Market Basket Analysis Assessment

-Suhita Goswami

1. Problem statement:

The dataset is provided by a retailer and comprises of 100K rows of transactions. Our goal is the find the group of items that are purchased together.

1. Understanding and preparing the dataset:

1. There are fifty items.
2. Each row represents a transaction.
3. '1' represents the item being bought in that transaction.
4. The dataset is first read in as a csv file into a variable DF.
5. There are no NA values present in the data set.
6. The values are converted to Boolean form so as to allow conversion to a sparse matrix format. This is required for the data to be seen as transactional data and be analyzed using the required packages.
7. Drop the id column by setting it to null.
8. The transactional data is now stored and accessed through the variable 'Raw'

2. Exploratory Analysis:

1. Using the packages: arules, arulesViz
Convert dataframe to a transactional sparse matrix

Density = Total number of non empty cells/ total number of cells
Density = total no of items purchased / total number of items possible in the matrix = 0.0701566

```
> summary(Raw)
transactions as itemMatrix in sparse format with
 100000 rows (elements/itemsets/transactions) and
 50 columns (items) and a density of 0.0701566

most frequent items:
 item_2  item_7 item_29  item_5 item_22 (Other)
  28326   28304   28302   28244   28167  209440

element (itemset/transaction) length distribution:
sizes
   0    1    2    3    4    5    6    7    8    9   10   12
1186 4105 13891 36084 21681 16048 5626 1141  210   25    2    1
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   3.000   3.508   4.000  12.000

includes extended item information - examples:
  labels variables levels
1 item_0    item_0   TRUE
2 item_1    item_1   TRUE
3 item_2    item_2   TRUE

includes extended transaction information - examples:
  transactionID
1           1
2           2
3           3
```

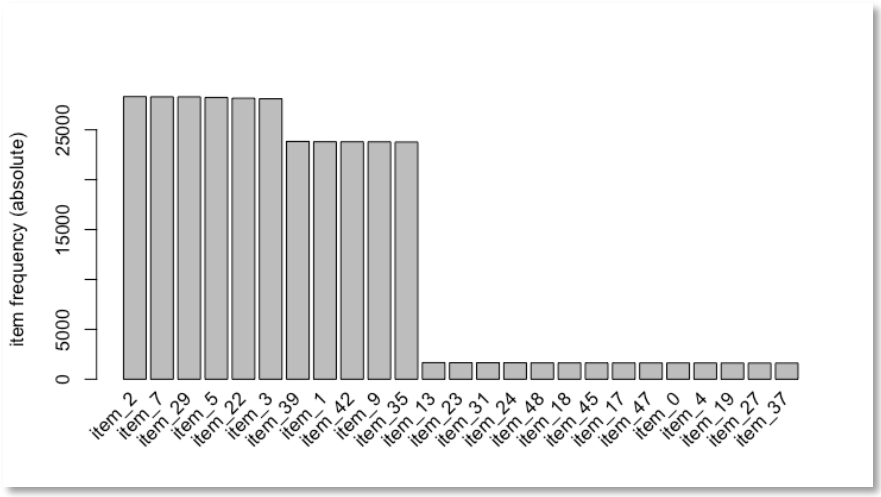
Observations:
Density*rows*columns = 350783 items purchased in all the transactions.
The most frequently purchased item at the store:
item_2 (28326 transactions)

Verification: Summing up the item count from the summary,

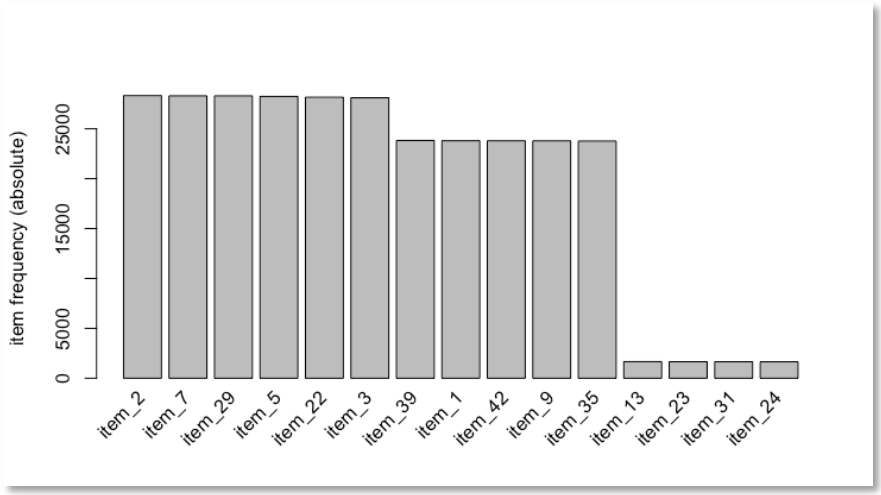
28326 + 28304 + 28302 + 28244 + 28167 + 209440 = 350783

As the mean (3.508) is greater than the median (3.0), the distribution is **right skewed**. Also, it is noted that as the 1st quartile is 3 and 3rd quartile is at 4, roughly 50% of the transactions have 3 to 4 items.

Item Frequency Plots:



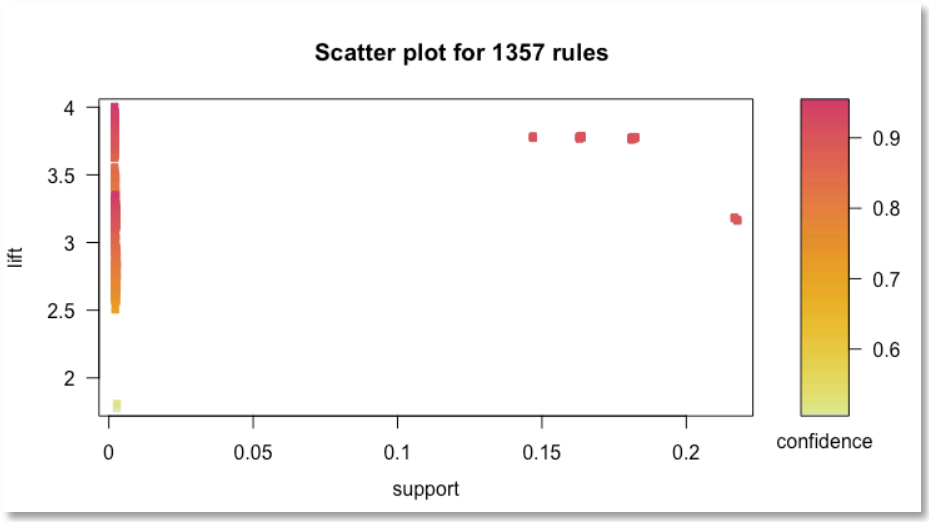
First plotting the top 25 items with the highest occurrence in the transactions.



We now narrow down to the top fifteen. Here the item frequency is in terms of number of transactions. As these are the most commonly bought items, there is a higher probability of finding patterns between these items.

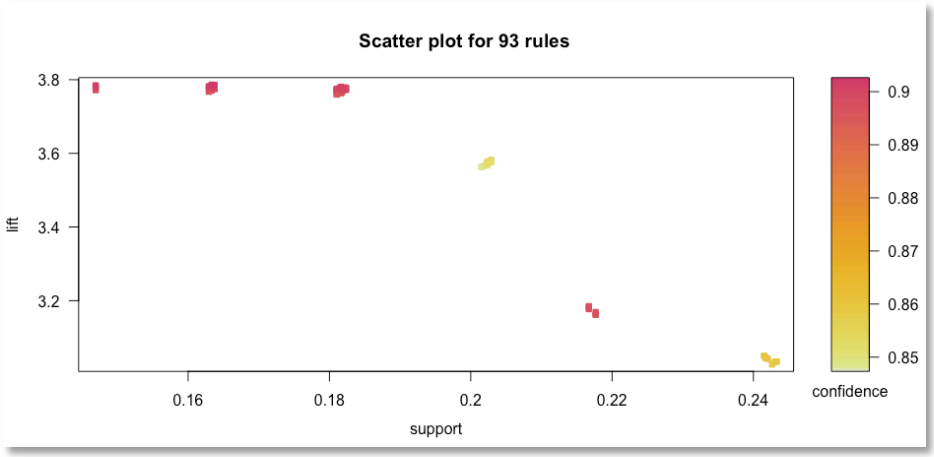
3. Data Modeling:

Rules:
Let us first generate a set of rules with support as 0.002 and confidence as 0.5. We get 1357 rules. On plotting a scatter plot to better visualize these rules, we get the following scatterplot:



We see that most of the rules from this set fall in a very low support range, implying that there are fewer transactions that follow these rules. Thus we can increase our support range to identify rules that will be of greater benefit to us.

The new set of rules has support of 0.1, i.e. the transactions which occur is more than 10% of the total transactions and confidence above 70%. This generates 93 rules.



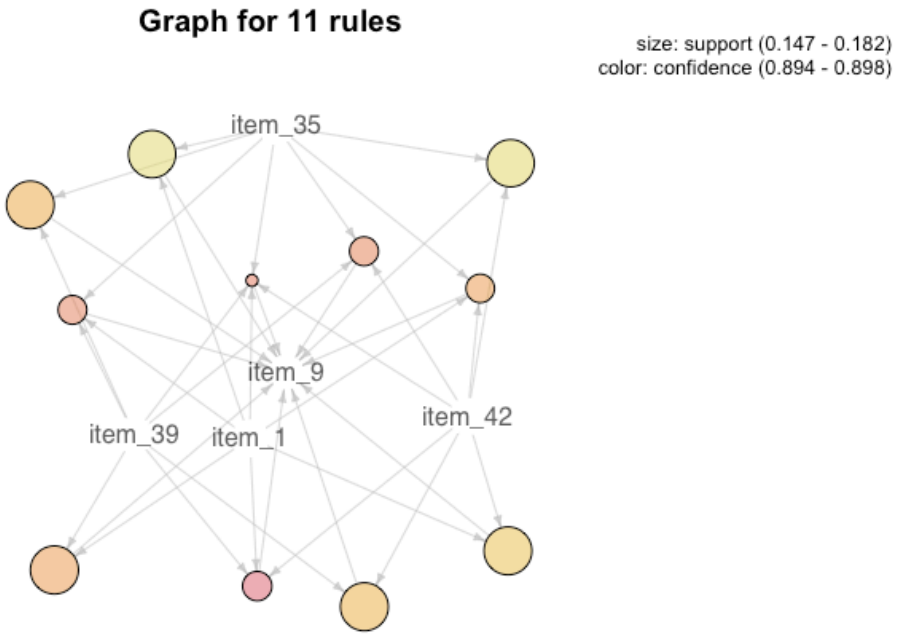
There are 93 rules having an element distribution between 2 and 5. As at average, transactions have three or four items at a time, let us filter out only those rules having three or four items. Also, the item of highest frequency is item_2, so let us include that in the filter as well:
We see there is only one rule with these conditions,

If Item_7, item_29 are being bought, item_2 may be bought as well.
There is strong support as over 20% transactions from the provided data show this and the confidence of item_2 being bought when item_7 and item_29 are bought is nearly 90%.

Similarly, some of the rules are as follows, in order of confidence:

lhs	rhs	support	*confidence	lift
73 {item_9, item_35, item_42} =>	{item_39}	0.16332	0.9023703	3.786699
87 {item_1, item_9, item_39} =>	{item_42}	0.16374	0.9008087	3.785547
68 {item_1, item_39} =>	{item_42}	0.18235	0.8991174	3.778439
76 {item_9, item_39, item_42} =>	{item_35}	0.16332	0.8986959	3.780481
84 {item_1, item_39, item_42} =>	{item_35}	0.16374	0.8979435	3.777316
33 {item_3, item_22} =>	{item_5}	0.21670	0.8972713	3.176856
35 {item_5, item_22} =>	{item_3}	0.21670	0.8954175	3.185179
44 {item_1, item_9} =>	{item_35}	0.18108	0.8953718	3.766498

We can generate a graph keeping one item at the center and observing how the sales of other items is related to it.



Here taking item_9 at the center, we see the interaction of the different items based on association rules with it. The size of the bubbles indicates how often these items appear in transactions with one another. These graphs allow trends and patterns to be more apparent and ease the process of decision making based on the insights acquired from the data.
Similarly, item_5, item_22 and item_3 tend to be bought together and are also in the higher Item frequency range in the dataset.

We can also mine the data to find relationships to boost the sales of other items which are not selling as well but are usually bought with a higher selling item.
For instance, restructuring our rule set by reducing the minimum support to 0.002 to accommodate more transactions to be observed while keeping the confidence level at 90.

We get the following observation:

	lhs	rhs	support	confidence	lift
118	{item_3, item_4, item_22} =>	{item_5}	0.00240	0.9302326	3.293558
119	{item_4, item_5, item_22} =>	{item_3}	0.00240	0.9160305	3.258504

We can use this information to boost the sales of item_4 as item_3, item_5 and item_22 are high selling items by offering discounts on item 4 when bought with item_3 or item_5. Although there are fewer transactions to support this, the confidence of them being bought together is very high. Similar observations are seen for other items as well. Thus, this can be considered as a useful decision making insight. Offers such as “buy two get one free” or “buy one get 50% off on the second” would prove useful in boosting the sales when using these insights to design these campaigns.

Key Insights:

1. Item_2, which has the highest frequency in transactions, is bought usually with item_7 and item_22
2. Item_9 may be bought with a variety of of other items, item_35, item_39, item_1 and item_42
3. Item_3, item_5 and item_22 are always bought together.

Conclusion:

In this manner several patterns can be found in the dataset.
The above rules can be seen as useful information for decision making in terms of providing discounts for one item when bought with another to encourage their collective sales. Another decision can be to have the items stocked closer together in the store or sending discount coupons keeping the above rules in mind to customers.
Thus, several patterns can be revealed by mining data which can aid in clearly decision making.