Machine Learning & Data Mining
Miniproject 2: MovieLens Visualization

Caltech CS/CNS/EE 155
Released February 16st, 2024

---

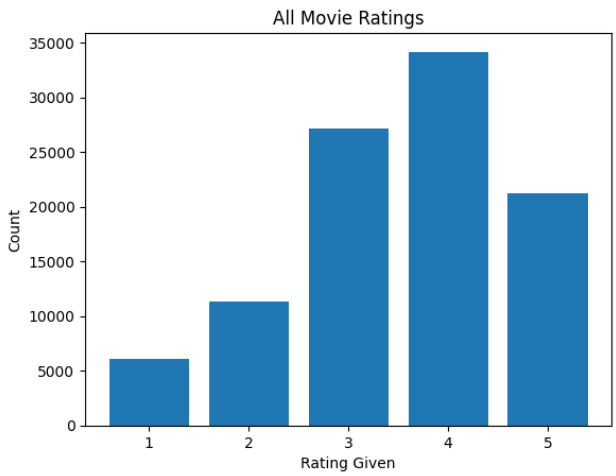
1 Introduction

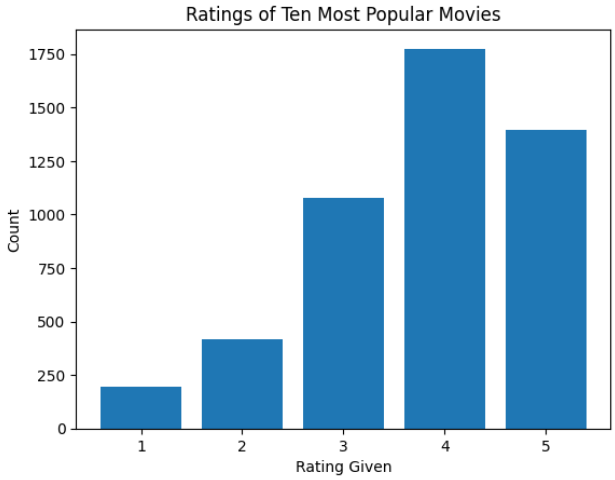
Team name: Beavers
Names: Brendan Flaherty, Sujit Iyer, Sayuj Choudhari
Work Division: Worked together on all parts of the project.
Packages Used: pandas, numpy, matplotlib, seaborn, counter, surprise
Google Colab Link: beavers_mp2.ipynb
Piazza Link: Beavers' Piazza Post

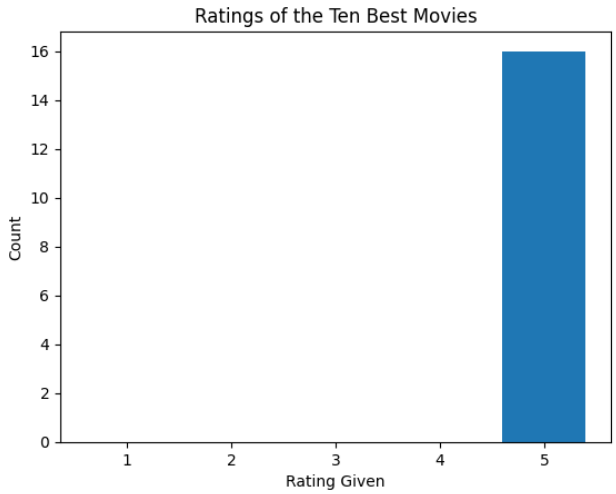
2 Basic Visualizations

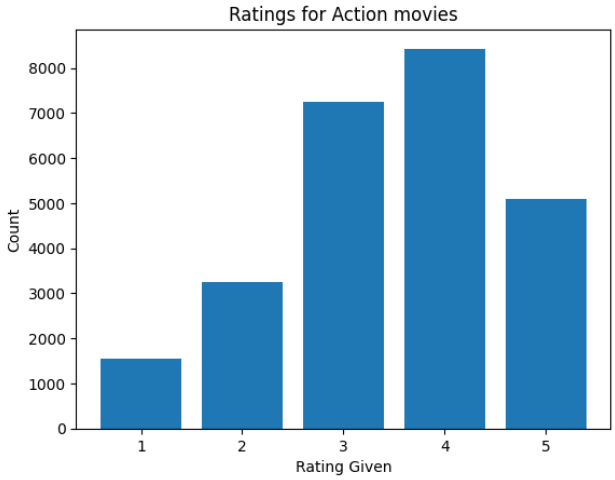
We created visualizations for all ratings in the MovieLens Dataset, for the ten most popular movies, for the ten best movies, and all ratings from Action, Comedy, and Crime genres. First, the histogram of all ratings had a mode rating of 4 and the frequencies were skewed right – favoring higher ratings. We expected to see this skewed distribution due to positivity bias and our belief that people would be more likely to give very good ratings compared to very bad ratings. The ratings of the most popular movies followed a similar distribution pattern of the overall ratings as well. One difference was a higher frequency of rating 5, but the mode was still 4. These results matched our expectations because logically the most popular movies make up a large subset of the entirety of movie ratings and therefore the top 10 most popular movies distribution should roughly follow the distribution for all movies. We also plotted the ratings of the ten best movies and found that each of the 16 ratings was a 5 (the frequency of all other ratings was 0). This matched our expectations because we never established a lower bound on the number of user ratings for each movie, which is why some movies have a high mean rating despite only 1-2 reviews. The distributions of the ratings for the best movies and the most popular movies were very different.

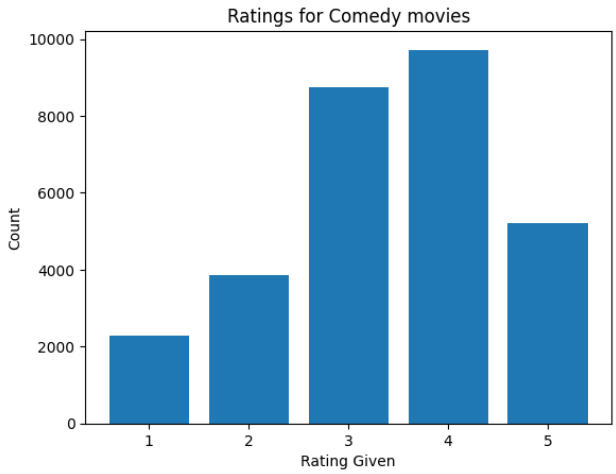
We plotted the ratings of Action, Comedy, and Crime genres. The distributions of the ratings for all the genres were all very similar. For all genres, the rating 4 was the most common. Comedy and Action genres both had many more ratings than the Crime genre. There may not be lots of variation between the genre distributions since the genres are all very popular and enjoyed by many. The six graphs are included below:

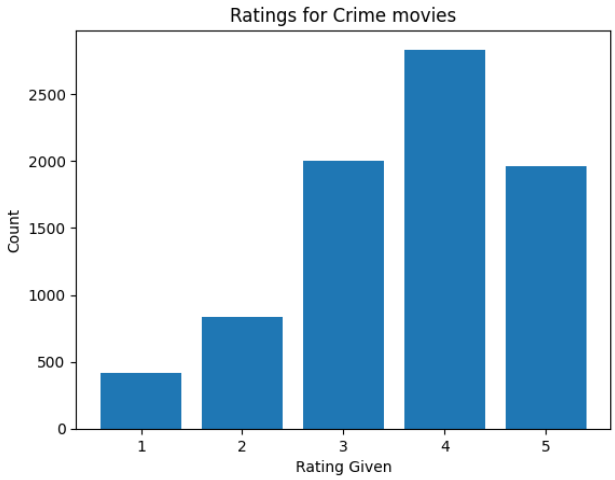

All Movie Ratings
Count
Rating Given
0
5000
10000
15000
20000
25000
30000
35000
1
2
3
4
5


Ratings of Ten Most Popular Movies
Count
Rating Given
0
250
500
750
1000
1250
1500
1750
1
2
3
4
5


Ratings of the Ten Best Movies
Count
Rating Given
0
2
4
6
8
10
12
14
16
1
2
3
4
5


Ratings for Action movies
Count
Rating Given
0
1000
2000
3000
4000
5000
6000
7000
8000
1
2
3
4
5


Ratings for Comedy movies
Count
Rating Given
0
2000
4000
6000
8000
10000
1
2
3
4
5


Ratings for Crime movies
Count
Rating Given
0
500
1000
1500
2000
2500
1
2
3
4
5

3 Matrix Factorization Visualizations

Method choices and derivations

We created visualizations for 3 different matrix factorization methods: a self-produced SVD trained on the regularized loss function, a self-produced SVD with Bias variables for each movie and user trained on the regularized loss function, and Sci-kit Learn's surprise library functions for SVD. The first two SVD's were trained using stochastic gradient descent on the loss functions. The first method used the code from project 5, as we could apply the same loss function and gradient descent method to the SVD of the MovieLens dataset just with different settings of matrix and latent factor sizes. However, for the SVD with bias, the loss function changes to incorporate the bias terms at each element for user i and movie j, as the estimation for each matrix entry at i, j is given by $u_i^T v_j + a_i + b_j$ where a_i is user i's bias and b_j is movie j's bias. This results in new loss function:

$$\arg\min_{U,V} \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + \frac{1}{2}\sum_{i,j}\left(y_{ij} - u_i^T v_j - a_i - b_j\right)^2$$

Given the new loss function, the gradient with respect to each of the variables now changes and we must now take a gradient step for each of the biases as well. For the non-bias variables, notice that the chain rule is not affected by the addition of either bias term as both bias terms are added to the chain term and not directly multiplied or applied to any of the non-bias variables. Thus, we can make simple edit of just adding the bias terms to the chain term of the original SVD gradient:

$$\partial_{u_i} = \lambda u_i + \sum_j (-v_j)(y_{ij} - u_i^T v_j - a_i - b_j)$$

$$\partial_{v_j} = \lambda v_j + \sum_i (-u_i)(y_{ij} - u_i^T v_j - a_i - b_j)$$

For the gradients with respect to the bias terms, notice that a_i, b_j are lone-standing terms of degree 1 with a -1 coefficient, thus it is a simpler chain rule and additionally the regularization term is removed as it has no dependency on the bias terms so it goes to 0 when differentiating, the resulting gradient gives:

$$\partial_{a_i} = -\sum_j (y_{ij} - u_i^T v_j - a_i - b_j)$$

$$\partial_{b_j} = -\sum_i (y_{ij} - u_i^T v_j - a_i - b_j)$$

Notice for the code of the gradient step with stochastic gradient descent only takes steps with respect to a single point y_{ij} so the gradient computation is essentially the same for a_i, b_j so they share the same gradient descent step. This makes the computation a bit easier as we have 1 less gradient to compute but lessens the degrees of freedom the model has to train on as both biases are moving in the same direction.

The third method, Sci-kit Learn's surprise library was an off-the-shelf implementation which was very easy to implement given the ample online documentation. Like the majority of Sci-kit Learn's libraries, the

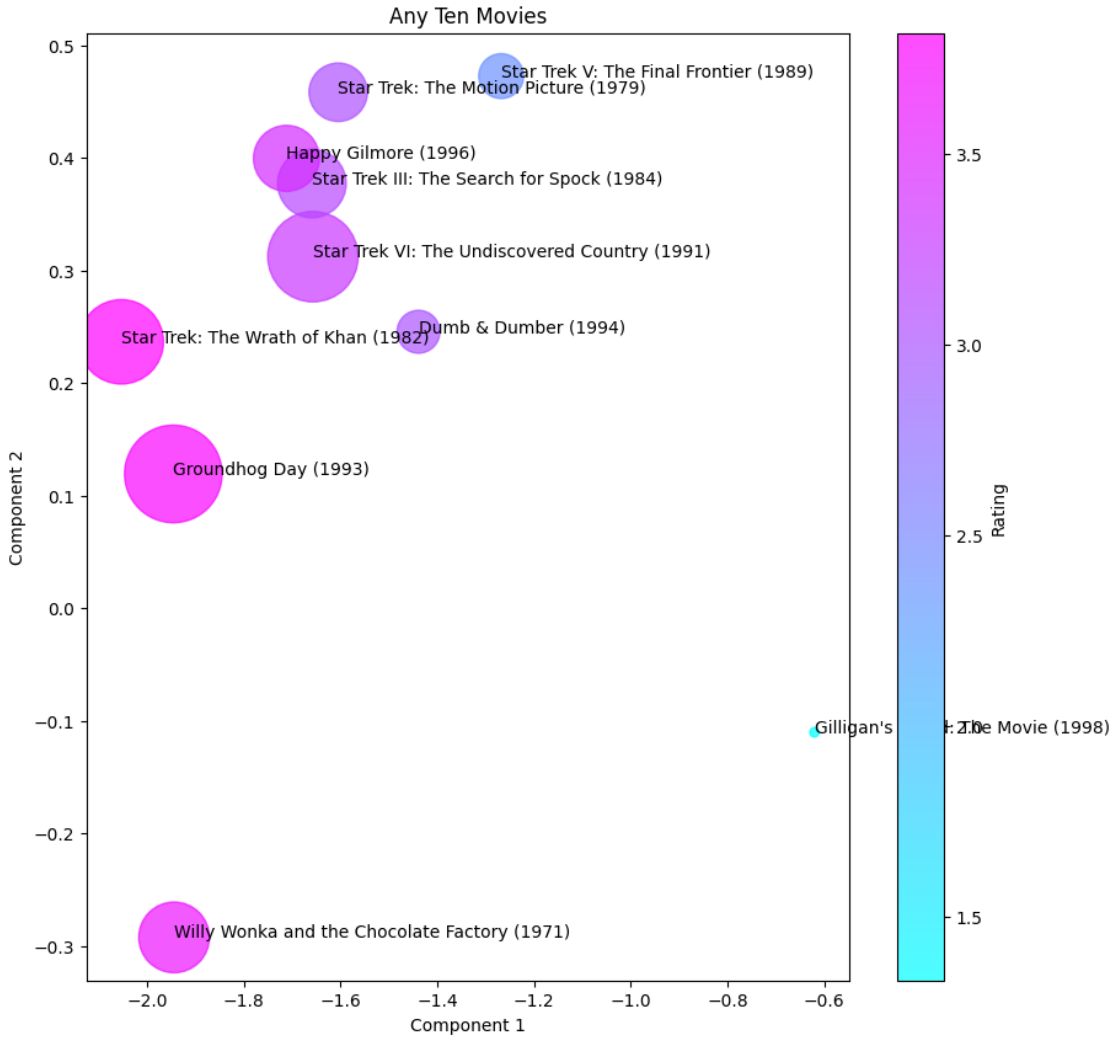
process was straightforward: Initialize an SVD object with desired hyperparameters, fit to the training data, and then find RMSE error by comparing the predictions on the test set. We also performed cross-validation on the entire data set to familiarize ourselves with the SVD library and compare the error to the other two methods.

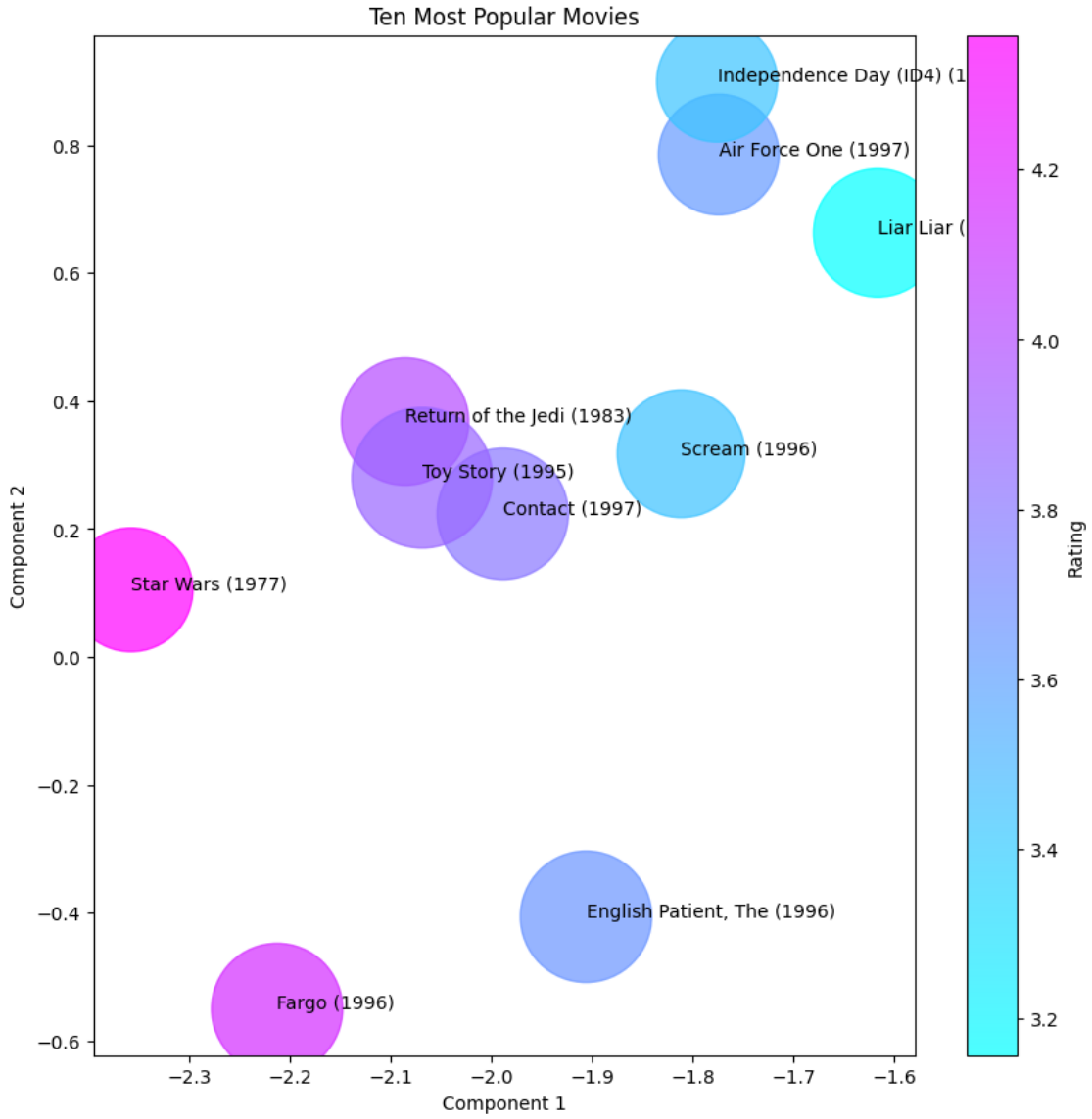
Generally, we saw that the test error was minimized for method 1 compared to methods 2 and 3. The most likely explanation is that methods 2 and 3 overfit the data. Methods 2 and 3 both had bias terms; adding bias terms adds a degree of freedom to the models, and therefore model complexity increases in expense to test accuracy.

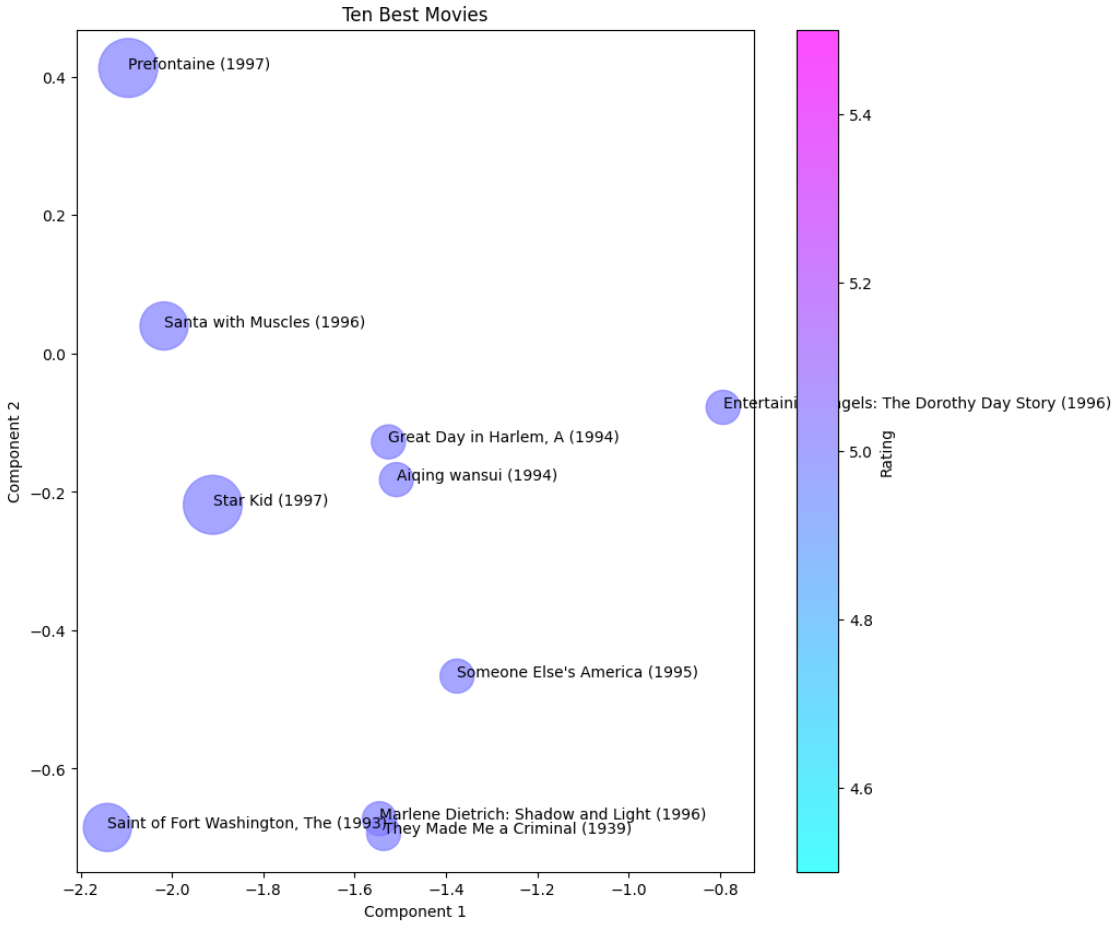
Using the methods described above, we produced 6 different plots for each method to compare the methods against each other. These visualizations plot given movies data projected against the first 2 latent factors of the SVD trained on the dataset. 10 movies were plotted on each graph and the choices of the 10 movies for each of the 6 graph types were: 5 star-trek movies and 5 comedy movies (to see if the latent factors clustered and split based off similarity), top 10 most popular movies (most ratings), top 10 best movies (highest average rating), 10 random action movies, 10 random comedy movies, 10 random crime movies. The plots are inserted below.

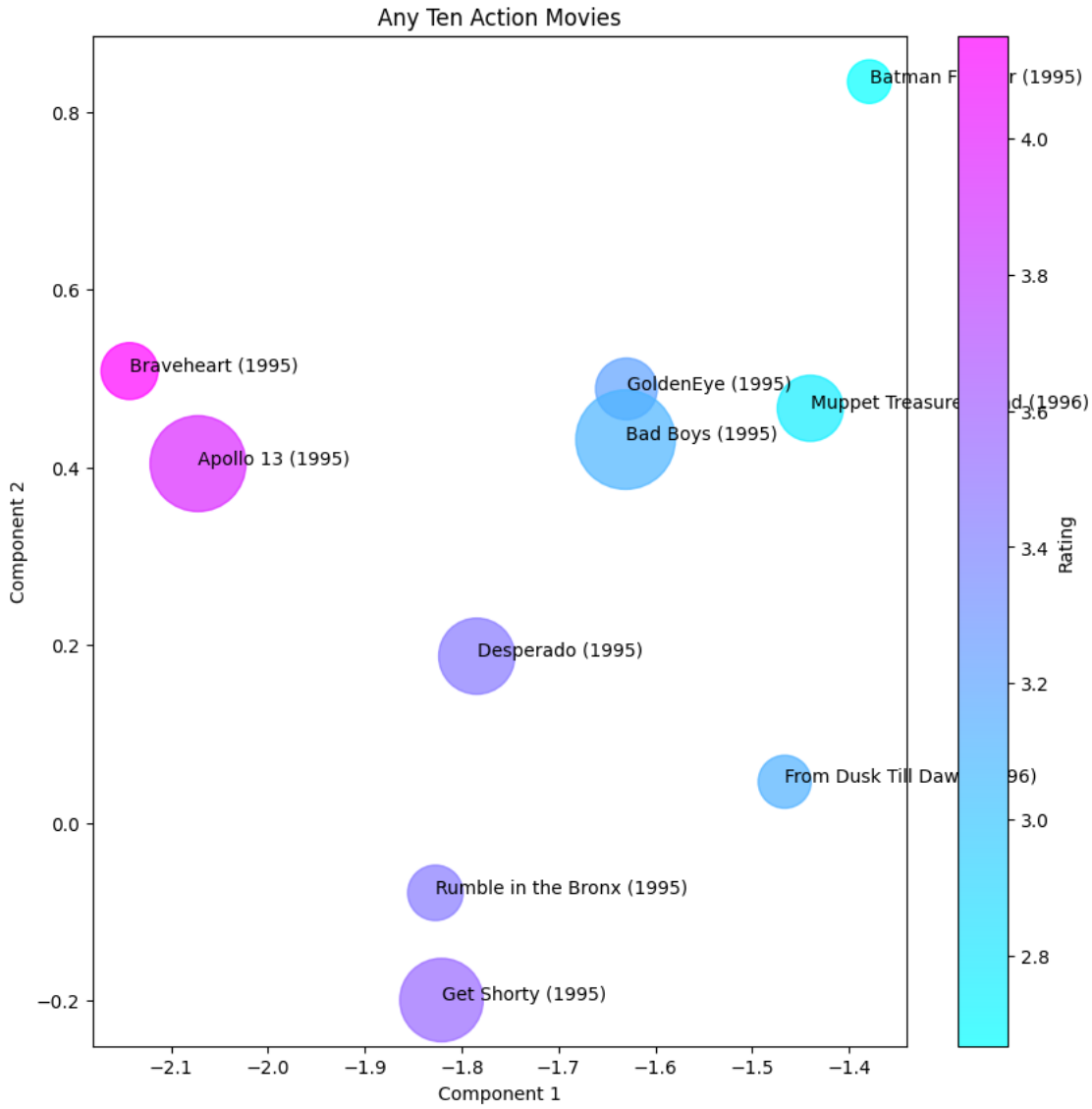
Method 1 Plots

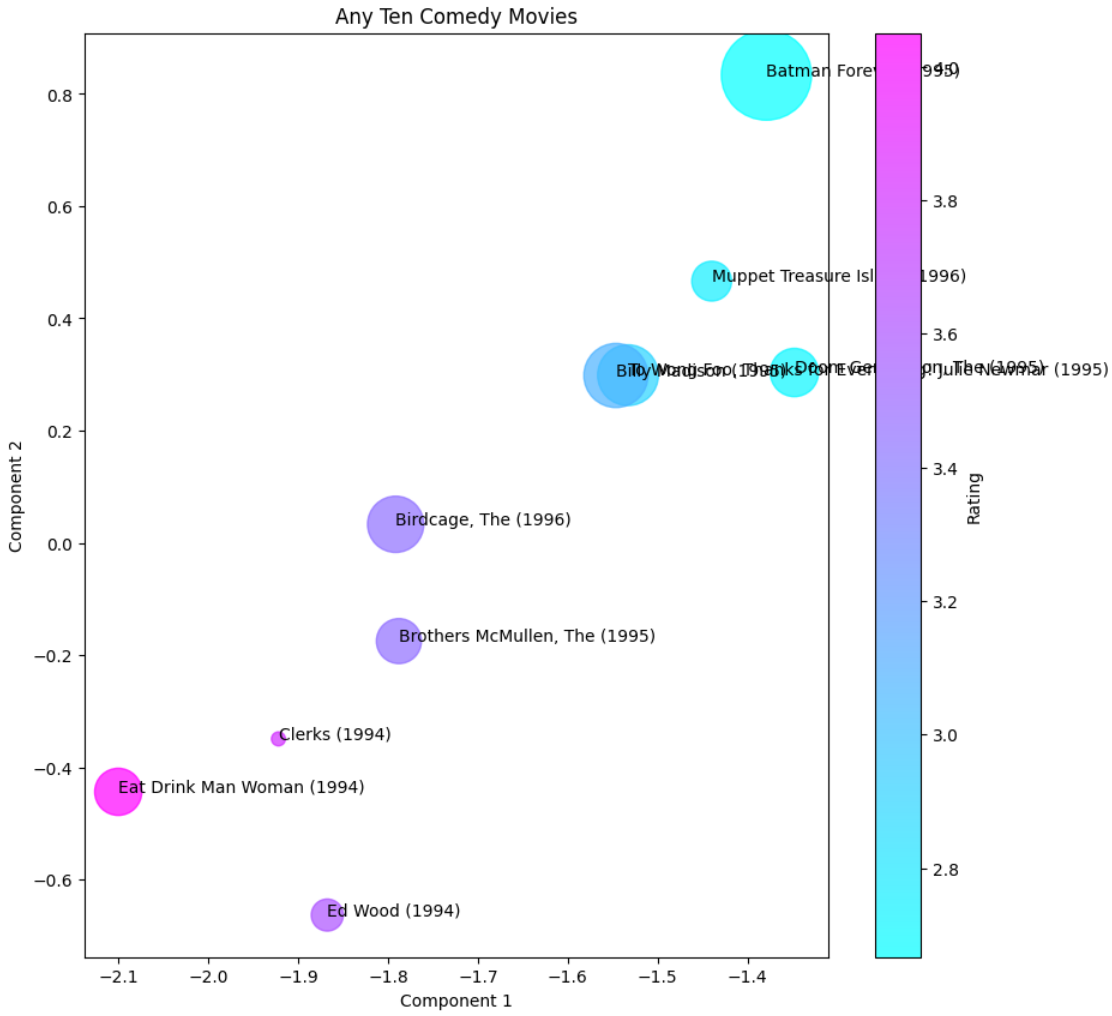
Below are the six visualizations resulting from Method 1.

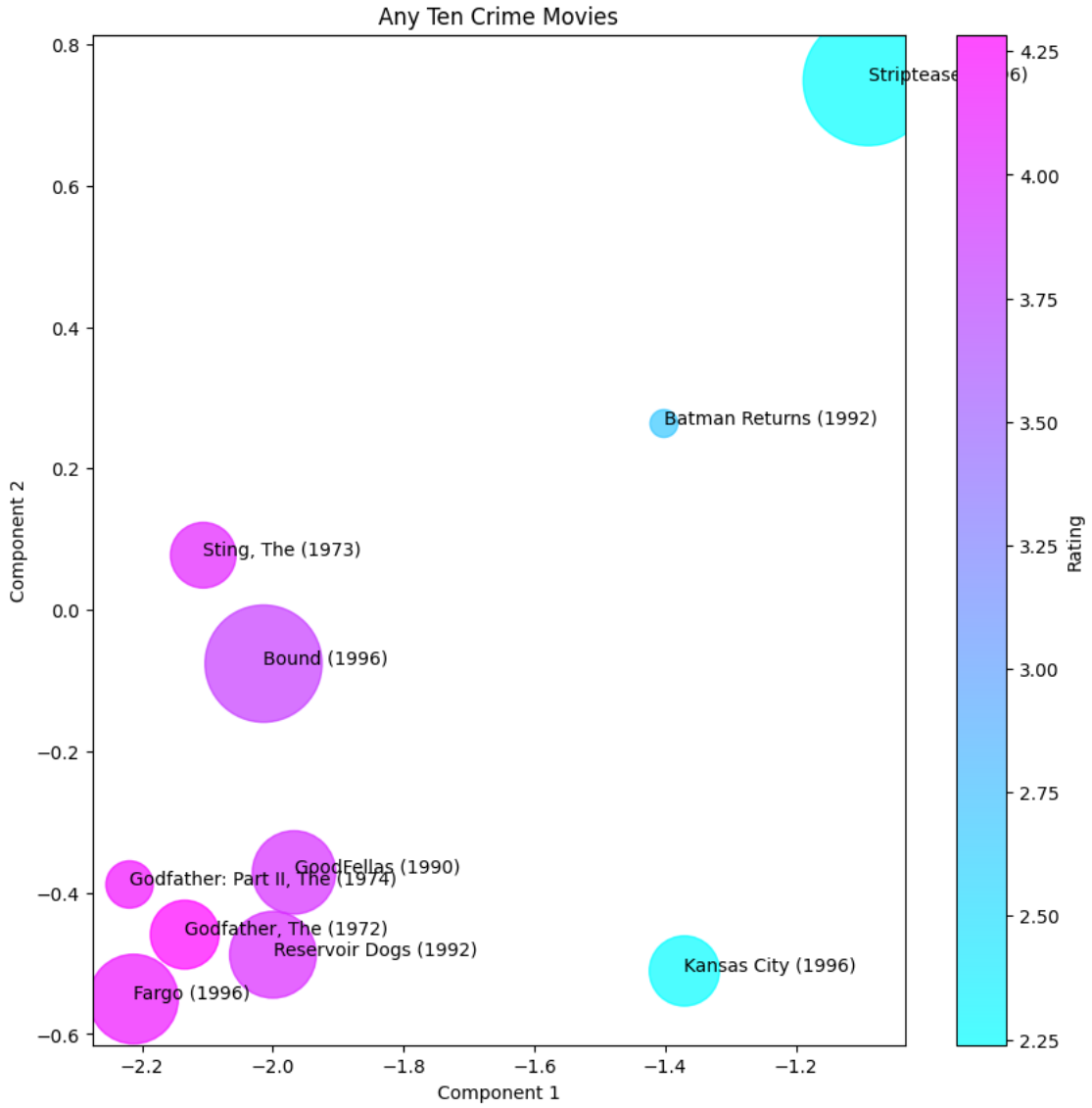



Ten Most Popular Movies
Independence Day (ID4) (1
Air Force One (1997)
Liar Liar (
Return of the Jedi (1983)
Scream (1996)
Toy Story (1995)
Contact (1997)
Star Wars (1977)
English Patient, The (1996)
Fargo (1996)
Component 2
Component 1
Rating
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−2.3
−2.2
−2.1
−2.0
−1.9
−1.8
−1.7
−1.6
4.2
4.0
3.8
3.6
3.4
3.2


Ten Best Movies
Prefontaine (1997)
Santa with Muscles (1996)
Great Day in Harlem, A (1994)
Aiqing wansui (1994)
Star Kid (1997)
Someone Else's America (1995)
Saint of Fort Washington, The (1993)
Marlene Dietrich: Shadow and Light (1996)
They Made Me a Criminal (1939)
gels: The Dorothy Day Story (1996)
Component 1
Component 2
Rating
0.4
0.2
0.0
−0.2
−0.4
−0.6
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
−1.0
−0.8
5.4
5.2
5.0
4.8
4.6

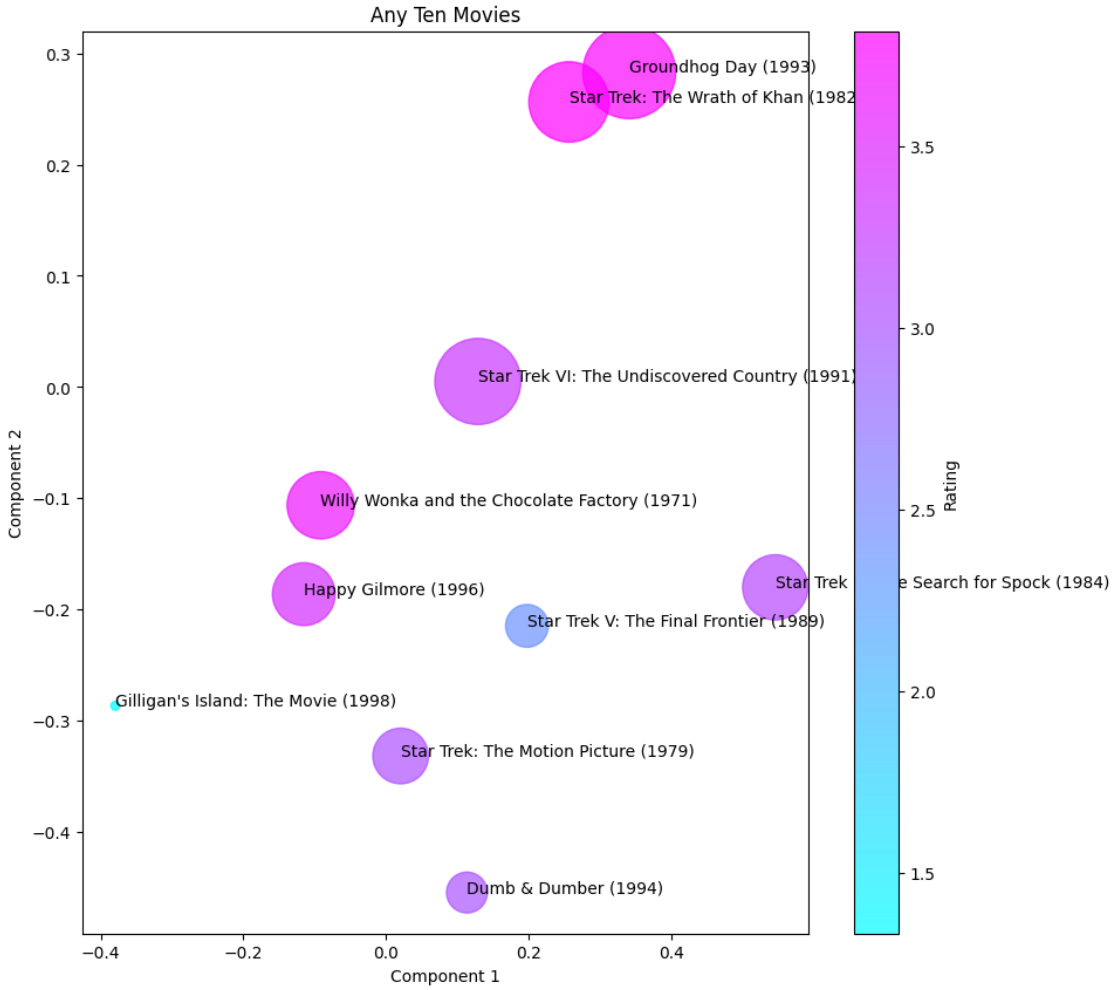

Any Ten Action Movies
Batman F r (1995)
Braveheart (1995)
GoldenEye (1995)
Muppet Treasure d (1996)
Bad Boys (1995)
Apollo 13 (1995)
Desperado (1995)
From Dusk Till Daw 96)
Rumble in the Bronx (1995)
Get Shorty (1995)
Component 2
Component 1
Rating
0.8
0.6
0.4
0.2
0.0
−0.2
−2.1
−2.0
−1.9
−1.8
−1.7
−1.6
−1.5
−1.4
4.0
3.8
3.6
3.4
3.2
3.0
2.8

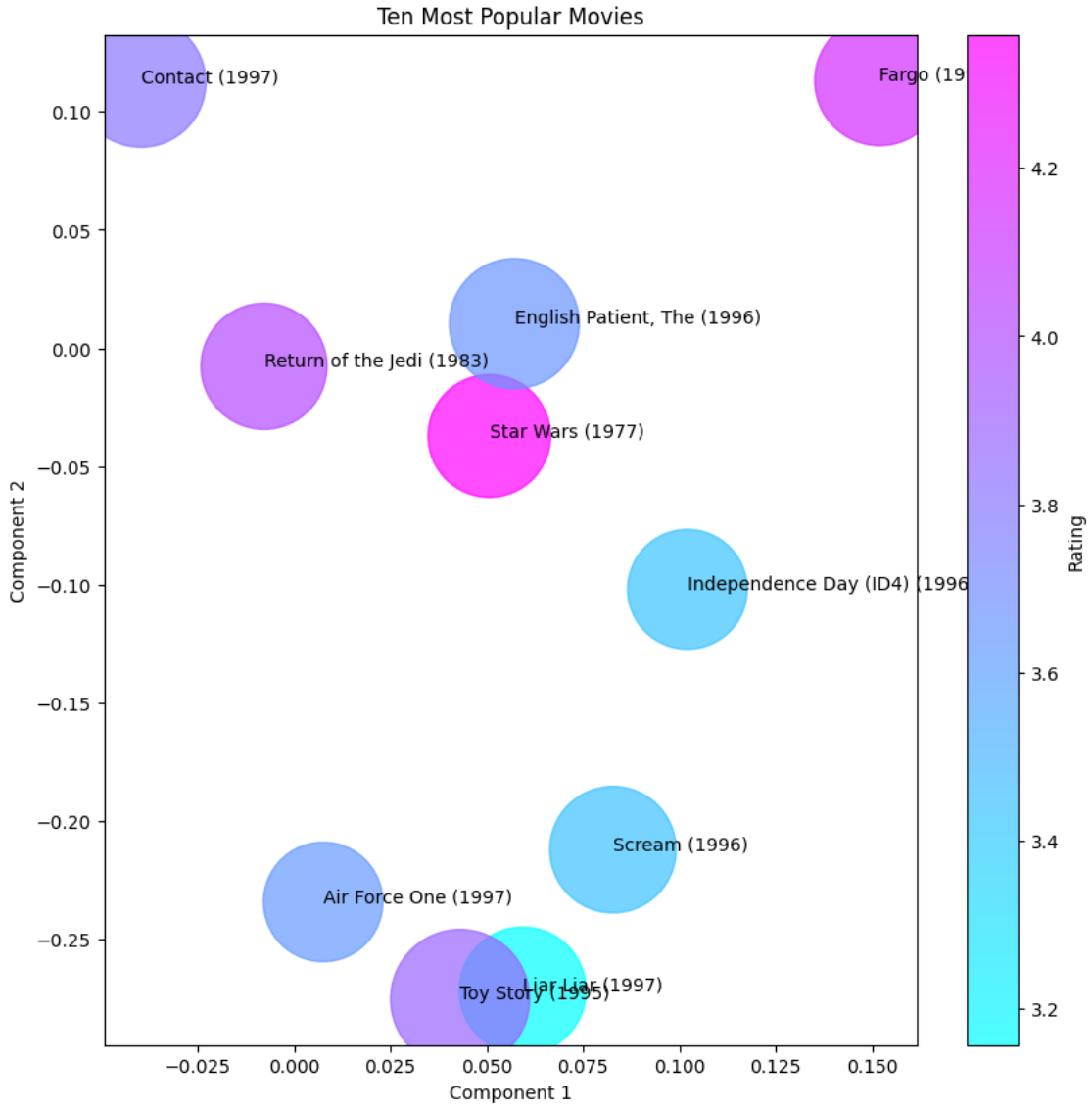

Any Ten Comedy Movies
Component 2
Component 1
Rating
Birdcage, The (1996)
Brothers McMullen, The (1995)
Clerks (1994)
Eat Drink Man Woman (1994)
Ed Wood (1994)
Muppet Treasure Isl
1996)
Julie Newmar (1995)

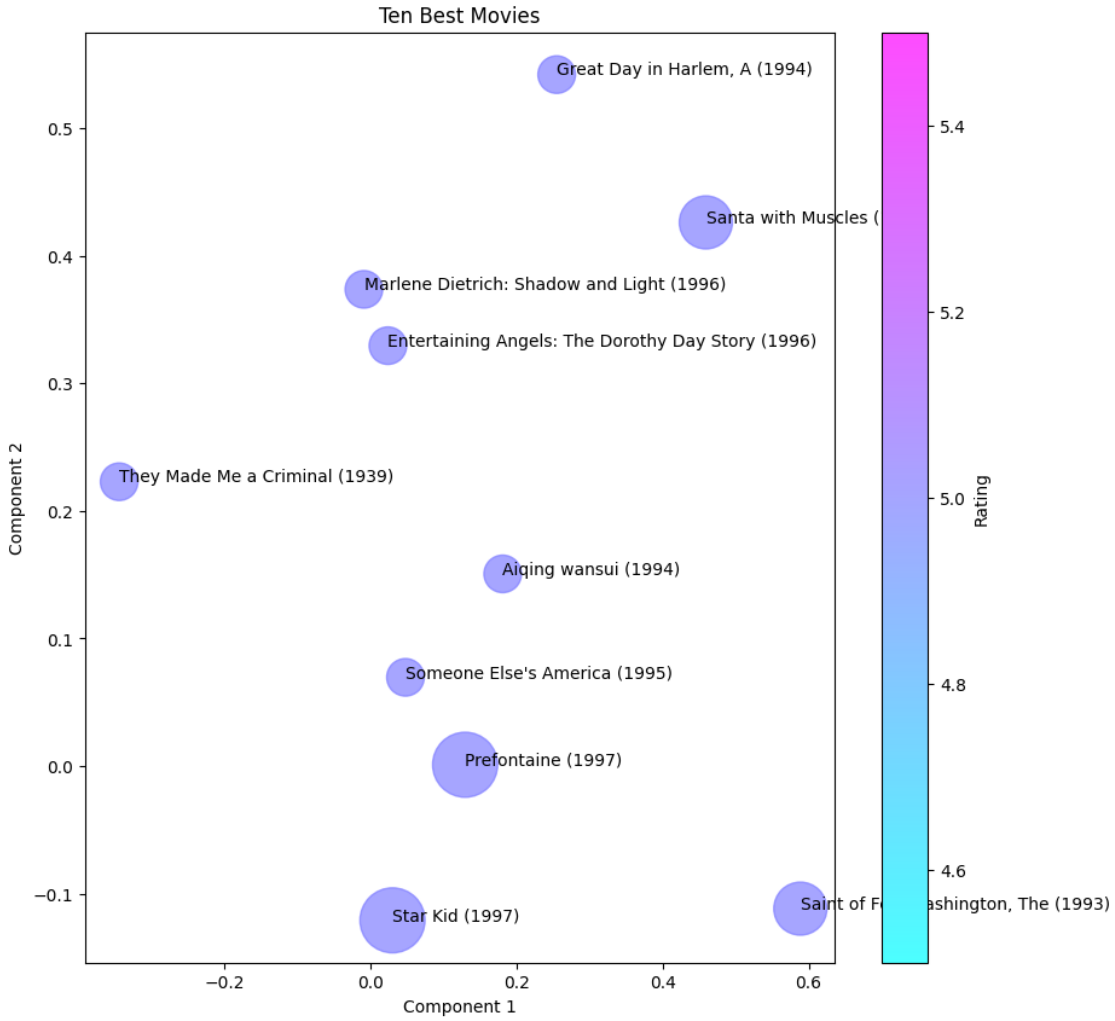

Any Ten Crime Movies
Striptease (1996)
Batman Returns (1992)
Sting, The (1973)
Bound (1996)
GoodFellas (1990)
Godfather: Part II, The (1974)
Godfather, The (1972)
Reservoir Dogs (1992)
Fargo (1996)
Kansas City (1996)
Component 1
Component 2
Rating
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25

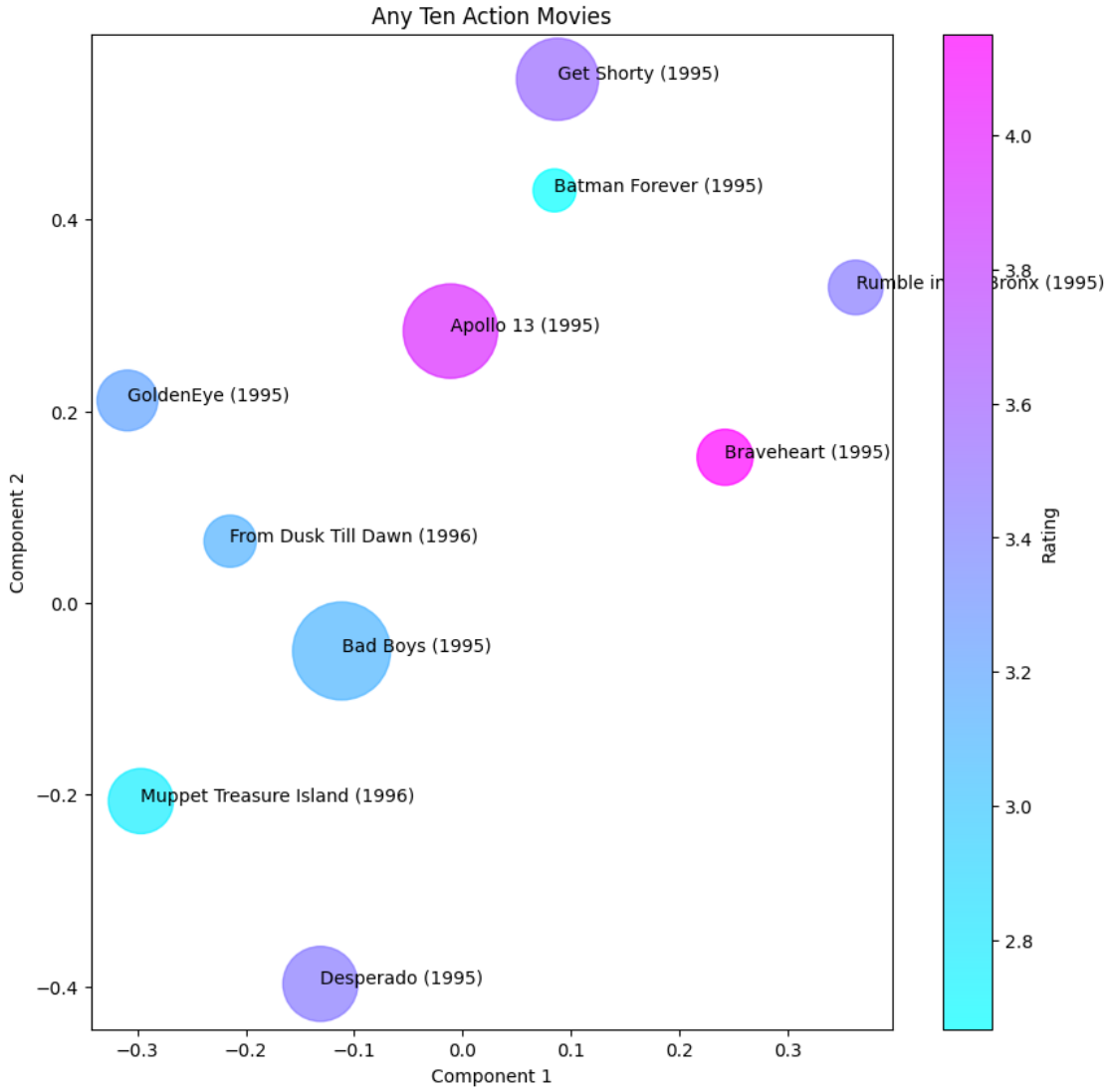
Method 2 Plots

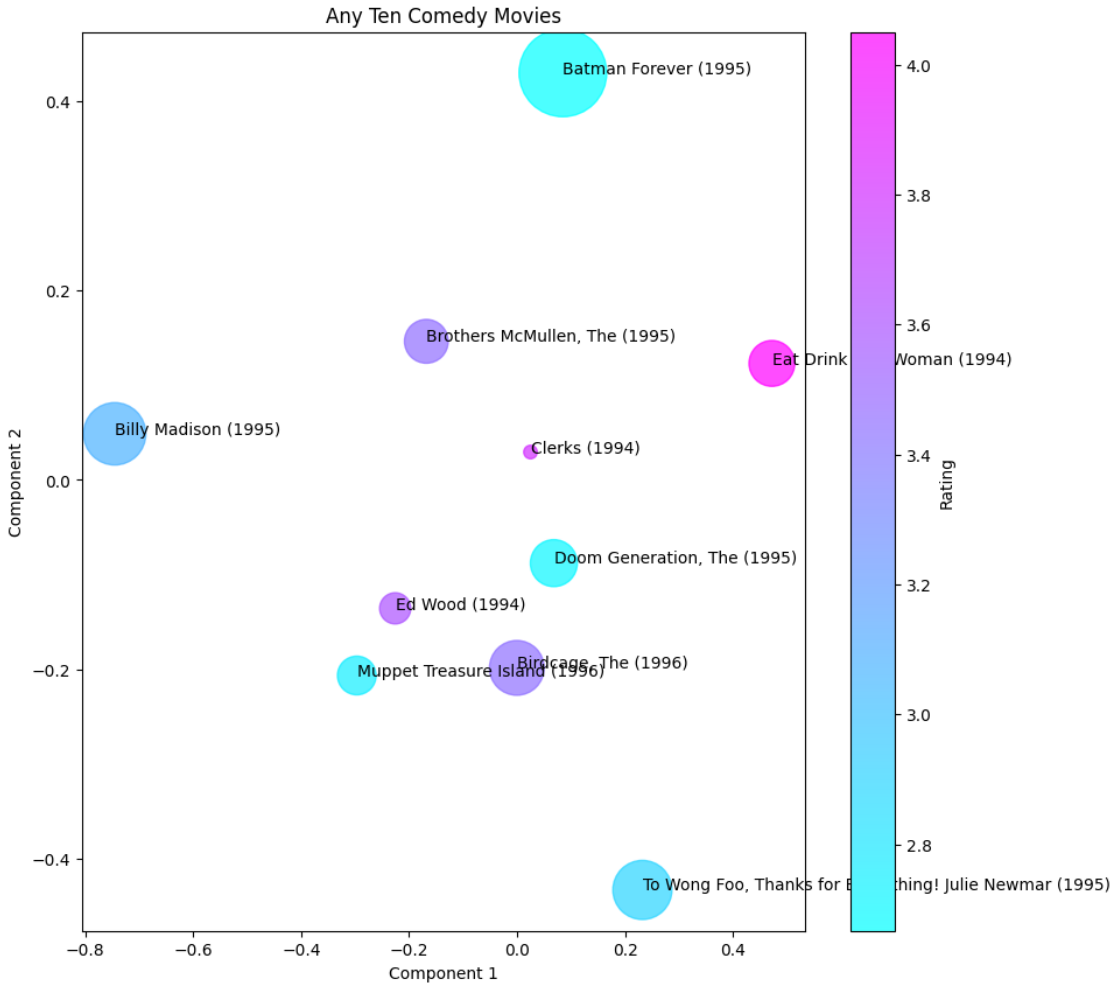
Below are the six visualizations resulting from Method 2.

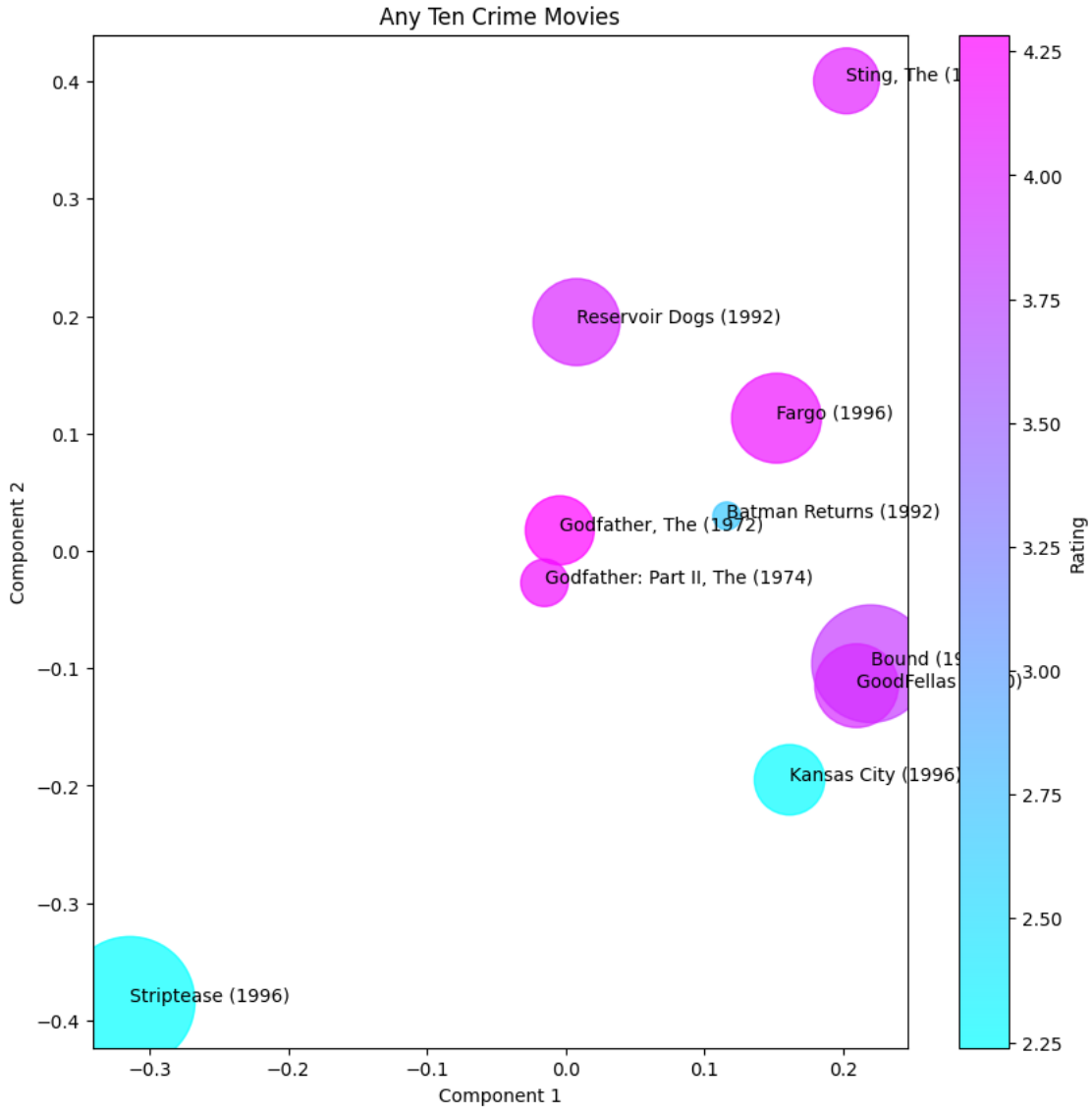



Ten Most Popular Movies
Contact (1997)
Fargo (19
English Patient, The (1996)
Return of the Jedi (1983)
Star Wars (1977)
Independence Day (ID4) (1996
Scream (1996)
Air Force One (1997)
Toy Story (1995)
Liar Liar (1997)
Component 1
Component 2
Rating
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.025
0.000
0.025
0.050
0.075
0.100
0.125
0.150
4.2
4.0
3.8
3.6
3.4
3.2


Ten Best Movies
Great Day in Harlem, A (1994)
Santa with Muscles (
Marlene Dietrich: Shadow and Light (1996)
Entertaining Angels: The Dorothy Day Story (1996)
They Made Me a Criminal (1939)
Aiqing wansui (1994)
Someone Else's America (1995)
Prefontaine (1997)
Star Kid (1997)
Saint of F
ashington, The (1993)
Component 2
Component 1
Rating
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
0.0
0.2
0.4
0.6
5.4
5.2
5.0
4.8
4.6

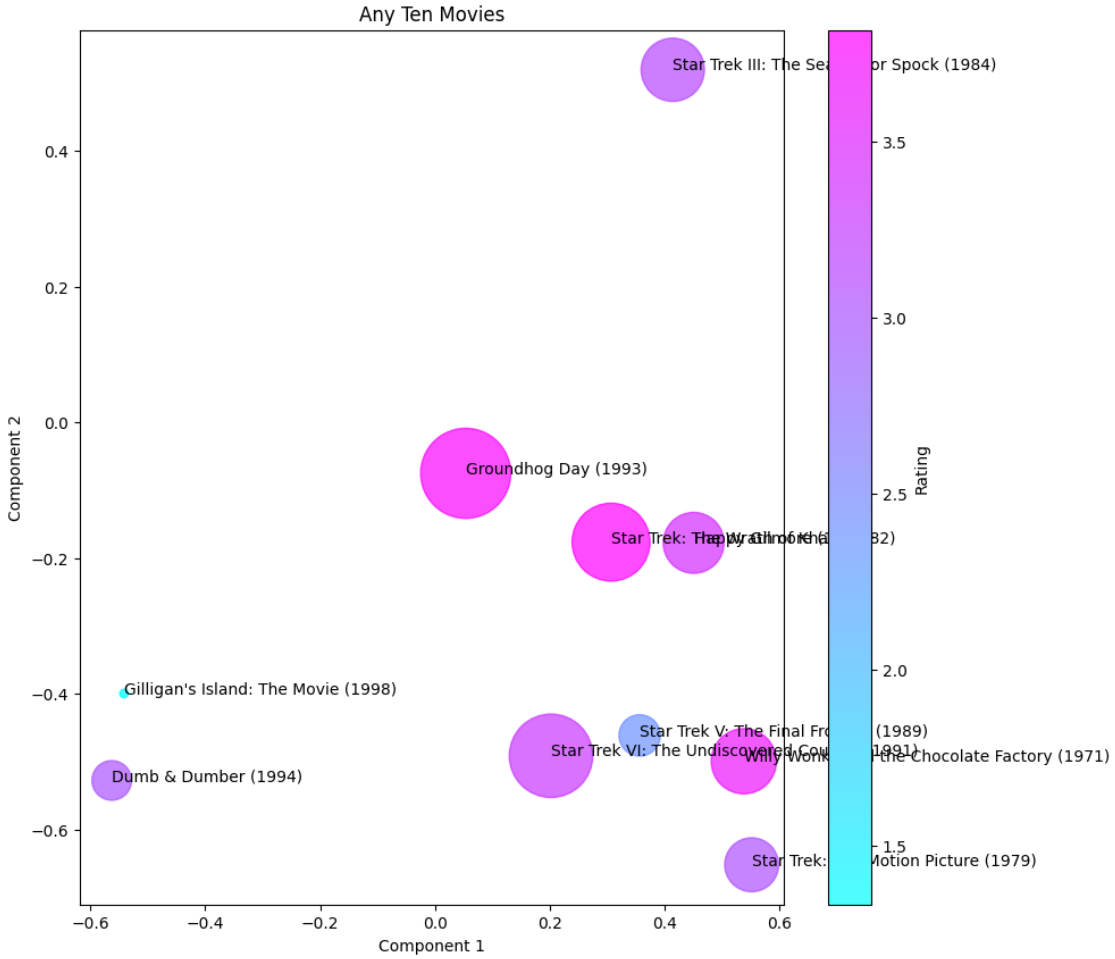

Any Ten Action Movies
Get Shorty (1995)
Batman Forever (1995)
Rumble in the Bronx (1995)
Apollo 13 (1995)
GoldenEye (1995)
Braveheart (1995)
From Dusk Till Dawn (1996)
Bad Boys (1995)
Muppet Treasure Island (1996)
Desperado (1995)
Component 1
Component 2
Rating
0.4
0.2
0.0
−0.2
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
4.0
3.8
3.6
3.4
3.2
3.0
2.8

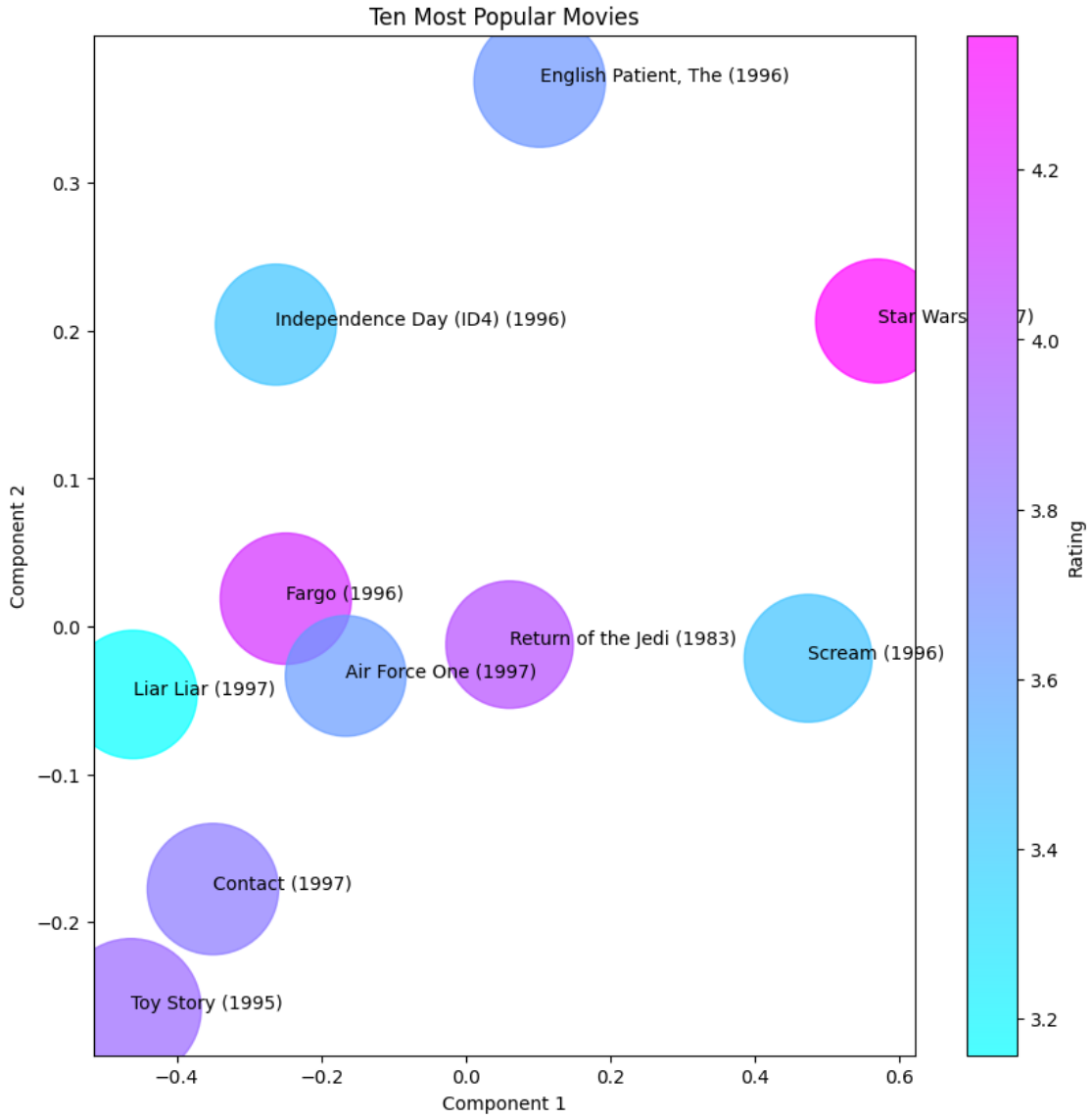

Any Ten Comedy Movies
Batman Forever (1995)
Brothers McMullen, The (1995)
Eat Drink
Woman (1994)
Billy Madison (1995)
Clerks (1994)
Doom Generation, The (1995)
Ed Wood (1994)
Muppet Treasure Island (1996)
Birdcage, The (1996)
To Wong Foo, Thanks for
thing! Julie Newmar (1995)
Component 1
Component 2
Rating
0.4
0.2
0.0
−0.2
−0.4
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
4.0
3.8
3.6
3.4
3.2
3.0
2.8

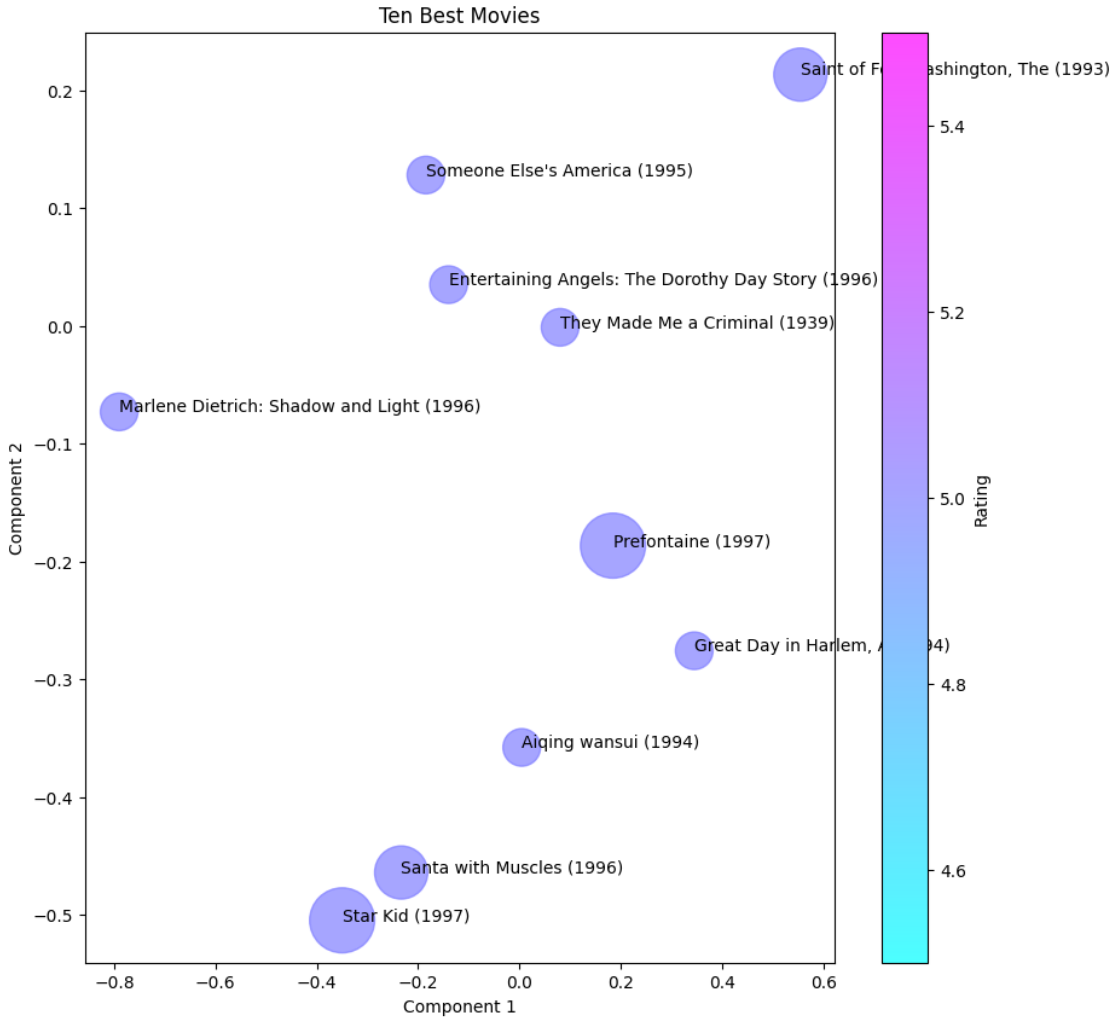

Any Ten Crime Movies
Sting, The (1
Reservoir Dogs (1992)
Fargo (1996)
Batman Returns (1992)
Godfather, The (1972)
Godfather: Part II, The (1974)
Bound (19
GoodFellas
Kansas City (1996)
Striptease (1996)
Component 1
Component 2
Rating
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25

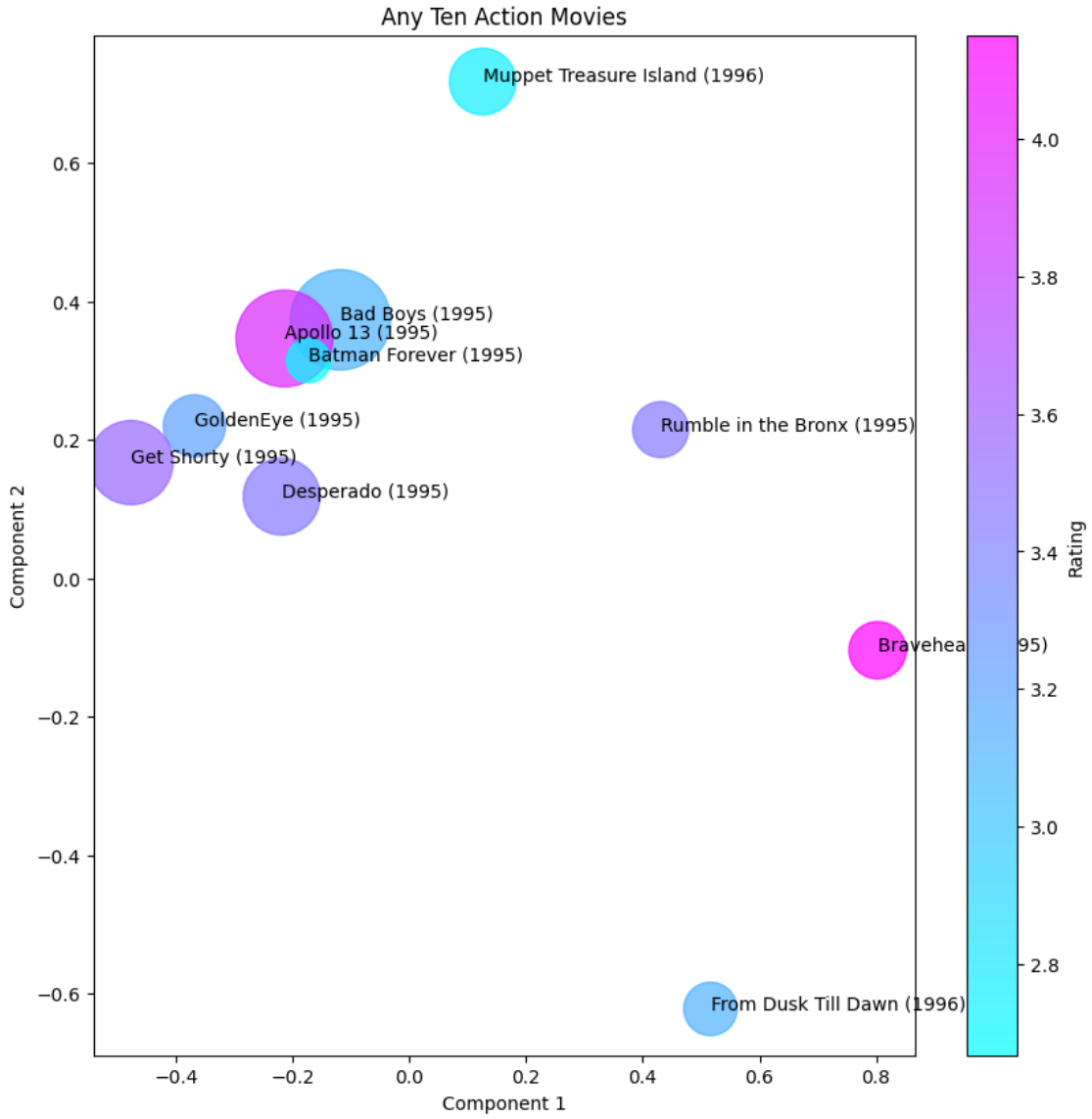
Method 3 Plots

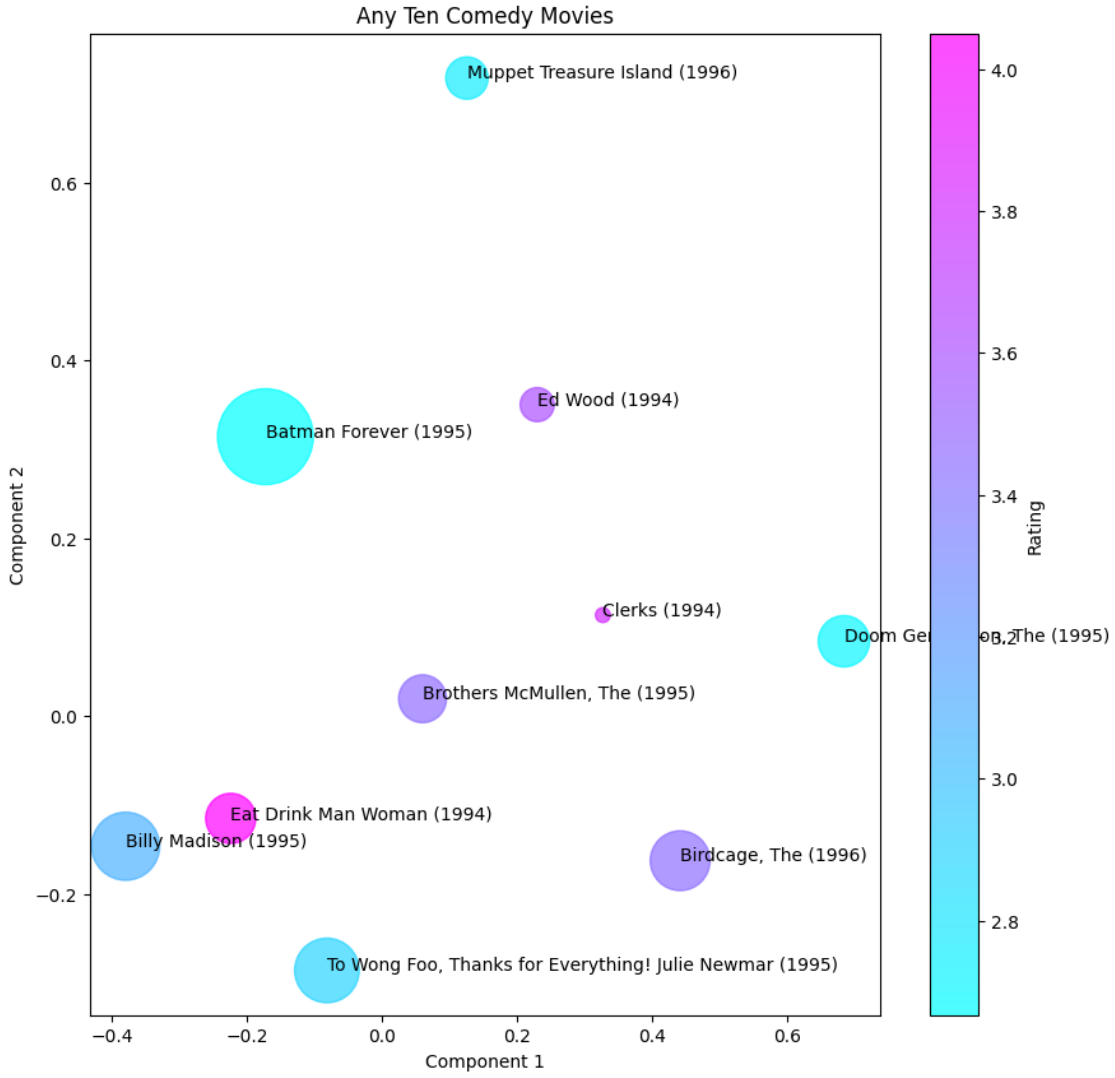
Below are the six visualizations resulting from Method 3.

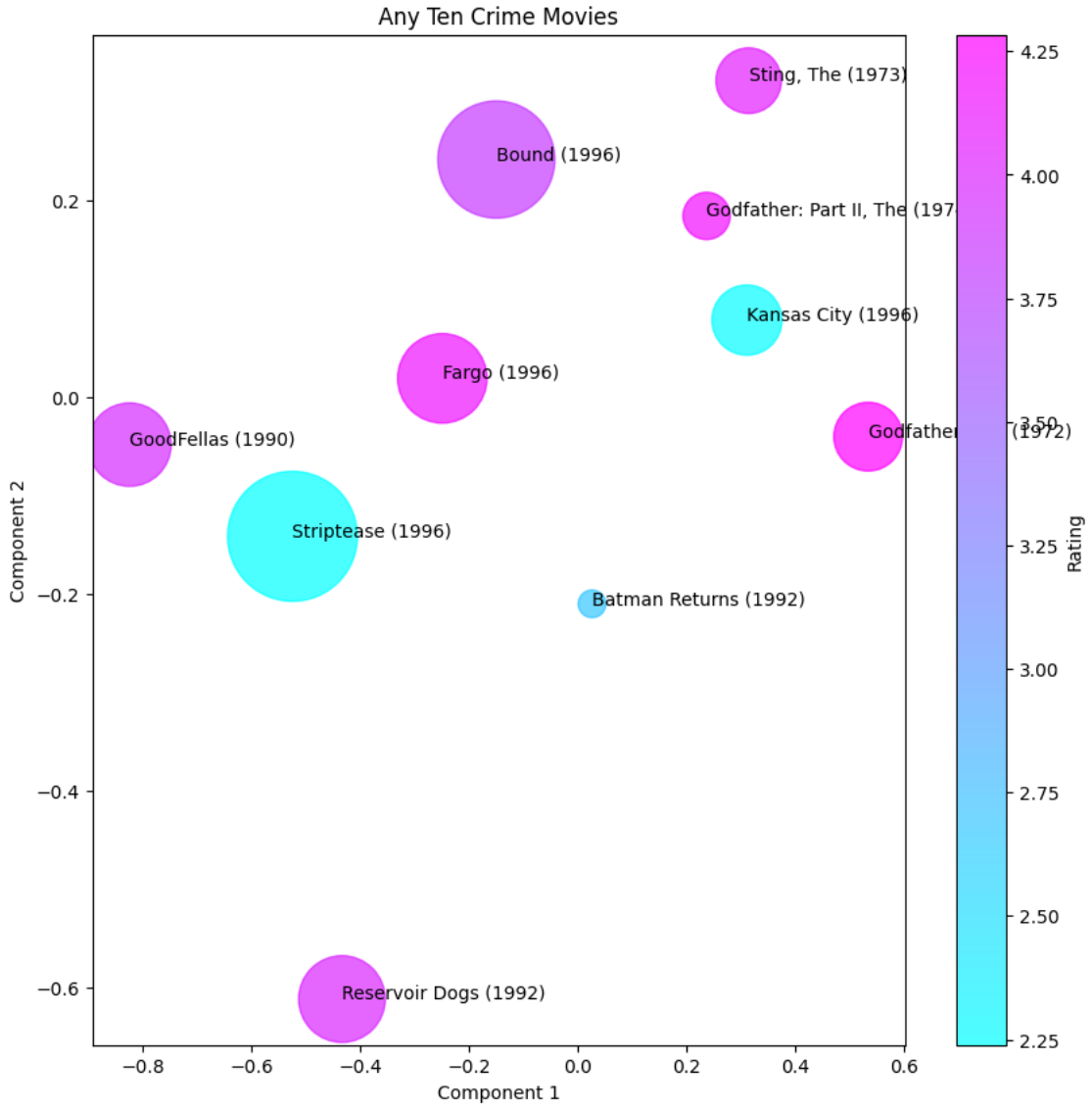



Ten Most Popular Movies
English Patient, The (1996)
Independence Day (ID4) (1996)
Star Wars
7)
Fargo (1996)
Return of the Jedi (1983)
Air Force One (1997)
Scream (1996)
Liar Liar (1997)
Contact (1997)
Toy Story (1995)
Component 1
Component 2
Rating
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
4.2
4.0
3.8
3.6
3.4
3.2


Ten Best Movies
Saint of F ashington, The (1993)
Someone Else's America (1995)
Entertaining Angels: The Dorothy Day Story (1996)
They Made Me a Criminal (1939)
Marlene Dietrich: Shadow and Light (1996)
Prefontaine (1997)
Great Day in Harlem, A 94)
Aiqing wansui (1994)
Santa with Muscles (1996)
Star Kid (1997)
Component 1
Component 2
Rating
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
5.4
5.2
5.0
4.8
4.6


Any Ten Action Movies
Muppet Treasure Island (1996)
Bad Boys (1995)
Apollo 13 (1995)
Batman Forever (1995)
GoldenEye (1995)
Get Shorty (1995)
Desperado (1995)
Rumble in the Bronx (1995)
Bravehea
95)
From Dusk Till Dawn (1996)
Component 1
Component 2
Rating
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
4.0
3.8
3.6
3.4
3.2
3.0
2.8


Any Ten Comedy Movies
Muppet Treasure Island (1996)
Ed Wood (1994)
Batman Forever (1995)
Clerks (1994)
Brothers McMullen, The (1995)
Eat Drink Man Woman (1994)
Billy Madison (1995)
Birdcage, The (1996)
To Wong Foo, Thanks for Everything! Julie Newmar (1995)
Component 1
Component 2
Rating
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6
4.0
3.8
3.6
3.4
3.2
3.0
2.8



Conclusions from plots

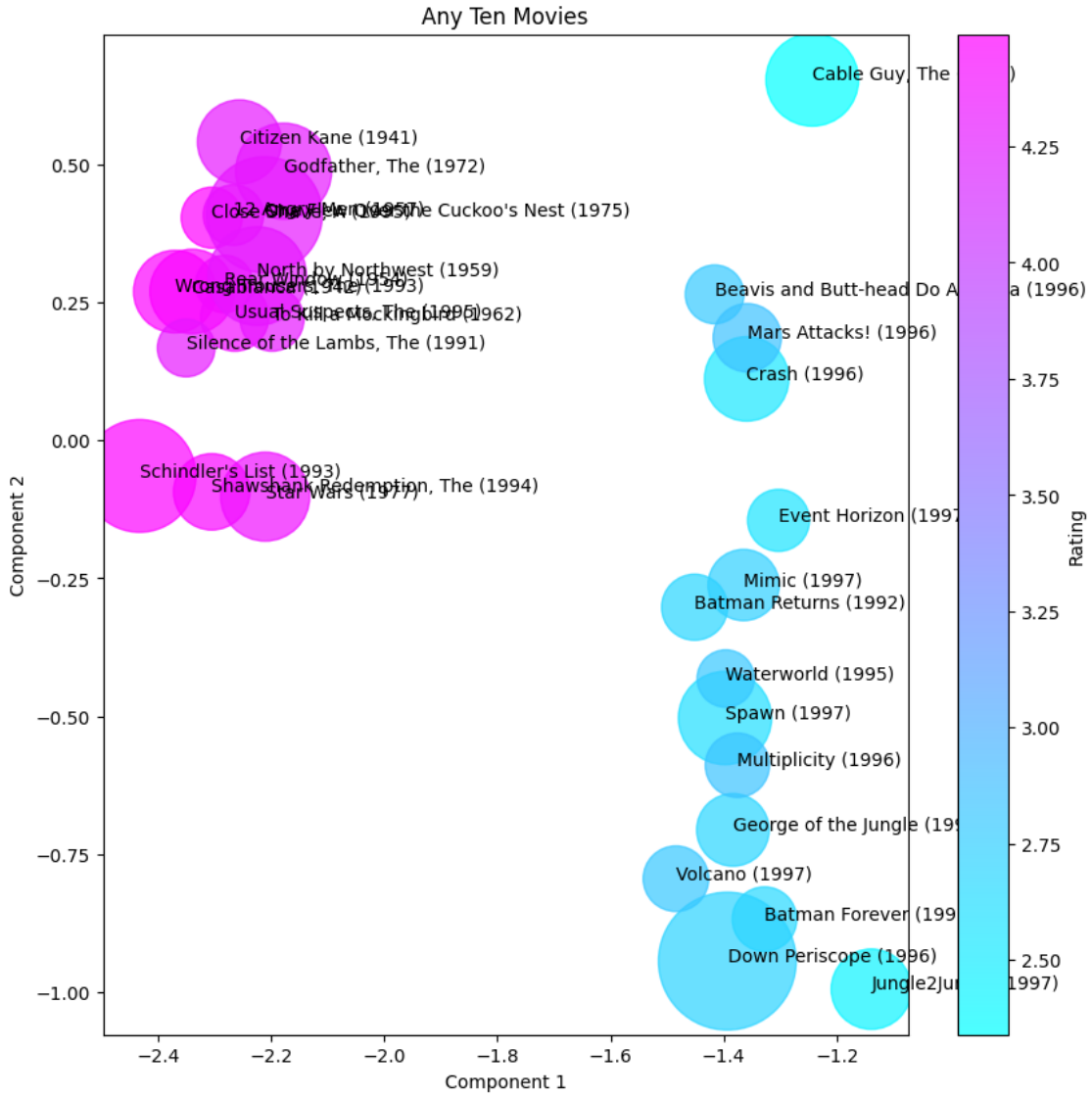
First, we expected to not see any kind of strong trend when it came to the visualization of the best movies. There aren't many reviews included for the best movies so a development of a strong trend would be difficult and not very likely.

For the visualization of the most popular movies, we expected some sort of general trend due to the large amount of ratings. For method 1, we notice a general trend where movies in the top right were the popular movies with relatively lower ratings (around 3 and colored blue in the visualization) and movies with

higher relative ratings were towards the left and lower. A line could be drawn from top right to the lower left to show this trend. The use of colors in the visualization make it easily visible. In method 2 and method 3, the overall presence of movies along this diagonal is followed but the trend with ratings is not. Combining these two, we can suggest that popular movies are more closely related to the values of the upper right and lower left regions of the graph. There again was no observed trend in the distribution of the best movies in methods 2 and 3.

The visualizations of the three genres also provide interesting insights. Many of the action movies were clumped together in Method 3. Other methods had less observed clustering and this may just be due to chance of randomly selecting ten action movies. The other genres had no observed trends when analyzed by Method 3. In method 1, a trend can be seen in all three genres. Movies with lower ratings are farther right on the graph, while movies with higher ratings are more likely to be on the left side of the graph pointing to positioning based on Component 1. Additionally, the crime genre visualization from Method 1 shows clustering by rating in the lower left corner specifically. Compared to the other methods, method 1 seems to show the most obvious trends and clustering. In method 2, the genres, comedy and action, do not seem to show many trends. For crime movies, all movies except one are on the right side of the graph. This could be indicative of a trend or could just be random based on picking of the movies. In method 3, for action movies, some clustering on the left side of the graph can be seen. This could tell us that this is where the majority of action movies are located. Overall, the plots of the movies of a specific genre were not similar for the methods. Specific movies were not in the same general areas and the different methods led to very different graphs.

The methods all offer valuable information and we are pleased with our visualizations. Our special plot is attached below and an in-depth discussion of it is included in our Piazza post.


Any Ten Movies
Cable Guy, The
Citizen Kane (1941)
Godfather, The (1972)
One Flew Over the Cuckoo's Nest (1975)
North by Northwest (1959)
Usual Suspects, The (1995)
To Kill a Mockingbird (1962)
Silence of the Lambs, The (1991)
Beavis and Butt-head Do A
a (1996)
Mars Attacks! (1996)
Crash (1996)
Schindler's List (1993)
Shawshank Redemption, The (1994)
Star Wars (1977)
Event Horizon (199
Mimic (1997)
Batman Returns (1992)
Waterworld (1995)
Spawn (1997)
Multiplicity (1996)
George of the Jungle (19
Volcano (1997)
Batman Forever (199
Down Periscope (1996)
Jungle2Jun
1997)
Component 1
Component 2
Rating
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−2.4
−2.2
−2.0
−1.8
−1.6
−1.4
−1.2
4.25
4.00
3.75
3.50
3.25
3.00
2.75
2.50