Data Collection & Preparation
SIS Project Assignment

Group members: Jakupov Dias, Kabiyev Zhanbek
The topic of the work: The relationship between audience ratings and film box office
The main goal: To get real data about films from open sources, to clean up, combine and visualize key relationships

1. Introduction

We chose this topic because we were interested in checking whether the rating of a film affects its box office. It can be assumed that films with high ratings from viewers and critics earn more, but we decided to check whether this is really the case in practice.
We decided to see if there is a correlation between ratings and fees and what factors can influence the success of the film.

2. Sources of Data

For this project, we used two data sources to ensure the reliability of analytical work. The first source was The Movie Database (TMDB) API, which provides structured data about films, including their titles, ratings, number of votes, release dates, original languages, and genres. Using the API made it possible to receive data in a normalized format, which simplified their further processing and integration.
Additional financial data was obtained from the website www.boxofficemojo.com by web scraping. This resource is an authoritative platform that publishes up-to-date information about budgets, box office receipts and commercial results of films. Before parsing, the rules for using the data specified in the file were checked. robots.txt , in order to comply with the ethical standards and restrictions set by the site owners.
The integration of data from the TMDB API and Box Office Mojo made it possible to combine the audience and financial indicators of films into a single analytical database, providing an integrated approach to the study of the relationship between ratings, budgets and box office.

3. Stages of work

The project work took place in stages. Initially, we collected data on films, including information on ratings, budgets, and box office receipts. After receiving the data, we checked them for omissions and duplicates. The check showed that the data was complete and did not require additional cleaning. Even if some of the values were missing, they would not be included in the final set, since the “inner join” was used when combining the tables. After checking the correctness of the data structure, we combined the sets into one table and moved on to the analysis.

4. Analysis and visualization

We analyzed the data and built visualizations showing the main relationships between ratings, budgets, and box office receipts of films. We also created charts for the distribution of ratings,

budget and profit comparisons, the dynamics of the number of films by year, as well as the top films by fees. The analysis showed that a high rating does not guarantee large collections, but there is a positive relationship between the quality of the film and its financial success. Additionally, we examined the distribution of indicators by genre, which helped to identify connectivity within individual categories.

5. Conclusions

We tested the hypothesis that there is a connection between the rating of films and their collections. The analysis results showed that the correlation coefficient is 0.2589, which indicates a weak positive relationship. This means that films with higher ratings usually earn more, but the impact of the rating on financial success is small. The analysis also showed that the commercial results of a film are shaped by many factors, such as genre, budget, marketing, and the popularity of the actors. Consequently, a high audience rating does not guarantee significant box office receipts, although a certain positive relationship between these indicators is still observed.