TIME SERIES ANALYSIS

DENNIS MULUMBI KYALO

04-12-2022

Contents

PART A **2**

PART B **3**

Section 1 : Plastic Sales Data 3

Section 2 : Electricity Data 6

2a. Stationarity 6

2b. ARIMA models 7

2c. Model residuals 8

2d. Forecasting 9

2e. Intervals 9

REFERENCES **10**

PART A

1. Explain Autocorrelation and Partial Autocorrelation in the case of Time Series?
Autocorrelation is obtained by computing the correlation between time series measurements taken at different points in time.
Partial autocorrelation refers to the association between observations in a time series and observations at preceding time steps (or lags).The partial autocorrelation statistic shows just the relationship between the two observations that are not explained by the shorter time lags between those observations.

2. Why does a Time Series require to be Stationary?
A stationary time series is one whose features are independent of the observation time, i.e., the mean, variance, autocorrelation, and other statistical features remain constant over time. This is useful because stationarity is a fundamental element in the area of time series analysis, and it has a significant impact on how data is viewed and forecasted. Time series models, which are used to forecast or predict the future, assume that each point is independent of the previous point.

What test do we use to confirm if the Time Series is stationary?
We use the Augmented Dickey–Fuller test (ADF).

What are the Null Hypothesis and Alternative Hypothesis considered in that test?
The Null Hypothesis for this test is that a unit root exists.
The Alternative Hypothesis for this test is that the time series is stationary.

3. What are the criteria we use to compare ARIMA models?
We can use the following comparison criteria:
1. Akaike Information Criterion (AIC).
2. Bayesian Information Criteria (BIC).
3. The Corrected Akaike Information Criteria (AICc).

The lower the test statistic, the the better the model fits the data.

4. Explain ARIMAx with all its components.
The Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) model is a more advanced variant of the ARIMA model. The model simply includes the covariate on the right-hand side of the equation, as follows:

$$\boldsymbol{y_t = \beta x_t + \phi_1 y_{t-1} + ... + \phi y_{t-p} - \theta_1 z_{t-1} - ... - \theta_q z_{t-q} + z_t}$$

Where $\boldsymbol{x_t}$ is a covariate at time $\boldsymbol{t}$ and $\boldsymbol{\beta}$ is its coefficient.

5. Discuss Simple Moving Average, Cumulative Moving Average, and Exponential Moving Average.
A simple moving average (SMA) is an arithmetic moving average calculated by adding the price of an instrument over a number of time periods and then dividing the sum by the number of time periods.

The Cumulative Moving Average (CMA) is defined as the unweighted mean of all prior values up to the present period (t).

An exponential moving average (EMA) is a kind of moving average (MA) that assigns a larger weight and relevance to the most recent data points, allowing the data to adapt faster to new information.

PART B

Section 1 : Plastic Sales Data

Table 1: Data summary

Name	plastics_tbl
Number of rows	60
Number of columns	2
Key	NULL
Column type frequency:	
Date	1
numeric	1
Group variables	None

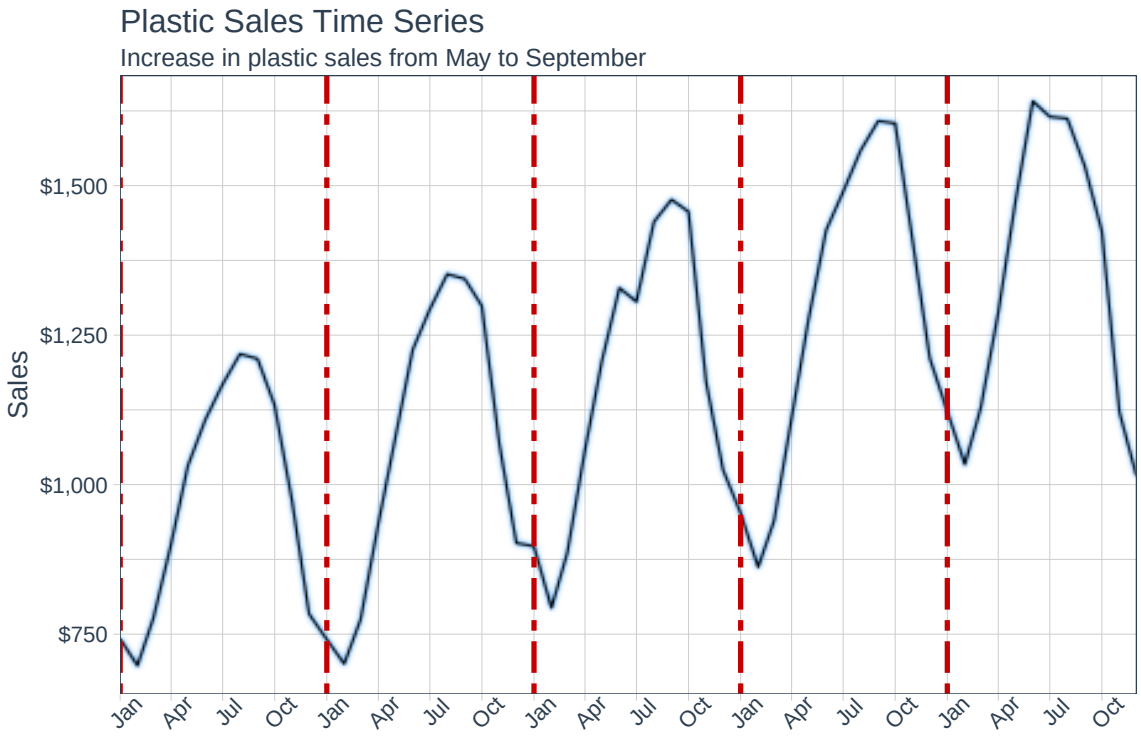
Variable type: Date

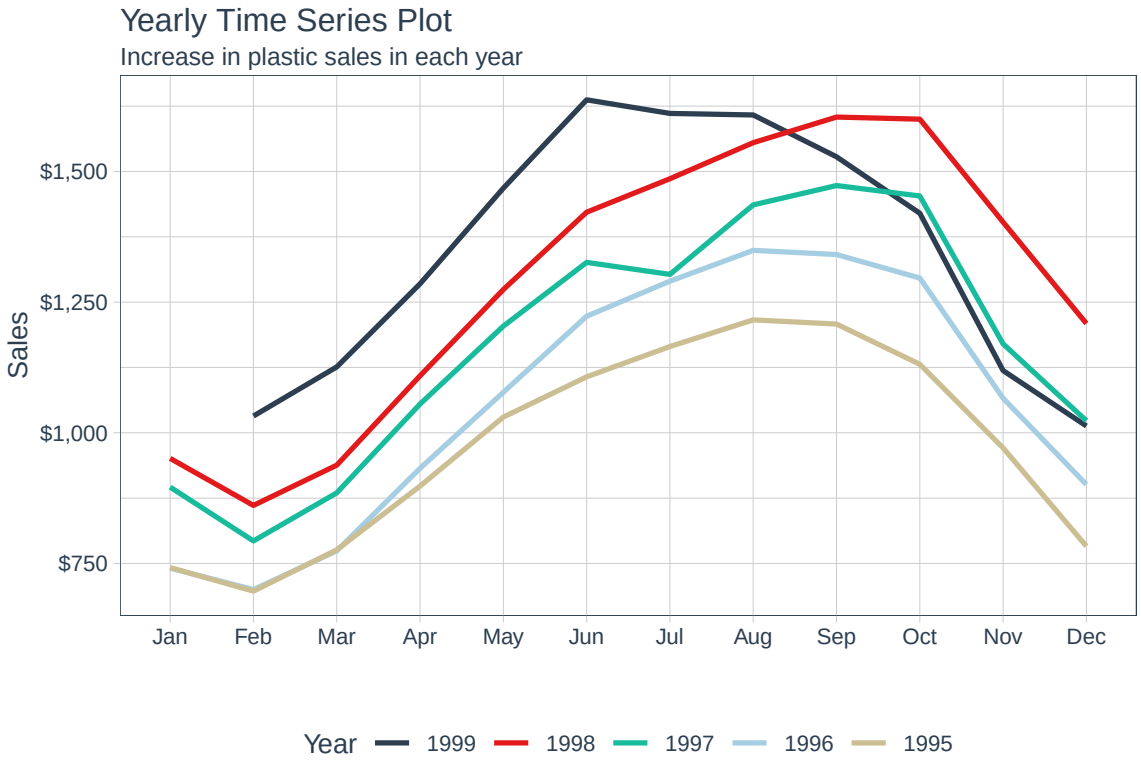
skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	1995-01-01	1999-12-01	1997-06-16	60

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
sale	2	0.97	1165.74	270.35	697	941.25	1167.5	1389.5	1637

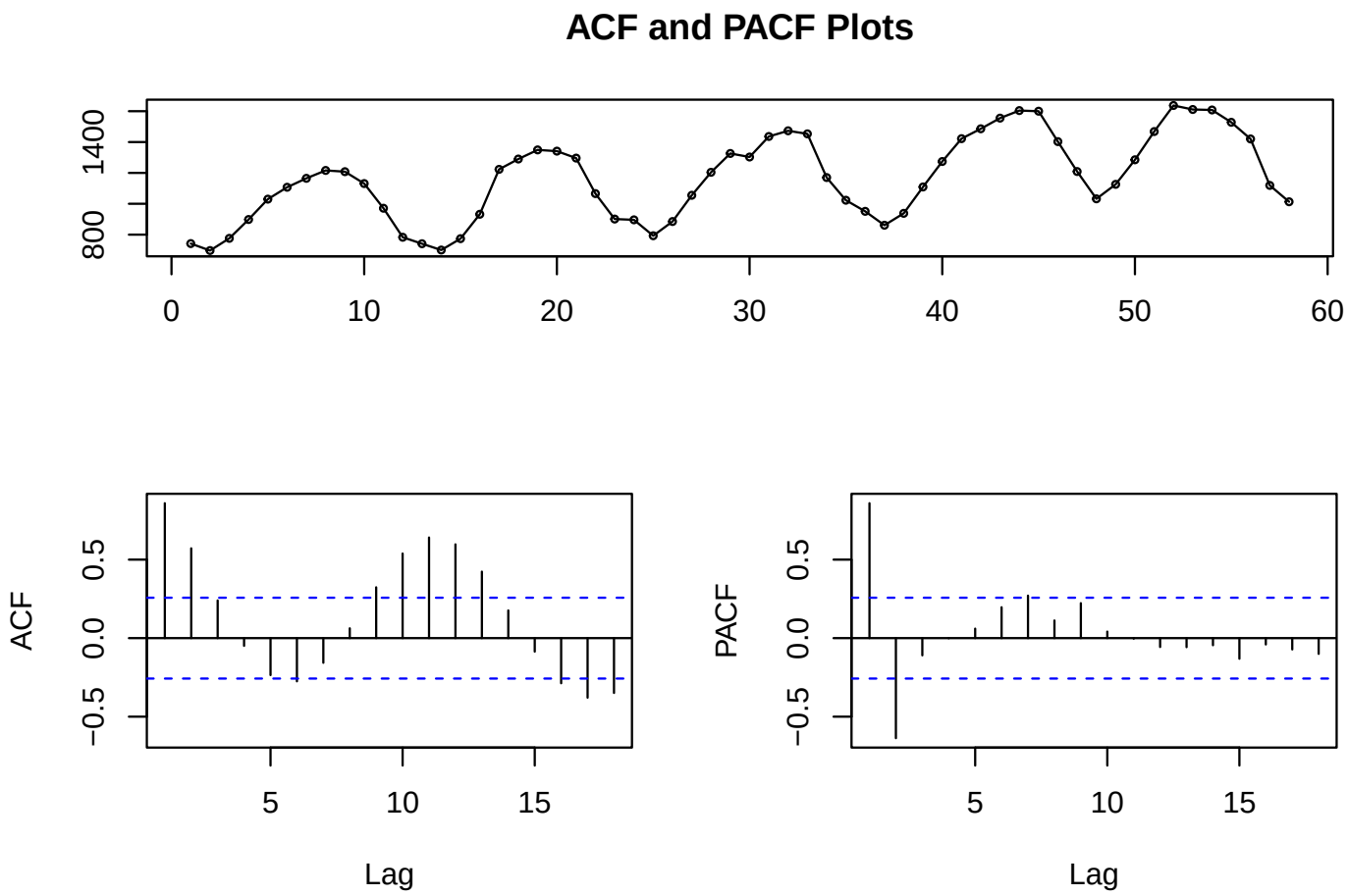
The plastic sales data is a time series data consisting of two columns, the date and sales variables. The time period starts from January 1995 to December 1999.





We can evidently see an increase in plastic sales each year. It is also evident that most of the sales take place starting from the month of May to around October. It may be anticipated that the majority of these sales take place during the summer months and that the number of sales begins to fall from October until roughly March of the following year, which is during the winter season of the year.

The autocorrelation and partial autocorrelation plots.



We can notice a spike in the twelfth lag on the autocorrelation plot, which corresponds to a positive correlation. This clearly demonstrates that the majority of the correlations occur on an annual basis, from the same month of one year to the next. When it comes to the sixth lag, there is a significant negative correlation; this is true because there are either less sales or more sales during the sixth lag, depending on the month's seasonality i.e. the month of June has higher sales of plastic as compared to the month of December, which is the sixth lag, sales at this time period are at their lowest.

Section 2 : Electricity Data

Table 4: Data summary

Name	electric_tbl
Number of rows	486
Number of columns	2
_____	
Column type frequency:	
Date	1
numeric	1
_____	
Group variables	None

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	1973-01-01	2013-06-01	1993-03-16	486

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
value	0	1	259.56	68.87	139.59	195.83	261.24	311.98	421.8

The electricity data is a time-series data containing two columns, date and value. The dataset starts from January 1973 to June 2013.

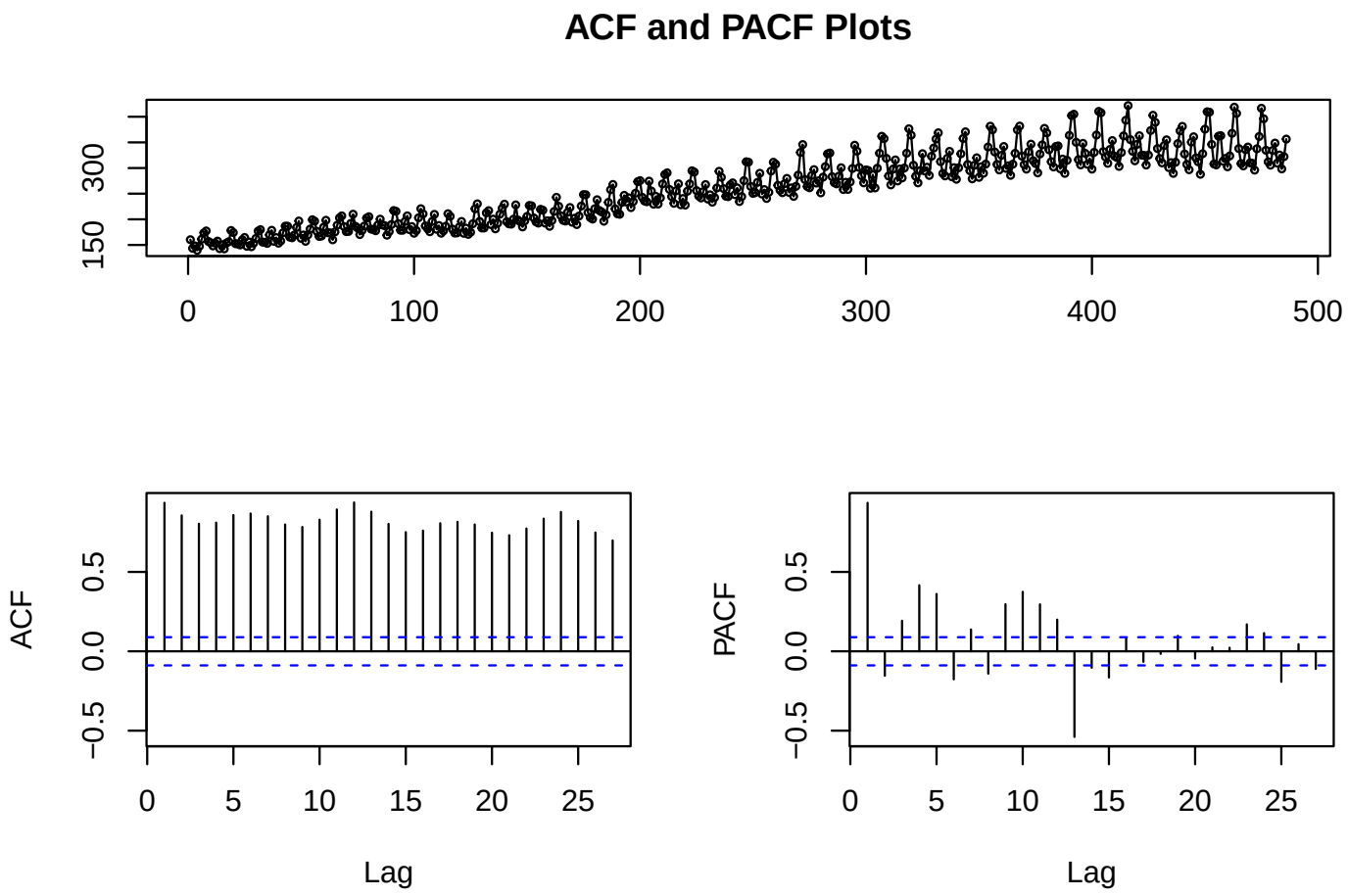
2a. Stationarity

We use the Augmented Dickey–Fuller test (ADF) to test for stationarity. The test hypothesis for this test is:
H_o : The unit root exists
H_1 The time series is stationary

```
    Augmented Dickey-Fuller Test

data:  electric_ts
Dickey-Fuller = -8.8683, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Since the p-value (0.01) is less than the significance level of significance 0.05, we reject the Null Hypothesis and conclude that the time series is stationary. Therefore, no differencing is required.



Based on the autocorrelation plot, we can evidently see that there are no correlations between time lags hence supporting the Augmented Dickey Fuller test that the data is stationary.

2b. ARIMA models

We shall now be using different ARIMA models that would be useful in describing the time series. We shall try to tweak the values of p (the number of autoregressive terms), d (differencing), and q (the Moving Average lags).

Table 7: ARIMA Models Test Statistics

models	AIC	BIC	sigma	logLik	nobs
Arima (5, 1, 2)	4150.884	4184.357	17.13395	-2067.442	485
Arima (4, 1, 1)	4190.332	4215.437	17.93198	-2089.166	485
Arima (2, 1, 2)	4233.838	4254.759	18.79015	-2111.919	485
Arima (3, 2, 1)	4302.871	4323.781	20.23839	-2146.435	484
Arima (2, 0, 3)	4428.720	4458.024	22.65045	-2207.360	486

Arima (5,1,2) model had the least AIC and BIC values of 4150.884 and 4184.357, respectively, thus making it the most suitable model for forecasting our dataset. The table above also shows the other models that were tested with their significant results.

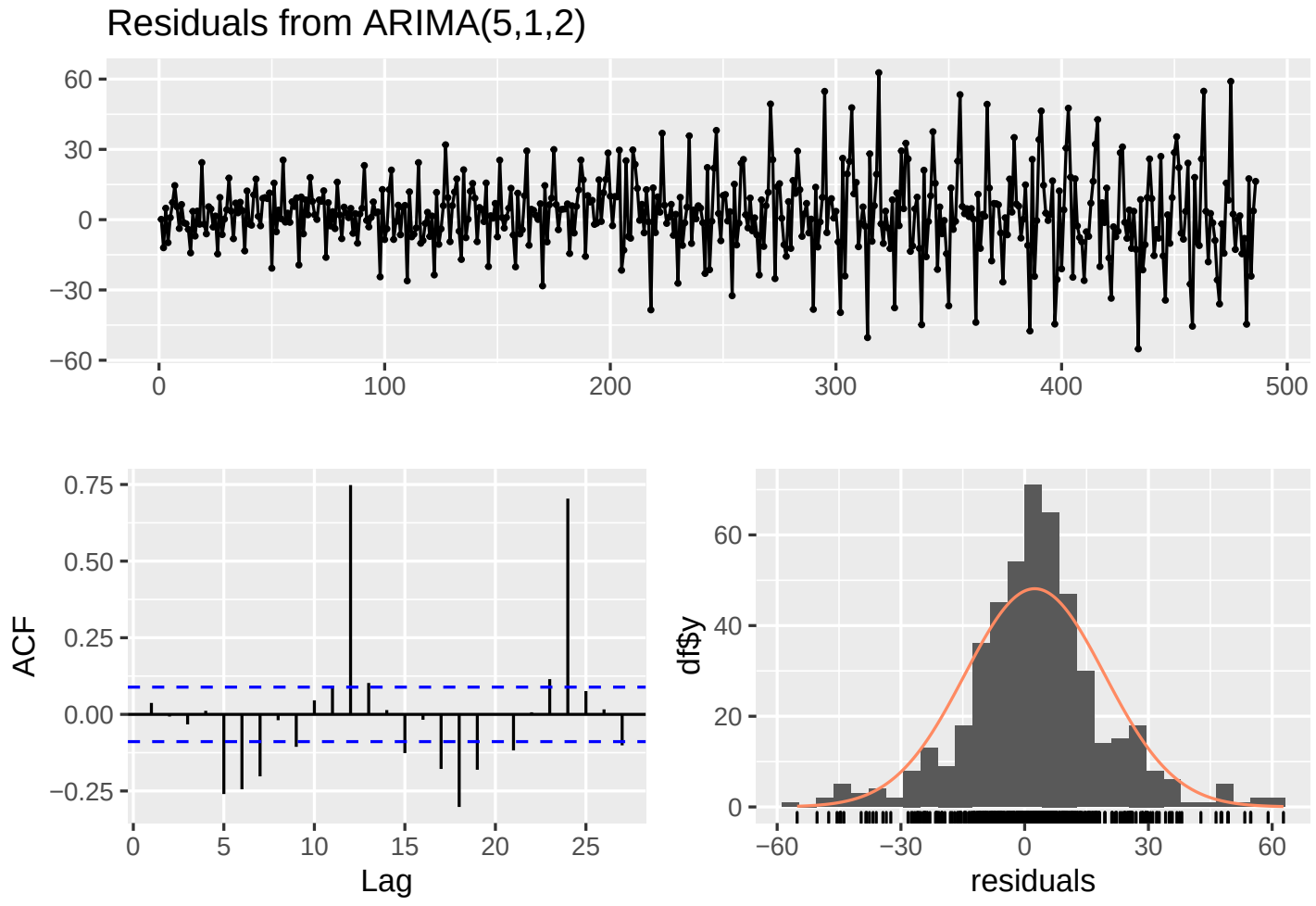
2c. Model residuals

We now estimate the parameters of our best model and do diagnostic testing on the residuals.

Table 8: Arima (5,1,2) Parameters

term	estimate	std.error
ar1	0.1055892	0.0604971
ar2	0.2550930	0.0485001
ar3	-0.4023029	0.0402649
ar4	-0.2659966	0.0435949
ar5	0.3685442	0.0466825
ma1	-0.2748559	0.0482917
ma2	-0.5819661	0.0409923

The table above shows Arima (5,1,2) parameters that will then be used to forecast our dataset.




```
    Ljung-Box test

data:  Residuals from ARIMA(5,1,2)
```

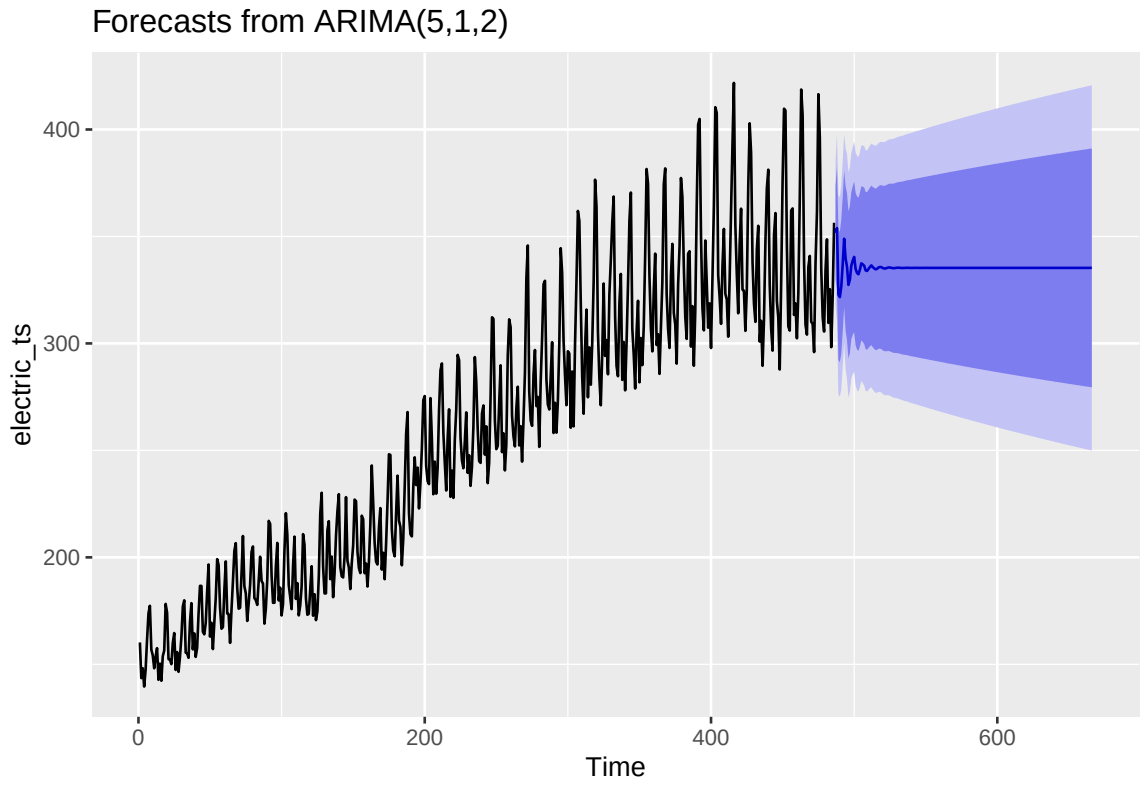
```
Q* = 91.19, df = 3, p-value < 2.2e-16

Model df: 7.   Total lags used: 10
```

We use the Ljung-Box test to test for our model fit. The test hypothesis for this test is:
H_o : The residuals are independently distributed.
H_1 : The residuals are not independently distributed; they exhibit serial correlation.

Since the p-value (2.2e-16) is less than the level of significance of 0.05, we, therefore, reject the Null Hypothesis and conclude that the residuals are not independently distributed; hence they exhibit serial correlation.

2d. Forecasting



2e. Intervals

A common feature of prediction intervals is that they increase in length as the forecast horizon increases. The further ahead we forecast, the more uncertainty is associated with the forecast, and thus the wider the prediction intervals.

REFERENCES

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. Journal of statistical software, 27, 1-22.

Cryer, J. D., & Chan, K. S. (2008). Time series analysis: with applications in R (Vol. 2). New York: Springer.