CSE 351 - Introduction to Data Science (Summer 2021)
Course Projects
Deadline: June 30, Wednesday, 11:59 PM EST, 2021.

Hey Jeff & Zhe -

I just realized that **popularity**, from a practical standpoint, might not be a great variable to use to *predict* revenue, because one can only measure the popularity of a movie AFTER its creation and release. On the other hand, **budget, runtime, genre, release day**, etc are better in that regard because companies are in control of this before the making and release of a movie. Let me know what you guys think.

Oliver

Baseline Models:

1) Simple linear regression. Using budget to linearly predict revenue. This one is preferred.

$$projected\,revenue \;=\; a\,(budget)\;+\;b$$

2) Mean/Median - Makes it much easier to come up with models better than this, so I find this one more and more appealing

$$projected\,revenue \;=\; median\,revenue\,of\,all\,movies$$

Advanced Models:

1) Linear regression after applying a nonlinear function to budget (or another variable)

$$projected\,revenue \;=\; a\,(budget')\;+\;b$$
$$where\,budget' \;=\; budget^{2}, budget^{3}, \sqrt{budget},\; etc$$

2) Groupby genre, release day of the week, movie language, etc, and use simple linear regression from our baseline model number 1. Might be a few challenges in regards to implementation of our testing, but we'll figure that out. Groupby genres is more complex because as many movies have more than one genre, we'll have to somehow combine the values from each genre, by computing the avg, or picking the highest value of the genres, etc. We can do a few of those.

3) Combining numbers 1 and 2. Should be easy if we already did 1 and 2.

4) Use machine learning to create a model. Uncertain about how much work that's going to take