Analysis and Predictors of NHL Playoff Hockey Success

Sports has always been a hobby and interest of mine. They are entertaining and provide a great playground for small data analysis projects. The National Hockey League is the highest level of hockey played in the world, and the league is notorious for exciting and random playoff matches. 16 of 30 teams make the playoffs, and teams play a best-of-7 series against one another before advancing to the next round. To showcase the randomness of these best-of-7 series, in 2019 the Tampa Bay Lightning tied an NHL record 62 wins during the regular season. In the first round of the playoffs, they were swept (lost 4 games in a row) to the worst team to make the playoffs in their conference. This result is not an anomaly, as all 4 division winners in 2019 during the regular season lost in the first round of the playoffs.

I want to look at regular season data from these hockey teams to try and find predictors of playoff success, because it is clear we cannot look at win-loss record alone. If we can find a statistic (ex: shots per game) that translates to positive (or negative) playoff success, this means teams, gamblers, and fans can have a better idea of what teams are more likely to succeed in playoff contention. This could even impact team construction and coaching decisions. I hope to learn whether this inherent randomness has always existed in the league, or whether this is a recent phenomenon. We can also begin to hypothesize why certain aspects of regular season play do not translate to playoff performance.

I found a dataset on Kaggle called Professional Hockey Database. It contains a collection of historical statistics from men's professional hockey leagues, from 1909 – 2011. It includes files containing data on teams, players, goalies, playoff series matchups, and other useful information. They are all csv files, and I have loaded them successfully into RStudio.

I plan to judge playoff performance on the basis that they win a playoff best-of-7 series. To measure this, I plan to use logistic regression to determine a team's probability of winning the series (winning the series corresponds to 1, losing to 0). I hope to train the logistic regression model with the actual historical playoff results, in hopes of finding the regressors that can best explain the actual results. Using t-tests and F-tests I will determine which variables have the largest effect on playoff performance. Evaluating the outcome will be determined by comparing my model with a trivial model that includes only one regressor, points acquired during the regular season. This is trivial model because this is simply picking the 'better' team according to their win-loss record in each series. I am unsure of other dimensions of evaluating the outcome, and I am open to suggestions.

Over spring break (3/23-3/27) I hope to Tidy the data and begin initial exploratory data analysis. From 3/30 – 4/3, I hope to continue exploring the data through transformation and visualization and begin modeling. From 4/6 – 4/10, I hope to refine the model and variable selection and begin writing the report. From 4/13 – 4/24 I hope to complete the modeling and report process.