|  |
| --- |
| WPI LOGO.png |
| [**Movie Analysis with MovieLens Data** ] |
| DS501 Case Study 1 : Analyzing Data from MovieLens  10/26/2017 |
|  |
|  |
|  |

**Team 6 Haowen Zhu, Dekun Geng, Yin Hang, Yixin Luo, Weiqing Li**

### Introduction:

### Help a Movie Company with Analyzing the Movie Data

In this report, we used data from MovieLens to do some analysis to help a movie company in choosing movie script and doing precision marketing. MovieLens (<https://movielens.org)> is a a movie recommendation service run by GroupLens, a research lab at the University of Minnesota. In the business report, we use the Full MovieLens Latest Dataset (<https://grouplens.org/datasets/movielens/latest/)> and an older dataset, MovieLens 1M Dataset (<https://grouplens.org/datasets/movielens/1m/>).

For a movie company, let’s take Netflix as an example, it will need to choose a good script to make it to a movie or buy a promising movie. It may be interest if we could tell what kind of movie is popular and have good ratings or which kind of movie is popular but people are still waiting for a good one.

For this question, we use the latest dataset. It contains more data and the conjecture from the newest data will be more meaningful.

When it has a certain movie to promote, it will be very happy if we could find targeted customers for this movie then do some precision marketing. As a movie company as Netflix, it will need a recommender service for its users. Thus we also do a recommender system.

In this part, we use the older one to have the users’ demographic to do the analysis.

### Dataset Overview

The latest dataset describes 5-star rating and free-text tagging activity from MovieLens. It contains 26,000,000 ratings and 750,000 tag applications applied to 45,000 movies by 270,000 users. Includes tag genome data with 12 million relevance scores across 1,100 tags. These data is created between January 09, 1995 and August 2017. Users were selected at random for inclusion, but all of them had at least rated 1 movies, which means we can filter out a lot of fake ratings done by fake reviewer. But it doesn’t contain demographic information of each user, which means we can only have the user’s ID but no more information in this dataset.

Thus, we use the older data set to do the second part, precision marketing. The MovieLens 1M Dataset contains 1 million ratings from 6000 users on 4000 movies. Also we have simple demographic info for the users (age, gender, occupation, zip).

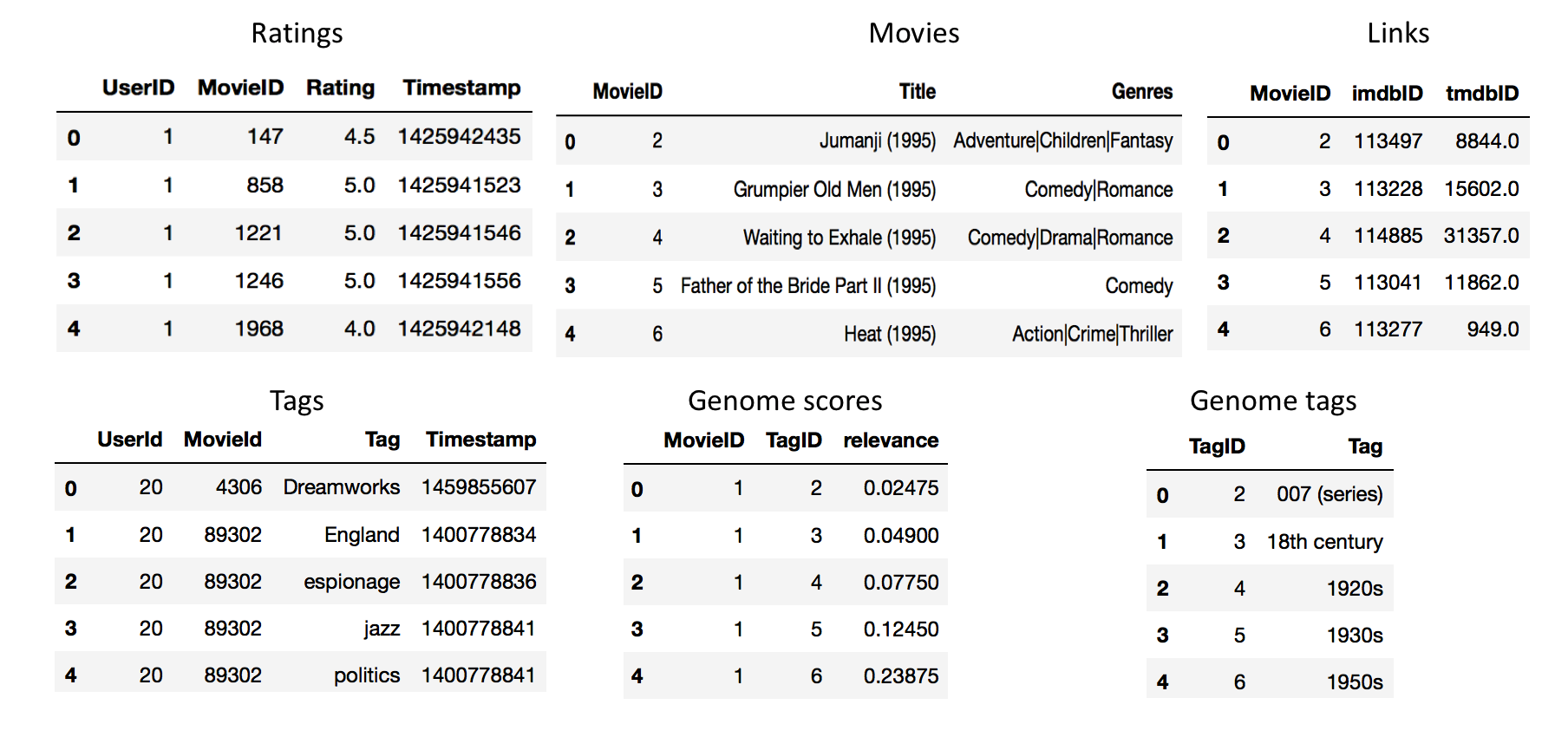
### Analyze the Data

Data cleaning

We have to first do the data cleaning to build our dataframe and exclude some abnormal or missing data. We use pandas package to do the data analysis.

*Latest Dataset*

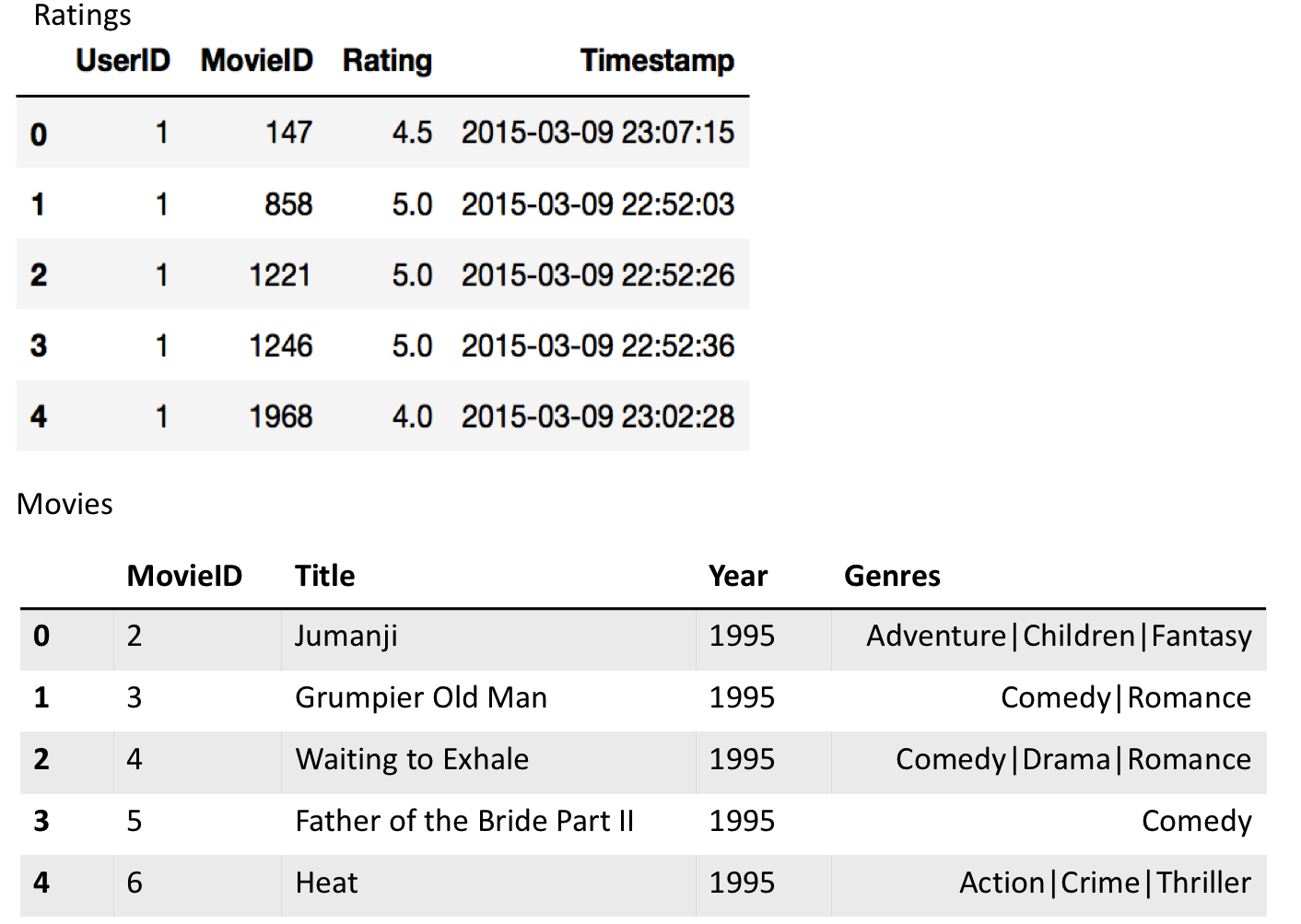
After loading the data, we have six dataframe from 6 different csv.



From the overview of the data, we will:

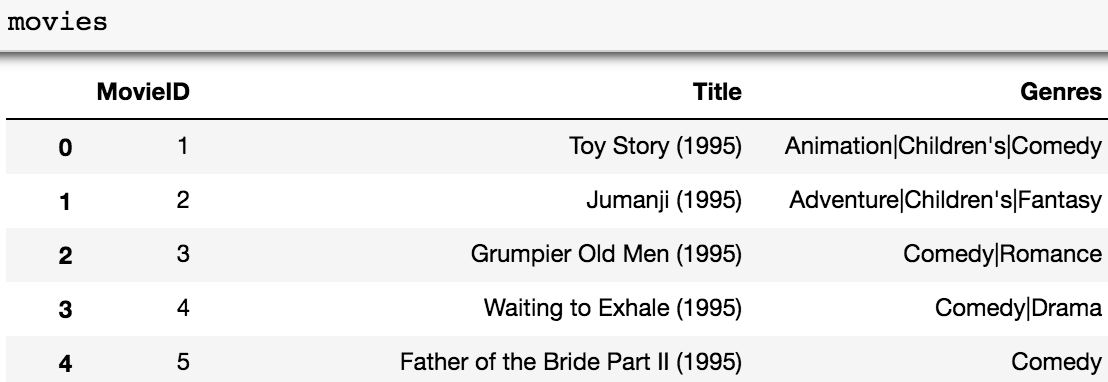
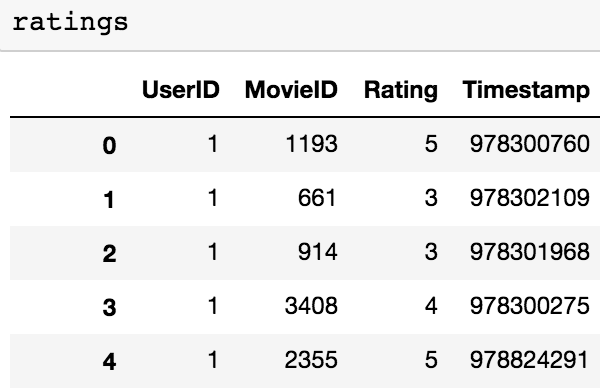
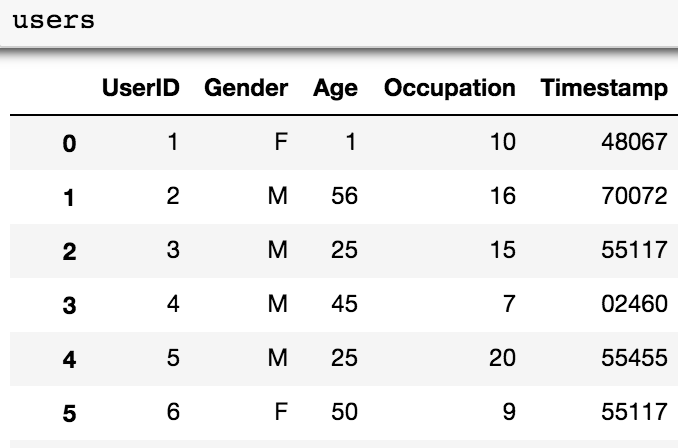
* Change the Timestamp in Ratings and Tags table to a readable date-time format.
* Split the Title from name(year) to two columns Title, Year.
* For the Genres, we will save to it later to generate a another table for the multivalue of the genres.
* Clean the missing data.

When 1 and 2 are done, the table is like



Here we found 405 missing-value rows in Movies. As we have 45842 movies, we will delete the 405 rows. For other tables, there is no missing value.

*Older* *Datase*



Data Analysis

*Latest* *Dataset*

1. How many movies are made every year?

In this part, we do the groupby of movies on ‘Year’ and count the number of movies in each year. As we could see the table of Year and Movie Count, a line plot will be more clear to see the popularity of movies changing with time.

1. What kind of movie was the most popular year by year?

In this part, we have to split the data and generate a new dataframe contains genres of the movie and the year of the movie.

Then, we may get interest about the popularity of different genres changing with year. We do the groupby on the new dataframe to make it a multi-index dataframe. By extracting a certain kind of genres, e.g. Action, we could have the count of this kind of movie in different year. Thus, we could see the popularity of a certain genre movie changing with time.

### What kind of movie was the best year by year?

### This part is similar with the second one as we just change the count of movies to the average rating of a genre.

### What kind of movie is people get interest in year by year?

In this part, we will see if people's attention of different genre changes with time. people's attention is indicated by the number of ratings. Also, we will see that if there taste which is indicated by mean ratings will change.

We generate a new dataframe named gen\_ratings containing genres, MovieID, rating and the rate time. The rate time is in year and get from the Timestamp. Then we do the groupby as in the second part.

1. What kind of movie is popular and rating high?

In this part, we use the gen\_ratings dataframe as well. But this time, we do a new groupby on Genres and MovieID to see what movie are popular and favored by people. Also, we will see the average rating and total rating numbers of different genres by do the aggregation.

1. What word will people use to describe a certain genre?

Let's assume we get a movie screenplay of a certain genre, we would like to know whether the script is qualified to succeed. Then, knowing what people expected about this genre will help us a lot.

Here we use the wordcloud package to draw the word cloud. We use the Tags table here.

We draw two kind of the word cloud. For the first one, we only calculate the frequencies of the tags. We generate the new frame gen\_tags containing movieID, tags, tiltle, year. Then, we extract a certain genre to get its tags.

For the second one, we include the relevance from the genome scores and genome tags to calculate the adjusted frequencies of the tags. We first inner join the genomescores an genometags to a new frame named genome. Then we inner join it with gen\_tags on MovieID and Tag. After extracting a certain genre and groupby on tag with sum the relevance, we got the tags and its adjusted frequencies of a certain genre.

*Older Dataset*

1. Explore Gender Preference for different film genres:

Just use those genres with sufficient views, let's just get those genres which have at least 250 reviews. See which genre men and women prefer. Explore the most diverse genres by using variable ‘diff’. measure the differential is through the spread, otherwise known as the standard deviation.

1. Explore ratings among different age groups:

Calculate the averaged rating genre for different age groups.

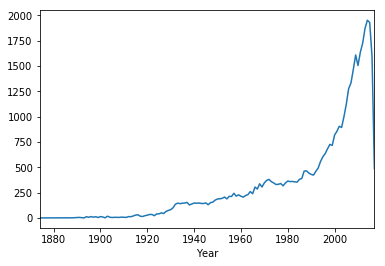
1. Recommender System

Outcomes and Conjectures from the Data

Latest Dataset

1. How many movies are made every year?

Here the line plot showing the number of the movies changing with year is in below.



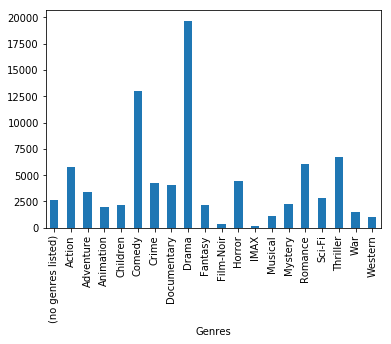
We can see there is an exponential growth of the movie number. However this could be caused by our dataset only choosing the recent review data. If there is really an exponential growth, it may attribute to the growth of people's interest on movies and the explode of technology which causes movie producing easier and easier.

There is a slight drop in 2010. The sudden drop in 2017 is because the data set is updated on 8/2017.

But we can make sure one thing, that the movie market is quite large today. So it is very likely to win a huge profit if we just do the right movie.

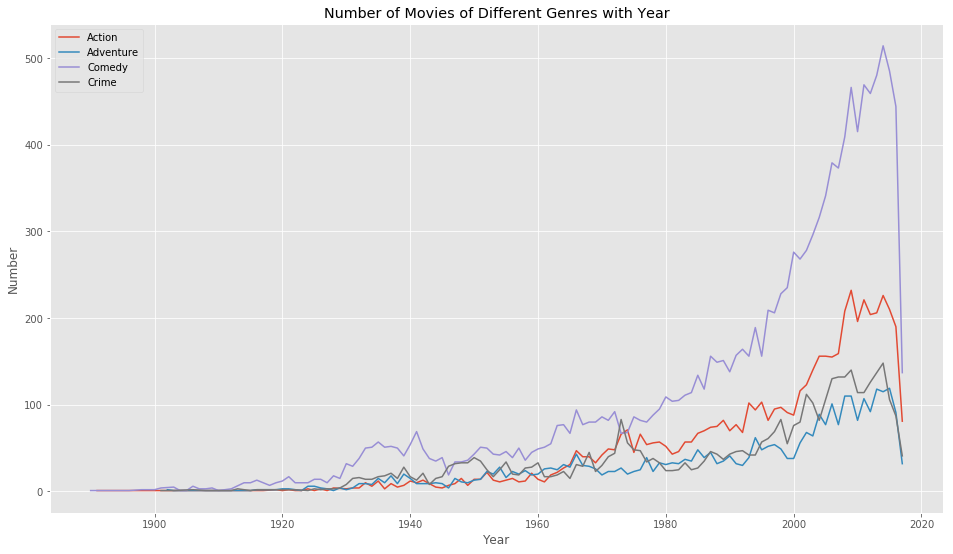
1. What kind of movie was the most popular year by year?

From the new dataframe, we could first get the bar plot of numbers of movie in different genres.



It is clear that Drama and Comedy are the genres which are most popular, followed by Thriller, Romance, Action and Horror. But War, Western, Musical is not so popular. This may because that they all need a lot of effort and money to make it. Just take a think of the war scene.

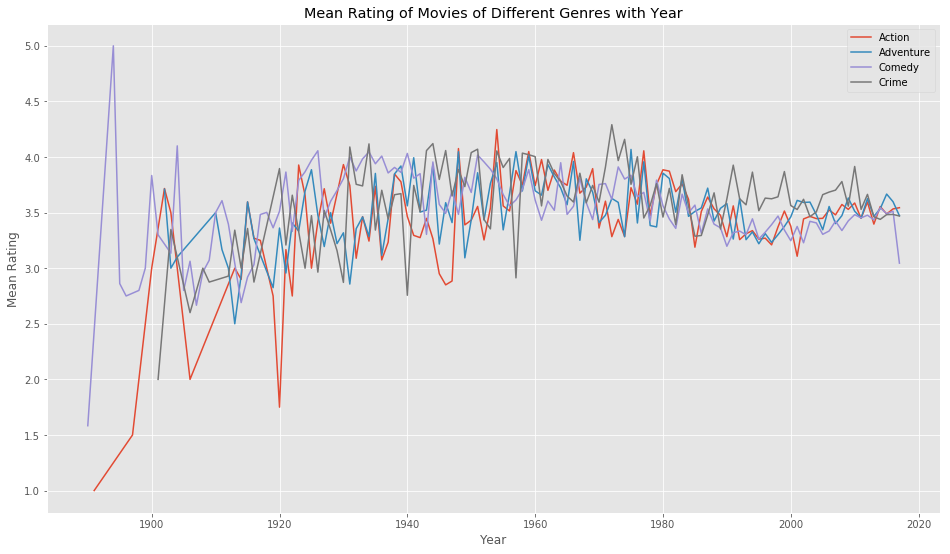
When it comes to a certain kind of movie, we could get the line plot of its count changing with year. For readability, we only show four genres.



The changing trend of this four genres is similar. They all get a rise in late 1900s and early 2000s. But the comedy has an exponential growth in this time. However, the other three genres grows slowly and reach a platform after 2010.

### What kind of movie was the best year by year?

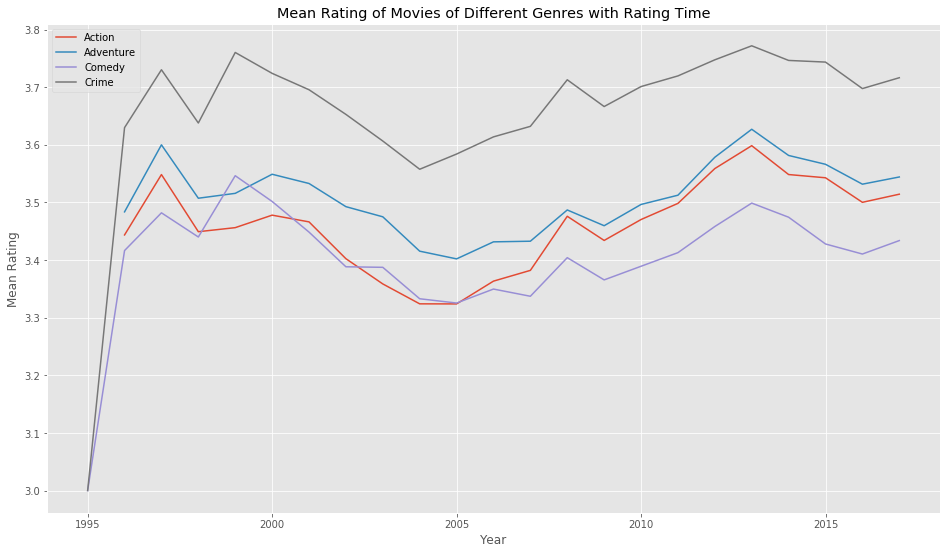
Similarly, we get this plot. For readability it will just contains four genres.

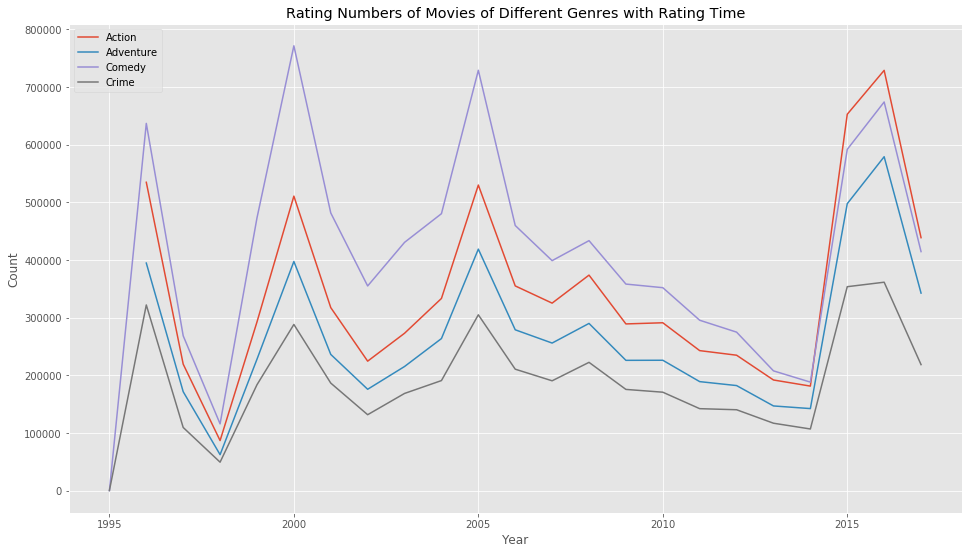


The mean rating doesn’t change too much for different year. But we could see a little down in the 2000.

1. What kind of movie is people get interest in year by year?

From the gen\_ratings dataframe, we get these two plot.

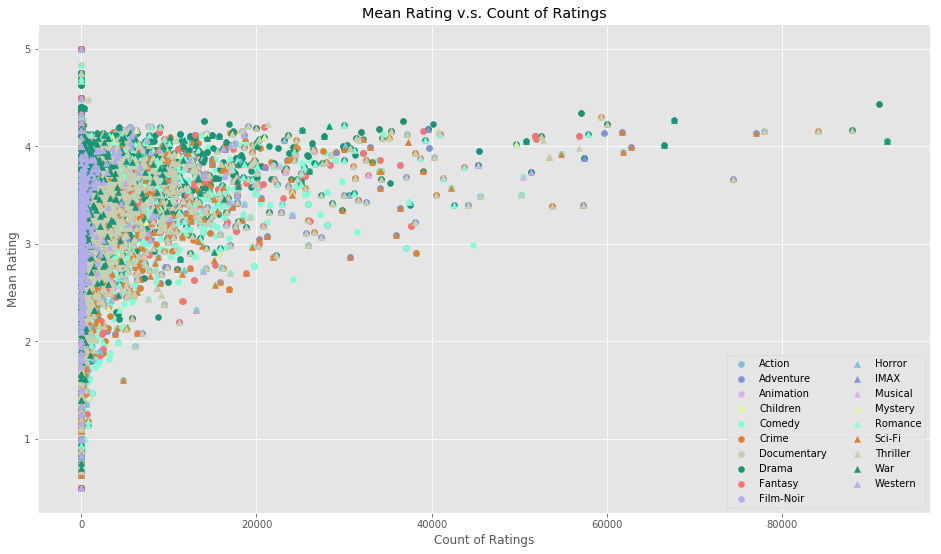




It seems people’s interest and taste doesn’t have a special trend when time goes. But there is some wicked things. Their interest have several peaks. In 2000, 2005, 2015, people review a lot of action, adventure, comedy and crime movies. However in 2015, people tend to give relative low ratings on these movies.

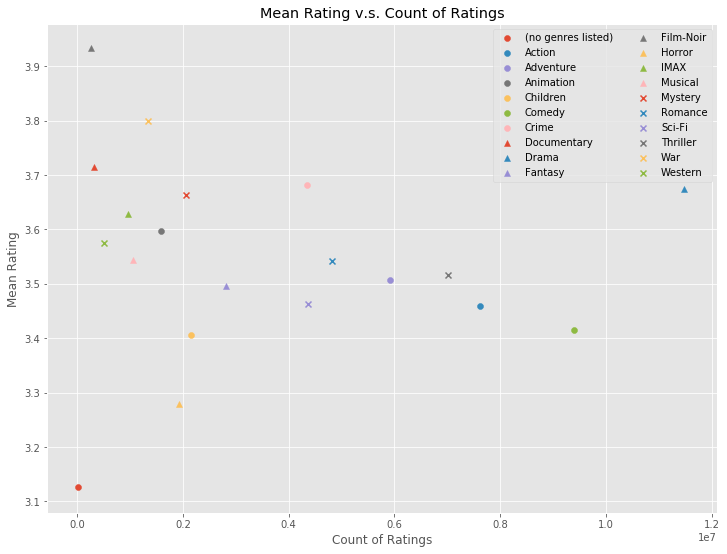
1. What kind of movie is popular and rating high?

First, we have this plot of the movies with rating number and average rating.



However, this plot is hard to see anything. We could only tell that some dramas, thrillers are both popular and good. If one want to see it close to a certain part of the plot, e.g. movies with high rating and low rating numbers, he could modify the x and y axis.

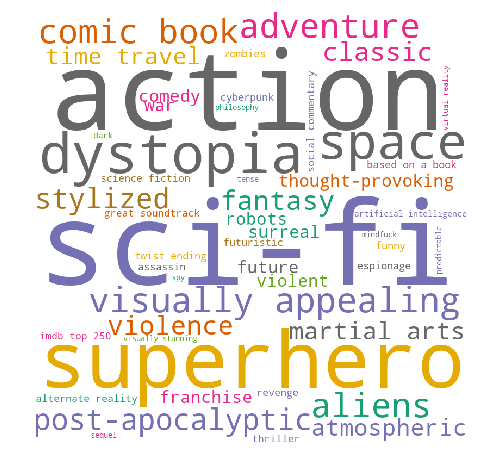
Thus we have this plot with average rating and rating numbers.



From it we could see there is several genres with low count of ratings and low rating, e.g. Horror. However, there is also some genres with high count and high rating, e.g. drama. Here we could see Comedy movies tend to get a lot of ratings but relatively low ratings. Action movies, etc. are also in this group. This means that people are familiar and willing to see comedies, but there is little movie satisfying them. Thus, there is enough room for us to make a succeed comedies.

1. What word will people use to describe a certain genre?

The first and second word cloud of Action is like this (left and right).

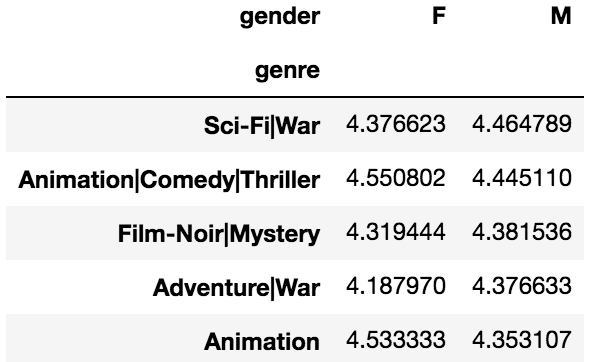
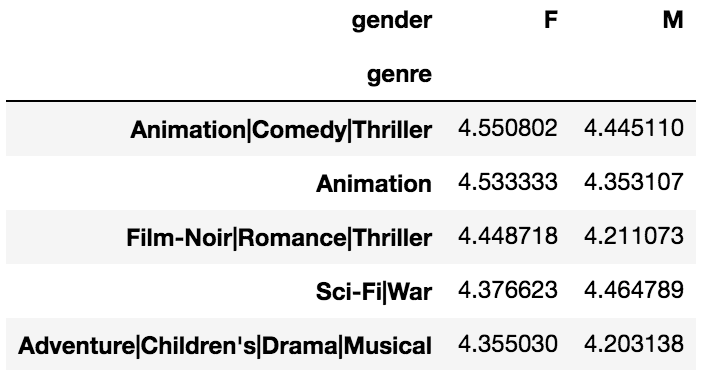
There is some difference between them. In the first one, we still can see some name of character or actor, e.g. James Bond and Brad Pitt. But in the second one, these names disappears and the tags in it are now more meaningful. We could see that people like superhero and sci-fi in the action movies.

Also we will give some more word cloud with adjusted frequencies.



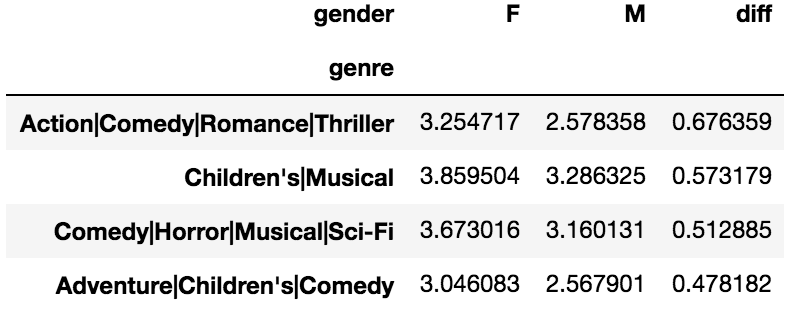
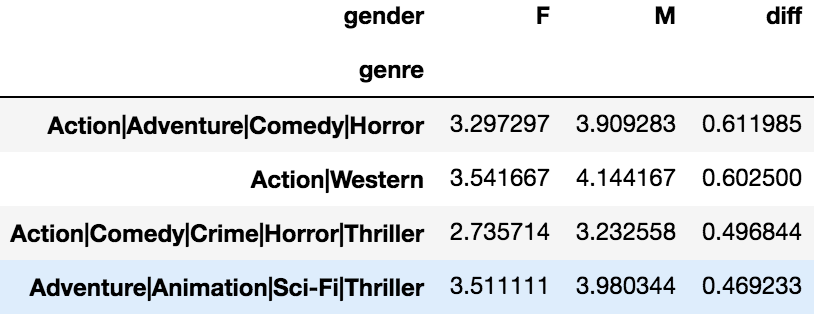
Older Dataset

1. See which genre men and women prefer

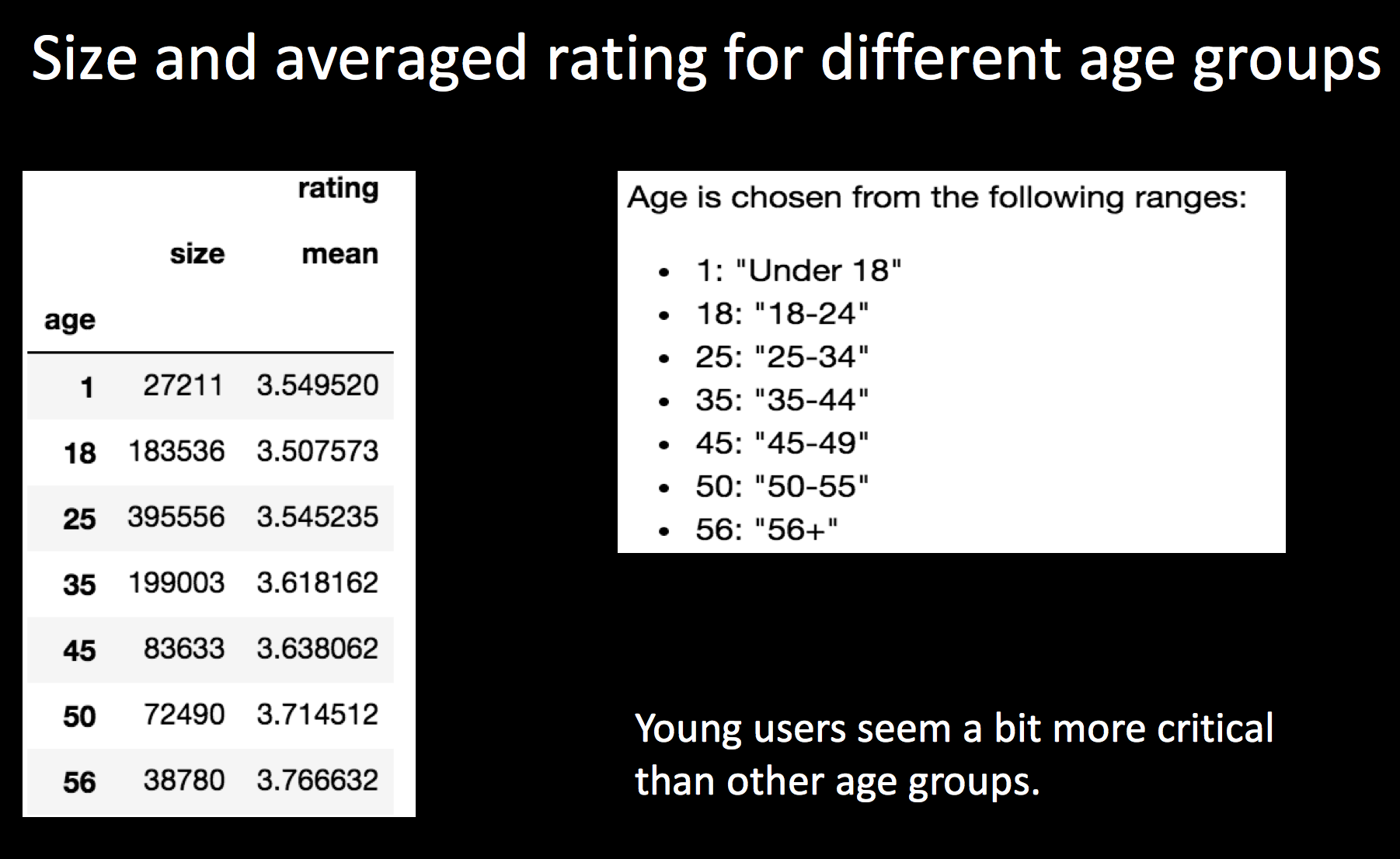
 

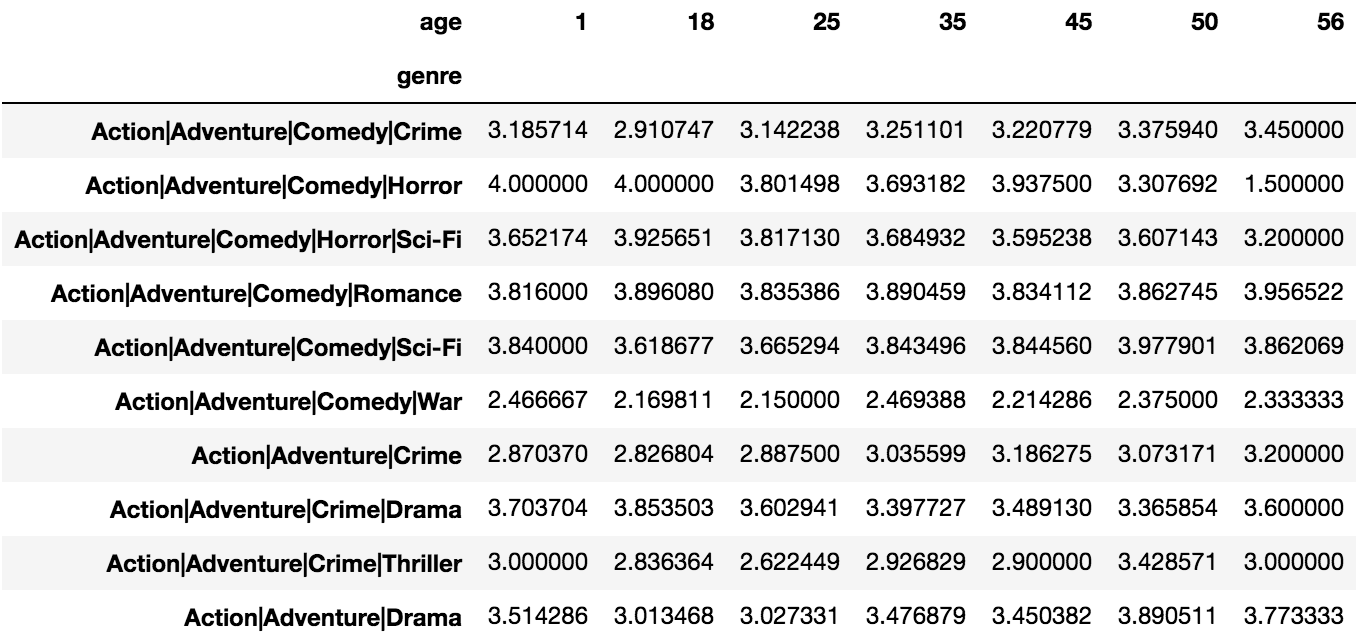
1. From this refined data, we can reorder and display the top rated genres by females and male. The top female averaged rated genre goes to 'Animation|Comedy|Thriller', with 'Animation' coming in second.
2. 'Sci-Fi|War', 'Animation|Comedy|Thriller', 'Animation' are all listed in TOP 5 averaged rated genres of both female and male. So, users enjoying a movie would be more inclined to rate a movie, thus this subset of movies with 250+ ratings is slightly more biased than the original total data.

2. Explore the most diverse genres:

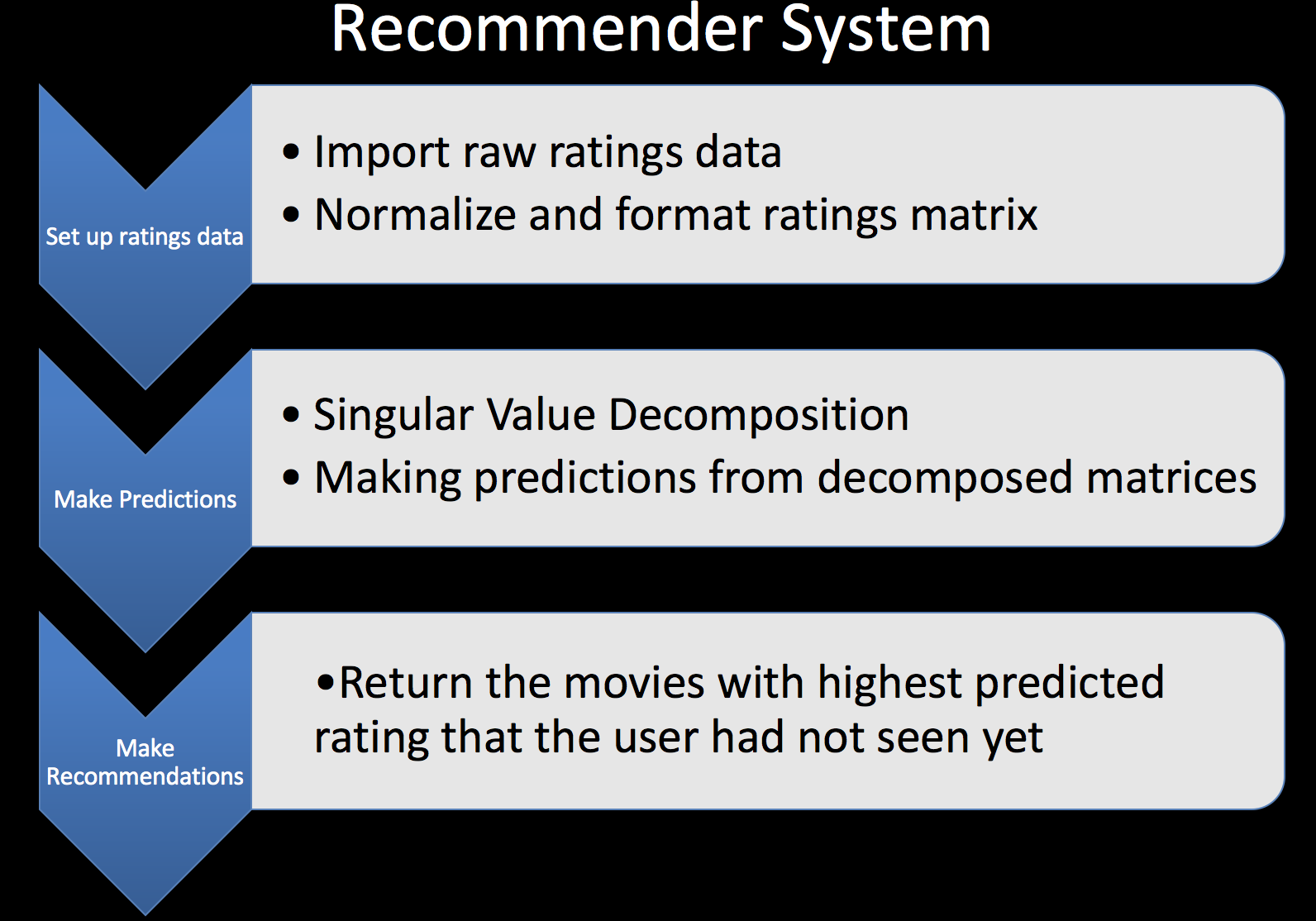
 

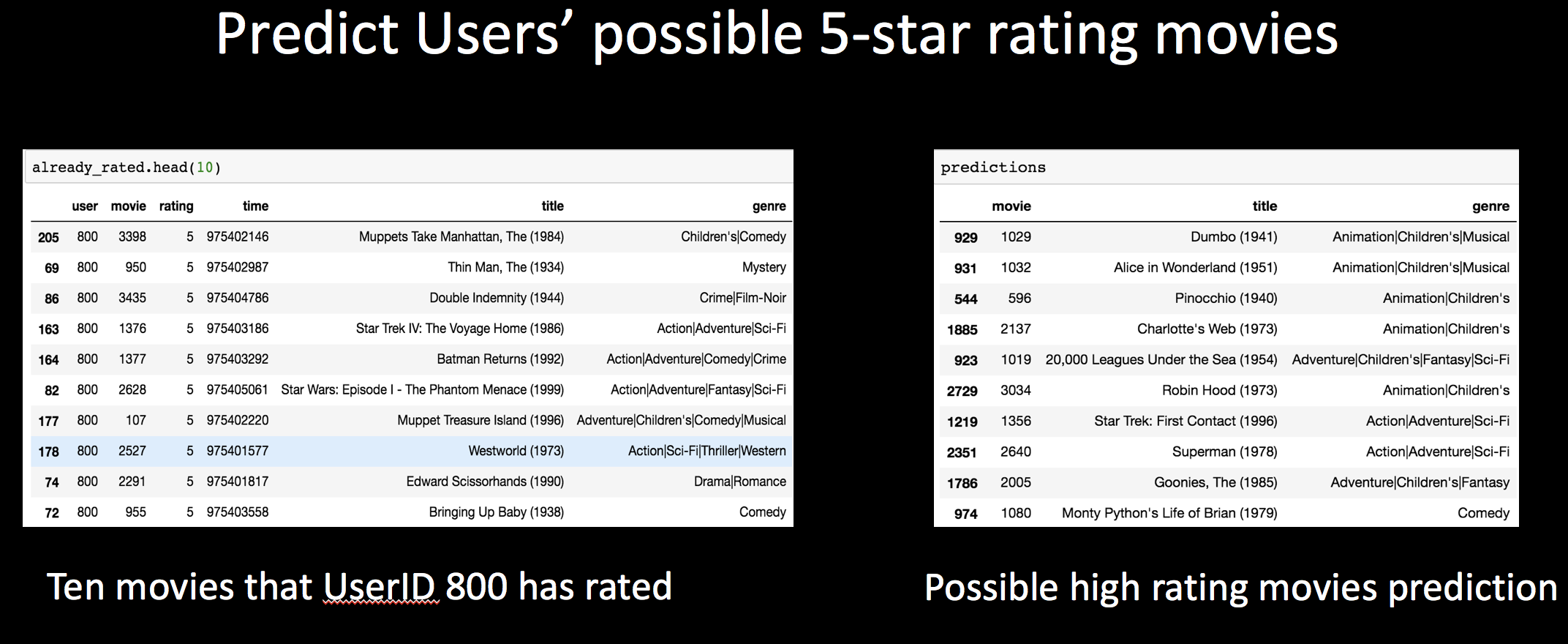
1. However, females and males did not rate different movie genres similarly. We can create a new variable, 'diff', for rating difference of males - females. Then we can sort the mean\_ratings with the largest difference first. The biggest disparity in rating where female movie genre ratings measured higher than males belonged to 'Action|Comedy|Romance|Thriller', 'Children's|Musical', 'Comedy|Horror|Musical|Sci-Fi'. Conversely, if we flipped the order around, where the male ratings were higher than female ratings, the differential pointed to 'Action|Adventure|Comedy|Horror', 'Action|Western', and 'Action|Comedy|Crime|Horror|Thriller'. Both sets of films' genre, while good, cater stereo-typically to different genders.
2. Another way to measure the differential is through the spread, otherwise known as the standard deviation. For male, 'Action|Horror' took the top prize with the highest standard deviation of 1.271567. Therefore, the ratings male users assigned to 'Action|Horror' movies differed more among each other than for other genres; For female, 'Sci-Fi' undoubtly ranked the TOP 1 movie genre with the highest standard deviation of 1.386618.
3. Business Intelligence: when the company produces a particular film, we can find its genre and tell them whether their marketing strategy should focus more on male or female users and the company will make profit when advertising to appropriate potential fans.
4. Explore age groups





1. Recommender System





Business Decision

For choosing a film or a script, we can :

1. Choose Crime, Romance, Adventure, Thriller for a conservative choice.
2. Choose Sci-Fi, Action, Comedy if you are confidence for the script. These genres have a great market space to succeed.
3. Choose a combination of popular genres and unpopular genres with high average ratings which shares a lot common from the word cloud.

For promoting, we can:

1. For any movie companies, when they produce new movies, before they do marketing strategy decision making., it’s important to find their target fans .
2. Our exploration of gender and age group preference for different movie genres will help movie companies to know their target high-rating users in general and make specific strategy for them.
3. We build such recommendation system to predict potential high rating movies that users have not rated yet. Movie companies can find users' individual taste or preference for the films and make particular marketing strategy for them. In this way the company could save a lot of money in advertising and achieve "precision marketing" which is the best approach to do data-driven business decision.

Prospect

Interesting things

In the analysis before, we could find there are a lot of similarity in Adventure, Action, Comedy and Crime, although some of them are popular but some are not. This inspires us that we could find the commons between different genres. Actually many films are combinations of different genres.

So when choosing a film or a film script, it is reasonable to choose a combination with hot genres and cold ones. For example, a crime film but also a comedy.

Also remember the year 2005. In 2005 people review a lot of movies but rate them really low.

Improvement

Remember we didn’t use the links table. This table allows us to get more information of a movie from IMDB and TMDB. For example, we could find the budget , box office, running time, directors and cast. When we get these data, it is likely for us to find which genre could use little budget to get high box office, which genre tend to get high box office, which director is best for a certain genre, which cast is the ultimate cast for a certain genre, etc. These findings could help the company a lot in choosing script or making films.