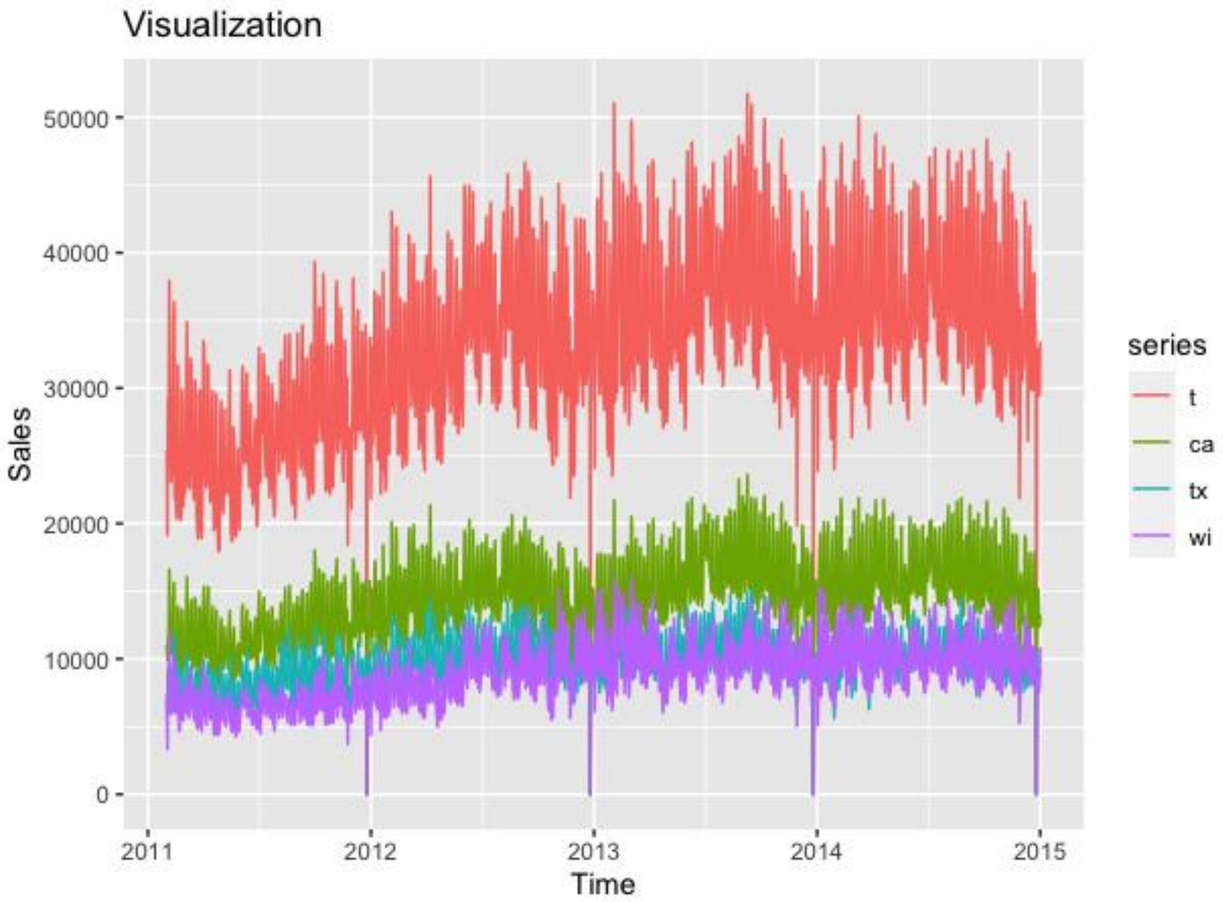
Analysis and Prediction of Walmart Sales using R

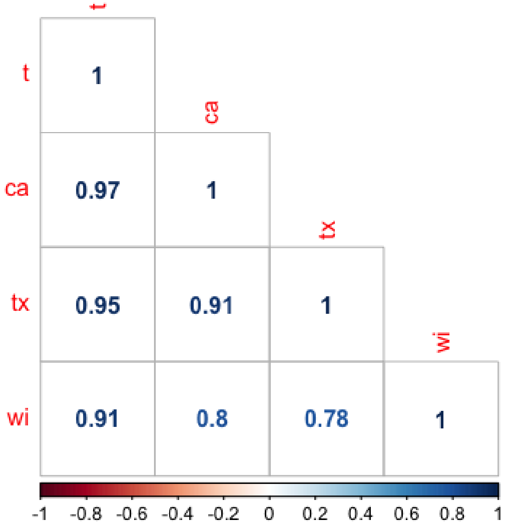
The goal of this project is to analyze and predict the sales of Walmart in the three states California, Texas and Wisconsin from from “2011-02-01” to “2014-12-31”.

Firstly, I convert my daily data of Walmart’s Sales using function ts() with frequency = 365 and as start date = c(2011, 32). Then, I create a plot the total sales of three states and total sales of all, such as CA, TX, WI , Total. According to the chart below, we can try to identify some things about the characteristics:

- **Trend:** Increasing trend.
- **Seasonality:** Sure exists.
- **Outliers:** We can clearly identify some outliers during the last month of every year.
- **Cyclical:** No cyclic pattern.
- **Irregular-Random:** There are irregular variations.



Next, I evaluated the linear correlations of time series. Values between 0.7 and 1. indicate a strong positive linear relationship via a firm linear rule.

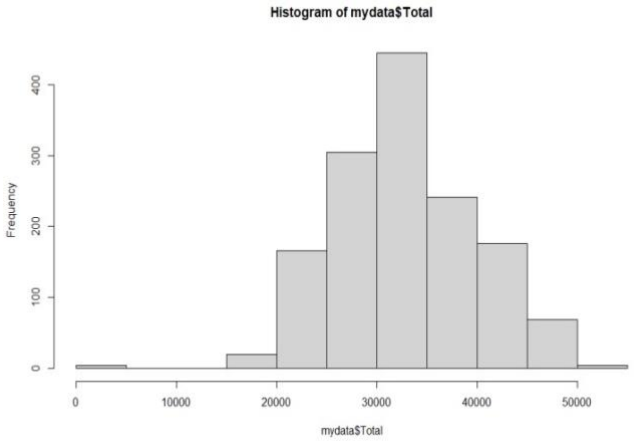


The percentages of sales of each state on the total sales of the company are:

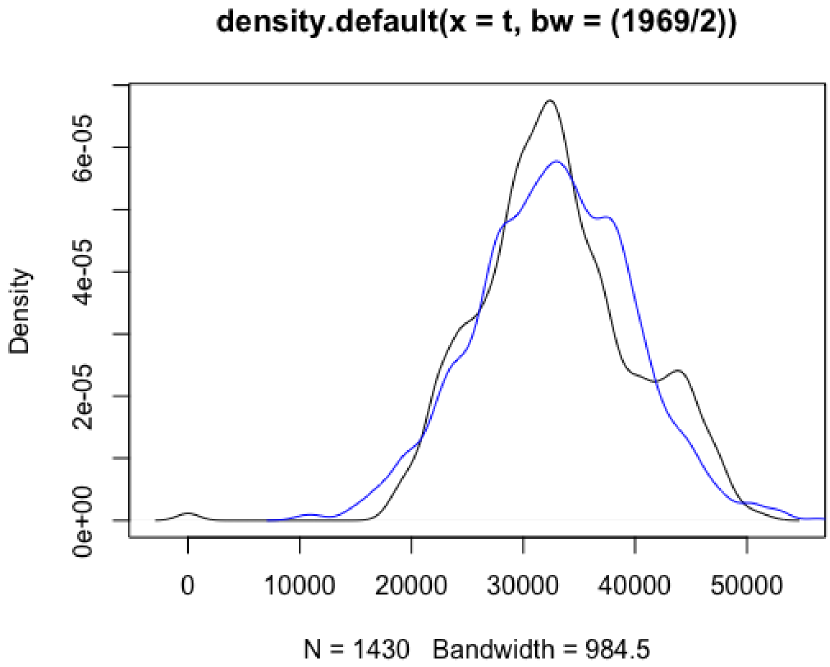
- Rate of California: **43.89865**
- Rate of Wisconsin: **26.88471**
- Rate of Texas: **29.21664**

Analyzing the percentages of sales per year, per month and per day of the week, it is clear that California is the leader of the three states, Texas comes second and Wisconsin is the weaker one.

The histogram of total sales of Walmart:



Then, I create the density plot of total sales and I generate a normal distribution based on my dataset so I can have same length, same mean and same standard deviation.



The sales of the company don't follow the normal distribution. Moreover, I identified some characteristics of non-normality that show my data has differences from a normal distribution:

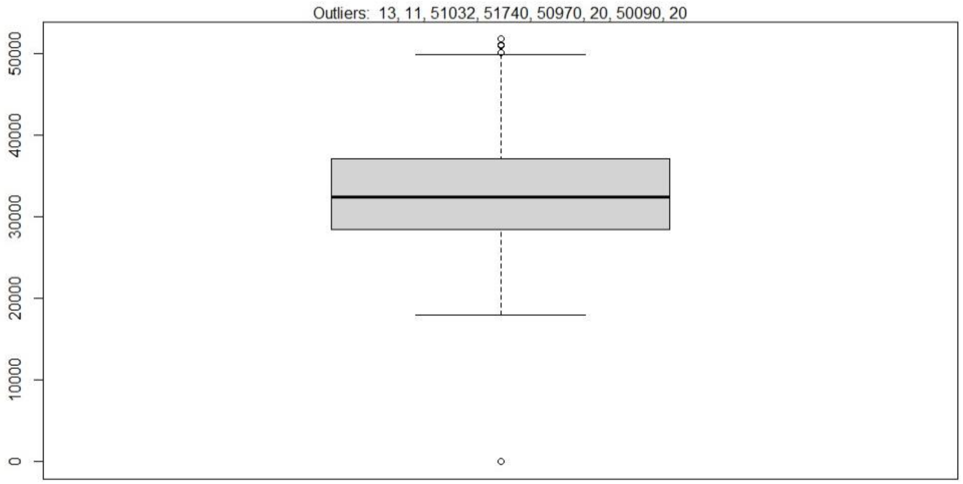
The histogram does not look bell shaped. Instead, it is skewed negatively.

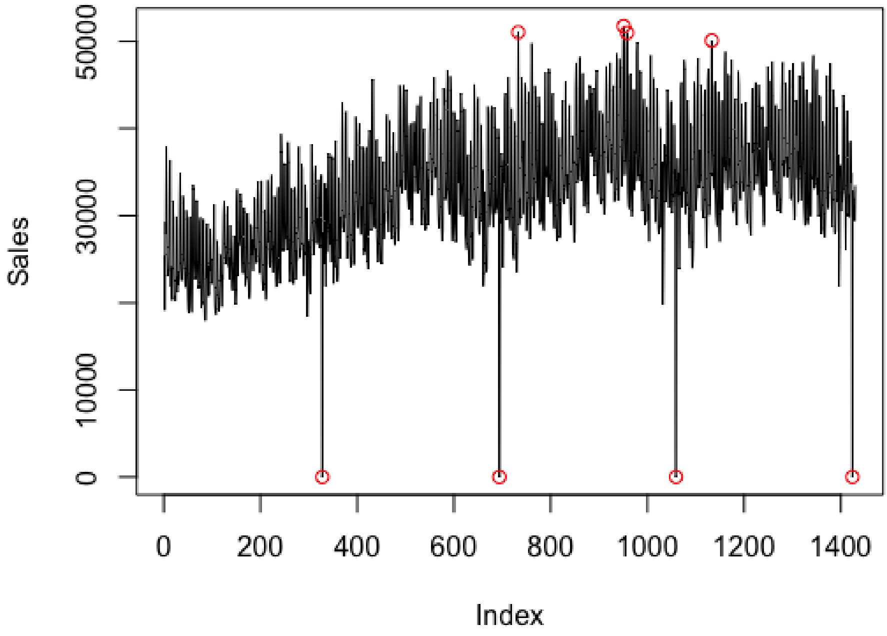
The existence of extreme values in our dataset.

Some values tend to zero or to a natural limit.

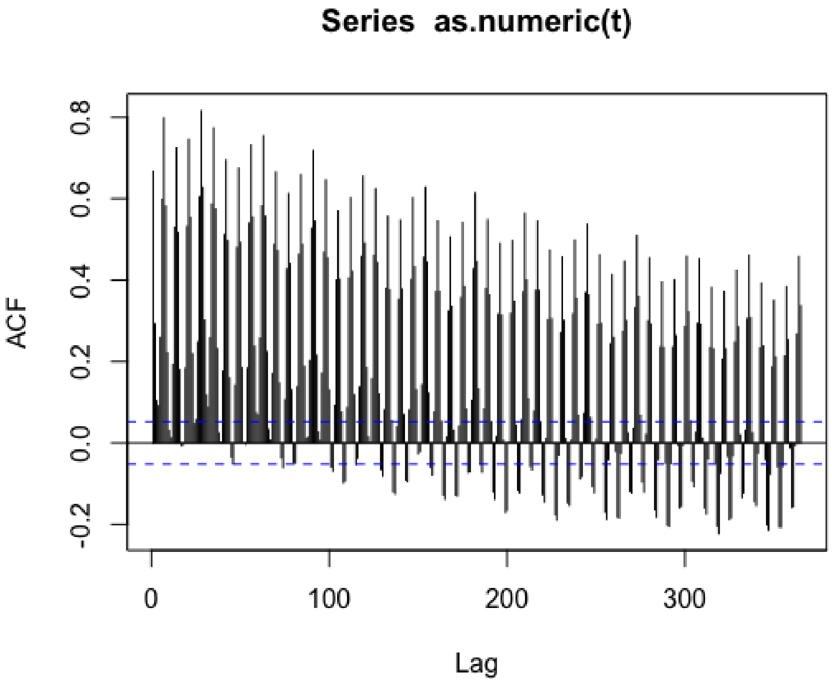
There is known seasonal process in our data such as Holidays, National days etc.

Furthermore, I identify the outliers of the time series. The outliers are shown at the top of the boxplot and are also visualized in the chart below.

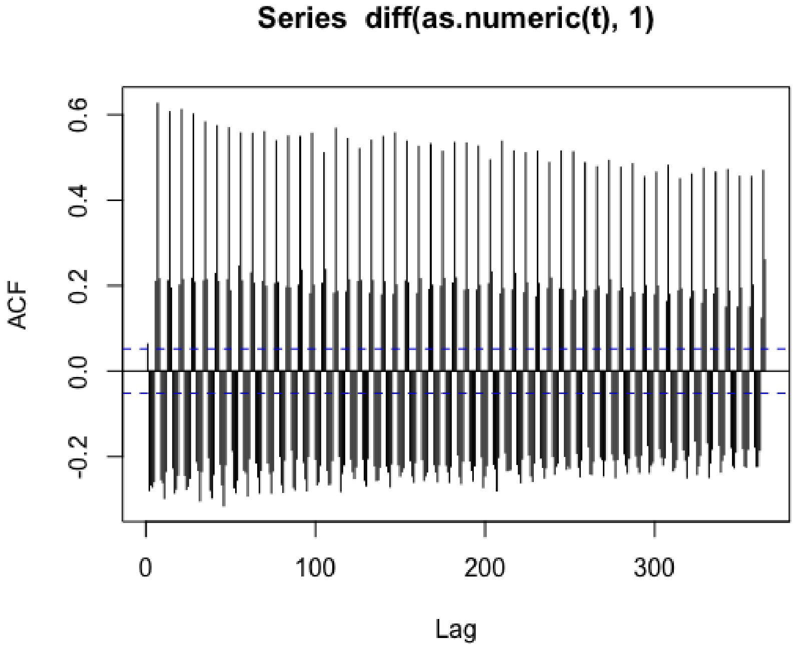




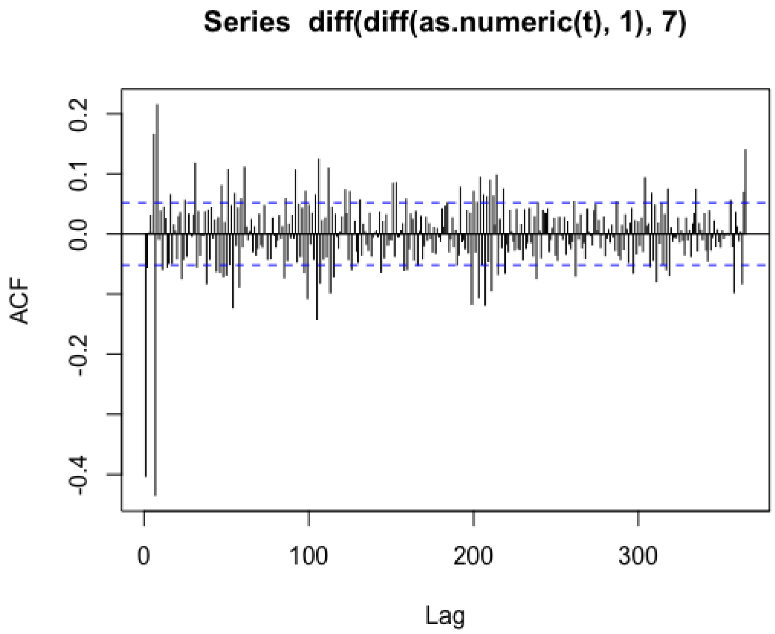
After that, I calculatedc the first and seasonal differences of the total sales time series, considering daily seasonality, as well as the combination of the above differences.



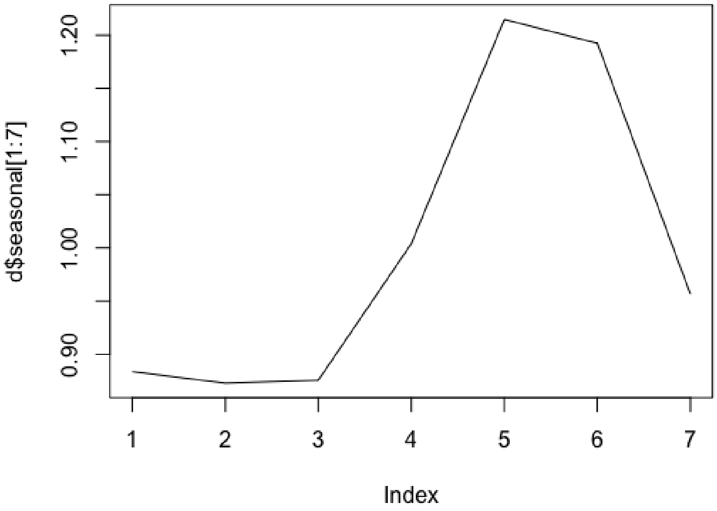
Obviously, we can see seasonal and trend effects. So I removed the trend:

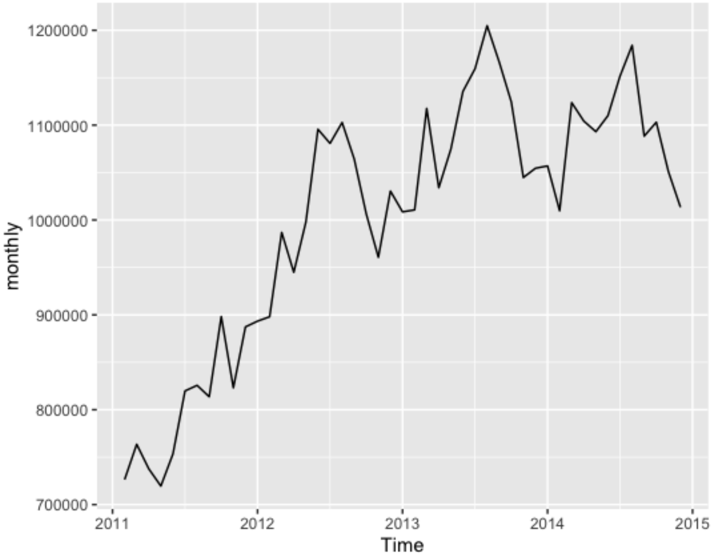


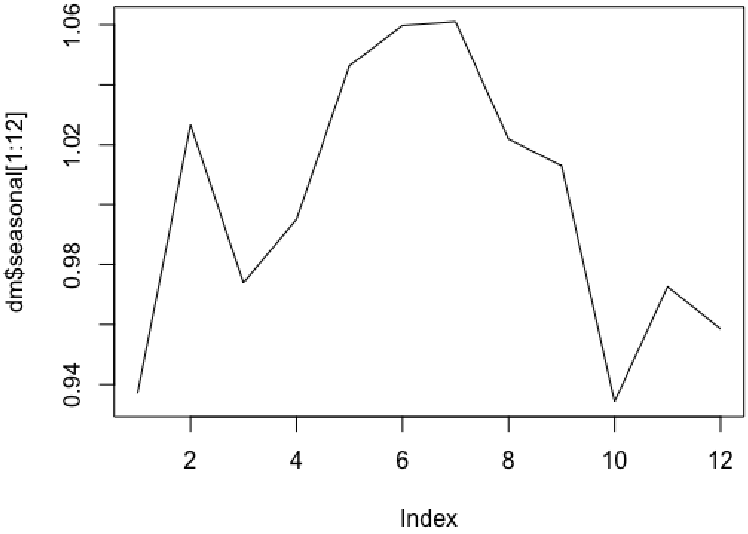
Then I removed the seasonality:



My time series is a good example of a multiplicative time series. As the time series increases in magnitude, the seasonal variation increases as well. Here why I use the multiplicative model(Time series = Trend * Seasonal *Random).

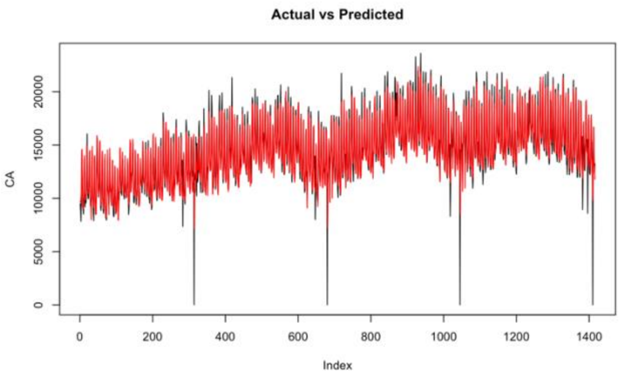


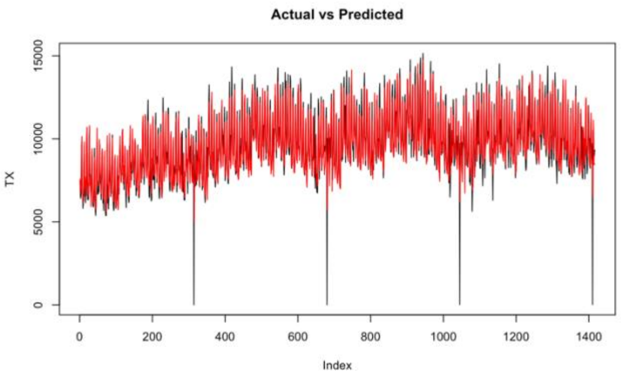


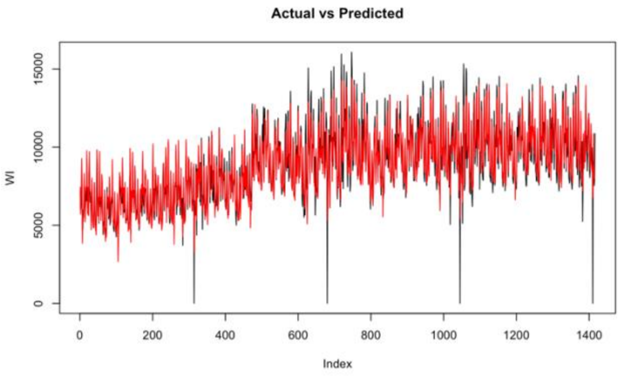


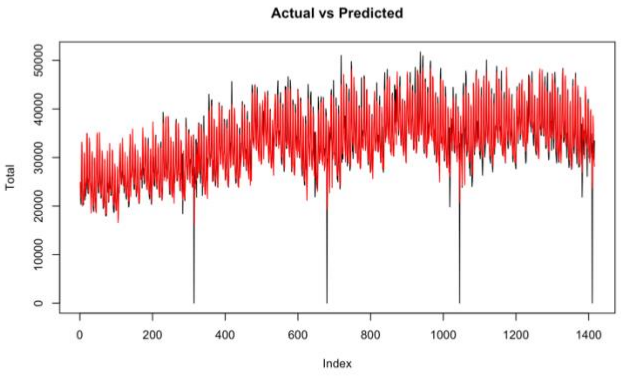
Based on the these charts, it is clear that the last days of the week show the highest sales such as the months during summer and spring.

After that, I started making predictions using linear regression models for each state (CA,TX,WI) and for Total Sales at the end.


Actual vs Predicted


Actual vs Predicted


Actual vs Predicted


Actual vs Predicted