ISYE 7406 Project Tasks

1. **Create full dataset**
   1. Gather any additional data on coaches and travel distance
   2. Combine original dataset with additional data
   3. Transform round eliminated column into number of games won
2. **Data Exploration**
   1. Explain the data in detail. Paragraph(s) about what each metric means and what is its intuition in basketball.
      1. Make sure to explain that we dropped the play in games.
      2. Do we treat each year separately or can we pool the entire dataset and consider each year to have identical properties?
   2. How much data do we have?
   3. What are the relationships between the variables (scatterplots, boxplots, QQplots, etc.)
   4. Are there outliers? How should we deal with outliers?
3. **Methods**
   1. Manipulate the data into a frame and/or matrix form useful for the prediction algorithms.
   2. Start with methods that can predict continuous values (Linear/Lasso/Ridge Regression, Random Forest/Bagging, etc.) and build a function for each method that can be passed into a cross-validation function.
   3. **Cross validation to tune model parameters and decide on the best modeling approach**
      1. Split data into k-folds for training and validation
      2. Input:
         1. Training Data
         2. Testing Data
         3. Algorithm function developed in part 3.b
      3. Output:
         1. MSE/AIC/BIC (choose one) of each model in a dataframe
         2. Optimal Parameters
      4. Use the following libraries to speed up for loop cross validation using multiple cores:
         1. Foreach
         2. doParallel
         3. parallel
         4. <https://medium.com/@najari.vahab/parallel-computing-in-r-and-python-1c916a2803d7>
4. Results
   1. Show performance improvement over parameter tuning for each model
   2. Display dataframe comparing the performance of each model.
   3. Explain why the final model was chosen
5. Conclusion
   1. What was the final model?
   2. How accurate is it trained on the entire dataset?
   3. How well did it predict the actual results of the 201X tournament(s)?
   4. What else can be done to improve this analysis?
   5. Where can the analysis go in the future?

Paper:

Abstract should say that we are attempting to predict how far a given team will go in a March Madness Tournament based on team metrics over the past year. Using XYZ modeling techniques, we found that we were able to predict team success with %%% accuracy. With this level of individual team accuracy, we were able to successfully predict the outcome of X/Y ratio of games in the 201X tournament(s). 🡨This sentence is important because we need to relate our results back to something meaningful in the real world as an application beyond doing machine learning well.