University of Nottingham

Department of Mathematical Sciences

MATH 4022- Time Series Analysis

Coursework, Spring 2024

by

Raonak Shukla

Student Id-20601487

"If I can't picture it, I can't understand it."- ***Albert Einstein***

I have read and understood the School and University guidelines on plagiarism. I confirm that this work is our own, apart from the acknowledged references.

Executive Summary

For this coursework, two datasets are provided for time series modeling. The first dataset contains the annual mean temperature in degrees Celsius for the years 1900 to 2021 (inclusive) for the Midlands region of England, recorded by the UK Meteorological Office Hadley Climate Centre. The second dataset contains mean house sale prices in East Midlands (in £GBP), calculated monthly, from January 2010 to December 2019 and the job is to forecast some future average house prices.

The approach for both problems is to check for stationarity, if the series is not stationary then go for transformation-like differencing. For the first problem, first-order differencing is used, and the second problem requires second-order differencing to achieve stationarity. To statistically prove stationarity has been achieved Augmented Dickey-Fuller (ADF) Test and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test are used. The next step involves plotting the Auto Correlation Function(ACF) and Partial Auto Correlation Function(PACF) for both series. Based on observations made from both plots it is found that the differenced series are ARMA(0,1,1) and SARIMA$(2, 2, 1)(0, 1, 1)_{12}$ respectively. For time series modeling Maximum Likelihood Method is used to find coefficients. It should be noted that the above models mentioned are the models obtained after a lot of iterations and analysis of residuals.

Finally, residual plots show that all the residuals are less than $\pm 2/\sqrt{n}$. Apart from the plots, the Ljung-Box Test proves the hypothesis that residuals are uncorrelated. For both problems, it was found that residuals are uncorrelated and normally distributed with mean zero for the final models. Additionally for the second problem forecasting is done with a confidence interval of 95% for the next six months. It might be visualized as shaded areas around the forecasted line, representing the range where the actual future values are expected to fall with 95% probability.

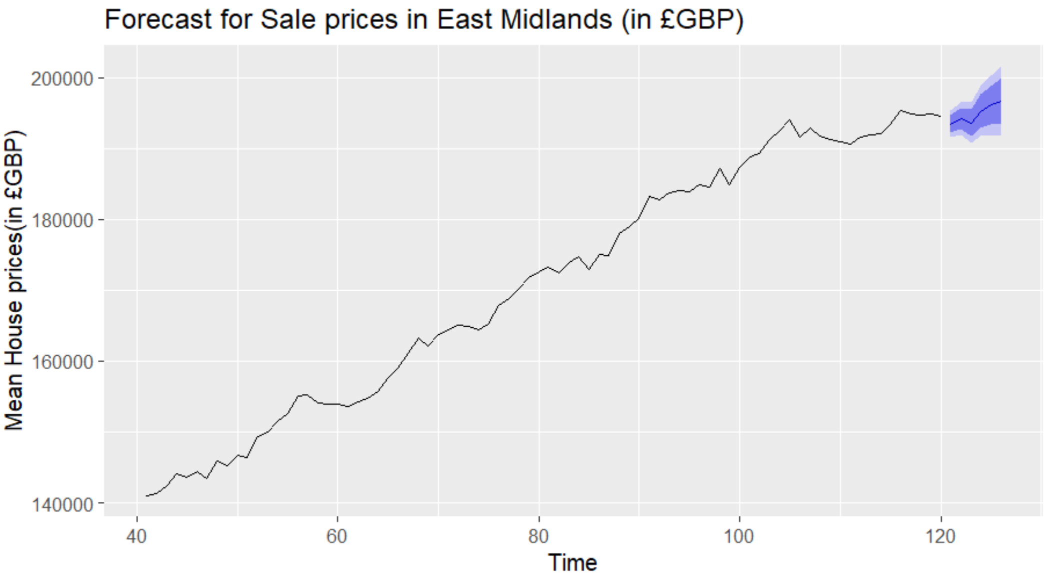

Figure 1: Forecast for next six months

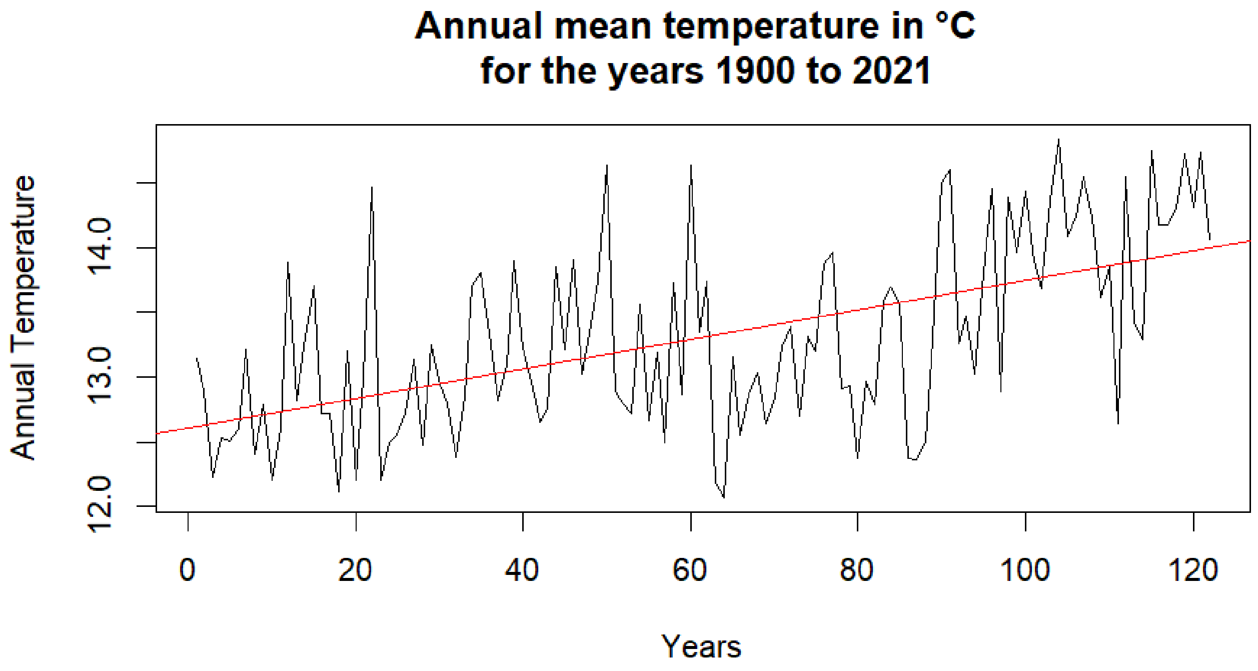

Figure 2: Original Time Series plot for Temperature

Problem 1

In the first problem, we are required to fit a model on a dataset. The dataset contains the annual mean temperature in degrees Celsius for the years 1900 to 2021 (inclusive) for the Midlands region of England, recorded by the UK Meteorological Office Hadley Climate Centre. The methodology adopted is to first plot the time series and check for stationarity. If not present then go for basic transformation. Once the time series is nearly stationary plot Auto Correlation Function(ACF) and Partial Auto Correlation Function(PACF). Based on the observation made from the ACF and PACF plots the order of the time series will be determined to fit the data. Finally, a correlogram is generated for residuals to check whether residuals are correlated also Ljung-Box test will statistically prove that residuals of the best model behave like white noise.

From the plot(Figure 2) above it can be easily determined that the Average annual temperature ranges from 12.07 to 14.84 Celsius with a minimum for the years 1963 and 2003 respectively. The trend line shown in red color in the plot shows the linear trend for the time series. The mean of the series appears to be increasing from 1900-2021 making it non-stationary. Stationarity is important before moving forward as there is an underlying assumption in ARIMA that the time series is stationary. Fitting a nonstationary time series to ARIMA models can lead to incorrect coefficients for AR and MA terms.

A time series can be considered stationary if the following statistical properties don't vary over time.

1. **Constant Mean:** The average value of the series should stay roughly the same throughout

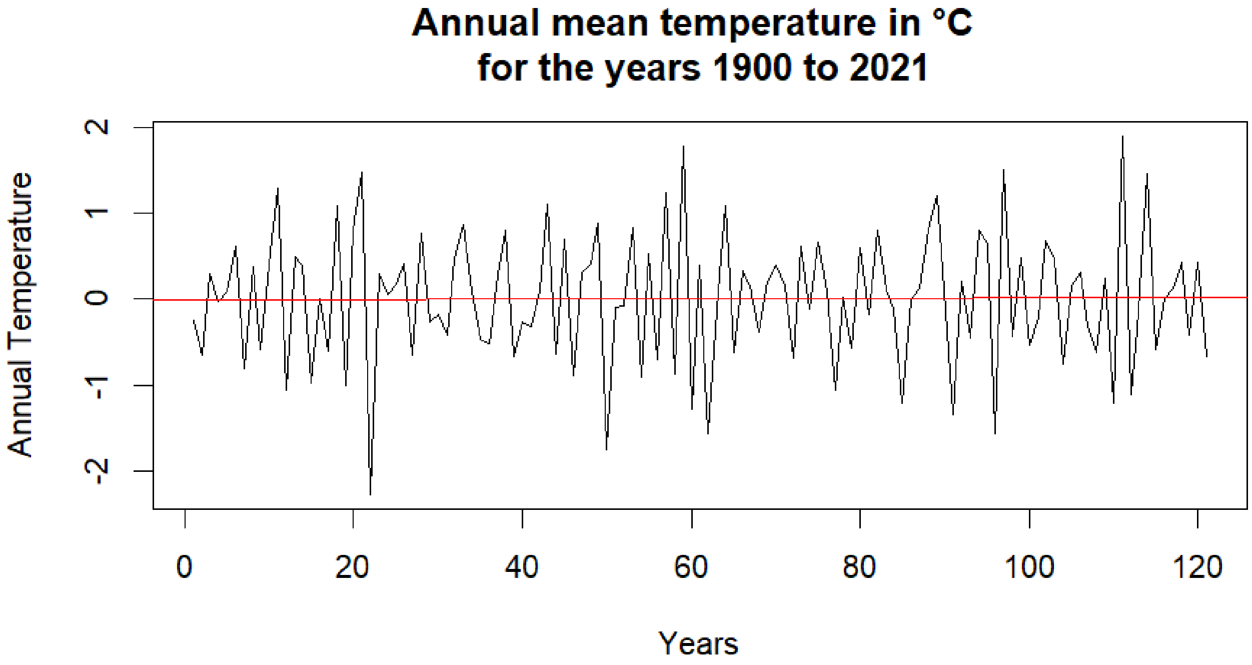

Figure 3: Time Series plot after differencing for Temperature

the whole series. There shouldn't be an upward trend which is present in our data.

2. **Constant variance :**The spread of the data around the mean should be consistent.

3. **Constant covariance :**If past values influence future values in a way that isn't consistent over time, the series may not be stationary.

There are many techniques to make time series stationary these include differencing, detrending, and filtering. To achieve stationarity for this problem, differencing of first order is used for this problem. It is visible from the plot below(Figure 3) that all the three conditions required for stationarity are met after differencing i.e. the new time-series has constant mean, variance, and autocorrelation only depends only on the lag. Further to prove our hypothesis that the differenced time series is stationary Augmented Dickey-Fuller (ADF) Test was used and a p-value of 0.01 was obtained. A low value of p-value implies that our series is stationary. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is also employed to check the stationarity of the process the p-value obtained is 0.03 which also shows that differenced time series is stationary.

After differencing we plot the histogram(Figure 4) for the differenced series to check the distribution of temperature for the given period. From the plot, it is visible that temperatures are normally distributed with mean equal to zero which is also one of the conditions for white noise. The normality plot(Figure 5) depicts the same picture as histogram. The reason for having histograms and normality plots in exploratory data analysis is to get a general sense of the distribution of your time series data. There are no outliers in the time-seires.

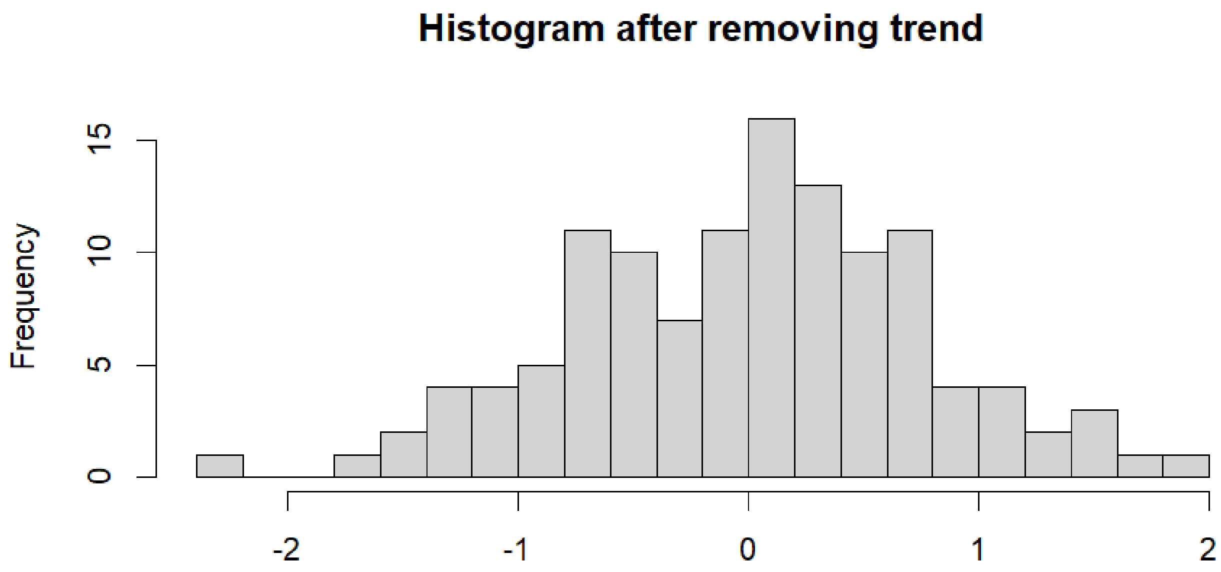

Figure 4: Histogram after removal of Trend

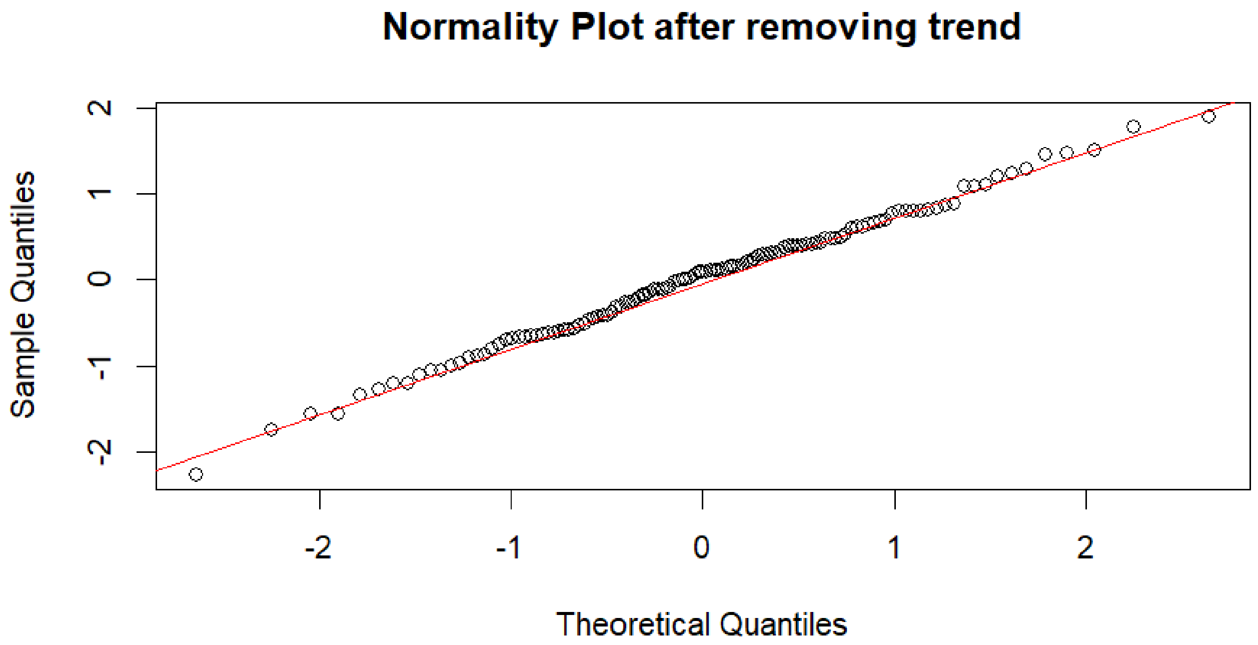

Figure 5: Normality Plot after removal of Trend

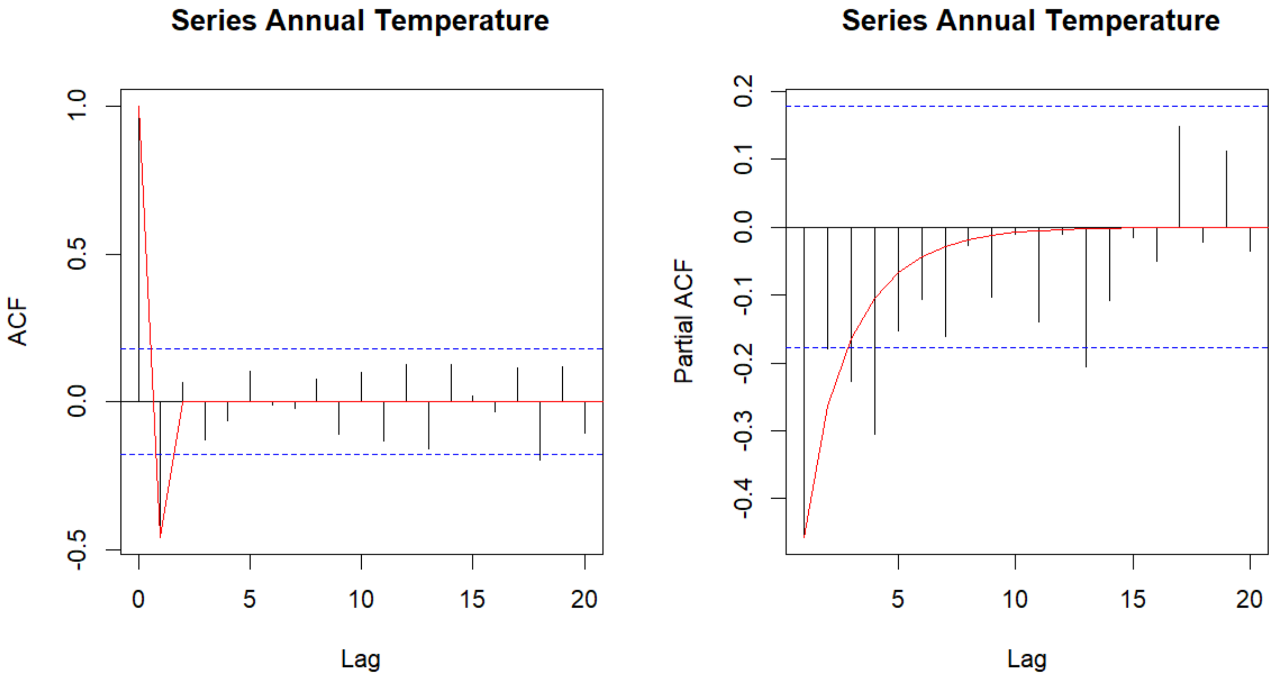

Figure 6: ACF and PACF Plots for differenced time series

Modelling: After achieving stationarity the next step in the model fitting is to plot ACF and PACF plots(Figure 6). From the plots, it is visible that ACF becomes zero after lag 1 and there are no significant peaks in the PACF plot after lag 1, which shows the process is MA(1). The red line in the ACF and PACF plot shows theoretical ACF and PACF for MA(1) process. Once the order of the process has been identified from the ACF and PACF plots the *arima* function from library *forecast* has been used to determine the coefficients θ_1 for equation 1.
The value of autocorrelation function ρ_1 can be read from the ACF plot which is equal to -0.457. Here, the negative sign indicates that there exists a negative correlation between X_t and X_{t-1}.

Mathematically the model,

$$X_t = Z_t + \theta * Z_{t-1} + \mu \qquad (\theta = -0.65, \mu = 0.007) \qquad (1)$$

Here, X_t represents observed value at time t, Z_t represents white noise term at time t. θ is the coefficient of the moving average (MA) part. From the equation, it is evident that the current value in the series is dependent on the error term from the current and previous term.

Now, we try to fit the MA(2) process using the same function *arima* from library *forecast*. The cofficients θ_1 and θ_2 obtained are -0.7726 and -0.0847 with a standard deviation of 0.0848 and 0.0807. Now, the ratio of coefficient and standard deviation is calculated for θ_1 and θ_2. It is found that the ratio is greater than 2 for θ_1 but less than 2 for θ_2. This forces us to choose MA(1)

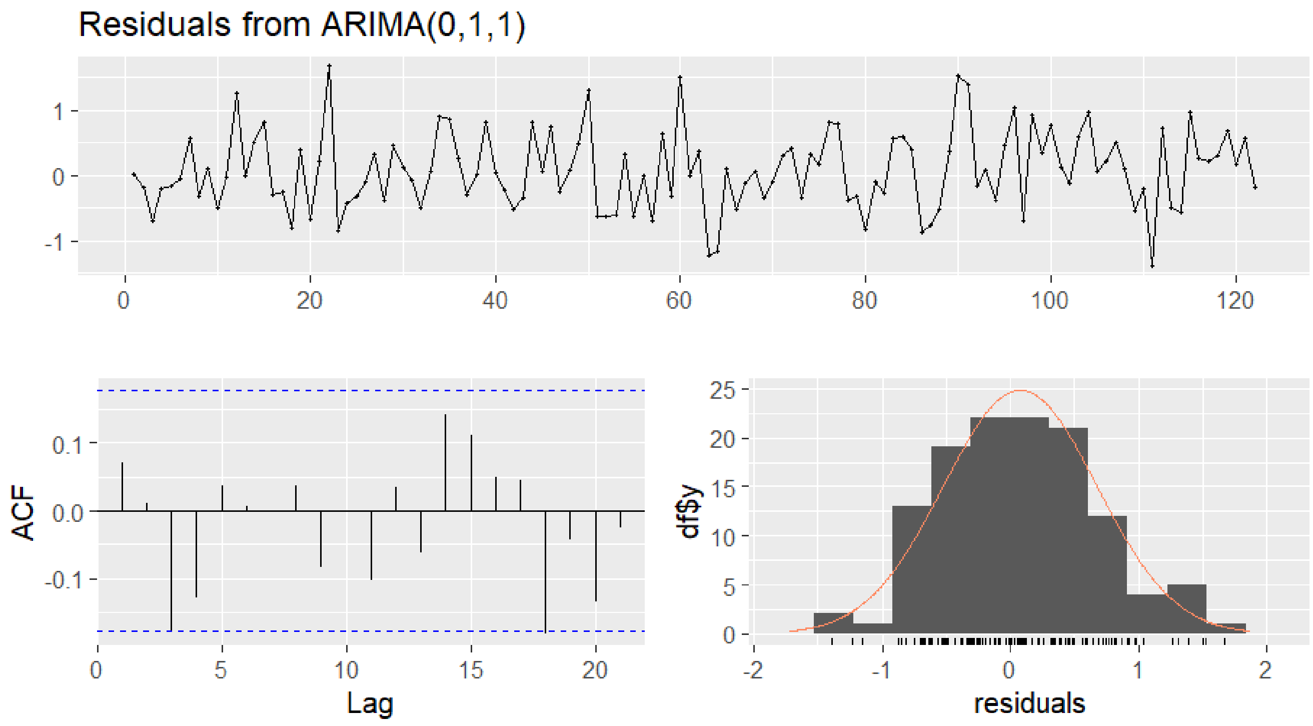

Figure 7: Residual Plots for ARIMA$(0, 1, 1)$

process over MA(2).Also, the Akaike Information Criteria(AIC) increased from 226.86 to 227.77 which shows higher order model is a poor choice.

To support the analysis further an autocorrelation plot for residuals is produced. It is evident from the plot (Figure 7) that residuals are independent and are below the significance level i.e. with no significant autocorrelation. From the correlogram of residuals, one can tell that all the residuals are less than $\pm 2/\sqrt{n}$ which proves the model fitted is adequate. Also, the residuals are normally distributed with mean zero making it practically a white noise. The series plot of the residuals looks stationary with constant variance.

Apart from the autocorrelation plot for the residual, the Ljung-Box test statistically checks if the errors (residuals) in the data are random and independent of each other.

1. **Null hypothesis** (H_0)**:** The residuals are independently and identically distributed (white noise). There's no serial correlation.

2. **Alternate hypothesis** (H_1)**:**The residuals exhibit serial correlation. There's a pattern in the errors.

The p-value obtained is greater than 0.05 which leads to failure to reject the null hypothesis. It indicates there's no strong evidence of serial correlation, supporting the model's adequacy. The final model proposed for Problem 1 is MA(1).

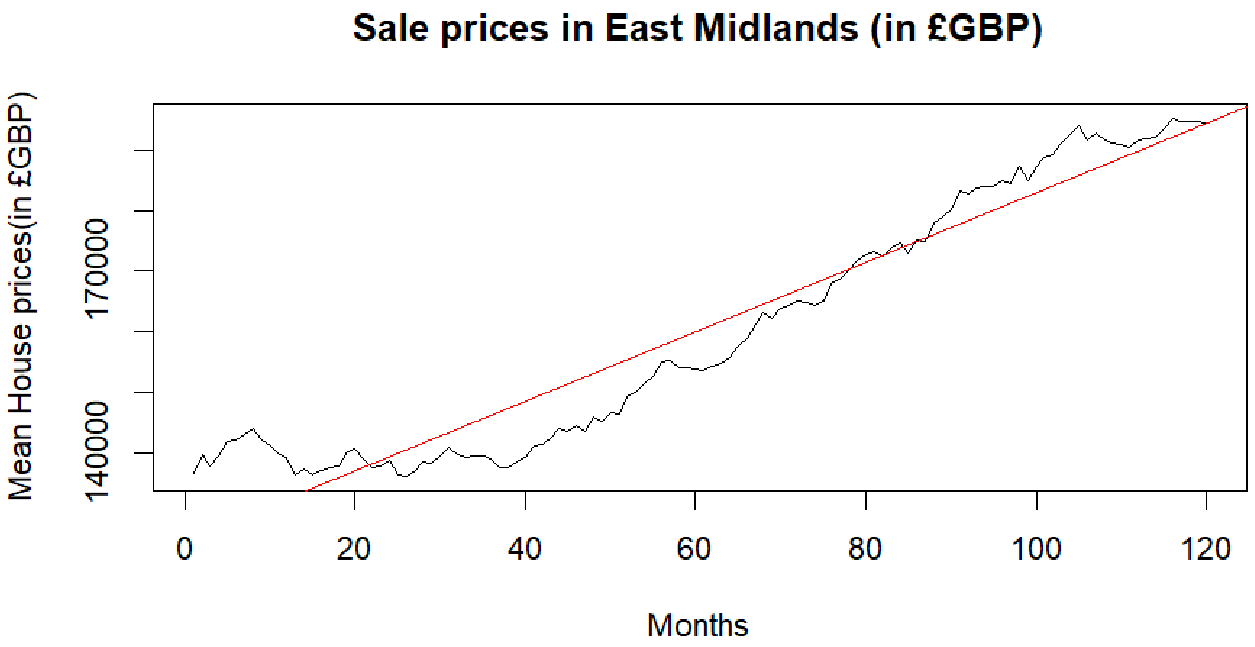

Figure 8: Original Time Series plot for Mean Sale Prices

Problem 2

The second dataset contains mean house sale prices in East Midlands (in £GBP), calculated monthly, from January 2010 to December 2019 and the job is to forecast future average house prices for the next six months. The methodology adopted is to first check for stationarity in time series both graphically and statistically using tests like ADF and KPSS. If stationarity is found in the time series then differencing will be done. To further our analysis ACF and PACF plots are generated for differenced series. These plots are used to ascertain the order of the process and also to check the presence of any seasonality. The Kruskal-Wallis test is used to check for seasonality. The final stage is to fit the model through iteration to get the best fit. A correlogram is generated for residuals to check whether residuals are correlated also Ljung-Box test will statistically prove that residuals of the best model behave like white noise.

From the plot(Figure 8) one can easily tell that there is an upward trend in the Mean Housing Price for East Midlands. The housing price in East Midland varies from £ 195345 to £ 136102 with the highest for August 2019 and the lowest for February 2012. The red line in the plot shows the linear trend for the mean housing prices. It can be inferred from the plot that the time series is not stationary. A time series is considered stationary if its statistical properties (mean, variance, and autocorrelation) remain constant over time i.e. there is no upward or downward trend, and the variability around the mean stays consistent. Non-stationary data can lead to misleading results in many time series analysis techniques. There are many techniques to make time series stationary these include differencing, detrending, and filtering. To achieve stationarity for this

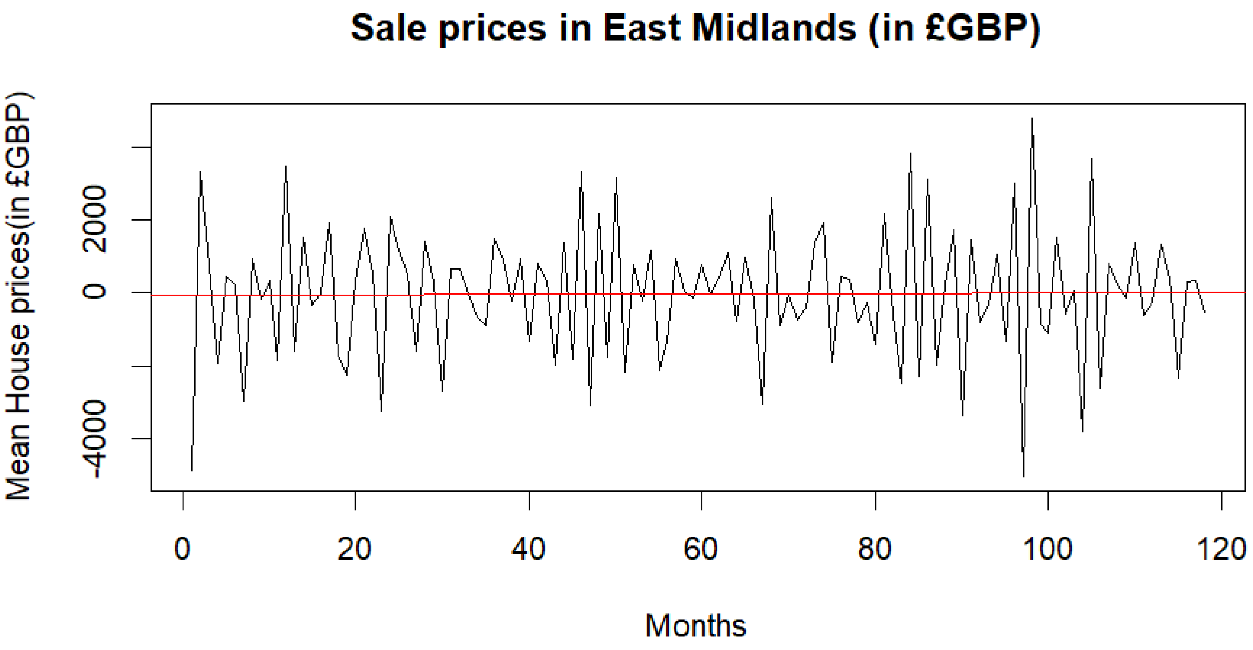

Figure 9: Time Series plot for Mean Sale Prices after differencing

problem, differencing of second order is used for this problem. This can also be verified through the Augmented Dickey-Fuller Test:

1. **Null Hypothesis** H_0**:** Series is non-stationary, or series has a unit root.

2. **Alternate Hypothesis** H_1**:** Series is stationary, or series has no unit root.

Before differencing the p-value for the test was found to be 0.167 which forced us to reject the null hypothesis. However, after second-order differencing, it was found that the p-value decreased from 0.167 to 0.01 making the time series stationary. This is also visible from the plot (Figure 9 for the differenced series.

Before moving further let's do the exploratory data analysis(EDA) for the time series. From the histogram shown in(Figure 10) and the normality plot(Figure 10) one can infer that data is normally distributed for the differenced time series which is one of the assumptions for fitting and ARIMA model.

Modelling: After achieving stationarity the next step in the model fitting is to plot ACF and PACF plots(Figure 16). From the ACF plot, it can be seen that a.c.f cuts off after lag 1 which shows the presence of MA(1) element in the time series. The red line in the ACF and PACF plot shows theoretical ACF and PACF. Also from the same plot we can see that there is a significant peak after lag 12 which shows the presence of seasonality in the time series. To prove the presence of seasonality statistically, the Kruskal-Wallis test is used.

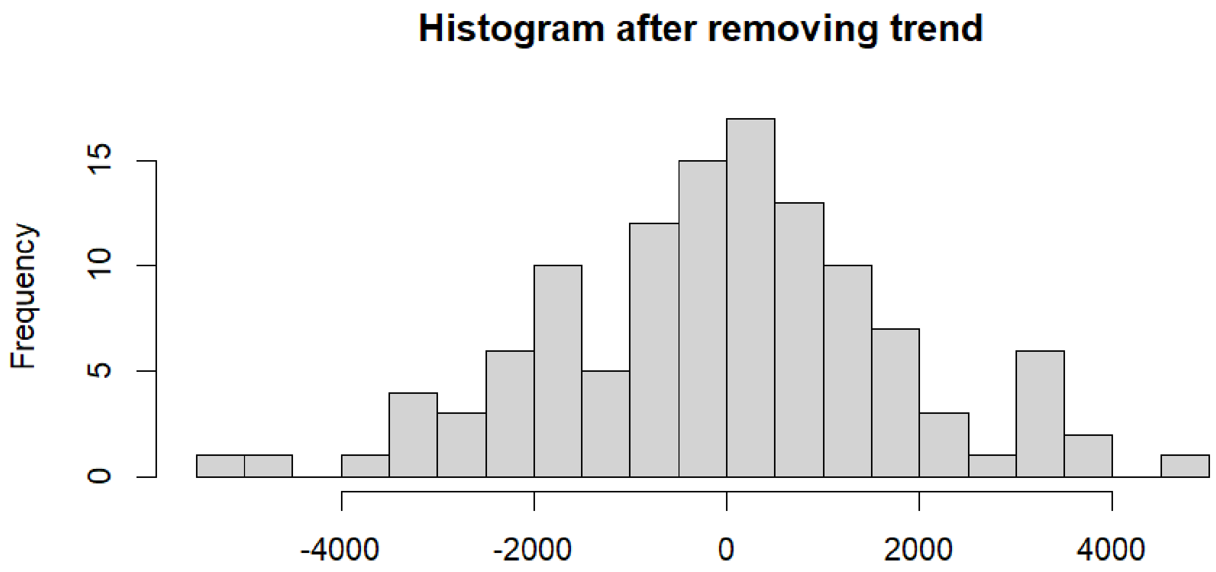

Figure 10: Distribution of differenced series

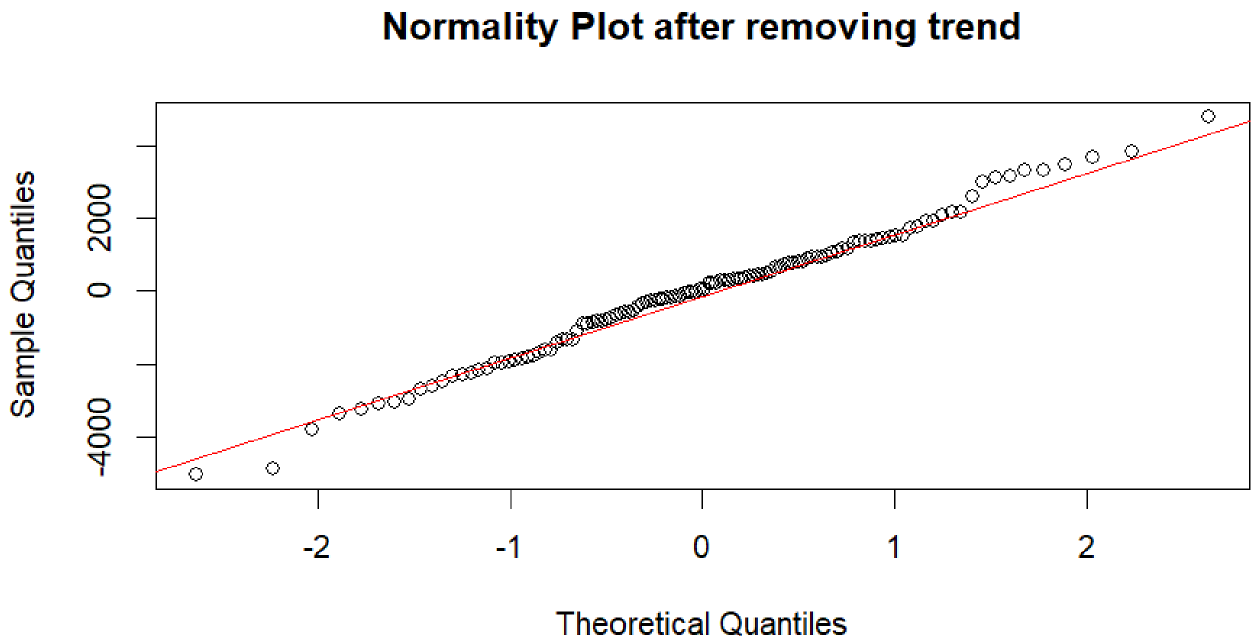

Figure 11: Normality plot for differenced series

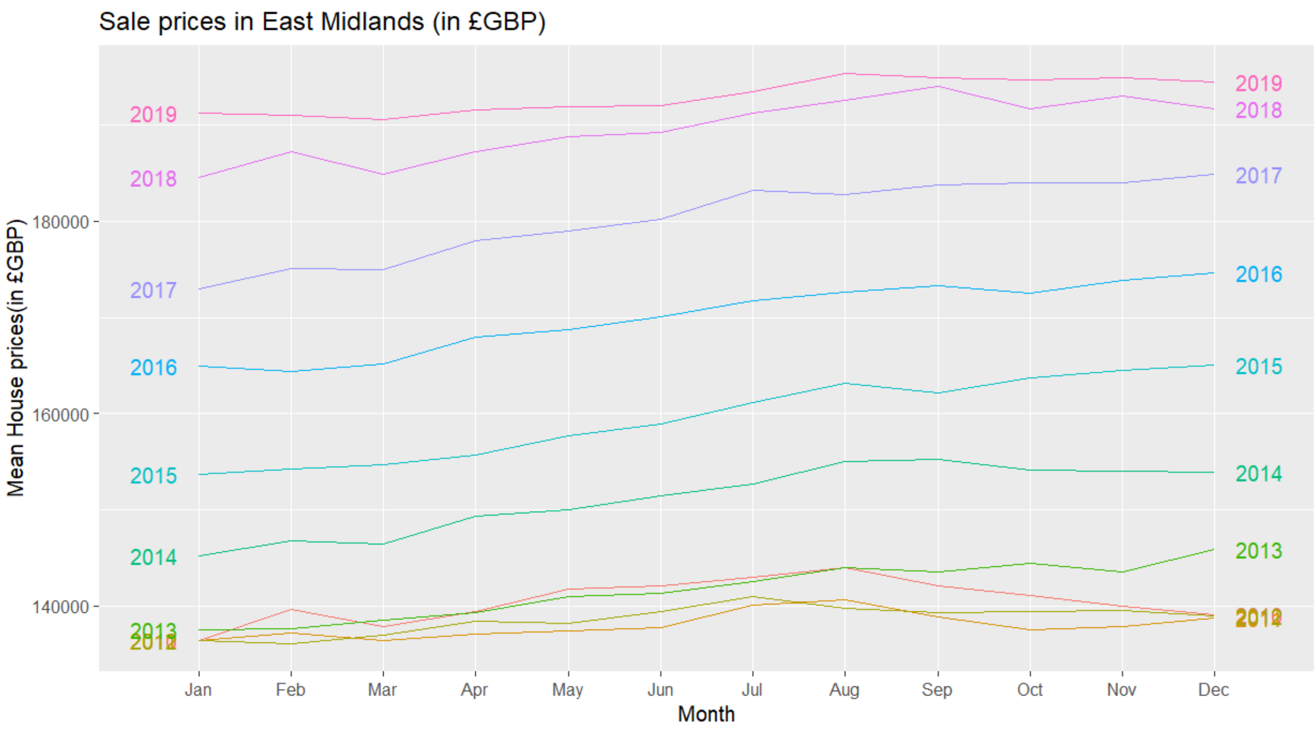

Figure 12: Seasonal Plot in Cartesian Coordinates

1. **Null Hypothesis H_0:** There is no recurring pattern (seasonality) affecting the median values across different periods. Refer to the plot(Figure 14).
2. **Alternate Hypothesis H_1:** There is a recurring pattern (seasonality) affecting the median values across different periods

It was found that after conducting the test p-value was below 0.05 which again forced us to reject the null hypothesis. Hence there is a presence of seasonality in the time series.

Seasonality can be visualized through seasonal plots available in library *ggplot2*. It can be seen in the plot(Figure 12) that every year the price increases steadily from January to December and then drops down again in January. This is also shown in the polar plot(Figure 13). The waviness present in the polar plot shows that there is seasonality in the time series.

Correlation plot(Figure 15) for different lags are plotted using *gglagplot*. It can be seen from the plot that correlation is strongest for lag 1. This is also the source of motivation for choosing MA(1) along with ACF plot shown earlier.

Now the basic model starts with MA(1) with seasonality of 12 i.e. SARIMA$(0, 2, 1)(0, 1, 1)_{12}$. At every iteration, the ratio of coefficient and standard deviation is calculated to check whether adding additional order to the time series is significant or not. The Akaike Information Criteria(AIC) obtained for the model is 1803.2 which is high compared to the final model proposed. For SARIMA$(0, 2, 1)(0, 1, 1)_{12}$ the MA coefficient were found out to be $ma1 = -0.9207$, $sma1 = -0.8650$ with standard deviation of 0.0385, 0.1738 respectively. Since the ratio of coefficients and standard deviations is greater than 2 making it statistically significant. As seen

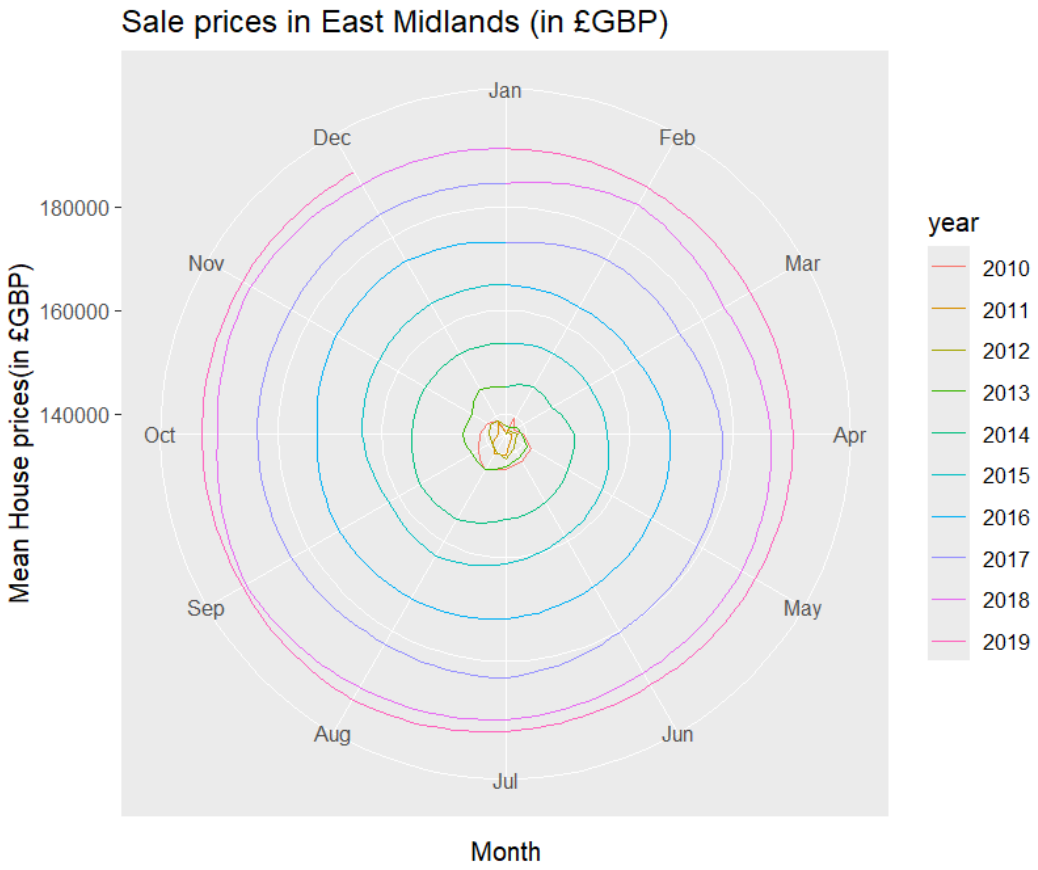

Figure 13: Seasonal Plot in Polar Coordinates

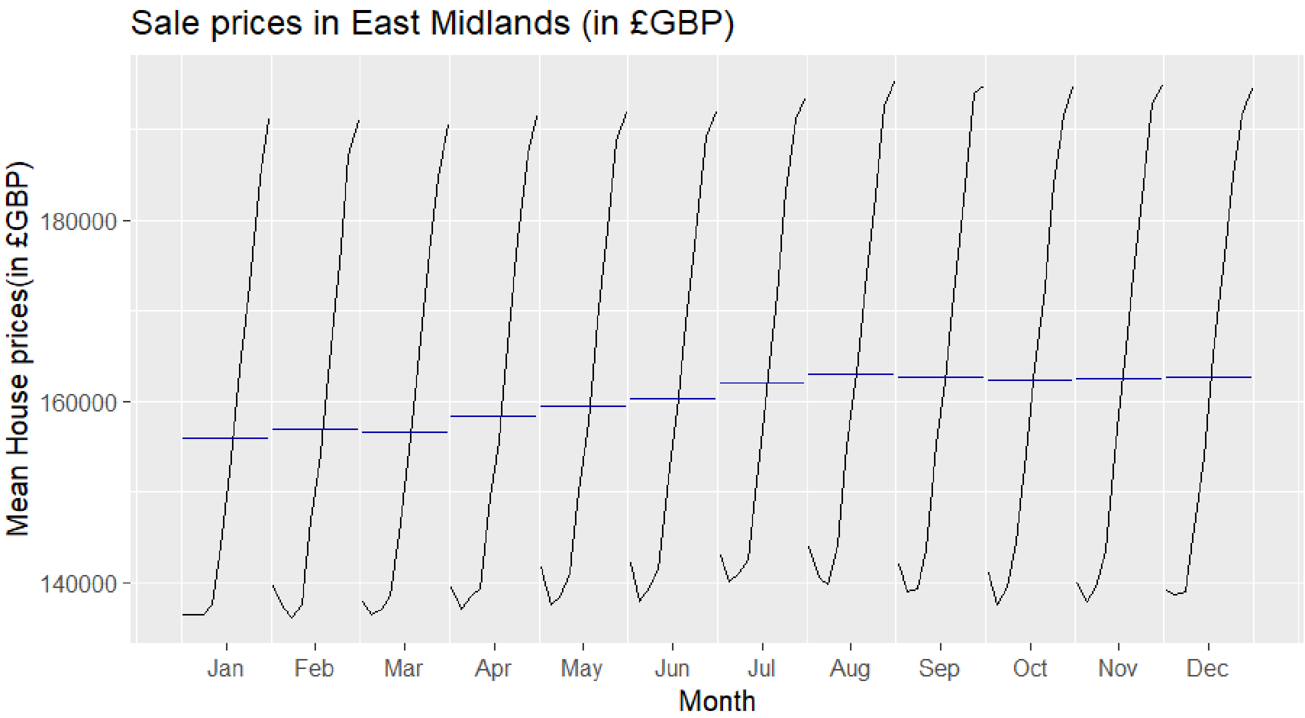

Figure 14: Sub-series Plot for time series

from ACF and PACF plots(Figure: 16) the value of ACF fluctuates from positive to negative it is necessary to add an autoregressive element in the next iteration. Now, the model becomes SARIMA$(1, 2, 1)(0, 1, 1)_{12}$ which reduces AIC to 1788.19. For SARIMA$(1, 2, 1)(0, 1, 1)_{12}$ the AR and MA coefficients were found out to be $ar1 = -0.4319$ $ma1 = -0.8445$, $sma1 = -0.8484$ with standard deviation of 0.0966, 0.0614, 0.1574 respectively. Since the ratio of coefficients and standard deviations is greater than 2 making it statistically significant. However, even this model obtained is not up to the mark. Next we move to SARIMA$(2, 2, 1)(0, 1, 1)_{12}$ which further reduces AIC to 1787.75.

For SARIMA$(2, 2, 1)(0, 1, 1)_{12}$ both AR and MA coefficient were found out to be $ar1 = -0.5618$, $ar2 = -0.2029$, $ma1 = -0.7537$ with standard deviation of 0.1266, 0.1018 and 0.1056 respectively. This makes all three significant for the time series. To check whether this model is optimum, it was tried to fit the data to a higher order if the coefficients are found insignificant then the higher order model will be rejected. The next model tried was SARIMA$(3, 2, 2)(0, 1, 1)_{12}$ the cofficients were found insignificant for $ar3$ and $ma2$. Now that we have stabilized the non-seasonal part of the time series which is visible from ACF and PACF plots, the next step is to find the optimum order of AR and MA terms for the seasonal part. For this AIC score is used as a criterion to find the best model. It is found that the least AIC score is obtained for (0,1,1) for the seasonal part. The AIC score increased from 1787.75 for SARIMA$(2, 2, 1)(0, 1, 1)_{12}$ to 1789.74 for SARIMA$(2, 2, 1)(0, 1, 2)_{12}$. So the final model selected for forecasting is SARIMA$(2, 2, 1)(0, 1, 1)_{12}$.

Mathematically the model can be written as,

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)^2(1 - B^{12})X_t = (1 + \theta_1 B)(1 + \Theta_1 B^{12})Z_t \tag{2}$$

Here, X_t represents observed value at time t, Z_t represents white noise term at time t. θ and ϕ are coefficients of the auto-regressive(AR) and moving average (MA) part for the non-seasonal part of the time series respectively.Φ and Θ are coefficients of the auto-regressive(AR) and moving average (MA) part for the seasonal part of the time series respectively.

The $SARIMA(2, 2, 1)(0, 1, 1)_{12}$ is used to forecast the price of the houses in East Midland for the next six months the black line in the plot(Figure 18) shows the forecasted values while the blue zone around the predictions shows the confidence interval.

The forecast trend shows an increase in the mean House prices for East Midland in the next six months. The mean forecast for the next six months is 193483.8, 194208.0, 193642.2, 195286.5, 195565.1, and 196707.4 respectively. The Root Mean Square Error and Mean Percentage error are 897.02 and 0.036 respectively which is quite low compared to other models.

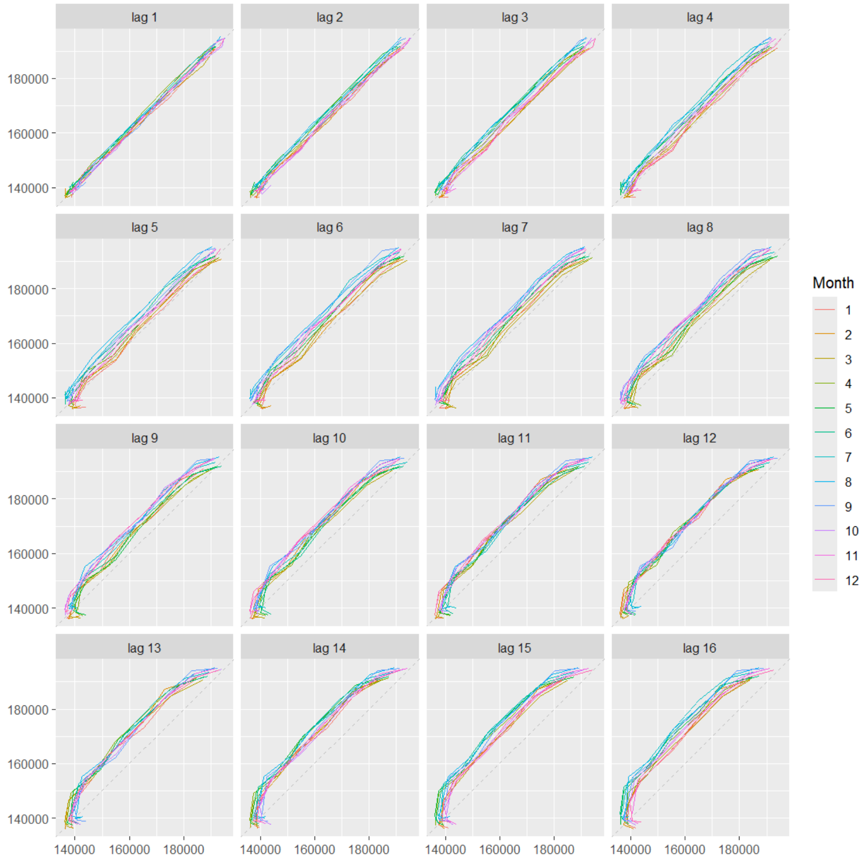

Figure 15: Correlation Plot for different Lags

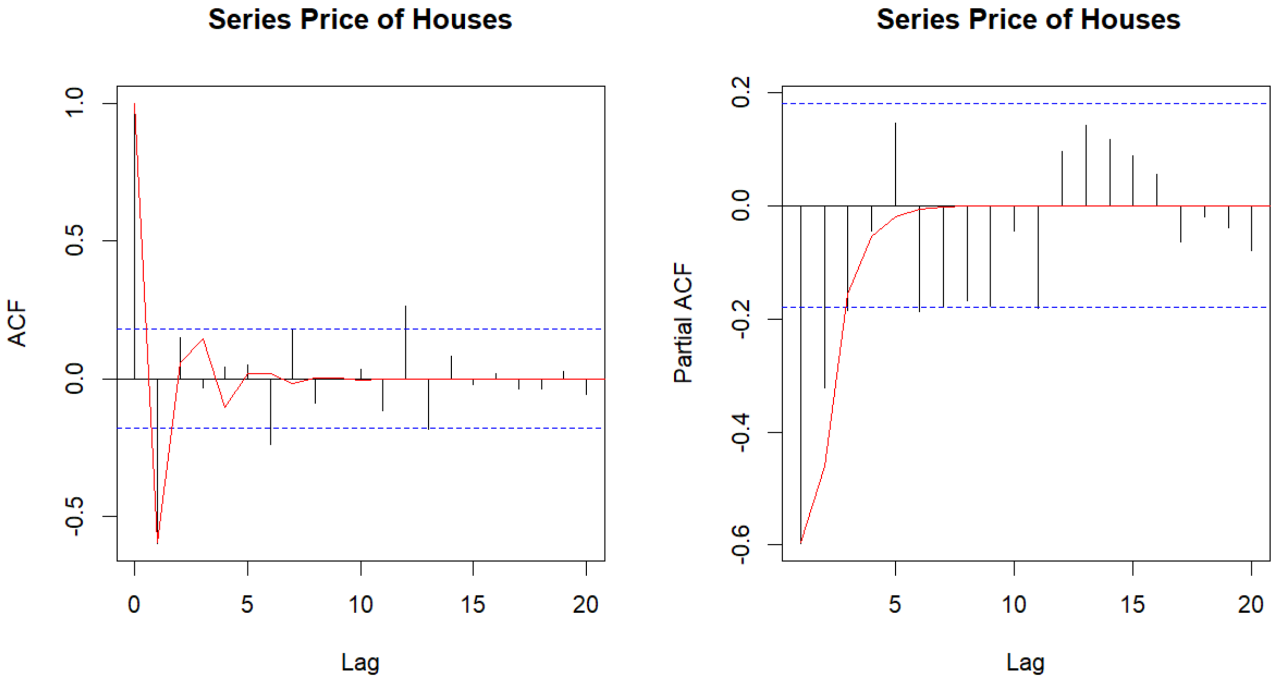

Figure 16: ACF and PACF plots for differenced timeseries

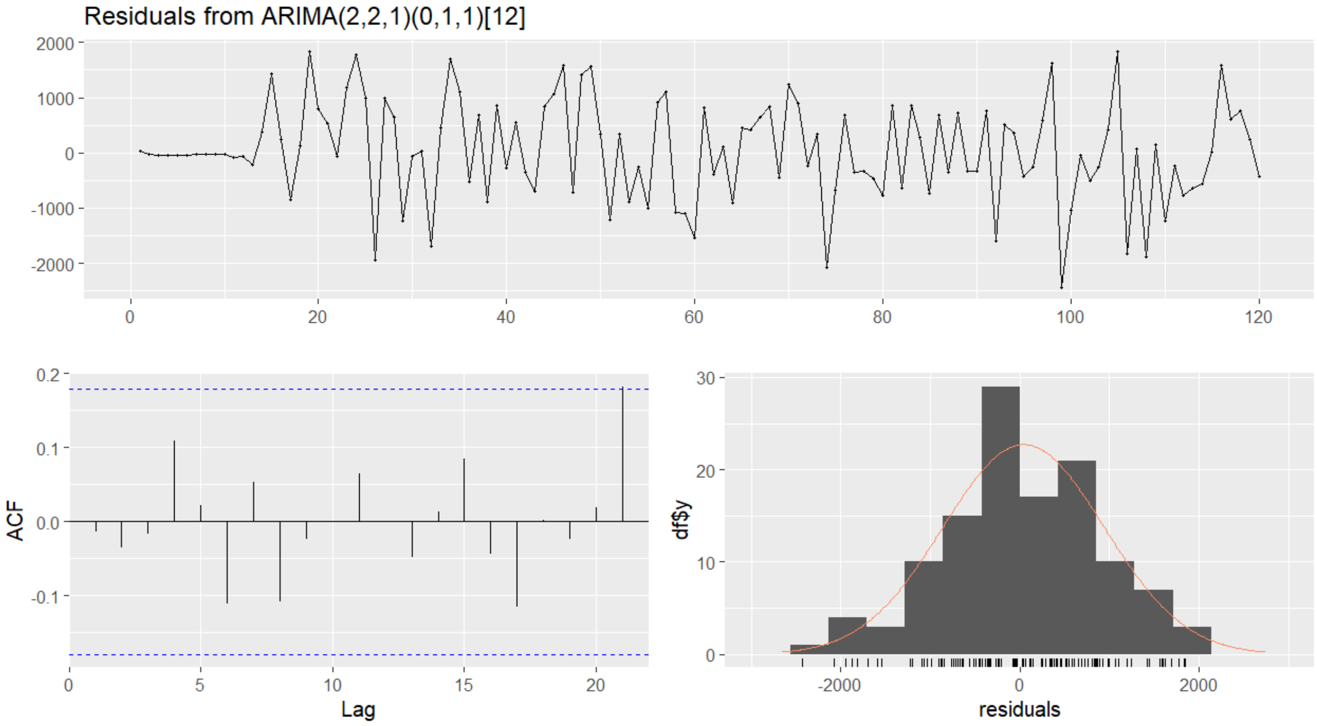

Figure 17: Residual Plots for SARIMA$(2, 2, 1)(0, 1, 1)_{12}$

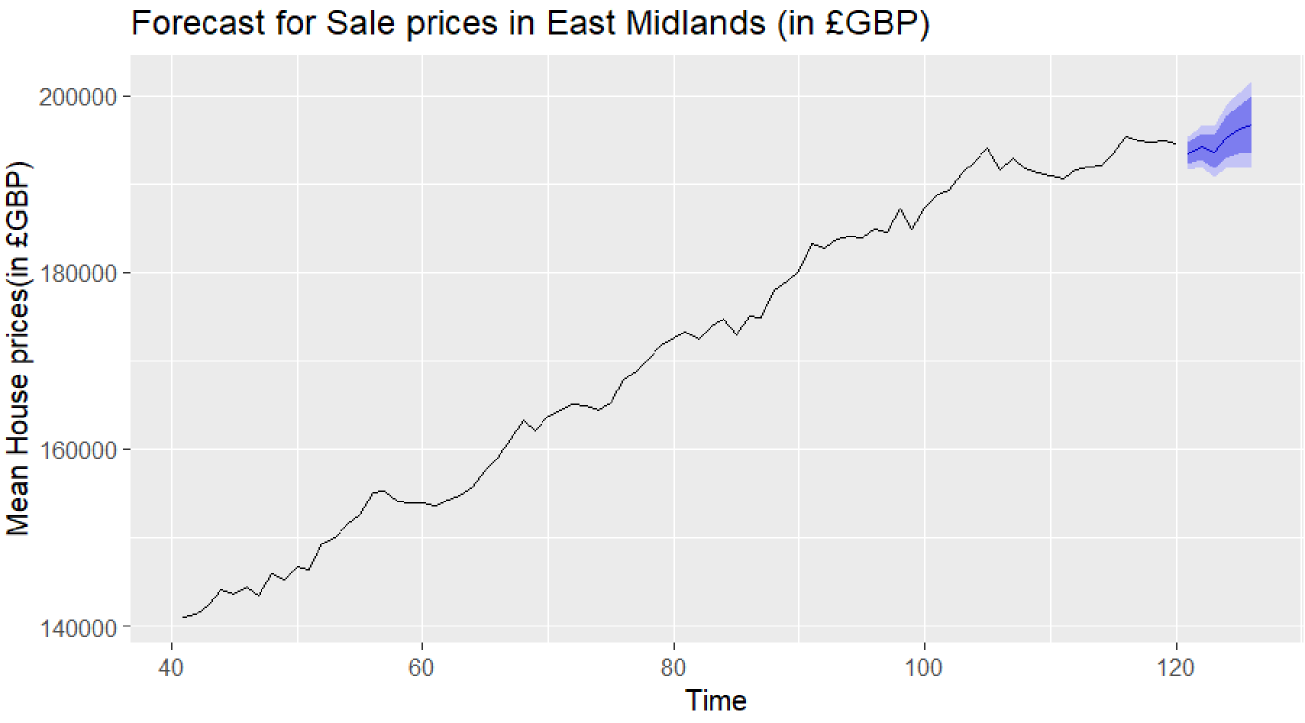

Figure 18: Forecast for next six months

```
#-----------------------------------PROBLEM 1------------------------------------------#

# Setting the working directory and importing libraries
setwd('C:\\Users\\raona\\OneDrive\\Documents\\Time Series')
install.packages('tseries')
install.packages('forecast')
install.packages('urca')
library('forecast')
library('tseries')
library('urca')

# Reading the csv file
cet_temp <- read.csv('cet_temp.csv')

# Temperature column as time series data
z_1 <- ts(cet_temp$avg_annual_temp_C,start=1900,frequency=1)

# Plotting temperature Vs time
ts.plot(cet_temp$avg_annual_temp_C,ylab= "Annual Temperature", xlab='Years', main="Annual
mean temperature in °C \n for the years 1900 to 2021")
abline(reg=lm(data=cet_temp,avg_annual_temp_C~time(avg_annual_temp_C)),col='red')

# Time series is not stationary need to take difference
y_1 <- diff(cet_temp$avg_annual_temp_C,differences = 1)
ts.plot(y_1,ylab= "Annual Temperature", xlab='Years', main='Annual mean temperature in °C
\n for the years 1900 to 2021')
abline(reg=lm(data=cet_temp,y_1~time(y_1)),col='red')

# Visualization after detrending
hist(y_1,breaks =20,main = "Histogram after removing trend",xlab = "")
boxplot(y_1,horizontal= TRUE,main = "Box plot after removing trend",xlab = "")
qqnorm(y_1,main = "Normality Plot after removing trend")
qqline(y_1,col = "red")

# Augmented Dickey-Fuller Test to check whether time series is Stationary
adf.test(y_1)

# Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test
summary(ur.kpss(y_1))

# Theoretical ACF and PACF
t_acf<-ARMAacf(ar=c(),ma=c(-0.65),lag.max=121)
t_pacf<-ARMAacf(ar=c(),ma=c(-0.65),lag.max=121,pacf=TRUE)

# Plot for sample ACF and PACF
par(mfrow=c(1,2))
acf(y_1, main = 'Series Annual Temperature')
lines(c(0:121),t_acf,col="red")
pacf(y_1, main=' Series Annual Temperature')
lines(c(1:121),t_pacf,col="red")

# Fitting an ARMA(0,1,1)
model<-arima(cet_temp$avg_annual_temp_C,order=c(0,1,1),method="ML")

# Residual plot for ARIMA(0,1,1)
res<- residuals(model)
checkresiduals(model)

# Performing Ljung Box test on residuals
Box.test(res,type=c("Ljung-Box"))
```

```
#----------------------------------PROBLEM 2----------------------------------------#

# Setting the working directory and importing libraries
setwd('C:\\Users\\raona\\OneDrive\\Documents\\Time Series')
install.packages('tseries')
install.packages('forecast')
install.packages('urca')
install.packages('seastests')
library('forecast')
library('tseries')
library('ggplot2')
library('urca')
library('seastests')

# Reading the csv file
em_house_prices <- read.csv('em_house_prices.csv')

# Housing price column as time series data
z<-ts(em_house_prices$average_price_gbp,start=2010,frequency=12)

#Augmented Dicky Fuller test for stationarity
adf.test(z)

#Krsukall Wallis test for Seasonality
isSeasonal(z, test = "kw", freq = 12)
kw(z, freq = 12, diff = T, residuals = F, autoarima = T)

# Plotting Housing Price Vs time
ts.plot(em_house_prices$average_price_gbp,ylab= "Mean House prices(in £GBP)",
xlab='Months', main="Sale prices in East Midlands (in £GBP)")
abline(reg=lm(data=em_house_prices,average_price_gbp~time(average_price_gbp)),col='red')

# Time series is not stationary need to take difference of order 2
y <- diff(em_house_prices$average_price_gbp,differences=2)
ts.plot(y,ylab= "Mean House prices(in £GBP)", xlab='Months', main="Sale prices in East
Midlands (in £GBP)")
abline(reg=lm(data=em_house_prices,y~time(y)),col='red')

# Visualization after detrending
hist(y,breaks =20,main = "Histogram after removing trend",xlab = "")
boxplot(y,horizontal= TRUE,main = "Box plot after removing trend",xlab = "")
qqnorm(y,main = "Normality Plot after removing trend")
qqline(y,col = "red")

# Seasonal Plots for checking Seasonality
ggseasonplot(z, year.labels=TRUE, year.labels.left=TRUE,main="Sale prices in East Midlands
(in £GBP)",ylab= "Mean House prices(in £GBP)")
ggseasonplot(z, polar=TRUE,main="Sale prices in East Midlands (in £GBP)",ylab= "Mean House
prices(in £GBP)")
ggsubseriesplot(z)
gglagplot(z)

# Augmented Dickey-Fuller Test to check whether time series is Stationary
adf.test(y)

# Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test after differencing
summary(ur.kpss(y))

# Theoretical ACF and PACF
t_acf<-ARMAacf(ar=c(-0.6,-0.3),ma=c(-0.35),lag.max=120)
t_pacf<-ARMAacf(ar=c(-0.6,-0.3),ma=c(-0.35),lag.max=120,pacf=TRUE)

# Plot for sample ACF and PACF
par(mfrow=c(1,2))
```

```
acf(y, main = 'Series Price of Houses')
lines(c(0:120),t_acf,col="red")
pacf(y, main=' Series Price of Houses')
lines(c(1:120),t_pacf,col="red")

# Fitting an ARIMA(2,2,1)(0,1,1)[12]
model1<-arima(em_house_prices$average_price_gbp,order=c(3,2,2), seasonal
=list(order=c(0,1,2),period=12),method="ML")
summary(model1)

# Residual Plots for the above model
res<- residuals(model1)
checkresiduals(model1)

#Forecast for next six  months
model1 %>% forecast(h=6) %>% autoplot(include=80,ylab= "Mean House prices(in
£GBP)",main="Forecast for Sale prices in East Midlands (in £GBP)",transform.pars=TRUE)
model1 %>% forecast(h=6)

#Ljung Box test on residuals
Box.test(res,type = "Ljung-Box")
```