UNITED KINGDOM · CHINA · MALAYSIA

MATH4022 - TIME SERIES AND FORECASTING

Coursework Project Report

SCHOOL OF MATHEMATICAL SCIENCES
SPRING SEMESTER 2023-2024

Name: **OMKAR MAINKAR**
Student ID: **20550557**

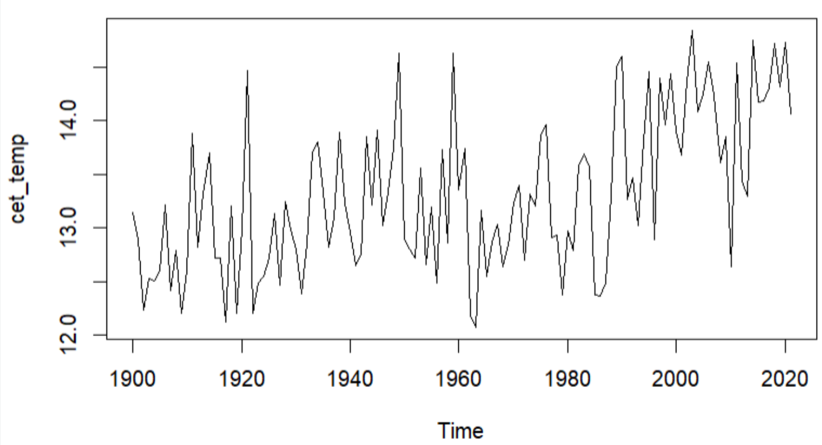
Statistical Analysis of Annual Mean Temperature Data for the Midlands Region of England: 1900-2021

Introduction:

Climate data analysis is critical in comprehending long-term trends, variability, and the potential consequences of climate change. In this research, we analyse the annual mean temperature data for the Midlands region of England from 1900 to 2021. These temperature records, obtained from the UK Meteorological Office Hadley Climate Centre, lend insightful details about the region's climatic conditions over the last century. The foremost objective of this statistical study is to fit a suitable time series model to the dataset, which will allow us to accurately depict and forecast on an annual basis mean temperature. Throughout this investigation, we will explain our process for model identification, selection, and validation.

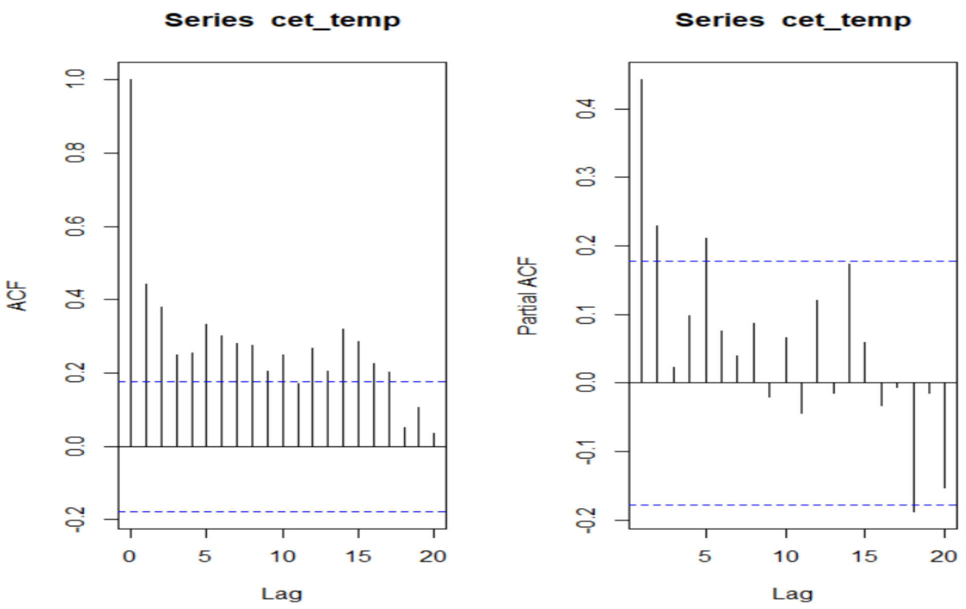
Model Fitting and Interpretation:

In the initial stage of our analysis, we used R's read.csv() method to import the dataset 'cet_temp.csv'. To acquire a visual grasp of the temporal trends in the data, we created a time series plot with the ts.plot() function. The following visualization presents a complete overview of yearly mean temperatures throughout time, allowing us to discover any notable trends, seasonality, or abnormalities in the dataset.



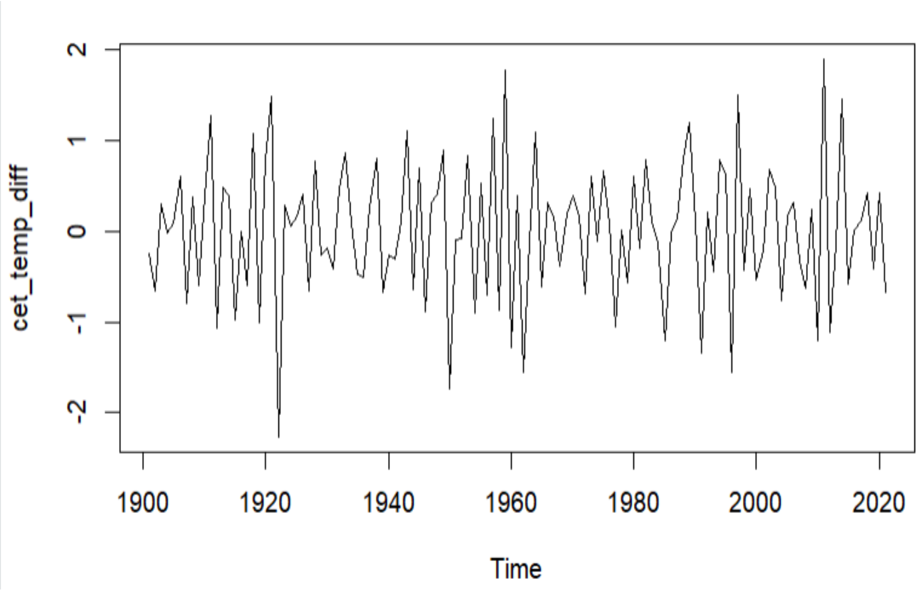
Examining the time series plot revealed that the underlying process was non-stationary. Stationarity is a fundamental assumption in time series analysis, implying that the statistical features of the data, such as mean and variance, remain constant across time.
The series' mean increased between 2001 and 2021 than it was during earlier periods (e.g., 1900-1930), as shown in the time plot.

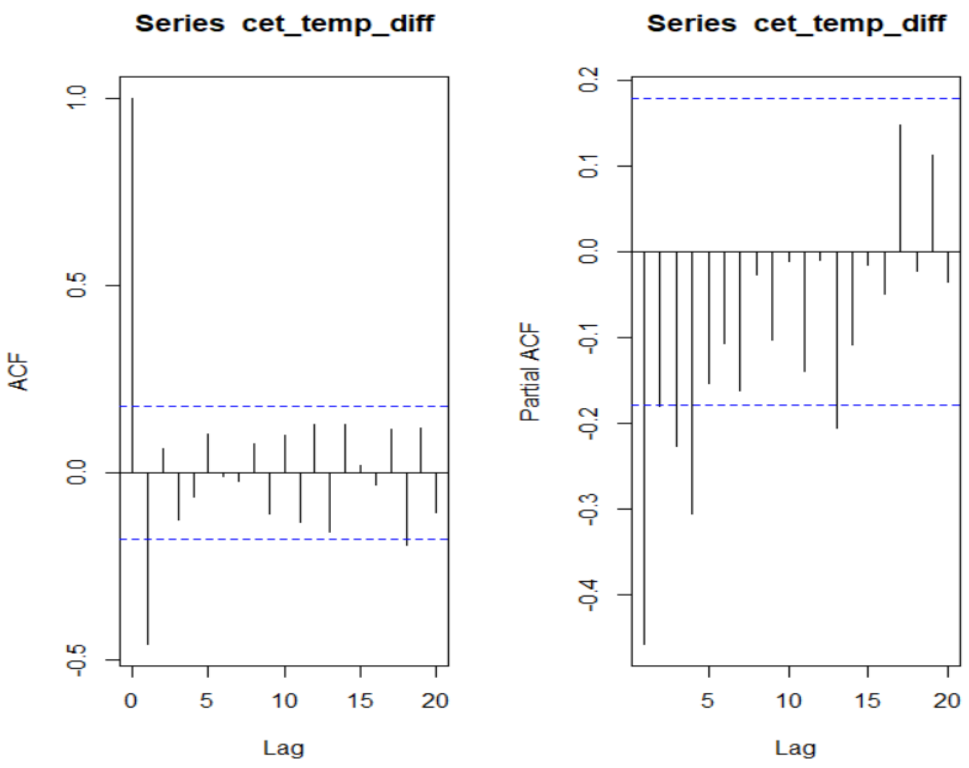
Following the time series data was initially visualized, it was examined further using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.



Furthermore, the sample ACF values weren't declining rapidly as the lag increased indicating the presence of autocorrelation in the data. For example, the ACF peak height at lag 14 is similar to that at lag 5) so it is not certain that the series is stationary. Similarly, the PACF figure lacked a distinct cutoff after lag 1, demonstrating residual connections between observations. The sample PACF plot provides less information about whether or not the series is stationary compared to the time plot and sample ACF plot.

After noticing the non-stationarity in the original time series data, the first difference of the series, labelled as 'cet_temp_diff', was calculated. This modification is frequently used for acquiring stationarity through eliminating trend and seasonality components from the data.
A subsequent analysis of the ACF and PACF plots of the differenced data demonstrated additional insight into its autocorrelation pattern

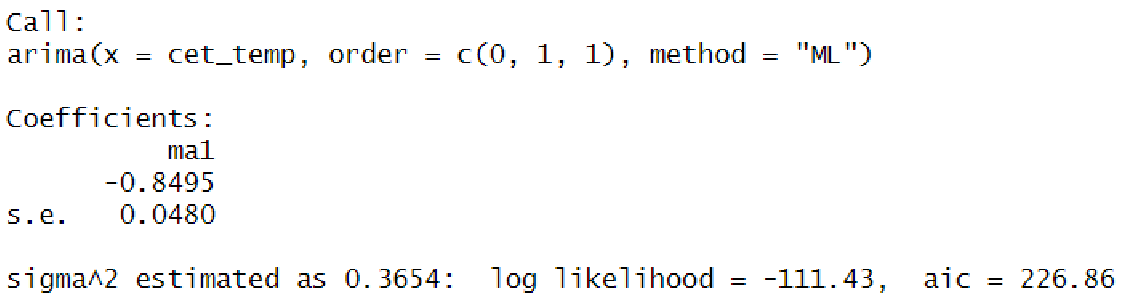




The resulting time series plot of the differenced data revealed a rather consistent pattern, indicating a potential improvement in stationarity over the original series. The figure showed a constant mean (zero owing to differencing) and appeared to have consistent variability across time. As the lag rises, the sample ACF rapidly decreases to zero (with a few spikes near ±2/√n, but no consistent departure from zero). The sample PACF drops fast to zero as the lag rises.

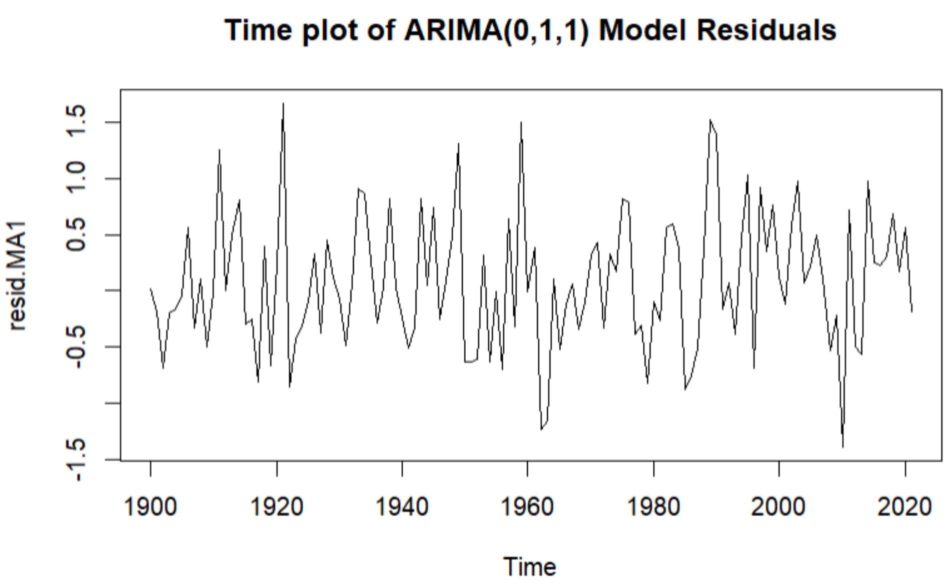
These results indicated a (weakly) stationary process, which was ideal for time series modelling. Furthermore, the rapid decay in both the ACF and PACF plots suggested the possibility of a Moving Average (MA) process in the data, in which the present observation is linearly reliant on previous white noise terms. The plot of the sample ACF, which cuts off at zero and shows a downward spike at lag 1, encouraged the possibility of an MA (1) component. Furthermore, the absence of a prominent peak in the sample PACF revealed that there is no major AR component in the process.

An ARIMA (0,1,1) model was implemented to the differenced time series data, designated as 'cet_temp', signifying a first-order moving average with first differencing. The resulting ARIMA (0,1,1) model was assessed and the summary results were produced. This output comprises data pertaining to the model's estimated coefficients, standard errors, and other necessary statistics.

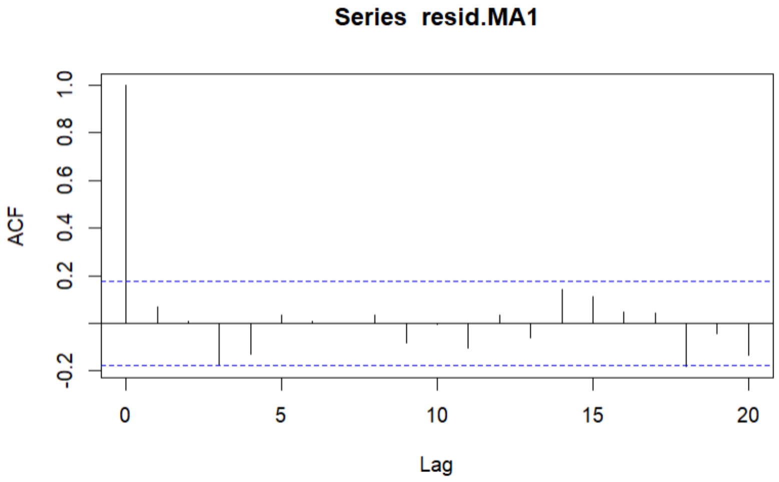


```
Call:
arima(x = cet_temp, order = c(0, 1, 1), method = "ML")

Coefficients:
          ma1
      -0.8495
s.e.   0.0480

sigma^2 estimated as 0.3654:  log likelihood = -111.43,  aic = 226.86
```

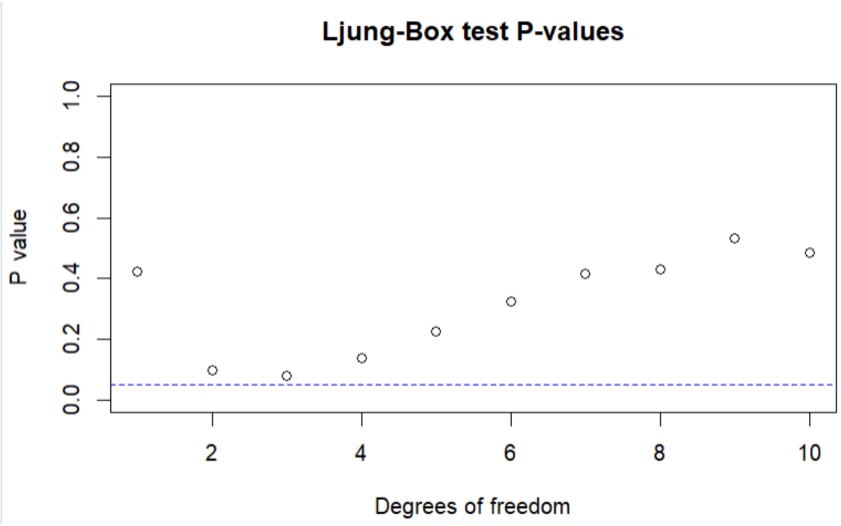
The residuals of the ARIMA(0,1,1) model, indicated as 'resid.MA1', were subsequently extracted and plotted to determine their behaviour. The temporal plot of the residuals resembled that of white noise, demonstrating randomness and no noticeable pattern.



The autocorrelation function (ACF) plot of the residuals was also analysed. The sample ACF was near to zero for all lags beyond lag 0, indicating that the residuals are independent.

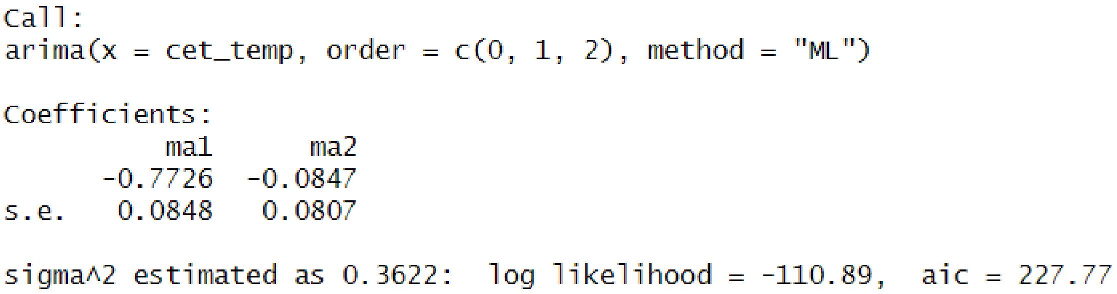


In addition, a plot of the Ljung-Box test p-values against degrees of freedom for ARIMA (0,1,1) was produced to depict the findings. A blue dashed line at the 5% significance level was added to facilitate comprehension.



All P-values are larger than the threshold 0.05, with the exception of a few that are somewhat modest but not less than 0.05. Thus it implies that ARIMA(0,1,1) is a good fit for the data.
After fitting the ARIMA(0,1,1) model and determining that it serves as a good fit to the data, we investigate whether there could be improvement in model fit through introducing another moving average (MA) component. To accomplish so, we propose fitting an ARIMA(0,1,2) model to the differenced time series data. We will speculate on the parameters of the ARIMA(0,1,2) model while comparing its results to the ARIMA(0,1,1) model. More precisely, we aim to determine if the new MA term significantly improves the model fit.

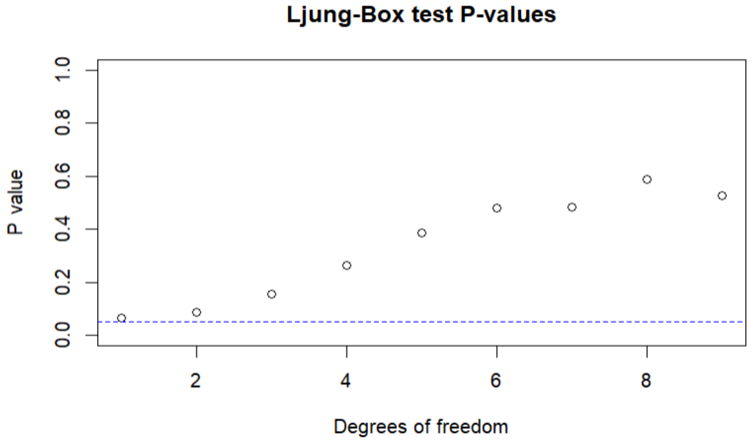
Model output for ARIMA(0,1,2) is shown below



```
Call:
arima(x = cet_temp, order = c(0, 1, 2), method = "ML")

Coefficients:
          ma1      ma2
      -0.7726  -0.0847
s.e.   0.0848   0.0807

sigma^2 estimated as 0.3622:  log likelihood = -110.89,  aic = 227.77
```

A number of tests were conducted to determine which of the ARIMA(0,1,1) and ARIMA(0,1,2) models better suited the data.
First, we calculated the Akaike Information Criterion (AIC) values for both models. The AIC represents the trade-off between model complexity and goodness of fit, with lower values suggesting a more favorable fit. In this scenario, the ARIMA(0,1,1) model had a lower AIC (226.86) than the ARIMA(0,1,2) model, which had an AIC of 227.77. This indicates a preference for the simpler ARIMA(0,1,1) model.
In addition, the test statistic for a test of the hypotheses: H0 : θ2 = 0 versus H1 : θ2 != 0 is |-0.0847/0.0807| = 1.04 which is not greater than 2. Hence we failed to reject H0 and conclude that the MA(2) term should not be included in the model.

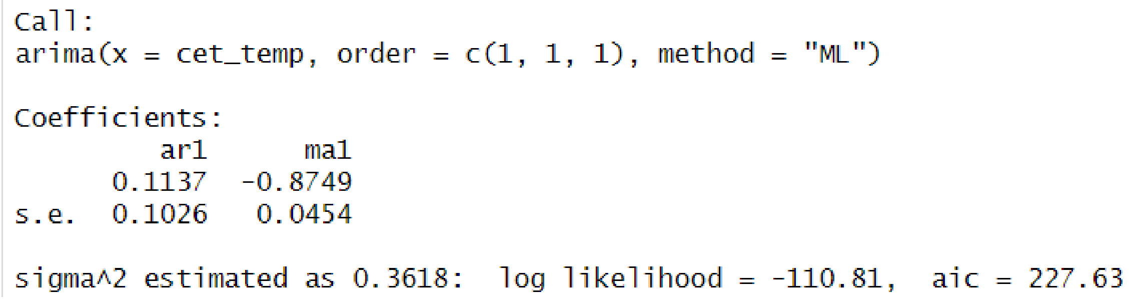
As a final check , we'll examine the Ljung-Box test P-values with respect to the model residuals.



After thoroughly examining the coefficients and accompanying p-values in both the ARIMA(0,1,1) and ARIMA(0,1,2) models, we discovered that the ARIMA(0,1,1) model had slightly improved p-values overall.

Despite the AIC values differed only slightly between the two models, the hypothesis test results and p-value distribution confirmed the choice for the simpler ARIMA(0,1,1).

As a final evaluation, in comparison to the previously evaluated ARIMA(0,1,1) model, the ARIMA(1,1,1) model has a higher Akaike Information Criterion (AIC) score (227.63) than the ARIMA(0,1,1) model (226.86). Furthermore, with hypothesis H0: ϕ1 = 0 and H1: ϕ1 ≠ 0, the resulting test statistic provided |0.1137/0.1026| = 1.10, which falls below the critical threshold of 2. As a result, at the 5% significance level, the hypothesis test analysing the coefficients of the autoregressive (AR) term in the ARIMA(1,1,1) model provided insufficient evidence to reject the null hypothesis, implying that the AR term should be excluded from the model.



```
Call:
arima(x = cet_temp, order = c(1, 1, 1), method = "ML")

Coefficients:
         ar1      ma1
      0.1137  -0.8749
s.e.  0.1026   0.0454

sigma^2 estimated as 0.3618:  log likelihood = -110.81,  aic = 227.63
```

Finally, fitting the ARIMA(1,1,1) and ARIMA(0,1,2) models, yielded higher AIC values and larger p-values for hypothesis testing, suggesting an inclination for the simpler ARIMA(0,1,1) model. Relying on these evaluations, the ARIMA(0,1,1) model was determined to be the best choice for explaining the temperature data, as it struck a reasonable balance between model complexity and goodness of fit while accurately expressing the dataset's underlying patterns.

<u>Equation of ARIMA(0,1,1) model:</u>

$$\phi(B)\,[\,(1-B)^{d}\,X_t\,] = \theta(B)\,Z_t$$

$$1^*\,[\,(1-B)^{1}\,X_t\,] = (1+\theta_1 B)\,Z_t$$

$$X_t - X_{t-1} = Z_t + \theta_1 Z_{t-1}$$

R Code:

```
#reading cet_temp.csv file
cet_df<-read.csv("cet_temp.csv",header=TRUE)

#producing time series plot using avg_annual_temp_C starting from year 1900
cet_temp<- ts(cet_df$avg_annual_temp_C,start=1900,frequency=1)
ts.plot(cet_temp)

#plotting acf and pacf
x11()
par(mfrow=c(1,2))
acf(cet_temp)
pacf(cet_temp)

#differencing the data and plotting its timeplot
cet_temp_diff<- diff(cet_temp)
ts.plot(cet_temp_diff)

#acf and pacf of differenced data
x11()
par(mfrow=c(1,2))
acf(cet_temp_diff)
pacf(cet_temp_diff)

#fitting arima model

#start with ARIMA(0,1,1) process

model1.MA1<- arima(cet_temp,order=c(0,1,1),method="ML")
model1.MA1

#residuals of model1
resid.MA1<-residuals(model1.MA1)
#timeplot of model1 residuals
ts.plot(resid.MA1,main='Time plot of ARIMA(0,1,1) Model Residuals')
#acf of model1 residuals
acf(resid.MA1)

#Ljung-Box test for p-values
MA1.LB<-LB_test(resid.MA1,max.k=11,p=0,q=1)
MA1.LB

#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(MA1.LB$deg_freedom,MA1.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P�value",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

#FITTING ARIMA(0,1,2) MODEL

model3.ARIMA<- arima(cet_temp,order=c(0,1,2),method="ML")
```

```
model3.ARIMA

#residuals of model3
resid3.ARIMA<-residuals(model3.ARIMA)
#timeplot of model2 residuals
ts.plot(resid3.ARIMA,main='Time plot of ARIMA(0,1,2) Model Residuals')
#acf of model2 residuals
acf(resid3.ARIMA)

#Ljung-Box test for p-values
ARIMA3.LB<-LB_test(resid3.ARIMA,max.k=11,p=0,q=2)
ARIMA3.LB

#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(ARIMA3.LB$deg_freedom,ARIMA3.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P�value",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

#FITTING ARIMA(1,1,1) MODEL

model2.ARIMA<- arima(cet_temp,order=c(1,1,1),method="ML")
model2.ARIMA

#residuals of model2
resid2.ARIMA<-residuals(model2.ARIMA)
#timeplot of model2 residuals
ts.plot(resid2.ARIMA,main='Time plot of ARIMA(1,1,1) Model Residuals')
#acf of model2 residuals
acf(resid2.ARIMA)

#Ljung-Box test for p-values
ARIMA.LB<-LB_test(resid2.ARIMA,max.k=11,p=1,q=1)
ARIMA.LB

#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(ARIMA.LB$deg_freedom,ARIMA.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P�value",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)
```

Analysing Trends and Forecasting Housing Prices in the East Midlands: A Time-Series Analysis from 2010 to 2019

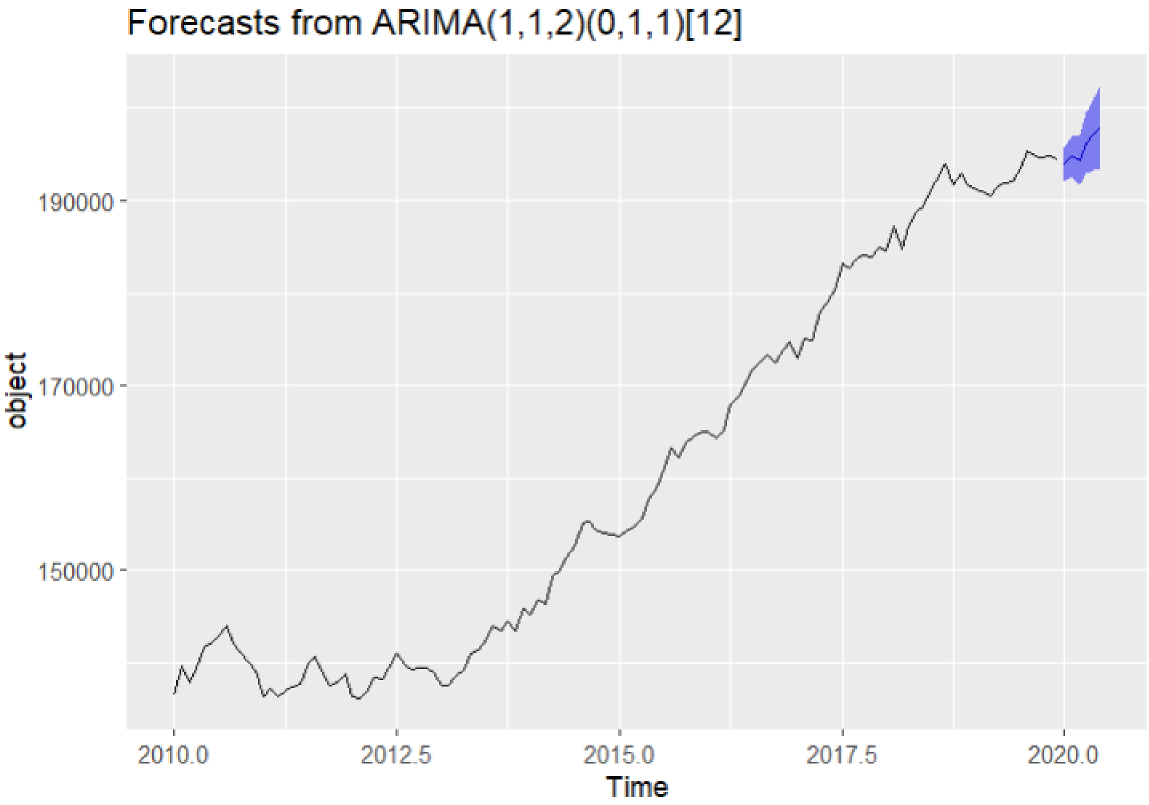
Executive Summary:

The average housing prices in the United Kingdom's East Midlands region from January 2010 to December 2019 are thoroughly examined in this research. The dataset, which included monthly average cost of housing expressed in GBP, was analysed to find hidden trends, patterns, and possible reasons that could affect real estate prices. We wanted to offer insightful information to investors, homeowners, and politicians using time-series analytic approaches.

The housing market data clearly demonstrated seasonality, which our study showed was a non-stationary process. We implemented differencing techniques to get rid of seasonality and attain stationarity in order to tackle this. Upon meticulous analysis of multiple SARIMA models, we discovered that the SARIMA(1, 1, 2) × (0, 1, 1)$_{12}$ model yielded the most accurate predictions for future housing prices and had the best fit to the data.

In addition, our prediction for the first half of 2020, which relies on the SARIMA(1, 1, 2) × (0, 1, 1)$_{12}$ model, indicates that home prices will continue to rise. For those involved in the industry of real estate, these forecasts and their prediction intervals offer significant insights into anticipated trends in prices and related uncertainties.

In general, this research is a useful tool for homeowners, shareholders, and legislators as it provides practical understanding of the characteristics of the East Midlands housing market and facilitates well-informed decision-making in the real estate industry.

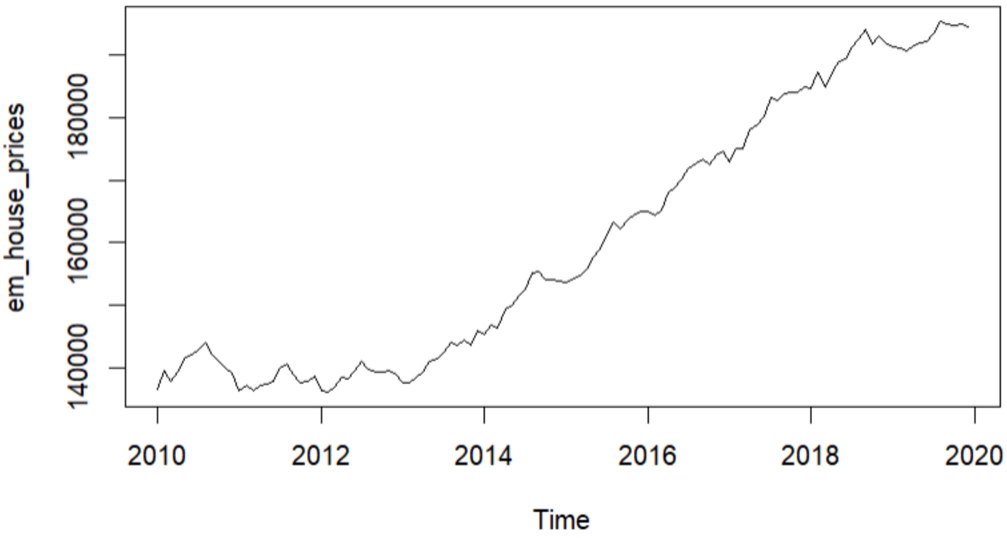


Introduction:

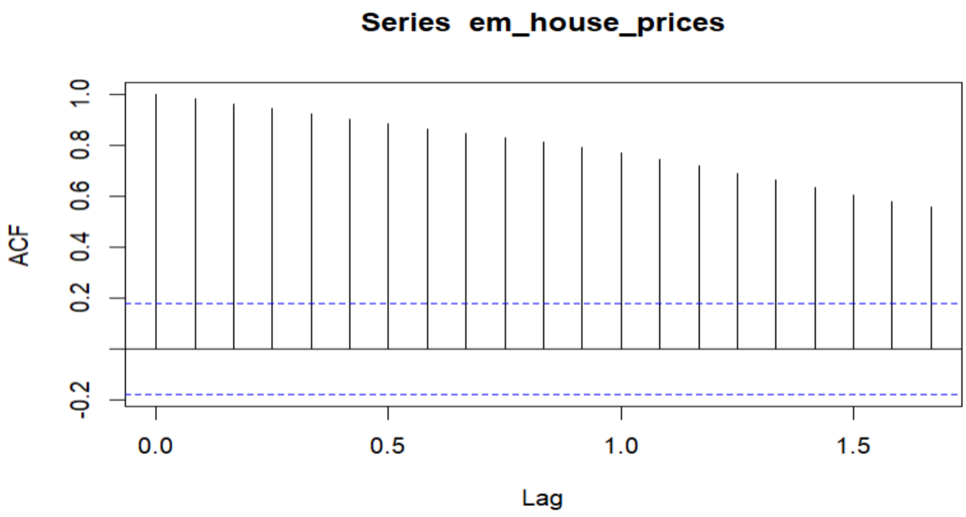
The analysis of housing market changes is critical for policymakers, investors, and homeowners. In this research, we analyse average real estate prices in the East Midlands region of the United Kingdom from January 2010 to December 2019.
The dataset, named "em_house_prices.csv", comprises of monthly average house prices in GBP (British Pounds) for the designated time period. Each record comprises the month and year of the data point, as well as the associated mean housing price.
The objective of our research would be to undertake a thorough examination of housing market data employing time-series analysis techniques to uncover that underlies patterns, trends, and potential causes impacting house prices.

Model Fitting and Interpretation:

After importing and observing the data containing average home prices in the East Midlands region, it became apparent that the time series plot discovered clear patterns that repeated at regular intervals, showing an existence of seasonality. This finding implies that the process is not stationary, i.e. it fails to have a steady mean or variance across time.

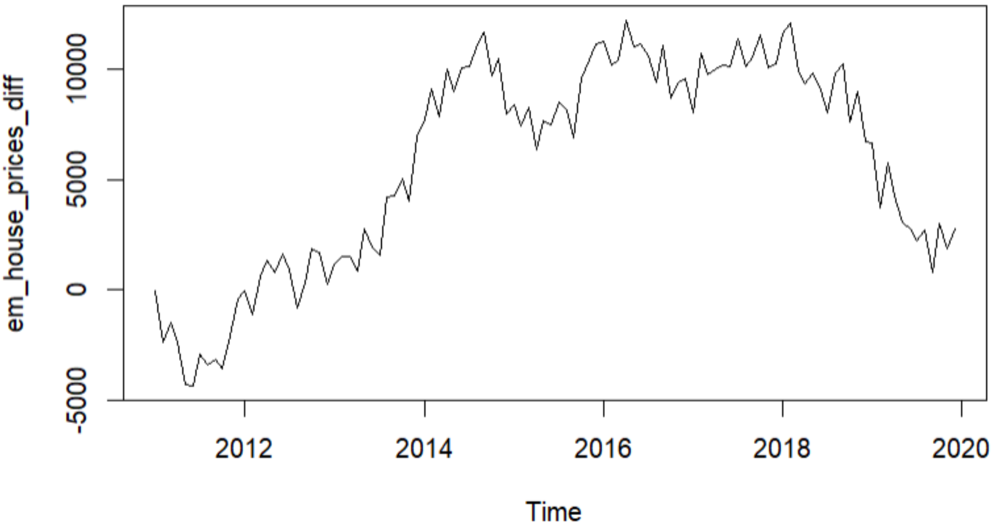


Moreover, the autocorrelation function (ACF) plot was utilized to investigate the data's autocorrelation structure. The ACF plot illustrated that the autocorrelation coefficients did not rapidly cut off to zero as the lag grew. On the contrary, they showed considerable autocorrelation over many delays, revealing persisting patterns and relationships in the data.

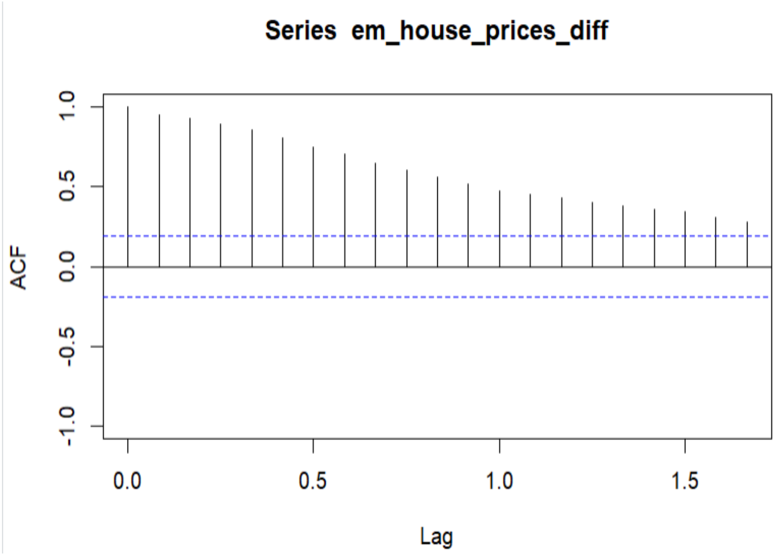


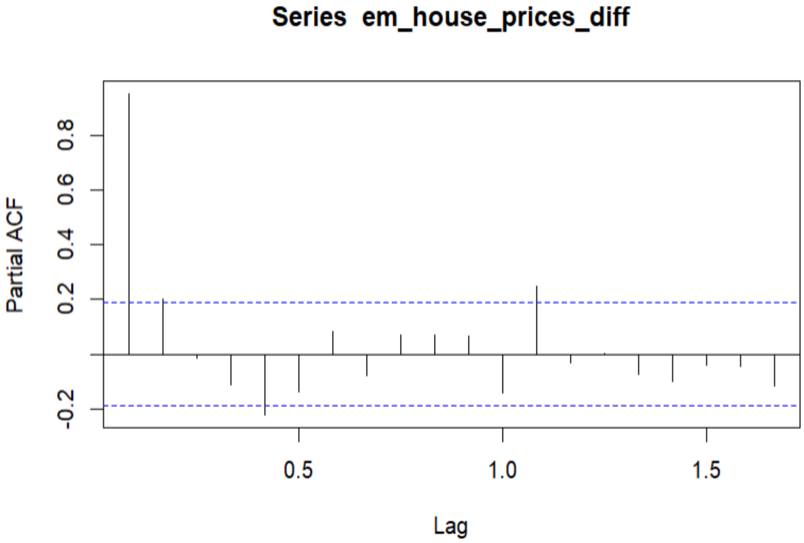
The aforementioned results indicate that the average house price data is seasonal and non-stationary, emphasizing the necessity for further preprocessing processes, which might include differencing or transformation, to produce stationarity. Handling non-stationarity is essential for ensuring the reliability of future modelling efforts and producing credible forecasts.

To overcome the seasonality evident in the above time series plot, we used differencing with a 12-month lag. This method attempts to eliminate the seasonal component from the time series. Plotting the differenced time series revealed that seasonality had been successfully eliminated, as shown by the lack of distinct patterns occurring at monthly time frames. However, it is important to highlight that the resulting time series is still not stationary.

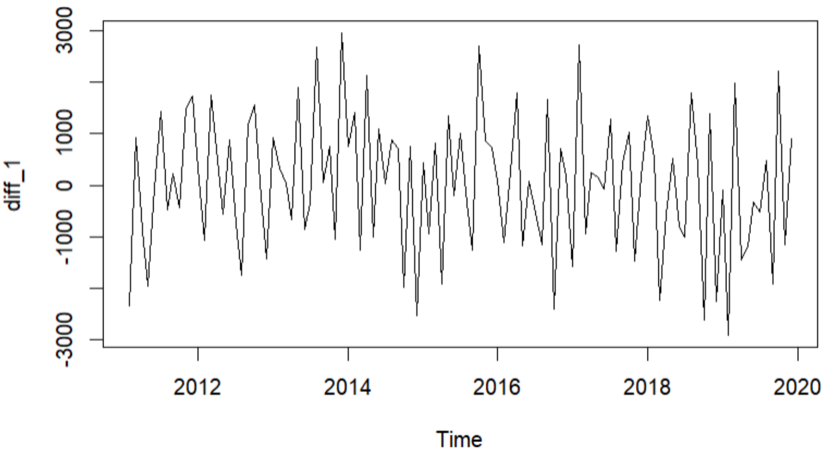


A closer look at the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced data revealed much more information. Regardless of the elimination of seasonality, both plots showed patterns consistent with autocorrelation, indicating that the differenced data wasn't stationary.



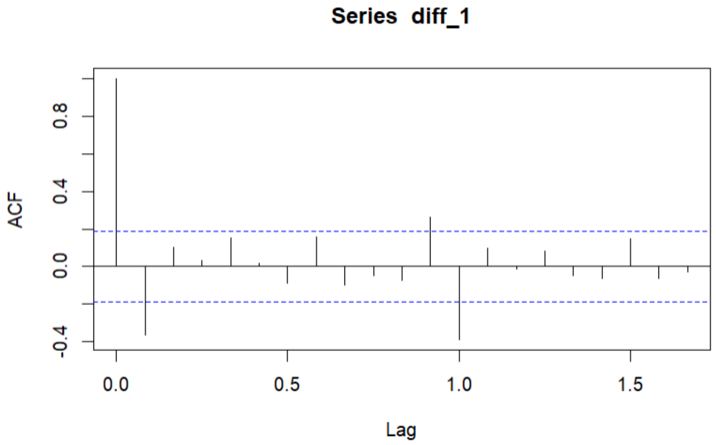


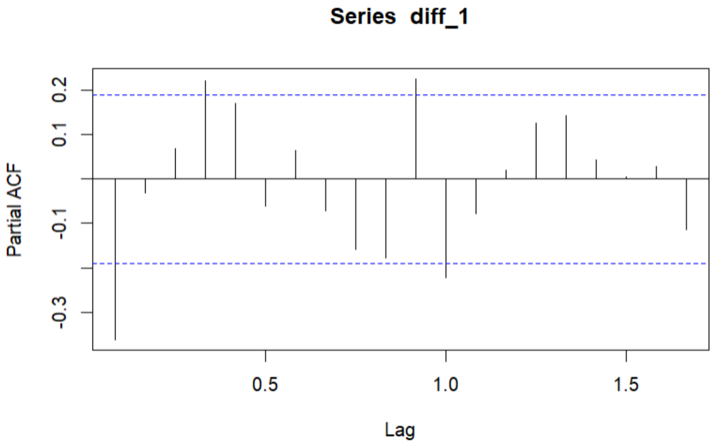
In order to produce stationarity in the seasonally differenced data, we imposed a first difference to the existing differenced series. This additional differencing phase was designed to get rid of any unwanted seasonal dependencies and ensure that the final time series matched the standards for stationarity.



After visualizing, the time series seemed to have stationarity behaviours. There were no discernible patterns or trends, indicating that the temporal dependencies had been efficiently erased.

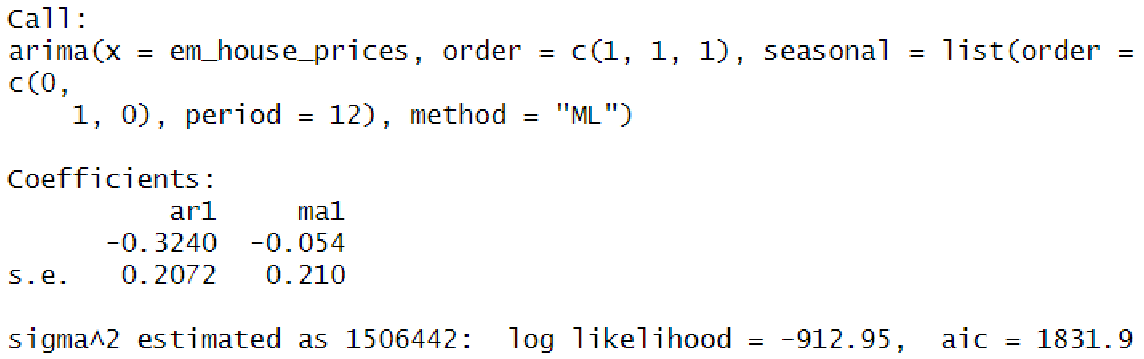
Further evaluation of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots confirmed stationarity. The ACF plot quickly decayed to zero after lag 1, showing that there were not any noteworthy autocorrelations after the first lag. This ACF plot suggests that there might be an MA (Moving Average) component. However, the PACF plot doesn't gradually dampens to zero rather dropped rapidly after lag 1, indicating that the model is not a pure MA(1) model, there might be an AR (AutoRegressive) component in it.





These findings indicate that the first difference in the seasonally differentiated data efficiently produced stationarity, which makes it acceptable for future time series modelling.

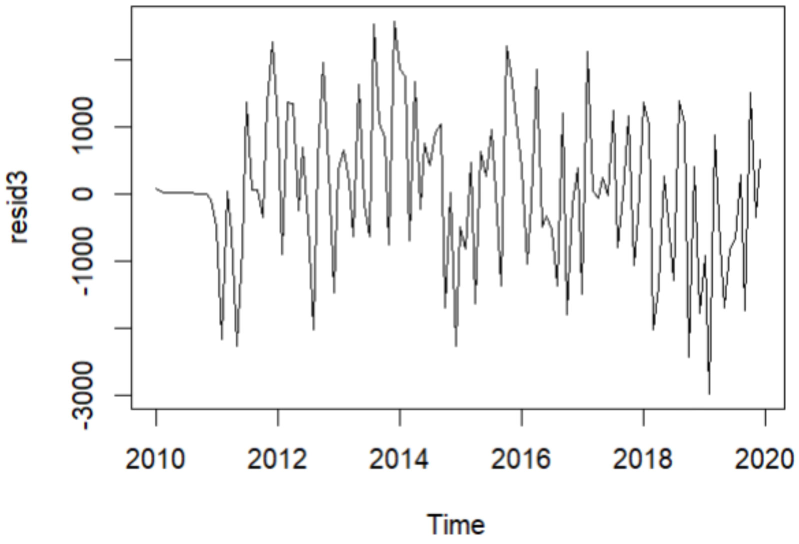
To begin the modelling procedure, we fitted an SARIMA(1, 1, 1) × (0, 1, 0)$_{12}$ model to the seasonally differenced average house price data. We computed the SARIMA(1, 1, 1) × (0, 1, 0)$_{12}$ model and produced a summary of the parameters.



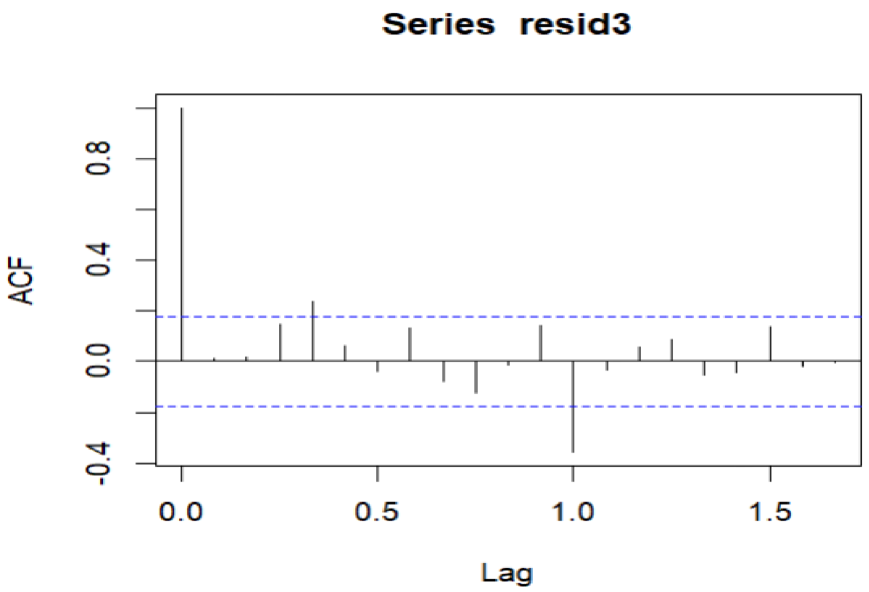
```
Call:
arima(x = em_house_prices, order = c(1, 1, 1), seasonal = list(order =
c(0,
    1, 0), period = 12), method = "ML")

Coefficients:
          ar1      ma1
      -0.3240   -0.054
s.e.   0.2072    0.210

sigma^2 estimated as 1506442:  log likelihood = -912.95,  aic = 1831.9
```

The generated model had an Akaike Information Criterion (AIC) value of 1831.9, showing the model' s quality of fit to the data.

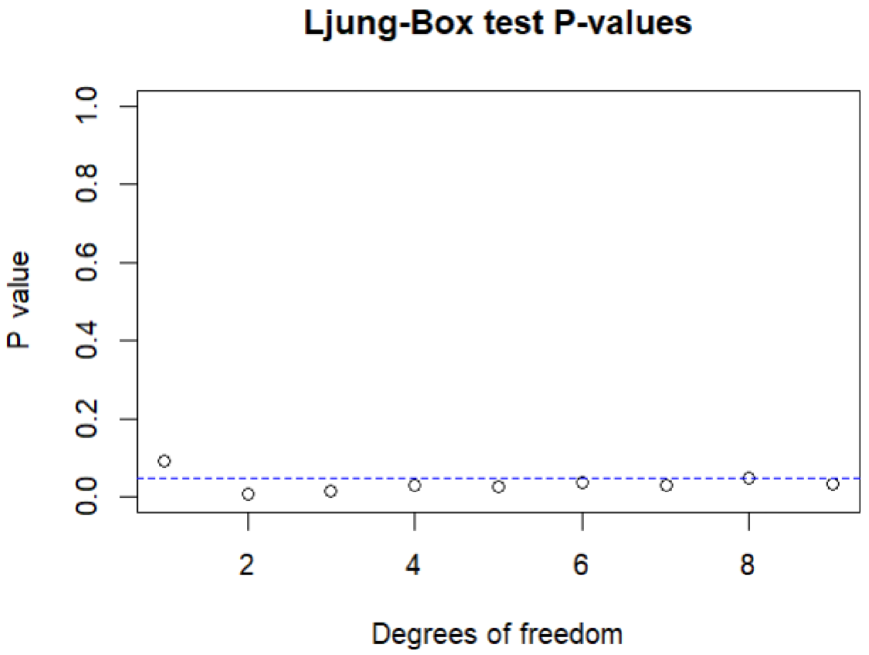
The residuals of the SARIMA(1, 1, 1) × (0, 1, 0)$_{12}$ model, designated as 'resid1', were analysed. The time plot of the residuals revealed a resemblance to white noise, with no visible patterns or trends.



Likewise, the autocorrelation function (ACF) plot of the residuals revealed values close to zero at all lags greater than lag 0, confirming residual independence.



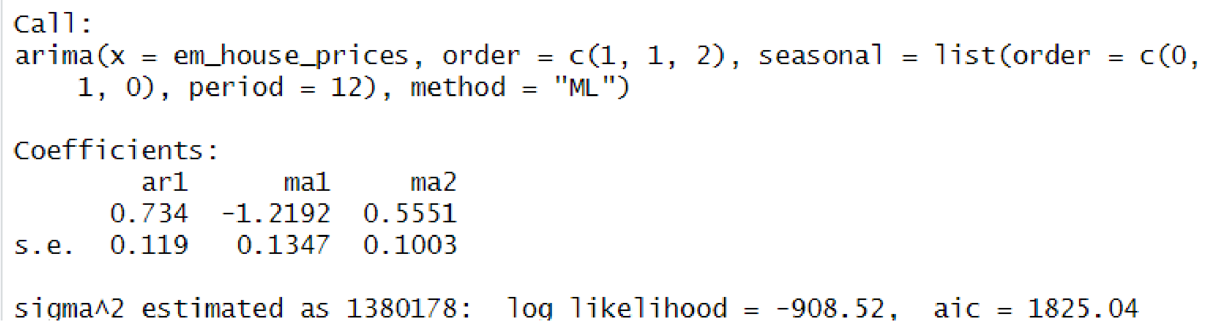
The Ljung-Box test was used to formally assess the absence of autocorrelation in the residuals. The corresponding p-values, shown against degrees of freedom, revealed that only few p-values were exceeding the standard 5% significance level.



Although the SARIMA(1, 1, 1) × (0, 1, 0)$_{12}$ model performed well, the tiny p-values indicated that alternate model parameters should be explored in order to enhance data fit.

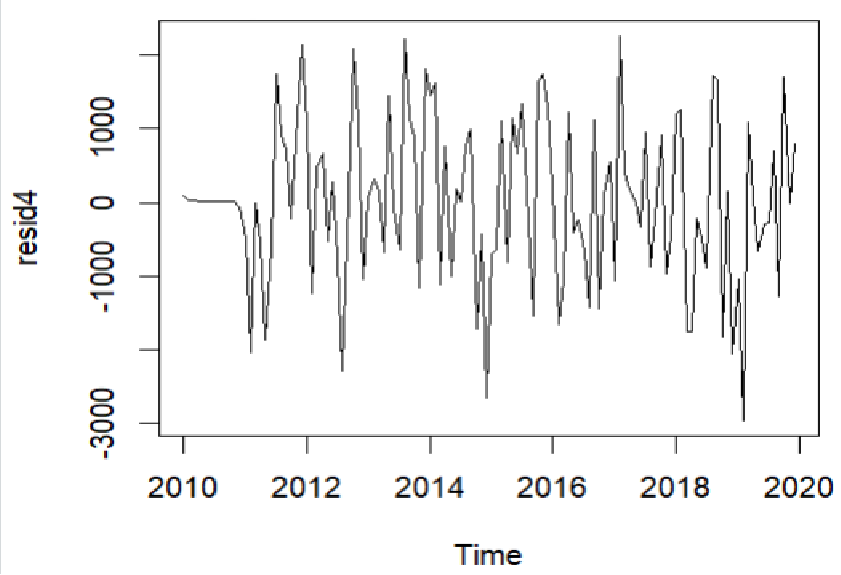
Now, to check whether an extra MA component should be added to the model, we implemented an SARIMA(1, 1, 2) × (0, 1, 0)$_{12}$ model to analyse seasonally differenced average house prices data. This model builds on the previous SARIMA(1, 1, 1) model by incorporating a second moving average (MA) factor. The resulting model had a bit lower Akaike Information Criterion (AIC) value of 1825.04 than t he SARIMA(1, 1, 1) model with AIC = 1831.9, implying a better fit to the data.

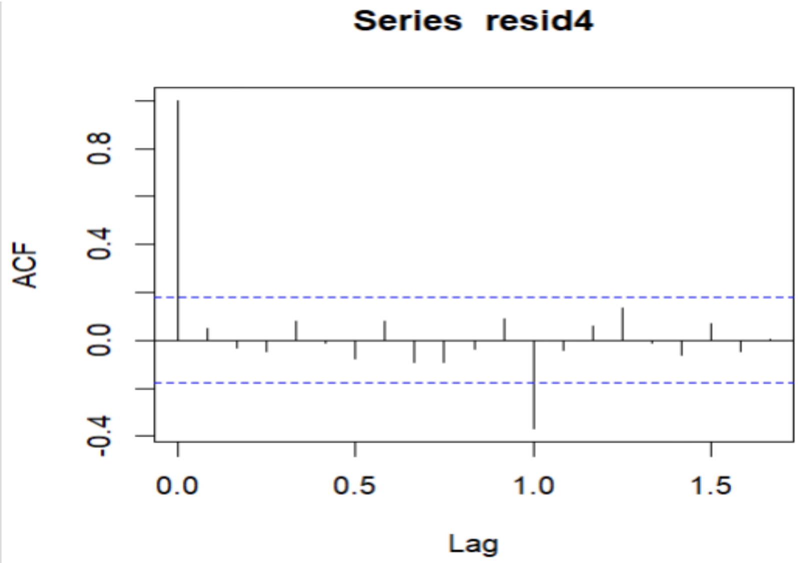
A hypothesis test is performed to test whether a second MA component should be included in the model or not. A test statistic of |0.5551/0.1003|=5.532, which is greater than 2, was obtained from the hypothesis test for the second moving average term MA(2) in the model. This indicates that we reject H0: θ2=0 and suggests that the MA(2) parameter is statistically significant.



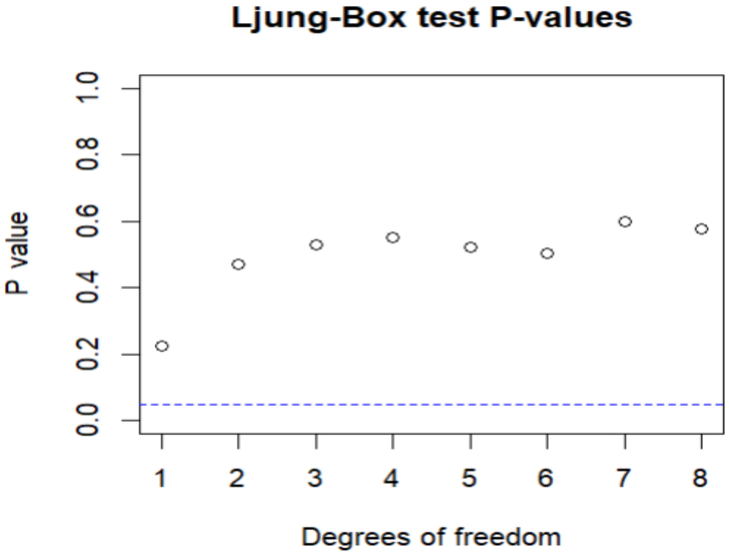
```
Call:
arima(x = em_house_prices, order = c(1, 1, 2), seasonal = list(order = c(0,
    1, 0), period = 12), method = "ML")

Coefficients:
        ar1      ma1     ma2
      0.734  -1.2192  0.5551
s.e.  0.119   0.1347  0.1003

sigma^2 estimated as 1380178:  log likelihood = -908.52,  aic = 1825.04
```

The time plot of the model residuals shows random variations close to zero, indicating absence of a pattern or trend.
The residuals' sample autocorrelation function (ACF) contains no notable spikes at any lags, suggesting that they are not correlated.



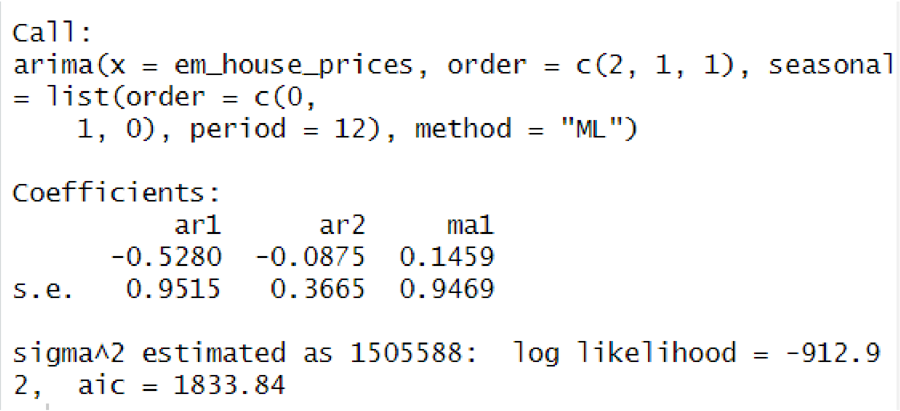


The model residuals were subjected to the Ljung-Box test. All of the test's p-values are above the significance level of 0.05, suggesting that there is no indication of residual autocorrelation at any lag.



With consideration to all these results, it indicated that model SARIMA(1, 1, 2) × $(0, 1, 0)_{12}$ was a better fit to the data.

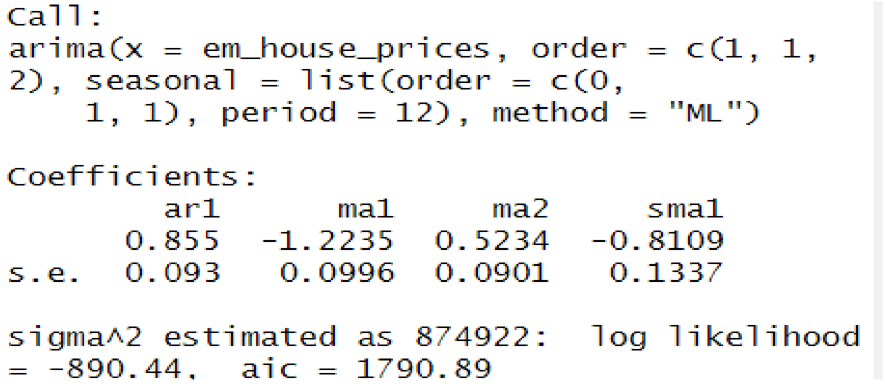
Similarly, to check whether an extra AR component should be added to the model, we implemented an SARIMA(2, 1, 1) × $(0, 1, 0)_{12}$ model to analyse seasonally differenced average house prices data. This model had a greater Akaike Information Criterion (AIC) value of 1833.84 than the currently best model SARIMA(1, 1, 2) with AIC = 1825.04, implying that it might not be a better fit to the data. Moreover, a hypothesis test was performed to test whether a second AR component should be inclu ded in the model or not. A test statistic of |-0.0875/ 0.3665|= 0.2387, which is less than 2, was obtai ned from the hypothesis test for the second AutoRegressive term AR(2) in the model. This indicates t hat we fail to reject H0: ϕ2=0 and suggests that the AR(2) parameter is not statistically insignificant.



```
Call:
arima(x = em_house_prices, order = c(2, 1, 1), seasonal
= list(order = c(0,
    1, 0), period = 12), method = "ML")

Coefficients:
          ar1      ar2     ma1
      -0.5280  -0.0875  0.1459
s.e.   0.9515   0.3665  0.9469

sigma^2 estimated as 1505588:  log likelihood = -912.9
2,  aic = 1833.84
```

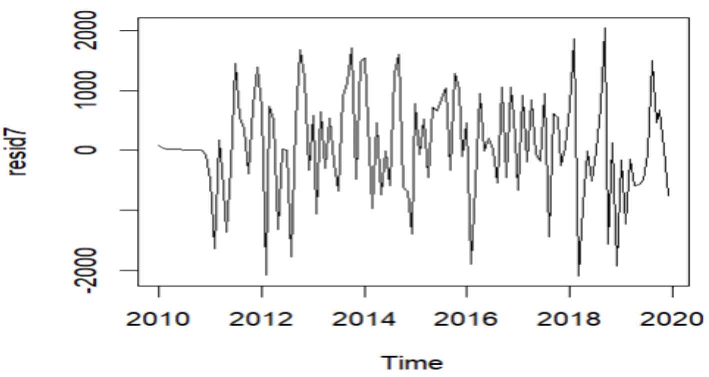
Moving ahead, we tried to add a seasonality MA component (i.e. Q=1) to the current best model SARIMA(1, 1, 2) × (0, 1, 0)$_{12}$ with AIC = 1825.04 to check whether adding this seasonal MA component to the current best model improves the quality of fit to the data.

An improved model fit was demonstrated by the SARIMA(1, 1, 2) × (0, 1, 1)$_{12}$ model's AIC value of 1790.89, which is smaller compared to the AIC of the previously examined SARIMA (1, 1, 2) × (0, 1, 0)$_{12}$ model. Additionally, a hypothesis test for seasonal MA component labelled as 'sma1' yields a test statistic of |-0.8109/0.1337|=6.06. This value is greater than 2 which indicated that we reject the hypothesis test for H0: Θ1 = 0 (i.e. sma1=0). This result illustrated that the seasonal MA component was statistically significant and can be the added to the model.

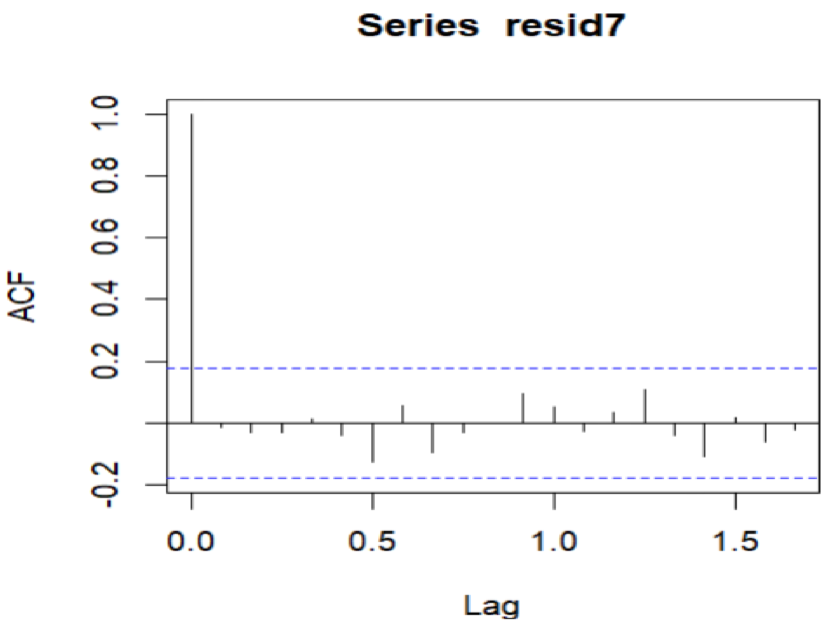


```
Call:
arima(x = em_house_prices, order = c(1, 1,
2), seasonal = list(order = c(0,
    1, 1), period = 12), method = "ML")

Coefficients:
        ar1      ma1     ma2     sma1
      0.855  -1.2235  0.5234  -0.8109
s.e.  0.093   0.0996  0.0901   0.1337

sigma^2 estimated as 874922:  log likelihood
= -890.44,  aic = 1790.89
```

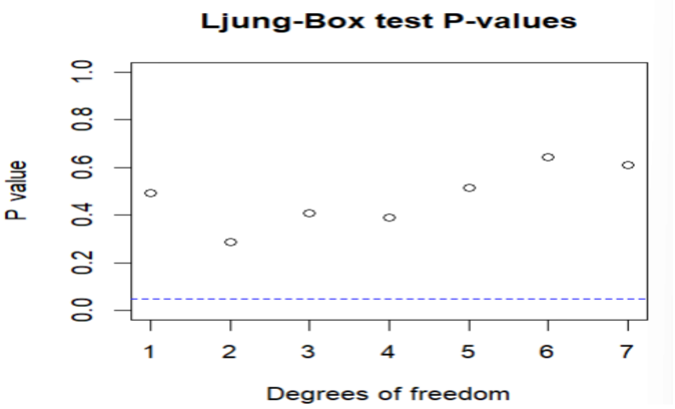
Also, the model residuals' time plot indicates that they show erratic swings around zero, with no discernible pattern or trend.



There are no noticeable spikes in the sample autocorrelation function (ACF) of the residuals at any lag, suggesting that the residuals are not correlated and acts as white noise.



Finally, to conclude, the residuals of the model were subjected to the Ljung-Box test. The Ljung-Box test yields p-values that are all greater than the significance level of 0.05, meaning that residuals are independent and don’t exhibit any autocorrelation.

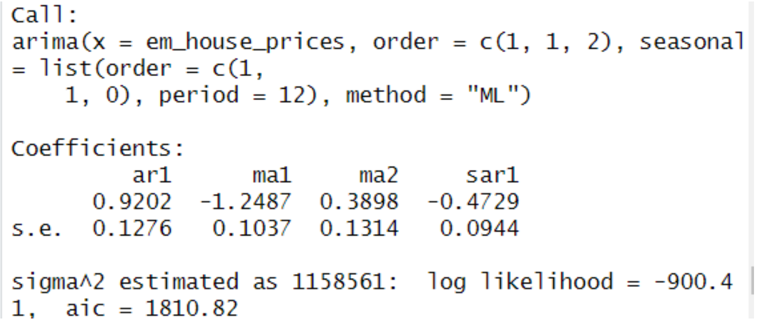


These results aided us to come to a conclusion that the SARIMA(1, 1, 2) × (0, 1, 1)$_{12}$ model might be a better fit when compared to the previous best fit SARIMA(1, 1, 2) × (0, 1, 0)$_{12}$ model.

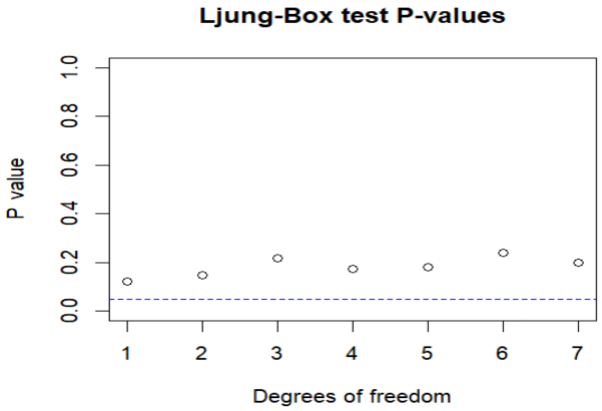
As a final evaluation, by taking into account various values for the seasonal autoregressive and seasonal moving average parameters (P and Q), we investigated additional seasonal components instead of seasonal MA component in an effort to improve the SARIMA(1, 1, 2) × (0, 1, 0)$_{12}$ model that was previously fitted to the average house prices data.

The AIC of model SARIMA(1, 1, 2) × (1, 1, 0)$_{12}$ is 1810.82, greater than the AIC of the SARIMA(1, 1, 2) × (0, 1, 0)$_{12}$ model that was previously examined, indicating that the introduction of the seasonal autoregressive term (SAR) did not enhance the model's fit.

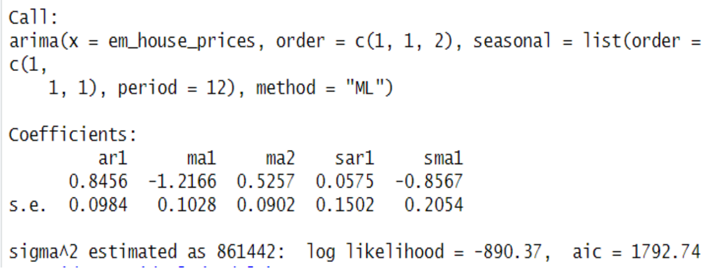
Also, the p-values of this model were a bit smaller as compared to the model with seasonal MA(sma1) component which lacked the evidence that it might be a better fit.



```
Call:
arima(x = em_house_prices, order = c(1, 1, 2), seasonal
= list(order = c(1,
    1, 0), period = 12), method = "ML")

Coefficients:
         ar1      ma1     ma2      sar1
      0.9202  -1.2487  0.3898   -0.4729
s.e.  0.1276   0.1037  0.1314    0.0944

sigma^2 estimated as 1158561:  log likelihood = -900.4
1,  aic = 1810.82
```



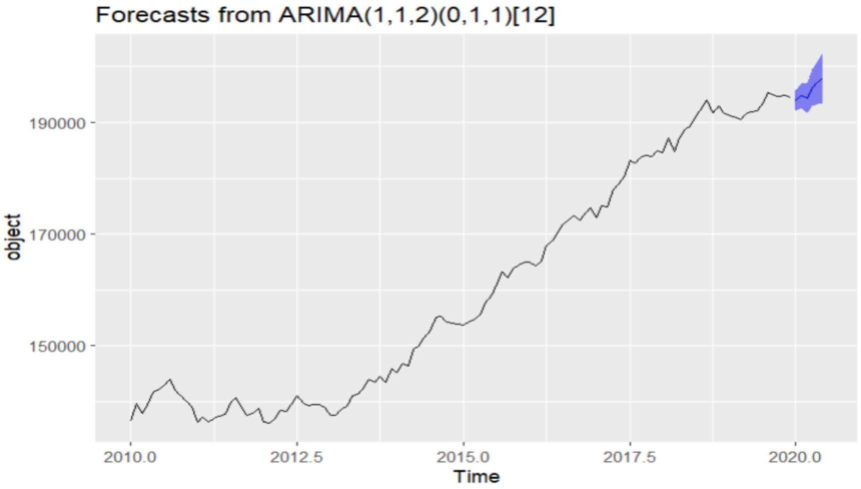
Similarly, we tried to introduce both the seasonal components in the model(i.e. P=1 and Q=1), resulting to a model as SARIMA(1, 1, 2) × $(1, 1, 1)_{12}$. An AIC of 1792.74 is produced by model SARIMA(1, 1, 2) × $(1, 1, 1)_{12}$, which is greater than the AIC of the SARIMA(1, 1, 2) × $(0, 1, 1)_{12}$ model that was previously under consideration. Furthermore, the seasonal autoregressive term's (SAR1) hypothesis test (H0: Φ1=0) proves that the term shouldn't be a part of the model as the test statistic obtained is |0.0575/0.1502|= 0.3838(< 2).



```
Call:
arima(x = em_house_prices, order = c(1, 1, 2), seasonal = list(order =
c(1,
    1, 1), period = 12), method = "ML")

Coefficients:
         ar1      ma1     ma2    sar1     sma1
      0.8456  -1.2166  0.5257  0.0575  -0.8567
s.e.  0.0984   0.1028  0.0902  0.1502   0.2054

sigma^2 estimated as 861442:  log likelihood = -890.37,  aic = 1792.74
```

Based on its lower AIC value (1790.89) than the alternatives, the SARIMA(1, 1, 2) × $(0, 1, 1)_{12}$ model continues to be the best appropriate choice among the studied models. This model is more suited for predicting future values of average house prices because it effectively reflects the underlying patterns and seasonality in the information provided.

For our next task, we aimed to produce a time series model to forecast the monthly house prices for the first six months of 2020. The model made use of previous data on housing prices spanning from January 2010 to December 2019. We used a SARIMA(1, 1, 2) × (0, 1, 1)12 model to provide estimates for January–June 2020 based on this data. From the plot of forecasted values, it is quite evident that the real estate values were initially a bit low with a few fluctuations in the prices from 2010 to approximately mid of 2012. The housing prices started to rise from 2013 to 2020, it follows an upward trend. Based on these observations, we predicted home prices for next six months. It illustrated that the prices continue to rise for January 2010 to December 2019 and also for the predicted next six months. The dark blue line indicates true values of our prediction. With a certain degree of confidence, the prediction intervals show the range that real home values are expected to fall inside. The shaded areas give uncertainty bounds (95%) around the forecasted estimates.



We have generated relevant projections for the monthly housing prices for the first half of 2020 by utilizing the SARIMA(1, 1, 2) × (0, 1, 1)12 model. For those involved in the real estate market, these predictions with prediction intervals provide insightful information that helps them come to conclusions based on expected price trends and related uncertainties.

R Code for Question2:

```
#reading/ importing em_house_prices.csv file
em_house_df<-read.csv("em_house_prices",header=TRUE)

#producing timeplot using average_price_gbp starting from year 2010 for 12 months
em_house_prices<- ts(em_house_prices$average_price_gbp,start=2010,frequency=12)
ts.plot(em_house_prices)

#acf of the house
acf(em_house_prices)


#we’d first difference the data with lag 12.
em_house_prices_diff<- diff(em_house_prices, lag = 12)
#timeplot of differenced data
ts.plot(em_house_prices_diff)

#acf and pacf of differenced data
acf(em_house_prices_diff,ylim=c(-1,1))
pacf(em_house_prices_diff)

#we’ll take the first difference of the seasonally differenced data.
diff_1<- diff(em_house_prices_diff)
#timeplot of seasonally differenced data.
ts.plot(diff_1)

#acf and pacf of seasonally differenced data.
acf(diff_1)
pacf(diff_1)
```

```
#fitting an ARIMA(1, 1, 1) × (0, 1, 0)12 model.

model3 <- arima(em_house_prices,order=c(1,1,1),
          seasonal=list(order=c(0,1,0), period=12),
          method="ML")
model3

#model residuals and its timeplot
resid3<-residuals(model3)
ts.plot(resid3)

#sample acf of model3 residuals
acf(resid3)

#Ljung-Box tests for the model residuals
model3.LB<-LB_test_SARIMA(resid3, max.k=11, p=1, q=1, P=0, Q=0)
model3.LB

#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(model3.LB$deg_freedom,model3.LB$LB_p_value,xlab="Degrees of freedom",ylab="P�value",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

#fitting an ARIMA(1, 1, 2) × (0, 1, 0)12 model.

model4 <- arima(em_house_prices,order=c(1,1,2),
          seasonal=list(order=c(0,1,0), period=12),
          method="ML")
model4
```

```
#model residuals and its timeplot
resid4<-residuals(model4)

ts.plot(resid4)

#sample acf of model4 residuals
acf(resid4)

#Ljung-Box tests for the model residuals
model4.LB<-LB_test_SARIMA(resid4, max.k=11, p=1, q=2, P=0, Q=0)
model4.LB

#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(model4.LB$deg_freedom,model4.LB$LB_p_value,xlab="Degrees of freedom",ylab="P-value",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

#fitting an ARIMA(2, 1, 1) × (0, 1, 0)12 model.

model5 <- arima(em_house_prices,order=c(2,1,1),
           seasonal=list(order=c(0,1,0), period=12),
           method="ML")
model5

#trying seasonality components for model 4 with Q=1

model7 <- arima(em_house_prices,order=c(1,1,2),
           seasonal=list(order=c(0,1,1), period=12),
           method="ML")
model7
```

```
#model residuals and its timeplot

resid7<-residuals(model7)

ts.plot(resid7)


#sample acf of model7 residuals

acf(resid7)


#Ljung-Box tests for the model residuals

model7.LB<-LB_test_SARIMA(resid7, max.k=11, p=1, q=2, P=0, Q=1)

model7.LB


#To produce a plot of the P-values against the degrees of freedom and

#add a blue dashed line at 0.05, we run the commands

plot(model7.LB$deg_freedom,model7.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P⍰value",main="Ljung-Box test P-values",ylim=c(0,1))

abline(h=0.05,col="blue",lty=2)


#Additionally trying seasonality components for model 4 with P=1


model6 <- arima(em_house_prices,order=c(1,1,2),

            seasonal=list(order=c(1,1,0), period=12),

            method="ML")

model6


#model residuals and its timeplot

resid6<-residuals(model6)

ts.plot(resid6)


#sample acf of model6 residuals

acf(resid6)
```

```
#Ljung-Box tests for the model residuals
model6.LB<-LB_test_SARIMA(resid6, max.k=11, p=1, q=2, P=1, Q=0)
model6.LB

#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(model6.LB$deg_freedom,model6.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P⍰value",main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

#current best model --------- model7

#now trying seasonality components for model 7 with P=1

model8 <- arima(em_house_prices,order=c(1,1,2),
            seasonal=list(order=c(1,1,1), period=12),
            method="ML")
model8

#model residuals and its timeplot
resid8<-residuals(model8)
ts.plot(resid8)

#sample acf of model8 residuals
acf(resid8)

#Ljung-Box tests for the model residuals
model8.LB<-LB_test_SARIMA(resid8, max.k=11, p=1, q=2, P=1, Q=1)
model8.LB
```

```r
#To produce a plot of the P-values against the degrees of freedom and

#add a blue dashed line at 0.05, we run the commands

plot(model8.LB$deg_freedom,model8.LB$LB_p_value,xlab="Degrees of freedom",ylab="P�value",main="Ljung-Box test P-values",ylim=c(0,1))

abline(h=0.05,col="blue",lty=2)


#Finally, model 7 seems to be the best fit

# model ARIMA(1, 1, 2) × (0, 1, 1)12 model is best fit


#forecasting for next six months of 2020 using model7


#Forecast for the next 6 months (h=6)

fc_6m<-forecast(em_house_prices,h=6,model=model7,level=95)


#Plot the forecasted values

autoplot(fc_6m)
```