# Topic: Association Rules

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name SHARANNBASAPPA**

**Batch Id** 050121 10AM

**Topic: - Association Rules.**

# Hints:

1. Business Problem
   1. Objective
   2. Constraints (if any)
2.  Work on each feature of the dataset to create a data dictionary as displayed in the below image**:**

 **Using R and Python codes perform:**

3.Data Pre-processing

* 1. Data Cleaning, Feature Engineering, etc.

4.Model Building

4.1 Application of Apriorism Algorithm.

* 1. Build most frequent item sets and plot the rules.
  2. Work on both R and Python Codes.

5.Deployment

* 1. Deploy solutions using R shiny and Python Flask.

6. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment, the solution should be submitted in the above format
2. Research and Perform all possible steps for improving the rules and also check if you can take out sub rules from main rules.
3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here that is commenting is necessary in the codes.
5. Please send all files at once whilst submitting assignments.

**Problem Statement: -**

Kitabi Duniya , a famous book store in India, which was established before Independence, the growth of the company was incremental year by year, but due to online selling of books and wide spread Internet access its annual growth started to collapse, seeing sharp downfalls, you as a Data Scientist help this heritage book store gain its popularity back and increase footfall of customers and provide ways the business can improve exponentially, apply Association Rule Algorithm, explain the rules, and visualize the graphs for clear understanding of solution.

**1.) Books.csv**

**1)**

**solution: - 1.1) objectives: - Maximize profit**

**1.2) Constraints: - Make less set of rules for buyers**

**2)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** |
| **All columns (all are same data type)** | **If book is sold out 1 else 0** | **Binary** | **Use full for finding solution for business** |

**3) data preprocessing: - for data pre- processing we check is there any NA values, and if yes, replace by null.**

**Removing nominal and unwanted columns.**

**But in this question, there is no nominal and no unwanted columns.**

**4)**

**4.1) support, interest, lift ratio are makes decision**

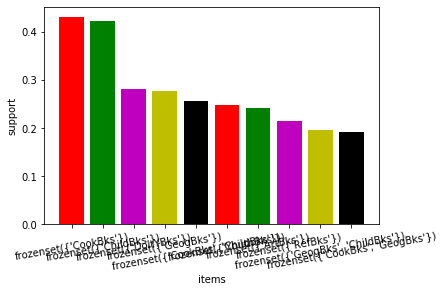
**from given data maximum sales of book is decided by 3 factors which are interest, support and lift ratio.**

**Why we are doing: - from given set of data we analyze how many peoples took which books, and try to make combination of those and using association techniques, with minimum rules and maximum convenience to customers so that they buy and improve in business.**

**Bar plot: - it is plot drawn between support and items.**

**Which has more lift ratio (more frequency of sale) that combination is more selling items in the market**

**Make those combination to improve business**



**5)**

**antecedents consequents ... leverage conviction**

**8038 (ItalArt, RefBks, DoItYBks) (ItalAtlas) ... 0.009093 7.062000**

**6478 (ItalArt, RefBks, GeogBks) (ItalAtlas) ... 0.011000 6.500250**

**3636 (ItalArt, RefBks) (ItalAtlas) 0.015760 5.502857**

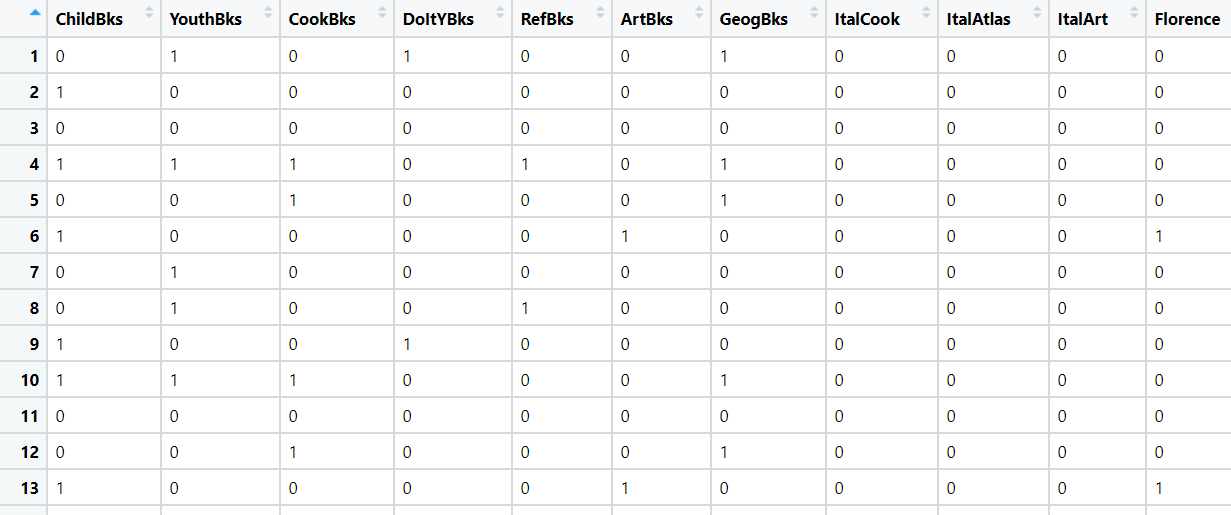
**3668 (ItalArt, RefBks, ArtBks) (ItalAtlas) ... 0.015760 5.502857**

**5484 (CookBks, ItalArt, RefBks) (ItalAtlas) ... 0.012389 4.539857**

**[5 rows x 9 columns]**

**Top 5 combination of books is taken into consideration-**

**ItalArt, RefBks, DoItYBks this combination of book is more selling, so combine those books and offer customers with attracted price, make improvement in business .**



**Problem Statement:**

**The Departmental Store, has gathered the data of the products it sells on a**

**Daily basis. Using Association Rules concepts, provide the insights on the rules and the plots.**

**2.) Groceries.csv**

**1)**

**solution: - 1.1) objectives: - Maximize sale with profit**

**1.2) Constraints: - Make less set of rules for buyers**

**2)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** |
| **All columns (all are same data type), except first one is nominal** | **If a person buy item 1 else 0** | **Categorical** | **Use full for finding solution for business, but need to convert into binary** |

**3)data pre-processing: data is to convert from discrete to binary using dummy variable creation method.**

**Check for NA and replace them by empty**

**4) 4)**

**4.1) support, interest, lift ratio are makes decision**

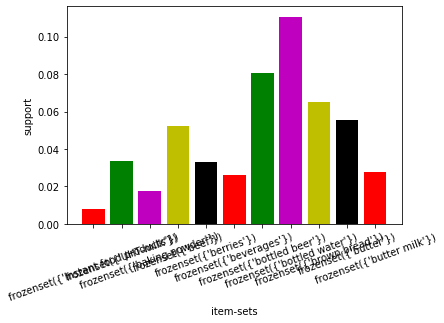
**from given data maximum sales of groceries is decided by 3 factors which are interest, support and lift ratio.**

**Why we are doing: - from given set of data we analyze how many peoples took which groceries , and try to make combination of those and using association techniques, with minimum rules and maximum convenience to customers so that they buy and improve in business.**

**Bar plot: - it is plot drawn between support and items.**

**Which has more lift ratio (more frequency of sale) that combination is more selling items in the market**

**Make those combination to improve business**



**antecedents consequents ... leverage conviction**

**24 (whipped/sour cream) (berries) ... 0.006666 1.106428**

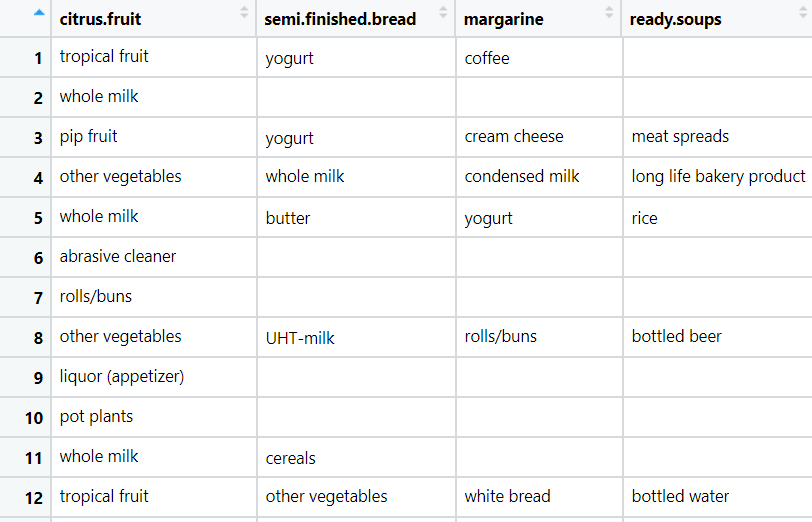
**24 (whipped/sour cream) (berries) ... 0.006666 1.106428**

**632 (other vegetables, beef) (root vegetables) ... 0.005781 1.490123**

**632 (other vegetables, beef) (root vegetables) ... 0.005781 1.490123**

**644 (whole milk, beef) (root vegetables) ... 0.005716 1.432456**

**From the above result whipped/sour cream berries has more lift ratio, and rest of all are sorted as shown above, make those combination to improve business.**



**Problem Statement:**

**A film distribution company wants to target audience based on their likes and dislikes, you as a Chief Data Scientist Analyze the data and come up with different rules of movie list so that the business objective is achieved.**

**3.) my\_movies.csv**

**1)**

**solution: - 1.1) objectives: - Maximize customer to more profit**

**1.2) Constraints: - Make less set of rules for buyers**

**2)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** |
| **From 1 to 5th colums are movie names**  **Remaining are equivalent of first 5 columns** | **Given different movies names wrt watched columns** | **Categorical (1 to 5) remaining converted to binary** | **First 5 columns are removed. Rest all Use full for finding solution for business, but need to convert into binary** |

**3)data pre-processing: first 5 columns are already converted into binary, so no need of first 5 columns, so removed first 5 columns.**

**Check for NA and replace them by empty**

**4) 4)**

**4.1) support, interest, lift ratio are makes decision**

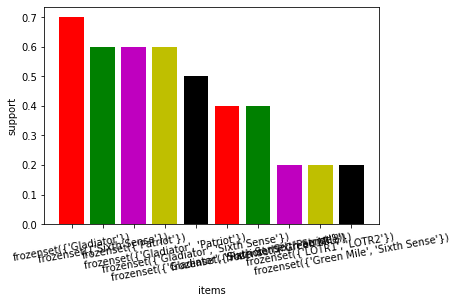
**from given data maximum watch of movies is decided by 3 factors which are interest, support and lift ratio.**

**Why we are doing: - from given set of data we analyze how many peoples took which movies, and try to make combination of those and using association techniques, with minimum rules and maximum convenience to customers so that they buy and improve in business.**

**Bar plot: - it is plot drawn between support and items.**

**Which has more lift ratio (more frequency of sale) that combination is more selling items in the market**

**Make those combination to improve business**



**antecedents ... conviction**

**80 (Green Mile, Gladiator) ... inf**

**130 (Green Mile, Gladiator, Sixth Sense) ... inf**

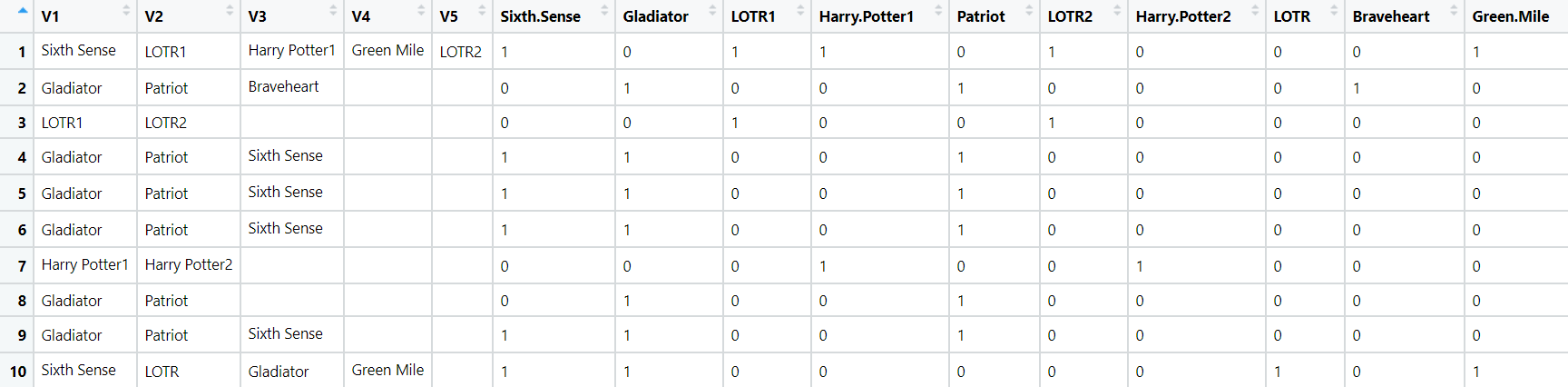
**118 (LOTR1, Sixth Sense) ... inf**

**208 (Green Mile, LOTR1, Sixth Sense, Harry Potter1) ... inf**

**50 (Sixth Sense, LOTR2) ... inf**

**From the above (Green Mile, Gladiator) these combinations of movies are watching most of the customers**

**So make these arrangement to improve business**



**Problem Statement: -**

**A Mobile Phone manufacturing company wants to launch its three brand new phone into the market, but before going with its traditional marketing approach this time it want to analyze the data of its previous model sales in different regions and you have been hired as an Data Scientist to help them out, use the Association rules concept and provide your insights to the company’s marketing team to improve its sales.**

**4.) myphonedata.csv**

**1)**

**solution: - 1.1) objectives: - Maximize sale to more profit**

**1.2) Constraints: - Make less set of rules for buyers**

**2)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** |
| **From 1 to 3th colums are names**    **Remaining are equivalent of first 3 columns** | **Given different names wrt color columns** | **Categorical (1 to 3) remaining converted to binary** | **First 3 columns are removed. Rest all Use full for finding solution for business, but need to convert into binary** |

**3)data pre-processing: first 3 columns are already converted into binary, so no need of first 3 columns, so removed first 3 columns.**

**Check for NA and replace them by empty**

**4) 4)**

**4.1) support, interest, lift ratio are makes decision**

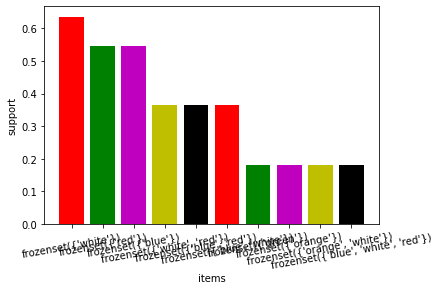
**from given data maximum watch of movies is decided by 3 factors which are interest, support and lift ratio.**

**Why we are doing: - from given set of data we analyze how many peoples took which movies, and try to make combination of those and using association techniques, with minimum rules and maximum convenience to customers so that they buy and improve in business.**

**Bar plot: - it is plot drawn between support and items.**

**Which has more lift ratio (more frequency of sale) that combination is more selling items in the market**

**Make those combination to improve business**



**antecedents consequents ... leverage conviction**

**8 (green, white) (red) ... 0.041322 inf**

**14 (orange, red) (white) ... 0.033058 inf**

**6 (orange) (white) ... 0.066116 inf**

**2 (blue) (red) ... 0.066116 1.363636**

**0 (white) (red) ... 0.016529 1.060606**

**green, white are the ,more selling items, make those combination to more profit**

**Problem Statement: -**

**A retail store in India, has its transaction data, and it would like to know the buying pattern of the consumers in its locality, you have been assigned this task to provide the manager with rules on how the placement of products needs to be there in shelves so that it can improve the buying patterns of consumes and increase customer footfall.**

**5.) transaction\_retail.csv**

**)**

**solution: - 1.1) objectives: - Maximize sale to more profit**

**1.2) Constraints: - Make less set of rules for buyers**

**2)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Feature** | **Description** | **Type** | **Relevance** |
| **All columns item name colurs** | **Colur and name of the items** | **catigorical** | **Need to convert binary for analysis** |

**3)data pre-processing: first 3 columns are already converted into binary, so no need of first 3 columns, so removed first 3 columns.**

**Check for NA and replace them by empty**

**4) 4)**

**4.1) support, interest, lift ratio are makes decision**

**from given data maximum watch of movies is decided by 3 factors which are interest, support and lift ratio.**

**Why we are doing: - from given set of data we analyze how many peoples took which movies, and try to make combination of those and using association techniques, with minimum rules and maximum convenience to customers so that they buy and improve in business.**

**Bar plot: - it is plot drawn between support and items.**

**Which has more lift ratio (more frequency of sale) that combination is more selling items in the market**

