Dataset

There are thousands of anime out there, all wildly varying in style, genre, and quality. With such a variety out there, it can sometimes give choice paralysis when it comes to choosing what to watch next. In this case, a recommendation system would be very useful. By analyzing the shows you liked, it should be possible to figure out some recommendations for things you'd probably like.

The dataset I'll be using for this assignment is the Anime Recommendations Database. This dataset actually contains two specific datasets, one carrying a list containing information on specific anime, such as name, genre, and # of members. The other dataset is a record containing the rating users of a rating website gave to these anime, ranging from 1-10 (-1 if tracked, but not rated). We'll mainly be using the second dataset, and referencing the first for obtaining specific information (such as names or genres).

Since there's a vast amount of shows in the dataset, I've opted to prune some of the anime in order to reduce the runtime of the script. I removed anything other than TV shows, as many specials and movies tend to be sequels or spin offs of their respective shows.

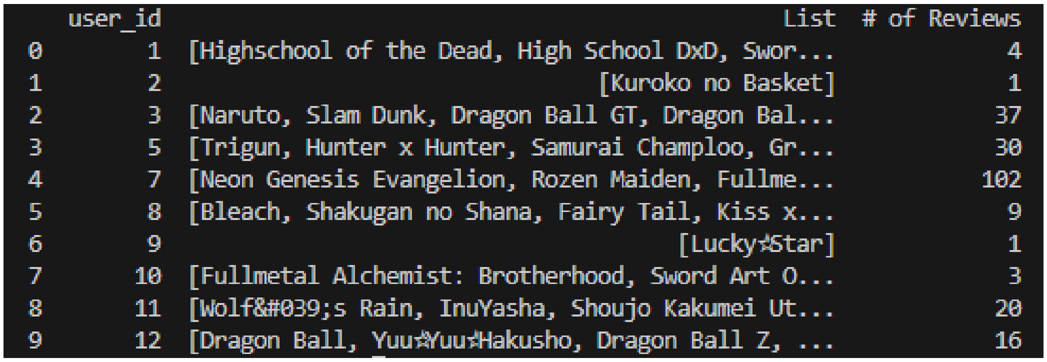
Association Analysis

One idea for creating a recommendation system is to see which shows other users like that are similar to what you may like. After all, if users A and B like Show 1, and user B likes Show 2, then user A may enjoy Show 2 as well if it's similar enough to Show 1. To determine this, we can use association analysis, specifically with the apriori algorithm.

Methodology

To perform association analysis using the apriori algorithm, we have to create itemsets to analyze. To do this, we can group the individual ratings by each user and collate them into a list of reviews per user. After that, we have to prune each list of the more negative reviews, so that each user only has a list of shows they like. As the average rating is ~7.8, we'll only take reviews scored 8 or higher (not only for simplicity, but also to ease overall calculations).

After that preprocessing, we get a list of shows per user that they enjoyed.



```
   user_id                                               List  # of Reviews
0        1  [Highschool of the Dead, High School DxD, Swor...             4
1        2                                  [Kuroko no Basket]             1
2        3  [Naruto, Slam Dunk, Dragon Ball GT, Dragon Bal...            37
3        5  [Trigun, Hunter x Hunter, Samurai Champloo, Gr...            30
4        7  [Neon Genesis Evangelion, Rozen Maiden, Fullme...           102
5        8  [Bleach, Shakugan no Shana, Fairy Tail, Kiss x...             9
6        9                                        [Lucky☆Star]             1
7       10  [Fullmetal Alchemist: Brotherhood, Sword Art O...             3
8       11  [Wolf&#039;s Rain, InuYasha, Shoujo Kakumei Ut...            20
9       12  [Dragon Ball, Yuu☆Yuu☆Hakusho, Dragon Ball Z, ...            16
```

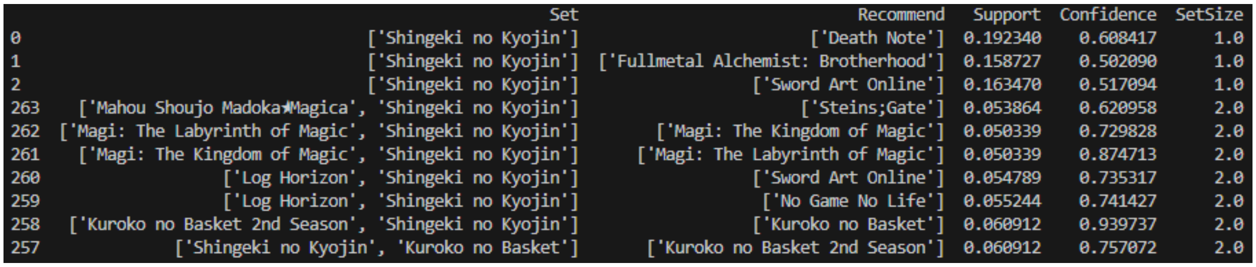
From here, we can use these lists as itemsets for the apriori algorithm.

We take the list of shows and run them through the apriori algorithm, generating a list of rules. For confidence, I set the minimum to 0.5. Due to having several thousands of items and itemsets, I opted to set the minimum support to a very low 0.05, so that we can generate some rules that don't include some of the more popular shows while also keeping runtime relatively low. I have also generated a ruleset with minimum support set to 0.03, and saved it to rules03.csv if you want to check it out.

With our list of rules, we can now analyze them to see some patterns.

Results

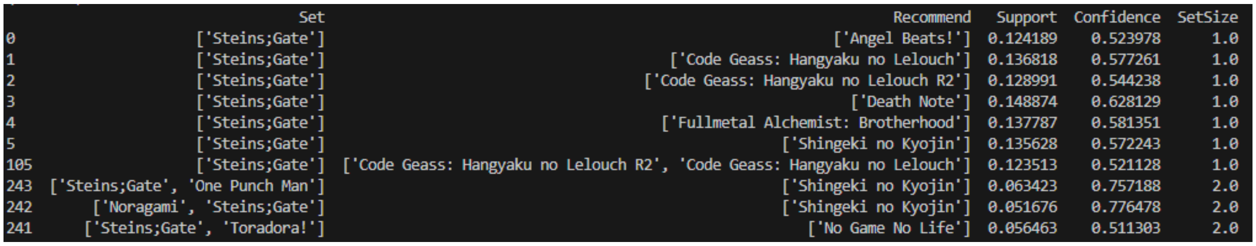
Our ruleset comes out to about 5583 different rules (96859 rules with 0.03 min support). Using this ruleset, we can query for specific anime and see what kind of recommendations it will give us. For example, let's say you enjoyed Attack on Titan. Querying for it as a selection will return the following rules (the first 10):



```
                                                 Set                            Recommend   Support  Confidence  SetSize
0                                ['Shingeki no Kyojin']                      ['Death Note']  0.192340    0.608417      1.0
1                                ['Shingeki no Kyojin']  ['Fullmetal Alchemist: Brotherhood']  0.158727    0.502090      1.0
2                                ['Shingeki no Kyojin']               ['Sword Art Online']  0.163470    0.517094      1.0
263    ['Mahou Shoujo Madoka★Magica', 'Shingeki no Kyojin']                    ['Steins;Gate']  0.053864    0.620958      2.0
262 ['Magi: The Labyrinth of Magic', 'Shingeki no Kyojin']     ['Magi: The Kingdom of Magic']  0.050339    0.729828      2.0
261   ['Magi: The Kingdom of Magic', 'Shingeki no Kyojin']   ['Magi: The Labyrinth of Magic']  0.050339    0.874713      2.0
260                 ['Log Horizon', 'Shingeki no Kyojin']               ['Sword Art Online']  0.054789    0.735317      2.0
259                 ['Log Horizon', 'Shingeki no Kyojin']                ['No Game No Life']  0.055244    0.741427      2.0
258  ['Kuroko no Basket 2nd Season', 'Shingeki no Kyojin']               ['Kuroko no Basket']  0.060912    0.939737      2.0
257            ['Shingeki no Kyojin', 'Kuroko no Basket']    ['Kuroko no Basket 2nd Season']  0.060912    0.757072      2.0
```

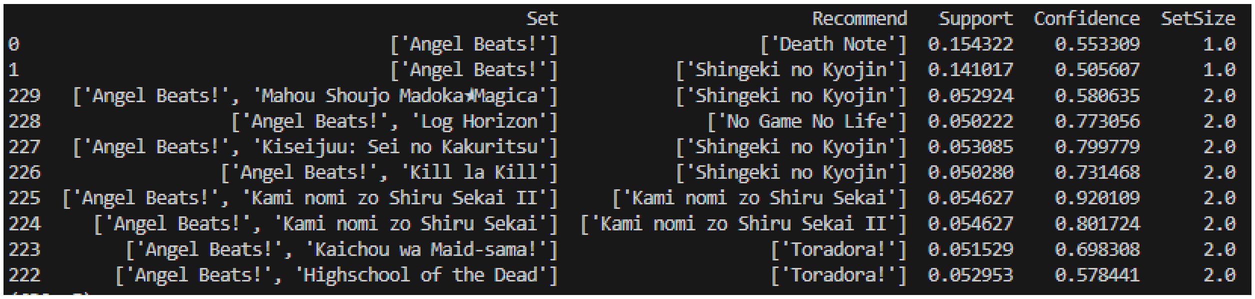
With this, we can see that if you liked Attack on Titan, you might enjoy Death Note, FMA:B, and Sword Art Online as well (at 0.6, 0.5, and 0.51 confidence values). If you watched other shows along with AoT, it will give rules for that as well (If you liked Log Horizon as well as AoT, then you'll likely enjoy SAO more, now at 0.73 confidence).

Querying again for Steins;Gate:



```
                                Set                                                          Recommend   Support  Confidence  SetSize
0                    ['Steins;Gate']                                                   ['Angel Beats!']  0.124189    0.523978      1.0
1                    ['Steins;Gate']                                  ['Code Geass: Hangyaku no Lelouch']  0.136818    0.577261      1.0
2                    ['Steins;Gate']                               ['Code Geass: Hangyaku no Lelouch R2']  0.128991    0.544238      1.0
3                    ['Steins;Gate']                                                    ['Death Note']  0.148874    0.628129      1.0
4                    ['Steins;Gate']                                ['Fullmetal Alchemist: Brotherhood']  0.137787    0.581351      1.0
5                    ['Steins;Gate']                                          ['Shingeki no Kyojin']  0.135628    0.572243      1.0
105                  ['Steins;Gate']  ['Code Geass: Hangyaku no Lelouch R2', 'Code Geass: Hangyaku no Lelouch']  0.123513    0.521128      1.0
243 ['Steins;Gate', 'One Punch Man']                                          ['Shingeki no Kyojin']  0.063423    0.757188      2.0
242      ['Noragami', 'Steins;Gate']                                          ['Shingeki no Kyojin']  0.051676    0.776478      2.0
241     ['Steins;Gate', 'Toradora!']                                             ['No Game No Life']  0.056463    0.511303      2.0
```

And then one more example, Angel Beats:

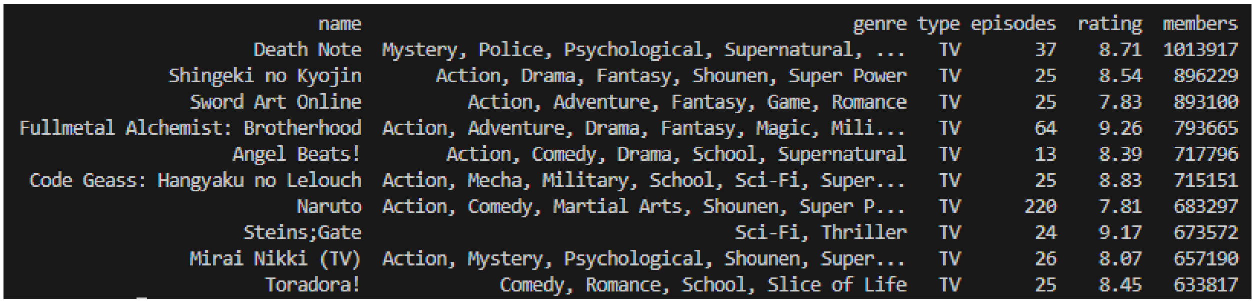


```
                                               Set                        Recommend   Support  Confidence  SetSize
0                                  ['Angel Beats!']                   ['Death Note']  0.154322    0.553309      1.0
1                                  ['Angel Beats!']           ['Shingeki no Kyojin']  0.141017    0.505607      1.0
229  ['Angel Beats!', 'Mahou Shoujo Madoka★Magica']           ['Shingeki no Kyojin']  0.052924    0.580635      2.0
228                  ['Angel Beats!', 'Log Horizon']              ['No Game No Life']  0.050222    0.773056      2.0
227  ['Angel Beats!', 'Kiseijuu: Sei no Kakuritsu']           ['Shingeki no Kyojin']  0.053085    0.799779      2.0
226                 ['Angel Beats!', 'Kill la Kill']           ['Shingeki no Kyojin']  0.050280    0.731468      2.0
225 ['Angel Beats!', 'Kami nomi zo Shiru Sekai II']     ['Kami nomi zo Shiru Sekai']  0.054627    0.920109      2.0
224    ['Angel Beats!', 'Kami nomi zo Shiru Sekai']  ['Kami nomi zo Shiru Sekai II']  0.054627    0.801724      2.0
223       ['Angel Beats!', 'Kaichou wa Maid-sama!']                    ['Toradora!']  0.051529    0.698308      2.0
222      ['Angel Beats!', 'Highschool of the Dead']                    ['Toradora!']  0.052953    0.578441      2.0
```

As you might have noticed, some patterns are starting to emerge.

The first notable thing to note here is that certain shows keep popping up in the recommend column. We queried three shows and got 30 results, but only 17 distinct ones.

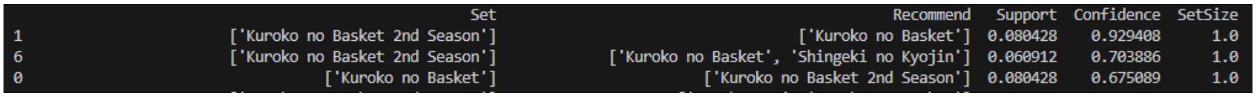
The reason for this is that our itemsets aren't exactly equal in representation. Sorting by anime popularity:



```
                             name                                              genre type  episodes  rating  members
                       Death Note  Mystery, Police, Psychological, Supernatural, ...   TV        37    8.71  1013917
               Shingeki no Kyojin       Action, Drama, Fantasy, Shounen, Super Power   TV        25    8.54   896229
                 Sword Art Online          Action, Adventure, Fantasy, Game, Romance   TV        25    7.83   893100
 Fullmetal Alchemist: Brotherhood  Action, Adventure, Drama, Fantasy, Magic, Mili...   TV        64    9.26   793665
                     Angel Beats!       Action, Comedy, Drama, School, Supernatural   TV        13    8.39   717796
  Code Geass: Hangyaku no Lelouch  Action, Mecha, Military, School, Sci-Fi, Super...   TV        25    8.83   715151
                           Naruto  Action, Comedy, Martial Arts, Shounen, Super P...   TV       220    7.81   683297
                      Steins;Gate                                   Sci-Fi, Thriller   TV        24    9.17   673572
                 Mirai Nikki (TV)  Action, Mystery, Psychological, Shounen, Super...   TV        26    8.07   657190
                        Toradora!               Comedy, Romance, School, Slice of Life   TV        25    8.45   633817
```

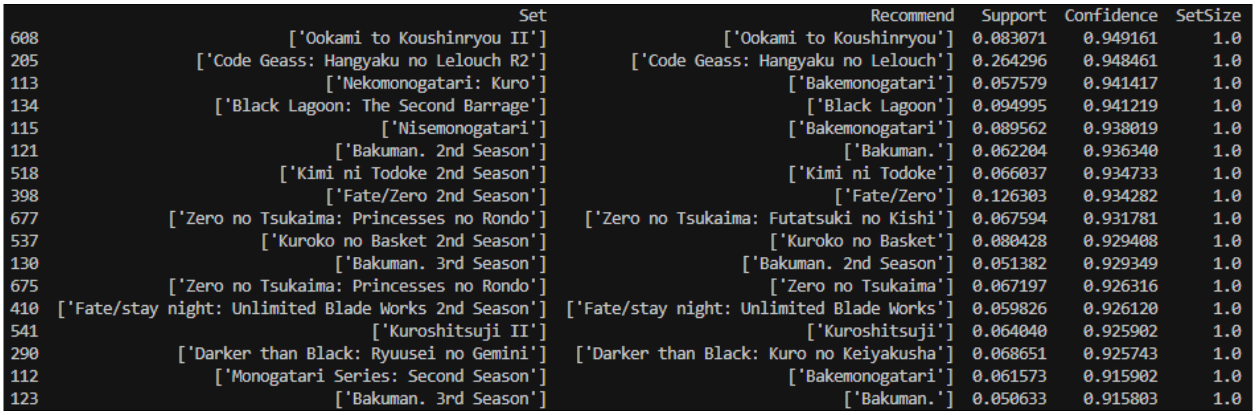
We can see that some shows may have a lot more members than others, and thus more reviews. This will not only create more rules involving popular anime due to the increased support, but it will decrease the support of less popular anime simply by increasing the total size of itemsets. We can see this if we query for rules involving these shows. Of the top 3 most popular shows, Death Note is part of 946 of 5583 itemsets, Attack on Titan involved in 807, and Sword Art Online involved in around 403.

The second thing to note is the confidence score of certain rulesets, specifically between a show and its sequel. For example, if we query Kuroko no Basket:



```
                                    Set                                     Recommend   Support  Confidence  SetSize
1       ['Kuroko no Basket 2nd Season']                          ['Kuroko no Basket']  0.080428    0.929408      1.0
6       ['Kuroko no Basket 2nd Season']    ['Kuroko no Basket', 'Shingeki no Kyojin']  0.060912    0.703886      1.0
0                  ['Kuroko no Basket']               ['Kuroko no Basket 2nd Season']  0.080428    0.675089      1.0
```

We can see that the confidence score between the 2nd Season to the 1st is much higher than the confidence score of the 1st season to the 2nd. This isn't an isolated pattern, as if we query for the highest confidence:



```
                                                      Set                                  Recommend   Support  Confidence  SetSize
608                          ['Ookami to Koushinryou II']                  ['Ookami to Koushinryou']  0.083071    0.949161      1.0
205                ['Code Geass: Hangyaku no Lelouch R2']        ['Code Geass: Hangyaku no Lelouch']  0.264296    0.948461      1.0
113                              ['Nekomonogatari: Kuro']                         ['Bakemonogatari']  0.057579    0.941417      1.0
134                  ['Black Lagoon: The Second Barrage']                           ['Black Lagoon']  0.094995    0.941219      1.0
115                                    ['Nisemonogatari']                         ['Bakemonogatari']  0.089562    0.938019      1.0
121                               ['Bakuman. 2nd Season']                               ['Bakuman.']  0.062204    0.936340      1.0
518                         ['Kimi ni Todoke 2nd Season']                         ['Kimi ni Todoke']  0.066037    0.934733      1.0
398                              ['Fate/Zero 2nd Season']                              ['Fate/Zero']  0.126303    0.934282      1.0
677             ['Zero no Tsukaima: Princesses no Rondo']   ['Zero no Tsukaima: Futatsuki no Kishi']  0.067594    0.931781      1.0
537                       ['Kuroko no Basket 2nd Season']                       ['Kuroko no Basket']  0.080428    0.929408      1.0
130                               ['Bakuman. 3rd Season']                    ['Bakuman. 2nd Season']  0.051382    0.929349      1.0
675             ['Zero no Tsukaima: Princesses no Rondo']                       ['Zero no Tsukaima']  0.067197    0.926316      1.0
410 ['Fate/stay night: Unlimited Blade Works 2nd Season'] ['Fate/stay night: Unlimited Blade Works']  0.059826    0.926120      1.0
541                                   ['Kuroshitsuji II']                           ['Kuroshitsuji']  0.064040    0.925902      1.0
290              ['Darker than Black: Ryuusei no Gemini']  ['Darker than Black: Kuro no Keiyakusha']  0.068651    0.925743      1.0
112                  ['Monogatari Series: Second Season']                         ['Bakemonogatari']  0.061573    0.915902      1.0
123                               ['Bakuman. 3rd Season']                               ['Bakuman.']  0.050633    0.915803      1.0
```

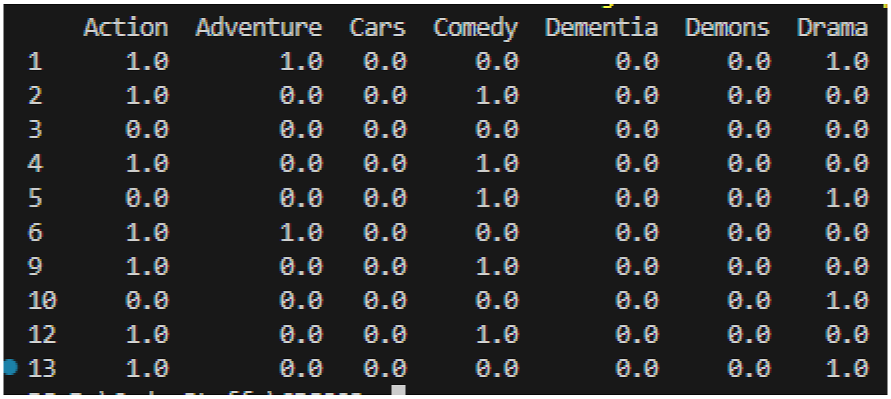
Every high confidence rule is simply a sequel -> prequel rule. This may be because a sequel would require watching the prequel, and as such the number of members of a sequel will be less than the prequel. If someone didn't enjoy the prequel, then their list may not include the sequel, thus reducing the confidence score. Vice versa, someone who enjoyed the sequel would have likely enjoyed the prequel, thus increasing the confidence score for that rule.

Clustering Analysis

Okay, association analysis has some problems. Perhaps a different solution. Well, users tend to have very similar sets of anime based on genre. Action shows like Naruto, Bleach, and One Piece are all shounen action anime, and all are liked by a similar group of users. On the other hand, someone who enjoys a romance show like Fruits Basket might not fall in that group. We could try grouping users based on their preferred genres, and use that to help recommend shows of specific genres to users.

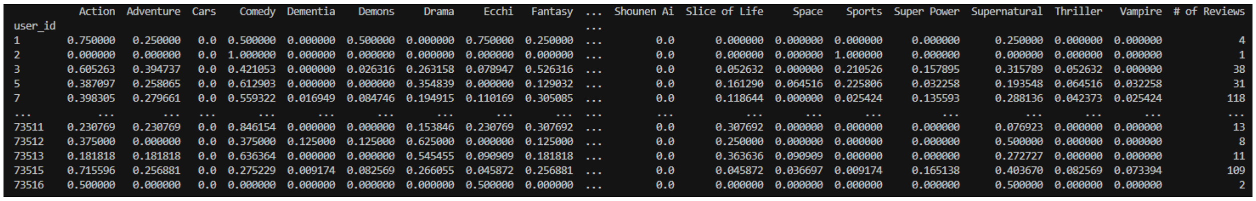
Methodology

Before we start that, we're going to transform the genre column from the anime dataset into dummy variables, so we can accurately score an anime based on its genre:



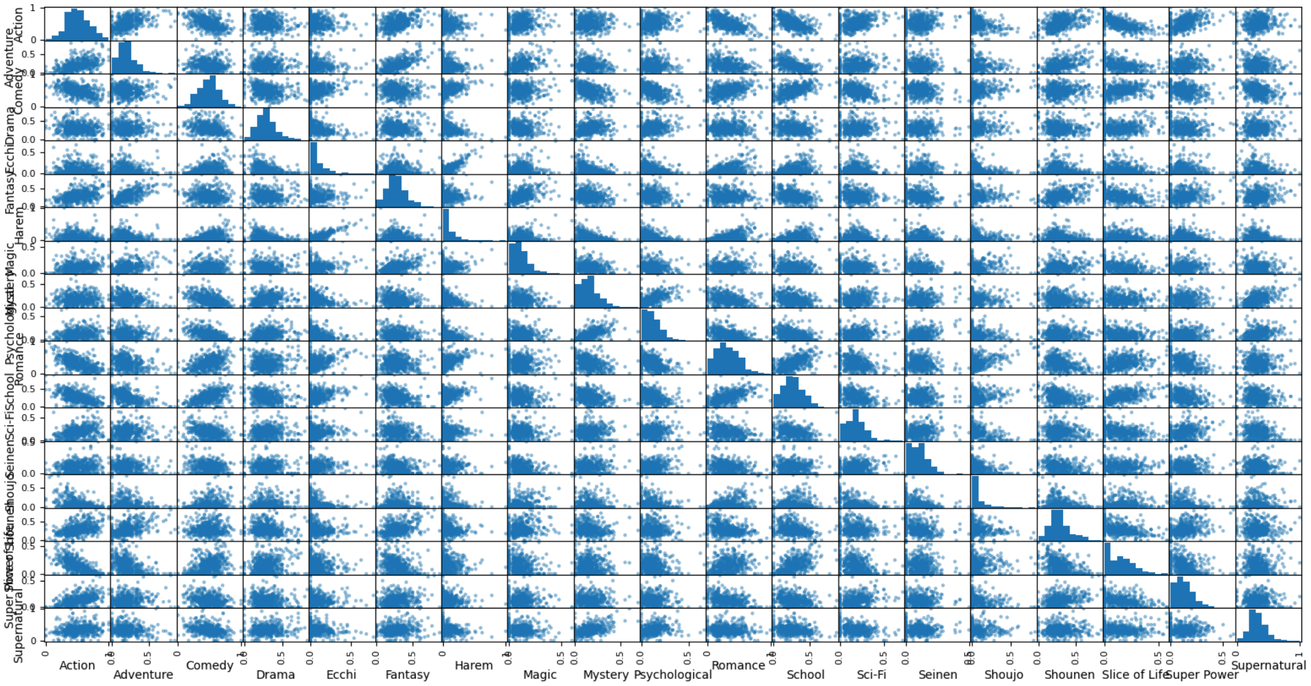
```
    Action  Adventure  Cars  Comedy  Dementia  Demons  Drama
1      1.0        1.0   0.0     0.0       0.0     0.0    1.0
2      1.0        0.0   0.0     1.0       0.0     0.0    0.0
3      0.0        0.0   0.0     0.0       0.0     0.0    0.0
4      1.0        0.0   0.0     1.0       0.0     0.0    0.0
5      0.0        0.0   0.0     1.0       0.0     0.0    1.0
6      1.0        1.0   0.0     0.0       0.0     0.0    0.0
9      1.0        0.0   0.0     1.0       0.0     0.0    0.0
10     0.0        0.0   0.0     0.0       0.0     0.0    1.0
12     1.0        0.0   0.0     1.0       0.0     0.0    0.0
13     1.0        0.0   0.0     0.0       0.0     0.0    1.0
```

After that, we're going to do something similar as we did in association analysis and group all reviews by user. Instead of creating a list of anime, we're going to sum up the genre count of each user's reviews and then divide each by the total number of reviews, so we can get a general picture of what genres of anime a user watches adjusted for their review counts.

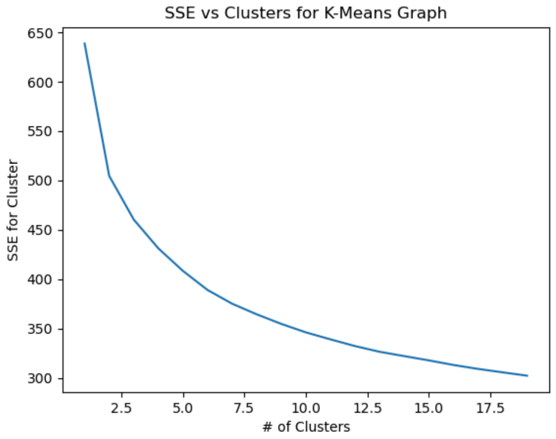


```
           Action  Adventure  Cars    Comedy  Dementia    Demons     Drama     Ecchi   Fantasy  ...  Shounen Ai  Slice of Life     Space    Sports  Super Power  Supernatural  Thriller   Vampire  # of Reviews
user_id                                                                                         ...
1        0.750000   0.250000   0.0  0.500000  0.000000  0.500000  0.000000  0.750000  0.250000  ...         0.0       0.000000  0.000000  0.000000     0.000000      0.250000  0.000000  0.000000             4
2        0.000000   0.000000   0.0  1.000000  0.000000  0.000000  0.000000  0.000000  0.000000  ...         0.0       0.000000  0.000000  1.000000     0.000000      0.000000  0.000000  0.000000             1
3        0.605263   0.394737   0.0  0.421053  0.000000  0.026316  0.263158  0.078947  0.526316  ...         0.0       0.052632  0.000000  0.210526     0.157895      0.315789  0.052632  0.000000            38
5        0.387097   0.258065   0.0  0.612903  0.000000  0.000000  0.354839  0.000000  0.129032  ...         0.0       0.161290  0.064516  0.225806     0.032258      0.193548  0.064516  0.032258            31
7        0.398305   0.279661   0.0  0.559322  0.016949  0.084746  0.194915  0.110169  0.305085  ...         0.0       0.118644  0.000000  0.025424     0.135593      0.288136  0.042373  0.025424           118
...           ...        ...   ...       ...       ...       ...       ...       ...       ...  ...         ...            ...       ...       ...          ...           ...       ...       ...           ...
73511    0.230769   0.230769   0.0  0.846154  0.000000  0.000000  0.153846  0.230769  0.307692  ...         0.0       0.307692  0.000000  0.000000     0.000000      0.076923  0.000000  0.000000            13
73512    0.375000   0.000000   0.0  0.375000  0.125000  0.125000  0.625000  0.000000  0.125000  ...         0.0       0.250000  0.000000  0.000000     0.000000      0.500000  0.000000  0.000000             8
73513    0.181818   0.181818   0.0  0.636364  0.000000  0.000000  0.545455  0.090909  0.181818  ...         0.0       0.363636  0.090909  0.000000     0.000000      0.272727  0.000000  0.000000            11
73515    0.715596   0.256881   0.0  0.275229  0.009174  0.082569  0.266055  0.045872  0.256881  ...         0.0       0.045872  0.036697  0.009174     0.165138      0.403670  0.082569  0.073394           109
73516    0.500000   0.000000   0.0  0.000000  0.000000  0.000000  0.000000  0.500000  0.000000  ...         0.0       0.000000  0.000000  0.000000     0.000000      0.500000  0.000000  0.000000             2
```

From here, we can start pruning some features. We have 39 different genres, so it'd be best to remove some of the less popular ones. After pruning those, we create a scatter matrix to check the correlation of each set of values (made with the first 500 users to reduce runtime).



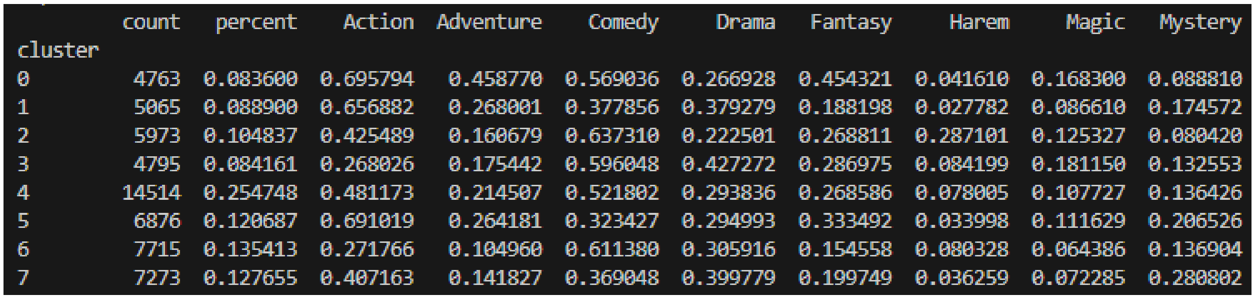
Since Ecchi and Harem have a similar enough distribution, I'll drop the former, leaving us with 18 different features to test. Still a lot, but at least it's not the 39 we started with.



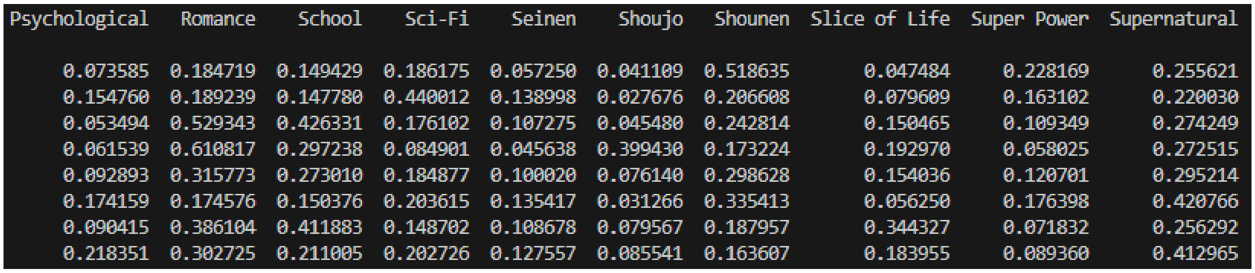
I had decided on using K-Means clustering, as I believed the spread of certain features wouldn't be good with DBSCAN. I determined the optimal number of clusters by testing each one, up to 20 clusters. Graphing the SSE for each cluster resulted in the graph to the left. Here, I determined that the optimal number of clusters would be 8.

Results

After fitting the data and assigning each review to a cluster, we can start analyzing these clusters. By grouping them together, we can see the percentage of the total number of users for each cluster, as well as the average percent of a genre they watch.



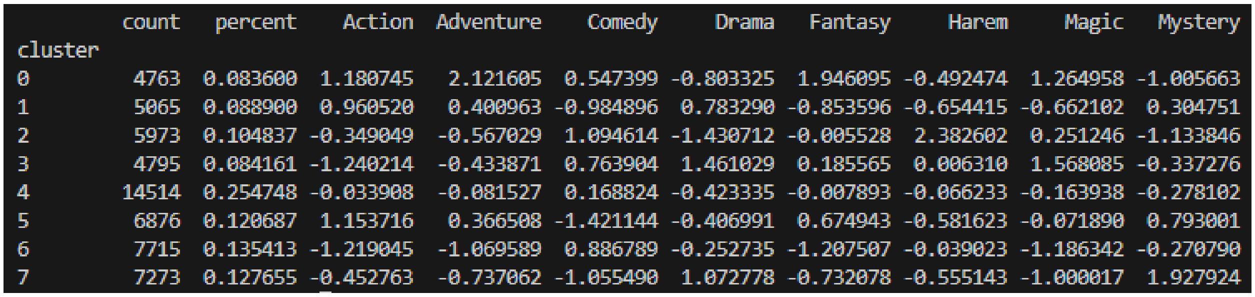
```
         count   percent    Action  Adventure    Comedy     Drama   Fantasy     Harem     Magic   Mystery
cluster
0         4763  0.083600  0.695794   0.458770  0.569036  0.266928  0.454321  0.041610  0.168300  0.088810
1         5065  0.088900  0.656882   0.268001  0.377856  0.379279  0.188198  0.027782  0.086610  0.174572
2         5973  0.104837  0.425489   0.160679  0.637310  0.222501  0.268811  0.287101  0.125327  0.080420
3         4795  0.084161  0.268026   0.175442  0.596048  0.427272  0.286975  0.084199  0.181150  0.132553
4        14514  0.254748  0.481173   0.214507  0.521802  0.293836  0.268586  0.078005  0.107727  0.136426
5         6876  0.120687  0.691019   0.264181  0.323427  0.294993  0.333492  0.033998  0.111629  0.206526
6         7715  0.135413  0.271766   0.104960  0.611380  0.305916  0.154558  0.080328  0.064386  0.136904
7         7273  0.127655  0.407163   0.141827  0.369048  0.399779  0.199749  0.036259  0.072285  0.280802
```



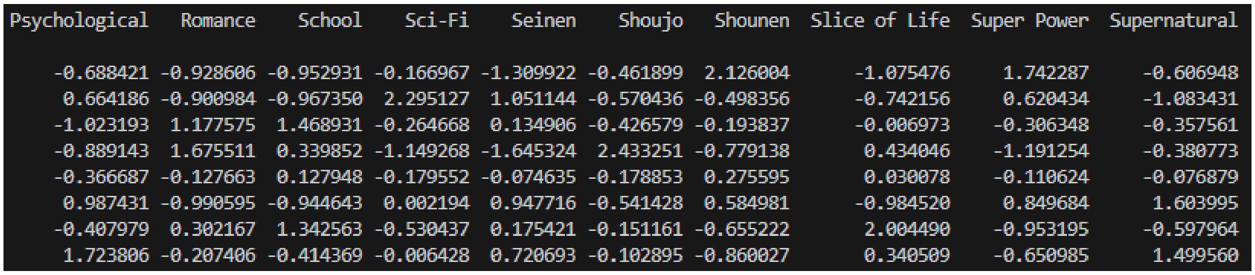
```
Psychological   Romance    School    Sci-Fi    Seinen    Shoujo   Shounen  Slice of Life  Super Power  Supernatural

     0.073585  0.184719  0.149429  0.186175  0.057250  0.041109  0.518635       0.047484     0.228169      0.255621
     0.154760  0.189239  0.147780  0.440012  0.138998  0.027676  0.206608       0.079609     0.163102      0.220030
     0.053494  0.529343  0.426331  0.176102  0.107275  0.045480  0.242814       0.150465     0.109349      0.274249
     0.061539  0.610817  0.297238  0.084901  0.045638  0.399430  0.173224       0.192970     0.058025      0.272515
     0.092893  0.315773  0.273010  0.184877  0.100020  0.076140  0.298628       0.154036     0.120701      0.295214
     0.174159  0.174576  0.150376  0.203615  0.135417  0.031266  0.335413       0.056250     0.176398      0.420766
     0.090415  0.386104  0.411883  0.148702  0.108678  0.079567  0.187957       0.344327     0.071832      0.256292
     0.218351  0.302725  0.211005  0.202726  0.127557  0.085541  0.163607       0.183955     0.089360      0.412965
```

The clusters are roughly even in size, save for cluster 4 containing a quarter of all users. We can also convert these numbers to a Z-score in order to get a better exact picture on how these clusters differ from one another.



```
         count   percent    Action  Adventure    Comedy     Drama   Fantasy     Harem     Magic   Mystery
cluster
0         4763  0.083600  1.180745   2.121605  0.547399 -0.803325  1.946095 -0.492474  1.264958 -1.005663
1         5065  0.088900  0.960520   0.400963 -0.984896  0.783290 -0.853596 -0.654415 -0.662102  0.304751
2         5973  0.104837 -0.349049  -0.567029  1.094614 -1.430712 -0.005528  2.382602  0.251246 -1.133846
3         4795  0.084161 -1.240214  -0.433871  0.763904  1.461029  0.185565  0.006310  1.568085 -0.337276
4        14514  0.254748 -0.033908  -0.081527  0.168824 -0.423335 -0.007893 -0.066233 -0.163938 -0.278102
5         6876  0.120687  1.153716   0.366508 -1.421144 -0.406991  0.674943 -0.581623 -0.071890  0.793001
6         7715  0.135413 -1.219045  -1.069589  0.886789 -0.252735 -1.207507 -0.039023 -1.186342 -0.270790
7         7273  0.127655 -0.452763  -0.737062 -1.055490  1.072778 -0.732078 -0.555143 -1.000017  1.927924
```



```
Psychological   Romance    School    Sci-Fi    Seinen    Shoujo   Shounen  Slice of Life  Super Power  Supernatural

    -0.688421 -0.928606 -0.952931 -0.166967 -1.309922 -0.461899  2.126004      -1.075476     1.742287     -0.606948
     0.664186 -0.900984 -0.967350  2.295127  1.051144 -0.570436 -0.498356      -0.742156     0.620434     -1.083431
    -1.023193  1.177575  1.468931 -0.264668  0.134906 -0.426579 -0.193837      -0.006973    -0.306348     -0.357561
    -0.889143  1.675511  0.339852 -1.149268 -1.645324  2.433251 -0.779138       0.434046    -1.191254     -0.380773
    -0.366687 -0.127663  0.127948 -0.179552 -0.074635 -0.178853  0.275595       0.030078    -0.110624     -0.076879
     0.987431 -0.990595 -0.944643  0.002194  0.947716 -0.541428  0.584981      -0.984520     0.849684      1.603995
    -0.407979  0.302167  1.342563 -0.530437  0.175421 -0.151161 -0.655222       2.004490    -0.953195     -0.597964
     1.723806 -0.207406 -0.414369 -0.006428  0.720693 -0.102895 -0.860027       0.340509    -0.650985      1.499560
```

With this, we can see more clearly how each cluster differs on average from the others.

(Thing to note: Shounen, Shoujo, and Seinen are more so demographic labels than genres specifically. Shounen for boys, Shoujo for girls, and Seinen for adult men. There is also Josei for adult women, but that genre was removed before due to lack of representation. While these demographics aren’t set rules for what genres they represent, they do trend towards what you would expect them to be).

Cluster	Analysis
0	Can be considered the de facto Shounen group as the only cluster with a Z-score > 1 and reaching above 2 in this demographic. Also trends slightly negative to Shoujo and very negative to Seinen. Trends towards Action, Adventure, Fantasy, and Super Power anime. Trending somewhat away from Mystery, Drama, and Slice of Life works. Shounen Action and Adventure shows would be the best recommendations for this group, like One Piece or Naruto.
1	The de facto Sci-fi group with a high z-score (2.29). They slightly trend toward Action, Drama, and Psychological works, while trending away from what you would expect from Sci-fi lovers (Fantasy, Magic, Slice of Life, and Supernatural), but no big major opinions apart from Sci-Fi. Demographics trend to Seinen and away from the other two. Sci-Fi recommendations are obvious. Probably enjoys Gundam.
2	The de facto Harem group with a high z-score (2.38). Similarly enjoys School, Comedy, and Romance based (likely shows that contain some combination of the four). Trends away from harsher genre topics like drama, psychological and mystery. No larger opinions on any of the other genres. Demographics are roughly equal. Highschool harem rom coms are the best bet for recommendations.
3	The de facto Shoujo group with a high z-score (2.43). Trends harshly away from Shounen and Seinen demographics. Trends towards the expected categories (Romance, Drama) and a focus on Magic as well. Trends away from Action, Psychological, Sci-Fi, and Super Power shows. Shoujo shows, especially focused on Romance and Drama would be the best recommendation here.
4	The largest group, and seemingly the least polarized at the same time. The highest scores are slightly negative trends to Drama and Psychological (two categories that match well together), but otherwise have no real significant trend. Likely a group of fans that would enjoy a variety of shows regardless of genre.
5	A group with no massive trends towards a specific genre, but it does have some significant ones. Trends positively to Action, Psychological, and Supernatural, while trending away from more light-hearted genres like Comedy, Romance, School, and Slice of Life. Demographics skew towards Seinen and slightly Shounen, and no big negative trend away from Shoujo. More serious action thriller shows would be a good recommendation, such as Tokyo Ghoul or Psycho-Pass.

6	The de facto Slice of Life group with a high z-score (2.00). Besides Slice of Life, trends positively towards School and Comedy genres, and slightly to romance. Notably, it actually negatively trends somewhat badly with almost every other genre (mystery, psychological, and harem are the exceptions). Demographics skew away from Shounen, but no significant bias for the other two. Best recommendations would probably be regular Slice of Life shows: Nichijou, Yuru Camp, etc.
7	The de facto Mystery (1.92) group, with a high trend towards Psychological (1.72) as well. Has a positive bias for Drama and Supernatural, and a negative bias to everything except for Romance, Sci-Fi, and kinda School. Demographics skew positively to Seinen, negatively to Shounen, and no real opinion for Shoujo. Best recommendations would be mystery and psychological focused shows, like Monster or Odd Taxi.

Conclusion

For recommending anime, both association and clustering analysis have their benefits and drawbacks, and it's important to know them best to see what's better for a recommendation system.

Association analysis would be a great system for a newcomer, someone who hasn't watched more than 2-3 shows (the largest itemset for a rule from the apriori algorithm was 5). They likely have watched a more popular show, and as such there is more likely to be a rule given the support needed. The added bonus is that this method is focused specifically on direct users: a rule A -> B is made specifically because a lot of users enjoyed Shows A and B, so a recommendation is more likely to be accurate.

The drawbacks become apparent as someone watches more shows. Due to the vast breadth of shows and popularity skew, a large portion of the rules involve the more popular shows like Death Note, and include very few lesser known shows. We could lower our minimum support in order to accommodate these less popular shows, but that comes with its own drawback: time. Doing so massively increases the time it takes to calculate and generate these rules. A 0.05 min support took 30 seconds to generate. 0.03 min took 25 minutes. Lowering this any further increases this time exponentially. As such, if you're someone who has already watched a lot of the popular shows, this analysis won't work as well for you.

Clustering analysis, on the other hand, can work better for someone who has watched a variety of shows. As someone watches more shows, they might skew towards a specific genre(s) (Action vs Romance), and may be more willing to let go of faults in said genre(s). As such, by clustering users by these genres, we can assign a user to a cluster and cater to them their desired genres.

The downside to this method, of course, is that it’s much more generalized, and not as specific. A person has their own individual likes and dislikes, and while we can approximate that to a genre, clustering will never be specific enough to nail it down exactly. We could recommend a show that they like, but just because a user fits into a genre doesn’t mean that it would be well received (even if we only account for positive reviews). There’s also the issue of Cluster 4. This group is effectively a dead group in terms of recommendations, as there’s nothing specific that they enjoy. This group encompasses a quarter of the total number of users we fitted. Basically, clustering is useless to 1 in every 4 users we check. Increasing the number of clusters may help, but it could also negatively split an existing cluster into two similar ones as well.

Each type of analysis has its benefits and drawbacks, so if you want to make an anime recommendation system, you have to figure out which benefits you want and which drawbacks you’re willing to take in order to build the best system you can.

References

CooperUnion. Anime Recommendations Database. Retrieved April 10, 2024 from https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database/data.