Data Mining Assignment 2

Submitted By:

Dileshwori Joshi (28385)

Importing the file

✓ Firstly, all the environment variables are removed from the environment.

✓All the required packages and libraries are installed and loaded which consists of tidyverse, arules and arulesviz.

✓tidyverse package is used here for data manipulation, and tidying.

✓arules and arulesviz package are used for association mining and visualizing association respectively.

✓Then the csv file is loaded, checked and analyzed.

```
#loading the package for data manipulation, tidy and visualization
library(tidyverse)

#installing and loading package for association mining and association visualization
install.packages("arules")
library(arules)
install.packages("arulesViz")
library(arulesviz)

#importing data
marketdata <- read.csv2('exercise_WS2020_data.csv')
marketdata
```

Fig 1: code snippet for installing packages and loading the file

Data Conversion

- ✓ The basket data is converted as factor to input as parameters to the split function.
- ✓ After the split is done, the data frame is converted to transactions to ease the association rule mining.

```
#converting basket as a factor
marketdata$article_group <- factor(marketdata$article_group)
marketdata$article_group

#converting baskets into different kinds of objects
shopping_basket <- split(marketdata$article_group,marketdata$basket_id)
shopping_basket
retail_transactions <- as(shopping_basket, "transactions")
retail_transactions
summary(retail_transactions)
```

Fig 2: Code snippet for conversion of the data into transaction type

Data Visualization

- ✓ The absolute item frequency of the top 30 items in the transactions is shown in the figure at the right.
- ✓ The kitchen gadgets and dishes are highest bought item and household helpers is 30th frequently bought item.
- ✓ So, to increasing the sales of cutlery items, retailer can put it near kitchen gadgets.
- ✓ The summary(retail_transactions) function gave the top five items as Kitchen gadgets, dishes, lighting, bathroom accessories and bags with the frequencies 20250, 16898, 15312, 14375, and 12680 respectively.

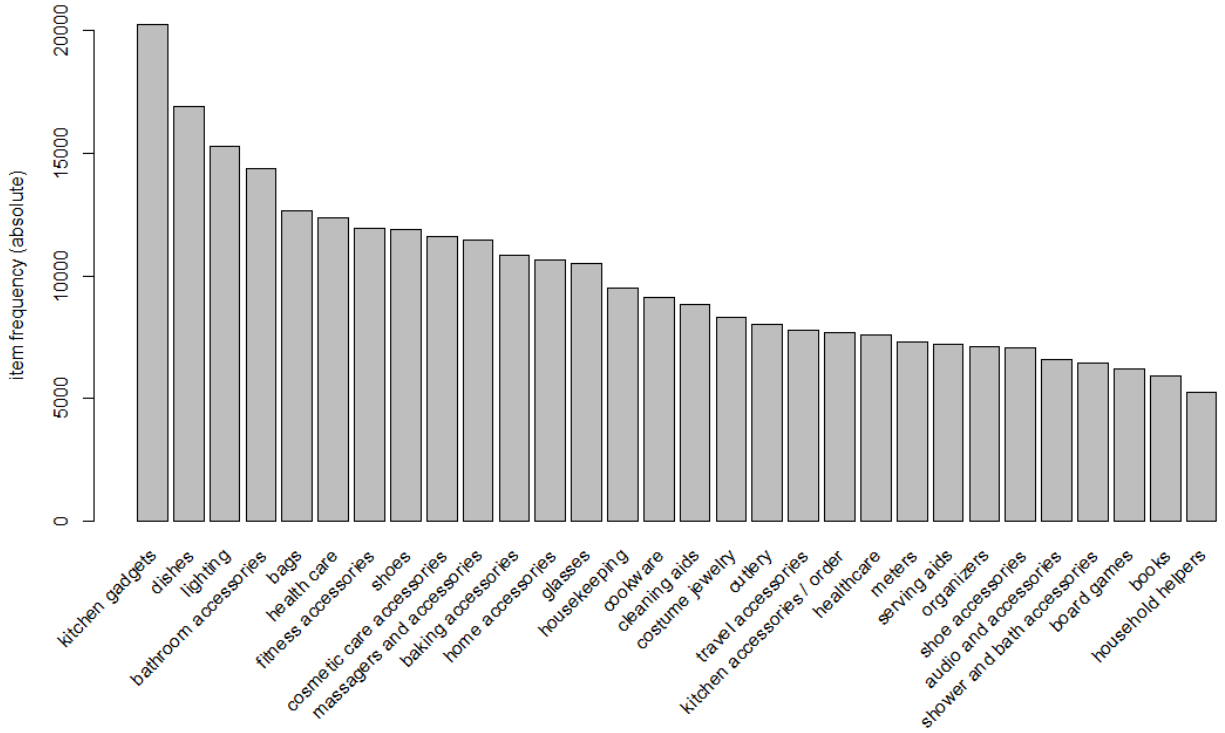

Fig 3: Data visualization for top 30 frequently bought items

Association Rule visualization - I

- ✓ The scatter plot illustrates the 85 rules produced by apriori algorithm at support=0.001 and confidence=0.1
- ✓ The darker the color of the circle, higher the value of the lift.
- ✓ According to fig 4, the higher the correlation between the itemset (lift value), lower is the support.

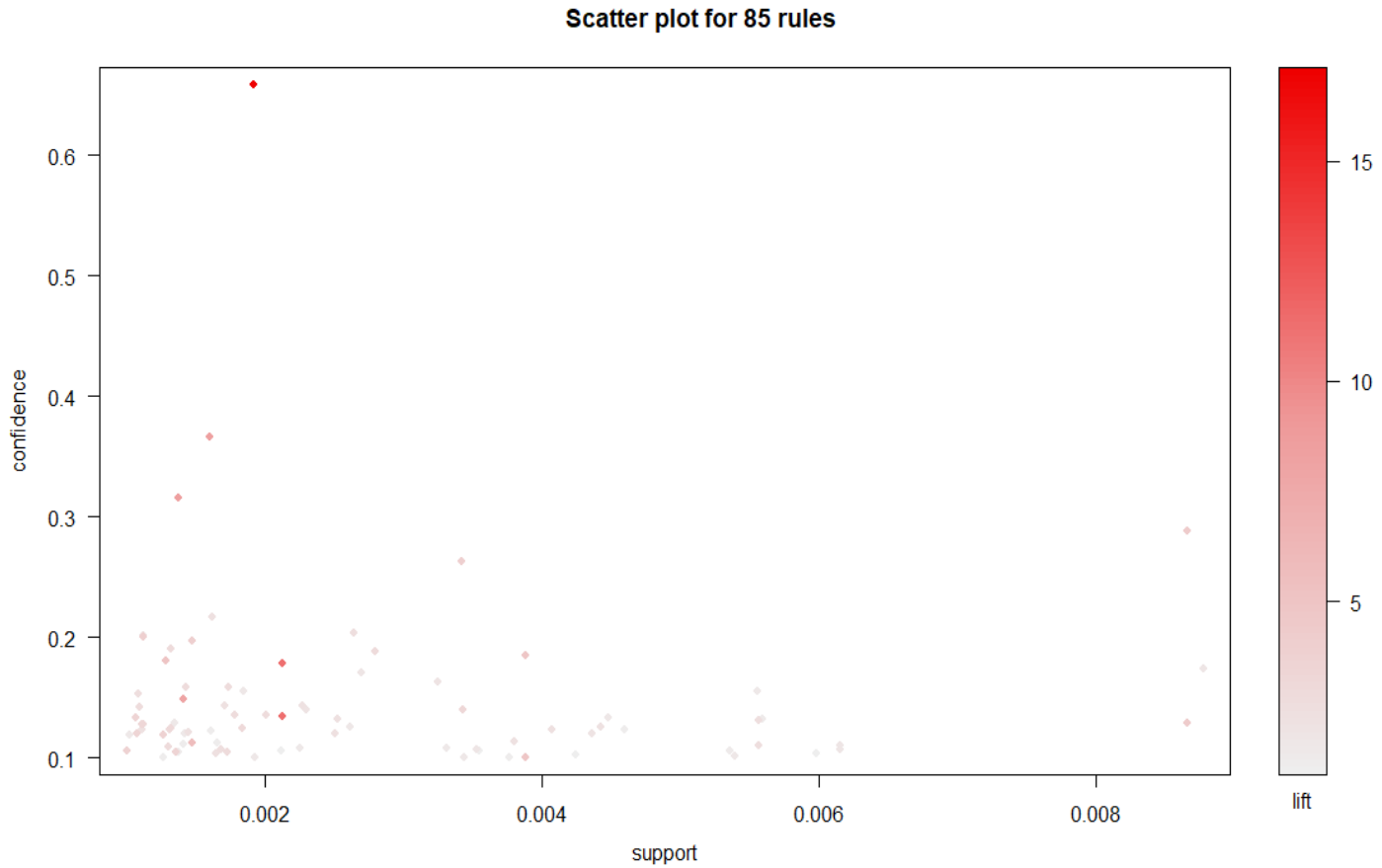

Fig 4: Visualization of association rules with threshold support=0.001 and confidence=0.1

Association Rule Visualization - II

- ✓ The scatter plot illustrates the 3 rules produced by apriori algorithm at support=0.002 and confidence=0.2.
- ✓ It shows only 3 rules at the minimum threshold support of 0.002 and min. threshold confidence of 0.2.
- ✓ There is only one rule with high lift according to the fig 4.

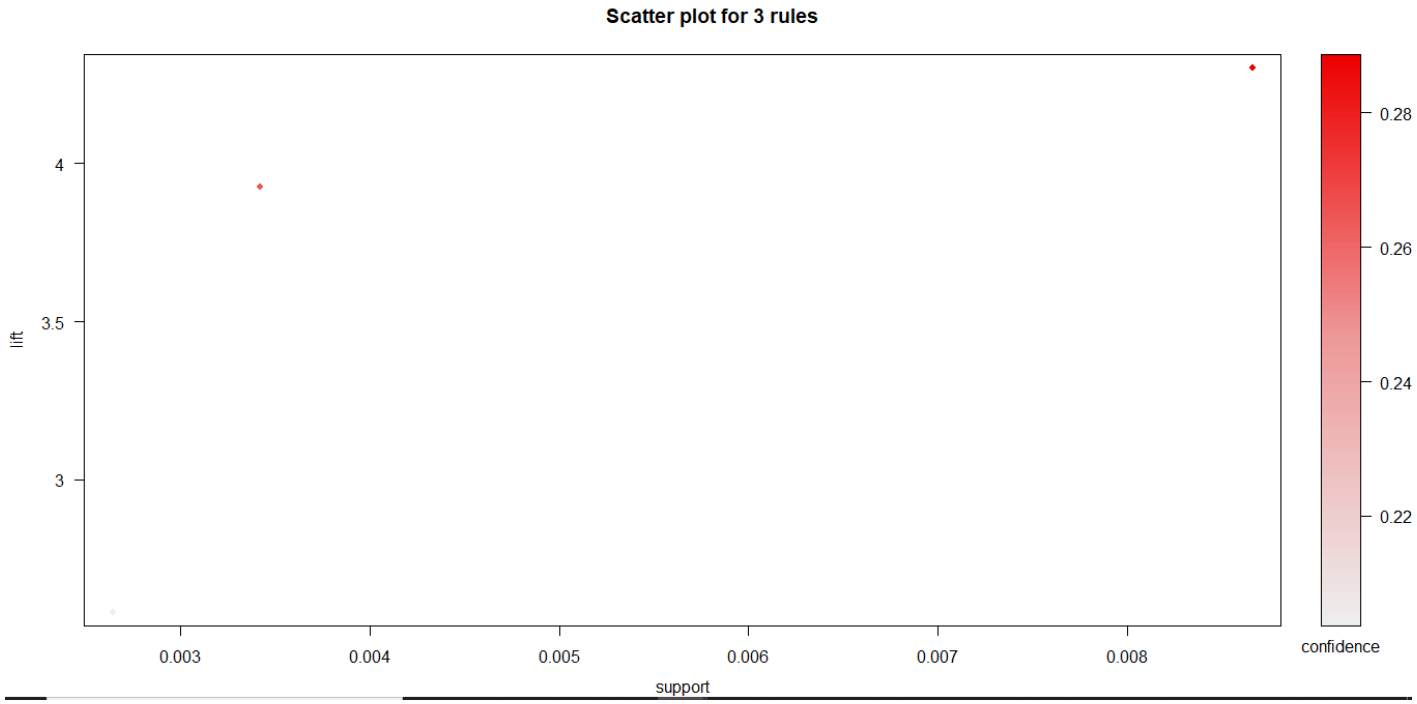

Fig 4: Visualization of association rules with threshold support=0.002 and confidence=0.2

Thank You