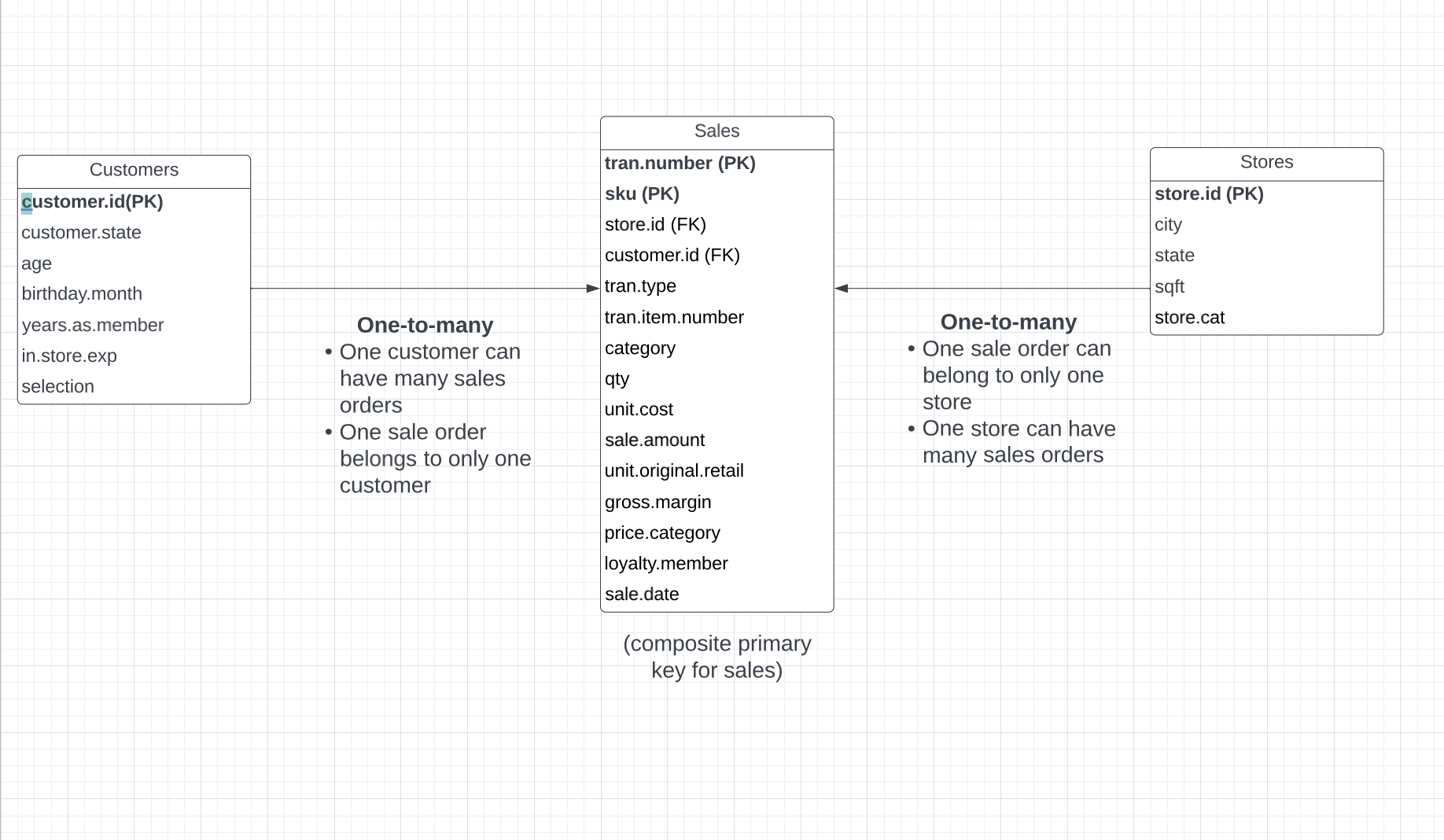
**Green River Outdoor Analytics Report**

Leticia Teixeira da Costa

**PART 1**

Entity Relationship Diagram

Cleaning and wrangling steps done for the ‘Customer’ dataset

These were the cleaning and wrangling steps completed before the data was analyzed:

* Fix inconsistencies in the ‘State’ column: We adjusted categorical labels to conform to ‘acceptable values’ of state as only two letter abbreviations using the mutate and str\_detect functions.
* Fix inconsistencies in the "Birthday Month" column: We adjust categorical labels to conform to ‘acceptable values’ of month as the numerical correspondent using the mutate and str\_detect functions. We also identified incorrect values (0) for the birthday month in two rows. Due to the low occurrence, the incorrect rows were removed.
* Checked for problematic numerical data in "Age" and "Years as Member" columns, identifying two rows where age was listed as 0. Created a function to identify age outside of acceptable bounds for store consumers and remove those rows.
* The columns "In Store Experience" and "Selection Satisfaction" both had 78% empty cells, indicating some problem in the collection of this data. As both columns regard customers' opinions to the in store experience, we decided that it would not be appropriate to assume or impute values for that column and chose to remove both of them.

Summary Statistics for ‘Sales Amount’

|  |  |  |  |
| --- | --- | --- | --- |
| Mean | Median | Standard Deviation | Skewness Coefficient |
| 71.95856 | 68 | 38.73058 | 2.859538 |

Handling outliers in ‘Sales Amount’

Using a z-score threshold of 3, we identified the presence of 75 outliers in the ‘Sales Amount’ column. Among these outliers, 70 are associated with the sleeping gear category, suggesting that they may represent higher-priced products rather than anomalies. However, a single outlier with a sales value of $940 in the travel category is identified as skewing the data significantly. To address this issue, we will selectively remove only this particular outlier to mitigate the impact on our analysis and retain the potential value represented by the sleeping gear category outliers. See **Exhibit 3** for the adjusted ‘Sale Amount’ Boxplot for all categories after removing the one identified outlier value.

Findings from the exploratory analysis

The analysis of sales data has unveiled several significant findings.

Distribution Characteristics:

* The distribution of sale amounts is positively skewed, indicating a prevalence of transactions with lower sale amounts and a scarcity of those with very high sale amounts.
* The mean sale amount stands at approximately $71.93, surpassing the median of $68.00. This suggests the existence of higher-value transactions that contribute to pulling the mean above the median.
* The standard deviation of sale amounts is approximately $38.74, signaling a moderate spread of transaction values around the mean.
* The skewness coefficient, calculated at 2.86, confirms the positive skew in the distribution of sale amounts.

Boxplot Analysis:

* See **Exhibit 1 and 2** for Boxplots
* A thorough examination of the boxplot identified a total of 380 outliers in the sale amount data.
* These outliers may be attributed to rare but legitimate high-value transactions.
* The recommendation is to exercise caution before removing these outliers, as they might represent valid sales data.

**PART 2**

The hypothesis test

Our hypothesis test looks to explore if there is a relevant variation in gross margin performance across departments. Our hypothesis test will evaluate the difference in mean gross margin for the category with highest ‘Sale Amount’ (Sleeping Gear) and the category with highest ‘Gross Margin’ (Backpacks)

**Null Hypothesis (H0):** The mean gross margin (GM$) for Backpacks is equal to the mean GM$ for Sleeping Gear. Null Hypothesis (H0): μ(Backpacks) =μ(Sleeping Gear)

**Alternative Hypothesis (H1):** The mean GM$ for Backpacks is significantly different from the mean GM$ for Sleeping Gear. Alternative Hypothesis (H1):μ(Backpacks)=μ(Sleeping Gear)

​

Hypothesis test results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| t-value | p-value | Lower bound | Upper bound | Mean  “Backpacks" | Mean  "Sleeping Gear" |
| -4.6425 | 3.629e-06 | -0.07207477 | -0.02926846 | 0.5335330 | 0.5842046 |

With such a low p-value, lower than a confidence level of 0.05, we would reject the null hypothesis. In this case, the null hypothesis is that there is no difference in means between "Backpacks" and "Sleeping Gear."

Based on the results, we can conclude that there is a statistically significant difference in gross margin between "Backpacks" and "Sleeping Gear." The negative sign of the confidence interval suggests that the mean gross margin for "Backpacks" is lower than that for "Sleeping Gear."

Some recommendations for them company would be to:

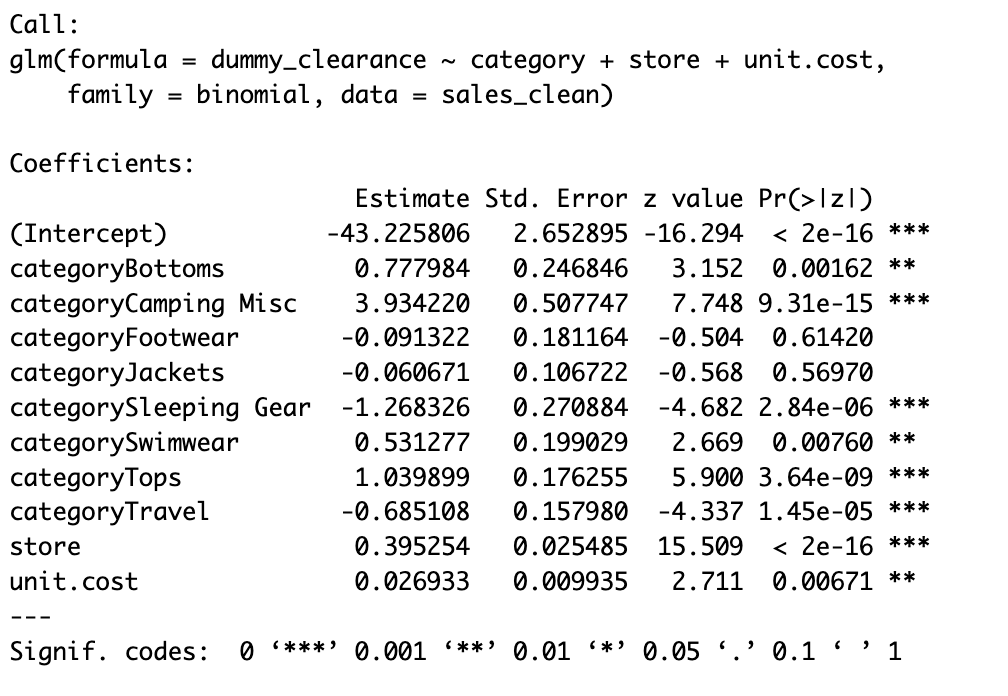
* Evaluate the pricing strategy for both "Backpacks" and "Sleeping Gear." If GRO wants to increase the profit margin for "Sleeping Gear," they may want to consider reevaluating the pricing of backpacks to make it more profitable.
* Examine the product mix and assortment within each category. Determine if there are specific products within "Sleeping Gear" that contribute significantly to its higher gross margin. Consider optimizing the product mix in "Backpacks" to include more profitable items.
* Conduct a detailed cost analysis for both categories. Evaluate if there are cost differences or operational efficiencies that contribute to the observed gross margin disparity. Identifying cost drivers can guide decisions on cost reduction.

The logistic regression model to predict the variable 'clearance'

We created a new binary variable called ‘Dummy\_Clearance’ based on the ‘Price.Category’ variable that returns the value 1 if a product was sold under clearance and 0 otherwise.

The model includes all product categories (bottoms, camping misc, footwear, jackets, sleeping gear, swimwear, tops, travel) with 'backpacks' as the reference category. Additionally, the unit cost and store variable is considered in the model to understand its impact on predicting whether a product falls under clearance.

Logistic regression results and model performance

****

The reference variable for our model was ‘Backpacks’. The low p-values (α < 0.05) suggest that the corresponding variables are statistically significant in predicting the clearance outcome. The coefficients for each category represent the change in probability of clearance for that category compared to the reference category. For example, for category ‘Bottoms’, the predicted change of an item going to clearance is higher by 0.7779 units compared to the ‘Backpacks’.

Items in these categories are more likely to go on clearance compared to ‘Backpacks’:

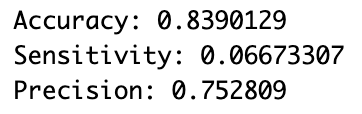
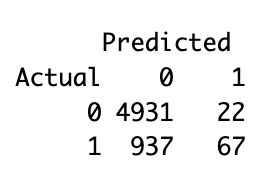
* "Camping Misc" (Coefficient: 3.9342)
* "Tops" (Coefficient: 1.0399)
* “Bottoms” (Coefficient: 0.7779)
* "Swimwear" (Coefficient: 0.5313)

Items in these categories are less likely to go on clearance compared to ‘Backpacks’:

* Jackets (Coefficient: -0.0607)
* Footwear (Coefficient: -0.0913)
* Travel (Coefficient: -0.6851)
* Sleeping Gear (Coefficient: -1.2683)

Analyzing the other predictor variables that are not ‘Category’. The coefficient for ‘Store’ is 0.3953, indicating that the store in which an item is located influences the likelihood of clearance. The positive coefficient suggests that items in certain stores are more likely to go on clearance. The coefficient for ‘Unit.Cost’ is 0.0269, indicating that the cost of the item influences the likelihood of clearance. The positive coefficient suggests that more expensive items are more likely to go on clearance.

To evaluate our model performance we constructed the confusion matrix and calculated the accuracy and sensitivity scores:



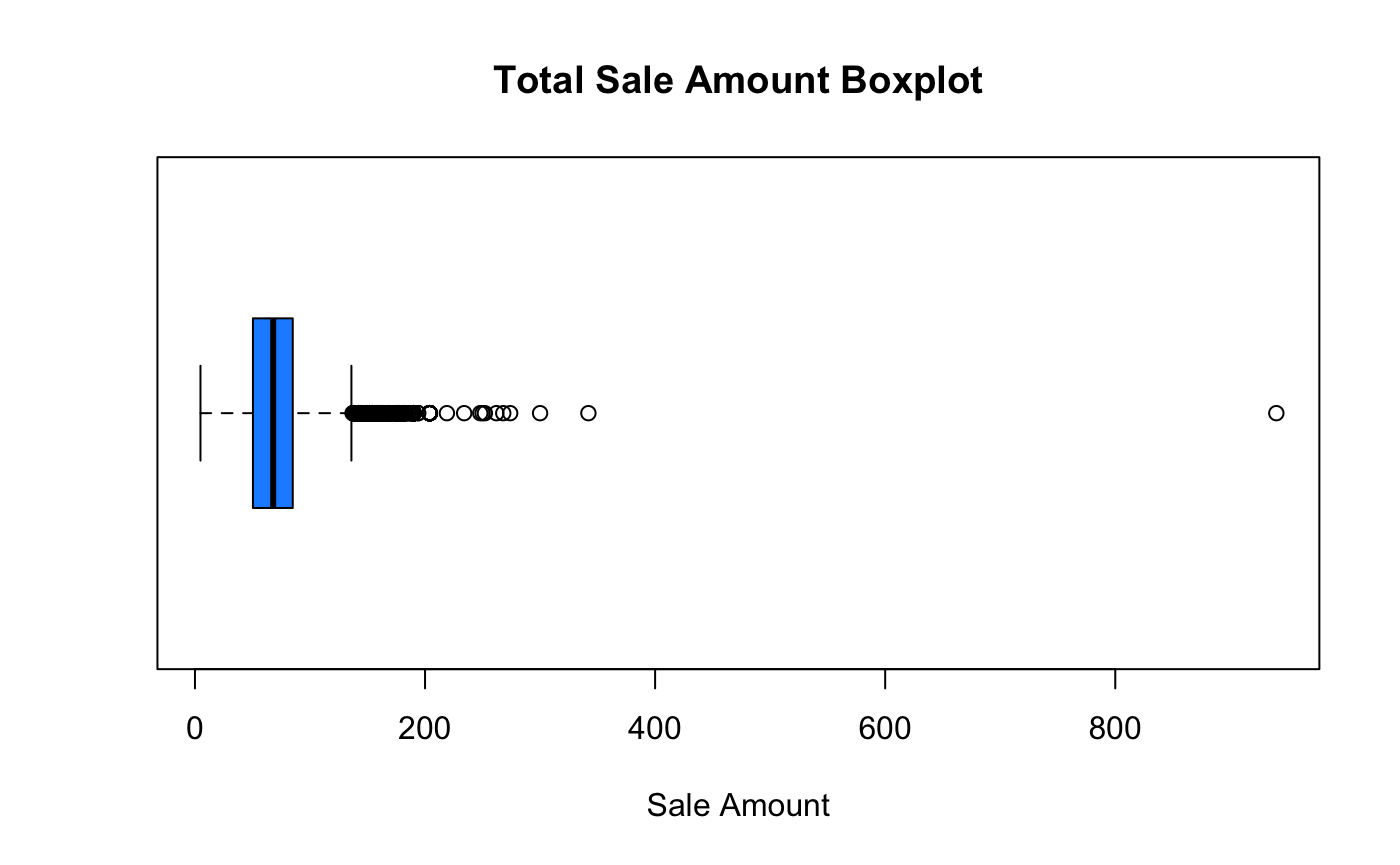
The overall accuracy of the model is 83.9%, which indicates the proportion of correctly classified instances (both 0s and 1s). However, accuracy on its own might not be able to assess the model’s ability to predict. The low sensitivity score of 6.67%, indicates that the model struggles to correctly identify positive instances. It is missing a significant portion of actual positive cases. Still, the precision of 75.28% indicates that, when the model predicts positive, it is accurate most of the time. Future analysis could focus on other variables that potentially impact an item going to sale or not to achieve a higher sensitivity.

Recommendations:

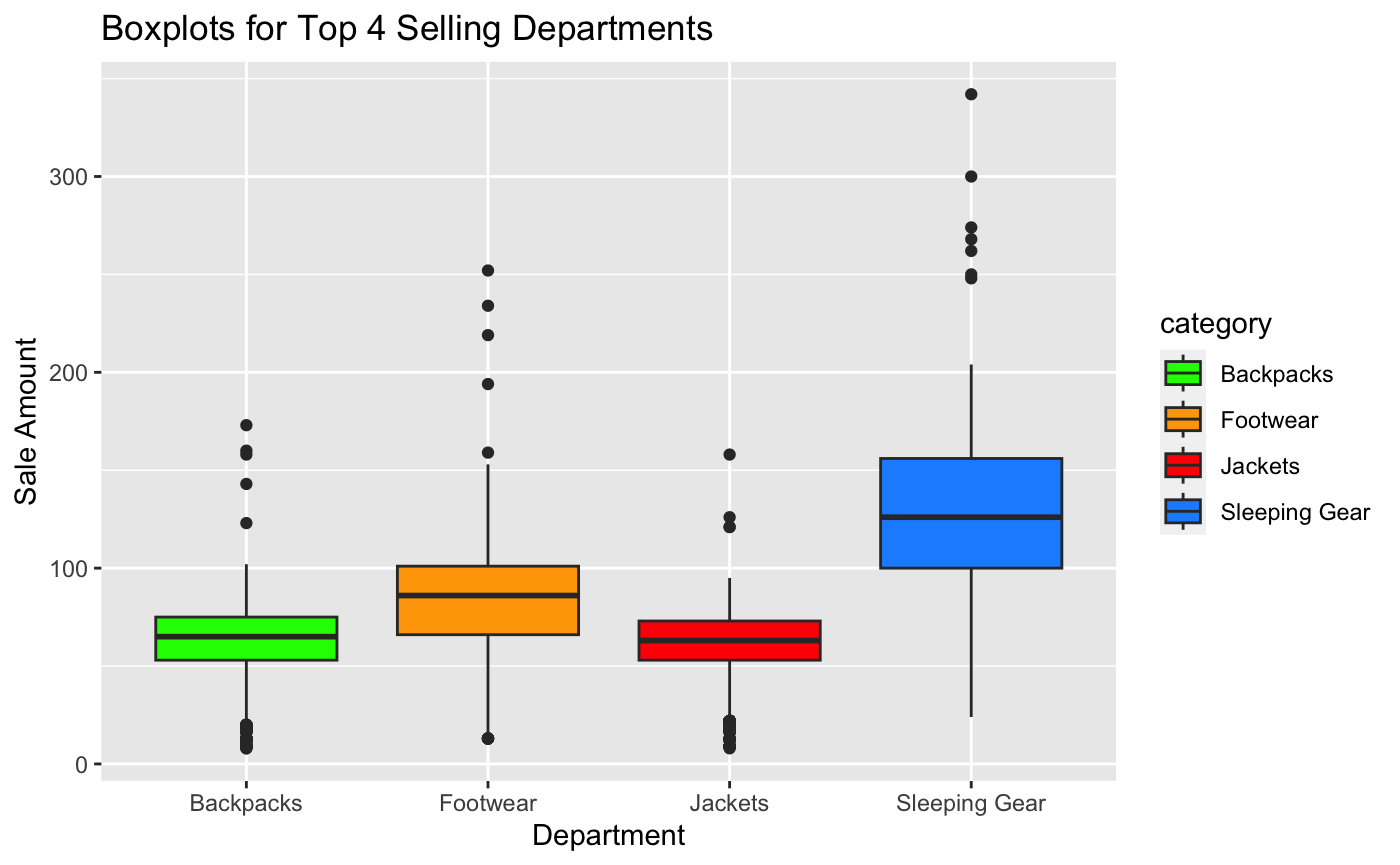
* Focus on understanding and managing clearance for categories with positive coefficients (e.g., "Camping Misc," "Tops," "Swimwear"). Try to figure out if there are specific trends within these categories that contribute to clearance.
* Consider further investigating the categories with significant negative effects
* Since the store variable and unit.cost also contribute significantly to the prediction of clearance, consider store-specific strategies for managing clearance based on the positive coefficient for the store variable.
* Evaluate whether certain stores have unique factors influencing clearance and adapt strategies accordingly.
* Explore discounting strategies for items at different price points.

**Exhibits**

**Exhibit 1** Boxplots for ‘Sales Amount’



**Exhibit 2** Boxplots for ‘Sales Amount for Top Selling Departments’



**Exhibit 3** Boxplots for ‘Sales Amount’ after Handling Outliers

