KBTU KAZAKH-BRITISH TECHNICAL UNIVERSITY

Kazakh-British Technical University
School of Information Technology and Engineering

Data collection and preparation
SIS 1 Report

Prepared by: Amirgali Sanzhar
Duman Bayron

Almaty, 2025

Approach

This project was aimed at the comparative analysis of Top 250 movies on the two large film rating websites: IMDb and Kinopoisk. The main aim was to find the statistical associations between ratings by the audience, source of production and the duration of the film and also to determine the extent of correlation between the rating system of the two platforms. Python was used to process and analyze the data with the help of pandas that operates data manipulation and matplotlib that visualized the results. The variables included the title of the movie, the countries of production, IMDb rating, and Kinopoisk rating as well as the length of the movie. The process of data cleaning included the normalization of the format of the ratings, the management of values that were not available, and the confirmation of the numerical integrity. After preprocessing, the exploratory data analysis (EDA) was performed to obtain the descriptive statistics, determine rating distributions, and calculate the pairwise correlation between the parameters under consideration.

Results

The statistical analysis produced several notable findings:

1) Average Ratings: The average IMDb rating of all 250 movies was 8.8 and the average rating of Kinopoisk was 8.55. This points to the fact that the two sites have a high number of highly rated movies, with IMDb ratings generally being higher (Figure 1).

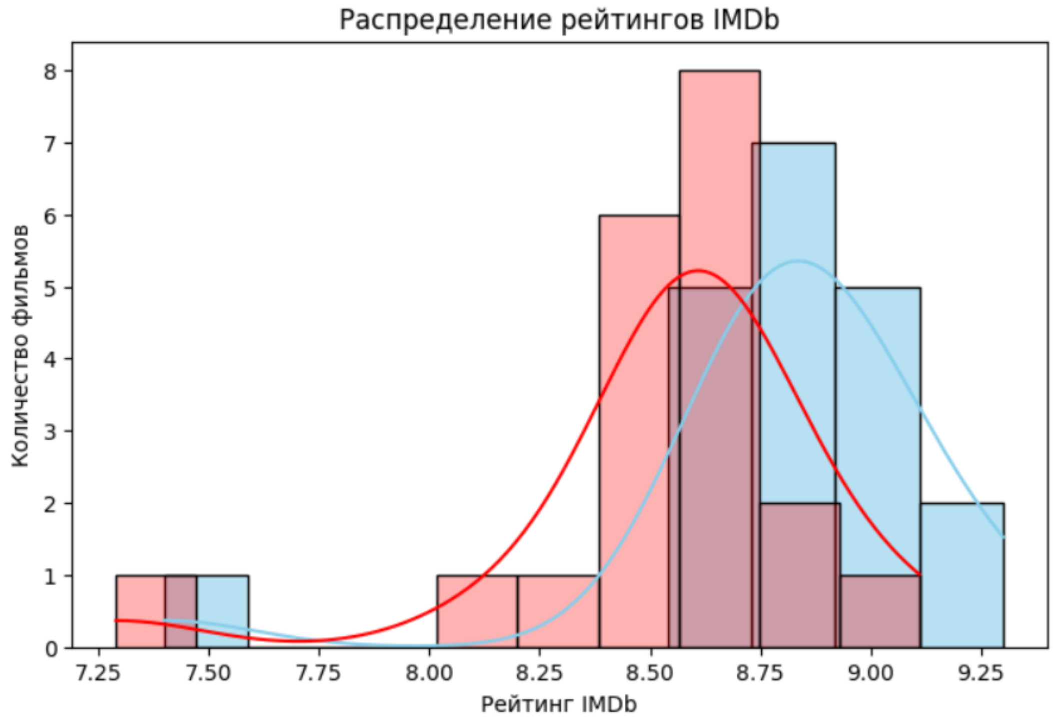

Figure 1 – Distribution of IMDb and Kinopoisk ratings

2) Rating Range: IMDb rating was between 7.4 and 9.3, and the Kinopoisk rating between 7.29 and 9.11 which were similar in terms of rating distribution and dispersion on both platforms. The bar chart shown below illustrates top 10 movies of all time ranked by IMDb ratings (Figure 2).

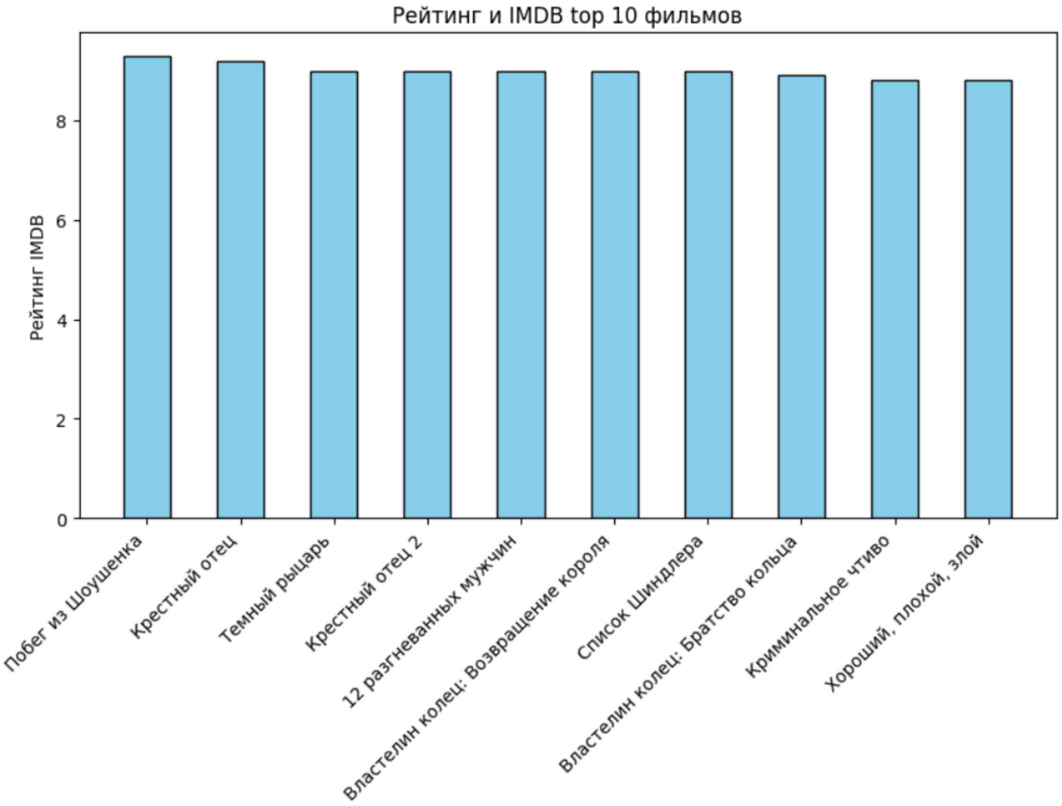

Figure 2 – Top 10 films ranked by IMDb ratings

Looking at the information in more detail, “The Shawshank Redemption” has the highest rating of 9.3 among these films. On the other hand, the legendary film “Il Buono, Il Brutton, Il Cattivo” placed at 10th place with IMDb rating of 8.8.

3) Ratings over time: The change in the IMDb and the Kinopoisk ratings was examined to understand the temporal changes in the preference of the audience in relation to the year of film release. Figure 3 shows that the average rating has been changing with time across the two platforms. The charted lines demonstrate that the ratings are quite steady over decades, and minor changes take place, which can be related to changes in the approaches to cinematography or expectations of the audience. The horizontal lines indicate the general average figures (IMDb = 8.80, Kinopoisk = 8.55), and it is important to note that in the majority of cases, the films of any release date have ratings that are close to the average rating in the long-term. Such consistency means that the parameters of evaluation in the audience have been quite stable across various periods of film production. It could also be a sign that classic themes, good narrative and quality of production can still appeal to the audience irrespective of the

year of release. These minor variations that can be noticed during certain periods might be attributed to the historical events, development of the film technology or the emergence of the new genres. On balance, the trend demonstrates the fact that, really brilliant movies are rated high irrespective of their date of production (Figure 3).

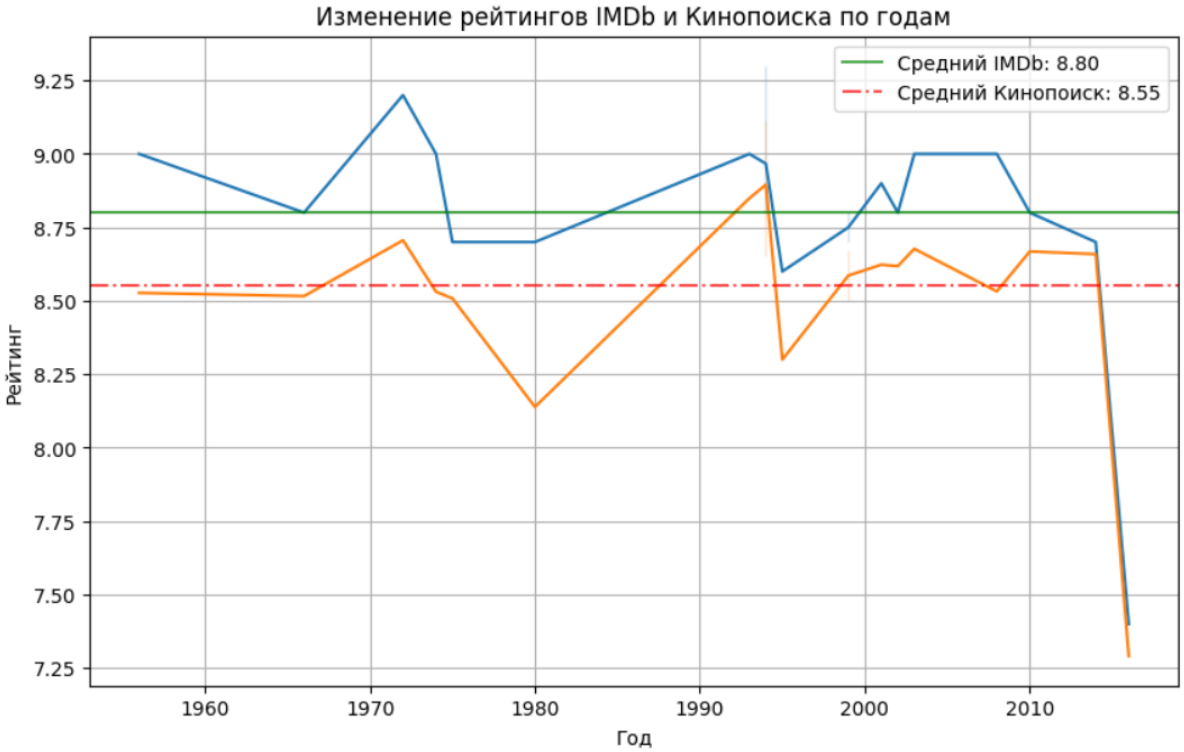

Figure 3 – Fluctuation of ratings throughout the time period

4) Country-Level Patterns: Most of the movies were produced by United States either on its own or in partnership with other nations. As an example, joint productions between USA and United Kingdom were rated at 8.40 and 8.16 on IMDb and Kinopoisk, respectively, whereas films merely produced in the USA were rated at 8.91 and 8.62, respectively. These results indicate that the U.S. cinema is dominant in the world as far as the top rated categories are concerned.

5) Film Duration: The overall mean of the analyzed movies was 154.25 minutes. The Godfather Part II (202 minutes) was the longest movie, and 12 Angry Men (96 minutes) was the shortest movie (Figure 4).

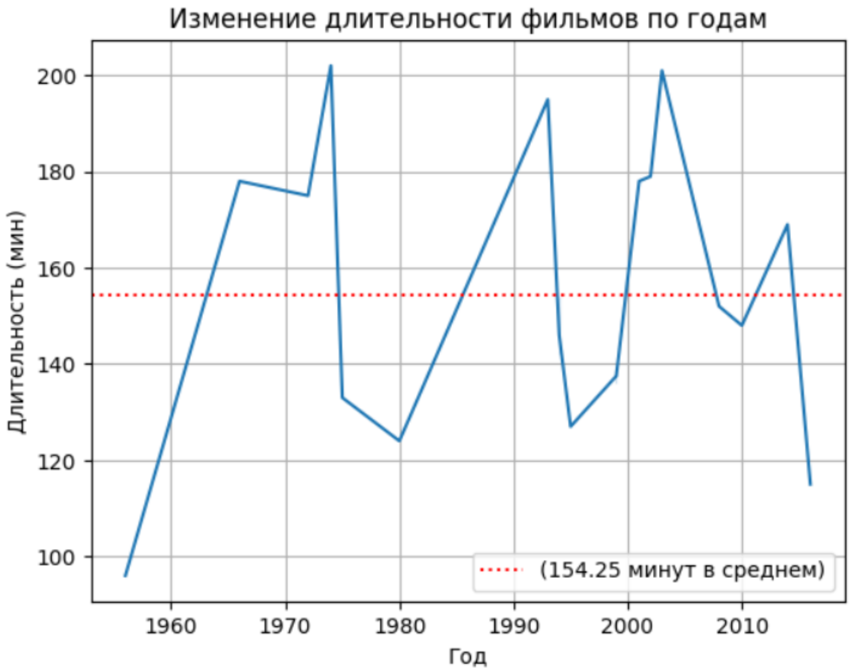

Figure 4 – Change in film length over years

6) Correlation Analysis: There was a Pearson correlation coefficient of 0.888 between the IMDb and the Kinopoisk ratings denoting a strong positive correlation between the ratings of the two sites. Also, the correlation between the runtime and Kinopoisk rating was 0.409, whereas runtime and IMDb rating had the correlation of 0.408, which suggests the moderate positive correlation between the rank of the film length of a film and its rating (Figure 5).

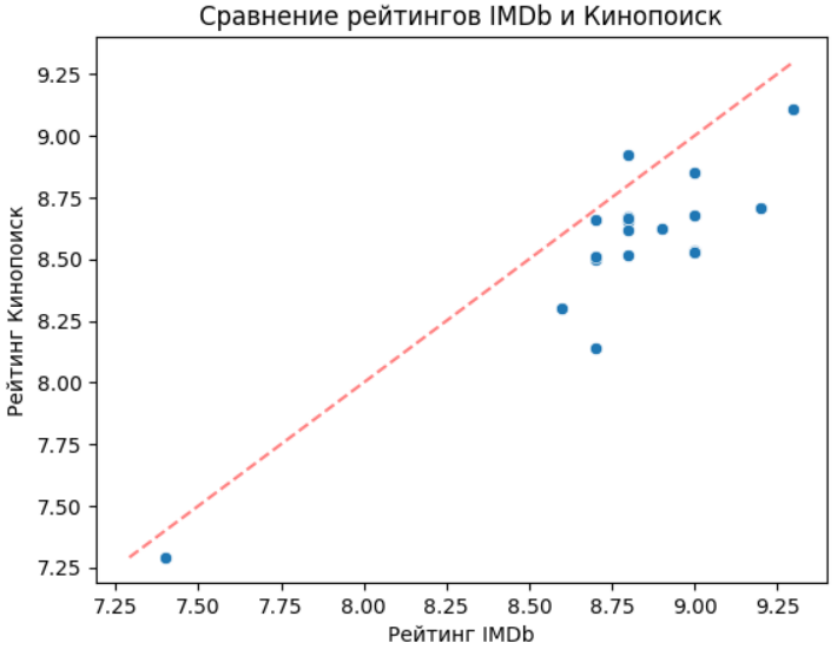

Figure 5 - Comparison of IMDb and Kinopoisk Ratings

Key insights

The comparison shows that the user rating of IMDb and Kinopois is very similar, which indicates that people globally have a similar vision of the quality of the film, irrespective of the culture. The high correlation coefficient (= 0.89) proves the stability of the audience mood of the platforms. Moreover, the evidence shows that the rating of longer movies is slightly higher, which could be connected with the preference to complex plots and characters development that is characteristic of the longer duration of filming. The fact that the list of the top 250 movies includes predominantly U.S.-made ones also underscores the impact of the American cinema on the world audience in terms of creating audience norms and standards of critical evaluation. All in all, the results give a quantitative understanding of preferences of the audience and rating trends of two major film databases. The fact that the correlation levels between platforms are very high is indicative of the universality of critical acclaim and serves to confirm the idea that cinematic excellence is not limited by the national and linguistic boundaries.