Assignment Overview

In this week, we have introduced the concepts of basic regression analyses and learned to use regression analyses to identify relationships among variables using National Hockey League (NHL) data.  In this assignment, you will explore a National Football League (NFL) dataset and practice using regression analyses to explore the impact of several external factors on game outcomes.

For this assignment, you will explore an NFL historical game result dataset. This dataset also contains basic information on NFL teams and the stadiums where the games took place. Additionally, this dataset contains game day weather information. You will test if there is home game advantage for the teams and if external factors such as stadium condition and weather have any impact of teams’ ability to score.

To complete the assignment you will need to perform similar-steps in the lecture.  You will be asked to answer several multiple choice questions based on your data and result. After passing each quiz, you will be provided sample code that produces the correct answer for that portion of the assignment. You should use this sample code as the basis to move onto the next part of the assignment if your code differs greatly.

**Beware:** even though your code might get you to the correct answer at a given point, it is sometimes possible that the way you write it might interfere with completing a further step. So even if you get the answer right, you should look at the code we supply to check if you are going the same way. In practice, there are often many ways to get to answer in Python, and we do not insist that you follow our approach exactly – but simply warn you to be aware that differences could turn out to be problematic later.

# Assignment Instructions - Part 1

## **I. Import and Explore Data**

1) Import the “nfl\_game.csv” data file and name the dataframe as “NFL\_Game” in Jupyter Notebook. Descriptions of selected variables - Unit of measurement of weather variables: - temperature - degree Fahrenheit; - humidity – relative percentage; - wind – mph - The variable “score” is the score earned by the team specified in the “team” variable. The variable “opponent\_score” is the score earned by the team specified in the “opponent” variable. - The variable “score\_diff” is defined as “score – opponent\_score” for the given team. - The variable “stadium\_age” is defined as the difference between the year of the season and the year the stadium opened. - The variable “stadium\_neutral” indicates if the game was played in a third stadium, which is neither the home team’s nor the away team’s own stadium.

2) Use the “describe” function to calculate summary statistics for the “date” variable. Use the “describe” function to calculate summary statistics for the “score” variable based on whether it is a home or an away game for the team.

3) Find the correlation efficients between the following pairs of variables: - “win” and “home”        - “score\_diff” and “home” - “score” and “weather\_temperature” - “score” and “weather\_humidity”        - “score” and “weather\_wind\_mph”

# Assignment Instructions- Part 2

## **II. Regression Analysis 1 – Test of Home Game Advantage**

In this regression analysis, you will try to determine if playing at home gives teams any advantage in their performance.

Run the following regressions where the variable “score\_diff” is the dependent variable.

* Reg1\_1: Include a single variable “home” as the independent variable.
* Reg1\_2: Include the “home,” “stadium\_capacity,” “stadium\_neutral” variables and the interaction between “home” and “stadium\_neutral” as the independent variables.
* Reg1\_3: Include the “home”, “stadium\_capcity,” and “stadium\_neutral” variables, the interaction between “home” and “stadium\_neutral,” as well as “team” and “opponent” as the independent variables.

# Assignment Instructions- Part 3

## **III. Regression Analysis 2 – Impact of Outside Factors on Scores**

In this regression analysis, you will investigate if the final score earned by each team is affected by outside factors such as the size and condition of the stadium as well as weather conditions at the stadium.

Run the following regressions where the variable “score” is the dependent variable.

* Reg2\_1: Include “season” and “home” variables as independent variables.
* Reg2\_2: Include “season,” “home,” “weather\_temperature,” “weather\_wind\_mph,” and “weather\_humidity” as independent variables.
* Reg2\_3: Include “season,” “home,” “weather\_temperature,” “weather\_wind\_mph,” “weather\_humidity,” “stadium\_capacity,” “stadium\_age”, “stadium\_type,” “stadium\_neutral”, and the interaction between “home” and “stadium\_neutral” as independent variables.
* Reg2\_4: Include “season,” “home,” “weather\_temperature,” “weather\_wind\_mph,” “weather\_humidity,” “stadium\_capacity,” “stadium\_age”, “stadium\_type,” “stadium\_neutral”, and the interaction between “home” and “stadium\_neutral” as independent variables. Additionally, add both “team” and “opponent” in the regression.