Sales Data Forecasting

(Auto Regression vs Linear Regression)

Executive Summary: Sales data is being analysed with an Auto regressive model and Linear model to compare the performance of predicting the future. We will explore how auto regressive models do not always guarantee superior performance. So, having a core understanding of how these models work can give us more insights in choosing the appropriate models for our datasets.

Data set: The csv file contains Sales values against dates from Jan 2014 to December 2023

## Process flow in Rstudio:

I have applied AR 1 model to the original data and have modeled using a linear model including seasonality to have a comparison of performance.

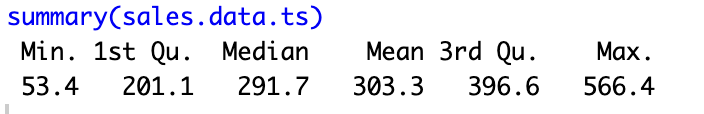
# Applying AR (1) model to Original Data:

A white background with green text

Description automatically generated

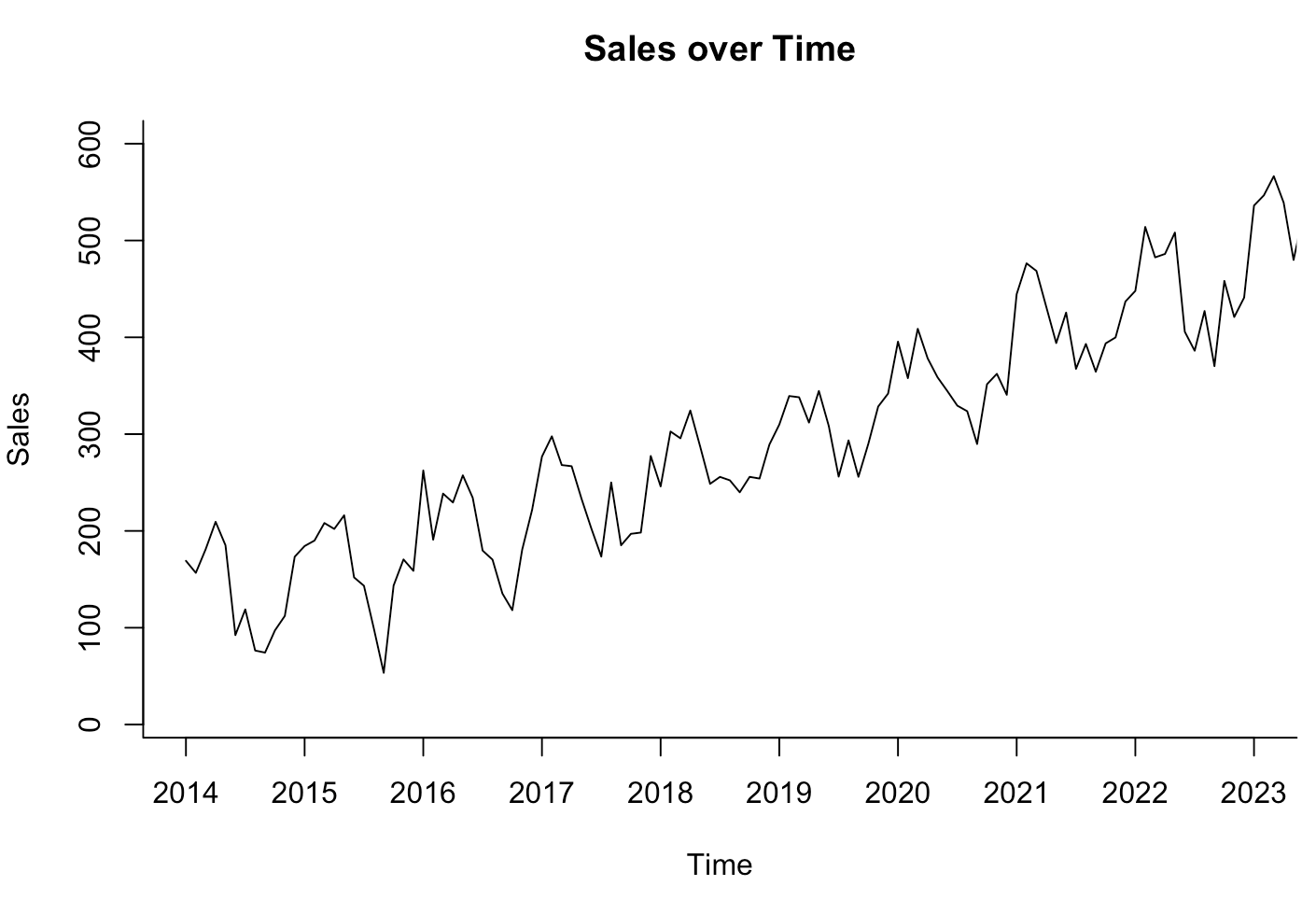
A computer screen shot of a computer code

Description automatically generated



A screenshot of a computer code

Description automatically generated



**Comments:** The data shows seasonality and an Increasing trend. The range of the sales figures through Summary is understood as (53.4,566.4), which helps us in setting our plot boundaries.

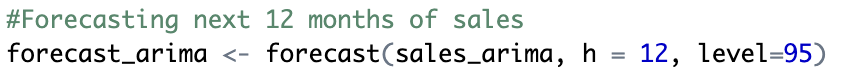
A white background with green text

Description automatically generated

A screenshot of a computer

Description automatically generated

AR1 coefficient of 0.9516 which is very high indicates the high Correlation between the predicted values and last recorded data. It also informs that the model might not be capturing the seasonality present in the data. AIC value (1240.63) is high, which tells us that the model is not the right fit for the data. The AIC value is relative, so we can only judge model performance in comparison to others. The mean (314) might be the long-term average or base level of the time series.

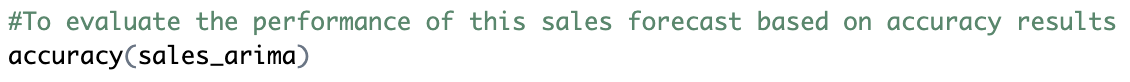


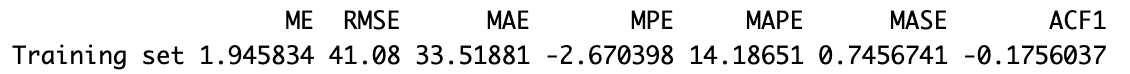
A screenshot of a computer code

Description automatically generated

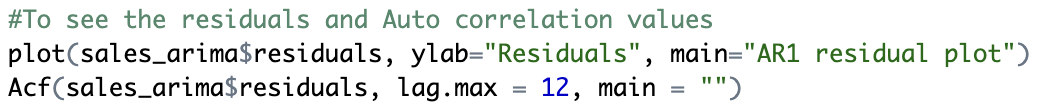
A graph with lines and numbers

Description automatically generated





Analysis: RMSE is so high implying the presence of large forecast errors. A negative MPE shows that forecasts tend to be 2.7% lower than actual values. MASE is lower than 1, meaning the model is better than naïve forecast, and also the ACF1 which is closer to 0, usually indicates a good model fit, but we cannot be sure unless we plot the residuals. Also we can see from the plot that the Confidence Interval itself is so high as there is lot of variance in correlation.



A graph of a plot

Description automatically generated with medium confidence

We can see significant forecast errors as well as a seasonality in the errors. So, the model seems to not be capturing the seasonality of the data.

A graph with lines and numbers

Description automatically generated

The Autocorrelation plot seems to be showing multiple spikes, meaning the predicted values are not exactly correlated to the previous period.

Assumptions of an AR(1) model:

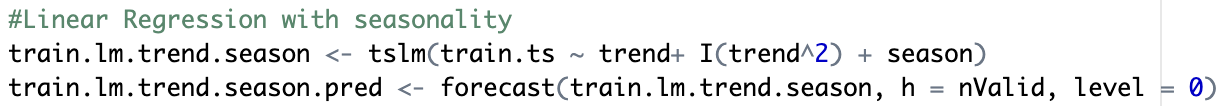
1. Linear Correlation between current value to lagged values. Which is not the case with the current dataset, as the dependency is random.
2. Normal distribution of errors, which doesn’t seem to be the case with our data.

Limitations:

1. The model doesn’t explicitly handle seasonality unless we use higher order and much complex model.
2. Constant parameters such as auto regressive coefficient, which is not static as values change independent of time in reality.

To conclude, this forecast is not complete and needs refinement.

# Applying a Linear Model with Seasonality and AR(1) to Residuals:

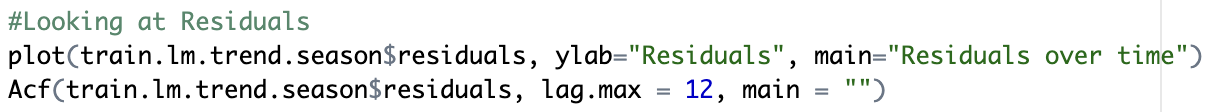


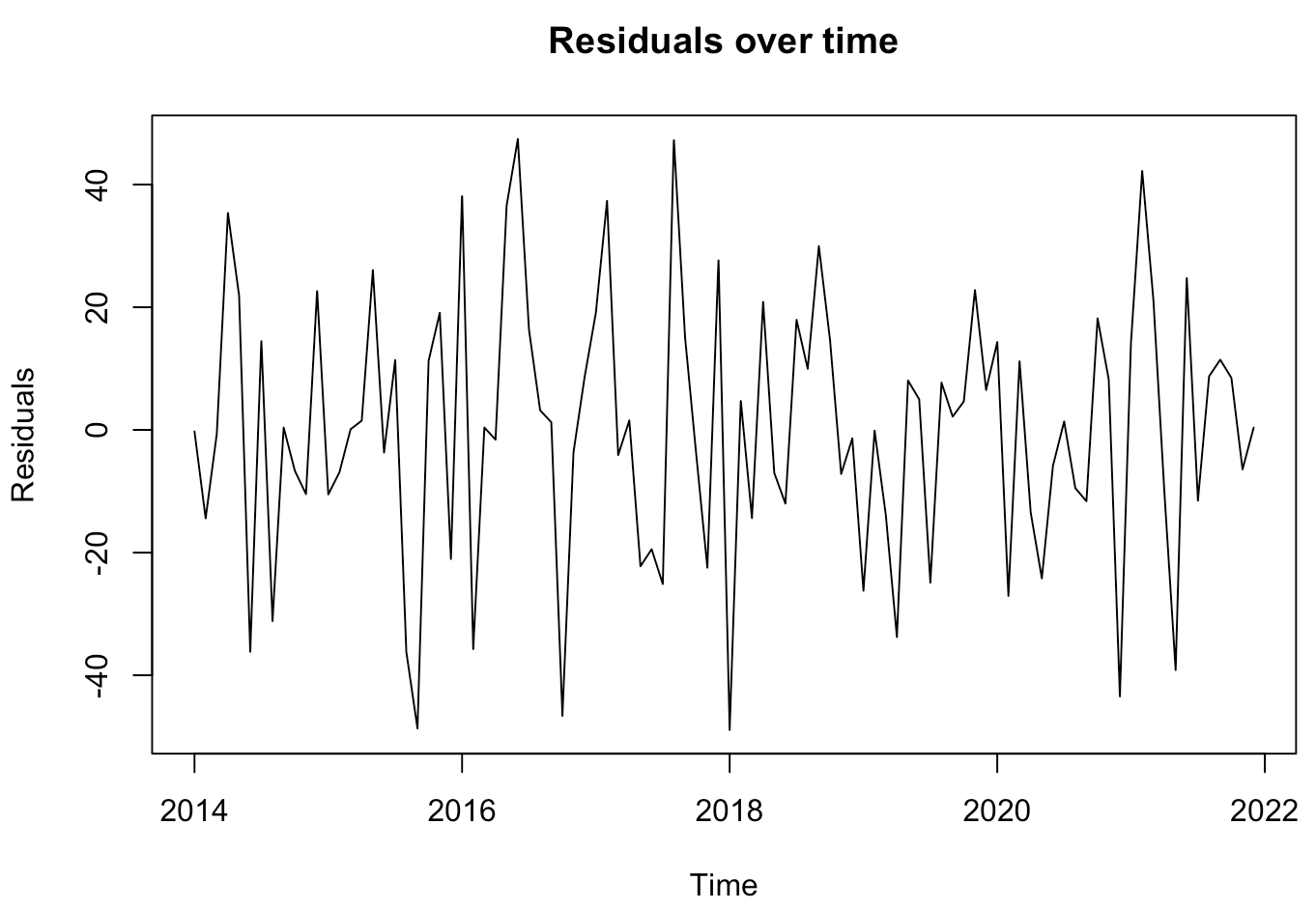
A close-up of a code

Description automatically generated

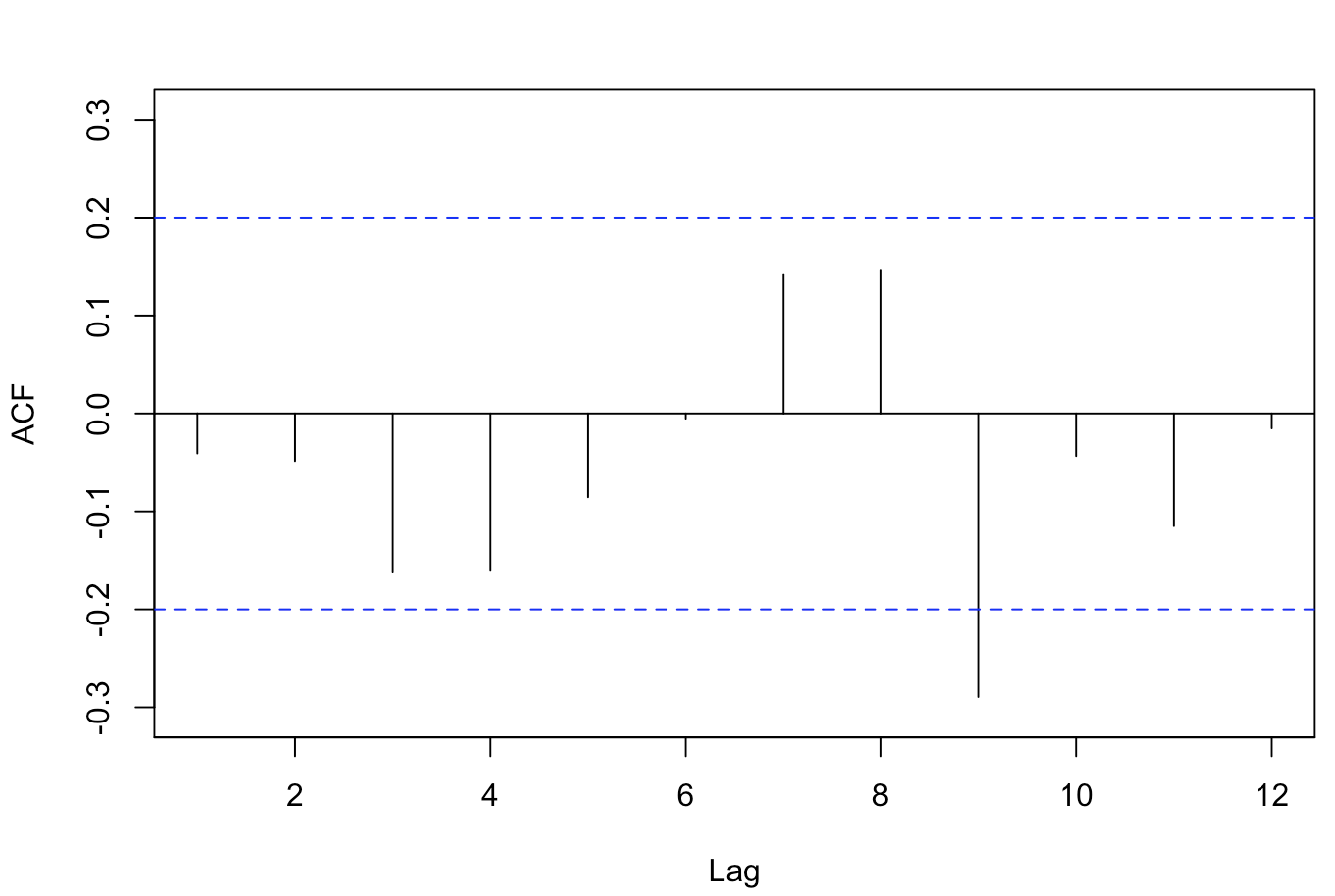
A graph of sales over time

Description automatically generated

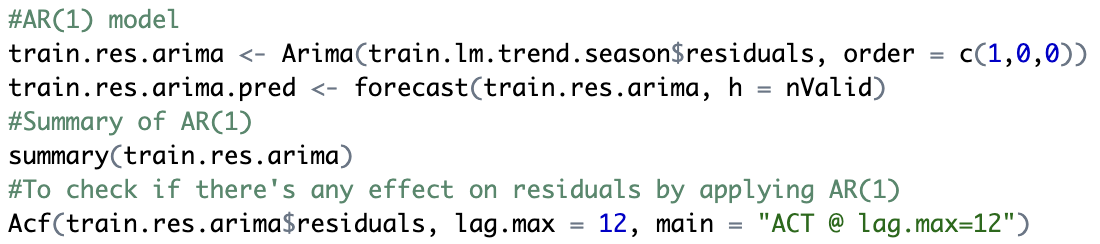


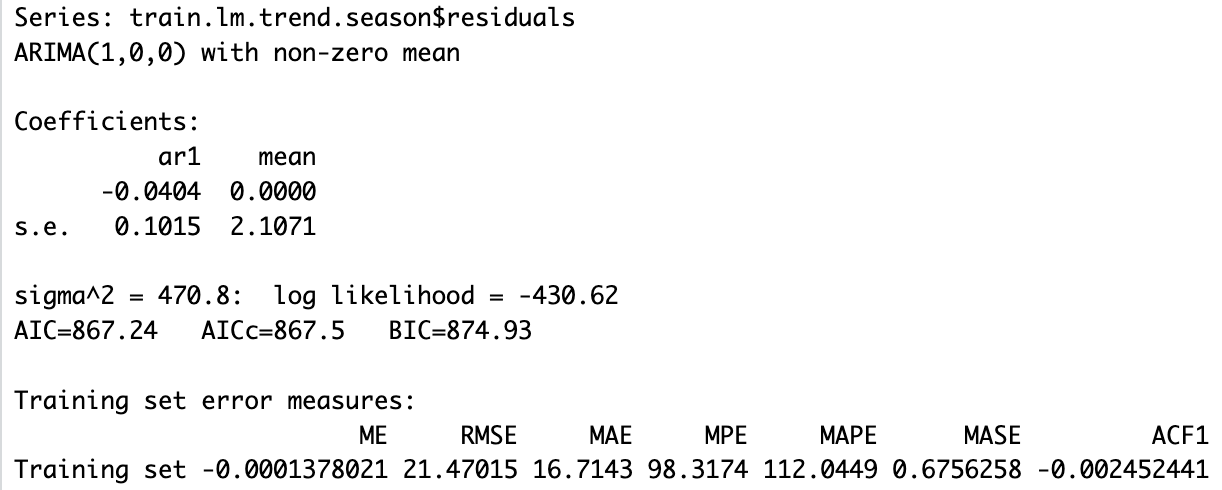


We can see how residuals are much lower with this model (swing between-40 to 40), Arima(1,0,0) gave us residuals in -100 to 100 range. Also we can see a difference in Autocorrelation in the plot below,

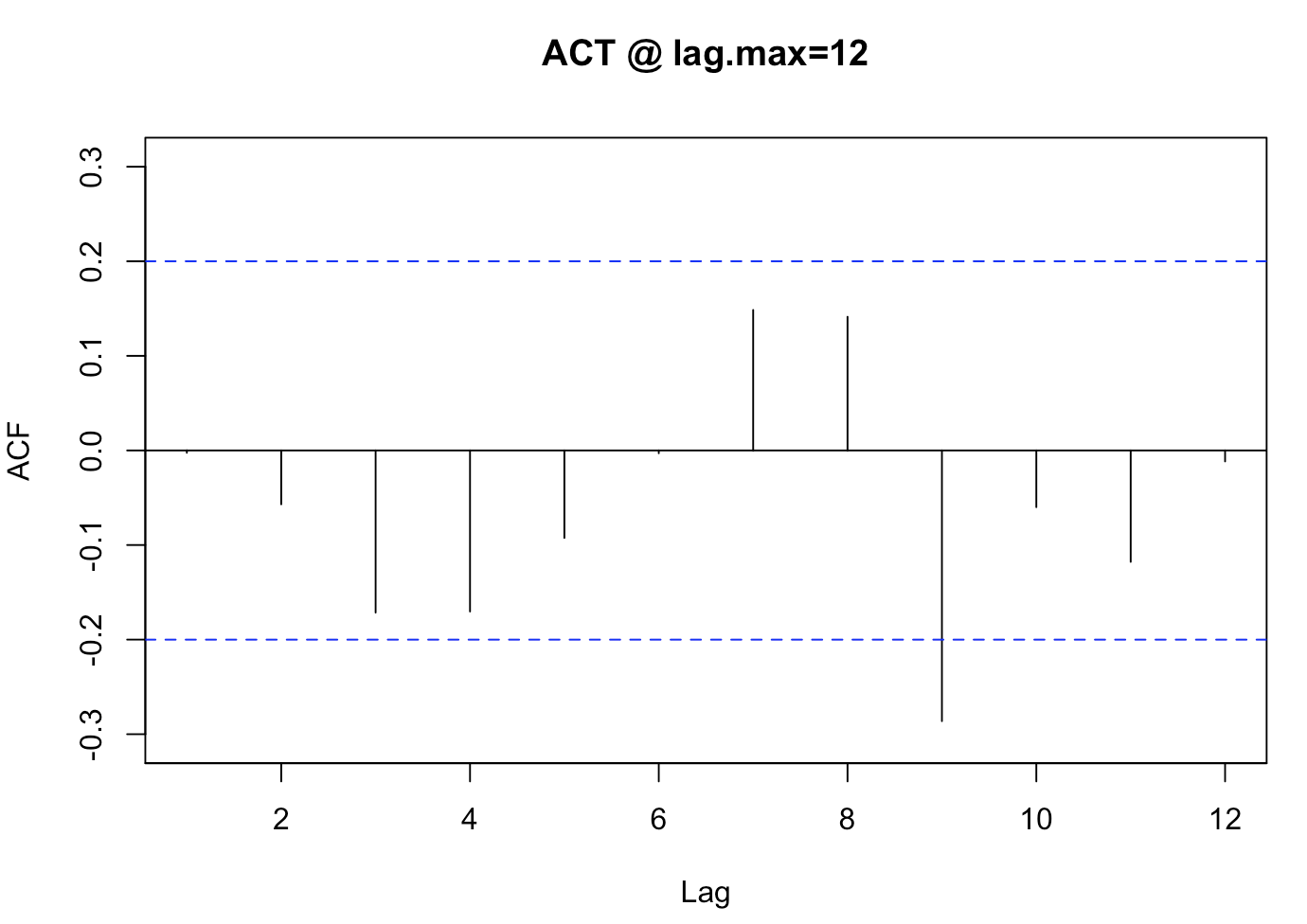


Only one significant autocorrelation at lag 9, which makes it much simpler to understand and make changes.





The Autocorrelation coefficient is almost 0, which indicates its very less dependence on last value and its also evident from the Acf plot showing Lag 9 as important rather than lag 1 which is considered by AR(1), second is the lower error values across board in comparison to Arima on original data. Also from following Acf we can conclude that AR(1) would have reduced the residual errors if there was correlation at lag 1 period, which is not true in our case and its shown by the ACF plot after AR(1), which is almost unchanged as seen below,



Conclusion: Based on my understanding, the data has seasonality and trend which are not captured by applying an AR (1) model on the original data, and much simpler models such as linear model with seasonality are able to do a better job at forecasting for our dataset.

Green Box: R code, Blue Box: Output