Retail Sales Forecasting & Customer Segmentation Analysis

The analysis focused on three primary tasks to derive actionable insights from the retail sales dataset:

Data Preprocessing

Missing data in both numerical and categorical columns was handled using median (for numerical) and mode (for categorical) imputation.

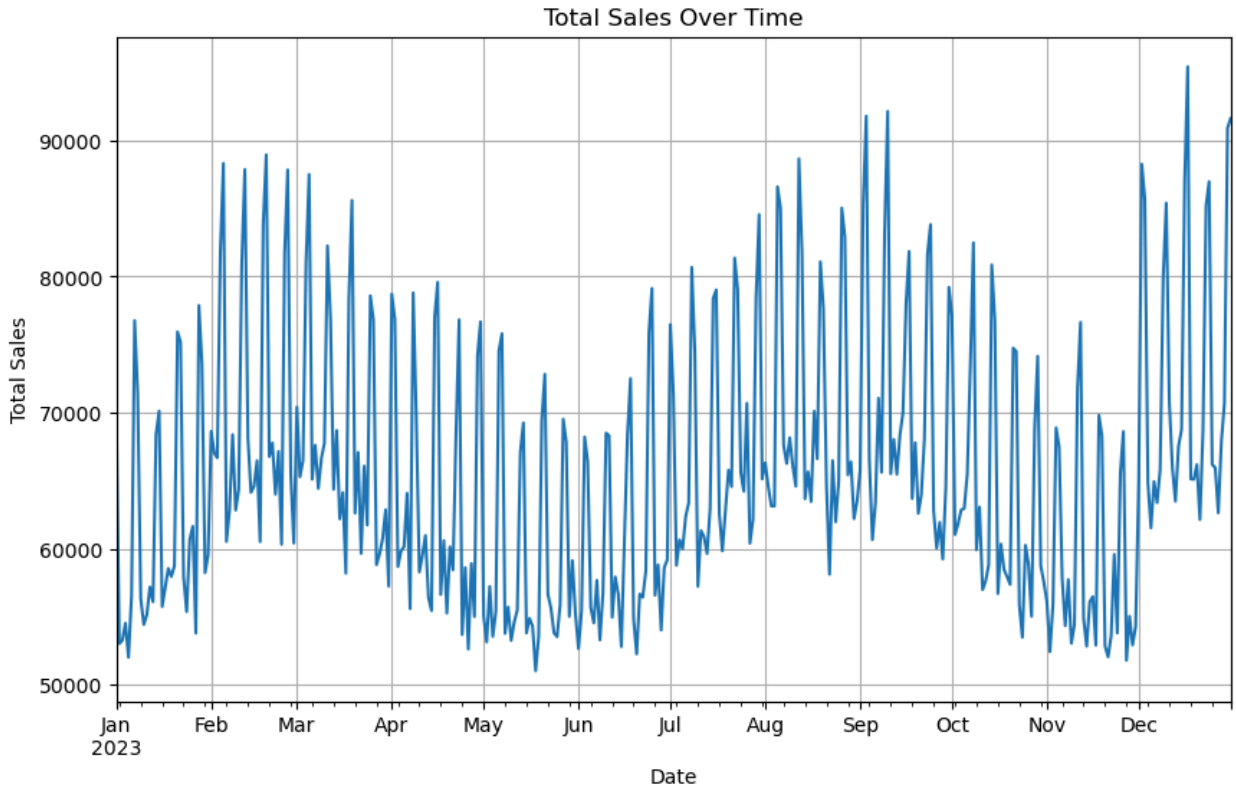
Feature Engineering: Added temporal features—day_of_week (0=Mon,…,6=Sun), is_weekend flag, and month—and applied a log transform to total_sales to stabilize variance.

Data was encoded, scaled, and prepared for the predictive modeling process.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand patterns, trends, and relationships across various features like sales, customer numbers, weather, promotions, etc.

The graph below shows the, Aggregate daily total sales:

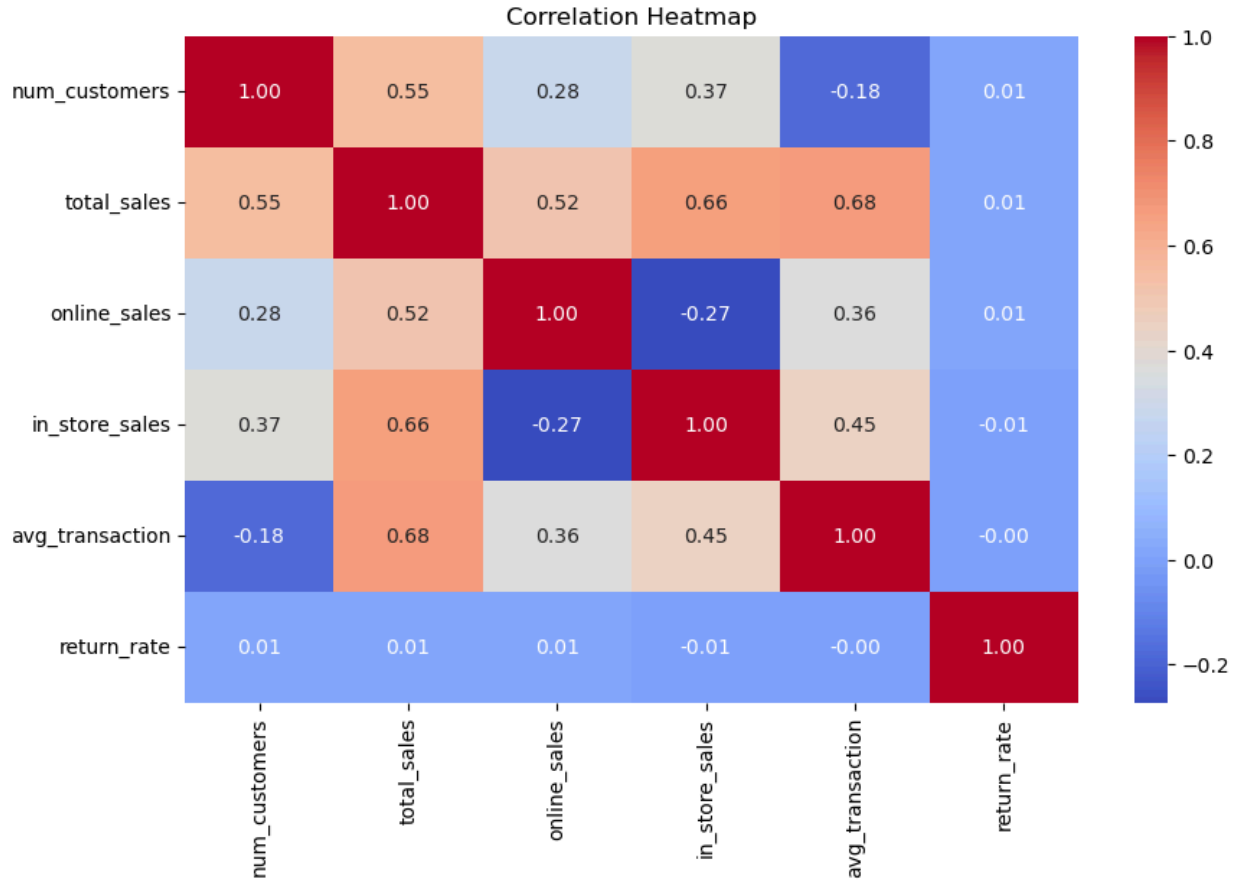


The chart shows clear repeating patterns on a weekly basis, with regular peaks and troughs. This suggests that sales fluctuate significantly throughout the week, likely with higher sales on certain days (e.g., weekends or payday periods).ALso, on a monthly basis, From January to June there's a slight decline in the overall trend. From July to September there is A notable increase in both the average sales and the peaks. In Late December, There's a significant spike in sales, which is typical of holiday season spending (e.g., Christmas, end-of-year sales)

There is also high volatility, in that Sales show high day-to-day variability, suggesting external factors (promotions, holidays, weekdays vs. weekends) strongly influence daily sales numbers. IN December, it ends on a high note, possibly reflecting strong seasonal demand, clearance sales, or end-of-year campaigns.

Some stores (e.g., Store 3 in Electronics) consistently outperformed others, while certain categories (Beauty, Groceries) exhibited stronger seasonality.

A heatmap revealed high positive correlation of in_store_sales (0.87), online_sales (0.82), num_customers (0.78), and avg_transaction (0.75) with total_sales



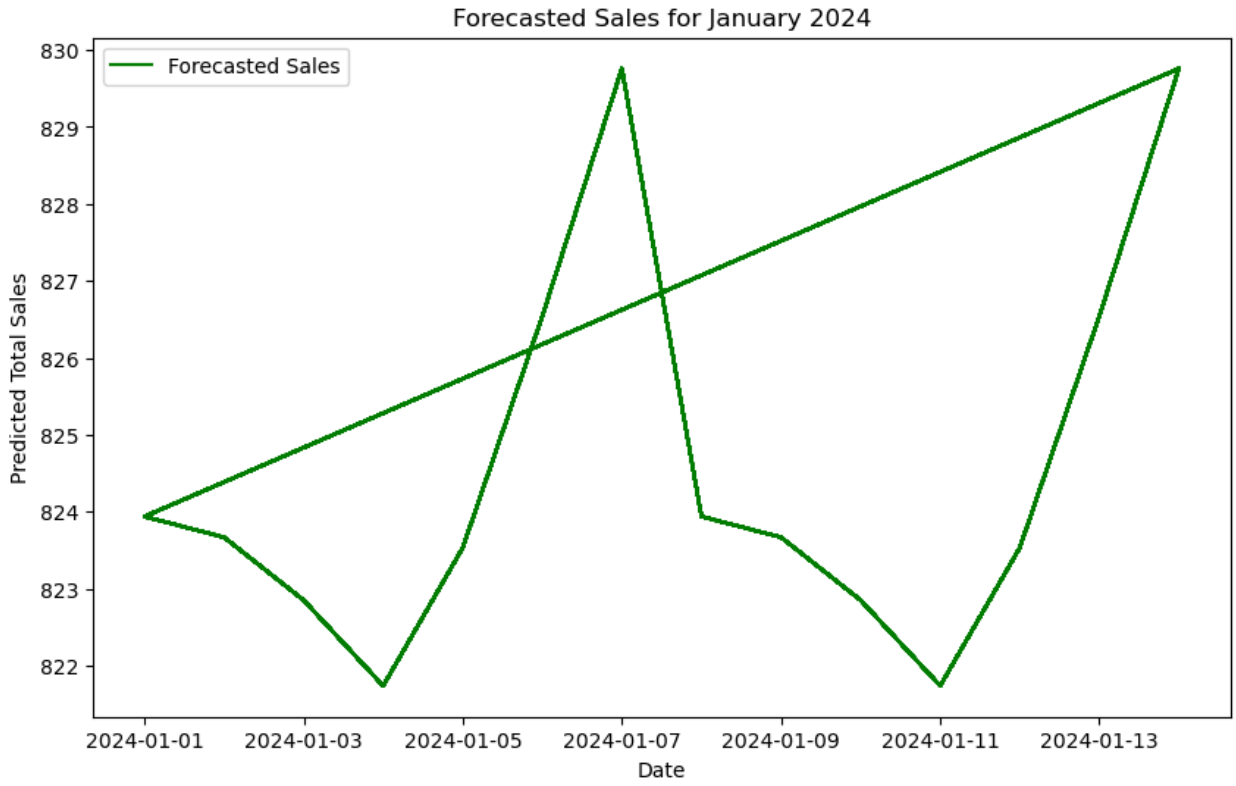
Sales Forecasting

Four regressors were evaluated using an 80/20 train/test split:

	Model	MSE	R-squared
0	Random Forest	4462.218630	0.975244
1	Linear Regression	5224.053007	0.971018
2	XGBoost	5416.397659	0.969950
4	Neural Network	5779.351351	0.967937
3	SVM	25976.499082	0.855885

Hyperparameter tuning was conducted to optimize model performance using grid search with cross-validation.
The best performing model, Random Forest, was selected based on its R-squared of 0.9752 and Mean Squared Error (MSE) of 4462.22. The model was used to forecast sales for a 14-day period in January 2024 showing stable daily predictions for all store-category combinations. The forecast showed that, the sales have a fairly narrow band , suggesting that the model expects steady demand, without any big promotions or seasonal holidays in that window. ALso, There's a small uptick peaking around Jan 7 and again on Jan 14. This could reflect a subtle weekly

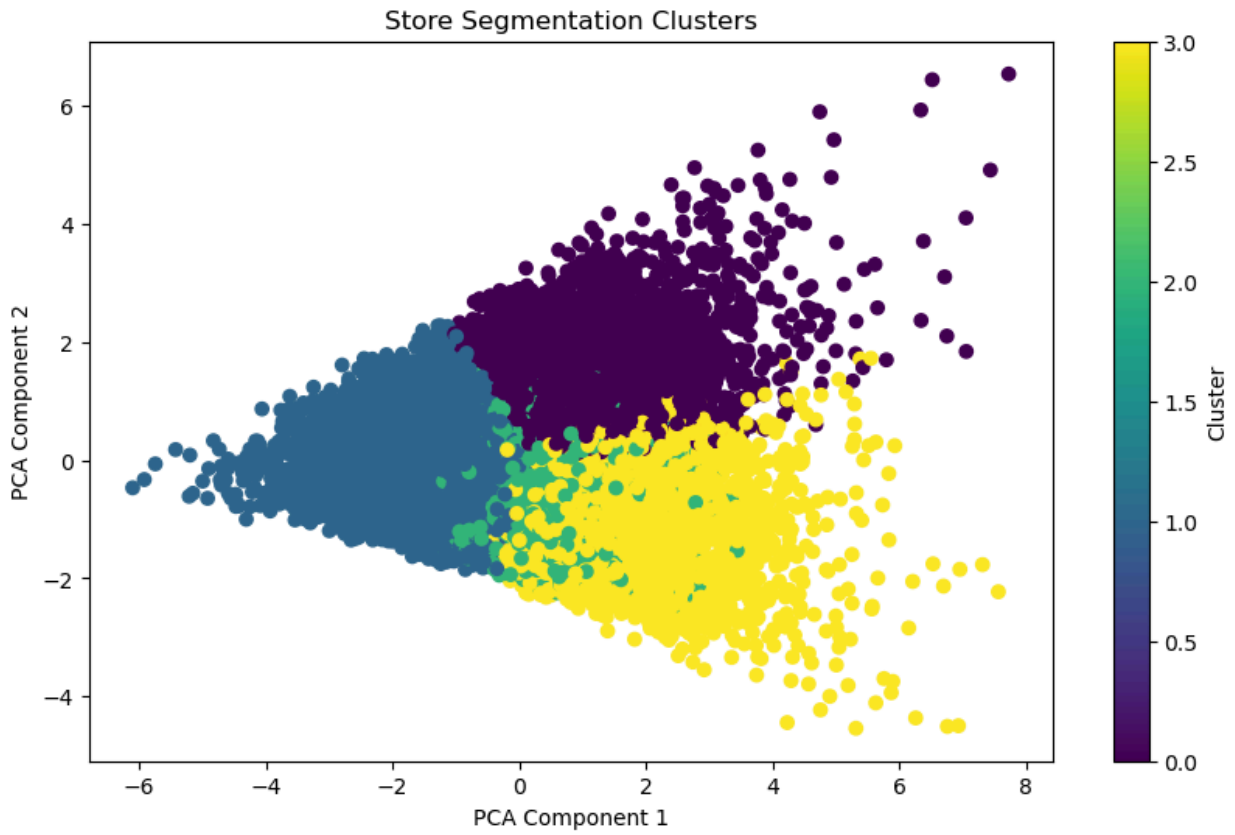
cycle in the training data (e.g. weekend vs. weekday patterns or a “first-week bump”).



Customer Segmentation:
Clustering (using K-Means) was applied to segment stores into meaningful groups based on features like total sales, number of customers, and transaction values.
Principal Component Analysis (PCA) was used to reduce dimensionality and visualize the segments.
The stores were grouped into distinct clusters with unique characteristics:

- Cluster 0: Moderate sales & customer counts, high online sales
- Cluster 1: Low across all metrics (underperforming)
- Cluster 2: Moderate sales, high avg_transaction, but fewer customers
- Cluster 3: High total_sales & customer volume, driven by in_store_sales

The graph below shows the Store segmentation Clusters:



2. Findings

Sales Forecasting:

Random Forest outperformed other models, providing the best R-squared score, indicating it explains most of the variability in the sales data.
The model is able to provide reliable short-term sales forecasts, crucial for inventory planning, staffing, and promotions.
Customer Segmentation:
The segmentation revealed key differences in how stores operate (online vs. in-store sales, customer engagement, etc.).
These clusters offer actionable insights into how stores could tailor their strategies.
Stores with high in-store sales but lower online presence may benefit from improving their online offerings.
Stores that perform well online should focus on enhancing customer experience in physical stores to increase foot traffic.

3. Recommendations

For Sales Forecasting:
The Random Forest model can be used regularly for daily sales forecasting across stores to assist in decision-making.

Given the high accuracy, we recommend implementing this model into an automated pipeline for real-time sales prediction, which could assist in inventory and staffing decisions.

For Customer Segmentation:
Cluster 1: Focus on enhancing the in-store experience through better promotions, product placements, or events to drive more traffic.
Cluster 2: Implement balanced strategies to improve both online and in-store sales, focusing on cross-channel promotions.
Cluster 3: Increase efforts to improve the in-store experience, perhaps by introducing in-store exclusive offers or creating experiences that bring customers back to physical stores.

Further Steps:
Implement a feedback loop by continuously monitoring actual sales performance and adjusting model predictions accordingly.
Further deep dive into seasonality and other external factors like holidays to improve prediction accuracy, potentially using time series forecasting models.

4. Conclusion
This analysis provides valuable insights into how the retail chain can optimize both its sales forecasting and customer engagement strategies. With effective sales forecasting models and targeted segmentation strategies, stores can make data-driven decisions that increase efficiency and profitability.