**VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY**

**INTERNATIONAL UNIVERSITY**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A blue and white logo

Description automatically generated

**REPORT**

**SALES OF SUMMER CLOTHES**

**IN E-COMMERCE WISH**

A coursework of Data Mining

By Pham Le Thanh Nhan

Semester 02, 2023 - 2024

**TABLE OF CONTENTS**

[**LIST OF TABLES** 3](#_Toc168943421)

[**Abstract** 4](#_Toc168943422)

[**I. Introduction** 5](#_Toc168943423)

[**II. Implementation** 7](#_Toc168943424)

[**1.** **Data Pre-processing** 7](#_Toc168943425)

[**1.1.** **Data Cleaning** 7](#_Toc168943426)

[**1.2.** **Data Reduction** 8](#_Toc168943427)

[**1.3.** **Data Transformation** 10](#_Toc168943428)

[**2.** **Sales Volume of Summer Clothes Analysis** 10](#_Toc168943429)

[**2.1.** **Influence of Pricing on Sales** 10](#_Toc168943430)

[**3.** **Sales Volume Prediction** 11](#_Toc168943431)

**LIST OF FIGURES**

[Figure 1. Pricing Distributions 10](#_Toc168944202)

[Figure 2. Correlation between Pricing and Sales 11](#_Toc168944203)

[Figure 3. Discounting Proportion 11](#_Toc168944204)

# **LIST OF TABLES**

[Table 1. Shape of Data Frame after Removing Duplicates 7](#_Toc168943210)

[Table 2. Columns contain missing values 7](#_Toc168943211)

[Table 3. Shape of Data Frame after Handling Missing Values 8](#_Toc168943212)

[Table 4. Shape of Data Frame after Data Reduction 10](#_Toc168943213)

[Table 5. Shape of Data Frame after Data Transformation 10](#_Toc168943214)

# **Abstract**

Founded in 2010, Wish has rapidly evolved into a prominent global e-commerce platform connecting millions of merchants directly with consumers through its unique model of delivering affordable products from manufacturers. Despite its growth, Wish faces significant challenges related to customer satisfaction, delivery times, and product quality. Addressing these challenges is critical for maintaining its competitive edge in the e-commerce market.

This report presents an in-depth analysis of the sales volume of summer clothes on the Wish website, focusing on the influence of pricing, delivery time, customer satisfaction, and product quality. The primary objectives include integrating and analyzing sales data from the past month, identifying the key factors affecting sales volume, and constructing predictive models to forecast sales for the upcoming month. Data for this study is sourced from sales records, customer feedback, shipping logs, and product quality assessments.

Descriptive statistics provide insights into sales trends, pricing strategies, delivery performance, customer satisfaction levels, and product quality metrics. Inferential statistics, including correlation and regression analyses, quantify the impact of each factor on sales volume. Additionally, predictive modeling techniques are utilized to forecast next month's sales.

Key findings reveal significant correlations between sales volume and factors such as pricing, delivery time, customer satisfaction, and product quality. The predictive models demonstrate high accuracy in forecasting future sales, offering valuable insights for strategic planning. Based on these findings, actionable recommendations are provided to optimize pricing strategies, improve delivery times, enhance product quality, and boost customer satisfaction.

This comprehensive data analysis equips Wish with a deeper understanding of the factors influencing the sales volume of summer clothes, enabling the implementation of effective strategies to enhance its product offerings, customer experience, and overall market competitiveness.

Objectives summary:

1. Integrate and analyze sales data of summer clothes over the past month.
2. Identify key factors influencing sales volume: pricing, delivery time, customer satisfaction, and product quality.
3. Develop predictive models to forecast next month’s sales volume for summer clothes.

# **I. Introduction**

1. **Dataset Information**

The dataset titled "Summer Products with Rating and Performance" contains detailed information on various summer clothes available on the Wish website. Collected in August 2020, this dataset comprises 1,573 entries and includes 43 columns, each representing different aspects of the products and their performance. The dataset provides a comprehensive overview of product attributes, pricing, sales performance, customer feedback, and merchant details.

Key Columns and Their Descriptions:

|  |  |  |
| --- | --- | --- |
| **No.** | **i. Product Information** | |
|  | title | Name of the product. |
|  | title\_orig | Original name of the product. |
|  | tags | Tags associated with the product. |
|  | product\_color | Color of the product. |
|  | product\_variation\_size\_id | Size variation of the product. |
|  | product\_picture | URL of the product picture. |
|  | product\_url | URL of the product page on Wish. |
|  | **ii. Pricing and Sales** | |
|  | price | Price of the product. |
|  | retail\_price | Retail price of the product. |
|  | units\_sold | Number of units sold. |
|  | currency\_buyer | Currency used by the buyer. |
|  | uses\_ad\_boosts | Indicator of whether the product uses advertisement boosts. |
|  | **iii. Customer Feedback** | |
|  | rating | Average rating of the product. |
|  | rating\_count | Total number of ratings. |
|  | rating\_five\_count | Count of 5-star, 4-star, 3-star, 2-star, and 1-star ratings, respectively. |
|  | rating\_four\_count |
|  | rating\_three\_count |
|  | rating\_two\_count |
|  | rating\_one\_count |
|  | **iv. Shipping Information** | |
|  | shipping\_option\_name | Name of the shipping option. |
|  | shipping\_option\_price | Price of the shipping option. |
|  | shipping\_is\_express | Indicator of whether the shipping option is express. |
|  | countries\_shipped\_to | Number of countries the product is shipped to. |
|  | delivery\_option\_price | Price of the delivery option. |
|  | **v. Product Quality and Urgency** | |
|  | badges\_count | Number of badges the product has. |
|  | badge\_local\_product | Indicators of specific badges (local product, product quality, fast shipping). |
|  | badge\_product\_quality |
|  | badge\_fast\_shipping |
|  | has\_urgency\_banner | Indicator of whether there is an urgency banner. |
|  | urgency\_text | Text displayed in the urgency banner. |
|  | **vi. Inventory and Origin** | |
|  | product\_variation\_inventory | Inventory of the product variation. |
|  | inventory\_total | Total inventory of the product. |
|  | origin\_country | Country of origin of the product. |
|  | **vii. Merchant Information** | |
|  | merchant\_title | Title of the merchant. |
|  | merchant\_name | Name of the merchant. |
|  | merchant\_rating\_count | Number of ratings for the merchant. |
|  | merchant\_rating | Average rating of the merchant. |
|  | merchant\_id | Unique identifier for the merchant. |
|  | merchant\_has\_profile\_picture | Indicator of whether the merchant has a profile picture. |
|  | merchant\_profile\_picture | URL of the merchant’s profile picture. |
|  | merchant\_info\_subtitle | The subtitle text as shown on a seller's info section to the user. (raw, not preprocessed). |
|  | **viii. Miscellaneous** | |
|  | theme | Theme of the product. |
|  | crawl\_month | Month during which the data was collected (August 2020). |

This dataset provides a rich source of information for analyzing the performance of summer clothes on Wish, focusing on factors such as pricing, delivery time, customer satisfaction, and product quality. The comprehensive nature of the data enables detailed statistical analysis and predictive modeling to inform strategic decisions for optimizing product offerings and enhancing customer experience. ​

1. **Technology in used**

Utilized Python packages such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn for efficient data manipulation, numerical computations, generating visualizations, model construction and evaluation.

# **II. Implementation**

1. **Data Pre-processing**
   1. **Data Cleaning**

Data cleaning is a crucial step to ensure the dataset's accuracy and consistency. This process involved handling missing values, identifying and removing duplicates, and reducing the data by removing unnecessary columns.

*Removing duplicates:*

**Identifying Duplicates:** The dataset was checked for duplicate entries to ensure each record represented a unique product.

**Dropping Duplicate Entries:** Duplicate entries by product\_id were removed and the first one is kept, which means that each product appears only once in the dataset, ensuring that the dataset is lean.

Table 1. Shape of Data Frame after Removing Duplicates

|  |  |  |
| --- | --- | --- |
| **Value** | **Rows** | **Columns** |
| Before | 1573 | 43 |
| After | 1341 | 43 |

*Handling missing values:*

**Identifying Missing Values:** The dataset was examined for missing values across all columns. These are identified to contain missing entries.

Table 2. Columns contain missing values

|  |  |
| --- | --- |
| **No.** | **i. Product Information** |
|  | title |
|  | title\_orig |
|  | tags |
|  | product\_color |
|  | product\_variation\_size\_id |
|  | product\_picture |
|  | product\_url |
|  | **ii. Pricing and Sales** |
|  | price |
|  | retail\_price |
|  | units\_sold |
|  | currency\_buyer |
|  | uses\_ad\_boosts |
|  | **iii. Customer Feedback** |
|  | rating |
|  | rating\_count |
|  | rating\_five\_count |
|  | rating\_four\_count |
|  | rating\_three\_count |
|  | rating\_two\_count |
|  | rating\_one\_count |
|  | **iv. Shipping Information** |
|  | shipping\_option\_name |
|  | shipping\_option\_price |
|  | shipping\_is\_express |
|  | countries\_shipped\_to |
|  | delivery\_option\_price |
|  | **v. Product Quality and Urgency** |
|  | badges\_count |
|  | badge\_local\_product |
|  | badge\_product\_quality |
|  | badge\_fast\_shipping |
|  | has\_urgency\_banner |
|  | urgency\_text |
|  | **vi. Inventory and Origin** |
|  | product\_variation\_inventory |
|  | inventory\_total |
|  | origin\_country |
|  | **vii. Merchant Information** |
|  | merchant\_title |
|  | merchant\_name |
|  | merchant\_rating\_count |
|  | merchant\_rating |
|  | merchant\_id |
|  | merchant\_has\_profile\_picture |
|  | merchant\_profile\_picture |
|  | merchant\_info\_subtitle |
|  | **viii. Miscellaneous** |
|  | theme |
|  | crawl\_month |

**Imputing Missing Values:** For columns with significant missing data but crucial for analysis, imputation was performed. Missing values in 'rating\_five\_count', 'rating\_four\_count', 'rating\_three\_count', 'rating\_two\_count', 'rating\_one\_count', and 'has\_urgency\_banner' were filled with the zero value.

**Dropping Missing Values:** The others had excessive missing values which were not essential for the analysis and lightly accounted for the whole dataset, those columns or specific rows were dropped. This helped in maintaining a clean and manageable dataset without losing important information.

Table 3. Shape of Data Frame after Handling Missing Values

|  |  |  |
| --- | --- | --- |
| **Value** | **Rows** | **Columns** |
| Before | 1341 | 43 |
| After | 1274 | 41 |

* 1. **Data Reduction**

Data reduction involves removing unnecessary columns that do not contribute to the analysis or predictive modeling, thus simplifying the dataset and enhancing computational efficiency.

|  |  |
| --- | --- |
| **No.** | **i. Product Information** |
|  | title |
|  | title\_orig |
|  | tags |
|  | product\_color |
|  | product\_variation\_size\_id |
|  | product\_picture |
|  | product\_url |
|  | **ii. Pricing and Sales** |
|  | price |
|  | retail\_price |
|  | units\_sold |
|  | currency\_buyer |
|  | uses\_ad\_boosts |
|  | **iii. Customer Feedback** |
|  | rating |
|  | rating\_count |
|  | rating\_five\_count |
|  | rating\_four\_count |
|  | rating\_three\_count |
|  | rating\_two\_count |
|  | rating\_one\_count |
|  | **iv. Shipping Information** |
|  | shipping\_option\_name |
|  | shipping\_option\_price |
|  | shipping\_is\_express |
|  | countries\_shipped\_to |
|  | delivery\_option\_price |
|  | **v. Product Quality and Urgency** |
|  | badges\_count |
|  | badge\_local\_product |
|  | badge\_product\_quality |
|  | badge\_fast\_shipping |
|  | has\_urgency\_banner |
|  | **vi. Inventory and Origin** |
|  | product\_variation\_inventory |
|  | inventory\_total |
|  | origin\_country |
|  | **vii. Merchant Information** |
|  | merchant\_title |
|  | merchant\_name |
|  | merchant\_rating\_count |
|  | merchant\_rating |
|  | merchant\_id |
|  | merchant\_has\_profile\_picture |
|  | merchant\_info\_subtitle |
|  | **viii. Miscellaneous** |
|  | theme |
|  | crawl\_month |

* Some variables of type object are represented to products or merchants. They do not contribute to sale volume prediction. So, they should be removed.

'title', 'title\_orig', 'currency\_buyer', 'tags', 'merchant\_title', 'merchant\_name', 'merchant\_info\_subtitle', 'merchant\_id', 'product\_url', 'product\_picture', 'product\_id', 'theme', 'crawl\_month'

* Columns with very few value count in most indices (less than 5 indices) meaning that it is imbalanced distributed should be removed to avoid bias.

'badge\_local\_product','badge\_fast\_shipping','shipping\_is\_express'

Table 4. Shape of Data Frame after Data Reduction

|  |  |  |
| --- | --- | --- |
| **Value** | **Rows** | **Columns** |
| Before | 1274 | 41 |
| After | 1274 | 25 |

* 1. **Data Transformation**

The columns stored as object types were converted to appropriate numerical types (e.g., float or integer). This ensured that mathematical operations and analyses could be performed accurately.

'product\_color', 'product\_variation\_size\_id', 'shipping\_option\_name', 'origin\_country'

**Feature Engineering:**

Create a new categorical variable named sales\_status, indicating how well products be sold, high or low level.

Table 5. Shape of Data Frame after Data Transformation

|  |  |  |
| --- | --- | --- |
| **Value** | **Rows** | **Columns** |
| Before | 1274 | 25 |
| After | 1274 | 26 |

1. **Sales Volume of Summer Clothes Analysis**
   1. **Influence of Pricing on Sales**

A graph of a graph

Description automatically generated

Figure 1. Pricing Distributions

**Price Distribution:** Most products are priced between $0 and $20, with a notable concentration around $10. The distribution is right-skewed, indicating a larger number of lower-priced items.

**Retail Price Distribution:** The retail prices show a wider spread, with many items below $50 but some extending up to $250. This distribution is also right-skewed.

**Price Difference Distribution:** The price difference (Price - Retail Price) shows that the majority of the items are sold below their retail prices, as indicated by the concentration of values on the negative side of the distribution. A significant peak is observed near zero, meaning many items are sold at prices close to their retail values, but the bulk is sold at a discount.

A graph showing sales by pricing

Description automatically generated

Figure 2. Correlation between Pricing and Sales

There is a visible clustering of sales with a retail price below $50 and actual prices below $20. As the retail price increases, the number of sales decreases, and there is a wider dispersion in the actual sale prices.

High and low sales statuses are distributed across different price points, but a higher density of low sales statuses is observed in the lower price ranges. There are few high sales statuses for items with high retail prices, indicating that higher-priced items may not sell as well.

A pie chart with numbers and a percentage

Description automatically generated

Figure 3. Discounting Proportion

A larger proportion (63.8%) of products are sold below their retail price, while 36.2% are sold above. This suggests that discounting is a common practice and might be necessary to drive sales.

Conclusions:

1. There is a clear price sensitivity among consumers, with a majority of sales occurring at lower price points. Items priced below $20 are more frequently purchased.
2. Selling below the retail price appears to be a common and possibly effective strategy to boost sales, as evidenced by the larger proportion of sales occurring below the retail price.
3. Higher retail prices are associated with fewer sales, and significant discounts might be required to move these items.
4. High sales statuses are more prevalent at lower price points, indicating that competitive pricing is critical for driving higher sales volumes.

In summary, the pricing strategy significantly influences sales, with lower prices and discounts being key factors in promoting the sale of summer clothes on the e-commerce platform.

* 1. **Customer Satisfaction**

A graph of sales by product rating

Description automatically generated

A graph of sales by merchant rating

Description automatically generated

* 1. **Delivery Time**
  2. **Product Quality**
  3. **Advertising**

1. **Sales Volume Prediction**