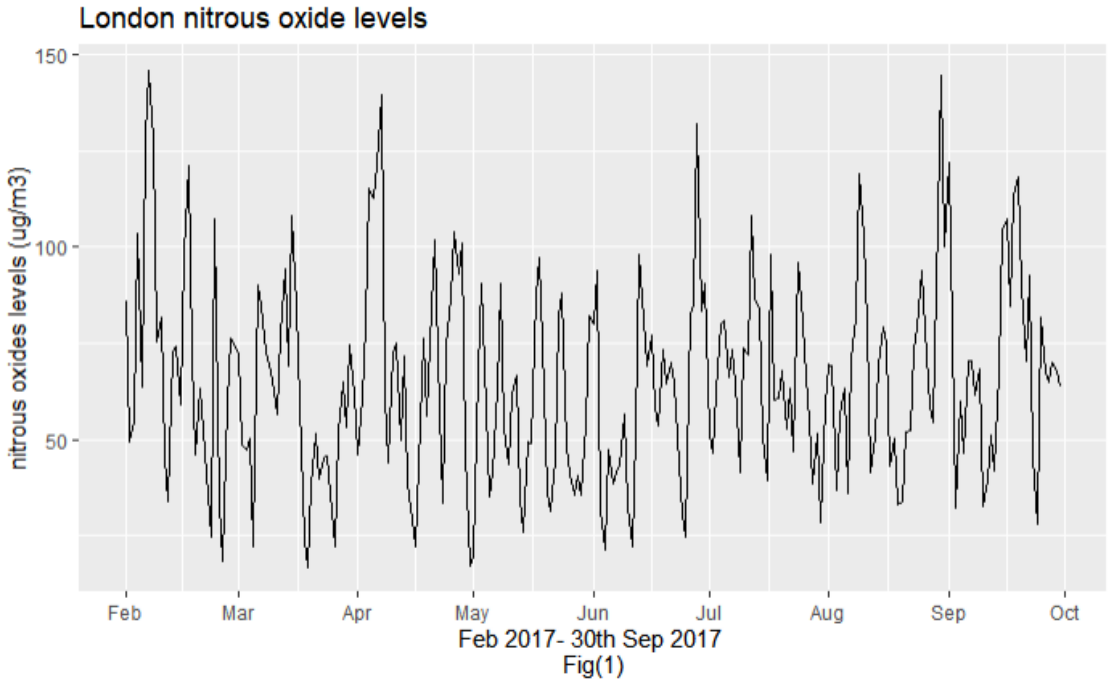
Lokeshwaran Arunachalam

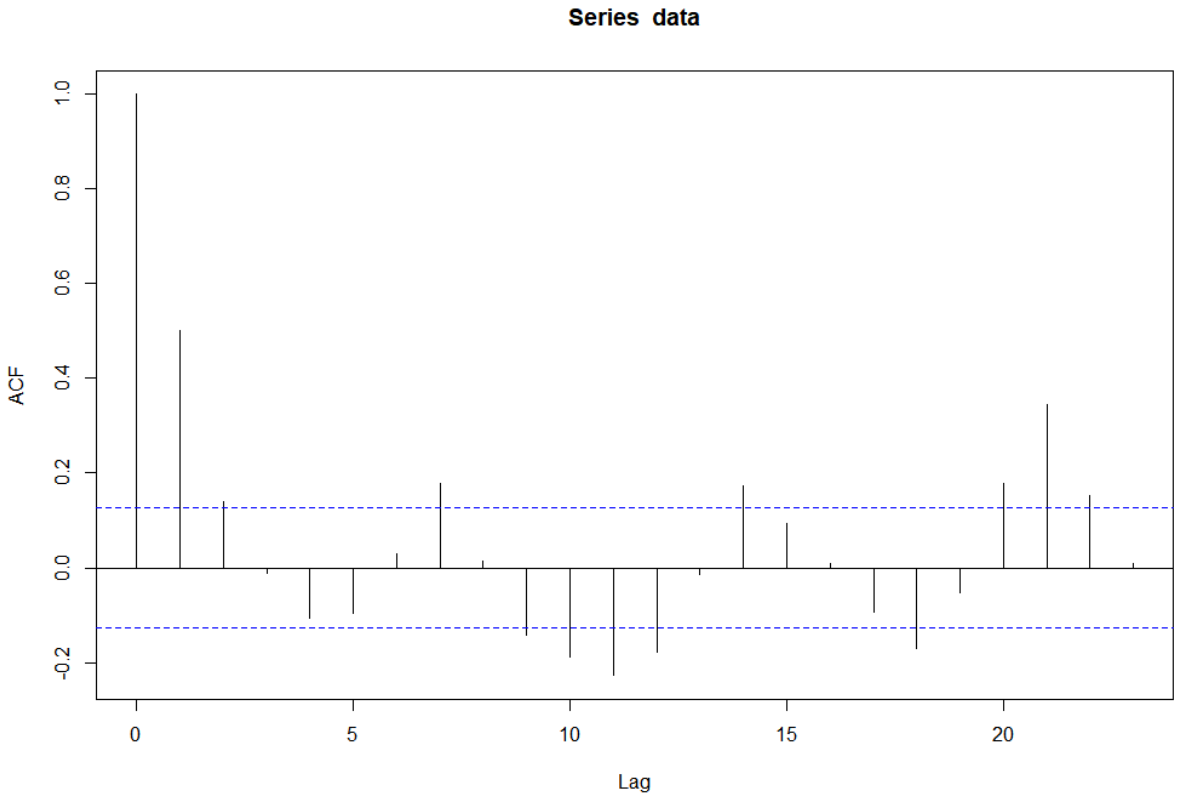
Report 1

The mean nitrous oxide levels of London city were recorded from 1st February 2017 to 30th September 2017. In this report we would analyse the time series data and try to find an optimal time series model which can be used to predict nitrous oxide level.



Fig(1)

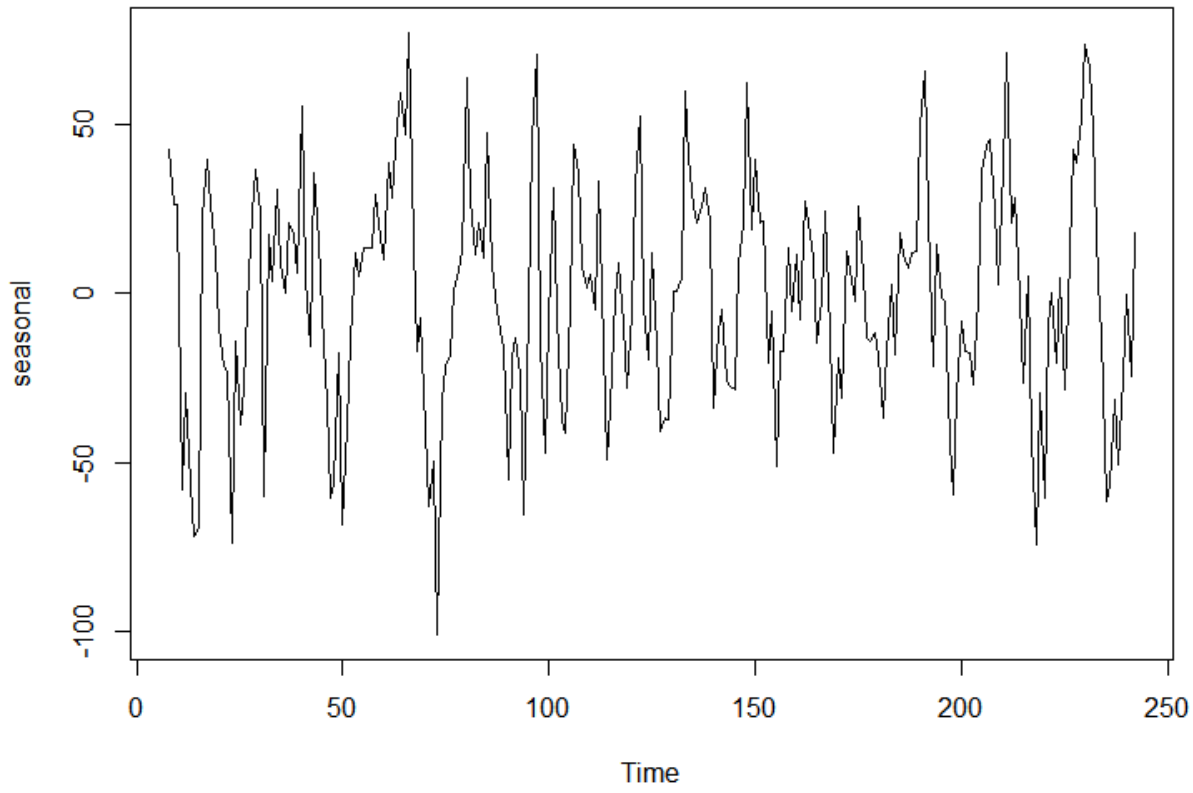
The line chart shows that there is some seasonality in the data, but it is not very evident because there is some variation in seasonality. From February to March there is very slight decreasing trend in nitrous oxide but after March there is no clear increasing or decreasing trend in nitrous oxide. Overall, there is no clear evidence that the data is stationary. Hence, we will plot ACF plot to find whether the data is stationary or not.



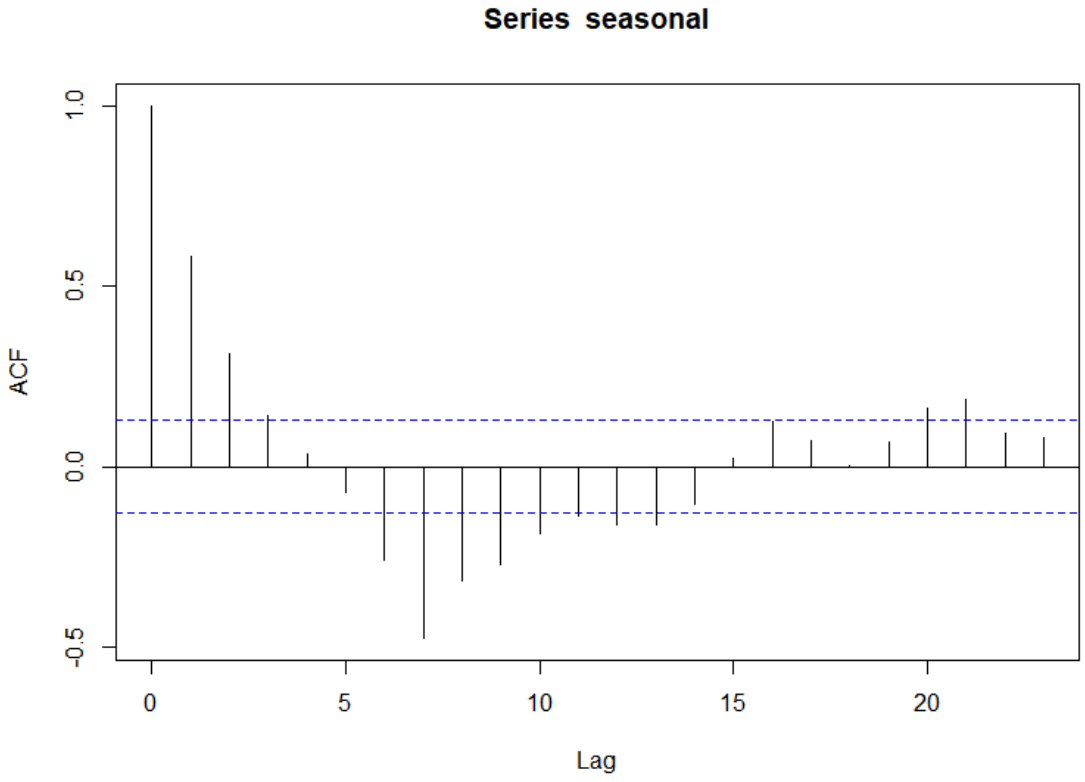
The ACF plot shows the lag 1 shows the highest correlation of 0.5 it is obvious that the contribution of previous day nitrous oxide level is used to predict nitrous oxide level for today. From the plot we can also observe that there is some contribution from the lag 7, 14, and 21. Therefore, there is a seasonal pattern that is used to repeat every week. The ACF value does not falls below 95% confidence interval of the correlation coefficients over a period of lag. Therefore, the data is not stationary.

Due to seasonal pattern that is used to repeat every week we will difference the data at lag 7. Let's consider X_t is nitrous oxide level at a particular day then X_{t-1} is the previous day nitrous oxide level.

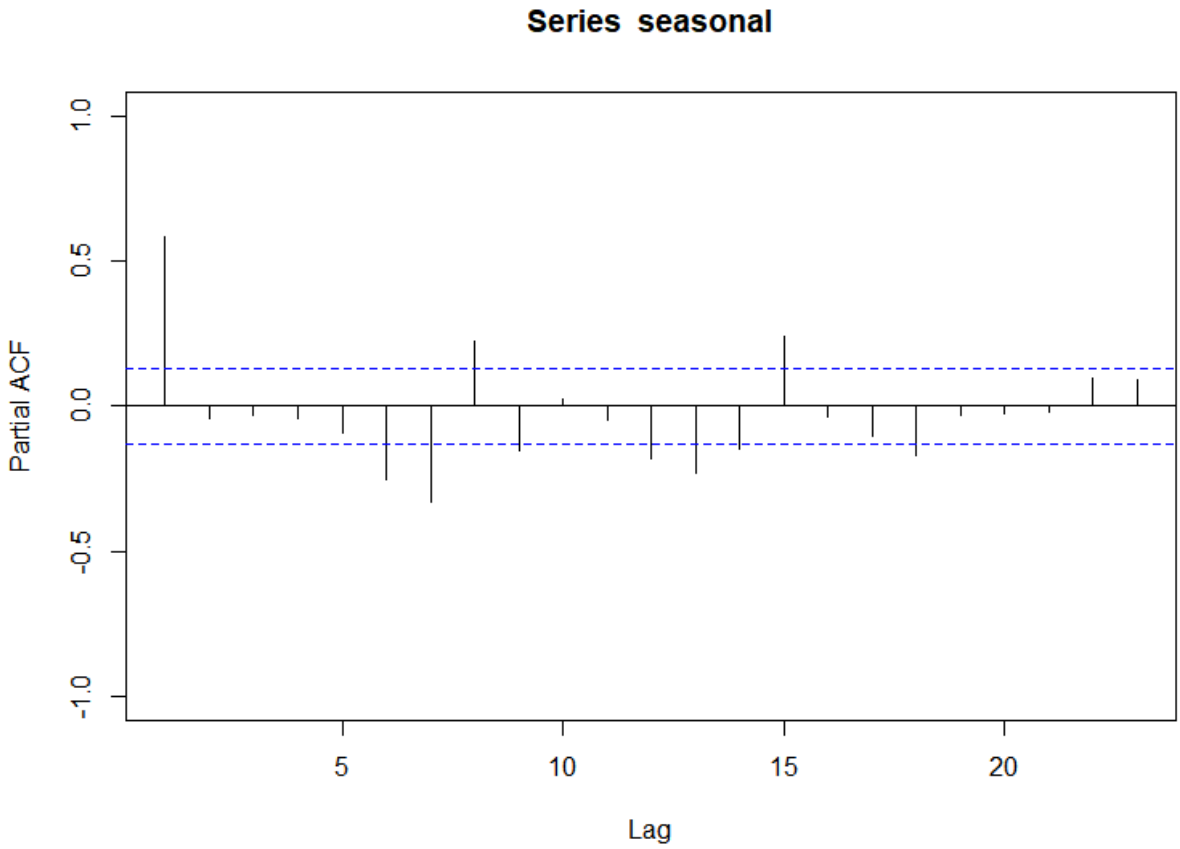
$$Y_t = X_t - X_{t-7}$$



From the weekly differenced data we cannot observe any obvious trend and it seems the mean value is constant and equal to zero there is no obvious change in variance over a period of time. There is some outlier near 70 but it is negligible. Overall, the data looks stationary.



The ACF plot shows that the lag values fall exponentially after certain lags. Hence, it is not an MA process. We may observe there is some seasonality in the data, but it falls below 95% confidence interval correlation over a period of time. Hence, it can be considered stationary.

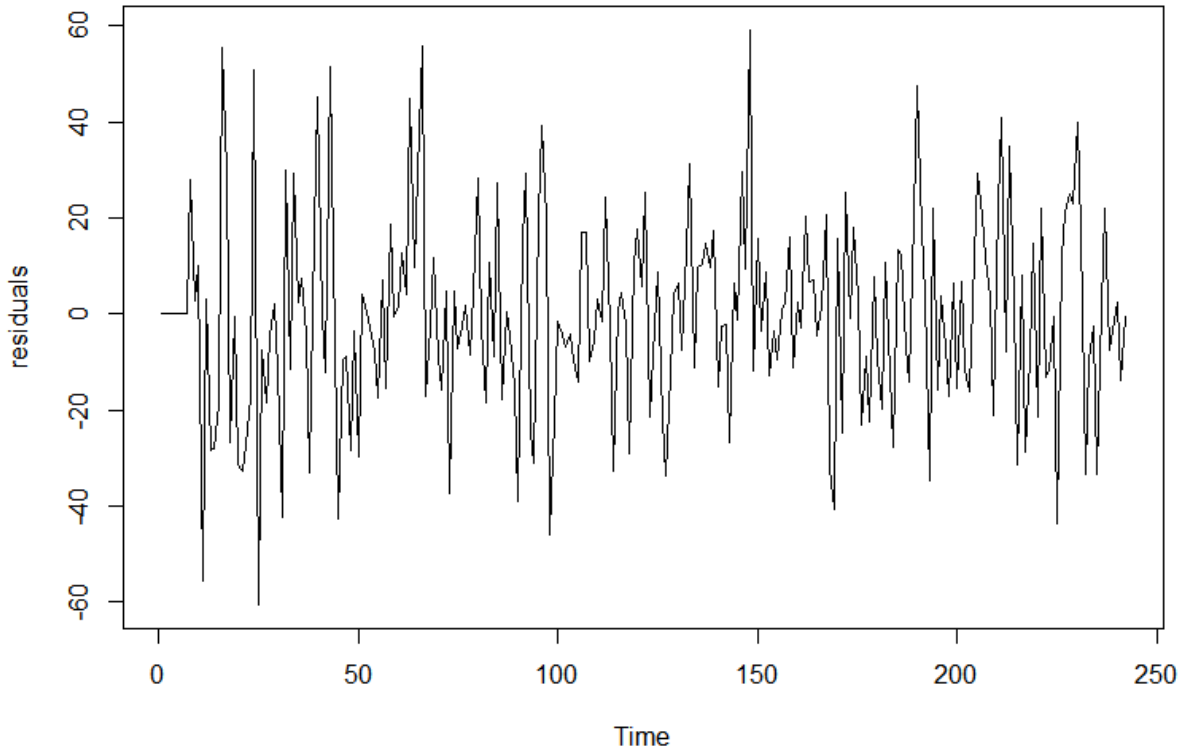


The plot shows non-seasonal lag falls to zero after lag 1. Therefore we can fit the AR(1) model. We can observe there is two seasonal lag from lag 5 to 10 and 10 to 15. Hence, we can try fit two seasonal AR(2) model, one non-seasonal AR(1) model and with seasonal difference of lag 7.

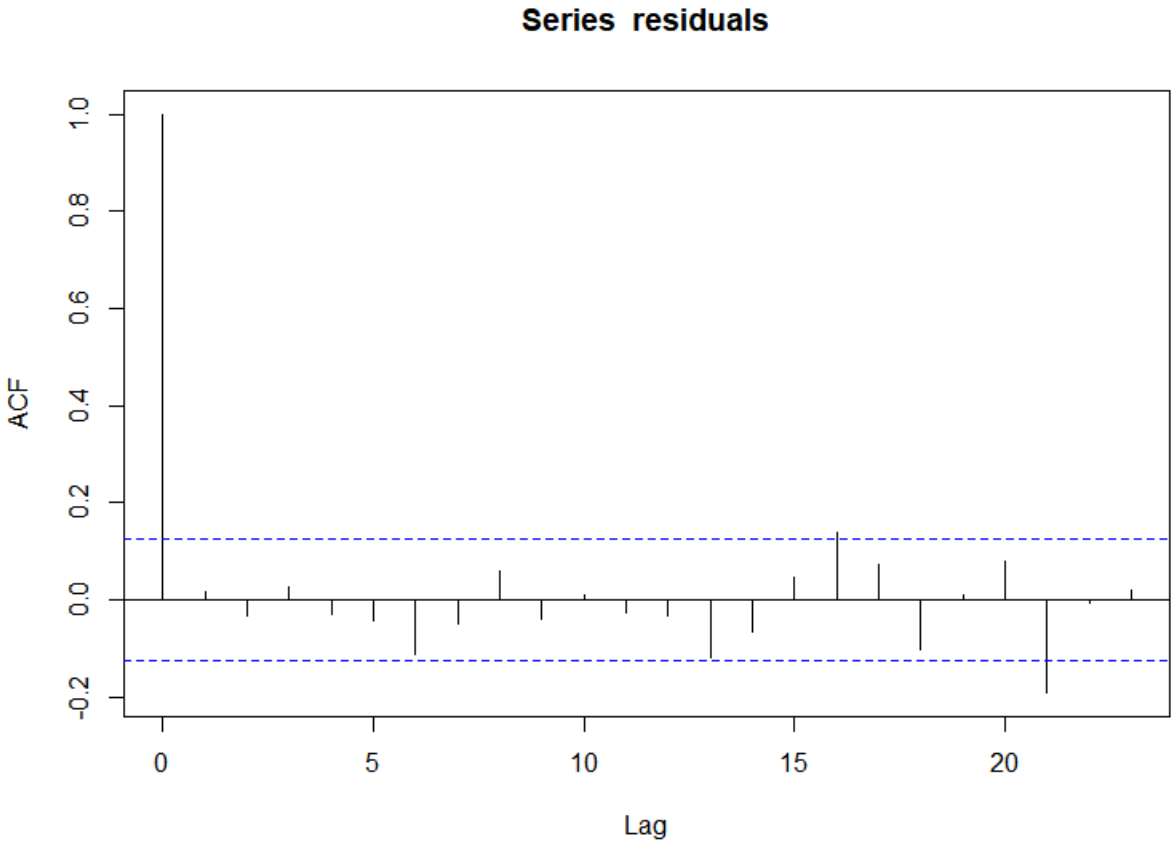
(i.e ARIMA(1, 0, 0) X $(2, 1, 0)_{12}$ model)

```
Call:
arima(x = data, order = c(1, 0, 0), seasonal = (list(order = c(2, 1, 0), period = 7)),
    method = "ML")

Coefficients:
         ar1     sar1     sar2
      0.5422  -0.7044  -0.4529
s.e.  0.0550   0.0603   0.0614

sigma^2 estimated as 442:  log likelihood = -1051.89,  aic = 2111.77
```

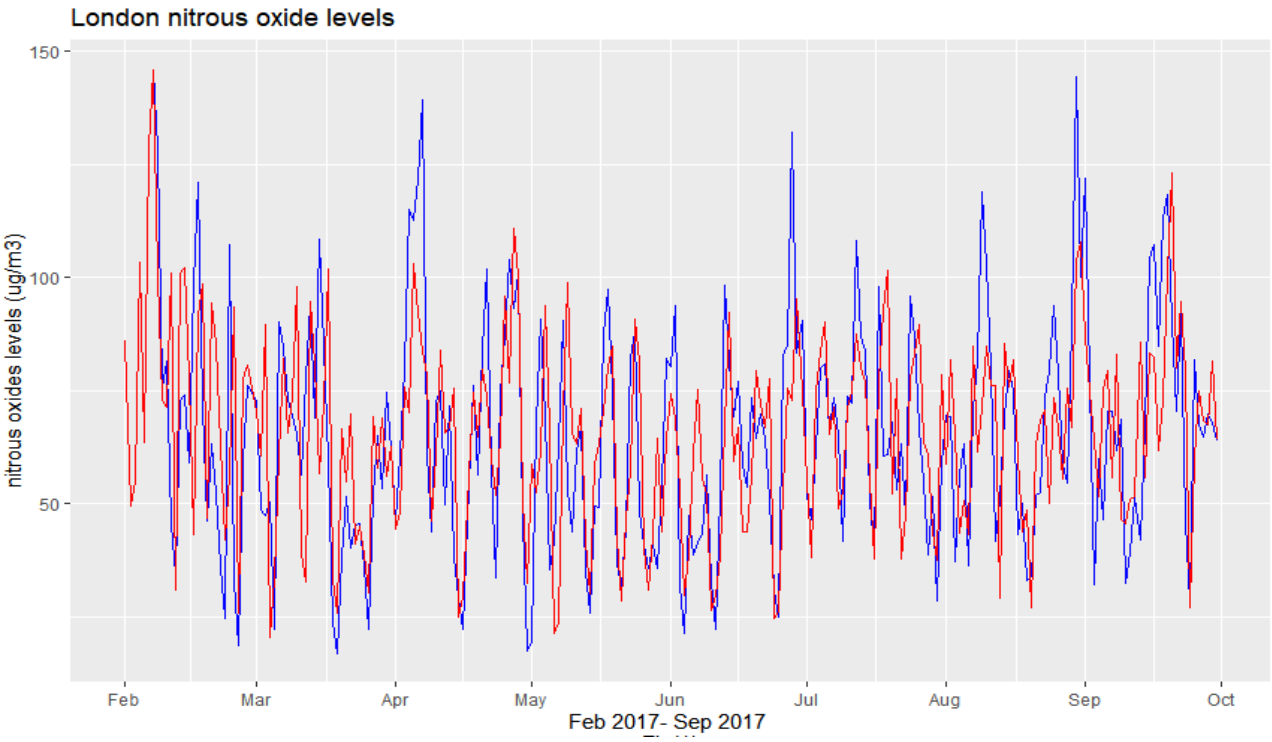
We can check whether the residuals are white noises by plotting the residuals plot.



The residual of the model looks stationary and there is no pattern. We can also verify by plotting ACF plot for residuals.



The residuals plot shows the more than 95% of data falls below the 95% confidence interval correlation. Therefore, the residuals are white noise.



In the above chart the blue line represents the actual value and the red line represents the predicted value. We can observe that the model is a good fit.

The parameters of the model were.

$\emptyset_1 = 0.54, \Phi_1 = -0.7, \Phi_2 = -0.45$

The Fitted model equation

$$(1 - 0.54B)(1 + 0.7B + 0.45\,B^2)(1 - B^7)X_t = Z_t$$

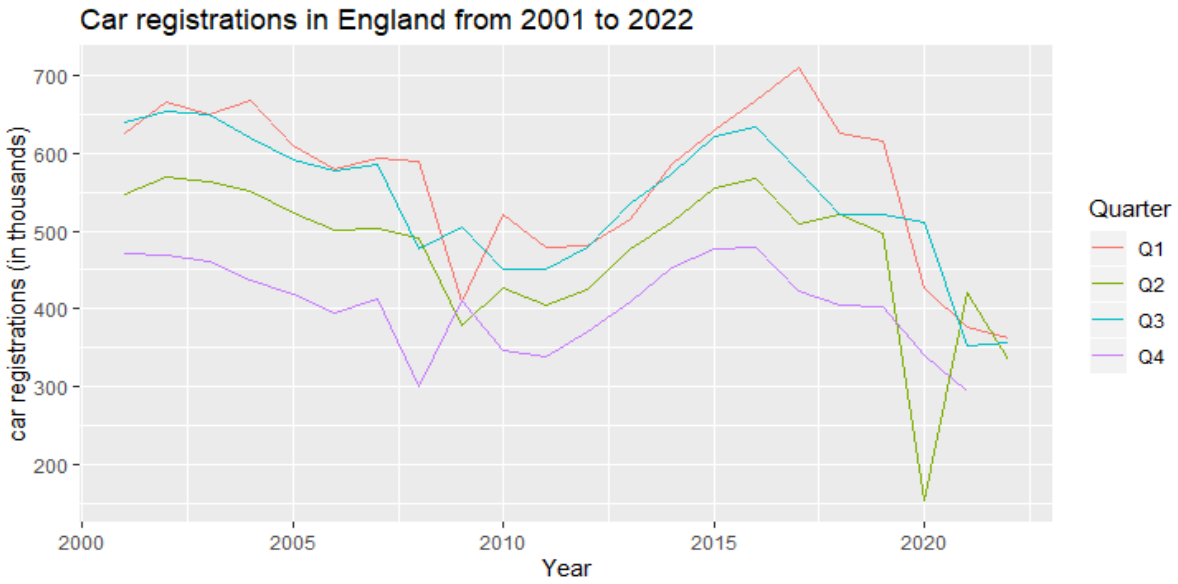
B is the Backward shift operator

$$BX_t = X_{t-1}$$

Report 2

Introduction:

In this report we are going to analyse and create a time series model to forecast the number of new car registrations (in thousands) in England using the recorded car registrations data in England from 2001 to 2022. The values were recorded for each of the four quarters of the year where Q1 = January to March, Q2 = April to June, Q3 = July to September, Q4 = October to December. We have to find the forecasted number of new cars registered for Q4 of 2022 and Q1-Q3 of 2023.



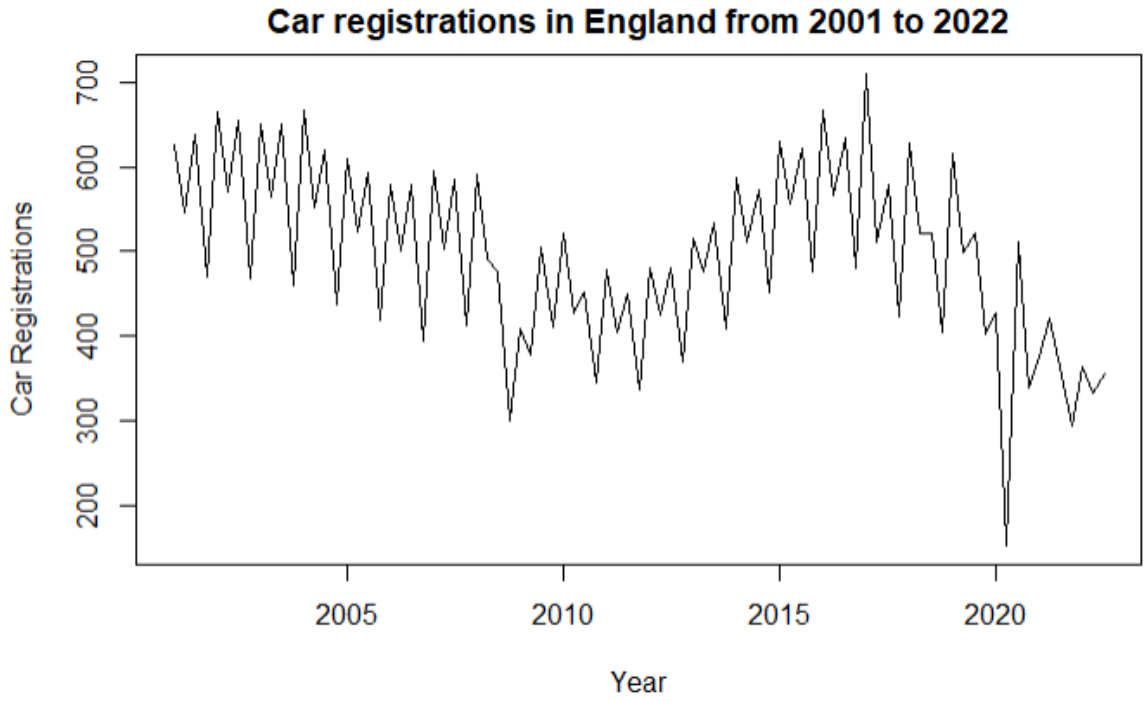
Analysis:

From the data we can clearly see there is a seasonal behaviour based on quarterly data. It is found that people show more interest to buy cars during the first quarter of the year when compared to other quarterly data and in Q4 the people interest is very low to buy cars. The Q3 pattern of car registrations is similar to Q1. There is a decreasing trend of car registrations from 2001 to 2007. From 2008 to 2009 there is a sudden fall in the number of car registrations this may be due to Global financial crisis occurred during that period so most of the people at the time don't have enough money to buy cars. After, 2009 there is an increasing trend from 2010 to 2017 where Q1 of 2017 achieve a highest number of car registrations which is more than 700,000 cars but after 2017 there is a steady decline in car registrations in England. The Q2 of 2020 saw the lowest number of 100,000 car registrations this is due to corona virus pandemic and after that from Q3 2020 it started to recover.

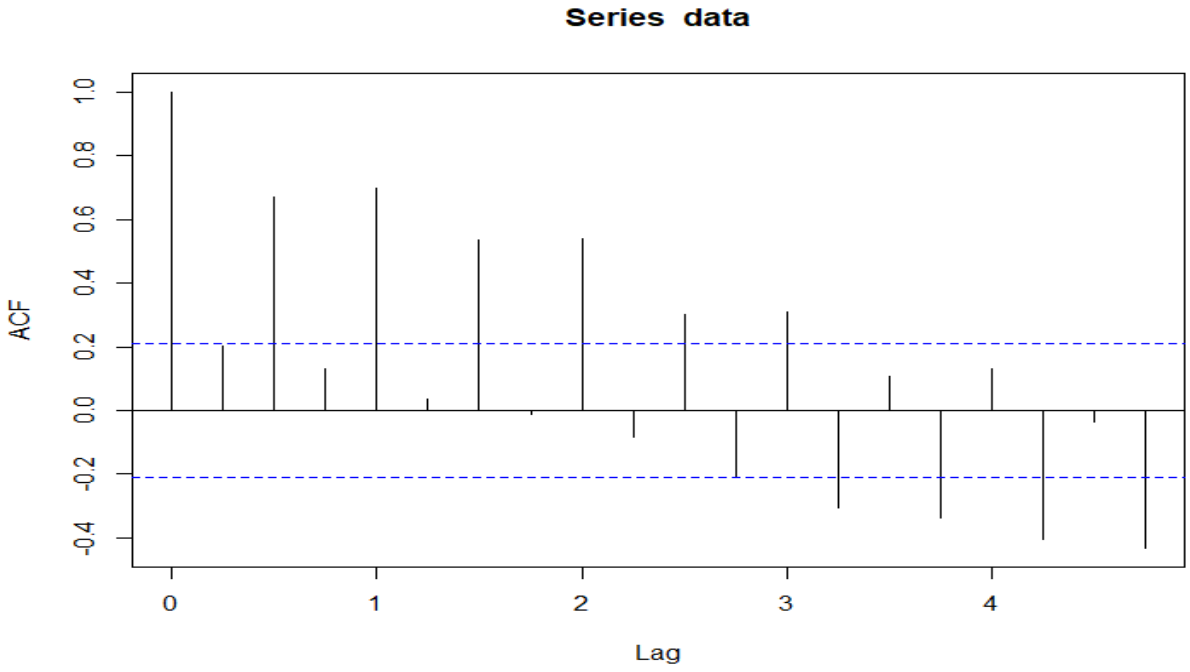
Results of time series model:

$$X_t = X_{t-4} + 0.98\,X_{t-1} - 0.98\,X_{t-5} - 0.94\,Z_{t-4} - 0.54\,Z_{t-1} + 0.51\,Z_{t-5} + Z_t$$

For example, if we want to predict the number of cars registered in Q1 2005 then it mostly depends upon the cars registered on Q1 2004 (X_{t-4}) and the unpredictable pattern occurred in Q1 2004 ($0.94\,Z_{t-4}$). The previous quarter (Q4 2003) of Q1 2004 also plays a role in predicting Q1 2005 value (i.e $-0.98\,X_{t-5}, 0.51\,Z_{t-5}$). Finally, Q4 2004 the previous quarter contribute 98% ($0.98\,X_{t-1}$) of its value to predict the car registered on Q1 2005. Z_t is the random pattern can occur in X_t.



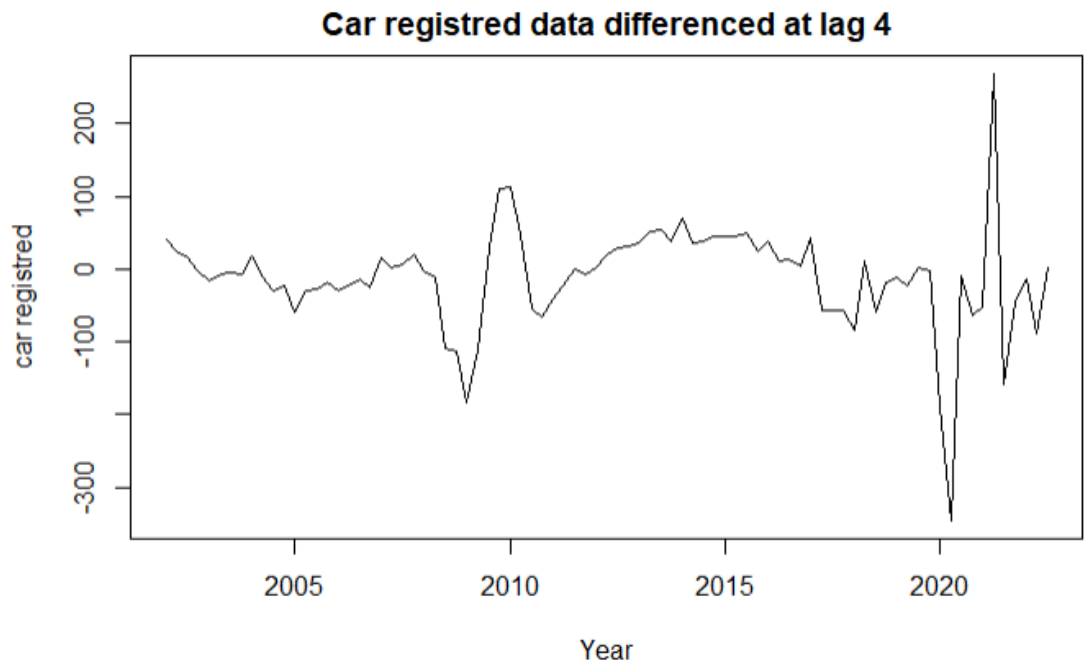
From the data we can observe there is a seasonal pattern and there is slight increasing and decreasing trend and it is not stationary data. However, we will plot the ACF plot to confirm it.



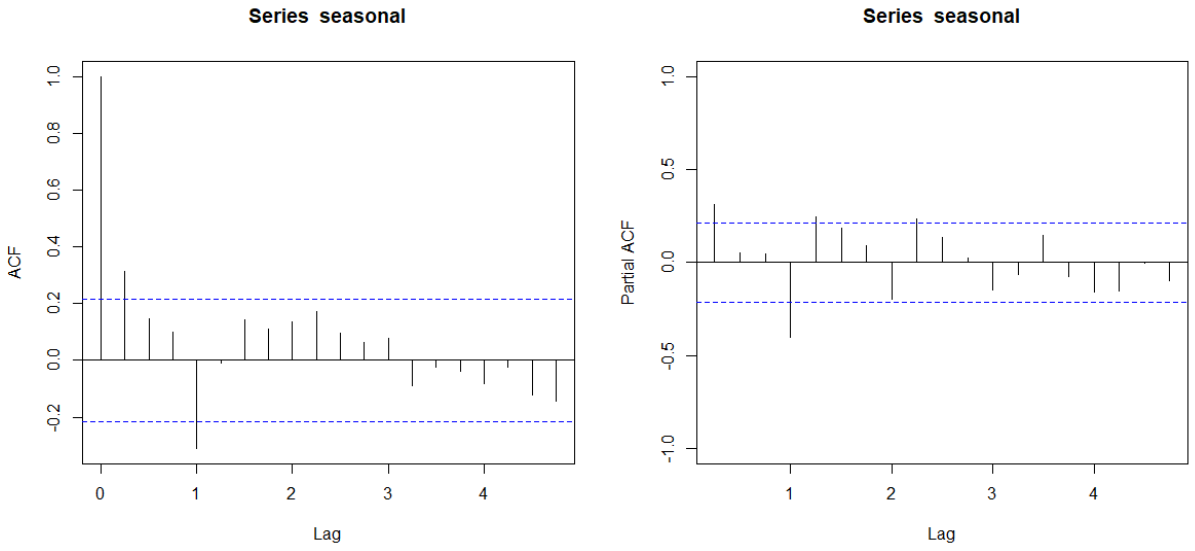
From the ACF plot against the lag we can notice that the correlation coefficient does not falls to zero over a period of lags therefore the data is not stationary. In order to remove the seasonal and pattern and make the data stationary we can difference the data at lag 4. X_t is the number of car registered in a quarter then X_{t-1} is the car registered in the previous quarter. Y_t is the differenced data at lag 4.

$$Y_t = X_t - X_{t-4}$$

$$Y_t = (1 - B^4)\, X_t$$



In the data we can observe that the mean is constant and equal to zero. Overall, the variance is constant. However, there are some outliers in 2010 and 2020 that is due to external factors. Overall, the data is stationary. We can plot the ACF and PACF against the lag to analyse the data.



From the ACF and PACF plot we can observe the data is stationary because over a period of time the correlation drops to zero. So that we are able to create a time series model that is able to predict value based on few lags which were close. If the correlation coefficient does not drops to zero in ACF of PACF we have to consider many parameters which will be a complex model and it is very hard to interpret. The ACF plot shows there is an cut-off at lag 1 which shows its an MA(1) process. The PACF plot against the lag also cut-off at lag 1 this property shows that this can be AR(1) process. It appears

that there is some seasonal pattern that we can observe in ACF and PACF plot but it is not very obvious. Hence we will fit the ARIMA(1, 0, 1) X $(1, 1, 1)_4$ model.

```
Series: data
ARIMA(1,0,1)(1,1,1)[4]

Coefficients:
         ar1      ma1     sar1     sma1
      0.9841  -0.5405  0.0084  -0.9373
s.e.  0.0563   0.1006  0.1279   0.1659

sigma^2 estimated as 3023:  log likelihood=-451.52
AIC=913.04   AICc=913.82   BIC=925.13
```

We can see from the fitted parameters of the model that the contribution of seasonal AR(1) is very less. Therefore we will do a hypothesis test where we check whether we need seasonal AR(1) model(Φ_1) or not

$$H_0: \Phi_1 = 0$$

$$H_1: \Phi_1 \neq 0$$

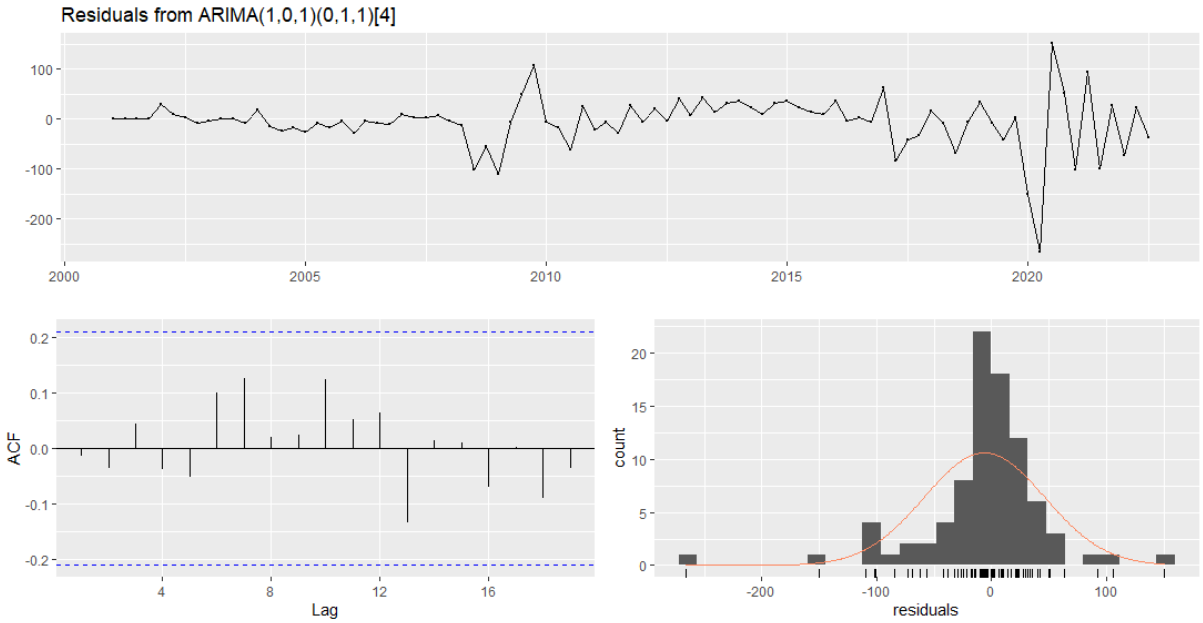
We would reject H_0 if $\left|\frac{\Phi_1}{s.e.(\Phi_1)}\right| > 2$

$s.e.(\Phi_1)$ is the standard error of Φ_1

According to test statistics |0.0084 / 0.1279| = 0.066. Therefore 0.066 is less than 2 so we do not reject the Null Hypothesis H_0. The seasonal AR(1) parameter contribution to the model is negligible so we can remove that parameter and we can fit an ARIMA(1, 0, 1) X $(0, 1, 1)_4$ model.

```
Series: data
ARIMA(1,0,1)(0,1,1)[4]

Coefficients:
         ar1      ma1     sma1
      0.9826  -0.5387  -0.9318
s.e.  0.0504   0.0968   0.1388

sigma^2 estimated as 2991:  log likelihood=-451.52
AIC=911.04   AICc=911.56   BIC=920.72
```

From the ARIMA(1, 0, 1) X $(0, 1, 1)_4$ model we can see that AIC value 911.04 is less than the ARIMA(1, 0, 1) X $(1, 1, 1)_4$ model AIC value 913.04. Hence the newly fitted model is optimal and less complex than the previous model. We can check the residuals plot and residuals ACF plot whether the residuals of the model were white noises (i.e residuals are independent and do not have any correlation which can be used to predict the model).



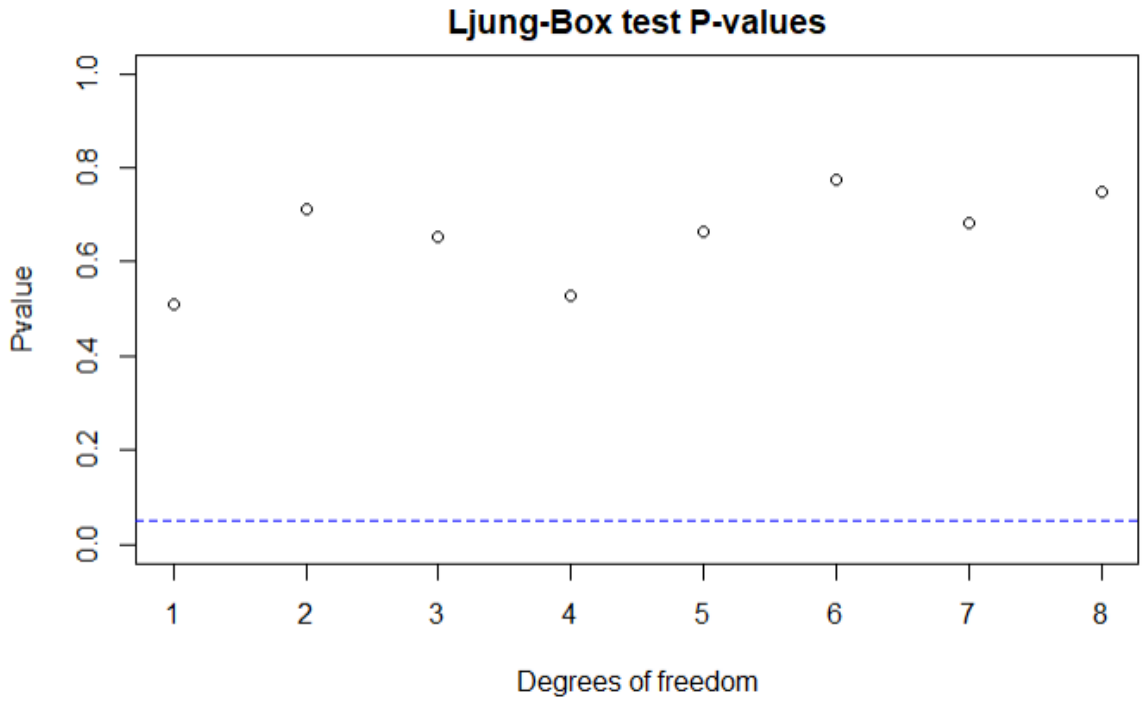
The residuals plot shows there is no correlation between residuals and in the ACF residuals plot against the lag shows that correlation are within 95% confidence interval of correlation coefficients therefore residuals were white noises.

We will also verify residuals are white noise or not using Ljung-Box test. Ljung-Box test is used to check randomness based on number of lags.

The Null Hypothesis H_0 : Residual series is white noise or independent and identically distributed.

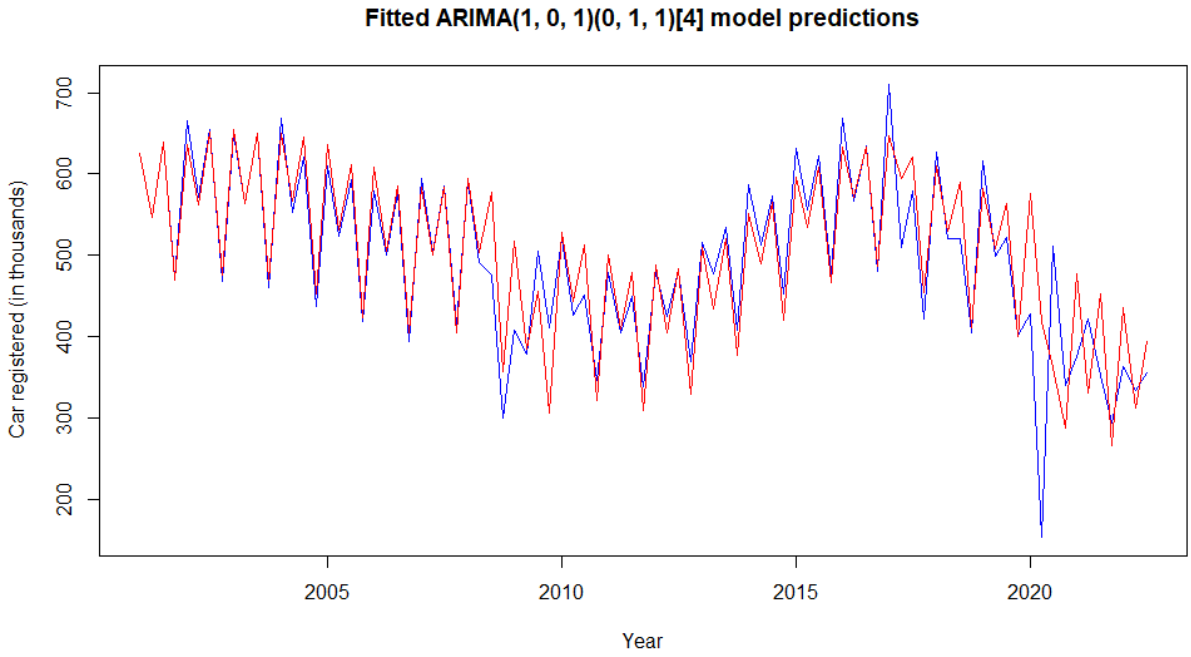
Alternative Hypothesis H_1 : Residual series shows serial correlation.

If Ljung-Box test P-values are less than 0.05 then the series is not white noise but if the P-values are greater than 0.05 then the series is a white noise process.



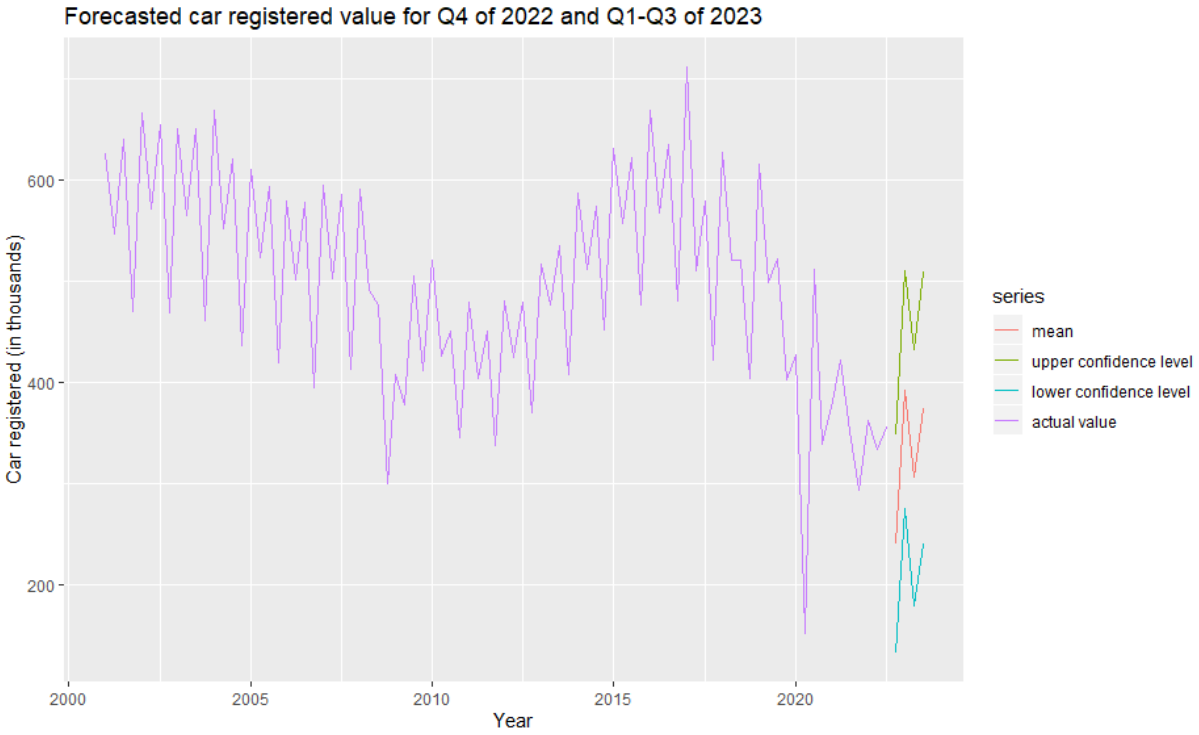
The Ljung-Box test plot shows that p values are above 0.05. Therefore, the null hypothesis is not rejected, and the residual series is a white noise.

We can plot the fitted model in the plot with the actual values of number of cars registered to analyse the reliability of the model



In the fitted plot the blue lines shows the actual value and the red line shows the predicted value. Mostly the model is able to predict the actual values of the time series data. However, it could not predict the sudden fall in the car registered in 2020 because of corona virus pandemic it is an outlier which occur in rare occasion. We can neglect the outlier because there is no pattern for it. Therefore, the model looks like a good fit to the data.

Using this fitted model we will try to forecast the number of cars will be registered for Q4 of 2022 and Q1-Q3 of 2023.



The mean value shows the actual forecast value that we can expect and there may be possibility that the values can oscillate between upper and lower level (i.e we are 95% confident that the values will be between the upper and lower confidence level)

	Point Forecast <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2022 Q4	241.0849	133.6034	348.5663
2023 Q1	392.5959	275.1014	510.0903
2023 Q2	306.1909	179.7903	432.5915
2023 Q3	374.8418	240.4267	509.2568

As we know about the seasonal pattern that Q4 will have low value when compared to other Quarters in a year that is reflected in the forecasted value of Q4 2022 which has 241,000 cars registered. The Q1 2023 has the highest number of car registered among the forecasted values. The difference of the upper and lower confidence value is approximately more than 200,000 cars it is because the pattern in previous years show both increasing and decreasing trend. Therefore, the value can go up or down that's the reason for upper and lower confidence level are far away from the forecasted value.

The fitted model is

$$(1 - 0.98B)(1 - B^4)X_t = (1 - 0.54B)(1 - 0.93B^4)Z_t$$

$$X_t = X_{t-4} + 0.98\,X_{t-1} - 0.98\,X_{t-5} + Z_t - 0.94\,Z_{t-4} - 0.54\,Z_{t-1} + 0.51\,Z_{t-5}$$

Appendix:

Report 1 R code :

```
#add library ggplot which is used to plot graphs
library(ggplot2)

#import the file using read.csv function
df = read.csv("a23_nox.csv")
#new attribute cale data class is created the values of this attribute will be a date class
df$date_class = as.Date(df$ï..date, format="%d/%m/%y")

#the graph is plotted based on date and nitrous oxide level
ggplot(df, aes(x = date_class, y = daily_mean_nox)) + geom_line() +
  xlab("Feb 2017- Sep 2017 \nFig(1)") + ylab("nitrous oxides levels (ug/m3)") + # x and y axis labels
  scale_x_date(date_breaks = "1 month", date_labels = "%b") + # y aixs values format displayed
based on months
  ggtitle("London nitrous oxide levels")

#the dataset is changed to time series data for further analysis
data = ts(df$daily_mean_nox)
#autocorrelation against lag is plotted in chart
acf(data)

#the data is differenced at lag 7
seasonal = diff(data, lag = 7)
#seasonally differenced data is plotted
plot(seasonal)
#autocorrelation of seasonally differenced data
acf(seasonal)
#partial autocorrelation of seasonally differenced data
pacf(seasonal, ylim=c(-1,1))

#ARIMA model with non-seasonal AR(1) and seasonal AR(2) parameters
model = Arima(data, order = c(1,0,0), seasonal = (list(order = c(2,1,0), period = 7)), method = "ML")
# residuals of the model
residuals = residuals(model)
# residuals chart
plot(residuals)
#autocorrelation plot for residuals
acf(residuals)

#Fitted graph of the model
ggplot(df, aes(x =date_class )) + geom_line(aes(y = model$x), color = 'blue') + geom_line(aes(y =
fitted(model)), color = 'red') + xlab("Feb 2017- Sep 2017 \nFig(1)") + ylab("nitrous oxides levels
(ug/m3)") + scale_x_date(date_breaks = "1 month", date_labels = "%b") + ggtitle("London nitrous
oxide levels")
```

Report 2 R code :

```
#Load the libraries
library(dplyr)
library(ggplot2)
library(forecast)

#Load the data
df = read.csv("eng_car_reg.csv")

#The graph of car registrations over years based on four quarters in a year
ggplot(df, aes(x = year, y = no_new_regs, color = quarter)) + geom_line() + labs(y = "car registrations (in thousands)", color = "Quarter", x = "Year")+ ggtitle("Car registrations in England from 2001 to 2022")

#dataset is converted into time series data
data = ts(df$no_new_reg,start=c(2001,1),frequency=4)

#time series data is plotted
plot(data, ylab = "Car Registrations", xlab = "Year")
title(main = "Car registrations in England from 2001 to 2022")

#autocorrelation agains the lag is plotted
acf(data)

# the data is differenced at lag 4
seasonal = diff(data, lag = 4)

# the differenced data at lag 4 is plotted
plot(seasonal, xlab="Year", ylab="car registred")
title(main = "Car registred data differenced at lag 4")

#the ACF and PACF were plotted against the lag
par(mfrow=c(1,2))
acf(seasonal)
pacf(seasonal, ylim= c(-1, 1))

#ARIMA model is fitted for non seasonal AR(1) , MA(1), and seasonal MA(1) with seasonal difference at lag 4
model = Arima(data, order = c(1,0,1), seasonal = (list(order = c(0,1,1), period = 4)), method = "ML")

#residuals of the model were plotted
checkresiduals(model)

#Ljung-Box-test where k is the max value we test
#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)
 }

#Since p+q=1, we run the following command to perform the first ten
#Ljung-Box tests for the model residuals where k is the degress of freedom
#p,q represents number of non-seasonal AR and MA parameters
#P,Q represents seasonal AR and MA parameters
SARIMA_LB<-LB_test_SARIMA(residuals(mod2),max.k=11, p=1, q=1 ,P=0, Q=1)

#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(SARIMA_LB$deg_freedom,SARIMA_LB$LB_p_value,xlab="Degrees of freedom",ylab="Pvalue",
main="Ljung-Box test P-values",ylim=c(0,1))
abline(h=0.05,col="blue",lty=2)

#the forecasted value for the next quarters were obtained using forecast function, level = 95
indicates 95% confidence interval
forecasted_data = forecast(data, h = 4,model = model, level = 95)

#Fitted ARIMA model is plotted with actual values
plot(model$x, col = "blue", xlab = "Year", ylab = "Car registered (in thousands)")
lines(fitted(model), col = "red")
title(main = "Fitted ARIMA(1, 0, 1)(0, 1, 1)[4] model predictions")

# summary of the forecasted data
summary(forecasted_data)

#the upper , lower, predicted and the actual values of the data were merged
combined_data = cbind(forecasted_data$mean, forecasted_data$upper, forecasted_data$lower,
data)

#the forecasted value is plotted
autoplot(combined_data, xlab = "Year", ylab = "Car registered (in thousands)",main = "Forecasted
car registered value for Q4 of 2022 and Q1-Q3 of 2023") + scale_color_discrete(labels = c("mean",
"upper confidence level", "lower confidence level", "actual value"))
```