Problem Statement - MovieLens Case Study

Background: The GroupLens Research Project is a research group in the Department of Computer Science and Engineering at the University of Minnesota. The data is widely used for recommendation systems. However, we will be using this data to act as a means to demonstrate our skill in using Python to “play” with data.

Objective: To extract valuable insights from the dataset.

Key Questions to be answered (and steps):

1. Import all the necessary packages.
2. Read all the datasets into pandas dataframes.
3. View the first 5 rows of all the datasets.
4. What are the shapes of all the datasets?
5. What are the data types of all the columns in datasets?
6. Provide a statistical summary of all the datasets.
7. Find the number of movies per genre in the movie dataset.
8. How many movies have more than one genre such as Action, Drama, etc?
9. Which 25 movies have got the highest average ratings?
10. What is the gender distribution across the genres?

Datasets:

1. rating.csv: It contains information on ratings given by the users to a particular movie.

user id: id assigned to every user

movie id: id assigned to every movie

rating: the rating given by the user

timestamp: Time recorded when the user gave a rating

2. movie.csv: The file contains information related to the movies and their genre.

movie id: id assigned to every movie

movie title: Title of the movie

release date: Date of release of the movie

Action: Genre containing binary values (1 - for action 0 - not action)

Adventure: Genre containing binary values (1 - for adventure 0 - not adventure)

Animation: Genre containing binary values (1 - for animation 0 - not animation)

Children's: Genre containing binary values (1 - for children's 0 - not children's)

Comedy: Genre containing binary values (1 - for comedy 0 - not comedy)

Crime: Genre containing binary values (1 - for crime 0 - not crime)

Documentary: Genre containing binary values (1 - for documentary 0 - not documentary)

Drama: Genre containing binary values (1 - for drama 0 - not drama)

Fantasy: Genre containing binary values (1 - for fantasy 0 - not fantasy)

Film-Noir: Genre containing binary values (1 - for film-noir 0 - not film-noir)

Horror: Genre containing binary values (1 - for horror 0 - not horror)

Musical: Genre containing binary values (1 - for musical 0 - not musical)

Mystery: Genre containing binary values (1 - for mystery 0 - not mystery)

Romance: Genre containing binary values (1 - for romance 0 - not romance)

Sci-Fi: Genre containing binary values (1 - for sci-fi 0 - not sci-fi)

Thriller: Genre containing binary values (1 - for thriller 0 - not thriller)

War: Genre containing binary values (1 - for war 0 - not war)

Western: Genre containing binary values (1 - for western - not western)

3. user.csv: It contains information about the users who have rated the movies.

user id: id assigned to every user

age: Age of the user

gender: Gender of the user

occupation: Occupation of the user

zip code: Zip code of the user