The University of Nottingham

SCHOOL OF MATHEMATICAL SCIENCES

SPRING SEMESTER 2023-2024

MATH4022 - TIME SERIES AND FORECASTING

Coursework Assignment

Name: Brendan Ezekiel Agnelo Vaz

Student id: 20610206

Course: MSc Data Science

Analysis of The Annual Mean Temperature Data for the Midlands Region of England

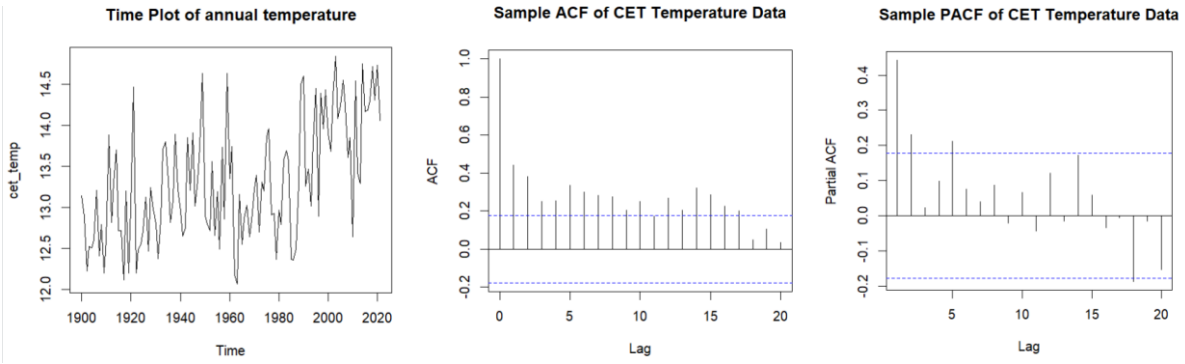
Introduction

The aim of the analysis is to find a suitable time series model to characterize the annual mean temperature data for the Midlands Region of England. Sourced from the UK Meteorological Office Hadley Climate Centre, the dataset utilized for this analysis includes yearly mean temperature from 1900 up to 2021.

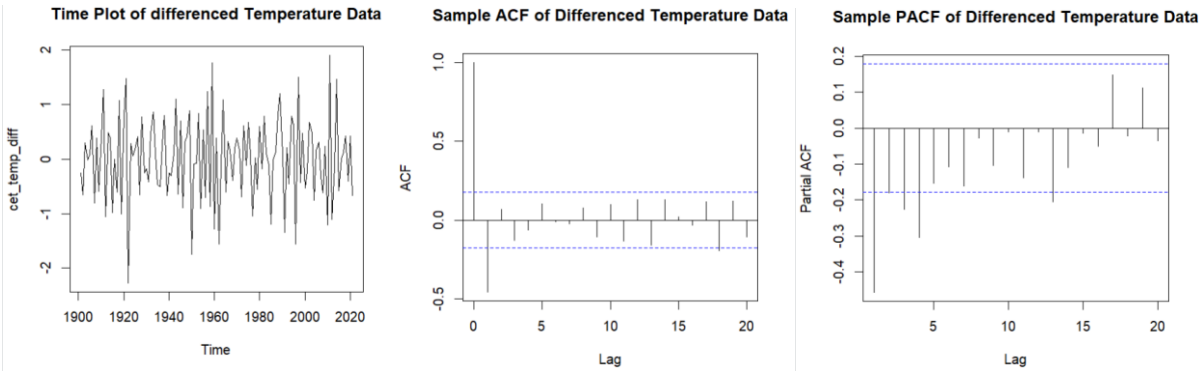
The dataset includes the Midlands Region of England's annual mean temperature values in degrees Celsius ranging from 1900 to 2021. The average temperature for a specific year is represented by each observation. The 'cet_temp.csv' file contains the data which comprises of two columns namely 'year' and 'avg_annual_temp_C' and a total of one hundred twenty-two observations.

Exploratory Data Analysis (EDA)

The dataset does not contain any missing values. Therefore, it is then converted to time series data. The Time plot, sample Autocorrelation Function (ACF) and sample Partial Autocorrelation Function are then plotted. Initial visualizing suggests that the data is not stationary; there is an overall increasing trend over the years as well as variability. The sample ACF values do not decline sharply as the lag increases and the PACF does not cut off after lag 1 as well.



Hence, to achieve stationarity, the data is differenced. On differencing the new time plot, sample ACF and sample PACF are plotted.



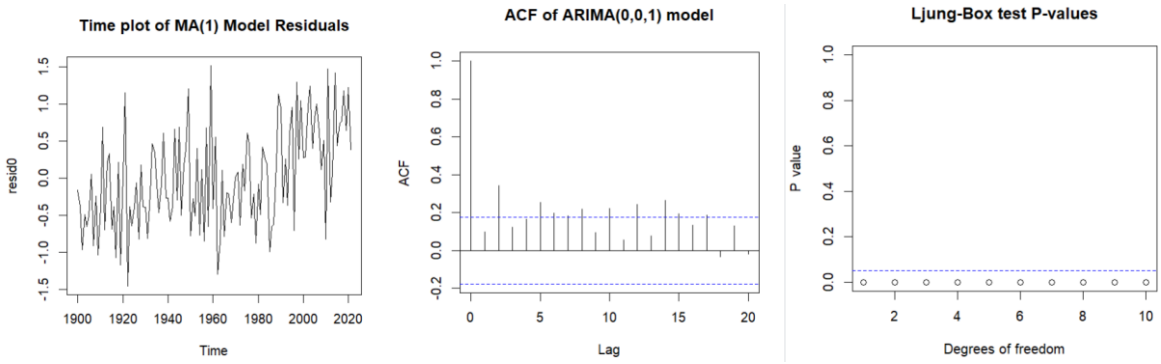
On comparing the plots, its is clear that the first difference of the data is (weakly stationary). There is a constant mean and variability over time. The ACF declines to zero after the first lag. And similarly, The PACF also tends to zero with increase in lag. This indicates a Moving Average or MA (1) process.

Model Fitting

Model 0 - ARIMA (0,0,1) process

To analyse the data, the Autoregressive Integrated Moving Average (ARIMA) model is used. It aims to determine short term dependencies in the data using a single Moving Average Term. The ARIMA (0,01) model is then fitted to the temperature data using maximum likelihood estimation.

Next, the residuals of the ARIMA (0,0,1) model were calculated and plotted to study its behaviour over time. Additionally, the ACF of the residuals is also plotted to identify any dependencies. On examining the plots, it is deduced that the time plot of the residuals looks like white noise, the sample ACF seems to be close to zero at all lags greater than 0. Hence, the residuals are independent.



To assess the residuals' independence at various lags, the Ljung-Box test was used. The test yielded p-values for every lag, which showed how important autocorrelation was in the residuals.

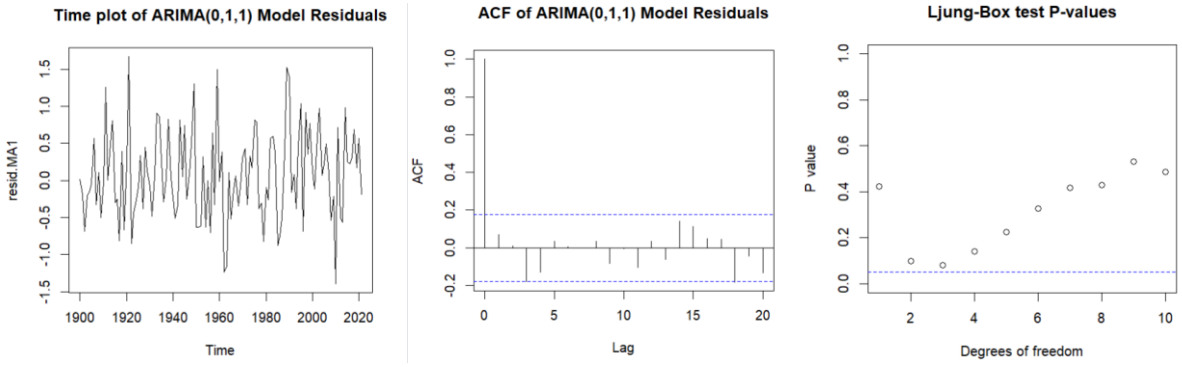
Result

The time plot does is not stationary for the model. The sample ACF of residuals does not cut off to zero as lag increases. Additionally, the p-values are below the threshold off 0.05. This suggests that this model might not be a good fit.

Model 1 - ARIMA (0,1,1) process

The next model used is the ARIMA (0,1,1) model. With only one moving average term, this model aims to attain stationarity by taking first-order differencing into account and capturing short-term dependencies in the data.
This model is then fit onto the data.
To evaluate their behaviour over time, the residuals of the ARIMA (0,1,1) model were computed and shown. The residuals' time plot revealed that they have an independent and random appearance, like white noise. Further evidence of the

residuals' independence came from the autocorrelation function (ACF), which had values near zero at most lags.



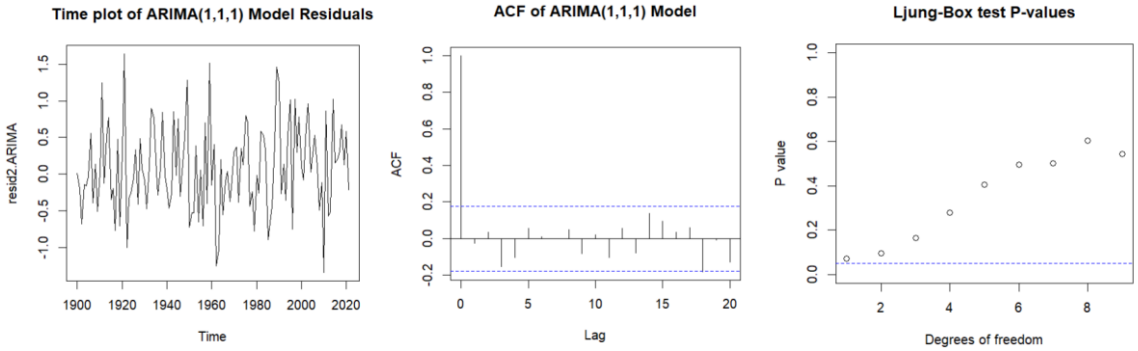
To assess the residuals' independence at various lags, the Ljung-Box test was used. Each lag's computed p-value showed that all the p-values were higher than 0.05, indicating non-significance at the 5% level. This provided more evidence in favour of the independent residual's thesis.

Result

The results of the investigation showed that the autocorrelation in the temperature data for the Midlands region was well captured by the ARIMA (0,1,1) model. The model's suitability for predicting temperature patterns was confirmed by the non-significant p-values and independence of residuals in the Ljung-Box test.
The model is appropriate for temperature variation forecasting because it can capture short-term relationships while maintaining stationarity through differencing. However, we will assess the fit of additional models before concluding.

Model 2 - ARIMA (1,1,1) process

The yearly mean temperature data for the English Midlands region were analysed using the ARIMA (1,1,1) model. This model seeks to capture temporal patterns and short-term dependencies in the data through first-order differencing, a moving average term, and an autoregressive term.
To evaluate the behaviour of the ARIMA (1,1,1) model over time, the residuals were calculated. The residuals' autocorrelation function (ACF) was studied to find any residual autocorrelation, and their time plot was looked at to spot any patterns or trends.



To assess the residuals' independence at various lags, the Ljung-Box test was used. The p-values that were calculated for every lag were compared with the generally used significance level of 0.05 to ascertain whether there was a substantial amount of autocorrelation in the residuals.

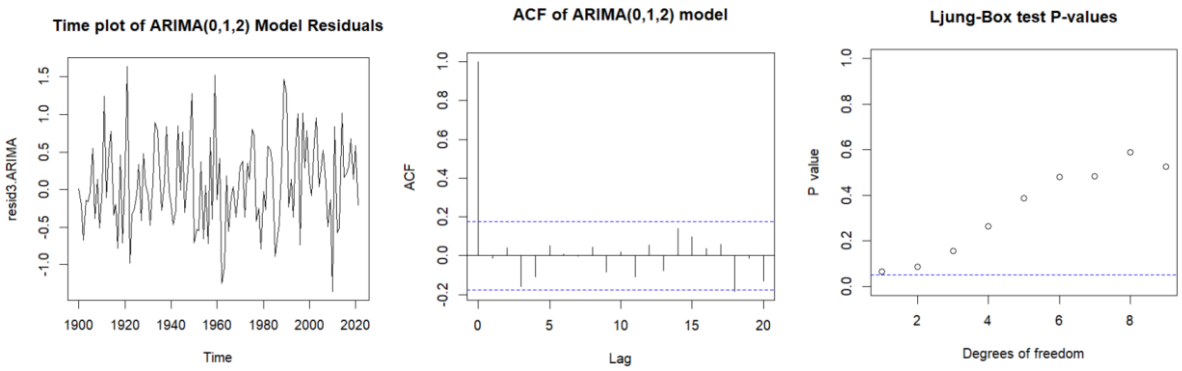
Result

The model ARIMA (1,1,1) was compared with the previously assessed model ARIMA (0,1,1) using the Akaike Information Criterion (AIC). With lower AIC values suggesting better fit, the models' relative goodness-of-fit was revealed by the AIC values. The AIC = 227.63 for ARIMA (1,1,1) model whereas AIC = 226.86 for ARIMA (0,1,1) model. Hence, ARIMA (0,1,1) is a better fit.

Model 3 - ARIMA (0,1,2) model

The ARIMA (0,1,2) model was used to analyse the annual mean temperature data for the English Midlands region. This model aims to capture both short-term dependencies and temporal patterns in the data using first-order differencing and two moving average terms.

To examine their behaviour over time, the residuals of the ARIMA (0,1,2) model were computed and shown. The residuals' autocorrelation function (ACF) was studied to find any residual autocorrelation, and their time plot was looked at to spot any patterns or trends.



To assess the residuals' independence at various lags, the Ljung-Box test was used. The p-values that were calculated for every lag were compared with the generally used significance level of 0.05 to ascertain whether there was a substantial amount of autocorrelation in the residuals.

The hypothesis test for MA (2) |-0.0847/ 0.0807|=1.049 (< 2), hence, this model should not be considered.

We fit an ARIMA (1,1,1) and an ARIMA (0,1,2) model as an additional check to see if adding more AR or MA terms could enhance the model fit. The AIC value of every one of these models is greater than that of the ARIMA (0,1,1) model. Furthermore, the ARIMA (0,1,1) model appears to be the better option based on the large p-values obtained from hypothesis tests conducted to determine whether extra terms should be included.

Conclusion

In summary, the annual mean temperature data for the Midlands region showed that the ARIMA (0,1,2) model was effective in capturing both temporal trends and short-term dependency. But the model's higher AIC value than the ARIMA (0,1,1) model implies that the latter would be a better choice for temperature variation predictions. Taking all factors into consideration, ARIMA (0,1,1) is the most suitable model fit for the Midlands region's annual mean temperature data.
The results of the evaluation of several ARIMA models indicated that the ARIMA(0,1, 1) model was the most suitable to describe the annual mean temperature data for th e England's Midlands region. The final fitted model has the following equation:

$$Yt= \theta et-1+et + c$$

where:

Yt = differenced series of annual mean temperatures.

θ = Coefficient of the lag residual term (et - 1).

et = the error term at time t.

c = constant term.

Executive Summary

Objective

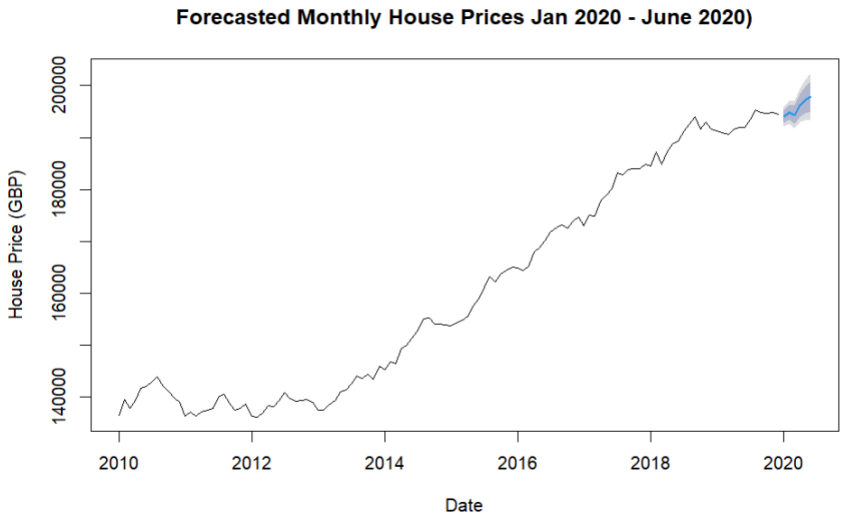
The goal of this analysis is to project monthly average home prices in the East Midlands area for the first half of 2020 by using historical data from January 2010 to December 2019.

Key Findings from the analysis

- Time series modelling approaches were required due to the dataset's non-stationarity, which was identified through an upward trend and seasonality during the exploratory study.
- The ARIMA (1, 1, 2) × (0, 1, 1)12 model was found to be the best appropriate after a thorough evaluation based on the Akaike Information Criterion (AIC) and Ljung-Box tests. It showed the lowest AIC value and good residual diagnostics.
- Using the chosen model for forecasting provides information about the expected direction of home prices over the next six months.

Forecasted House prices from January 2020 to June 2020:

The forecasted values depict an upwards trend beginning from January 2020 until June 2020. This suggests that the house prices will continue to increase in the first half of 2020. However, a slight dip in pricing is forecasted in March 2020 where the price drops to 194401.0 GBP from 194836.4 GBP in February 2020. Thereon, the values increase once again.
The time plot of the forecasted data can be visualized in a graph having X-axis as time in years vs Y axis showing House prices (GBP). The forecasted values are illustrated by a blue line.



Conclusion:

Using the ARIMA (1, 1, 2) × (0, 1, 1)12 model provides insightful predictions about how East Midlands house prices will develop in the future, giving the local government agency useful information to deal with upcoming obstacles and take advantage of new opportunities.

Report on Forecasting the Monthly House Prices in the East Midlands Region

Introduction

The state of the housing market affects many different sectors and stakeholders, making it an important gauge of the health and stability of the economy. Our goal in this analysis is to project East Midlands average house prices for the first half of 2020. We aim to offer useful insights for stakeholders, such as investors, legislators, and real estate experts, by utilizing time series analysis and statistical modelling approaches.

The monthly average house prices in the East Midlands from January 2010 to December 2019 are included in the dataset 'em_house_prices.csv'. This dataset is a useful tool for comprehending the trends in the local property market over the last ten years. The dataset comprises off 3 columns namely: month, year, average_price_gbp and a total of 120 observations. average_price_gbp gives insight on the monthly average house prices in Pounds (GBP) for properties in the East Midlands Region.

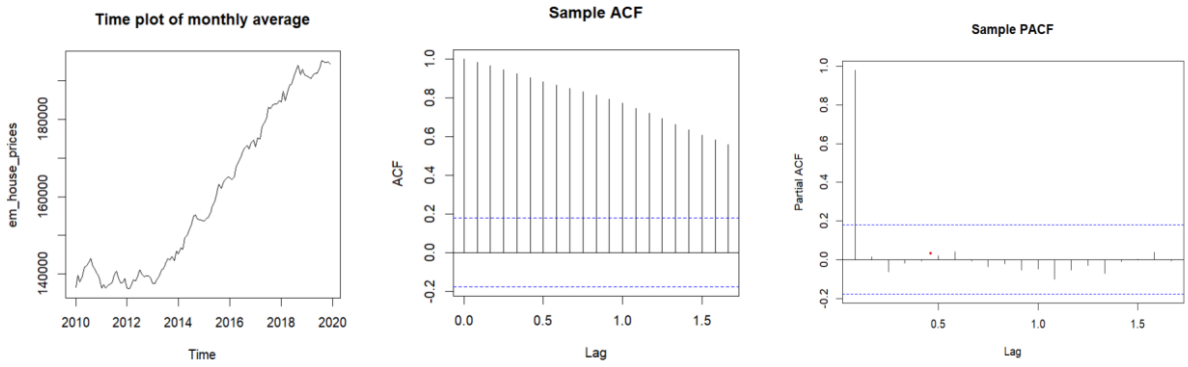
Significance

Policymakers, investors, real estate agents, and potential homeowners are among the many stakeholders who must comprehend the patterns and trends in housing pricing. Price fluctuations have an impact on the economy, investment choices, affordability of homes, and general market attitude. Our goal is to produce reliable forecasting models and actionable insights to help with well-informed housing market decision-making by evaluating this dataset.

Exploratory Data Analysis (EDA)

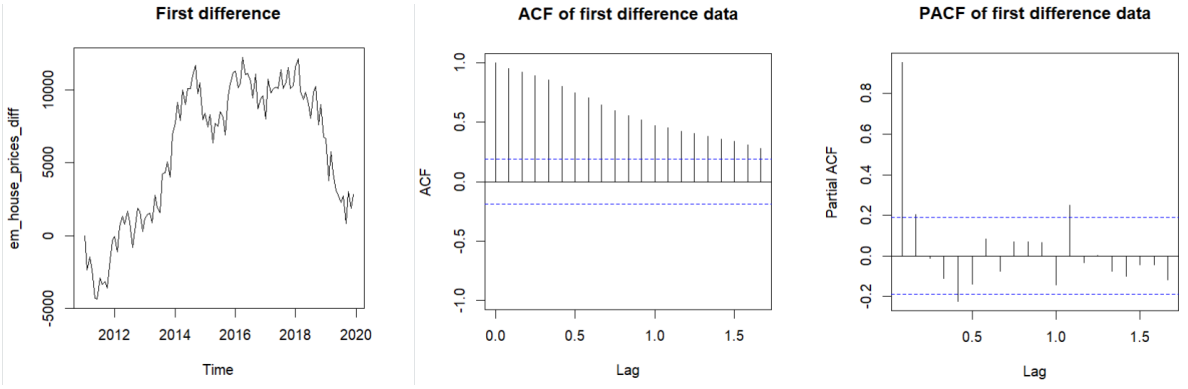
There are no missing values in the dataset, hence, the analysis can proceed to achieve accurate results. The monthly average prices are then transformed to time series data.

The next objective is to make sure the time series data is stationary. This is done as modelling non-stationary time series data leads to inaccurate estimates and unreliable forecasts. To determine stationarity, a time plot, sample Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) is plotted.

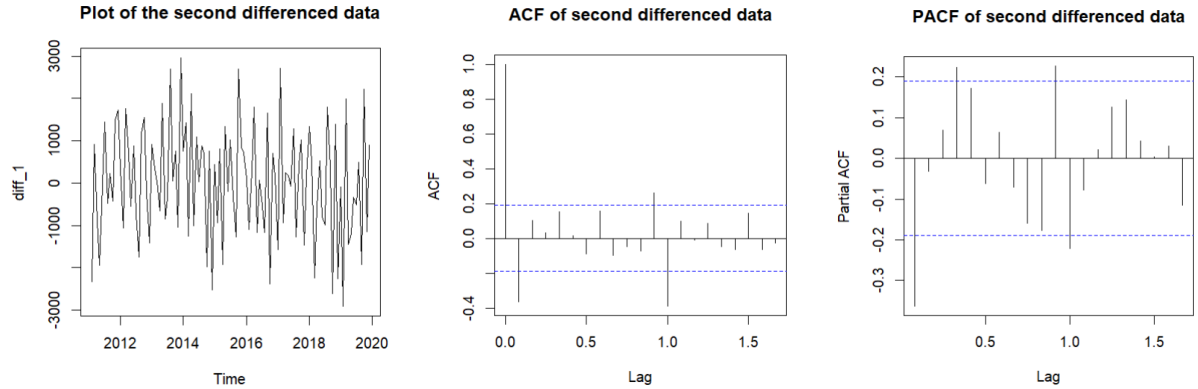


On visualizing the plots, there is a gradually increasing trend as well as seasonality depicted in the time plot which are characterisitics of a non-stationary data. The sample Autocorrelation Function (ACF) shows all values are significantly greater than zero which indicates strong correlation at all lags. The sample Partial Autocorrelation Function (PACF) values drop to zero after the first lag. Thus, the data is non-stationary.

The technique of Differencing is used to achieve stationary. A lag of 12 is applied to account for the monthly frequency of the data, and then the first difference is calculated. The first differenced data is then visualized using a time plot, sample Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).



The data still seems to be non-stationary as the time plot does not exhibit constant mean and seasonality. The plots of ACF and PACF tell us that there is strong autocorrelation at various lags indicating non stationary components in the data. As the data is still not stationary, we difference this data once again. Similar to the last observation, a time plot, the ACF and the PACF is plotted.



The second difference performed show much better results. On inspecting the Time plot, a constant mean and no seasonaltiy is observed. The ACF and PACF prove that there autocorrelation appears to be close to zero most lags, which indicates stationary data. Therefore, the time series data is now stationary and is suitable for further analysis. The ACF cuts off to zero after the first lag which suggests that an MA(1) model may be appropriate to model the data.

Model Evaluation

Various ARIMA models are fit to the data and their performance is evaluated. The following tests are calculated on each model to select which one fits the data best :

1]Akaike Information Criterion (AIC)

An indicator of a statistical model's relative quality for a certain dataset is its AIC. It penalizes for model complexity, promoting more straightforward models that can nevertheless describe the data well. A better balance between goodness of fit and model complexity is indicated by lower AIC values.

2]Ljung-Box test for Residual Autocorrelation

To determine whether a model's residuals show considerable autocorrelation, one can do the Ljung-Box test. It aids in assessing how well the model represents the temporal dependencies present in the data. The test's null hypothesis is the absence of residual autocorrelation. Significant residual autocorrelation is shown by a low p-value (< 0.05), which implies rejection of the null hypothesis.
The residuals are independent when the p-value is non-significant, indicating that the model does a good job of capturing the temporal dependencies in the data.

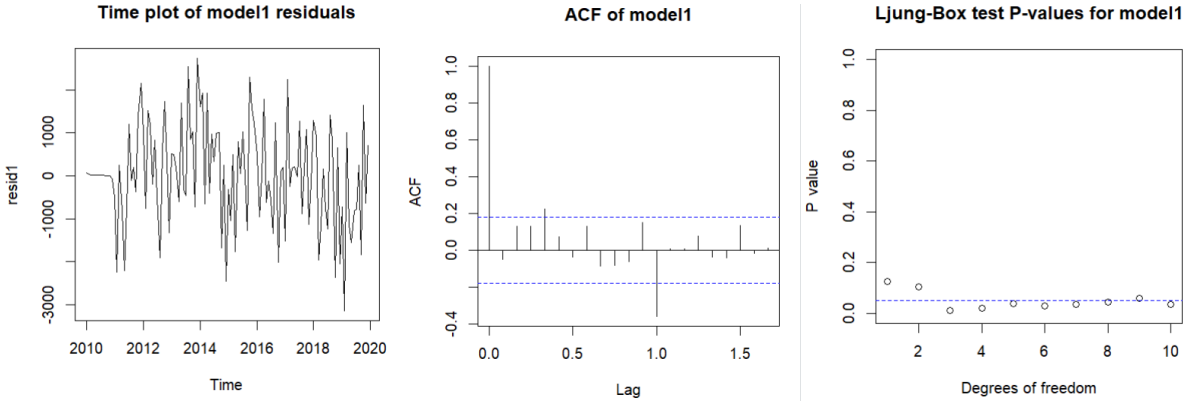
Model Selection

Models are thoroughly assessed by considering both the AIC values and the outcomes of the Ljung-Box test.

By choosing the model that best fits the data, the least amount of residual autocorrelation and AIC value are taken into consideration, which ensures a healthy balance between model simplicity and data dynamics accuracy.

Model Fitting

Model 1 - ARIMA (0,1,1) * (0,1,0)12 model

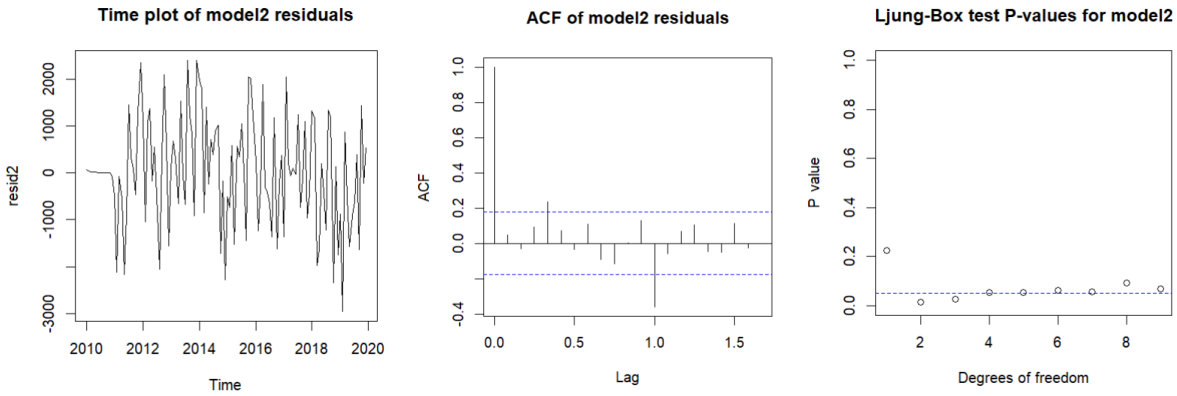


On examination of the model, the following observations are made:

- The AIC value for this model is 1832.14.
- The time plot of residuals for this model is similar to that of white noise.
- The Sample Autocorrelation(ACF) does not show a significant correlation after the first lag.
- The Ljung-Box test p-values are very low (<0.05) and below the threshold line, indicating there is residual independence.

Result: Due to the low p-values, the model is not a good fit for the data. Hence, alternative models should be explored to get an accurate estimate.

Model 2 - ARIMA (0,1,2) * (0,1,0)12 model

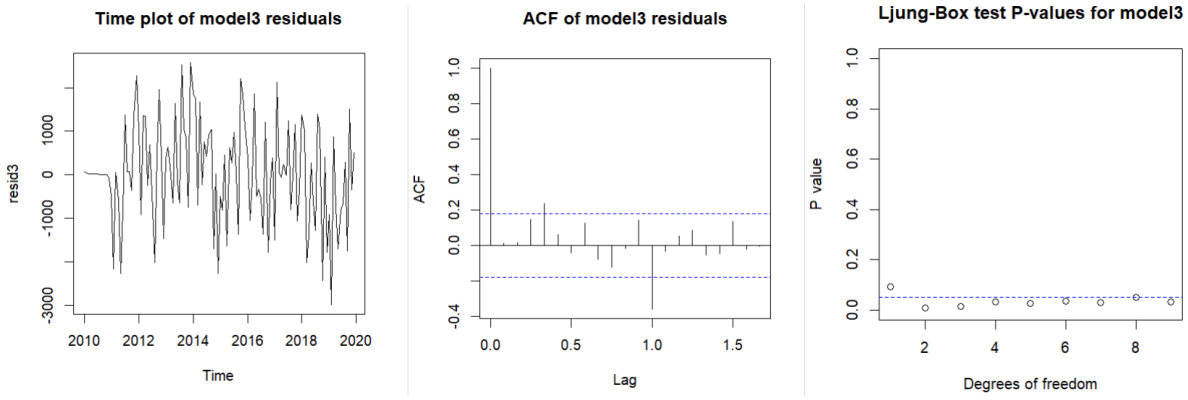


On examination of the model, the following observations are made:

- The AIC value for this model is 1830.26 which is lesser than that of Model1 (AIC = 1832.14).
- A Null Hypothesis test is calculated for MA(2) |0.1893/0.0960|= 1.9718 (< 2).
- The Ljung-Box test results do not reject the null hypothesis.
- The time plot of residuals does not show any significant pattern.
- The Sample Autocorrelation(ACF) does not show a significant correlation after the first lag.

Result: This model does not give any significant improvement over the previous model.

Model 3 - ARIMA (1,1,1) * (0,1,0)12 model

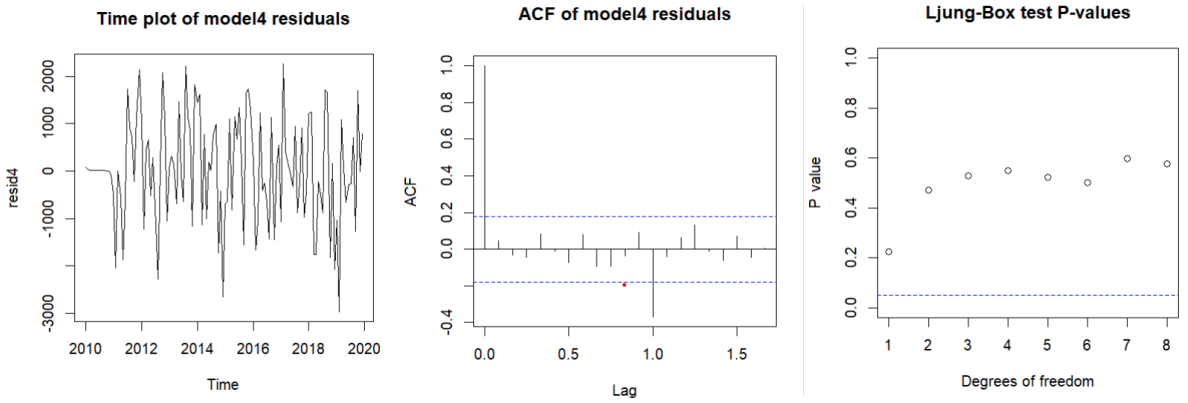


On examination of the model, the following observations are made:

- The AIC value for this model is 1831.9
- The time plot of residuals for this model is similar to that of white noise.
- The Sample Autocorrelation(ACF) does not show a significant correlation after the first lag.
- The Ljung-Box test p-values are very low (<0.05) and below the threshold line, indicating there is residual independence.

Result: Due to the low p-values and high AIC value, the model is not a good fit for the data.

Model 4 - ARIMA (1,1,2) * (0,1,0)12 model



On examination of the model, the following observations are made:

- The AIC value for this model is 1825.04 which is lesser than that of all the models tested until now. It indicates that this model might be a good fit.
- A Null Hypothesis test is calculated for MA(2) |0.5549/0.1003|= 5.532 (>2)
- The Ljung-Box test p-values are all greater than 0.05.
- The time plot of residuals does not show any significant pattern.
- The Sample Autocorrelation(ACF) does not show a significant correlation after the first lag.

Result: The ARIMA (1,1,2) * (0,1,0)12 model is a good fit for the data as it has a lesser AIC value, additionally, all the p-values are quite significant(>0.05). Thus, it is the best model fit when compared to the previous models, however, the remaining models are also tested in order to get the best fit.

Model 5 - ARIMA (2,1,1) * (0,1,0)12 model

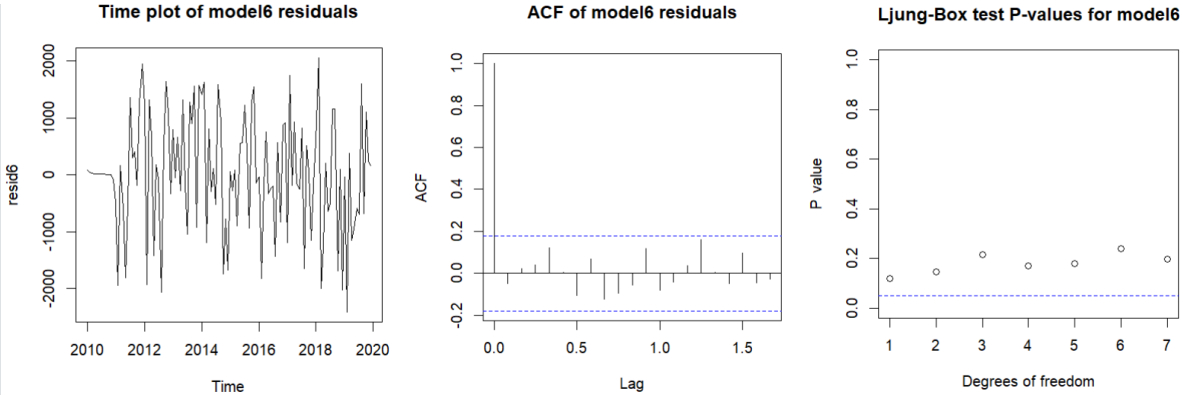
On examination of the model, the following observations are made:

The AIC value for this model is 1833.84 which is greater than our best model, ARIMA (1,1,2) * (0,1,0)12 model with AIC = 1825.04.

We know that models with higher AIC are not a good fit for the data.

Result: The ARIMA (2,1,1) * (0,1,0)12 model is not a good fit for the data.

Model 6 - ARIMA (1,1,2) * (1,1,0)12 model (Seasonality components with P=1)

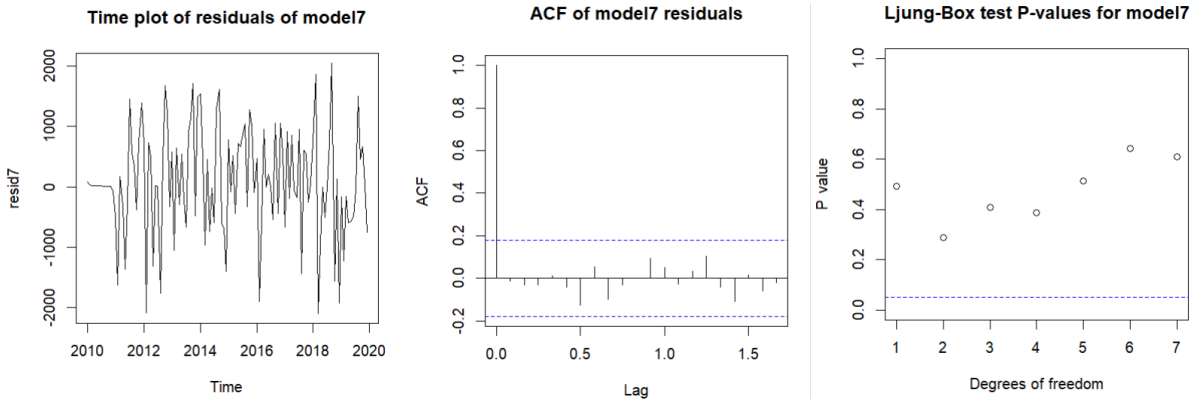


On examination of the model, the following observations are made:

- The AIC value for this model is 1810.82 which is lesser than that of all the models tested until now. It indicates that this model might be the best fit.
- A Null Hypothesis test is calculated for SAR1 |-0.4729/0.0944|= 5.0095(greater than 2)
- The Ljung-Box test p-values are all greater than 0.05.
- The time plot of residuals does not show any significant pattern.
- The Sample Autocorrelation(ACF) does not show a significant correlation after the first lag.

Result: The ARIMA (1,1,2) * (1,1,0)12 model is a good fit for the data as it has a lesser AIC value, additionally, all the p-values are quite significant(>0.05). So far, it is the best model fit found, however, the remaining models are also tested in order to get the best fit.

Model 7 - ARIMA (1,1,2) * (0,1,1)12 model(Seasonality components with Q = 1)

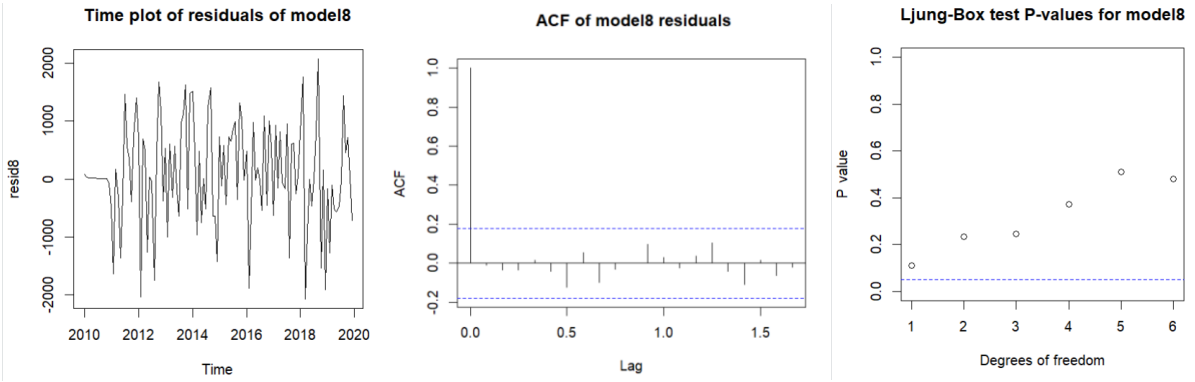


On examination of the model, the following observations are made:

- The AIC value for this model is 1790.89 which is lesser than that of all the models tested. It indicates that this model might be the best fit for the data.
- A Null Hypothesis test is calculated for SMA1 |-0.8109/0.1337|= 6.06(greater than 2)
- The Ljung-Box test p-values are all greater than 0.05.
- The time plot of residuals does not show any significant pattern.
- The Sample Autocorrelation(ACF) does not show a significant correlation after the first lag.

Result: The ARIMA (1,1,2) * (0,1,1)12 model is a good fit for the data as it has a smaller AIC value than the remaining, additionally, all the p-values are quite significant(>0.05). It is the best model fit found this far. Another model is studied before coming to a conclusion.

Model 8 - ARIMA (1,1,2) * (1,1,1)12 model (Seasonality components with P=1)



On examination of the model, the following observations are made:

- The AIC value for this model is 1792.74 which is greater than that of the best model we have (ARIMA (1,1,2) * (0,1,1)12 model with AIC = 1790.89).
- A Null Hypothesis test is calculated SAR1 |0.0575/0.1502|= 0.3838(less than 2)
- The Ljung-Box test p-values are all greater than 0.05.
- The time plot of residuals does not show any significant pattern.

- The Sample Autocorrelation(ACF) does not show a significant correlation after the first lag.

Result: The ARIMA (1,1,2) * (1,1,1)12 model is not a good fit for the data as its AIC value is less than the AIC value of the ARIMA (1,1,2) * (0,1,1)12 model (AIC value = 1790.89).

Conclusion of Model Fitting

- Variations in the non-seasonal ARIMA structure, including various combinations of AR and MA components, were investigated in Models 1 through 5. Although several models outperformed others in terms of AIC, no model offered a residual analysis-based fit that was deemed satisfactory.
- To account for any seasonality in the data, models 6 through 8 added additional seasonal components (seasonal AR and MA terms).
- Model selection was done using AIC values, where models with lower values were better fits. But additional procedures were also taken into consideration, such as the Ljung-Box test for residual autocorrelation.
- Based on the diagnostic tests that were sufficient for residual analysis and the lowest AIC value, Model 7 (ARIMA (1, 1, 2) × (0, 1, 1)12) was determined to be the best fit for the data.
- Model 7's ability to accurately identify both seasonal and non-seasonal patterns in the data suggests that it can be used to estimate house prices.

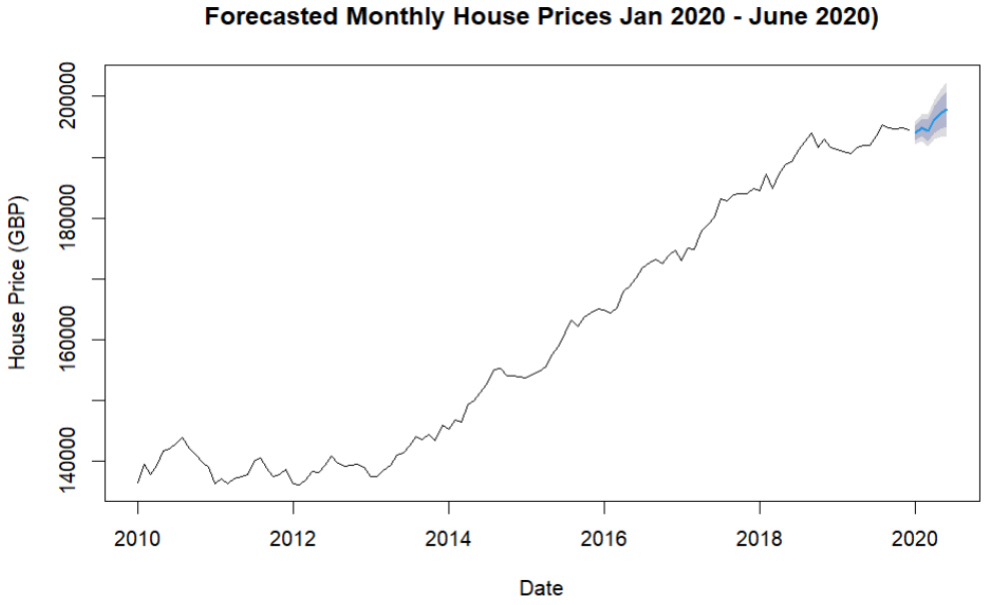
Forecasting the Monthly House Prices

Aim: To forecast the monthly house prices in the East Midlands Region from January 2020 to June 2020.

Model Selection
Model 7, an ARIMA (1, 1, 2) × (0, 1, 1)12 model, has been shown to have the best fit for the data after several models were analysed.

The ‘forecast’ package in R was used to predict future home prices. The chosen ARIMA model was used to create the predicted values for the upcoming six months. The projected figures shed light on the anticipated pattern of home prices between January 2020 and June 2020.

The projected trajectory of house prices throughout the forecast horizon was then visualized by plotting the forecasted values next to the historical data. Based on the chosen ARIMA model, this graphical depiction provides a clear picture of the expected movement of housing prices.



The following are the anticipated monthly home prices for the months of January through June 2020, along with the corresponding uncertainty intervals:

January 2020: The forecast for January is £193,930.5, with an 80% confidence interval ranging from £192,727.4 to £195,133.6 and a 95% confidence interval ranging from £192,090.5 to £195,770.5.

February 2020: The point forecast for February is £194,836.4, with an 80% confidence interval ranging from £193,413.9 to £196,258.9 and a 95% confidence interval ranging from £192,660.8 to £197,012.0.

March 2020: The point forecast for March is £194,401.0, with an 80% confidence interval ranging from £192,656.7 to £196,145.3 and a 95% confidence interval ranging from £191,733.3 to £197,068.7.

April 2020: The point forecast for April is £196,204.1, with an 80% confidence interval ranging from £194,073.5 to £198,334.7 and a 95% confidence interval ranging from £192,945.6 to £199,462.5.

May 2020: The point forecast for May is £197,202.1, with an 80% confidence interval ranging from £194,649.2 to £199,754.9 and a 95% confidence interval ranging from £193,297.8 to £201,106.3.

June 2020: The point forecast for June is £197,933.9, with an 80% confidence interval ranging from £194,940.8 to £200,926.9 and a 95% confidence interval ranging from £193,356.4 to £202,511.4.

The estimated prices depict an increasing trend in the prices from January to June 2020. However, a slight decline is foreseen during March 2020 wherein the house prices drop to 194401.0 GBP. Thereafter the price increases April 2020 onwards. All things considered, the ARIMA (1, 1, 2) × (0, 1, 1)12 model offers a solid foundation for predicting home values and insightful information for real estate market decision-making.

Conclusion

The goal of the investigation was to determine which ARIMA model fit the data the best for predicting monthly home prices. We fitted and assessed several ARIMA models using residual diagnostics, Ljung-Box tests, and AIC values. Among the models that were examined, ARIMA(1, 1, 2) × (0, 1, 1)12 was found to be the most appropriate model because it had the lowest AIC value and all of the p-values were more than 0.05, which indicated that the model fit the data well.

Next, for the months of January 2020 through June 2020, monthly housing prices were predicted using the chosen model. While there was significant diversity in the predicted values, there was no discernible pattern for the given time frame.

Based on the provided dataset and conducted research, the ARIMA (1, 1, 2) × (0, 1, 1)12 model offers a dependable framework for predicting monthly property prices. But it's important to understand that other variables, like the state of the economy and the dynamics of the housing market, that are outside the purview of this analysis may also have an impact on future home prices. Consequently, to guarantee accurate forecasts in a dynamic context, continuous monitoring and model upgrades might be required.

Appendix

```
#---------- Question 1 ----------
#The Task is to analyse this dataset by fitting a suitable time
#series model to describe the data.

#Reading the cet_temp.csv dataset
cet_df<-read.csv("cet_temp.csv",header=TRUE)

#Printing the data
print(cet_df)

#Setting the annual mean temperature to time series data
cet_temp<- ts(cet_df$avg_annual_temp_C,start=1900,frequency=1)

#Produce a time plot of the simulated data 'cet_temp'.
ts.plot(cet_temp, main = "Time Plot of annual temperature")

#Set the R graphics device to contain two plots (1 row, 2 columns)
x11()
par(mfrow=c(1,2))

#Plot the sample ACF
acf(cet_temp, main = "Sample ACF of CET Temperature Data")


#Plot the sample PACF
pacf(cet_temp,main = "Sample PACF of CET Temperature Data")

#On visualizing the plots to check stationarity, it is seen
#that the process does not appear to be stationary.
#The time plot depicts an increasing trend and has variation over time
```

```
#whereas the sample ACF values do not decline sharply as the lag increases.
#The pacf does not cut off after lag 1

#Hence, to achieve stationarity the data is differenced
cet_temp_diff<- diff(cet_temp)
# Plot the differenced data
ts.plot(cet_temp_diff, main = "Time Plot of differenced Temperature Data")

x11()
par(mfrow=c(1,2))
#Plot the ACF for 'cet_temp_diff'
acf(cet_temp_diff, main = "Sample ACF of Differenced Temperature Data")

#Plot the PACF for 'cet_temp_diff'
pacf(cet_temp_diff,main = "Sample PACF of Differenced Temperature Data")

#In comparison, the above plots show that the first difference
#of the data is (weakly) stationary.
#The time plot illustrates constant mean and variability over time.
#The ACF declines rapidly to zero after lag 0.
#Similarly, the PACF also tends to zero as the lag increases.
#This is indicative of an MA process

#Begin with an ARIMA(0,0,1) process

#Fit an ARIMA(0,0,1) model to the data
model0.MA1<- arima(cet_temp,order=c(0,0,1),method="ML")

#Print the ARIMA model summary
model0.MA1
```

```
#Calculate the residuals of the ARIMA(0,0,1) model
resid0<-residuals(model0.MA1)

#Plot the time series of the residuals
ts.plot(resid0,main='Time plot of MA(1) Model Residuals')

#Plot ACF of the residuals
acf(resid0,main='ACF of ARIMA(0,0,1) model')

#Function to produce P-values for the Ljung-Box test for different lags
#where an ARMA(p,q) model has been fitted.
#Note that k must be > p+q
LB_test<-function(resid,max.k,p,q){
  lb_result<-list()
  df<-list()
  p_value<-list()
  for(i in (p+q+1):max.k){
    lb_result[[i]]<-Box.test(resid,lag=i,type=c("Ljung-Box"),fitdf=(p+q))
    df[[i]]<-lb_result[[i]]$parameter
    p_value[[i]]<-lb_result[[i]]$p.value
  }
  df<-as.vector(unlist(df))
  p_value<-as.vector(unlist(p_value))
  test_output<-data.frame(df,p_value)
  names(test_output)<-c("deg_freedom","LB_p_value")
  return(test_output)
}

# Compute Ljung-Box test p-values for an ARMA(0,1) model with max lag 11
MODEL0.LB<-LB_test(resid0,max.k=11,p=0,q=1)
#Print the p-values for each lag
```

```
MODEL0.LB


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

plot(MODEL0.LB$deg_freedom,MODEL0.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P□value",main="Ljung-Box test P-values",ylim=c(0,1))

#Adding a blue dashed line at y = 0.05

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


#The time plot does is not stationary for above model.

#The sample ACF of residuals does not cut off to zero as lag increases.

#Additionally, the p-values are below the threshold off 0.05.

#This suggests that this model might not be a good fit.


#Begin with an ARIMA(0,1,1) process


#Fit an ARIMA(0,1,1) model to the data

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


#Print the summary of the ARIMA model

model1.MA1


#Calculate the residuals of the ARIMA(0,1,1) model

resid.MA1<-residuals(model1.MA1)

#Time plot of the residuals

ts.plot(resid.MA1,main='Time plot of ARIMA(0,1,1) Model Residuals')

#Plot ACF of the residuals

acf(resid.MA1,main='ACF of ARIMA(0,1,1) Model Residuals')


#The time plot of the residuals looks similar to white noise.

#The sample ACF seems to be close to zero at all lags > 0.

#Hence,the residuals are independent.
```

```
#Applying the LB_test function to the residuals of the ARIMA(0,1,1) model
MA1.LB<-LB_test(resid.MA1,max.k=11,p=0,q=1)
#Print the p-values for each lag value.
MA1.LB

#Produce a plot of the P-values against the degrees of freedom
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))
#Add a blue dashed line at y=0.05
abline(h=0.05,col="blue",lty=2)

#It is noted that all p-values are > 0.05 (and hence non-significant at the 5% level)
some
#The p-values are fairly small and we may
#consider fitting an alternative model to the data.

#Fitting an ARIMA(1,1,1) model to the data
model2.ARIMA<- arima(cet_temp,order=c(1,1,1),method="ML")
#Print the summary of the ARIMA model
model2.ARIMA

#Calculate the residuals of the ARIMA(1,1,1) model
resid2.ARIMA<-residuals(model2.ARIMA)
#Time plot of the residuals
ts.plot(resid2.ARIMA,main='Time plot of ARIMA(1,1,1) Model Residuals')
#Plot ACF of the residuals
acf(resid2.ARIMA, main = "ACF of ARIMA(1,1,1) Model")

#Applying the LB_test to the residuals of the ARIMA(1,1,1) model
ARIMA.LB<-LB_test(resid2.ARIMA,max.k=11,p=1,q=1)
```

```
ARIMA.LB


#Produce a plot of the P-values against the degrees of freedom

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))

#Add a blue dashed line at 0.05

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

par(mfrow=c(1,1))


#aic = 227.63 FOR ARIMA(1,1,1) MODEL

#aic = 226.86 FOR ARIMA(0,1,1) MODEL


#The test statistic for a test of the hypotheses: H0 : φ = 0 versus

#H1 : φ != 0 is 0.1137/0.1026 = 1.1081 which is NOT greater than 2.

#Hence we do not reject H0 and conclude that the

#AR term should NOT be included in the model


#Fitting an ARIMA(0,1,2) Model

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

model3.ARIMA



#Calculate the residuals of the ARIMA(0,1,2) model

resid3.ARIMA<-residuals(model3.ARIMA)

#Time plot of the residuals

ts.plot(resid3.ARIMA,main='Time plot of ARIMA(0,1,2) Model Residuals')

#Plot ACF of the residuals

acf(resid3.ARIMA,main = "ACF of ARIMA(0,1,2) model")


#Applying the LB test to residuals of the ARIMA(0,1,2) model

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

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))

#Add a blue dashed line at 0.05

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


#aic = 227.77 FOR ARIMA(0,1,2) MODEL

#ma2 |-0.0847/ 0.0807|=1.049 (less than 2.....should not be included)


#As a further check we consider fitting an ARIMA(1,1,1) and an ARIMA(0,1,2)

#model to see if the addition of further AR or MA terms might improve the model

#fit. Each of these models has an AIC value higher than the ARIMA(0,1,1) model.

#Moreover, hypothesis tests to check whether or not additional terms should be
included result in large P-values,

#thereby suggesting that the ARIMA(0,1,1) model is preferable.


#At this point, choose the ARIMA(0,1,1) model as the most appropriate

#for the above dataset.



#---------- Question 2 ----------

#The Task is to produce a time series model to forecast the monthly house prices

#for the first six months of 2020.


#Read the 'em_house_df' dataset

em_house_df<-read.csv("em_house_prices.csv",header=TRUE)


#Print the data

print(em_house_df)
```

```
#Set the monthly average price to time series data
em_house_prices<- ts(em_house_df$average_price_gbp,start=2010,frequency=12)

#Produce a time plot of the simulated data
ts.plot(em_house_prices, main = "Time plot of monthly average")

#The data is not stationary as there is an upward
#trend(increasing mean) and seasonality present.

#Plot the sample ACF
acf(em_house_prices,main = "Sample ACF")

#The sample ACF does not cut off to zero

#Plot the sample PACF
pacf(em_house_prices, main = "Sample PACF")

#The sample PACF drops to zero after the first lag.

#In order to achieve stationarity, the data is differenced
#A first order difference is used.

#As the data have been collected monthly
#The first difference is done with lag 12.
em_house_prices_diff<- diff(em_house_prices, lag = 12)

#Plot the first-differenced data
ts.plot(em_house_prices_diff,main = "First difference")

#Even though there isn't any seasonality, the mean is still
```

```
#not constant which suggests its not stationary.

#Plot the ACF of the first-differenced data
acf(em_house_prices_diff,ylim=c(-1,1), main = "ACF of first difference data")

#Plot the PACF of the first-differenced data
pacf(em_house_prices_diff, main = "PACF of first difference data")

#These plots illustrate that the first differenced data is not stationary.
#A first difference of the seasonally differenced data is done to
#achieve stationarity.
diff_1<- diff(em_house_prices_diff)

#Plot the second differenced data
ts.plot(diff_1, main= "Plot of the second differenced data")

#Plot the ACF of the differenced data
acf(diff_1, main = "ACF of second differenced data")
#Plot the PACF of the differenced data
pacf(diff_1, main = "PACF of second differenced data")

#On observing the above plots the time series appears to be stationary
#as it has a constant mean, no seasonality and constant variance over time.

#The sample ACF may cut off to zero after lag 1, which suggests that an
#MA(1) model may be used.

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

model1 <- arima(em_house_prices,order=c(0,1,1),
          seasonal=list(order=c(0,1,0), period=12),
```

```
        method="ML")
model1

#Note that the model1 aic = 1832.14

#Calculate the residuals of the model
resid1<-residuals(model1)
#Time plot of the residuals
ts.plot(resid1, main = "Time plot of model1 residuals")

#Plot ACF of model1 residuals
acf(resid1, main = "ACF of model1")



#Function to produce p-values for the Ljung_Box test for different lags
#where an ARIMA(p,d,q)x(P,D,Q)_h model has been fitted.
#Note that k must be > p+q+P+Q
#Number of degrees of freedom for the test = k-p-q-P-Q

#Arguments for the function "LB_test"
#resid = residuals from a fitted ARIMA(p,d,q)x(P,D,Q)_h model

#max.k = the maximum value of k at which we perform the test
#Note that the minimum k is set at p+q+P+Q+1 (corresponding to a test with one
degree
#of freedom)

#p = Order of the non-seasonal AR part of the model
#q = Order of the non-seasonal MA part of the model
#P = Order of the seasonal AR part of the model
#Q = Order of the seasonal MA part of the model
```

```
#The function returns a table with one column showing the number of degrees
#of freedom for the test and the other the associated P-value.
LB_test_SARIMA<-function(resid,max.k,p,q,P,Q){
  lb_result<-list()
  df<-list()
  p_value<-list()
  for(i in (p+q+P+Q+1):max.k){
    lb_result[[i]]<-Box.test(resid,lag=i,type=c("Ljung-Box"),fitdf=(p+q+P+Q))
    df[[i]]<-lb_result[[i]]$parameter
    p_value[[i]]<-lb_result[[i]]$p.value
  }
  df<-as.vector(unlist(df))
  p_value<-as.vector(unlist(p_value))
  test_output<-data.frame(df,p_value)
  names(test_output)<-c("deg_freedom","LB_p_value")
  return(test_output)
}



#Perform Ljung-Box tests for the model1 residuals
model1.LB<-LB_test_SARIMA(resid1, max.k=11, p=0, q=1, P=0, Q=0)
model1.LB


#Produce a plot of the p-values against the degrees of freedom
plot(model1.LB$deg_freedom,model1.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P□value",main="Ljung-Box test P-values for model1",ylim=c(0,1))
#Add a blue dashed line at 0.05
abline(h=0.05,col="blue",lty=2)


#The time plot of the residuals appears to look similar to that for white noise.
```

```
#The sample ACF seems to be close to zero at all lags > 0.
#Therefore, the residuals are independent.
#The p-values are also quite less.
#An alternative model should be considered to fit the data.

#Fitting an ARIMA(0, 1, 2) × (0, 1, 0)12 model.
model2 <- arima(em_house_prices,order=c(0,1,2),
            seasonal=list(order=c(0,1,0), period=12),
            method="ML")
model2

#Note that the model2 aic = 1830.26

#Calculate the residuals of model2
resid2<-residuals(model2)
#Time plot of model2 residuals
ts.plot(resid2,main = "Time plot of model2 residuals")

#Plot ACF of model2 residuals
acf(resid2, main = "ACF of model2 residuals")

#Ljung-Box tests for model2 residuals
model2.LB<-LB_test_SARIMA(resid2, max.k=11, p=0, q=2, P=0, Q=0)
model2.LB

#Produce a plot of the p-values against the degrees of freedom
plot(model2.LB$deg_freedom,model2.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P□value",main="Ljung-Box test P-values for model2",ylim=c(0,1))
#Add a blue dashed line at 0.05
abline(h=0.05,col="blue",lty=2)
```

```
#Even though the model1 aic = 1832.14 is greater than model2 aic = 1830.26,
#hypothesis test for ma2 |0.1893/0.0960|= 1.9718 (which is less than 2)
#Do not reject H0: theta2=0
#Hence ARIMA(0, 1, 2) × (0, 1, 0)12 model is not the best fit for the data.



#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

#Note that model3 aic = 1831.9

#Calculate the residuals of model3
resid3<-residuals(model3)
#Time plot of model3 residuals
ts.plot(resid3,main = "Time plot of model3 residuals")

#Plot ACF of model3 residuals
acf(resid3, main = "ACF of model3 residuals")

#Ljung-Box test for model3 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
plot(model3.LB$deg_freedom,model3.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P□value",main="Ljung-Box test P-values for model3",ylim=c(0,1))
#Add a blue dashed line at 0.05
abline(h=0.05,col="blue",lty=2)
```

```
#The model3 aic = 1831.9, however its p-values are fairly small,
#hence, ARIMA(1, 1, 1) × (0, 1, 0)12 model is not a good fit.

#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

#Note that model4 AIC = 1825.04

#Calculate the residuals of model4
resid4<-residuals(model4)
#Time plot of model4 residuals
ts.plot(resid4, main= "Time plot of model4 residuals")

#Plot ACF of model4 residuals
acf(resid4, main = "ACF of model4 residuals")

#Ljung-Box test for model4 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
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))
#Add a blue dashed line at 0.05
abline(h=0.05,col="blue",lty=2)

#The model4 aic = 1825.04 is lesser than model1 aic = 1832.14
```

```
#Hypothesis test for ma2 |0.5549/0.1003|= 5.532 (greater than 2)
#All p values > 0.05
#Therefore, ARIMA(1, 1, 2) × (0, 1, 0)12 model is the best fit until now.
#Howver more models can be checked for best result.

#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

#Note that model5 aic = 1833.84

#The current best model(model4 - aic = 1825.04) is lesser than that of model 5 - aic =
1833.84
#Therefore, ARIMA(2, 1, 1) × (0, 1, 0)12 model is not a good fit.
#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

#Note that model6 aic = 1810.82

#Calculate the residuals of model6
resid6<-residuals(model6)
#Time plot of model6 residuals
ts.plot(resid6, main = "Time plot of model6 residuals")

#Plot ACF of model6 residuals
```

```
acf(resid6, main = "ACF of model6 residuals")

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

#Produce a plot of p-values against the degrees of freedom
plot(model6.LB$deg_freedom,model6.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P□value",main="Ljung-Box test P-values for model6",ylim=c(0,1))
#Add a blue dashed line at 0.05
abline(h=0.05,col="blue",lty=2)

#The model6 aic = 1810.82 is lesser than model4 aic = 1825.04,
#Hypothesis test for SAR1 |-0.4729/0.0944|= 5.0095(greater than 2)
#All thep values > 0.05

#Therefore, the best model until now is model6.

#Similarly, try 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

#Note that model7 aic = 1790.89

#Calculate the residuals of model7
resid7<-residuals(model7)
#Time plot of Residuals of model7
ts.plot(resid7, main = "Time plot of residuals of model7")
```

```
#Plot sample ACF of model7 residuals
acf(resid7, main = "ACF of model7 residuals")

#Ljung-Box test for model7 residuals
model7.LB<-LB_test_SARIMA(resid7, max.k=11, p=1, q=2, P=0, Q=1)
model7.LB

#Produce a plot of the p-values against the degrees of freedom
plot(model7.LB$deg_freedom,model7.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P□value",main="Ljung-Box test P-values for model7",ylim=c(0,1))
#Add a blue dashed line at 0.05
abline(h=0.05,col="blue",lty=2)

#The model7 aic=1790.89 is lesser than model6 aic=1810.82
#Hypothesis test for SMA1 |-0.8109/0.1337|= 6.06(greater than 2)
#All p values > 0.05
#Model7 is the best fit model uptil now

#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

#Note that model8 aic = 1792.74

#Calculate the model8 residuals
resid8<-residuals(model8)
#Time plot of residuals of model8
ts.plot(resid8, main = "Time plot of residuals of model8")
```

```
#Plot ACF of model8 residuals
acf(resid8, main =" ACF of model8 residuals")

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

#Produce a plot of the p-values against the degrees of freedom
#Add a blue dashed line at 0.05
plot(model8.LB$deg_freedom,model8.LB$LB_p_value,xlab="Degrees of
freedom",ylab="P□value",main="Ljung-Box test P-values for model8",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

#The model8 aic = 1792.74 is greater than model7 aic = 1790.89

#Hypothesis test for SAR1 |0.0575/0.1502|= 0.3838(< 2)
#Hence, sar1 should not be included
#All p values > 0.05
#Model8 is not a good fit

#On analysis of multiple models, model 7 seems to be the best fit for the data with
#The least AIC and all p-values > 0.05.
#ARIMA(1, 1, 2) × (0, 1, 1)12 model is best fit.

#Forecasting using the selected ARIMA model
install.packages("forecast")

#Access the library forecast
library(forecast)

#Forecasting values for the next 6 months
```

```
forecast_values <- forecast(model7, h = 6)


#Display the forecasted values

print(forecast_values)


#Plot the forecasted values along with the the data

plot(forecast_values, main = "Forecasted Monthly House Prices Jan 2020 - June
2020)",

    xlab = "Date", ylab = "House Price (GBP)")
```