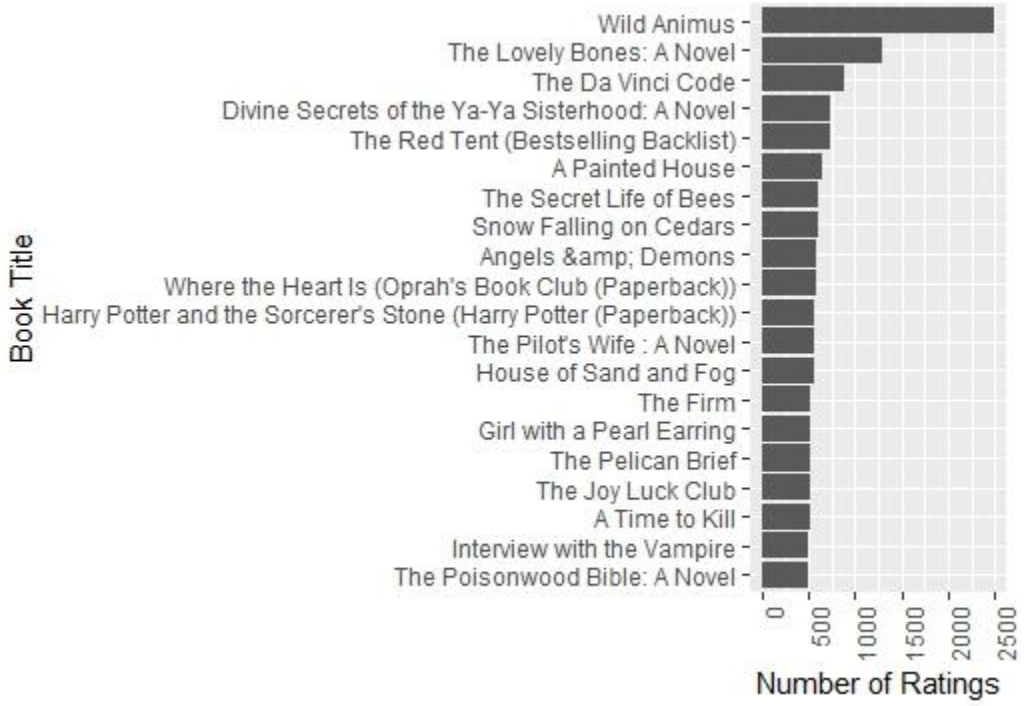
Today I wanted to use the arules and the arulesvis packages in R in order to see the results of the book data that I downloaded. I wanted to use these packages to explore the data and then look into setting the minimum support to 0.005, and minimum confidence of 0.70. I wanted to first get to know my data and then use the apriori algorithms to display the top ten rules and summary information. I then wanted to adjust the rules and use the aruleViz library to then display rules. This would help to then describe and discuss the association rules of the dataset and how they can be changed after transforming the data into a transaction set.

First, we were want to see the structure of the data and describe what we saw along with some initial inferences. First since we have to look into the top ten books, the top 20 books were visualized out of curiosity. This was done after libraries that we are going to need to help us summarize the dataset (arules, arulesViz, and datasets) were uploaded.



We then saw the top 20 books to get a better idea of their ratings. After seeing the above, we then want to use the arules package to find the associations in the data. So, in order to do that we first need to transform the data into a transaction data format. However, once I did this I found that the data, since it was over a million observations, needed to be sampled in order to get R to run faster and correctly. A sample of the data was then taken in order to help the program run a bit more smoothly and faster.

Once this was accomplished a data frame named samplebooks was created with 50,000 observations. This dataset included the transformed data in order to use it in the apriori steps. The apriori algorithm was then used on the sample dataset named samplebooks in order to continue with the assignment. The initial parameters for the rules dataset for support and confidence was set too high to capture any rules. A resizing of the confidence/supports had to be completed in order to produce rules. This had to happen 3 times before I received 174 rules to proceed with the assignment. All those attempts are below:

```
> rules <- apriori(samplebooks, parameter = list(supp = 0.2, conf = 0.5, target = "rules"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5     0.2      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 10000

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[52262 item(s), 50000 transaction(s)] done [0.08s].
sorting and recoding items ... [1 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 done [0.00s].
writing ... [1 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

```
> rules <- apriori(samplebooks, parameter = list(supp = 0.005, conf = 0.7, target = "rules"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.7    0.1    1 none FALSE            TRUE       5   0.005      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 250

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[52262 item(s), 50000 transaction(s)] done [0.10s].
sorting and recoding items ... [12 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 done [0.00s].
writing ... [1 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> summary(rules)
set of 1 rules
```

```
> summary(rules)
set of 1 rules

rule length distribution (lhs + rhs):sizes
2
1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support         confidence       lift            count
 Min.   :0.0071   Min.   :1    Min.   :1.607   Min.   :355
 1st Qu.:0.0071   1st Qu.:1    1st Qu.:1.607   1st Qu.:355
 Median :0.0071   Median :1    Median :1.607   Median :355
 Mean   :0.0071   Mean   :1    Mean   :1.607   Mean   :355
 3rd Qu.:0.0071   3rd Qu.:1    3rd Qu.:1.607   3rd Qu.:355
 Max.   :0.0071   Max.   :1    Max.   :1.607   Max.   :355

mining info:
        data ntransactions support confidence
 samplebooks          50000   0.005        0.7

> inspect(rules)
    lhs                 rhs              support confidence lift     count
[1] {User.ID=198711} => {Book.Rating=0} 0.0071  1          1.607407 355
```

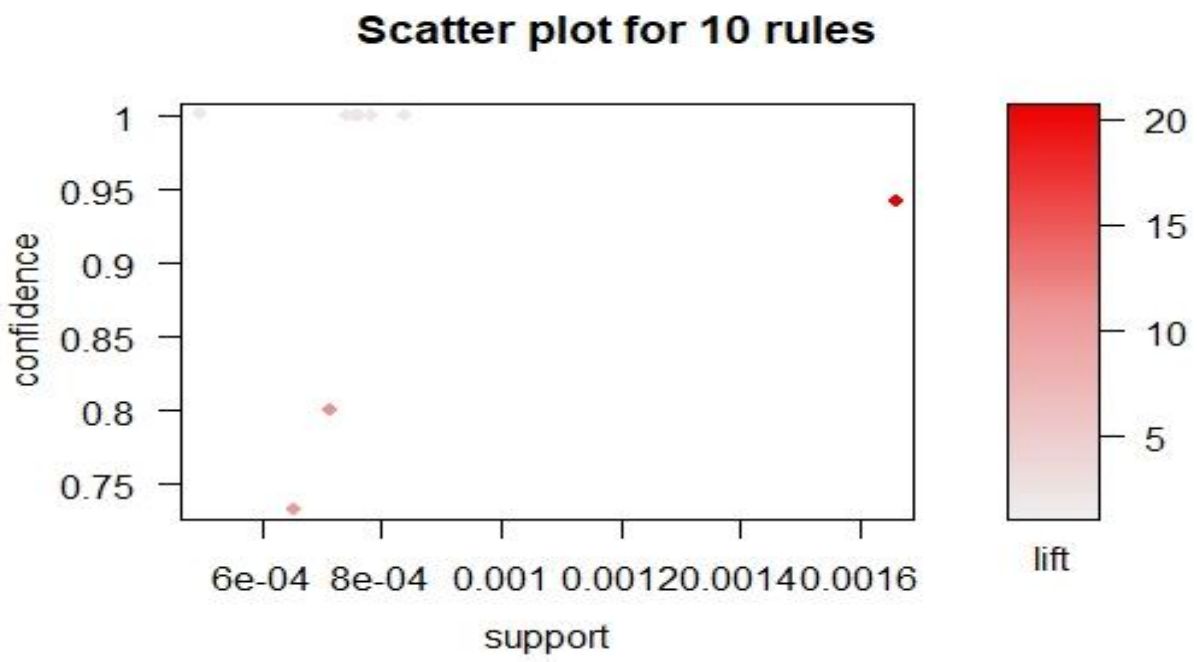
```
> rules <- apriori(samplebooks, parameter = list(supp = 0.0005, conf = 0.5, target = "rules"))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
        0.5    0.1    1 none FALSE            TRUE       5   5e-04      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 25

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[52262 item(s), 50000 transaction(s)] done [0.16s].
sorting and recoding items ... [283 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 done [0.00s].
writing ... [174 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].

> summary(rules)
set of 174 rules

rule length distribution (lhs + rhs):sizes
  1   2
  1 173

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   1.994   2.000   2.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.000500   Min.   :0.5714   Min.   : 0.9185   Min.   :   25.00
 1st Qu.:0.000625   1st Qu.:0.8000   1st Qu.: 1.2874   1st Qu.:   31.25
 Median :0.000850   Median :0.8779   Median : 1.4151   Median :   42.50
 Mean   :0.004661   Mean   :0.8643   Mean   : 1.6101   Mean   :  233.03
 3rd Qu.:0.001260   3rd Qu.:0.9615   3rd Qu.: 1.5508   3rd Qu.:   63.00
 Max.   :0.622120   Max.   :1.0000   Max.   :20.6025   Max.   :31106.00

mining info:
        data ntransactions support confidence
 samplebooks         50000   5e-04        0.5
```

Once enough rules were gathered to do an analysis, they were sorted by the lift in order to add some structure to the data before plotting any of it. This is shown below.

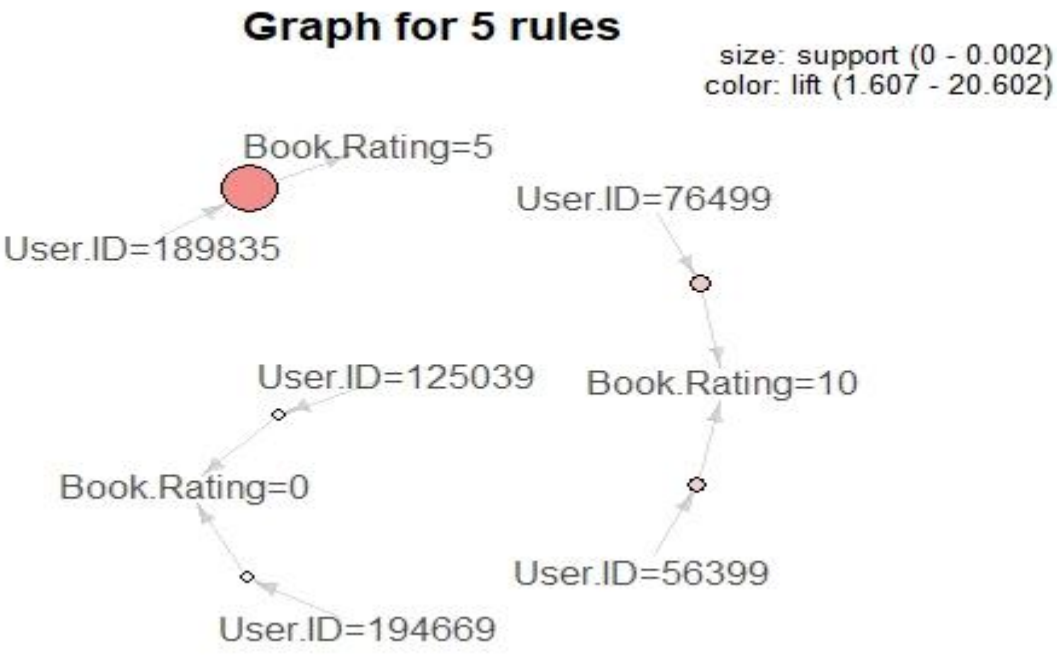
```
> rules.sorted <- sort(rules, by="lift")
> inspect(rules.sorted[1:20])
     lhs                   rhs                 support confidence lift      count
[1]  {User.ID=189835} => {Book.Rating=5}  0.00166 0.9431818  20.602486  83
[2]  {User.ID=76499}  => {Book.Rating=10} 0.00072 0.8000000  11.376564  36
[3]  {User.ID=56399}  => {Book.Rating=10} 0.00066 0.7333333  10.428517  33
[4]  {User.ID=125039} => {Book.Rating=0}  0.00050 1.0000000   1.607407  25
[5]  {User.ID=194669} => {Book.Rating=0}  0.00050 1.0000000   1.607407  25
[6]  {User.ID=203968} => {Book.Rating=0}  0.00074 1.0000000   1.607407  37
[7]  {User.ID=87746}  => {Book.Rating=0}  0.00076 1.0000000   1.607407  38
[8]  {User.ID=32195}  => {Book.Rating=0}  0.00076 1.0000000   1.607407  38
[9]  {User.ID=182987} => {Book.Rating=0}  0.00078 1.0000000   1.607407  39
[10] {User.ID=175886} => {Book.Rating=0}  0.00084 1.0000000   1.607407  42
[11] {User.ID=166596} => {Book.Rating=0}  0.00088 1.0000000   1.607407  44
[12] {User.ID=170229} => {Book.Rating=0}  0.00092 1.0000000   1.607407  46
[13] {User.ID=228998} => {Book.Rating=0}  0.00094 1.0000000   1.607407  47
[14] {User.ID=170518} => {Book.Rating=0}  0.00100 1.0000000   1.607407  50
[15] {User.ID=213350} => {Book.Rating=0}  0.00130 1.0000000   1.607407  65
[16] {User.ID=242824} => {Book.Rating=0}  0.00134 1.0000000   1.607407  67
[17] {User.ID=127429} => {Book.Rating=0}  0.00154 1.0000000   1.607407  77
[18] {User.ID=212898} => {Book.Rating=0}  0.00386 1.0000000   1.607407 193
[19] {User.ID=198711} => {Book.Rating=0}  0.00710 1.0000000   1.607407 355
[20] {User.ID=76352}  => {Book.Rating=0}  0.00278 0.9858156   1.584607 139
```

It was found at the end of the parameter shifting that the lower the confidence and the lower the support the more rules you will be able to get out of the data. Also, when it is sorted it is easier to analyze and to understand.

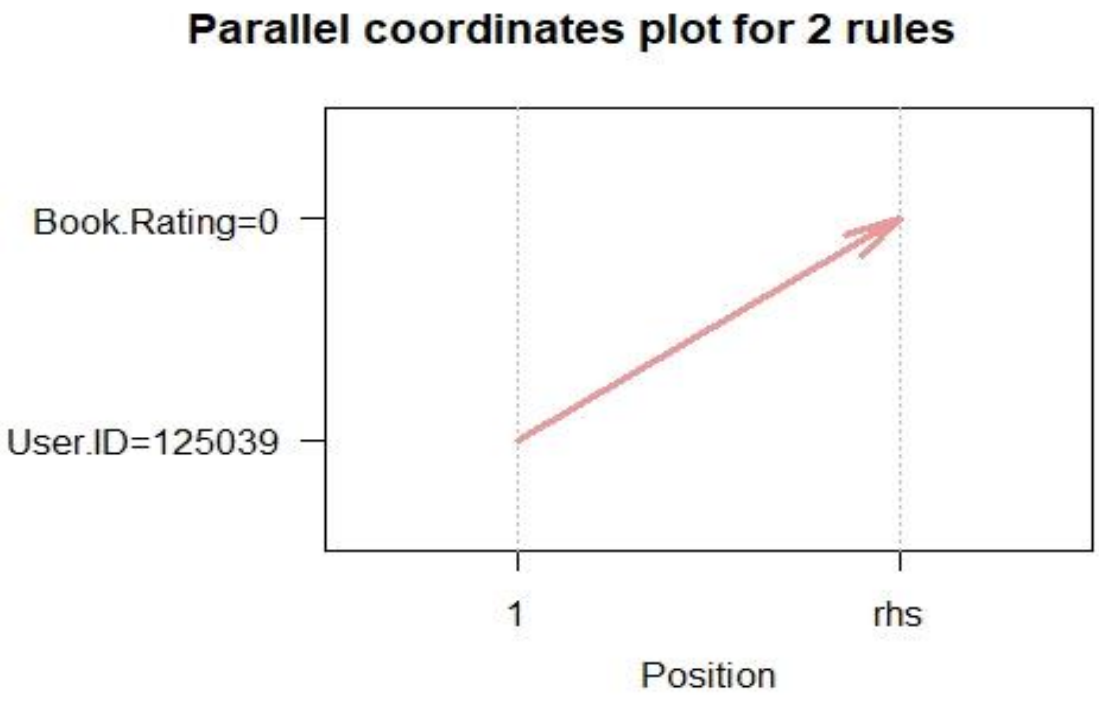
We then move to plot the points by using the library aruleViz to display the rules. I used a matrix as well as a paracoord method to look at the directionality as well as what the rules were doing (once sorted). I also looked at the plots and tried to determine where the ratings are going. The way these analyses were run, we would need to go through several iterations of these tests before feeling confident about book reviews by users. From the charts below, I believe the most helpful is the first, displaying a clear indication for lift, support and confidence. It showed that the lift is important because as the lift rises, so does the need for the support to be lower and the confidence to be higher as well.

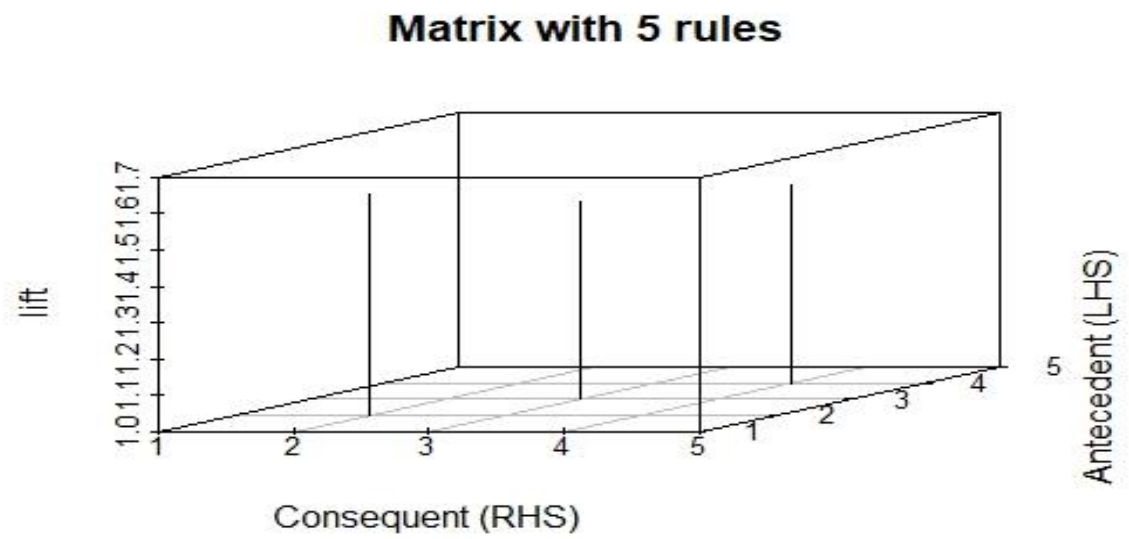



The second chart is also interesting as it shows that there is at least some overlap in to the top rules.



Also, only 2 of the rules for the visualization of the paracoord were don't to get the idea of the general direction of the rules. Also, a 3D matrix was done in order to see in a spatial context the data.






In conclusion, with more analysis as well as data points and the whole dataset one could figure out who exactly did what review for what book. Next analysis will be extended to the full dataset for this reason. There was a lot that we could figure out between the two methods (visualization and analysis). We also see that there is a lot of processing that goes into the understanding of this data after transformation. A rule that surprised me was that for book rating number 5 and how large of an impact it had on the data (graph 2). This was the most overwhelming rule out of the rules analyzed. In the future, the recommendations are to increase the lift, decrease the support, and decrease the confidence. This will help to get more rules and help us to visualize the reviews and ratings more accurately.

References:

1. R guides. Package arules. Available at: https://cran.r-project.org/web/packages/arules/arules.pdf.

2. R-Documentation. Transactions-class. Available at:

https://www.rdocumentation.org/packages/arules/versions/1.6-3/topics/transactions-class