My Colab link is:
https://colab.research.google.com/drive/10V---AkJKw4uwyYtpQDtHfNgLRsIvoxz

Overview

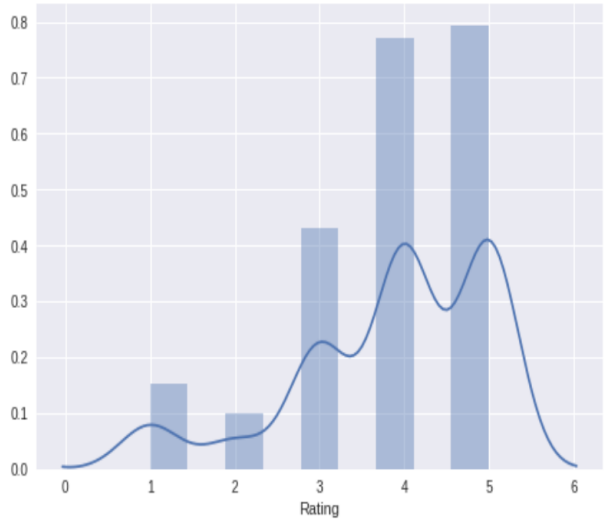
In this assignment, I use the data collected from our class last year. The data describes the class's ratings of 50 movies that won Oscars. I will make recommendations based on my analysis.

Technique

To make recommendations, I use two methods, collaborative filtering, and Pearson's R correlations.

First, I load the data and clean all missing values. Because missing values, in this case, cannot give helpful information in this case, I just delete the rows containing a missing value.

To have a better understanding of the data, I draw the distribution of 5 rating levels and find the majority of ratings are concentrated on 4 and 5 scores. I also list the top 10 movies that have the highest ratings.



Movie ID	Rating
12	35
5	32
35	32
7	31
39	25
1	23
47	21
49	20
4	20
32	20

When doing recommendation with collaborative filtering, I split the whole data into three folds and use algorithm SVD. I also measure the RMSE and MAE of the model. The result is shown below:

Fold N	RMSE	MAE
1	0.8710	0.7343
2	0.8605	0.7069
3	0.8425	0.6760
Mean	0.8580	0.7057

Then, I use this method to find movies I would like to watch. So, I list all the movies I

have rated and calculated estimate scores for the rest of 50 films. There are 10 films I may rate more than 4 and maybe I can watch these recommend films to test how well this algorithm works.

The second method I use is Pearson's R correlations, which measures the linear correlation between review scores of all pairs of movies. When I give the algorithm a movie name, it will then list the top 10 movies with the highest correlations. For example, when I input *La La Land*, it recommends 10 movies including *Precious* and *A Separation*.

```
For movie (La La Land)
- Top 10 movies recommended based on Pearsons'R correlation -
PearsonR                   Movie Name  count      mean

     1.0                      Precious      3  3.333333
     1.0                  A Separation      3  3.333333
     1.0     The Secret in Their Eyes      2  2.500000
     1.0   Blue is the Warmest Colour      2  2.500000
     1.0                  Toni Erdmann      2  2.500000
     1.0                         Amour      2  2.500000
     1.0          Inside Llewyn Davis      2  2.500000
     1.0  Beasts of the Southern Wild      2  2.500000
     1.0                    La La Land     32  4.187500
     1.0                           Ida      3  3.000000
```

Conclusion

1. The majority of ratings are concentrated on 4 and 5 scores.
2. The two methods are based on different principles, one is to calculate the score of other films by similarity, the other is to find out which movies are most related to the favorite movies through the correlation between them. The predicted results obtained by different methods are not completely consistent.