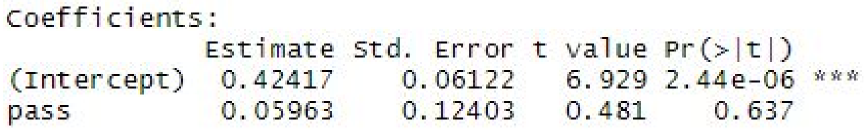
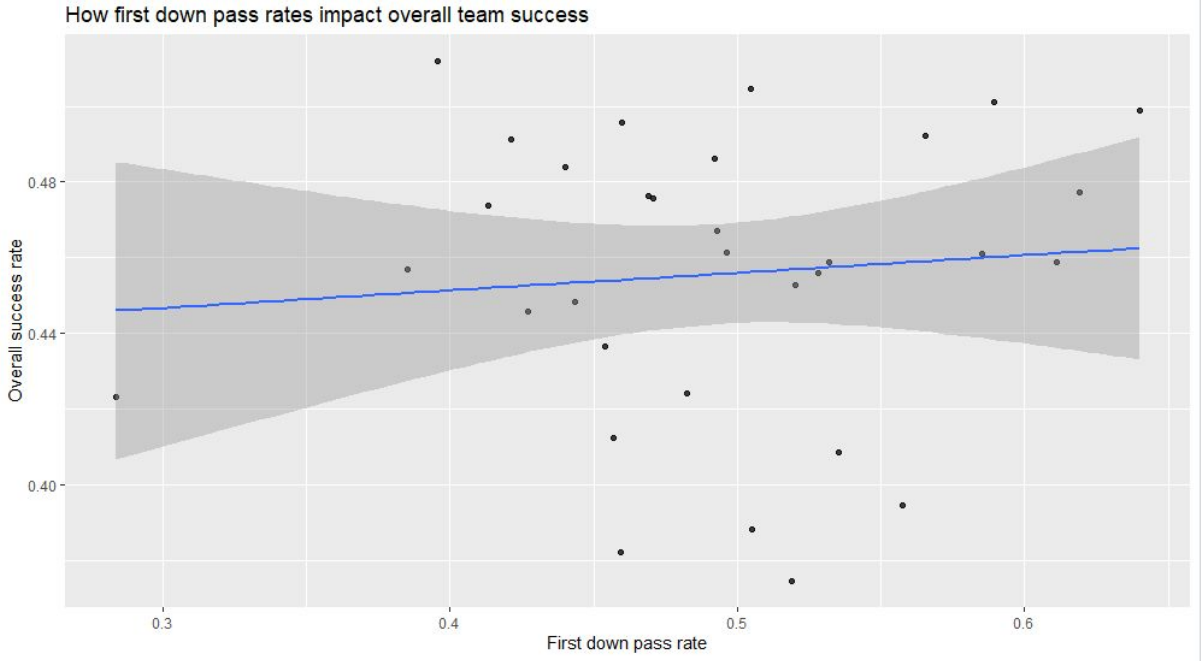
Nathan Marzion
Data Exploration/Visualization
5 November 2020

I am using the NFL play-by-play data set for 2020 to look at different variables that create successful plays. This is something that teams could use to examine the best strategies, patterns, and decisions in order to be as successful as possible. To clean my data set, I first filtered it down to only run and pass plays and plays that do not have penalties, which eliminated a lot of unnecessary rows. I then had a smaller (but still rather large overall) data set that I could condense down and create smaller data frames from when needed.

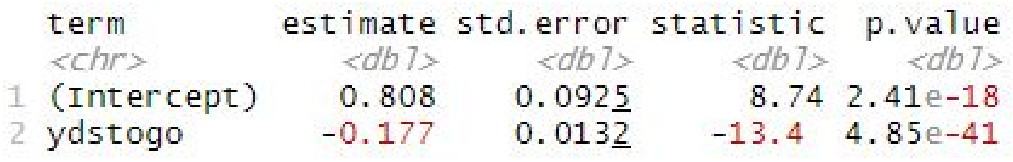
One of the first things I wanted to examine was 1st down play calling. I first had to create a data frame for all 32 teams with each team's first down play call rate (percentage of plays they run or pass) and overall success rate. This could help me look at how running/passing more on first down might cause changes in success rates for teams. When I created a linear regression model for team success rate vs team pass rate, the results showed that 1st down pass rate had a p-value of 0.637 and that it was not a statistically significant regressor for success rate. This seems to imply that in general, running or passing on 1st down is not a clear better option, but it depends more on the team and players you have. A team with a great running back and offensive line will likely succeed more from running on 1st down, whereas a team with a great quarterback and receivers would probably benefit more from passing. But in general, there is no clear "better" strategy.



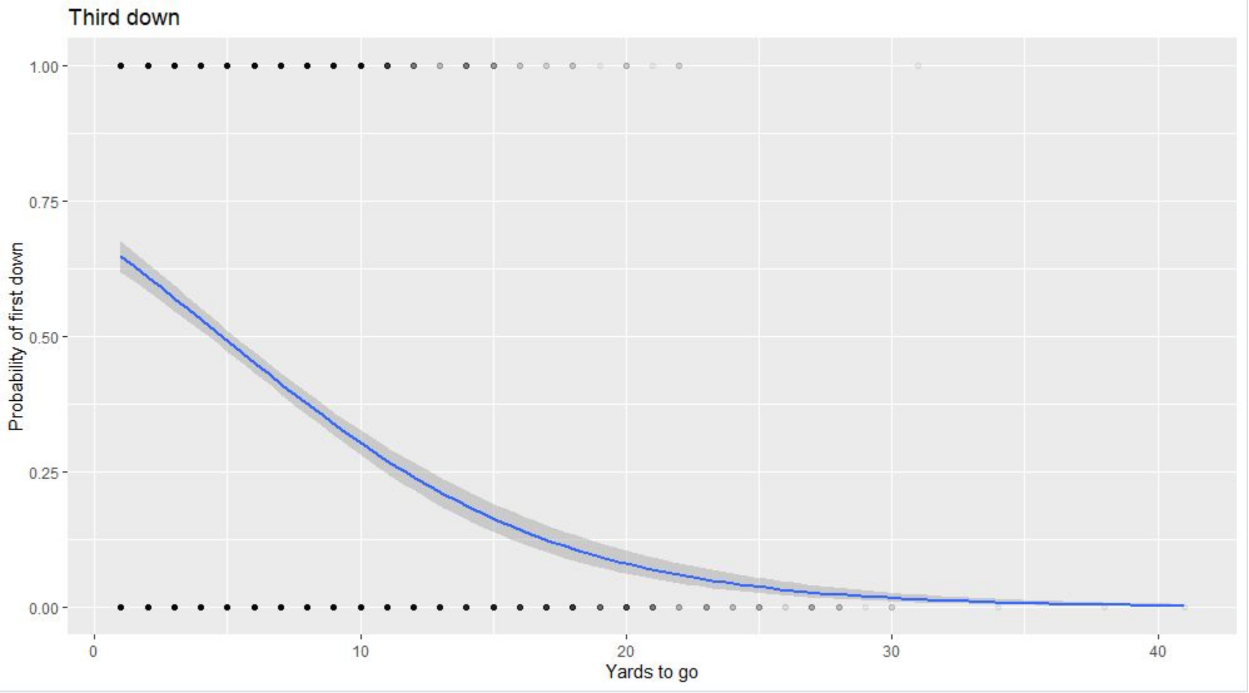
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.42417    0.06122   6.929 2.44e-06 ***
pass         0.05963    0.12403   0.481    0.637    
```



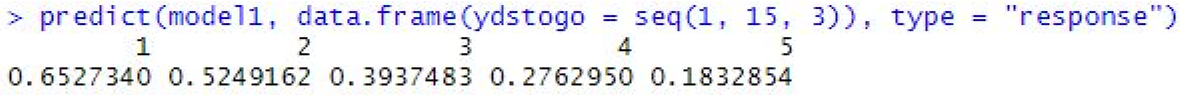
Next, I wanted to examine third downs, which are arguably the most important down in football. How important is it to get into good third down position in order to convert and keep the drive going? To do this, I looked at third down plays and how many yards to go were needed on each play. I could then use logistic regression to see if needing more yards was significant in causing fewer third down conversions.



```
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     0.808    0.0925      8.74 2.41e-18
2 ydstogo        -0.177    0.0132    -13.4  4.85e-41
```



From this model, it can be concluded that third down yards to go is a statistically significant regressor for third down conversion rate (getting a first down). It has a p-value of 4.85e^-41, which is less than .05. From the graph, it can be seen that once we are at about 5 yards to go or more, the model will predict failure to get a first down (regression line is under 50%), and once we get to 10 yards to go, the model predicts only about a 30% chance of success. In order to get a further gauge of the impact of yards to go, I used the predict function in R.

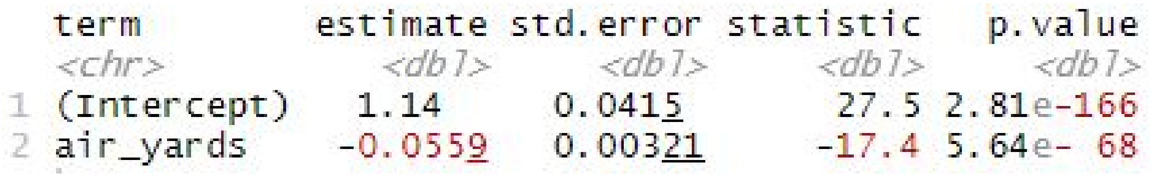


```
> predict(model1, data.frame(ydstogo = seq(1, 15, 3)), type = "response")
        1         2         3         4         5 
0.6527340 0.5249162 0.3937483 0.2762950 0.1832854 
```

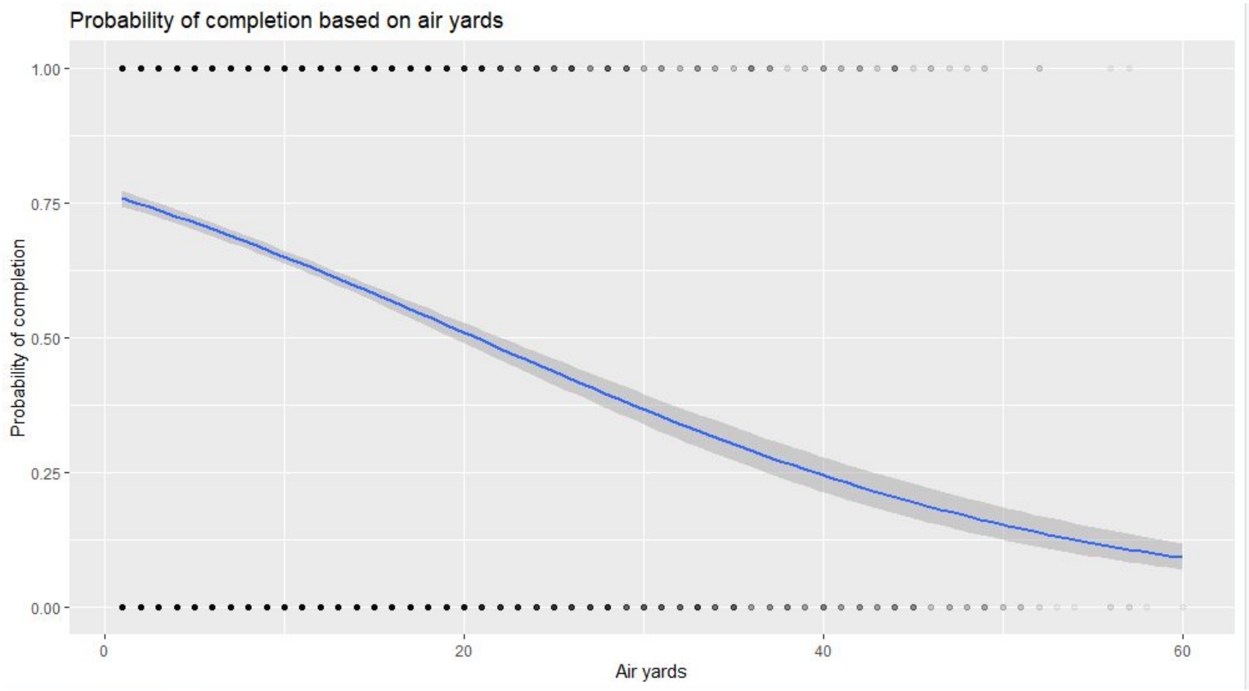
This looks at the model’s predictions for intervals of 3 up to 15 yards to go. So “1” represents 1-3 yards to go, “2” represents 4-6, “3” represents 7-9, “4” represents 10-12, and “5” represents

13-15 yards to go. From this, we can see that for every 3 yards more that are needed on third down, our model predicts the chances of succeeding to decrease somewhere from 9-13%. We can also say that from this model, the chance of converting a 3rd down with 1-3 yards to go are over twice as good as the chance of doing so with 10-12 yards to go. This shows the importance of doing well on 1st and 2nd down plays so that a team can get in good position on 3rd down and have a better chance of converting and extending a drive.

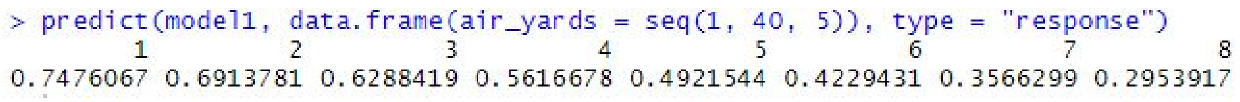
Another interesting and important variable that I wanted to look at which might contribute to success or failure is air yards. Air yards are the yards that a pass travels in the air before it reaches the receiver. This is something I used a logistic regression on to evaluate if pass distance has a significant effect on whether that pass is complete (the binomial variable).



```
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)   1.14     0.0415       27.5 2.81e-166
2 air_yards    -0.0559   0.00321     -17.4 5.64e- 68
```



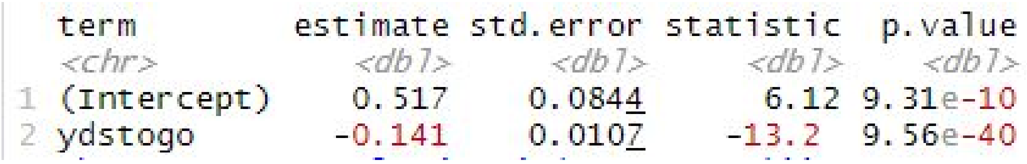
As you can see from the p-value, air yards is a significant regressor in this model for whether or not a pass is complete. This would mean that, in general, pass distance will have a big impact on completion percentage, not just how covered the receiver is. It seems that even if a receiver is fairly open on a deep pass, that pass still isn't likely to be completed. Maybe quarterbacks just aren't as accurate on deep passes, or maybe defensive players are faster at getting to the ball. From the graph, it can be shown that for passes which travel fewer than 10 yards have around a 65% chance or greater of being completed, but once a pass is being thrown 20 or more yards, that chance falls below 50% and the model would predict an incomplete pass. I once again used the predict function to generate some values and see the difference.



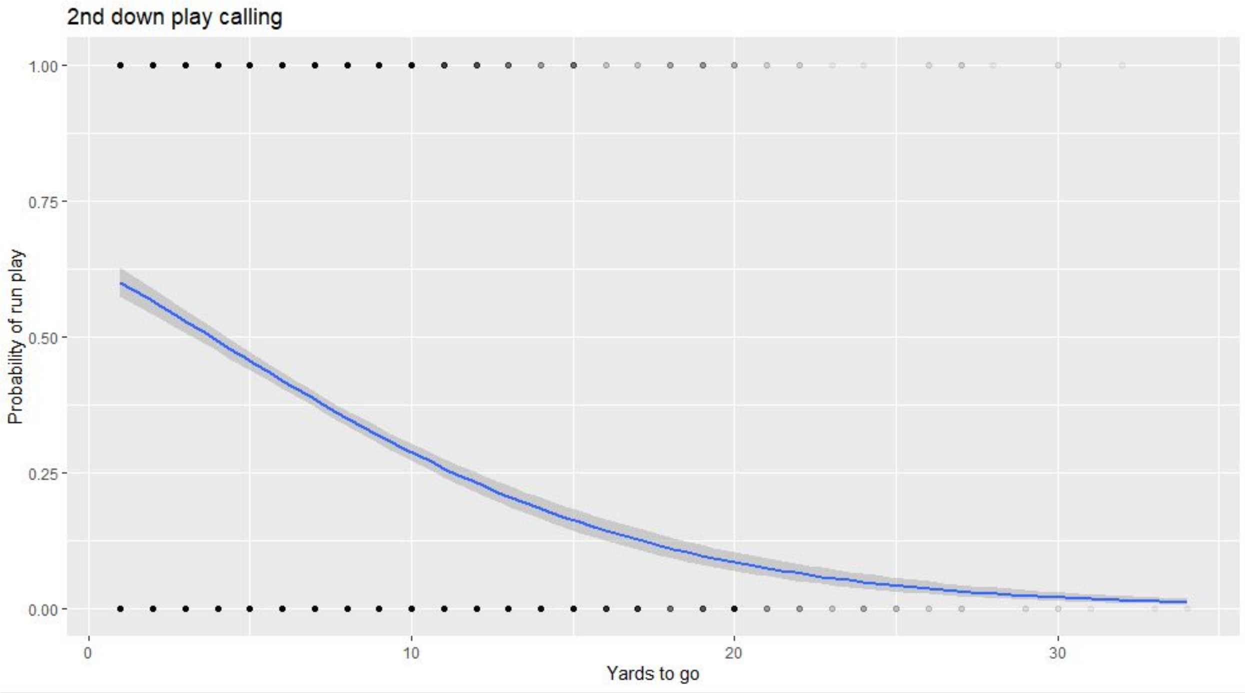
```
> predict(model1, data.frame(air_yards = seq(1, 40, 5)), type = "response")
        1         2         3         4         5         6         7         8 
0.7476067 0.6913781 0.6288419 0.5616678 0.4921544 0.4229431 0.3566299 0.2953917 
```

Here, each value represents 5 more air yards (1-5, 6-10, 11-15, etc.). From this, it can be seen that for every 5 more air yards that a pass travels, the probability of completion will drop about 7% for the most part in our model. The chances of completing a pass between 35-40 yards (#8 value) is about half of the chance of completing a 16-20 yard pass (#4 value). It is worth noting that a typical QB completion percentage is in the mid-60s, so having a 50% probability of success is rather poor, and the #1 value of .747 for 1-5 yard passes is very good.

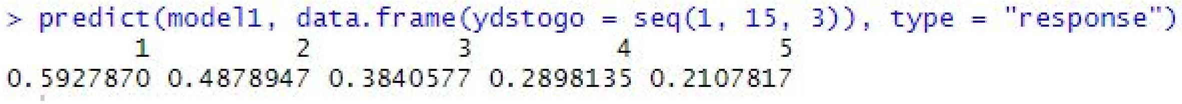
One other thing I wanted to look at was potential patterns in offensive decision making. What type of play do teams typically call on a certain down and distance? I created a logistic regression model for all 2nd and 3rd downs, looking at yards to go and resulting run rate. I didn't use first down, because almost all 1st downs start with 10 yards to go. Here is 2nd down.



```
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.517    0.0844      6.12 9.31e-10
2 ydstogo       -0.141    0.0107    -13.2  9.56e-40
```

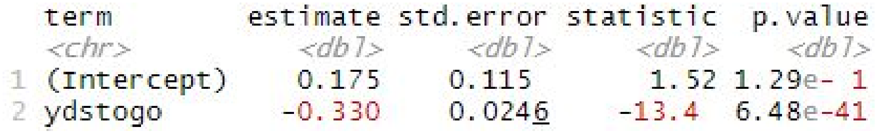


P-value is statistically significant, so it looks like yards to go is a significant indicator of play calling on 2nd down. From the graph above, you can see that on 2nd and short (3 yards or less), the model would predict a run, otherwise it will probably predict a pass. On 2nd and 10 or more, the probability falls below 25% of a run call. It is clear that the worse a team does on first down (more yards to go), the more likely they are to pass and not run the ball on 2nd down. Here are the predicted values in intervals of 3 from 1 yard to go up until 20.

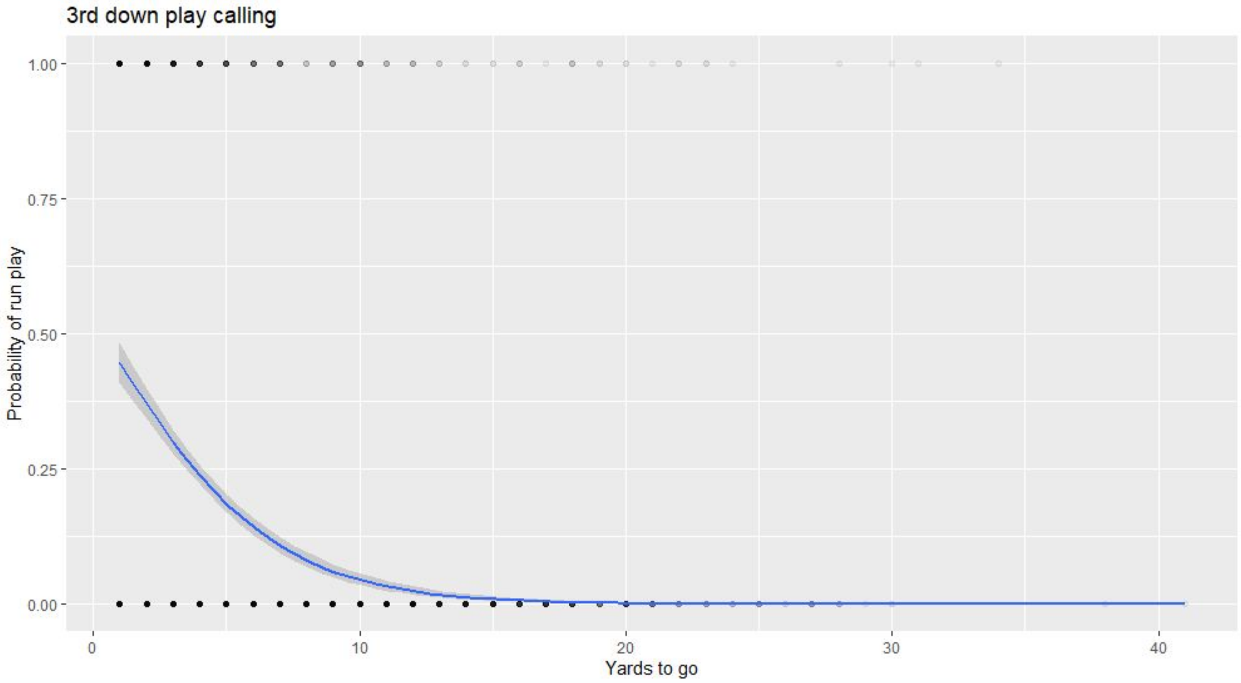


```
> predict(model1, data.frame(ydstogo = seq(1, 15, 3)), type = "response")
        1         2         3         4         5 
0.5927870 0.4878947 0.3840577 0.2898135 0.2107817 
```

Our model predicts that the chances of a team running the ball will more than double when there are 3 yards or less to go compared to needing 10-12 yards. Lastly, let’s look at this for 3rd down.



```
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    0.175    0.115       1.52 1.29e- 1
2 ydstogo       -0.330    0.0246    -13.4  6.48e-41
```



The p-value here is also statistically significant, so yards to go is also significant in determining play calling on 3rd down. The graph illustrates how teams generally do not run the ball on 3rd down no matter what, but even on 3rd and very short (1-2 yards) there is still only about a 40-45% predicted chance of teams running. As yards to go increases, that percentage continues to dwindle a lot, which shows the significance. It is pretty clear that the only time defenses should be worried about their opponent running on 3rd down is when it is 3rd and 1 or maybe 3rd and 2.

References

Ben Baldwin. A beginner's guide to nflfastR. Retrieved October 8, 2020 from https://mrcaseb.github.io/nflfastR/articles/beginners_guide.html

Ben Baldwin. An R package to quickly obtain clean and tidy NFL play by play data. Retrieved October 8, 2020 from https://mrcaseb.github.io/nflfastR/

guga31bb. 2020. guga31bb/nflfastR-data. (October 2020). Retrieved October 3, 2020 from https://github.com/guga31bb/nflfastR-data/blob/master/data/play_by_play_2020.zip