k-means Clustering of Movie Ratings

June 26, 2018

1 k-means Clustering of Movie Ratings

Say you're a data analyst at Netflix and you want to explore the similarities and differences in people's tastes in movies based on how they rate different movies. Can understanding these ratings contribute to a movie recommendation system for users? Let's dig into the data and see.

The data we'll be using comes from the wonderful MovieLens user rating dataset. We'll be looking at individual movie ratings later in the notebook, but let us start with how ratings of genres compare to each other.

1.1 Dataset overview

The dataset has two files. We'll import them both into pandas dataframes:

```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        import numpy as np
        from scipy.sparse import csr_matrix
        import helper

        # Import the Movies dataset
        movies = pd.read_csv('ml-latest-small/movies.csv')
        movies.head()
```

```
Out[1]:    movieId                               title  \
        0        1                    Toy Story (1995)
        1        2                      Jumanji (1995)
        2        3             Grumpier Old Men (1995)
        3        4            Waiting to Exhale (1995)
        4        5  Father of the Bride Part II (1995)

                                            genres
        0  Adventure|Animation|Children|Comedy|Fantasy
        1                   Adventure|Children|Fantasy
        2                               Comedy|Romance
        3                         Comedy|Drama|Romance
        4                                       Comedy
```

```
In [2]: # Import the ratings dataset
        ratings = pd.read_csv('ml-latest-small/ratings.csv')
        ratings.head()
```

Out[2]:

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205

Now that we know the structure of our dataset, how many records do we have in each of these tables?

```
In [3]: print('The dataset contains: ', len(ratings), ' ratings of ', len(movies), ' movies.')
```

```
The dataset contains:  100004  ratings of  9125  movies.
```

1.2 Romance vs. Scifi

Let's start by taking a subset of users, and seeing what their preferred genres are. We're hiding the most data preprocessing in helper functions so the focus is on the topic of clustering. It would be useful if you skim helper.py to see how these helper functions are implemented after finishing this notebook.

```
In [4]: # Calculate the average rating of romance and scifi movies

        genre_ratings = helper.get_genre_ratings(ratings, movies, ['Romance', 'Sci-Fi'], ['avg_r
        genre_ratings.head()
```

Out[4]:

userId	avg_romance_rating	avg_scifi_rating
1	3.50	2.40
2	3.59	3.80
3	3.65	3.14
4	4.50	4.26
5	4.08	4.00

The function `get_genre_ratings` calculated each user's average rating of all romance movies and all scifi movies. Let's bias our dataset a little by removing people who like both scifi and romance, just so that our clusters tend to define them as liking one genre more than the other.

```
In [5]: biased_dataset = helper.bias_genre_rating_dataset(genre_ratings, 3.2, 2.5)

        print( "Number of records: ", len(biased_dataset))
        biased_dataset.head()
```

```
Number of records:  183
```

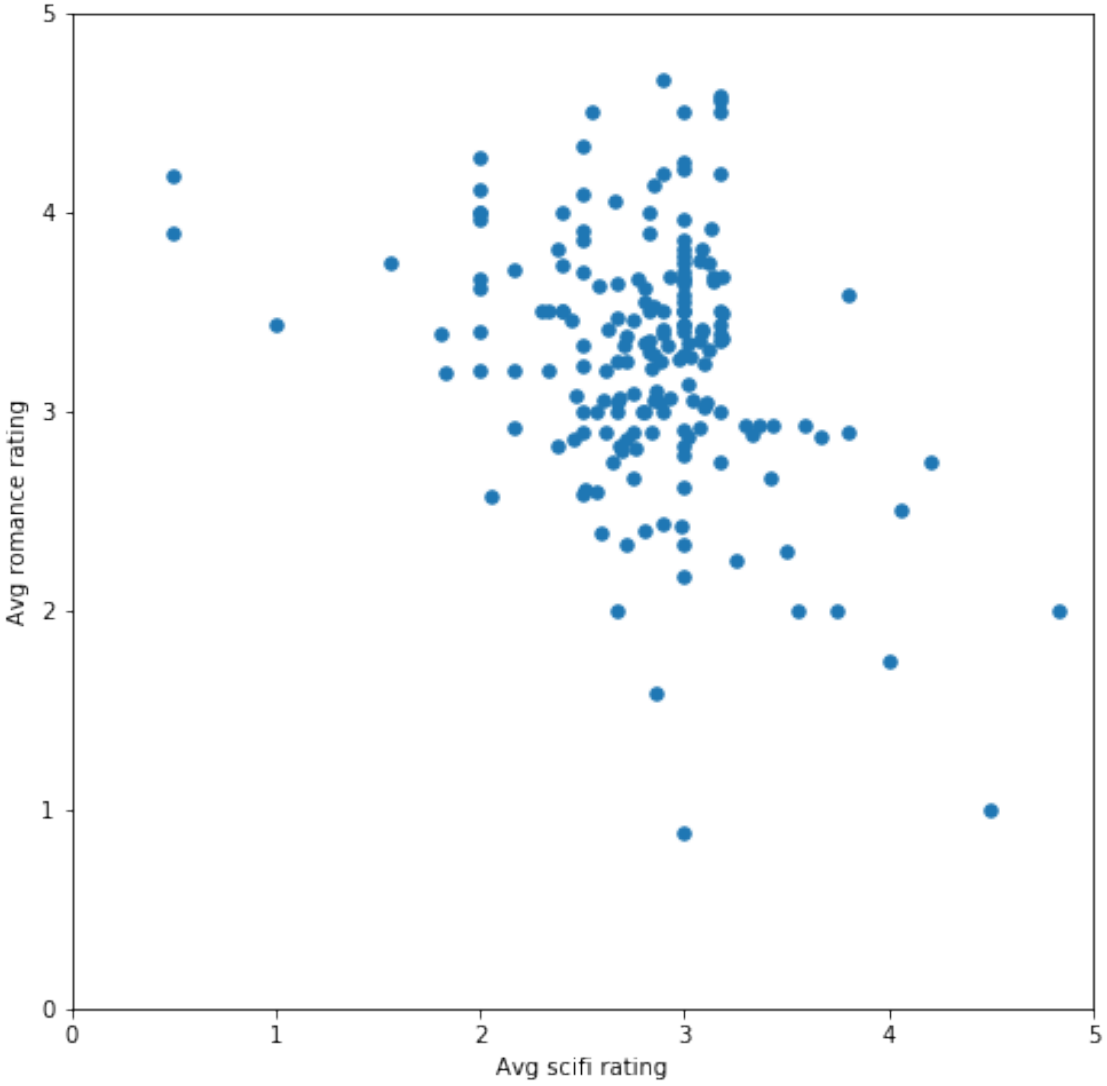
Out[5]:

	userId	avg_romance_rating	avg_scifi_rating
0	1	3.50	2.40
1	3	3.65	3.14
2	6	2.90	2.75
3	7	2.93	3.36
4	12	2.89	2.62

So we can see we have 183 users, and for each user we have their average rating of the romance and sci movies they've watched.

Let us plot this dataset:

In [6]:
```
%matplotlib inline

helper.draw_scatterplot(biased_dataset['avg_scifi_rating'],'Avg scifi rating', biased_da
```

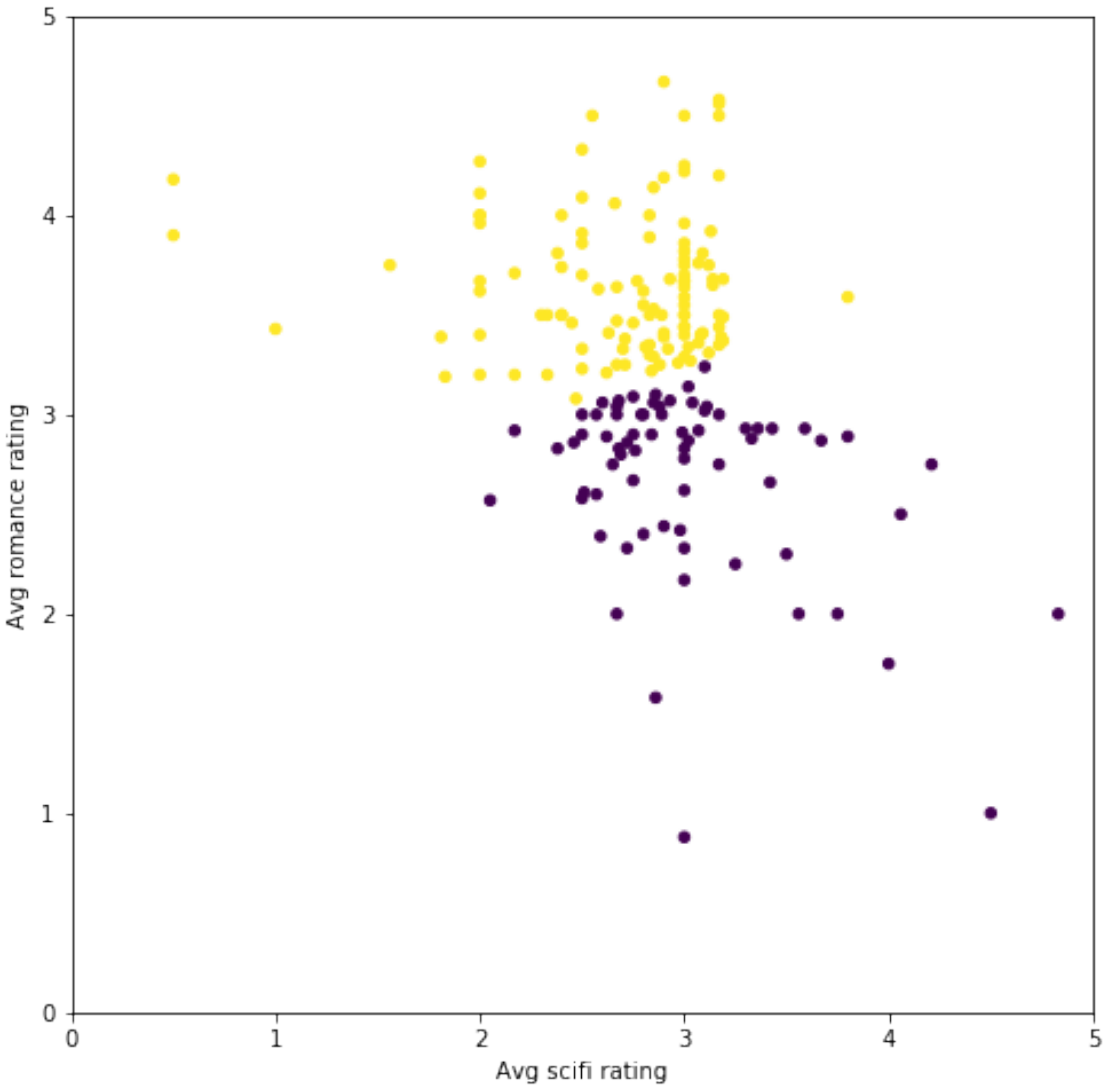


We can see come clear bias in this sample (that we created on purpose). How would it look if we break the sample down into two groups using k-means?

```
In [11]: # Let's turn our dataset into a list
         X = biased_dataset[['avg_scifi_rating','avg_romance_rating']].values
```

- Import KMeans
- Prepare KMeans with n_clusters = 2
- Pass the dataset **X** to KMeans' fit_predict method and retrieve the clustering labels into *predictions*

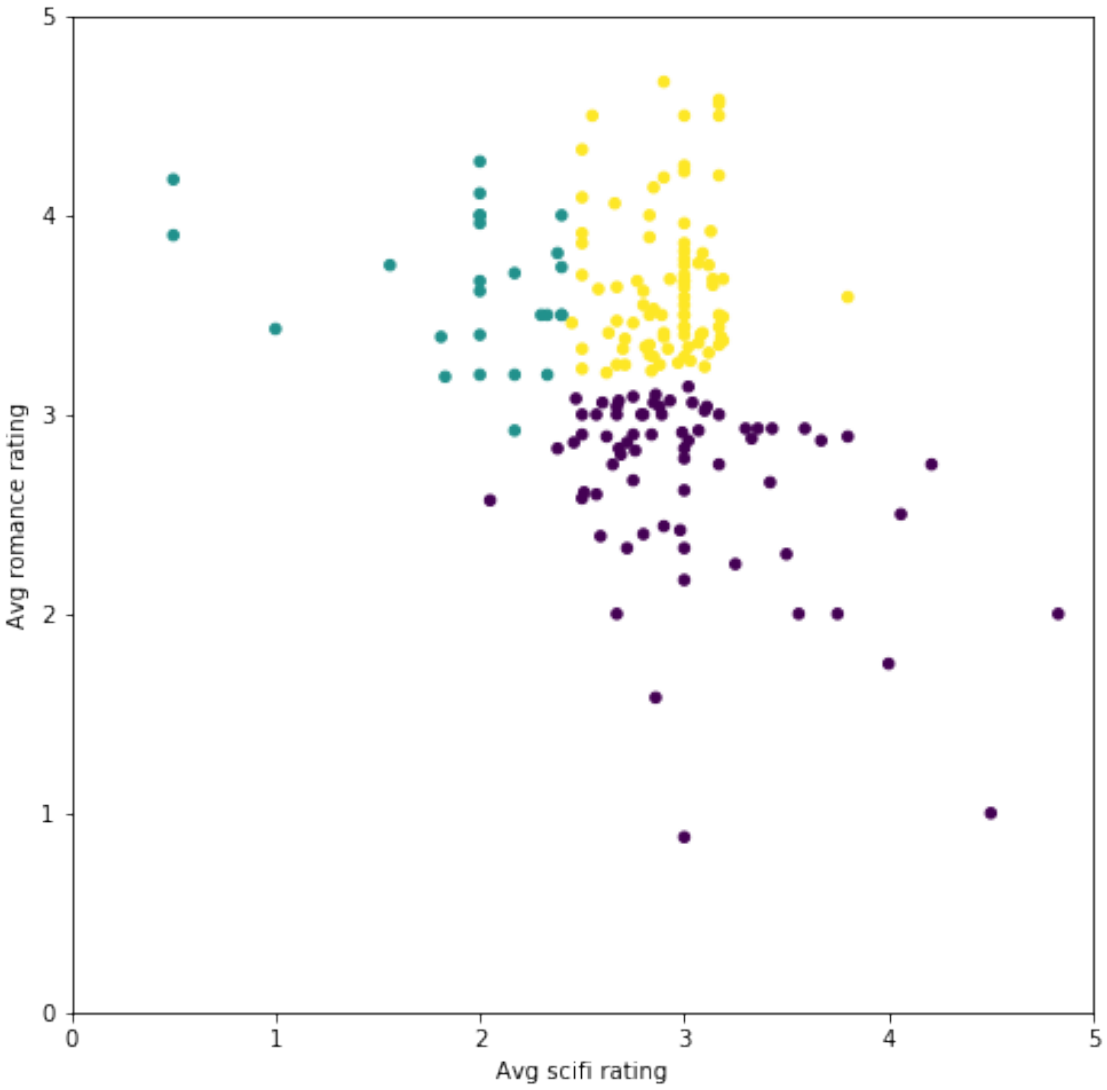
```
In [19]: # TODO: Import KMeans
         from sklearn.cluster import KMeans

         # TODO: Create an instance of KMeans to find two clusters
         kmeans_1 = KMeans( n_clusters = 2, n_init =30)

         # TODO: use fit_predict to cluster the dataset
         predictions = kmeans_1.fit_predict(X)

         # Plot
         helper.draw_clusters(biased_dataset, predictions)
```



We can see that the groups are mostly based on how each person rated romance movies. If their average rating of romance movies is over 3 stars, then they belong to one group. Otherwise, they belong to the other group.

What would happen if we break them down into three groups?

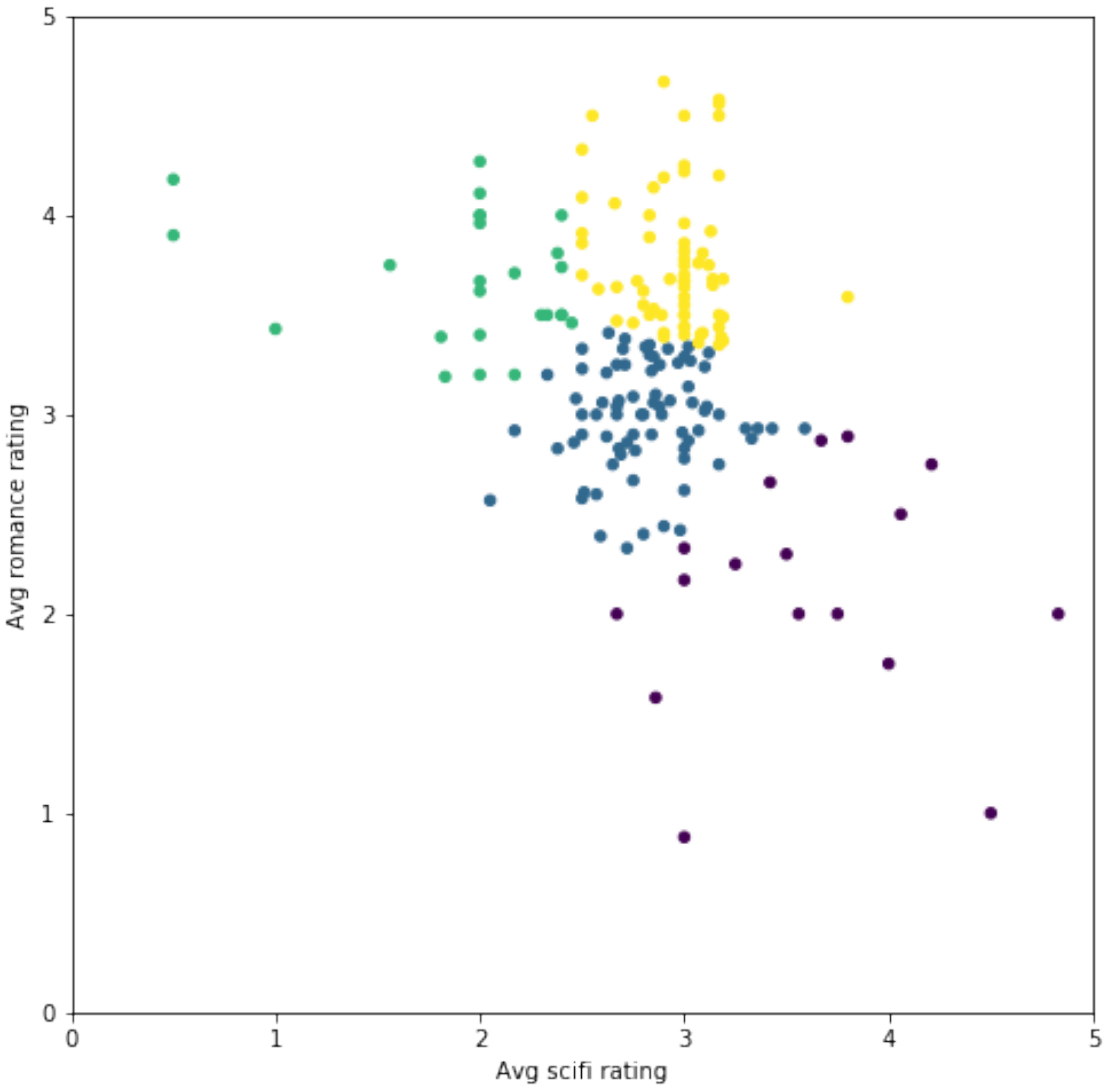
```
In [22]: # TODO: Create an instance of KMeans to find three clusters
         kmeans_2 = KMeans(n_clusters = 3)

         # TODO: use fit_predict to cluster the dataset
         predictions_2 = kmeans_2.fit_predict(X)

         # Plot
         helper.draw_clusters(biased_dataset, predictions_2)
```



Now the average scifi rating is starting to come into play. The groups are: * people who like romance but not scifi * people who like scifi but not romance * people who like both scifi and romance

Let's add one more group

```
In [23]: # TODO: Create an instance of KMeans to find four clusters
         kmeans_3 = KMeans(n_clusters =4 )

         # TODO: use fit_predict to cluster the dataset
         predictions_3 = kmeans_3.fit_predict(X)

         # Plot
         helper.draw_clusters(biased_dataset, predictions_3)
```



We can see that the more clusters we break our dataset down into, the more similar the tastes of the population of each cluster to each other.

1.3 Choosing K

Great, so we can cluster our points into any number of clusters. What's the right number of clusters for this dataset?

There are several ways of choosing the number of clusters, k. We'll look at a simple one called "the elbow method". The elbow method works by plotting the ascending values of k versus the total error calculated using that k.

How do we calculate total error? One way to calculate the error is squared error. Say we're calculating the error for k=2. We'd have two clusters each having one "centroid" point. For each point in our dataset, we'd subtract its coordinates from the centroid of the cluster it belongs to. We then square the result of that subtraction (to get rid of the negative values), and sum the values.

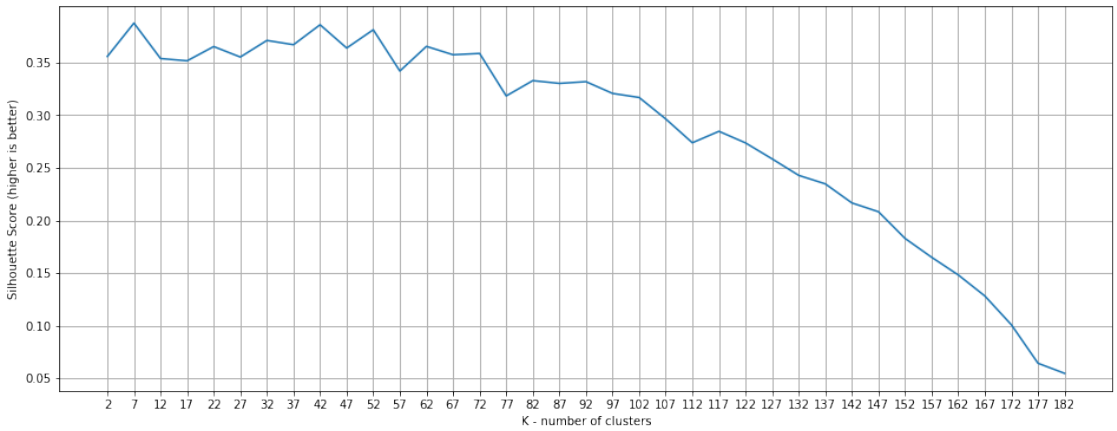
This would leave us with an error value for each point. If we sum these error values, we'd get the total error for all points when k=2.

Our mission now is to do the same for each k (between 1 and, say, the number of elements in our dataset)

```
In [24]: # Choose the range of k values to test.
         # We added a stride of 5 to improve performance. We don't need to calculate the error f
         possible_k_values = range(2, len(X)+1, 5)

         # Calculate error values for all k values we're interested in
         errors_per_k = [helper.clustering_errors(k, X) for k in possible_k_values]

In [25]: # Optional: Look at the values of K vs the silhouette score of running K-means with tha
         list(zip(possible_k_values, errors_per_k))

Out[25]: [(2, 0.35588178764728268),
          (7, 0.38761257576589386),
          (12, 0.35390967793441436),
          (17, 0.35174387289992504),
          (22, 0.36529470541241388),
          (27, 0.35532881339368294),
          (32, 0.371121228513068),
          (37, 0.36704249292489327),
          (42, 0.38602306594022961),
          (47, 0.36389008769471737),
          (52, 0.38127007839649324),
          (57, 0.34202460913244365),
          (62, 0.36547622352996068),
          (67, 0.35751273821516916),
          (72, 0.35876538000811714),
          (77, 0.3183652056534455),
          (82, 0.33290578008894822),
          (87, 0.33028011501959464),
          (92, 0.33190488605195823),
          (97, 0.32074799440095264),
          (102, 0.31687697978205726),
          (107, 0.29644718071806836),
          (112, 0.2738439924808409),
          (117, 0.2847341602945056),
          (122, 0.27371019906509997),
          (127, 0.25851481264382214),
          (132, 0.24285659085330791),
          (137, 0.23476827629600011),
          (142, 0.21664844402715644),
          (147, 0.20826924160850682),
          (152, 0.18288943960429158),
          (157, 0.16496749756636125),
          (162, 0.14820613734314372),
```

```
 (167, 0.12820427579529109),
 (172, 0.10075966098920461),
 (177, 0.064230120163174503),
 (182, 0.054644808743169397)]
```

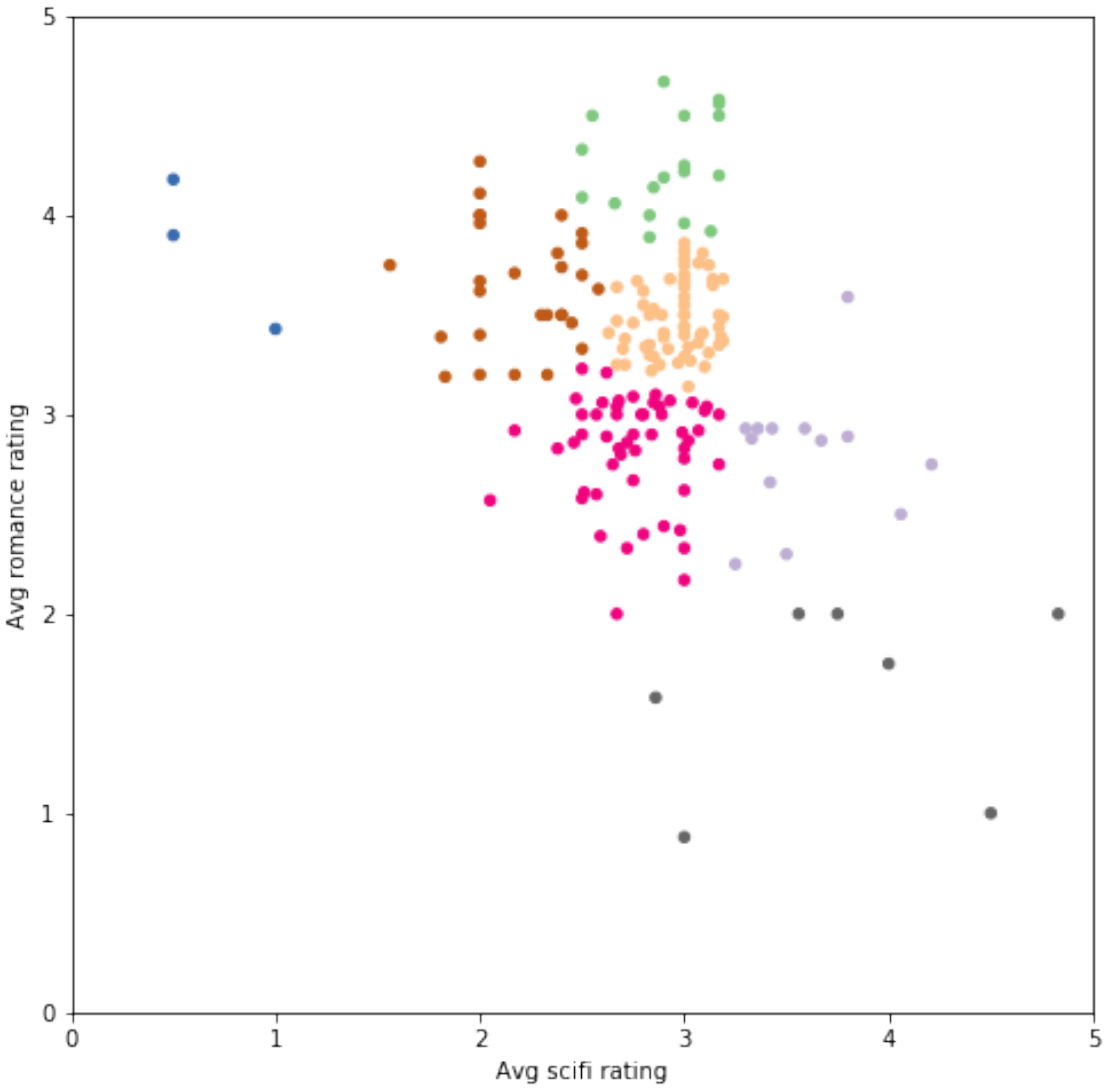
In [26]:
```
# Plot the each value of K vs. the silhouette score at that value
fig, ax = plt.subplots(figsize=(16, 6))
ax.set_xlabel('K - number of clusters')
ax.set_ylabel('Silhouette Score (higher is better)')
ax.plot(possible_k_values, errors_per_k)

# Ticks and grid
xticks = np.arange(min(possible_k_values), max(possible_k_values)+1, 5.0)
ax.set_xticks(xticks, minor=False)
ax.set_xticks(xticks, minor=True)
ax.xaxis.grid(True, which='both')
yticks = np.arange(round(min(errors_per_k), 2), max(errors_per_k), .05)
ax.set_yticks(yticks, minor=False)
ax.set_yticks(yticks, minor=True)
ax.yaxis.grid(True, which='both')
```



Looking at this graph, good choices for k include 7, 22, 27, 32, amongst other values (with a slight variation between different runs). Increasing the number of clusters (k) beyond that range starts to result in worse clusters (according to Silhouette score)

My pick would be k=7 because it's easier to visualize:

In [28]:
```
# TODO: Create an instance of KMeans to find seven clusters
kmeans_4 = KMeans(n_clusters = 7)

# TODO: use fit_predict to cluster the dataset
predictions_4 = kmeans_4.fit_predict(X)

# plot
helper.draw_clusters(biased_dataset, predictions_4, cmap='Accent')
```



Note: As you try to plot larger values of k (more than 10), you'll have to make sure your plotting library is not reusing colors between clusters. For this plot, we had to use the matplotlib colormap 'Accent' because other colormaps either did not show enough contrast between colors, or were recycling colors past 8 or 10 clusters.

1.4 Throwing some Action into the mix

So far, we've only been looking at how users rated romance and scifi movies. Let's throw another genre into the mix. Let's add the Action genre.

Our dataset now looks like this:

```
In [29]: biased_dataset_3_genres = helper.get_genre_ratings(ratings, movies,
                                                            ['Romance', 'Sci-Fi', 'Action'],
                                                            ['avg_romance_rating', 'avg_scifi_
```

```
biased_dataset_3_genres = helper.bias_genre_rating_dataset(biased_dataset_3_genres, 3.2

print( "Number of records: ", len(biased_dataset_3_genres))
biased_dataset_3_genres.head()
```

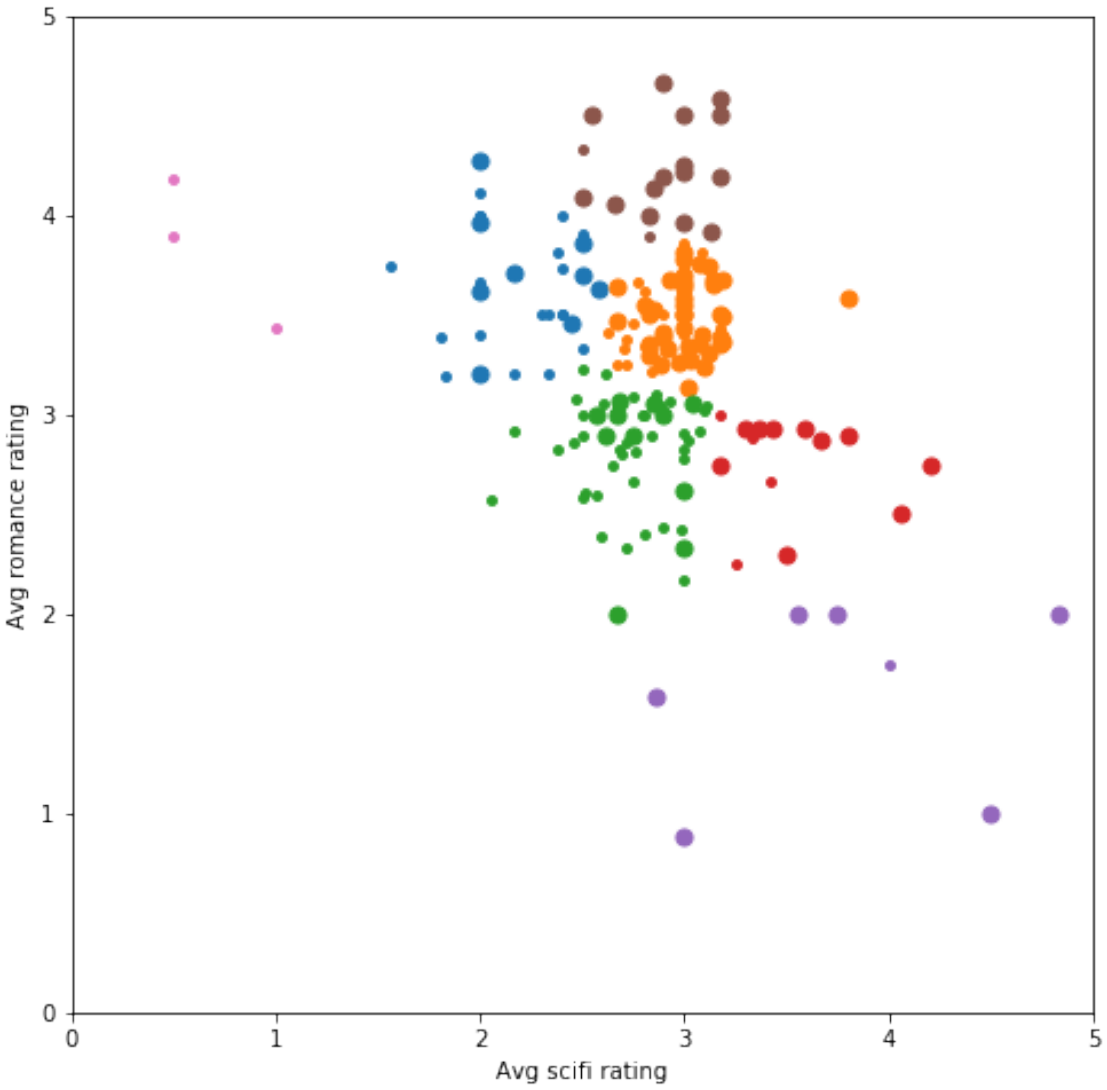
```
Number of records:  183
```

Out[29]:

	userId	avg_romance_rating	avg_scifi_rating	avg_action_rating
0	1	3.50	2.40	2.80
1	3	3.65	3.14	3.47
2	6	2.90	2.75	3.27
3	7	2.93	3.36	3.29
4	12	2.89	2.62	3.21

In [30]:
```
X_with_action = biased_dataset_3_genres[['avg_scifi_rating',
                                         'avg_romance_rating',
                                         'avg_action_rating']].values
```

In [32]:
```
# TODO: Create an instance of KMeans to find seven clusters
kmeans_5 = KMeans(n_clusters = 7)

# TODO: use fit_predict to cluster the dataset
predictions_5 = kmeans_5.fit_predict(X)

# plot
helper.draw_clusters_3d(biased_dataset_3_genres, predictions_5)
```



We're still using the x and y axes for scifi and romance respectively. We are using the size of the dot to roughly code the 'action' rating (large dot for avg ratings over than 3, small dot otherwise).

We can start seeing the added genre is changing how the users are clustered. The more data we give to k-means, the more similar the tastes of the people in each group would be. Unfortunately, though, we lose the ability to visualize what's going on past two or three dimensions if we continue to plot it this way. In the next section, we'll start using a different kind of plot to be able to see clusters with up to fifty dimensions.

1.5 Movie-level Clustering

Now that we've established some trust in how k-means clusters users based on their genre tastes, let's take a bigger bite and look at how users rated individual movies. To do that, we'll shape the dataset in the form of userId vs user rating for each movie. For example, let's look at a subset of the dataset:

In [33]:
```
# Merge the two tables then pivot so we have Users X Movies dataframe
ratings_title = pd.merge(ratings, movies[['movieId', 'title']], on='movieId' )
user_movie_ratings = pd.pivot_table(ratings_title, index='userId', columns= 'title', va

print('dataset dimensions: ', user_movie_ratings.shape, '\n\nSubset example:')
user_movie_ratings.iloc[:6, :10]
```

```
dataset dimensions:  (671, 9064)

Subset example:
```

Out[33]:

title	"Great Performances" Cats (1998)	$9.99 (2008)
userId		
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN
5	NaN	NaN
6	NaN	NaN

title	'Hellboy': The Seeds of Creation (2004)
userId	
1	NaN
2	NaN
3	NaN
4	NaN
5	NaN
6	NaN

title	'Neath the Arizona Skies (1934)	'Round Midnight (1986)
userId		
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN
5	NaN	NaN
6	NaN	NaN

title	'Salem's Lot (2004)	'Til There Was You (1997)	'burbs, The (1989)
userId			
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	NaN	NaN	NaN
4	NaN	NaN	NaN
5	NaN	NaN	NaN
6	NaN	NaN	4.0

```
title   'night Mother (1986)  (500) Days of Summer (2009)
userId
1                         NaN                          NaN
2                         NaN                          NaN
3                         NaN                          NaN
4                         NaN                          NaN
5                         NaN                          NaN
6                         NaN                          NaN
```

The dominance of NaN values presents the first issue. Most users have not rated and watched most movies. Datasets like this are called "sparse" because only a small number of cells have values.

To get around this, let's sort by the most rated movies, and the users who have rated the most number of movies. That will present a more 'dense' region when we peak at the top of the dataset.

If we're to choose the most-rated movies vs users with the most ratings, it would look like this:

```
In [34]: n_movies = 30
         n_users = 18
         most_rated_movies_users_selection = helper.sort_by_rating_density(user_movie_ratings, n

         print('dataset dimensions: ', most_rated_movies_users_selection.shape)
         most_rated_movies_users_selection.head()

dataset dimensions:  (18, 30)


Out[34]: title  Forrest Gump (1994)  Pulp Fiction (1994)  \
         29                     5.0                 5.0
         508                    4.0                 5.0
         14                     1.0                 5.0
         72                     5.0                 5.0
         653                    4.0                 5.0

         title  Shawshank Redemption, The (1994)  Silence of the Lambs, The (1991)  \
         29                                  5.0                               4.0
         508                                 4.0                               4.0
         14                                  2.0                               5.0
         72                                  5.0                               4.5
         653                                 5.0                               4.5

         title  Star Wars: Episode IV - A New Hope (1977)  Jurassic Park (1993)  \
         29                                           4.0                  4.0
         508                                          5.0                  3.0
         14                                           5.0                  3.0
         72                                           4.5                  4.0
         653                                          5.0                  4.5

         title  Matrix, The (1999)  Toy Story (1995)  Schindler's List (1993)  \
         29                    3.0               4.0                      5.0
```

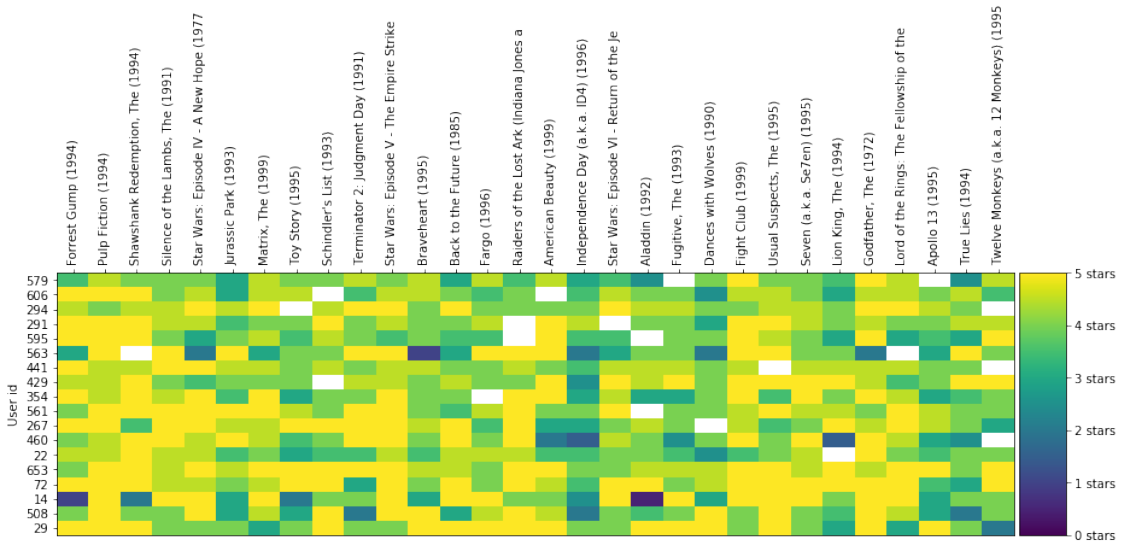
```
508                  4.5              3.0                    5.0
14                   5.0              2.0                    4.0
72                   4.5              5.0                    5.0
653                  5.0              5.0                    5.0

title  Terminator 2: Judgment Day (1991)  \
29                                   4.0
508                                  2.0
14                                   4.0
72                                   3.0
653                                  5.0

title                    ...                    Dances with Wolves (1990)  \
29                       ...                                          5.0
508                      ...                                          5.0
14                       ...                                          3.0
72                       ...                                          4.5
653                      ...                                          4.5

title  Fight Club (1999)  Usual Suspects, The (1995)  \
29                   4.0                         5.0
508                  4.0                         5.0
14                   5.0                         5.0
72                   5.0                         5.0
653                  5.0                         5.0

title  Seven (a.k.a. Se7en) (1995)  Lion King, The (1994)  \
29                             4.0                    3.0
508                            4.0                    3.5
14                             5.0                    4.0
72                             5.0                    5.0
653                            4.5                    5.0

title  Godfather, The (1972)  \
29                       5.0
508                      5.0
14                       5.0
72                       5.0
653                      4.5

title  Lord of the Rings: The Fellowship of the Ring, The (2001)  \
29                                                          3.0
508                                                         4.5
14                                                          5.0
72                                                          5.0
653                                                         5.0

title  Apollo 13 (1995)  True Lies (1994)  \
```

```
29                  5.0                 4.0
508                 3.0                 2.0
14                  3.0                 4.0
72                  3.5                 3.0
653                 5.0                 4.0

title  Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
29                                            2.0
508                                           4.0
14                                            4.0
72                                            5.0
653                                           5.0

[5 rows x 30 columns]
```

That's more like it. Let's also establish a good way for visualizing these ratings so we can attempt to visually recognize the ratings (and later, clusters) when we look at bigger subsets.

Let's use colors instead of the number ratings:

```
In [35]: helper.draw_movies_heatmap(most_rated_movies_users_selection)
```



Each column is a movie. Each row is a user. The color of the cell is how the user rated that movie based on the scale on the right of the graph.

Notice how some cells are white? This means the respective user did not rate that movie. This is an issue you'll come across when clustering in real life. Unlike the clean example we started with, real-world datasets can often be sparse and not have a value in each cell of the dataset. This makes it less straightforward to cluster users directly by their movie ratings as k-means generally does not like missing values.

For performance reasons, we'll only use ratings for 1000 movies (out of the 9000+ available in the dataset).

```
In [36]: user_movie_ratings =  pd.pivot_table(ratings_title, index='userId', columns= 'title', v
         most_rated_movies_1k = helper.get_most_rated_movies(user_movie_ratings, 1000)
```

To have sklearn run k-means clustering to a dataset with missing values like this, we will first cast it to the sparse csr matrix type defined in the SciPi library.

To convert from a pandas dataframe to a sparse matrix, we'll have to convert to SparseDataFrame, then use pandas' `to_coo()` method for the conversion.

Note: `to_coo()` was only added in later versions of pandas. If you run into an error with the next cell, make sure pandas is up to date.

```
In [37]: sparse_ratings = csr_matrix(pd.SparseDataFrame(most_rated_movies_1k).to_coo())
```

1.6 Let's cluster!

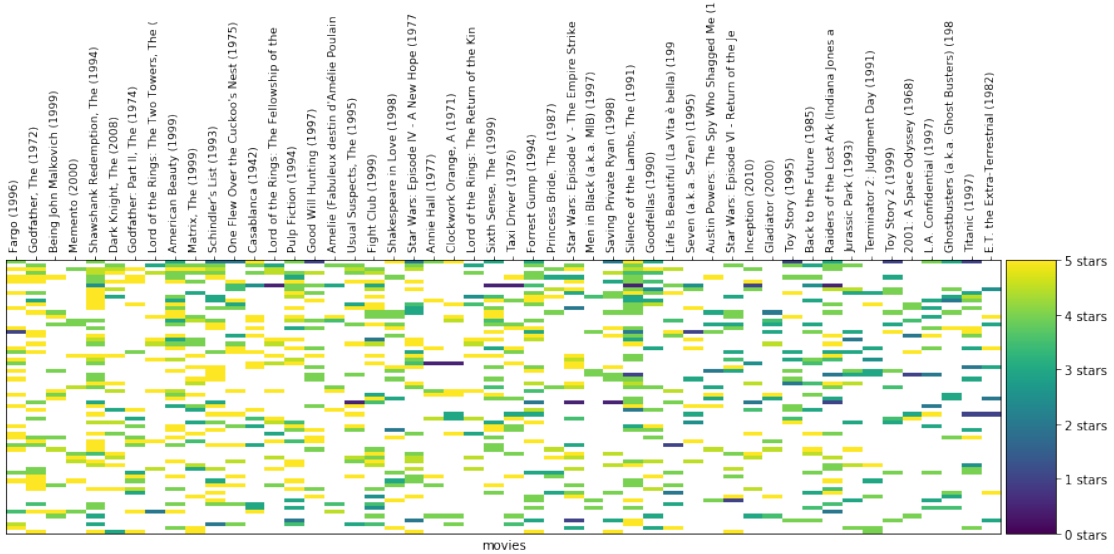
With k-means, we have to specify k, the number of clusters. Let's arbitrarily try k=20 (A better way to pick k is as illustrated above with the elbow method. That would take some processing time to run, however.):

```
In [38]: # 20 clusters
         predictions = KMeans(n_clusters=20, algorithm='full').fit_predict(sparse_ratings)
```

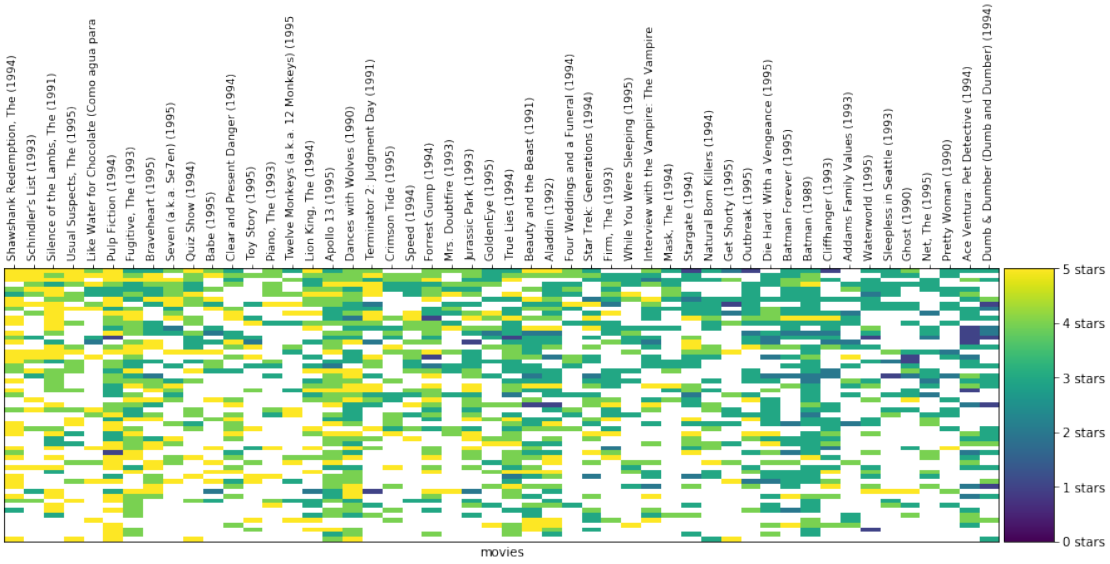
To visualize some of these clusters, we'll plot each cluster as a heat map:

```
In [39]: max_users = 70
         max_movies = 50

         clustered = pd.concat([most_rated_movies_1k.reset_index(), pd.DataFrame({'group':predic
         helper.draw_movie_clusters(clustered, max_users, max_movies)
```

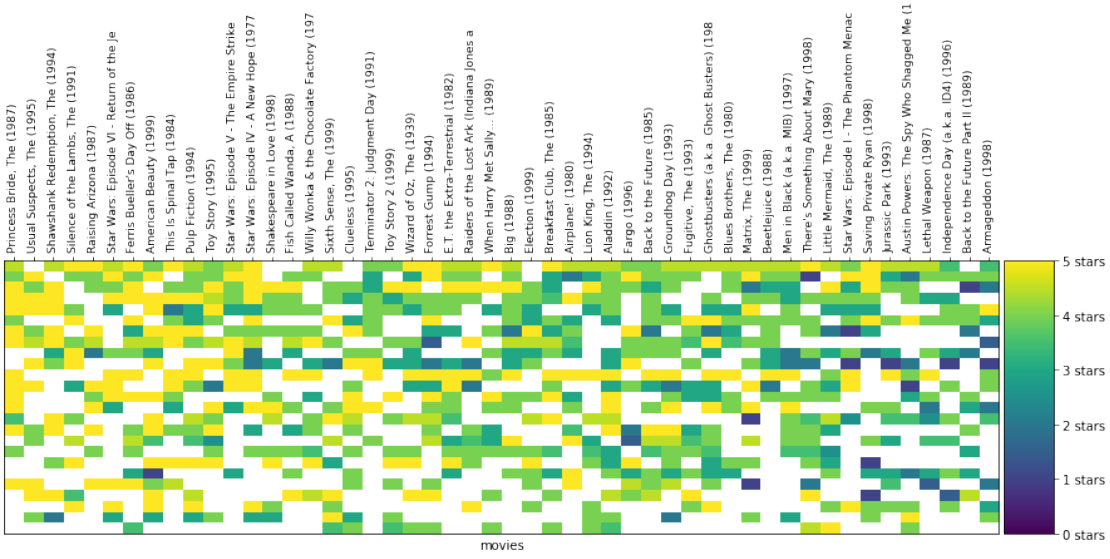
```
cluster # 15
# of users in cluster: 265. # of users in plot: 70
```



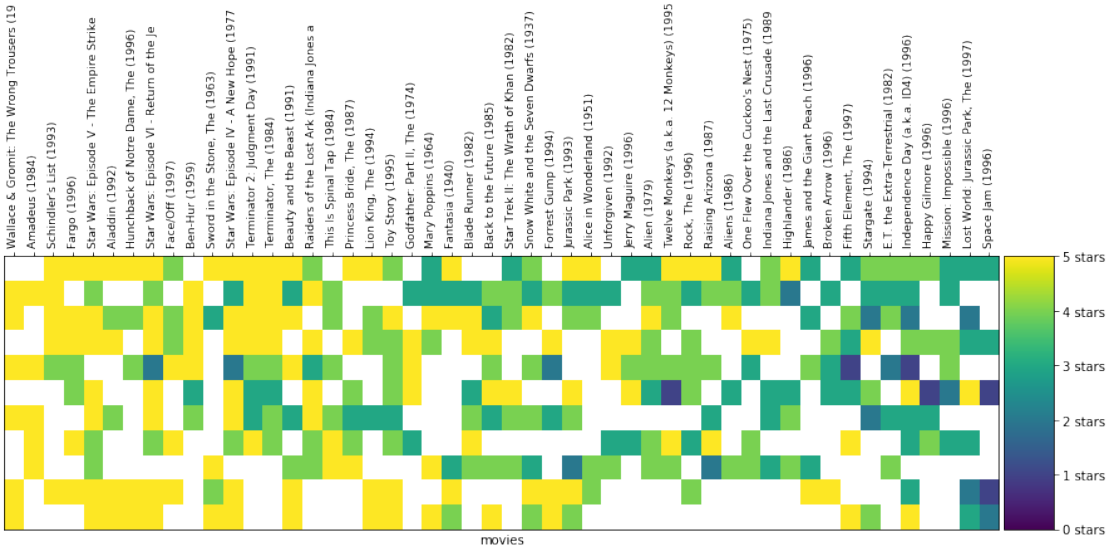
cluster # 4
# of users in cluster: 57. # of users in plot: 57



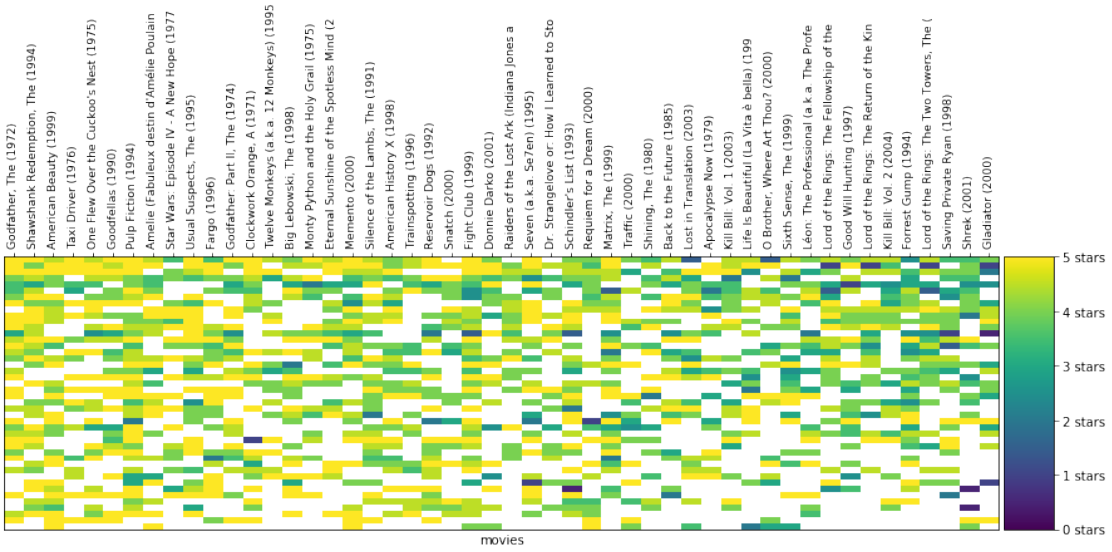
cluster # 7
# of users in cluster: 25. # of users in plot: 25



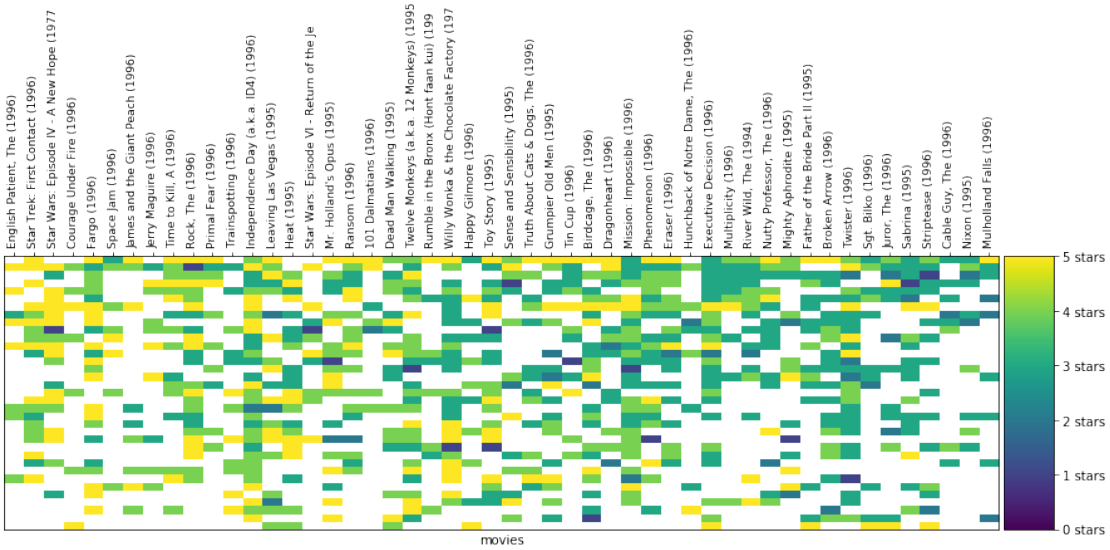
cluster # 1
# of users in cluster: 11. # of users in plot: 11



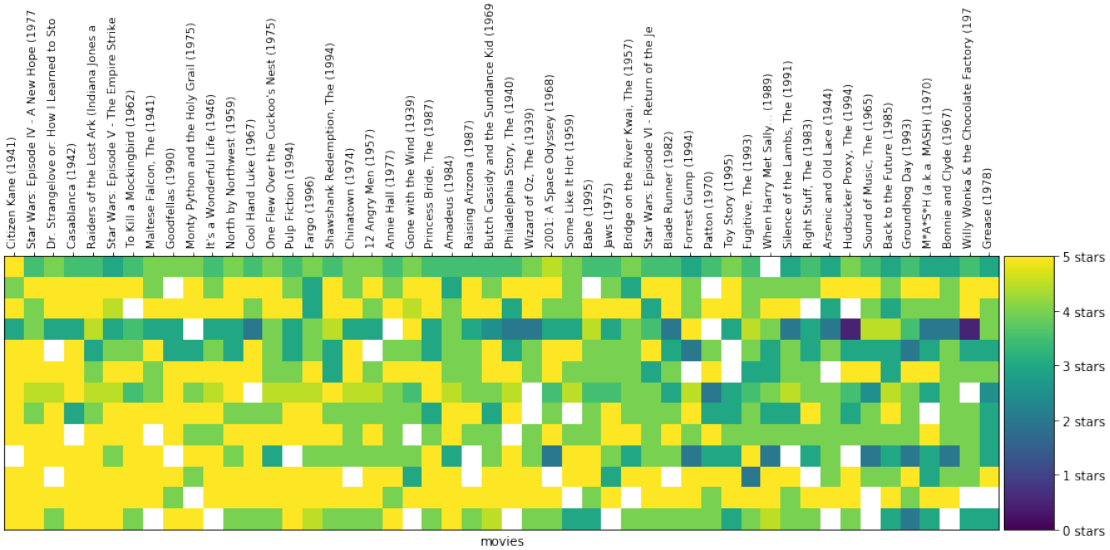
cluster # 8
# of users in cluster: 44. # of users in plot: 44



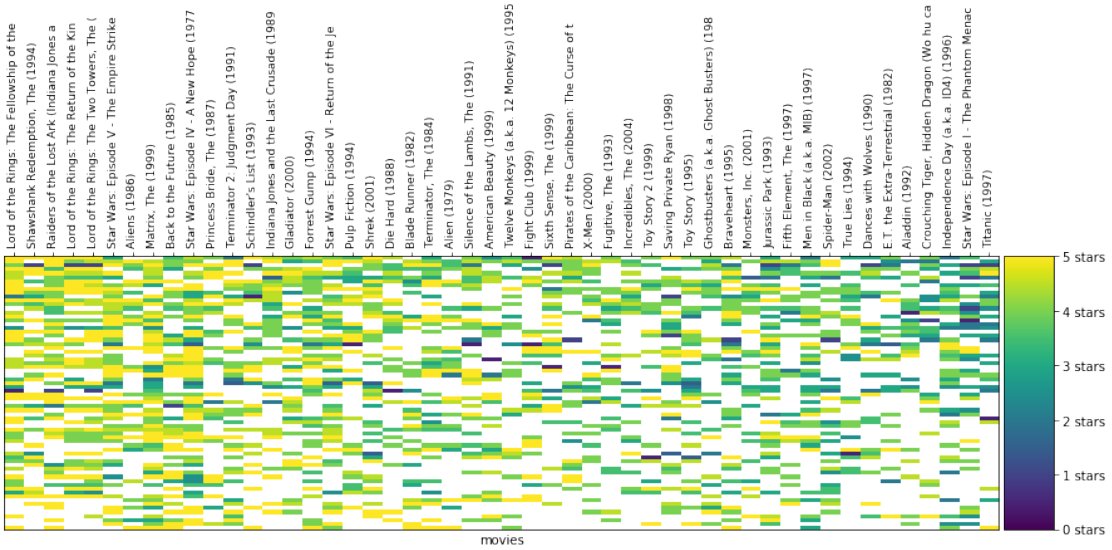
cluster # 5
# of users in cluster: 35. # of users in plot: 35



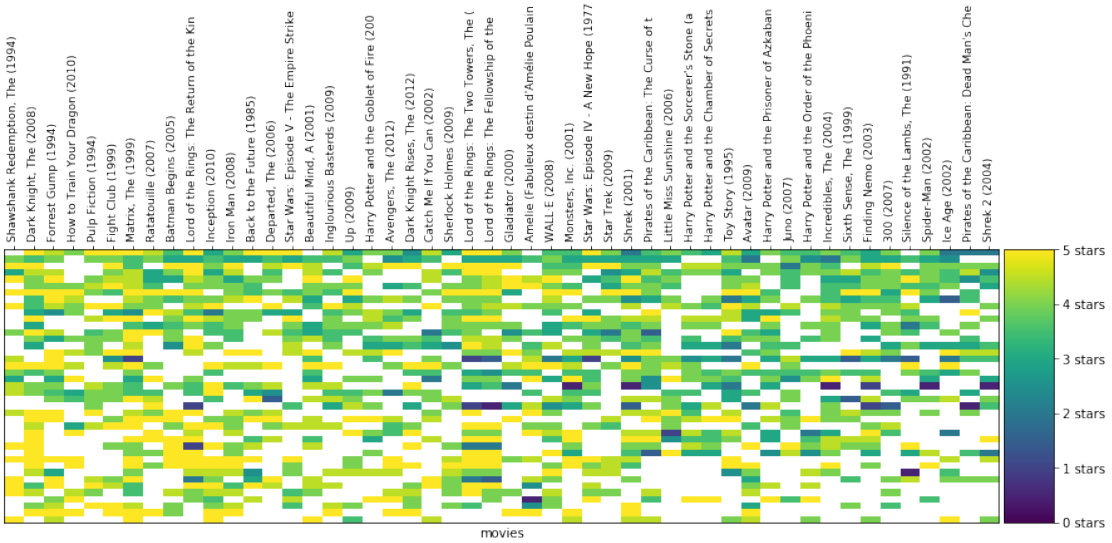
cluster # 12
# of users in cluster: 13. # of users in plot: 13



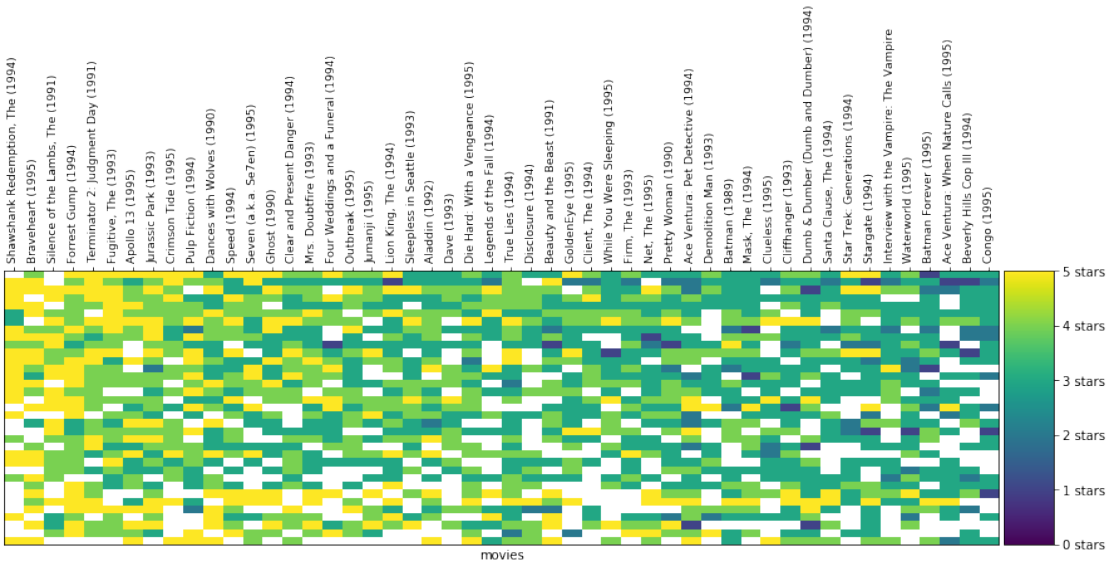
cluster # 19
# of users in cluster: 78. # of users in plot: 70



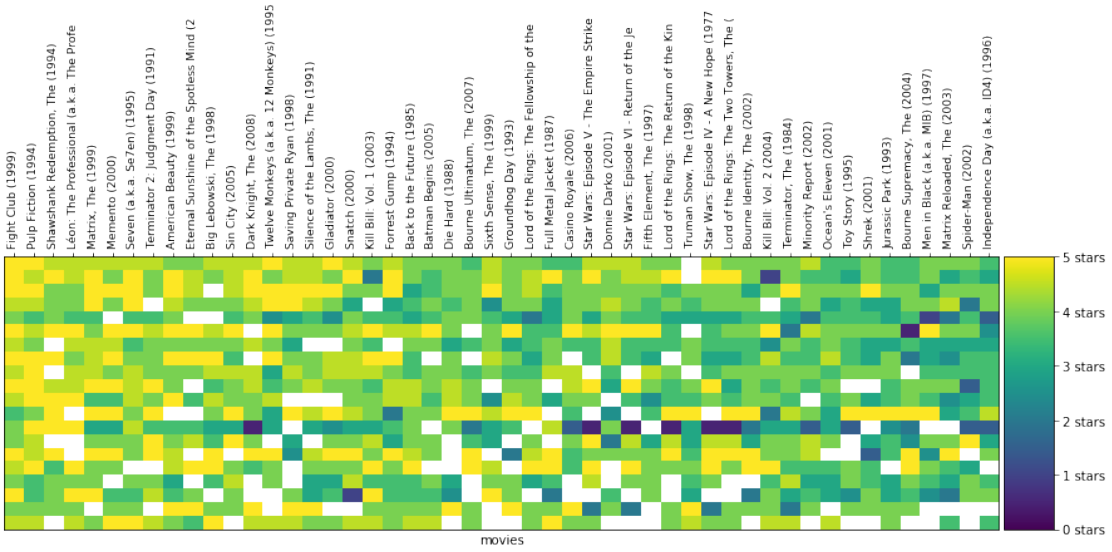
cluster # 16
# of users in cluster: 41. # of users in plot: 41



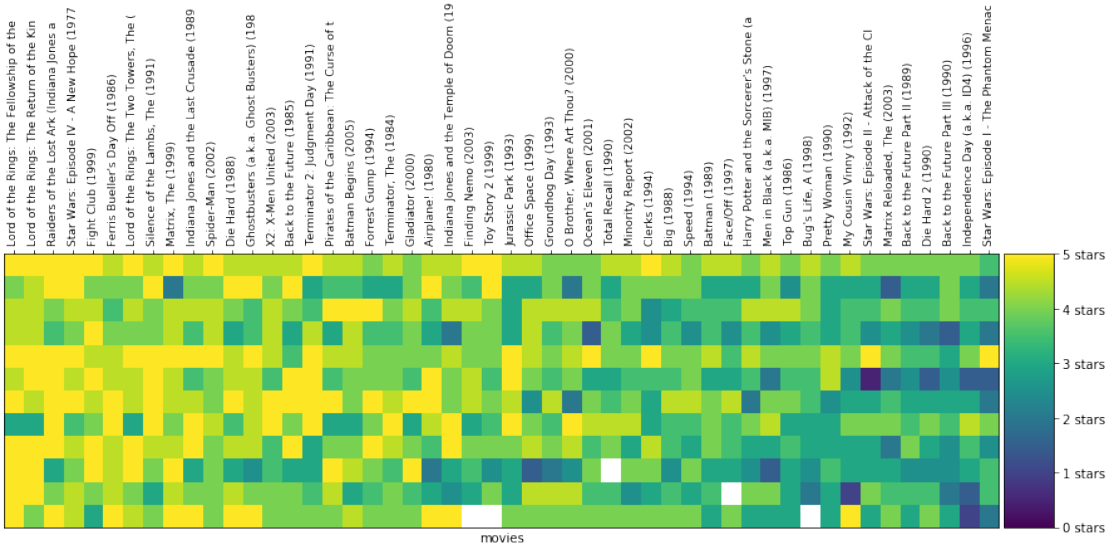
cluster # 14
# of users in cluster: 35. # of users in plot: 35



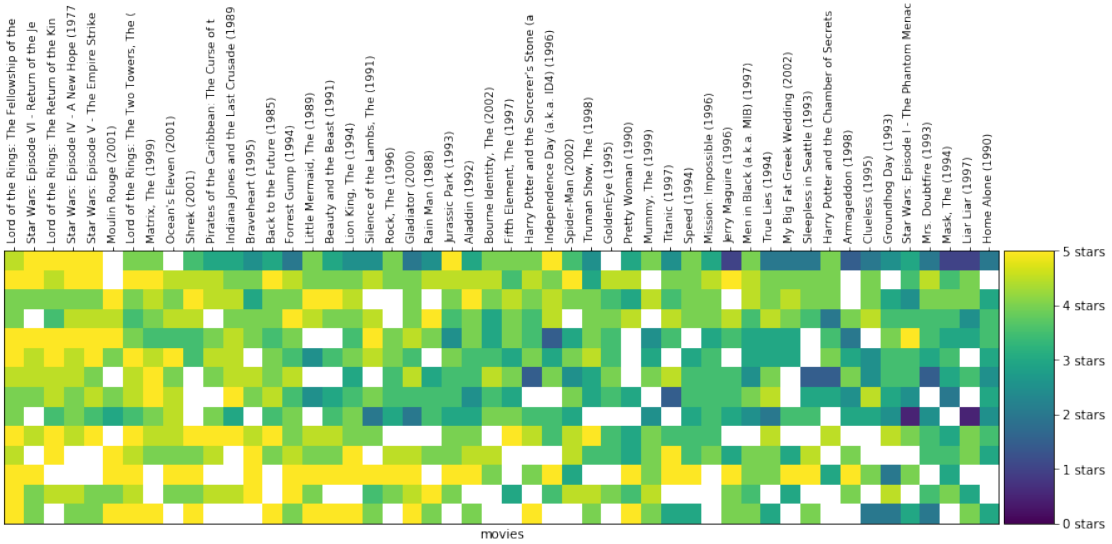
cluster # 2
# of users in cluster: 20. # of users in plot: 20



cluster # 3
# of users in cluster: 12. # of users in plot: 12




```
cluster # 18
# of users in cluster: 14. # of users in plot: 14
```



There are several things to note here: * The more similar the ratings in a cluster are, the more **vertical** lines in similar colors you'll be able to trace in that cluster. * It's super interesting to spot trends in clusters: * Some clusters are more sparse than others, containing people who probably watch and rate less movies than in other clusters. * Some clusters are mostly yellow and bring together people who really love a certain group of movies. Other clusters are mostly green or navy blue meaning they contain people who agree that a certain set of movoies deserves 2-3 stars.

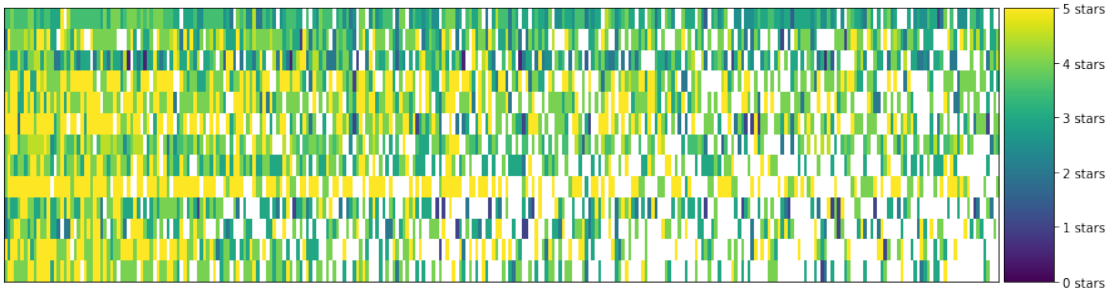
* Note how the movies change in every cluster. The graph filters the data to only show the most rated movies, and then sorts them by average rating. * Can you track where the Lord of the Rings movies appear in each cluster? What about Star Wars movies? * It's easy to spot **horizontal** lines with similar colors, these are users without a lot of variety in their ratings. This is likely one of the reasons for Netflix switching from a stars-based ratings to a thumbs-up/thumbs-down rating. A rating of four stars means different things to different people. * We did a few things to make the clusters visibile (filtering/sorting/slicing). This is because datasets like this are "sparse" and most cells do not have a value (because most people did not watch most movies).

1.7 Prediction

Let's pick a cluster and a specific user and see what useful things this clustering will allow us to do.

Let's first pick a cluster:

```
In [40]: # TODO: Pick a cluster ID from the clusters above
         cluster_number = 12

         # Let's filter to only see the region of the dataset with the most number of values
         n_users = 75
         n_movies = 300
         cluster = clustered[clustered.group == cluster_number].drop(['index', 'group'], axis=1)

         cluster = helper.sort_by_rating_density(cluster, n_movies, n_users)
         helper.draw_movies_heatmap(cluster, axis_labels=False)
```



And the actual ratings in the cluster look like this:

```
In [41]: cluster.fillna('').head()
```

```
Out[41]:    Groundhog Day (1993)  Amadeus (1984)  North by Northwest (1959)  \
         7                   3.5             3.5                        4.0
         9                   2.0             5.0                        4.0
         5                   3.5             4.0                        3.0
         8                   4.0             4.0                        5.0
         2                   5.0             4.0                        5.0
```

```
   Princess Bride, The (1987)  Fugitive, The (1993)  \
7                          3.5                   3.5
9                          4.0                   4.0
5                          3.0                   4.5
8                          5.0                   4.0
2                          5.0                   3.0

   Star Wars: Episode VI - Return of the Jedi (1983)  \
7                                                 3.5
9                                                 4.0
5                                                 3.5
8                                                 3.0
2                                                 3.0

   Butch Cassidy and the Sundance Kid (1969)  \
7                                        3.5
9                                        4.5
5                                        2.5
8                                        5.0
2                                        5.0

   One Flew Over the Cuckoo's Nest (1975)  Fargo (1996)  \
7                                     4.0           4.0
9                                     4.0           4.5
5                                     4.0           4.0
8                                     5.0           3.0
2                                     4.0           5.0

   Back to the Future (1985)       ...        Terms of Endearment (1983)  \
7                        3.0       ...
9                        3.0       ...
5                        4.5       ...
8                        3.0       ...                                 4
2                        3.0       ...                                 4

   Crouching Tiger, Hidden Dragon (Wo hu cang long) (2000)  \
7                                                      3.5
9                                                      4.5
5                                                        4
8                                                        4
2

   Pee-wee's Big Adventure (1985)  Caddyshack (1980) Crocodile Dundee (1986)  \
7                             3.5                2.5                       3
9                               3                4.5                       3
5                                                                          4
8                               4                                          4
2                               4                  4

   Tootsie (1982)  Blues Brothers, The (1980)  \
7             3.5                         3.5
9             4.0                         4.0
5             4.0                         3.5
8             3.0                         3.0
2             3.0                         3.0

   Silence of the Lambs, The (1991)  Wizard of Oz, The (1939)  \
7                               3.5                       3.5
9                               4.0                       4.0
5                               4.0                       4.0
8                               4.0                       3.0
2                               4.0                       3.0

   Cinderella (1950)  Dances with Wolves (1990)  \
7                3.5                        3.5
9                4.0                        4.0
5                4.0                        4.0
8                3.0                        3.0
2                3.0                        3.0

   Nightmare Before Christmas, The (1993)  Raiders of the Lost Ark (1981)  \
7                                     3.5                             3.5
9                                     4.0                             4.0
5                                     4.0                             4.0
8                                     3.0                             3.0
2                                     3.0                             3.0

   Mr. Smith Goes to Washington (1939)  Big Night (1996)  \
7                                  3.5               3.5
9                                  4.0               4.0
5                                  4.0               4.0
8                                  3.0               3.0
2                                  3.0               3.0

   Day the Earth Stood Still, The (1951)  Monty Python and the Holy Grail (1974)  \
7                                    3.5                                     3.5
9                                    4.0                                     4.0
5                                    4.0                                     4.0
8                                    3.0                                     3.0
2                                    3.0                                     3.0

   Wizard of Oz, The (1939)  Cinderella (1950)  Dances with Wolves (1990)  \
7                       3.5                3.5                        3.5
9                       4.0                4.0                        4.0
5                       4.0                4.0                        4.0
8                       3.0                3.0                        3.0
2                       3.0                3.0                        3.0

   Nightmare Before Christmas, The (1993)  Raiders of the Lost Ark (1981)  \
7                                     3.5                             3.5
9                                     4.0                             4.0
5                                     4.0                             4.0
8                                     3.0                             3.0
2                                     3.0                             3.0

   Mr. Smith Goes to Washington (1939)  Big Night (1996)  \
7                                  3.5               3.5
9                                  4.0               4.0
5                                  4.0               4.0
8                                  3.0               3.0
2                                  3.0               3.0

   Day the Earth Stood Still, The (1951)  Monty Python and the Holy Grail (1974)  \
7                                    3.5                                     3.5
9                                    4.0                                     4.0
5                                    4.0                                     4.0
8                                    3.0                                     3.0
2                                    3.0                                     3.0

   Wizard of Oz, The (1939)
7                       3.5
9                       4.0
5                       4.0
8                       3.0
2                       3.0
```

```
  Brothers McMullen, The (1995) Doors, The (1991) Cast Away (2000)  \
7                                                                 3
9                             4                                   3
5                                                 2               4
8                             4                                   3
2                                                 3

  Don Juan DeMarco (1995) Rounders (1998)
7
9                       3              4
5                       3            4.5
8                                      3
2                       4

[5 rows x 300 columns]
```

Pick a blank cell from the table. It's blank because that user did not rate that movie. Can we predict whether she would like it or not? Since the user is in a cluster of users that seem to have similar taste, we can take the average of the votes for that movie in this cluster, and that would be a reasonable predicition for much she would enjoy the film.

```
In [54]: # TODO: Fill in the name of the column/movie. e.g. 'Forrest Gump (1994)'
         # Pick a movie from the table above since we're looking at a subset

         movie_name = 'Back to the Future (1985)'

         cluster[movie_name].mean()
```

```
Out[54]: 3.7692307692307692
```

And this would be our prediction for how she'd rate the movie.

1.8 Recommendation

Let's reiterate what we did in the previous step. We have used k-means to cluster users according to their ratings. This lead us to clusters of users with similar ratings and thus generally a similar taste in movies. Based on this, when one user did not have a rating for a certain movie we averaged the ratings of all the other users in the cluster, and that was our guess to how this one user would like the movie.

Using this logic, if we calculate the average score in this cluster for every movie, we'd have an understanding for how this 'taste cluster' feels about each movie in the dataset.

```
In [55]: # The average rating of 20 movies as rated by the users in the cluster
         cluster.mean().head(20)
```

```
Out[55]: Groundhog Day (1993)                                    3.692
         Amadeus (1984)                                          4.230
         North by Northwest (1959)                               4.384
```

```
Princess Bride, The (1987)                                                    4.269
Fugitive, The (1993)                                                          3.884
Star Wars: Episode VI - Return of the Jedi (1983)                             4.076
Butch Cassidy and the Sundance Kid (1969)                                     4.192
One Flew Over the Cuckoo's Nest (1975)                                        4.346
Fargo (1996)                                                                  4.307
Back to the Future (1985)                                                     3.769
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)    4.500
Star Wars: Episode V - The Empire Strikes Back (1980)                         4.461
Shawshank Redemption, The (1994)                                              4.307
Star Wars: Episode IV - A New Hope (1977)                                     4.615
Grease (1978)                                                                 3.541
M*A*S*H (a.k.a. MASH) (1970)                                                  3.666
Bonnie and Clyde (1967)                                                       3.666
Citizen Kane (1941)                                                           4.750
When Harry Met Sally... (1989)                                                3.875
2001: A Space Odyssey (1968)                                                  4.166
dtype: float64
```

This becomes really useful for us because we can now use it as a recommendation engine that enables our users to discover movies they're likely to enjoy.

When a user logs in to our app, we can now show them recommendations that are appropriate to their taste. The formula for these recommendations is to select the cluster's highest-rated movies that the user did not rate yet.

In [56]:
```
# TODO: Pick a user ID from the dataset
# Look at the table above outputted by the command "cluster.fillna('').head()"
# and pick one of the user ids (the first column in the table)
user_id = 7

# Get all this user's ratings
user_2_ratings  = cluster.loc[user_id, :]

# Which movies did they not rate? (We don't want to recommend movies they've already ra
user_2_unrated_movies =  user_2_ratings[user_2_ratings.isnull()]

# What are the ratings of these movies the user did not rate?
avg_ratings = pd.concat([user_2_unrated_movies, cluster.mean()], axis=1, join='inner').

# Let's sort by rating so the highest rated movies are presented first
avg_ratings.sort_values(ascending=False)[:20]
```

Out[56]:
```
Blues Brothers, The (1980)                  4.555556
Schindler's List (1993)                     4.350000
Lone Star (1996)                            4.312500
Shakespeare in Love (1998)                  4.312500
Killing Fields, The (1984)                  4.285714
Bullets Over Broadway (1994)                4.214286
```

```
African Queen, The (1951)                                                    4.136364
Terms of Endearment (1983)                                                   4.083333
Grosse Pointe Blank (1997)                                                   4.071429
High Fidelity (2000)                                                         4.071429
Piano, The (1993)                                                            4.071429
Delicatessen (1991)                                                          4.071429
Dead Man Walking (1995)                                                      3.937500
Grifters, The (1990)                                                         3.900000
When Harry Met Sally... (1989)                                               3.875000
Star Trek: First Contact (1996)                                              3.857143
Rounders (1998)                                                              3.833333
Good, the Bad and the Ugly, The (Buono, il brutto, il cattivo, Il) (1966)    3.833333
Rocky (1976)                                                                 3.750000
Secrets & Lies (1996)                                                        3.714286
Name: 0, dtype: float64
```

And these are our top 20 recommendations to the user!

1.8.1 Quiz:

- If the cluster had a movie with only one rating. And that rating was 5 stars. What would the average rating of the cluster for that movie be? How does that effect our simple recommendation engine? How would you tweak the recommender to address this issue?

1.9 More on Collaborative Filtering

- This is a simplistic recommendation engine that shows the most basic idea of "collaborative filtering". There are many heuristics and methods to improve it. The Netflix Prize tried to push the envelope in this area by offering a prize of US$1,000,000 to the recommendation algorithm that shows the most improvement over Netflix's own recommendation algorithm.
- That prize was granted in 2009 to a team called "BellKor's Pragmatic Chaos". This paper shows their approach which employed an ensemble of a large number of methods.
- Netflix did not end up using this $1,000,000 algorithm because their switch to streaming gave them a dataset that's much larger than just movie ratings – what searches did the user make? What other movies did the user sample in this session? Did they start watching a movie then stop and switch to a different movie? These new data points offered a lot more clues than the ratings alone.

1.10 Take it Further

- This notebook showed user-level recommendations. We can actually use the almost exact code to do item-level recommendations. These are recommendations like Amazon's "Customers who bought (or viewed or liked) this item also bought (or viewed or liked)". These would be recommendations we can show on each movie's page in our app. To do this, we simple transpose the dataset to be in the shape of Movies X Users, and then cluster the movies (rather than the users) based on the correlation of their ratings.
- We used the smallest of the datasets Movie Lens puts out. It has 100,000 ratings. If you want to dig deeper in movie rating exploration, you can look at their Full dataset containing 24 million ratings.