Capstone project 1

Recommender systems

Recommender systems

- A group of machine learning algorithms, generally used to enhance the search quality by taking into account the interest and preferences of the user
- Used by major tech companies such as Amazon, Netflix, Youtube or social media like Facebook and Instagram

Data

- A kaggle data set containing over 45000 movies
- 26 million ratings from 270000 users
- Ratings on a scale of 1-5
- Including cast, crew, plot keywords, budget, revenue, posters, release dates, languages, TMDB vote counts …

Data wrangling

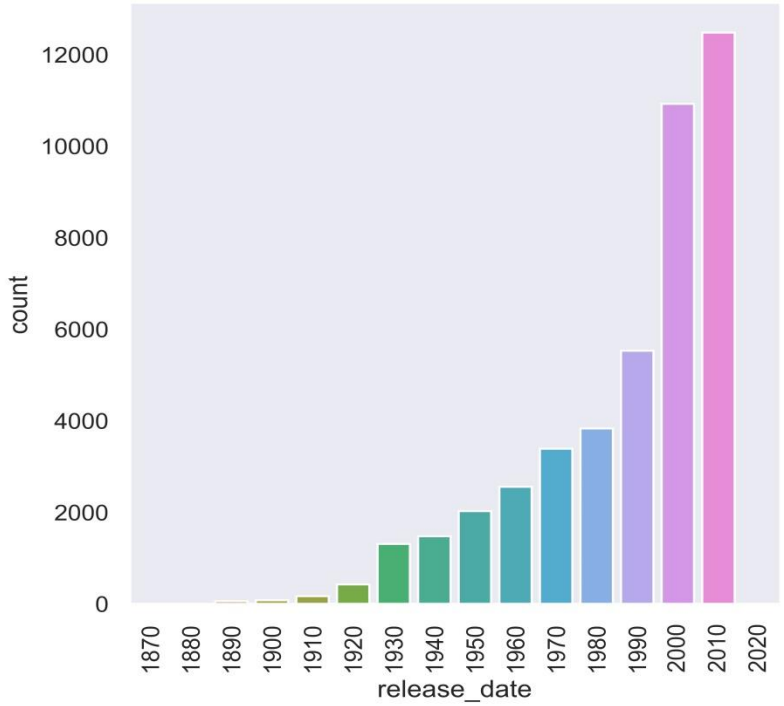
General procedure used to clean the data:

1. Check for the null values
2. Check for the duplicated values
3. Validating the correct data type of each column
4. Dropping the irrelevant columns
5. Changing the data type to categorical, when applicable
6. Turning the string-formatted date and time to datetime format
7. Setting the datetime column as the index
8. Sorting the index

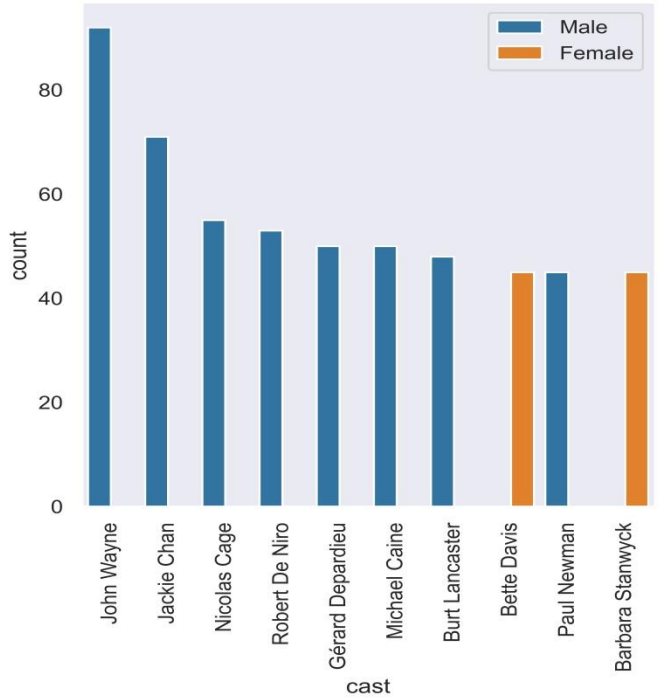
Exploratory data analysis

Visualization, an effective way to get familiarize with the data

Movies released per decade

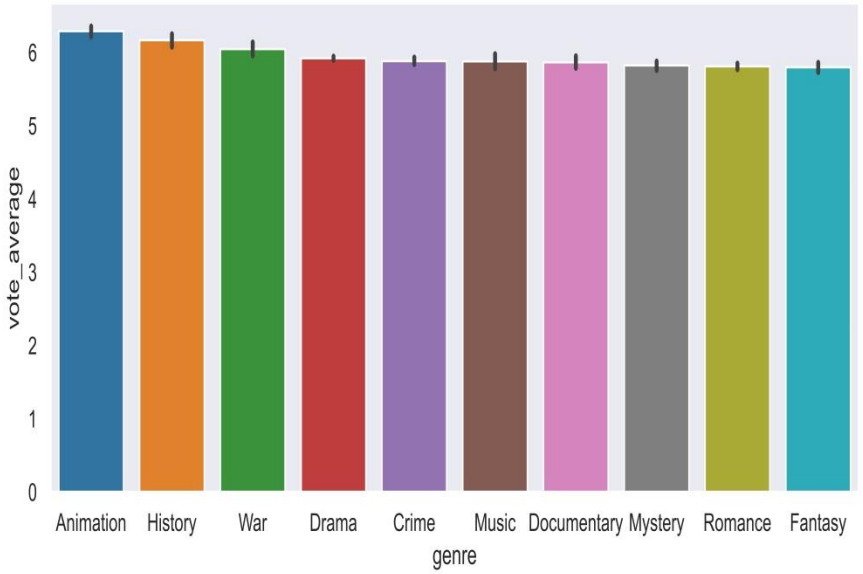


Top 10 leading actors by the number of movies they played

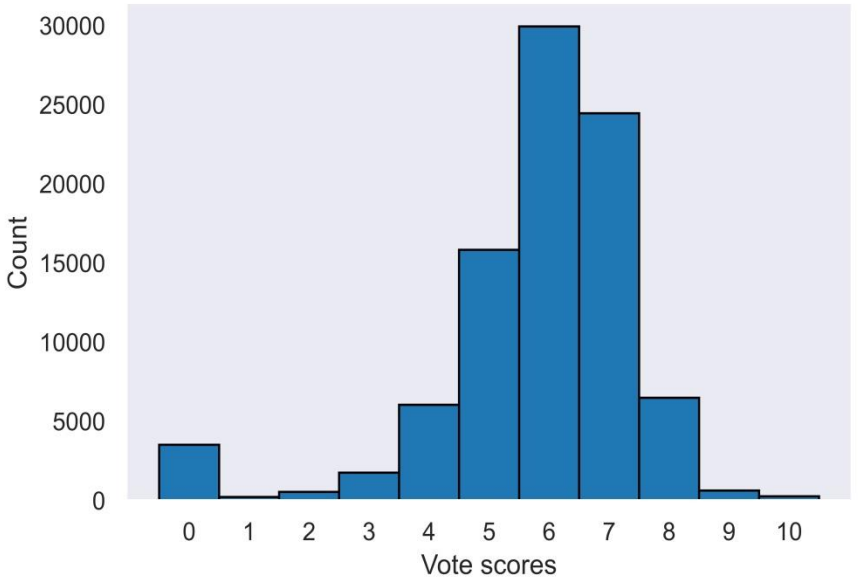


Exploratory data analysis

Top 10 genres based on the average ratings they received



A histogram of the vote scores people tend to give to the movies

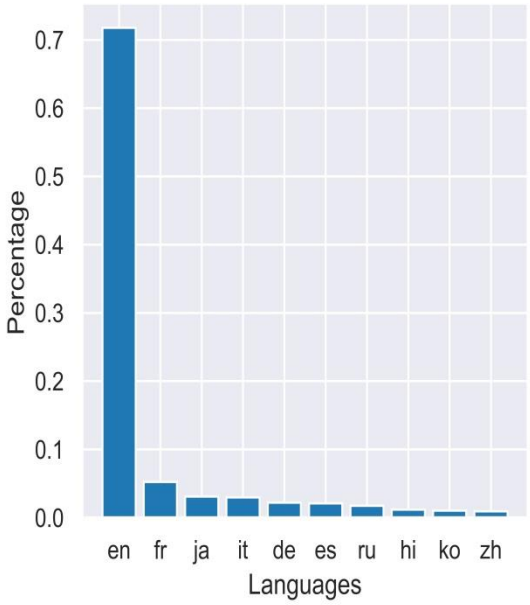


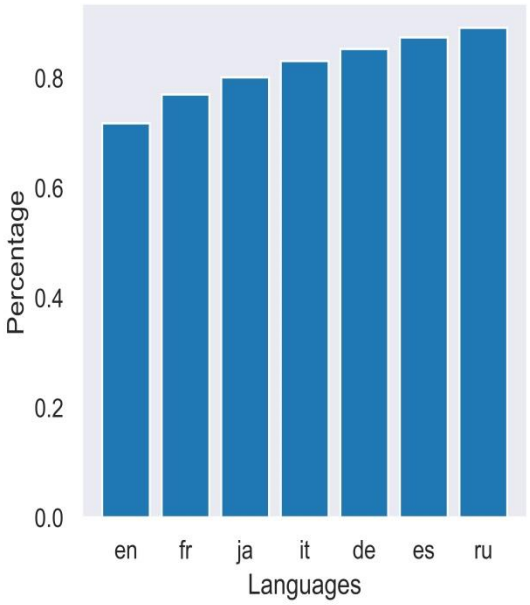
Exploratory data analysis

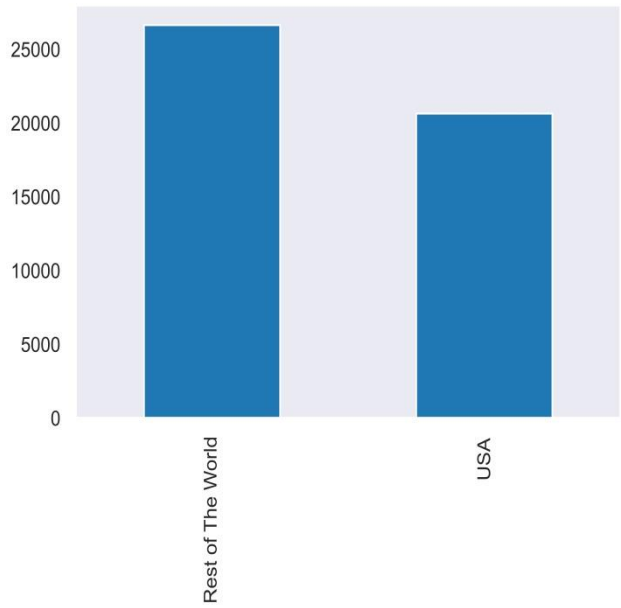
Percentages of top languages of the movie industry, by number of movies produced, in the left image

Seven languages are responsible for more than 90% of all movies produced

Movies produced in the USA vs. rest of the world

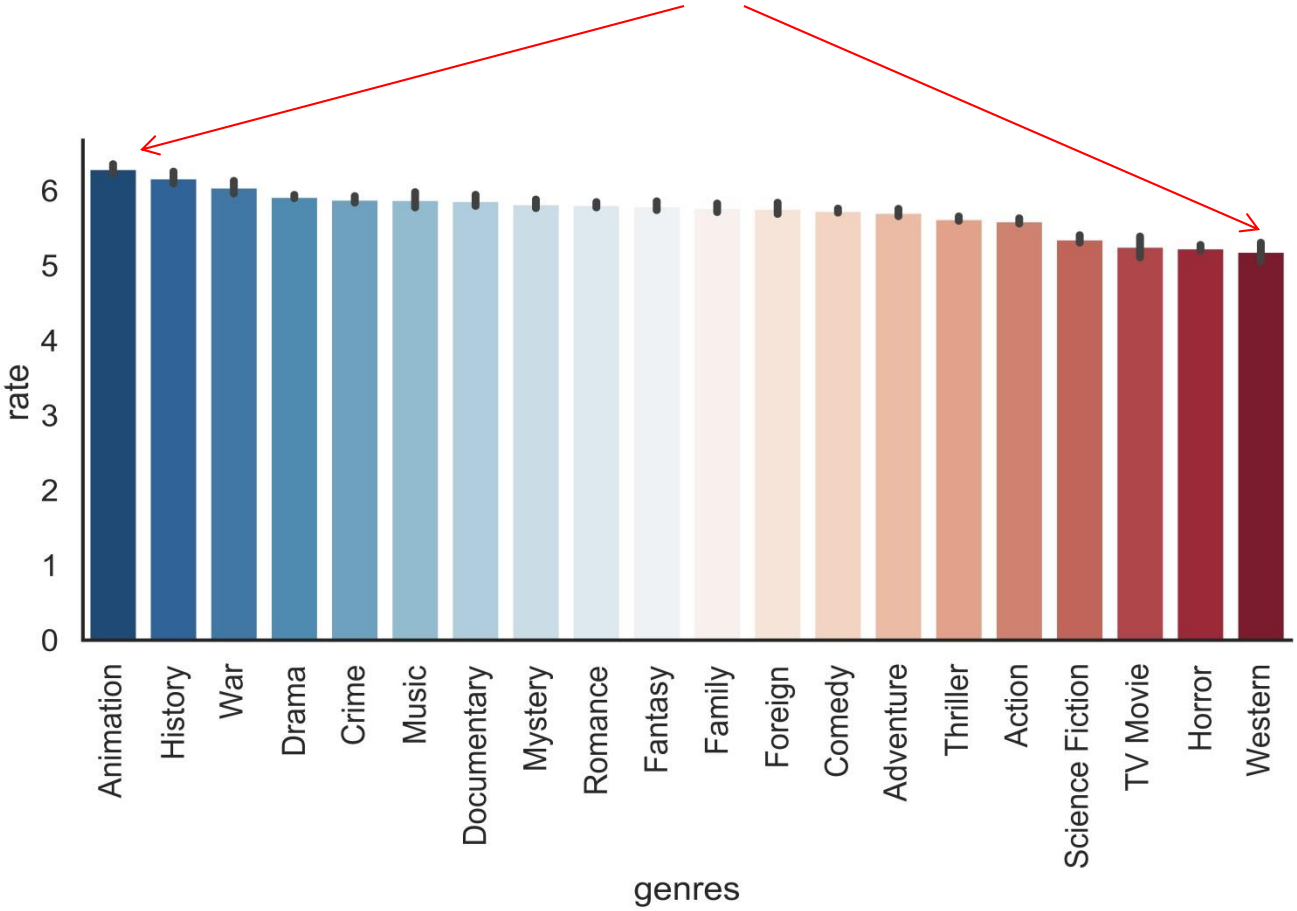






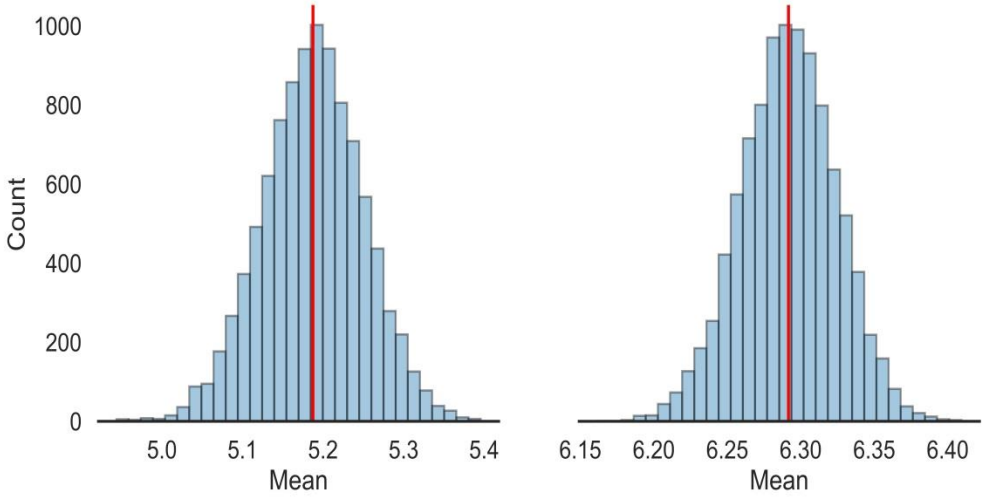
Two sample hypothesis test

Is the difference between the ratings of animation genre (best ratings) and the western genre (worst ratings) significant?

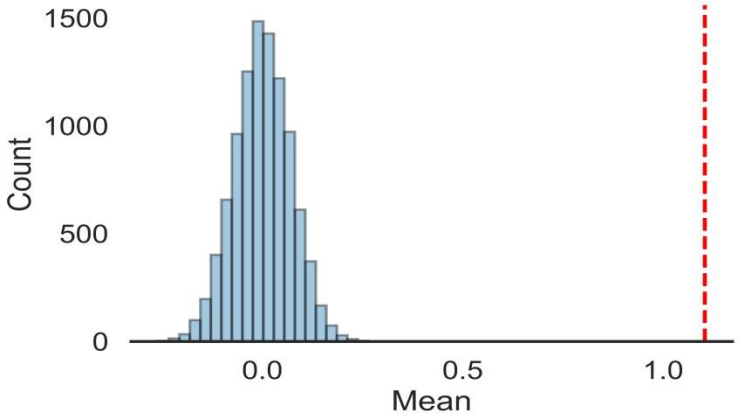


Two sample hypothesis test

Western movies means distribution on the left (mean around 5.19)
Animation movies means distribution on the right (mean around 6.29)

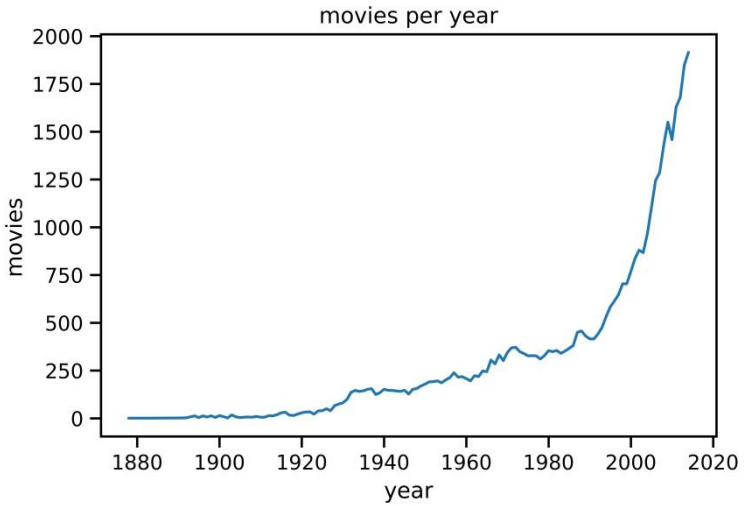


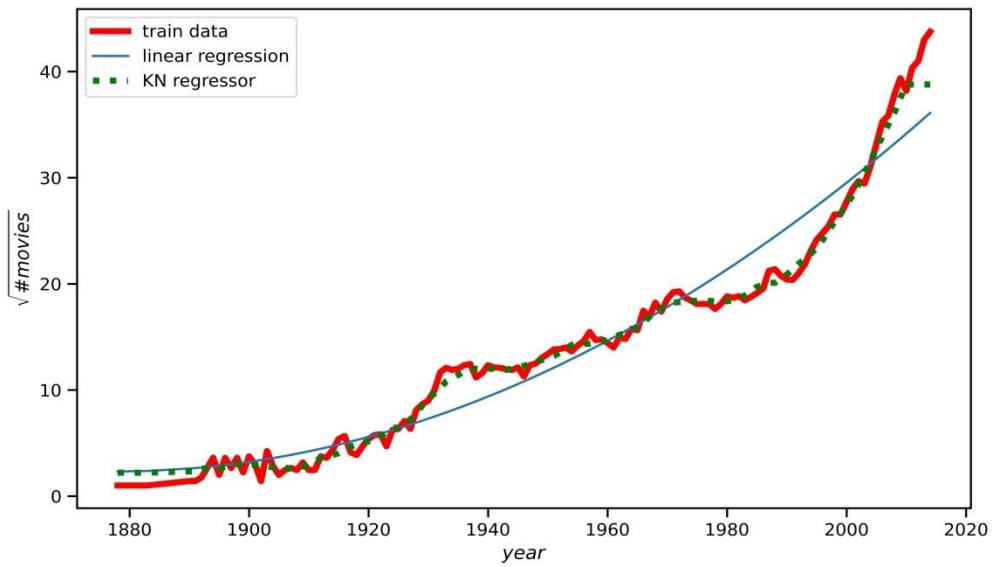
Distribution of mean differences between animations and western movies
The infinitesimally small p-value (red dashed line) tells us there is almost zero chance of observing the difference between the ratings, just by chance

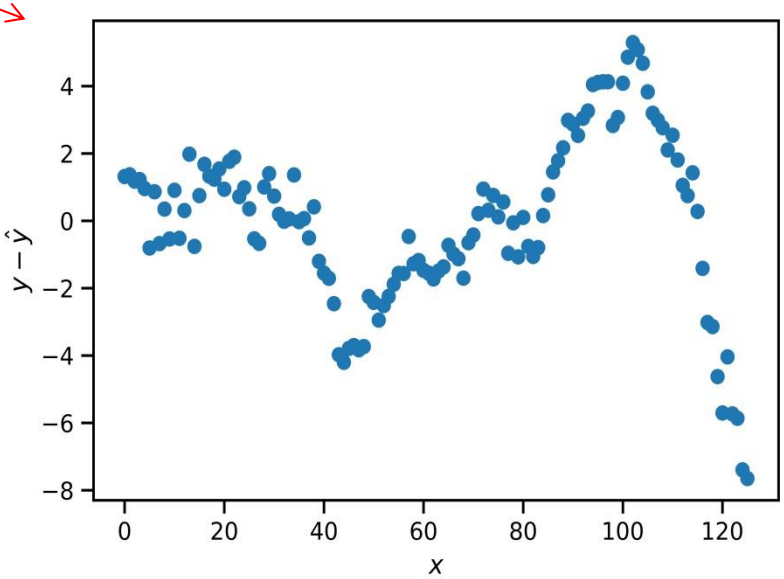


Regression analysis

Movies produced as a function of year
In order to make the relationship linear, sqrt transformation is applied on the y axis (image below)
Considering the non-linear behaviour of our data, as show in the residual plot, applying non-linear regression methods seems reasonable (methods like KNN-regressor as shown below)

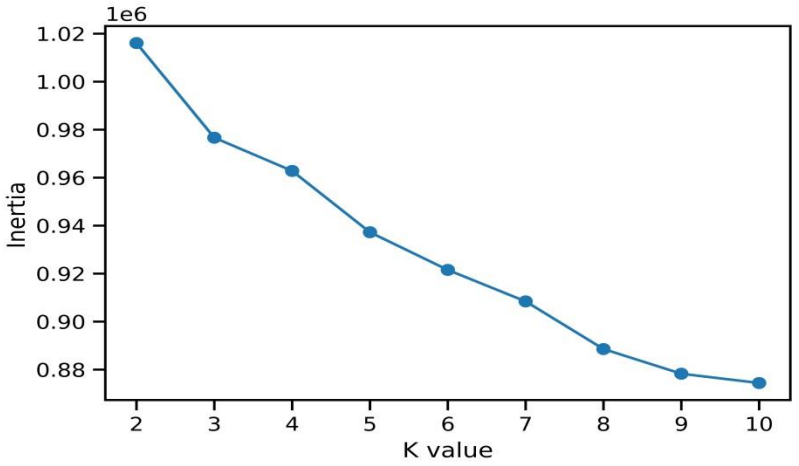




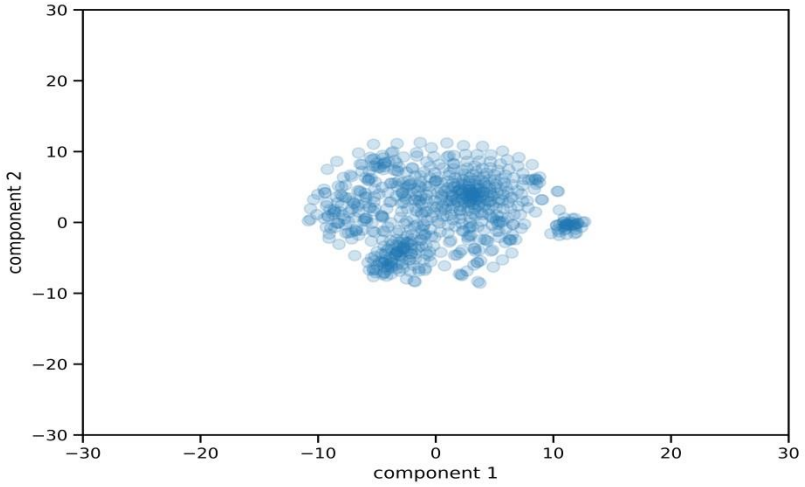


Cluster analysis

Clustering the users based on their ratings to each movie (671 users and 9000 movies)
PCA has been applied to the data to lower the dimensions from 9000 to 300 with 90% of variance preserved
The plot of cluster numbers vs inertia doesn't show a clear elbow point



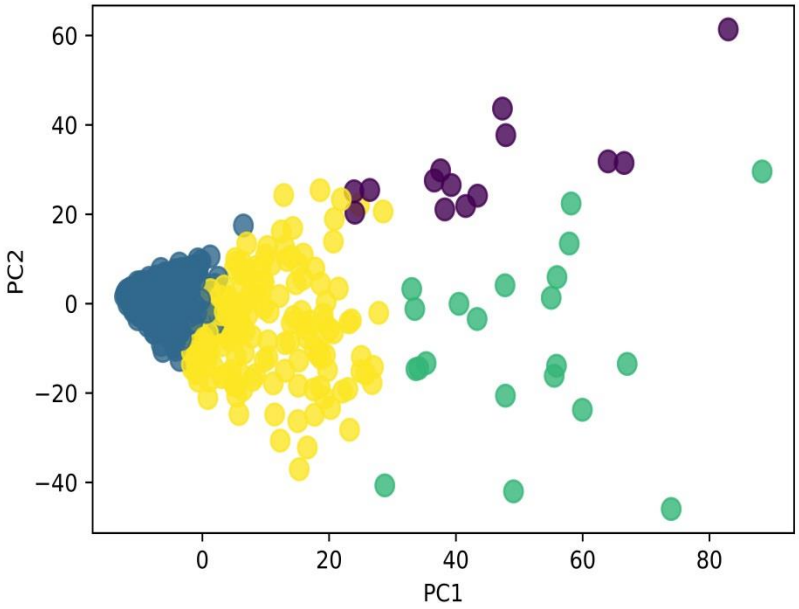
The t-SNE plot of the high-dimensional data in order to give a clear visualizations of the different clusters.
Still not a clear answer on the number of clusters...



Cluster analysis

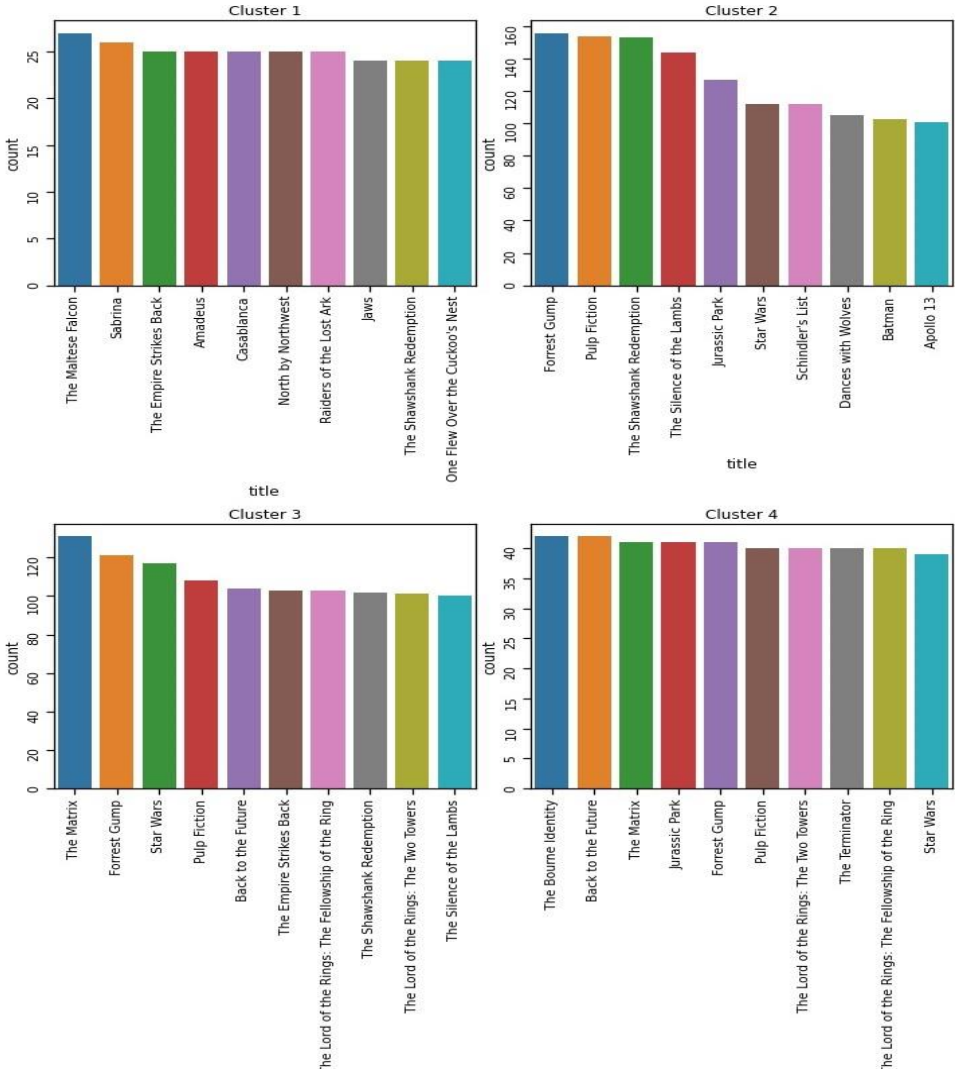
Another effective method to specify the number of clusters is the silhouette score of the data for different cluster numbers. In this case silhouette score for 4 clusters is higher than the rest, so we go with 4

In the end, I chose 4 clusters and used K-means clustering algorithm to label the data.
The labels are then used in visualization of the first two principal compoents of the date



Cluster analysis

4 clusters and top 10 viewed movies for each cluster is presented in the form of bar charts



Recommender systems

- Algorithms used are categorized as **Collaborative item-based Filtering**
- Surprise package of python is used to build the models
- The models are bulit based on the Singular Value Decomposition **SVD** and Non-negative Matrix Factorization **NMF**
- Root mean square error RMSE is the metric of choice to evaluate the performance of each algorithm

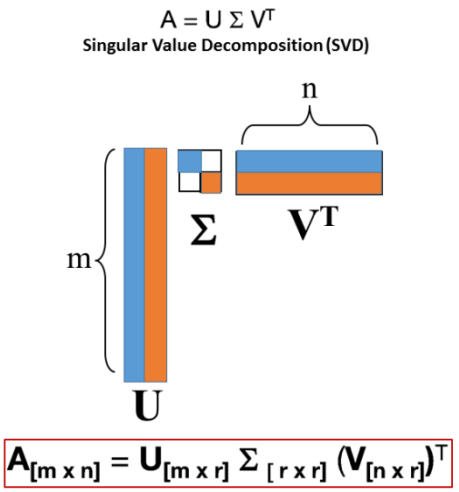
SVD

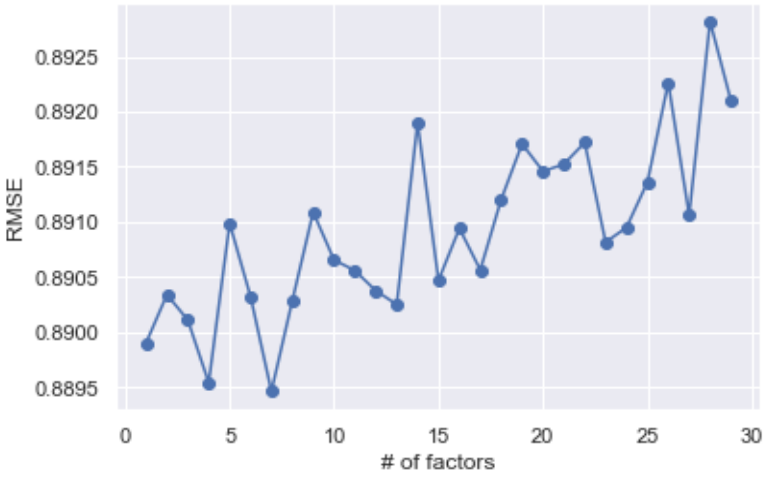
SVD expresses the input matrix as the product of 3 matrices

U represents the user-to-concept similarity

V represents the movie-to-concept similarity

The diagonal matrix represents the strength of each concept



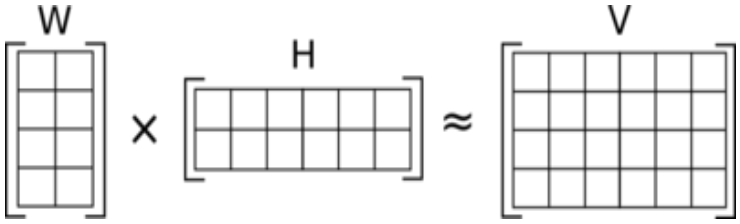


One hyper parameter to choose is the number of concepts (factors) we want for the model

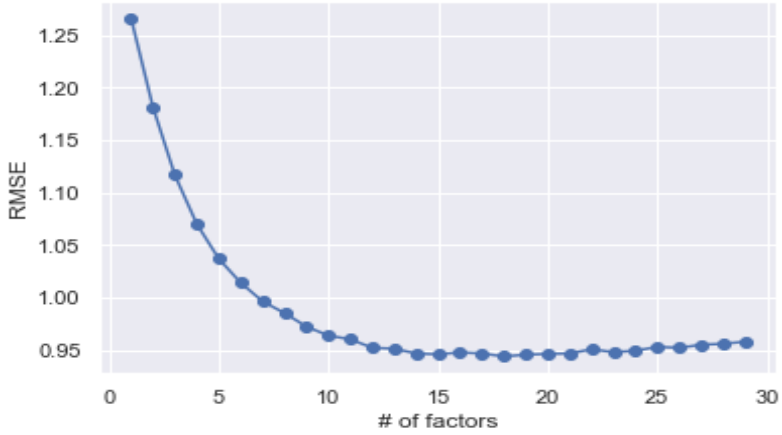
Plot on the left shows the lowest RMSE for 7 factors, which yield the **RMSE** of **0.89**

NMF

NMF decomposes the **non-negative** input matrix **V** of m*n dimensions into 2 matrices of **W** and **H** with dimensions with m*r and r*n respectively, where r < min(m, n)



Using cross validation method to find the best number of components which gives the lowest **RMSE** possible (**0.95**). Based on the plot below, 17 gives a suitable number for the components value



SVD vs. NMF

SVD gave a slightly lower RMSE (0.89 compared to 0,95)
To compare the results of each method, the first 10 recommendations of each algorithm is given below for one of the users plus the highest rated movies by the same user