R Notebook

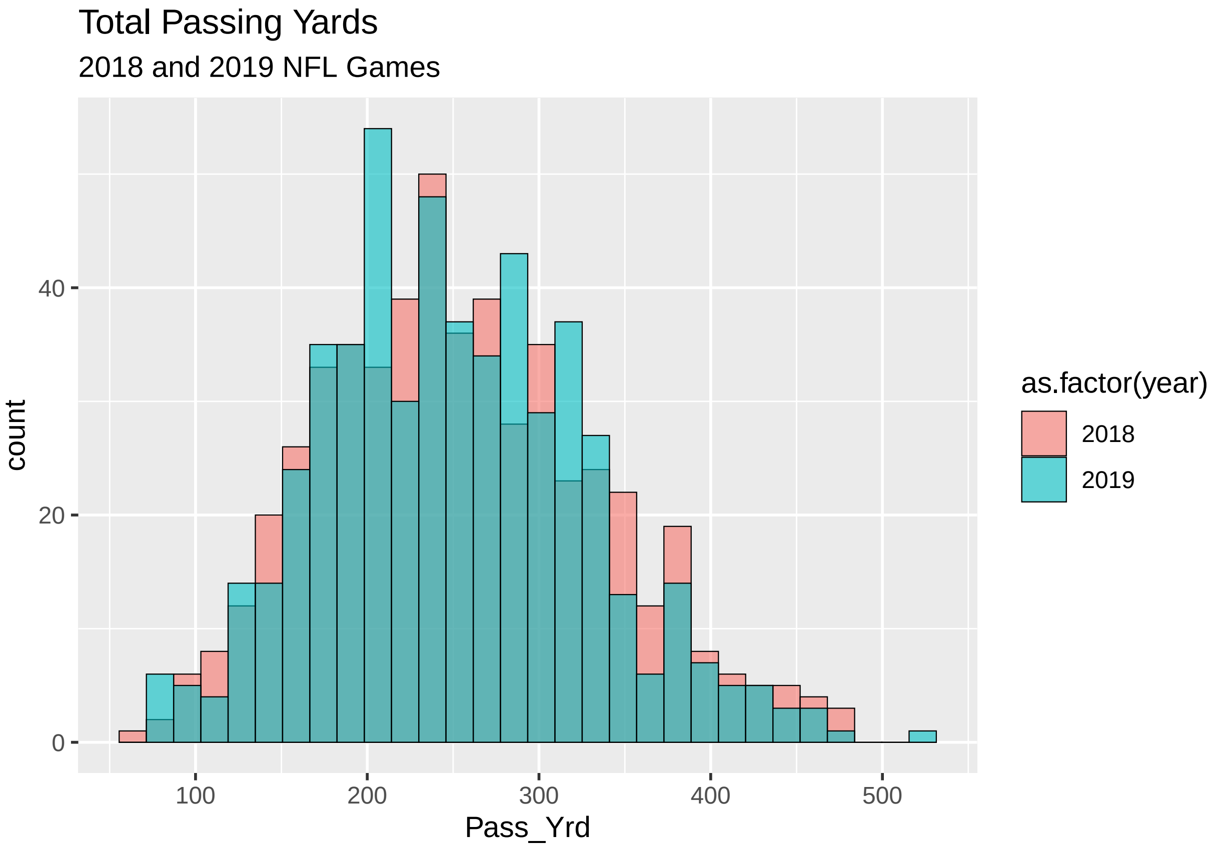
Alex Baker, James Kitch, Jackson Smith, Matthew Sheridan

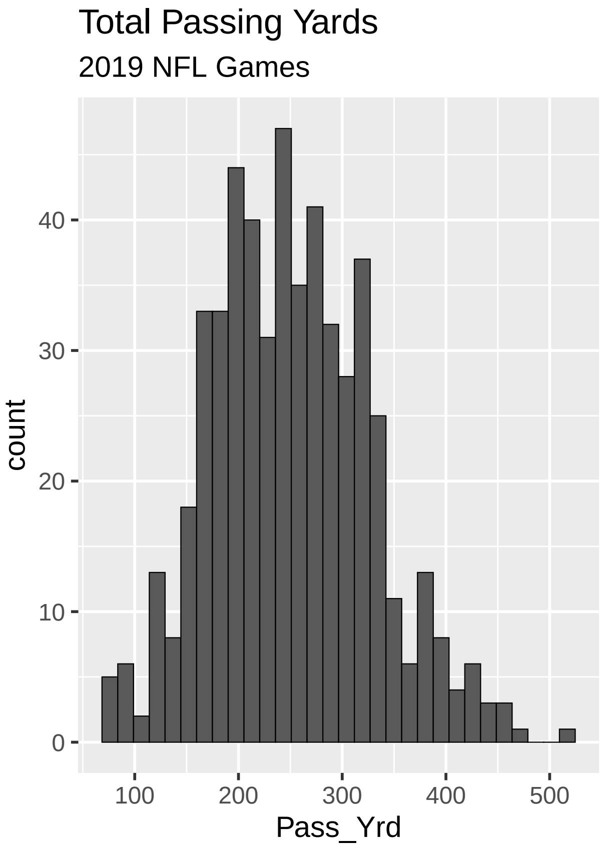
Introduction:

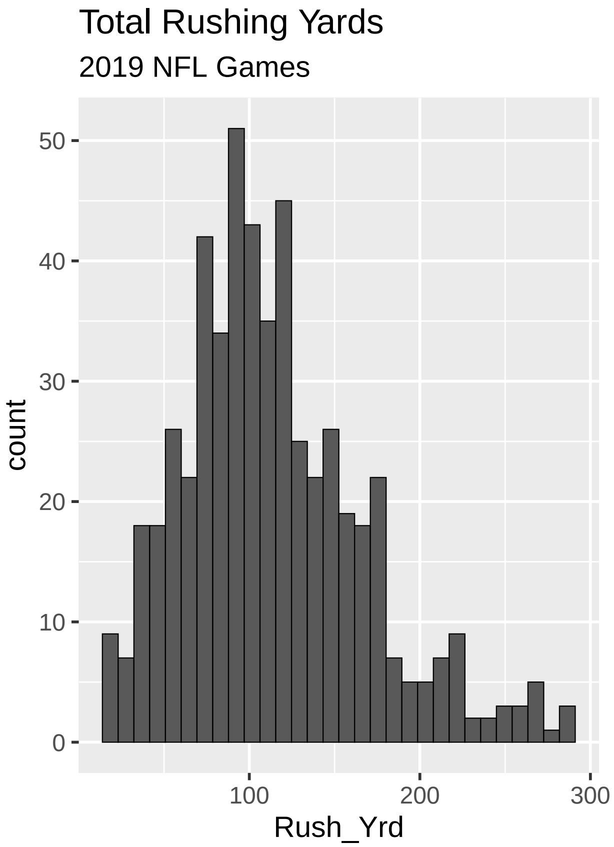
Data Cleaning, EDA:

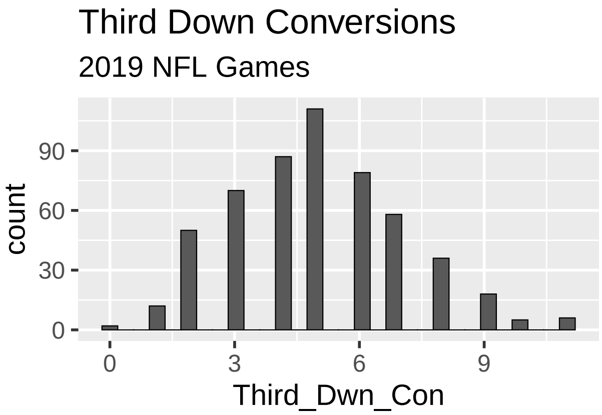
Below, we graphed the distributions for a number of different potential predictors, as well as their relationship with the response variable of interest. Most predictors are slightly right-skewed, as they have a positive support and have a slight "bell" shape for the most common values but are stretched out by the few exceptional games where a team throws for 500 yards or rushes for 250 yards. However, it is mild right-skewness as it is not to the extent that log-transforming them would make the distributions more symmetric. For example, if we log-transform `Pass_Yrd` we actually find that the the resulting distribution is left-skewed, suggesting that the log transformation was actually too strong!

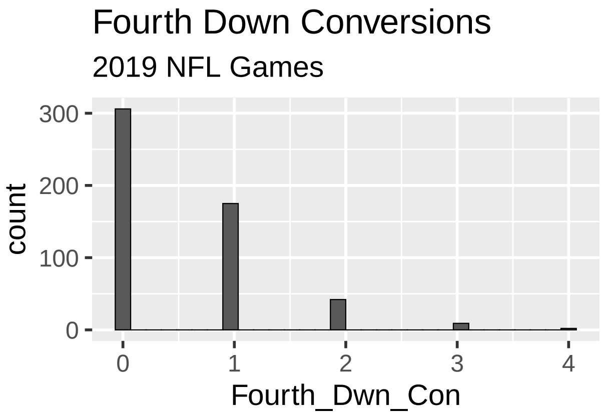
Furthermore, we can perform a quick check and see that the distribution of passing yards is roughly the same in 2019 as it was in 2018.

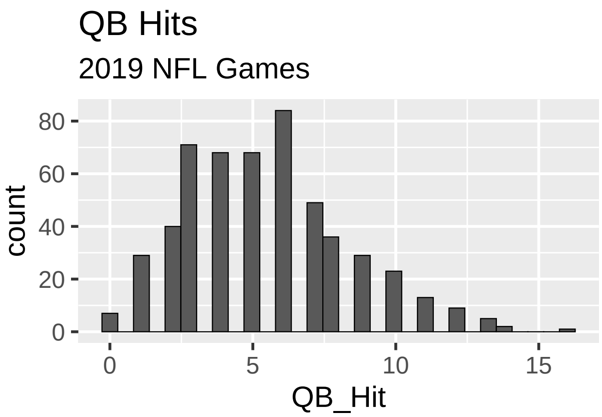


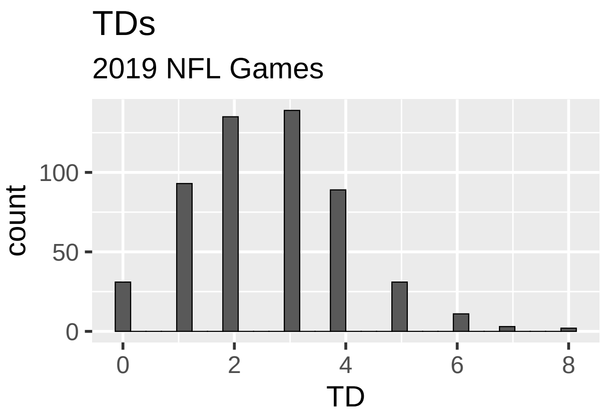

Total Passing Yards
2019 NFL Games
count
0
10
20
30
40
100
200
300
400
500
Pass_Yrd


Total Rushing Yards
2019 NFL Games
count
0
10
20
30
40
50
100
200
300
Rush_Yrd









Initial Modeling:

The code for initializing the models has been hidden for space constraints. The linear model was using all predictors, the stepwise model stepped from an intercept only model to all predictors, and the randomforest has not been tuned (YET) and uses all predictors, as well as the default parameters.

```
##        rmse_2017 rmse_2018 rmse_2019 rmse_2020 rmse_2021
## Linear  9.992705  9.833446  9.919896  9.924871  10.14683
## Step    9.986854  9.799207  9.939556  9.923551  10.13224
## RF1    10.729163 10.304991 10.782059 10.505457  11.11002
```

Conclusion:

National Football League data provides a ripe opportunity for statistical analysis, especially when able to access play-by-play level data using the `nflfastR` package. While most variables have a slight right skew to them and only take on positive values, this did not present tremendous challenges for our models. Going forward, we're interested to continue investigating the impact of variables that provide a clear relationship towards "success" in a game–such as total passing yards or total rushing yards–as well as more "strategic" variables such as the number of third and fourth down conversions. Given the violations of symmetric distributions in many of the predictors and the potential for non-linear relationships to arise, examining the performance and interpretations of Random Forest models will also be a focus in our analysis.