Quarterbacks: Why Football (Kind of) Isn't a Team Sport

You've probably heard of the quarterback. It's the most famous position in all of American sports. He's the guy who throws the football! When a team wins, the quarterback generally receives a lot of praise, and when the team loses, the quarterback receives a lot of blame. This would indicate that the quarterback is largely responsible for whether a team wins or loses.

However, football has also been called the "ultimate team sport." A football *team* consists of a lot of players. Besides the quarterback, there are an additional ten players on offense: the running back, the wide receivers, the linemen, etc. There's also the defense: an entirely different set of eleven players who are also responsible for victory. There are even an additional eleven players that comprise special teams.

So: just how responsible is a team's quarterback responsible for victory? I ran an exploratory data analysis (EDA) to try to provide some answers to this question.

There are two items I would like to address:

1. Are quarterback performance metrics positively correlated with winning football games?
2. If so, can we use quarterback performance metrics to predict how often a team will win?

The first step of the project was to determine how I would operationalize "quarterback performance." This part was fairly straightforward. The NFL tracks plenty of stats for quarterbacks. Some simple examples include touchdowns thrown in a season, interceptions thrown in a season, etc. For more information on what these statistics mean, you can reference this glossary.

As it turns out, football fans can be nerds too. Pro Football Reference keeps an extensive online database of NFL statistics. They do a great job at gathering data and uploading it to a (relatively) easy to read source.

They even include an option that allows users to download the data as an Excel file. I gathered the quarterback passing statistics from the 2010-2019 seasons. I chose this specific range of data for three reasons:

1. I wanted to use a sufficiently large sample size, so I selected a decade of quarterback data.
2. The NFL has changed a lot over the years and the 2010s would contain the data most closely resembling today's sport.
3. In 2021, the NFL changed the season length from sixteen games to seventeen games. I wanted to keep my data consistent.

Once I downloaded all 10 spreadsheets from 2010 through 2019, it was time to clean the data.

First: I had to eliminate unnecessary data. For example, in its "Passing" spreadsheets, Pro Football Reference includes *every player* that completed a pass, not just quarterbacks. This can include some pretty funny names in the spreadsheet.

96	Rex Burkhead	HOU	32	RB	16	0	0	1	0.0	0	0	0.0	0	0.0	0	0.0	0	0.0	0.0	0.0	39.6	3.2	0	0	0.0	0.00	0.00
97	Amari Cooper	CLE	28	WR	17	17	0	1	0.0	0	0	0.0	1	100.0	0	0.0	0	0.0	-45.0	0.0	0.0	0.0	0	0	0.0	0.00	-45.00
98	DeeJay Dallas	SEA	24	RB	15	0	0	1	0.0	0	0	0.0	1	100.0	0	0.0	0	0.0	-45.0	0.0	0.0	0.0	0	0	0.0	0.00	-45.00
99	Phillip Dorsett	HOU	29	WR	15	4	0	1	0.0	0	0	0.0	0	0.0	0	0.0	0	0.0	0.0	0.0	39.6	2.7	0	0	0.0	0.00	0.00
100	Leonard Fournette	TAM	27	RB	16	9	0	1	0.0	0	0	0.0	1	100.0	0	0.0	0	0.0	-45.0	0.0	0.0	0.0	0	0	0.0	0.00	-45.00
101	Chad Henne	KAN	37	QB	3	0	0	2	0.0	0	0	0.0	0	0.0	0	0.0	0	0.0	0.0	0.0	39.6	3.8	0	0	0.0	0.00	0.00
102	Christian Kirk	JAX	26	WR	17	17	0	1	0.0	0	0	0.0	0	0.0	0	0.0	0	0.0	0.0	0.0	39.6	10.7	0	0	0.0	0.00	0.00
103	Cooper Kupp	LAR	29	WR	9	9	0	1	0.0	0	0	0.0	0	0.0	0	0.0	0	0.0	0.0	0.0	39.6	3.3	0	0	0.0	0.00	0.00
104	James Proche	BAL	26	WR	15	0	0	1	0.0	0	0	0.0	1	100.0	0	0.0	0	0.0	-45.0	0.0	0.0	0.0	0	0	0.0	0.00	-45.00
105	Tommy Townsend*+	KAN	26	P	17	0	0	1	0.0	0	0	0.0	0	0.0	0	0.0	0	0.0	0.0	0.0	39.6		0	0	0.0	0.00	0.00
106	Garrett Wilson	NYJ	22	WR	17	12	0	1	0.0	0	0	0.0	0	0.0	0	0.0	0	0.0	0.0	0.0	39.6	0.8	0	0	0.0	0.00	0.00

I'm sorry Leonard Fournette – you don't belong in a "Quarterbacks" analysis.

I eliminated Fournette and the rest of the non-QBs from the spreadsheet. I also cut out some stats that didn't make sense for the purposes of this analysis. For example, there were some columns such as age that were largely irrelevant.

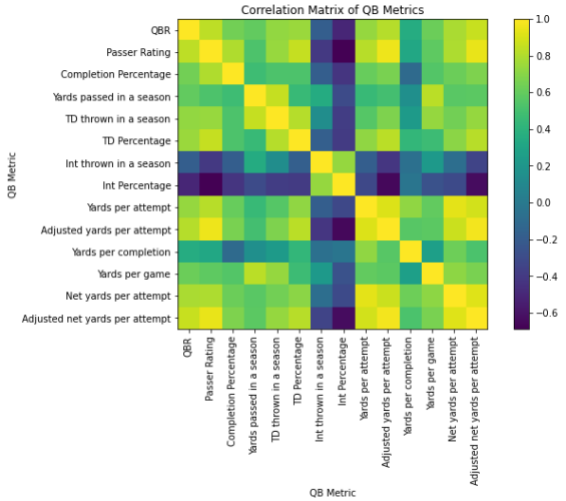
I wanted my analysis to focus on "starting quarterbacks," or the quarterbacks who play the majority of the time. To establish what constituted a starting quarterback, I only included quarterbacks who started at least 8 games – or half of a team's season. While this would reduce my sample size from 457 to 319 quarterbacks (retain 69%), I felt that an n of 319 would be sufficiently large to establish trends.

Finally, to measure "winning," instead of using the number of wins a quarterback had, I chose to calculate each quarterback's win percentage in a separate column. I did this because only counting wins relies on the assumption that every QB played in all 16 games of the season. I didn't want to reward or punish players for having different sample sizes.

For example, imagine a situation where Quarterback A and Quarterback B both had 5 wins, but played in 10 and 16 games respectively:

- If Quarterback A was assumed to have played a full 16 game season, he would be incorrectly considered a "losing quarterback."
- Meanwhile, Quarterback B would be considered to lose a lot of games, rightfully so.

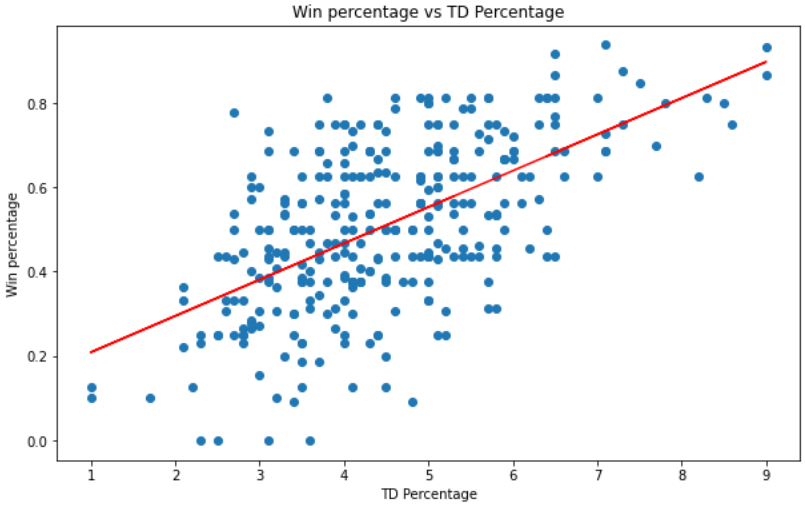
Now that I had my full database of quarterbacks that I would evaluate, it was time to visualize the data to better understand it. I created a correlation matrix to determine how correlated every quarterback metric was with each other.



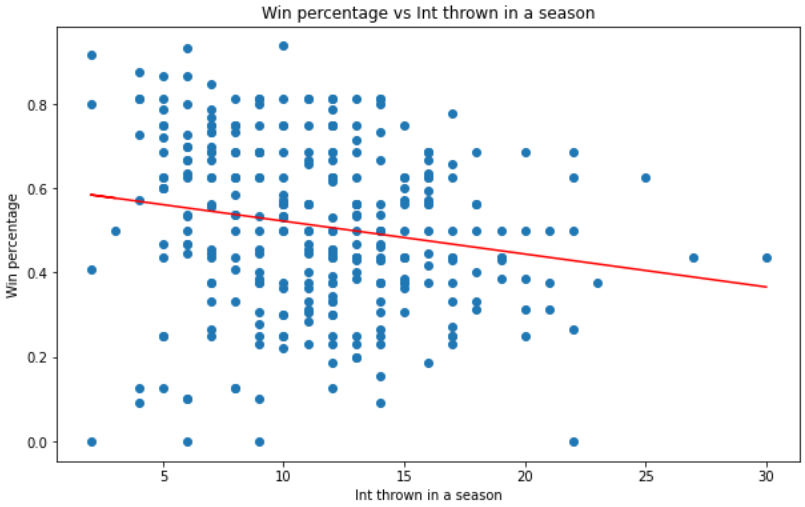
In general, most of these variables are positively correlated with each other. This intuitively makes sense. As an example: yards thrown in a season is very positively correlated with touchdowns passed in a season (Correlation coefficient of 0.850) – if you're throwing for a lot of distance, there are a lot more opportunities to land in the end zone.

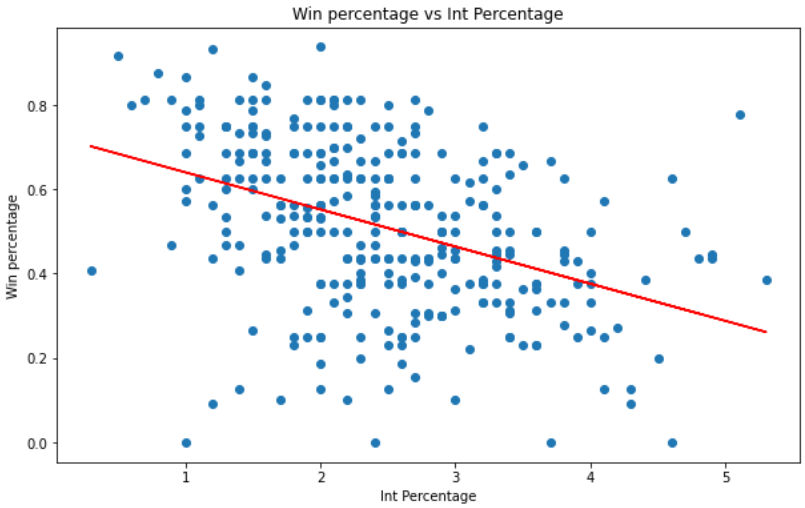
Interceptions thrown in a season and interception percentage are also mostly negatively correlated with other variables. This also intuitively makes sense: if you're throwing the ball to the wrong team, it counts against your completion percentage, removes chances to throw touchdowns, etc.

I also created various scatter plots of quarterback metrics on the x-axis and their win percentage on the y-axis. Most quarterback metrics appeared to have a positive linear relationship with win percentage.



Two variables were negatively correlated with win percentage, and both involved interceptions.

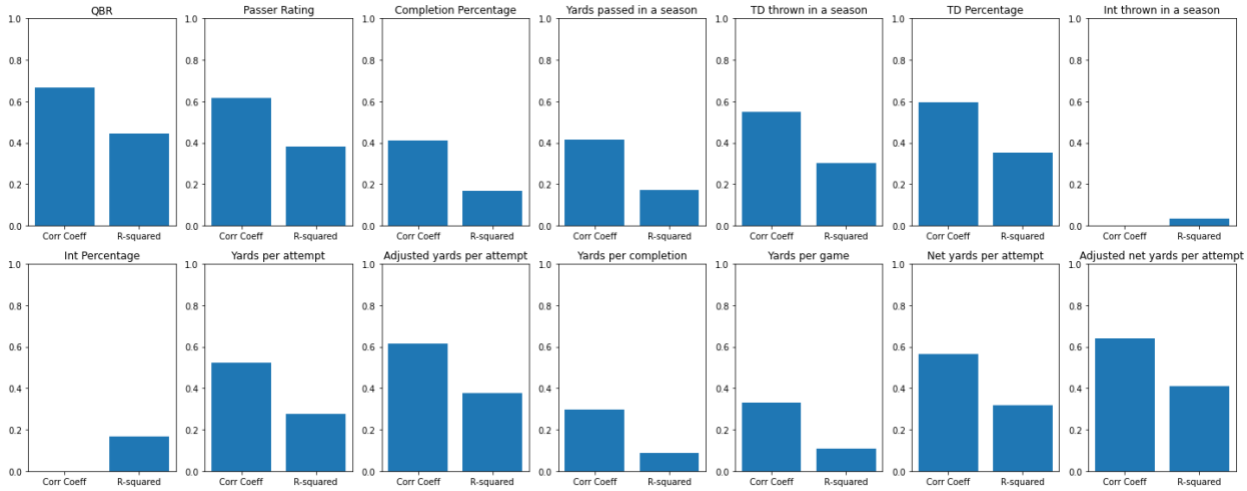




The plot for interceptions thrown in a season has a steeper slope than interception percentage. On a high level, this appears to be a case of volume vs. ratio. In the context of quarterbacks, the number of interceptions a quarterback throws tells an incomplete story. If he throws the ball a lot, he also will be throwing for a lot of yards and have lots of opportunities for touchdowns, both of which are positively correlated with winning. Inevitably, he will also throw a lot of interceptions. You could consider this the "price of doing business." But, if a quarterback is throwing interceptions at a high rate, he is frequently sabotaging his own team, regardless of high or low volume.

Throwing a lot of interceptions is bad, but throwing interceptions at a high rate is worse.

I also ran a linear regression with each variable and win percentage. From this, I calculated correlation and r-squared values for each plot.



I wanted to use one of these stats to perform a linear regression, so I looked for which metric had the greatest "correlation coefficient and r-squared value duo" – i.e. which singular stat was most correlated with winning and explained the most variance.

Of these values, QBR (quarterback rating) has the greatest correlation coefficient and r-squared value duo. However, I will not be using QBR in my analysis, because it is a black box formula. QBR is a formula developed by ESPN used to measure quarterback performance. However, ESPN keeps how the statistic is calculated hidden from the public. No one knows what goes into the formula: touchdowns, yards thrown, etc., nor to what degree each stat is weighted. I chose to exclude this from my analysis because I did not want to base my conclusions off this suspicious statistic.

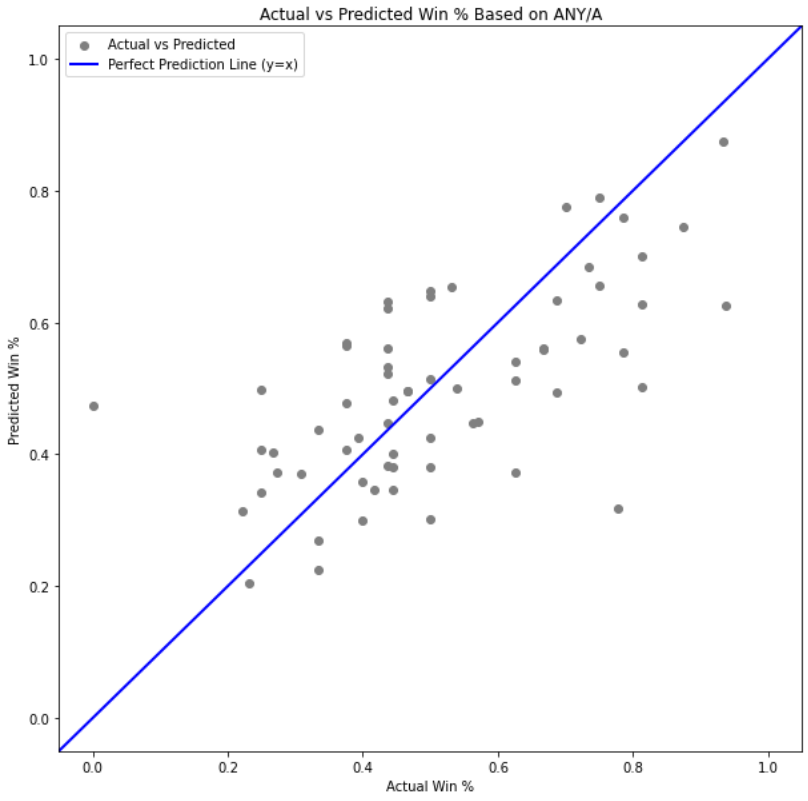
The second greatest correlation coefficient and r-squared value duo was adjusted net yards per attempt (ANY/A) with a correlation coefficient of 0.641 and an r-squared value of 0.412. Like QBR, ANY/A is a calculated formula. However, the formula for ANY/A is transparent.

ANY/A takes in passing yards, sack yards lost, passing touchdowns, interceptions thrown, pass attempts, and times sacked to calculate a corresponding value that is a numerical value of overall how "well" a quarterback played.

$$\text{ANY/A} = \frac{\text{Passing Yards} - \text{Sack Yards Lost} + (20 \times \text{Passing Touchdowns}) - (45 \times \text{Interceptions})}{\text{Pass Attempts} + \text{Sacks}}$$

ANY/A is not a perfect measure of quarterback play because it only considers passing statistics. A quarterback's rushing yards and rushing touchdowns are not counted. Still, as a singular entity, ANY/A takes more variables into account than any singular metric.

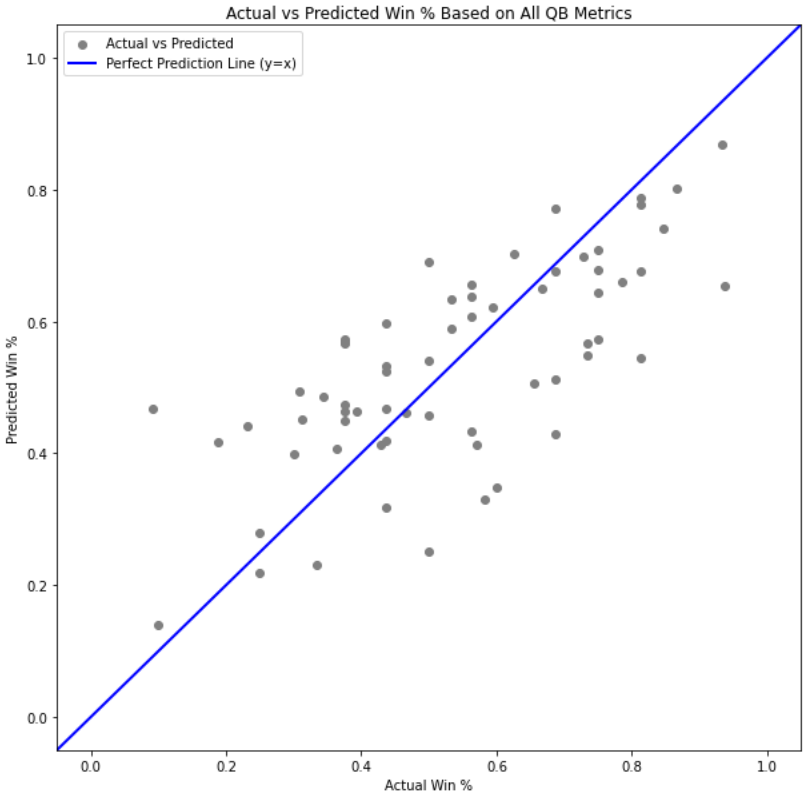
I used ANY/A and win percentage to create a linear regression prediction model. I used a train-test split of 80-20 with my data. This is a plot of my test set predicted accuracy vs. actual accuracy.



```
Intercept [-0.19931152]
Coefficient [[0.11441481]]
Mean Absolute Error: 0.1204685173248754
Mean Squared Error: 0.023042293340541944
Root Mean Squared Error: 0.15179688185381787
R-squared 0.39800902591547926
```

This data had some interesting initial results. First: the RMSE was 0.152 (I will expand on what this means later). Additionally, ANY/A appears to account for a large portion of variance with an r-squared value of 0.398. While the above operation was a linear regression in name, in practice it was a multiple regression. This is because ANY/A uses the aforementioned 6 variables as components. Really, we are using the six variables as multiple regression predictors with modified weightings.

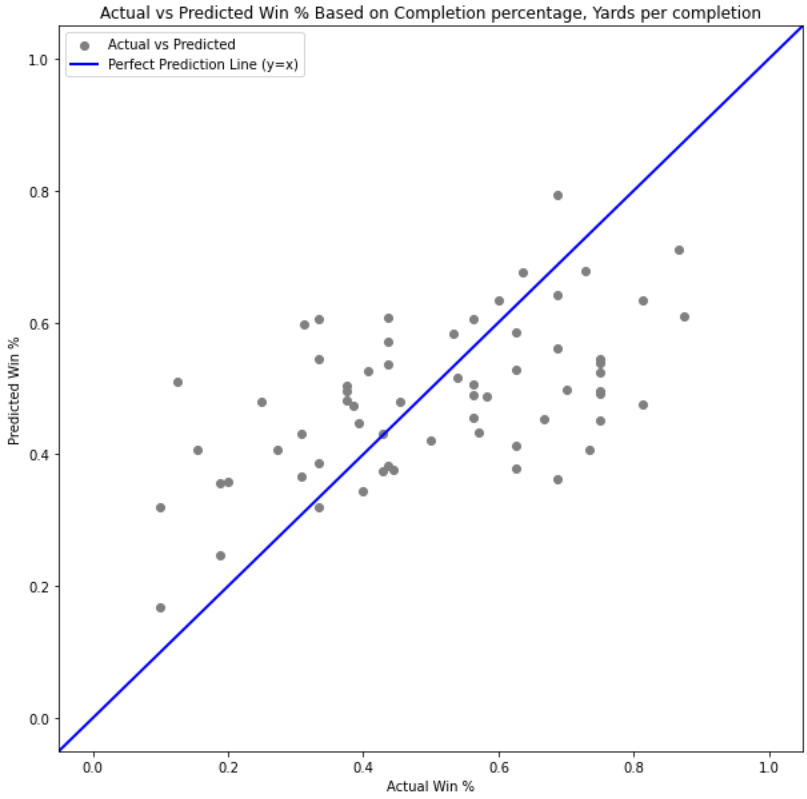
With this, I wanted to investigate if actually using multiple variables to predict win percentage would have higher accuracy. Correspondingly, I created a multiple regression prediction model with all quarterback metric variables.




```
Intercept: 0.665937569745041
Coefficients: [ 6.09154284e-03  4.54722292e-02 -4.95468982e-02  1.36249288e-04
 -6.35566156e-03 -5.48646650e-02 -8.79766127e-03  1.25164558e-01
  3.65375837e-02 -2.29178748e-01 -5.41487572e-02 -2.26106376e-03
  8.62045748e-02  3.69563788e-02]
Mean Absolute Error: 0.11637698532522966
Mean Squared Error: 0.020213315164886017
Root Mean Squared Error: 0.14217353890540257
R-squared 0.5068023960184189
```

I ran this multiple regression knowing full well that there was going to be high multicollinearity. The correlation matrix that I previously calculated would indicate as much. I still ran this multiple regression because I thought that the results would still be interesting to look at. Technically, I have reduced the RSME to 0.142 and increased the r-squared to 0.507.

After testing this multiple regression with all ten variables, I wanted to select two independent variables to use in a separate multiple regression. I went back to the correlation matrix I calculated above and selected two uncorrelated variables to contribute distinct additions to the overall variance. I found the variables with the lowest correlation coefficient to be completion percentage and yards per completion at 0.073.[1] I ran a multiple regression prediction model with these two variables.




```
Intercept: -1.6883684268740031
Coefficients: [0.02093507 0.07652682]
Mean Absolute Error: 0.1421002706995771
Mean Squared Error: 0.029093893862845065
Root Mean Squared Error: 0.17056932274839184
R-squared 0.284439672763222
```

This multiple regression got an interesting result: an increased RMSE value and a decreased r-squared value. While this model avoids issues of multicollinearity, it is a worse predictor than ANY/A alone.

[1] Other variable pairs had lower correlation coefficients, but one variable contained the other in the calculation. I excluded these such pairs to avoid multicollinearity.

Since I have just reported my least accurate model yet: I believe it's a fitting time to discuss my interpretations of the RSME. Because RMSE is directly interpretable in terms of units, this meant that my predictions had an error between 0.142 - 0.171 win percentage.

Earlier I stated that quarterbacks with 8 or more starts would be included in this dataset. I also gave an example of how using the number wins as our variable could be misleading. But for the sake of putting win percentages into perspective, I'll temporarily convert win percentage back into games won / team record.

To convert the calculated win percentage error value into games, you simply multiply the win percentage by games in a season. Roughly half of all the quarterbacks in my reduced dataset have started 16 games (a full season), so this is a decent approximation. Let's go back and use the ANY/A error value: 0.152 x 16 = 2.43 games. In other words, the model predicted a team's record to be ±2-3 games off from their real record.

A ±2-3 game difference is an enormous gap for NFL teams. For example:

- An 8-8 team is average. They usually can't make the playoffs and can't get top draft picks.
- A 6-10 team is below average and should focus on doing better next season.
- A 10-6 team is in playoff contention and has a nonzero chance to win the Super Bowl.

If all you had to predict a team's success was the quarterback's ANY/A, you could get somewhat close to guessing their team's real record. But if the effective guessing range is this wide, then it is not that accurate of a predictor. With a quarterback's statistics, you can get *somewhat close* to guessing a team's win percentage, but you will rarely get a truly correct guess.

Consider the structure of how a football team runs. Every football team has an offense, a defense, and special teams. The quarterback only plays on one of those three subsets. And even on offense, the quarterback *doesn't always throw the ball:* the running back can take the ball and run with it!

Even when the quarterback throws the ball, it's not a one-man-band. He needs the offensive line to block and give him enough time to throw the ball. His receivers need to catch the ball. His coach needs to design a play to put the players in position to even make the throw and catch possible. And none of the quarterback stats will directly tell you this information.

In summary: quarterback performance is very correlated with winning. However, while a quarterback almost certainly has a causal impact[2] on winning, it is not the only variable that affects winning. Still, if you're going to only use one variable to determine if a team will win or lose, quarterback performance is a very good choice.

[2] Performing causal interference calculations is something I would like to revisit in the future.

While the purposes of my original EDA were to determine the correlation and potential causality between quarterback metrics and winning, I also wanted to take the opportunity to try to classify quarterbacks into different tiers. To do this, I started by running a K-means algorithm on ANY/A and win percentage to determine the ideal number of clusters to consider. I selected ANY/A for the combination of simplicity and consistency from the calculations above.

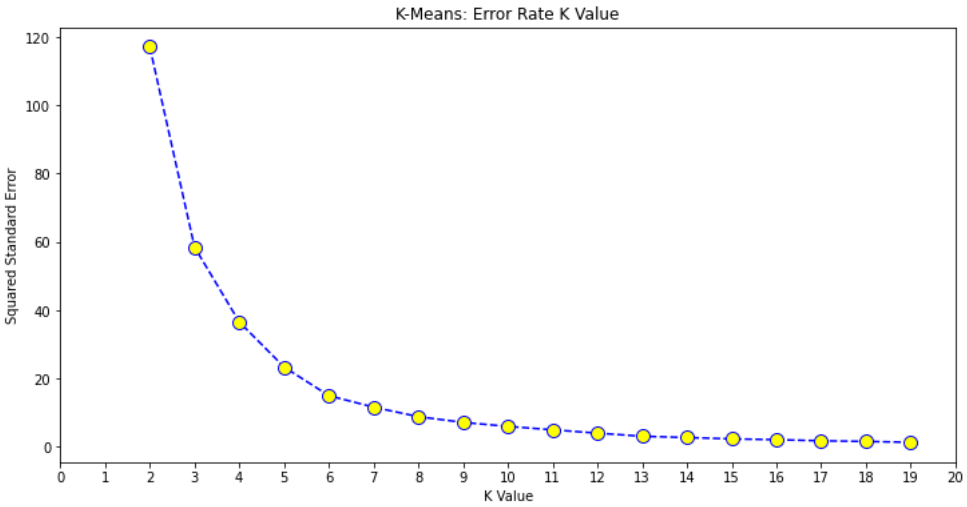
To determine the ideal k-value, I used the elbow method and calculated silhouette scores.

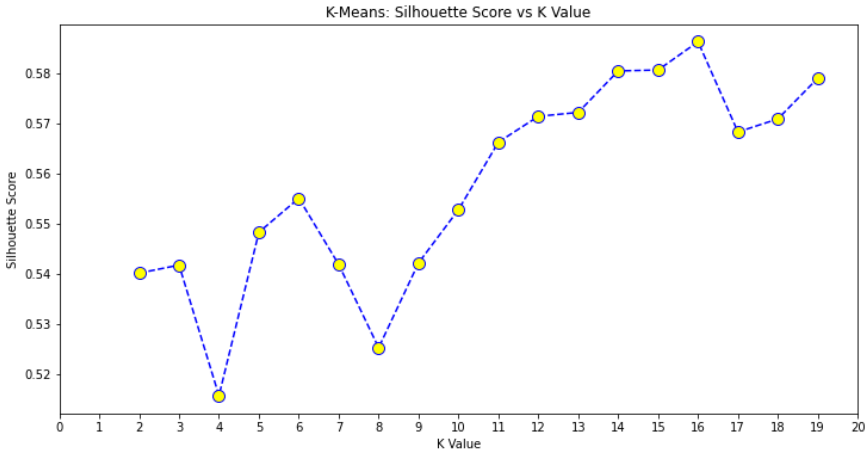
Using the elbow method, I found that the ideal k-value was 4. However, the greatest silhouette score was for k= 16. This is tricky because my results contradict each other.

I tried comparing each method's ideal k-value to the other test. k = 4 on the silhouette score graph had the lowest silhouette score and therefore the worst possible k-value. Meanwhile, k = 16 on the elbow graph is at a point where the trajectory has long since turned linear, also making it a poor choice.

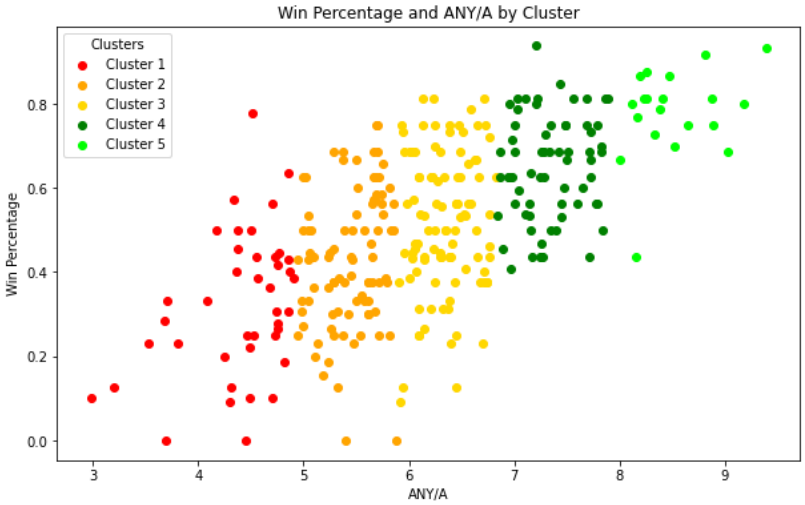
With this, I looked at the context of the data. I am using k-means to classify quarterbacks into groups. There are 32 teams in the league and correspondingly ~32 starting quarterbacks in a given season. If we choose k = 16 and create 16 categories of quarterbacks, the average group size would be very small (i.e. 2 QBs per season). This would be too specific and there would be very little different from one group of quarterbacks to the next. Conversely, choosing k = 4 creates fewer classifications and more members per group. This does a better job in creating distinct groups of quarterbacks where there are notable differences between consecutive groups. However, k = 4 also comes with its own issues: k = 4 is an even number of groups, meaning there is no true "average" quarterback. There would only be "above average" and "below average" quarterbacks.

Therefore, I chose k = 5 as my k-means k-value. This would allow for distinct, moderately sized groups of quarterbacks to form, while also enabling the creation of an "average quarterback" group. It's also important to mention that while k = 5 does not have the best silhouette score, the difference between k = 21 and k = 5 is relatively minor (0.038).





Once I determined my k-means k-value, I re-plotted the data with the clusters and created a table with clusters, ANY/A, and median win percentage:



Cluster	ANY/A	Win Percentage
1	4.5	0.307692
2	5.47	0.428571
3	6.33	0.5
4	7.28	0.666667
5	8.395	0.8

I then named each of these clusters as such:

- Cluster 1 = “Terrible”
- Cluster 2 = “Bad”
- Cluster 3 = “Average”
- Cluster 4 = “Good”
- Cluster 5 = “Great”

In some unsupervised machine learning experiments, there isn't any accompanying data to check if the results are accurate. In my case however, the NFL awards quarterbacks with honors based on their performance.

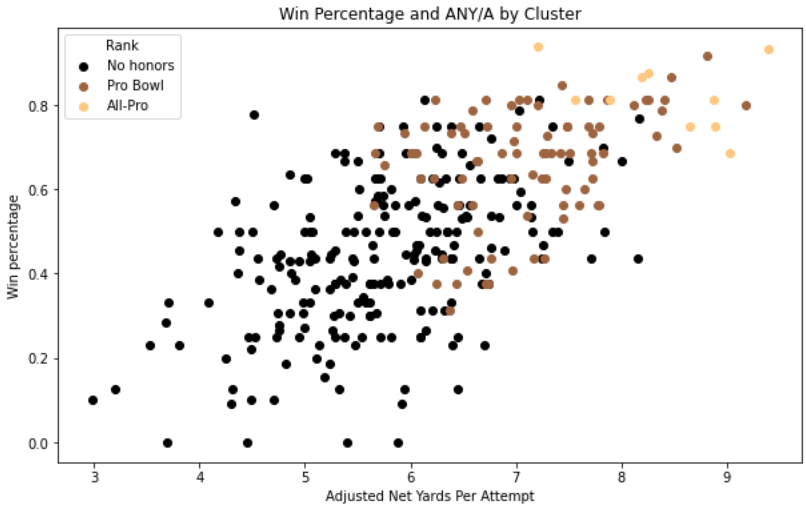
- Most quarterbacks in a season do not receive any honors.
- A group of quarterbacks with good performance in a season are awarded with a "Pro Bowl" nomination.
- The "best" quarterback in a given season gets voted as a "First Team All-Pro."

I chose to use these honors as my reference point for determining if the 5 clusters were accurate to reality. I went back to the dataset and assigned each quarterback a rank depending on the highest honor they received.

- If a player received no honors = rank of 0.
- If a player received a Pro Bowl = rank of 1.
- If a player received a First-Team All-Pro = rank of 2.

I plotted every player based on their ANY/A and win percentage, then assigned each data point to their rank.[3]

The first thing I observed was that every single All-Pro came from the top right corner of the graph: high ANY/A and high win percentage. I also observed that Pro Bowl quarterbacks had a surprising range of win percentages. While the Pro Bowl median win percentage was 0.6875, a quarterback with a win percentage as low as 0.3125 still received an award (Philip Rivers in 2016).



Rank	ANY/A	Win Percentage
No honors	5.7	0.4444
Pro Bowl	7.1	0.6875
All-Pro	8.45	0.8125

[3] Players can receive both Pro Bowl and First-Team All-Pro awards. If a player received both, I gave the player a rank of 2.

I created a GIF to help visualize how the k-means generated clusters mesh with the actual awarded honors. I uploaded the GIF to my GitHub here.

I also created a table combining the k = 5 clusters and counts of awards. I wanted to see how many quarterbacks in each group received awards.

K-means cluster	No honors	Pro Bowl	All-Pro	Total
Terrible	41	0	0	41
Bad	86	4	0	90
Average	74	29	0	103
Good	21	41	3	65
Great	3	10	7	20

I made several observations from this table:

- Terrible quarterbacks don't get awards. Even over a decade, there were no "fluke awards."

- There are fewer Terrible quarterbacks than Bad quarterbacks.
    - This is probably due to Bad quarterbacks being more justifiable than Terrible quarterbacks.
    - If a quarterback is Terrible, they put up awful performances, the coach benches them, and the quarterback never plays again.
    - Lots of circumstances can lead to a Bad quarterback. They could be a rookie, in which case they're not expected to perform but still get multiple chances. Or, there might not be a better option on the team: despite the quarterback on the field playing poorly, the other quarterbacks on the bench belong in the Terrible tier.

- Average quarterbacks generally don't receive awards – only 28% (29/103) did. Of the Average quarterbacks that received an award, 80% (23/29) had some combination of above average stats + above average winning. And even then, none of them received an All-Pro nomination.

- Most Pro Bowl quarterbacks fall in the Good tier. In a given season, about 6-8 quarterbacks make the Pro Bowl, meaning there can be a bit of variance as to which guys are selected. Additionally, the Pro Bowl is not the highest award given – therefore, it makes sense that you only have to be "Good" to receive one.

- 95% of Pro Bowlers come from the Average, Good and Great clusters (80/84).

- Only 5% of Pro Bowlers come from the Bad cluster (4/84).
    - Quarterbacks who make it from here most likely have suboptimal win percentage but extremely good stats.

- 3 players have received the All-Pro award despite coming from the Good cluster:
    - Cam Newton (2015): Despite having the worst passing statistics of all the included All-Pro quarterbacks, Newton had the best *rushing statistics* with 636 yards and a whopping 10 rushing touchdowns.
    - Tom Brady (2017): Brady attempted the most passes out of any quarterbacks – therefore ANY/A penalized him for this suboptimal efficiency. He still led the league this year in yards thrown and yards per game, so the raw numbers probably looked impressive enough to vote for.
    - Peyton Manning (2012): Manning led the league in ANY/A and completion percentage this year. It's worth noting that in 2012, Manning came back from a potentially career ending injury from the prior year – possibly endearing him more to the voters.

- 3 quarterbacks from the Great cluster received *no awards.* All three cases are due to injury: the quarterback was performing very well but suffered a season ending injury, and the voters selected other players for awards.
    - Andy Dalton (2015): Broken thumb.
    - Matthew Stafford (2019): Spinal fractures.
    - Aaron Rodgers (2013): Broken collarbone.

- Great quarterbacks are truly rare. In 10 seasons, there were only 20 great quarterback seasons. On the extreme ends of the spectrum: there are twice as many Terrible quarterbacks as Great quarterbacks!

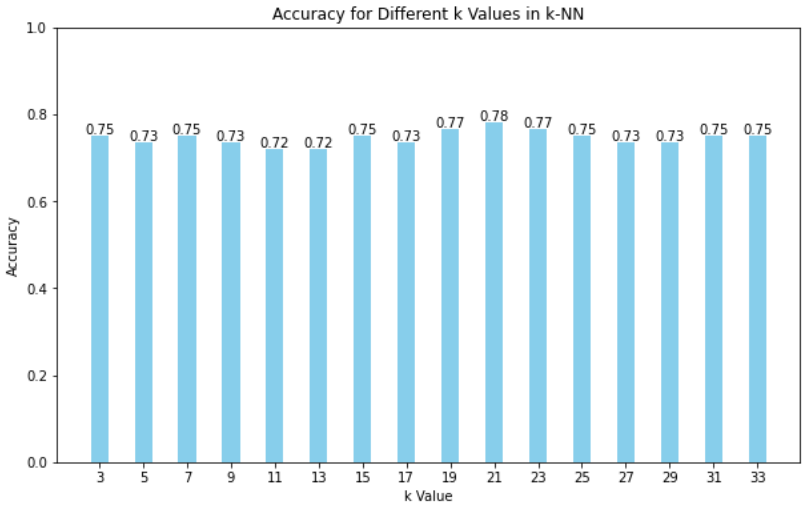
After making these observations, I took a step back to evaluate the legitimacy of checking the number of k-means clusters by referencing NFL awards.

- The awards are helpful in determining the validity of the upper tiers: Great and Good. Great quarterbacks receive the best awards, and Good quarterbacks receive the second best awards.
- The awards are also plausible for the Average tier: a supermajority of Average quarterbacks do not receive any awards, and only a small portion of Average quarterbacks win the second best awards (Pro Bowls).
- However, using awards to identify the legitimacy of the Bad and Terrible tiers is tricky. Bad and Terrible quarterbacks don't receive positive awards, and there is no "worst of the worst" award given to these quarterbacks. I only have an absence of evidence to work with here. The validity of these two tiers is somewhat inconclusive.

After running the k-means algorithm, I wanted to build a k-NN algorithm to take in a quarterback's ANY/A and win percentage, then predict whether they received no award, Pro Bowl, or All-Pro.

Once again using an 80-20 train-test split, I ran several k values to determine what the most optimal k-value to use in the k-NN algorithm was.

k = 21 had the highest prediction accuracy at 0.78. Interestingly, differing k-values did not have a drastic impact on prediction accuracy.



To conclude this extension section: if a quarterback performs poorly and doesn't win, they will almost never win an award. If a quarterback performs above average and has an above average winning percentage, they have a good chance of receiving a Pro Bowl nomination. And if a quarterback performs extremely well *and* wins a lot while doing it, they'll be in line for the All-Pro nomination.

Some extra commentary on my overall EDA:

Quarterbacks have improved as time has gone on. Superior training, dieting, coaching, film review, etc., have all contributed to quarterbacks throwing further distances and more accurately. Statistics that would have earned a Pro Bowl nomination in the 1990s would fall short of the bar set in the 2020s. Correspondingly, I would imagine that the categories established in my analysis would become less and less accurate over time. However, for the immediate future of the 2020s, I believe these conclusions and values would hold true.

Finally, a word for football fans:

- If your team has a Great cluster quarterback, the team better do everything in their power to go all in and push for a Super Bowl. These guys are rare.
- If your team has a quarterback who would fit in the Bad or Terrible cluster, they better replace him as fast as possible.
- The worst possible situation is if your team has a Good cluster quarterback. These quarterbacks are certainly better than most in the league, but probably can't take a team to the promised land and win a Super Bowl.
    - Trying to replace such a quarterback will be a gamble. The chance that the replacement will be worse than his predecessor is very likely. But keeping the above average quarterback will cap the potential of the team.

Win or lose, the quarterback is in the spotlight. Every Sunday, Americans across the country watch him throw the ball. He's under an enormous amount of pressure to perform and win. Even if he's making tens of millions of dollars to play a sport and routinely making mistakes worthy of the Terrible tier, we should give these guys some empathy – they're human too.