# Association Rules

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: DHEERAJ MISHRA Batch ID:**  DS\_ 01072021

**Topic: Association Rules**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

Hints:

1. Business Problem
   1. **What is the business objective?**
   2. **Are there any constraints?**
2. Work on each feature of the dataset to create a data dictionary as displayed in the below image**:**



1. Data Pre-processing
   1. Data Cleaning, Feature Engineering, etc.
2. Model Building

4.1 Application of Apriori Algorithm

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

5.Deployment

5.1 Deploy solutions

6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?

**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 RuleAlgorithm, explain the rules, and visualize the graphs for clear understanding of solution.

**1.) Books.csv**

1. **Business objective :**

**Max:- Frequent and dependent books**

**Min:- Not sold books**

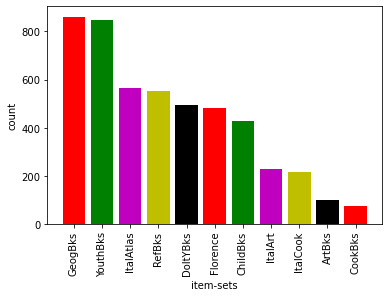
**Constraints:- Skill of employer**

1. Data understanding :

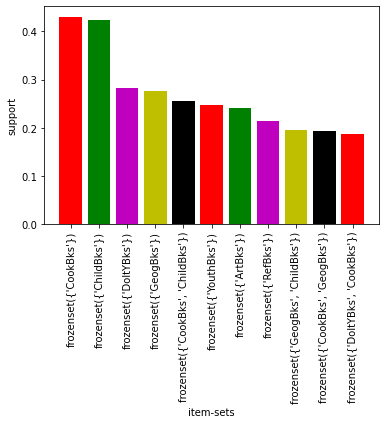
|  |  |  |  |
| --- | --- | --- | --- |
| N Name of Feature | D Description | TyTypes | R Relevance |
| ID ChildBks | U Books for kids | Discrete,  Binary | Relevant |
| BaYouthBks | Books for youth | Discrete,  Binary | R Relevant |
| Q CookBks | Books for cooking | Discrete,  Binary | R lRelevant |
| Cc DoItYBks | Picture books | Discrete,  Binary | Relevant |
| Cc RefBks | N Books for reference | Discrete,  Binary | R Relevant |
| Cc ArtBks | Books for arts | Discrete,  Binary | R Relevant |
| B GeogBks | Books for geography | Binary | R Relevant |
| ItalCook | Books for cooking in Italy | Discrete,  Binary | R Relevant |
| ItalAtlas | Historical atlas for italy | Discrete,  Binary | R Relevant |
| D ItalArt | Art books in Italy | Discrete,  Binary | R Relevant |
| A I Florence | Travel guide books | Discrete,  Binary | Relevant |

1. DATA CLEANSING

1. No null values found in each feature
2. Duplicate row found and retained
3. All features are of type int64
4. Dataset consists of 2000 rows and 11 columns
5. Frequencies of each item were calculated



1. Model building :-
2. Model builded on dataset by apriori algorithm
3. From frequent\_items we can find dependency



1. DEPLOYMENT :-
2. Most frequent books

|  |  |
| --- | --- |
| SR NO | Most frequent books |
| 1. | **Cooking books** |
| 2. | **Kids book** |
| 3. | **Picture books** |
| 4. | **Geographical books** |
| 5. | **Cooking and kids book** |
| 6. | **Youth books** |
| 7. | **Arts book** |
| 8. | **Reference books** |
| 9. | **Geographical and kids book** |
| 10 | **Cooking and geographical books** |

1. **Item sets of book with antecedents and consequents**

|  |  |
| --- | --- |
| ANTECEDENTS | CONSEQUENTS |
| Reference , Italy arts , kids books | **Italy atlas books** |
| Reference , Italy atlas , geographical ,arts , Italy cook books | **Italy arts books** |
| Reference ,cook , Italy atlas ,Italy cook ,geographical ,arts books | **Italy arts books** |
| Italy atlas ,Italy cook ,geographical ,arts books | **Italy arts books** |
| Reference , Italy atlas ,Italy cook , picture ,arts books | **Italy arts books** |
| Reference ,cook , Italy atlas ,geographical ,arts , pictures books | **Italy arts books** |
| Reference ,cook , Italy atlas ,kids ,arts , pictures books | **Italy arts books** |
| Italy atlas ,Italy cook ,arts , pictures books | **Italy arts books** |
| Reference , Italy atlas ,Italy cook ,geographical ,arts, pictures books | **Italy arts books** |
| Reference ,cook , Italy atlas ,geographical ,arts , kids books | **Italy arts books** |

1. **Impact :**

**From above store can recognize which books are sold more frequently.**

**Also which books are purchased more frequently with other books .**

**2 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.

1. **Business objective :**

**Max:- Frequent and dependent items**

**Min:- Not sold groceries items**

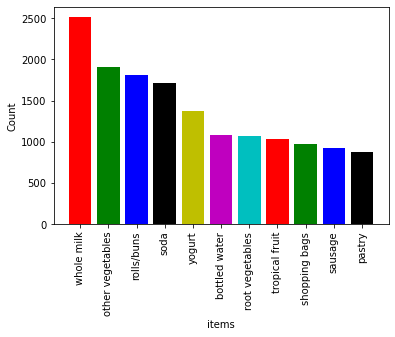
**Constraints:- Skill of employer**

1. Data understanding :

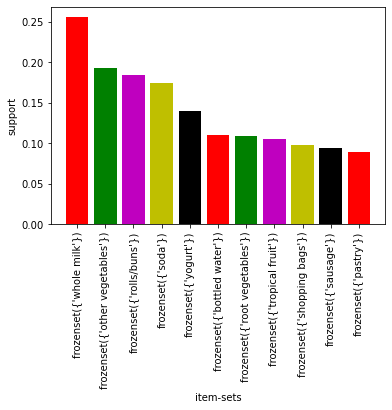
|  |  |  |  |
| --- | --- | --- | --- |
| N Name of Feature | D Description | TyTypes | R Relevance |
| ID value | U all items names | String ,text | Relevant |

1. DATA CLEANSING

1. List is created for separate items
2. Dictionary is created and get count of each items
3. First key element deleted as containing empty count
4. All features are of type string
5. Last row deleted for null transactions
6. Dataset consists of 9835 rows and 1 columns
7. Frequencies for each items are shown below



1. Model building :-
2. Model builded on dataset by apriori algorithm
3. From frequent\_items we can find dependency



1. DEPLOYMENT :-
2. Most frequent groceries items

|  |  |
| --- | --- |
| SR NO | Most frequent groceries items |
| 1. | **Whole milk** |
| 2. | **Other vegetables** |
| 3. | **Rolls or burns** |
| 4. | **Soda** |
| 5. | **Yogurt** |
| 6. | **Bottled water** |
| 7. | **Root vegetables** |
| 8. | **Tropical fruits** |
| 9. | **Shopping bags** |
| 10 | **Sausage** |

1. **Item sets of groceries with antecedents and consequents**

|  |  |
| --- | --- |
| ANTECEDENTS | CONSEQUENTS |
| Whipped or sour cream | **Berries** |
| Whole milk ,yogurt | **Curd** |
| Whole milk, other vegetables , yogurt | **Tropical fruits** |
| Root vegetables | **Beef** |
| Whole milk , root vegetables | **Butter** |
| Root vegetables ,citrus fruits | **Other vegetables** |
| Root vegetables ,tropical fruits | **Other vegetables** |
| Whole milk ,yogurt | **Butter** |
| Root vegetables | **onions** |
| Whole milk, root vegetables , yogurt | **Other vegetables** |

1. **Impact :**

**From above groceries store can recognize which grocery items are sold more frequently.**

**Also which groceries items are purchased more frequently with other groceries items .**

**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. **Business objective :**

**Max:- Frequent and dependent movies**

**Min:- Not liked movies**

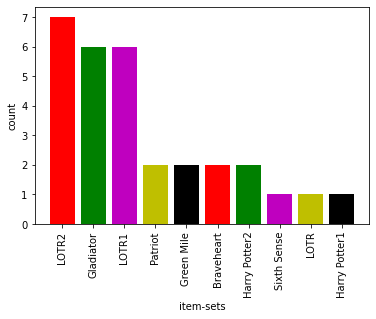
**Constraints:- Skill of employer**

1. Data understanding :

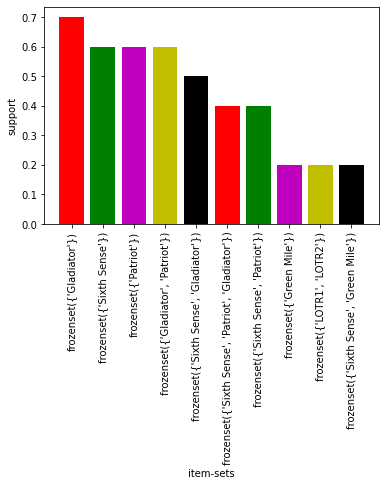
|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| V1 | Type of movies | Strings ,char | Not relevant |
| V2 | Type of movies | Strings ,char | Not relevant |
| V3 | Type of movies | Strings ,char | Not relevant |
| V4 | Type of movies | Strings ,char | Not relevant |
| V5 | Type of movies | Strings ,char | Not relevant |
| Sixth Sense | Sixth sense movie | Discrete , binary | Relevant |
| Gladiator | Gladiator movie | Discrete , binary | Relevant |
| LOTR1 | Lords of the rings part 1 | Discrete , binary | Relevant |
| Harry Potter1 | Harry potter part1 | Discrete , binary | Relevant |
| Patriot | Patriot movie | Discrete , binary | Relevant |
| LOTR2 | Lords of the rings part 2 | Discrete , binary | Relevant |
| Harry Potter2 | Harry potter part2 | Discrete , binary | Relevant |
| LOTR3 | Lords of the rings part 3 | Discrete , binary | Relevant |
| Brave Heart | Brave heart movie | Discrete , binary | Relevant |
| Green Mile | Green mile movie | Discrete , binary | Relevant |

1. DATA CLEANSING

1. There are five nominal columns so dropped it
2. There is no null value
3. Duplicate records found but retained
4. All features are of type string and discrete
5. Dataset consists of 10 rows and 15 columns
6. Frequencies for each movies are shown below



1. Model building :-
2. Model builded on dataset by apriori algorithm
3. From frequent\_items we can find dependency



1. DEPLOYMENT :-
2. Most frequent movies

|  |  |
| --- | --- |
| SRNO | MOVIES NAME |
| 1. | Lords of the rings part 2 |
| 2. | Gladiator , lords of the rings part1 |
| 3. | Patriot , green mile, brave heart , harry potter part 2 |
| 4. | Lords of gladiator , harry potter part 21 |

1. Item sets of movies **with antecedents and consequents**

|  |  |
| --- | --- |
| ANTECEDENTS | CONSEQUENTS |
| Green mile ,gladiator | **Lords of the rings** |
| Sixth sense , green mile, lords of the rings part 1 | **Lords of the rings part2** |
| Sixth sense , lords of the rings part 1 | **Harry potter part 1** |
| Sixth sense , green mile | **Lords of the rings** |
| Sixth sense , harry potter part 1 | **Lords of the rings part2** |
| Green mile | **Lords of the rings** |
| Sixth sense , green mile, harry potter part 1 | **Lords of the rings part2** |
| Sixth sense , lords of the rings part1, harry potter part 1 | **Lords of the rings part2** |
| Sixth sense , green mile, lords of rings part 1 and part 2 | **Harry potter part 1** |
| Lords of the rings part1 | **Lords of the rings part2** |

1. **Impact :**

**From above we can recognize which movies are liked more frequently.**

**Also which movies are liked more frequently with other movies.**

**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. **Business objective :**

**Max:- Frequent and dependent faceplates**

**Min:- Not sold faceplates**

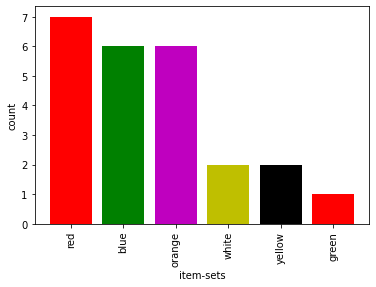
**Constraints:- Skill of employer**

1. Data understanding :

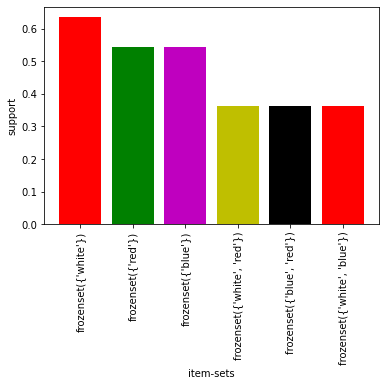
|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| V1 | Type of faceplates | Strings ,char | Not relevant |
| V2 | Type of faceplates | Strings ,char | Not relevant |
| V3 | Type of faceplates | Strings ,char | Not relevant |
| Red | Red face plate | Discrete , binary | Relevant |
| white | White face plate | Discrete , binary | Relevant |
| green | Green face plate | Discrete , binary | Relevant |
| Yellow | Yellow face plate | Discrete , binary | Relevant |
| Orange | Orange face plate | Discrete , binary | Relevant |
| blue | Blue face plate | Discrete , binary | Relevant |

1. DATA CLEANSING

1. There are three nominal columns so dropped it
2. There is no null value
3. Duplicate records found but retained
4. All features are of type string and discrete
5. Dataset consists of 11 rows and 9 columns
6. Frequencies for each faceplates colour are shown below



1. Model building :-
2. Model builded on dataset by apriori algorithm
3. From frequent\_items we can find dependency



1. DEPLOYMENT :-
2. Most frequent faceplates color

|  |  |
| --- | --- |
| SRNO | FACE PLATE COLOR |
| 1. | Red |
| 2. | Blue and orange |
| 3. | White and yellow |
| 4. | Green |

1. Item sets of movies **with antecedents and consequents**

|  |  |
| --- | --- |
| ANTECEDENTS | CONSEQUENTS |
| White | **Orange** |
| White , red | **Orange** |
| White , red | **Green** |
| Blue | **Red** |
| White | **Blue** |
| White | **Red** |

1. **Impact :**

**From above we can recognize which faceplate color are liked more frequent.**

**Also which faceplate are liked more frequently with other faceplate color.**

**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**

1. **Business objective :**

**Max:- Frequent and dependent items**

**Min:- Not sold items**

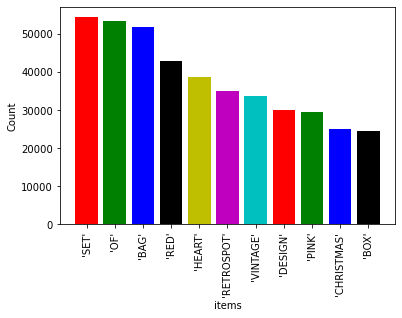
**Constraints:- Skill of employer**

1. Data understanding :

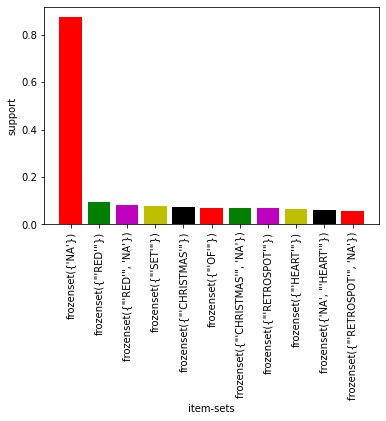
|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Value | Text data | Strings ,char | Relevant |

1. DATA CLEANSING

1. List is created for separate items
2. Dictionary is created and get count of each items
3. First key element deleted as containing empty count
4. All features are of type string
5. Last row deleted for null transactions
6. Dataset consists of 9835 rows and 1 columns
7. Frequencies for each items are shown below



1. Model building :-
2. Model builded on dataset by apriori algorithm
3. From frequent\_items we can find dependency



1. DEPLOYMENT :-
2. Most frequent items

|  |  |
| --- | --- |
| SRNO | FACE PLATE COLOR |
| 1. | Set |
| 2. | Of |
| 3. | Bag |
| 4. | Red |
| 5. | Heart |
| 6. | Retrospot |
| 7. | Vintage |
| 8. | Design |
| 9. | Pink |
| 10 | Christmas |

1. Item sets of movies **with antecedents and consequents**

|  |  |
| --- | --- |
| ANTECEDENTS | CONSEQUENTS |
| Play house | **Poppy** |
| NA , play house | **Poppy** |
| Ribbon | **Reel** |
| NA , ribbon | **Reel** |
| NA, tin, in | **Plasters** |
| Tin ,in | **Plasters** |
| Alaram | **Bakelike** |
| Pencils | **Tube** |
| Dolly | **Girl** |
| NA , bakelike | **Clock** |
|  |  |

1. **Impact :**

**From above we can recognize which item are purchased more frequent.**

**Also which item are purchased more frequently with other items.**