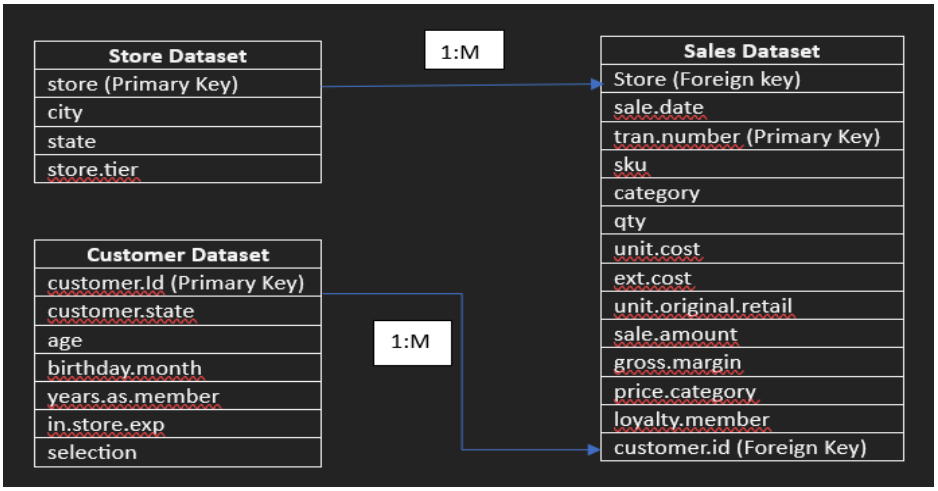
Analytics Problem Set

Part 1: Descriptive Analytics

I. Data Modeling: ERD diagram of Customer, Sales and Stores:



II. Inspection and cleaning:

Cleaning Steps Performed on the Customers Data file:

Customer Dataset Profiling Results:

1. The customer.state field contains inconsistent abbreviations for states such as "Mass.", "Massachusetts," and "Massachusets" for MA, as well as "Connecticut" and "Conn." for CT. These need to be standardized.
2. Empty or missing values in the selection satisfaction column are identified and need to be replaced with NA.
3. The birthday.month column contains invalid values, such as 0, which should be omitted since only 1-12 are acceptable.
4. A check on unique values in fields such as age, years.as.member, and in.store.exp reveals the need for validation and consistent formatting.

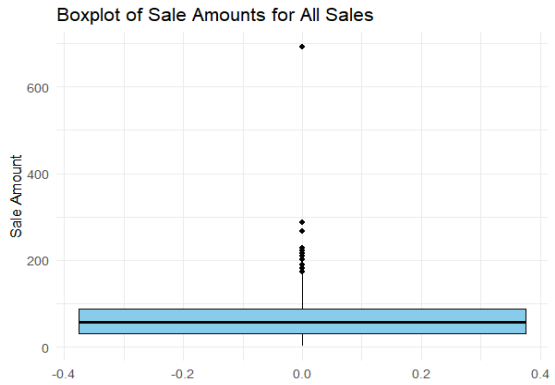
Customer Dataset Cleansing Results:

1. The customer.state column was cleaned by mapping all variations of state names and abbreviations to their correct two-letter format (e.g., "Massachusetts" → "MA", "Connecticut" → "CT").
2. Empty or missing values in the selection column were replaced with NA to signify the absence of information.
3. Invalid entries in the birthday.month column, such as 0, were removed, and inconsistent values (e.g., "March", "Mar.") were transformed into numerical equivalents (e.g., "3").
4. Rows with missing or invalid data in critical columns were excluded to enhance dataset integrity.

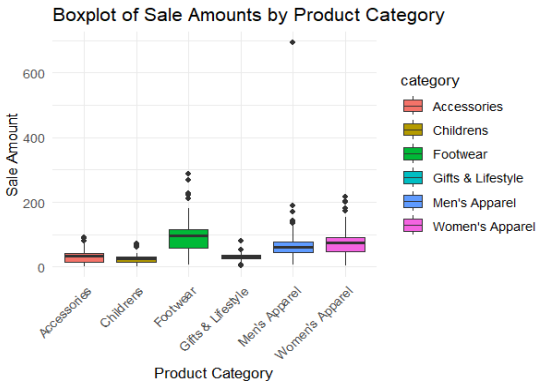
III. Summary statistics and visualization

Sale.amount Statistic	
Mean	60.5993059378687
Median	56.2
SD	36.2619973648122
Skewness	1.00587608786905

(a) Boxplot with Sale amount for all sales records in the data



(b) Boxplot with Sale amount separately for each product category

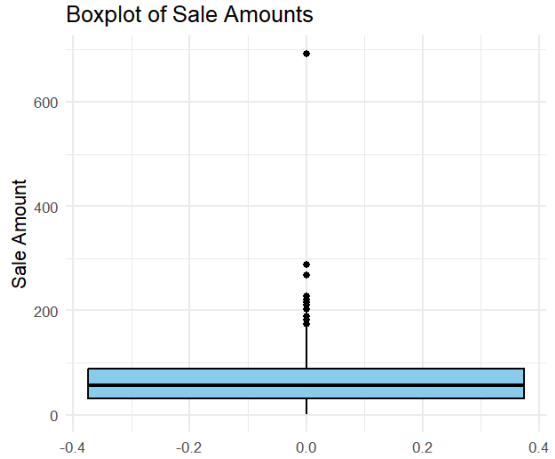


Blended gross margin for each product category:

	category	total_sale_amount	total_ext_cost	blended_gross_margin
1	Accessories	45033.53	16930.06	62.40566
2	Childrens	27191.00	10036.94	63.08727
3	Footwear	204102.96	74701.97	63.39986
4	Gifts & Lifestyle	9967.50	4213.47	57.72792
5	Men's Apparel	103846.60	35795.64	65.53027
6	Women's Apparel	226274.55	83021.90	63.30922

IV. Identifying Outliers in the Sale.amount variable using boxplot and z-score method:

Boxplot output:



Z-Score output:

25 entries out of the 10172 are the outliers.

Conclusion on the outliers:

The Z-score method works well for symmetric distributions where 25 outliers have been identified and these can be ignored from the available data. The 25 outliers in the "sale.amount" variable are errors and don't significantly affect the analysis, we can eliminate them or limit the extreme numbers. To reduce their influence on the outcomes, we may use robust models or data transformations.

Part 2: Statistical Inference

V. Hypothesis Testing Statements:

Null Hypothesis (H_0):
There is no significant difference in the gross margin percentage (GM%) across the four seasons (Winter, Spring, Summer, Fall).
Alternative Hypothesis (Ha):
There is a significant difference in the gross margin percentage (GM%) across the four seasons.
Results:
The conducted ANOVA tested and examined whether there are significant differences in gross margin percentage (GM%) across the four seasons: Winter, Spring, Summer, and Fall. This hypothesis test

helps address the business problem of understanding how seasonality affects GM%, as inconsistent performance in gross margin could be linked to seasonal variations in sales. By determining if GM% varies across seasons, the company can identify potential areas for improvement, adjust pricing or discount strategies, and better forecast future performance.

List of other Hypothesis tests that can be performed and can give useful insights for the Business:
Test for GM% across seasons: To understand the impact of seasonality on GM%.
Test for GM% across stores: To assess performance variations across stores.
Test for GM% across price categories: To evaluate how price category affects GM%.
Multiple regression analysis: To understand the relationship between GM% and GM$ across stores.
Time-series analysis or regression with seasonality: For improved GM$ forecasting based on seasonal patterns.
By conducting these tests, the retailer can gain valuable insights into the root causes of inconsistent GM% and GM$, as well as identify strategies for improving forecasting and financial performance across different stores and seasons.

VI. Model Overview:

With a significant positive coefficient, the model demonstrates that “sale.amount” has a favorable impact on the gross margin. Its negative coefficient indicates that ext.cost have a negative effect on gross margin. When compared to other product categories, footwear, men's apparel, and women's apparel have lower gross margins. The non-significant coefficient demonstrates that loyalty membership has a minimal effect on gross margin, although both full price and markdown pricing schemes have a beneficial influence. All things considered, the model accounts for 72.75% of the variation in gross margin, emphasizing the significance of pricing and sales volumes.

Co-efficient name	Estimate
(Intercept)	-0.245744228
sale.amount	0.009805729
ext.cost	-0.02632864
categoryChildrens	-0.389687076
categoryFootwear	-0.056901347
categoryGifts & Lifestyle	-0.163847256
categoryMen's Apparel	0.052797733
categoryWomen's Apparel	-0.141870853
price.categoryFull Price	1.045229317
price.categoryMarkdown	0.913819144
loyalty.member	-0.004388389
interaction_terms_priceprice.categoryClearance	0.02292657
interaction_terms_priceprice.categoryFull Price	-0.001674343
interaction_terms_categorycategoryAccessories	-0.001718304
interaction_terms_categorycategoryChildrens	0.010468004
interaction_terms_categorycategoryFootwear	-0.000959357
interaction_terms_categorycategoryGifts & Lifestyle	0.001961498
interaction_terms_categorycategoryMen's Apparel	-0.002564582

The model explains approximately 72.75% of the variance in the dependent variable (gross.margin), as indicated by the Multiple R-squared value. The Adjusted R-squared value of 72.70% suggests that the model generalizes well to new data, given the number of predictors included.

The F-statistic is highly significant (p-value < 2.2e-16), indicating that the model as a whole is statistically significant.

VII. Findings to the Management Team:

Regression analysis demonstrates that price category, ext costs, and sale amount have a considerable impact on gross margin. Margin is increased by higher revenue and decreased by higher external expenditures. Discounted and full-priced goods have larger profit margins than clearance goods. Different categories have different margins; women's clothing and footwear have lower margins. Additionally, different product categories and prices have different correlations between gross margin and sale quantity, as indicated by interaction terms.

Actionable Insights:

1.Revenue vs. Cost Dynamics: High external costs reduce gross margins, while higher revenue improves them. Focus on reducing costs while maintaining or increasing revenue.

2.Leverage Loyalty Members: Loyalty members show minor margin impact. Boost profitability with tailored incentives to encourage high-margin purchases.

3.Price-Category Interaction: Gross margins vary by product category and price point. Use targeted pricing strategies, such as premium pricing for high-demand items and markdowns for slower-moving ones.

The hypothesis test rejects the null hypothesis (H_0) by showing significant variations in gross margin percentage (GM%) between seasons. Additional evidence from regression analysis shows that product categories, pricing policies, and sales volumes all have a significant impact on gross margins. Interaction terms demonstrate that the influence of these factors differs according to pricing and category.

Recommendations:

1.Seasonal Strategy Alignment: To maximize profitability during failing seasons, adjust inventory and promotions for high-margin products.

2.Category-Specific Focus: To increase margins, improve pricing and cost control for less lucrative categories such as women's apparel and footwear.

3.Dynamic Pricing: Target categories and seasons with the biggest potential impact and use markdown tactics sparingly to increase sales without reducing margins.