Project #2 - Data Exploration and Design

Jacob M. Lundeen

2021/02/23

Purpose

The purpose of this assignment is to introduce you to the process of exploring and visually analyzing data without even having to develop a visualization tool. You will pick a domain and data set that you are interested in. The data should have at least 10 variables (i.e. columns) and 1,000 records (i.e. rows). The purpose of this assignment is to design different visualizations to illustrate individual aspects of the data set under consideration.

Data

For my project, I will be using the data set created by Mr. Ben Baldwin called 'nflfastR' (www.nflfastr.com). It is a database containing NFL play-by-play data back to 1999. It even includes more advanced metrics, like Completion Probability (CP), Completion Percentage over Expected (CPOE) and others. The data set contains 361 variables and, depending on how many season your are looking at, millions of observations. For this project, I will be using a varying range of years, but only going as fas back as 2015 (which is well over 100,000 observations). Using the skimr package, I will show a summary of the first ten variables (I chose not to display all 361 variables for obvious reasons).

Table 1: Data summary

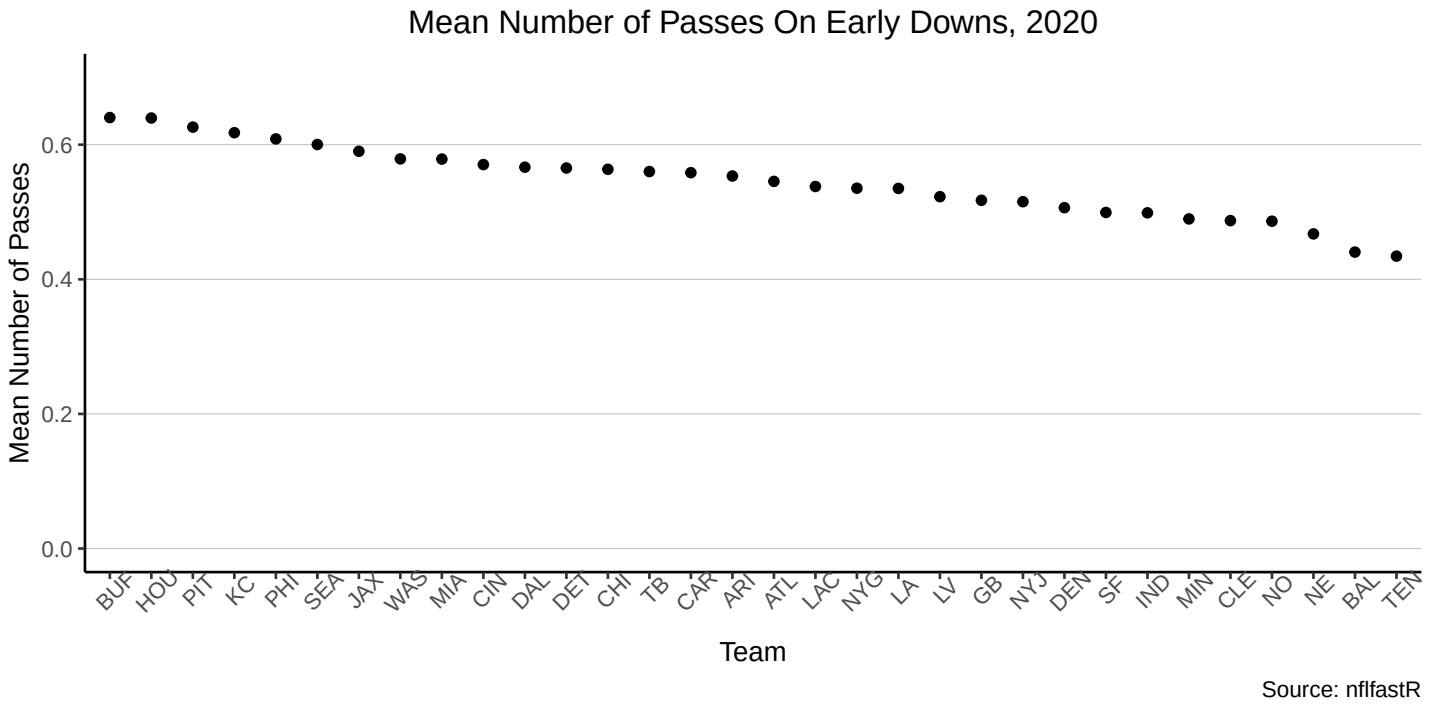
Name	df
Number of rows	48514
Number of columns	363
Column type frequency:	
character	10
Group variables	None

Variable type: character

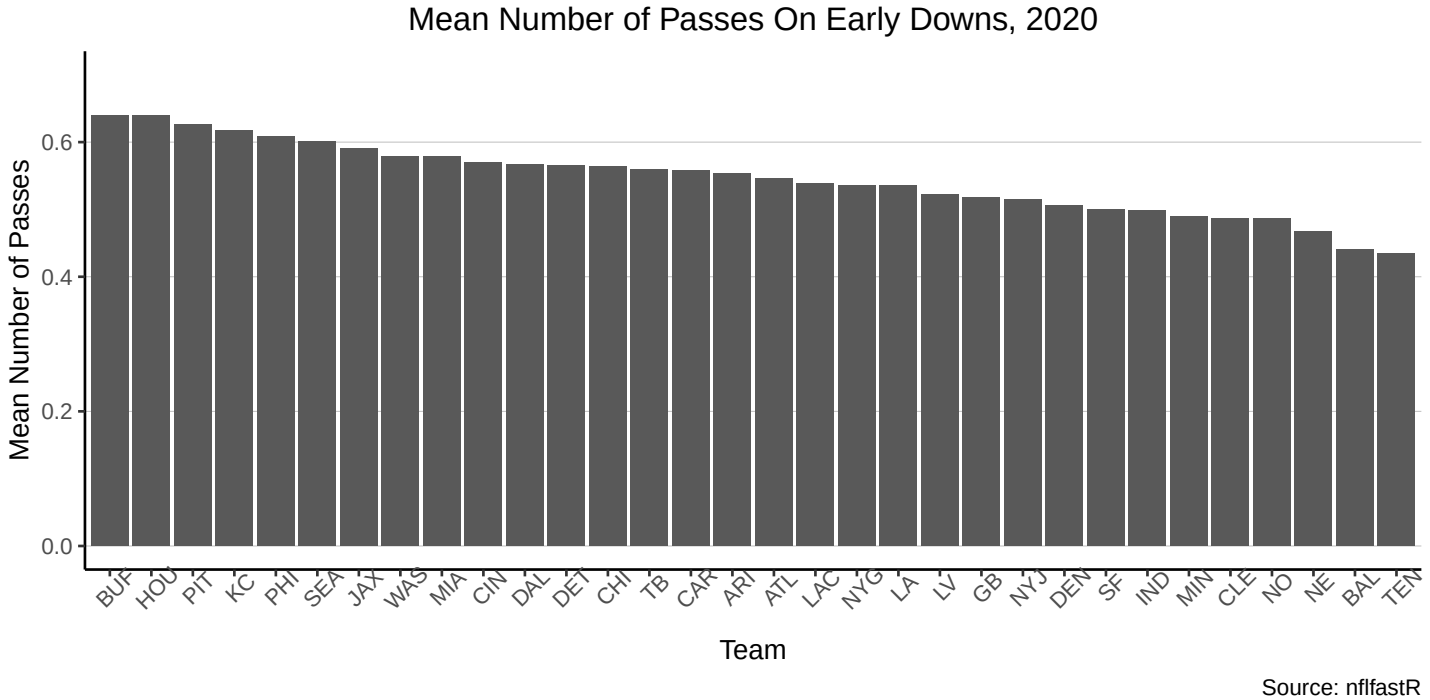
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
game_id	0	1.00	13	15	0	269	0
old_game_id	0	1.00	10	10	0	269	0
home_team	0	1.00	2	3	0	32	0
away_team	0	1.00	2	3	0	32	0
season_type	0	1.00	3	4	0	2	0
posteam	3336	0.93	2	3	0	32	0
posteam_type	3336	0.93	4	4	0	2	0
defteam	3336	0.93	2	3	0	32	0
side_of_field	4150	0.91	2	3	0	33	0
game_date	0	1.00	10	10	0	58	0

Questions

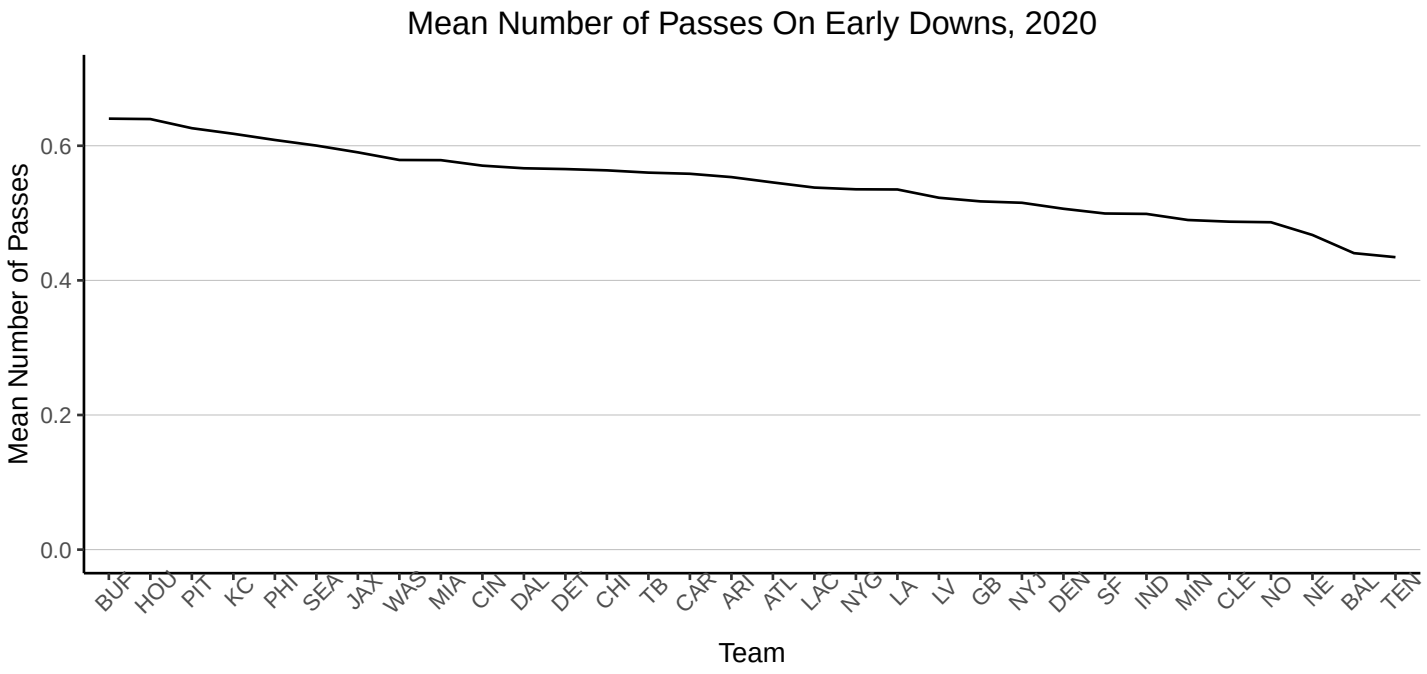
#1. Which teams were the most pass heavy on early downs excluding the last two minutes of the half/game?



The chart above shows how aggressive NFL teams are on early downs (1st and 2nd down) with passing the ball. The Y-axis is the mean number of passes thrown on early downs, with the teams on the X-axis (which are ordered). It is a simple point plot with no coloring as I felt that adding any sort of color would not provide any additional context.



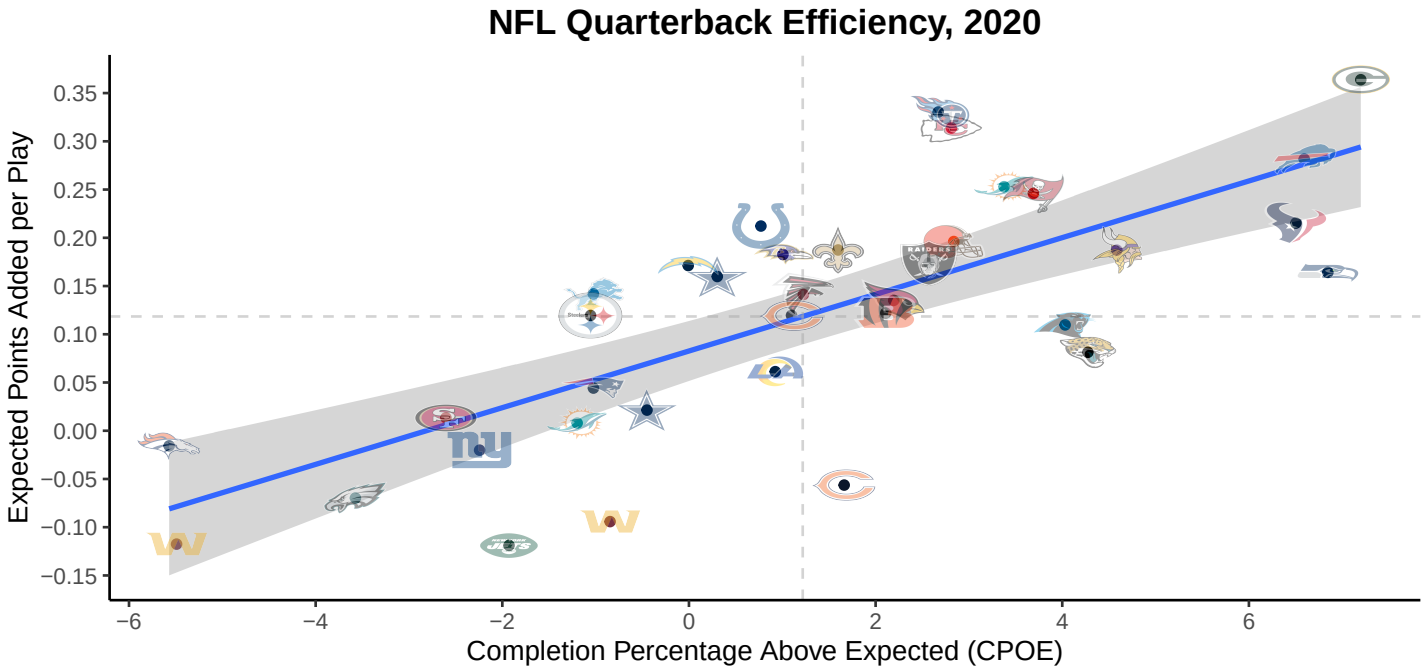
Here we have the same data, but presented as bar chart. It tells the same story as the previous point plot, and I still feel that adding color isn't necessary, but because there are 32 teams the chart is pretty crowded so a bar chart isn't the best option here.



Source: nflfastR

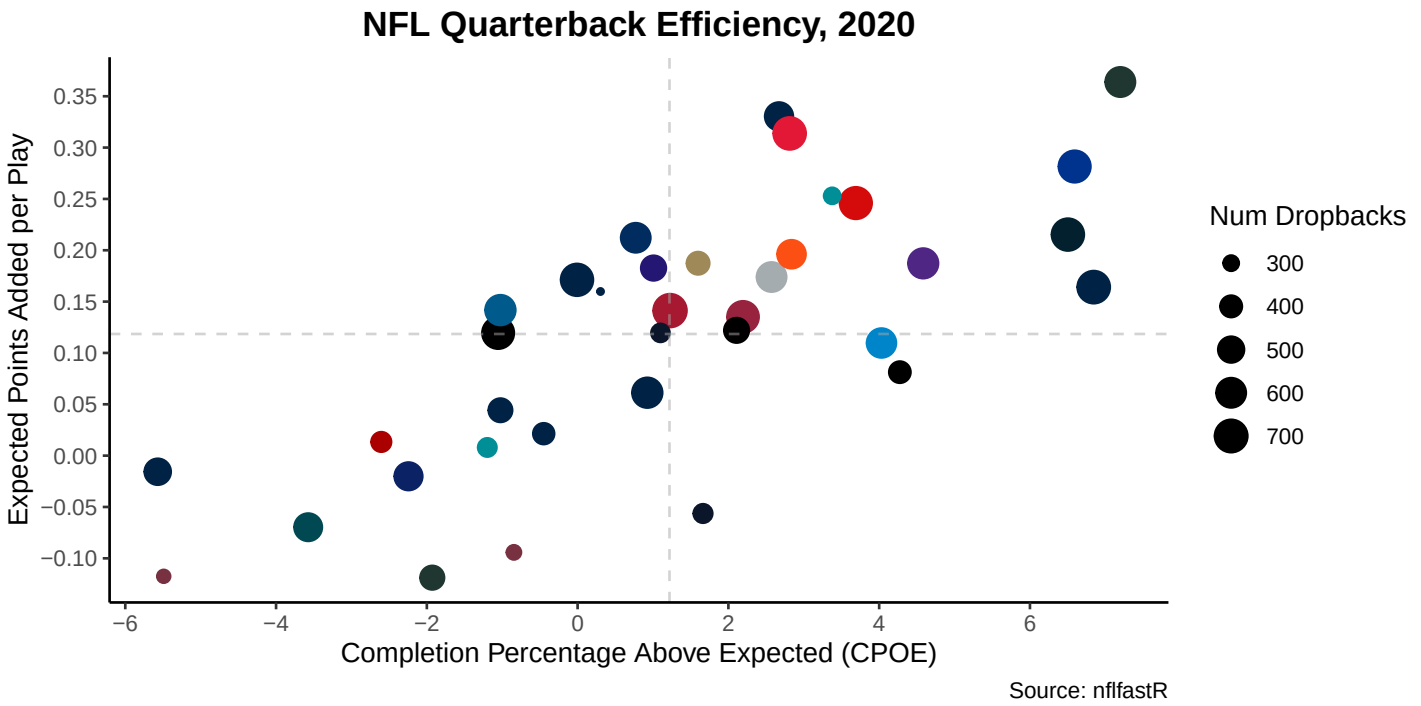
My third chart here is simply a line chart of the data. Again, no color included and the line shows the downward slop of aggressiveness. I like this better then the bar chart, but I think it is difficult to get a feel where each individual team is without the points to act as a reference.

#2. How does a Quarterback's CPOE compare to Expected Points Added (EPA)?

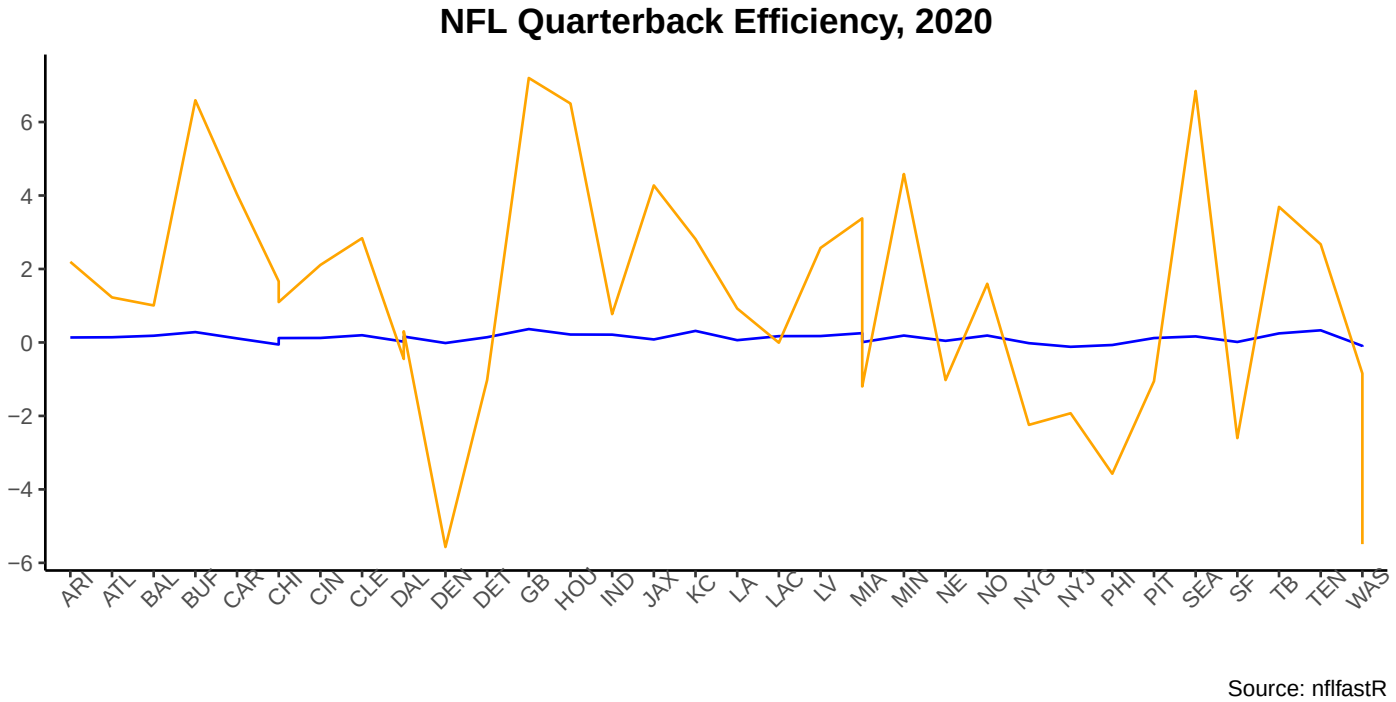


Source: nflfastR

This chart is a little on the fancy side, but for the general audience I figured they would like it. What it is showing is the comparison of quarterback's EPA versus CPOE. These are some advanced metrics for quarterbacks that not everyone may be familiar with, but one can research them easily. What I have done is taken a basic scatter-plot, added a regression line, mean lines for each axis and included the team logos as the points. The logos can make things a little muddy, but for the average fan it would be appreciated.

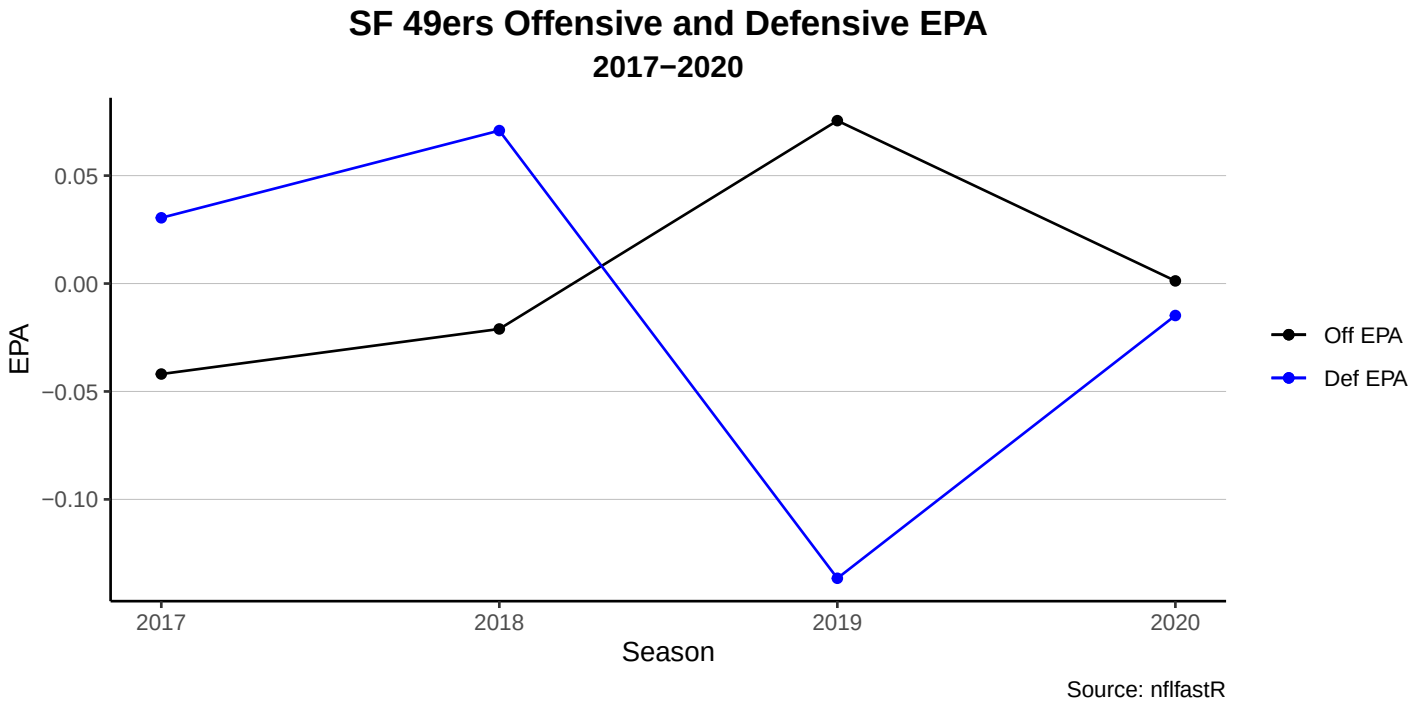


Here I've taken the EPA vs. CPOE data and converted it to a bubble plot. The bubble size is based on the number of dropbacks each quarterback had over the course of the season. The actual points are colored by team color. Using team color instead of logo makes the chart much cleaner, but the colors are not well defined / do not differentiate greatly enough to accurately tell which dot is which team.



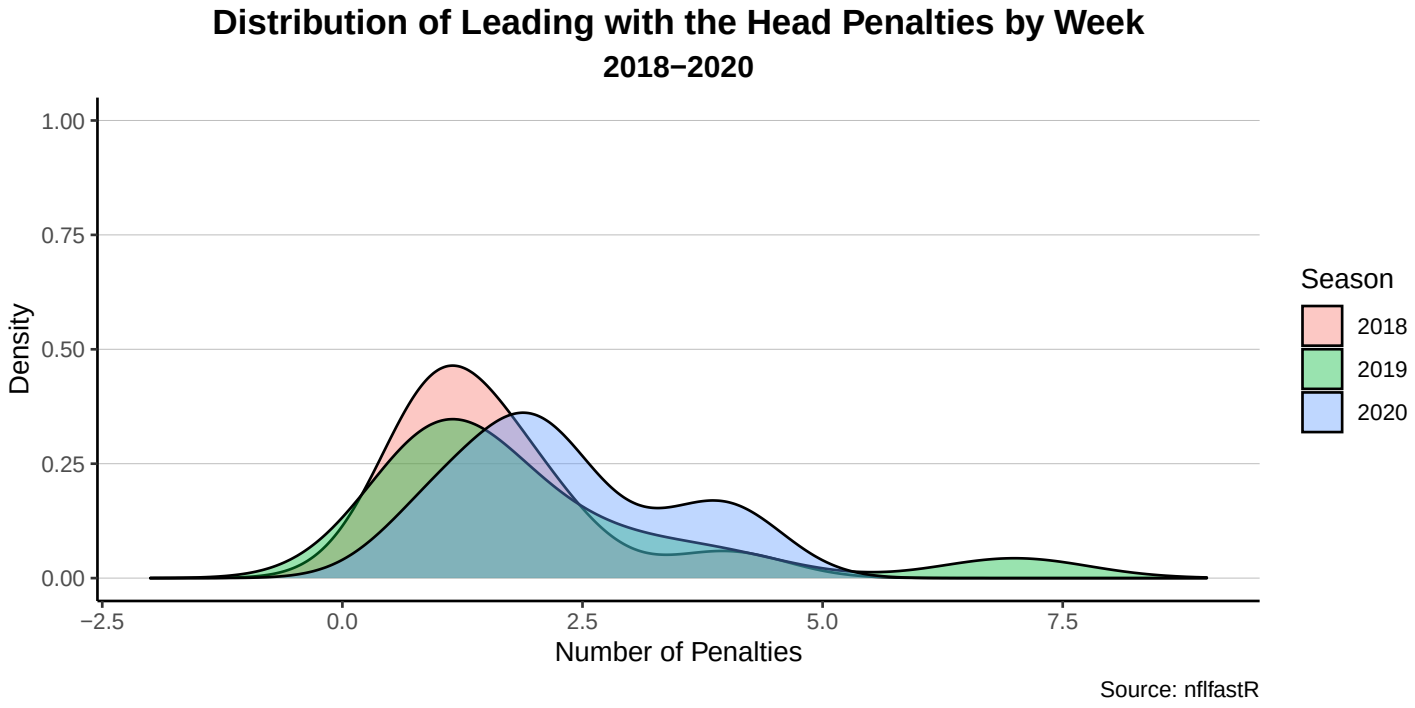
Lastly we have a stacked line chart showing EPA and CPOE by team (couldn't get the legend to work, blue is EPA and orange is CPOE). This presents the data in an interesting light, EPA is fairly constant across the league, whereas CPOE is highly variable. Seeing the data presented like this could lead to further investigation as to why. With having two lines of data, I needed to color them to differentiate them for the reader. I could include some grid lines, and even the points, if desired to help the reader.

#3 What Does the San Francisco 49ers Offensive and Defensive EPA look like under Head Coach Kyle Shanahan?



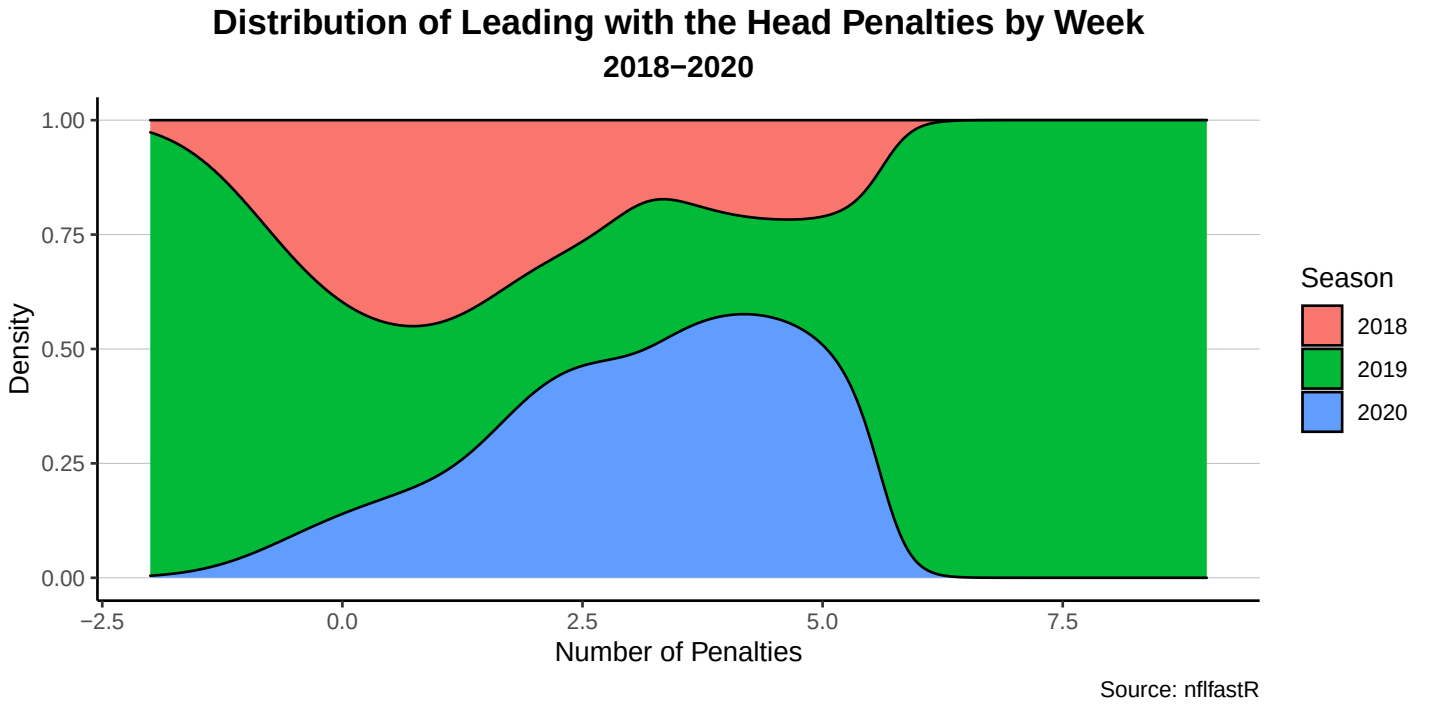
This chart shows a line comparison of the San Francisco 49ers offensive and defensive EPA since Head Coach Kyle Shanahan was hired. For defense, negative EPA is better because it indicates that it is stopping the opposing offense. The immediate thing that jumps out is just how good the team was 2019, especially the defense (0.08 offensive EPA and -0.14 defensive EPA for the season!). The plot is simple and clean, with one of the lines colored to differentiate offense from defense.

#4 What is the distribution of "lowering the head penalties" since 2018?

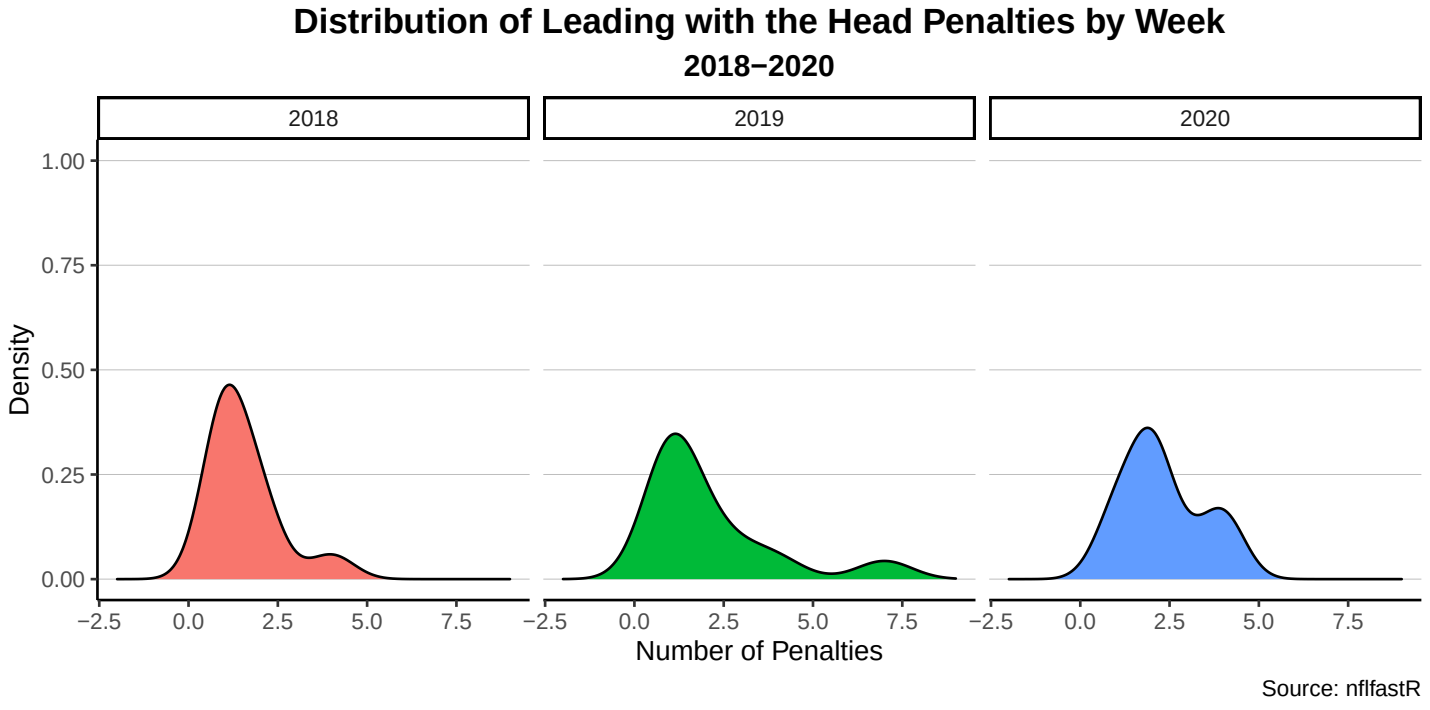


In 2018, the NFL introduced a new rule that made it illegal for any player (offense or defense) to lead with the head/helmet to hit another player (essentially the targeting rule from college football). The rule has been controversial as it can lead to an ejection from the game. What I am showing in the chart above is the number of these penalties called per week over the

last three seasons. This density chart does a nice job of indicating just how often this penalty has been called each year. I adjusted the alpha so the reader can see the separate plots.

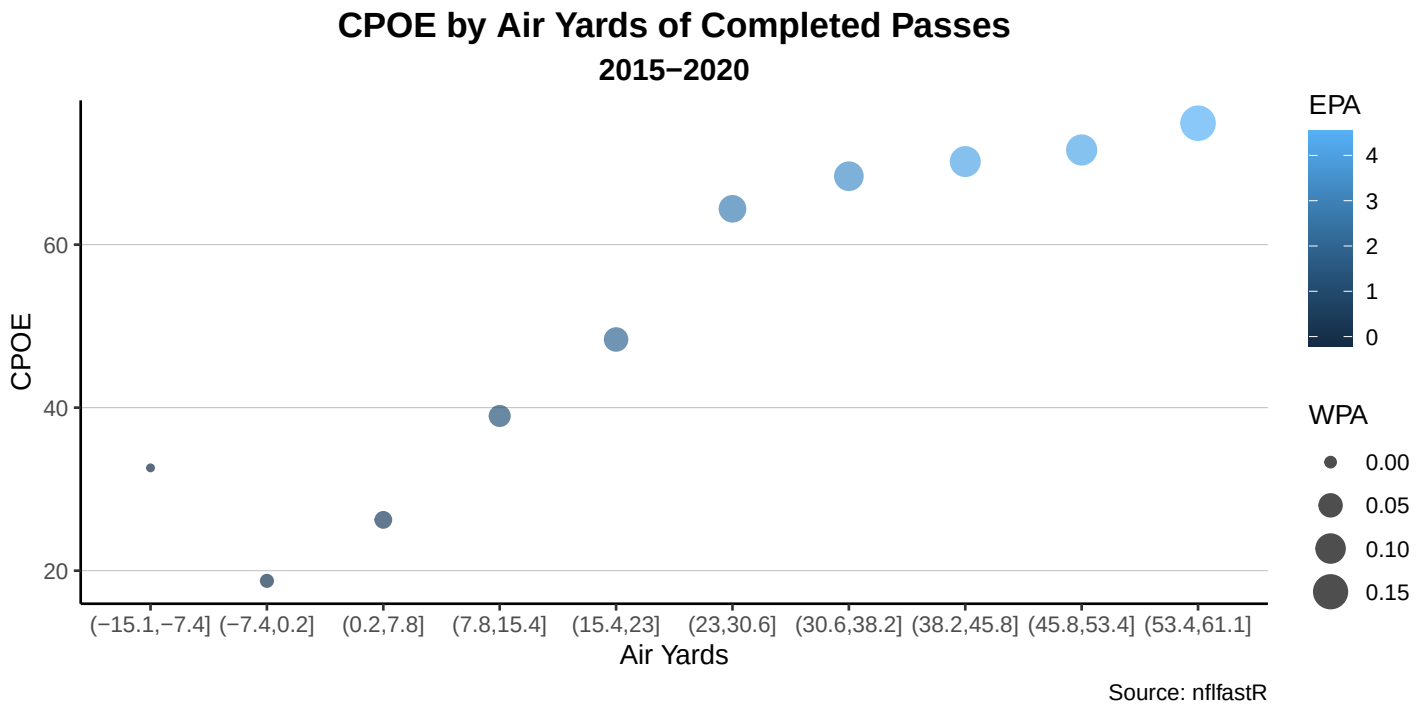


Here we have the same data but in a stacked format. It's pretty, but confusing. This makes it look like the penalty was called with reckless abandon in 2019.



Lastly, we show the same data by utilizing the principle of small multiples. I like this one better then the stacked chart, but I am not sure if I like it better than the original chart. While this makes it easier to see each year, you lose the clarity on how they compare to each other since they are not stacked on top of each other.

#5 How does CPOE and air yards of a completed pass affect Win Probability Added (WPA)?



This last chart is a take on the earlier chart with CPOE and EPA. This is a bubble chart that shows CPOE versus air yards with WPA as the filler and EPA as the color. It is a lot for a single chart, and not something I would normally go with.