`all_sites_scores.csv` contains every film that has a Rotten Tomatoes rating, a RT User rating, a Metacritic score, a Metacritic User score, and IMDb score, and at least 30 fan reviews on Fandango. The data from Fandango was pulled on Aug. 24, 2015.

Column	Definition
FILM	The film in question
RottenTomatoes	The Rotten Tomatoes Tomatometer score for the film
RottenTomatoes_User	The Rotten Tomatoes user score for the film
Metacritic	The Metacritic critic score for the film
Metacritic_User	The Metacritic user score for the film
IMDB	The IMDb user score for the film
Metacritic_user_vote_count	The number of user votes the film had on Metacritic
IMDB_user_vote_count	The number of user votes the film had on IMDb

fandango_scape.csv

`fandango_scrape.csv` contains every film 538 pulled from Fandango.

Column	Definiton
FILM	The movie
STARS	Number of stars presented on Fandango.com
RATING	The Fandango ratingValue for the film, as pulled from the HTML of each page. This is the actual average score the movie obtained.
VOTES	number of people who had reviewed the film at the time we pulled it.

TASK: Import any libraries you think you will use:

In [173]:

```
# IMPORT HERE!
```

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Part Two: Exploring Fandango Displayed Scores versus True User

Ratings

Let's first explore the Fandango ratings to see if our analysis agrees with the article's conclusion.

TASK: Run the cell below to read in the fandango_scrape.csv file

In [2]:

```
fandango = pd.read_csv("fandango_scrape.csv")
```

TASK: Explore the DataFrame Properties and Head.

In [3]:

```
fandango.head()
```

Out[3]:

	FILM	STARS	RATING	VOTES
0	Fifty Shades of Grey (2015)	4.0	3.9	34846
1	Jurassic World (2015)	4.5	4.5	34390
2	American Sniper (2015)	5.0	4.8	34085
3	Furious 7 (2015)	5.0	4.8	33538
4	Inside Out (2015)	4.5	4.5	15749

In [4]:

```
fandango.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 4 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   FILM     504 non-null    object 
 1   STARS    504 non-null    float64
 2   RATING   504 non-null    float64
 3   VOTES    504 non-null    int64  
dtypes: float64(2), int64(1), object(1)
memory usage: 15.9+ KB
```

In [5]:

```
fandango.describe()
```

Out[5]:

	STARS	RATING	VOTES
count	504.000000	504.000000	504.000000
mean	3.558532	3.375794	1147.863095
std	1.563133	1.491223	3830.583136
min	0.000000	0.000000	0.000000
25%	3.500000	3.100000	3.000000
50%	4.000000	3.800000	18.500000
75%	4.500000	4.300000	189.750000
max	5.000000	5.000000	34846.000000

TASK: Let's explore the relationship between popularity of a film and its rating. Create a scatterplot showing the relationship between rating and votes. Feel free to edit visual styling to your preference.

In [179]:

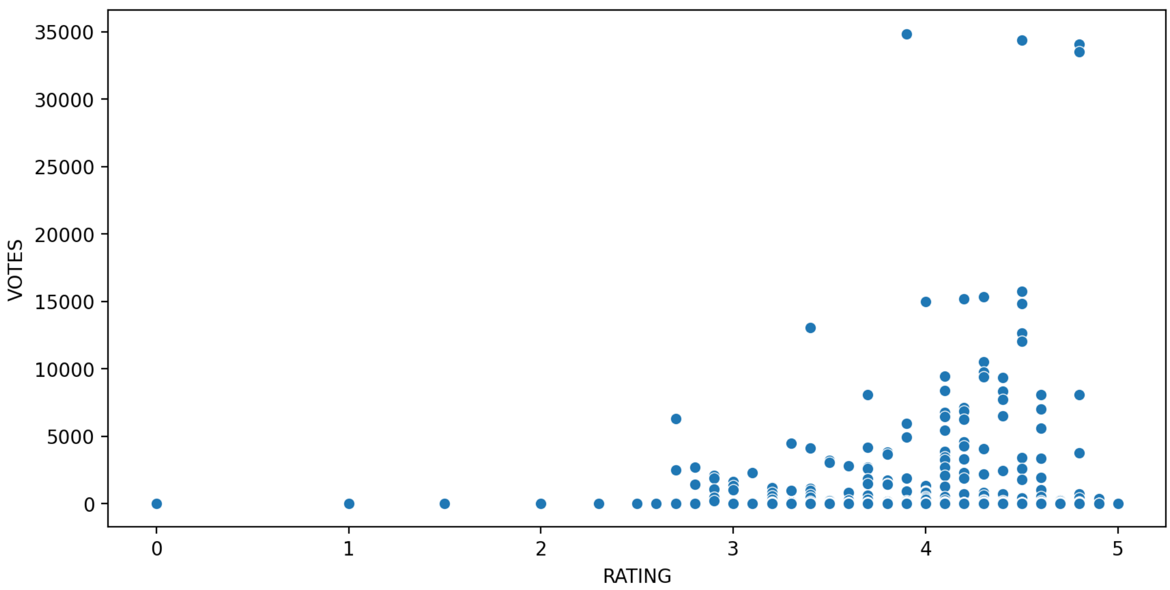
```
# CODE HERE
```

In [6]:

```
plt.figure(figsize=(10,5),dpi = 200)
sns.scatterplot(x = 'RATING',y = 'VOTES',data=fandango)
```

Out[6]:

<AxesSubplot:xlabel='RATING', ylabel='VOTES'>



TASK: Calculate the correlation between the columns:

In [181]:

```
# CODE HERE
```

In [7]:

```
df = fandango.drop('FILM',axis = 1)
df.corr()
```

Out[7]:

	STARS	RATING	VOTES
STARS	1.000000	0.994696	0.164218
RATING	0.994696	1.000000	0.163764
VOTES	0.164218	0.163764	1.000000

TASK: Assuming that every row in the FILM title column has the same format:

```
Film Title Name (Year)
```

Create a new column that is able to strip the year from the title strings and set this new column as YEAR

In [183]:

```
# CODE HERE
```

In [8]:

```
fandango['YEAR'] = fandango['FILM'].apply(lambda title:title.split('(')[-1].replace(')',''))
fandango['YEAR']
```

Out[8]:

```
0      2015
1      2015
2      2015
3      2015
4      2015
       ...
499    2015
500    2015
501    2015
502    1964
503    2012
Name: YEAR, Length: 504, dtype: object
```

TASK: How many movies are in the Fandango DataFrame per year?

In [185]:

```
#CODE HERE
```

In [9]:

```
fandango['YEAR'].value_counts()
```

Out[9]:

```
2015    478
2014     23
2016      1
1964      1
2012      1
Name: YEAR, dtype: int64
```

TASK: Visualize the count of movies per year with a plot:

In [187]:

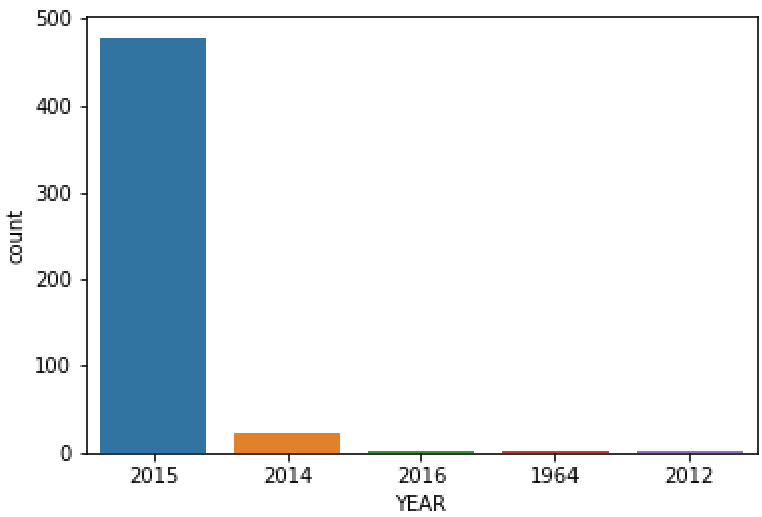
```
#CODE HERE
```

In [10]:

```
sns.countplot(x = 'YEAR',data =fandango)
```

Out[10]:

```
<AxesSubplot:xlabel='YEAR', ylabel='count'>
```



TASK: What are the 10 movies with the highest number of votes?

In [189]:

```
#CODE HERE
```

In [11]:

```
df = fandango.sort_values('VOTES',ascending = False)
df.iloc[0:10]
```

Out[11]:

	FILM	STARS	RATING	VOTES	YEAR
0	Fifty Shades of Grey (2015)	4.0	3.9	34846	2015
1	Jurassic World (2015)	4.5	4.5	34390	2015
2	American Sniper (2015)	5.0	4.8	34085	2015
3	Furious 7 (2015)	5.0	4.8	33538	2015
4	Inside Out (2015)	4.5	4.5	15749	2015
5	The Hobbit: The Battle of the Five Armies (2014)	4.5	4.3	15337	2014
6	Kingsman: The Secret Service (2015)	4.5	4.2	15205	2015
7	Minions (2015)	4.0	4.0	14998	2015
8	Avengers: Age of Ultron (2015)	5.0	4.5	14846	2015
9	Into the Woods (2014)	3.5	3.4	13055	2014

TASK: How many movies have zero votes?

In [191]:

```
#CODE HERE
```

In [12]:

```
df = fandango.sort_values('VOTES')
df['VOTES'].value_counts()
```

Out[12]:

```
0        69
1        35
4        22
2        19
5        15
         ..
409       1
449       1
450       1
496       1
34846     1
Name: VOTES, Length: 210, dtype: int64
```

TASK: Create DataFrame of only reviewed films by removing any films that have zero votes.

In [193]:

```
#CODE HERE
```

In [13]:

```
fandango = fandango.replace(0,np.nan)
fandango = fandango.dropna()
fandango
```

Out[13]:

	FILM	STARS	RATING	VOTES	YEAR
0	Fifty Shades of Grey (2015)	4.0	3.9	34846.0	2015
1	Jurassic World (2015)	4.5	4.5	34390.0	2015
2	American Sniper (2015)	5.0	4.8	34085.0	2015
3	Furious 7 (2015)	5.0	4.8	33538.0	2015
4	Inside Out (2015)	4.5	4.5	15749.0	2015
...	...	...	...	...	...
430	That Sugar Film (2015)	5.0	5.0	1.0	2015
431	The Intern (2015)	5.0	5.0	1.0	2015
432	The Park Bench (2015)	5.0	5.0	1.0	2015
433	The Wanted 18 (2015)	5.0	5.0	1.0	2015
434	Z For Zachariah (2015)	5.0	5.0	1.0	2015

435 rows × 5 columns

As noted in the article, due to HTML and star rating displays, the true user rating may be slightly different than the rating shown to a user. Let's visualize this difference in distributions.

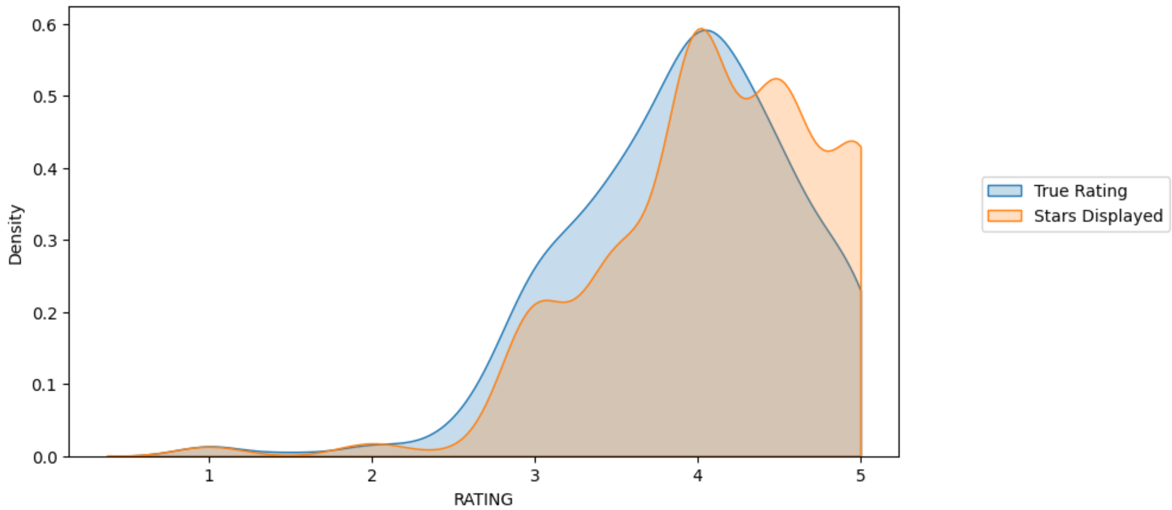
TASK: Create a KDE plot (or multiple kdeplots) that displays the distribution of ratings that are displayed (STARS) versus what the true rating was from votes (RATING). Clip the KDEs to 0-5.

In [14]:

```
plt.figure(figsize=(9,5),dpi = 100)
sns.kdeplot(x = 'RATING',data = fandango,clip = [0,5],shade = True,label = 'True Rating')
sns.kdeplot(x = 'STARS',data = fandango,clip = [0,5],shade = True , label = 'Stars Displaye
plt.legend(loc = [1.1,0.5])
```

Out[14]:

<matplotlib.legend.Legend at 0x1ab1ac1a340>

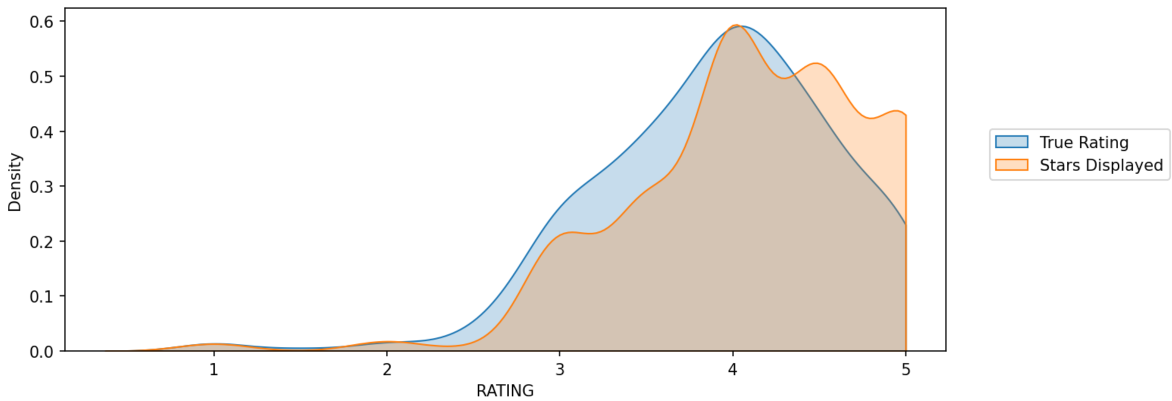


In [196]:

```
```

Out[196]:

<matplotlib.legend.Legend at 0x1aa0110cdc8>



TASK: Let's now actually quantify this discrepancy. Create a new column of the different between STARS displayed versus true RATING. Calculate this difference with STARS-RATING and round these differences to the nearest decimal point.

In [15]:

```
fandango['STARS_DIFF'] = (fandango['STARS'] - fandango['RATING']).round(1)
fandango = fandango.sort_values('VOTES',ascending = False)
fandango
```

Out[15]:

	FILM	STARS	RATING	VOTES	YEAR	STARS_DIFF
0	Fifty Shades of Grey (2015)	4.0	3.9	34846.0	2015	0.1
1	Jurassic World (2015)	4.5	4.5	34390.0	2015	0.0
2	American Sniper (2015)	5.0	4.8	34085.0	2015	0.2
3	Furious 7 (2015)	5.0	4.8	33538.0	2015	0.2
4	Inside Out (2015)	4.5	4.5	15749.0	2015	0.0
...	...	...	...	...	...	...
410	One Cut, One Life (2015)	3.0	3.0	1.0	2015	0.0
411	The Face of an Angel (2015)	3.0	3.0	1.0	2015	0.0
412	The Living (2015)	3.0	3.0	1.0	2015	0.0
414	Buggs Bunny (2015)	4.0	4.0	1.0	2015	0.0
434	Z For Zachariah (2015)	5.0	5.0	1.0	2015	0.0

435 rows × 6 columns

In [198]:

```

```

In [199]:

```

```

Out[199]:

	FILM	STARS	RATING	VOTES	YEAR	STARS_DIFF
0	Fifty Shades of Grey (2015)	4.0	3.9	34846	2015	0.1
1	Jurassic World (2015)	4.5	4.5	34390	2015	0.0
2	American Sniper (2015)	5.0	4.8	34085	2015	0.2
3	Furious 7 (2015)	5.0	4.8	33538	2015	0.2
4	Inside Out (2015)	4.5	4.5	15749	2015	0.0
...	...	...	...	...	...	...
430	That Sugar Film (2015)	5.0	5.0	1	2015	0.0
431	The Intern (2015)	5.0	5.0	1	2015	0.0
432	The Park Bench (2015)	5.0	5.0	1	2015	0.0
433	The Wanted 18 (2015)	5.0	5.0	1	2015	0.0
434	Z For Zachariah (2015)	5.0	5.0	1	2015	0.0

435 rows × 6 columns

TASK: Create a count plot to display the number of times a certain difference occurs:

In [200]:

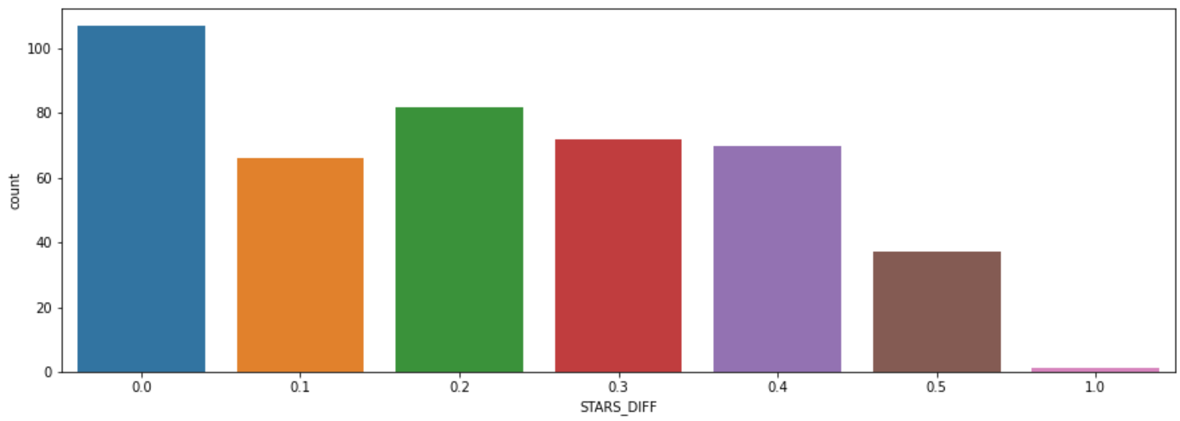
```
#CODE HERE
```

In [16]:

```
plt.figure(figsize = [15,5])
sns.countplot(x = 'STARS_DIFF',data = fandango)
```

Out[16]:

```
<AxesSubplot:xlabel='STARS_DIFF', ylabel='count'>
```



TASK: We can see from the plot that one movie was displaying over a 1 star difference than its true rating! What movie had this close to 1 star differential?

In [202]:

```
#CODE HERE
```

In [203]:

```

```

Out[203]:

	FILM	STARS	RATING	VOTES	YEAR	STARS_DIFF
381	Turbo Kid (2015)	5.0	4.0	2	2015	1.0

Part Three: Comparison of Fandango Ratings to Other Sites

Let's now compare the scores from Fandango to other movies sites and see how they compare.

TASK: Read in the "all_sites_scores.csv" file by running the cell below

In [17]:

```
all_sites = pd.read_csv("all_sites_scores.csv")
```

TASK: Explore the DataFrame columns, info, description.

In [18]:

```
all_sites.head()
```

Out[18]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacrit
0	Avengers: Age of Ultron (2015)	74	86	66	7.1	7.8	
1	Cinderella (2015)	85	80	67	7.5	7.1	
2	Ant-Man (2015)	80	90	64	8.1	7.8	
3	Do You Believe? (2015)	18	84	22	4.7	5.4	
4	Hot Tub Time Machine 2 (2015)	14	28	29	3.4	5.1	

In [205]:

```

```

Out[205]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacrit
0	Avengers: Age of Ultron (2015)	74	86	66	7.1	7.8	
1	Cinderella (2015)	85	80	67	7.5	7.1	
2	Ant-Man (2015)	80	90	64	8.1	7.8	
3	Do You Believe? (2015)	18	84	22	4.7	5.4	
4	Hot Tub Time Machine 2 (2015)	14	28	29	3.4	5.1	

In [19]:

```
all_sites.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 146 entries, 0 to 145
Data columns (total 8 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   FILM                        146 non-null    object 
 1   RottenTomatoes              146 non-null    int64  
 2   RottenTomatoes_User         146 non-null    int64  
 3   Metacritic                  146 non-null    int64  
 4   Metacritic_User             146 non-null    float64
 5   IMDB                        146 non-null    float64
 6   Metacritic_user_vote_count  146 non-null    int64  
 7   IMDB_user_vote_count        146 non-null    int64  
dtypes: float64(2), int64(5), object(1)
memory usage: 9.2+ KB
```

In [20]:

```
all_sites.describe()
```

Out[20]:

	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacrit
count	146.000000	146.000000	146.000000	146.000000	146.000000	
mean	60.849315	63.876712	58.808219	6.519178	6.736986	
std	30.168799	20.024430	19.517389	1.510712	0.958736	
min	5.000000	20.000000	13.000000	2.400000	4.000000	
25%	31.250000	50.000000	43.500000	5.700000	6.300000	
50%	63.500000	66.500000	59.000000	6.850000	6.900000	
75%	89.000000	81.000000	75.000000	7.500000	7.400000	
max	100.000000	94.000000	94.000000	9.600000	8.600000	

Rotten Tomatoes

Let's first take a look at Rotten Tomatoes. RT has two sets of reviews, their critics reviews (ratings published by official critics) and user reviews.

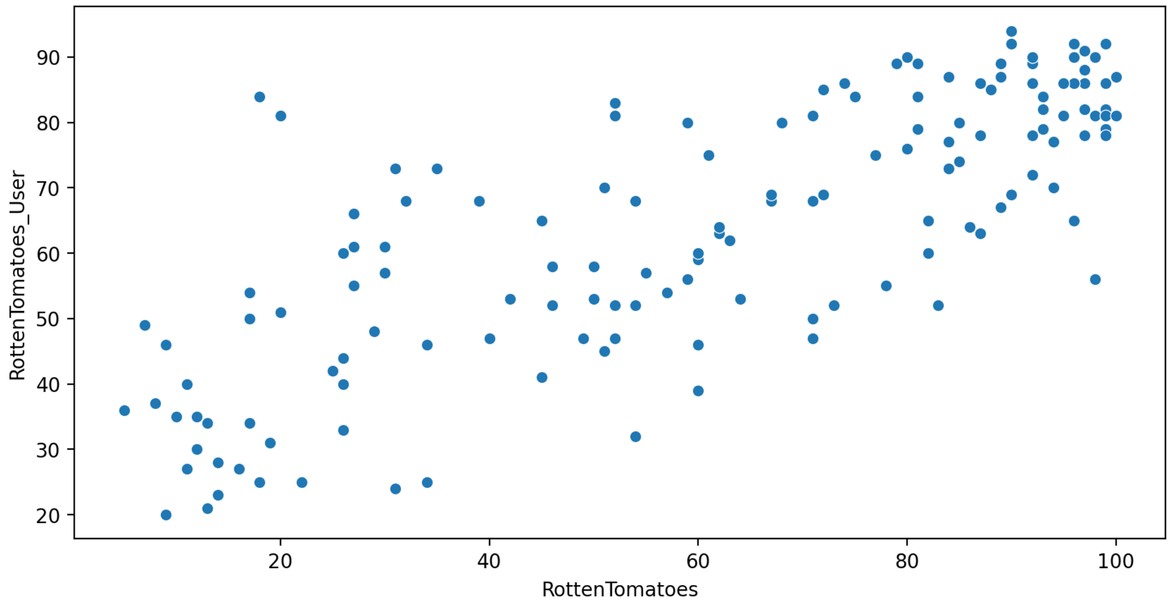
TASK: Create a scatterplot exploring the relationship between RT Critic reviews and RT User reviews.

In [21]:

```
plt.figure(figsize = (10,5),dpi = 200)
sns.scatterplot(x = 'RottenTomatoes',y = 'RottenTomatoes_User',data = all_sites)
```

Out[21]:

```
<AxesSubplot:xlabel='RottenTomatoes', ylabel='RottenTomatoes_User'>
```



Let's quantify this difference by comparing the critics ratings and the RT User ratings. We will calculate this with RottenTomatoes-RottenTomatoes_User. Note: Rotten_Diff here is Critics - User Score. So values closer to 0 means aggrement between Critics and Users. Larger positive values means critics rated much higher than users. Larger negative values means users rated much higher than critics.

TASK: Create a new column based off the difference between critics ratings and users ratings for Rotten Tomatoes. Calculate this with RottenTomatoes-RottenTomatoes_User

In [22]:

```
all_sites['Rotten_Diff'] = all_sites['RottenTomatoes'] - all_sites['RottenTomatoes_User']
all_sites
```

Out[22]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metac
0	Avengers: Age of Ultron (2015)	74	86	66	7.1	7.8	
1	Cinderella (2015)	85	80	67	7.5	7.1	
2	Ant-Man (2015)	80	90	64	8.1	7.8	
3	Do You Believe? (2015)	18	84	22	4.7	5.4	
4	Hot Tub Time Machine 2 (2015)	14	28	29	3.4	5.1	
...	...	...	...	...	...	...	
141	Mr. Holmes (2015)	87	78	67	7.9	7.4	
142	'71 (2015)	97	82	83	7.5	7.2	
143	Two Days, One Night (2014)	97	78	89	8.8	7.4	
144	Gett: The Trial of Viviane Amsalem (2015)	100	81	90	7.3	7.8	
145	Kumiko, The Treasure Hunter (2015)	87	63	68	6.4	6.7	

146 rows × 9 columns

Let's now compare the overall mean difference. Since we're dealing with differences that could be negative or positive, first take the absolute value of all the differences, then take the mean. This would report back on average to absolute difference between the critics rating versus the user rating.

TASK: Calculate the Mean Absolute Difference between RT scores and RT User scores as described above.

In [23]:

```
all_sites['Rotten_Diff_abs']=abs(all_sites['Rotten_Diff'])
all_sites['Rotten_Diff_abs'].mean()
```

Out[23]:

```
15.095890410958905
```

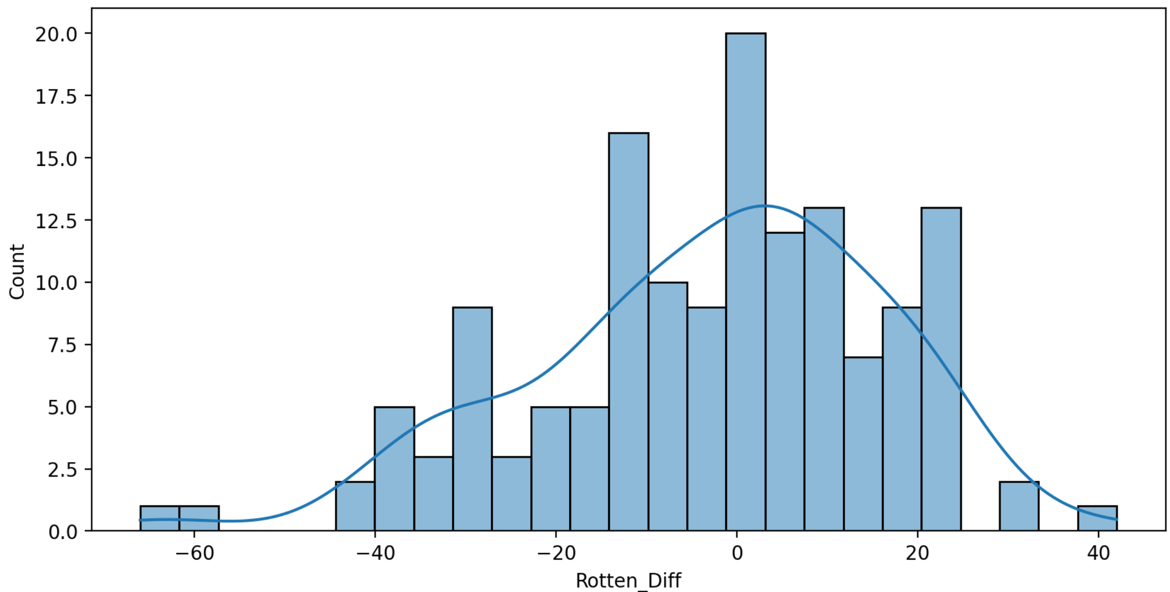
TASK: Plot the distribution of the differences between RT Critics Score and RT User Score. There should be negative values in this distribution plot. Feel free to use KDE or Histograms to display this distribution.

In [24]:

```
plt.figure(figsize = (10,5),dpi = 200)
sns.histplot(data = all_sites,x = 'Rotten_Diff',bins = 25,kde = True)
```

Out[24]:

```
<AxesSubplot:xlabel='Rotten_Diff', ylabel='Count'>
```



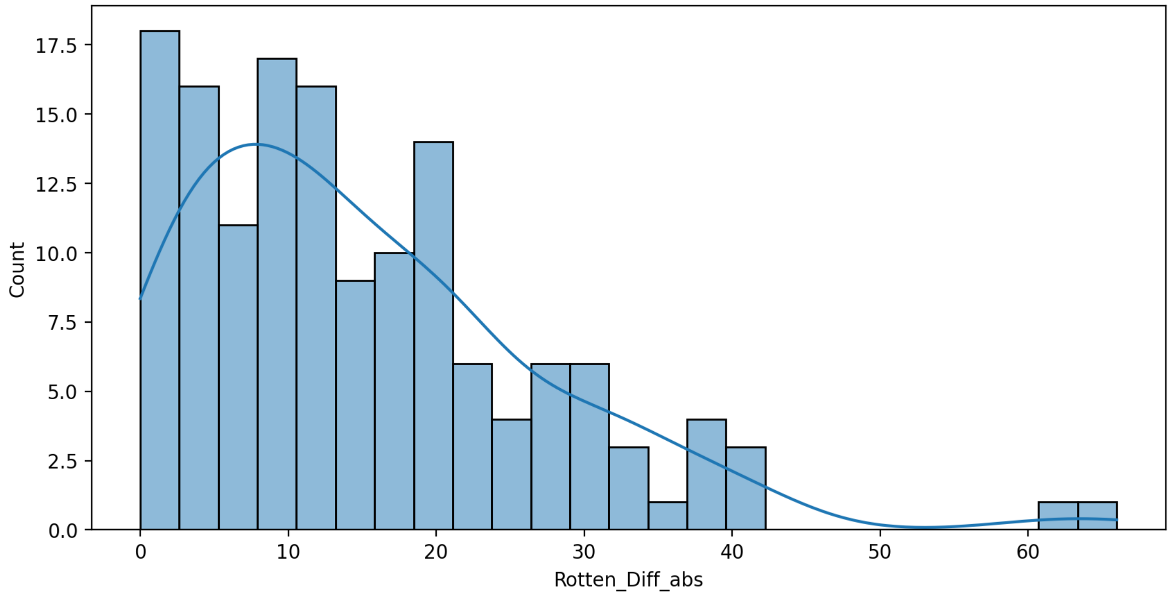
TASK: Now create a distribution showing the *absolute value* difference between Critics and Users on Rotten Tomatoes.

In [25]:

```
plt.figure(figsize = (10,5),dpi = 200)
sns.histplot(data = all_sites,x = 'Rotten_Diff_abs',bins = 25,kde = True)
```

Out[25]:

<AxesSubplot:xlabel='Rotten_Diff_abs', ylabel='Count'>



Let's find out which movies are causing the largest differences. First, show the top 5 movies with the largest *negative* difference between Users and RT critics. Since we calculated the difference as Critics Rating - Users Rating, then large negative values imply the users rated the movie much higher on average than the critics did.

TASK: What are the top 5 movies users rated higher than critics on average:

In [26]:

```
df = all_sites.sort_values('Rotten_Diff')
df.iloc[0:5]
```

Out[26]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacr
3	Do You Believe? (2015)	18	84	22	4.7	5.4	
85	Little Boy (2015)	20	81	30	5.9	7.4	
134	The Longest Ride (2015)	31	73	33	4.8	7.2	
105	Hitman: Agent 47 (2015)	7	49	28	3.3	5.9	
125	The Wedding Ringer (2015)	27	66	35	3.3	6.7	

In [ ]:

```

```

TASK: Now show the top 5 movies critics scores higher than users on average.

In [27]:

```
df = all_sites.sort_values('Rotten_Diff',ascending = False)
df.iloc[0:5]
```

Out[27]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metaci
69	Mr. Turner (2014)	98	56	94	6.6	6.9	
112	It Follows (2015)	96	65	83	7.5	6.9	
115	While We're Young (2015)	83	52	76	6.7	6.4	
145	Kumiko, The Treasure Hunter (2015)	87	63	68	6.4	6.7	
37	Welcome to Me (2015)	71	47	67	6.9	5.9	

In [221]:

```

```

```
Critics love, but Users Hate
```

Out[221]:

	FILM	Rotten_Diff
69	Mr. Turner (2014)	42
112	It Follows (2015)	31
115	While We're Young (2015)	31
37	Welcome to Me (2015)	24
40	I'll See You In My Dreams (2015)	24

MetaCritic

Now let's take a quick look at the ratings from MetaCritic. Metacritic also shows an average user rating versus their official displayed rating.

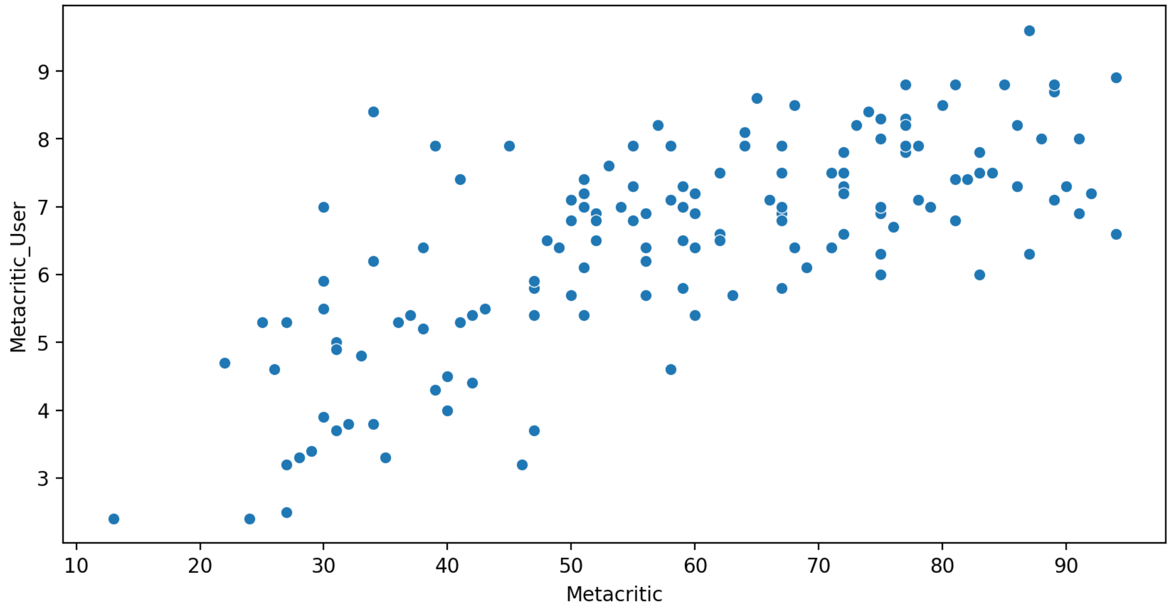
TASK: Display a scatterplot of the Metacritic Rating versus the Metacritic User rating.

In [28]:

```
plt.figure(figsize = (10,5),dpi = 200)
sns.scatterplot(x = 'Metacritic',y = 'Metacritic_User',data = all_sites)
```

Out[28]:

```
<AxesSubplot:xlabel='Metacritic', ylabel='Metacritic_User'>
```



IMDB

Finally let's explore IMDB. Notice that both Metacritic and IMDB report back vote counts. Let's analyze the most popular movies.

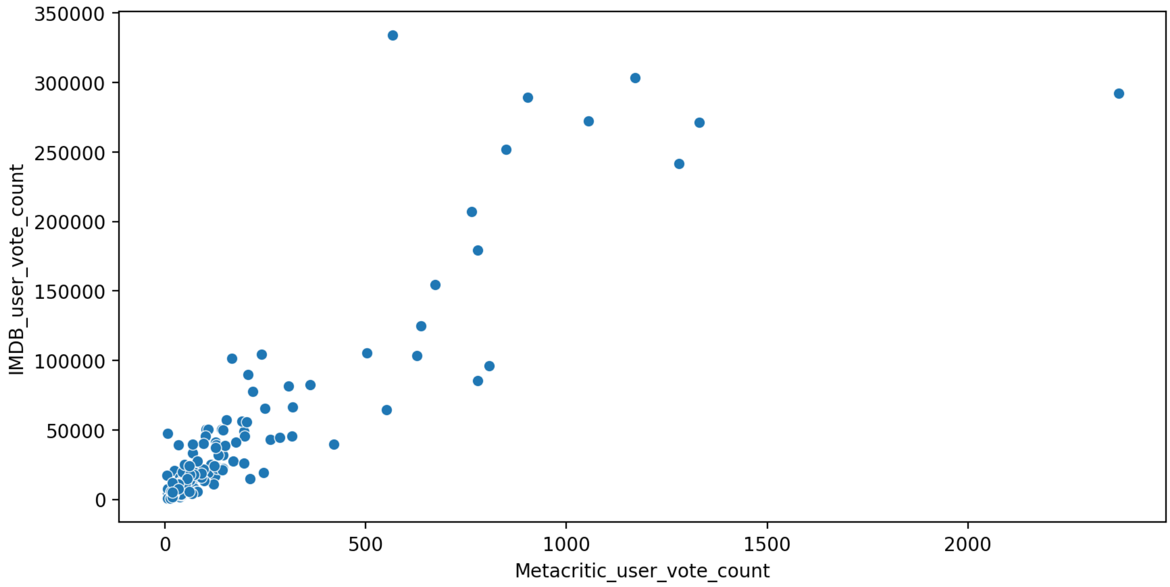
TASK: Create a scatterplot for the relationship between vote counts on MetaCritic versus vote counts on IMDB.

In [29]:

```
plt.figure(figsize = (10,5),dpi = 200)
sns.scatterplot(x = 'Metacritic_user_vote_count',y = 'IMDB_user_vote_count',data = all_site
```

Out[29]:

<AxesSubplot:xlabel='Metacritic_user_vote_count', ylabel='IMDB_user_vote_cou
nt'>



Notice there are two outliers here. The movie with the highest vote count on IMDB only has about 500 Metacritic ratings. What is this movie?

TASK: What movie has the highest IMDB user vote count?

In [33]:

```
all_sites.nlargest(1,'IMDB_user_vote_count')
```

Out[33]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacriti
14	The Imitation Game (2014)	90	92	73	8.2	8.1	

In [227]:

```

```

Out[227]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacriti
14	The Imitation Game (2014)	90	92	73	8.2	8.1	

TASK: What movie has the highest Metacritic User Vote count?

In [34]:

```
all_sites.nlargest(1,'Metacritic_user_vote_count')
```

Out[34]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacritic_
88	Mad Max: Fury Road (2015)	97	88	89	8.7	8.3	

In [229]:

```

```

Out[229]:

	FILM	RottenTomatoes	RottenTomatoes_User	Metacritic	Metacritic_User	IMDB	Metacritic_
88	Mad Max: Fury Road (2015)	97	88	89	8.7	8.3	

Fandago Scores vs. All Sites

Finally let's begin to explore whether or not Fandango artificially displays higher ratings than warranted to boost ticket sales.

TASK: Combine the Fandango Table with the All Sites table. Not every movie in the Fandango table is in the All Sites table, since some Fandango movies have very little or no reviews. We only want to compare movies that are in both DataFrames, so do an *inner* merge to merge together both DataFrames based on the FILM columns.

In [42]:

```
df = pd.merge(fandango,all_sites,on='FILM',how='inner')
```

In [44]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 142 entries, 0 to 141
Data columns (total 15 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   FILM                         142 non-null    object 
 1   STARS                        142 non-null    float64
 2   RATING                       142 non-null    float64
 3   VOTES                        142 non-null    float64
 4   YEAR                         142 non-null    object 
 5   STARS_DIFF                   142 non-null    float64
 6   RottenTomatoes               142 non-null    int64  
 7   RottenTomatoes_User          142 non-null    int64  
 8   Metacritic                   142 non-null    int64  
 9   Metacritic_User              142 non-null    float64
 10  IMDB                         142 non-null    float64
 11  Metacritic_user_vote_count  142 non-null    int64  
 12  IMDB_user_vote_count         142 non-null    int64  
 13  Rotten_Diff                  142 non-null    float64
 14  Rotten_Diff_abs              142 non-null    float64
dtypes: float64(8), int64(5), object(2)
memory usage: 17.8+ KB
```

In [232]:

```

```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 145 entries, 0 to 144
Data columns (total 13 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   FILM                         145 non-null    object 
 1   STARS                        145 non-null    float64
 2   RATING                       145 non-null    float64
 3   VOTES                        145 non-null    int64  
 4   YEAR                         145 non-null    object 
 5   RottenTomatoes               145 non-null    int64  
 6   RottenTomatoes_User          145 non-null    int64  
 7   Metacritic                   145 non-null    int64  
 8   Metacritic_User              145 non-null    float64
 9   IMDB                         145 non-null    float64
 10  Metacritic_user_vote_count  145 non-null    int64  
 11  IMDB_user_vote_count         145 non-null    int64  
 12  Rotten_Diff                  145 non-null    int64  
dtypes: float64(4), int64(7), object(2)
memory usage: 15.9+ KB
```

In [233]:

Out[233]:

	FILM	STARS	RATING	VOTES	YEAR	RottenTomatoes	RottenTomatoes_User	Metacritic
0	Fifty Shades of Grey (2015)	4.0	3.9	34846	2015	25	42	46
1	Jurassic World (2015)	4.5	4.5	34390	2015	71	81	59
2	American Sniper (2015)	5.0	4.8	34085	2015	72	85	72
3	Furious 7 (2015)	5.0	4.8	33538	2015	81	84	67
4	Inside Out (2015)	4.5	4.5	15749	2015	98	90	94

Normalize columns to Fandango STARS and RATINGS 0-5

Notice that RT,Metacritic, and IMDB don't use a score between 0-5 stars like Fandango does. In order to do a fair comparison, we need to *normalize* these values so they all fall between 0-5 stars and the relationship between reviews stays the same.

TASK: Create new normalized columns for all ratings so they match up within the 0-5 star range shown on Fandango. There are many ways to do this.

Hint link: https://stackoverflow.com/questions/26414913/normalize-columns-of-pandas-data-frame (https://stackoverflow.com/questions/26414913/normalize-columns-of-pandas-data-frame)

Easier Hint:

Keep in mind, a simple way to convert ratings:

- 100/20 = 5
- 10/2 = 5

In [50]:

```
df['RT_Norm'] = (df['RottenTomatoes']/20).round(1)
df['RTU_Norm'] = (df['RottenTomatoes_User']/20).round(1)
```

In [55]:

```
df['Meta_Norm'] = (df['Metacritic']/20).round(1)
df['Meta_U_Norm'] = (df['Metacritic_User']/2).round(1)
```

In [56]:

```
df['IMDB_Norm'] = (df['IMDB']/2).round(1)
```

In [57]:

```
df.head()
```

Out[57]:

	FILM	STARS	RATING	VOTES	YEAR	STARS_DIFF	RottenTomatoes	RottenTomatoes_U
0	Fifty Shades of Grey (2015)	4.0	3.9	34846.0	2015	0.1	25	
1	Jurassic World (2015)	4.5	4.5	34390.0	2015	0.0	71	
2	American Sniper (2015)	5.0	4.8	34085.0	2015	0.2	72	
3	Furious 7 (2015)	5.0	4.8	33538.0	2015	0.2	81	
4	Inside Out (2015)	4.5	4.5	15749.0	2015	0.0	98	

In [238]:

```

```

Out[238]:

	FILM	STARS	RATING	VOTES	YEAR	RottenTomatoes	RottenTomatoes_User	Metacritic
0	Fifty Shades of Grey (2015)	4.0	3.9	34846	2015	25	42	46
1	Jurassic World (2015)	4.5	4.5	34390	2015	71	81	59
2	American Sniper (2015)	5.0	4.8	34085	2015	72	85	72
3	Furious 7 (2015)	5.0	4.8	33538	2015	81	84	67
4	Inside Out (2015)	4.5	4.5	15749	2015	98	90	94

TASK: Now create a norm_scores DataFrame that only contains the normalizes ratings. Include both STARS and RATING from the original Fandango table.

In [60]:

```
new_df = df[['STARS','RATING','RT_Norm','RTU_Norm','Meta_Norm','Meta_U_Norm','IMDB_Norm']]
```

In [61]:

```
new_df.head()
```

Out[61]:

	STARS	RATING	RT_Norm	RTU_Norm	Meta_Norm	Meta_U_Norm	IMDB_Norm
0	4.0	3.9	1.2	2.1	2.3	1.6	2.1
1	4.5	4.5	3.6	4.0	3.0	3.5	3.6
2	5.0	4.8	3.6	4.2	3.6	3.3	3.7
3	5.0	4.8	4.0	4.2	3.4	3.4	3.7
4	4.5	4.5	4.9	4.5	4.7	4.4	4.3

In [241]:

```

```

Out[241]:

	STARS	RATING	RT_Norm	RTU_Norm	Meta_Norm	Meta_U_Norm	IMDB_Norm
0	4.0	3.9	1.2	2.1	2.3	1.6	2.1
1	4.5	4.5	3.6	4.0	3.0	3.5	3.6
2	5.0	4.8	3.6	4.2	3.6	3.3	3.7
3	5.0	4.8	4.0	4.2	3.4	3.4	3.7
4	4.5	4.5	4.9	4.5	4.7	4.4	4.3

Comparing Distribution of Scores Across Sites

Now the moment of truth! Does Fandango display abnormally high ratings? We already know it pushs displayed RATING higher than STARS, but are the ratings themselves higher than average?

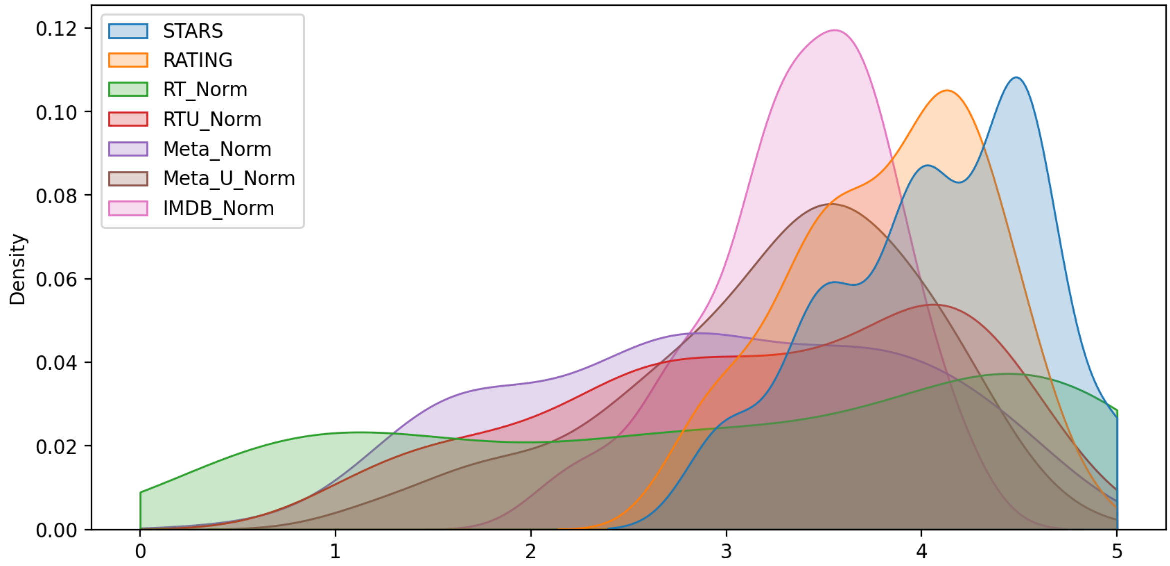
TASK: Create a plot comparing the distributions of normalized ratings across all sites. There are many ways to do this, but explore the Seaborn KDEplot docs for some simple ways to quickly show this. Don't worry if your plot format does not look exactly the same as ours, as long as the differences in distribution are clear.

Quick Note if you have issues moving the legend for a seaborn kdeplot: [https://github.com/mwaskom/seaborn/issues/2280 (https://github.com/mwaskom/seaborn/issues/2280)](https://github.com/mwaskom/seaborn/issues/2280)

In [83]:

```
def move_legend(ax, new_loc, **kws):
    old_legend = ax.legend_
    handles = old_legend.legendHandles
    labels = [t.get_text() for t in old_legend.get_texts()]
    title = old_legend.get_title().get_text()
    ax.legend(handles, labels, loc=new_loc, title=title, **kws)

fig,ax = plt.subplots(figsize = (10,5),dpi = 200)
sns.kdeplot(data = new_df,shade = True,clip=[0,5])
move_legend(ax,"upper left")
```

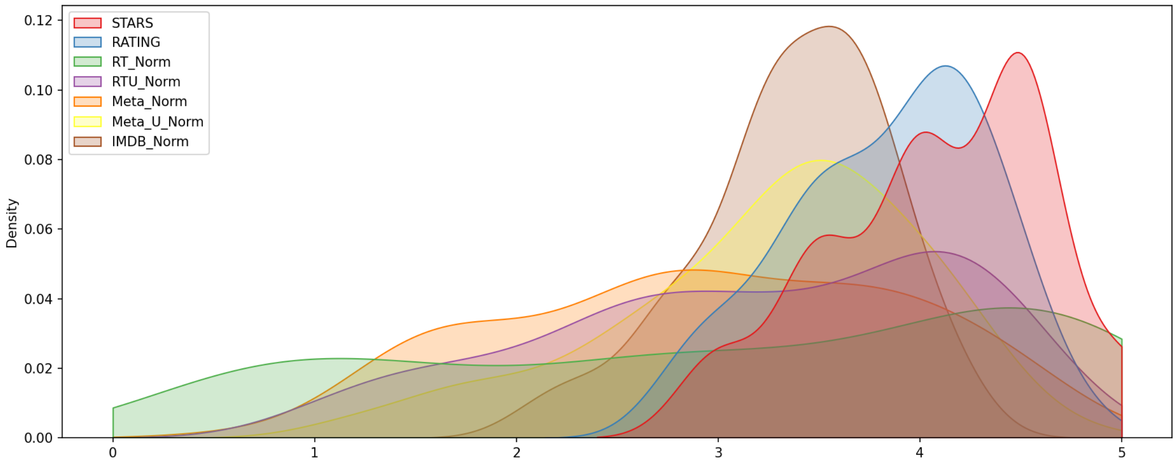


In [ ]:

In [244]:

```

```



Clearly Fandango has an uneven distribution. We can also see that RT critics have the most uniform distribution. Let's directly compare these two.

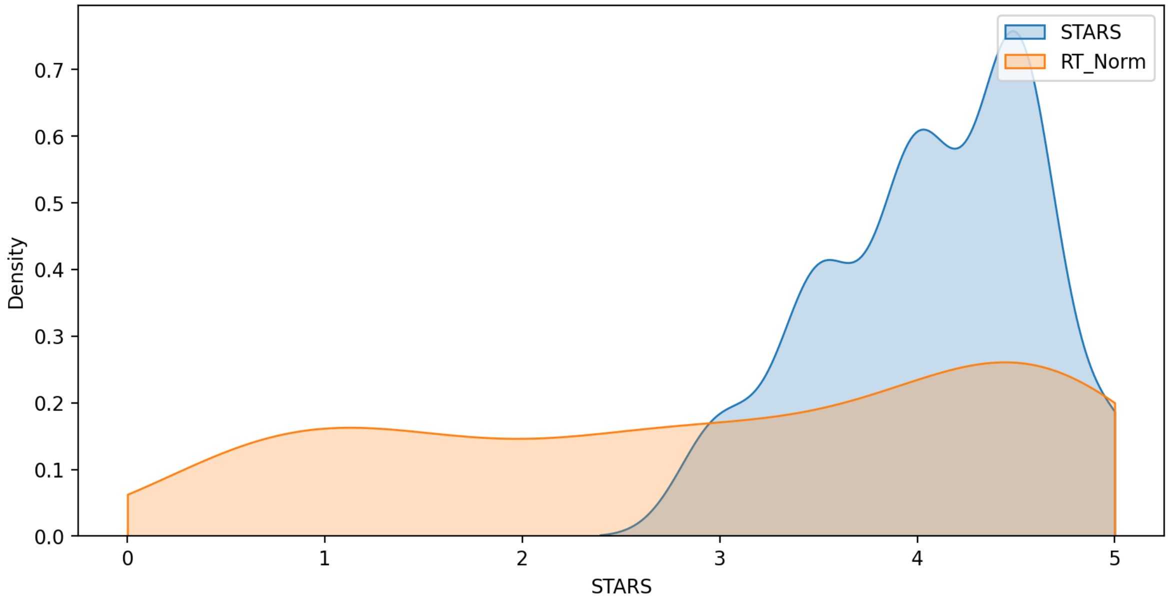
TASK: Create a KDE plot that compare the distribution of RT critic ratings against the STARS displayed by Fandango.

In [90]:

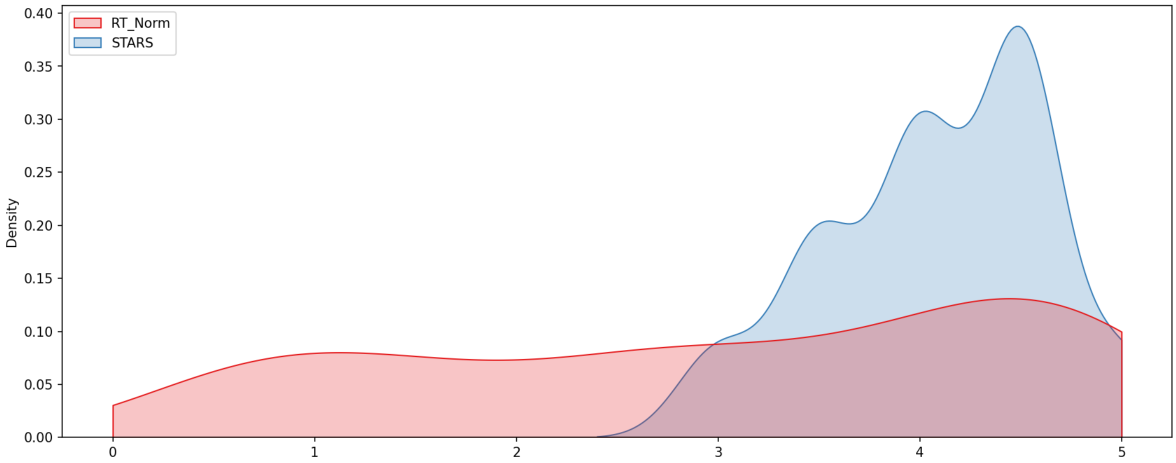
```
plt.figure(figsize=(10,5),dpi = 200)
sns.kdeplot(data = new_df,x = 'STARS',shade = True,label = 'STARS',clip = [0,5])
sns.kdeplot(data = new_df,x = 'RT_Norm',shade = True,label = 'RT_Norm',clip = [0,5])
plt.legend()
```

Out[90]:

```
<matplotlib.legend.Legend at 0x1ab2be1c4c0>
```



In [167]:



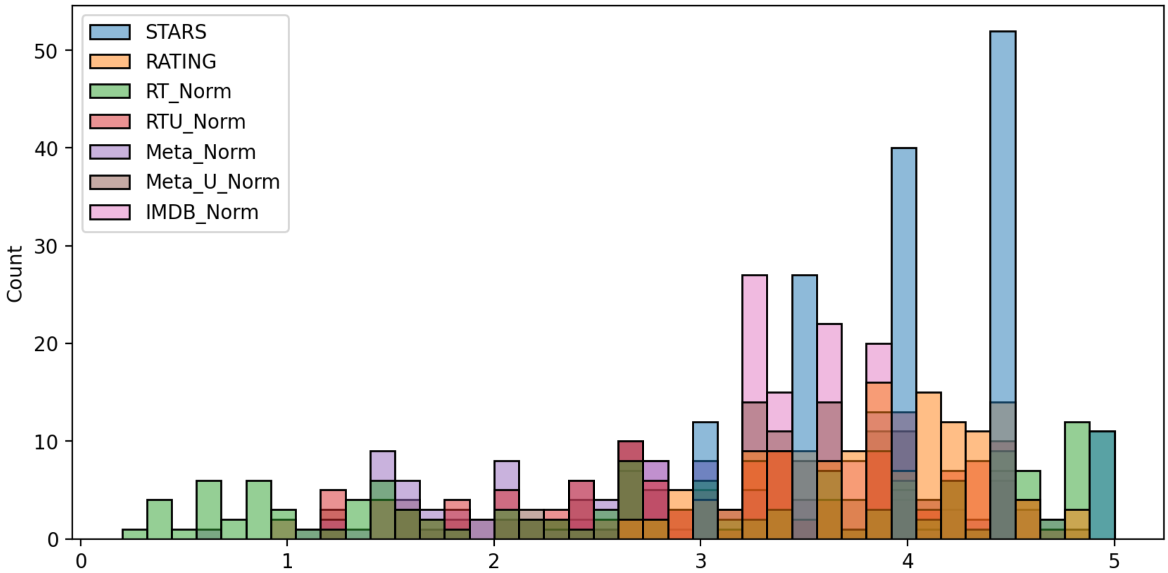
OPTIONAL TASK: Create a histplot comparing all normalized scores.

In [97]:

```
plt.figure(figsize = (10,5),dpi = 200)
sns.histplot(data = new_df,bins = 40)
```

Out[97]:

<AxesSubplot:ylabel='Count'>



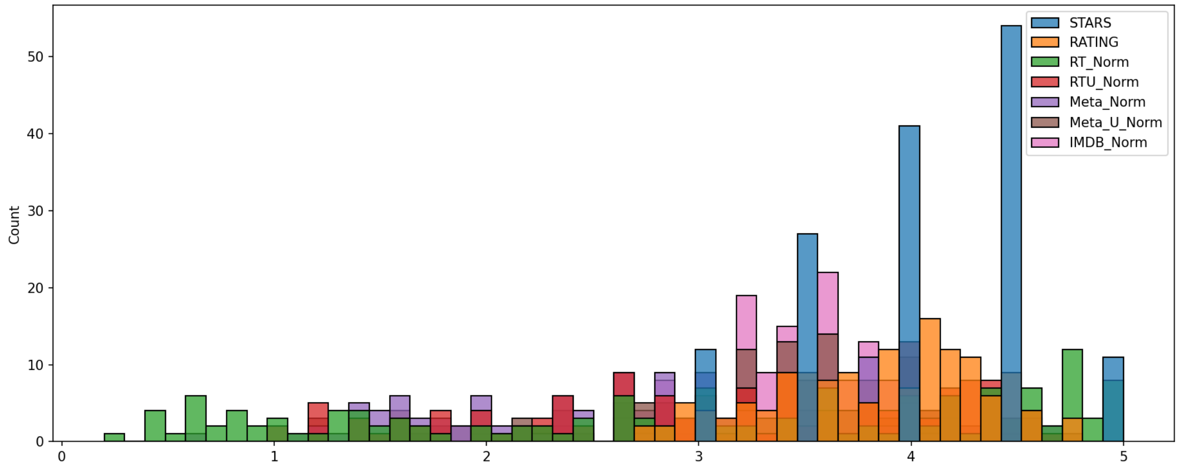
In [168]:

```

```

Out[168]:

```
<AxesSubplot:ylabel='Count'>
```



How are the worst movies rated across all platforms?

TASK: Create a clustermap visualization of all normalized scores. Note the differences in ratings, highly rated movies should be clustered together versus poorly rated movies. Note: This clustermap does not need to have the FILM titles as the index, feel free to drop it for the clustermap.

In [ ]:

```
# CODE HERE
```

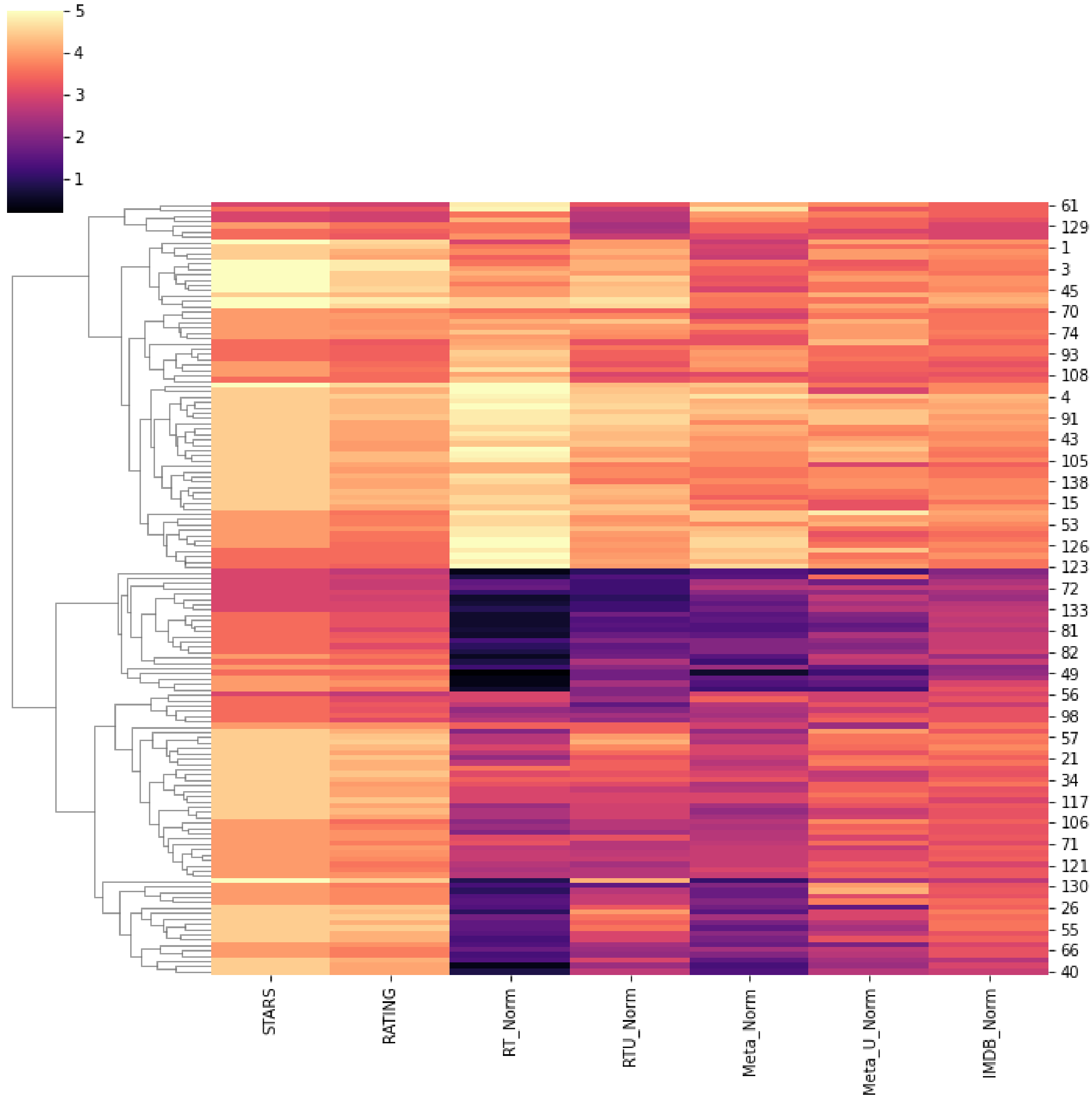
In [169]:

```

```

Out[169]:

<seaborn.matrix.ClusterGrid at 0x1aa7cb2b548>



TASK: Clearly Fandango is rating movies much higher than other sites, especially considering that it is then displaying a rounded up version of the rating. Let's examine the top 10 worst movies. Based off the Rotten Tomatoes Critic Ratings, what are the top 10 lowest rated movies? What are the normalized scores across all platforms for these movies? You may need to add the FILM column back in to your DataFrame of normalized scores to see the results.

In [104]:

```
new_df1 = df[['STARS','RATING','RT_Norm','RTU_Norm','Meta_Norm','Meta_U_Norm','IMDB_Norm','
```

In [107]:

```
temp = new_df1.nsmallest(10,'RT_Norm')
temp
```

Out[107]:

	STARS	RATING	RT_Norm	RTU_Norm	Meta_Norm	Meta_U_Norm	IMDB_Norm	FILM
49	3.5	3.5	0.2	1.8	0.6	1.2	2.2	Paul Blart: Mall Cop 2 (2015)
25	4.5	4.1	0.4	2.3	1.3	2.3	3.0	Taken 3 (2015)
28	3.0	2.7	0.4	1.0	1.4	1.2	2.0	Fantastic Four (2015)
54	4.0	3.7	0.4	1.8	1.6	1.8	2.4	Hot Pursuit (2015)
84	4.0	3.9	0.4	2.4	1.4	1.6	3.0	Hitman: Agent 47 (2015)
50	4.0	3.6	0.5	1.8	1.5	2.8	2.3	The Boy Next Door (2015)
77	3.5	3.2	0.6	1.8	1.5	2.0	2.8	Seventh Son (2015)
78	3.5	3.2	0.6	1.5	1.4	1.6	2.8	Mortdecai (2015)
83	3.5	3.3	0.6	1.7	1.6	2.5	2.8	Sinister 2 (2015)
87	3.5	3.2	0.6	1.4	1.6	1.9	2.7	Unfinished Business (2015)

In [248]:

```

```

Out[248]:

	STARS	RATING	RT_Norm	RTU_Norm	Meta_Norm	Meta_U_Norm	IMDB_Norm	FILM
49	3.5	3.5	0.2	1.8	0.6	1.2	2.2	Paul Blart: Mall Cop 2 (2015)
25	4.5	4.1	0.4	2.3	1.3	2.3	3.0	Taken 3 (2015)
28	3.0	2.7	0.4	1.0	1.4	1.2	2.0	Fantastic Four (2015)
54	4.0	3.7	0.4	1.8	1.6	1.8	2.4	Hot Pursuit (2015)
84	4.0	3.9	0.4	2.4	1.4	1.6	3.0	Hitman: Agent 47 (2015)
50	4.0	3.6	0.5	1.8	1.5	2.8	2.3	The Boy Next Door (2015)
77	3.5	3.2	0.6	1.8	1.5	2.0	2.8	Seventh Son (2015)
78	3.5	3.2	0.6	1.5	1.4	1.6	2.8	Mortdecai (2015)
83	3.5	3.3	0.6	1.7	1.6	2.5	2.8	Sinister 2 (2015)
87	3.5	3.2	0.6	1.4	1.6	1.9	2.7	Unfinished Business (2015)

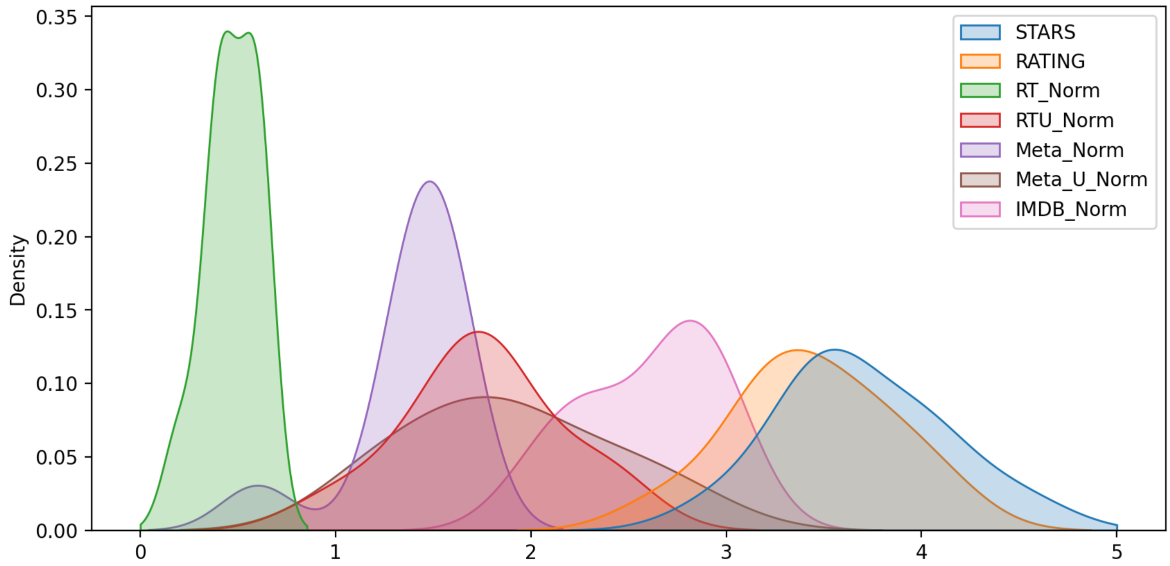
FINAL TASK: Visualize the distribution of ratings across all sites for the top 10 worst movies.

In [112]:

```
temp.drop('FILM',axis = 1)
plt.figure(figsize=(10,5),dpi = 200)
sns.kdeplot(data = temp,clip = [0,5],shade = True)
```

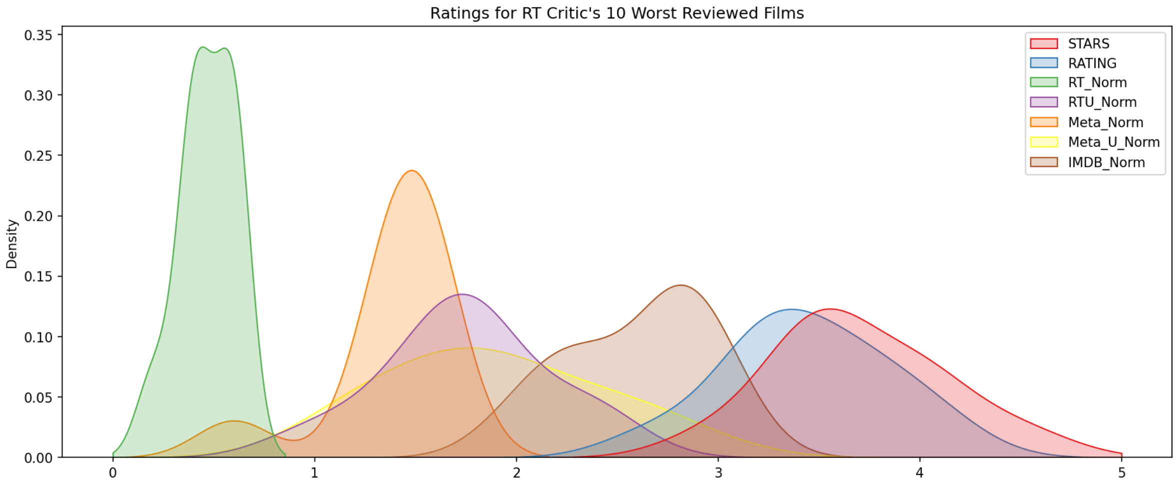
Out[112]:

```
<AxesSubplot:ylabel='Density'>
```



In [251]:

```
```


Ratings for RT Critic's 10 Worst Reviewed Films
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
STARS
RATING
RT_Norm
RTU_Norm
Meta_Norm
Meta_U_Norm
IMDB_Norm


IT ENDS HERE
LIAM NEESON
TAKEN 3
JANUARY 9

Final thoughts: Wow! Fandango is showing around 3-4 star ratings for films that are clearly bad! Notice the biggest offender, [Taken 3! (https://www.youtube.com/watch?v=tJrfImRCHJ0)](https://www.youtube.com/watch?v=tJrfImRCHJ0). Fandango is displaying 4.5 stars on their site for a film with an [average rating of 1.86 (https://en.wikipedia.org/wiki/Taken_3#Critical_response)](https://en.wikipedia.org/wiki/Taken_3#Critical_response) across the other platforms!

In [253]:

```

```

Out[253]:

```
STARS                    4.5
RATING                   4.1
RT_Norm                  0.4
RTU_Norm                 2.3
Meta_Norm                1.3
Meta_U_Norm              2.3
IMDB_Norm                  3
FILM           Taken 3 (2015)
Name: 25, dtype: object
```

In [254]:

```
0.4+2.3+1.3+2.3+3
```

Out[254]:

```
9.3
```

In [255]:

```
9.3/5
```

Out[255]:

```
1.86
```