Football Analytics Blitz Case

Villanova Sports Analytics Club

Harry Lazerwitz, Joshua Mould, Kevin O'Donnell, Noah Swan

Overview

Context

- Our approach
- Success rates
- First play of drives

Main Findings

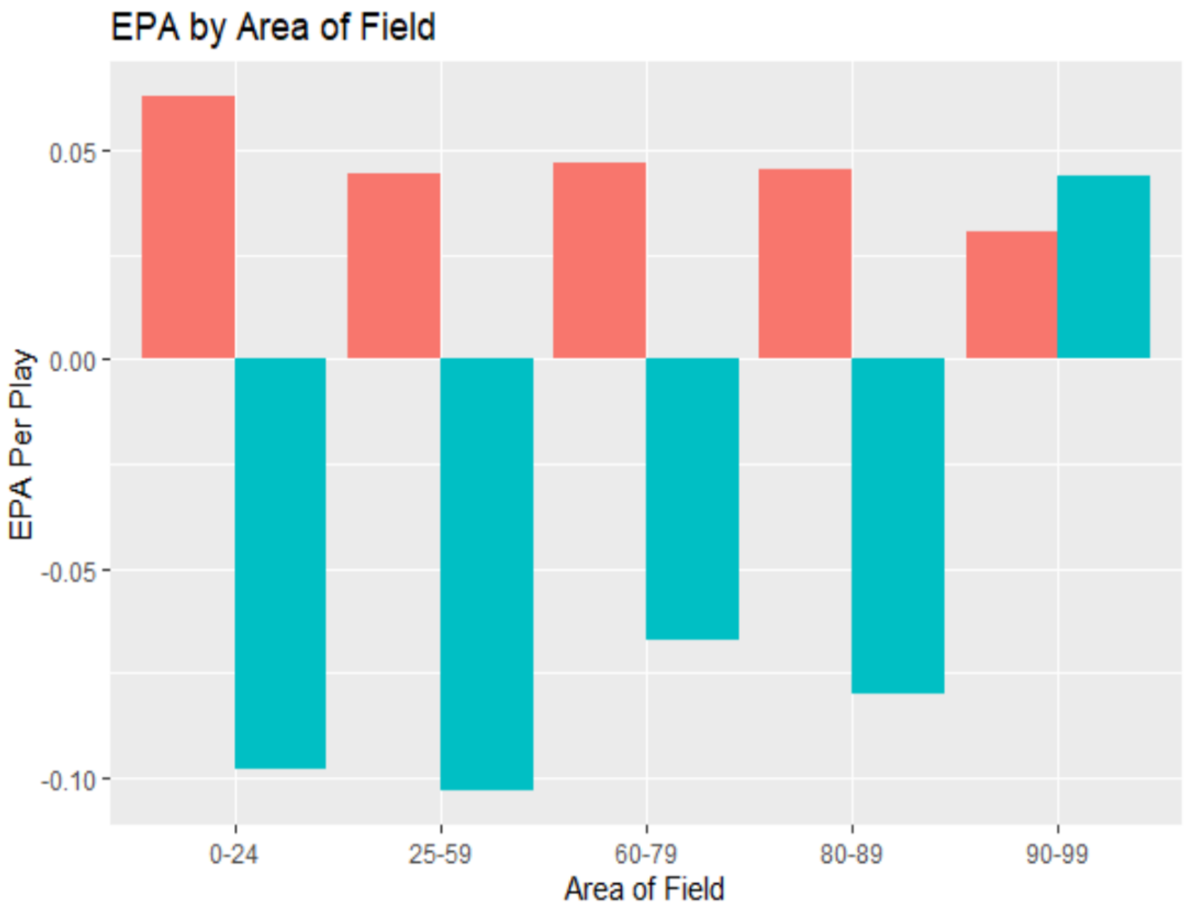
- Leaguewide ratios
- Team analysis
  - Play Action

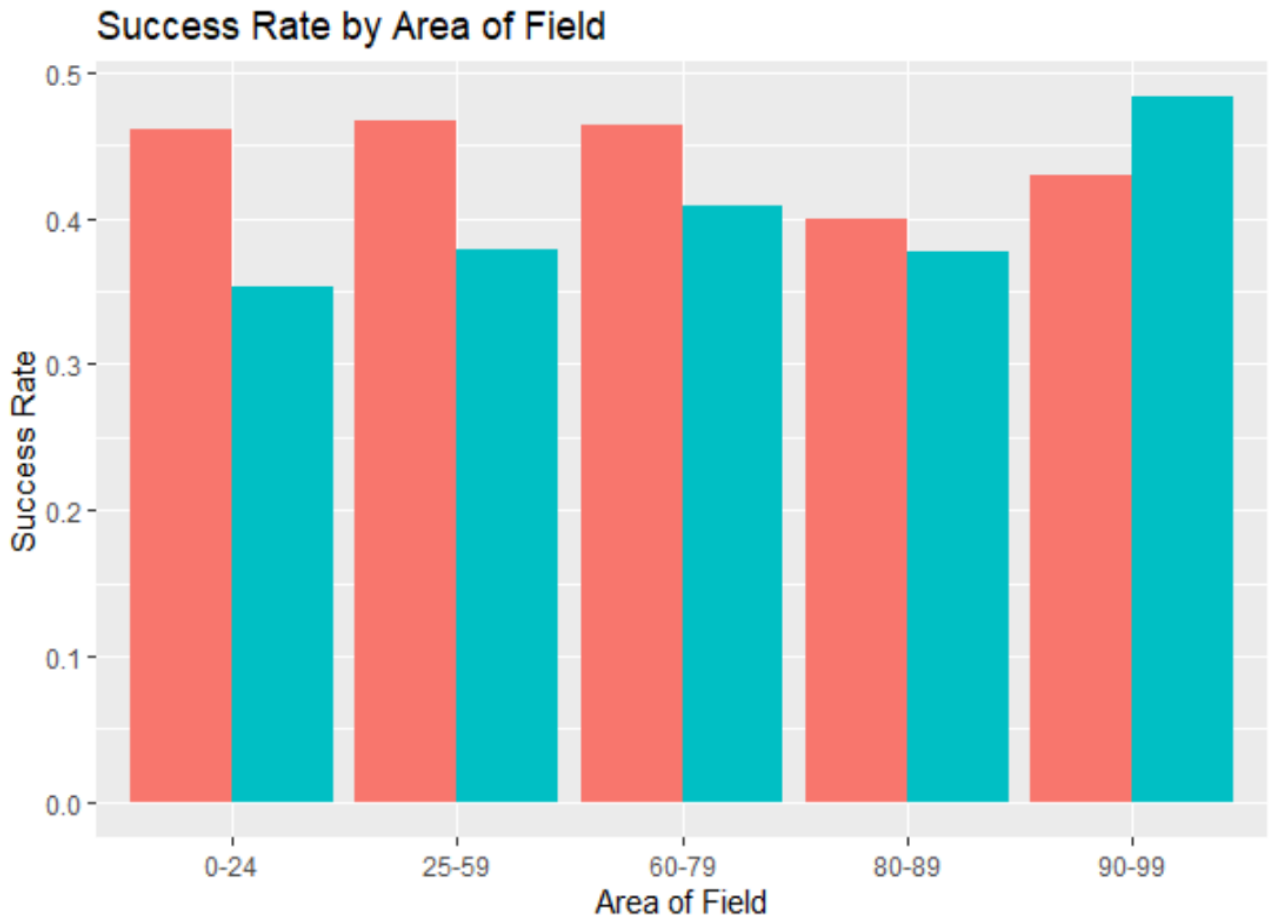
Other Notes

- Play sequencing
- Close range play calling

Our Approach

- Analysis in R
- Combined provided data with nflfastR
    - Change scrambling to pass play
- Some graphing in Tableau

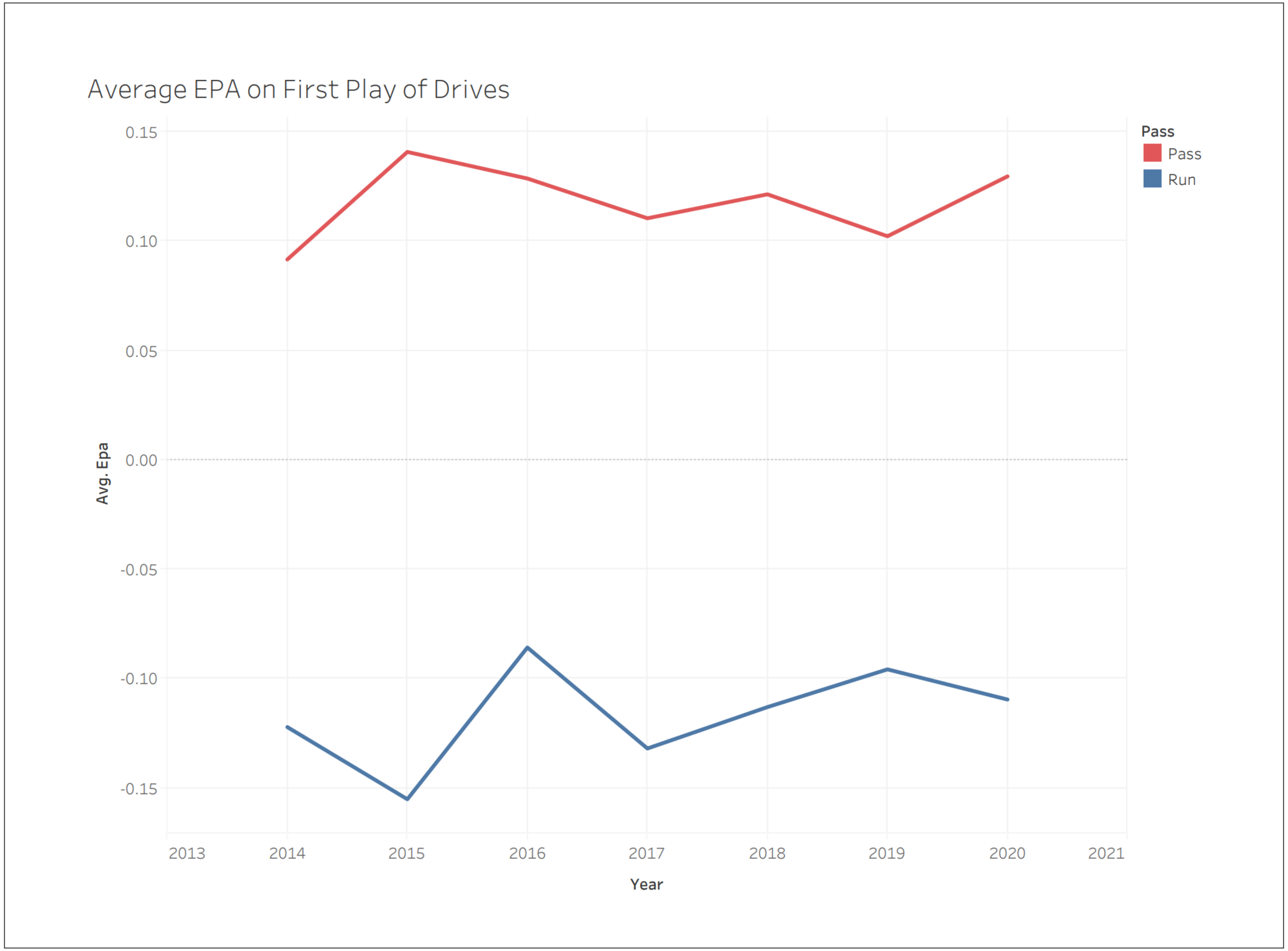


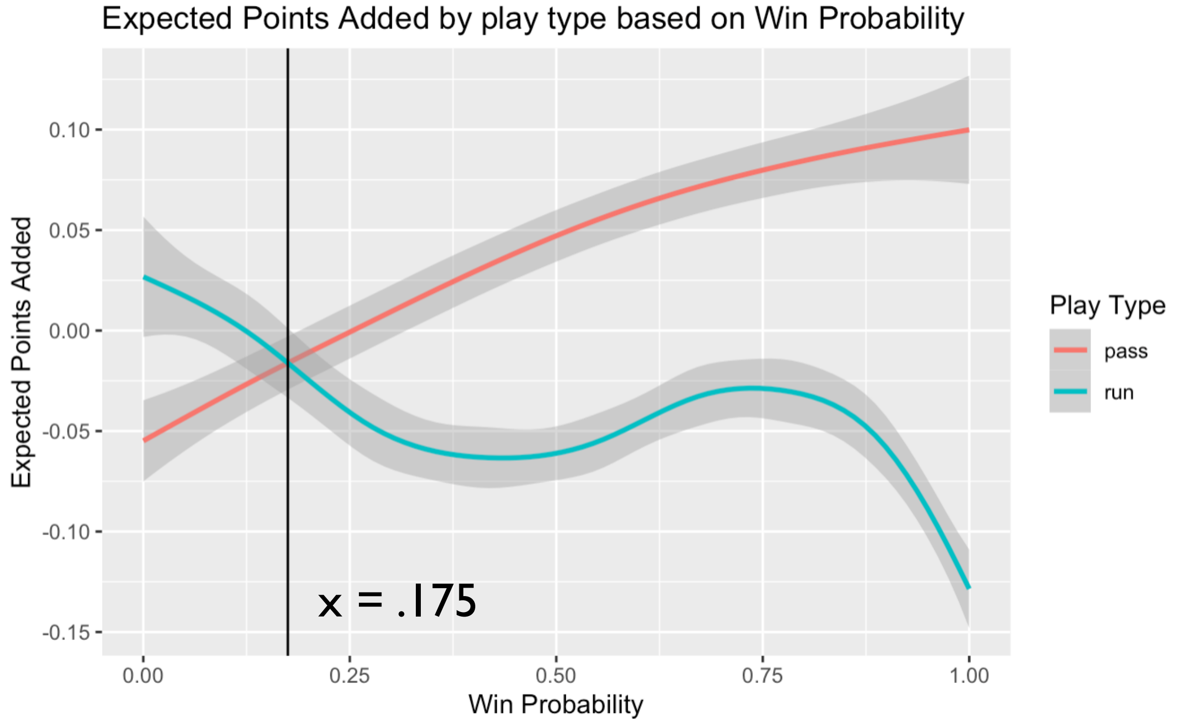


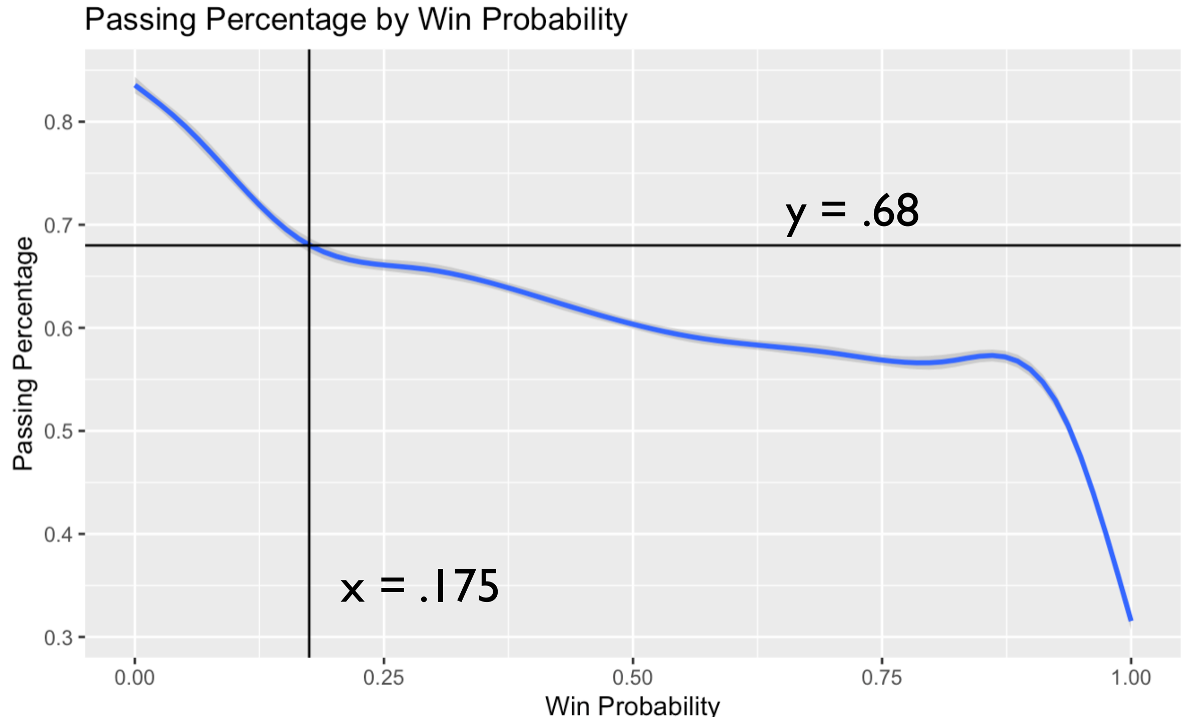
Leaguewide Success by Area of Field

First Play of Drives

- Convention is to establish the run
- Passing is much more successful
  - Expectation of defense





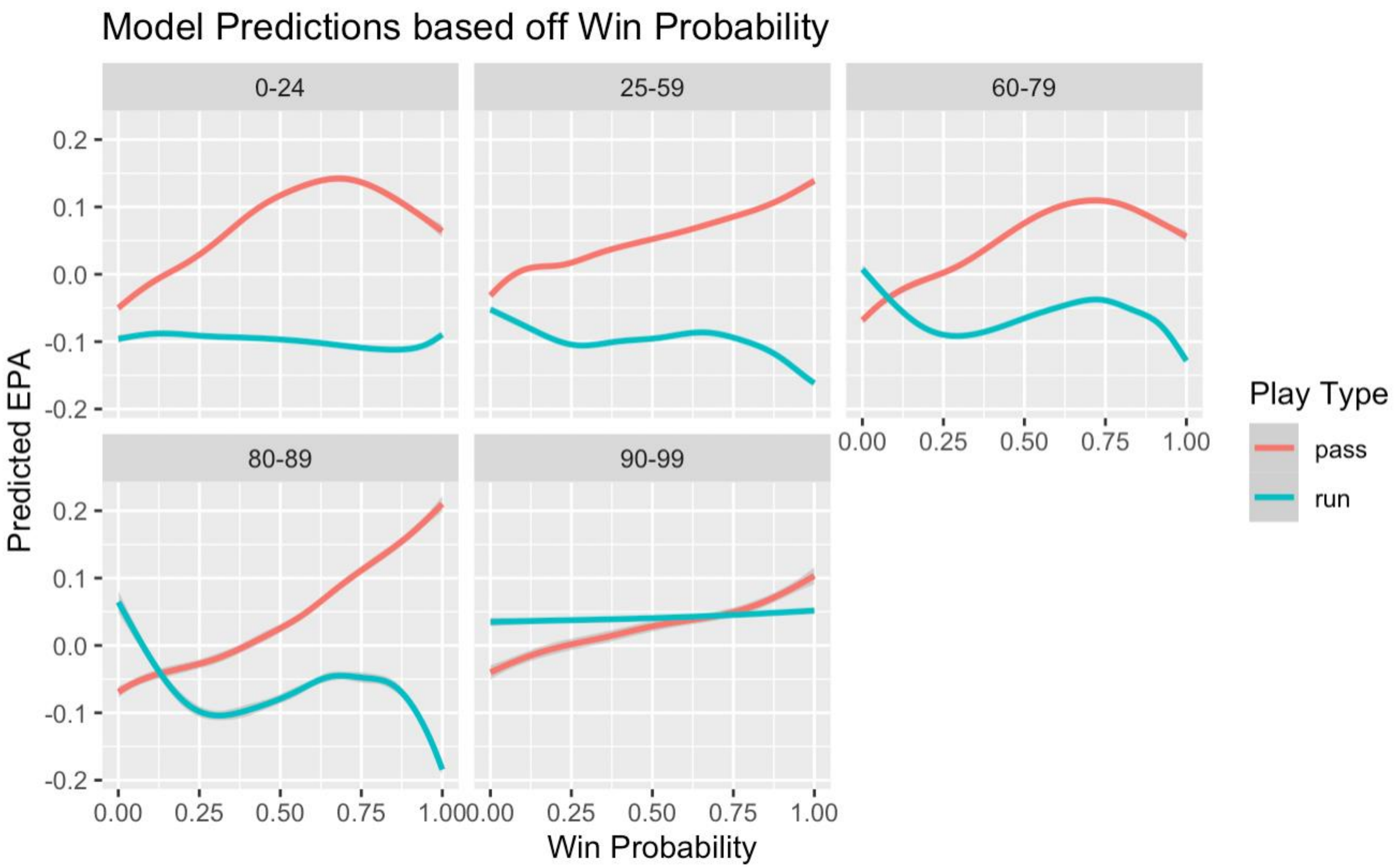


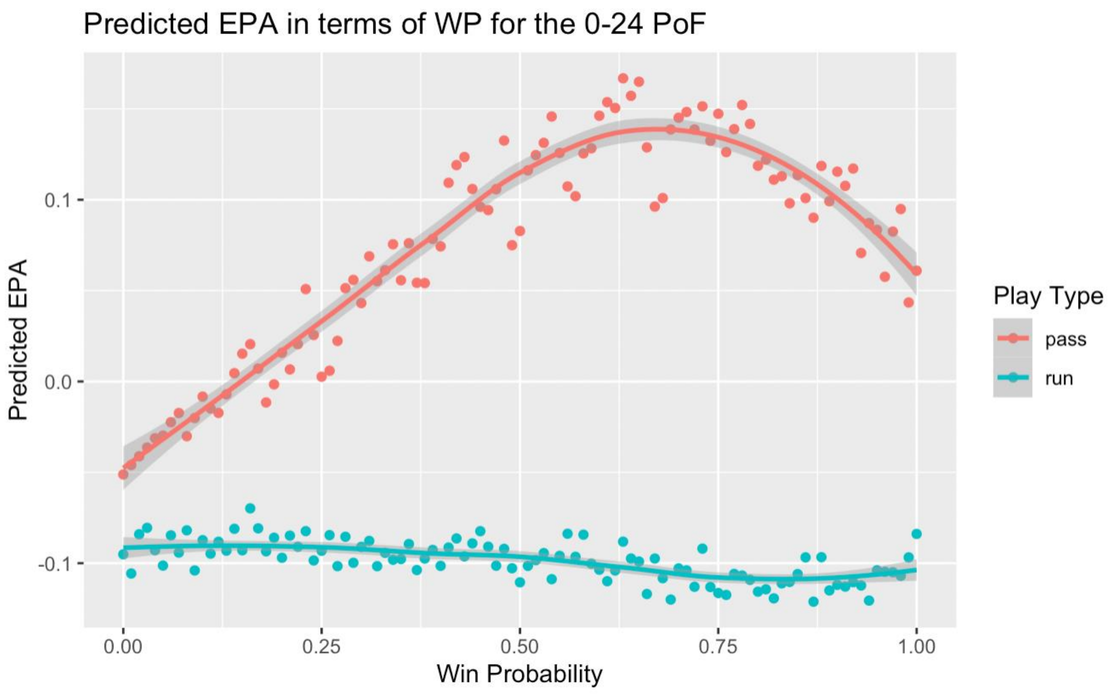
Optimal Ratios

- Took the approach of David Schmerfeld: https://davidschmerfeld.github.io/nfl-optimum-pass-run-ratio/
- Based our models on graphs of WP and EPA in order to cross reference with passing percentages
- Smoothed very noisy data
- Separated by part of field

By Part of Field

- Split up by part of field
- Built models for each play type in each zone
- Smoothed
- Parameters:
  - Win Probability
  - Down
  - Distance
  - Facet Grades
  - Previous play (pass/rush)





Finding Optimal WP

- Rounded WP of predictions
- Grouped according to WP
- Averaged predicted EPAs
- Loess models to predict EPA
- Predicted for each WP
- Found minimum difference between pass and run EPA
  - WP at that point is optimal

Optimal Passing %

- Loess models for Pass % for each part of the field
- Predicted Optimal Pass % for Optimal WP

Part of Field	Breakeven WP	Pass Percentage
0-24	0%	80.4%
25-59	0%	82.1%
60-79	8%	79.3%
80-89	14%	75.5%
90-99	69%	49.0%

Titans 2020

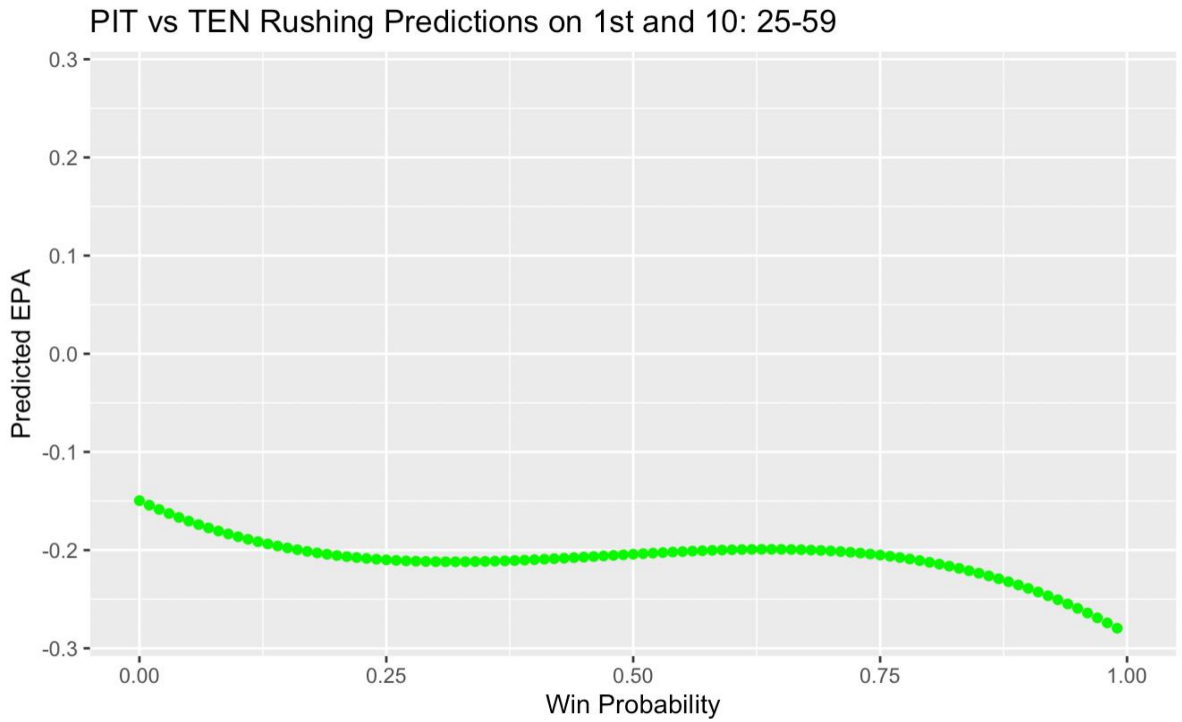
- Much lower than the average Optimal Pass %
- Heavy Rushing offense

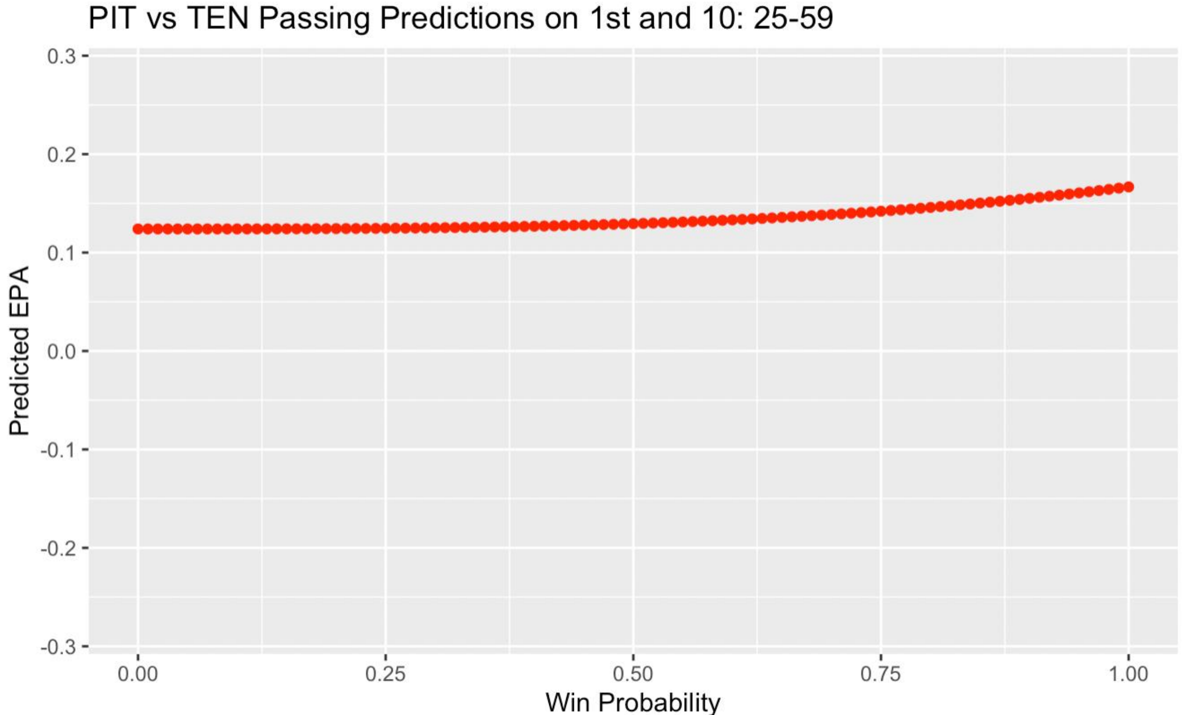
Part of Field	Breakeven WP	Passing Percentage
0-24	11%	52.6%
25-59	4%	63.7%
60-79	10%	65.6%
80-89	45%	41.5%
90-99	78%	39.7%

Steelers 2020

- Very high optimal Pass %
- Heavy Passing Offense

Part of Field	Breakeven WP	Passing Percentage
0-24	0%	79.1%
25-59	0%	87.7%
60-79	78%	78.9%
80-89	30%	90.6%
90-99	83%	47.7%





Situational Applications

- Model is not exclusively big picture
- 2020 Steelers against the 2020 Titans
  - Steelers have the ball in the 25-59 yard range on 1st and 10
- Results in Optimal WP of 0, Optimal Pass % of 87.7%

Optimal Play Action

- Same method to find Optimal Play Action %
- Loess model for Play Action % and predicted from Optimal WP

Part of Field	Breakeven WP	Play Action Percentage
0-24	0%	12.8%
25-59	0%	10.7%
60-79	8%	12.6%
80-89	14%	11.2%
90-99	69%	28.4%

TEN 2020 Optimal PA

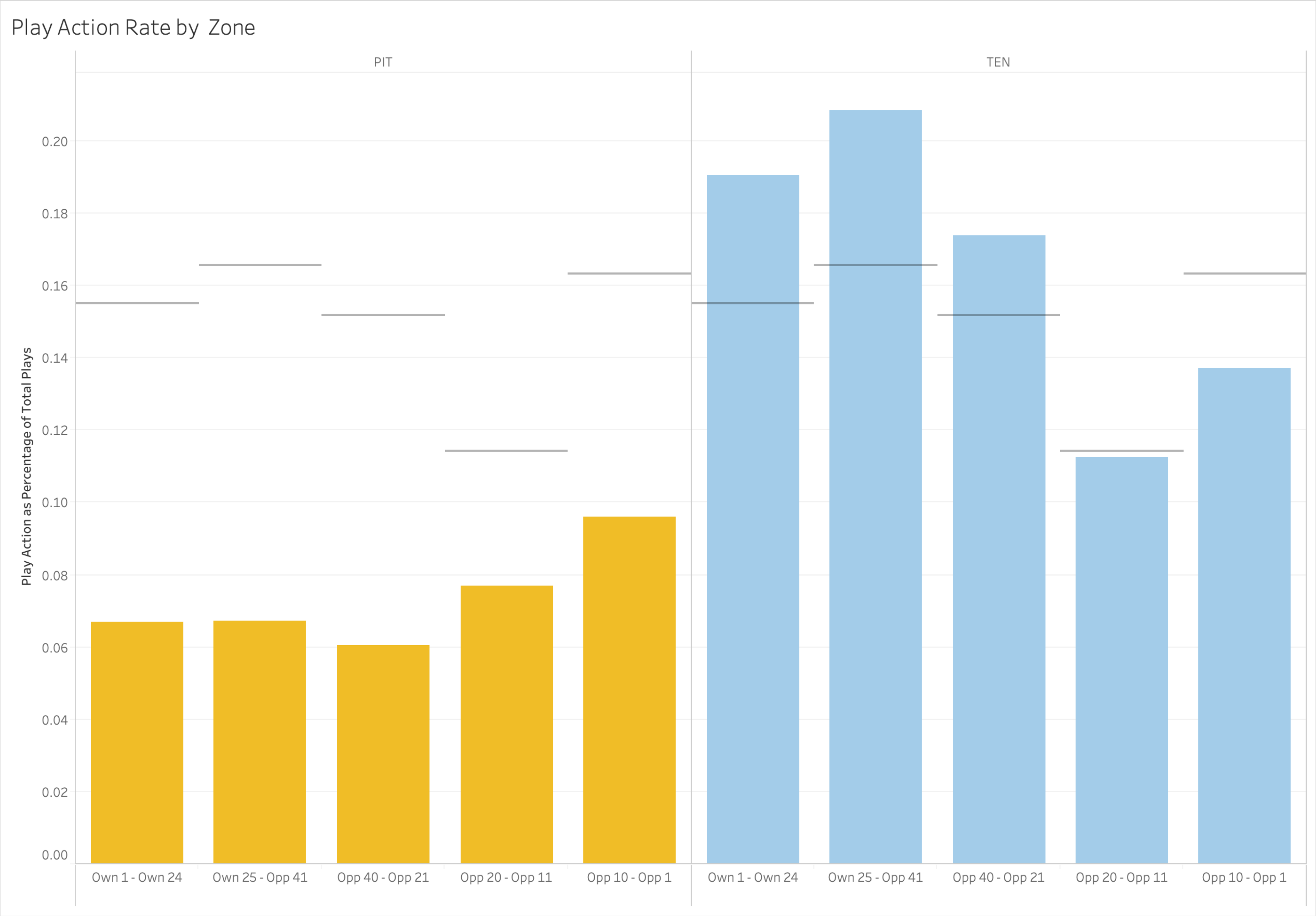
- Might expect more runs, so Play Actions would be more successful

Part of Field	Breakeven WP	Play Action Percentage
0-24	11%	11.3%
25-59	4%	30.3%
60-79	10%	19.3%
80-89	45%	22.9%
90-99	78%	31.9%

PIT 2020 Optimal PA

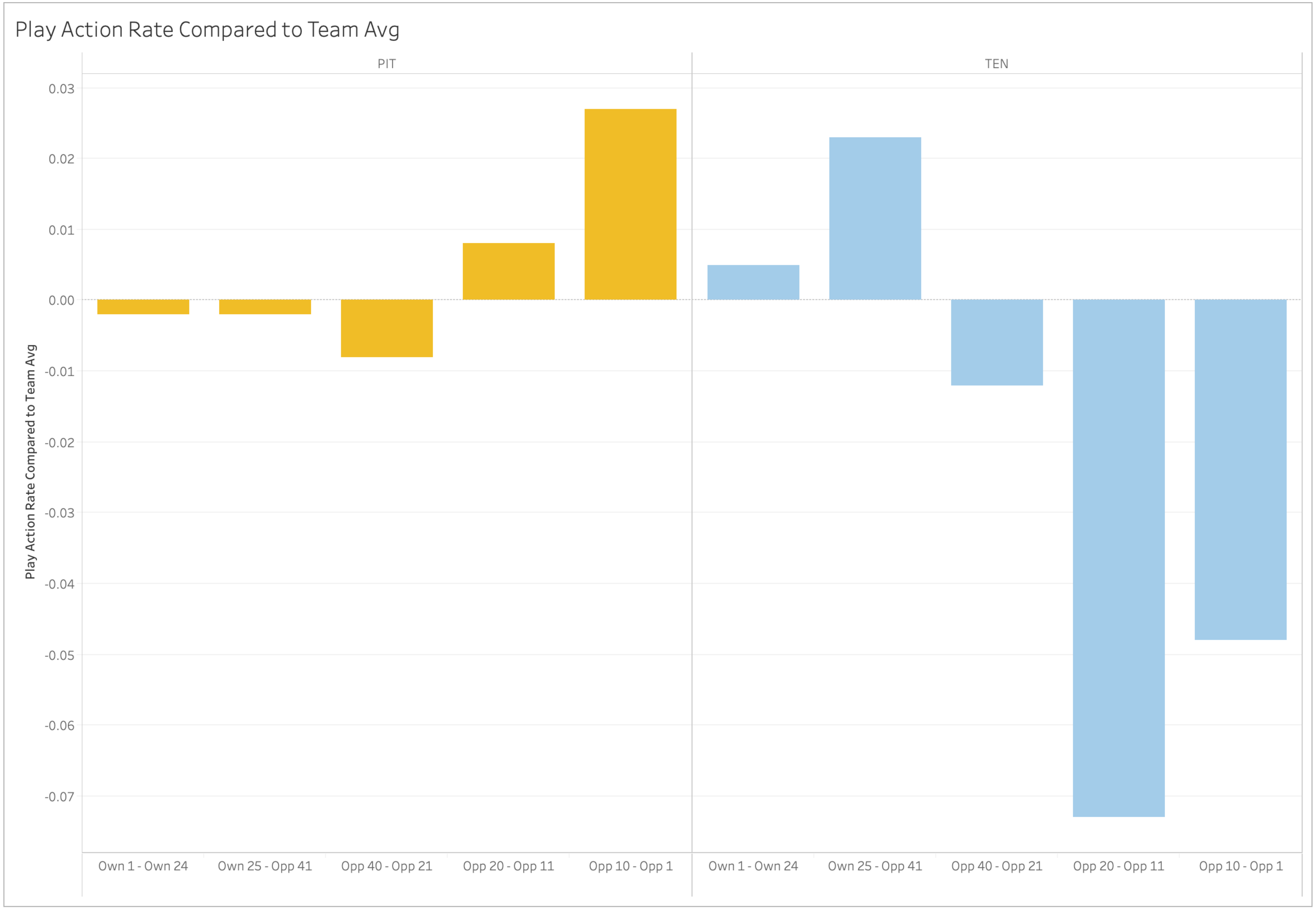
- More likely to pass, faking run won't fool anyone

Part of Field	Breakeven WP	Play Action Percentage
0-24	0%	16.3%
25-59	0%	2.0%
60-79	78%	12.5%
80-89	30%	10.9%
90-99	83%	17.1%



Play Action

Gray bars are league average

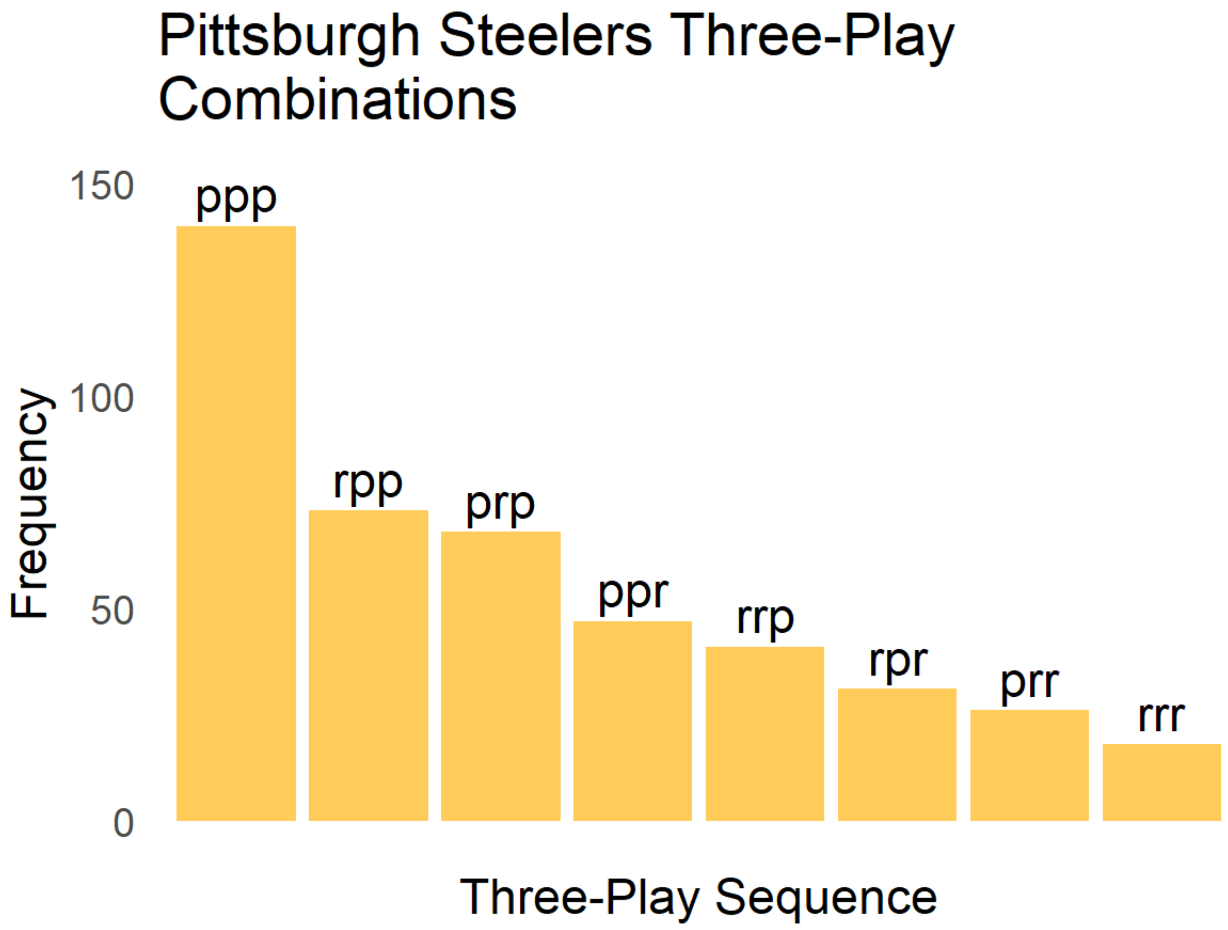


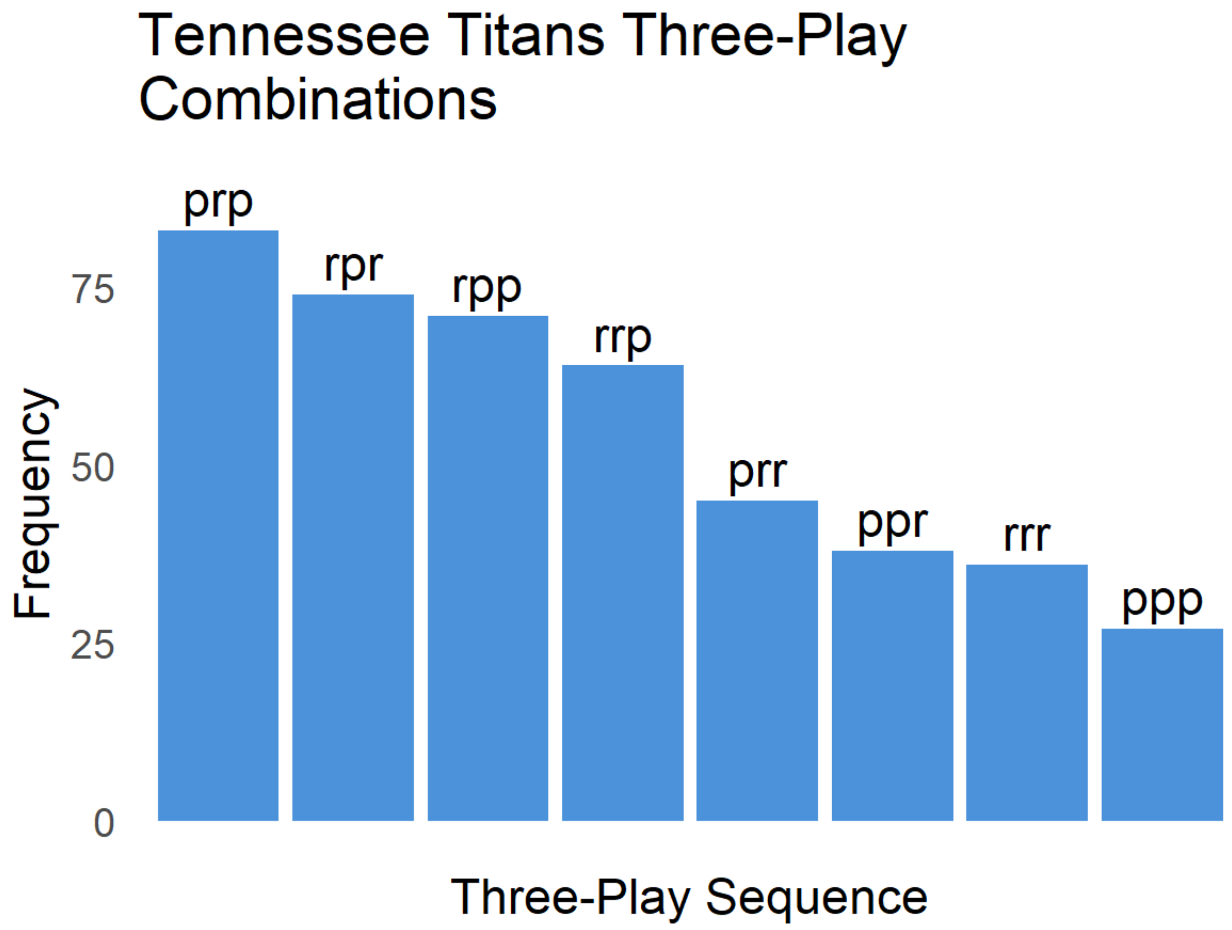
Play Action

Zone play action rates compared to overall rates for each team

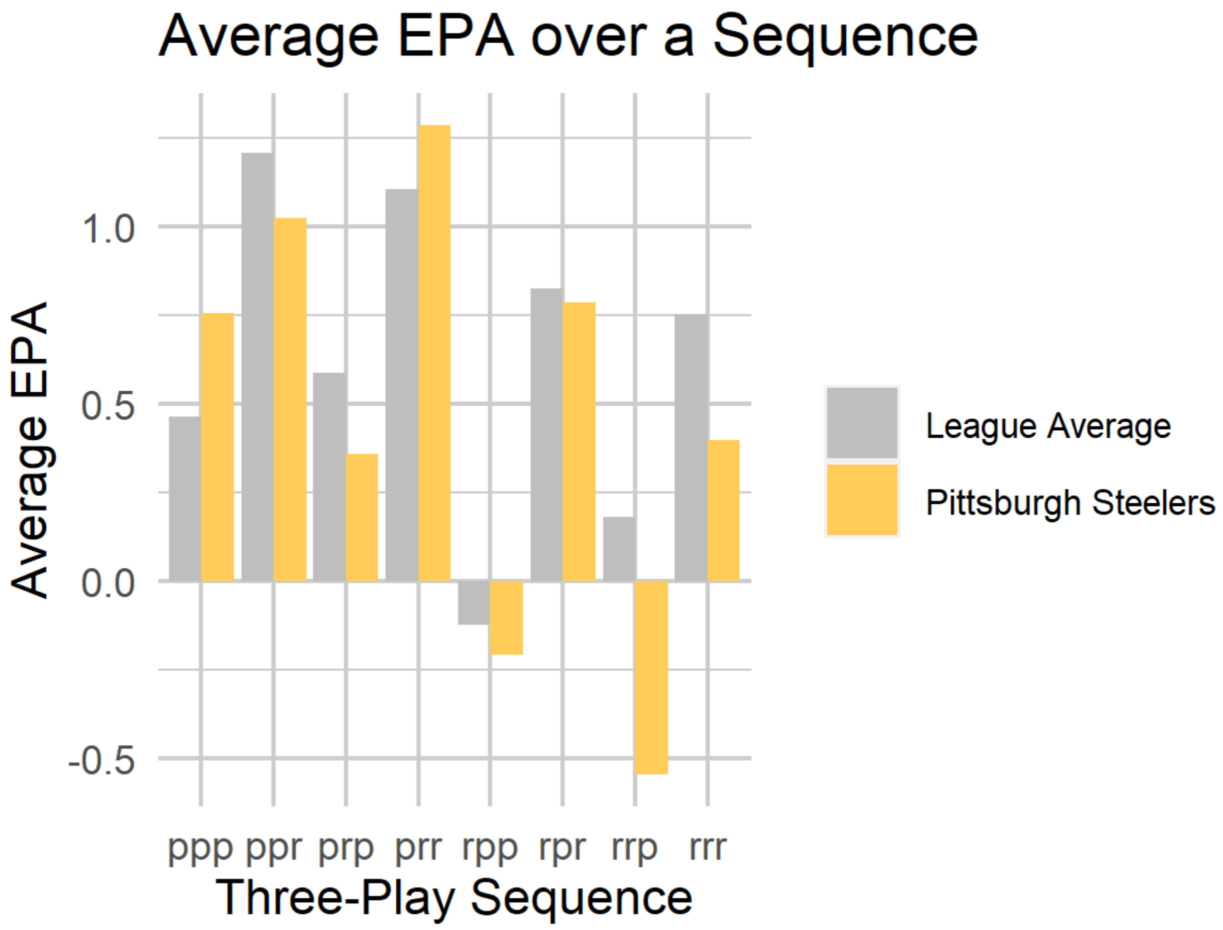
Play Action

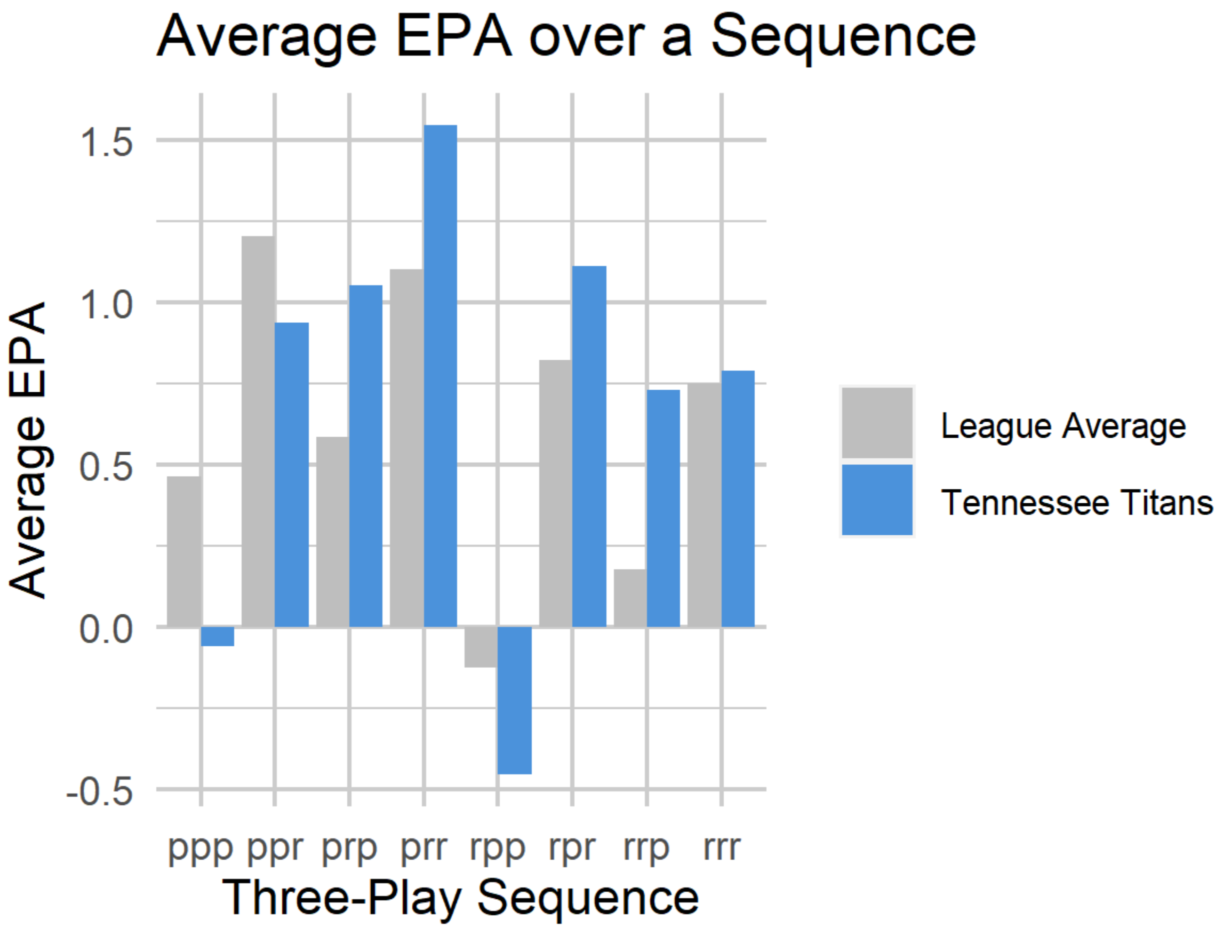
Green bars are optimal play action rate for each team and zone



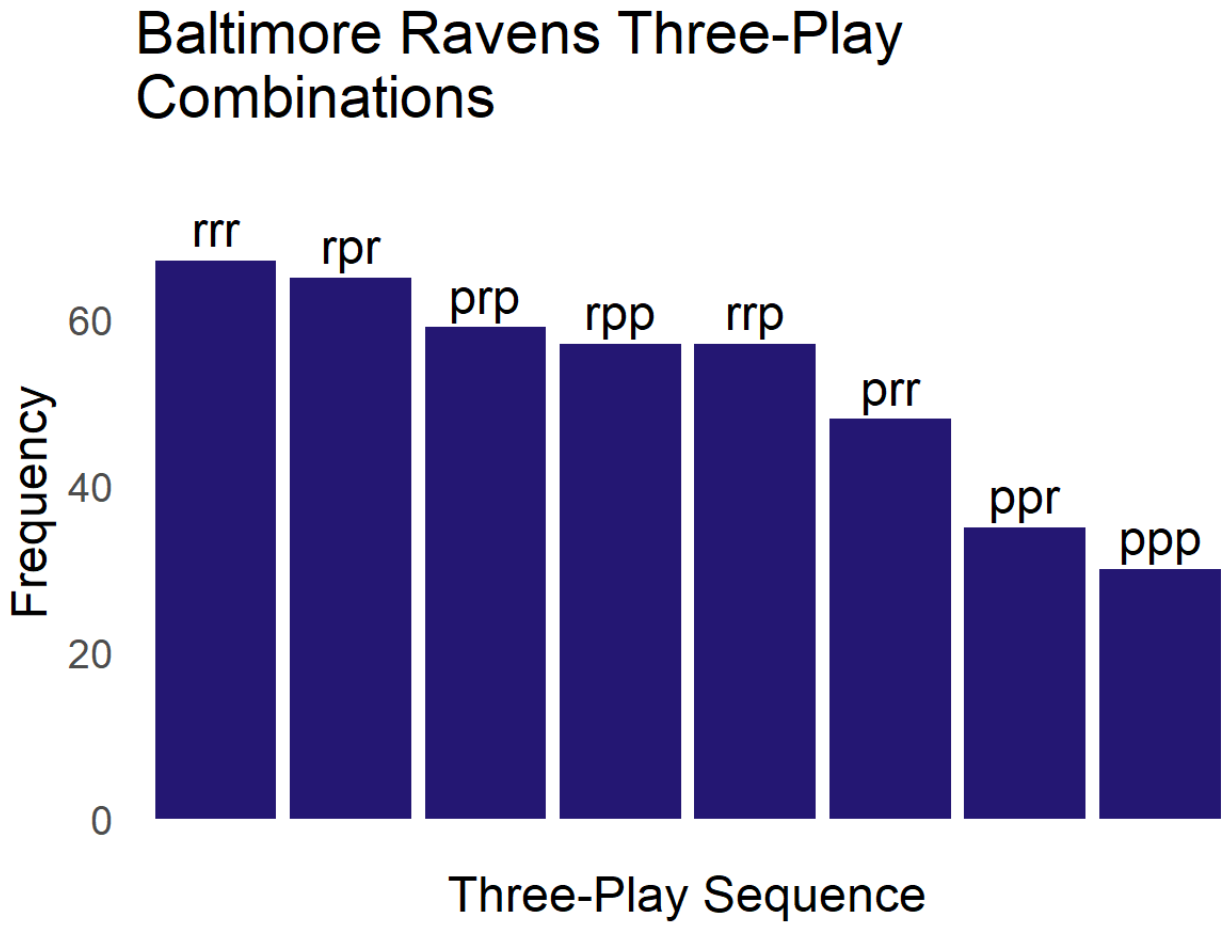


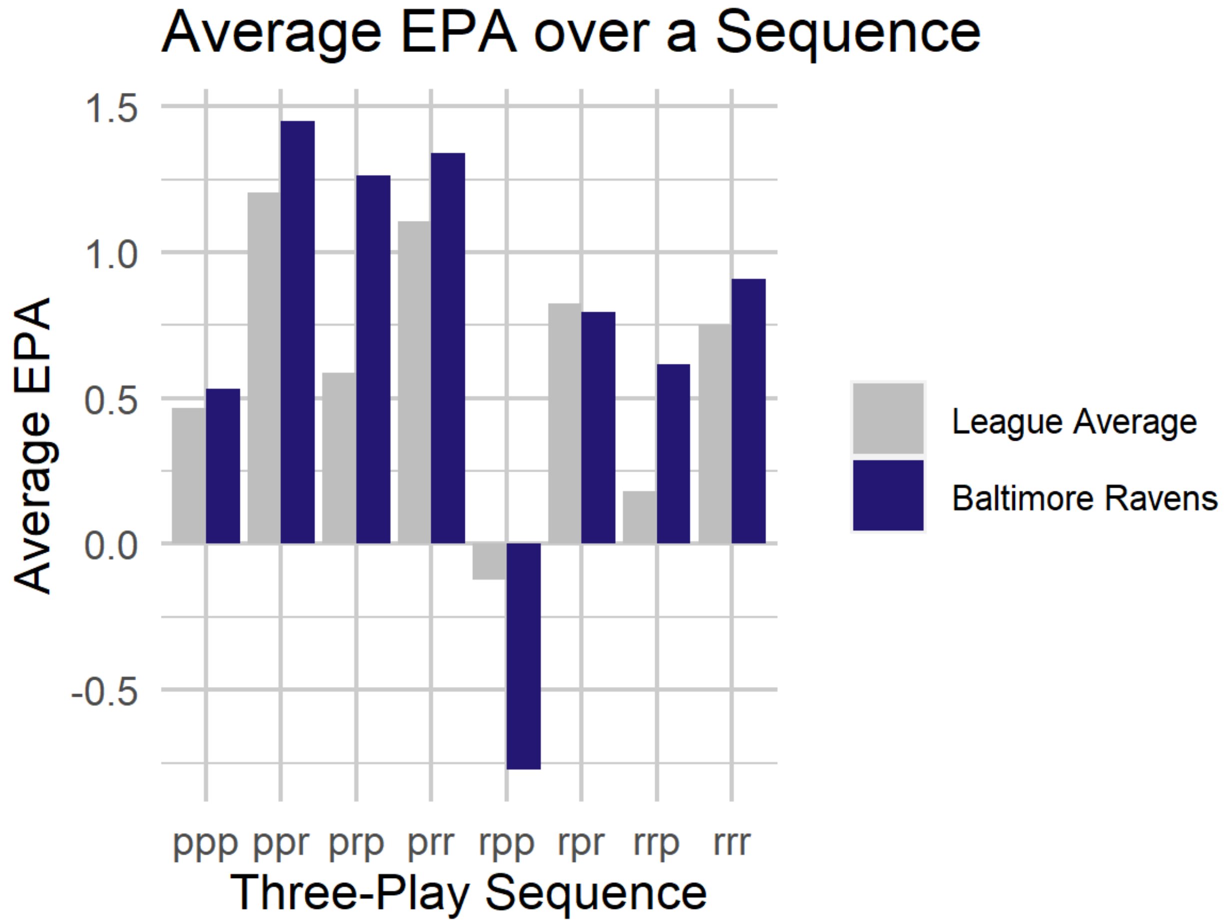
Play Sequencing



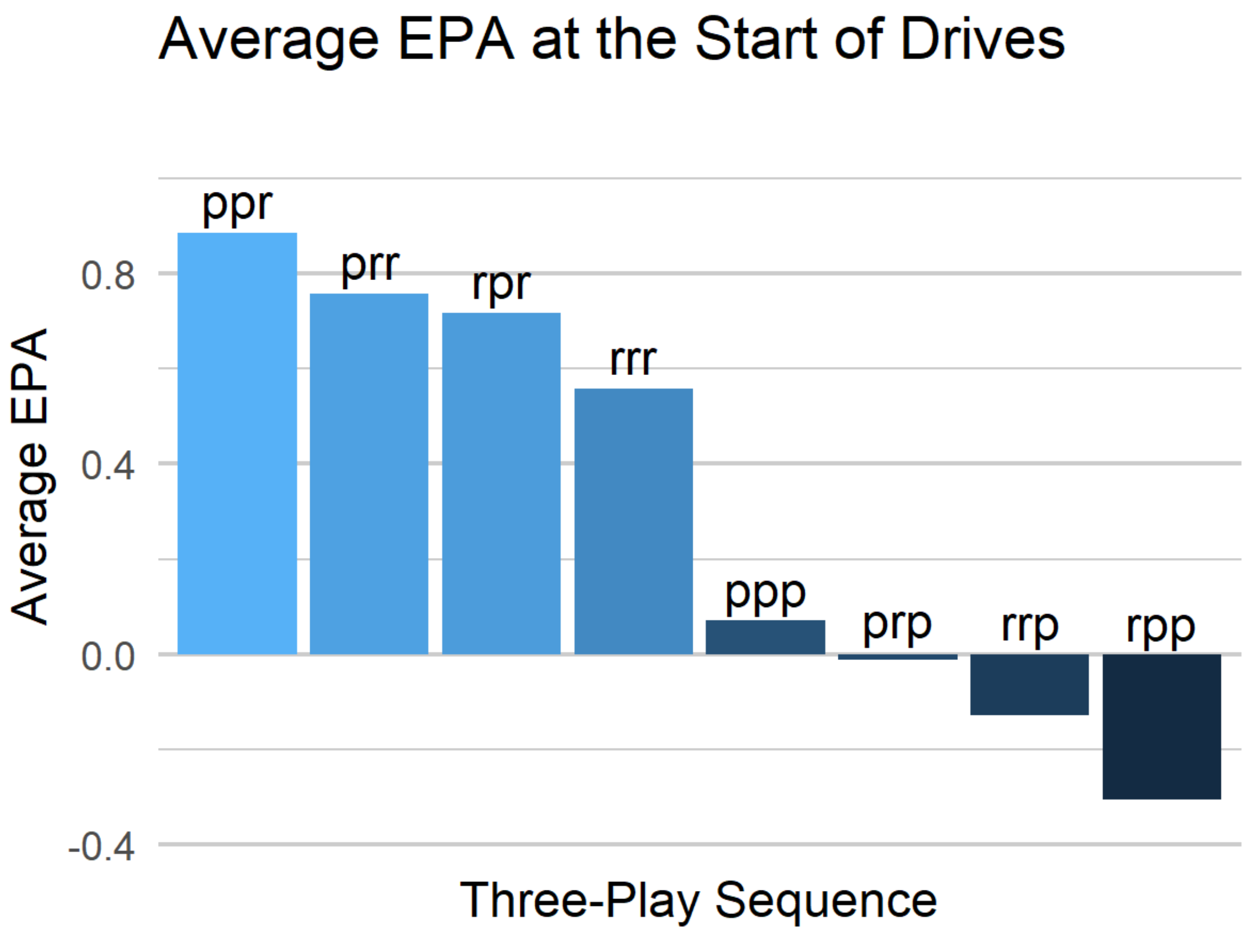


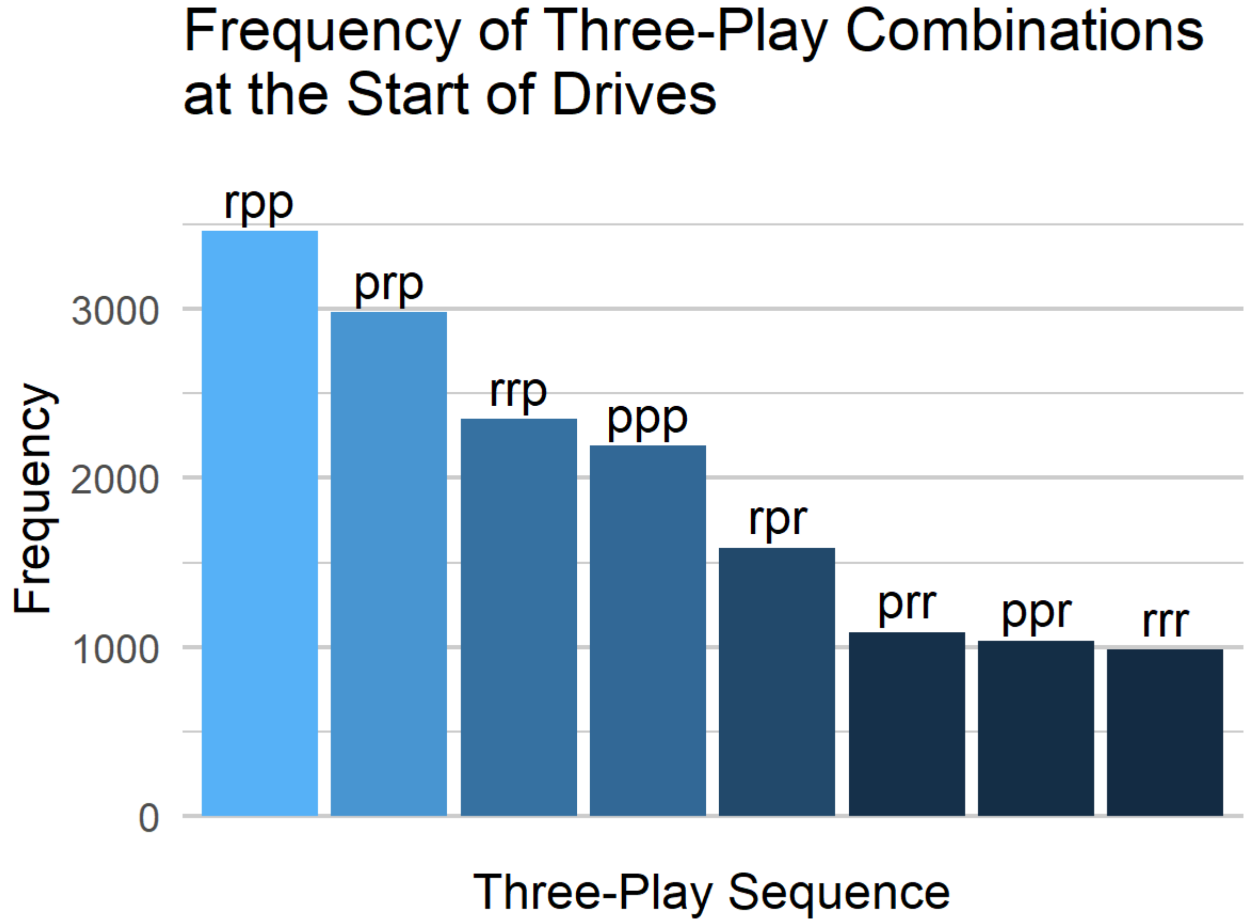
Sequence Efficiency



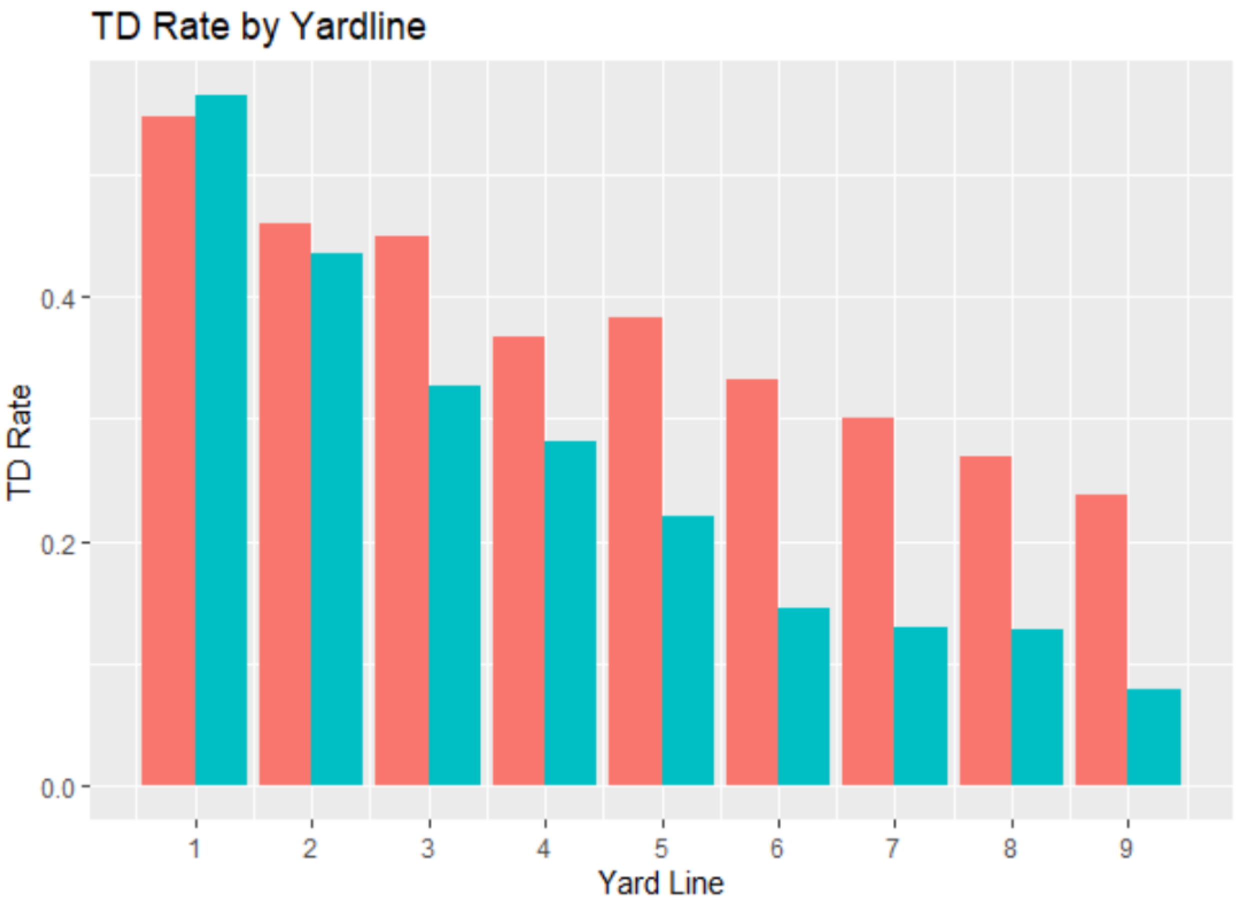


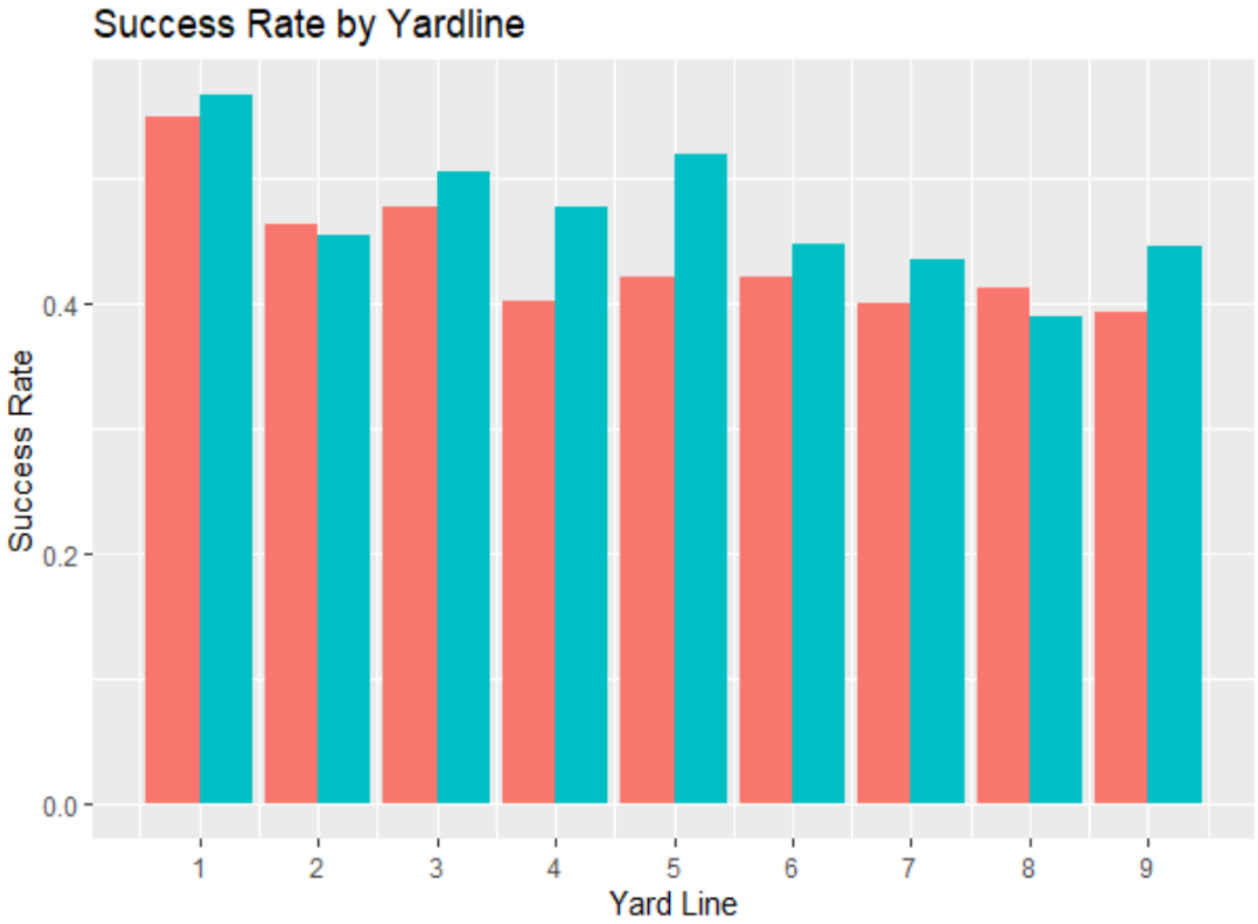
One More Consideration





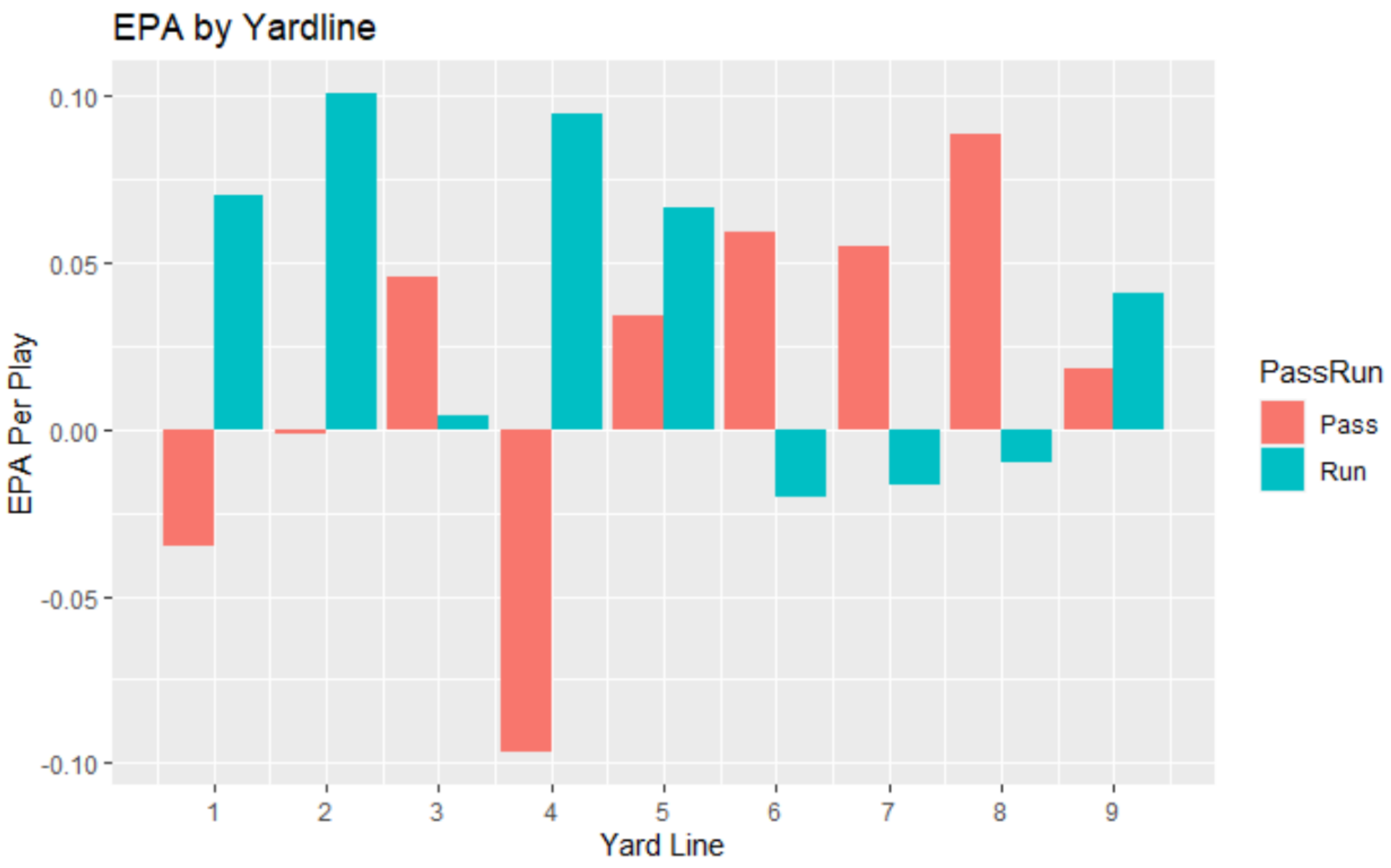
Applicability





Close Range Success

Close Range EPA



Potential Pitfalls

- Sample size is too small, especially for selected teams
  - Win probability may not be optimal
- Models are big picture
- Context is needed for play sequences
  - Hard to isolate the efficiency without game situation

Conclusion and Next Steps

- Model to find the optimal passing/play action percentages for each team
- Do what the defense won't expect
  - Equilibrium in each situation will hide what you're going to do
- Add game situation context to provide a more applicable analysis of sequence efficiency

Appendix

- Logistic Win Probability Model made on nflfastR data to apply to play action data
- AUC of 0.841
- Methodology from Stephen Hill
  - https://medium.com/@technocat79/building-a-basic-in-game-win-probability-model-for-the-nfl-54600e57fe1c

```
Call:
glm(formula = poswins ~ qtr + down + ydstogo + game_seconds_remaining +
    yardline_100 + score_differential, family = "binomial", data = wprob_data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.97542  -0.81215   0.08008   0.84599   2.99813

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.049e+00  5.738e-02  18.273  < 2e-16 ***
qtr2                   -1.934e-02  2.069e-02  -0.934  0.35010
qtr3                   -4.962e-02  3.426e-02  -1.448  0.14758
qtr4                   -1.168e-01  4.979e-02  -2.345  0.01903 *
down2                  -7.995e-02  1.155e-02  -6.920 4.53e-12 ***
down3                  -1.894e-01  1.346e-02 -14.074  < 2e-16 ***
down4                  -3.820e-01  1.677e-02 -22.778  < 2e-16 ***
ydstogo                -9.155e-03  1.231e-03  -7.439 1.01e-13 ***
game_seconds_remaining -4.776e-05  1.736e-05  -2.751  0.00595 **
yardline_100           -8.979e-03  1.994e-04 -45.039  < 2e-16 ***
score_differential      1.808e-01  7.538e-04 239.811  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 386238  on 278638  degrees of freedom
Residual deviance: 272120  on 278628  degrees of freedom
AIC: 272142

Number of Fisher Scoring iterations: 5
```

```
lm(formula = epa ~ wp2 + down + ydstogo + pass.y + prsh.y + cov.y,
    data = all_pass_80_89)

Residuals:
     Min       1Q   Median       3Q      Max
-12.3202  -0.7792  -0.3878   0.8982   4.9830

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.262468   0.198377   1.323 0.185839
wp2          0.187386   0.054839   3.417 0.000635 ***
down        -0.102521   0.020230  -5.068 4.09e-07 ***
ydstogo     -0.012597   0.005199  -2.423 0.015405 *
pass.y       0.007892   0.001410   5.596 2.25e-08 ***
prsh.y      -0.004554   0.002259  -2.016 0.043790 *
cov.y       -0.003040   0.001006  -3.020 0.002531 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.641 on 10478 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.009711,  Adjusted R-squared:  0.009144
F-statistic: 17.12 on 6 and 10478 DF,  p-value: < 2.2e-16
```

```
lm(formula = epa ~ wp + ydstogo + run.x + rblk.x + def.y + as.factor(prevplay),
    data = all_rush_0_24)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1977 -0.4015 -0.1213  0.3323  7.8718

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.2634907  0.1023943  -2.573 0.010083 *
wp                   -0.0674471  0.0252292  -2.673 0.007517 **
ydstogo              -0.0117313  0.0019876  -5.902 3.66e-09 ***
run.x                 0.0045861  0.0007382   6.213 5.34e-10 ***
rblk.x                0.0024324  0.0007455   3.263 0.001105 **
def.y                -0.0024467  0.0007016  -3.488 0.000489 ***
as.factor(prevplay)p -0.0172134  0.0179789  -0.957 0.338368
as.factor(prevplay)r  0.0455878  0.0182035   2.504 0.012278 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8855 on 15814 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.007334,  Adjusted R-squared:  0.006895
F-statistic: 16.69 on 7 and 15814 DF,  p-value: < 2.2e-16
```

APPENDIX

Sample output from EPA models