Preprocessing and analysis in assignment 2

Dataset

I used the **MovieLens100k** dataset that contains 100k ratings from 973 users to 1682 movie. This dataset contains not only the information about ratings, but the information about users, like occupation and age, and about movies, like genre or video release year.

Data preprocessing

I removed the following information from the data:

Users dataset

- Zipcode

Movies dataset

- Link to imdb
- Title of the movie
- video release year

They were removed because they do not contribute much to the choice of movies by users.

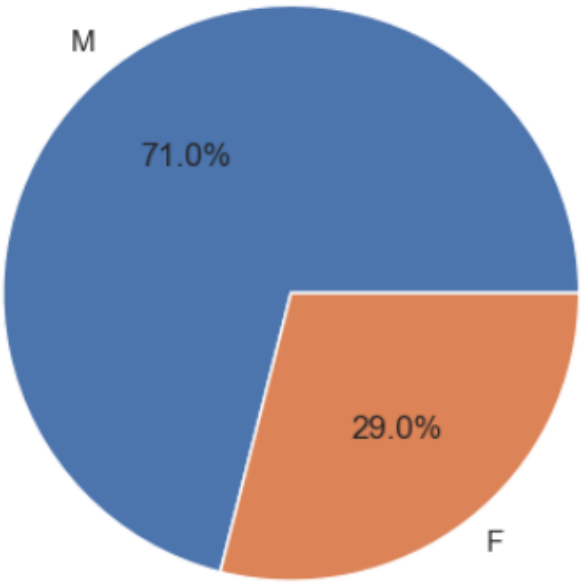
A lot of other information was one-hot encoded, like the age of the users (separated into categories), gender of the user, occupation of the user etc.

Also, for the future needs, all the datasets were transformer into tensors and saved as tensors

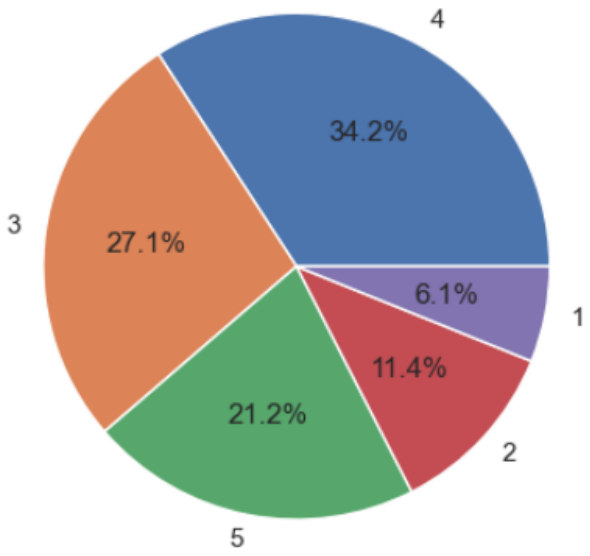
Data analysis

Here you can see the distributions of the gender and ratings.

/imag



Gender distribution

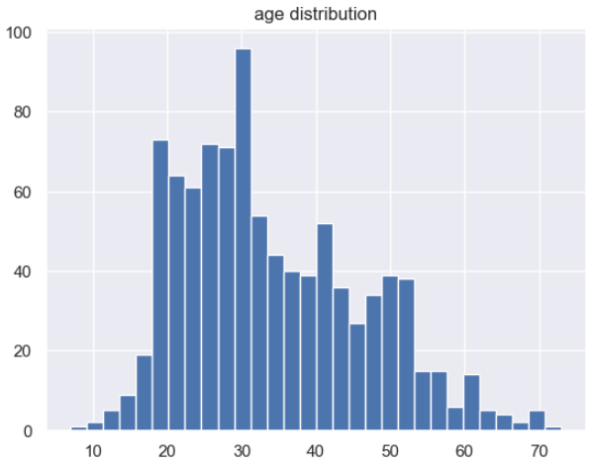


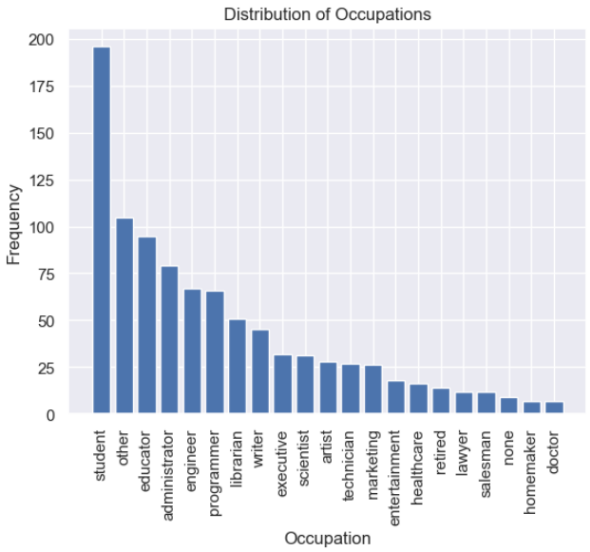
Rating distribution

You see here, that there are more males than females among users.

Also, users use rating 4 most frequently.

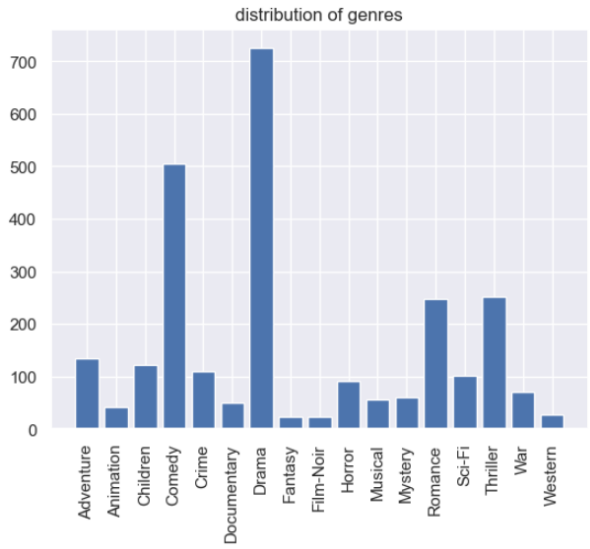
Here is the distribution of the age and occupations

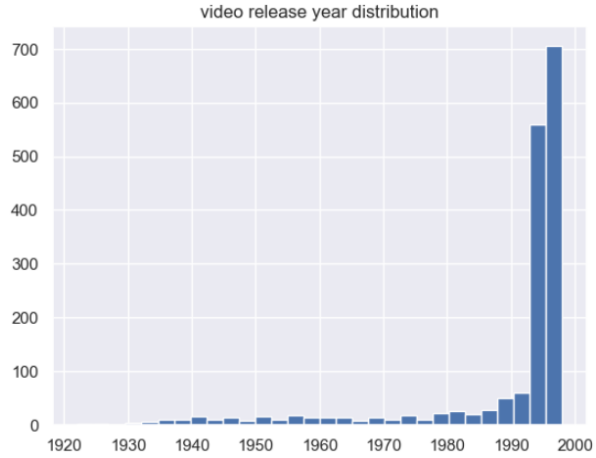




So, students are the majority of the users. Moreover, there are more young users up to 30 years than old users.

Finally, here is the distribution of movie genres and video release years





Here we see that most of the movies are dramas and comedies. Also, most of the movies are released in videos in 1990s.