**Project Report**

**Overall Status:**

**Brief overview**

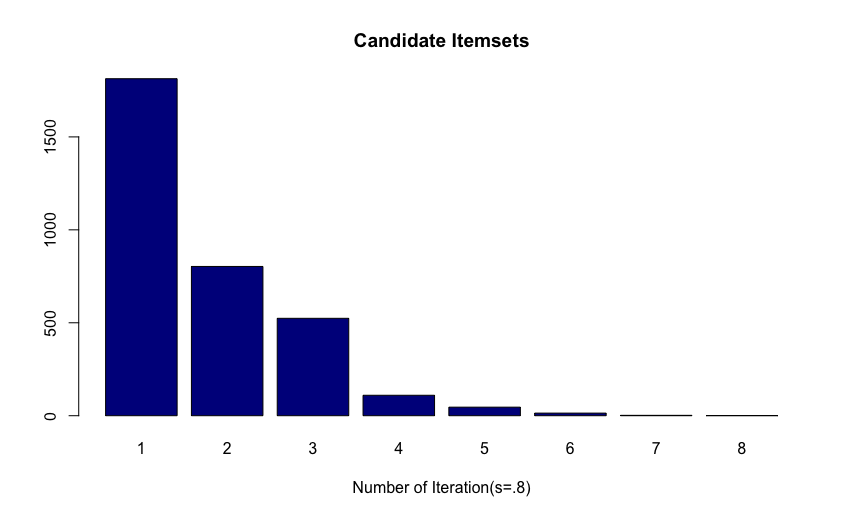
* **At first, we started with data pre-processing.**
* **We have started by combining the data of both the file. Then refine the necessary data for our team for the project. That is data for the years 2010 to 2013, but our file had data from 2009 till 2013.**
* **We then separated all the column's data with respective Tid's. Removed all the '?' and duplicate records.**
* **Then combine all the records with building 'single' format data set.**
* **Then converted this single data set in the form of a basket.**
* **Then calculated the frequent itemset , candidate itemset, and transition rules using support and confidence values provided**

**Status:**

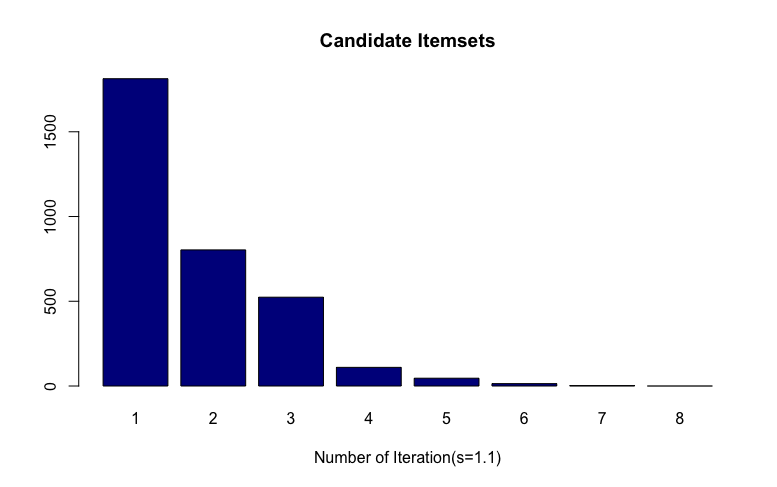
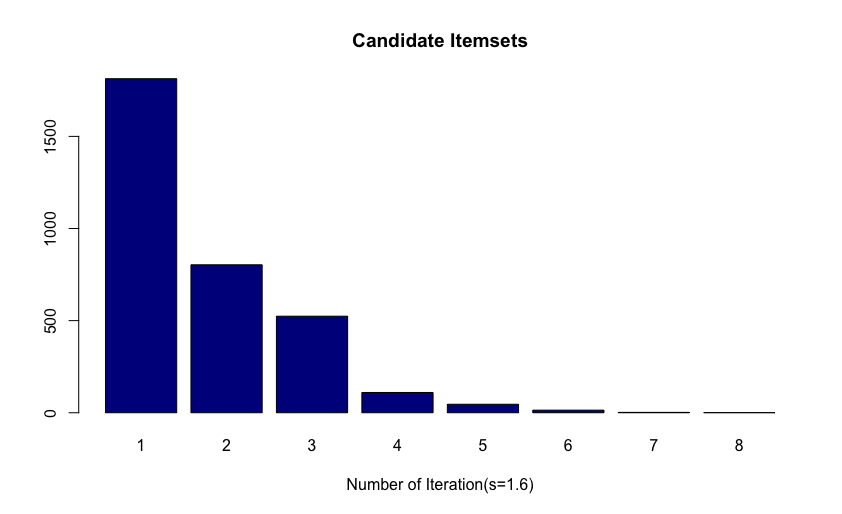
* The project is complete by matching all the parameters mentioned in the document.

**Analysis:**

1. By using each min\_sup value given to us we have, ploted the candidate itemsets and showed how it changes over iterations.
   1. When support = .0008

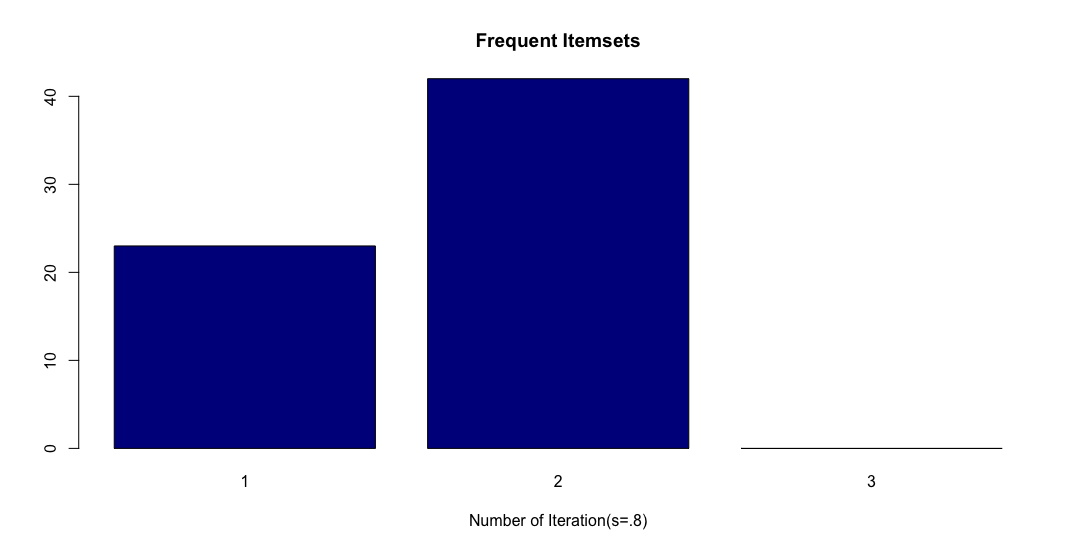


* 1. When support = .0001

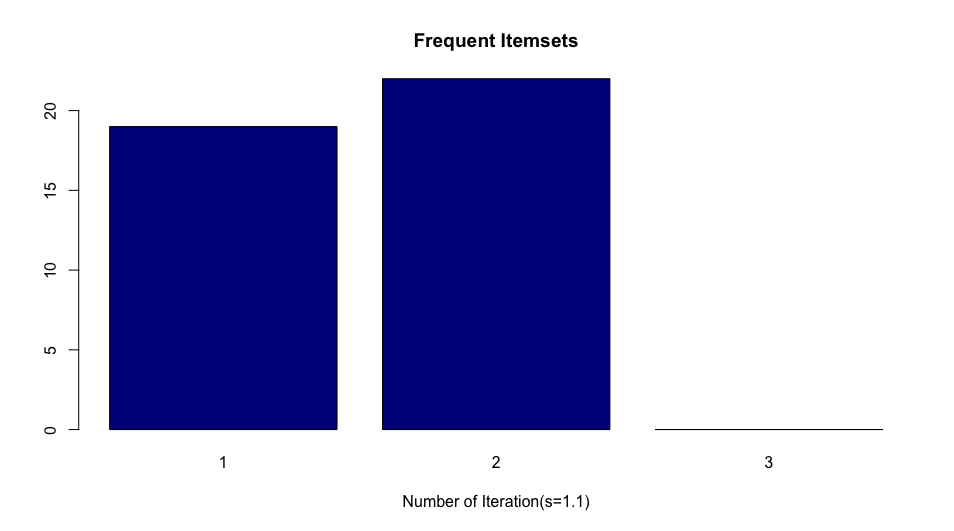


From this, we can tell that as the support values decrease, the item set increases for each iteration. After some iteration, it is observed that there is no itemset. But if we have very little support value, then the item set gradually decreases to zero.

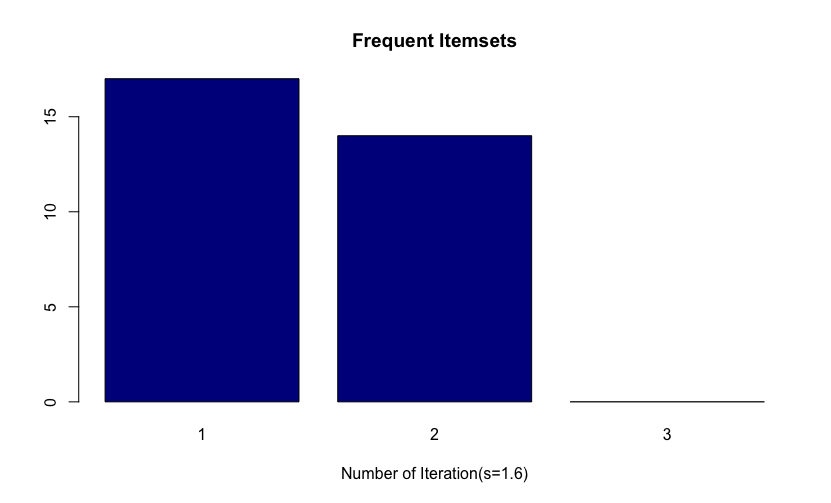
1. By using each min\_sup value given to us we have, ploted the Frequent itemsets and showed how it changes over iterations.
   1. When support = 0.8% and confidence=0.5



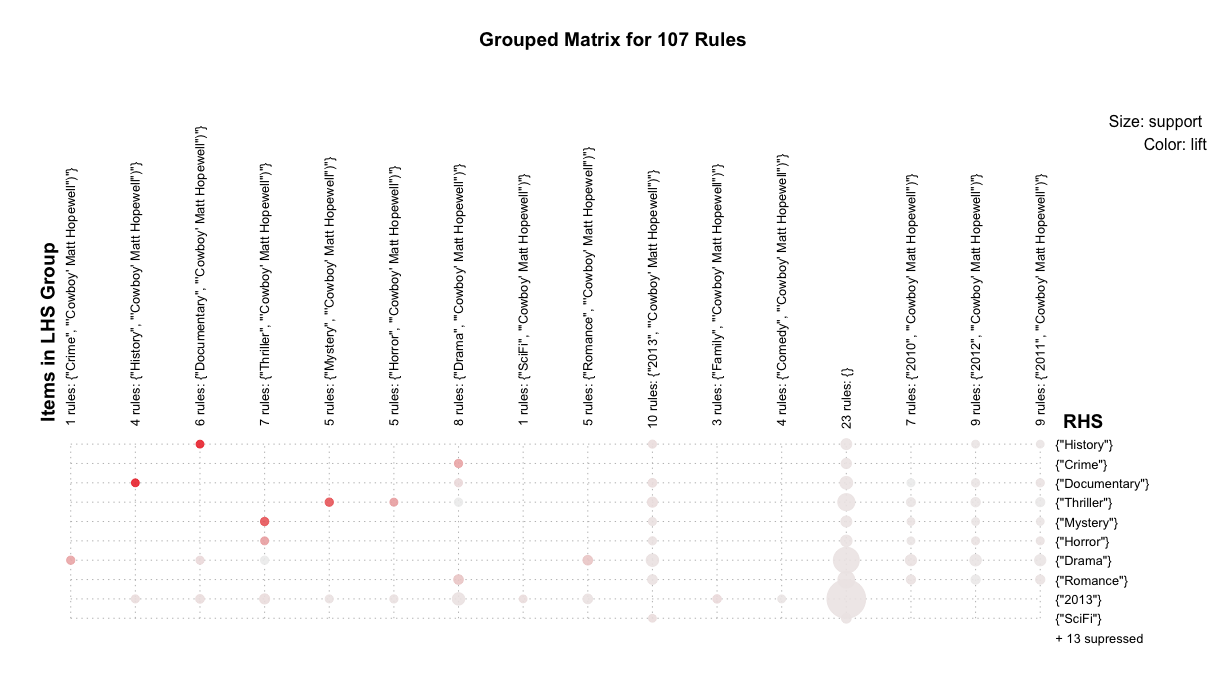
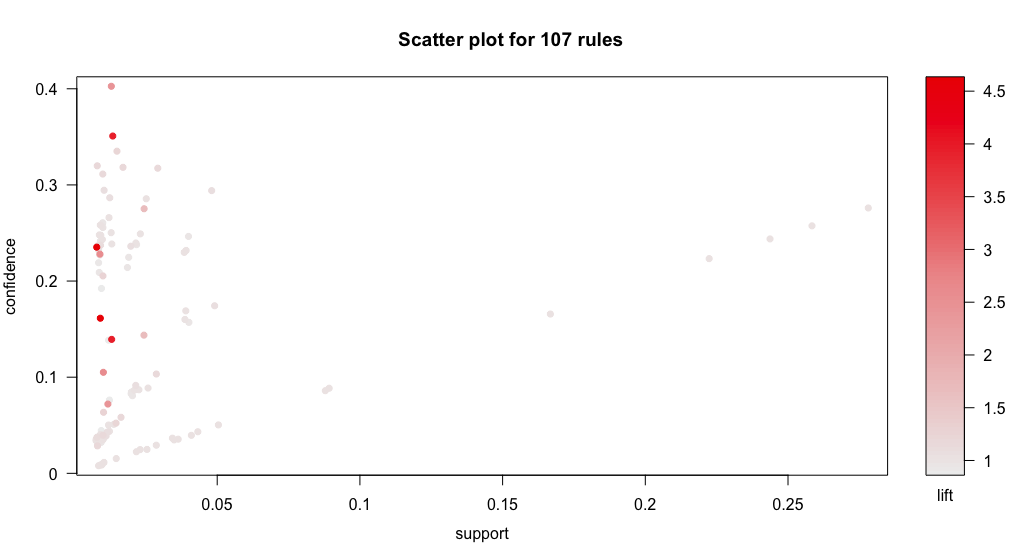
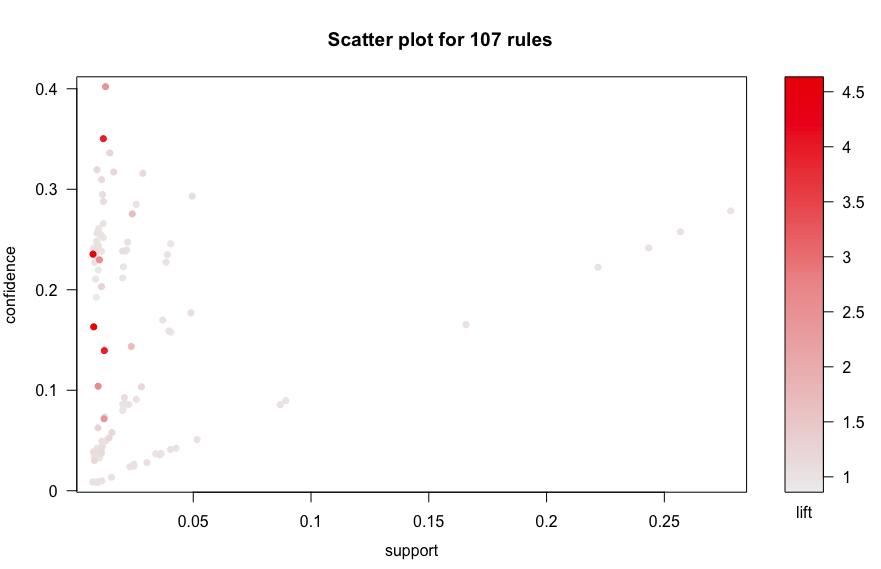
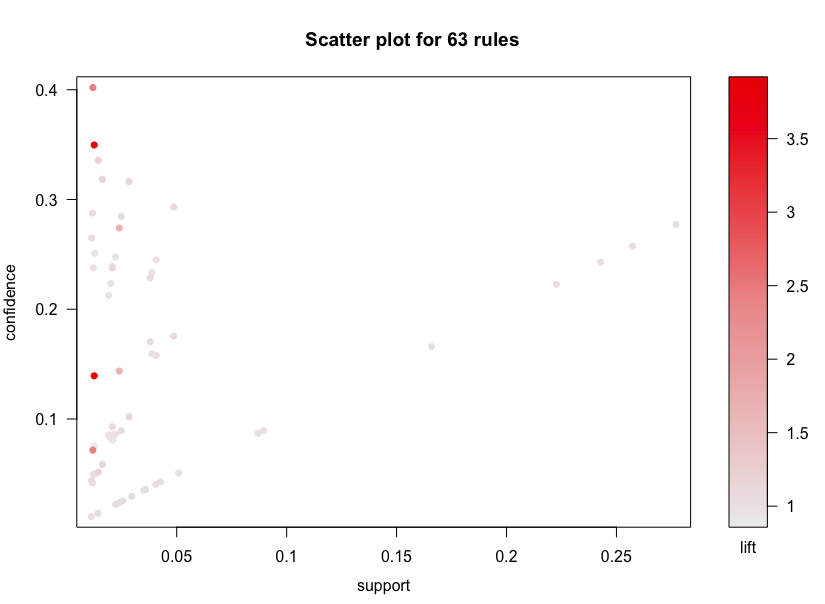
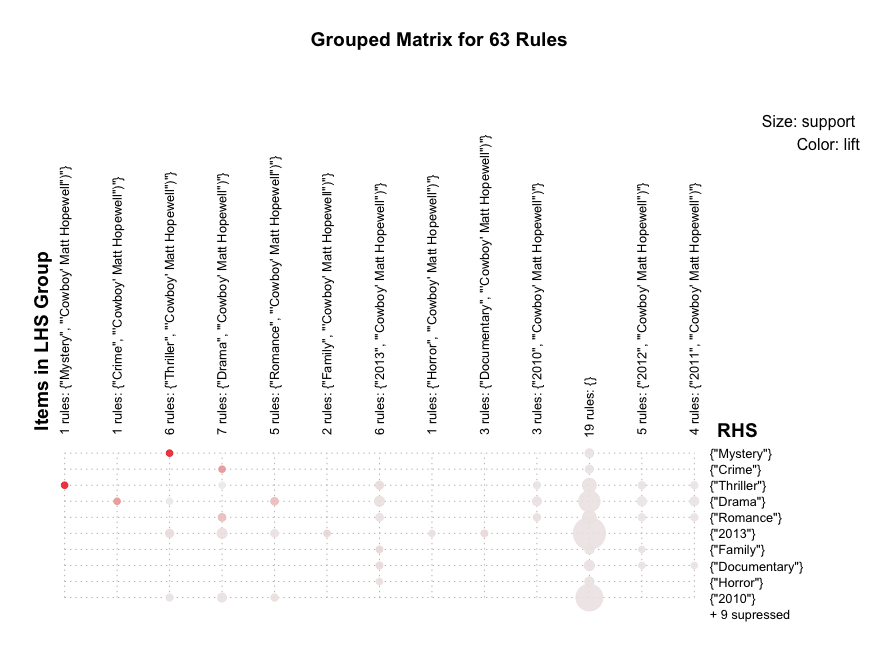
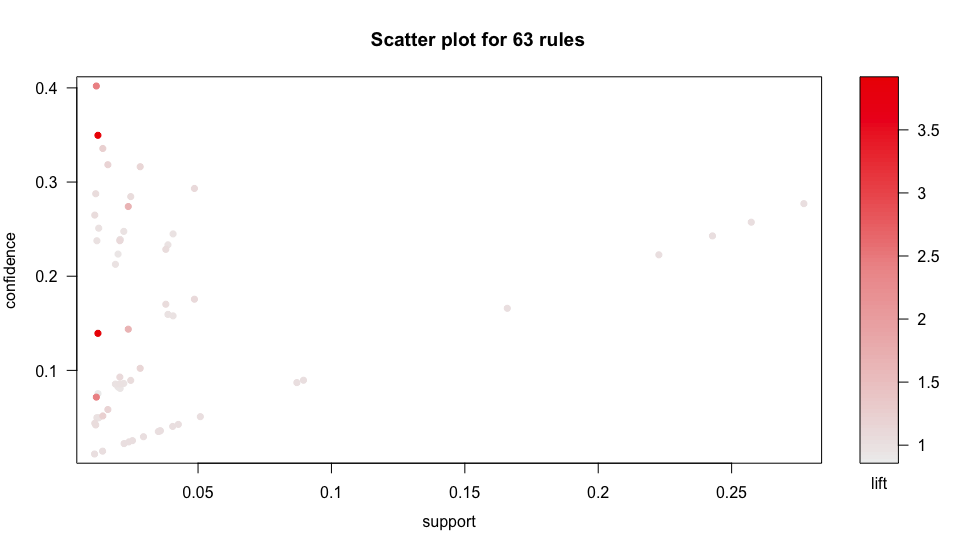
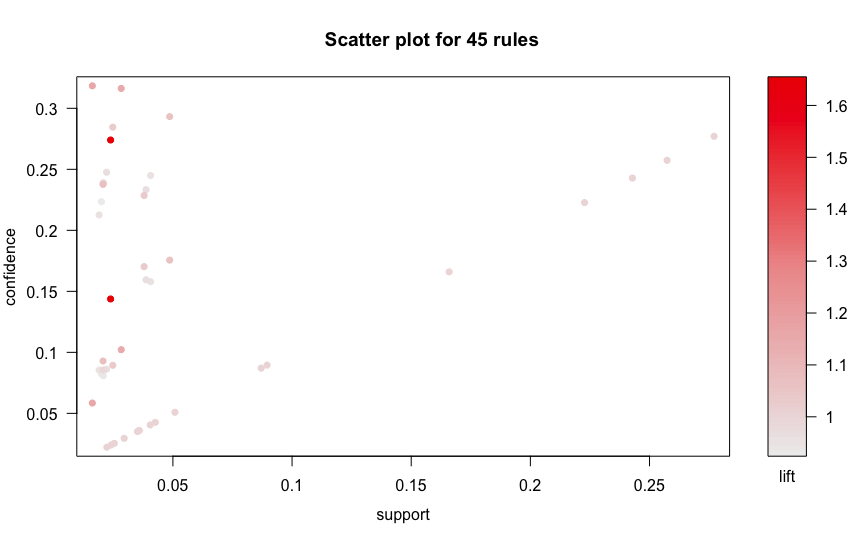
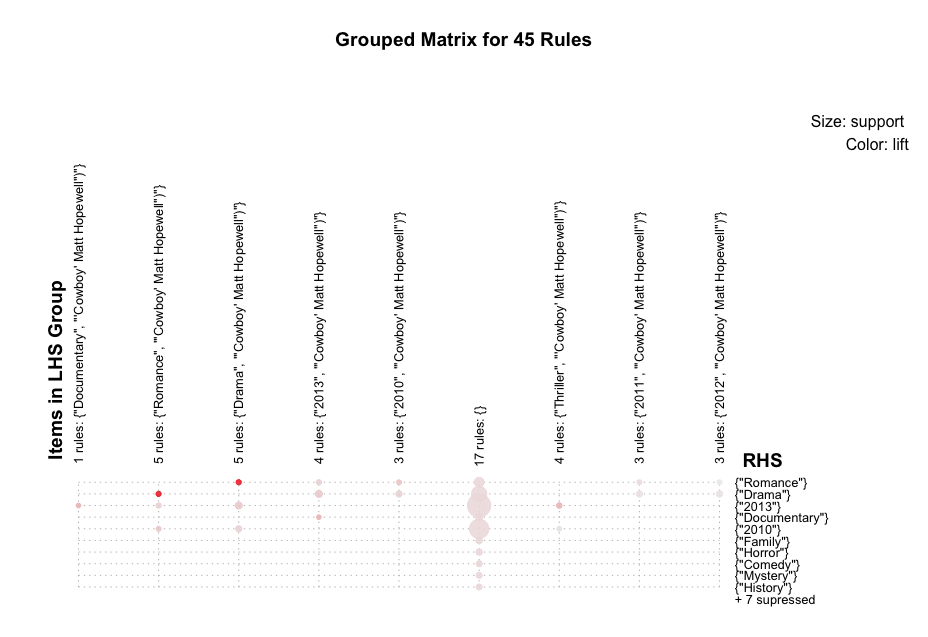
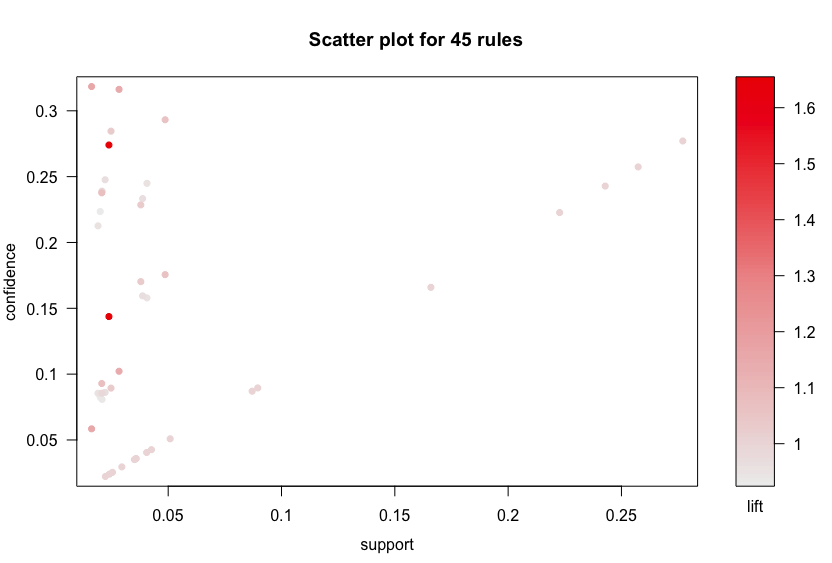
* 1. When support = 1.1% and confidence=0.7



* 1. When support = 1.6% and confidence=0.8



From this, we have analyzed that as the support increase, the frequency of item sets decreases.

1. Comparison of rules:
   1. Rules generated for support=.8% and confidence =.5:
   2. Rules generated for support=1.1% and confidence =.7:
   3. Rules generated for support=1.6% and confidence =.8:

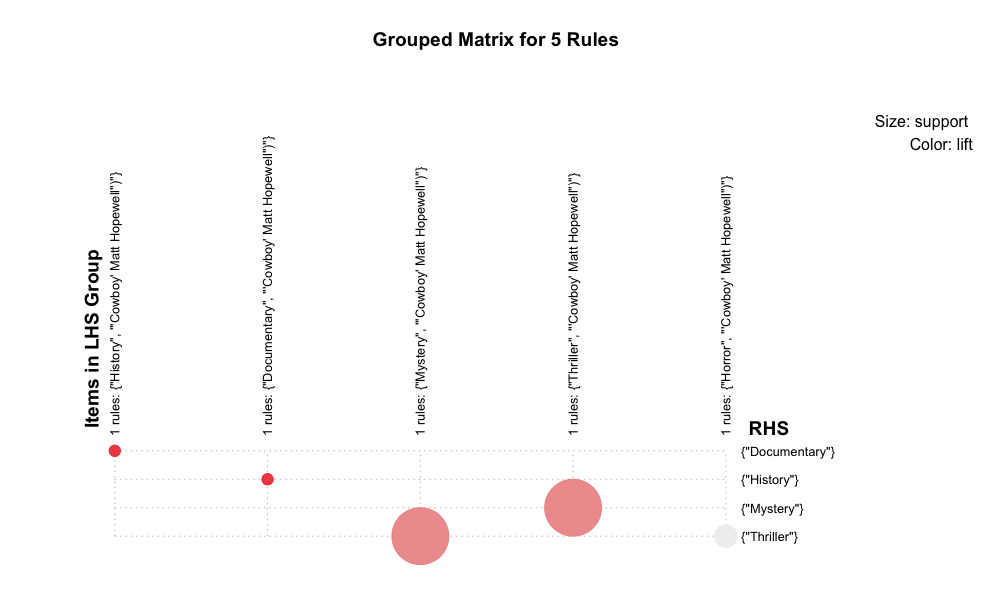
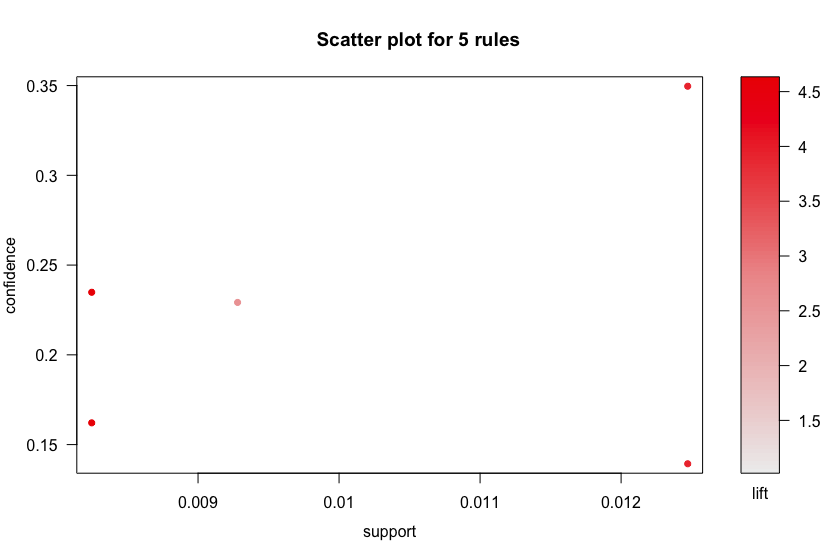
From this, we have analyzed that as the support increase, the number of rules decreases decreases.

Top 5 rules based on lift:

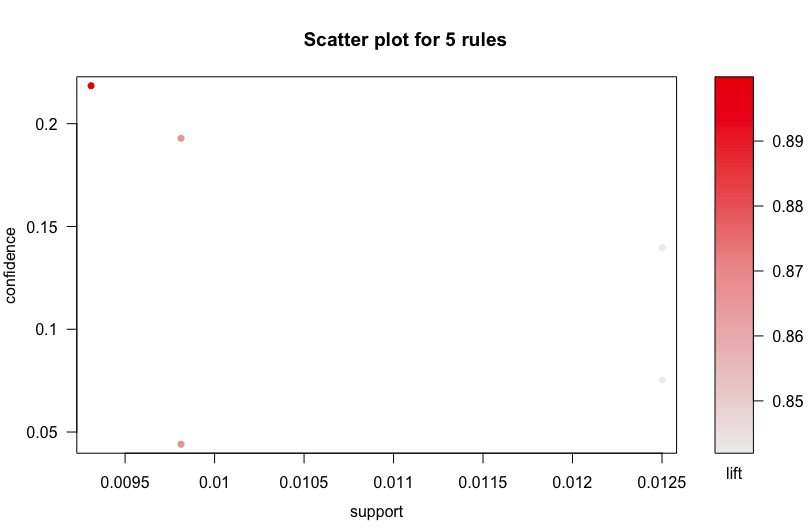
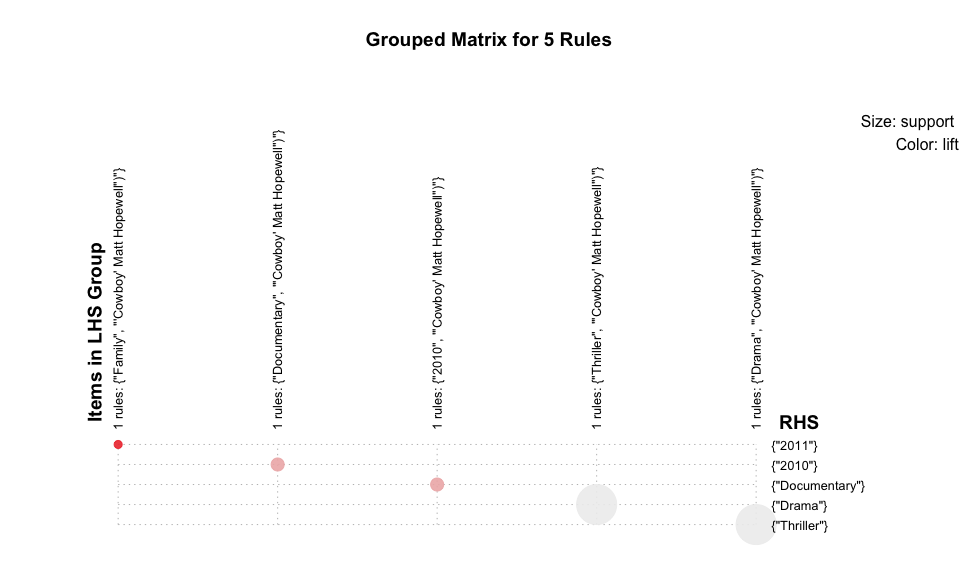
1. When support = 0.8% and confidence=0.5

* **Lift > 1:**

lhs                rhs             support     confidence lift      
[1] {"History"}     => {"Documentary"} 0.008245899 0.2348485  4.617140  
[2] {"Documentary"} => {"History"}     0.008245899 0.1621150  4.617140  
[3] {"Mystery"}     => {"Thriller"}    0.012472292 0.3496272  3.905459  
[4] {"Thriller"}    => {"Mystery"}     0.012472292 0.1393199  3.905459  
[5] {"Horror"}      => {"Thriller"}    0.009280331 0.2291971  2.560212



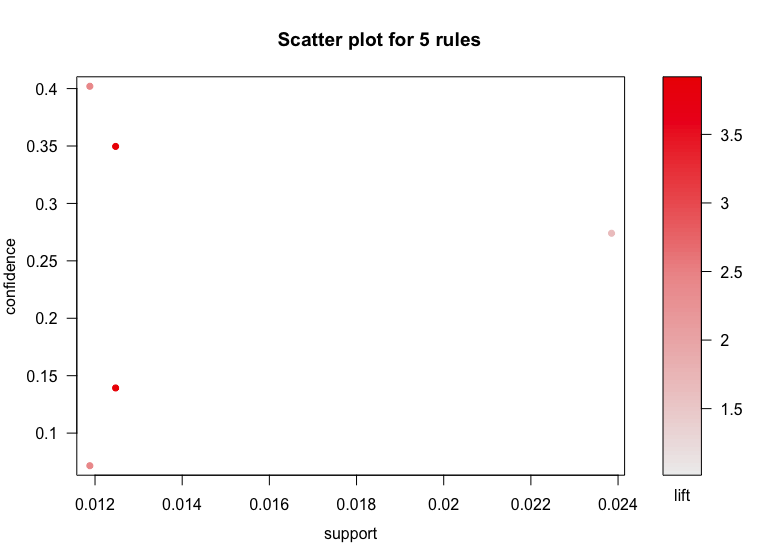
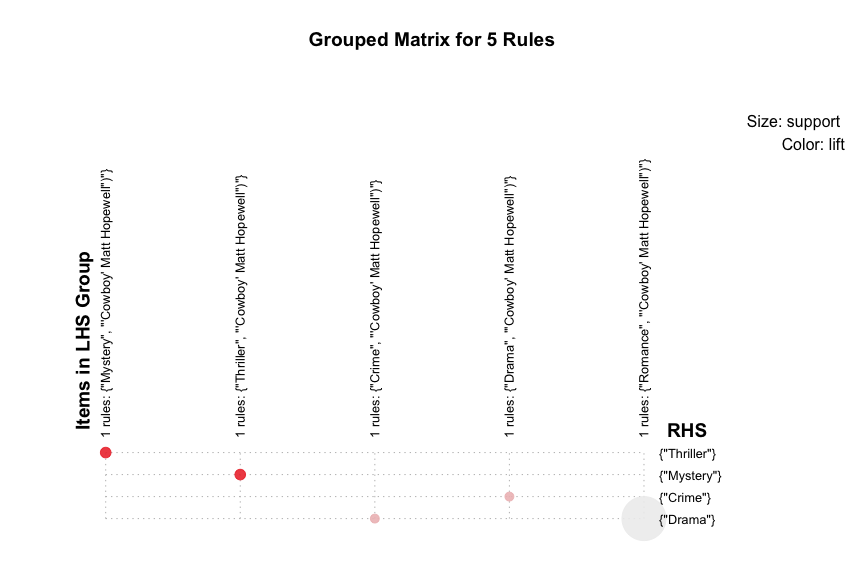
* **.Lift < 1**

lhs                rhs             support     confidence lift      
[1] {"Thriller"}    => {"Drama"}       0.012501847 0.13965005 0.8416565  
[2] {"Drama"}       => {"Thriller"}    0.012501847 0.07534735 0.8416565  
[3] {"2010"}        => {"Documentary"} 0.009812325 0.04405520 0.8661288  
[4] {"Documentary"} => {"2010"}        0.009812325 0.19291110 0.8661288  
[5] {"Family"}      => {"2011"}        0.009309886 0.21844660 0.8996033

1. When support = 1.1% and confidence=0.7

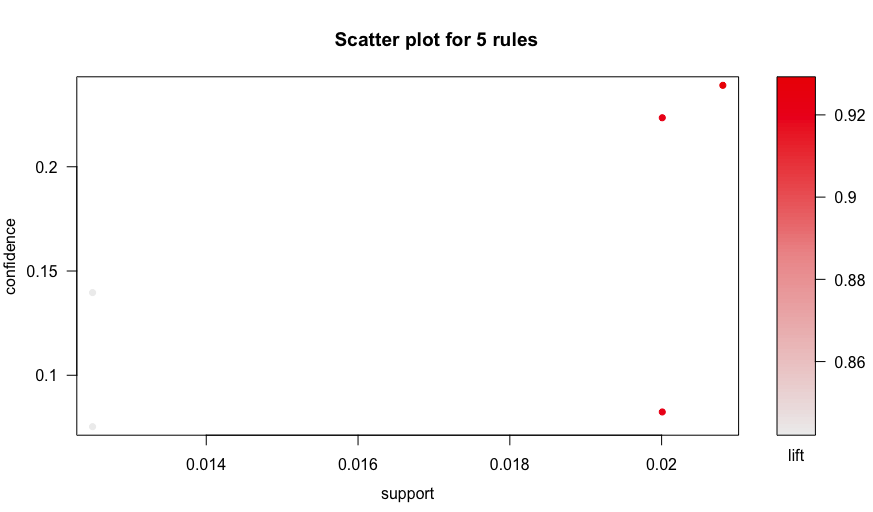
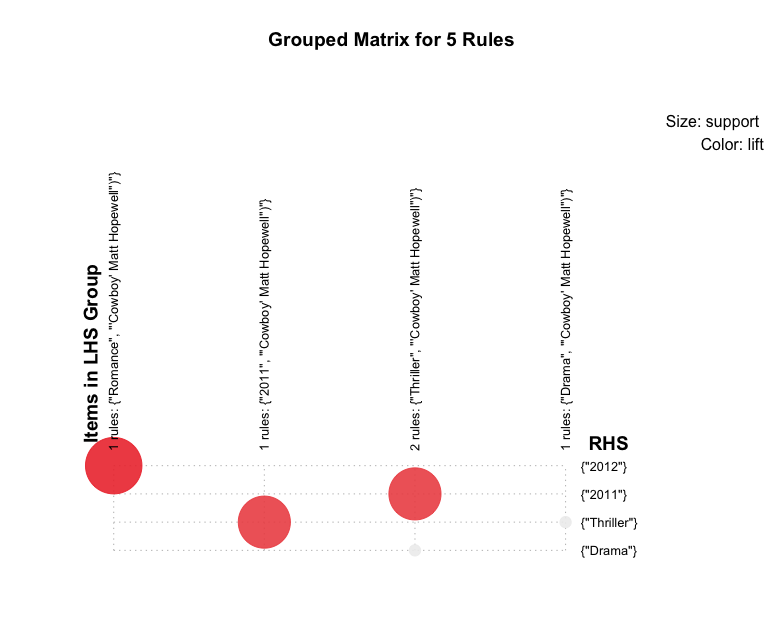
* **Lift > 1:**

hs             rhs          support    confidence lift     count  
[1] {"Mystery"}  => {"Thriller"} 0.01247229 0.3496272  3.905459 422    
[2] {"Thriller"} => {"Mystery"}  0.01247229 0.1393199  3.905459 422    
[3] {"Crime"}    => {"Drama"}    0.01188119 0.4020000  2.422813 402    
[4] {"Drama"}    => {"Crime"}    0.01188119 0.0716067  2.422813 402    
[5] {"Romance"}  => {"Drama"}    0.02385104 0.2740238  1.651513 807



* Lift < 1

lhs             rhs          support    confidence lift      count  
[1] {"Thriller"} => {"Drama"}    0.01250185 0.13965005 0.8416565 423    
[2] {"Drama"}    => {"Thriller"} 0.01250185 0.07534735 0.8416565 423    
[3] {"Thriller"} => {"2011"}     0.02000887 0.22350611 0.9204393 677    
[4] {"2011"}     => {"Thriller"} 0.02000887 0.08240019 0.9204393 677    
[5] {"Romance"}  => {"2012"}     0.02080686 0.23904924 0.9288276 704



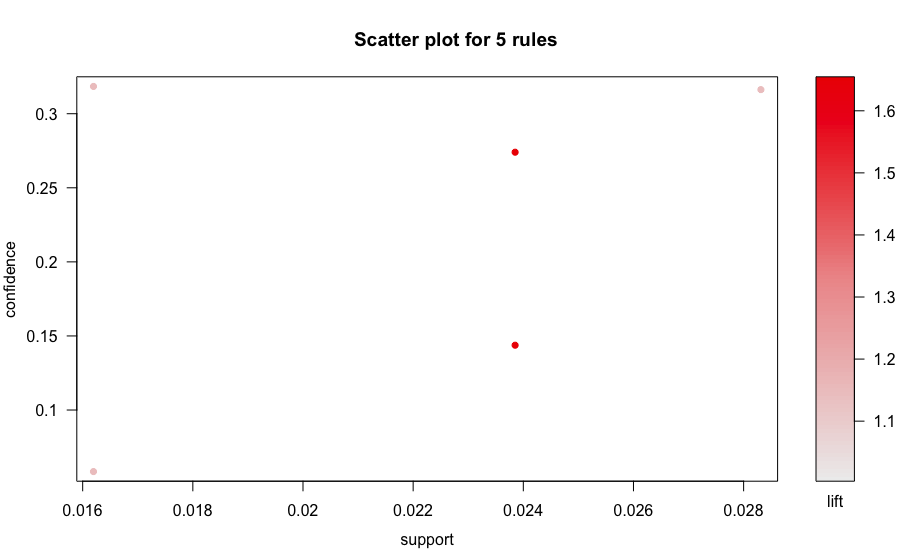
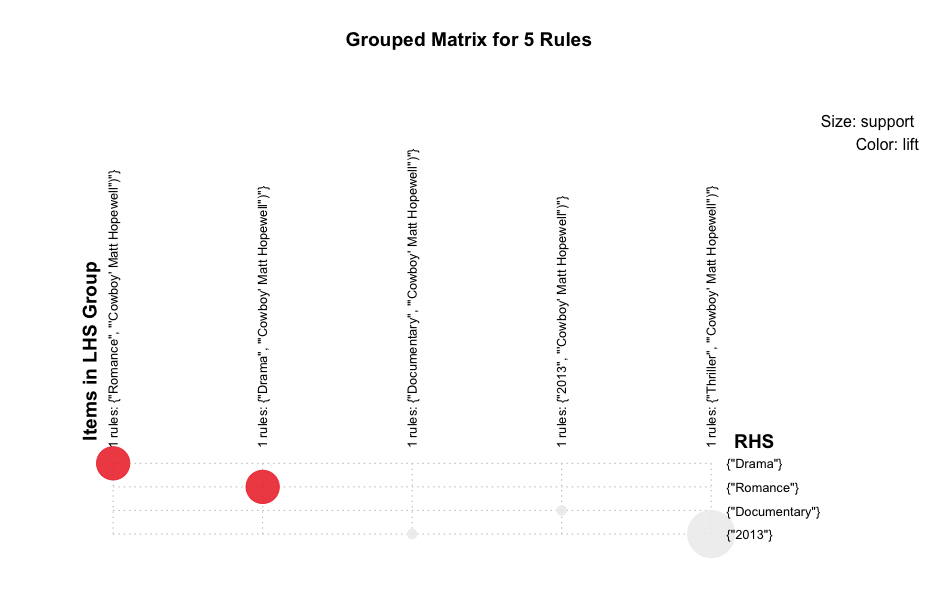
* Lift = 0:
* Because all the rules generated are having lift more than 1

1. When support = 1.6% and confidence=0.8

* **Lift > 1:**

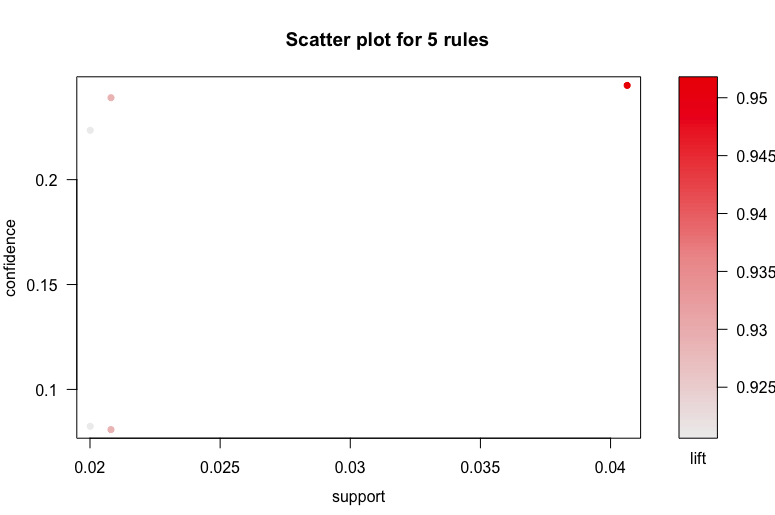
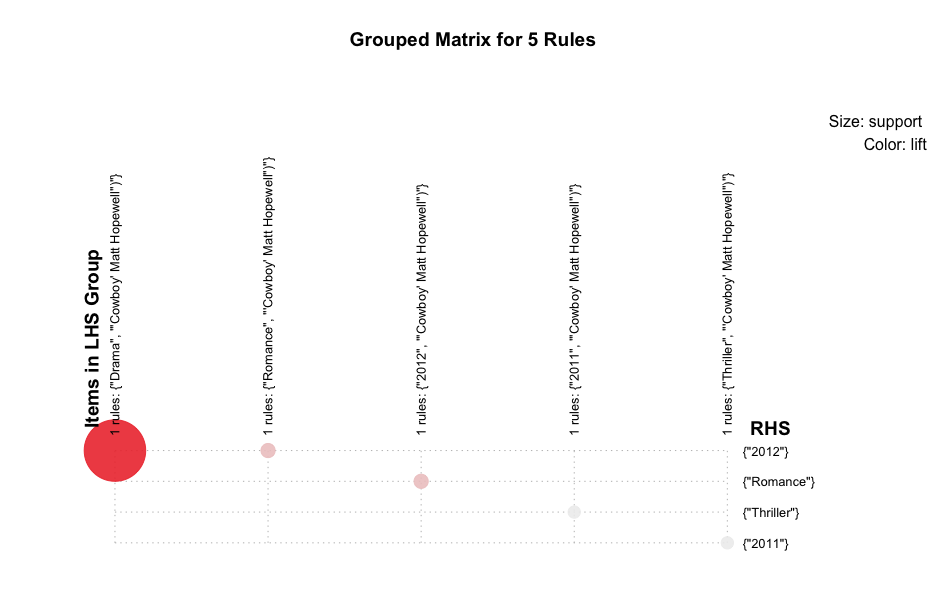
lhs                rhs             support    confidence lift      
[1] {"Romance"}     => {"Drama"}       0.02385104 0.27402377 1.651513  
[2] {"Drama"}       => {"Romance"}     0.02385104 0.14374777 1.651513  
[3] {"Documentary"} => {"2013"}        0.01619625 0.31841952 1.149197  
[4] {"2013"}        => {"Documentary"} 0.01619625 0.05845333 1.149197  
[5] {"Thriller"}    => {"2013"}        0.02831388 0.31627600 1.141461





* Lift < 1

  lhs             rhs          support    confidence lift      count  
[1] {"Thriller"} => {"2011"}     0.02000887 0.22350611 0.9204393  677  
[2] {"2011"}     => {"Thriller"} 0.02000887 0.08240019 0.9204393  677  
[3] {"Romance"}  => {"2012"}     0.02080686 0.23904924 0.9288276  704  
[4] {"2012"}     => {"Romance"}  0.02080686 0.08084520 0.9288276  704  
[5] {"Drama"}    => {"2012"}     0.04063839 0.24492341 0.9516517 1375



Problem Encountered:

1. For preprocessing we first did single format. Later on realized it is not producing proper result , So we had to generate basket format of data.
2. How to find candidate item set was tricky. It was generating any.So generated it as per the explanation given by professor on discussion board.
3. To figure out what format to use to visualize result.