Data Mining (Section 001)

# Project III: ASSOCIATION RULES

## Team No: 11

## Members: Vamsee Krishna B, Nikhileshwar I

## What are given?

This project is about understanding the concept of association rule mining and its application on a real-world data set and analyzing the results obtained. We are given the data extracted from the IMDb database

## Pre-Processing:

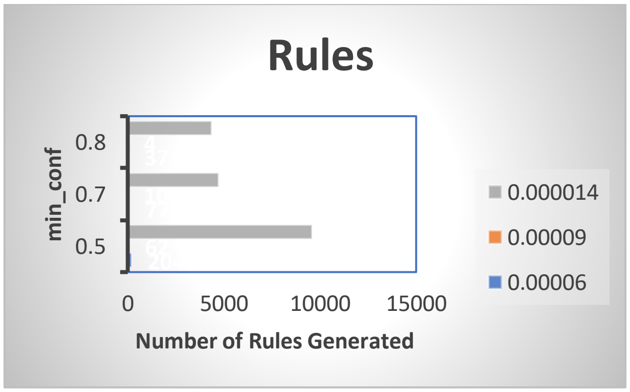
We are given two individual datasets each of a different format.We read both the csv files into two dataframes.We cleaned the datasets by removing the unwanted values.We the merged both dataframes into one and also subsetted the frame where year != 1999

The data sets are preprocessed in such a way that ?? are removed from the actor names and then both the data sets are merged on movieid_tid.

Then the attribute movieid_tid is removed and frequent itemsets along with candidate itemsets are generated for the given minimum support and confidance values.Thus rules will get generated for the minimum support and confidence values given.

We can obtain the lift value upon inspecting the generated rules

| **Rules** | | | | |
|---|---|---|---|---|
| | | Minimum support | | |
| | | 0.00006 | 0.00009 | 0.000014 |
| Minimum Confidance | 0.5 | 204 | 62 | 9593 |
| | 0.7 | 77 | 10 | 4733 |
| | 0.8 | 37 | 04 | 4364 |

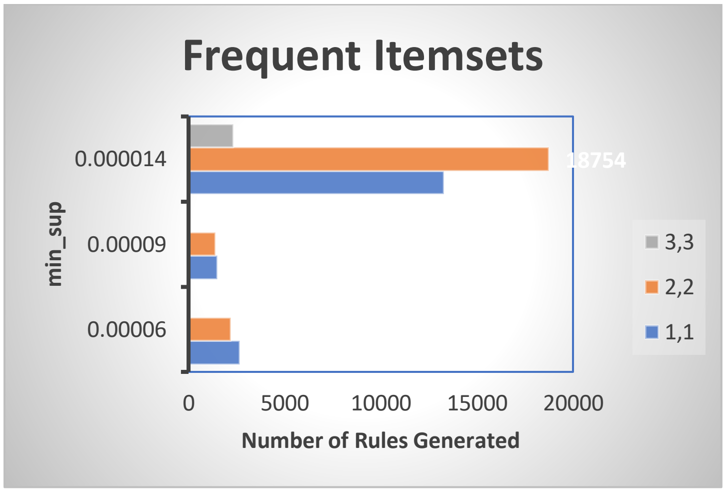


By keeping the minimum support same and with changing the confidence to 50%, 70%, 80% we see that there is a decrease in the number of rules that are being generated. From this we can infer that when the confidence is being increased the number of rules that will be generated will decrease.

**Frequent Itemsets generated for each iteration:**

Frequent Itemsets for each iteration

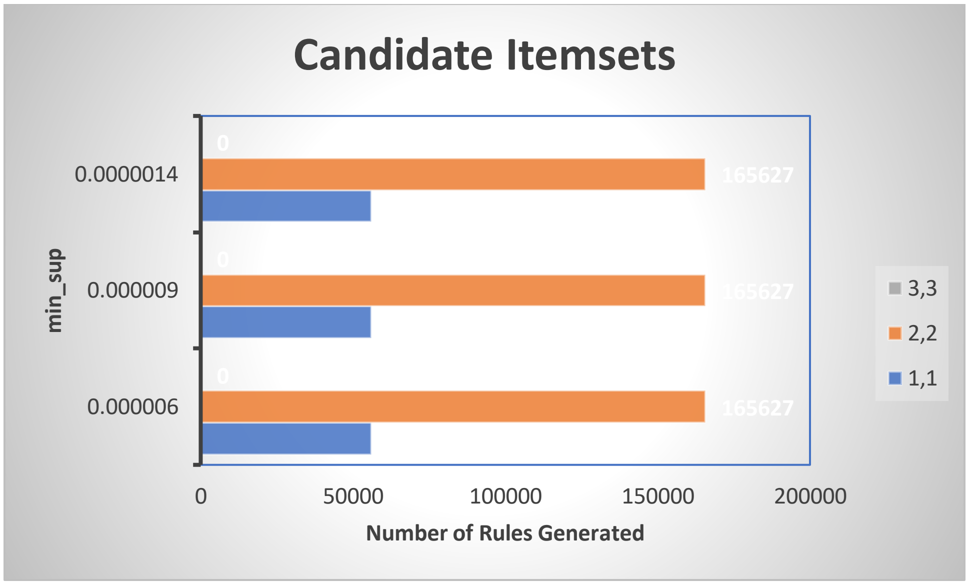
| | | Minimum support | | |
|---|---|---|---|---|
| | | 0.00006 | 0.00009 | 0.000014 |
| | (1,1) | 2656 | 1485 | 13272 |
| Iterations | (2,2) | 2180 | 1385 | 18754 |
| | (3,3) | 59 | 16 | 2321 |



For each iteration on changing the min and max length for the same support, we see a pattern that the number of itemsets being generated are decreaing.

**Candidate itemsets generated for each iteration:**

| | | min_sup | | |
|---|---|---|---|---|
| | | 0.000006 | 0.000009 | 0.0000014 |
| | (1,1) | 55903 | 55903 | 55903 |
| Iterations | (2,2) | 165627 | 165627 | 165627 |
| | (3,3) | 0 | 0 | 0 |

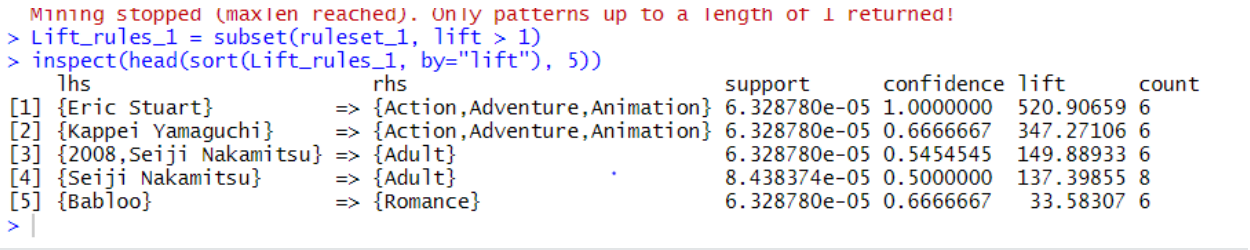


For each iteration on changing the min and max length for the same support, we see a pattern that the number of itemsets being generated are decreaing.

Here we couldn't generate the candidate sets for min_sup 0.000006, 0.000009 and 0.00000014 with min and max length (3,3) .

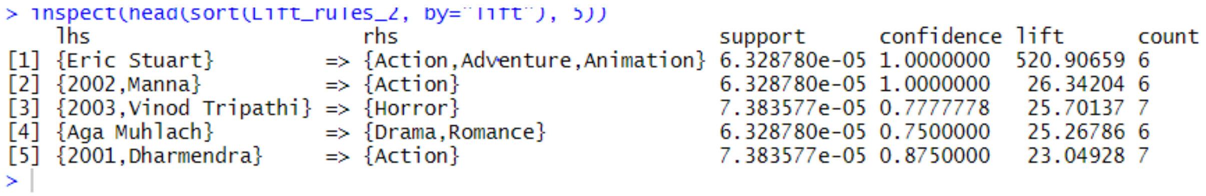
**Top 5 rules with lift>1**

1. Min_sup= **0.00006** , Min_conf= **0.5**



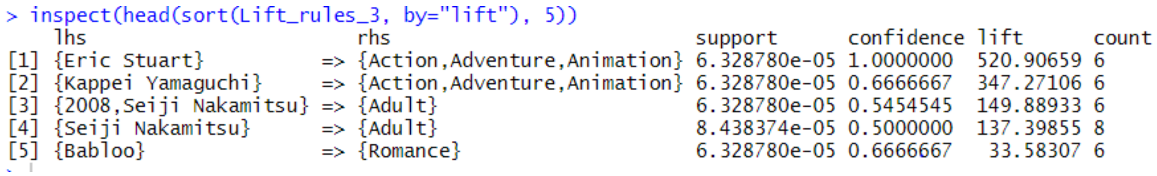
```
 Mining stopped (maxlen reached). Only patterns up to a length of 1 returned!
> Lift_rules_1 = subset(ruleset_1, lift > 1)
> inspect(head(sort(Lift_rules_1, by="lift"), 5))
    lhs                         rhs                          support      confidence lift      count
[1] {Eric Stuart}            => {Action,Adventure,Animation} 6.328780e-05 1.0000000  520.90659 6
[2] {Kappei Yamaguchi}       => {Action,Adventure,Animation} 6.328780e-05 0.6666667  347.27106 6
[3] {2008,Seiji Nakamitsu} => {Adult}                        6.328780e-05 0.5454545  149.88933 6
[4] {Seiji Nakamitsu}        => {Adult}                      8.438374e-05 0.5000000  137.39855 8
[5] {Babloo}                 => {Romance}                    6.328780e-05 0.6666667   33.58307 6
> |
```

2. Min_sup= **0.00006** , Min_conf= **0.7**



```
> inspect(head(sort(Lift_rules_2, by="lift"), 5))
    lhs                        rhs                          support      confidence lift      count
[1] {Eric Stuart}           => {Action,Adventure,Animation} 6.328780e-05 1.0000000  520.90659 6
[2] {2002,Manna}            => {Action}                     6.328780e-05 1.0000000   26.34204 6
[3] {2003,Vinod Tripathi} => {Horror}                       7.383577e-05 0.7777778   25.70137 7
[4] {Aga Muhlach}           => {Drama,Romance}              6.328780e-05 0.7500000   25.26786 6
[5] {2001,Dharmendra}       => {Action}                     7.383577e-05 0.8750000   23.04928 7
> |
```

3. Min_sup= **0.00006** , Min_conf= **0.8**



```
> inspect(head(sort(Lift_rules_3, by="lift"), 5))
    lhs                        rhs                          support      confidence lift      count
[1] {Eric Stuart}           => {Action,Adventure,Animation} 6.328780e-05 1.0000000  520.90659 6
[2] {Kappei Yamaguchi}      => {Action,Adventure,Animation} 6.328780e-05 0.6666667  347.27106 6
[3] {2008,Seiji Nakamitsu} => {Adult}                        6.328780e-05 0.5454545  149.88933 6
[4] {Seiji Nakamitsu}       => {Adult}                        8.438374e-05 0.5000000  137.39855 8
[5] {Babloo}                => {Romance}                      6.328780e-05 0.6666667   33.58307 6
```

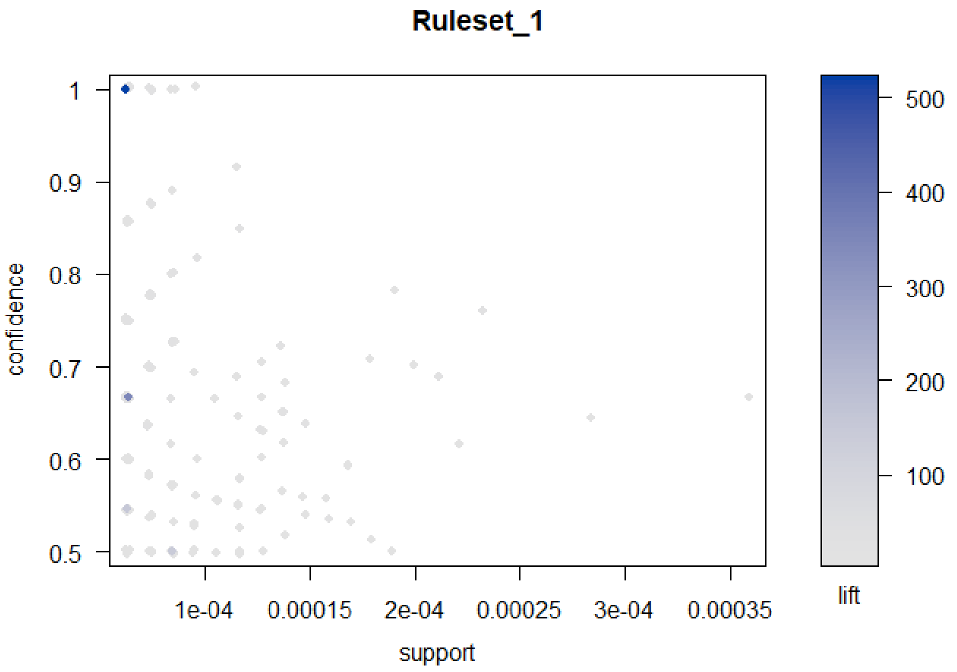
In the same manner, the top 5 rules for the the other support and confidence values can be computed.

Lift > 1 shows that the rules are in high co relation amongst each other.

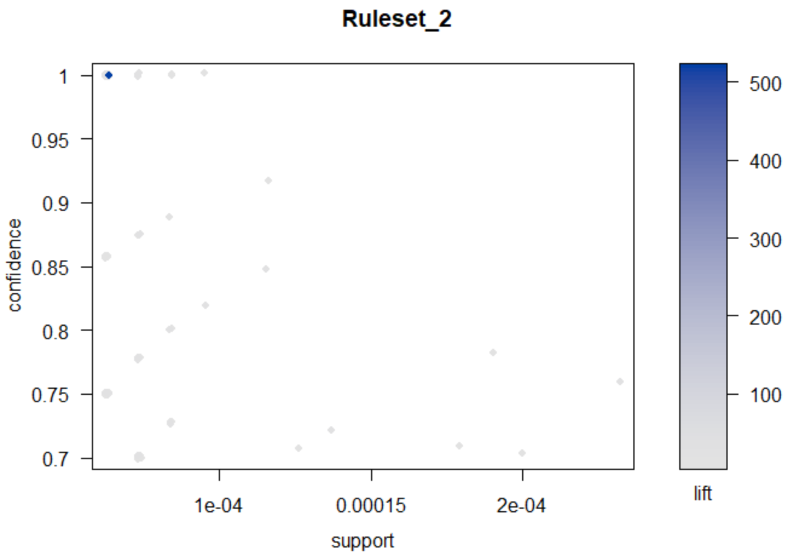
Unable to generate the rules for Lift < 1 and Lift = 1 because of the minimum support values

**Plots:**

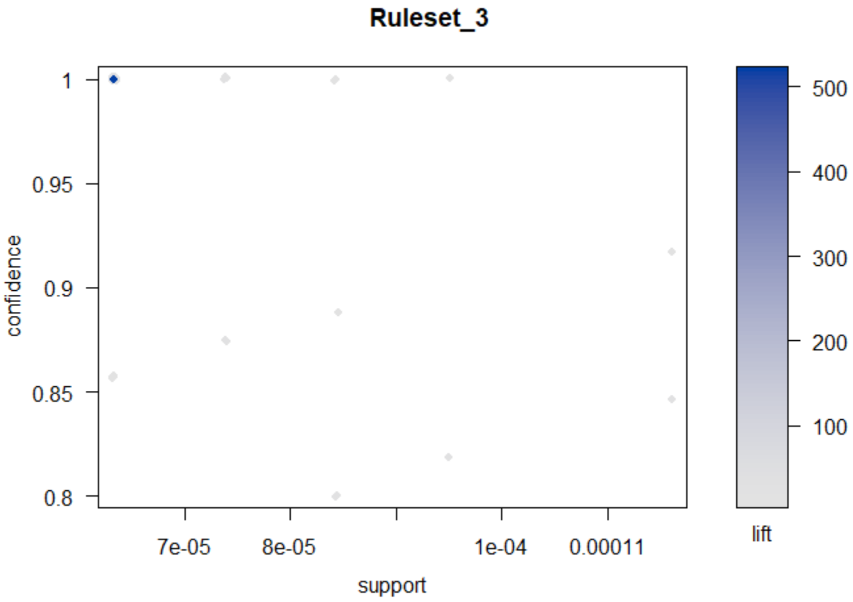
1. Min_sup= **0.00006** , Min_conf= **0.5**



2. Min_sup= **0.00006** , Min_conf= **0.7**



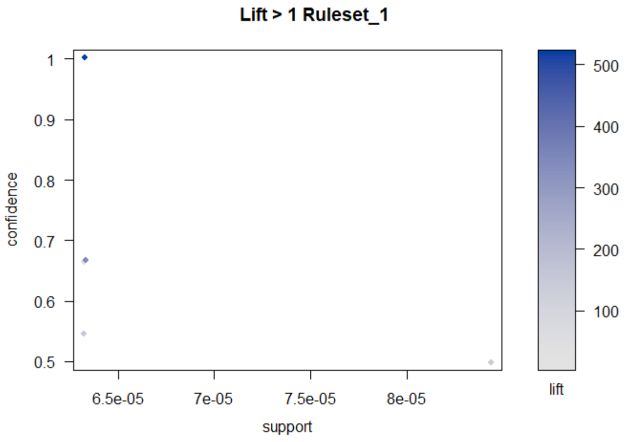
3. Min_sup= **0.00006** , Min_conf= **0.8**

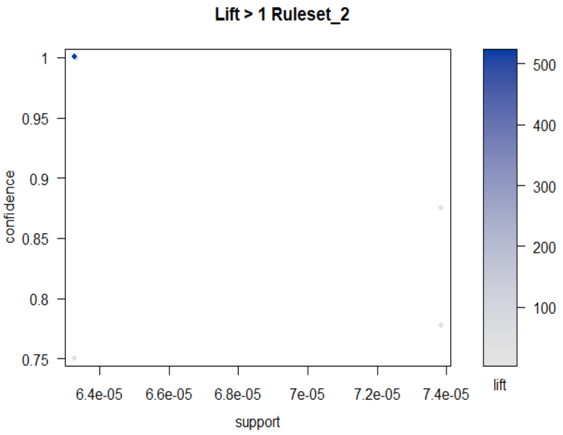


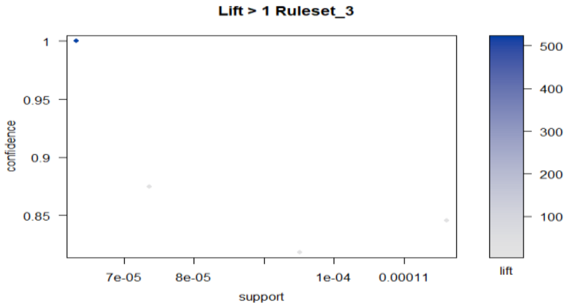
In the same manner, all the plots for various combinations for support and confidence are generated

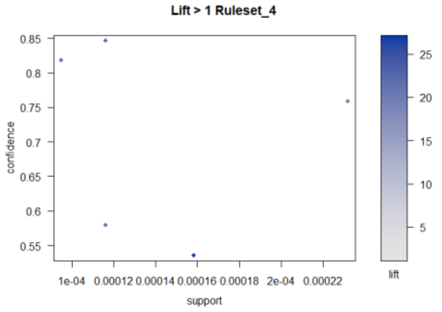
And from the above graphs, it can be inferred that the greater the lift and max conf the points are represented in darker blue shade and the one with less confidence pale shade.

## Plots between min_sup and min_conf (given) for lift values greater than 1:









Thus in the similar manner, all other combinations for lift > 1 are plotted.

In all the above plots the darker points reflects that for the lift >1 and higher confidence close to 1.

## Division Of Labor:

- We divided the entire project into two few components and have individually approached each component and came up with a feasible solution to the problems by carrying out discussions.

## Problems encountered:

- During pre-processing initially, we could not figure out how to clean the dataset and merge based on tid's.
- As, with the given minimum support values we could not generate the rules, we kept on lowering the minimum value until we were able to generate rules.