|  |  |
| --- | --- |
| A picture containing banana, fruit, indoor, apple  Description automatically generated  **Report on Analysis of Avocado Prices**  TBANLT 540: Applied Regression Models | Group 4  Gunthala Archana  Barani Bhoopalan  Nitika Raj  Sharmistha Ganguly Ghosh  Sonam Misra |

Contents

[Abstract 2](#_heading=h.qsh70q)

[Introduction: 2](#_heading=h.30j0zll)

[Dataset Description: 3](#_heading=h.1fob9te)

[Data Analysis Plan: 3](#_heading=h.3znysh7)

[Data PreProcessing 4](#_heading=h.2et92p0)

[Exploratory Data Analysis 4](#_heading=h.tyjcwt)

[Descriptive Statistics 5](#_heading=h.1t3h5sf)

[Time Series Output and Analysis 11](#_heading=h.4d34og8)

[Augmented Dickey Fuller Test for type Organic: 12](#_heading=h.2s8eyo1)

[Augmented Dickey Fuller Rest For Conventional Type: 13](#_heading=h.17dp8vu)

[KPSS Test for type Organic: 13](#_heading=h.3rdcrjn)

[KPSS Test For type Conventional: 13](#_heading=h.26in1rg)

[Correlograms for type Organic: 14](#_heading=h.lnxbz9)

[Correlograms for type Conventional: 14](#_heading=h.35nkun2)

[Modeling using auto.arima() for type Organic: 16](#_heading=h.1ksv4uv)

[Modeling using auto.arima() for type Conventional: 17](#_heading=h.44sinio)

[Forecast for type Organic: 17](#_heading=h.gjdgxs)

[Forecast for type Conventional: 18](#_heading=h.2jxsxqh)

[Residual Diagnostic 19](#_heading=h.z337ya)

[ACF and PACF Plot of Residuals for type Organic: 19](#_heading=h.3j2qqm3)

[Histogram of the Residual for type Organic: 20](#_heading=h.1y810tw)

[ACF and PACF Plots of Residual for type Conventional: 20](#_heading=h.4i7ojhp)

[Histogram of the residual for type Conventional: 21](#_heading=h.2xcytpi)

[Conclusion 21](#_heading=h.1ci93xb)

[Limitations 21](#_heading=h.3whwml4)

[Limitations of Time Series Analysis 22](#_heading=h.2bn6wsx)

# Abstract :

This report will investigate the price of avocados in the US market to determine whether there is a price difference between conventional and organic avocados. The data used in this report is an open data from kaggle.com. It is originally from Hass Avocado Board website in May of 2018 & compiled into a single CSV.

Summary statistics and boxplots of the price in each year will be produced by using RStudio functions to explore any price fluctuations throughout the years from 2015 to 2018.

We did Time Series Analysis to understand the variation in average price for organic and conventional avocado across Chicago from the year 2015 to 2018.

**Objectives:** This Analysis Plan focuses on documenting the steps and the analyses techniques that will be used to answer our main research question:

**“The extent to which the average price of avocados is affected by years, regions, types.”**

Initiallywe decided to proceed with OLS regression but later after the inputs from Dr. Margo, we decided to go forward with Time Series Analysis.

**Interesting Findings after the Analyses:**

1. In 2017, the maximum average price for conventional avocados went higher than organic avocados.
2. Variation in average prices was very high in 2016 and very low in 2018.
3. Year 2017 has seen highest average prices for avocados and was the most volatile year.
4. Based on our AUTO ARIMA model, we expect a short downward trend in case of conventional avocados and organic avocados looks stagnant for next few weeks.

In this report we have briefly discussed about Data Analysis Plan, Data Preprocessing and our different analyses on the data. We have included code snippets, graphs and plots of our Analyses.

# Introduction:

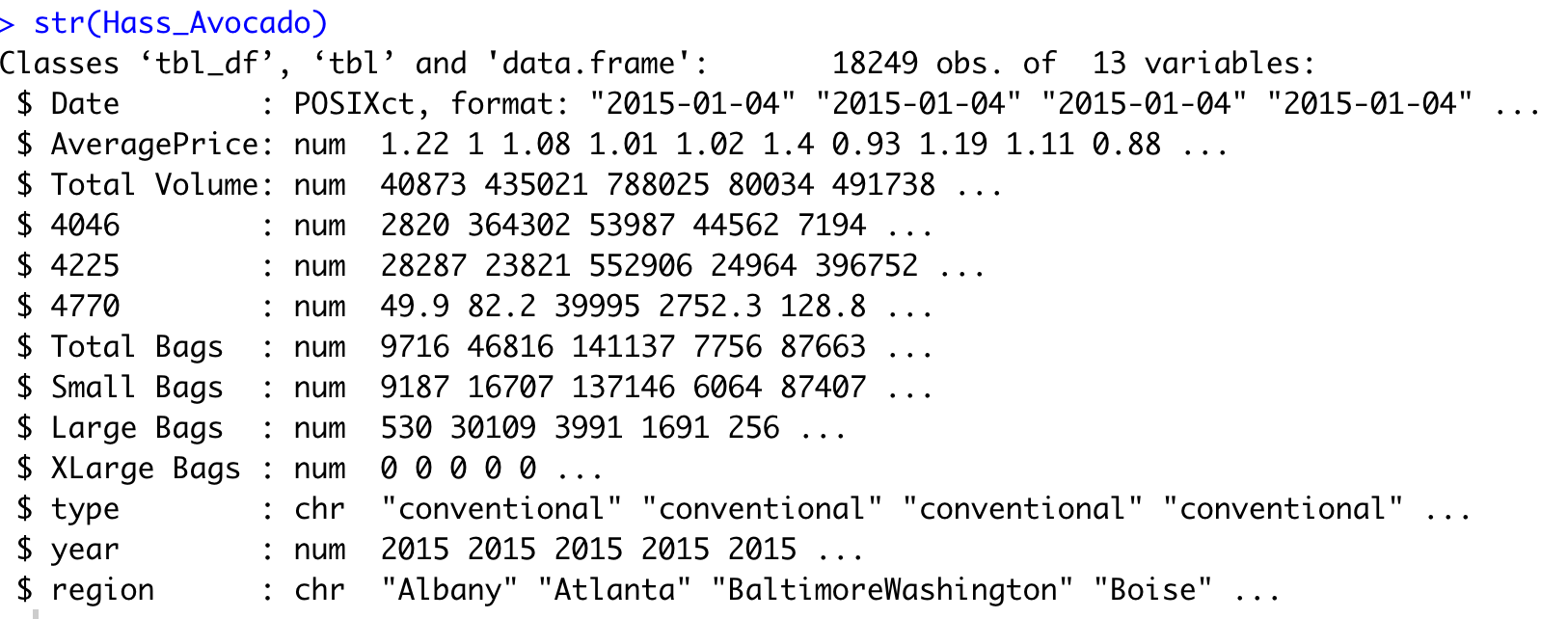
Avocado is a popular fruit in the United States and its price has continued to spike over the years due to its year-round demand, especially from millennials. According to the Department of Agriculture weekly retail price report, the average national price of a Hass avocado was $1.6 on March 4, 2018 compared to $1 from March 6, 2016. In order to better understand the demand and price trend over the years, we decided to proceed with time series analysis. The chosen Hass Avocado Board (HAB) data has over 18,000 observations and is comprised of weekly time series data spanning 168 weeks from January 4, 2015, to March 25, 2018. The data provides average price and total volume of two main Hass avocado types; conventional and organic avocados in 54 regions of the United States. The following time series report intends to analyze,

* the extent to which the average price of avocados is affected by years.
* variation in the average price for organic and conventional avocado across Chicago.
* and predict the next 10 weeks of data for both organic and conventional types with the best parsimonious models.

# Dataset Description:

The data on Hass avocado prices and sales volume in multiple US markets was downloaded from Kaggle, which was originally from Hass Avocado Board website dated March,2018. The below variables are weekly 2018 retail scan data for National retail volume (avocado units) and price.

This study uses statistical methods on data acquired from multi-outlet retail sales details of Hass avocados. The data used in the analysis was extracted from Kaggle and corresponds to 18249 rows and 13 variables of interest. Below is the structure of the dataset:



# Data Analysis Plan:

To analyze the average price of different types of avocado in the Chicago region, we explored how the prices have varied over a period of 4 years (2015-2018) and forecasted the prices for the next 10 weeks.Below steps were performed :

1. Data Preprocessing
2. Data Exploration
3. Modeling :

* Analyzed the time Series Plots (Acf and Pacf)
* Performed tests to find out non-stationarity, drift, seasonality in the data.
* Determined the parameters of the models

p = AR order (the number of lags)

d = Degree of differencing

q = MA order

* Build the model using auto.arima() and verified the results on the basis

of parameters we figured out using acf and pacf plots.

1. Tested the residuals by analyzing the ACF/PACF of the residuals. We found that all autocorrelations for the residual series were non-significant.

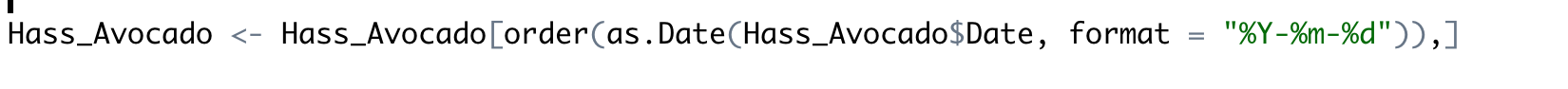
Data PreProcessing:

This dataset was clean and it did not have any null values. Below modifications were made to the dataset for the purpose of the Time series analysis.

1. Below fields were filtered out as per the research question of our analysis. Columns in the dataset and their brief description:

* Date - observation date. The observed dates are weekly.
* Average Price -Average Price of Single Avocado
* Type - Conventional or Organic
* Region- The city or region of the observation

1. Date was not sorted. It is ordered from ascending to descending by using below code:



# Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics with visual methods. Here we have box plots and summarize function.

**Box plot:** A box plot is a graph that gives us five number summaries of a set of data(minimum, first quartile, median, third quartile and maximum). Here we have used box plots for year and type of avocados.

**Summary function:** Summary function is used to produce results of various model fitting functions.

We start the analysis by Data Filtering.

**Data Filtering:** We filtered the data for the descriptive analysis based on type of Avocado and year.

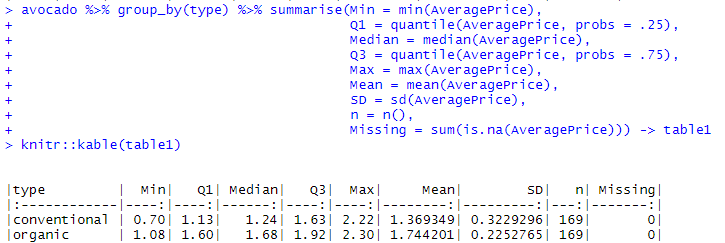
A screenshot of a cell phone

Description automatically generated

### Descriptive Statistics

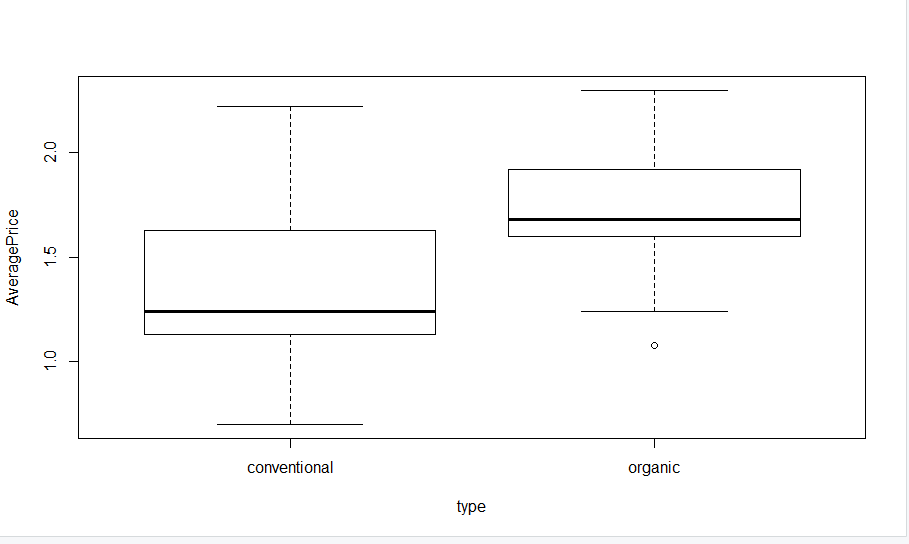
# We did the descriptive statistics based on the type and Year:

* **Type:**

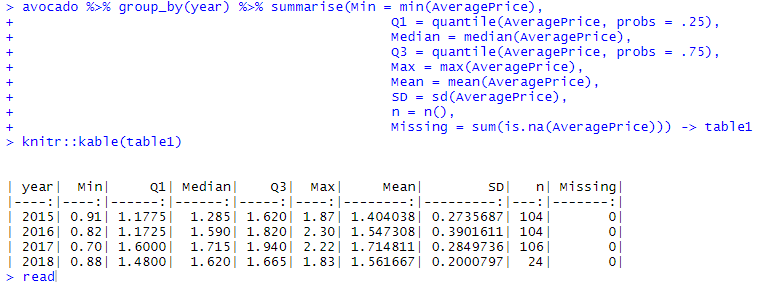


**Boxplot:**

avocadodesc %>% boxplot(AveragePrice ~ type, data = ., y1ab = "Price(US$)")



* **Year:**

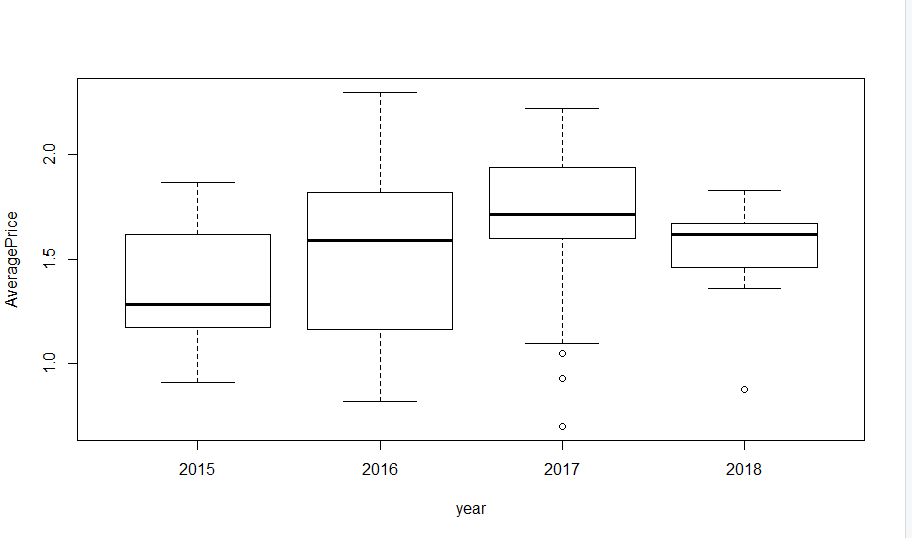


**Boxplot:**

The boxplot below represents average avocado prices across all years for Chicago City.

avocado <- Hassavocado1 %>% filter(region == "Chicago")

boxplot(data = avocado, AveragePrice ~ year)



We can notice a slight variation in the prices over years, though 2017 was noted as the lowest. The maximum price was observed in 2016 i.e. $2.3

Now let’s go deeper and see how the price varied for each year in Chicago

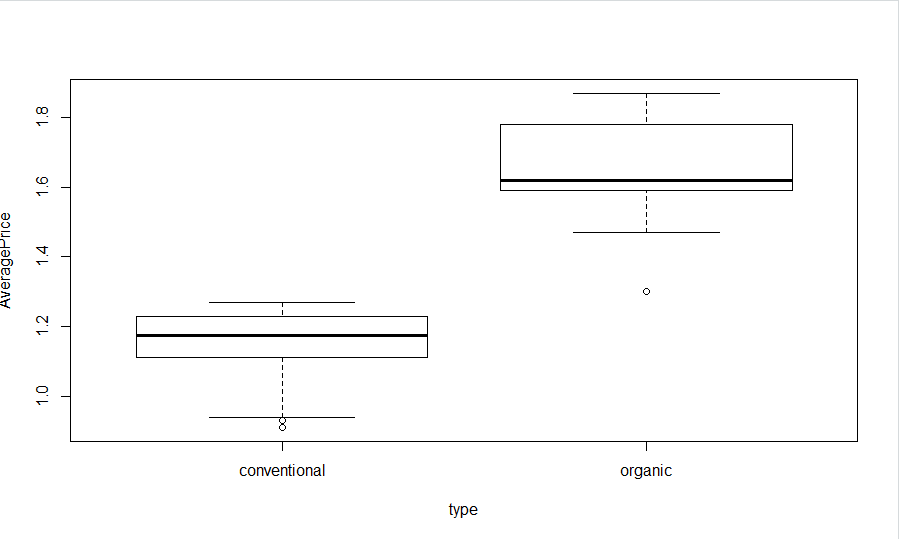
* **Type and Year together:**
  + - * **For the year 2015:**

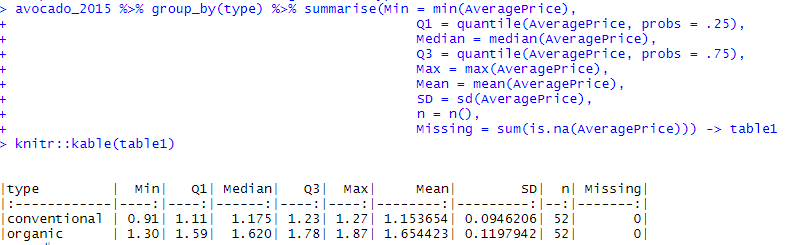
The boxplot below represents organic avocado prices in 2015 for Chicago state.

#avocado data year wise for Chicago.

organic <- Hassavocado1 %>% filter(region == "Chicago", year == 2015)

boxplot(data = organic, AveragePrice ~ type)





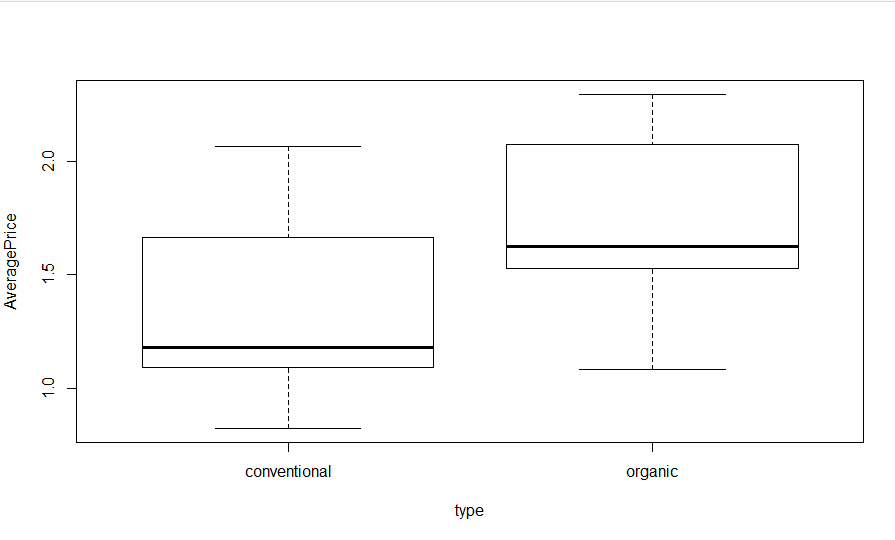
In 2015 we can see a good difference in the average prices of organic and conventical avocados as assumed.

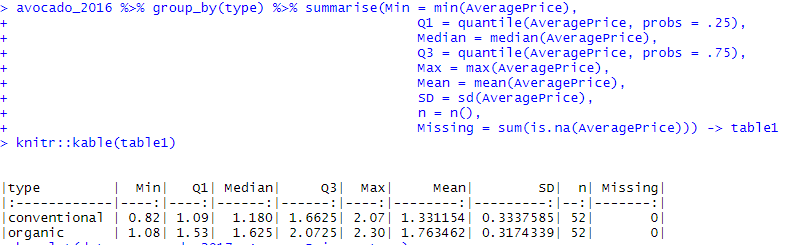
* + - * **For the year 2016:**

The boxplot below represents organic avocado prices in 2016 for Chicago state.

organic <- Hassavocado1 %>% filter(region == "Chicago", year == 2016)

boxplot(data = organic, AveragePrice ~ type)





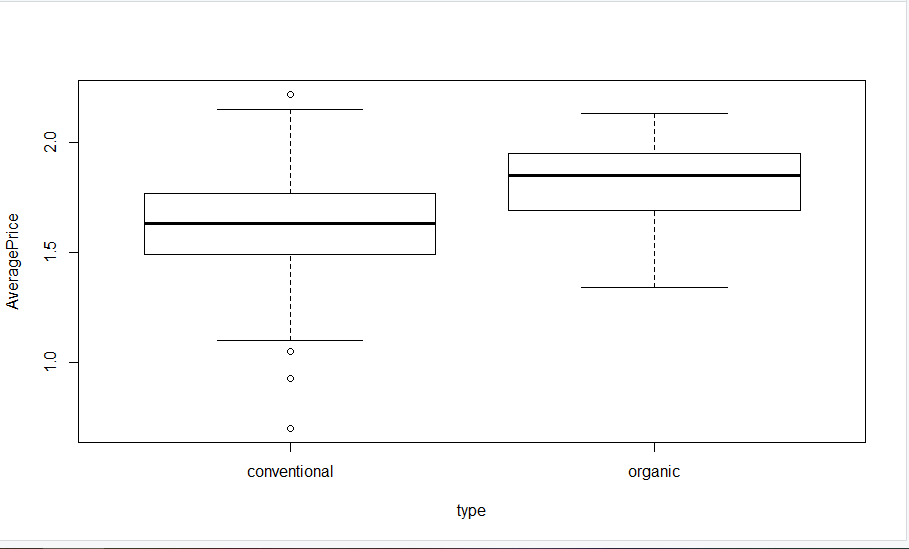
Here in 2016 the deviation in the average prices is slight less compared to the prices in 2015. And the prices for both type of avocados went as high as $2.30 which is much more than in 2015. For conventional we can see a huge difference in average price when compared to 2015. Therefore, we can say that the prices went high in a year.

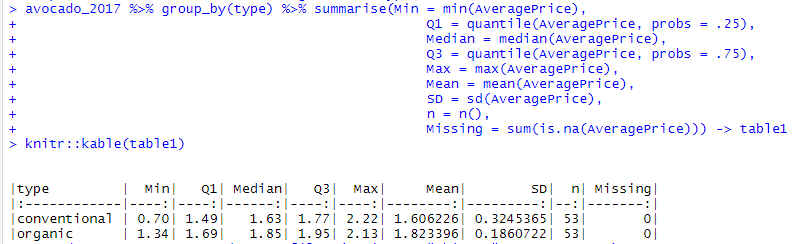
* + - * **For the year 2017:**

The boxplot below represents organic avocado prices in 2017 for Chicago state.

organic <- Hassavocado1 %>% filter(region == "Chicago", year == 2017)

boxplot(data = organic, AveragePrice ~ type)

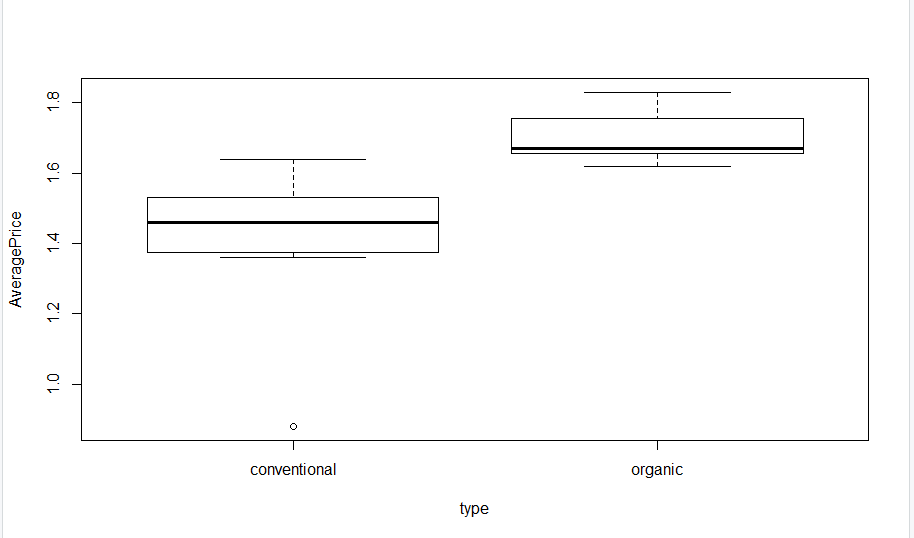


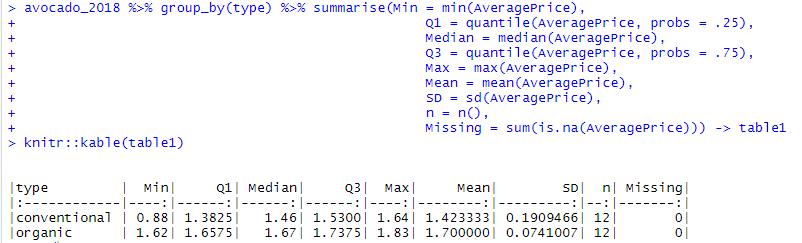


We can see something very unusual thing in 2017, the maximum average price for conventional avocados went high than organic avocados.

* + - * **For the year 2018:**

avocado\_2018 %>% boxplot(AveragePrice ~ type, data = ., y1ab = "Price(US$)")

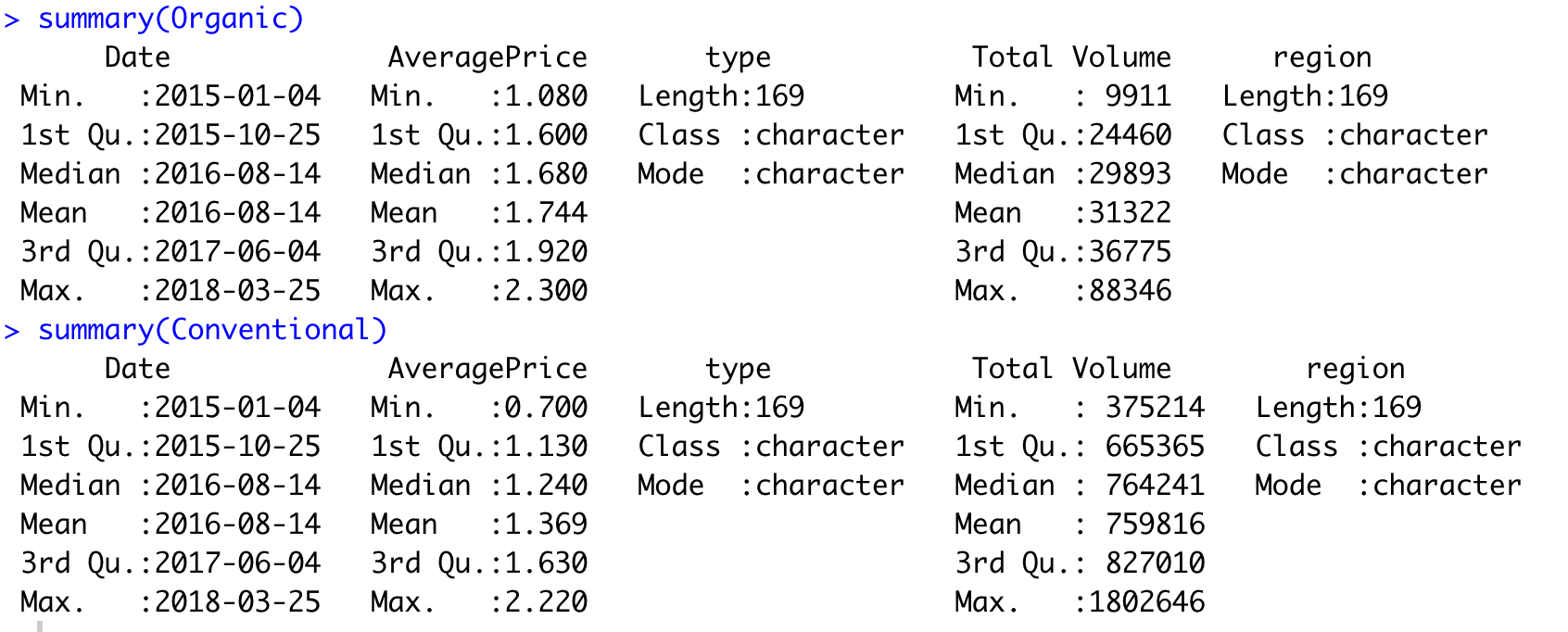




Again in 2018 the prices went up with a slight variation. As we can see in case of organic avocados the variation is very less in average minimum price and average maximum price. That means the prices for organic avocados did not go up with huge variations.

Overall, we have seen that the variation in average prices was very high in 2016 and very low in 2018.

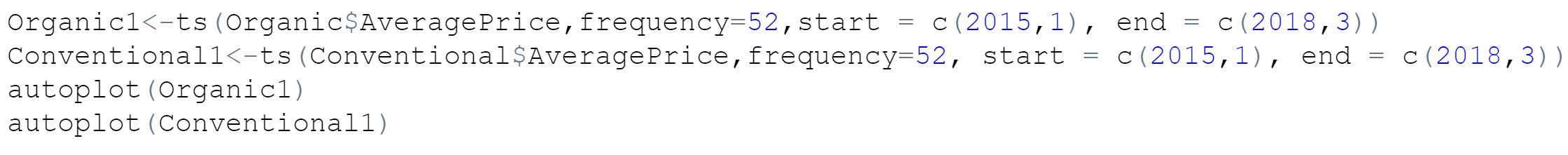
**Descriptive statistics for Organic and Conventional Avocado for the chosen city (Chicago)**



# Timeseries Output and Analysis:

To understand the variation in average price for organic and conventional avocado across Chicago from the year 2015 to 2018 and find the most parsimonious model for prediction, we have followed the below mentioned steps.

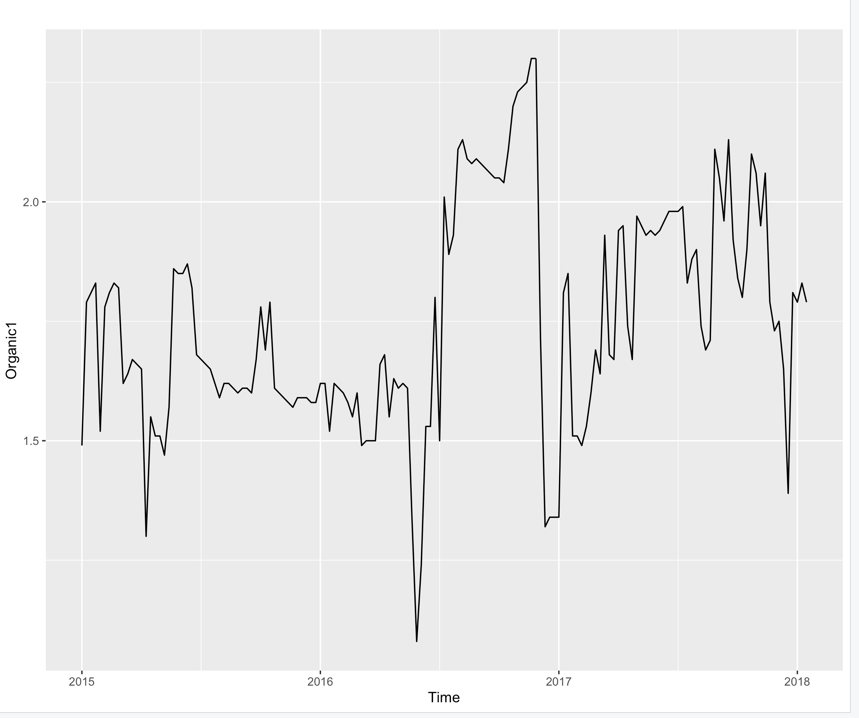
**Step 1:** Converted the dataset into time series data and created plot for each type of avocado using R code mentioned below:



Note: As our dataset is a weekly data, we have used frequency value as 52.

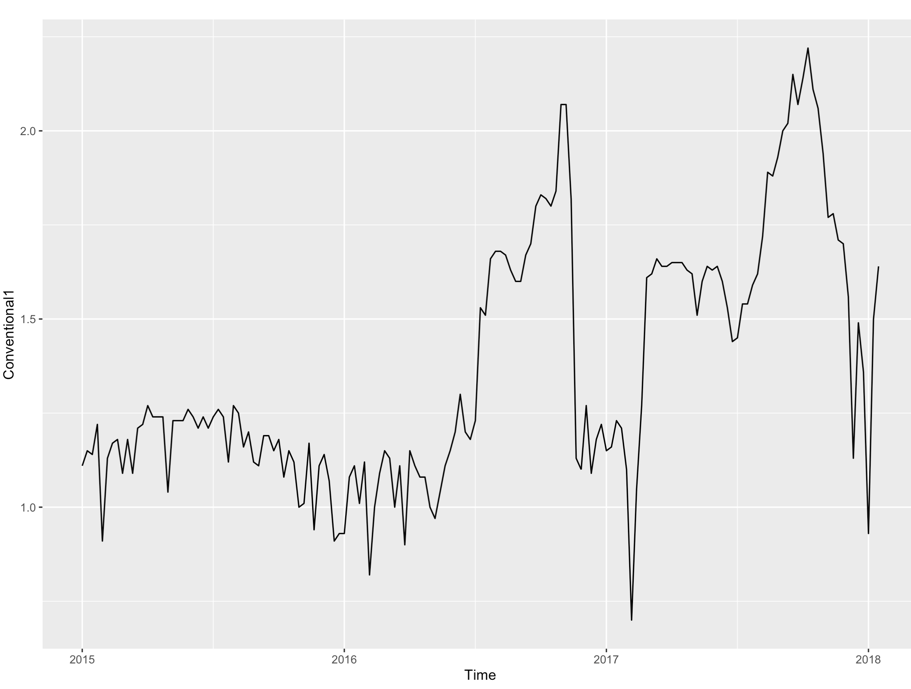
**Output:**

* **Timeseries plot for Average Price and Organic Avocado Type:**

****

We can observe 2017 is the year with highest prices and price seems to be volatile during this year.

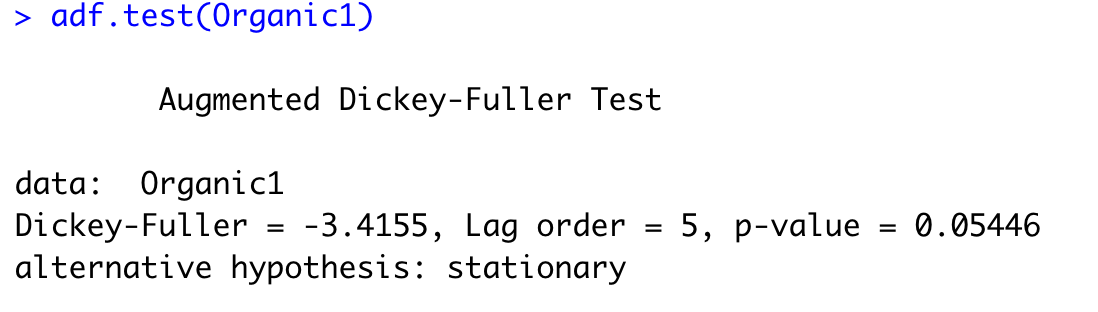
* **Timeseries plot for Average Price and Conventional Avocado Type:**

****

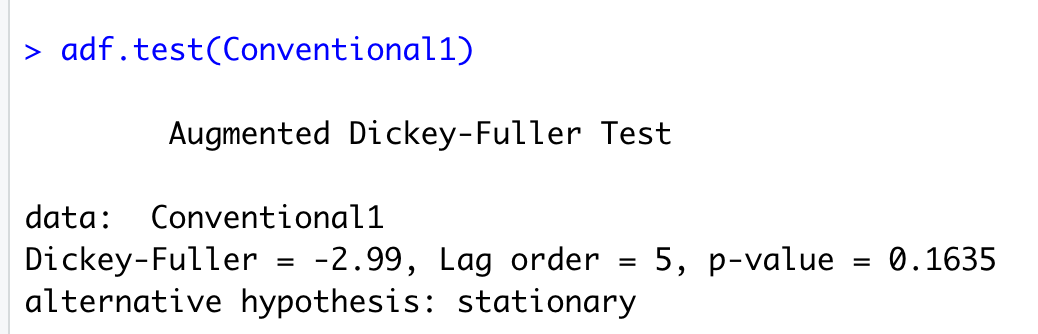
In case of conventional avocados also 2017 has seen the highest average price and it took a sharp dip in the same year. We can say that the price in this year seen high volatility.

**Step 2:** For the time series analysis, the data need to be stationary. From the above plots we cannot say that the data is stationary or not. So, we have performed stationarity test on both organic and conventional type dataset.

### Augmented Dickey Fuller Test for type Organic:



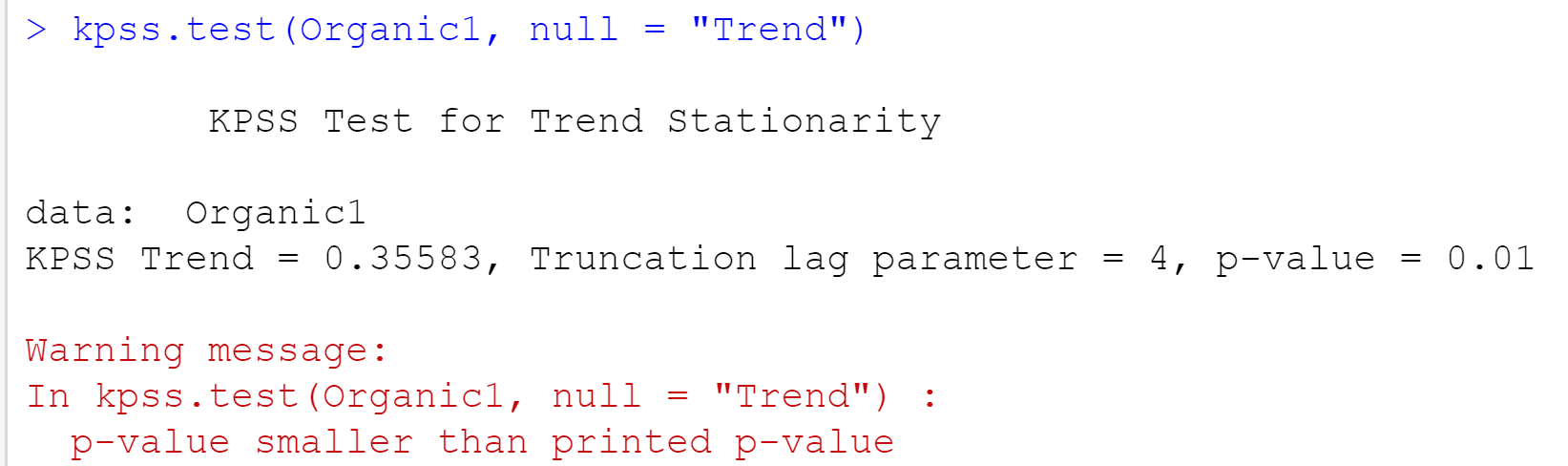
### Augmented Dickey Fuller Rest For Conventional Type:

****

From the above results, we can say that the data is non-stationary for both types of avocados as the p-value for each case is greater than 0.05 and thus didn’t reject the null hypothesis for stationary.

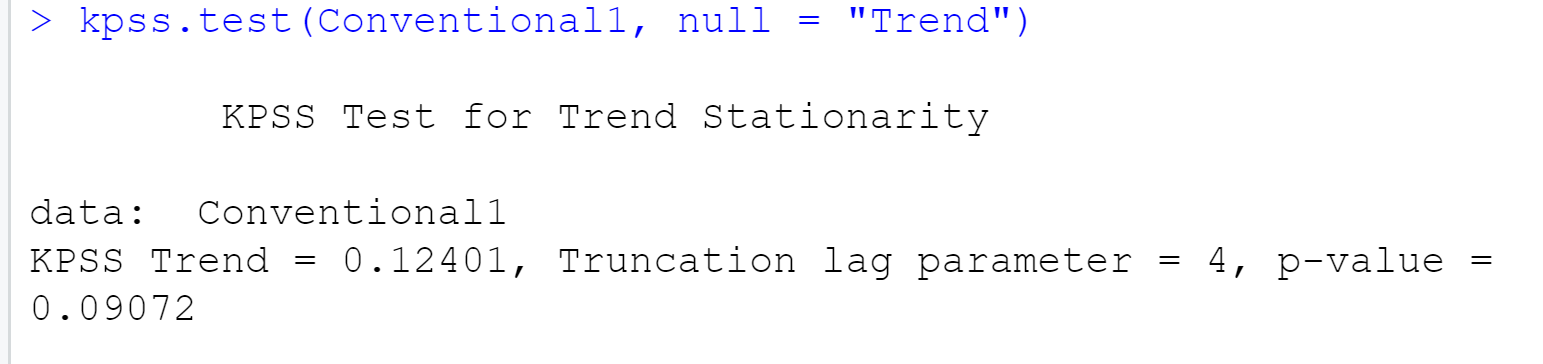
**Step 3:** Further we have checked for Trend Stationarity using KPSS test for the data. The R codes and outputs are mentioned below:

### KPSS Test for type Organic:

****

From the above result we can say that the data with type organic is not trend stationary as the p value is smaller than 0.05 and thus rejects the null hypothesis.

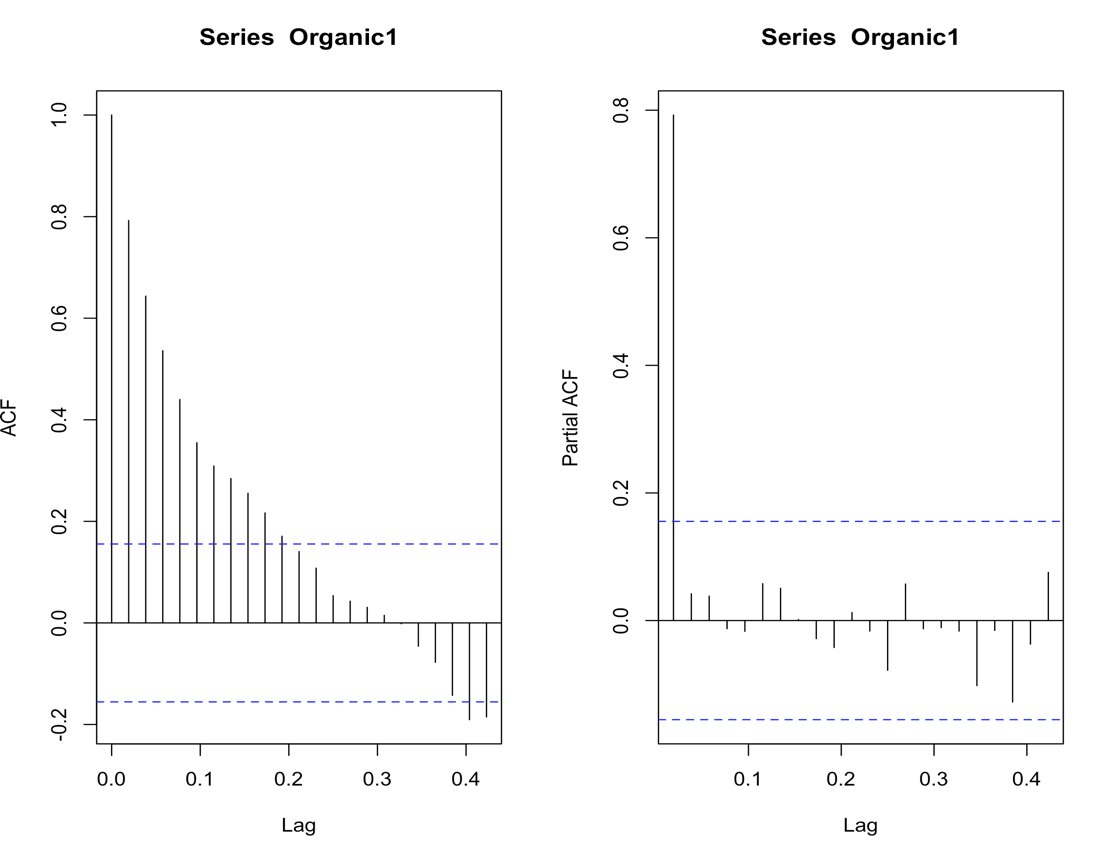
### KPSS Test For type Conventional:

****

From the above result we can say that the data with type conventional is trend stationary as the p value is greater than 0.05 and thus didn’t reject the null hypothesis.

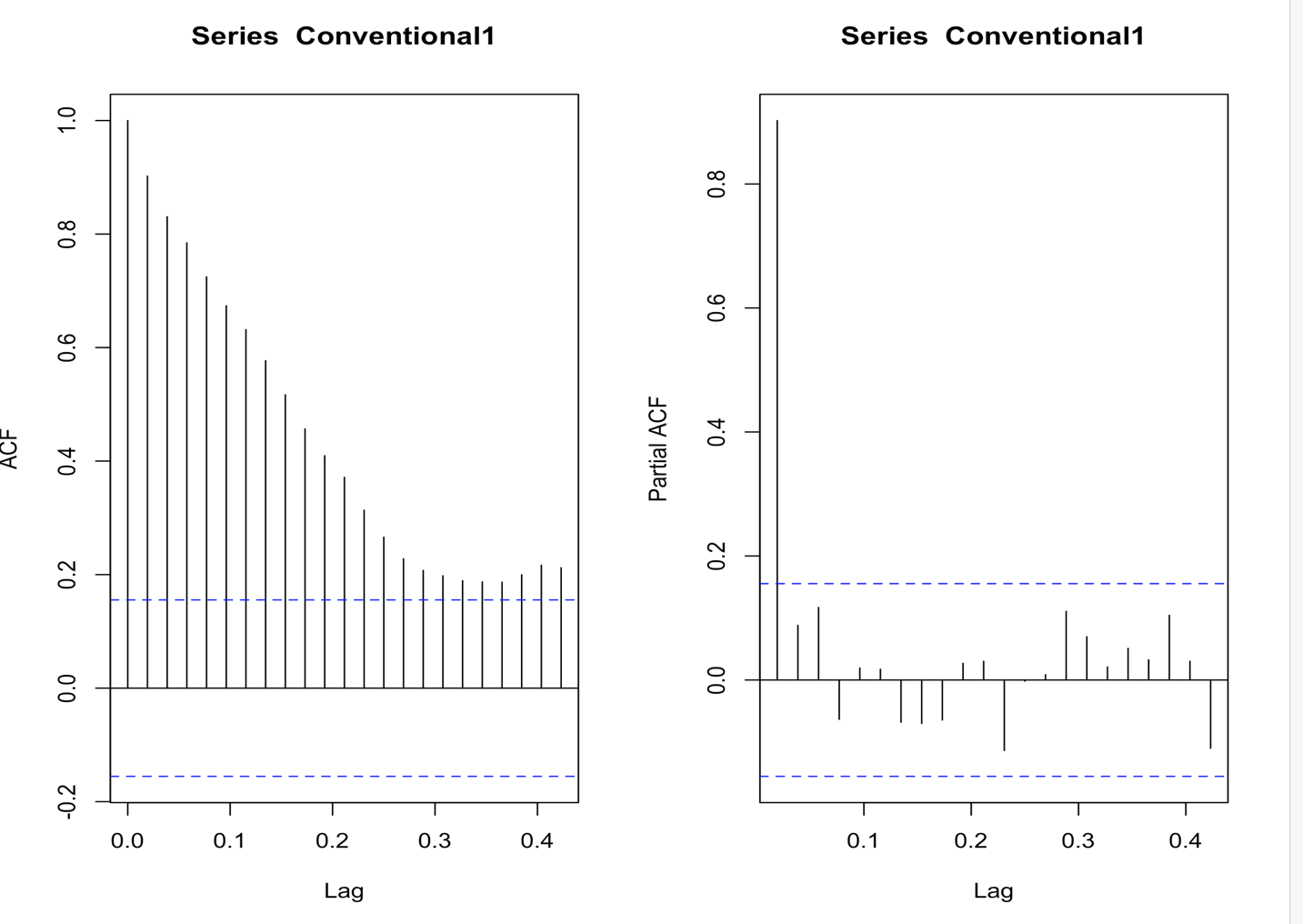
**Step 4:** To further analyze the data for error and lags, we have created thecorrelograms for both organic and conventional type.

### Correlograms for type Organic:

****

From the above ACF and PACF graph, we can say that the model for type organic could be an Autoregressive model as ACF trails off after a lag and has a hard cut-off in the PACF after a lag. Also, the model has potential for seasonal component. We will further analyse our hypothesis using auto.arima() function.

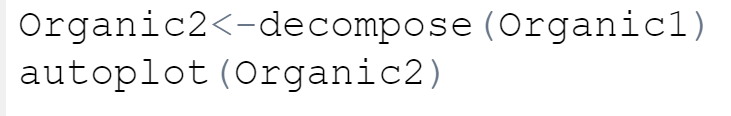
### Correlograms for type Conventional:

****

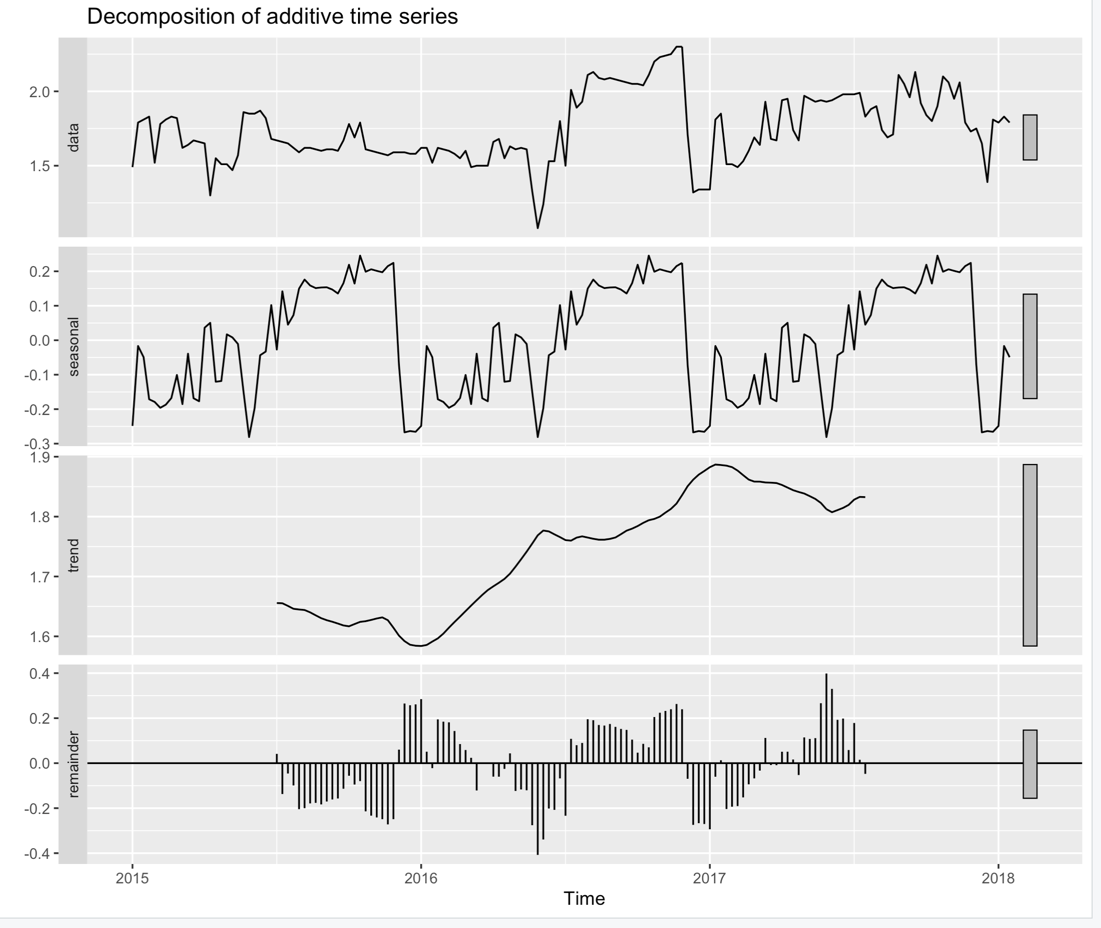
From the above ACF and PACF graph, we can say that the model for type conventional could be an Autoregressive model as ACF trails off after a lag and has a hard cut-off in the PACF after a lag. Also, the model has potential for seasonal component. We will further analyse our hypothesis using auto.arima() function.

**Step 5:** To identify the randomness, Trend and Seasonality in the data graphically, we have created the plots using decompose function for both organic and conventional type using the R codes mentioned below:

* **For type Organic:**

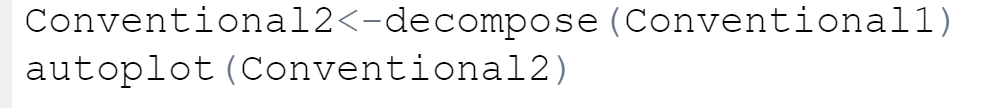
****

**Output:**

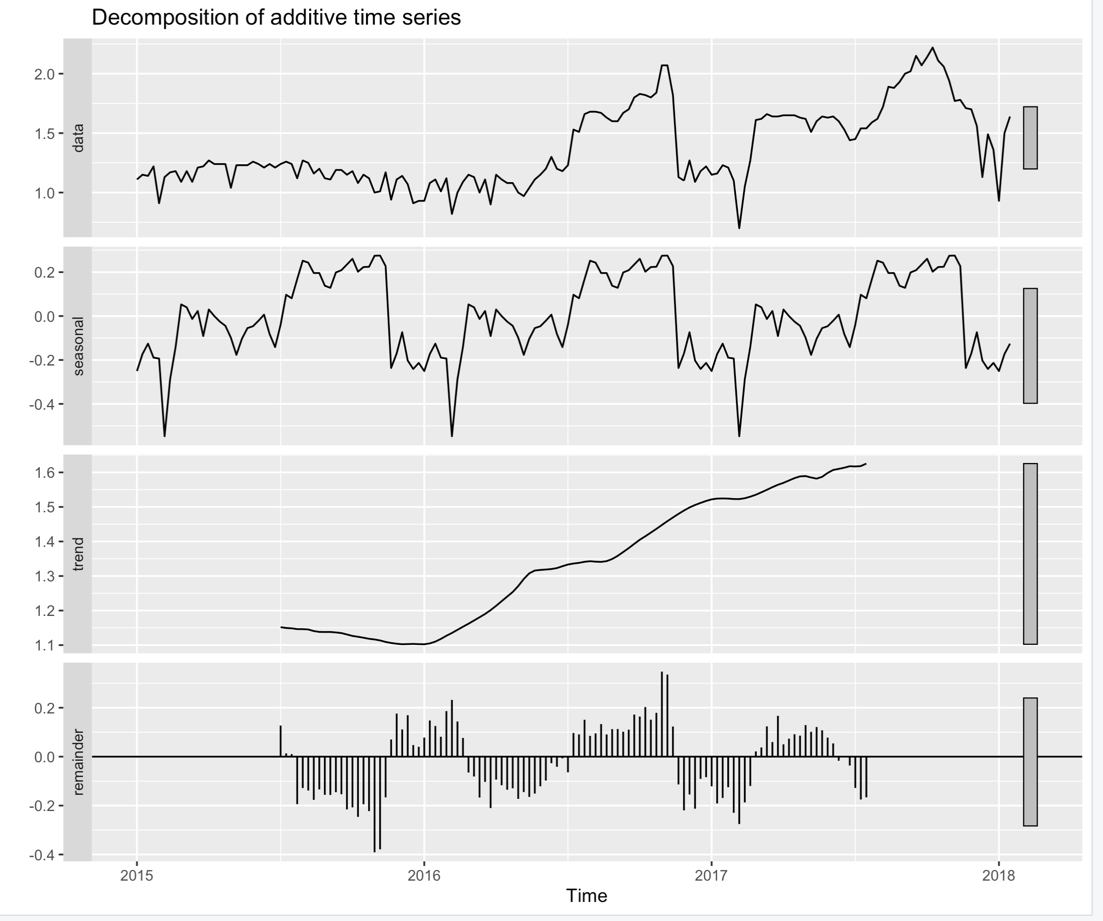
****

From the above graph, we can say that the type organic data has an upward trend and has seasonal pattern. Also, there is some randomness in the data but is not exactly replicating the dataset and thus we can say that there might be some other factors contributing to this randomness apart from trend and seasonality. We will further explore our analysis using auto.arima() function.

* **For type Conventional:**

****

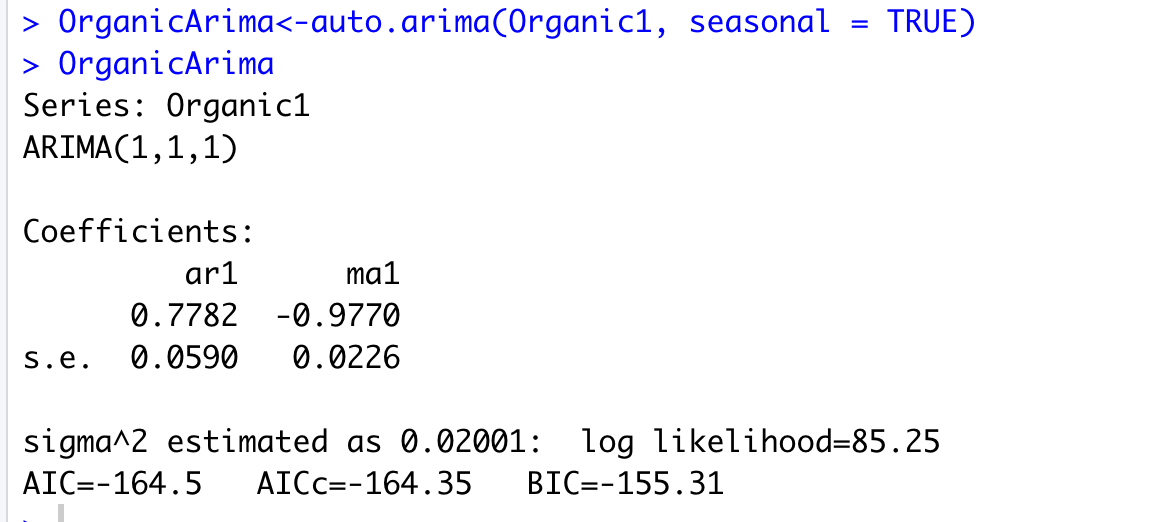
**Output:**

****

From the above graph, we can say that the type organic data has an upward trend and has seasonal pattern. Also, there is some randomness in the data but is not exactly replicating the dataset and thus we can say that there might be some other factors contributing to this randomness apart from trend and seasonality. We will further explore our analysis using auto.arima() function.

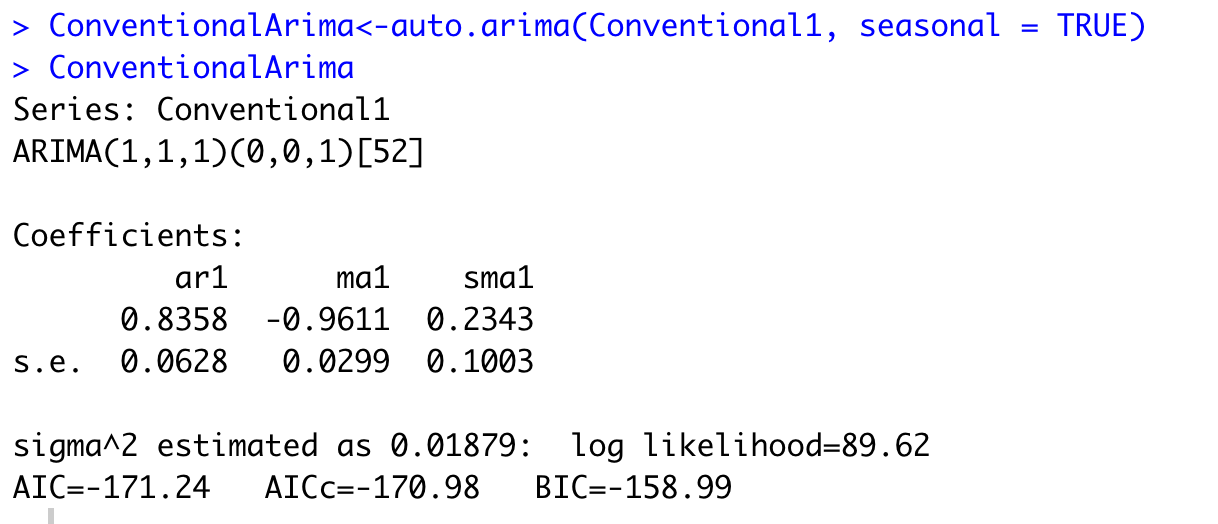
**Step 6:** To identify the most parsimonious model for both organic and conventional type, we have used auto.arima() function. The R codes and the model with AIC value are shown below:

### Modeling using auto.arima() for type Organic:

****

From the above result, it was found that the model for type organic is mix AR and MA model with one difference and has no seasonal component. The p and q values for the model are 1 and 1 respectively and the AIC value is -164.5

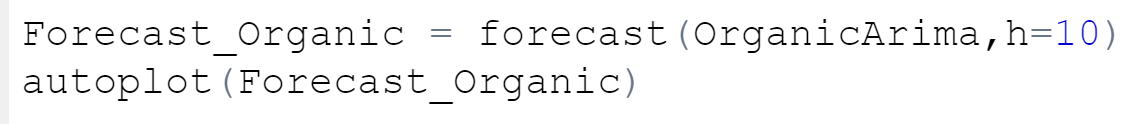
### Modeling using auto.arima() for type Conventional:

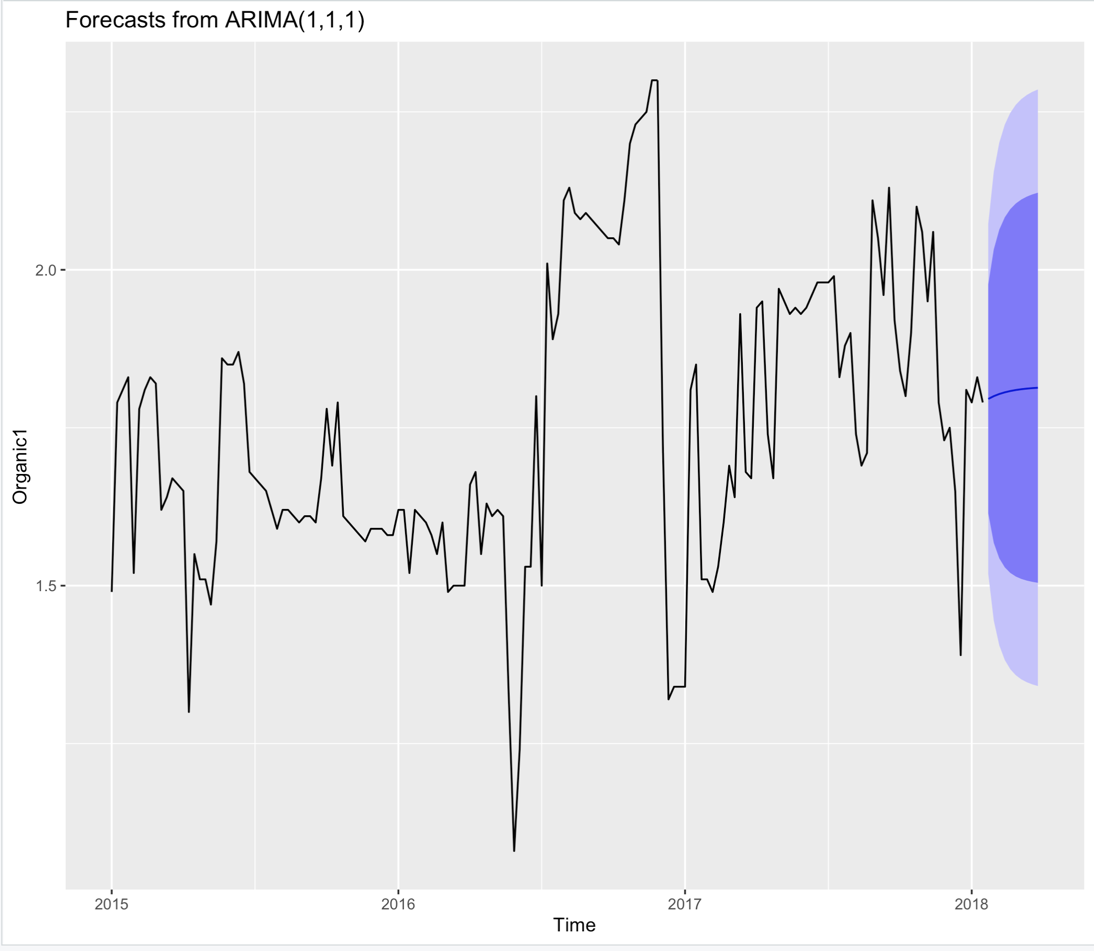
****

From the above result, it was found that the model for type Conventional is mix AR and MA model with one difference and has seasonal component. The AIC value for this model is -171.24. The p and q values for the model are 1 and 1 respectively including the seasonal component with q=1.

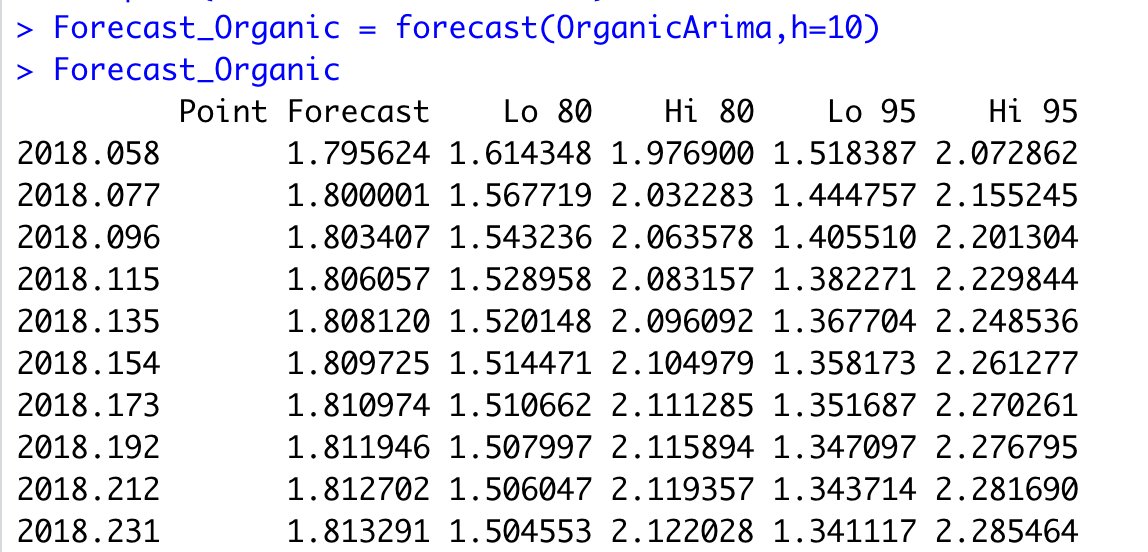
**Step 7:** To forecast and predict the next 10 weeks data for both organic and conventional type with the parsimonious models found above, we have used below mentioned R codes:

### Forecast for type Organic:

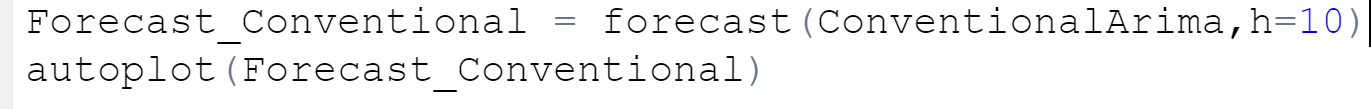
****

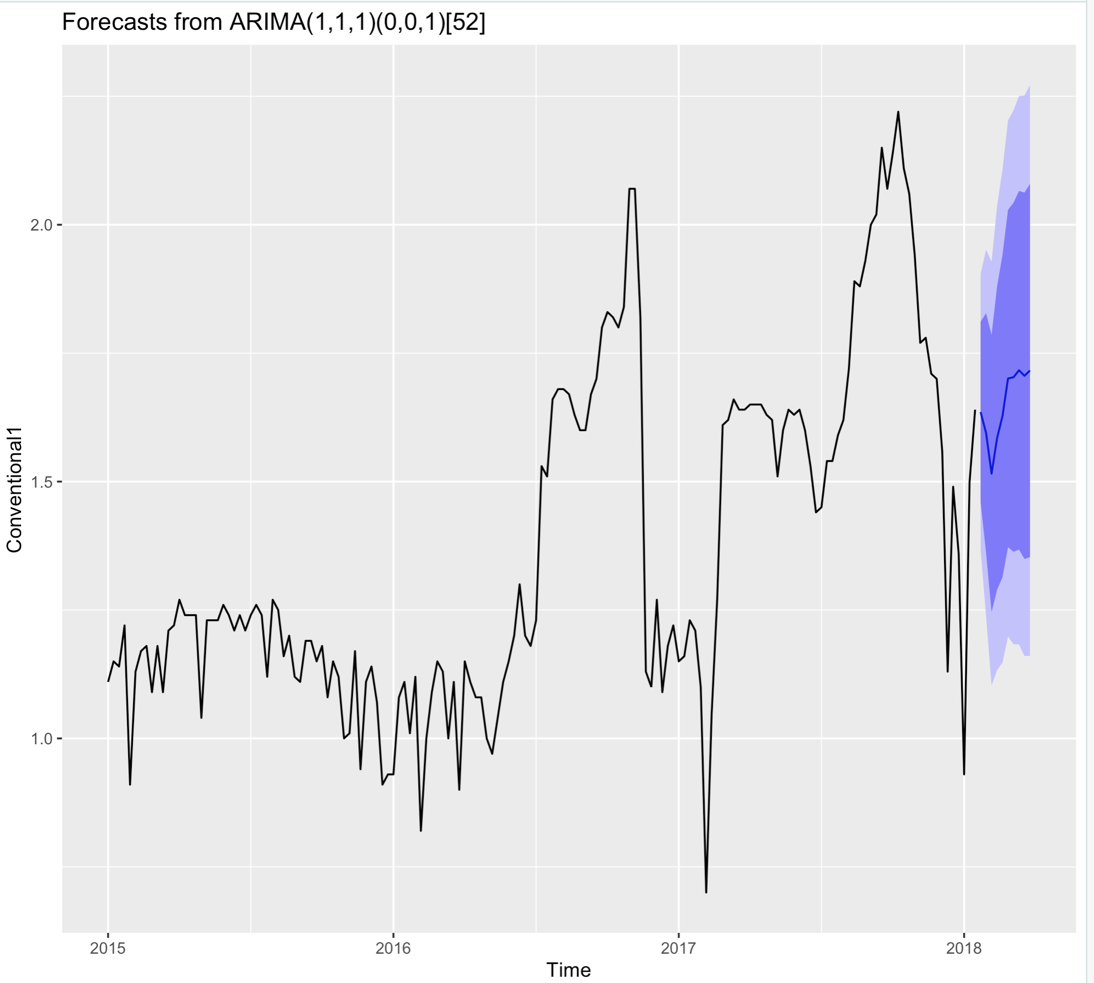
****

**Forecast:**

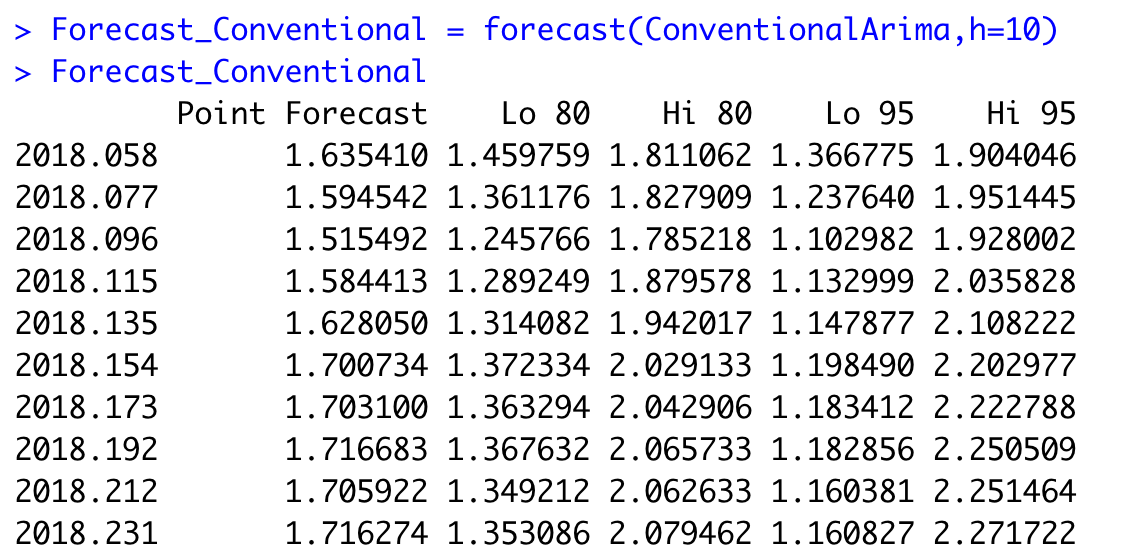
****

### Forecast for type Conventional:

****

****

**Forecast:**

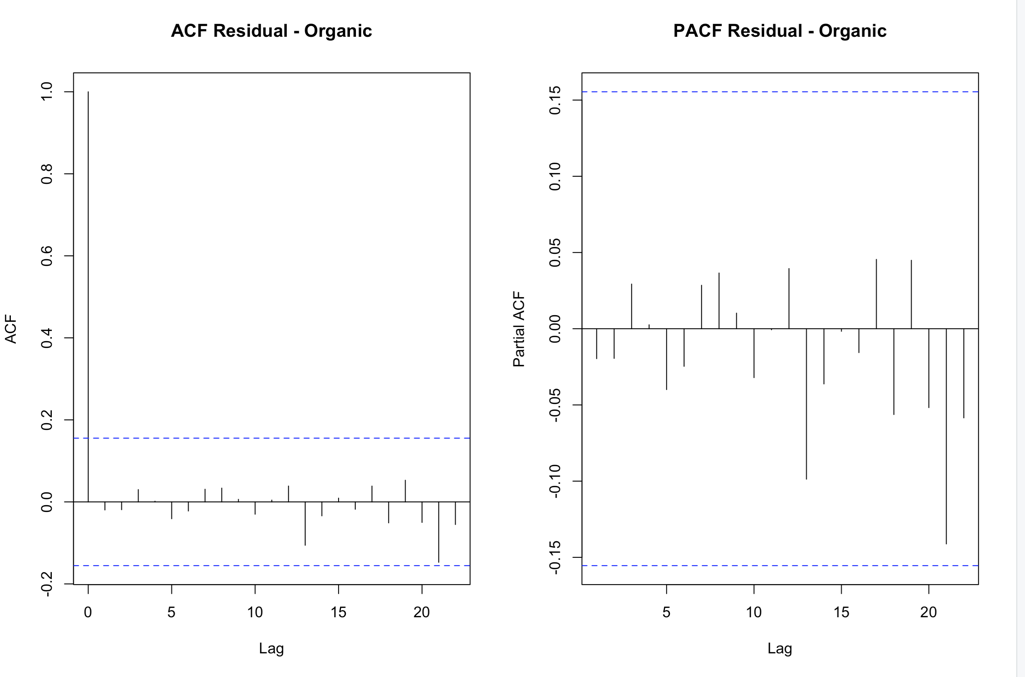
****

# Residual Diagnostic

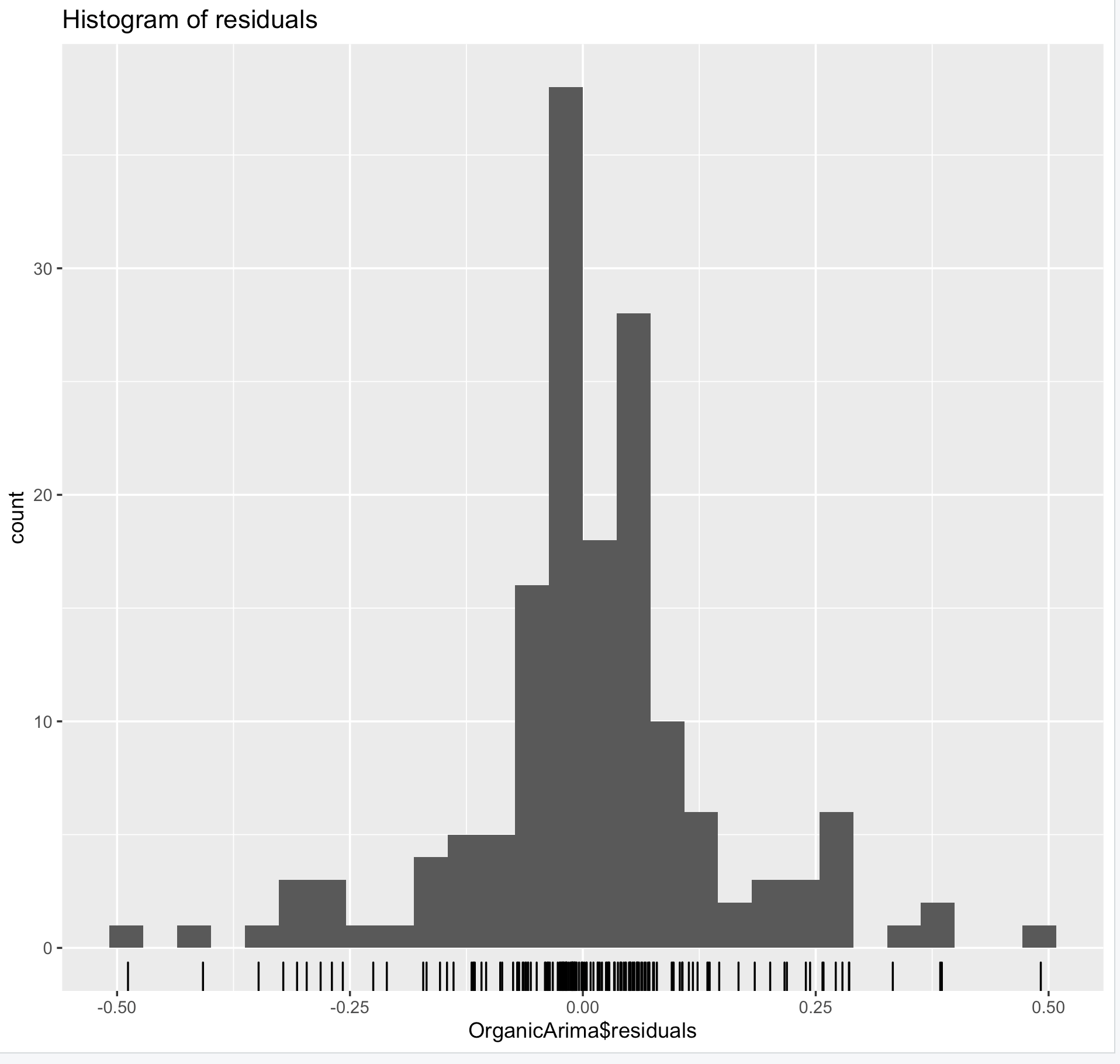
To check the accuracy of the model we have performed residual diagnostic by analyzing the correlation, mean and distribution of the residuals. Below are the observations:

* Acf/Pacf plots of the residuals for organic and conventional type shows that mean of the residual is close to 0 and there is no significant correlation in the residual as the spikes where within the significance bounds represented by blue dotted line.
* Histogram plots for both organic and conventional suggested that the residuals are normally distributed and have constant variance.

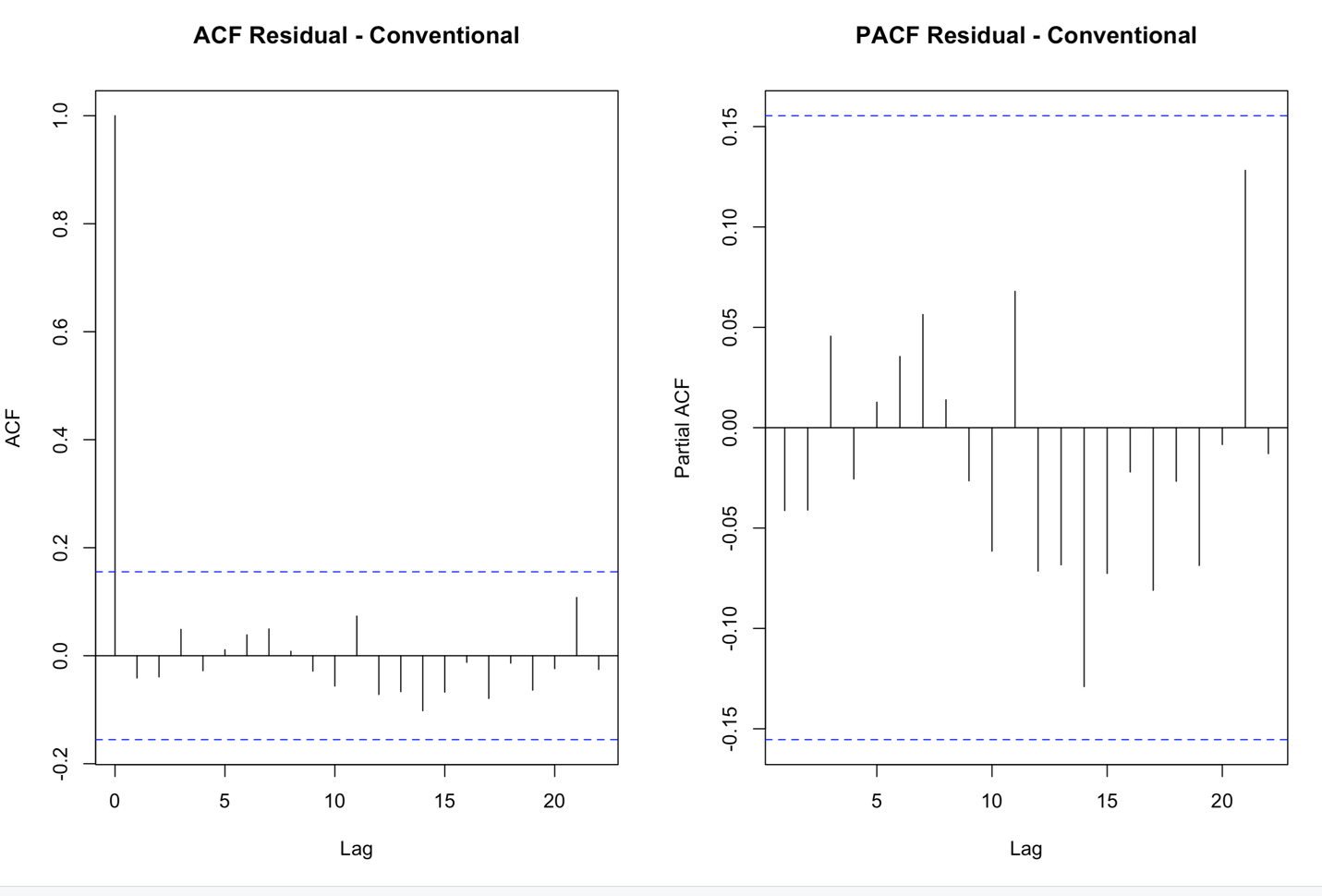
### ACF and PACF Plot of Residuals for type Organic:



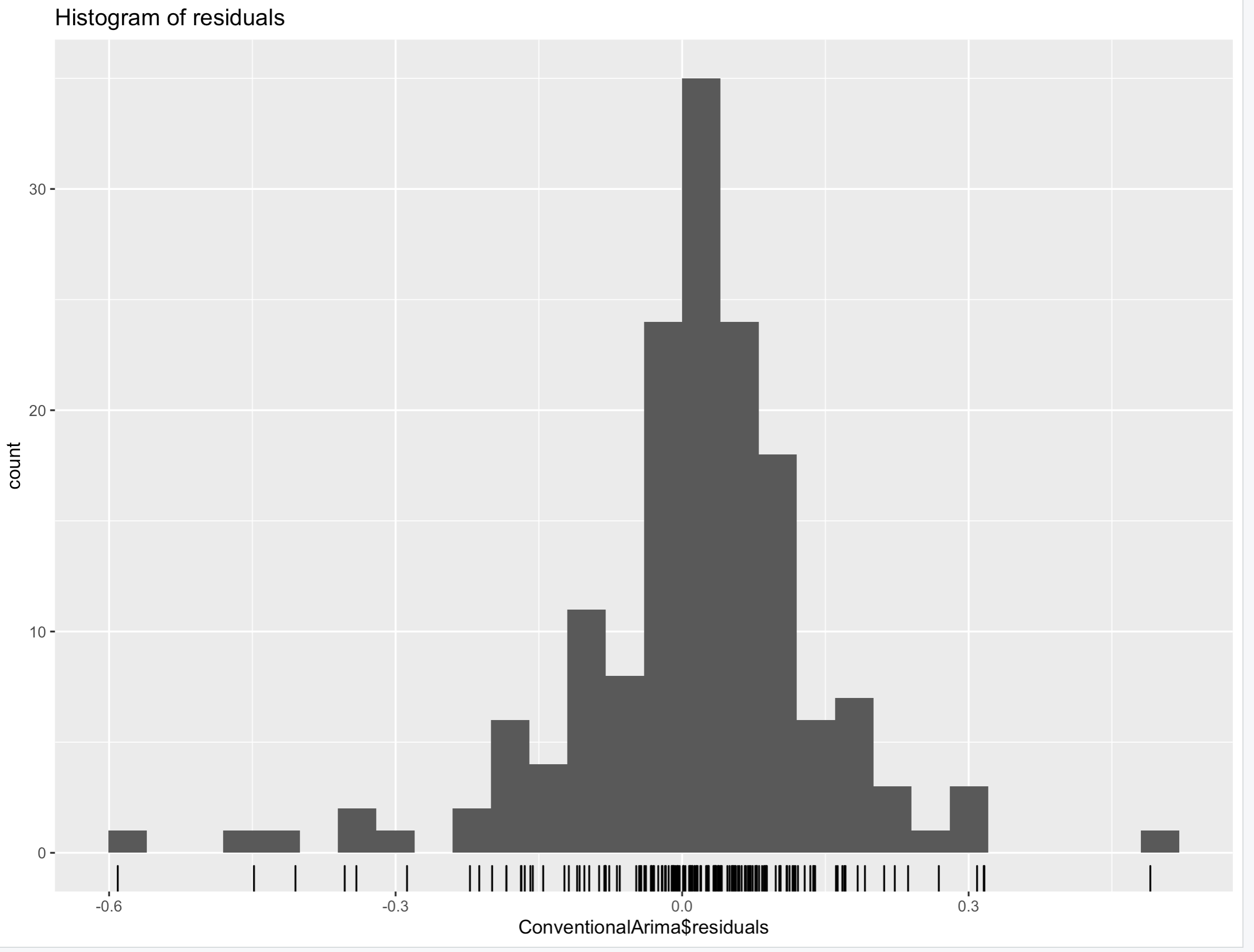
### Histogram of the Residual for type Organic:



### ACF and PACF Plots of Residual for type Conventional:



### Histogram of the residual for type Conventional:



From the residual plots we can see that the final model seems to be a good fit. However, further accuracy of the model can be determined by analyzing it’s performance on the new test data.

# Conclusion:

We have observed several observations while working and analyzing the data. These observations are limited to the avocados in Chicago;

1. As expected, before starting our analysis, we have seen that organic avocados are much more expensive compared to the conventional.
2. Though we have noticed few variations in the patterns between these two types of avocados, we have observed that most of the patterns look similar.
3. As discussed before, year 2017 has seen highest average prices for avocados and was the most volatile year. As in 2017, the average price for both the types of avocados was high we can say that good economy could be the reason for it.
4. Based on our AUTO ARIMA model, we expect a short downward trend in case of conventional avocados and organic avocados looks stagnant for next few weeks.

# Limitations:

* The average price in the table denotes the per unit price of avocado even though there are many units in sold in a bag. So, the price we are taking in consideration is not the actual price but the average of all other pieces in the bag put together. This may lead to a slight inaccuracy in the results.
* We are only considering Hass avocados for our analysis; other varieties are not included in our dataset. Therefore, this analysis will be only limited to Hass type avocados but not for all other varieties.
* The sample size contains 18k records i.e. it is not big enough to get the most accurate results due to variability.
* The data set contains the records starting from 2015 till 2018. As we do not have records before 2015 it is not possible to see if the data has any cyclic patterns and hence it might not be a good idea to perform time series analysis on this data set.

### Limitations of Time Series Analysis

* The various factors that influence the time series may not remain the same for an extended period and so forecasting made on this basis may become unreliable.
* The various factors that affected the fluctuations of price for avocados cannot be fully adjusted by the time series analysis.
* Sometimes the increasing trend in the time series data may be due to the increase in population. So, unless a necessary modification is made to the data it would be difficult to understand the trend.

**References:**

<https://www.kaggle.com/neuromusic/avocado-prices>

<https://en.wikipedia.org/wiki/Avocado>