Abstract: The goal of this project was to model and interpret profit vs a range of features for the top 1000 grossing movies that never made it to #1 on the weekend box office. Virtually every weekend box office #1 movie is cut from the same cloth. Half(!) of the top 20 highest domestic grossing movies are either from Star Wars or The Avengers alone. This project is for the dreamers out there who might have a movie script in mind that's not a Super Hero trilogy, who might benefit from learning about trends that the most successful, but not mega-blockbuster movies had.

Design: Data was scraped from **boxofficemojo.com** using BeautifulSoup. A Baseline Linear Regression was performed and then various methods of feature engineering, feature selection, cross-validation and model selection were employed.

Algorithms: Simple Linear Regression, Multiple Linear Regression, Exploratory Data Analysis.

Feature engineering including polynomial features and interaction effects. Simple Validation and testing, and 5-Fold and 10-Fold Cross-Validation techniques.

Feature Selection and Regularization techniques used included Forward Stepwise Selection, Lasso and Ridge Regression.

Tools: Pandas, Numpy, BeautifulSoup, Sci-kit Learn, Statsmodels, Matplotlib, Seaborn.

Results and Communication: Results were communicated in a 5 minute presentation to the Metis online DS bootcamp.

I found this to be a very well designed Data Science project by Metis, at least in terms of helping me to become a better Data Scientist.

So while I do not feel confident in hardly any of my modeling or inferential results, I learned a lot and got to explore a ton of Data Science tools and got my first experience with Statistical Learning.