Final Year Project Presentation
MARKET BASKET ANALYSIS

By Group 15

Ossama Bin Hassan (1805210032)
Shivam Sachan (1805210048)
Shubham Gupta (1805210052)

Supervised by:

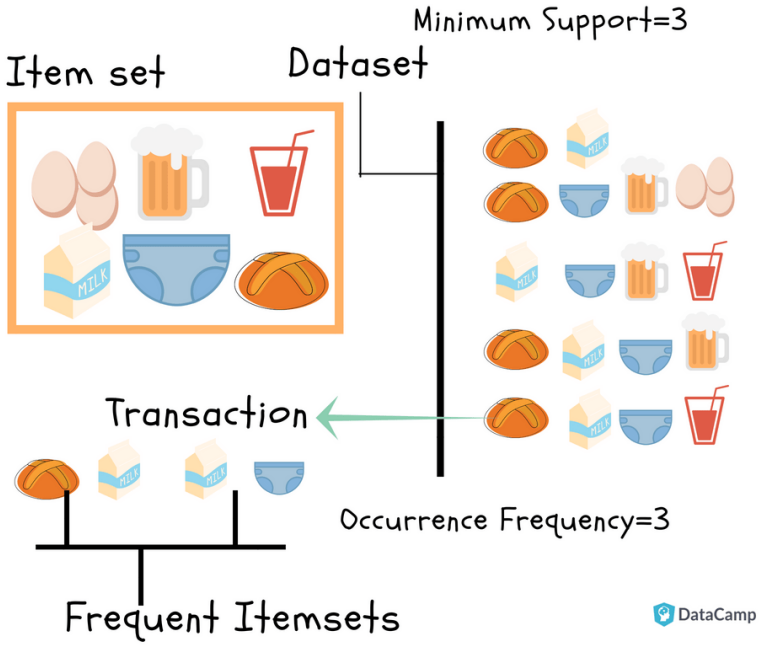
Prof. D.S. Yadav
Ms. Deepa Verma

Department of Computer Science and Engineering
Institute of Engineering and Technology, Lucknow

Market Basket Analysis

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

Implementation of market basket analysis requires a background in statistics and data science, as well as some algorithmic computer programming skills.



Problem Statement

In order to improve the performance of a retail store regarding sales and to reduce store inventory we need a method that can give best optimistic result we apply Market Basket Analysis.With growing demands of items and rapid lifestyle of an individual it helps us in saving a lot of time and effort.

- Market basket analysis can increase sales and customer satisfaction. Using data to determine that products are often purchased together, retailers can optimize product placement, offer special deals and create new product bundles to encourage further sales of these combinations.

- These improvements can generate additional sales for the retailer, while making the shopping experience more productive and valuable for customers. By using market basket analysis, customers may feel a stronger sentiment or brand loyalty toward the company.

Significance of the Project

The specific objectives of the project are listed below:

- To understand the purchasing pattern of products that comprise the customers' basket.
- To study about many products usually purchased by the customers.
- To study the most likely products purchased by the customers along with a particular product category.
- To recommend and suggest products to individual customers.

Layout of the Project

- **Data Collection**: The data was collected from http://archive.ics.uci.edu/ml/datasets/online+retail due to the unavailability of data from the supermarkets.

- **Data Pre-processing**: The data collected was mapped manually as integer values.
  For example “Fruit” was labeled as 1, “Bread” as 2, “Soups” as 3, and so on. The mapped integer values were then saved and given as the input to the system.

- **Apriori Algorithm:** The Apriori algorithm was used for processing the input data and the result was produced as the list of rules that are strongly associated with each other.

Apriori Algorithm

Apriori algorithm is used for finding frequent itemsets in a dataset for boolean association rule. Name of the algorithm is Apriori because it uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called **Apriori property** which helps by reducing the search space.

Apriori Property :- All subsets of a frequent itemset must be frequent(Apriori property).

If an itemset is infrequent, all its supersets will be infrequent.

Working of Apriori Algorithm

Few terms before the example

1. ### Confidence

   Given two items x,y confidence measure the percentage of times that item y is purchased, given that x was purchased

   Confidence Value always ranges between 0 and 1. Where 0 indicates y is never purchased when x is purchased 1 indicates y is always purchased whenever x is purchased

   Confidence{x,y} = freq(x,y)/freq(x)

2. **Support:** Support is the minimum probability for the product to get sold. Percentage of orders that contain the item set is explained

Support{x} = freq(x)/n [x → item, n → total no of transactions]

3. **Lift:** This is unlike confidence metric whose value may vary depending on direction

Example : Confidence{x,y} is different from Confidence{y,x}

Lift{x,y} = support{x,y}/(support{x}*support{y})

When Lift = 1; implies no relationship between x and y

When Lift > 1; implies there is positive relationship between x and y i.e., x and y are more likely to be purchased together.

When Lift < 1; implies there is negative relationship between x and y i.e., x and y are unlikely to be purchased together.

Imagine we have transactions of n customers who purchased something from retail store

Transactions are:-

Customer 1: Bread, egg, papaya, oats

Customer 2: Papaya, bread, oat packet and milk

Customer 3: Egg, bread, and butter

Customer 4: Oat packet, egg, and milk

Customer 5: Milk, bread, and butter

Customer 6: Papaya and milk

Customer 7: Butter, papaya, and bread

Customer 8: Egg and bread

Customer 9: Papaya and oat packet

……

Customer n:

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

Code Output for first 6 rules generated

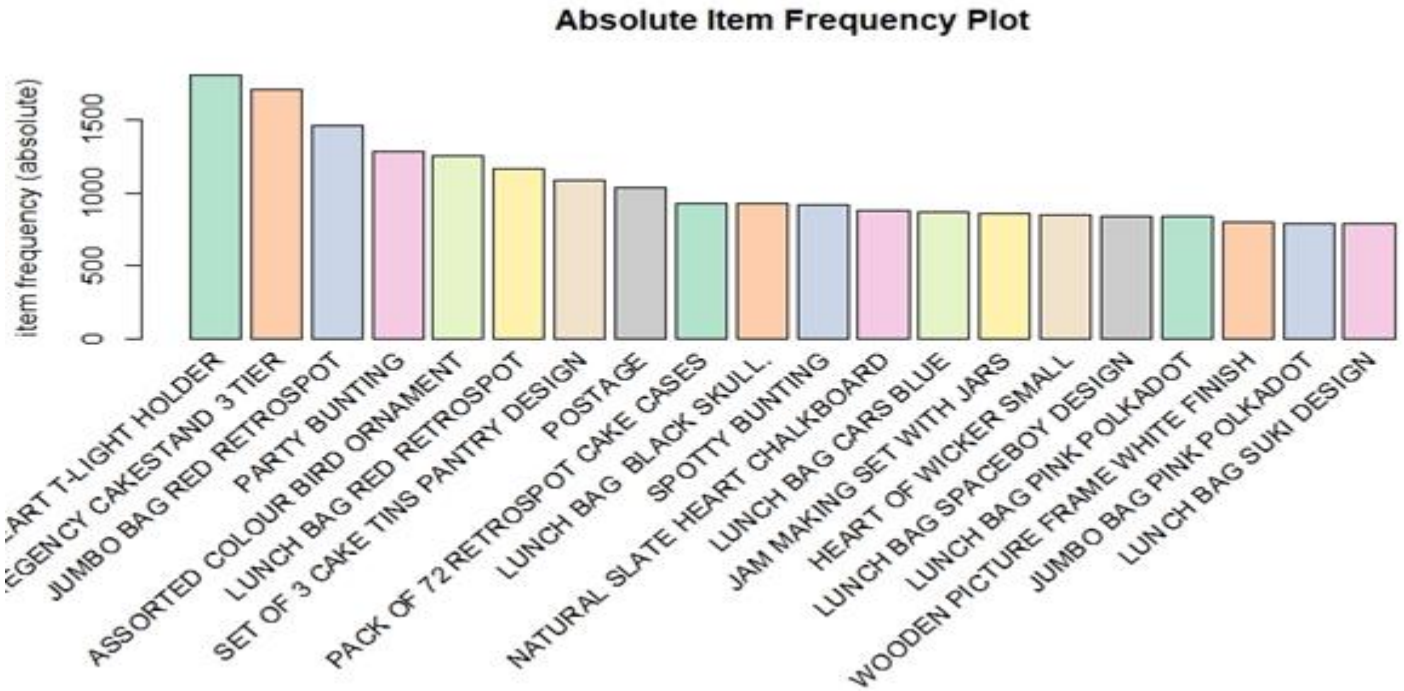
```
> inspect(head(association.rules[1:10]))
    lhs                           rhs                            support     confidence coverage    lift     count
[1] {WOBBLY CHICKEN}           => {DECORATION}                   0.001261773 1.0000000  0.001261773 443.8200 28
[2] {WOBBLY CHICKEN}           => {METAL}                        0.001261773 1.0000000  0.001261773 443.8200 28
[3] {DECOUPAGE}                => {GREETING CARD}                0.001036456 1.0000000  0.001036456 389.3158 23
[4] {BILLBOARD FONTS DESIGN}   => {WRAP}                         0.001306836 1.0000000  0.001306836 715.8387 29
[5] {WRAP}                     => {BILLBOARD FONTS DESIGN}       0.001306836 0.9354839  0.001396963 715.8387 29
[6] {ENAMEL PINK TEA CONTAINER} => {ENAMEL PINK COFFEE CONTAINER} 0.001396963 0.8157895  0.001712406 385.1741 31
```

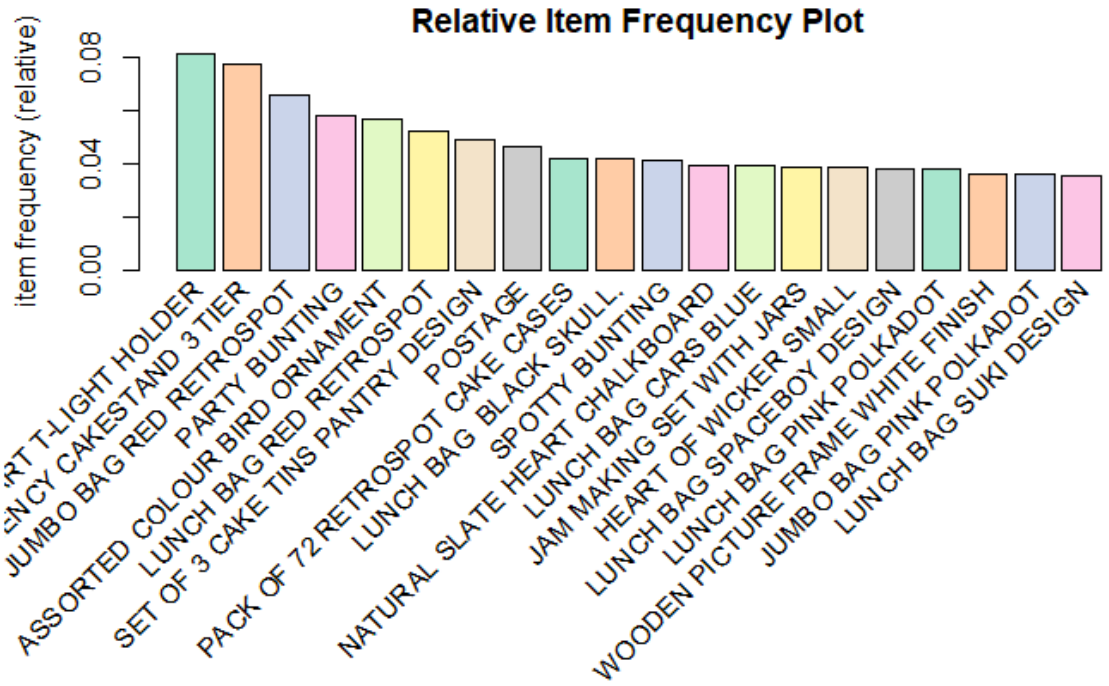
Visualizing Association Rules

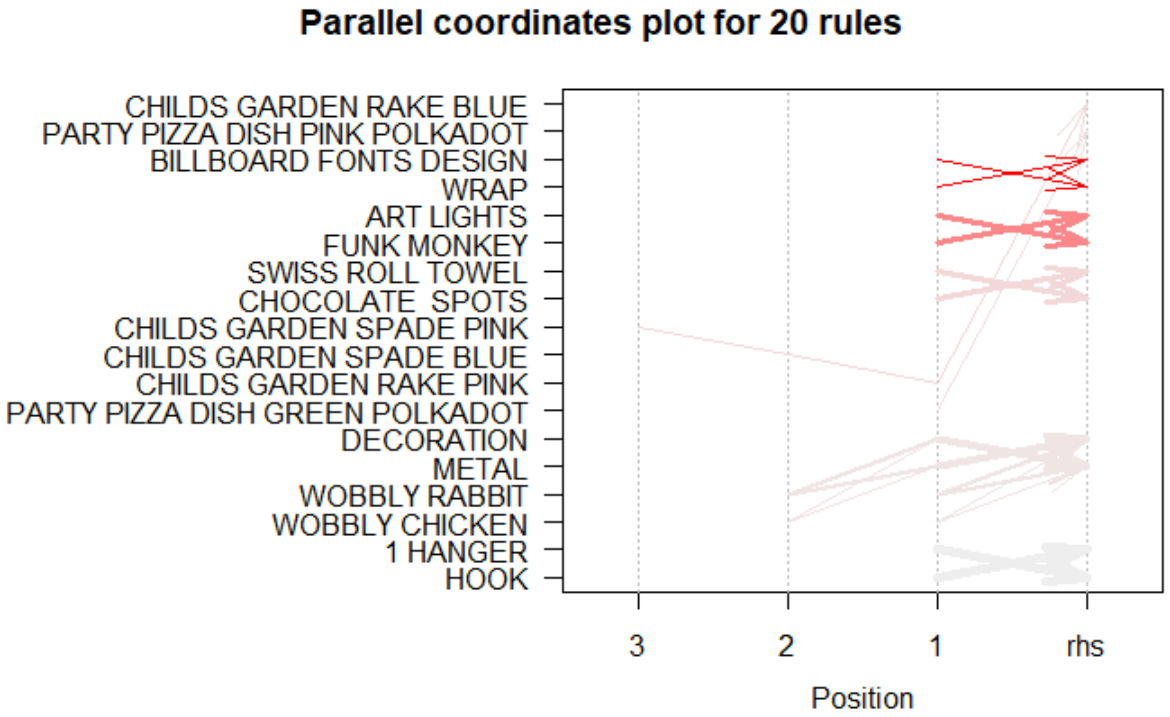
1. Since there will be hundreds or thousands of rules generated based on data, you need a couple of ways to present your findings.

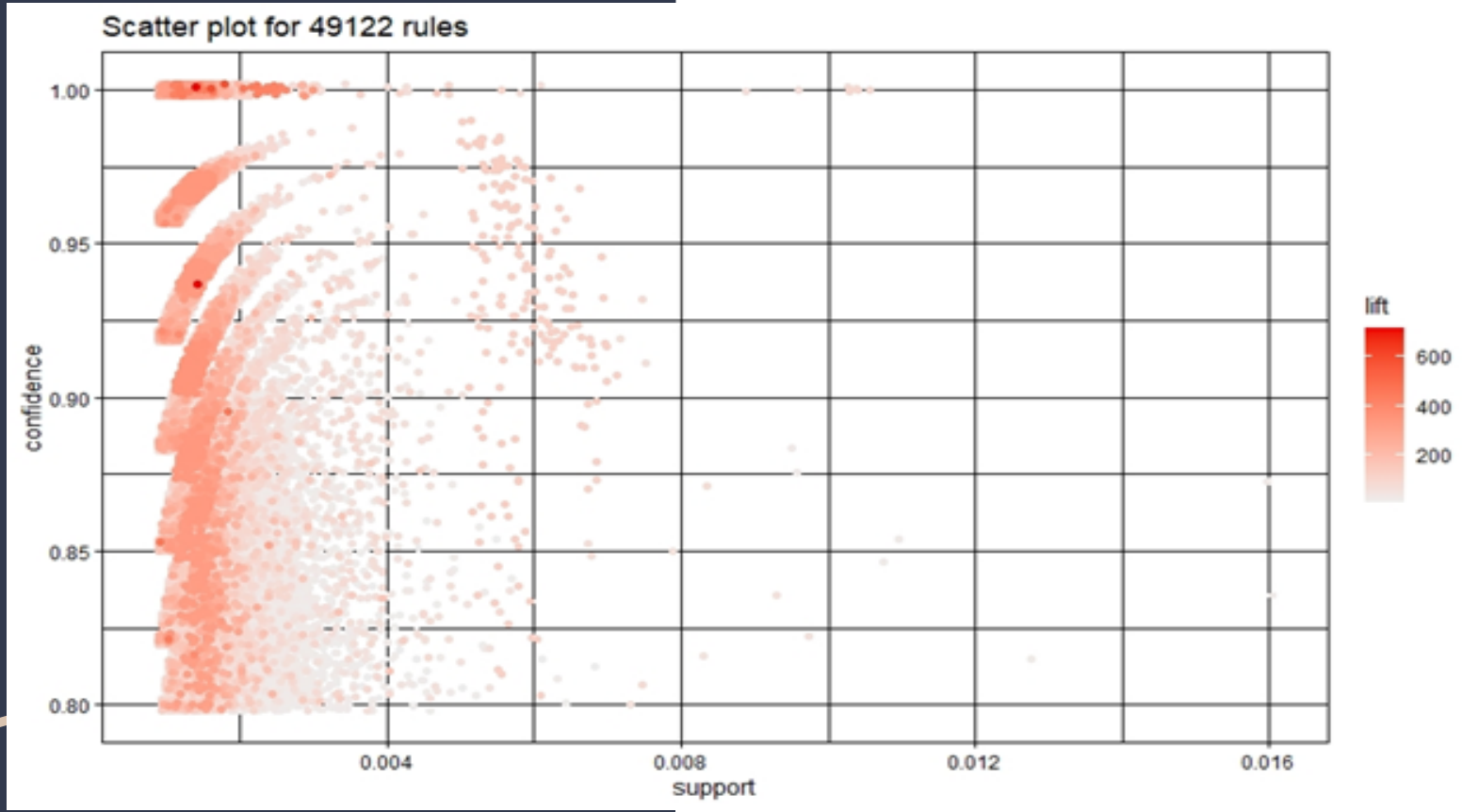
1. `ItemFrequencyPlot` has been used, which is also a great way to get top sold items.
2. Following visualization are used:
    - Scatter-Plot
    - Interactive Scatter-plot
    - Individual Rule Representation

Results and Observations from the project




Relative Item Frequency Plot
item frequency (relative)
0.00
0.04
0.08
RT T-LIGHT HOLDER
ENCY CAKESTAND 3 TIER
JUMBO BAG RED RETROSPOT
PARTY BUNTING
ASSORTED COLOUR BIRD ORNAMENT
LUNCH BAG RED RETROSPOT
SET OF 3 CAKE TINS PANTRY DESIGN
POSTAGE
PACK OF 72 RETROSPOT CAKE CASES
LUNCH BAG BLACK SKULL.
SPOTTY BUNTING
NATURAL SLATE HEART CHALKBOARD
LUNCH BAG CARS BLUE
JAM MAKING SET WITH JARS
HEART OF WICKER SMALL
LUNCH BAG SPACEBOY DESIGN
LUNCH BAG PINK POLKADOT
WOODEN PICTURE FRAME WHITE FINISH
JUMBO BAG PINK POLKADOT
LUNCH BAG SUKI DESIGN


Parallel coordinates plot for 20 rules
CHILDS GARDEN RAKE BLUE
PARTY PIZZA DISH PINK POLKADOT
BILLBOARD FONTS DESIGN
WRAP
ART LIGHTS
FUNK MONKEY
SWISS ROLL TOWEL
CHOCOLATE SPOTS
CHILDS GARDEN SPADE PINK
CHILDS GARDEN SPADE BLUE
CHILDS GARDEN RAKE PINK
PARTY PIZZA DISH GREEN POLKADOT
DECORATION
METAL
WOBBLY RABBIT
WOBBLY CHICKEN
1 HANGER
HOOK
3
2
1
rhs
Position


Scatter plot for 49122 rules
confidence
1.00
0.95
0.90
0.85
0.80
0.004
0.008
0.012
0.016
support
lift
600
400
200

References

- https://www.geeksforgeeks.org/apriori-algorithm/
- https://searchcustomerexperience.techtarget.com/definition/market-basket-analysis
- https://www.datacamp.com/community/tutorials/market-basket-analysis-r
- https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/
- https://paginas.fe.up.pt/~ec/files_1112/week_04_Association.pdf

THANK YOU