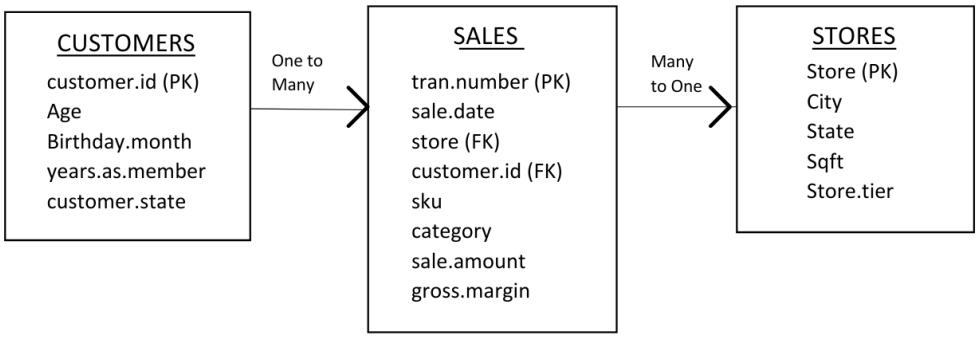
Part A

(1) Entity Relationship Diagram



(2) The steps taken in order to clean the ***'customers'*** data file on excel for better analysis is listed below:

- **Customer.state** – As per the data dictionary, modified the state names to two-letter U.S. state abbreviation codes.
- **Age** – Since the age variable cannot be 0, the fields were modified by calculating the median age and replacing the values where 0 was mentioned with the median.
- **Birthday.month** – The fields with 0 mentioned was replaced with the mode of the column.
- **In.store.experience & Selection** – The blank columns in these two fields were replaced by the following 'N/A' for better understanding of the data.

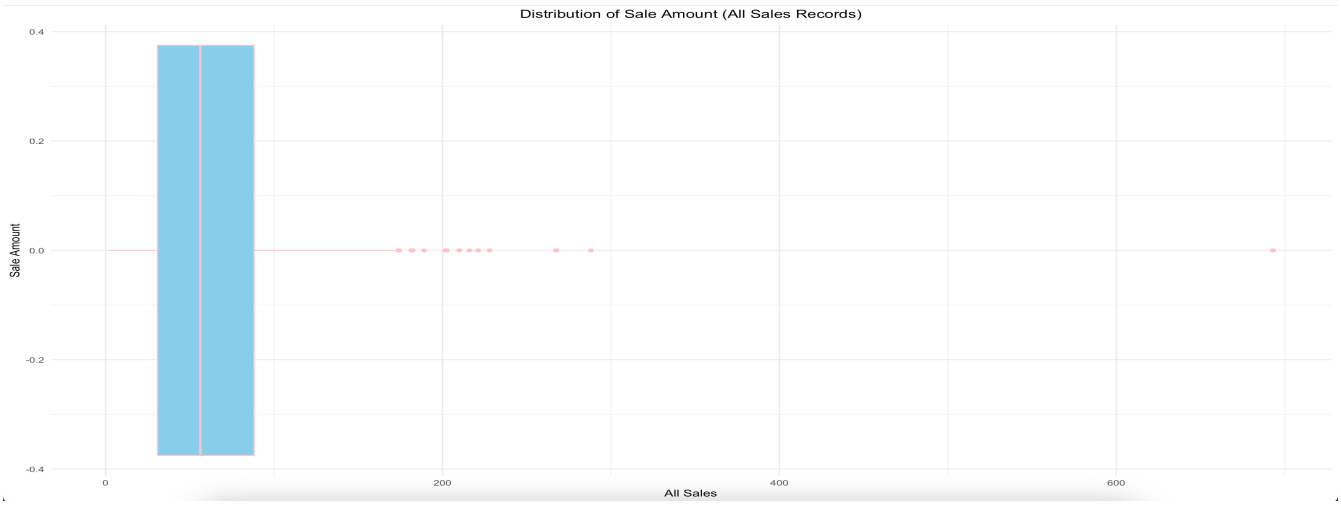
(3) The summary statistics and boxplots created, including a well-formatted table showing blended gross margin by category is given below :

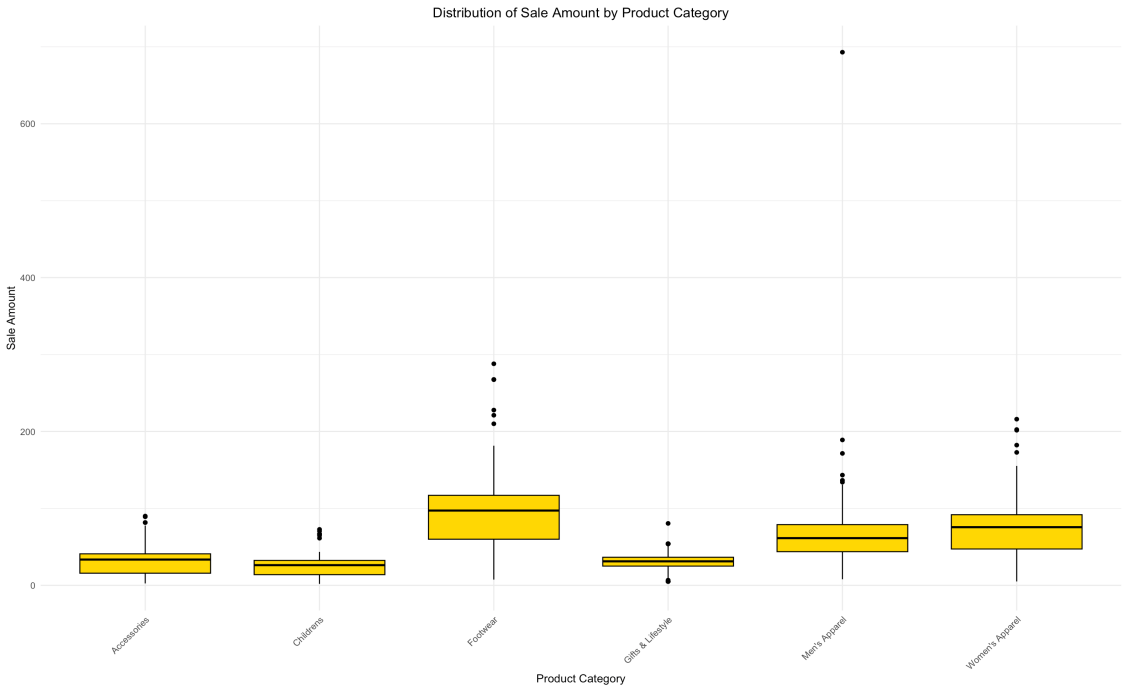
- **Summary Statistics**

  The mean, median, standard deviation was found in R-studio and the code used is mentioned below and the skewness was found using excel formula

Skewness	1.01
mean	60.60
median	56.2
standard deviation	36.26

- **Boxplots created**





category	total_sale_amount	total_ext_cost	blended_gross_margin
Accessories	45033.53	16930.06	62.40565641
Children	27191	10036.94	63.08727152
Footwear	204102.96	74701.97	63.39985956
Gifts & Lifestyle	9967.5	4213.47	57.72791573
Men's Apparel	103846.6	35795.64	65.53027254
Women's Apparel	226274.55	83021.9	63.30921882

(4) **Detection of outliers**

There are ***21 outliers*** present in the '***sale.amount***' variable in the sales data file. In order to handle the outliers we can either remove them in case they do not have a significant impact in the analysis or we can use the ***clustering/binning*** method and isolate the outliers. We can also use the ***imputation technique*** and replace the outliers with estimated values (e.g. Mean, median).

Part B

(5) **Hypothesis Testing**

The hypothesis test chosen is a ***one sampled t-test*** (two-tailed test) with a ***level of significance of 0.05***.

- ➔ **Null Hypothesis (H_0): $\mu=0$** (The mean gross margin percentage (GM%) of footwear is equal to 0)
- ➔ **Alternative Hypothesis (H_a): $\mu\neq0$** (The mean gross margin percentage (GM%) of footwear is not equal to 0)

The aim of the hypothesis is to determine whether the mean gross margin percentage (GM%) of the "Footwear" category differs significantly from 0, providing insights into its profitability. Basis the hypothesis test results (t = 64.624, df = 2246, p-value = 0.0000000000000002), ***we reject the null hypothesis (H_0)*** and conclude that the mean GM% for "Footwear" is significantly different from 0. The 95% confidence interval for the mean GM% is between 0.5027 and 0.5342, with a sample mean of 0.5185. This result confirms that the "Footwear" category is generating a positive gross margin, highlighting its profitability.

Statistic	Value
t-value	64.624
Degrees of Freedom (df)	2246
p-value	0.0000000000000002
Alternative Hypothesis	True mean is not equal to 0
95% Confidence Interval	[0.5027487, 0.5342153]
Sample Mean	0.518482

(6) **Regression analysis**

- Summary

Residuals Summary	
Statistic	**Value**
Min	-3.5599
1Q	-0.0377
Median	0.0073
3Q	0.0809
Max	0.827

Coefficients Summary			
Variable	**Estimate**	**Std. Error**	**Pr(>|t|)**
(Intercept)	0.2465784	0.0283766	0.0000000000000002
qty	-0.6025244	0.0245375	0.0000000000000002
price.categoryFull Price	1.1132235	0.0127559	0.0000000000000002
price.categoryMarkdown	0.9899426	0.0217271	0.0000000000000002
sale.amount	0.0334303	0.0003522	0.0000000000000002
unit.cost	-0.0219688	0.0012452	0.0000000000000002
unit.original.retail	-0.0012483	0.0003168	0.0000915
price.categoryFull Price:sale.amount (Interaction variable)	-0.0261442	0.0003198	0.0000000000000002
price.categoryMarkdown:sale.amount (Interaction variable)	-0.0244473	0.0004127	0.0000000000000002

Model Summary	
Metric	**Value**
Residual standard error	0.2612
Multiple R-squared	0.7127
Adjusted R-squared	0.7124
F-statistic	2801
p-value	0.0000000000000002

The regression model explains ***71.27%*** of the ***variability in gross margin***, as indicated by the R-squared value, making it a strong fit for the data. The key findings from the model are as follows - ***sale.amount*** has a positive and significant impact on gross margin, with higher sales contributing strongly to profitability. The impact of Selling at ***full price*** has greater positive effect, while ***markdown*** sales also improve gross margin but at a

slightly lesser extent. The variables ***qty*** and ***unit.cost*** have significant negative effects which indicates that higher quantities and costs reduce profitability.
The interaction terms ***price.category:sale.amount*** reveal diminishing returns on gross margin as sales volume increases, suggesting a need for optimizing pricing strategies. In conclusion, the model is robust and provides actionable insights into pricing and sales strategies to improve gross margins.

(7) **Findings from Exploratory Analysis**

- ***Gross Margin Differences Across Categories*** - Significant differences in GM% across product categories ($p < 0.05$) highlight varying profitability. Some categories outperform like – Footwear, Women's apparel and men's apparel, while others underperform like – gift and lifestyle, children and accessories, requiring focused analysis.
- ***Pricing Strategy Insights*** - Full-price sales significantly boost gross margin (Estimate = 1.1132, 0.0000000000000002), while markdowns contribute less (Estimate = 0.9899, 0.0000000000000002). Reducing markdown reliance is key to improving profitability as it will only boost revenue temporarily, relying heavily on them will affect the profitability over time due to reduced per-unit margins.
- ***Operational Challenges*** - Bulk sales (Estimate = -0.6025, 0.0000000000000002) and higher unit costs (Estimate = -0.0219, 0.0000000000000002) reduce gross margin, indicating inefficiencies in pricing and cost management.

Findings from Hypothesis Test and Regression Analysis

- ***Hypothesis Test*** **-** The test confirmed significant differences in GM% among product categories, emphasizing the need to address underperforming segments.
- ***Regression Analysis*** **-** The regression model explains 71.27% of the variability in gross margin, identifying key drivers like sale amount (positive impact) and quantity sold (negative impact), along with the influence of pricing strategies.
- ***Operational Insights*** - Diminishing returns on gross margin with increasing sales volumes, especially in markdown categories, highlight the need to balance profitability and volume.

Actionable Recommendations

- ***Optimize Pricing Strategies*** – Focus on selling more products at full price rather than relying heavily on markdowns. Use discounts only when absolutely necessary to clear out excess stock while still protecting profits. Adjust pricing based on customer demand, what competitors are doing, and seasonal trends to balance making money with selling more.
- ***Focus on Underperforming Categories*** Take a closer look at categories that aren't doing well. Figure out if the problem is too many discounts, slow-moving stock, or products that don't match what customers want. Use tools like customer surveys or sales data to understand why these categories aren't popular and tweak pricing, product selection, and promotions to make them more attractive.
- ***Refine Bulk Sales and Cost Management*** – Review how discounts for large purchases are being offered. Instead of focusing only on volume, make sure bulk sales still contribute to profits. Look for ways to cut costs, such as negotiating better deals with suppliers or streamlining operations, to make selling in bulk more worthwhile.

The above mentioned recommendations can help the company improve gross margins and address key profitability challenges.