# 1 Introduction [5 points]

- Group members:
  1. Sri Aditya Deevi
  2. Palak Purohit
  3. Princekumar Kothadiya
- Team Name : Popcorn Party!
- Colab link:
  - Basic Visualization –
    https://colab.research.google.com/drive/1Ku8h984HtybKBSoZ0h9METEF2b6YGRRO
  - Matrix Factorization Techniques and Corresponding Visualization –
    https://colab.research.google.com/drive/1LPOipBxJtg-71tgYCsmfhVe5bR0sETbJ#scrollTo=thWZUNDt9OXl
- Piazza link:
  https://piazza.com/class/lbv0docn6037fw/post/550
- Division of labor:
  - Basic Visualization* : Sri Aditya Deevi, Palak Purohit and Princekumar Kothadiya
  - Data Preprocessing and Visualization Code Structure : Sri Aditya Deevi
  - Basic SVD : Princekumar Kothadiya
  - SVD+Bias and SVD+Bias+Global Bias : Palak Purohit
  - SVD++ : Sri Aditya Deevi
  - Piazza Visualization and Post * : Sri Aditya Deevi, Palak Purohit and Princekumar Kothadiya
  - Report and Colab * : Sri Aditya Deevi, Palak Purohit and Princekumar Kothadiya
- Packages used:
  - Numpy
  - Matplotlib
  - Seaborn
  - Surprise
  - Sklearn
  - Pandas

*Equal Contributions

# 2 Basic Visualizations [20 points]

## Discussion

Visualizing all the ratings of all movies in a single plot would make the plot crowded and extremely unclear. So we came up with multiple visualizations to capture multiple aspects of the given dataset as follows:

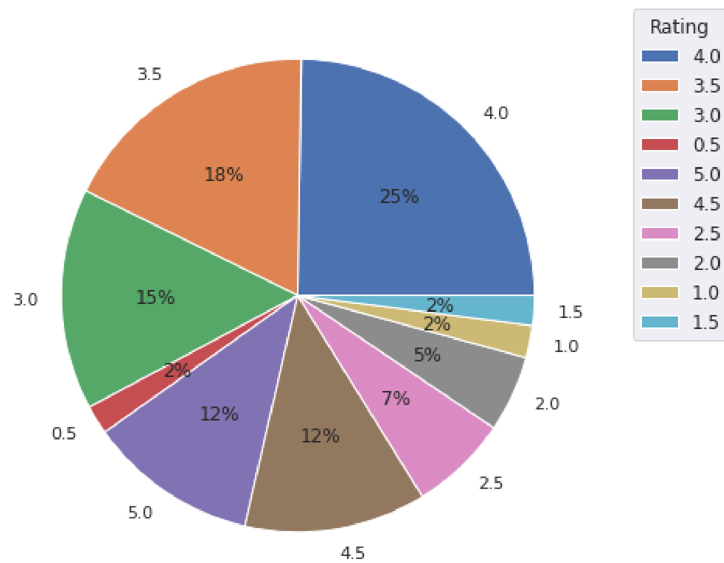


Figure 1: All ratings in the MovieLens Dataset : Types of Ratings

From Figure. 1 we can observe that the most popular rating by users is 4.0 (25 % of all ratings). We can say (based on the data in this dataset), users tend to give high ratings or they generally don't rate at all. (Users giving ratings less than 2.0 is only 6 % percent of all ratings).

To construct the heatmap in Figure.2, we fill the missing entries with zeros. So the black portion (which is the majority) in the heatmap represents the missing entries. This shows the sparsity of the given data.

Figure. 3 indicates that the given dataset has very few movies recent i.e. movies that were released after 2015.

Both Figure. 3 and Figure. 4 clearly show that the average rating given by users is increasing as time passes. This can indicate either better movies being made (or) an increase in the audience that give a good rating to movies.

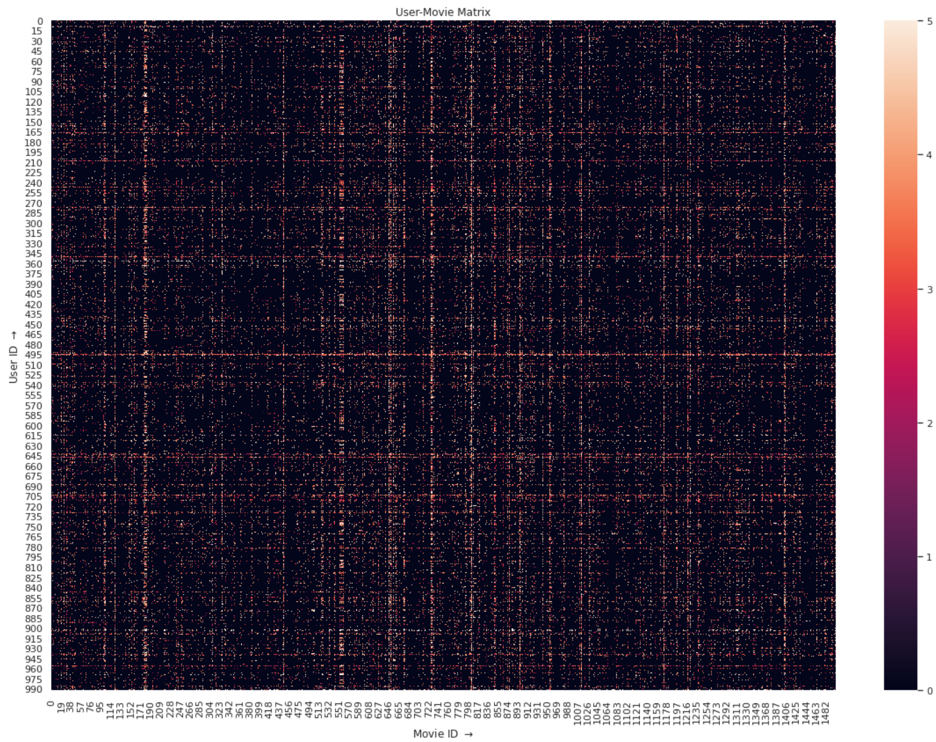


Figure 2: All ratings : Heatmap Representation *Sparse* User Movie Matrix

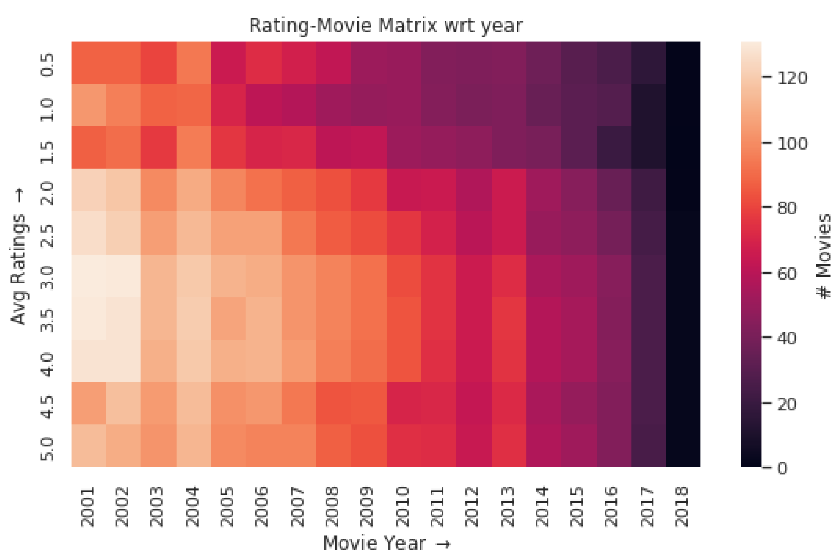


Figure 3: All ratings : Heatmap Representation of Rating-Movie Matrix (w.r.t year)

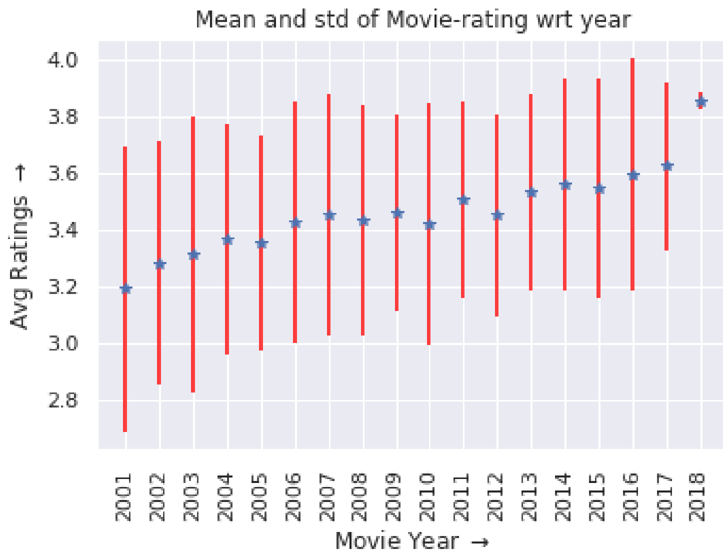


Figure 4: All ratings : Average Rating vs. Year

We considered plotting stacked histograms for the following basic visualization plots, since in each category we have just 10 movies.

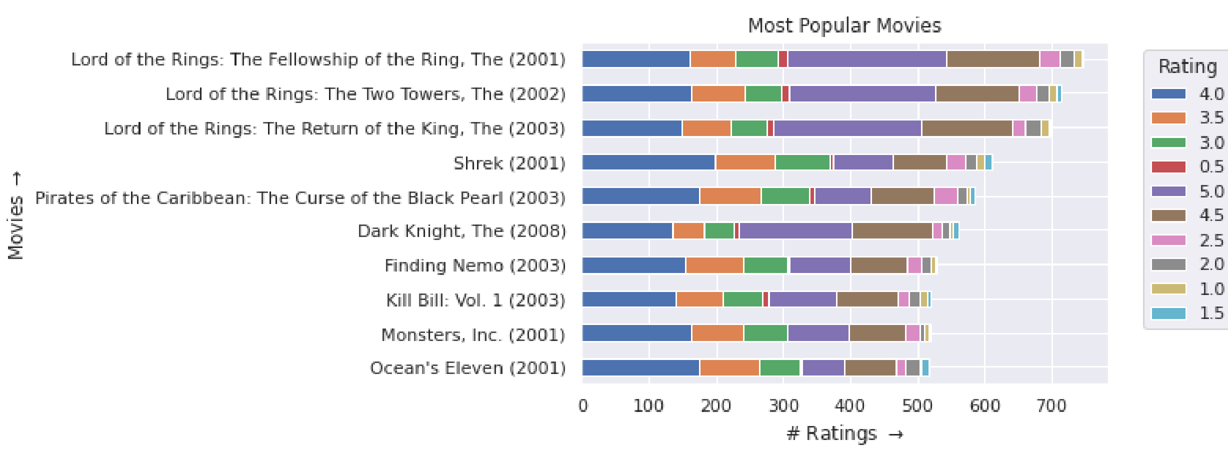


Figure 5: All ratings of the ten most popular movies (movies which have received the most ratings).

Comparing Figure. 5 and Figure. 6, we can see that best movies are having few (but high) ratings as compared to most popular movies. Almost all of the top rated movies have less than 50 ratings (*City of God* being an exception with around 250 ratings). Also for both of them, number of rating-4, rating-4.5 and rating-5 are higher compared to other rating value.

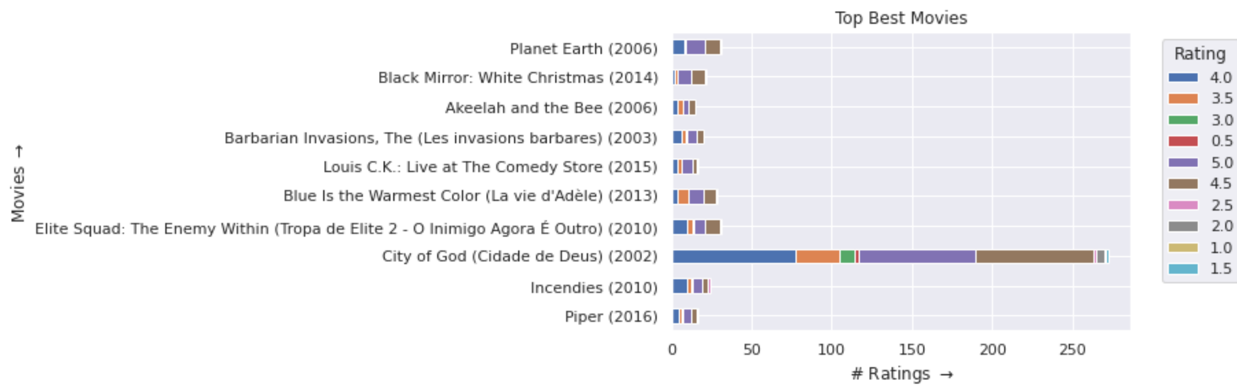


Figure 6: All ratings of the ten best movies (movies with the highest average ratings).

Comparing figure 7, 8 and 9, we can see that in-spite of having higher number of movies in genre *Documentary*, the highest number of rating is around 200 where as for *Musical* on the second position has around highest number of rating value of 350 and at last, *Western* is having highest number of rating of 300.

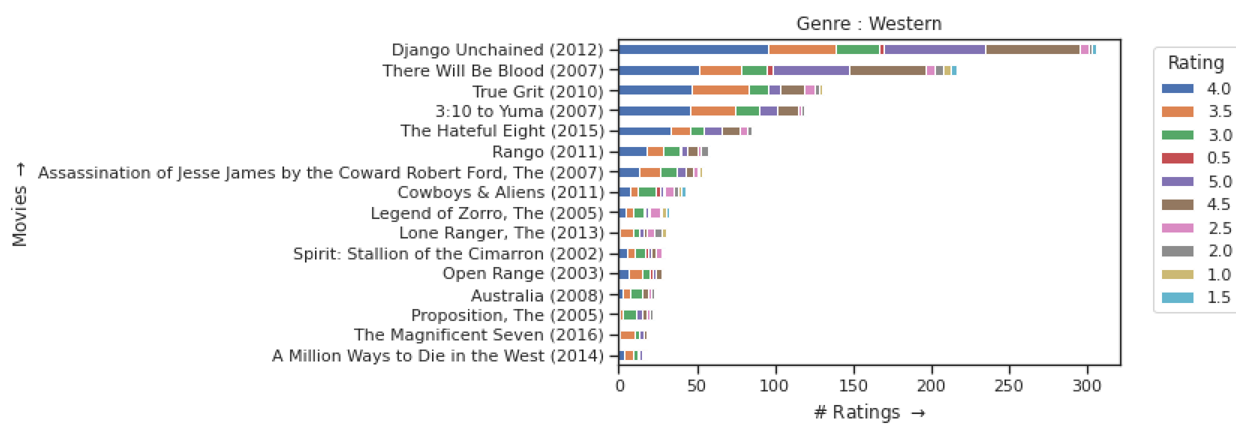


Figure 7: All ratings of movies from the genre *Western*.

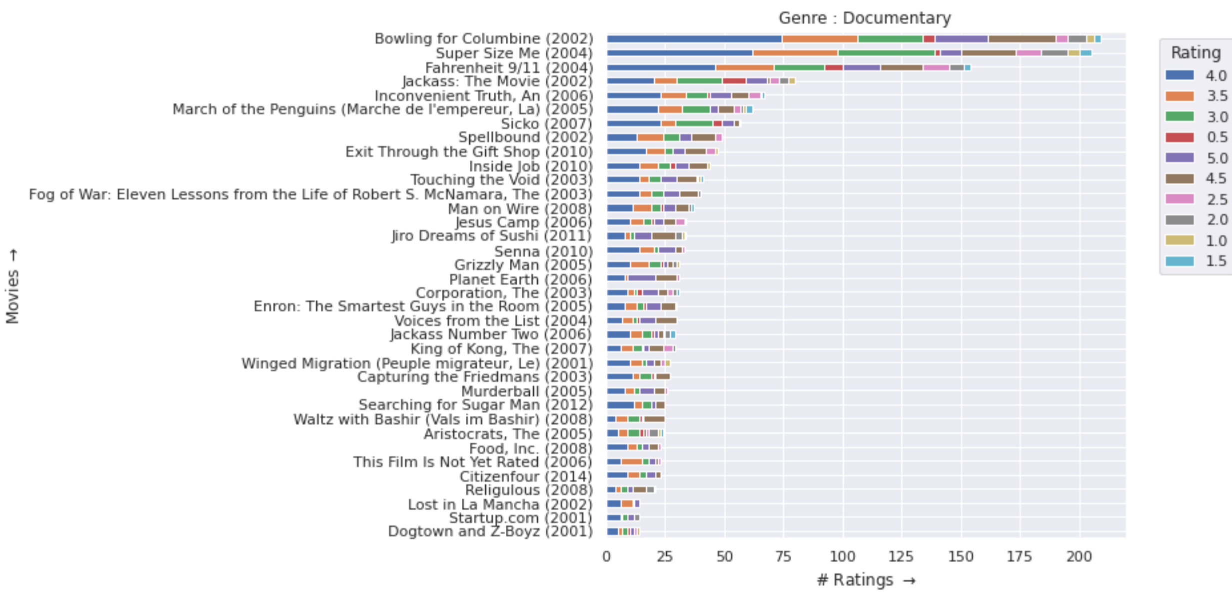


Figure 8: All ratings of movies from the genre *Documentary*.

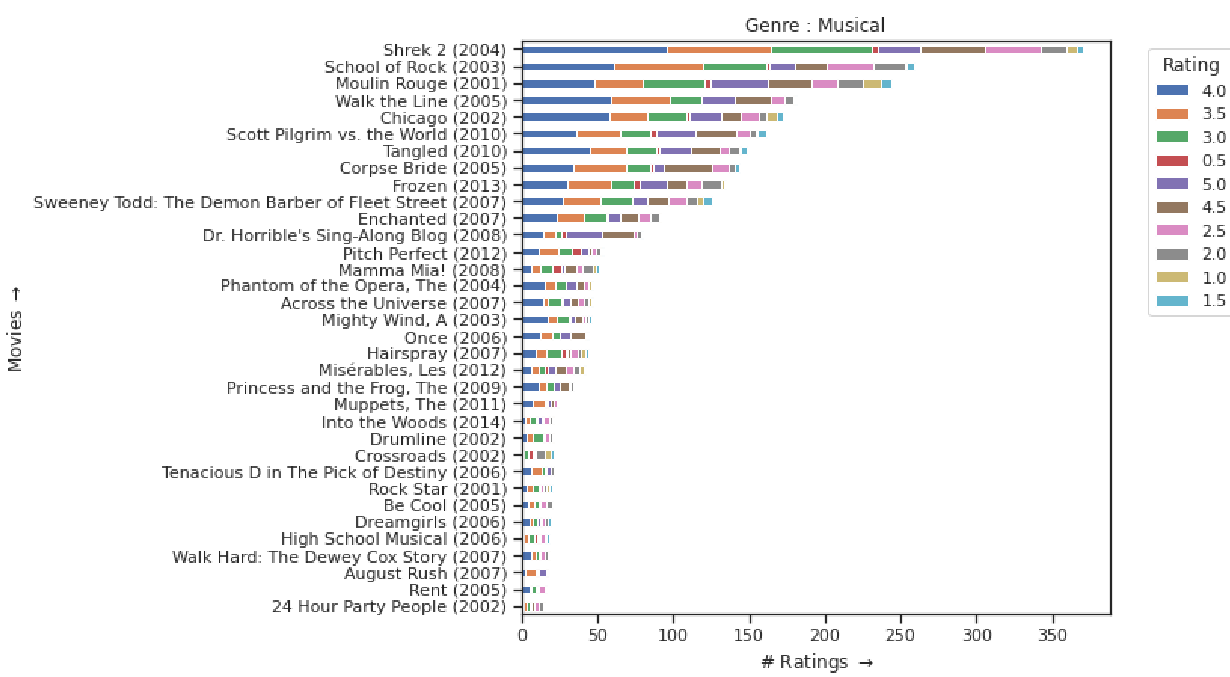


Figure 9: All ratings of movies from the genre *Musical*.

## 3 Matrix Factorization Visualizations [60 points]

### Matrix Factorization Methods

In this section, we describe the various matrix factorization methods we considered in this project. The notation that we will using is as follows:

$Y_{ij} \mapsto$ the true rating of user $i$ for Movie $i$
$\hat{Y}_{ij} \mapsto$ the estimated rating of user $i$ for movie $j$
$u_i \mapsto$ Latent Vector corresponding to user $i$
$v_i \mapsto$ Latent Vector corresponding to movie $j$
$a_i \mapsto$ Bias corresponding to user $i$
$b_i \mapsto$ Bias corresponding to movie $j$
$\mu \mapsto$ Global Bias = Average of all obsv. in $Y$

$z_j \mapsto$ Implicit Rating Factor for movie $i$
$N_i \mapsto$ the set of all items rated by user $i$.
$\lambda \mapsto$ Regularization Strength
$\eta \mapsto$ Learning Rate
$U \mapsto$ User (Latent) Matrix
$V \mapsto$ Movie (Latent) Matrix
$Y \mapsto$ True Rating Matrix

Some implementation details are as follows:

1. Splitting the given dataset into training and test sets was done. We made sure that atleast one instance of all users and all movies are included in the training split. This is to ensure that the model can learn the corresponding vectors for both the users and movies in a good manner.
2. We used K-Fold $(K = 10)$ Cross Validation to determine the optimal hyperparameters $(\eta, \lambda)$. We set the latent vector size $k = 20$.

#### Basic SVD

The basic SVD method is the same as what studied in lectures and implemented in Homework-5. The prediction expression for this is:

$$\hat{Y}_{ij} = u_i^T v_j$$

Using this predicted value and regularization term, we can write the regularized mean-square error function as follows:

$$E(U, V) = \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + \sum_{(i,j)\in S}\left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

The gradient using the error function has been calculated and the SGD update rule* for this method is as follows:

*We used the same gradient as in homework-5 while considering effective $\eta$ as $2\eta$ and effective $\lambda$ as $\frac{\lambda}{2}$ since there is a difference of multiplication factor of $\frac{1}{2}$ in squared error term compared to homework-5.

$$u_i \leftarrow u_i - 2\eta\left(\frac{\lambda}{2}\, u_i - v_j\left(y_{ij} - u_i^T v_j\right)\right)$$
$$v_j \leftarrow v_j - 2\eta\left(\frac{\lambda}{2}\, v_j - u_i\left(y_{ij} - u_i^T v_j\right)\right)$$

We can observe that train and test errors are higher than other methods (From Table 1).

**SVD + Bias**

SVD with bias is similar to the Basic SVD algorithm except that it also incorporates the global expectation of a movie's average rating and the average of each user's rating. This is done by introducing vectors ai and bj which represent this deviation.
The predicted value now becomes:

$$\hat{Y}_{ij} = u_i^T v_j + a_i + b_j$$

The error function is again the MSE with the regularized error as follows:

$$E(U, V, a, b) = \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + \sum_{(i,j)\in S}\left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

The SGD update equations then become:

$$u_i \leftarrow u_i - 2\eta\left(\frac{\lambda}{2}\, u_i - v_j\left(y_{ij} - a_i - b_j - u_i^T v_j\right)\right)$$
$$v_j \leftarrow v_j - 2\eta\left(\frac{\lambda}{2}\, v_j - u_i\left(y_{ij} - a_i - b_j - u_i^T v_j\right)\right)$$
$$a_i \leftarrow a_i - 2\eta\left(y_{ij} - a_i - b_j - u_i^T v_j\right)$$
$$b_j \leftarrow b_j - 2\eta\left(y_{ij} - a_i - b_j - u_i^T v_j\right)$$

We can see from Table 1 that the training loss is lower than Basic SVD, thus meaning that adding the parameters a and b leads to better performance.

**SVD + Bias + Global Bias**

This method also includes a global bias $\mu$ of the model which is the average of all observations in Y. Then, a can be thought of as the user-specific deviation from $\mu$ and b can be thought of as movie-specific deviation from $\mu$. The prediction in this case is:

$$\hat{Y}_{ij} = u_i^T v_j + \mu + a_i + b_j$$

The error function is slightly different from SVD + Bias in that is also includes the norm of a and b in the regularization. It can be written as:

$$E(U, V, a, b, \mu) = \frac{\lambda}{2}\left(\|U\|^2 + \|V\|^2 + \|a\|^2 + \|b\|^2\right) + \sum_{(i,j)\in S}\left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

The gradient update equations for u and v remain the same but those for a and b become as follows:

$$a_i \leftarrow a_i - 2\eta\left(\frac{\lambda}{2}\, a_i - \left(y_{ij} - \mu - a_i - b_j - u_i^T v_j\right)\right)$$
$$b_j \leftarrow b_j - 2\eta\left(\frac{\lambda}{2}\, b_j - \left(y_{ij} - \mu - a_i - b_j - u_i^T v_j\right)\right)$$

**SVD++**

SVD++ is an extension of the SVD algorithm that takes into account the *implicit* ratings of users. The prediction equation is as follows:

$$\hat{Y}_{ij} = \mu + a_i + b_j + v_j^T\left(u_i + |N_i|^{-\frac{1}{2}}\sum_{k\in N_i} z_k\right)$$

In this method, as we can see the user is modellled as $\left(u_i + |N_i|^{-\frac{1}{2}}\sum_{k\in N_i} z_k\right)$ where $u_i$ indicates the explicit ratings aspects and $|N_i|^{-\frac{1}{2}}\sum_{k\in N_i} z_k$ captures the implicit feedback concept.

This method is most effective when the dataset has implicit feedback information about the users, in addition to explicit feedback. In general, implicit feedback data is much higher as it depends on the preferences of the user. For example, if a user buys an item but leaves no rating. This strategy is also considered to be effective, even in circumstances where independent implicit feedback is lacking, one can capture a meaningful signal by accounting for which objects users rate, regardless of their rating value.

The error function used is :

$$E(U, V, a, b, z) = \lambda\left(\|U\|^2 + \|V\|^2 + \|a\|^2 + \|b\|^2 + \|z\|^2\right) + \sum_{(i,j)\in S}\left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

The SGD update equations for the parameters are as follows:

$$u_i \leftarrow u_i + \eta\left(2v_j\left(Y_{ij} - \hat{Y}_{ij}\right) - 2\lambda u_i\right)$$
$$v_j \leftarrow v_j + \eta\left(2\left(u_i + |N_i|^{-\frac{1}{2}}\sum_{k\in N_i} z_k\right)\left(Y_{ij} - \hat{Y}_{ij}\right) - 2\lambda v_j\right)$$
$$a_i \leftarrow a_i + \eta\left(2\left(Y_{ij} - \hat{Y}_{ij}\right) - 2\lambda a_i\right)$$
$$b_j \leftarrow b_j + \eta\left(2\left(Y_{ij} - \hat{Y}_{ij}\right) - 2\lambda b_j\right)$$
$$z_k \leftarrow z_k + \eta\left(2\,|N_i|^{-\frac{1}{2}}\,v_j\left(Y_{ij} - \hat{Y}_{ij}\right) - 2\lambda z_k\right)$$

As shown in Table 1, the SVD++ performs very similar to the SVD with bias terms cases but performs better then Basic SVD. This is because of smaller size of given Dataset (subset of MovieLens) and the fact that there is no implicit feedback signal present in the given dataset.

| **Method** | **Training Error** | **Testing Error** | **Optimum Hyperparams** |
|---|---|---|---|
| **Basic SVD** | 0.4652 | 0.7685 | $2\eta$ = 0.01, $\frac{\lambda}{2}$ = 0.1 |
| **SVD + Bias** | 0.4241 | 0.7354 | $2\eta$ = 0.01, $\frac{\lambda}{2}$ = 0.1 |
| **SVD + Bias + Global Bias** | 0.4249 | 0.7379 | $2\eta$ = 0.01, $\frac{\lambda}{2}$ = 0.1 |
| **SVD++** | 0.4118 | 0.7446 | $\eta$ = 0.01, $\lambda$ = 0.1 |

Table 1: Quantitative Results from Various Matrix Factorization Methods

## Visualizations

### Visualization Plots

In this section, we present the results obtained by visualizing the latent vectors learnt using various matrix factorization methods. It should be noted that since we are using SGD, each time we run the algorithm may converge differently and absolute configuration of the axes (with semantic meaning) may not be the same. So in the plots (even within the same matrix factorization method) relative configuration of axes (with semantic meaning) is preserved across different runs.

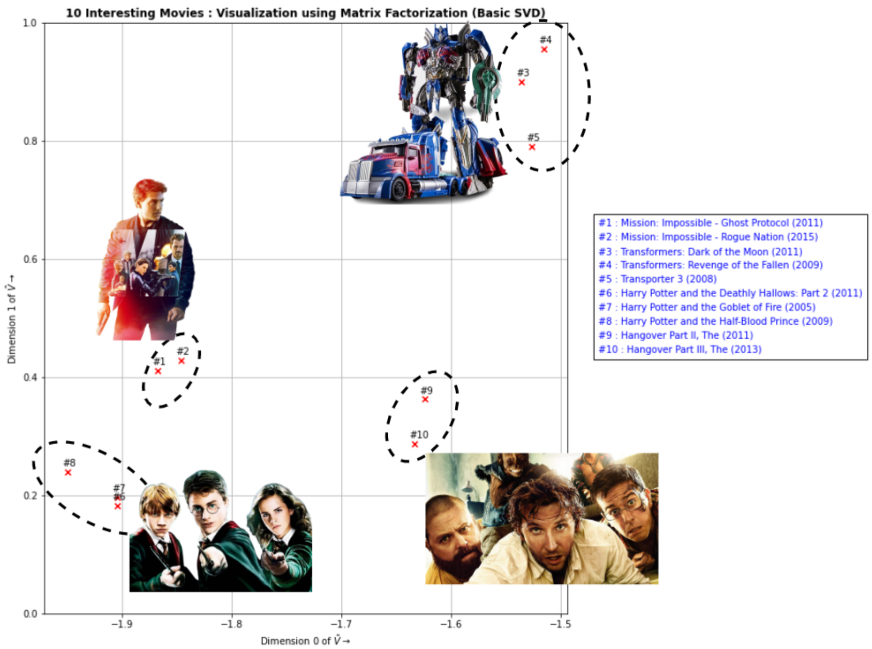


Figure 10: **HW5 A:** Ten movies from the MovieLens dataset.

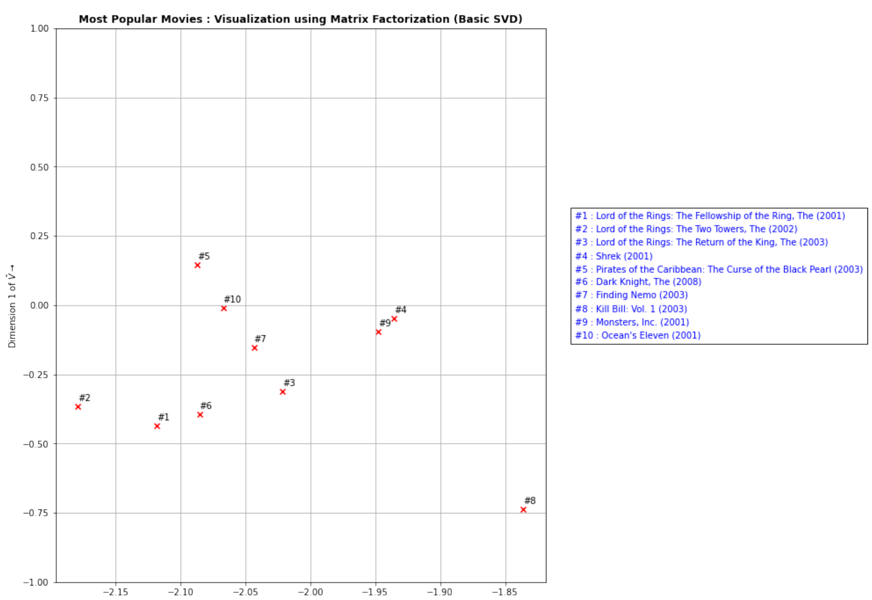


Figure 11: **HW5 B:** The ten most popular movies (movies which have received the most ratings).

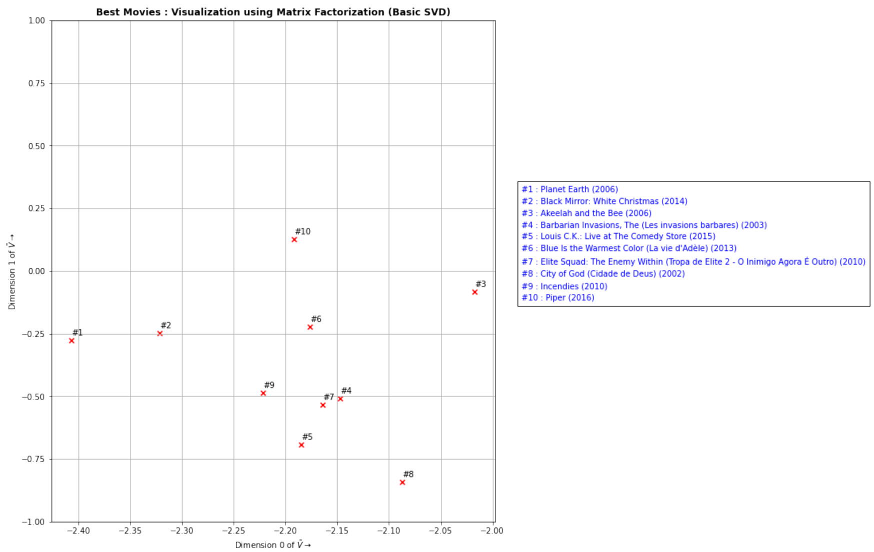


Figure 12: **HW5 C:** The ten best movies (movies with the highest average ratings).

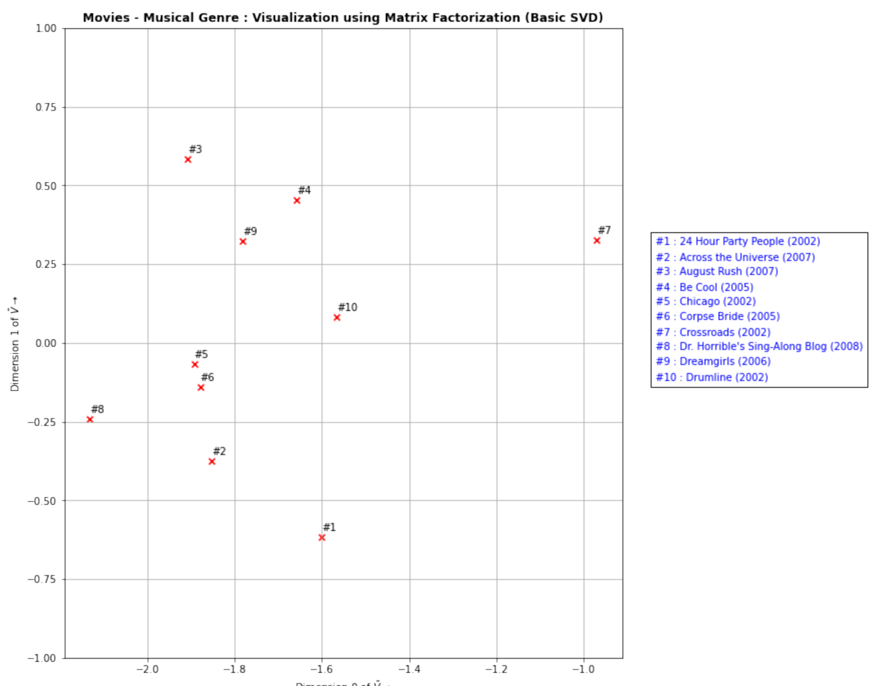


Figure 13: **HW5 D1:** Ten movies from the genre **Musical**.

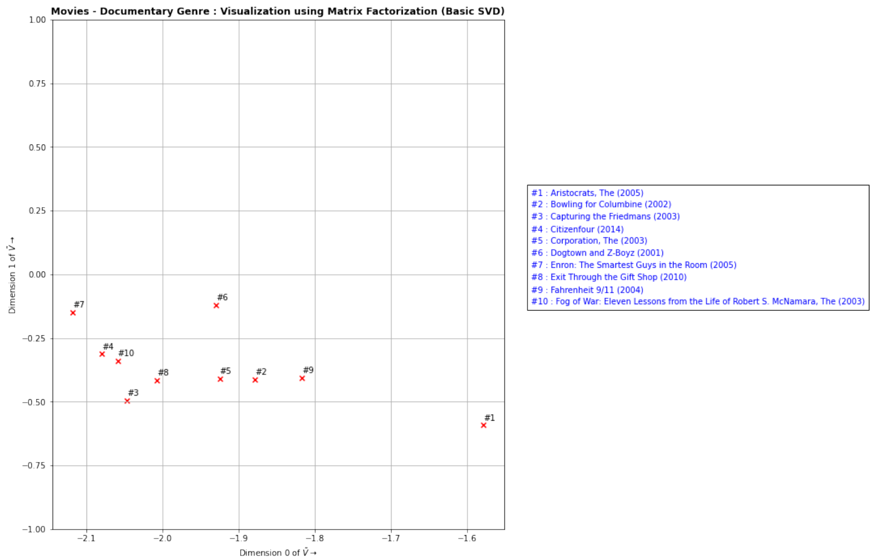


Figure 14: **HW5 D2:** Ten movies from the genre **Documentry**.

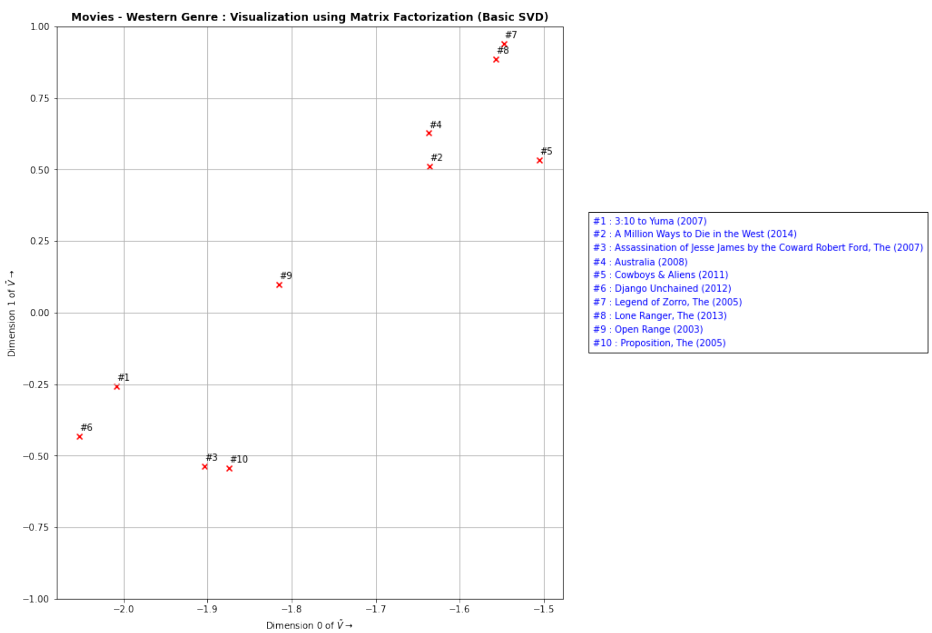


Figure 15: **HW5 D3:** Ten movies from the genre **Western**.

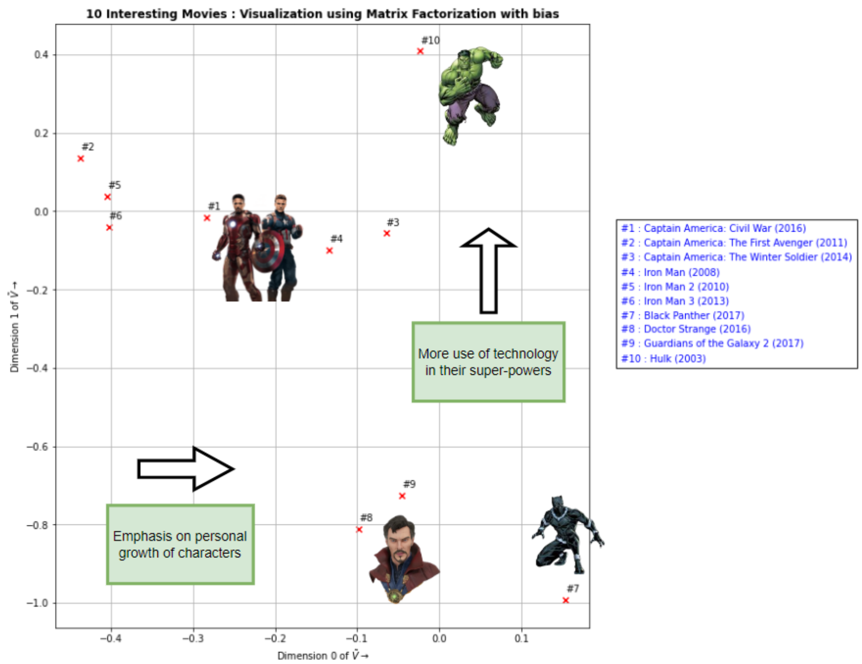


Figure 16: **Bias A:** Any ten movies of your choice from the MovieLens dataset.

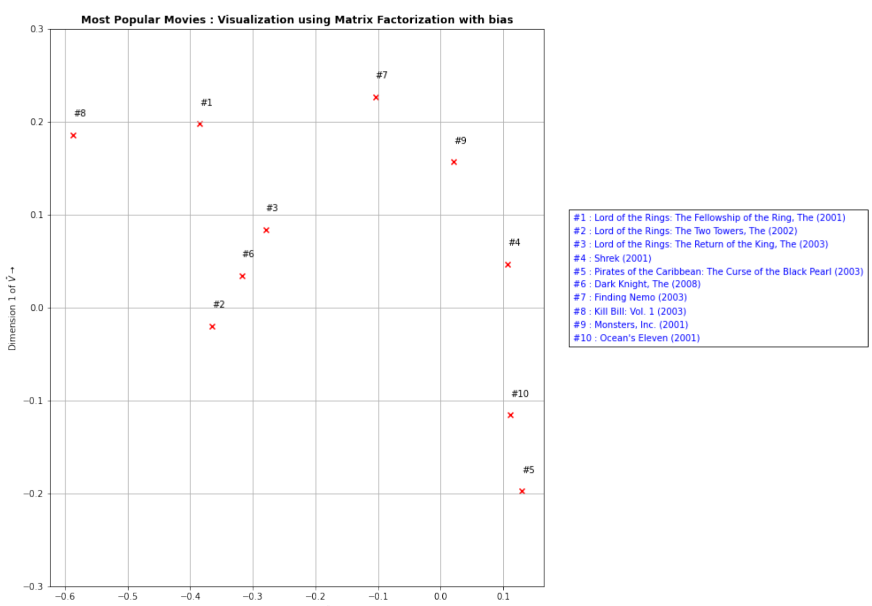


Figure 17: **Bias B:** The ten most popular movies (movies which have received the most ratings).

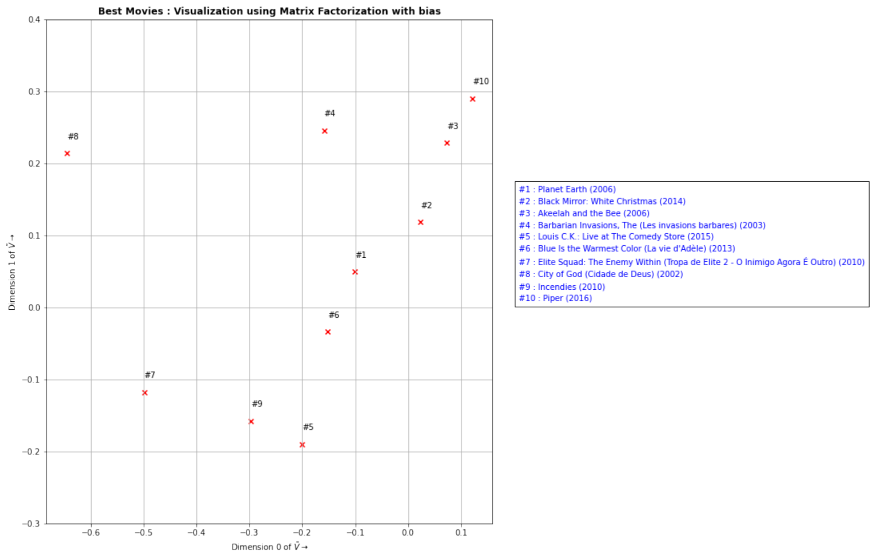


Figure 18: **Bias C:** The ten best movies (movies with the highest average ratings).

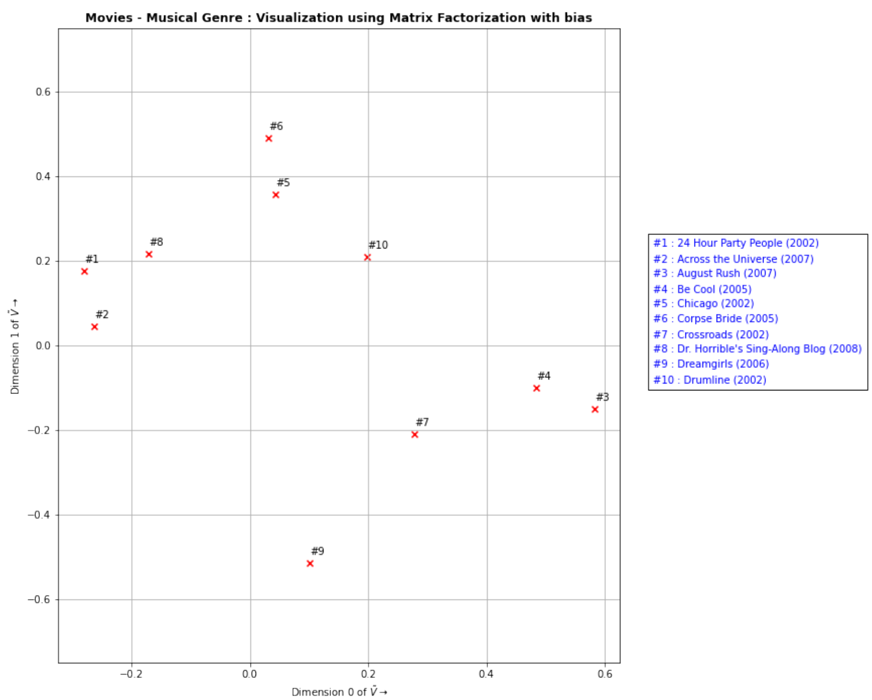


Figure 19: **Bias D1:** Ten movies from the genre [Musical].

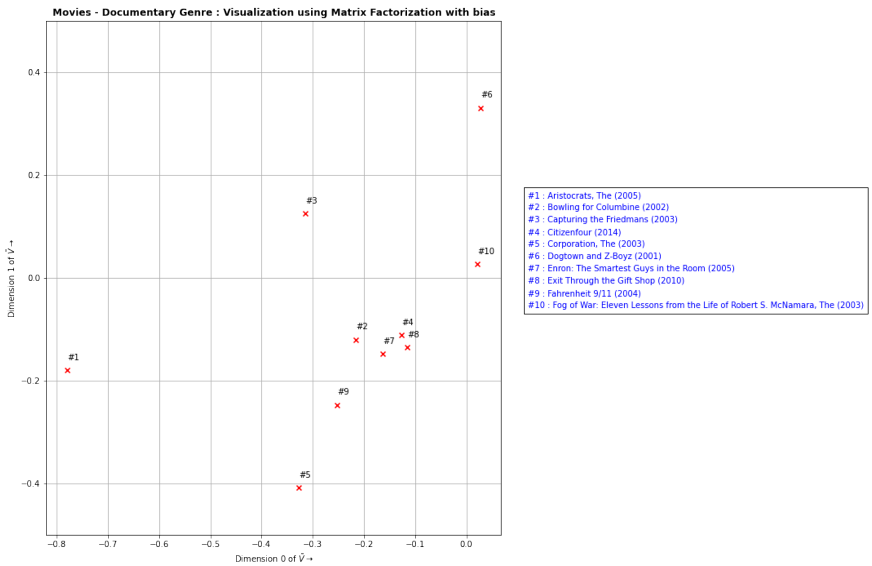


Figure 20: **Bias D2:** Ten movies from the genre [Documentary].

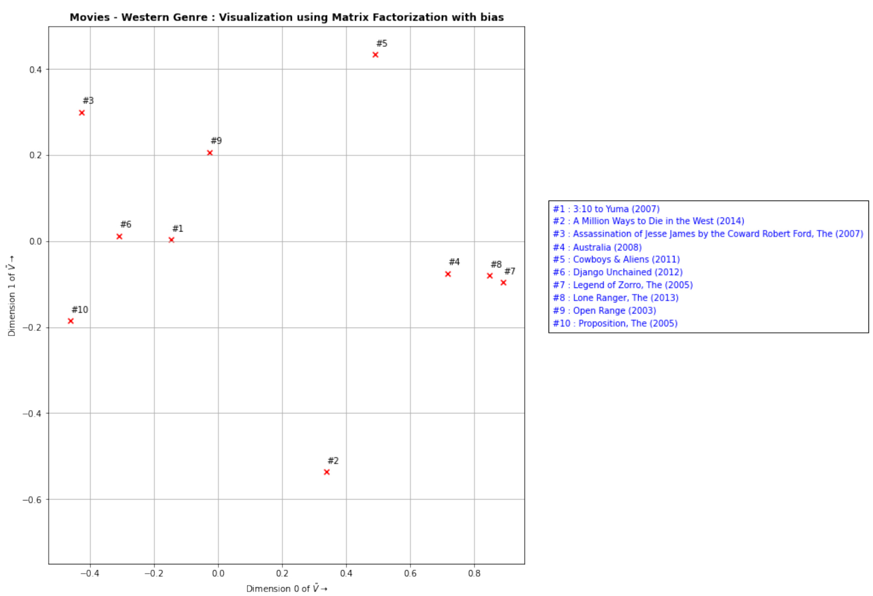


Figure 21: **Bias D3:** Ten movies from the genre [Western].

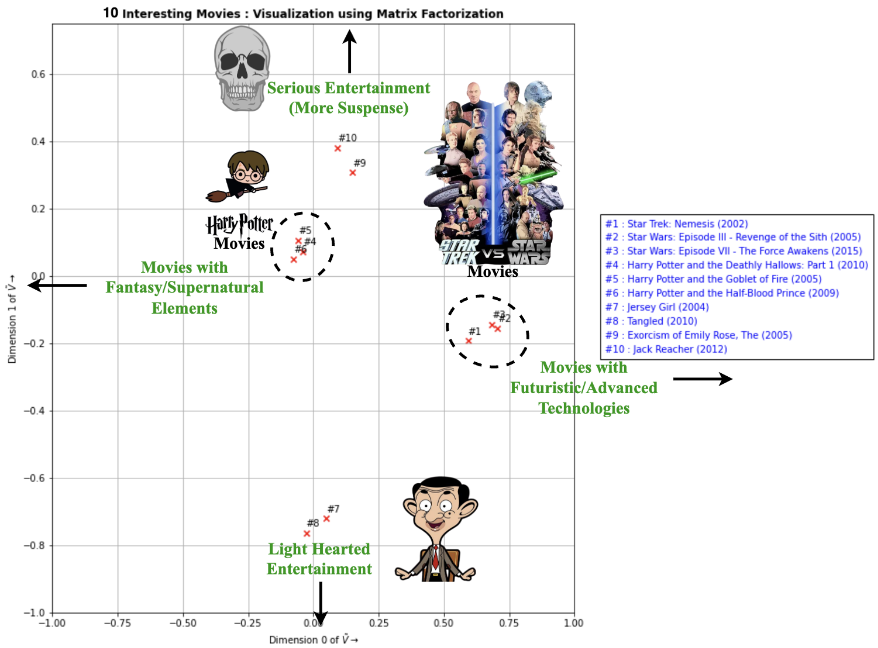


Figure 22: **COTS A:** Any ten movies of your choice from the MovieLens dataset.

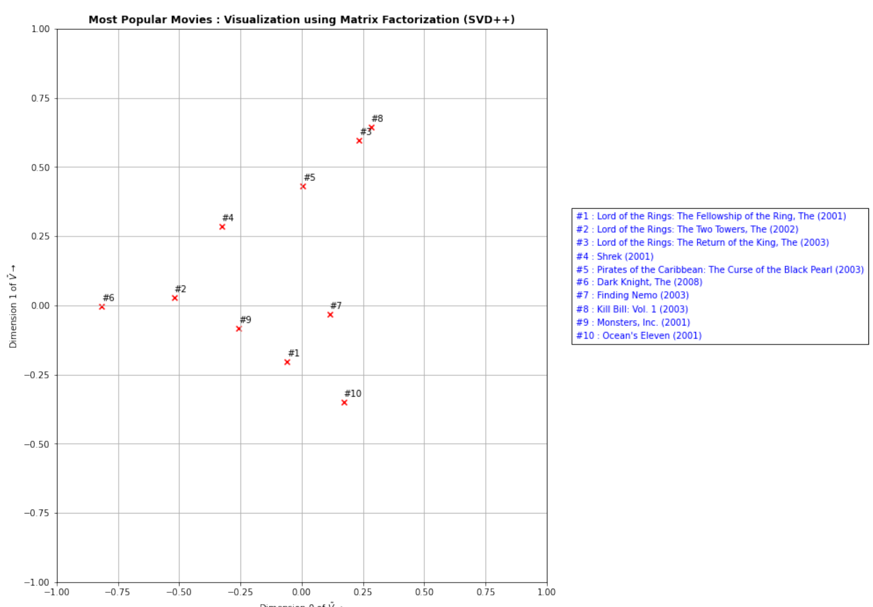


Figure 23: **COTS B:** The ten most popular movies (movies which have received the most ratings).

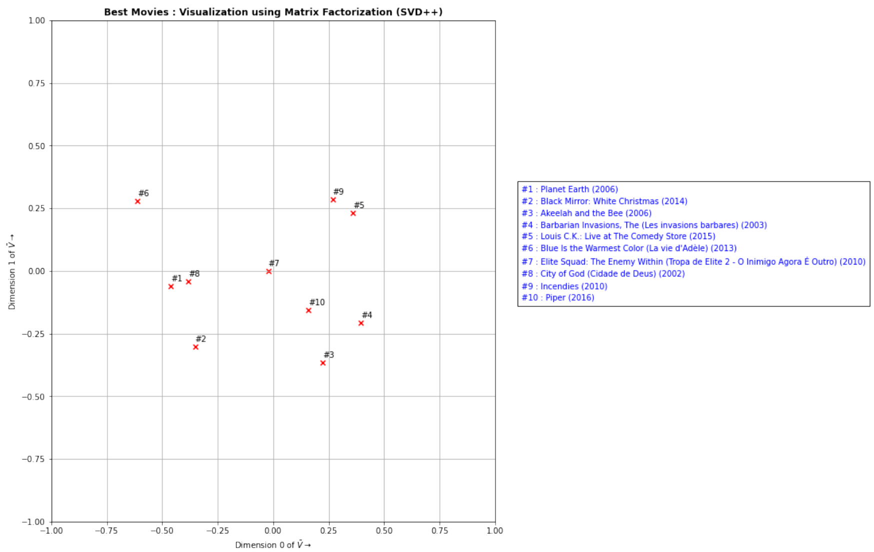


Figure 24: **COTS C:** The ten best movies (movies with the highest average ratings).

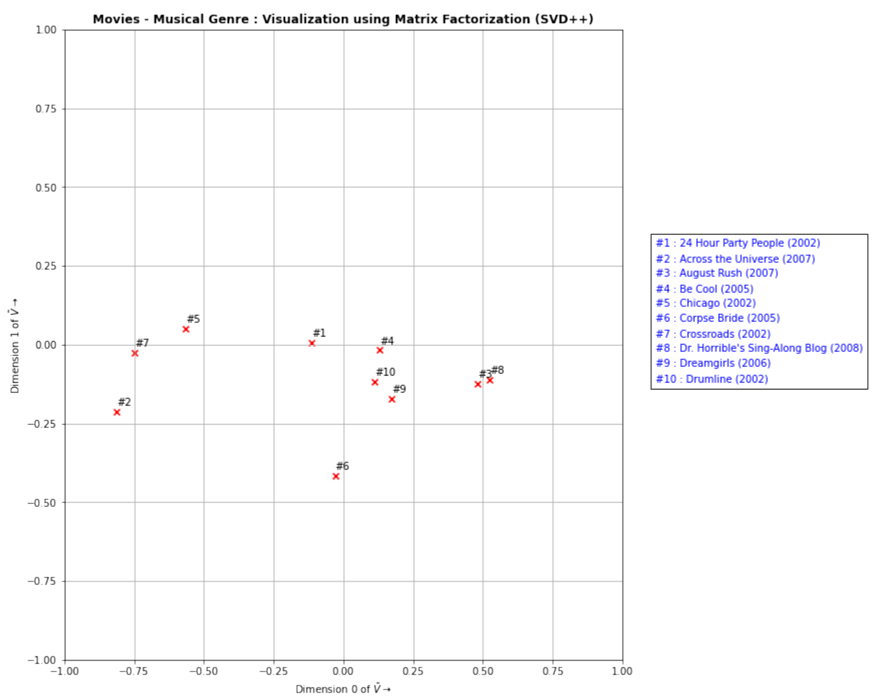


Figure 25: **COTS D1:** Ten movies from the genre [Musical].

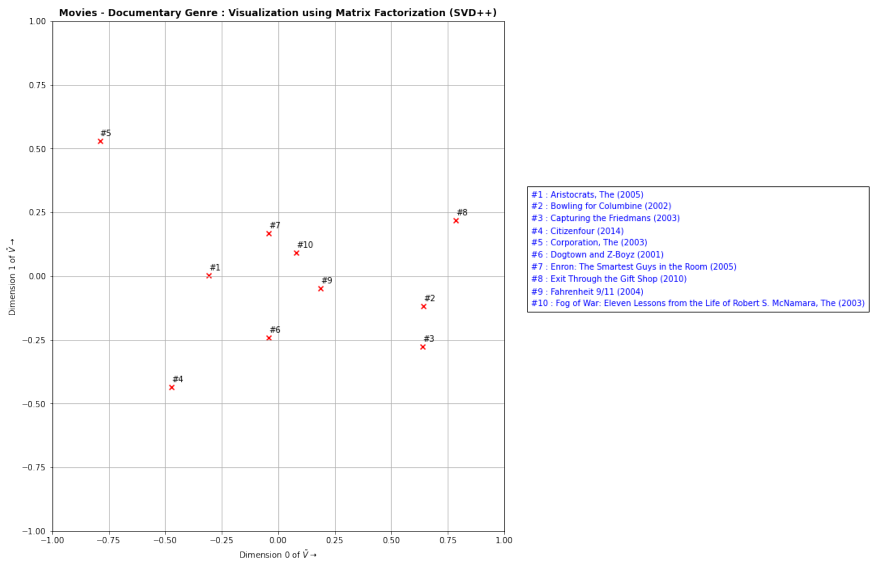


Figure 26: **COTS D2:** Ten movies from the genre [Documentary].

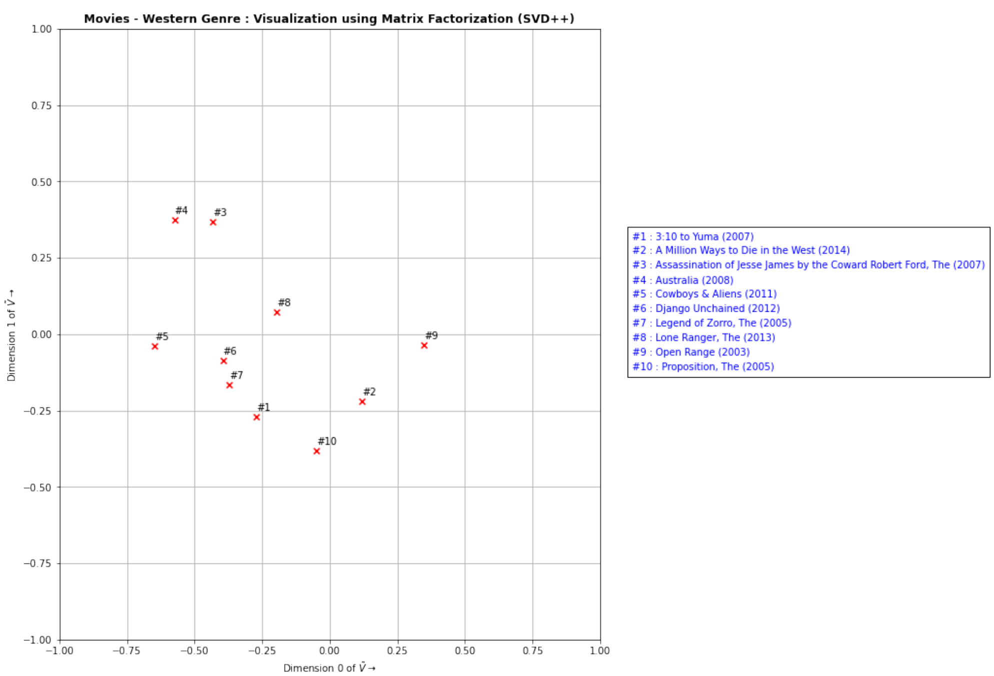


Figure 27: **COTS D3:** Ten movies from the genre [Western].

### Visualization Discussion

The following are some of our observations and corresponding analysis based on the visualizations obtained:

1. **Basic SVD** :–

   - From the figure 10, we can see that the movie sequels like harry potter movies, Mission impossible movies, Hangover movies and Transformer movies form a cluster as expected. This is because of the dedicated fan-bases of this famous movies who gives similar ratings to all sequels.
   - Other important point is that from the same figure 10 that, these movies are of different genes so they form the cluster at different locations in the plan. As an example, harry potter movies are having of fantasy/supernatural elements whereas transformers are more of sci-fi and action movie so they are diagonal apart in the plan. Similarly, MI is more of action and thriller movies whereas hangover is full of funny and comedy movies so they are also located far apart in the plane.
   - From the figure 11 of top 10 most popular movies, It can be observed that the Monsters, Inc. and Shrek movies are very close to each other since both are similar animated films and both feature non-human characters that have personalities and emotions that are relatable to human audiences. Another observation from the same plot is that movie kill bill is action and crime based movie which is opposite of almost all from the plot which ended up locating it far away from others.
   - In the plot 12 of 10 best movies, The movie Elite Squad: The Enemy Within and The Barbarian Invasions are closer to each other because in-spite of having different theme and different country audience, both movies share a similar idea of societal issues and the impact that they have on people's lives. So, both movies ended up rated similarly.
   - Now, from the figure 13, 14 and 15, we can see that *Musical* genre movies are having random pattern in the plane, Movies of *Documentary* genre are almost together and forms a pattern whereas *Western* genre movies are divided into two different regions in the plane with some exceptions. These shows that some of the picked movies from this genres are having few similarities and opposites.

2. **SVD+Bias** :–

   - In Figure 16, We plotted some of our favourite movies from the MCU and came up with this plot, which seemed fairly interesting to us (just like the plot of these movies). As we move towards the bottom on the y-axis, we find characters that use more mystical/supernatural powers (such as Doctor Strange and Guardians of the Galaxy) while moving towards the top, we see superheroes with powers that arise from technological transformations (such as Iron Man and Captain America).
   - We also see in Figure 16 that as we move towards the right on the x-axis, the movies with more emphasis on personal growth of the characters increases (like Hulk, Doctor Strange). However, if we move towards the left the focus is more on their interaction with other characters and the Universe (such as Captain America: Civil War and Iron Man 2).

- In Figure 18, we see that movies in the upper half, such as Planet Earth (2006), Black Mirror: White Christmas (2014), Akeelah and the Bee (2006), and Piper (2016) have more focus on cinematic and visual aspects while movies at the bottom such as Louis C.K.: Live at The Comedy Store (2015), Blue Is the Warmest Color (La vie d'Adèle) (2013) and Elite Squad lay more emphasis on dialogues.

3. **SVD++** :–

- From Figure 22, we can see that that all Harry Potter movies form a cluster that lie slightly to the right (along +X Axis) and all-Star Wars & Star Trek movies form a cluster that lie slightly to the left (along -X Axis). This is expected because such movies have dedicated fanbases who rate the movies almost consistently. . Then we plotted a few more fiction movies and among these, we found that movies with more fantasy/supernatural elements lie towards right (For example, Fantastic Beasts, Pirates of the Caribbean etc.). Also, movies with more futuristic and advanced technology elements often related to space, tend to lie on the left half (For example, Gravity, Equilibrium etc.),
- In Figure 22, we also find that, lighthearted and comedy movies were mostly in the top half of the plot. We expected that the bottom half of the plot would contain movies with serious themes (that have suspense elements).
- In Figure 24, we can see that Black Mirror and Louis C.K: Live at The Comedy Store are two very different types of movies and are far away from each other in the plot. Black Mirror is a science-fiction anthology series that explores the impact of technology on society, Louis C.K: Live at The Comedy Store is a stand-up comedy film that provides a humorous take on everyday life.

4. **General Analysis** :–

- Considering the most popular movies and looking at Figures 11, 17 and 23 we can see that only SVD++ algorithm groups the movies *The Lord of the Rings: The Return of the King* and *Kill Bill* together. Indeed they are similar too. Both films feature battle/action scenes and have themes related to revenge.
- In Musical Genre, comparing Figures 13, 25 and 19 we can see that Basic SVD and SVD+Bias groups them together even though they are different movies. But SVD++ captures this showing its superiority. *Chicago* is a musical drama, while *Corpse Bride* is an animated musical fantasy.
- In Documentary Genre, comparing Figures 14 and 20 we can see that in case of SVD+Bias the movies *Bowling for Columbine*, *Citizenfour*, *Enron: The Smartest Guys in the Room* and *Exit Through the Gift Shop* are clustered together. This can be because the model seems to capture similarities among them such as the fact they are popular documentaries that deal with controversial topics. They also use innovative methods to communicate their stories around the nuances of society.