Brad Nott
(310) 923-6337
bradley.nott@gmail.com

NFL Big Data Bowl 2019

Introduction

Speed is exciting. It should be no surprise that many of the metrics that football analysts and fans love to quote have an inherent speed component. Sometimes this fascination with speed is explicit. A good example is the Fastest Ball Carriers metric. More often though speed is simply implied like when ranking quarterbacks by Time to Throw or receivers by Yards After Catch. Regardless of the perspective, speed is a crucial aspect of the game.

In some situations understanding the value of speed is straightforward. For example, it is just as important for wide receivers to have explosive acceleration and high top-end speed as it is for the opposing defensive backs who are charged with disrupting pass routes. Additionally, a quarterback who can fire off a pass quickly also needs receivers who can get to their positions on the field rapidly.

In other situations the way that speed contributes to the success of a play is less obvious. As a result, speed metrics might not immediately seem useful. In order to appreciate the impact of speed we need context, and we can establish that context with data. We already collect a lot of data on NFL players and games in order to gain insights and improve our decisions. Now that we have the ability to take substantially more measurements than ever before during an NFL game, we face an exciting new challenge: how do we find value in the data that is actually useful for teams?

To answer that question we need to think critically about what we should measure, and can actually measure, during a game. Once we properly frame the problems we want to solve, we can formulate meaningful questions that reveal what the collected data can explain.

The Effectiveness of Speed

There are ways to be effective using speed other than simply being the fastest player to run a 40 yard dash. It is definitely exciting to watch an elite player expertly evade defenders and open up the throttle for an uncontested touchdown. It certainly helps to be the fastest player on the field, but the gametime opportunities to get up to top-speed without interruption are limited. It is more useful to think about speed as a capability. In other words, it is not always top speed, but more often optimal speed and optimal movement that win the day.

If a more general way to think about speed is optimal speed, we need to think about what that means. A combination of various decisive factors such as decision-making, speed, balance, maneuverability, momentum, and strength contribute to the success of a football play. While some factors, like decision-making, might be hard to measure, we can study the outcomes of plays as a means to learn about how these factors might interact. Moreover, with

Next Gen Stats, we can specifically dive into the necessary details to learn about what encourages the outcomes we want to reproduce.

Focus of This Study

In recent NFL seasons rule changes, as well as breakout players, have placed the focus on quarterbacks and wide receiver metrics. We are currently in an era of record-breaking offensive performances. As a result, the focus of metrics and analysis is heavily weighted toward understanding, and often revering, NFL teams with a pass-heavy offense. This study takes an alternate approach and considers optimal speed in the rushing game.

Rushing plays fail for many reasons, and many of those reasons are very difficult, if not impossible, to consistently predict. There is legitimate value in determining why a play did not work. Unfortunately, that can often be a fairly complicated and subjective task. So instead, as a more straightforward initial approach, we will look primarily to plays that were successful at gaining yards. We want to understand the speed-related factors that keep a play alive and help propel a drive to the end zone. We want to know what encourages that success.

To help us find our answer, we will focus on several fundamental themes:

1. Efficiency
2. Momentum

Data Preparation

The Big Data Bowl *plays.csv* data is our roadmap. It serves as the guide to plan an analysis of the tracking data files. The goal is to find all running plays and associated ball carriers in order to dive into player-level game data. While each of the data sets could have been merged into one large data set, it is more intuitive to start with a specific purpose and then gather only what is needed from each of the 91 game tracking data sets.

First, a subset of all non-penalty and non-special teams plays is taken. Two-point conversion plays are also removed. While these do contain some rushing plays, they represent a special-case and correspond to only a small portion of the data. Therefore they are not included. For similar reasons, fumble plays are excluded as well.

Once this is done, a column containing the names of the ball carriers on each play is inserted into the play data in order to facilitate pairing ball carriers with plays.

Theme 1: Efficiency

Players who use speed in an optimal way are efficient. Their decision-making and agility often lead to additional yards when they are most needed. To operationalize the concept of rushing efficiency we will in part use the popular Running Back Success Rate and Rushing Efficiency metrics.

Running Back Success Rate (credit: Football Outsiders):

- A play counts as a "Success" if it gains 40% of yards to go on first down, 60% of yards to go on second down, and 100% of yards to go on third or fourth down.
- If the team is behind by more than a touchdown in the fourth quarter, the benchmarks switch to 50%/65%/100%.
- If the team is ahead by any amount in the fourth quarter, the benchmarks switch to 30%/50%/100%.

Rushing Efficiency (credit: Next Gen Stats):

- Rushing Efficiency is calculated by taking the total distance a player traveled on rushing plays as a ball carrier according to Next Gen Stats (measured in yards) per rushing yards gained. The lower the number, the more of a North/South runner.

Running Back Success Rate

To calculate success rate a *RunSuccess* column was first added to the *plays.csv* data set to represent yards gained divided by yards needed for each rushing play. Then, using a function that incorporated the associated decision rules, a new column was generated to indicate if a run was a deemed a success. Determining success rate then became a simple task of grouping the *plays.csv* data by ball carrier, and dividing a player's total successful runs by their total carries.

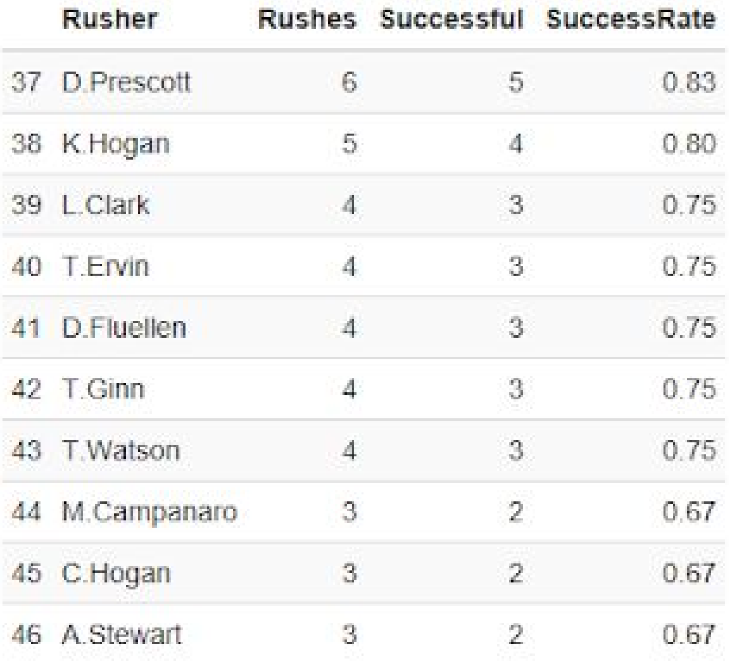
You could stop here and simply sort players by their success rates. However, this presents a few problems. For one, some rushers appear more frequently in the data than others. This imbalance results in their success rate being more representative since it is based on more data. Additionally, other players have few total carries but many of them are successful. This results in a high success rate based on little evidence.

If success rate is going to be used as criteria for sorting and ranking running backs then we must weight the ranking with a confidence measure. Think of it like this: at an online retailer's website you naturally trust a 4.5-star rating with 1,000 reviews more than a 5-star rating with 10 reviews. When there is less evidence there is greater uncertainty. In our case, the necessary correction can be made using a clever technique that several websites employ to allow users to sort content by “best.”

Running Back Ranking Scheme

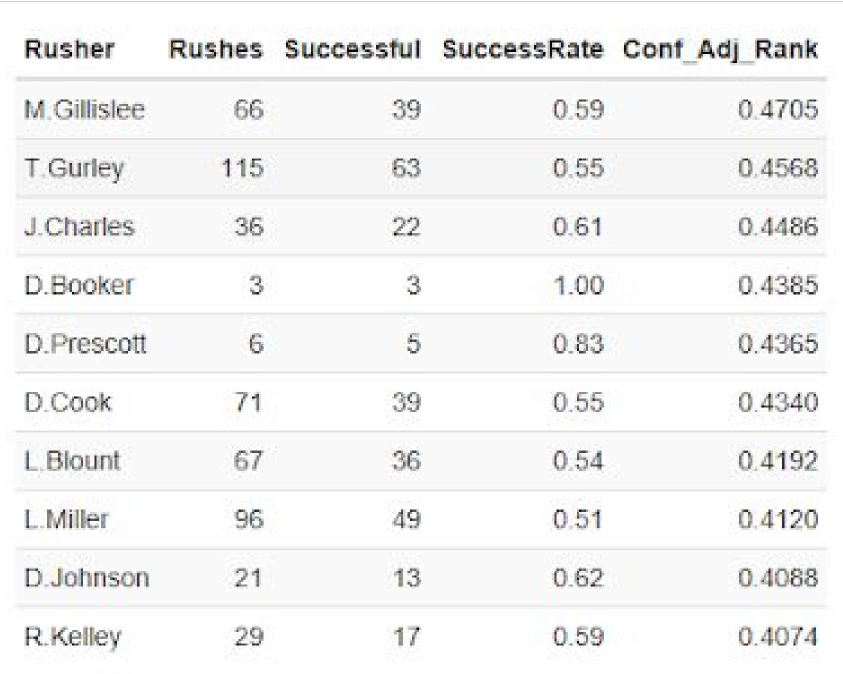
Since the *RunSuccess* column only takes on values of true or false, we can create an adjusted ranking scheme using a method known as the Lower bound of a Wilson score confidence interval for a Bernoulli parameter. The method takes a set of positive and negative ratings (e.g., true or false) and estimates the real fraction of positive ratings provided a desired confidence level.

While the success rate metric seeks to capture the value of a rushing play, this sorting scheme provides a confidence-adjusted ranking. In the following tables note how the top ten rushers correctly sort to the top when the ranking scheme is employed. In the first table only rushers with a success rate less than one were included in the top ten. This was done to facilitate recognition of the fact that a small sample size and the successful rush proportion can result in the metric not being very useful. While more elaborate ranking schemes exist, usually based on the influence of additional metrics, this approach is a very simple way to add context to a commonly used metric.



	Rusher	Rushes	Successful	SuccessRate
37	D.Prescott	6	5	0.83
38	K.Hogan	5	4	0.80
39	L.Clark	4	3	0.75
40	T.Ervin	4	3	0.75
41	D.Fluellen	4	3	0.75
42	T.Ginn	4	3	0.75
43	T.Watson	4	3	0.75
44	M.Campanaro	3	2	0.67
45	C.Hogan	3	2	0.67
46	A.Stewart	3	2	0.67

Figure 1: non-adjusted running back ranking by success rate



Rusher	Rushes	Successful	SuccessRate	Conf_Adj_Rank
M.Gillislee	66	39	0.59	0.4705
T.Gurley	115	63	0.55	0.4568
J.Charles	36	22	0.61	0.4486
D.Booker	3	3	1.00	0.4385
D.Prescott	6	5	0.83	0.4365
D.Cook	71	39	0.55	0.4340
L.Blount	67	36	0.54	0.4192
L.Miller	96	49	0.51	0.4120
D.Johnson	21	13	0.62	0.4088
R.Kelley	29	17	0.59	0.4074

Figure 2: confidence-adjusted ranking by success rate

Rushing Efficiency

Rushing efficiency is a metric that serves as a proxy for many aspects of a run play. It is the product of a player's athleticism and agility. It is also reflective of a player's effort to fight and weave through traffic to gain more yards. Ultimately, the rushing efficiency metric attempts to account for all of the unique movements related to a particular position.

To calculate Rushing Efficiency, or *EFF*, a column was generated by iterating over the game tracking data sets and calculating a rusher's total distances traveled on all plays. This distance column was then divided by the *PlayResult* column to derive the *EFF* column.

Theme 2: Momentum

Rushers who use speed in an optimal way are also tenacious. No matter the situation, they find a way to maintain their momentum until the defense forces them to stop, or until they score a touchdown. Sometimes being a battering ram gets the extra yards. In other situations it is all about the agility necessary to avoid or shed tackles. For some players, given their size and strength, the likelihood that they will gain yards after contact might increase after they accelerate up to a certain optimal speed.

The best way to gauge a player's ability to maintain momentum is to measure the yards they gain after first contact with the defense. Measuring yards gained after contact is a way to identify and understand players who best utilize their remaining speed to make forward progress. Additionally, measuring a player's speed at the moment they make contact is vital to understanding what it means to use speed effectively throughout a rushing play.

Yards After Contact

- Reveals the influence of momentum and maneuvering in gaining additional yards before the play is over.

Contact Speed

- Determined by calculating a rushing ball carrier's speed at the moment they make contact with defensive resistance.

The Yards After Contact metric is derived from each of the run plays a particular ball carrier participated in throughout all of the game tracking data files. For a particular game and play, a ball carrier's data is examined in order to determine the total distance they traveled following first contact with a member of the opposing team. A corresponding contact speed was also extracted from the appropriate game tracking files in a similar fashion.

Additional Variables

In addition to the variables and columns described, additional variables were calculated or derived and inserted into the *plays.csv* data set to facilitate answering questions and generating exploratory plots. The entire procedure is detailed and available for review in the Python code appendix.

Results

We might as well begin with the most-interesting observation. In the scatterplot (figure 3) we see the relationship between a player's speed at the time of contact and the yards gained after contact.

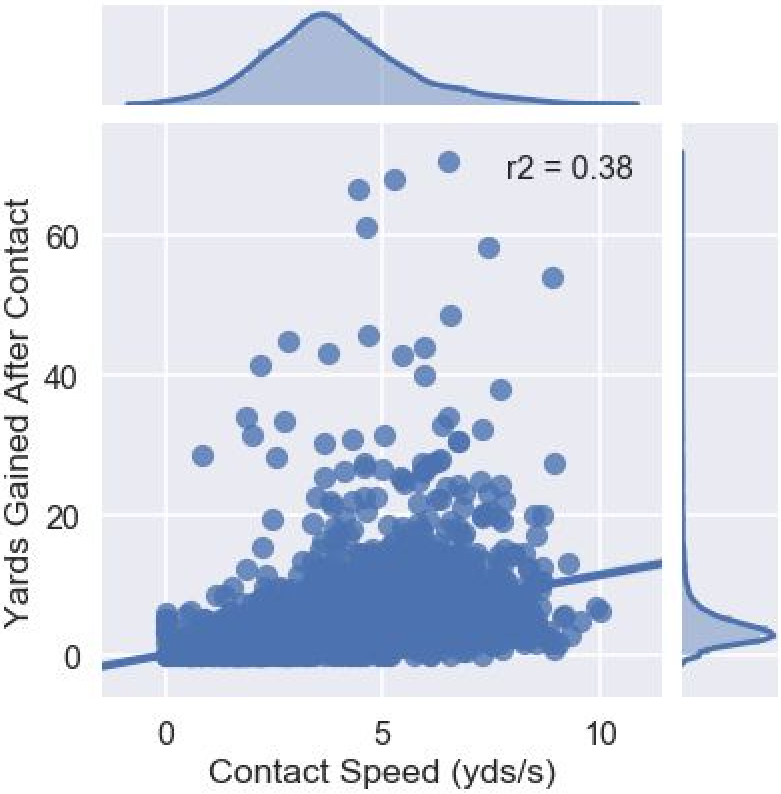


Figure 3: contact speed vs. Yards After Contact

Note that while the correlation is positive, and moderately strong, we still have quite a few observations that clearly exceed the trend. These apparent outliers should not be a surprise. Their presence simply demonstrates that the majority of the time it is difficult for a rusher to escape the defense.

So we know a relationship exists between a ball carrier's contact speed and the yards they gain after contact. But contact speed might be slower and less-representative of what a talented rusher is capable of during a play. To see how else speed might influence the yards gained after contact we will look at the max player speed on a play and compare it to the yards gained after contact.

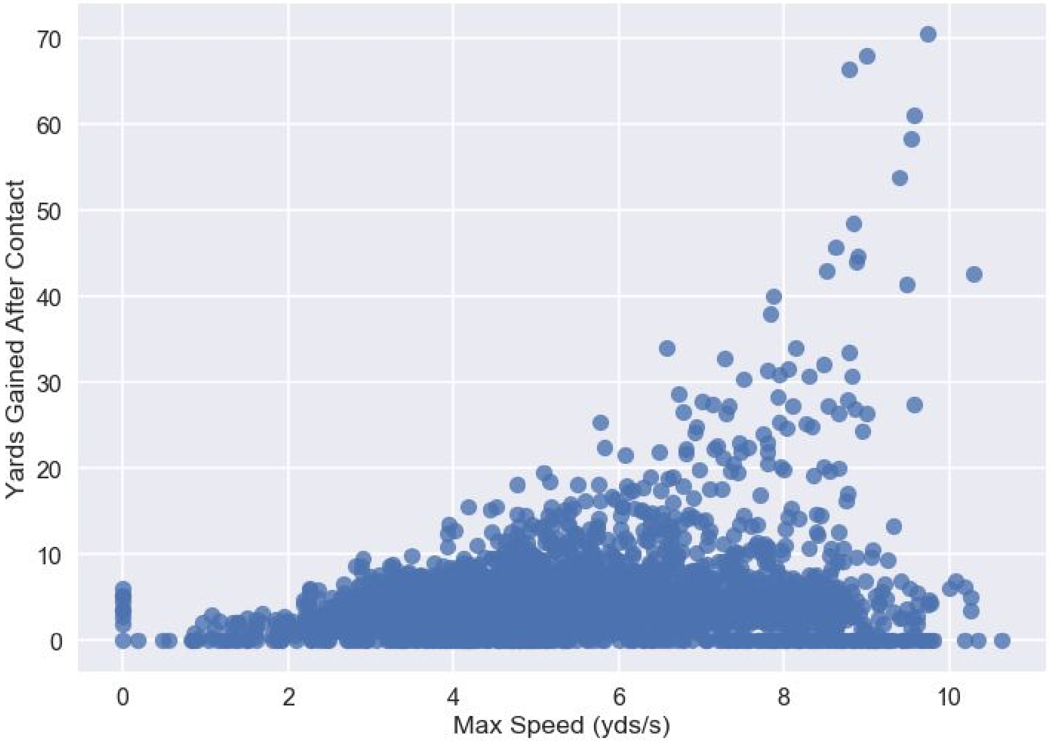


Figure 4: Maximum player speed per play vs yards gained after contact

It is pretty clear to see that there is a trend suggesting that higher max speeds are related to gaining more yards. However some of these observations appear to be rare events, so attempting to model this relationship with a non-linear approach might not prove very useful.

As anticipated, speed appears to be at least partially related to gaining additional yards after contact. However, don’t be confused by this next diagram (figure 5). At first glance this plot seems to be suggesting that the rushing efficiency metric and yards gained after contact do not have a strong association.

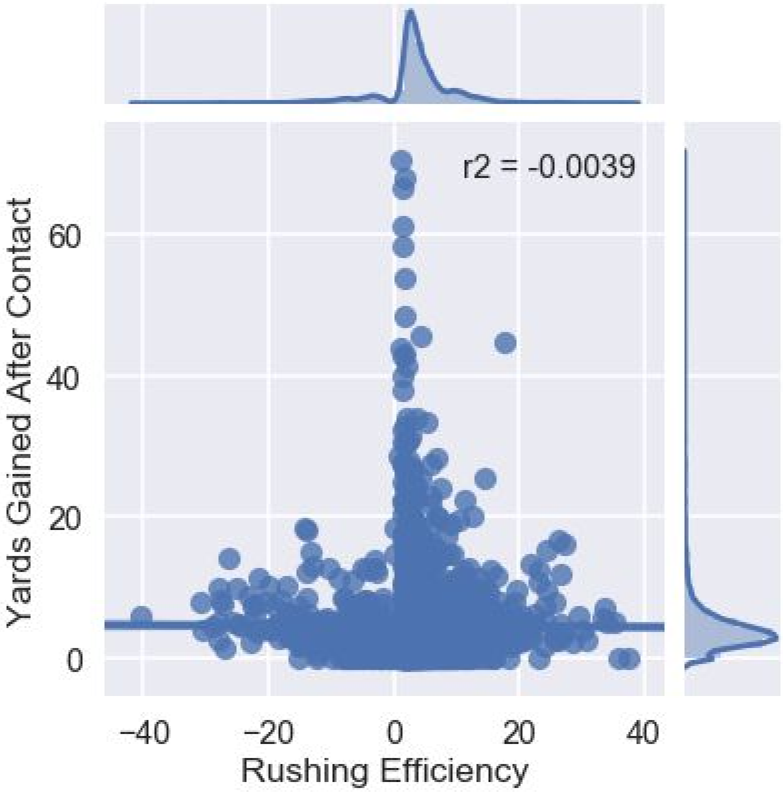


Figure 5: Rushing Efficiency vs Yards After Contact

Looks can be deceptive. The Rushing Efficiency metric actually takes on positive, negative, and zero values. This forces the plot to appear like it does not define a strong relationship between Rushing Efficiency and yards gained after contact. But if you look closer you will see that the diagram is making a very clear claim: players with a Rushing Efficiency value near zero consistently gain yards after contact, while those with Rushing efficiency values far from zero do not gain many yards after contact. In other words efficient runners gain more yards after contact with the defense.

Finally, given we know that contact speed is important, we should examine how contact speed is related to whether or not a run is a success.

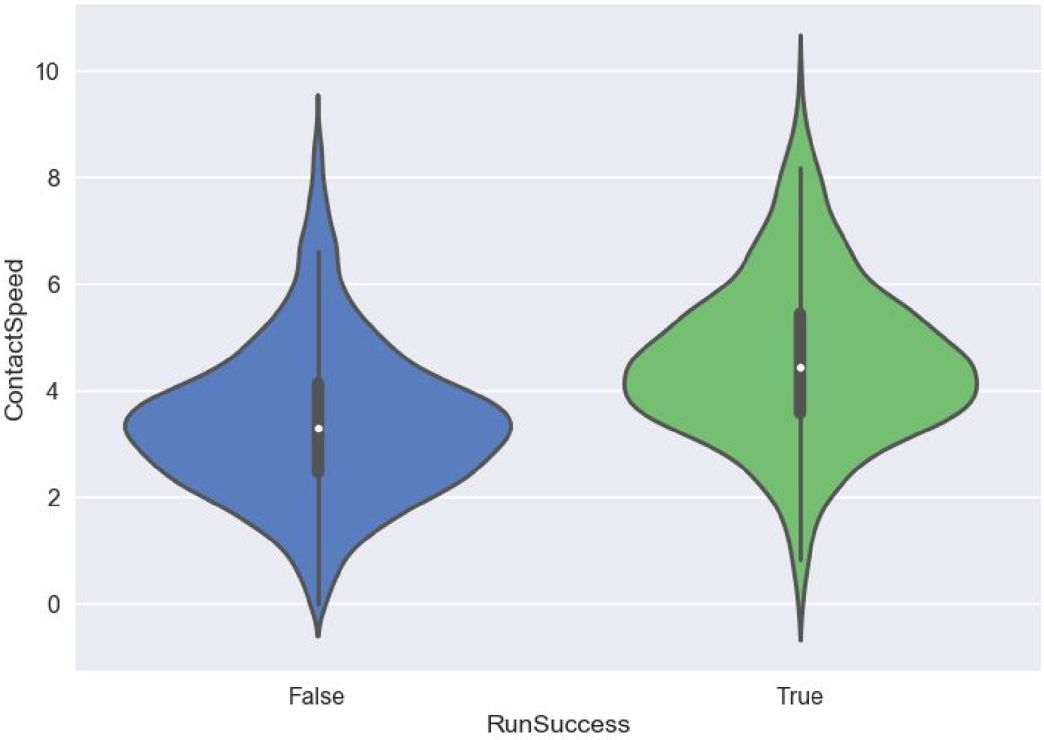


Figure 6: Contact Speed based on Run Success

It is clear to see that, among the observations in this data set, player’s with a successful run have higher contact speeds.

Conclusion and Recommendations

Data should be used to help make better decisions, and discover opportunities to improve efficiency and decision-making in organizations. Given the research agenda and data available in this study it is interesting to see that higher speeds are most definitely correlated with more yards. However, it is far from a perfectly linear relationship. On many plays it is not the highest top speed that matters. Rather, speed varies according to the needs of the situation. This provides at least some support for the idea that player’s try to use their speed in an optimal way whether or not they consciously realize it.

Exploring the relationships in the data suggests that successful runs have a higher contact speed when they meet the defense. That higher contact speed is also correlated with yards gained after contact. It is important to remember that while gaining yards after contact is important, gained yards are simply the result of getting other critical aspects of a run play correct.

We note that max speeds on a play have a somewhat non-linear relationship with yards gained after contact. This could potentially be an indication that teams should run more motion plays (e.g., jet motion) in order to allow their rushers to accelerate to a higher speed prior to moving downfield. In that way pre-snap motion might represent much more than simply deception.

Finally, the data seems to suggest that more efficient runners tend to have successful runs.

Recommendations for Rushers

1. Strive to make runs efficient
2. High contact speed matters
3. Train to get to your max speed quickly

Sources

https://www.evanmiller.org/how-not-to-sort-by-average-rating.html

https://nextgenstats.nfl.com/glossary

Code

https://github.com/bradleynott/NFL-Big-Data-Bowl/blob/master/master.py