Calculating the Odds of Pass Completion in the National Football League

2023-12-16

Introduction

Motivation

When thinking about the sport of football as a whole, the ability to complete a pass is one of the most important skills a team can possess. The game of football has changed drastically over the years and teams are throwing the ball more than ever. The days of “ground and pound” football are over, meaning teams no longer rely on running backs and defense to win games. Now, teams rely on throwing the ball, stretching the defense, and being versatile. Because of this, the quarterback position is the most important position on the field in today’s game. Every team in the National Football League (NFL), which is the highest level of competitive football, will go to extraordinary measures in order to find their “franchise quarterback”. These teams take massive risks and mortgage their future on finding this player. If a team is lucky enough to find this special quarterback, then they will pay a premium price to keep him on their team for a long period of time.

In the 2017 NFL draft, the Kansas City Chiefs traded three valuable future draft picks in order to select a 22 year old quarterback named Patrick Mahomes. By trading these future picks, Kansas City gave up a large chunk of their future in the hopes that Mahomes would be their best player and the face of their franchise for years to come. Mahomes did just this, winning the league Most Valuable Player award in his second season and then leading Kansas City to a Super Bowl victory the next year. Following these feats, Kansas City refused to let Mahomes play for any other team, so they offered him nearly half of a billion dollars to be their quarterback for the next decade.

This is not the only example of a team giving out hundreds of millions of dollars to their quarterback, as the 15 largest contracts in the all of football belong to quarterbacks. Clearly, there is an incredible amount of money in this sport, so much so that the NFL is the most profitable sports league in America. According to Statista, the NFL generated nearly 19 billion dollars in revenue in 2022, which is almost double that of the National Basketball Association and Major League Baseball, and this number continues to grow every year. In summary, football is a massive industry, so this project is our attempt to provide data-backed opinions and takeaways surrounding the most valuable aspect of one of the most valuable modes of entertainment today - completing passes in an NFL game.

Research Question

Simply put, our research is predicated on determining the odds that a pass is completed in the NFL, based on data from the 2022 season. We recognize that there are a multitude of variables within the game of football that have an impact on whether a pass is completed, many of which we attempt to capture in our modeling. For example, certain situations such as the time remaining and the score differential dictate the plays that are called, the tempo of the game, and overall scheme. Additionally, certain characteristics of a pass can make the odds of completion more difficult, such as how far and to what area of the field the pass is thrown.

Even still, we cannot account for the baseline talent and decision making ability of each individual quarterback in our dataset by simply using a logistic regression model. In the same way, the talent of the surrounding teammates, especially the lineman and wide receivers, understandably has an impact on the odds of completion. With a logistic model, the assumption would be that because we are observing the highest level of football, each quarterback and his teammates must be equally talented to play professionally, and therefore the odds of completing a pass are solely based on the football-specific contexts we include as predictor variables. We believe this is an overgeneralization. In fact, the baseline talent and decision making ability of each individual quarterback can and should be thought of as correlated with the odds of completing a pass. Because of this, as well as the fact that our data contains repeated measurement of the same quarterback (passes thrown throughout a season), the following investigation is from a “Correlated Data” framework and utilizes Longitudinal Data techniques.

Data

Context

Our data comes from nflfastR, a dataset built into R that has play-by-play data for every game from the last eight NFL seasons. We chose to focus on 2022 for our project since it is the most recent, fully completed season. Our dataset has 372 variables for every single play that happened in the 2022 season - which was 50,147 plays. These variables range from the type of play, the success of the play, the score at that time, where the play took place on the field, where the game was played, the fantasy football implications of a play, and even the weather of where the game took place. One could pretty much attempt any project imaginable with this data, so we just had to specify our needs. Based on our own knowledge of football as both fans and college football players, we have clear expectations of which variables alter the probability of a pass being completed. Generally speaking, this includes game situation and pass difficulty.

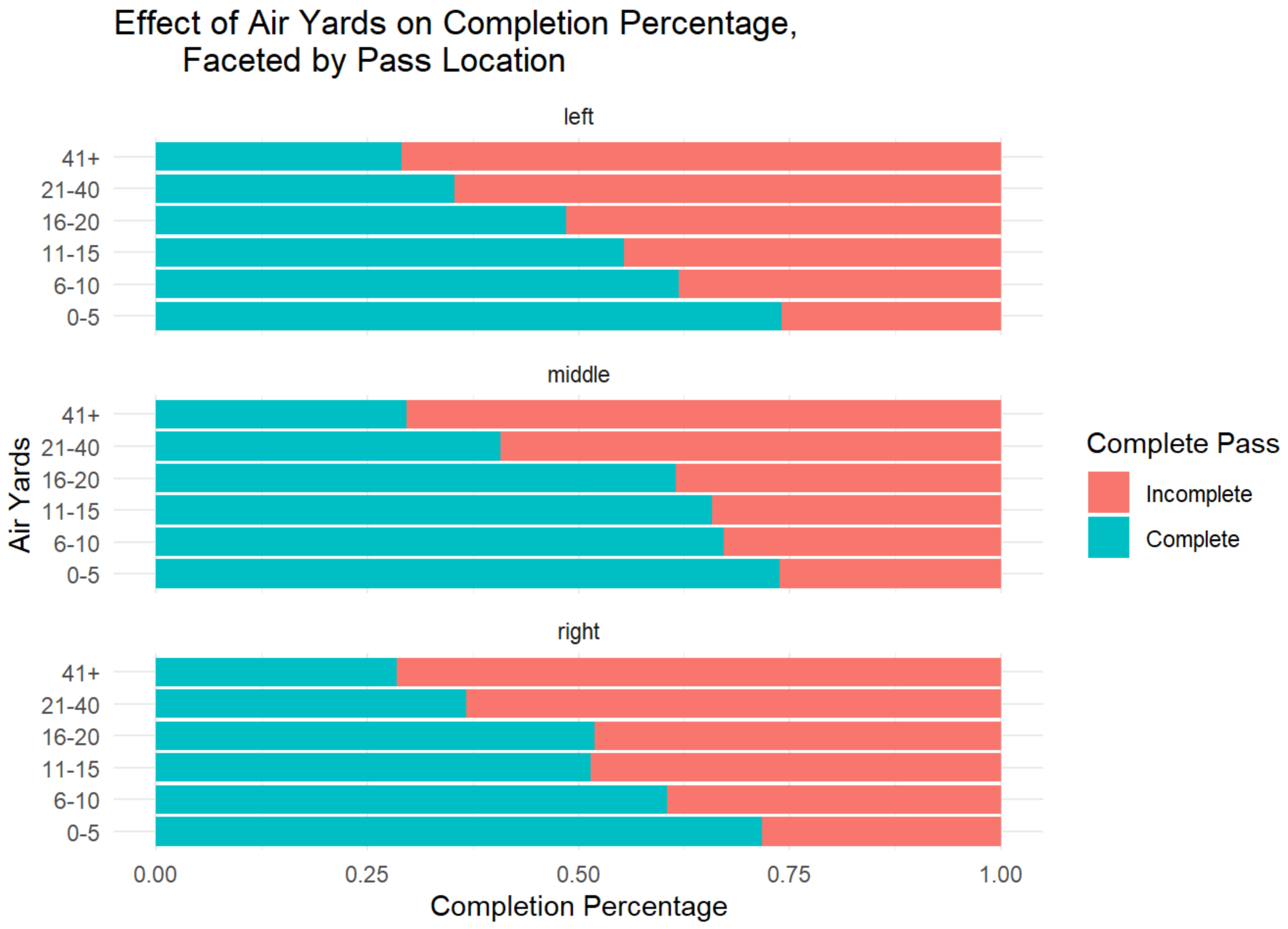
Filtering

When looking at the odds of a pass being completed in the NFL, there are certain plays that will not matter for our search. We decided to filter our data in a number of different ways. First, we filtered in order to make sure that we were only looking at pass plays, leaving out all rushing and kicking plays. We also wanted to make sure that the pass went beyond the line of scrimmage, or forward in other words, so that screen plays were not in our data set. A screen play is just a quick pass to a running back or wide receiver very close to the quarterback and should be completed nearly every single time. We also only wanted to look at passing attempts in regulation game time, so we filtered out any overtime plays due to the rules changes associated with this added playing time. The final filtering we did was to get rid of passing attempts from non-starting quarterbacks. We filtered out any player who had less than 190 pass attempts on the season, leaving us with 33 players. This gave us the starting quarterback for 31 of the NFL teams and the two quarterbacks who split time for the final team. Our final data set consisted of 12,753 passing plays.

Variables

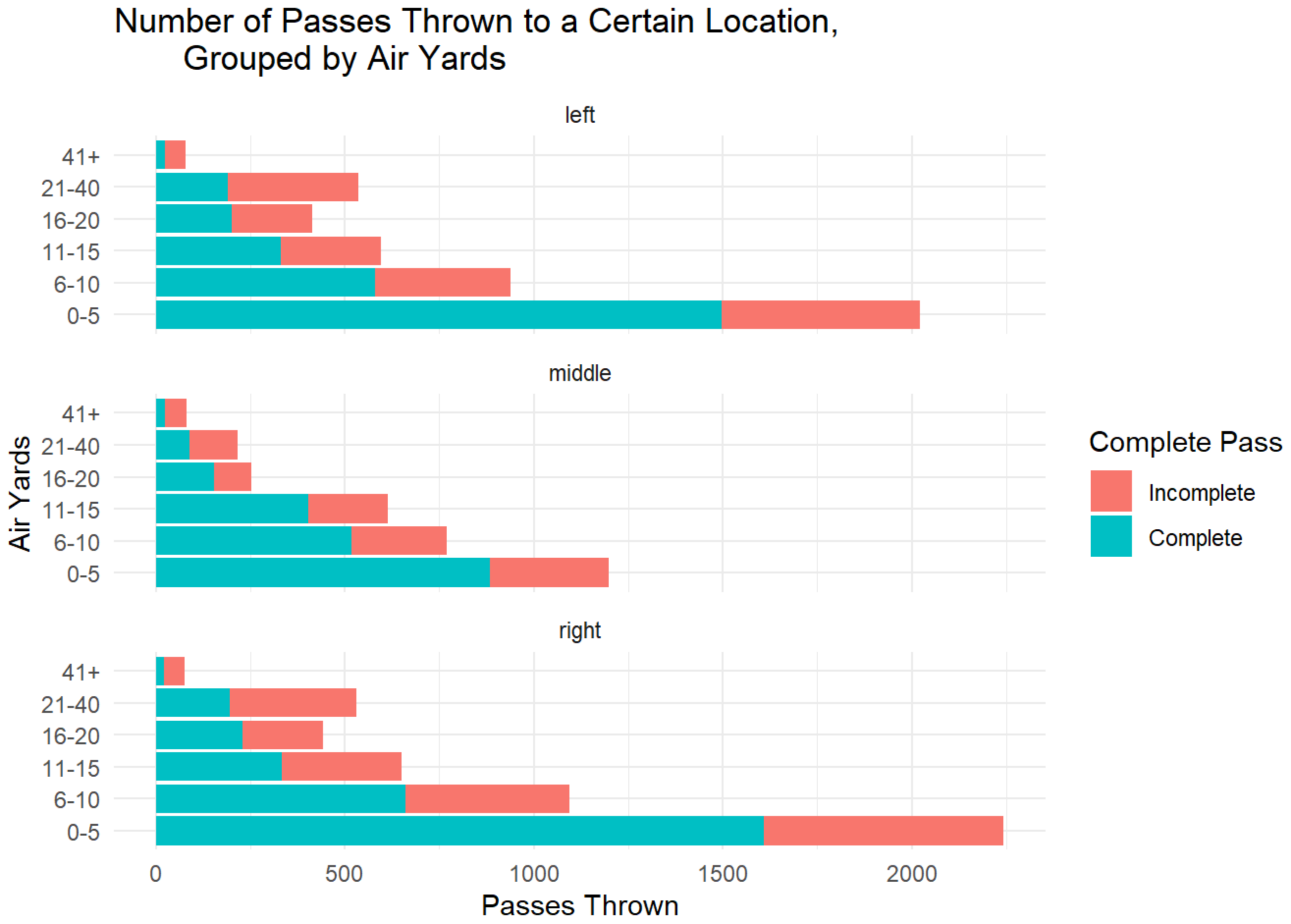
Pass Location

One variable that we expect to impact the odds of completion is whether the ball is thrown to the middle of the field or not. We can see from the graph below that passes thrown to the middle of the field tend to be completed at a higher rate. Contextually, this makes sense. We know that passes thrown to the middle of the field are typically shorter and considered safer throws. We also know that the defensive players in the middle of the field have responsibilities to stop the run and compete against offensive lineman, so they cannot fully focus on defending the pass. The defensive players on the outside parts of the field are exclusively focused on not letting the offense complete passes. With this being said, we created a binary variable called *not_mof* which equals 1 for passes thrown to the outer parts of the field and 0 for passes thrown to the middle of the field. We can see that left and right have very similar tendencies so we grouped those two together.



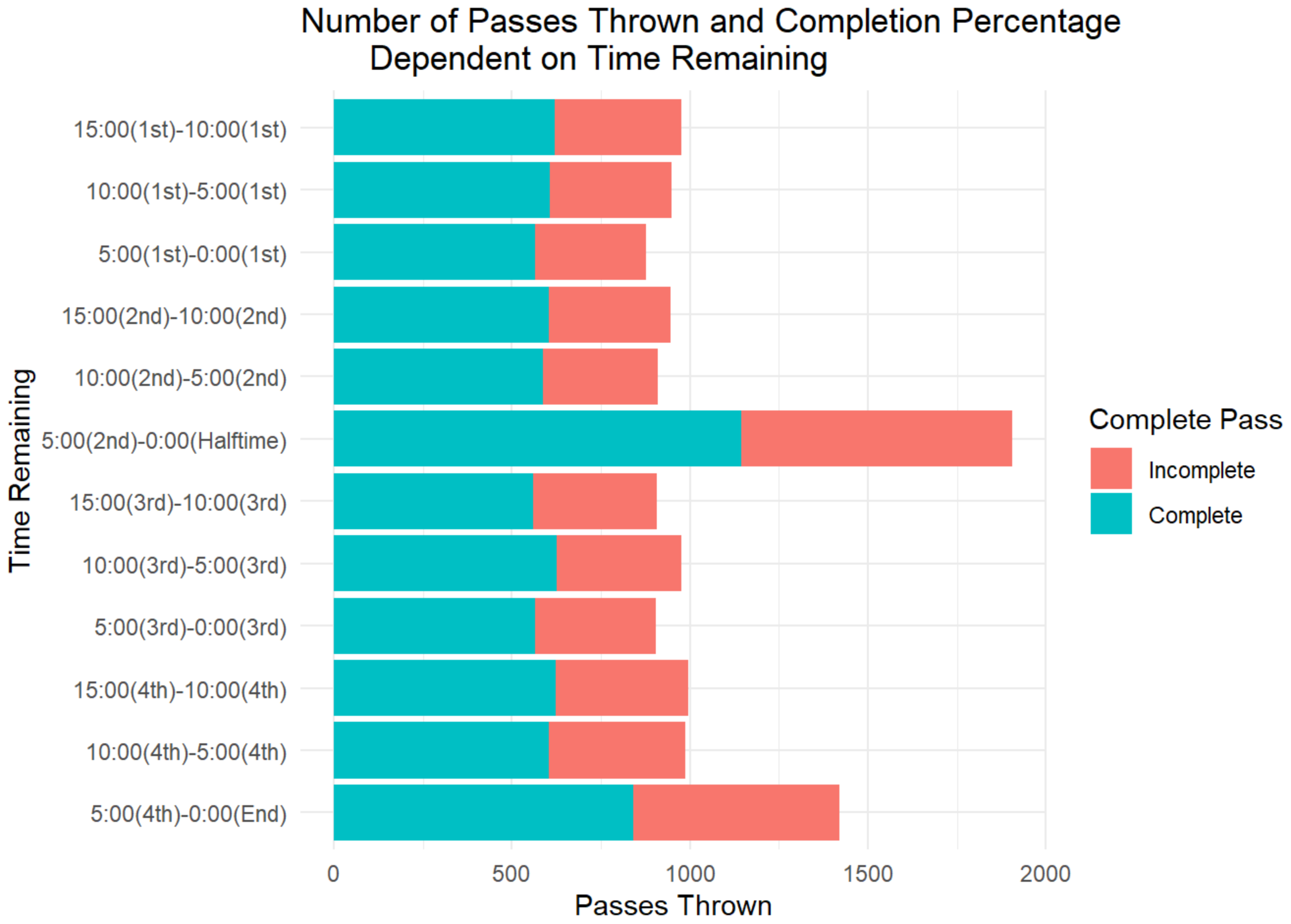
Air Yards

Furthermore, we can also see in the graph above that air yards contributed to completion percentage, so we created our own variable to account for this. Based on our exploratory data analysis and football knowledge, we concluded that a one unit increase in air yards is insignificant and contextually irrelevant, so we chose to account for the distance in yards a ball is thrown categorically. Once again with a combination of data analysis and football knowledge, we formed six categories of pass distances that make up our variable *air_yards_cat*. Passes that are 0-5 yards down the field are known as “check downs” and should be completed at a high level. We then increased in increments of 5 yards in order to show “chunk plays” and “intermediate passes” which are passes designed to be completed for more yards. Certain plays get schemed for these different levels of the field, so we wanted to separate them until 20 yards. Once we get to 20 yards, anything from 20-40 yards down the field is considered a “deep pass” and these are plays designed for teams to make a big, explosive play. Passes thrown 40 or more yards down the field are usually a “hail mary”, or a ball thrown out of desperation with a low likelihood of completion. From the visualization below, we can see that there are vastly different amounts of passes thrown to certain areas of the field. Quarterbacks generally want to complete as many passes as they can, so most of the time, they will throw the ball shorter distances in the middle of the field.



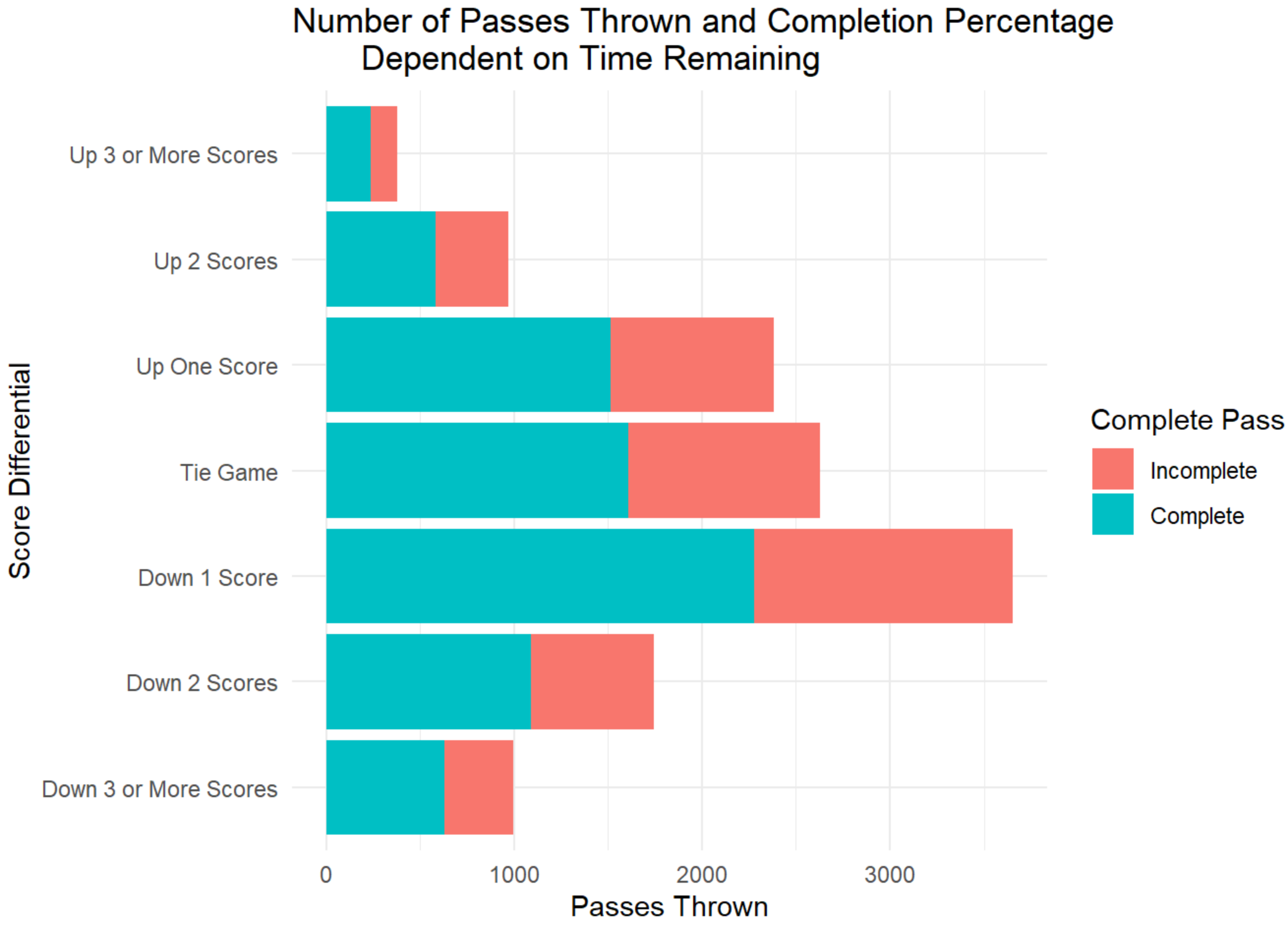
Time Remaining

When considering the game situation, we first wanted to see how much time remaining had an effect on the number of passes being thrown as well as the rate of completion. The graph below shows that there are much more passes thrown at the end of the first half and the end of the second half. Therefore, we created a variable called *end_of_half* because these two periods of time are vastly different than the rest. All football offenses have a scheme called “4 minute offense”. This style of play occurs when offenses are reaching the end of a half and need to score. They will throw the ball far more often because they can gain yards much faster, but more importantly, the clock will stop when an incomplete pass occurs. This allows for the maximum amount of yards gained in the shortest amount of time. Because of this, *end_of_half* is a binary variable where a pass attempt in the last four minutes of the first half and the last four minutes of the second half equal 1, while all other pass attempts equal 0.



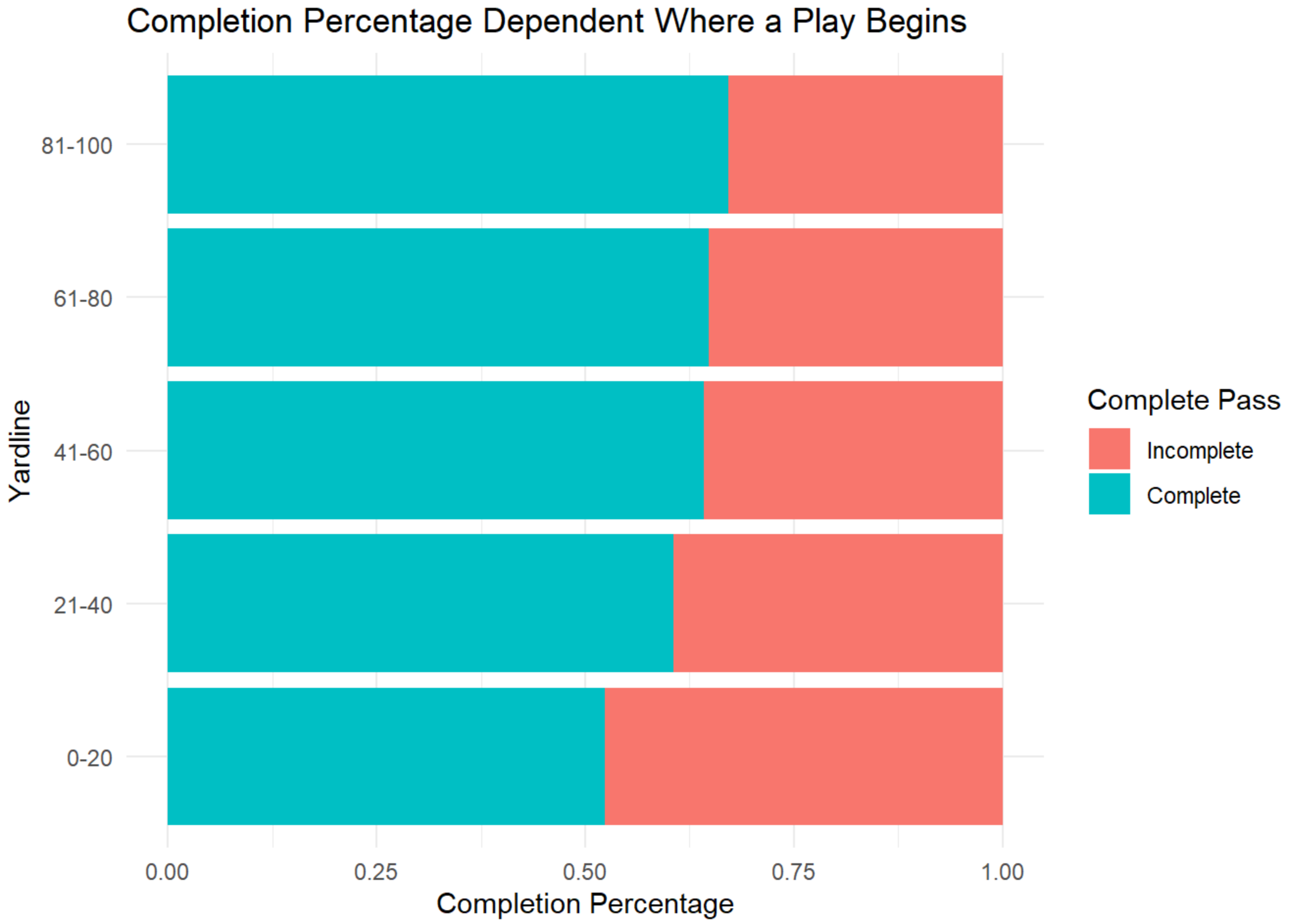
Score Differential

In addition, if a team is losing in the game, they will be more inclined to throw the ball in these situations, while the team that is winning may run the ball to waste time. This leads us to our next variable that accounts for score differential. We cut the variable *score_diff_cat_manual* into categories based off of the NFL scoring system and the ability to score a certain amount of points. Teams will act and call plays noticeably different based on how many scores they are up or down in a game. Teams losing by multiple scores will have to throw the ball with more desperation, and teams up multiple scores tend to try and use as much clock as they can. We can see from our graph below that teams down one score throw the ball much more than teams that are up one score. We can see the overall increase in pass attempts from teams that are trailing, and we expect this to change the probability in a pass being completed.



Field Position

Our last game situation variable is called *field_position_cat*. This splits the field into 20 yard intervals. 0-20 yards encompasses pass attempts from quarterbacks who are backed up close to their own endzone and have to go the length of the field. The 81-100 category contains passes that occur when a team is in the “red zone”, meaning they need 20 yards or less to score a touchdown. From this graph we can see that completion percentage increases as a team gets closer to their endzone.



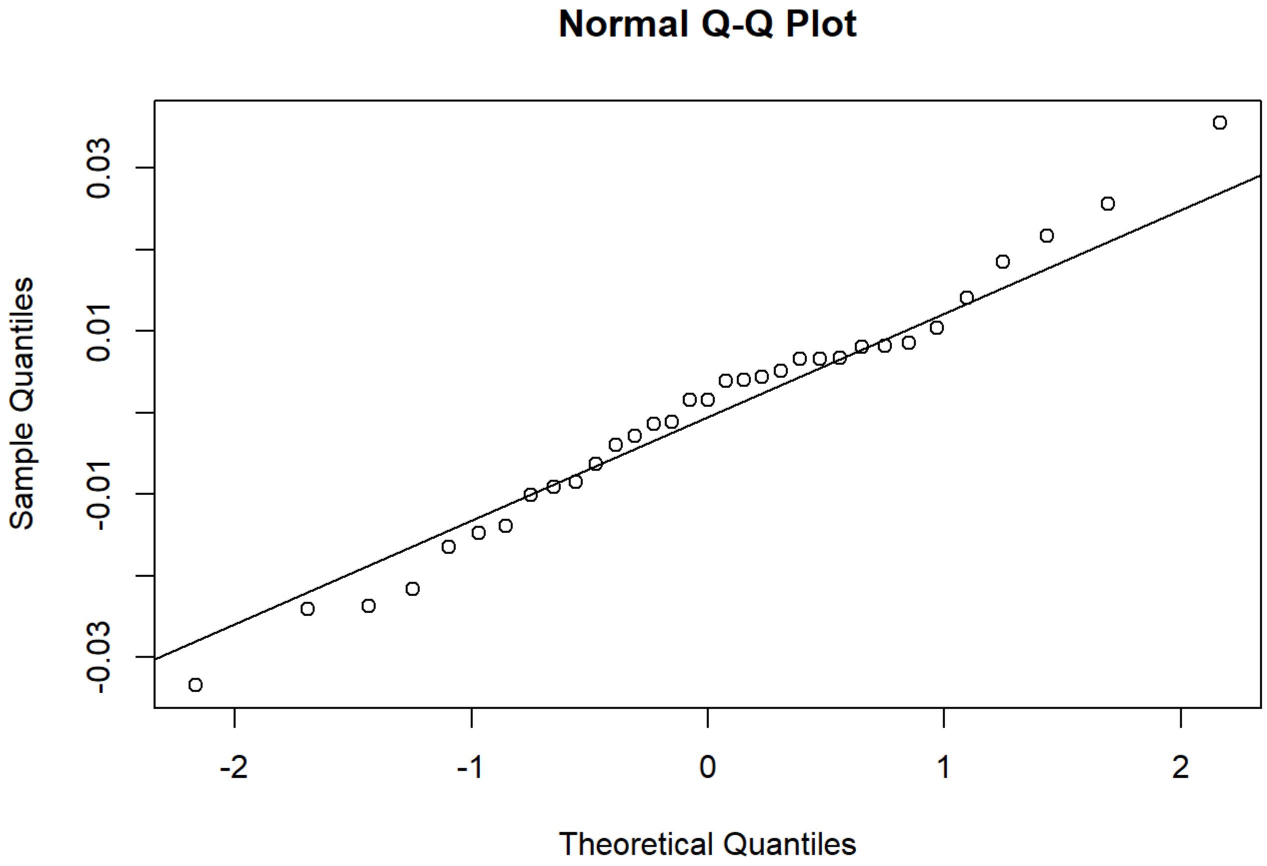
Outcome and Random Intercept

The last variables used in our modeling are *passer_player_name* and *complete_pass*. The variable *complete_pass* is a binary variable, 1 if the pass was completed and 0 if the pass was not completed. This is the outcome variable in our modeling. The variable *passer_player_name* identifies which player threw each pass. The random intercept in our model is based on each quarterback in our dataset, represented by this variable.

Modeling

Description

As mentioned in the introduction, in order to properly model the odds of pass completion, we must recognize the correlation that exists between the quarterback who threw the pass and whether the pass was completed. The specific model that allows us to account for each quarterback, specifically their baseline talent and decision making ability, is a Random Intercept Mixed Effects Model. Mixed Effects Models are one of the two main ways to model longitudinal data, which are situations in which there are repeated measurements over time. The “units” being repeatedly measured in this context are the individual quarterbacks, and the time is the 2022 NFL Season. With this model, each quarterback has their own random intercept, and that intercept deviates from the overall average intercept according to a Normal distribution with mean 0 and variance σ^2. Before continuing, we confirmed this assumption with a QQ plot, checking for linearity.



In regards to the full model, the equation is as follows. Xiβ represents the fixed effects of the model, Zibi represents the random effects, and ϵi represents the error.

Yi = Xiβ + Zibi + ϵi

Another assumption of this model is that the errors are independent with constant variance (σ2e), or in other words, follow an exchangeable correlation structure. This means that the variance stays the same as points grow further apart in space and time. Contextually, the correlation between the quarterback who threw the pass and whether the pass was completed is constant, regardless of if we are comparing passes by that quarterback that are back-to-back or if they happened at completely different points of the season.

Final Model

Of the proposed final models, we made our decision through a balance of football context and goodness of fit measurements. The measure of goodness of fit that we used was the Bayesian Information Criterion (BIC), where the model with the lowest BIC value is selected. At no point did we remove any of the dependent variables previously explained because we valued their contextual importance. However, we attempted to add complexity and quality through interactions of the air yards and score differential variables, as well as an interaction between end of half and score differential. Based on BIC, these changes did not provide enough of an improvement in fit and performance. Below is our final model and its summary output.

VarName	OddsRatio	ConfInt	P.Value
Intercept	3.62	(3.1,4.24)	0

VarName	OddsRatio	ConfInt	P.Value
Pass Outside Hashes	0.78	(0.71,0.87)	0
End Of Half	0.88	(0.79,0.98)	0.015
Down 3 or More Scores	1.15	(0.96,1.38)	0.129
Down 2 Scores	1.17	(1,1.36)	0.043
Down 1 Score	1.16	(1.03,1.32)	0.017
Up One Score	1.2	(1.04,1.38)	0.009
Up 2 Scores	1.06	(0.88,1.27)	0.547
Up 3 or More Scores	1.01	(0.77,1.32)	0.938
Air Yards[6-10]	0.6	(0.54,0.68)	0
Air Yards[11-15]	0.47	(0.41,0.53)	0
Air Yards[16-20]	0.4	(0.34,0.47)	0
Air Yards[21-40]	0.19	(0.16,0.22)	0
Air Yards[40+]	0.11	(0.08,0.16)	0
Inside Own 20 Yardline	0.46	(0.4,0.53)	0
Own 20-40 Yardline	0.86	(0.75,0.98)	0.02
Plus 40-20 Yardline	0.98	(0.87,1.09)	0.679
Redzone	1.1	(0.92,1.31)	0.305

Results and Conclusions

Coefficent Interpretation

There is a significant amount of information to be discussed based on our model. First, there is a statistically significant difference in the odds of pass completion based on pass location. This is seen in the *Pass Outside Hashes* odds ratio that is less than 1. Specifically, the odds of completing a pass “outside the hashes” are on average 0.78 times as high as when thrown in the middle of the field, holding all else constant. Second, as air yards increase, the odds of pass completion continually decreases. In reference to the odds ratios, these values decrease as the air yards increase. Third, the end of each half is significantly different from the rest of the game. On average, the odds of pass completion are 0.88 times as high as compared to the rest of the game, holding all else constant. Fourth, as a team gets further from scoring, the odds of pass completion decrease. Stated differently, starting a play inside a team’s own 40 yard line is predicted to make the odds of completion less than if at midfield. Lastly, the impact of score differential is generally unexpected. Of the possible score differential categories, only being down two scores, down one score, or up one score are significantly different than a tie game.

Random Intercept

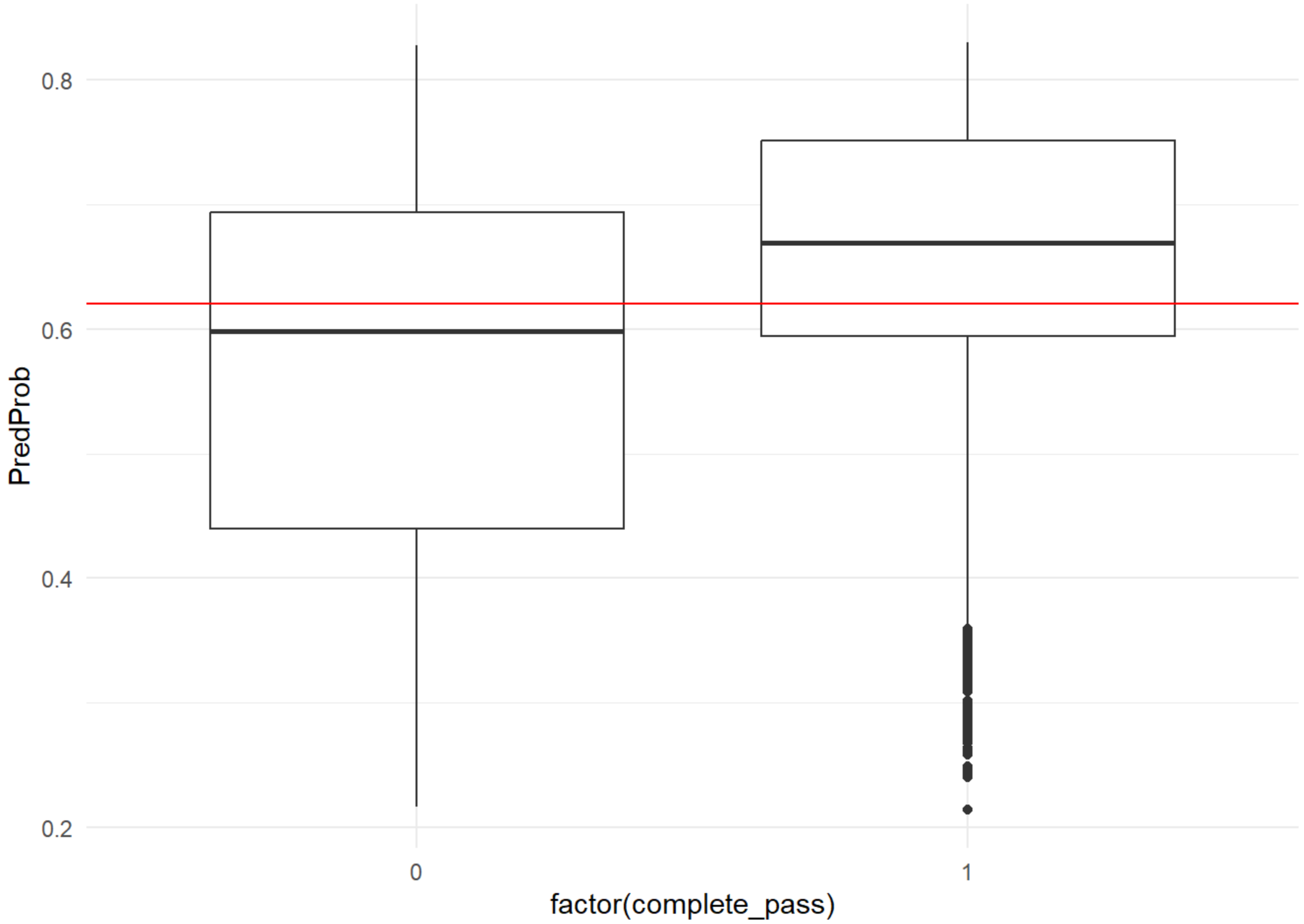
Once again, we accounted for the baseline talent and decision making ability of each individual quarterback with the random intercept. At this point, we can add more context to this value. Because this value is calculated as the log odds of completion for a quarterback divided by the odds of completion of an average quarterback in our dataset, the exponentiated value is an interpretable odds ratio. For example, a 1.03 odds ratio equates to a 3% increase in the odds of completion as compared to the average QB in the dataset. There is one more added piece of detail that is also important. This ratio is based on a very specific situation, that is, our reference categories, because this value is an intercept. The reference categories are passes thrown in the middle of the field, the score is tied, the play begins within 10 yards of midfield, it is not the end of a half, and the air yards of the pass are less than 5 yards. Below is a list of the odds ratios based on our final model, in descending order. We were thrilled to see that this list closely mimicked a professionally developed statistic that is commonly used by the NFL, known as Passer Rating. Besides noticing that this list seemed to generally order quarterbacks based on our perception of their talent and decision making, this similarity was highly encouraging.

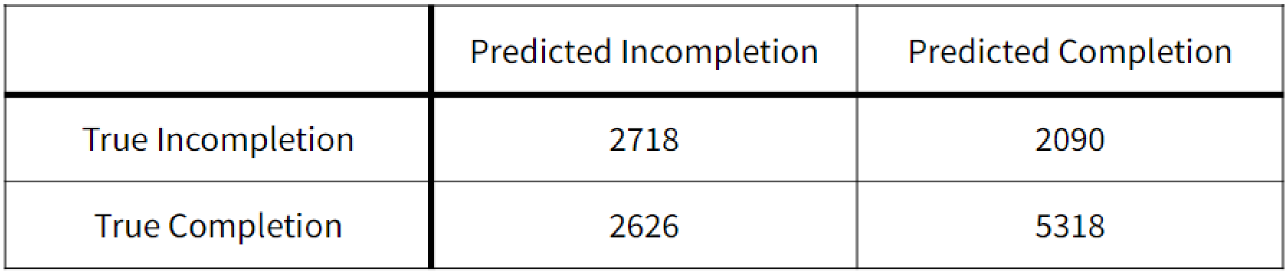
	(Intercept)
G.Smith	1.0361086
J.Allen	1.0259761
K.Cousins	1.0218698
J.Burrow	1.0186745
T.Tagovailoa	1.0141357
J.Herbert	1.0104015
K.Murray	1.0085985
A.Dalton	1.0082404
D.Prescott	1.0080602
M.Mariota	1.0066728
J.Brissett	1.0065820
J.Hurts	1.0065605
R.Tannehill	1.0050691
T.Heinicke	1.0043981
M.Stafford	1.0040173
P.Mahomes	1.0038634
T.Lawrence	1.0015796
J.Garoppolo	1.0015699
A.Rodgers	0.9989100
M.Jones	0.9986213
D.Jones	0.9971419
L.Jackson	0.9960382
C.Wentz	0.9937777
T.Brady	0.9915708
J.Goff	0.9909536
M.Ryan	0.9899360
K.Pickett	0.9861954
D.Carr	0.9853467
R.Wilson	0.9836436

	(Intercept)
B.Mayfield	0.9785633
J.Fields	0.9765232
D.Mills	0.9761612
Z.Wilson	0.9671439

Predictions

Lastly, we wanted to assess our model based on its predictions. First, we had to set a threshold of predicted complete pass probability. Based on the boxplots below, we chose a threshold of 0.62, where if the predicted catch probability was greater than 62%, our model would predict a complete pass. This led to our confusion matrix, also shown below.





	Predicted Incompletion	Predicted Completion
True Incompletion	2718	2090
True Completion	2626	5318

The overall accuracy of our model was 63%. 63% of the time the model correctly predicted if the pass would be complete or incomplete. The sensitivity or true positive rate of our model was 66.9%. 66.9% of truly complete passes are predicted to be complete. The specificity or true negative rate of our model was 56.5%. 56.5% of truly incomplete passes are predicted to be incomplete.

Limitations and Advancements

Unfortunately, our model was only able to accurately predict 63% of the passing plays as completions or incompletions based on our variables. There are still a myriad of factors that are difficult to account for. The original data set had hundreds of variables for each specific play, so creating a model that was simple yet effective was tough using only a handful of variables. There are so many things in football that can affect a ball being completed like dropped passes, passes purposefully thrown incomplete, and defensive talent level. If a quarterback is playing a team that is constantly pressuring him, making him throw the ball quicker, with bodies all around him, he will be less accurate. If he is playing against defensive backs that guard receivers better, then he will not have anybody open to throw to. The quarterback himself has to deal with injuries, travel, and external pressures that vary on a weekly basis, which will affect his performance. There is a certain human element of the game that would take a lot more time and energy to even come close to representing.

We wanted to be able to create a model that is simple enough for people without a football background to understand, while also accounting for the major components of throwing the ball in the NFL. The continuation of this project would mostly consist of added complexity. We would want to add a multitude of other variables to our model, and even make the ones we used more specific. This data is also exclusive to 2022 so our random intercept that accounts for baseline talent and decision making are solely based on one season. It would be nice if we could have more data on each quarterback in order to more accurately represent this. Also, having more years of data would allow us to get a better sense for completing passes in the NFL on a larger scale, rather than just the 2022 season. We would be able to look at passing trends over the years in order to see just how much the game has really changed.