Pollution
Dirty air: how India became the most polluted country on earth
© Illustration by Steven Bernard/FT

TIME SERIES ANALYSIS AND FORECASTING ASSIGNMENT

SUBMITTED BY:

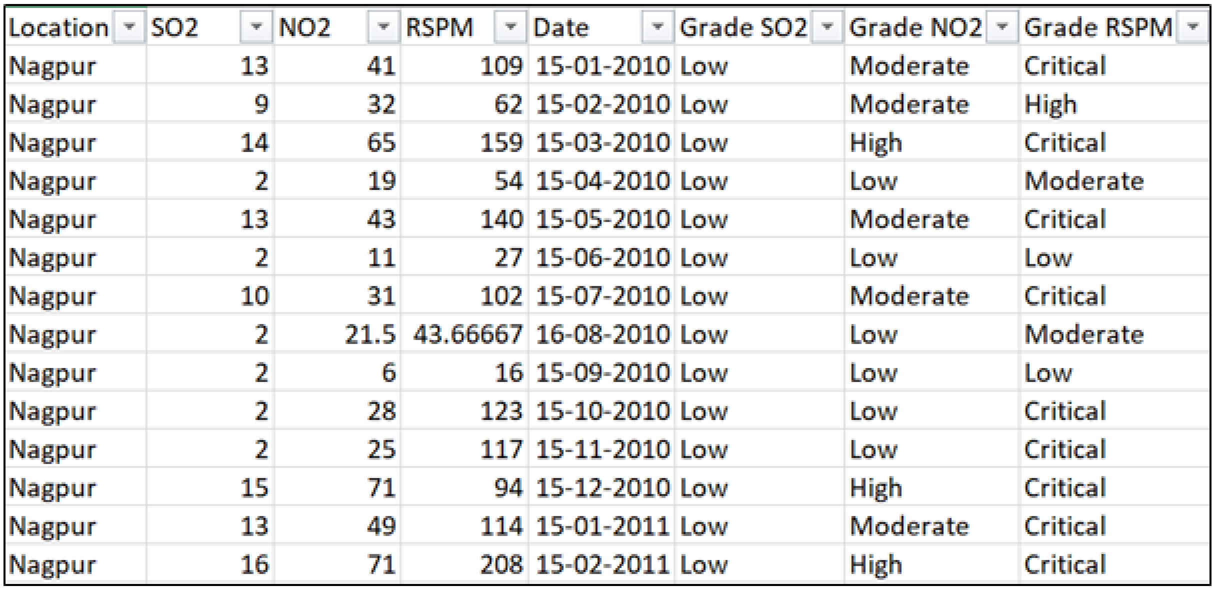
Amisha Dhawan 1640837
Devanshi Saini 1640850
Mahek Mowar 1640859

OBJECTIVE OF STUDY:

Air pollution is a global problem and can be perceived as a modern-day curse. One way of dealing with it is by finding economical ways to monitor and forecast air quality.
India has far more people living in heavily polluted areas. At least 140m people in India are breathing air 10 times or more over the WHO safe limit. The results are alarming: not just the number of people breathing in polluted air, but those breathing air contaminated with particulates that are multiple times over the level deemed safe — 10 micrograms of PM2.5 per cubic metre — by the World Health Organisation. Top officials in prime minister Narendra Modi's government have suggested New Delhi's air is little dirtier than that in other major capitals such as London.
Maharashtra is the second-most populous state and third-largest state by area in India. It is the wealthiest state by all major economic parameters and also the most industrialized state in India.
Air pollution cost Maharashtra 1,08,038 lives in 2017, the second highest in the country, according to a study published in the Lancet Planetary Health. The state saw 86.9 air pollution-related deaths for every 1 lakh people, making it the 15th in the average deaths in the country. The study also revealed the average life expectancy of Maharashtra would have been 1.5 years higher if the air pollution levels were less.

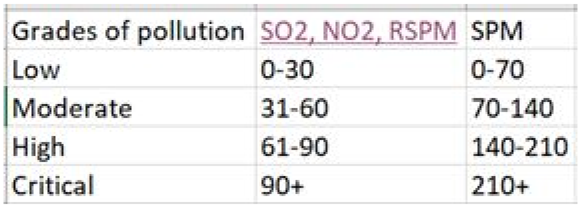
ABOUT THE DATASET:

The data given below has been extracted from data.gov.in. The city of choice is Nagpur to draw conclusions about its transformation from a clean city in 2010 to one of the world's polluted cities in the World. It was ideal to select Nagpur from our data of cities like Amravati, Aurangabad, Chandrapur, Nanded, Mumbai, etc since Nagpur has seen most hikes in pollution levels. Also, seasonal changes in pollution index is observed.
A sample of pollution levels of each month from 2010 till 2015 is taken.



Location	SO2	NO2	RSPM	Date	Grade SO2	Grade NO2	Grade RSPM
Nagpur	13	41	109	15-01-2010	Low	Moderate	Critical
Nagpur	9	32	62	15-02-2010	Low	Moderate	High
Nagpur	14	65	159	15-03-2010	Low	High	Critical
Nagpur	2	19	54	15-04-2010	Low	Low	Moderate
Nagpur	13	43	140	15-05-2010	Low	Moderate	Critical
Nagpur	2	11	27	15-06-2010	Low	Low	Low
Nagpur	10	31	102	15-07-2010	Low	Moderate	Critical
Nagpur	2	21.5	43.66667	16-08-2010	Low	Low	Moderate
Nagpur	2	6	16	15-09-2010	Low	Low	Low
Nagpur	2	28	123	15-10-2010	Low	Low	Critical
Nagpur	2	25	117	15-11-2010	Low	Low	Critical
Nagpur	15	71	94	15-12-2010	Low	High	Critical
Nagpur	13	49	114	15-01-2011	Low	Moderate	Critical
Nagpur	16	71	208	15-02-2011	Low	High	Critical

The given grades are given via approved scales depicted below:



Grades of pollution	SO2, NO2, RSPM	SPM
Low	0-30	0-70
Moderate	31-60	70-140
High	61-90	140-210
Critical	90+	210+

METHODOLOGY:

Data smoothing is often required within the environmental data analysis. Moving average smoothing is a time series constructed by taking the averages of several sequential values of another time series.
Simple exponential smoothing is the method used with data showing no trend or seasonality. Since pollution index shows variability according to seasons, simple exponential smoothing cannot be used. The data shows seasonality and hence, Winter's Seasonal forecasting is employed.
Suppose that the data follows the model $Zt = mt + st + et$ with $E(et) = 0$ and the trend component mt is a linear function with $mt = mn + (t-n)\beta$, $\forall$ t= n, n+1,….

THEORY:

The arrangement of data in accordance with their time of occurrence is a Time Series. It is essentially the *chronological arrangement* of data.

The are various forces which affect the values of an observation in a time series. They are the components of a time series, namely,

1. **Trend:** The trend shows the general tendency of the data to increase or decrease during a long period of time.
2. **Seasonal Variation:** These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year.
3. **Random or Irregular movements:** They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic.

We have used the following concepts from Time Series Analysis for our project:

1. **Stationary Time Series:** We are aware that a time series model can only be built, if Time Series is stationary. Now, a time series is said to be stationary only if it holds the following conditions true:
   - The mean value of time-series is constant over time, which implies, the trend component is nullified.
   - The variance does not increase over time. (This property is known as *homoscedasticity*.)
   - The seasonality effect is minimal. This means that it is devoid of trend or seasonal patterns, which makes it looks like a random *white noise* irrespective of the observed time interval.
   - The covariance of the ith term and the (i + m) th term should not be a function of time.

There are two ways of making data stationary. They are:

- **Detrending** (Removing **trend**, removing **seasonality**).
- **Differencing**.

Trend: The variation of observations in a time series over a **long** period of time is known as Trends.

Seasonality: A repetitive pattern which can be predicted is termed as Seasonality. The variation of observations in a time series caused due to **regular or periodic** time variations is known as Seasonality.

Differencing: *Differencing* a time series means, to subtract each data point in the series from its successor.
For most time series patterns, 1 or 2 differencing is necessary to make it a stationary series. But if the time series appears to be seasonal, a better approach is to difference with respective season's data points to remove seasonal effect. If needed, differencing can be done again with successive data points.
But, we need to know about the number of differencing which is needed. The "nsdiffs" and "ndiffs" from *forecastpackage* can help find out the number of *seasonal differencing* and *regular differencing* respectively, needed to make the series stationary.

2. Simple moving average: A simple moving average (SMA) is an arithmetic moving average which is calculated by adding recent closing prices and then dividing that by the number of time periods in the calculation average.

A simple, or arithmetic, moving average that is calculated by adding the closing price of the security for a number of time periods and then dividing this total by that same number of periods.

3. Simple Exponential Smoothing: The simple exponential smoothing method provides a way of estimating the level at the current time point. Smoothing is controlled by the parameter alpha; for the estimate of the level at the current time point. The value of alpha lies between 0 and 1. If the values of alpha are close to 0, it means that little weight is placed on the most recent observations when making forecasts of future values.

4. Holt Winters Exponential Smoothing: Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three

parameters: alpha, beta, and gamma, for the estimates of the level, slope b of the trend component, and the seasonal component, respectively, at the current time point.
The parameters alpha, beta and gamma have values between 0 and 1, and values that are close to 0 mean that relatively little weight is placed on the most recent observations when making forecasts of future values.

5. Dickey fuller test: Dickey Fuller Test is used to test if a time series stationary or not. A P-Value of less than “0.05” in adf.test(), i.e, indicates that it is stationary.
If the null hypothesis gets rejected, we’ll get a stationary time series.

6: Box test: Ljung-Box test is used to test if a time series stationary or not.
A P-value of less than 0.05 indicates that it is stationary.

7. Lag: When the time base is shifted by a given number of periods, a Lag of time series is created.

8. Autocorrelation: The correlation between values of data points at different time intervals is known as Autocorrelation. It is also sometimes termed as Lagged Correlation.ACF is the coefficient of correlation between the value of a point at a current time and its value at lag p. i.e. correlation between Y(t) and Y(t-p).

9. Partial autocorrelation: Correlation of the time series with a lag of itself, with the linear dependence of all the lags between them removed.
PACF is same as ACF just that the intermediate lags between t and t-p is removed i.e. correlation between Y(t) and Y(t-p) with p-1 lags excluded.

10. ARIMA model: ARIMA is generally represented as ARIMA(p,d,q) where,

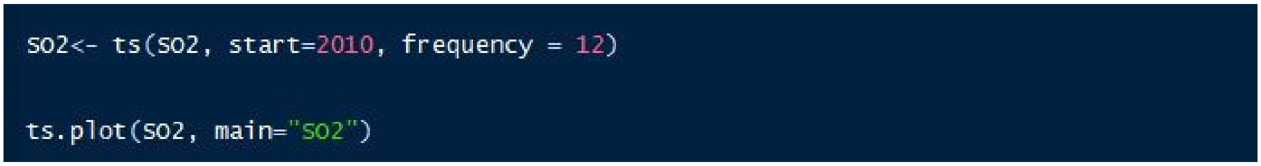
- p is past values
- d is differentiating order
- q is error term

While exponential smoothing methods do not make any assumptions about correlations between successive values of the time series, in some cases a better predictive model can be made by taking correlations in the data into account.

Autoregressive Integrated Moving Average (ARIMA) models include an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.

PROCEDURE:

1. We start the procedure of analysis of the given time series data by conducting the *visual exploratory data analysis.* It is essential to analyse the presence of any trend or seasonality or random behaviour in the data in order to choose appropriate time series models.
   To obtain the plot of the data, we first convert it into a time series object.



```
SO2<- ts(SO2, start=2010, frequency = 12)

ts.plot(SO2, main="SO2")
```

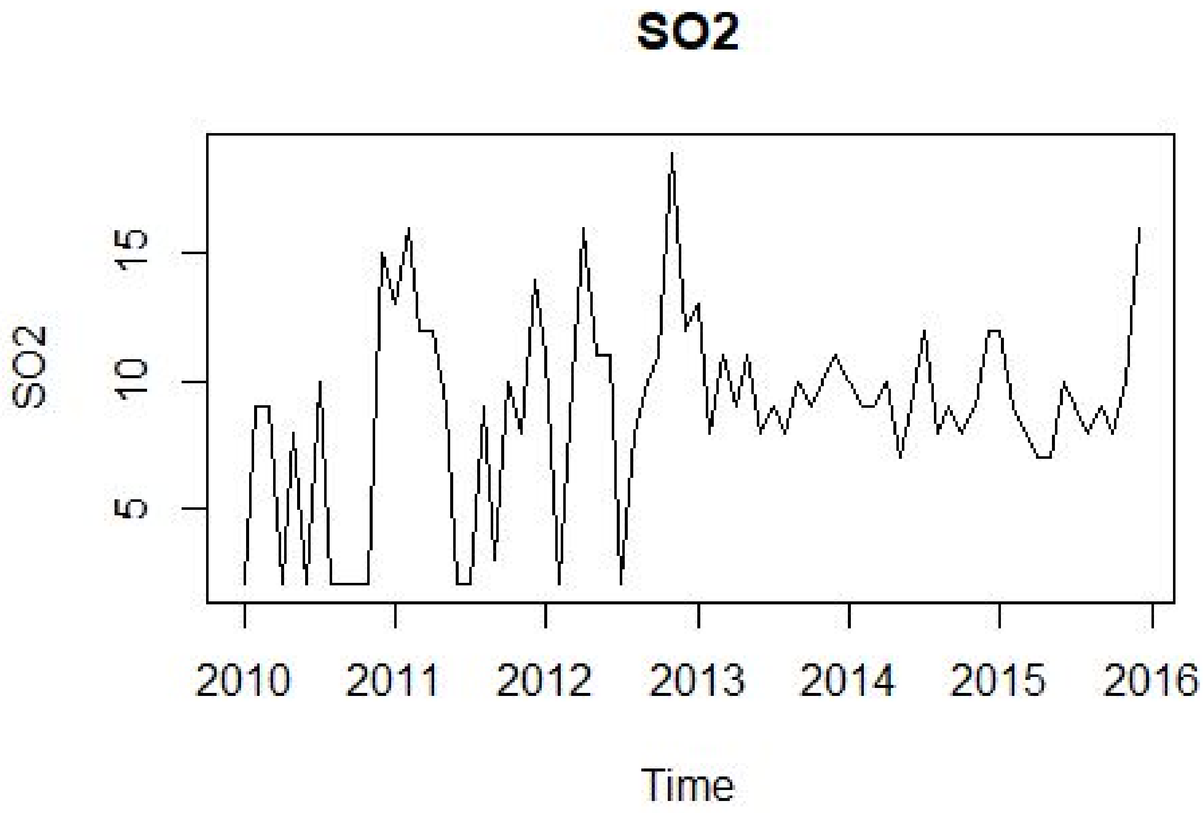
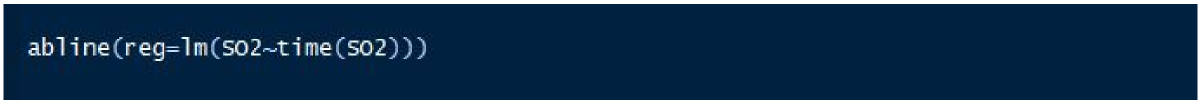


Figure 1: This is the plot which is obtained from the original dataset.

Here we see that the data has irregular patterns and even though there is presence of trend and seasonality, it is not explicitly visible.

We now construct an aggregate line to adjudge the consistency of the mean.



```
abline(reg=lm(SO2~time(SO2)))
```

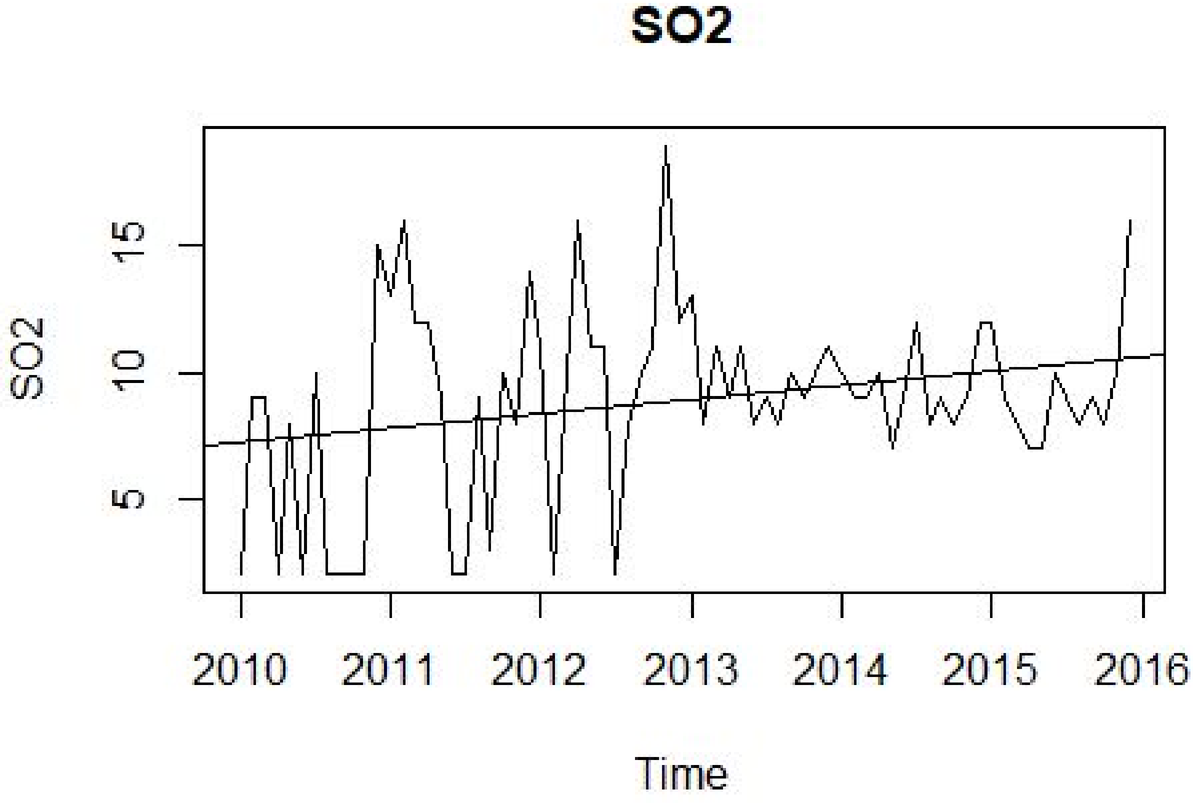
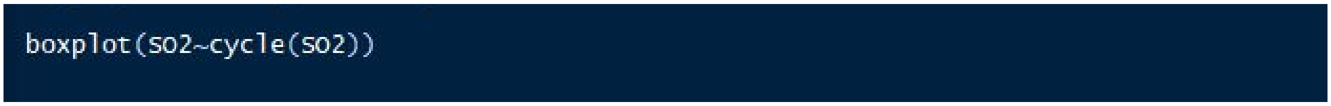
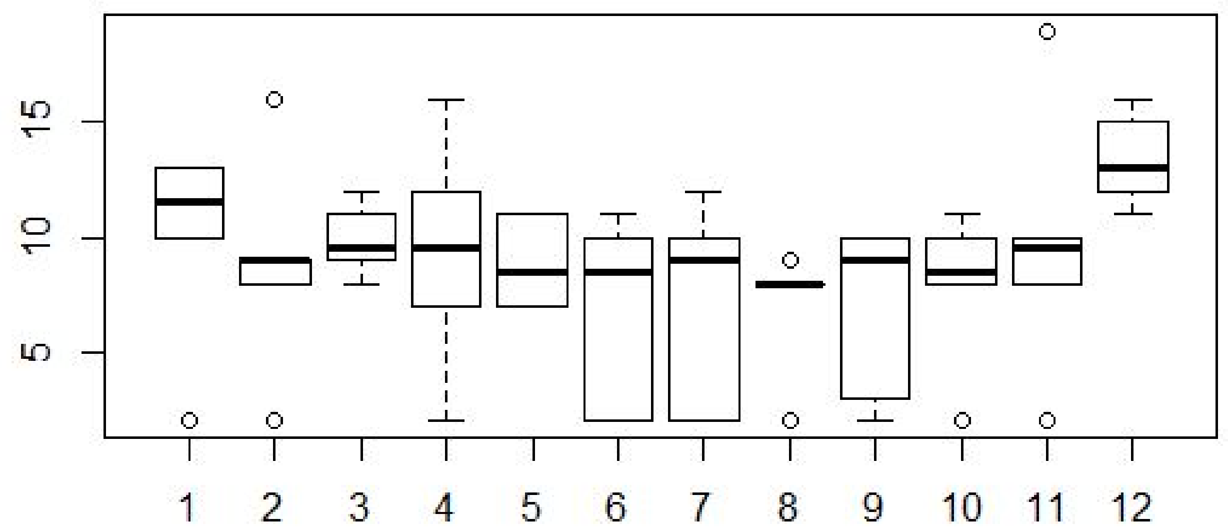


Figure 2: This is the SO2 plot, with the mean line showing the presence of trend in the graph.

Here we see that the mean is increasing with time. This is a clear indication of the non stationarity of data. Now we plot a boxplot of the time series data to get a sense of the seasonal effect of the data.

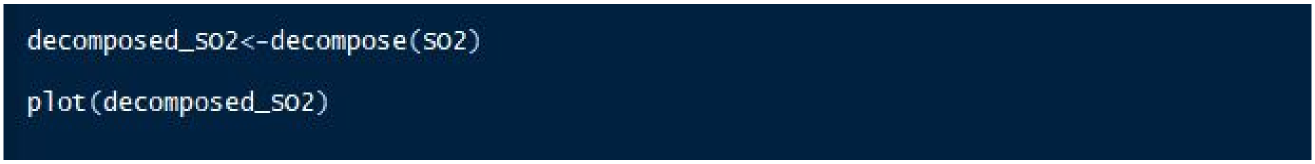


```
boxplot(SO2~cycle(SO2))
```

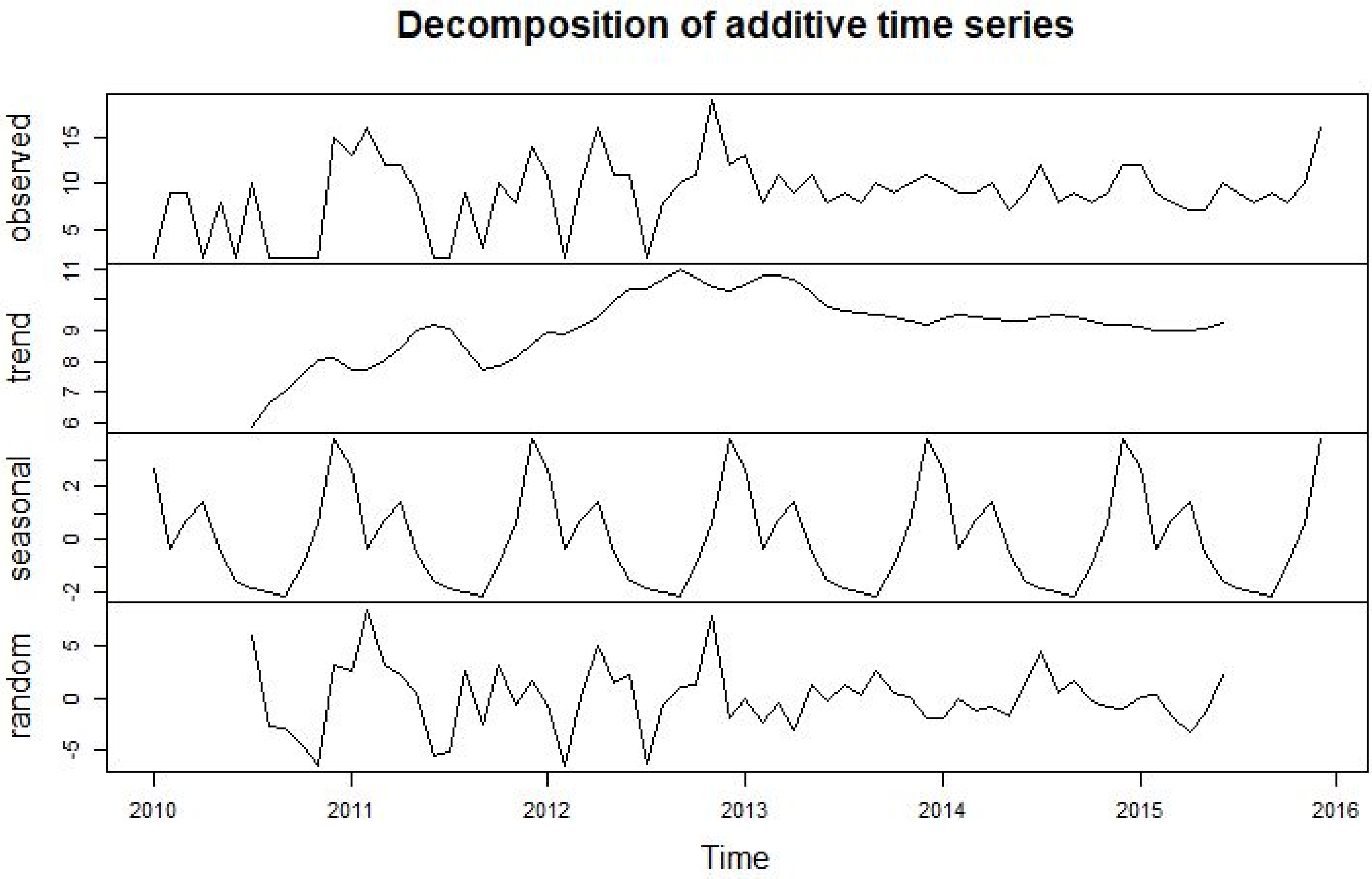


In the given boxplots, we can see that the highest boxplot with the highest median is the data point of the 12th month, and a close second is the data point of the 1st month. We can derive that there exists seasonality towards the very end and the beginning of the year.

Furthermore, we decompose to split the time series data into its components that is trend, seasonality, observed value and random value.

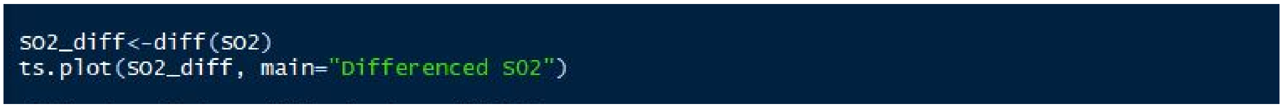


```
decomposed_SO2<-decompose(SO2)

plot(decomposed_SO2)
```

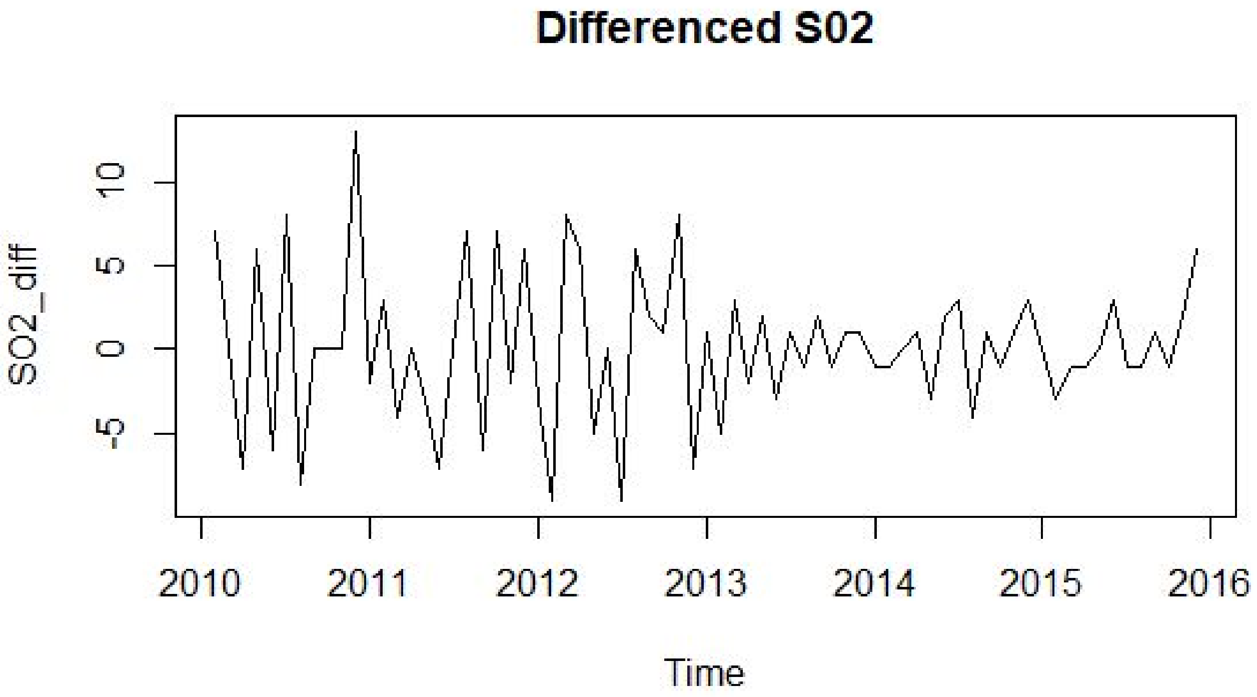


.

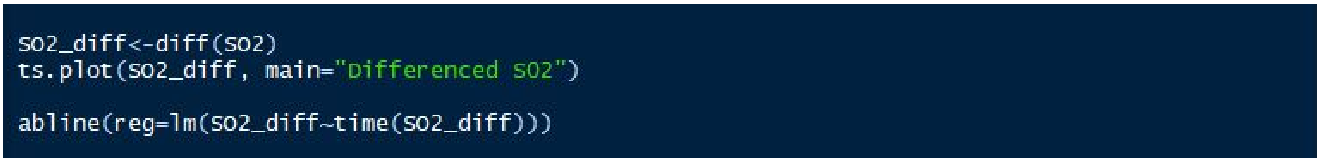
Here we can examine the different components of the given time series data. This is strongly indicative of the data being non-stationary. Hence we now have to make the data stationary. We use two different methods of making the data stationary i.e. 1) By differencing it 2) By individually removing trend and removing seasonality succeedingly
The data is differenced in the following manner.



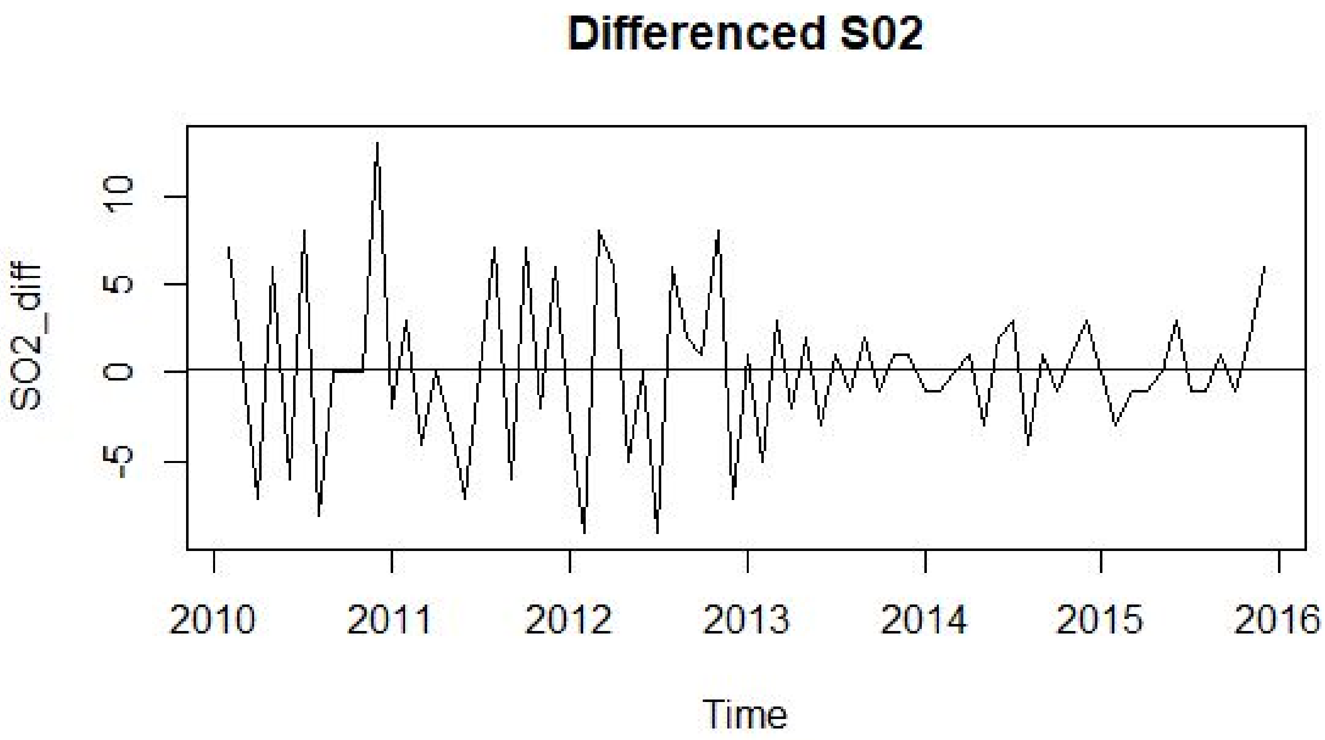
```
SO2_diff<-diff(SO2)
ts.plot(SO2_diff, main="Differenced SO2")
```



We now construct the aggregate line to adjudge the consistency of the mean.

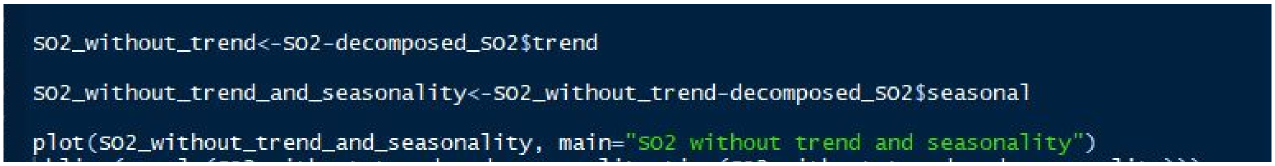


```
SO2_diff<-diff(SO2)
ts.plot(SO2_diff, main="Differenced SO2")

abline(reg=lm(SO2_diff~time(SO2_diff)))
```

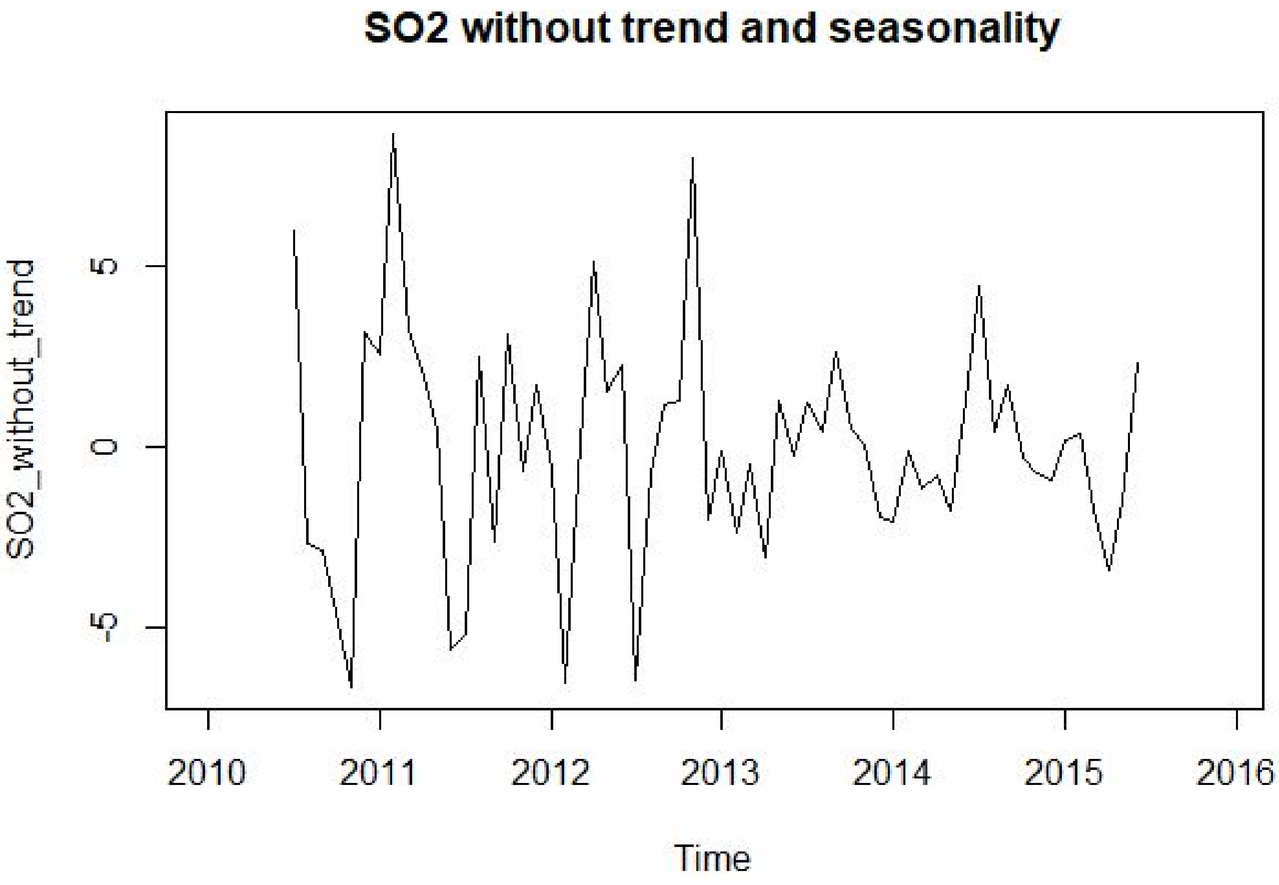


Here we see that the mean is constant ie. mean = 0. Hence we can roughly say that the data is now stationary, according to the visual EDA.

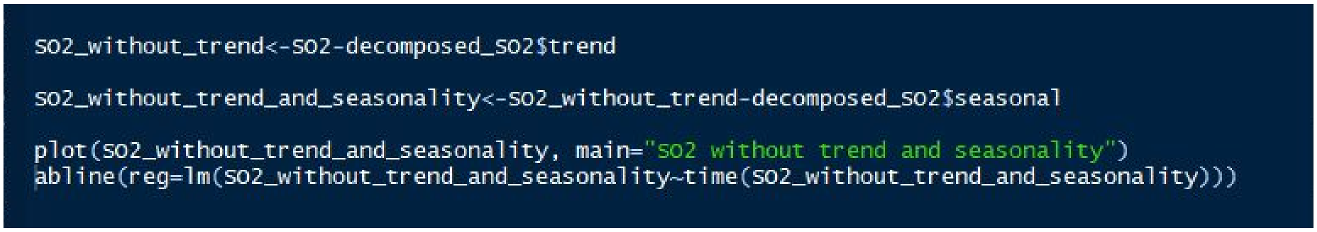
Alternately, we individually remove trend and succeedingly remove seasonality. This is done by removing these components from the decomposition of the time series data that we previously obtained. Then we obtain the plot of the data in which trend and seasonality has been removed.



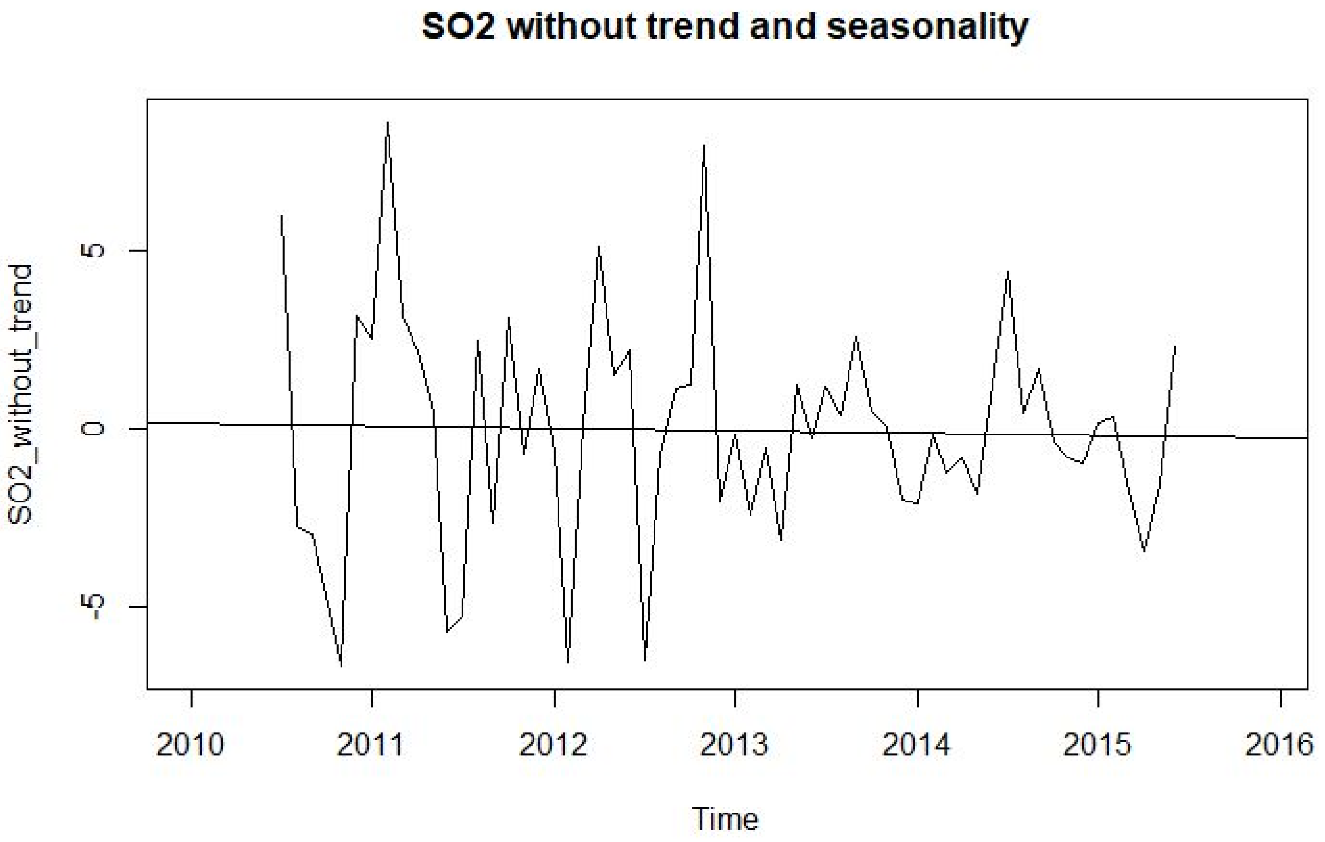
```
SO2_without_trend<-SO2-decomposed_SO2$trend

SO2_without_trend_and_seasonality<-SO2_without_trend-decomposed_SO2$seasonal

plot(SO2_without_trend_and_seasonality, main="SO2 without trend and seasonality")
```



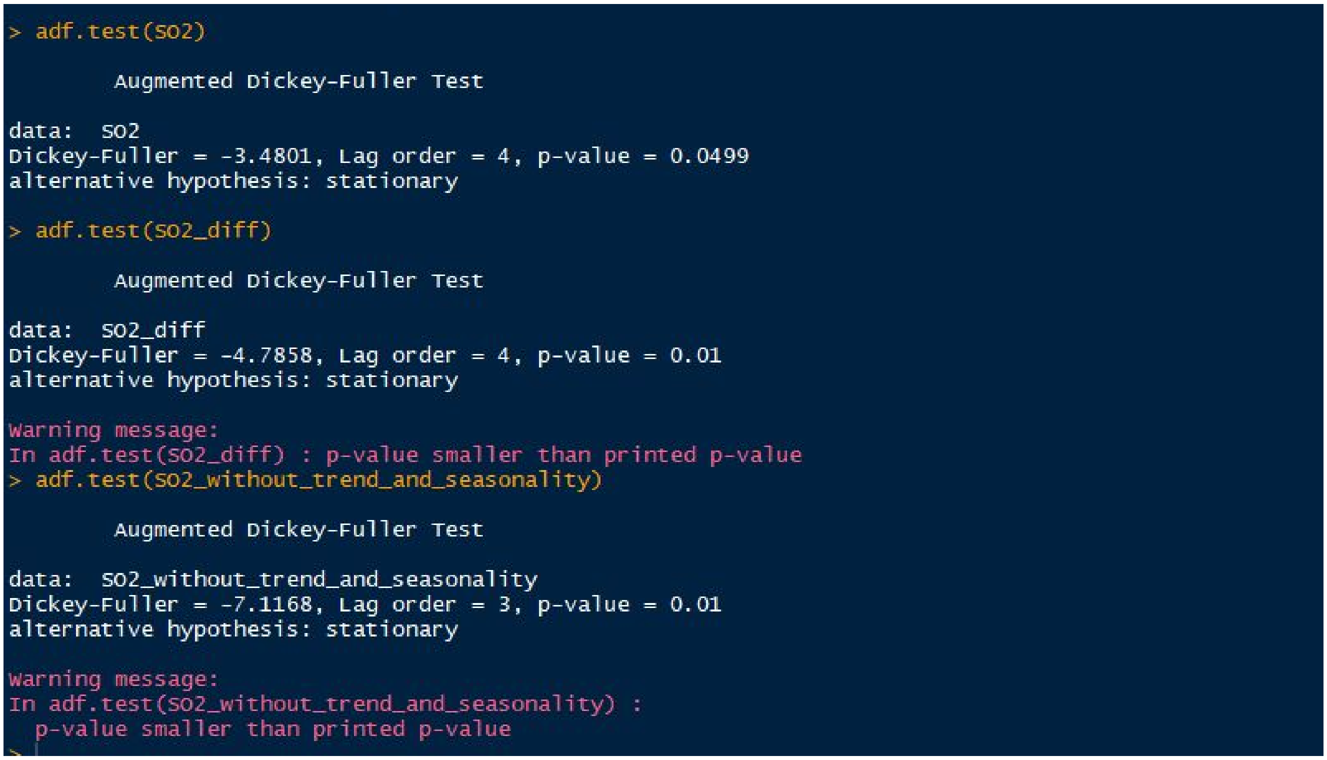
Here again we construct the aggregate line to judge the consistency of mean



```
SO2_without_trend<-SO2-decomposed_SO2$trend

SO2_without_trend_and_seasonality<-SO2_without_trend-decomposed_SO2$seasonal

plot(SO2_without_trend_and_seasonality, main="SO2 without trend and seasonality")
abline(reg=lm(SO2_without_trend_and_seasonality~time(SO2_without_trend_and_seasonality)))
```



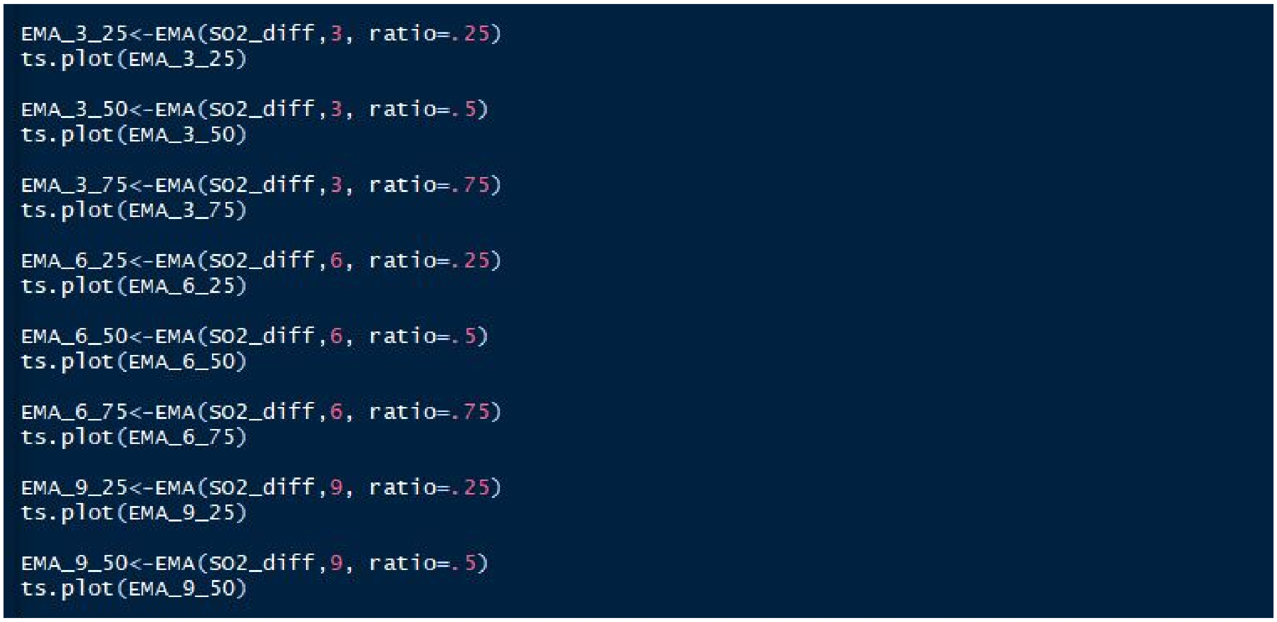
Here again we see that the mean is constant ie. mean = 0. Hence we can roughly say that the data is now stationary, according to the visual EDA, using the alternate detrending method as well.
We now test the stationarity of data using the adf test to technically validate the stationarity of the data.



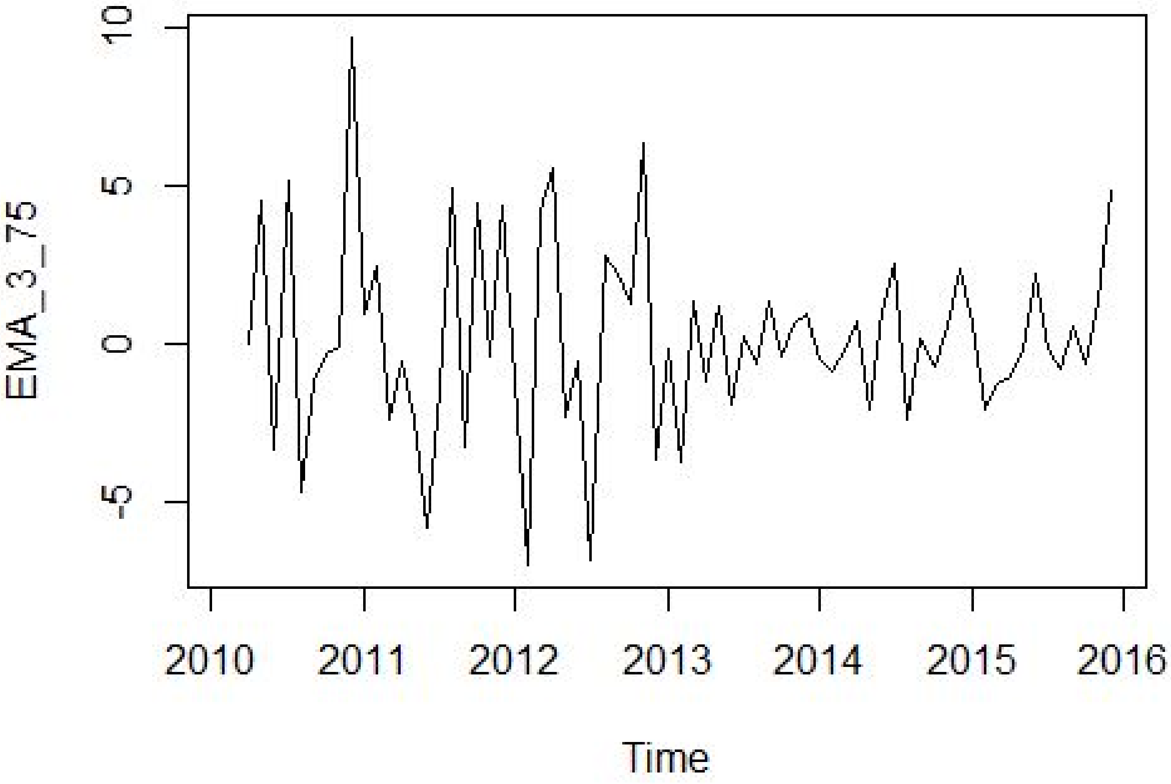
```
> adf.test(SO2)

        Augmented Dickey-Fuller Test

data:  SO2
Dickey-Fuller = -3.4801, Lag order = 4, p-value = 0.0499
alternative hypothesis: stationary

> adf.test(SO2_diff)

        Augmented Dickey-Fuller Test

data:  SO2_diff
Dickey-Fuller = -4.7858, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(SO2_diff) : p-value smaller than printed p-value
> adf.test(SO2_without_trend_and_seasonality)

        Augmented Dickey-Fuller Test

data:  SO2_without_trend_and_seasonality
Dickey-Fuller = -7.1168, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(SO2_without_trend_and_seasonality) :
  p-value smaller than printed p-value
```

For the original data, the p-value is 0.05. Therefore we accept the null hypothesis that the data is non stationary. By using the differencing method, the p-value for the test becomes 0.01. Therefore the data is now stationary. Since the p-value for the adf test of detrending method is also 0.01, the data is stationary using the alternate method as well.
Since the result for both the methods is same, for the sake of convenience, we use the differenced data.

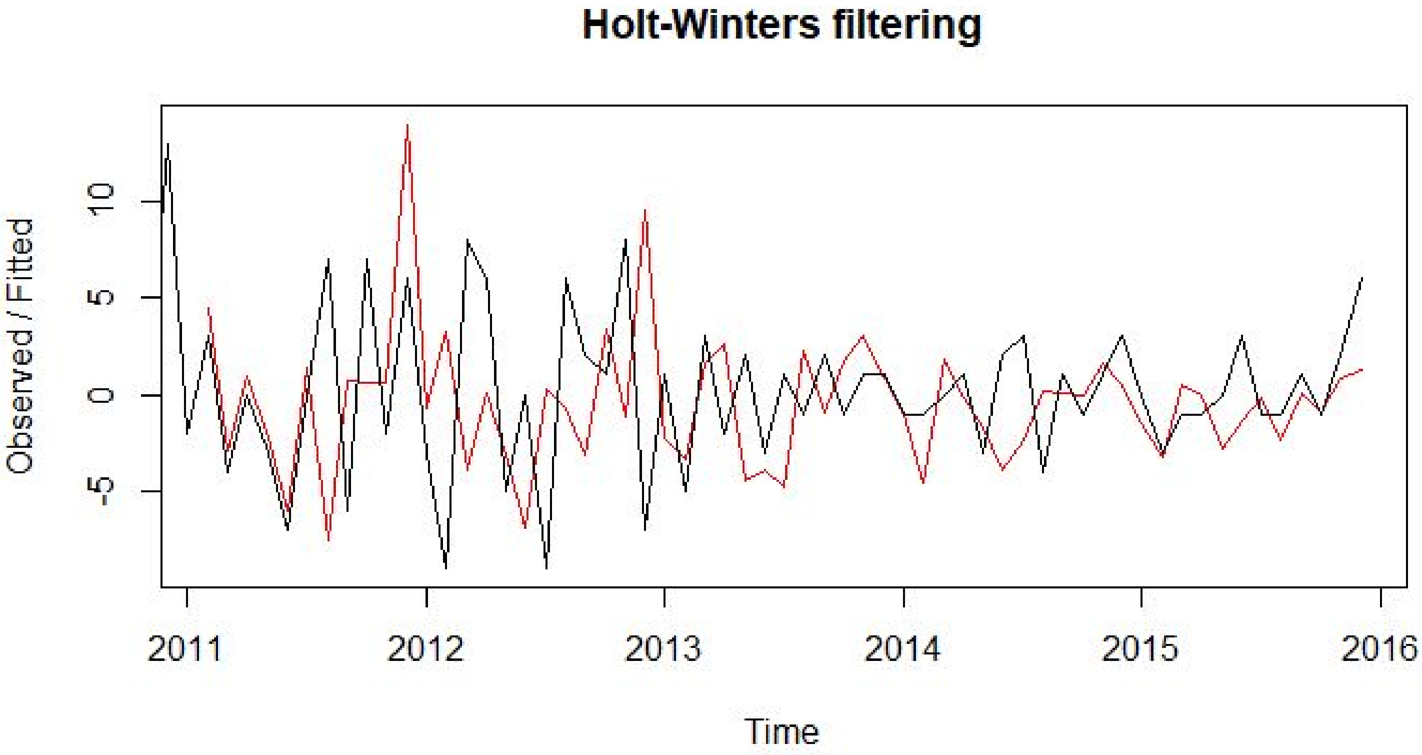
The next step is to conduct smoothing of the data. Using trial and error of various smoothing methods (simple moving average, exponential weighted average smoothing and holt winter's smoothing), we choose the optimal method that smoothens the data without causing loss of values. In Exponentially weighted moving average aka EMA, it exponentially weights SMA. EMAs have faster response to recent value changes than SMAs in a way that SMA caused loss of values. In EMA, the determining factors are the time period and the ratio which depicts the weight given the to the previous values. On running various iterations of EMA, we could adjuge that the most optimal EMA smoothing model is when EMA takes 3 periods and the ratio as 0.75. The ratio is the degree of weighting decrease, a constant smoothing factor between 0 and 1. A higher ratio discounts older observations faster.



```
EMA_3_25<-EMA(SO2_diff,3, ratio=.25)
ts.plot(EMA_3_25)

EMA_3_50<-EMA(SO2_diff,3, ratio=.5)
ts.plot(EMA_3_50)

EMA_3_75<-EMA(SO2_diff,3, ratio=.75)
ts.plot(EMA_3_75)

EMA_6_25<-EMA(SO2_diff,6, ratio=.25)
ts.plot(EMA_6_25)

EMA_6_50<-EMA(SO2_diff,6, ratio=.5)
ts.plot(EMA_6_50)

EMA_6_75<-EMA(SO2_diff,6, ratio=.75)
ts.plot(EMA_6_75)

EMA_9_25<-EMA(SO2_diff,9, ratio=.25)
ts.plot(EMA_9_25)

EMA_9_50<-EMA(SO2_diff,9, ratio=.5)
ts.plot(EMA_9_50)
```

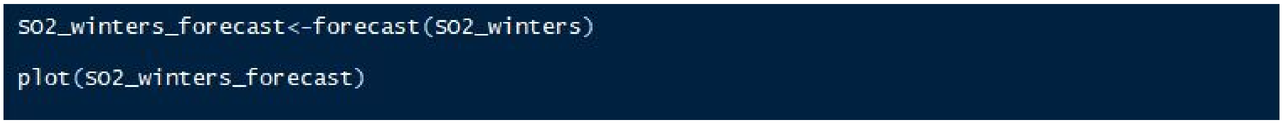


Now we employ Winter’s exponential smoothing. Winter’s method conducts triple exponential smoothing. Here the function assumes three parameters, namely alpha, beta and gamma which represent level, trend and seasonality respectively. In order to conduct single level smoothing, we put beta=False and gamma=False. But here, we want the Winter’s method to smoothen the data on all three levels, hence we use the following code. We get the following smoothened series.

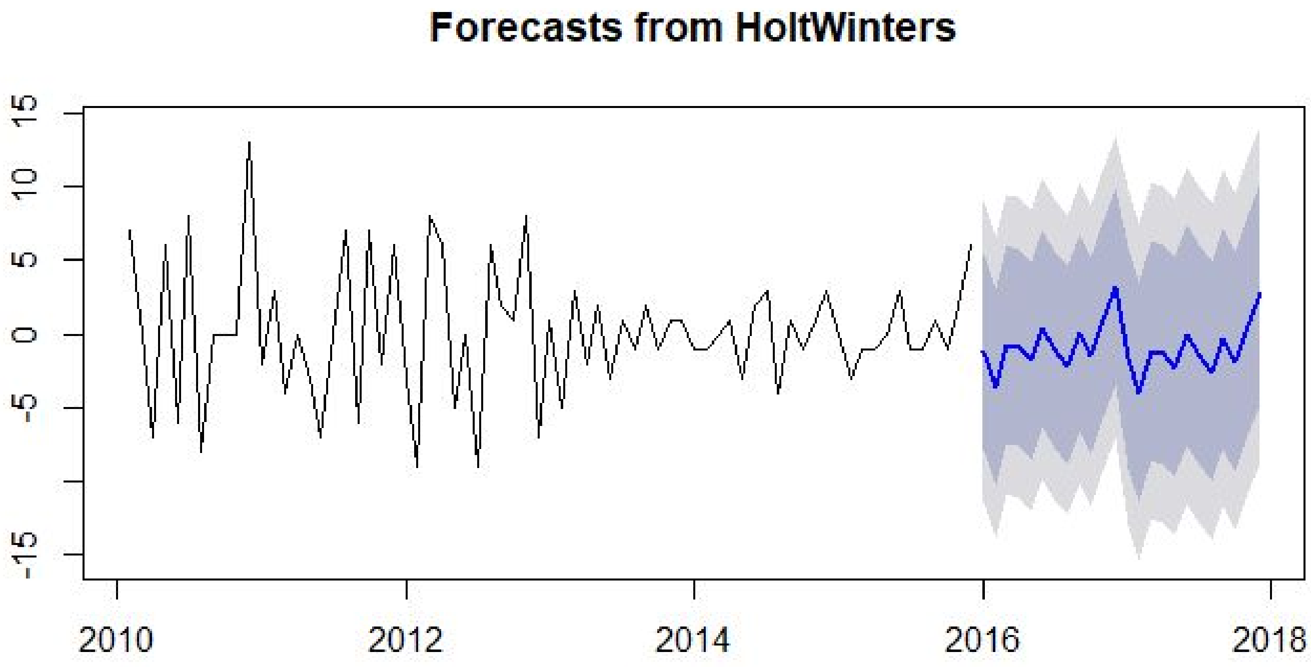


On comparing the three methods of smoothing time series data, namely Simple Moving Average, Exponential Moving Average and Holt Winter’s Exponential smoothing, we see that the most accurate model is the Winter’s method wherein the data gets smoothened without the loss of data.

We proceed with the forecasting on the differenced data which is smoothened using the winter’s method. We forecast the future values using two models: Holt Winter’s and ARIMA model and then we compare the residual values. In general, if forecast errors shows constant variance over a period of time with no fluctuations or no specific pattern, it means that it's a good model to fit.

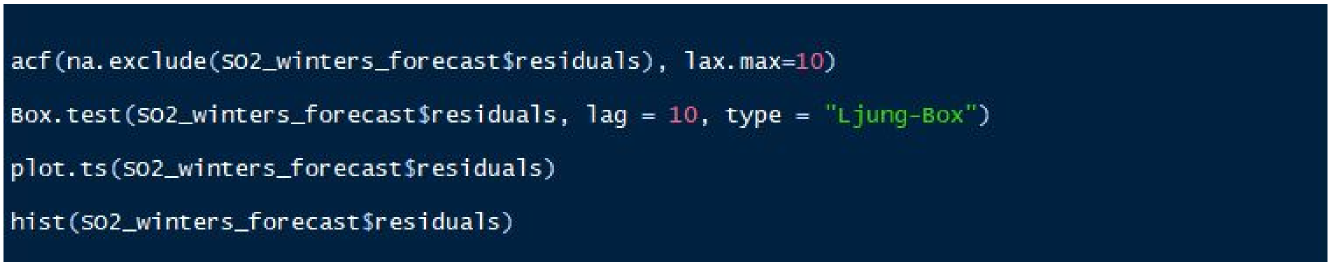


```
SO2_winters_forecast<-forecast(SO2_winters)

plot(SO2_winters_forecast)
```

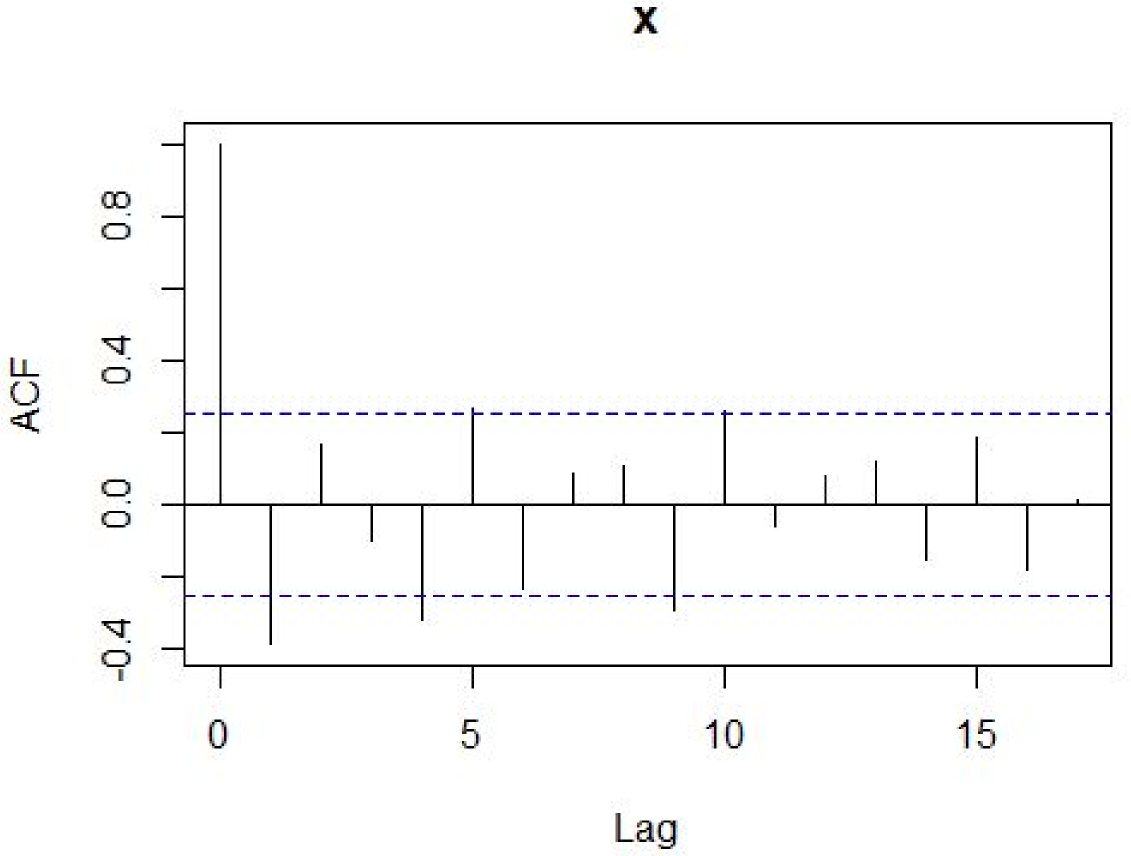


Holt winters method does the prediction by placing the weight to the nearer previous observed value and by connecting the current and the next value. We get the forecast with 80% confidence interval depicted by the dark grey shade and 95% confidence interval shown by the light gray shade surrounding the forecast value.

After obtaining the forecasted values, we attempt to test the accuracy of the predicted values by judging the error terms using acf, Ljung Box test, and a simple error distribution plot.

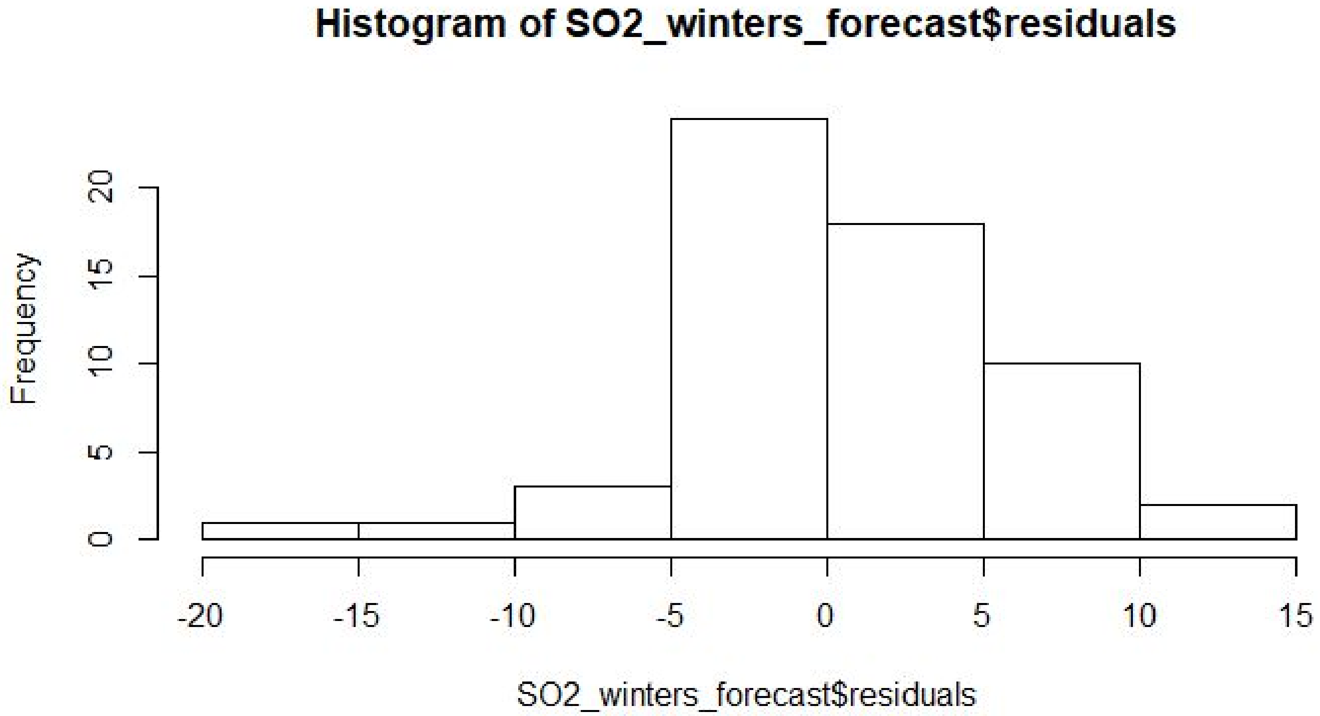


```
acf(na.exclude(SO2_winters_forecast$residuals), lax.max=10)

Box.test(SO2_winters_forecast$residuals, lag = 10, type = "Ljung-Box")

plot.ts(SO2_winters_forecast$residuals)

hist(SO2_winters_forecast$residuals)
```

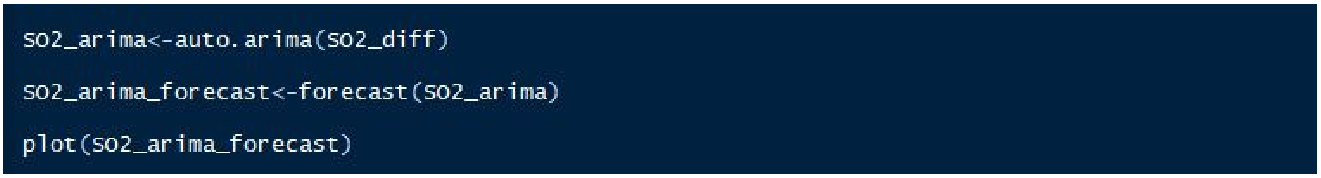


The basic idea behind the ACF function is that it plots the forecast errors that have been plotted on the next predictive values. If our model is a good fit, then the forecast error should not be related to the successive values or the predicted values. If there is a correlation between the forecast errors and the predictive values, then definitely the model is not the best fit and needs transformation or revision in methodology.

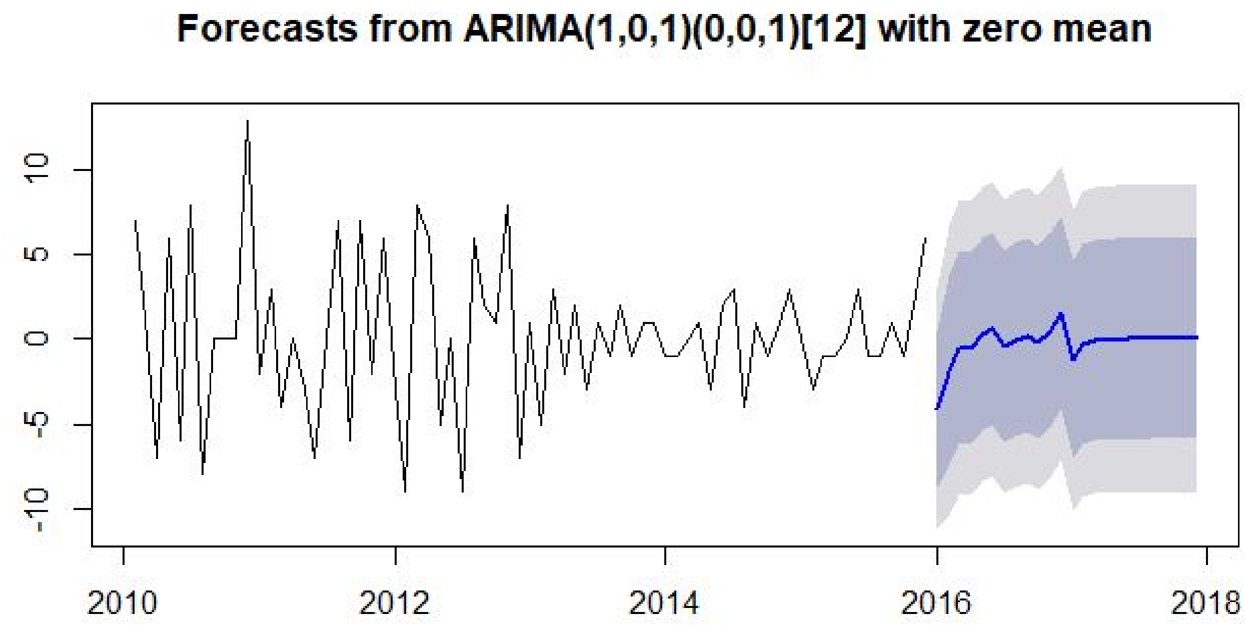
We also plot a simple error distribution plot that is histogram to assess the error values. We can conclude that the residuals do not follow the normal distribution, and there is a negative skew in the distribution. We also see that the modal value of the error distribution is 0.



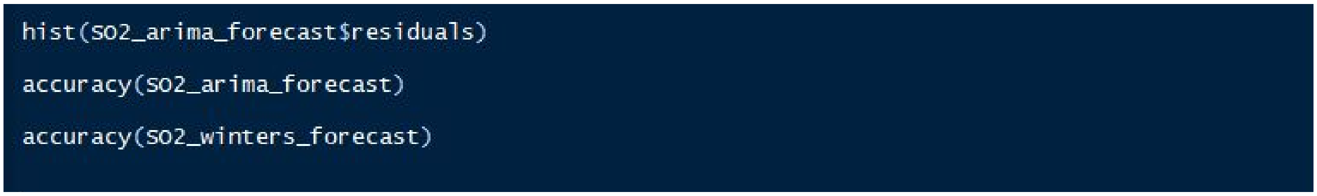
We now forecast the values of the differenced data using the ARIMA model. ARIMA model uses three parameters: p, d, and q depicting acf, differencing order and pacf respectively. We can either enter the optimal parameter values manually deriving from the acf plot and the pacf plot, or we can use auto.arima( ) function which automatically applies the optimal parameter values pertaining to the time series data. The plot hence obtained from the ARIMA modelling is given as follows.



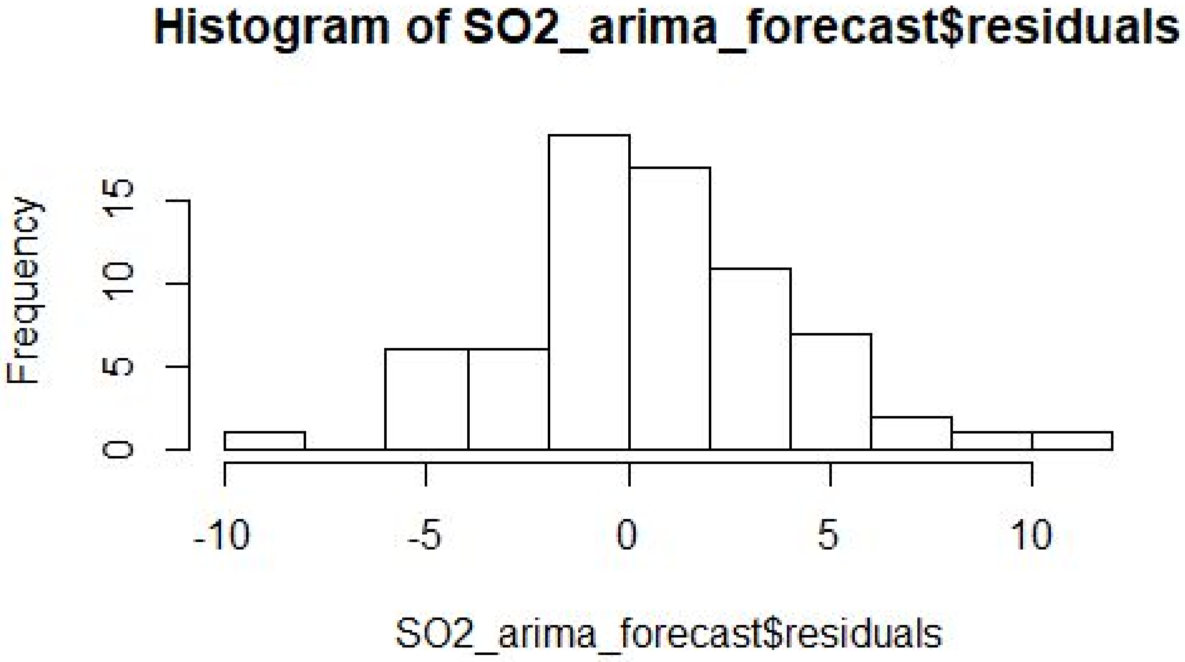
```
SO2_arima<-auto.arima(SO2_diff)

SO2_arima_forecast<-forecast(SO2_arima)

plot(SO2_arima_forecast)
```



Just by method of visual EDA of the two forecasts obtained from the Holt winter's model and the ARIMA model, we find the initial plot more accurate and representative of the preceding patterns. We then obtain a histogram plot of the error distribution and in order to technically analyse the two models, we run the function accuracy( ) which does a residual analysis of the given models.



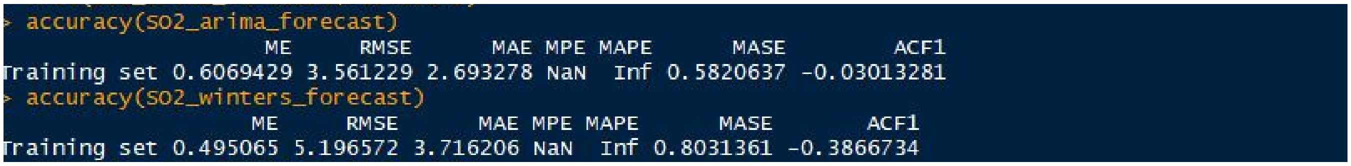
```
hist(SO2_arima_forecast$residuals)

accuracy(SO2_arima_forecast)

accuracy(SO2_winters_forecast)
```



The given error histogram of the ARIMA model appears to follow a normal distribution. Also the modal error value is 0, frequencies of other error values is also quite high. This can be indicative of the fact that this particular model may not be a good fit. However, we now technically compare the two models by conducting residual analysis.

The accuracy( ) gives us the following measures:

- ME: Mean Error
- RMSE: Root Mean Squared Error
- MAE: Mean Absolute Error
- MPE: Mean Percentage Error
- MAPE: Mean Absolute Percentage Error
- MASE: Mean Absolute Scaled Error
- ACF1: Autocorrelation of errors at lag 1.

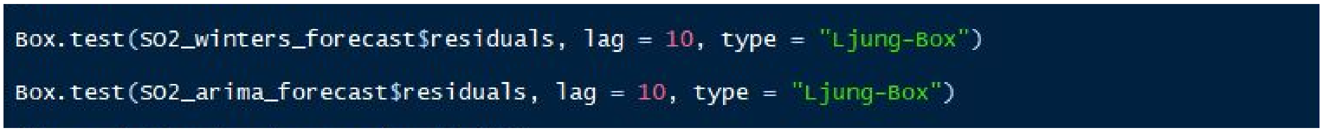


```
> accuracy(SO2_arima_forecast)
                   ME     RMSE      MAE MPE MAPE      MASE        ACF1
Training set 0.6069429 3.561229 2.693278 NaN  Inf 0.5820637 -0.03013281
> accuracy(SO2_winters_forecast)
                  ME     RMSE      MAE MPE MAPE      MASE       ACF1
Training set 0.495065 5.196572 3.716206 NaN  Inf 0.8031361 -0.3866734
```

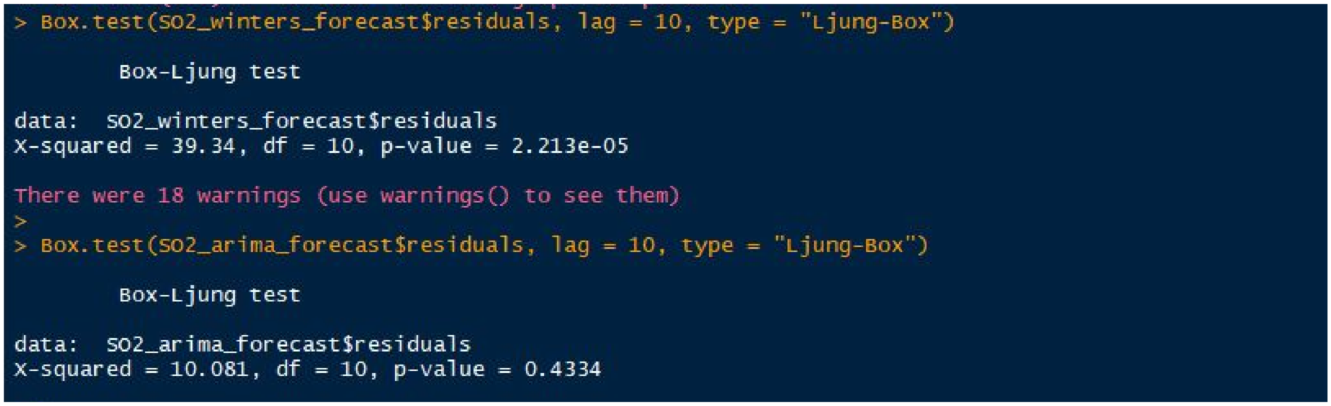
We see that mean error is lower in case of winter’s forecast. Hence, the Winter’s forecast is more accurate.

Given below is the Ljung-Box test conducted on the two models.

The p-value of the winter’s forecast is comparatively very small than the p-value of the arima forecast.



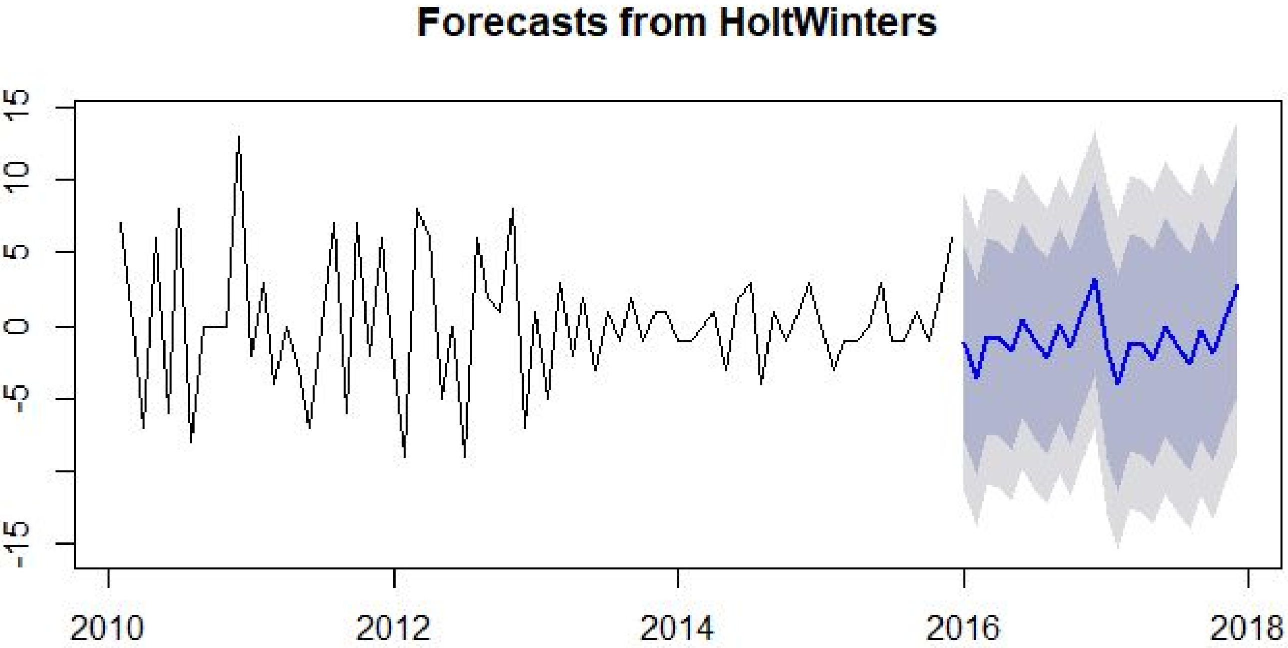
```
Box.test(SO2_winters_forecast$residuals, lag = 10, type = "Ljung-Box")

Box.test(SO2_arima_forecast$residuals, lag = 10, type = "Ljung-Box")
```



```
> Box.test(SO2_winters_forecast$residuals, lag = 10, type = "Ljung-Box")

        Box-Ljung test

data:  SO2_winters_forecast$residuals
X-squared = 39.34, df = 10, p-value = 2.213e-05

There were 18 warnings (use warnings() to see them)
>
> Box.test(SO2_arima_forecast$residuals, lag = 10, type = "Ljung-Box")

        Box-Ljung test

data:  SO2_arima_forecast$residuals
X-squared = 10.081, df = 10, p-value = 0.4334
```

Hence our final conclusion is to use the Winter’s forecast given the higher accuracy of the model.

CONCLUSION:



From the forecasted time series values of SO2 levels in the industrial city of Nagpur, we see that the levels are expected to remain same as they have been since 2014, despite some variability. The reasons can be attributed to the reiteration of environmental concerns particularly pertaining to the air quality in industrial cities of India. There has been a strong emphasis on the Corporate Social Responsibility measures to be undertaken by the companies' production and manufacturing plants by the government and independent non-profit pro environment bodies. And we may conclude that the measures are being mandated just by analysing the status of air quality of the region.