**Coursework Report – Annual Temperature Model Fitting**

**MATH4022 – Time Series and Forecasting**

**Student ID: 20387234**

**Introduction to the report:**

This report analyses a dataset that contains the annual mean temperatures in degree Celsius for the years 1900 to 2021 for the Midlands region of England, recorded by the UK Metrological Office Hadley Climate Centre and is sourced from the website - [www.metoffice.gov.uk/hadobs/hadcet](http://www.metoffice.gov.uk/hadobs/hadcet). We are expected to infer the conclusions from the following actions that will be conducted on the dataset:

* **Checking for Stationarity of the Time Series**: We will check if the Time series is stationary or not stationary. If the time series is not stationary we will perform differencing on the time series data to make it stationary and hence, feasible for further evaluation.
* **Differencing**: We will difference the data if the data that we are dealing with is not stationary. After doing that we will replot our time series data and check for stationarity and form an opinion on it by checking on the ACF and PACF and where it cuts off. If the data is still not stationary and ACF takes a lot of time to decay or does not decay at all , we will repeat this step.
* **Auto-Correlation Function:** We will check the ACF of the time series and check for the presence of MA(Moving Average) components for our possible fitted model.
* **Partial Auto-Correlation Function**: We will check the PACF of the time series and check for the presence of AR(Auto Regressive) components for our possible fitted model.
* **Finalising a resultant model fit and verification**: We will fit the data based on the steps that we have conducted above.

**Initial look at the Dataset:**

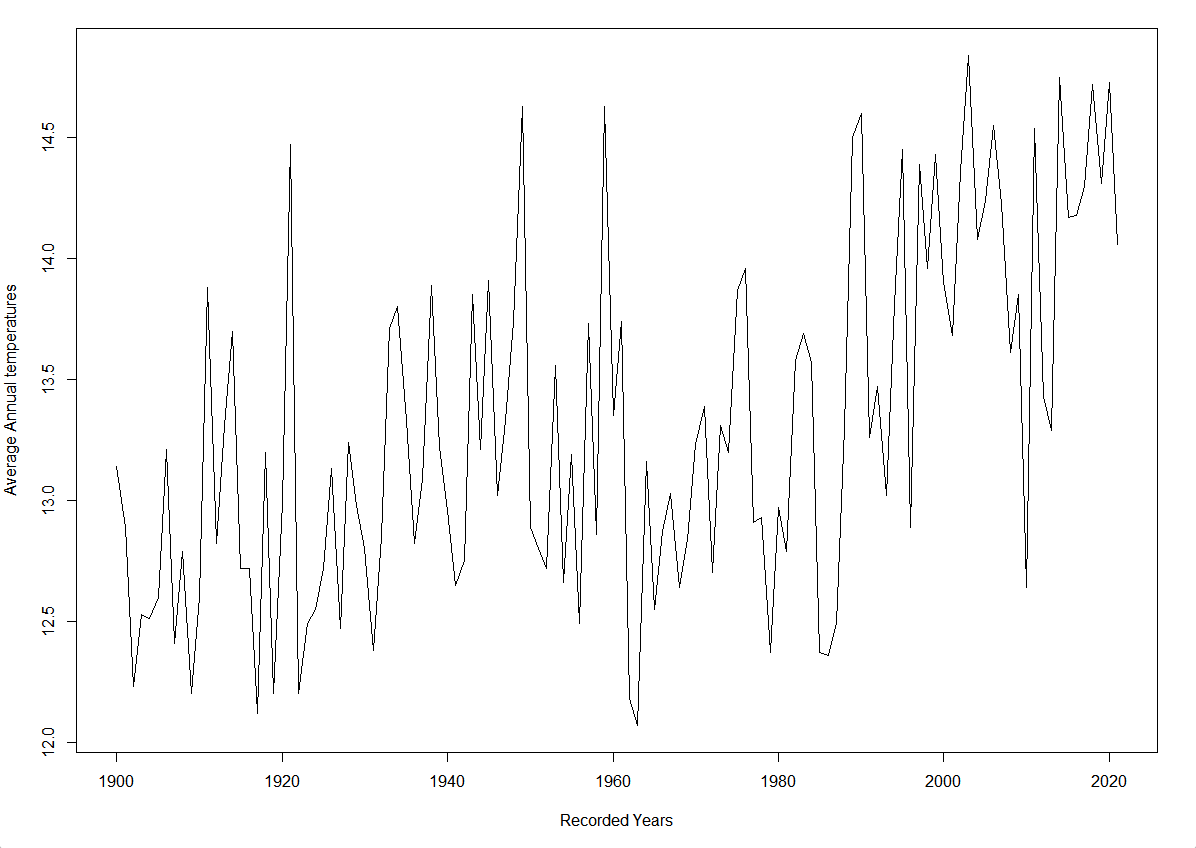
The dataset seems to have two columns that are as follows:

1. **Year:** The year when the temperature was actually measured
2. **Annual-Temperature**: The temperature measured that corresponds to that particular year.

After taking a look at our dataset it will be safe to continue and perform initial stationarity checks and check for ACF and PACF as we continue towards finding a suitable model for our time series data.

**Stationarity Checks:**

We can start checking for stationarity by plotting a time plot on the time series data. The time plot is shown below as follows:



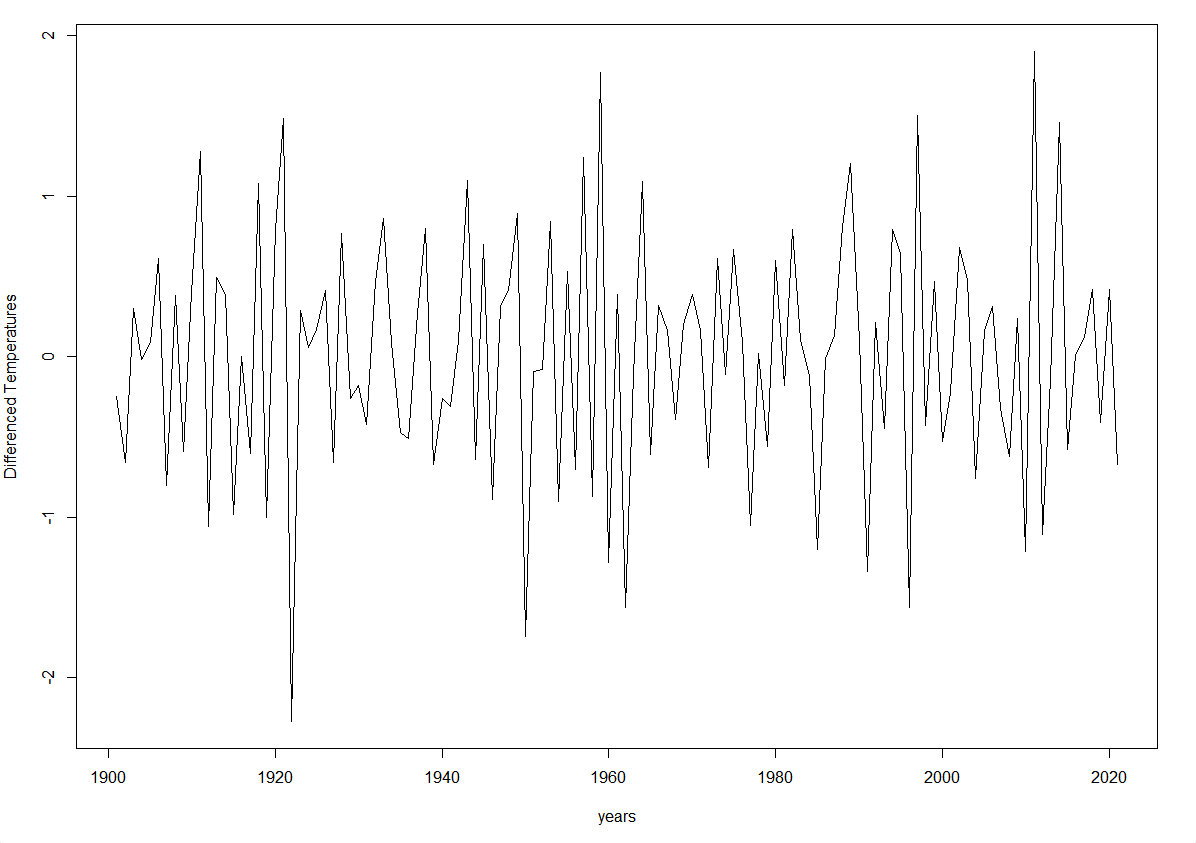
After taking a look at the time plot we can make a note of the following things:

* This plot is not stationary with a non-constant mean and variance and as a result we will need to perform differencing on this data.
* As the years go by and we enter the 20th Century we see glimpses of possible trends in the time plot.

The main inference to take from here would probably be that since this data is following an increasing trend and the mean and variance of the data is not constant, it is not stationary and we will need to perform first order of differencing to convert this data into non-stationary and then check for the ACF and PACF for the time series plot.

**Differencing:**

After the stationary checks are complete and we find that the data we are dealing with is not stationary by nature, we will then use first order of differencing on it and check if the data is stationary now.



After taking a look at this time plot, it will be safe to infer that the time plot is stationary with a constant mean. Hence, differencing the time series worked and we have successfully converted the time series data from non-stationary to stationary.

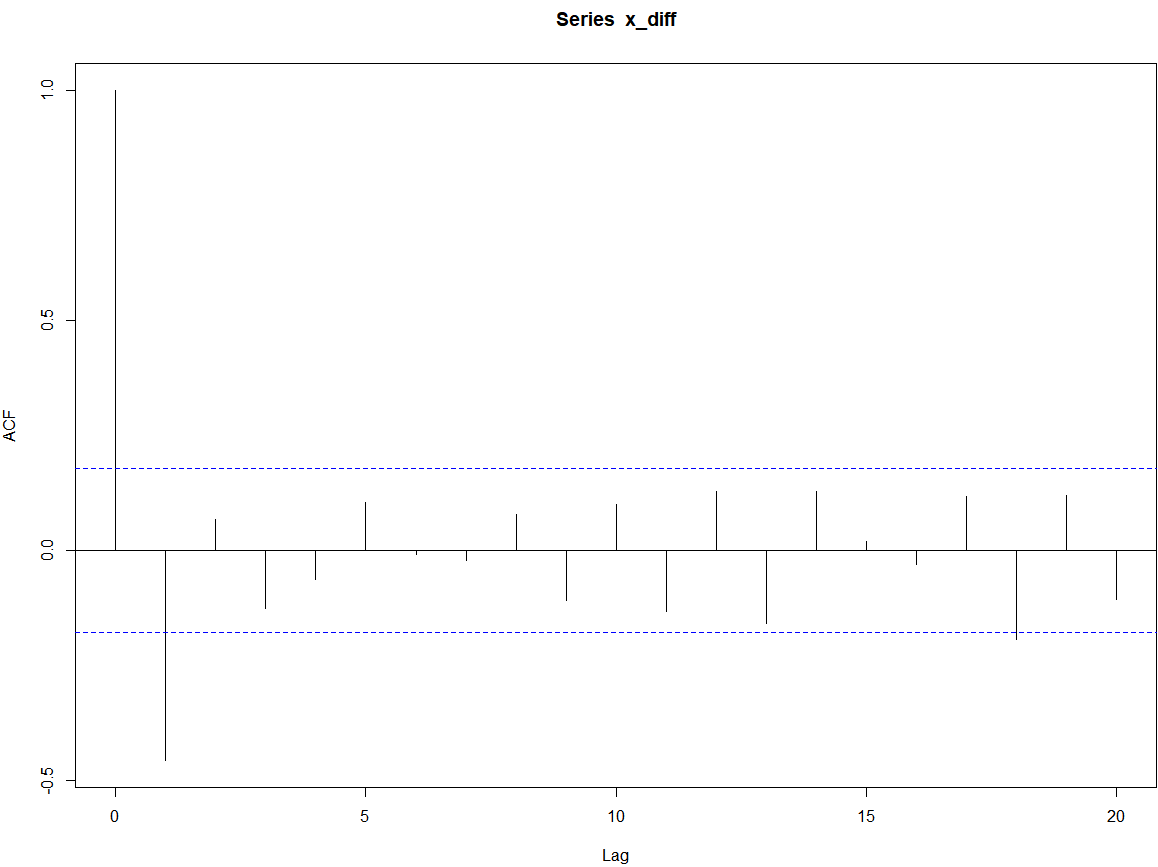
Now we can further our analysis by using this plot and calculating the acf and pacf for the same.

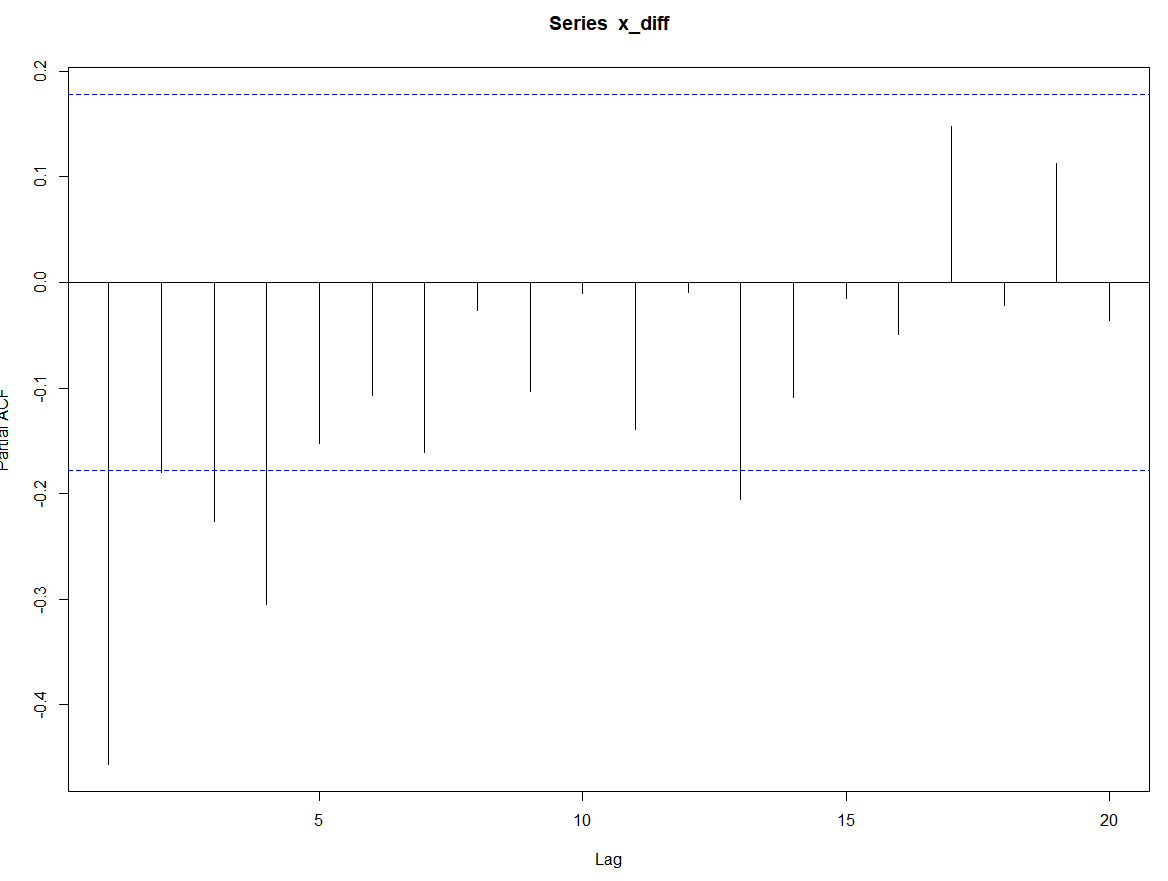
**AUTO-CORRELATION AND PARTIAL AUTO-CORRELATION FUNCTION:**

It is important to talk about these functions because of the following reasons:

* For a weakly stationary time series, the auto correlation function will decay pretty quickly and will show us the presence of a moving average component
* In the case of partial auto correlation function cutting off pretty quickly, it will show us the presence of an auto regressive component.

Now we will keep these things in mind and take a look at the ACF and PACF of our time series

****



From the following acf and pacf plots we can make the following assumptions:

* The ACF decays off fairly quickly on an early lag 1 which not only denotes a moving average component but it also denotes that this is a MA(1) process.
* The PACF does not seem to decay off at all which indicates an absence of any auto regressive components.

We now have enough information to fit a subsequent model for this time series data

**FITTING A MODEL:**

A brief overview of what we have done so far could be summarised as follows:

* The data was non-stationary so we differenced it and converted it to a stationary time series model
* After the conversion we can infer from the ACF and PACF that since the ACF decays off fairly quickly there is a possible presence of MA components and absence of AR components because the PACF does not decay at all.

**The model that can be suggested based on these points will be a pure MA(1) model or in the case of denoting it as an ARIMA model we can write it as (p,d,q) where p and q are AR and MA components and d is the times its differenced. So it will be a (0,1,1) model since it was differenced one time and it has one MA component.**

**Model-Checking:**

To check if the chosen model fits the data perfectly, we will find the theoretical ACF and fit it along the lines of the sample ACF. To find the theoretical ACF, we would need to find the parameter values for our MA(1) process that we have chosen as our suitable fit.

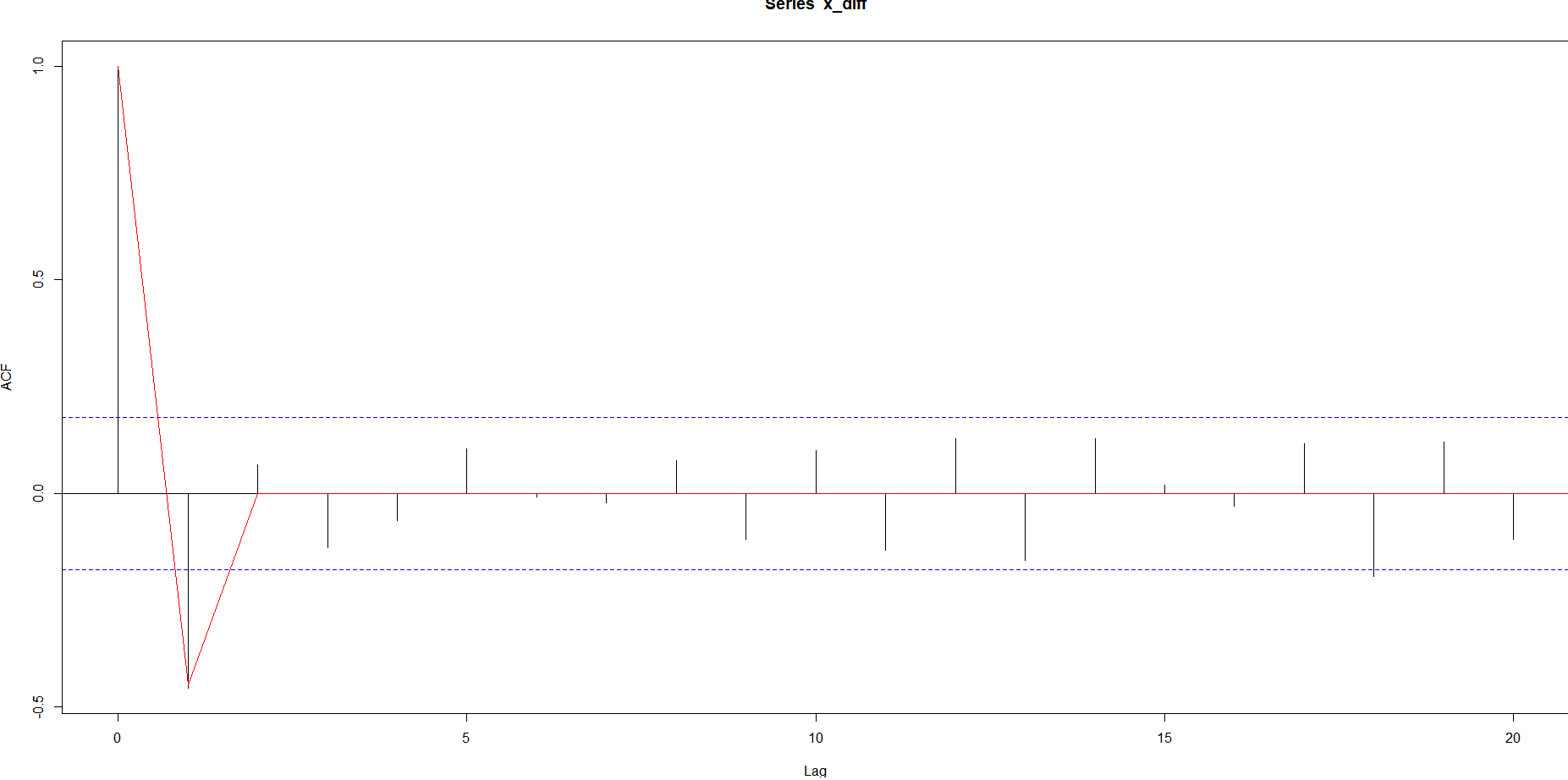
Suppose that we want to fit an MA(1) model to these data.

The parameter of interest is θ where Xt = Zt + θZt-1 .The theoretical ACF at lag 1 can be written as: ρ(1) = θ/ 1 + θ2.Looking at our sample ACF plot, we can assign ρ(1) the value of approximately -0.45.

-0.45= θ/ 1 + θ2 .

-0.45(1 + θ2)= θ

After solving for theta, we get two values of θ=-0.62 and θ=-1.59. Since, the value of θ=-1.59 lies outside the unit circle, we will choose θ=-0.62 so that the process is invertible.

We will plot the theoretical ACF along the lines of the sample ACF with the theta value as θ=-0.62.

Since the theoretical ACF and the sample ACF align correctly, the model (0,1,1) that we have chosen is a correct fit.

**CONCLUSION:**

After converting the model from non-stationary to stationary and checking on the ACF and PACF, we inferred from our observations that the model is a MA(1) and this claim is further supported by carrying out theoretical ACF checks which align with the sample ACF as well.

**Appendix :**

#QUESTION 1

#LOAD THE DATA

cet\_temp<-read.csv("cet\_temp.csv",header = TRUE)

cet\_temp

#A LOOK AT THE DATA

head(cet\_temp)

summary(cet\_temp)

#PLOT THE DATA AS A TIME SERIES MODEL

x=ts(cet\_temp$avg\_annual\_temp\_C,start=1900,frequency=1)

ts.plot(x)

#THE TIME SERIES DOES NOT SEEM STATIONARY SO WE WILL USE FIRST ORDER DIFFERENCING TO CONVERT THE DATA TO A STATIONARY MODEL

x\_diff<-diff(x)

ts.plot(x\_diff)

#DATA SEEMS STATIONARY NOW,SO WE WILL TAKE A LOOK AT THE ACF AND PACF OF THE DIFFERENCED DATA AND FORM A MODEL DEPENDING ON THE ACF AND PACF CUTOFFS

#SEEMS LIKE THE ACF DECAYS OFF PRETTY QUICKLY AT LAG 1 AND PACF DOES NOT DECAY AT ALL.SO WE CAN SAY THAT THE MODEL FITTED IS ARIMA(0,1,1)

acf(x\_diff)

pacf(x\_diff)

#COMPARING THEORATICAL AND SAMPLE ACF.THE VALUE OF THETA IS CALCULATED BY ρ(1) = θ/1+θ^2. THE VALUE OF ρ(1) IS -0.45 ACCORDING TO THE ACF AND THETA IS CALCULATED ACCORDINGLY

tacf<-ARMAacf(ar=c(0),ma=c(-0.62),lag.max=100)

x\_lag<-c(0:100)

lines(x\_lag,tacf,col="red")

**Coursework Report – MODEL FITTING AND FORECASTING**

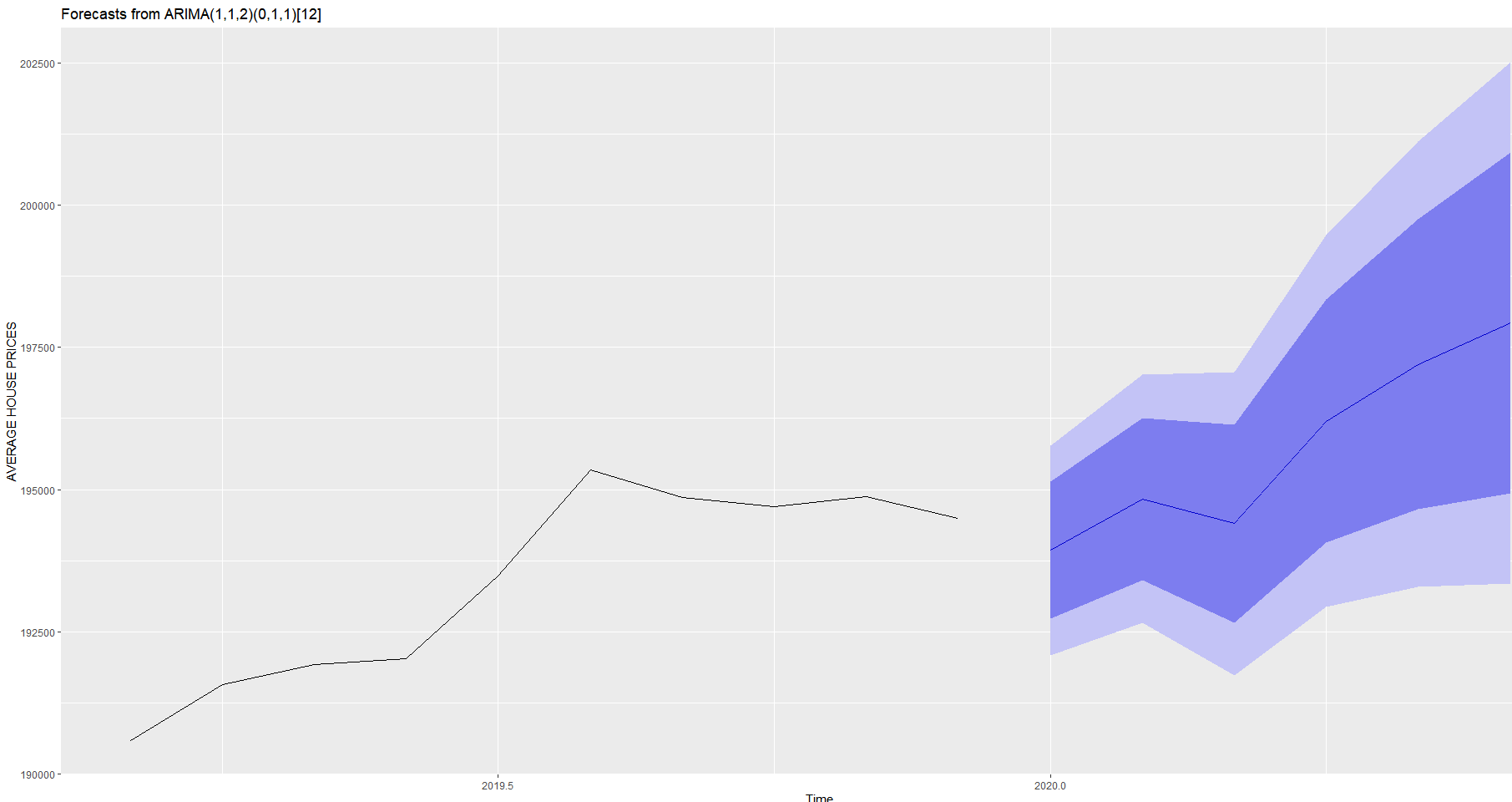
**MATH4022 – Time Series and Forecasting**

**Student ID: 20387234**

**EXECUTIVE SUMMARY OF THE REPORT:**

This is a report on the data for the monthly average house prices in the East Midlands Region of the United Kingdom which analyses different time series models and chooses the most suitable model to get the forecasted values for the house prices 6 months into the future.

Several statistical procedures and methods were used to convert the original time series data into a type of time-series data on which forecasting for every month of 2020 was possible. The forecast shows that albeit a small decrease in the prices from February 2020 to March 2020 there is generally an upward trend observed in the forecasted prices in the next 6 months.

****

The forecasted data shows that the house prices from Jan 2020 to June 2020 reach from 193930.5 to 197933.9 pounds. This forecasting plot is bound to contain some uncertainties and hence for example in June 2020, it would be only right to infer from the data that there could be a possible maximum increase in the house prices by 2.2 percent in and a possible decrease of 2.31 percent from the original forecasted prices. The time-series model on which an extensive preliminary research was done and which was used to make this forecasted model is called SARIMA or Seasonal Auto Regressive Integrated Moving Average Model.

The dataset is cleaned and adjusted with respect to all the assumptions to use the SARIMA model for model forecasting.

**Aim of the report:**

This report analyses a dataset that contains the monthly average house prices in the East Midlands region. We are supposed to produce a fitted time series model and then forecast some future average house prices for a local government agency. We are expected to infer the conclusions from the following actions that will be conducted on the dataset:

* **Checking for Stationarity of the Time Series**: It is important to check if the time series data that we are dealing with is stationary for a constant mean and variance. If it is not stationary it is important to deal with non-stationarity using different orders of differencing.
* **Differencing:** If the time series plot is not stationary we can convert it into one by differencing it. If the time series has seasonality and trends involved we will seasonally difference the data initially and we will difference the seasonally differenced data to finally get the time series plot that can be further used to study the proper fit of the model and hence can be used for forecasting purposes.
* **ACF and PACF:**  After we are done with the differencing, we will use the ACF and PACF plots to give us a proper idea of the model that we can fit for forecasting purposes.
* **Finalising a SARIMA Model**  We will finalise a seasonal arima model by taking a look at the ACF and PACF of the time series model which will help us in computating the seasonal and non-seasonal parts of the model
* **Forecasting:**  After we have finalised a model we will use it for forecasting to predict how the model will behave 6 months in the future. We will be using the forecasting package in R to do this.

**Initial Glimpse of the Dataset:**

The dataset consists of the following columns:

Month: Shows the month when that particular mean house price was recorded

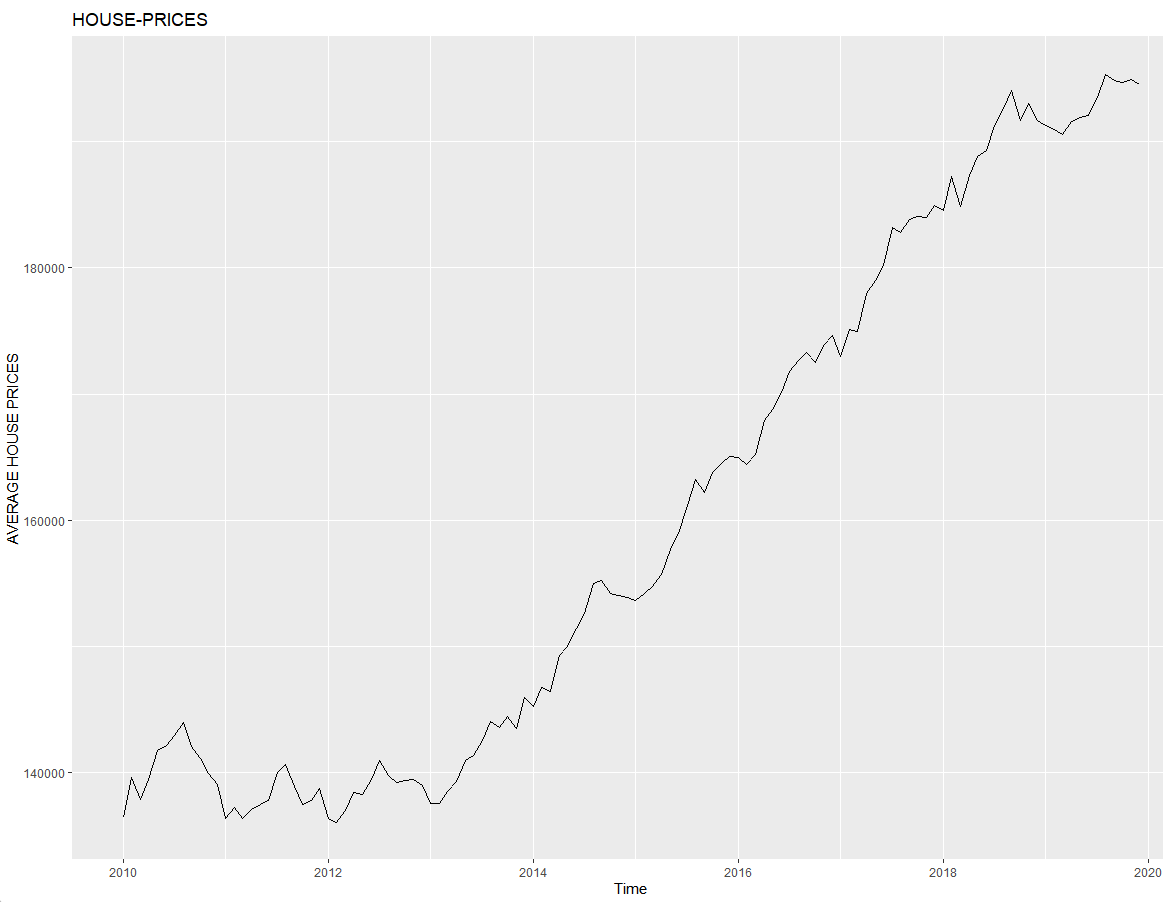
Year: Shows the year when that particular mean house price was recorded

Average\_Price\_GBP: Shows the mean house prices in GBP calculated monthly from the period of January 2010 to December 2019.

After taking a look at the dataset we are eligible to start with initial plotting to decide the nature of the dataset and decide if it is stationary or not.

**Stationarity Checks:**

Initially, we can start checking for stationarity by plotting the time series data. The following is the plot that shows the raw time-series data :

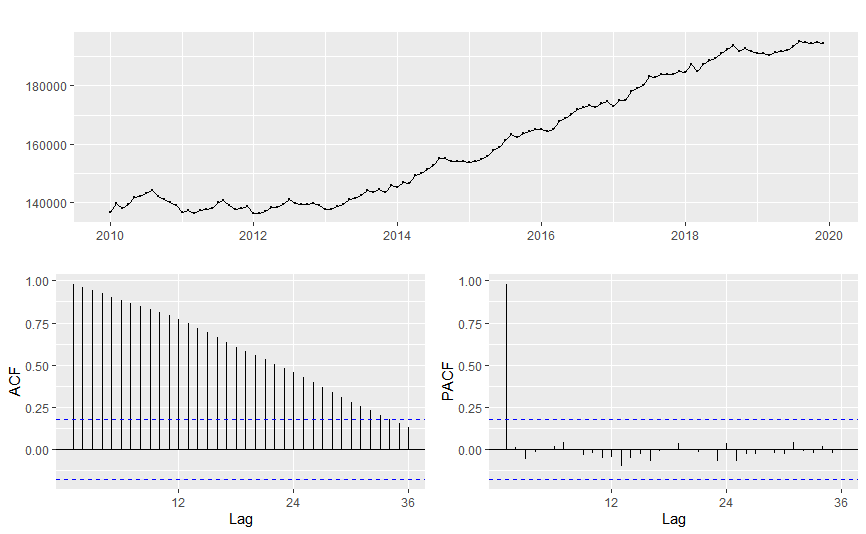


From the following time plot we can tell the following things:

* The time plot is not stationary because the mean and variance of the plot are not constant
* We can see emerging patterns of trend and seasonality in this plot which are important time series metrics that we need to deal with and get rid of,
* We see an upward trajectory trend of average house prices throughout the decade which means the mean is varied over that time
* We also see that our data is experiencing regular predictable changes that are constant and reoccurring throughout every calendar year. We see that for every year there is a peak in somewhere in the middle of the year which would suggest that the average house prices are maximum at that point of the year after which they slowly start to regress.

The main inference to be taken from this would be that we will be differencing our data twice, one would be classified as a seasonal difference and the other would be the first order difference of the seasonally differenced data.

**PRELIMINARY ACF AND PACF CHECK:**

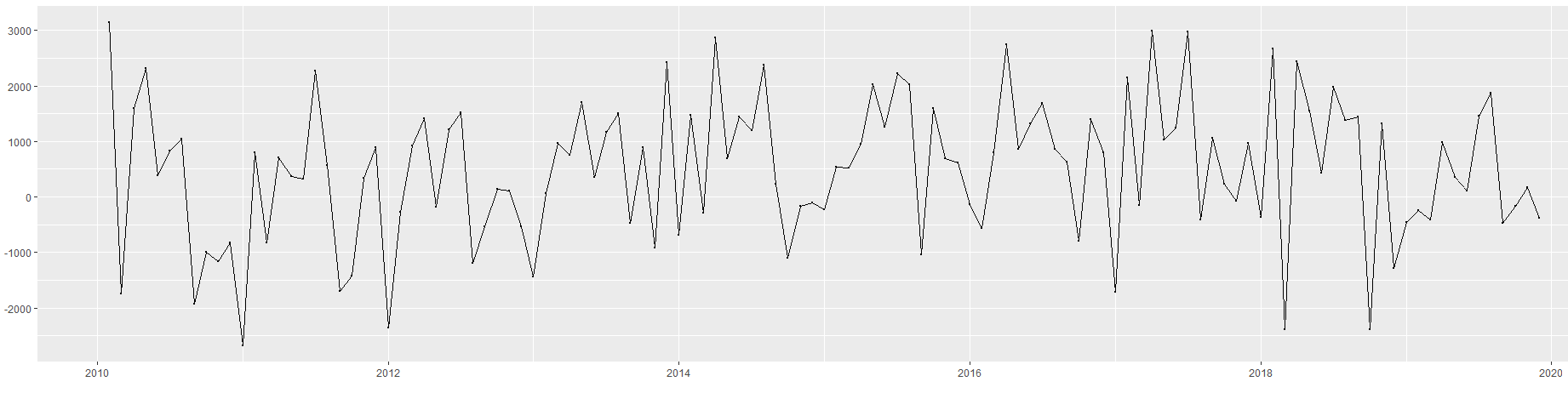
****

When the plot such as the one we have seen above is non-stationary and has trends and seasonality associated to it, we can infer by observing the ACF that instead of decaying quickly it plummets to zero on a rather gradual basis or does not plummet at all. The spikes in the plot that exceed the significance level tell us that the there are significant autocorrelations with the lagged values.

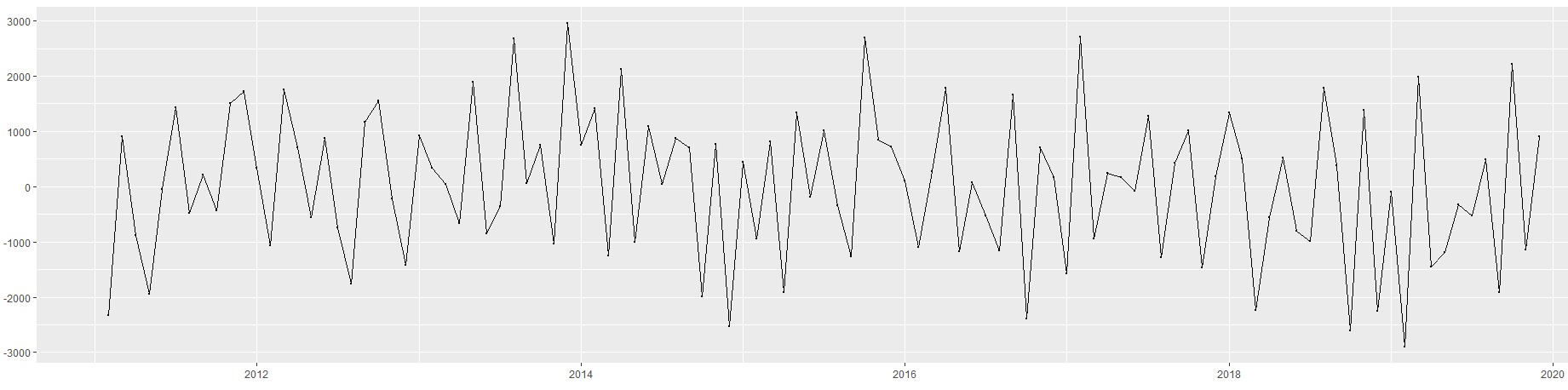
In the case of PACF, we see that adding more lags will decrease the degree of freedom and power.

**DIFFERENCING:**

After our stationarity checks are complete and we find that the data is not stationary, differencing twice usually removes all the drift and seasonal trends from the model and gives us a stationary plot

After taking first difference of the time series at lag 1

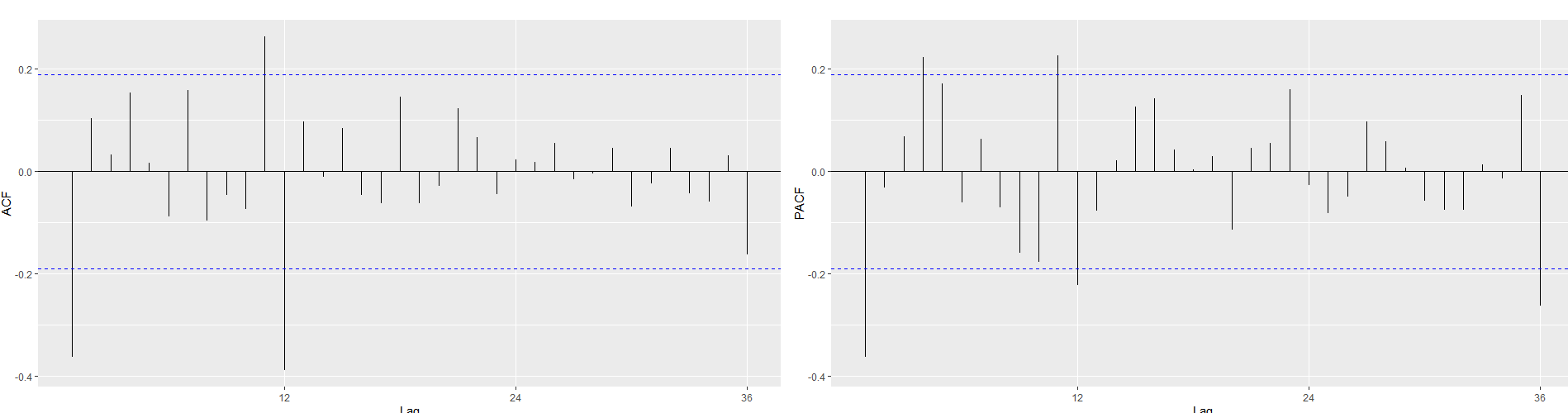
After taking seasonal difference at lag 12 on the first order differenced plot :



It is important to note that the order of differencing does not matter. However in cases of strong seasonality trends it is advisable to seasonally difference the data first and then carry out the first difference.

**ACF AND PACF:**

Since the time series is stationary now, we can go ahead and talk about the ACF and PACF and try to form a model using the information we infer from the plots. The plots are shown below :



From the following plots a lot of information can be inferred and it can be summarised in a succinct manner below:

* Since there are significant spikes at lag1 in the ACF of the differenced time-series, it suggests a non-seasonal MA(1) component
* Also since, there is also a significant spike at lag 12 in the ACF of the differenced time series, it suggests a seasonal MA(1) component

Consequently, it will be right to begin with an ARIMA (0,1,1)(0,1,1)[12] model which indicates a first and a seasonal difference, and a non-seasonal and a seasonal MA(1) component.

**FITTING THE SARIMA MODEL:**

We know that a SARIMA models which are seasonal arima models can be written as per the general formula:

**(p,d,q) x (P,D,Q)s**  where **,**

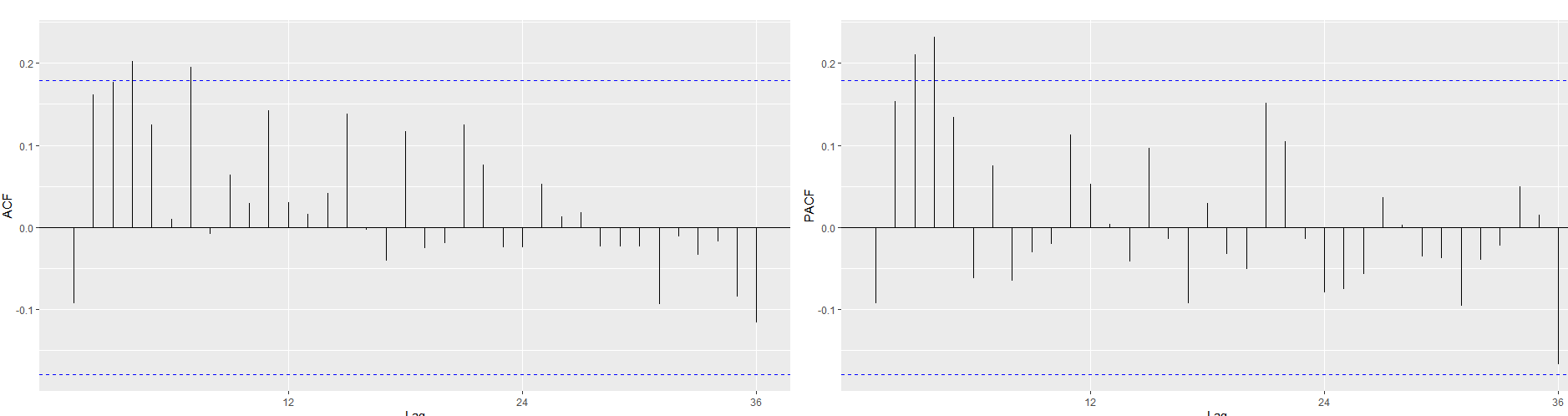
* the first term denotes the non-seasonal component with AR(p) component and MA(q) component with d denoting the times the time series has been differenced
* the second term denotes the non-seasonal component with AR(P) component and MA(Q) component with D denoting the times the time series has been seasonally differenced with s denoting the lag of the time series.

After taking inference from the ACF and PACF we know that for the non-seasonal components we are dealing with a model that has AR(1) and MA(1) components whereas for the seasonal components we are only dealing with MA(1) components because of the regular peak repetition after every multiple of 12.

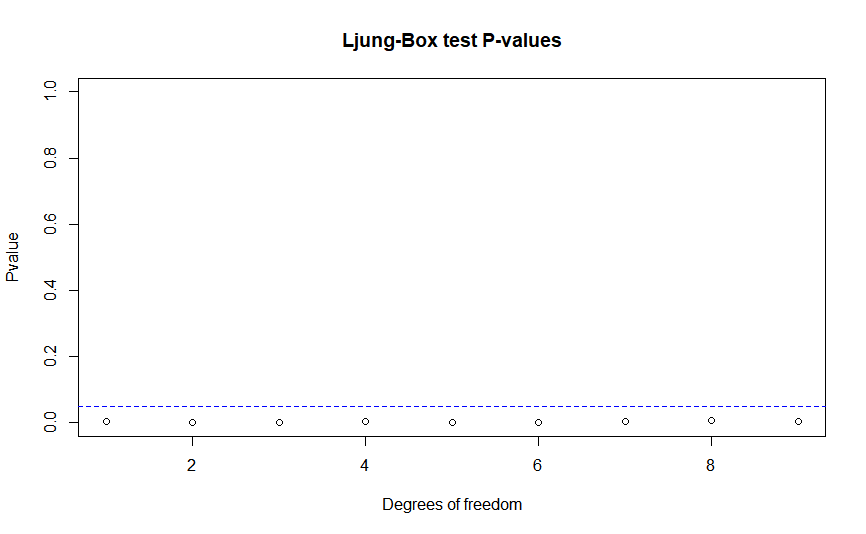
So going by the general formula of the SARIMA model we can write our fitted model as :

**(0,1,1)X(0,1,1)[12]**

To verify if this is in-fact the correct model, we will now take a look at how the ACF and PACF of the residuals act and also perform the Ljung-Box Test to check on the p-values and if they lie above the 0.05 mark.

 Above are the plotted ACF and PACF of the residuals that we get after we plot the SARIMA model. We see that there are peaks in lag 2 and lag 3 in both the ACF and PACF plots which would signify that even after fitting the model there is a lot of information that is adequately left over and not captured properly.

For final verification, we will conduct a Ljung-Box test to check how the p-values are plotted against the degrees of freedom and what we can infer from it.

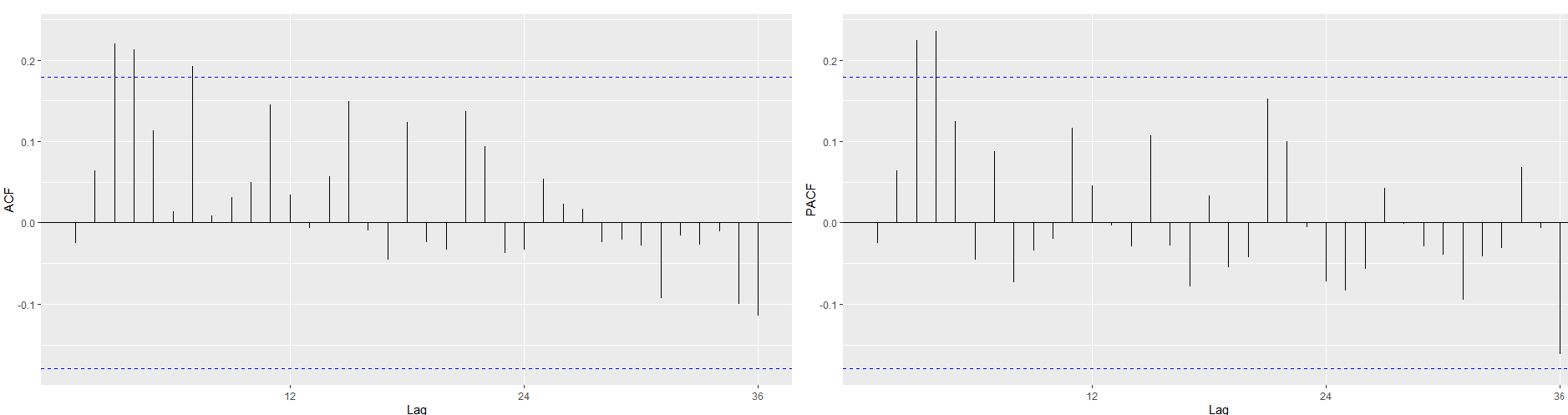


All P-Values appear to be smaller than 0.05 which would suggest that we should reject the Null-hypothesis and that the residuals are independent to each other. Hence we have verified that the SARIMA model (0,1,1)X(0,1,1)[12] will not be a good-fit for this model.

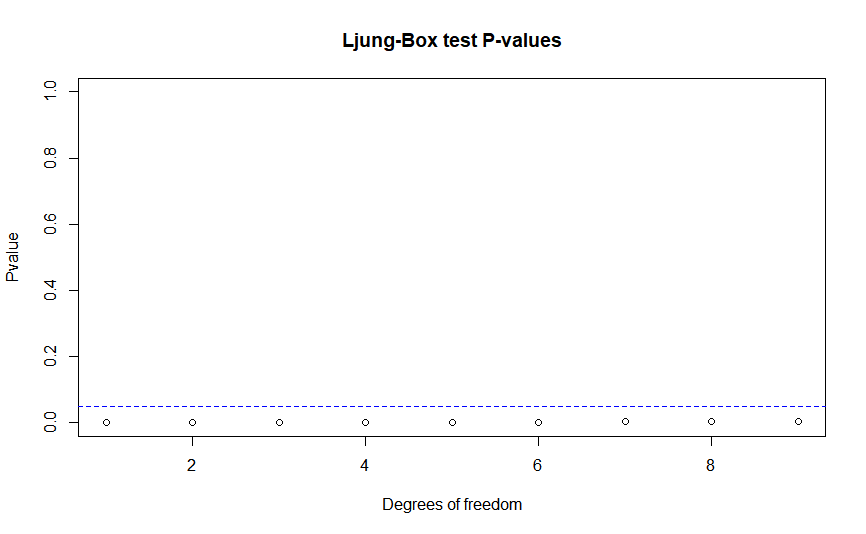
Now we will try fitting a model that is in the neighbourhood of our previous SARIMA model and check how the residuals are acting and how the Ljung-Box test performs on it.Having fitted the (0,1,1)X(0,1,1)[12] model previously, we will check if (1,1,1)X(0,1,1)[12] by adding a non-seasonal AR(1) component and check for residuals and verify using Ljung Box test.

**TESTING THE MODEL(1,1,1)X(0,1,1)[12]**

To verify if this is the correct model we will take a look at the ACF and the PACF of the residuals and also verify using the Ljung-Box Test.



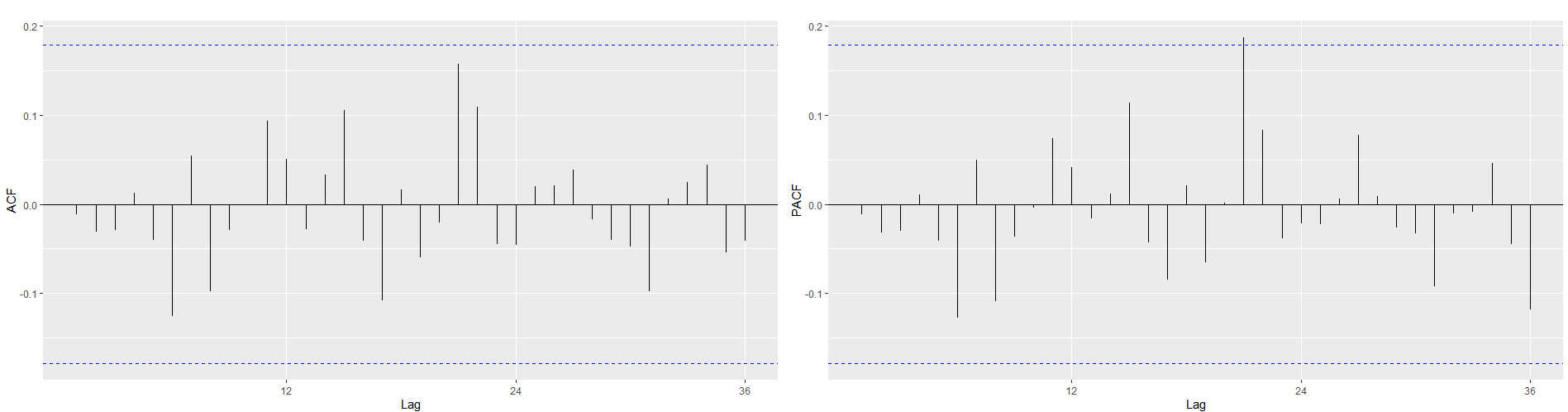
As we can see from the plots above, there are lags at 2 and 3 that exceed the boundaries which would mean that even though this model is better than the previous one, the plot of the residuals show that there is some information that is not captured properly and hence is adequately left over. So this is probably not a fit model. To support our claim, we will do the Ljung-Box Test and check on the plotted P-values.



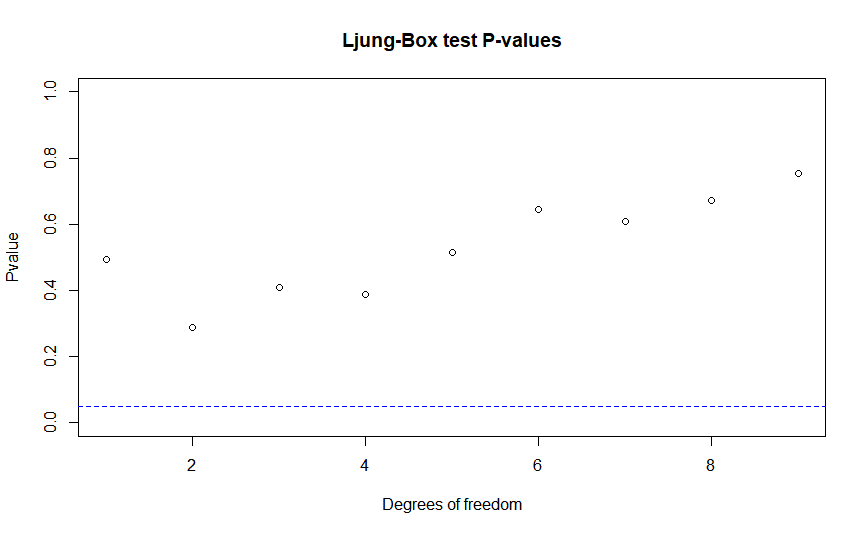
As we can see, all of the P-values fall below the 0.05 significance value and hence we will reject the null hypothesis. So we will select another model in the neighbourhood of the previous one and do the same procedure that we have already been doing.

**TESTING THE MODEL(1,1,2)X(0,1,1)[12]**

Now we will check if (1,1,2)(0,1,1)[12] can be fitted successfully and we will do this by checking how the ACF and the PACF of the residuals behave with respect to lags and also do the Ljung-Box Test to confirm our observations.



As we can see all the peaks here are within the boundaries. There is a peak at lag 21 in the PACF that slightly exceeds the boundary line but that can be ignored as the difference is instantaneously small so it is not statistically significant. Apart from that this model looks like the perfect fit but we need to support this claim by doing a Ljung-Box test on it and check if how the p-values perform against various degrees of freedom.



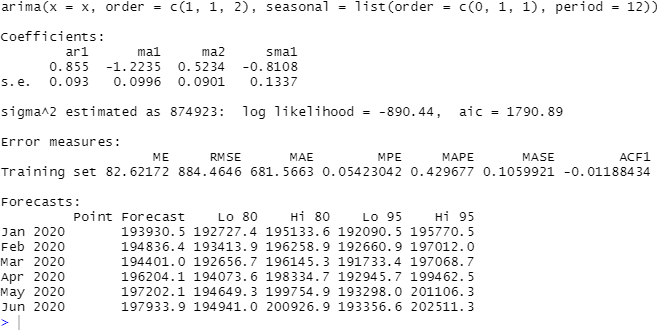
We see that for the first time, all our p-values are bigger than the 0.05 significance level, hence it would make sense to not reject the null hypothesis and conclude that the model **(1,1,2)X(0,1,1)[12]** is actually a better model to fit than the ones we have tried to fit before.

Furthermore, if we take a look at the AIC scores of all the models we have tried to fit, we will realise that the model **(1,1,2)X(0,1,1)[12]** has a score of 1790.89 which is less than the AIC of **(1,1,1)X(0,1,1)[12]** (1808.48) and the AIC of **(0,1,1)X(0,1,1)[12]** (1809.08), and it is a fact that a lower AIC score is always preferred when we are fitting a model.

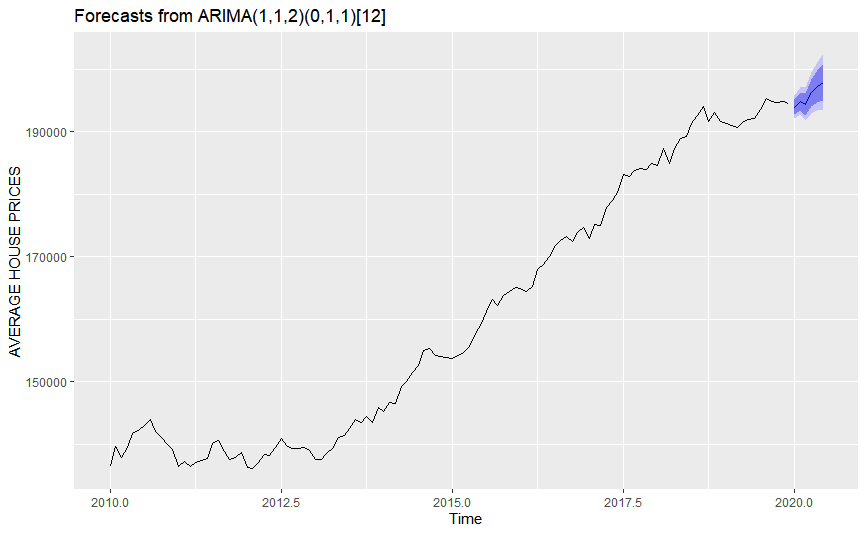
**FORECASTING:**

Now that we have finalised a SARIMA model, we can go ahead and use it for forecasting purposes. The R package “forecast” can be used to make forecasts for time-series models. We are supposed to forecast the monthly house prices for the first 6 months of 2020. We can do this by using the forecast function and insert inputs to the function which will include the fitted SARIMA model, the amount of time we need to forecast for and a confidence interval levels which we have selected as 80 and 95. We will then plot the forecasted values using autoplot function.

The forecasted values for the 6 months from January-June 2020 are as follows:

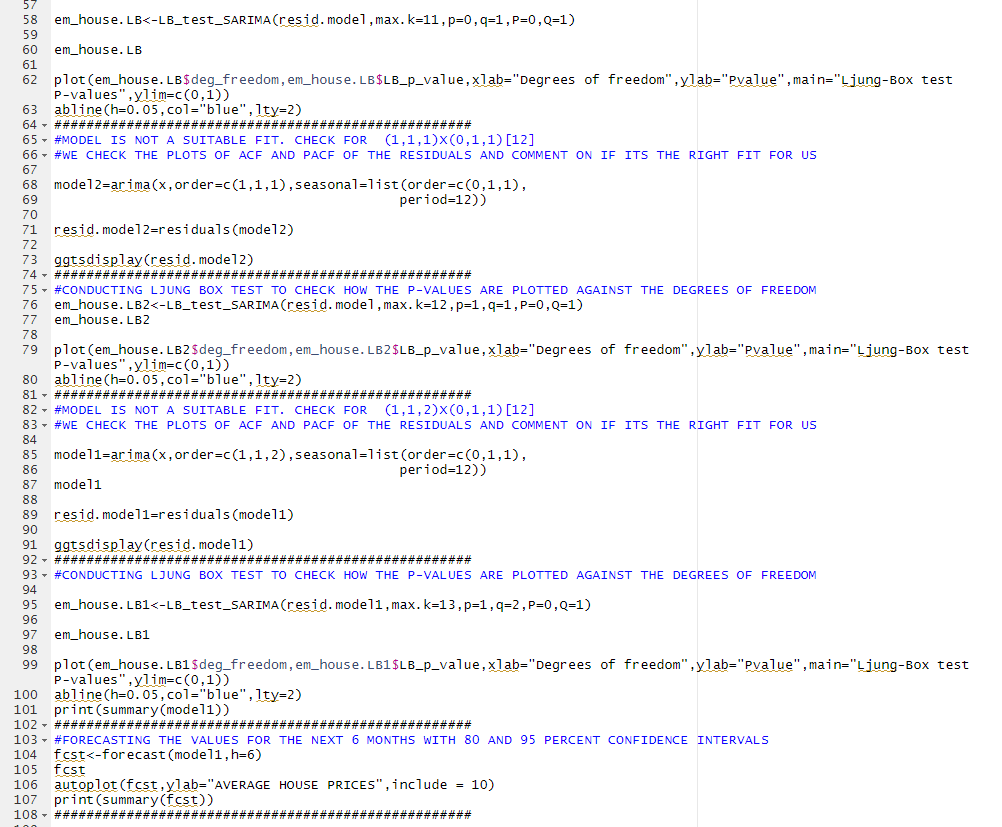


Now, we will plot these values using the auto-plot command to plot the time-series along with the forecasts



The line in between the highlighted confidence intervals signifies how the forecasted values will move along with the time. We see that the resultant forecast is expected to follow the normal seasonal behaviour with a dip in the prices from February 2020 to March 2020 but will rise up again from March to June. Due to the associated uncertainty regarding the forecasts we have designed confidence intervals of 80 and 95 percent. The light blue boundary signifies a 95 percent confidence interval and the dark blue boundary signifies a 80 percent confidence interval.

**APPENDIX: **

****