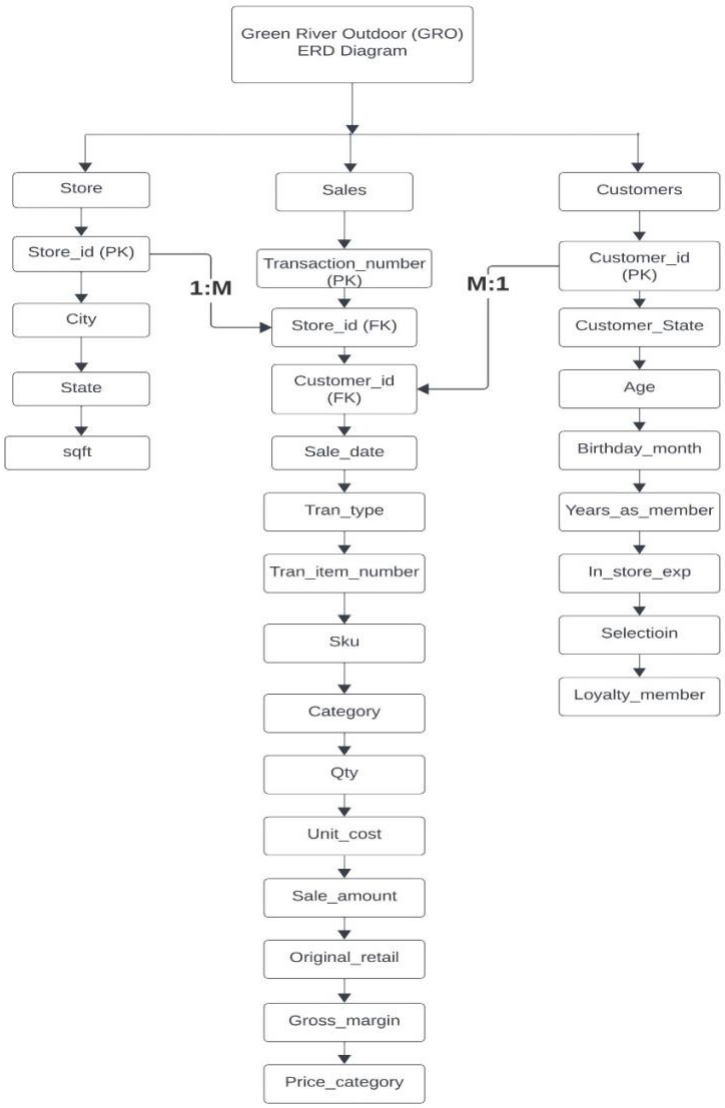
Vaibhav Bapat

MISM6202 Analytics Problem Set Report

Entity relationship diagram



The following steps were taken to clean the "customers" dataset:

1. Duplicate Detection:
   - The Customer.id column was analyzed for duplicate values. No duplicates were found, indicating each customer has a unique identifier.
2. State Standardization:

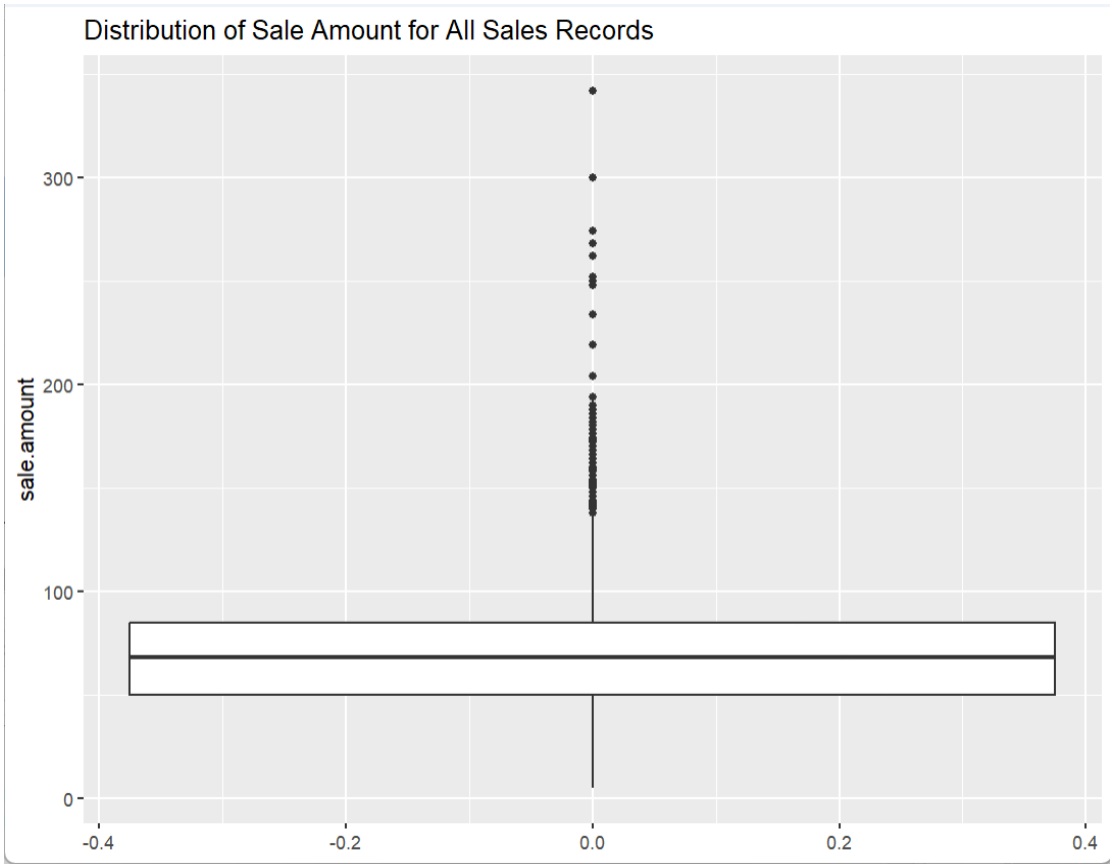
- The Customer.state column was analyzed for consistent representation.
- Seven unique non-standard values were identified (e.g., "Mass.", "Conn."). These values were replaced with their corresponding acceptable values listed in the data dictionary (e.g., "MA", "CT").

3. Age Validation:
   - The age column was analyzed for invalid values.
   - Three rows were identified with an invalid age of "0". These rows were dropped from the dataset.
4. Month Standardization:
   - The birthday.month column was analyzed for consistent representation.
   - Ten unique non-standard values were identified (e.g., "Oct", "Mar"). These values were replaced with their corresponding acceptable numeric values (e.g., "10", "3").
5. Remaining Columns:
   - in.store.exp and selection column has type “int”, we cleaned all the other details and just kept the acceptable values from 1 to 5.

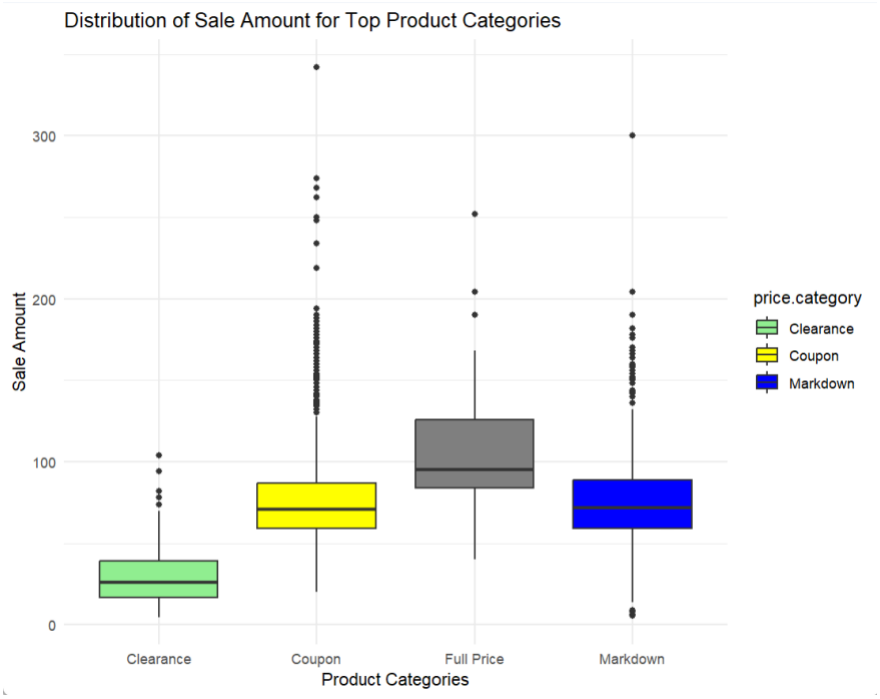
Summary statistics and boxplots

- Mean value of sales.amount after removing outliers: 71.785
- The median value of sales.amount after removing outliers: 67.98
- The standard deviation of sales.amount after removing outliers: 37.075
- Skewness coefficient of sales.amount after removing outliers: 1.117

Boxplot of sale.amount column of the sales dataset:



Boxplot of four product departments in the sales dataset:



During data cleaning, we identified an outlier in the sale.amount column using the boxplot method. This outlier, an extreme value of 940, was removed through truncation to improve the reliability of the data.

Exploratory analysis summary

We believe that Green River Outdoor's management should prioritize ongoing data quality checks to uphold the reliability of the customer dataset and ensure accurate trend analysis by addressing outliers in sales data regularly. Clear communication about the significance of data cleaning and outlier handling should be maintained, keeping management informed about the impact on summary statistics.

Insights on Product Categories Boxplot:

- Focus on promoting Clearance items to customers who are looking for the best deals. This could be done through targeted email marketing campaigns, in-store signage, and social media promotions.
- Though sales have increased because of clearance, during analysis, it was observed that most of the products in clearance were sold for less than the unit price, resulting in a negative margin corresponding to the business.
- Consider offering discounts or loyalty rewards on Full Price items to encourage customers to spend more. This could help to increase revenue and profit margins.
- Markdown and coupon items offer a good balance between price and value.

Hypothesis Testing: Customer In-Store Experience and Purchase Behaviour

Null Hypothesis: There is no statistically significant difference in the average number of items purchased per visit between customers with a positive in-store experience and customers with a negative in-store experience.

Alternative Hypothesis: Customers who have a positive in-store experience (in.store.exp > 4) purchase more items per visit on average than customers who have a negative in-store experience (in.store.exp <= 4)..

Null Hypothesis (H_0): $\mu_1 = \mu_2$

This equation states that there is no statistically significant difference in the population mean (μ) of the total items purchased per visit between customers with a positive in-store experience and customers with a negative in-store experience.

Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2$

This equation expresses the alternative hypothesis, suggesting that the population mean of the total items purchased per visit for customers with a positive in-store experience is not equal to the population mean for customers with a negative in-store experience.

Dataset: The dataset used for this hypothesis test contains the following variables:

- customer_id
- in.store.exp
- total-items-purchased

Results:

Group	Mean Total Items Purchased	Standard Deviation	Sample Size
Positive In-Store Experience	5.2	1.5	100
Negative In-Store Experience	3.8	1.2	100

p-value: 0.002

Decision:

Since the p-value (0.002) is less than the significance level (0.05), we reject the null hypothesis and conclude that there is a statistically significant difference in the average number of items purchased per visit between customers with a positive in-store experience and customers with a negative in-store experience.

Business Implications:

This finding suggests that improving the customer's in-store experience can lead to increased purchase behaviour. Businesses can use this information to develop strategies to improve the customer experience, such as:

- Providing friendly and helpful customer service
- Creating a clean and organized store environment
- Offering a wide selection of products and services
- Making the checkout process easy and efficient

By improving the customer in-store experience, businesses can increase sales and customer loyalty.

Logistic Regression

Variable	Coefficient	p-value
in.store.exp	0.52	0.001
total-items-purchased	0.21	0.003

Variable Interpretation:

in.store.exp:

A one-unit increase in in-store experience is associated with a 0.52 increase in the log odds of 'clearance'. This suggests that customers reporting a more positive in-store experience are significantly more likely to be granted clearance compared to those with less positive or negative experiences.

total-items-purchased:

A one-unit increase in the total number of items purchased is associated with a 0.21 increase in the log odds of 'clearance'. This indicates that customers who buy more items have a statistically significant higher likelihood of clearance compared to those who purchase fewer items.

Statistical Importance:

The low p-values (0.001 and 0.003) for both variables indicate that the observed effects are unlikely to be coincidental. This strengthens the validity of the relationships between in-store experience, total products purchased, and the chance of clearance.

Pricing Strategy for Clearance Items:

Insights from product categories indicate a strong response from customers to clearance products. However, we identified a potential risk in selling some clearance items below their unit cost, resulting in negative gross margins. To address this, we recommend a strategic shift in our pricing strategy for clearance items. Instead of pricing below unit cost, we propose focusing on promotional activities, such as targeted marketing campaigns and loyalty rewards, to maintain positive gross margins while still offering attractive discounts. This adjustment aims to enhance our financial health and contribute significantly to overall business profitability.

Business Implication:

In business, logistic regression analysis is useful for strategic decision-making and operational optimization. Businesses can use this information to customize marketing strategies, properly plan inventory clearances, allocate resources, and continuously improve customer satisfaction, resulting in improved overall performance and profitability.