Yash Nakadi

Dr. Hitt

OIDD 315

May 8, 2022

OIDD 315 Project

**Description of attributes (many are self-explanatory):**

Play\_id: primary key; unique for every instance in the table

Game\_id: unique identifier for every game; displays in order the away team, home team, year

Week: what week the game takes place in

Desc: a description of the play

Posteam/Defteam: the team on offense/defense

Game\_date: as a date variable, provides the day, month and year the game took place in

Game\_seconds\_remaining: how many seconds are left in the game

Passer\_player\_me: name of the quarterback involved in the play (if applicable)

receiver\_player\_me: name of the wide receiver involved in the play (if applicable)

rusher\_player\_me: name of the running back involved in the play (if applicable)

wpa: win-probability added; shows how much a team’s chance of winning the game changed as a results of the play

EPA: expected points added; advanced stat that determines how well the play was relative to the “expected points” of that play. This stat is very important for this project. Better definition [here](https://www.the33rdteam.com/epa-explained/)

Posteam\_type: indicates ‘home’ if the team on offense is the home team, or ‘away’ if the team on offense is the visiting team

Score\_differential: at the time of the play, how many points the HOME TEAM is ahead by (this value can be negative)

Yards\_gained: how many yards were gained as a result of this play

Home\_score/away\_score: how many points the home team/away team has at the end of the game

Result: (Home\_score) – (Away\_score)

Play\_type: play that was ran; main types are ‘pass’, ’run’, ’field\_goal’, ’extra\_point’, ’punt’

Down: is either 1, 2, 3, 4; represents what down the play took place during

**Guiding Question and question set**

**Guiding question:**

Using NFL play-by-play data, analyze offensive tendencies, explore which offenses are the most successful and why, along with what defenses focus on in order to be successful, with the overall hope of determining offensive and defensive strategies and goals that lead to a high winning percentage.

Part 0: Setup and adding useful columns

0.1) Currently there is one dataset featuring seasons 2001-2010, 2012, 2014, 2016, 2018-2021 combined, and then there are 4 separate datasets for the play-by-play data for 2011, 2013, 2015, 2017. Assuming that all have the same columns (identified above), merge the datasets together under a new table (using UNION will make things easy).

0.2) With this table, add 2 new columns: season and winner. Be careful for season; it is not just the year the game takes place in because you need to consider late regular season games and playoff games that are in January (month(game\_date) = 1) and February (month(game\_date) = 2). For winner, this should determine whether the team ON OFFENSE won the game (can determine this by looking at result).

Part 1: basic queries, getting used to the data

1.1) How many pass and run plays are in the dataset?

1.2) What are some of the largest plays by wpa since 2010 (do not include field goals)?

1.3) Who were some of the best qbs, wide receivers and running backs, in terms of average EPA. Make sure to specify the season

Part 2: DE-FENSE

2.1) Which defenses forced the most turnovers (look for interceptions and fumbles, though consider that not all fumbles are recovered by the defense)? Rank every defense by this metric.

2.2) Which defense forced the most negative plays (no yards gained)? Rank each defense.

2.3) Which defenses allowed the fewest points per game? Rank each defense.

2.4) Determine which defenses allowed the fewest yards per play. Again, rank the defense.

2.5) For the moment, let’s assume all these indicators are weighted equally. Using a ranking system, determine the best defenses (team that averages the highest rank).

2.6) Looks like scoring lines up nicely with the overall best defenses. That begs the question, is there a strong relationship between scoring defense and winning? Determine the winning percentage for every defense, compare to the corresponding points allowed by graphing the points on a scatter plot in Excel, adding the trendline and showing the equation and R^2.

2.7) On second thought, we can find this information right here on SQL. Write a function that takes in a table with two columns of numbers and outputs the R^2 value (note: you have to make a unique table datatype for SQL to accept a table parameter; this part is not easy).

Part 3: Offense Time

3.1) Find the NFL teams that were top 10 in points per drive. Are any of them notable? Now do this for the worst teams.

3.2) Determine the winning percentages for all NFL offenses. From there, determine how often they ran the ball and how often they passed it.

3.3) A key idea in Moneyball was that you needed a certain number of runs to get into the playoffs (have a great winning percentage). Can we do something similar with football? How many successful plays are needed to reach a certain winning percentage? We will define a successful play as a play that has more than twice the epa of the average play, specifically for both passes and runs. Is this a good way to think about teams, or should this thinking only stick with baseball?

3.4) Teams that win love passing, so league-wide have teams begun passing it more and more? Write a CTE and determine, on a year-by-year basis, how much the rate of passing on first down has changed over time.

3.5) A big topic in football is how often teams should go for it on 4th down. Do something similar to 3.4 and determine how the 4th down attempt rate has changed season-by-season

3.6) Typically the first 15 or so plays (almost all in the first quarter) are scripted, while plays in the fourth quarter come solely from adjustments based on how the game has gone. Is there a significant difference in play calls, success, yards gained or lost, for great offenses that is?

**Analysis:**

Much of my analysis for this project (mainly how I answered the questions above and tried to solve the overall problem) was through constant comparison of various data points. For instance, when evaluating NFL defenses, I compared them in terms of points, yards, turnovers forced and bad plays forced to see which measurement was most useful. With the offense, the key comparison was with passing plays vs running plays, seeing how valuable each one was and how often teams utilized them. In addition, EPA was crucial for me, as I used that statistic to determine what made a play successful and associate a high EPA with a successful offense or offensive player. Perhaps in the future I will find other ways to holistically gauge how good an offense is. As one last point, since the overall goal was to find strategies and ideas that directly lead to winning football games, I experimented with linear regression and R^2 values. While it did not work out the way I would have hoped in problem 2.6, I can use this function to compare other variables and see if I can find something. Below are some of the graphs and queries I was able to make as part of this project.

**Key Findings and Conclusions:**

* For defense, allowing the fewest points is the most important trait, over field position, time of possession, etc.
* That being said, the R^2 value is not that high, meaning that we cannot say that the link between allowing fewer points and winning games is very strong
* Teams with the highest winning percentage, along with the most efficient offenses, tend to pass the ball more often than they throw it
* Passing plays have far more value (epa per play) than running plays
* As such it might be easier to have more successful running plays than passing plays, but it is true that for the same number of successful plays, run plays yield a higher win percentage than pass percentage, perhaps a fallacy of some kind
* Despite this emphasis on passing the ball, the yearly change in passing rate, particularly on first down, is not what we might expect, since it is not strictly positive.

**Information about the Data and Data Preprocessing:**

The zipped file contains the dataset I used for this project, though the season and winner columns are included if you want to remove those. As mentioned earlier, the source of this data comes from [this Github repository](https://github.com/nflverse/nflfastR), which has play-by-play data for each NFL season dating back to 1999. This database keeps track of every feature one can think of, from basic information like the down and distance, what type of play was called, time on the clock, if a touchdown occurred, etc. to more advanced statistics like EPA and WPA. I only used a few of these attributes, and those are the ones mentioned at the top. In terms of preprocessing, as mentioned already, I did not run a script because each season was its own data table, so much of the work cleaning the data and combining across seasons before using the import wizard had to be done outside of SQL. Nonetheless I was able to show on the main SQL file how I merged the data from the 2011, 2013, 2015 and 2017 seasons into the main dataset with the other years.

**Appendix (Some Useful Results as part of Analysis):**

**From Problem 2.6:**

**From Problem 3.2**

**Table

Description automatically generated**

**From Problem 3.6**

**A picture containing application

Description automatically generated**