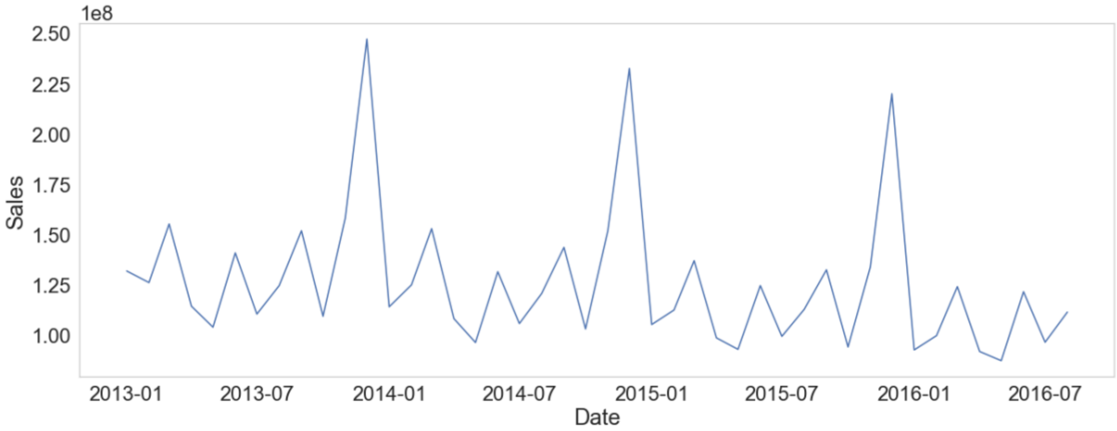
Data Challenge Summary Page

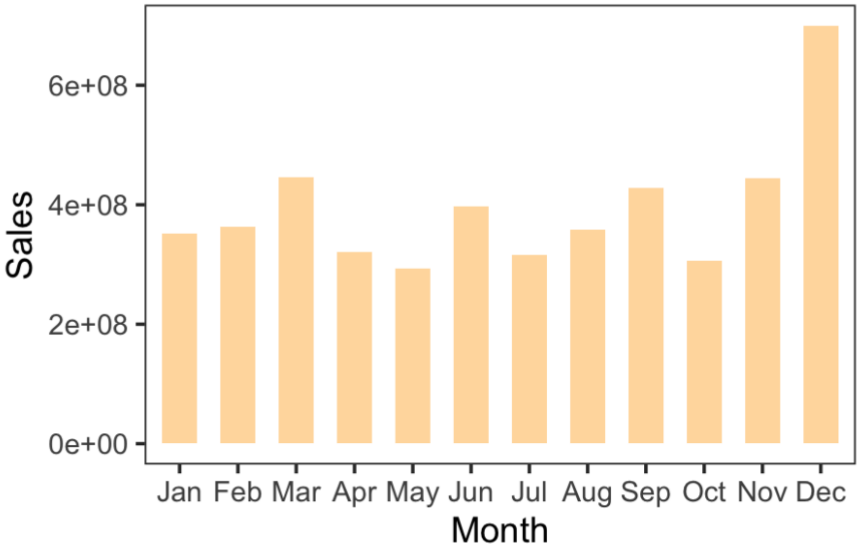
Karen Chen   June 15, 2020

Data Exploration

The sales data starts in January 2013 ends at August 2016.

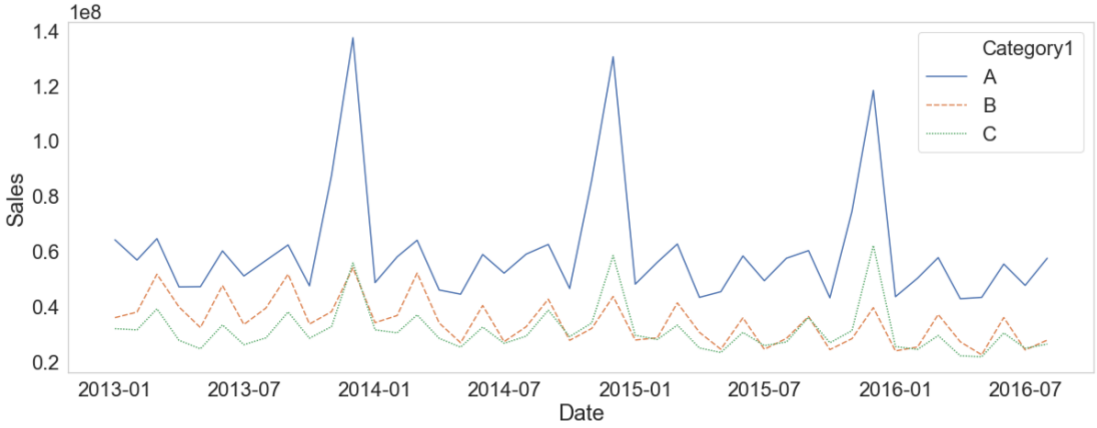


Total Sales



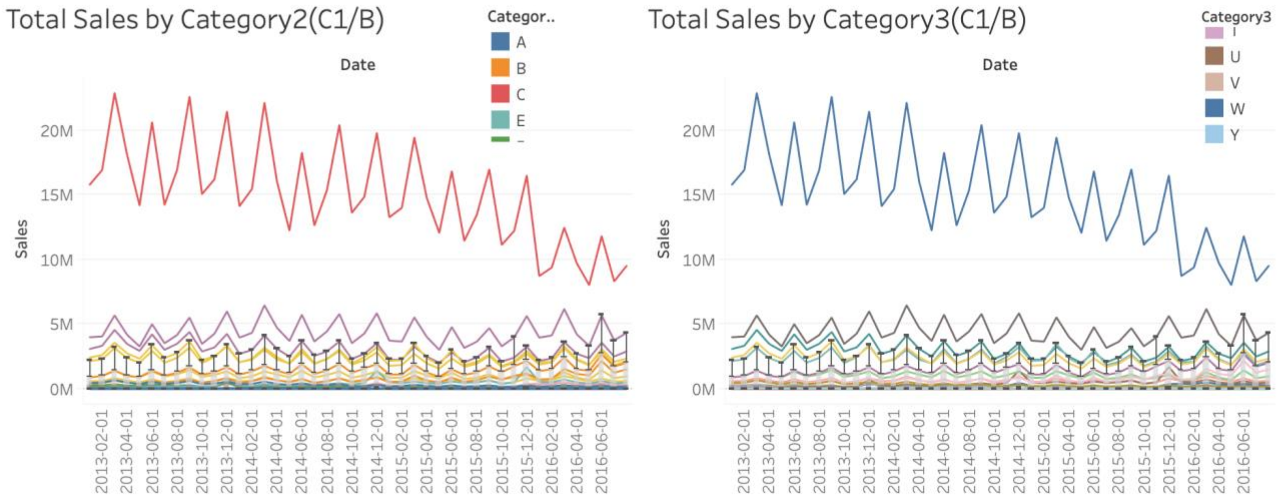
Total Sales by Months in 2013, 2014, 2015

Understanding how the sales change over time give us an idea of how the company is performing. It shows strong seasonality with an interval of 12. Sales hit all time high at each December over three years. The total sales have a slight downward trend with a rate of -0.053 in 2014 and -0.077 in 2015. We should find out why sales are decreasing over time. Let's first take a closer look at each type in Category1.



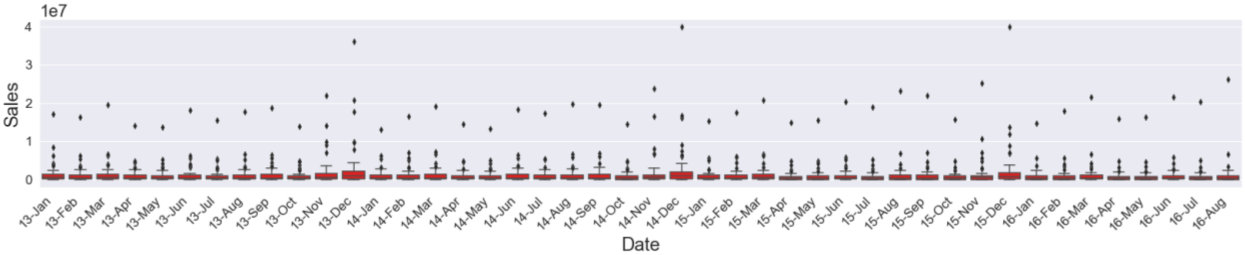
Total Sales by Category1

Analyzing sales by Category1 or the general category provides insights into how well vs not well each type of product is sold and what types of products customers are into. Obviously, we can tell that Category1/ A drives the sales spike in each December. But like the total sales, sales of each type products are slightly decreasing.



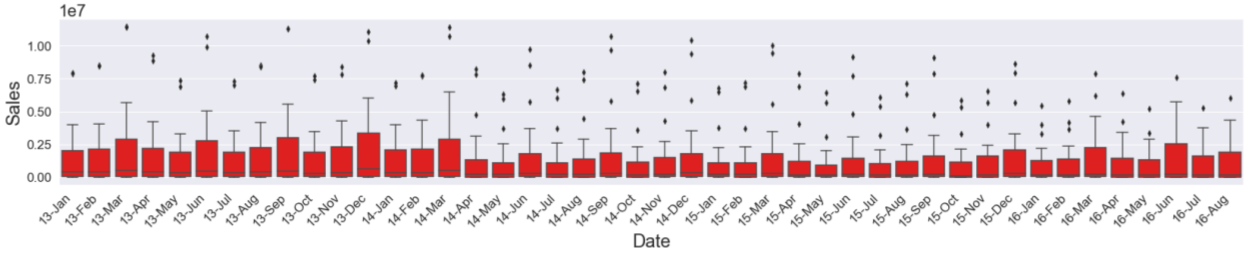
Total Sales under Category1/B

The sales for each type in Catgory1 show downward trend. In order to find out the reason behind this, let’s dive deeper into Category2 and Catgory3 for each type in Category1. Obviously we can tell that sales of products belonging to Category1/B and Category2/C drop significantly from 15m to 9m over three years, same for products belonging to Category1/B and Category3/W. Category2/C products and Category3/W products act as the best sellers under Category1/B, but their sales drop significantly over time, leading to decrease in total sales. We finally find the cause! If the company want to turn things around, they should look into these products and find out why people buy these products less as time goes by.

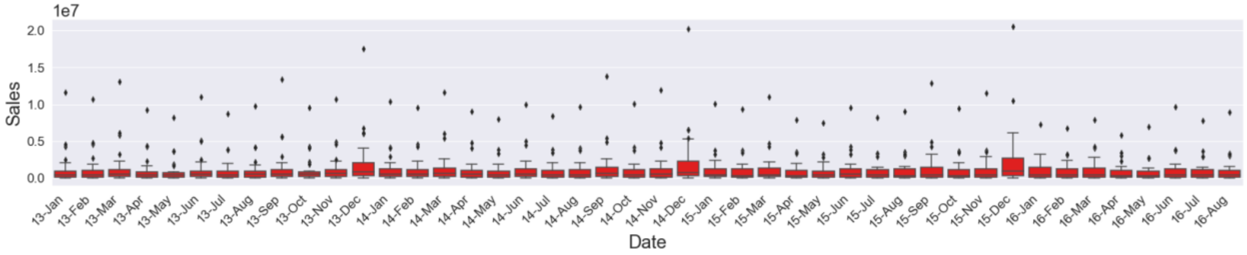


Category1/A Boxplot

We found out the reasons why sales go down. Now we want to analyze some outliers, meaning the sales of the products are either too high or too low. Having a deep understanding into these outliers will better allow us to narrow down the range for analysis and customize our efforts for specific product development or marketing strategy. For Category1/A, one type of product stands out of the other products. That is, the product belonging to Category1/A, Category2/A and Category3/M generates the highest sales almost every month among other products and reached all-time high in December. Its sales continue to grow over the three years even though the total sales are decreasing overall. This is a type of lucrative product in the company so we are interested to know how its sales will go in the future and will do forecasting on this type of product.



Sales under Category1/B



Sales under Category1/C

Under both Category1/B and Category1/C, products that belong to Category2/C and Category3/W generate the highest sales revenue among other products. Both are regarded as outliers though, sales in Category1/C are increasing while sales in Category1/B are decreasing. In other words, for the products with the same Category2 and Category3, customers tend to buy Category1/C more and more and buy Category1/C less and less over time. This is very interesting. The company should consider them both together, looking for commonalities and differences in order to develop different marketing or sales operations strategy.

Suggestions:

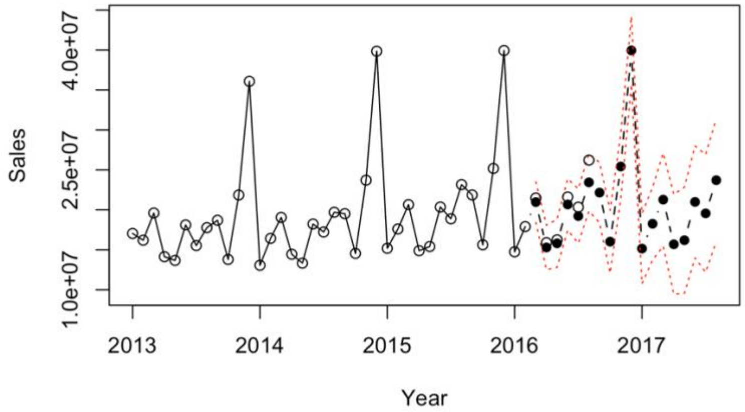
In order to prevent total sales dropping further, the company should look into products in Category2/C under Category1/B and Category3/W under Category1/B. They are the main drivers that drag down total sales. The company should be well prepared for each December when the sales hit all time high throughout that year. More efforts should be made to the outliers including products in Category1/A, Category2/A and Category3/M, in Category1/B, Category1/C and Category1/W and in Category1/C, Category1/C and Category1/W. They can have huge impact on how total sales go.

Forecasting

Goal: Forecast next 12-month sales for products that belong to Category1/A, Category2/A and Category3/M

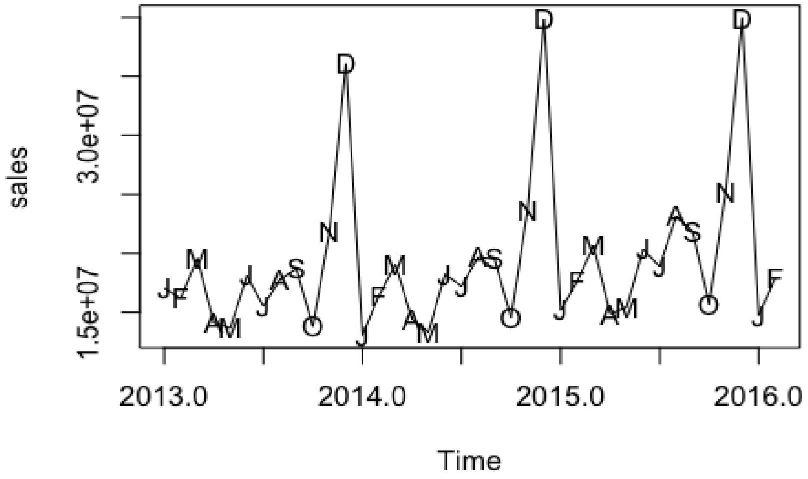
Model: ARIMA, SARIMA

Result:



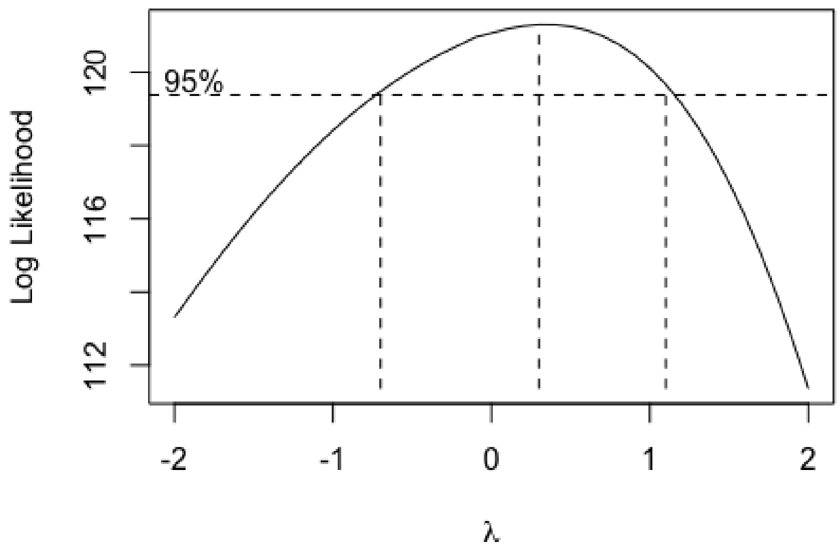
Steps:

1, Plot the data

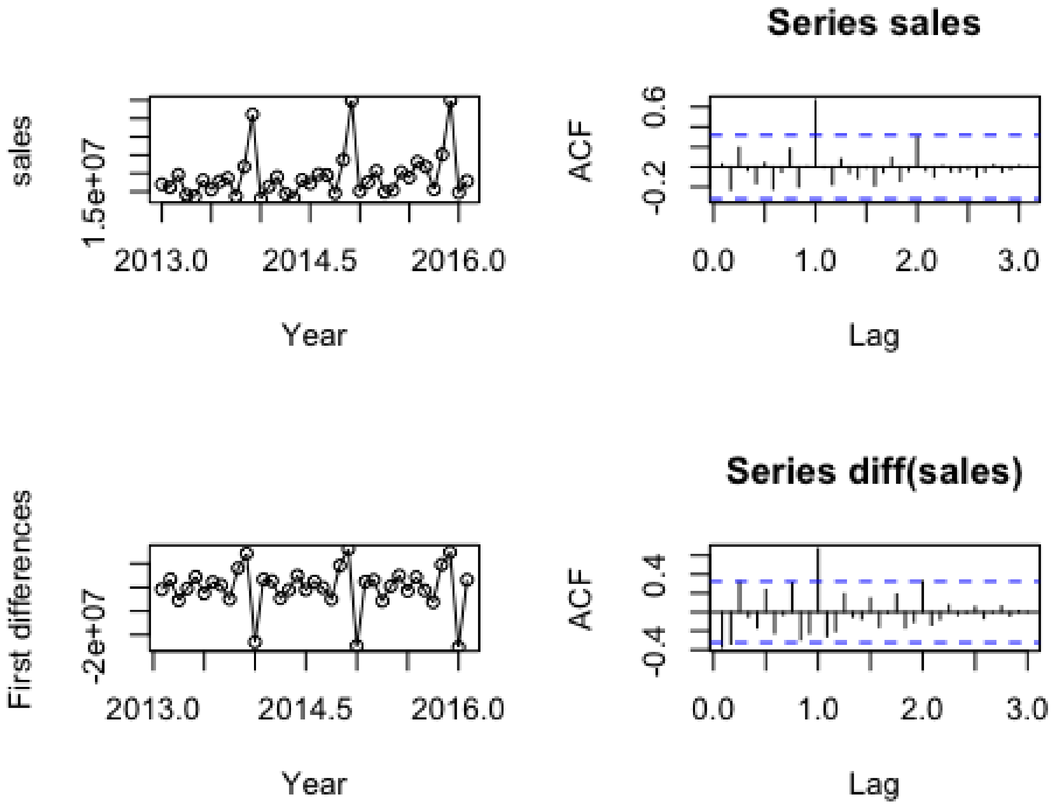


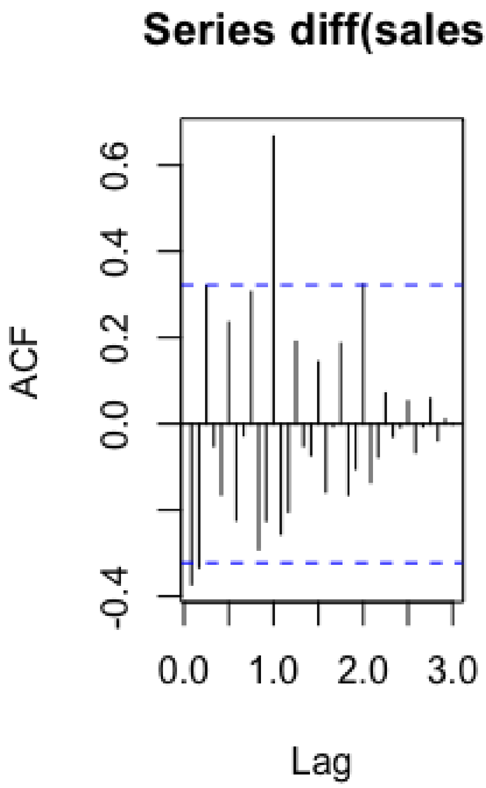
The time series plot suggests that this process is not stationary. It has standard variance and there are no obvious outliers.

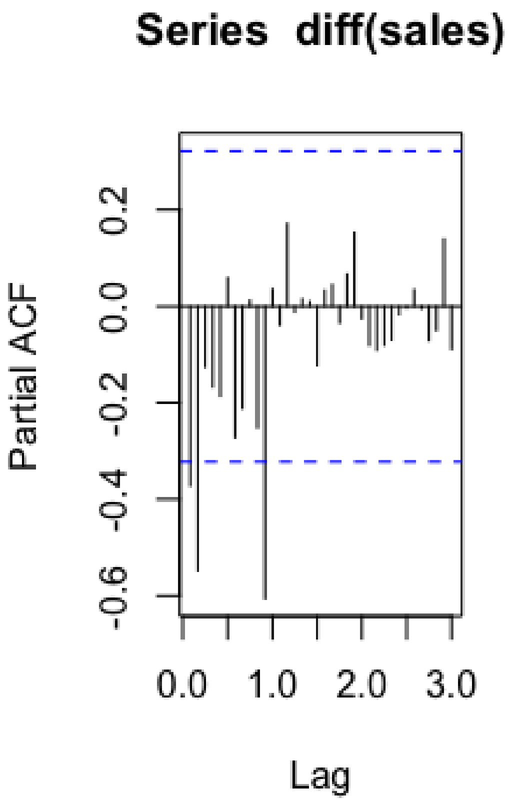
2, Model Specification



It turns out that there is no need for transformation since the confidence interval for lambda contains 1.







Here are some models I can start trying based on ACF plot and PACF plot.

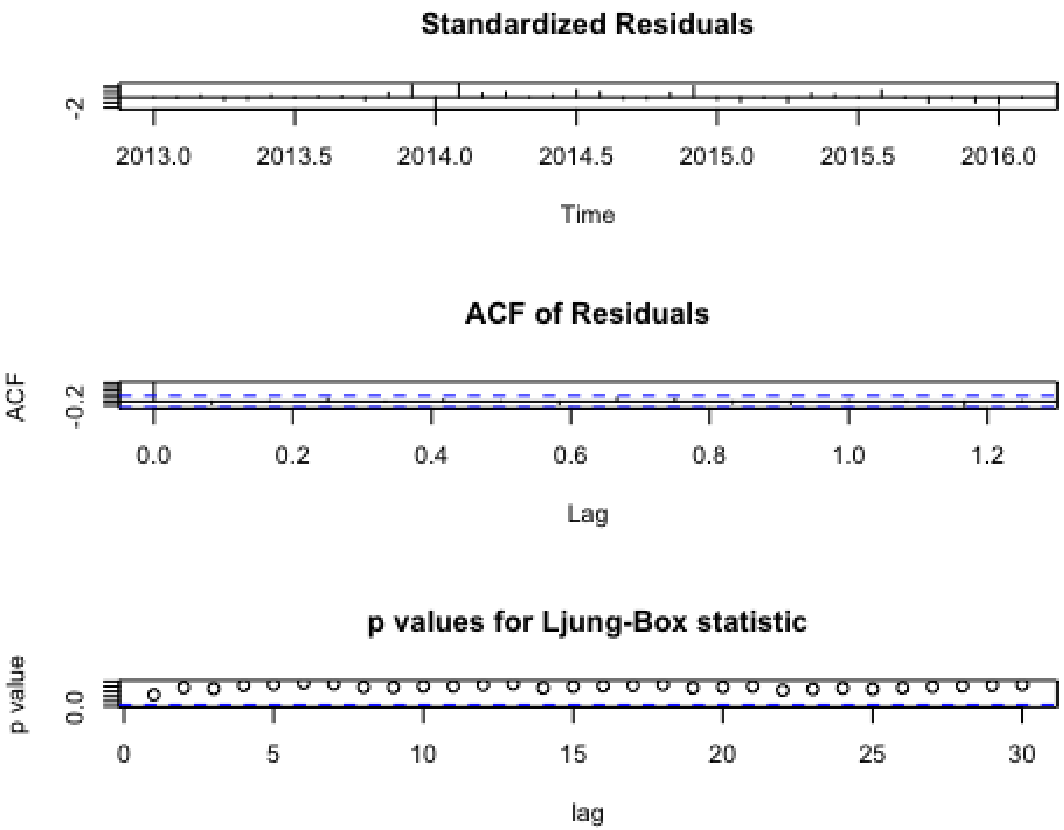
SARMA model:(0,0,2), (1,0,0), (1,0,1), (1,0,0), (0,0,1).

ARIMA model:(2,1,0), (0,1,3), (2,1,3).

By increasing and decreasing components based on performances of the models, here are four good models I ended up with temporarily.

```
Fit arima(0,1,0) * sarma(1,0,0)_{12} → Correlated residues
Fit arima(2,1,0) * sarma(0,0,2)_{12} → Correlated residues
Fit arima(0,1,1) * sarma(1,0,0)_{12} → selected model
Fit arima(2,1,3) * sarma(1,0,0)_{12} → Reject normality
```

3, Model Diagnostics

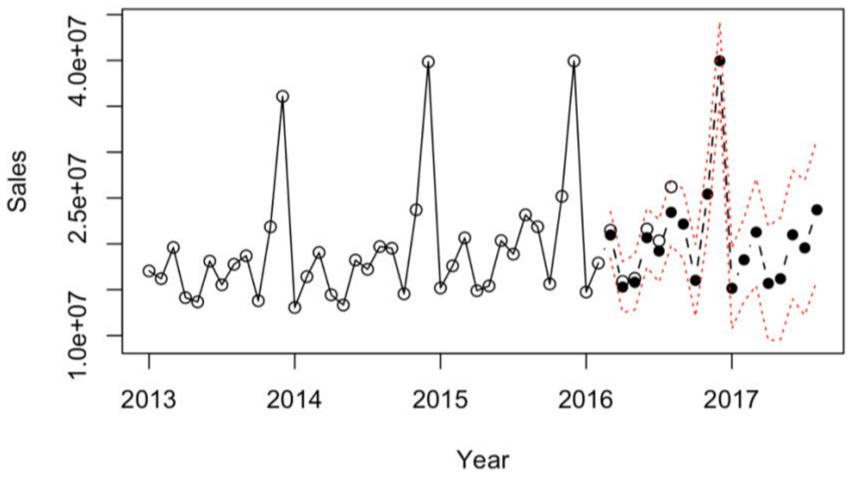

Standardized Residuals

ACF of Residuals

p values for Ljung-Box statistic

*The histogram and qq plot of the standardized residuals in generally supports the normality assumption. In addition, when further examining the standardized residuals from the model fit, the Shapiro-Wilk test does not reject normality (p-value = 0.3143) and the runs test does not reject independence (p-value = 0.625). Ljung-Box p-values do not suggest lack of fit. Therefore, I decided to go with arima(0,1,1) * sarma(1,0,0) model.*

4, Forecasting



Test	Pred
21483559	20968087
15900085	15292669
16247858	15809611
21614813	20676744
20341383	19224662
26216774	23439429
MAPE: 1333379.8	