Inquiry Line 4

James Klimavicz, EB Wlezien, Coskun Erden, Mauricio Serrano-Porras, and Valeria Velasquez-Zapata

11/21/2018

Line of Inquiry

Do the number of corner kicks and conceded free kicks associated with a higher number of goals scored per game?

Here, we consider the number of home team and away team committed fouls, the home and away corner kicks, and the number of home and away goals scored.

Data Loading and Tidying

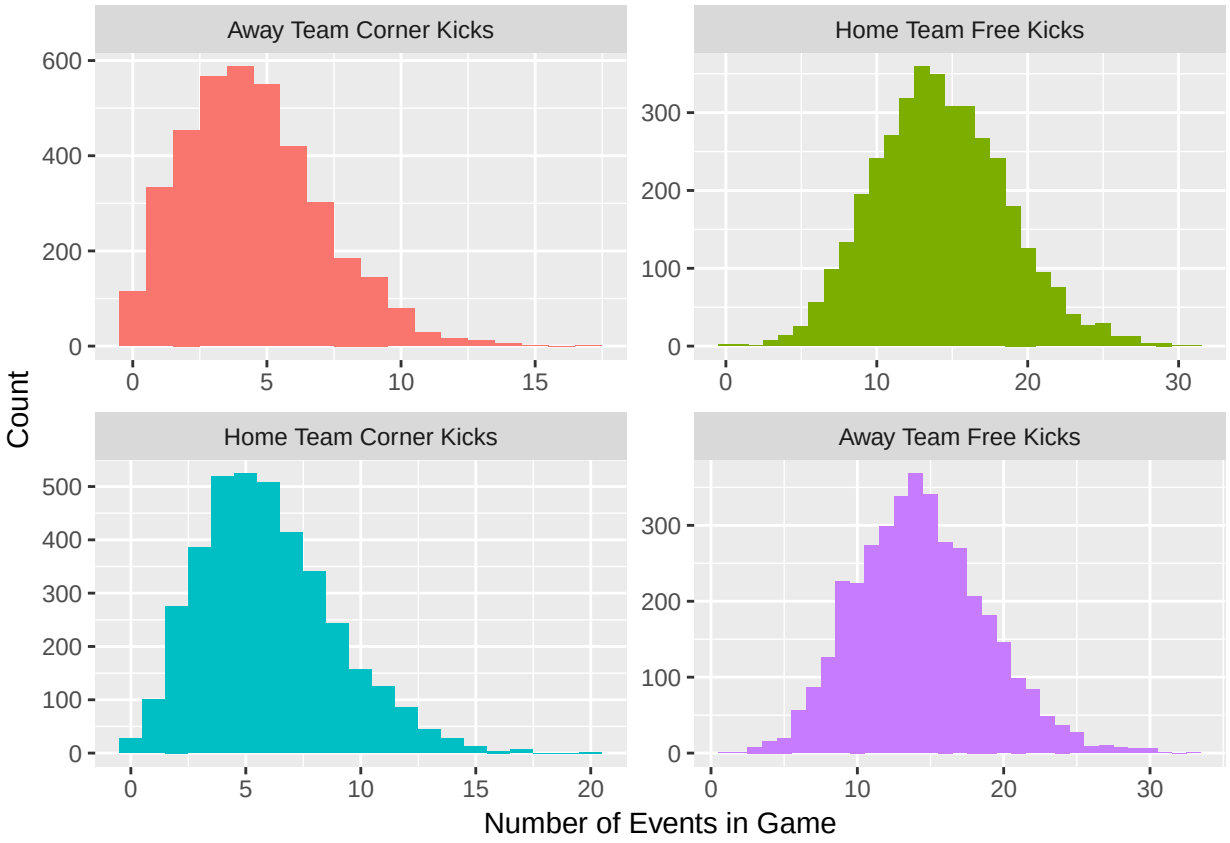
First, load the data, and perform some tidying.

```
#read data
football = read.table("allseasons.csv", header = T, sep=",")
#determine total goals (TG)
football$TG=with(football,FTHG+FTAG)
#gather based on the number of corner in free kicks, our variables of interest.
football_long_in <- football %>% gather(key = "param", value = "count", c("HF","AF","HC","AC"))
football_long <- football_long_in[,c("FTHG","FTAG","TG","param","count")]
football_long$param.name <- factor(football_long$param)
levels(football_long$param.name) <- c("Away Team Corner Kicks",
                                      "Home Team Free Kicks",
                                      "Home Team Corner Kicks",
                                      "Away Team Free Kicks")
#gather based on types of scores
football_long_scores <- football %>% gather(key = "goal_type",
                                            value = "score", c("FTHG", "FTAG", "TG"))
football_long_scores$goal.name <- factor(football_long_scores$goal_type)
levels(football_long_scores$goal.name) <- c("Away Goals", "Home Goals", "Total Goals")
#gather based on all parameters of interest:
param_list <- c("HF","AF","HC","AC","FTHG", "FTAG", "TG")
football_all_long <- football %>% gather(key="param", value = "count", param_list)
football_all_long <- football_all_long[,c("param","count")]
football_all_long$param.name <- factor(football_all_long$param)
#note that AF (away fouls) becomes Home Team Free Kicks, and HF becomes Away Team Free Kicks
levels(football_all_long$param.name) <- c("Away Team Corner Kicks",
                                          "Home Team Free Kicks",
                                          "Away Team Goals",
                                          "Home Team Goals",
                                          "Home Team Corner Kicks",
                                          "Away Team Free Kicks",
                                          "Total Goals")
```

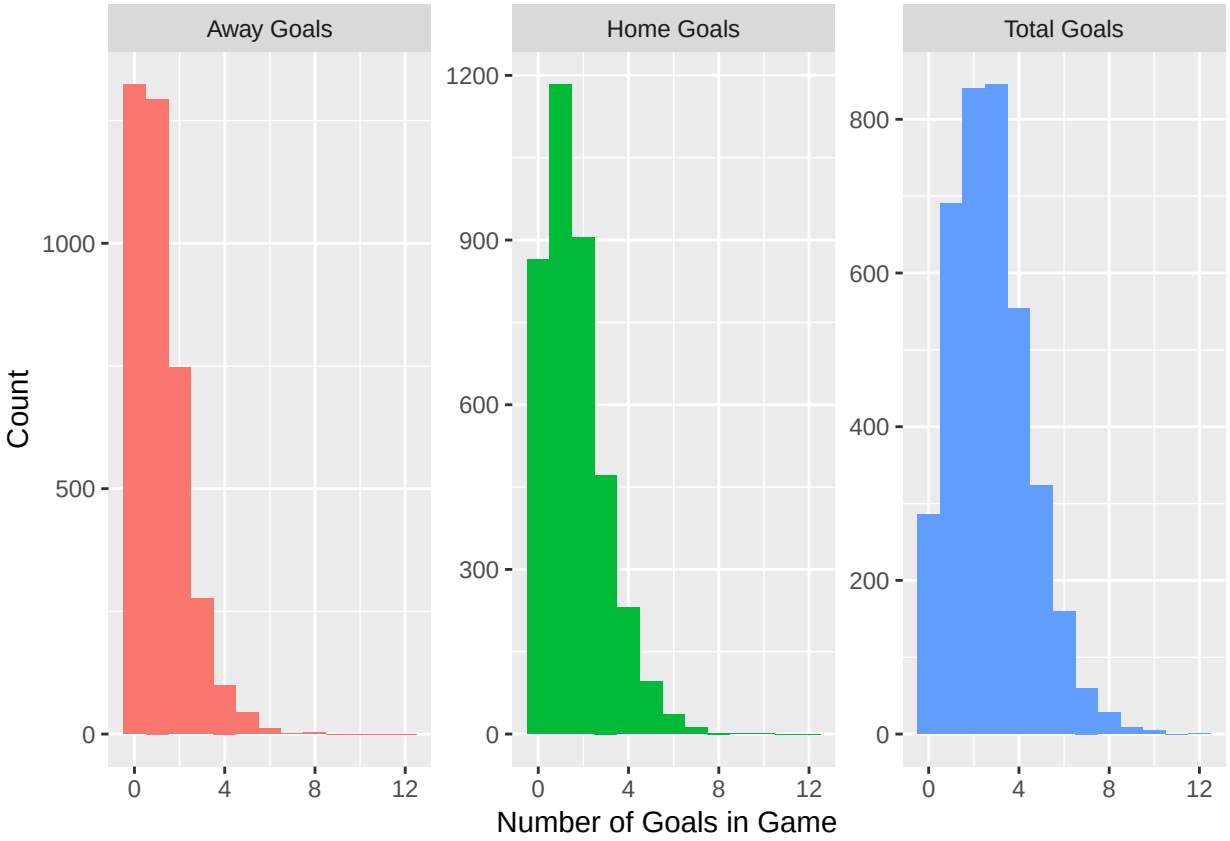
Preliminary Dataset Analysis

Now, analyze data, particularly data that will be used for the analysis of this project. First, we'll look at the parameters of the number of free kicks and corner kicks, as well as the number of home, away, and total goals.

```
football_long %>% ggplot(aes(x = count, fill=param)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~param.name, scales="free") +
  theme(legend.position="none") +
  xlab("Number of Events in Game") +
  ylab("Count")
```

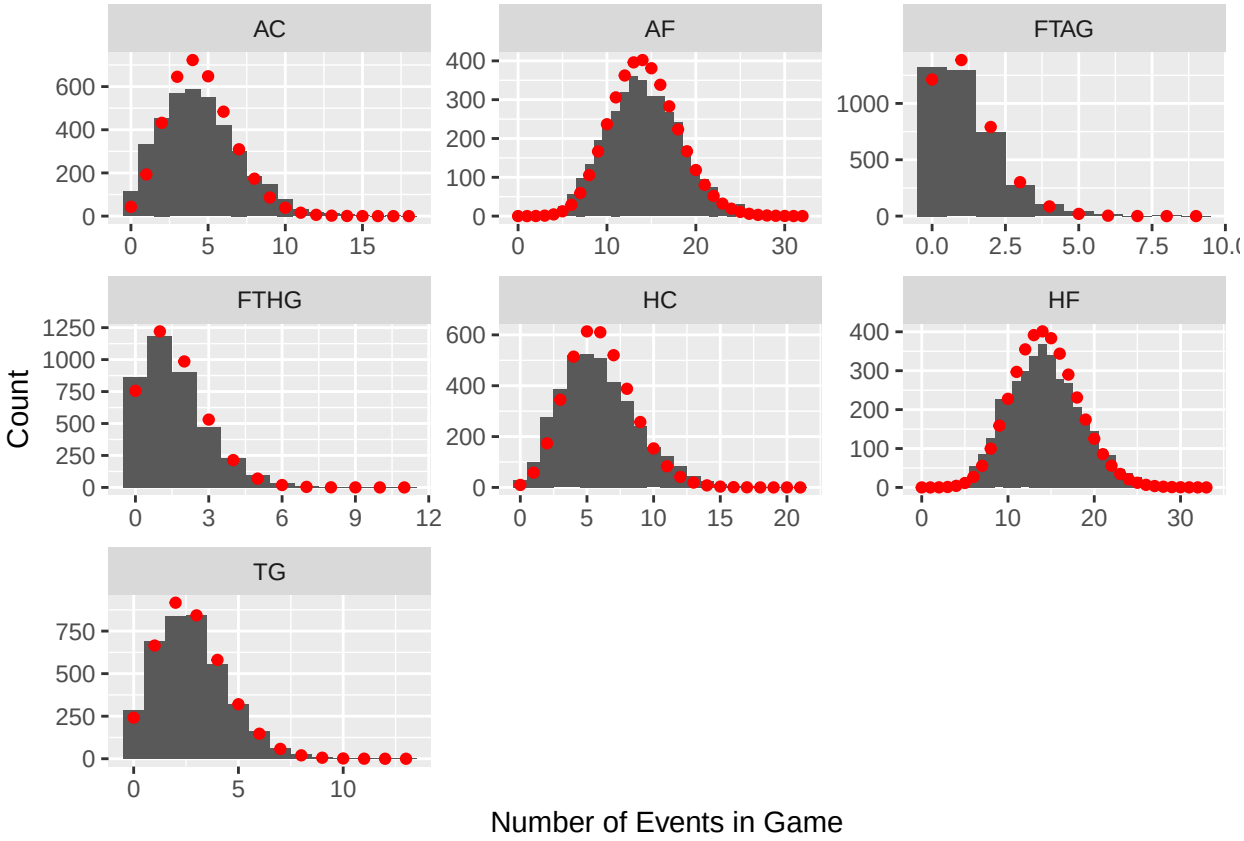



```
football_long_scores %>% ggplot(aes(x = score, fill=goal_type)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~goal.name, scales="free_y") +
  theme(legend.position="none") +
  xlab("Number of Goals in Game") +
  ylab("Count")
```



The data is slightly skewed, and superficially the number of free and corner kicks by the home and away team look like Poisson distributions. In fact, they look close enough to Poisson distributions to look at this further.

```
#Find lambda for a fitted Poisson curve, as well as other statistics, for each game stat.
params <- football_all_long %>% group_by(param) %>%
  dplyr::summarise(lambda = MASS::fitdistr(count, "Poisson")$estimate[1],
                   max = max(count),
                   count = n())
#Find the predicted values of a Poisson curve with the estmated lambda for each game stat.
#inspired by https://stackoverflow.com/questions/1376967/
grid <- with(football_all_long, seq.int(min(count), max(count), 1))
pois_dens <- ddply(football_all_long, "param", function(df) {
  data.frame(
    predicted = grid,
    density = dpois(grid, MASS::fitdistr(df[,2], "Poisson")$estimate[1])*length(df[,2])
  )
})
#modify the predicted data to fit with free-scale facet wrap when plotting
pois_dens_mod <- left_join(pois_dens,params, by = "param") %>% filter(predicted <= max + 1)
#plot the game stat with the Poisson predicted curves.
football_all_long %>% ggplot(aes(x = count)) + geom_histogram(binwidth = 1) +
  geom_point(aes(x=predicted, y=density), data=pois_dens_mod, color="red") +
  facet_wrap(~param, scale="free") +
  xlab("Number of Events in Game") +
  ylab("Count")
```



It's not a great fit, but it's not terrible.

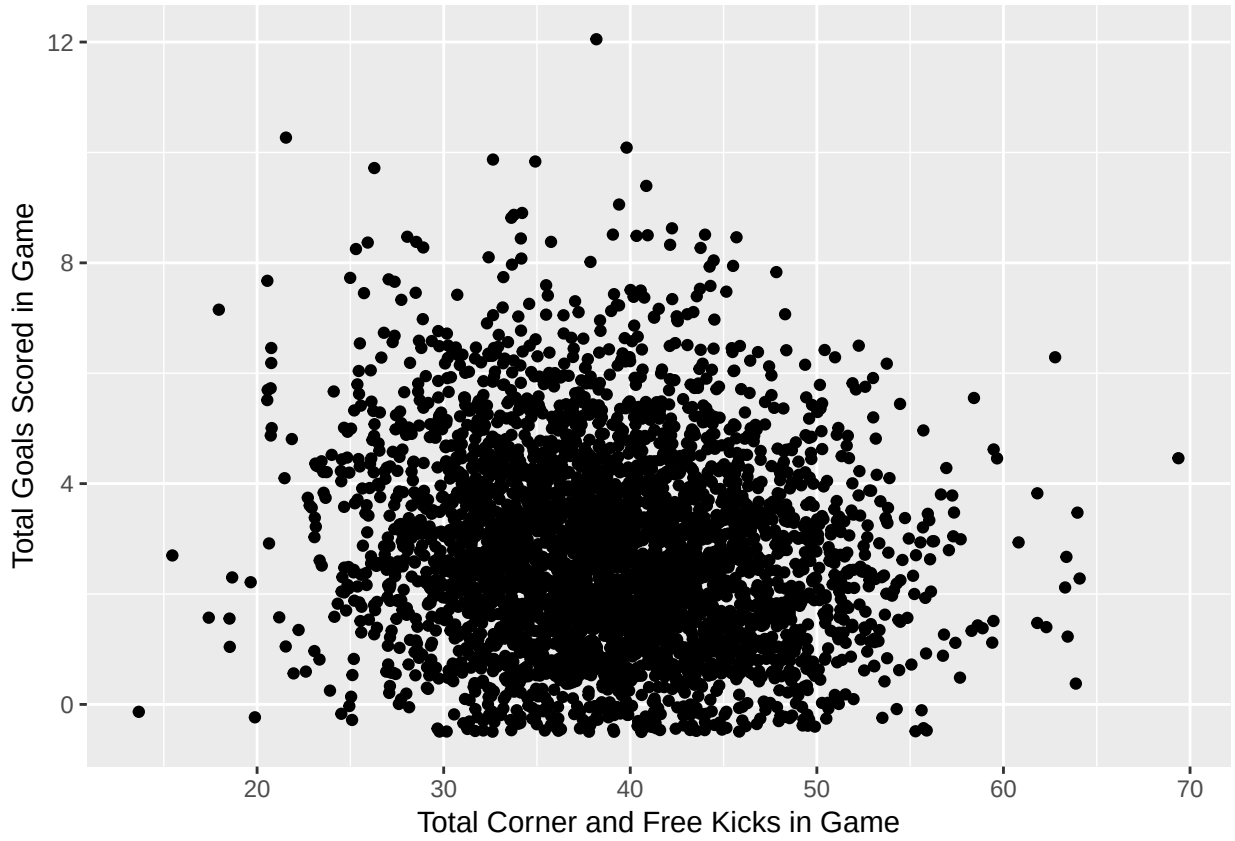
Association of Kicks and Scoring

We can now look at any potential association between free and corner kicks with scoring goals.

Preliminary Dataset Analysis

First, let's look at a simple plot of the total goals scored in the game against the total number of free and corner kicks.

```
#use geom_jitter with 0.5 for height and width to show density of data better
football %>% ggplot(aes(y=TG, x=HF+AF+HC+AC)) + geom_jitter(width=0.5, height=0.5) +
  ylab("Total Goals Scored in Game") +
  xlab("Total Corner and Free Kicks in Game")
```



This plot shows that there is very little association immediately present. If we squint, there is perhaps some very minor downward trend that is nearly obfuscated by the noise. Perhaps we'd have better luck looking at simple correlation, with the hope there are some linear trends.

```
cor(football[,c("FTHG","FTAG","TG")],football[,c("HF","AF","HC","AC")])
```

```
##                HF          AF           HC          AC
## FTHG -0.145074334 -0.03146761  0.006572098 -0.04230196
## FTAG -0.004331261 -0.07009970 -0.022909255  0.02392529
## TG   -0.118909946 -0.07225975 -0.010137294 -0.01774926
```

Unfortunately, our correlation matrix shows essentially no evidence of a simple linear trend between home, away, or total goals scored in a game with the number of home or away free or corner kicks.

Linear Model Results

We may expect that the number of goals scored in a game is a function of the number of free kicks and corner kicks, i.e.

$$G = f(AF, HC, HF, AC) + \varepsilon,$$

where G is the number of a specific type of goals in a game (i.e. total goals, away team goals, or home team goals),AF is the number of away team fouls (penalty kicks awarded to the home team), HC is the number of home team corner kicks, HF is the number of home team fouls (penalty kicks awarded to the away team), AC is the number of away team corner kicks, and ε is an error term.

As a last attempt, perhaps we can fit linear models to predict the number of home, away, and total goals scored, based on the number of corner and penalty kicks. Because a foul committed by one team results in a

penalty kick for the other team, we might expect a synergistic effect between corner kicks on one team with penalty kicks on the other team. We therefore could develop the linear model

$$G_i = \mu + AF_i + HC_i + (AF \cdot HC)_i + (HF)_i + (AC)_i + (HF \cdot AC)_i + \varepsilon_i,$$

where μ is a constant, $(AF \cdot HC)$ and $(HF \cdot AC)$ are interaction terms, and $\varepsilon \sim \mathcal{N}(0, \sigma)$ is assumed to be i.i.d. random error. As a reminder, we would likely expect that the number of goals for a team would increase when that team is awarded penalty kicks and corner kicks, as these allow for formation of strategy.

Total Goals

We'll first look at the number of total goals in the game.

```
#Total goals
fit_all = lm(TG ~ HC*AF + HF*AC, data=football)
summary(fit_all)
```

```
##
## Call:
## lm(formula = TG ~ HC * AF + HF * AC, data = football)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4967 -1.2985 -0.0479  1.1009  9.1426
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0806616  0.2931321  13.921  < 2e-16 ***
## HC          -0.0177805  0.0324271  -0.548   0.5835
## AF          -0.0260366  0.0145321  -1.792   0.0733 .
## HF          -0.0521792  0.0128592  -4.058 5.06e-05 ***
## AC          -0.0373356  0.0371722  -1.004   0.3153
## HC:AF       -0.0001442  0.0021692  -0.066   0.9470
## HF:AC        0.0012461  0.0024796   0.503   0.6153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.727 on 3793 degrees of freedom
## Multiple R-squared:  0.01995,    Adjusted R-squared:  0.0184
## F-statistic: 12.87 on 6 and 3793 DF,  p-value: 1.896e-14
```

```
#We see that interation is not significant, so try with no interaction.
# fit_all = lm(TG ~ HC + AF + HF + AC, data=football)
# summary(fit_all) #AC and HC are not significant
fit_all = lm(TG ~ AF + HF, data=football)
summary(fit_all)
```

```
##
## Call:
## lm(formula = TG ~ AF + HF, data = football)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3148 -1.3022 -0.0308  1.0945  9.0802
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.770934   0.127547  29.565  < 2e-16 ***
## AF          -0.026029   0.006428  -4.049 5.24e-05 ***
## HF          -0.044897   0.006293  -7.134 1.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.727 on 3797 degrees of freedom
## Multiple R-squared:  0.01838,    Adjusted R-squared:  0.01786
## F-statistic: 35.55 on 2 and 3797 DF,  p-value: 5.075e-16
```

With this model with only the statistically significant componenents, we see that both home and away team fouls significantly reduce the number of total goals scored in these games, though the effect is rather small.

Away Team Goals

```
#Away Team goals
fit_away = lm(FTAG ~ HC*AF + HF*AC, data=football)
summary(fit_away)
```

```
##
## Call:
## lm(formula = FTAG ~ HC * AF + HF * AC, data = football)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7218 -1.0840 -0.1383  0.8246  6.7694
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.806221   0.197974   9.124  < 2e-16 ***
## HC          -0.088487   0.021900  -4.040 5.44e-05 ***
## AF          -0.052260   0.009815  -5.325 1.07e-07 ***
## HF           0.007218   0.008685   0.831  0.40598
## AC           0.030652   0.025105   1.221  0.22218
## HC:AF        0.005674   0.001465   3.873  0.00011 ***
## HF:AC       -0.001804   0.001675  -1.077  0.28149
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.166 on 3793 degrees of freedom
## Multiple R-squared:  0.009736,   Adjusted R-squared:  0.008169
## F-statistic: 6.215 on 6 and 3793 DF,  p-value: 1.663e-06
```

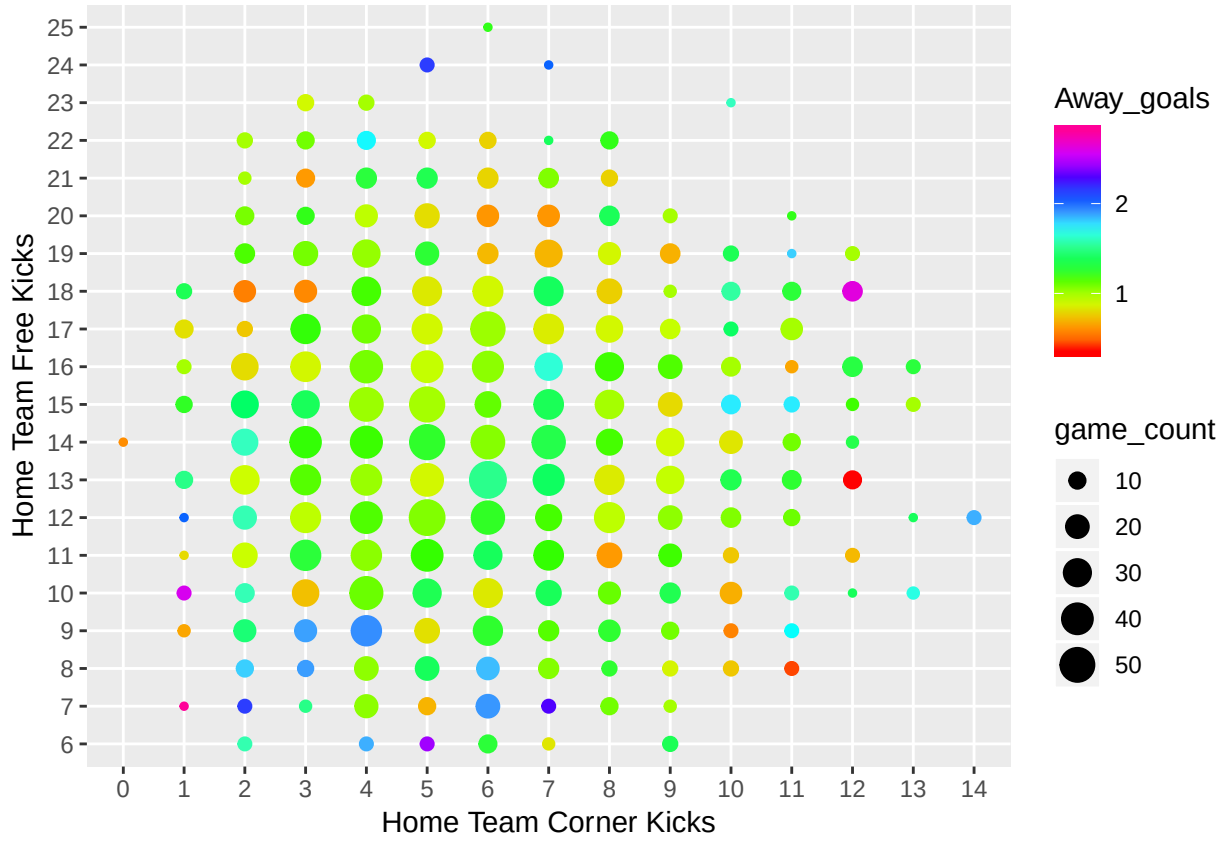
```
# We see that HF, AC, and HF:AC are not significant. We will try again without the HF*AC term.
fit_away = lm(FTAG ~ HC*AF , data=football)
summary(fit_away)
```

```
##
## Call:
## lm(formula = FTAG ~ HC * AF, data = football)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7215 -1.0888 -0.1349  0.8260  6.7435
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.941541   0.144812  13.407  < 2e-16 ***
## HC          -0.089327   0.021769  -4.104 4.15e-05 ***
## AF          -0.052729   0.009782  -5.391 7.45e-08 ***
## HC:AF        0.005682   0.001464   3.882 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.166 on 3796 degrees of freedom
## Multiple R-squared:  0.009313,   Adjusted R-squared:  0.00853
## F-statistic:  11.9 on 3 and 3796 DF,  p-value: 9.441e-08
```

Here, we expected to see a positive correlation between HF and AC with the number of away team goals scored. However, these effects were not statistically significant, and instead, we find that HC and AF are statistically significant linear predictors, with decreases of 0.096 and 0.057 goals per game per event, respectively. There is also a statistically significant interaction between HC and AF:

```
football$HC.f <- factor(football$HC)
football$AF.f <- factor(football$AF)
football %>% group_by(HC.f,AF.f) %>% dplyr::summarise(Away_goals = mean(FTAG),
                                                      game_count = n()) %>%
  filter(game_count >=5) %>% #to remove counts with few data points
  ggplot(aes(x=HC.f,y=AF.f, colour = Away_goals, size=game_count)) +
  geom_point() +
  scale_colour_gradientn(colours=rainbow(10)) +
  ylab("Home Team Free Kicks") +
  xlab("Home Team Corner Kicks")
```



Looking at this graph, we see that, in general terms, the away team scores more goals when the home team has low numbers of corner kicks and free kick, while more home team corner and free kicks typically work together to produce fewer goals for the away team.

Home Team Goals

```
#Home Team goals
fit_home = lm(FTHG ~ HC*AF + HF*AC, data=football)
summary(fit_home)
```

```
##
## Call:
## lm(formula = FTHG ~ HC * AF + HF * AC, data = football)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3135 -0.9076 -0.3367  0.6665  8.2403
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.274441   0.233501   9.741  < 2e-16 ***
## HC           0.070707   0.025831   2.737 0.006223 **
## AF           0.026224   0.011576   2.265 0.023546 *
## HF          -0.059397   0.010243  -5.799 7.23e-09 ***
## AC          -0.067987   0.029610  -2.296 0.021726 *
## HC:AF       -0.005818   0.001728  -3.367 0.000768 ***
## HF:AC        0.003050   0.001975   1.544 0.122646
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.375 on 3793 degrees of freedom
## Multiple R-squared:  0.02733,    Adjusted R-squared:  0.0258
## F-statistic: 17.77 on 6 and 3793 DF,  p-value: < 2.2e-16
```

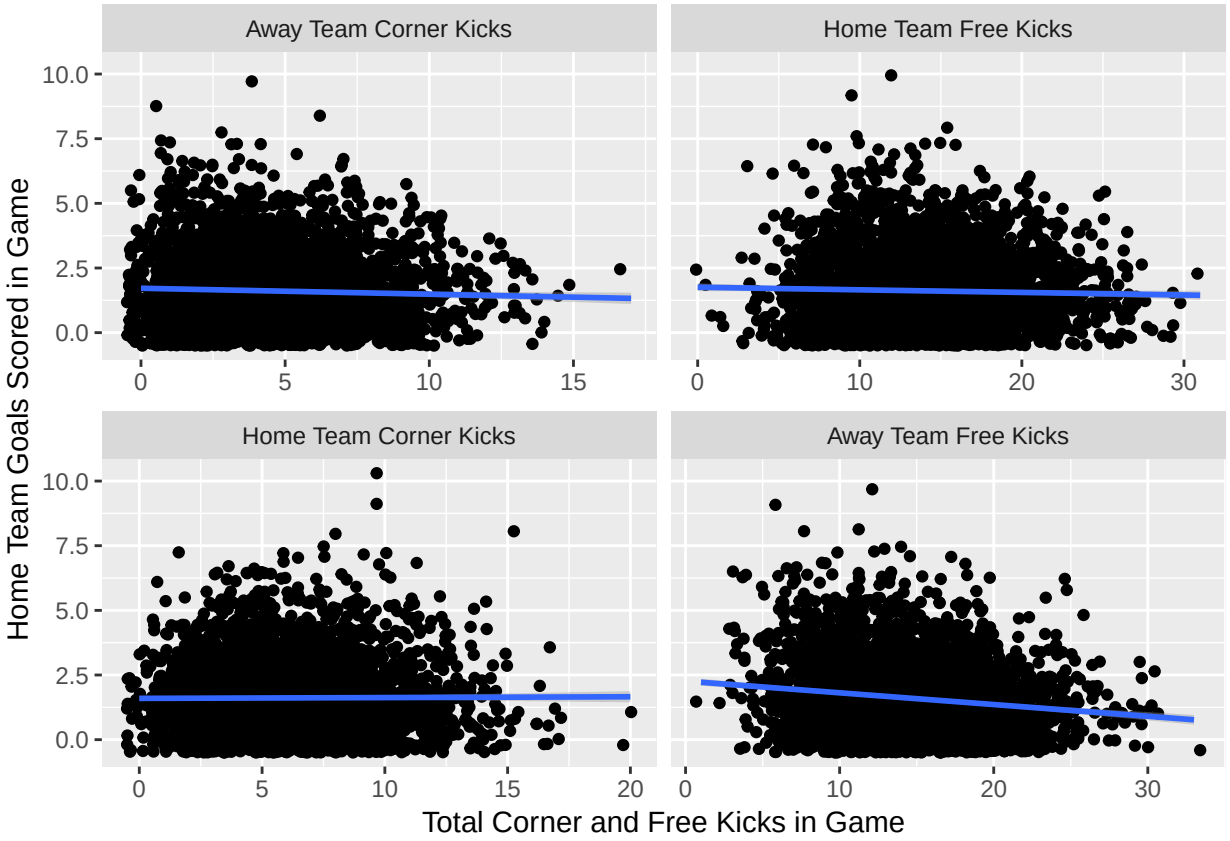
Most of the parameters are significant at the $\alpha = 0.05$ level in this model. HC and AF were associated with higher numbers of home team goals, while HF and AC were associated with lower numbers of home team goals. We try a modified model to better fit the data.

```
fit_home = lm(FTHG ~ HC*AF + HF + AC, data=football)
summary(fit_home)
```

```
##
## Call:
## lm(formula = FTHG ~ HC * AF + HF + AC, data = football)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2561 -0.8951 -0.3353  0.6775  8.2434
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.080371   0.196825  10.570  < 2e-16 ***
## HC           0.070181   0.025833   2.717 0.006623 **
## AF           0.026129   0.011578   2.257 0.024078 *
## HF          -0.045657   0.005074  -8.997  < 2e-16 ***
## AC          -0.024412   0.008968  -2.722 0.006513 **
## HC:AF       -0.005780   0.001728  -3.345 0.000831 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.376 on 3794 degrees of freedom
## Multiple R-squared:  0.02672,    Adjusted R-squared:  0.02544
## F-statistic: 20.83 on 5 and 3794 DF,  p-value: < 2.2e-16
```

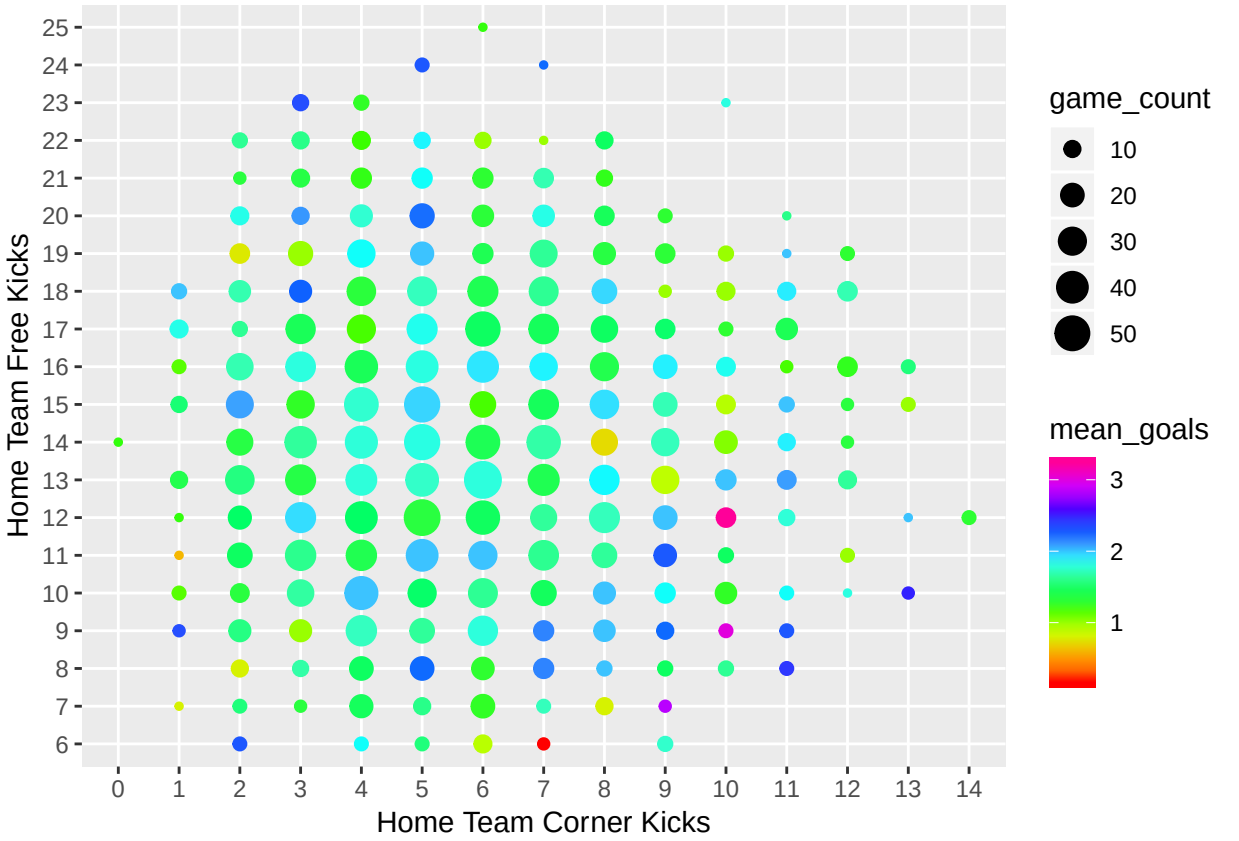
We can look at how the indivaidual parameters are related to the number of home team goals scored.

```
football_long %>%
  ggplot(aes(y=FTHG, x=count)) + geom_jitter(width=0.5, height=0.5) +
  ylab("Home Team Goals Scored in Game") +
  xlab("Total Corner and Free Kicks in Game") +
  geom_smooth(method=lm, se=TRUE) +
  facet_wrap(~param.name, scales = "free_x")
```



We can also look at the HC:AF interaction term, which we also observed in the Away Team goals.

```
football %>% group_by(HC.f,AF.f) %>% dplyr::summarise(mean_goals = mean(FTHG),
                                                      game_count = n()) %>%
  filter(game_count >=5) %>% #to remove counts with few data points
  ggplot(aes(x=HC.f,y=AF.f, colour = mean_goals, size=game_count)) +
  geom_point() +
  scale_colour_gradientn(colours=rainbow(10)) +
  ylab("Home Team Free Kicks") +
  xlab("Home Team Corner Kicks")
```



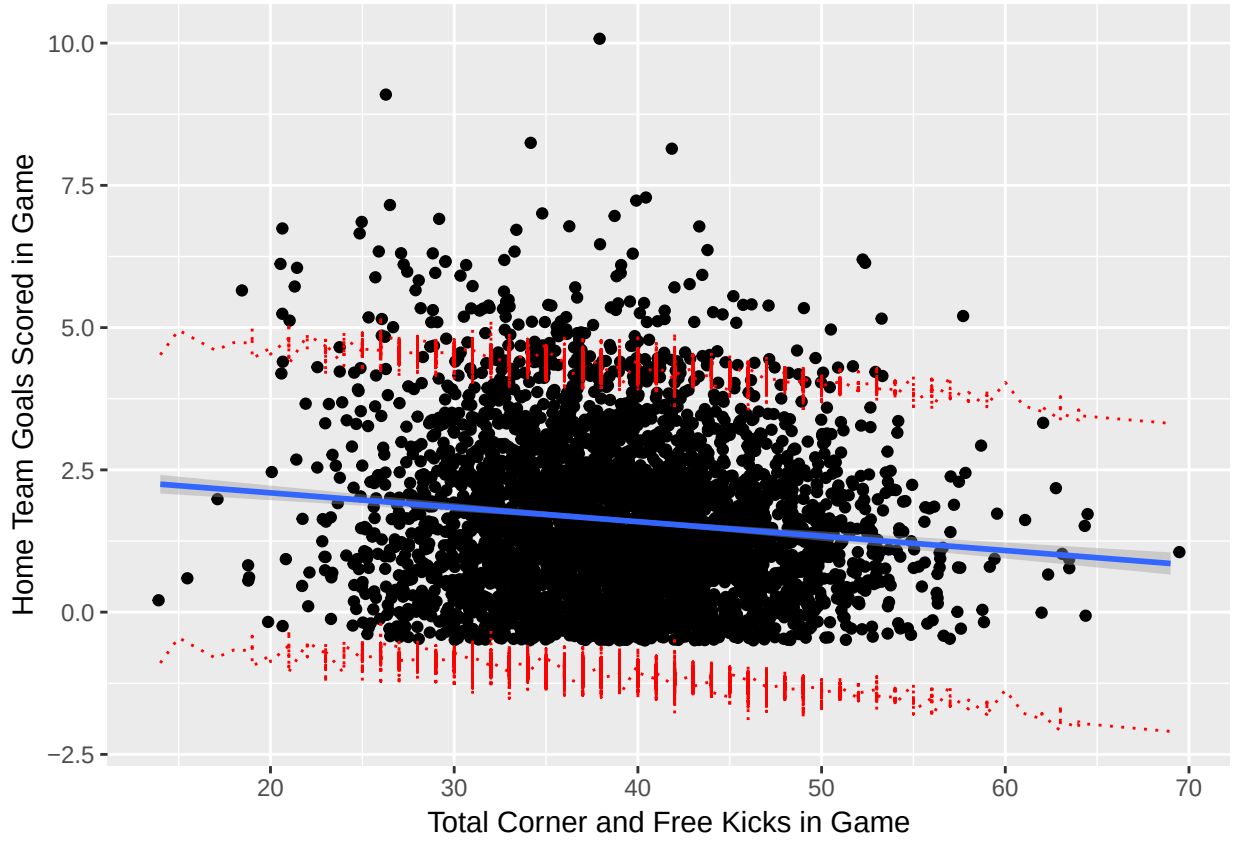
It's not obvious what the interaction is, other than the fact that there are differences of the mean number of goals between different combinations of the number of corner kicks and free kicks given to the home team, thought perhaps a higher corner kick to free kick ratio for the home team often is associated with more goals for the home team.

Linear Model Discussion

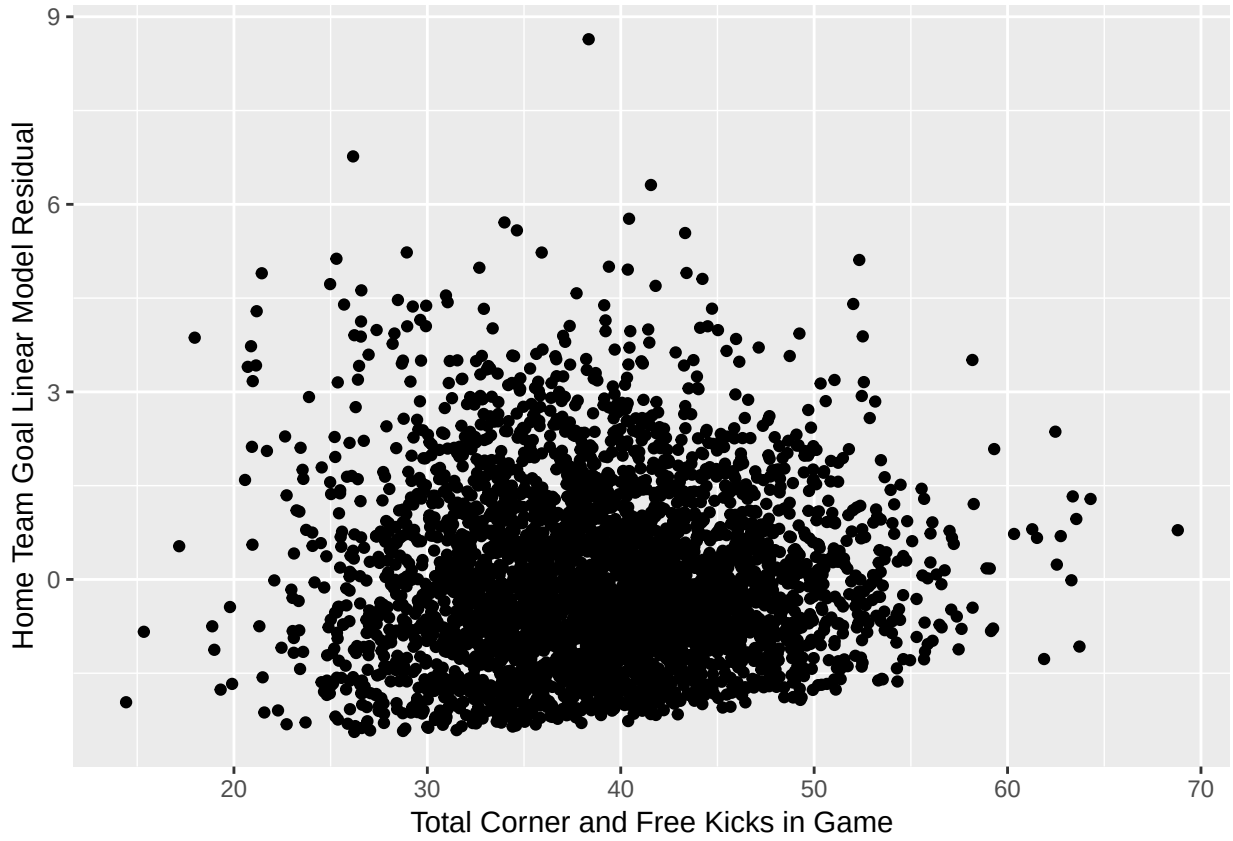
Perhaps the greatest surprise was the persistent trend of lower numbers of goals scored with greater numbers of free and penalty kicks, as opposed to increases in scores as expected. In soccer, the game timer is never stopped except at half time. Anytime play is stopped, the timer is continued, and at the end of each 45-minute half, up to 5 minutes of play is added to compensate for some of this lost play time; however, the amount of time lost to play stoppage is typically much greater than the time added back at the end of the half. This means than in increase in the number of corner kicks and free kicks, which were the result of restarting play after play stopped, produces a game with less actual playing time, and therefore less time to score goals. Unfortunately, these effects are quite small compared to the error term in the linear models. To firther explote this, we will consider the Home Team Goals, as this linear model had several statistically significant terms. The prediction margins are quite large, as expected.

```
FTHG_predict <- predict(fit_home, interval="prediction")
football_predict <- cbind(football, FTHG_predict)
football_predict %>%
  ggplot(aes(y=FTHG, x=HF+AF+HC+AC)) + geom_jitter(width=0.5, height=0.5) +
  ylab("Home Team Goals Scored in Game") +
  xlab("Total Corner and Free Kicks in Game") +
  geom_smooth(method=lm, se=TRUE) +
  geom_line(aes(y=lwr), color = "red", linetype = "dotted")+
```

```
  geom_line(aes(y=upr), color = "red", linetype = "dotted")
```




```
football_predict$resid <- football_predict$FTHG - football_predict$fit
football_predict %>%
  ggplot(aes(y=resid, x=HF+AF+HC+AC)) + geom_jitter(width=0.5, height=0.5) +
  ylab("Home Team Goal Linear Model Residual") +
  xlab("Total Corner and Free Kicks in Game")
```




```
abs_res_mean <- mean(abs(football_predict$resid))
FTHG_mean <- mean(football_predict$FTHG)
```

The residuals are also quite substantial. Indeed, the mean absolute residual was 1.09 goals, which is nearly as large as the mean number of goals scored by the home team, which was 1.61 goals. As a result, we conclude that while more free kicks and corner kicks are associated with lower home team, away team, and total goals scored, possibly due to less playing time from game stoppage, the number of these kicks has little predictive power in determining the number of these goals. Alternatively, more fouls may be committed in low-scoring games due to frustration over inability to score. Without temporally resolved data, it is unfortunately not possible to determine these types of causal relationships.