MA678 Final Project Report

Jinran Li

2024-12-10

Abstract

This report investigates the product sales and pricing strategies of Temu, the U.S. subsidiary of Pinduoduo, one of China's largest e-commerce platforms. The analysis uses a detailed sales data dataset, employing statistical methods and exploratory techniques to reveal patterns in consumer buying behavior and pricing dynamics. The study provides high-level insights into market trends, contributes to a deeper understanding of competitive strategies in the e-commerce space, and offers actionable recommendations for optimizing product offerings and pricing.

Introduction

Temu is a fast-growing online marketplace that, as the U.S. subsidiary of China's Pinduoduo, provides a unique platform for analyzing e-commerce trends. By combining a competitive pricing strategy with a diverse product catalog, Temu has quickly captured the attention of consumers and competitors alike.

This report explores the underlying factors that drive Temu.com's product sales and pricing strategies. Using data sets that capture product performance and pricing details, the analysis identifies key trends and patterns in consumer behavior. These insights provide a foundation for understanding how pricing strategies affect market outcomes and highlight optimization opportunities in the e-commerce industry.

The findings of this study are intended to facilitate strategic decision-making in e-commerce and provide actionable recommendations for improving product offerings, increasing customer engagement, and enhancing market competitiveness.

Method

Data Processing

The dataset used in this study was downloaded from Kaggle and contains detailed information on product sales and pricing from Temu.com. It includes various features such as product names, categories, prices, sales volumes, ratings, and comments. Before conducting the analysis, several preprocessing steps were undertaken to ensure data quality and relevance.

Column Name	Description
Leve_1_category_name	The main category of the product (e.g., Electronics, Clothing).
Leve_2_category_name	A subcategory under `leve_1_category_name`, providing finer classification (e.g., Smartphones).
Sales_volume	Total units sold for a product, indicating its popularity.
Price	The product's selling price, influencing sales performance.
Comment_num	Number of customer comments or reviews, reflecting feedback and engagement.
Good Score	Product rating on a scale from 1 (lowest) to 5 (highest), representing customer satisfaction.

Data Cleaning

I first checked the dataset for missing values and found 14,119 missing entries in the good_score and comment_num columns and 1,709 incomplete values in the sales_volume column. A high incidence of NA values indicates potential barriers such as low engagement or lack of motivation to leave feedback, especially for customers with a neutral experience, so instead of removing the NA values for the ratings and comments, I made them new columns to mine later.

Exploratory Data Analysis

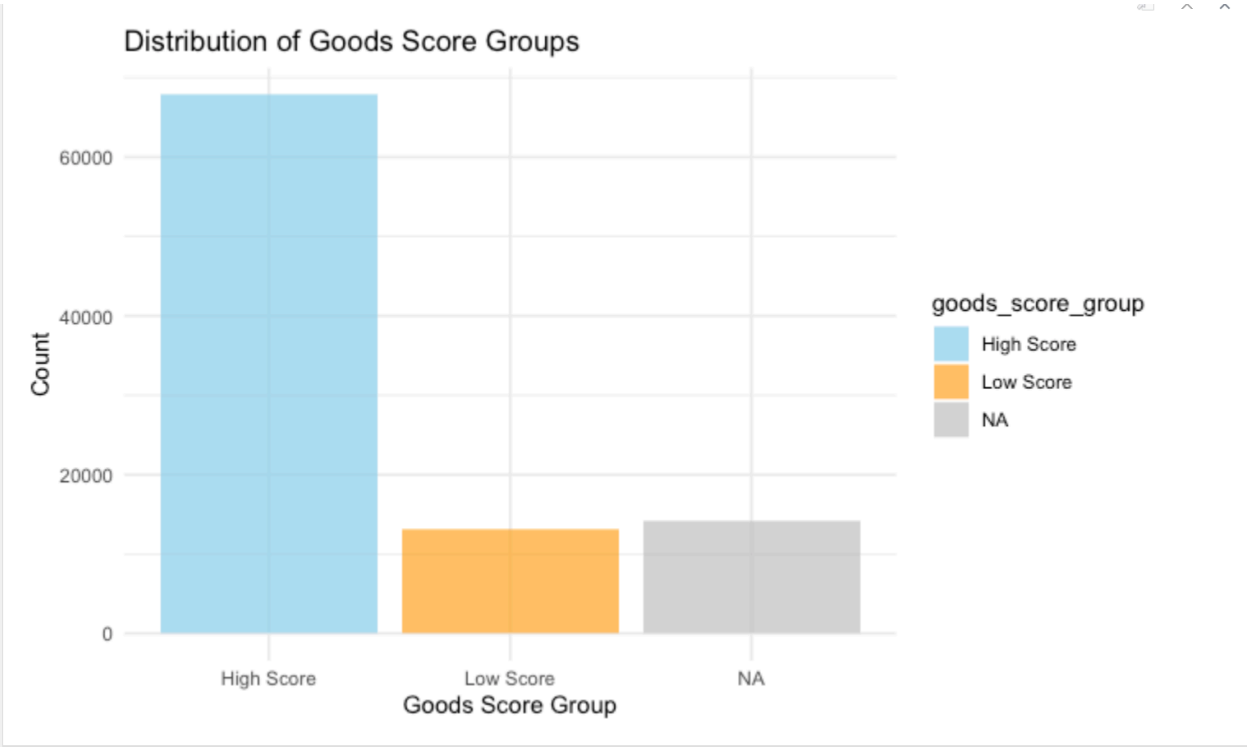


Figure1:Distribution of Goods Score

Our total data is more than 90,000, in which I set the rating above 4.5 as a high rating, below 4.5 is set to a low rating, and NA value for another column, from the chart we can see Most of the goods still get a high rating, while the other not rating and a low rating is nearly at the same level

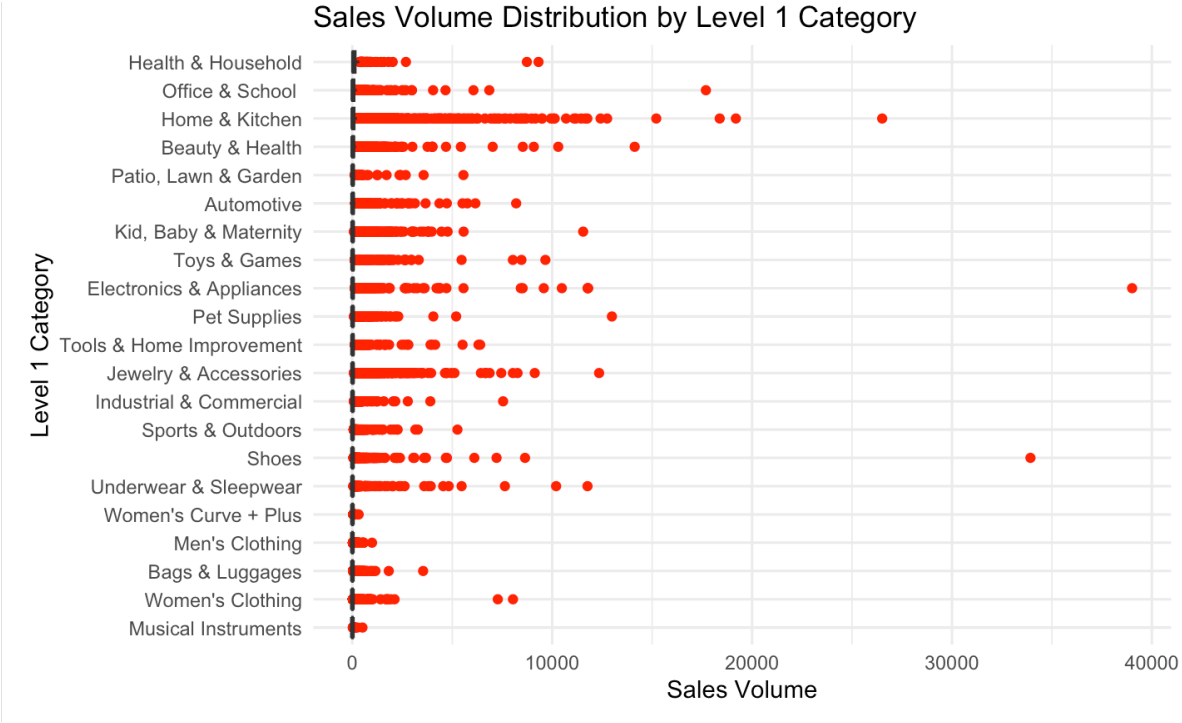


Figure 2: Distribution Of Sales Volume by Level 1 Category

Figure 2 shows the distribution of sales by level 1 category. Each red dot represents the sales volume of one product, the horizontal axis represents the sales volume, and the vertical axis lists the categories.

The graph shows significant differences between categories.” Electronic & Appliances“” and “Shoes” are categories in which product sales are particularly high, while categories such as “Musical Instruments” and “Women’s Curve + Plus” products have consistently lower sales. The majority of products are concentrated in lower sales volumes, with a few outliers dominating the higher range.

The level 2 category is a subdivision of the first-level category, so the number of second-level categories is huge. There are a total of 189, and the object of our research is a high-demand product, so I only selected the total sales volume of the top 20 categories.

Figure 3: Distribution of Sales Volume for Top 20 Level 2 categories

Figure 3 shows the distribution of sales volume for the top 20 secondary categories, providing a more detailed view of product performance within the subcategories. Each red dot represents the sales volume of one product, with categories such as “Women's Slippers” and “Headphones, Earbuds & Accessories” having the highest sales volume, including outliers of over 30,000 units.

Most of the products in these categories had lower sales volumes, indicating that only a few products in each category had higher sales volumes. Categories such as “Storage & Organizations” and “Kitchen & Dining” had a more even distribution of sales, while other categories such as “Pet Apparel & Accessories” had lower overall demand.

This segmentation highlights the importance of identifying high-performing products in specific subcategories and understanding the unique factors driving their success. It provides actionable insights for targeted marketing and inventory management.

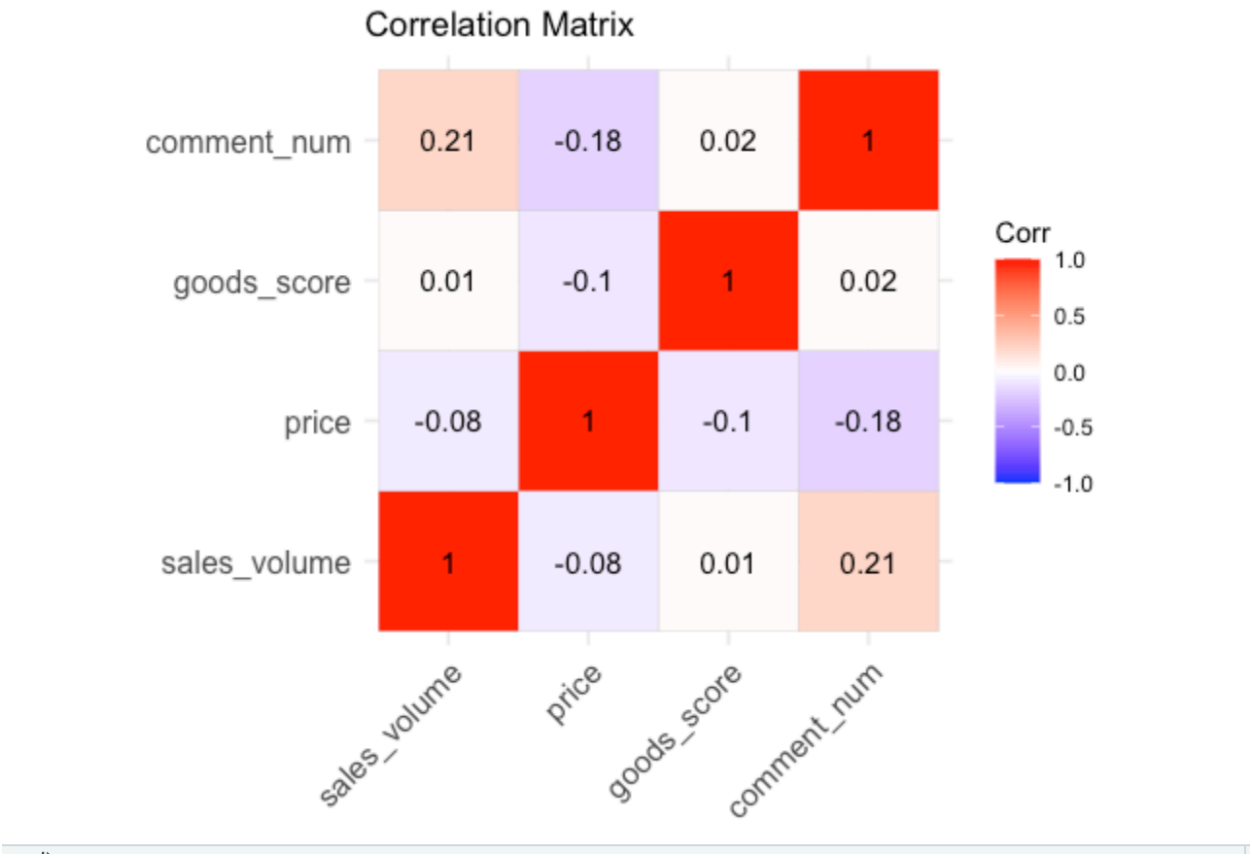


Figure4: Correlation Matrix

Figure 4 shows the correlation matrix reveals weak linear relationships among the variables. Notably, `sales_volume` has a slight positive correlation with `comment_num` (0.21), suggesting that higher comments may be associated with increased sales volume. Conversely, there is a weak negative correlation between `sales volume` and `price` (-0.08), indicating that price reductions might have a minimal impact on sales volume. The correlations between other variables, such as `goods_score` and the rest, are negligible, reflecting limited direct linear associations in the dataset. These results highlight the need for further exploration beyond simple correlation to understand the underlying dynamics.

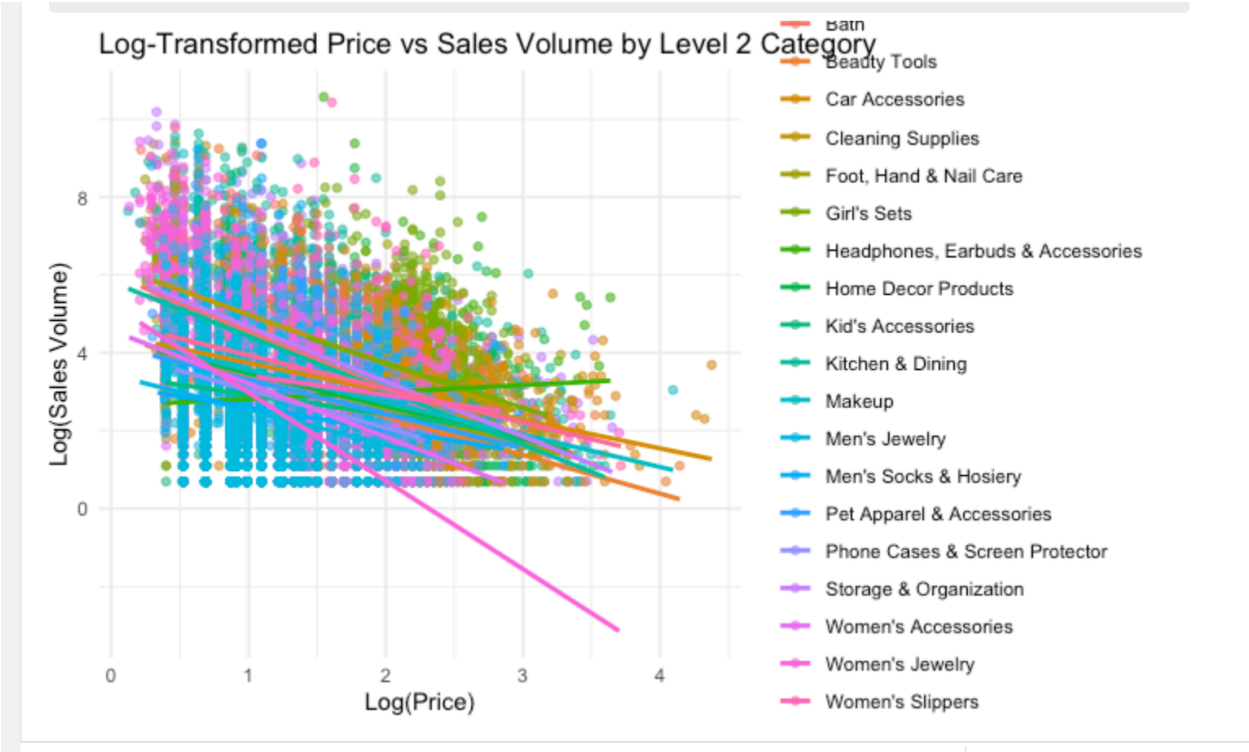


Figure 5 : Log Transformed

The scatter plot illustrates the relationship between log-transformed prices and log-transformed sales volumes for different secondary product categories. The right skewness of the price distribution is evident as many of the data points are clustered at lower price levels. There is an overall negative trend in the correlation between price and sales volume, indicating that higher prices are associated with lower sales volumes.

However, this trend is heavily influenced by a few extreme outliers, as the outliers (high sales volumes) were intentionally retained in the analysis. These outliers significantly affected the observed relationship, resulting in different slopes for different categories and highlighting the need for careful interpretation of the results in the presence of such data points.

Model Building

Based on explorations of correlation prior, we could not capture this variability with simple linear regression, so I chose a linear mixed effects model that by including random effects for price and level_2_category_name, we can account for unobserved

heterogeneity and ensure that fixed effects estimates are not biased by missing group-specific factors.

This approach provides both more accurate parameter estimates and accounts for the multilevel structure of the data.

```
Linear mixed model fit by REML ['lmerMod']
Formula: sales_volume ~ (1 | price) + high_score + comment_num + (1 |      leve_2_category_name)
   Data: top5_data

REML criterion at convergence: 93854.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.4170 -0.0944 -0.0196  0.0504 27.9885

Random effects:
 Groups               Name        Variance Std.Dev.
 price                (Intercept) 1091826  1044.90
 leve_2_category_name (Intercept)    2907    53.92
 Residual                          611640   782.07
Number of obs: 5775, groups:  price, 200; leve_2_category_name, 5

Fixed effects:
             Estimate Std. Error t value
(Intercept) 443.86750   90.48299   4.906
high_score  -13.53257   35.98940  -0.376
comment_num   0.27985    0.01843  15.181

Correlation of Fixed Effects:
            (Intr) hgh_sc
high_score  -0.346
comment_num -0.052 -0.087
```

Figure 6 : Lmer output

Interpretation

The model highlights the significant variability in sales volume driven by price and customer engagement across different products. The random effect for price (Variance = 1,122,289) shows that sales performance varies greatly depending on pricing strategies, emphasizing the need for tailored approaches. Similarly, the random effect for comment number (Variance = 66,099) underscores the importance of customer engagement, as its impact on sales is not uniform across products. In contrast, the small variance for product categories (Variance = 1,040) indicates minimal differences at the category level. These results suggest that while pricing and customer comments

are critical drivers, their influence is highly context-dependent, requiring product-specific strategies to maximize sales.

Result

Focus on Customer Engagement:
To maximize sales, businesses should prioritize collecting and showcasing customer reviews, especially for products that benefit more from engagement. Strategies include incentivizing reviews or improving review visibility on product pages.

Dynamic Pricing Strategies:
Pricing strategies should consider product-specific sensitivities. For high-priced variability products, dynamic pricing or targeted promotions can better align with consumer behavior and maximize revenue.

Leverage Predictive Insights:
This model demonstrates the need for data-driven decision-making, as simple linear relationships (e.g., between ratings and sales) may not provide actionable insights. Businesses should adopt advanced models to identify nuanced patterns and guide strategies effectively.

What‘s the next

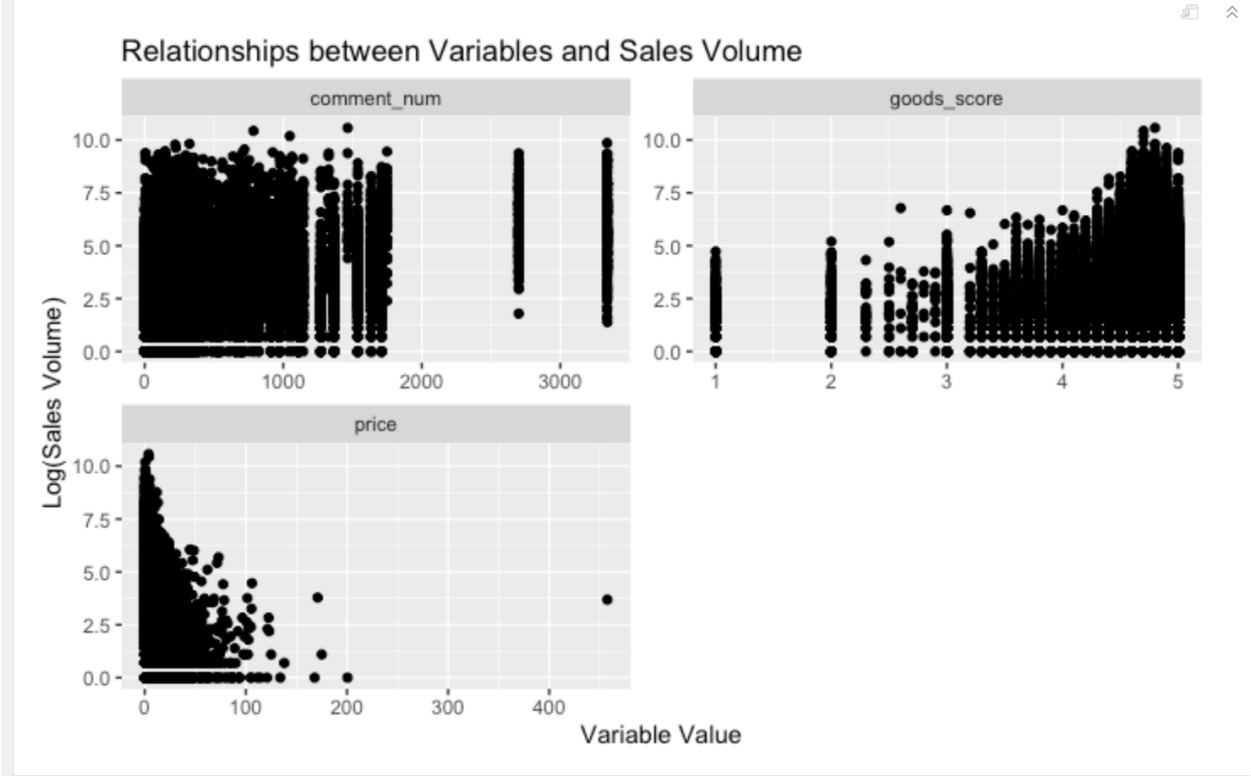
The current model provides actionable insights into the impact of pricing and engagement on sales, but the exploration of interactions, outlier effects, and external variables could be further improved to see the decisions. Future work should focus on refining the model, experimenting with pricing strategies, and expanding the scope of the analysis to ensure full business impact. Address limitations, presence of influential outliers in prices and sales volumes, perform sensitivity analysis or use robust modeling approaches. Examining the problem through further alternative models (random forest model)

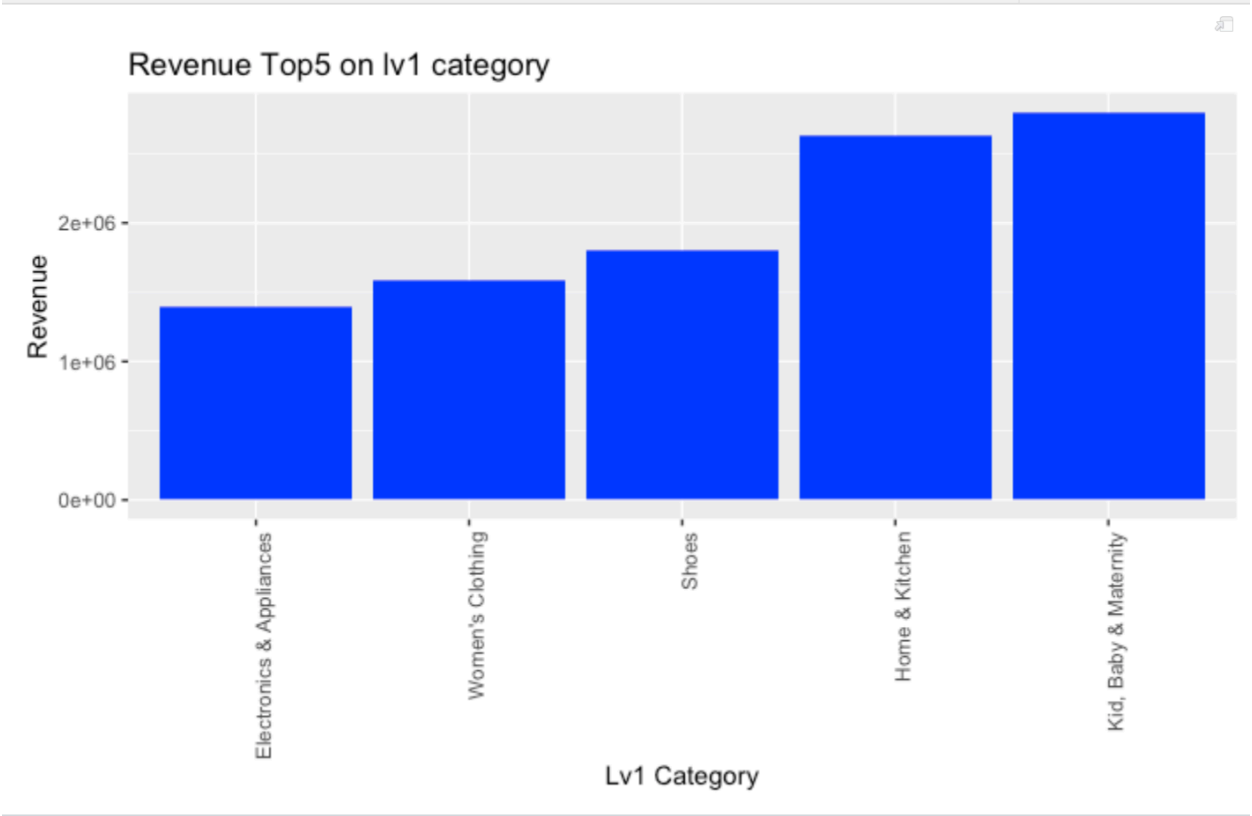
Appendix:

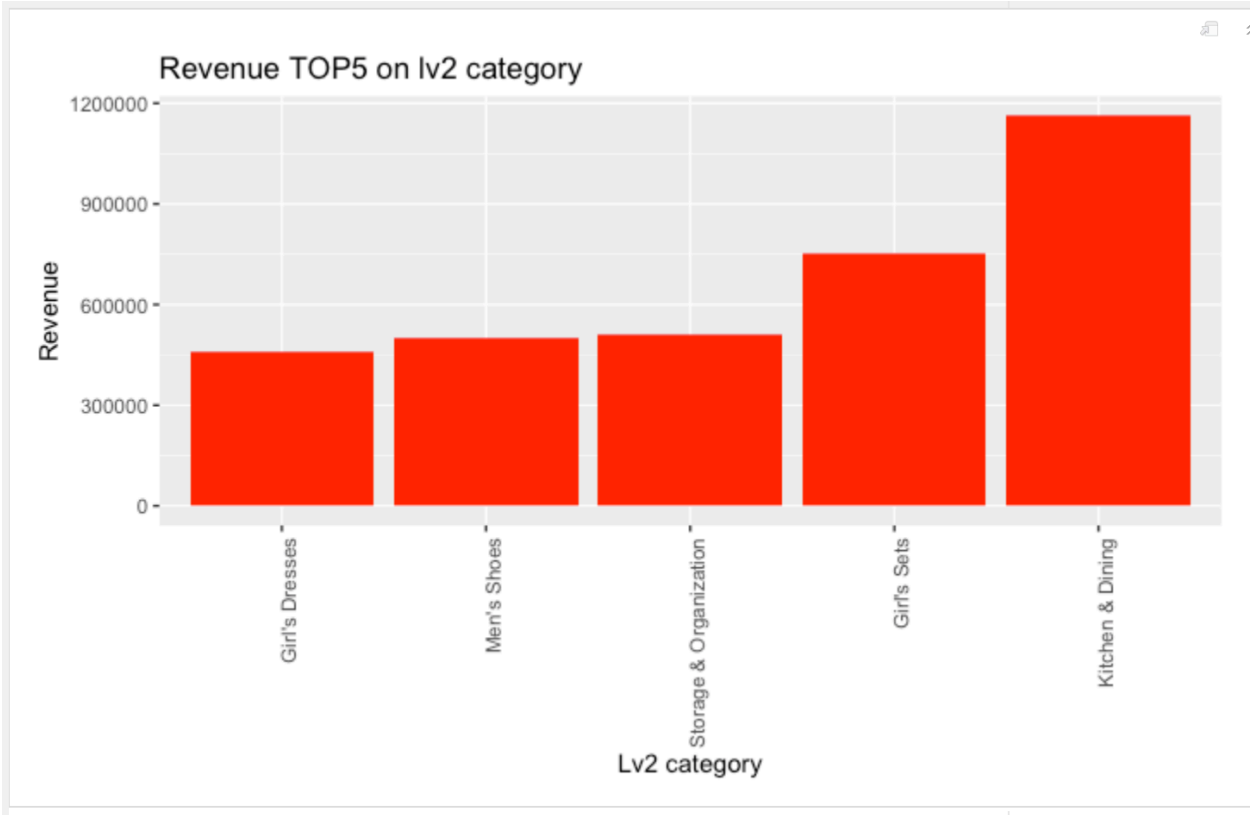
Data:
https://www.kaggle.com/datasets/polartech/temu-dataset-us-online-market-place

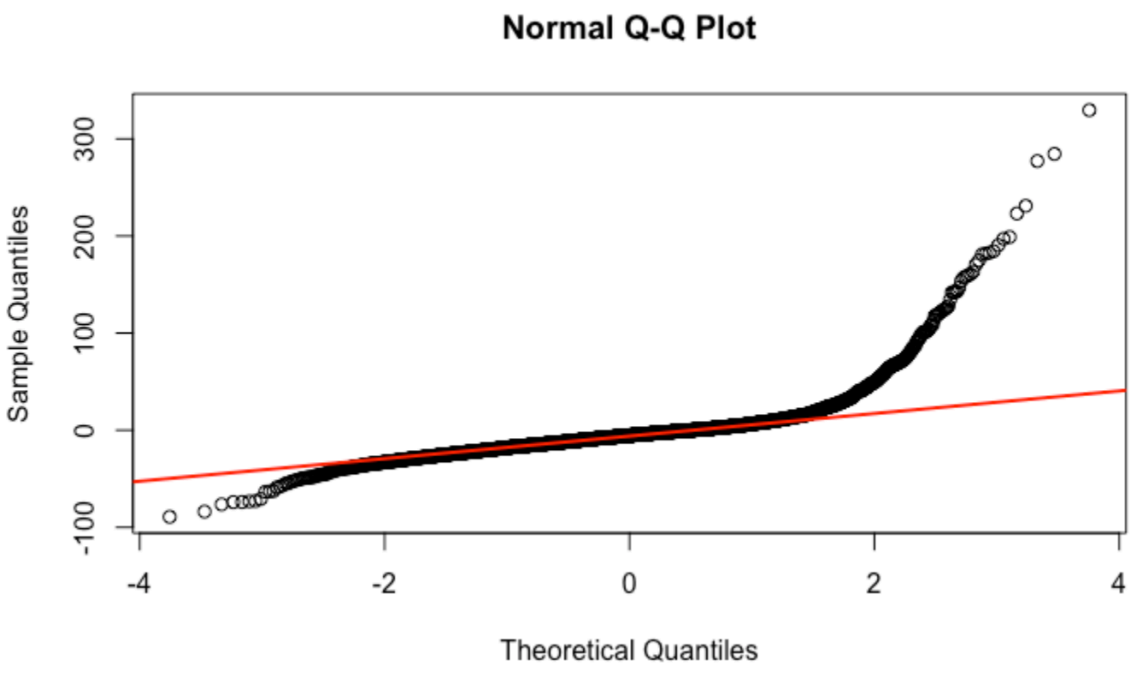
GitHub Code: https://github.com/KinokoZ/Temu

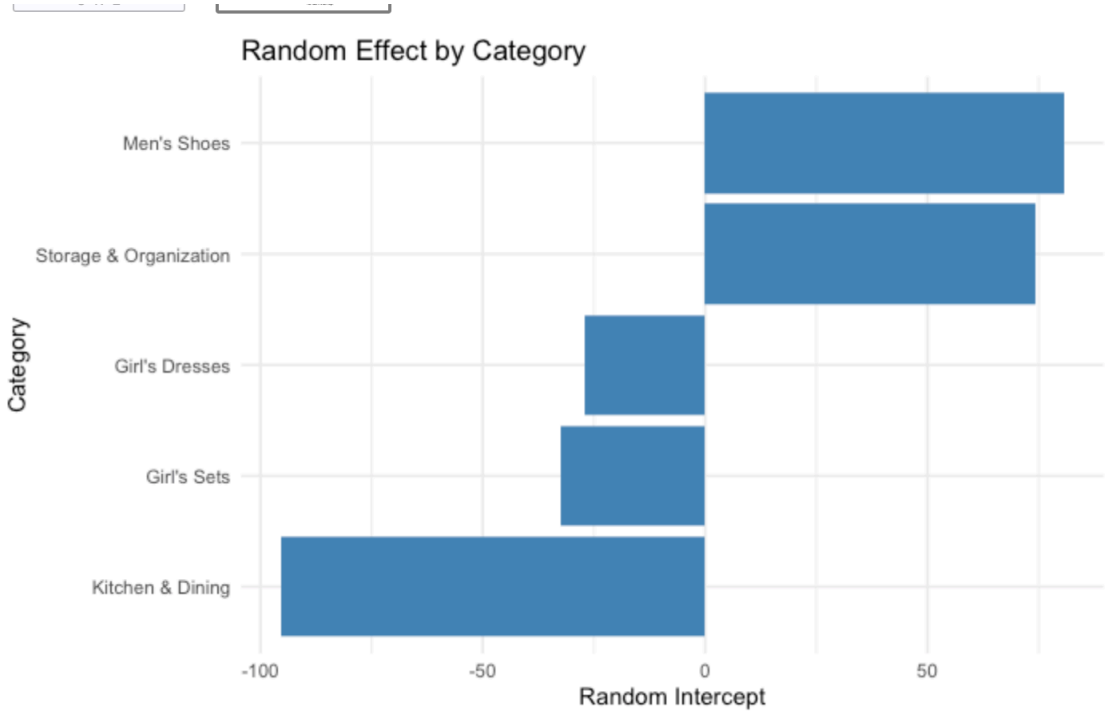
More EDA




Revenue Top5 on lv1 category
Revenue
2e+06
1e+06
0e+00
Electronics & Appliances
Women's Clothing
Shoes
Home & Kitchen
Kid, Baby & Maternity
Lv1 Category


Revenue TOP5 on lv2 category
Revenue
1200000
900000
600000
300000
0
Girl's Dresses
Men's Shoes
Storage & Organization
Girl's Sets
Kitchen & Dining
Lv2 category


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-100
0
100
200
300
-4
-2
0
2
4


Random Effect by Category
Men's Shoes
Storage & Organization
Girl's Dresses
Girl's Sets
Kitchen & Dining
Category
-100
-50
0
50
Random Intercept