A Study on Factors Influencing the Yearly Amount Spent by Customers of a Fashion and Cosmetic Brand

Dinghan Zhang

March 30, 2025

Introduction

With the development of e-commerce, an increasing number of brands are committed to building and improving their own e-commerce platforms like App and website. At the same time, the importance of in-store service to the brand remains undiminished. This report focuses on identifying the factors influencing the yearly amount spent by customers of a fashion and cosmetic brand. To achieve this objective, two main questions need to be addressed. Firstly, among three factors—average time spent in-store, average time spent on App, average time spent on website—which has the most significant impact on yearly amount spent? Secondly, this report will explore whether there is a significant difference in yearly amount spent by Males compared with Females. The findings of this report will assist the brand in determining whether to prioritize investment in website improvement or App enhancement or in-store services. Furthermore, it will help the brand identify its target customer groups, enabling them to provide more tailored services and products, which could optimize profitability.

Research Design/Methodology

Research Design

This report analyses the relationship between yearly amount spent and three factors, including average time spent in store, average time spent on the App, and average time spent on the website. Additionally, this report also examines whether there is a difference in yearly amount spent between Males and Females. It is worth noting that in addition to the four factors mentioned above, there are other lurking variables that may also affect the yearly amount spent such as the frequency of promotional campaigns, the advertising expenditure of the brand, and the average income level of target cus-

tomers. In order to enhance the representativeness of selected data and improve the precision of estimates, a stratified random sampling method was applied. Firstly, the overall population was stratified into two groups according to gender. Secondly, given that the total sample size is 50, 25 samples were randomly selected from each group. Through this method, the representativeness of selected samples can be ensured.

Methodology

After completing the data sampling, a numerical summary was conducted to evaluate the quality of the data, in order to determine whether these data can be used for subsequent analysis and research. A table containing statistical information such as data range, IQR, and standard deviation will be created. Boxplot and histogram will be constructed to intuitively present the distribution of data. By conducting numerical summary, critical information about the selected data such as the distribution, level of dispersion, range of values, and presence of outliers can be clearly identified. And this key information is crucial for evaluating the quality of the selected data. Linear regression method will be applied to analyse the relationship between the three factors, that average time spent in store, average time spent on the App, and average time spent on the website, and yearly amount spent. By observing the generated linear regression line generated from the scatterplot, it can be clearly determined whether the qualitative relationship between each factor and the yearly spent amount is positive, negative or non-existent. Furthermore, by comparing the slopes of the regression lines, it can be determined that which factor has a more significant impact on the yearly amount spent. After constructing the linear regression model, residual analysis is needed to assess the validity of the constructed model. If the residual points are randomly scattered and distributed around zero without any noticeable curve or systematic pattern, it indicates that the linear regression model is valid; if a distinct curve or another non-linear pattern appears, it may suggest the existence of a non-linear relationship. Since the sample size is fewer than 30, a t-test is employed to ascertain whether gender differences influence yearly amount spent. However, it should be noted that the use of t-test assumes that the population follows a normal distribution. If the population distribution is not normal, the t-test may be invalid.

Result

Full Numerical Summary (T)	**Time in store (Minutes)**	**Time on APP (Minutes)**	**Time on Website (Minutes)**	**Yearly Amount Spent (£)**
Min	1.11	0.19	0.18	0.17
Q1	3.43	3.32	3.21	3.01
Q2	4.62	4.61	4.60	3.79
Q3	5.55	5.24	5.23	5.21
Max	6.67	6.55	6.54	6.35
SD	7.26	7.13	7.55	7.33
IQR	13.23	12.13	12.54	13.31
Range	4.23	4.12	4.54	4.36
Skew	7.12	7.41	7.53	7.32
1.5IQR	11.22	11.14	12.55	12.33
Lower Limit	7.24	7.15	6.15	7.93
Upper Limit	17.42	17.17	17.55	17.36

Table 1: Numerical summary

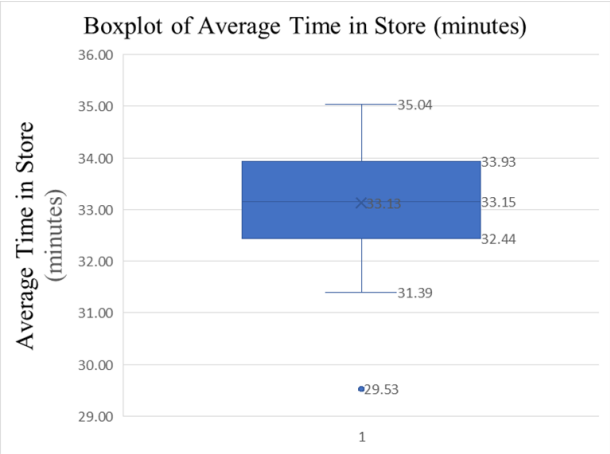

Figure 1: Boxplot of average time in store

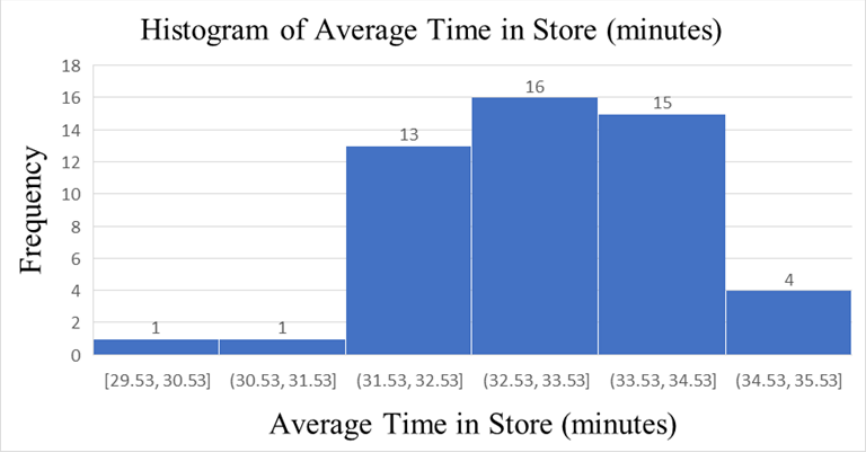

Figure 2: Histogram of average time in store

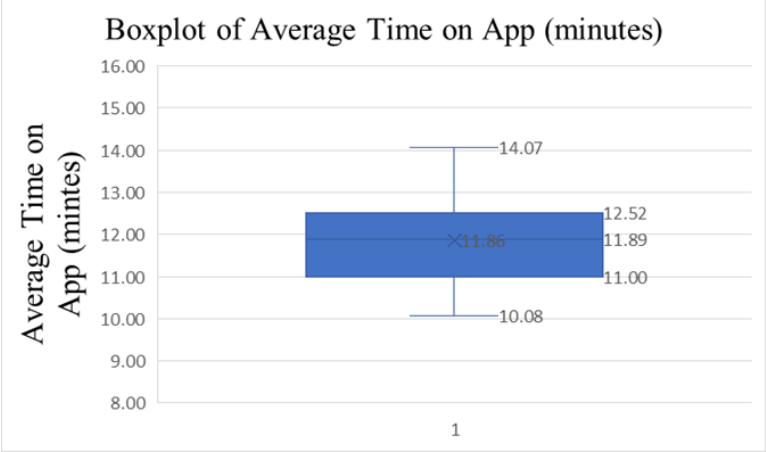

Figure 3: Boxplot of average time in App

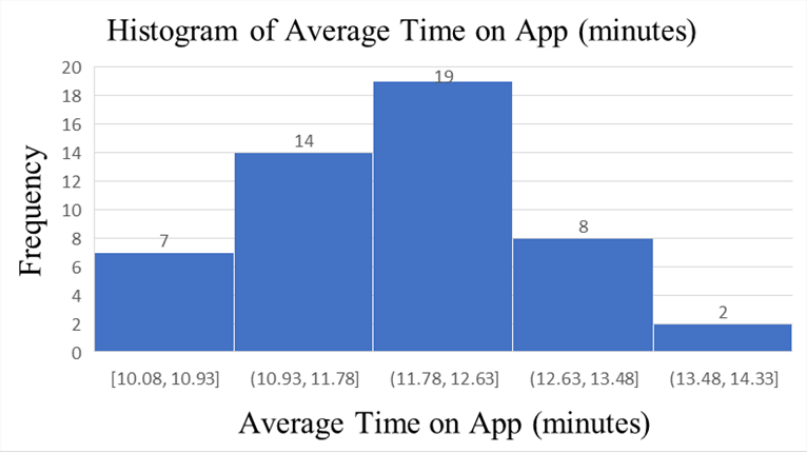

Figure 4: Histogram of average time in App

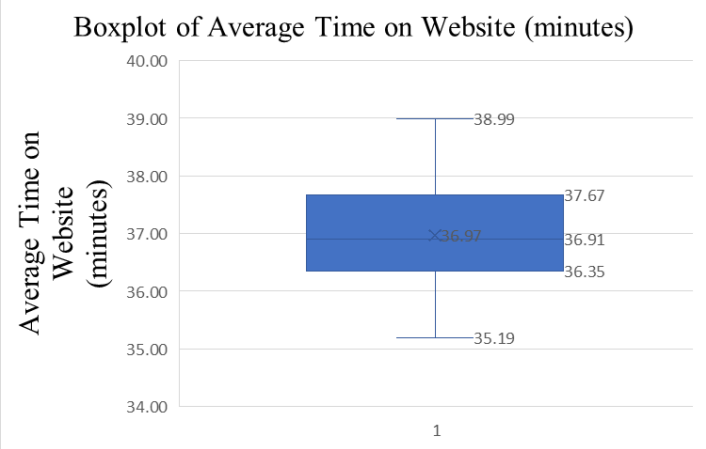

Figure 5: Boxplot of average time in Website

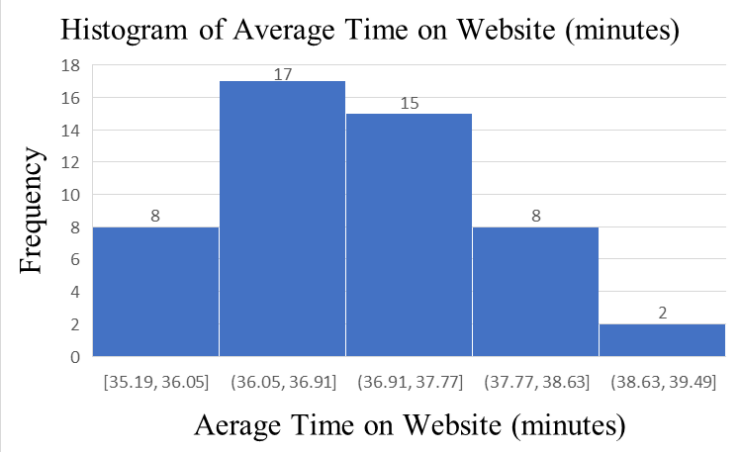

Figure 6: Histogram of average time in Website

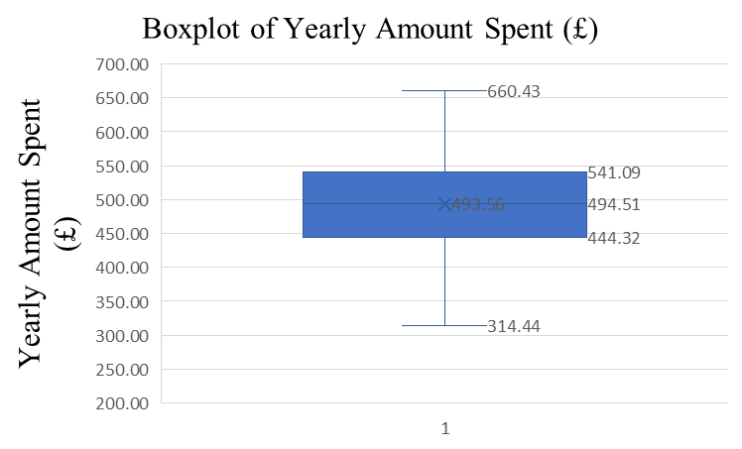

Figure 7: Boxplot of Yearly Amount Spent

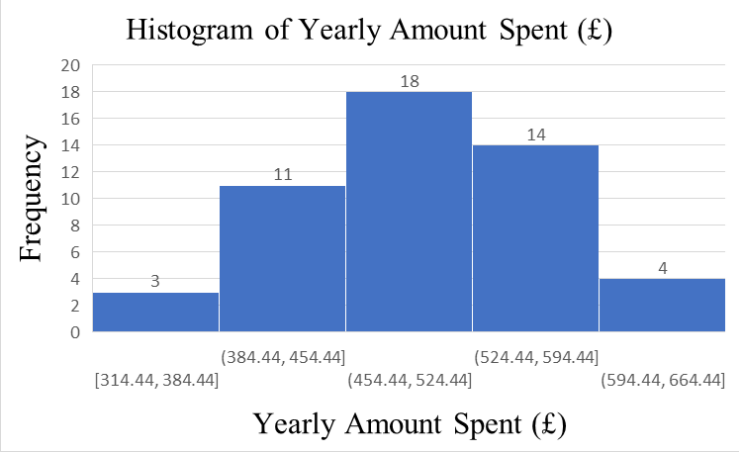

Figure 8: Histogram of Yearly Amount Spent

	Time in Store
Least Square Regression Line	$\hat{y} = 30.338x - 511.55$
Pearson's r	0.438
Coefficient of Determination	0.192

Table 2: Summary of results of linear regression model of time in store.

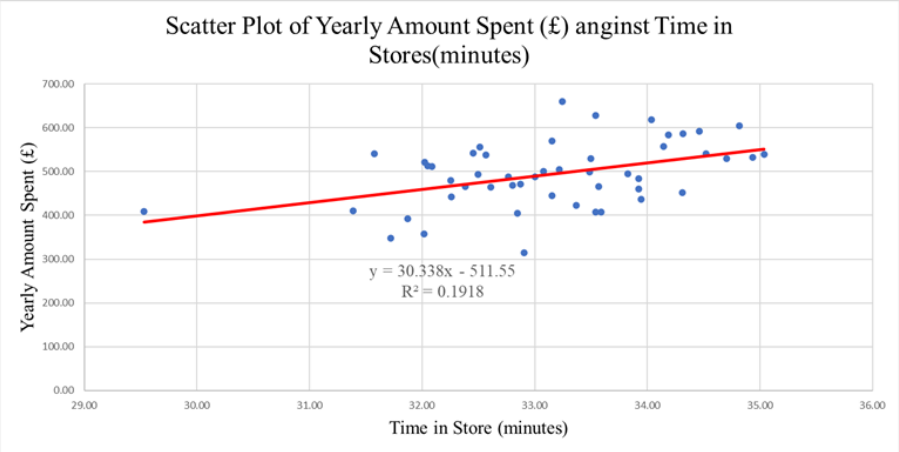

Figure 9: The scatter plot of yearly amount spent against time in store.

	Time on App
Least Square Regression Line	$\hat{y} = 49.607x - 94.881$
Pearson's r	0.600
Coefficient of Determination	0.360

Table 3: Summary of results of linear regression model of time on App.

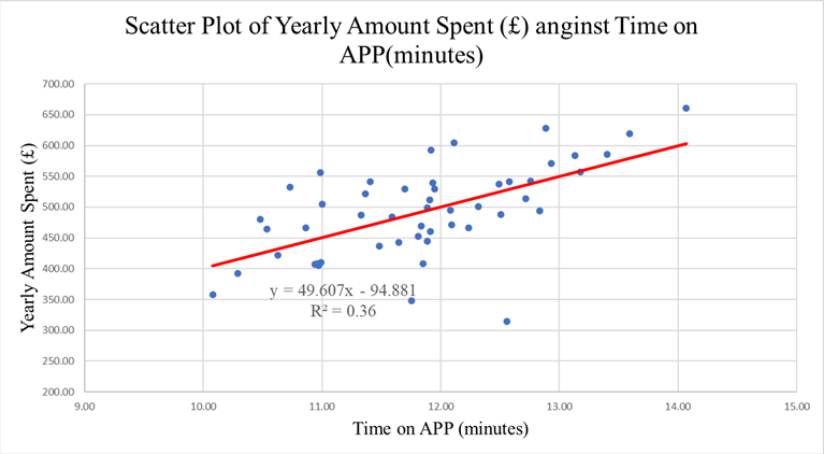

Figure 10: The scatter plot of yearly amount spent against time on App.

	Time on Website
Least Square Regression Line	$\hat{y} = -3.6342x + 627.92$
Pearson's r	-0.044
Coefficient of Determination	0.002

Table 4: Summary of results of linear regression model of time on Website.

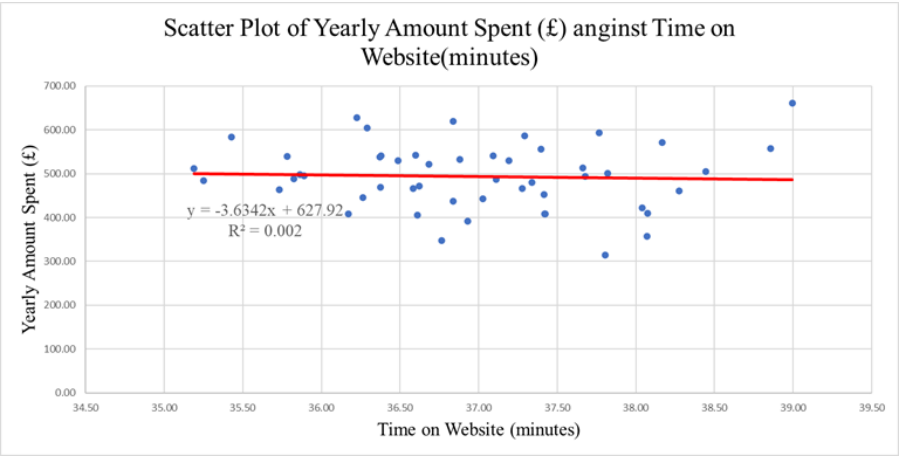

Figure 11: The scatter plot of yearly amount spent against time on Website.

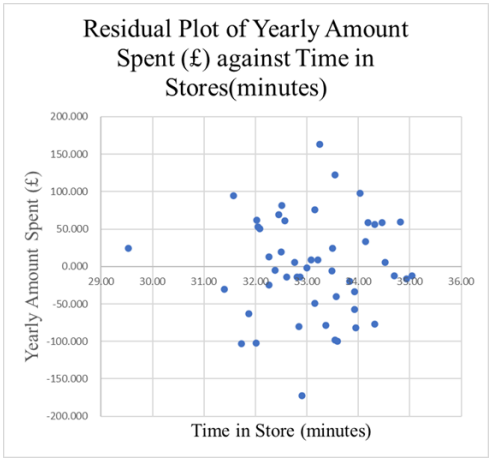

Figure 12: The residual plot of yearly amount spent against time in stores.

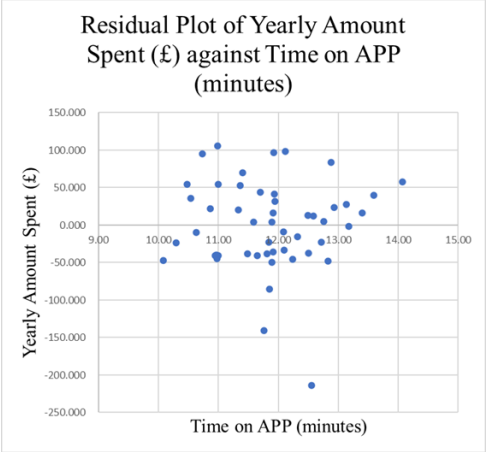

Figure 13: The residual plot of yearly amount spent against time on App.

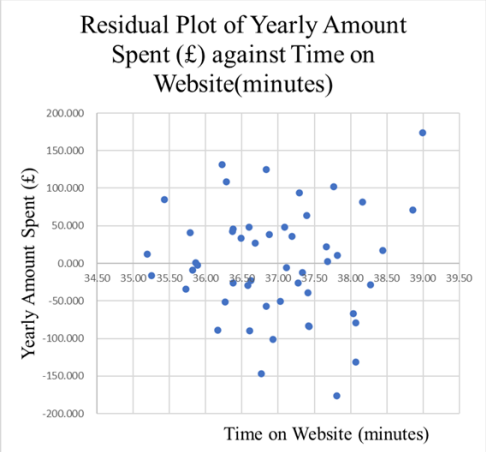

Figure 14: The residual plot of yearly amount spent against time on Website.

Discussion and Findings

Numerical Summary

By comparing the data in **Table 1**, among the three factors that average time in store (ATS), average time in App (ATA), and average time in website (ATW), the following observations can be made:

1. The mean of ATA is noticeably smaller than the means of the other two factors and the means of ATS and ATW are similar to each other.

2. The range of ATS is largest at 5.505, which indicates ATS has the greatest variability among the three factors.

3. Both interquartile range (IQR) and standard deviation are similar across all three factors, suggesting that the data sets are comparably dispersed.

However, it is worth noting that there is an outlier in ATS data (29.53), which could impact the mean and standard deviation and therefore deserves careful consideration. As shown in **Figure1**, the boxplot for ATS clearly displays this outlier. Calculations reveal that after removing the outlier, the mean and standard deviation of ATS are 33.204 and 0.946, respectively, compared to 33.131 and 1.071 for the original data. Since the impact is not very significant, the outlier is retained for subsequent analysis.
Furthermore, as shown in **Figure 2**, the histogram of ATS exhibits a slightly left (negative) skew with a skewness of -0.579, which indicates the mean of ATS is smaller than the median. In contrast, **Figures 3, 4, 5, and 6** show that there are no outliers in the ATA and ATW data sets, and both data sets exhibit positive skewness, indicating that, for each data set, the mean is greater than the median. Lastly, for the yearly amount spent (YAS), from **Figures 7 and 8**, the range of YAS is 345.987 and the IQR is 96.778, there is no outlier and the histogram of YAS exhibits a slightly left (negative) skew with a skewness of -0.103, indicating the mean is smaller than the median

Correlation and Residual Analysis

As show in **Figures 9, 10, and 11**, a linear regression line can be generated from every scatter plot by utilizing Least Squares Method (LSM). These linear regression lines indicate the relationships between YAS and each of the three factors are nearly linear. However, as observed in Table 2, 3, and 4, there are two notable distinctions among these three linear regression lines:

- The slopes of the lines differ.
- The determination coefficient (R^2) of the models differ.

Regarding the differences in the slopes, let β denote the slope of a linear line. In this context, β_{ATA} and β_{ATS} are positive, whereas β_{ATW} is negative. This suggests a positive linear correlation between YAS and ATA as well as between YAS and ATS, and a negative linear correlation between YAS and ATW. In terms of magnitude, β_{ATA} is largest at 49.607, which indicates the change rate of YAS with ATA is greatest compared with ATS and ATW. Additionally, the slope of the linear regression line between YAS and ATW is very close to zero, making it difficult to determine whether the linear relationship between the two factors is positive or negative due to the inherent randomness of the experiment. More sets of experiments are required to obtain more reliable conclusions.
For determination coefficient, the determination coefficient between YAS and ATA is largest at 0.360, which means 36% of the variation in the YAS can

be explained by the linear regression model while the remaining 64% of the variation may be attributed to some lurking factors. In the contrast, the determination coefficient between YAS and ATW is lowest at 0.002, which indicates only 0.2% of the variation in the YAS can be explained by the linear regression model. Furthermore, the Pearson's correlation coefficient (r) can be obtained by taking the square root of determination coefficient as its magnitude and taking the sign of β as its sign. Calculations reveal that r_{ATA} is largest at 0.6, indicating a moderate positive correlation between the YAS and ATA, while r_{ATW} is smallest at -0.044, indicating a negligible negative correlation between the YAS and ATW. The r_{ATW} is in the middle at 0.438, indicating a weak positive correlation between the YAS and ATW. Given that both the β_{ATA} and r_{ATA} are largest compared with others', it can be concluded that average time on App has the most significant impact on the yearly amount spent. However, it must be noted that some lurking factors may influence these results, therefore, it is necessary to expand the sample size and conduct multiple trials to rule out this randomness.
The residual plot is used to ensure the validity of the regression models. Observing the Figures 12,13, and 14, it is evident that all residuals randomly distribute around zero without any noticeable curve or systematic pattern, which indicate the linear regression model is suitable for the selected data.

Two Sample Confidence Intervals Analysis

Comparing the Male customers and Female customers:

- Male Customers
    1. Mean: 488.686
    2. Standard Deviation: 59.147
    3. Sample Size: 25
- Female Customers
    1. Mean: 498.442
    2. Standard Deviation: 87.707
    3. Sample Size: 25

According to the equation (1)

$$\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 \pm t_k \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \tag{1}$$

in this case, 99% confidence level is required for higher precision, then, $\mu_{male} - \mu_{female} = (-68.933, 49.421)$. Since, the result includes zero, it can be concluded that *there is no significant difference between the YAS of male and female customers with 99% probability.*

Conclusion

Based on the analysis of the data, two main conclusions can be drawn: first, the among the three factors, average time on App has the most significant impact on the yearly amount spent; second, through the two sample confidence intervals analysis, it can be concluded that there is no significant difference in yearly amount spent by Males compared with Females. These conclusions imply that: first, the fashion and cosmetic brand should prioritize its investment in App construction, which impact the yearly amount spent most significantly. Also, they need to improve their service quality because its impact on yearly amount spent cannot be ignored. Conversely, investment in the website can be reduced since its impact on sales is negligible. Second, gender does not substantially impact the yearly amount spent, so the brand should equally consider the needs of both male and female customers to prevent customer attrition. Last but not least, some issues may affect the validity of the data analysis and conclusions. First, the sample size is too small; 50 samples are insufficient to fully represent the overall population, which may lead to results that are merely coincidental. Therefore, the experiment should be conducted with a larger sample size. Second, the experiment was conducted only once, which might lead to a chance result. Hence, multiple experiments should be performed to rule out the influence of random factors on the results.