Michael Rocchio

MSDS 411 Project Proposal

1. I will be utilizing the Moneyball data set as I have had experience with it in previous courses I have taken at Northwestern. I also need to get around to watching the movie. The dataset is highly aggregated as it is at the team level. It has a Boolean operation that signifies if the team went to the playoffs as well as the rank for the season and playoffs. There are also the main predictors: Games Played, OBP, SLG, & BA.
2. I will, unlike they did in Moneyball, use every type of unsupervised learning method taught in this class to determine seasonal performance of each baseball club. I would also like to do a PCA analysis to determine the weights of various metrics used to assess a teams performance.
3. I am going to initially focus on a clustering analysis to determine club ranking by segmenting the predictor variables, ranking, to allow a classification. I also need to worry about how I am going to deal with the different leagues. Perhaps the movie will have some tips.
4. **Questions:** First, is it alright if we use Python? Also, I am concerned about the lack of the dimensionality of the data and may look for more player by player data, maybe I could compare the movie's analysis with mine once if I can find player data.