**Portfolio Milestone- Module 7- Final Paper**

Shawn Poling

Colorado State University Global

MIS 581 – Capstone: Business Intelligence and Data Analytics

Dr. Justin Bateh

2022, October 2

**Abstract**

This paper is focused on exploring the potential relationships that exist between weather and player performance in the National Football League. This research is intended to be a source of information for both front and back-office operations and provides an opportunity to consider improvements on in-game strategy to align with expect weather conditions. Using data gathered on NFL play-by-play data, linear regression is used to produce a model to compare the potential relationship between yards gained per play and various weather variables. While the model was not shown to be a reliable predictor of performance, there is evidence to show that wind has a negative effect. This project provides the opportunity for further research into other components of weather that can potentially have an effect on yards gained. Additional areas of research could include play type, biomechanical data from IoT devices, and nutritional factors.

*Keywords*: Sports Analytics, National Football League, Linear Regression, Weather

**Portfolio Milestone- Module 7- Final Paper**

Insights from data is nothing new in the world of sports. Like many industries, the volume, velocity, veracity, and variety of data has been a key driver in improving many important areas from operations to customer engagement. The National Football League is committed to taking a similar approach in their attempt to find value in data. According to NFL Football Operations (n.d.), sensor technology has improved to the point of being able to capture more than 200 data points on every play of every game. These tracking systems have the capability to record speed, location, acceleration, and also chart individual movements. While there is focus on the predictive capabilities of these metrics to improve outcome probability and fan experience, there is an opportunity to expand on this analysis to include the effects of weather on overall performance.

The ability for individual teams to have insights into the effect that weather has on player performance could lead to various benefits for the organization. This could lead to improvements in safety protocols, influence play selection, gain insight into the opposing team, and even lead to modification of nutritional planning and training. With every team being run as an organization, there is also the potential to add value towards improving overall operations. According to Ozanian and Settimi (2021), the average team in the NFL is worth over 3.5 billion dollars and the Dallas Cowboys lead the league with revenues in 2021 of over eight-hundred million dollars.

The geographic diversity of team location presents unique challenges given the time of the year professional football takes place. As climate change continues to be a major contributor to changing weather patterns, teams need to appropriately plan for the effects on both player and operational performance. Having insight into how weather events could affect various performance metrics could lead to substantial changes in strategy for both sides of the operation.

**Objectives**

The purpose of this paper is to quantify the impact weather has on teams in the National Football League and highlight the potential challenges and opportunities for individual teams and the overall organizations in preparing for more extreme weather. The purpose statement is condensed to measuring the impact of weather on performance in the National Football League.

This research is intended to be a source of information for teams and front office operations. For organizations with stadiums subject to the elements, this will provide a better view on the overall impact weather can have on overall performance. Having insight into the effects of extreme weather on player and operational performance will give teams an opportunity to plan more effectively. This could lead to more accurate financial budgeting and financial projections or helping with financing of projects or enhancements to stadium infrastructure. There is also an opportunity to add value to player performance. This could include on-field infrastructure such as turf selection and temperature control devices or be more focused on playing strategy and modifications to player training. Playing strategy would be focused on adjustments to play calling and player selection where training could encompass nutritional and physical modifications.

**Research Question and Hypotheses**

This research paper aims to quantify the impact weather has had on prior seasons and incorporate a model to plan more effectively given projections in climate change. Literature review on the research area of the impact of weather on player performance has yielded few results and presents an opportunity to provide work in this focus of study. The formal research question is listed as follows:

* What is the impact of weather on player performance in the National Football League?

In an attempt to help answer this question, there are several hypotheses that will be tested. The importance of providing a hypothesis is to take the research question a step further in an attempt to offer a statement as to the potential relationship amongst the variables (O’Leary, 2021). The focus on hypotheses development is centered around discovery of the potential correlations between weather conditions and performance. Performance will be analyzed in terms of player performance. A null and alternate hypothesis will be used to test the association between weather and performance. The null will be used to describe a result that shows a lack of association between the predictor and outcome variable. The following hypotheses will be tested:

* Ho: There is no relationship that exists between player performance and weather
* Ha: There is a statistically significant relationship that exists between player performance and weather

**Literature Review**

Depending on the specific team location, seasonality will have a potential effect on different ends of this spectrum. There is a significant amount of research into the physiological effects that weather has on professional athletes. In an article by Borresen (2008), the impact of environmental conditions on athletes for the 2008 Beijing Olympics games is discussed. Some of the factors covered that effect performance include temperature, allergens, ultraviolet radiation, pollution, and altitude (Borresen, 2008). Additional support can be found in the works of Peiser et. Al (2006), where the seasonality of weather and surface conditions are analyzed in relation to the effect it has on injury risk and athletic performance. There is also research supporting the differences in the effects of extreme heat and cold on player performance. According to Phetteplace (2000), fine motor skills, accuracy in processing information, physical strength, awareness, and reaction time are significantly reduced for players in the presence of low temperatures. Grundstein et. al (2014) provide research into the hazards of extreme heat for American Football Players. This research covers the dangers of extreme heat on the southern tier of the United States and proposes modification of outdoor training times to reduce the likelihood of heat related illness. Further research into the area of modification of game times for NFL team located in these regions could be beneficial.

Additional studies included research more focused on outcomes and potential advantages based on geographic location. In research by Brocherie et. al (2015), there was a strong correlation between successful outcomes and temperature increases for soccer clubs in the Gulf Cooperation Council region when playing clubs in other regions. In additional research by Borghesi (2007), results from analysis suggest that game day temperature significantly affects team performance. The literature review highlighted areas of team performance but there is an opportunity to explore the effects on individual performance. This research project focuses on the more granular view of individual performance and how it can vary according to environmental conditions.

**Methodology**

The R package used in this research is called nflfastR and was developed by Sebastian Carl and Ben Baldwin. According to Carl and Baldwin (2022), this package was created by utilizing a set of functions to scrape play-by-play data going back to 1999. Given the large size of the potential data available from this scraping program, there is a need to focus on narrowing the scope.

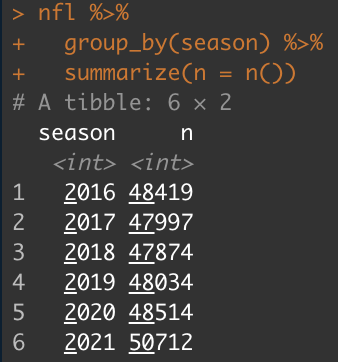
The dataset for this project will be manually created based on specific criteria set within the nflfastR package. The data that can be gathered from this procedure is wide enough to obtain the measures applicable to the intended analysis. Some of the important considerations to arrive at this determination included the ability to obtain all play-by-play data on individual players, teams, various values relating to the type of play and total yards, playing facility details, and weather conditions.

To begin, the decision was made to analyze the output from several time periods starting with the most recent season, 2021. The first attempt spanned a total of ten years and included over six-hundred thousand lines. This also presented an opportunity to learn more about the various dimensions and potential issues within some of the data. The total dimensions in this dataset are 372. There is a clear need to reduce this to a smaller number given a large amount of the dimensions are not critical to the intended analysis. The dimensions represent various types of variables, including nominal, categorical, numeric, continuous, and discrete.

Another important factor was viewing the potential nulls and irregular values within the data. Upon further analysis, it was discovered that games prior to 2016 had inconsistent weather data. The decision was made to reduce this period from the 2016 to 2021 seasons. This reduction still provides a significant volume of data. The totals number of plays for each season is highlighted in Figure 1 and the total dimensions of the reduced dataset are shown in Figure 2.

**Figure 1**

*Totals Plays in Dataset by Season*

**

**Figure 2**

*Dimensions of NFL Dataset*

*Text

Description automatically generated*

***Methods***

In order to test the hypotheses, a quantitative methodology will be used. The goal is to utilize the raw data to discover the potential relationship that exists between performance and weather conditions. This will be done through the use of descriptive and inferential statistics in R Studio. Descriptive statistics will focus on summarizing variables through measures of dispersion and central tendency. This will also be helpful in understanding if certain statistical tests can be used on the data.

In addition to the descriptive tests, predictive statistics will be used to test the chosen hypotheses. The exact model to test the hypotheses will be clearer after EDA and descriptive tests. A potential solution is to use multivariate analysis to determine if there is a statistical relationship between the dependent variable of performance and the predictor variables associated with weather conditions. However, if the descriptive test violates the assumptions needed for model reliability, it may be necessary to approach the question with a different method such as a non-parametric test.

It is possible that multiple models will be used and compared to arrive at a conclusion about the association between performance and weather. When engaged in the predictive modelling tests, the data will be broken into a training and validation set made up by sixty percent and forty percent of the full data set, respectively. The purpose of the data split is to refine and validate the model. If successful, the model could be applied to new data for the upcoming season.

***Limitations***

The scope of this research is narrowed to fit within the confines of data accessibility. The data within this project is legally obtained and publicly available. Further research is possible to look into the use of sensor data on individual players but that is outside the capacity of this research. Limitations in obtaining sensor data stem from budgetary and time constraints, NFL team approval, and individual player approval.

Another consideration is in relation to the accuracy of weather data. This study will attempt to highlight the effects of weather on player performance. As such, there is subjectivity towards what is considered favorable versus unfavorable temperature. For example, eighty degrees will feel different depending on the relative humidity of the location. Attempts are being made to appropriately account for this variable but may be constrained due to time and complexity of data transformations.

***Ethical Considerations***

Data for this research project is publicly available so there is not a need to consider consent, confidentiality, potential for harm, or results communication. However, consideration into how the data was obtained was researched. According to the NFL (2017), there is not direct mention prohibiting web scraping unless it is obtained for commercial use or substantially interferes with the service. In a recent case around the legality of web scraping publicly available data, the United States Ninth Circuit of Appeals affirmed its decision that it is not a violation of the Computer Fraud and Abuse Act (Whittaker, 2022). The overall use of the data and the proposed methodology would not require ethical review.

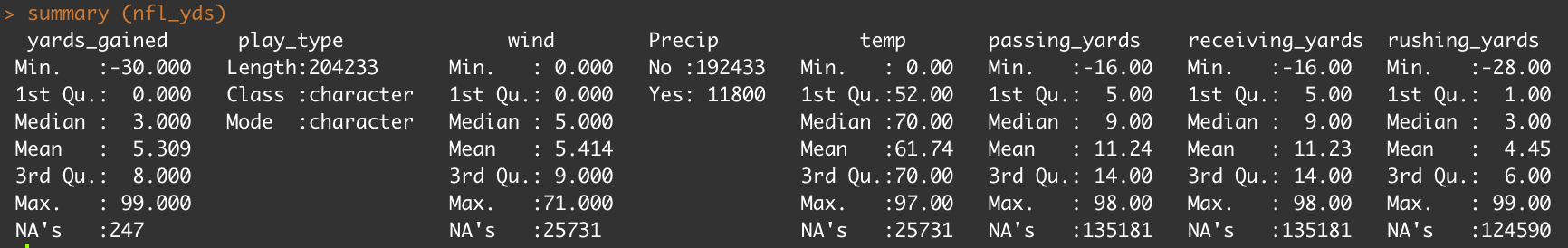
**Findings**

Prior to analysis, significant time was spent analyzing available data and getting a better understanding of the variables within the datasets. The final data set for analysis included the values player\_id, yards\_gained, play\_type, wind, Precip, temp, passing\_yards, receiving yards, and name. Filters were placed on the data to only include plays with a rush or pass and excluded null player\_id values. The mutate function was used to create case statements to adjust values for conditions related to playing football in an indoor stadium. This included adjustments for wind, precipitation, and temperature. These steps were necessary to normalize the data.

Analysis of the data began with a view of the summary statistics. As shown in Figure 3, the number of null values within the data is greatly reduced due to transformations in the cleaning step. Play type such as passing, receiving, and rushing maintain null values as the granularity of the data is a single play. Each play will only be associated a pass, rush, or reception. For the dependent variable, yards\_gained, we can see that the max value includes 99.00 where the min does include negative yards. It is important to note that yards\_gained represents the total yards involved in a single play so values cannot exceed 100 but can be negative if there was a loss of yards. There was a notable difference in the number of plays with precipitation compared to no precipitation. This is further represented in Figure 4.

**Figure 3**

*Summary Statistics of NFL Yards*

**

*Note.* Restricted to show the variables used for the linear regression model

**Figure 4**

*Plays with Precipitation Versus No Precipitation*

*Chart, scatter chart

Description automatically generated*

Next, the potential relationships amongst the variables were explored. Given the volume of data with the data set, it was appropriate to use scatterplots to visualize the potential relationship. Figure 5 highlights the relationship between Temperature and Distance. It appears as if there is a slight increase in yards\_gained with temperatures near the median range. In areas higher temperatures, values exist but the concentration is lower. There appears to be a strong negative correlation between yards\_gained and wind levels. As shown in Figure 6, yards\_gained drops as wind increases in intensity. The relationship between field goal result in different temperatures and wind speeds was also explored as shown in Figure 7 and Figure 8. This appears to indicate the colder temperatures and higher winds tend to be associated with a higher percentage of missed field goals.

**Figure 5**

*Scatter Plot of Yards Gained and Temperature*

*Chart, scatter chart

Description automatically generated*

**Figure 6**

*Scatter Plot of Yards Gained and Wind*

Chart, scatter chart

Description automatically generated

**Figure 7**

*Field Goal Result by Temperature*

*Chart, scatter chart

Description automatically generated*

**Figure 8**

*Field Goal Result by Wind Condition*

*Chart, scatter chart

Description automatically generated*

The predictive module for this data will be applied using linear regression. To begin, a training and testing data set was created at a split of sixty and forty percent respectively. Yards\_gained was compared with play\_type, wind, Precip, and temp. The results of the output, as shown in Figure 9, indicate that wind does test significant with a p-value of less than 0.05. While this variable appears to be a strong indicator of yards\_gained, the low R-squared value indicates that the model is not a good predictor of the target. This is further evidenced by the computation of Root Mean Square Error using the predict () function on the training data set. RMSE explains how much, on average, the predicted value will be from the actual value. As shown in Figure 10, a value of 456.11 is quite large and would not be useful for predicting yards\_gained.

**Figure 9**

*Linear Regression Model Results*

*Text

Description automatically generated with low confidence*

**Figure 10**

*Root Mean Square Error and R-Square*

*Graphical user interface, text

Description automatically generated*

***Summary***

While the results show the model is not a great predictor of yards\_gained, the analysis did show that wind is a variable of importance. The result show that as wind increases, yards\_gained does decrease. With a p-value of less than 0.05, we can reject the null hypothesis that there is not a relationship between player performance and weather. There is enough evidence to suggest that wind plays a role in player performance.

With this research, coaches and players can have more insights into how they have performed under various weather conditions. This is another data point that can be used for training and equipment adjustments, as well as modifications to in-game strategy. There is also an opportunity to explore other components of weather that can potentially have an effect on yards\_gained. Other analysis relating to the effect of weather on various play types is a possible extension of this research.

**References**

Borghesi, R. (2007). The home team weather advantage and biases in the NFL betting market.

Journal of Economics & Business, 59(4), 340–354. https://doi.org/10.1016/j.jeconbus.2006.09.001

Borresen, J. (2008). Environmental considerations for athletic performance at the 2008 Beijing

Olympic Games. International SportMed Journal, 9(2), 44–55.

Brocherie, F., Girard, O., Farooq, A., & Millet, G. P. (2015). Influence of weather, rank, and

home advantage on football outcomes in the Gulf region. Medicine and Science in Sports and Exercise, 47(2), 401–410. https://doi.org/10.1249/MSS.0000000000000408

Grundstein, A., Cooper, E., Ferrara, M., & Knox, J. A. (2014). The geography of extreme heat

hazards for American football players. Applied Geography, 46, 53–60.

https://doi.org/10.1016/j.apgeog.2013.10.007

NFL (2017). *NFL.com – Terms and Conditions*. https://www.nfl.com/legal/terms

NFL Football Operations (n.d.). *NFL Next Gen Stats*.

<https://operations.nfl.com/gameday/technology/nfl-next-gen-stats/>

O’Leary, Z. (2021). *The Essential Guide to Doing Your Research Project*, *4th Edition*. London:

Sage Publications, Ltd.

Ozanian, M., & Settimi, C. (2021). *The NFL’s Most Valuable Teams 2021: Average Team Value*

*Soars To $3.5 Billion As League Shrugs Off Pandemic Year*. Forbes.

https://www.forbes.com/sites/mikeozanian/2021/08/05/the-nfls-most-valuable-teams-

2021-average-team-value-soars-to-35-billion-as-league-shrugs-off-pandemic-

year/?sh=2edda572654e

Peiser, B., Reilly, T., Atkinson, G., Drust, B., & Waterhouse, J. (2006). Seasonal changes and

physiological responses: Their impact on activity, health, exercise and athletic

performance. International SportMed Journal, 7(1), 16–32.

Phetteplace, G. (2000). Integrating cold weather impacts on human performance into army M&S

applications. *Winter Simulation Conference Proceedings, 1*, 1020-1024.

<https://doi.org/10.1109/WSC.2000.899906>

Whittaker, Z. (2022). *Web scrapping is legal, US appeals court reaffirms*. TechCrunch.

https://techcrunch.com/2022/04/18/web-scraping-legal-court/#:~:text=Good%20news%20for%20archivists%2C%20academics,a%20U.S.%20appeals%20court%20ruling.