Movie Recommendation System

With this project we aim to build a movie recommendation system. We are using the MovieLens 25M dataset available on grouplens.org. The data is present across 6 csv files : `genome-scores.csv`, `genome-tags.csv`, `links.csv`, `movies.csv`, `ratings.csv` and `tags.csv’. The dataset contains 25 million ratings and one million tag applications applied to 62000 movies by 162,000 users. This data was created between year 1995 and 2019. For each movie we have the title and the genre it belongs to. Movie titles are entered manually or imported from other sources and include the year of release in parentheses. Genres are a pipe-separated list and falls into 1 or multiple of the 18 categories(Action, Adventure, etc.). This data is present in ‘movies.csv’ file. For each user - movie pair we have the movie rating the user has provided along with the timestamp. Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars). Timestamps represent seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970. This information is contained in ‘ratings.csv’ file. All selected users have rated at least 20 movies. For each user - movie pair we also have the associated tags(sci-fi, fiction, tense, climate, etc.) along with the timestamp. Tags are user-generated metadata about movies. Each tag is typically a single word or short phrase. The meaning, value, and purpose of a particular tag is determined by each user. This information is present in ‘tags.csv’ file. We also have the tag genome data available. The tag genome is a data structure that contain tag relevance score for movies. The tag genome basically encodes how strongly movies exhibit particular properties represented by tags. This information is present in ‘genome-scores.csv’ and ‘genome-tags.csv’ files. Additionally, we also have ‘links.csv’ file which contains identifiers that can be used to link to other sources of movie data such as IMDB website.

From this extensive dataset, we aim to answer question like i) Which genres has the maximum number of movies and how does it change with time? ii) Which genres receive the highest ratings and how do these ratings change with time? iii) distribution of star ratings across genres iv) what tags best summarize a movie genre? v) best movie of every decade vi) best years for a genre vii) genre preference for specific users viii) association rules for users watching movies.

We would start off with tidying the data(if necessary) using R such as i) dealing with NAs ii) separating genres which has values like “Action | Comedy” into separate rows iii) timestamp conversion into readable format iv) extracting year from movie title

We then aim to do data visualization in R using Exploratory data analysis techniques. We plan on utilizing cosine similarity to find the similarity between 2 users and 2 movies. We then plan on using Market basket analysis to find association rules.