Part 1: Data

This dataset is a random sample of movies that includes information from two popular movie database and rating sites Rotten Tomatoes and IMDB.

*Generalizability*:

We do not have information regarding the sampling method used. For this exercise, we assume it is a valid random sample.

*Causality:*

On the same note, there is no indication that random assignment has been performed in drawing this sample. Therefore, we cannot draw/infer conclusions on causality. At most, we can generalize on our findings.

*Bias and reservations:*

Considering many successful movies tend to generate sequels, we acknowledge potential bias on sequels rating from original movie rating. This requires taking into account temporal data that it is out of scope for this research topic.

For the project at hand, I assume the observations in the sample dataset are independent. However, some variables in the dataset could be dependent of each other. I will analyse the variable collinearity to identify such potential dependencies and remove them from the model.

Part 2: Research question

My research question is to analyse if it’s possible to predict the IMDB rating score based on the variables in the sample dataset (genre, ratings, actors, length, etc.). Such analysis could potentially aid movie theaters what movies to show and promote, and assign viewing rooms ahead of the movie release.

Part 3: EDA

Since I will be looking to fit a multiple linear regression model the exploratory data analysis will address the research question in context of validating the conditions required to develop a linear regression model.

Conditions for linear regression:

1. Linearity
2. Nearly normal residuals
3. Constant variability of residuals
4. Independence of residuals

Modeling

Inference for linear regression:

* Significant Predictors
* HT and CI for the slope
* Conditions for inference| Interpretations

Variability partitioning (ANOVA)?

Model Selection:

* Stepwise models selection
* P-value & adjusted R2
* Expert opinion?

Model Diagnostic

* Linear relationships ship between x and y
* Nearly normal residuals
* Constant variability of residuals
* Independence of residuals

This dataset is a random sample of movies that includes information from two popular movie database and rating sites Rotten Tomatoes and IMDB. \n

\*\*Generalizability:\*\*

We do not have information regarding the sampling method used. For this exercise, we assume it is a valid random sample. \n

\*\*Causality:\*\*

On the same note, there is no indication that random assignment has been performed in drawing this sample. Therefore, we cannot draw/infer conclusions on causality. At most, we can generalize on our findings.\n

\*\*Bias and reservations:\*\*

Considering many successful movies tend to generate sequels, we acknowledge potential bias on sequels rating from original movie rating. This requires taking into account temporal data that it is out of scope for this research topic.\n

For the project at hand, I assume the observations in the sample dataset are independent. However, some variables in the dataset could be dependent of each other. I will analyse the variable collinearity to identify such potential dependencies and remove them from the model.\n

I would like to investigate if a movie IMDB rating score can be predicted based on the variables in the sample dataset (genre, ratings, actors, length, etc.). Such analysis can aid movie theaters make better decisions regarding the movies they show and promote, or assign viewing rooms ahead of the movie release.

Since I will be looking to fit a multiple linear regression model, the exploratory data analysis will address the research question in context of validating the conditions required to develop a linear regression model.\n

The data set is comprised of 651 randomly sampled movies produced and released before 2016.