**Capstone Project Proposal**

**Topic**: Analyzing the effect of batting order on baseball games

**Problem:** Managers often place their best hitters in the 3-5 spots, with the worst hitters in the 8th position in the lineup. However, the order players come up to bat is not the same for every inning; only the 1st inning is guaranteed to have batters come up in the original 1-9 order. Do innings where the listed leadoff hitter actually leads off contribute to more runs and a better chance at winning? Is it possible to predict the number of runs scored in an inning based on features such as batting position leading off, inning, etc.?

**Hypothetical client(s)**: Baseball teams and managers would clearly be interested in finding out what influences run-scoring. Different strategies can be developed from data that shows whether specific methods of setting batting order would increase runs scored, thus increasing chances of winning baseball games. Perhaps “casual” fans could determine which innings are more likely to feature scoring depending on what kind of hitter is leading off, and thus pay more attention in innings with higher scoring-chance.

**Data**: Play-by-play data for games is needed in order to find out who was leading off in games for each inning, and how many runs, hits, etc. occurred in each inning of a game. This data is available through Retrosheet.org. It is also possible that specific player or team data may be useful; the Lahman database could provide some season statistics. Baseball-reference has easy-to-access data, but there doesn’t appear to be a free method to obtain the data as a download for processing in Python or R.

**Approach**: First, a data story could be created by simply looking through the data available. The game files provided by Retrosheet contain loads of different potential features for data science algorithms to use. Models can be created to predict whether a team will win, predict number of runs scored in a particular inning, etc. Various regression models such as logistic and linear regression could be employed for these models, either classifying wins/losses and/or predicting number of runs, hits, etc.

**Deliverables**: Deliverables for the project include the code used to create the models and clean the data, a report paper on the results, and a slide deck with useful plots and graphs that summarize the data for those not as familiar with analytics. Links can be provided to sources that were used to obtain data and help clean the data, allowing the client to further continue exploration of the data on their own if they desire.