Name: Sreeharsha Sistla
Github Repo: https://github.com/sistlasre/football_analysis

Breakdown of NFL Plays Over a 10-Year Span

Introduction:

I have been a huge fan of the NFL since I've been a little child. I watch the games religiously, and more recently, I have been following the NFL analysts in how they describe the changes in the NFL. The following analysis explores some of the trends that have been emphasized over the past 10 years and determines whether what people have been saying is actually the case. The input to the notebook is a CSV containing all the plays in the NFL from 2009-2018. The output I have produced is a set of graphs that explore the different trends in the NFL , and I will add my own interpretation on why the trends are as they are. Some of the visuals produced through the Jupyter Notebook could not be included in this specific report as they were created in the Matplotlib interactive environment within Jupyter.

Dataset:

The dataset that I wanted to explore was a full play-by-play breakdown over the years. I was able to pull a fairly comprehensive dataset from this Kaggle link. The dataset itself contains a fairly good breakdown of each play, including things like play type, play location, yards gained, and most importantly, the play description. Some of the columns are very generalized, but with the play description, I was able to get more granular values for some of the columns. In order to push the specific data I needed for my analysis into the Github repo, I first had to explore the different columns available. Understanding the columns in the dataset was not included in the Jupyter Notebook as the notebook contains the final version of my code. However, after determining what columns I wanted to use, I split the larger dataset, which was ~700 MB, into 10 files, split by season. The code for the splitting is included in the notebook, but it is currently marked as a Raw cell so it won't execute.

Literature Review:

In order to determine what other research has been done in this area, I took a look at the NFL's Big Data Bowl. Many of the projects included in this competition explore formations, efficiency, player physique, and delve deep into machine learning in an order to predict what teams should do in different scenarios or what kind of players make teams most effective. While these analyses are extremely interesting, this analysis won't delve too much into the machine learning side of things. It will instead focus on what I, as a lifelong obsessive football fan, believe has changed in the NFL over the years. And based on the graphs produced, I will attempt to either debunk or confirm whether what I have seen and heard holds true based on this analysis.

Quality of Cleaning:

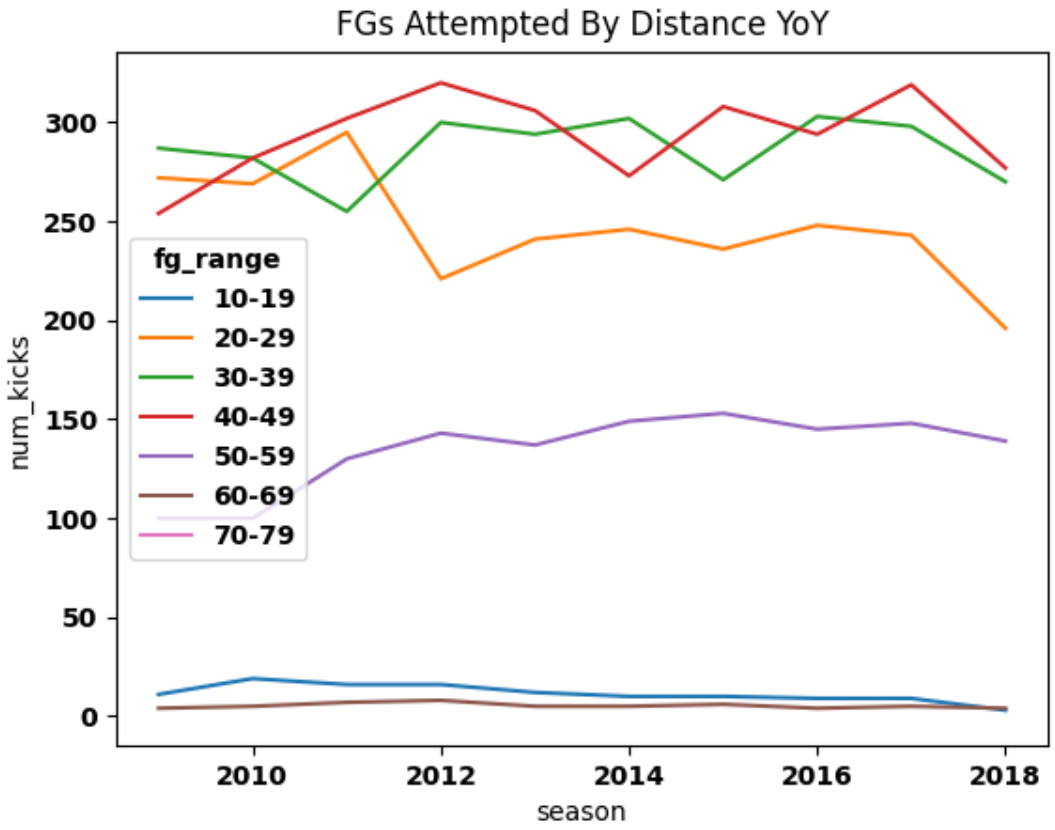
My analysis was mostly focused on plays of the following type: pass, run, field goal, or extra point. I initially broke the bigger dataset down into a subset of columns and filtered it to just plays that were of these 4 types. The initial data contained a game date column. However, it was a string. I wanted to create field, season, based on the game date so I could group the rows by season. In order to do this, I had to convert the game date column into a Pandas DateTime object. The most interesting aspect of my data cleaning/processing was the "description" column. I wanted to get a more granular value for the passing location which required some inspection of the field for different types of rows. While the notebook only includes the final regex I produced, I iterated over several options till I figured out the one that was most optimal, based on the percentage of rows that produced what I deemed a correct location.

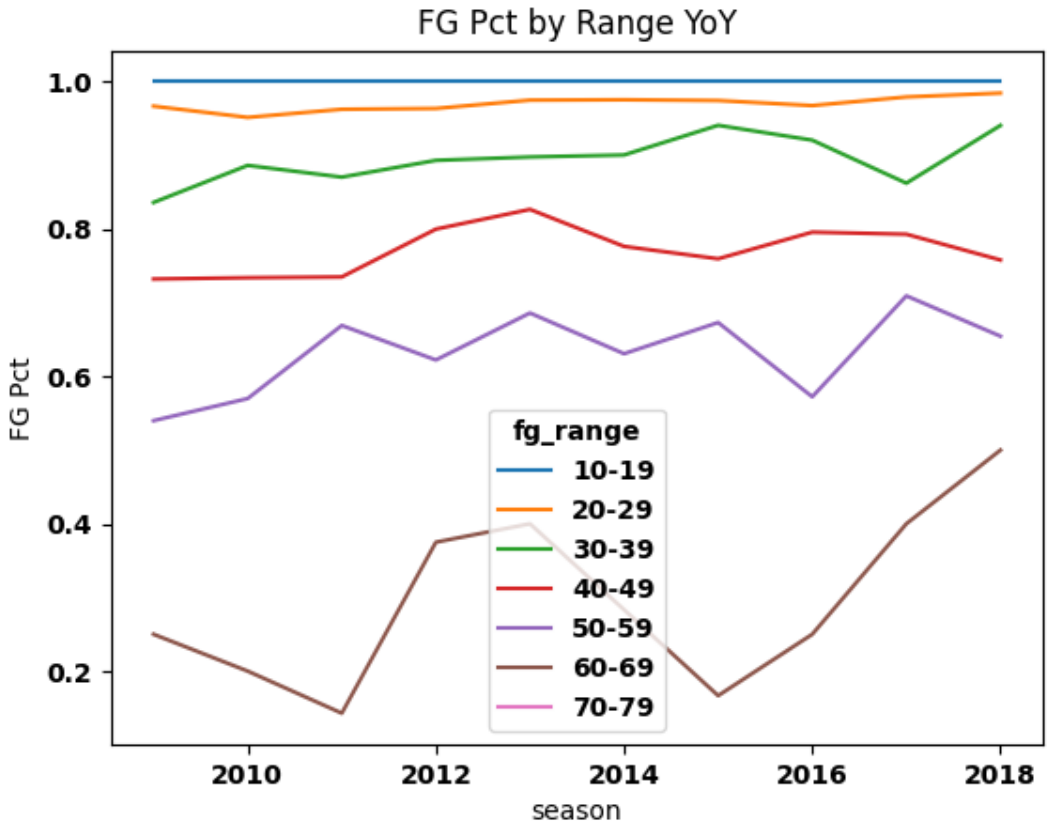
Even after applying this regex, there were still a few cases where I had missing values, or values which just didn't make sense. I took a look at these rows, and more often than not, the description didn't provide the extra granularity I wanted, so I simply had to convert these to be NaN. I did something similar for when I was parsing the extra point and field goal data. Extra points and field goals can be categorized as "Made", "Missed", "Blocked", and "Aborted". Since I was looking at the efficiency of the kickers, the "Aborted" value doesn't make sense since the kicker never had a chance to even kick the ball in this case. I also combined "Missed" and "Blocked" into the same value so the only thing to distinguish the rows was whether a kick was made or missed.

One of the major difficulties I had during this analysis was my attempt to produce an interactive set of graphs that could show trends over the years. I struggled quite a lot to determine how to use the interactive mode provided by the Matplotlib library. And even when I was able to enable this mode, I took quite a long time to figure out how to programmatically disable it throughout the rest of the notebook. However, I did feel that the interactive mode provided the benefit I wanted: I was able to see changes in the stats produced over the years very clearly, even when these changes were very subtle.

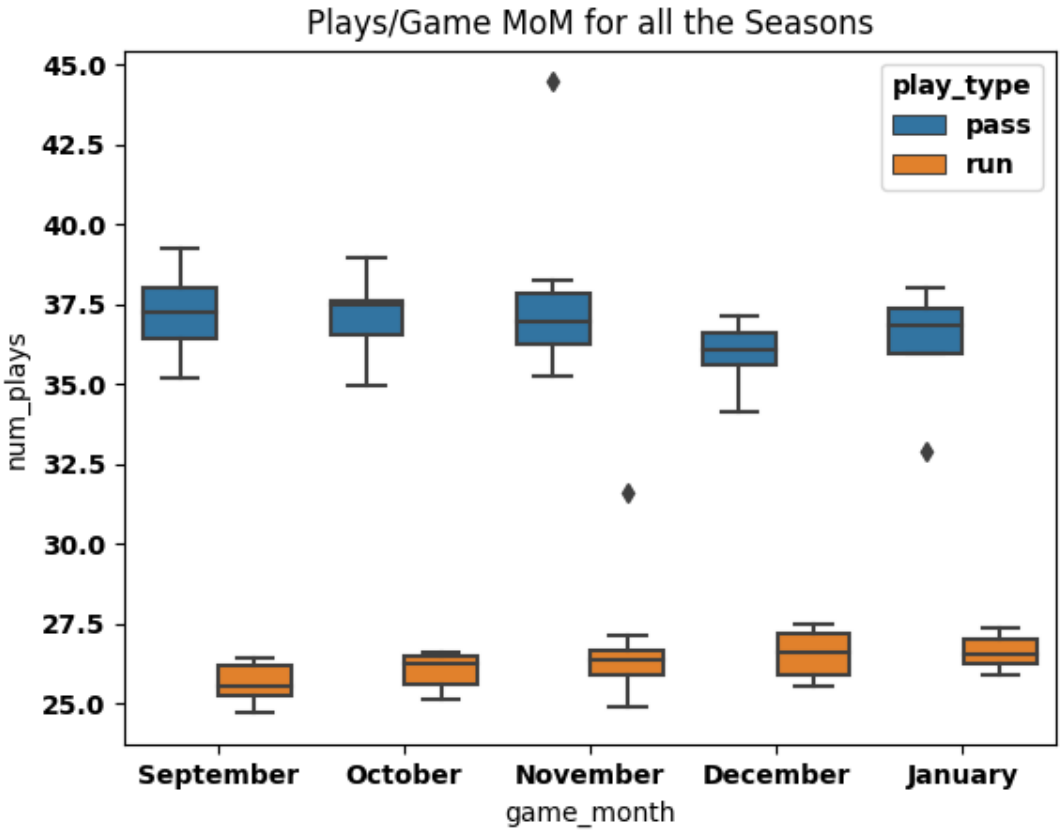
Data Visualizations:

The corresponding Jupyter Notebook produced several different graphs. As mentioned above, some of visuals in the notebook are interactive so there wasn't a good way to save those specific cells. But the following images produced help provide more insight into how the NFL has changed over the 10 year span.

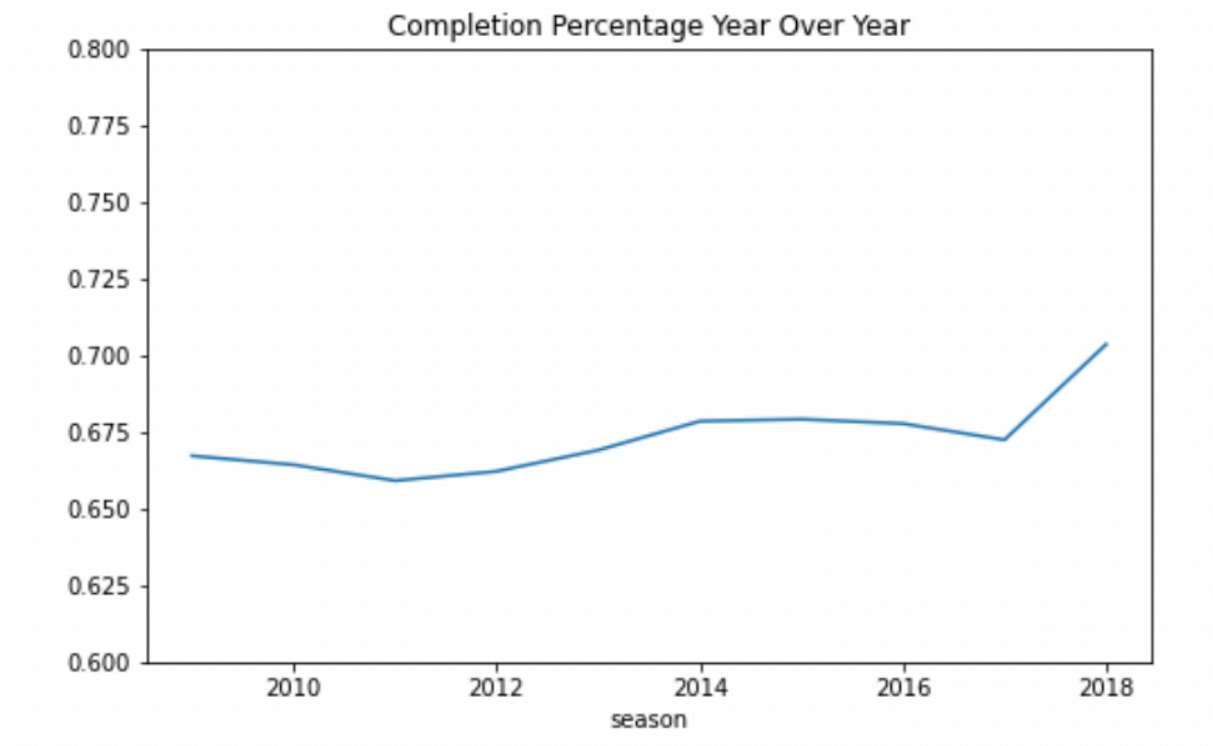




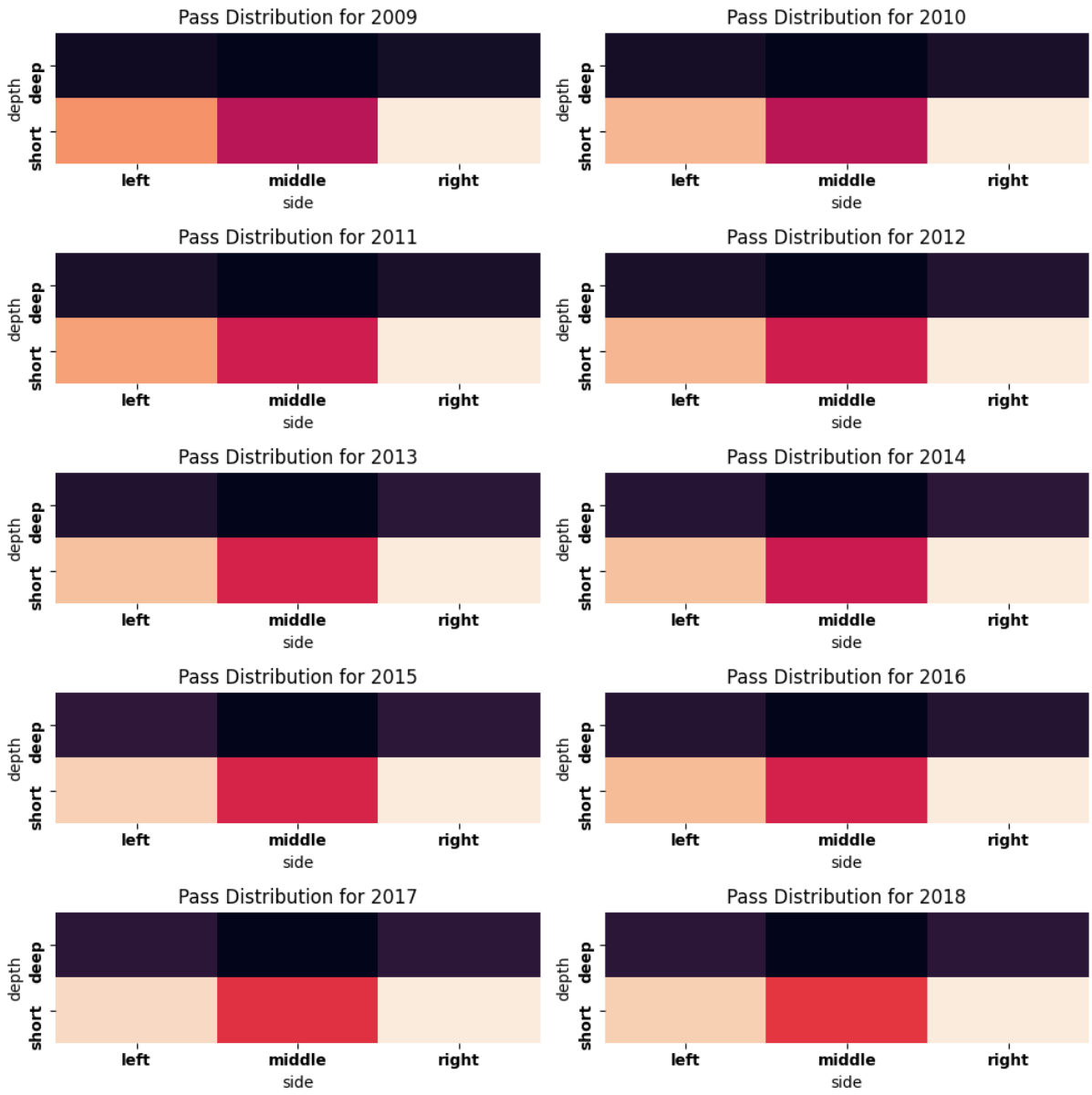
One of the common complaints I've heard (and also made myself) is that field goal (FG) kicking has deteriorated quite a bit over the past 10 years. Many fans, including myself, felt that FGs were fairly automatic back in the day, and more recently, kickers have been struggling quite a lot. The above graphs seem to debunk this complaint to some extent. The graph on the right shows the FG percentages at different ranges. Looking at it, the percentages have stayed fairly consistent. More so, it looks like these percentages have even improved a bit over the years. The one noticeable trend that we see above is the big drop of FGs attempted from 20-29 yards after 2011, and another drop after 2017. This aligns with what I have seen in the NFL. More teams are starting to be aggressive, especially when they get close to the end zone. One wonders why fans, like myself, have been extremely critical of the FG kickers these days. It's possible the kicks came in high pressure situations, which is why fans noticed it more. However, the dataset I worked with did not have the appropriate data to classify kicks as high pressure or not.



One of the other trends that I wanted to check was whether or not teams pass more than they used. Analysts and fans keep saying that the NFL is a passing league these days. So I decided to take a look at the distribution of runs and passes per game for each month over the seasons. As we can see above, teams tend to run a bit more as the season progresses. Pass plays tend to drop a bit in December. But from the overall perspective, there isn't a drastic increase in the number of pass plays over the years, which is surprising because many people still have the belief that teams only pass. So why is the league labeled as a "passing league" these days if the distribution of pass and run plays hasn't drastically changed? I decided to take a look at another metric that may provide us more insight into this notion.



Coming back to the question why people always say the NFL has turned into a passing league, I decided to take a look the completion percentage over the years. As seen above, there seems to be a small increase every year initially until we see the bigger jump in 2017. Why is there such a noticeable increase in completion percentage starting in 2017? Many people think it's because the QBs have become dual threats meaning that defenses have fewer people to stop the pass plays when the QB has room to run. Another reason is that the rules have changed so defenses are not allowed to be aggressive as they used to. So even though the play distribution hasn't changed much over the years, the notion that the NFL is a passing league does have some truth to it based on the graph of completion percentage above.



Another visual that I thought would be interesting to see was the change in passing distribution over the years. The corresponding notebook has a cell where we can see this in an interactive fashion. So while it's tough to discern differences in the visual above, we can see the subtle changes when we run these in interactive mode. For the most part, we see fairly similar trends throughout the years. As someone who also QBs a lot in a flag football league, I find that these distributions do reflect my own throwing pattern, as a right-handed QB, which is what nearly all QBs in the NFL are. While my own anecdotes don't really provide much of a basis for a valid analysis, I added this visual more for my personal curiosity to see how NFL QBs throw when compared to me.

Conclusion:

The graphs that the Jupyter Notebook produced seems to say that several of the assumptions I made about changes in the NFL aren't necessarily true. I thought FG kicking has become worse over the past 10 years, but the graphs provided by this analysis suggest otherwise. If anything, it suggests that the kicking has actually gotten better. I thought teams tend to pass much more than they used to in the past. My assumption was worded incorrectly prior to this analysis. Teams don't tend to pass more than they used to. Rather, teams are able to be more effective when passing than they used to. Overall, I learned quite a lot about changes in the NFL. In the future, I would love to delve more into this with a more machine learning and statistical analysis approach.