



ANIME STUDIOS

ANIME RECOMMENDATION DATABASE 2020

DATA SOURCES

kaggle™

Anime Recommendation Database 2020

Recommendation data from 320.0000 users and 16.000 animes at
myanimelist.net



Anime Recommendation Database 2020 | Kaggle

<https://www.kaggle.com/datasets/hernan444/anime-recommendation-database-2020?select=anime.csv>

COLUMN/SCHEMA

	Name	Type
1	index	int
2	mal_id	int
3	name	string
4	score	float
5	englishname	string
6	japanesename	string
7	type	string
8	episodes	bigint
9	aired	string
10	premiered	string
11	producers	string
12	licensors	string
13	studios	string
14	source	string
15	duration	string
16	rating	string
17	ranked	int
18	popularity	int
19	members	int
20	favorites	int
21	watching	int
22	completed	int
23	onhold	int
24	dropped	int
25	plantowatch	int
26	score10	int
27	score9	int
28	score8	int
29	score7	int
30	score6	int
31	score5	int
32	score4	int
33	score3	int
34	score2	int
35	score1	int



Hive Metastore : animelist

COLUMN/SCHEMA

	Name	Type
1	 user_id	bigint
2	 anime_id	bigint
3	 rating	smallint
4	 watching_status	smallint
5	 watched_episodes	smallint



animelist.csv

Hive Metastore : reviewlist

COLUMN/SCHEMA

COLUMNS (3)		
	Name	Type
1	index	bigint
2	name	string
3	genre	string

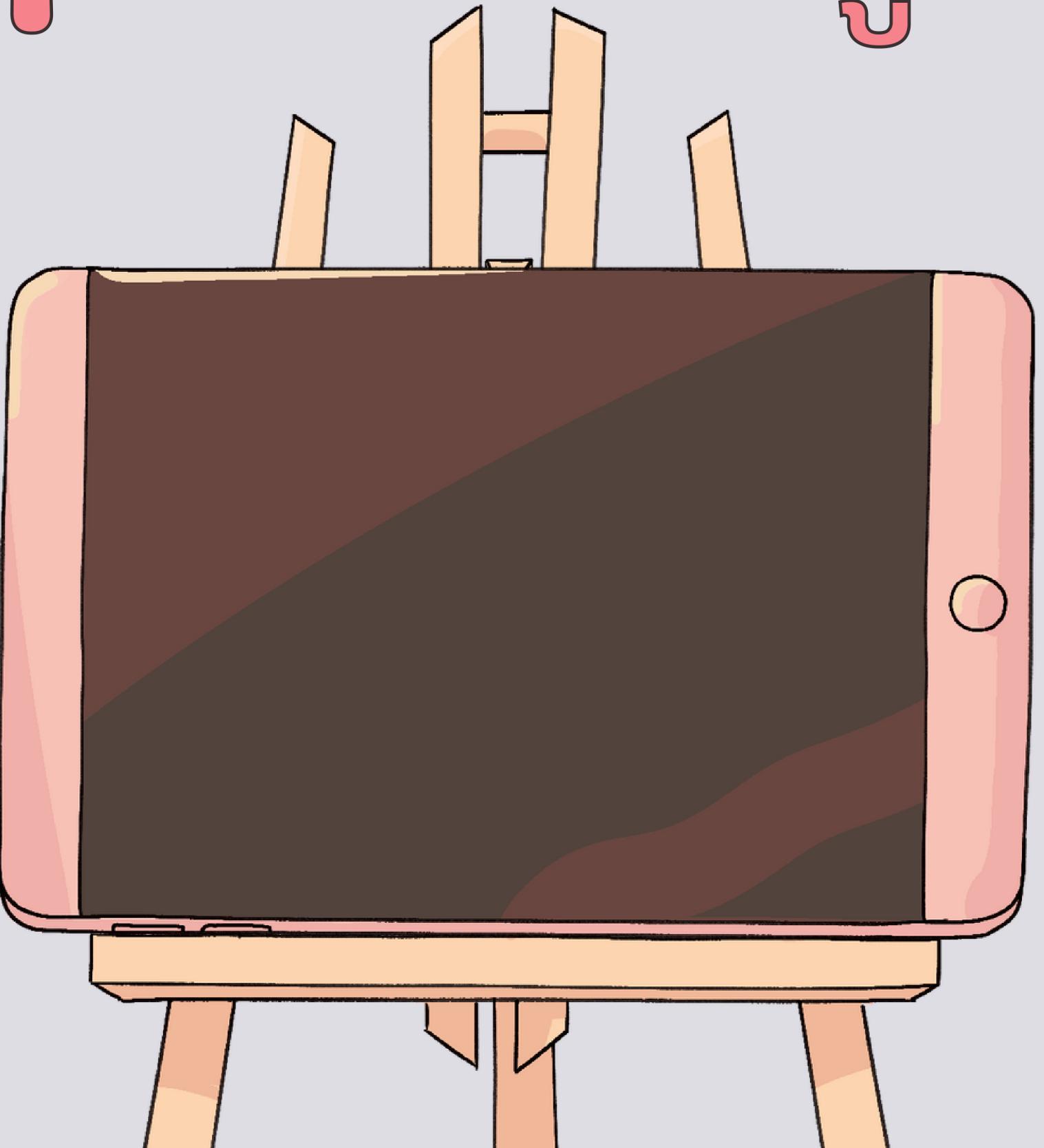
Hive Metastore : anime_genres

IMPORT DATA



ขั้นตอนการ import ข้อมูล

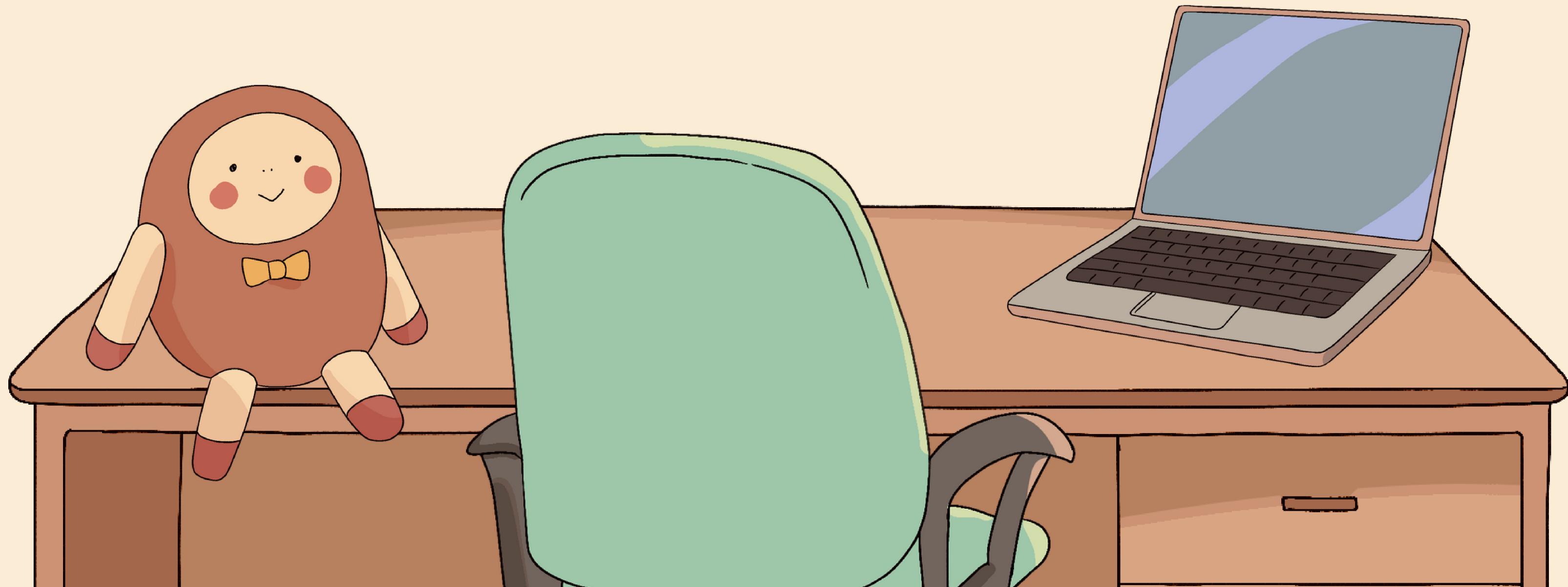
- นำข้อมูลบางชุดมาทำ Normalize ให้เป็น 1NF
เนื่องจากข้อมูลมีบาง Column เป็น Multivalue
โดยใช้ Pandas
- Upload ข้อมูลลงใน HDFS
- ใช้ Hive Metastore อ่านไฟล์และสร้างตาราง



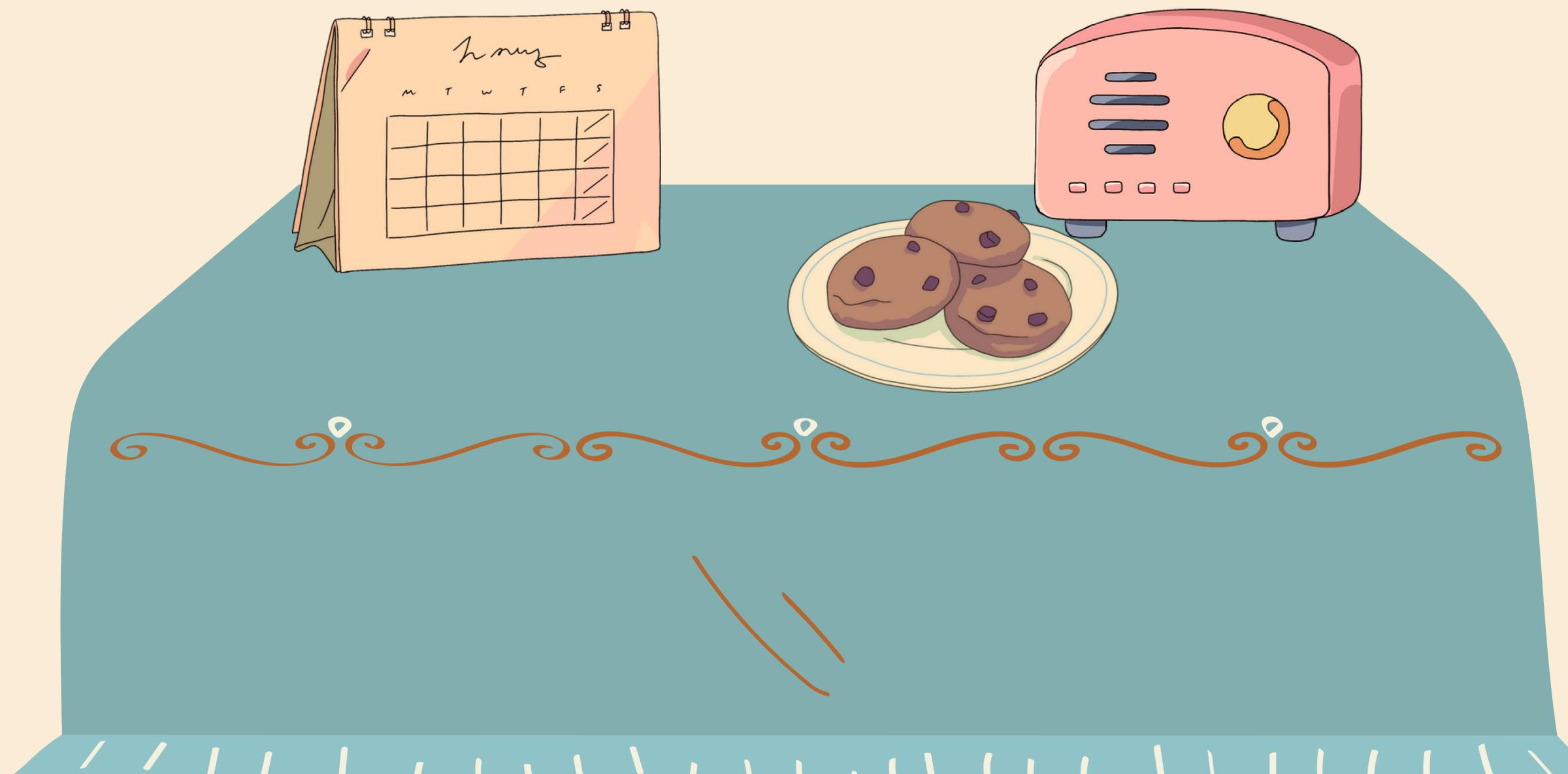


IMPALA

Apache Impala is the open source, native analytic database
for Apache Hadoop.



ແຕ່ລະປົມອນີເມະໄໝມໍກີເຮືອງ



ແຕ່ລະປີນຶອບີເມະໄທມໍກີເຮື່ອງ

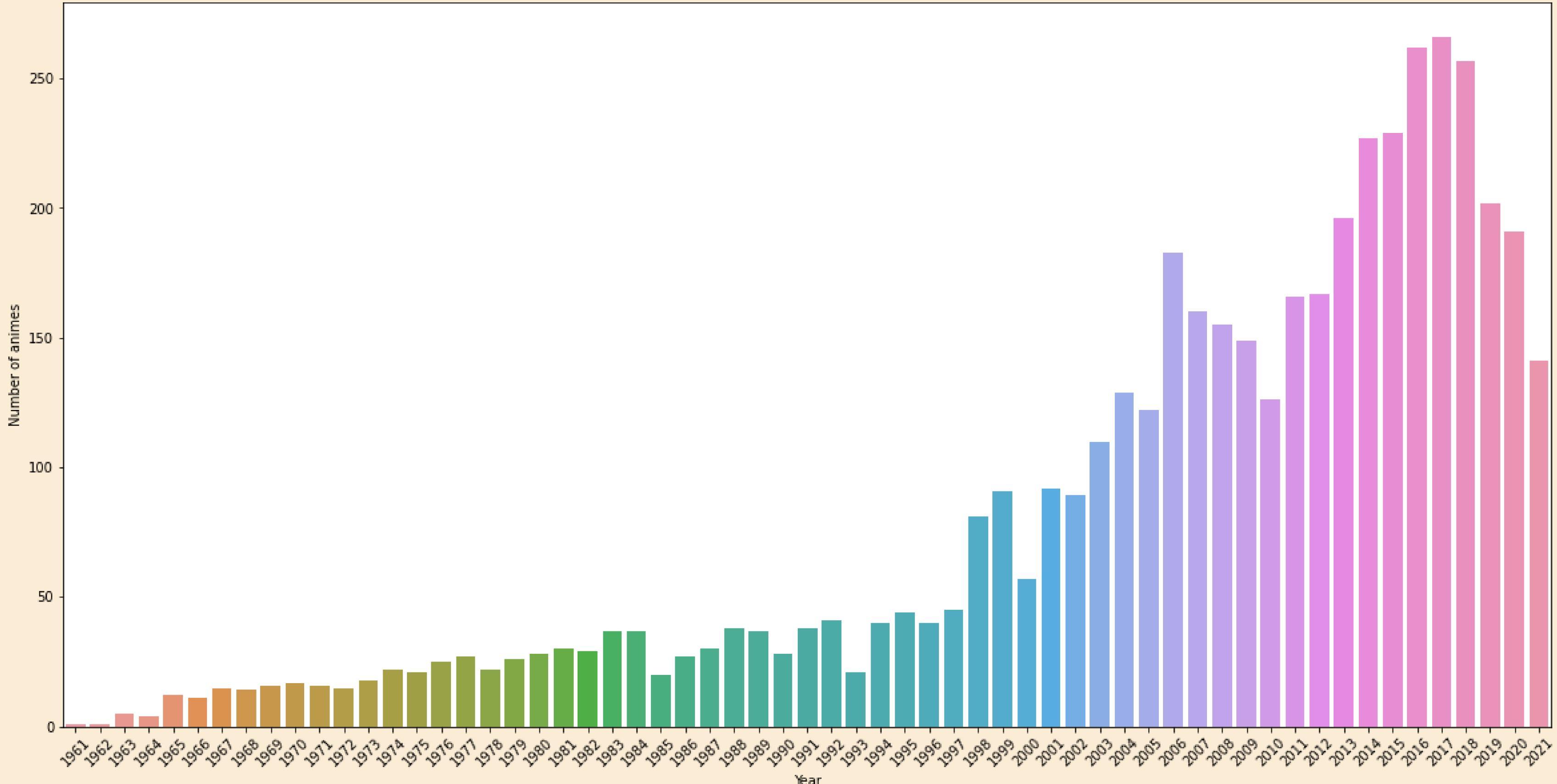
ອນິເມະໄທມໍແຕ່ລະປີ ຄົວ ໃນແຕ່ລະປີ ຜູ້ພລັຕ Anime ສຕູດໂວຕ່າງໆ ຈະປລ່ອຍພລງານໃຫ້ໄດ້ຮັບ
ໝາກນ ເພື່ອໃຫ້ຜູ້ໜ້າຮັບຜູ້ປະກອບການສາມາຄເໜີນແນວໂນ້ມວຸຕສາກຣມອນິເມະໄດ້

CODE

```
SELECT sub.Year, COUNT(*) as count_of_anime
FROM
(SELECT substring(premiered, -4,4) AS Year
FROM animelist
WHERE premiered != 'Unknown' ) sub
GROUP BY sub.Year
ORDER BY sub.Year
```

Result

	Year	Number of animes
0	1961	1
1	1962	1
2	1963	5
3	1964	4
4	1965	12
...
56	2017	266
57	2018	257
58	2019	202
59	2020	191
60	2021	141



ອົບເມະຄະແລ້ວ TOP 10 (1961-2021)



อนิเมะคะแนน TOP 10 (1961-2021)

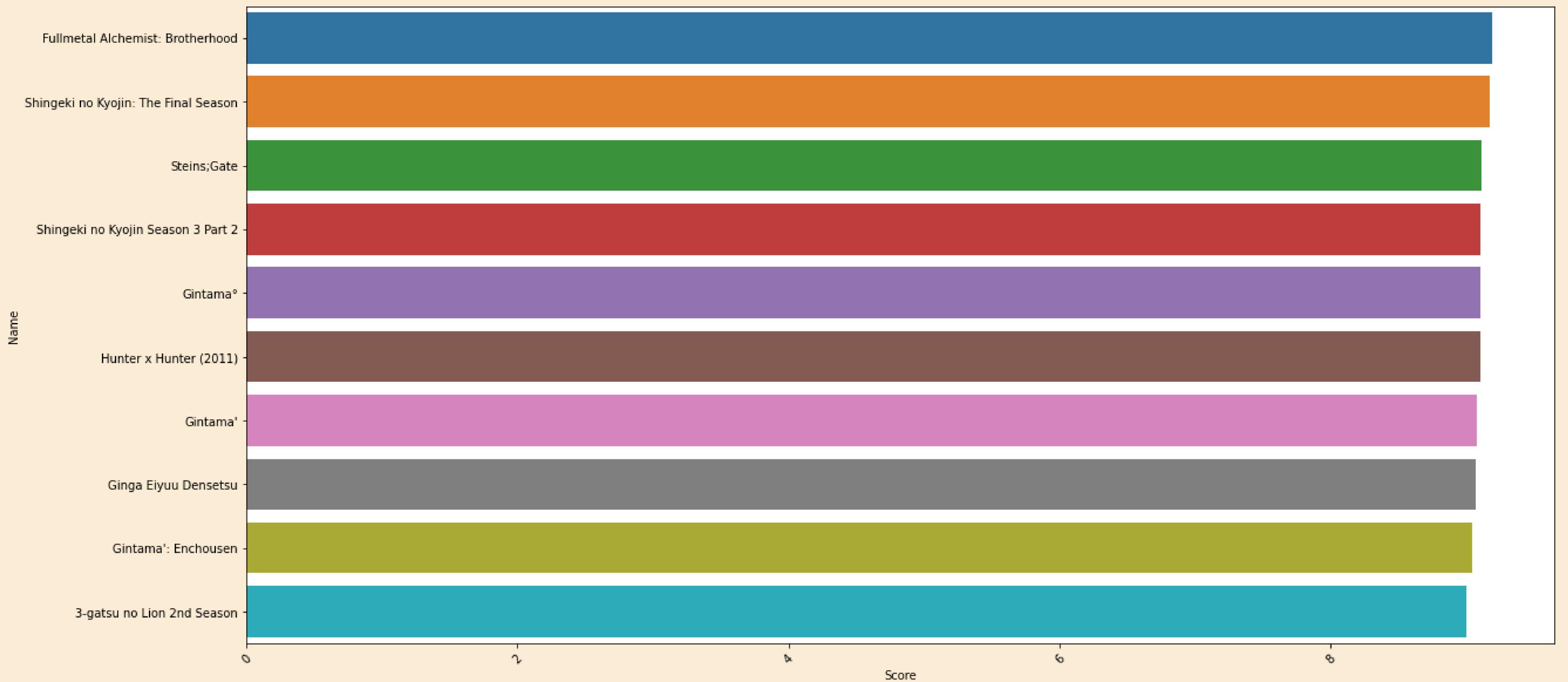
การให้คะแนนของผู้ชมต่อ Anime คือ ภายในเว็บไซต์ MyAnimeList User สามารถให้คะแนน/รีวิวอนิเมะในเรื่องต่างๆ ได้ และแสดงให้เห็นถึงความชื่นชอบของผู้คนส่วนใหญ่ กับอนิเมะเรื่องนั้น โดยเป็นการให้คะแนนสรุปตั้งแต่ปี 1961 จนถึง 2021

CODE

```
SELECT name,score FROM animelist  
Where score > 1  
ORDER BY score DESC  
LIMIT 10;
```

Result

	Name	Score
0	Fullmetal Alchemist: Brotherhood	9.19
1	Shingeki no Kyojin: The Final Season	9.17
2	Steins;Gate	9.11
3	Shingeki no Kyojin Season 3 Part 2	9.10
4	Gintama°	9.10
5	Hunter x Hunter (2011)	9.10
6	Gintama'	9.08
7	Ginga Eiyuu Densetsu	9.07
8	Gintama': Enchousen	9.04
9	3-gatsu no Lion 2nd Season	9.00



ອົບົມະຄະແບນ TOP ຂອງແຕ່ລະເຈັ້ນ



อนิเมะคะแนน TOP ของแต่ละซีซั่น

ภายในเว็บไซต์ MyAnimeList User จะสามารถให้คะแนน/รีวิวอนิเมะในเรื่องต่างๆ ได้ การแสดงคะแนน Top ของแต่ละซีซั่นนั้นแสดงให้เห็นถึงความชื่นชอบของผู้คนกับอนิเมะในแต่ละซีซั่นที่ถูกปล่อยให้รับชมอุ่นมา

CODE

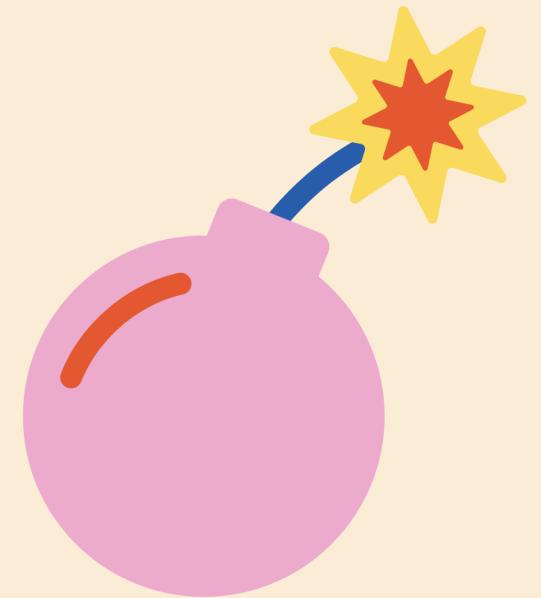
```
select animelist.name, sub.premiered, max_score as score FROM (
    select animelist.premiered, substring(animelist.premiered, -4, 4) as year, MAX(score) AS
    max_score from animelist
    where animelist.premiered is not NULL and score is not NULL and animelist.premiered != "Unknown"
    group by animelist.premiered
) sub
join animelist on animelist.score = sub.max_score and animelist.premiered = sub.premiered
order by sub.year
```

Result

	Name	Premiered	Score
0	Tetsuwan Atom	Winter 1963	7.11
1	Tetsujin 28-gou	Fall 1963	6.20
2	Yuusei Shounen Papii	Summer 1965	5.87
3	Jungle Taitei	Fall 1965	6.56
4	Jungle Taitei: Susume Leo!	Fall 1966	6.49
...
216	Haikyuu!!: To the Top 2nd Season	Fall 2020	8.57
217	Kaguya-sama wa Kokurasetai?: Tensai-tachi no R...	Spring 2020	8.74
218	Re:Zero kara Hajimeru Isekai Seikatsu 2nd Season	Summer 2020	8.50
219	Haikyuu!!: To the Top	Winter 2020	8.39
220	Shingeki no Kyojin: The Final Season	Winter 2021	9.17

221 rows × 3 columns

ตรวจหา ANIME ที่ถูก REVIEW BOMB



REVIEW BOMB

การ Review Bomb คือการที่มีผู้ Review จำนวนมากเข้าไปกลุ่มให้คะแนนที่ต่ำ จนทำให้คะแนนโดยรวมของ Anime เรื่องนั้นตกลงอย่างมาก โดยมีสาเหตุก็จากการต้องการกำลังชื่อเสียง และปัญหาของอนิเมะเอง การที่รู้ว่ามี Review Bomb เกิดขึ้นจะทำให้ผู้ผลิต/ให้บริการสามารถรับรู้และเข้าไปตรวจสอบปัญหาระหว่าง Anime เรื่องนั้นกับผู้ชมได้ และแก้ไขปัญหาที่อยู่ใน Platform ของผู้ให้บริการจะเต็มไปด้วยความคิดเห็นด้านลบที่ไม่สมเหตุสมผลและไม่เป็นธรรมกับผู้ผลิต Anime

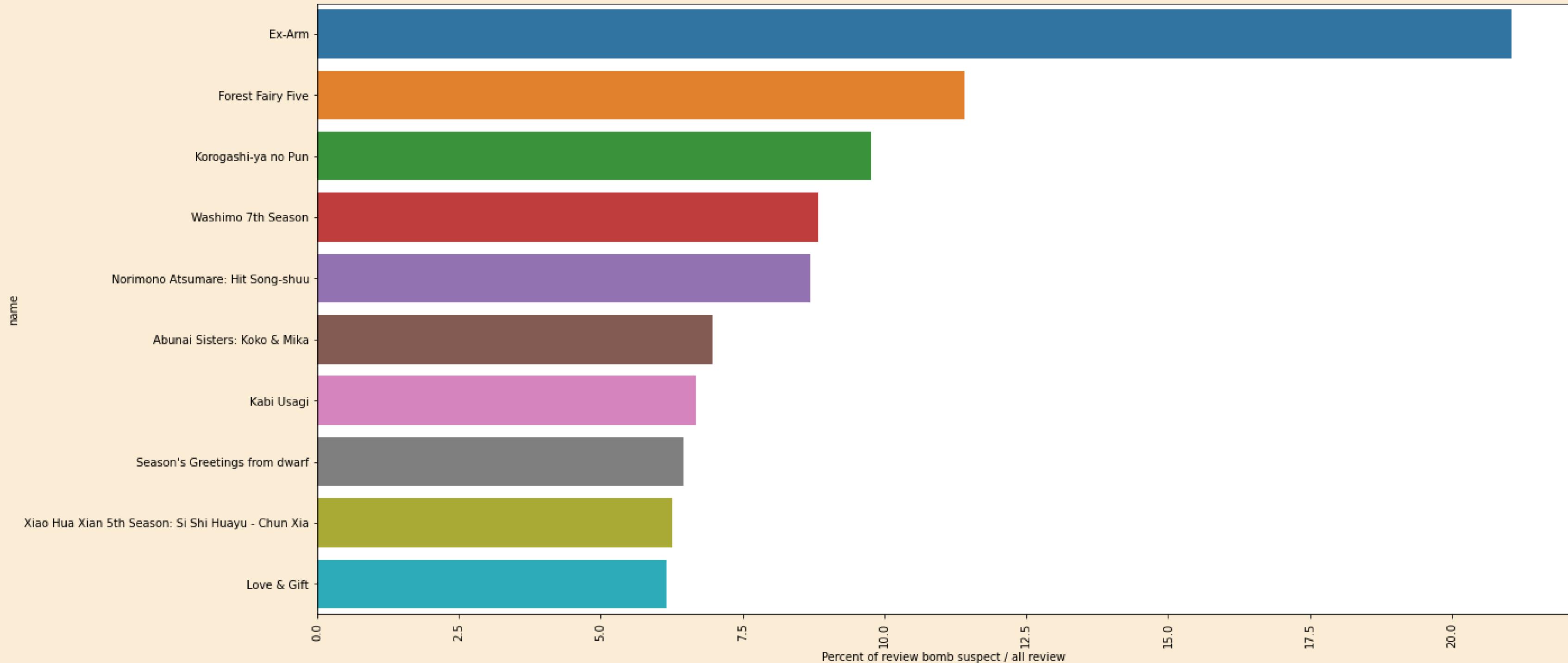
CODE

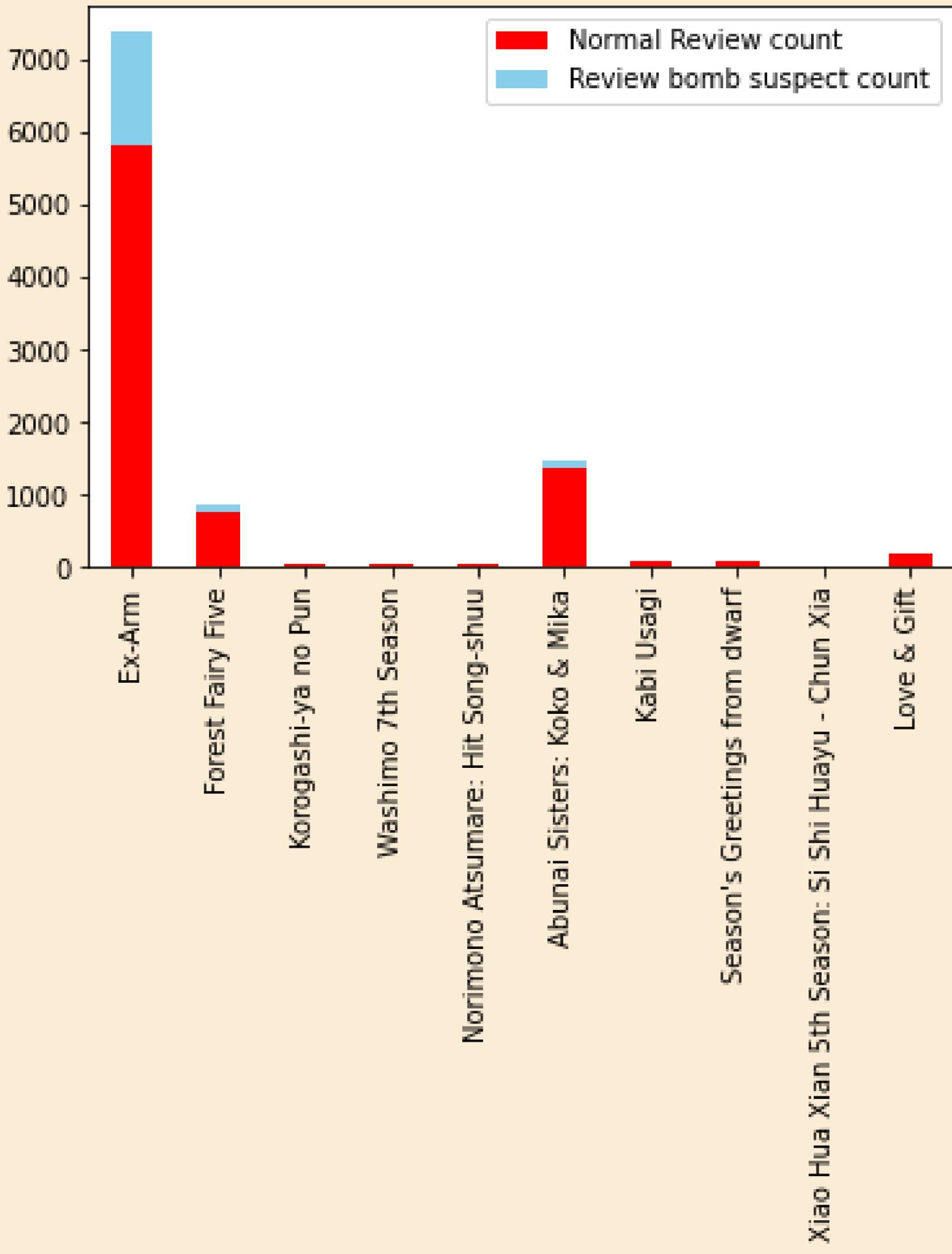
```
SELECT
h.name AS name,
h.n/COUNT(animelist.rating)*100 AS "Percent of review bomb suspect / all review",
h.n AS "Review bomb suspect count",
COUNT(animelist.rating) AS "Review count",
h.episodes AS "Episode(s)",
AVG(animelist.watched_episodes) AS "Average episode watch",
AVG((animelist.watched_episodes/h.episodes)*100) AS "Average watched progress"
FROM (
    SELECT COUNT(r.rating) AS n, anime.name, anime.mal_id, anime.episodes
    FROM animelist r
    JOIN anime ON anime.MAL_ID = r.anime_id
    WHERE
        (r.watched_episodes BETWEEN 0 AND 1) AND (r.rating BETWEEN 1 AND 2) AND anime.episodes >
        GROUP BY anime.name, anime.mal_id, anime.episodes
) h
JOIN reviewlist ON h.MAL_ID = reviewlist.anime_id
GROUP BY h.name, animelist.anime_id, h.n, h.episodes
ORDER BY h.n/COUNT(animelist.rating) DESC
```

Result

		name	Percent of review bomb suspect / all review	Review bomb suspect count	Review count	Episode(s)	Average episode watched	Average watched progress
0		Ex-Arm	21.045504	1554	7384	12	1.536430	12.803584
1		Forest Fairy Five	11.411765	97	850	13	4.352941	33.484163
2		Korogashi-ya no Pun	9.756098	4	41	2	0.512195	25.609756
3		Washimo 7th Season	8.823529	3	34	40	3.588235	8.970588
4		Norimono Atsumare: Hit Song-shuu	8.695652	4	46	6	1.673913	27.898551
...	
6329		Black Lagoon: Roberta's Blood Trail	0.004749	2	42118	5	4.247186	84.943730
6330		Hataraku Saibou!!	0.004705	1	21255	8	2.670148	33.376853
6331		Gekkan Shoujo Nozaki-kun Specials	0.004516	1	22145	6	5.048769	84.146158
6332		Danshi Koukousei no Nichijou Specials	0.004311	1	23194	6	4.979865	82.997758
6333		Durarara!!x2 Shou	0.003125	2	64005	12	8.744770	72.873083

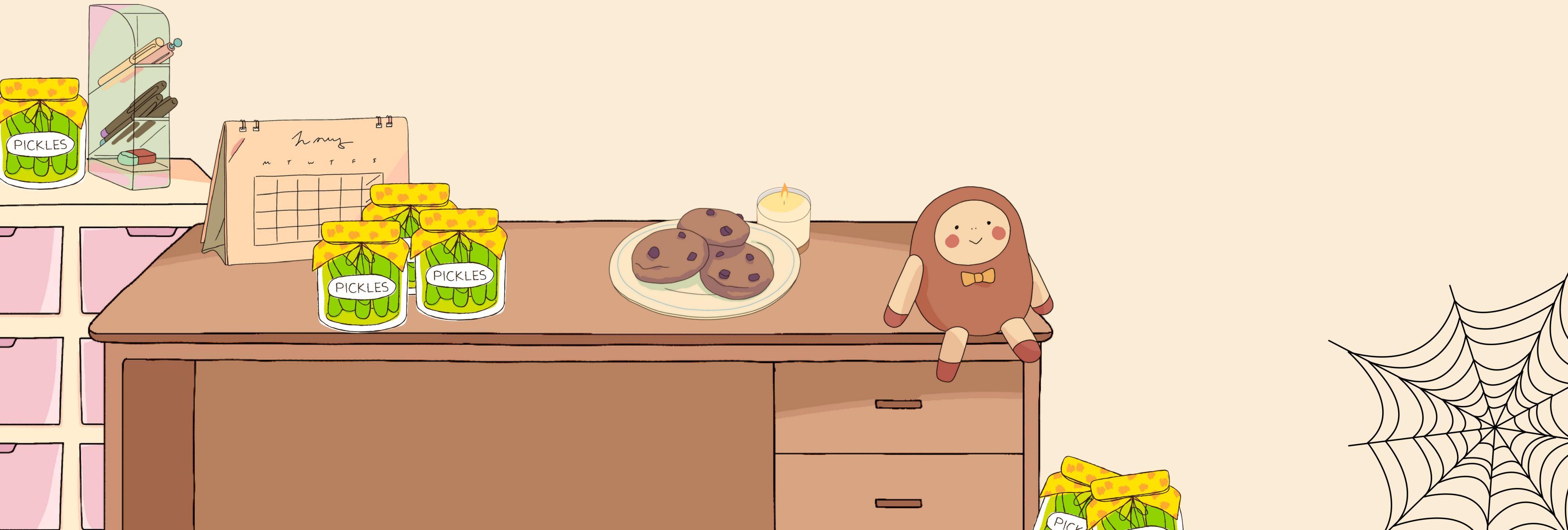
6334 rows × 7 columns





ຕរវជាតា ប្រហេក ANIME

កែគណនោងមាតកក់ស្តុ



ประเภท ANIME ที่คบชอปดอง

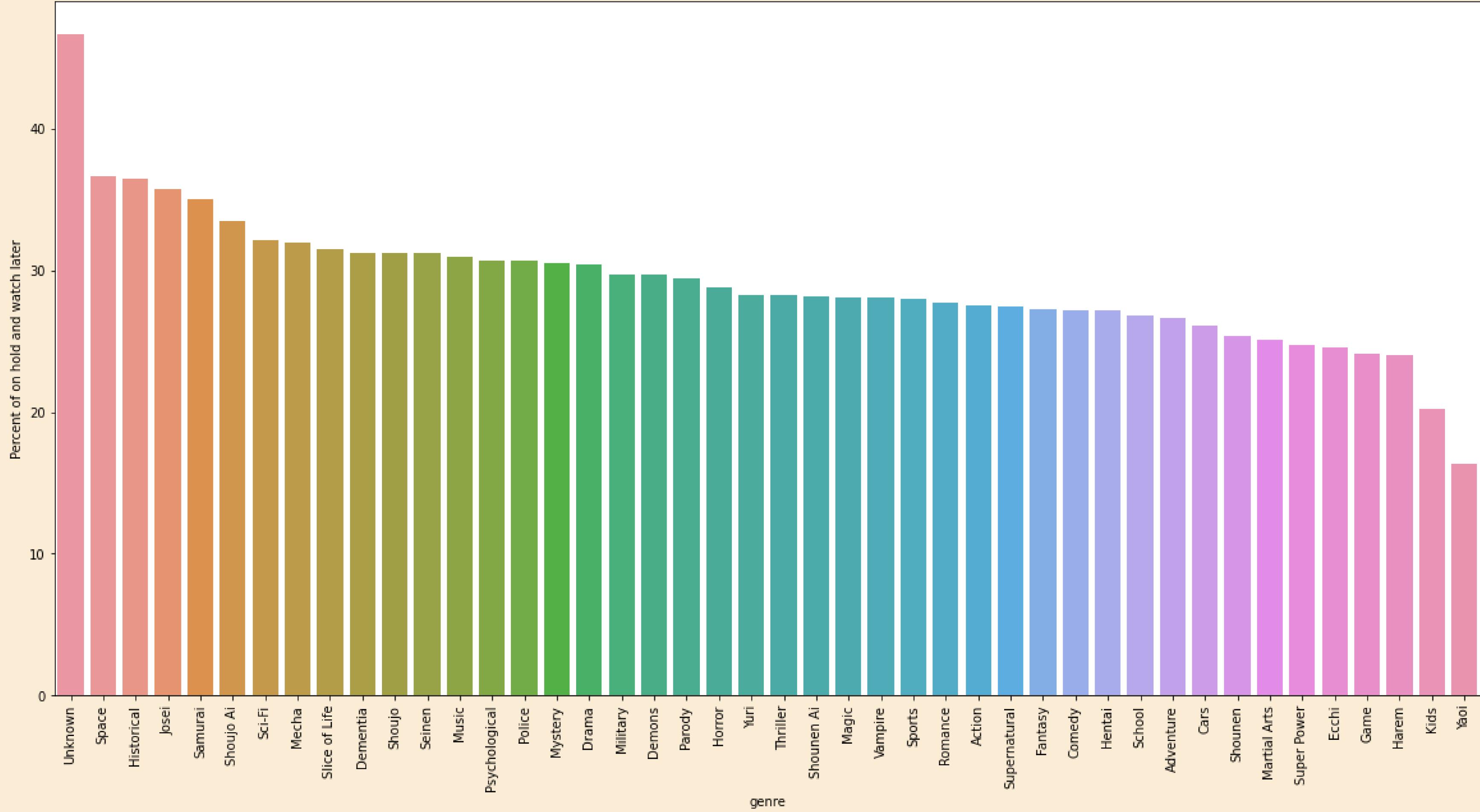
การดอง anime คือการที่มีผู้ชมทำการวางแผนที่จะดูเอาไว้ (Plan to watch) หรือพักการดู anime เรื่องนั้นไป (On hold) โดยการที่รู้ว่า Anime ประเภทไหนถูกดองเป็นจำนวนเปอร์เซ็นต์มาก จะทำให้ผู้ผลิต/ให้บริการสามารถรับรู้และสามารถเลือกประเภท Anime เรื่องต่อไปที่จะทำ เพื่อให้เข้ากับประเภทของคนดูหรือกระแสในตอนนั้นได้

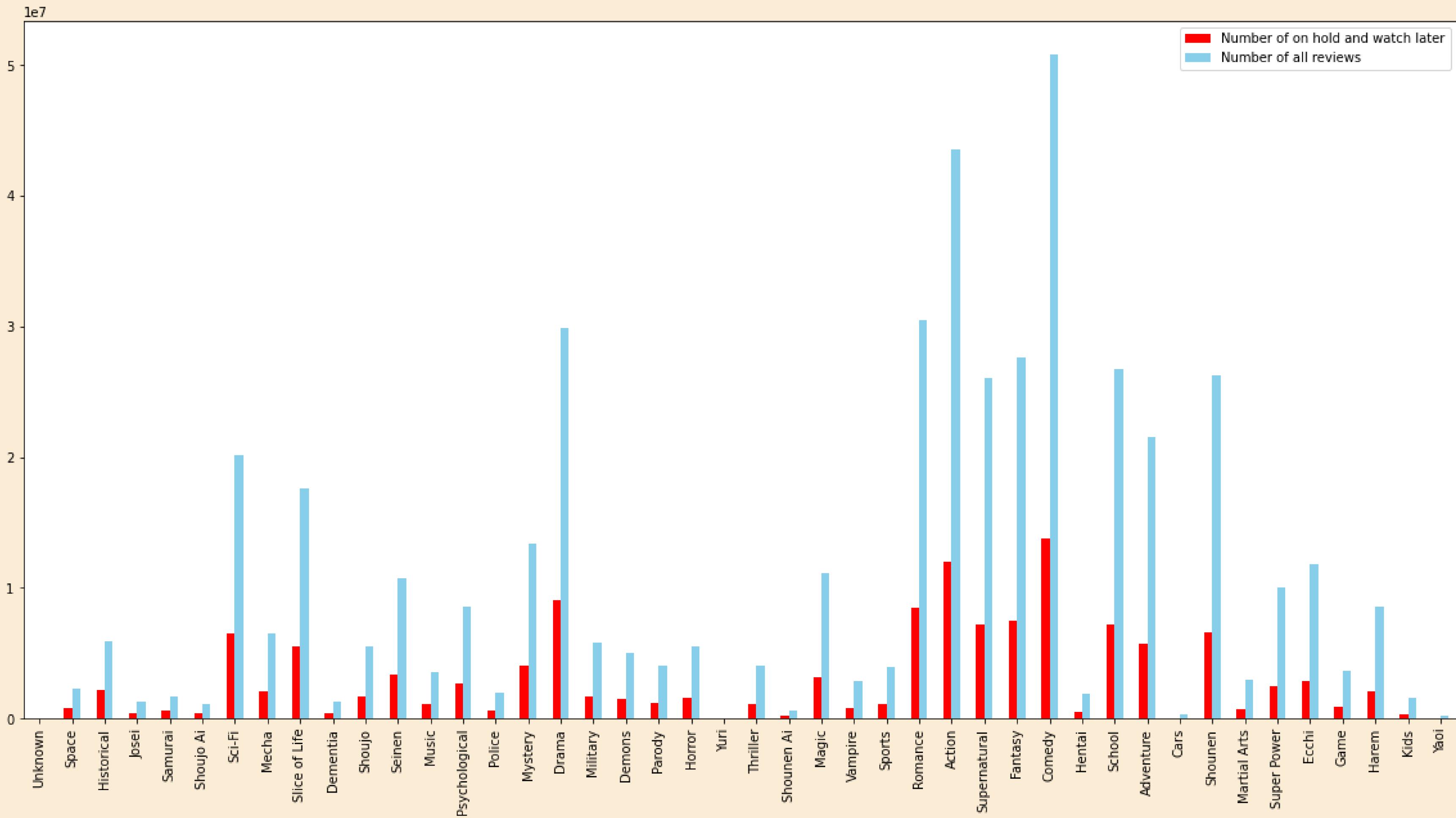
```
SELECT genre, dongCount AS "Number of on hold and watch later", total AS "Number of all reviews",
dongCount*100/total AS "Percent of on hold and watch later" FROM (
    SELECT SUM(sub.c) AS dongCount, anime_genres.genre
    FROM (
        SELECT animelist.name, sub.c
        FROM (
            SELECT COUNT(*) AS c, sub2.anime_id
            FROM (
                SELECT watching_status, reviewList.anime_id
                FROM reviewList
            ) sub2
            WHERE sub2.watching_status = 6 OR sub2.watching_status = 3
            GROUP BY sub2.anime_id
        ) sub
        INNER JOIN animelist ON sub.anime_id = animelist.mal_id
    ) sub
    INNER JOIN anime_genres ON anime_genres.name = sub.name
    GROUP BY anime_genres.genre
) Dong
INNER JOIN (
    SELECT SUM(sub.c) AS total, anime_genres.genre AS g
    FROM (
        SELECT animelist.name, sub.c
        FROM (
            SELECT COUNT(*) AS c, sub2.anime_id
            FROM (
                SELECT watching_status, reviewList.anime_id
                FROM reviewList
            ) sub2
            GROUP BY sub2.anime_id
        ) sub
        INNER JOIN animelist ON sub.anime_id = animelist.mal_id
    ) sub
    INNER JOIN anime_genres ON anime_genres.name = sub.name
    GROUP BY anime_genres.genre
) AS allReview ON allReview.g = Dong.genre
ORDER BY dongCount*100/total DESC
```

CODE

Result

	genre	Number of on hold and watch later	Number of all reviews	Percent of on hold and watch later
0	Unknown	5032	10775	46.700696
1	Space	826909	2254612	36.676333
2	Historical	2160283	5920665	36.487168
3	Josei	451635	1262849	35.763183
4	Samurai	605551	1727350	35.056647
5	Shoujo Ai	373857	1114997	33.529866
6	Sci-Fi	6483322	20199880	32.095844
7	Mecha	2079152	6496561	32.003886
8	Slice of Life	5533347	17569281	31.494442
9	Dementia	420057	1342973	31.278142
10	Shoujo	1715688	5487048	31.267960
11	Seinen	3349540	10729322	31.218562
12	Music	1098551	3552435	30.923887
13	Psychological	2637506	8590930	30.701053
14	Police	620600	2021582	30.698730
15	Mystery	4082338	13375774	30.520387
16	Drama	9077794	29846140	30.415303
17	Military	1743946	5868335	29.717901
18	Demons	1508565	5080882	29.691006
19	Parody	1190820	4047520	29.420979
20	Horror	1584685	5496595	28.830303
21	Yuri	20723	73249	28.291171
22	Thriller	1147126	4061201	28.245979
23	Shounen Ai	185706	658382	28.206421
24	Magic	3130417	11147444	28.081926
25	Vampire	810911	2889810	28.061049
26	Sports	1104082	3943300	27.998935
27	Romance	8433555	30421107	27.722709
28	Action	11966251	43519118	27.496538





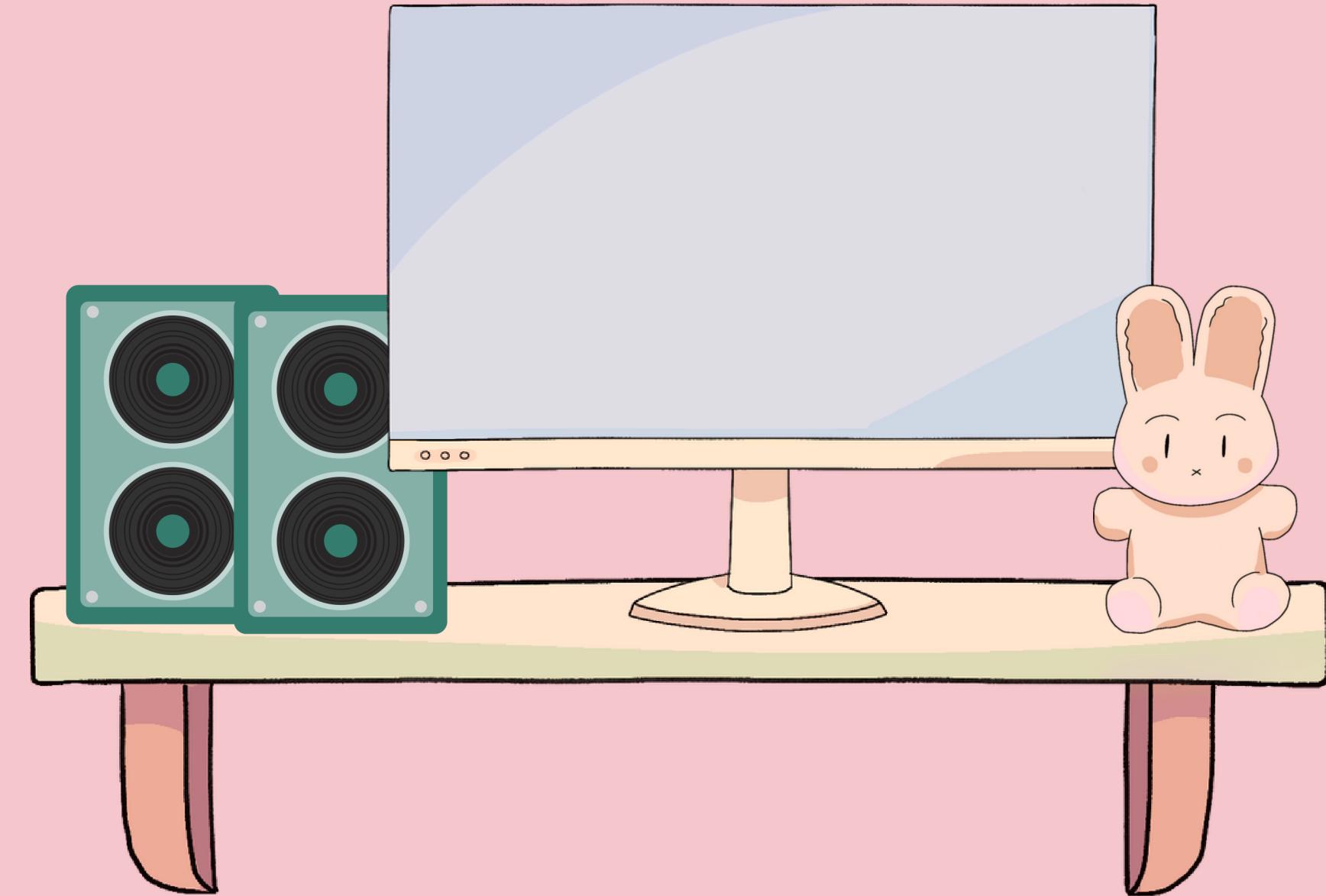


SPARK

a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.



เรียงอันดับ 10 อันเบะที่มี ตอนเยอะสุด



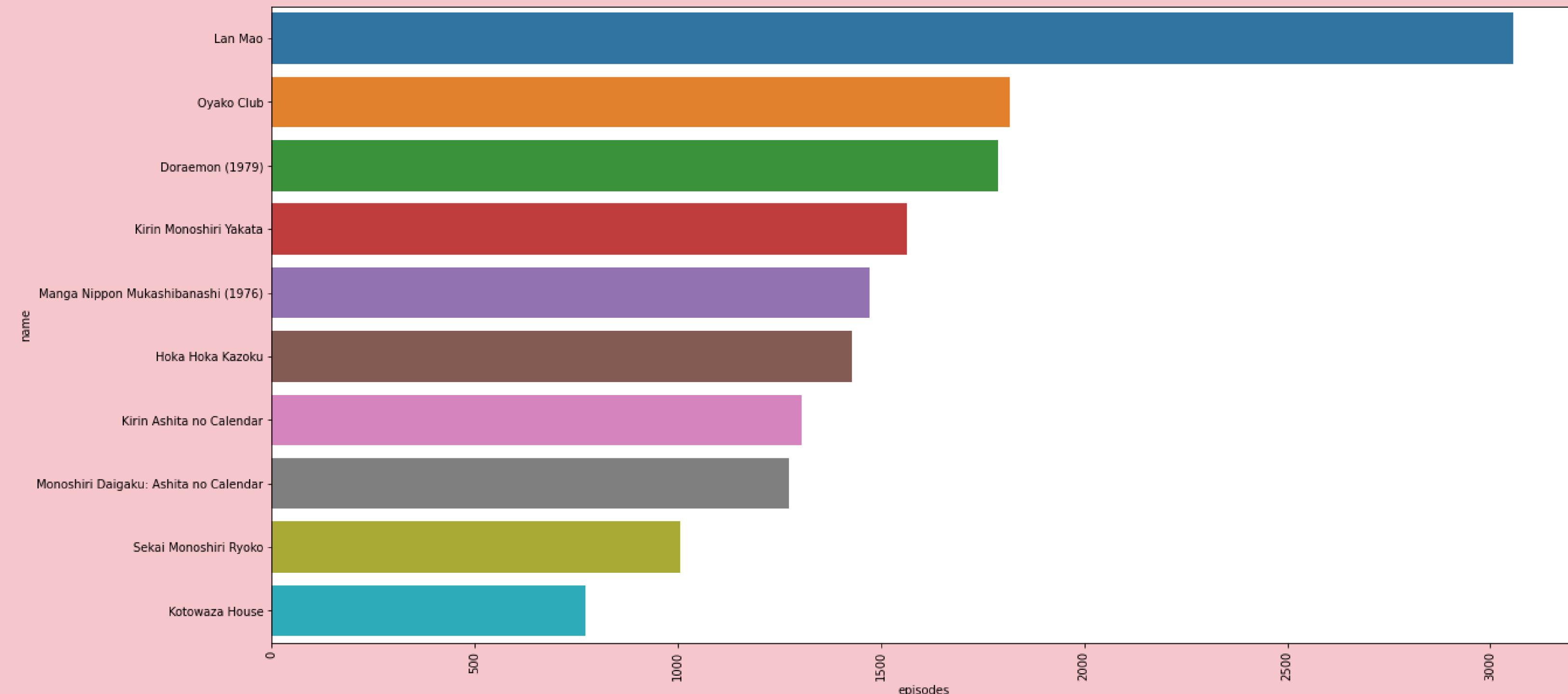
เรียงอันดับ 10 อันเมะที่มีตอน เยอะสุด

10 อันดับอันเมะที่มีตอนเยอะสุด ทำให้เห็นความนิยมและความต่อเนื่องของทั้งตัวมังงะและอันเมะเรื่องนี้ๆ ทั้งตัวอาจารย์ผู้เขียนมังงะ และสตูดิโอผู้รับผิดชอบ แสดงให้เห็นว่าผู้ชมยังคงติดตามอยู่เสมอ

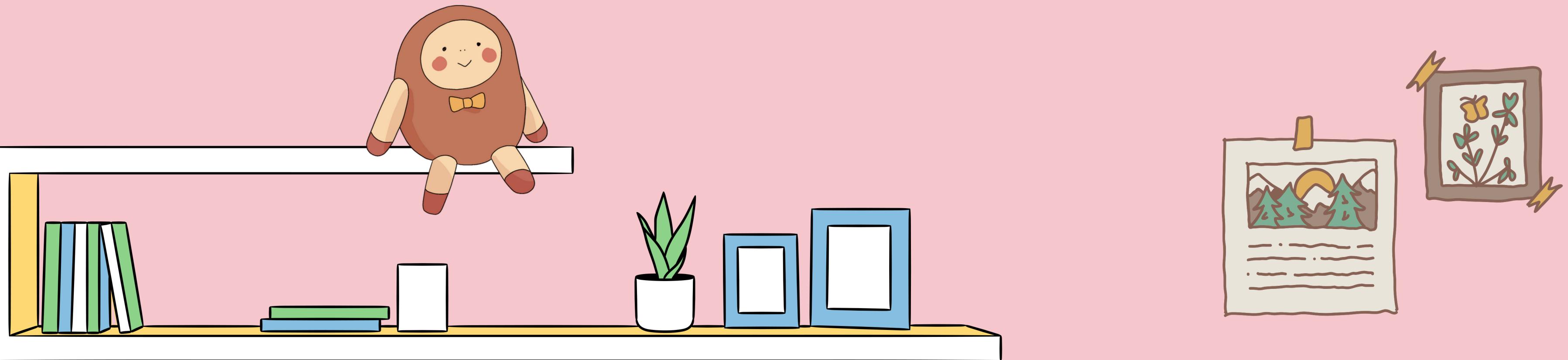
CODE

```
SELECT name, episodes  
FROM anime  
ORDER BY episodes DESC  
LIMIT 10;
```

Result



ช่วงเวลาที่เมื่อ ANIME ออก แล้วมีคนดูมากสุด 10 อันดับ



ช่วงเวลาที่เมื่อ ANIME ออกแล้วมี คนดูมากสุด 10 อันดับ

ช่วงเวลาที่ Anime ออกแล้วมีคนดูมากสุด แสดงให้เห็นถึงความเป็นที่สนใจในอนิเมะนั้นๆ ทั้งทางด้านกระแสและการตั้งตารอของผู้ชม ซึ่งโดยปกติแล้วจะปล่อยวายสปีด่าหลาบตอน เพื่อให้ผู้ผลิตได้ผลตอบแทนกัน และประกอบกระแสของเรื่องนั้นๆ

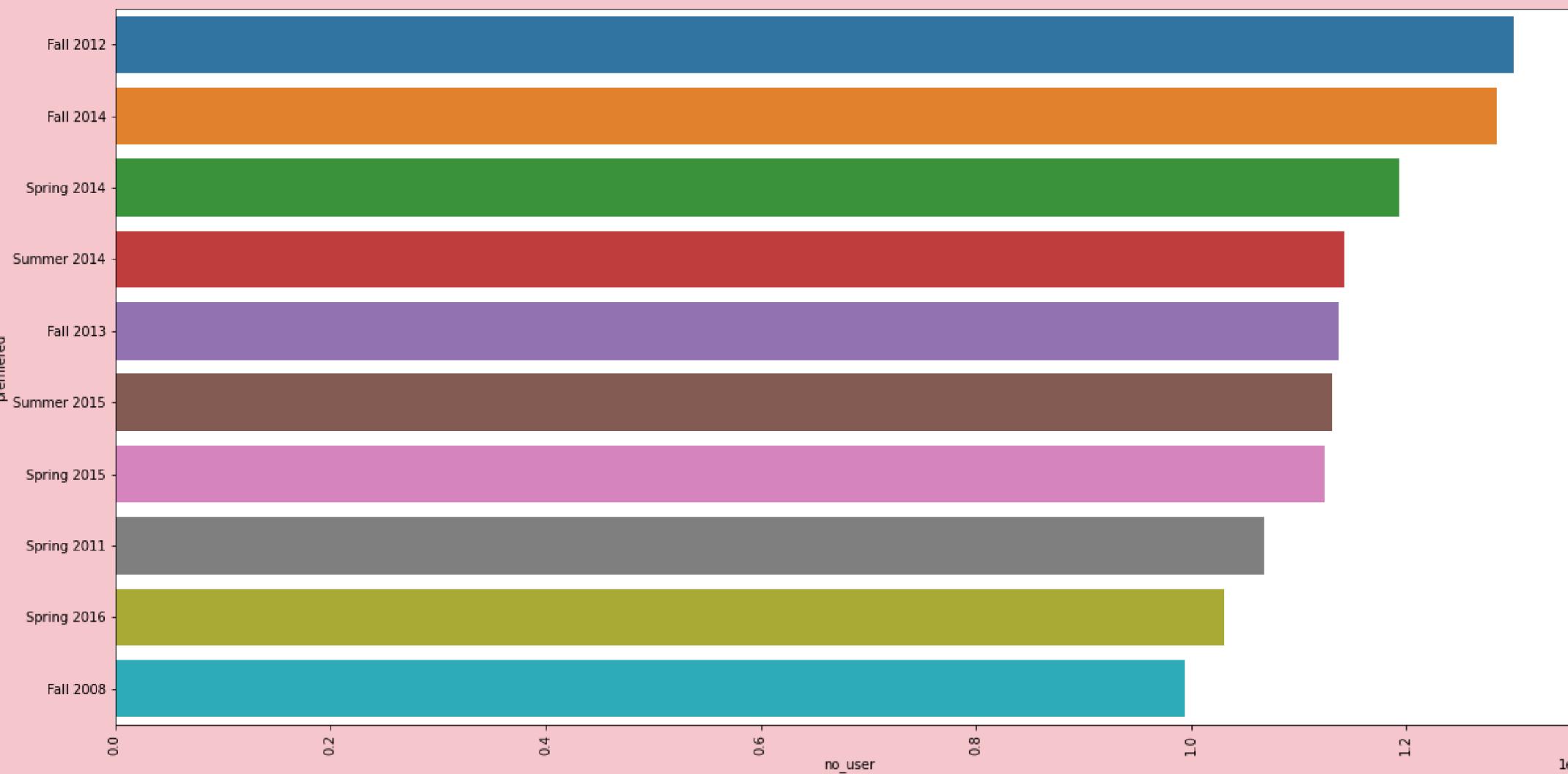
CODE

```
scala> val result = sqlContext.sql("SELECT count(reviewlist.user_id) as no_user, animelist.premiered FROM reviewlist LEFT JOIN animelist ON reviewlist.anime_id = animelist.mal_id WHERE reviewlist.watching_status = 1 OR reviewlist.watching_status = 2 AND animelist.premiered <> 'Unknown' GROUP BY animelist.premiered ORDER BY count(reviewlist.user_id) DESC LIMIT 10")
```

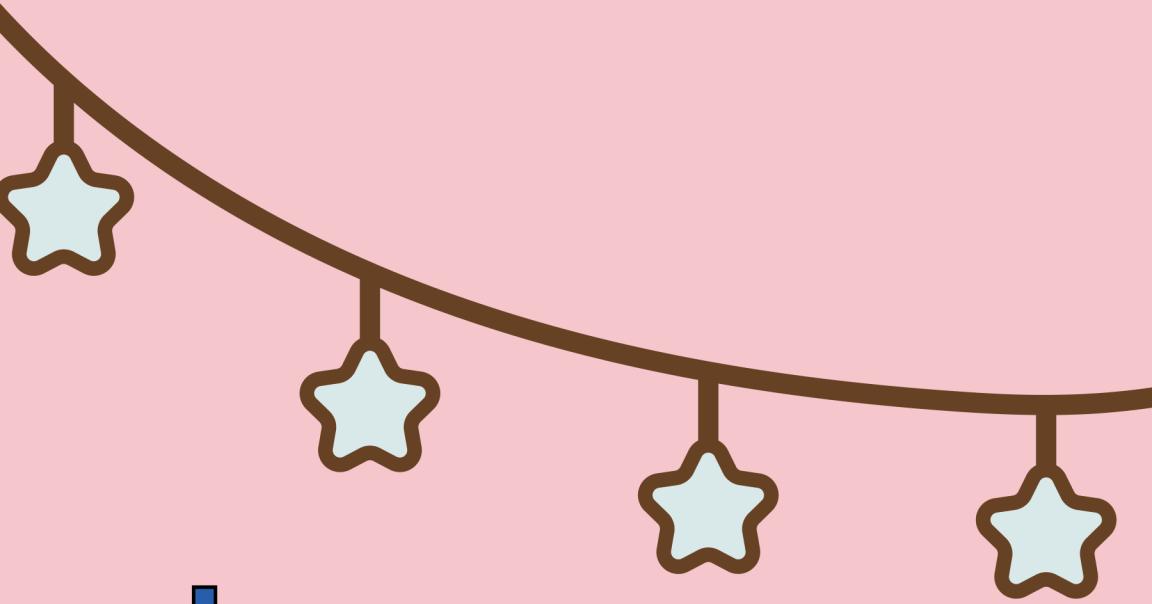
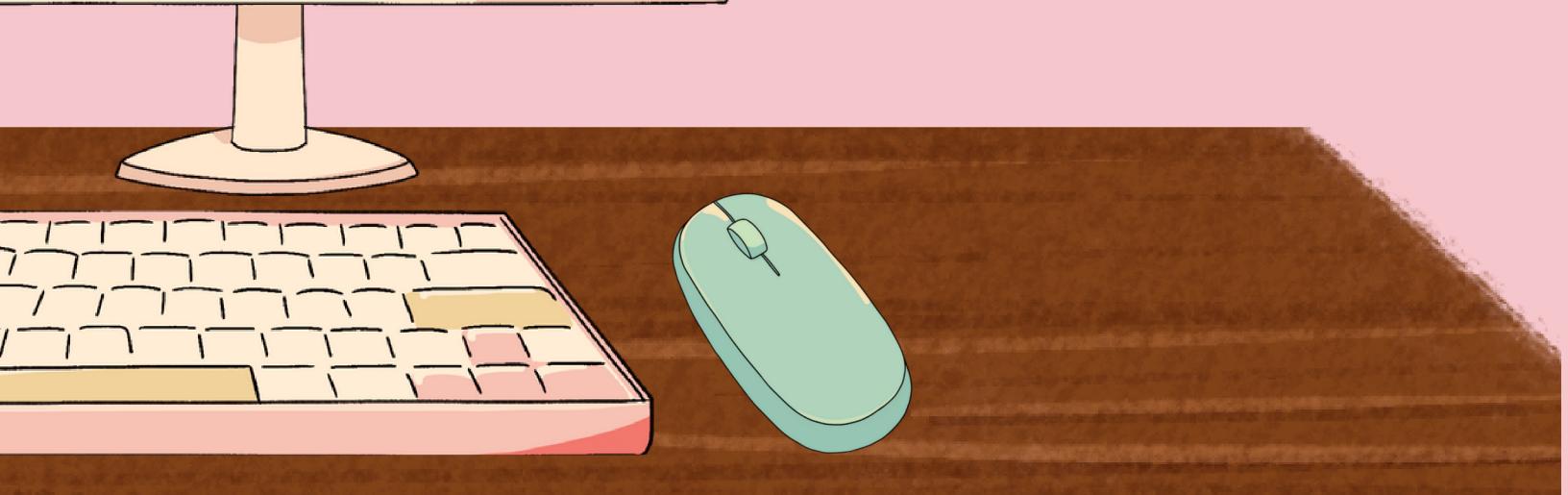
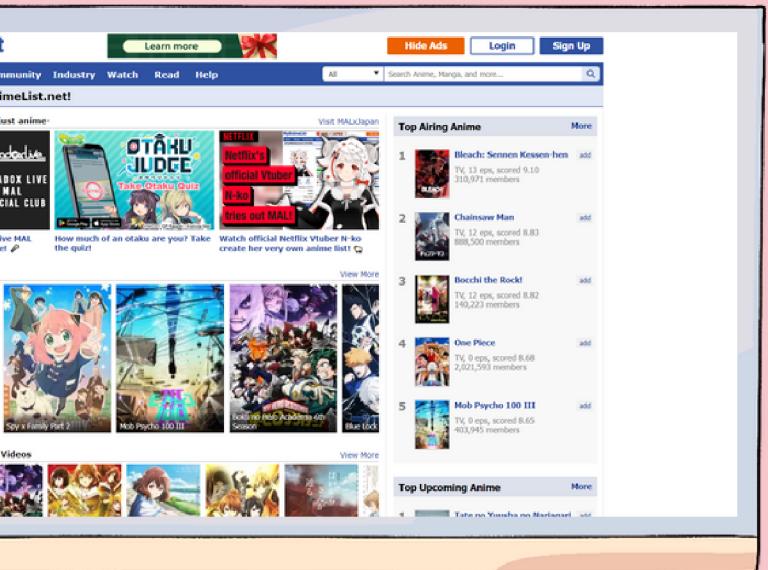
```
SELECT count(animelist.user_id) as no_user,  
       anime.premiered  
  FROM animelist  
LEFT JOIN anime  
    ON animelist.anime_id = anime.mal_id  
 WHERE animelist.watching_status = 1  
   OR animelist.watching_status = 2  
   AND anime.premiered <> 'Unknown'  
 GROUP BY anime.premiered  
 ORDER BY count(animelist.user_id) DESC  
LIMIT 10;
```

Result

no_user	premiered
1299557	Fall 2012
1283442	Fall 2014
1194106	Spring 2014
1142776	Summer 2014
1136863	Fall 2013
1130878	Summer 2015
1123861	Spring 2015
1067306	Spring 2011
1031537	Spring 2016
994237	Fall 2008



ຕរចកា REVIEW នៃ គុណភាព



ตรวจหา REVIEW ที่มีคุณภาพ

การตรวจหา review ที่มีคุณภาพ เป็นสิ่งที่ต่อเนื่องมาจากเรื่องของ Review Bomb การตรวจหา Review ที่มีคุณภาพคือการหา Review ที่วิจารย์อนิเมะเรื่องนั้นๆ อย่างตรงไปตรงมา พูดถึงตัวคุณภาพของอนิเมะ วิเคราะห์ใบແง່ນຸມຕ່າງໆ ไม่ใช่การຈົງໃຈกำລາຍຊື່ເສີຍ

CODE

```
scala> val result = sqlContext.sql("SELECT ep.user AS user, ep.name AS aniname, ep.rating, ROUND((ep.watched_episodes*100)/ep.epi, 2) AS watchComp FROM (SELECT reviewlist.user_id  
as user, reviewlist.rating as rating,reviewlist.watched_episodes as watched_episodes, animelist.mal_id as id,animelist.name as name, animelist.episodes as epi FROM animelist INNER  
JOIN reviewlist ON reviewlist.anime_id = animelist.mal_id) AS ep WHERE ep.watched_episodes > 1 AND ROUND((ep.watched_episodes*100)/ep.epi, 2) < 101 AND ROUND((ep.watched_episodes  
*100)/ep.epi, 2) > 85 ORDER BY ep.user ASC")
```

```
SELECT ep.user AS user, ep.name AS aniname, ep.rating, ROUND((ep.watched_episodes*100)/ep.epi, 2) AS watchComp  
FROM (SELECT reviewlist.user_id as user,  
reviewlist.rating as rating,  
reviewlist.watched_episodes as watched_episodes,  
animelist.mal_id as id,  
animelist.name as name,  
animelist.episodes as epi  
FROM animelist  
INNER JOIN reviewlist ON reviewlist.anime_id = animelist.mal_id) AS ep  
WHERE ep.watched_episodes > 1  
AND ROUND((ep.watched_episodes*100)/ep.epi, 2) < 101  
AND ROUND((ep.watched_episodes*100)/ep.epi, 2) > 85  
ORDER BY ep.user ASC
```

Result

		user		aniname	rating	watchComp
0	0			Tokimeki Tonight	9	85.29
1	0			Black Cat (TV)	6	100.00
2	0			Erementar Gerad	7	100.00
3	0			Jungle no Ouja Taa-chan	9	100.00
4	0			Fate/stay night	9	100.00
...
51692970	353404			Yuu☆Yuu☆Hakusho	9	100.00
51692971	353404			Gravitation: Lyrics of Love	7	100.00
51692972	353404			Dragon Ball	9	100.00
51692973	353404			Grappler Baki (TV)	9	100.00
51692974	353404		Grappler Baki: Saidai Tournament-hen		8	100.00

51692975 rows × 4 columns

user	aniname	rating	watchComp
0	Shoukoujo Sara	7	100.0
0	Legend of Duo	6	100.0
0	Shingetsutan Tsuk...	7	100.0
0	Fullmetal Alchemist	9	100.0
0	Kaiketsu Zorro	7	100.0
0	Mushishi	0	100.0
0	Samurai Deeper Kyou	8	100.0
0	Byousoku 5 Centim...	6	100.0
0	Fate/stay night	9	100.0
0	Igano Kabamaru	9	100.0
0	Jungle no Ouja Ta...	9	100.0
0	Lovely★Complex	8	100.0
0	Muka Muka Paradise	6	100.0
0	Naruto	0	100.0
0	Robin Hood no Dai...	7	100.0
0	Tokimeki Tonight	9	85.29
0	Black Cat (TV)	6	100.0
0	Daisougen no Chii...	6	100.0
0	Erementar Gerad	7	100.0
0	Ghost Hunt	10	100.0

only showing top 20 rows

THANK YOU FOR LISTENING!



ສມາຊິກ

นางสาว ณิชกานต์ สุขุมจิตพิทักษ์
นายนริชญ์ อยู่บัว
นายมหามุಥ
นายวีรภัทร
นายศุภกร

รหัสนักศึกษา 62010299
รหัสนักศึกษา 62010465
รหัสนักศึกษา 63010782
รหัสนักศึกษา 63010895
รหัสนักศึกษา 63010921