NEW-PROJECT-SET-4330.R

Justin A. Calcada

2021-12-11

Introduction

I am a huge fan of sports and they have been an important part of my life in helping me become the person I am today. As a result, I've always tried to find a way to mix sports and mathematics into a career. What I came up with was finding a career in sports statistics, which lead me to doing this project based around that.

In this case, with hockey being my favourite sport, I tried to find a question that I could ask which would help me to use the content of this course in an analysis of certain data. So what I did is I found team data for the 2018-19 NHL season.

In order to satisfy a categorical and causal response variable condition, I decided to ask the question of what it takes a team to make the playoffs and decided to view it from the standpoint of what variables could lead to that result and so the following analyses were performed.

Analysis

```r
nhl_team_stats <- read.csv("~/PROJECT STUFF/nhl_team_stats.csv")
# View(nhl_team_stats)
# nhl_team_stats
salarydata <- read.csv("~/salarydata.csv")
# View(salarydata)
```

```r
gg <- within(nhl_team_stats,
      {
          wins <- W
          loss <- L + OT
          goaldiff <- GF.GP - GA.GP
          GF_g <- GF.GP
          GA_g <- GA.GP
          GF_g_c <- scale(GF.GP,scale = FALSE)
          GA_g_c <- scale(GA.GP,scale = FALSE)
          pp <- PP.
          pk <- PK.
          pp_c <- scale(pp, scale = FALSE)
          pk_c <- scale(pp, scale = FALSE)
          shots_on <- Shots.GP
          shots_against <- SA.GP
          shots_on_c <- scale(Shots.GP, scale = FALSE)
          shots_against_c <- scale(SA.GP, scale = FALSE)
          shotdiff <- shots_on - shots_against
          shotdiff_c <- scale(shotdiff, scale = FALSE)
          faceoff <- FOW.
          goaldiff_c <- scale(goaldiff,scale = FALSE)
          faceoff_c <- scale(faceoff, scale = FALSE)

})
wins <- gg$wins
loss <- gg$loss
goaldiff <- gg$goaldiff
pp <- gg$pp
pk <- gg$pk
shots_on <- gg$shots_on
shots_against <- gg$shots_against
faceoff <- gg$faceoff
GF_g <- gg$GF_g
GA_g <- gg$GA_g
GF_g_c <- scale(gg$GF.GP,scale = FALSE)
GA_g_c <- scale(gg$GA.GP,scale = FALSE)
shots_on_c <- scale(gg$Shots.GP, scale = FALSE)
shots_against_c <- scale(gg$SA.GP, scale = FALSE)
faceoff_c <- gg$faceoff_c
pp_c <- gg$pp_c
pk_c <- gg$pk_c
shotdiff <- shots_on - shots_against
shotdiff_c <- scale(shotdiff, scale = FALSE)
ggg <- merge(gg,salarydata)
ggg
```

```
##                       Team   Season GP  W  L  T OT   P    P. RW ROW S.O.Win  GF
## 1           Anaheim Ducks 20182019 82 35 37 -- 10  80 0.488 27  32       3 196
## 2         Arizona Coyotes 20182019 82 39 35 --  8  86 0.524 30  35       4 209
## 3           Boston Bruins 20182019 82 49 24 --  9 107 0.652 38  47       2 257
## 4          Buffalo Sabres 20182019 82 33 39 -- 10  76 0.463 21  28       5 221
## 5          Calgary Flames 20182019 82 50 25 --  7 107 0.652 45  50       0 289
## 6     Carolina Hurricanes 20182019 82 46 29 --  7  99 0.604 39  44       2 243
## 7      Chicago Blackhawks 20182019 82 36 34 -- 12  84 0.512 24  33       3 267
## 8      Colorado Avalanche 20182019 82 38 30 -- 14  90 0.549 33  36       2 258
## 9  Columbus Blue Jackets 20182019 82 47 31 --  4  98 0.598 37  45       2 256
## 10          Dallas Stars 20182019 82 43 32 --  7  93 0.567 36  42       1 209
## 11     Detroit Red Wings 20182019 82 32 40 -- 10  74 0.451 20  29       3 224
## 12       Edmonton Oilers 20182019 82 35 38 --  9  79 0.482 24  32       3 229
## 13      Florida Panthers 20182019 82 36 32 -- 14  86 0.524 26  33       3 264
## 14     Los Angeles Kings 20182019 82 31 42 --  9  71 0.433 22  28       3 199
## 15        Minnesota Wild 20182019 82 37 36 --  9  83 0.506 33  36       1 210
## 16    Montréal Canadiens 20182019 82 44 30 --  8  96 0.585 37  41       3 246
## 17   Nashville Predators 20182019 82 47 29 --  6 100 0.610 38  43       4 236
## 18     New Jersey Devils 20182019 82 31 41 -- 10  72 0.439 24  28       3 219
## 19    New York Islanders 20182019 82 48 27 --  7 103 0.628 37  43       5 223
## 20      New York Rangers 20182019 82 32 36 -- 14  78 0.476 23  26       6 221
## 21       Ottawa Senators 20182019 82 29 47 --  6  64 0.390 23  29       0 242
## 22   Philadelphia Flyers 20182019 82 37 37 --  8  82 0.500 28  34       3 241
## 23   Pittsburgh Penguins 20182019 82 44 26 -- 12 100 0.610 37  42       2 271
## 24       San Jose Sharks 20182019 82 46 27 --  9 101 0.616 38  46       0 289
## 25       St. Louis Blues 20182019 82 45 28 --  9  99 0.604 36  42       3 244
## 26   Tampa Bay Lightning 20182019 82 62 16 --  4 128 0.780 49  56       6 319
## 27   Toronto Maple Leafs 20182019 82 46 28 --  8 100 0.610 40  46       0 286
## 28     Vancouver Canucks 20182019 82 35 36 -- 11  81 0.494 22  29       6 219
## 29  Vegas Golden Knights 20182019 82 43 32 --  7  93 0.567 36  40       3 246
## 30   Washington Capitals 20182019 82 48 26 --  8 104 0.634 39  44       4 274
## 31         Winnipeg Jets 20182019 82 47 30 --  5  99 0.604 38  45       2 270
##     GA GF.GP GA.GP  PP.  PK. Net.PP. Net.PK. Shots.GP SA.GP FOW.  faceoff_c
## 1  248  2.39  3.02 17.0 79.7    12.3    81.6     27.7  33.2 51.3  1.3129032
## 2  220  2.55  2.68 16.3 85.0    12.8    92.0     30.7  30.8 47.9 -2.0870968
## 3  212  3.13  2.59 25.9 79.9    19.9    82.8     32.7  29.5 50.7  0.7129032
## 4  268  2.70  3.27 19.5 80.9    15.7    82.3     32.9  33.0 47.9 -2.0870968
## 5  223  3.52  2.72 19.3 79.7    16.7    87.0     32.4  28.1 52.4  2.4129032
## 6  221  2.96  2.70 17.8 81.6    14.6    84.9     34.4  28.6 49.0 -0.9870968
## 7  291  3.26  3.55 20.2 72.7    17.2    75.3     32.5  34.8 49.5 -0.4870968
## 8  244  3.15  2.98 22.0 78.7    20.3    82.0     32.6  31.9 48.1 -1.8870968
## 9  231  3.12  2.82 15.4 85.0    12.7    89.0     31.6  29.5 50.3  0.3129032
## 10 200  2.55  2.44 21.0 82.8    20.1    84.5     30.7  31.6 51.9  1.9129032
## 11 272  2.73  3.32 18.1 77.1    14.8    80.0     29.2  33.7 50.8  0.8129032
## 12 271  2.79  3.30 21.2 74.8    18.0    78.9     29.2  31.7 48.4 -1.5870968
## 13 273  3.22  3.33 26.8 81.3    21.9    82.6     33.0  30.7 49.8 -0.1870968
## 14 259  2.43  3.16 15.8 76.5    10.0    78.3     28.8  31.4 50.5  0.5129032
## 15 233  2.56  2.84 20.3 81.7    18.7    83.8     31.3  29.5 49.4 -0.5870968
## 16 236  3.00  2.88 13.3 80.9    11.5    83.0     34.1  31.1 49.4 -0.5870968
## 17 212  2.88  2.59 12.9 82.1     9.8    85.3     32.7  30.0 51.5  1.5129032
## 18 271  2.67  3.30 17.7 84.3    13.8    88.2     30.3  31.6 49.4 -0.5870968
## 19 191  2.72  2.33 14.5 79.9    14.1    82.7     28.8  30.9 47.4 -2.5870968
## 20 267  2.70  3.26 19.4 78.2    17.6    80.8     29.2  33.8 46.9 -3.0870968
## 21 301  2.95  3.67 20.4 79.2    16.8    81.0     29.6  35.7 49.6 -0.3870968
## 22 280  2.94  3.41 17.1 78.5    12.4    80.2     31.5  32.5 54.7  4.7129032
## 23 238  3.30  2.90 24.6 79.7    18.0    85.1     33.3  33.3 50.6  0.6129032
## 24 258  3.52  3.15 23.7 80.8    19.9    84.6     33.0  28.3 50.3  0.3129032
## 25 220  2.98  2.68 21.1 81.5    18.1    83.6     31.8  28.6 51.4  1.4129032
## 26 221  3.89  2.70 28.2 85.0    27.1    89.5     32.0  32.1 51.2  1.2129032
## 27 249  3.49  3.04 21.8 79.9    17.5    82.4     33.4  33.1 53.0  3.0129032
## 28 248  2.67  3.02 17.1 81.1    13.9    83.9     29.7  31.7 49.5 -0.4870968
## 29 228  3.00  2.78 16.8 80.9    16.0    85.7     34.3  29.3 50.4  0.4129032
## 30 248  3.34  3.02 20.8 78.9    18.6    80.8     30.4  31.5 45.7 -4.2870968
## 31 243  3.29  2.96 24.8 79.2    22.0    83.2     31.2  33.4 50.7  0.7129032
##    goaldiff_c faceoff   shotdiff_c shotdiff shots_against_c shots_on_c
## 1  -0.6296774    51.3 -5.503225806     -5.5       1.7516129 -3.7516129
## 2  -0.1296774    47.9 -0.103225806     -0.1      -0.6483871 -0.7516129
## 3   0.5403226    50.7  3.196774194      3.2      -1.9483871  1.2483871
## 4  -0.5696774    47.9 -0.103225806     -0.1       1.5516129  1.4483871
## 5   0.8003226    52.4  4.296774194      4.3      -3.3483871  0.9483871
## 6   0.2603226    49.0  5.796774194      5.8      -2.8483871  2.9483871
## 7  -0.2896774    49.5 -2.303225806     -2.3       3.3516129  1.0483871
## 8   0.1703226    48.1  0.696774194      0.7       0.4516129  1.1483871
## 9   0.3003226    50.3  2.096774194      2.1      -1.9483871  0.1483871
## 10  0.1103226    51.9 -0.903225806     -0.9       0.1516129 -0.7516129
## 11 -0.5896774    50.8 -4.503225806     -4.5       2.2516129 -2.2516129
## 12 -0.5096774    48.4 -2.503225806     -2.5       0.2516129 -2.2516129
## 13 -0.1096774    49.8  2.296774194      2.3      -0.7483871  1.5483871
## 14 -0.7296774    50.5 -2.603225806     -2.6      -0.0483871 -2.6516129
## 15 -0.2796774    49.4  1.796774194      1.8      -1.9483871 -0.1516129
## 16  0.1203226    49.4  2.996774194      3.0      -0.3483871  2.6483871
## 17  0.2903226    51.5  2.696774194      2.7      -1.4483871  1.2483871
## 18 -0.6296774    49.4 -1.303225806     -1.3       0.1516129 -1.1516129
## 19  0.3903226    47.4 -2.103225806     -2.1      -0.5483871 -2.6516129
## 20 -0.5596774    46.9 -4.603225806     -4.6       2.3516129 -2.2516129
## 21 -0.7196774    49.6 -6.103225806     -6.1       4.2516129 -1.8516129
## 22 -0.4696774    54.7 -1.003225806     -1.0       1.0516129  0.0483871
## 23  0.4003226    50.6 -0.003225806      0.0       1.8516129  1.8483871
## 24  0.3703226    50.3  4.696774194      4.7      -3.1483871  1.5483871
## 25  0.3003226    51.4  3.196774194      3.2      -2.8483871  0.3483871
## 26  1.1903226    51.2 -0.103225806     -0.1       0.6516129  0.5483871
## 27  0.4503226    53.0  0.296774194      0.3       1.6516129  1.9483871
## 28 -0.3496774    49.5 -2.003225806     -2.0       0.2516129 -1.7516129
## 29  0.2203226    50.4  4.996774194      5.0      -2.1483871  2.8483871
## 30  0.3203226    45.7 -1.103225806     -1.1       0.0516129 -1.0516129
## 31  0.3303226    50.7 -2.203225806     -2.2       1.9516129 -0.2516129
##    shots_against shots_on       pk_c       pp_c   pk   pp        GA_g_c
## 1           33.2     27.7 -2.7032258 -2.7032258 79.7 17.0  0.0390322581
## 2           30.8     30.7 -3.4032258 -3.4032258 85.0 16.3 -0.3009677419
## 3           29.5     32.7  6.1967742  6.1967742 79.9 25.9 -0.3909677419
## 4           33.0     32.9 -0.2032258 -0.2032258 80.9 19.5  0.2890322581
## 5           28.1     32.4 -0.4032258 -0.4032258 79.7 19.3 -0.2609677419
## 6           28.6     34.4 -1.9032258 -1.9032258 81.6 17.8 -0.2809677419
## 7           34.8     32.5  0.4967742  0.4967742 72.7 20.2  0.5690322581
## 8           31.9     32.6  2.2967742  2.2967742 78.7 22.0 -0.0009677419
## 9           29.5     31.6 -4.3032258 -4.3032258 85.0 15.4 -0.1609677419
## 10          31.6     30.7  1.2967742  1.2967742 82.8 21.0 -0.5409677419
## 11          33.7     29.2 -1.6032258 -1.6032258 77.1 18.1  0.3390322581
## 12          31.7     29.2  1.4967742  1.4967742 74.8 21.2  0.3190322581
## 13          30.7     33.0  7.0967742  7.0967742 81.3 26.8  0.3490322581
## 14          31.4     28.8 -3.9032258 -3.9032258 76.5 15.8  0.1790322581
## 15          29.5     31.3  0.5967742  0.5967742 81.7 20.3 -0.1409677419
## 16          31.1     34.1 -6.4032258 -6.4032258 80.9 13.3 -0.1009677419
## 17          30.0     32.7 -6.8032258 -6.8032258 82.1 12.9 -0.3909677419
## 18          31.6     30.3 -2.0032258 -2.0032258 84.3 17.7  0.3190322581
## 19          30.9     28.8 -5.2032258 -5.2032258 79.9 14.5 -0.6509677419
## 20          33.8     29.2 -0.3032258 -0.3032258 78.2 19.4  0.2790322581
## 21          35.7     29.6  0.6967742  0.6967742 79.2 20.4  0.6890322581
## 22          32.5     31.5 -2.6032258 -2.6032258 78.5 17.1  0.4290322581
## 23          33.3     33.3  4.8967742  4.8967742 79.7 24.6 -0.0809677419
## 24          28.3     33.0  3.9967742  3.9967742 80.8 23.7  0.1690322581
## 25          28.6     31.8  1.3967742  1.3967742 81.5 21.1 -0.3009677419
## 26          32.1     32.0  8.4967742  8.4967742 85.0 28.2 -0.2809677419
## 27          33.1     33.4  2.0967742  2.0967742 79.9 21.8  0.0590322581
## 28          31.7     29.7 -2.6032258 -2.6032258 81.1 17.1  0.0390322581
## 29          29.3     34.3 -2.9032258 -2.9032258 80.9 16.8 -0.2009677419
## 30          31.5     30.4  1.0967742  1.0967742 78.9 20.8  0.0390322581
## 31          33.4     31.2  5.0967742  5.0967742 79.2 24.8 -0.0209677419
##           GF_g_c GA_g GF_g goaldiff loss wins Cap.space.left.over
## 1  -0.5906451613 3.02 2.39    -0.63   47   35                 Low
## 2  -0.4306451613 2.68 2.55    -0.13   43   39                High
## 3   0.1493548387 2.59 3.13     0.54   33   49              Medium
## 4  -0.2806451613 3.27 2.70    -0.57   49   33              Medium
## 5   0.5393548387 2.72 3.52     0.80   32   50                 Low
## 6  -0.0206451613 2.70 2.96     0.26   36   46                High
## 7   0.2793548387 3.55 3.26    -0.29   46   36              Medium
## 8   0.1693548387 2.98 3.15     0.17   44   38                High
## 9   0.1393548387 2.82 3.12     0.30   35   47              Medium
## 10 -0.4306451613 2.44 2.55     0.11   39   43                 Low
## 11 -0.2506451613 3.32 2.73    -0.59   50   32                 Low
## 12 -0.1906451613 3.30 2.79    -0.51   47   35                 Low
## 13  0.2393548387 3.33 3.22    -0.11   46   36              Medium
## 14 -0.5506451613 3.16 2.43    -0.73   51   31              Medium
## 15 -0.4206451613 2.84 2.56    -0.28   45   37              Medium
## 16  0.0193548387 2.88 3.00     0.12   38   44                High
## 17 -0.1006451613 2.59 2.88     0.29   35   47              Medium
## 18 -0.3106451613 3.30 2.67    -0.63   51   31                High
## 19 -0.2606451613 2.33 2.72     0.39   34   48                High
## 20 -0.2806451613 3.26 2.70    -0.56   50   32              Medium
## 21 -0.0306451613 3.67 2.95    -0.72   53   29                High
## 22 -0.0406451613 3.41 2.94    -0.47   45   37                High
## 23  0.3193548387 2.90 3.30     0.40   38   44                 Low
## 24  0.5393548387 3.15 3.52     0.37   36   46                 Low
## 25 -0.0006451613 2.68 2.98     0.30   37   45                 Low
## 26  0.9093548387 2.70 3.89     1.19   20   62                 Low
## 27  0.5093548387 3.04 3.49     0.45   36   46              Medium
## 28 -0.3106451613 3.02 2.67    -0.35   47   35                High
## 29  0.0193548387 2.78 3.00     0.22   39   43              Medium
## 30  0.3593548387 3.02 3.34     0.32   34   48                 Low
## 31  0.3093548387 2.96 3.29     0.33   35   47              Medium
```

An important factor for teams making the playoffs rests on how many games they can win throughout the regular season, the more they can win, the better their chances are of making the playoffs, so that is where we'll start our analysis.

Figuring out what causes wins is a very complex question in the NHL these days , as it can depend on many different variables, such as how well individual players perform throughout the season in many categories, how the schedule is formed–does a team have back to back games often or not, or even based on team stats overall. In our case here, we have data on overall team statistics so our analysis will be mainly focused on those, while keeping other factor contingencies in mind.

Like with many sports, scoring more goals than the other team is a tried-and-true way to win games and so, we'll see now if that is the case with our data:

```
library(latticeExtra)
```

```
## Warning: package 'latticeExtra' was built under R version 4.0.5
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 4.0.5
```

```
plot(goaldiff, wins, data = ggg,col = "blue", xlab = "Goal Differential Per Game",ylab = "Wins",
     main = "Team goal differential per game .vs. wins for 2018-19 NHL season")
```

```
## Warning in plot.window(...): "data" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter
```

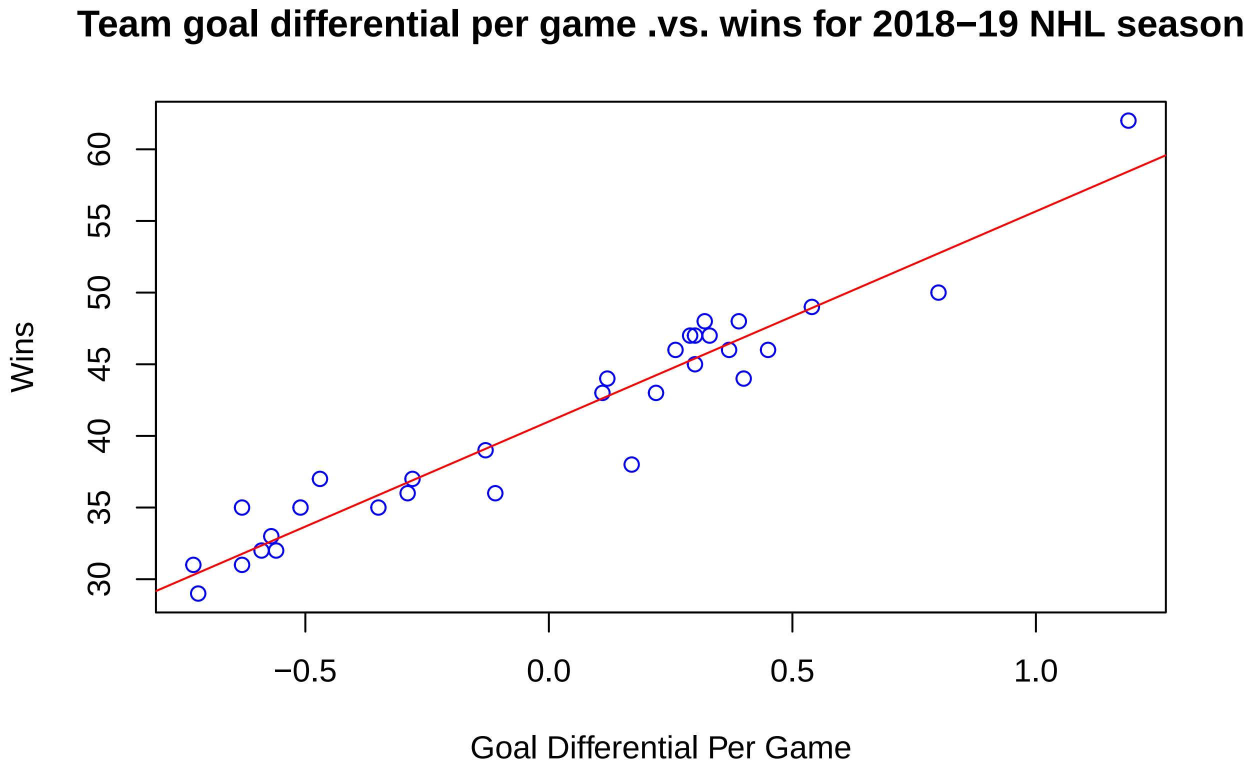
```
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a
## graphical parameter
```

```
## Warning in box(...): "data" is not a graphical parameter
```

```
## Warning in title(...): "data" is not a graphical parameter
```

```
abline(lm(wins~goaldiff), col = "red")
```



Arguably, the most important causal factor on wins would have to be goal differential per game (Goals scored per game - Goals against per game), and given the plot developed above, that exact trend is shown, that the higher the goal differential per game, the more wins a team will most likely secure. This lines up our intuition and provides a solid level of confidence that goals affect wins. Moving forward, our analysis will take into account other factors that may play a part in this relationship with goal differential and wins.

The first idea that we will tackle with that relationship in mind, starts at faceoff dot. Faceoffs in hockey are a situation where two players line up across from one another and wait until the referee drops the puck to fight for possession of it. These usually happen at the start of periods and are used to resume play after a stoppage. They can also occur in any zone of the ice so they can affect both the offensive and defensive side of the game. As a result, with winning these faceoffs being a way to gain possession of the puck, more time of possession in many sports gives a team more chances to score goals and acquire a win. We will now perform some statistical tests to see if faceoff win percentage, represented by the variable “faceoff”, is as important as we may think in terms of goals and wins.

```r
goaldiff_model <- glm(cbind(wins,loss) ~ goaldiff,data = ggg,family = binomial)
summary(goaldiff_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ goaldiff, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.23035  -0.16961  -0.00501   0.26092   1.02790
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.0008412  0.0402781   0.021    0.983
## goaldiff    0.7325941  0.0848443   8.635   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.5268  on 30  degrees of freedom
## Residual deviance:  6.3122  on 29  degrees of freedom
## AIC: 160.06
##
## Number of Fisher Scoring iterations: 3
```

```r
faceoff_model <- glm(cbind(wins,loss) ~ faceoff,data = ggg,family = binomial)
summary(faceoff_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ faceoff, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5990  -1.3135   0.0891   1.1227   4.5324
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.03687    1.09114  -1.867   0.0619 .
## faceoff      0.04075    0.02181   1.868   0.0618 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 80.029  on 29  degrees of freedom
## AIC: 233.78
##
## Number of Fisher Scoring iterations: 3
```

Individually, these summaries show that goal differential per game is highly significant in predicting wins with other variables free to change, but faceoff win percentage is less so, as it’s p-value is relatively much larger than that for goal differential per game in its effect on wins.

When putting these predictors together in a model and perform an Anova test, we find the following result:

```
basic_model <- glm(cbind(wins,loss) ~ faceoff + goaldiff,data = gg,family = binomial)
summary(basic_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ faceoff + goaldiff, family = binomial,
##     data = gg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.18296  -0.16525   0.02944   0.30222   1.01942
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.261661   1.122137  -0.233    0.816
## faceoff      0.005253   0.022439   0.234    0.815
## goaldiff     0.728952   0.086260   8.451   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.5268  on 30  degrees of freedom
## Residual deviance:  6.2575  on 28  degrees of freedom
## AIC: 162
##
## Number of Fisher Scoring iterations: 3
```

```
library(car)
```

```
## Warning: package ’car’ was built under R version 4.0.5
```

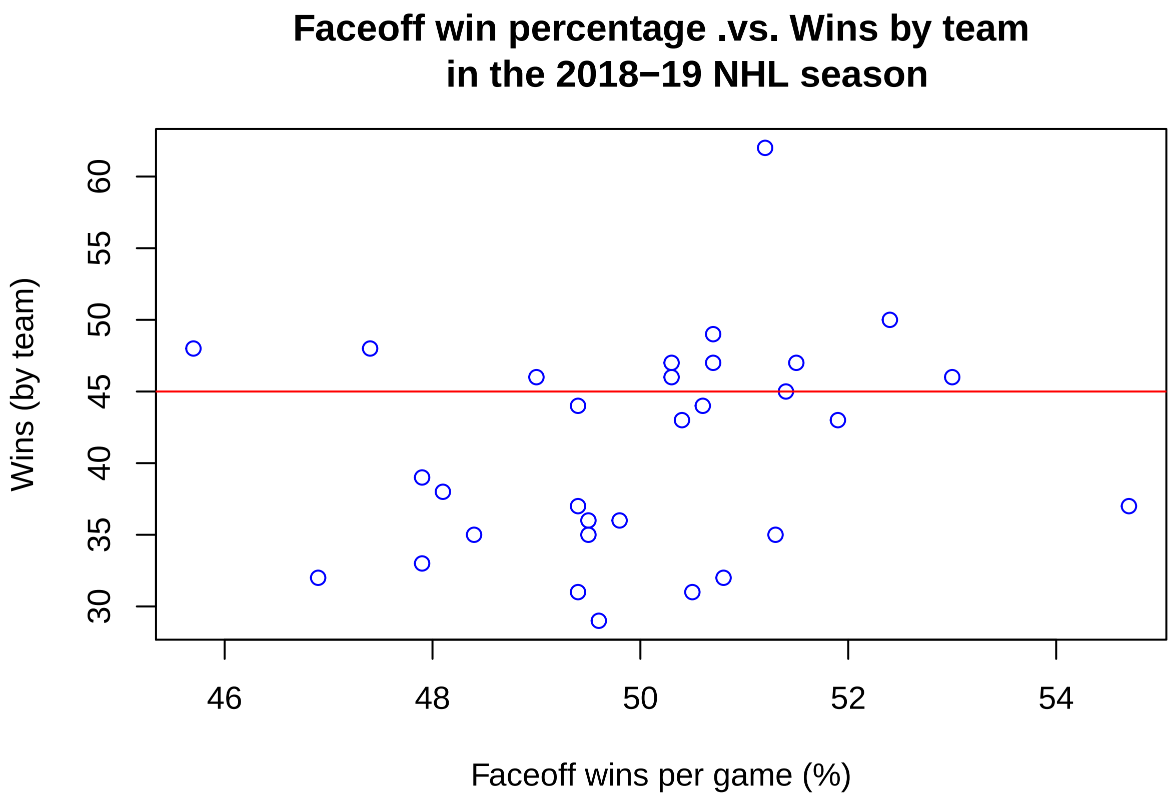
```
## Loading required package: carData
```

```
Anova(basic_model, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##           LR Chisq Df Pr(>Chisq)
## faceoff      0.055  1      0.815
## goaldiff    73.772  1     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that adding goal differential to a formula for faceoff win percentage makes it a much better model in having evidence of a causal effect on wins, but given that the p-value in Anova test for "faceoff" is not significant, we have evidence that faceoffs are not as important in gaining wins as we first thought. Essentially, this means that when holding the goal differential of each team constant, faceoff win percentage does not have a significant impact on a team winning a game. When we look at a plot of faceoff win percentage .vs. wins, we see a similar result, that faceoff win percentage does not significantly affect a team's number of wins or not.

```
plot(faceoff,wins, col = "blue", xlab = "Faceoff wins per game (%)",
     ylab = "Wins (by team)", main = "Faceoff win percentage .vs. Wins by team
     in the 2018-19 NHL season")
abline(h = 45, col = "red")
```



No relationship whatsoever. So now we can move onto check for interactions, as they may play a part in this result. For the interactions, we will be centering the predictors as the extrapolation to zero that happens when interpreting interactions does not hold any value here. Therefore, we achieve the following results:

```
int_c_model <- glm(cbind(wins,loss) ~ goaldiff_c*faceoff_c,data = ggg,family = binomial)
summary(int_c_model)
```

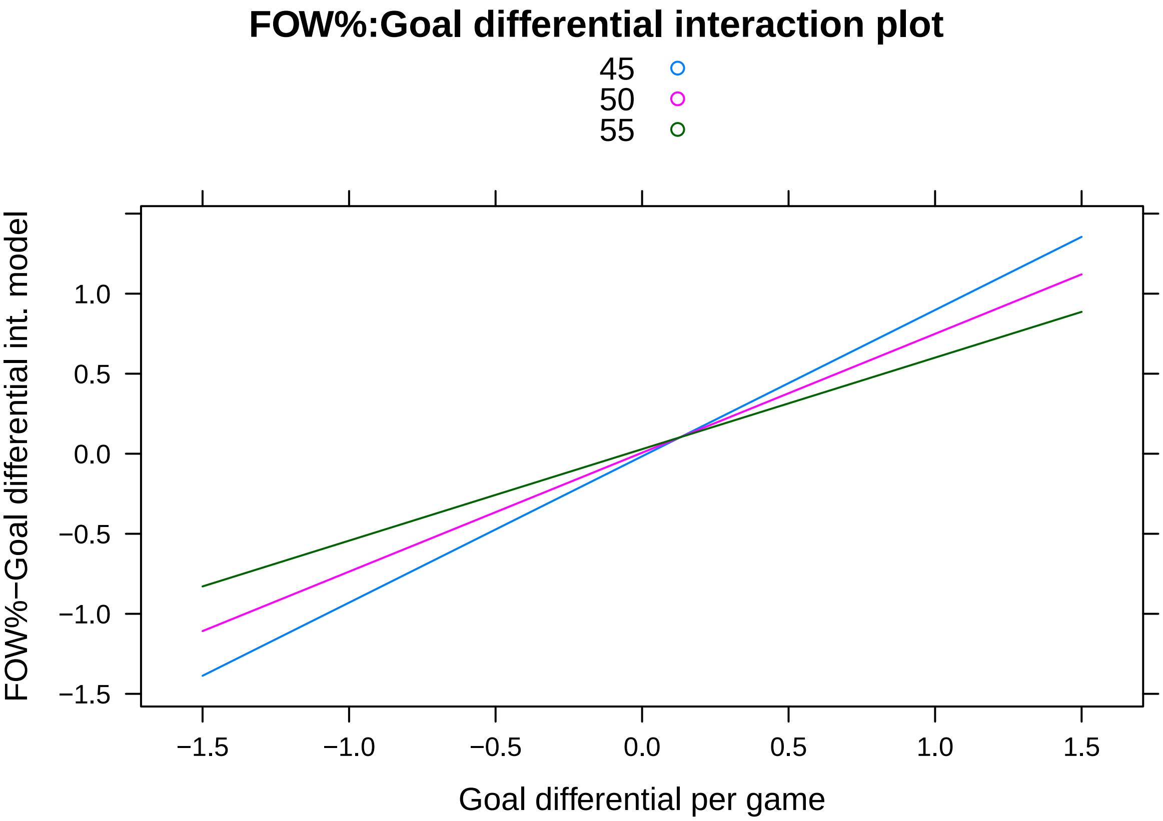
```
## 
## Call:
## glm(formula = cbind(wins, loss) ~ goaldiff_c * faceoff_c, family = binomial, 
##     data = ggg)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.27445  -0.15073   0.05156   0.22947   1.13254  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          0.006013   0.041023   0.147    0.883    
## goaldiff_c           0.743335   0.088712   8.379   <2e-16 ***
## faceoff_c            0.004502   0.022463   0.200    0.841    
## goaldiff_c:faceoff_c -0.034232   0.049663  -0.689    0.491    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 83.5268  on 30  degrees of freedom
## Residual deviance:  5.7822  on 27  degrees of freedom
## AIC: 163.53
## 
## Number of Fisher Scoring iterations: 3
```

```
Anova(int_c_model, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
## 
## Response: cbind(wins, loss)
##                      LR Chisq Df Pr(>Chisq)    
## goaldiff_c             73.772  1     <2e-16 ***
## faceoff_c               0.055  1     0.8150    
## goaldiff_c:faceoff_c    0.475  1     0.4906    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
fGAGF_model <- glm(cbind(wins,loss) ~ goaldiff*faceoff,data = ggg,family = binomial)

par(mfrow = c(1,1), oma = c(0,0,0,0))
pred <- expand.grid(goaldiff = c(-1.5,0,1.5), faceoff = c(45,50,55))
pred$fGAGF_model <- predict(fGAGF_model,newdata = pred,type = 'link')
library(latticeExtra)
xyplot(fGAGF_model~goaldiff, groups = faceoff,pred,type = "l", auto.key = TRUE,
       xlab = "Goal differential per game",ylab = "FOW%-Goal differential int. model"
       ,main = "FOW%:Goal differential interaction plot")
```

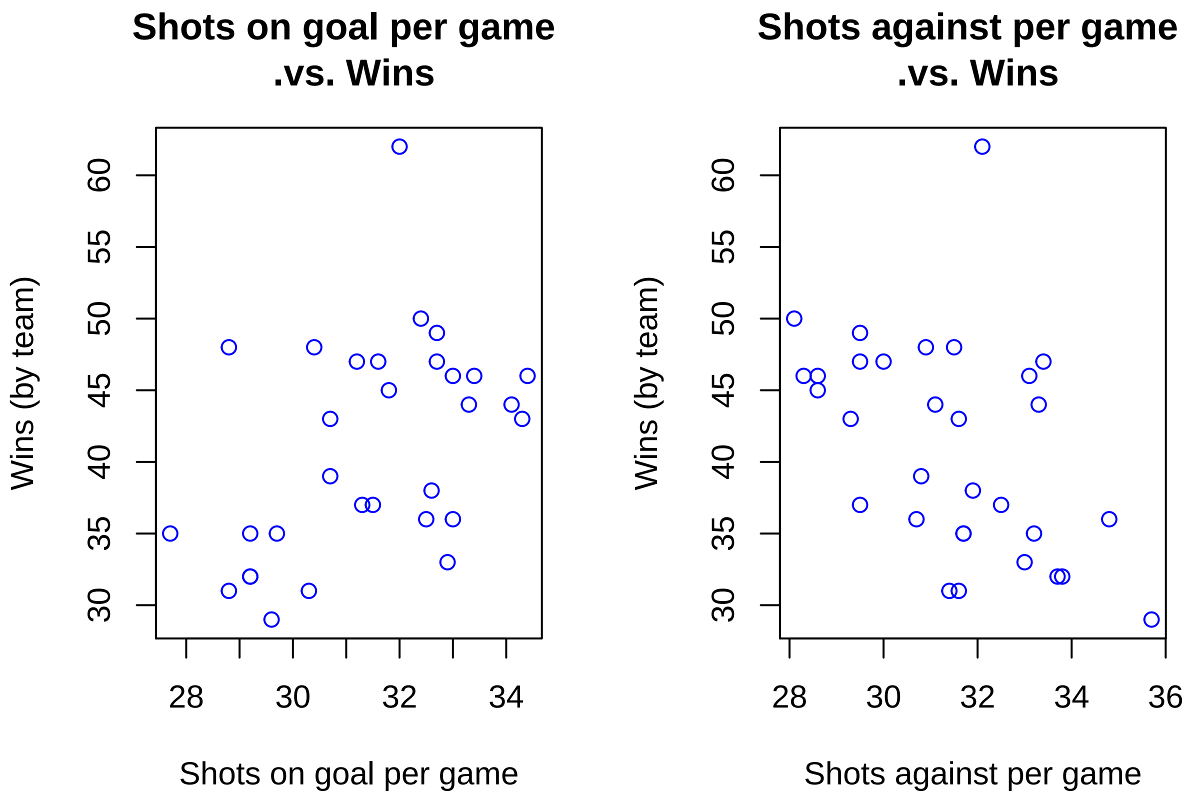


All of the tests just performed give us the idea that the interaction between goal differential per game and faceoff win percentage has no significant effect on wins, as the estimates of those parameters in these tests turned out to be highly non-significant, along with the fact that the plot for testing the interaction effect of goal differential per game and faceoff win percentage shows nearly parallel, if not parallel, lines which means that no significant interaction exists given our data based on these variables, giving us the idea that holding each variable at a certain level does not have a specific effect when changing the other.

With an individual significant effect on wins, but an opposite result when adding goal differential per game to the equation, we have evidence that goal differential must be a mediating factor on faceoff win's percentage effect on wins, but the strength of this result is not very high. This result allows us to believe that the result of faceoff win percentage affecting possession, in turn developing a better goal differential on the way to effecting wins is a definitely a possibility. Although, there are tons of factors that could have confounding effects on faceoff win percentage, such as which two players are facing off against one another–are they skilled at faceoffs or not–and being able to control for variables like that would show a better overview of the effect that faceoffs truly have on the game and their relationships within it.

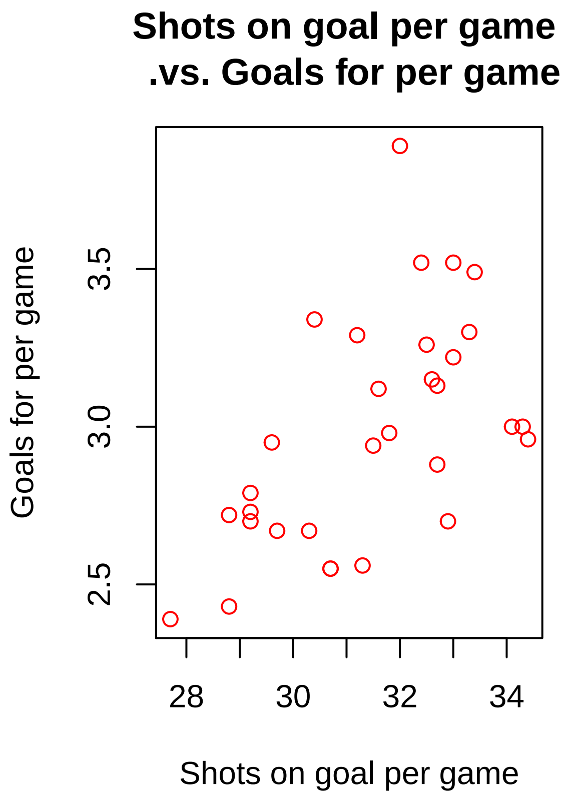
The ideas from this analysis on faceoff win percentage's effect on wins opens up so many more questions about how certain aspects of hockey can help teams gain wins and how other factors can affect their goal differential. With this in mind, faceoffs may not be significant on wins directly, but they do lead to more possession and with more possession of the puck, they have more of a chance of scoring goals, or keeping them out of their goal, and a major factor in these situations results from shots on goal and shots against respectively. Therefore, we can measure the effect of shots on goal per game in accordance with goals for per game and shots against per game in accordance with goals against per game to see how shots can be an effective measure of goals here, which again, may lead to wins. In this case, we are splitting up goals for and goals against when dealing with the shots variables, rather than dealing with overall goal differential because faceoffs effect offense and defense in complex ways and so, faceoffs effect goal differential rather than each separate variable, while the shots on goal affects offense, and shots against affects defense, individually.

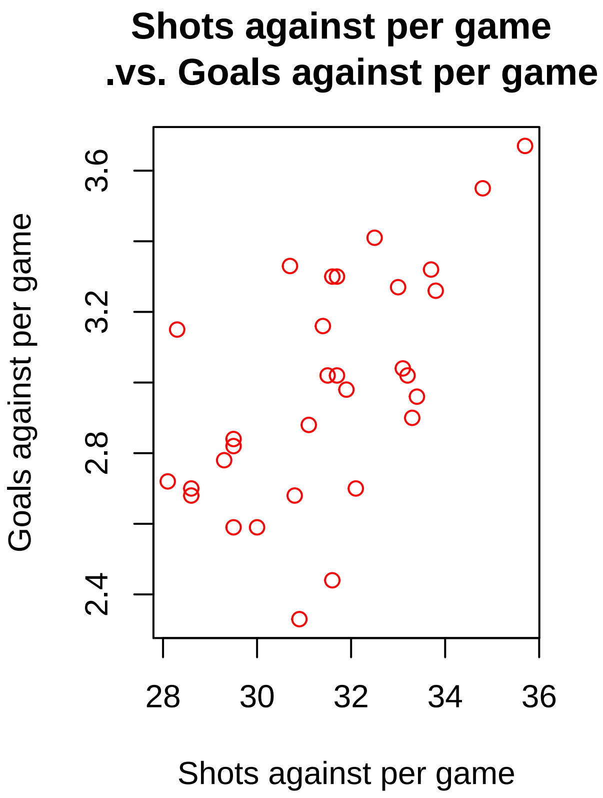
```
par(mfrow = c(1,2), oma = c(0,0,0,0))
plot(shots_on,wins,col = "blue", xlab = "Shots on goal per game", ylab = "Wins (by team)",
     main = "Shots on goal per game \n .vs. Wins")
plot(shots_against,wins, col = "blue", xlab = "Shots against per game", ylab = "Wins (by team)",
     main = "Shots against per game \n .vs. Wins")
```


Shots on goal per game .vs. Wins

Shots against per game .vs. Wins

```
plot(shots_on, GF_g, col = "red", xlab = "Shots on goal per game", ylab = "Goals for per game",
     main = "Shots on goal per game \n .vs. Goals for per game")
plot(shots_against,GA_g, col = "red", xlab = "Shots against per game", ylab = "Goals against per game",
     main = "Shots against per game \n .vs. Goals against per game")
```





In the plots above we see that shots on goal per game tend to have a, weak, but positive linear relationship with wins and goals for per game, and shots against per game show a negative linear relationship with that of wins, but a positive relationship with that of goals against per game, which is in line with our intuition, that more shots on goal, in general, lead to more goals, and more shots against, in general, lead to more goals against.

We'll now perform formal analyses to see their effects in these cases.

```
shots_on_model <- glm(cbind(wins,loss)~shots_on,data = ggg,family = binomial)
summary(shots_on_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ shots_on, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3706  -1.0694  -0.1303   0.7829   4.5295
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.87712    0.70659  -4.072 4.66e-05 ***
## shots_on     0.09148    0.02243   4.079 4.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 66.779  on 29  degrees of freedom
## AIC: 220.53
##
## Number of Fisher Scoring iterations: 3
```

```
sht_goal_model <- lm(GF_g~shots_on)
summary(sht_goal_model)
```

```
##
## Call:
## lm(formula = GF_g ~ shots_on)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44584 -0.24563  0.00615  0.14111  0.84681
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.60646    0.96113  -0.631 0.532993
## shots_on     0.11405    0.03051   3.738 0.000811 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3025 on 29 degrees of freedom
## Multiple R-squared:  0.3252, Adjusted R-squared:  0.3019
## F-statistic: 13.97 on 1 and 29 DF,  p-value: 0.0008105
```

The shots_on model that tests if the predictors individually are highly significant, shows that it is important in its effect on wins when the values of the other factors in this data are changing with them, but testing the effect of shots on goal per game on goals for per game shows that shots on goal have a significant effect on goals, again, when not keeping other factors constant, which develops the idea that goals for per game could possibly be a mediating factor on shots on goal per game, and that would correlate with our intuition stated above.

```
acc_add_model <- glm(cbind(wins,loss)~GF_g + shots_on,data = ggg,family = binomial)
summary(acc_add_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ GF_g + shots_on, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4285  -0.8846   0.1448   0.8632   2.5285
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.50084    0.71216  -3.512 0.000445 ***
## GF_g         0.64336    0.13799   4.663 3.12e-06 ***
## shots_on     0.01855    0.02731   0.679 0.496869    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 44.709  on 28  degrees of freedom
## AIC: 200.46
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(acc_add_model, test ="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##          LR Chisq Df Pr(>Chisq)    
## GF_g      22.0705  1  2.628e-06 ***
## shots_on   0.4617  1     0.4969    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
accuracy_model <- glm(cbind(wins,loss)~GF_g_c*shots_on_c,data = ggg,family = binomial)
summary(accuracy_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ GF_g_c * shots_on_c, family = binomial, 
##     data = ggg)
##
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.4369  -0.8874   0.1466   0.8602   2.5281  
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        0.001494   0.051119   0.029    0.977    
## GF_g_c             0.643760   0.138411   4.651  3.3e-06 ***
## shots_on_c         0.018064   0.030357   0.595    0.552    
## GF_g_c:shots_on_c -0.003288   0.089226  -0.037    0.971    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 44.707  on 27  degrees of freedom
## AIC: 202.45
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(accuracy_model, test ="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##                     LR Chisq Df Pr(>Chisq)
## GF_g_c               22.0705  1  2.628e-06 ***
## shots_on_c            0.4617  1     0.4969
## GF_g_c:shots_on_c     0.0014  1     0.9706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With these two tests, we see that the interaction between shots on goal per game and goals for per game is insignificant. Since this is a test of accuracy, as goals for per game represents how many shots became goals, and so, accuracy (as in the interaction) proves to not matter as much to wins, or has a non-significant effect on it.

Furthermore, with the Anova tests, we see that predicting the cause of wins with these two predictors in the same model, displays the following idea: when goals are kept constant between all teams, our model is showing that more shots on goal on it's own does not necessarily play a big role in predicting wins, although with keeping shots on goal per game constant between teams, we see that goals for per game shows us a measure of how accurate or how good players are at scoring on certain teams, which may give us the idea that including better goal scorers, or more accurate shooters can help a team win, and gives us evidence that goals for per game is the mediating factor on shots on goal per game that we first thought about. Now it makes sense why they pay the high level goal scorers the big bucks.

Overall, this provides evidence that shots on goal per game has a significant effect on goals for per game, but only has an effect on wins through the goals for per game variable. Shots lead to goals and goals lead to wins, simply put.

```
shots_against_model <- glm(cbind(wins,loss)~shots_against,data = ggg,family = binomial)
summary(shots_against_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ shots_against, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2390  -1.0806  -0.2525   0.7083   5.0057
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.78066    0.65685   4.233 2.30e-05 ***
## shots_against -0.08842    0.02085  -4.241 2.23e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 65.389  on 29  degrees of freedom
## AIC: 219.14
##
## Number of Fisher Scoring iterations: 3
```

```
goals_against_model <- glm(cbind(wins,loss)~GA_g,data = ggg,family = binomial)
summary(goals_against_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ GA_g, family = binomial, data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5796  -0.8920  -0.1845   0.7018   3.7704
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.3685     0.3780   6.266 3.70e-10 ***
## GA_g         -0.7946     0.1261  -6.301 2.96e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 43.069  on 29  degrees of freedom
## AIC: 196.82
##
## Number of Fisher Scoring iterations: 3
```

```
shtag_goalag_model <- lm(GA_g~shots_against)
summary(shtag_goalag_model)
```

```
##
## Call:
## lm(formula = GA_g ~ shots_against)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59489 -0.16591  0.01874  0.15717  0.49101
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.23519    0.77302  -0.304 0.763109
## shots_against  0.10227    0.02453   4.168 0.000253 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2621 on 29 degrees of freedom
## Multiple R-squared:  0.3747, Adjusted R-squared:  0.3531
## F-statistic: 17.37 on 1 and 29 DF,  p-value: 0.0002529
```

These tests show us those same results that we have seen above, that shots against per game has a significant effect on goals against per game and wins, so we can see that goals against per game must be a mediating factor between the effect of shots against per game on wins. To ensure that this is the case, we can perform the following test in an overall model:

```
defend_model <- glm(cbind(wins,loss)~GA_g + shots_against, data = ggg, family = binomial)
summary(defend_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ GA_g + shots_against, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6184  -0.8263  -0.1461   0.6496   3.8623
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.62441    0.66034   3.974 7.06e-05 ***
## GA_g          -0.74895    0.15868  -4.720 2.36e-06 ***
## shots_against -0.01247    0.02635  -0.473    0.636
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 42.846  on 28  degrees of freedom
## AIC: 198.59
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(defend_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##               LR Chisq Df Pr(>Chisq)
## GA_g           22.5439  1  2.054e-06 ***
## shots_against   0.2237  1     0.6362
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
defend_int_model <- glm(cbind(wins,loss)~GA_g_c*shots_against_c, data = ggg, family = binomial)
summary(defend_int_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ GA_g_c * shots_against_c, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6718  -0.8235  -0.0543   0.6482   3.8070  
## 
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)    
## (Intercept)              0.009534   0.046887   0.203    0.839    
## GA_g_c                  -0.738080   0.160969  -4.585 4.53e-06 ***
## shots_against_c         -0.010820   0.026671  -0.406    0.685    
## GA_g_c:shots_against_c -0.026581   0.066617  -0.399    0.690    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 42.686  on 27  degrees of freedom
## AIC: 200.43
## 
## Number of Fisher Scoring iterations: 3
```

```
Anova(defend_int_model)
```

```
## Analysis of Deviance Table (Type II tests)
## 
## Response: cbind(wins, loss)
##                        LR Chisq Df Pr(>Chisq)    
## GA_g_c                  22.5439  1  2.054e-06 ***
## shots_against_c          0.2237  1     0.6362    
## GA_g_c:shots_against_c   0.1595  1     0.6896    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Once again, the results from the models above are in line with what we first assumed ot be the case.

```
shots_add_model <- glm(cbind(wins,loss)~shots_on + shots_against, data = ggg, family = binomial)
summary(shots_add_model)
```

```
## 
## Call:
## glm(formula = cbind(wins, loss) ~ shots_on + shots_against, family = binomial, 
##     data = ggg)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8416  -0.8040  -0.4147   0.7075   4.7830  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)   
## (Intercept)    0.03256    1.22672   0.027  0.97883   
## shots_on       0.06436    0.02430   2.648  0.00809 **
## shots_against -0.06541    0.02258  -2.896  0.00377 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 58.356  on 28  degrees of freedom
## AIC: 214.1
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(shots_add_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##               LR Chisq Df Pr(>Chisq)
## shots_on        7.0330  1   0.008002 **
## shots_against   8.4226  1   0.003706 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Given these tests, we see that the model that includes both shots on goal per game and shots against per game fits the data better than the models that only include the predictors separately, which develops the interesting idea that much like with goal differential, shot differential per game (shots on goal per game - shots against per game) is a more important predictor than the two predictors separately. Also, the relationship between these two variables should be included in the model because if a team were to have more shots on goal per game, they would be more likely to be in the offensive zone more often than in their own defensive zone, which is a measure of the skill of a given team or their offensive abilities in a sense. Of course there are times when both teams that face one another may have a low number of shots on goal each, or a high number of shots on goal each, but that is not as common in this sport, so we are not as worried about it in the analysis. This brings up many more questions as well, as this shot differential could be affected by injuries throughout the season, or even what teams play against one another, as a weaker team would tend to be outshot when facing a stronger team, and so it may well be very dependent on factors that we cannot explore within this dataset.

```
shotdiff <- shots_on - shots_against
shotdiff_model <- glm(cbind(wins,loss)~shotdiff,data = ggg, family = binomial)
summary(shotdiff_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ shotdiff, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8391  -0.8040  -0.4114   0.7103   4.7803
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.0002232  0.0398653  -0.006    0.996
## shotdiff     0.0649129  0.0130115   4.989 6.07e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

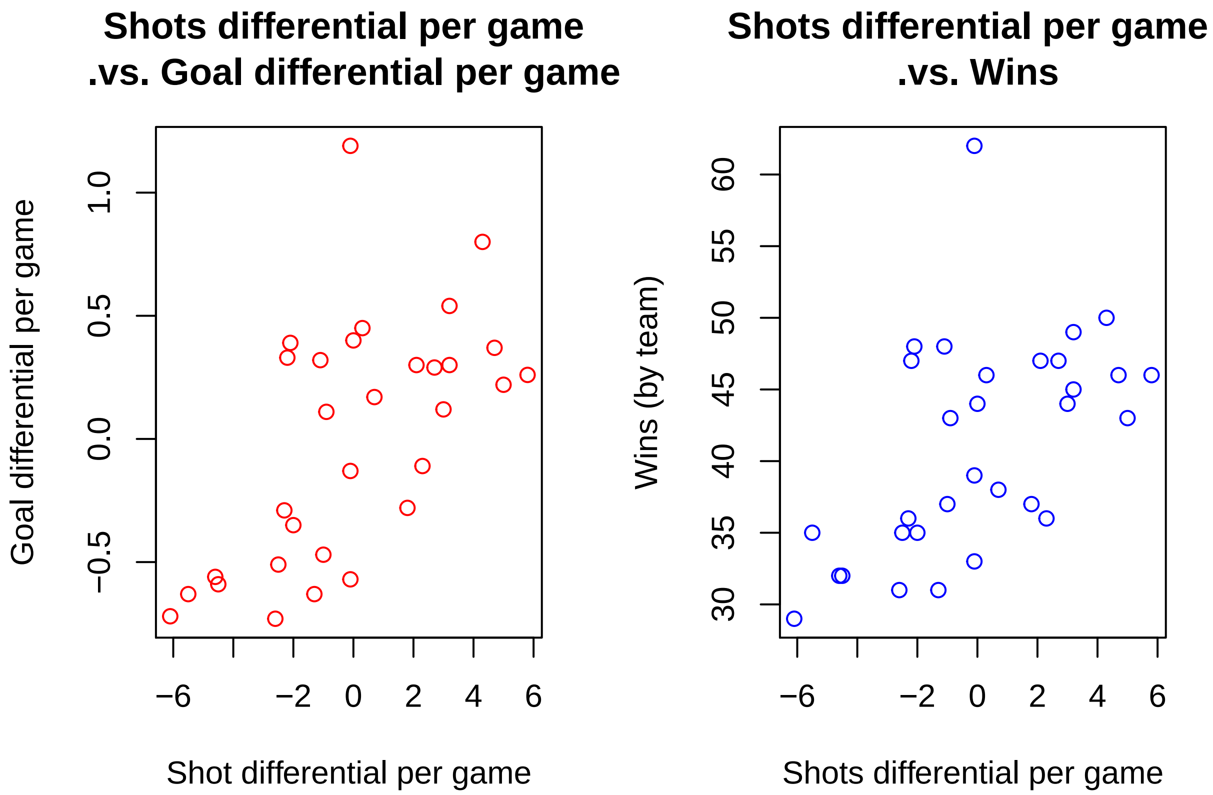
```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 58.357  on 29  degrees of freedom
## AIC: 212.1
##
## Number of Fisher Scoring iterations: 3
```

```
summary(lm(goaldiff~shotdiff))
```

```
##
## Call:
## lm(formula = goaldiff ~ shotdiff)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55965 -0.26576 -0.09523  0.21395  1.20035
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0006359  0.0706992  -0.009 0.992886
## shotdiff     0.0971156  0.0229312   4.235 0.000211 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3936 on 29 degrees of freedom
## Multiple R-squared:  0.3821, Adjusted R-squared:  0.3608
## F-statistic: 17.94 on 1 and 29 DF,  p-value: 0.0002106
```

These final tests, with an even smaller AIC, provides more evidence that shot differential per game has a significant effect on wins (while other variables are permitted to change), but through goal differential per game (a mediating factor of it). This can also be seen in the following plots:

```
par(mfrow = c(1,2), oma = c(0,0,0,0))
plot(shotdiff,goaldiff, col = "red", xlab = "Shot differential per game",
     ylab = "Goal differential per game",
     main = "Shots differential per game \n .vs. Goal differential per game")
plot(shotdiff,wins, col = "blue", xlab = "Shots differential per game",
     ylab = "Wins (by team)",
     main = "Shots differential per game \n .vs. Wins")
```

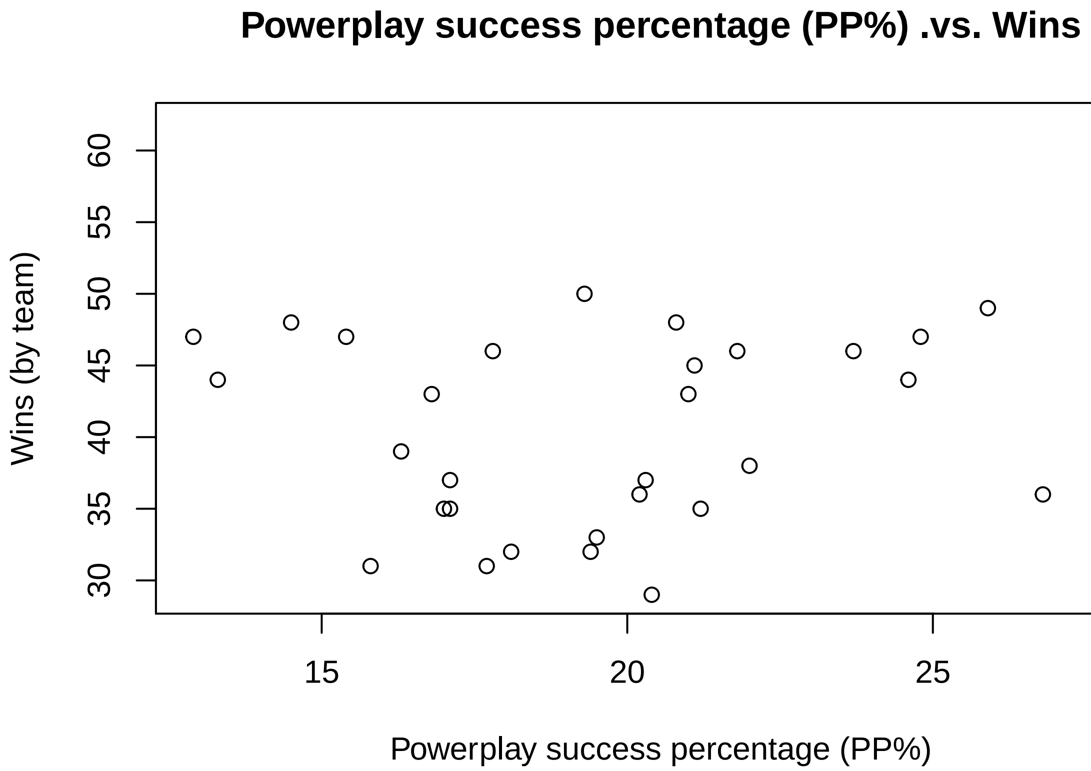


Overall, we see that, just simply through these associations, that shots on goal and shots against paint a vivid picture of the game of hockey and how we can truly interpret the scoring ability of a team, along with why wins would follow as a result. Our analysis does not end here though, we'll now move on to our final factor in our quest of figuring out what it takes to win hockey games in the NHL. In all sports, momentum swings can have a very significant effect on the direction of the game and can revert the score in the blink of an eye, to put a losing team in a winning position. In hockey, these so-called momentum swings are largely the work of powerplays, in which one team is given an advantage over the other by virtue of the opposition losing a player for a certain amount of time due to that player violating a rule. With a team gaining a man-advantage, or sometimes a 2-man advantage over the opposition, scoring a goal becomes an easier task and so having an easier pathway to a goal can cause a team to score more, which in turn, allows for a higher probability of winning games.

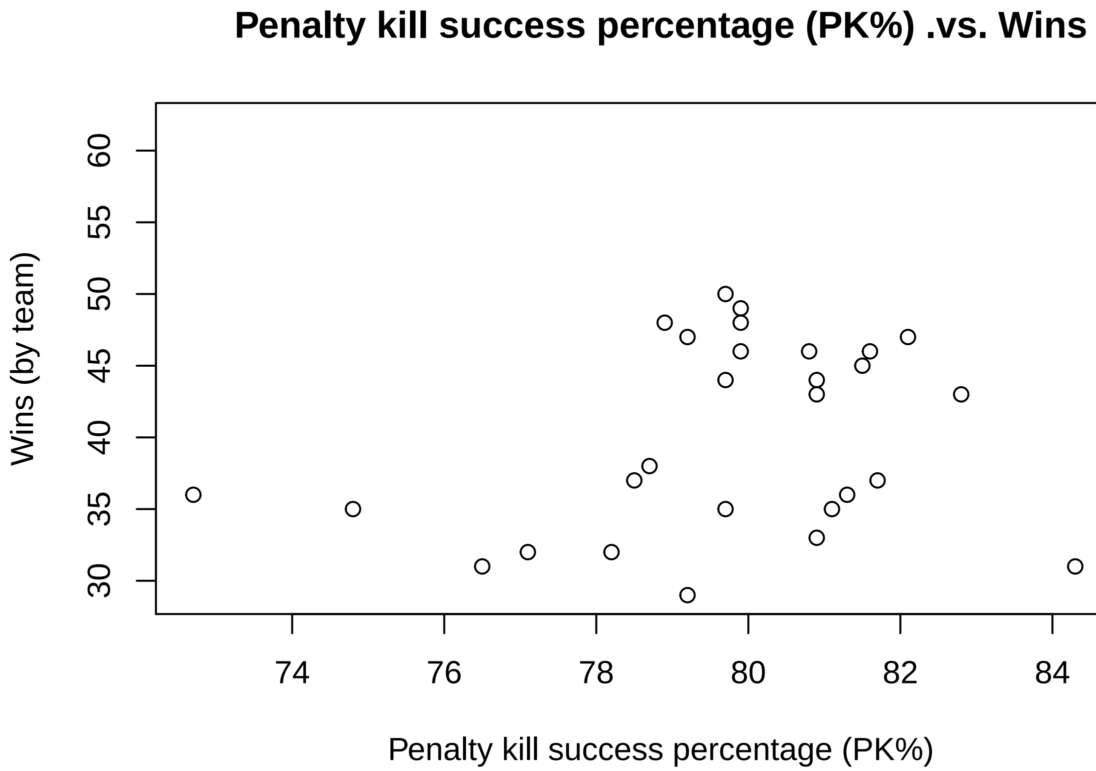
Therefore, we can discuss the factors of powerplay success percentage (PP%), and success is determined when a team scores a goal on the powerplay , and penalty kill success percentage (PK%), in which success is determined by whether a team does not let up a goal when facing a powerplay, to see how they may affect wins.

We can start with a simple plot to see their relationship:

```
plot(pp,wins,col = "black", xlab = "Powerplay success percentage (PP%)",
     ylab = "Wins (by team)",
     main = "Powerplay success percentage (PP%) .vs. Wins")
```



```
plot(pk,wins,col = "black", xlab = "Penalty kill success percentage (PK%)",
     ylab = "Wins (by team)",
     main = "Penalty kill success percentage (PK%) .vs. Wins")
```



Simply from the plots alone, we see that PP% does not have a linear relationship with wins, but the data points tend to mostly fall below about 50 wins with a varied distribution and the PK% also do not show a linear relationship with wins alone.

When looking at these variables based on real life applications, this makes sense as the values that are developed for PP% are dependent on goals for, as a successful powerplay is designated by a team scoring on the opponent within the duration of the powerplay, along with how long the powerplays are on average and how many powerplays the teams get per game. On the other hand, the PK% is affected by goals against, as a failed penalty kill comes from the fact that a team is scored on by their opponent within the duration of the powerplay, and so that affects the percentage. In effect, the less they get scored on on those penalty kills, the higher their PK% will be, alongside with the other factors that also affect PP% that have an effect on PK% as well.

Overall, there is definitely a confounding factor that is affecting the causal relationship between PP% and PK% and there may be interaction that a model which includes them may need to develop a proper model.

```
pp_model <- glm(cbind(wins,loss)~pp, data = ggg, family = binomial)
summary(pp_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pp, family = binomial, data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7530  -1.3361   0.0762   1.1833   3.7617
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.52189    0.21088  -2.475   0.0133 *
## pp           0.02649    0.01051   2.520   0.0117 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 77.158  on 29  degrees of freedom
## AIC: 230.91
##
## Number of Fisher Scoring iterations: 3
```

```
pk_model <- glm(cbind(wins,loss)~pk, data = ggg, family = binomial)
summary(pk_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pk, family = binomial, data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1778  -1.2972   0.2052   0.9314   3.6574
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.20169    1.18410  -3.548 0.000388 ***
## pk           0.05236    0.01475   3.551 0.000384 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 70.801  on 29  degrees of freedom
## AIC: 224.55
##
## Number of Fisher Scoring iterations: 3
```

Both individual predictor models seem to be significant predictors of wins when nothing else is held constant, but we’ll need more analysis to truly understand the relationship, as many factors changing can affect this result. We’ll now see how goals for and against can affect PP% and PK%:

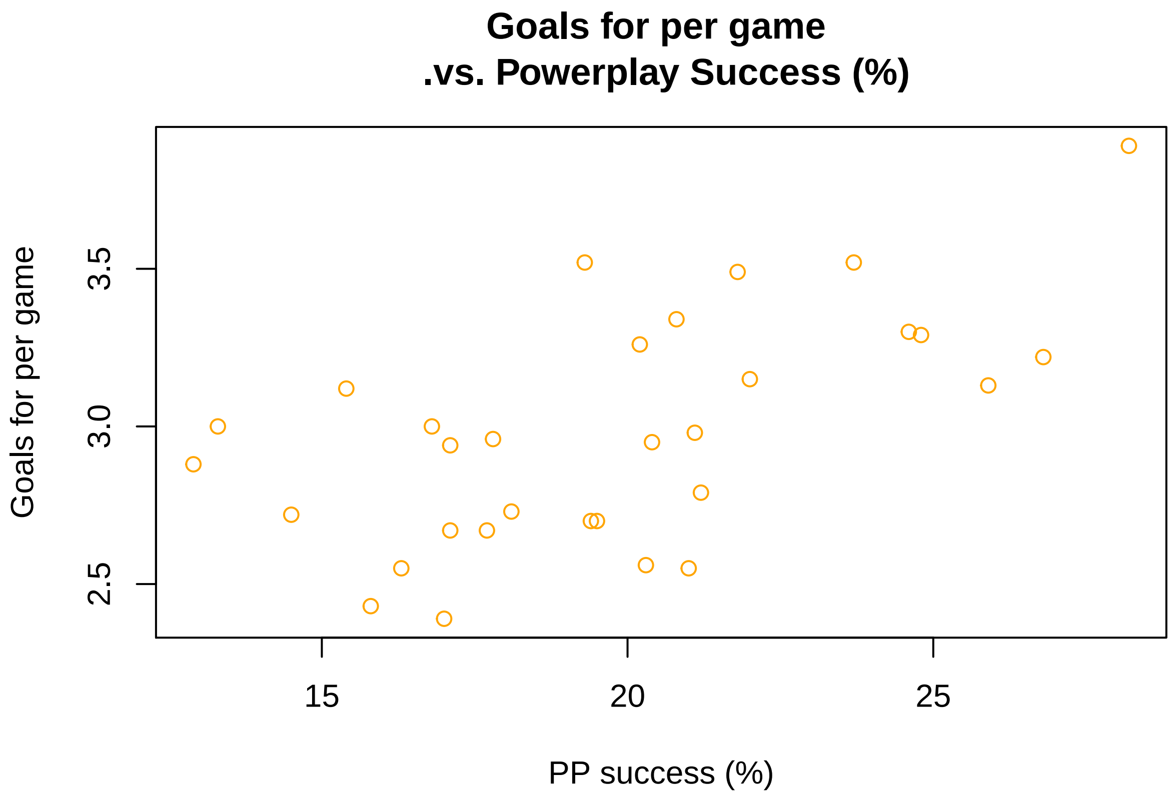
```
par(mfrow = c(1,1), oma = c(0,0,0,0))
pp_GF_model <- glm(cbind(wins,loss)~pp + GF_g,data = ggg,family = binomial)
summary(pp_GF_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pp + GF_g, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.50007  -0.85928  -0.00769   0.78271   2.22921
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.06027    0.34086  -6.044 1.50e-09 ***
## pp          -0.01885    0.01312  -1.437    0.151
## GF_g         0.81595    0.14079   5.796 6.81e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 43.103  on 28  degrees of freedom
## AIC: 198.85
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(pp_GF_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##      LR Chisq Df Pr(>Chisq)
## pp      2.067  1     0.1505
## GF_g   34.054  1  5.359e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
plot(pp, GF_g, col = "orange", xlab = "PP success (%)", ylab = "Goals for per game",
     main = "Goals for per game \n .vs. Powerplay Success (%)")
```



```r
pk_GA_model <- glm(cbind(wins,loss)~pk + GA_g, data = ggg, family = binomial)
summary(pk_GA_model)
```

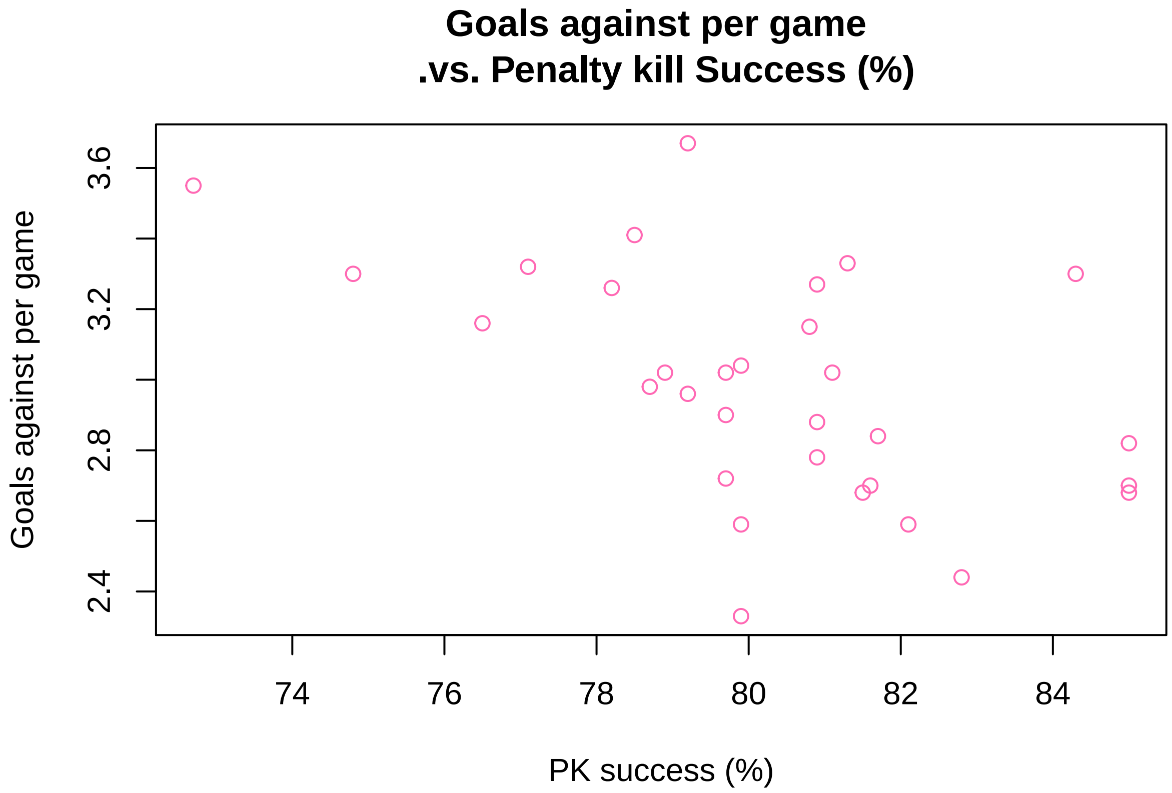
```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pk + GA_g, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5951  -0.8806  -0.0999   0.6575   3.6989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.929548   1.667593   1.157    0.247
## pk           0.004691   0.017362   0.270    0.787
## GA_g        -0.773617   0.147984  -5.228 1.72e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 42.996  on 28  degrees of freedom
## AIC: 198.74
```

```
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(pk_GA_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##      LR Chisq Df Pr(>Chisq)
## pk      0.073  1      0.787
## GA_g   27.805  1  1.342e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(pk,GA_g, col = "hot pink", xlab = "PK success (%)", ylab = "Goals against per game",
     main = "Goals against per game \n .vs. Penalty kill Success (%)")
```



```
pp_GA_model <- glm(cbind(wins,loss)~pp + GA_g,data = ggg,family = binomial)
summary(pp_GA_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pp + GA_g, family = binomial,
##     data = ggg)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8192  -0.7583  -0.3503   0.9857   2.4240
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.82543    0.41161   4.435 9.22e-06 ***
## pp           0.03514    0.01070   3.286  0.00102 **
## GA_g        -0.84460    0.12726  -6.637 3.20e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 32.205  on 28  degrees of freedom
## AIC: 187.95
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(pp_GA_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##      LR Chisq Df Pr(>Chisq)
## pp     10.864  1  0.0009804 ***
## GA_g   44.953  1  2.019e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
pk_GF_model <- glm(cbind(wins,loss)~pk + GF_g,data = ggg,family = binomial)
summary(pk_GF_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pk + GF_g, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.34249  -0.53317   0.09299   0.61072   2.43898
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.12667    1.24073  -4.938 7.90e-07 ***
## pk           0.05064    0.01486   3.407 0.000657 ***
## GF_g         0.69251    0.11465   6.040 1.54e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 33.472  on 28  degrees of freedom
## AIC: 189.22
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(pk_GF_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##      LR Chisq Df Pr(>Chisq)
## pk     11.698  1  0.0006255 ***
## GF_g   37.329  1  9.977e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results from these models are very interesting as when we add goals for per game to the model with PP%, we see that the effect of PP% on wins becomes non-significant, which develops the idea that goals for per game may be a mediator rather than a confounding factor on PP%, and the same goes for PK% on wins when adding goals against per game to that respective model. Although, when we instead apply goals for per game to the model with PK% and goals against per game to the model with PP%, each predictor in each model becomes significant. Each of these new models reinforces the ideas from the first two and those are as follows: when we keep goals against per game constant between all teams, we can see the effect that goals for has on PP% and that when goals for are allowed to change with that of PP%, we see a significant effect that each of them have on a team winning. In other words, having more success on the powerplay (higher PP%) means scoring more goals and affecting wins in a significantly positive way. A similar result holds when holding goals for per game constant and allowing PK% to change as goals against per game changes because this shows that the more goals against per game can have an effect on a team's ability to defend their goal on the penalty kill, or said another way, if teams tend to let in more goals, they are not as good defensively as another team and their PK% will suffer due to their weak defensive abilities leading to a team having a lower chance of winning. Essentially, goals for per game is a mediating factor of PP%'s effect on wins and goals against per game is a mediating factor on PK%'s effect on wins, so a team improving their proficiency on their powerplays and penalty kills give teams a better chance to win. As a sidenote, these statistics do depend on what teams were playing against one another and the number of powerplays that resulted in each of those games for each team, along with the factor of which team was the home team as fan support plays a motivating role in hockey games that helps teams play better in front of their home fans; these variables could have an effect on these results, but what we have shown here is the general situation with the data that we have.

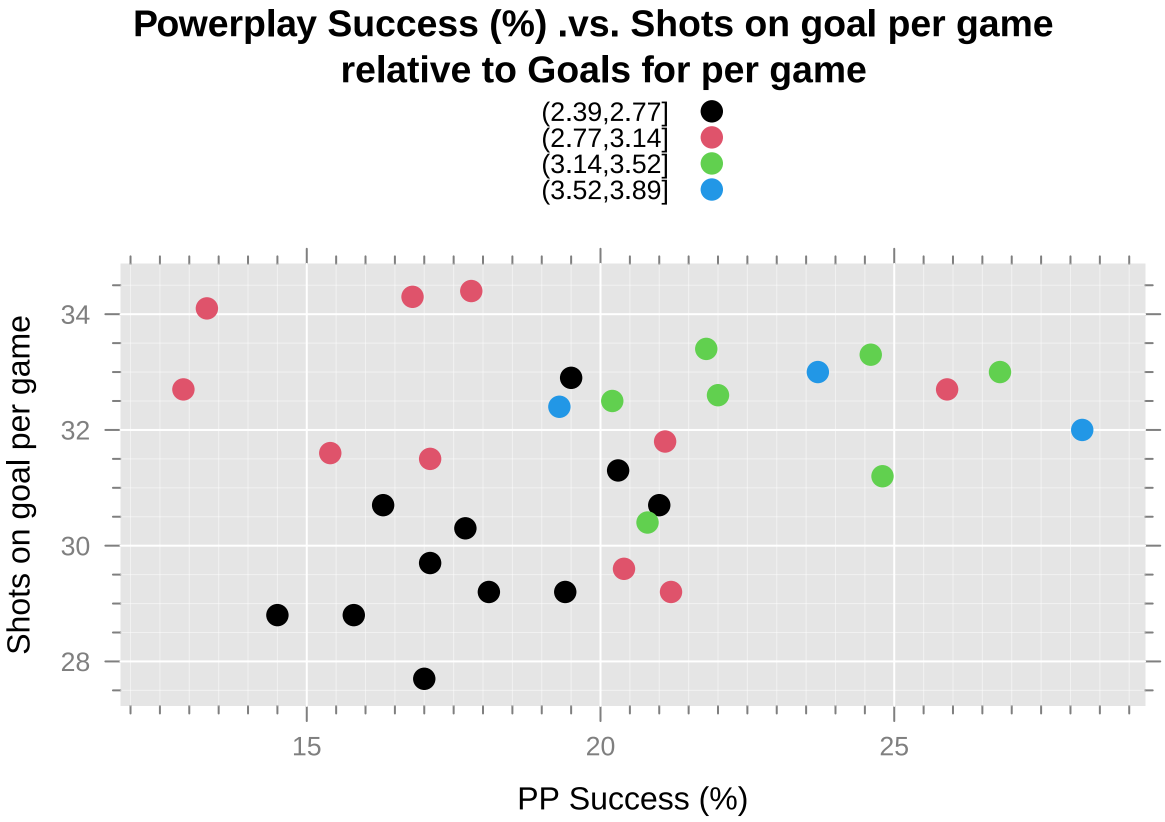
This develops an interesting idea. It seems that since goals against affects PP% more than PK% on an effect on wins, we have evidence to suggest that goals against affect PP% and wins simultaneously, and that goals for affect wins and PK% simultaneously and so we need to include those predictors in the model when discovering evidence of a causal effect of PP% and PK% on wins.

Given that we see that goals for per game and goals against per game seem to be mediators for PP% and PK%, respectively, it brings up another question: If goals have an effect on PP% and PK%, and shots affect goals, would shots on goal per game and shots against per game affect these factors too?

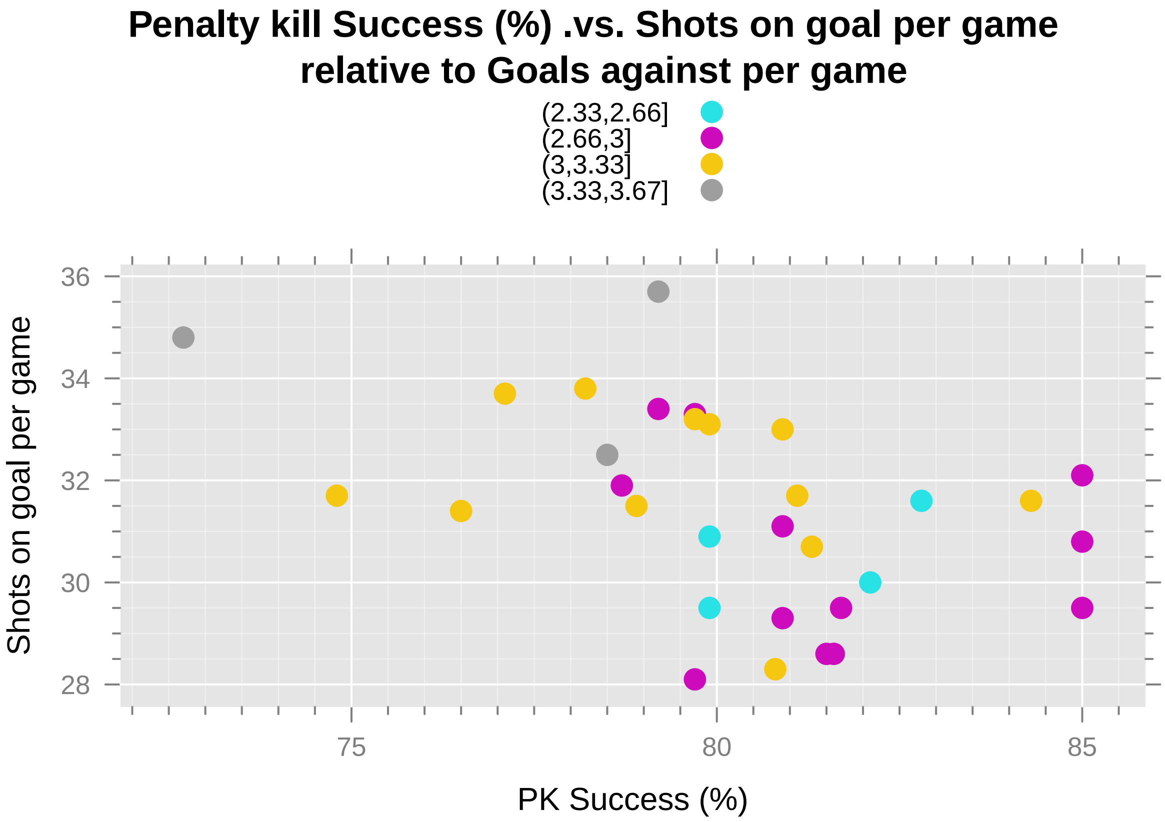
It seems very reasonable that it should, but we will use statistical analysis to see if that is the case:

```
library(latticeExtra)
library(spida2)
```

```
gd(col = 1:4,pch = 16)
xyplot(shots_on~pp, auto.key  = TRUE, groups = cut(GF_g, 4),
       xlab = "PP Success (%)", ylab = "Shots on goal per game",
       main = "Powerplay Success (%) .vs. Shots on goal per game \n relative to Goals for per game")
```



```
gd(col = 5:8, pch = 16)
xyplot(shots_against~pk, auto.key  = TRUE, groups = cut(GA_g, 4),
       xlab = "PK Success (%)", ylab = "Shots on goal per game",
       main = "Penalty kill Success (%) .vs. Shots on goal per game \n relative to Goals against per ga
```



From the first plot above, we see that teams in the upper echelon of goals for per game tend to get more shots on goal per game and have a relatively high percentage of success, but for teams that have a lower number of goals for per game, despite more or less shots per game, their success on the powerplay is relatively equivalent, which develops the idea that more shots on the powerplay could be beneficial in its success, but this relationship is not prominent. Although, in the second plot, we see a much stronger relationship between shots against per game and success on the penalty kill (this includes with goals against per game at different levels, and the less goals against per game, the higher the penalty kill success percentage). This gives us the idea that less shots against per game could have a significant effect on improving a team's penalty killing ability, which allows them to keep goals out of their net. They seem to be interconnected, but we cannot acquire the full picture of these relationships because we do not have statistics on powerplay goals and powerplay shots individually from their overall totals, or number of penalty kills or powerplays that each team endured throughout the season, so we cannot truly see if more shots on the powerplay, or less against on penalty kills truly does influence PP% and PK% respectively, but this gives us a general idea and motivates further research for sure. Also, there is no stats on what teams had powerplays against any other team and so the skill that each team has in these situations is not accounted for here, but once again, the general case takes these assumptions into account.

```
pp_shton_model <- glm(cbind(wins,loss)~pp + shots_on, data = ggg, family = binomial)
summary(pp_shton_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pp + shots_on, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.2994  -0.9687   0.1198   0.7208   3.8757
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.97207    0.70906  -4.192 2.77e-05 ***
## pp           0.01807    0.01078   1.676 0.093651 .
## shots_on     0.08317    0.02296   3.623 0.000291 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 63.966  on 28  degrees of freedom
## AIC: 219.71
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(pp_shton_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##          LR Chisq Df Pr(>Chisq)
## pp         2.8135  1  0.0934738 .
## shots_on  13.1922  1  0.0002811 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
pk_shtag_model <- glm(cbind(wins,loss)~pk + shots_against, data = ggg, family = binomial)
summary(pk_shtag_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pk + shots_against, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7350  -1.0957  -0.1799   1.0052   4.3043
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.29455    1.76320  -0.167  0.86733
## pk             0.03080    0.01641   1.877  0.06052 .
## shots_against -0.06923    0.02320  -2.984  0.00284 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 61.854  on 28  degrees of freedom
## AIC: 217.6
##
## Number of Fisher Scoring iterations: 3
```

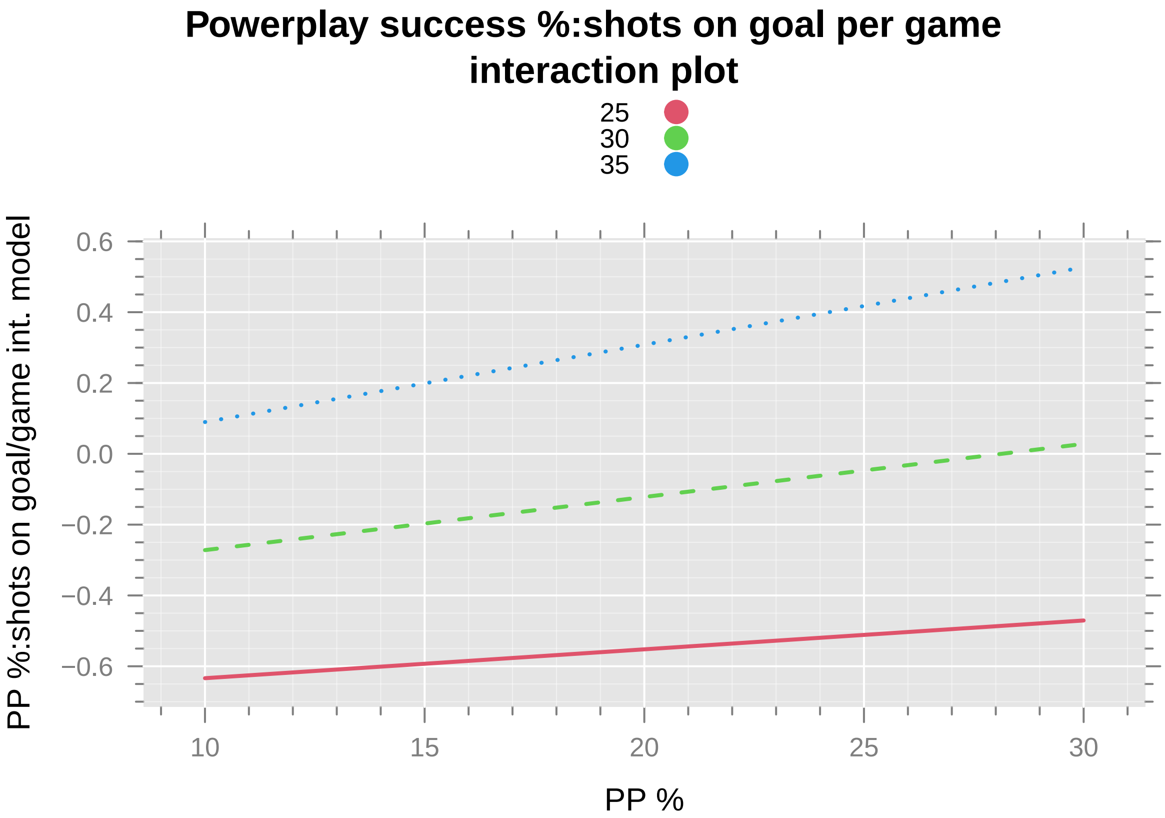
```
Anova(pk_shtag_model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##               LR Chisq Df Pr(>Chisq)
## pk              3.5355  1   0.060069 .
## shots_against   8.9472  1   0.002779 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Finally, we try a model with the previous predictors but add shots on goal per game for the PP% model and shots against per game for the PK% model as those are known to directly affect the PP% and PK% respectively, so by adding them into these models, we find an even better model (based on the AIC), where each of the predictors are significant and this tells us that wins can be affected by PP% levels and PK% levels, but only when we control for shots in these cases, as PP% and PK% are built on those values, or shots in both factions confound their effects on wins. Given these results, it begs the question whether there could be significant interaction between these respective sets of variables given the confounding nature of shots on goal per game on powerplay success percentage and shots against per game on penalty kill percentage.

```
pp_shots_int_model <- glm(cbind(wins,loss)~pp*shots_on, data = ggg, family = binomial)
pp_shots_int_model_c <- glm(cbind(wins,loss)~pp_c*shots_on_c, data = ggg, family = binomial)
```

```
par(mfrow = c(1,1), oma = c(0,0,0,0))
predshots <- expand.grid(pp = c(10,20,30), shots_on = c(25, 30, 35))
predshots$pp_shots_int_model <- predict(pp_shots_int_model,newdata = predshots,type = 'link')
library(latticeExtra)
gd(col = 2:4)
xyplot(pp_shots_int_model~pp, groups = shots_on,predshots,type = "l", auto.key = TRUE,
       lwd = 2, xlab = "PP %", ylab = "PP %:shots on goal/game int. model",
       main = "Powerplay success %:shots on goal per game \n interaction plot")
```



```
summary(pp_shots_int_model_c)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pp_c * shots_on_c, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3088  -0.9889   0.1046   0.7143   3.8907
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.002112   0.041535  -0.051  0.95944
## pp_c             0.016987   0.012565   1.352  0.17640
## shots_on_c       0.085631   0.027252   3.142  0.00168 **
## pp_c:shots_on_c  0.001370   0.008181   0.167  0.86704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 63.938  on 27  degrees of freedom
## AIC: 221.68
```

```
##
## Number of Fisher Scoring iterations: 3
```

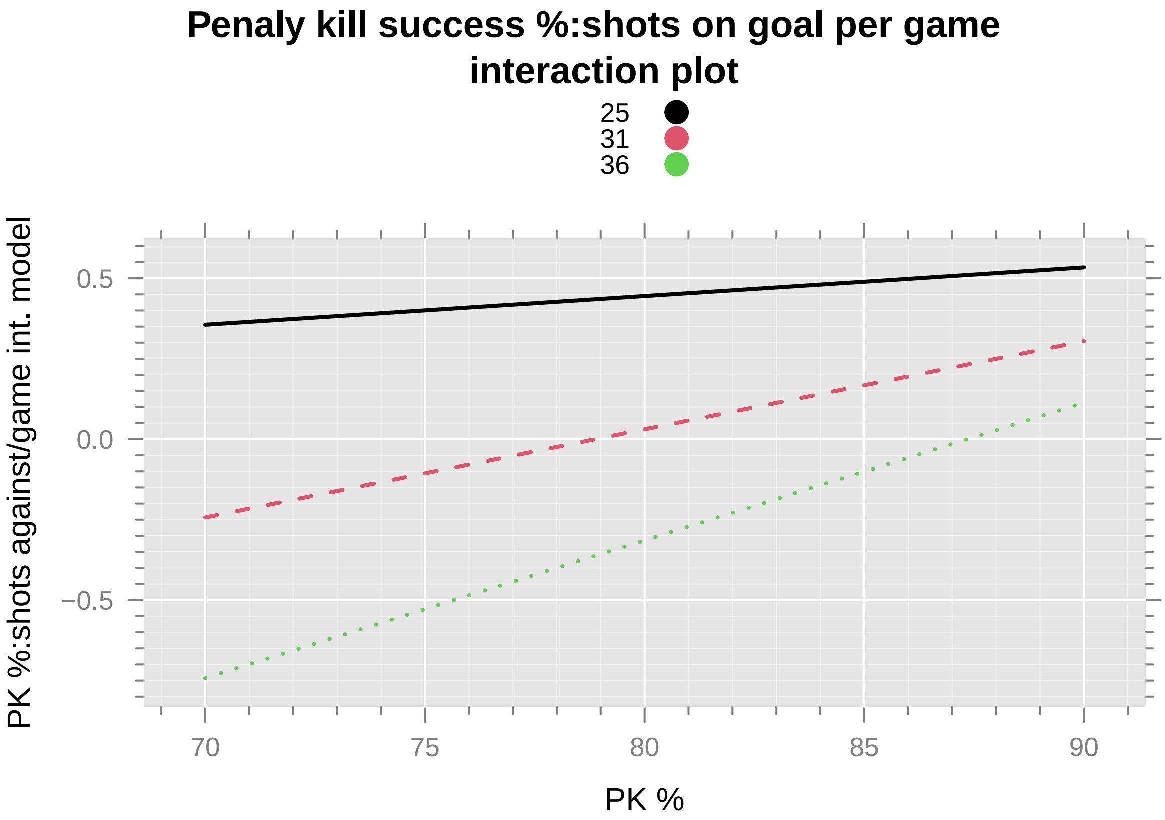
```
Anova(pp_shots_int_model_c)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##                  LR Chisq Df Pr(>Chisq)
## pp_c               2.8135  1  0.0934738 .
## shots_on_c        13.1922  1  0.0002811 ***
## pp_c:shots_on_c    0.0280  1  0.8670447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This interaction plot between powerplay success percentage and shots on goal per game showed slight deviations from parallelism between the lines, but when we dealt with the interaction numerically, the Anova test showed us that there is no significant interaction. Given the additive model results and the results from this interaction model, it's fairly clear to see that the shots on goal per game a team gets has a confounding effect on the success of a powerplay, when keeping shots on goal at it's zero, or centered level in this case, but that doesn't carry over to it's effect on powerplays when keeping it constant at certain levels–it is pretty much the same effectiveness at each different level of shots on goal per game. To note, the scaled variables were used for the numerical analysis as extrapolating to zero for those variables would make no sense for analysis of their effects.

```
pk_shots_int_model_c <- glm(cbind(wins,loss)~pk_c*shots_against_c, data = ggg, family = binomial)
pk_shots_int_model <- glm(cbind(wins,loss)~pk*shots_against, data = ggg, family = binomial)
```

```
par(mfrow = c(1,1), oma = c(0,0,0,0))
predshotsag <- expand.grid(pk = c(70,80,90), shots_against = c(25, 31, 36))
predshotsag$pk_shots_int_model <- predict(pk_shots_int_model,newdata = predshotsag,type = 'link')
library(latticeExtra)
gd(col = 1:3)
xyplot(pk_shots_int_model~pk, groups = shots_against,predshotsag,type = "l", auto.key = TRUE,
       lwd = 2, xlab = "PK %", ylab = "PK %:shots against/game int. model",
       main = "Penaly kill success %:shots on goal per game \n interaction plot")
```



```
summary(pk_shots_int_model_c)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pk_c * shots_against_c, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.02249  -0.82328  -0.01161   0.66151   2.77120
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.026112   0.040520  -0.644    0.519
## pk_c                  0.044315   0.011094   3.994 6.49e-05 ***
## shots_against_c      -0.114234   0.021596  -5.290 1.23e-07 ***
## pk_c:shots_against_c  0.032229   0.008022   4.018 5.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 40.037  on 27  degrees of freedom
## AIC: 197.78
```

```
##
## Number of Fisher Scoring iterations: 3
```

```
Anova(pk_shots_int_model_c)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(wins, loss)
##                      LR Chisq Df Pr(>Chisq)
## pk_c                   9.1231  1   0.002524 **
## shots_against_c       20.8914  1  4.861e-06 ***
## pk_c:shots_against_c  16.2289  1  5.613e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We performed similar tests for penalty kill success percentage and shots against per game as we did with powerplay success percentage and shots on goal per game, but we got opposite results in this case. The interaction between the PK% variable and shots against per game variable was much stronger and significant numerically, as well as can be seen on its interaction plot, as the lines are clearly not parallel. Therefore, the confounding effect of shots against per game on PK% tells us that by keeping shots against per game to a certain level, a team can achieve success on the penalty kill and translate that to wins. In other words, there is a significant effect of penalty kill percentage on wins when changing the levels of shots against per game.

Overall, powerplay and penalty kill success definitely hinges on of shots, and this provides us with another stepping stone to understanding what can help a team to win, even if our data is minimal in this case, it gives us insight of where to look, which is truly a spectacular result.

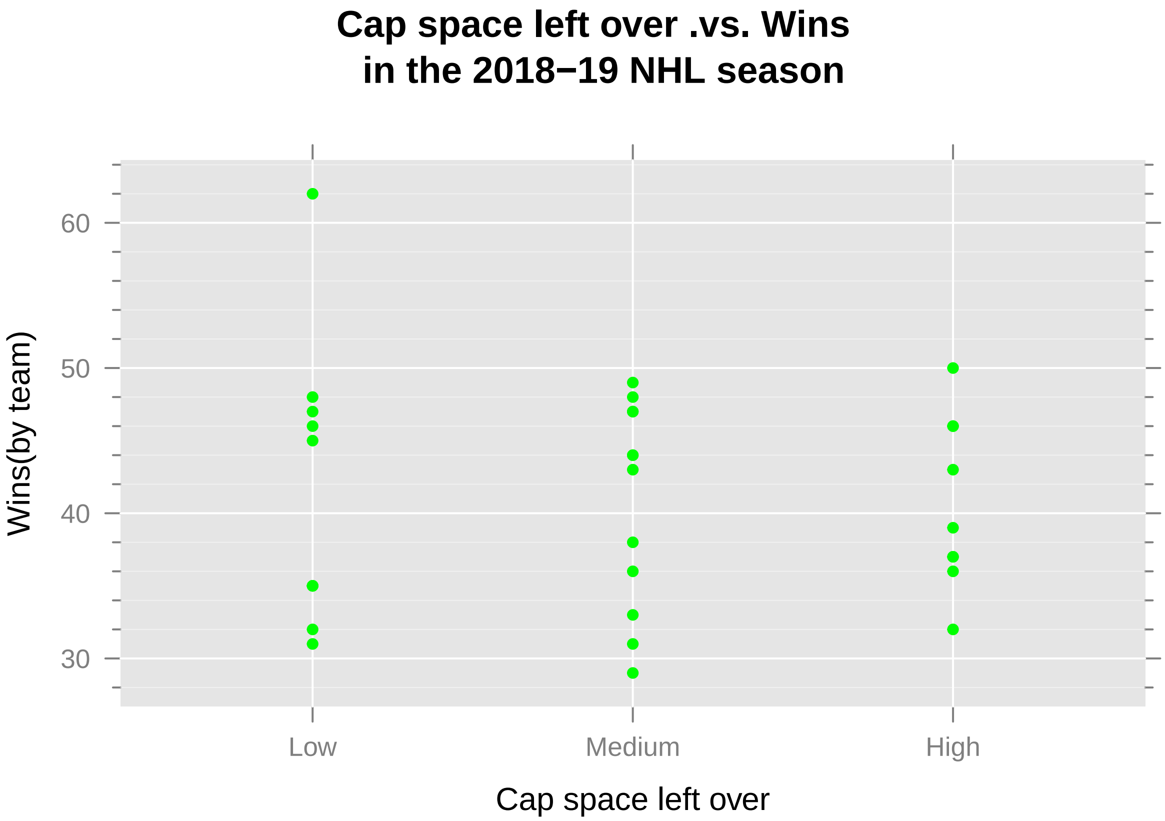
Thus far, we have dealt with each numerical variable in our data set that is determined within the hockey games themselves and how they may affect wins, but we will now test to see how a factor outside of the game itself may develop an important connection to winning.

```
cap_space <- factor(ggg$Cap.space.left.over, levels = c("Low", "Medium", "High"))
```

We will be looking at how the salary cap, or how teams spent their money on players in this season, affected their ability to win games. The idea behind testing the effect of a variable like this is that usually the more skilled players are paid a higher salary, and so if a team were to spend more of their cap in a season, we are testing this variable under the assumption that that team has more skilled players, and with more skilled players, wins tend to be easier to come by, as we have seen with the Edmonton Oilers organization throughout the 1980s.

Therefore, the code above (about "cap_space") deals with introducing the variable of cap space into the data set in which it's levels are determined by how much salary cap space a team has left after they had paid all of their players in that year. In this case, "low" represents a team having less that $2 million in cap space left over, "Medium" deals with having $2-5 million in cap space left over, and "High" represents having more than $5 million in cap space left over.
This variable may help us figure out if spending more or less in cap space for a team's players is beneficial in terms of winning, within these given categories.

```
library(latticeExtra)
xyplot(wins~cap_space, col = "green", xlab = "Cap space left over", ylab = "Wins(by team)",
       main = "Cap space left over .vs. Wins \n in the 2018-19 NHL season")
```



After looking at the plot above, we see that, despite the one outlier of a team winning over 60 games with a low amount of salary cap space left over, all of the levels show fairly similar results, which means that we don’t have a clear indication that a certain amount of cap space left over is beneficial in terms of winning.

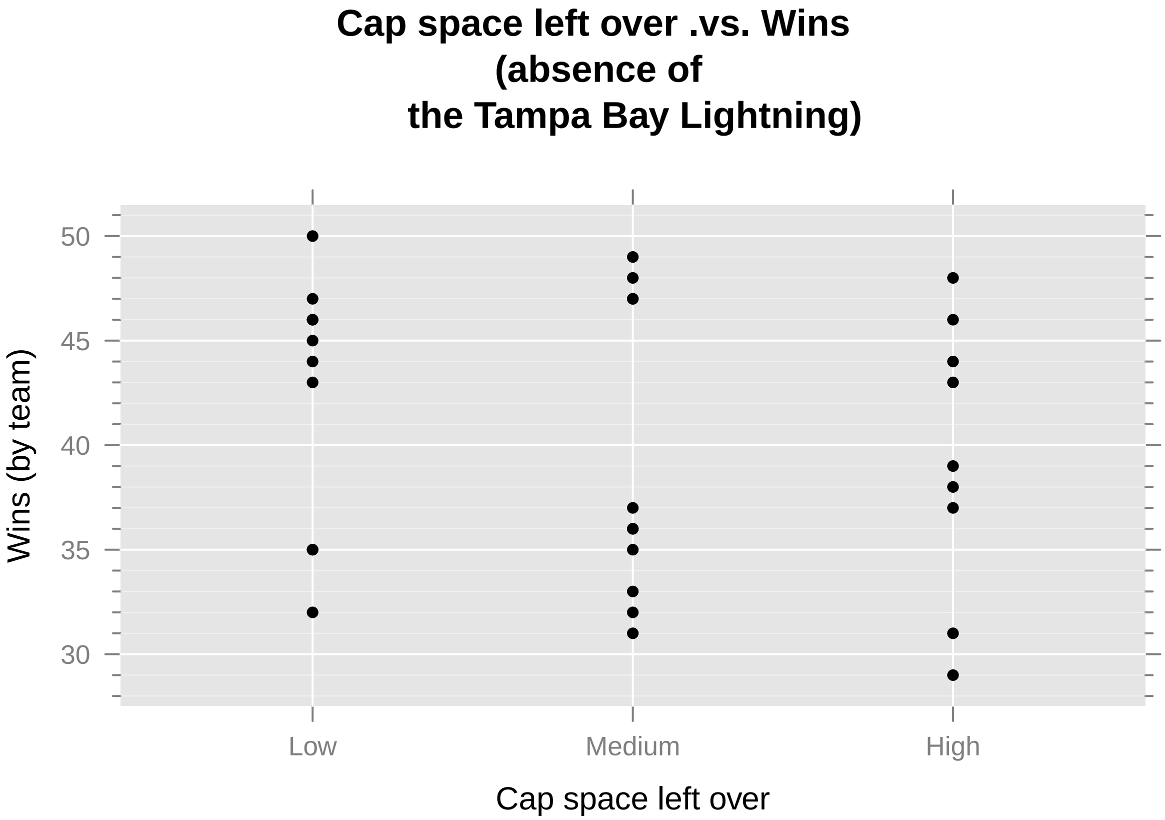
```
cap_model <- glm(cbind(wins,loss)~ cap_space, data = ggg, family = binomial)
summary(cap_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ cap_space, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.658  -1.292   0.000   1.294   4.105
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.14660    0.07003   2.093  0.03631 *
## cap_spaceMedium -0.17913    0.09471  -1.891  0.05859 .
## cap_spaceHigh   -0.26599    0.10170  -2.615  0.00891 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 76.248  on 28  degrees of freedom
## AIC: 231.99
##
## Number of Fisher Scoring iterations: 3
```

Although, after developing the model to test it against wins, we see that this variable may have an interesting affect after all, but before we can interpret these results, it seems that the outlier we spotted in the plot earlier may be affecting our results, so we will develop an analysis of this variable of interest after removing that outlier from our dataset.

```
no_light <- ggg[-c(26),]
cap_space_no_light <- factor(no_light$Cap.space.left.over, levels = c("Low", "Medium", "High"))
xyplot(wins~cap_space, data = no_light, xlab = "Cap space left over",
       ylab = "Wins (by team)", main = "Cap space left over .vs. Wins \n (absence of
       the Tampa Bay Lightning)")
```



When we look at the plot that excludes the outlier (Tampa Bay Lightning who had over 60 wins and a low amount of salary cap space left over), we can see a clearer picture of how the salary cap space left over does not produce as much of an effect on wins as I had initially thought. To see if this is truly the case, we will perform a more technical analysis:

```
no_light_model <- glm(cbind(wins,loss)~ cap_space_no_light, data = no_light, family = binomial)
summary(no_light_model)
```

```
##
```

```
## Call:
## glm(formula = cbind(wins, loss) ~ cap_space_no_light, family = binomial,
##     data = no_light)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2160  -1.4008   0.1596   1.3104   2.0893
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               0.04879    0.07364   0.663    0.508
## cap_space_no_lightMedium -0.08131    0.09741  -0.835    0.404
## cap_space_no_lightHigh   -0.16817    0.10422  -1.614    0.107
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 60.242  on 29  degrees of freedom
## Residual deviance: 57.635  on 27  degrees of freedom
## AIC: 208.82
##
## Number of Fisher Scoring iterations: 3
```

Interestingly enough, the outlier from our data affected our results quite strongly and now to interpret these results we have evidence to say that the difference between the medium "cap space left over" level and its respective low level is insignificant; the same goes for the difference between the high level and the low level. What this means is that the amount that a team would spend on players to improve it does not improve their probability of winning games all that much. Once again, these results stem from a minimal number of available factors and observations, so the types of players they spend money on, or even the experience and past success of the players that they are spending money on, may quite easily change these results and make this variable a significant in terms of its effect on wins, but with what we have available, that is not the case.

What this analysis above has shown us though, is that the Tampa Bay lightning was a team that had a beyond outstanding regular season in comparison to the other teams in the NHL in the 2018-19 season. As a result, this leads us to think about how our analysis for the other variables we tested earlier may have been affected by their outstanding play in those categories that could have led to results that did not illustrate the average tendencies of the league on a more regular basis. We will test one of those previous models with the data set that the Tampa Bay Lightning has been removed from to see if their effect on the data would have affected our results fro those models, much like it did with the cap space models.

```
faceoff_model <- glm(cbind(wins,loss) ~ faceoff,data = ggg,family = binomial)
summary(faceoff_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ faceoff, family = binomial,
##     data = ggg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5990  -1.3135   0.0891   1.1227   4.5324
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.03687    1.09114  -1.867   0.0619 .
## faceoff      0.04075    0.02181   1.868   0.0618 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 80.029  on 29  degrees of freedom
## AIC: 233.78
##
## Number of Fisher Scoring iterations: 3
```

```
no_light_faceoff_model <- glm(cbind(wins,loss) ~ faceoff,data = no_light,family = binomial)
summary(no_light_faceoff_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ faceoff, family = binomial,
##     data = no_light)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4717  -1.2885   0.1607   1.2642   2.2545
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.46618    1.09727  -1.336    0.181
## faceoff      0.02867    0.02195   1.306    0.192
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 60.242  on 29  degrees of freedom
## Residual deviance: 58.534  on 28  degrees of freedom
## AIC: 207.72
##
## Number of Fisher Scoring iterations: 3
```

Initially, the model that was testing the effect faceoff win percentage on wins was significant when using the full data set, but weakly at best, although when we test this same model but using the data set that has excluded the Tampa Bay Lightning,the effect of faceoff win percentage on wins becomes insignificant on wins, which provides evidence to show us that this outlier had a strong effect on the data and gave us inflated results. We see similar results with the powerplay success percentage model as well, but along with becoming not significant, the estimated effect of powerplay success percentage became smaller (much weaker):

```
pp_model <- glm(cbind(wins,loss)~pp, data = ggg, family = binomial)
summary(pp_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pp, family = binomial, data = ggg)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7530  -1.3361   0.0762   1.1833   3.7617
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.52189    0.21088  -2.475   0.0133 *
## pp           0.02649    0.01051   2.520   0.0117 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.527  on 30  degrees of freedom
## Residual deviance: 77.158  on 29  degrees of freedom
## AIC: 230.91
##
## Number of Fisher Scoring iterations: 3
```

```
no_light_pp_model <- glm(cbind(wins,loss)~pp, data = no_light, family = binomial)
summary(no_light_pp_model)
```

```
##
## Call:
## glm(formula = cbind(wins, loss) ~ pp, family = binomial, data = no_light)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5492  -1.1734   0.1805   1.2709   2.1540
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.178561   0.226802  -0.787    0.431
## pp           0.007436   0.011492   0.647    0.518
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 60.242  on 29  degrees of freedom
## Residual deviance: 59.823  on 28  degrees of freedom
## AIC: 209.01
##
## Number of Fisher Scoring iterations: 3
```

Despite this being a very basic approach to illustrating the effect that an outlier can have on a dataset, it does display how the effects of variables in this dataset on wins could be swayed one way or the other by the influence and leverage that a superior team’s ability in certain categories is. Therefore, our analysis performed above with the full data set on the variables of interest and determining their effects on wins holds valid, as it takes into account all teams and still displays how strongly performing at optimal levels in many aspects of hockey can optimize wins, but by excluding that outlying team’s performance we could show even stronger relationships that could affect a team’s ability to win that does not have the skill or players that the superior team does; how a mid to low tier team can improve their play and find success, essentially, which could provide very interesting connections if we had the data to do so. Very exciting nonetheless.

Conclusion

The statistical nature of the NHL is quite vast and complex, with a multitude of causal questions that breed excitement and intrigue. In the case of this report, I only ventured into the shallow waters of statistical analysis and what greases the gears of success in the NHL, but was able to find some interesting results nonetheless. With the overall question of what it takes to make the playoffs, I was able to reduce this complex question into the nature of winning. Through this I found that shots, goals, and powerplays were effective on a team's ability to win games, but faceoffs were less so. With interaction testing and looking for confounders and mediators, goal differential seemed to be the most important factor in finding evidence of causing wins, with the other variables playing more minor roles, but important nonetheless. Also, many causal questions were faced and I was able to decipher some interesting answers, be them significant or not. Finally, with 2018-19 being an interesting year in the NHL, as the Tampa Bay Lightning won an extraordinary amount of games (62 out of 82), I was able to see how outliers in our data could affect results and what that could mean moving forward. Overall, statistics are an important foundational tool to help analyze uncertainty, or almost anything that we face in our lives, and being able to show that within the context of my favourite sport and something that has changed my life was a fruitful and awe-inspiring adventure.

References

1. https://www.tsn.ca/nhl/statistics
2. https://www.capfriendly.com/archive/2020