M6: Spark Assignment

Cheryl Ngo, Daniel Jagdmann, Thomas Fatale, & Tyler Johnson

The direction and goal of our assignment in Spark was to analyze the popularity of movies from the MovieLens data set by age. There were three files from the data set that were needed: *u.data*, *u.items*, and *u.user*. The *u.data* file provided the user's rating value per movie where the user and movie are numeric IDs. The *u.items* file was joined to the data to provide the movie names and the *u.user* file was used to provide the age of the specific user.

Methods:

Our group decided to explore the dataset by deriving insight based on age bracketing of 10 year windows (10-19, 20-29, etc.). We wanted to explore 3 key aspects; how many ratings did the overall top 10 rated films receive from each age bracket, what was the average rating for the overall top 10 rated films from each bracket, and what were the top 10 rated films for each age bracket? We also decided to add additional parameters that films must have at least 50 ratings total to be included in our first two pivot tables, and films must have at least 30 ratings from the specific age bracket to be included in an age bracket recommendation list.

Results:

The highest rated movie among each age group had no higher than a 4.6 rating. It was difficult for a movie to score higher than 4.6 unless it had a particularly low number of reviews. Similarly, the age groups 0-9, 60-69, and 70-79 do not have an age bracket top ten list because no movies had over 30 reviews from that group.

The overall top rated film with over 50 ratings was *Wallace & Gromit: The Best of Aardman Animation*. Notably, this film would not have made the cut if the required number of reviews was too much higher than 50 as the rest of the films in the list had many more reviews. The high ratings of this film were driven in quantity mostly by the 20-29 age range but it was also highly rated by the 40-49 age range. From this, we can conclude that it was not widely popular--it even has a 3.6 average rating from the 50-59 age group--but was a favorite of its core audience.

Similarly, *12 Angry Men* yielded interesting results. While ranking 9th in the overall top 10 films pivot tables, the film failed to make the top 10 recommended films for any of the age brackets. This is likely due to the film's limited number of reviews from the 10-19 and 40-49, and higher age brackets.

The film that drove the most widespread popularity would be *Star Wars*. It is one of three movies in the overall top ten list that had a rating from the 70-70 age group and the only one to have a rating from the 0-9 group. While the reviewers do skew younger, the movie still appears as the highest rated movie in the 50-59 top ten list. For this combination of factors, we would consider it to be the most popular movie of those included in the data.