

Department of Computer Science and Electrical Engineering

CMSC 491 – Introduction to Data Science Assignment 4 Student: Sanaa Mironov

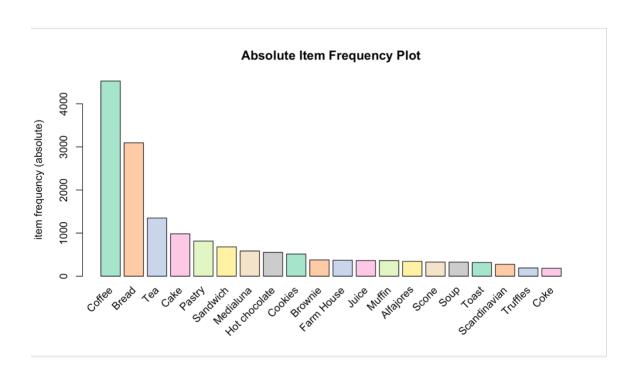
1. Background: The Bread Basket dataset lists all transactions taking within one cafe. We are using to determine the transactions that happen and to list all the specific items in the dataset. The dataset must if from a café shop. The form of data collection is through a point of sale tool. This data is useful for analyzing the most popular item sold and identifying the business's busiest time, week, and season. With this type of data, there is so much to analyze, predicate, and improve the business. Some examples include what kind of customers they have; are they repeat customers or new every time. Is the customer staying in the cafe or come back later in the day. Learn about how the customer tips. The data collected can determine the customer service experience, and with the right information, create a better marketing strategy to improve and increase sales and return customer rate.

We can observe that the dataset collects the Date, Time, Transaction Number, and Item.

From our data we can see that the most frequent item bought is Coffee followed by bread:

```
most frequent items:
Coffee Bread Tea Cake Pastry (Other)
4526 3094 1349 983 814 8105
```

The Bar graph displays the most frequent item bought, coffee.



2. Given the set of transactions from the café shop, the five rules that will predict an item's occurrence based on the occurrences of other items in the transaction include. The summary of the association rules is below. The data mining information has an 80% confidence using the data transaction:

```
> summary(association.rules)
set of 7 rules
rule length distribution (lhs + rhs):sizes
2 3
2 5
  Min. 1st Qu. Median
                          Mean 3rd Qu.
                                         Max.
  2.000 2.500
               3.000
                         2.714 3.000
                                        3.000
summary of quality measures:
   support
                    confidence
                                        lift
                                                       count
Min.
      :0.001056
                         :0.8095
                                   Min. :1.693
                                                   Min. :10.00
                   Min.
1st Qu.:0.001321
                   1st Qu.:0.8246
                                   1st Qu.:1.725
                                                   1st Qu.:12.50
Median :0.001479
                   Median :0.8333
                                   Median :1.743
                                                   Median :14.00
Mean :0.002203
                   Mean :0.8415
                                   Mean :1.760
                                                   Mean :20.86
3rd Qu.:0.002430
                   3rd Qu.:0.8619
                                   3rd Qu.:1.803
                                                   3rd Qu.:23.00
Max.
       :0.005388
                   Max.
                          :0.8750
                                   Max. :1.830
                                                   Max.
                                                        :51.00
mining info:
data ntransactions support confidence
  tr
              9466 0.001
                                 0.8
```

The five association rules set for this dataset:

- 1. Extra salami or Feta has an 81% chance they will buy a coffee.
- 2. keeping it local, there is an 80% chance they will buy a coffee as well.
- 3. cake, vegan, mince pie there is an 83% chance they will also buy a coffee.
- 4. Extra salami or Feta, salad there is an 87% chance they will buy a coffee.
- 5. Heart & Seasonal, a sandwich with an 85% chance to buy a coffee.

Check below for the output:

Definition of support, confidence, and lift to explain how it is used in association rules:

Support: this measure gives an idea of how frequently an *item* is in all the transactions. Consider *itemset1* = { Extra salami } and *itemset2* = {coffee}. There will be far more transactions containing coffee than those containing extra salami. Support is the fraction of the total number of transactions in which the itemset occurs. The number of items in the transaction we have is about 10,000. Rule number 2 the support is 50%. If the support for an association rule is low, we can state that we do not have enough information to predict a relationship.

Confidence: This measure defines the likelihood of predicting the item, given that the site already has the antecedents. Looking at the output if we have {Keeping it Local }, how many also bought {Coffee} with them? We can say by the common knowledge that {Keeping it Local } \rightarrow {Coffee} should be a high confidence rule. Technically, confidence is the conditional probability of occurrence of consequent given the antecedent.

Lift: Lift controls for the *support* (frequency) of items while calculating the conditional probability of occurrence of {Y} given {X}. *Lift* is the rise in the probability of having {Y} on the item with the knowledge of {X} being present over the probability of having {Y} on the cart without any knowledge about the presence of {X}. The value of lift will be greater than one is complimentary, and less is harmful, and it means increasing the chances of occurrence. Lift is used to predict if someone buys one thing the probability they will also buy coffee with it.