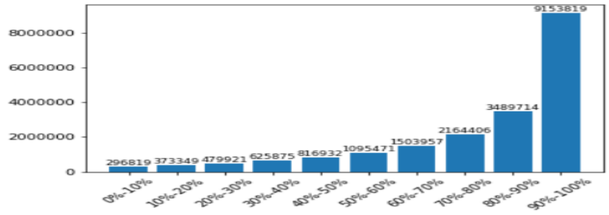
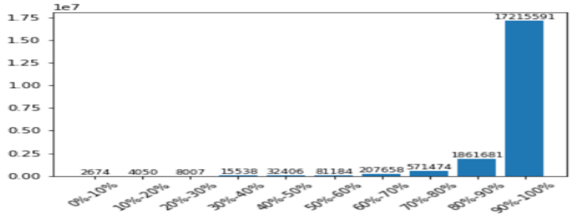
Answer of Q1: The ‘rating’ label in rating.csv in the dataset, MovieLens(ml-20m), can be used to represent user preference(how much a user likes the movie). The higher of the value of rating means higher appraise.

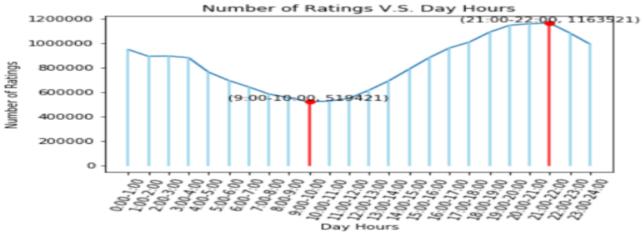
Answer of Q2: I define the interaction as the number of movies rated by a user. By sorting the interactions ascending and grouping each 10% class of all the users, the figure below is generated. X-axis is each of the 10% class with number of 13849 users(total 138493 users). Y-axis is the number of rated movies by each of the 10% group. The top 10% group(90%-100%) rated 9153819 movies, and the lowest 10% group(0%-10%) rated only 296819 movies.



Answer of Q3: I define the interaction as the number of users that rated a movie. By sorting the interactions ascending and grouping each 10% class of all the movies, the figure below is generated. X-axis is each of the 10% class with number of 2674 movies(total 26744 movies). Y-axis is the number of rating users that rate each of the 10% group of movies. The top 10% group(90%-100%) has a number of 17215591 users rated these movies, and the lowest 10% group(0%-10%) has only 2674 users rated these movies.

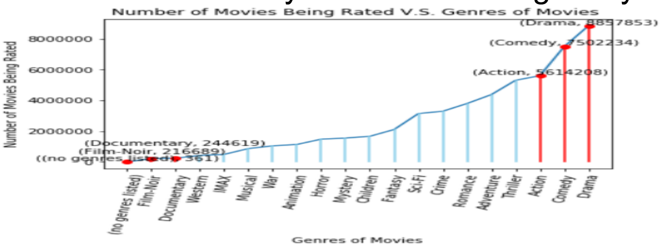


Answer of Q4: I define the interaction as the number of ratings of movies by users. The Y-axis means the total number of ratings of movies in different time slot. The X-axis is the time slot of an hour in a day. The figure shows the highest number of rating occurs during 21:00-22:00, which is 1163521 among totally 20000263 ratings. The lowest number of rating occurs during 09:00-10:00, which is 519421.



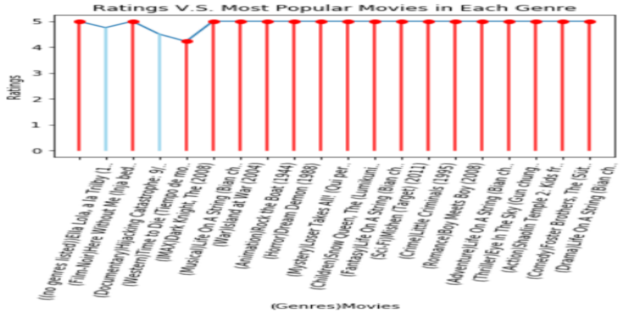
Q5: What are the three most popular and the three least popular movies in the dataset? This question is valuable because you can learn the taste of movies of most of the people. The result is expected to show the order of the three most popular and least popular genre of movies. For example, the order of the most three popular movies are Drama > Comedy > Action, and the order of the least three popular movies are (no genres listed) > Film-Noir > Documentary.

Answer of Q5: I use the number of being rated in a movie as the degree of popularity. The more number of the rating in a movie(column 'rating' in ratings.csv), the more count on each of the genre of that movie. As the below figure, the three most popular genres of movies are Drama(8857853 rating) > Comedy(7502234 rating) > Action(5614208 rating), and the three least popular genres of movies are (no genres listed)(361 rating) > Film-Noir(216689 rating) > Documentary(244619 rating). The result is as I expect because the genres of Drama, Comedy and Action are much more interesting and they make people relax. The Film-Noir and Documentary are much more gloomy and boring.



Q6: What is the most popular movie in each of the genres? What are their average ratings? The question is valuable because you can combine the result with **Q5.** For example, you already know that Drama is the genre that most people like, and you can dig deeper to find out which movie in the Drama genre is the most popular, and you may be interested in that movie.

Answer of Q6: I sort the average rating number of each movie in each genre to get the highest one as the most popular movie in that genre. So the Y-axis is the average rating number of the movie. The X-axis is the most popular movie in each of the genre, and the order is same as the X-axis in figure of Q5. The result may not be fully expected since I do not even know much of the movies like 'Life On A String'. I think the reason is that I do not use the 'number of rating' in the popularity calculation(only choose the highest average rating as popularity). Some unpopular movies may be only rated by 1 person but with the highest '5' rating. Adding a threshold to the 'number of rating' may make the result more convincible.



Feedback: This homework makes me practice the using of pandas dataframe and some numpy library manipulation for statistic calculation. The Q2-Q6 questions also needs the ability to use matplotlib to plot your figures for illustration your data mining results. The Q5 and Q6 also make me practice the ability the explore the dataset freely with my own imagination but not loosing conscientiousness in validating the mining results.