Description of the final project

**Introduction**

In this project, you will be asked to help a film-producing company. They would like to make sure that films, which they are creating are successful. The company has collected a large dataset, which includes information about more than 80'000 films: their description, genre, budget, rating, etc.

Your task will be to suggest and justify a way to measure the success of the film and build a machine learning model, which is capable of predicting this quantity.

**Detailed Task Description**

The solution should have the following parts:

[20 %] Formalise the task and suggest the evaluation criterion

* What is your target variable? Why is it a good measure of success and how the film-producing company can use it to make their decision?
* Which task are you going to solve (is it a regression, classification or something else)?
* Which metric will you use to evaluate the performance of the model? Do you have any idea, which values of the metrics can be considered good enough?

[40 %] Explore and preprocess the data

* Split data into train and test
* Fill missing values (if there are any)
* Remove irrelevant feature (if there are any) and create new ones (explore which features you have, what is their distribution and connection to the target, explain why newly created features are reasonable)
* Properly encode categorical features (if required)

[40 %] Train the models and present result of your work

* Use the models introduced during the course, which are relevant for the task you are solving
* You are supposed to compare at least 3 models
* Tune (or explain why there is no need to tune) hyperparameters of each model
* Use cross-validation to compare models and select hyperparameters
* Which model and setup is the best? How does it perform on the test dataset?

General Requirements

* You should provide a solution in the form of the Jupyter Notebook (use the template that we provide)
* Make sure that the instructor can run all the cells to reproduce your results
* Clearly answer each question / perform required actions from the task
* Support all your statements with calculations and plots.

For example: If you claim that the dataset does not have missing values, you need to provide code which shows that the total number of missing values is zero