DSC630 Milestone 2 – Erez Sarousi

**Introduction:**

**Background** - Hollywood is known to be a high-yielding and profitable business, worth over forty-two billion dollars as of 2019 (Watson, 2020). Just like any other business, their goal is to increase their profits by constantly producing high-quality content. Hollywood employs just under a million people (Sakoui, 2019), and roughly 13,500 people are actors in the United States (Gleeson, 2018). The financial side of the business is constantly poring through the data and are interested in producing more revenue while limiting their expenditures, resulting in heightened profits.

**Problem Statement** – Based on prior data, it’s easy to see how prior films fared, but is it possible to determine how they will do in the future? Using predictive analytics, this project seeks to discover how well future films will do based on films from the past and will extrapolate based on those figures to determine how well future films will perform.

**Scope** – Using the power of predictive analytics, Team Decibel will be predicting how to make the best film through thorough understanding of different factors. This research pores through features such as genre, performers, budget and more to learn what contributes to the success of a film and to which degree.

**Document Overview** – The information gathered here originates from *The Movie Database* (TMDB), which contains nearly 5,000 movies, each comprising a row in the dataset. There are also several columns, which appear to be useful in the analyses conducted in this project. Some of these columns include budget, genre, original language, and production companies.

**Preliminary Requirement:**

**Technical Approach** – This project is completed using two main coding programs. The first is by R, a free statistical coding language used for statistical analysis. The IDE (Integrated Developing Environment) used will primarily be Jupyter Notebook combined with the R kernel. R is planned to primarily be used in terms of data exploration and understanding the different variables planned.

The second part of the project, including data analysis, and data model execution will be done by Python, a free computer programming language known for its high readability. Like R, the Python portion of the project will be planned through the Python kernel of Jupyter Notebook.

**Data Sources or Plan for Data** - In order to proceed, it will be necessary to clean the data, such as removing unused columns, and properly dealing with the vagueness of some of the columns such as genres. It will be necessary to perform feature reduction, feature selection, and hyperparameter testing. In order to implement the predictive analytics, regression will be conducted; due to the substantial risk of multicollinearity, ridge regression will likely be pursued.

References:

Gleeson, P. P. D. (2018, June 27). Statistics on People Getting Famous in Acting. Work - Chron.Com. <https://work.chron.com/statistics-people-getting-famous-acting-23946.html>

Sakoui, A. (2019, March 18). Hollywood Employs More Workers Than Mining and Farming, MPAA Says. Bloomberg. https://www.bloomberg.com/news/articles/2019-03-18/hollywood-tops-mining-crop-production-in-employment-mpaa-says

Watson, A. (2020, November 10). Film Industry - statistics & facts. Statista. https://www.statista.com/topics/964/film/