**Project Proposal**

A group proposal will be due 11:59pm Tuesday March 14. A single proposal should be submitted by the group. The proposal should be submitted via the course website. The proposal should include: 1. The names of the group members 2. Title of the project 3. Questions the group intends to answer with its analyses 4. Basic description of data including data source, number of variables in the data sets, variable descriptions (if there are many variables, rough description of the type of information in the data set and description of variables of most interest is fine), and number of data points 5. Description of the analysis each member intends to do and what technique(s) they intend to use; each group member is expected to do exploratory/descriptive and diagnostic analysis for their portion of the project in addition to the predictive analysis (e.g. linear regression, logistic regression, specific type(s) of generalized linear model, ANOVA, PCA, cluster analysis, discriminant analysis, etc.). Group members may either use different methods to answer related questions about the data, or answer different questions of interest for different aspects of the data.

**Group Members**

**Title of Project**

Analysis of Movies

**Questions Intended to Answer**

We will be looking to answer many questions involving movie statistics, such as: Which star actors and star directors are producing the highest gross revenue for their movies? Which genres and content ratings (ie. pg, pg13, r) result in the highest rated movies? What variables best predict the gross revenue for the movie? (production cost, star actor, star director, color vs. black and white…) Do American/english speaking movies generate a higher gross revenue than movies produced in other nations? Does the length of the movie predict the ratings that the movie will receive? And finally, how to Facebook likes, if at all, correlate with gross revenue and ratings for movies?

**Data Set Description**

This movie data has been scraped off of IMDb.com, and published on Kaggle.com as a downloadable csv file. It includes data on a little over five thousand movies and has 28 variables for each one. Categorical variables include whether or not the movie was in color, the director, the 3 lead actors, genre, content rating, language, and country of origin. Numerical variables include duration, gross revenue, number of cast members, production budget, imdb rating score, and Facebook likes, among others.

**Each group member’s description of analysis**

**Member one:** I intend to look at many of the top directors and star actors and analyze whether or not there are differences in the true means of their gross revenue. The null hypothesis is that each actor and each director has the same true mean while the alternative hypothesis is that there are difference in the gross revenue true means among different star actors and directors. I will need to test for normality in the data, and then either use a t test or nparway1 test along with also looking at the confidence limits.

**Member two :** linear regression between budget and imdb\_score and logistic regression between countries both budget and imdb\_score

**Member three :** PCA - actor/director/movie facebook likes

**Member four(me)：gross**