NFL Sports Betting Analysis: Betting for Starters

Stadium Type Analysis

By: Justin Butler

For my section of our project, I chose to analyze how betting on an NFL game may be affected by the type of stadium the games are played in. Whether games played indoors in dome style stadiums are approached differently than games played outdoors in open-air style stadiums. Considering that harsher weather generally plays a factor on games by making it harder for teams to score. For e.g.: High wind speeds making it harder for the quarterbacks to throw to their receivers or more difficult for kickers to kick field goals and extra points after touchdowns. From this topic I will analyze our group’s acquired data from the 2013 to 2023 NFL seasons and pose two more precise questions and in my conclusion make a recommendation on how I would bet on a future NFL game. Now let’s get into the data!

Question 1) – Is stadium type, being either indoor or outdoor factored in when the over/under line is set for games ?

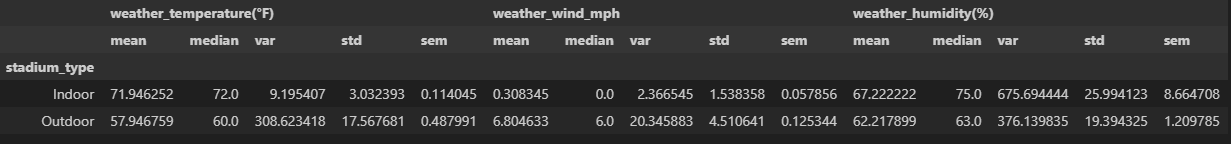
To answer my questions I first used the value.counts() function on the stadium column in our csv data file, which contained the names of all the stadiums played in over the past 10 years to get the full stadium list so I could later filter it by indoor and outdoor and place these values in a newly created column titled “stadium\_type”. I used the following for loop to achieve the desired result:

A picture containing text, screenshot, font

Description automatically generated

With this new data column, I could now analyze all the rest of the data from the csv file through the prism of games played indoors and outdoors. For instance, the below summary statistics table of the tracked weather metrics for all the games to provide some evidence for my consideration of weather being a large factor on outdoor games opposed to indoor games. Achieved by performing the groupby() and aggregation (agg([“mean”, “median”, “var”, “std”, “sem”]) functions on the stadium\_type column against the 3 weather columns for temperature, wind and humidity.

Weather Summary Statistics Table:



The table supports the idea that the weather in outdoor stadiums is much harsher and unpredictable than indoor stadiums with such high statistic measurements such as variance and standard deviation. With an outdoor temperature variance 30 times greater than indoor temperatures, a standard deviation almost 6 times greater than indoor, outdoor wind speed variance almost 10 times greater than indoor and standard deviation almost 3 times greater, the analysis clearly shows tremendous variability and value spread across the weather data. Meaning the overall weather at outdoor stadiums can be significantly different than indoor stadiums due to things like changing weather conditions, time of year and geography.

To answer Question 1) better I formed the following hypotheses:  
Alternate Hypothesis: If stadium type is factored into how the over/under lines are set. Then the average over/under line for games played indoors should be higher than the games played outdoors.

Null Hypothesis: If stadium is NOT factored into how the over/under lines are set. Then the average over/under line for games played indoors will NOT be higher than the games played outdoors.

To answer these and see how the stadium type affects a betting measure such as over/under through statistical analysis I again used the groupby() and aggregation functions to analyze the over/under line of each game against the stadium type. To produce the following summary statistics table:

Over/Under Line by Stadium Type Summary Statistics Table:

A screenshot of a graph

Description automatically generated with low confidence

To visualize the data between these 2 columns and the summary table I chose a box-and-whisker plot for its ability to show mean, quartiles, outliers and overall potential skew of the data. Seen below:

A picture containing diagram, screenshot, text, line

Description automatically generated

Once plotted I was able to see what was also displayed by the weather summary statistics table, a much higher degree of spread in the Outdoor values than the Indoor values. Shown in one example, with the much higher number of outliers in the Outdoor values than the Indoor Values. This finding makes sense in connection with the weather analysis: much more variability in the potential weather during Outdoor games ties into the concept of much more variability in over/under lines set for Outdoor games. It also shows the in a clear visual sense that the Indoor mean is higher than the Outdoor mean.

To test my hypotheses, I performed an independent 2-sample T-test as I was comparing 2 distinct groups, the over/under lines of games played indoor versus those of games played outdoors.

A screen shot of a computer program

Description automatically generated with low confidence

This produced an extremely low P-value, assuring a very high confidence level and proving what my alternative hypothesis stated. That the average over/under line of Indoor stadiums is higher than that of Outdoor stadiums. Allowing us to at least assume that the stadium type in some capacity is factored in when the over/under lines are set for games. Although it is important to note that, due to the extensive data and other factors considered by those who set these betting over/under lines, our results thus far are not enough for me to make a definitive statement for or against our original question, only a recommendation leaning towards a small correlation.

Question 2) - Is there a homefield advantage to teams playing their home games indoors versus outdoors ?

To analyze this question, I used another for loop to iterate through the “score\_home” and “score\_away” columns and count the column with the higher value as the “game\_winner”. Whose value would be then stored in the newly created column of the same name with the same 3 letter abbreviation as seen in the “team\_home” or “team\_away” columns.

A screenshot of a computer program

Description automatically generated with low confidence

The next step was to determine if the home team won, this was done by creating a new column titled “home\_team\_won” and applying the anonymous “lambda” function to it by comparing the values of the “team\_home” and “game\_winner” columns row by row. When the two values are the same then that means the home team won that game, and a 1 is placed in the “home\_team\_won” column. If the away team won then a 0 is placed. 1s and 0s are used here so I can later divide the total column sum() of the “home\_team\_won” column by the count() (tota)l of it. As seen below.

A picture containing text, screenshot

Description automatically generated

Which produced a nice clean DataFrame:

A picture containing text, screenshot, font

Description automatically generatedTo visualize this data, I chose a simple bar plot. I felt this was the best choice as the 2 vertical bars this would create, one displaying the percentage of home games won by NFL teams whose home field is Indoors and the other showing home games won by NFL teams whose home fields are outdoors, would allow for easy visual comparison of the percentage values, thus any potential conclusions could be easily drawn.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Important to note that most teams in the NFL play in Outdoor Stadiums. Only 10/32 teams have Indoor Stadiums for their home fields as of 2023. This can be reflected in the data as 1983 of the 2707 games in the csv data were played in Outdoor stadiums over the past 10 seasons. Even still, the bar plot shows a very even playing field (no pun intended) over the past 10 seasons worth of games. With game data recorded over thousands of games, the comparison of the 2 values reveals almost identical percentages at 54% and 56% respectively. The data and subsequent bar plot show more of a skew towards home teams in general having an advantage than an advantage to the stadium type the home teams play in. The corresponding data to these figures would mean 46% of teams have lost their away games that they played indoors and 44% of away teams have lost games played in Outdoor stadiums. Therefore, I must conclude that I see no significant correlation towards a home field advantage for either home stadium type. The slight 2% higher value for the Outdoor stadium teams can be nullified by the fact that there have been 2.73 times more Outdoor games than Indoor games over the past 10 seasons, a slightly higher value is not a surprise or statistically significant given the size discrepancy in games played Outdoor vs. Indoor.

CONCLUSION & LIMITATIONS:

After all this analysis, I believe I can make some modest recommendations towards placing a bet on an NFL game based on the stadium type the game is played in. First, I would advise betting on the over/under line opposed to placing a bet on one team to win over the other. Based on some positive correlation results towards over/under line based on stadium type. Whereas no real correlation was found on a home field advantage. When making your bet I would look for 2 things based on my findings, expected weather for the game and corresponding stadium type. Based on the extreme variability I uncovered with the weather of outdoor NFL games and the data I found on average over/under lines. I would look for games with poor expected weather that will be played outdoors and bet the games point total to be under the over/under line.

Some limitations to my analysis include:

* Overall data limitations, I originally chose to examine growth in the NFL betting industry and was going to try to predict future growth in the market based on my findings. Although I was unable to find any accessible or clean data.
* Data limitations in the data my group found. Several rows in the weather columns possessed Nan values and limited the scope of this section of data. For ex: more data in the “weather\_detail” column such as rainy or sunny could have allowed for more in-depth analysis. E.g., average score for games labeled as rainy vs. sunny.
* Ability/Time constraints. Due to the need to change project topics, a few days were lost at the beginning of the project timeline. Also, my experience in Data Analytics limited my ability to analyze our csv data, or any data for that matter, at a more detailed level. In the future with more skills such as “scrapping” more detailed data from multiple sources would likely be available, thus allowing for a more detailed analysis.