You Can’t Win From The Penalty Box

An In-Depth Exploration of The Effects of Penalties in the NHL

Bella Salter

2024-11-27

1 Abstract

In this study, I aim to discover the nuances within the relationship of penalties and game outcomes in the National Hockey League. Using win probability models, hypothesis testing, and linear regression models, I determined that the effects of penalties in hockey is highly variable, and in many cases, marginal.

2 Introduction

In hockey, penalties are considered a vital part of the game. Due to the adverse effects of punishments on scoring, it is widely believed that they help determine the outcome of each game. As such, it becomes increasingly important to determine how these situations affect the game on a micro- and macro- level.

In the first component of my study, I investigated how penalties affect single-game situations. I trained win probability models on the current goal differential and the current penalties in minutes for each team, and studied the differences between models with penalty information and those without. I found that they were overall unsuccessful, but that penalties could occasionally increase the win probability in certain situations. Additionally, I investigated the effects of the current goal differential on how many penalties were committed. I determined that winning teams gained slightly more penalties in minutes than losing teams. However, teams that came back from a deficit had the same amount of penalties as those who did not. Similarly, teams that lost despite a lead had the same number of penalties as those who stayed winning.

However, it is important to note my investigation of teams’ behavior at a goal differential of 0. I found that teams who ended up winning the game had more penalties in minutes during the time in which they were tied. I would hesitate to assume a causal relationship, though, since penalties can be a side effect of an aggressive play style.

In the second component of my study, I determined the relationship between penalties in minutes and various end-of-season statistics. I trained various linear models on wins, rank, penalties in minutes per game, goals per game, and goals against per game. I found that the only statistic of which penalties was a good predictor was goals scored per game. Thus, I generally found a high variance in the ways that penalties can affect success statistics.

3 Background

Penalties in the National Hockey League can arise from a player hitting another with their stick, tripping an opposing player, fighting another player, and more. Given the wide range of events that can spur a penalty, they are split into three different categories– minors, majors, and game ejections. For the purposes of this study, minors and majors are the most important. When these penalties occur, the player must not play for 2 minutes for a minor, or 5 minutes in the event of a major penalty. During this time, one team plays with five men, while the other has only four, meaning that penalties can play a large role in determining the game winner. Capitalizing on the opposing team’s penalty is so vital to a team’s success that the time in which an opposing skater is serving a penalty was dubbed a “power play”. On the other hand, the opposing team is on a “penalty kill” during that time. Success rates on the power play vary every year, but NHL teams scored during 19% of power plays on average in 2018[3].

Despite the disadvantage of temporarily having one less player than another team, fans and commentators alike often brag about one team's physicality. Additionally, they have been a source of wide debate in recent years given the rise in popularity of hockey. Proponents of keeping punishments for penalties mild argue that they hold the other teams' players accountable for their actions, while others think that they are purely an unsafe practice[2]. As a fan, one must wonder if the effects of physical penalties are exclusively disadvantageous to winning, given their large role in hockey culture.

Previous work has considered the effects of penalties in minutes on goal-scoring for both players and teams. However, they often lack in depth and consideration of nuance within the dynamics of hockey. For example, one found a correlation between penalties in minutes and goal scoring per player, but did not consider the effects of ice time on this relationship[3]. These studies also lack consideration of effects of penalties on both in-game success and season success.

Furthermore, given the wide range of events that can trigger a penalty, it could be worth asking whether this can cause discrepancies in their effects on winning games. For example, "roughing" and "too many men on the ice" are both penalties, with the same sentence. This means that not only can the result of the penalty affect the game, but it could be possible that the effects of more physical penalties could affect the game differently. Similarly, fighting is somewhat allowed in hockey. A team can initiate a fight and receive a few penalties, and occasionally no players are outright ejected. Thus, it could be worth asking whether or not winning fights affects the outcome of a particular game.

Additionally, as in many sports, different teams have different play styles. Some may be more adapted to a skillful penalty kill, while others may be strongest when they minimize their penalties. Thus, I will also consider deviation from a team's regular performance during that season in analyzing the effects of penalties in a singular game.

4 Implementation

Using the NHL API, I was able to request data for the years from 2011-2018 using a Python script. I requested schedule data for each team, box scores, and the play-by-play for each regular season game. I compiled the data in various csv files, which I then imported into my R code.

4.1 Single Game Analysis

4.1.1 Win Probability Models

First, I trained two separate win probability models. The first included the time on ice, penalties in minutes for each team, physical penalties in minutes for each team, and the goals each team had at the time. On the other hand, the second model did not include the penalty information. This was primarily to be able to compare the predictions with the extra information of penalties, to the predictions of a model without this information.

```r
mod_PIM <- glm(tm1_win ~ ., family ="binomial", data = train)
mod_noPIM <- glm(tm1_win ~ ., family ="binomial", data = select(train, -tm1pPIM,
                                                              -tm2pPIM,-tm1PIM, -tm2PIM))
```

My final win probability model considered data on the team's current PIM in physical penalties. I classified a penalty as 'physical' if it was not a delay of game or too many men on the ice penalty. This is because nearly every other penalty involved physical contact with another player, or was so exceedingly rare that it does not occur in the dataset.

```r
mod_pPIM <- glm(tm1_win ~ ., family ="binomial", data = select(train, -tm1PIM, -tm2PIM))
```

4.1.2 Effects of Current Goal Differential on Penalties Committed

To determine the effects of current goal differential on penalties, I collected aggregate statistics on the typical penalties in minutes added at each goal difference. Since there were very many games in the dataset, I had sufficient data to to analyze. I completed various t-tests to determine if the number of physical penalties committed truly changed at different goal differentials, assuming that the true difference in means was equal to 0 as the null hypothesis.

```
pop1 <- diff_vs_ppims[diff_vs_ppims$goal_diff > 0,]
pop2 <- diff_vs_ppims[diff_vs_ppims$goal_diff < 0,]
t.test(pop1$pims_acquired, pop2$pims_acquired)
```

Additionally, I did a t-test to determine whether or not teams who ended up winning committed more penalties during a tied game than those who did not.

```
pop1_tie <- diff_vs_ppims[diff_vs_ppims$goal_diff == 0 & diff_vs_ppims$tm1_won == 1,]
pop2_tie <- diff_vs_ppims[diff_vs_ppims$goal_diff < 0 & diff_vs_ppims$tm1_won == 0,]
t.test(pop1_tie$pims_acquired, pop2_tie$pims_acquired)
```

4.1.3 Considering Comebacks and Fumbles

I also wanted to investigate how the style of play differs in teams who make a comeback from those who did not. Thus, I conducted various t-tests to determine how physical penalties change in those who make a comeback from those who did not. I partitioned this into cases where teams made a close comeback(returning to win the game with a goal differential of less than 3) and those who made a drastic comeback(returning to win with a differential of greater than 3).

```
close_pop1 <- diff_vs_ppims[diff_vs_ppims$goal_diff < 0 & diff_vs_ppims$goal_diff > -3
                            & diff_vs_ppims$tm1_won == 1,]
close_pop2 <- diff_vs_ppims[diff_vs_ppims$goal_diff < 0 & diff_vs_ppims$goal_diff > -3
                            & diff_vs_ppims$tm1_won == 0,]
t.test(close_pop1$pims_acquired, close_pop2$pims_acquired)

drastic_pop1 <- diff_vs_ppims[diff_vs_ppims$goal_diff < -3 & diff_vs_ppims$tm1_won == 1,]
drastic_pop2 <- diff_vs_ppims[diff_vs_ppims$goal_diff < -3 & diff_vs_ppims$tm1_won == 0,]
t.test(drastic_pop1$pims_acquired, drastic_pop2$pims_acquired)
```

Additionally, I considered “fumbles”, or a team losing a game in which they had a lead at some point.

```
close_pop1_ahead <- diff_vs_ppims[diff_vs_ppims$goal_diff > 0
                                & diff_vs_ppims$goal_diff < 3
                                & diff_vs_ppims$tm1_won == 1,]
close_pop2_ahead <- diff_vs_ppims[diff_vs_ppims$goal_diff > 0
                                & diff_vs_ppims$goal_diff < 3
                                & diff_vs_ppims$tm1_won == 0,]
t.test(close_pop1_ahead$pims_acquired, close_pop2_ahead$pims_acquired)

drastic_pop1_ahead <- diff_vs_ppims[diff_vs_ppims$goal_diff > 3
                                  & diff_vs_ppims$tm1_won == 1,]
drastic_pop2_ahead <- diff_vs_ppims[diff_vs_ppims$goal_diff > 3
                                  & diff_vs_ppims$tm1_won == 0,]
t.test(drastic_pop1_ahead$pims_acquired, drastic_pop2_ahead$pims_acquired)
```

4.2 Season-Wide Analysis

4.2.1 Linear Models

Next, I trained an array of linear models on season-wide data. This included seasons 2011-2018, with statistics for each team such as wins/losses, points(which depends on how a win was achieved), shootout wins/losses, ranking, strength of schedule, goals, penalties in minutes, and various shot and save statistics.

```
#remove clearly correlated info
szn_data <- subset(all_szns, select = -c(Tm, L, OL, PTS, SOL, GP, SOW, `PTS%`, Season))

lmod_all <-  lm(W ~ ., data = select(szn_data, -c(Rk)))
lmod_ga <- lm(`GA/G` ~. , data = select(szn_data, -c(W, Rk)))
lmod_gf <- lm(`GF/G` ~. , data = select(szn_data, -c(W, Rk)))
lmod_rk <- lm(Rk ~ . , data = select(szn_data, -c(W, SRS, SOS)))
```

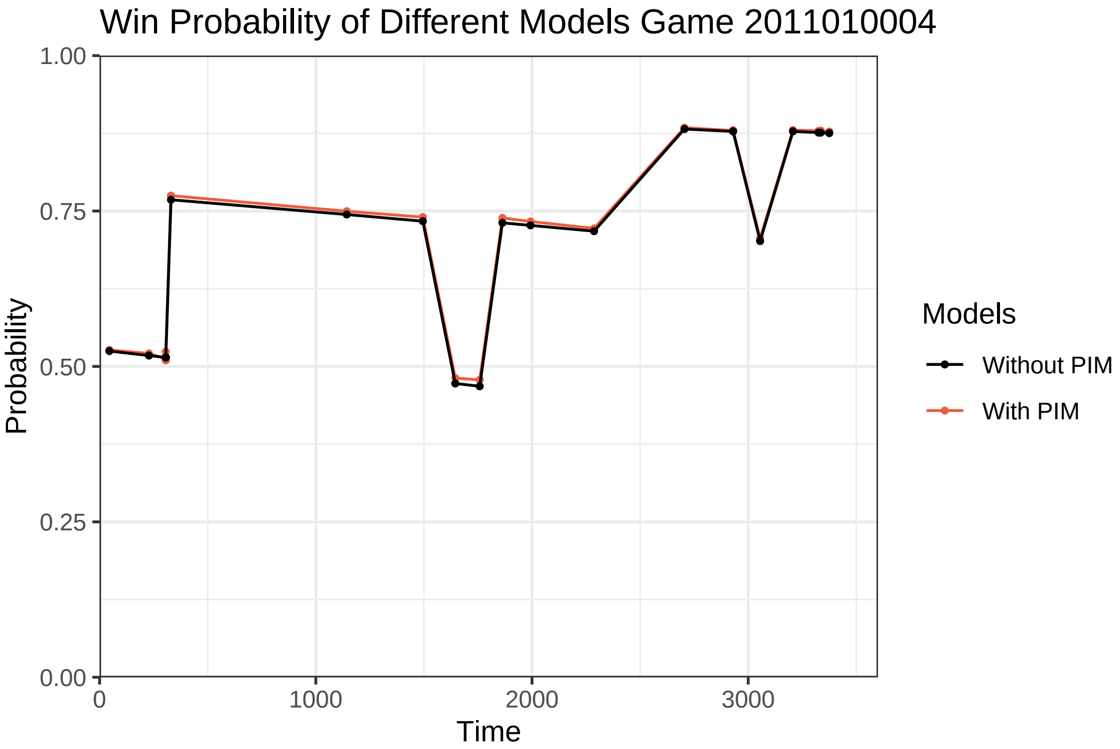
After analyzing the correlation plot of these statistics, I created final linear models to describe wins, goals for/against per game, and rank, based on the other season-long statistics collected for each team.

5 Results

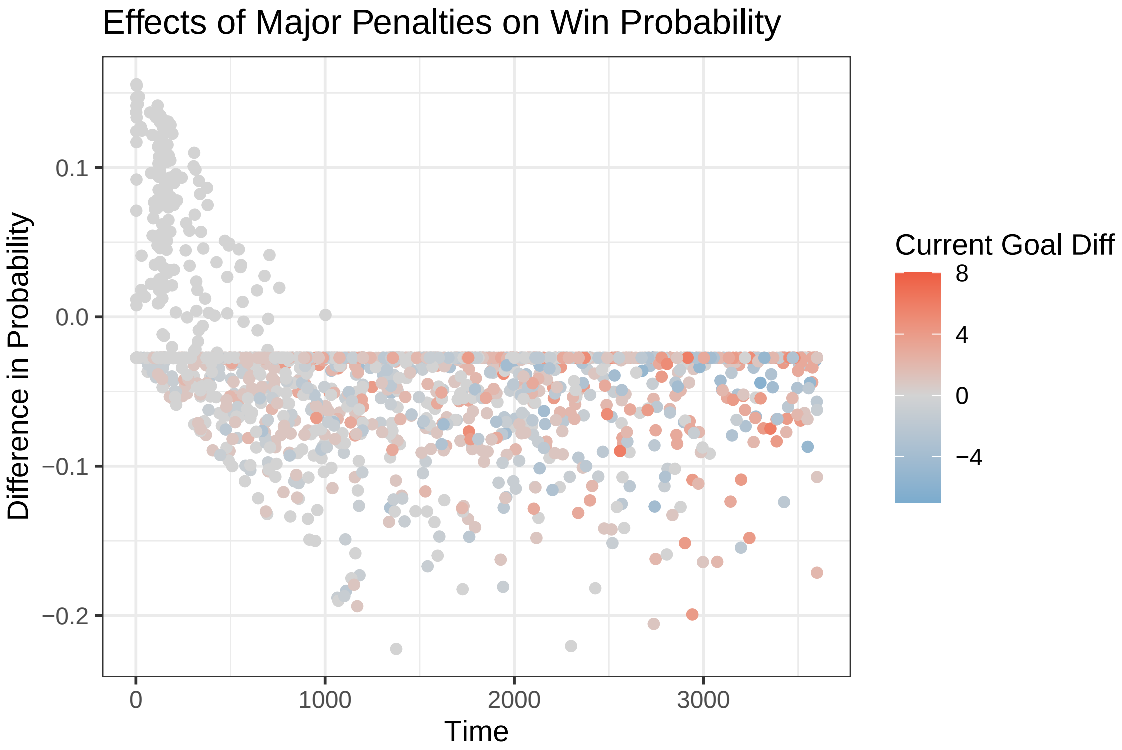
5.1 Single Game Analysis

5.1.1 Win Probability Models

After training my probability models, I decided to determine how win probabilities are affected during specific scenarios. I found a specific game that had a significant amount of fighting.



Notice that the probabilities do not differ much at all. Furthermore, the model with all penalties agreed with the model with only physical penalties. This is probably an indication that penalties do not significantly alter win probability.



Notice that while the fighting can occasionally increase the win probability, especially in the first period, it can often decrease the probability of winning. However, it appears that the closer it is towards the beginning of the game, the better fighting is.

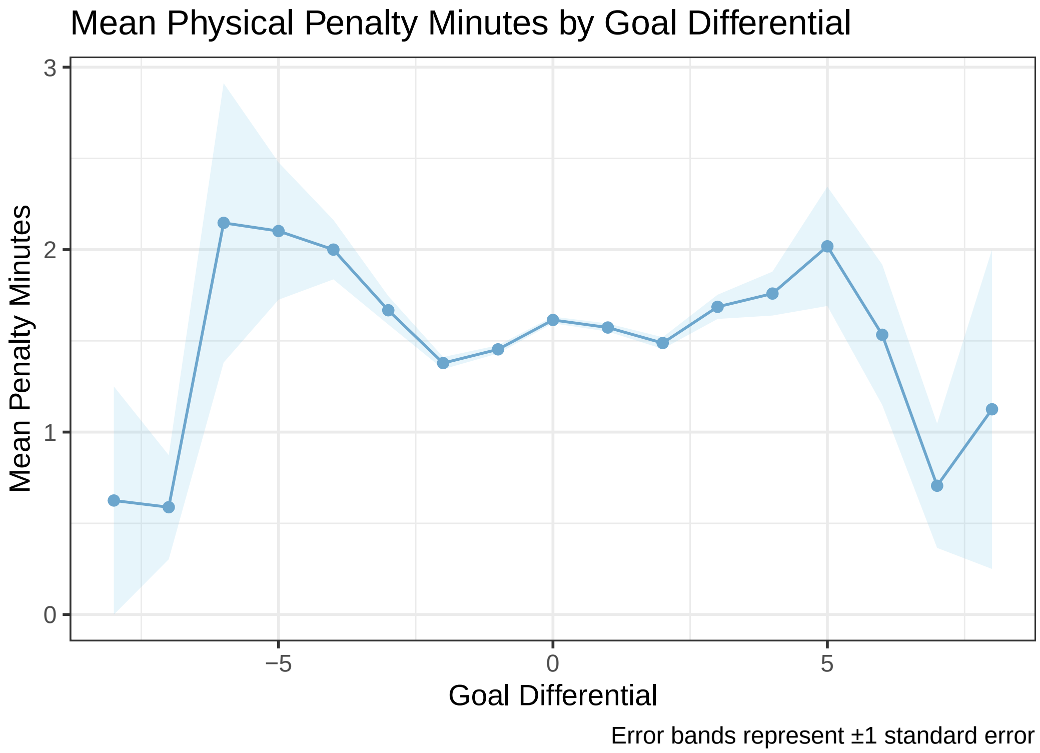
One method of evaluating a win probability mode is using the R^2 metric. This computes the explained deviance by the model in a ratio to the deviance that was not explained.

```
## [1] "R squared including physical penalties: 0.299833792075795"
```

```
## [1] "R squared without penalties: 0.299798709484469"
```

Here, we can see that the R^2 for both models were far from ideal. This indicates that the relationship described by the win probability model is not accurate, and can be improved further.

5.1.2 Effects of Current Goal Differential on Penalties Committed



From the general trends in goal differentials, we can see that penalties committed vary greatly depending on current goal differential. Additionally, there is a lot of variance in this data, some of which we expect. Of course, games where one team is winning by over 6 goals are exceedingly rare, so we expect that this are a will have a high degree of variance.

For the t-test determining the difference in means for the teams currently winning or losing, the results were

```
t.test(pop1$pims_acquired, pop2$pims_acquired)
```

```
##
##  Welch Two Sample t-test
##
## data:  pop1$pims_acquired and pop2$pims_acquired
## t = 3.4054, df = 49085, p-value = 0.0006611
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03785953 0.14053548
## sample estimates:
## mean of x mean of y
##  1.572071  1.482873
```

Thus, we are 95% confident that the true difference in penalties in minutes acquired by teams who are currently winning and teams who are currently losing is in the interval (0.038, 0.14). Thus, we reject the null hypothesis. However, this is an average, and since teams cannot commit penalties that are less than 2 minutes long, this is still not a strong indicator that teams who are winning commit more penalties.

For the tied game situation, the results were

```r
t.test(pop1_tie$pims_acquired, pop2_tie$pims_acquired)
```

```
##
##  Welch Two Sample t-test
##
## data:  pop1_tie$pims_acquired and pop2_tie$pims_acquired
## t = 3.3829, df = 25054, p-value = 0.0007184
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05250381 0.19715951
## sample estimates:
## mean of x mean of y
##  1.613919  1.489087
```

Thus, we reject the null hypothesis. This means that we are 95% sure that teams who ended up winning commit between 0.05 and 0.19 more minutes of penalties while tied than teams who ended up losing.

5.1.3 Considering Comebacks and Fumbles

For the t-test determining the difference in means for the comeback teams, the results were

```r
t.test(close_pop1$pims_acquired, close_pop2$pims_acquired)
```

```
##
##  Welch Two Sample t-test
##
## data:  close_pop1$pims_acquired and close_pop2$pims_acquired
## t = -0.83208, df = 20682, p-value = 0.4054
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10378193  0.04192665
## sample estimates:
## mean of x mean of y
##  1.413761  1.444689
```

```r
t.test(drastic_pop1$pims_acquired, drastic_pop2$pims_acquired)
```

```
##
##  Welch Two Sample t-test
##
## data:  drastic_pop1$pims_acquired and drastic_pop2$pims_acquired
## t = 1.844, df = 1205.4, p-value = 0.06543
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0345751  1.1157916
## sample estimates:
## mean of x mean of y
##  2.261538  1.720930
```

Thus, we are 95% confident that the true difference in the physical penalties in minutes for teams who come back from a deficit is between -0.1 and 0.04 for close comebacks, and -0.03 and 1.11 for drastic comebacks. Again, note that since extreme goal deficits in hockey are rare, the large confidence interval for the more dramatic games makes sense in the context of hockey. Since both of these intervals contain 0, we fail to reject the null hypothesis. Thus, we also cannot assume that physical penalties are committed more or less often by teams who do or do not make a comeback.

The results for the fumble analysis were

```
t.test(close_pop1_ahead$pims_acquired, close_pop2_ahead$pims_acquired)
```

```
##
##  Welch Two Sample t-test
##
## data:  close_pop1_ahead$pims_acquired and close_pop2_ahead$pims_acquired
## t = 0.27158, df = 20726, p-value = 0.7859
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06120491  0.08089339
## sample estimates:
## mean of x mean of y
##  1.549794  1.539950
```

```
t.test(drastic_pop1_ahead$pims_acquired, drastic_pop2_ahead$pims_acquired)
```

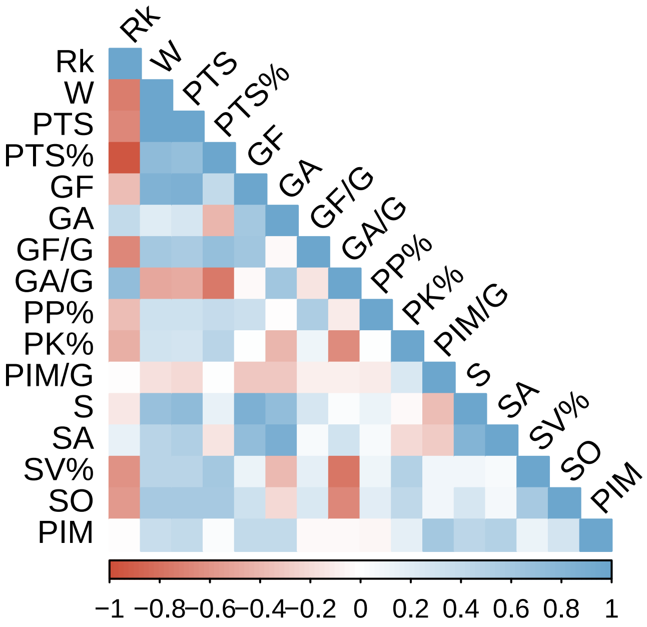
```
##
##  Welch Two Sample t-test
##
## data:  drastic_pop1_ahead$pims_acquired and drastic_pop2_ahead$pims_acquired
## t = 1.6168, df = 1249, p-value = 0.1062
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07800077  0.80887733
## sample estimates:
## mean of x mean of y
##  1.958462  1.593023
```

Thus, we are 95% confident that teams who fumbled a lead of less than 3 goals had roughly the same number of physical penalties in minutes as those who did not, and the same is true for those with a more dramatic lead. This implies that again, physical penalties in minutes are not strongly associated with keeping or losing a lead.

5.2 Season-Wide Analysis

5.2.1 Linear Models

As mentioned above, I first created a correlation plot to determine any confounding relationships between my data points.



Clearly, many stats such as goals against per game and goals against total are highly correlated. Thus, I threw these out when training my models.

```
szn_data <- subset(all_szns, select = -c(Tm, Season, AvAge, GP, L, OL,
                                          SOS, SRS, PPO, PPOA, SH, SHA, `S%`, SOW, SOL,
                                          `oPIM/G`, PP, PPA, OL, PTS, SOL, GP, GA, GF,
                                          `PTS%`, SA, SO))

lmod_all <-  lm(W ~ ., data = select(szn_data, -c(Rk)))
lmod_ga <- lm(`GA/G` ~. , data = select(szn_data, -c(W, Rk)))
lmod_gf <- lm(`GF/G` ~. , data = select(szn_data, -c(W, Rk)))
lmod_rk <- lm(Rk ~ . , data = select(szn_data, -c(W)))
summary(lmod_all)
```

```
##
## Call:
## lm(formula = W ~ ., data = select(szn_data, -c(Rk)))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.852 -1.704 -0.086  1.813  8.409
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.302e+01  3.589e+01   0.363    0.717
## ‘GF/G‘       1.614e+01  8.536e-01  18.906   <2e-16 ***
## ‘GA/G‘      -1.281e+01  1.101e+00 -11.643   <2e-16 ***
## ‘PP%‘       -8.024e-02  7.898e-02  -1.016    0.311
## ‘PK%‘       -1.968e-02  8.333e-02  -0.236    0.814
## ‘PIM/G‘     -2.985e+00  3.231e-01  -9.240   <2e-16 ***
## S            9.414e-04  1.342e-03   0.701    0.484
## ‘SV%‘        1.770e+01  3.616e+01   0.489    0.625
## PIM          4.078e-02  3.832e-03  10.640   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.699 on 202 degrees of freedom
## Multiple R-squared:  0.9202, Adjusted R-squared:  0.917
## F-statistic: 291.1 on 8 and 202 DF,  p-value: < 2.2e-16
```

```
summary(lmod_ga)
```

```
##
## Call:
## lm(formula = ‘GA/G‘ ~ ., data = select(szn_data, -c(W, Rk)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57615 -0.11815 -0.00043  0.10942  0.50268
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.611e+01  1.371e+00  19.046  < 2e-16 ***
## ‘GF/G‘      -2.954e-02  5.440e-02  -0.543    0.588
## ‘PP%‘       -9.445e-04  5.037e-03  -0.188    0.851
## ‘PK%‘       -3.710e-02  4.632e-03  -8.009 8.88e-14 ***
## ‘PIM/G‘      5.378e-03  2.060e-02   0.261    0.794
## S            3.767e-05  8.556e-05   0.440    0.660
## ‘SV%‘       -2.257e+01  1.676e+00 -13.467  < 2e-16 ***
## PIM          1.400e-04  2.442e-04   0.573    0.567
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1721 on 203 degrees of freedom
## Multiple R-squared:  0.7083, Adjusted R-squared:  0.6983
## F-statistic: 70.43 on 7 and 203 DF,  p-value: < 2.2e-16
```

```
summary(lmod_gf)
```

```
##
## Call:
## lm(formula = ‘GF/G‘ ~ ., data = select(szn_data, -c(W, Rk)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66926 -0.15095  0.00743  0.13507  0.55027
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.121e-01  2.951e+00  -0.241    0.810
## ‘GA/G‘      -4.910e-02  9.043e-02  -0.543    0.588
## ‘PP%‘        4.722e-02  5.585e-03   8.454 5.46e-15 ***
## ‘PK%‘        2.000e-03  6.850e-03   0.292    0.771
## ‘PIM/G‘      1.546e-01  2.425e-02   6.374 1.21e-09 ***
## S            7.438e-04  9.724e-05   7.648 8.00e-13 ***
## ‘SV%‘        7.690e-01  2.973e+00   0.259    0.796
## PIM         -1.911e-03  2.852e-04  -6.700 2.01e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2219 on 203 degrees of freedom
## Multiple R-squared:  0.4652, Adjusted R-squared:  0.4467
## F-statistic: 25.22 on 7 and 203 DF,  p-value: < 2.2e-16
```

```
summary(lmod_rk)
```

```
##
## Call:
## lm(formula = Rk ~ ., data = select(szn_data, -c(W)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3105  -2.1468   0.1784   2.2669  12.0435
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.466021  47.823646   1.181    0.239
## ‘GF/G‘      -16.760596   1.137446 -14.735   <2e-16 ***
## ‘GA/G‘       17.199451   1.466527  11.728   <2e-16 ***
## ‘PP%‘        -0.013273   0.105247  -0.126    0.900
## ‘PK%‘         0.140205   0.111035   1.263    0.208
## ‘PIM/G‘       0.122886   0.430549   0.285    0.776
## S             0.001085   0.001789   0.606    0.545
## ‘SV%‘       -60.340865  48.181393  -1.252    0.212
## PIM          -0.002727   0.005107  -0.534    0.594
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.596 on 202 degrees of freedom
## Multiple R-squared:  0.8364, Adjusted R-squared:   0.83
## F-statistic: 129.1 on 8 and 202 DF,  p-value: < 2.2e-16
```

Via the summaries, we can see that the only statistic that penalties in minutes per game is a good predictor of is goals per game. Furthermore, this linear model did not have a large R^2 value. Thus, due to the high variability in the play styles and penalties that each team receives, this is likely not a good indicator of any of the success statistics I evaluated.

6 Conclusion

I attempted to understand the relationship between penalties in minutes on both a small and large timescale. Within games, I determined that win probability models based solely off of goal differentials and penalties were very inaccurate, but that physical penalties often increased the win probability. Furthermore, home teams that fought within the first half of the first period often saw a spike in their win probability.

Additionally, teams who are winning tend to commit slightly more penalties than the other teams. Also, teams who ended up fumbling a lead committed the same number of penalties as teams who did not, as did teams who made a comeback or did not make a comeback. I found no difference in the outcome of games when teams were not tied. However, an important result was that teams who ended up winning a game had more penalties during the tied game on average than those who did not. In other words, teams who won were more likely to have more penalties in minutes during the time in which the goal differential was 0 than teams who did not.

Finally, on a season-long scale, I determined that penalties in minutes was only strongly related to the number of goals made per game. Intuitively, this contradicts previous thoughts regarding penalties. However, the R^2 value for this specific linear model was particularly high. Also, it is unclear how penalties can significantly affect the goal differential for each game, yet do not affect the number of games won.

Despite the possible interpretation of some of my results, I would hesitate to assume that penalties create more opportunities to score. First and foremost, penalties can occur due to a "scrappy" style of play – one that is common in teams that make a comeback from a deficit. Secondly, consideration of the current goal differential when such penalties are committed is crucial. At first glance, it may seem as though penalties are correlated with wins, but it is true that teams who have a positive goal differential tend to commit more penalties.

Additionally, it is important to remember that penalties are a subjective measure of the aggressiveness of a team. That is, referees are imperfect in their judgement, and could have attitudes that may affect their allocation of penalties to the currently winning or losing team. Thus, a better metric to measure in-game physicality could be hits. Future work could potentially include this in analysis, as well as consideration of the location in which a penalty occurs.

#References 1. Christensen, A. (n.d.). *Penalties and Their Effect on Goal-Scoring.* Symposium by ForagerOne. https://symposium.foragerone.com/utc-spring-research-and-arts-conference-2024/presentations/62311

2. Encyclopædia Britannica, inc. (2024, November 8). *Fighting in Hockey.* Encyclopædia Britannica. https://www.britannica.com/procon/fighting-in-hockey-debate
3. *Facts and figures: Power-play success rate on the rise.* NHL.com. (2018, February 4). https://www.nhl.com/news/nhl-facts-and-figures-power-play-success-rate-295620666