# NOTES DELETE

EDA

Explore data (groupby money and category) (groupby money and language) (plot budget vs revenue) (create derived column of revenue\_budget\_ratio) (plot categorye by revenue\_budget\_ratio)

Convert multilabel data to one-hot-encoding. Like Language.

Use for, while and if statements for transforming data.

Trend of release dates. Extract year and month.

Two features plots (budget vs. revenue, popularity vs. vote avarage) use .hist and the sns plots.

Any observable clusters in the structured data?

Stemming and Lemmatized columns of the title+overview

Classify a new movie description.

Perform similarity search.

Predict popularity and revenue given a movie description.

Concepts to represent, variables, for, while, if/else statments, conditionals, defining functions, .at, .loc, .iloc, dataframe slicing, plotting.

Notes on Classes and Classificaion: In some cases there may be a movie that was miss-classified. Depending on how the classification was done (e.g., manually) there may be some biases towards a class in some cases. Classification is not an exact science and there may be some overlap and in cases a persom making the classification decision may say this movie is adventure while another may say is action or in some cases the both classes maybe applied. In some cases there were no genre applied. This may cause challenges if trying to create an algorithm to classify. Perhaps a null class is also desirable and and SME may be able to provide input. Another note which may be applicable to genres is that the genres may change overtime, for example, a genre that exist today may not have existed 50 years ago. If trending genres by time it may also be affected. For example, there may be a genre that was applied starting in the 1950's but movies prior to 1950's could also be applicable to that genre but the genre was not applied. It is important to understand the history of a database and its features. Databases and its features can be revamped overtime. Could also trend the genres overtime to explore if this may be an issue.