

Anime Recommendation Database 2020

Big Data and Cloud Computing
Prof. Hossein Amini

Yusheng Xiang 502611

1. Executive summary.....	1
2. Introduction.....	1
i. Description of the data	1
a) anime	1
b) Animelist.....	1
c) rating_complete.....	1
ii. Why is this Big Data?	1
3. Problem Statement	2
i. Question 1: Which genre is most popular?	2
ii. Question 2: The relationship between user rating how many episodes did they Watch?.....	2
iii. Question 3: Are there any highly-biased users?.....	4
iv. Question 4: Which sources are best among all the sources based on their average score.....	4
v. Question 5: Which sources are best among all the sources based on their popularity.	5
vi. Question 6: Based on the average score of all the anime, which studios have the best reputations?.....	6
vii. Question 7: Which anime have the highest drop-rate	7
viii. Question 8: What pattern can we find among the anime with the highest drop rate?.....	8
ix. Question 9: Can we explain why specific anime has the highest drop rate?	8
4. Conclusion	8
5. Appendices.....	9
6. Reference	13

1. Executive summary

This report provides a series of analyses of Anime Recommendation Database 2020. It extracts the most crucial information from 17,652 anime and the preferences of different users. The purpose of the report is to discover what factors will significantly affect the animation's ratings and views. In order to get the result, this report has nine questions focused on anime.

The first question is to find out which anime genre is most popular. Question two and three are focused on MyAnimeList.net user and user rating. The four to six questions discover the relationship between anime sources, average user rating scores, popularity, and studio. The seven and eight questions are focused on which anime and anime patterns have the highest drop rate. The last question explains why specific anime has the highest drop rate.

This report first describes the dataset and then shows the method to solve all questions. After visualizing the results, it makes some recommendations for audiences and producers. All analysis methods are based on Hadoop, Linux, and Tableau. All codes are in the appendix.

2. Introduction

i Description of the dataset

The reporting dataset is named Anime Recommendation Database 2020, created by Hernan Valdivieso from Kaggle on July 13th, 2021. This semi-structured dataset has five CSV (Comma Separated Values) files, which amount to 2.87 GB (gigabytes), and it contains information about 17,562 anime and the preferences from 325,772 different users.

Therefore, our team aims to transfer these data to display into a file that audiences can easily read. Although this dataset has five files, our analysis is only based on three files: anime, animelist, and rating_complete. The following are the descriptions of the three files.

a) anime

The file named anime is 1.5 MB (megabyte), contains 12 different general information of 17,542 different anime like genre, stats, and studio. The explained of each column are in *Appendix 1*.

b) animelist

The file named animelist is 2.03 GB (gigabytes), contains 5 different information which are all animes registered by the user with the respective score, watching status, and numbers of episodes watched. The explained of each column are in *Appendix 2*.

c) rating_complete

The file named rating_complete is 817.9 MB (megabyte), contains 3 different information, which is a list of ratings given by the user to animes with already been completely watched. The explained of each column are in *Appendix 3*.

ii Why is this Big Data?

Big data can be defined as data that we cannot analyze using traditional processes or tools.

And three famous V's for big data are volume, variety and velocity. The datasets we chose have a total size of nearly 3G, and the animelist.csv contains 109 million rows. So the volume is quite huge. Meanwhile, the datasets have different types of data such as integer, datetime, string values and so on, which creates a high level of variety. Last but not least, the datasets are in the form of csv files, so we can process the information we need after massaging the datasets by certain techniques quickly.

3. Problem Statement

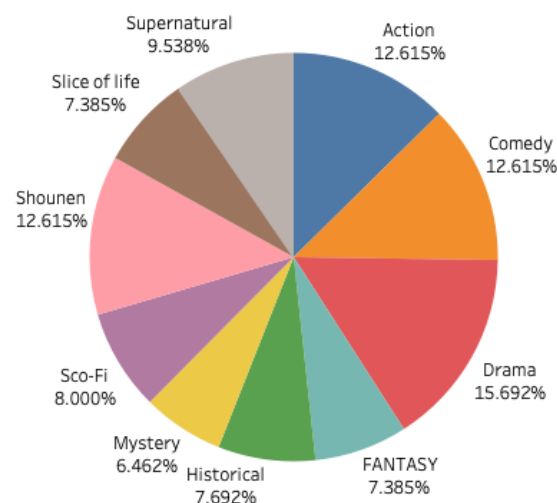
i Question 1: Which genre is most popular?

In order to find the most popular genre, we decided to consider two aspects: rating and number of times being watched. After cleaning the data with MapReduce technique, we utilized Apache Hive to sort the average value of both rating and number of times being watched for each anime. We then selected the top 100 anime for each strategy and applied MapReduce again to obtain a text file containing all the genres of anime. After that, we applied Linux commands to find unique genres and count the number of each genre.

```
[x.yusheng@ip-172-31-95-86 ~]$ cat genre.txt | sort | uniq -c | sort -rn | head -10
51 Drama
41 Shounen
41 Comedy
41 Action
31 Supernatural
26 Sci-Fi
25 Historical
24 Slice of Life
24 Fantasy
21 Mystery
```

From the result of applying average score, the most popular genres are action, supernatural, drama, shounen and drama. From the result of applying the number of times being watched, the most popular genres are shounen, supernatural, comedy and adventure.

The visual plot:



The MapReduce and HIVE codes are in Appendix 4

ii **Question 2: The relationship between user rating and how many episodes did they watch?**

By applying HIVE on the rating_complete.csv we uploaded, we group the dataset by individual user id and calculate the average value of ratings they gave and the sum of episodes they had watched.

	user_id	avg(rating)	sum(watched_episodes)
1	148295	6.453310696095077	4331006
2	115661	3.4710982658959537	3546727
3	100624	5.2988802756244615	3290706
4	264961	6.916666666666667	1733769
5	217753	2.0407969639468693	1334051
6	189037	6.800537735259176	1186630
7	342683	0	1182420
8	348328	5.467018469656992	1058140
9	13478	0.1527777777777778	1049862
10	266170	5.731481481481482	987333

The result shows that for the top 10 users who watched the most number of episodes, their average score ranges from 2 to 6.9, which is relatively low. Such a phenomenon can be explained by the fact that the more anime a person has watched, the more strict a person he will be.

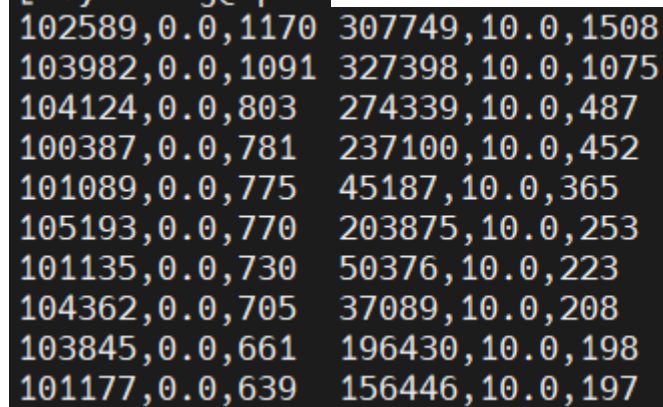
The visual plot:



The code for HIVE is in Appendix 5.

iii Question 3: Are there any highly-biased users?

By applying MapReduce technique on complete_rating.csv, we created a dictionary containing all users, and we created a dictionary for each user containing several lists including the anime id, user's rating, user's watching status and the number of episodes the user has watched. Then we calculated the average rating given by each user and used Linux commands to find highly_biased users.



```
102589,0.0,1170 307749,10.0,1508
103982,0.0,1091 327398,10.0,1075
104124,0.0,803 274339,10.0,487
100387,0.0,781 237100,10.0,452
101089,0.0,775 45187,10.0,365
105193,0.0,770 203875,10.0,253
101135,0.0,730 50376,10.0,223
104362,0.0,705 37089,10.0,208
103845,0.0,661 196430,10.0,198
101177,0.0,639 156446,10.0,197
```

The screenshot above represents the information we need. The first column is the User_id, the second column is the user's average rating and the third column is the number of anime the user has watched. We found that some users gave 0s for all the anime he has watched and some users gave 10s for all the anime he has watched.

The MapReduce function is in Appendix 6.

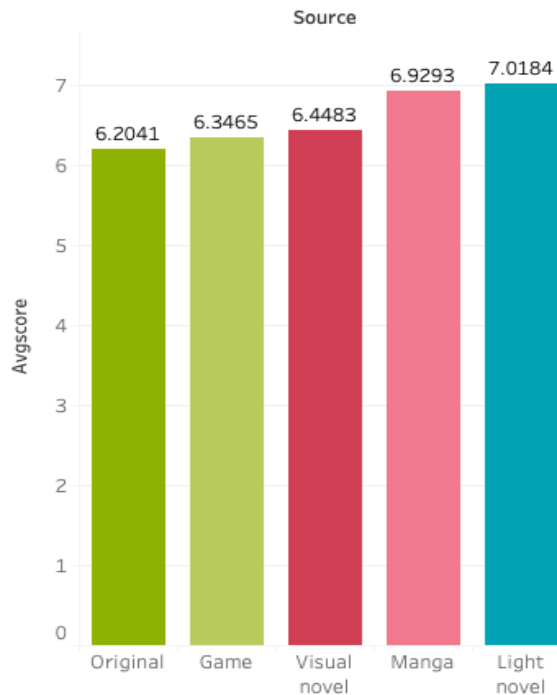
iv Question 4: Which sources are best among all the sources based on their average score.

By applying HIVE on the anime.csv we uploaded, we grouped the dataset by the source and calculated the average score for each source of anime. The result shows that anime with light novels as source scores highest among other anime, the second best source is manga and then visual novels.

	avgscore	source
1	7.018446866485016	Light novel
2	6.929264380530981	Manga
3	6.448270181219104	Visual novel
4	6.346522842639596	Game
5	6.20406175771972	Original

The top 4 sources (Light novel > Manga > Visual novel > Game) all have an accumulated fanbase, who are highly likely to rate the adapted animation high. Besides, since they are adapted from script foundation, their plot and character design tend to be more logical and attractive. However, original anime doesn't have such an accumulated fanbase and may have a worse plot and character design due to a lack of script foundation, which leads to its low average score.

The visual plot:



The code for HIVE is in Appendix 7.

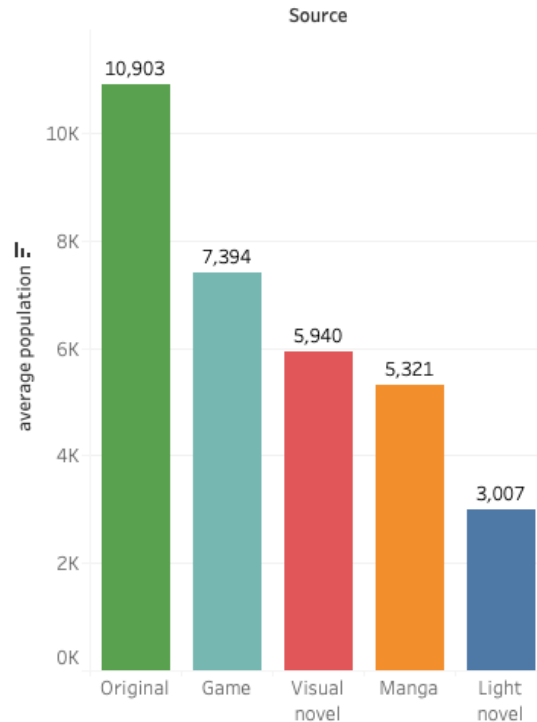
v **Question 5: Which sources are best among all the sources based on their popularity.**

Like the problem above, we took a similar technique except that we are calculating the average popularity rankings for each source.

	avgpopularity	source
1	3006.9520202020203	Light novel
2	5321.0079443892755	Manga
3	5940.252019386106	Visual novel
4	7393.769547325103	Game
5	10903.470790378007	Original

Combined with the result in question 4, it turns out that anime sources with higher average scores also have higher average popularity rankings. This result can be explained with reasons given in question 4. Compared with original anime, the top 4 sources (Light novel > Manga > Visual novel > Game) all have an accumulated fanbase, who are highly likely to watch the anime. This contributes to higher average popularity rankings of these 4 anime sources.

The visual plot:



The code for HIVE is in Appendix 8.

vi **Question 6: Based on the average score of all the anime, which studios have the best reputations?**

For this problem, we still applied HIVE on anime.csv and grouped the dataset by studios and calculated the average score for anime produced by each studio.

Studios	round(avg(Score), 2)
▶ Madhouse', MAPPA	8.6
Studio Bind	8.37
Gainax', Tatsunoko Production	8.32
Gallop', Studio Comet	8.29
J.C.Staff', Egg Firm	8.28
Tezuka Productions', MAPPA	8.19
Artland', Magic Bus	8.15
B.CMAY PICTURES', Colored-Pencil Animation Design	8.03
Nice Boat Animation', Samsara Animation Studio	8
Gainax', Production I.G	8
Satelight', 8bit	7.97
Artland', Madhouse	7.95
Gonzo', Production I.G	7.94
Madhouse', TMS Entertainment	7.93
AIC Spirits', BeSTACK	7.84
Sunrise', Ascension	7.83

We found that MAPPA has the highest reputation and the result makes sense because MAPPA is one of the most famous anime studios and it has produced many great anime such as 'Attack on Titan'. The second best studio is Studio Bind, and by searching this studio online

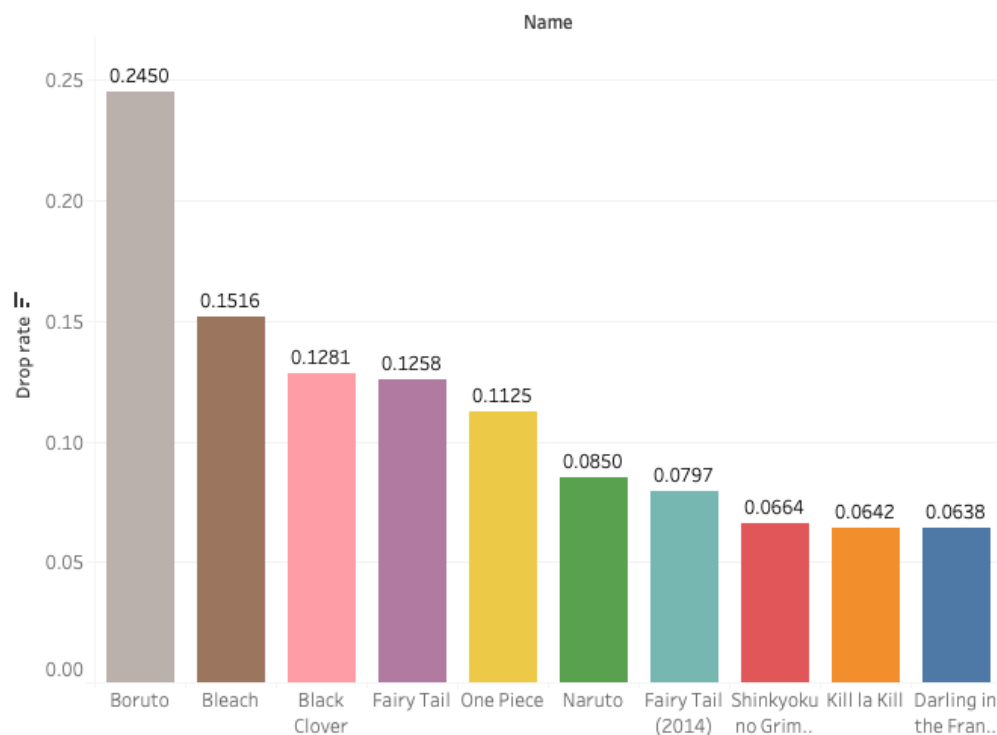
we found it only produced two anime but both of them received great feedback. So we recommend Studio Bind to produce more anime.
The code for HIVE is in Appendix 9.

vii **Question 7: Which anime have the highest drop-rate?**

By using HIVE, we divided the number of dropped by the sum of the number of dropped, watching, completed and on-hold to calculate the drop rate for each anime. In this question, we specified anime with a dropped number greater than 50000 to ensure that there are enough viewers and can hence avoid meaningless data.

MAL_ID	Dropped	Dropped+Watching+Completed+On-Hold	Drop_rate
34566	113677	463963	0.24501307216308196
269	174710	1152578	0.15158193198204373
34572	89594	699508	0.12808145153450712
6702	148408	1179664	0.12580531405552767
21	136245	1211530	0.11245697588999035
1735	124253	1461727	0.08500424497871353
22043	50700	636430	0.07966312084596892
3588	65962	993956	0.06636309856774344
18679	67845	1057345	0.06416543323134832
35849	53880	843920	0.06384491420987772

We found that ‘Boruto: the next generation’ has the highest drop rate.
The code for HIVE and web scrapping is in Appendix 10.
The visual plot:



viii **Question 8: What pattern can we find among the anime with the highest drop rate?**

Other than ‘Boruto: the next generation’, which is still serializing and only has a score of 5.9, all other anime in the list have a score over 7.5.

	MAL_ID	Episodes
▶	269	366
	1735	500
	6702	175
	22043	102
	34566	Unknown
	34572	170

By slicing anime.csv with the index of anime in the list, we found that most anime have more than 100 episodes. So we assume having too many episodes might be the reason for people dropping the anime.

ix **Question 9: Can we explain why specific anime has the highest drop rate?**

In order to find out why ‘Boruto: the next generation’ has the highest drop rate, we decided to check comments of users who have watched this anime. We used web scraping to crawl 30 random comments from IBDM. We then applied MapReduce to clean the data and sorted all the words in the comment.

sucks
style
stupid flaws hated worse garbage

dwehdwen
dwedbwhdbwh
dwedbwhdbhwebdbwhwebdbhewbhdhewbdhebwhdew

We found several negative words such as garbage, stupid, worth etc. And it looks like most people like comparing ‘Boruto’ with ‘Naruto’, which is the precursor of this anime. So the reason might be people have high expectations about ‘Boruto’ but it did not satisfy their expectations.

The code for MapReduce and Linux commands is in Appendix 11.

4. Conclusion

Based on question 4 and 5, the average scores and average popularity rankings are the same, which is Light novel > Manga > Visual novel > Game. So we suggest that when audiences want to explore new anime, they can select anime based on this order of sources ranking. Based on question 6, studios (MAPPA, Studio Bind, gainax, Tatsunoko Production) have the best reputations according to the average score of the anime they produced. We suggest audiences pay more attention to these companies’ newly released anime because they may

have a higher quality. Besides, not only big companies like MAPPA produces great anime, start-up companies like Studio Bind also produce popular anime.

Based on question 7 and 8, we find that there are two main causes for the high drop rate of anime. One is the low quality of the anime, which can be reflected by the 5.9 score of the anime“Boruto”. The other cause is that these anime have too long episodes. Therefore, we suggest audiences don’t just quit watching anime at a glance of the drop rate, they should also pay attention to anime’s score.

Based on question 9, through web scraping of 30 random comments from IBDM, apart from the negative words, we find many people compare ‘Boruto’ with its precursor ‘Naruto’. Since audiences are disappointed due to ‘Boruto’ ’s worse performance than ‘Naruto’, the director and producer company should listen to audiences’ voices and try to improve the plot and character design of ‘Boruto’ by learning lessons from ‘Naruto’.

5. Appendices

Appendix 1:

MAL_ID	MyAnimelist ID of the anime.
Name	Full name of the anime
Score	Average score of the anime given from all users in MyAnimelist database.
Genres	comma separated list of genres for this anime.
Type	TV, movie, OVA, etc.
Episodes	Number of chapters.
Producers	Comma separated list of producers
Studios	Comma separated list of studios
Source	Manga, Light novels, books, etc.
Popularity	Position based on the number of users who have added the anime to their list.
Watching	Number of users who are watching the anime
Completed	Number of users who have completed the anime

On-Hold	Number of users who have the anime on Hold
Dropped	Number of users who have dropped the anime

Appendix 2:

User_id	Non identifiable randomly generated user id
Anime_id	MyAnemlist ID of the anime
Rating	Score between 1 to 10 given by the user. 0 if the user didn't assign a score
Watching_status	State ID from this anime in the anime list of this user
Watched_episodes	Numbers of episodes watched by the user

Appendix 3:

User_id	Non identifiable randomly generated user id
Anime_id	MyAnimelist ID of the anime that this user has rated
Rating	Rating that this user has assigned

Appendix 4:

Mapper:

```
#!/usr/bin/env python
import sys
for line in sys.stdin:
    line = line.rstrip("\n")
    line = line[0:-3]
    words = line.split("\t", 1)
    for word in words:
        print word

$ cat genre.csv | python genre_mapper.py > genre.txt
```

Hive:

```
SELECT * FROM anime_zhuanyi1
WHERE Score != 'Unknown'
```

```
ORDER BY Score DESC
LIMIT 100;

SELECT MAL_ID, Score, Genres, Rating, Watching, Completed FROM
anime_zhuanyil
WHERE Score != 'Unknown'
ORDER BY (Watching + Completed) DESC
LIMIT 100;
```

Appendix 5:

```
SELECT user_id, avg(rating), sum(watched_episodes) FROM animelist
GROUP BY user_id
ORDER BY sum(watched_episodes) DESC
LIMIT 10;
```

Appendix 6:

Mapper:

```
#!/usr/bin/env python
import sys
for line in sys.stdin:
    line = line.strip()
    words = line.split()
    for word in words:
        print word
```

Reducer:

```
#!/usr/bin/env python
import sys
rating_all = {}
for line in sys.stdin:
    user_id, anime_id, rating, watching_status, watched_episodes =
line.strip().split(',')
    try:
        rating = float(rating)
    except ValueError:
        continue
    try:
        rating_all[user_id]['anime_id'].append(anime_id)
        rating_all[user_id]['rating'].append(rating)
        rating_all[user_id]['watching_status'].append(watching_status)
        rating_all[user_id]['watched_episodes'].append(watched_episodes)
```

```

except:
    rating_all[user_id] = {'anime_id':[anime_id], 'rating':[rating],
'watching_status':[watching_status], 'watched_episodes':[watched_episodes]}

for user_id in rating_all.keys():
    rating_list = rating_all[user_id]['rating']
    average_rating = sum(rating_list)/len(rating_list)
    print '%s\t%s\t%s' %(user_id, average_rating, len(rating_list))

```

Appendix 7:

```

select avg(score) as avgscore, source
from default.group8_anime4
where source = "Manga" or source = "Original" or
source = "Light novel" or source = "Visual novel" or source = "Game"
group by source
order by avgscore desc;

```

Appendix 8:

```

select avg(popularity) as AvgPopularity, source
from default.group8_anime4
where source = "Manga" or source = "Original" or
source = "Light novel" or source = "Visual novel" or source = "Game"
group by source
order by AvgPopularity desc;

```

Appendix 9:

```

SELECT Studios, round(avg(Score), 2) FROM anime_zhuanyi
GROUP BY Studios
ORDER BY avg(Score) DESC;
LIMIT 10

```

Appendix 10:

```

SELECT MAL_ID, Dropped, Dropped+Watching+Completed+`On-Hold`,
Dropped/(Dropped+Watching+Completed+`On-Hold` ) AS Drop_rate
FROM `anime_noname`
WHERE Dropped > 50000
ORDER BY Drop_rate DESC
LIMIT 10

```

Appendix 11:

Web scrapping:

```
import lxml.etree as le

review_lists = ['BORUTO_ Reviews', 'BORUTO_ Reviews2', 'BORUTO_
Reviews3']
content = []
for review_list in review_lists:
    with open(r'C:\Users\chung\Desktop\Big Data Final\%s.html'%review_list, 'r',
encoding='utf-8') as f:
        html = f.read()
        html_x = le.HTML(html)
        reviews = html_x.xpath("//div[@class='trunc']/text()")

        for review in reviews:
            content.append(review.strip())

with open(r'C:\Users\chung\Desktop\Big Data Final\comment.txt', 'w') as f:
    f.write(str(content))
```

MapReduce:

```
#!/usr/bin/env python
import sys
for line in sys.stdin:
    line = line.strip('[,]')
    words = line.split(',')
    for word in words:
        wordlist = word.split(' ')
        for i in wordlist:
            print
i.strip("").strip("").strip('.').strip(' ').strip('(').strip('?')
```

Linux command:

```
cat reduced_comment.txt | sort | uniq -c | sort -rn
```

6. Reference

link for the dataset (from Kaggle):

<https://www.kaggle.com/hernan4444/animerecommendation-database-2020?select=animelist.csv>

link for the source of data (myanimelist.net):

<https://myanimelist.net/>