Estimate the Unit Sales of Walmart Retail Foods

1. Problem statement

Walmart, the world's largest company by revenue, operates numerous stores across the United States. The company aims to predict the daily sales of selected food items in three key states: California, Texas, and Wisconsin. Precise sales forecasts are essential for Walmart to maintain optimal inventory levels, preventing stockouts or overstocking. By leveraging these forecasts, Walmart can adopt data-driven approaches for inventory management, make informed pricing choices, and strategically plan promotions based on projected sales trends. This optimization enhances inventory turnover rates and boosts overall operational efficiency, enabling Walmart to maximize revenue while minimizing inventory holding expenses.

This project aims to estimate the point forecasts of unit sales for ten food products sold by Walmart in three US States (California, Texas, and Wisconsin) and provides detailed insights at the item level and specific store attributes. The data contains 1913 days from 2011 to 2016, and the goal is to forecast daily sales for the next 28 days.

2. Data Wrangling

2.1Data Sources

The data is sourced from Kaggle and consists of the following filles: calendar.cvs, sell_prices, sales_train.cvs, and sales_train_evaluation.csv.

- calendar.csv - Contains information about the dates on which the products are sold (2011-2016).
- sell_prices.csv - Contains information about the price of the products sold per store and date.
- sales_train.csv - Contains the historical daily unit sales data per product and store for day-1 to day-1913.
- sales_train_evaluation.csv - Includes sales data for day-1 to day-1941 (contains the next 28 days).

This data contains three categories (Hobbies, Food, and Household) in various departments. For this project, we only use 10 items from the Food category.

2.2 Data Organization and missing values

Upon merging, our dataset is formed, focusing on key features, day, store id, item id, and id of the week (wm_yr_wk).

Moreover, a check for missing values was conducted, sell price, event type and event name have some missing values. The null columns for the event type and name columns are the ones with no event. We fill these cells with 'None'.

There are 14,819 rows with missing values in the 'sell_price' column for the above items/stores and years. First, we fill the missing values with the mean of each year for each item at each store. It appears that there are still 7,751 rows with missing values in the sell_price column for certain items in 2011. Since there is no data available for these items in 2011, we will use the data from the following year to fill in the missing values (we checked other items price and compared the price from 2011 and there is not any significant differences between the prices.)

Furthermore, we are consolidating the information from the 'event_type_1', 'event_name_1','event_type_2', and 'event_name_2' columns into a new column named 'event', where the value is 0 when there is no event ('None' in 'event_type_1' and 'event_type_2') and 1 if there is at least one event.

The data is organized by day in ascending order due to its temporal nature. In total, our dataset contains 13 features.

2.3 Unique Values

Additionally, the number of unique values was determined for certain features:

- Food item: 10
- Stores:10
- States: 3
- years:6

The data contain 10 food items available for sale in 10 stores across three states: 4 stores in California, 3 stores in Texas, and 3 stores in Wisconsin, spanning from the years 2011 to 2016.

3. Exploratory Data Analysis

3.1individual feature distributions

The overall distribution of numeric values is shown in figure 3-1.

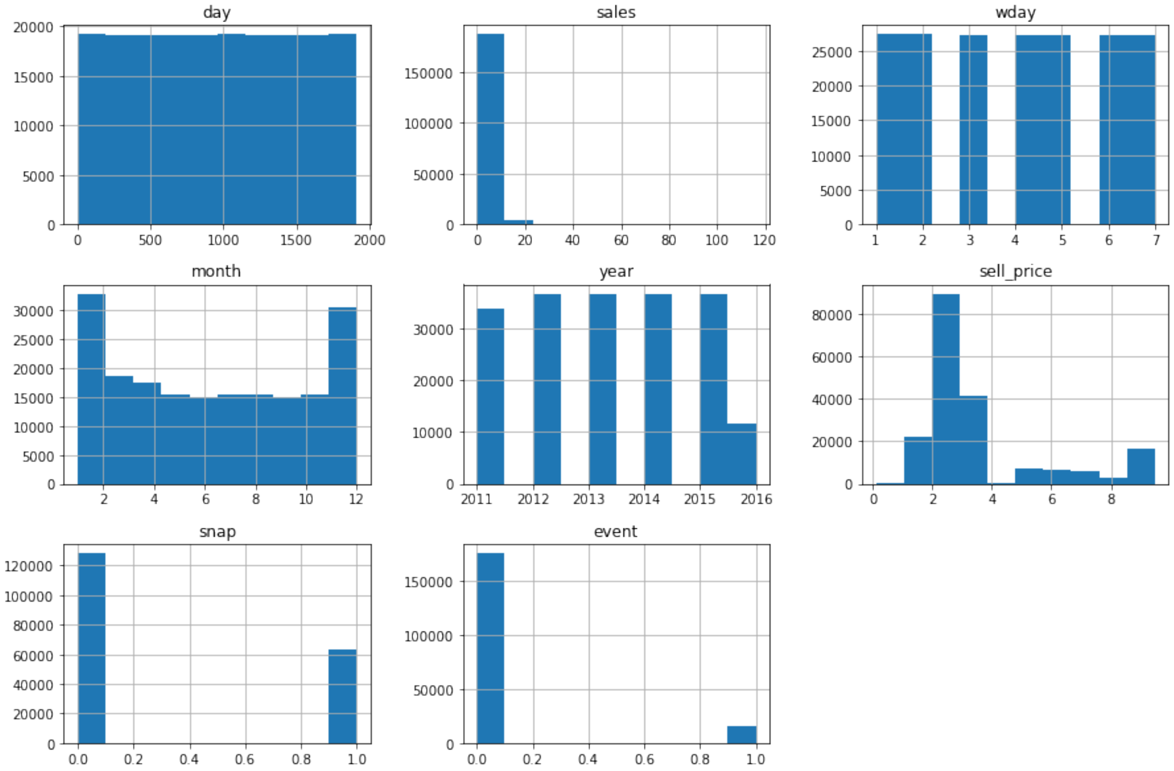

Figure 3-1: Distribution Of Numeric Values

From Figure 3-1, the sales distribution shows that the majority of sales fall below 25 units, accounting for 99.56% of all sales. Only food item 4 stands out with sales exceeding 35 units, while all other items have sales below this threshold. Moreover, unique values for categorical features are shown in Figure 3-2.

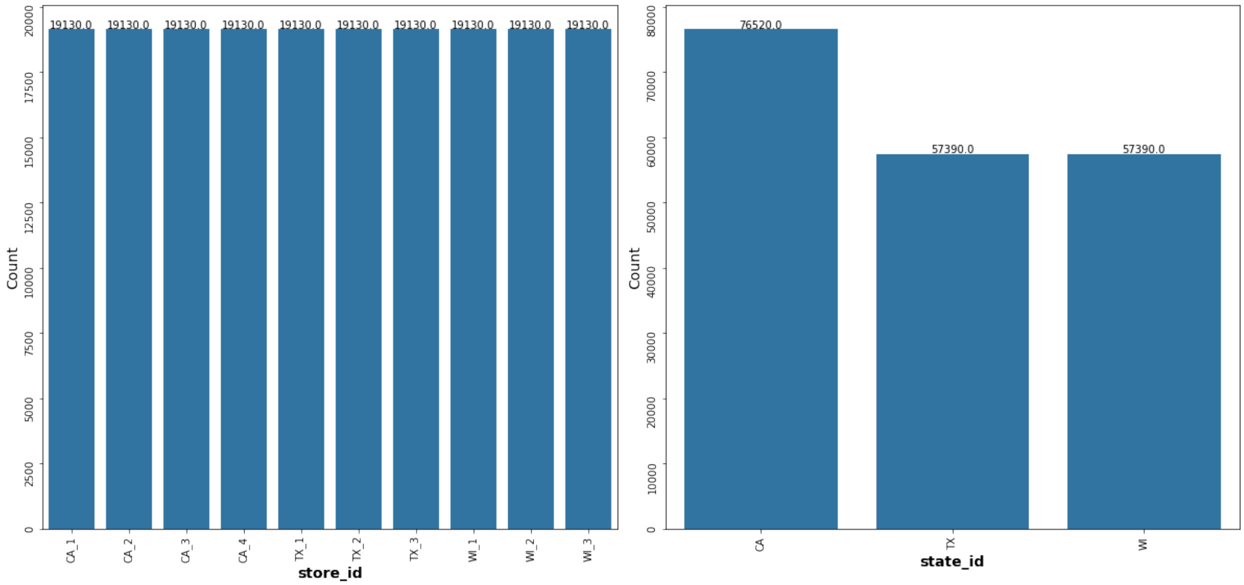

Figure 3-2: Distribution of Unique Values in Categorical Variables

Store and state distribution shows that for all stores, the data is available for all items and dates.

3.2 Summarize daily sales

We visualize the daily sales, to understand the amounts of sales spread over time.

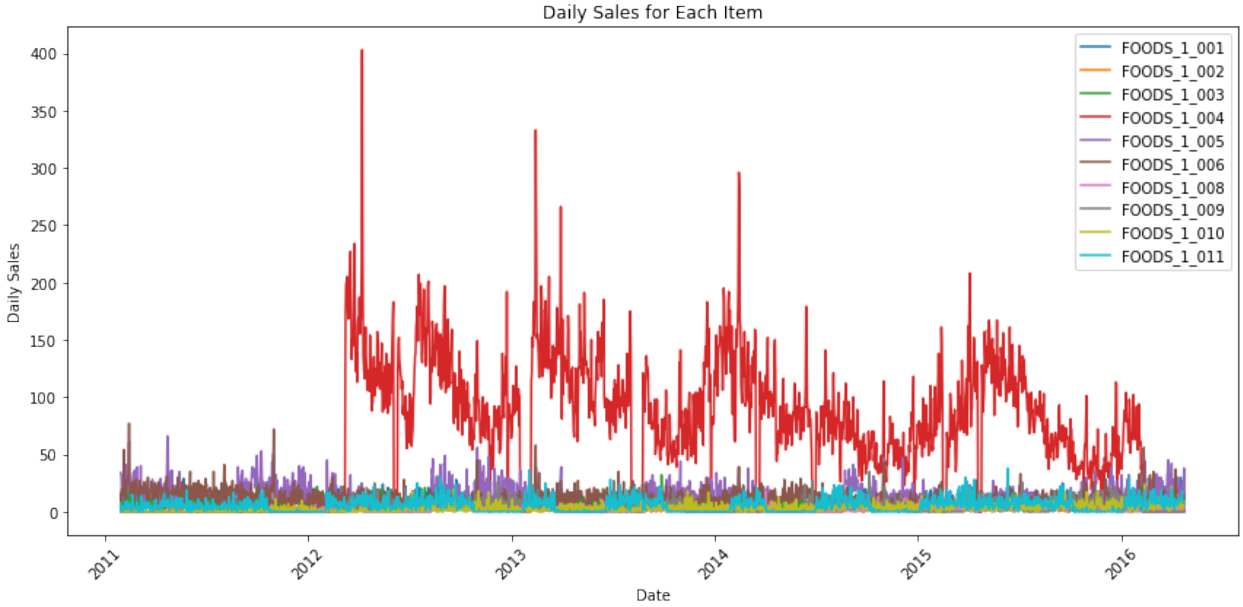

Figure 3-3: Daily Sales of Food Items

The food item_004 has a significantly higher number of sales compared to others. In addition, the figures indicate that sales for certain items, such as Food_004, Food_008, and Food_009, commenced in the year 2012. Also, items like Food_001, Food_005, and others experienced periods with zero sales.

3.3 Summarize monthly sales

Monthly sales of food items show in the following figure.

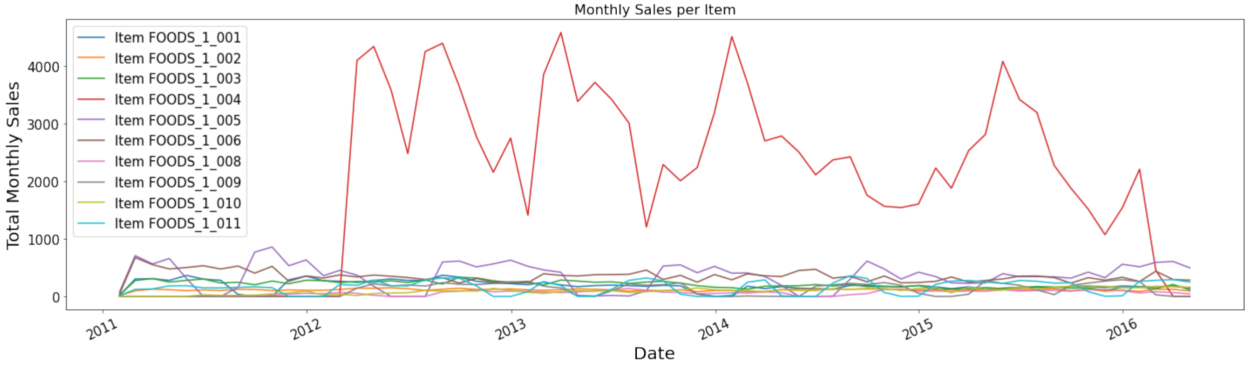

Figure 3-4: Monthly Sales of Food Items

The analysis of monthly sales over multiple years reveals fluctuations in sales volumes without a consistent seasonal pattern, indicating variable sales trends throughout the year across the items.

3.4 Sales per year

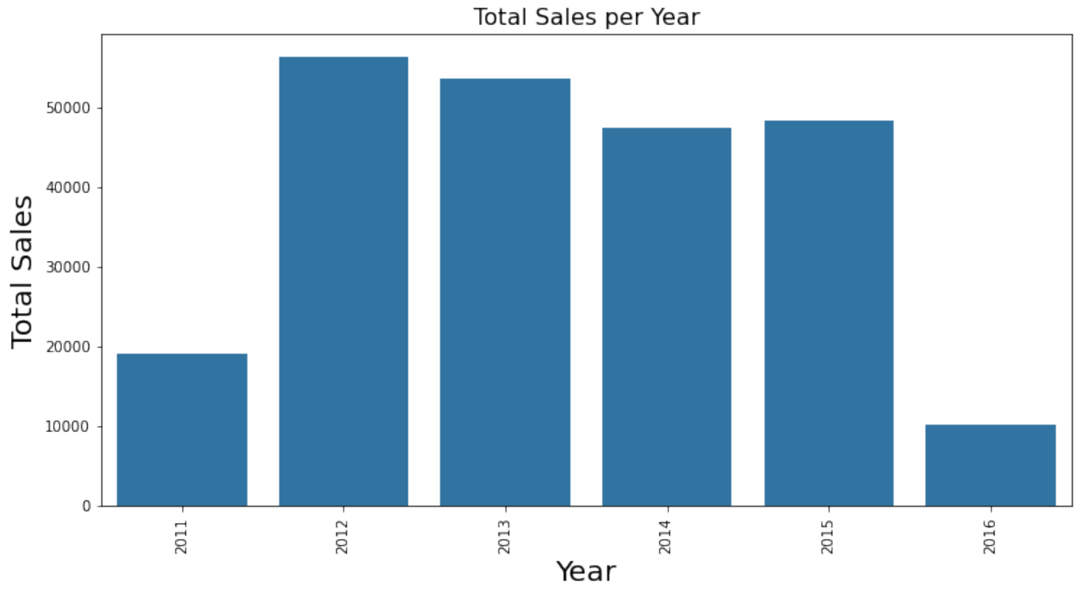

Figure 3-5: Total Sales per year

Figure 3-5 shows the total sales per year. Years 2012 and 2013 have the highest number of sales, respectively.

3.5 Sales per state

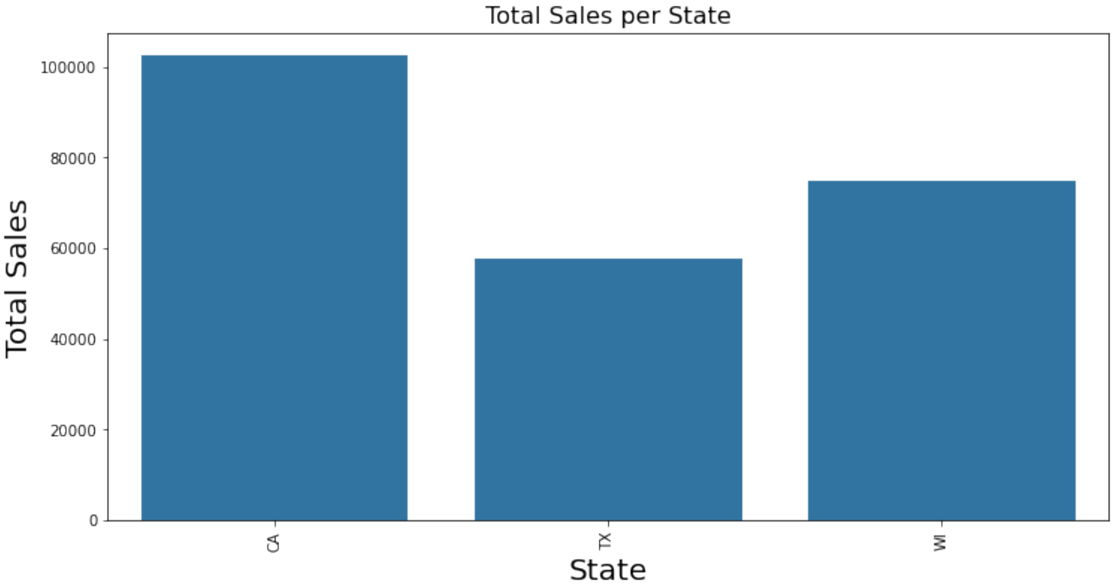

Figure 3-6: Total Sales per State

The state of California has the highest number of sales among Texas, Wisconsin, and California.

3.6 Sales per weekday

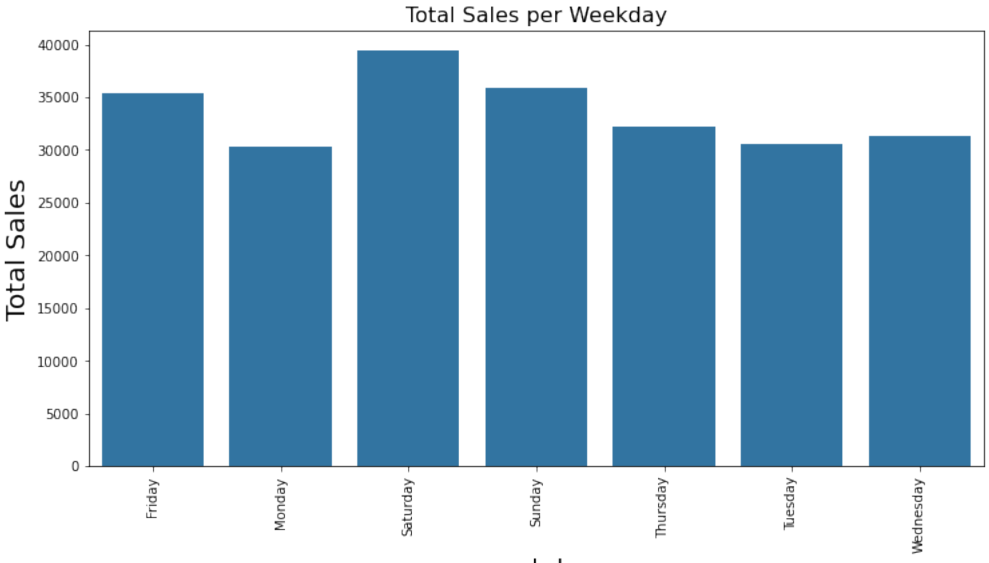

Figure 3-7: Total Sales per Weekday

As expected, Saturday has the highest number of sales among weekdays overall and for most items.

3.7 Sales per store

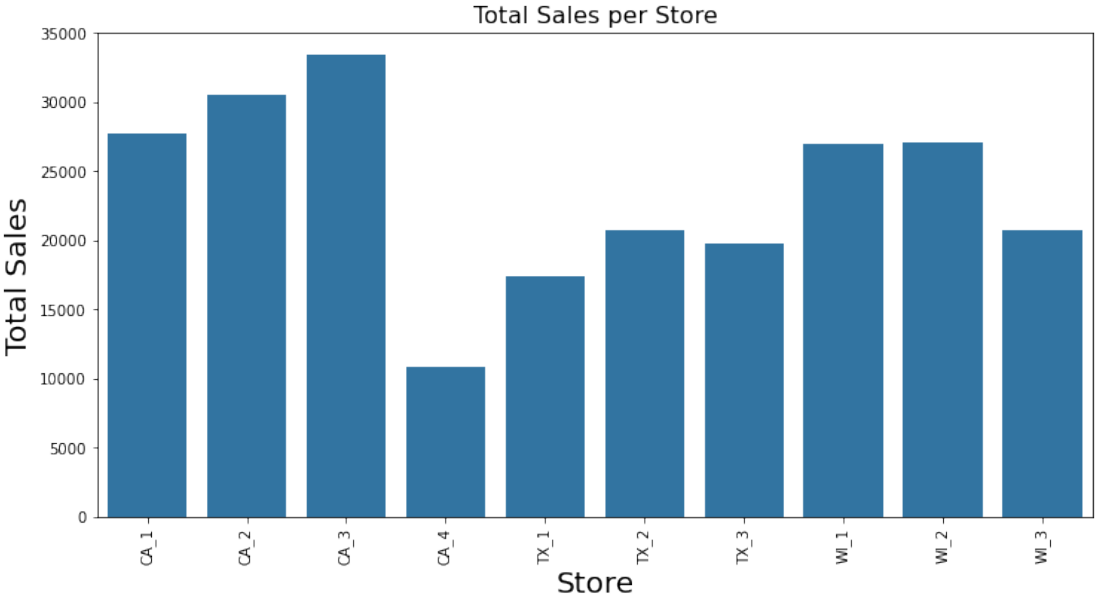

Figure 3-8: Total Sales per Store

In California: CA_3 has sold the greatest number of items while, CA_4 has sold the least number of items.
In Texas: TX_2 has sold the maximum number of items. TX_1 has sold the least number of items.
In Wisconsin: WI_2 has sold the maximum number of items while, WI_3 has sold the least number of items.
In the US the stores CA_3 and CA_2 have the highest number of sales overall while, CA_4 has sold the least number of items.

3.8 Sales per item

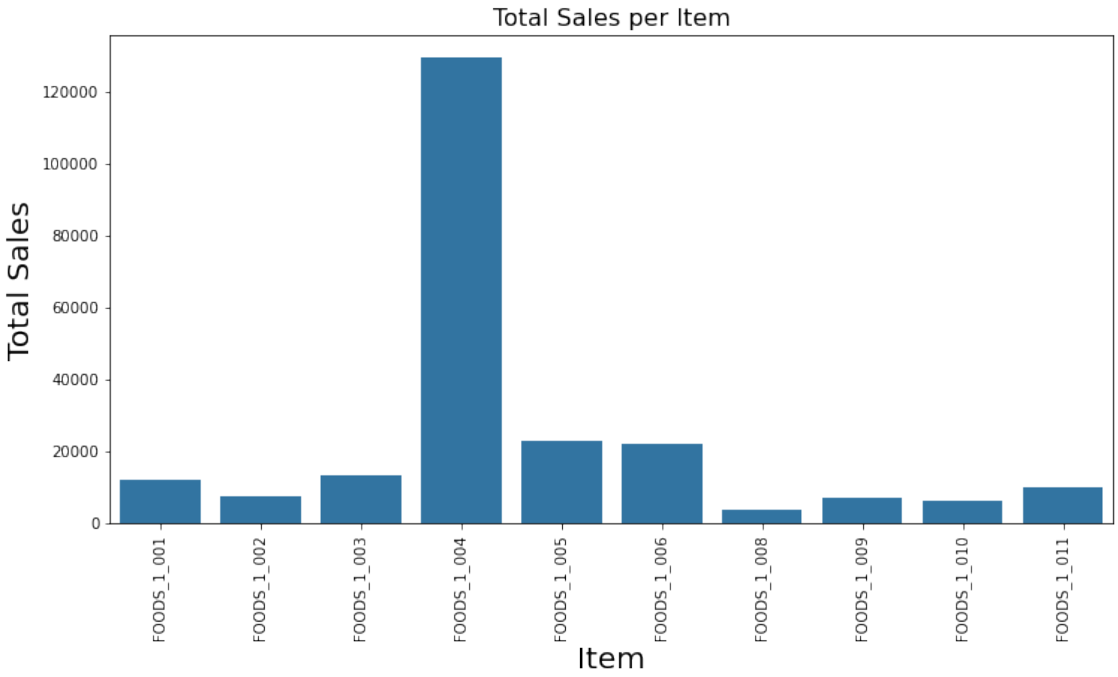

Figure 3-9: Total Sales per Item

It's evident that Food_004 consistently exhibits the highest sales volume across all stores and dates, totaling more than 120,000 sales overall.

By closely examining the number of sales for items, stores, years, and states, we can identify those that consistently contribute the most significant sales figures. This insight is crucial for refining our models and allocating resources effectively, as it directs our focus to key features that have a substantial impact on overall sales distribution. Furthermore, understanding the importance of these specific elements provides a foundation for strategic decision-making and targeted improvements in forecasting and planning processes.

4. Modeling

4.1Data preprocessing and train/test split:

When working with time-series data, the sequential order of data points holds significance. Unlike traditional machine learning models, where a random train-test split is common, time-series data requires a sequential split to maintain the temporal order of observations. This consideration arises from the inherent dependence of future data points on past ones in time series data. Random splits can potentially lead to using future information to predict past events, which is inappropriate.

To address this, we adopt a sequential split for time series, dividing it into training and test sets based on a specified proportion of the data.

For the remainder of the analysis, we will use store TX_1 as an illustrative example.

4.2 Decomposition:

Decomposition of time series data involves breaking down the series into its constituent components, typically including trend, seasonality, and noise (or residual). Time series data often exhibit various patterns over time, such as overall trends, repeating seasonal patterns, and irregular fluctuations. Decomposition helps in identifying and separating these patterns, which can provide valuable insights into the underlying behavior of the data.

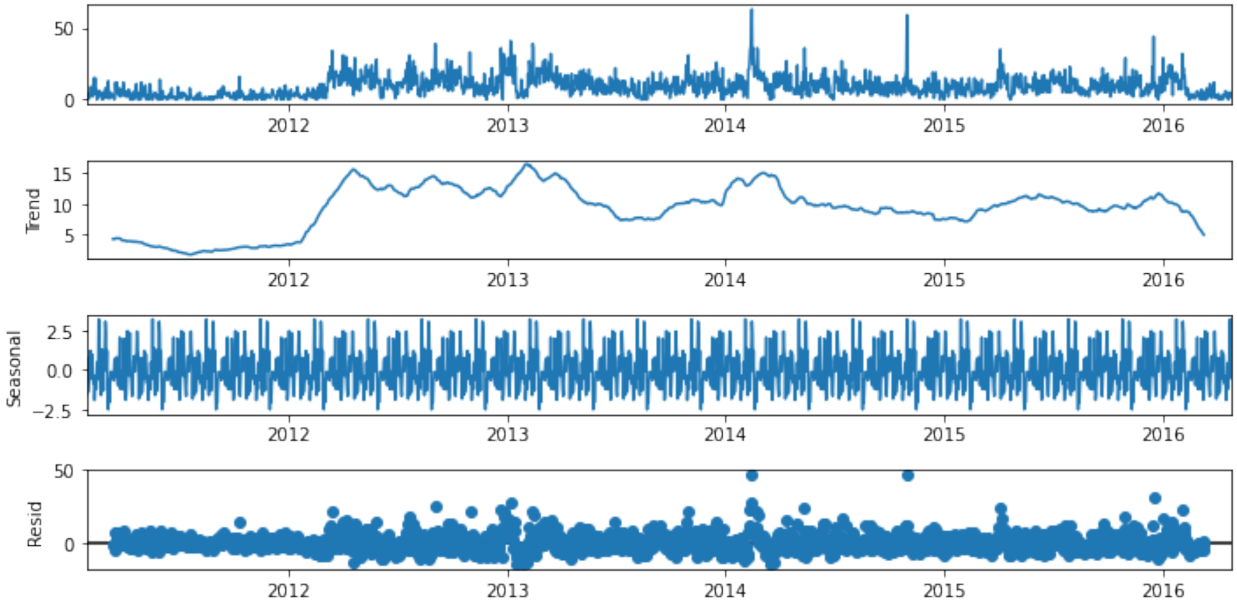

Figure 4-1: Seasonal Decomposition plot of Store TX_1

Original sales (Top Plot): The top plot shows the daily sales over a period of several years. We observe an increasing after 2012 which is due to the availability of some items after 2012 and then even trend over years. There are also noticeable seasonal spikes, particularly during the beginning of each year.

Trend Component (Second Plot): The second plot displays the trend component, which smooths out the short-term fluctuations and highlights the long-term movement. We can see an upward trend specially after 2012.

Seasonal Component (Third Plot): In the third plot, we see the seasonal component that captures the recurring patterns within each year.

Residual Component (Bottom Plot): The bottom plot represents the residual component, which consists of random fluctuations and irregularities not explained by the trend or seasonality. We can observe random spikes or dips in the residual plot, which could be due to one-time events, outliers, or measurement errors. Analyzing the residuals helps us assess the model's goodness of fit and identify any unusual patterns that require further investigation.

4.3Model selection

We studied the performance of four timeseries models: ARIMA model, SARIMA model, ARIMAX model, and Prophet model. The summary of the model results is shown in table 4.1.

Table 4-1 Model Performance Comparison

Models	RMSE	AIC
ARIMA	5.12	11,875
SARIMA	7.86	13,321
ARIMAX	6.05	11,855
Prophet	6.31	-

1- ARIMA Model

ARIMA (AutoRegressive Integrated Moving Average) models are a popular choice for time series analysis due to their ability to capture both the autocorrelation and trend components present in the data. The ARIMA model is versatile and can handle a wide range of time series patterns, including stationary and non-stationary processes.

We found the optimum ARIMA model parameters for TX_1 store. The best p,d,q parameters for our ARIMA model are 2,0,1. The ARIMA(2, 0, 1) model was applied to the sales data of TX_1 store, comprising 1913 observations. The log likelihood of -5932.731 indicates a good fit of the model to the data. The Akaike Information Criterion (AIC) of 11875.462 and Bayesian Information Criterion (BIC) of 11903.244, along with the Hannan-Quinn Information Criterion (HQIC) of 11885.686, suggest that this model strikes a balance between goodness of fit and complexity, with lower values indicating better model performance. The model's coefficients show a significant influence of the autoregressive (AR) and moving average (MA) terms on sales, as evidenced by their low p-values (<0.05). The Ljung-Box test indicates that the model captures autocorrelation well, with a high p-value (0.95) suggesting no significant autocorrelation in residuals. However, the Jarque-Bera test indicates a departure from normality in residuals, which might warrant further investigation. The root mean squared error (RMSE) of 5.12 suggests that the model's predictions are generally close to the actual sales values, considering that the mean sales for this store are around 9. This indicates a reasonably accurate fit of the model to the data, with deviations from the actual values being relatively small.

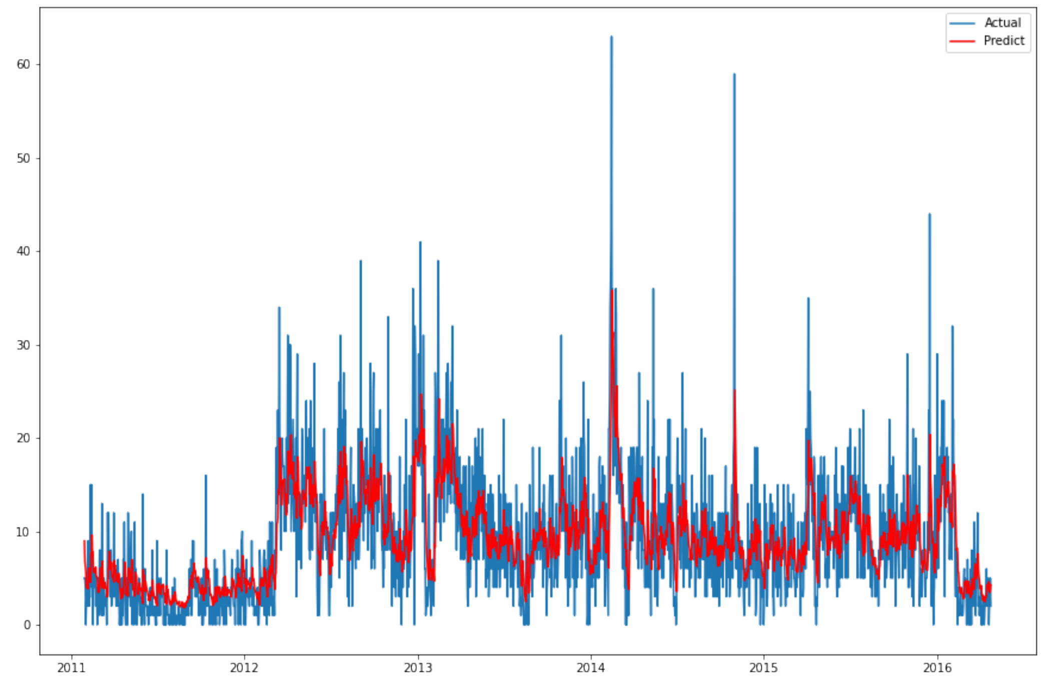

Figure 4-2: ARIMA Model Predictions for Store TX_1

Overall, while the ARIMA(2, 0, 1) model demonstrates strong predictive power and captures autocorrelation effectively, the departure from normality in residuals suggests potential areas for refinement to enhance model accuracy.

To improve the result, we can incorporate additional variables that could influence sales, such as events, SNAP sales, or sell price. This can be achieved by using SARIMA (Seasonal ARIMA) models or including exogenous variables in the ARIMA model.

2- The SARIMA Model

SARIMA model is an extension of the ARIMA model that incorporates seasonality into the time series analysis. Incorporating a SARIMA model can lead to improved accuracy and reliability in forecasting time series data, especially when dealing with seasonal trends that impact the underlying patterns and behaviors.

To find the best parameters for a SARIMA model for TX_1 store, we used grid search, to try different combinations of parameters and evaluate their performance using a metric such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). After the grid search the best parameters are (0, 0, 0) for order and (0, 2, 1) for seasonal order based on AIC. The model's convergence and optimization process indicate a successful fitting procedure. However, the diagnostic tests suggest potential areas for model improvement, such as addressing residual autocorrelation, non-normality, and heteroskedasticity.

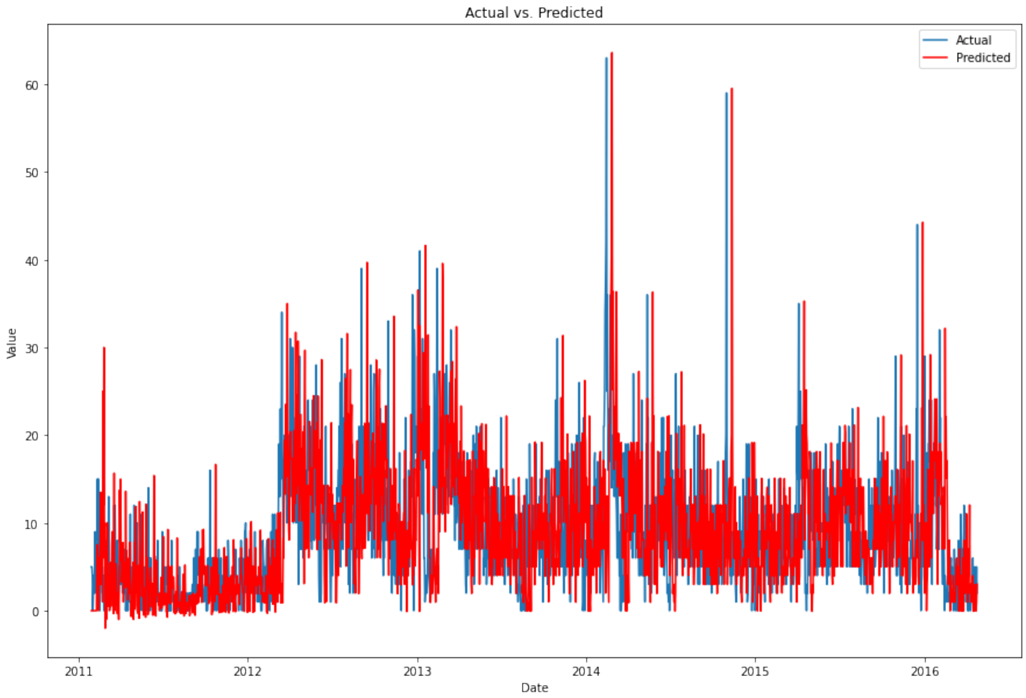

Figure 4-3: SARIMA Model Predictions for Store TX_1

The model's performance metrics (MAE= 5.75 and RMSE=7.86) are reasonable but could be further evaluated in comparison to alternative models. In conclusion, while the model shows promise, further refinement and evaluation may be necessary to determine its effectiveness and suitability for the intended application.

ARIMA and SARIMA comparison:

The ARIMA model demonstrates better performance by its higher log likelihood, lower AIC/BIC values, and lower RMSE, indicating better forecasting accuracy. Both models exhibit satisfactory results regarding residual autocorrelation (Ljung-Box) and display similar levels of heteroskedasticity. An advantage of the ARIMA model lies in its simplicity and ease of interpretation due to its non-seasonal nature. On the other hand, the SARIMA model, while more complex due to its inclusion of seasonal effects, may offer improved accuracy, particularly for datasets with pronounced seasonal patterns. However, both models exhibit some weaknesses such as residual autocorrelation in SARIMAX and non-normality in ARIMA residuals, indicating areas for potential improvement or further model refinement.

3- ARIMAX Model

ARIMAX is an extension of the ARIMA model that allows for the inclusion of exogenous variables, which are additional time series that may influence the sales variable. ARIMAX can capture the impact of external factors like events, or price changes on sales.

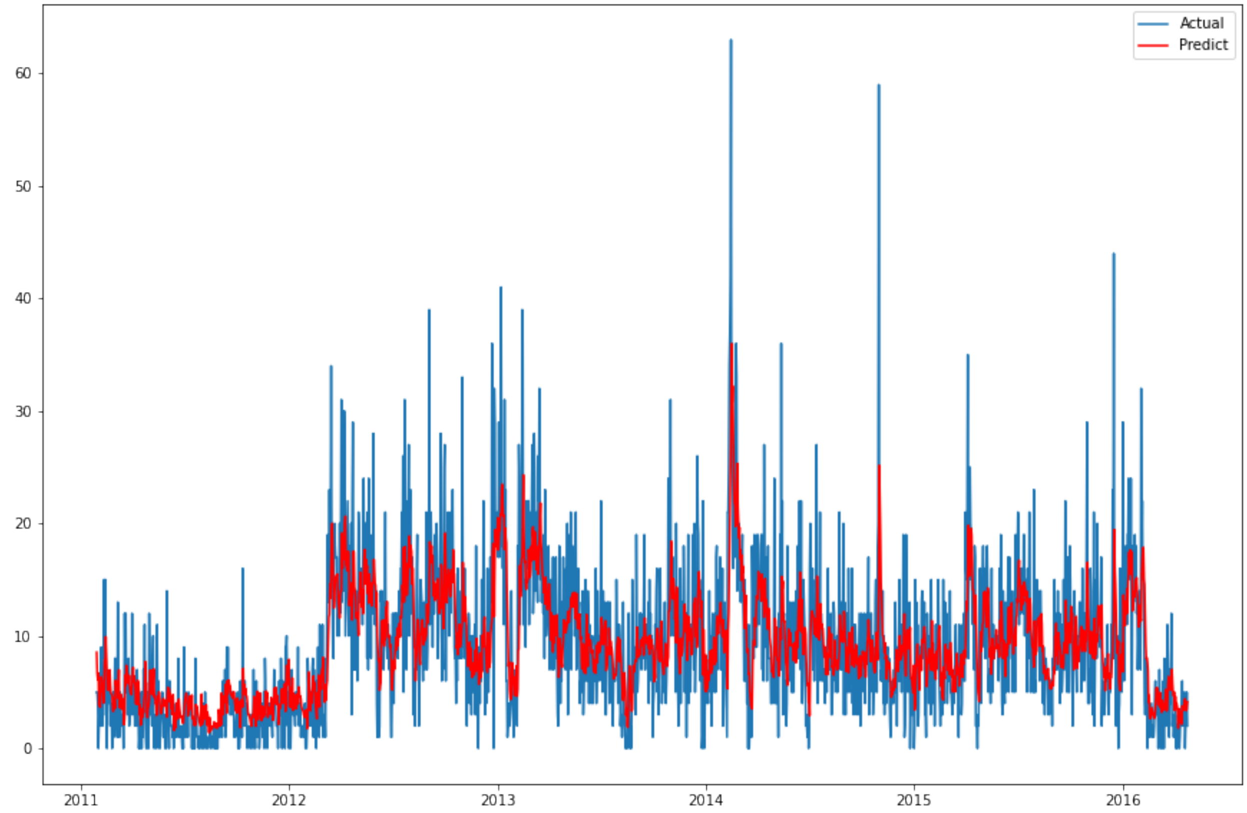

Figure 4-4: ARIMAX Model Predictions for Store TX_1

Best Parameters for ARIMAX model are (1,0,1) for the RMSE of 6.05. The comparison between the ARIMAX (1, 0, 1) and ARIMA (2, 0, 1) models for forecasting sales reveals interesting insights. The ARIMAX model incorporates exogenous variables like events and snap, which slightly increases its complexity compared to the ARIMA model without exogenous factors. Despite this, the ARIMA model performs marginally better with a lower root mean squared error (RMSE) of 5.122, indicating superior predictive accuracy. Both models exhibit relatively similar log likelihood, AIC, and BIC values, suggesting comparable goodness of fit to the data. However, further analysis shows that the ARIMA model's residuals have a slightly better behavior, as indicated by the Ljung-Box test's p-value of 0.00, indicating no significant autocorrelation, and the Jarque-Bera test's p-value of 0.00, indicating non-normality in residuals. Therefore, based on these evaluations, the ARIMA model appears to be a more suitable choice for forecasting sales in this scenario.

4- The Prophet Model

Prophet model Developed by Facebook, is designed to handle time series data with strong seasonal effects and multiple seasonality. It's robust to missing data and outliers, making it a popular choice for many applications. Additionally, Prophet can handle datasets with many zero values, which makes it particularly suitable for our data, where many zero sales occurrences are present.

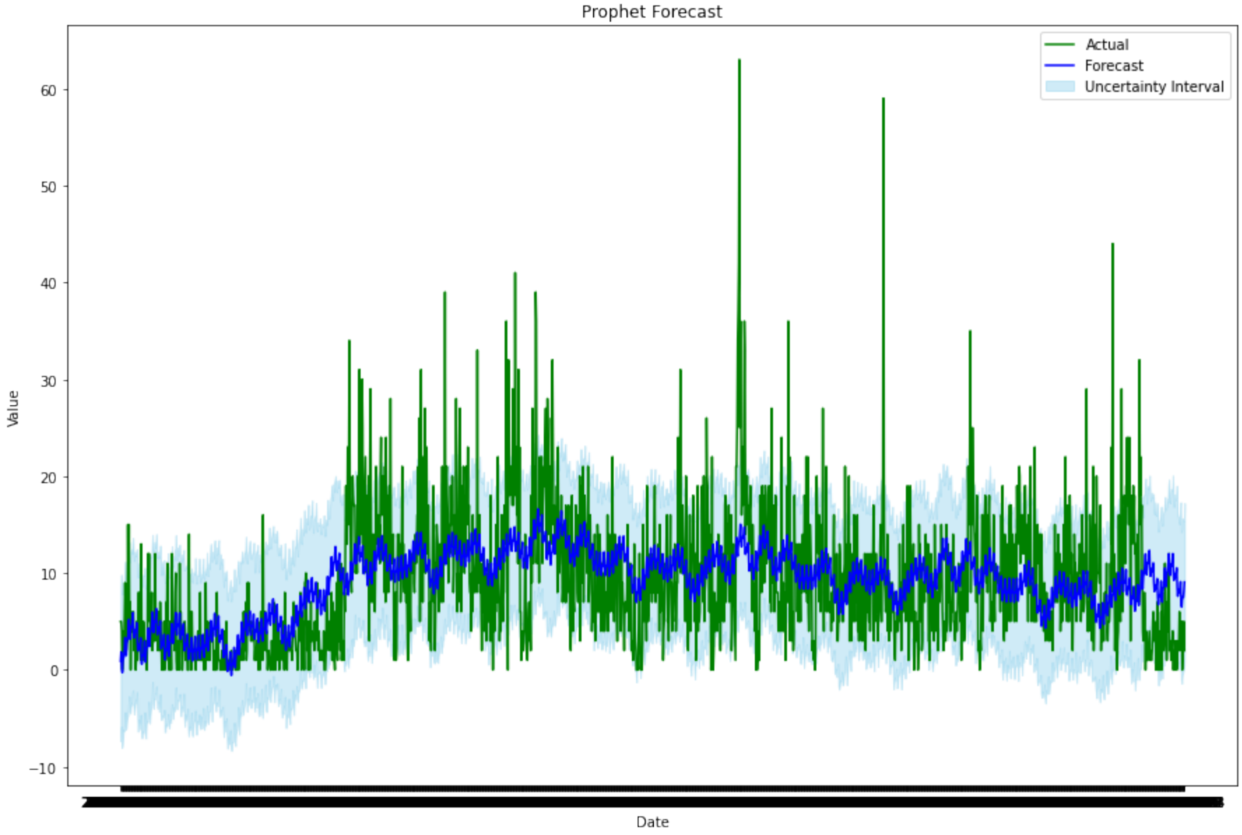

Figure 4-5: Prophet Model Predictions for Store TX_1

The tunning process on 'changepoint_prior_scale', 'holidays_prior_scale', 'seasonality_mode', and 'seasonality_prior_scale' variables show the best model is achieved for the following parameters:{ 'changepoint_prior_scale': 0.01, 'holidays_prior_scale': 0.1, 'seasonality_mode': 'multiplicative', 'seasonality_prior_scale': 0.01}

The RMSE value of 6.31 for the best model indicates that, on average, the Prophet model's predictions are off by approximately 6.31 units from the actual values. In other words, the model's accuracy in predicting the target variable is moderate.

The R-squared (R2) value of 0.18 indicates that approximately 18% of the variance in the target variable is explained by the model. A higher R-squared value closer to 1 indicates a better fit of the model to the data, while a lower value suggests that the model does not explain much of the variance in the data.

Overall, based on these metrics, the Prophet model performs moderately in predicting the target variable, capturing some but not all of the variability in the data. Further model evaluation and refinement may be necessary to improve predictive performance.

4.4Summary and Final Model Selection

Based on the analysis and comparison of different time series models, including ARIMA, SARIMA, ARIMAX, and Prophet, the ARIMA model emerges as the most suitable choice for predicting the unit sales. The ARIMA model demonstrated strong predictive power, capturing both autocorrelation and trend components effectively. It achieved a relatively low root mean squared error (RMSE) of 5.12, indicating accurate predictions compared to actual sales values. However, a comprehensive comparison would require additional metrics and evaluation on test data to determine the most suitable model. The computational complexity of each model can vary. Prophet is known for its ease of use and automatic feature selection, making it simpler for users. On the other hand, ARIMA and SARIMA models may require more manual tuning and expertise but can offer flexibility in modeling various time series patterns.

5. Future direction:

Moving forward, several avenues can be explored to improve the forecasting models and enhance decision-making regarding inventory management and sales optimization:

1. **Feature Engineering:** Incorporate additional features such as promotional events, seasonal trends, competitor pricing, and economic indicators to capture more comprehensive factors influencing sales variability.
2. **Ensemble Modeling:** Explore ensemble techniques that combine the strengths of multiple models (e.g., ARIMA, SARIMA, and Prophet) to leverage their collective predictive power and mitigate individual model weaknesses.
3. **Machine Learning Algorithms:** Explore advanced machine learning algorithms like Gradient Boosting Machines (GBM), Long Short-Term Memory (LSTM) networks, or neural networks to capture complex temporal patterns and nonlinear relationships in the data.