Video Game Recommendation System and Analysis

SB Capstone 3

Trenten Beram

START MENU

Main Objectives

1. Web Scrape Data from MetaCritic All Time Game List
2. Clean the Data
3. Perform EDA
4. Build Content or Collaborative Recommendation Model
5. Improvements

1.Web Scraping

MetaCritic.com

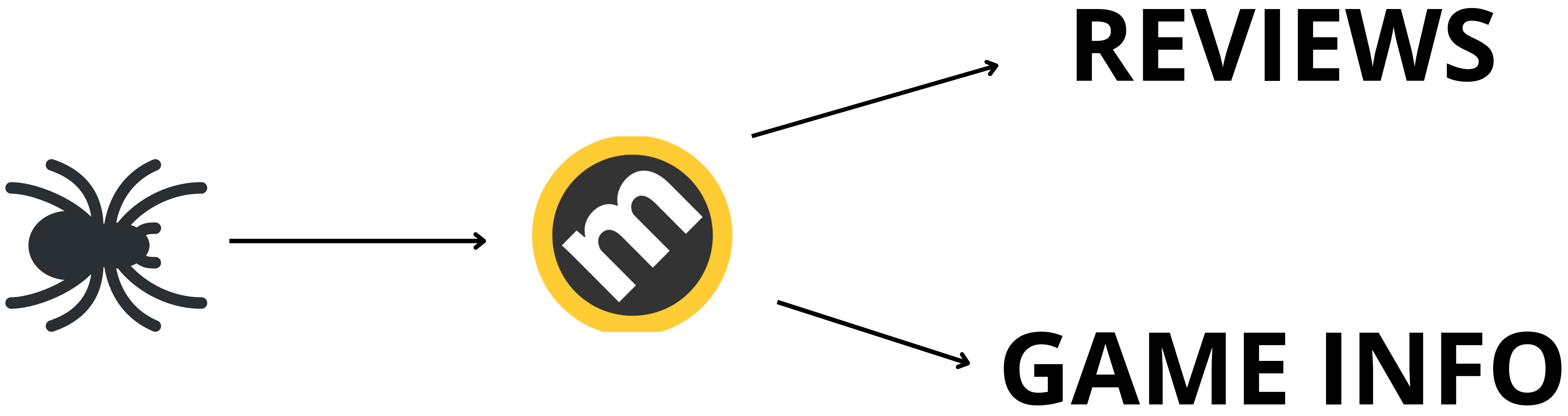
Web Scraping

Libraries

requests - retrieves HTML from each link

BeautifulSoup - Parses the HTML for info

pandas - stores the info obtained in a DataFrame

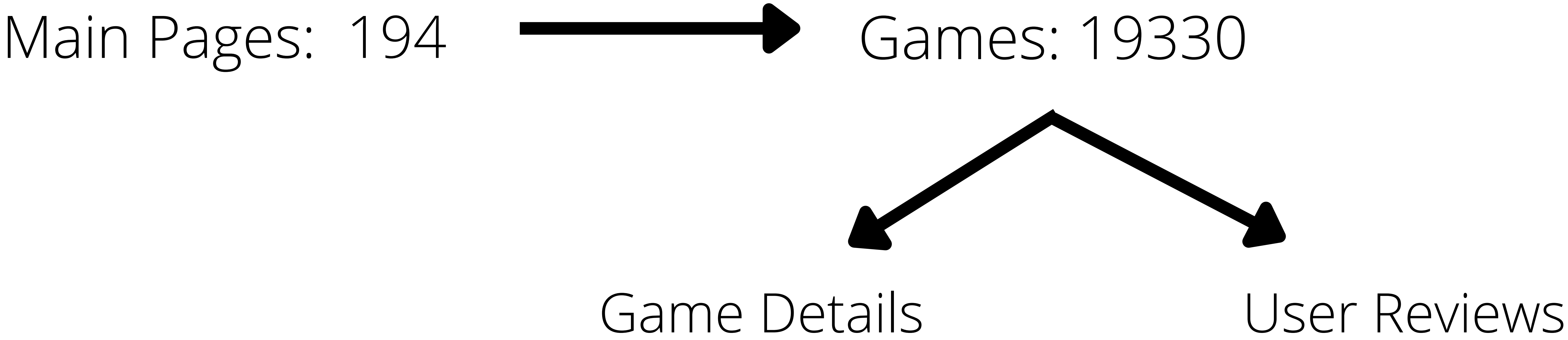


REVIEWS

GAME INFO

Web Scraping

Obtaining Links



Main Pages: 194

Games: 19330

Game Details

User Reviews

Objective: Get the Details and User Review links for each game. The 194 main pages were parsed to get the links to each game. Then the strings "details" or "user-reviews" were concatenated to the trunk of each game link

Total Number of Links Parsed: 194 + 19,330x2 = 38,854

Web Scraping

Game Set

- title
- release_date
- genres
- platform
- developer
- esrb_rating (e.g. E)
- ESRBs (e.g. Violence)
- metascore
- userscores
- critic_reviews (amount)
- user_reviews (amount)
- num_players (e.g. single-player)
- summary

Reviews Set

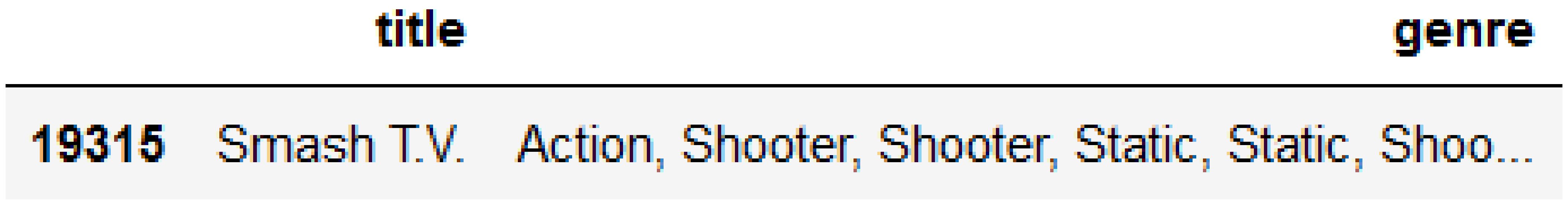
- User ids
- game title
- rating
- review

2.Data Cleaning

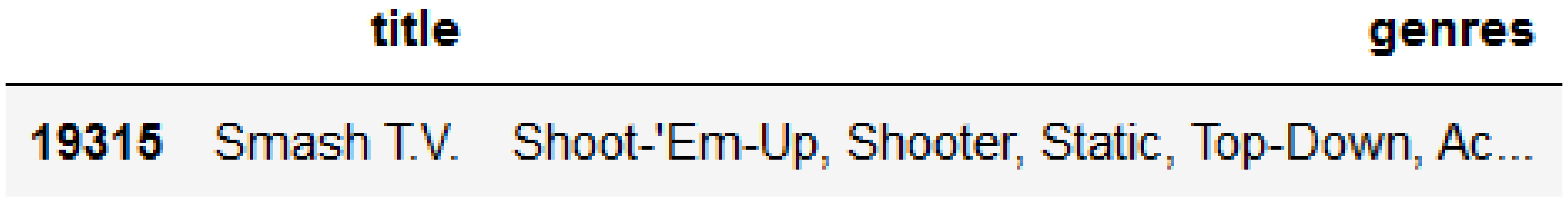
Data Cleaning

Repeat Genres

Duplicate genres were removed from the genre column in the Game Set



	title	genre
19315	Smash T.V.	Action, Shooter, Shooter, Static, Static, Shoo...

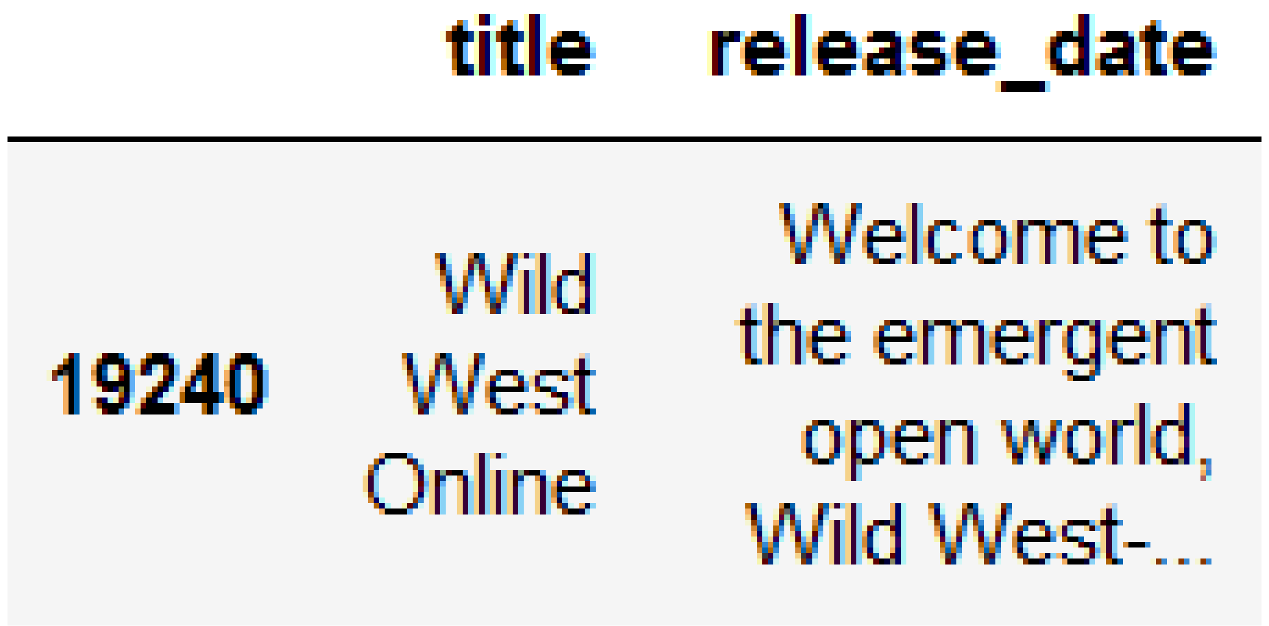


	title	genres
19315	Smash T.V.	Shoot-'Em-Up, Shooter, Static, Top-Down, Ac...

Data Cleaning

Release Date to DateTime

The Release Date column contained incompatible date formats.



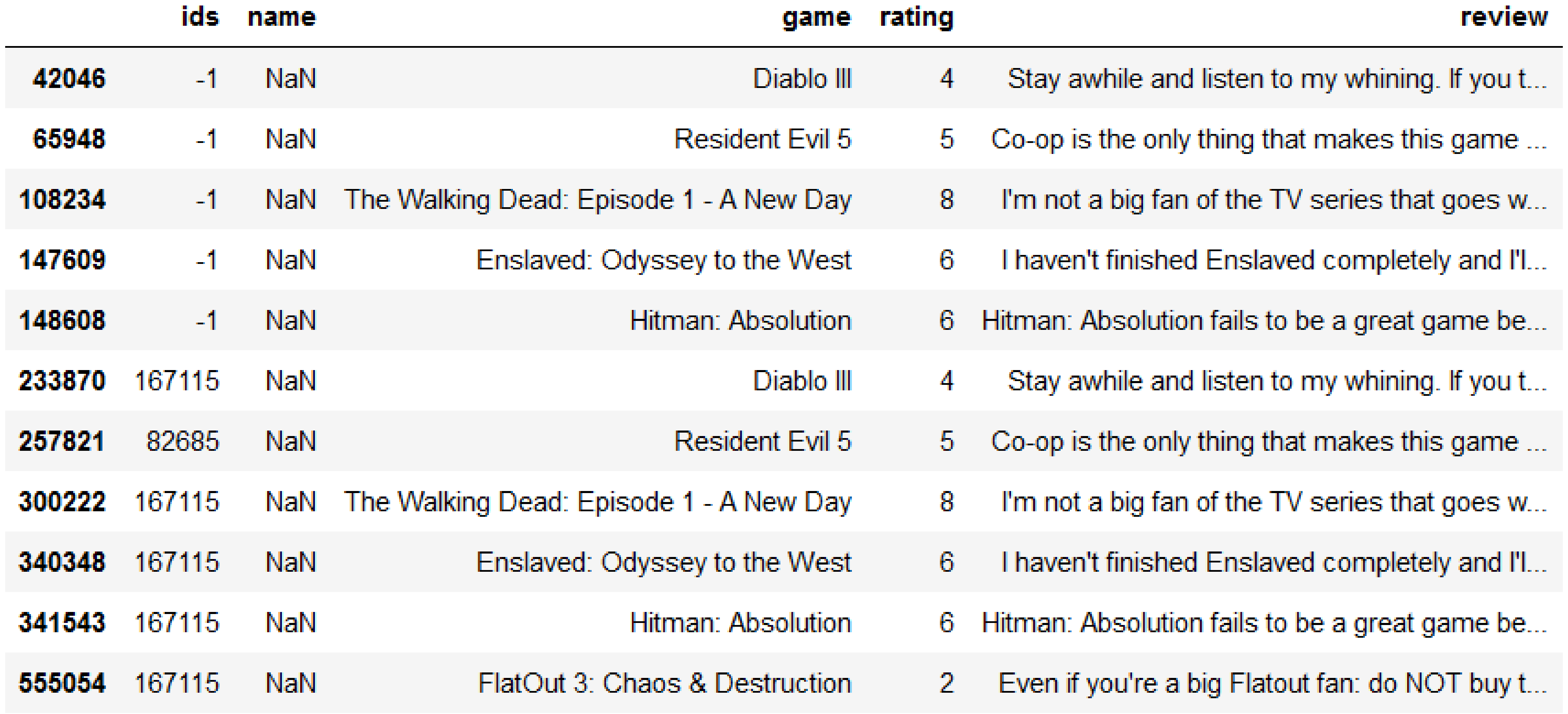
	title	release_date
19240	Wild West Online	Welcome to the emergent open world, Wild West-...

This might have been an issue if there were many rows like this. Thankfully there were only 11. so the release dates were manually searched and inserted. After the dates were inserted the release date column was converted to a datetime type. Some of these rows also had summaries in the release date column. So those summaries were moved over to the summary column.

Data Cleaning

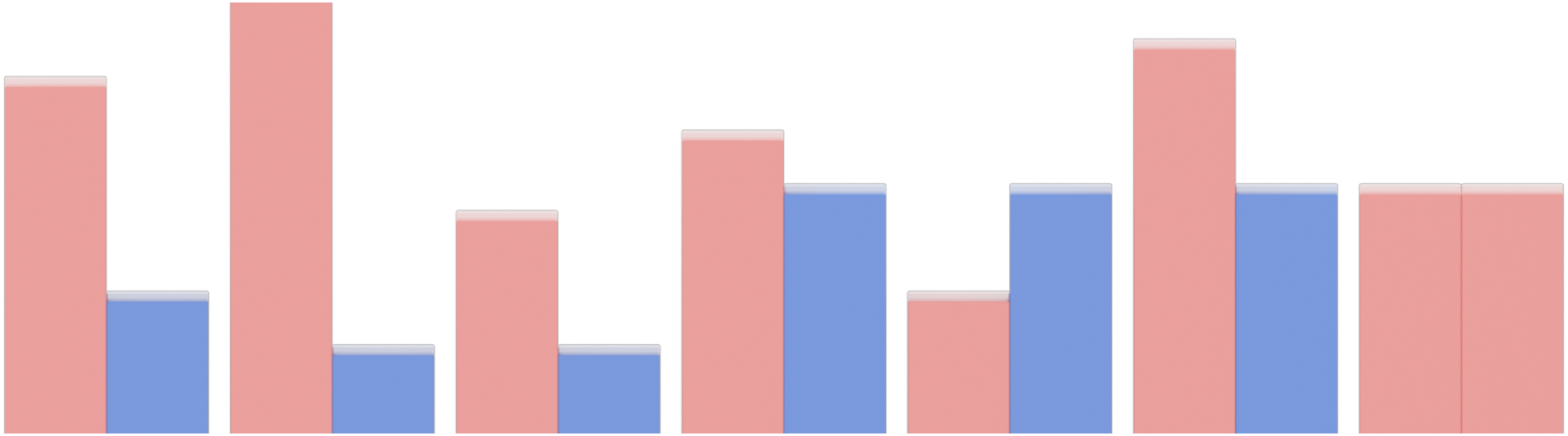
Negative IDs in Review Set

For an unknown reason there were duplicate reviews that did not have duplicate user ids. Interestingly, Pandas somehow was able to distinguish between the Nan values. These duplicate reviews were then dropped.

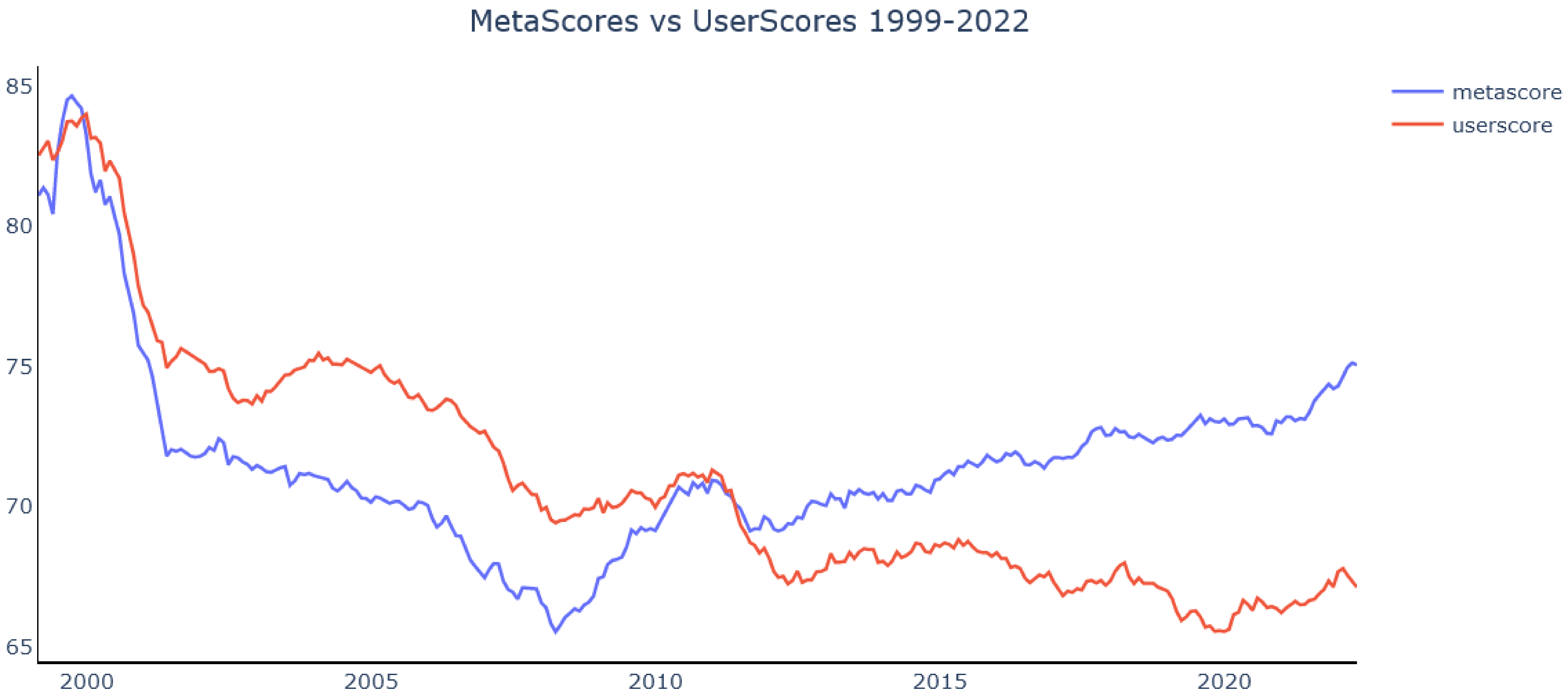


	ids	name	game	rating	review
42046	-1	NaN	Diablo III	4	Stay awhile and listen to my whining. If you t...
65948	-1	NaN	Resident Evil 5	5	Co-op is the only thing that makes this game ...
108234	-1	NaN	The Walking Dead: Episode 1 - A New Day	8	I'm not a big fan of the TV series that goes w...
147609	-1	NaN	Enslaved: Odyssey to the West	6	I haven't finished Enslaved completely and I'l...
148608	-1	NaN	Hitman: Absolution	6	Hitman: Absolution fails to be a great game be...
233870	167115	NaN	Diablo III	4	Stay awhile and listen to my whining. If you t...
257821	82685	NaN	Resident Evil 5	5	Co-op is the only thing that makes this game ...
300222	167115	NaN	The Walking Dead: Episode 1 - A New Day	8	I'm not a big fan of the TV series that goes w...
340348	167115	NaN	Enslaved: Odyssey to the West	6	I haven't finished Enslaved completely and I'l...
341543	167115	NaN	Hitman: Absolution	6	Hitman: Absolution fails to be a great game be...
555054	167115	NaN	FlatOut 3: Chaos & Destruction	2	Even if you're a big Flatout fan: do NOT buy t...

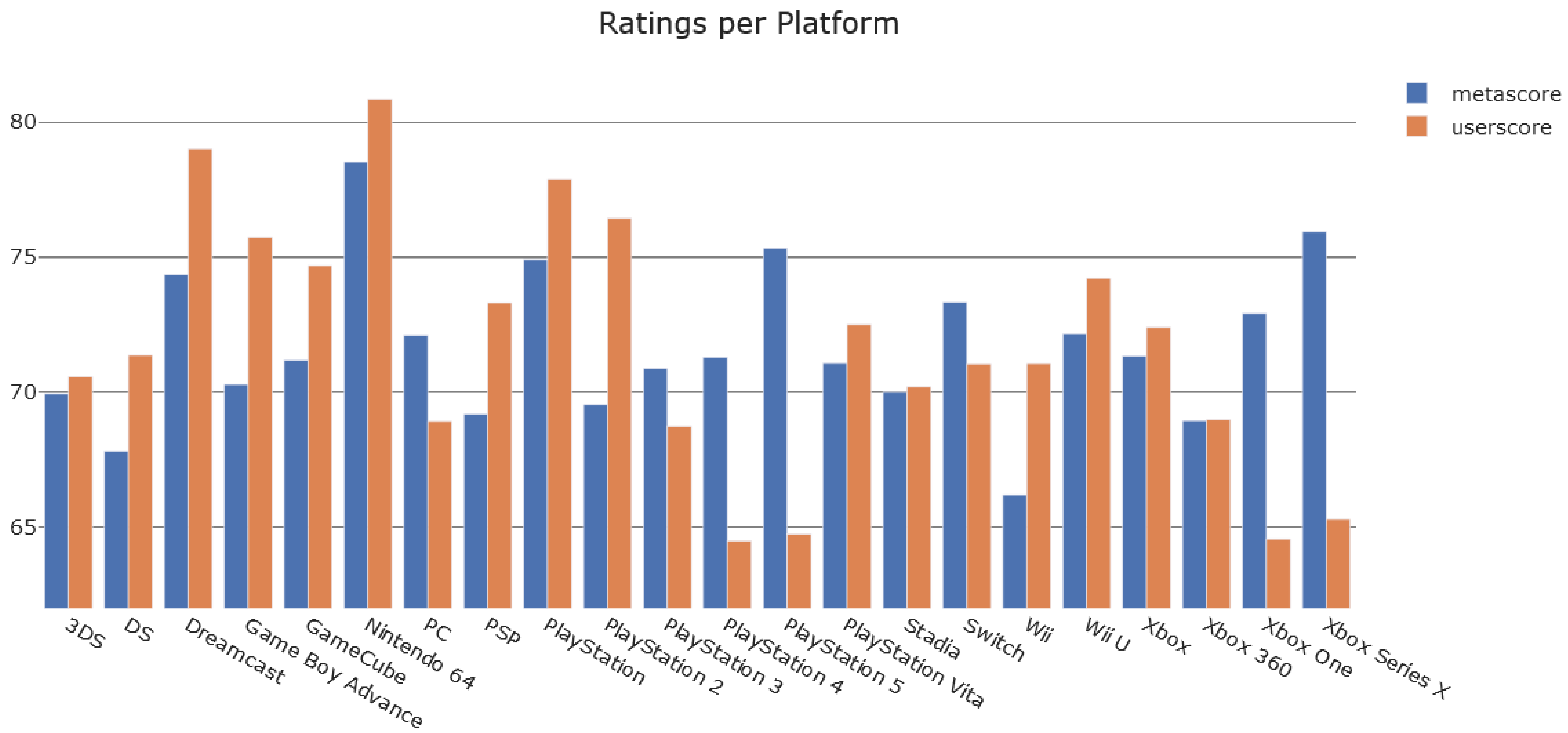
3.EDA
(Interesting Visualizations)



EDA

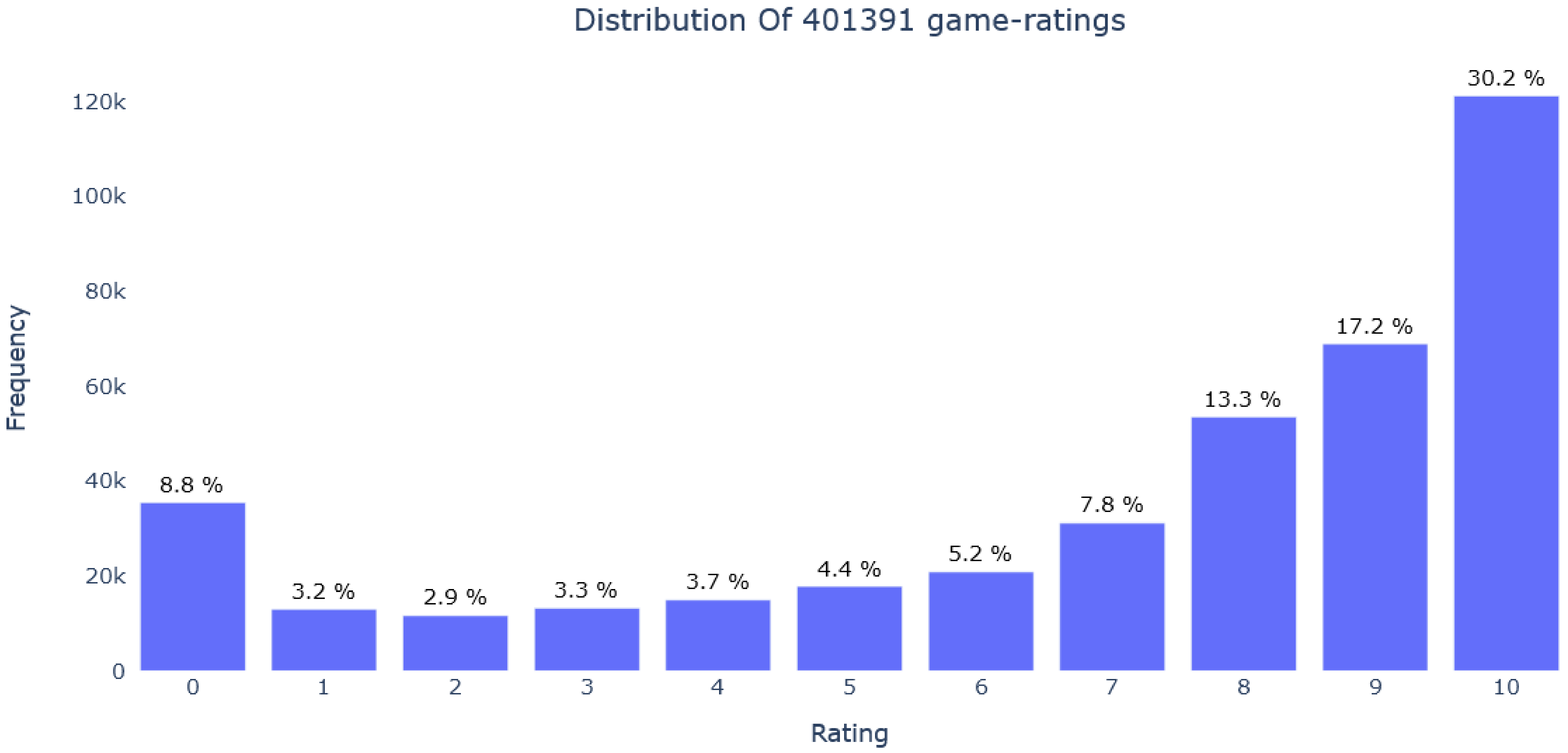

MetaScores vs UserScores 1999-2022
85
80
75
70
65
2000
2005
2010
2015
2020
metascore
userscore

EDA



Ratings per Platform
metascore
userscore
80
75
70
65
3DS
DS
Dreamcast
Game Boy Advance
GameCube
Nintendo 64
PC
PSP
PlayStation
PlayStation 2
PlayStation 3
PlayStation 4
PlayStation 5
PlayStation Vita
Stadia
Switch
Wii
Wii U
Xbox
Xbox 360
Xbox One
Xbox Series X

EDA


Distribution Of 401391 game-ratings
Frequency
120k
100k
80k
60k
40k
20k
0
8.8 %
3.2 %
2.9 %
3.3 %
3.7 %
4.4 %
5.2 %
7.8 %
13.3 %
17.2 %
30.2 %
0
1
2
3
4
5
6
7
8
9
10
Rating

4.Modeling

Modeling

Content Filtering: Genre Similarities

Goal: Get titles that are most similar to other titles based off genres

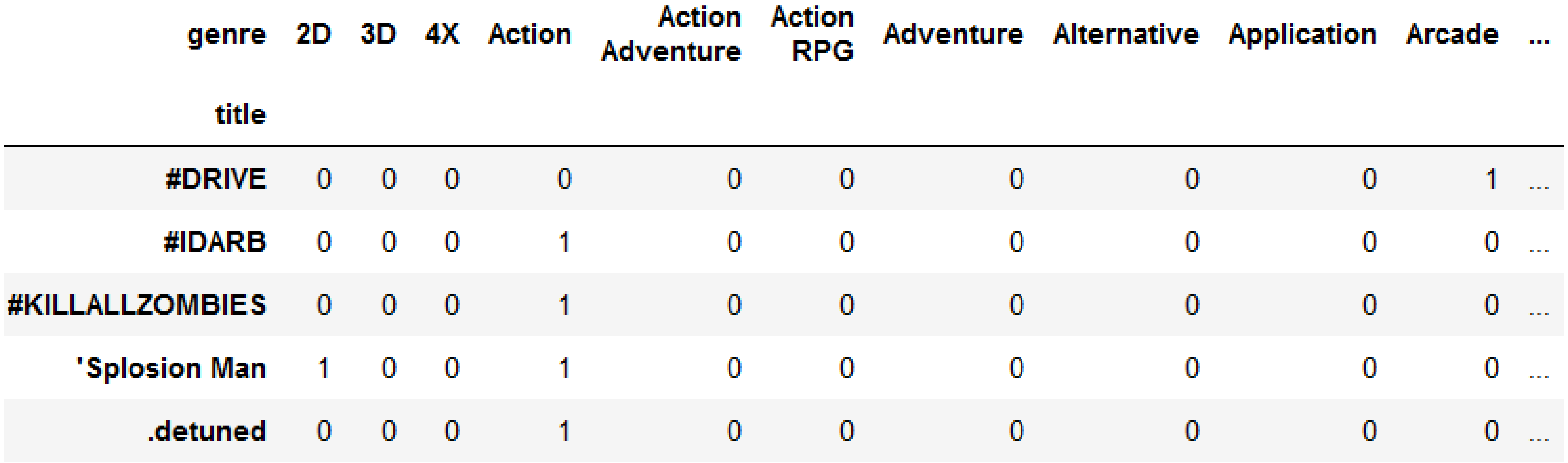
Steps

1. Create Dummy Matrix from genres
2. Get Jaccard similarity coefficients matrix
3. Create function to get titles with similar genres

Modeling

Content Filtering: Genre Similarities

Dummy matrix for all the unique genres

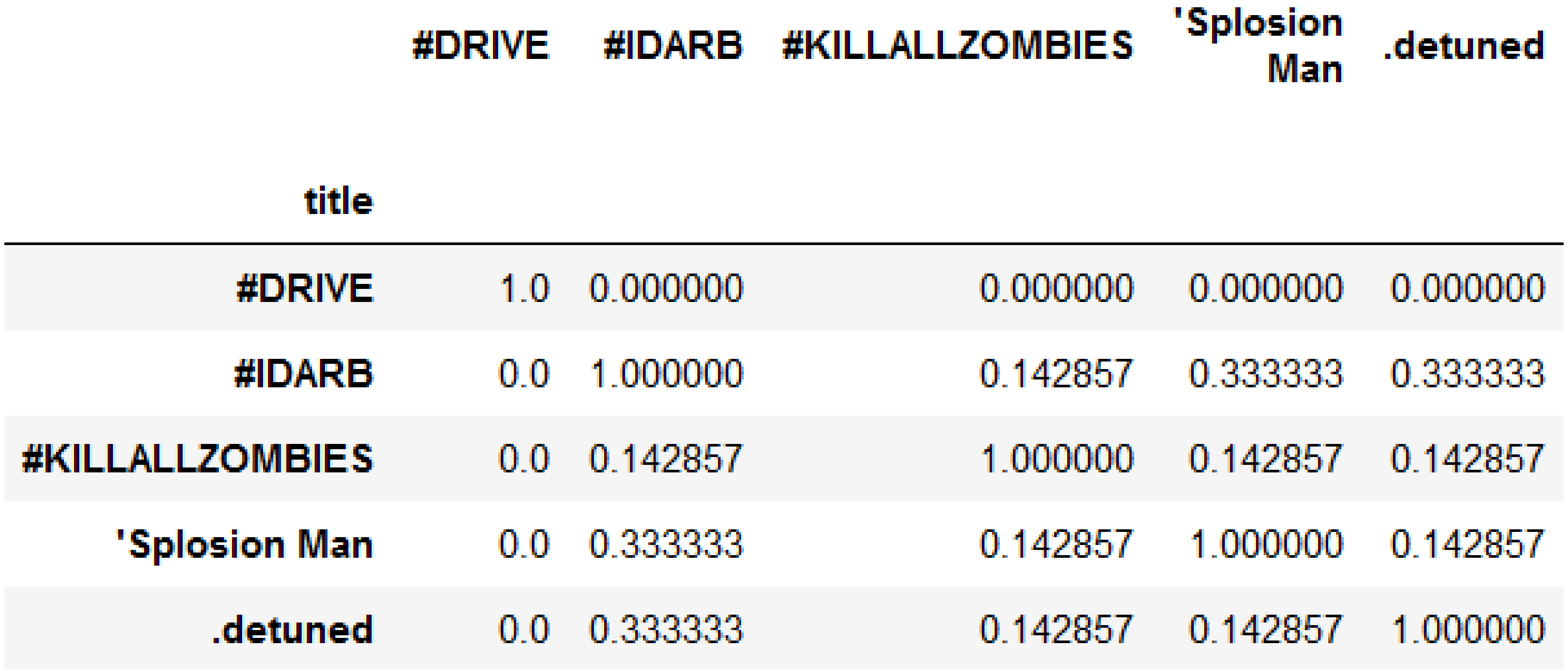


genre	2D	3D	4X	Action	Action Adventure	Action RPG	Adventure	Alternative	Application	Arcade	...
title											
#DRIVE	0	0	0	0	0	0	0	0	0	1	...
#IDARB	0	0	0	1	0	0	0	0	0	0	...
#KILLALLZOMBIES	0	0	0	1	0	0	0	0	0	0	...
'Splosion Man	1	0	0	1	0	0	0	0	0	0	...
.detuned	0	0	0	1	0	0	0	0	0	0	...

Modeling

Content Filtering: Genre Similarities

Calculated Jaccard Similarity Coefficients

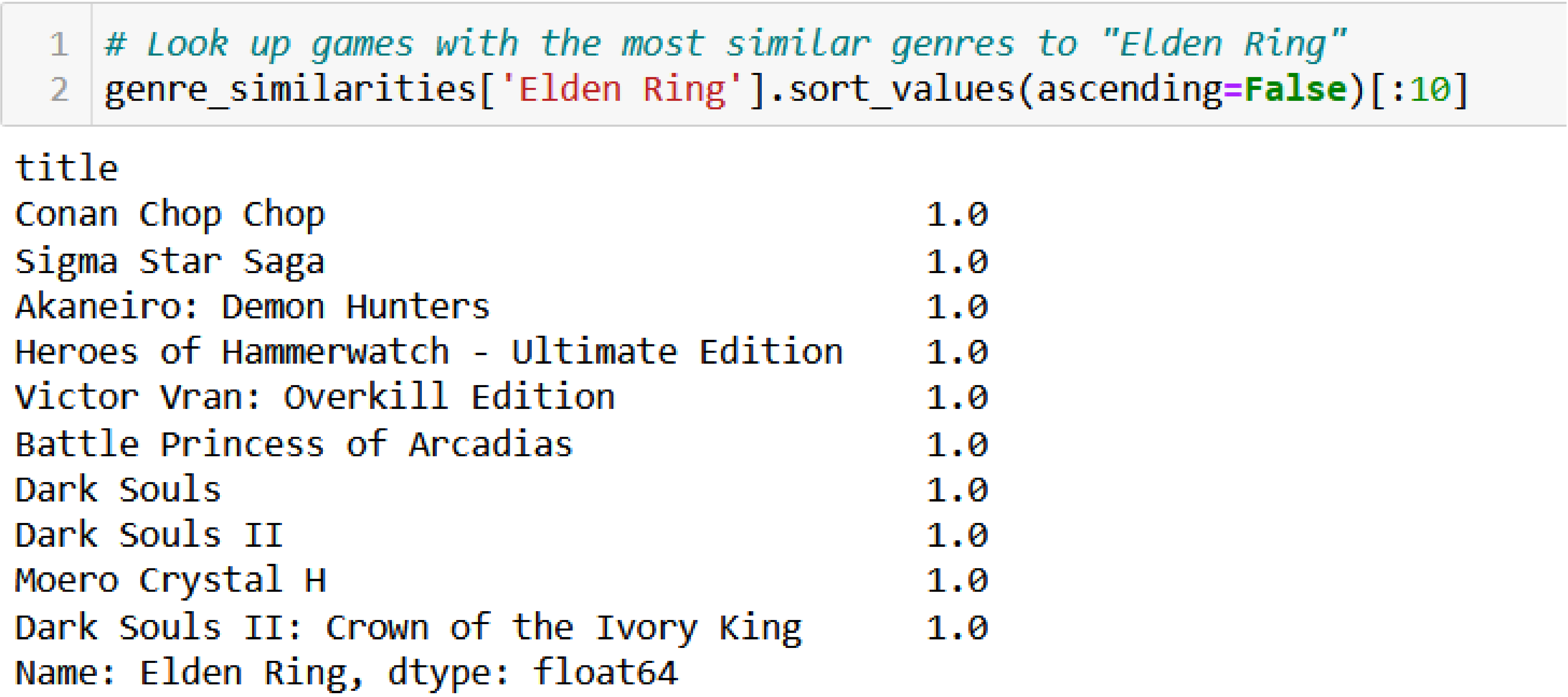


title	#DRIVE	#IDARB	#KILLALLZOMBIES	'Splosion Man	.detuned
#DRIVE	1.0	0.000000	0.000000	0.000000	0.000000
#IDARB	0.0	1.000000	0.142857	0.333333	0.333333
#KILLALLZOMBIES	0.0	0.142857	1.000000	0.142857	0.142857
'Splosion Man	0.0	0.333333	0.142857	1.000000	0.142857
.detuned	0.0	0.333333	0.142857	0.142857	1.000000

Modeling

Content Filtering: Genre Similarities

Games with similar genres to Elden Ring. All the Jaccard scores are 1 here because these games have the exact same genres



```python
# Look up games with the most similar genres to "Elden Ring"
genre_similarities['Elden Ring'].sort_values(ascending=False)[:10]
```

```
title
Conan Chop Chop                                1.0
Sigma Star Saga                                1.0
Akaneiro: Demon Hunters                        1.0
Heroes of Hammerwatch - Ultimate Edition       1.0
Victor Vran: Overkill Edition                  1.0
Battle Princess of Arcadias                    1.0
Dark Souls                                     1.0
Dark Souls II                                  1.0
Moero Crystal H                                1.0
Dark Souls II: Crown of the Ivory King         1.0
Name: Elden Ring, dtype: float64
```

Modeling

Content Filtering: Summary Comparisons

Goal: Get titles that are most similar to other titles based off summaries

Steps

1. TfidfVectorize the summaries

2. Create matrix of cosine similarities

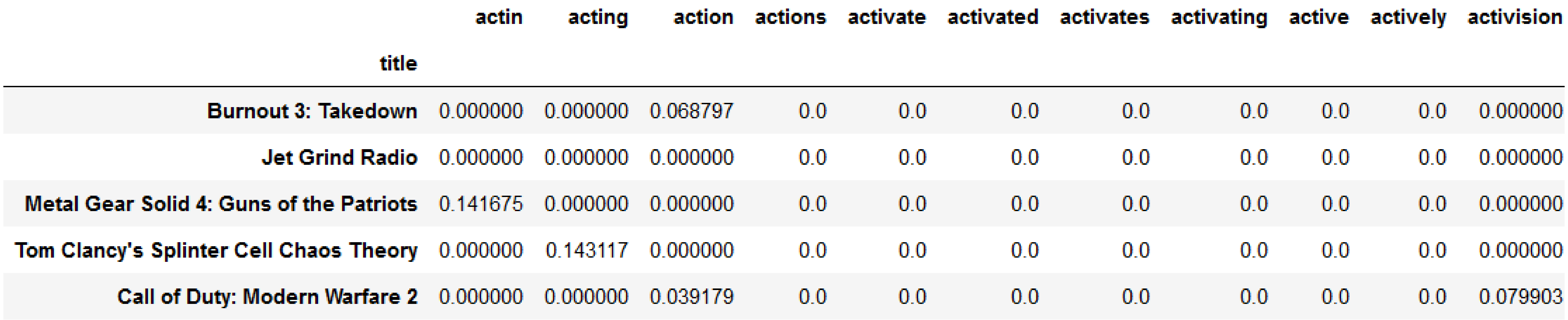
3. Create function to get titles with similar summaries

Modeling

Content Filtering: Summary Comparisons

TfidfVectorizer Matrix

The values are scores of importance of the features in each summary. For this matrix, zero values mean the term does not appear in the summary. Any non-zero value means the term does appear. The closer it is to 1, the more important or unique across documents that term is. Features are chosen if it appeared in atleast 2 summaries and appeared in less than 70% of the all the summaries



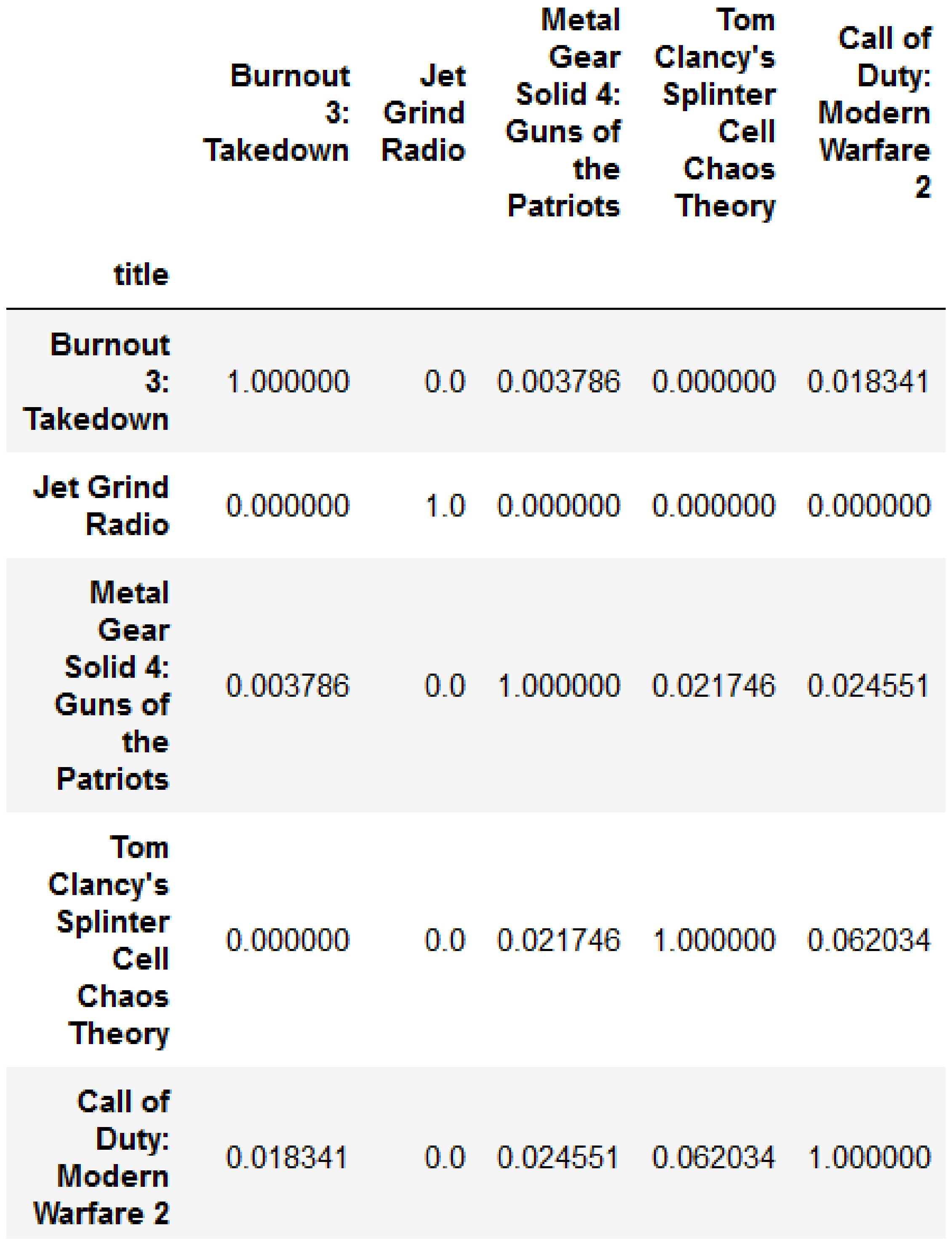
	actin	acting	action	actions	activate	activated	activates	activating	active	actively	activision
title											
Burnout 3: Takedown	0.000000	0.000000	0.068797	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
Jet Grind Radio	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
Metal Gear Solid 4: Guns of the Patriots	0.141675	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
Tom Clancy's Splinter Cell Chaos Theory	0.000000	0.143117	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
Call of Duty: Modern Warfare 2	0.000000	0.000000	0.039179	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.079903

Modeling

Content Filtering: Summary Comparisons

Cosine Similarity Matrix

Notice the 1s along the diagonal. This tells us that computing the cosine similarity between a game and itself will output a score of 1, meaning the summaries are exactly the same. The closer to 1, the more similar the summaries are

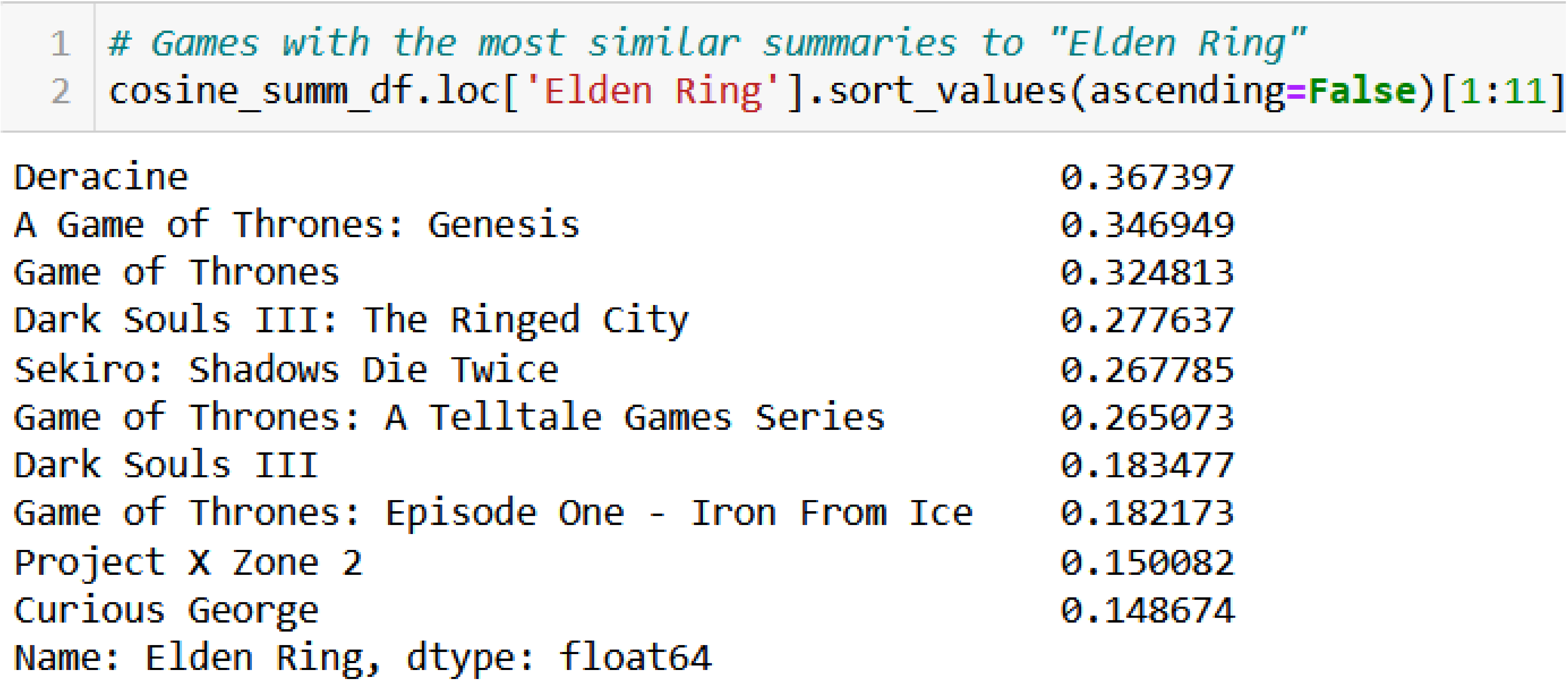


	Burnout 3: Takedown	Jet Grind Radio	Metal Gear Solid 4: Guns of the Patriots	Tom Clancy's Splinter Cell Chaos Theory	Call of Duty: Modern Warfare 2
title					
Burnout 3: Takedown	1.000000	0.0	0.003786	0.000000	0.018341
Jet Grind Radio	0.000000	1.0	0.000000	0.000000	0.000000
Metal Gear Solid 4: Guns of the Patriots	0.003786	0.0	1.000000	0.021746	0.024551
Tom Clancy's Splinter Cell Chaos Theory	0.000000	0.0	0.021746	1.000000	0.062034
Call of Duty: Modern Warfare 2	0.018341	0.0	0.024551	0.062034	1.000000

Modeling

Content Filtering: Summary Comparisons

These are the games with the most similar summaries to Elden ring. For anyone that doesn't know, the plot for Elden Ring and Game of Thrones was written by the same person, George R.R. Martin.



```
# Games with the most similar summaries to "Elden Ring"
cosine_summ_df.loc['Elden Ring'].sort_values(ascending=False)[1:11]
```

```
Deracine                                        0.367397
A Game of Thrones: Genesis                      0.346949
Game of Thrones                                 0.324813
Dark Souls III: The Ringed City                 0.277637
Sekiro: Shadows Die Twice                       0.267785
Game of Thrones: A Telltale Games Series        0.265073
Dark Souls III                                  0.183477
Game of Thrones: Episode One - Iron From Ice    0.182173
Project X Zone 2                                0.150082
Curious George                                  0.148674
Name: Elden Ring, dtype: float64
```

Modeling

Content Filtering: Imrovements

Combine Genre and Summary

Make a recommendation system that takes both genres and summaries into account. Genres mainly describe the gameplay where as the Summaries tend to describe the plot of the game.

Utiilize other features

Other features from the game set could also be utilized somehow. Platform, Developer, and Number of Players could definitely be of interest to a user.

Modeling

Collaboritive Filtering: Scikit Surprise

The Review set was utilized here along with algorithms from the **Scikit Surprise** library.

Models

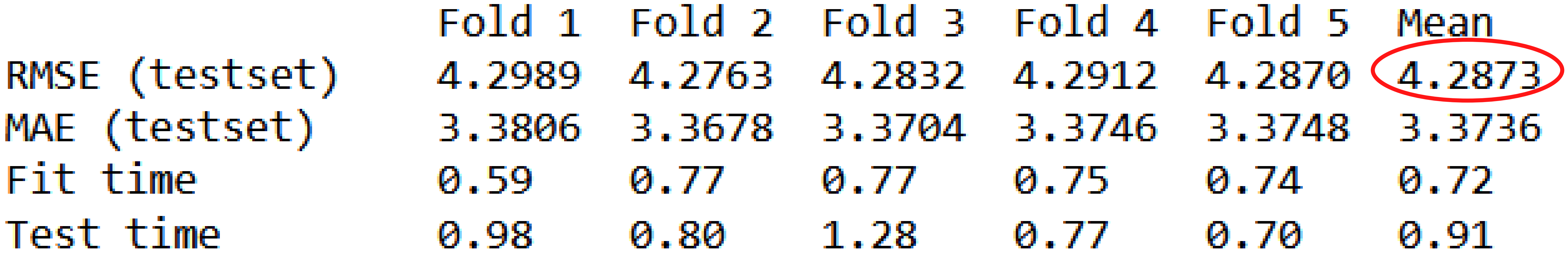
- NormalPredictor
- KNNBasic (User-Based)
- KNNBasic (Item-Based)
- SVD

A general idea behind recommendation algorithms is that they predict user-ratings for items that were not rated by that user. How the predictions are made is what sets the models apart.

Modeling

Collaboritive Filtering: Baseline Model

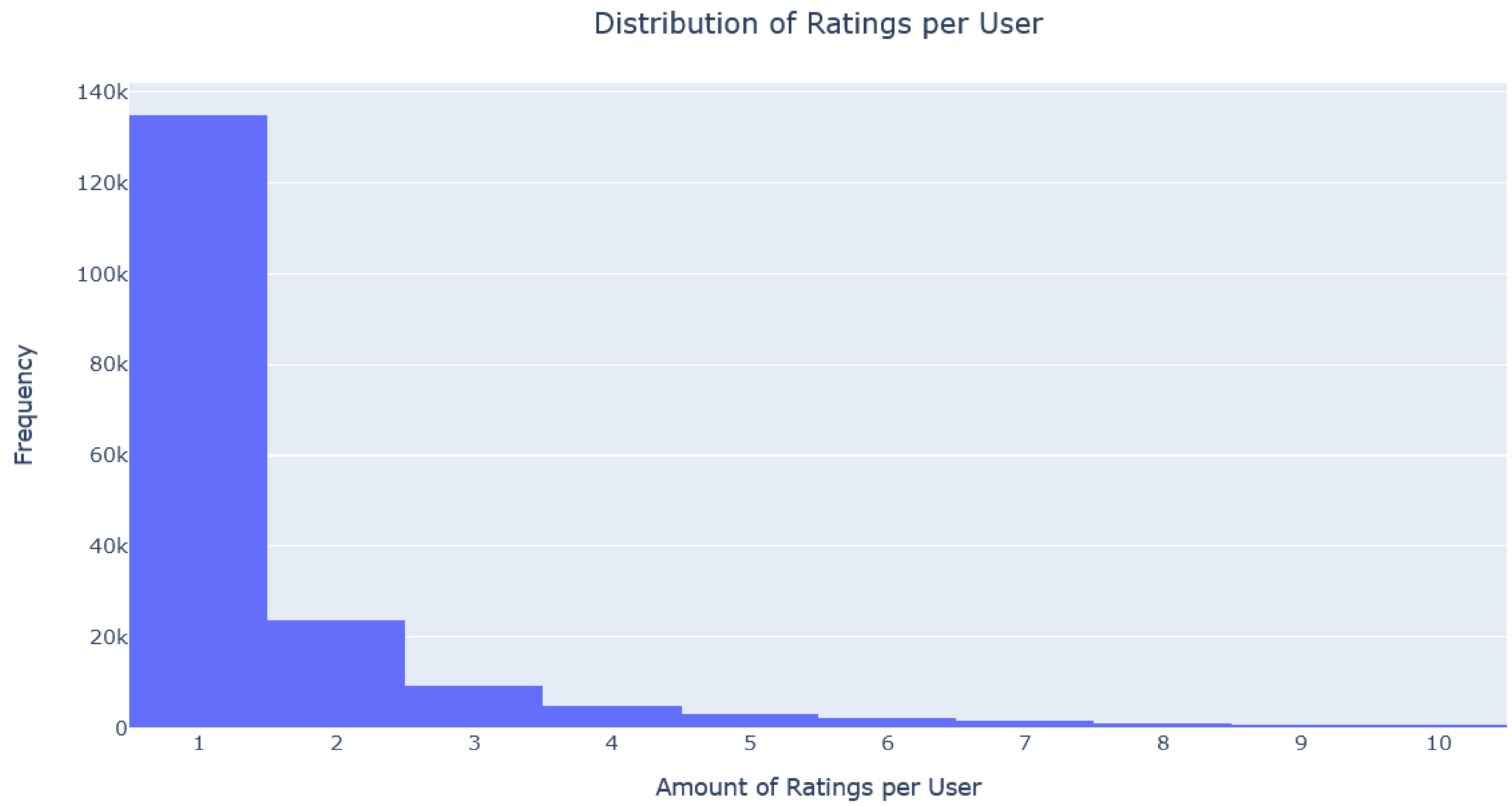
This NormalPredictor algorithm was used for our baseline model, meaning we expect the other 3 algorithms to perform, at the very least, better if not much better. This algorithm predicts random ratings for users based off an assumed normal distribution of the ratings. Obviously, random predictions are not meant to perform well, hence why this was used as the baseline. The output from the first run is seen here. We'll focus on the mean RMSE. This means our rating predictions were off, on avg, by 4.29. Not a great score.



```
                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean
RMSE (testset)    4.2989  4.2763  4.2832  4.2912  4.2870  4.2873
MAE (testset)     3.3806  3.3678  3.3704  3.3746  3.3748  3.3736
Fit time          0.59    0.77    0.77    0.75    0.74    0.72
Test time         0.98    0.80    1.28    0.77    0.70    0.91
```

Modeling

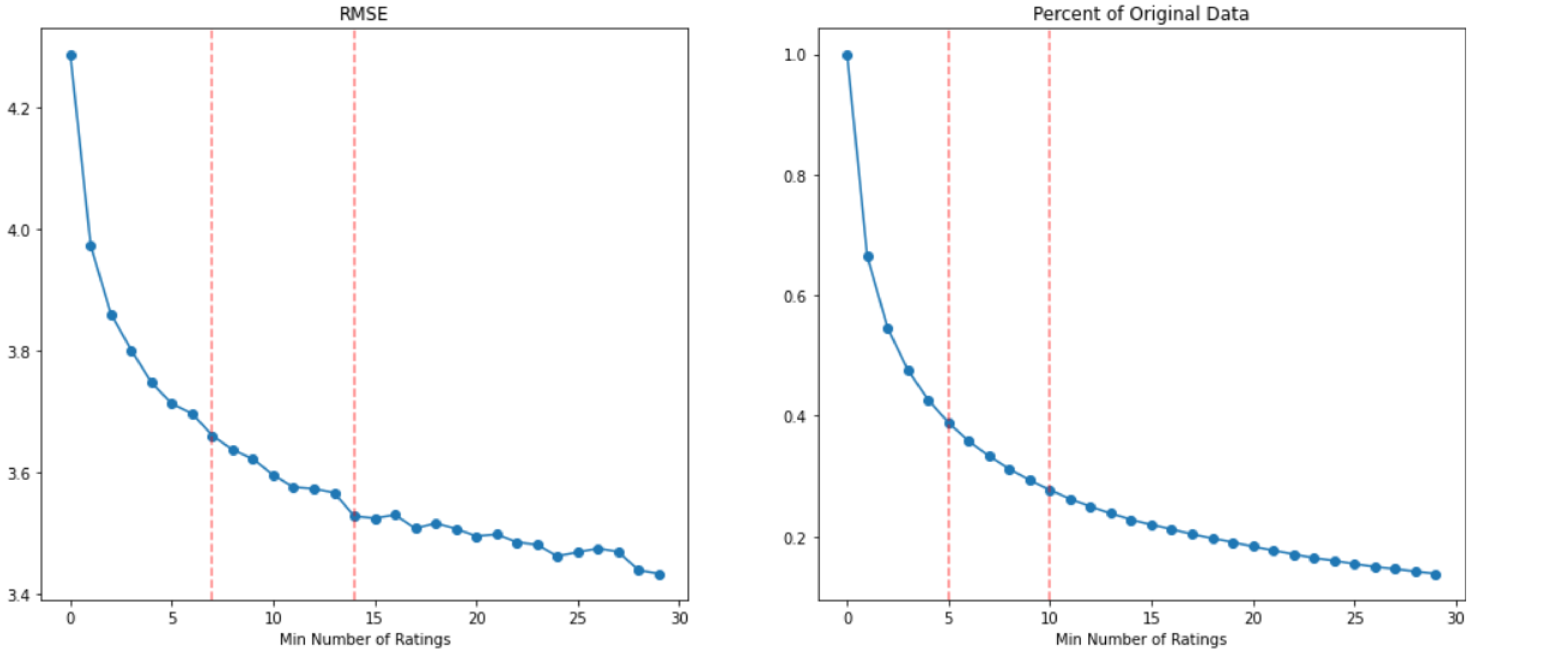
Collaboritive Filtering: Reducing Dimensions



The distribution of ratings per user is extremely skewed to the right. Given there are about 400k ratings, we can see that more than a quarter of the ratings are from users who only rated 1 game.

Modeling

Collaboritive Filtering: Reducing Dimensions

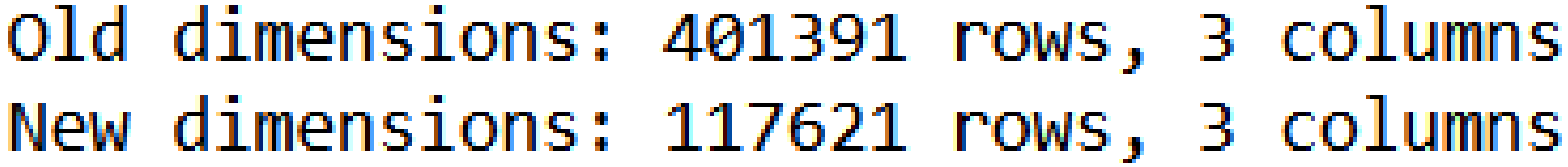


As the data shrinks to only users with more than "n" amount of ratings, the RMSE scores get better. But the percentage of data left drops significantly

Modeling

Collaboritive Filtering: Reducing Dimensions

Now we needed to decide where to set the threshold for minimum number of ratings. The intervals in which the rate of decrease for the RMSE score and the percent of data left slowed between [7,14] and [5,10] respectively. Taking the mean of the intervals gives us 10.5 and 7.5. Taking the mean of these numbers gives us 9, which is what the threshold was set to.

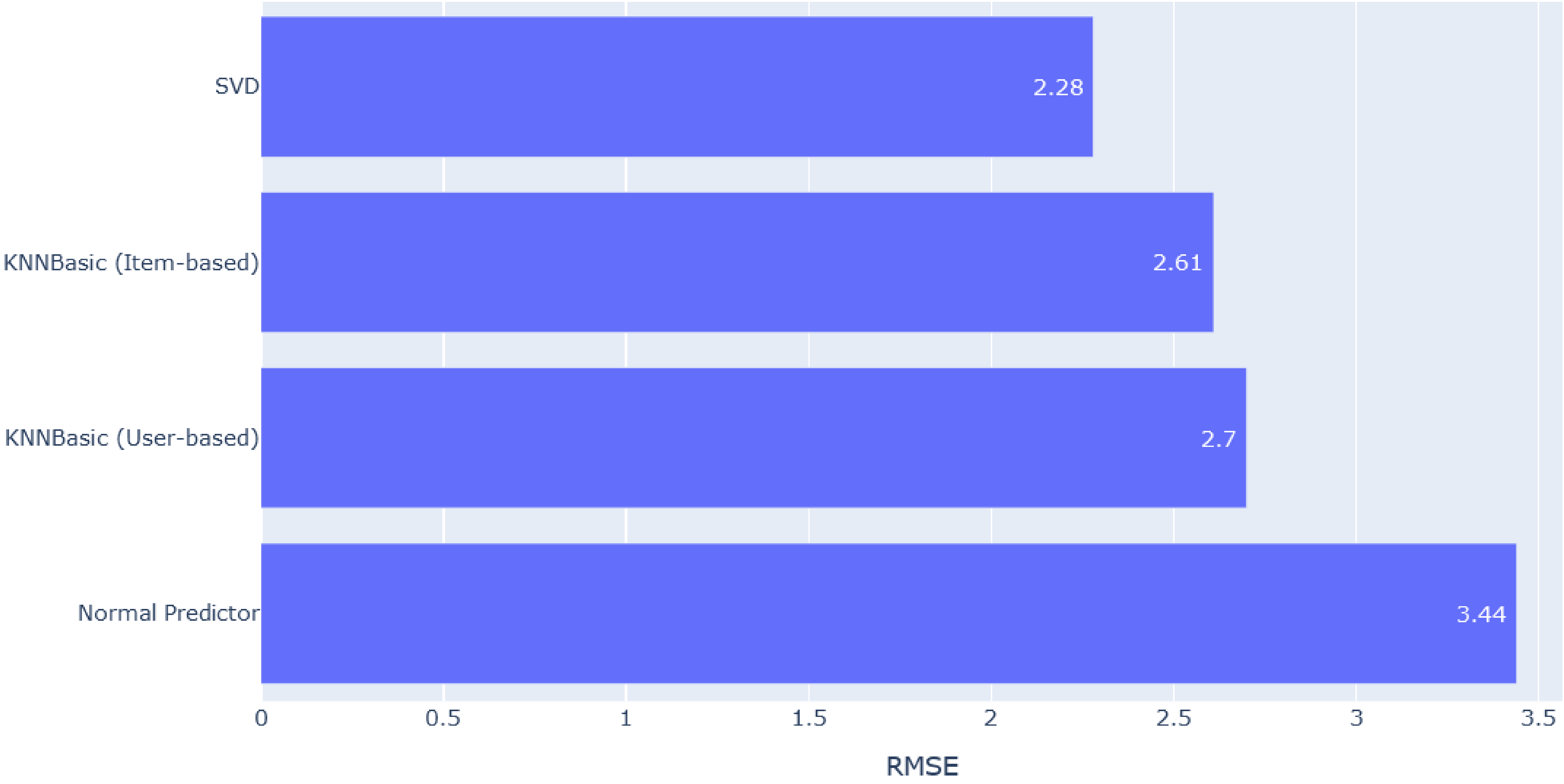


```
Old dimensions: 401391 rows, 3 columns
New dimensions: 117621 rows, 3 columns
```

We're still left with over a quarter of the ratings which is decent.

Modeling

Collaboritive Filtering: Comparing Models

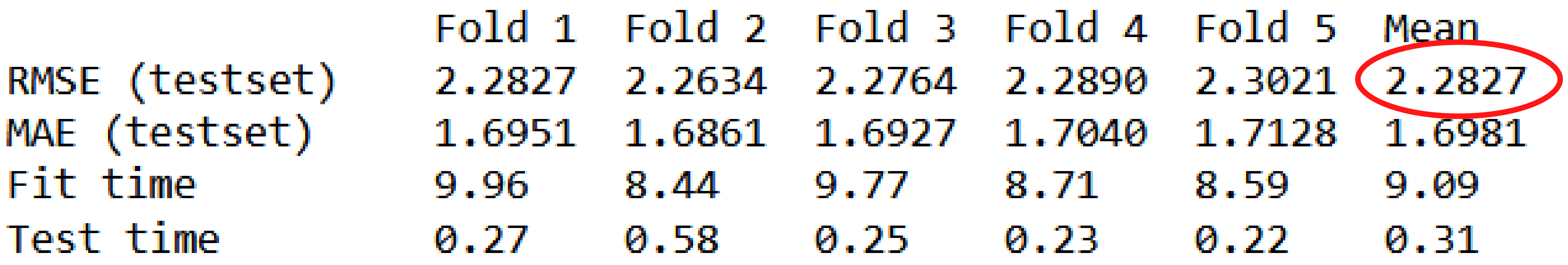


SVD wins!

Modeling

Collaboritive Filtering: Tuned SVD

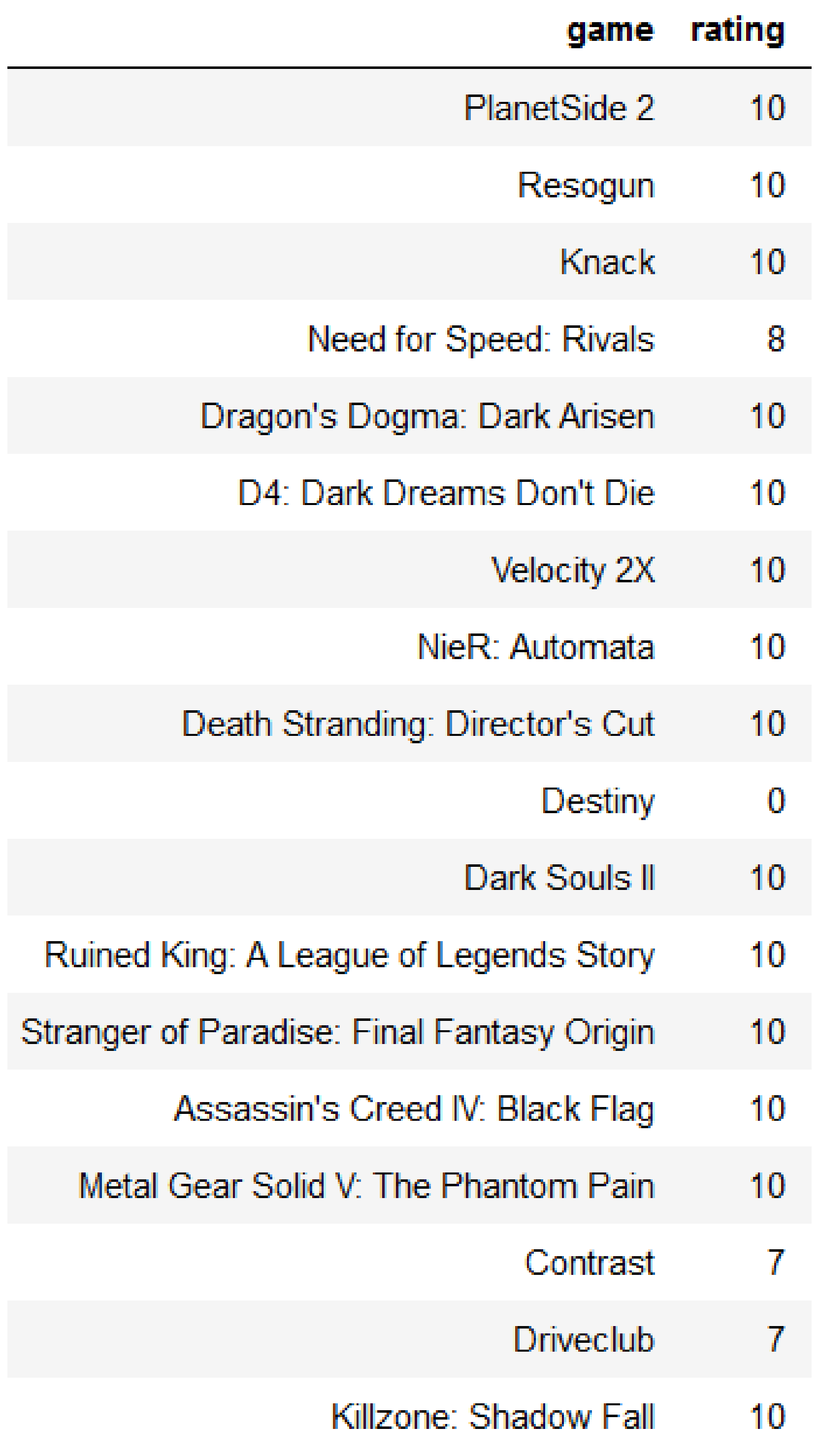
After tuning the SVD model with a GridSearchCV, we were able to marginally improve the RMSE from 2.295 to 2.283



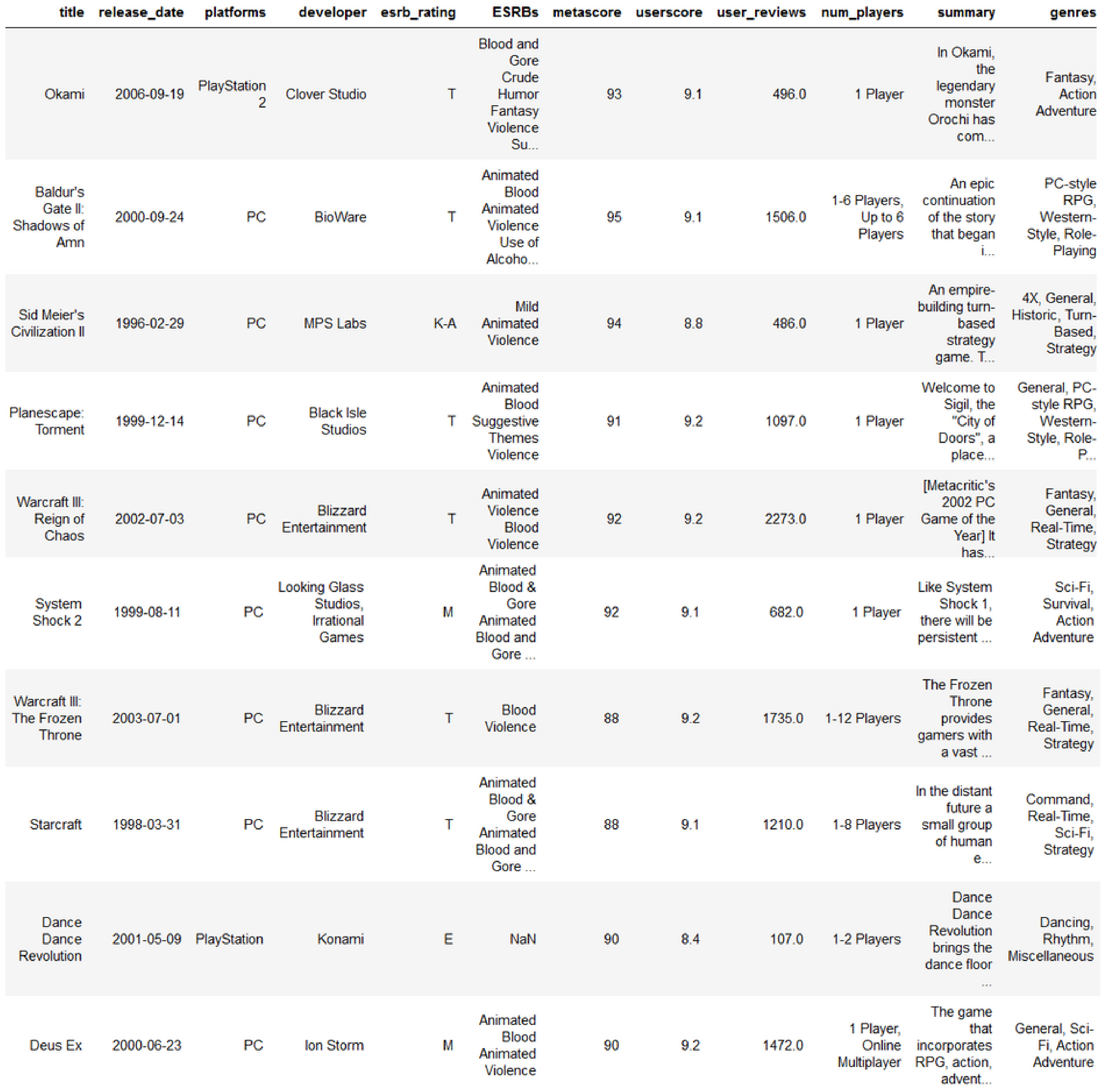
```
                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean
RMSE (testset)    2.2827  2.2634  2.2764  2.2890  2.3021  2.2827
MAE (testset)     1.6951  1.6861  1.6927  1.7040  1.7128  1.6981
Fit time          9.96    8.44    9.77    8.71    8.59    9.09
Test time         0.27    0.58    0.25    0.23    0.22    0.31
```

Modeling

Collaboritive Filtering: Test User recommendations



game	rating
PlanetSide 2	10
Resogun	10
Knack	10
Need for Speed: Rivals	8
Dragon's Dogma: Dark Arisen	10
D4: Dark Dreams Don't Die	10
Velocity 2X	10
NieR: Automata	10
Death Stranding: Director's Cut	10
Destiny	0
Dark Souls II	10
Ruined King: A League of Legends Story	10
Stranger of Paradise: Final Fantasy Origin	10
Assassin's Creed IV: Black Flag	10
Metal Gear Solid V: The Phantom Pain	10
Contrast	7
Driveclub	7
Killzone: Shadow Fall	10



title	release_date	platforms	developer	esrb_rating	ESRBs	metascore	userscore	user_reviews	num_players	summary	genres
Okami	2006-09-19	PlayStation 2	Clover Studio	T	Blood and Gore Crude Humor Fantasy Violence Su...	93	9.1	496.0	1 Player	In Okami, the legendary monster Orochi has com...	Fantasy, Action Adventure
Baldur's Gate II: Shadows of Amn	2000-09-24	PC	BioWare	T	Animated Blood Animated Violence Use of Alcoho...	95	9.1	1506.0	1-6 Players, Up to 6 Players	An epic continuation of the story that began i...	PC-style RPG, Western-Style, Role-Playing
Sid Meier's Civilization II	1996-02-29	PC	MPS Labs	K-A	Mild Animated Violence	94	8.8	486.0	1 Player	An empire-building turn-based strategy game. T...	4X, General, Historic, Turn-Based, Strategy
Planescape: Torment	1999-12-14	PC	Black Isle Studios	T	Animated Blood Suggestive Themes Violence	91	9.2	1097.0	1 Player	Welcome to Sigil, the "City of Doors", a place...	General, PC-style RPG, Western-Style, Role-P...
Warcraft III: Reign of Chaos	2002-07-03	PC	Blizzard Entertainment	T	Animated Violence Blood Violence	92	9.2	2273.0	1 Player	[Metacritic's 2002 PC Game of the Year] It has...	Fantasy, General, Real-Time, Strategy
System Shock 2	1999-08-11	PC	Looking Glass Studios, Irrational Games	M	Animated Blood & Gore Animated Blood and Gore ...	92	9.1	682.0	1 Player	Like System Shock 1, there will be persistent ...	Sci-Fi, Survival, Action Adventure
Warcraft III: The Frozen Throne	2003-07-01	PC	Blizzard Entertainment	T	Blood Violence	88	9.2	1735.0	1-12 Players	The Frozen Throne provides gamers with a vast ...	Fantasy, General, Real-Time, Strategy
Starcraft	1998-03-31	PC	Blizzard Entertainment	T	Animated Blood & Gore Animated Blood and Gore ...	88	9.1	1210.0	1-8 Players	In the distant future a small group of human e...	Command, Real-Time, Sci-Fi, Strategy
Dance Dance Revolution	2001-05-09	PlayStation	Konami	E	NaN	90	8.4	107.0	1-2 Players	Dance Dance Revolution brings the dance floor ...	Dancing, Rhythm, Miscellaneous
Deus Ex	2000-06-23	PC	Ion Storm	M	Animated Blood Animated Violence	90	9.2	1472.0	1 Player, Online Multiplayer	The game that incorporates RPG, action, advent...	General, Sci-Fi, Action Adventure

5.Improvements?

Scrape rest of reviews

Only a maximum of a 100 reviews were taken from each game. Scraping the rest of the reviews off meta would most likely make the recommendations more accurate.

Collect date of each review

The review site, MetaCritic, is over 2 decades old. Many of the user-reviews for older games were posted closer to the release dates. With the date of the review, I could do some adjustment in the recommendations based off the release dates of the title.

Collect platform of title for each review

This was by far my biggest mistake. Without the platform to each of the titles in the review set, I had no way of distinguishing between sets of reviews for the same title on different platforms

Set minimum ratings per title

I could have also tried limiting the titles used to only titles that received a minimum amount of reviews. Instead I only limited the users to users who rated at least 10 games.