Exploratory Data Analysis

Basic Statistics (#Ratings, #Users, and #Movies):

```
Total data
-------------------------------------------------

Total no of ratings : 20000263
Total No of Users   : 138493
Total No of movies  : 26744
```

By seeing the above information I'm just understanding that how many ratings, users and movies are there in my dataset. I've almost 20000263 ratings given by the 138493 users on 26744 movies. And also this gives me very high level overview.

Number of movies rated by a user:

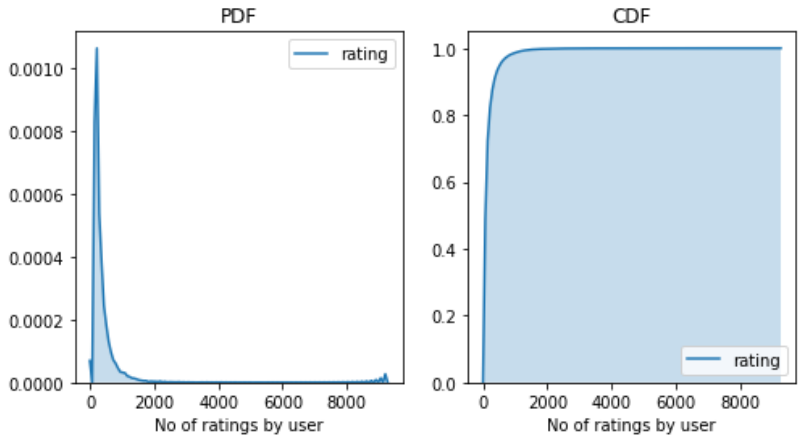
```
no_of_rated_movies_per_user = ratings_data.groupby(by='userId')['rating'].count().sort_values(ascending=False)

no_of_rated_movies_per_user.head()
```

```
userId
118205    9254
8405      7515
82418     5646
121535    5520
125794    5491
Name: rating, dtype: int64
```

The above data says that the number of movies rated by a given user. If you see the above user Id with 118205 gave 9254 movies that seems to be very large to me for a single user to give. Likewise for every user how many movies he rated with corresponding count we calculated.

Finding PDF & CDF for no of ratings per user:



We plotted CDF & PDF here, we quickly noticed that the peak here tells us that, most of the users rate only a few movies. But there are few users here giving lots of rating. And if you look at the CDF also, almost 90% of people gave very few ratings.

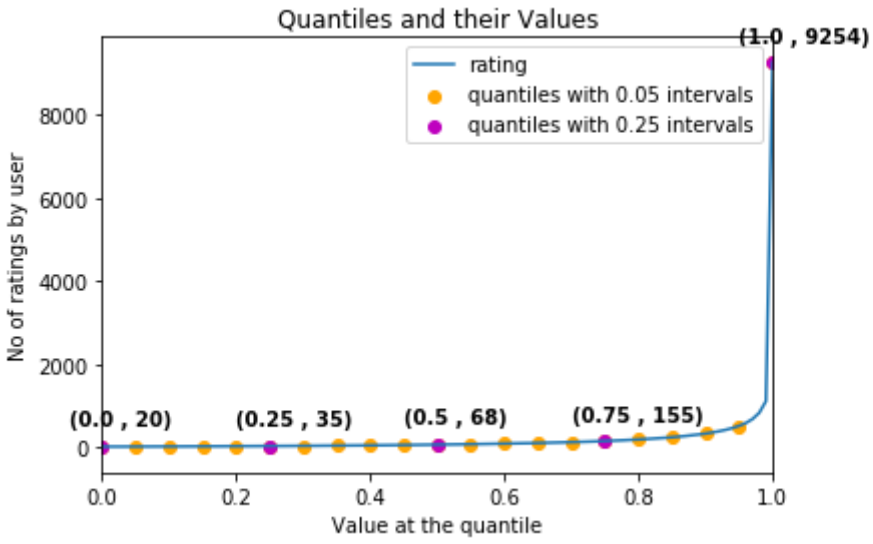
```
no_of_rated_movies_per_user.describe()
```

```
count    138493.000000
mean        144.413530
std         230.267257
min          20.000000
25%          35.000000
50%          68.000000
75%         155.000000
max        9254.000000
Name: rating, dtype: float64
```

From the above data, we observe that the average number of movies that are rated per user is about 144. This shows that most of the users rated lot of movies. Max number of ratings are 9254 and minimum number of ratings are 20. And if you see the median number of movie rated by a user are 68. That means almost 50% of customers have rated more than 68 movies.

We thought by looking at PDF & CDF, we're not able to get it so well. So we went on understanding about percentiles.

Let's get all the percentile values.



If you notices, what's happening here is, each of the violet circle represents 0.25 intervals. And also we plotted yellow circle it represents 0.05 interval. At 0.25 percentile there are 35 movies

rated by the users. And if you closely observe, at 0.95 percentile is also quite low, only 100% percentile is very large.

We actually printed those values.

```
quantiles[::5]
```

```
0.00      20
0.05      21
0.10      24
0.15      27
0.20      30
0.25      35
0.30      39
0.35      45
0.40      51
0.45      59
0.50      68
0.55      79
0.60      93
0.65     108
0.70     127
0.75     155
0.80     193
0.85     246
0.90     334
0.95     520
1.00    9254
Name: rating, dtype: int64
```

At 0.95 percentile also, it is showing 520 movies, there are 5% of users who rated more than 520 movies.

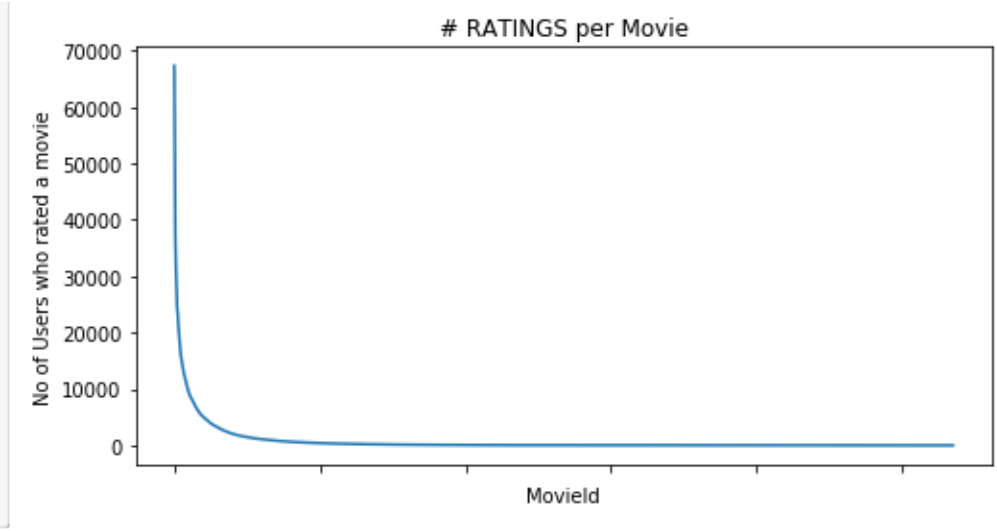
```
print('\n No of ratings at last 5 percentile : {}\n'.format(sum(no_of_rated_movies_per_user>= 520)) )
```

```
 No of ratings at last 5 percentile : 6940
```

We also saw that No of ratings at last 5% percentile are 6940 movies. But this gave us good sense on how many ratings that each user has provide.

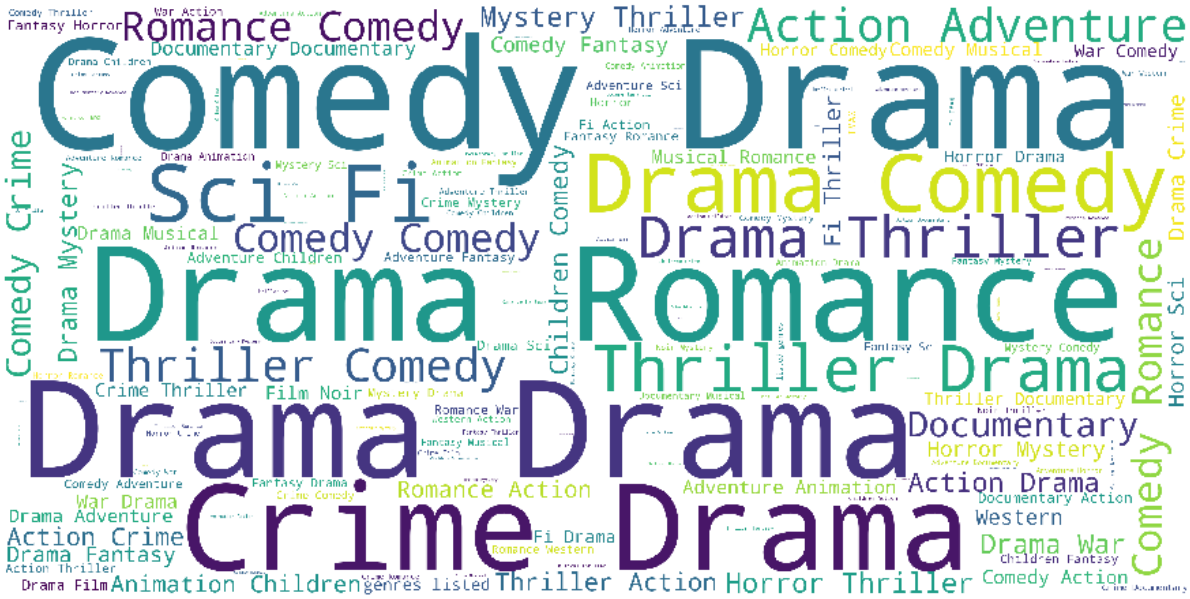
Analysis of rating of a movie given by a user:

For a given movie let's find the no of users who rated a movie. Because there will be some movies like titanic which are liked by millions of people across the world and hence there will be millions of the rating for a movie like that. But there are some other movies which are liked by very few of them.



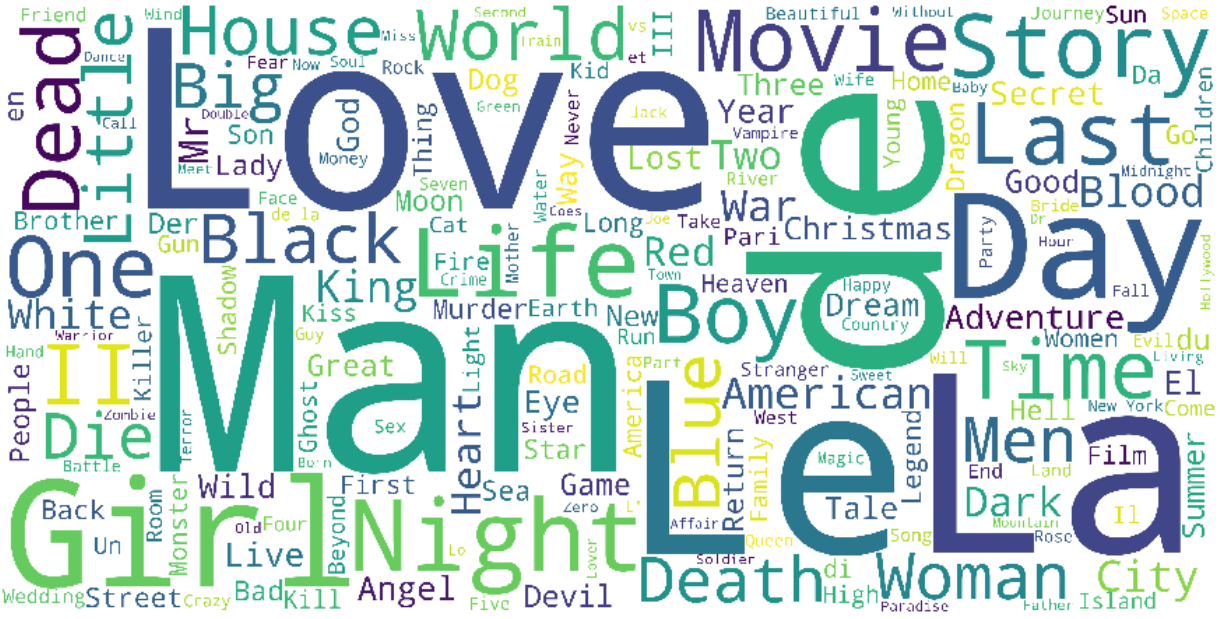
From the above fig, X-axis will be the Movieid, and Y-axis will be no of users and we plotted it. As you'll see there are very small no of movies which are rated by lots of people. But most of the movies rated by a fewer no of people.

Movie genres and overview wordclouds:



Drama is the most commonly used word in Movie genres. **Comedy** and **Romance** are also popular in Movie genres. Together with **Drama**, **Comedy** and **Romance**, these wordclouds give us a pretty good idea of the most popular genres present in movies.

Title and overview wordclouds:



Love is the most commonly used word in Movie titles. **Man** and **Girl** are also popular in Movie Blurbs. Together with **Love**, **Man** and **Girl**, these wordclouds give us a pretty good idea of the most popular themes present in movies.