Walmart's Weekly Sales Analysis Part 3: Spike Analysis and Store Segmentation

By

Temidayo Olowoyeye

Data and Business Analyst

1. Introduction

As the title implies, spike analysis in sales involves identifying and understanding transactions that deviate significantly from normal sales patterns. For large retail organizations like Walmart, there are periods where weekly sales can experience dramatic increases. These spikes can be due to various factors such as seasonal demand, promotional events, or even external economic influences.

Analyzing these spikes effectively involves adopting advanced machine-learning techniques such as the Isolation Forest model, a tool commonly used in fraud detection due to its ability to identify anomalies, which plays a crucial role in our analysis. Its versatility beyond fraud detection makes it ideal for analyzing Walmart's weekly sales data.

This analysis is not limited to identifying sales spikes but also employs Hierarchical clustering for store segmentation. This segmentation will help differentiate normal sales fluctuations from significant spikes. Hence, understanding how different stores perform weekly can influence strategies to suit specific segments, leading to more targeted and effective interventions.

1.1. Objective

The main objective of this analysis is to comprehensively understand and analyze Walmart's weekly sales patterns. The specific objectives are to:

i. Detect the spike in Walmart's weekly sales
ii. Quantify the weekly sales spike
iii. Assess store weekly sales performance

1.2. About the Dataset

The data was extracted from this link. It contains the Export sales records from 2010 to 2012, with 6435 rows and eight (8) columns.

1.3. Data Processing

Part 2 of this analysis explains the data processing. It can be accessed here. This and all other analysis processes were done using Python, and the codes can be accessed via this link.

2. Data Analysis

2.1. Spike Modeling and Detection

This analysis adopted an isolation forest (ISF) model for spike detection. To achieve this, an essential component called the contamination parameter must be carefully chosen. The weekly_sales data was carefully explored using statistical methods and visualization, including a histogram with a KDE plot, a box plot, a Q-Q plot, and a combination of histogram and percentiles, as shown in Figure 1.

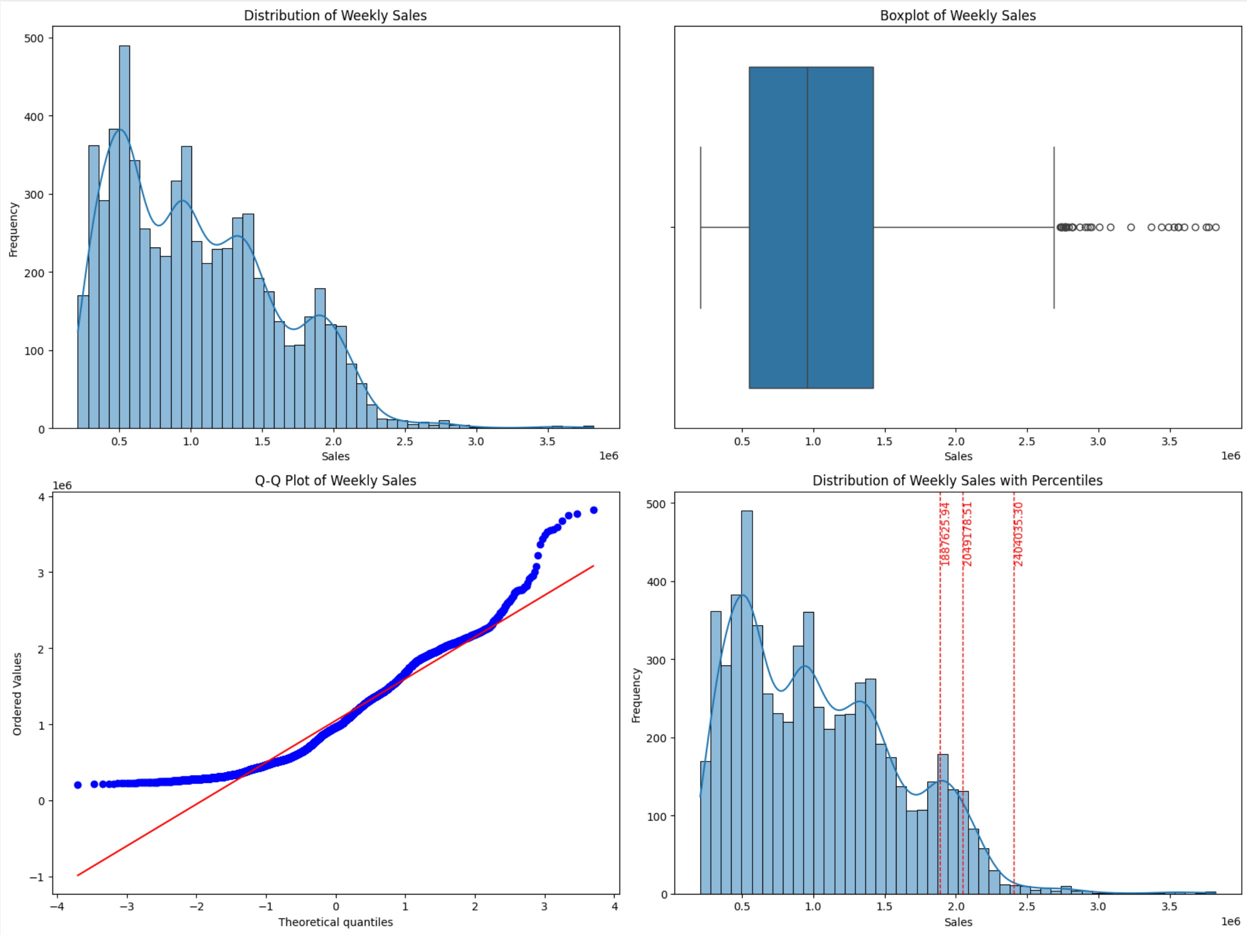

Figure 1: Showing Statistical analysis of Weekly Sales

From the statistical analysis, the following was observed:

i. The histogram with a density plot showed a right-skewed distribution with most sales clustering between 0 and 1.5 million and a long tail of higher values

ii. the boxplot highlighted the median, quartiles, and potential outliers, indicating several high-end outliers

iii. The Q-Q plot showed a significant deviation from a normal distribution, especially in the tails, suggesting a heavy-tailed distribution;

iv. The histogram annotated with percentiles showed the dataset's 95th, 97th, and 99th percentile values.

Based on this observation, the 99th percentile for contamination in the isolation forest was adopted to ensure that only the top 1% of sales values are captured as spikes. This approach allows for the precise detection of the most unusual sales periods without misclassifying regular high sales, is supported by the data's significant skewness and heavy tails, and ensures that the analysis remains highly accurate and relevant by pinpointing critical patterns or events without being overly inclusive.

To apply this, the dataset was split into training and testing sets. The model was fit on the training set and subsequently tested on the testing set, with the results presented in Figure 2 below.

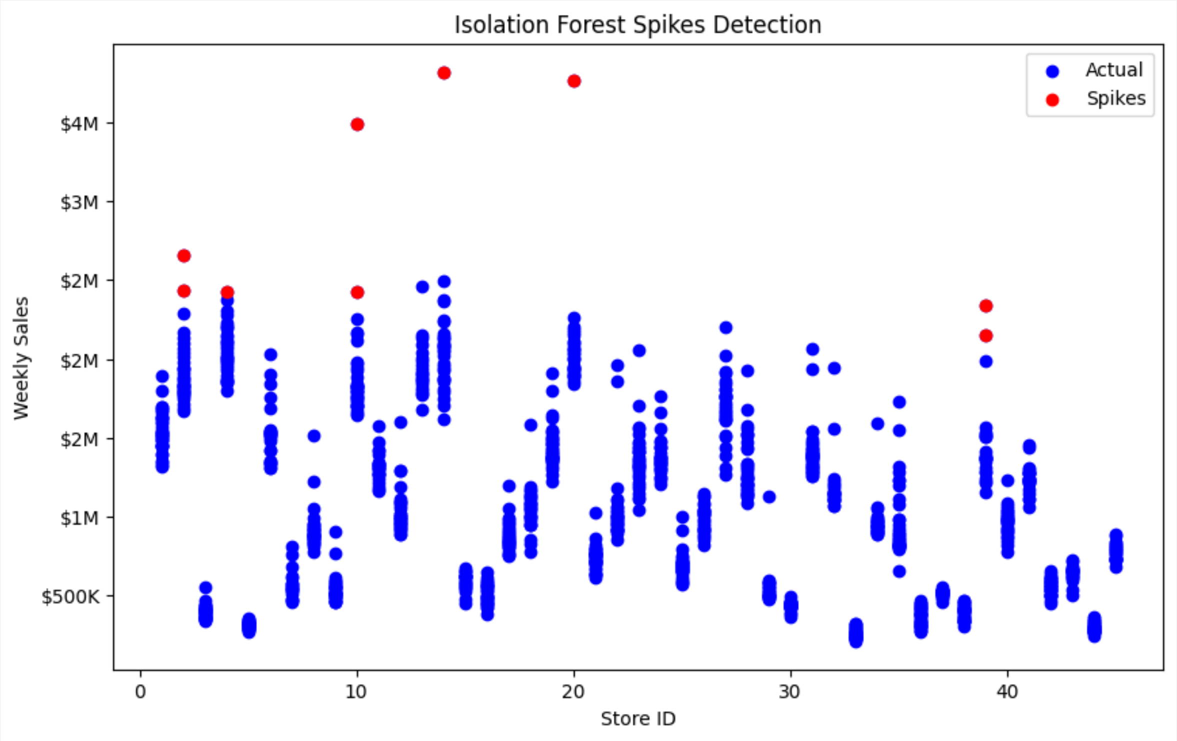

Figure 2: Showing the Result ISF model on Tested data

Furthermore, the model was saved and fitted on the entire dataset to predict sales spikes between 2010 and 2012. The results indicate multiple sales spikes during this period, as shown in the Figure 3 below.

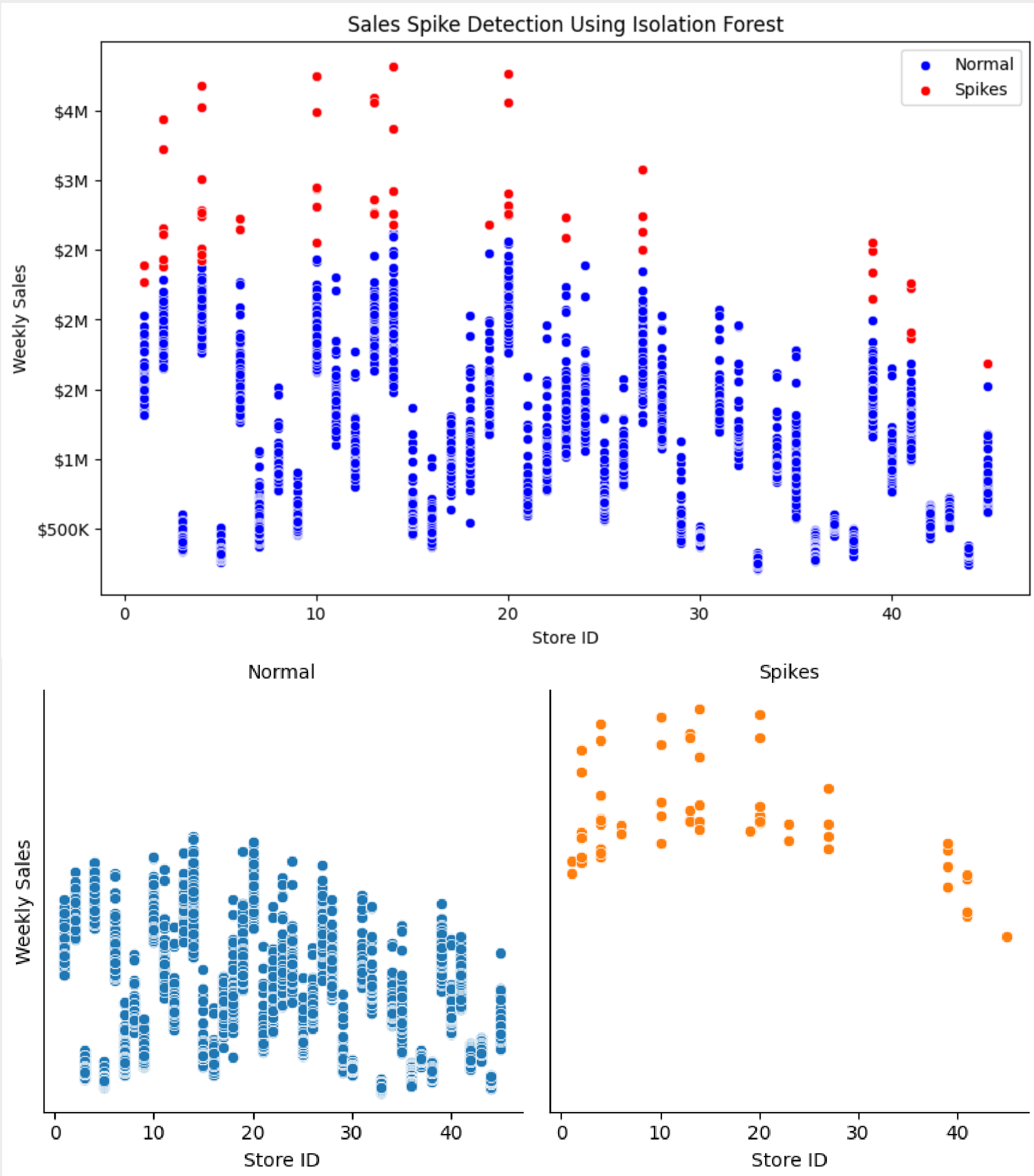

Figure 3: Showing Predicted Sales Spike from 2010 - 2012

2.2. Spike Analysis

i. **Spike Occurrence**: From the analysis (Figure 4), it was observed that 14 Walmart stores out of 45 experienced sales spikes between 2010 and 2012. Store 4 had the highest sales spikes during this period, occurring nine (9) times. This was followed by Store 2 and Store 20, with seven (7) occurrences each. These spikes can be attributed to various factors, such as the geographical locations of the stores, store-specific factors (such as store size, product range, and customer service quality), and seasonal variations, among others.

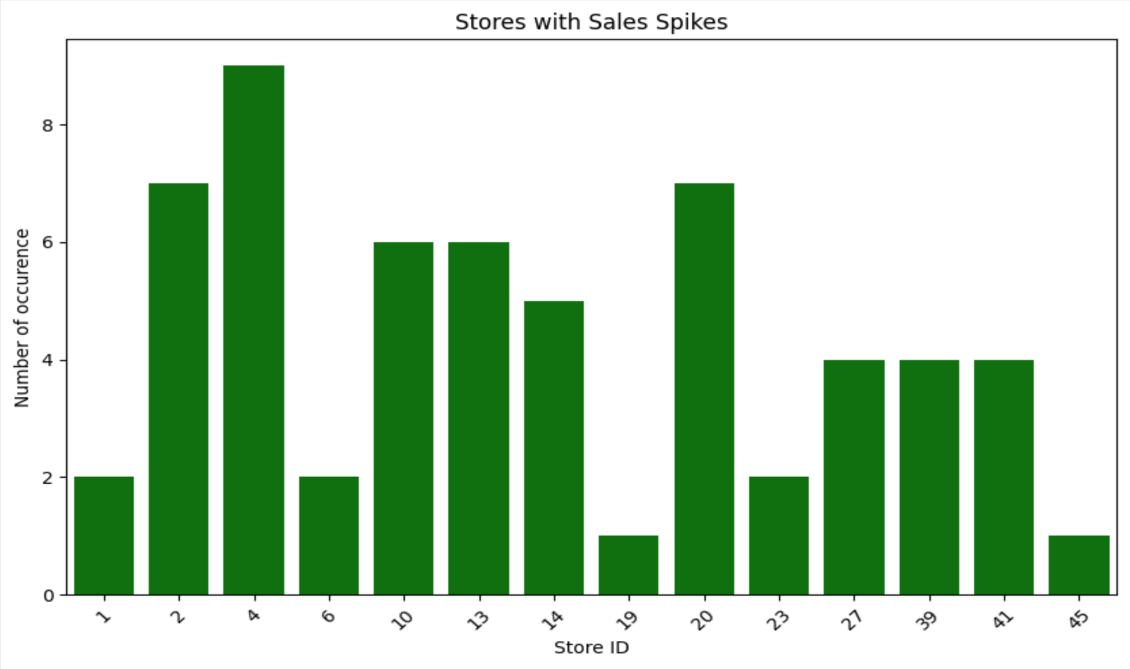

Figure 4: Showing Spike Occurence per Store

ii. Sales Spike Per Year: Between 2010 and 2012, total sales from spikes amounted to \$167,473,735 (see Figure 5). In 2010, the highest spike in sales was recorded at \$87,079,848.03, followed closely by 2011, with sales of \$75,496,041. However, a drastic fall in sales spikes was observed in 2012, with sales falling to \$4,897,846. This decline could be attributed to various factors, including changes in economic conditions, competitive actions, or shifts in consumer behaviour.

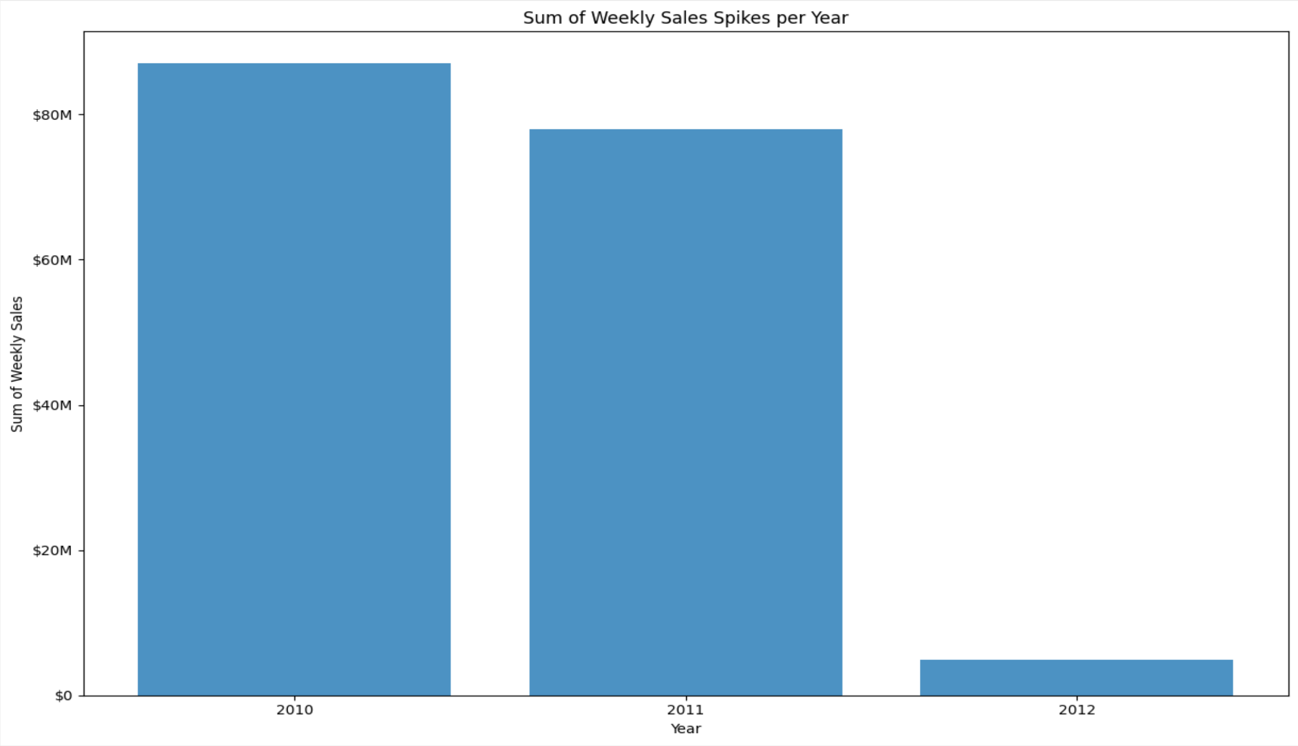

Figure 5: Showing Sales Spikes per Year

iii. **Sales Spike per Month:** Similar to yearly sales spikes, monthly sales spikes were observed in 6 out of the 12 months, with the highest in December, followed by November. These spikes are caused by seasonality in sales, accompanied by promotions and increased demand during festive periods. The holiday season in November and December typically drives higher consumer spending due to festivities, gift-giving, and holiday promotions.

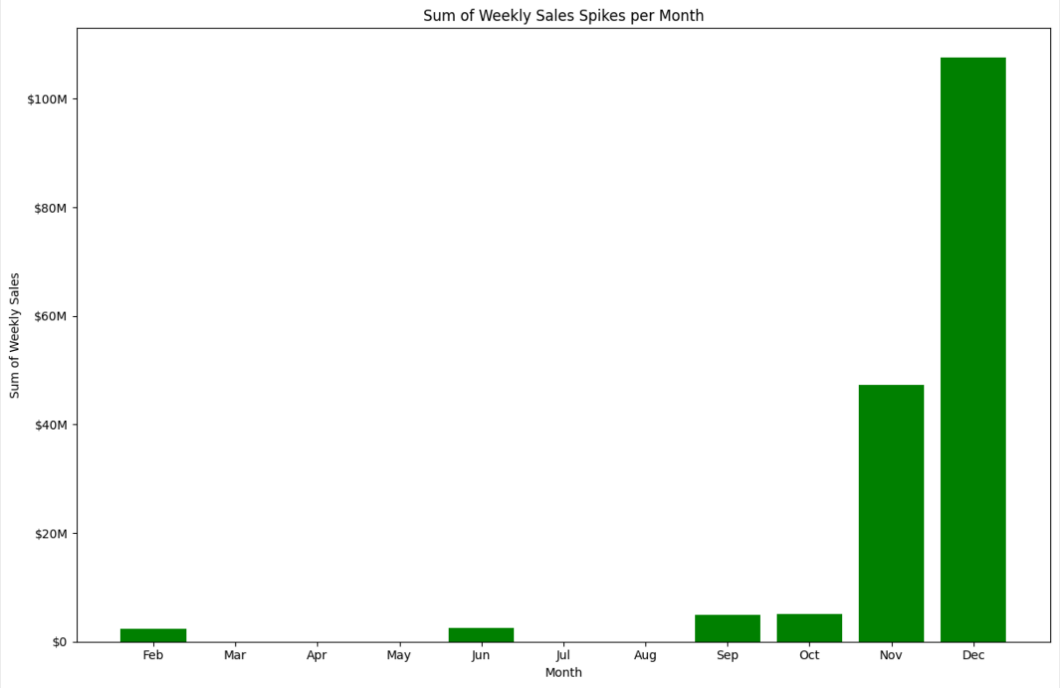

Figure 6: Showing Sales Spikes per Month

iv. **Store-Specific Spikes**: This analysis is drilled down to how these spikes occur in stores on a yearly basis. In 2010, Store 20 recorded a $12.2 million sales spike, followed closely by Stores 2, 10, and 13. Similarly, in 2011, Store 4 had the highest sales spikes of $12 million, closely followed by Stores 13, 2, and 10. In 2012, only Store 4 experienced a spike in sales, amounting to $4.5 million. Store 4 recorded the highest total sales spikes between 2010 and 2012.

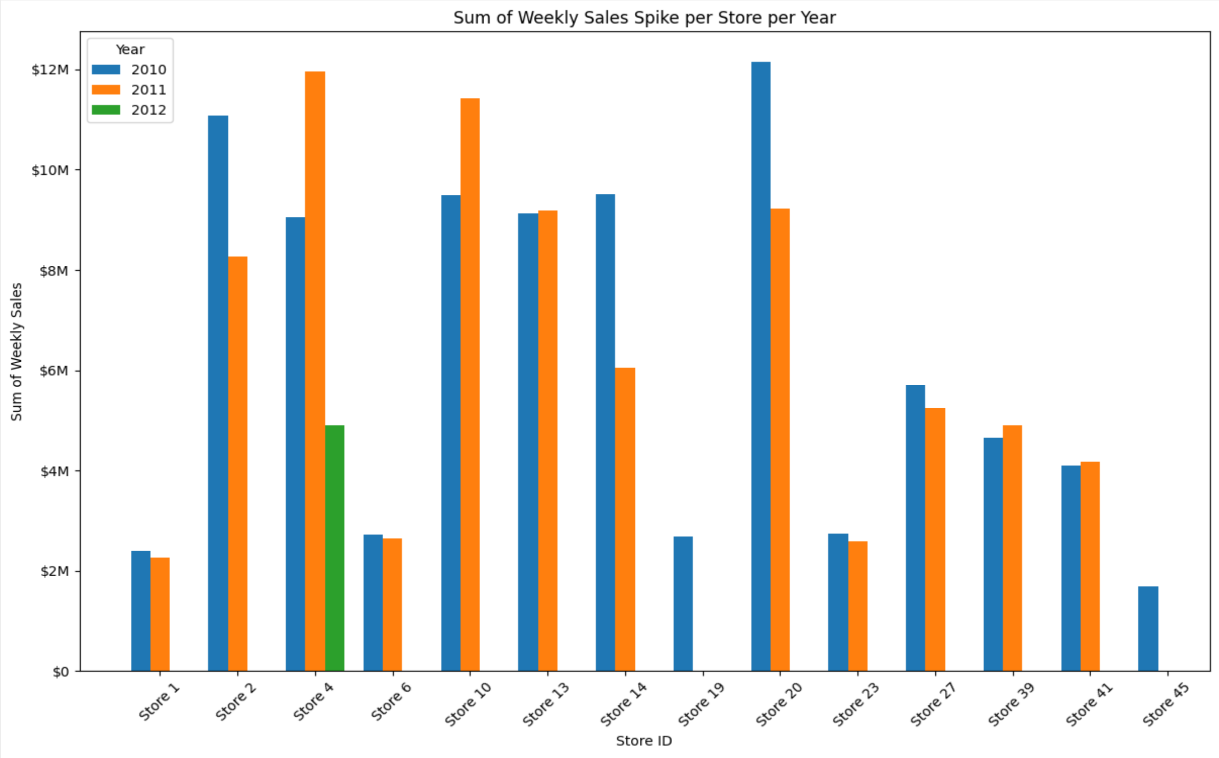

Figure 7: Showing Store Specific Sales Spikes per Year

2.3. Store Segmentation

The Store segmentation was carried out in two stages;

Stage 1: Hierarchical Clustering

The first stage of hierarchical clustering allows for identifying nested groupings within the data. It thus provides insights into the relationships between individual stores based on their weekly sales. This was done by visualizing the clustering using a dendrogram (refer to the figure). This tree-like diagram records the sequences of splits or mergers in hierarchical clustering. The length of the branches represents the distance or dissimilarity between clusters (measured using Euclidean distance). Short branches indicate the merged clusters are very similar, while longer branches indicate more significant dissimilarity.

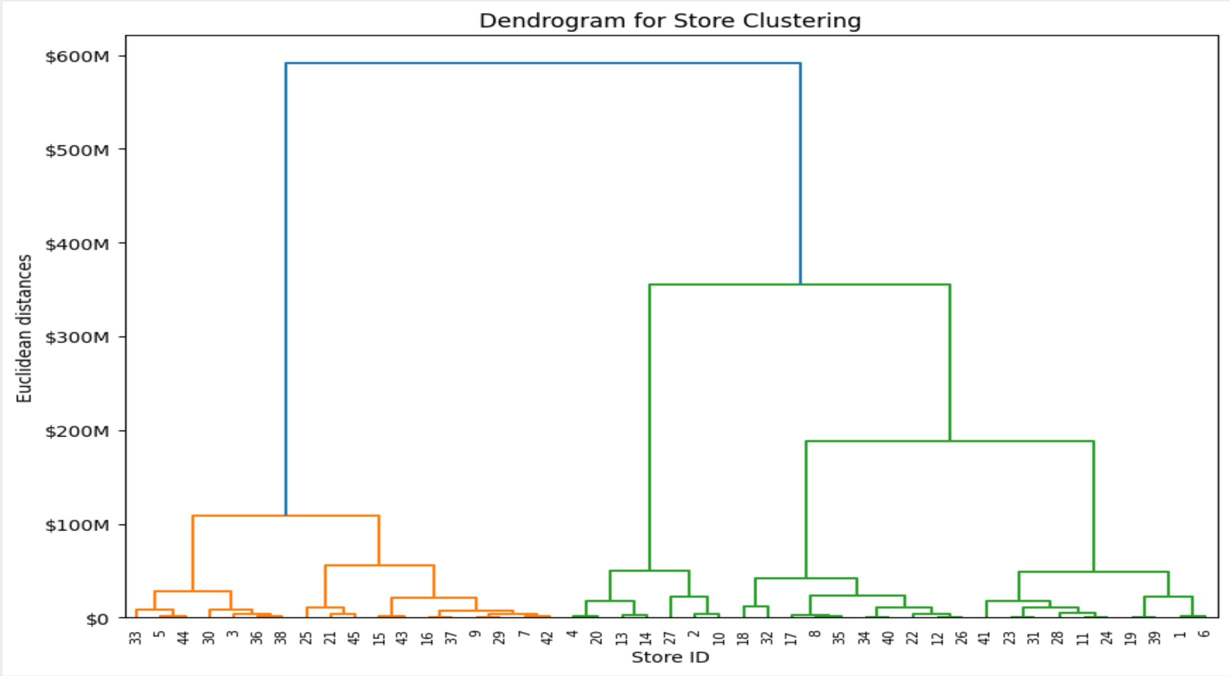

Figure 8: Showing Hierarchical clusters of Stores

Probing further at the bottom of the dendrogram, each store is its cluster. This represents the most granular level where no clustering has taken place. The first merges occur at the lowest points on the Euclidean distance scale, indicating that the merged stores have similar weekly sales figures. For example, stores 33 and 5 merge early, indicating similar weekly sales. The second and subsequent merges occur as we move up the dendrogram, where small clusters of similar stores merge into larger clusters. For example, a sub-cluster involves stores 4, 20, 14, 13, and 27, 2, 10.

Further sub-clusters are formed at lower heights within these main clusters, indicating closer similarities among stores within these sub-clusters. Lastly, the dendrogram revealed two major store groups: low-performing stores (orange cluster) and high-performing stores (green cluster). These groups indicate significant differences in weekly sales.

Stage II: Store Classification

Based on the store sales cluster analysis, the weekly sales were classified into 5 clusters: very low, low, average, high, and very high, as shown in figure 9 below. From the analysis, 15.56% of the stores fell into very low and very high weekly sales categories, 22.22% fell into average and high sales categories, and most fell into low sales categories (24.44%). Specifically, Stores 33, 44, 5, 36, 38, 3 and 30 have very low sales performance. In contrast, stores 27,10, 2, 13, 14, 4 and 20 have very high weekly sales performance, with the other stores falling within low, medium and high weekly sales performance categories, respectively.

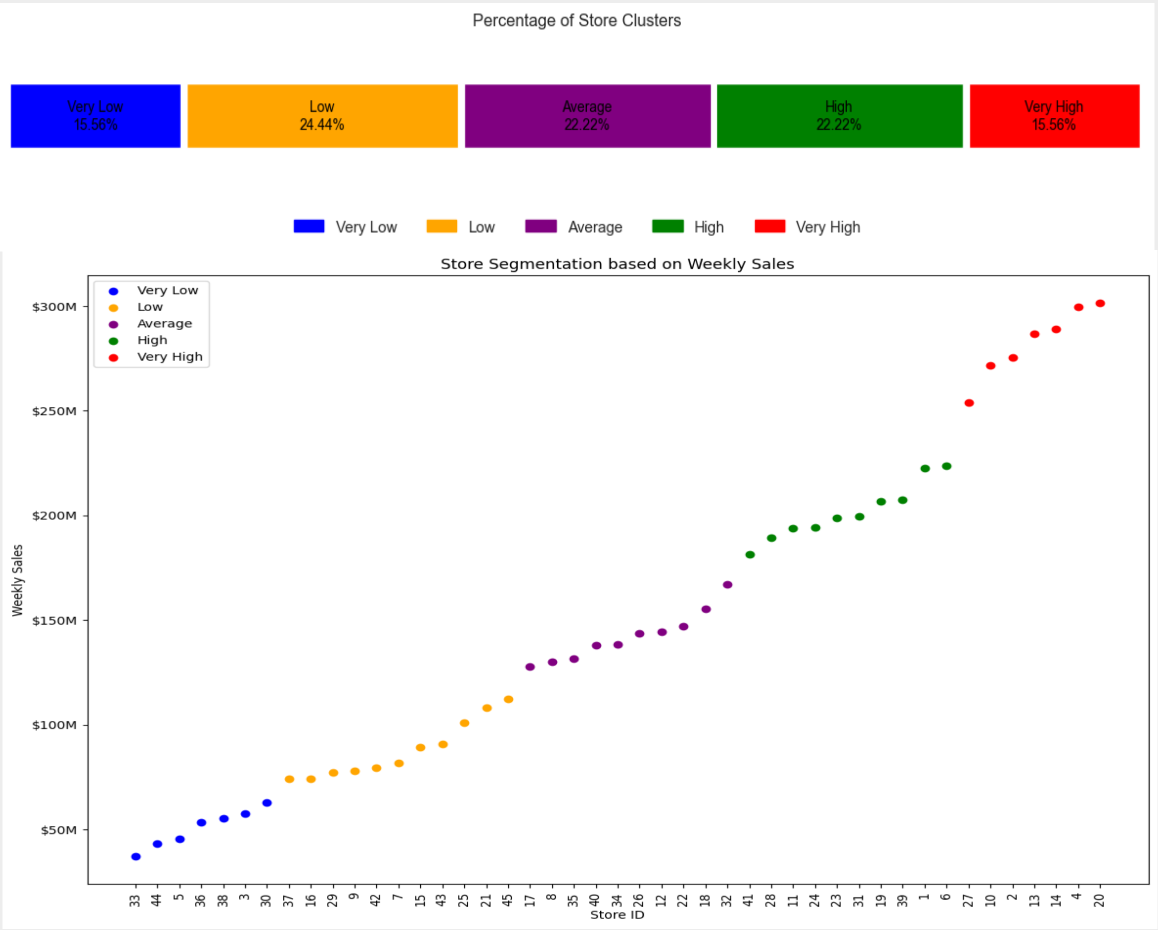

Figure 9: Showing Store Classification by Weekly Sales

2.4. Effect of Sales Spike on Store Segmentation

One key observation from this analysis is a correlation between the stores that experience sales spikes and those that fall into very high sales performance. Seven (7) stores out of 14 that experienced sales spikes were classified as very high performing, with Store 20, which experienced seven (7) occurrences of sales spikes, topping this category. Others with sales spikes were in the high-performing clusters except for Store 45, which, despite experiencing sales spikes, fell to the low-performing cluster.

This observation suggests that sales spikes can significantly indicate overall sales performance. Stores that frequently experience sales spikes tend to have higher overall sales, possibly due to successful marketing campaigns, promotional events, or a larger customer base. Store 20, in particular, might be employing effective strategies that lead to repeated spikes in sales, positioning it as a top performer.

More investigation will be required to better understand these dynamics. This will include assessing the causes of sales spikes and the factors influencing overall store performance. Analyzing marketing strategies, customer demographics, and operational practices will also provide further insights into why some stores leverage sales spikes into sustained high performance.

3. Conclusion and Recommendation

3.1. Recommendations

i. For stores like 4, 20, and 2 that frequently experience spikes, continuously leverage effective marketing campaigns and promotions, especially around the holiday season, to sustain and boost performance.

ii. Probe the factors leading to spikes in stores and develop targeted strategies to convert these spikes into sustained sales growth.

iii. Increase promotional activities in November and December across all stores, as these months historically show higher spikes due to festive demand.

iv. Evaluate operational practices, customer service quality, and product ranges in high-performing stores to identify best practices that can be applied to lower-performing stores.

v. Develop strategies for each store cluster. For High-performing clusters, focus on innovation and expanding successful promotions, while low-performing clusters will benefit from foundational improvements in inventory management, customer engagement, and localized marketing.

3.2. Conclusions

The analysis of Walmart's weekly sales from 2010 to 2012 reveals significant insights into sales spike patterns and store performance. The Isolation Forest model effectively identified spikes, highlighting the importance of seasonal demand and promotional activities. Store segmentation through hierarchical clustering provided a clear distinction between low and high performers, with a notable correlation between frequent sales spikes and overall high performance.

Further improvement can boost performance by implementing approaches to enhance sales, mainly focusing on high-impact periods and optimizing practices in underperforming stores. In addition, investigation into store-specific factors and monitoring of sales patterns will be essential in maintaining and improving sales performance across all stores.