Usable AI :Final Project Report

Topic : Time series analysis on Grocery sales trend

Team : Mitali Tavildar, Shefali Luley

ARGUMENT:

Problem Statement:

The main objective of the project is to understand the implementation of the time series model and apply it in an appropriate context. The dataset that we have chosen contains details of purchase of a variety of grocery items over time & we aim to utilize this dataset to predict the demand for these different products in future. This will help manage inventory and increase efficiency for the distributors.

DATA COLLECTION & MANAGEMENT:

- Link to the original dataset: https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset

- Link to the Python notebook: https://colab.research.google.com/drive/1EVwpCloUHRMIQgHg63ytQqZaj2m0Wq8G#scrollTo=70l2vzTo3kFY

Data Description:

- The data which will be utilized throughout this project is of Groceries dataset.
- There are 38,765 rows in the dataset that contain purchase orders made by customers at grocery stores.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38765 entries, 0 to 38764
Data columns (total 3 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Member_number    38765 non-null  int64 
 1   Date             38765 non-null  object
 2   itemDescription  38765 non-null  object
dtypes: int64(1), object(2)
memory usage: 908.7+ KB
```

Figure 1

To ensure effective organization and formatting of data, we employ data frames as a means of storage. Leveraging Python libraries such as pandas, we have successfully achieved efficient data management.

Columns:

- Member_number: Unique id of the customer which is of the data type ‘int’
- Date: Provides the date of purchase for the corresponding commodity. This is of the data type ‘object’
- itemDescription: It is the description/name of the commodity being purchased. This is in text format with the datatype ‘object’.

DATA ANALYSIS:

Observations from the data:

- The data we have covers commodities purchased for the year 2014 and 2015.
- There are a total of 38765 rows, and 3 columns. Out of which,there are a total 167 unique products in the itemDescription column.

Data preprocessing/Feature Engineering:

- The data was in the form of a csv, which using Pandas, we were able to import into a dataframe.
- We observed in the data description(Figure 1), that there are no null values.

- Since there are 167 unique values, and felt the need to categorize them into different food product types that they belong to.
    - We engineered a new column called ‘Category’ that broadly described the kind of the product.
    - For example, ‘chicken’, ‘beef’ and ‘sausage’ would be classified as ‘Meat’.
- There were missing dates present in the data. We were able to generate records for these dates using the reindex() method and fill their corresponding count values to zero.

Summary statistics:

- The data present is for the items purchased between the dates ‘01-01-2014’and ‘31-10-2015
- We listed out the frequency of the categories of the products based on the count in Figure 2.
- The visual representation of the same can be seen in Figure 3. Through the heatmap,we observed that Dairy, Beverages and Vegetables are some of the categories with the highest bought products in the last 2 years.

Number	Category	Count
1	Dairy	5725
2	Beverages	5083
3	Vegetables	4572
4	Meat	4313
5	Sweets	4064
6	Bakery	3417
7	Fruits	3167
8	Condiments	2910
9	Kitchen supplies	2006
10	Misc products	1913

Figure 2

Figure 3

- Figure 4 represents the count of each product category in the year 2014 and 2015 separately. This plot provides an overview of the demand of each category year wise.

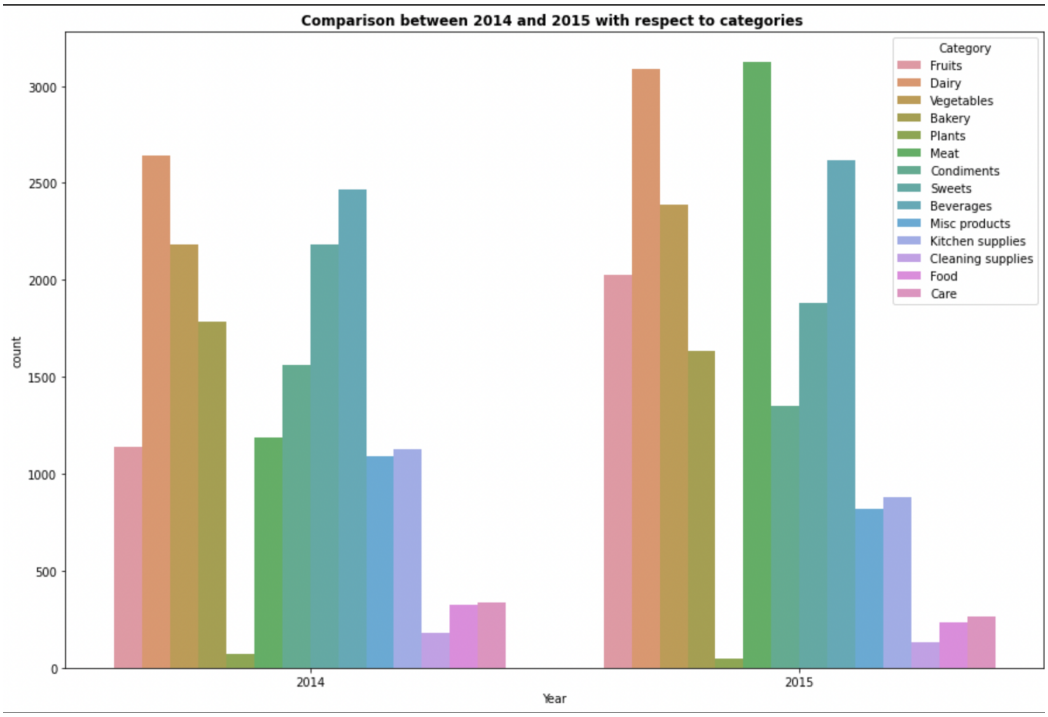


Figure 4

Basic visualizations for data analysis:

- In order to forecast the demand for each category, we needed to understand and analyze the demand of the historical data.
- We began analyzing if there exists any trend in the demand pattern from a very high i.e, yearly level to a very low, i.e daily basis.
- Figure 4 clearly shows the yearly demand for each category. We observe the count for meat,dairy having highest demands.
- In Figure 5, we then drilled down to analyzing monthly trends for demand for each country. We observe a noticeable difference in the range of count for each category.

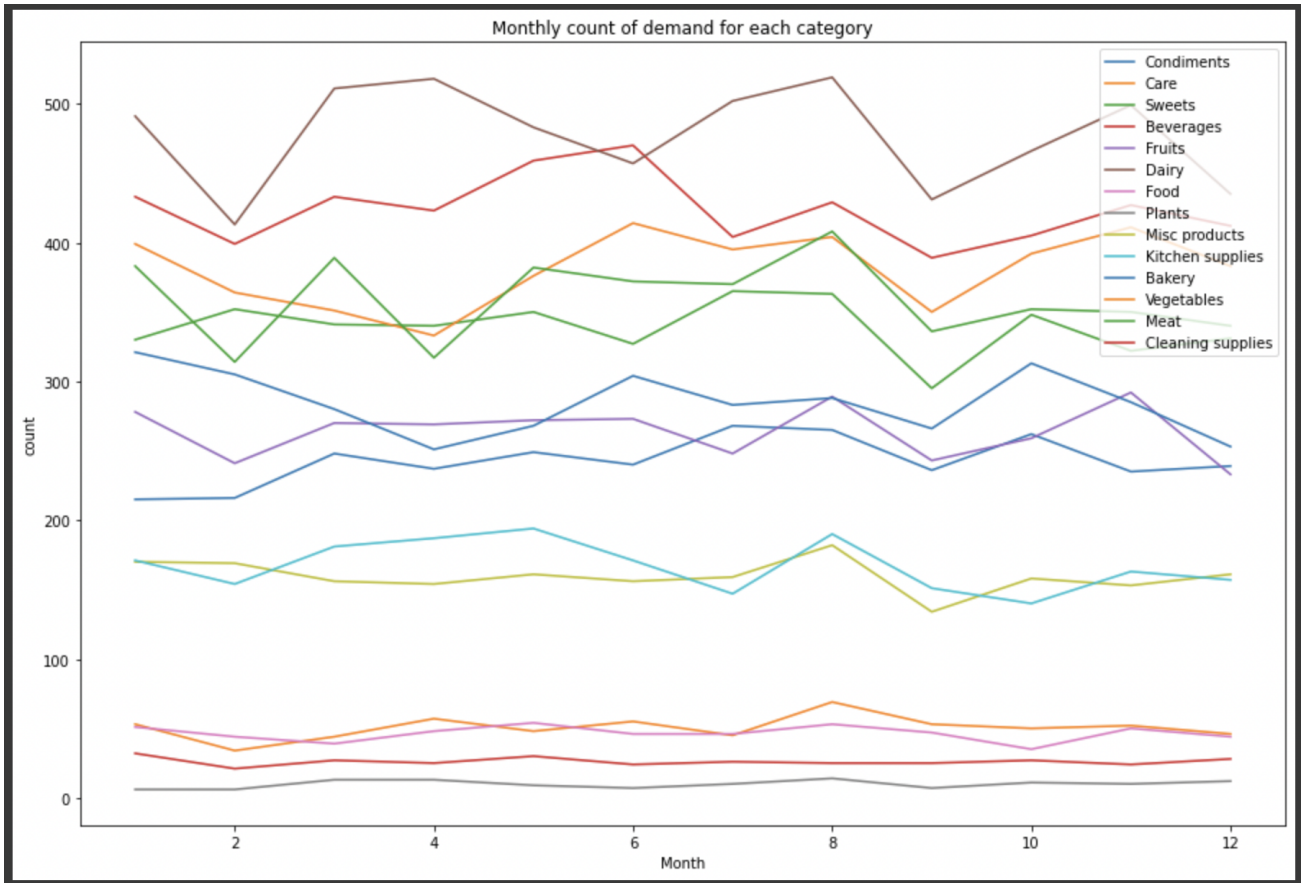


Figure 5

- In Figure 6, we drilled further down and analyzed a plot for demand over 2 years for each category.
- Figure 6a represents these demands for the year 2014, similarly Figure 6b represents these demands for the year 2015.
- We observe an overlap in certain categories, and notice the demand trends to be similar for these.
- In figure 7,The graph above indicates a consistent pattern in daily sales from 2014 to 2015, with the exception of a dip in January 2015.
- The cause of this dip is uncertain and may be attributed to various factors such as inflation, among other factors.

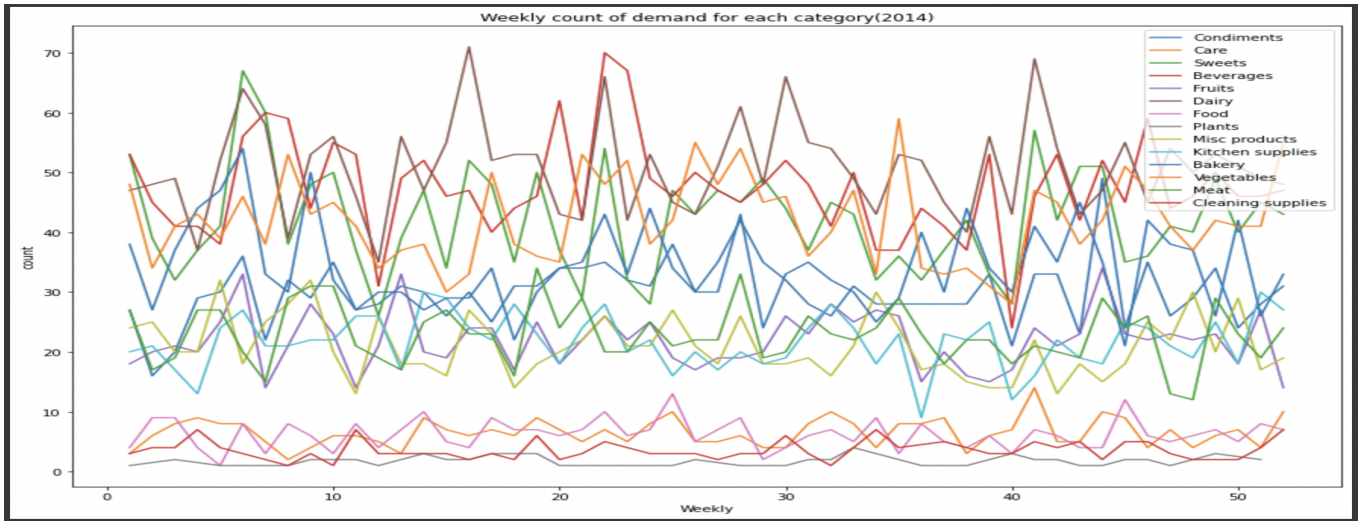


Figure 6a

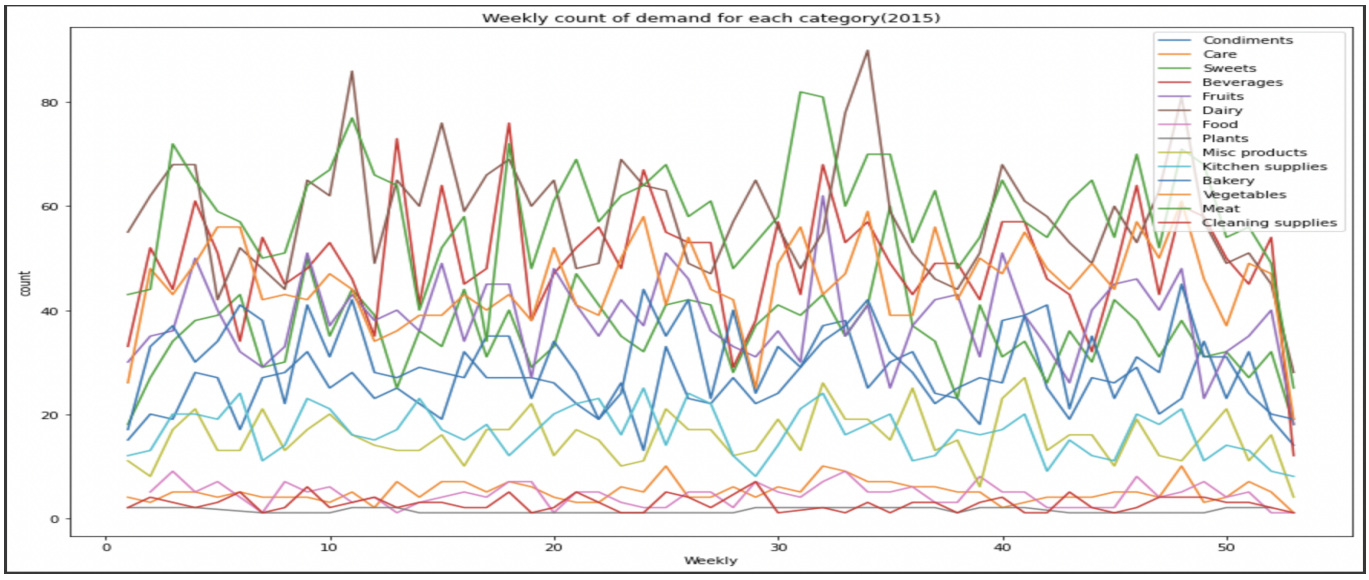


Figure 6b

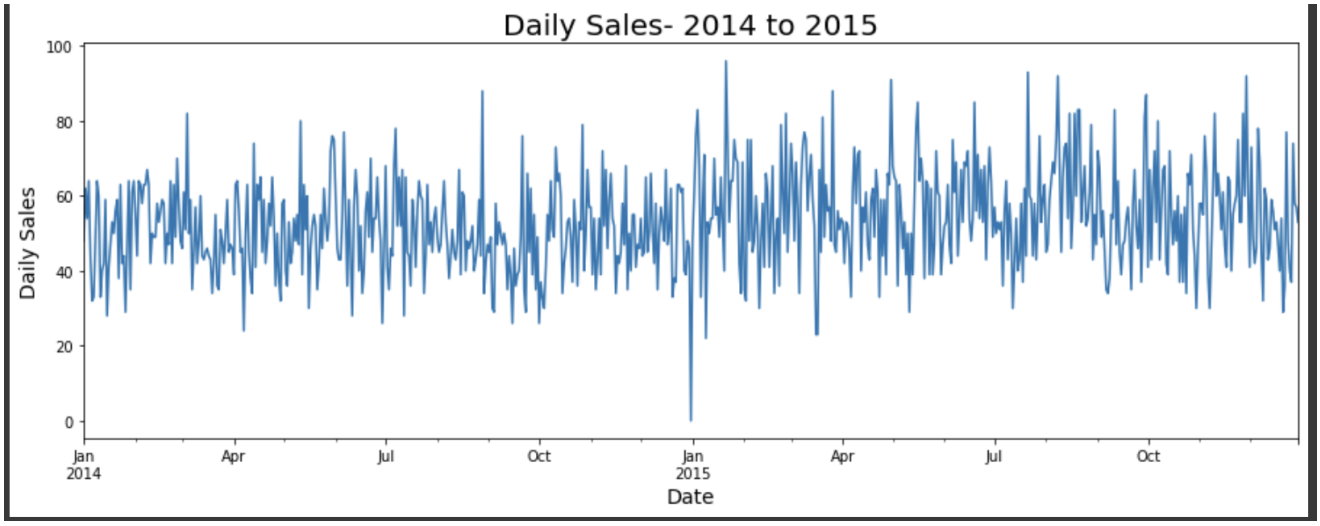


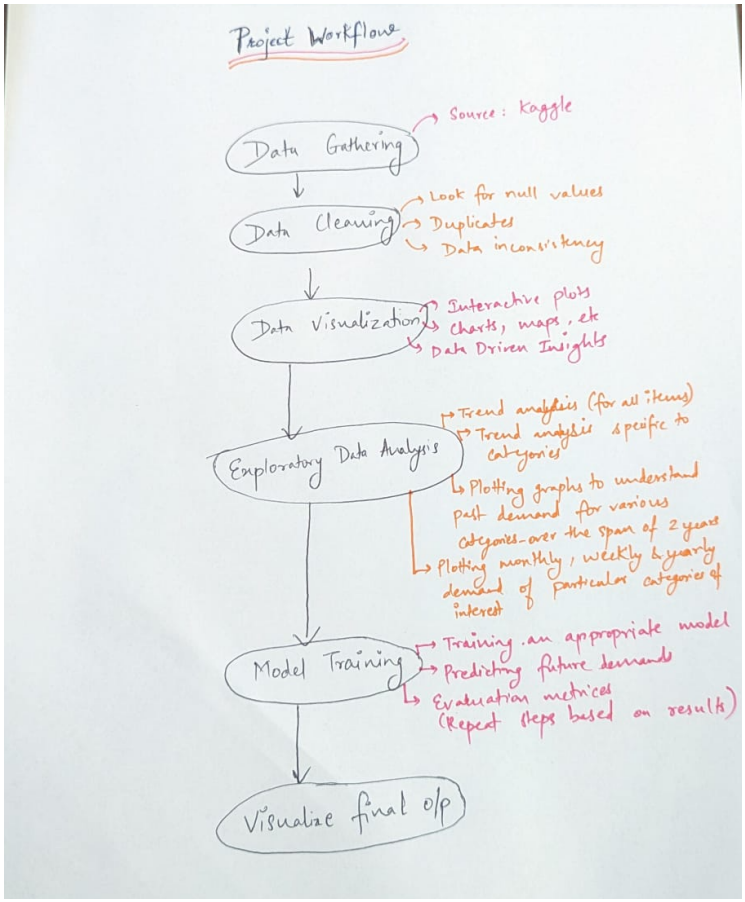
Figure 7

Result:

- In all the above graphs and statistical analysis, where we are trying to look into variation in demand of different categories over months, weeks and on a day to daily basis.
- It can be observed that there are few factors and categories which have a significant amount of dominance on the whole patterns and trends.

Research and Design Questions:

- Are there seasonal or trend-based patterns in customers' purchasing behavior that can be identified and accounted for in a forecasting model?
- What is the most appropriate forecasting method for this type of data (e.g. time series analysis using ARIMA, SARIMA,SARIMAX), and how can its accuracy be evaluated?
- How can data preprocessing techniques such as feature scaling, outlier detection, or missing value imputation improve the accuracy of a forecasting model?
- What are the trade-offs between a simpler, more interpretable forecasting model and a more complex, potentially more accurate model, and how can these trade-offs be optimized for the specific business problem at hand?
- Experimenting with the timeframe factor - whether to consider daily or monthly data makes a difference in the accuracy of the predictions.
- Using automated models/libraries and different methods to find the most accurate predictions.



INTERVENTION

Expected/Target behavior of our intervention

For stakeholders who are interested in conducting market research such as surveys,etc can use this data to analyze trends and patterns in the retail industry. This will also give them a benefit where they can develop further strategies to stay ahead of their competitiors.As we said, " develop strategies", stakeholders can also consider investing in other retail industries to evaluate financial performance and growth potential of other companies like Target, Kroger etc. They can also identify the potential areas of growth or expansion ,evaluate the feasibility of entering new markets

Form of delivery of our intervention

The sharing of information solely depends on the preference of the retailers/shareholders. Some of the information might be shared in forms of reports, presentations, some data visualizations showcasing the trends and patterns , in terms of differences in every year, month and more. Another way could be accessing a remote/ online available database that provides access to chunks of information from that company where you can provide stakeholders to manipulate and analyze the data.

Following is the design solution and detailed design,description and visualizations of our intervention:

SECTION I :

Focusing further on the research and design questions:

- Are there seasonal or trend-based patterns in customers' purchasing behavior that can be identified and accounted for in a forecasting model?
- What is the most appropriate forecasting method for this type of data (e.g. time series analysis using ARIMA, SARIMA,SARIMAX), and how can its accuracy be evaluated?
- How can data preprocessing techniques such as feature scaling, outlier detection, or missing value imputation improve the accuracy of a forecasting model?

DESIGN

Approach :

In this section, our focus will be on the three primary categories with the highest demand, as identified in the previous analysis. These categories include:

- Beverages
- Dairy
- Vegetables

To gain insight into the seasonal and trend-based patterns within these categories, we have employed time series analysis. Our initial step involves reading the original data and

subsequently plotting the monthly sales data. We then proceed to decompose the time series into its constituent components: trend, seasonal, and residual. Below are the graphical representations of our implemented analysis:

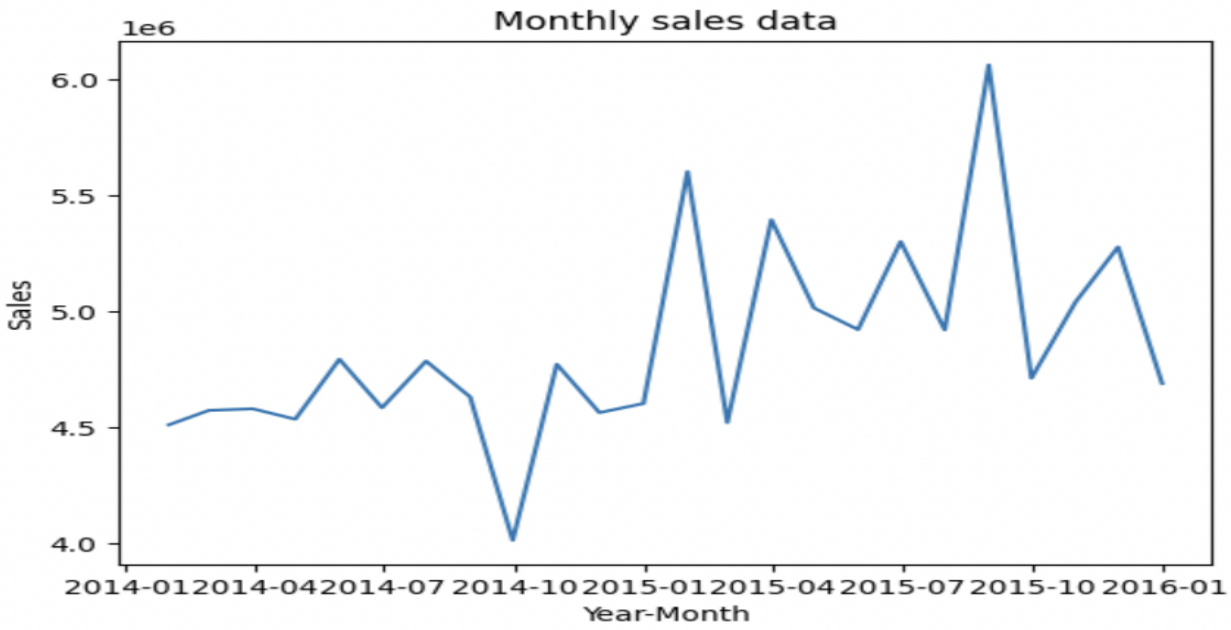


Figure 8

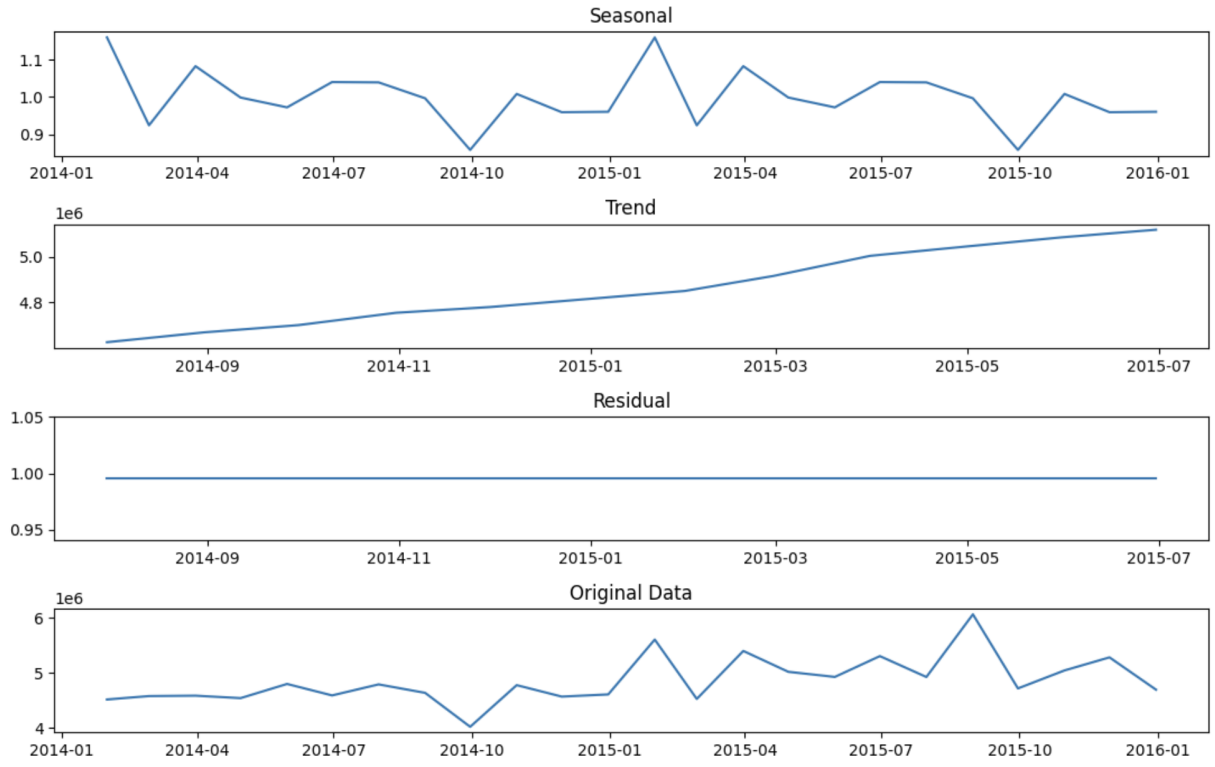


Figure 9

During the decomposition process as seen in Figure 9, we extract three essential elements from the time series data:

- Trend: This component represents the overall direction or tendency of the series over a span of 24 months (2 years) for all the available data items. It helps us identify any long-term changes or patterns.
- Seasonality: Seasonality refers to a recurring pattern that occurs at regular intervals due to various seasonal factors. In our analysis, we observed a monthly seasonality, indicating predictable variations in the data that repeat monthly.
- Residual: The residual component captures the irregular and unpredictable fluctuations in the time series data that cannot be explained by the trend or seasonal patterns. It represents the random or residual variability left after removing the trend and seasonality components.

Decomposition Analysis for top 3 Category (Monthly)

In order to conduct a more comprehensive trend analysis of the top three categories, we further decomposed the sales trend of these corresponding categories and analyzed the same. The time frame we chose initially for monthly data. The following the results :-

Analysis for Beverages category (Monthly)

Figure 10 depicts the demand for beverages during the years 2014 and 2015. Through the analysis of these graphs, a noticeable trend can be observed, indicating a potential dip in demand during the months of March to May and around September to October. These periods suggest a possible seasonality or external factors influencing consumer behavior within the given timeframes. This insight provides a general understanding of the fluctuations in beverage demand during the specified years, aiding in future forecasting and strategic decision-making for businesses operating in the beverage industry.

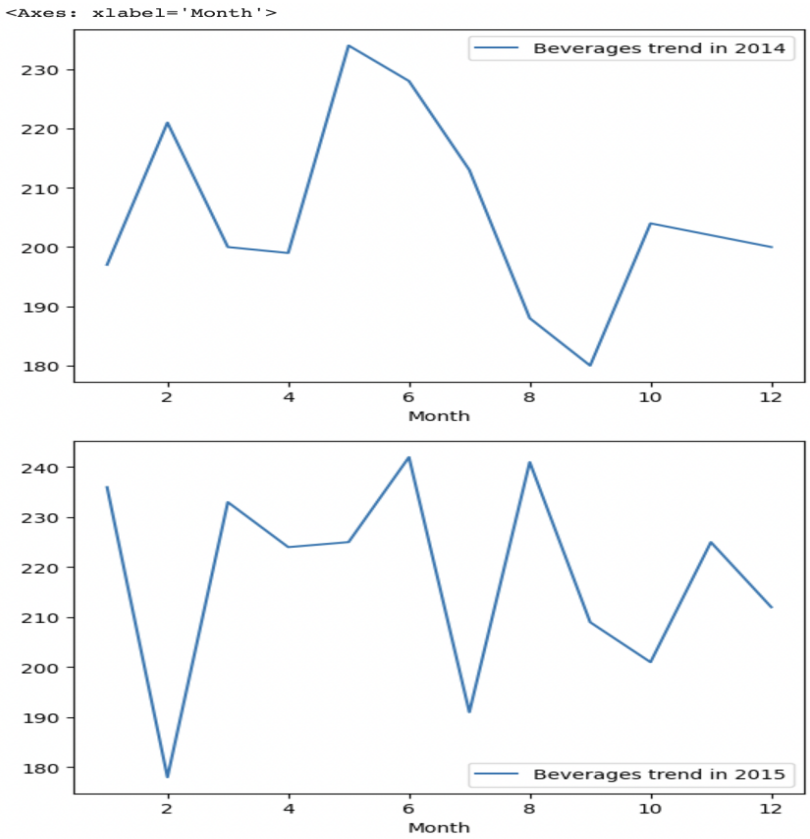


Figure 10

Next, we then proceed to decompose the time series into its constituent components: trend, seasonal, and residual for the category “Beverages”. Figure 11 is the graphical representations of our implemented analysis:

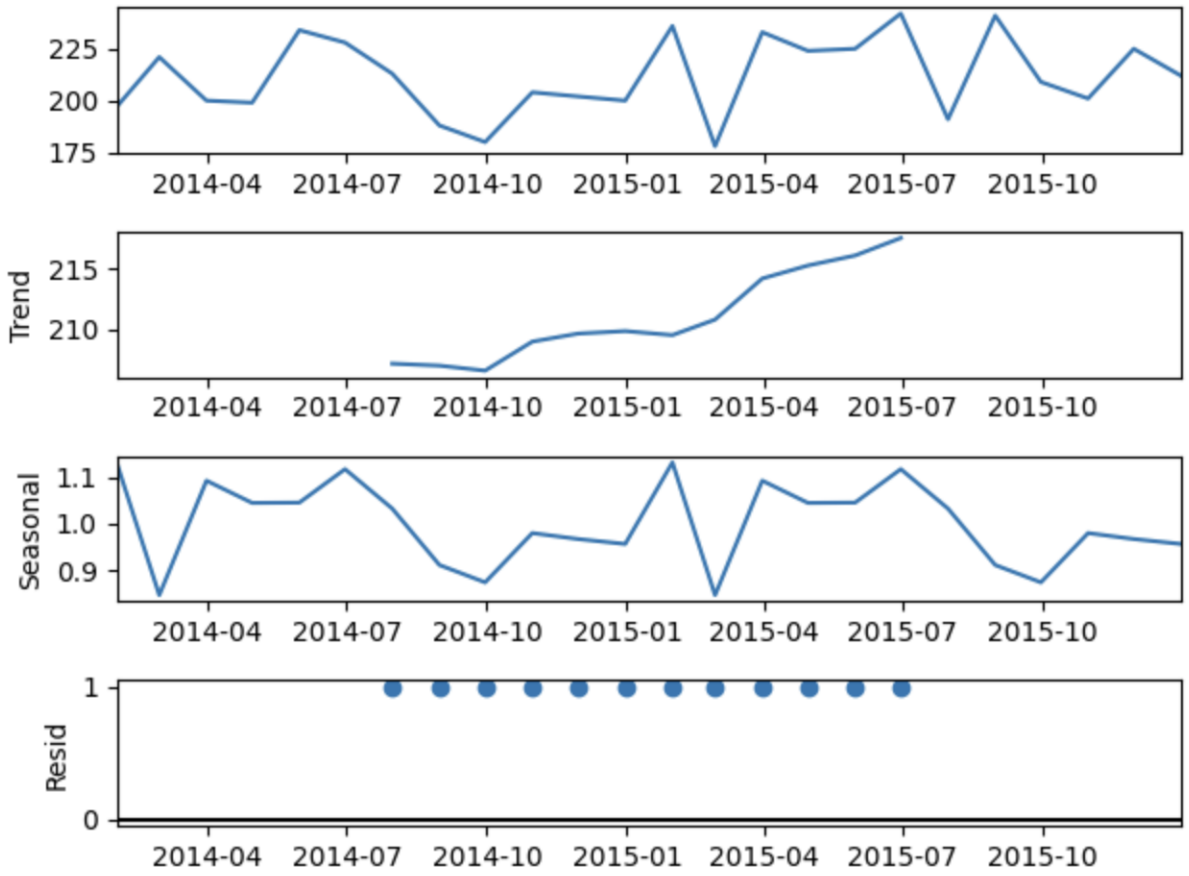


Figure 11

The decomposition graph presented in Figure 11 further reinforces the observed trends in monthly demand. However, it is important to note that the seasonal pattern identified may not be entirely reliable due to the limited historical data available, spanning only the past two years. With a relatively short time frame, the seasonal component may not capture all potential seasonal patterns accurately. Therefore, it is crucial to exercise caution when interpreting and relying solely on the observed seasonal pattern, considering the limited historical data available for analysis. Supplementing the analysis with additional data or considering other factors could

provide a more comprehensive and accurate understanding of the seasonality in the demand for beverages.

Analysis for Dairy Category(Monthly)

To conduct a comprehensive trend analysis, we specifically focused on the "Dairy" category within a monthly time frame. Our analysis involved sorting the data and filtering out records exclusively related to dairy, thereby isolating the relevant data for further examination. We then created separate data frames for the years 2014 and 2015, allowing us to plot the demand for dairy during these specific periods(Figure 12). By visualizing and analyzing the demand trends within the "Dairy" category during 2014 and 2015, we gained valuable insights into the consumption patterns and trends specific to this category. This approach facilitates a deeper understanding of the dynamics within the dairy market during the specified time frame.

Figure 12

Similarly to the analysis conducted for the beverages category, we attempted to understand the overall trend in demand for the Dairy category. However, it's important to note that the available data for only a two-year period may not be sufficient to make definitive generalizations. The demand patterns within the Dairy category can be influenced by various factors that may vary over time. Therefore, while the analysis provides insights into the observed trends during the specific two-year period, caution should be exercised when extrapolating these findings to make broader conclusions about the long-term demand patterns for Dairy products. Supplementing the analysis with more extensive and diverse historical data would help to improve the accuracy and reliability of any generalizations made.

Next, we then proceed to decompose the time series into its constituent components: trend, seasonal, and residual for the category “Dairy”.Below is the graphical representations of our implemented analysis:

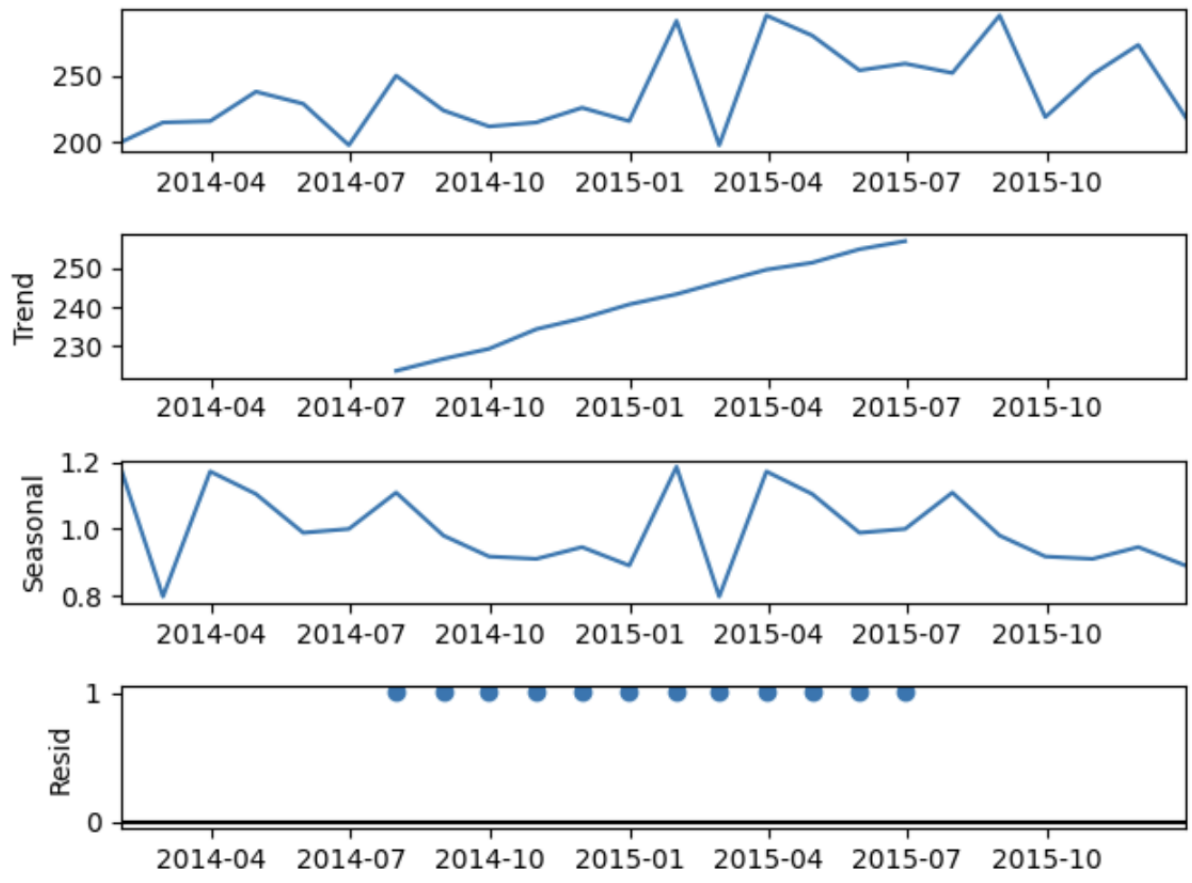


Figure 13

The decomposition of the demand for the Dairy category,in Figure 13, reveals a general positive trend over time. This indicates an overall increasing demand for dairy products. Additionally, the seasonal pattern analysis reveals a noticeable dip in demand around the month of March. This observation suggests a potential seasonality in the consumption of dairy products, with a decrease in demand during that specific period. Understanding these trends and seasonal patterns can assist businesses in optimizing their production and marketing strategies, taking into account the fluctuating demand patterns throughout the year. However, it is important to note that these insights are based on the available data and may not necessarily capture all underlying factors influencing the trends and seasonality in the Dairy category.

Analysis for Vegetable Category(Monthly)

We conducted a comprehensive trend analysis specifically focused on the "Vegetable" category within a monthly time frame. By sorting and filtering the data, we isolated records related to vegetables for further examination. Separate data frames were created for the years 2014 and 2015, enabling us to plot the demand for vegetables during those periods. This analysis provided valuable insights(Figure 14) into the consumption patterns and trends specific to the vegetable category, deepening our understanding of the market dynamics within the specified timeframe.

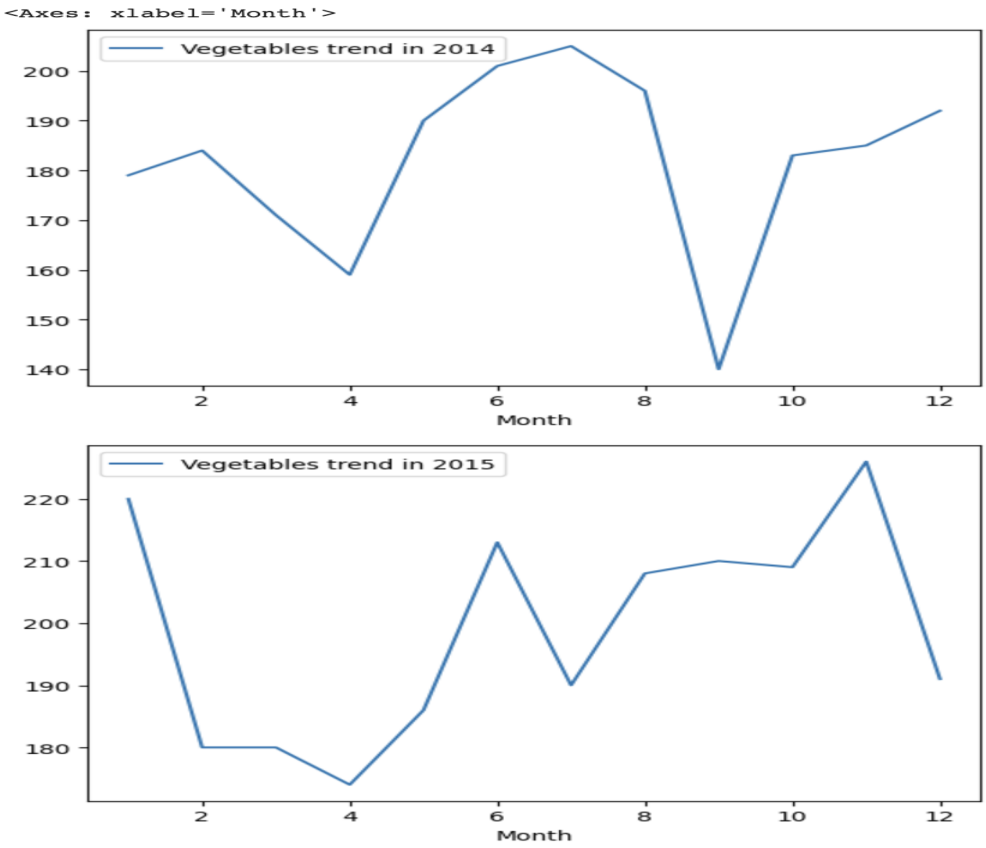


Figure 14

Similar to the previous categories, our analysis aimed to understand the overall trend in demand for the Vegetables category. By examining the data and conducting trend analysis, we sought to gain insights into the general patterns and changes in demand for vegetables.

Next, we then proceed to decompose the time series into its constituent components: trend, seasonal, and residual for the category “Vegetables”.Below is the graphical representations of our implemented analysis:

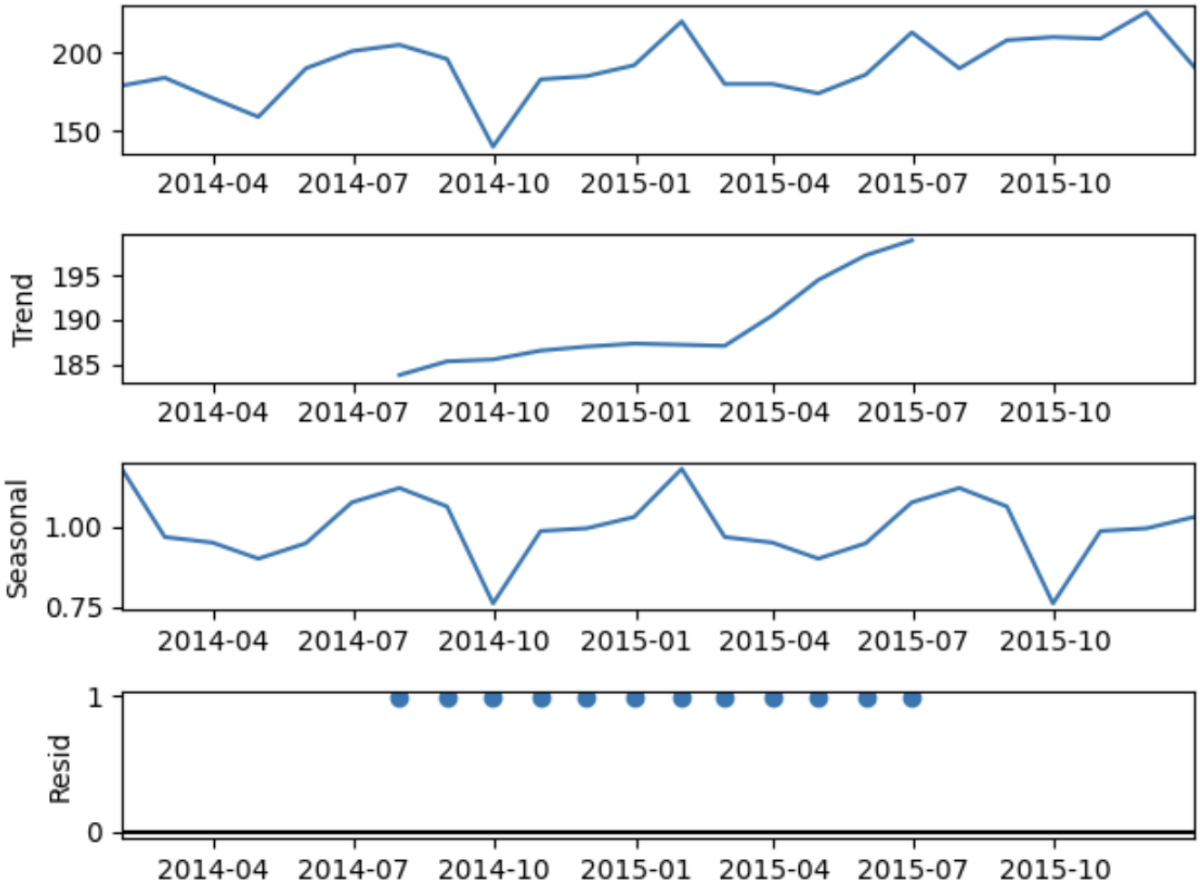


Figure 15

The trend analysis for the "Vegetables" category,in Figure 15, reveals a gradual increase in demand, followed by exponential growth after approximately March. This suggests a positive and potentially accelerating trend in consumer demand for vegetables. Additionally, the seasonal graph illustrates a dip in demand around the month of October, while indicating higher demand at the beginning of the year. These findings highlight the seasonality of vegetable consumption, with a potential decrease during October and heightened demand during the early months of the year.

Results of Section I:

- Our analysis revealed distinct seasonal and trend-based patterns in customers' purchasing behavior across the three identified categories.

- However, due to the significant volume of data and the potential for data reliability issues, our next step involves developing a robust forecasting model.
- This forecasting model will consider the observed trend patterns in the top three categories, enabling us to make more accurate predictions and projections for future demand.
- By incorporating these trend patterns into our forecasting model, we aim to improve the accuracy and reliability of our predictions, assisting businesses in making informed decisions regarding inventory management, production planning, and overall strategy.

SECTION II

Following are the research and design questions which are the further steps and will be the focus on the next part of the study :

- What are the trade-offs between a simpler, more interpretable forecasting model and a more complex, potentially more accurate model, and how can these trade-offs be optimized for the specific business problem at hand?
- Experimenting with the timeframe factor - whether to consider daily,weekly or monthly data makes a difference in the accuracy of the predictions.
- Using automated models/libraries and different methods to find the most accurate predictions.Following is the summary of steps undertaken further:
    - Using auto arima to find the best hyperparameters.
    - Training SARIMA model to fit the data with params identified in a.
    - Optimizing results using the 'Walk forward optimisation' method.
    - Experimenting with Prophet model predictions.

Approach :

For experimental purposes, we have successfully implemented all the aforementioned steps specifically for the 'Beverages' category- Monthly time frame. These steps included

- Identifying seasonal and trend-based patterns, sorting and filtering the data to isolate the 'Beverages' category
- Creating separate dataframes for the years 2014 and 2015
- Plotting the demand for beverages during these periods.

By completing this implementation, we have gained valuable insights into the consumption patterns and trends within the 'Beverages' category. This experimentation serves as a foundation for further analysis and provides a basis for future forecasting and decision-making in the beverage industry.

We used auto arima to select the most appropriate values for p and q, the highest lags with the greatest sum from each plot for p (autoregressive order) and q (moving average order). Additionally, considering that the data is in a monthly time frame, a value of m = 12 is chosen. This value represents the number of periods in a year and can be used to account for any seasonality or periodicity in the data. By determining the appropriate values for p, q, and m, we can better model and forecast the data using autoregressive integrated moving average (ARIMA) or seasonal ARIMA (SARIMA) models.

AUTO ARIMA

To test for stationarity in the data, we perform statistical tests such as the Augmented Dickey-Fuller (ADF) test . These tests help determine if the data exhibits stationary properties or if it requires differencing to achieve stationarity. The null hypothesis for the ADF test is that the data is non-stationary. By conducting this test and examining the resulting p-value, we can make a determination about the stationarity of the data.

```
ADF Statistic: -2.058803
p-value: 0.261399
Critical Values:
        1%: -3.788
        5%: -3.013
        10%: -2.646
```

Based on the observation that the p-value is greater than 0.05, we reject the null hypothesis and conclude that the series is non-stationary. Since we also noticed a periodic component in the data, it is necessary to apply transformations to make the data stationary.Next, we plot the monthly demand for beverages category.

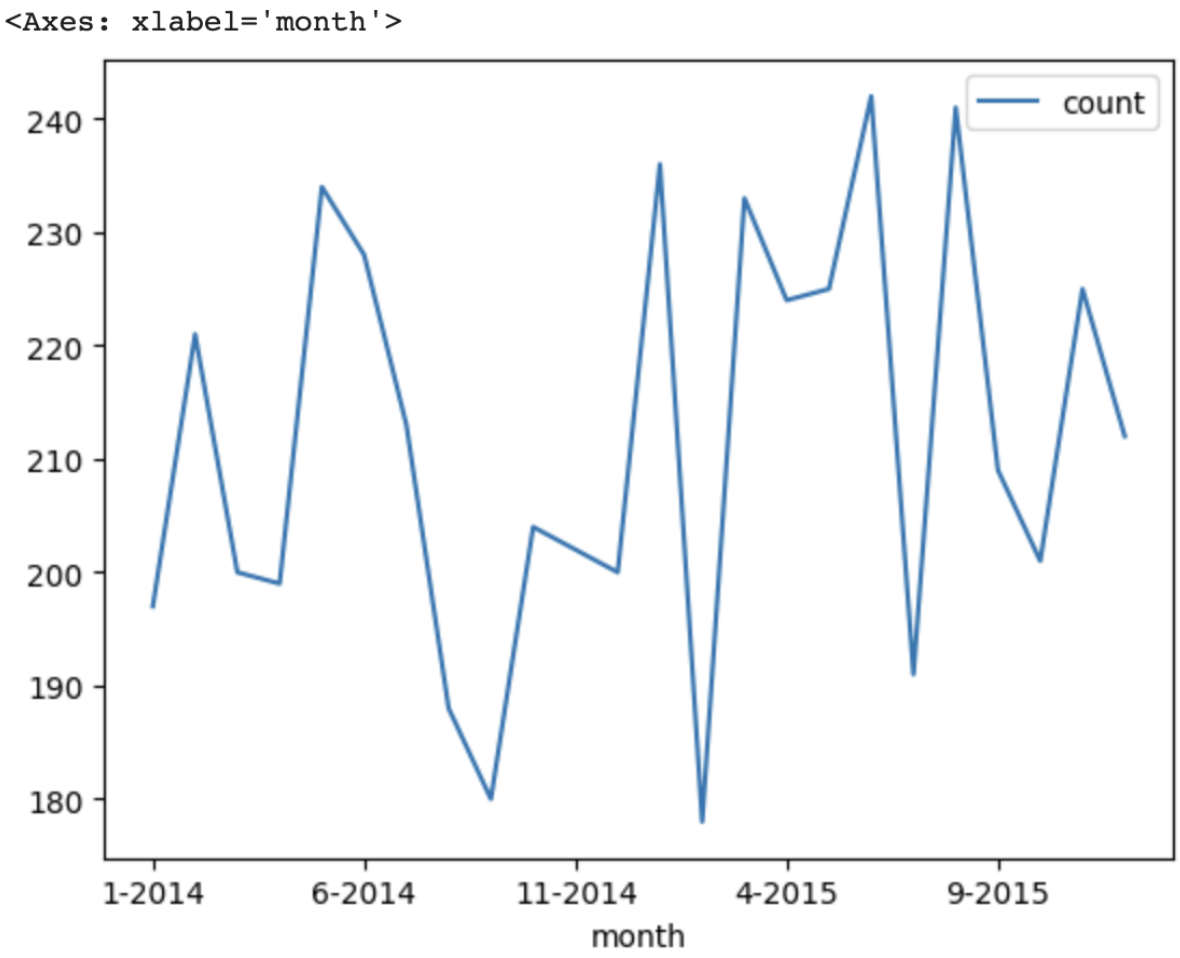


Figure 16

To determine the appropriate value for the differencing parameter (d), we can apply the differencing method. This involves taking the difference between consecutive observations in the data.By iteratively applying differencing until the resulting series becomes stationary, we can determine the minimum number of differencing operations required to achieve stationarity. Each differencing step reduces the trend and seasonality in the data, moving towards a stationary series.

Once the differencing parameter (d) is identified, the resulting differenced series can be used in combination with the identified values for p and q (from ACF and PACF analysis) to build an appropriate ARIMA or SARIMA model for forecasting and analysis.

Figure 17

After performing second-order differencing, it can be observed,in Figure 17, that there is less noise in the data. This suggests that the differenced series has become more stationary. However, it is important to verify this observation using an autocorrelation plot.

By examining the autocorrelation plot of the second-order differenced series, we can assess the autocorrelation patterns and determine if the values fall within the confidence interval. If the autocorrelation values decay significantly and fall within the confidence bounds, it provides further evidence that the second-order differencing has successfully made the series stationary.

Verifying the effectiveness of the second-order differencing using the autocorrelation plot helps confirm the appropriate value for the differencing parameter (d). In this case, since the second-order differencing exhibits improved stationarity and autocorrelation patterns, it suggests that d=2 is a suitable choice for the ARIMA or SARIMA modeling process.

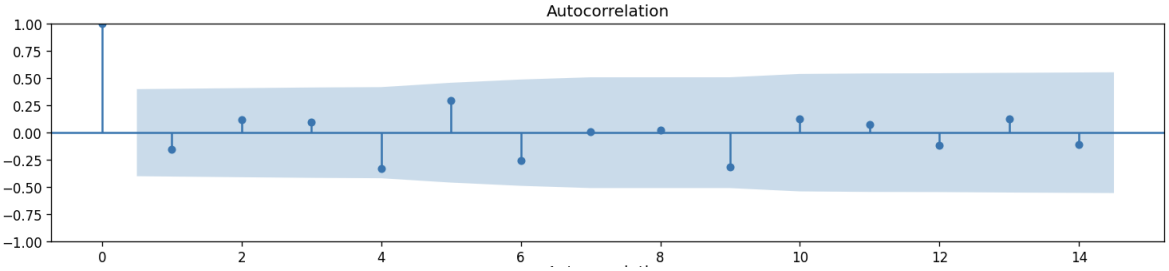


Figure 18

To determine the appropriate value for the autoregressive order (p) in the ARIMA or SARIMA model, we analyzed the partial autocorrelation function (PACF) plot(Figure 19), which further refined the value for p.The PACF plot helps identify the direct relationship between observations at different lags, considering the intermediate relationships accounted for by earlier lags.

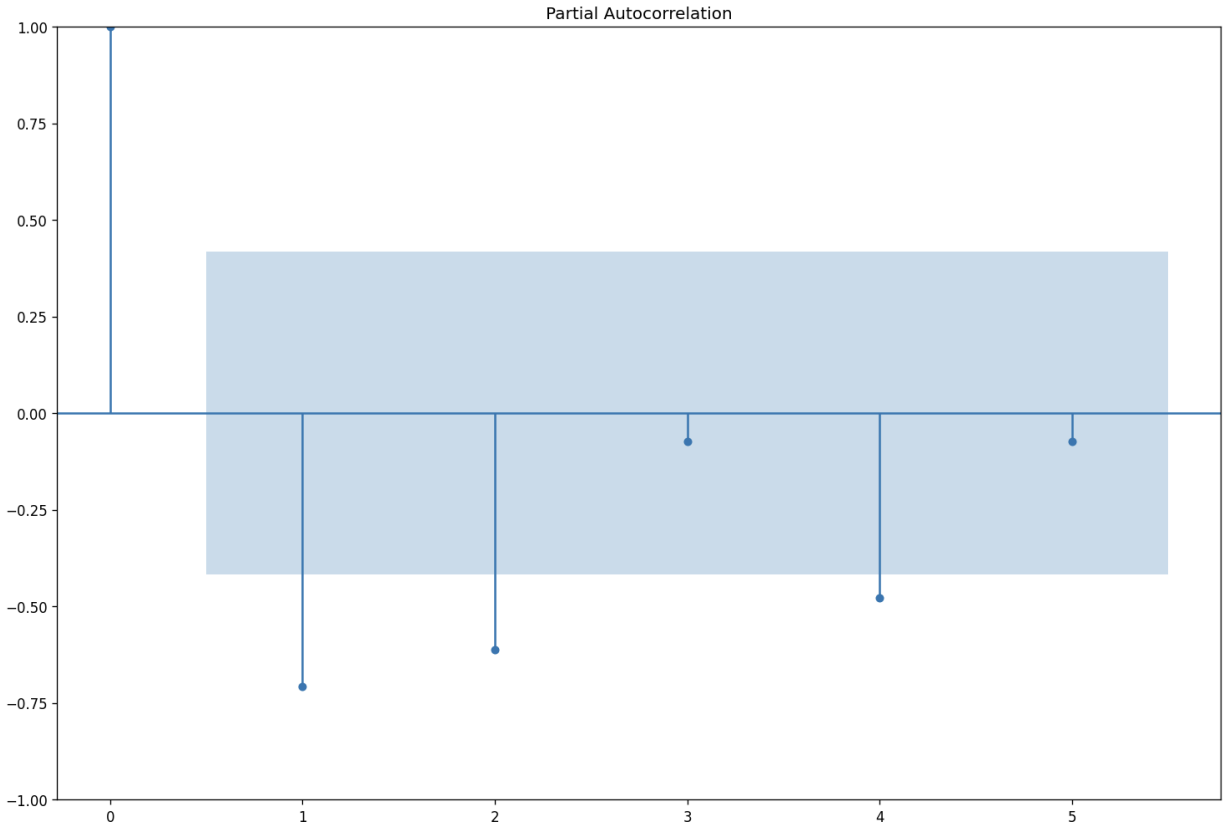


Figure 19

To determine the appropriate value for the moving average order (q) in the ARIMA or SARIMA model, we can analyze the autocorrelation function (ACF) plot(Figure 18). The ACF plot displays the correlation between the time series observations at different lags.By examining the ACF plot, we look for significant spikes or correlations that fall outside the confidence intervals. These spikes indicate the presence of autocorrelation and help identify the value for the q parameter. The lag at which the significant spikes occur can be used as the value for q.

To find the q parameter, we plot the ACF plot and observe the lag at which the significant spikes occur. These spikes outside the confidence intervals indicate the presence of significant autocorrelation at those lags. The lag associated with the last significant spike can be considered as the value for q.

Based on the analysis of the ACF plot in Figure 18, we have determined that the appropriate value for the moving average order (q) is 2. Additionally, for the autoregressive order (p),in Figure 19, we have chosen a value of 2 based on the significant spikes observed in the ACF plot.

To further confirm and determine values for all the hyperparameters, we proceed to fit an ARIMA model to the data. The model fitting process involves estimating the model coefficients and other parameters to best capture the patterns and behavior of the data. It is important to note that the model assumes the data to be white noise, indicating that the residual errors are uncorrelated and have constant variance.

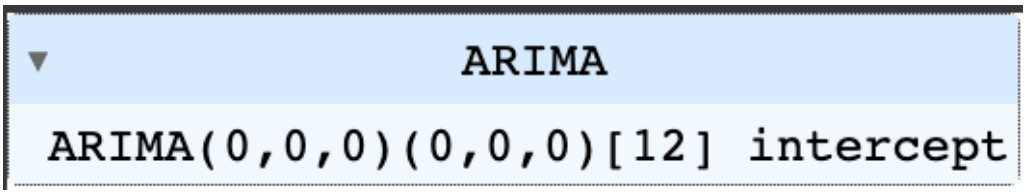


Figure 20

Conclusion from Section II:

The above results(Figure 20) from Auto ARIMA indicated that the data for the monthly time frame is equivalent to white noise. Hence, to gain a more detailed understanding of the 'Beverages' category, we decided to analyze the daily data instead of the monthly data. We

recognized that the daily data may provide additional insights and capture any finer patterns or trends that were not evident in the monthly data.

To identify the appropriate parameters for the ARIMA model on the daily data, we initially performed manual analysis using plots such as the ACF and PACF. However, we also utilized the 'auto.arima()' function, which automates the parameter selection process for ARIMA modeling. The 'auto.arima()' function uses statistical algorithms to determine the optimal values for the ARIMA model parameters based on the given data.

In general, we place more trust in the 'auto.arima()' function as it utilizes advanced algorithms and statistical techniques to identify the most suitable parameters for the ARIMA model. This automated approach reduces human bias and can lead to more accurate and reliable model selection.

Beverages (Daily)

In order to handle missing dates and ensure a complete date range, we utilized the reindex function. By reindexing the data, we can fill in the missing dates within the minimum and maximum date range mentioned.Firstly, we identified the minimum and maximum dates in the dataset. Then, using the reindex function, we can create a new index with a complete date range from the minimum date to the maximum date.

To fill the corresponding count values for the missing dates, we can set those values to 0. This ensures that all dates within the range have a count value assigned, even if it is zero.By performing this feature engineering step, we can ensure a consistent and complete dataset with no missing dates, enabling more accurate analysis and modeling of the data.

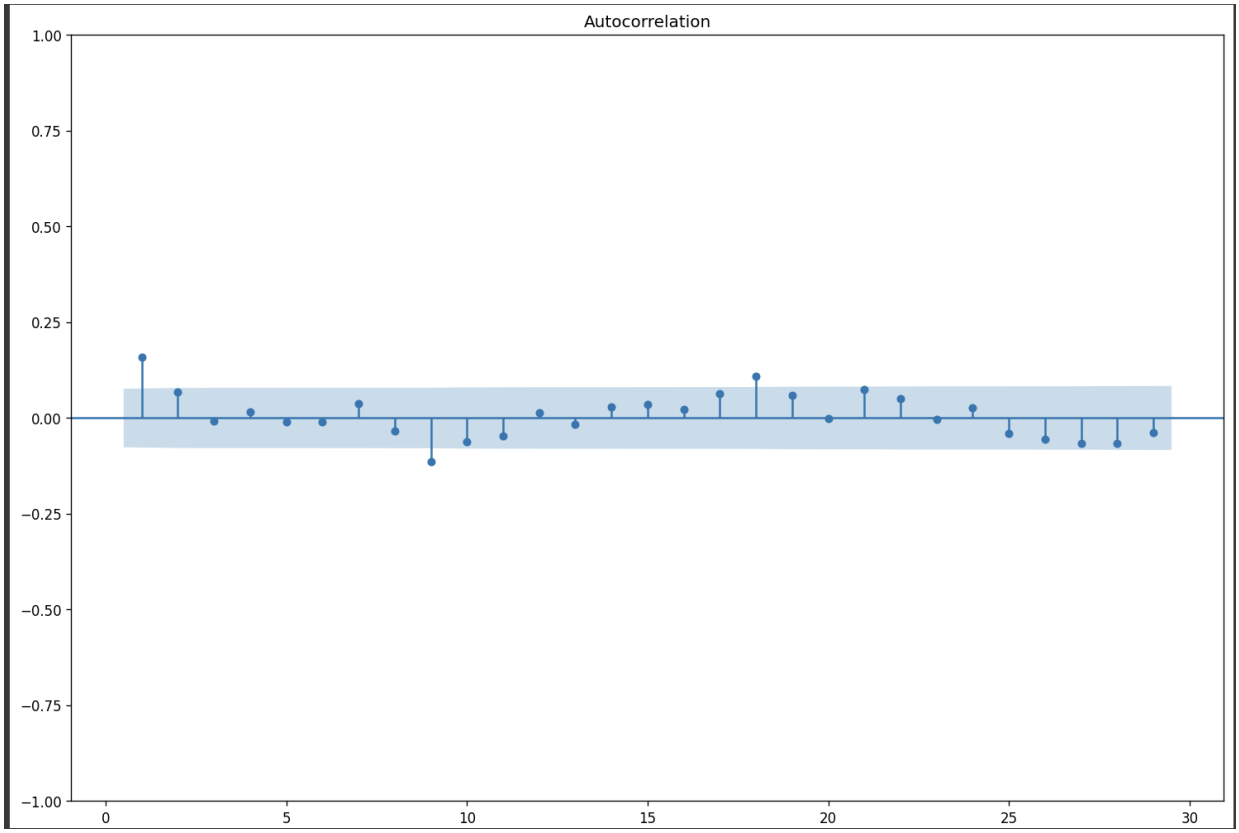


Figure 21

Based on the autocorrelation function (ACF) plot in Figure 21, we can observe that the autocorrelation values decrease gradually and become insignificant after lag 1. This indicates that there is a significant correlation between the current observation and its lag 1 value, suggesting a possible autoregressive (AR) component.

Therefore, we can set the parameter p to 1, indicating that we consider the lag 1 value in the AR component of the SARIMA model. This helps capture the dependence on the previous observation in predicting future values.

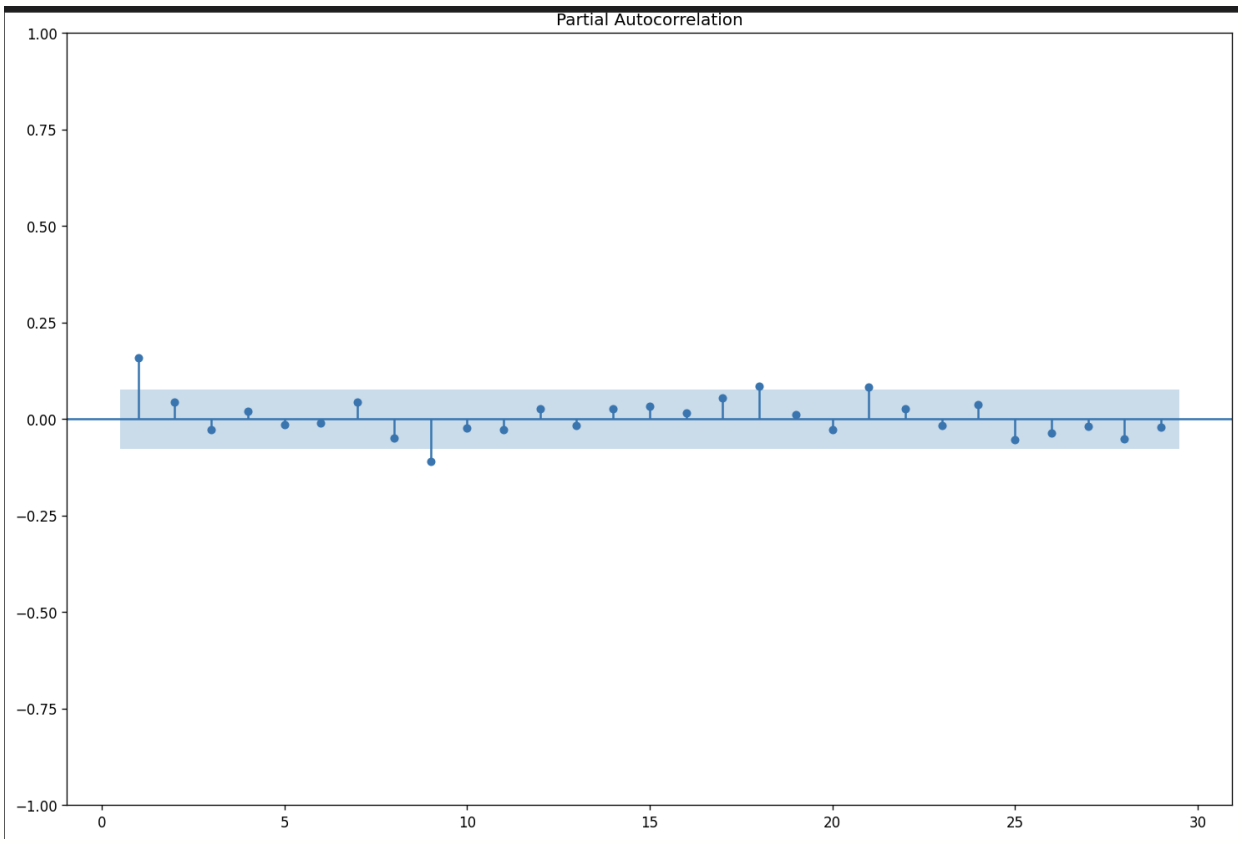


Figure 22

The q parameter, on the other hand, is the moving average component of the model. In other words, it represents the lagged value of the forecast errors in the model and allows for better forecasting of sudden shifts in the trajectory of the time series. From Figure 22, we observe the q value to 1.

```
ADF Statistic: -21.795225
p-value: 0.000000
Critical Values:
        1%: -3.440
        5%: -2.866
        10%: -2.569
```

Figure 23

The p-value obtained from the statistical test,the Augmented Dickey-Fuller (ADF) test,in Figure 24, is zero.It indicates strong evidence against the null hypothesis of non-stationarity. In this case, we can conclude that the data is already stationary and does not require any further processing or transformations to achieve stationarity.

Additionally, since the data is already stationary,as observed from the results of the ADF test, meaning there is no need for differencing, we can set the parameter d to 0. This implies that we do not perform any differencing on the data.

Steps undertaken:

- Visualizing the daily demand for beverages over the course of two years, we can plot a line graph with the dates on the x-axis and the corresponding demand values on the y-axis.
- By using the complete dataset with the filled missing dates and their corresponding count values, we generated a comprehensive representation of the daily demand trends for beverages.

The resulting line graph,in Figure 24, provided insights into the overall patterns, seasonal variations, and any notable trends in the daily demand for beverages over the two-year period. This visualization will help stakeholders understand the demand patterns and make informed decisions based on the observed trends.

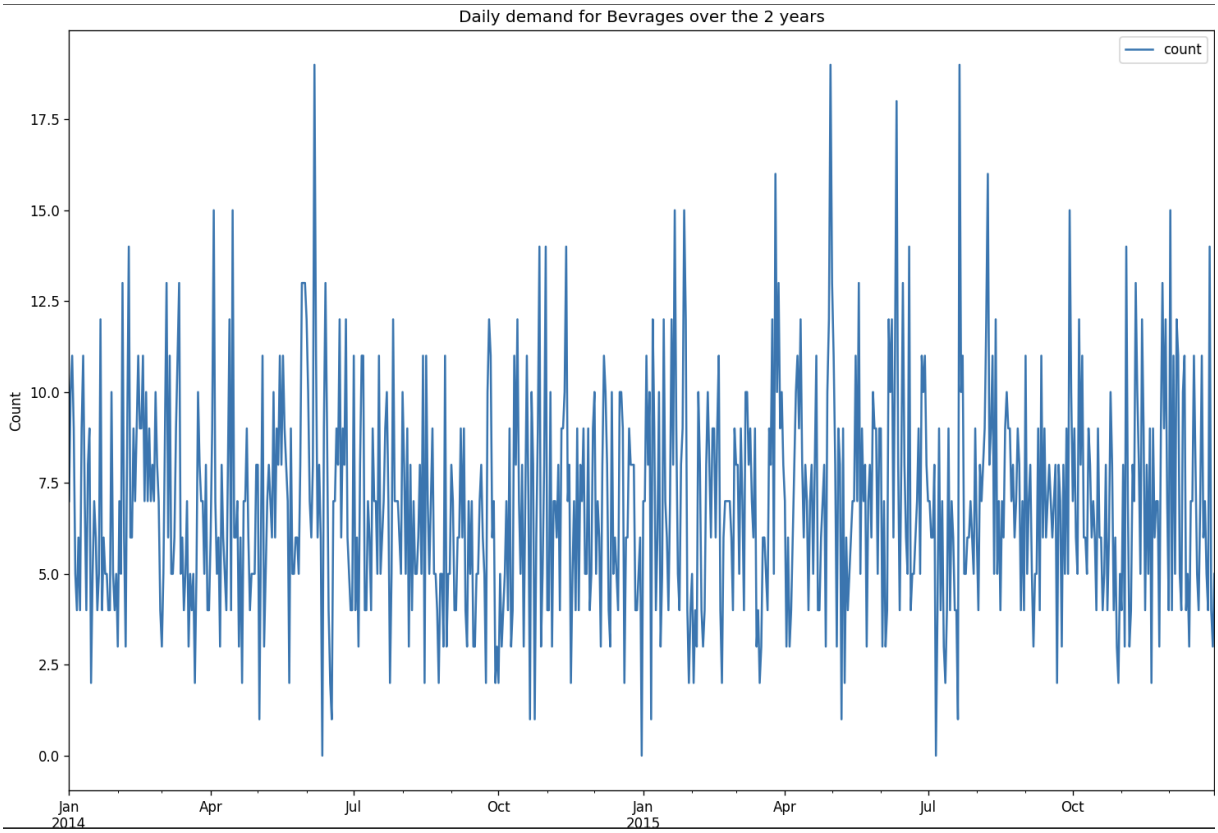


Figure 24

- The next step included splitting the data into training and testing datasets, we used a specific time period as the cutoff point. Typically, we reserve a portion of the data, such as the most recent period, as the test dataset, while the remaining data is used for training.
- Next, we checked if the data is stationary. Stationarity refers to the property of a time series where the statistical properties, such as mean and variance, remain constant over time. If the data is not stationary, we need to apply differencing to make it stationary.

This process of splitting the data and checking for stationarity, followed by differencing if necessary, ensures that we have appropriate datasets for training and testing, and that the data is in a suitable form for modeling and analysis.

Additionally, if no periodic component or seasonality is observed in the data, it further confirms that the data is already in a stationary form. This means that the statistical properties of the data, such as mean and variance, are relatively constant over time.

With a stationary dataset,as observed from results in Figure 23, and no need for further transformations, we can proceed with the modeling and analysis of the data. This allows us to build models based on the available stationary data and make accurate predictions and forecasts.

- To gain a deeper understanding of the underlying components within our training dataset, we further performed a decomposition analysis. Decomposition helps us separate the time series data into its individual components, namely trend, seasonality, and residual.

By decomposing the training dataset, in Figure 25,we identified the long-term trend, recurring seasonal patterns, and the random fluctuations or residuals present in the data. This analysis provided valuable insights into the underlying patterns and dynamics that helped us make more informed decisions and predictions.

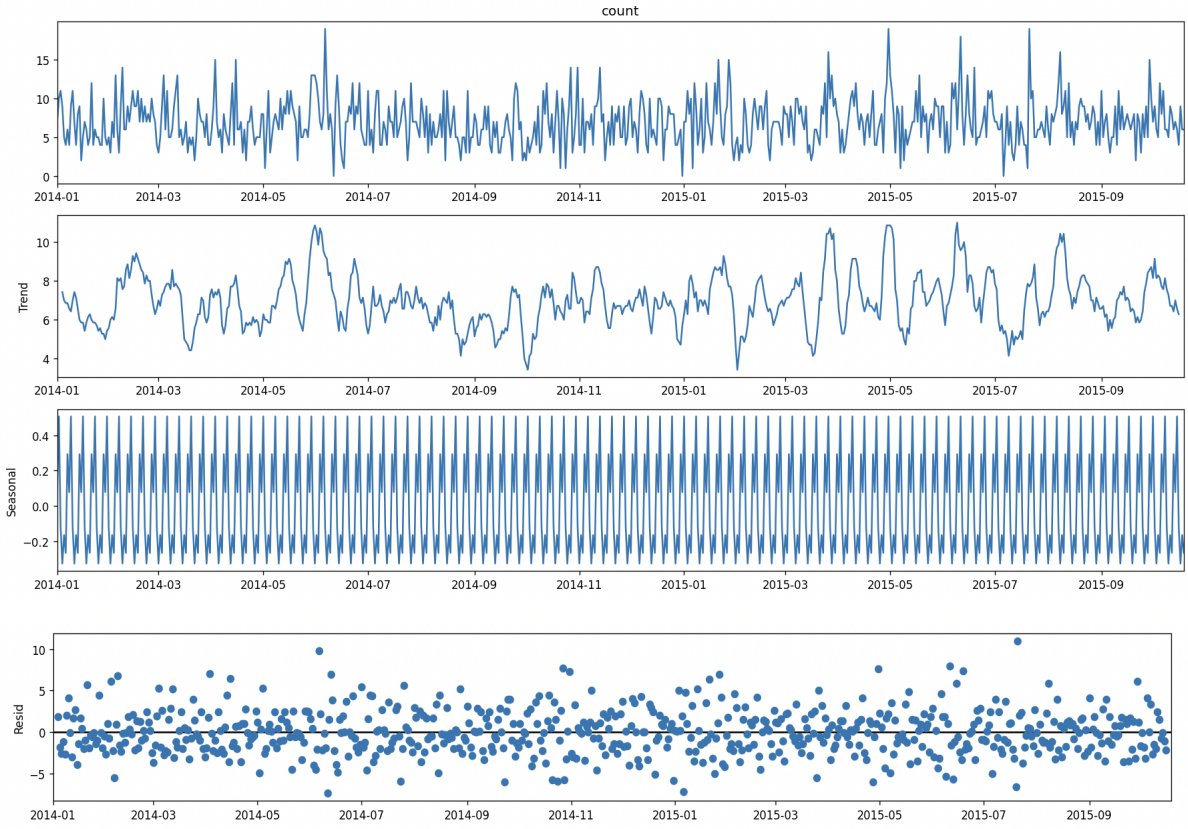


Figure 25

- To determine the optimal parameters for our time series model, we utilized the auto_arima function. This function automates the process of selecting the best parameters for an ARIMA model based on statistical criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion).

By applying auto_arima to our training dataset, we let the function search through various combinations of parameters, including different orders of autoregressive (AR), differencing (I), and moving average (MA) terms. The function evaluates each combination and selects the one that minimizes the chosen criterion.

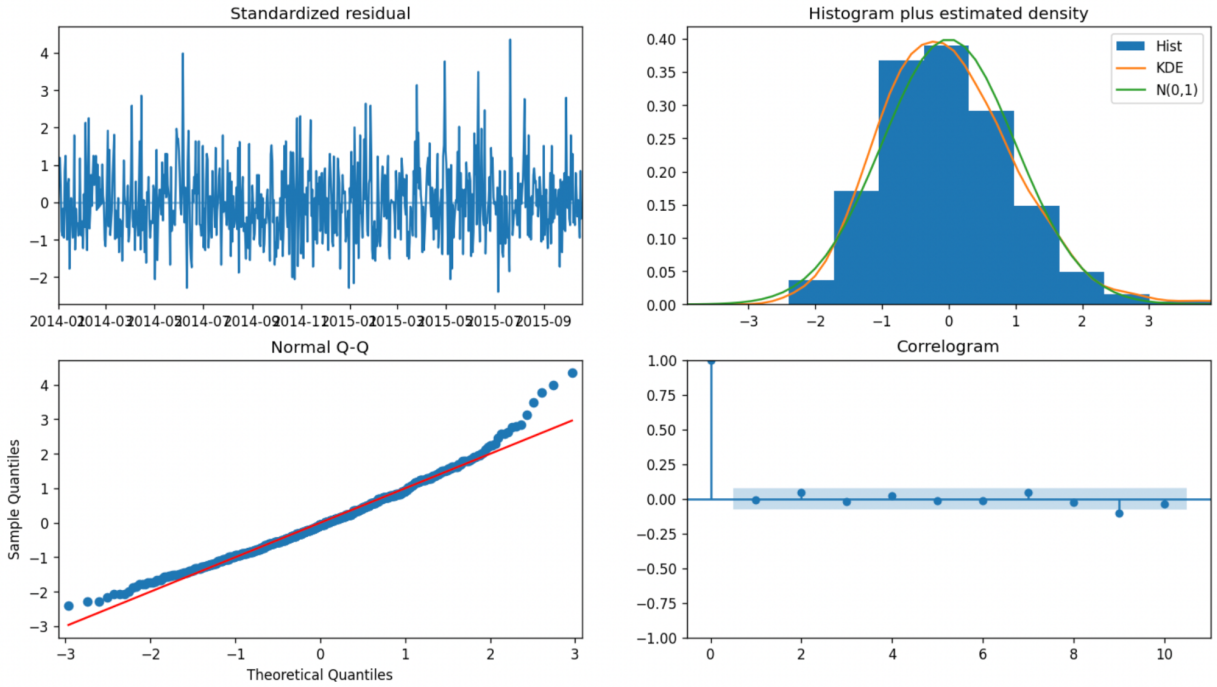


Figure 26

- The next step included predicting the demand (count) for the timeframe of the test data, we can use the trained model and the parameters obtained from the training phase. These parameters include the order of autoregressive (AR), differencing (I), and moving average (MA) terms.The order and seasonal order were found to be :-

```
(0, 0, 0, 0)
(1, 0, 0)
```

- Using the above selected parameters, we then fit the model to the training data and then apply it to the test data to make predictions. The model will use the historical information from the training data to forecast the future demand values for the test period.
- The predicted demand values will provide insights into the expected count for the test time frame based on the patterns and trends captured by the model during the training phase. These predictions can be compared against the actual demand values in the test dataset to evaluate the performance and accuracy of the model.

- To fit the SARIMA model with the given parameters obtained from the training phase, we used the SARIMAX class from the statsmodels library in Python. SARIMAX allows us to incorporate both the autoregressive (AR), moving average (MA), and seasonal components into the model.

- Once the SARIMA model is fitted to the training data, we then evaluated its performance by calculating the error metrics. Common error metrics for time series forecasting include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics provide insights into the accuracy and precision of the model's predictions compared to the actual values.
- Error for SARIMA model: 7.363592305352441

- After obtaining the error metrics, we combined the predicted values from the SARIMA model with the corresponding test data. This allows us to plot and visualize the predicted demand alongside the actual demand values,as seen in Figure 27, enabling us to assess the performance and reliability of the model visually.

- By fitting the SARIMA model, calculating the error metrics, and plotting the predicted and actual data, we gained a comprehensive understanding of the model's effectiveness in capturing the demand patterns and making accurate forecasts.

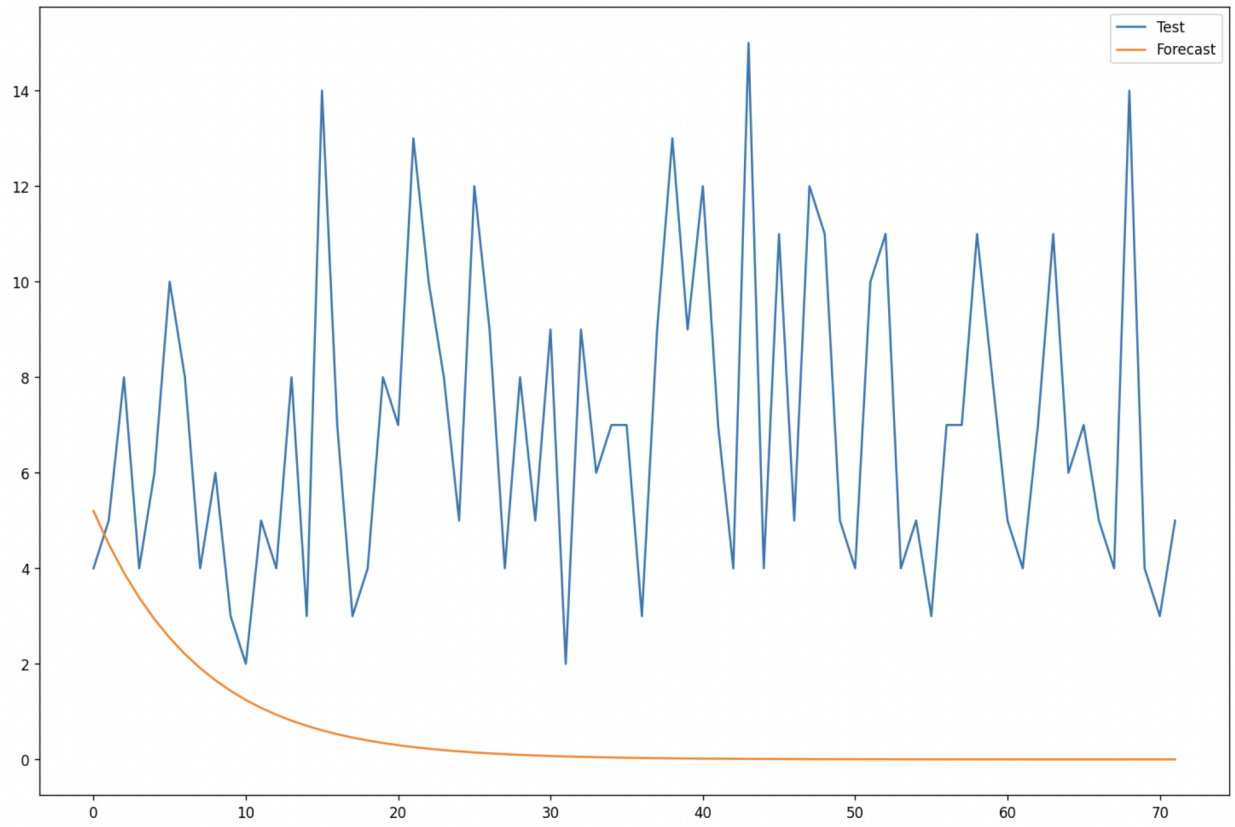


Figure 27

Walk Forward Cross Validation

Walk forward optimization is a process for testing a trading strategy by finding its optimal trading parameters in a certain time period (called the in-sample data) and checking the performance of those parameters in the following time period (called the out-of-sample data).

- In this step,we iterate over the test set in a walk forward manner, where we sequentially train the model on the training set and make predictions for the next observation in the test set.
- Further,Calculate the out-of-sample error, such as the Mean Absolute Error (MAE), Mean Squared Error (MSE), or any other appropriate error metric, by comparing the predicted values with the actual values in the test set. In this case it is RMSE.Next, we append the predicted value to the list of predicted values and further evaluate the performance of the walk forward optimized model by calculating the overall out-of-sample error using the error metric.
- Error for SARIMA model: 4.469798

Enclosed below(Figure 28) is the plot for predictions from Walk Forward Optimization.

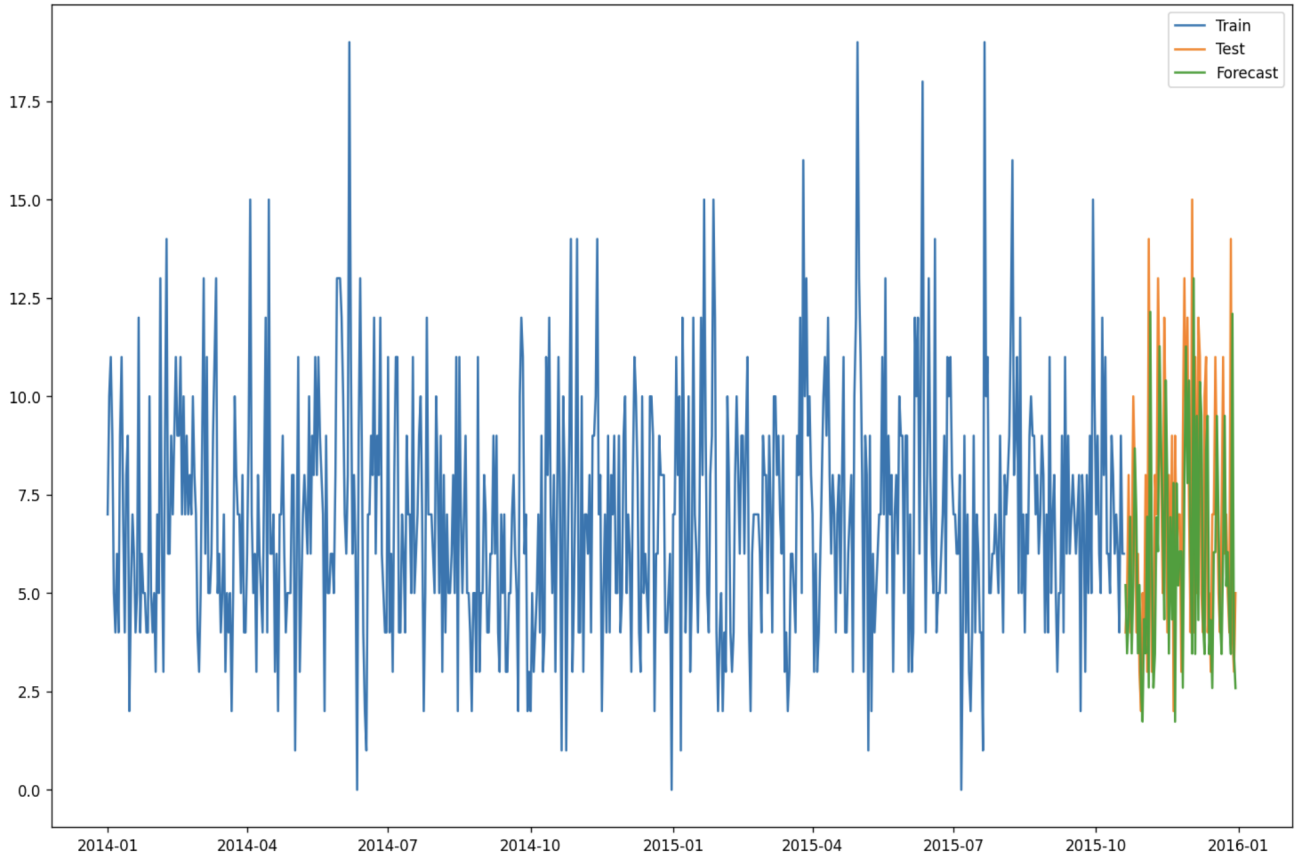


Figure 28

Results from SARIMA model

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	657
Model:	SARIMAX(1, 0, 0)	**Log Likelihood**	-1649.478
Date:	Mon, 24 Apr 2023	**AIC**	3304.956
Time:	22:19:28	**BIC**	3318.419
Sample:	01-01-2014 - 10-19-2015	**HQIC**	3310.175
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	5.8678	0.301	19.493	0.000	5.278	6.458
ar.L1	0.1584	0.038	4.185	0.000	0.084	0.233
sigma2	8.8757	0.436	20.345	0.000	8.021	9.731

Ljung-Box (L1) (Q):	0.03	**Jarque-Bera (JB):**	61.27
Prob(Q):	0.86	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.13	**Skew:**	0.60
Prob(H) (two-sided):	0.38	**Kurtosis:**	3.89

Experimentation with Prophet (Beverage demand)

Prophet is a forecasting technique that utilizes an additive model to predict time series data. It excels in handling time series with prominent seasonal effects and a substantial history of data. The method fits non-linear trends, incorporates yearly, weekly, and daily seasonality, and considers holiday effects. Prophet is known for its robustness in the presence of missing data, trend shifts, and outliers.
In this, we created a copy of beverages , consisting of Date and count columns.In order to work with the time series data, we can create a pandas DataFrame called 'ts' that includes two columns: 'ds' for the date values and 'y' for the corresponding count values.

To utilize the SARIMA model for forecasting, we start by instantiating the model and fitting it to the time series data. This involves training the model on the historical data to capture the underlying patterns and dependencies.

After fitting the model, we create a future dataframe that contains the dates for which we want to make predictions. This allows us to generate forecasts for future time periods.Once we have

the forecasts, we display the most important output columns, such as the predicted values, upper and lower confidence intervals, and any other relevant information.

Finally, we plot the forecasts for the timeframe of 2 months for the year 2016,along with the actual data to visually assess the model's performance and gain insights into the future trends and patterns of the time series.

Results of Experimentation with Prophet

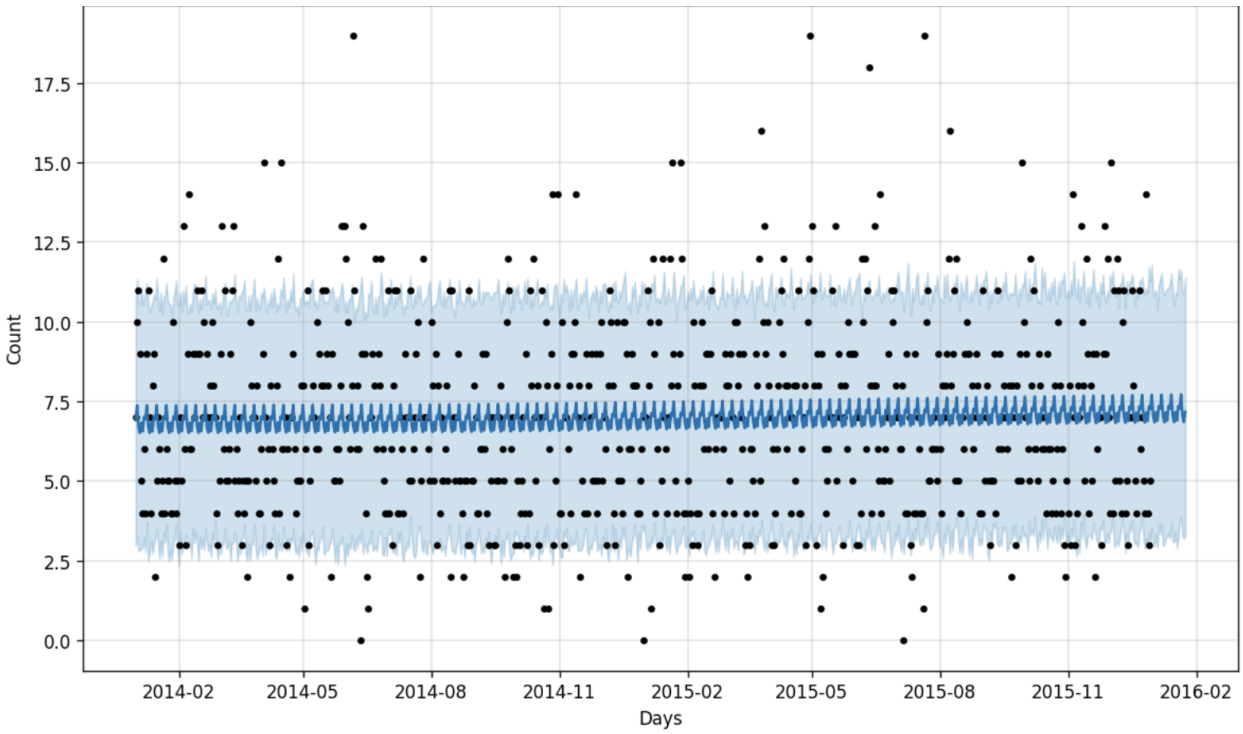


Figure 29

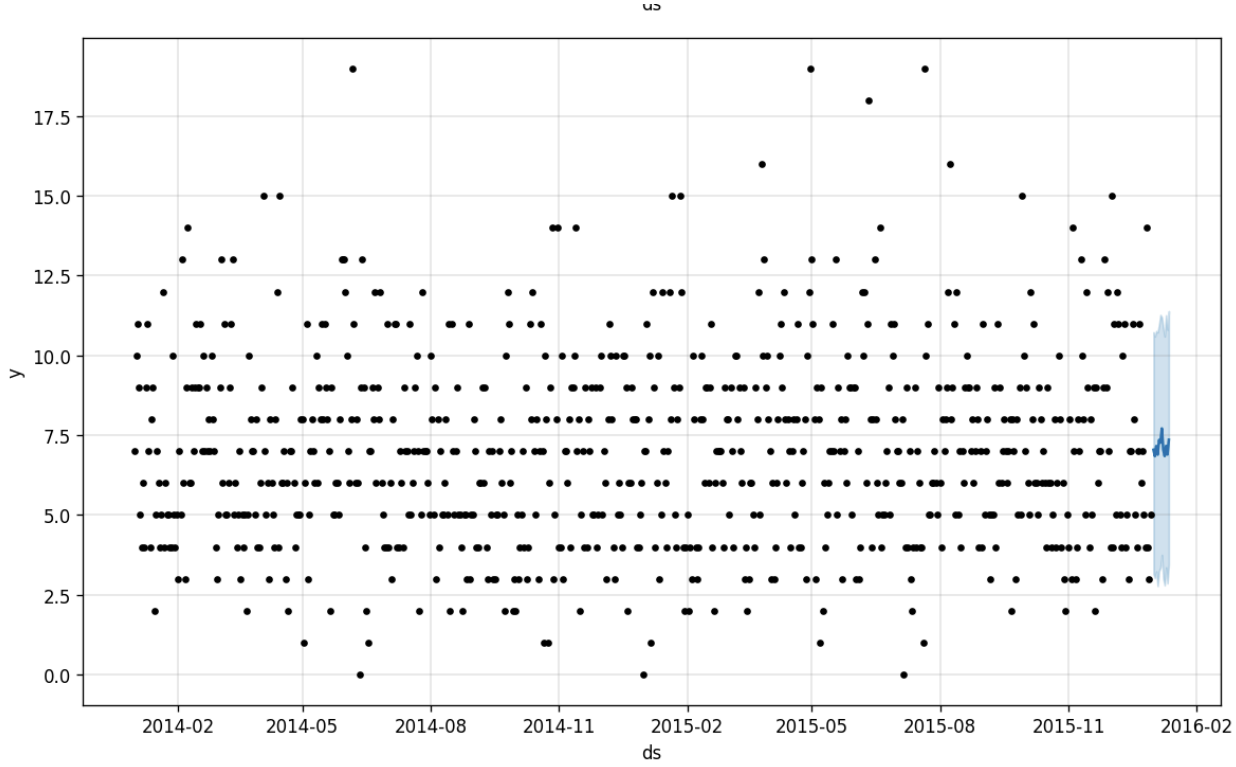


Figure 30

ETHICS:

What values are relevant to your design?

The two most relevant values would be Data Privacy and Data Quality. Because we are dealing with user information including demographics and user choices, there needs to be a layer of security added such that user information is not misused.Data Quality is important because the whole trend analysis and prediction is based on the actual demand which needs to be right/true.

How does your design portray your values?

The visualizations that we plan to implement would not include user demographics or such. It will only include the choices that the user has made,based on which different visualizations would be developed, thereby abiding Data Privacy.We would try to stick to the original data received instead of modifying it to keep it true and protecting the original quality of it.

What are some of the ethical and unethical ways of using your data?
Ethical ways as mentioned would be protecting the user demographics and other information.And on the other hand,unethical ways would be exposing user data such as user demographics,payment method,payment details,etc.

Does your design disproportionally affect underserved, marginalized, low-resourced, and underrepresented populations?
This design would be generic to all the retailers, which is our target audience and is not expected to disproportionally affect underserved, marginalized, low-resourced, and underrepresented populations.

Does your design impact the world's environment, resources, and climate?
Our design is predicting the future trend for grocery items that are frequently bought by users, and this design could only indirectly affect the world's resources.

Are there ways to accomplish your personal and organization's mission and values while promoting positive change in the society?
Yes. Since we are working on trend analysis for grocery items, we can emphasize on analysis in terms of sustainability, inclination towards healthy food choices and reduction in waste material,thereby considering impact on the environment. Based on the output, we could incentivize the audience in buying more sustainable items, thereby promoting positive change in the society.

Future Scope :

- Expand the analysis done for the 'Beverages' category to all the other categories.
- Make predictions for all the other categories.
- Built a real time system to intake data and make predictions periodically

References :

1. https://tsanggeorge.medium.com/a-semi-auto-way-to-determine-parameters-for-sarima-model-74cdee853080
2. https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8
3. https://xang1234.github.io/prophet/
4. https://audhiaprilliant.medium.com/walk-forward-optimization-cross-validation-technique-for-time-series-data-61739f58f2c0
5. https://pypi.org/project/fbprophet/