NFL Matchups - Data Wrangling

In order to gather the data I needed for this project, I used an NFL API https://profootballapi.com/. The API was able to generate a broad range of information and statistics so before I began I had to decide the features I wanted and formalize my approach.

First, I decided that I wanted features that could be easily controlled by coaches and organizations making the predictive model useful as a tool for insight. By using controllable features coaches could adjust their approach to what the model values in order to better their team's performance. For this reason, I stayed away from features such as team record or previous matchup outcomes and instead tried to generate features representing team composition such as a strong running game or a good pass defense.

I began by collecting the raw data from the API, which was returned in JSON format and had to be parsed and converted into pandas dataframes. Games and team statistics were simple and involved basic manipulation of the columns or grouping before they were saved to CSV files. However, individual player data was stored beneath multiple layers of dictionary keys and needed a team to be specified for each request. I created a function to parse the players data for each team and aggregate them in the relevant dataframe passing, receiving, rushing, or defense where it could then be saved and manipulated later. Next I had to merge the player data with the game data to correlate the year with the player statistics before perform various grouping, merging, and column manipulations to get the fields I desired. Once the game outcomes, team, and player data were collected the datasets were merged together to form a final dataset. Each row represents a game between 2009 and 2017 where the scores and yearly statistics of the home and away teams are the columns. The model will use the yearly team data for each team to try and predict the outcome of a given matchup.

After some validation queries I could see that the API was missing player statistics for 5 teams since 2009. Games which involved one of these teams were excluded, however, this does not seem to be an issue as there are still over 2,000 matchups for training the model. Because the data had been collected in a custom way no extra cleaning steps were needed.