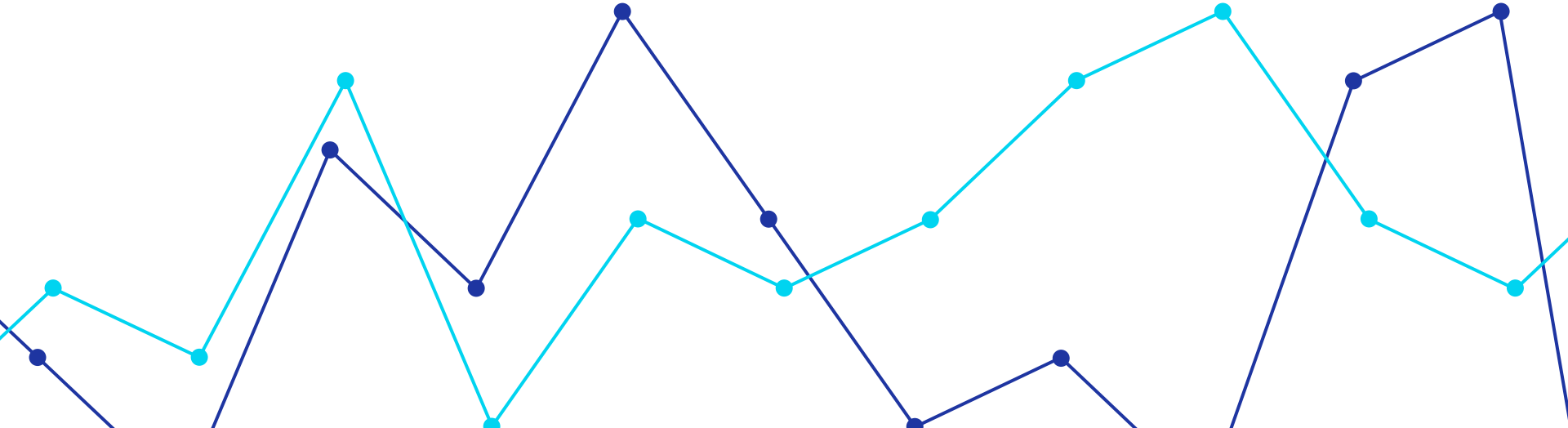
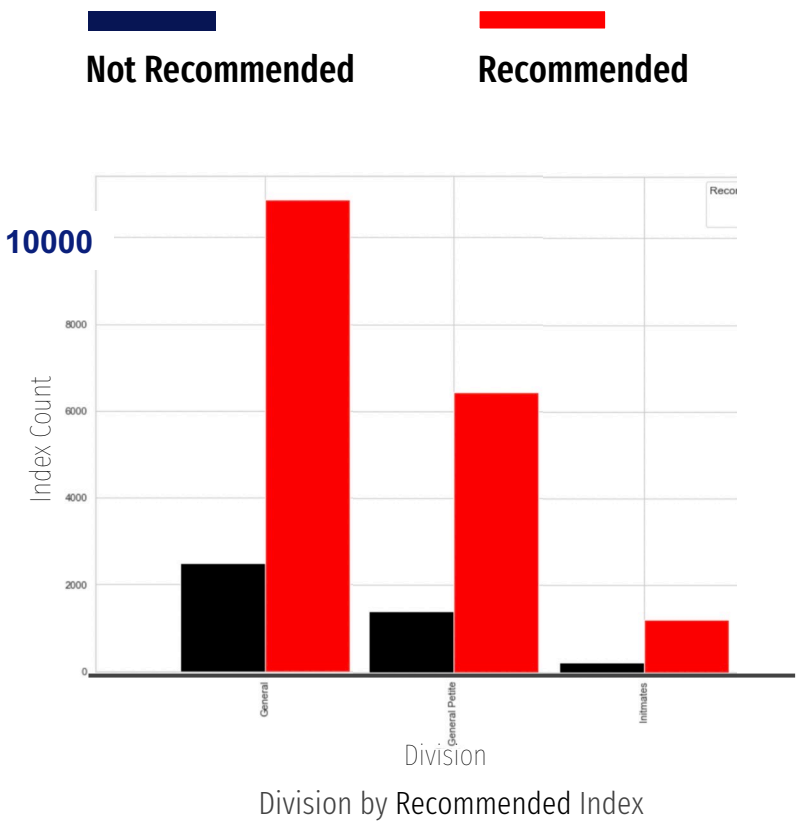
Data Analysis Report

Sentiment Analysis And Classification Modelling On E-Commerce Reviews



Recommended Rate Higher Than The Unrecommended



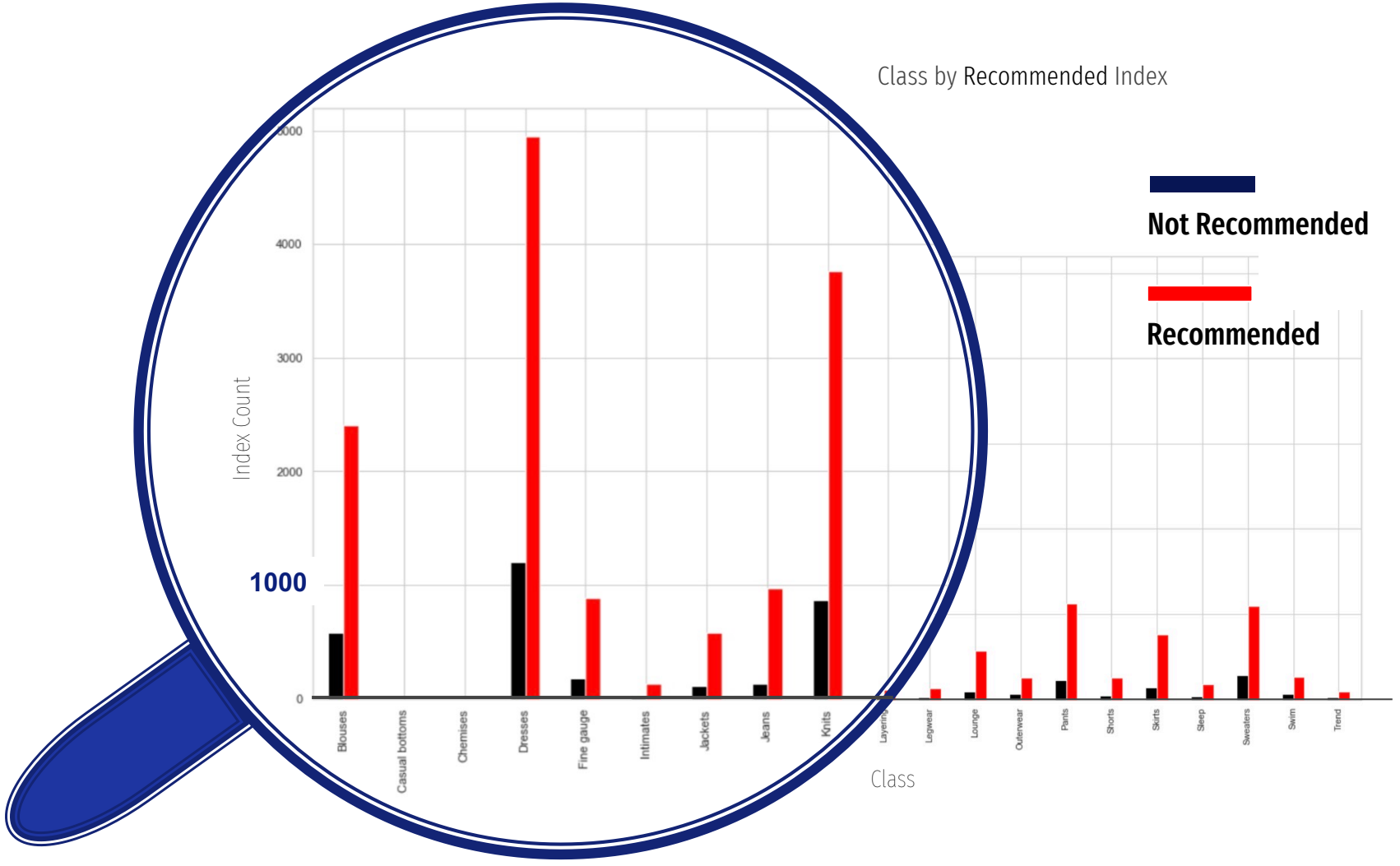
Division by Recommended Index

- The **sold units are decreasing as the sizes of the clothes get smaller**

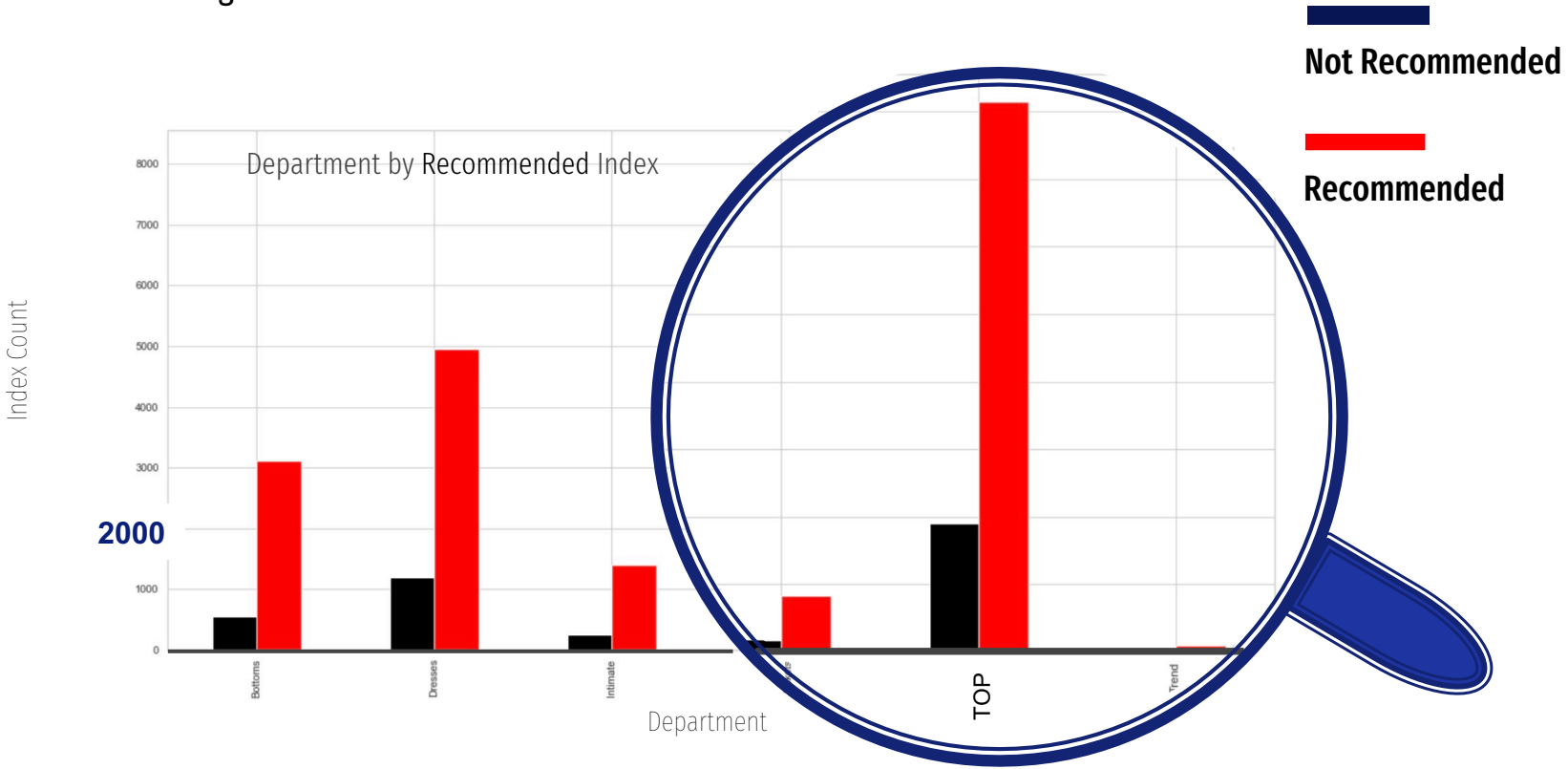
- "General" division has the highest sold units, "General Petite" ranks second, and "Intimates" has **the lowest unit sold**.
  - In other words, compared to other different divisions, a larger proportion of customers recommend "General" than "General Petite" and "Intimate". Based on these metrics, we should design the products in larger sizes for the following seasons, such as medium, large and extra-large.

- There will be a more promising future for the e-commerce platform when it **provides a wider variety of types of clothes**.
  - The women's e-commerce platform has to embrace the new trend of body positivity, though it could be a sharp turn away from the styles that defined the women's apparel industry for decades.

- Product **quality consistency** is a principle to the overall success of every business.
  - Providing consistency allows customers to know what to expect every time they merchandise and every product they purchase, which could increase the trust of the customers towards the brand and increase sales units in the long term.


Class by Recommended Index
Not Recommended
Recommended
Index Count
1000
0
Blouses
Casual bottoms
Chemises
Dresses
Fine gauge
Intimates
Jackets
Jeans
Knits
Layering
Legwear
Lounge
Outerwear
Pants
Shorts
Skirts
Sleep
Sweaters
Swim
Trend
Class

Based on charts, "Class by recommended Index" above and "Department by recommended Index" below, we can find out the unrecommended rate in **Blouses, Dresses, Knits, and Top are relatively higher than in other categories**. Customers are able to observe and realize the quality of clothing products. By improving the product consistency aim for these four types of segments could be a significant increase to the recommended rate.

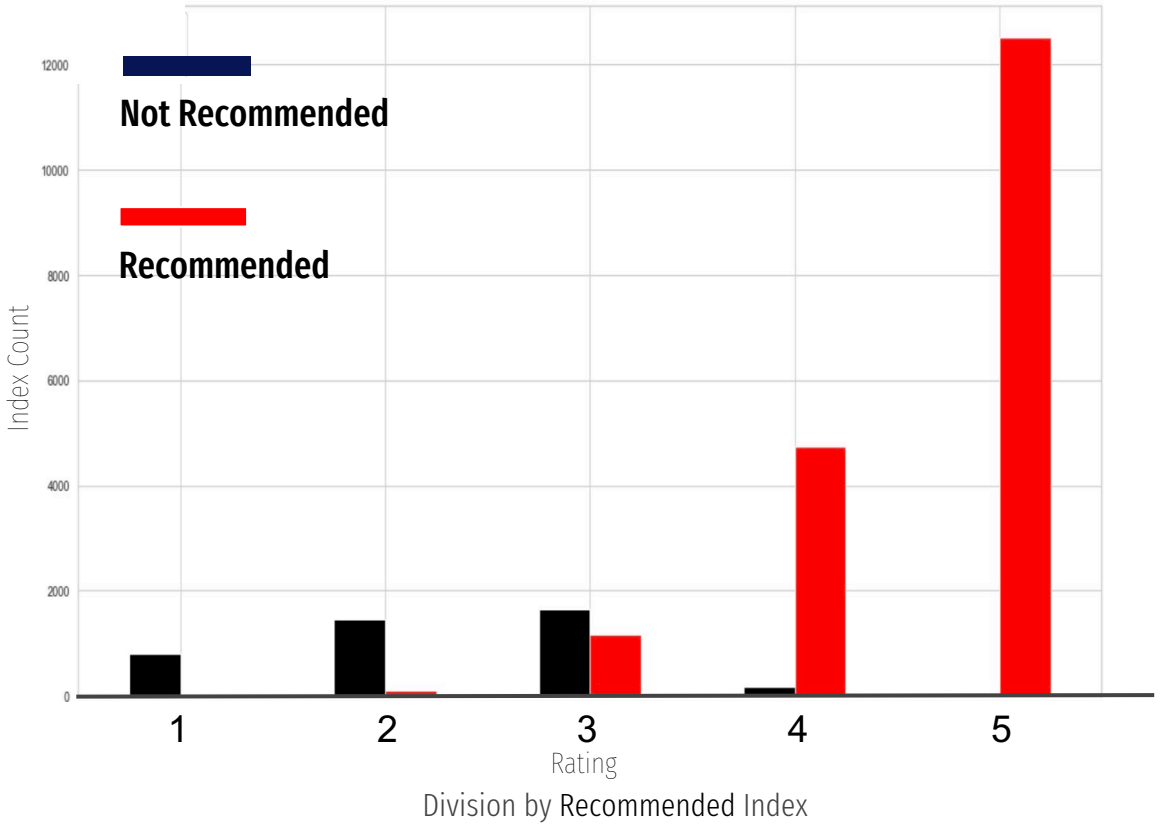


Rating Score (From 1 To 5) Is Positively Related To The Customers' Recommended Propensity

Customers who rate the platform less than 3 stars, their recommendation inclination is inconsistent. So, the platform **could enforce the customer relations management** with this segment of customers, which could potentially turn their negative shopping experience into a positive one.

For the customers who rate the platform at 4 or 5, **product consistency** is a key to maintaining customer retention; Also, for the customers who rate the platform under 3, which could be the churn rate that is a loss for the platform.

In addition, the proportion of ratings more than 3 stars on the condition of customers who recommend (81%) is higher than ratings less than 3 stars on the condition of customers who don't recommend (19%).



Fabric And Sweaters Are The Words That Show Up More Frequently in both positive and negative comments

The platform should **narrow the customer segmentation**, to clarify which types of customers could be the target audience depending on the type of fabric they choose.

The platform can **expand the business**: create multiple product lines, to target different types of customers, such as a high-end line for customers at the age of 28-35, with better financial capability and a classic line for the customer at the age of 18-25, with relatively less money to spend on apparel.

Pants are the category of products that have been mentioned significantly in the positive comments. Based on the metrics, the women's e-commerce platform could **take the product categories as advertised** to attract new customers based on the high customer reputation.

Size is the biggest concern of customers, which could be related to the bar chart of "division name". The women's e-commerce platform produces mainly smaller sizes for the customers, which could harm the platform's reputation. Therefore, the platform could benefit from producing the medium, large and extra-large sizes of apparels.

Positive words

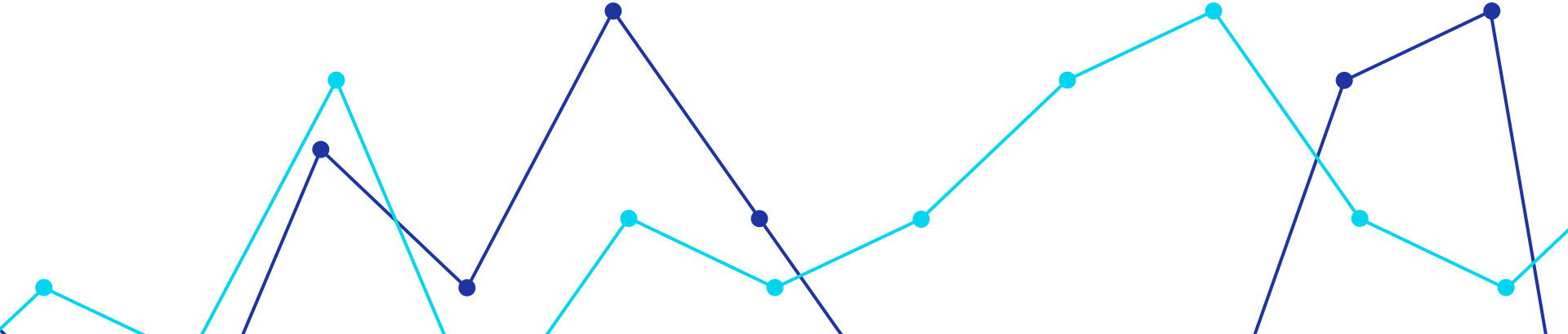
Negative words

The Project Challenges

There were two main challenges during the journey of working on this projects.

Firstly, I found that the dataset was imbalanced between the recommended and not recommended of reviewer comments. In order to solve the problem, we used SMOTE to oversample the data in the “negative” class. After the oversampling, we were able to build further different models, and compared and evaluated their performances.

Second, I also encountered the challenge of how to evaluate different models as a consistent benchmark. Due to the advantages of the ROC curve, I decided to use it as a criterion to indicate the performance of multiple models at all classification thresholds. With the higher AUC scores, which compound score is higher than 0.5, I am able to measure the better performances of the models.



Conclusion

The Performance of the models

Model	Accuracy	Recall	F-score	AUC
Logistic Regression	**0.9627**	**0.9627**	**0.9678**	**0.81**
Naive Bayes	0.9274	0.9274	0.9445	0.75
Decision Tree	0.9466	0.9466	0.9550	0.68
Random Forest	0.9621	0.9621	0.9634	0.65

Compared to the different performance of each model I used, the model using **Logistic Regression performs the best**. Because its AUC score and F1-Score are the highest, 0.81 and 0.9627 among all.

From those charts and tables presented above, we conclude the following points. Firstly, the women's e-commerce platform can expand the amount of product by producing larger sizes of apparels. Moreover, customers do value the importance of product quality, which reflects on higher ratings and more positive reviews. The platform should make more efforts to improve and maintain its consistent product quality so that the platform can attract new potential customers.

Hopefully, as long as the platform is willing to adjust and adapt to the trends nowadays, it leads to higher profits, better reputation and loyalty to the platform.