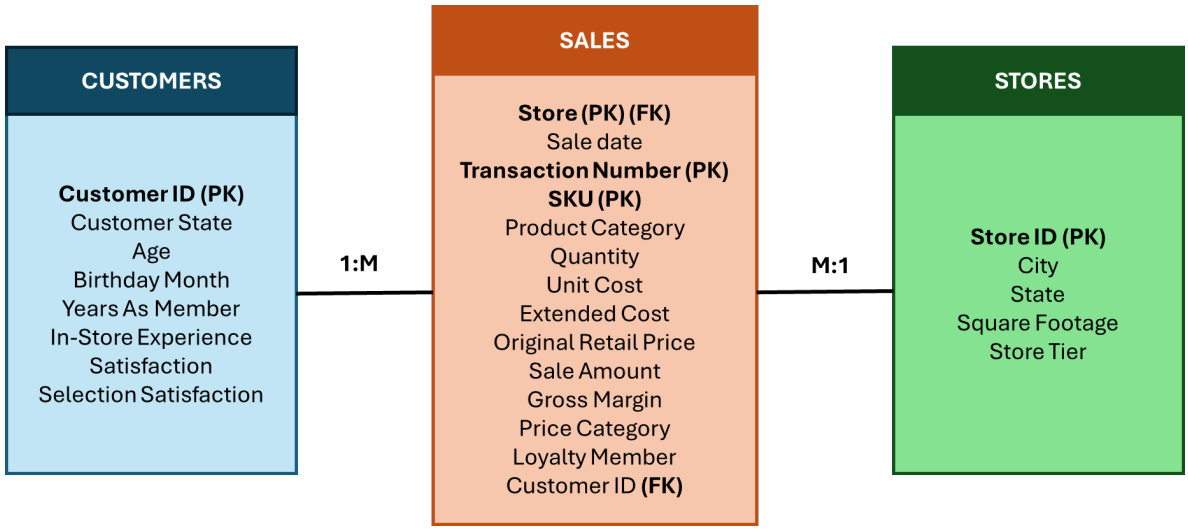
Analytics Problem Set Report

(1) Entity-Relationship Diagram (ERD)



The ERD Diagram has three entities, Customer, Sales and Stores. Sales entity will have a composite primary key made up of Store, Transaction number and SKU. It also has two foreign keys – (i) Store which is the primary key of Stores entity (ii) Customer ID which is the primary key of Customers entity.

(2) Data Cleaning Steps for 'customers' Data

The following cleaning steps were performed on the customers dataset using excel:

- Missing Values Check:Verified the row count for each column to identify any missing data. No missing values (empty cells) were found.
- The States column contained inconsistent categorical formats, e.g., variations like "CT," "Connecticut," and "Conn" or "MA" with four other variations. Used Excel's filter feature to identify and standardize these entries by changing all variations to the correct state abbreviations ("CT" and "MA").
- Cleaning Numerical Data (Age and Birthday Month Columns)
- The Birthday Month column had erroneous values such as "0" (acceptable range: 1-12).
- Identified "0" likely resulted from missing birthday data, affecting the Age column.
- Replaced missing birthday months with the most frequent value (mode), January.
- Replaced erroneous Age entries with the median age of 35.

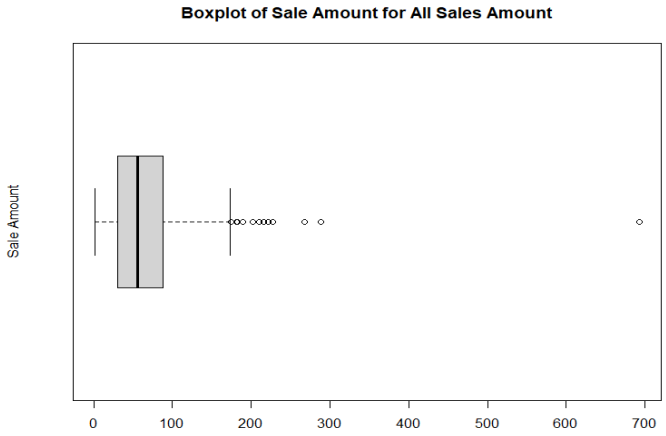
These steps ensured the dataset was consistent and ready for analysis.

(3) Summary Statistics and Boxplots

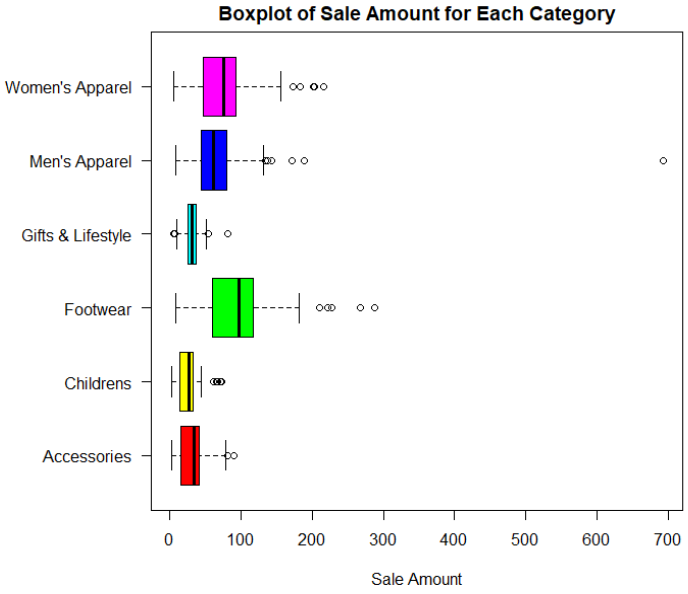
Sale Amount Summary Statistics	
Mean	60.599
Median	56.200
Standard Deviation	36.262
Skewness	1.006

Boxplots:

- **Sale Amount (All Records)**: A boxplot revealed significant outliers, particularly in the higher range of sale amounts. This is indicative of high-value transactions that may need further investigation or adjustments.



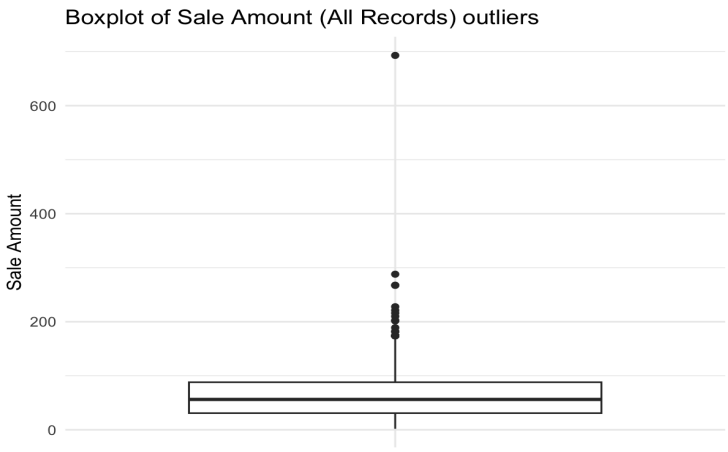
- **Sale Amount by Product Category**: The boxplot by category displayed noticeable variability across different product categories. Categories such as "Footwear" and "Children's Apparel" had higher median sale amounts compared to others.



Blended Gross Margin by Category:

Blended Gross Margin for each Category				
Sr. No	Category	Extended Cost	Sale Amount	Gross Margin
1	Accessories	$ 16,930.06	$ 45,033.53	0.624
2	Childrens	$ 10,036.94	$ 27,191.00	0.631
3	Footwear	$ 74,701.97	$ 204,102.96	0.634
4	Gifts & Lifestyle	$ 4,213.47	$ 9,967.50	0.577
5	Men's Apparel	$ 35,795.64	$ 103,846.60	0.655
6	Women's Apparel	$ 83,021.90	$ 226,274.55	0.633

(4) Presence of Outliers in Sale Amount



A box plot analysis revealed 21 outliers in the sales data, including one extreme case. Due to the data's positive skewness, the box plot method was used for its robustness. The value 174 (frequency: 8) was retained. Manual inspection showed most outliers occurred when the quantity sold exceeded 1 or items were sold at full price, indicating valid sales rather than errors.

(5) Hypothesis Test: ANOVA on Gross Margin by Price Category

The hypothesis test chosen was **ANOVA (Analysis of Variance)** to examine whether the **gross margin** varies significantly across different **price categories** (Full Price, Markdown, etc.). This test is essential for understanding how pricing strategies influence profitability, which is a critical business question for optimizing product pricing.

- **Null Hypothesis (H_0)**: There is no significant difference in gross margin across price categories.
- **Alternative Hypothesis (H_1)**: There is a significant difference in gross margin across price categories.

The ANOVA results showed a highly significant difference between price categories ($F = 4028$, $p < 2e\text{-}16$), suggesting that the pricing strategy does impact the gross margin.

Source of Variation	df	Sum of Squares (SS)	Mean Square (Mean Sq)	F-value	Pr(>F)
price.category	2	1,066.00	533.2	4028	<2e-16 ***
Residuals	10,169.00	1,346.00	0.1		

(6) Regression Model Summary

Regression Coefficients:

Predictor	Estimate	Std. Error	t Value	Significance Level
Intercept	1.118	0.054	20.760	***
Sale Amount	0.007	0.000	38.530	***
Price Category - Full Price	0.388	0.012	33.380	***
Price Category - Markdown	0.389	0.015	26.054	***
Quantity	-0.455	0.031	-14.650	***
Unit Cost	-0.024	0.001	-20.690	***
Store	-0.034	0.003	-10.592	***
Category - Childrens	-0.110	0.013	-8.216	***
Category - Footwear	0.112	0.028	4.003	***

Predictor	Estimate	Std. Error	t Value	Significance Level
Category - Gifts & Lifestyle	0.003	0.021	0.160	
Category - Mens Apparel	0.075	0.016	4.560	***
Category - Women's Apparel	0.036	0.020	1.840	
Loyalty Member	-0.009	0.007	-1.413	

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Key Findings from the Model:

The regression model indicates that the **sale amount**, **price category**, and **unit cost** are significant predictors of **gross margin**. The most noteworthy findings include:

- **Sale Amount**: Higher sales amounts lead to higher gross margins.
- **Price Category**: Products sold at **Full Price** or **Markdown** categories have higher gross margins than those in lower price categories.
- **Quantity Sold**: A negative relationship between quantity and gross margin indicates that higher-volume sales often have lower margins.

(7) Summary of Key Findings for Management

The analysis of customer and sales data revealed several important insights that can guide the company's strategies:

- **Pricing Strategy and Gross Margin**:
  - The **gross margin** is significantly impacted by the pricing category (Full Price, Markdown), and optimizing the pricing strategy can enhance profitability.
  - **High-value customers** (those with higher satisfaction and higher sales amounts) should be targeted for retention efforts, as they contribute more to revenue.
- **Purchase Behavior Insights**:
  - Customers with higher satisfaction ratings tend to purchase more, especially in Full Price or Markdown categories. This suggests that improving customer experience could directly increase sales.
- **Outlier Management**:
  - Extreme outliers in sale amounts may skew profitability analysis. To account for this, consider separating high-value transactions from regular ones, or applying strategies to handle outliers without losing valuable data.

In conclusion, the company should focus on **refining pricing strategies** to improve gross margins, **target high-value customers** through personalized marketing, and consider **handling outliers** in high-value transactions for more accurate financial planning.