# Assignment 1C

## Problem 1

### Dataset

The data for this problem comes from the MovieLens user rating dataset. The relevant files being used are rating.csv which contain user ratings from movies and movies.csv which relate is used to refence movie names. This dataset contains 100836 ratings of 9742 movies from 610 users.

### Dataset Pre-processing

As the objective of the problem is to find movie recommendations for users the data in its current form will not be able to produce this. As users are the object that will be clustered with the information used to cluster them will be their ratings of varies movies. The data will be transformed to create a table where the row of the dataset will be each individual user and the columns will represent a specific movie rating. In figure 1 below a visual representation of the data can be seen.

It should be noted that a lot of the data is unrated or NaN as most users will not have rated every movie.

It was decided to use all movies and users in the model as it seemed unlikely that every user will have watched a high number of the 9742 movies available causing a sparse dataset to be the most unbiased dataset able to be produced.

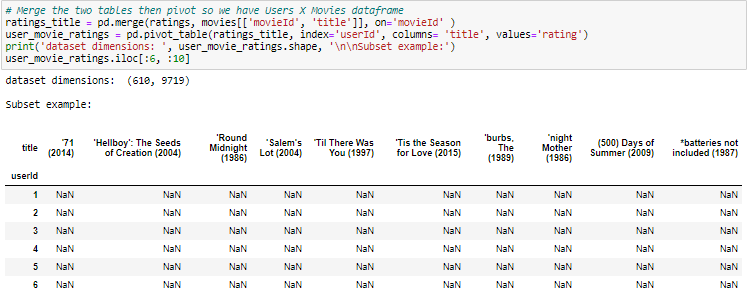


Figure : Data after Pre-processing

### Model

To cluster this dataset the k-means model/algorithm was chosen to be used. This model was chosen because due to the fact the there are so many dimensions (movies) to this data it is hard to visualise, and it is therefore hard for me to know how this data should be classified due to this the simplest option was chosen being k-means. This method was also chosen to limit the number of movies recommendation per cluster if a method like GMM was chosen then multiple movies could belong to many clusters if there is a lot of overlap movie recommendation could simply become the most highly rated and watched movies. k-means algorithms uses hard assignment and avoids this issue.

The algorithm uses all default values except for k or number of clusters. To find an optimal number of clusters for this problem a grid search like method was used creating 14 k-means models increasing the increment of cluster by 1 then calculating the Silhouette Score to determine how well users are clustered. A Silhouette Score measure of how similar an object is to its own cluster compared to other clusters. A higher value indicates that a user is more accurately depicted in a cluster on average. As seen in figure 2 below the Silhouette Score (Distortion) falls after 4 clusters however after running this grid search method a few times the location of this fall varies around 4 to 6 and therefore a k of 5 was chosen.

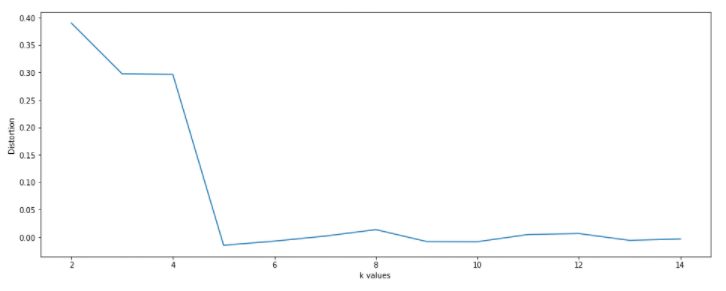


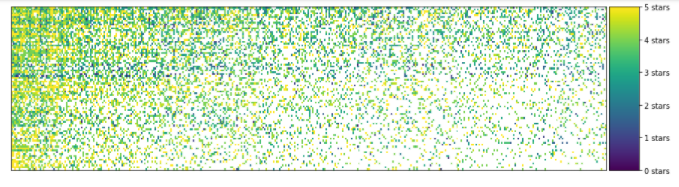
Figure : silhouette score of n clusters

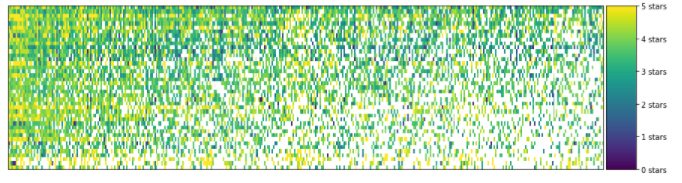
### How recommendations are formed

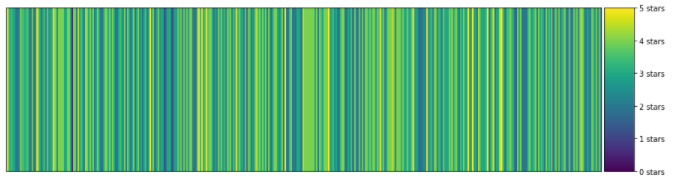
Using the model stated previously the model can generate predictions for users by locating the cluster of which and individual user belongs to then finding the highest rated movies that there cluster has rated that they themselves have not rated.

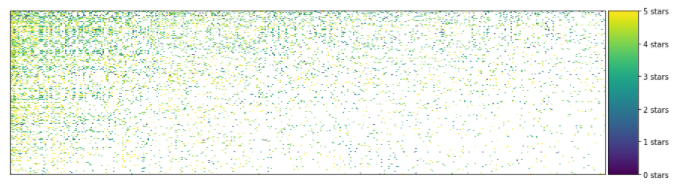
### Results of the clustering

In figure 3 below it shows a visual of the resulting cluster that the model and dataset produce. Each of the plots in this figure shows the rating of movies (columns) in colour (5-star rating) by users (rows). Note that white means no rating given. There are cluster with both unproportionally huge and small group showing the groups are not sorted very evenly. To determine how well this model fits the problem an analysis of individuals will be given below.









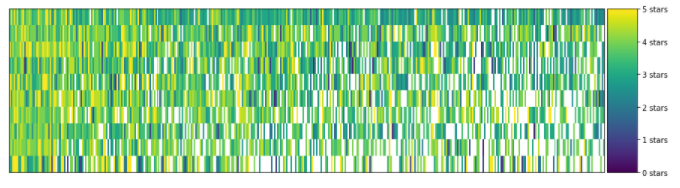


Figure : Resulting Clusters of Model.

### Results/User Recommendations

**User 4**

In the below figure or figure 4 is shows users 4 most watched movies genres that was rated above 3. This user enjoys mostly drama, romance and comedy type movies.

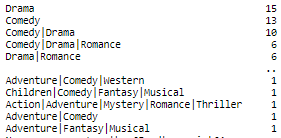


Figure : User 4's count of genres rated above 3.

In Figure 5 below it shows the top 5 recommendations the model has produced, and as seen Drama or Comedy is in the movie Genres. This show for User 4 the cluster can group similar movies to what user 4 has shown a liking towards.

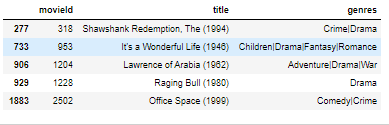


Figure : User 4's Top 5 recommendations

**User 42**

In the below figure or figure 6 it shows users 42 most watched movies genres that was rated above 3. This user enjoys mostly Drama and Comedy type movies and seems very similar to User 4 however it should be noted that this user was sorted into a different cluster. This may be because this user has shown a liking to other genres where User 4 seemed to be exclusively into comedy drama and Romance.

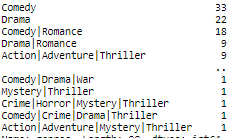


Figure : User 42's count of genres rated above 3.

In Figure 7 below it shows the top 5 recommendations the model has produced, and as seen the movies in the drama genre does show however unlike User 4 this cluster seems to have more action and adventure movies included in the selection.



Figure : User 42's Top 5 recommendations

**User 314**

In the below figure or figure 4 is shows users 314 most watched movies genres that was rated above 3. This user enjoys mostly drama, comedy, and romance type movies. This preference for these types of movies genres to the point where they are the only genres that the user has rated above 5 more then once is very similar to User 4 and the model was able to recognise this as they are in the same cluster.

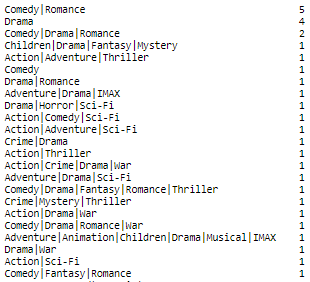


Figure : User 314's count of genres rated above 3.

In Figure 9 below it shows the top 5 recommendations the model has produced, and as seen movies in the drama and romance genres are frequently suggested for this user showing a reasonable recommendation list however this is one suggestion in this list where the User has shown no interest in its genres. This was likely because movies in clusters are sorted by most highly rated and will therefore have a bias for movies with high average rating in their cluster.

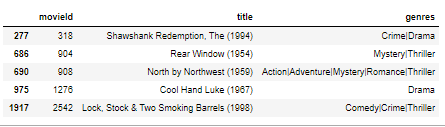


Figure : User 314's Top 5 recommendations

**Overall thoughts**

Overall, I think the recommendation where reasonable as the model was able to recognise Users 4 and 314 with very similar tastes as they both clearly did not like much outside of 3 genres, they would only highly rate. User 34 also like those same genres but showed some interest in other genres and was able to be sorted into a group with a more diverse movie selection. As the model was able to easily categized similar users the rating of each group would allow the model to find recommendation of user with similar preference. However there where some issues as there was one group with one user assigned to it. This is an issue as this user would not be able to find recommendation due to this issue. To fix this in the future the number of k would need to be selected to find a number that would have more then 1 amount of people per cluster.

### 