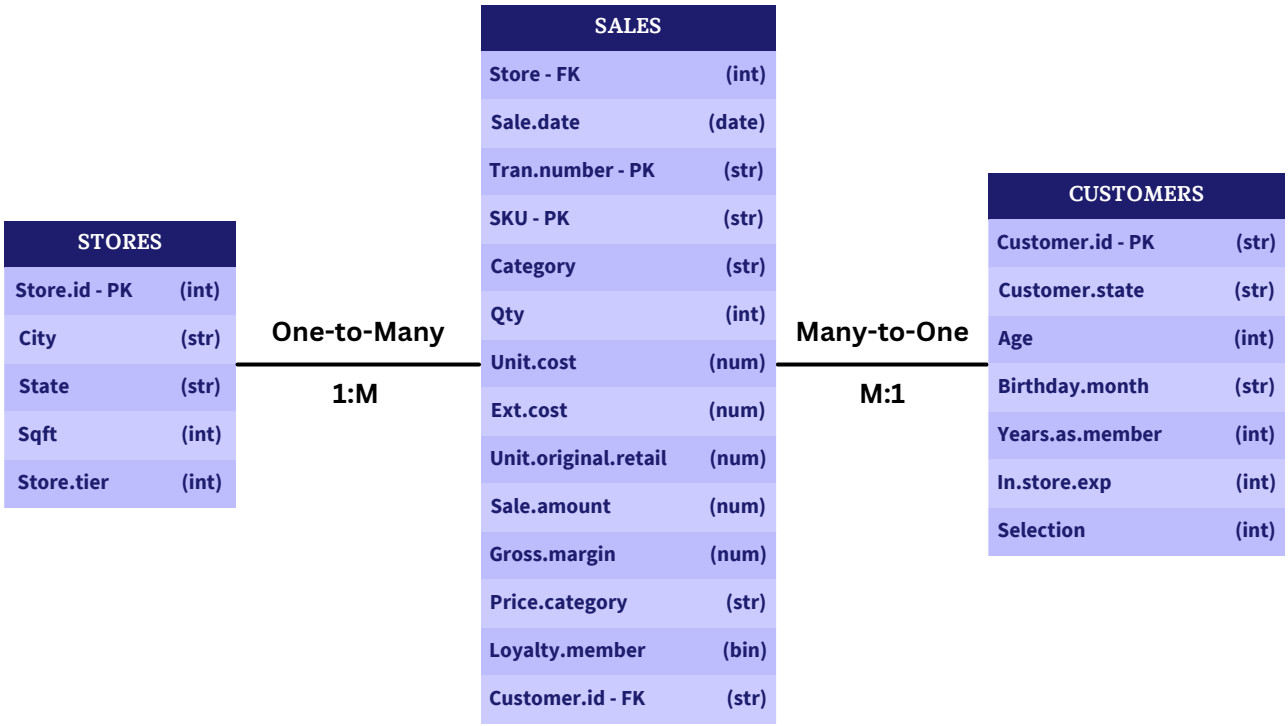
Analytics Problem Set by Group 2

(1) Entity Relationship Diagram



(2) Data Cleaning Process for 'customers' data file:

- Standardized naming convention for 'customer.state' column to adhere to data integrity:
  - First, converted all values to upper case
  - Then, replaced all redundant values to its appropriate two-letter U.S. state abbreviation
- Converted 'birthday.month' values to numerical format:
  - First, created a month mapping dictionary to assign each month (also its short-hands) with its corresponding numerical value
  - Then, utilized lambda function to apply these numerical values to each record, matching values from 1-12. Any records with 0 values were replaced with the mode of 'birth.month'
  - Column data type was then converted to string to follow data requirement
- Replaced 0 values in column 'age' with median value (since there were outliers in the dataset)

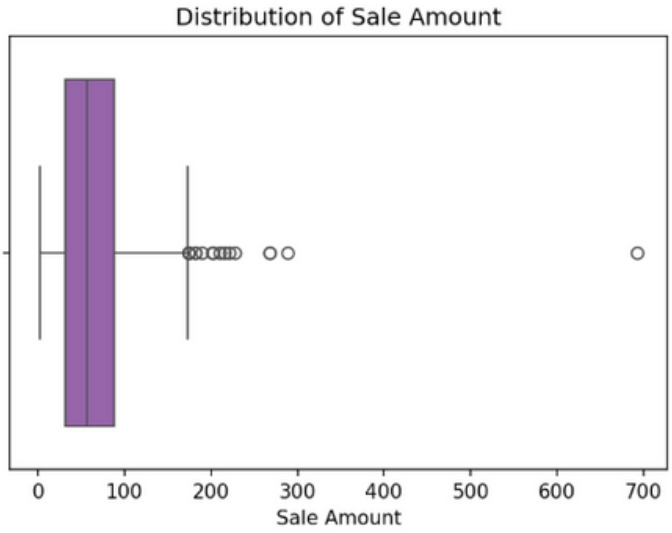
(3) Summary Statistics of 'sale.amount'

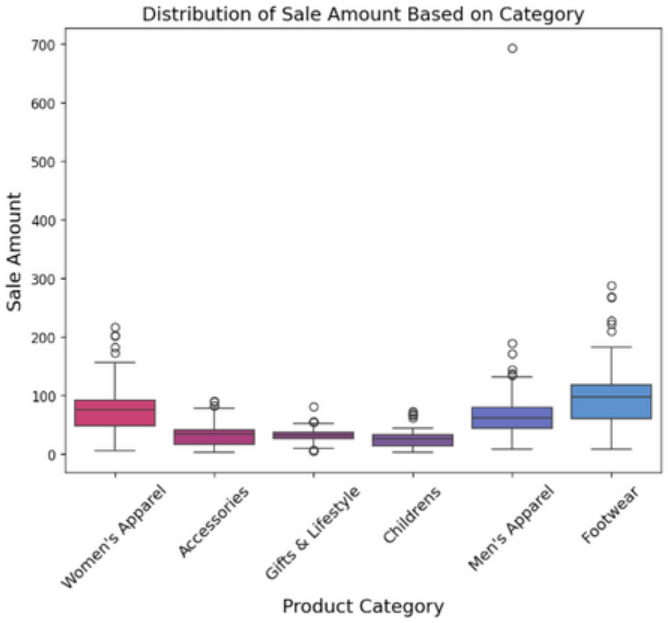
Count	10172
Mean	60.599306
Std	36.261997
Min	1.870000
25%	30.830000
50%	56.200000
75%	88.035000
Max	693.000000
Skew	1.005876

Blended gross margin by category

Category	Total Sale Amount	Total Ext Cost	Blended Gross Margin
Accessories	45033.53	16930.06	0.624057
Children	27191.00	10036.94	0.630873
Footwear	204102.96	74701.97	0.633999
Gifts & Lifestyle	9967.50	4213.47	0.577279
Men's Apparel	103846.60	35795.64	0.655303
Women's Apparel	226274.55	83021.90	0.633092

Boxplot of 'sale.amount' distribution (Overall & Product Category):





(4) Presence of Outliers & Recommendations:

- By using the z-score method, we noticed 25 observations identified as outliers (z-score > 3)

With less than 1% of outliers in the sales dataframe, handling them can be done as follows:

- The extreme value in 'sale.amount' (693) with a z-score of 17.4 can be capped as upper bound level to maintain data integrity given its bulk purchases nature. Other outliers (z-scores of 3 to 6) in categories like "Footwear," "Men's Apparel," and "Women's Apparel" (likely due to higher-priced products), should be capped at the upper bound to preserve data integrity as well.

(5) Hypothesis Testing:

H0	**μsale.amount_Winter-μsale.amount_Summer = 0** The average sale amount per unit in Winter equals to that in Summer.
Ha	**μsale.amount_Winter-μsale.amount_Summer =/= 0** The average sale amount per unit in Winter differs to that in Summer.

	Winter	Summer
Mean	58.28763402	63.96376489
Variance	0.097074282	0.20740798
Observations	2798	3190
df	5986	
T-Critical Value	1.645108	
T-Statistic	-6.178561279	
P(T<=t) two-tail	6.89494E-10	

Winter: October, November, December
Summer: May, June, July

t-Test: Two-Sample Assuming Unequal Variances

Level of significance: alpha = 0.05

- Since the p-value < alpha (0.05), we reject H0. There is sufficient evidence to conclude that the mean sale amount per unit of Winter season (October to December) significantly differs from that in Summer season (May to July). Understanding that there is seasonality effect in the overall per-unit sale amount can guide forecasting and pricing strategies to optimize gross margin during different periods of the year. Further analysis is needed to better understand if discounts are effective during Winter/Holiday seasons or if excessive markdowns/clearance leads to degradation in GM%, directly addressing the business challenges in understanding GM% inconsistency. The business should analyze whether the Winter season is underperforming (with lower GM$ and GM%) and plan to optimize the discounting strategies, inventory management, or product mix during that period to maintain profitability.

(6) Summary of Regression Model

R-squared	0.174
Adj. R-squared	0.173

	coef	std err	p-value
intercept	0.4295	0.041	0.000
loyalty.member	-0.0145	0.009	0.100
ext.cost	-0.0184	0.001	0.000
qty	0.3100	0.046	0.000
category_Accessories	-0.1182	0.027	0.000
category_Gifts & Lifestyle	-0.0644	0.034	0.062
category_Childrens	-0.2443	0.030	0.000
category_Men's Apparel	0.0546	0.015	0.000
category_Footwear	0.1333	0.017	0.000
store.tier_2	-0.0883	0.017	0.000
store.tier_3	-0.0262	0.030	0.382
high_price_items_store2_int	0.0437	0.020	0.032
high_price_items_store3_int	0.0710	0.035	0.040
discount_amount_qty_interaction	0.0068	0.000	0.000
seasonality_Holiday_discount_int	0.0864	0.012	0.000

Key findings from the Regression Model:

The regression model highlights several significant variables impacting gross margin (GM%) with an R2 of 0.174. Key predictors include quantity (qty) and seasonality_Holiday_discount_interaction, both of which have positive coefficients (0.3100 and 0.0864), indicating positive contribution to GM%. Conversely, ext.cost (-0.0184), store.tier_2 (-0.0883), and product categories such as Children's (-0.2443), Accessories (-0.1182), and Gifts & Lifestyle (-0.0644) negatively affect GM, with significant p-values. Men's Apparel and Footwear categories contribute positively to GM with 0.0546 and 0.1333 units higher than Women's Apparel, holding all other factors constant. Interaction terms, like high_price_items_store2_int, reveal positive effects of high-price items on GM%, depending on the store tier. The model underscores the critical role of both product categories and store characteristics in determining GM%. Statistically insignificant variables like loyalty.member suggests further analysis on loyalty membership program's impact on GM% is needed to identify areas of improvement (e.g: should the business offer member-exclusive discounts to entice higher sale volume?). These findings can inform targeted strategies for improving margins by focusing on product categories and store tiers.

***Note**: New variables (not included in dataset) explanation:*

- *Discount_amount: developed hypothetically based on unit.original.retail variable for Full-Price items with a random drawing of discount percentage from 5-75% (typical discount value found on the dataset). This was added to analyze the business's site-wide discounting strategy (not specific to item level)*
- *Discount: a binary variable indicating whether or not a customer would use coupons on a Full-Price item.*
- *seasonality_Holiday: a binary variable indicating whether transactions occurred in October, November, December*
- *high_price_items: a binary variable indicating whether the item is in Women's/Men's Apparel or Footwear*

(7) Synthesis of Insights for Company's Management Team:

- **Summary Findings from Exploratory Analysis:**
  - Upon analysis, Footwear and Men's Apparel categories have the highest blended gross margins (0.64 to 0.65), indicating strong profitability. However, categories like Children's and Accessories experience frequent markdowns/clearance, which inflate the difference between extended cost and sale amount, leading to negative gross margins. This indicates that high discounting or unsuccessful product offerings can severely affect margins. Although Women's Apparel also faces markdowns, it maintains high GM% (0.63) due to strong sales volume, helping offset lower margins. Gifts & Lifestyle products, while underperforming throughout the year, reach higher sales during the holiday season.

  - The store performance analysis reveals that Store 12 (Tier 1) achieved the highest GM$ of $184,162.37 and the second-highest GM% (0.64), supported by its location in Watertown, MA, and a large customer base. Store 13 (Tier 3) achieved the highest GM% (0.67) despite its lower GM$ compared to Store 12. This can be attributed to the low-cost product offerings and fewer items being sold at markdowns (65 items compared to 355 in Store 14). Store 14 frequently have items on clearance/markdown, further deteriorating its overall profitability.

- From the sample data provided, loyalty members have lower total sum of sale amount ($292,695.42) than non-members ($323,720.72), suggesting a more nuanced analysis needed to assess the influence of loyalty program on the business's GM$ & GM% and further optimize customer retention strategies.

- **Summary Findings of Hypothesis Test & Regression Model**

  - Our hypothesis test shows that there is a difference between sale amount in Winter and Summer season. However, further analysis is needed to assess whether or not holiday seasons have positive impacts on the business's GM% due to increased customer demand. On the other hand, the model reveals that discounts can erode margins, but high sales volumes can offset margin loss through economies of scale. The interaction between discount amount and quantity sold shows a positive coefficient (0.0068), indicating that large volumes of full-priced products at appropriate discounts can still drive profitability. The seasonality interaction term (seasonality_Holiday_discount_interaction) indicates that holiday promotions (October to December) improve gross margin if discounts being offered, but monthly sample sales data shows that holiday season performance (0.60 GM%) lags behind summer months (0.67 GM%). This is likely due to increased markdowns and clearance items during the holiday season, which diluted overall profitability despite the seasonal effect. On the other hand, high-priced products in Tier 2 and 3 stores, such as those in Women's Apparel, Men's Apparel, and Footwear, show positive influences on GM% (coefficients of 0.0437 and 0.0710), suggesting that premium products perform well in these store tiers, yet the sample data does not reflect high stocking for these items.

- **Business Recommendations:**
  - A detailed analysis should be conducted on the underperforming categories (Children's, Accessories, Gifts & Lifestyle) to identify potential causes, such as overstocking, product issues, changing market/seasonal trends, or inaccurate demand forecasting to take preemptive measures early, further reducing its impacts on the overall margin. After that, the business should consider reducing stock in categories with consistently low GM%, especially those with high markdowns (Children's and Accessories). The retailer may need to adopt higher-quality products that better meet customer demands. For slow-moving items, the company could implement targeted promotions or bundling strategies that pair low-margin items with higher-margin products (e.g.: Women's Apparel can go with Accessories at exclusive deals), instead of applying blanket discounts to improve overall profitability.

  - For stores with low gross margins (e.g., Store 14), consider adjusting the product mix, offering more premium products, or optimizing inventory levels. Training staff, improving store-level operations, and enhancing local marketing strategies can also help boost sales in this tier. Tier 1 stores can justify premium pricing for higher-margin categories, while Tier 2 and Tier 3 stores should focus on more competitive pricing for lower-margin categories while diversifying its high-margin product portfolio. More marketing initiatives, such as running geo-targeted ads or store-exclusive promotions on high-priced items, can be developed to drive local customers & encourage higher sales volume.

  - The company should focus on targeted promotions and volume-based discounting that encourage higher purchases without significantly reducing per-unit profitability. For example, the stores can deploy "Buy X Get Y" or tiered promotions (Spend $X to get $Y on next order) on high-priced items to increase sales volume and GM$. While holiday sales boost sales volumes, they also increase markdowns, which negatively impacts gross margin. The company should consider balancing full-price and markdown/clearance products during the holidays. Each store can run a Flash 24-Hour Sale a few days/weeks before peak holiday sales to clear out inventory while ensuring that these items do not influence the sales of other high gross margin items. The company should also ensure that full-price items are prioritized during high-demand months (e.g., June and July) to preserve annual gross margins.