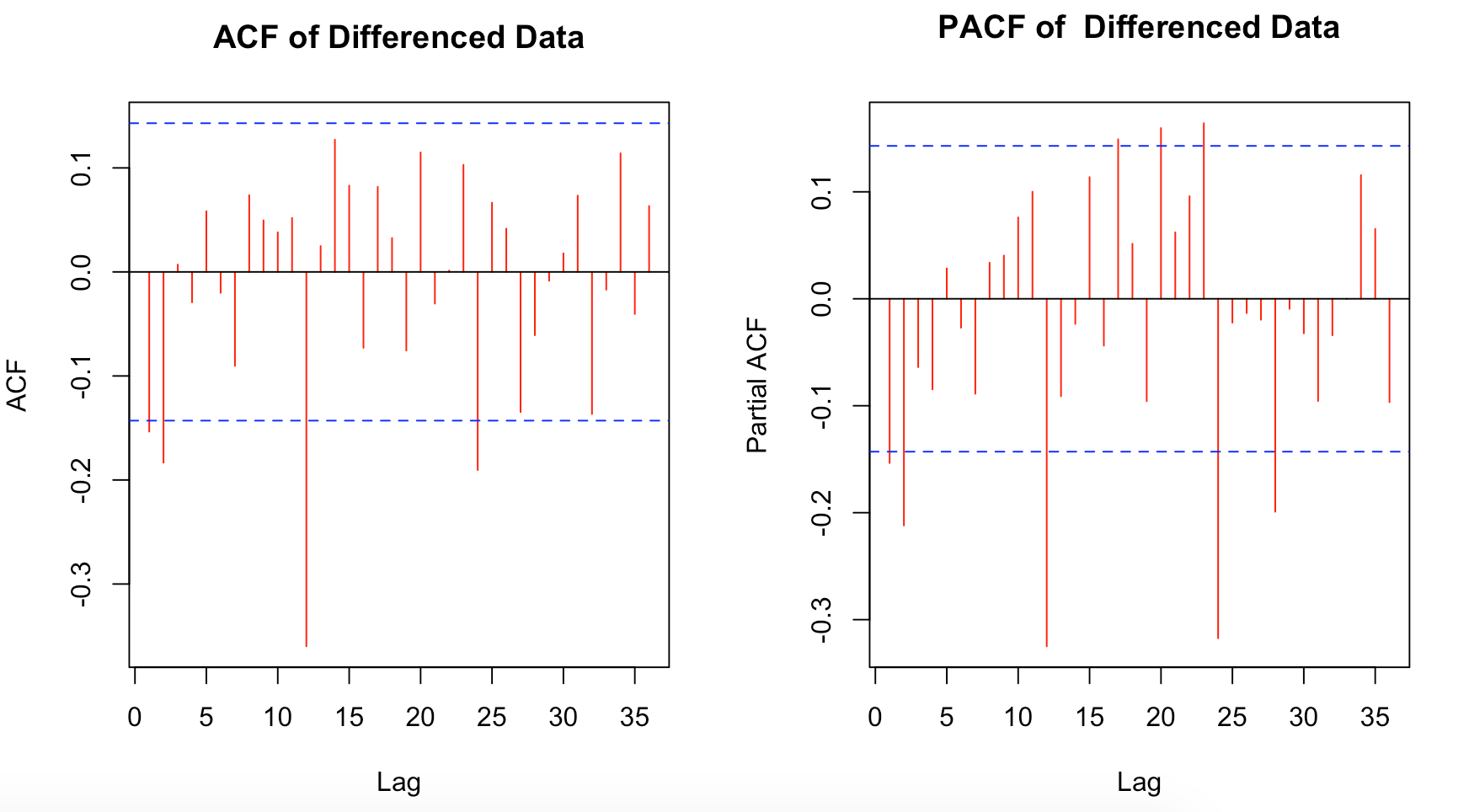
The process seems stationary around mean 0 after taking differences. But, it is better to check this with KPSS test.



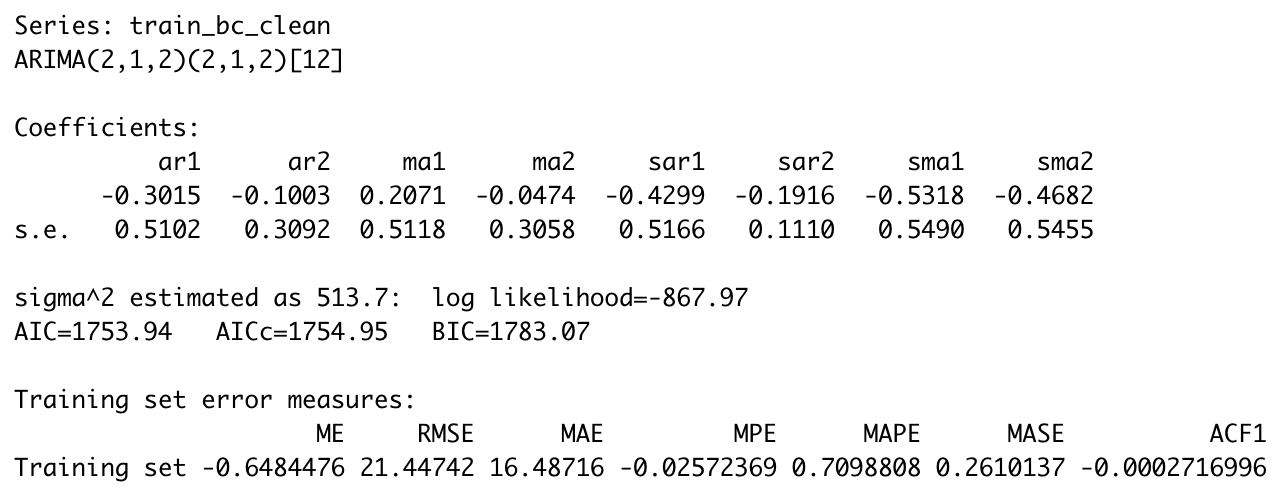
Since the p-value is greater than 0.05,the process is stationary.



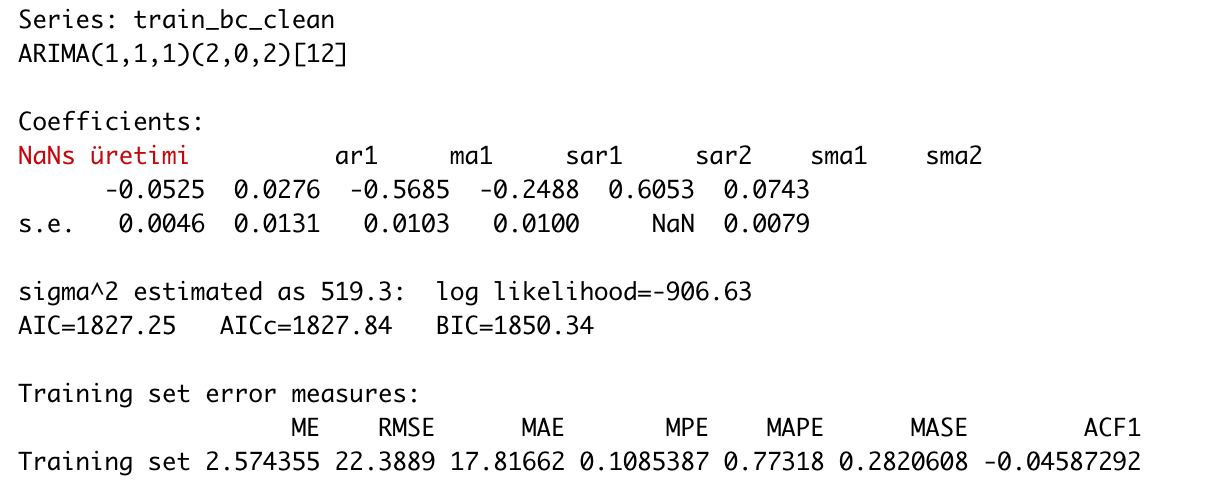
Taking both regular and seasonal differences solved both regular and seasonal unit root problems (p-values of tpi\_1 and Fpi\_11:12 are less than 0.05)



The model *SARIMA(2,1,2)(2,1,2)[12]* can be suggested by looking at ACF and PACF graphs of differenced data. In addition, from *auto.arima* the model *SARIMA(1,1,1)(2,0,2)[12]* can be suggested. Now, it is time to check significance of these models.To do so, we will look at the last estimated parameters of each component. If the ratio between these estimates and their standard errors (s.e) are greater than +2 or less than -2 ,it can said that these parameters are significant and the model is significant.

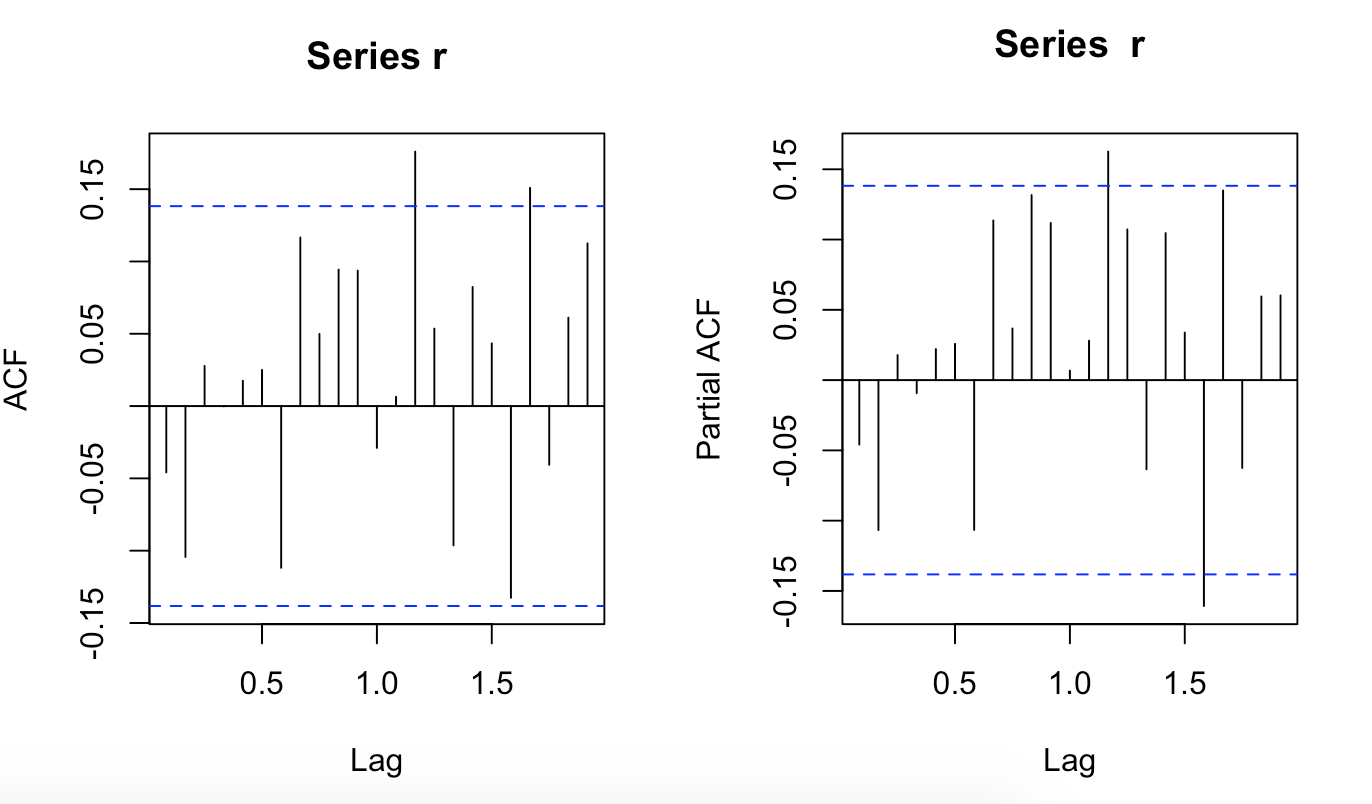


So, this model is not significant.

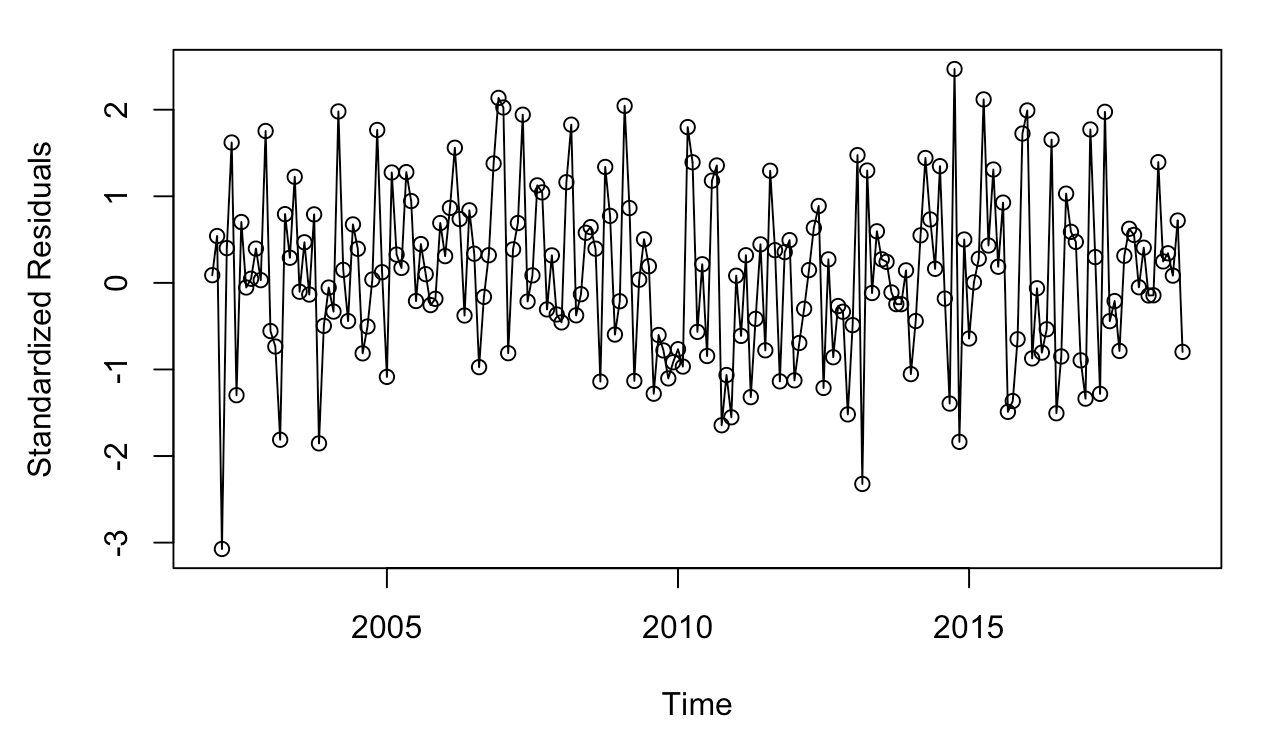


On the contrary, this model is significant.

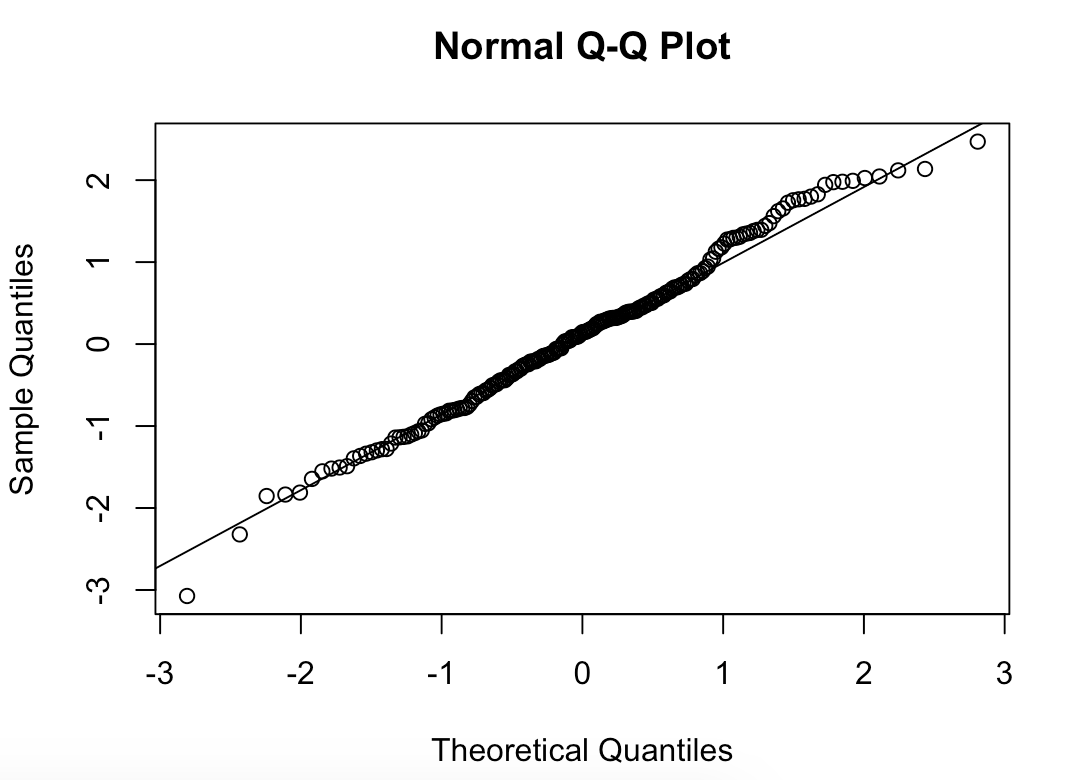
Since only SARIMA(1,1,1)(2,0,2)[12] model is significant, we continue diagnostic check with this model. Let us look at ACF and PACF plots of residuals.



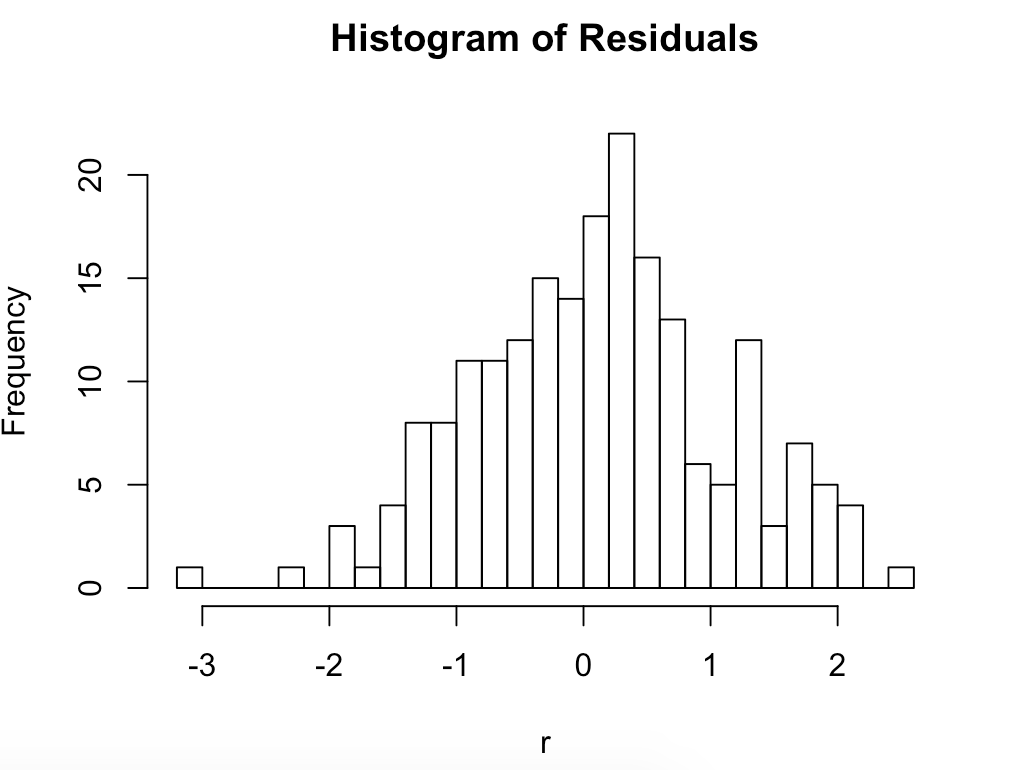
For all time points, ACF and PACF values should be in the white noise bands. However, some of the spikes are outside of the bands. This means there might be some outliers.



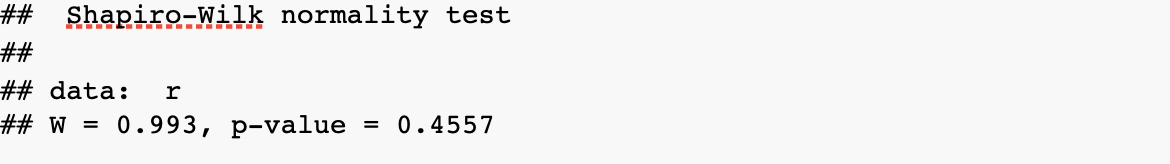
The standardized residual vs time plot also shows some outliers.

*Checking Normality of Residuals*

By looking at Normal Q-Q Plot, it can be said that residuals are normally distributed.



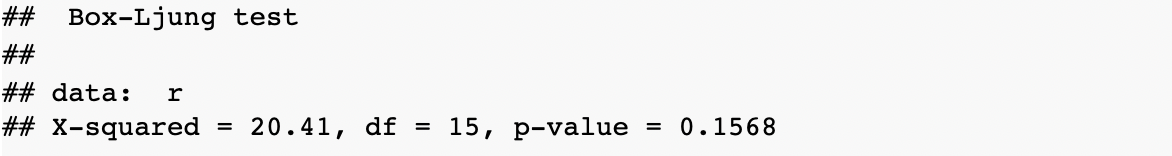
The histogram of residuals shows that they might have a normal distribution with outliers.



*Ho: Residuals have normal distribution.H1: Residuals do not have normal distribution.*

p-value is greater than 0.05, so fail to reject H0. We don’t have enough evidence to claim that residuals don’t have normal distribution.

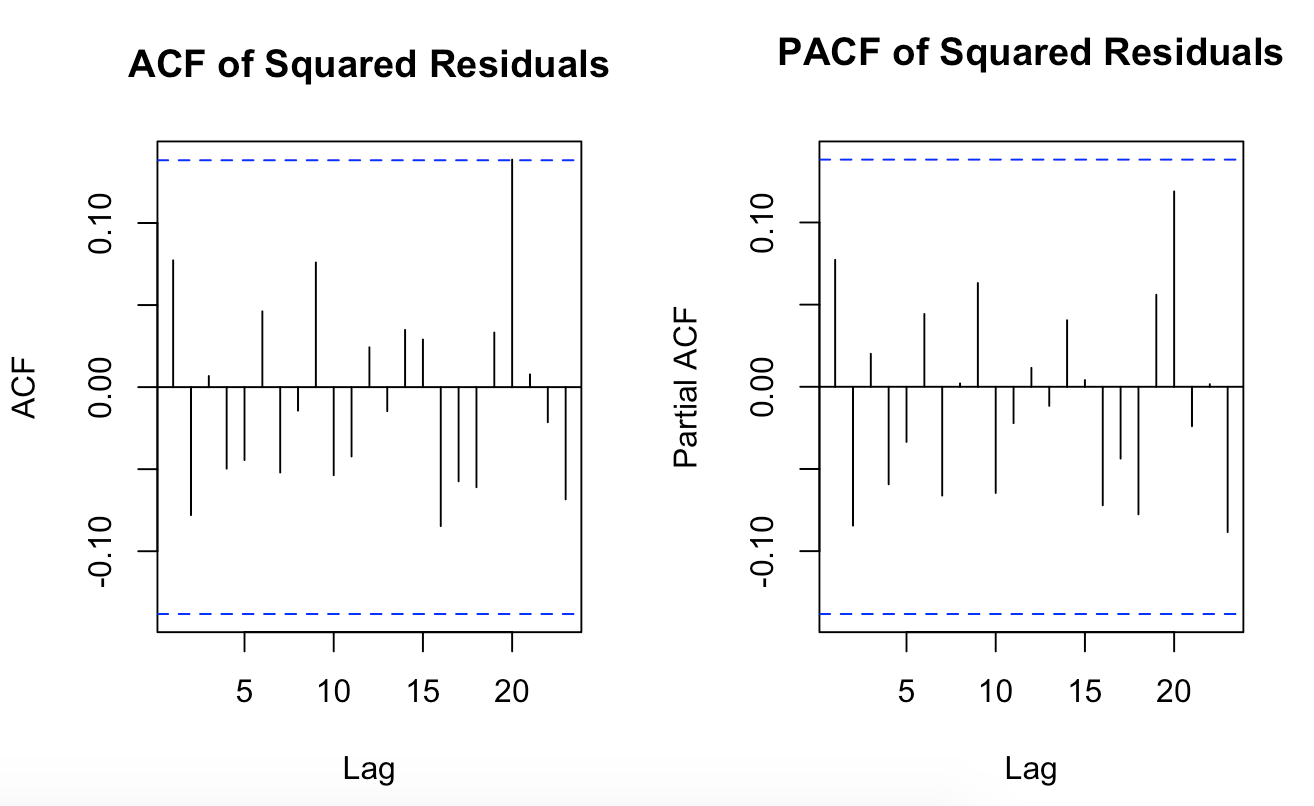
*Checking Autocorrelation of Residuals*



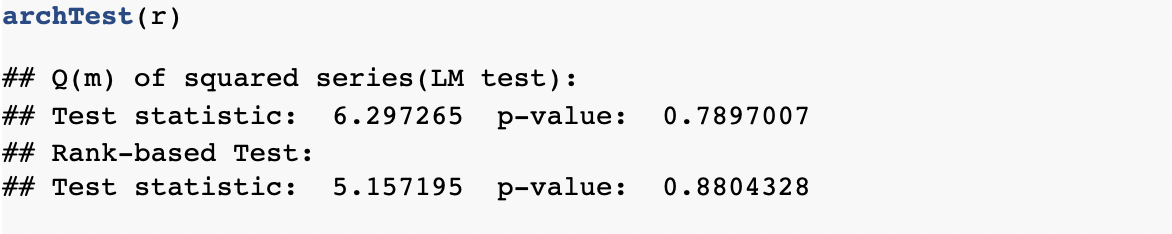
Since p value is greater than α, we are 95% confident that the residuals of the model are uncorrelated, according to results of Box-Ljung Test.

*Checking Heteroscedasticity*

For the heteroscedasticity, the ACF-PACF plots of the squared residuals should be looked.



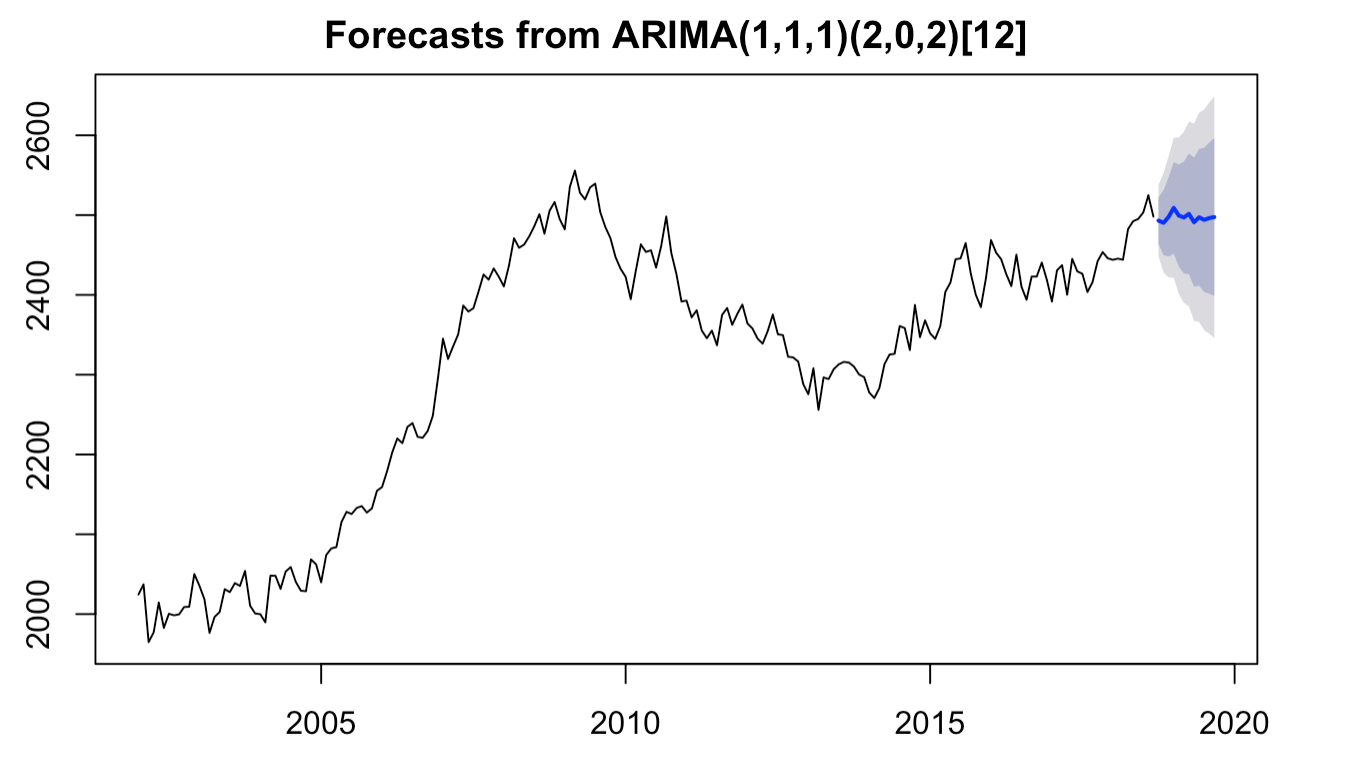
Both plots shows that there is no spike out of the white noise bands that is an indication of homoscedasticity.

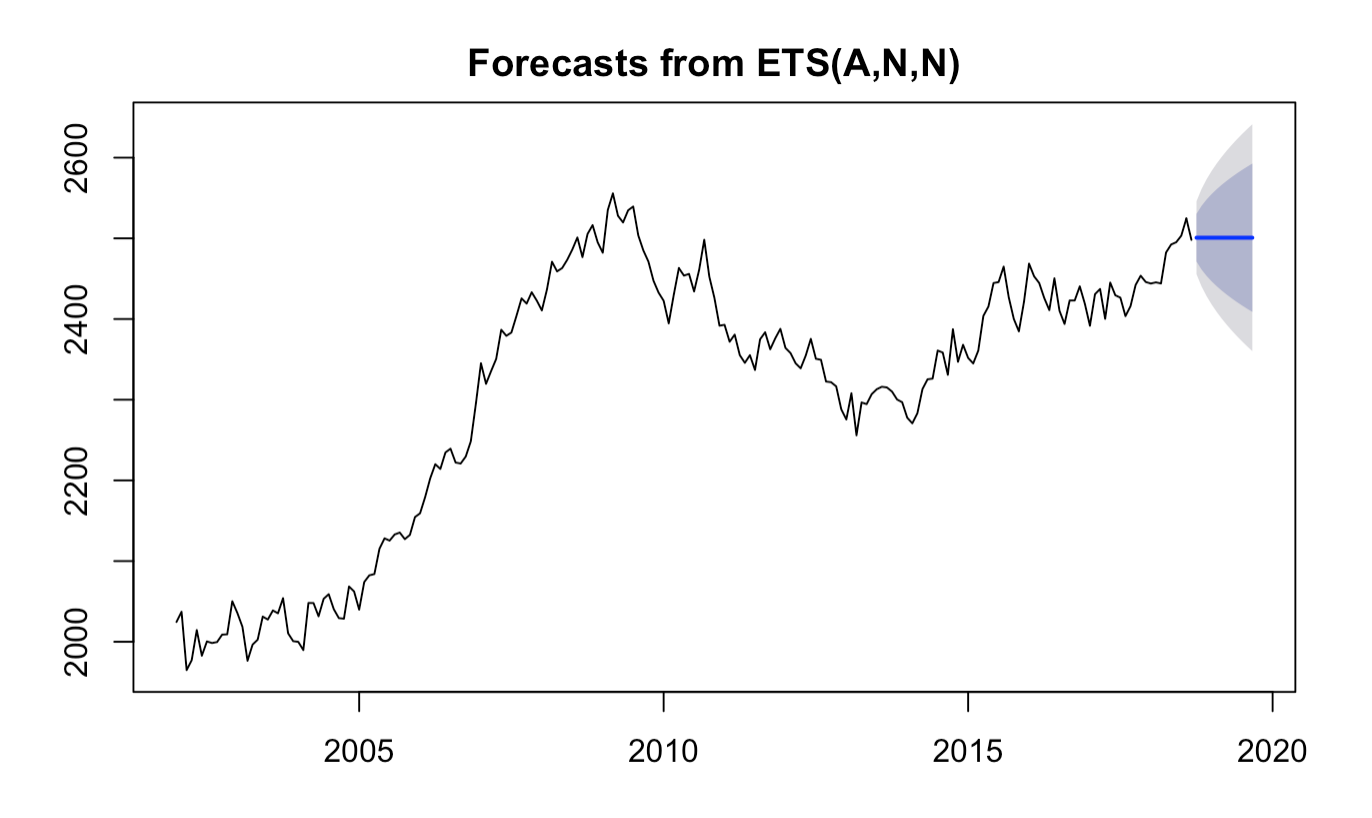


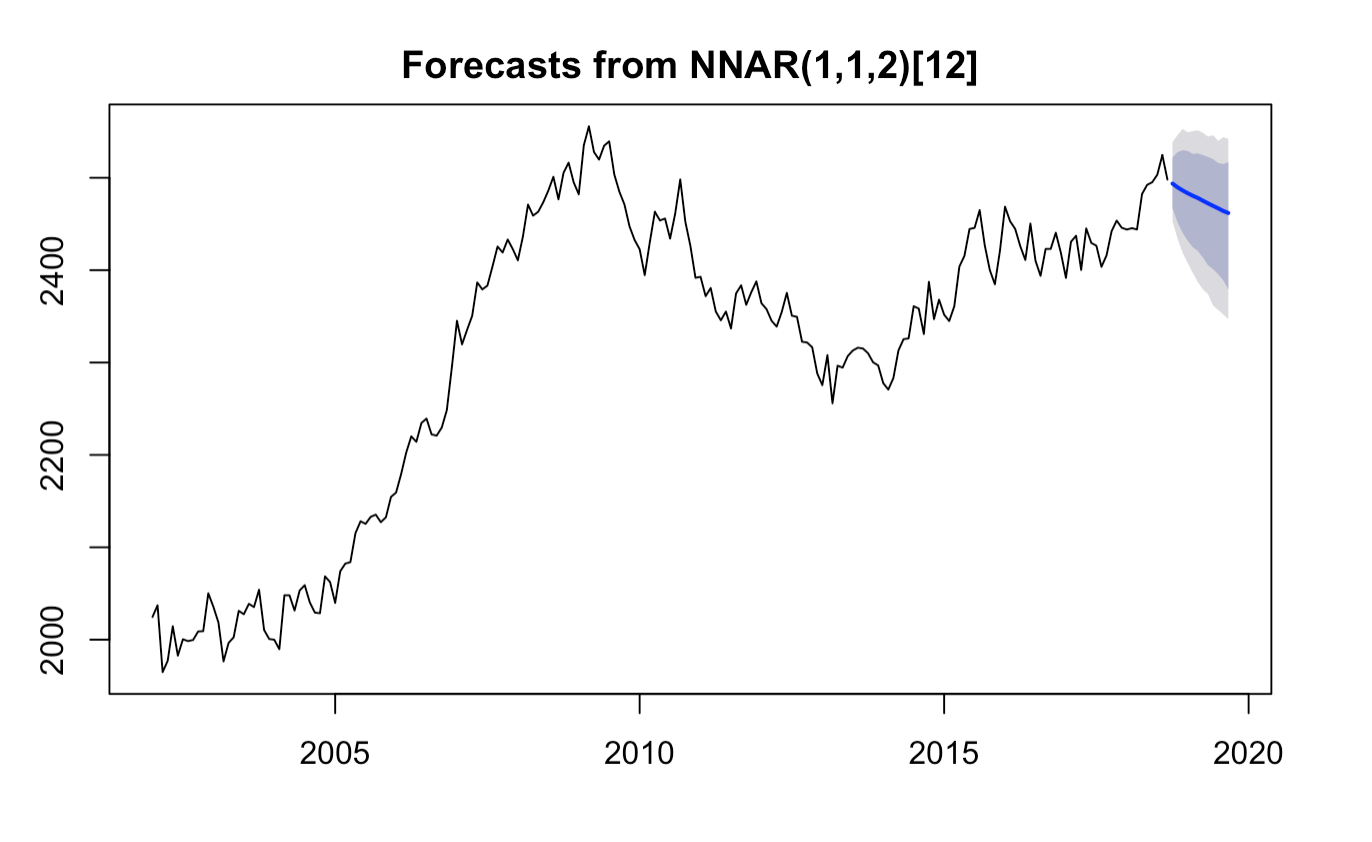
It is homoscedastic, according to archTest. All tests show that there is no heteroscedasticity problem.

As a result, all of the assumptions checked.

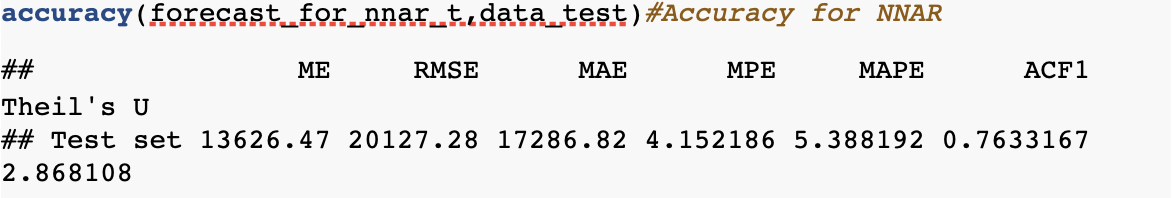
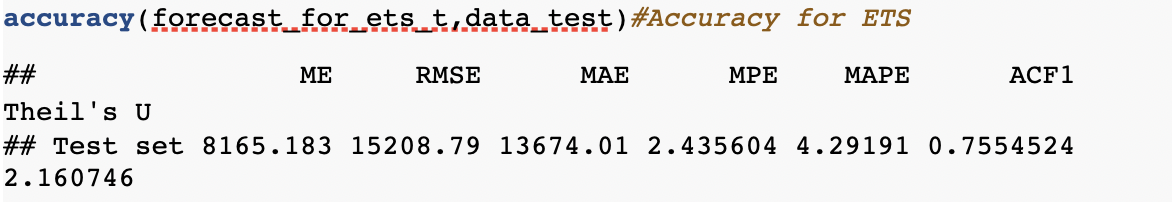
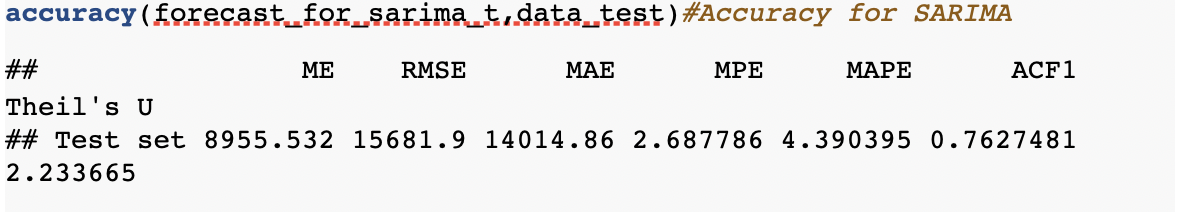
Forecasts are obtained from SARIMA(1,1,1)(2,0,2)[12], Holt’s exponential smoothing method, and neural networks (*nnetar*).

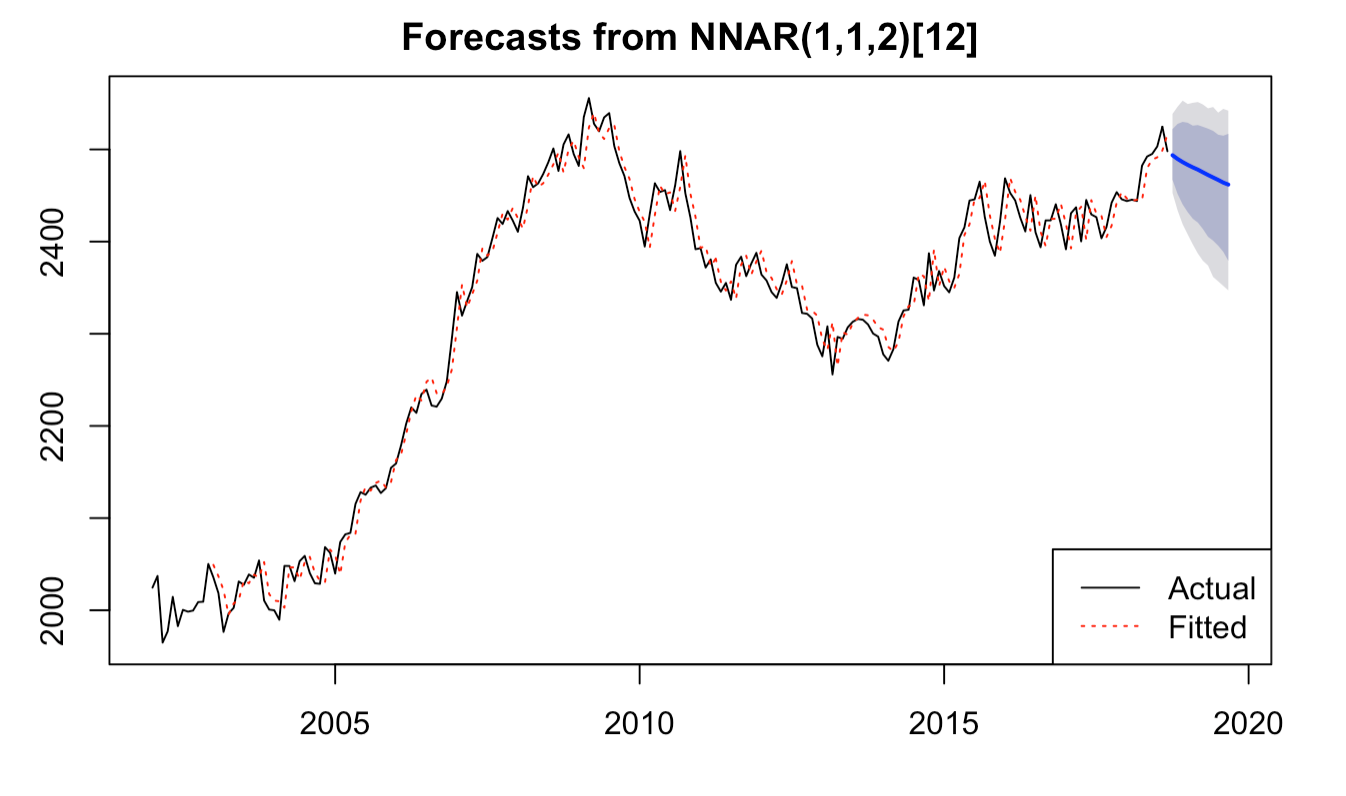
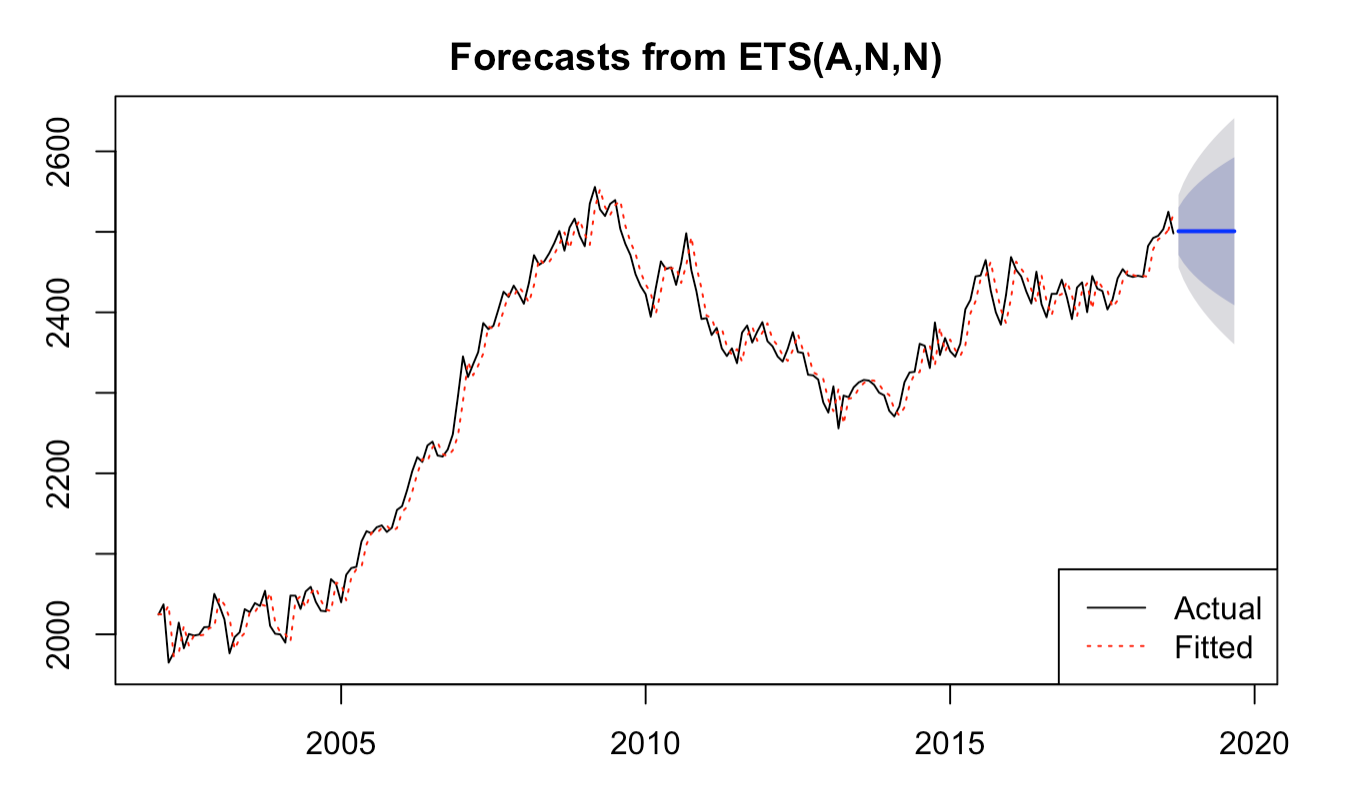
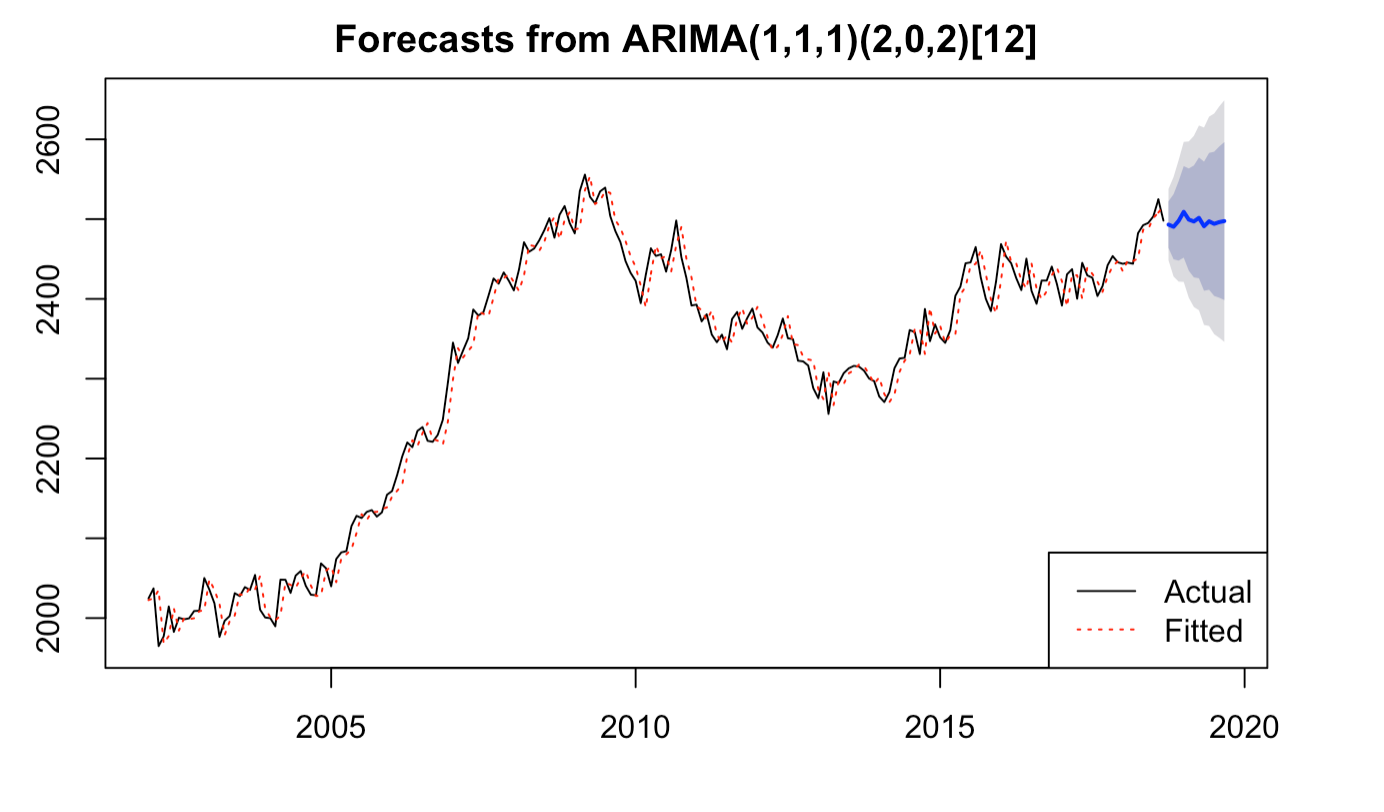






Since at the beginning the series is transformed for SARIMA model, back transform is done the series to reach the estimates for the original units.

At the end,the forecast accuracy measures are calculated for each model.

ETS is the best technique to forecast since its MAPE value is the lowest.

**Conclusion**

In conclusion, firstly the time series is plotted to understand behaviour. And it is found that series is non-stationary and follows an increasing trend. Then, data is splitted into two parts in order to measure accuracy of forecasts. Box Cox transformation is applied to stabilize variance after that. Later on, anomaly detection is made and if the series is cleaned from anomalies. KPSS test shows that the system is non-stationary and follows stochastic trend. Moreover, ADF and PP tests also show the series is non-stationary(unit root). For unit roots, HEGY test is used and found that there are both regular and seasonal unit roots.By taking both regular and seasonal differences are solved both regular and seasonal unit root problems. The series also became stationary. Then, by looking ACF and PACF plots model is suggested. Also, from auto.arima function a model is suggested. These models are *SARIMA(2,1,2)(2,1,2)[12]* and *SARIMA(1,1,1)(2,0,2)[12].*But only *SARIMA(1,1,1)(2,0,2)[12]* is significant. Moreover, diagnostics are checked for this model and all of the assumptions are checked.

Secondly, forecasts are obtained with SARIMA model, ETS and neural networks. And back transformation is done because at first Box-Cox transformation is applied. Accuracy measurements are done for forecasts. Moreover, it is found that ETS is the best technique to forecast since its MAPE value is the lowest.