Final Project

Collin McMahon

Abstract

The recommendation systems and the algorithms behind them on streaming services such as Netflix and YouTube are paramount for streamlining the content users see so they do not have to spend immense effort finding what they want. On streaming services such as Netflix, anime is a category of content that is included, as despite being Japanese in origin, anime is quite popular with American audiences. I feel that the recommendation systems are not as effective when it comes to recommending the appropriate anime for me. In this analysis I will be looking at anime data and the anime watch histories of various individuals to try and find relationships between different anime and their associations with what people watch in order to provide better recommendations. I find this especially relevant because many platforms that solely stream anime lack any sort of recommendation system at all, so it is hard to find content that matches my interests without doing a decent amount of research. To find these relationships, I used both k-means clustering with PCA and association rule mining with the a priori algorithm, to cluster anime into related groups and to find which anime go together based on the lists of what anime individuals have watched. Through the clustering, I discovered two subgroups of anime, the highly mainstream anime and longer anime series, which could potentially be used to give very rough recommendations. If someone mainly watches anime from a given cluster, then recommending them anime from that same cluster would be effective. The association rules proved to be interesting also, yielding many high confidence rules, as well as a decent amount of high lift rules. These high lift rules tended to associate one season of a series with another though, limiting the scope of their use as recommendations, but regardless, many rules were mined that associated different series of anime together. These rules frequently included the most popular anime too, which makes them helpful for recommendation if an individual has not already seen all the popular anime. Overall, this analysis is a starting point at looking at the relationships and associations between different anime, which can be used to construct better performing recommendation systems.

Introduction

The stay-at-home order has likely shifted the activities people engage in during their free-time. Personally, I started to watch a lot more Netflix as I found myself stuck inside. As someone who watches a mix of American and Japanese content on Netflix, I felt that its recommendation system was missing the mark in terms of recommending anime (Japanese animation) to me. Unlike with American content, where the recommendations appear to be more suited to my specific interests, the anime recommendations tended to simply be what is popular, regardless of the specific preferences I have shown in my history.

My goal is to gain insight into the relationships between different anime and try to find possible connections or trends in the anime people watch. This information is at the core of building a good recommendation algorithm. I find the recommendations paramount to my experience on these platforms because it helps me easily find content I would enjoy amongst the vast amount of content provided. I would like to use this analysis as a starting point in figuring out what anime are the best to recommend to an individual given their watch history, in order to improve their experience on services that stream anime such as Netflix.

Methods

The first data set used in this analysis is data about 12,294 anime from myanimelist.net. This data set includes name, genre, type (movie, TV, OVA, etc.), episode count, average rating out of 10, and the number of community members on the site that “follow” each anime. The second data set, also from myanimelist.net, is user preference data from 73,516 users. myanimelist.net allows each user to add anime they have completed to a list on their profile and give it a rating. This data set includes user ID, anime ID, and rating out of 10 (-1 if no rating was assigned). Entries with the same user ID are all the anime one specific individual has listed as watching on the site.

I used two different methods to analyze the relationships between different anime and analyze common patterns in watch histories.

The first method is clustering with k-means++ in conjunction with PCA in an attempt to group similar anime together based on the metrics in the first data set such as episode count, average rating, and number of community members. I used PCA here because it would make visualization of the clusters easier, as the clusters would be built from 3 dimensions. Additionally, relatively strong correlations exist in the data set (see Figure 1), especially between rating and community members (see Figure 5 in the Appendix), which makes sense as highly rated anime are expected to be more popular on average. Ideally, PCA would be used on a higher dimension data set, but unfortunately the data on each anime was limited to these three numeric variables. After visual inspection of the clusters, as well as analysis of their centers, I decided 3 clusters best represented this data. These clusters would help summarize this large set of different anime and help group similar anime together beyond simply their genre or type. With this, basic recommendations could be made by recommending anime from the same cluster as the majority of previous anime the user has watched.

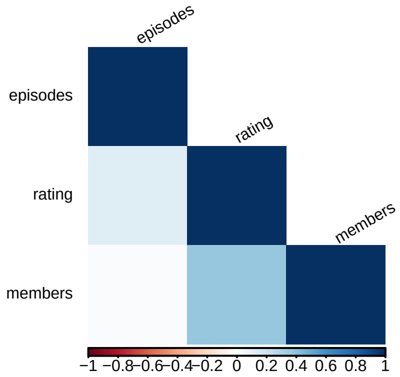


Figure 1: Correlations of Anime Metrics

The second method is association rule mining in an attempt to find pairs or groups of anime that were commonly watched together based on the user data. Since each entry of the second data set is a single rating made by a user for an anime, I had to transform the data to group all the anime rated by each user together. After this, I ran the a priori algorithm on the watch lists of every user. Because of the massive amount of potential association rules, I set the a priori algorithm to only consider rules with a support greater than 0.1 and a confidence greater than 0.25. Additionally, I set the maximum amount of anime in a single rule to be 5, in order to keep the number of rules to a manageable amount to analyze.

Results

Figure 2 shows the results of using kmeans++ in conjunction with PCA with 3 clusters. Clusters 2 and 3 specifically are quite compact, while cluster 1 is more spread out. Despite this, cluster 1 as well as cluster 3 turn out to represent two important groupings of anime. Table 1 shows the cluster centers represented as z-scores, as well as their sizes. Cluster 1 is the smallest cluster with 467 anime, while cluster 2 is the largest with 7091 anime. Cluster 1 has relatively average episode count, above average ratings, and much above average members. Cluster 2 is below average in all metrics. Cluster 3 has above average episode counts and is slightly above average in ratings. After visual inspection of the cluster scatterplots and analysis of the cluster centers, I decided 3 clusters was the most appropriate for representing this data. Figure 6 in the Appendix shows the results of k-means++ and PCA with 2 and 4 clusters, which provide a lesser visual summary of the data in comparison to 3 clusters.

Figure 3 shows the results of a priori on the user preference data set. Out of the 7302 association rules found within the given parameters, many of them appear to have high confidence. The majority of the rules have support between 0.1 and 0.2 and lift between 1 and 5. However, there are a significant amount of rules with lift greater than 5, which will be looked at during the conclusive analysis. Finally, Figure 4 is a graph of these high lift association rules. There is notable clustering in the rules that go together, and incredibly popular anime such as Death Note, Angel Beats, and Sword Art Online can be seen involved in many of these rules. Figure 7 in the Appendix shows the 20 most popular anime, based on the frequency of their appearance within the user watch histories. Included in the folder with this report are a .graphml file and an interactive HTML document that can be used to visualize this graph in better detail, and look into specific rules with more precision.

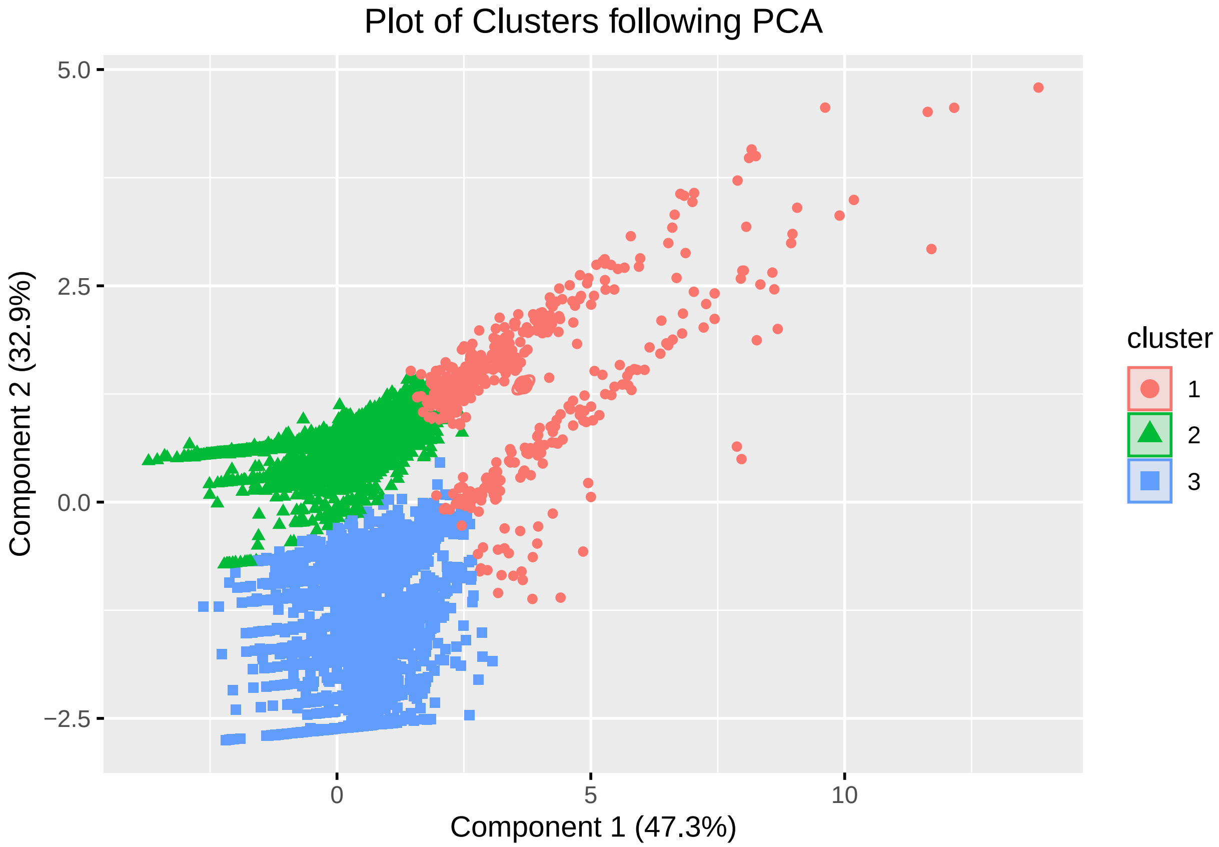


Figure 2: Clusters obtained from k-means++ and then summarized using PCA. The percentages in parentheses represent the variation of the data captured by that principle component

Table 1: Anime Cluster Centers and Sizes

	episodes	rating	members	size
Cluster 1	0.0168000	1.4611924	3.9836772	467
Cluster 2	-0.7378271	-0.1651028	-0.1510058	7091
Cluster 3	1.1593623	0.1083815	-0.1752319	4506

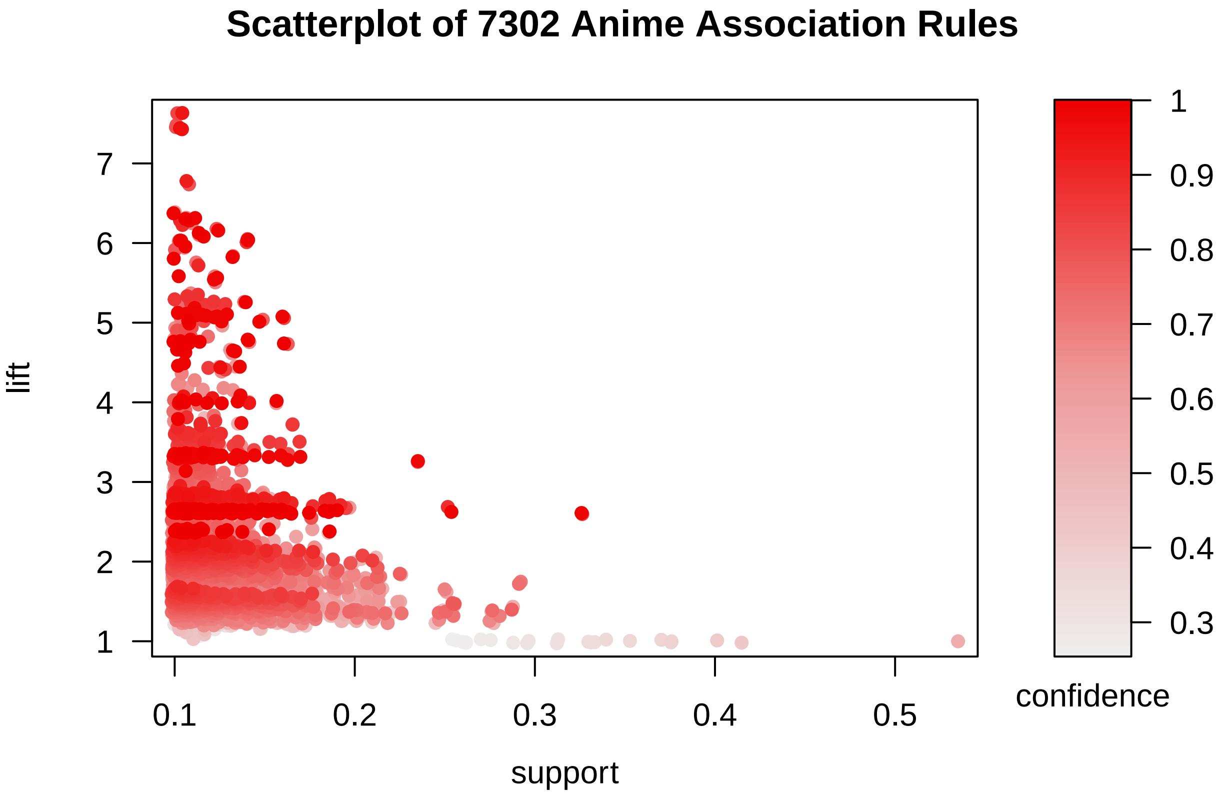
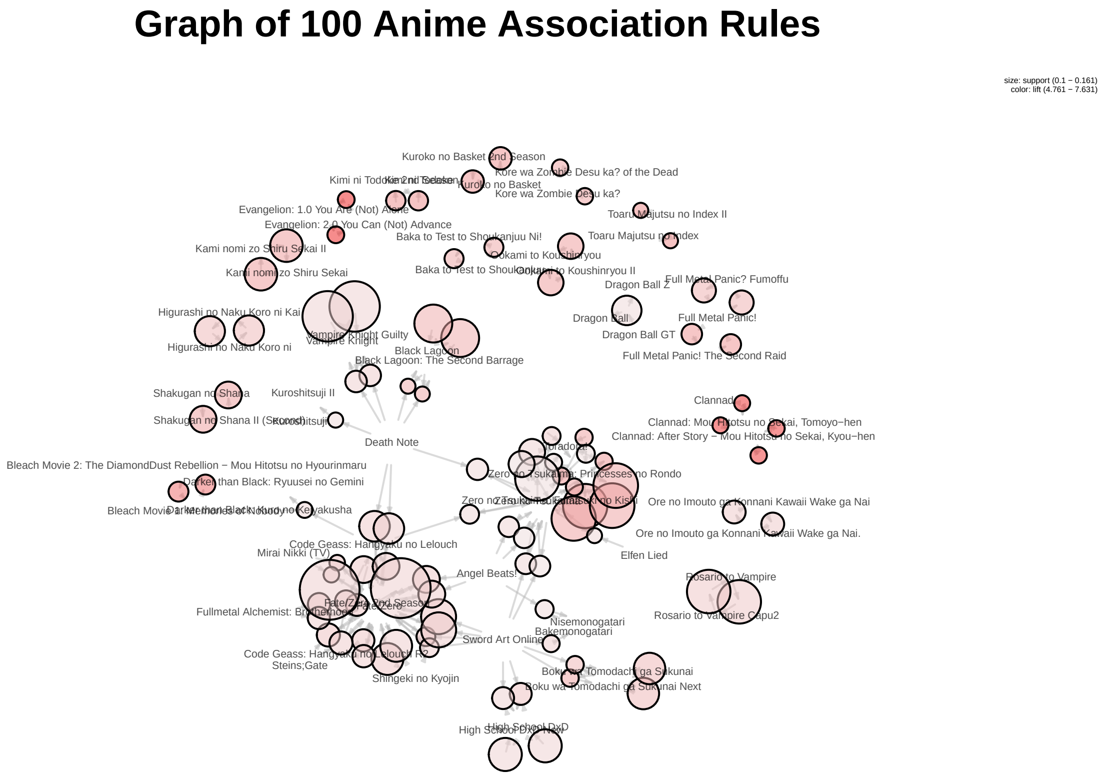


Figure 3: Plot of association rules from the a priori algorithm with min support of 0.1, min confidence of 0.25, and max length of 5

Conclusions

Both methods of analysis on these data sets provided some interesting insights into the relationships between different anime and the patterns in user watch histories with respect to the anime they watched.

Analyzing the results from Table 1, we are able to come up with a rough profile of the types of anime that would be included in each of the clusters. Cluster 1, with above average rating and much above average members, is made up of the highly popular and well-known anime. A potential reason that this cluster is so much smaller than the other two clusters is because highly rated anime are the tail of the ratings distribution, which can be seen in Figure 8. Cluster 2 is below average in all aspects and is the largest cluster, which is unfortunate because using it for recommendations would mean recommending lower rated anime. Cluster 3, with above average episode count and relatively average ratings and members, appears to be made up of the longer anime. Therefore, it is reasonable that this cluster is separated for



Graph of 100 Anime Association Rules

size: support (0.1 − 0.161)
color: lift (4.761 − 7.631)
Kuroko no Basket 2nd Season
Kore wa Zombie Desu ka? of the Dead
Kimi ni Todoke
Kimi ni Todoke 2nd Season
Kuroko no Basket
Kore wa Zombie Desu ka?
Evangelion: 1.0 You Are (Not) Alone
Toaru Majutsu no Index II
Evangelion: 2.0 You Can (Not) Advance
Baka to Test to Shoukanjuu Ni!
Toaru Majutsu no Index
Kami nomi zo Shiru Sekai II
Ookami to Koushinryou
Baka to Test to Shoukanjuu
Ookami to Koushinryou II
Kami nomi zo Shiru Sekai
Dragon Ball Z
Full Metal Panic? Fumoffu
Higurashi no Naku Koro ni Kai
Dragon Ball
Full Metal Panic!
Vampire Knight Guilty
Vampire Knight
Dragon Ball GT
Higurashi no Naku Koro ni
Black Lagoon
Full Metal Panic! The Second Raid
Black Lagoon: The Second Barrage
Shakugan no Shana
Kuroshitsuji II
Clannad
Shakugan no Shana II (Second)
Kuroshitsuji
Clannad: Mou Hitotsu no Sekai, Tomoyo−hen
Death Note
Clannad: After Story − Mou Hitotsu no Sekai, Kyou−hen
Toradora!
Bleach Movie 2: The DiamondDust Rebellion − Mou Hitotsu no Hyourinmaru
Zero no Tsukaima: Princesses no Rondo
Darker than Black: Ryuusei no Gemini
Zero no Tsukaima
Zero no Tsukaima: Futatsuki no Kishi
Ore no Imouto ga Konnani Kawaii Wake ga Nai
Bleach Movie 1: Memories of Nobody
Darker than Black: Kuro no Keiyakusha
Ore no Imouto ga Konnani Kawaii Wake ga Nai.
Code Geass: Hangyaku no Lelouch
Mirai Nikki (TV)
Elfen Lied
Angel Beats!
Rosario to Vampire
Fate/Zero 2nd Season
Fate/Zero
Fullmetal Alchemist: Brotherhood
Rosario to Vampire Capu2
Nisemonogatari
Bakemonogatari
Sword Art Online
Code Geass: Hangyaku no Lelouch R2
Steins;Gate
Shingeki no Kyojin
Boku wa Tomodachi ga Sukunai
Boku wa Tomodachi ga Sukunai Next
High School DxD
High School DxD New

Figure 4: Graph of top 100 anime association rules by lift

the other two. Though they are quite large and built on only 3 features, these clusters can be used to generate recommendations by recommending an anime from the same cluster as the majority of anime the user has watched in the past, preferably the highest rated ones. For example, an anime movie that falls in cluster 1 is Koe no Katachi, so if a user has watched a majority of anime belonging to cluster 1, such as Steins:Gate, Clannad, and Kimi no Na Wa, then Koe no Katachi would likely be an appropriate recommendation to make to them. Table 2 shows the top-rated anime in cluster 1 used for the example in this analysis. Additionally, the top-rated anime in cluster 3 are shown in Table 3, though cluster 2 is left out as it is below average in every metric and would not be ideal to use for recommendations. With more features, I feel the precision of the clusters would be increased, which would improve the ability of these clusters to be used for recommendation purposes. However, I feel that the results of the clustering from this analysis provide a good starting point for looking at the differences between anime at a large scale.

The results of the association rule mining yielded several thousand rules, many of which were high confidence. High confidence means that there is a high probability of coexistence of the consequent with the antecedent, and these high confidence rules are incredibly important in knowledge discovery. Additionally, there were quite a few rules with lifts above 6 or 7 (see Table 4), which is important because lifts greater than 1 indicate that the antecedent and consequent appear more together than expected. Both high lift and high confidence rules can be used for recommendations because these rules show that the anime they pertain to appear to go well together and appear more often together based on the user preference data. Therefore, if a user has watched the consequent anime in an association rule, recommending them the antecedent anime would be a good idea because those anime appear to go together often. One issue with this analysis was that the highest lift rules tended to be the obvious ones. These rules tended to associate different seasons of the same anime together, which is logical because if a user watched season one, it is highly likely that they watched season 2 if it exists. In the event that they did not watch season 2, then these rules could be used for making good recommendations, but otherwise, the amount of helpful rules is decreased by the rules associating the same series of anime with each other. Another observation that is clear from both the high lift association rules and the graph in Figure 4 is the dominance of the incredibly popular anime in them. This could potentially be because of their high support, but for the users who have already watched a majority of the popular anime, using these rules for recommendations could be less effective. If they have not watched all the popular anime though, these rules will likely serve to be remarkably effective.

In conclusion, this analysis helped gain insight into the relationships between different anime and found concrete connections and trends in the anime people watch that could serve as the basis for making recommendations. The heavy associations between different seasons and content of the same anime series were confirmed, but also associations linking unrelated anime were found, which are perfect for using as recommendations. Additionally, a rough clustering of anime into the popular and long groups was made, to further help provide recommendations. I believe that with further research and potentially more factors than the three looked at here, these clusters could be better refined and a better framework for making recommendations could be made. This analysis was a good start and revealed some interesting trends, but with the ever growing amount of content, recommendation algorithms need to continue improving to provide users with content tailored to their interests so they do not have to search in the endless sea of shows and movies.

Appendix

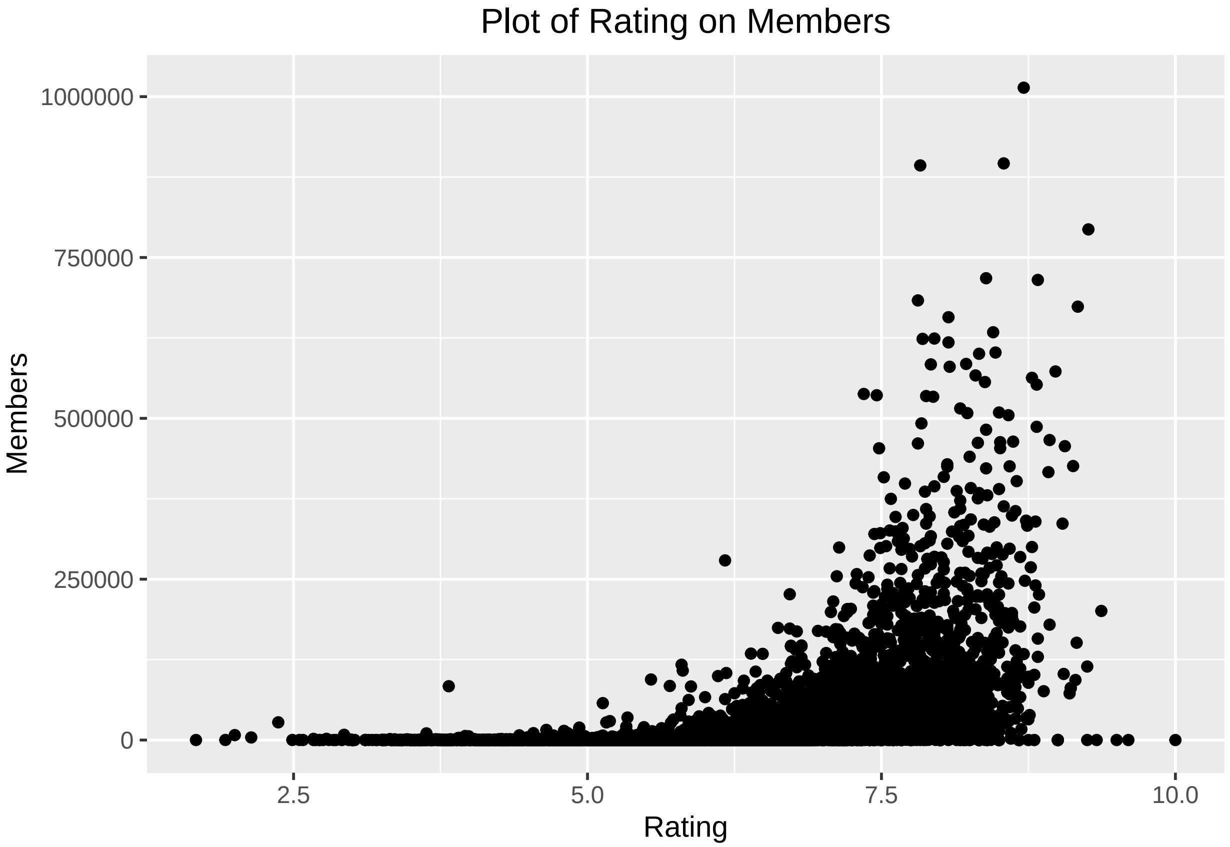


Figure 5: Scatterplot of the ratings and number of members of the anime from this data set. The relationship appears to be somewhat exponential, with the majority of anime with large followings having a rating between 7 and 9

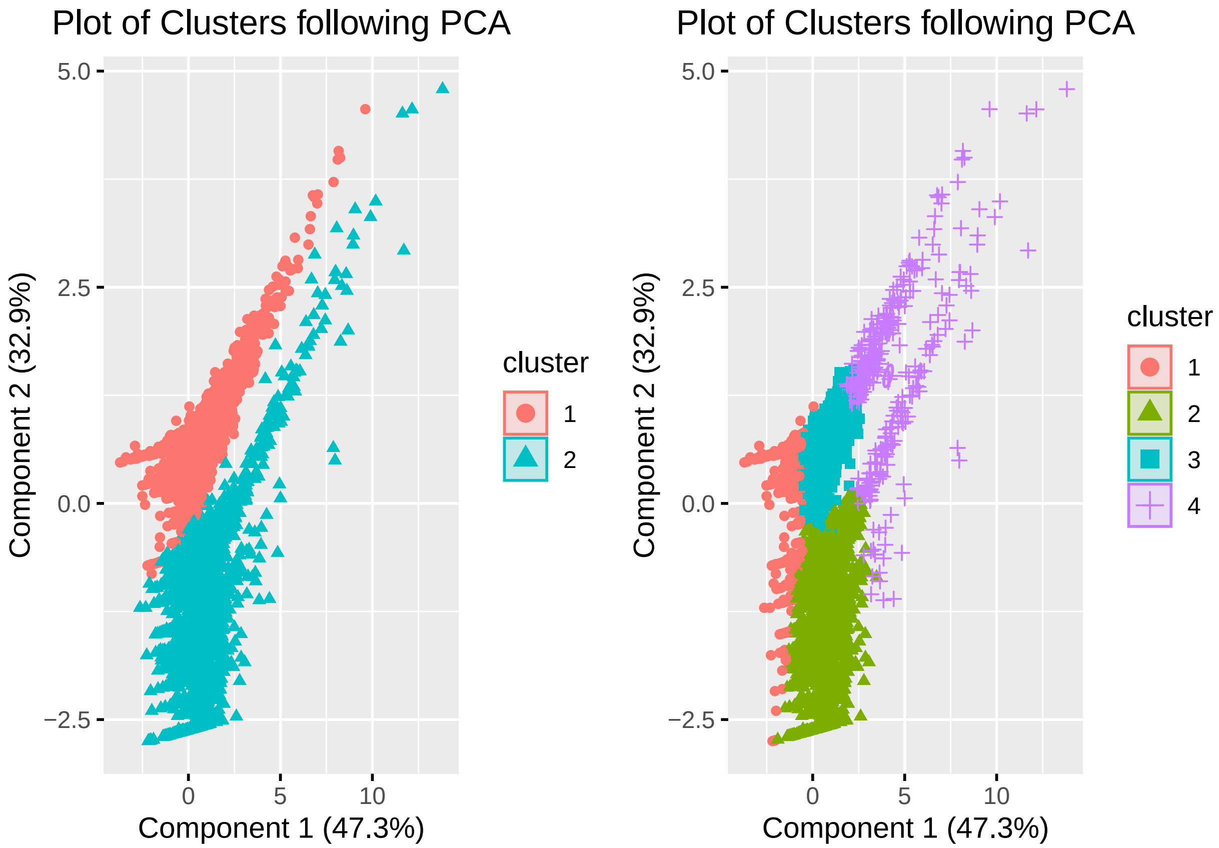


Figure 6: Clusters obtained from k-means++ and then summarized using PCA. The percentages in parentheses represent the variation of the data captured by that principle component. These clusters are less interpretable than when k = 3

Table 2: Top 10 Cluster 1 Anime by Rating

name	episodes	rating	members
Kimi no Na wa.	1	9.37	200630
Fullmetal Alchemist: Brotherhood	148	9.26	793665
GintamaÂ°	133	9.25	114262
Steins;Gate	85	9.17	673572
Gintama&#039;	133	9.16	151266
Haikyuu!!: Karasuno Koukou VS Shiratorizawa Gakuen Koukou	2	9.15	93351
Hunter x Hunter (2011)	42	9.13	425855
Clannad: After Story	85	9.06	456749
Koe no Katachi	1	9.05	102733
Gintama	76	9.04	336376

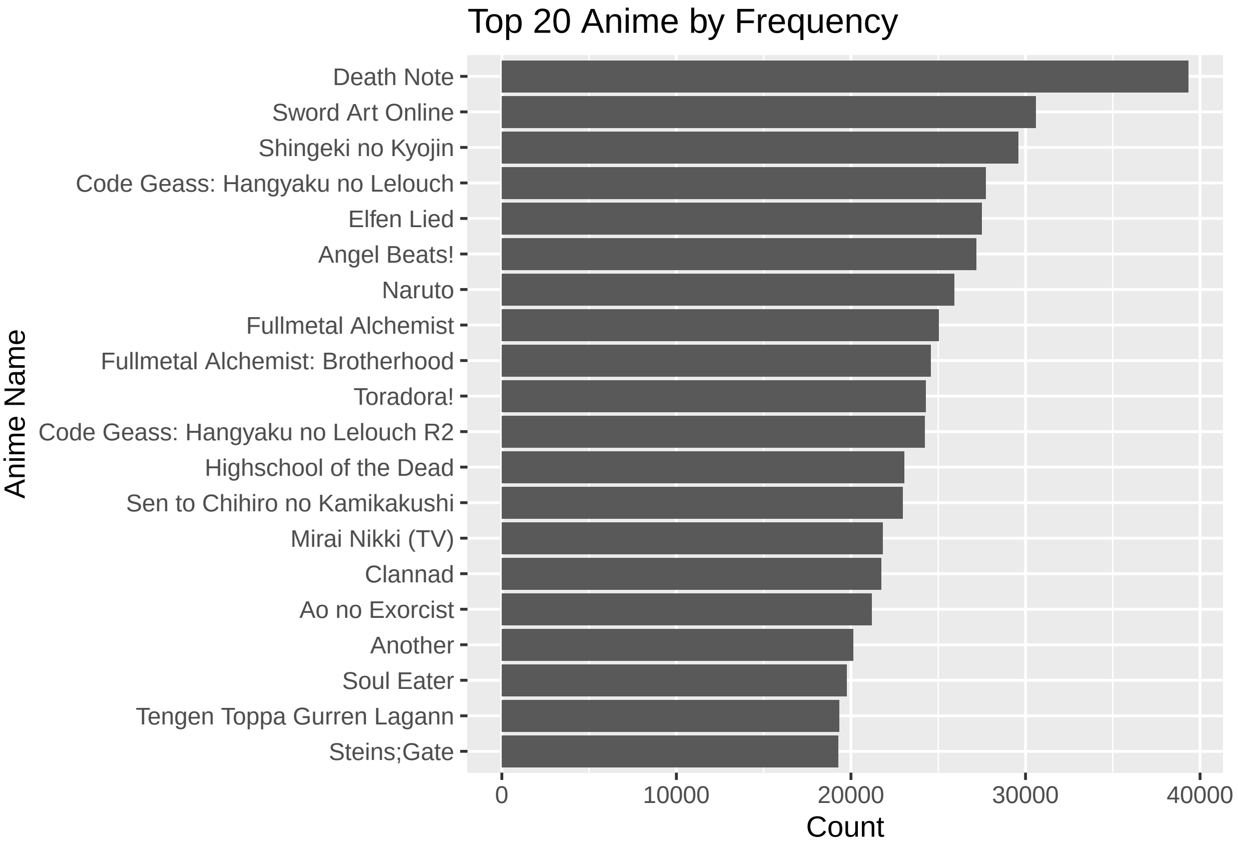


Figure 7: Names of the 20 most frequent anime found in the lists of anime previously watched by users on myanimelist.net

Table 3: Top 10 Cluster 3 Anime by Rating

name	episodes	rating	members
Spoon-hime no Swing Kitchen	187	9.60	47
Ryouma 30 Seconds	131	8.80	54
Hajime no Ippo: New Challenger	89	8.75	88995
GintamaÂ°: Aizome Kaori-hen	73	8.69	16947
Hajime no Ippo: Rising	88	8.68	66756
Minna no Doutoku	143	8.67	41
Warera Salaryman Tou	89	8.67	57
Kamisama Hajimemashita: Kako-hen	119	8.64	33422
Gintama: Yorinuki Gintama-san on Theater 2D	73	8.60	11104
JoJo no Kimyou na Bouken: Stardust Crusaders 2nd Season	85	8.60	93657

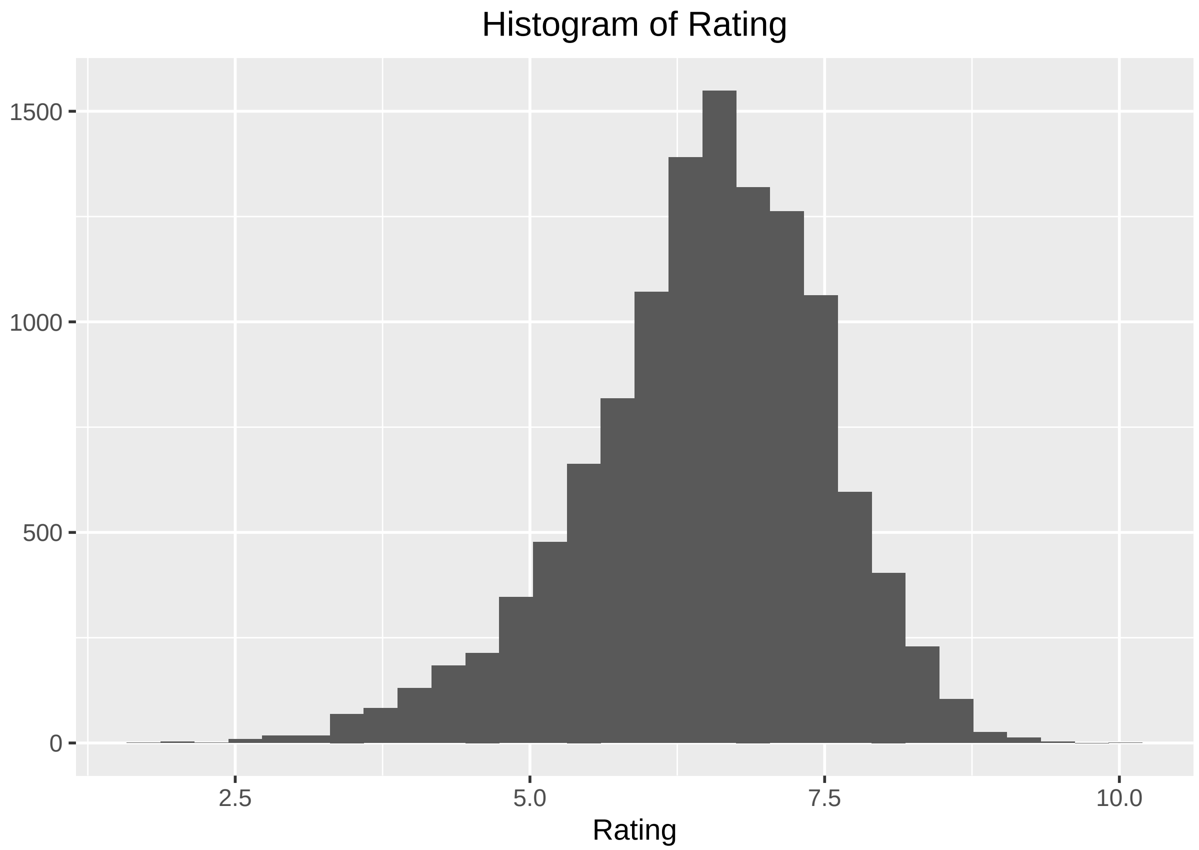


Figure 8: Histogram of the ratings given to various anime by users on myanimelist.net. Note the peak around 6-6.5 and the major drop on either end.

Table 4: Anime Association Rules with Highest Lift

rules	support	confidence	lift
{Evangelion: 1.0 You Are (Not) Alone} => {Evangelion: 2.0 You Can (Not) Advance}	0.10	0.83	7.63
{Evangelion: 2.0 You Can (Not) Advance} => {Evangelion: 1.0 You Are (Not) Alone}	0.10	0.94	7.63
{Clannad,Clannad: Mou Hitotsu no Sekai, Tomoyo-hen} => {Clannad: After Story - Mou Hitotsu no Sekai, Kyou-hen}	0.10	0.80	7.46
{Clannad,Clannad: After Story - Mou Hitotsu no Sekai, Kyou-hen} => {Clannad: Mou Hitotsu no Sekai, Tomoyo-hen}	0.10	0.95	7.46
{Clannad: After Story - Mou Hitotsu no Sekai, Kyou-hen} => {Clannad: Mou Hitotsu no Sekai, Tomoyo-hen}	0.10	0.95	7.43
{Clannad: Mou Hitotsu no Sekai, Tomoyo-hen} => {Clannad: After Story - Mou Hitotsu no Sekai, Kyou-hen}	0.10	0.80	7.43
{Bleach Movie 2: The DiamondDust Rebellion - Mou Hitotsu no Hyourinmaru} => {Bleach Movie 1: Memories of Nobody}	0.11	0.92	6.75
{Bleach Movie 1: Memories of Nobody} => {Bleach Movie 2: The DiamondDust Rebellion - Mou Hitotsu no Hyourinmaru}	0.11	0.79	6.75
{Toaru Majutsu no Index} => {Toaru Majutsu no Index II}	0.10	0.65	6.39
{Toaru Majutsu no Index II} => {Toaru Majutsu no Index}	0.10	0.99	6.39
{Kimi ni Todoke 2nd Season} => {Kimi ni Todoke}	0.11	0.99	6.33
{Kimi ni Todoke} => {Kimi ni Todoke 2nd Season}	0.11	0.68	6.33
{Kuroko no Basket} => {Kuroko no Basket 2nd Season}	0.11	0.71	6.30
{Kuroko no Basket 2nd Season} => {Kuroko no Basket}	0.11	0.99	6.30
{Full Metal Panic! The Second Raid} => {Full Metal Panic!}	0.11	0.98	6.27
{Full Metal Panic!} => {Full Metal Panic! The Second Raid}	0.11	0.69	6.27
{Toradora!,Zero no Tsukaima,Zero no Tsukaima: Futatsuki no Kishi} => {Zero no Tsukaima: Princesses no Rondo}	0.10	0.89	6.26
{Toradora!,Zero no Tsukaima: Futatsuki no Kishi} => {Zero no Tsukaima: Princesses no Rondo}	0.10	0.89	6.24
{Kami nomi zo Shiru Sekai II} => {Kami nomi zo Shiru Sekai}	0.12	0.99	6.16
{Kami nomi zo Shiru Sekai} => {Kami nomi zo Shiru Sekai II}	0.12	0.77	6.16
{Ookami to Koushinryou II} => {Ookami to Koushinryou}	0.11	0.99	6.11
{Ookami to Koushinryou} => {Ookami to Koushinryou II}	0.11	0.71	6.11
{Shakugan no Shana II (Second)} => {Shakugan no Shana}	0.12	0.99	6.08
{Shakugan no Shana} => {Shakugan no Shana II (Second)}	0.12	0.72	6.08
{Zero no Tsukaima,Zero no Tsukaima: Futatsuki no Kishi} => {Zero no Tsukaima: Princesses no Rondo}	0.14	0.86	6.05
{Toradora!,Zero no Tsukaima,Zero no Tsukaima: Princesses no Rondo} => {Zero no Tsukaima: Futatsuki no Kishi}	0.10	0.99	6.05
{Zero no Tsukaima,Zero no Tsukaima: Princesses no Rondo} => {Zero no Tsukaima: Futatsuki no Kishi}	0.14	0.99	6.04
{Toradora!,Zero no Tsukaima: Princesses no Rondo} => {Zero no Tsukaima: Futatsuki no Kishi}	0.10	0.99	6.04
{Zero no Tsukaima: Princesses no Rondo} => {Zero no Tsukaima: Futatsuki no Kishi}	0.14	0.99	6.03
{Zero no Tsukaima: Futatsuki no Kishi} => {Zero no Tsukaima: Princesses no Rondo}	0.14	0.86	6.03
{Boku wa Tomodachi ga Sukunai,Sword Art Online} => {Boku wa Tomodachi ga Sukunai Next}	0.10	0.75	6.03
{Kore wa Zombie Desu ka? of the Dead} => {Kore wa Zombie Desu ka?}	0.10	0.98	6.01
{Kore wa Zombie Desu ka?} => {Kore wa Zombie Desu ka? of the Dead}	0.10	0.62	6.01