# Experiment Design for Data Science

Teodor Chakarov - 12141198

## Question 1 - Dataset exploration

We load the files with names *u.data* (containing data with ratings, users-id and item-id), *u.item* (containing data for a movies) and *u.data* (containing data from users).

In the dataset about ratings we can see the number of ratings. Based on **Figure 1** we can state that there are not so many negative comments and the most are around the middle (3 and 4s).

In the movies dataset we can see that the majority of the movies are being created after 1990 **(Figure 2)**. So we have not well distributed movies based on the year of creation as well. Maybe this is because of the advanced technologies and investments in the film industry after 1990s.

Another unbalanced attribute is the gender. We can see on **Figure 3** the male accounts are 71%. That is why we cannot rely on numbers based on gender.

In **Figure 4** we can see that women rate more thriller, western and musical movies and men rate more Film-Noir, Action and Comedy movies. But the difference is not so big when it comes to genres rated more from males. That could mean that men tend to give worse rating for every genre than the women.

Potential ethical issues could be showing people with specific occupations not liking a specific genre as much and because of that stereotypes can occur towards jobs.

Figure 1 - Number of ratings Figure 2 - Number of movies being released for year

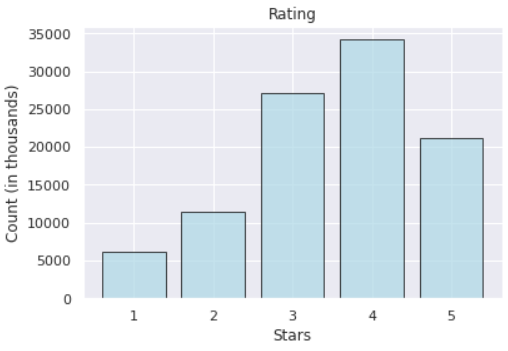
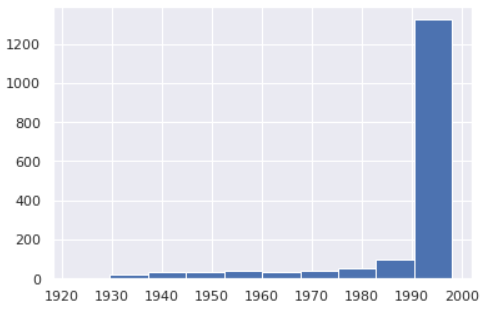


Figure 3 - Gender count

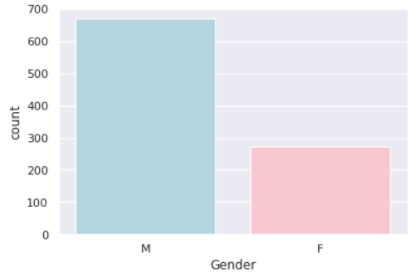
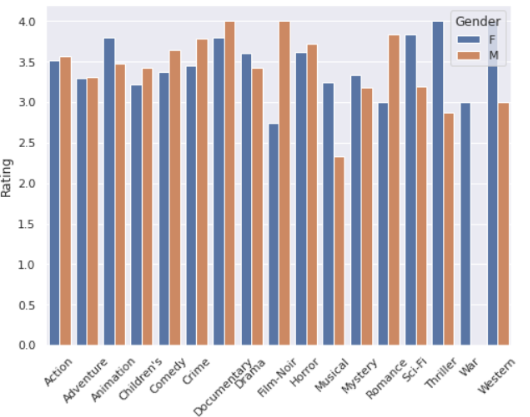


Figure 4 - Average ratings by genre for each gender



Task 2

Hypothesis: People who has the same gender, age and occupation based on the genre of the movie will have the same rating.

Independent Variable: age, gender, occupation and move genre

Dependent Variable: rating

Control:

Group A (treatment group): People with same age, gender and occupation will rate movies with the same genre.

Groups B: People with different age, gender and occupation will rate movies with the same genre.

Performance Indicator: Comparing the results on different groups. If Group A is valid (people rate the movies +-1 rating). And if Group B will validate that people with different characteristics tend to give different rating.

Prediction Model: kNN will be good model based on its property to cluster our data based on the difference between people and predict based on the k - nearest data points

Experimental Setup: We will train our kNN model based on our supervised data. We will encode the categorical variables like gender, occupation and movie genres. We will evaluate how good our model can cluster the data and distinguish

Task 3

Hypothesis: The positive ratings are given more from older people than younger once

Independent Variable: gender, occupation, rating and move genre

Dependent Variable: age

Control:

Group A: Old people with different gender and occupation will rate different genres of movies.

Groups B: Young people with different gender and occupation will rate movies with the same genre.

Performance Indicator: Comparing the results on different groups. If results in group A are better in general in comparison with Group B so our Hypothesis is valid.

Prediction Model: We can estimate the age of the people with linear regression or decision tree regressors.

Experimental Setup: We will train our linear regression model based on our supervised data. We will encode the categorical variables like gender, occupation and movie genres. After we predict we will see how well our data is performing.

Task 4

Hypothesis: The budget, actors and film crew is the main reason for good rating.

Independent Variable: actors, budget, film crew

Dependent Variable: rating

Data Missing and Strategy for Data Collection: actors, budget, film crew. In the datasets with movies we will store additional information about the film itself. The cast, budget and the producers.

Possible Issues: Not enough information for determining the rating.