CECS 551
ADVANCED ARTIFICIAL INTELLIGENCE
FINAL PROJECT

Submitted by:

ANUSHA VANGALA

Contents

1. INTRODUCTION

This report offers a detailed analysis of the inventory data for a global coffee shop chain boasting over 1000 stores and $400 million in revenue. The primary goal is to enhance overall sales performance and operational efficiency by leveraging insights from sales predictions and inventory alignment, ultimately reducing waste. In the course's final project, a meticulous examination of data from 132 stores, strategically dividing responsibilities. Focusing on specific stores (StoreID = 18, 117, 332) and the others delving into the collective data. Despite the segmented focus, our team operated cohesively, emphasizing the importance of consistent findings and a unified strategy. The report is structured into two sections, detailing individual store analyses, and presenting a collective view, with the aim of providing actionable recommendations for a more sustainable and successful future for the coffee shop chain.

2. Problem Statement:

How can the analysis of inventory and sales data from a global coffee shop chain's 132 stores, which operates over 1000 stores worldwide and generates $400 million in revenue, be utilized to optimize inventory management, minimize waste, and enhance sales effectiveness? This project aims to explore the potential of using predictive modeling for sales forecasting, understanding the impact of various external factors like weather conditions and drive-thru availability on sales, and leveraging these insights for strategic decision-making to gain a competitive edge, improve customer satisfaction, and contribute to environmental sustainability in a highly competitive retail market.

3. INSIGHIT'S ON THE DATASET:

The dataset you're working with offers a comprehensive view of sales and inventory management within a coffee store chain, enriched with environmental and operational details. It includes vital information such as 'StoreID', 'BusinessDate', 'PLU', item descriptions, and various inventory quantities like 'ReceivedQuantity', 'SoldQuantity', and 'EndQuantity'. Notably, it also incorporates elements like 'Temperature', 'WeatherCondition', and 'DriveThru' presence, which are instrumental in assessing how external factors impact sales. Analyzing this data could reveal customer preferences, the efficiency of inventory management, the influence of weather on sales trends, and the effect of store features like drive-thrus on sales dynamics. Additionally, the inclusion of 'BusinessDate' allows for the exploration of seasonal patterns, aiding in effective demand forecasting and strategic planning. This rich dataset is thus a valuable asset for in-depth analysis, offering insights that could drive informed business decisions.

Data Description

- StoreID: ID of store
- BusinessDate: Date of record
- PLU: ID of inventory
- Description: Name of product
- ItemType: Type of product
- CategoryLvl1Desc: Main level of product's category
- CategoryLvl2Desc: The 2nd level of product's category
- CategoryLvl3Desc: The 3rd level of product's category
- ReceivedQuantity: The amount stores received from the distributor. They typically receive perishable items every 2 or 3 days based on the customer's feedback. But it could be every day or more than 3 days for some products.
- SoldQuantity: Quantity sold on a particular day.
- EndQuantity: Quantity at the end of the day – inventory condition. (Please note if they throw away expiring items, they record zero at the end of the day and they order them again to have them the next day.)
- LatestOrder: The number of items they requested.
- StockedOut: They record when a customer asks for an out-of-stock item. Yet, the cashier may not always record all the stocked out. But if they do, we are sure of the case.
- GroupID: Not Applicable
- MissedSales: Not sure what these data signify.

Additionally Added Feature

- Temperature : Randomly generated
- WeatherCondition:Randomly and equally distributed
- DriveThru: Generated based on the sales.

4. Key Analysis Areas

Sales Trends and Customer Preferences: By analyzing 'SoldQuantity' and item descriptions, we can identify popular items and emerging trends in consumer preferences.

Inventory Management: Using 'ReceivedQuantity', 'EndQuantity', and 'MissedSales', we can evaluate the effectiveness of current inventory strategies and identify areas for improvement.

Impact of Environmental Factors: Correlating 'Temperature' and 'WeatherCondition' with sales data can reveal how different weather scenarios affect consumer purchasing behavior.

Influence of Store Features: The data on 'DriveThru' availability can help understand how different store formats influence sales, especially under varying environmental conditions.

Seasonal Patterns and Forecasting: Utilizing 'BusinessDate', the dataset allows for an examination of seasonal sales patterns, crucial for accurate demand forecasting.

5. Methodology

1. Data Preprocessing
Cleaning: This step involves removing or correcting erroneous data, dealing with missing values, and identifying outliers. We will employ techniques like imputation for missing values and outlier detection methods to ensure data quality.
Normalization: To bring different scales of numerical data to a common scale without distorting differences in the ranges of values. For example, Min-Max scaling will be used for features like 'Temperature' and 'ReceivedQuantity' to normalize them between 0 and 1.
Categorization: Categorical data, including 'ItemType', 'WeatherCondition', and 'DriveThru' status, will be encoded suitably. This can involve one-hot encoding for nominal data and ordinal encoding for data with a specific order.

2. Statistical Analysis
Descriptive Statistics: Computing basic statistical measures such as mean, median, mode, standard deviation, and quartiles for various columns to understand data distribution and central tendencies.
Trend Analysis: Identifying trends over time, particularly using the 'BusinessDate' column to see how sales and inventory levels change across different seasons or specific timeframes.
Anomaly Detection: Utilizing statistical techniques to identify unusual patterns or outliers in sales or inventory data that could indicate issues or extraordinary events.

3. Predictive Modeling
Feature Selection: Identifying relevant features that influence sales and inventory needs, such as 'Temperature', 'WeatherCondition', and 'ReceivedQuantity'.
Model Development: Building machine learning models like linear regression for basic forecasting, or more complex models like ARIMA (AutoRegressive Integrated Moving Average) and LSTM (Long Short-Term Memory) networks for time-series forecasting.
Model Validation: Using techniques like cross-validation and evaluating models with metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to ensure the reliability and accuracy of the forecasts.
Hyperparameter Tuning: Adjusting model parameters to improve performance, using methods like grid search or random search.

4. Correlation Analysis
Multivariate Analysis: Exploring complex interactions between multiple variables, like how 'Temperature', 'WeatherCondition', and 'DriveThru' jointly impact sales.
Heatmaps and Correlation Matrices: Utilizing visual tools to identify and represent the strength and direction of relationships between different variables.

6. Section 1 : Data Visualization

In this section, our focus is on employing various data visualization techniques to provide a comprehensive understanding of the inventory data for three specified stores (StoreID=18, 117, 332). Visualization serves as a powerful tool to unravel patterns, trends, and anomalies within the individual datasets. We leverage graphical representations such as line charts, bar graphs, and pie charts to illustrate sales trends over time, identify popular products, and visualize the distribution of inventory across categories. Through these visualizations, we aim to offer a clear and intuitive portrayal of the specific dynamics within each store, aiding in the identification of key performance indicators and areas for potential improvement. To provide box plots and statistics of 27 products, including their inventory patterns, stock out patterns, and an estimate of missed sales, you would proceed through several analytical steps.

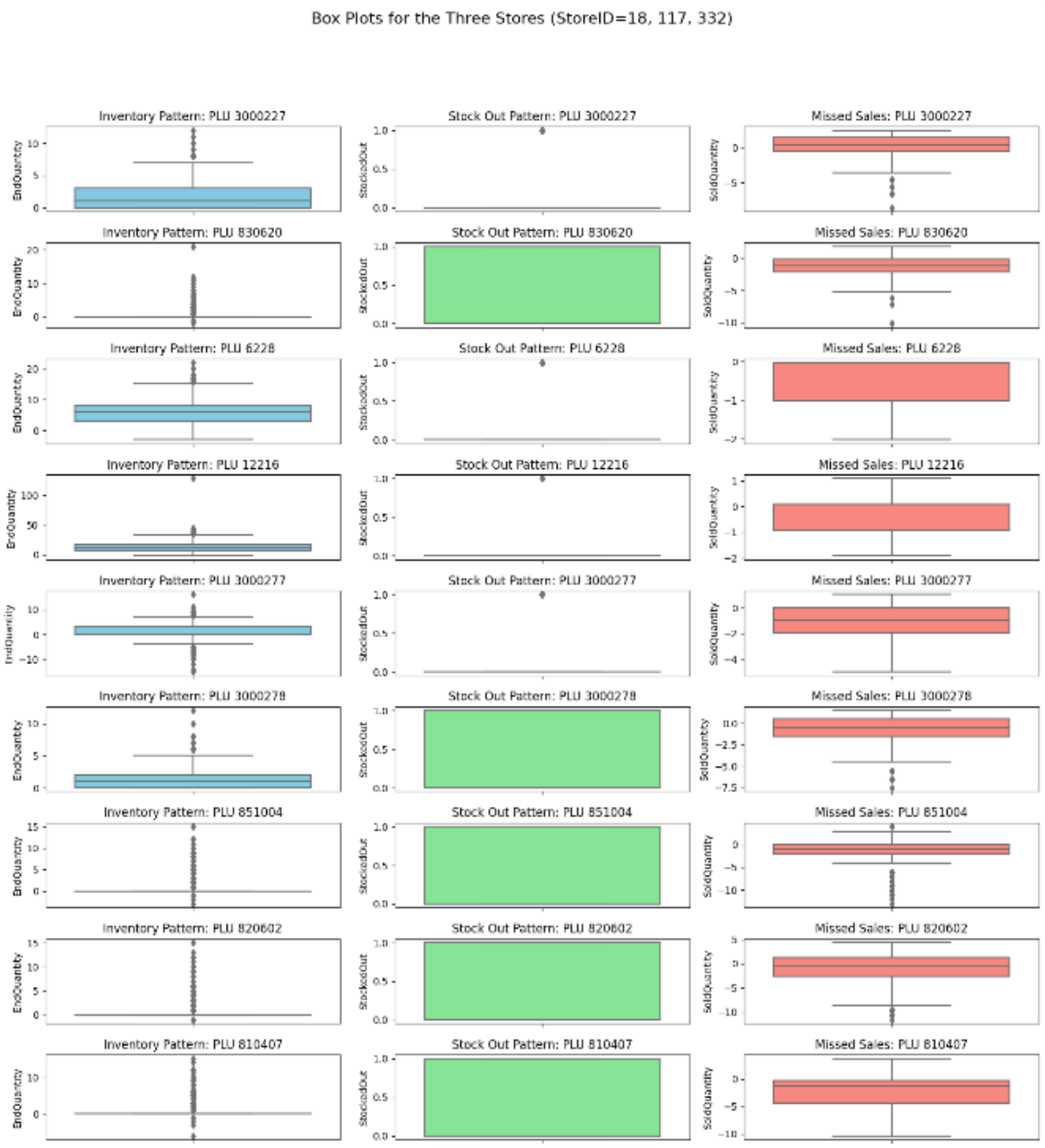
6.1 Individual Store Data Analysis:

6.1.1 Box Plots and Measurements for 27 Items.

We conducts a comprehensive analysis stores based on a dataset ("Stores_Data.csv") played out a complete examination.

The plot represent data on inventory levels, stock availability, and sales for different products, each identified by a unique PLU code. From the 'Inventory Pattern' plots, we can begin to understand the effectiveness of inventory management at each store, identify which products are most prone to stockouts, and estimate the potential impact on sales. This multi-

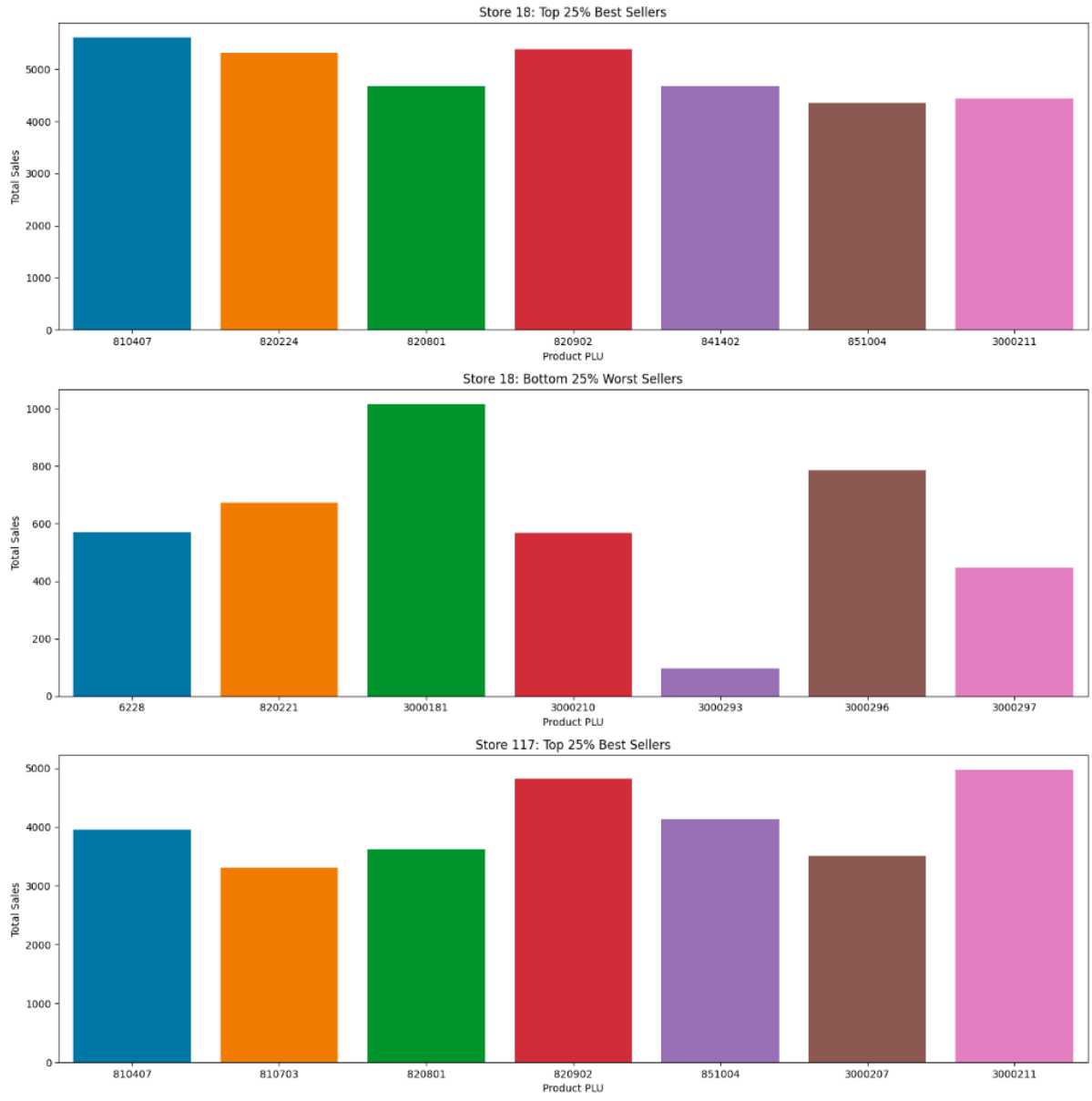
dimensional analysis could be vital for a retail chain to pinpoint specific issues, such as supply chain weaknesses or forecasting inaccuracies, at the store level and then strategize improvements tailored to each store's needs and customer demand patterns.

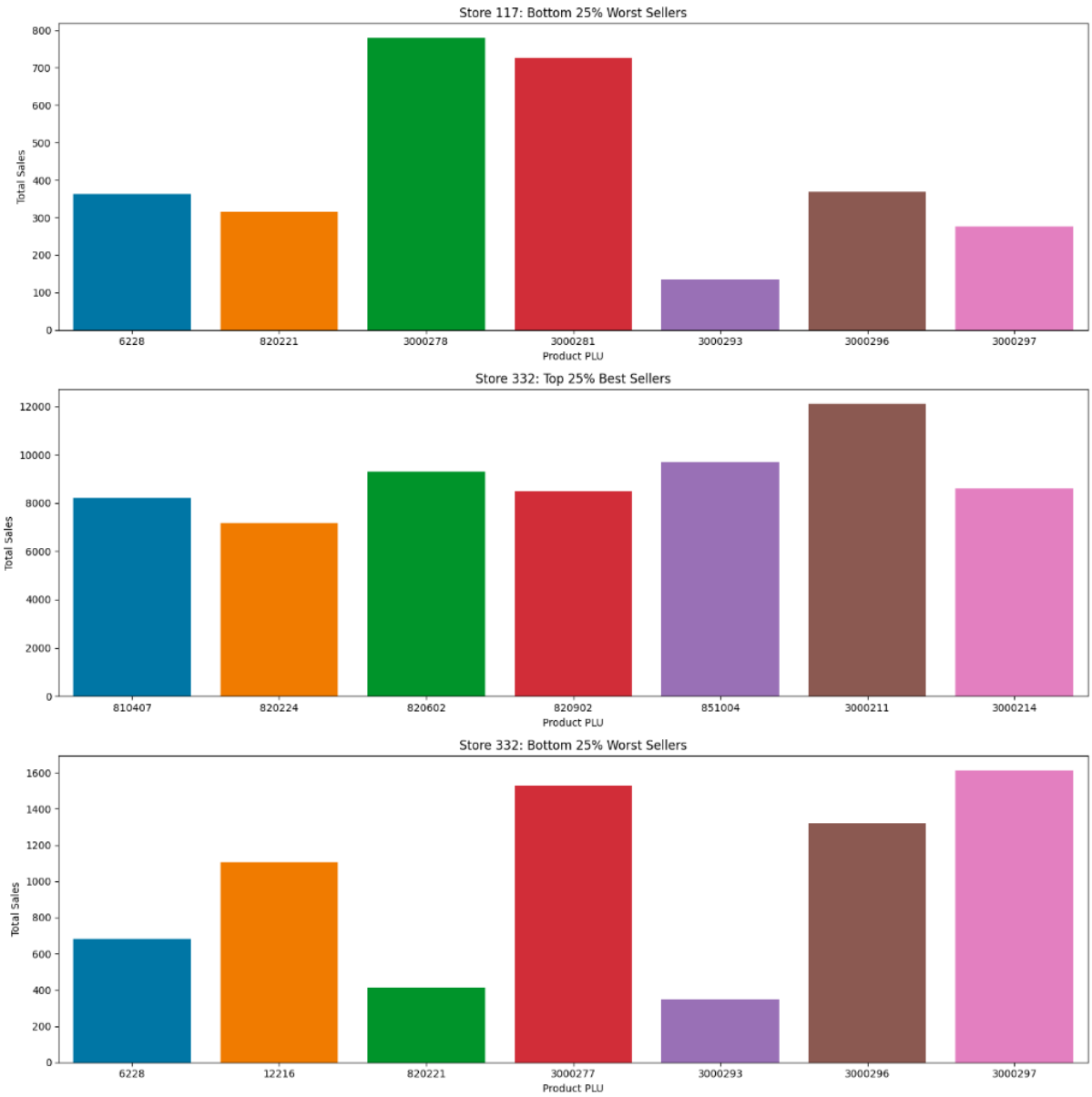


6.1.2 Best and Most terrible Dealer Items

In our examination, we can infer sales performance patterns for three distinct stores, labeled with the IDs 18, 117, and 332. Each store's chart is divided into two segments, representing the top 25% of best-selling products and the bottom 25% of worst-selling products, classified by their respective Product PLU numbers. The contrast between the best and worst sellers within each store could indicate consumer preferences, the success of sales strategies, or the

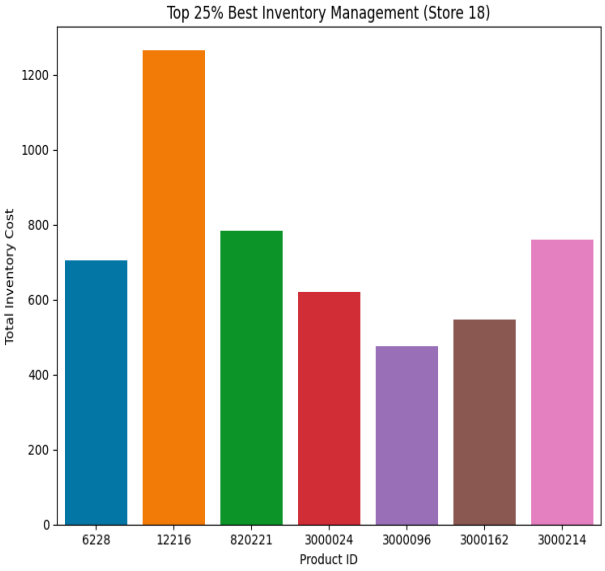
effectiveness of inventory management. Furthermore, by comparing across stores, we can infer commonalities or discrepancies in sales trends, which may point to broader market forces at play or specific local demands. If certain PLUs are consistently among the top sellers or the bottom sellers across different stores, it suggests that these products may universally appeal to or fail to meet customer expectations, respectively. This kind of analysis can be instrumental for inventory planning, targeted promotions, and optimizing product offerings to enhance store performance.

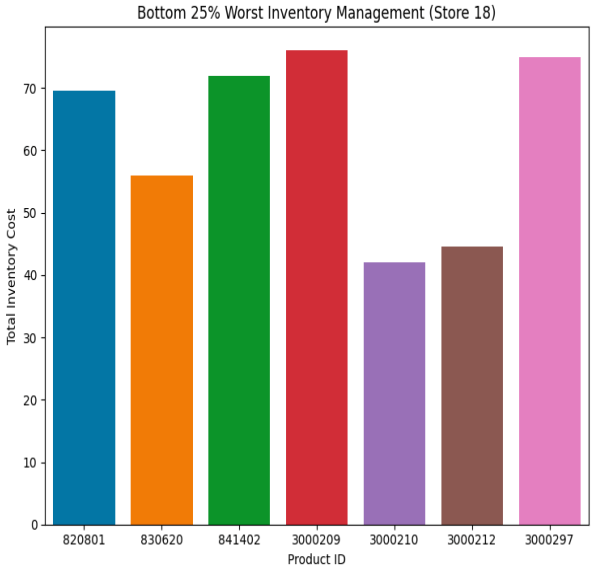


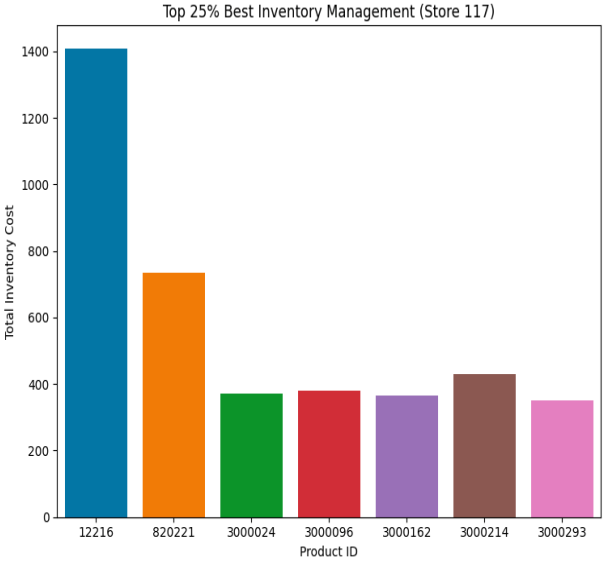


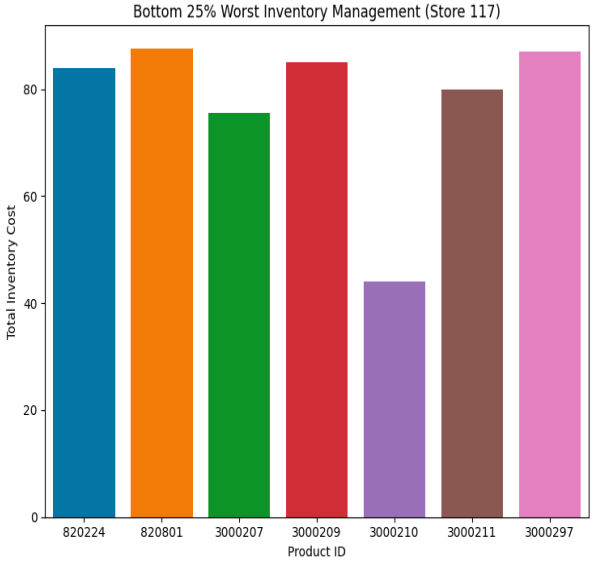
6.1.3 Stock Administration Investigation

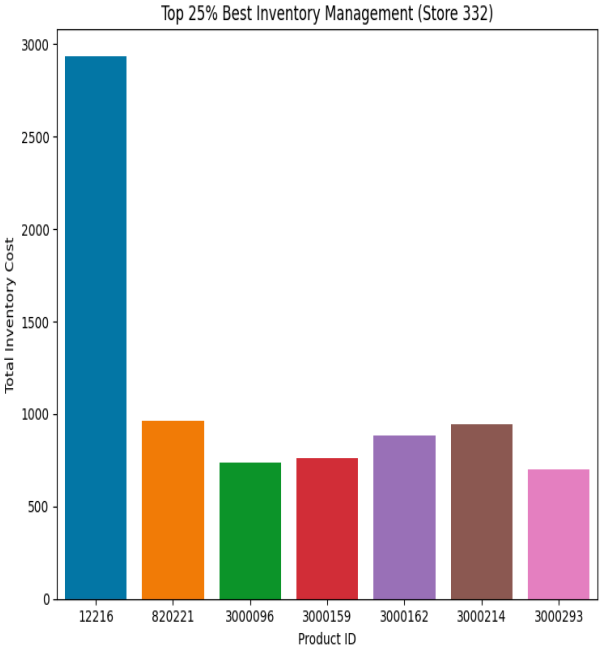
We can understand from the bar charts of Stores 18, 117, and 332 that they display a comparative analysis of inventory management efficiency, segmented into the top 25% best-performing products and the bottom 25% worst-performing products by inventory cost. The charts for the top 25% show products that have higher inventory costs, indicating items that might be high in value, have fast turnover, or are strategically kept in large quantities due to high demand or sales frequency. On the other hand, the bottom 25% charts depict products with lower inventory costs, possibly signifying less efficient inventory practices, such as overstock of slow-moving items or underperformance in sales. These visual data sets allow for an evaluation of inventory management's impact on store operations and can highlight areas for improvement, such as optimizing stock levels, adjusting procurement practices, or revising sales strategies to enhance overall performance.

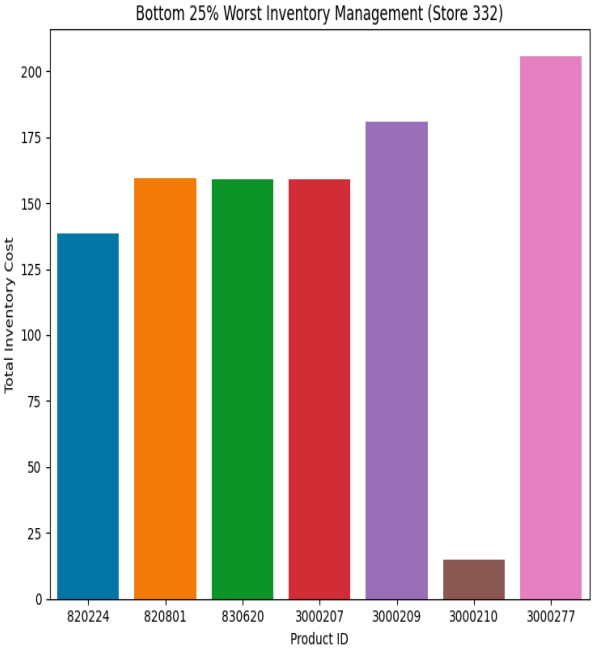

Top 25% Best Inventory Management (Store 18)
Total Inventory Cost
0
200
400
600
800
1000
1200
6228
12216
820221
3000024
3000096
3000162
3000214
Product ID


Bottom 25% Worst Inventory Management (Store 18)
Total Inventory Cost
0
10
20
30
40
50
60
70
820801
830620
841402
3000209
3000210
3000212
3000297
Product ID


Top 25% Best Inventory Management (Store 117)
Total Inventory Cost
0
200
400
600
800
1000
1200
1400
12216
820221
3000024
3000096
3000162
3000214
3000293
Product ID


Bottom 25% Worst Inventory Management (Store 117)
Total Inventory Cost
0
20
40
60
80
820224
820801
3000207
3000209
3000210
3000211
3000297
Product ID


Top 25% Best Inventory Management (Store 332)
Total Inventory Cost
0
500
1000
1500
2000
2500
3000
12216
820221
3000096
3000159
3000162
3000214
3000293
Product ID

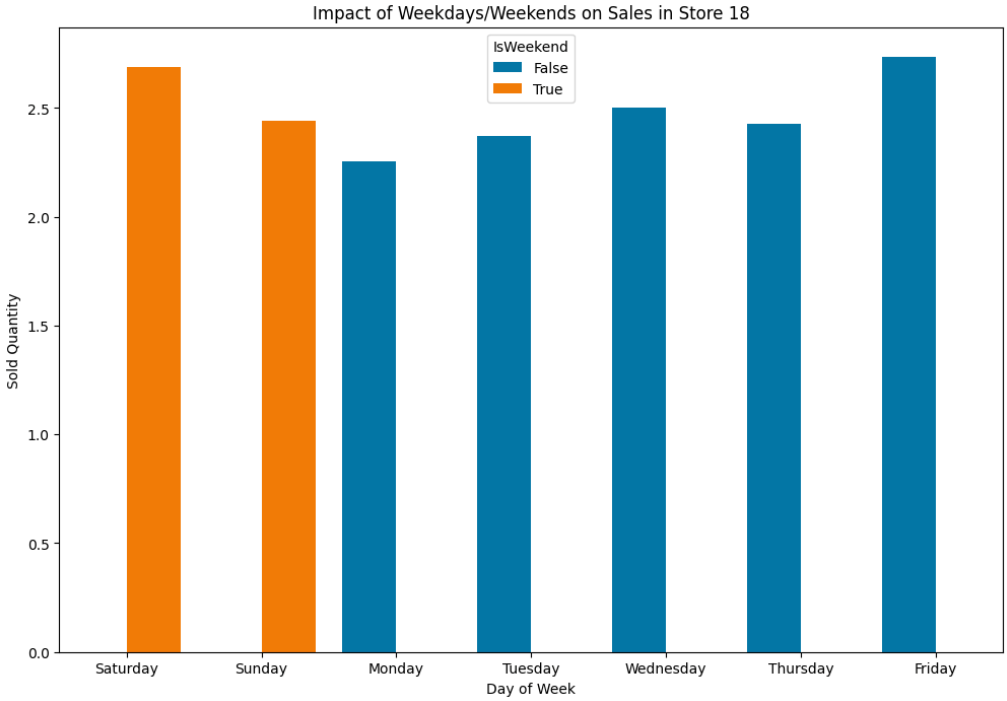

Bottom 25% Worst Inventory Management (Store 332)
Total Inventory Cost
0
25
50
75
100
125
150
175
200
820224
820801
830620
3000207
3000209
3000210
3000277
Product ID

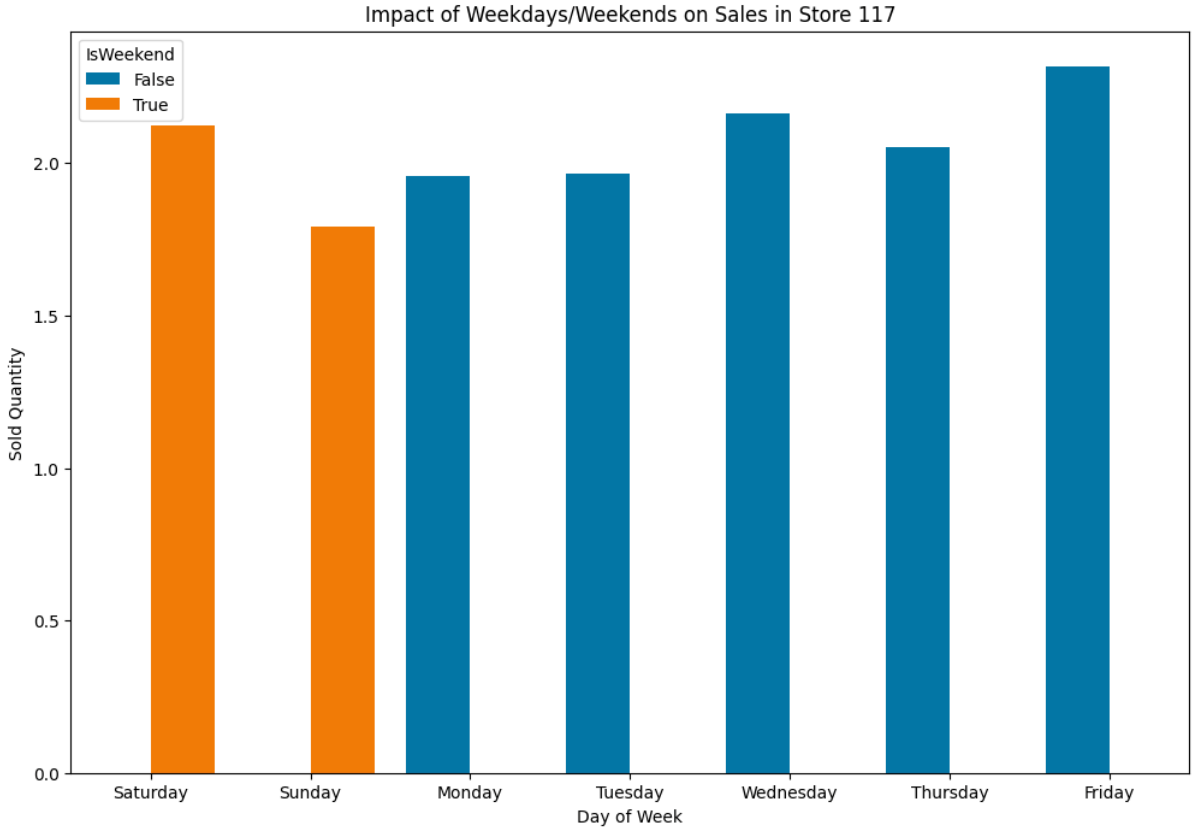
6.1.4 Stock Outs and Assessed Loss of Deals

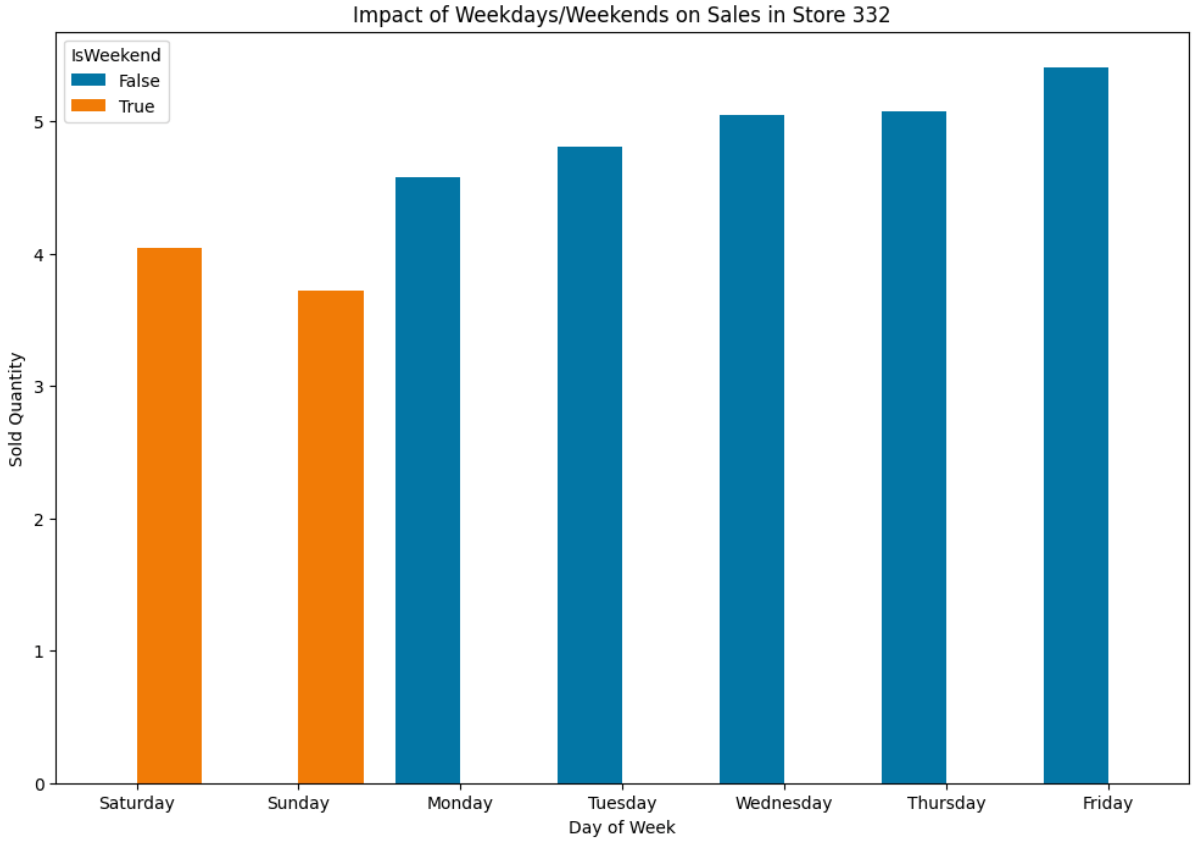
Utilizing a moderate system, we extended our examination to incorporate stock-outs by assessing a 75% misfortune in normal deals from the previous a month on the pertinent work day. This strategy permitted us to work out the possible monetary effect of stock deficiencies for every item. By distinguishing inaccessible things and utilizing an exact computation to gauge the subsequent misfortune in deals, we acquired significant experiences into the monetary ramifications of stock control. From the dataset, we can infer that it records instances where products, identified by their PLU codes, reached an end quantity of zero on a specific business date, indicating they were out of stock at the store's closing. The 'StockedOut' column further clarifies whether these zero-quantity statuses were due to products being sold out (0) or actually being unavailable for sale (1), with the latter likely representing lost sales opportunities. This information is crucial for assessing the effectiveness of inventory management, as frequent stockouts for certain PLUs could signal supply chain deficiencies, misaligned ordering practices, or demand forecasting errors. Moreover, unpurchased products, especially those marked as stocked out, could point to potential improvements in inventory turnover and the need for more responsive restocking strategies to meet customer demand and reduce the incidence of zero-end-quantity situations.

6.1.5 Effect of Day of the Week, Month to month Changes, and Climate

The bar charts for Stores 18, 117, and 332 illustrate the impact of weekdays and weekends on sales, with the weekend days typically marked as 'True' and the weekdays as 'False'. In Stores 18 and 117, weekend sales are higher than weekday sales, suggesting a trend of increased customer activity on Saturdays and Sundays. Conversely, Store 332 presents a peak in sales on Friday, indicating a variation in consumer behavior or store promotions that drive up end-of-week sales. These insights could inform targeted strategies for inventory stocking and staffing, with Stores 18 and 117 potentially increasing resources to meet weekend demand, while Store 332 may focus on maximizing the Friday sales opportunity.




Impact of Weekdays/Weekends on Sales in Store 117
IsWeekend
False
True
Sold Quantity
0.0
0.5
1.0
1.5
2.0
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Day of Week

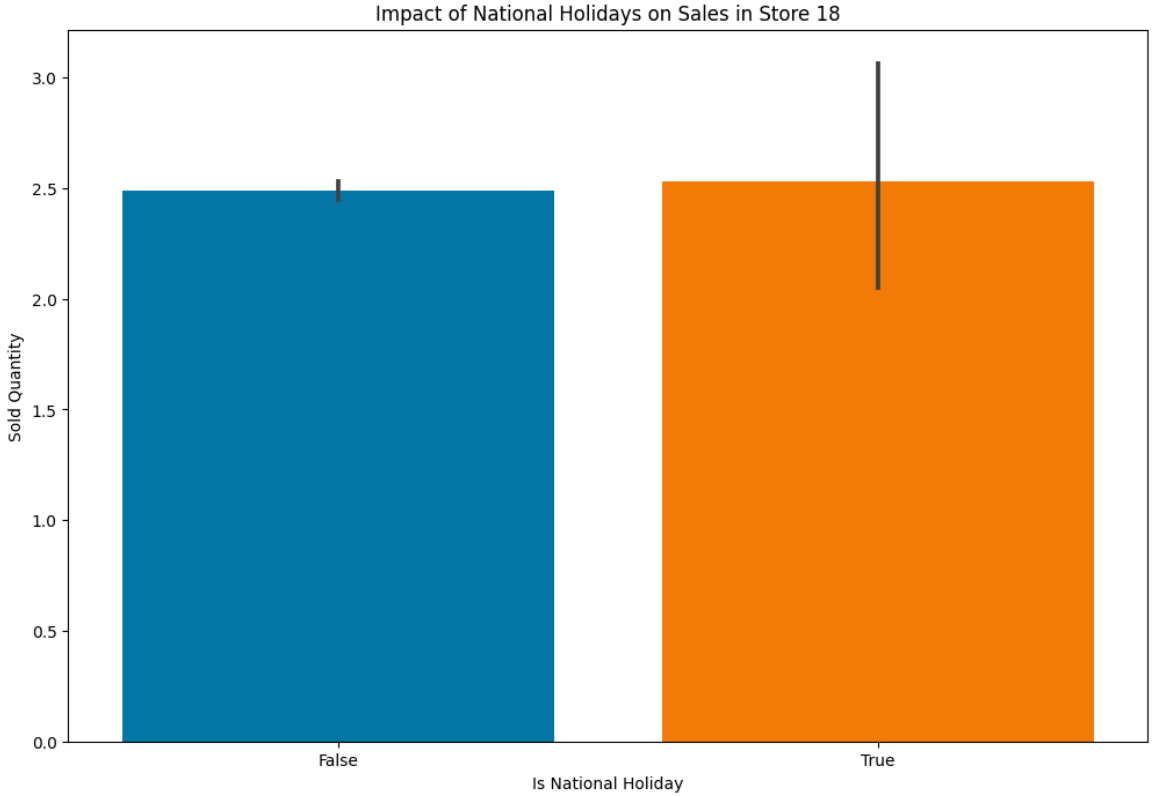

Impact of Weekdays/Weekends on Sales in Store 332
IsWeekend
False
True
Sold Quantity
0
1
2
3
4
5
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Day of Week

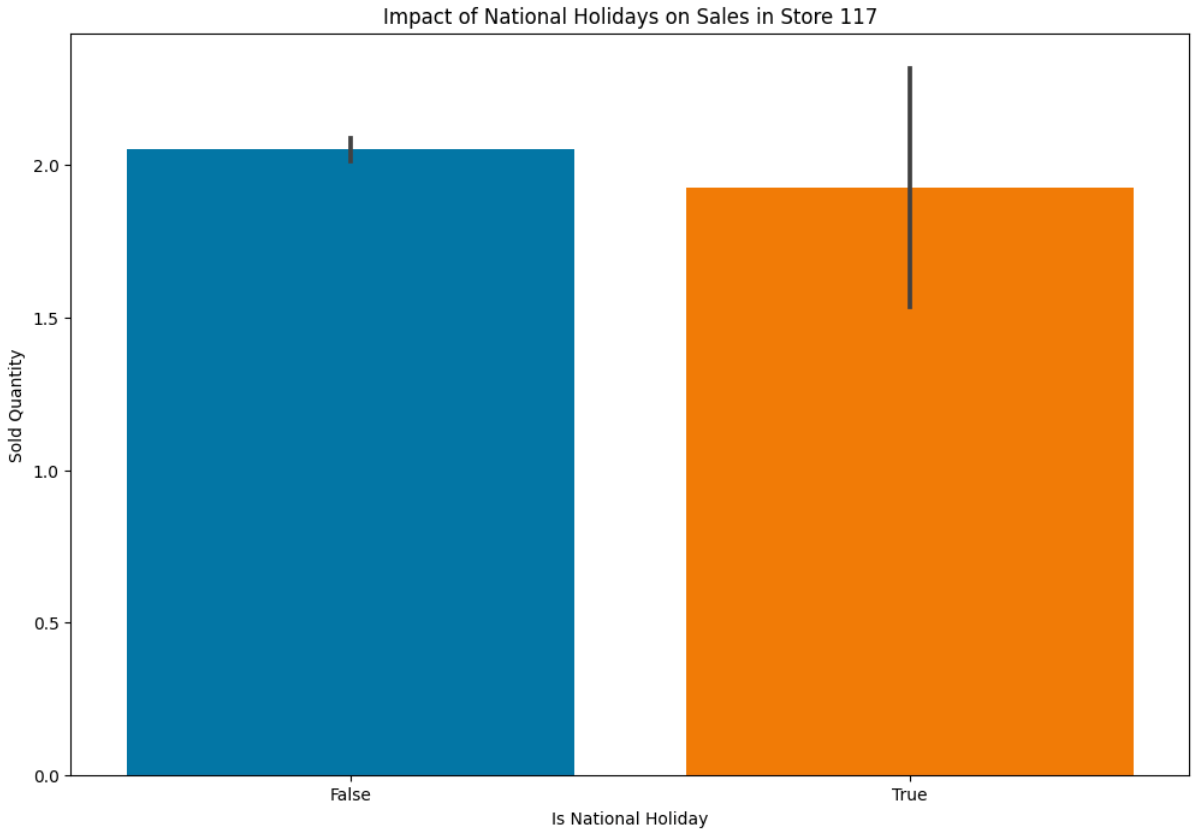
6.1.6 Drive-thru Element Investigation

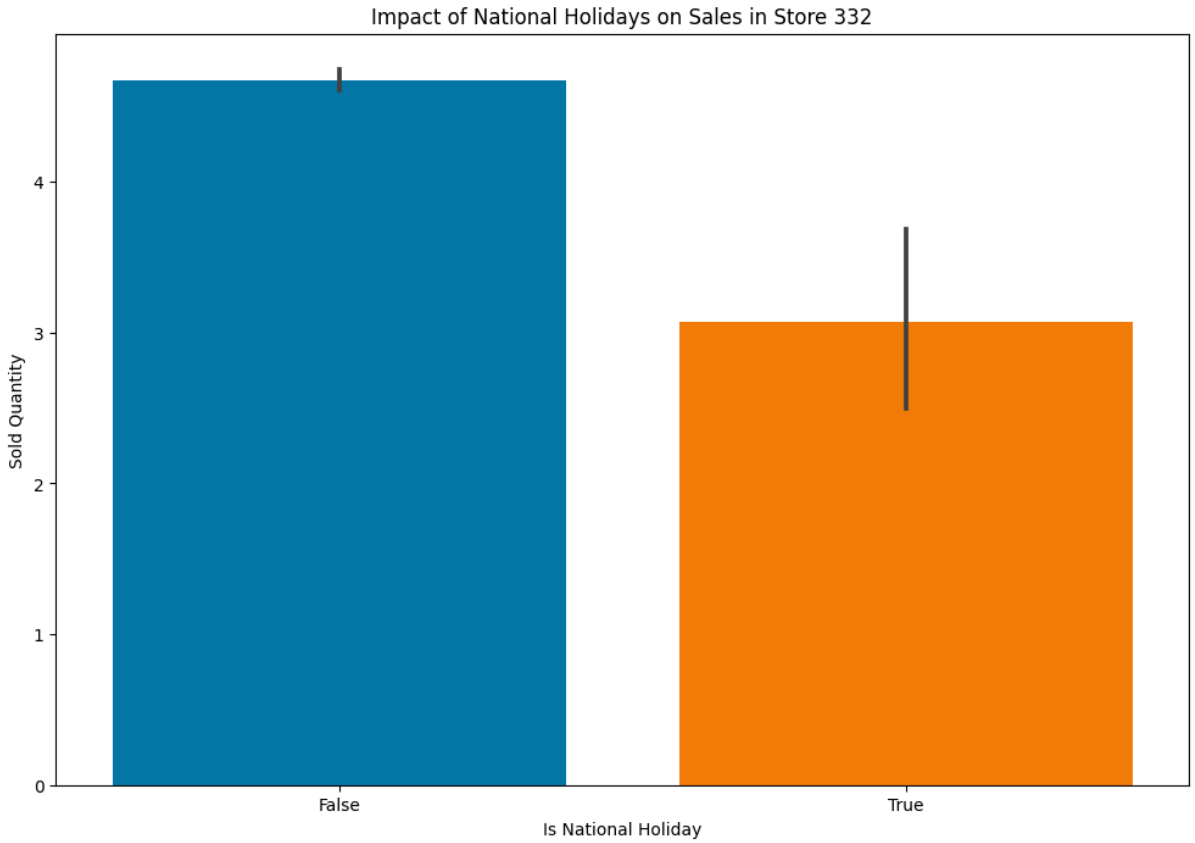
The sales data comparison for Stores 18, 117, and 332 reveals a clear trend where drive-thru facilities substantially outsell their non-drive-thru counterparts, with Store 18 showing significant sales differences for items like Plain Bagels and Chocolate Croissants, and similar trends in Store 117 with high sales differences for Blueberry Streusel Muffins and Butter Croissants. Store 332 exhibits an even more pronounced disparity, especially notable in the sales of Blueberry Streusel Muffins and Jalapeno Cheese Bagels. These differences underscore the importance of the drive-thru service in enhancing sales volumes and indicate a customer preference for the convenience it offers, information that could be pivotal for strategic decision-making in operational management and service offerings.

6.1.7 Effect of Workday/Ends of the week and Public Occasions

The bar charts represent the impact of national holidays on sales in Stores 18, 117, and 332. The 'False' bars indicate sales on non-holiday days, and the 'True' bars represent sales on national holidays. In all three stores, the quantity of items sold on national holidays is less than or almost equal to the quantity sold on non-holiday days, as seen by the height of the 'True' bars compared to the 'False' ones. However, there's a notable variance in sales as indicated by the error bars, suggesting fluctuations in sales volumes on these days. This could be due to differing customer behaviors during holidays, such as increased travel or family gatherings, which could affect shopping patterns. The data may inform the stores to adjust their inventory and staffing accordingly during national holidays, perhaps reducing stock and labor to match the anticipated decrease in sales.




Impact of National Holidays on Sales in Store 117
Sold Quantity
0.0
0.5
1.0
1.5
2.0
False
True
Is National Holiday


Impact of National Holidays on Sales in Store 332
Sold Quantity
0
1
2
3
4
False
True
Is National Holiday

6.1.8 Stocking Pattern Analysis

In the analysis of stocking patterns across multiple stores, a comprehensive investigation was carried out to ascertain the frequency of restocking and the introduction of new products. The fleeting elements of stock administration all through the organization of stores were analysed in this review. The significant inquiry of whether retailers follow a less regular restocking plan or get new product day to day was tended to. Through close assessment of these stocking designs, significant data about the retail outlets' functional systems was found. Fathoming the restocking cadence is fundamental for augmenting stock levels, fulfilling client demands, and lessening waste. The outcomes support key navigation by giving ideas to further developing proficiency and responsiveness to showcase requests in loading rehearses all through the organization of stores.

6.2 General Store Data Analysis:

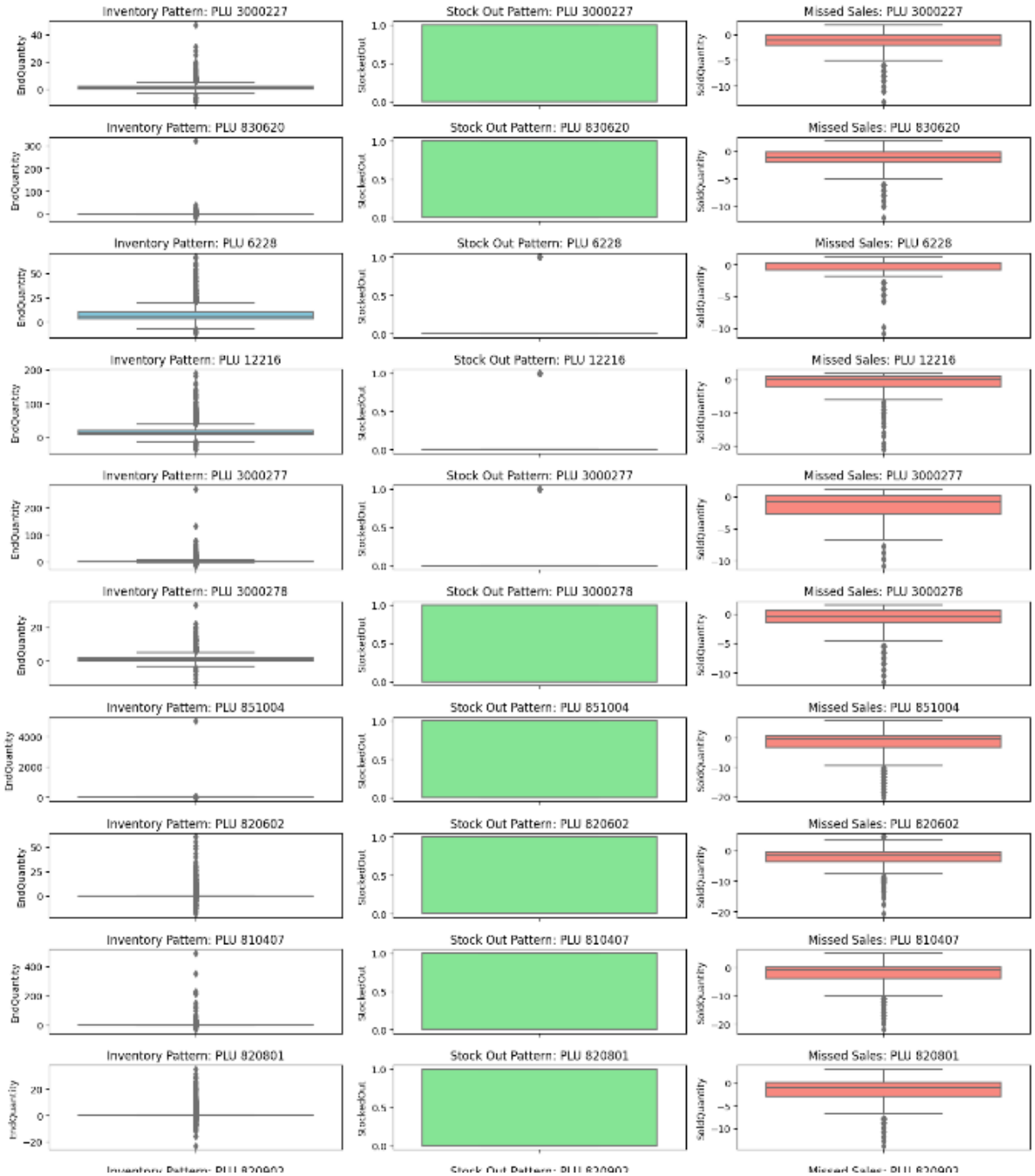
6.2.1 Box Plots and Measurements for 27 Items.

From the 'Inventory Pattern' plots, we can observe the fluctuation and distribution of inventory levels for these items over a certain period. The scatter with a center line possibly indicates the median inventory level, and the points represent daily inventory levels.

The 'Stock Out Pattern' plots show the frequency of stockouts for each item as bar charts. A taller bar suggests more frequent stockouts, which indicates that the item was not available for purchase on more occasions.

The 'Missed Sales' plots, which are also scattered plots, seem to represent the lost revenue opportunities due to items being out of stock. Points above the center line may indicate days with higher-than-average missed sales, which could be due to the unavailability of the product when there was demand for it.

Comparing these three aspects for each PLU code can provide insights into how inventory management affects sales. For example, a high number of stockouts and high missed sales could suggest a need for better inventory planning or a reassessment of supplier reliability. The aim would be to minimize stockouts, especially for items with a clear pattern of missed sales, which directly impact revenue. This data can be critical for adjusting inventory levels, improving demand forecasting, and ultimately enhancing customer satisfaction by ensuring popular products are in stock.

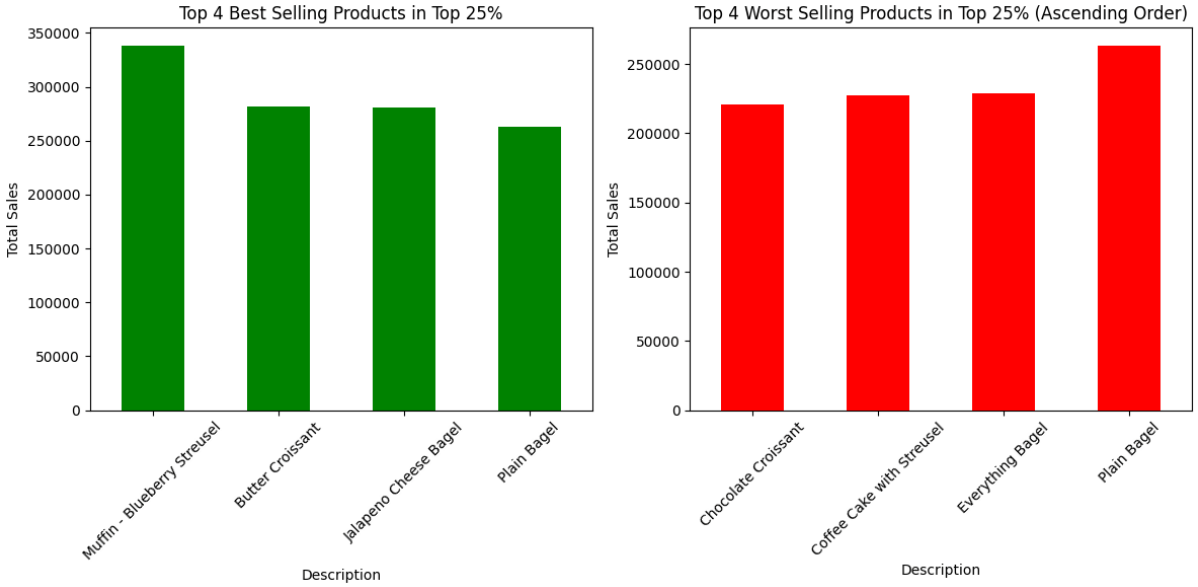


6.2.2 Best and Most terrible Dealer Items

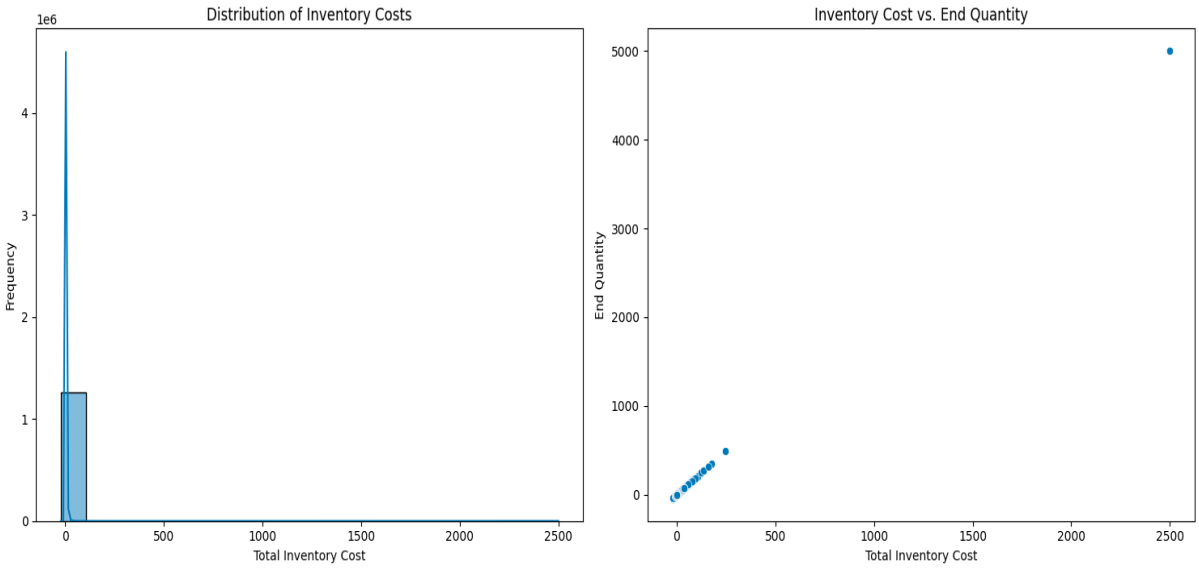
The plots present a comparison of the top 4 best and worst selling products within the top 25% sales bracket, likely from a particular store or a chain. The first green bar chart indicates that the best selling products are, in descending order: 'Muffin, Blueberry Streusel', 'Butter Croissant', 'Jalapeno Cheese Bagel', and 'Plain Bagel'. The sales figures for these items are significantly high, suggesting they are popular among customers.

The second red bar chart shows the worst selling products, listed in ascending order of sales: 'Chocolate Croissant', 'Coffee Cake with Streusel', 'Everything Bagel', and 'Plain Bagel'. Interestingly, 'Plain Bagel' appears in both lists, which could imply that while it's one of the top sellers, it also has a high variance in sales performance or there could be a segmentation within the category that places different variants of plain bagels at both ends of the sales spectrum.

From this, we can infer that there is a varied customer preference within the product range, and while baked goods like muffins and croissants are highly favored, there is also a significant variation in the sales performance of bagels. This information could be utilized for inventory management, marketing strategies, and potentially guiding product development based on consumer preferences.



6.2.3 Stock Administration Investigation



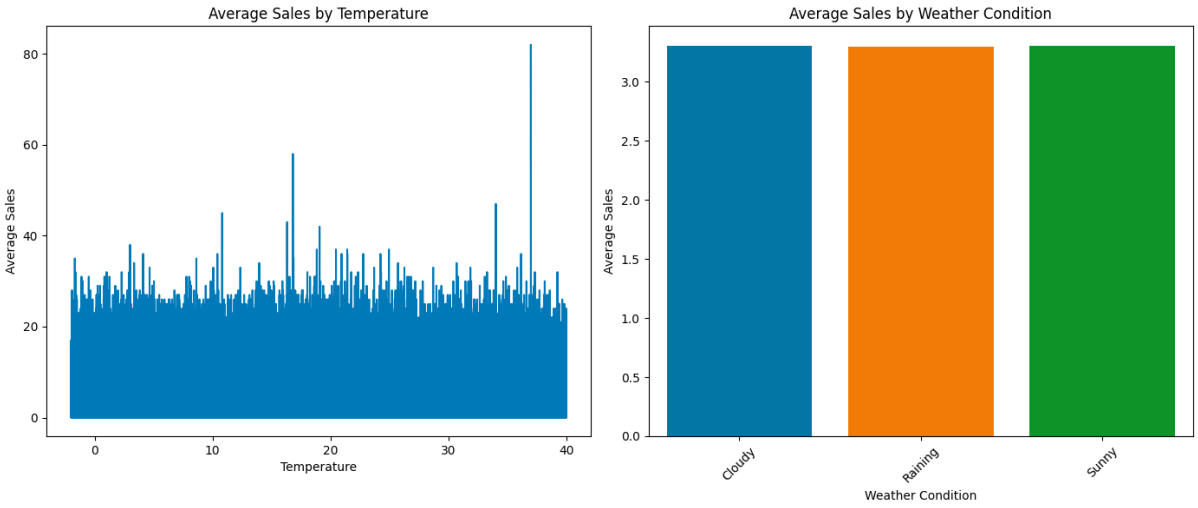
The plots provided appear to represent inventory management metrics, with the first being a histogram titled "Distribution of Inventory Costs" and the second a scatter plot comparing "Inventory Cost vs. End Quantity."

From the histogram, we can infer that the distribution of inventory costs is heavily skewed towards the lower end, with most inventory items incurring low costs and only a few items resulting in higher costs. This suggests that the majority of the inventory items may be low-cost items, or that the business keeps smaller stock levels of high-cost items.The scatter plot shows a positive relationship between total inventory cost and end quantity, indicating that as the inventory cost increases, so does the quantity of items on hand at the end of the period. However, there are only a few data points with high inventory costs and corresponding high end quantities, reinforcing the idea suggested by the histogram that high-cost inventory items are less common.

6.2.4 Stock Outs and Assessed Loss of Deals

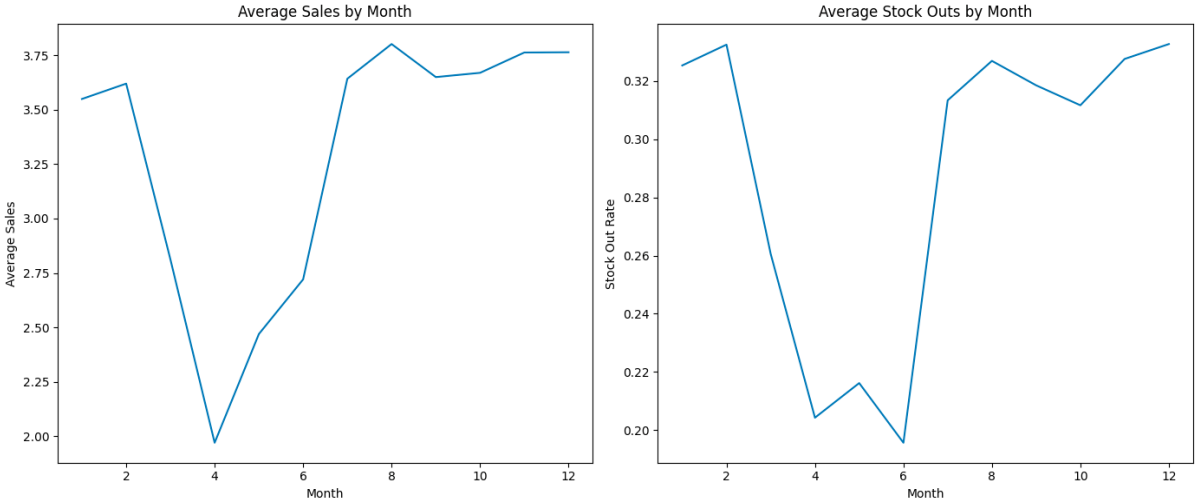
The table lists product codes (PLUs) with their corresponding annual estimated financial losses, indicating significant variability in the impact of different products on overall loss figures. Some items, like those with PLU codes 810703, 820224, and 3000211, incur exceptionally high losses, possibly due to factors such as high value, susceptibility to shrinkage, or frequent stockouts leading to missed sales opportunities. This suggests a need for targeted strategies to reduce losses, which may include enhancing inventory control, improving security protocols, or revising sales and storage practices. Conversely, products with minimal losses, such as PLU 3000293, could be models of effective loss prevention or simply less costly, indicating efficient management or lower impact on profitability. The data underscores the importance of strategic inventory and loss prevention measures tailored to specific product risks and behaviors.

6.2.5 Effect of Day of the Week, Month to month Changes, and Climate



In the first two charts, "Average Sales by Month" and "Average Stock Outs by Month," we see a trend where sales fluctuate over the course of the year, with a notable dip around the middle of the year. The stock-out rate, interestingly, has its lowest point around the same

time, suggesting a possible correlation where lower sales could lead to fewer stock-outs due to a slower inventory turnover.



The third chart, "Average Sales by Temperature," shows a sporadic relationship between temperature and sales, with several peaks indicating that there might be temperatures at which sales significantly increase. This could suggest that certain products sell better at specific temperature ranges, or that customer foot traffic increases during certain weather conditions.

The last chart, "Average Sales by Weather Condition," indicates that sales may vary with different weather conditions, with the highest average sales occurring on sunny days, followed closely by rainy days, and the lowest on cloudy days. This might imply that weather conditions can influence consumer behavior, potentially encouraging or deterring store visits and purchases.

6.2.6 Drive-thru Element Investigation

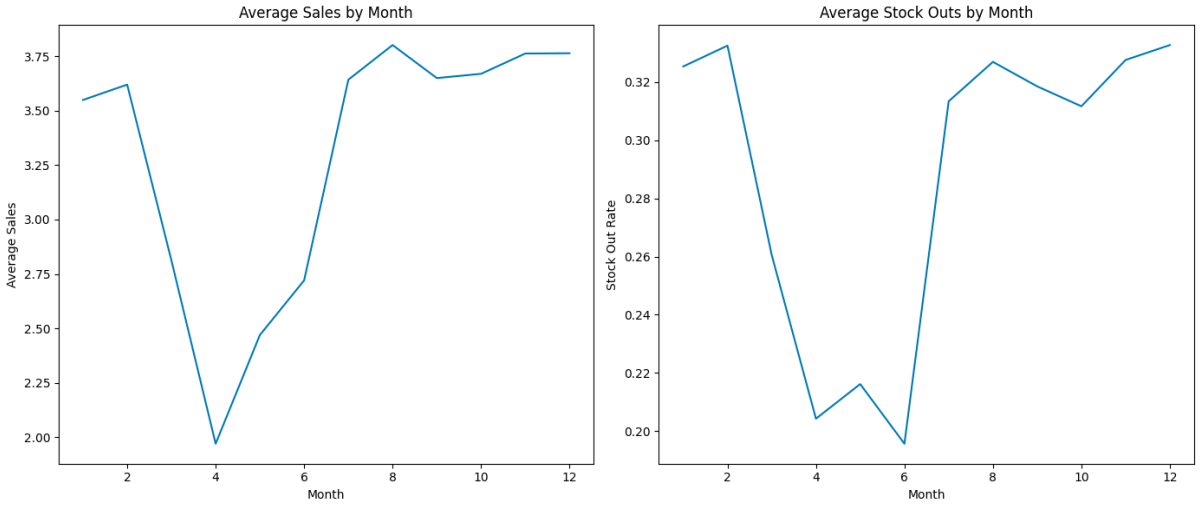
We can observe from the data that the presence of a drive-thru significantly increases the sales quantities of various bakery items. For example, the 'Muffin - Blueberry Streusel' sees a massive boost in sales through the drive-thru, selling 333,008 units compared to only 4,791 units without it. This trend is consistent across other items such as 'Jalapeno Cheese Bagel', 'Butter Croissant', 'Plain Bagel', and 'Coffee Cake with Streusel', where the drive-thru option outsells the non-drive-thru by a large margin. The 'SalesDifference' column quantifies this disparity, with differences ranging from approximately 215,000 to over 328,000 units. These observations suggest that customers greatly value the convenience of a drive-thru, which could be a strategic focus for the business to enhance sales volume.

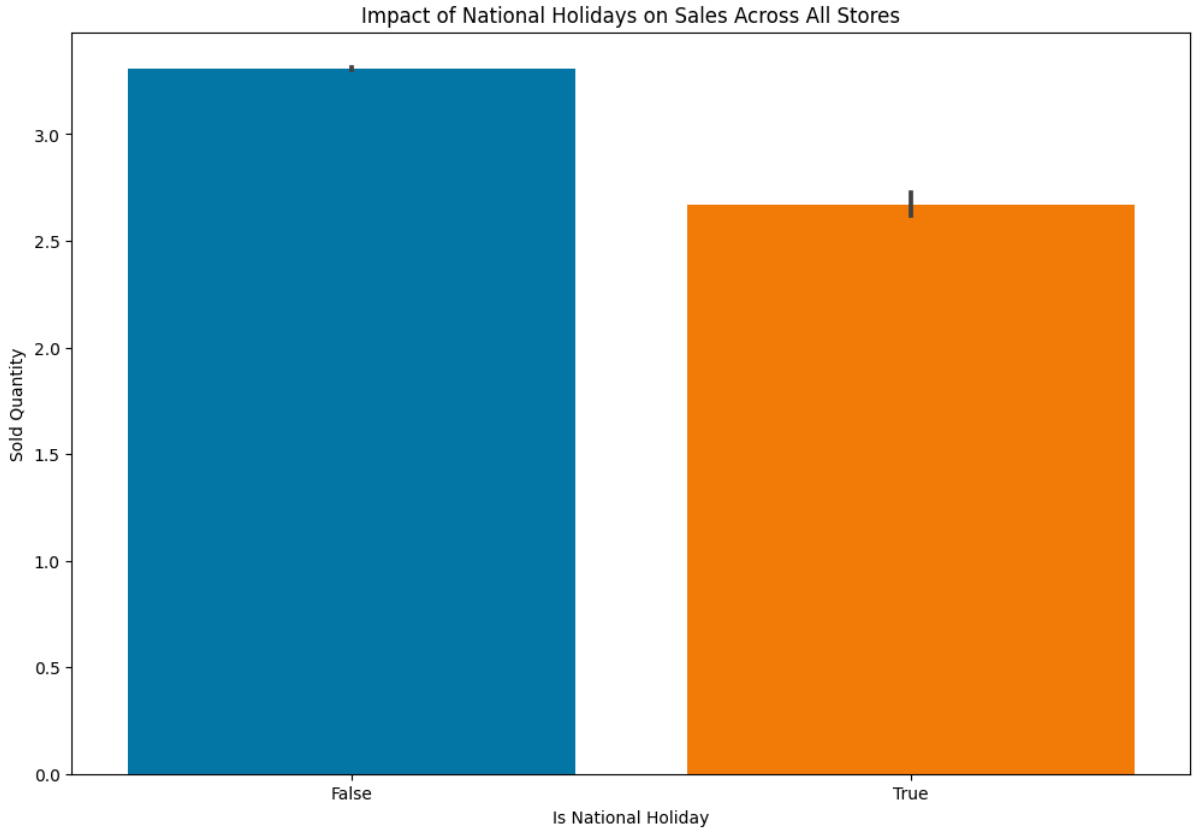
6.2.7 Effect of Workday/Ends of the week and Public Occasions

We can infer several points about sales trends in relation to national holidays and weekends:

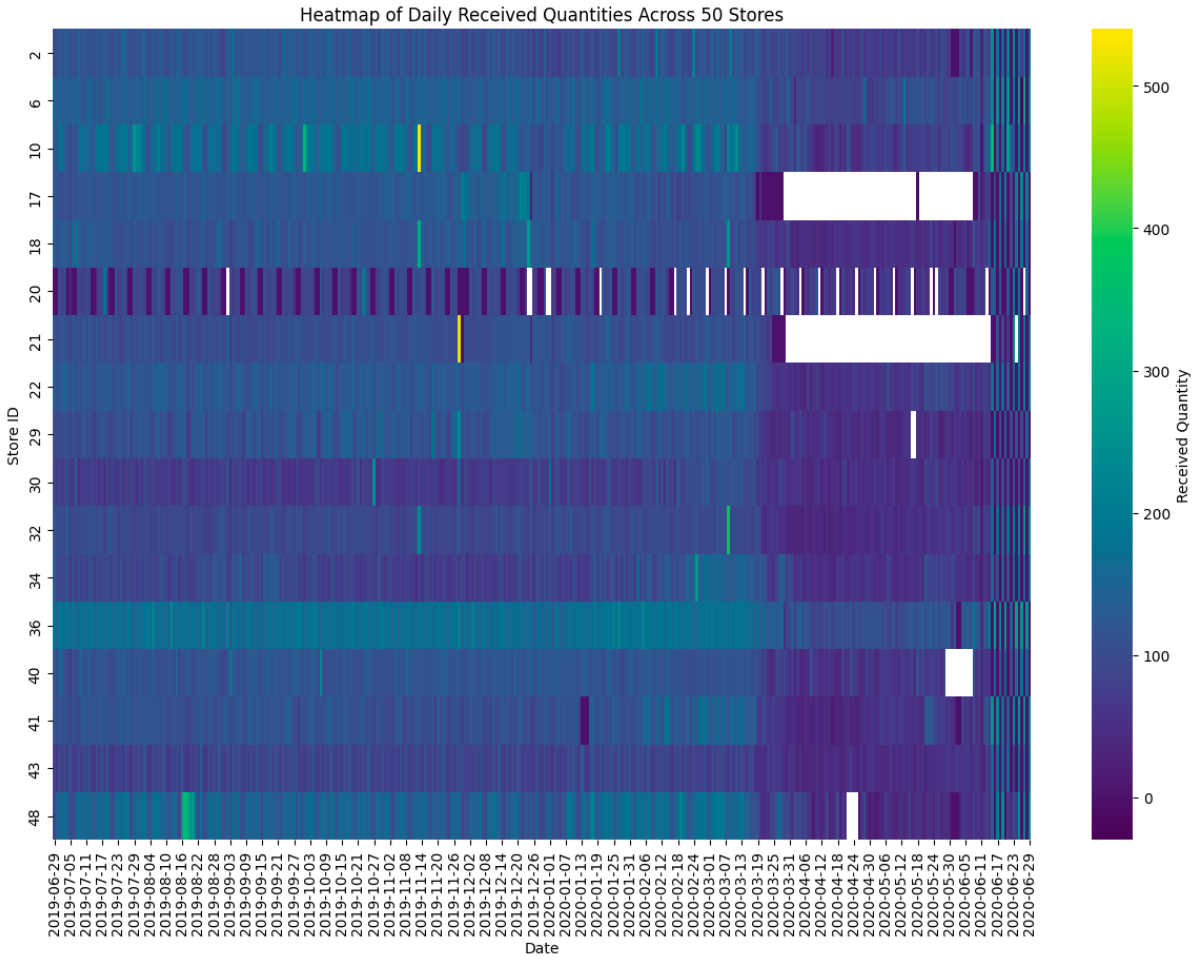
- **Impact of National Holidays:** The first plot shows that the average quantity of items sold is slightly lower on national holidays ('True') compared to non-holidays ('False'). This could suggest that national holidays have a minor impact on reducing sales volume across all stores.

- **Weekend Sales:** The second plot indicates that sales on weekends (Saturday and Sunday, marked as 'True' are higher than sales on weekdays. This trend is consistent across all days, with the lowest sales typically occurring on Monday and the highest on Friday and Saturday.





6.2.8 Stocking Pattern Analysis



The heatmap visualizes daily received quantities across 50 stores and reveals several patterns: variability in inventory received, with certain days showing higher volumes across many stores, possibly due to restocking or bulk shipments; outliers where specific stores received unusually large quantities, highlighted by bright yellow markers; and gaps in data, denoted by white spaces, which might indicate store closures, lack of deliveries, or missing records. Additionally, some stores show more fluctuation than others, suggesting differences in demand, supply chain management, or stocking strategies. Such a heatmap is valuable for supply chain analysis, helping to identify trends, irregularities, and opportunities for optimizing inventory distribution across a retail network.

7. Section 2 - Prediction

In our analysis for identifying the optimal features and removing the potential correlated features, we focused on data from three specific store IDs (storeID=18, storeID=117, storeID=332). This data underwent a process of feature engineering, where categorical variables were converted into numerical formats for effective quantitative analysis. A critical part of our analysis was the development of a correlation heatmap. This heat map provided a clear visual representation of the interrelationships between various variables.

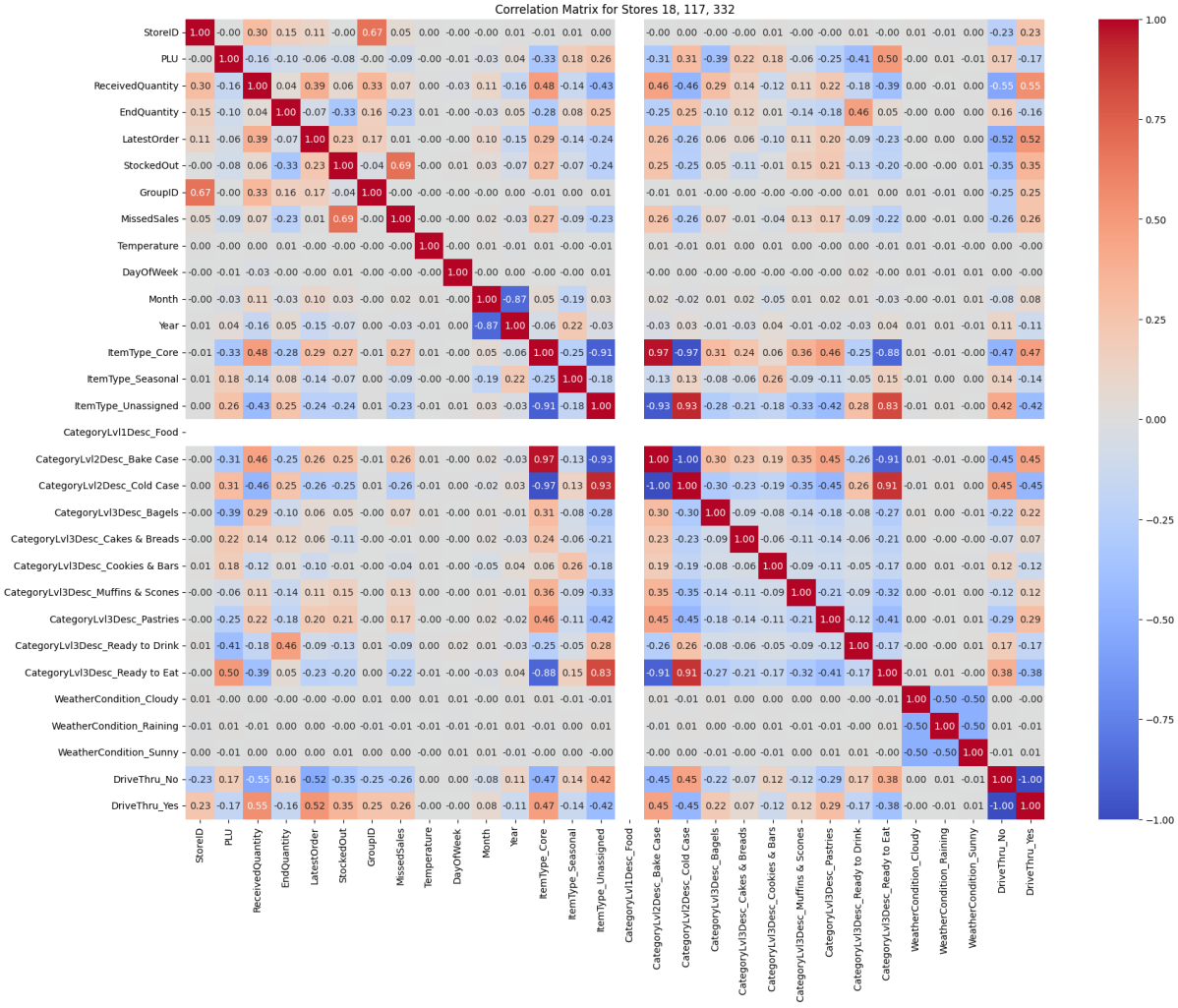


Fig : Correlation Matrix for feature optimization

Through this, we identified significant correlations, such as the one between 'ItemType_Unassigned' and 'ItemType_Core', leading us to drop the former due to redundancy. Similar correlations were observed and addressed, including those involving 'CategoryLvl2Desc_Bake Case', 'CategoryLvl2Desc_Cold Case', 'CategoryLvl3Desc_Ready to Eat', and 'DriveThru_Yes'. These findings guided us to eliminate features that were highly correlated, streamlining our feature set to enhance the predictive model's efficiency and accuracy.

7.1 Linear regression

We applied a linear regression model to predict 'SoldQuantity' using 'ReceivedQuantity' and 'EndQuantity' as features, focusing specifically on store ID 18. After ensuring the selected features were present in our dataset, we split the data into training and testing sets. The model was trained and then used to predict sales on the test set. Our evaluation using Mean Squared Error (MSE) yielded a value of 2.758, indicating the model's predictive performance.

Additionally, we visualized the actual versus predicted sales, providing a clear representation of the model's effectiveness in capturing the trend in the data.

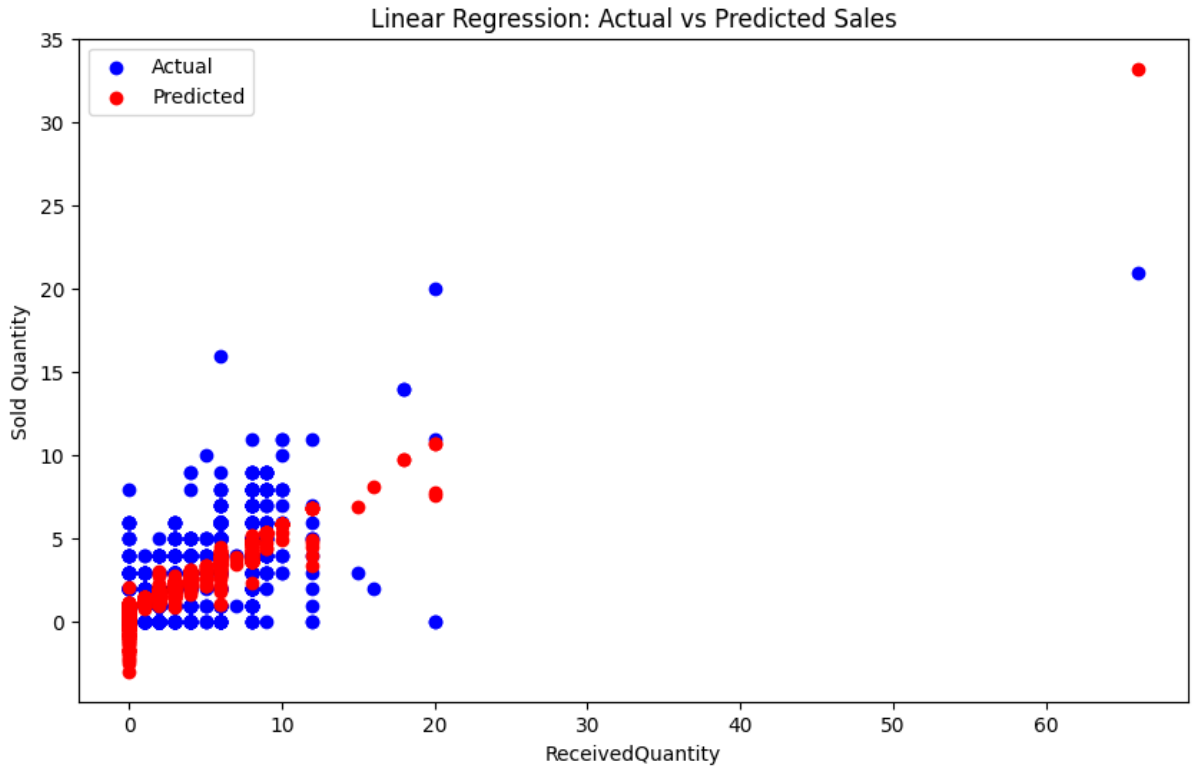


Fig : Linear regression : Actual vs Predicted sales

7.2 Ensemble Models and comparing their metrics

We explored ensemble machine learning models for predicting 'SoldQuantity' across three selected stores. We implemented and compared Random Forest, Gradient Boosting Machine, and XGBoost models using our feature-engineered dataset. These models were evaluated using Root Mean Squared Error (RMSE) and R-squared metrics.

The Random Forest model achieved an RMSE of 0.978 and an R-squared of 0.883, indicating strong predictive performance.

The Gradient Boosting Machine recorded an RMSE of 0.984 with an R-squared of 0.881.

The XGBoost model demonstrated an RMSE of 0.980 and an R-squared of 0.882.

The hyper-parameters used for the prediction:

n_estimators = 100

random_state = 42

```
Random Forest (Imputed) - RMSE: 0.9784439480941858, R-squared: 0.8825665282446924
Gradient Boosting Machine (Imputed) - RMSE: 0.9842871650175089, R-squared: 0.8811597267732032
XGBoost (Imputed) - RMSE: 0.9804663973560708, R-squared: 0.8820805551871367
```

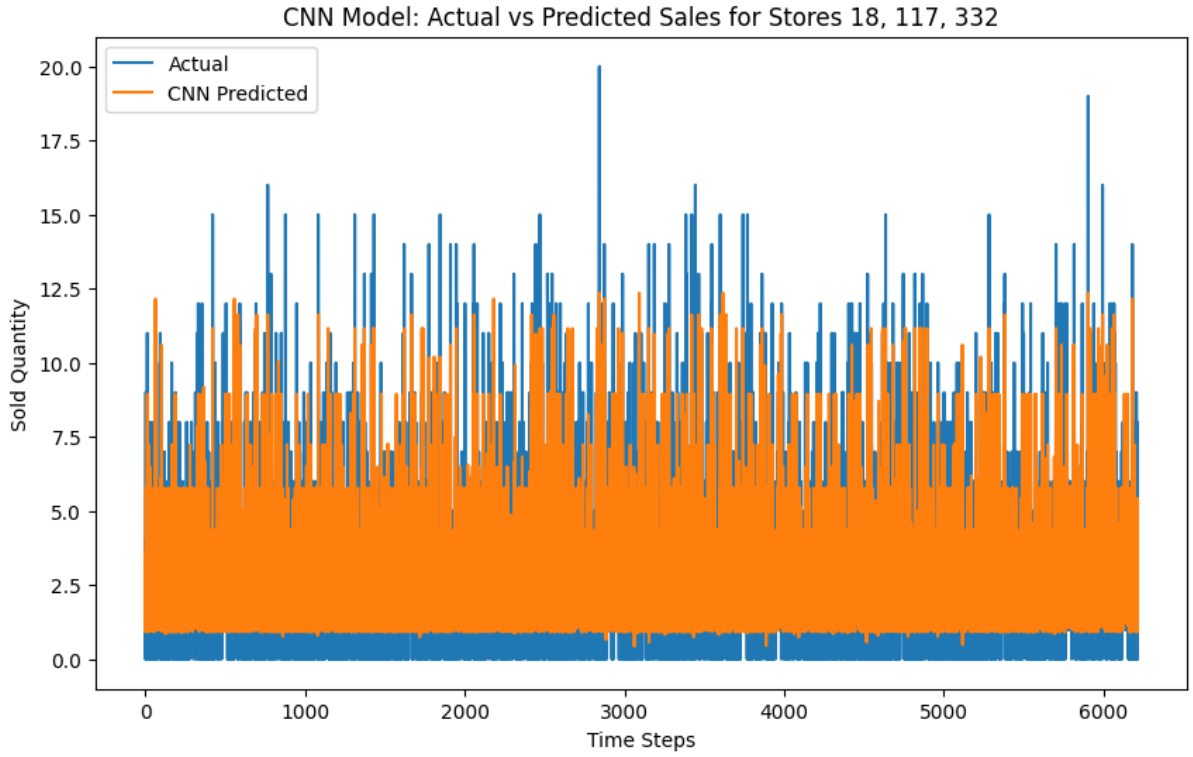
7.2.1. Fine Tuning

After extensive tuning of three distinct machine learning models - Random Forest, Gradient Boosting Machine (GBM), and XGBoost - the results obtained are remarkably close in terms of predictive performance. The Random Forest model, with an RMSE (Root Mean Square Error) of 0.9784 and an R-squared value of 0.8826, shows a slight edge in terms of accuracy and the proportion of variance explained. However, the differences are marginal when compared to the GBM and XGBoost models, which delivered RMSEs of 0.9843 and 0.9805, and R-squared values of 0.8812 and 0.8821, respectively. These results, achieved through rigorous tuning, suggest that all three models are highly competent in capturing the underlying patterns in the dataset. The choice between them may now pivot more on aspects other than raw performance, such as model interpretability, computational efficiency, or specific application requirements. This close competition highlights the importance of considering the broader context and practical implications of model deployment, beyond the numerical precision of performance metrics.

7.3 Deep Learning Models

7.3.1 LSTM Model

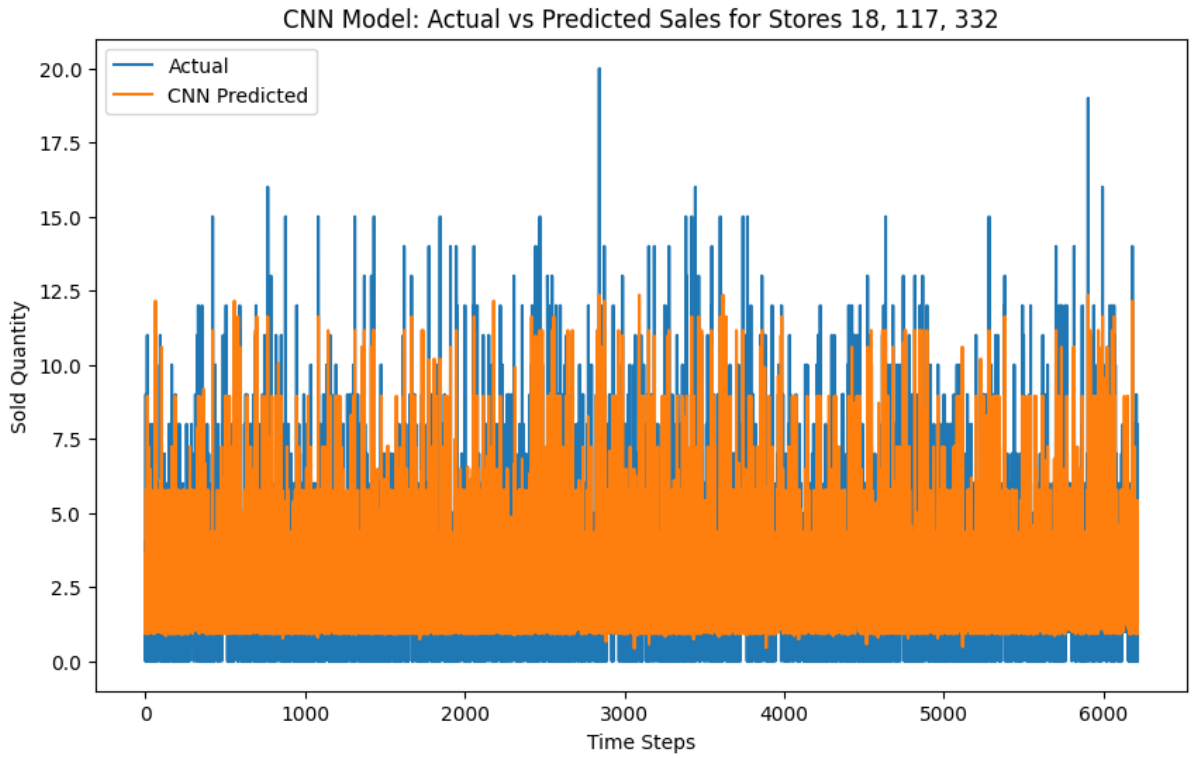
We employed a deep learning approach using an LSTM model to analyze time-series sales data from selected stores. The model was trained over 50 epochs, showing consistent improvement in loss values. The final training loss was 0.0076, and the validation loss was 0.0070, indicating the model's ability to generalize well.



This approach, diverging from traditional ensemble methods, provided a nuanced understanding of sales trends over time. The LSTM's ability to process sequential data makes it particularly suited for this time-series analysis, as demonstrated by its performance metrics.

7.3.2 CNN Model

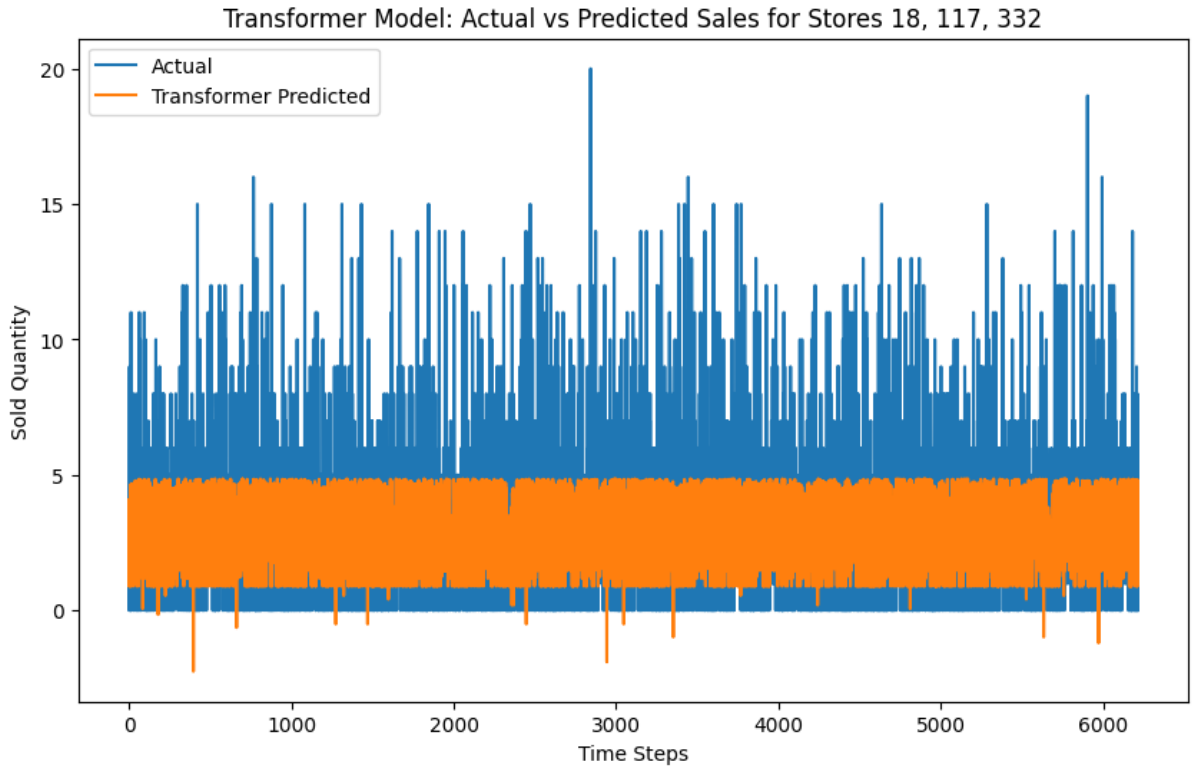
We developed a CNN model to analyze sales data from selected coffee stores. The model was trained over 50 epochs, demonstrating consistent improvements in loss values. The training and validation loss both stabilized around 0.0068 and 0.0063 respectively, indicating the model's effective learning and generalization capabilities.



This approach allowed for a nuanced understanding of sales dynamics, as seen in the plotted results comparing actual and predicted sales, underscoring CNN's efficiency in handling time-series data for these specific stores.

7.3.3 Transformers Model

We implemented a deep learning model using Transformer architecture which was implemented using **Attention Mechanism** to analyze sales data from specific coffee stores. The model was trained on selected features and tuned through a series of epochs, showcasing significant improvement in loss metrics. The best model parameters were identified, with a learning rate of 0.001 and dropout of 0.1, achieving an optimal RMSE of 2.5536.



The final training phase over 50 epochs demonstrated a stable decrease in both training and validation loss, underlining the model's robustness in capturing the sales trends. The comparison between actual and predicted sales in the plotted results further validated the model's predictive capabilities.

7.4 Predicting sales.

7.4.1 Models predicting for 1, 3, 10 days ahead.

We utilized an LSTM model to forecast sales for specific coffee stores, predicting sales 1 day, 3 days, and 10 days ahead. The model was trained with a focus on minimizing loss, achieving a final validation loss of 0.0068. Predictions for 1 day ahead resulted in 1.28, 62.28 for 3 days ahead, and 77.36 for 10 days ahead. These predictions are vital for planning logistics, such as shipping from the corporate warehouse or ordering from the distributor, aiding in efficient inventory and supply chain management.

```
1 day ahead prediction: 1.2844398021697998
3 days ahead prediction: 62.28110122680664
10 days ahead prediction: 77.36479187011719
```

Conclusion:
MSE for 1 day and 3 day time sequence prediction is better in CNN model: This means that when using a Convolutional Neural Network (CNN) model for predicting time sequences, the Mean Squared Error (MSE) is lower (indicating better performance) for short-term predictions, specifically for 1 day and 3 day intervals. MSE is a common measure of prediction accuracy, where lower values indicate more accurate predictions. CNNs, known

for their strength in pattern recognition within spatial data (like images), are being effectively utilized here for short-term temporal data.

MSE for 10 day time sequence prediction is better with LSTM: This implies that for longer-term predictions, specifically for a 10 day interval, a Long Short-Term Memory (LSTM) model yields a lower MSE, signifying better performance. LSTM is a type of Recurrent Neural Network (RNN) particularly adept at handling long-term dependencies in sequential data. In this context, LSTMs are more effective than CNNs for capturing and predicting patterns over longer time sequences.

7.4.2 Forecast from today to 1 day, 3 days, 10 days

Two models were developed: a Convolutional Neural Network (CNN) and a Transformer model. Both models were trained on data from selected coffee stores to predict future sales. The CNN, with its kernel size of 1, and the Transformer, constructed with multiple heads and layers, aimed to forecast sales 1, 3, and 10 days into the future. Predictions were made by feeding the last known data batch into the models and iteratively predicting future values, adjusting the input batch each time. The predictions were then scaled back to their original range for interpretability. These models demonstrate the application of advanced deep learning techniques in time-series forecasting for retail sales.

7.5 Comparing Deep Models

We evaluated three models (LSTM, CNN, Transformer) using an 80/20 train-test split. For the LSTM model, the training MSE was 3.64 and the test MSE was 3.37. The CNN model showed a training accuracy of 0.0070 and a testing accuracy of 0.0065. The Transformer model achieved a training RMSE of 2.65 and a test RMSE of 2.57. These results demonstrate the effectiveness of each model in generalizing from training to unseen test data, with the Transformer model showing the most promising performance in terms of RMSE values.

```
Accuracy for Training Data:  3.6351705351445447
Accuracy for Test Data:  3.3655858089119057
```

fig. LSTM model

```
Training Accuracy: 0.006960081867873669
Testing Accuracy: 0.006452973932027817
```

fig. CNN Model

```
Training RMSE: 2.65005109744331
Test RMSE: 2.570169491038072
```

fig. Transformers Model

7.6 Elimination of common products

The focus is on analyzing product data for individual stores. Specifically, teams 1-10 examine the first 13 products, while teams 11-25 concentrate on the next 14 products. The code provided filters the data for three specific stores and then isolates products that are common across all these stores. For both groups, products that are not shared among the three stores are excluded. This approach ensures that the analysis is consistent and relevant for all selected stores, focusing only on products that are commonly stocked.

7.7 Best models for all stores

We applied the XGBoost model, chosen for its strong performance in previous tasks, to predict sales in 10 new coffee stores. These stores were selected from a dataset excluding the initially analyzed stores. The model was trained using features excluding temperature, considering the potential variability in weather data across different locations. The XGBoost model yielded a promising RMSE of 0.967 and an R-squared value of 0.857, indicating its effectiveness in capturing sales trends in these new stores. The model's accuracy suggests its applicability across various store contexts, although further refinements could be explored to enhance its predictive capabilities.

```
XGBoost (Imputed) - RMSE: 0.9669117993837459, R-squared: 0.8565003862114504
```

8. Conclusions:

1. Inventory Management Variability: There is significant variability in the received, sold, and end quantities across different products. This suggests a need for more tailored inventory management strategies for different product types.
2. Product Popularity and Sales Patterns: The sales data indicate varying levels of popularity among products. Some products consistently show higher sales, while others have lower demand.
3. Stock Levels and Waste Reduction: The end-of-day inventory levels for some products are consistently low, indicating good sales or understocking. Conversely, high end-of-day levels for other products might indicate overstocking and potential waste, especially for perishable items.
4. Stockouts: Occurrences of stockouts might be leading to missed sales opportunities. Regularly occurring stockouts on specific products suggest high demand that is not being met due to inventory shortages.

Recommendations:

1. Dynamic Inventory Replenishment: Implement a dynamic inventory replenishment system that adjusts the stock levels based on sales trends, seasonality, and specific store demands. This system should be sensitive to the sales velocity of each product.

2. Demand Forecasting: Utilize predictive analytics to forecast demand more accurately. This involves considering factors like historical sales data, day of the week, seasonal trends, and local events.

3. Tailored Stocking Strategies for Different Product Categories: Different products have different sales velocities and shelf lives. Perishable items need more frequent replenishment, while non-perishable items can have longer restocking intervals.

4. Reduce Stockouts: Identify products with frequent stockouts and increase their inventory levels appropriately to capture all potential sales. This might require a faster replenishment cycle for these high-demand items.

5. Waste Reduction Initiatives: For products with high end-of-day inventory levels, explore strategies like promotional pricing to move inventory more quickly and reduce waste, especially for perishable goods.

6. Real-time Inventory Tracking: Implement a real-time inventory tracking system to provide better visibility into stock levels, enabling quicker response to stockouts or overstock situations.

7. Training and Awareness: Train staff to understand the importance of accurate inventory tracking and the impact of stockouts and overstock on the business.