7COM1079-0901-2024 - Team Research and Development Project

Final report title: Is there a difference in the proportion of highly rated products between young women and older women”

Group ID: A177

Dataset number: DS 147

Prepared by: *Roshwin Johny* - 23095204
Gowri Shankar Kamalakshan sugandhi – 23097222
Ajzal Bin Faisal Madathil – 23067047
Anandakrishna Sivaprabha – 230965639
Abhay Shankar Alakkal - 23107161

University of Hertfordshire
Hatfield, 2024

Table of Contents

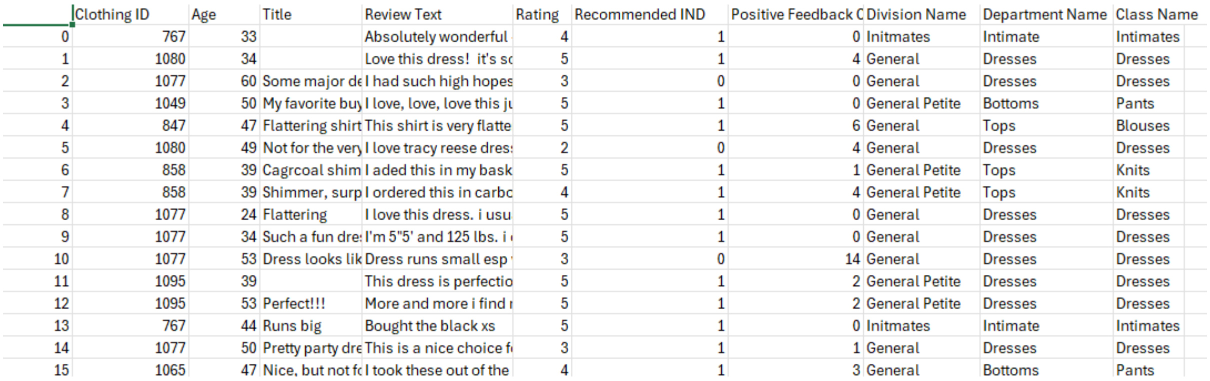
1. Introduction

1.1. Problem statement and research motivation

Customers Reviews plays a crucial role in maintaining an e-commerce business sector. These reviews help build a brands reputation and even in promoting the brands products. However after careful consideration of the dataset Women's E-commerce clothing reviews we found out that age plays an important role in product ratings. This leads us to our question "Is there a difference in the proportion of highly rated products between young and old women?".Prior studies suggest that consumer age significantly affects preferences and perceptions of clothing fit, service, and quality. Understanding this can result in valuable insights such as targeted marketing and other strategies and improve customer satisfaction

1.2. The data set

The dataset used for this research is "Women's E-Commerce Clothing Reviews," available on kaggle.com. This dataset revolves around the written reviews provided by customers. It explains how the ratings and feedback for a clothing product are distributed across different categories. The dataset includes the variables: Clothing ID, Age, Title, Review Text, Rating, Recommended Index, Division Name, Department Name, and Class Name of the products. It contains a total of 23,486 customer reviews and ratings.



	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback C	Division Name	Department Name	Class Name
0	767	33		Absolutely wonderful	4	1	0	Initmates	Intimate	Intimates
1	1080	34		Love this dress! it's sc	5	1	4	General	Dresses	Dresses
2	1077	60	Some major de	I had such high hopes	3	0	0	General	Dresses	Dresses
3	1049	50	My favorite buy	I love, love, love this ju	5	1	0	General Petite	Bottoms	Pants
4	847	47	Flattering shirt	This shirt is very flatte	5	1	6	General	Tops	Blouses
5	1080	49	Not for the very	I love tracy reese dres	2	0	4	General	Dresses	Dresses
6	858	39	Cagrcoal shim	I aded this in my bask	5	1	1	General Petite	Tops	Knits
7	858	39	Shimmer, surp	I ordered this in carbc	4	1	4	General Petite	Tops	Knits
8	1077	24	Flattering	I love this dress. i usu	5	1	0	General	Dresses	Dresses
9	1077	34	Such a fun dre	I'm 5"5' and 125 lbs. i	5	1	0	General	Dresses	Dresses
10	1077	53	Dress looks lik	Dress runs small esp	3	0	14	General	Dresses	Dresses
11	1095	39		This dress is perfectio	5	1	2	General Petite	Dresses	Dresses
12	1095	53	Perfect!!!	More and more i find	5	1	2	General Petite	Dresses	Dresses
13	767	44	Runs big	Bought the black xs	5	1	0	Initmates	Intimate	Intimates
14	1077	50	Pretty party dre	This is a nice choice f	3	1	1	General	Dresses	Dresses
15	1065	47	Nice, but not fo	I took these out of the	4	1	3	General	Bottoms	Pants

Fig 1. The Data set

1.3. Research question

“Is there a difference in the proportion of highly rated products between young women and older women ”

To answer this question, we categorized the dependent variable "Rating" into high (>=4) and low (<4) ratings, and selected high-rated products while categorizing the independent variable "Age" into young (<=40) and old (>40). We will create a contingency table to compute the p-value using the chi-square test. Then, we will check if the p-value is less than or greater than 0.05 to determine whether to reject or fail to reject the null hypothesis.

1.4. Null hypothesis and alternative hypothesis (H0/H1)

Null Hypothesis : *There is no difference in the proportion of highly rated products between young and old women.*

Alternative Hypothesis : *There is a difference in the proportion of highly rated products between young and old women*

We perform the chi-square test to calculate the p-value, which helps determine whether there is a statistical relationship between two categorical variables. If the resulting p-value is less than the significance level of 0.05, there is strong evidence to reject the null hypothesis. Rejecting the null hypothesis means the data supports the alternative hypothesis, indicating that age influences customer rating behavior.

2. Background research

2.1. Research papers

1) Data Analysis on Women's Clothing Reviews – Sabrina Lee

- This Research analysis the dataset to examine customer feedback by class and department. The analysis is structured into three main components

* Class and Department analysis

*Sentiment Analysis

*Predictive Analysis

The findings highlights the importance of product quality, size inclusivity and sentiment in driving recommendations and customer satisfaction

2) Sentiment Analysis of Women's Clothing Reviews on E-commerce Platforms: A Machine Learning Approach - by Masfiq Mahmud

- This Research Paper focuses on using machine learning for sentiment analysis of the dataset. It explored text preprocessing, feature extraction and model selection to classify reviews as positive, negative and neutral, which helps the business understand customer sentiment more effectively

***3) How the Age Impacts Women*'s Rating on Clothing Based on Women's Clothing E-Commerce Review**

- This research focuses on how age influences ratings. It examines demographic data, statistical analysis, and sentiment analysis to reveal how women of different ages rate clothing products. The findings provide insights into marketing strategies, customer segmentation, and more.

2.2. Why RQ is of interest (research gap and future directions according to the literature

It is interesting because, as a young adult and part of a significant consumer demographic, analyzing this data helps us understand how the opinions of younger generations differ from those of older ones. We found the research question intriguing because it addresses a gap in understanding how women's age influences product ratings. This gap is important since businesses can develop marketing strategies based on these insights. Future research could explore factors such as customers' income, geographic location, body structure, or other personal preferences, which may further impact product review patterns and lead to more personalized product recommendations.

3.Visualisation

3.1 Appropriate plot for the RQ

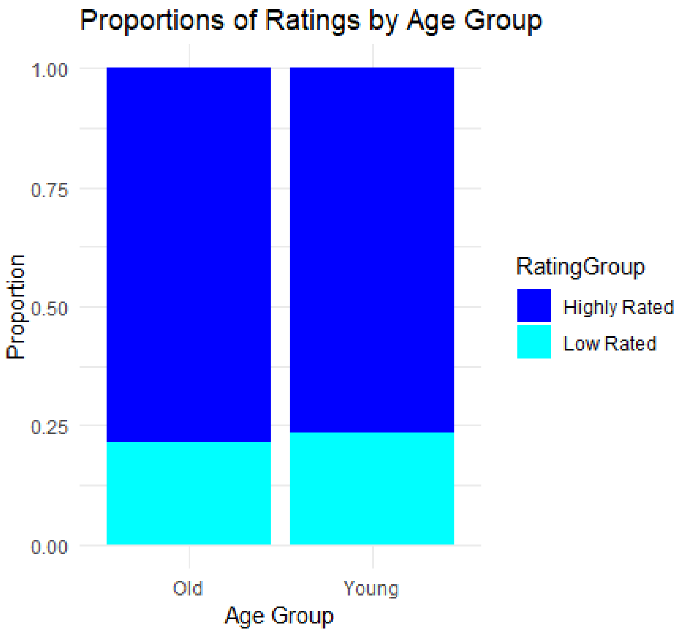


Fig 2 . Bar chart

Since both our independent variable and depended variable age and rating act as nominal variable. Our Research question focuses on the difference in proportions. Since it related to proportions, we have to visualize the data using a bar chart to represent the distribution which makes it easier to compare the proportions more effectively

3.2 Useful information for the data understanding

This Bar chart displays the proportions of ratings categorized by young and old age groups. are divided into highly rated and low rated. While Both age groups have a majority of highly rated the Old Aged women tend to give more high ratings than the younger category. This proves our alternative hypothesis.

4. Analysis

4.1. Statistical test used to test the hypotheses and output

For the analysis Chi-Square Test for independence is used, since this statistical test finds For the analysis, the Chi-Square Test for independence is used, as this statistical test determines whether there is a significant association between two categorical variables. In this case, the variables are age, categorized into young and old, and rating, categorized into high and low. The test compares the values in the contingency table and provides a p-value, which can be used to test the hypothesis. If the p-value is less than 0.05, we reject the null hypothesis.

4.2 The null hypothesis is rejected /not rejected based on the p-value

Null Hypothesis: "There is no difference in the proportion of highly rated products between young and old women."

The null hypothesis is rejected if the p-value is less than the significance level which is 0.05. This also means that there is a proof that the alternative hypothesis is accepted.

The following is our Chi-Square test results

Chi-squared test with Yates' continuity correction

data: contingency_table
Chi-Square Statistic (X^2) = 10.448
Degrees of Freedom (df) = 1
p-value = 0.001228

Since the p-value is 0.001228 which is less than the significance level (i.e., 0.05) we **Reject** the null hypothesis

5. Evaluation

5.1. What went well

The group effectively collaborated and contributed to the Research, which lead to understanding each other's capabilities and skills. Clear and scheduled communications and meetings were made to meet the project goal quickly and effectively. The team members encouraged each other's and mutual support helps us to overcome the challenges and finish the research and documentations before the deadline. Overall working as a group helped us to know each other and improved everyone's capability to work as a team

5.2. Points for improvement

Even the team successfully finished the work in time, there were some areas that needs improvements which could improve future efforts. The unclear project role assessment resulted in overlapped tasks and missed some tasks which lead to a confusion in the project. The feedbacks provided by the tutors were greatly appreciated but feedbacks given among other groups were very less, it was essential for finding the common doubts about the Research and Structure of the Project

5.3. Group's time management

Time management was not satisfactory during the initial stage due to poor scheduling made by the team Members. But the group managed to overcome this challenge and finish the task at time. The group had some last minute rush to submit the Initial Tasks of the Project and the improved time management among the members and helped us to submit the projects final tasks before the deadline

5.4. Project's overall judgement

The Project was a Success and the group was fantastic. We overcome many challenges including a phase where we had to change the research question multiple times. The Groups strong collaboration and efforts helped to finish the project in time and improved Skills such as time management, communication, presentation among the members

5.5. Changes to group since submission of Assignment 1

Not Applicable

5.6. Comment on the GitHub log output

1) **Commit Message**: barchart new

Explanation: This commit, Authored by *gshankar631*, introduced a new bar chart visualization.

2) **Commit Message**: analysis chi-square

Explanation: This commit, authored by *gshankar631*, implemented the chi-square analysis.

3) **Commit Message**: Research Question Template

Explanation: This commit, authored by ROSHWIN JOHNY, added a presentation on the research questions.

6. Conclusions

6.1. Results explained

The findings of this research prove that age plays a significant role in customer reviews, with older women tending to give more high ratings to clothing products than younger women. This result was identified using statistical analysis, specifically a Chi-Square test, which provides evidence that age does influence customer ratings. The analysis revealed a significant difference in the proportion of highly rated products between young and old women, proving that age affects the way clothing products are reviewed. These findings contribute valuable insights into understanding customer behavior based on age.

6.2. Interpretation of the results

The result of this research analysis shows that age influence on customer ratings, with older women are more likely to give higher ratings to clothing products than younger women. This findings shows the result to our research question and these result can be used by the marketers and retailers to understand the priority of age in marketing, which can lead to better customer satisfaction and can influence sales strategies

6.3. Reasons and/or implications for future work, limitations of your study

As we analyzed the dataset, future work could explore other demographic factors, such as geographic location and income, to better understand how these factors affect clothing product reviews. Limitations of this study include a focus solely on age and clothing products, potentially overlooking other important factors.

7. Reference list

1)**Lee, S. (2022)** 'Data Analysis on Women Clothing Reviews'.

2)**Mahmud, M., Mullick, R.A. & Anas, M. (2023)** 'Sentiment Analysis of Women's Clothing Reviews on E-commerce Platforms: A Machine Learning Approach'.

3) **Li, S. & Wang, Z. (2022)** 'How the Age Impacts Women’s Rating on Clothing Based on Women’s Clothing E-Commerce Review', *Proceedings of the 2022 6th International Seminar on Education, Management and Social Sciences (ISEMSS 2022)*, pp. 2156–2167.

8. Appendices

A. R code used for analysis and visualisation

Visualisation (R Code)

```
# Load required libraries
library(ggplot2)
library(dplyr)
# Load your datadata <- read.csv("C:/Users/user/OneDrive/Desktop/Womens Clothing E-Commerce Reviews.csv")
# Classify age groups
```

```
data$AgeGroup <- ifelse(data$Age <= 40, "Young", "Old")
# Classify rating groups
data$RatingGroup <- ifelse(data$Rating >= 4, "Highly Rated", "Low Rated")
# Create a contingency table
contingency_table <- table(data$AgeGroup, data$RatingGroup)
# Calculate proportions for visualization
proportions <- prop.table(contingency_table, margin = 1)
proportions_df <- as.data.frame(as.table(proportions))
colnames(proportions_df) <- c("AgeGroup", "RatingGroup", "Proportion")
# Plot proportions using ggplot2
ggplot(proportions_df, aes(x = AgeGroup, y = Proportion, fill = RatingGroup)) +
geom_bar(stat = "identity", position = "fill") +
labs(
title = "Proportions of Ratings by Age Group",
x = "Age Group",
y = "Proportion"
) +
scale_fill_manual(values = c("Low Rated" = "cyan", "Highly Rated" = "blue")) +
theme_minimal()
```

Analysis (R code)

```
# Load required libraries
library(ggplot2)
library(dplyr)
# Load your data
data <- read.csv("C:/Users/user/OneDrive/Desktop/Womens Clothing E-Commerce Reviews.csv")
# Classify age groups
data$AgeGroup <- ifelse(data$Age <= 40, "Young", "Old")
# Classify rating groups
data$RatingGroup <- ifelse(data$Rating >= 4, "Highly Rated", "Low Rated")
# Create a contingency table
contingency_table <- table(data$AgeGroup, data$RatingGroup)
```

```
# Perform the chi-square test
chi_square_test <- chisq.test(contingency_table)
# Print results
print("Contingency Table:")
print(contingency_table)
print("Chi-Square Test Results:")
print(chi_square_test)
```

B. GitHub log output.

```
'f7df057,2025-01-06T22:42:54Z,ROSHWIN JOHNY,Add files via upload'
'f856f4a,2025-01-06T21:28:55Z,anandakrishna485,Update README.md'
'1cb3fdf,2025-01-06T21:28:14Z,anandakrishna485,Update README.md'
'7a35096,2025-01-06T21:06:46Z,ROSHWIN JOHNY,Update README.md'
'489bf44,2025-01-05T11:42:24Z,anandakrishna485,Add files via upload'
'71110ba,2025-01-03T23:39:29Z,gshankar631,barchart new'
'd348b0c,2025-01-03T21:28:40Z,gshankar631,analysis chi-square'
'27374a3,2025-01-03T21:25:34Z,gshankar631,just space on 3rd line'
'd72917e,2025-01-03T21:22:32Z,gshankar631,new barchart'
'5056c42,2025-01-03T21:19:44Z,gshankar631,bar plot'
'5e0cd3e,2025-01-03T20:25:28Z,gshankar631,Add Rplot files to .gitignore'
'287d7af,2025-01-03T20:23:59Z,gshankar631,Save local changes to analysis, scatterplott, and team scripts'
'75b38e3,2024-12-13T14:31:14Z,ROSHWIN-JOHNY,Update README.md'
'6f21c57,2024-12-13T14:30:44Z,ROSHWIN-JOHNY,Add files via upload'
'f4e9995,2024-12-13T14:07:28Z,ROSHWIN-JOHNY,Add files via upload'
'892f272,2024-12-11T11:44:44Z,ROSHWIN-JOHNY,Add files via upload'
'66e01de,2024-12-04T09:33:46Z,anandakrishna485,Update README.md'
'5b98b8d,2024-11-25T10:51:09Z,gshankar631,Add files via upload'
'b0bf62a,2024-11-25T06:45:53+05:30,ROSHWIN-JOHNY,Add files via upload'
'6a11808,2024-11-24T21:00:03Z,gshankar631,Add files via upload'
'3e9d217,2024-11-24T15:17:10Z,ajzalfaizal,Update analysis.R'
'bcef5a9,2024-11-24T14:12:25Z,ajzalfaizal,Update README.md'
```

```
'cbe4473,2024-11-24T13:51:44Z,ajzalfaizal,Update README.md'
'40ed677,2024-11-24T19:21:22+05:30,ROSHWIN-JOHNY,Add files via upload'
'3add1fb,2024-11-24T13:48:32Z,ajzalfaizal,Update README.md'
'189cb86,2024-11-24T19:18:16+05:30,ROSHWIN-JOHNY,Update README.md'
'22d2868,2024-11-24T19:15:07+05:30,ROSHWIN-JOHNY,Update README.md'
'e77c9a0,2024-11-24T13:39:20Z,ajzalfaizal,Update histogram.R'
'7e6c888,2024-11-24T13:38:57Z,ajzalfaizal,Update histogram.R'
'0839bb4,2024-11-24T13:32:23Z,ajzalfaizal,Update scatterplott.R'
'e461105,2024-11-24T13:31:12Z,ajzalfaizal,Update README.md'
'c4b27d6,2024-11-24T13:29:50Z,ajzalfaizal,Update README.md'
'2b91212,2024-11-24T18:52:26+05:30,ROSHWIN-JOHNY,Add files via upload'
'f81cc76,2024-11-24T18:40:47+05:30,ROSHWIN-JOHNY,Add files via upload'
'34e1a1d,2024-11-23T22:11:38Z,abhaysa10prog,Add files via upload'
'58ec7df,2024-11-22T19:44:41Z,gshankar631,analysis of s,pvalue,alt hypothesis'
'4d1b731,2024-11-22T16:55:58Z,gshankar631,unfilled circle'
'bf21be8,2024-11-22T16:40:34Z,gshankar631,histogram'
'cc78d42,2024-11-22T16:38:33Z,gshankar631,Merge branch 'main' of
https://github.com/ajzalfaizal/Team-Research-Group-A177'
'774e02d,2024-11-22T16:37:51Z,gshankar631,scatterplott'
'af21af8,2024-11-22T15:54:06Z,gshankar631,scatterplot'
'90750fb,2024-11-22T20:46:50+05:30,ROSHWIN-JOHNY,Add files via upload'
'4f7a0cc,2024-11-22T15:11:08Z,gshankar631,Update README.md'
'8618480,2024-11-22T15:00:49Z,ajzalfaizal,Update README.md'
'0fe044d,2024-11-22T14:57:24Z,ajzalfaizal,Update README.md'
'd4bf888,2024-11-21T22:09:46Z,gshankar631,person's r and p value'
'bae5add,2024-11-21T15:48:18Z,gshankar631,changes'
'f78efb5,2024-11-21T15:44:34Z,gshankar631,dataset analysis code'
'8153d76,2024-11-21T15:43:36Z,gshankar631,Merge branch 'main' of
https://github.com/ajzalfaizal/Team-Research-Group-A177'
'fe90b22,2024-11-21T15:42:40Z,gshankar631,dataset analysis code.'
'47c53a2,2024-11-21T15:24:17Z,gshankar631,Add files via upload'
'5b891c6,2024-11-21T15:16:29Z,gshankar631,code'
'c592dec,2024-11-21T14:33:57Z,gshankar631,research question ppt'
'5cf4b30,2024-11-15T14:44:53Z,ajzalfaizal,Initial commit'
```