Time series analysis on 5 distinct datasets- Gold price in India, SBI stock price, House Price, Fertility Rate of women in India and Rainfall aggregate data.

Time series Analysis

PRN-2038

Rasika Bose

METHOD:

A time series data is defined as a set of observations or events arranged in a chronological order, i.e, a sequence of observations or events usually ordered in time. The time series is used to draw patterns of changes in statistical data over regular intervals of time.

COMPONENTS OF TIME SERIES:

Trend: An upward or downward tendency of the time series phenomenon is known as secular trend or trend.

Seasonal Variations: In some time series phenomenon, we observe, the periodic fluctuations or regular fluctuations within a year's period during particular time called seasonal variations.

Cyclic Variations: Oscillatory movements of time series with period of oscillation more than a year which causes due to recurring ups and down and uniformly periodic are called Cyclic Variations.

Irregular variations: Irregular variations do not exhibit any definite patterns and there is no regular period time of occurrence. These are erratic, accidental, random, unforeseen and unpredictable events.

MODEL OF TIME SERIES:

Additive Model: $\mathbf{Y_t = T_t + S_t + C_t + I_t}$

Where Y_t is time series, T_t is trend value, S_t is seasonal variations, C_t is cyclic variations and It is irregular variation at time t.

In Additive model, all four components are assumed to be operating independently of one another.

Multiplicative Model: $\mathbf{Y_t = T_t \times S_t \times C_t \times I_t}$

Where Y_t is time series, T_t is trend value, S_t, C_t, It are Seasonal, Cyclic, Irregular indices respectively and expressed as decimal percentage at time t.

In Multiplicative model, all four components are due to different causes but they are not necessarily operating independently of one another.

ACF & PACF PLOTS:

ACF is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values. We plot these values along with the confidence band and we get an ACF plot. In simple

terms, it describes how well the present value of the series is related with its past values. A time series can have components like trend, seasonality, cyclic and residual. ACF considers all these components while finding correlations hence it's a 'complete auto-correlation plot'.

PACF is a partial auto-correlation function. Basically, instead of finding correlations of present with lags like ACF, it finds correlation of the residuals with the next lag value hence 'partial' and not 'complete' as we remove already found variations before we find the next correlation. So, if there is any hidden information in the residual which can be modelled by the next lag, we might get a good correlation and we will keep that next lag as a feature while modelling.

HOLT WINTERS MODEL:

Holt-Winters forecasting is a way to model and predict the behaviour of a sequence of values over time. Holt-Winters is one of the most popular forecasting techniques for time series. Holt-Winters is a way to model three aspects of the time series: level, trend and seasonality. Holt-Winters uses exponential smoothing to obtain lots of values from the past and use them to predict values for the present and future. The three aspects of the time series behaviour such as level, trend, and seasonality are expressed as three types of exponential smoothing, so Holt-Winters is called triple exponential smoothing.

Holt Winters Multiplicative model algorithm:

for $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$ and $0 \leq \gamma \leq 1$ where α, β, γ are smoothing parameters.

Level $L_t = \alpha\ (Y_t / S_{t-m}) + (1-\alpha)\ (L_{t-1} + T_{t-1})$

Trend $T_t = \beta\ (L_t - L_{t-1}) + (1-\beta)\ T_{t-1}$

Seasonal $S_t = \gamma\ (Y_t / L_t) + (1-\gamma)\ S_{t-m}$

Forecast $F_t = (L_t + T_t)\ S_{t+1-m}$

This is done until no more observation Y_t are available and the subsequent forecasts are:

$F_{t+k} = (L_t + k\ T_t)\ S_{t+k-s}$

Value of L_0 can be assumed as the first observation of Y_t or the average of the observations of Y_t & value of T_0 is taken as the slope.

MEASURE OF FORECAST ACCURACY:

Mean Absolute Forecast Error (MAFE): $\frac{\sum_{t=1}^{m} |y_t - \hat{y}_t|}{m}$

Mean Absolute Percentage Error (MAPE): $\frac{\sum_{t=1}^{m} [\frac{|y_t - \hat{y}_t|}{y_t}]}{m} \times 100$

Root Mean Square Error (RMSE): $\sqrt{\frac{\sum_{t=1}^{m} (y_t - \hat{y}_t)^2}{m}}$

Strict Stationarity:

A series $\{X_t\}$ is said to be strictly stationary, if joint distribution of $(X_{t1}, X_{t2}, ..., X_{tn})$ is same as that of $(X_{t1+h}, X_{t2+h}, ..., X_{tn+h})$ for any (t1, t2, ..., tn) and h > 0. Stationarity of a time series implies that, for any n, the joint distribution of n consecutive random variables is the same, no matter where we start.

Weak Stationarity (Second order stationarity):

$E(X_t) = \mu < \infty$ (a constant free from t), $Var(X_t) < \infty$ (free from t)

& $Cov(X_t , X_s)$ should be a function of the time difference (or lag) $|(t - s)|$

SEASONAL AUTO REGRESSIVE INTEGRATED MOVING AVERAGE (SARIMA) MODEL

Auto regressive Integrated Moving Average, or ARIMA, is one of the most widely used forecasting methods for univariate time series data forecasting. Although the method can handle data with a trend, it does not support time series with a seasonal component. An extension to ARIMA that supports the direct modelling of the seasonal component of the series is called SARIMA. **SARIMA(p, d, q) × (P, D, Q)s** is defined by

$$\Phi(B^s)(1 - B^s)^D \varphi(B)(1 - B)^d X_t = \Theta(B^s)\theta(B)Z_t , Z_t \sim WN(0, \sigma^2)$$

Where d = non-seasonal order of differencing

D = seasonal order of differencing

p = non-seasonal AR order

P = seasonal AR order

q = non-seasonal MA order

ARCH and GARCH effect

An ARCH (autoregressive conditionally heteroscedastic) model is a model for the variance of a time series. ARCH models are used to describe a changing, possibly volatile variance. Although an ARCH model could possibly be used to describe a gradually increasing variance over time, most often it is used in situations in which there may be short periods of increased variation.

An ARCH(m) process is one for which the variance at time t is conditional on observations at the previous *m* times, and the relationship is

$$\mathbf{Var}(y_t|y_{t-1},\ldots,y_{t-m})=\sigma_t^2=\alpha_0+\alpha_1 y^2_{t-1}+\cdots+\alpha_m y^2_{t-m}.$$

With certain constraints imposed on the coefficients, the y_t series squared will theoretically be AR(m).

A GARCH (generalized autoregressive conditionally heteroscedastic) model uses values of the past squared observations and past variances to model the variance at time t. As an example, a GARCH(1,1) is

$$\sigma_t^2=\alpha_0+\alpha_1 y^2_{t-1}+\beta_1\sigma^2_{t-1}$$

In the GARCH notation, the first subscript refers to the order of the y^2 terms on the right side, and the second subscript refers to the order of the σ^2 terms.

1. Gold Price in India

Data link: https://www.gold.org/goldhub/data/gold-prices

Price discovery is crucial for any market. Gold not only has a spot price, but it also has the LBMA Gold Price, as well as several regional prices. The LBMA Gold Price is used as an important benchmark throughout the gold market, while the other regional gold prices are important to local markets.

This data set provides the gold price over a range of timeframes (daily, weekly, monthly, annually) going back to 1978, and in the major trading, producer, and consumer currencies.

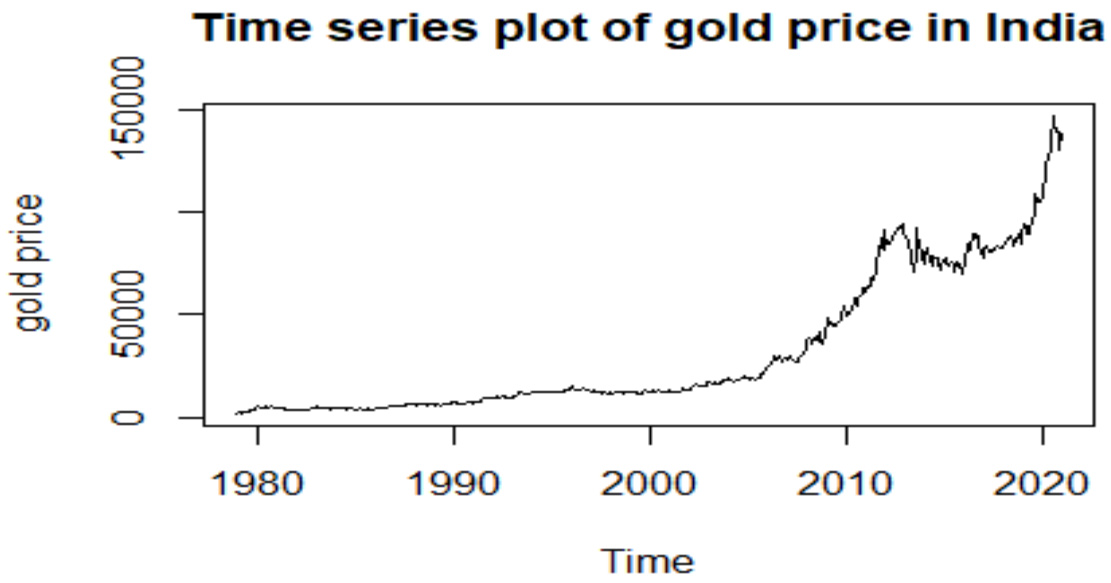
Report:

The data consists of end of month gold prices in India from January 1979 to December 2021. There are in total 516 points with no missing values. For the time series analysis I have taken 504 points from January 1979 to December 2020 as the training set for the model. The remaining 12 points from January 2021 to December 2021 is our test set.

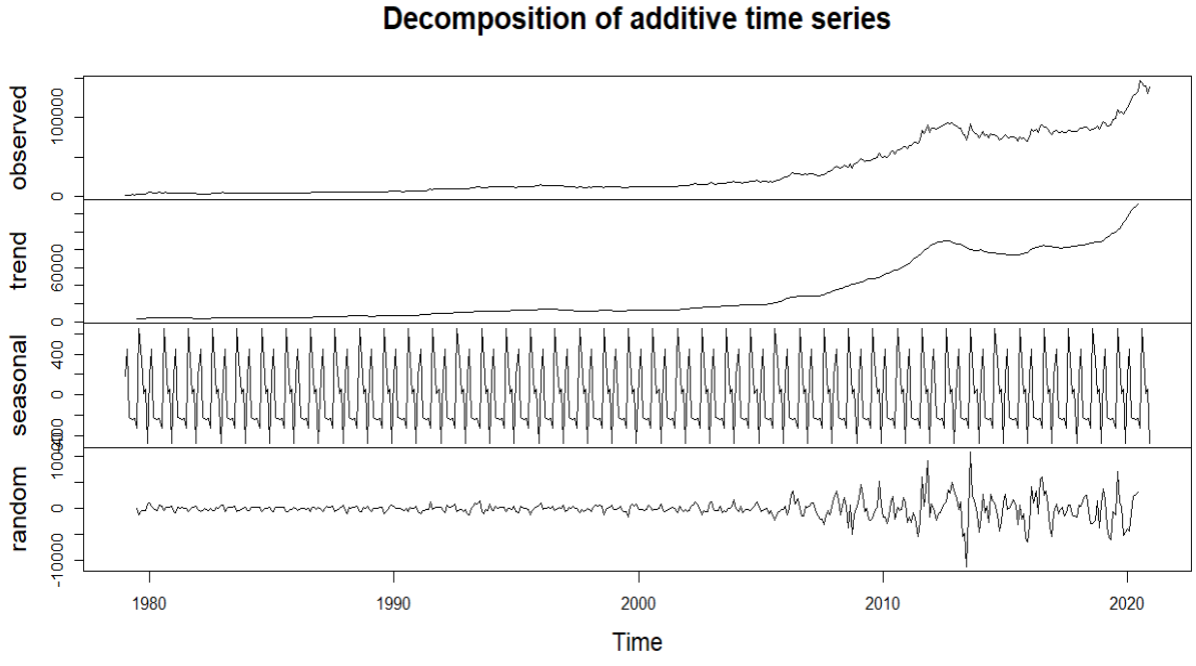
Descriptive Statistics of the data:

Mean	31752
1st Quartile	6216
Median	12868
3rd Quartile	56434
Min	1841
Max	146999

After plotting the data I have observed that the data has an upward trend.

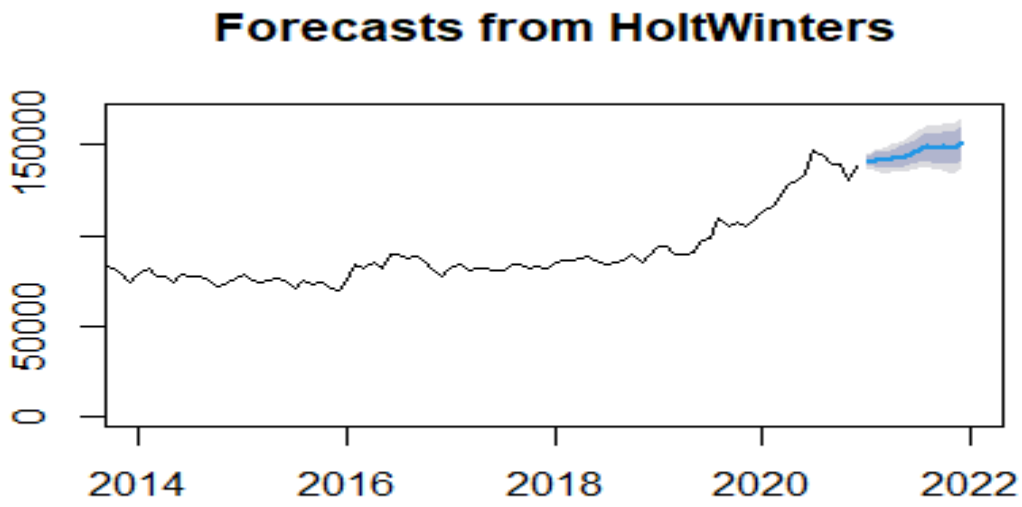


After decomposing the data I observe the presence of seasonality in the data.



Since the data has both trend and seasonality I will perform **Holt Winter's Exponential Smoothing** on the data. After applying Holt Winter's Smoothing I got the parameter values $\alpha = 0.8135872$, $\beta = 0.02813567$ and $\gamma = 0.6445282$.

I made a 12 step ahead forecast using the Holt Winter's model.



The RMSE of the test data for Holt Winter's forecast is 13754.68.

A better model can be obtained whose forecast should give a lower test RMSE.
Since the data has seasonality, I will try to fit a SARIMA$(p,d,q)(P,D,Q)_s$ where the parameters will be decided using the ACF and PACF plots.
Primary step of fitting an ARIMA/SARIMA model is checking if the data is stationary.
I performed ADF test, PP test and KPSS test for stationarity checking.

- **ADF test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: > 0.99
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **PP test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: > 0.99
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **KPSS test:**
  H_0: The time series is stationary

H_1: The times series is non-stationary
Obtained p-value: < 0.01
Thus, we reject the null hypothesis and conclude data is non-stationary.

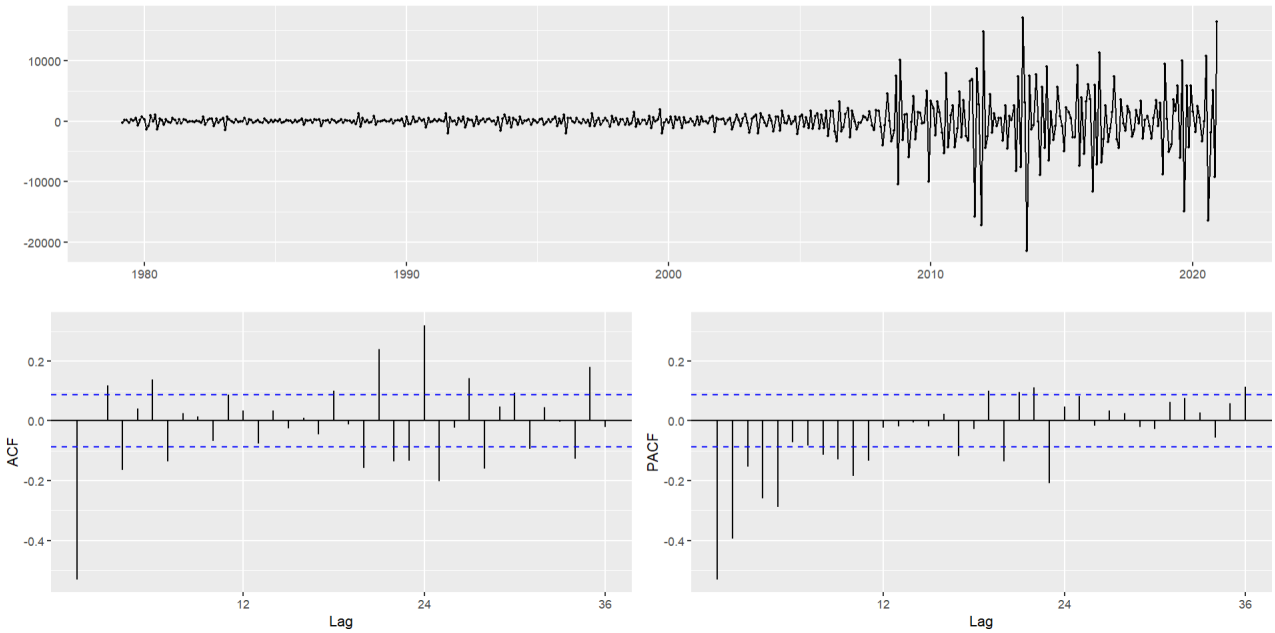
All the three tests indicated that the data is **non stationary**.

I applied first order differencing i.e. $Y_t^{(1)} = Y_t - Y_{t-1}$ and performed all the three tests where only the KPSS test indicated that the data is non-stationary.
After applying the second order differencing $Y_t^{(2)} = Y_t^{(1)} - Y_{t-1}^{(1)}$ and performing all the tests of stationarity I observed that the second order differenced data is indicated stationary by all the tests. Hence, we get the value of *d* as 2 for our SARIMA model.

Finding the required seasonal differencing (*D*) manually for the SARIMA model is hard unlike finding (*d*). I have used the *'nsdiffs()'* function to determine the number of seasonal differences required. It estimates the number of seasonal differences necessary to make the time series stationary. *R* provided the required seasonal differencing as zero.

So, I get the differencing schemes as $d = 2$ and $D = 0$.

Now I plot the stationary time series and its ACF and PACF plots.



ACF and PACF of stationary data

From the ACF plot I can see a significant spike at lag 1 which suggests an *MA(1)* model. I also see a significant spike at lag 24 in the ACF plot which suggests a seasonal *MA(2)* process. The PACF plot shows an exponential decay more or less both non-seasonally and seasonally so no AR process is present in the model.
So, by looking at the ACF and PACF plots we can guess that the appropriate model is SARIMA$(0,2,1)(0,0,2)_{12}$. The AIC of the manually chosen model was 9146.119.
Next I used *'auto.arima()'* to find out the model suggested by *R*. The software provided the appropriate model as SARIMA$(0,2,2)(0,0,2)_{12}$ with AIC 9142.364.

The SARIMA model is given by,

$$\Phi(B^s)(1-B^s)^D\phi(B)(1-B)^d X_t = \Theta(B^s)\Theta(B)Z_t$$

where *p=0, d=2, q=2 , P=0, D=0* and *Q=2*
Thus the model is,

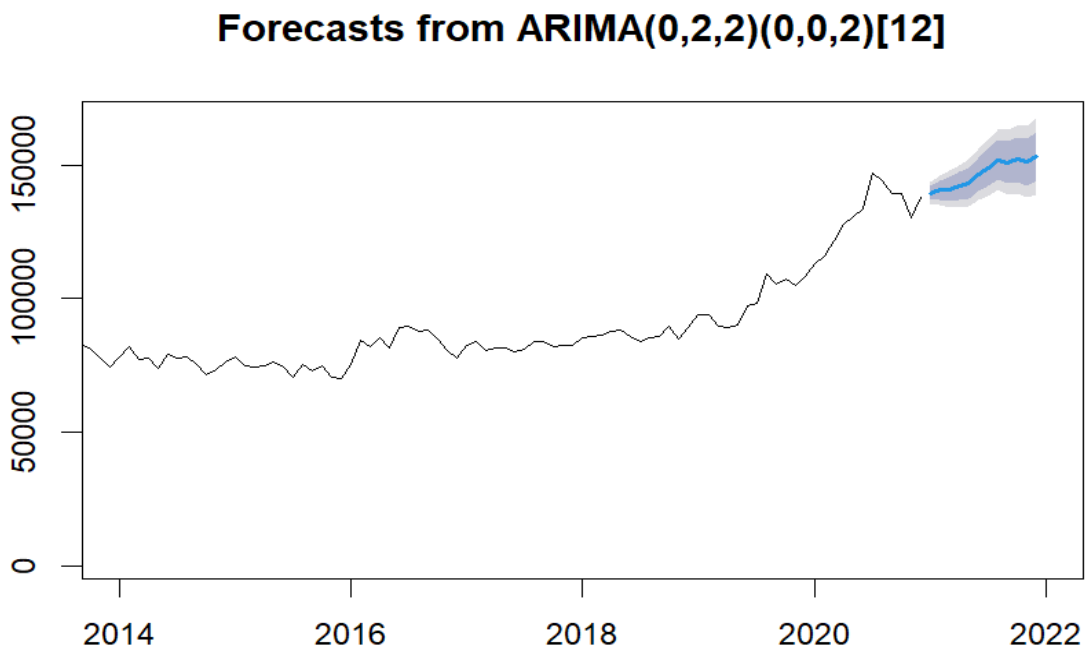
$$(1-B)^2 X_t = \Theta(B^{12})\Theta(B)Z_t$$

Next I have tested the significance of the model. I have done that by checking the p-values of the coefficients.

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ma1  -1.098053   0.045674 -24.0413 < 2.2e-16 ***
ma2   0.110261   0.045904   2.4020   0.01631 *
sma1  0.114534   0.050276   2.2781   0.02272 *
sma2  0.275782   0.046009   5.9941 2.047e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the coefficients have p-values less than 0.05 which means the model coefficients are significant.

Using the above model suggested by *R* I have forecasted 12 steps ahead i.e. forecast for the prices in the months of year 2021. The test RMSE obtained is 15493.52.

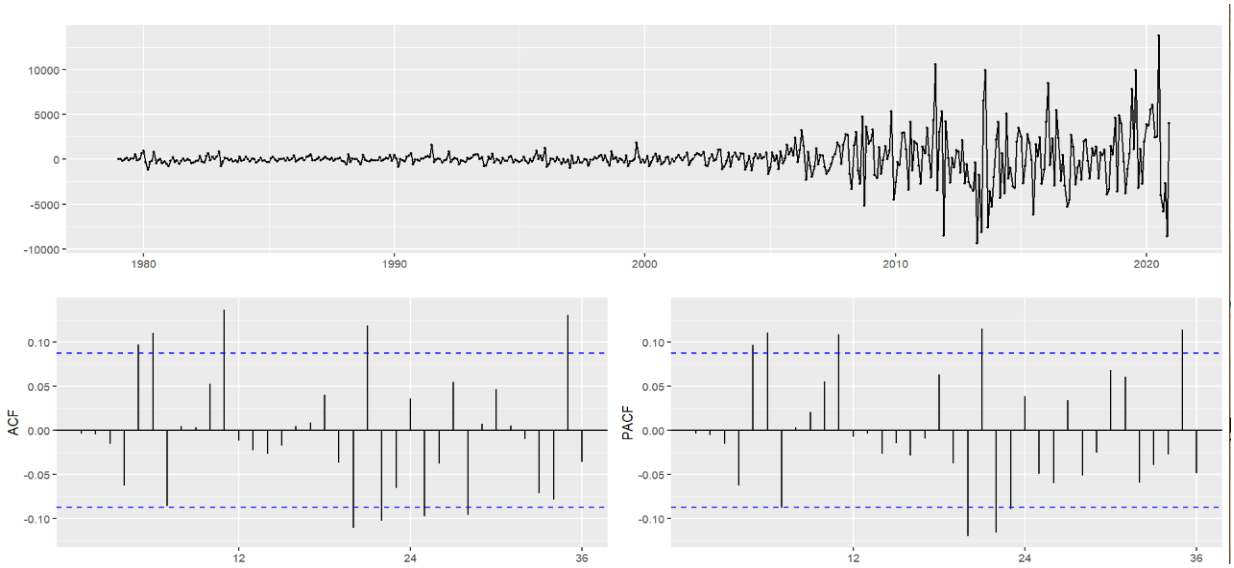


From the comparison RMSEs I can say that Holt Winter's Exponential Smoothing method has provided better forecasts than the SARIMA model.

Residual Analysis

Next I have analysed the residuals of the models.
First I have checked for the stationarity of the residuals of the model using the KPSS test. The p-value obtained was > 0.1 which indicated that the residuals are stationary.

Next I have plotted the ACF and PACF plots of the model residuals.



ACF and PACF of model residuals

The residuals of the model should follow i.i.d sequence with mean zero and constant variance. To check the mean of the residuals equal to zero I have used the '*wilcox.test()*'.

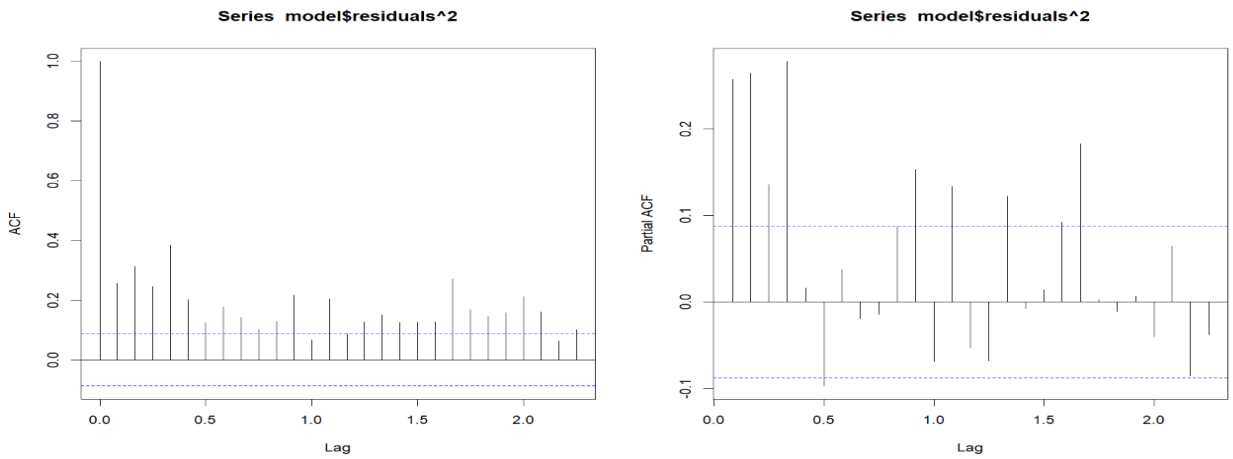
```
        Wilcoxon signed rank test with continuity correction

data:  model$residuals
V = 65733, p-value = 0.5204
alternative hypothesis: true location is not equal to 0
```

From the above test I see that the residuals have mean equal to zero.

Next, using the Bartlett test I can conclude that the variance of the residuals is not constant.

```
        Bartlett test of homogeneity of variances

data:  model$residuals and a
Bartlett's K-squared = 909.39, df = 41, p-value < 2.2e-16
```

Next, I observe the ACF and PACF plots of the squared residuals of the model.



ACF and PACF plots of squared residuals

Observing the above plots I can infer that there is ARCH effect present in the model.

By applying the Ljung-Box test on the squared residuals I will confirm my observations.

```
        Box-Ljung test

data:  model$residuals^2
X-squared = 285.19, df = 12, p-value < 2.2e-16
```

The Ljung-Box test confirms my observation and I conclude that there is ARCH effect present in the model.

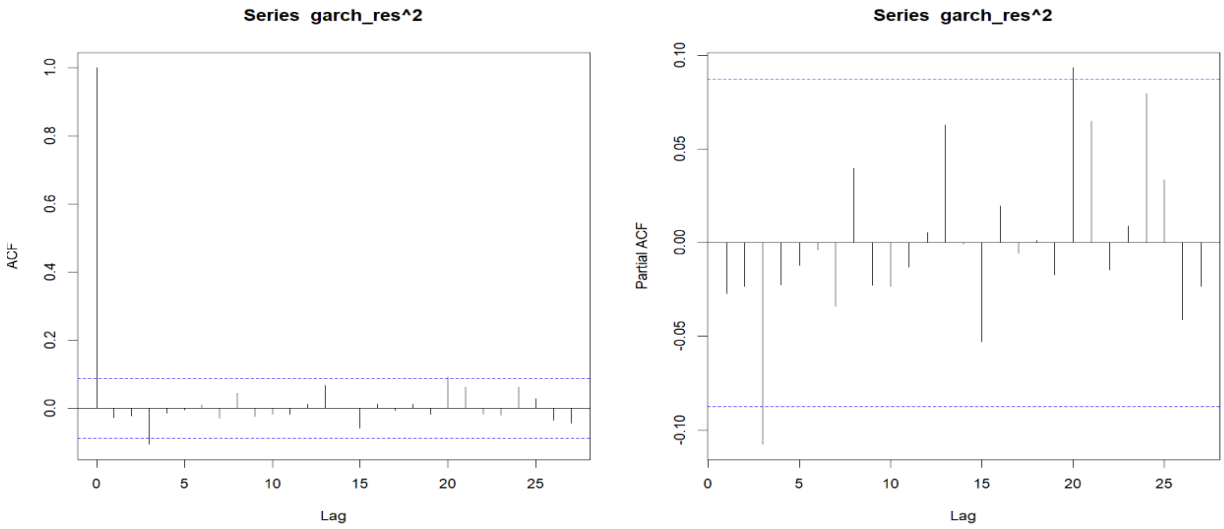
From the PACF plot I can see a significant spike till lag 4. I will start with the ARCH(4) model till I get a model with all significant coefficients and lowest AIC. After trying various ARCH models I conclude that ARCH(4) is the best model.

Now I will fit the ARCH(4) model.

```
Title:
 GARCH Modelling

Call:
 garchFit(formula = ~1 + garch(4, 0), data = model$residuals,
    trace = F)
```

```
Conditional Distribution:
 norm

Coefficient(s):
         mu        omega       alpha1       alpha2       alpha3       alpha4
  -32.28319  44961.15106      0.32844      0.15359      0.53710      0.34968

Std. Errors:
 based on Hessian

Error Analysis:
         Estimate  Std. Error  t value Pr(>|t|)
mu     -3.228e+01   1.937e+01   -1.667  0.09555 .
omega   4.496e+04   1.112e+04    4.043 5.28e-05 ***
alpha1  3.284e-01   8.299e-02    3.958 7.57e-05 ***
alpha2  1.536e-01   4.782e-02    3.212  0.00132 **
alpha3  5.371e-01   8.858e-02    6.063 1.33e-09 ***
alpha4  3.497e-01   8.478e-02    4.125 3.71e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log Likelihood:
 -4151.692    normalized:  -8.237484


Standardised Residuals Tests:
                                Statistic p-Value
 Jarque-Bera Test   R    Chi^2  53.5072   2.404743e-12
 Shapiro-Wilk Test  R    W      0.9836355 1.904116e-05
 Ljung-Box Test     R    Q(10)  6.614144  0.7613005
 Ljung-Box Test     R    Q(15)  13.86374  0.5358876
 Ljung-Box Test     R    Q(20)  16.12115  0.7090805
 Ljung-Box Test     R^2  Q(10)  8.397015  0.5901172
 Ljung-Box Test     R^2  Q(15)  12.71918  0.623978
 Ljung-Box Test     R^2  Q(20)  17.43656  0.624471
 LM Arch Test       R    TR^2   9.19691   0.686026

Information Criterion Statistics:
     AIC      BIC      SIC     HQIC
16.49878 16.54905 16.49850 16.51850
```

The model coefficients are significant. The Ljung-Box test in the Standardised Residuals test shows that there is no ARCH effect present now in the model.



ACF and PACF plots of residuals of ARCH model

From the ACF and PACF we can observe that there is no ARCH effect present in the model. This implies that the ARCH(4) model fitted above on the residuals has captured maximum volatility.

2. SBI stock price

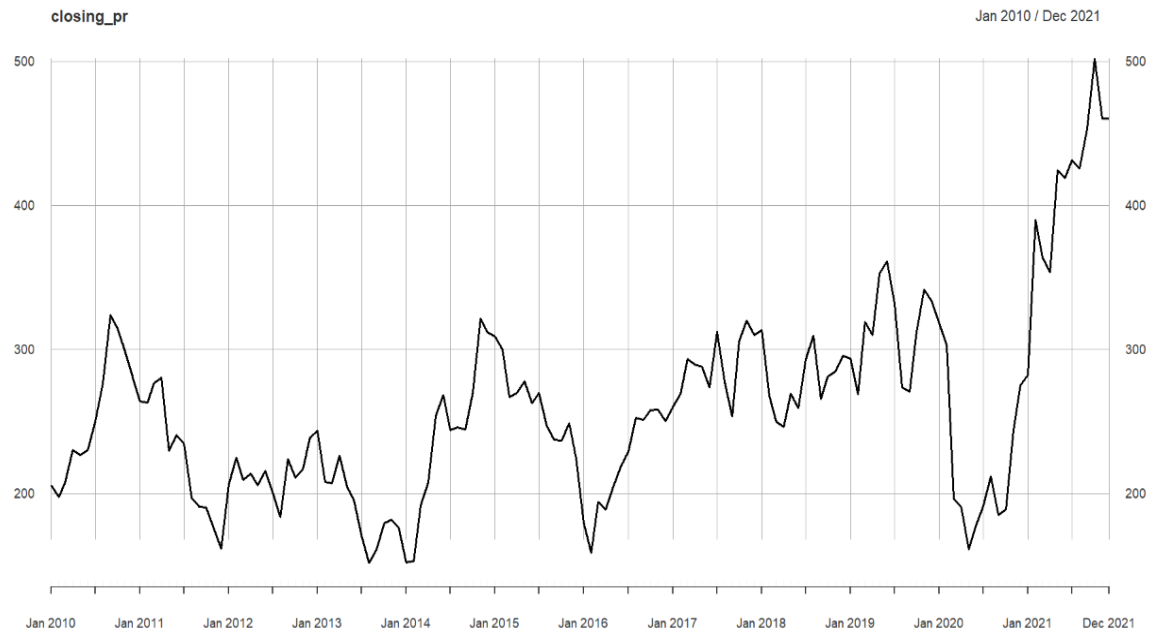
The stock market is a marketplace that allows for the seamless exchange of corporate stock purchases and sales. Every Stock Exchange has its own value for the Stock Index. The index is the average value derived by adding up the prices of various equities. This aids in the representation of the entire stock market as well as the forecasting of market movement over time. The stock market can have a significant impact on individuals and the economy as a whole. As a result, effectively predicting stock trends can reduce the risk of loss while increasing profit.

The data is collected from the *'quantmod'* library of the *'R'* software where the source used is 'yahoo'.

Report:

The data consists of the stock price of State Bank of India from 1st Jan 2010 to 31st Dec 2021. I will consider only the close price of the stock for the analysis. There are in total 2698 data points with 7 missing data points. I have removed the rows i.e. dates with missing data points. Since there are an inconsistent number of data points for each year I will restrict the analysis to the end of month closing price. There are a total of 144 months in the span of the year 2010 to 2021 and hence 144 data points are considered for analysis.
Let us observe the difference between the plot of daily closing price and end of month closing price of SBI.

Plot of daily closing price of SBI from 2010 to 2021



Plot of end of month closing price of SBI from 2010 to 2021

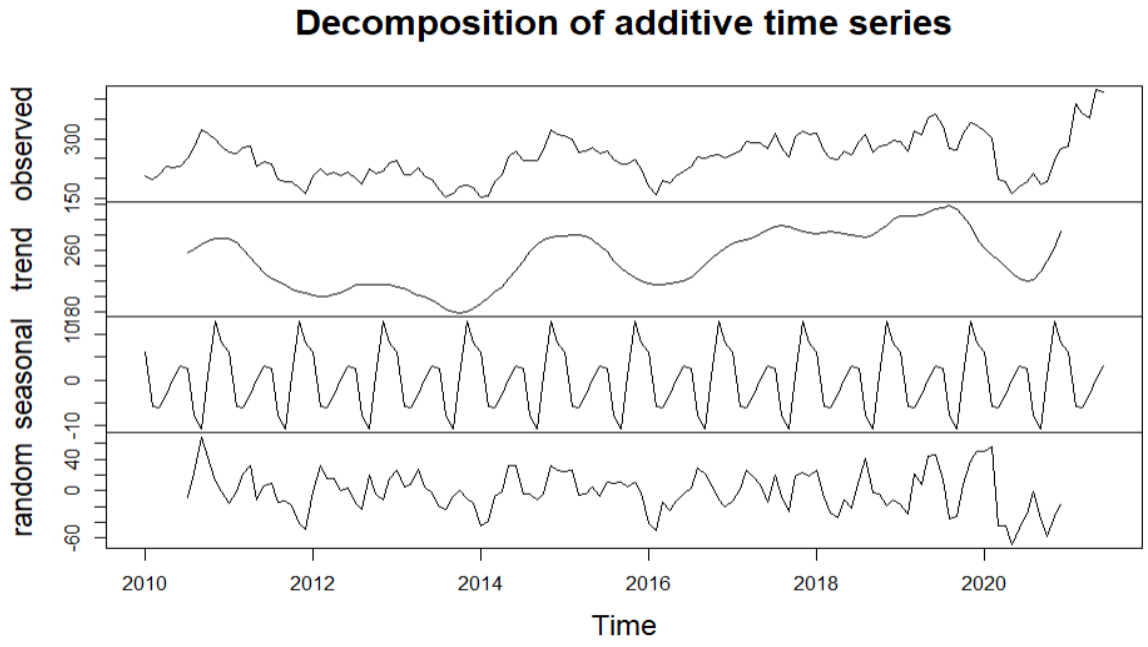
Comparing the two plots I can say that the end of month plot retains the behaviour of the daily time series and hence minimal information is lost. I can safely compare the end of month analysis to daily analysis and can make forecasts accordingly.

I have taken 138 points from Jan 2010 to June 2021 as my training data and remaining 6 points from July 2021 to Dec 2021 as my test data.

Descriptive Statistics of the data:

Mean	260.1
1st Quartile	207.8
Median	253.2
3rd Quartile	293.5
Min	151.9
Max	502.1

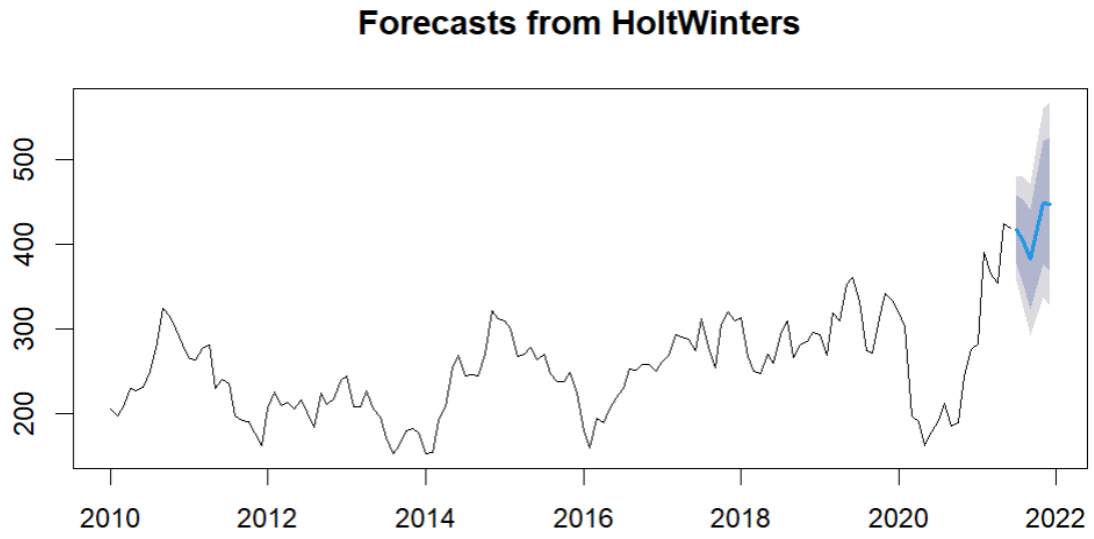
Next I will decompose the data to observe patterns of trend and seasonality.



From the decomposition I can observe negligible presence of trend and some presence of seasonality.

First I will perform **Holt Winter's Exponential Smoothing** on the data.
I got the parameters values as $\alpha = 0.7551427$, $\beta = 0.008374501$ and $\gamma = 1$.

I made a 6 step ahead forecast using the Holt Winter's method.



The test RMSE for Holt Winter's method obtained is 203.45

Next I will try to find a better model in terms of forecasting i.e. a model with lower test RMSE.
Since the data has seasonality I will try to fit a SARIMA$(p,d,q)(P,D,Q)_s$ where the parameters will be decided using the ACF and PACF plots.
Primary step of fitting an ARIMA/SARIMA model is checking if the data is stationary.
I performed ADF test, PP test and KPSS test for stationarity checking.

- **ADF test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: 0.2253
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **PP test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: 0.05164
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **KPSS test:**
  H_0: The time series is stationary
  H_1: The times series is non-stationary

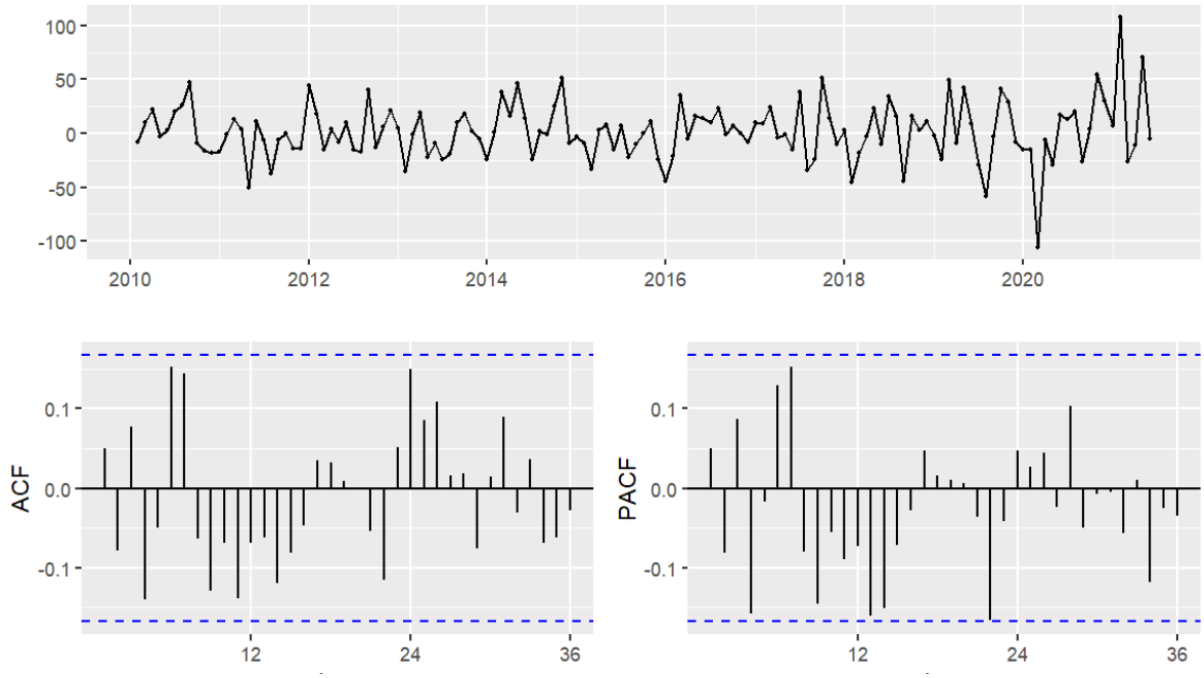
Obtained p-value: < 0.01
Thus, we reject the null hypothesis and conclude data is non-stationary.

All the three tests indicate that the data is non-stationary.

I applied first order differencing i.e. $Y_t^{(1)} = Y_t - Y_{t-1}$ and performed all the three tests.
I got the following results: p-value for ADF test < 0.01,
p-value for PP test < 0.01,
p-value for KPSS test > 0.1.
Thus from all the tests I came to a conclusion that the data became stationary after first order differencing. Hence, I got $d = 1$ for my SARIMA model.

I will find the required seasonal differencing (D) for SARIMA using the *'nsdiffs()'* function. *R* provided the required seasonal differencing as zero.

So, I get the differencing schemes as $d = 1$ and $D = 0$.
Now I plot the stationary time series and its ACF and PACF plots.



ACF and PACF plot of the stationary data

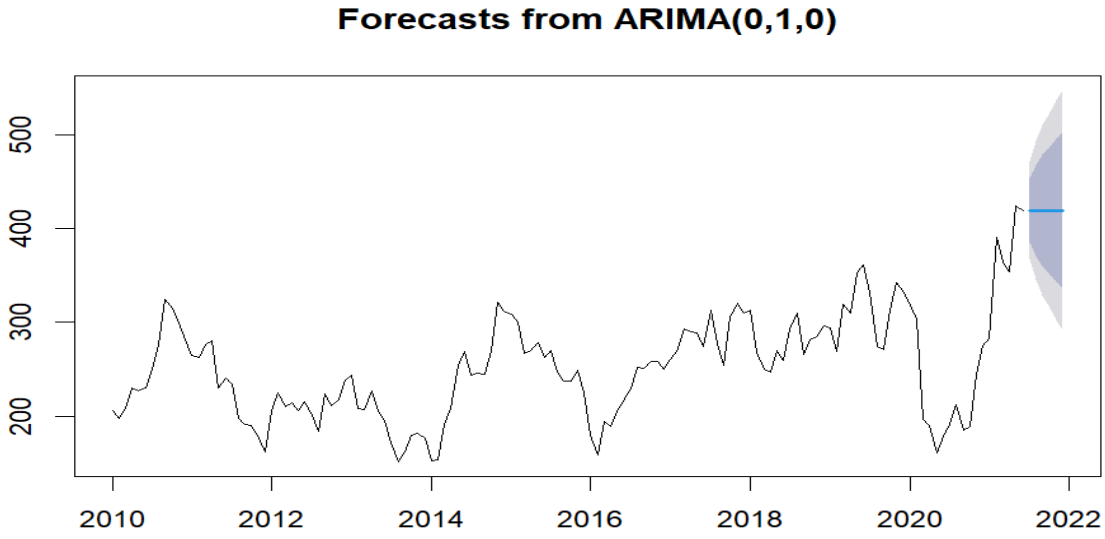
Observing the ACF and PACF plots I can see there are no significant lags in both the plots. So there is no AR and MA process involved both seasonally and non-seasonally. Hence the appropriate model is ARIMA(0,1,0). The AIC of the manually chosen model is 1290.193.
Next I used *'auto.arima()'* to find out the model suggested by *R*. The software provided the appropriate model as ARIMA(0,1,0) which is the same as the manually chosen one.
Thus, the model is given by,

$$Y_t^{(1)} = \epsilon_t \Rightarrow Y_t = Y_{t-1} + \epsilon_t$$

The model becomes a random walk with no drift.

Using the above model I now forecast 6 steps ahead i.e. forecast the end of month closing prices for the last 6 months of year 2021. The test RMSE obtained is 203.2818.

Both ARIMA and Holt Winter's method have produced more or less the same results but the confidence interval of forecasts in Holt winter's method is less wider than that of ARIMA forecasts.



Residual Analysis

The residuals of the model should follow i.i.d sequence with mean zero and constant variance. To check the mean of the residuals equal to zero I have used the '*wilcox.test()*'.

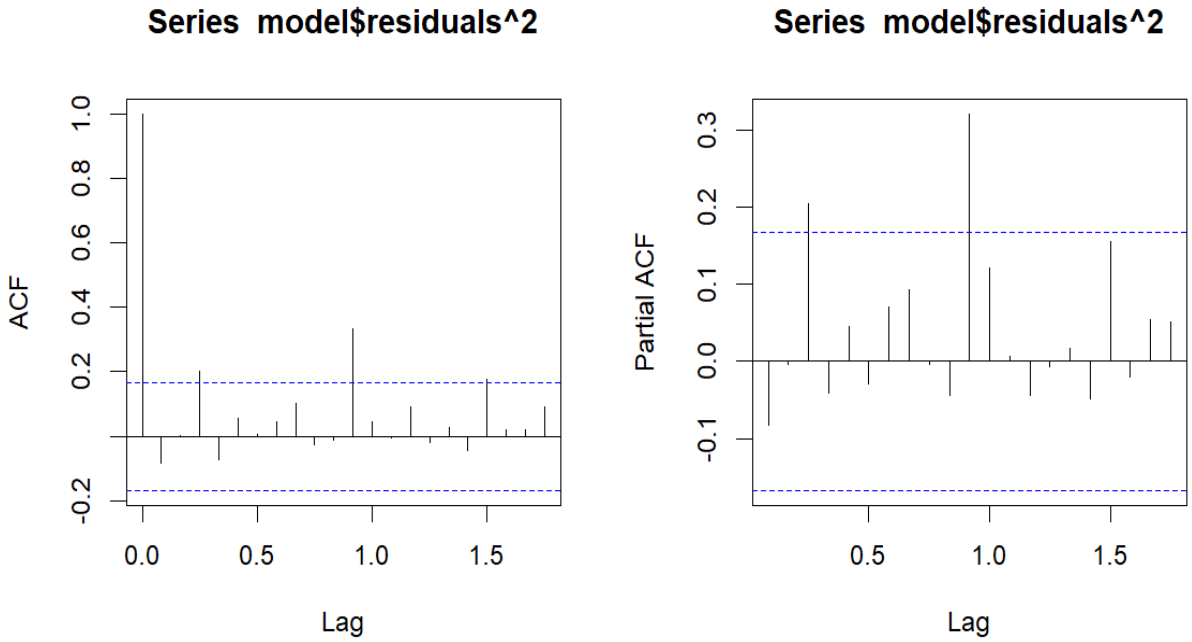
```
        Wilcoxon signed rank test with continuity correction

data:  model$residuals
V = 4991.5, p-value = 0.6778
alternative hypothesis: true location is not equal to 0
```

From the above test I see that the residuals have mean equal to zero.

Next, using the Bartlett test I can conclude that the variance of the residuals is not constant.

```
        Bartlett test of homogeneity of variances

data:  model$residuals and a
Bartlett's K-squared = 28.21, df = 11, p-value = 0.003006
```

Now we can look forward to checking for the presence of ARCH effect.
Next I will observe the ACF and PACF plots of squared residuals of the model.



ACF and PACF of squared residuals

From the ACF and PACF of squared residuals there is minimal evidence of the presence of ARCH effect i.e. the residuals are uncorrelated. I will confirm my observation by performing the Ljung-Box test.

```
        Box-Ljung test

data:  model$residuals^2
X-squared = 10.018, df = 10, p-value = 0.4389
```

The Ljung-Box test accepts the null hypothesis till lag 10 which indicates that there is no ARCH effect present in the model.

3. House price data

Data link: https://www.kaggle.com/htagholdings/property-sales

The data consists of property sales data for the 2007-2019 period for one specific region. The data contains sales prices for houses and units with 1,2,3,4,5 bedrooms. It consists of moving average of median price grouped by quarterly intervals per property type and number of bedrooms.

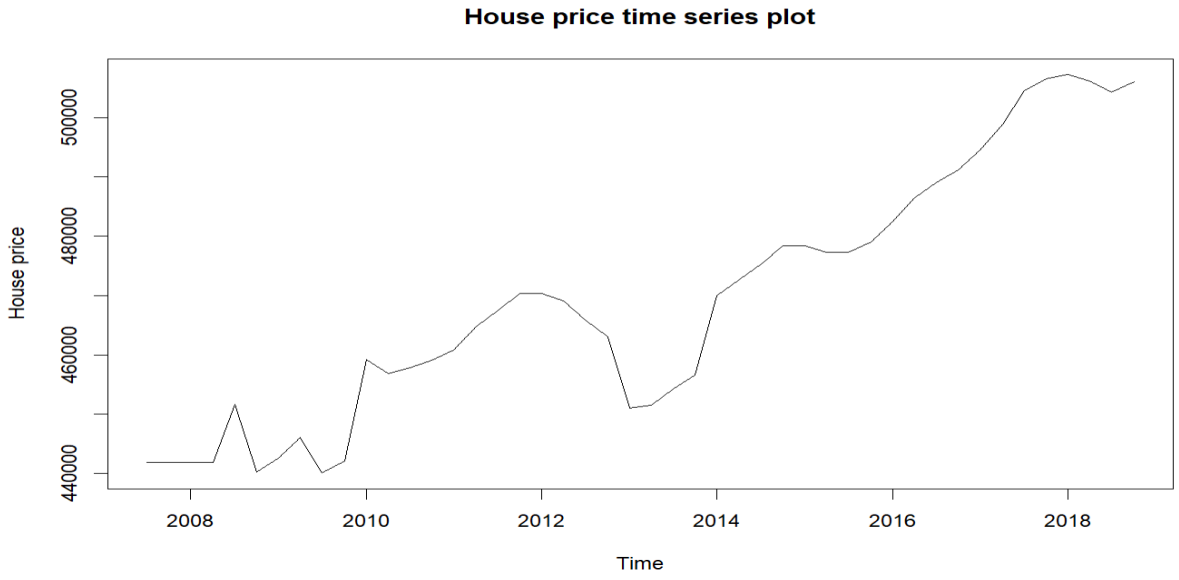
Report:

For my analysis I have restricted the data to type ='houses' and number of bedrooms = 2 since I want to model an univariate time series data. The data consists of median price of property sales from the 3rd quarter of 2007 to the 3rd quarter of 2019. There are a total of 49 data points with no missing values. For the time series analysis I have taken data from the 3rd quarter of 2007 to the 4th quarter of 2018 as my training set. The remaining data i.e. 3 quarters of 2019 is my test set.

Descriptive Statistics of the data:

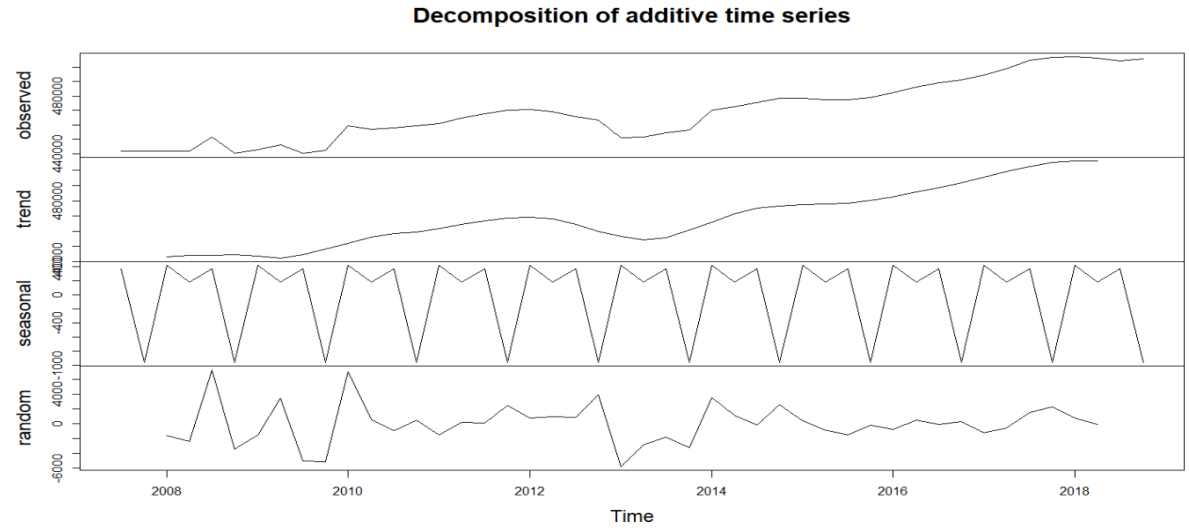
Mean	471454
1st Quartile	454270
Median	469920
3rd Quartile	489104
Min	440123
Max	510712

Let us observe the plot of the data.



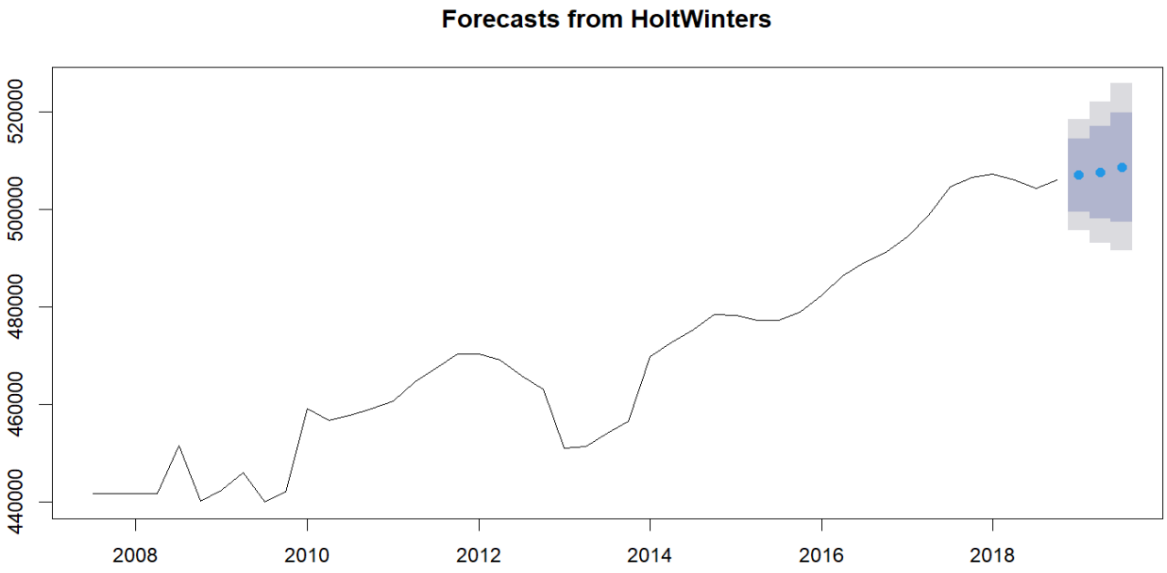
From the data plot I can observe the presence of trends in the data.

Next I will decompose the data to observe patterns of trend and seasonality.



From the decomposition of the data I can observe the presence of both trend and seasonality.
So, I will try to perform **Holt Winter's Exponential Smoothing** on the data.
I got the parameters values as $\alpha = 0.7673283$, $\beta = 0.01836596$ and $\gamma = 1$.

I made a 3 step ahead forecast using the Holt Winter's method.



The test RMSE obtained for Holt Winter's Exponential Smoothing method is 65920.24

Next I will try to find a better model in terms of forecasting i.e. a model with lower test RMSE.
Since the data has seasonality I will try to fit a SARIMA$(p,d,q)(P,D,Q)_s$ where the parameters will be decided using the ACF and PACF plots.
Primary step of fitting an ARIMA/SARIMA model is checking if the data is stationary.
I performed ADF test, PP test and KPSS test for stationarity checking.

- **ADF test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: 0.1546
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **PP test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: 0.3979
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **KPSS test:**
  H_0: The time series is stationary

H_1: The times series is non-stationary
Obtained p-value: < 0.01
Thus, we reject the null hypothesis and conclude data is non-stationary.

All the three tests indicate that the data is non-stationary.

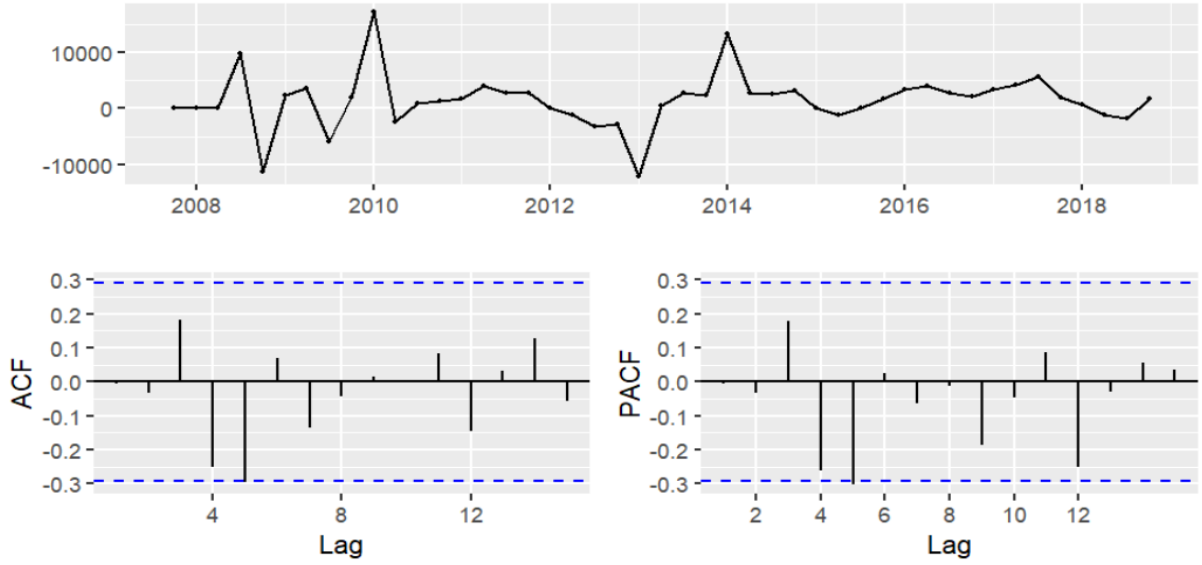
I applied first order differencing i.e. $Y_t^{(1)} = Y_t - Y_{t-1}$ and performed all the three tests.
I got the following results: p-value for ADF test = 0.04605,
p-value for PP test < 0.01 ,
p-value for KPSS test > 0.1.

Thus from all the tests I came to a conclusion that the data became stationary after first order differencing. Hence, I got *d =1* for my SARIMA model.

I will find the required seasonal differencing (*D*) for SARIMA using the *'nsdiffs()'* function. *R* provided the required seasonal differencing as zero.

So, I get the differencing schemes as *d = 1* and *D = 0*.

Now I plot the stationary time series and its ACF and PACF plots.



Observing the ACF and PACF plots I can see there are no significant lags in both the plots. So there is no AR and MA process involved both seasonally and non-seasonally.

Hence by manual search the appropriate model seems to be ARIMA(0,1,0). The AIC of the manually chosen model is 895.8018.

Next I used *'auto.arima()'* to find out the model suggested by *R*. The software provided the appropriate model as SARIMA$(0,1,0)(0,0,1)_4$. The AIC of the model suggested by the software is 892.5721.

The SARIMA model is given by,

$$\Phi(B^s)(1-B^s)^D\phi(B)(1-B)^d X_t = \Theta(B^s)\Theta(B)Z_t$$

where *p=0, d=1, q=1 , P=0, D=0* and *Q=1*
Thus the model is,

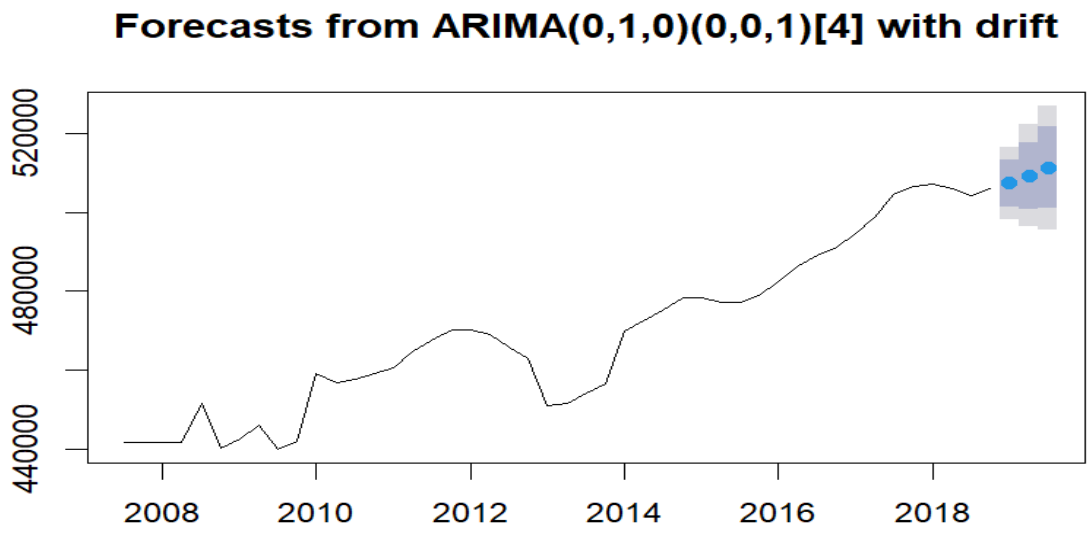
$$(1-B)\ X_t = \Theta(B^4)\Theta(B)Z_t$$

Next I have tested the significance of the model. I have done that by checking the p-values of the coefficients.

```
z test of coefficients:

        Estimate Std. Error z value Pr(>|z|)
sma1    -0.35125    0.17350 -2.0245 0.042922 *
drift 1448.39581  461.35432  3.1394 0.001693 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All the coefficients have p-values less than 0.05 which means the model coefficients are significant.

Using the above model suggested by *R* I have forecasted 3 steps ahead i.e. forecast for the house prices in the 3 quarters of year 2019. The test RMSE obtained is 67504.71.

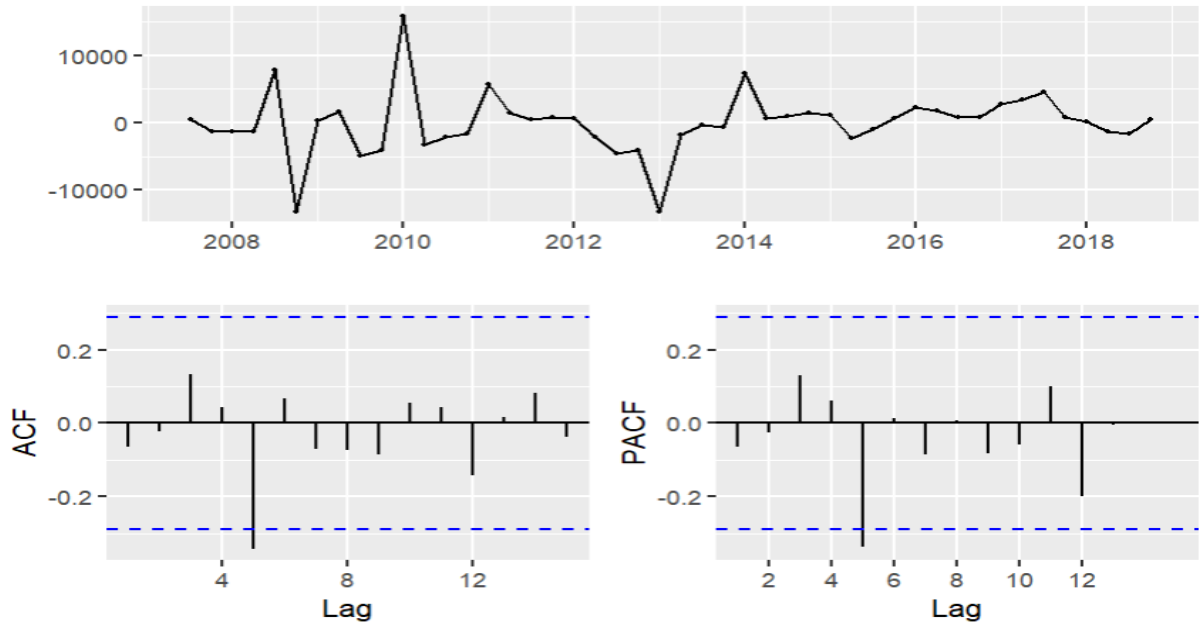


From the comparison RMSEs I can say that Holt Winter's Exponential Smoothing method has provided better forecasts than the SARIMA model.

Residual Analysis

Next I have analysed the residuals of the models.
First I have checked for the stationarity of the residuals of the model using the KPSS test. The p-value obtained was > 0.1 which indicated that the residuals are stationary.

Next I have plotted the ACF and PACF plots of the model residuals.



ACF and PACF of model residuals

The residuals of the model should follow i.i.d sequence with mean zero and constant variance. To check the mean of the residuals equal to zero I have used the '*wilcox.test()'*.

```
        Wilcoxon signed rank test with continuity correction

data:  model$residuals
V = 538, p-value = 0.9826
alternative hypothesis: true location is not equal to 0
```

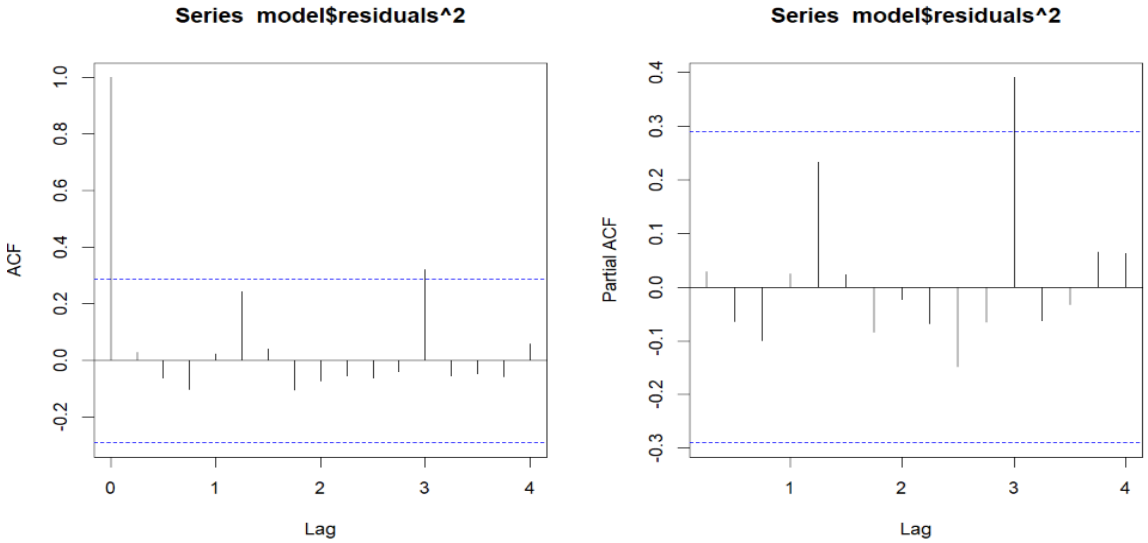
From the above test I see that the residuals have mean equal to zero.

Next, using the Bartlett test I can conclude that the variance of the residuals is not constant.

```
        Bartlett test of homogeneity of variances

data:  model$residuals and a
Bartlett's K-squared = 32.409, df = 11, p-value = 0.0006555
```

Now we can look forward to checking for the presence of ARCH effect.

Next I will observe the ACF and PACF plots of squared residuals of the model.



ACF and PACF of squared residuals

From the ACF and PACF of squared residuals there is almost no evidence of the presence of ARCH effect i.e. the residuals are uncorrelated. I will confirm my observation by performing the Ljung-Box test.

```
        Box-Ljung test

data:  res^2
X-squared = 4.9853, df = 8, p-value = 0.7591
```

The Ljung-Box test accepts the null hypothesis which indicates that there is no ARCH effect present in the model.

4. Fertility rate of woman in India

Data link: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN?locations=IN

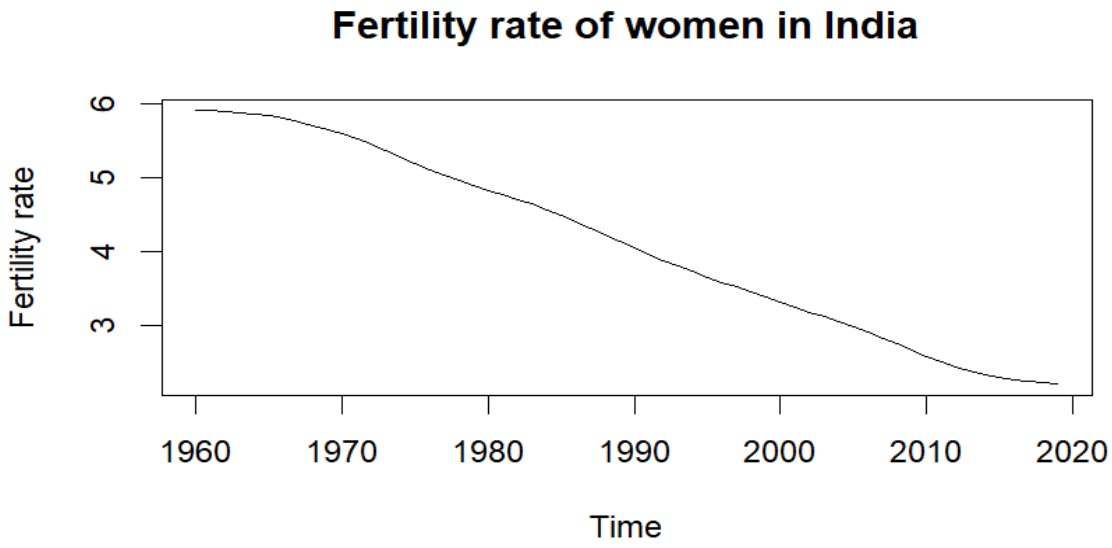
Report:

The data consists of the fertility rate of women in India from the year 1960 to 2019. So, there are a total of 60 data points with no missing data. For the time series analysis I have taken 55 data points i.e. data from 1960 to 2014 as my training data. The remaining 5 data i.e. data from 2015 to 2019 is my test data.

Descriptive Statistics of the data:

Mean	4.100
1st Quartile	3.024
Median	4.088
3rd Quartile	5.213
Min	2.202
Max	5.906

Let's plot the data.



From the plot we can clearly see there is a downward trend with no seasonality as the data is yearly.

For this data Holt Winter's Exponential smoothing is not possible because it can be applied on the data which exhibits both trend and seasonality.
Since the data has no seasonality I will try to fit an ARIMA *(p,d,q)* where the parameters will be decided using the ACF and PACF plots.
Primary step of fitting an ARIMA model is checking if the data is stationary.
I performed ADF test, PP test and KPSS test for stationarity checking.

- **ADF test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: 0.6161
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **PP test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: 0.4631
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **KPSS test:**
  H_0: The time series is stationary
  H_1: The times series is non-stationary
  Obtained p-value: < 0.01
  Thus, we reject the null hypothesis and conclude data is non-stationary.

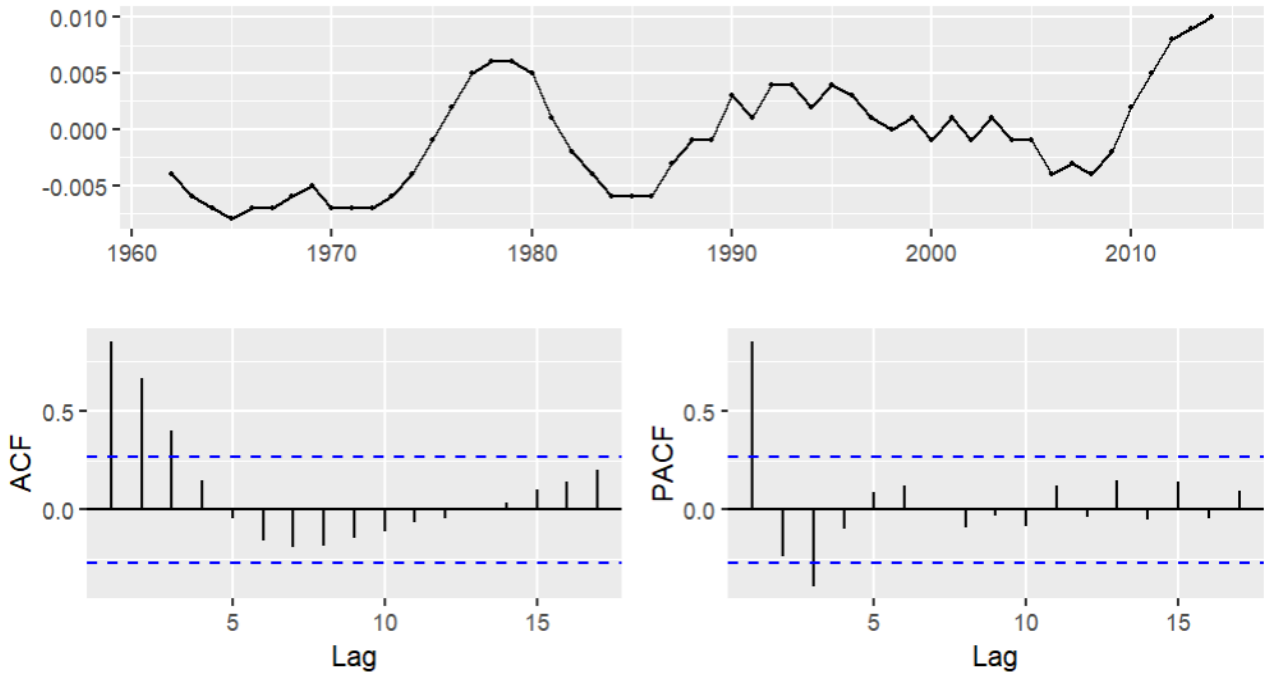
All the three tests indicate that the data is non-stationary.

I applied first order differencing i.e. $Y_t^{(1)} = Y_t - Y_{t-1}$ and performed all the three tests.
I got the following results: p-value for ADF test = 0.6146,
p-value for PP test = 0.8006,
p-value for KPSS test = 0.03933.

Thus, all the tests indicate the data is still non stationary.
After applying the second order differencing $Y_t^{(2)} = Y_t^{(1)} - Y_{t-1}^{(1)}$ and performing all the tests of stationarity,
I got the following results: p-value for ADF test < 0.01,
p-value for PP test = 0.0221,
p-value for KPSS test = 0.0835.

All the tests indicate that the data is stationary after second order differencing.
So, I get $d = 2$.

Now I plot the stationary time series and its ACF and PACF plots.



ACF and PACF of stationary data

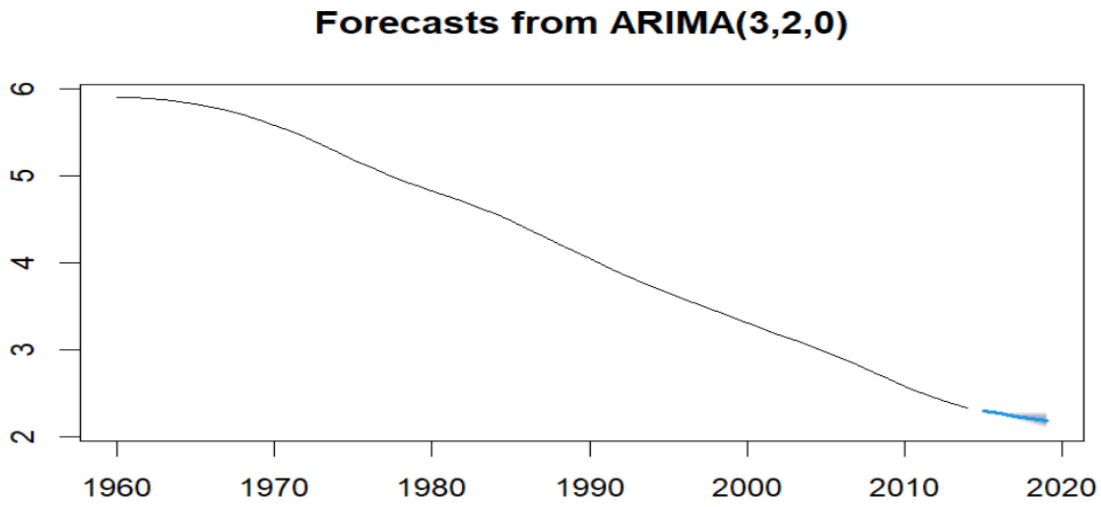
From the PACF plot I can see the last significant spike at lag 3 which suggests an AR(3) process. The ACF plot shows an exponential decay which suggests no presence of MA process in the model. So, by manually finding the ARIMA model by looking at the ACF and PACF plots it seems that the appropriate model is ARIMA(3,2,0). This model has an AIC of -524.523.
Next I used *'auto.arima()'* to find out the model suggested by *R*. The software provided the appropriate model as ARIMA(3,2,0) which is the same as the manually chosen one.

Next I have tested the significance of the model. I have done that by checking the p-values of the coefficients.

```
z test of coefficients:

    Estimate Std. Error z value  Pr(>|z|)
ar1  1.04543    0.10905  9.5866 < 2.2e-16 ***
ar2  0.32729    0.17388  1.8823   0.05979 .
ar3 -0.57022    0.11442 -4.9837 6.238e-07 ***
```

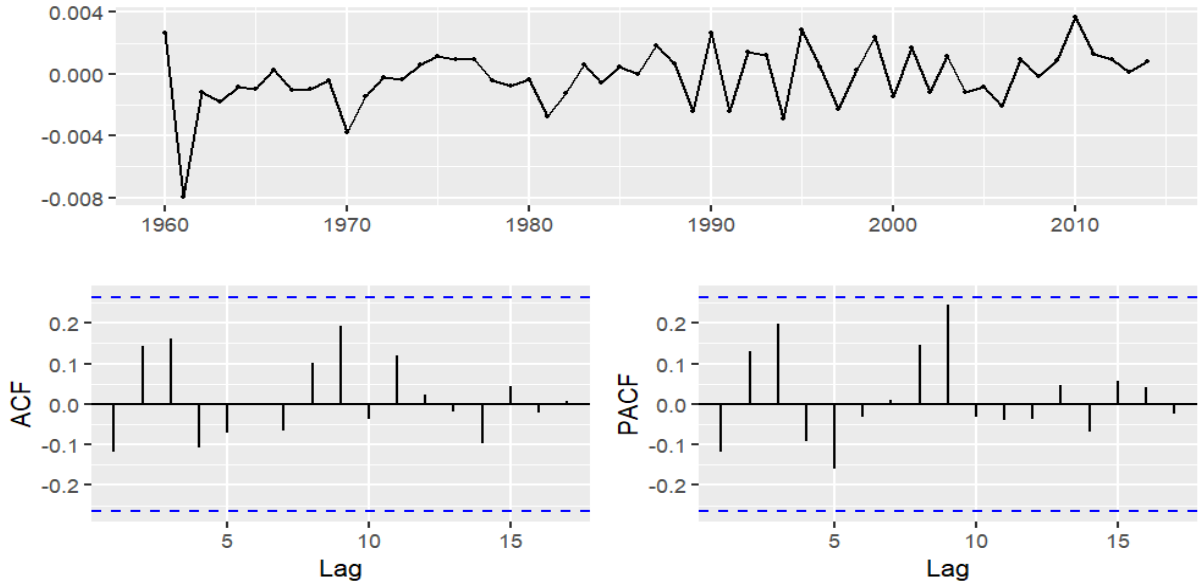
All the coefficients are significant except the second coefficients as it is slightly more than 0.05. Since the p-value is just 0.009 more than the significant level we can consider it significant.

Using the above model suggested by *R* I have forecasted 5 steps ahead i.e. forecast for the fertility rate of the years 2015 to 2019. The test RMSE obtained is 3.650599. Considering the test data values the obtained RMSE indicates fairly good forecast given the model.



Residual Analysis

Next I have analysed the residuals of the models.

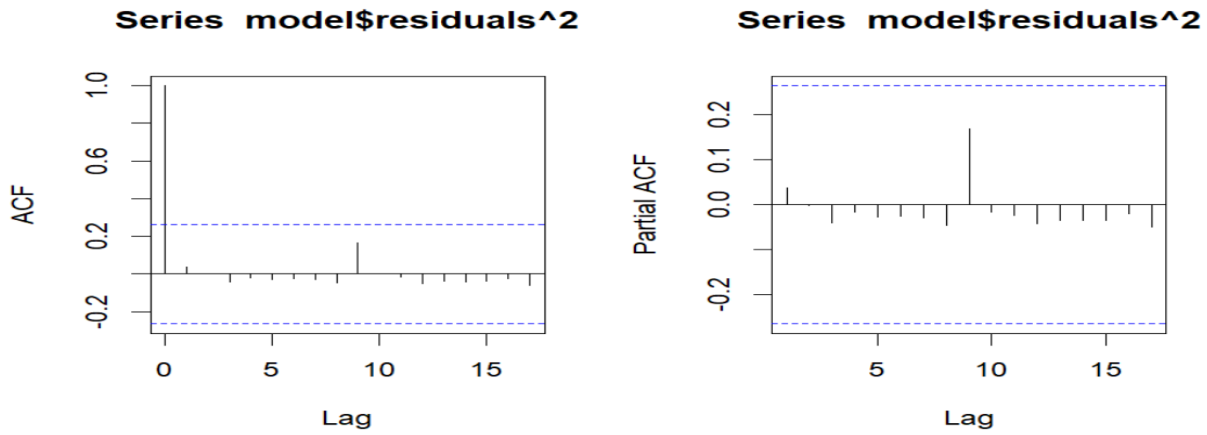


ACF and PACF of model residuals

The residuals of the model should follow i.i.d sequence with mean zero and constant variance. To check the mean of the residuals equal to zero I have used the '*wilcox.test()*'.

```
        Wilcoxon signed rank test with continuity correction

data:  model$residuals
V = 711, p-value = 0.624
alternative hypothesis: true location is not equal to 0
```

From the above test I see that the residuals have mean equal to zero.



ACF and PACF of squared residuals

From the ACF and PACF of squared residuals there is no evidence of the presence of ARCH effect i.e. the residuals are uncorrelated. I will confirm my observation by performing the Ljung-Box test.

```
        Box-Ljung test

data:  model$residuals^2
X-squared = 0.080752, df = 1, p-value = 0.7763
```

The Ljung-Box test accepts the null hypothesis which indicates that there is no ARCH effect present in the model.

5. Monthly rainfall in India

Data link: https://data.gov.in/resources/sub-divisional-monthly-rainfall-1901-2017

The data consists of monthly rainfall levels in mm for all subdivisions in India from 1901 to 2017.

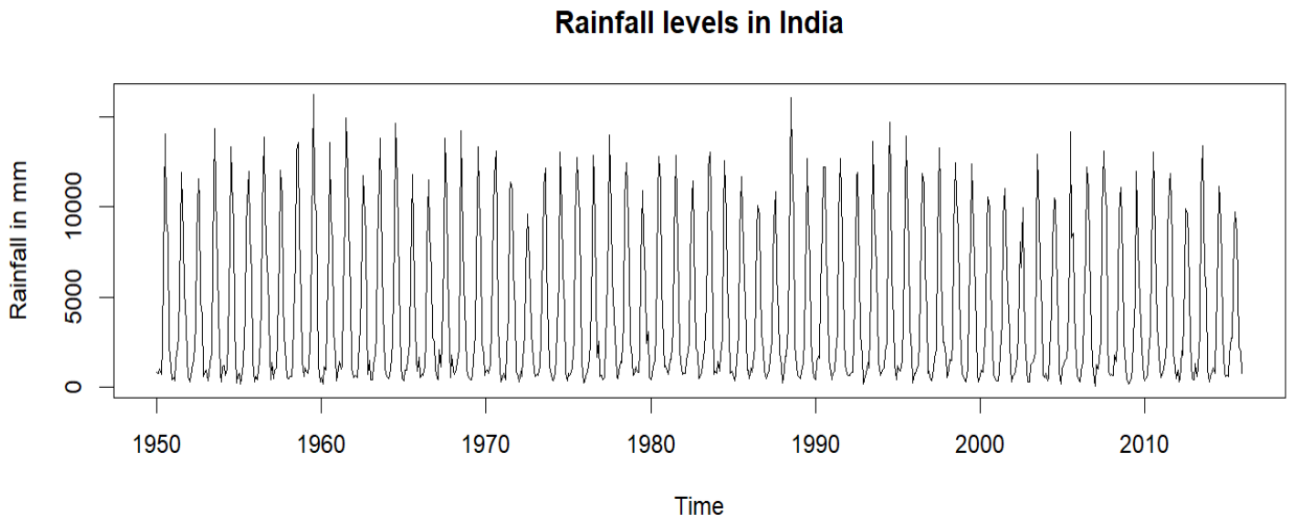
Report:

The data considered for the analysis consists of the aggregated sum over all the divisions for each month of each year and produced as a whole Indian data. There were some missing points which were replaced by 0 and aggregated over the sub-divisions. For my time series analysis I have taken data from the months of the year 1950 to 2015 i.e. a total of 792 data points. The remaining 24 data points i.e. year 2016 and 2017 are my test data points.

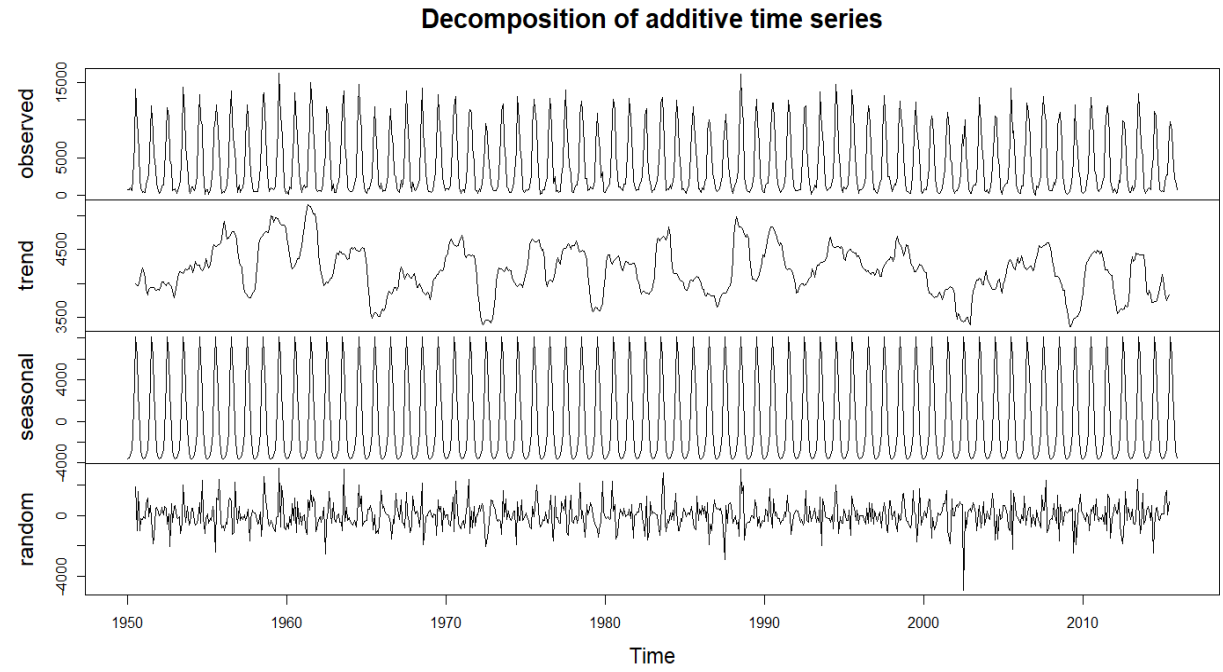
Descriptive Statistics of the data:

Mean	4189.1
1st Quartile	873.5
Median	2158.2
3rd Quartile	7550.8
Min	16196.5
Max	5.906

Let's plot the data.



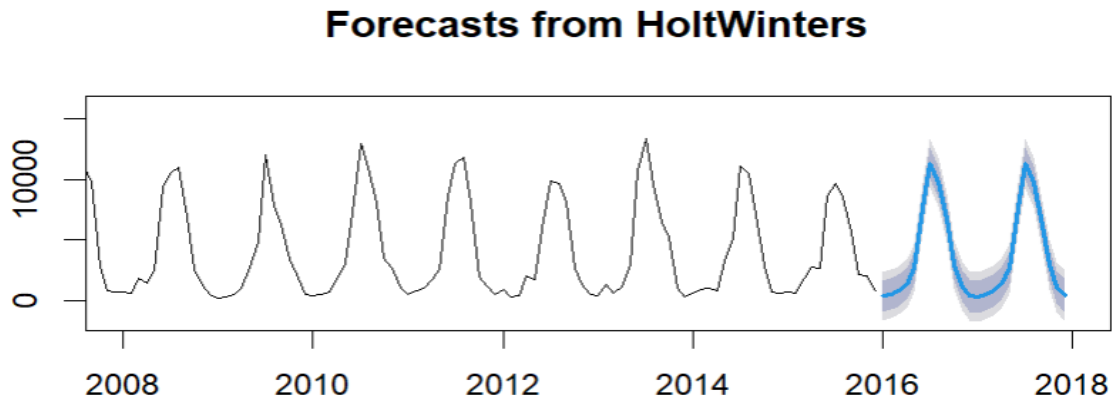
Next we will decompose the data to observe the presence of trend and seasonality.



We can clearly see there is a presence of seasonality and negligible presence of trend.

I will perform the **Holt Winter's Exponential Smoothing** method to see how the model performs. I got the parameters as $\alpha = 0.04905488$, $\beta = 0.005389185$ and $\gamma = 0.1194357$.

I made a 24 step ahead forecast using the Holt Winter's model.



The RMSE of the test data for Holt Winter's forecast is 897.1364

I will try to find a better model whose forecast should give a lower test RMSE.
Since the data has seasonality, I will try to fit a SARIMA$(p,d,q)(P,D,Q)_s$ where the parameters will be decided using the ACF and PACF plots.
Primary step of fitting an ARIMA/SARIMA model is checking if the data is stationary.
I performed ADF test, PP test and KPSS test for stationarity checking.

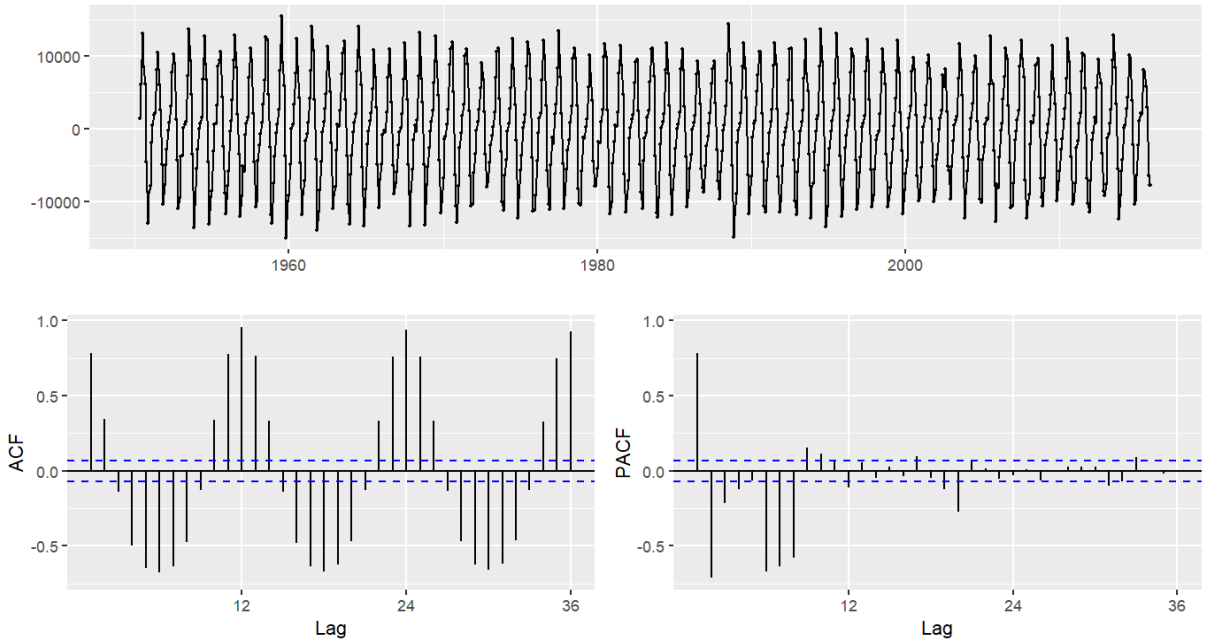
- **ADF test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: < 0.01
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **PP test:**
  H_0: The time series is non-stationary
  H_1: The times series is stationary
  Obtained p-value: < 0.01
  Thus, we fail to reject the null hypothesis and conclude data is non-stationary.
- **KPSS test:**
  H_0: The time series is stationary
  H_1: The times series is non-stationary
  Obtained p-value: > 0.1
  Thus, we reject the null hypothesis and conclude data is non-stationary.

All the three tests indicated that the data is **stationary**.

I will find the required seasonal differencing using the *'nsdiffs()'* function. It estimates the number of seasonal differences necessary to make the time series stationary.
R provided the required seasonal differencing as one.

So, I get the differencing schemes as $d = 0$ and $D = 2$.

Now I plot the stationary time series and its ACF and PACF plots.



ACF and PACF of stationary data

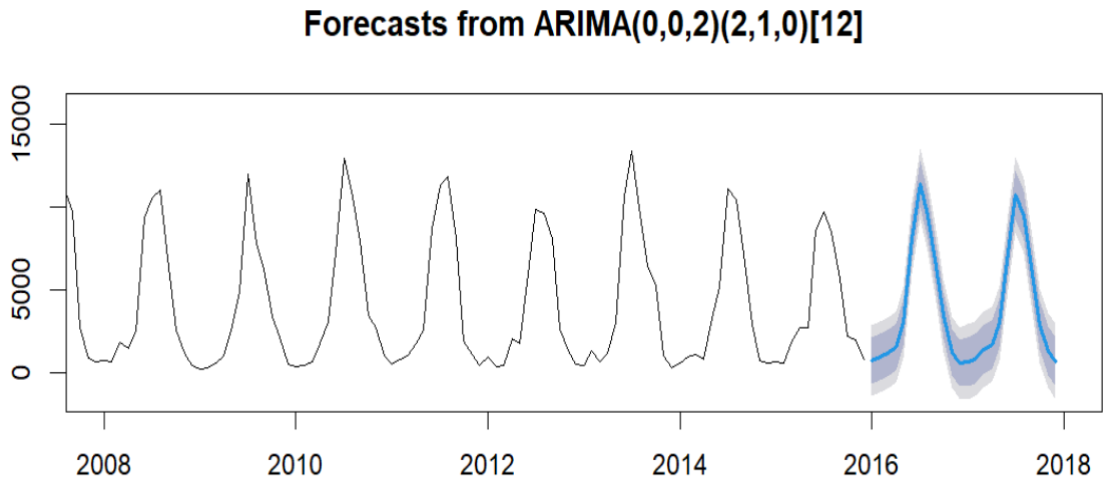
From the PACF plot we can see a significant spike at lag 2 after which it fades off which suggests non-seasonal AR(2) process. A small seasonal spike is present at lag 12 which may suggest a seasonal AR(1) process, From the ACF plot it is hard to infer the MA processes as it forms a sinusoidal curve.
I will use the *'auto.arima()'* to directly find the best model suggested by the software. The model suggested is SARIMA(0,0,2)(2,1,0)$_{[12]}$ with an AIC of 13128.21.

Next I have tested the significance of the model. I have done that by checking the p-values of the coefficients.

```
z test of coefficients:

      Estimate Std. Error  z value  Pr(>|z|)
ma1   0.127818   0.035949   3.5556 0.0003772 ***
ma2   0.095396   0.034168   2.7920 0.0052388 **
sar1 -0.667645   0.034356 -19.4330 < 2.2e-16 ***
sar2 -0.326876   0.034709  -9.4176 < 2.2e-16 ***
```

All the coefficients have p-values less than 0.05 which means the model coefficients are significant.

Using the above model suggested by *R* I have forecasted 24 steps ahead i.e. forecast for the rainfall levels in the months of year 2016 and 2017. The test RMSE obtained is 944.8866.

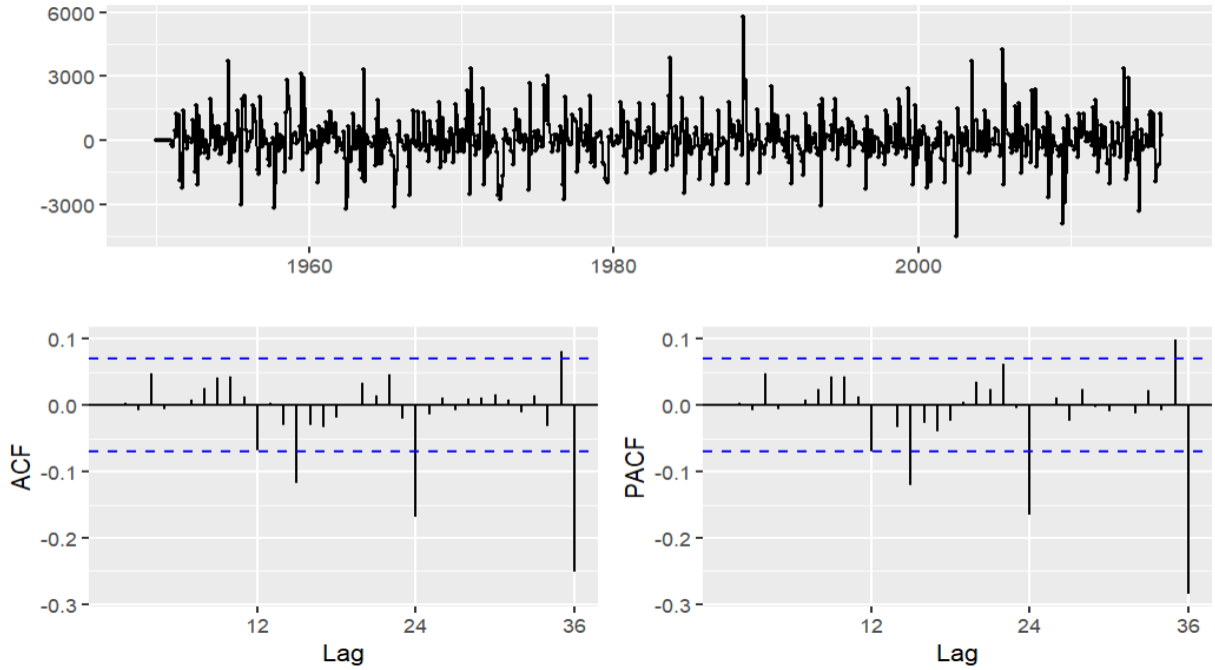


From the comparison RMSEs I can say that Holt Winter's Exponential Smoothing method has provided better forecasts than the SARIMA model.

Residual Analysis

Next I have analysed the residuals of the models.
First I have checked for the stationarity of the residuals of the model using the KPSS test. The p-value obtained was > 0.1 which indicated that the residuals are stationary.

Next I have plotted the ACF and PACF plots of the model residuals.



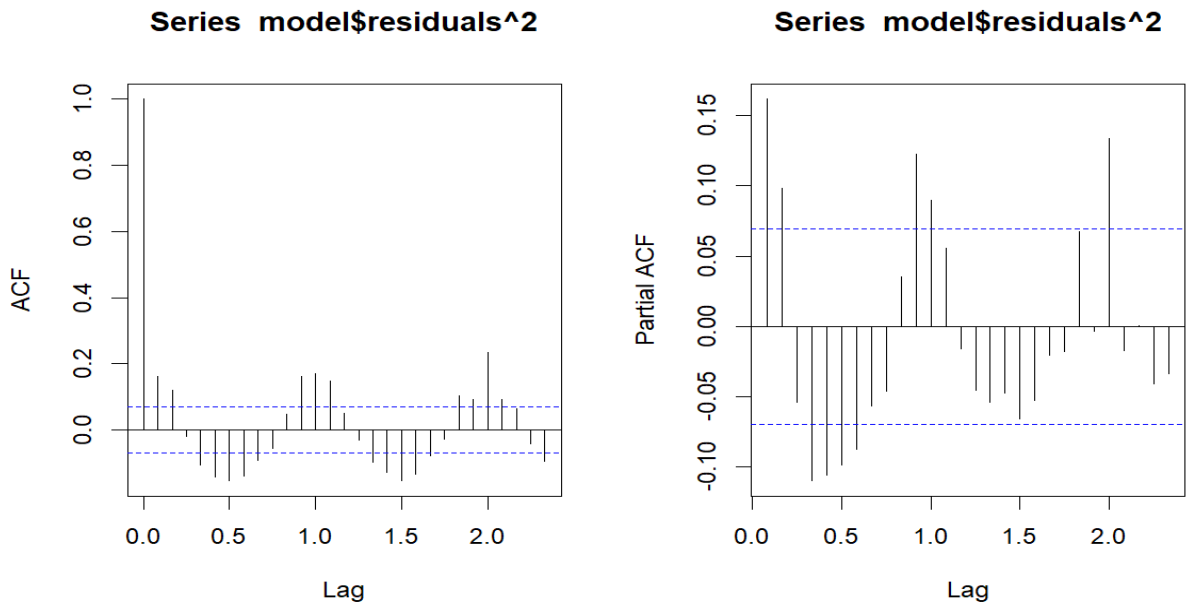
ACF and PACF of model residuals

The residuals of the model should follow i.i.d sequence with mean zero and constant variance. To check the mean of the residuals equal to zero I have used the '*wilcox.test()*'.

```
        Wilcoxon signed rank test with continuity correction

data:  model$residuals
V = 153367, p-value = 0.5713
alternative hypothesis: true location is not equal to 0
```

From the above test I see that the residuals have mean equal to zero.

Next, I observe the ACF and PACF plots of the squared residuals of the model.



ACF and PACF plots of squared residuals

Observing the above plots I can infer that there is ARCH effect present in the model.

By applying the Ljung-Box test on the squared residuals I will confirm my observations.

```
        Box-Ljung test

data:  model$residuals^2
X-squared = 146.37, df = 12, p-value < 2.2e-16
```

From the PACF plot I can see a significant spike till lag 2. I will start with the ARCH(2) model till I get a model with all significant coefficients and lowest AIC. By trial and error I get that ARCH(2) is the best model.

Title:
GARCH Modelling

Call:
garchFit(formula = ~1 + garch(2, 0), data = model$residuals,
trace = F)

Mean and Variance Equation:
 data ~ 1 + garch(2, 0)
<environment: 0x000001fe6f9a7db8>
 [data = model$residuals]

Conditional Distribution:
 norm

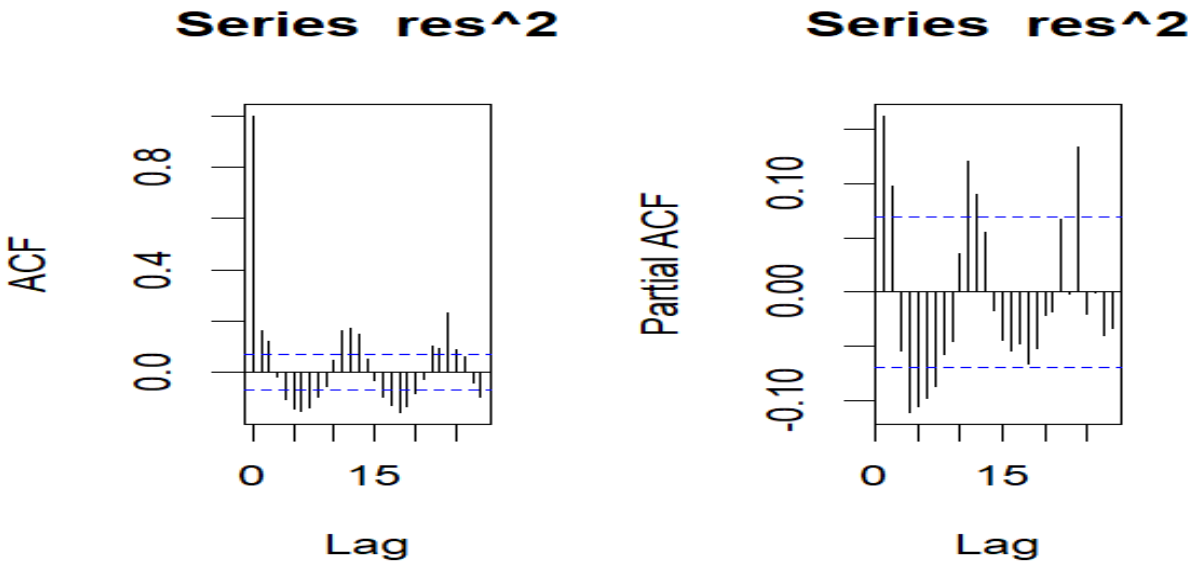
Coefficient(s):
       mu   omega   alpha1   alpha2
2.0861e+01 7.0782e+05 3.5405e-01 1.0395e-01

Std. Errors:
 based on Hessian

Error Analysis:
       Estimate Std. Error t value Pr(>|t|)
mu   2.086e+01   3.286e+01   0.635   0.5255
omega 7.078e+05   6.419e+04   11.026 < 2e-16 ***
alpha1 3.540e-01   8.602e-02   4.116 3.86e-05 ***
alpha2 1.040e-01   5.381e-02   1.932   0.0534 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log Likelihood:
 -6617.611   normalized: -8.35557



ACF and PACF plots of residuals of ARCH model

The ARCH(2) could not model the volatility properly and the residual modelling requires further analysis.

END OF REPORT