Cole Jennings
DATA 400 Mini-Project Idea Submission

Data tractability:
Find out what factors are most important in movies receiving a higher rating (on TMDB). Exploring features such as budget, genre, cast, revenue, etc.

Data retrieval:
I will use three datasets I found on Kaggle. I will merge them using common keys.
https://www.kaggle.com/datasets/aayushsoni4/tmdb-5000-movie-dataset-with-ratings?resource=download

Exploratory data analysis:
After removing unnecessary columns and dropping null values the shape of my dataset is (17258147, 25), which is subject to change if I decide to drop additional columns. I will also use a EDA package such as SweetViz or Pandas Profiling to look at the distributions of different variables. Given what is shown on Kaggle, "rating" which will likely be my variable of interest looks slightly left-skewed.

Implications for stakeholders:
Anyone involved in the movie making process will be interested to hear how different variables associate with higher or lower user ratings. Movie watchers might be interested to hear what is associated with them rating movies higher or lower as well.

Ethical, legal, societal implications:
This type of data will alter the movie-making process and might have societal implications in that studios will begin to create similar films that follow the output of the model.