Ideal Interception Percentage/Pass Depth

Andrew Moss

June 23, 2020

Libraries

```
library(tidyverse)
library(nflfastR)
library(gridExtra)
library(ggrepel)
```

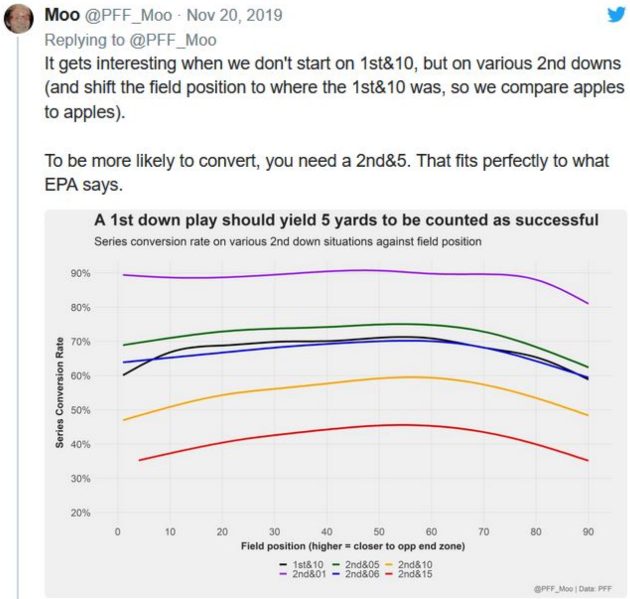
I used tidyverse to perform data frame manipulations and NFLfastR to scrape additional data to join on to the play-by play data in the zip folder. I also used gridExtra to make this markdown look better with side-by-side plots of mean and variance.

Research Question:

The question the analysis is examining is “Is there an optimal interception rate in the NFL?” However, I found a more appropriate phrasing of this question to be “Is there an optimal target depth for passes in the NFL environment.” From there, if I wanted to find an “optimal interception rate” I could find the rate that maps to the optimal target depth.

Data wrangling/cleaning/joining

To make my analysis easier, I modified passdepth and added the following columns: interception(binary), yards left, and success. I modified the passdepth column by lumping in throws that were thrown behind the line of scrimmage. For the sake of the analysis, there is no difference between throws at or behind the line of scrimmage; NFL teams also scheme these throws the same types of ways (as slip, bubble, and other screen varieties) The added INT column shows whether there was an interception on the play or not. Finally, the YardsLeft column shows how many yards were needed to get a first down on the play that follows the play in a given row. The YardsLeft and Down columns were used to generate the Success column, which shows if a team is ahead of the chains i.e. did that play keep or get the team on schedule to convert a first down more often than the originally given first and 10. Only second and <=5 and 3rd and 1 are more likely to be converted into a fresh set of downs than first and 10, per Timo Riske of PFF.



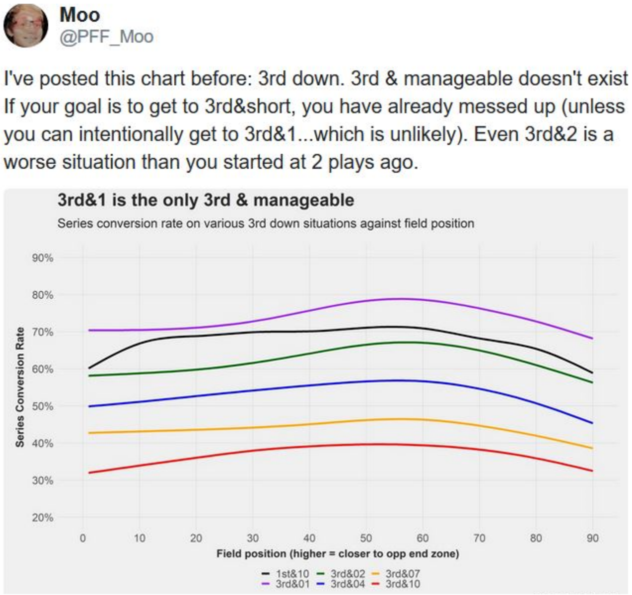
Moo @PFF_Moo · Nov 20, 2019

Replying to @PFF_Moo

It gets interesting when we don't start on 1st&10, but on various 2nd downs (and shift the field position to where the 1st&10 was, so we compare apples to apples).

To be more likely to convert, you need a 2nd&5. That fits perfectly to what EPA says.

A 1st down play should yield 5 yards to be counted as successful
Series conversion rate on various 2nd down situations against field position
Series Conversion Rate
90%
80%
70%
60%
50%
40%
30%
20%
0
10
20
30
40
50
60
70
80
90
Field position (higher = closer to opp end zone)
1st&10
2nd&05
2nd&10
2nd&01
2nd&06
2nd&15
@PFF_Moo | Data: PFF



Moo
@PFF_Moo

I've posted this chart before: 3rd down. 3rd & manageable doesn't exist If your goal is to get to 3rd&short, you have already messed up (unless you can intentionally get to 3rd&1...which is unlikely). Even 3rd&2 is a worse situation than you started at 2 plays ago.

3rd&1 is the only 3rd & manageable
Series conversion rate on various 3rd down situations against field position
Series Conversion Rate
90%
80%
70%
60%
50%
40%
30%
20%
0
10
20
30
40
50
60
70
80
90
Field position (higher = closer to opp end zone)
1st&10
3rd&02
3rd&07
3rd&01
3rd&04
3rd&10

```
pbp = read.csv("C:\\Users\\Andrew Moss\\Documents\\PackersPBP.csv")
pbp$passdepth = ifelse(pbp$passdepth<0,0,pbp$passdepth)
pbp$yardsgained = ifelse(is.na(pbp$yardsgained),0,pbp$yardsgained)

pbp = pbp %>%
  mutate(pbp,INT = ifelse(passresult=="INTERCEPTION",1,0)) %>%
  mutate(pbp,YardsLeft = ifelse(distance-yardsgained>0,distance-yardsgained,0)) %>%
  mutate(pbp, Success = ifelse(down==1,ifelse(YardsLeft<=5,1,0),
    ifelse(down==2,ifelse(YardsLeft<=1,1,0),
      ifelse(down==3|down==4,ifelse(YardsLeft==0,1,0),0))))
```

I knew that I was going to want to use Expected Points Added (EPA) as a part of my analysis. After seeing that the play IDs from NFLfastR matched the play IDs from the pbp data that I was given, I knew I wanted to join columns from the NFLfastR data frame. The function of this is to get each play's EPA as well as replacing the QB's Team/Name/Position abbreviation with their actual name.

```
games_2019 = readRDS(url(
  'https://raw.githubusercontent.com/guga31bb/nflfastR-data/master/data/play_by_play_2019.rds'))
games_2019 = games_2019 %>% rename(playid = play_id)
games_2019 = games_2019[c(1,2,320,322)]
```

After I used NFLFastR to scrape the additional data, I prepared the game results data from the zip folder by changing the team abbreviations from the sportsreference.com abbreviations to ones that matched NFLFastR. I then generated game codes in the NFLfastR format (year_week_awayteam_hometeam). From there, I used the original gamecode that came with the files to get the NFLfastR gamecodes appended onto plays from the appropriate game.Finally, I joined the NFLfastR additional data on to my main "pbp" dataframe on gameID and play_id.

```
results = read.csv("C:\\Users\\Andrew Moss\\Documents\\PackersGameResults.csv")
results = results[1:2]
results$game_id = as.character(results$game_id)
pbp = left_join(pbp,results,"gamekey")
pbp = left_join(pbp,games_2019,c("game_id","playid"))
pbp = na.omit(pbp)
```

I now had all of the data I wanted to use in my analysis joined. However, there were 16 NAs for EPA and 73

NAs for passer name. After further examination, I found that the NAs for EPA were all end of half-type situation where the team was unlikely to score and thus EP before the play was essentially 0 and the EP after the play was also 0 because the play triggered halftime or the end of the game. I could have imputed these NAs with 0, but the context of these plays aren't really helpful in my analysis of the generally most optimal target depth; as a result I omitted them. The 73 NAs for passer name were all spikes so I omitted those too.

In order to find ideal target depths, I needed to filter the dataset to passes that the scorekeeper is sure how deep the intended target was. For example, it is not always clear who the intended receiver is on a batted pass or "hit as threw" (quickly). To isolate passes that got to the intended depth of target, I only had AllPasses include, completions, incompletions, and interceptions.

```
unique(pbp$passresult)
```

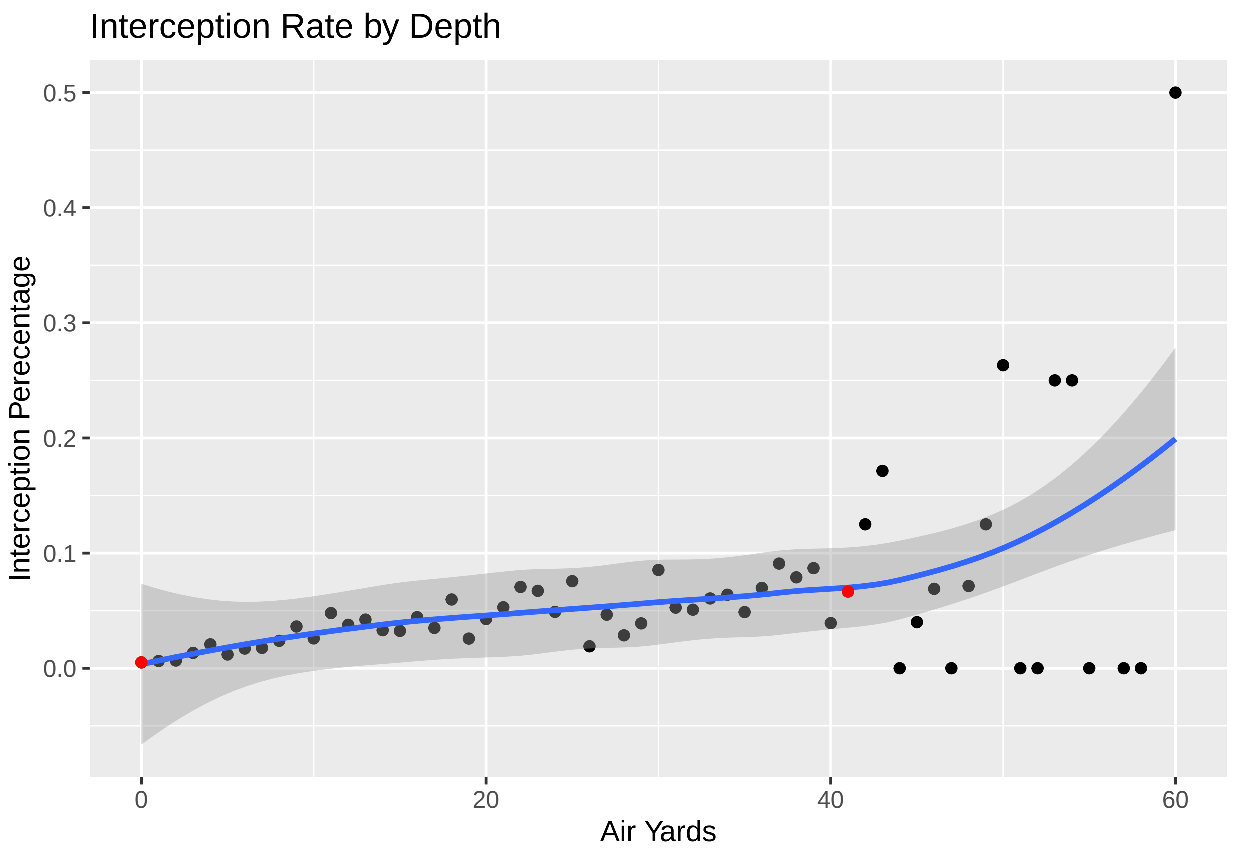
```
## [1] COMPLETE     SACK         BATTED PASS  RUN          INCOMPLETE  
## [6] HIT AS THREW THROWN AWAY  INTERCEPTION LATERAL     
## 10 Levels: BATTED PASS COMPLETE HIT AS THREW INCOMPLETE ... THROWN AWAY
```

```
AllPasses = filter(pbp, passresult =="COMPLETE" |
                     passresult == "INCOMPLETE" |
                     passresult == "INTERCEPTION")
AllPasses$passresult = ifelse(AllPasses$passresult =="COMPLETE", 1 , 0)
```

Analysis

I used interception percentage as a starting point for my analysis. I knew that interception percentage at each target depth alone would not be enough to determine the optimal target depth (and thus interception percentage), but it allowed me to visualize how much riskier intermediate or deep passes were than shorter passes.

```
INTs = AllPasses %>% group_by(passdepth) %>% summarise(INTPercent = mean(INT))
ggplot(INTs, aes(x=passdepth, y=INTPercent)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Air Yards", y = "Interception Perecentage") +
  ggtitle("Interception Rate by Depth") +
  geom_point(data=INTs[1,], colour="red")+
  geom_point(data=INTs[42,], colour="red")
```



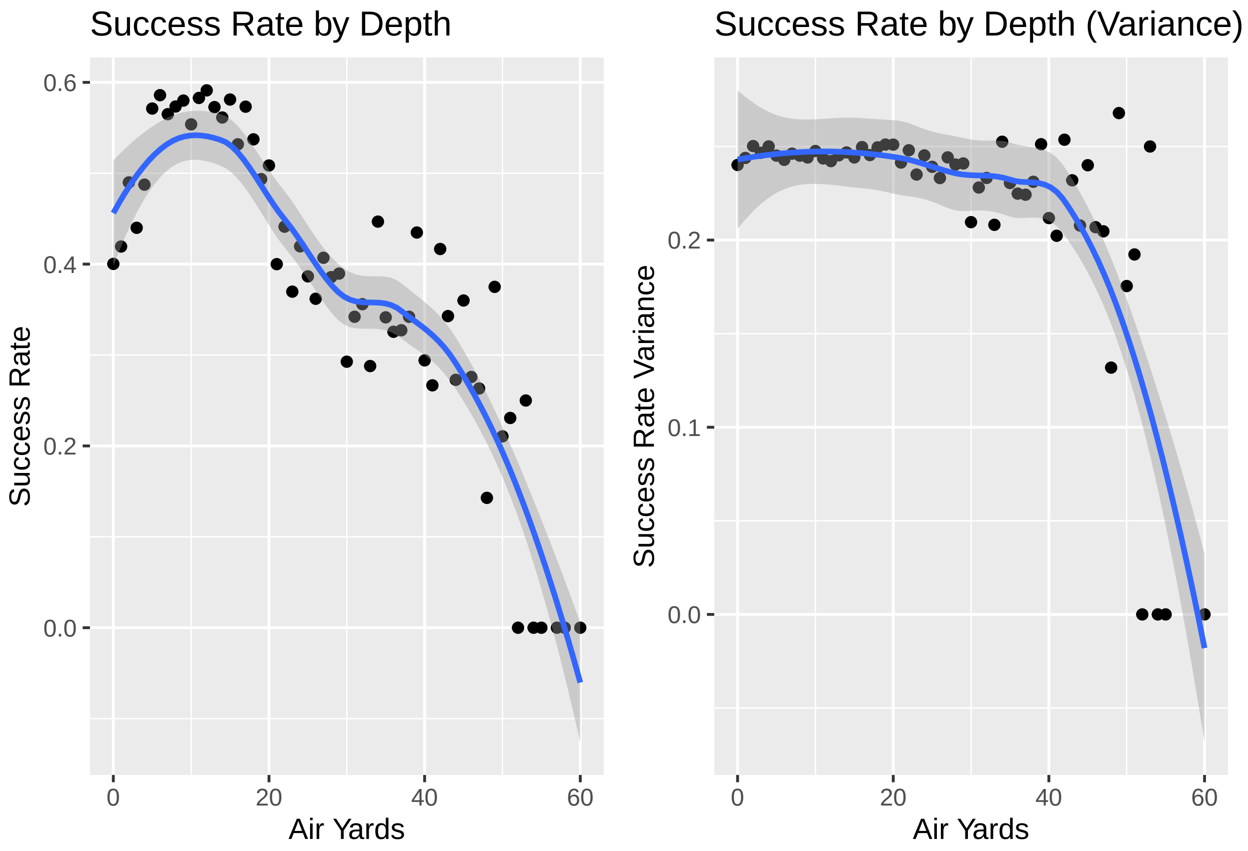
A roughly linear trend can be seen from throws at or behind the line of scrimmage to 41 air yard throws (in red). This encompasses almost 98.5% of passes in the 2019 season. This trend suggests that each additional yard a QB throws beyond the line of scrimmage comes with a .1484 of a percent increase in interception rate.

However, living in fear of interceptions is no way to conduct an NFL offense. To get a more detailed picture of the effectiveness of passes at different depths, I examined success rate and EPA as a function of air yards. While I conjectured that there would be some longer passes that would be more effective than shorter passes, I thought that increase in expectation would also come with an increase in variance. As a result I summarized and plotted both the mean and variance of EPA and Success rate by pass depth. EPA and success rate are both measures of a team moving closer toward scoring, and had a correlation of nearly .7 over the play-by-play data set. However, because Success is binary, the metric doesn't see a difference between an 8 yard completion on first down or an 80 yard completion. It sees both as a success. As a result, I expected Success rate to skew toward the most easily completable (shortest) passes within a reasonable distance of a successful play. EPA, on the other hand, recognizes that an 80 yard completion is almost always better for a team than an 8 yard completion. It also penalizes interceptions, which success rate sees as the same as incompletion, an unsuccessful play. Because deep passes are intercepted at a higher rate, but also create the most extremely successful plays per EPA, I expected the variance of EPA to increase as pass depth increased.

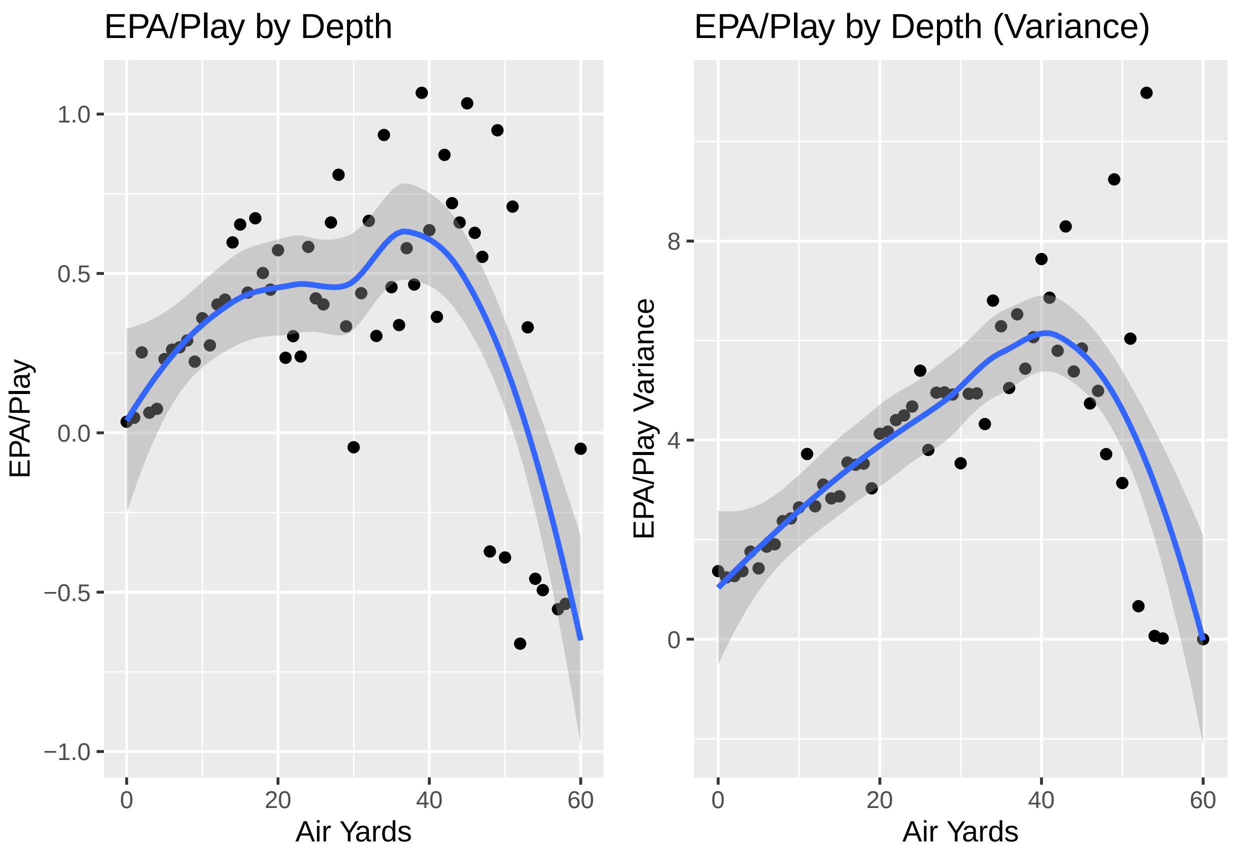
```
cor(pbp$qb_epa,pbp$Success)
```

```
## [1] 0.7027027
```

```
AllSuccess = AllPasses %>% group_by(passdepth) %>%
  summarise(mean = mean(Success), Variance = var(Success))

SuccessPlotEV = ggplot(AllSuccess, aes(x=passdepth,y=mean)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Air Yards", y = "Success Rate") +
  ggtitle("Success Rate by Depth")

SuccessPlotVar = ggplot(AllSuccess, aes(x=passdepth,y=Variance)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Air Yards", y = "Success Rate Variance") +
  ggtitle("Success Rate by Depth (Variance)")

grid.arrange(SuccessPlotEV,SuccessPlotVar, ncol=2)
```



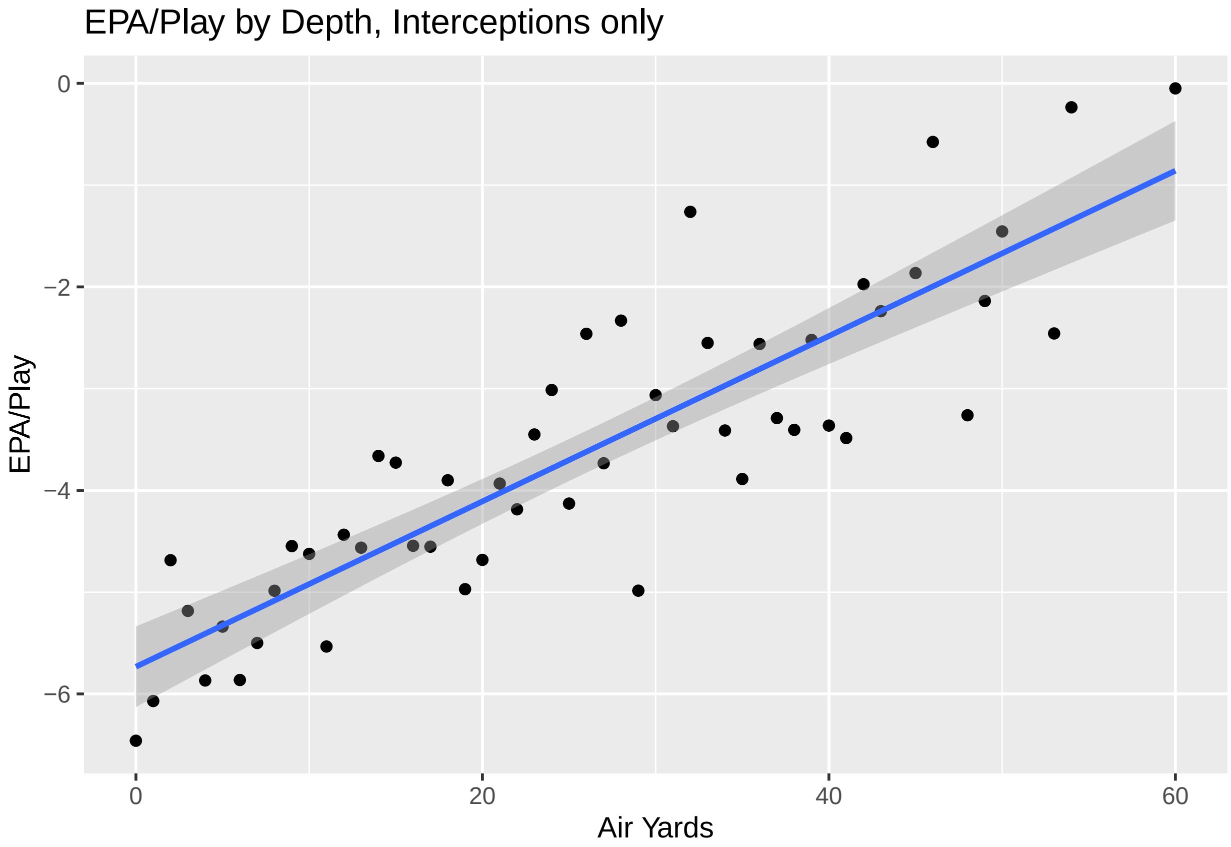
```
EPApPlayDepth = AllPasses %>% group_by(passdepth) %>%
  summarise(EPA = mean(qb_epa),Variance = var(qb_epa))

EPAPlotEV = ggplot(EPApPlayDepth, aes(x=passdepth,y=EPA)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(x = "Air Yards", y = "EPA/Play") +
  ggtitle("EPA/Play by Depth")

EPAPlotVar = ggplot(EPApPlayDepth, aes(x=passdepth,y=Variance)) +
  geom_point() +
  geom_smooth(method = "loess") +
  labs(x = "Air Yards", y = "EPA/Play Variance") + ggtitle("EPA/Play by Depth (Variance)")

grid.arrange(EPAPlotEV,EPAPlotVar, ncol=2)
```



These plots followed my expectation. Success rate was highest on passes 5 to 17 yards downfield, then dropped off. Variance was constant from 0 to 40 air yards and then there isn't a big enough sample size to generate a meaningful conclusion past 41 yards or so. EPA was a little more bi-modal, with a roughly constant EPA/Play from 13 to 28 air yards. There is a second, higher peak around 40 yards, but it is not as meaningful given the smaller sample size. The variance increases linearly from balls thrown 0 to 40 yards downfield, similar to the increase in interception rate. This reminds us that while throwing the ball further down the field increases the expectation of EPA, there is a wider range of common outcomes the further down the field you go.

```
AllPicks = filter(pbp, passresult == "INTERCEPTION")
AllPicks$down = as.factor(AllPicks$down)

EPAINTmodel = lm(qb_epa~passdepth+down+distance+fieldposition+down*distance,
                 data = AllPicks)
summary(EPAINTmodel)
```

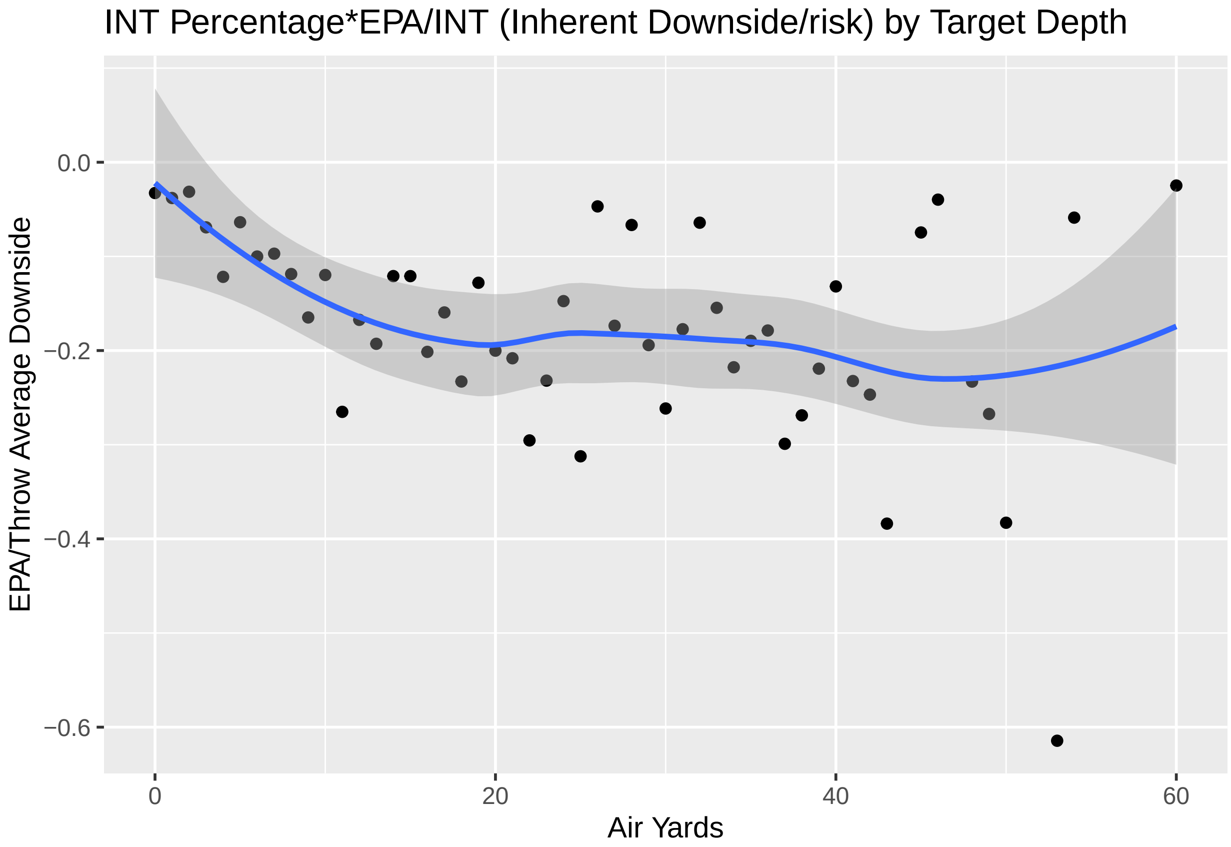
```
AllPicks = AllPicks %>%
  group_by(passdepth) %>%
  summarise(EPA = mean(qb_epa))

ggplot(AllPicks, aes(x=passdepth, y=EPA)) +
  geom_point() + geom_smooth(method = "lm") +
  labs(x = "Air Yards", y = "EPA/Play") +
  ggtitle("EPA/Play by Depth, Interceptions only")
```



A final examination I wanted to do in order to find the optimal target depth or interception percentage is to generally assess the risk that interceptions pose at different depths. Even though they happen less, interceptions on shorter passes on average, are more catastrophic than interceptions on deeper passes. This likely comes as a result of shorter pass-interceptions resulting from jumped routes, which often lead to long or scoring interception returns. Longer passes also are intercepted at a less field position-advantageous point, which makes the opponent's expected points to start out the new drive lower than an interception that puts them better field position.

After seeing that there was a linear relationship between Air Yards and EPA of an interception, I built a basic linear model with down, distance, the interaction between the two, field position, and air yards. The summary statistics of the model indicate that each additional yard down the field the ball was thrown mitigated .086691 of the negative EPA from the average interception.

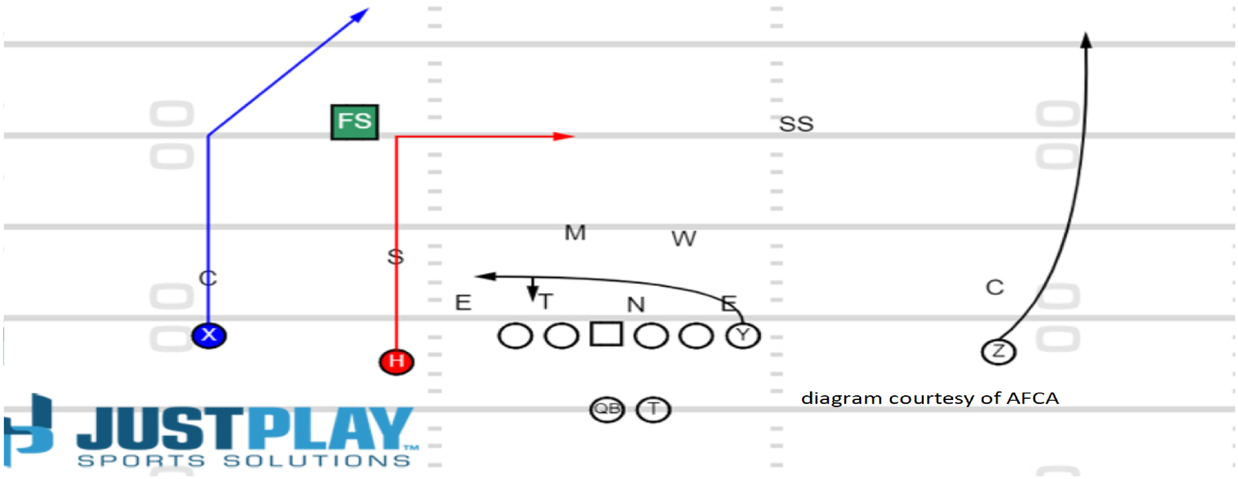
```
AllPicks = left_join(AllPicks,INTs,"passdepth")
AllPicks$TurnoverEPAThrow = AllPicks$EPA*AllPicks$INTPercent
ggplot(AllPicks, aes(x=passdepth, y=TurnoverEPAThrow)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "Air Yards", y = "EPA/Throw Average Downside") +
  ggtitle("INT Percentage*EPA/INT (Inherent Downside/risk) by Target Depth ")
```



The next step of this analysis was to see whether the decrease in interception rate is enough to mitigate the EPA downside of interceptions on shorter passes. So, I appended the interception rate by Depth next to each interception depth's respective EPA. I then multiplied the two figures for each row to get a "risk factor" or EPA/interception adjusted for how frequently it occurs. Throws under 10 yards were intercepted so infrequently at the NFL level in 2019 that there is still less downside inherent in a short throw than there is in a longer throw. So if the goal of optimal passing was to minimize potential downside (it isn't), the recommendation would be for route concepts that create shorter throws for the QB (less than 10 air yards)

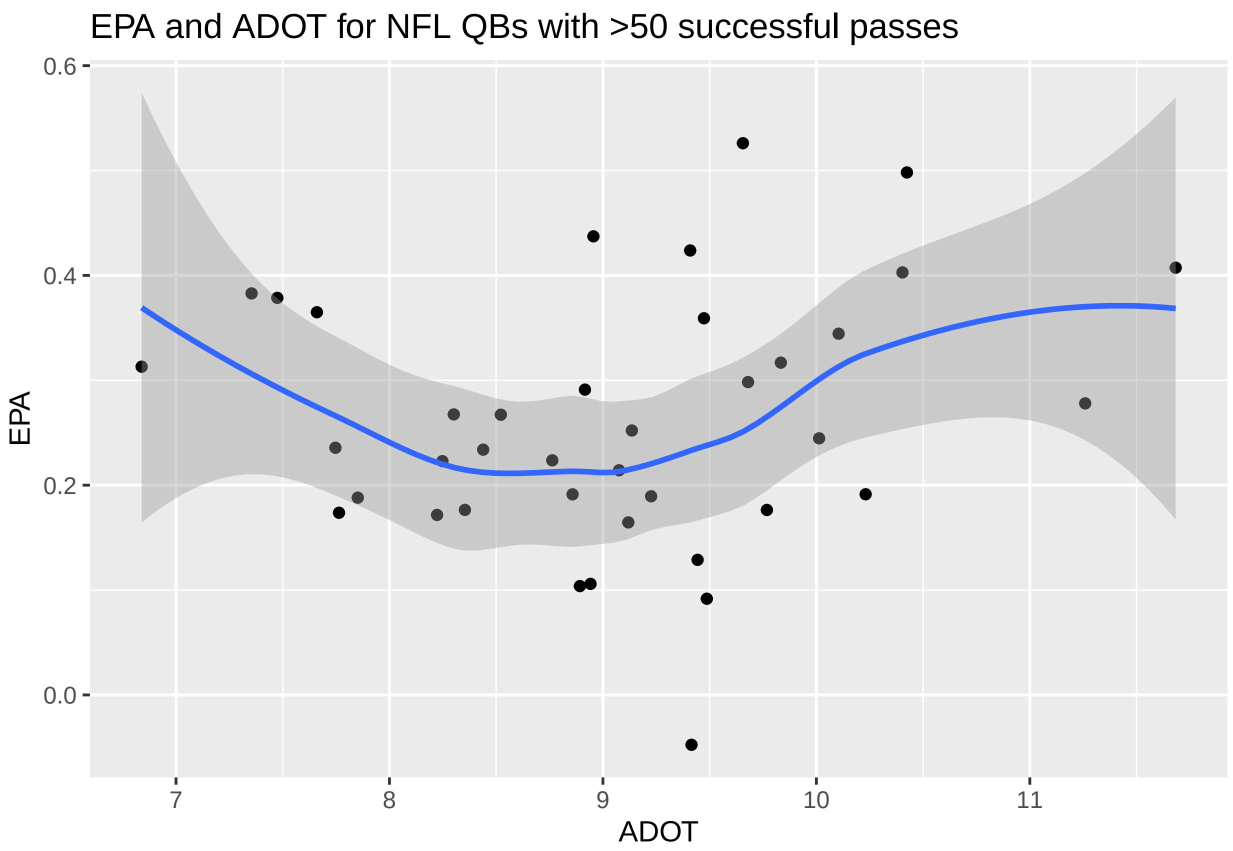
Conclusion and Future Work

The ultimate answer to the question of "is there an optimal interception rate (or target depth)" is "it depends." While EPA is used to maximize points scored on a given drive, it can have limitations in context. For example, if a team is up by 4 on their own side of the field in the last 4 minutes of a game, maximizing the opportunity to move the chains and minimizing both variance and inherent downside is more important than trying to maximize points on that drive. However, in a neutral game script situation, there isn't a huge difference between maximizing EPA and maximizing win probability. In today's NFL, passes of 12-15 yards maximize EPA without letting negative variance control a game. They are also completed enough to generally keep the chains moving (min success rate between throws 12-15 yards is .56). There is a reason that the Mills concept has been a staple in playbooks since Spurrier and Wuerrfel used it in the 1990s; it was a route combination that the quarterback was comfortable with that puts him in position to succeed with intermediate throws.



Another factor that can be explored on the team level that influences ideal target depth is quarterback strengths. For example, the plot below shows that there are top QBs per EPA on both the high side and low side of average depth of target (ADOT). When focusing on a single QB, it would be helpful to use more than one season of game data whi;e tracking changes or improvement over time. Practice data would also be helpful. Other data that would assist further inquiry, other than spatial tracking data which opens up another whole set of possibilities, includes the coverage the defense was in on a given play and the concept that the offense was running. Along a similar line of thought, including personnel groupings in the play-by-play data would further help identify mismatches that lead to exploitative opportunities for the offense in various down and distance situations. This additional information would allow for more situational analysis and the recommendation of certain concepts versus a coverage in a given down-and-distance situation.

```
ADOTbyQB = AllPasses %>%
  group_by(name) %>%
  summarise(ADOT = mean(passdepth),
            EPA = mean(qb_epa), Success = sum(Success))

ADOTbyQB = filter(ADOTbyQB,Success>50)

ggplot(ADOTbyQB, aes(x=ADOT, y=EPA, label = name)) +
  geom_point() + geom_smooth(method = "loess") +
  labs(x = "ADOT", y = "EPA") +
  ggtitle("EPA and ADOT for NFL QBs with >50 successful passes")
```


EPA and ADOT for NFL QBs with >50 successful passes
EPA
0.6
0.4
0.2
0.0
ADOT
7
8
9
10
11