Exploring the Effects of Weather on Points Scored in NFL Games

Peter Christenson

12/27/2023

Contents

1 Introduction

In this project, I explore the relationship between weather-related variables and the total amount of points scored by both teams combined in NFL games played from the 2000-2022 seasons. I approach this in two ways:

- Conducting a hypothesis test to assess whether more points are scored in games played in indoor stadiums than in games played in non-indoor stadiums
- Constructing a regression to assess the relationship between weather-related variables and total points scored in games played outdoors

This analysis is largely exploratory and a more rigorous analysis would be necessary to fully assess the relationship between weather and scoring. However, the results of this analysis suggest that:

- The mean number of total points scored by both teams combined per game in games played indoors is greater than the mean number of total points scored by both teams combined per game in games played outdoors
- Weather by itself is a weak predictor for total scoring

2 Data

I web scraped boxscores of NFL games from the 2000 through 2022 seasons from Pro Football Archives, which provides scores for games as well as other characteristics such as location and relevant weather information.[1]

[1] While I could not find any explicit statement that scraping from this website is a violation, I advise that any scraping done be done responsibly and with caution.

After some cleaning, I have a dataset with the scores of all games, the teams involved, the location, and weather information.

Table 1: Cleaned Web Scraped Data

season	team_1_pts	team_2_pts	date	location	venue	team_1	team_2	weather_type	temp_f	wind_mph	humidity_pct	indoors	total_pts
2000	16	21	2000-09-03	East Rutherford, NJ	Giants Stadium	Arizona Cardinals	New York Giants	Haze	84	0	75	0	37
2000	16	0	2000-09-03	Pittsburgh, PA	Three Rivers Stadium	Baltimore Ravens	Pittsburgh Steelers	Haze	81	5	74	0	16
2000	17	20	2000-09-03	Landover, MD	FedExField	Carolina Panthers	Washington Redskins	Overcast	79	7	79	0	37
2000	27	30	2000-09-03	Minneapolis, MN	Metrodome	Chicago Bears	Minnesota Vikings	Indoors	NA	NA	NA	1	57
2000	14	10	2000-09-03	New Orleans, LA	Louisiana Superdome	Detroit Lions	New Orleans Saints	Indoors	NA	NA	NA	1	24
2000	27	14	2000-09-03	Kansas City, MO	Arrowhead Stadium	Indianapolis Colts	Kansas City Chiefs	Clear	90	3	41	0	41

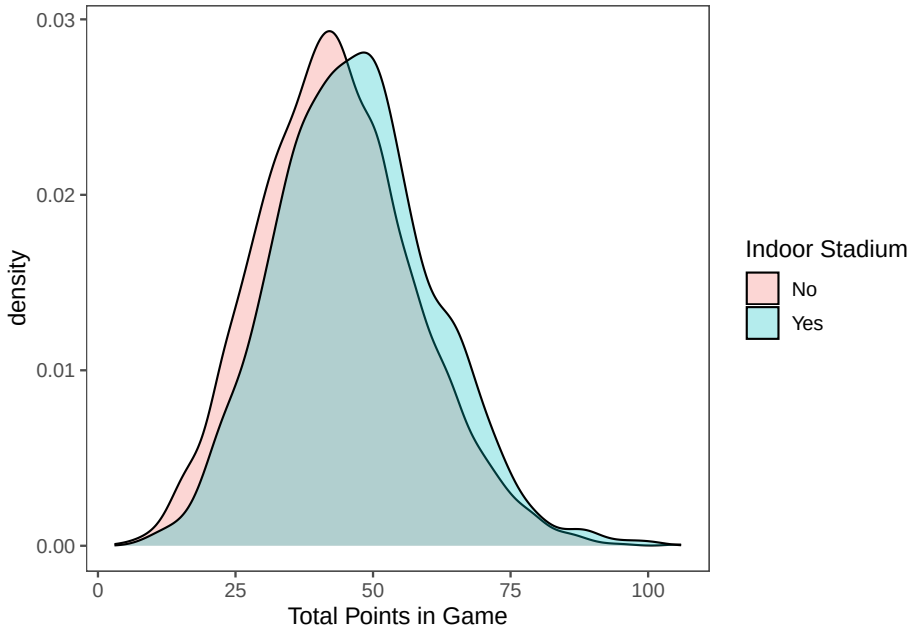
Note that the above table is just 5 rows of the cleaned data.

3 Analysis

3.1 Hypothesis Testing

I was first interested in exploring the difference in scoring between games played indoors versus games played outdoors. Specifically, I was interested in conducting a two-sample t-test comparing the mean number of points scored in both circumstances.

Figure 1: Kernel Density Plot of Total Points Scored by Stadium Type



Before I could perform the t-test, I checked:

1. Whether both groups followed a normal distribution
2. Whether there is homogeneity of variance

As both groups have a large amount of observations, I assume the Central Limit Theory holds and that both groups are normally distributed.

Table 2: Number of Games by Stadium Type

indoors	num_games
No	4894
Yes	1258

To compare the variances of both groups, I conduct an F test.

```
##
##  F test to compare two variances
##
## data:  points_indoor and points_non_indoor
## F = 1.0049, num df = 1257, denom df = 4893, p-value = 0.9061
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9215953 1.0982422
## sample estimates:
## ratio of variances
##           1.004891
```

Per the results of this test, I assume the variances are equal due to the p-value being greater than 0.05.

Since both conditions are satisfied, I can move forward with the t-test. Since indoor games provide a more controlled football-playing atmosphere (temperature, wind, precipitation, etc.), my hypothesis was that games played indoors have more points scored. Let group 1 represent games played indoors and group 2 represent games played outdoors, and μ represent the mean total number of points scored per game.

H_0: $\mu_1 \leq \mu_2$

H_a: $\mu_1 > \mu_2$

```
##
##  Welch Two Sample t-test
##
## data:  points_indoor and points_non_indoor
## t = 7.6405, df = 1949.6, p-value = 1.685e-14
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.683659      Inf
## sample estimates:
## mean of x mean of y
##  46.89746  43.47711
```

As the p-value is near zero, I reject the null hypothesis and conclude that the mean total number of points scored per game is greater in games played indoors than in games played outdoors.

3.2 Regression

I also performed regression analysis to assess the effects of weather on scoring. The first model I constructed considered only variables that pertain to weather as independent variables, specifically rain, snow, temperature, wind, and percent humidity. My dependent variable is total points scored in the game. The model is then as follows:

$$\text{Total Points Scored} = \beta_0 + \beta_1 D_{rain} + \beta_2 D_{snow} + \beta_3 Temperature + \beta_4 Wind + \beta_5 Humidity + \varepsilon$$

Where D_{rain} and D_{snow} refer to binary 'dummy' variables that indicate the presence of rain or snow, and $Temperature$, $Wind$, and $Humidity$ referring to the temperature (in degrees Fahrenheit), the wind speed (in mph), and percent humidity, respectively.

```
##
## Call:
## lm(formula = total_pts ~ rain + snow + temp_f + wind_mph + humidity_pct,
##     data = regression_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.460 -10.016  -0.693   8.960  62.158
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  46.861298   1.211339  38.686  < 2e-16 ***
## rain1        -4.040155   0.955035  -4.230 2.38e-05 ***
## snow1        -1.676979   1.691829  -0.991    0.322
## temp_f       -0.014763   0.012871  -1.147    0.251
## wind_mph     -0.255360   0.040148  -6.361 2.21e-10 ***
## humidity_pct -0.001511   0.011819  -0.128    0.898
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.04 on 4529 degrees of freedom
## Multiple R-squared:  0.01429,    Adjusted R-squared:  0.0132
## F-statistic: 13.13 on 5 and 4529 DF,  p-value: 1.022e-12
```

These results suggest weather by itself is a fairly weak predictor for the total amounts of points scored in a game.

To improve the regression, I introduced new variables: the location of the game, and the two teams involved.[2] The new model is then constructed as:

$$\begin{aligned}\text{Total Points Scored} = \beta_0 + \beta_1 D_{rain} + \beta_2 D_{snow} + \beta_3 \text{Temperature} + \beta_4 \text{Wind} \\ + \beta_5 \text{Humidity} + \beta_6 \text{Location} + \beta_7 \text{Team}_1 + \beta_8 \text{Team}_2 + \varepsilon\end{aligned}$$

The results of which are below.

```
## Multiple R-squared: 0.3838
## Adjusted R-squared: 0.1353
```

Comparing the R^2 values across both models, the inclusion of location and the teams playing improves the fit of the model quite a bit.

[2] I consider the interaction between the season of the game and two teams. For example, if a game took place in the 2017 season between the Packers and the Vikings, I would consider the 2017 Packers and 2017 Vikings as independent variables (as opposed to the teams by themselves without any season).