SI618 Project 2 Report

Alexandra Elias

December 10, 2019

Motivation

My father was a football player, my mother did cheer, my brother played baseball, and I played soccer. Being raised in an athletic family, I have always had a passion for sports. On the other hand, I am from Phoenix, Arizona where hockey is not the first sport someone considers playing. Prior to moving to the Midwest, I was not all that familiar with hockey but I have since become a fan. It is fast-paced, intense, and difficult to analyze according to the sports industry. Therefore, I used this two-part project as a platform to become more familiar with hockey and the National Hockey League (NHL) as it relates to sports analytics. In part one, my goal was to find and clean a useable dataset and then produce initial sanity checks to verify the accuracy of the data. I did so by analyzing the actions of individual hockey players and comparing them to the NHL database. In part two, my goal is to produce some exploratory data analyses that lead to team-level insights. More specifically, I want to answer the following questions:

1. **What is the distribution of total goals per player per season?**
2. **What does a winning team look like according to hockey metrics?**
3. **Which features, if any, have a linear relationship with wins?**

Data Source

On Kaggle (https://www.kaggle.com/martinellis/nhl-game-data), I found a directory of files that contain the official NHL metrics measured for each game from the previous nine seasons. It was last updated in June of 2019. From this directory, I used five datasets that were provided as csv files: player_info.csv, team_info.csv, game.csv, player_stats.csv, and game_teams_stats.csv. The player information file had 2409 observations and contained categorical variables like player name, nationality, and position. The team information file had 33 observations and contained categorical variables like team name and its abbreviation. The game file had 11434 observations and contained categorical variables like season, away team, home team, and outcome. The player statistics file had 411578 observations and contained aggregated data of quantitative player metrics like goals, shots, assists, hits, face off wins, giveaways and takeaways, for each game. The team statistics file had 14882 observations and contained aggregated data of quantitative team metrics like goals, shots, hits, and outcome, for each game.

Methods

Question 1: What is the distribution of total goals per player per season?

I used three joins to prepare a data frame for this analysis. I used a left join, to combine the player statistics table with the game information table, an additional left join to add the player information table to the previous join, and a final left join to add the team information table to the previous joins. I dropped 14 of the columns that were irrelevant to this question. Because I used this data set for part one of the project, the issue of the missing and incomplete data was previously rectified. There were no challenges faced. The data was prepared for the analysis.

Question 2: What does a winning team look like according to hockey metrics?

Similarly, to question one, I used the same joined dataset to answer question two. To reiterate, I used three joins to prepare a data frame for this analysis. I used a left join, to combine the player statistics table with the game information table, an additional left join to add the player information table to the previous join, and a final left join to add the team information table to the previous joins. I dropped 14 of the columns that were irrelevant to this question. Because I used this dataset for part one of the project, the issue of missing and incomplete data was previously rectified. There were no challenges faced. The data was prepared for the analysis.

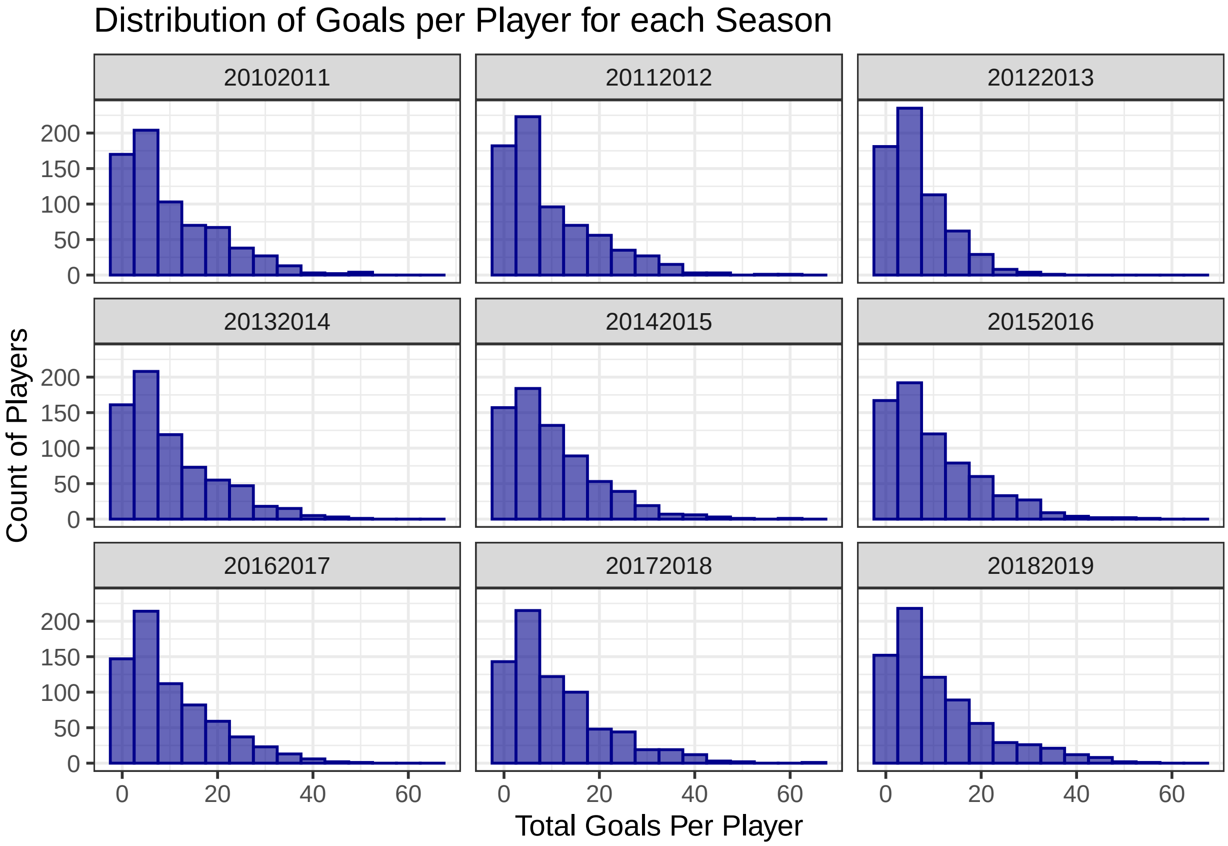
Question 3: Which features, if any, have a linear relationship with wins?

To answer question three, I used a left join to combine the team statistics table with the team information table. There were a handful of missing rows so I used a function to remove those rows from the dataset. After removing them, I still had over ninety percent of the data so I could continue the analysis. One challenge that I faced was that the team statistics file was new to me since I did not use it in part one of the project. In part one of the project, I focused on individual players in hockey but the goal of part two was to discover team-level insights. Therefore, this file was necessary. Because the file was not previously used, I had to take the additional time to understand it. I learned that there were two rows for each game: one row for the home team and one row for the away team. I observed that the dataset contained aggregated hockey metrics per team per game. Additionally, I cross referenced the data with the online NHL database to verify the accuracy. Other than that, the file was relatively clean and easy to work with.

Analysis and Results

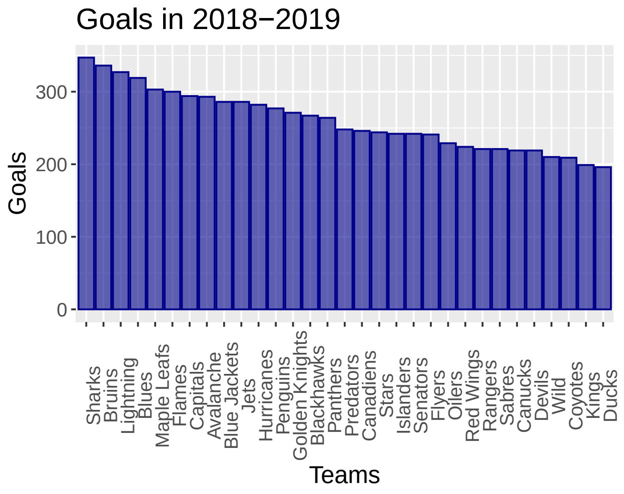
Question 1: What is the distribution of total goals per player per season?

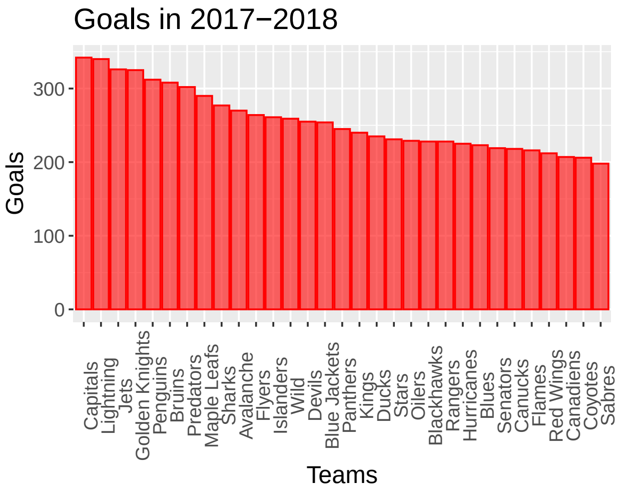
The purpose of this question was to see if there were seasonal differences for one of the most important hockey metrics, goals. If there were significant differences between seasons, I would have taken more of a time series approach for the remainder of this project. To answer this question, I grouped the data frame by player and season and created a column called goals that computed the total goals per player per season. Then I filtered out the players that did not score any goals in a season. To visualize this analysis, I created one histogram for each of the nine seasons. I found that the distributions of total goals per player is similar regardless of the season. Overall, the shape is right skewed as I expected. Additionally, the majority of the players score between one and twenty goals in a season and a few players are outliers scoring 50 or more goals in a season.

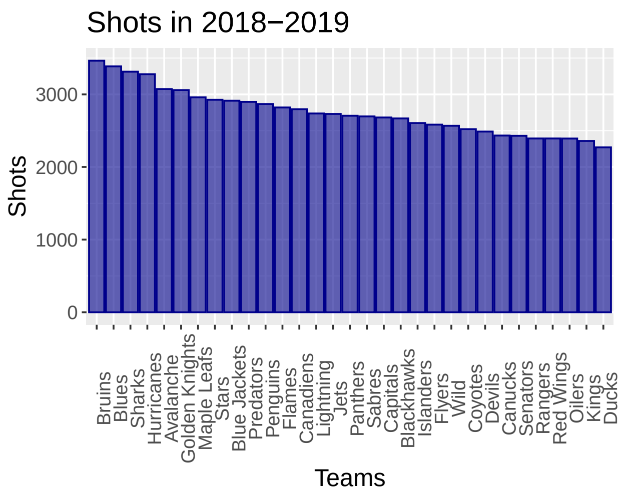


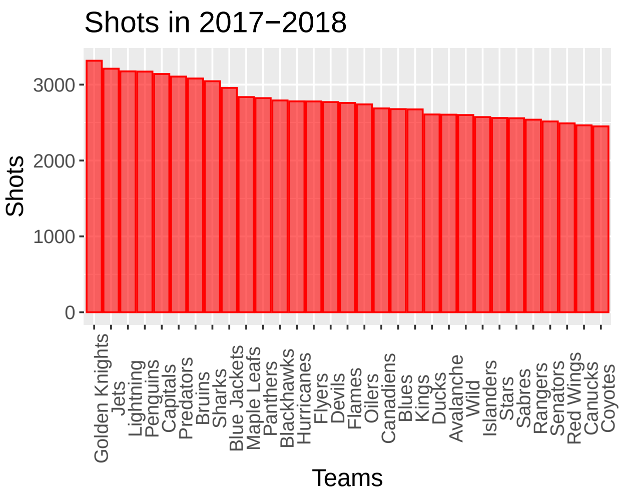
Question 2: What does a winning team look like according to hockey metrics?

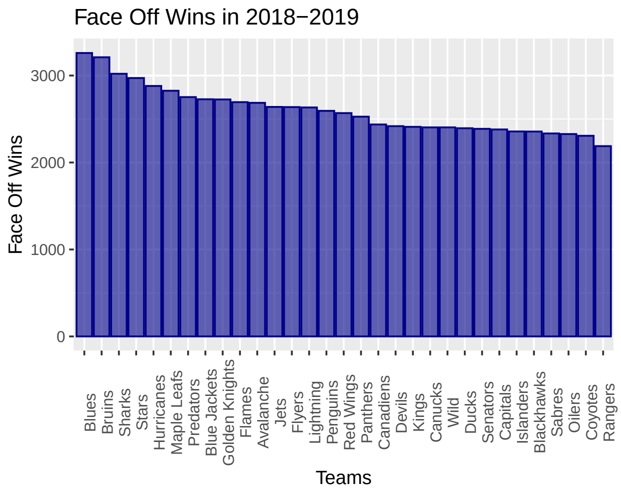
The purpose of this question was to reflect on previous Stanley Cup winners and observe what their aggregated totals for goals, shots, face off wins, takeaways, and giveaways looked like. This is an exploratory analysis to better understand which features of the dataset might be best utilized in predictive models (say, in a future project). To answer this question, I selected relevant variables from the data frame and grouped them by team name and season. Then I computed totals for the following variables: goals, shots, face off wins, takeaways, and giveaways. I filtered the data to focus on the two most recent seasons: 2018-2019 and 2017-2018. In 2018-2019, the Stanley Cup winner was the St. Louis Blues and the runner-up was the Boston Bruins. In 2017-2018, the Stanley Cup winner was the Washington Capitals and the runner-up was the Vegas Golden Knights. To visualize this analysis, I created an ordered bar chart for each metric for the two seasons. Looking at goals, the Bruins and Blues ranked second and fourth, respectively, during their winning season and the Capitals and Lightning ranked first and second, respectively, during their winning season. Similar observations can be seen for goals, shots, face off wins, and takeaways. In summary, the two teams that played in the Stanley Cup that season can also be seen as one of the top six teams for each of these metrics. The feature with a noteable difference in pattern is giveaways. Surprisingly, the teams that played in the Stanley Cup had more giveaways compared to the other teams. Prior to conducting the analysis, I expected to observe the contrary. One might infer that the best teams take risks, which makes them good but also makes them vulnerable to giveaways. On the other hand, it is possible that the best teams take risks but trust that their defensive lines and goalies can recover from giveaways.

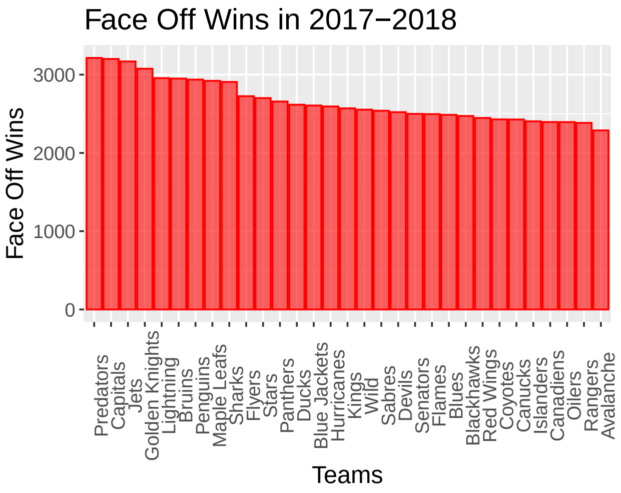


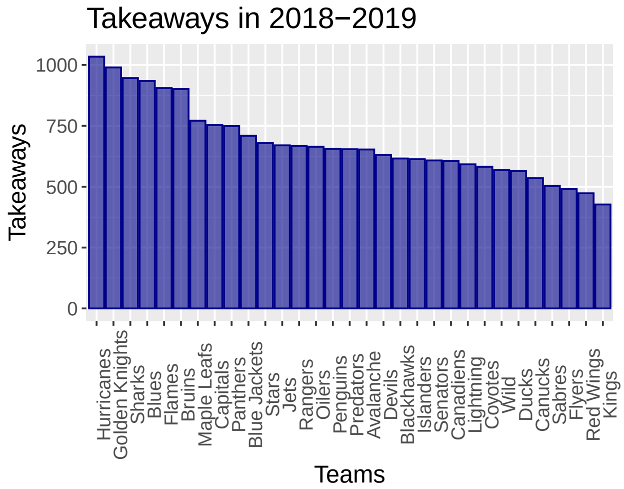


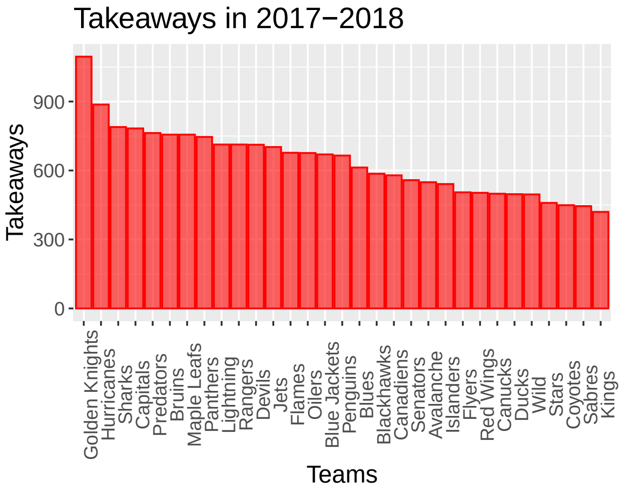


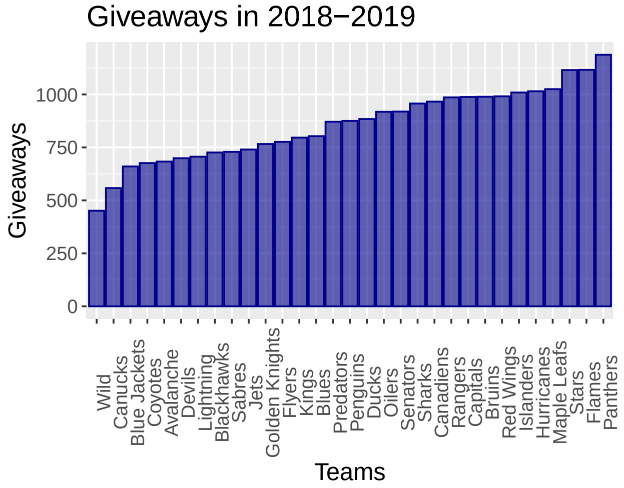


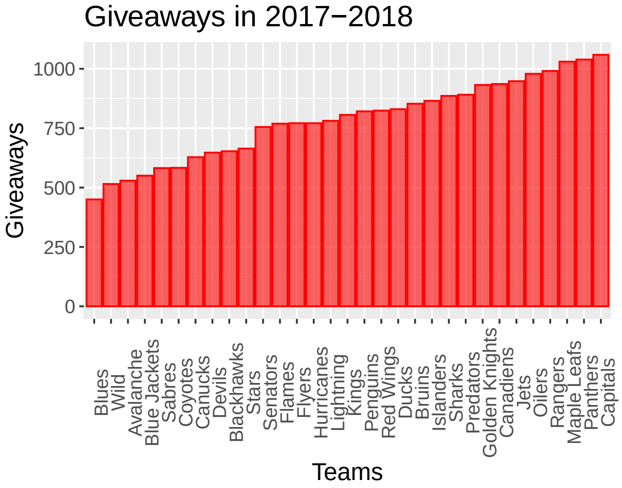

Face Off Wins in 2018−2019


Face Off Wins in 2017−2018


Takeaways in 2018−2019


Takeaways in 2017−2018


Giveaways in 2018−2019


Giveaways in 2017−2018

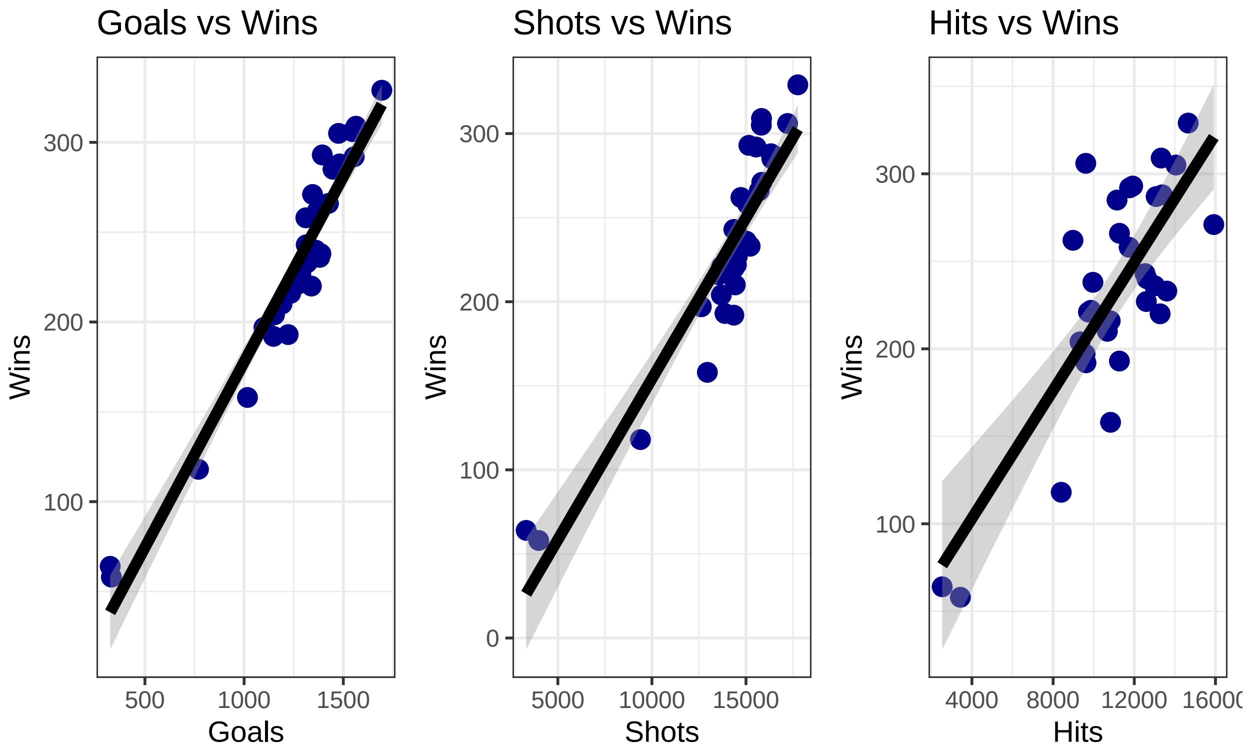
Question 3: Which features, if any, have a linear relationship with wins?

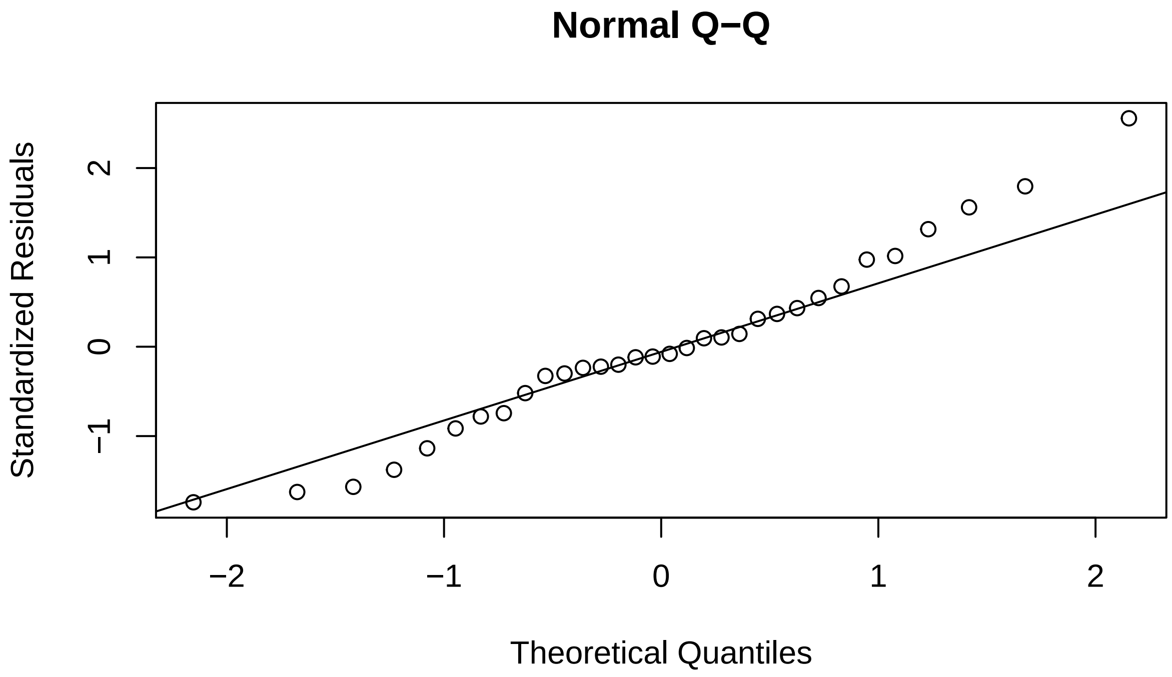
The purpose of this question was to explore linear regression related to hockey data. My intent was to use goals, shots, and hits as explanatory variables and wins as the response variable. Goals and shots are popular offensive metrics and hits is a popular defensive metric. To conduct this analysis, I created a numerical column of wins, selected the necessary features, grouped the data by team, and totaled the numerical columns. The resulting data frame contained one row for every team in the NHL and their totals for goals, shots, hits, and wins over the past nine seasons. As we learned, one assumption of linear regression is that the relationship between the explanatory variables and the response variable is truly linear. Therefore, I graphed each explanatory variable against wins, independently, and observed strong, positive linear relationships for each. I decided to take the analysis one step further and explore multiple linear regression including additive variables and interactions between variables. The three models can be seen below. To select the best model, I used the Akaike Information Criterion (AIC). This process of model selection chose the second model over the other two. In words, the expected number of total wins for a team can be regressed using the interaction of total goals and total shots and the additivity of total hits. As a sanity check, I graphed a QQ plot of the theoretical quantiles versus the standardized residuals and the data appeared normal. Therefore, the assumption of regression, that the true errors are normally distributed, was met. Then I plotted a fitted versus residuals plots. The data was centered around zero with no apparent pattern so the assumption that the true errors have equal variance and the assumption that the relationship between the variables is in fact linear were met. Although linear regression can be a considered a naïve approach in data science, I was content with the results for this exploratory data analysis.

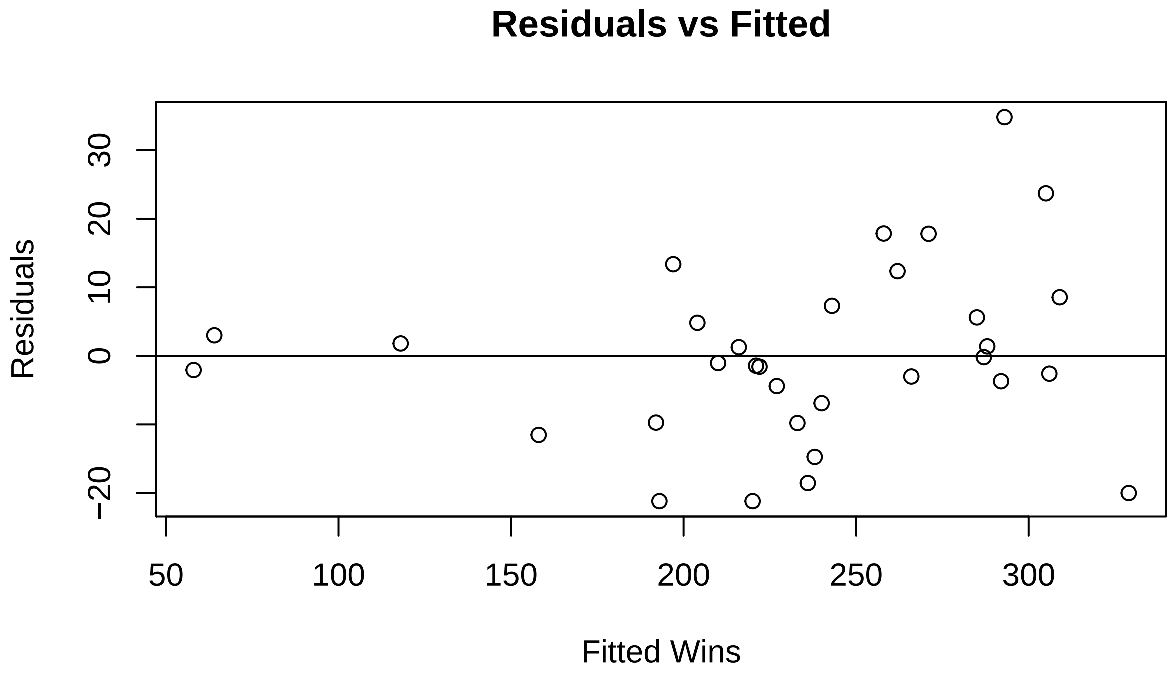
$$wins = goals + shots + hits$$

$$wins = goals * shots + hits$$

$$wins = goals * shots * hits$$




Normal Q−Q
Standardized Residuals
Theoretical Quantiles
−2
−1
0
1
2
−1
0
1
2


Residuals vs Fitted
Residuals
Fitted Wins
50
100
150
200
250
300
−20
0
10
20
30

References:

The data is from Kaggle: https://www.kaggle.com/martinellis/nhl-game-data

The course lecture materials for SI618 were reviewed while writing this report.

My previous work completed in part one of the project was referenced while writing this report.