**Time Series Forecasting Using ARIMA**

Contents

[Introduction 3](#_Toc171459131)

[Data Preparation 3](#_Toc171459132)

[EDA 9](#_Toc171459133)

[Monthly Sales Distribution by Product Family (using boxplots) 10](#_Toc171459134)

[Yearly Sales Distribution by Product Family (using boxplots) 10](#_Toc171459135)

[Impact of on promotions on Sales: 11](#_Toc171459136)

[Decomposition 12](#_Toc171459137)

[Poultry 14](#_Toc171459138)

[FROZEN FOODS 15](#_Toc171459139)

[LIQUOR 16](#_Toc171459140)

[PREPARED FOODS 17](#_Toc171459141)

[MEATS 18](#_Toc171459142)

[Analyzing Residuals and Model Diagnostics 19](#_Toc171459143)

[Comparing All 5 Families: 21](#_Toc171459144)

[BEST TREND, SEASONALITY, RANDOMNESS: 21](#_Toc171459145)

[Effect of Earthquake: 22](#_Toc171459146)

[Structural Break: 22](#_Toc171459147)

[Change point analysis: 23](#_Toc171459148)

[Smoothing 24](#_Toc171459149)

[Poultry 25](#_Toc171459150)

[Frozen Foods 26](#_Toc171459151)

[Liquor, Wine, Beer 27](#_Toc171459152)

[Prepared Foods 28](#_Toc171459153)

[Meats 29](#_Toc171459154)

[Comparison 29](#_Toc171459155)

[Aggregate 30](#_Toc171459156)

[Total sales Per Year 30](#_Toc171459157)

[Poultry 31](#_Toc171459158)

[Frozen Foods 32](#_Toc171459159)

[Liquor Wine Beer 33](#_Toc171459160)

[Prepared Foods 34](#_Toc171459161)

[Checking Seasonality 35](#_Toc171459162)

[Vector Autoregression (VAR) 39](#_Toc171459163)

[Phillips-Perron Unit Root Test- univariate time series 40](#_Toc171459164)

[Regression plots of similar product families: 41](#_Toc171459165)

[Structural Breaks 45](#_Toc171459166)

[Casual Relationships 46](#_Toc171459167)

[Shocks 47](#_Toc171459168)

[Poultry and Meats 47](#_Toc171459169)

[Frozen Foods and Prepared Foods 49](#_Toc171459171)

[Liquor and Poultry 50](#_Toc171459172)

[FEVD 51](#_Toc171459173)

[Relationships: 52](#_Toc171459174)

[Fancharts 53](#_Toc171459175)

[Multivariate Time Series Forecasting 56](#_Toc171459176)

[ACF plots: 56](#_Toc171459177)

[PACF plots: 56](#_Toc171459178)

[Dickey-Fuller test results: 57](#_Toc171459179)

[Conclusion 61](#_Toc171459180)

[Is there a trend? 61](#_Toc171459181)

[Is there a Seasonal effect? 62](#_Toc171459182)

# Introduction

The report examines the grocery sales trends and forecasts for Ecuador's Favorita stores, with a particular emphasis on the following product families: FROZEN FOODS, PREPARED FOODS, MEATS, LIQUOR, and POULTRY. Utilizing advanced machine learning models, such as Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregression (VAR), our analysis predicts sales for the upcoming 15 days and finds patterns in the dataset. Our goal is to provide insights into the sales dynamics of Favorita stores by comprehending these patterns.

Time series data containing important information like store number, product family, sales numbers, and promotional status are included in the dataset. We also take into account important external factors such as the bi-monthly wage payments to the public sector, which might affect consumer purchasing behavior, and the major disruption created by the earthquake in April 2016, which had a big effect on sales patterns. These elements are essential to our study since they guarantee that our models take possible data shifts and abnormalities into account.

Several features of the dataset are covered by our approach, such as trends, seasonal impacts, structural breaks, and causal linkages. Seasonal effects are especially important because they make it easier to spot trends that occur repeatedly over time, such as higher sales around holidays or during special promotions. We look at structural failures, like the earthquake mentioned earlier, to see how they affect sales and modify our models accordingly. To figure out how various elements interact and affect sales data, causal linkages are investigated. Through this comprehensive analysis, we evaluate the predictive performance of our models, ensuring their robustness and accuracy in forecasting future sales.

# Data Preparation

**combined\_sales <- rbind(sales\_sample1, sales\_sample2, sales\_sample3, sales\_sample4, sales\_sample5)**

The first step we took was combining all the categories of product families into one using rbind, the combined dataset is then sorted by date to ensure that the sales data is organized chronologically.

**grouped\_sales <- combined\_sales %>%**

**group\_by(date, family) %>%**

**summarise(total\_sales = sum(sales, na.rm = TRUE), .groups = 'drop')**

We then combined the date and family columns to calculate the total sales for each group.

**wide\_sales <- grouped\_sales %>%**

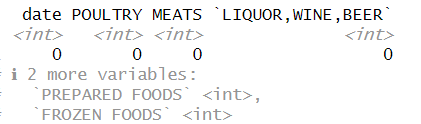
**pivot\_wider(names\_from = family, values\_from = total\_sales)**

Using pivot\_wider to reshape the data from long format to wide format is necessary for statistical analysis, time series analysis, and visualization. The sales data for each product family is separated into a single column in wide format, which makes it easier to apply time series models, create comparative visualizations, and run statistical tests. It is simpler to examine and understand patterns and connections between various product families throughout time with this framework in place.

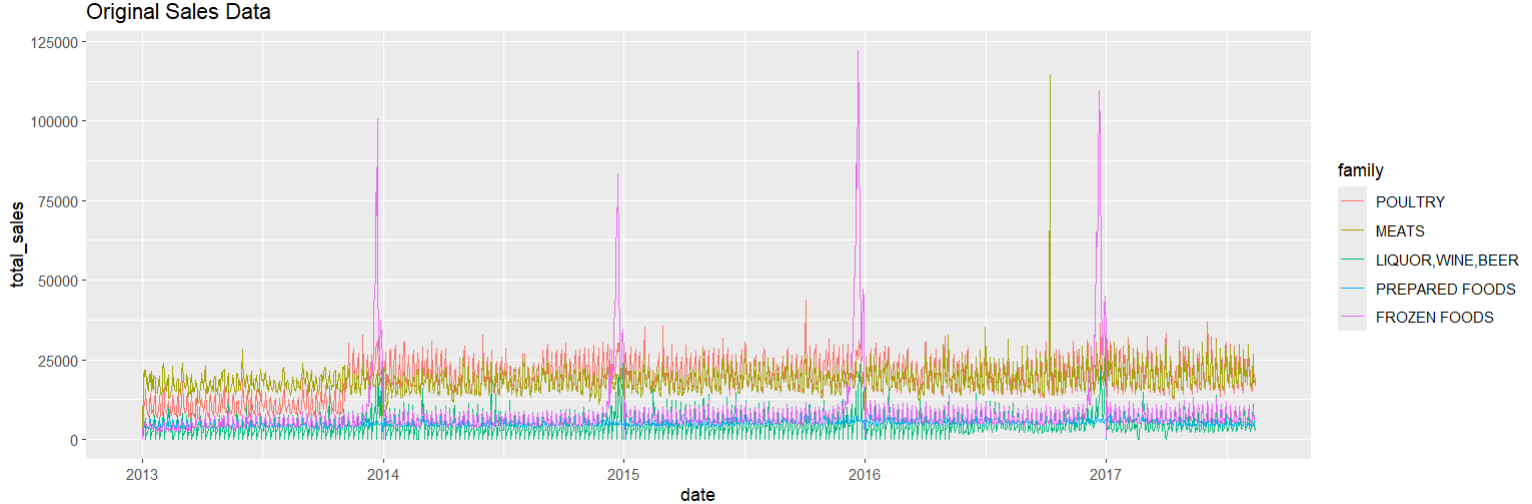
**missing\_values <- wide\_sales %>% summarise\_all(~sum(is.na(.)))**

**print(missing\_values)**

summarise\_all(~sum(is.na(.))) is to be used to check for missing values, and the findings should be printed. This is important because it finds any gaps in the data that may compromise the correctness and dependability of further analysis and modeling. By identifying and assessing the amount of missing data, we can choose the best handling techniques to maintain the validity for our research and the conclusions drawn from it.

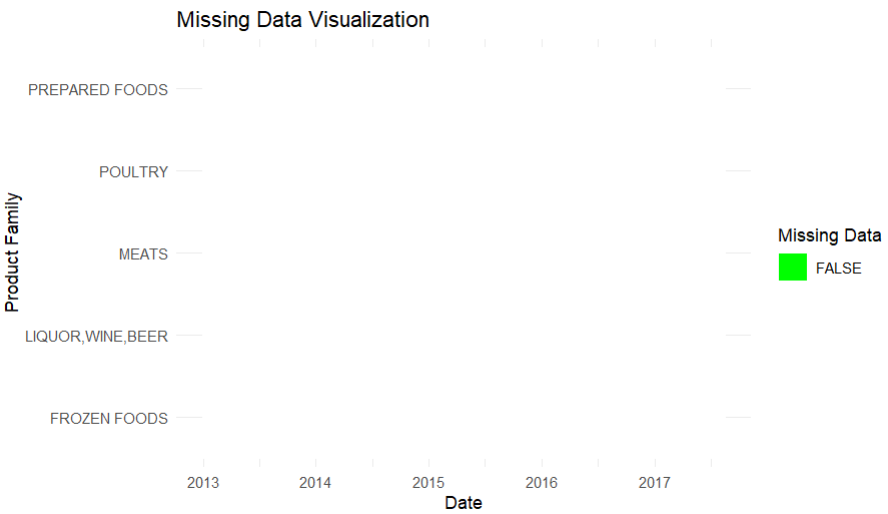


As you can see, there were no missing values indicated



The graph shows the first sales information for different product families between early 2013 and early 2017. The time period is displayed on the X-axis, while total sales up to about 125,000 units are displayed on the Y-axis. A distinct product family is represented by each colored line, and seasonal trends are indicated by distinct, recurrent peaks toward the end of each year, most likely as a result of the holiday seasons. Sales of FROZEN FOODS show notable surges, particularly in 2014, 2015, and 2016, which may indicate seasonal fluctuations or the impact of particular events on sales. Although less evident, MEATS also display peaks, and other families such as POULTRY, LIQUOR, WINE, BEER, and PREPARED FOODS show smaller, more consistent fluctuations. Most families' baseline sales don't clearly show any increase or decrease patterns over time. Instead, they stay steady. Strong seasonality is indicated by the data, particularly for FROZEN FOODS, necessitating the use of seasonally adjusted models in forecasting. The notable increases point to the necessity of anomaly detection in order to identify the root reasons and support promotional and inventory planning.

Using ggplot to plot the original sales data is a crucial step in preprocessing as it facilitates the identification of general trends, seasonal patterns, and anomalies within the data. Significant spikes, seasonal impacts, and comparisons across other product families are shown by visualizing the data, which helps to guide critical preprocessing procedures including handling missing data, treating outliers, and normalizing the data. This preliminary visual evaluation guarantees a thorough comprehension of the behavior of the data, resulting in more precise and customized analysis and modeling in the future.

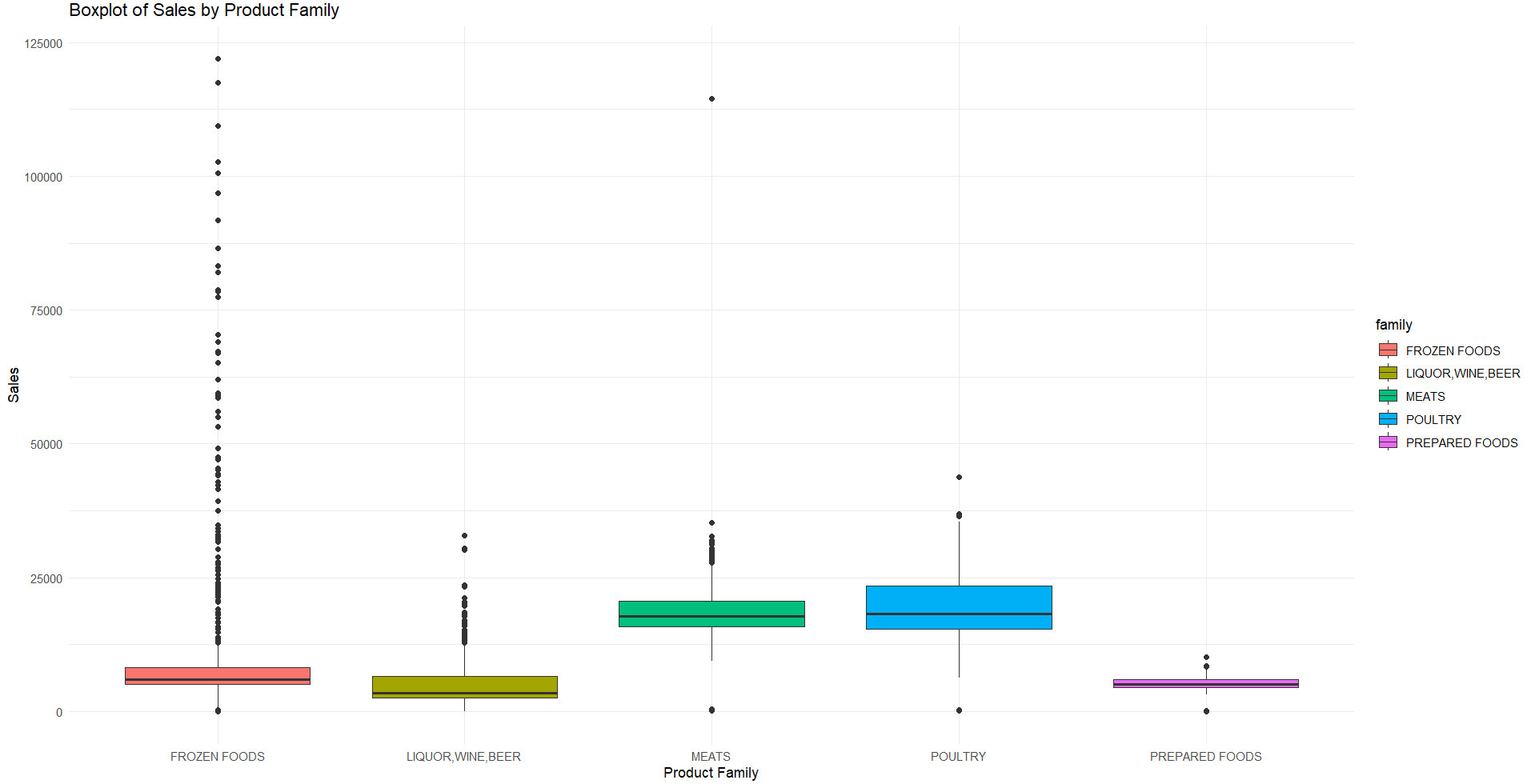
****

This visualizes that there is no missing data in the dataset

**wide\_sales\_long <- wide\_sales %>%**

**pivot\_longer(cols = -date, names\_to = "family", values\_to = "sales")**

Using the pivot\_longer function, the code sample converts the wide\_sales data from wide to long format. Cols = -date causes all columns to be modified, excluding the date column. Two new columns, family and sales, are created to carry the names of the product families and their matching sales values. Because many R functions and plotting packages, such ggplot2, are geared for processing data in long format, this transformation makes data analysis and visualization easier. This makes further analysis more straightforward and versatile.

****

This plot visualizes all of the outliers in the dataset for each product family

Sales data per product family are displayed in a boxplot, with a focus on FROZEN FOODS, LIQUOR, WINE, BEER, MEATS, POULTRY, and PREPARED FOODS highlighted. Sales values are displayed on the Y-axis up to about 125,000 units. The median sales are shown by the middle line, which reflects the interquartile range (IQR) for each box. The dots outside of the whiskers represent outliers, and the whiskers extend up to 1.5 times the IQR. The category with the highest number of outliers is FROZEN FOODS, with sales of up to 125,000 units and a short IQR that suggests substantial variability. While there are a few outliers in LIQUOR, WINE, and BEER as well, they are more concentrated around the median, indicating focused sales. Wider sales variability is indicated by the broader IQRs of MEATS and POULTRY with moderate outliers. With a smaller IQR and fewer outliers, PREPARED FOODS exhibits more stable sales.

**Identifying Outliers:**

**identify\_outliers <- function(data) {**

**q1 <- quantile(data, 0.25)**

**q3 <- quantile(data, 0.75)**

**iqr <- q3 - q1**

**lower\_bound <- q1 - 1.5 \* iqr**

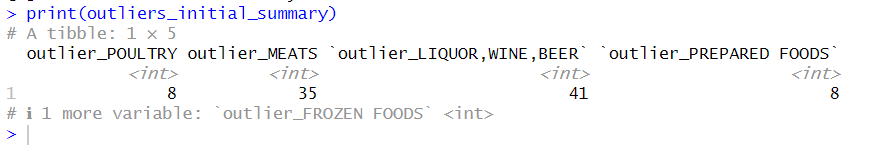
**upper\_bound <- q3 + 1.5 \* iqr**

**outliers <- data < lower\_bound | data > upper\_bound**

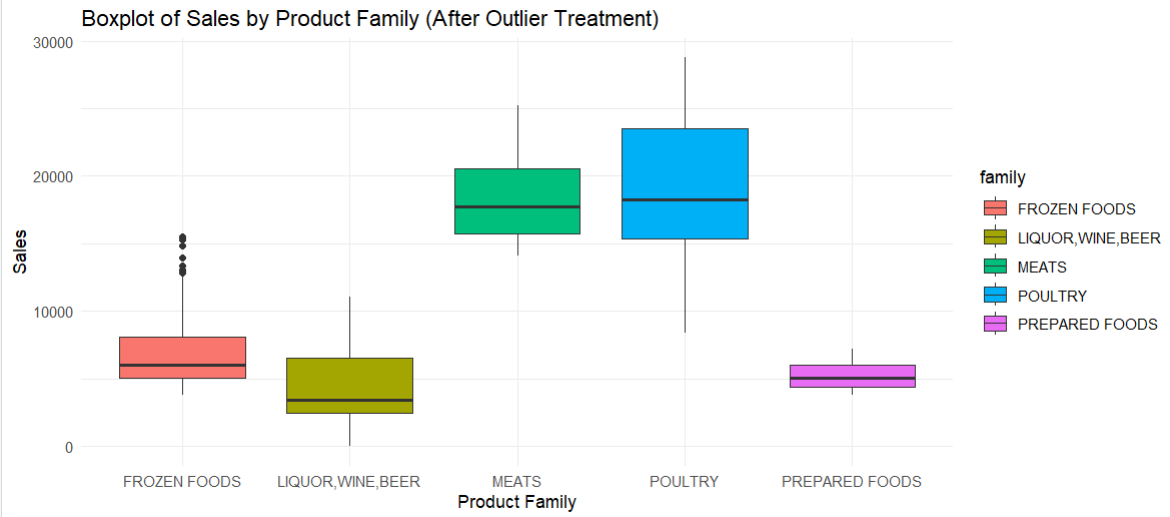
**return(outliers)**

**}**

Based on the IQR approach, the function will produce a logical vector showing which values in the data are outliers. Because of its ease of use and ability to quickly find points that significantly differ from the rest of the information, this approach is frequently used in datasets.

****

There are eight outliers in each category—poultry and prepared foods—showing largely stable sales trends with minimal notable fluctuations. Conversely, there are more outliers in the LIQUOR, WINE, and BEER and MEATS categories—35 and 41 outliers, respectively—which may indicate higher levels of fluctuation and erratic sales trends. Because there are more outliers in these categories, pretreatment handling must be done carefully to guarantee appropriate analysis and modeling. Determining the distribution of outliers among product families facilitates decision-making regarding outlier treatment, including transformation or removal, which improves the forecasting models' resilience.

****

The impact of managing outliers on the dataset is demonstrated by the boxplot graph, which displays sales by product family following outlier treatment. The outlier treatment has decreased extreme values and narrowed the interquartile ranges (IQRs) for FROZEN FOODS, LIQUOR, WINE, BEER, MEATS, POULTRY, and PREPARED FOODS, showing more consistent and predictable sales trends. By making this modification, data dependability for further analysis and modeling is increased, offering more precise insights and reflections of typical sales trends for all product groups.

**normalize <- function(x) {**

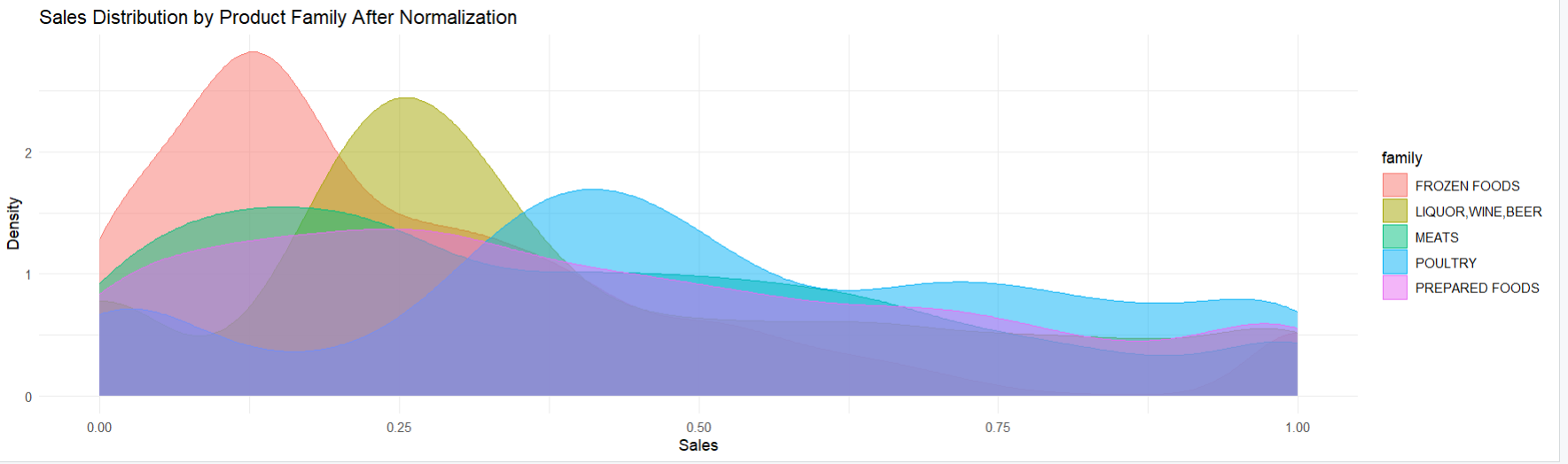
**return((x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)))**

**}**

**wide\_sales <- wide\_sales %>%**

**mutate(across(where(is.numeric), normalize))**

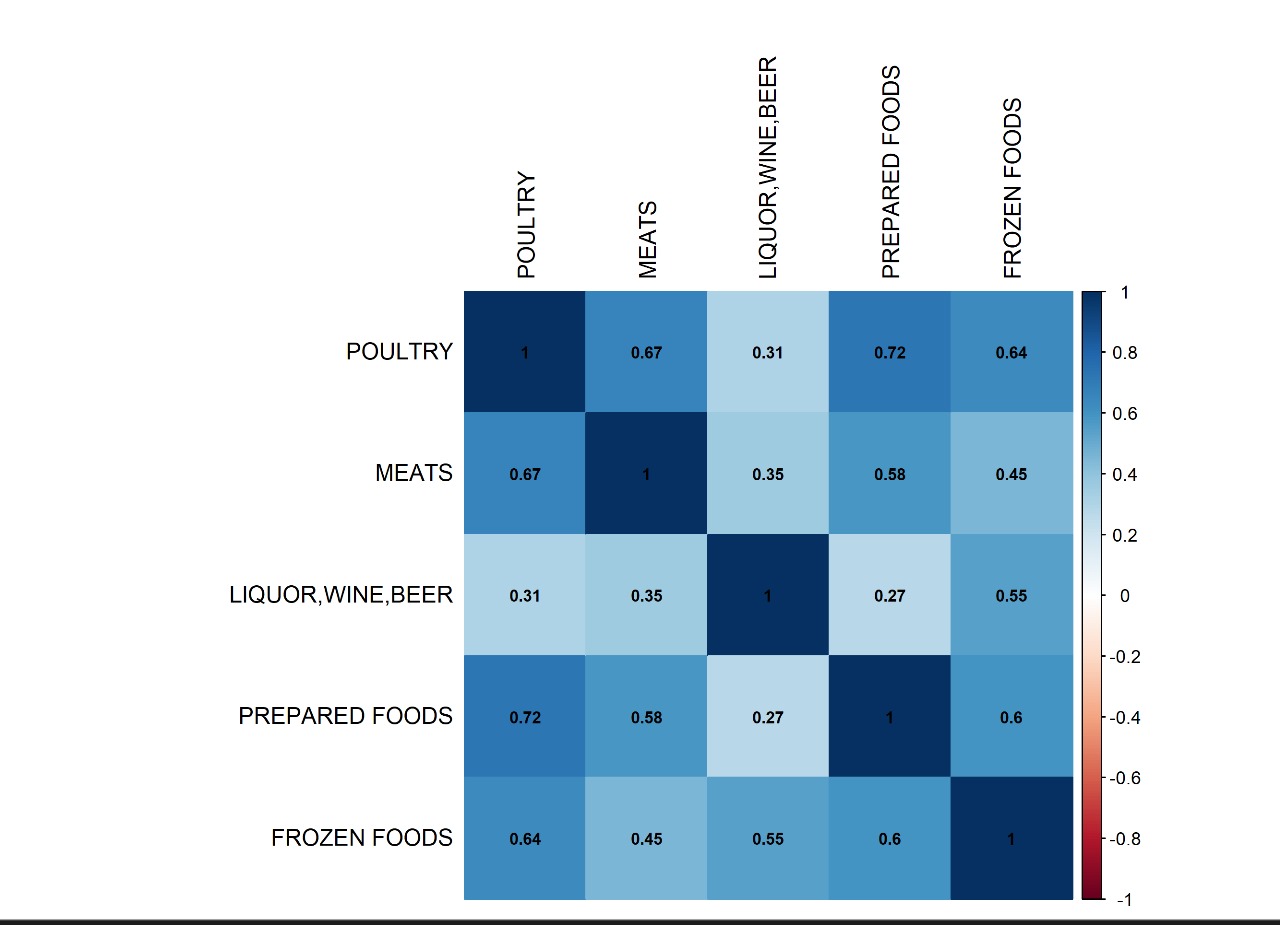
The normalize function uses min-max normalization, which bases each value transformation on the lowest and greatest values in the dataset, to scale the data to a range between 0 and 1. It is imperative for data preparation that all numerical data be on a comparable scale, which is ensured by applying this function. This stage is crucial for this dataset because it harmonizes sales figures from various product lines, making comparisons, visualizations, and further analysis—like machine learning algorithms, which depend on normalized input for best results—more precise and insightful.



After normalization, sales values are scaled between 0 and 1, allowing for direct comparison across families. The normalized sales are represented by the X-axis, while the density of sales values is represented by the Y-axis. Peak density for FROZEN FOODS (pink) is close to 0.1, which suggests that most sales are low. The yellow LIQUOR, WINE, and BEER peak is around 0.2, indicating a somewhat higher concentration of poor sales. The distributions of POULTRY (blue) and MEATS (green) are wider, with POULTRY peaking at 0.3 and MEATS at 0.25, respectively, suggesting more sales volatility. The purple PREPARED FOODS has a very high density overall, especially above 0.25, which indicates a more even dispersion of sales.

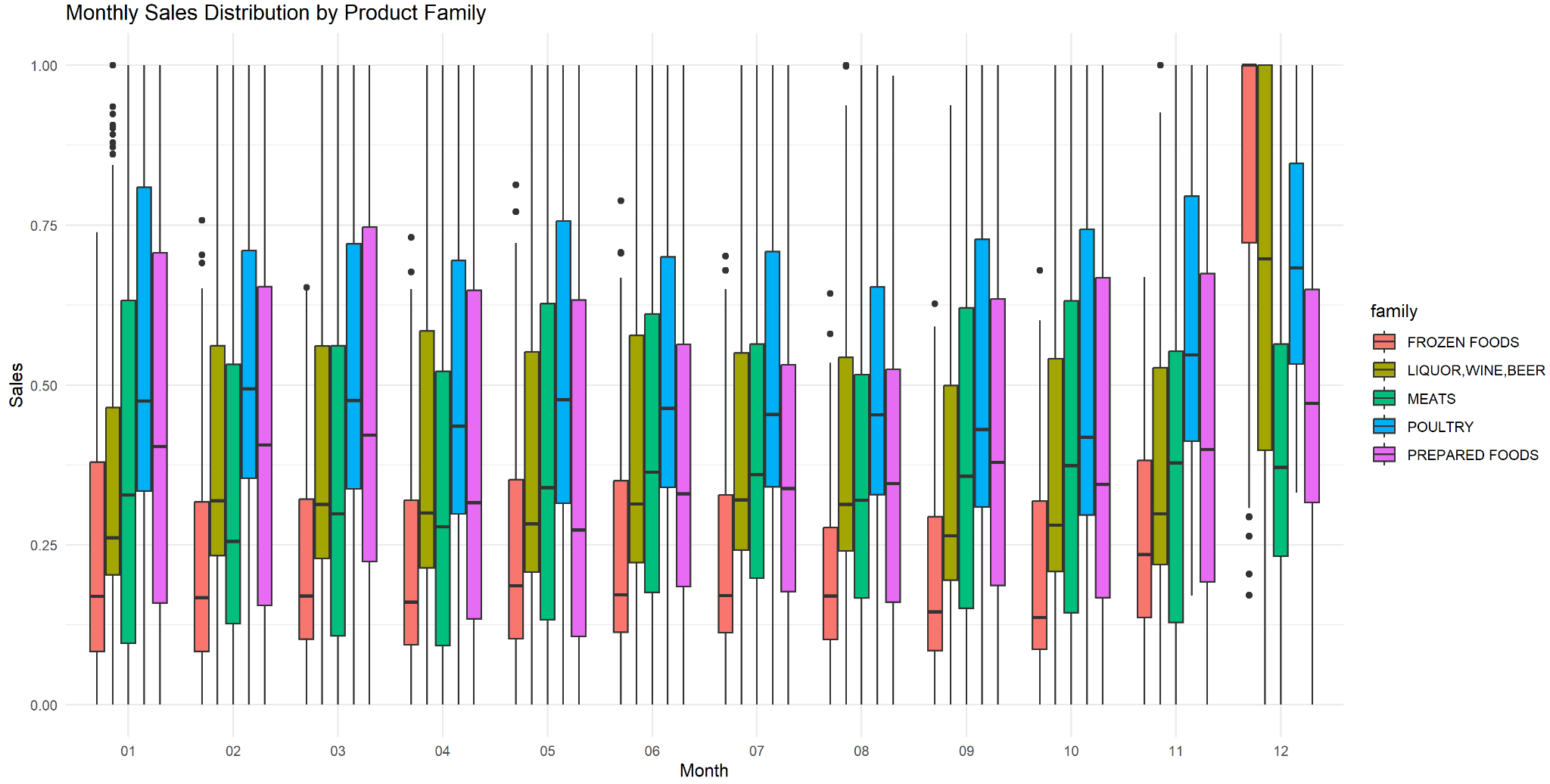
# EDA

Exploratory Data Analysis (EDA) is an approach to analyzing data sets in order to summarize their main characteristics using many different visual methods. It is a crucial step in our project because it helps us understand the data, detect anomalies, find patterns, test hypotheses, and check assumptions through graphical representations and statistical techniques. Throughout our entire code, we have implemented EDA to gain insights into our sales data for various product families. This includes steps such as normalization, decomposition, residual analysis, and structural break detection. We have many different EDA’s in our code. Below, we will show some examples of EDA.

****

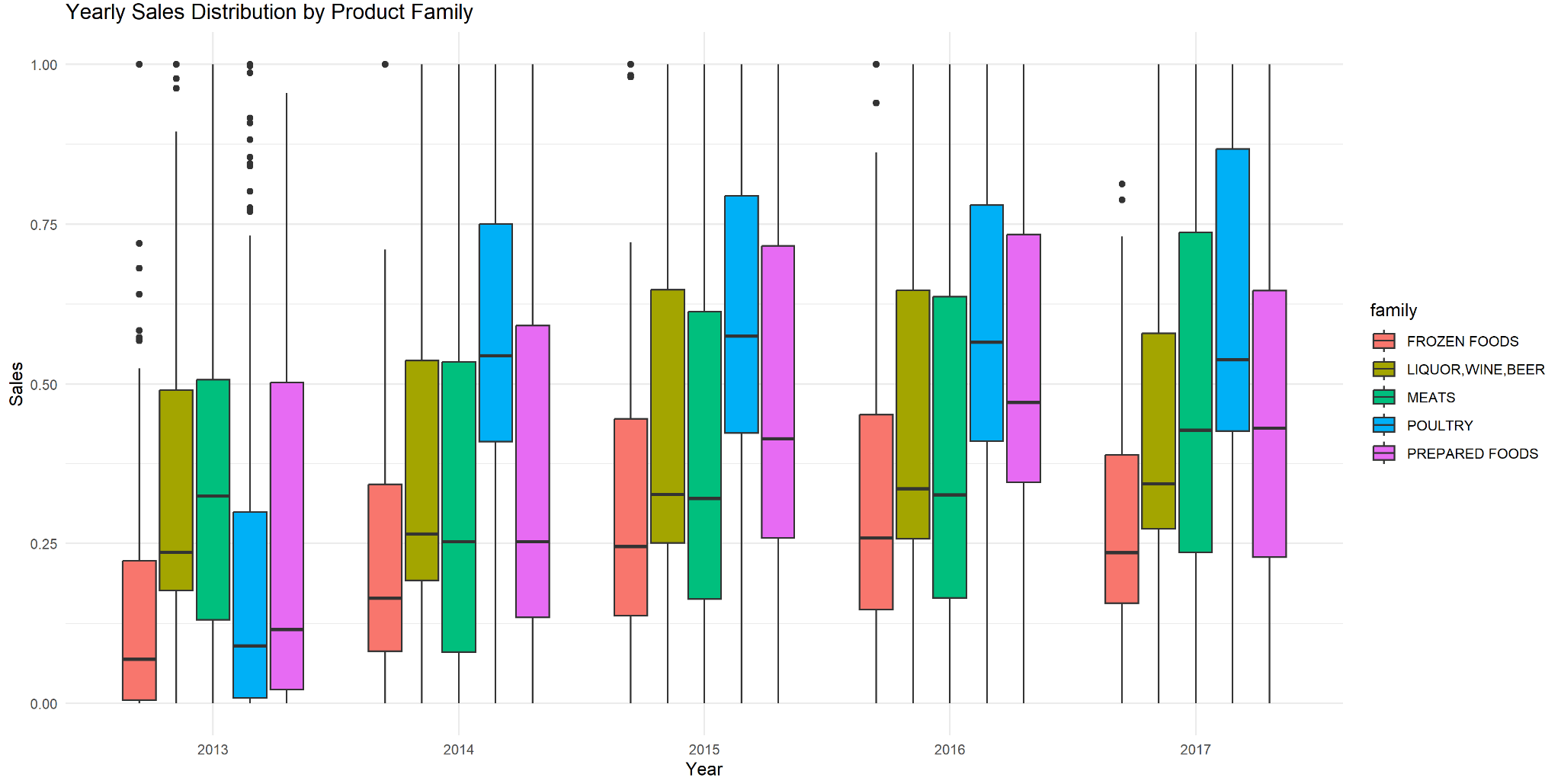
**Anomaly Detection Correlation Matrix**

|  |  |
| --- | --- |
| Anomaly detection plot shows significant sales anomalies (blue points), especially in Liquor, Wine, and Beer, indicating unexpected sales spikes. Anomalies are scattered across all product families. | The correlation matrix shows that Poultry and Prepared Foods have highest positive correlation (1 and 0.72), indicating their sales trends move similarly |

****

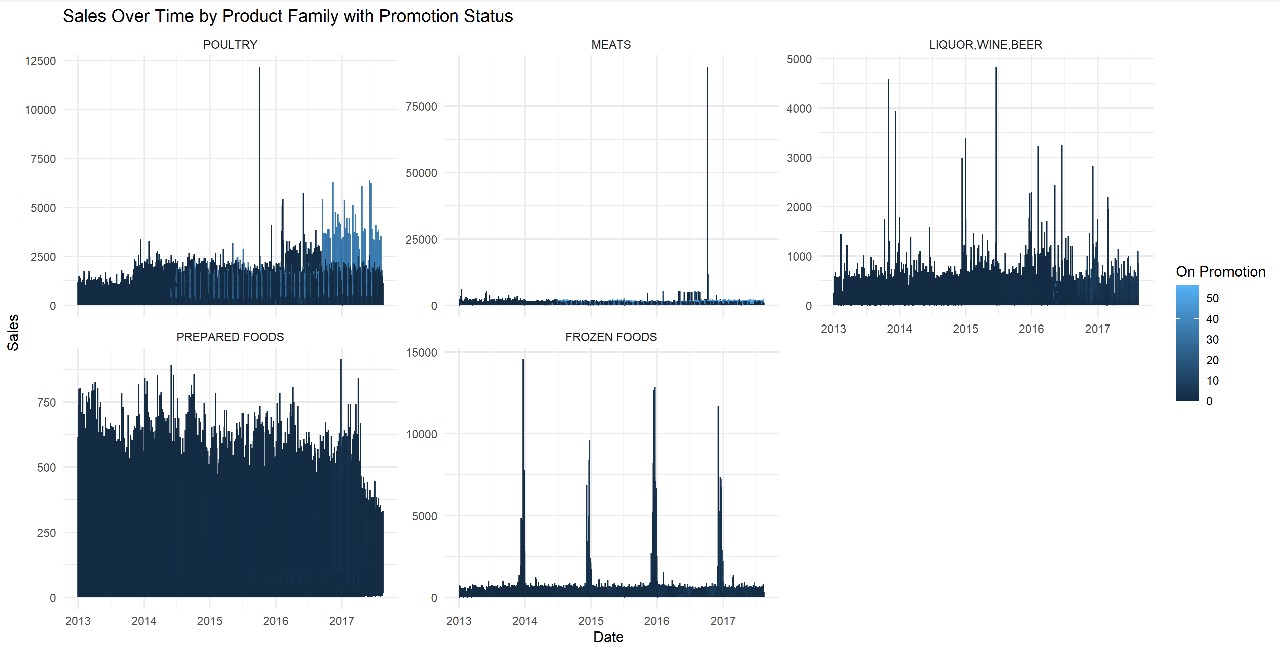
### Monthly Sales Distribution by Product Family (using boxplots)

The boxplot illustrates the monthly sales distribution for different product families targeting consistent sales patterns and variability. We can infer from the plot that "Poultry" and "Prepared Foods" show higher median sales throughout the months, with "Frozen Foods" and "Liquor, Wine, Beer" exhibiting more significant fluctuations and outliers.

****

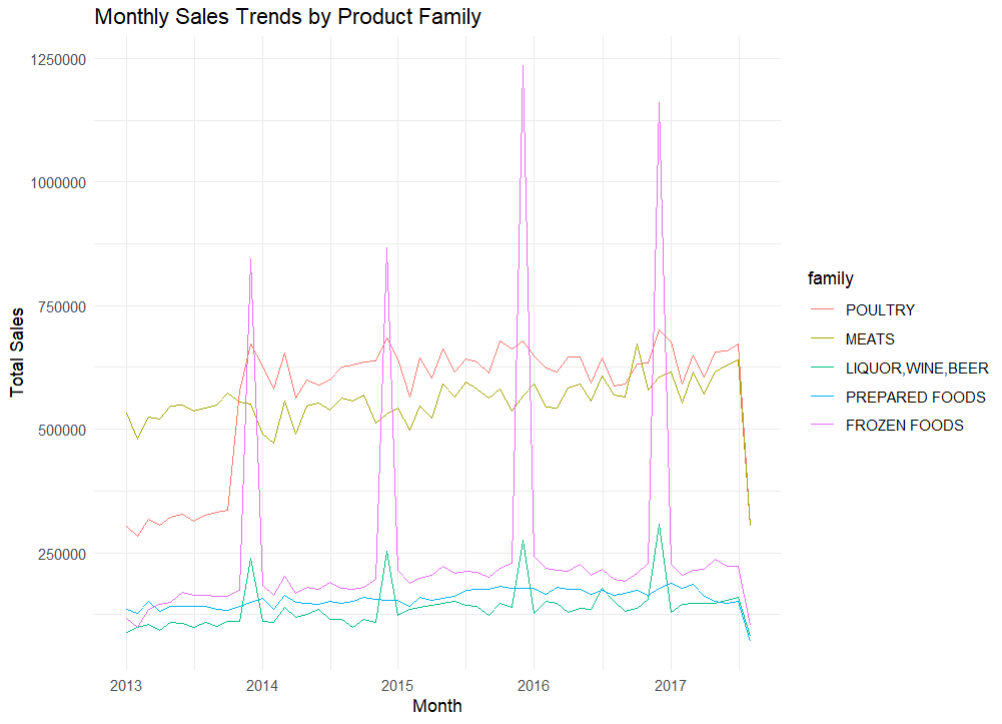
### Yearly Sales Distribution by Product Family (using boxplots)

The boxplot shows the yearly sales distribution for different product families from 2013 to 2017. It reveals significant variability in sales between different families and years, with "Prepared Foods" and "Poultry" showing consistently higher sales, while "Frozen Foods" and "Liquor, Wine, Beer" show more fluctuation and a higher presence of outliers.

****

### Impact of on promotions on Sales:

The plot shows sales trends over time for different product families, with varying shades indicating the number of promotions. Max points in sales are shown the most around periods of high promotion activity.

****

The monthly sales trends show that Frozen Foods experience the highest sales spikes, especially around holiday seasons. Poultry, Meats, and Liquor have stable sales trends, with Prepared Foods showing notable increases during peak periods.

# Decomposition

Decomposing is a time series process where the time series is broken into 3 major components: trend component, seasonal component and irregular component (residual). This process helped us understand the hidden patterns and structure within our data leading to more accurate forecasting analysis. The two main decomposition types that we focused on were both additive and multiplicative. In additive decomposition it sums the three time series components and displays it in the upper box which is the observed data. It is best used when the seasonal

variations are constant over time. As for Multiplicative decomposition it is used as the product of its components, where the observed data is = T (trend) x S (seasonality) x R (residuals). It is very useful when the seasonal variations increase or decrease in line with the level of time series. Both Types of decomposition were applied to the total sales and to each product family separately to have a comprehensive understanding.

**Total Sales**

|  |
| --- |
|  |

The decomposition plots consist of four panels: Observed, Trend, Seasonal, and Random (Residual). The total sales data from 2013 to 2017 is shown in the observed panel, with notable spikes and periodic patterns that point to significant seasonal variations. The trend component displays a clear upward trend progression in sles from 2013 to 2016. It stabilizes towards the end which tells us that the growth is likely influenced by external factors. Seasonality component shows patterns at regular intervals which is the biggest hint that there is seasonality in the plot. The random component tells us that there are irregular fluctuations not explained by the trend or seasonal components.

Factors affecting decomposition plots:

The bi-weekly wage payments ( on the 15th and last day of the month) significantly affect the seasonal component, with sales rising consistently around these dates. The earthquake on April 16, 2016, led to notable spikes in the residual component, which indicates an abnormal increase in sales of essential items. As shown above, both additive and multiplicative decompositions reveal similar patterns, suggesting that either model can be used for this data. The observed seasonality is impacted by wage payment cycles and holiday periods. The trend shows an overall growth stabilizing in later years. The residual component highlights significant anomalies around the earthquake period which focuses on the importance of considering these external factors in sales forecasting.

## 

## Poultry

|  |  |
| --- | --- |
| **Additive Decomposition** | **Multiplicative Decomposition** |

For the decomposition of family “POULTRY” the observed panel displays the sales that lasted from 2013 to 2017, conveying periodic patterns and variations. The trend component captures an upward trend in sales from 2013 to 2015. This was followed by a stabilized period which suggested that there is an overall growth in POULTRY sales. The seasonal component reveals consistent patterns which tells us that there is seasonality. The random component captures irregular fluctuations (randomness) which were not explained by the trend or seasonal components. The spikes are noticeable the most around March-April of every single year between 2014-2017 .

The payments issue (15th and last day of the month) have impacted the seasonal component, with sales reaching the highest point consistently around these dates. The earthquake in April, 2016, caused obvious spikes in the residual component. This indicates an abnormal increase in sales of essential items at that time. The observed seasonality is affected the most by wage payment cycles and holiday periods. There is an obvious trend showing overall growth which keeps stabilizing in later years. The residual component focuses on the irregularities like the earthquake that took place.

## FROZEN FOODS

|  |  |
| --- | --- |
| **Additive Decomposition** | **Multiplicative Decomposition** |

The decomposition plots for “FROZEN FOODS”once again are made of four panels: Observed, Trend, Seasonal, and Random (Residual). The observed panel showing “FROZEN FOODS” sales data from 2013 to 2017, which highlights for us the significant periodic spikes. The trend component here captures a gradual upward progression in sales from 2013 to 2015. It goes stable for a while and then increases then stabilizes one more time. We can see that the seasonal component reveals strong, consistent patterns occurring at regular intervals. The random component here just like all the others captures irregular fluctuations which were not explained by the trend or seasonal components

The payments given every 2 weeks (15th and last day of the month) impact the seasonal component the most with sales repeating the same pattern consistently around these dates. The earthquake on April 16, 2016, affected the residual component because it is the one responsible for any unusual behavior. Nevertheless, this indicates a huge increase in sales of essential items. Both additive and multiplicative decompositions reveal similar patterns, and this tells us that we can use any of them.

## LIQUOR

|  |  |
| --- | --- |
| **Additive Decomposition** | **Multiplicative Decomposition** |

The above plots show the decomposition for LIQUOR, WINE, BEER sales . The observed panel shows data from 2013 to 2017 summed and added. The trend component shows an upward trend from the time period 2013 to 2016 which is followed by a flat stabilized period. The seasonal component reveals strong, repeating patterns as a result of the bi-weekly wage payments and holiday periods. The random component captures irregular fluctuations which were skipped by the trend or seasonal components, with pronounced spikes. It is safe to say that there is randomness in the data.

The wages paid (15th and last day of the month) impacted the seasonal component, with sales going crazy consistently around these dates. The earthquake on April 16, 2016, and some other subsequent relief efforts caused notable up spikes in the residual component. The observed seasonality is driven by wage payment cycles and holiday periods, the trend shows overall growth stabilizing in later years. The residual component was triggered and has shown a massive change particularly around the earthquake period.

## PREPARED FOODS

|  |  |
| --- | --- |
| Additive Decomposition: | Multiplicative Decomposition: |

As for “ PREPARED FOODS” sales we have done both decompositions as well. The observed panel displays the raw sales from 2013 to 2017. The trend component clearly shows an upward trend in sales from 2013 to 2015, followed by a stabilization period and then a slight decline towards the end. This indicates that there is an overall growth in sales but with some fluctuations. The seasonal component shows patterns because of the major external factors acting on it. The random component captures randomness not explained by the trend or seasonality.

The payments paid ( on the 15th and last day of the month) impacted the seasonal component a lot during these dates. The sudden earthquake on April 16, 2016 caused an abnormal increase in sales of essential items that were needed when it all happened. The trend indicates overall growth stabilizing and then slightly declining in later years, and the residual component highlights significant anomalies, especially around the earthquake period. This points out the significance of taking such external factors into account in sales forecasting.

## MEATS

|  |  |
| --- | --- |
| Additive Decomposition: | Multiplicative Decomposition: |

The decomposition plots for MEATS sales show both additive and multiplicative decompositions which look exactly the same. The first panel displays the sales data from 2013 to 2017 with periodic patterns and variations. For The trend component there is an upward trend in sales almost throughout the entire time indicating consistent increase in sales. The seasonal component shows patterns that keep occurring at some point in every year due to some external factors. The random component focuses on the irregular fluctuations not explained by the trend or seasonal components.

Factors

The bi-weekly wage payments (15th and last day of the month) affected seasonal components. As we can see there is a pattern in the sales consistently happening around these dates. The abnormal increase in sales can be attributed to the earthquake that occurred on April 16, 2016. The trend indicates that overall growth stabilizes in later years, the residual component emphasizes significant anomalies, especially around the earthquake period, and the observed seasonality is driven by wage payment cycles and holiday periods.

## Analyzing Residuals and Model Diagnostics

|  |  |
| --- | --- |
| **Total Sales:**  The model has captured the main trends and seasonal patterns in the data. The residual plot shows fluctuations around zero which tells us the model has accounted for the primary key components of the time series. Even though there are some noticeable spikes around April 2016 this means some irregular fluctuations remain. The ACF plot (bottom left) shows that most of the residuals' correlations are within the confidence limits, indicating that the residuals are mostly not related to each other. However, there are some significant correlations at certain lags. These point to patterns that the model hasn't fully explained.The histogram (bottom right) of residuals shows an approximate normal distribution centered around zero. The density plot confirms this distribution, even though we see some deviations present, mainly in the tails. |  |
| **Poultry**:  The residual plot highlights some spikes, particularly around April 2016, indicating that certain irregularities remain. The ACF plot is showing that most residuals are within the expected range, meaning they are mostly uncorrelated. The histogram indicates that the residuals are approximately normally distributed, centered around zero, although there are some deviations in the tails. Overall, while the model performs well in getting the data's main features, some residual irregularities and correlations suggest room for improvement, particularly considering external events like the earthquake. |  |
| **FROZEN FOODS:**  The residual plot displays oscillations close to zero, but it also exhibits notable spikes, particularly around April 2016, which point to atypical patterns that the model is unable to fully explain. The ACF plot shows that most residuals' correlations fall within the confidence bounds, suggesting that they are mostly not similar, although some significant correlations at specific lags suggest unexplained patterns. In the histogram it is shown that residuals are approximately normally distributed around zero, with some deviations. |  |
| **LIQUOR:**  The model catches most of the data's key patterns. The histogram shows that the residuals are roughly normally distributed around zero, with some deviations at the tails. The ACF plot indicates that most residual correlations fall within the confidence limits, meaning they are mostly uncorrelated. A few lags show significant correlations which point to some remaining unexplained patterns. The residual plot shows that the residuals generally fluctuate around zero, but there are noticeable spikes, particularly around April 2016, highlighting irregular patterns not fully captured by the model. Overall, while the model effectively captures the primary trends and seasonal effects, some irregularities and correlations, especially around the earthquake period, suggest areas for further refinement. |  |
| **PREPARED FOODS:**  The histogram shows that the residuals are mostly normally distributed around zero, although there are some deviations (found in the tails). The ACF plot suggests that most residual correlations are mostly uncorrelated because they are within the confidence bounds; however, some significant correlations at particular lags suggest that there are still some patterns that the model isn't fully able to explain. The residual plot shows that the residuals generally fluctuate around zero, with noticeable spikes, particularly around April 2016, highlighting irregular patterns not fully captured by the model. |  |
| **MEATS:**  The ACF plot indicates that the residuals are essentially uncorrelated because most of their correlations are within the confidence bounds. However, some significant correlations at particular lags point to residual patterns that the model may not have been able to fully explain. The residuals are roughly normally distributed around zero, with some deviations, particularly in the tails, according to the histogram. The residual plot shows that while the residuals mostly oscillate around zero, there are some discernible spikes, especially around April 2016, which highlight atypical patterns that the model was unable to fully capture. |  |

## Comparing All 5 Families:

The decomposition of the five product families: “POULTRY”, “FROZEN FOODS”, “LIQUOR”, “PREPARED FOODS” and “MEATS” reveals both unique characteristics and interconnected responses to external factors. All families show an upward trend which indicates an overall growth in sales over time. Seasonal patterns are noticeable across all families caused by wage payments (on the 15th and last day of each month) and the excessive spending during holiday periods. The seasonal patterns for all families were regular and consistent, illustrating the influence of wage cycles and purchasing patterns across various product categories. The earthquake on April 16, 2016, had a distinct impact on all families, with significant spikes in the residual component (random). This event highlights how external shocks can uniformly change sales patterns across numerous product categories. Even though each family's sales dynamics are different, they are linked by how they react to major disruptive events like earthquakes and cyclical wage payments. This shows how these external factors influence the market in a cohesive way.

## BEST TREND, SEASONALITY, RANDOMNESS:

Among the five different families in our data, it is safe to say that “LIQUOR” has the best trend. Just by looking at it, we can see a strong and a consistent upward progression from 2013 to 2016 which only stabilizes towards the end of the period. As for seasonality, the best one we observed was “FROZEN FOODS”.It exhibits clear and regular seasonal patterns which are driven by salary payments and holiday periods. When considering randomness, “MEATS ” has the most residual fluctuations, which simply indicate higher variability in data that was not explained by the trend or seasonal components. Hence, we all agreed that “MEATS” sales are more influenced by irregular factors (earthquake) making it the most unpredictable family compared to the others.

## Effect of Earthquake:

**A graph of a earthquake

Description automatically generated with medium confidence**

The plot shows that the earthquake in April 2016 (displayed by the red dashed line) had a noticeable impact on sales across all product families. Sales initially declined around the earthquake period, particularly for Frozen Foods and Meats. However, in the months following the earthquake, there was a significant increase in sales across most product families, indicating a recovery and possible increased demand for certain goods. This shows that the earthquake briefly disrupted sales, but they later increased as things got back to normal.

## Structural Break:

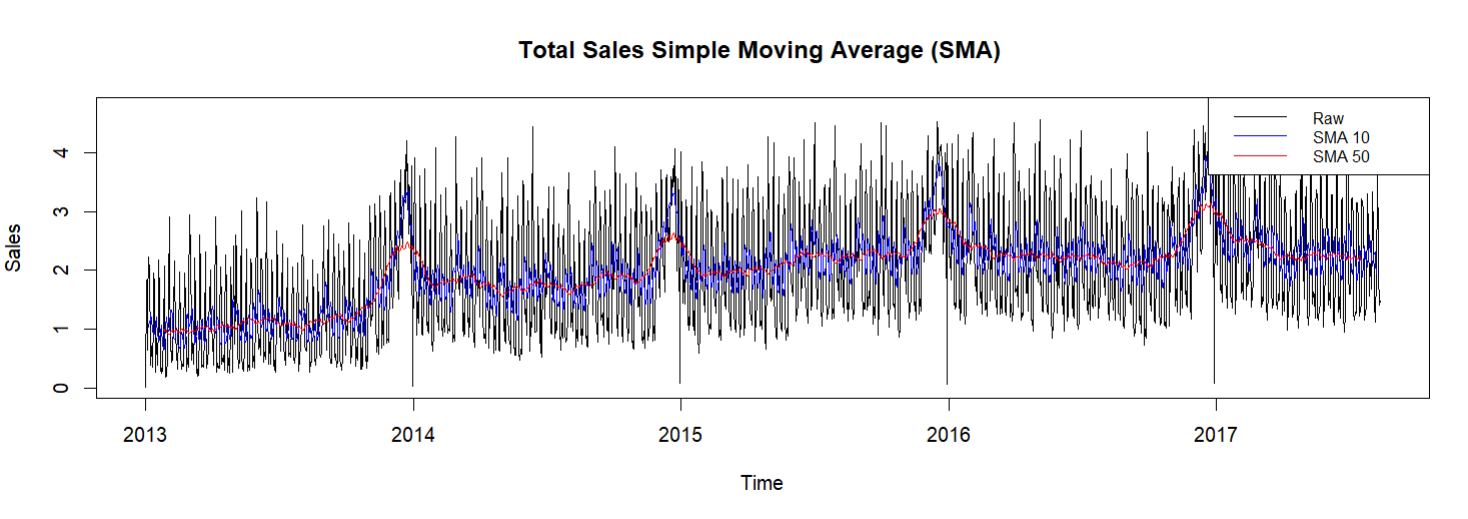
Structural breaks occur when there is a sudden change in a time series due to external events or significant disruptions in the real world. In our data, there is a clear indication of a structural break around April 2016, when a significant earthquake struck Ecuador. This event led to noticeable spikes, patterns and irregularities in the sales data across all product families, as observed in the residual plots. These spikes are evident in both the residual plots and the ACF plots, where significant correlations appear at certain lags. The structural break caused by the earthquake disrupted the usual sales patterns, resulting in abnormal increases in sales of essentials. This disruption highlights the importance of considering external shocks and structural breaks in time series analysis, as they can significantly either positively or negatively affect the data and the accuracy of forecasting models. In our case it's a mixture of both because people had to buy essentials due to the emergency.

# Change point analysis:

|  |  |
| --- | --- |
|  |  |
| The change point analysis for POULTRY sales from 2013 to 2017 shows significant shifts in sales levels. Notable changes occur around early 2014, with a massive increase, and another around 2016. The red lines highlight these shifts, indicating periods of stable sales interspersed with sudden changes. This underscores the need to account for structural breaks in sales forecasting. | The change point analysis for FROZEN FOODS sales shows significant shifts around major spikes, indicating seasonal peaks or events. The red lines highlight stable periods between these spikes. This analysis underscores the impact of seasonal changes on sales and the need to consider them in forecasting. |
|  |  |
| Liquor: Same thing is happening over and over in every change point plot. The graph is more compact because the sales volume is higher here for this family. | The change point model shows significant shifts in sales levels, with noticeable changes around 2015 and 2016.. This plot is more densely packed with data points compared to others, indicating frequent changes and higher sales volumes. |
|  | The change point analysis for MEATS sales shows a few significant shifts, with major spikes around 2015 and 2016. Nothing different is happening but the sudden changes happening are because of the external factors acting on our datasets. |

# Smoothing

**Total Sales:**

****

The graph uses Simple Moving Averages (SMA) to improve trend visualization and shows the cumulative sales of a whole product family over time. The normalized sales numbers are represented by the Y-axis, while the period span from 2013 to 2017 is represented by the X-axis.

**Raw Sales Data (Black Line)**: This is a highly volatile line that displays the daily sales data, which fluctuates a lot. It does not use any smoothing to represent the actual sales figures.

**SMA 10 (Red Line):** The Simple Moving Average for a 10-day period is shown by this line. It evens out short-term swings and draws attention to mid-range patterns in the sales data.

**SMA 50 (Blue Line):** The Simple Moving Average for a 50-day period is displayed on this line. It offers a more noticeable smoothing effect and makes seasonal patterns and longer-term trends in the sales data visible.

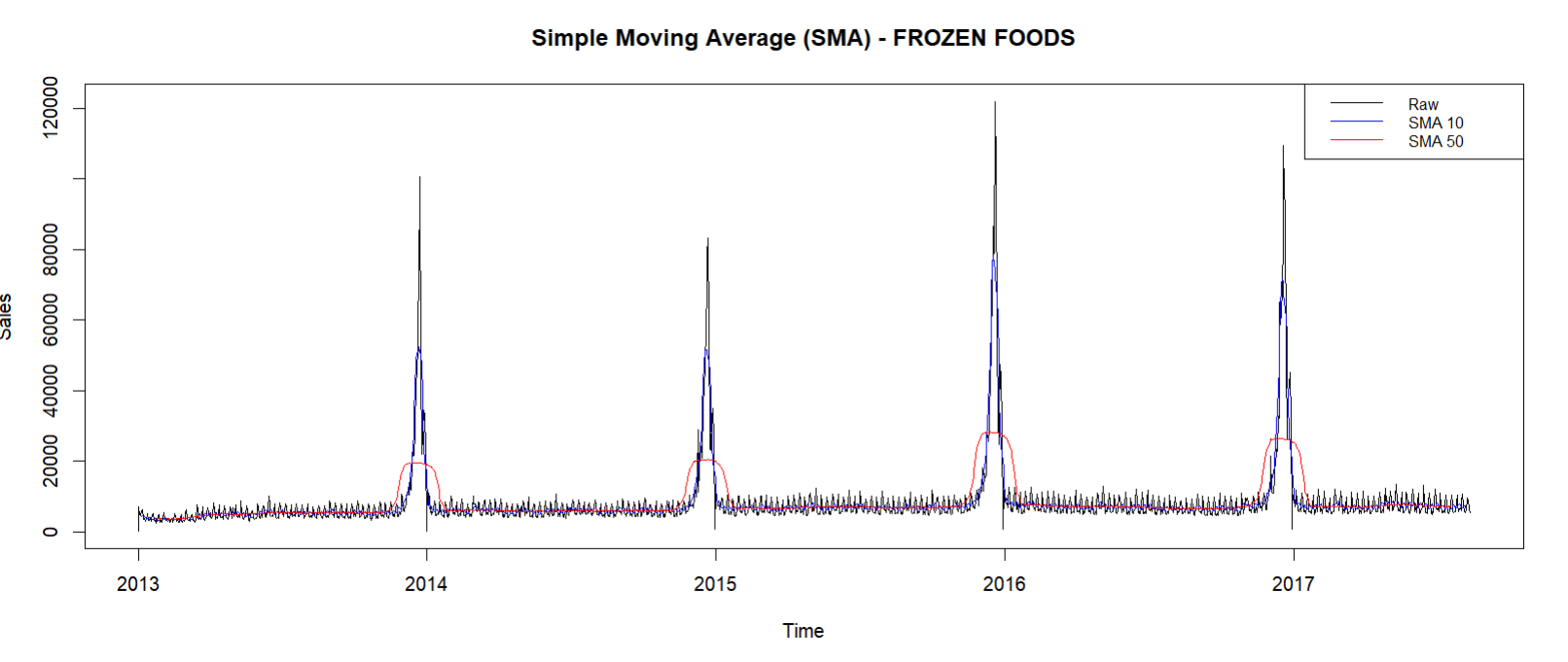
Because of daily volatility, the SMA aids in revealing underlying trends that are not immediately visible in the raw data. While the SMA 50 shows seasonality and longer-term trends, the SMA 10 emphasizes short- to medium-term trends. The SMA reduces the noise from random daily fluctuations, making it easier to observe and analyze consistent patterns and trends. When looking at the annual seasonality, strong seasonality in sales is shown by the graph's recurrent peaks and troughs at regular intervals. These peaks probably line up with particular seasons of the year, such the holidays, when sales usually rise. We understand that some of the regular, smaller surges in sales may coincide with the bi-weekly wage payments made to public sector employees on the 15th and last day of each month. When daily sales fluctuations are subtracted, longer-term trends in the data are revealed by the 50-day SMA (red line). The graph indicates that sales generally increased between 2013 and 2014, then stabilized for a while before slightly declining around 2016. The peak around the middle of 2016 is very noteworthy. Following a period of high volatility while the region recovered, the magnitude 7.8 earthquake on April 16, 2016, probably prompted a major jump in sales due to increased demand for necessary commodities.

## Poultry

****

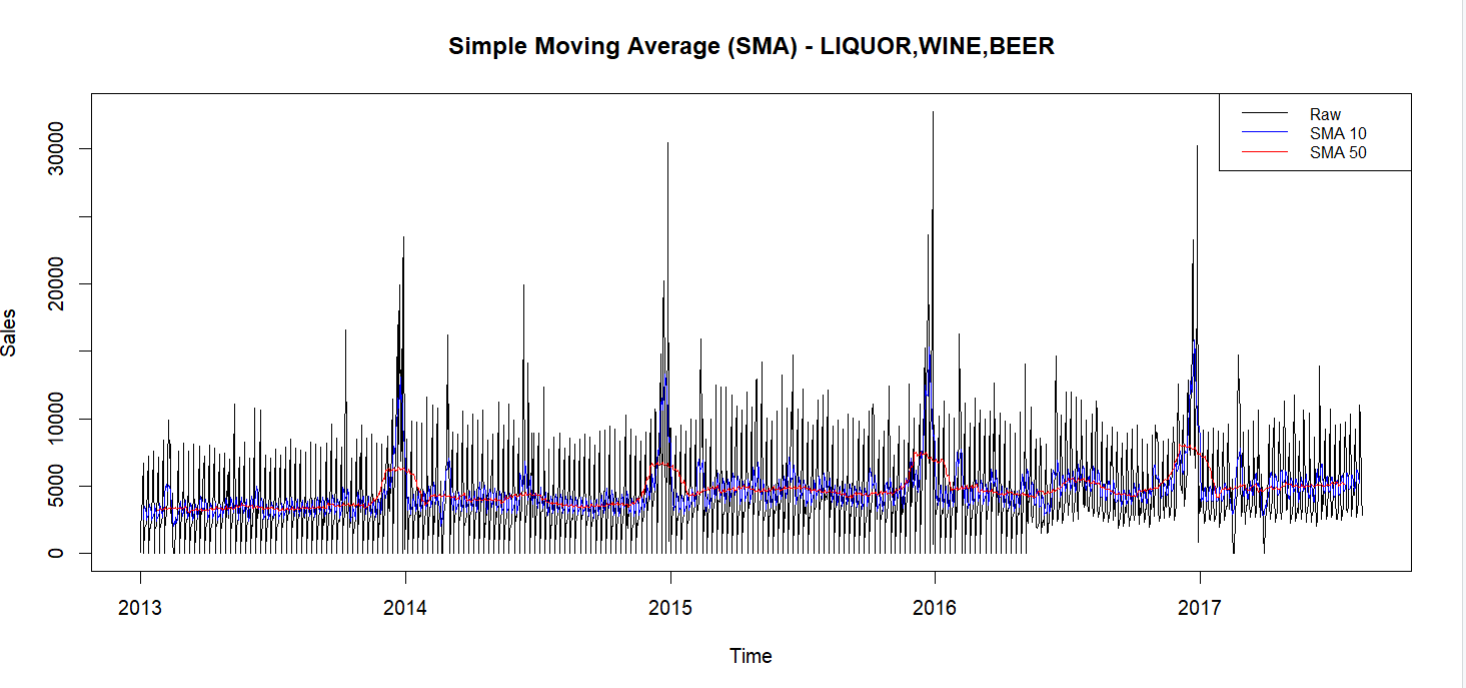
There is a sharp dramatic increase in sales from 2013 to 2014. There is a noticeable rise in sales from late 2013 to early 2014. There could be several reasons for this, including the success of marketing initiatives, seasonal promotions, or an overall rise in the demand for poultry at that time among consumers. It might also be a reflection of increased product availability or supply chain efficiency. After 2014, there were stabilized sales with regular fluctuations. The SMA lines show this steadiness, indicating that following the significant spike, the poultry market may have achieved a new equilibrium. The biweekly spikes and frequent oscillations are most likely caused by public sector wage payments on the fifteenth and last day of each month, as well as typical seasonal patterns. Sales exhibit a noticeable rise near the close of 2015. This increase is probably the result of higher demand over the holidays when people usually buy more food for parties and celebrations, including chicken. Furthermore, bonuses at the end of the year and extra money during the holidays might have increased consumer spending. The effects of the earthquake had a slight decline around the mid-2016. Although there was no spike in poultry sales after the earthquake, consumers may have temporarily cut back on non-essential expenditures at this time as they concentrated on healing and reconstruction. Both the SMA 10 and SMA 50 lines show a drop, suggesting a prolonged period of declining sales.

## Frozen Foods

****

For the total sales of Frozen foods, we can identify that there is a pattern of annual spikes in the sales. These spikes usually occur with the end of the year, when activities and festivities for the holidays lead to a major increase in demand for frozen meals. Sales stay constant at a lesser level in between these yearly peaks. Although there are small daily variations in the baseline sales, the general trend is stable, indicating a year-round demand for frozen goods. The sales data exhibits a constant pattern with few long-term increasing or decrease tendencies, with the exception of the annual surges. This stability points to a consistent demand for frozen meals throughout the year, with notable spikes occurring only at certain times.

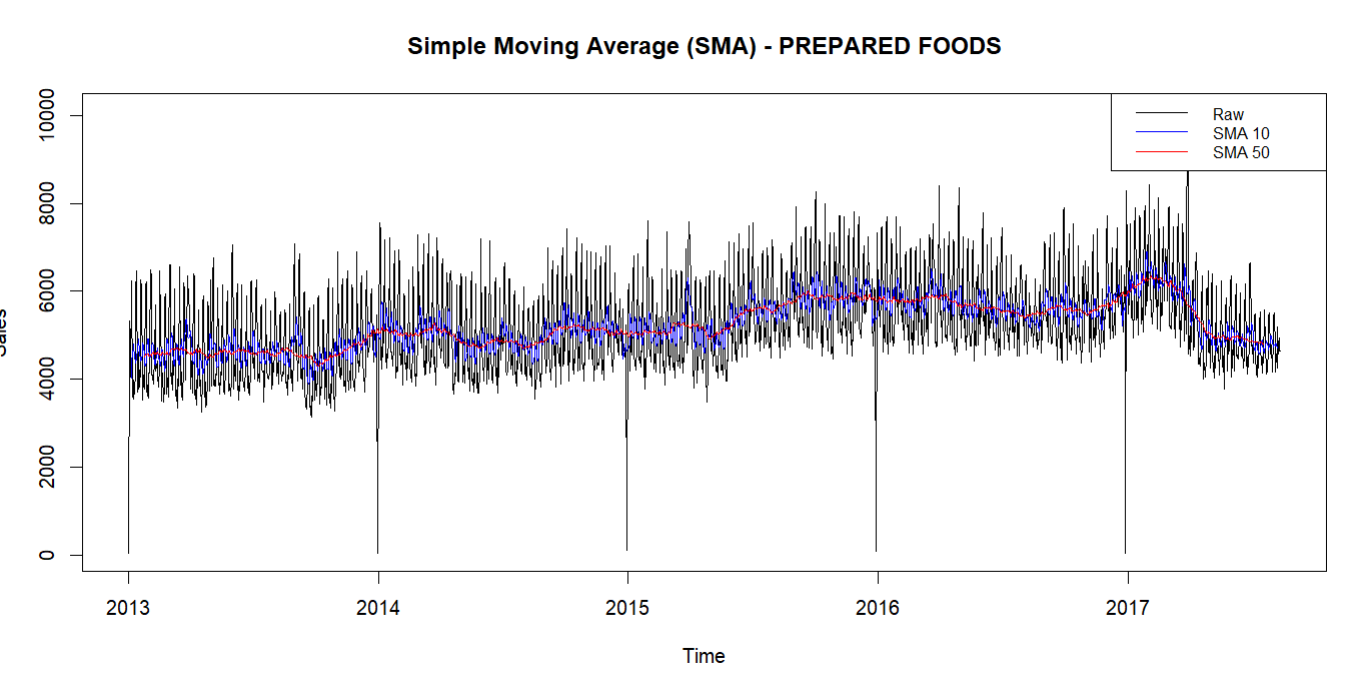
## Liquor, Wine, Beer

****

The LIQUOR, WINE, and BEER product family's overall sales data from 2013 to 2017 is shown on the graph, which uses Simple Moving Averages (SMA) to smooth out short-term variations and highlight underlying patterns. The graph indicates consistent, notable sales spikes that occur about once a year, usually in the latter part of the year. These spikes probably align with holidays and occasions such as New Year's Eve, when there is a significant increase in demand for alcoholic beverages. Sales hold steady at a lesser level in between these yearly peaks. While there are daily variations in the baseline sales, the general trend is stable, suggesting that there is a year-round demand for beer, wine, and liquor. Sales clearly fall following each peak, as shown by the SMA 10 and SMA 50 lines. After the high holiday activity, this post-spike period probably represents a normalization phase where demand returns to typical levels.

Significantly, the graph's lower region indicates a gradual decline in the smaller, baseline spikes over time, particularly starting in the middle of 2016. This pattern may point to a decrease in the daily sales fluctuation, potentially as a result of multiple factors such as market saturation and the effects of the earthquake.

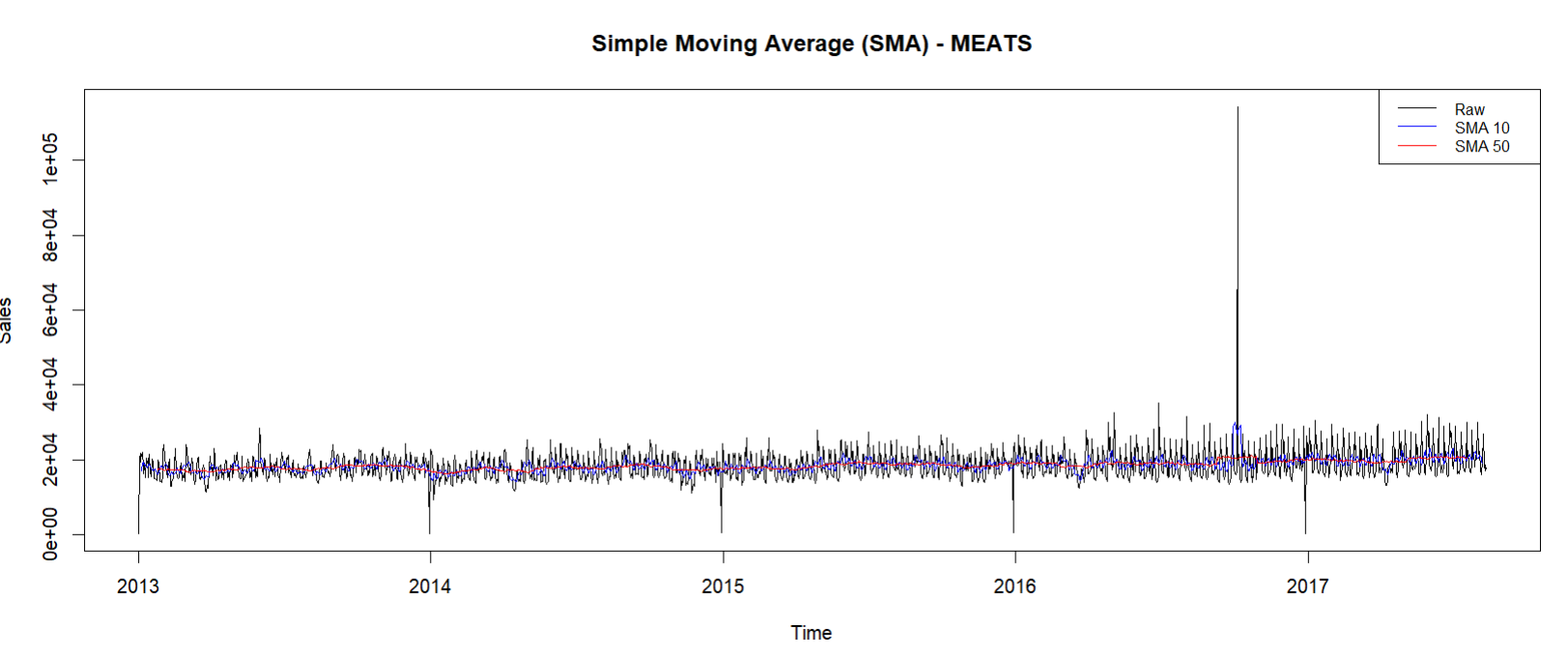
## Prepared Foods

****

The graph exhibits regular variations with slight peaks and troughs. These trends point to regular, although small, seasonal fluctuations in the demand for prepared foods. Sales of prepared foods don't see any significant yearly increases like those of other product families. This implies that the demand for prepared foods is more consistent and does not experience large seasonal fluctuations. The rising trend in the SMA 50 indicates a gradual increase in sales from 2013 to mid-2016. This implies that there was an increase in the demand for prepared foods at this time. Baseline sales show frequent tiny peaks and troughs, which are indicative of the daily or weekly fluctuations in customer buying patterns. These slight variations may be caused by things like sales events, weekdays versus weekends, and slight seasonal variations.

Sales did not significantly increase or decrease after the earthquake in April 2016, indicating that the demand for prepared foods held steady during the recovery phase. With no significant hiccups, the SMAs indicate that the current trend will continue.

## Meats

****

There are a few consistent, small variations in the graph that most likely reflect the biweekly wage payments to the public sector on the fifteenth and thirty-first of every month. These occasional spikes in sales point to a pattern in which customers buy more meat products as soon as they get paid. Small peaks and troughs are seen throughout the baseline sales period, which are generally stable. This suggests that there is a steady market for meat goods, with spikes probably brought on by marketing campaigns or small seasonal variations. Sales exhibit noticeable surges, especially towards the end of 2016. The SMA lines help to partially smooth out these surges, which are strongly reflected by the raw sales data. There wasn't a sudden, significant increase in sales after the earthquake in April 2016, but the trends are still steady with some minor variations. This suggests that, despite the immediate focus on durable commodities for relief efforts and the short lifespan of meat, the demand for meat products did not shift much in the wake of the earthquake.

## Comparison

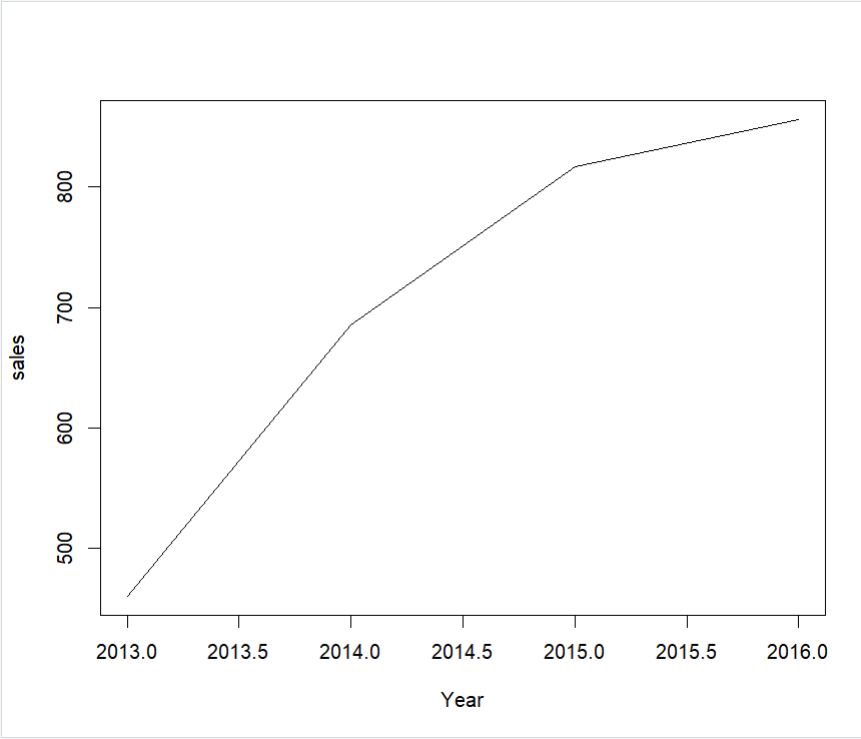
Different patterns and trends appear throughout all product families (POULTRY, FROZEN FOODS, LIQUOR, WINE, BEER, PREPARED FOODS, and MEATS). The demand for meats and poultry is comparatively steady, with only slight biweekly variations related to wage payments. Strong seasonal effects are shown by the considerable annual rises in FROZEN FOODS, LIQUOR, WINE, and BEER around the holidays. PREPARED FOODS show a growing but stable market with a slight increase followed by stabilization.

With a notable peak in 2016, LIQUOR, WINE, and BEER was the product family most impacted by the 2016 earthquake. Although there was no immediate post-earthquake surge, MEATS exhibited a notable spike around the end of 2016, probably as a result of the effects of the Christmas season and possible post-earthquake rebuilding consumption. With the exception of the noted spike patterns, no product family's sales trajectory showed an obvious structural split.

The most prominent seasonal rises were seen in FROZEN FOODS, suggesting strong holiday seasonality. Overall, external events, such as the earthquake, had the greatest impact on LIQUOR, WINE, and BEER sales patterns, showing both the immediate and seasonal effects on consumer behavior.

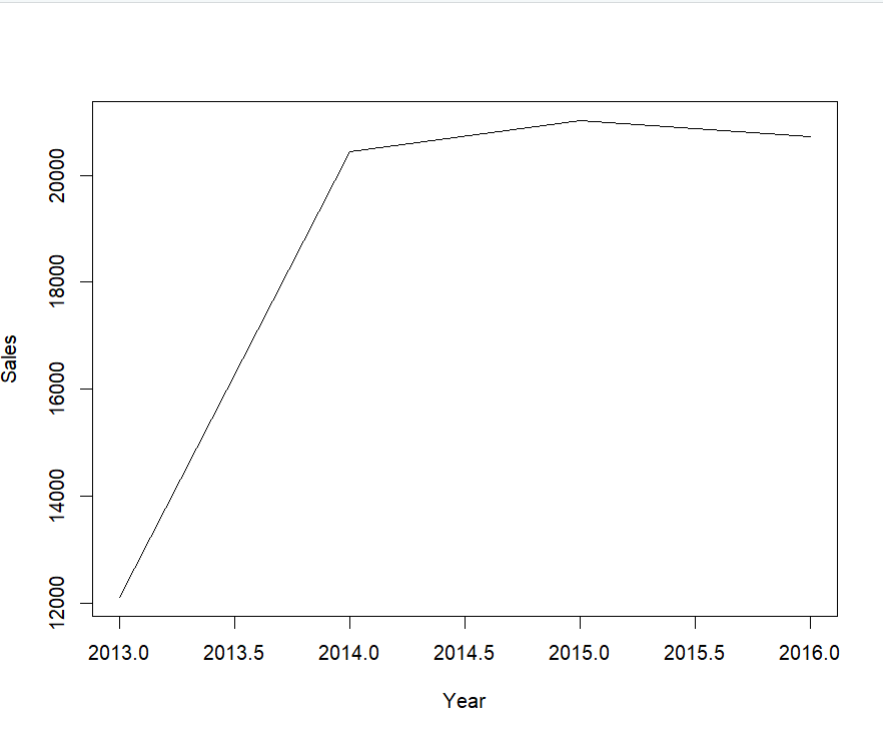
# Aggregate

## Total sales Per Year

****

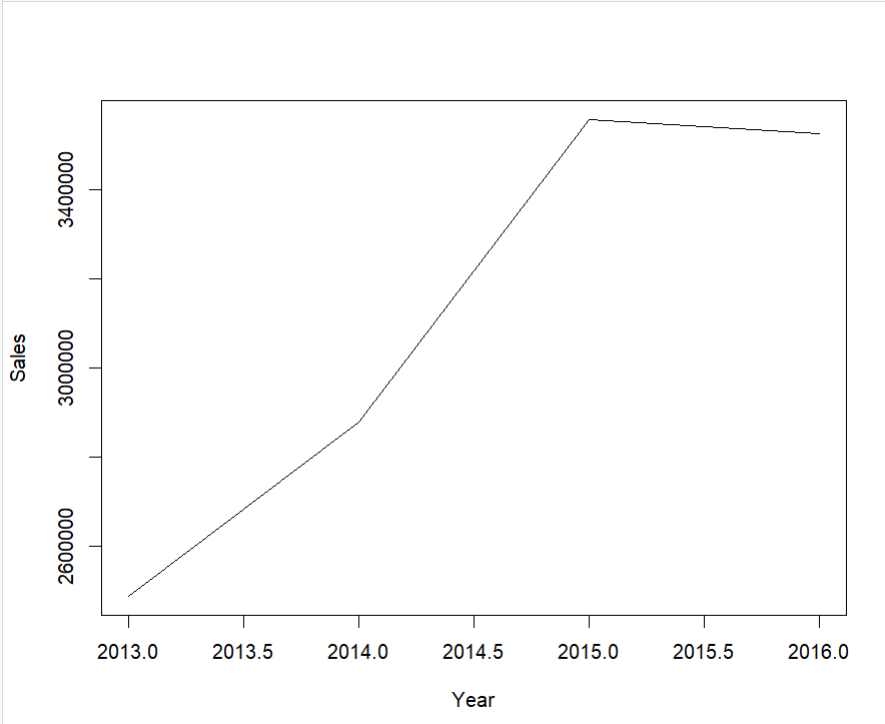
Plotting the total aggregated sales for each year between 2013 and 2016 reveals a distinct increasing trend that indicates sustained sales growth from year to year. Sales grew dramatically between 2013 and 2014, going from about 500 units to over 650 units, indicating a substantial positive growth rate during this time. Although at a slightly slower rate, the rise continued in 2014 and 2015, with sales increasing from 650 units to over 750 units. The sales growth continued in 2015 and 2016 to approach 800 units, although at a slower pace than in the preceding years. All things taken into account, the plot shows a strong and steady rise in sales, which may be ascribed to expanding markets, marketing plans, or both. The following years' decreased growth rates may indicate that the market has reached saturation or that fresh tactics are required to keep the previous momentum going.

## Poultry



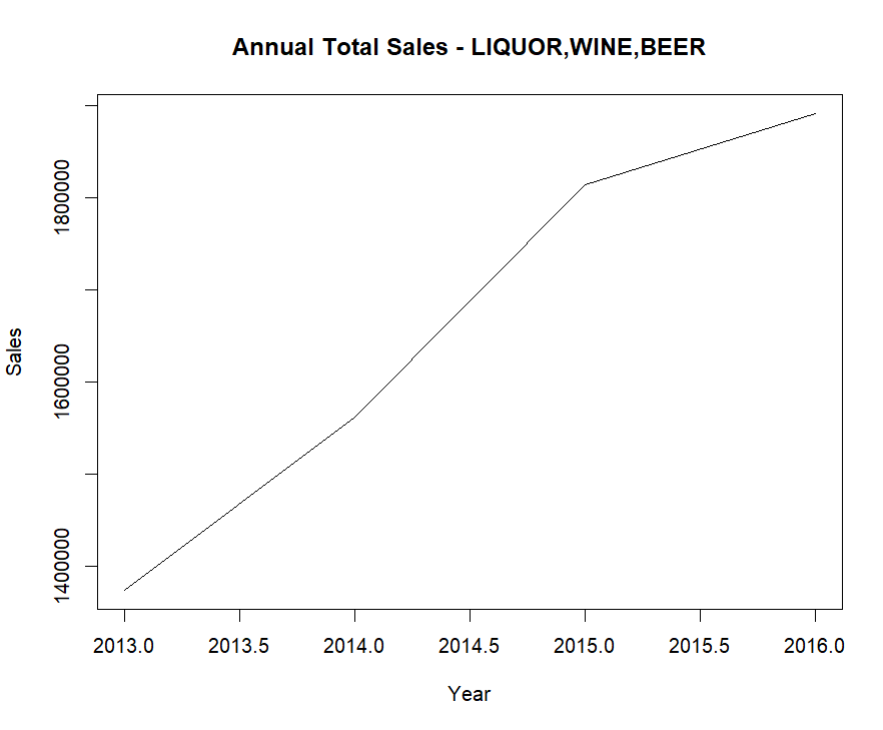
The annual total aggregated sales plot from 2013 to 2016 shows a notable rising tendency at first, which is followed by a plateau. Sales rose significantly from approximately 12,000 units to over 20,000 units in 2014, demonstrating a significant growth rate. However starting in 2014, the growth rate levels out, with sales remaining relatively stable at 20,000 units through 2015 and then slightly falling by 2016. This may indicate that the market has achieved saturation or that the growth rate has decreased, marking the end of the initial period of rapid expansion. This can suggest that fresh approaches are required to maintain growth or adjust to shifting market dynamics.

## Frozen Foods

****

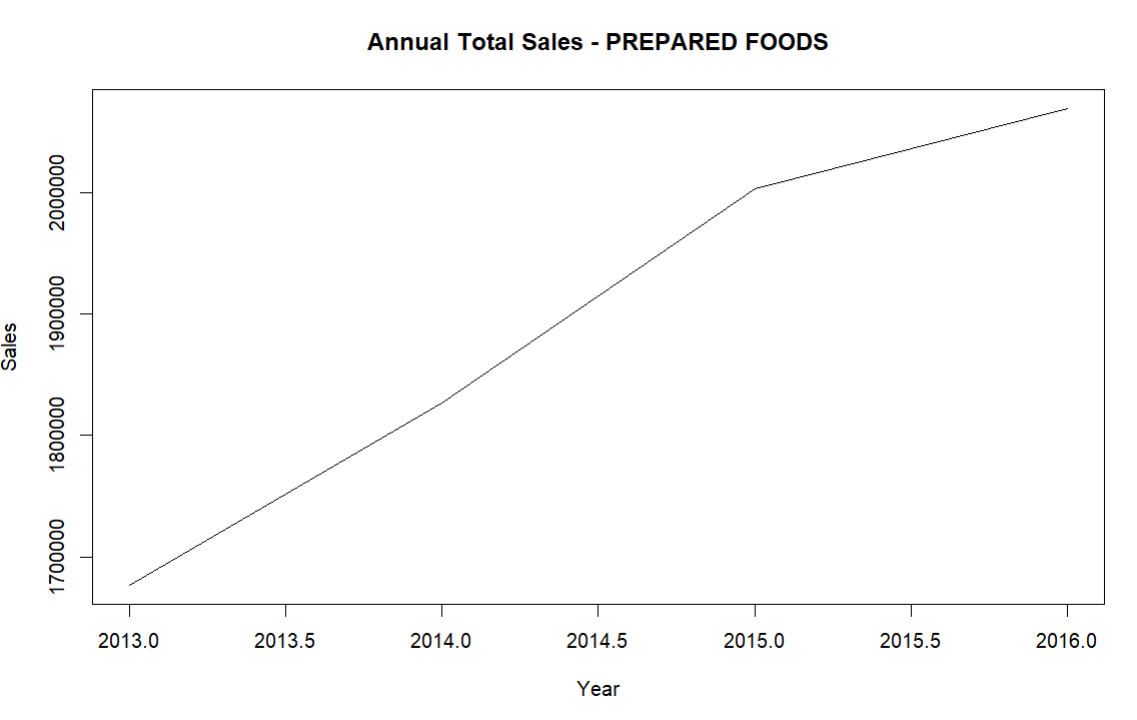
Plotting the total aggregated sales for frozen foods from 2013 to 2016 shows a clear upward trend that ends with a noticeable halt. 2013 had an initial sales volume of about 2.6 million copies. This number increases gradually, showing a clear rising trend, peaking at about 3.5 million units by 2015. This sudden spike points to a period of fast growth, possibly profitable company plans, or a demand for frozen foods in the market. After 2015, though, the growth rate levels off, and sales stay close to the 3.5 million threshold. The sales signify that the market demand has steadied and the growth phase has reached a saturation point.

## Liquor Wine Beer



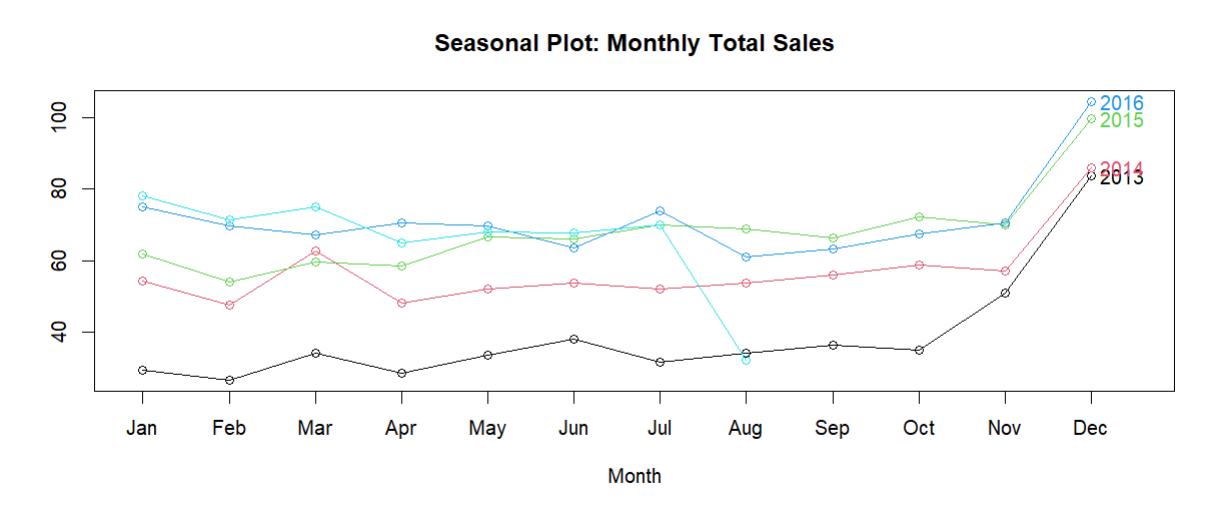
A steady increase trend can be seen in the plot of the annual total sales of LIQUOR from 2013 to 2016. Sales have been rising annually since 2013 when there were about 1.4 million devices sold. Sales increased to almost 1.6 million units by 2014, suggesting a robust expansion in this area. This increasing trend continued through 2015 and 2016, with an approximate total of 1.8 million units sold throughout that time. This steady expansion points to either increasing consumer demand for these goods or the use of effective sales techniques during these years. The steady rise in sales suggests that there is a sizable and maybe growing market for products related to LIQUOR.

## Prepared Foods

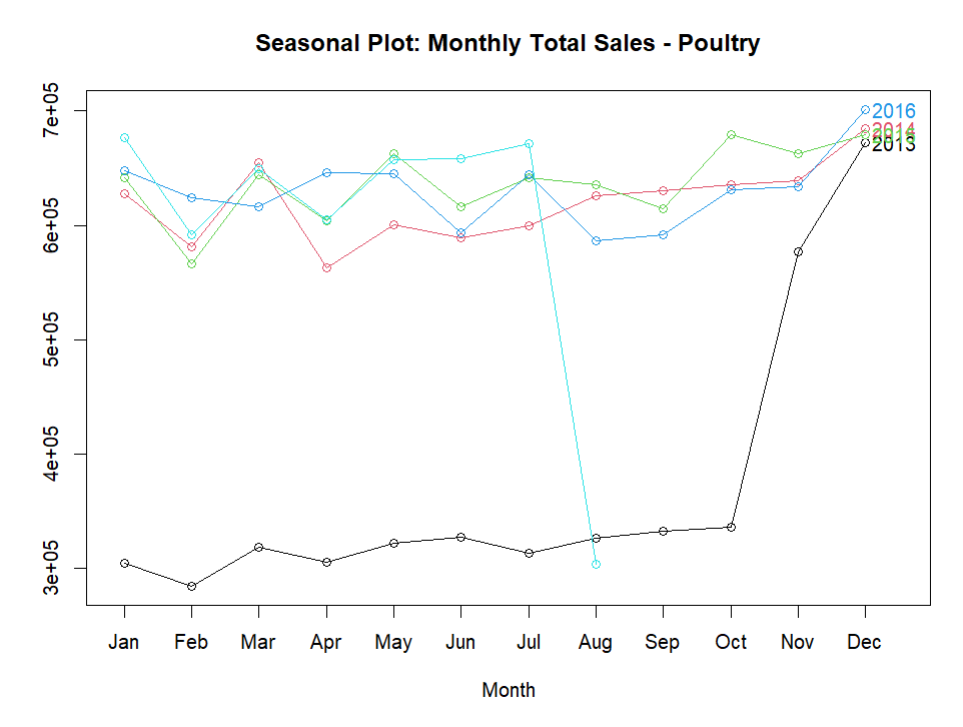


The yearly total sales of PREPARED FOODS line graph from 2013 to 2016 shows a steady increase trend. Sales have climbed yearly since 2013, when they were roughly 1.68 million. By 2016, they had surpassed 2 million. This consistent rise points to a steady and growing market for PREPARED FOODS, which may be fueled by reasons like rising customer demand, successful marketing campaigns, or more widely available products. There haven't been any notable abrupt fluctuations in the sales pattern, which suggests that this market segment has been strong and expanding steadily throughout this time.

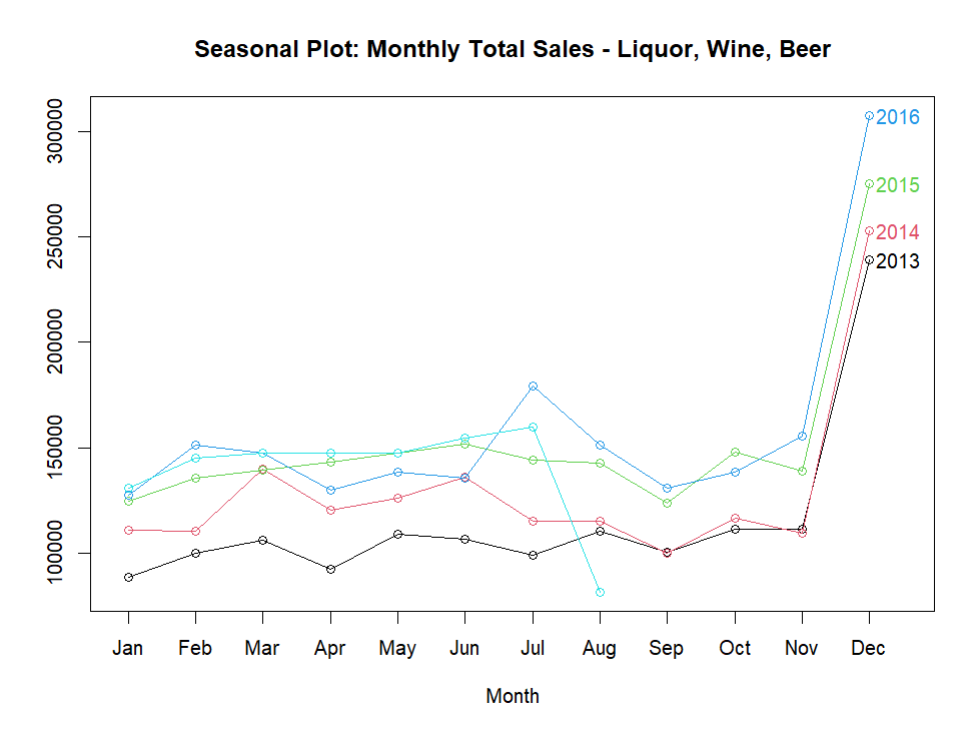
# Checking Seasonality

****

A notable abnormality in July 2016 can be seen in the seasonal plot of monthly total sales from 2013 to 2016, which is probably due to the impact of the earthquake. July 2016 sales data indicates a substantial fall in sales across all product families, in contrast to past years when sales were either marginally increased or stayed reasonably flat. This abrupt decline suggests a break in regular sales patterns, which is in line with what happens following a big catastrophe like an earthquake. After the July decline, sales eventually show signs of recovery and rise, particularly in the latter part of 2016. This recovery could indicate attempts to return to normalcy as well as a potential shopping boom as local economies recovered from the consequences of the earthquake.

****

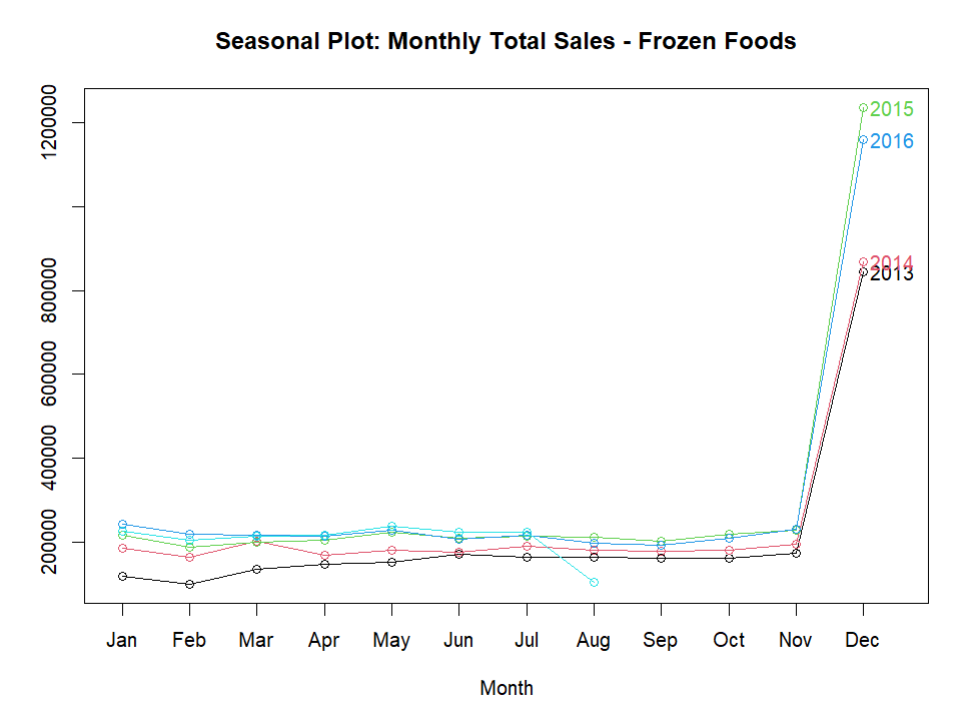
The monthly total sales for poultry in a seasonal plot from 2013 to 2016 show a distinct seasonal pattern, with a noticeable interruption in the middle of the year that is probably related to the earthquake. Poultry sales generally show some slight changes but generally follow a consistent trend. Nonetheless, there is a notable decline in sales in July 2016, departing significantly from the typical pattern seen in prior years. This drop suggests that the earthquake had an immediate effect on supply networks, consumer behavior, or both. Following this disruption, sales began to recover, particularly evident towards the end of 2016, where sales not only returned to normal levels but also showed an increase, suggesting a robust recovery phase. This recovery could be attributed to reconstruction efforts, increased consumer demand post-crisis, or market



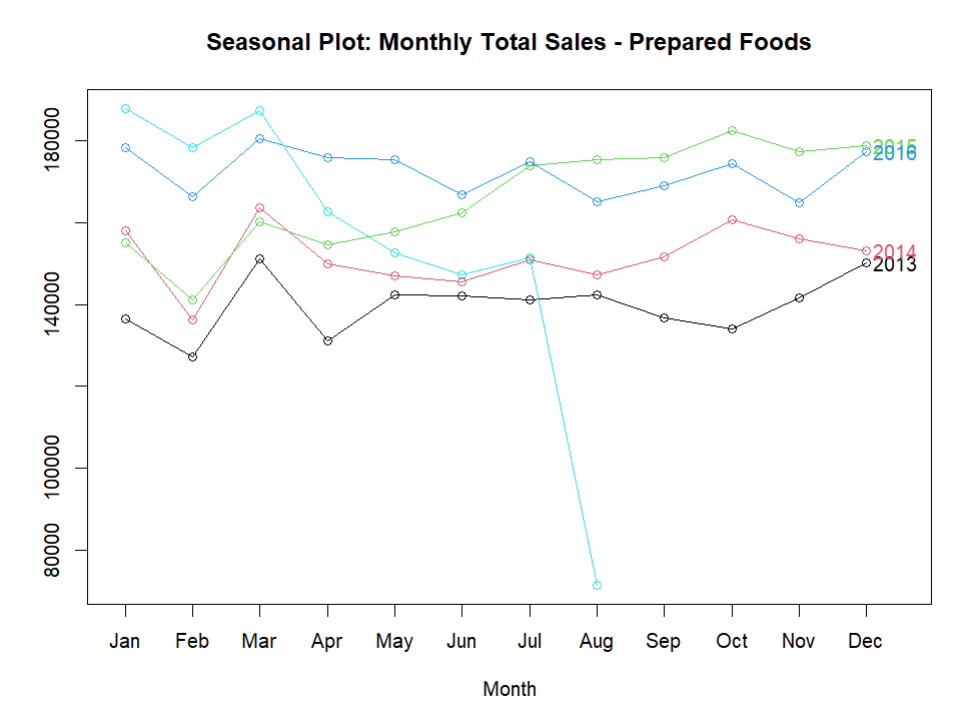
From 2013 to 2016, the seasonal plot of monthly total sales for wine, beer, and liquor reveals a steady sales pattern with a notable seasonal high in December every year. In contrast to prior years, the data for 2016 shows a notable decline in sales in July, which is probably related to the effects of the earthquake. After this decline, sales exhibit a recovery tendency, which is especially noticeable in the dramatic surge around the end of 2016, when December sales levels exceeded those of prior years. This rebound suggests a post-earthquake level of resilience and potentially higher consumer demand.



Some noteworthy patterns may be seen in the seasonal plot of Meats' monthly total sales from 2013 to 2016. With the exception of a notable decline in July that was probably caused by the earthquake, the sales in 2016 mainly followed a similar pattern to those in prior years. This decline, which is not observed in previous years, highlights the precise effect of the earthquake on meat sales. Sales in 2016 bounce back fast in the months that follow, reaching a notable peak in September, indicating a robust upturn in consumer demand, despite this disruption. With minor seasonal changes, the overall sales pattern for meats is consistent throughout the year. The consistent increase in sales towards the end of the year, especially in November and December, aligns with holiday season demand.



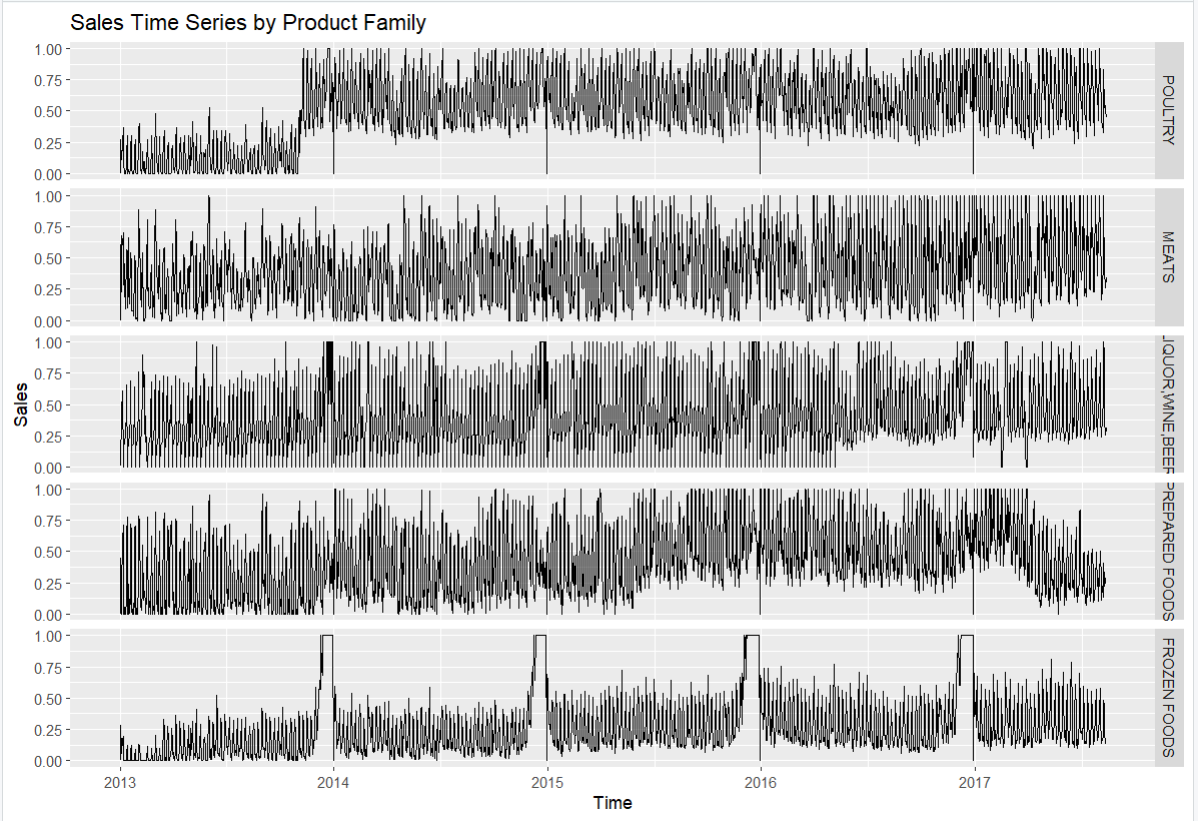
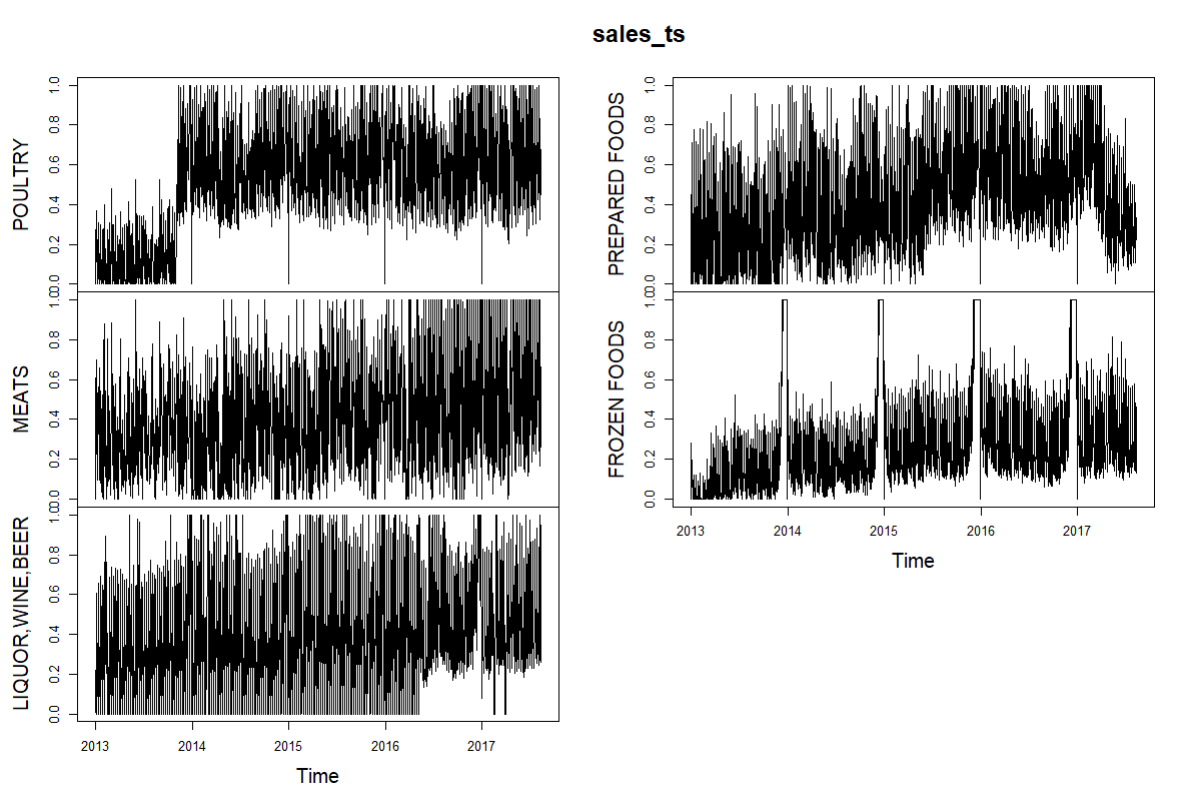
From 2013 to 2016, there were notable seasonal spikes and recurring trends in the monthly total sales for frozen foods, as shown by the seasonal plot. For the majority of the year, sales just slightly fluctuate and stay relatively consistent. Remarkably, the sales show a sharp rise in December every year, probably due to increased demand for the holidays. This pattern is particularly noticeable in 2015 and 2016, when sales peak much higher than in prior years, suggesting that frozen foods are becoming more and more popular throughout the Christmas season. Fascinatingly, the sales pattern for frozen foods does not appear to have been disrupted in spite of the 2016 earthquake, indicating that this sector was not greatly affected. All things considered, the results show a robust seasonal demand for frozen foods in December, with comparatively stable sales during other months.

****

From 2013 to 2016, the monthly total sales for prepared foods displayed a seasonal plot that displays clear seasonal trends along with a few oddities. Sales often peak in the first few months of the year, particularly in January and February, and then gradually decrease in the middle of the year. Notably, there is a notable decline in July 2016 that is probably caused by the effects of the earthquake because this decline is not seen in prior years. Despite this setback, sales pick significantly in the ensuing months, especially starting in August. In addition, sales in 2016 had a greater baseline than in prior years, suggesting that the Prepared Foods sector is generally growing.

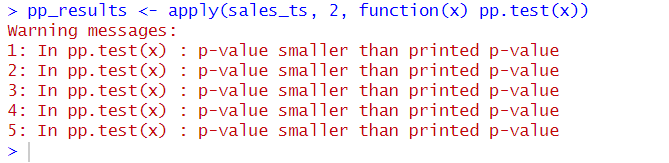
# Vector Autoregression (VAR)

In order to identify the relationships between the sales of various product families in our dataset, we employed Vector Autoregression (VAR) analysis. We were able to comprehend how variations in the sales of one product family affected the others thanks to this method. We took advantage of these interdependencies by using VAR to anticipate our sales data in a way that was more thorough and accurate. In order to determine how shocks, like the 2016 earthquake, affected our sales, we also used VAR impulse response analysis. This gave us the opportunity to watch how sales varied depending on the product family in response to these occurrences, giving us important information for improved decision-making and scenario planning.

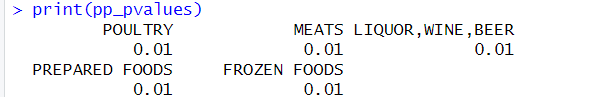
****

The sales time series data for the five product groups (POULTRY, MEATS, LIQUOR, WINE, BEER, PREPARED FOODS, and FROZEN FOODS) from 2013 to 2017 are shown in two distinct plots created with autoplot and plot.ts. Every product family is displayed in a separate subplot in a faceted plot created by the autoplot function. With reduced visual clutter and a stronger emphasis on particular series, this arrangement facilitates the comparison of trends and patterns across several product families. Plot.ts, on the other hand, displays each series in different panels but within the same plot frame by combining all product families into a single plot with a multi-panel structure. The overlapping data and different scales in this approach can make it more difficult to interpret. Plot.ts delivers a simple but less sophisticated representation of the data, whereas autoplot offers a better organized and visually appealing depiction overall.

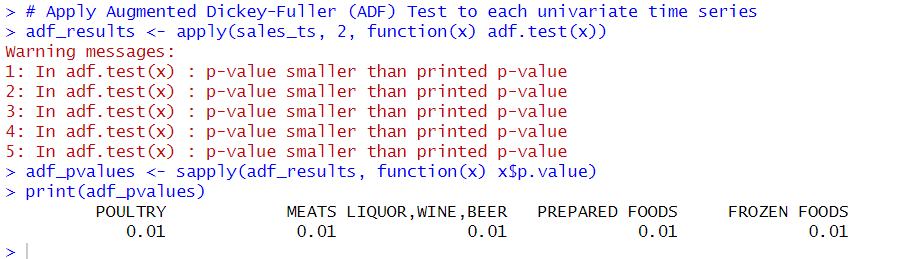
## Phillips-Perron Unit Root Test- univariate time series



When testing the Phillips-perron unit root test, The p-values for each time series in the sales\_ts dataset are extremely small, less than the values that R usually prints by default. This implies that the unit root null hypothesis can be rejected, indicating that the time series data is most likely stationary. These results are optimistic for more investigation because stationarity is a necessary assumption for moving forward with time series analysis, including Vector Autoregression (VAR).

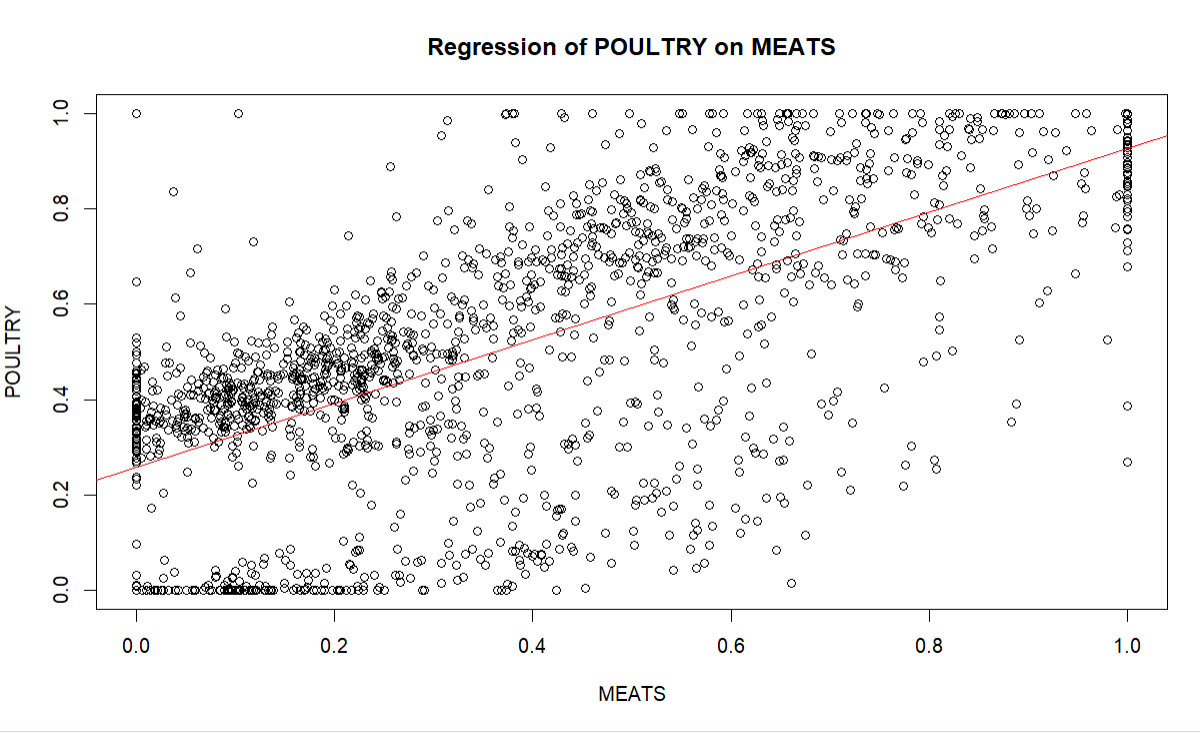


All of the p-values for the different product families in the dataset, by the Phillips-Perron (PP) test, are 0.01—a decrease from the generally accepted significance level of 0.05. This indicates that the unit root null hypothesis for each time series can be rejected, proving the stationary nature of the data for poultry, meats, liquor, wine, beer, prepared foods, and frozen foods.

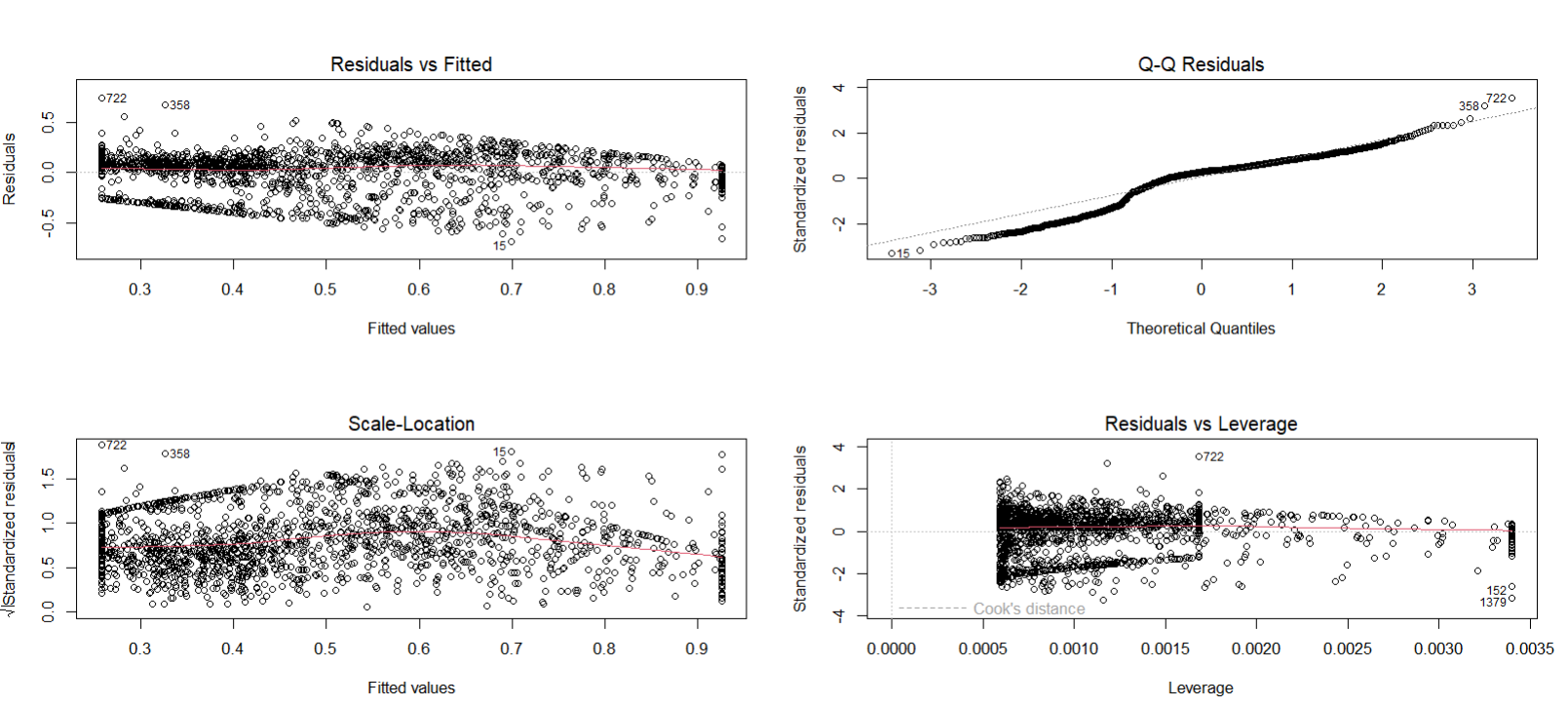


Consequently, this confirms that the time series data for all these product families are stationary.

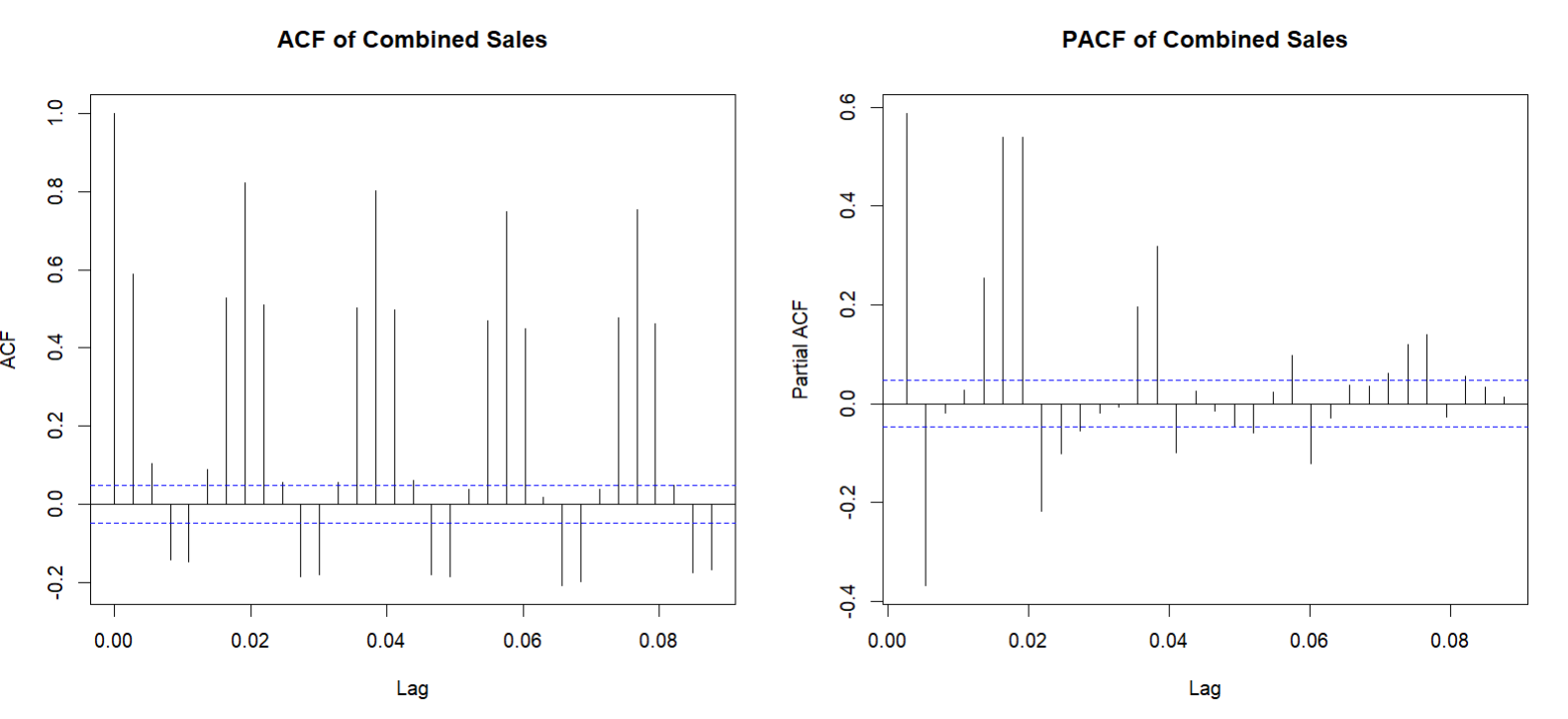
## Regression plots of similar product families:

****

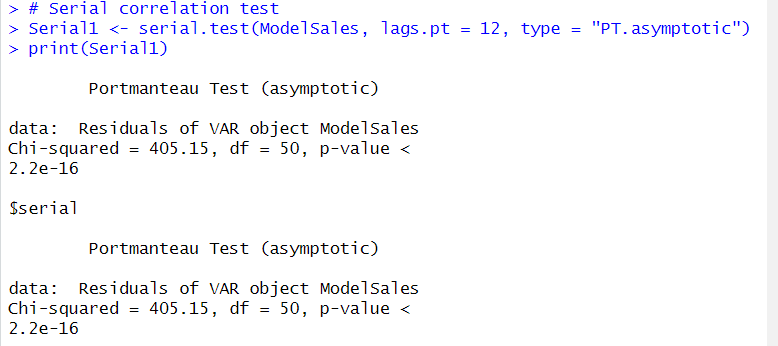
The residual plot, which is a scatter plot that shows the disparities between observed and values from the regression of POULTRY sales on MEATS sales, is what is being displayed. The MEATS sales numbers are represented on the x-axis, while the residuals are plotted on the y-axis. The regression line, or more accurately, the line that best fits the data, is represented by the red line. The residuals should ideally be dispersed randomly down the horizontal axis, signifying that the regression model fits the data well. However, a trend can be seen in this plot, particularly at the extremities of the MEATS sales values. This raises the possibility of problems with the model, such as non-linearity or the impact of extraneous factors. The upward-sloping regression line shows that POULTRY and MEATS sales have a noticeable positive association. The large dispersion of data points surrounding the regression line, however, indicates that there is substantial variability in POULTRY sales that cannot be fully explained by MEATS sales. Data point clusters at specific MEATS sales levels, especially at lower values, show times when POULTRY sales are regularly greater or lower than expected, which could be an indication of outside influences on sales. Furthermore, the existence of dense clusters and outliers points to possible model problems, such as heteroscedasticity or non-linearity, in which the variability of POULTRY sales rises as MEATS sales rise. This variation suggests that the inaccuracy of the model is not constant for all MEATS sales levels.

****

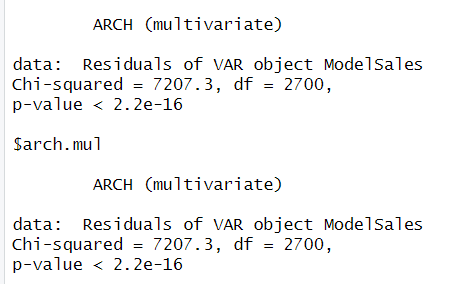
The residuals are dispersed randomly about the horizontal axis, as seen in the "Residuals vs. Fitted" plot, suggesting that while the model fits well in some areas, there is also some curves, suggesting the possibility of non-linearity. The "Q-Q Residuals" plot suggests that the residuals are not normally distributed since it shows departures from normality, especially in the tails where the dots diverge from the diagonal line. Heteroscedasticity is suggested by the "Scale-Location" plot's small upward trend, which shows that the residuals' variance rises with fitted values. Lastly, the "Residuals vs. Leverage" figure reveals that while the majority of data points have low leverage, certain high-leverage points—possibly significant influences on the regression results.

****

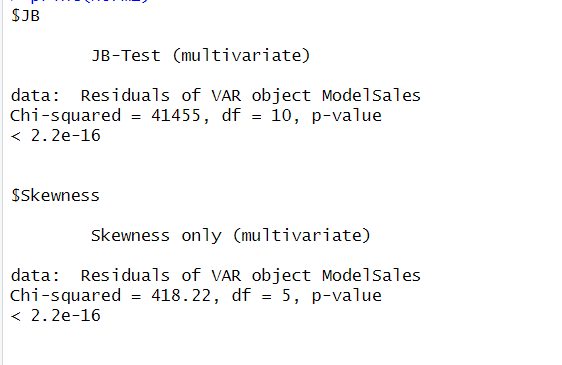
Significant partial autocorrelations are shown in the PACF plot, mostly at the first lag, suggesting a direct relationship between recent past values and current values. Since recurring patterns at lags of ACF are a defining characteristic of seasonal impacts, the pattern of these spikes may also point to possible seasonality or cyclic activity within the sales data.

****

We reject the null hypothesis that there is no serial connection based on the relatively small p-value (< 2.2e-16) found in the Portmanteau Test results for the VAR model ModelSales. This indicates that the residuals have a substantial amount of autocorrelation, indicating that the dependencies in the data have not been adequately represented by the model.

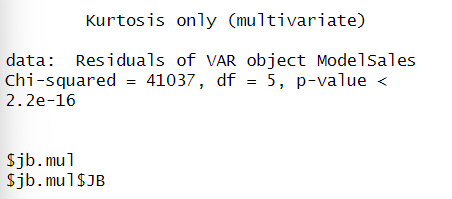
****

Since the p-value is very low, we reject the null hypothesis that there is no heteroskedasticity. This suggests that the residuals of the VAR model ModelSales exhibit a considerable degree of heteroskedasticity. Heteroskedasticity indicates that the residuals of the model have changing variances over time, which may have an impact on the model's predictive power and the validity of statistical conclusions drawn from it.



**JB-Test (Jarque-Bera Test):** The extremely small p-value suggests that we reject the null hypothesis of normally distributed residuals, indicating significant non-normality.

**Skewness Test:** The data are not symmetrically distributed, as indicated by the small p-value, which also suggests severe skewness in the residuals.



**Kurtosis Test:** Significant kurtosis in the residuals, indicated by the small p-value, suggests that the data has larger tails or more extreme values than a normal distribution.

## Structural Breaks

****

These plots indicate the the stability over time and the potential structural breaks. Movements outside the boundary suggest the structural breaks or changes in the regression.

**MEATS**: The initial steady drop is followed by variations that approach but do not cross the bottom border in the OLS-CUSUM plot for MEATS. This suggests a state of relative stability with no notable structural breaks.

**PREPARED FOODS**: This figure similarly depicts a process that is largely stable, but it also displays a noticeable rising trend that approaches the upper border at the end, suggesting that there may be instability or a structural breaks at that point.

**POULTRY**: Like PREPARED FOODS, the POULTRY plot exhibits a stable process at first, but toward the end, there is a noticeable rising trend that approaches the upper border and raises the possibility of a structural break or change.

**FROZEN FOODS**: This plot shows a slightly elevated area towards the conclusion and a more intricate pattern with fluctuations. Although the trend does not go beyond the bounds, it does point to possible instability.

**LIQUOR, WINE, AND BEER**: This plot displays an early downward trend, which is followed by a steady process, culminating in a notable upward trend that approaches the upper boundary. This suggests that there may have been structural instability.

Although FROZEN FOODS occasionally fluctuates, MEATS and FROZEN FOODS exhibit greater stability overall when compared to the other series. LIQUOR, POULTRY, AND PREPARED FOODS are possible structural breaks. Near the end, WINE.BEER exhibits notable upward trends that point to possible structural alterations or fractures. Prepared foods, poultry, and liquor exhibit the greatest degree of variation among all the series. There are more noticeable tendencies in LIQUOR that point to possible stability problems.

## Casual Relationships

**Poultry:** According to the findings of the Granger causality test, sales of poultry in the past can be used to forecast sales of MEAT, LIQUOR, PREPARED FOODS and FROZEN FOODS. The null hypothesis that POULTRY does not Granger-cause the other product groups is rejected based on an F-Test value of 8.3975 and a p-value of less than 2.2e-16. This indicates a strong predictive association between them and, to put it simply, shows that changes in POULTRY sales can be used to predict changes in the sales of these other product families.

**Meats:** The null hypothesis in the Granger causality test for MEATS states that MEATS does not Granger-cause POULTRY or LIQUOR, PREPARED FOODS as well as FROZEN.FOODS. With a p-value of less than 2.2e-16 and an F-Test value of 10.07, the null hypothesis is rejected. This indicates that MEATS sales can predict future sales of these other product families. Additionally, the instantaneous causality test's null hypothesis states there is no instantaneous causality between MEATS and the other product families. With a Chi-squared value of 595.87 and a p-value < 2.2e-16, this hypothesis is also rejected, indicating significant instantaneous causality.

**Frozen Foods**

The null hypothesis of the Granger causality test for FROZEN.FOODS states that it does not Granger-cause MEATS, POULTRY, or LIQUOR. With a p-value of less than 2.2e-16 and an F-Test value of 7.5454, the null hypothesis is rejected. This implies that sales of FROZEN.FOODS may be used to forecast sales of these other product groups.

**Prepared Foods:**

For PREPARED.FOODS, the Granger causality test indicates a causal relationship with other product families. The null hypothesis, states that PREPARED.FOODS does not Granger-cause MEATS, POULTRY, LIQUOR.WINE.BEER, and FROZEN.FOODS is rejected with an F-Test value of 5.2953 and a p-value < 2.2e-16. This means PREPARED.FOODS sales can help predict the sales of these other product families.

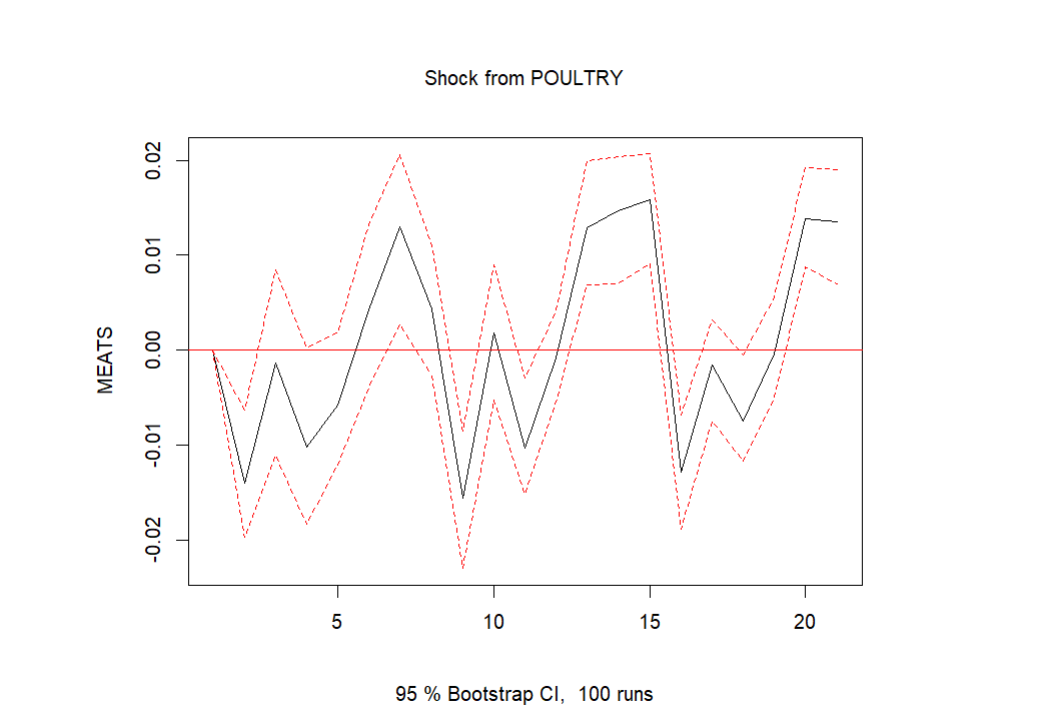
**LIQUOR:** The Granger causality test for LIQUOR.WINE.BEER shows a causal relationship with other product families. With an F-Test value of 10.203 and a p-value < 2.2e-16, the null hypothesis—which holds that LIQUOR.WINE.BEER does not Granger-cause MEATS, POULTRY, PREPARED.FOODS, and FROZEN.FOODS—is rejected. This implies that sales of LIQUOR, WINE, and BEER may be used to forecast sales of these other product groups.

## Shocks

### Poultry and Meats

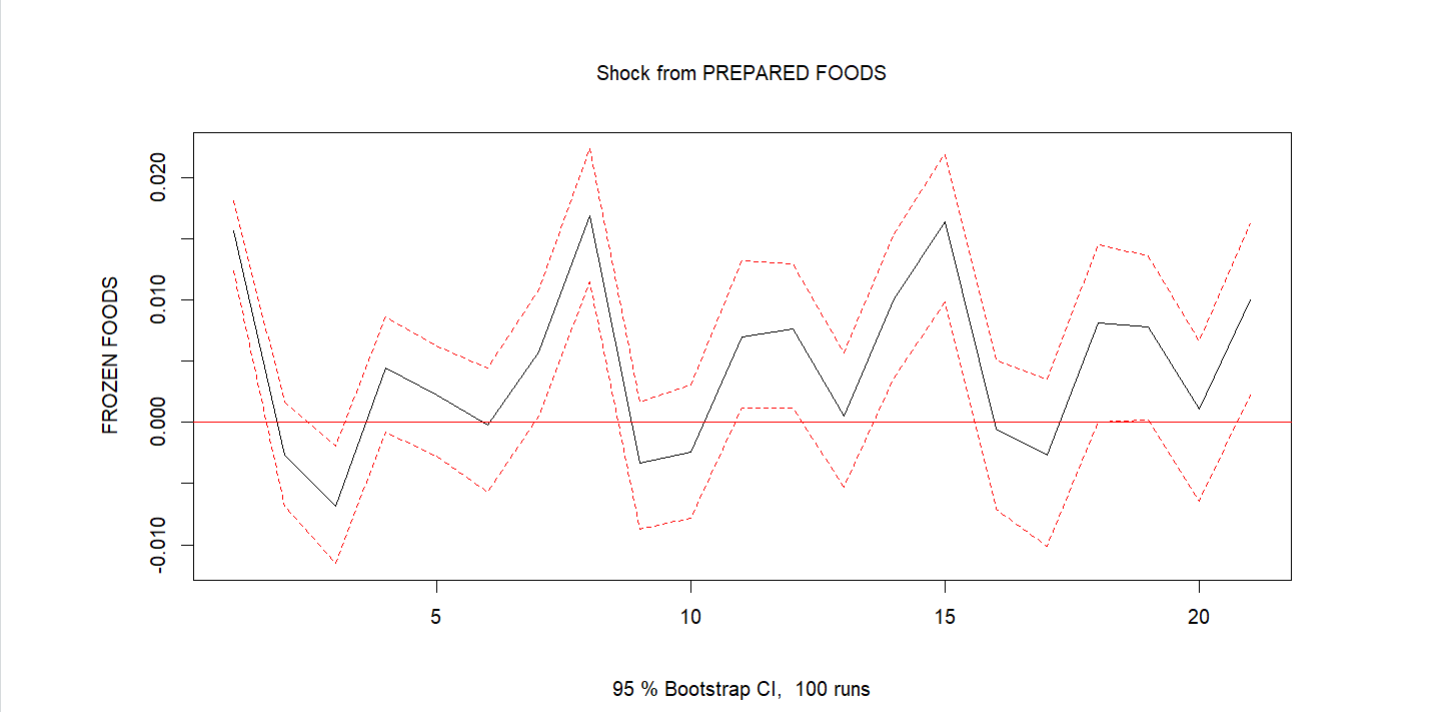
### 

This graph conveys the impulse response function of Poultry sales to Meat sales. This algorithm calculates the long-term effects of a sudden shift in MEATS sales on POULTRY sales. The estimated response of POULTRY sales is shown by the solid black line, and the 95% bootstrap confidence intervals based on 100 simulations are shown by the red dashed lines. Sales of POULTRY show a noticeable initial positive effect that quickly peaks. Nevertheless, this positive effect fades quickly and almost disappears in a short while. After the first decline, POULTRY sales response varies slightly about zero, showing brief intervals of favorable and unfavorable effects. These fluctuations settle and remain near zero over time. The main line's surrounding narrow red dashed lines show that these estimates are reasonably accurate. Overall, poultry sales are positively impacted by a quick surge in meat sales, but this effect is only temporary; poultry sales eventually level off.

****

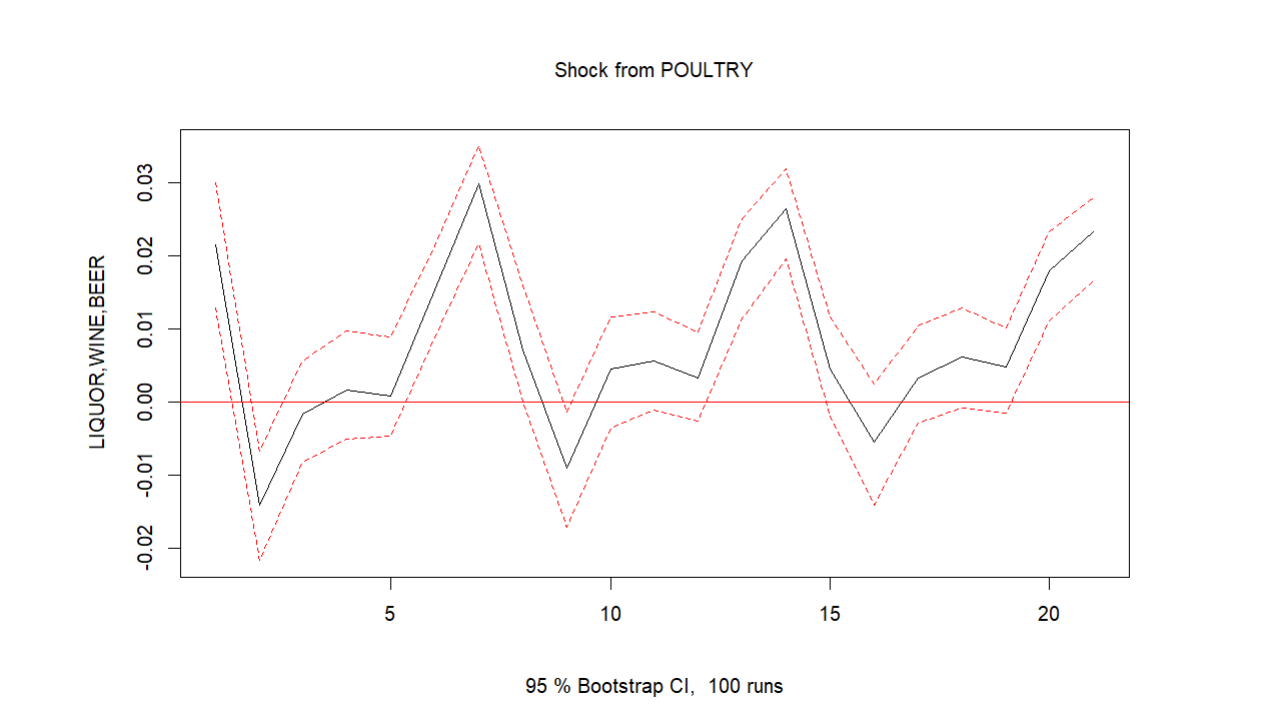
Following a shock to poultry sales, there is initially a slight negative impact on meats sales, but this effect quickly levels out. The reaction is almost constant throughout the course of the time period, with only little variations above and below the baseline, suggesting that the long-term impact of POULTRY sales on MEATS sales is minimal. Based on the confidence intervals, it appears that these variations lack statistical significance.

### Frozen Foods and Prepared Foods

****

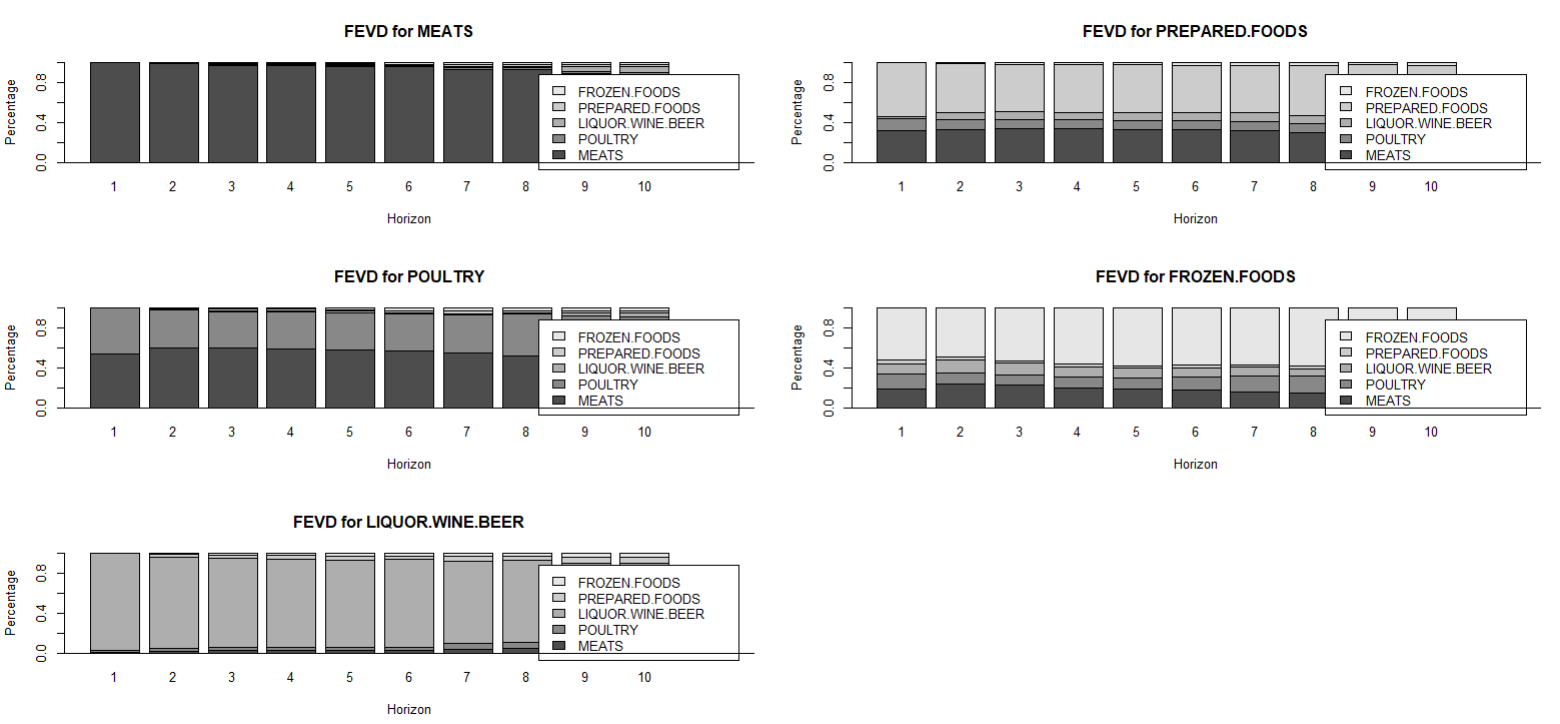
Sales of FROZEN FOODS are initially negatively impacted when sales of PREPARED FOODS are shocked, but this effect gradually approaches zero. Response patterns alternate between brief positive and negative effects, suggesting that variations in the sales of PREPARED FOODS have nonoticable, long-term impact on the sales of FROZEN FOODS. There appears to be no apparent or noteworthy long-term impact of higher sales of PREPARED FOODS on FROZEN FOODS, as the two goods have a weak and inconsistent association. This would suggest that, despite possible brief connections, the two product groups have rather autonomous influences on sales.

### Liquor and Poultry



As seen by the fall below zero, a positive shock to POULTRY sales initially causes a slight decline in LIQUOR.WINE.BEER sales. The response fluctuates around zero with both positive and negative deviations with time, although this impact is transient. The estimated impact is represented by the black line, while the 95% confidence intervals from 100 bootstrap runs are displayed by the red dashed lines. Given that the reaction varies and lacks a discernible pattern, the findings imply that there is no reliable, long-term correlation between shocks to POULTRY sales and variations in LIQUOR, WINE, and BEER sales.

## FEVD

****

In all charts, the majority of the forecast error variance is explained by the product itself, indicating strong self-dependence. Other products have made some noteworthy contributions, nevertheless. The impact of POULTRY on MEATS is noticeable, although still minimal. Meats and prepared foods have a more visible impact on poultry. Strong internal predictability is indicated by the decomposition of MEATS, which reveals that MEATS alone accounts for almost all of the forecast error variance. The forecast error variance for PREPARED.FOODS is mostly determined by PREPARED.FOODS; however, with time, contributions from FROZEN.FOODS and other product families become somewhat more apparent.

Regarding POULTRY, the breakdown indicates that although POULTRY is primarily responsible for the variance in its forecast, contributions from MEATS and other product groups increase gradually. In the same way, FROZEN.FOODS over the prediction, the impact of other families, such as POULTRY, becomes marginally relevant, but FOODS is predominantly impacted by its own values.

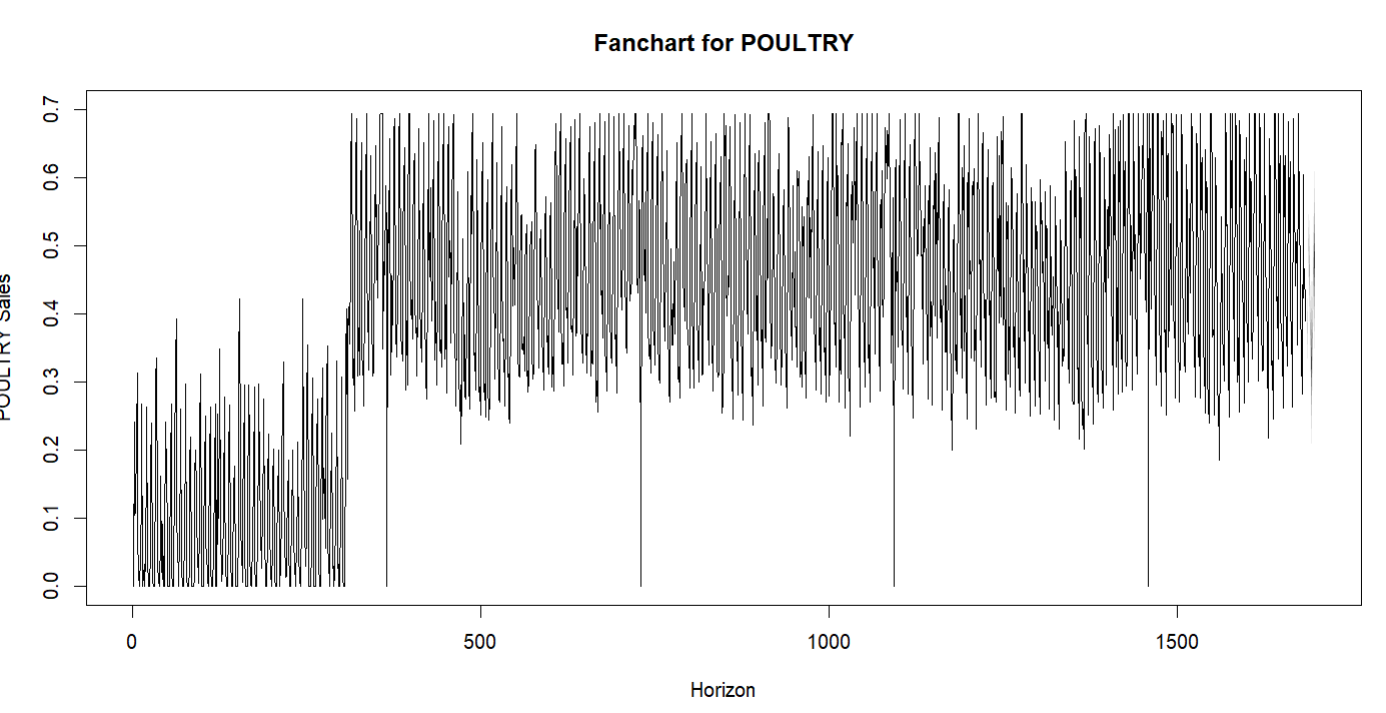
Ultimately, LIQUOR.WINE.BEER's FEVD shows that it is the most self-predictive product family, with minimal influence from other product families throughout the course of the forecast period. The LIQUOR.WINE.BEER sales data exhibits a high autoregressive tendency, as indicated by this consistency.

To summarize the prediction error variance of each product family can be mostly explained by its past values, while there is a progressive influence from other families as well. This indicates different levels of interdependence among the product sales.

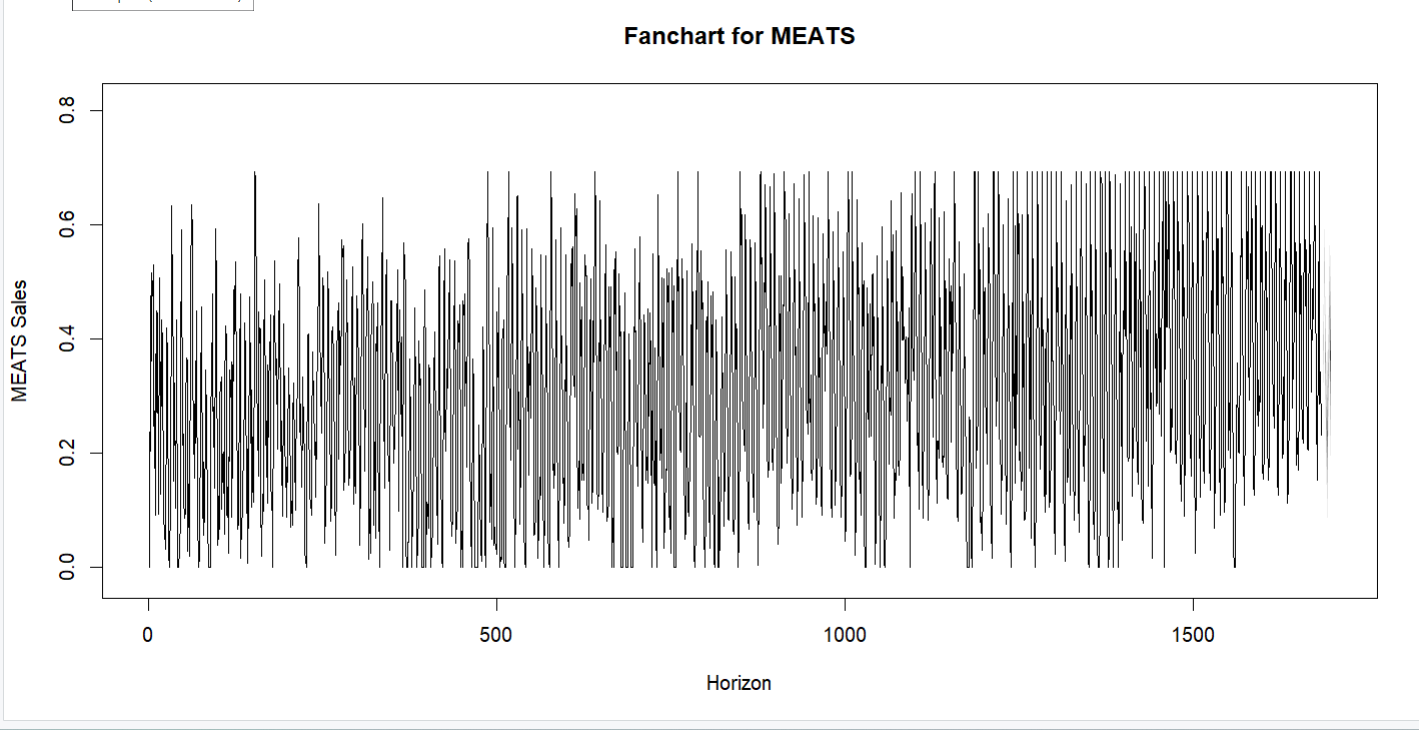
## Relationships:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Product Family** | **Self Influence** | **Influence from Poultry** | **Influence from Meats** | **Influence from Liquor, Wine , Beer** | **Influence from Prepared Foods** |
| **Meats** | High | Low |  | Very Low | Very Low |
| **Prepared Foods** | High | Low | Very Low | Very Low |  |
| **Poultry** | High |  | Very Low | Very Low | Very Low |
| **Frozen Foods** | High | Very Low | Very Low | Very Low | Very Low |
| **Liquor, wine, beer** | High | Very Low | Very Low |  | Very Low |

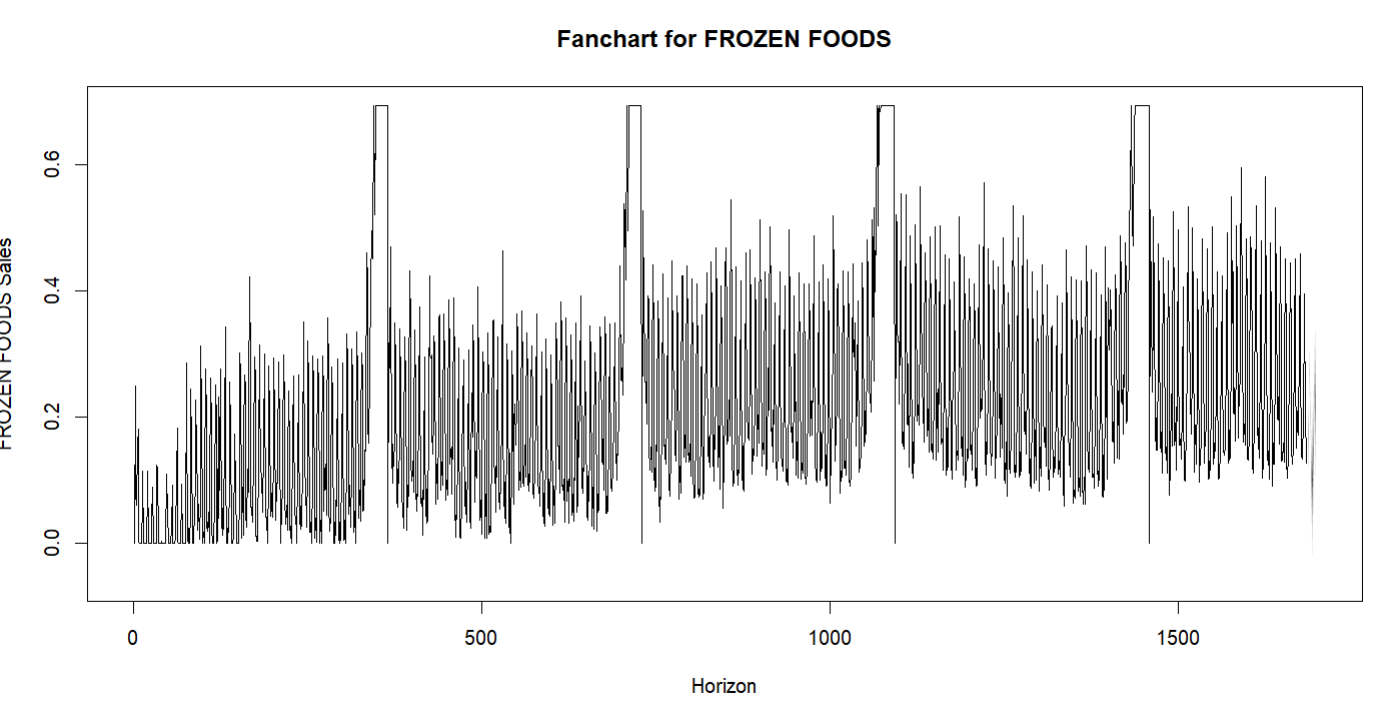
## Fancharts

****

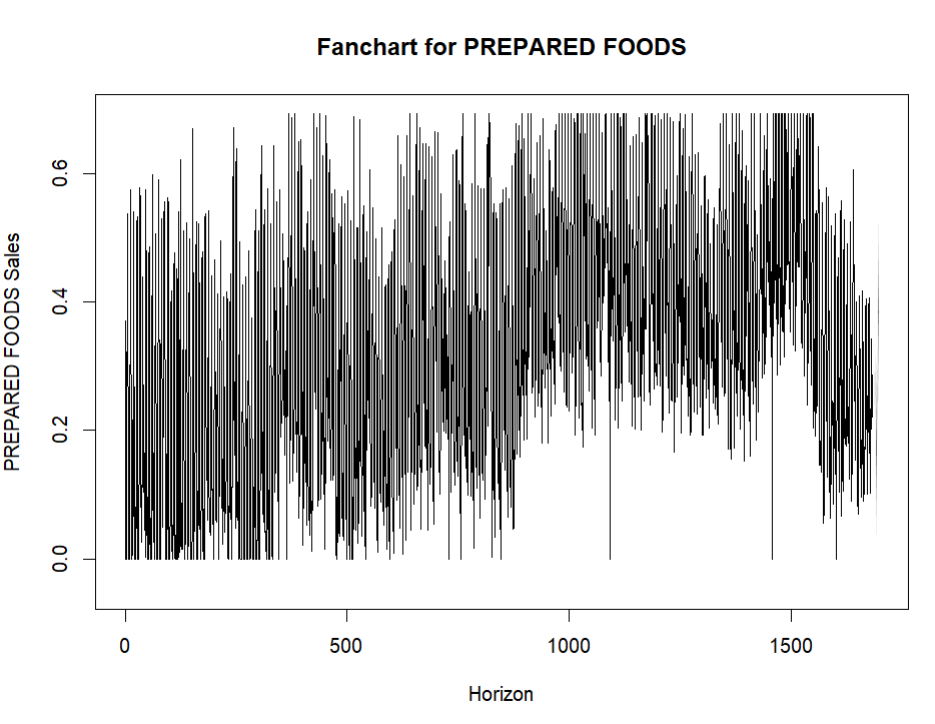
From the beginning of the horizon, the narrow spread of the lines indicates that the forecast uncertainty is relatively low. However, as we go further into the future, the less certain things become, and the lines get farther apart. This suggests that across extended time horizons, forecasts lose some of their certainty. The central tendency of the forecast is depicted by the thicker, darker region in the middle of the fan, while the outer, lighter regions represent the increasing levels of uncertainty



The comparatively narrow spread of the lines at the start of the forecast horizon indicates minimal forecast uncertainty. But as the horizon grows, the lines are farther apart, which indicates that the forecasts are less certain. This expanding spread indicates that we are less confident in the accuracy of the sales estimates the further into the future we project.

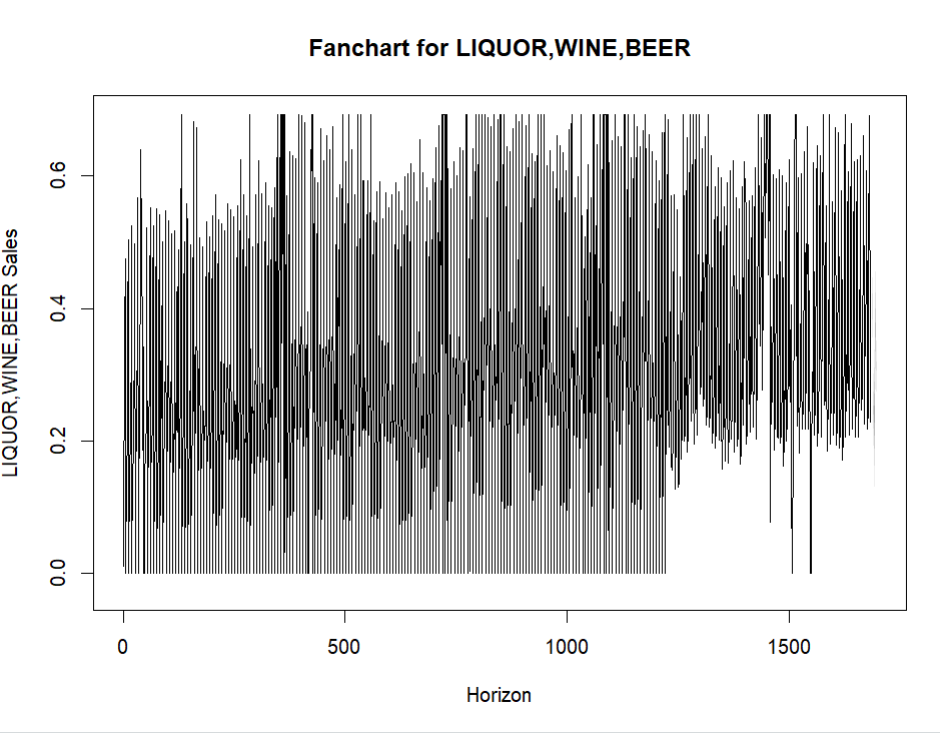


The chart shows some notable spikes at certain points, which might indicate seasonal peaks or events causing sudden increases in sales. These spikes are captured within the fan chart, showing how they contribute to the overall forecast uncertainty.

****

The projected range is initially quite small, which suggests that the short-term projections are less unclear. The range of anticipated values, however, widens as the horizon gets longer, indicating a rise in uncertainty in the long-term projections.

This expanding range suggests that our confidence in the precision of our forecasts declines as we project further into the future. The central tendency of the forecasts, which displays the most likely values for PREPARED FOODS sales, is represented by the dense clustering of lines in the middle of the chart. The fan chart's outside regions, which highlight the fluctuation and unpredictability over time, depict the range within which actual sales are anticipated to fall.

****

The projected range is initially quite small, which suggests that the short-term projections are less unclear. The range of anticipated values, however, widens as the horizon gets longer, indicating a rise in uncertainty in the long-term projections.

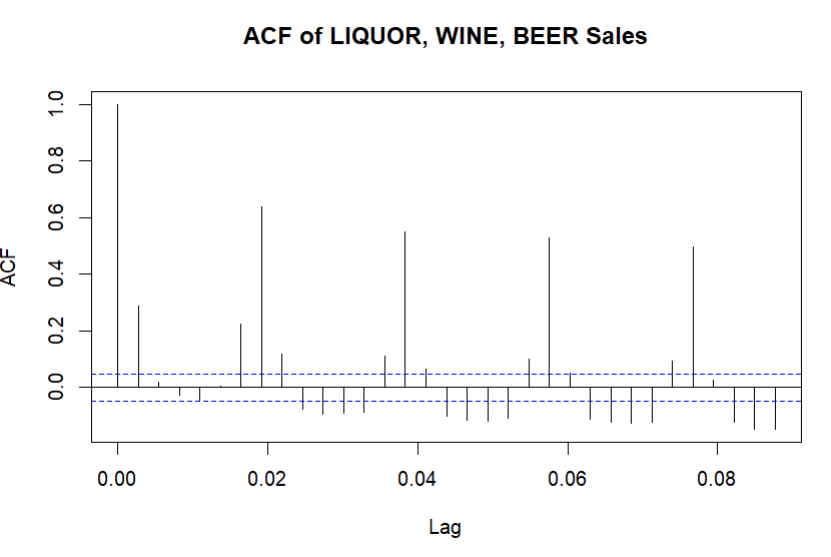
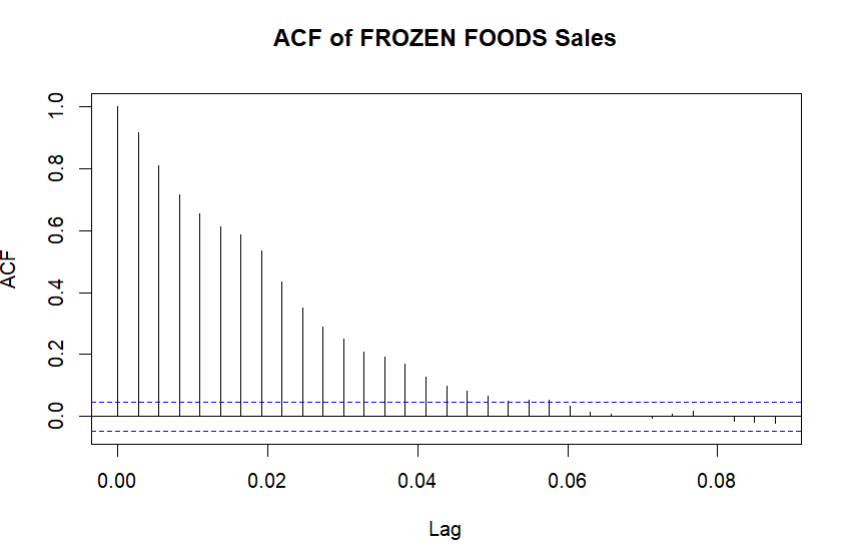
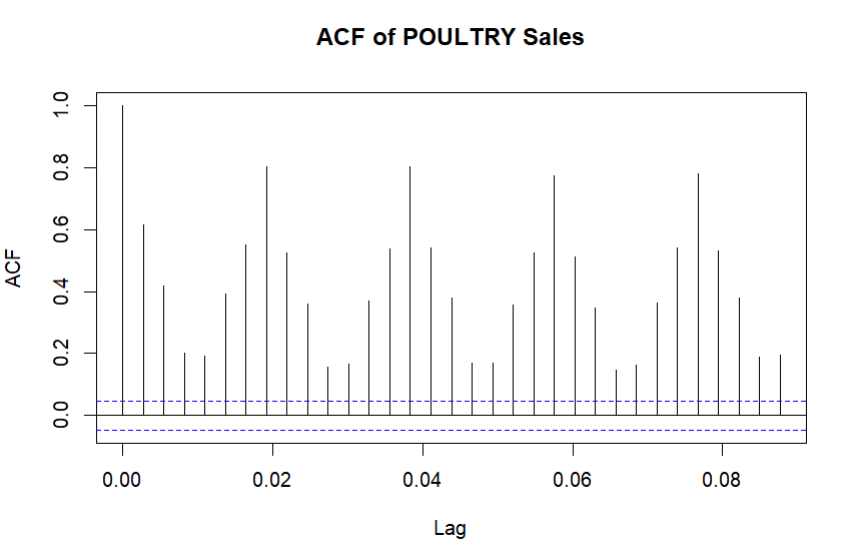
A core trend may be seen in the chart's middle, where a dense cluster of lines indicates the most likely sales numbers. The fan chart's outside edges illustrate the possible fluctuation over time by capturing the larger range within which actual sales are anticipated to fall.

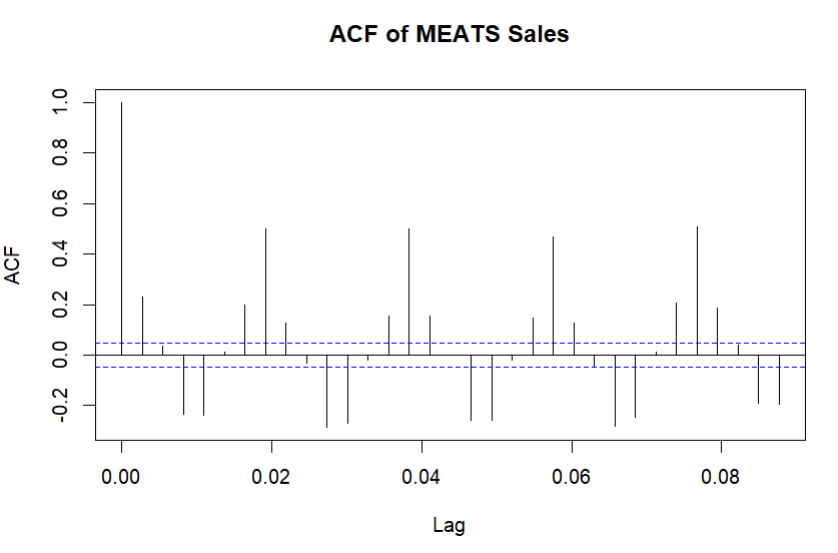
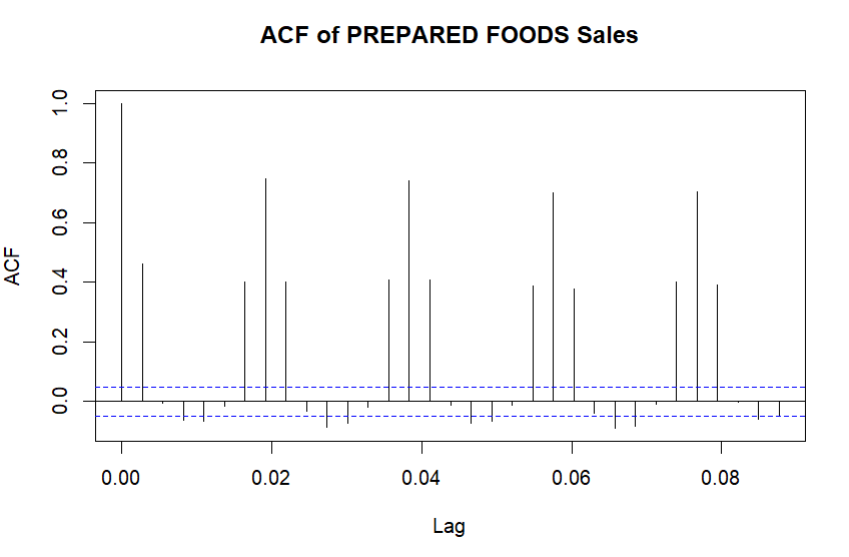
The figure is noteworthy because it highlights times when the predicted range varies more than others. This could mean that there are seasonal variations or important events that have an impact on sales.

# Multivariate Time Series Forecasting

Now for multivariate time series forecasting, we used the time series data to predict sales in the future, 15 days in particular. This was done using 2 different ways. The first one was using the library “MARIMA”, which required us to specify the p, d, and q values on which the model would run on. This meant that in order for the model to predict accurately, we had to first analyze the acf and pacf plots and determine what p and q values to set the model to run on.

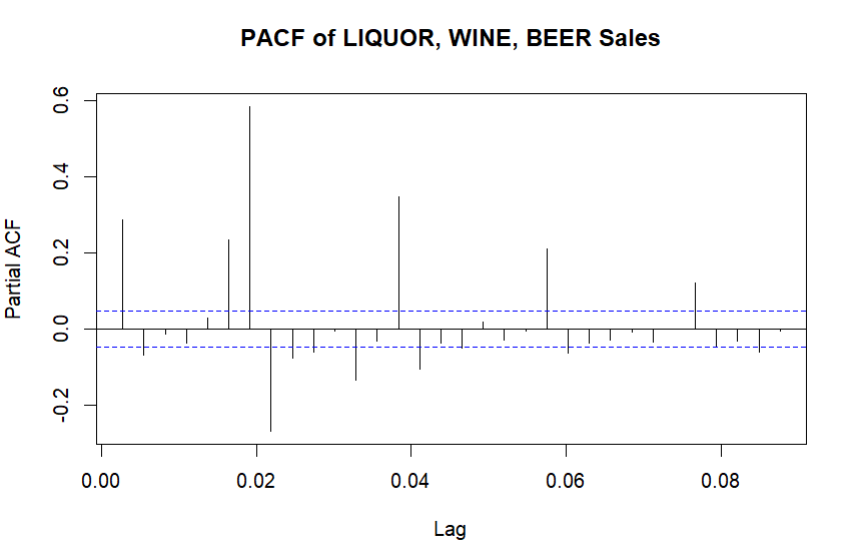
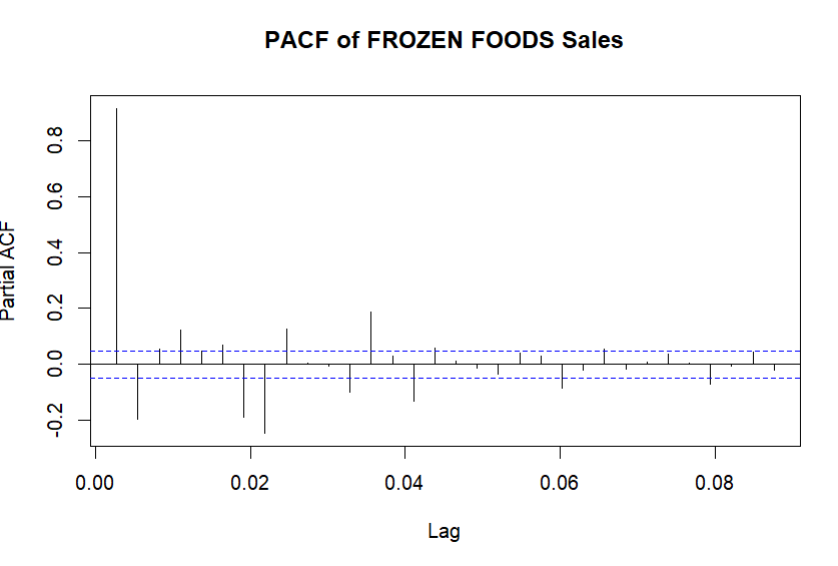
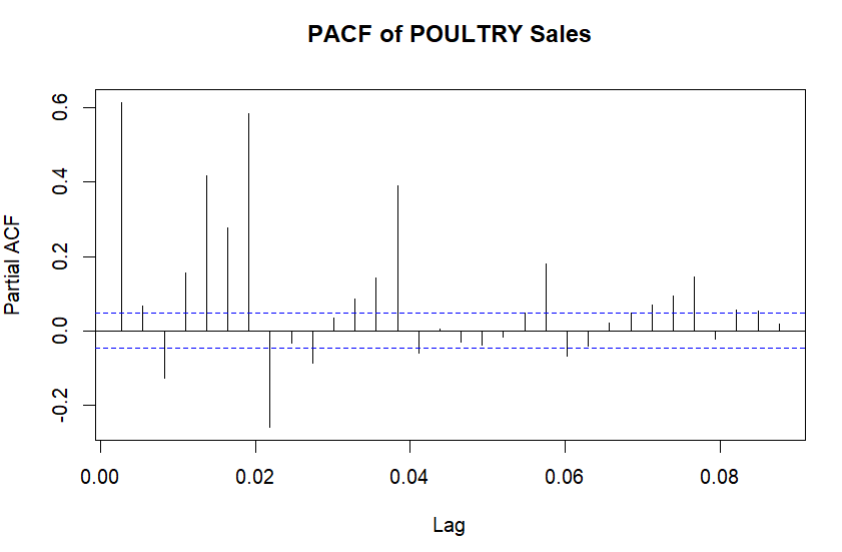
## ACF plots:

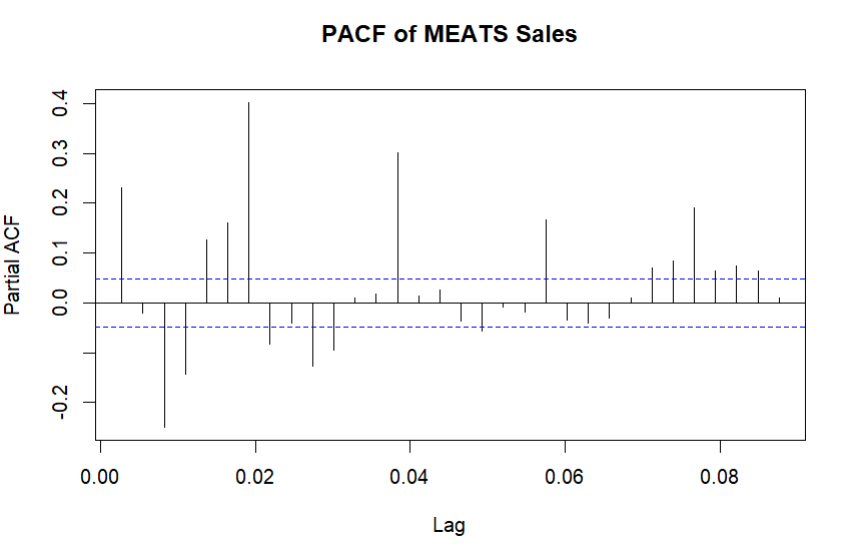
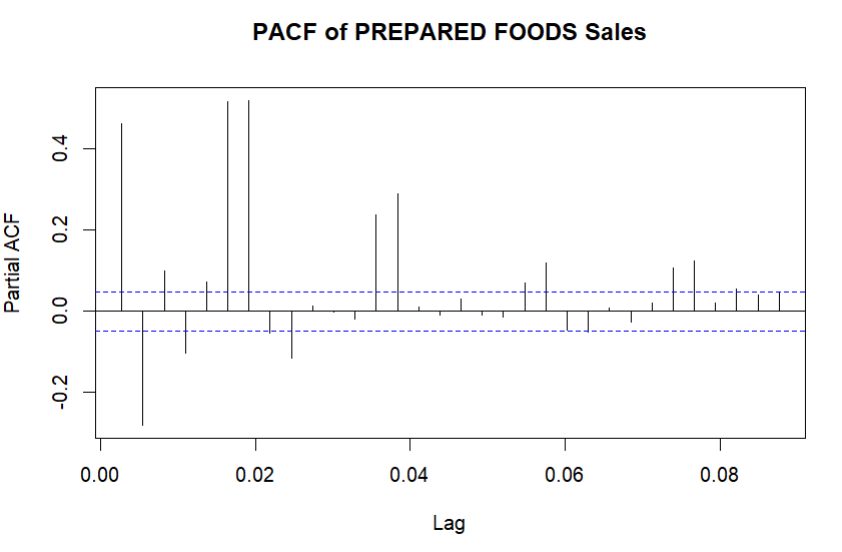




Looking at the acf plots, we can see that on average, the lag where the significant spikes outside the confidence bands usually stop is at the value 2. Which is why we set our q to be 2.

## PACF plots:

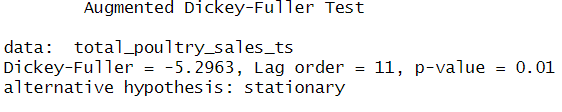


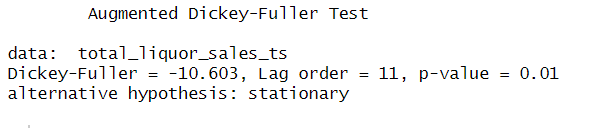
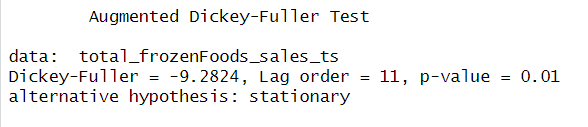


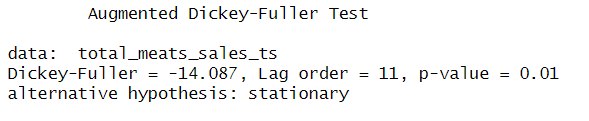
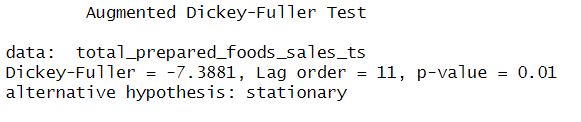
Similarly for the pacf plots, we can see that on average, after the lag 1-2 it usually stabilizes, meaning that we need to set the p value to be either 1 or 2. To determine which value to choose, we ran the model on both p=1 and p=2 and it turned out that p=2 provided the least error, so we went with that.

Finally, when deciding on the d value, which is responsible for ensuring a time series is stationary. We first needed to check if the time series was stationary to begin with. If it was, then our d value will be set to 0 as no more work needs to be done in order for it to be stationary. If it wasn’t stationary, then we will likely need to provide a d value of 1 or greater depending on how non-stationary it is. For this, we used the Dickey-Fuller test to test the null hypothesis, where if the p-value provided by this test is higher than a preset threshold (set to be 0.05), we assume that the timeseries is likely non-stationary.

## Dickey-Fuller test results:



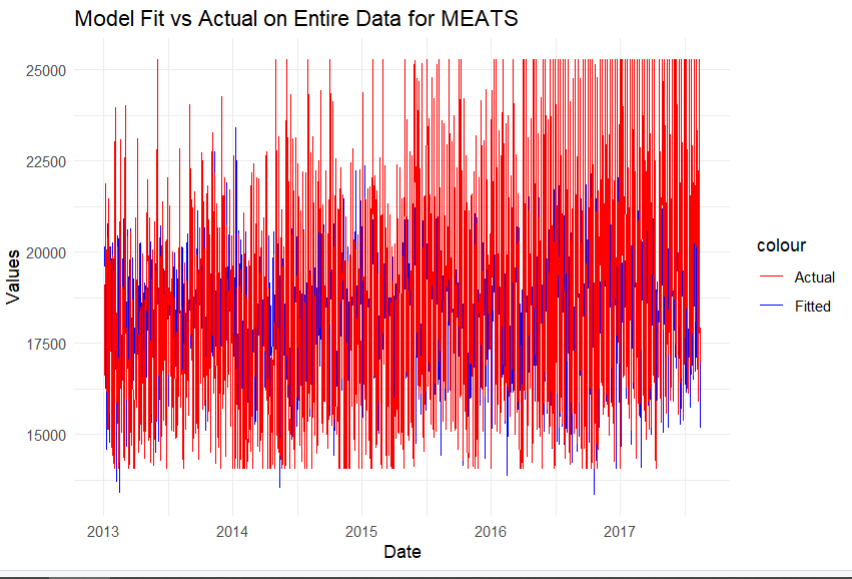
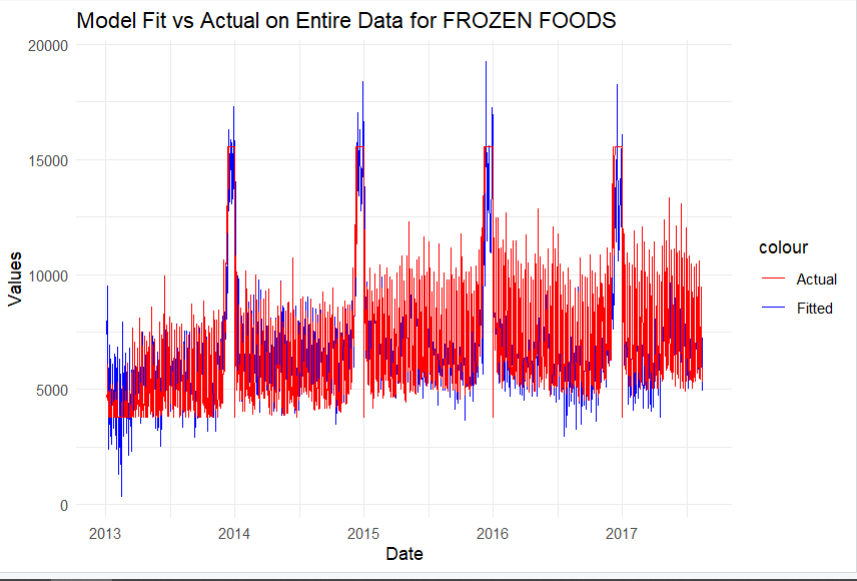




After conducting the Dickey-Fuller test on the total sales time series, as well as the time series containing the sales for each of the product families separately, we can see the null hypothesis is rejected in all cases, meaning in all our time series values, the data is stationary. Therefore, we set our d-value in our arima model to be 0.

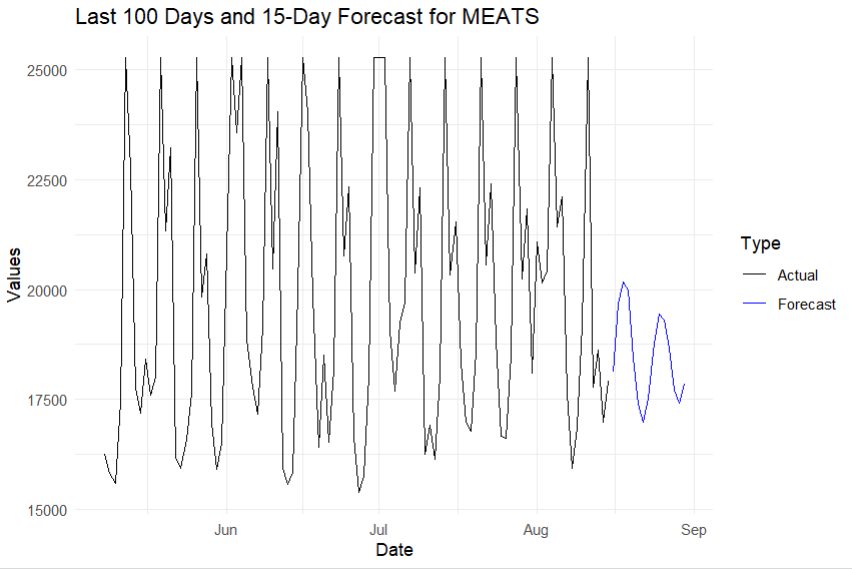
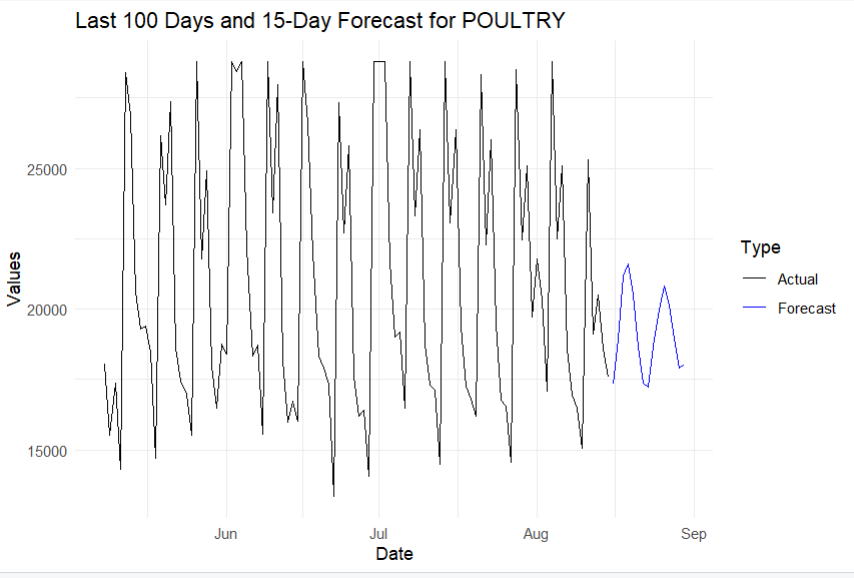
Now that we have decided the parameters for our arima model. We create a multivariate arima model, using “MARIMA”, with p = 2, d = 0, and q = 2. After creating this model, we fit the time series data on this model and compare the fitted data to our original data:

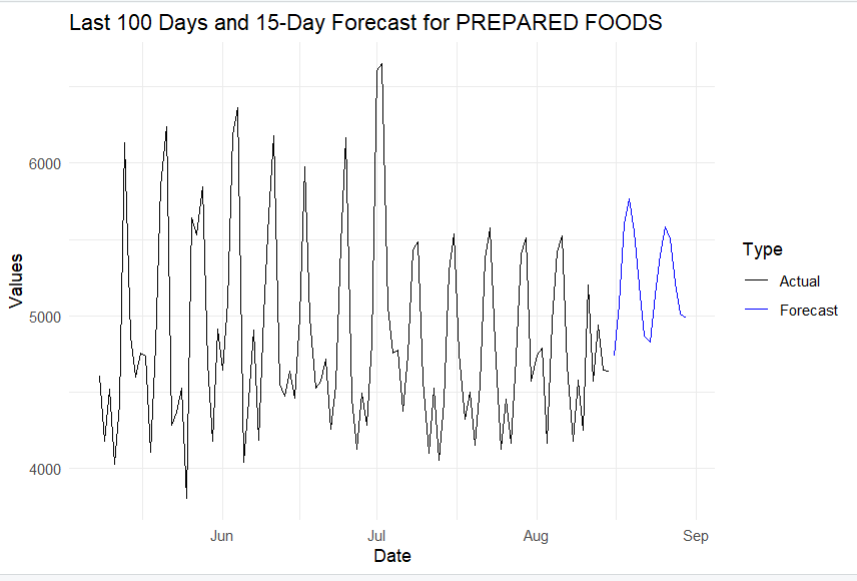
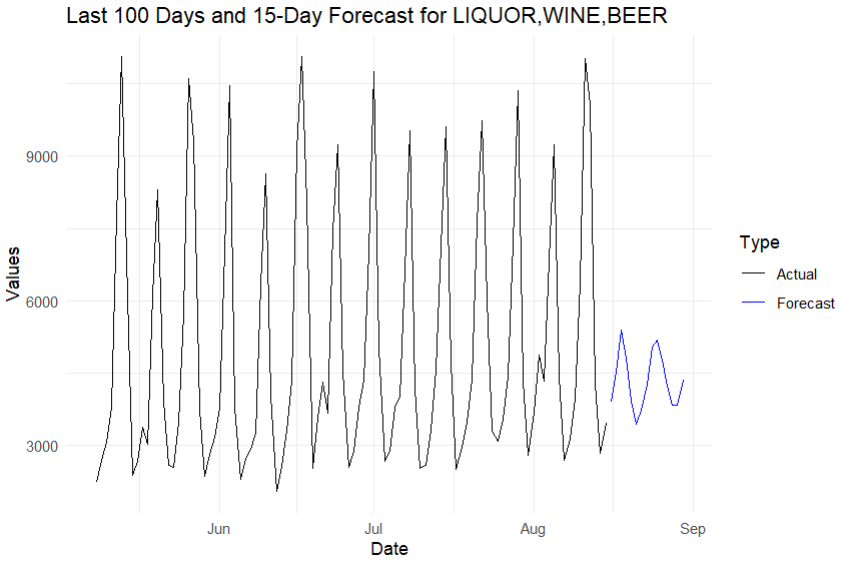
Examples:

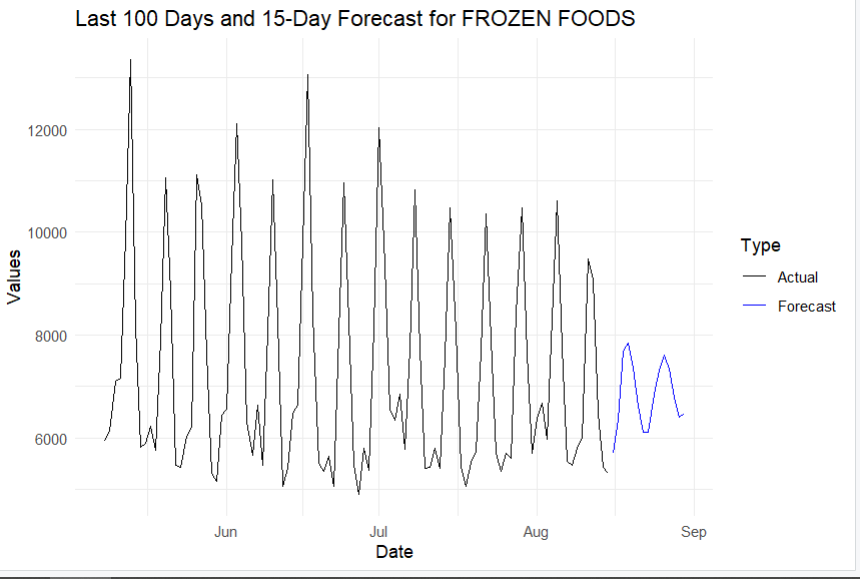


We can see from the examples shown here that the fitted data has more accuracy on some product families than others. For example, the fitted data for frozen foods is better than that of meats. This is because we chose for p and q values the average of all the product families, so it might be the case that some family names have actual p and q values very close to the averages, while others have further p and q values.

Now that we have a model with fitted data, we can use this model to forecast the future sales for the next 15 days. After setting the forecast date to be 15 days and running the forecast, we can plot the forecasted values, as shown here:

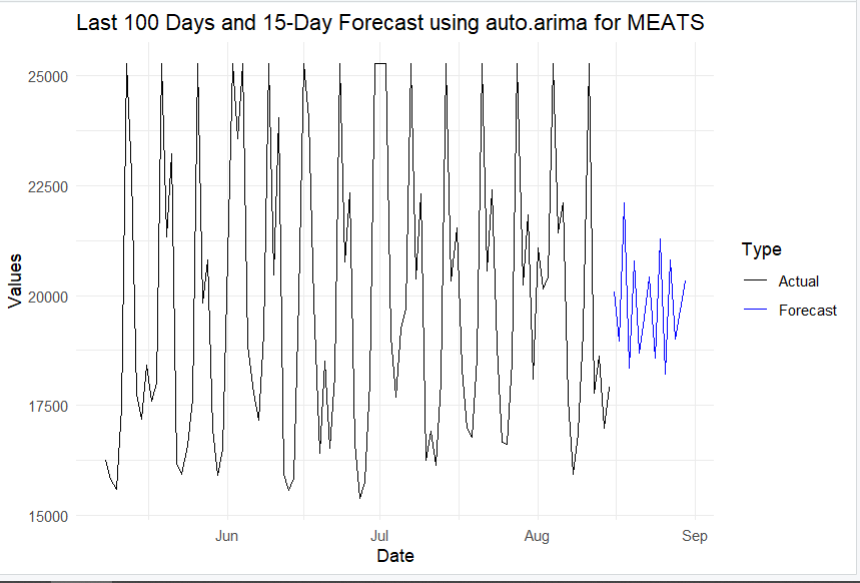
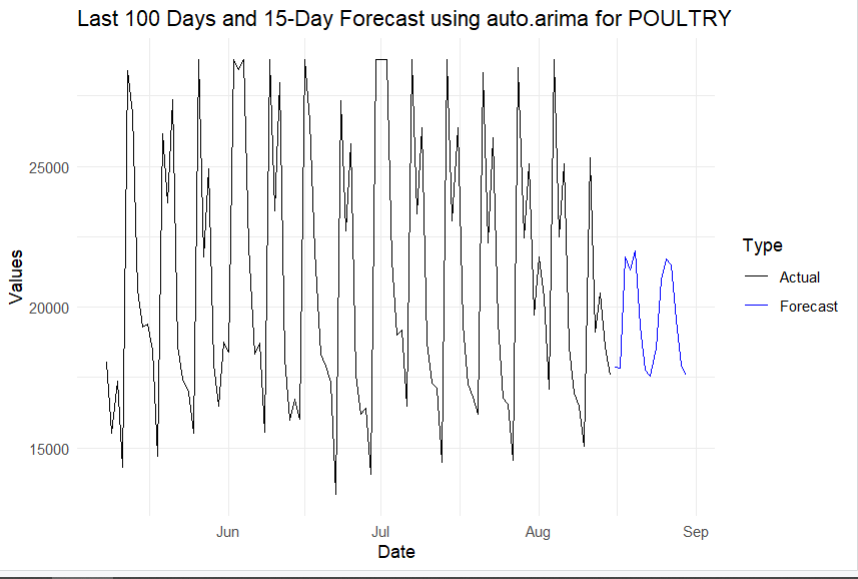


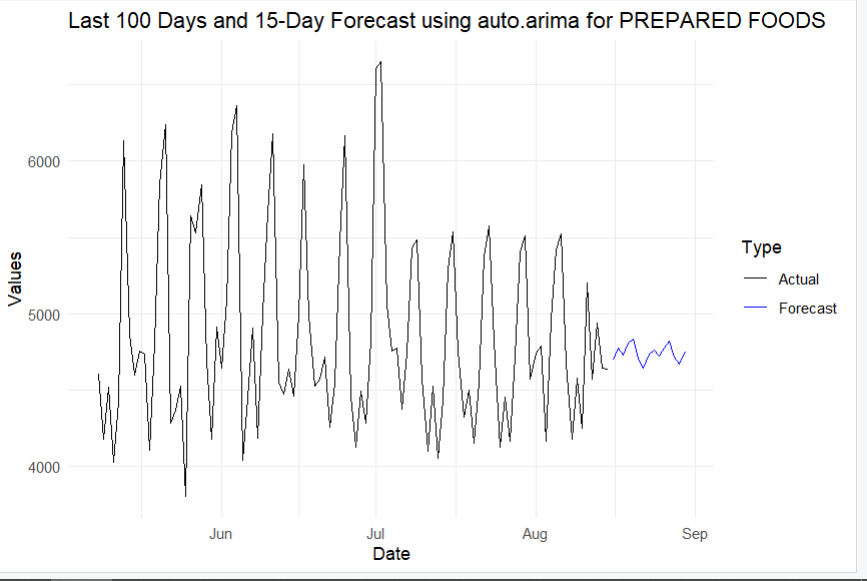
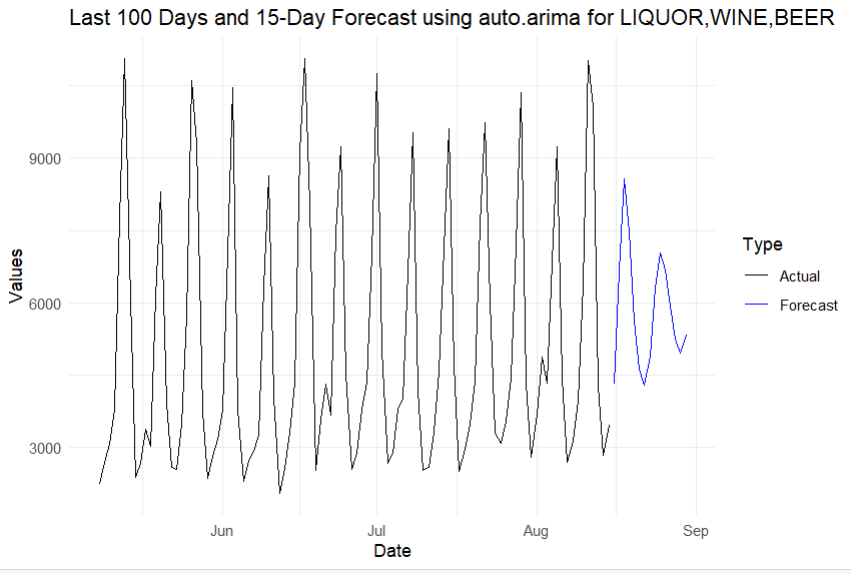


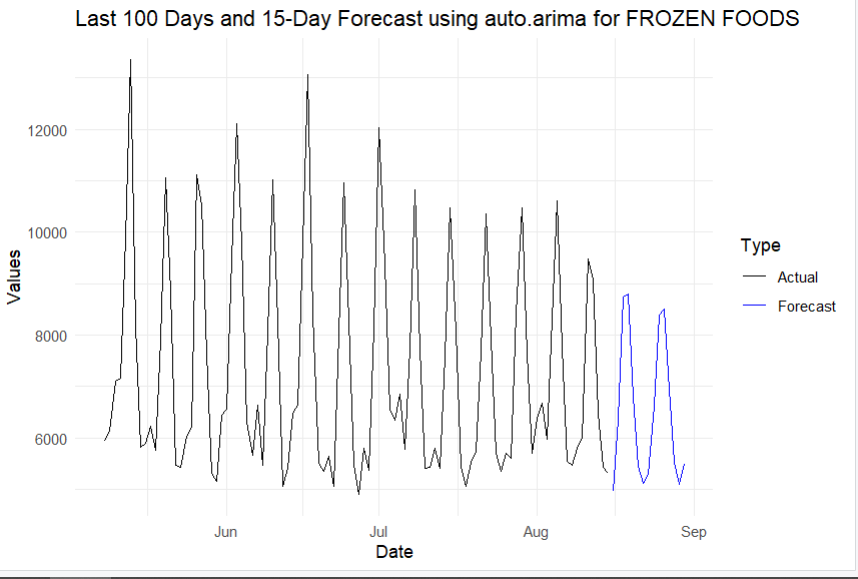


We can see from these plots that the forecasted data has some fluctuations, but some are greater than the rest. We can also see that the predicted values are very alike but on different scales in all plots, likely due to the p and q values being set to the average. In most cases the data stays at the same range, except for prepared foods, where the forecasted data showed an increase from the range of actual data.

Now that we have generated the forecasts using the “MARIMA” library, we can try another approach for multivariate arima forecasting, using auto.arima. This method simplifies the process of time series forecasting by automatically choosing the p, d, and q values in the arima model. Running auto.arima just requires the time series data and the forecast period. After running the auto arima model to generate forecasts, we plot these forecasts as shown here:

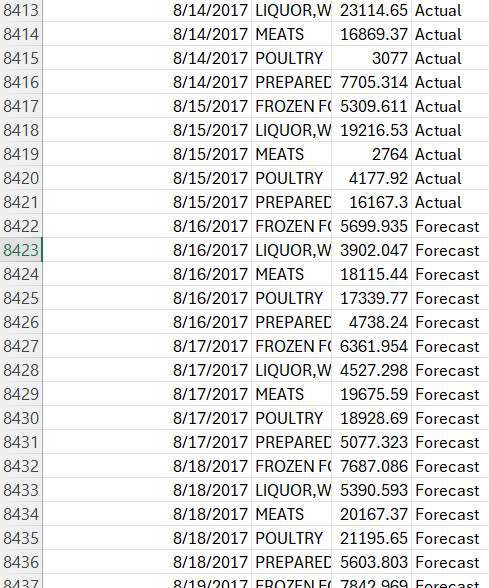






We can see the plots from the auto arima model show similar trends overall, however, since not all the product families are predicted using the same p and q values which are taken from the average like in the MARIMA model, we can see that the fluctuations in the forecasted data is different from one family to another. Since each family runs on its ideal q and p values, the data forecasted varies more than in MARIMA model, but this variation is a good thing since the accuracy should realistically increase if each prediction is made with parameters specific to that prediction, which is the case here.

Now that we have compared our multivariate arima model outputs, we saved the combined dataset, as well as the forecasted values for the next 15 days using the “MARIMA” model in a csv file, with an extra column specifying if this value is “Actual” or “Forecast” for better understanding. Below is a snippet of what the csv file contains:



# Conclusion

In summary, the examination of Favorita shop sales data from 2013 to 2017 indicates noteworthy patterns, the influence of seasonal fluctuations, and a noteworthy structural disruption in April 2016 as a result of the earthquake in Ecuador. The Granger causality tests demonstrate the interdependencies and predictive linkages between product families.

## Is there a trend?

A consistent upward trend across multiple product families is revealed by analyzing sales data from 2013 to 2016, demonstrating steady rise in sales over the period. Every product family's combined annual sales plots demonstrate a consistent upward trend, which is indicative of rising consumer demand and successful marketing tactics. The decomposition study, which reveals a progressive increase in the trend component for each product family—most notably, for frozen foods and liquor, wine, and beer—also supports this tendency. The product families appear to have grown steadily, as indicated by their upward trend. Sales may have peaked at specific times due to seasonal variations and outside events like the 2016 earthquake. Overall, the trend shows a strong pattern of growth in the sales data, indicating that Favorita stores did well in the market during this time.

## Is there a Seasonal effect?

Strong seasonal impacts are seen in the sales data from 2013 to 2016 for a number of different product families, highlighting the important influence of recurrent events on consumer purchase behavior. The time series data's seasonal decomposition shows distinct, dependable patterns that correspond to particular seasons of the year. For example, sales of Frozen Foods and Liquor, Wine, and Beer exhibit notable increases over the holiday season, especially in December, indicating a rise in demand for occasions. Additionally, bi-weekly wage payments to the public sector employees, occurring on the 15th and last day of each month, create noticeable periodic increases in sales, especially evident in Prepared Foods and Meats. These seasonal effects are crucial for understanding sales dynamics as they indicate predictable variations that can be leveraged. The fact that these patterns have remained consistent throughout time emphasizes how crucial it is to include seasonality in analytical models in order to correctly identify and forecast sales trends.

**Poultry & Meats:** Increased consumption during festive periods is reflected in higher sales witnessed at the beginning of the year and during the holiday season in December.

**Frozen Foods:** Towards the end of the year, especially in November and December, there are noticeable increases in sales, which correspond with the holiday season and the likelihood that customers will stock up on frozen goods.

**Prepared Foods:** Displays a distinct seasonal trend, with higher sales at the start and finish of the year.

**Liquor, Wine, and Beer:** Show biweekly surges in line with wage checks and seasonal peaks around the holidays.

**Is there any relationship including a causal relationship in the data?**

The data does, in fact, show correlations and causal relationships. Previous sales of a particular product family can be used to forecast future sales of other families, according to the Granger causality tests. Sales of meats, alcohol, prepared foods, and frozen foods, for instance, can be Granger-caused by sales of chicken, meaning that fluctuations in poultry sales can be used to forecast fluctuations in these other product categories.

There are a number of significant distinctions between the MARIMA and auto.arima models and the Vector Autoregression (VAR) model. Understanding the interdependencies and dynamic interactions between different time series is essential for comprehending how shocks in one series, such sales of a specific product family, affect other series. VAR models are especially good at capturing these relationships. This makes VAR a great option for doing impulse response analysis to assess the effects of occurrences like the 2016 earthquake and for examining the relationships between various product families. MARIMA and auto.arima models, on the other hand, fit individual series with optimal parameters with a greater emphasis on precise prediction. While MARIMA permits human parameter specification, auto.arima chooses the optimal parameters on its own, improving forecast accuracy for every product family. They might not, however, be as good as VAR at capturing the interdependencies between series. In conclusion, VAR is better if the objective is to comprehend and simulate the dynamic relationships and affects between numerous series. Because auto.arima offers customized parameter selection, it is suggested for obtaining the most accurate individual series forecasts, particularly in cases where the interdependencies are not as important.