PROJECT 3 REPORT

OVERALL STATUS:

The data was preprocessed into a single format file, frequent items were generated and candidate itemsets were generated manually from them. Graphs and data for each step have been compiled and rules were visualized as matrix-based grouped plots.

FILE DESCRIPTIONS:

preprocess.R – To preprocess data into a single file readable by read.transactions method

apriori.R – To perform apriori to generate rules and frequent itemsets

candidateGen.R – To generate candidate itemsets from frequent itemsets

dataplot.R – To plot data for frequent items and rules

DIVISION OF LABOR:

Single member team so no division of labor. About 15-20 hours were spent to complete the project

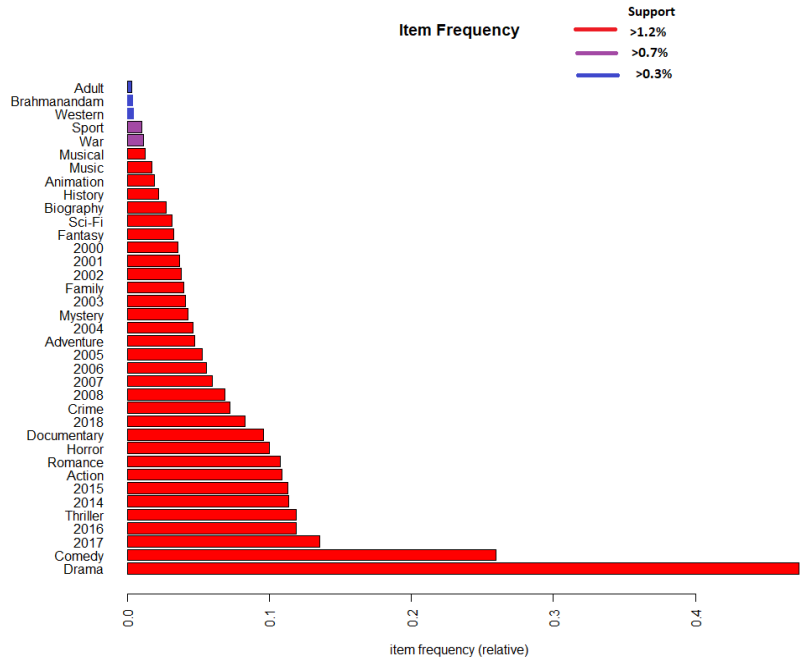
PROBLEMS ENCOUNTERED:

Mostly in preprocessing where it took time for trial and error methods to get the data in a suitable format.

Other problems were choices of plots for rule visualization and approach to find candidate itemsets since I was not satisfied with decreasing the support to get the candidate itemsets.

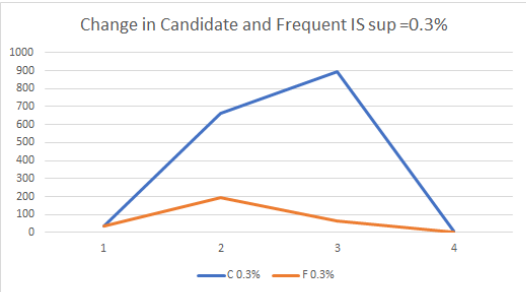
RESULTS AND ANLYSES:

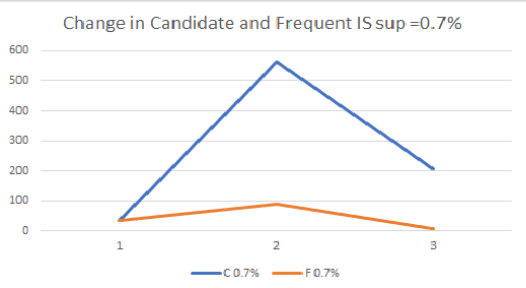
1. Change in candidate and frequent item sets

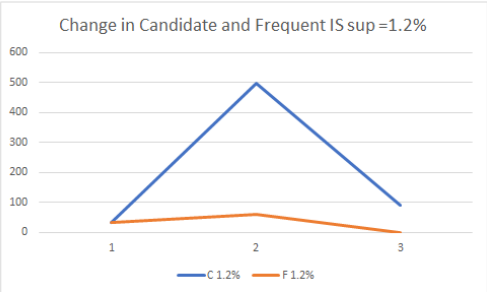


The above plot shows a frequency graph for all the frequent 1-itemsets by support. Lower support values have frequent itemsets that are supersets to frequent itemsets of higher support values.

A total of 37,34 and 32 frequent 1-itemsets were generated for the supports 0.3%,0.5% and 1.2% respectively with Drama and Comedy being the 2 highest by a large margin. Due to their frequency being more than 25%, we can expect them to be in one or more rules.





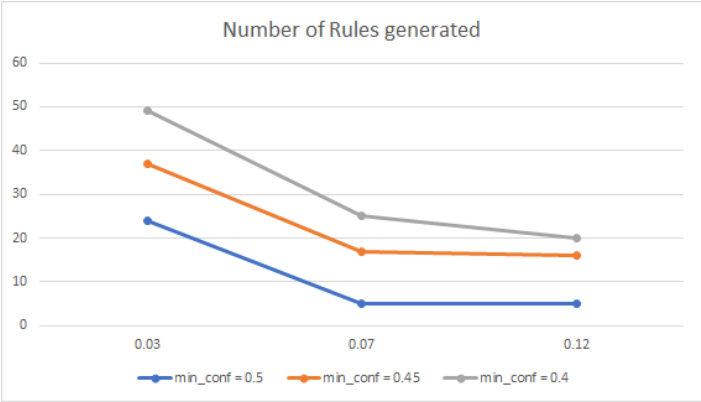


Each candidate itemset was generated using the frequent itemset of the previous iteration lexicographically (written as a function) without pruning included.

As we can expect, a lower support allows for more itemsets as shown in the difference between 0.3% and the rest for candidate and frequent 2,3-itemsets. However, the change in candidate itemsets between 0.7% and 1.2% is fairly small even though the change in support is larger. This can be attributed to the support count of most frequent 1-itemsets being well above the 1.2% mark thereby having decent joint support as well.

2. Number of Rules generated for each min_conf and min_sup

For the provided values of min_conf = 0.5,0.7,0.8 and min_sup=0.3,0.7,1.2, no rules were generated for confidence values =0.7 and 0.8. On decreasing the min_sup to lower than 0.1 there were not enough rules generated for the same. Hence, the **min_conf values were reduced to min_conf =0.4,0.45 and 0.5** to be able to generate rules.



The difference in number of rules generated across confidences is almost uniform especially between min_conf =0.5 and 0.45. While the difference is uniform, the difference in number of rules generated is considerably large, meaning that a fair amount of rules have a confidence between 45% and 50%.

Moreover, the number of rules for supports 0.7% and 1.2% is almost the same thereby supporting the previous assumption of majority of frequent itemsets having support more than 1.2%

3. Rules sorted by Lift

"**Lift >1"**

	lhs	rhs	support	confidence	lift	count
[1]	{2014,Biography}	=> {Documentary}	0.004585412	0.6980108	7.252352	386
[2]	{Biography}	=> {Documentary}	0.011404134	0.4150454	4.312333	960
[3]	{Biography,Documentary}	=> {2014}	0.004585412	0.4020833	3.537930	386
[4]	{2014,Romance}	=> {Comedy}	0.004906153	0.4033203	1.554627	413
[5]	{2008,Romance}	=> {Drama}	0.004454740	0.6117455	1.293206	375

From the above, we can deduce that Biographies that are Documentaries were very common movies in 2014 as each of their generated rules have high lift values. However, the reverse, i.e., Documentaries were not mostly Biographies.

Due to the High support of Comedy and Drama it is expected that they have at least one or more rules as shown. 2014 was popular for Romances movies that we comedies and 2008 was very popular for Romance movies that were Drama (60% confidence).

" Lift < 1"

	lhs	rhs	support	confidence	lift	count
[1]	{Musical}	=> {Drama}	0.005119981	0.4054563	0.8571183	431
[2]	{2018}	=> {Drama}	0.034141126	0.4098688	0.8664462	2874
[3]	{Comedy,Romance}	=> {Drama}	0.017355666	0.4151748	0.8776628	1461
[4]	{Crime,Thriller}	=> {Drama}	0.007412687	0.4216216	0.8912912	624
[5]	{Sport}	=> {Drama}	0.004502257	0.4282486	0.9053004	379

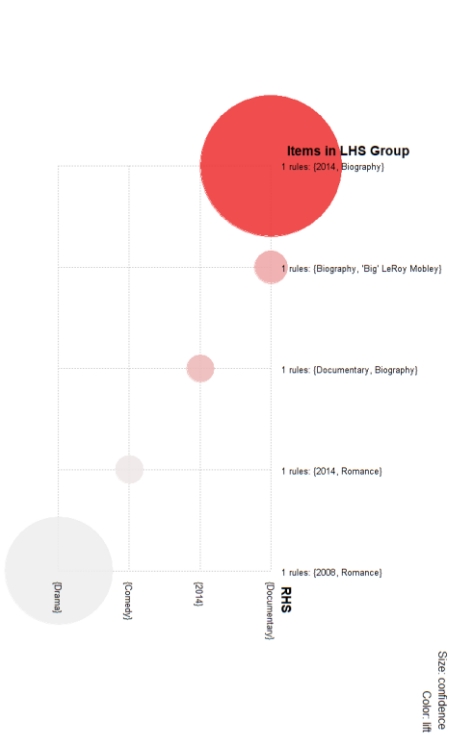
From the above Musicals, movies in 2018, Romance-Comedies, Crime-Thrillers and Sports movies, all of which were Dramas was not very likely although it would be intuitively expected to be otherwise. This may be because of the overwhelmingly large support of Dramas and lesser confidence of the other genres with Drama.

"Lift = 1"

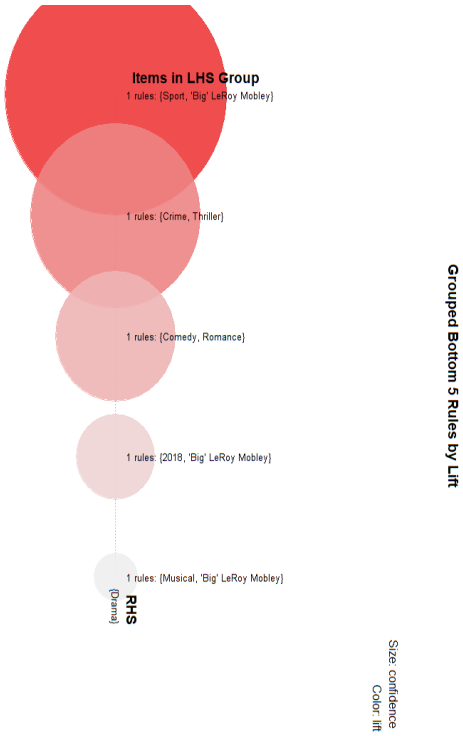
lhs	rhs	support	confidence	lift	count
[1] {Crime,Mystery}	=> {Drama}	0.003813257	0.4893293	1.0344225	321
[2] {2000}	=> {Drama}	0.017486339	0.4888741	1.0334603	1472
[3] {2015}	=> {Drama}	0.053302447	0.4711256	0.9959406	4487
[4] {Biography}	=> {Drama}	0.012758375	0.4643320	0.9815793	1074
[5] {2014}	=> {Drama}	0.052744120	0.4640953	0.9810789	4440

For Lift = 1, i.e. no assosciation, Crime-Mysteries, movies in 2000, 2014 or in 2015, and Biographies may were not associated with Dramas possibly meaning that while Dramas were popular these genres were not limited to Dramas and were likely to be better associated with a different genre

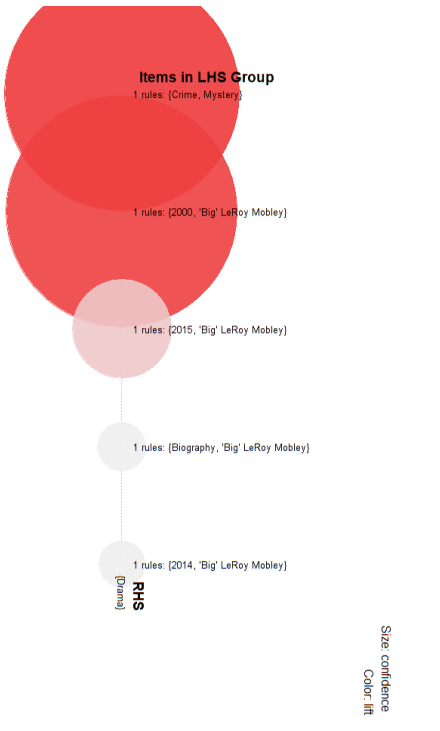
4. Visualization of Rules by confidence(size being relative)



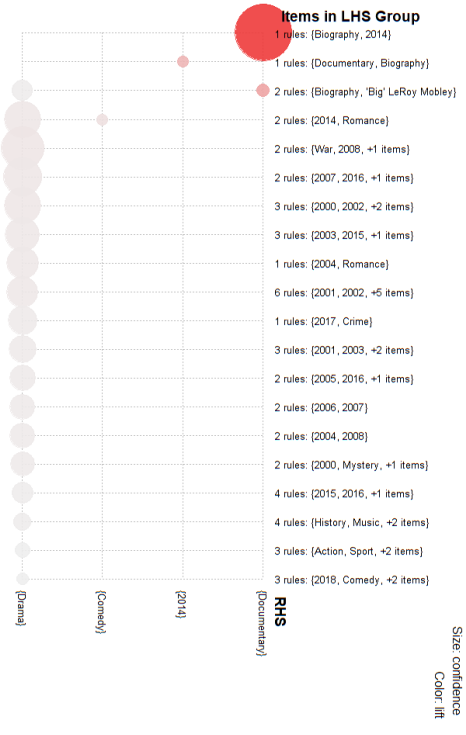
Grouped Top 5 Rules by Lift



Grouped Bottom 5 Rules by Lift



Grouped Rules with Lift closest to 1



Grouped All Rules by confidence