DAT Sample Modeling Challenge

Welcome to this week's bonus assignment! If you choose, you can spend time this week cutting your teeth on a sample modeling challenge.

Your Objective: With Linear Regression as your hammer, see if you can transform your dataset into a nail that it can pound to accurately predict the amount of money a particular movie will be able to gross.

How Your Model Will Be Evaluated: At the start of class on December 2nd, I'm going to run a script that grades everyone's final submissions live in class, to determine who the winner is.

Key Points to Pay Attention To:

- See if you can find a coherent way to extract information out of the many categorical columns in this dataset. You can't work with all of these, but pick at least one that looks meaningful to you and see if you can do something to make it meaningful.
- Start with a simple model and iteratively add to that, as opposed to starting with something very complicated
- There are empty values…..what are the most likely causes for these? Should we interpret them as being randomly generated, or do they mean something specific? Are they specific to particular periods of time?
- There is a date column. Can you extract information about what season it was released? If it was done on a Friday? Wednesday? How long was its opening weekend?
- Things like popularity and budget have meaningful correlations to final revenue, but what about movies that have higher or lower budgets compared to their categorical average?
- Maybe dramas tend to be rated more highly than other films, so columns that compare stats about a film to its category average are often useful ways to extract information out of different features.
- How do different validation strategies compare to one another?

Again…….you probably won't have enough time to go through **all of these,** but see if you can find at least one way to meaningfully improve your score.