**STAT 440 Project Proposal**

**Group Members**

…

…

…

**Title of Project**

Entertainment Yearly

**Data Set Description**

The first dataset contains the measurements taken at the NFL combine for all players drafted into the NFL from 1999-2015. It includes data on a little below five thousand observations and has 26 variables for each one. Categorical variables include position, college. Numerical variables include heightfeet, heightinches, heightinchestotal, weight, arms, hands, fortyyd, twentyyd, tenyd, twentyss, threecone, vertical, broad, bench, round, pick, pickround, pick total, wonderlic, nflgrade.

Variables are mostly player measurements and some variables which need additional explanation are as follows:

* fortyyd: The 40-yard dash is how long an athlete takes to sprint 40m. These athletes are timed at 10, 20 and 40-yard intervals.
* Twentyss: The short shuttle tests the athlete's lateral quickness and explosion in short areas. The athlete starts in the three-point stance, explodes out 5 yards to his right, touches the line, goes back 10 yards to his left, left hand touches the line, pivot, and he turns 5 more yards and finishes.
* ThreeCone: The 3 cone drill tests an athlete's ability to change directions at a high speed. Three cones in an L-shape. He starts from the starting line, goes 5 yards to the first cone and back. Then, he turns, runs around the second cone, runs a weave around the third cone, which is the high point of the L, changes directions, comes back around that second cone and finishes.
* Vertical: The athlete stands flat-footed and they measure his reach. It is important to accurately measure the reach, because the differential between the reach and the flag the athlete touches is his vertical jump measurement.
* Broad: The athlete starts out with a stance balanced and then he explodes out as far as he can. It tests explosion and balance, because he has to land without moving.
* Bench: The bench press is a test of strength -- 225 pounds, as many reps as the athlete can get.
* Round: (out of 7) Round drafted, first being the highest, 0 being player was not drafted
* Pick: # selected within round (1 being highest)
* Picktotal: # selected overall (disregarding round, 1 being highest)
* wonderlic: Popular group intelligence test used to assess the [aptitude](https://en.wikipedia.org/wiki/Aptitude) of prospective employees for learning and problem-solving in a range of occupations
* nflgrade: Rating of player’s skill (10 being the most skilled)

The second dataset relates to data from imdb.com and can be found on Kaggle.com. It includes data on about 5000+ movies and has 28 variables for each movie. A movie’s success is related to many factors, this set of data mainly analysis the film’s quality from the independent variables like actor’s popularity, director’s acclaim, genres, investment to the dependent variable like IMDB score and gross. Variables are listing as following：

* Categorical variables include director, 3 lead actors, genres, movie title, plot keywords, content rating, language, and country of origin.
* Numerical variables include gross, budget, number of voters, duration, year, number of faces in movie poster, number of critics, imdb rating, aspect ratio, and total Facebook likes for the cast, among others.

**Background Information**

The first data set can be directly downloaded from <http://www.nflsavant.com/dump/combine.csv?year=2015>

Explanation of variable nflgrade: <http://www.patriots.com/news/2016/02/23/nflcoms-grades-top-prospects-nfl-combine>

The second data set can be found at <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

However, you need a login to continue; we have uploaded the data to compass to your convenience.

**Research Interests**

For the NFL combine data set, we can subset the players based on character features, then find the descriptive statistics for the various subsets, such as drill performance and position. This could potentially help coaches with choosing between different players who play the same position for future drafts based on past performances.

For the IMDb data set, there are many factors which can potentially influence popularity before movies are released into theatres. Any conclusions drawn from the differences between our subgroups can be utilized by businesses to determine which advertising factors would be most influential on a movie’s success before public access.

**Data Preparation Techniques**

In the NFL combine data, there are some instances where players who posses a suffix in their name are incorrectly read into the csv file, which need to be addressed. Also, for numeric variables which have specific ranges for values, proc freq should be used to confirm there are no erroneous entries. The variable heightinchestotal should be correctly calculated from the sum of heightfeet (converted) and heightinches. NFLgrade can be formatted to explain what each grade means. The data can also be subsetted according to team, total height, and position.

As for the IMDb data, we can subset movies in various ways, based on several potential groups of interest like genre, country, content rating, and year(s). Also, it is possible for an actor to appear in multiple different movies, so we should check to see if this factor is of importance.