EXPLORING SALES DYNAMICS

A Time Series Study of

Favorita Stores' Product Families

Made by Ishan Sharma

INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Kanpur 208016 INDIA

Contents

1 Introduction

1.1 Problem Statement

The primary aim of this project is to conduct a comprehensive time series analysis and forecast the sales trends across product families available at Favorita stores in Ecuador. Our objective is to **gain insights into the temporal patterns of sales** and some statistics like the correlation coefficient between each category of products, which can help by providing information for strategic decision-making and **optimization of inventory management**.

1.2 About Dataset

The dataset used in this project comprises comprehensive information related to the sales and contextual details of Favorita stores located in Ecuador. The dataset includes the following key attributes:

- **Store_nbr:** This column indicates the store number corresponding to the data. A total of 54 stores are represented, with values ranging from 1 to 54.
- **Family:** Identifies the product family being sold, with diverse families such as 'AUTOMOTIVE,' 'BEVERAGES,' 'BEAUTY PRODUCT,' etc.
- **Sales:** Represents the total sales for a product family at a particular store on a given date. Fractional values are possible, reflecting the sale of fractional units of a product (e.g., 1.5 kg of cheese).
- **Onpromotion:** Indicates the total number of items in a product family that were being promoted at a store on a given date.

In addition to the sales-related information, metadata on different stores is provided, including details on the city and state of the store's location, the store type, and the cluster to which it belongs. Clusters are groupings of similar types of stores.

For further contextual understanding, the dataset also encompasses metadata on oil prices, providing daily information on the oil prices in Ecuador.

2 Techniques and Theory

2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to detect underlying patterns and characteristics of time series data. In this section, we employ various techniques to gain insights into the dataset.

2.1.1 Bar Plots and Line Charts

We utilized bar plots primarily for selection purposes. Figure 1 illustrates a bar plot depicting the total sales across all stores. Subsequently, we strategically selected 5 stores with low, moderate, and high total sales values. This selection aimed to ensure a well-representative sample of stores for our analysis. Moreover, we had information about clusters of stores, so the selection was done by even taking into account that stores are from different clusters. Selected store numbers are 3, 11, 24, 44 and 45.

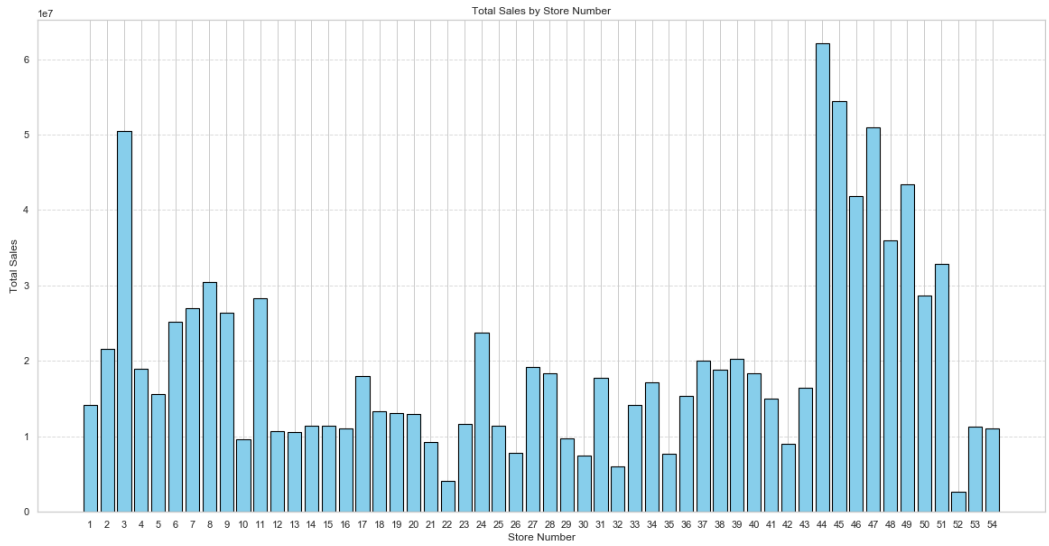

Figure 1: Total sales across all the stores

We employed a similar approach, using bar plots to select categories of products. By visualizing the distribution of products within each category, we could effectively identify and choose specific product categories for further analysis. Selected categories are "BEVERAGES", "GROCERYI", "CLEANING", "LIQUOR,WINE,BEER" and "DAIRY". See figure 2

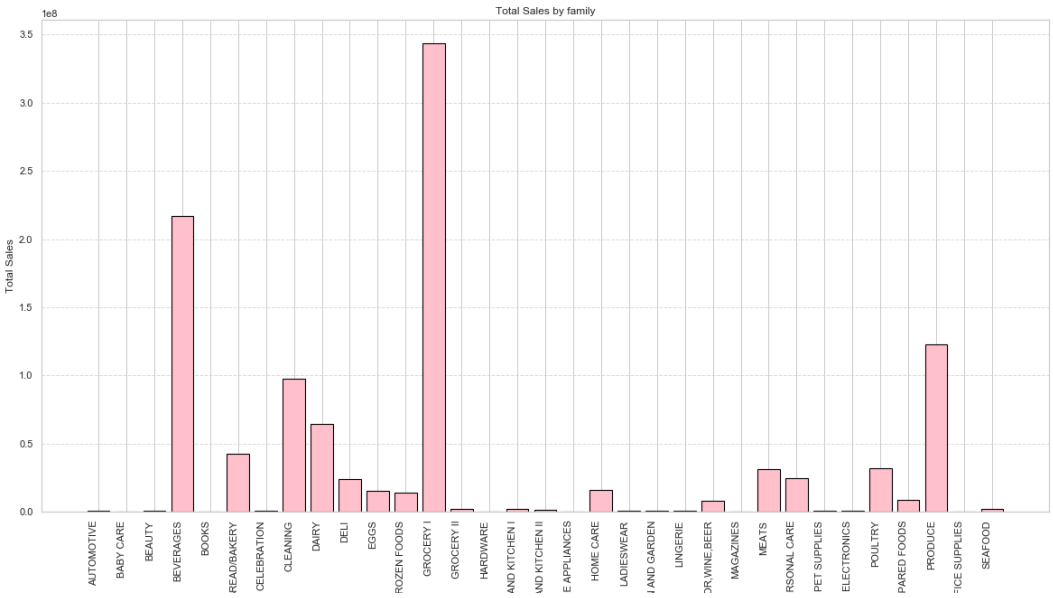

Figure 2: Total sales across all the categories of product

2.1.2 Outlier Detection

Outliers can signal errors or anomalies in the data collection process. Identifying and addressing outliers helps maintain data quality and accuracy. Outliers can distort the underlying patterns in time series data, leading to biased forecasts. Detecting and handling outliers is crucial for generating reliable and realistic predictions.

For every Time Series considered, outliers were identified using the Z-score technique, wherein the Z-score for each data point was calculated. Subsequently, data points identified as outliers, they are first checked for any pattern of their occurrence and if no seasonal occurrence is observed then they undergo a treatment process involving imputation through interpolation or median imputation.

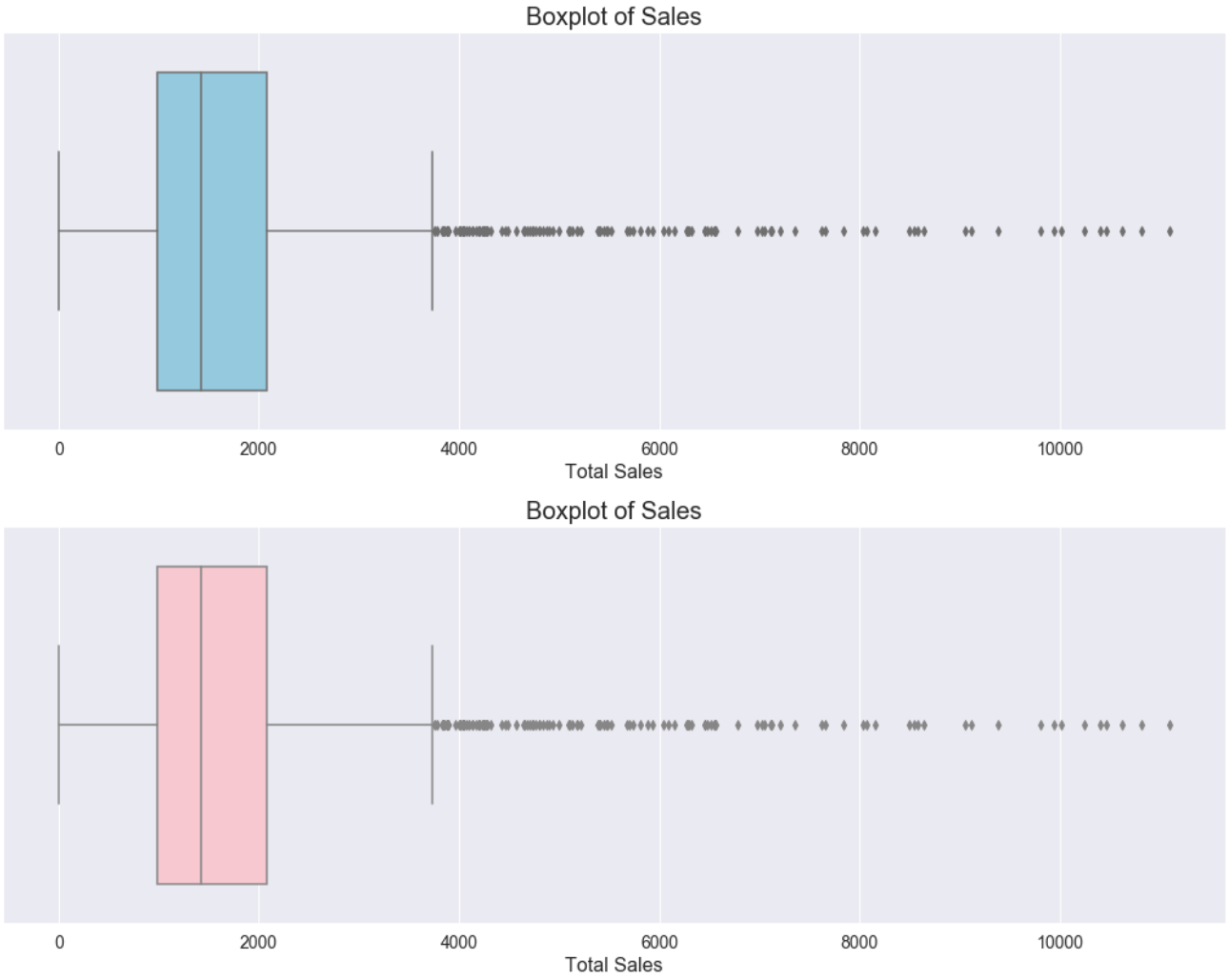

Figure 3: 'Sales' of beverages for Store 25- before and after outlier treatment

2.1.3 Correlation between Covariates

Understanding the correlation between different covariates can provide insights into potential relationships within the time series data. In Figure 5, is a correlation matrix heatmap, visually representing the strength and direction of relationships between variables.

2.1.4 ACF/PACF Plots

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are essential for identifying the autocorrelation structure in time series data. These plots assist in determining the order of autoregressive and moving average components in a time series model.

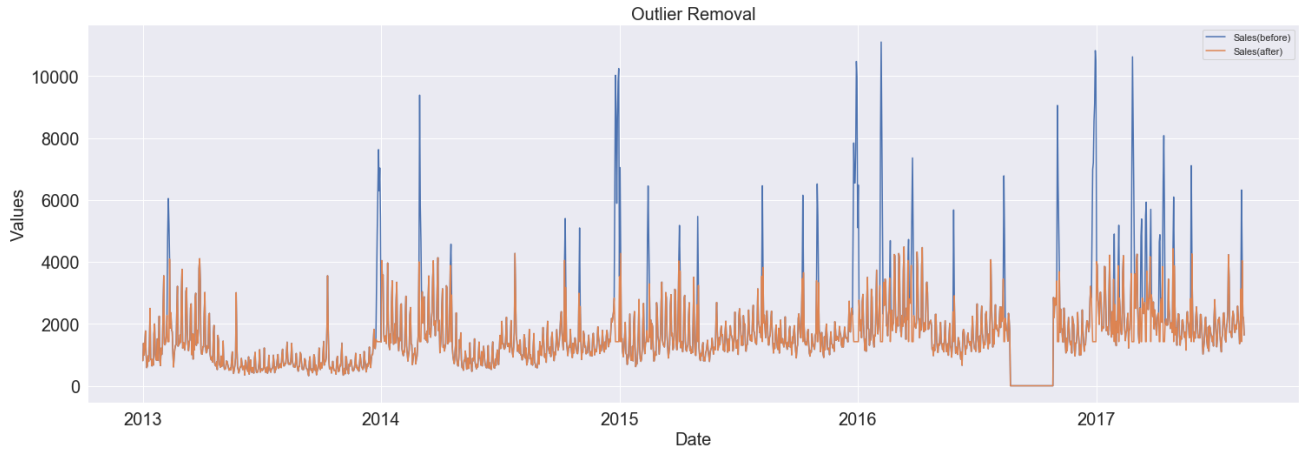

Figure 4: Boxplot of 'Sales' of beverages of Store 25 without outlier treatment and after outlier treatment.

2.2 Time Series Decomposition

2.2.1 Model assumed in this project

Different types of time series models include the additive model, which involves breaking down the time series data into distinct components, such as trend, seasonality, and residual, allowing for a more granular understanding of the underlying patterns. Multiplicative models, which involve considering the interaction between components as a multiplication rather than addition, and hybrid models that combine both additive and multiplicative features to capture a broader range of temporal structures. We applied the **additive model** approach to our time series analysis.

2.2.2 Deterministic components of Time Series and testing their presence

Components:

1. **Trend:** The long-term movement or general direction in the time series. It represents the underlying pattern or trajectory that the data follows over an extended period.

2. **Seasonality:** The repetitive and predictable patterns that occur at regular intervals, often corresponding to specific time units such as days, weeks, or months. Seasonality captures recurring fluctuations within the time series.

3. **Cycle:** The cyclical patterns that are not strictly tied to fixed time intervals. Unlike seasonality, cycles do not have a fixed duration and may represent longer-term oscillations in the data.

4. **Irregularity/Residual:** The random and unpredictable fluctuations that remain after removing the trend, seasonality, and cycle components. Residuals capture the unexplained variance in the time series.

Test for each Component:

Tests for Trend: (Relative Ordering Test)
This is a non-parametric test procedure used for testing the existence of a trend component in a time series.
Hypotheses:
H_0: Trend is not present

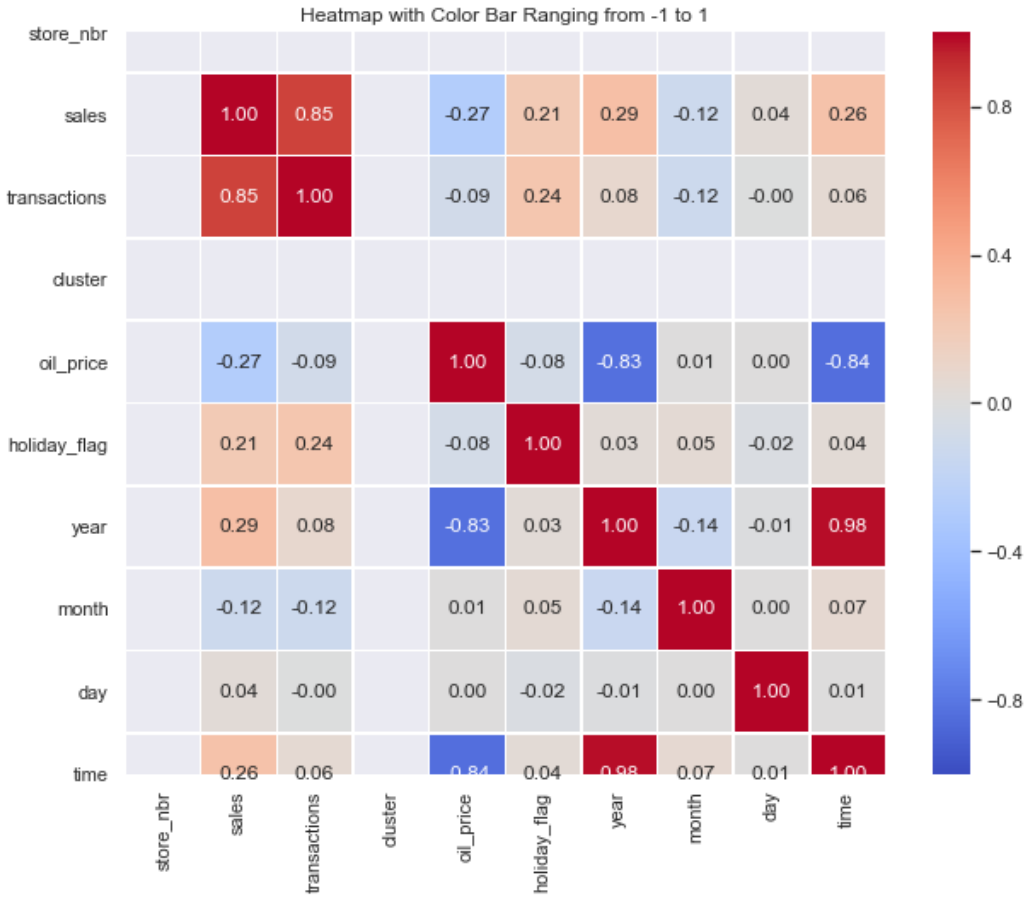

Figure 5: Correlation Matrix Heatmap for all covariates(Store 25)

H_1: Trend is present

Test Procedure:

Let $\{Y_1, \ldots, Y_n\}$ denote the time series at n time points. Define $q_{ij} = 1$ if $Y_i > Y_j$ when $i < j$, and 0 otherwise. $Q = \sum q_{ij}$ counts the number of decreasing points and discordances.

If there is no increasing or decreasing trend, then

$$P(q_{ij} = 0) = P(q_{ij} = 1) = \frac{1}{2}$$

. Therefore, under H_0,

$$E(Q) = \frac{n(n-1)}{4}$$

.

If observed $Q \ll E(Q)$, it indicates a rising trend; if $Q \gg E(Q)$, it indicates a falling trend. If Q does not differ significantly from $E(Q)$ (under H_0), it indicates no trend.

Q is related to Kendall's tau (τ), the rank correlation coefficient, through

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

. Using the standard results of Kendall's Tau, under H_0,

$$E(\tau) = 0$$

$$V(\tau) = \frac{2(2n+5)}{9n(n-1)}$$

.

Test Statistic: Asymptotic test for H_0 : no trend is based on

$$Z = \frac{\tau - E(\tau)}{\sqrt{V(\tau)}}$$

It follows asymptotically $N(0, 1)$ under H_0. Reject H_0 at α if $|Z| > Tau(\frac{\alpha}{2})$.

2.2.3 Stationarity and test to detect Stationarity

Stationarity: In time series analysis, stationarity is a concept that refers to the statistical properties of a time series remaining constant over time. A stationary time series exhibits constant mean, variance, and auto-correlation throughout its entire duration. Stationarity is essential for many time series models and analyses, as it simplifies the patterns and relationships within the data.

ADF (Augmented Dickey-Fuller) Test: The ADF test is a statistical test used to assess the stationarity of a time series. It evaluates whether a unit root is present in the data, indicating non-stationarity. The **null hypothesis of the ADF test is that the time series has a unit root (i.e., it is non-stationary)**, while the alternative hypothesis suggests stationarity. A low p-value (¡ 0.05) leads to the rejection of the null hypothesis, indicating that the time series is stationary.

KPSS (Kwiatkowski-Phillips-Schmidt-Shin) Test: The KPSS test is another method for testing the stationarity of a time series. Unlike the ADF test, KPSS focuses on the **null hypothesis of stationarity**. The test involves two hypotheses: the null hypothesis assumes that the series is stationary around a deterministic trend, while the alternative hypothesis assumes a unit root. A high p-value suggests stationarity, while a low p-value indicates non-stationarity.

In our analysis, we employed both the ADF and KPSS tests to assess the stationarity of the time series data. These tests play a crucial role in determining the appropriate transformations and models for our time series analysis, ensuring that the data's statistical properties remain consistent over time.

2.3 Forecasting

In practical implementation, a closer look at the working of each forecasting method provides insights into their effectiveness:

2.3.1 Linear Regression for Time Series Forecasting

In the context of time series analysis, the goal is often to predict future values of a target variable based on its historical observations.

Basic Concept: The basic concept involves fitting a linear equation to the historical data, where the independent variable represents time, and the dependent variable is the observed values of the time series. The linear regression model can then be used to make predictions for future time points.

Time Series Transformation: For time series forecasting using linear regression, it is common to transform the time series into a supervised learning problem. This is achieved by creating lag features, representing past observations, as independent variables. For instance, if predicting the value at time t, the model may use the values at $t-1$, $t-2$, and so on, as input features.

Training and Testing: The time series dataset is typically split into training and testing sets. The model is trained on the historical data and then evaluated on a separate set of data not seen during training. This allows assessing the model's performance on unseen future observations.

Evaluation Metrics: In this project evaluation metrics for linear regression in time series forecasting was Root Mean Squared Error (RMSE).

2.3.2 Autoregressive Integrated Moving Average (ARIMA)

ARIMA models operate by decomposing time series data into three main components: autoregressive (AR), integrated (I), and moving average (MA). The autoregressive component captures dependencies between an observation and its lagged values, the integrated component addresses non-stationarity by differencing the series, and the moving average component models the relationship between an observation and a residual error. The practical implementation involves selecting appropriate orders for these components (p, d, q) through a combination of visual inspection and statistical criteria.

2.3.3 Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA builds upon the ARIMA framework by incorporating seasonal components. It introduces additional parameters for seasonal autoregressive, seasonal differencing, and seasonal moving average components. In practical terms, identifying and selecting these seasonal components involves analyzing the seasonality patterns present in the data. Implementers need to assess whether the seasonality is additive or multiplicative and adjust the model accordingly.

2.3.4 Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are deep learning models designed for sequence data. Their working involves processing input sequences through recurrent connections, allowing the network to retain information from previous time steps. Practical implementation of RNNs includes defining the network architecture, specifying hyperparameters, and training the model using backpropagation through time. Implementers often grapple with challenges such as vanishing or exploding gradients, which can impact the network's ability to capture long-term dependencies.

In our practical application of these methods, each was tailored to the specific characteristics of our time series dataset, and their performance was evaluated based on their ability to accurately predict future values.

2.4 Data Preparation

2.4.1 Misisng Value treatment

In time series analysis, the occurrence of missing values is a common challenge that can significantly impact the accuracy of forecasting models.

Causes of Missing Values in Time Series

Several factors contribute to the presence of missing values in time series data:

- **Sensor Malfunction:** Malfunctioning sensors or data collection instruments can lead to gaps in the recorded time series.
- **Data Transmission Issues:** Problems during the transmission of data can result in missing or corrupted values.
- **Irregular Sampling Intervals:** Inconsistent or irregular time intervals between data points can introduce missing values.
- **Data Privacy Concerns:** In some cases, certain data points might be intentionally omitted due to privacy concerns.

Remedies for Missing Value Treatment

Several techniques can be employed to handle missing values effectively:

- **Interpolation:** Interpolation methods estimate the missing values based on the observed data points. Popular techniques include linear interpolation, spline interpolation, and polynomial interpolation. These methods assume a smooth transition between existing data points.
- **Forward or Backward Filling:** This method involves propagating the last observed value forward or the next observed value backward to fill in missing values. While simple, this approach might not be suitable if the time series exhibits abrupt changes.
- **Mean or Median Imputation:** Imputing missing values with the mean or median of the observed data can be a straightforward approach. However, it may not be suitable for time series with trend or seasonality.
- **Machine Learning Techniques:** Advanced techniques, such as regression models or machine learning algorithms, can be employed to predict missing values based on the available information. These methods are particularly useful when the relationships within the time series are complex.

Interpolation in the Project

For the current project, the chosen approach for missing value treatment was interpolation. Interpolation was choosen due to its ability to estimate missing values while preserving the overall trends and patterns present in the time series. (used linear interpolation)

3 Key Findings

3.1 Exploratory Data Analysis

Overall observations from data: 1. No missing value
2. there are 1684 unique dates
3. each date is repeated 1782 times
4. there are 54 stores with each store having 55572 entries
5. for each date we have data of all 33 family items
6. stores between 44 to 51 are having good sales
7. For evaluation taking 5 stores: 3,11,24,44 and 45 each from different cluster.

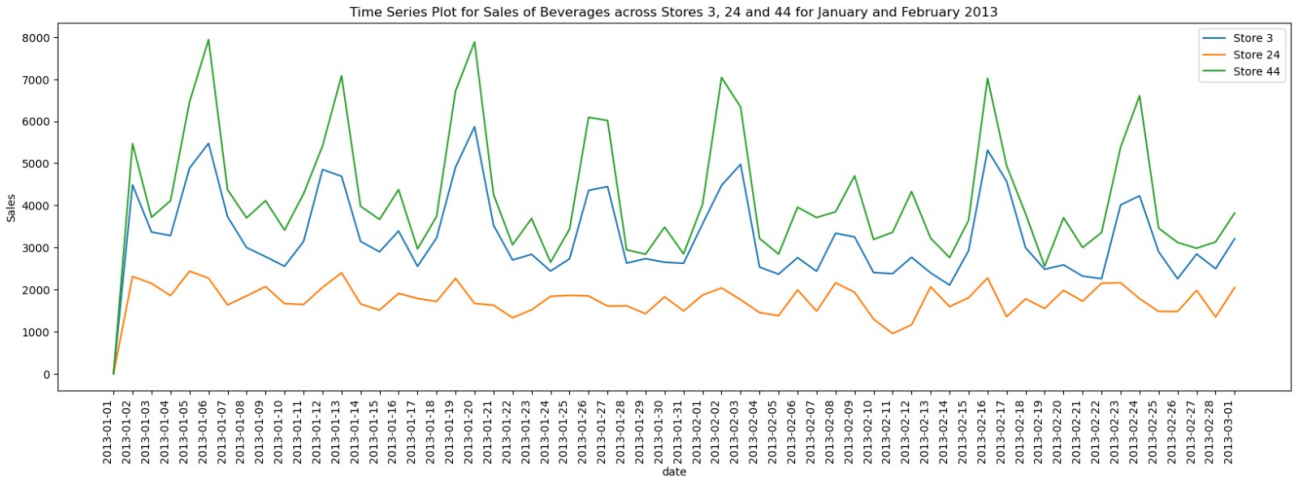

Figure 6: Boxplot of 'Sales' of beverages of Store 25 without outlier treatment and after outlier treatment.

Beverage sales of Store 25: ADF and KPSS tests were applied on the data. The results were:
Case 1: Both tests conclude that the series is not stationary - series is not stationary
Case 2: Both tests conclude that the series is stationary - series is stationary
Case 3: KPSS = stationary and ADF = not stationary - trend stationary, remove the trend to make series strict stationary
Case 4: KPSS = not stationary and ADF = stationary - difference stationary, use differencing to make series stationary.

First plot represents the component-wise decomposition of the time series data.

3.2 Forecasting - Sales

Linear Regression Results: From the plot of RMSE vs lag, we can observe that RMSE was minimum when our series was regressed on co-variates and its 35 lag values. (Figure 10) From the ACF and PCF plot, it seems that ARMA model will be a better fit for the data compared to the AR or MA model. (Figure 11)

ARIMA Results: Best model: ARIMA(3,0,2)(0,0,0)[0] Total fit time: 30.923 seconds Mean Absolute Error: 496.06303643947484 Root Mean Squared Error: 936.0269896225337 (Figure 12)

SARIMA Results: Best model: ARIMA(5,0,5)(0,0,0)[0] Total fit time: 89.286 seconds Mean Absolute Error: 644.1505273429959 Root Mean Squared Error: 1032.0985389617178 (Figure 13)

RNN Results: Mean Absolute Error: 644.1505273429959 Root Mean Squared Error: 910.679 (Figure 14)

3.3 Forecasting - Correlation

This is the time series of correlation between 'PREPARED FOODS' and'LIQUOR,WINE,BEER'. (Figure 16) Its decomposition is: (Figure **??**)

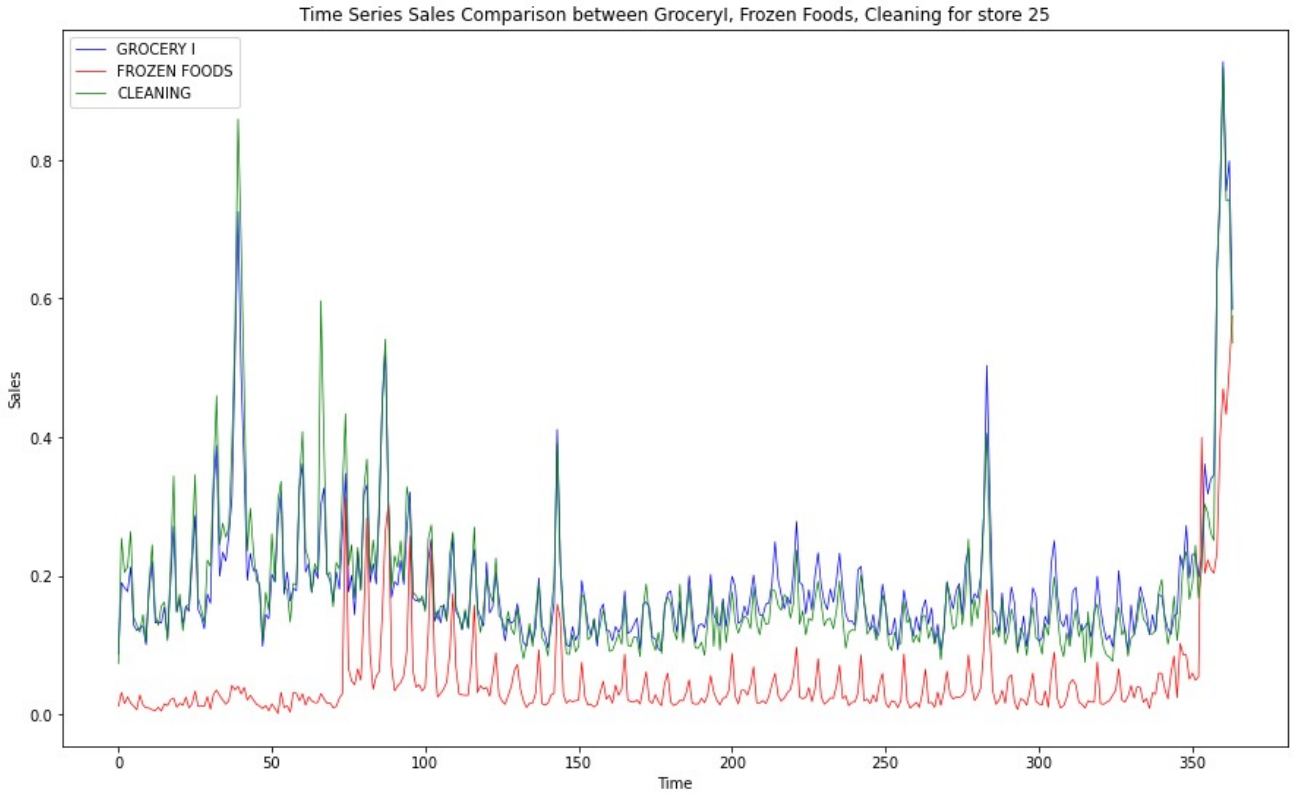

Figure 7: Comparison of the sales of 3 categories of Store 25

Figure 8: Component-wise decomposition of 'BEVERAGES' in store 25

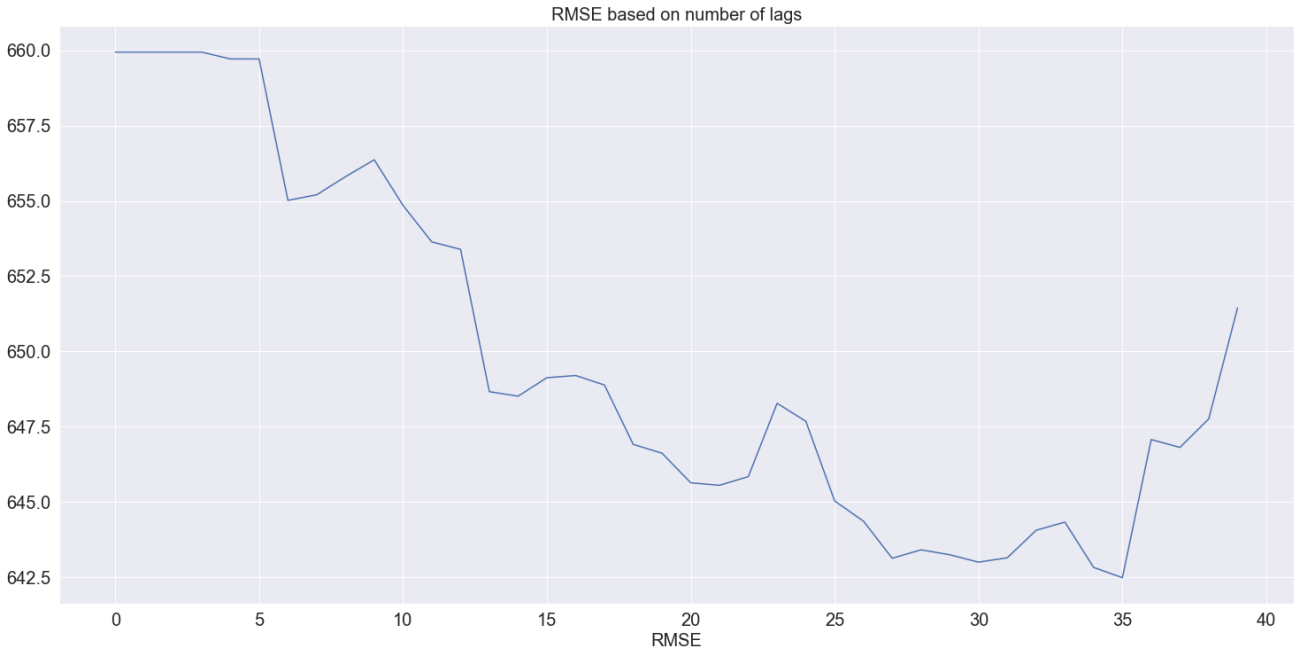

Figure 9: RMSE vs lag for Linear regression forecasting on sales of 'BEVERAGES' in store 25

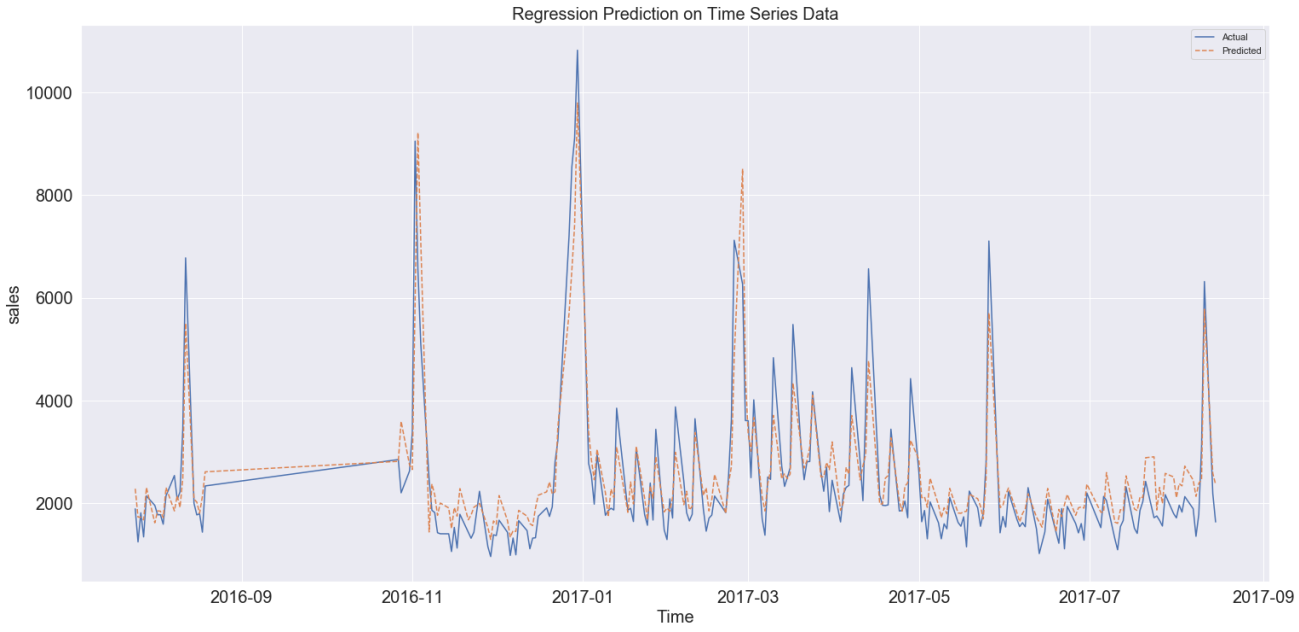

Figure 10: Linear Regression Forecast : prediction was based on 35 lag values

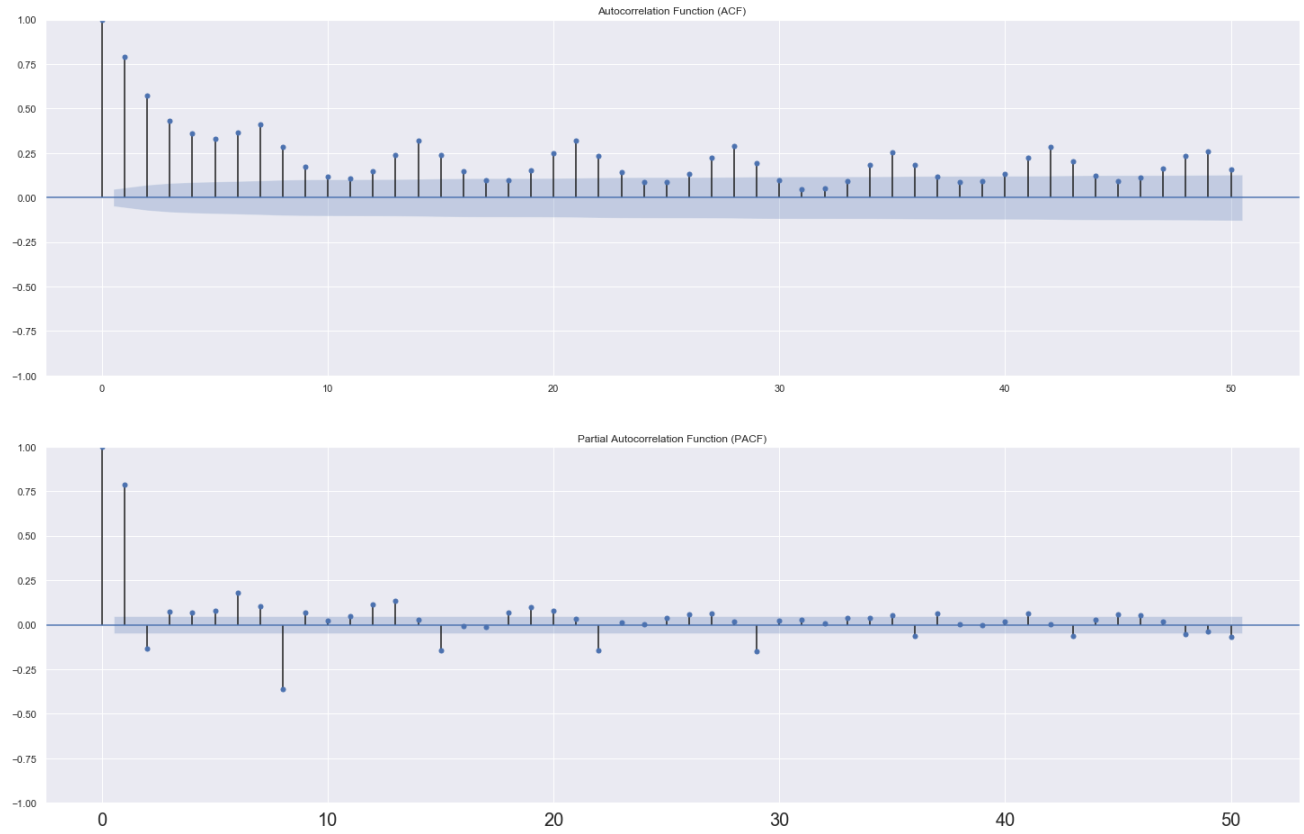

Figure 11: ACF/PACF Plots

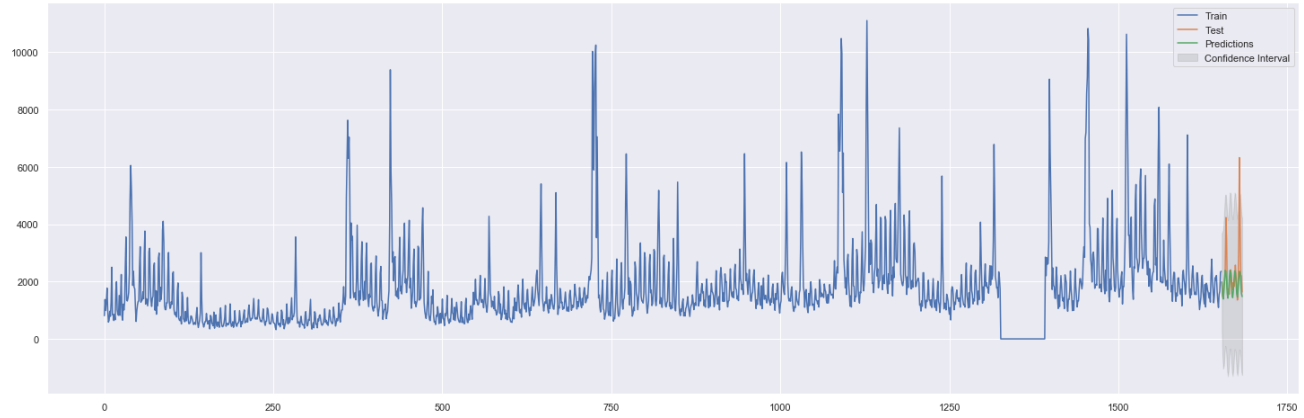

Figure 12: ARIMA Forecast: for next 34 days

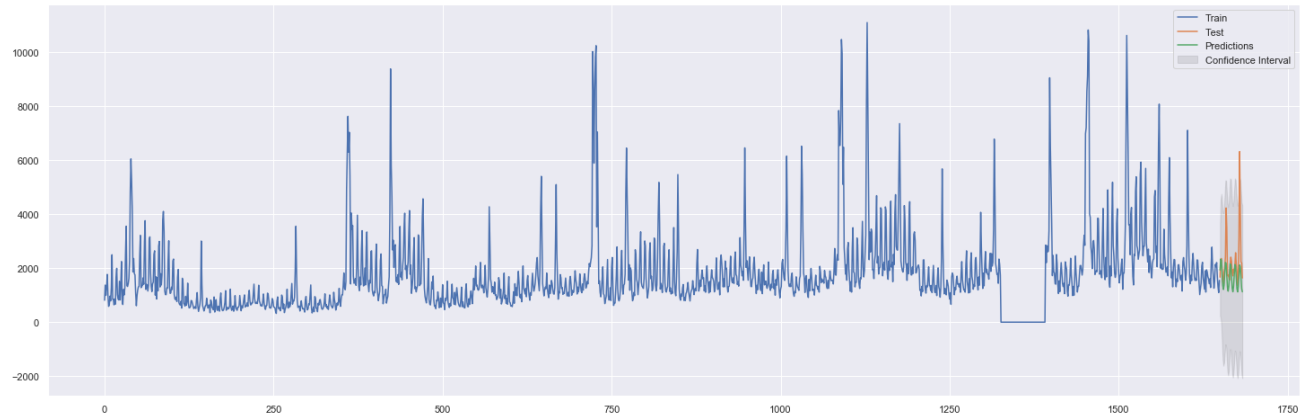

Figure 13: SARIMA Forecast: for next 34 days

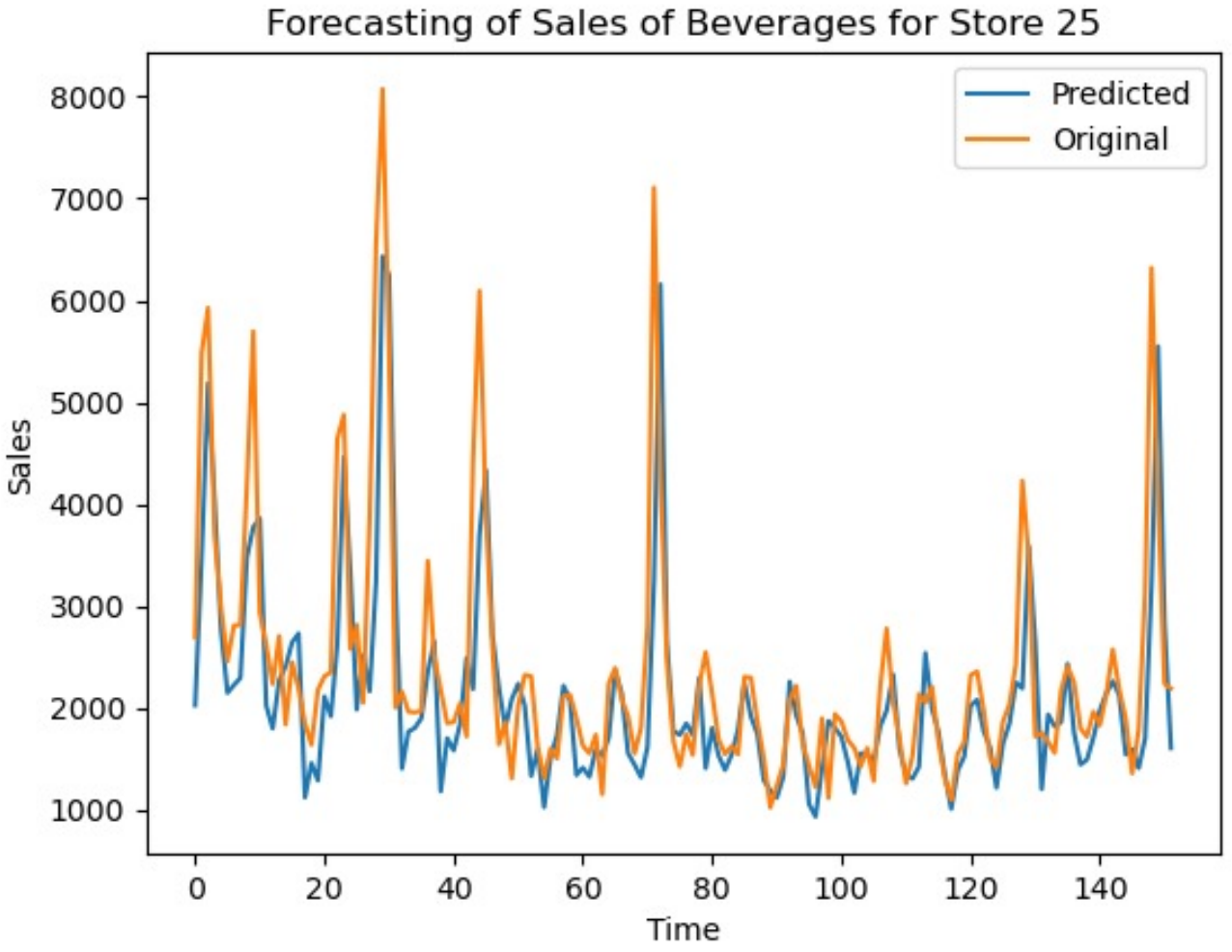

Figure 14: RNN Forecast: for next 34 days

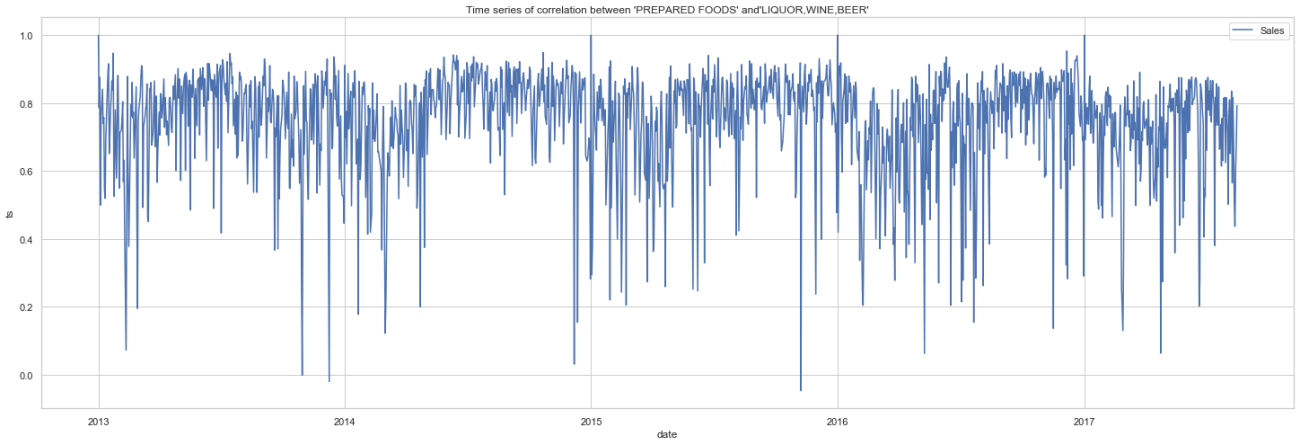

Figure 15: Time series plot of correlation coefficient of 'PREPARED FOODS' and'LIQUOR,WINE,BEER'.

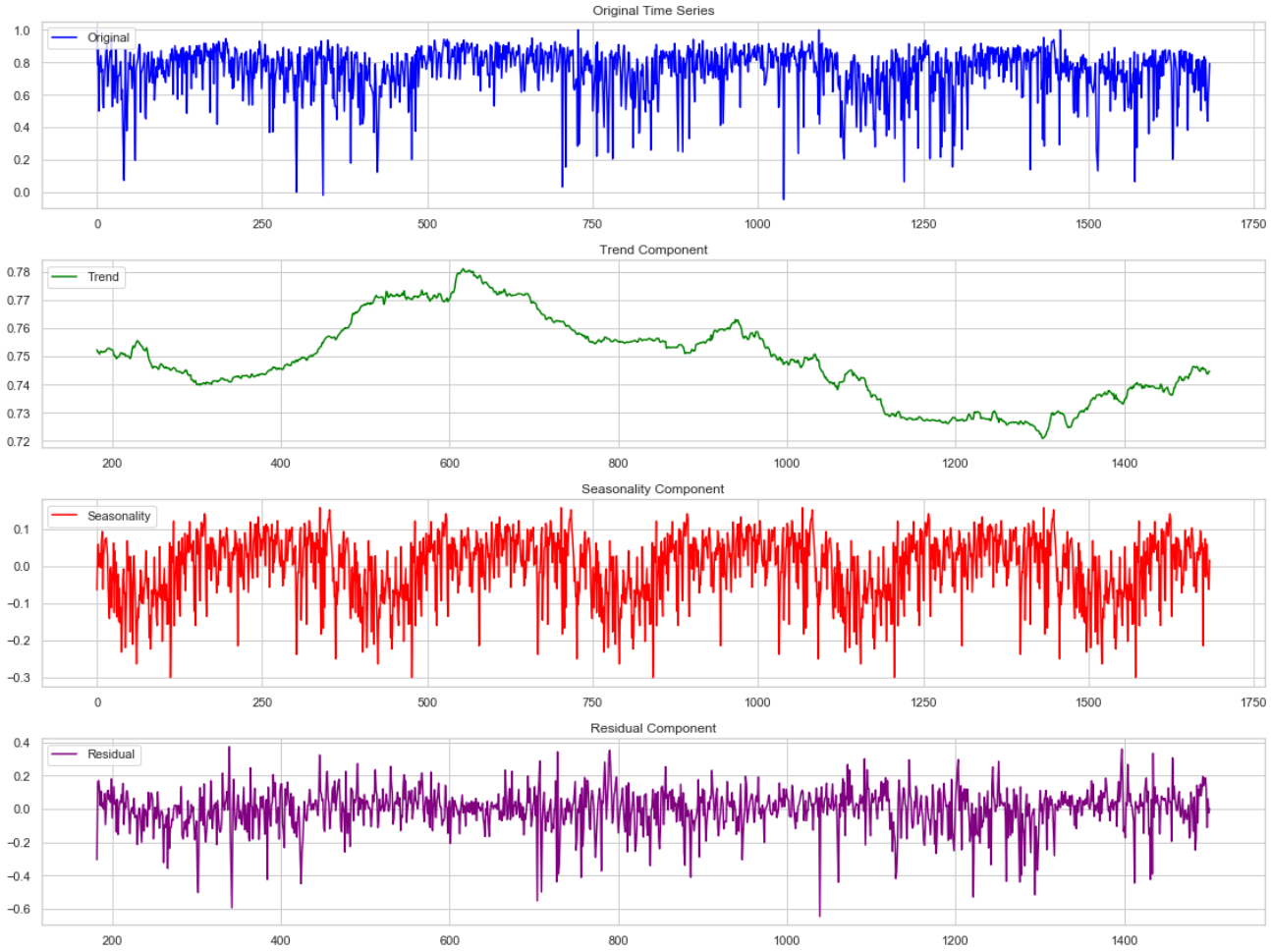

Figure 16: Time series plot of correlation coefficient of 'PREPARED FOODS' and'LIQUOR,WINE,BEER'.