Achraf Safsafi
DSC-530 Project Summary

In this project, we will answer the question of" What makes a movie popular?". We hear many explanations from different people. Some associate the success of the movie with the actor or actress, and others claim that the reason is the large budget. Of course, there are many and varied claims, but all of them do not exceed to be personal opinions. It can be false or true. It is not issued by experts and not the result of studies.

To deal with the weaknesses of those anecdotes, we will perform an exploratory data analysis to look for patterns and features that address the statistical question we defined above.

So, to conduct the project, I used a dataset, available in Kaggle, https://www.kaggle.com/danielgrijalvas/movies, of 6820 movies and 15 features, released from 1986 to 2016. But the data that I kept only 8 variables for my project :

budget: the budget of a movie. Some movies don't have this, so it appears as 0
genre: main genre of the movie.
gross: revenue of the movie
rating: rating of the movie (R, PG, etc.)
runtime: duration of the movie
score: IMDb user rating
votes: number of user votes
year: year of release

According to the correlation analysis, I can conclude that the linear relationship between score and budget is moderate negative (r= -0.54) and the budget variable explains about 29% of the score variable (r^2 = 0.29). The linear relationship between score and genre is negligible (r= 0.042) and the genre variable explains about 0.18 % of the score variable (r^2 = 0.00176).The linear relationship between score and rating is low negative (r= -0.46) and the rating variable explains about 21% of the score variable (r^2 = 0.21). the linear relationship between score and gross is low negative (r= -0.35) and the gross variable explains about 12% of the score variable (r^2 = 0.12). The linear relationship between score and runtime is low positive (r= 0.46) and the runtime variable explains about 21% of the score variable (r^2 = 0.21).And the linear relationship between score and votes is negligible (r= -0.54) and the votes variable explains about 2.9% of the score variable (r^2 = 0.0289).

Besides that, I performed Pearson's Correlation tests between the score variable and other variables. I conclude that the relationship between the score variable and other variables are likely to be dependent, and those correlations are unlikely to have occurred by chance.

Next, I built a binary logistic regression to predict whether a movie will be popular or not. For me, I considered a movie is popular if it is scored well. So I set the score variable as the target variable of the model. To do this, I encoded the score variable as one of these two numbers, 0 or 1. The number 0 means an unpopular movie, and the number 1 means a popular movie. Since the average score is 6.3, I considered any movie with a score higher than 6.3 a popular movie, and any movie with a score of less than 6.4 unpopular movies.

According to the predictors' coefficients, the predictors' budget, gross, and votes have a small effect in predicting the movie's popularity. However, the variables' Genre, Rating, and runtime are considered good predictors in determining movie popularity.

Finally, My project data were based on user ratings. We also know that users submit their ratings online. So the data doesn't include opinions of audiences who don't go to the website to vote.