# Influential Factors on Hockey Team Success

Vincent Latona

2023-03-16

## Problem Introduction

The National Hockey League was originally founded in 1917 as a professional ice-hockey association. Ice-hockey is by-far the most popular sport today in Canada and has only grown in popularity in the United States since the 1960's. In addition to increased popularity of the game, ice-hockey has also become a new frontier for data analytics. All kinds of statistics are collected in professional ice-hockey today, including: goals, assists, goals against average, save percentage, high-danger chances, expected goals against, goals saved above expected and many more. One of the most intriguing questions surrounding the sport today is: **What factors influence the success of a hockey team?** A statistical solution to this exact problem is something that many NHL general managers would love to have; giving them a statistical blueprint for building a great hockey team.

In the following report, *"success"* will be defined in the context of a data set of ice-hockey team statistics. The data set will be analyzed for trends in the data and various factors will be used to create possible models of *"success."* The data set to be used is "Teams.csv" from the Professional Hockey Database, courtesy of Open Source Sports.

### Necessary Libraries

Libraries that will be used for analysis include: `dplyr,` `tidyverse,` and `data.table` for cleaning/managing the imported data set, and `ggplot2` for data visualization.

```r
library(dplyr) #Data Wrangling
library(tidyverse)
library(data.table)
library(ggplot2) #Graphing
```

### Data Set Description

The data set to be analyzed is a collection of ice-hockey team statistics which were collected between 1909 and 2011. The data set includes all relevant information of teams including:
* Year the data was collected
* League the team is associated with
* Regular season division rank
* Number of games played
* Number of games won, lost and tied/lost in overtime
* Number of ranking points
* Number of goals for and against
* Final playoff position (if applicable)

# Data Wrangling

Before performing Exploratory Data Analysis, the data set that has been imported must first be standardized. Only necessary columns from the data set will be kept, including: Year, Name, Rank, Wins, Losses, Overtime losses, Points, Goals For, Goals Against, and Playoff results. Data analysis and modeling will be used on the standardized data set in order to draw possible conclusions for modern NHL team success. The number of games played during a current regular season is 82, thus only data points for 82-game seasons will be used for analysis. The current format of Rank within the original data set is the division rank, which must be changed to league ranking according to the team's regular season performance. Playoff results must also be standardized as another possible factor of success.

```r
df = read.csv("Teams.csv") #Read the csv file for the project
attach(df) #Attach the data frame for column referencing

stand = filter(df, G == 82 & lgID == "NHL") %>% #Filter 82-game seasons
  select(year, name, rank, W, L, OTL, Pts, GF, GA, playoff) #Select necessary fields

min = min(stand$year) #Get least recent year
max = max(stand$year) #Get most recent year

setorder(stand, year, -Pts) #Order data according to year and points

for (i in min:max) #Re-rank iterator
{
  stand$rank[stand$year == i] = #Re-rank according to standings
    1:length(stand$year[stand$year == i])
}

stand$playoff = recode_factor(stand$playoff, "SC" = 1, #Re-code playoffs
                              "F" = 2, "CF" = 3, "CSF" = 4,
                              "CQF" = 5, .default=6)

print.data.frame(head(stand, 5)) #Display standardized data frame
```

```
   year                name rank  W  L OTL Pts  GF  GA playoff
8  1995   Detroit Red Wings    1 62 13  NA 131 325 181       3
6  1995  Colorado Avalanche    2 47 25  NA 104 326 240       1
18 1995 Philadelphia Flyers    3 45 24  NA 103 282 208       4
19 1995 Pittsburgh Penguins    4 49 29  NA 102 362 284       3
16 1995    New York Rangers    5 41 27  NA  96 272 237       4
```

It is important that the data was collected in a similar context, thus the data points that were kept were collected during seasons that consisted of 82 games. It is also important that the data points are descriptive of ice-hockey teams at the same level of competition, thus there was a need to keep data points regarding teams in the NHL. Both of these factors provide a foundation for contemporary ice-hockey, providing possible insights into how the statistics that are currently collected during the course of modern 82-game NHL season may inform team success. The purpose of re-ranking teams according to league standings is a way to better quantify success that would not be possible with relative ranking by divisions. Playoff data was re-coded to allow for quantified analysis where the lower the number associated with the team, the further the team progressed during the playoffs. The best a team can do in the playoffs is to win a Stanley Cup championship, `playoff` = 1, and the worst a team can do is to not make the playoffs at all, `playoff` = 6.

# Exploratory Data Analysis

## Data Set Summary

To gain some elementary insights for each column of data from the standardized data set, the summary of the data set is displayed below:

```
summary(stand) #Display summary
```

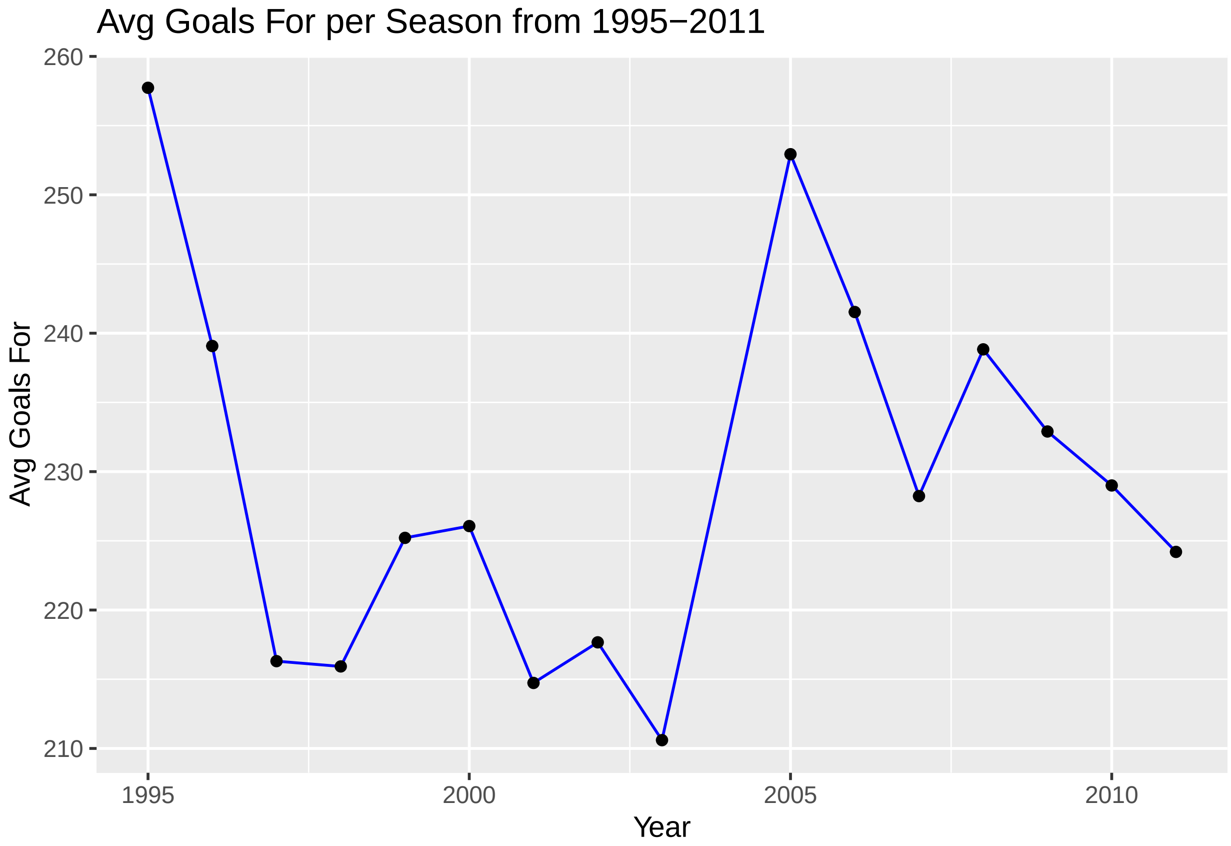
```
      year          name                rank             W        
 Min.   :1995   Length:463         Min.   : 1.00   Min.   :14.00  
 1st Qu.:1999   Class :character   1st Qu.: 8.00   1st Qu.:32.00  
 Median :2003   Mode  :character   Median :15.00   Median :39.00  
 Mean   :2003                      Mean   :15.02   Mean   :38.02  
 3rd Qu.:2008                      3rd Qu.:22.00   3rd Qu.:44.00  
 Max.   :2011                      Max.   :30.00   Max.   :62.00  
                                                                  
       L              OTL              Pts               GF        
 Min.   :13.00   Min.   : 0.000   Min.   : 39.00   Min.   :151.0  
 1st Qu.:27.00   1st Qu.: 5.000   1st Qu.: 77.50   1st Qu.:211.0  
 Median :31.00   Median : 7.000   Median : 90.00   Median :228.0  
 Mean   :32.25   Mean   : 7.461   Mean   : 87.77   Mean   :229.3  
 3rd Qu.:37.00   3rd Qu.:10.000   3rd Qu.: 99.00   3rd Qu.:247.0  
 Max.   :59.00   Max.   :18.000   Max.   :131.00   Max.   :362.0  
                 NA's   :105                                      
       GA        playoff
 Min.   :164.0   1: 16  
 1st Qu.:206.5   2: 16  
 Median :228.0   3: 32  
 Mean   :229.3   4: 64  
 3rd Qu.:249.0   5:128  
 Max.   :357.0   6:207  
```

According to the summary of the standardized data frame, the data was collected for teams between the 1995 and 2011 seasons. In a single season the maximum number of wins was 62 and the minimum was 14. The maximum number of losses was 59 and the minimum was 13. The average number of wins and losses in the standardized data set were 38.02 and 32.25 respectively. The average number of goals scored by teams was 229.3. The average number of standings points in the standardized data set was 87.77. It could be useful to view some of these statistics plotted per season from 1995 and 2011 to view trends per year, which could inform possible models of team success.

### Trends in Goals For

The following plot displays the trend of the average number of goals scored by teams during a season:

```
goalsfor = group_by(stand, year) %>% #Generate summary of goals for
  summarize(GF = mean(GF))
ggplot(goalsfor, aes(x=goalsfor$year, y=goalsfor$GF)) + #Plot the summary
  geom_line(color="blue") +
  geom_point() + labs(title="Avg Goals For per Season from 1995-2011", #Add labels
                      y="Avg Goals For",
                      x="Year")
```

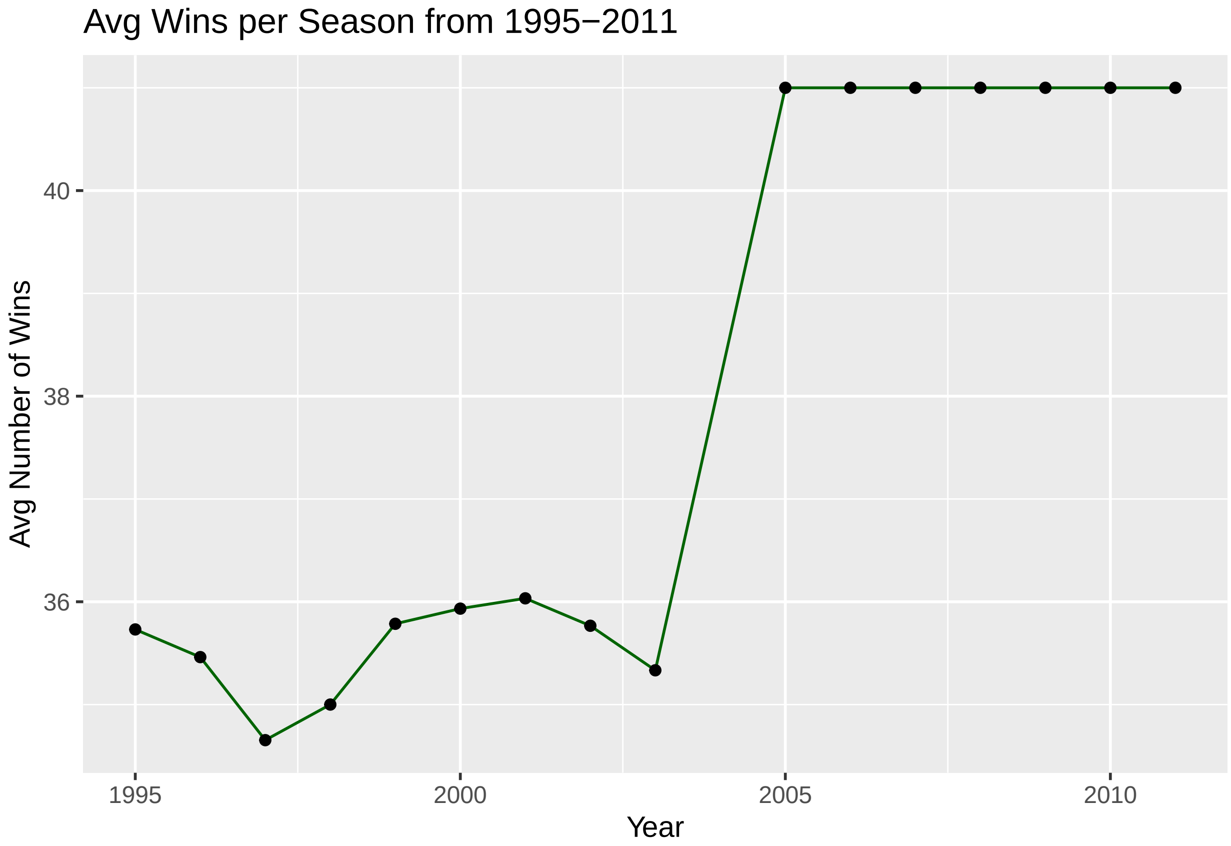


The trend of the average number of goals for tends to decrease between 1995 and 2003, and once again between 2005 and 2011. These trends in the data could be explained by improved goal-tending by goalies, decreased goal-scoring prowess, or some combination of both.

### Trends in Wins

The following plot displays the trend of the average number of wins per season:

```r
wins = group_by(stand, year) %>% #Generate summary of wins
  summarize(W = mean(W))
ggplot(wins, aes(x=wins$year, y=wins$W)) + #Plot the summary
  geom_line(color="darkgreen") +
  geom_point() + labs(title="Avg Wins per Season from 1995-2011", #Add labels
                      y="Avg Number of Wins",
                      x="Year")
```

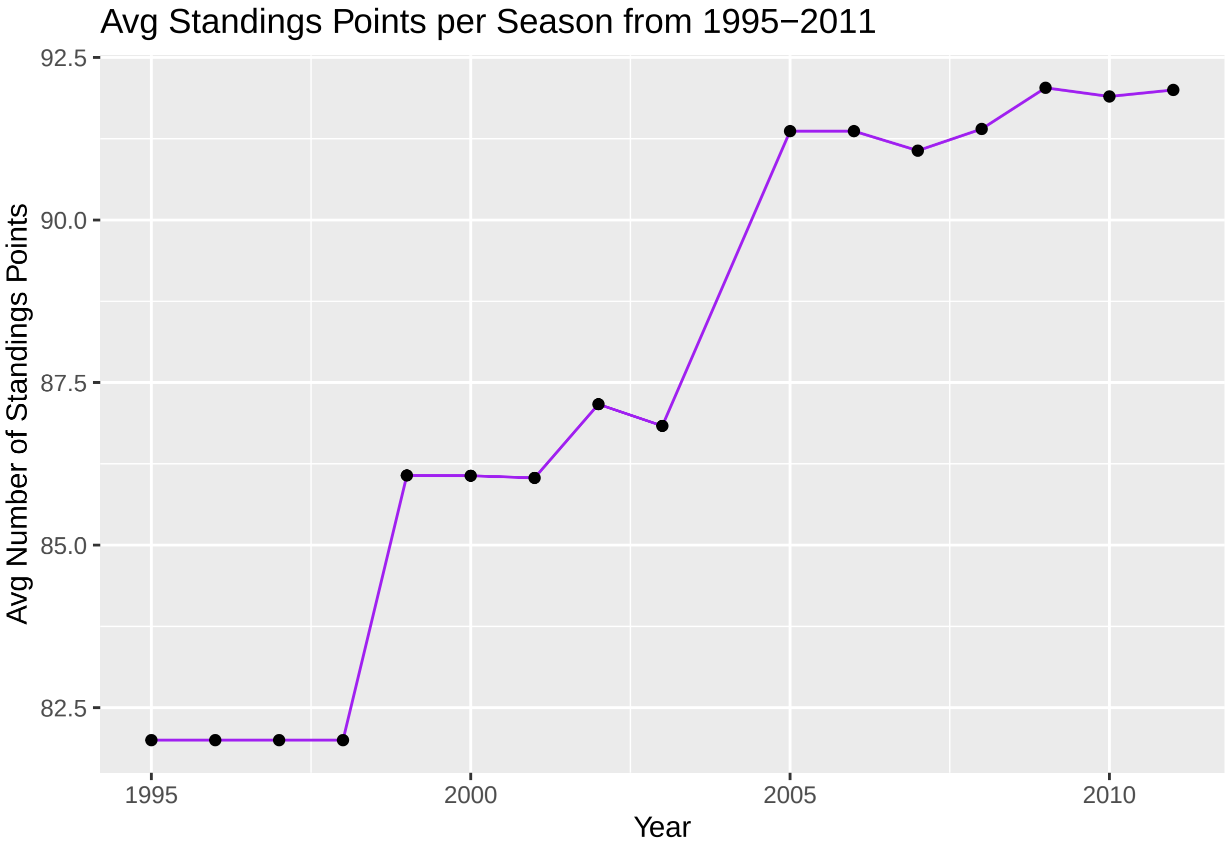


The trend of the average number of wins appears to cycle between increasing and decreasing until 2005. After 2005, teams earn about 41 wins per season on average. After 2000, the NHL had 30 teams, which may have adjusted after expansion drafts. Perhaps once teams adjusted to their placement in the league, they began to normalize and earn more wins per season.

### Trends in Standings Points

The following plot displays the trend of the average number of standings points earned per season:

```
pts = group_by(stand, year) %>% #Generate summary of points
  summarize(p = mean(Pts))
ggplot(pts, aes(x=pts$year, y=pts$p)) + #Plot the summary
  geom_line(color="purple") +
  geom_point() + labs(title="Avg Standings Points per Season from 1995-2011", #Add labels
                      y="Avg Number of Standings Points",
                      x="Year")
```



The initial trend of standings points remained constant until about 1999. In 1999, the NHL started to track the number of Losses in Overtime. An overtime loss counts as 1 point towards the standings. When this began, the number of points began to steadily increase.

# Modeling

## Quantifying Success

There are a variety of ways to classify success in the context of ice-hockey. For the purpose of this report, success will be defined in the context of the regular season and will also be defined in the context of the playoffs. Success in the regular season will be measured by the number of standings points earned by teams, which directly influence league rank, during the season. Success in the playoffs will be measured by the final positioning in the playoffs, with 1 resulting in a Stanley Cup Championship, 2 resulting in a loss in the Stanley Cup Finals, 3 resulting in a loss in the Conference Championship, and so on. 6 results in not making the playoffs at all.

## Regular Season - Standings Points from Goals For and Goals Against

In the following model, multivariate linear regression modeling will be used to determine whether Goals For and Goals Against have an effect on Standings Points. A significance level of $\alpha = 0.05$ will be used to evaluate the model(s).

```r
spModel = lm(Pts ~ GF + GA, stand) #Perform linear regression
summary(spModel) #Display ANOVA
```

```
Call:
lm(formula = Pts ~ GF + GA, data = stand)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6784  -3.9314  -0.2002   4.0854  14.5687

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.705233   2.934446   28.52   <2e-16 ***
GF           0.345681   0.009429   36.66   <2e-16 ***
GA          -0.327961   0.008438  -38.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.676 on 460 degrees of freedom
Multiple R-squared:  0.863, Adjusted R-squared:  0.8624
F-statistic:  1448 on 2 and 460 DF,  p-value: < 2.2e-16
```

According to the Analysis of Variance table, the p-values associated with Goals For and Goals Against are significantly small. Both have p-values of less than 2e-16, which indicates that there is a significant connection between Goals For and Goals Against with Standings Points. The Adjusted Multiple $R^2$ value is 0.862, which explains that about 86% of the variance in Standings Points is covered by Goals For and Goals Against. This provides a possible model for success in the regular season. The resulting model demonstrates that Goals For and Goals Against are highly influential on the number of Standings Points earned: $\text{Standings Points} = 83.71 + 0.35 \cdot \text{Goals For} - 0.33 \cdot \text{Goals Against} + \varepsilon$ with $\varepsilon$ error.

## Playoffs - Playoff performance from League Rank, Goals For, and Goals Against

In the following model, multivariate linear regression will be used to determine the independent variables (League Rank, Goals For, Goals Against) that affect performance of ice-hockey teams during the playoffs. A significance level of $\alpha = 0.05$ will be used to evaluate the model(s).

```
pModel = lm(as.numeric(playoff) ~ rank + GF + GA, stand) #Perform linear regression
summary(pModel) #Display ANOVA
```

```

Call:
lm(formula = as.numeric(playoff) ~ rank + GF + GA, data = stand)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7260 -0.3515  0.1373  0.5805  1.7516 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.637349   0.507667   7.165 3.12e-12 ***
rank         0.099728   0.012241   8.147 3.57e-15 ***
GF          -0.003134   0.002693  -1.164    0.245    
GA           0.002236   0.002608   0.857    0.392    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9158 on 459 degrees of freedom
Multiple R-squared:  0.5168,	Adjusted R-squared:  0.5137 
F-statistic: 163.7 on 3 and 459 DF,  p-value: < 2.2e-16
```

```
rpModel1 = lm(as.numeric(playoff) ~ rank + GF, stand) #Perform linear regression
summary(rpModel1) #Display ANOVA
```

```

Call:
lm(formula = as.numeric(playoff) ~ rank + GF, data = stand)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7634 -0.3624  0.1416  0.5596  1.7292 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.645374   0.507435   7.184 2.75e-12 ***
rank         0.108678   0.006389  17.011  < 2e-16 ***
GF          -0.001520   0.001924  -0.790     0.43    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9155 on 460 degrees of freedom
Multiple R-squared:  0.516,	Adjusted R-squared:  0.5139 
F-statistic: 245.3 on 2 and 460 DF,  p-value: < 2.2e-16
```

```
rpModel2 = lm(as.numeric(playoff) ~ rank, stand) #Perform linear regression
summary(rpModel2) #Display ANOVA
```

```
Call:
lm(formula = as.numeric(playoff) ~ rank, data = stand)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.7035 -0.3682  0.1788  0.5260  1.6377 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.250501   0.086910   37.40   <2e-16 ***
rank        0.111769   0.005048   22.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9151 on 461 degrees of freedom
Multiple R-squared:  0.5154,    Adjusted R-squared:  0.5143 
F-statistic: 490.3 on 1 and 461 DF,  p-value: < 2.2e-16
```

According to the first Analysis of Variance table, League Rank was the most significant independent variable with a p-value of 3.57e-15. The next step was to remove the least significant independent variable, Goals Against, which had a p-value of 0.392. Another round of linear regression was performed without Goals Against, resulting in a p-value of 0.43 and less than 2e-16 for Goals For and League Rank respectively. This model reaffirmed the significance of League Rank and resulted in the removal of Goals For from the model. The final round of linear regression resulted in a p-value of less than 2e-16 for League Rank. The final value of the Adjusted $R^2$ value was 0.514, which explains that only 51% of the variance in Playoff performance is determined by League Rank. The resulting model demonstrates that League Rank is highly influential on final playoff placement: $\text{Playoff Result} = 3.25 + 0.11 \cdot \text{League Rank} + \varepsilon$ with $\varepsilon$ error.

## Conclusions

After performing multivariate linear regression to determine influential factors on success in the regular season and playoffs, it was determined that both Goals For and Goals Against are highly influential on the number of Standings Points received during the regular season and that League Rank was highly influential on Playoff performance. Regular season success can be explained in terms of goals scored for your team and good goal-tending will result in more wins during the course of a season. Playoff performance could possibly be explained by the format of the playoffs where rank within a given division will result in more favorable match-ups in the post-season, as well as possible advantage for playing in the team's home arena for more games in each series. However, the model generated for playoff success was not a very adequate model and in order to derive a better model in the future, it would be necessary to analyze statistics collected during the playoffs. In general, this information could be useful for NHL general managers to determine how to build teams for better performance during the regular season, which could influence positioning for performance in the playoffs.

## Citations

"Professional Hockey Database." www.kaggle.com, 2020, www.kaggle.com/datasets/open-source-sports/professional-hockey-database?select=Teams.csv.