Market Basket Analysis for French retail store

Submitted by

Soumita Bandyopadhyay (211390), Soumya Paul (211391)

Under the guidance of

Prof. Amit Mitra

Department of Mathematics and Statistics, IIT Kanpur

Motivation

We have a Market Basket Dataset in which different products are given as the transactions over the course of a week at a French retail store. We wish to conduct the Market Basket Analysis on this data.

The main aim of Market Basket Analysis is

- to list items that are frequently purchased together,
- represent relationships in terms of association rules,

for example, if a customer buys bread, then he is likely to buy milk as well.

Association Rule Mining

- For any given set of transactions, our aim is to find a set of rules such that

1. support ≥ min sup
2. confidence ≥ min conf, where min conf is the predefined minimum confidence threshold value.

Here (min sup, min conf) depends on how much risk one wants to take.

We can conduct the Association Rule Mining by two methods- The Brute force method or the Apriori algorithm.

Since the former method is computationally prohibitive and hence we prefer

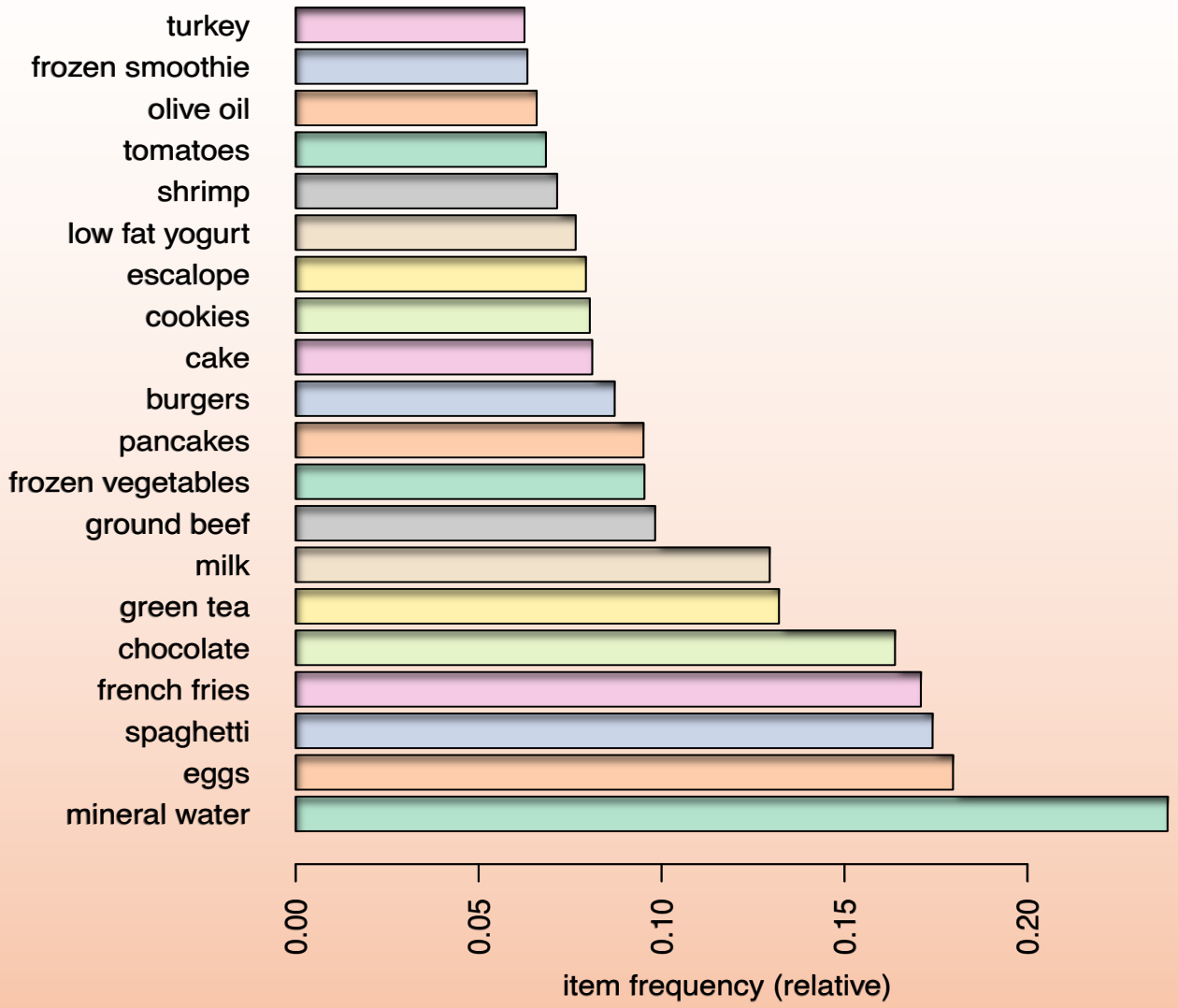
the Apriori Algorithm.

Dataset Exploration

- There are 7501 transactions and the maximum number of items purchased is 119.

```
    items
[1] {almonds,
     antioxydant juice,
     avocado,
     cottage cheese,
     energy drink,
     frozen smoothie,
     green grapes,
     green tea,
     honey,
     low fat yogurt,
     mineral water,
     olive oil,
     salad,
     salmon,
     shrimp,
     spinach,
     tomato juice,
     vegetables mix,
     whole weat flour,
     yams}
[2] {burgers,
     eggs,
     meatballs}
[3] {chutney}
[4] {avocado,
     turkey}
[5] {energy bar,
     green tea,
     milk,
     mineral water,
     whole wheat rice}
[6] {low fat yogurt}
```

Most frequent item sets



Results by applying Apriori Algorithm

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{frozen smoothie, spinach}	=>	{mineral water}	0.001066524	0.8888889	0.001199840	3.729058	8
[2]	{bacon, pancakes}	=>	{spaghetti}	0.001733102	0.8125000	0.002133049	4.666587	13
[3]	{nonfat milk, turkey}	=>	{mineral water}	0.001199840	0.8181818	0.001466471	3.432428	9
[4]	{ground beef, nonfat milk}	=>	{mineral water}	0.001599787	0.8571429	0.001866418	3.595877	12
[5]	{mushroom cream sauce, pasta}	=>	{escalope}	0.002532996	0.9500000	0.002666311	11.976387	19
[6]	{milk, pasta}	=>	{shrimp}	0.001599787	0.8571429	0.001866418	11.995203	12

We conclude the following:

- There are 8 baskets which satisfy the {frozen smoothie, spinach} -> {mineral water}
- 0.1066524 % of baskets satisfy the {frozen smoothie, spinach} -> {mineral water}
- In 88.88889% of the cases {mineral water} appears in transactions containing {frozen smoothie, spinach}
- 0.001199840 is the fraction of baskets that have the items {frozen smoothie, spinach}
- Consumers purchasing {frozen smoothie, spinach} are 3.729058 times more likely to buy {mineral water} than randomly selected consumers

Sorting by quality measures

```
> inspect(sort(data_rules, by="lift", decreasing=TRUE)[1:4])
    lhs                                  rhs         support     confidence coverage    lift     count
[1] {eggs, mineral water, pasta}      => {shrimp}    0.001333156 0.9090909  0.001466471 12.72218 10
```

The rule {eggs, mineral water, pasta} $\Longrightarrow$ {shrimp} has the highest lift value of 12.72218 which implies that consumers purchasing eggs, mineral water, and pasta together are 12.72218 times more likely to buy shrimp than randomly selected consumers.

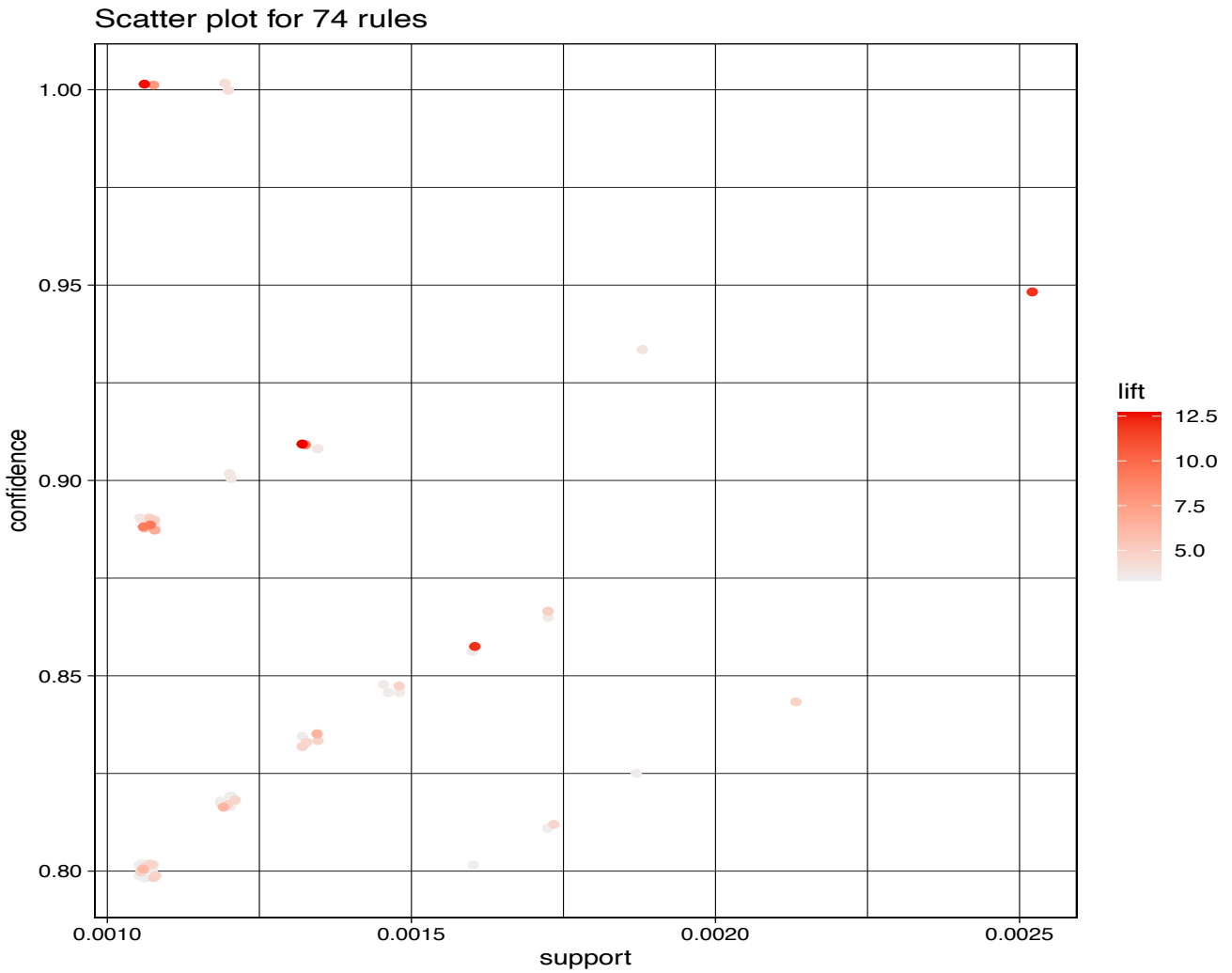
```
> inspect(sort(data_rules, by="confidence", decreasing=TRUE)[1:4])
    lhs                                        rhs            support     confidence coverage    lift      count
[1] {french fries, mushroom cream sauce, pasta} => {escalope}  0.001066524 1          0.001066524 12.606723 8
```

The rule {mushroom cream sauce, pasta} $\Longrightarrow$ {escalope} has the highest support value of 0.002532996 which implies that the itemset {mushroom cream sauce, pasta, escalope} has about 0.2532% chance of occuring in the whole transaction set.

Sorting by quality measures

```
> inspect(sort(data_rules, by="support", decreasing=TRUE)[1:4])
    lhs                                  rhs             support     confidence coverage    lift      count
[1] {mushroom cream sauce, pasta}     => {escalope}      0.002532996 0.9500000  0.002666311 11.976387 19
```

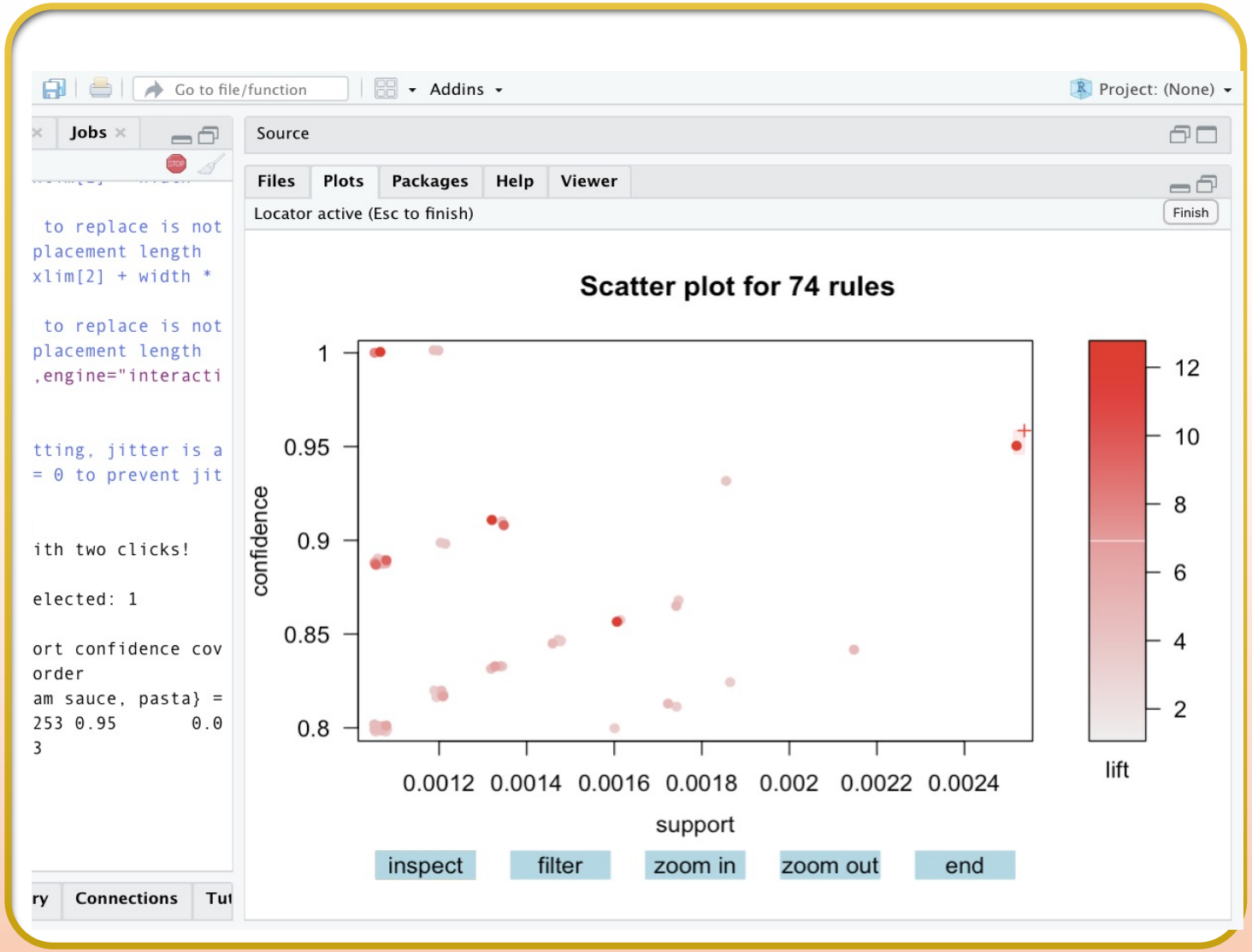
The rule {french fries, mushroom cream sauce,pasta} => {escalope} has the confidence value of 1 which implies that escalope always appears in transactions containing french fries, mushroom cream sauce and pasta together.



Scatter Plot

Scatter plot

- We can say that there is one rule whose support and confidence is moderately high.
- So, that rule is more most interesting than any other rules as it also possesses high lift value.
- The rule is {mashroom cream sauce, pasta} ⟹ {escalope}. We can see it from interactive plot.



Grouped matrix -Based visualization

Cluster Antecedents with k-means clustering algorithm using lift values as measures

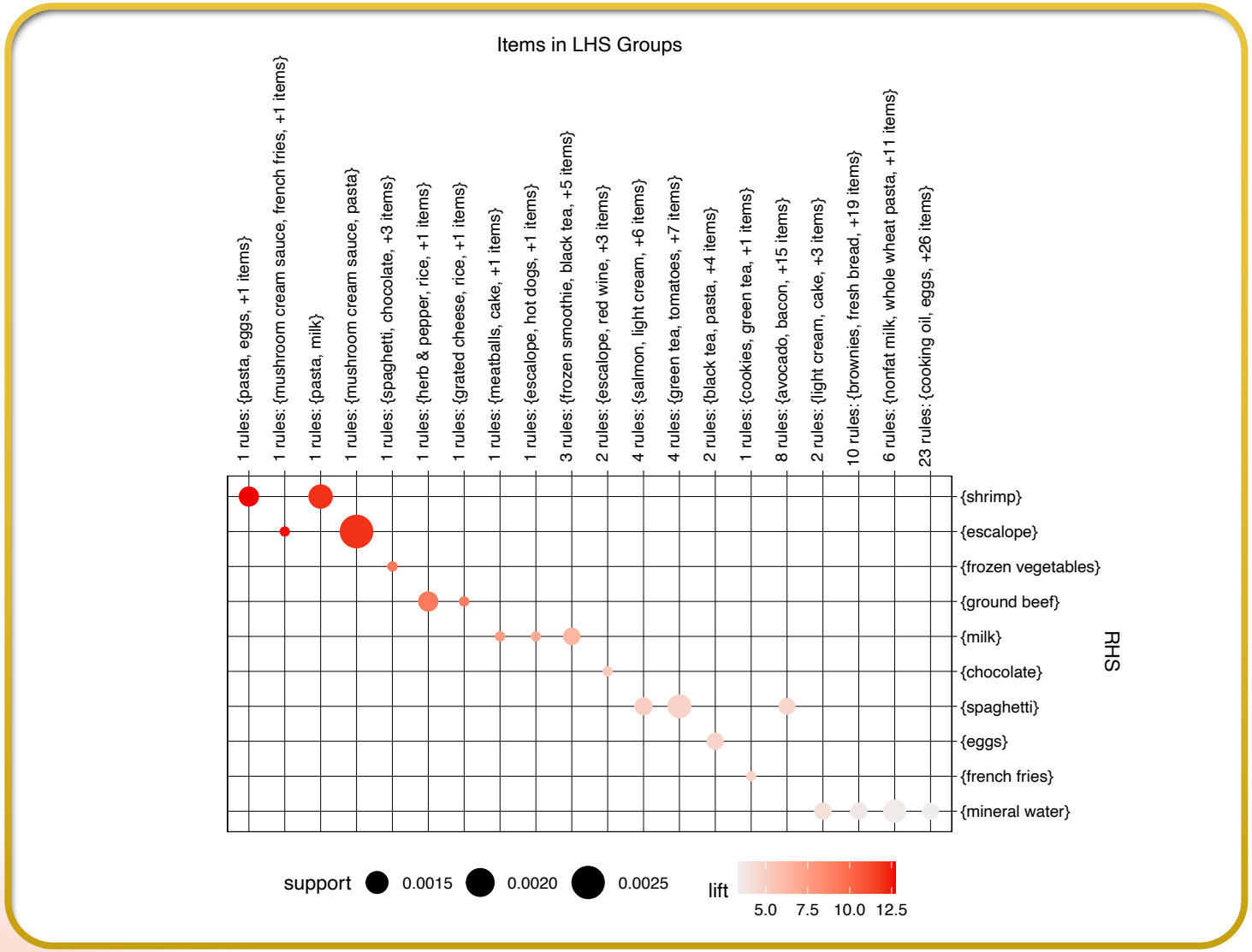
Form the collection of (Grouped antecedents, Consequent, aggregated lift, aggregated support) using median as aggregation method

Sort the collection with respect to aggregated lift values in decreasing order

Put the grouped antecedents into columns and consequents into the rows of a balloon plot

The colour and size of the balloons represent the aggregated lift and support respectively

Grouped matrix -Based visualization



Graph-Based visualization

- Here we are visualizing association rules using vertices and edges where vertices annotated with item labels represent items, and item sets or rules are represented as a second set of vertices.
- Items are linked to groups of items/rules using arrows.
- For rules arrows pointing from items to rule, vertices indicate LHS items and an arrow from a rule to an item indicates the RHS.