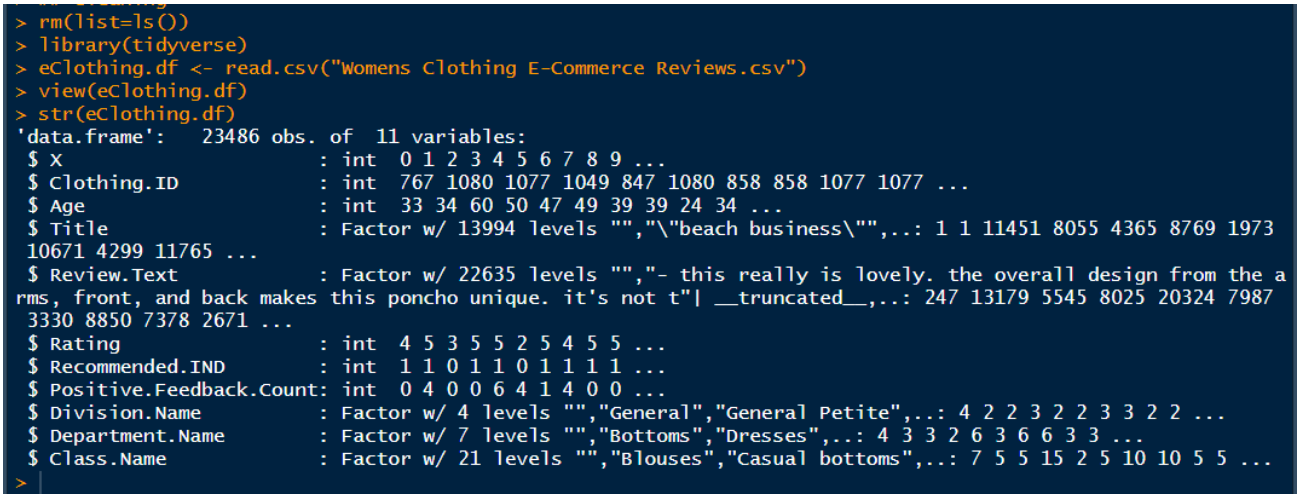
Women's E-Commerce Clothing Review Project – Version 2.
--Data Exploration and Visualization

There is a typical list of steps for data mining are as follows: [1]

1. Set the goals of the data mining projects
2. Get the data to analyze
3. Explore, clean, and preprocess the data
4. Shorten the data scale
5. Determine the data mining task (e.g. classification, prediction, clustering, etc.)
6. Partition the data (for supervised tasks)
7. Select the most critical, creating potential data mining techniques
8. Use the right algorithms to perform the task
9. Interpret the results of the algorithms
10. Choose the right model to deploy

The project objectives were made and the dataset was obtained in the first version. This project version provides implement data exploration, including exploring, cleaning, and preprocess the data. The data exploration is done with visual graphics. The 'Women's E-Commerce Clothing Review' dataset [2] is explored in R Studio. The first step is to clean the workspace with rm(list = ls()) command then upload library packages will be used later. Library(tidyverse) is including the base packages for data exploration and visualization in R, such as dplyr package, is for data manipulation, tidyr for data clean, and ggplot for data graphics creating. [4] [5]

After uploading the dataset with read.csv(), I checked the data structure and viewed it with the results shows below in the screenshot.



```
> rm(list=ls())
> library(tidyverse)
> eClothing.df <- read.csv("Womens Clothing E-Commerce Reviews.csv")
> view(eClothing.df)
> str(eClothing.df)
'data.frame':   23486 obs. of  11 variables:
 $ X                      : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Clothing.ID            : int  767 1080 1077 1049 847 1080 858 858 1077 1077 ...
 $ Age                    : int  33 34 60 50 47 49 39 39 24 34 ...
 $ Title                  : Factor w/ 13994 levels "","\"beach business\"",..: 1 1 11451 8055 4365 8769 1973
 10671 4299 11765 ...
 $ Review.Text            : Factor w/ 22635 levels "","- this really is lovely. the overall design from the a
rms, front, and back makes this poncho unique. it's not t"| __truncated__,..: 247 13179 5545 8025 20324 7987
 3330 8850 7378 2671 ...
 $ Rating                 : int  4 5 3 5 5 2 5 4 5 5 ...
 $ Recommended.IND        : int  1 1 0 1 1 0 1 1 1 1 ...
 $ Positive.Feedback.Count: int  0 4 0 0 6 4 1 4 0 0 ...
 $ Division.Name          : Factor w/ 4 levels "","General","General Petite",..: 4 2 2 3 2 2 3 3 2 2 ...
 $ Department.Name        : Factor w/ 7 levels "","Bottoms","Dresses",..: 4 3 3 2 6 3 6 6 3 3 ...
 $ Class.Name             : Factor w/ 21 levels "","Blouses","Casual bottoms",..: 7 5 5 15 2 5 10 10 5 5 ...
>
```

Then applied is. na() function to check the missing value and got all FALSE, which means it is a clean dataset. Based on the project's objectives and bar graphs showing the relationships between different data series independent of each other[5], I started with the bar chart to explore variables: "Age, Rating, Recommended.IND, Division.Name and Class.Name."

The plot below shows the women at the age of 38 – 40 most likely to give online clothing shopping comments.

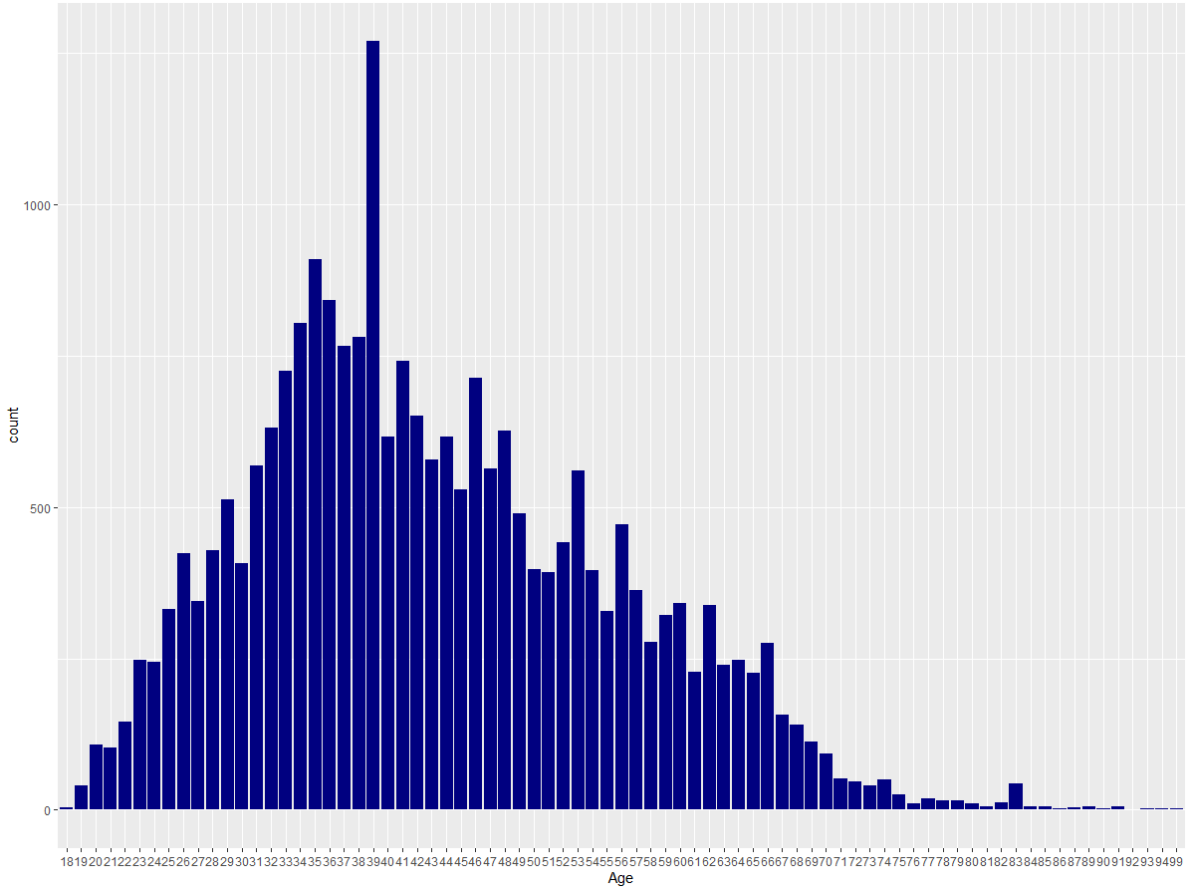


Figure 1.

Figure 2. bar chart shows the higher rating, and the more recommend it will make. Figure 3. bar chart represents that the positive recommended with 1 is around four times than the negative recommended with 0, which is a big found for this dataset.

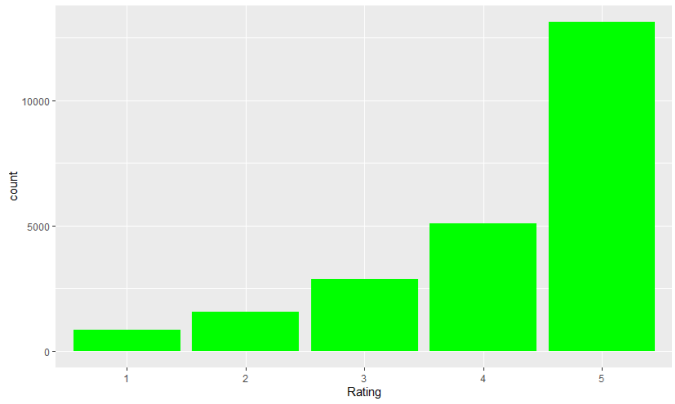


Figure 2.

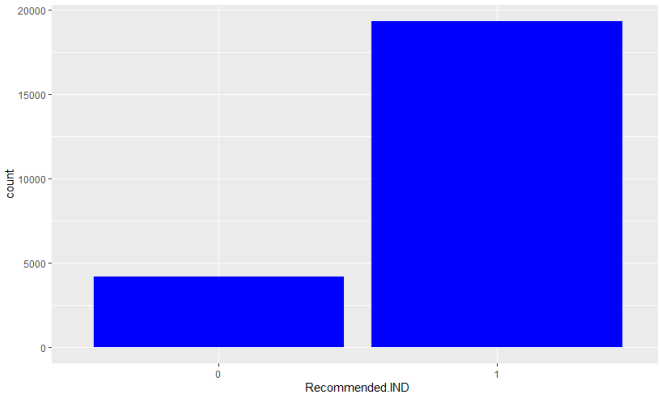


Figure 3.

The bar chart (Figure 4.) demonstrated clearly that the higher rating clothing got recommend to the others more likely.

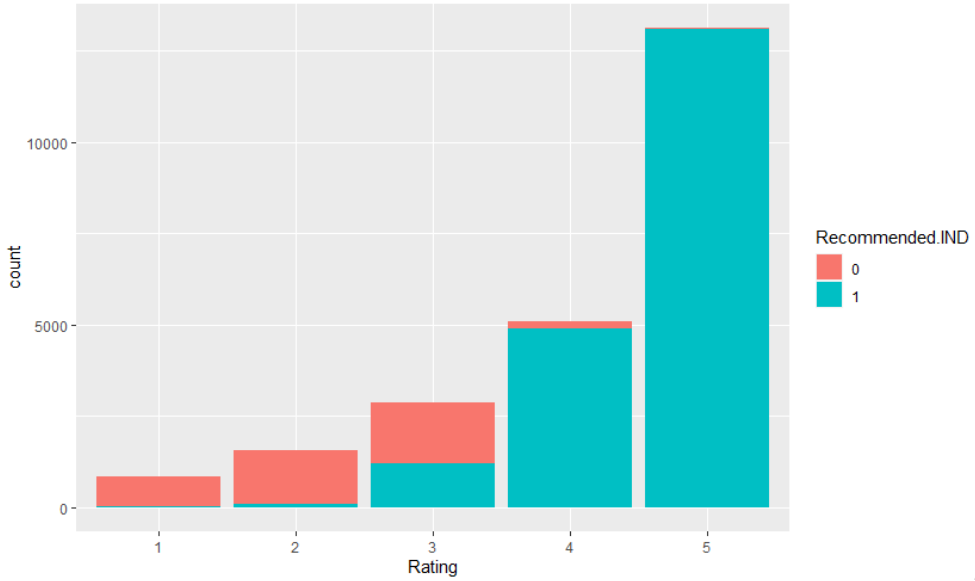


Figure 4.

The three plots below present the distribution by the department.name, division.name, and class.name on Recommended category. On the first bar chart – Figure 5., the department named Tops in purple color has the most positive recommendation, and the dress is the second proportion among them. The negative recommended bar shows Tops is also the biggest part of it. Figure 6. plot indicated the General's division name has the most proportion in positive and negative recommend. The third bar chart (Figure 7.) shows that Layering's class got the most significant proportion recommended both in positive and negative. The conclusion is the same type of clothing that gets the most positive comment will get the most negative recommendations.

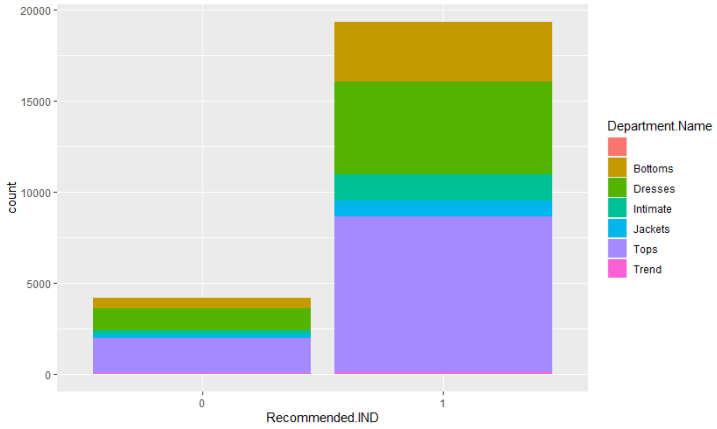


Figure 5.

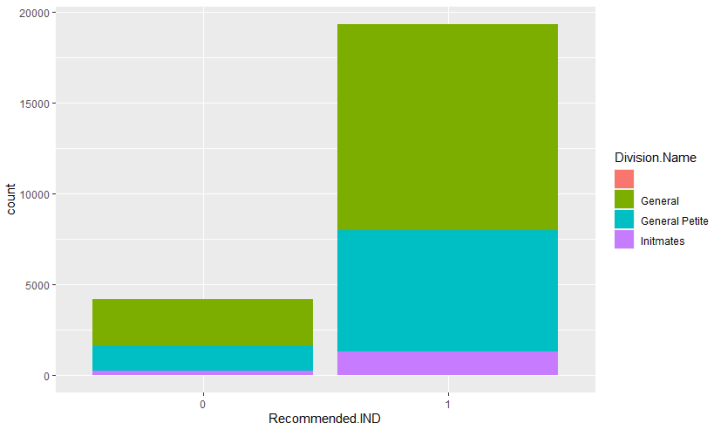


Figure 6.

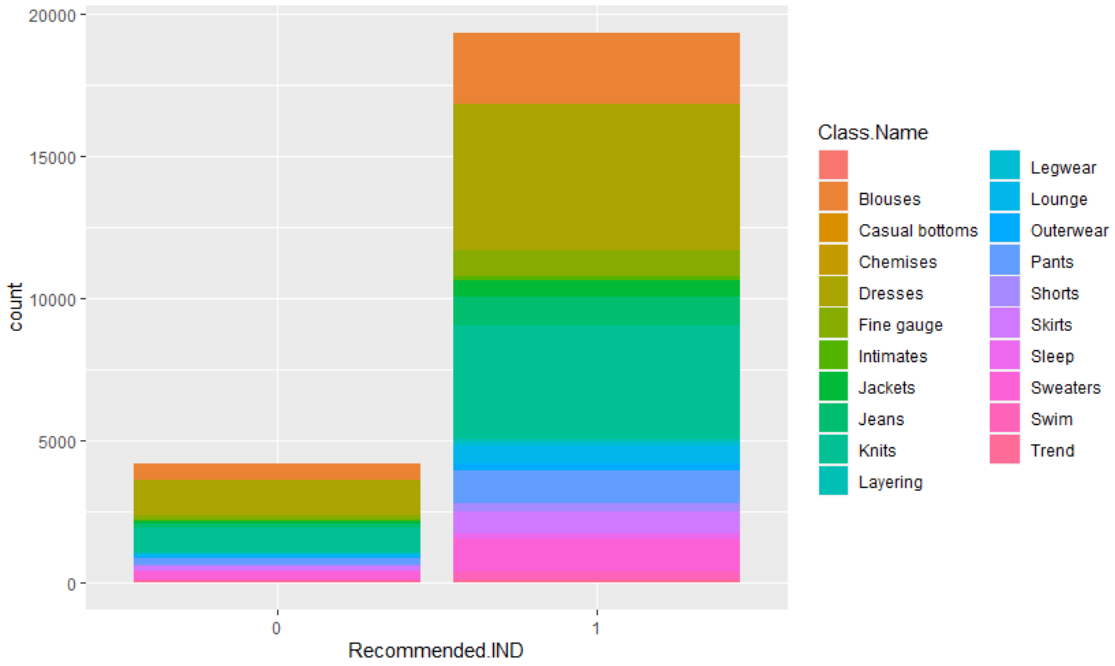


Figure 7.

The positive feedback graphic – Figure 8. Below shows most clothing got 1 positive feed back for the online shopping. Figure 9. highlights among the positive feedback, most of them are recommended.

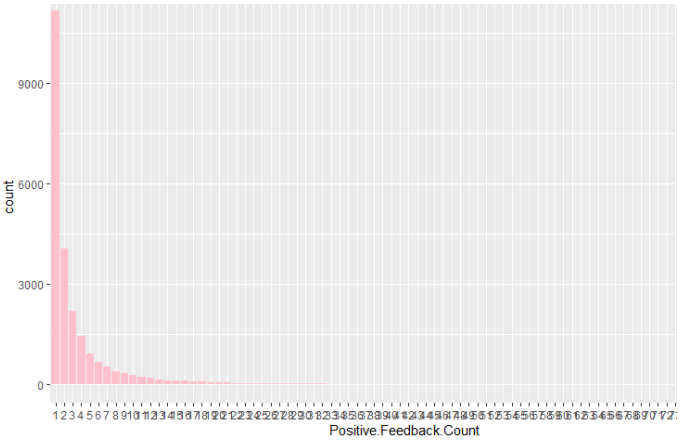


Figure 8.

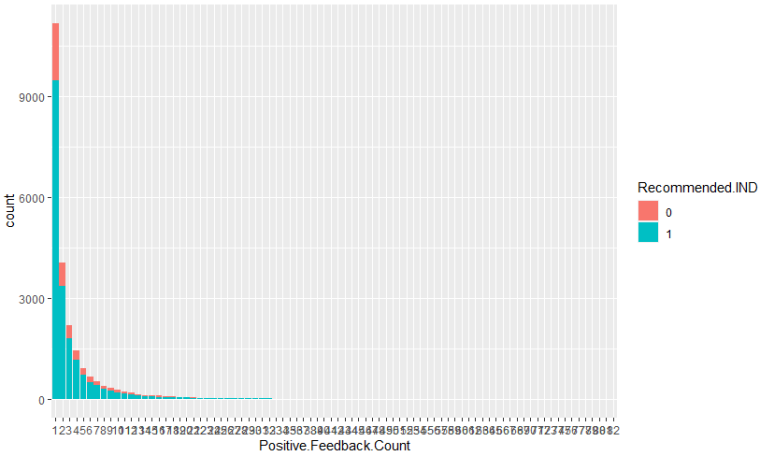


Figure 9.

Figure 10. contingency table - listed the rating counts on recommended positive 1 and negative 0.

	Rating				
Recommended	1	2	3	4	5
0	826	1471	1682	168	25
1	16	94	1189	4909	13106

Figure 10.

The summary for the contingency table (Figure 11.) shows p-value < 0.05, which means it can be a good model to deploy.

```
Number of cases in table: 23486
Number of factors: 2
Test for independence of all factors:
        Chisq = 16723, df = 4, p-value = 0
```

Figure 11.

The mosaic plot (Figure 12) below illustrated the rating identifying above 2.5 got most recommended. [6]

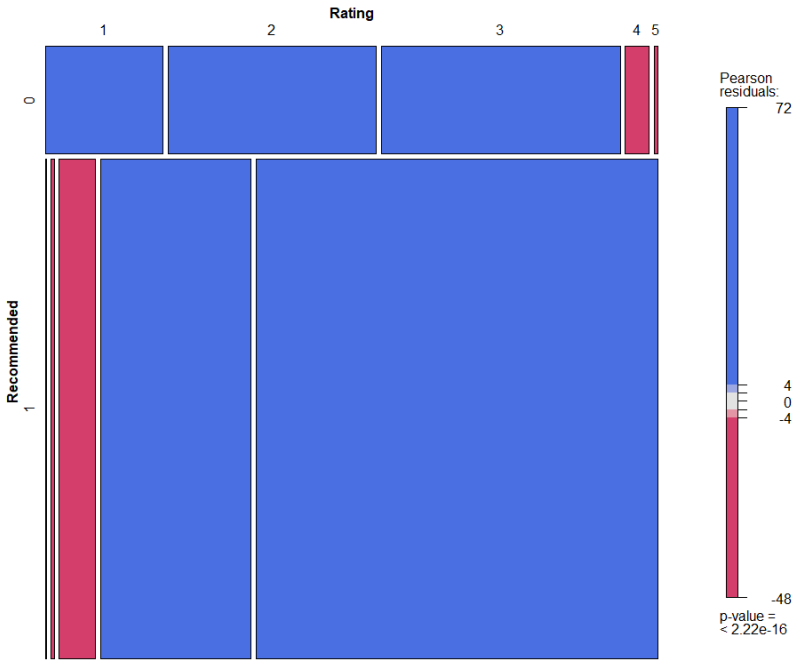


Figure 12.

Figure 13. is the contingency table showing the matrix of recommendation in the department.

```
            Department
Recommended      Bottoms Dresses Intimate Jackets Tops Trend
          0    0     565    1212      260     169 1935    31
          1   14    3234    5107     1475     863 8533    88
>
```

Figure 13.

The summary for the contingency table (Figure 14.) shows p-value < 0.05, which means it can be a good model to deploy.

```
Number of cases in table: 23486
Number of factors: 2
Test for independence of all factors:
        Chisq = 53.29, df = 6, p-value = 1.024e-09
        Chi-squared approximation may be incorrect
```

Figure 14.

The mosaic plot below in Figure 15. illustrated the positive recommended indicated with 1 for tops got the highest recommendation.

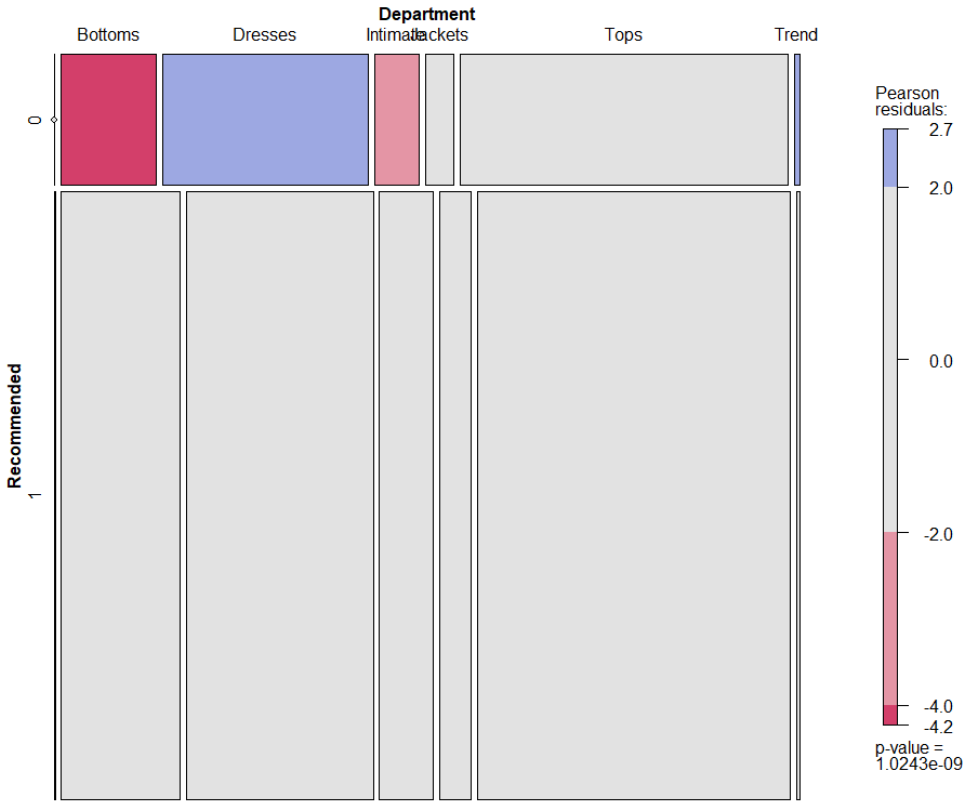


Figure 15.

References

[1]. Shmueli, G., Bruce,P., Yahav, I., Patel, N. and Lichtendahl, K. 2018. Overview of the Data Mining Process. Data mining for business analytics: concepts, techniques, and applications in R, John Wiley &Sons, Inc. Wiley, NJ, USA, 19 - 20.
[2]. nicapotato, 2017. Women's E-Commerce Clothing Reviews. Kaggle https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews/version/1.
[3]. Tidyverse packages. https://www.tidyverse.org/packages/.
[4]. How to describe charts, graphs, and diagrams in the presentation. Preply. https://preply.com/en/blog/2018/08/17/charts-graphs-and-diagrams-in-the-presentation/#scroll-to-heading-1.
[5]. ggplot2. Wikipedia. https://en.wikipedia.org/wiki/Ggplot2 https://towardsdatascience.com/data-cleaning-with-r-and-the-tidyverse-detecting-missing-values-ea23c519bc62.
[6]. Zavarella, L. 2018. Mosaic Plot and Chi-Square Test. Towards data science. https://towardsdatascience.com/mosaic-plot-and-chi-square-test-c41b1a527ce4.