For this assignment, my distance metric was the difference in the user's ages and the difference in the user's average rating for a specific genre. I also weighted the user's age difference more than the average rating in order to emphasize the age groups more. I ran my code for each genre that was represented in the data, so that I was able to get 17 outputs for all 17 genres.

The question that I was trying to answer was: "What genres do different age groups like?". I am able to answer this question by running my code for all different genres. By itself, my distance metric is able to answer the question: "How do the various age groups feel about this specific genre?". After running the code for all genres, I can tell how each age group feels about each genre, which in turn tells me what genres that each age group would prefer.

Since I wasn't sure what K value to use in my K-means implementation, I tried a variety of K values to determine one that worked well. I ended up using a K value of 16. This worked the best because with a higher value of K, I was able to get more precise rating averages. On the other hand, a value of 16 isn't too high that the program takes a long time to run. The value of 16 also split the users fairly nicely. Since we can split the users into 8 age groups of size 10, 16 centroids gives us 2 per age group (in theory), which I believe helps with splitting the groups nicely.

To analyze my results, I manually looked at my output files to determine how age groups felt about each movie. I could look at the output file for each genre, and see the average rating for each centroid and which age group that centroid fell into. This would tell me, roughly, how that age group felt about that move genre. It is important to note that if two or more centroids fell into the same age group, I took the average rating of all those centroids.

It also is important to note that not all age groups are represented equally, so this may skew the data slightly. I have included a bar graph that shows the distribution of users per age group to accompany the data, so that readers can see that some groups may not be fully accurate.

Additionally, one thing I noticed while analyzing the data is that users who did not rate any movies in a certain genre were assigned an average rating of 0 for that genre. This could also cause problems in the results and not as accurately reflect the true results. To try and counter this during my analysis, any average ratings that were less than 1 were not used. I believe that this would greatly counter the effect, because users with an average rating of 0 would be clustered together (so not counting this centroid's average would, in theory, remove such users).

There are a few interesting insights that we can gain from the data and my analysis. The first interesting thing to note is the output from the 'Crime' genre. If we look at the 30-39 age range, we can see that there are 2 centroids that ended at age 30 and 2 that ended at age 34. The interesting thing is that these two centroids had widely different rating averages. This didn't show up as heavily in my analysis since I took the averages of those, but this was cool to see that the centroids were actually highlighting rating differences even within the same age groups.

The second interesting thing I noticed from the analysis is that 20-29 year olds really like Documentaries compared to all other age groups. This was the most pronounced difference out of all of the outputs, since all other age groups had an average below 2. Film-Noir movies were highly preferred by 40-49 year olds, compared to the other age groups. Finally, Western movies were also highest rated among 20-29 year olds. The full analysis table is included in a separate PDF in this directory, along with some bar graphs that highlight some of the larger differences.