Time Series Analysis on Downtown Los Angeles Monthly Weather Data (1990-2022)

Name: Praveen Manimaran (praveenmanimaran@ucsb.edu)

Introduction

This project focuses on analyzing and exploring the monthly weather patterns in Downtown Los Angeles. The purpose of this study is to gain insights into the presence of global warming, understand seasonal variations, and identify any significant trends or patterns in the weather data. By examining monthly historical weather data, I hope to understand how the temperature in Los Angeles has changed over the years and see if global warming has impacted temperature levels in Los Angeles.

I chose this dataset to explore the idea that global warming has caused temperatures to rise over the last century or so and to see how it has specifically impacted my hometown, Los Angeles. There have been some studies conducted which predict that Los Angeles will become significantly hotter by 2050, such as a UCLA study that predicts that by 2050 the number of days with temperatures of 95 degrees Fahrenheit or hotter will reach 22 times a year which is more than double that of what Los Angeles County sees currently.

Some methods that I am applying include Exploratory Data Analysis (EDA) which is used to gain insights into the weather dataset and identify any outliers/missing values that currently exist. I also apply 2 time series analysis methods (SARIMA and Lagged Regression) to examine temperature trends and seasonal patterns.

An important discovery that I hope to find from the dataset is to see if temperatures have risen over the last 30 years and if the presence of global warming exists. I also hope to be able to predict monthly average temperatures for 2023 using 30 years of data as well. The findings from this project can be utilized by researchers, local authorities, and citizens living in Los Angeles to understand the climate characteristics of Los Angeles and see how temperatures are year-round as well as help in better decision-making, risk assessment, and planning for the future.

Dataset Description

The dataset used in this project consists of monthly average temperatures for Downtown Los Angeles, spanning the time range from 1990 to 2022. The dataset provides temperature values for each month of the year (1990-2022)

Time Range and Frequency:

The dataset covers a time range of 32 years, from 1990 to 2022. The average temperature values are recorded on a monthly basis (12 times per year).

Values and Size of the Dataset:

The dataset consists of monthly average temperature values for each year. There are 384 observations in the dataset (32 years × 12 months = 384 observations)

Data Source and Collection:

The dataset was obtained from the website of the National Weather Service (NWS) which is a government agency that collects weather data to help provide forecasts and more climate information about various regions across the United States (https://www.weather.gov/wrh/Climate?wfo=lox).

Importance and Background of the Dataset:

Weather data, such as monthly average temperature measurements, allows us to find climate patterns and identify trends in a region to help in risk assessment, and planning for the future. There are hundreds of datasets from the National Weather Service for multiple regions across each state and the nation from the 1900s to 2023. This dataset was specifically chosen to extract the monthly average temperatures for Downtown Los Angeles.

Purpose of Studying the Dataset:

The purpose of studying this dataset is to gain insights into the long-term temperature patterns and fluctuations in Downtown Los Angeles. By analyzing the monthly average temperatures, I hope to discover seasonal variations, identify potential climate trends, and explore the effects of global warming on Los Angeles. This dataset can also be used to predict future monthly average temperatures for the next few year as well and can be used by researchers to understand how Los Angeles might change in the upcoming years.

Methodology

SARIMA Model

In order to model and forecast the monthly temperature data, I decided to use a SARIMA (Seasonal Autoregressive Integrated Moving Average) model to incorporate seasonal patterns in the data. The (p, d, q) terms represent the order of the non-seasonal autoregressive, differencing, and moving average components while The (P, D, Q) terms represent the order of the seasonal autoregressive, differencing, and moving average components. To determine the parameters for the SARIMA model, I decided to use differencing to find out possible parameter values for (p, d, q) x (P, D, Q)$_s$.

I compared the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values associated of each model with different combinations of these parameters. AIC and BIC are used to determine the the model's goodness of fit, where lower values of AIC and BIC indicate a better fit. I selected SARIMA (3,0,3) x (1,1,1)$_4$ model as the model with the best fit for the monthly temperature data in Downtown Los Angeles since it had the lowest AIC and BIC values among the models considered, indicating a good fit to the data.

After selecting the ARIMA (3,0,3) x (1,1,1)$_4$ model, I decided to forecast the next 12 data points (12 months of 2023) and see how the SARIMA model forecasted values for 2023.

Lagged Regression Model

Another method I used to to model and forecast the monthly temperature data is lagged regression which uses lagged values of a variable(temperature) as predictors in a regression model. Lagged values means using past values of monthly average temperature as predictors to forecast future monthly average temperatures.

After looking at various lag plots of different lag values, I found that the best lag value was 12. I decided to create a lagged variable for the temperature data with a lag of 12 months, which is essentially introducing a predictor variable that represents the temperature 12 months ago. This will help me examine how lagged temperature impacts the current temperature.

I performed lagged regression using the lm() function in R to fit a linear regression model with the lagged temperature variable as the predictor and the current temperature as its response variable. From examining the summary of the output of the lagged regression analysis, I was able to determine that my model performed well.

Lastly I decided to forecast monthly average temperature values for the next year (2023).

Results of SARIMA Model

Original Time Series Plot of Temperature Data (1990-2022)

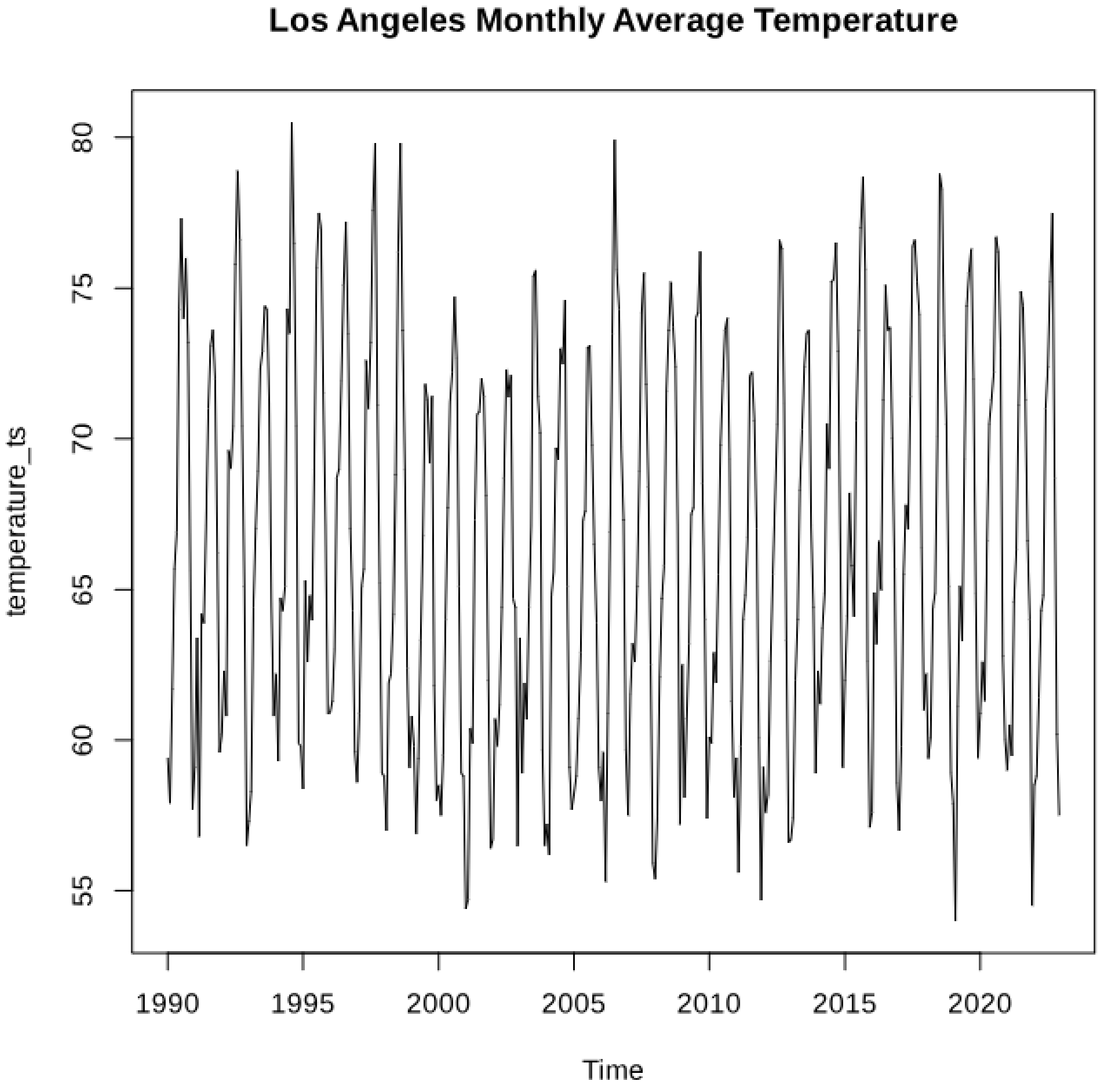


Figure 1

Autocorrelation Function and Partial Autocorrelation Function Plots

What are ACF and PACF plots?

The ACF plot shows the autocorrelation coefficients of a time series at different lags. Each bar on the plot represents the correlation between the time series and itself at that specific lag. The height of the bar shows the how strong the correlation is and the 0 horizontal line represents no correlation. The ACF plot can help us identify the presence of significant autocorrelation in the data and seasonal trends.

The PACF plot shows the partial autocorrelation coefficients of a time series at various lags. The partial autocorrelation represents the correlation between the time series and their lagged values, while controlling the correlations with lags in between.

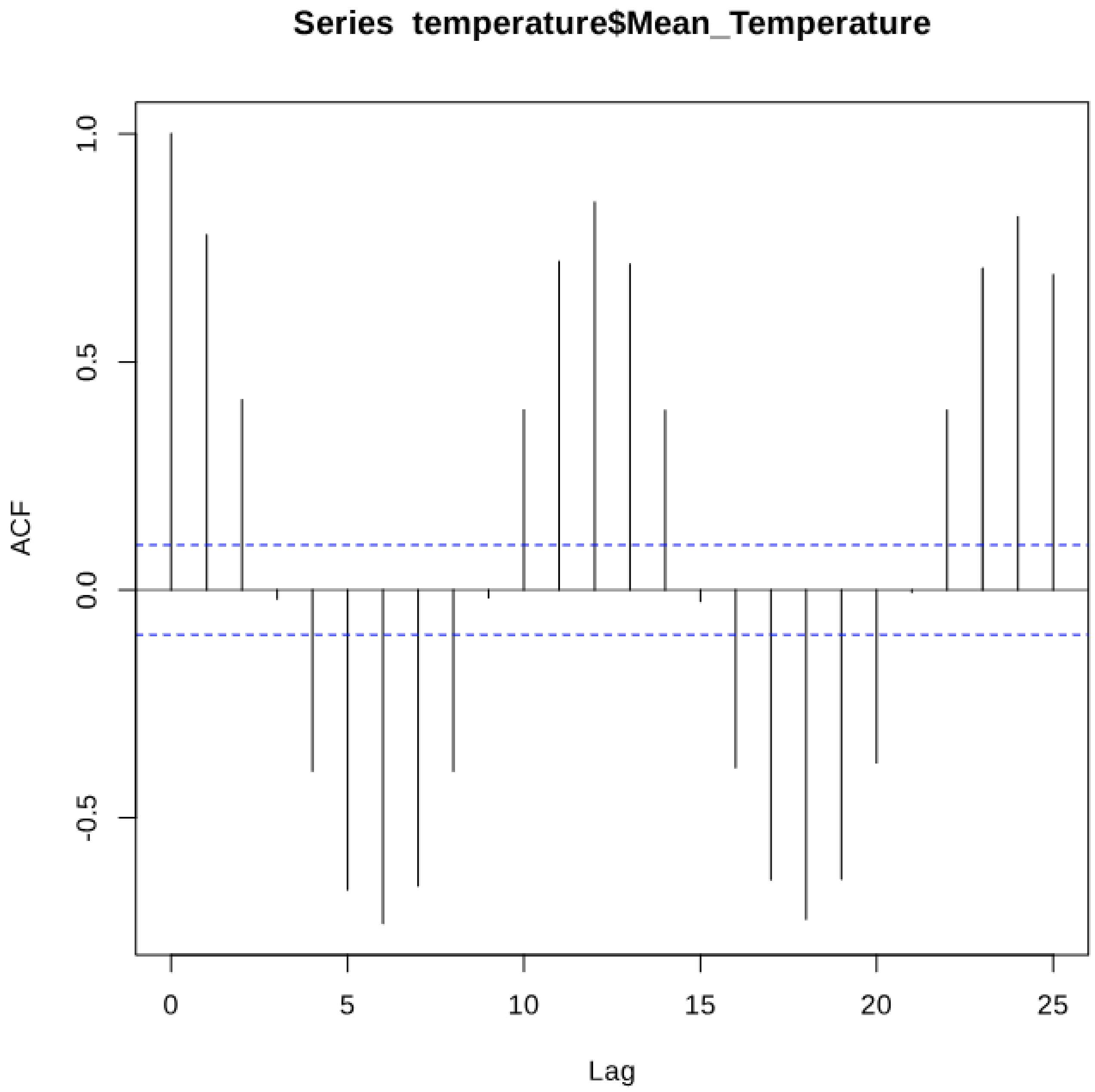


Figure 2

From the ACF Plot, we can see that the plot alternates between positive and negative values, which suggests a seasonal pattern in the data. At lag 12 we can see a large spike which can be interpreted as a high correlation between the time series data and itself at lag 12.

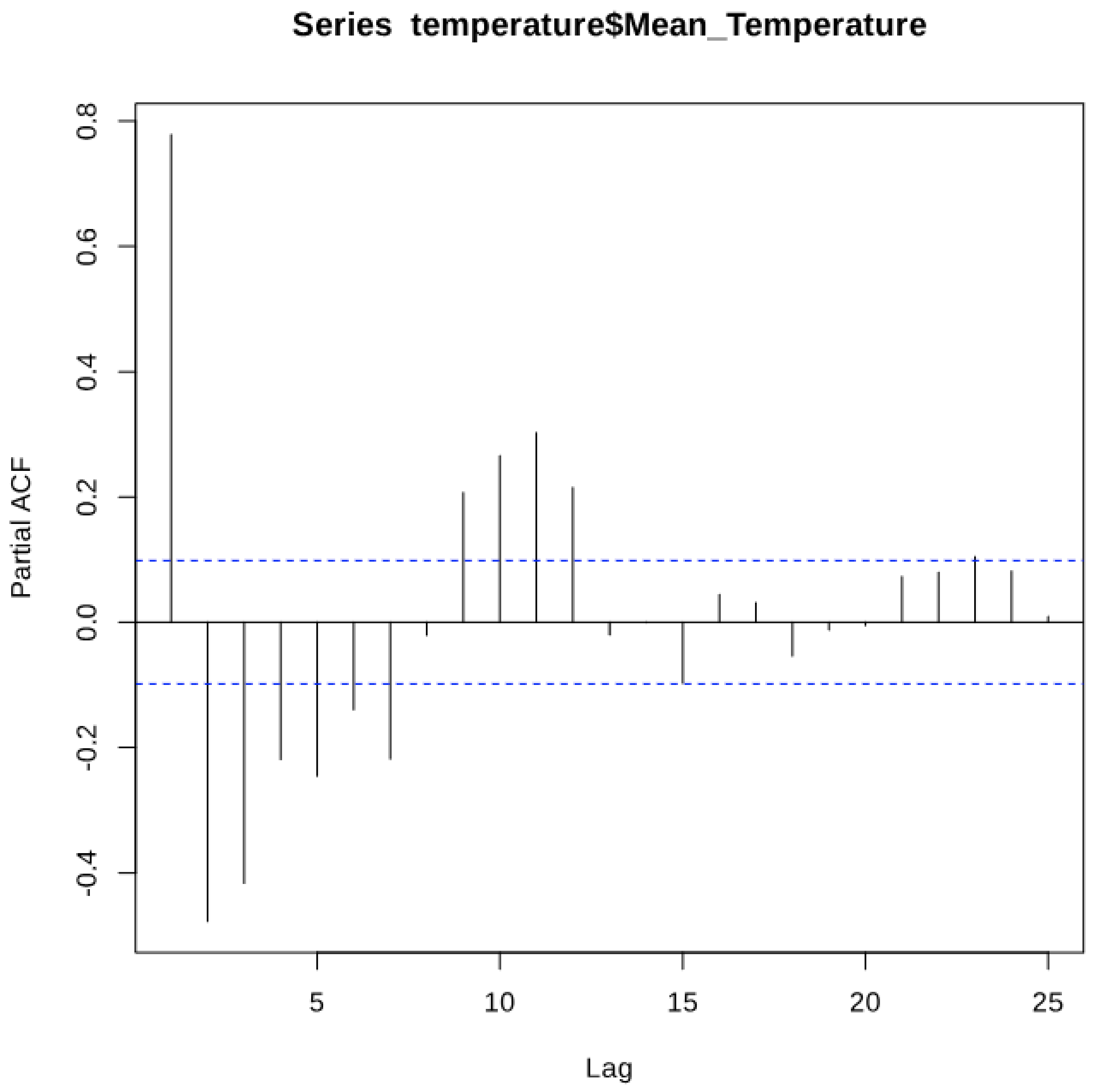
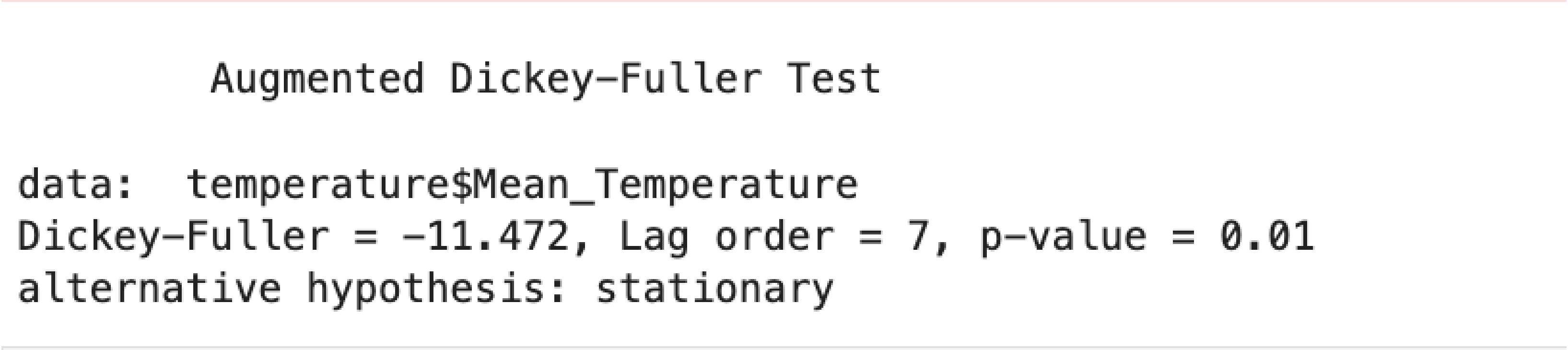


Figure 3

From the PACF Plot, we can see that at lag 12 there is a large spike, which can be interpreted as a high correlation between the time series data and itself at lag 12 (without including other intermediate lags).

Testing For Stationarity

Let us test for stationarity in the time series using the Augmented Dickey Fuller's test. By performing the ADF test, we can determine whether a time series is stationary and see if we need to use differencing to remove a trend.



```
	Augmented Dickey-Fuller Test

data:  temperature$Mean_Temperature
Dickey-Fuller = -11.472, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Figure 4

The output of the ADF test shows us that our p-value of 0.01 is lower than the significance level (typically 0.05) which tells us that we reject the null hypothesis and can conclude that our time series data is in fact stationary.

Deseasonalizing The Dataset

In order to perform time series analysis on the data, it is important to deseasonalize the data when there are seasonal patterns present in the data. We noticed that from the ACF and PACF plots that there are seasonal trends at lag 12, so every 12 months the cycle repeats.

We will use the method of differencing to remove the seasonality present in the dataset. The process involves taking the difference between consecutive data points to create a whole new time series. We will use a difference of 12 since every 12 months the cycle repeats.

Deseasonalized Plot

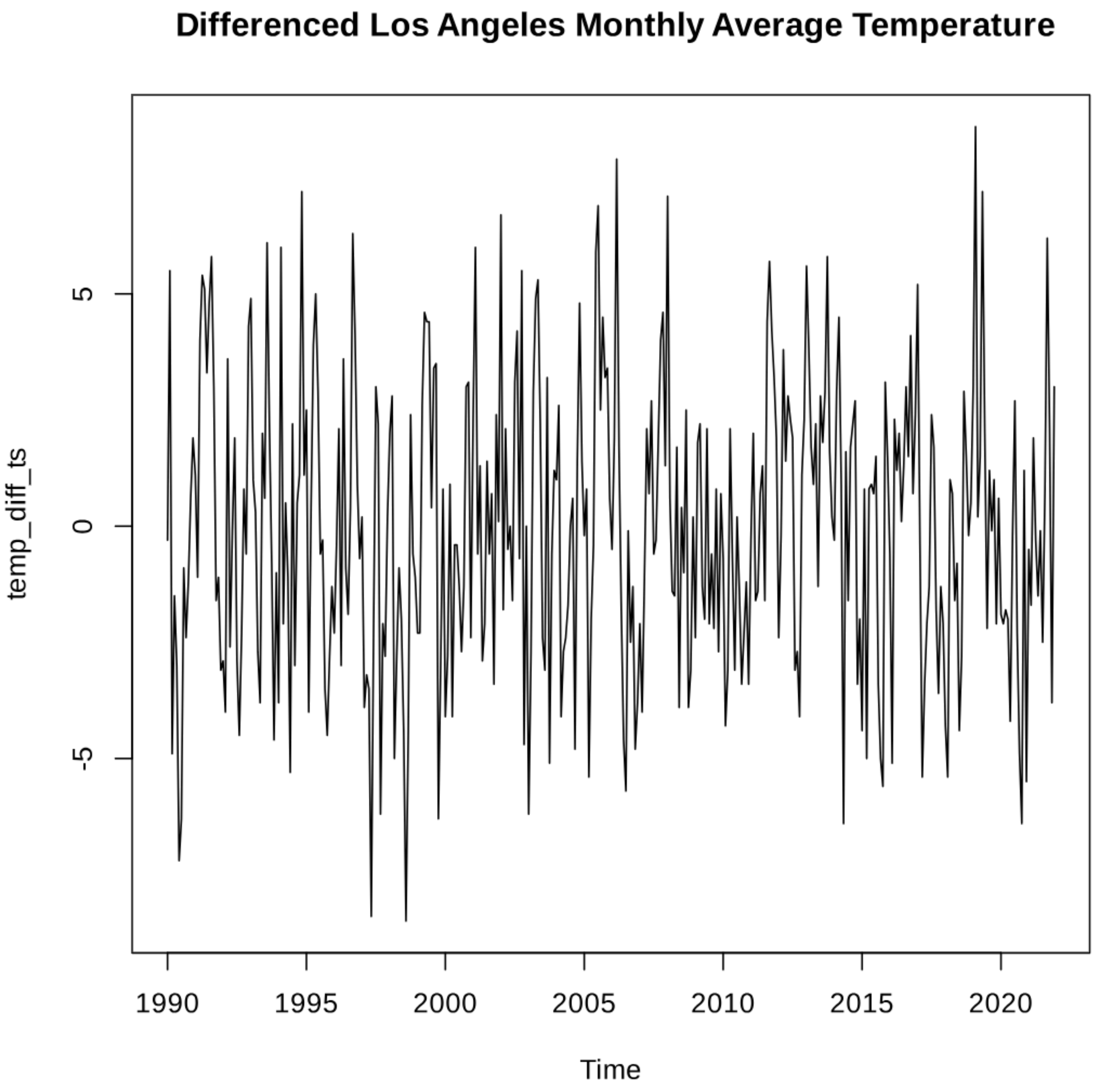


Figure 5

The plot of above shows the Downtown Los Angeles Dataset after it has been differenced by 12 lags (months) to remove seasonality.

Autocorrelation Function and Partial Autocorrelation Function Plots

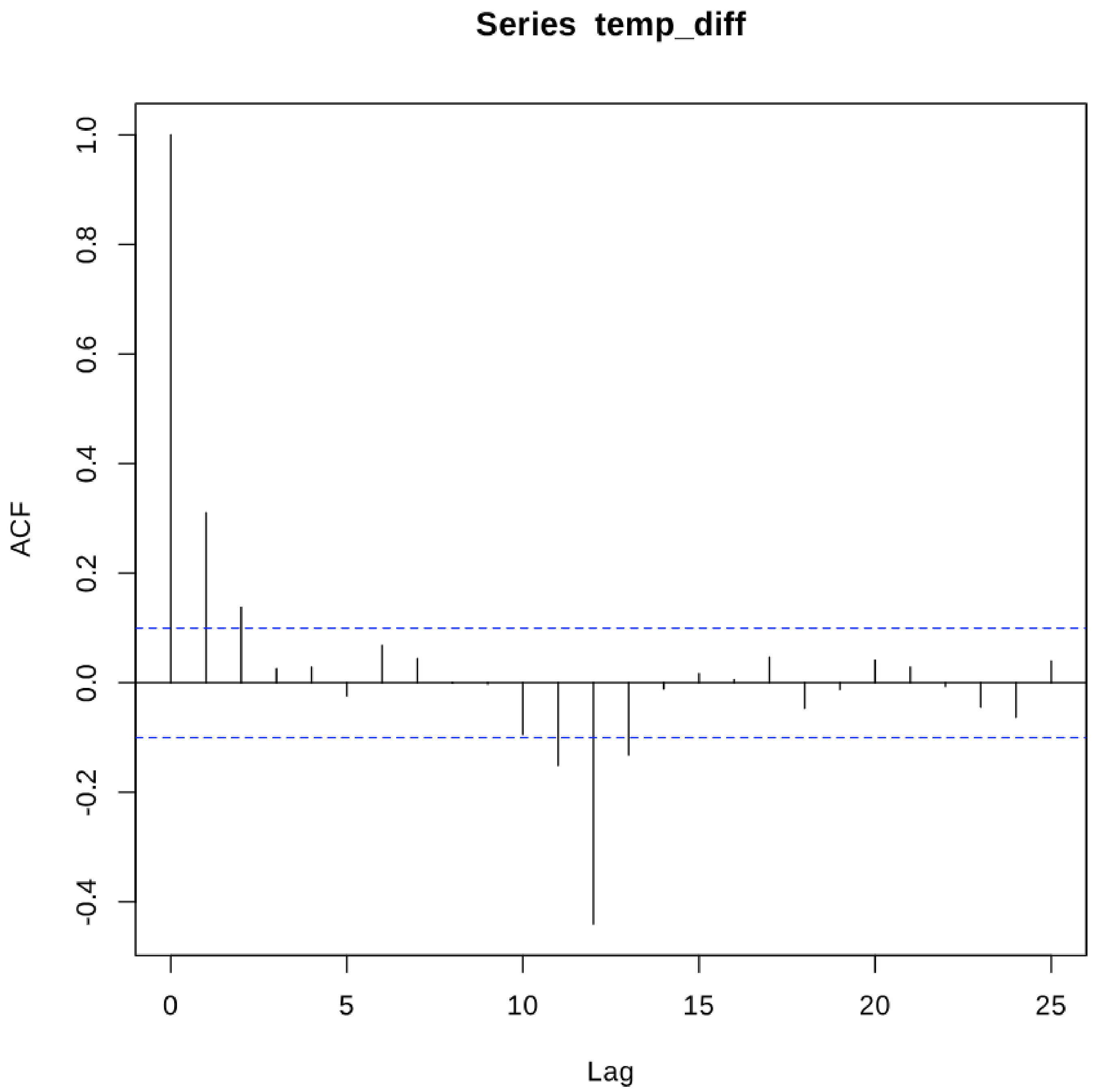


Figure 6

From the ACF Plot, we can see that there are no longer any patterns present in the plot and can use the plot to help us determine the seasonal MA order (Q).

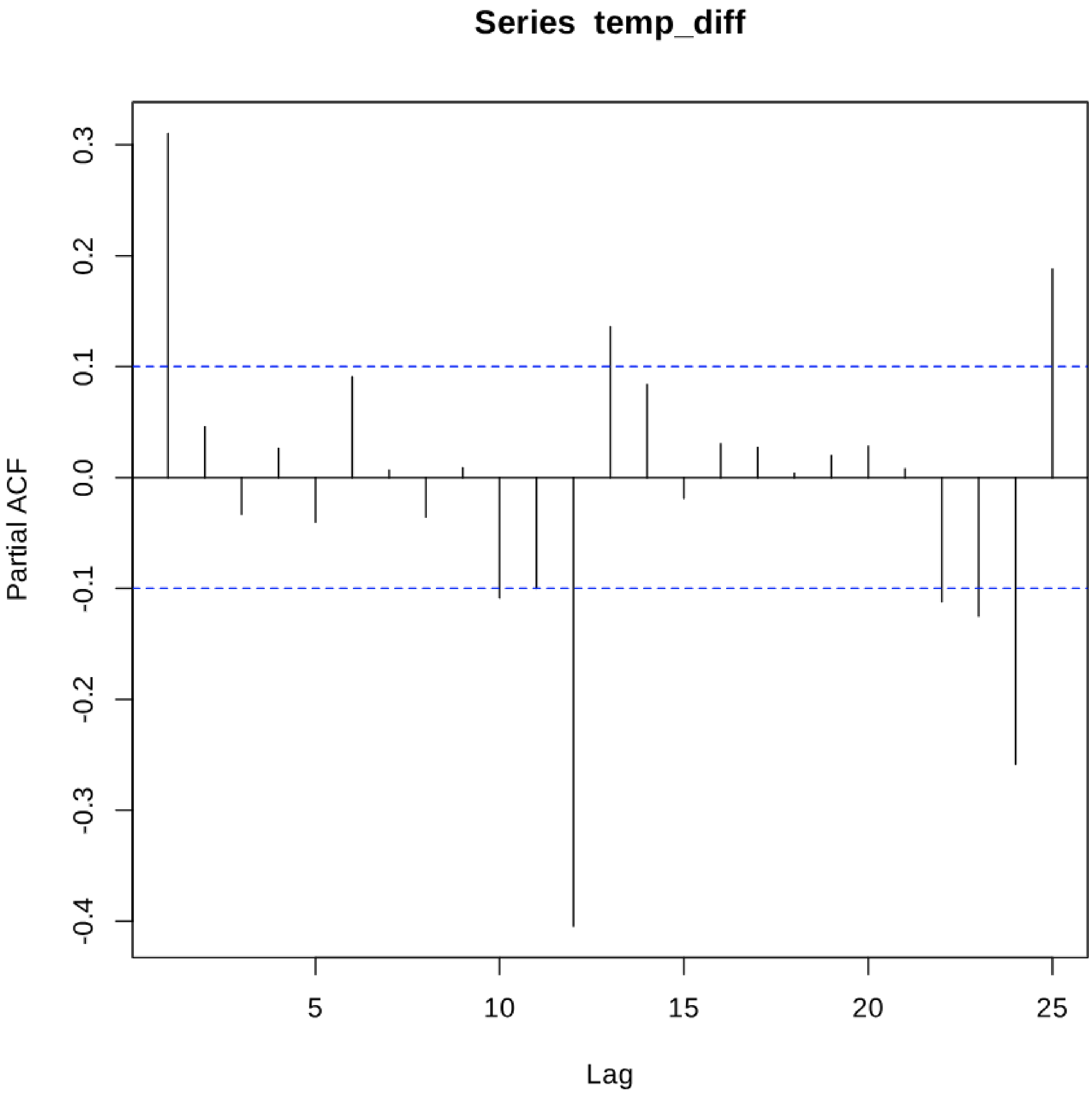
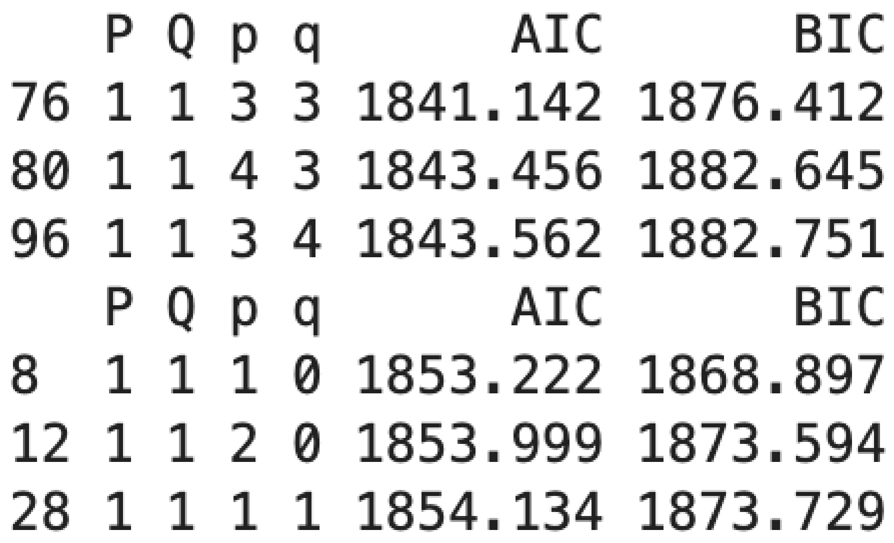


Figure 7

From the PACF Plot, we can see certain spikes and use it to determine the seasonal AR order (P).

Selecting Parameters for SARIMA (p,d,q) x (P,D,Q)$_S$

We will now try to select the parameters for our SARIMA model and figure out which values of p, d, q, P, D, and Q will have the best model performance in terms of AIC and BIC. We will use an interative process of trying various combinations with max values for p, q to be 3 and P, Q to be 3 with d being equal to 1.



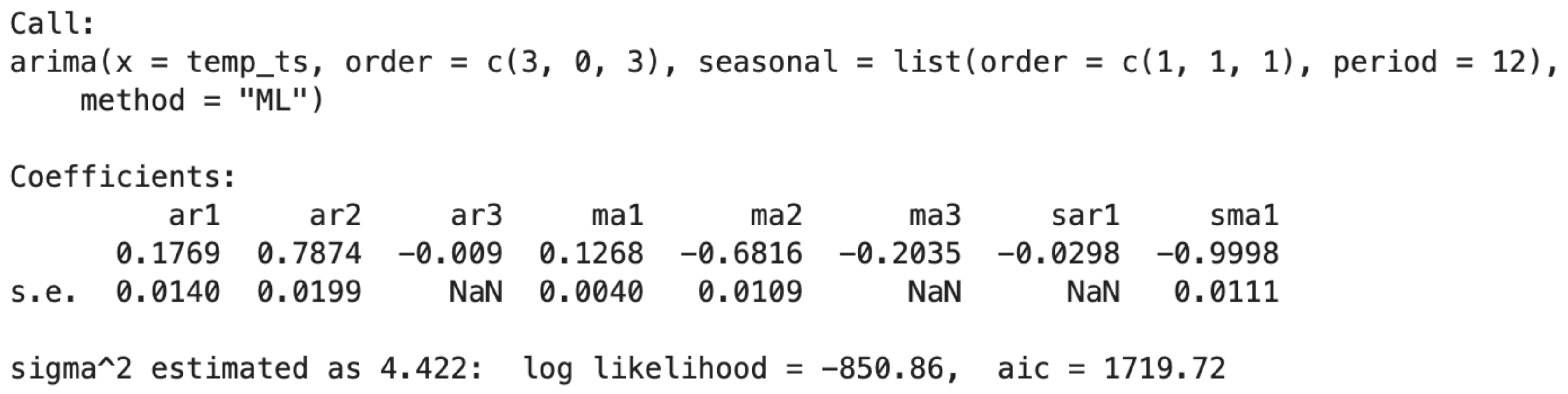
```
   P Q p q      AIC      BIC
76 1 1 3 3 1841.142 1876.412
80 1 1 4 3 1843.456 1882.645
96 1 1 3 4 1843.562 1882.751
   P Q p q      AIC      BIC
8  1 1 1 0 1853.222 1868.897
12 1 1 2 0 1853.999 1873.594
28 1 1 1 1 1854.134 1873.729
```

Figure 8

The table above shows us the top 3 parameter combinations for our SARIMA model in terms of AIC and the top 3 parameter combinations for BIC.

Comparing The Two Best Models

We notice that the parameters with lowest AIC are (3,0,3) x (1,1,1)$_{12}$ and the parameters for lowest BIC are (1,0,0) x (1,1,1)$_{12}$



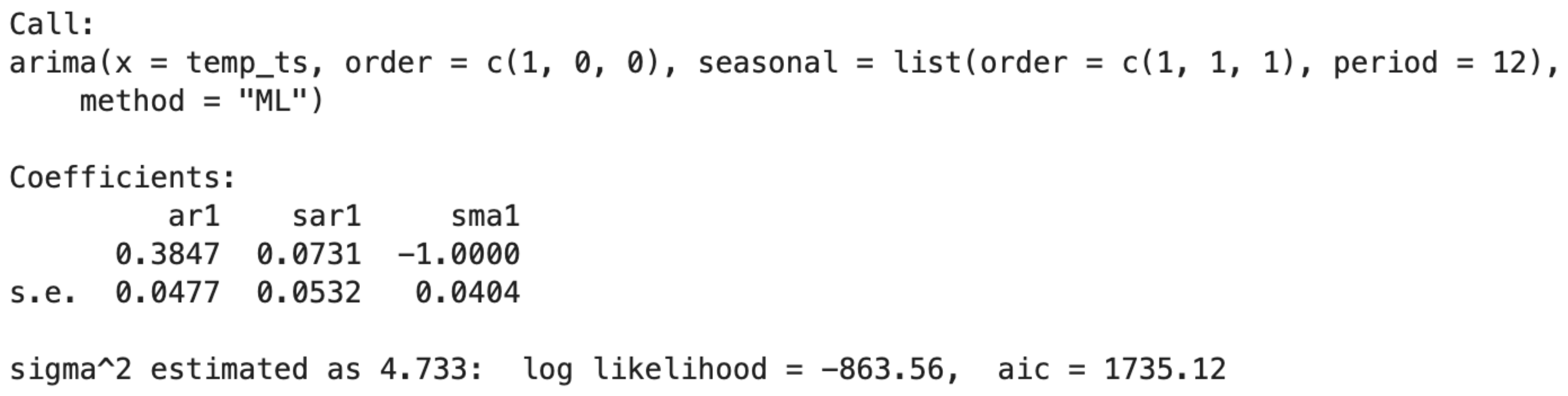
```
Call:
arima(x = temp_ts, order = c(3, 0, 3), seasonal = list(order = c(1, 1, 1), period = 12),
    method = "ML")

Coefficients:
         ar1     ar2     ar3     ma1      ma2      ma3     sar1     sma1
      0.1769  0.7874  -0.009  0.1268  -0.6816  -0.2035  -0.0298  -0.9998
s.e.  0.0140  0.0199     NaN  0.0040   0.0109      NaN      NaN   0.0111

sigma^2 estimated as 4.422:  log likelihood = -850.86,  aic = 1719.72
```

Figure 9

The figure above shows the SARIMA summary for (3,0,3) x (1,1,1)$_{12}$ with its AIC value.



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

Coefficients:
         ar1    sar1     sma1
      0.3847  0.0731  -1.0000
s.e.  0.0477  0.0532   0.0404

sigma^2 estimated as 4.733:  log likelihood = -863.56,  aic = 1735.12
```

Figure 10

The figure above shows the SARIMA summary for (1,0,0) x (1,1,1)$_{12}$ with its AIC value.

Residual Plots and ACF & PACF Plots

Residual plots can help us analyze the differences between the observed values and the predicted values (residuals) of a statistical model. If we see randomness in the plot and no patterns, we can can use the plot to determine that the assumptions of the model are met.

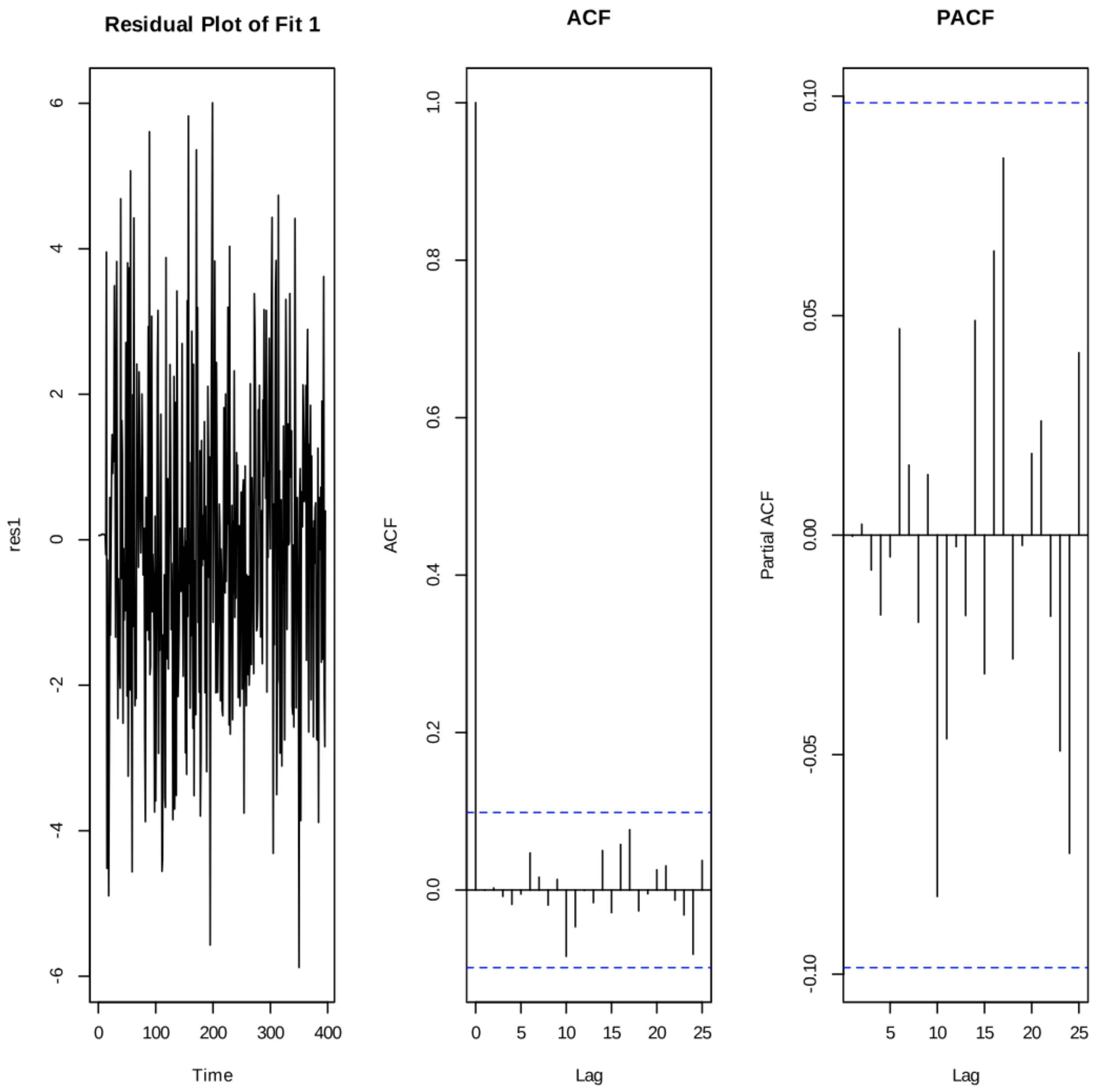


Figure 11

In the residual plot we can see that there are no patterns and the points appear to be random. In the ACF and PACF plots, we can see that no patterns exist and that all bars are within the boundary which indicate that there are no significant lags.

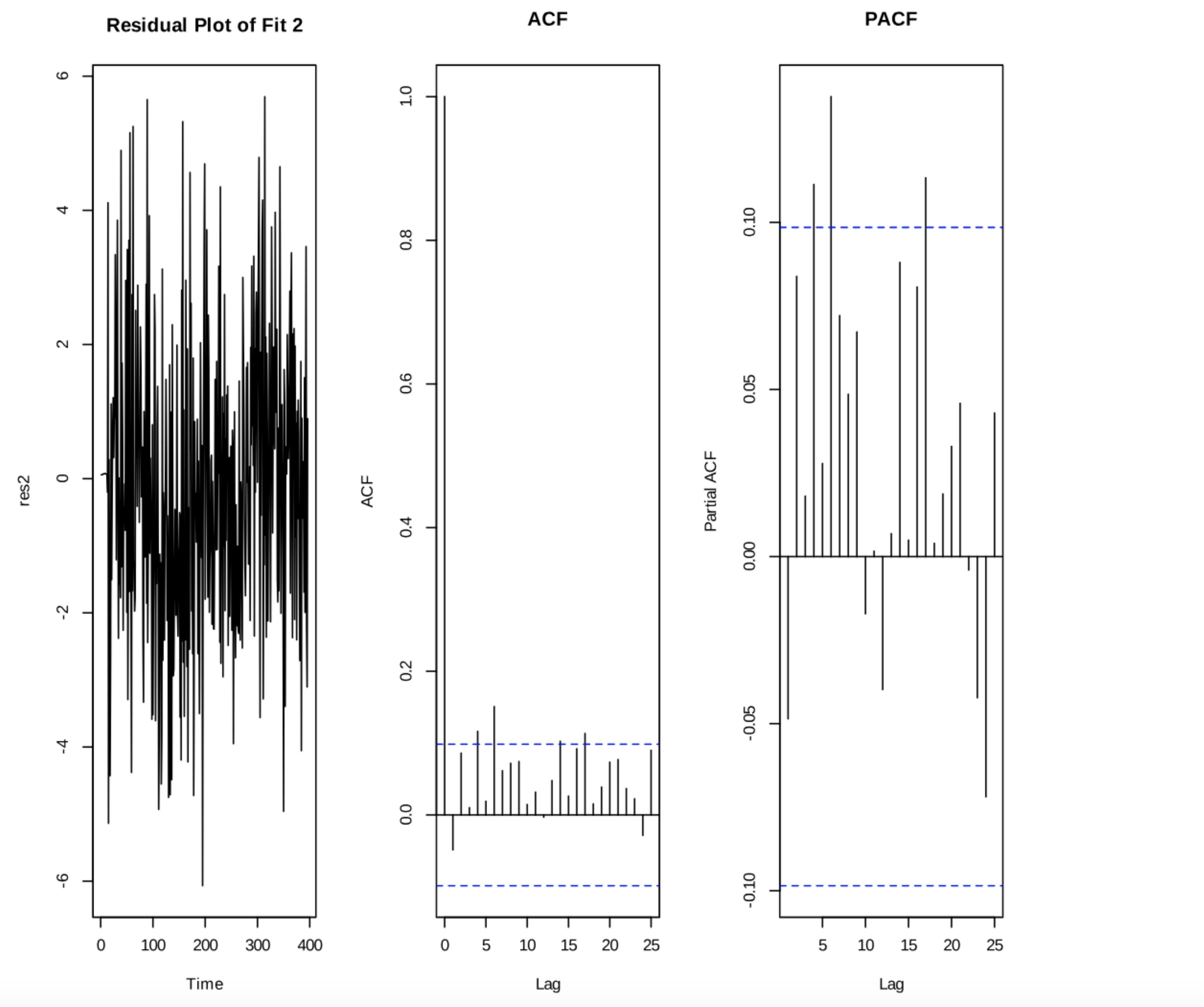


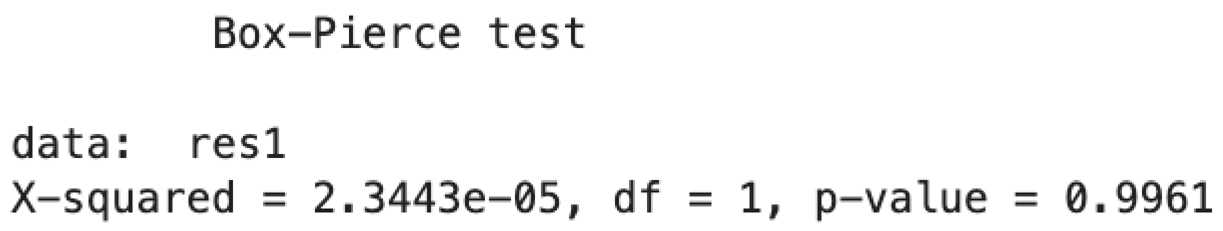
Figure 12

In the residual plot we can see that there are no patterns and the points appear to be random. In the ACF and PACF plots, we can see that no patterns exist; however, not all bars are within the boundary which indicate that there could be some significant lags.

Based off the figures above, I have decided that the best parameters for our SARIMA model are (3,0,3) x $(1,1,1)_{12}$.

Box-Pierce Test

The Box-Pierce Test can give us a statistical assement to determine if the residuals are independent of one another (randomness).



```
        Box-Pierce test

data:  res1
X-squared = 2.3443e-05, df = 1, p-value = 0.9961
```

Figure 13

Since the p-value is 0.9961, which is greater than the typical significance level of 0.05, we do not have enough evidence to reject the null hypothesis of independence. We can conclude that our residuals are independent.

QQ Plot for Normality w/ Histogram

A QQ plot (quantile-quantile plot) is a graph that can help us determine if our dataset satisfies the normality assumption. When we combine it with a histogram, we can see the distribution of the data and see if it has a bell shaped curve.

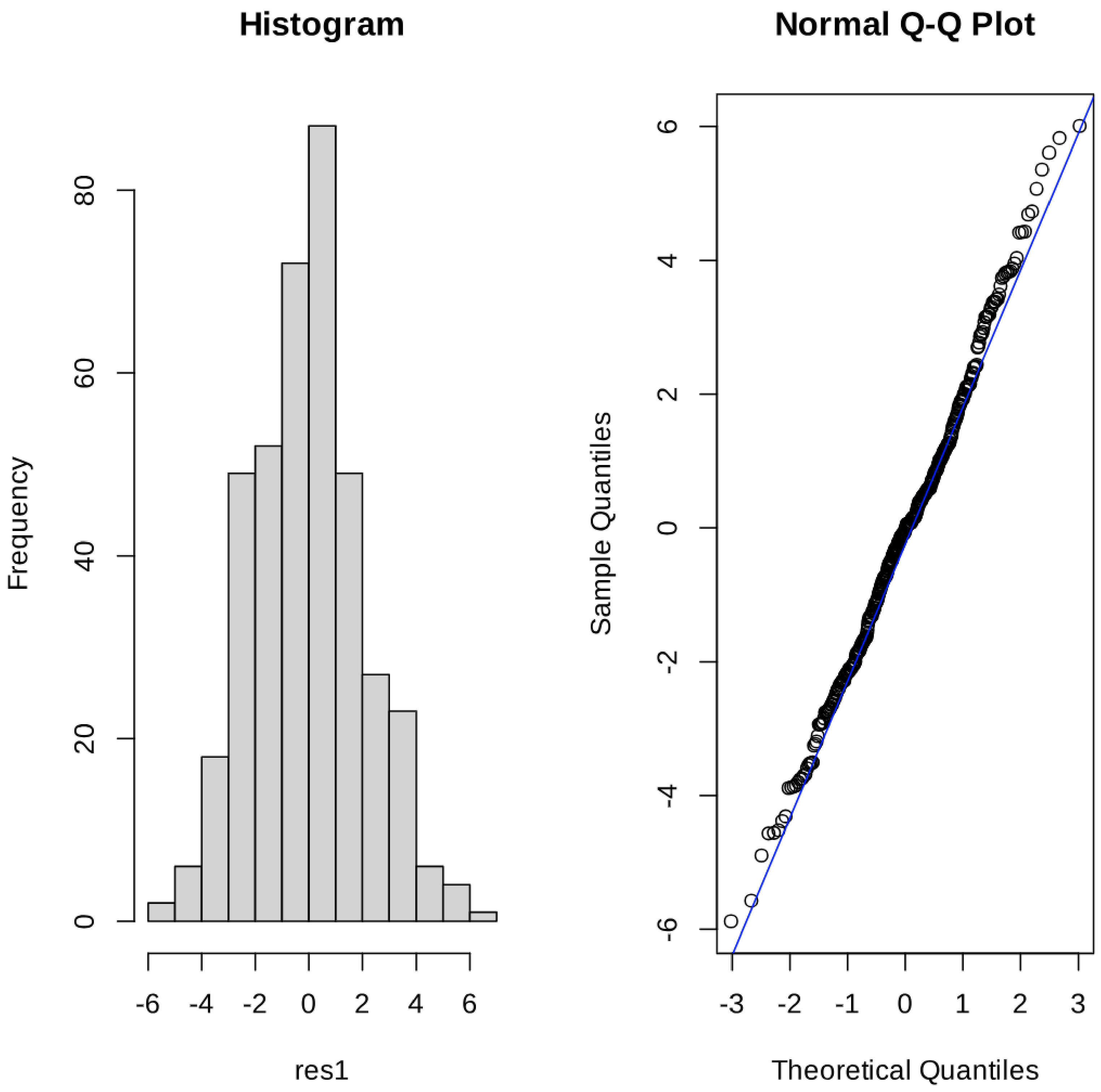


Figure 14

We can see in the QQ plot that most of the points on the plot roughly fall along a straight line, which means that our dataset meets the normality assumption. We also notice that the distribution of the data seems to resemble that of a bell curve in the histogram.

Forecasting the Next 12 Months (2023)

Let us now forecast the average temperatures for the next 12 months of 2023 using our SARIMA model.

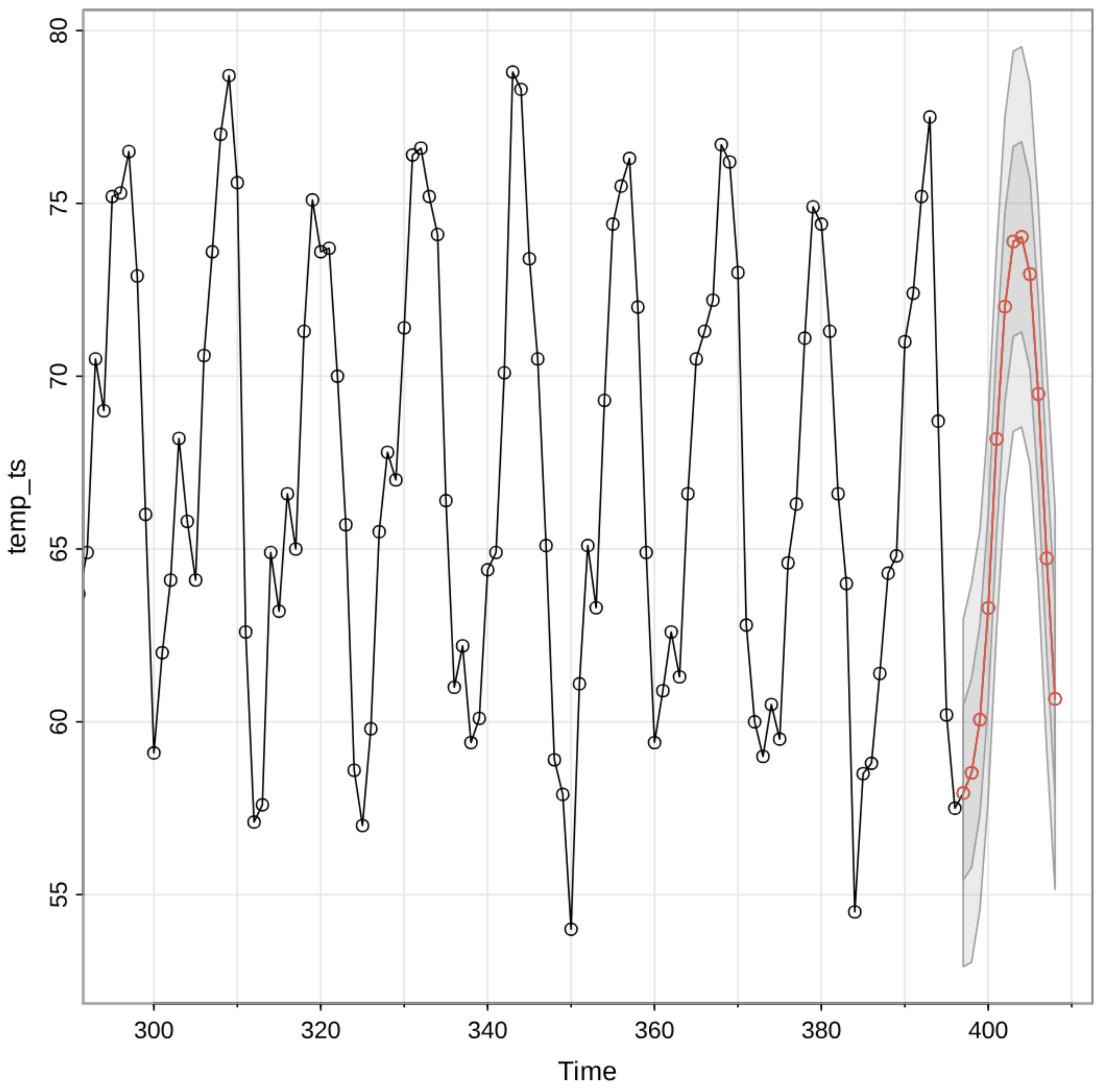


Figure 15

We can see our forecasted values for 2023 in red.

Results of Lagged Regression Model

Lag Plot

A lag plot is a scatterplot of lagged values, which can help us visualize the relationship between an observation and its lagged values. It helps in understanding the autocorrelation structure of a time series. I decided to create a 12 lag plots for lag values 1 to 12 to determine which lagged value formed the strongest linear relationship which can suggest the presence of autocorrelation.

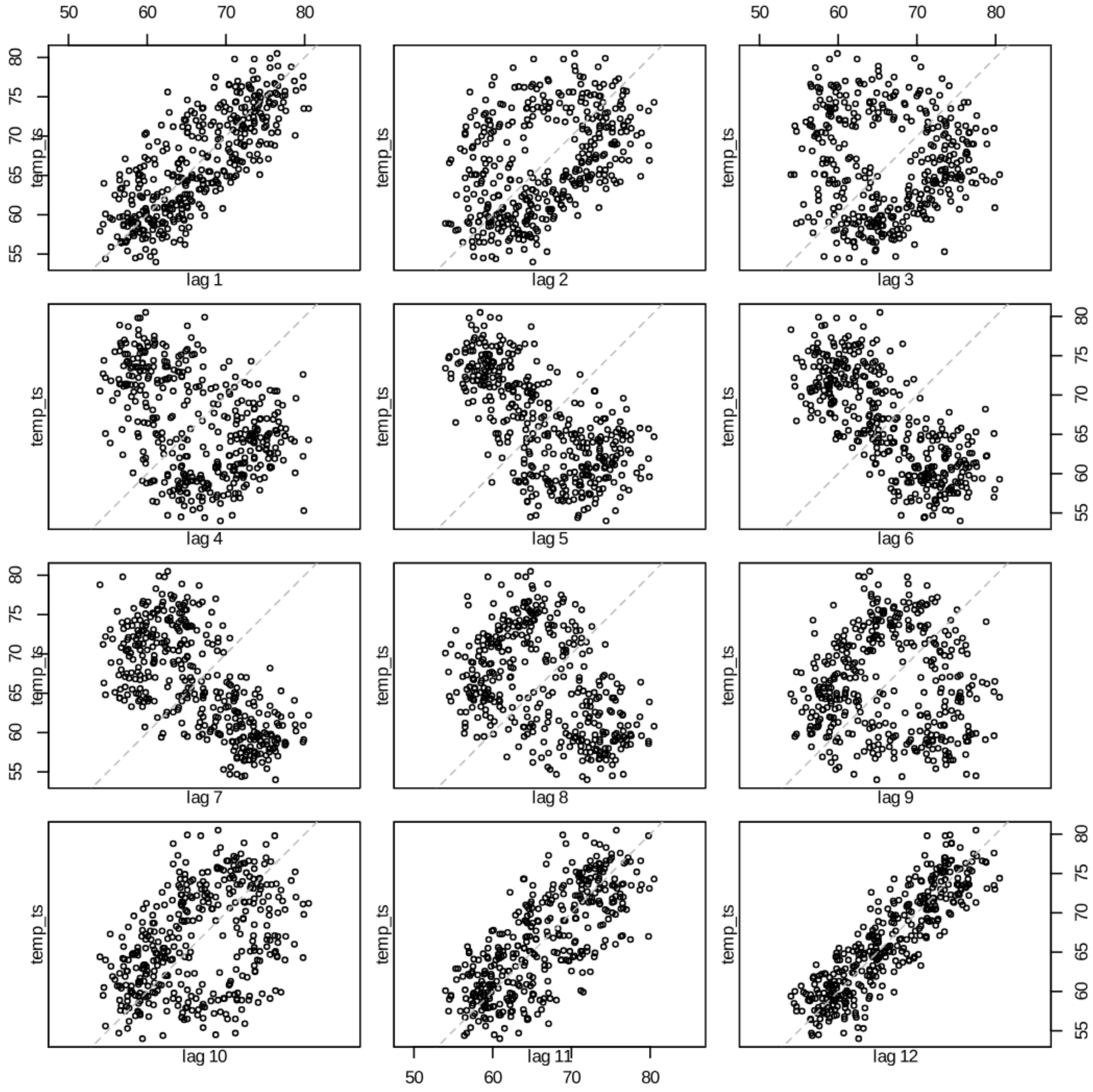
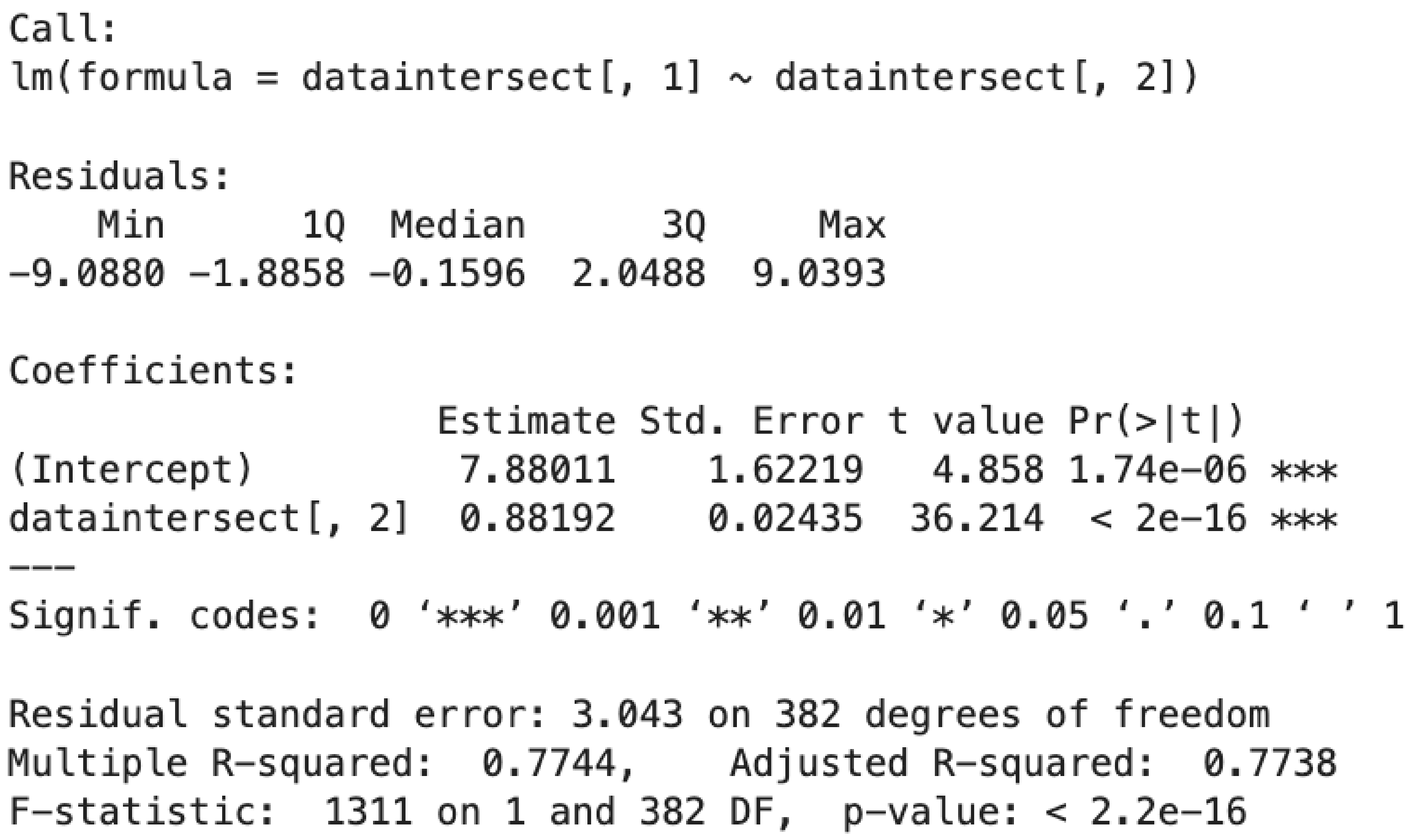


Figure 16

From the lag plot above, we can see that at lag 12 there is a very strong linear relationshop between the observation and its lagged value.

Fitting Lagged variable to lm()

After creating a lagged variable with a lag value of 12, I decided to fit the lagged variable to the lm() function in R and print its summary to determine the coefficients, standard errors, and p-values.



```
Call:
lm(formula = dataintersect[, 1] ~ dataintersect[, 2])

Residuals:
    Min      1Q  Median      3Q     Max 
-9.0880 -1.8858 -0.1596  2.0488  9.0393 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         7.88011    1.62219   4.858 1.74e-06 ***
dataintersect[, 2]  0.88192    0.02435  36.214  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.043 on 382 degrees of freedom
Multiple R-squared:  0.7744,	Adjusted R-squared:  0.7738 
F-statistic:  1311 on 1 and 382 DF,  p-value: < 2.2e-16
```

Figure 17

From the summary we see that the intercept value is 7.88011 and that the coefficient of the lagged variable is 0.88192, which is very close to 1, indicating that there is a good chance that the model is a good fit. The coefficient estimate is also statistically significant (p-value < 2.2e-16). Overall, the output indicates that the model is a good fit.

Forecasting the Next 12 Months (2023)

Let us now forecast the average temperatures for the next 12 months of 2023 using our lagged regression model.

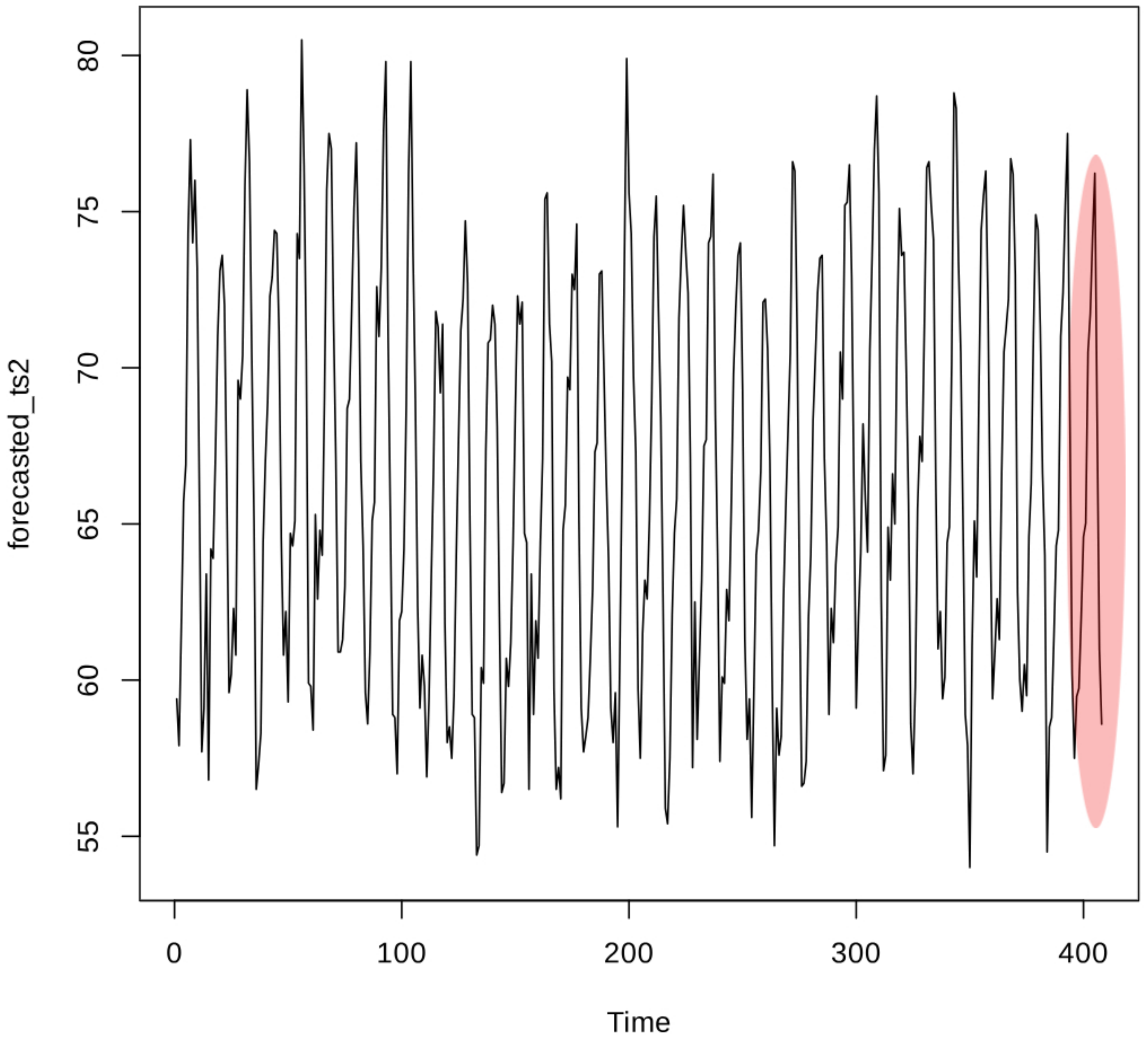


Figure 18

We can see that forecasted values in the red circle.

Conclusion and Future Study

In this project, the focus was primarily on analyzing the monthly average temperature data for Los Angeles over a period of approximately 30 years. The purpose of this time series analysis was to explore any patterns or trends in the temperature data and to see if there was any impact of global warming on Downtown Los Angeles.

One of the key discoveries from the analysis was that the average temperatures in Los Angeles did not show any increase over the past 30 or so years. This finding was unexpected, as I was under the assumption that global warming would have raised the local temperature by 1-3 degrees in Fahrenheit. I am able to successfully conclude that changes in Downtown Los Angeles temperature do not quite align with the increase in the global average temperature.

In order to improve upon this project, I would definitely collect more data on additional weather factors in Downtown Los Angeles such as precipitation, humidity, or wind patterns to get a more complete analysis on if global warming has impacted any of these factors in the past 30 years. I am also interested in collecting monthly global average temperature data to compare the relationship between global and local climate variations.

In conclusion, this project highlights the importance of analyzing and interpreting long-term temperature data to assess the impact of global warming accurately. It emphasizes the need for continued exploration of different climate factors to get a better understanding of the impacts of climate change at both global and regional scales.

References

Websites

https://www.weather.gov/wrh/Climate?wfo=lox

https://climatecheck.com/california/los-angeles#:~:text=Heat%20risk%20in%20Los%20Angeles%2C%20CA&text=By%202050%2C%20people%20in%20Los,more%20informatic

https://la.curbed.com/2019/8/19/20726773/los-angeles-hotter-temperature-climate-change

Appendix

In [1]:
```
#Final Project
#Author: Praveen Manimaran
```

In [46]:
```
#LOAD ALL PACKAGES

#install.packages("astsa")
#library(astsa)
#install.packages("sarima")
#library(sarima)
#install.packages("tseries")
#library("tseries")
```

In [53]:
```
#FIGURE 1 CODE

#Reading in LA Monthly Temperature File
temperature <- read.csv('Monthly_Mean_Temperature_LA.csv')

colnames(temperature) <- c("Year", "Month_Number", "Mean_Temperature")
#Exploring the dataset by plotting as a time series
temperature_ts <- ts(temperature[, 3], start=c(1990, 1), frequency=12)
#plot.ts(temperature_ts, type="l", main="Los Angeles Monthly Average Temperature")

#head(temperature)
```

In [31]:
```
#FIGURE 2 CODE
#ACF Plot of Dataset
#acf(temperature$Mean_Temperature)
```

In [32]:
```
#FIGURE 3 CODE
#PACF Plot of Dataset
#pacf(temperature$Mean_Temperature)
```

In [33]:
```
#FIGURE 4 CODE

#Let us test for Stationarity using Augmented Dickey Fuller's test
#adf.test(temperature$Mean_Temperature)
```

In [52]:
```
#FIGURE 5 CODE
#Deseasonalizing
#temp_diff <- diff(temperature$Mean_Temperature, 12)
```

```
#temp_diff_ts <- ts(temp_diff, start=c(1990, 1), frequency=12)
#plot.ts(temp_diff_ts, type="l", main=" Differenced Los Angeles Monthly Average Temperature")

#temp_diff
```

In [34]:

```
#FIGURE 6 CODE
#acf(temp_diff)
```

In [35]:

```
#FIGURE 7 CODE
#pacf(temp_diff)
```

In [60]:

```
#FIGURE 8 CODE


#SARIMA (p, d, q) x (P, D, Q) model

#temp_ts <- as.ts(temperature$Mean_Temperature)
#model_1 <- sarima(temp_ts, p = 1, d = 0, q = 1, P = 0, D = 1, Q = 1)
#model_2 <- sarima(temp_ts, p = 2, d = 0, q = 2, P = 0, D = 1, Q = 1)
#model_3 <- sarima(temp_ts, p = 0, d = 1, q = 1, P = 1, D = 0, Q = 1)

#model_1 <- arima(temp_ts, order=c(1, 0, 1),
#seasonal=list(order=c(0, 1, 1), period=12),
#method="ML")

#model_2 <- arima(temp_ts, order=c(2, 0, 2),
#seasonal=list(order=c(0, 1, 1), period=12),
#method="ML")

#model_3


#df <- data.frame(expand.grid(P = 0:1, Q = 0:1, p = 0:4, q = 0:4), AIC = NA, BIC = NA)
#for (i in 1:nrow(df)) {
#  m <- df[i, ]
#  fit <- arima(temp_diff, order = c(m$p, 0, m$q),
#                seasonal = list(order = c(m$P, 1, m$Q), period = 12),
#                method = "ML")
#  df[i, ]$AIC <- fit$aic
#  df[i, ]$BIC <- BIC(fit)
#}

# Print the summary of each model
#print(summary(model_1))
#print(summary(model_2))
#print(summary(model_3))
#model_2

# Top 3 models based on AIC
#top_3_aic <- df[order(df$AIC)[1:3], ]
#print(top_3_aic)

# Top 3 models based on BIC
#top_3_bic <- df[order(df$BIC)[1:3], ]
#print(top_3_bic)
```

In [61]:

```
#FIGURE 9 CODE


#(3,0,3) x (1,1,1)_4
model_1<-arima(temp_ts, order=c(3, 0, 3),
seasonal=list(order=c(1, 1, 1), period=12),
method="ML")

#model_1
```

In [38]:

```
#FIGURE 10 CODE

#(1,0,0) x (1,1,1)_4
model_2<-arima(temp_ts, order=c(1, 0, 0),
seasonal=list(order=c(1, 1, 1), period=12),
method="ML")

#model_2
```

In [39]:

```
#FIGURE 11 CODE

#Residual Plot 1
res1 <- residuals(model_1)
#par(mfrow=c(1, 3))
#plot.ts(res1, type="l", main="Residual Plot of Fit 1")
#acf(res1, main="ACF")
#pacf(res1, main="PACF")
```

In [40]:

```
#FIGURE 12 CODE
#Residual Plot 2
res2 <- residuals(model_2)
#par(mfrow=c(1, 3))
#plot.ts(res2, type="l", main="Residual Plot of Fit 2")
#acf(res2, main="ACF")
#pacf(res2, main="PACF")
```

In [15]:

```
#Model 1 seems to
```

In [41]:

```
#FIGURE 13 CODE

#Box Test
#Box.test(res1)
```

In [42]:

```
#FIGURE 14 CODE

#QQ Plot for Normality
#par(mfrow=c(1, 2))
#hist(res1, main = "Histogram")
#qqnorm(res1); qqline(res1, col="blue")
```

In [43]:

```
#Shapiro Test for Normality
#shapiro.test(res1)
```

In [48]:

```
#FIGURE 15 CODE

#Forecasting next 12 months

#pred_12 <- sarima.for(temp_ts, n.ahead=12, plot.all=F,
#p=3, d=0, q=3, P=1, D=1, Q=1, S=4)

df2 <- unlist(pred_12$pred[1:12])
forecasted_ts <- c(temp_ts,df2)
#plot.ts(forecasted_ts)
```

In [45]:

```
#LAGGED REGRESSION
```

In [54]:

```
#FIGURE 16 CODE
#lag.plot(temp_ts, lags = 12, do.lines = FALSE)
```

In [56]:

```
#FIGURE 17 CODE

# Create the lagged variable
lag12_temp <- lag(temperature$Mean_Temperature, 12)

#str(lag12_temp)
dataintersect = ts.intersect(temp_ts, lag12_temp, dframe= TRUE)
#head(dataintersect)

fit1 <- lm(dataintersect[,1] ~ dataintersect[, 2])
#summary(fit1)


#reg_data <- data.frame(temperature$Mean_Temperature, lag12_temp)

# Fit the lagged regression model
#model <- lm(temperature$Mean_Temperature ~ lag12_temp, data = reg_data)

# Print the model summary
#summary(model)

#reg_data
```

In [57]:

```
# Obtain the last 12 values of lagged temperatures
last_lagged_temps <- tail(lag12_temp, 12)

# Perform the forecast using the coefficient estimate
forecast <- 7.88011 + 0.88192 * last_lagged_temps

# Print the forecasted temperatures for the next 12 months
#print(forecast)
```

In [58]:

```
#FIGURE 18 CODE
forecasted_ts2 <- c(temp_ts,forecast)
#plot.ts(forecasted_ts2)

#plot.ts(temp_ts)
```