A New Metric to Evaluate Defensive Tackles in NFL

Chris Wang

Introduction

My project adapts from a Kaggle competition (NFL Big Data Bowl 2024). “American football is a complex sport, but once an offensive player receives a handoff or catches a pass, all 11 defenders focus on one task – tackle that ball carrier as soon as possible” [1]. This year’s competition focuses on creating a practical and novel metric to evaluate tackling performance of game plays. While tackling is an important element of an NFL football game, little effort has been made in tackling evaluations. Previous tackling statistics only records single-dimensional game summary statistics, such as solely or assisted tackle numbers of each defensive player, sacks, and forced fumbles. Complex evaluation metrics are previously only created for other positions, such as the passing rating to evaluate quarterbacks (QB). Such complex evaluation metrics are hard to create for analyzing tackles because a successful tackle usually involves with several defensive positions other than a single position. Thus, there is no universal standards for the tackle evaluations in a game, and leaves this evaluation open ended.

This year’s competition is mainly built upon the NFL Next Gen Stat data for Week 1-9 of the 2022 NFL season. Each week’s dataset (a csv. file) consists of frame-by-frame tracking data of each player’s field status (such as location, speed, acceleration, team formation.), game situations (such as game clock, field positions, real-time score and win probability of home/visitor’s team.), and play outcomes (such as resulted yardage, penalties, and fouls.). Other supplementary dataset includes player’s information, game’s information, and PFF scouting data.

The tackle evaluation metric I created is called “pivotal factor” and is constructed based on a variety of detailed play statistics, such as game clock, position of the football, and play results etc. The metric reflects the importance of each tackling play and its contribution to the entire game. “Pivotal factor” could be a metric that football fans are most interested and has the potential to holistically evaluate defensive players in terms of their tackling contribution to the team. Table 1 shows some tackling plays with my evaluation based on “pivotal factor”. The table also includes some other statistics of the dataset that are used to calculate the holistic metric.

Method

Preprocess and data wrangle

After reading in the original data, I preprocessed on four main datasets—-play description data, detailed tackle recording data, player's information, and 9 weeks' game tracking data of each player. Four kinds of datasets are linked together by each observation's "nflId", "ballCarrierName", "gameId", and "playId". Since each week's tracking data is very large and cost huge amount of computational resource, I preprocessed each week's data separately before combining them. For each week's data, I grouped them by each game and play, and focused on 26 frames' tracking of 22 on-filed players in any given play. I used two strategies to reduce frame numbers for further analysis. The first filter strategy was to filter out frames irrelevant to tackles by scrutinizing the "event" variable. Defensive tackling process usually starts when an offensive player possesses the football. Thus, for a passing play, only pass outcome frame (ex. a caught pass) was kept and all snap process and passing trajectory frames were filtered out. Other important frames included handoff and run events. This strategy reduced to 2 or 3 frames for each play. The second strategy was to limit my analysis on key players only. I used tackle dataset to identify players who was directly involved in a solely tackle or an assisted tackle and filter out other players. The strategy reduced to only 2 or 3 players' tracking data for each play. After each week's tracking data was reduced in size, I combined them to an integrated tracking file for all 9 weeks. The tracking data was then split into offensive players' dataset and defensive players' dataset based on play description data file that records defensive side and offensive ball carrier. As for this project, only defensive players' tracking data was considered for the sake to avoid complex network analysis between each offensive and defensive player.

Principle Component Analysis (PCA)

The defensive player's tracking data was reduced to nearly 40000 observations of 35 variables. I deployed PCA to explore collinearity between different variables and to summarize covariates' information to reduce dimensionality of the dataset. I either transformed all non-continuous variables into ordinary categorical variables or filtered out discrete variables for later usage, treating as cluster category reference in PC plots. Based on the summary of PCA in Table 2 and standardized variance plot in Figure 1, the first 15 PCs were chosen to reduce variable dimensions because they explained nearly 80% of remaining variance in the covariate matrix.

A remaining question is how these 15 PCs would assist my feature selection process. I further computed each variable's contribution to every PC separately. As in Table 3, some variables related to game situation and play outcomes were identified as relative high contributions to 15 PCs. For clearer visualization, I also plotted biplots of the first 2 PCs against each other and superimposed loadings of each variable on the first 2 PCs. Consistent feature contribution results were shown in Figure 2. Longer arrows mean larger feature contributions. For example, play result, distance to the endzone, and expected points of a play have large contributions to PC2, while play ID and pre-snap home/visitor team scores have large contributions to PC1. Furthermore, the angle between arrows of pre-snap home team and visitor team scores is very small, indicating high collinearity between the two variables. Thus, some simplified metric (like score difference) could later be created to summarize two variables information.

The "pivotal factor" metric

The "pivotal factor" was calculated based on factors that were deemed important features both in terms of feature contribution in Table 3 and the general knowledge of important tackling factors as football fans. In determining the "pivotal factor", I included the tracking of game time left in the game, home/visitor team score difference, current downs, play result, distance to the first down yard line, distance to the endzone, distance between the tackling spot and previous events of a catch, handoff, or run. I assigned weighted scores to each of these variables and summed them up for the final "pivotal factor" of each play.

This section below would provide the detailed calculation of each variable's weight:

1. Game time left in the game: The conclusion of a closely contested game often commands heightened attention from viewers. So, game clock near the end of the game should be given more weights. Only if the play is in fourth quarter, the game clock left in minutes are divided by 2 as the weighted score. And overtime game clock automatically gives 10 credits as the weighted score.

2. Home/visitor team score difference: Usually, a game within two scores would be considered as a close game. So, only when the score is within 14 points, a negative exponential function is created to assign more weights when the score difference is closer to zero.

3. Current down: Each down is given their numeric down number as their scores. For example, the first down contributes 1 credit and the fourth down contributes 4 credits.

4. Play result: Only negative play result is considered defensive gains. So, I apply a log function to the negative play results and multiply the result by 4 to give this element more weights, as play result is deemed to have more contributions to PCs in Figure 2.

5. Distance to the first down: For positive offensive play in third or fourth down, a distance less than 5 yards to the first down yard line would be considered a significant defensive stop. So, an exponential function of (5-yardsToGo) is created for the weighted credits.

6. Distance to the endzone: Since a usual scoring distance (like a field goal) is around 40-yard line, credits of distance to the endzone is only considered when the football moves within the 40-yard line. A negative exponential function is applied to give more weights when the football is closer to goal line.

7. Distance between the tackling spot and previous events of a catch, handoff, or run: A Euclidean distance is calculated for the distance between the tackling spot and the previous event's spot. The distance is then divided by 10 to give the more weighted credits for a defensive player that moves farther to tackle down the ball carrier.

Result

Biplot

Several biplots were generated with different categorical variables to distinguish different clusters of PCA processed dataset as shown in Figure 3-6. The reference PC biplots did not reveal obvious separation patterns across different categories. This was probably due to the diverse nature of a tackle process. A tackle could be made by any teammates who were close to the ball carrier on the field and a successful tackle usually resulted from the entire team's effort on creating pressures and coverages on the offense. Therefore, I did not expect perfectly separated clusters in the PC plots. However, biplots should only be read to analyze minor patterns within the dataset. Figure 3 showed minor separation between tackles made by primary defense positions (such as defensive line and linebacks) and secondary defense positions (such as cornerbacks and safeties). The primary defense were more likely to involve in run game defense, while the secondary defense were more likely to involve in passing defense. Figure 4 showed minor separation between tackles and previous events, indicating a successful tackle did change the game situations. Figure 5 showed a clear separation between shotgun formation and other offensive formations. The shotgun tactic is often characterized by swift and short yardage plays, designed to enhance completion probabilities while accepting the trade-off of potentially limited yardage gains. So, the shotgun formation tended to be associated with a higher tackle rate than other offensive formations. Figure 6 exhibited an interesting finding that each team's tackling pattern spread across the vertical PC other than the horizontal PC. As the philosophy to tackle was generally consistent for each NFL team—-tackling down the nearest ball carrier as soon as possible, each team's tackling pattern did not vary in different play situations.

The "pivotal factor" as a new metric

After computing each play's pivotal factor, I plotted the histogram of each tackling play's pivotal factor in Week 1-9 of 2022 NFL seasons. Figure 7 and 8 supported the validity of the "pivotal factor" metric in distinguishing NFL player's performance. Figure 7 exhibited that the "pivotal factor" histogram was under a left-skewed distribution with a mean around 10. The most pivotal play had a pivotal score over 30. I further compiled the tackle data to represent the average pivotal score a player had across the half season in 2022. Figure 8 showed the histogram of each player's tackling pivotal factor in Week 1-9 of 2022 NFL seasons.

The histogram showed a normal distribution of player's pivotal score with a similar mean around 10. The highest pivotal score was a little over 20. Since Figure 7 and 8 exhibit near normal or normal distributed pivotal factor score, the metric preserved the natural distribution of tackling performance of NFL players. The metric also successfully distinguished tackling performance of different players.

To further check the validity of the new metric, I listed out the top 10 individual play's tackling pivotal score in Table 1. The top play is made by James Smith-Williams who made a tackle for a loss of 4 yards in fourth quarter on 3rd down and 6 yards to the endzone. The play was credited with a high score because its game situation is a close game with nearly nothing left on the clock and the player made a huge play with a sack on a 3rd down for even a loss of 4 yards. A video clip of this play could be found at the "10:52" starting time on YouTube [2]. Table 4 listed the top 10 player's average tackling pivotal score, in which Jason Pierre-Paul performed the best. This validated my metric because Jason Pierre-Paul is a well-known defensive player, and the metric showed the ability to rank up well-known players in defensive teams.

Conclusion

This project addresses the need for a comprehensive and novel metric to evaluate tackling performance in the context of NFL. Even as a fundamental aspect of the game, tackling has lacked a standardized evaluation metric. The newly proposed metric could fill in the gap of the absence of tackling evaluation. The "pivotal factor" metric combines single factors such as game time, team score differences, downs, play results, and spatial relationships on the field, using combined weights to calculate the new metric in terms of player's tackling performance. The metric is further validated with the natural distribution of tackle performance.

One pitfall of this project is that the weight assignment process still involves with human judgment on the importance of a defensive tackle play, although with the help of PCA to reduce covariates related to tackling. Human judgment could lead to bias in reinforcing personal preferences on the standards of tackle evaluation metric. Besides, we should use the metric carefully when making any conclusions related to tackles. The metric solely analyzes player's tackle performance and does not indicate performance of other aspects of defensive performance. One should not extend the metric to holistic defensive evaluation.

There are still much more factors to consider in the NFL Next Gen Stat dataset. In the future, possible enrichment of the tackle evaluation metric could involve force factor analysis. A defensive player's physical condition, speed, acceleration, and angle relative to the ball carrier could be added to the metric to include interactions between each defensive and offensive players in determining tackling success rate.

Reference

1. Kaggle competition description page: https://www.kaggle.com/competitions/nfl-big-data-bowl-2024/data
2. (start from 10:52) https://youtu.be/fuIp98BPoG4?si=SY9yE86Tmg77Mpsl&t=652

Appendix - Figures

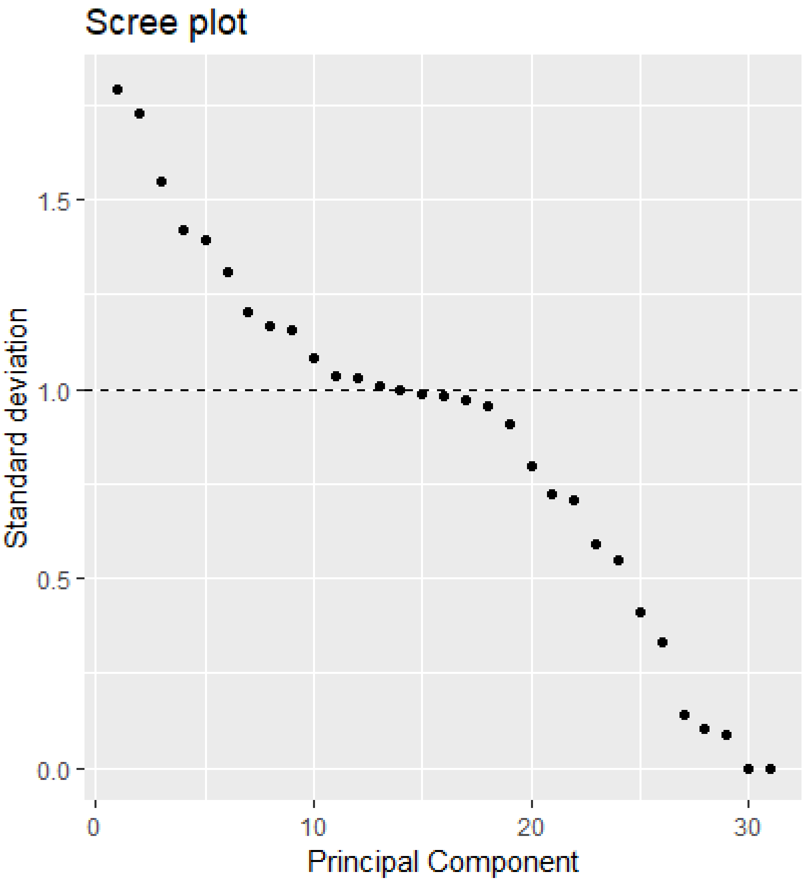


Figure 1: Figure 1

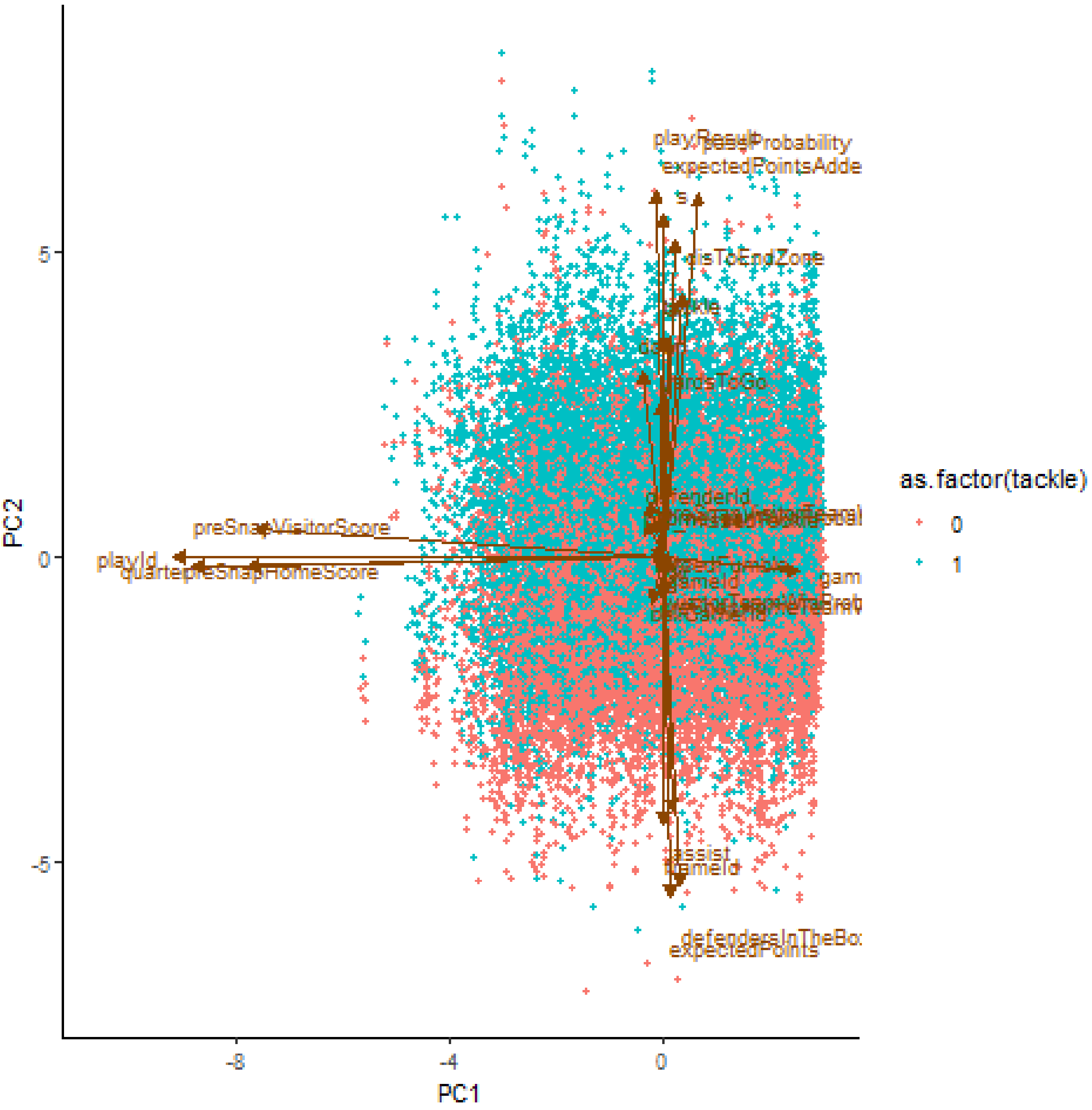


Figure 2: Figure 2

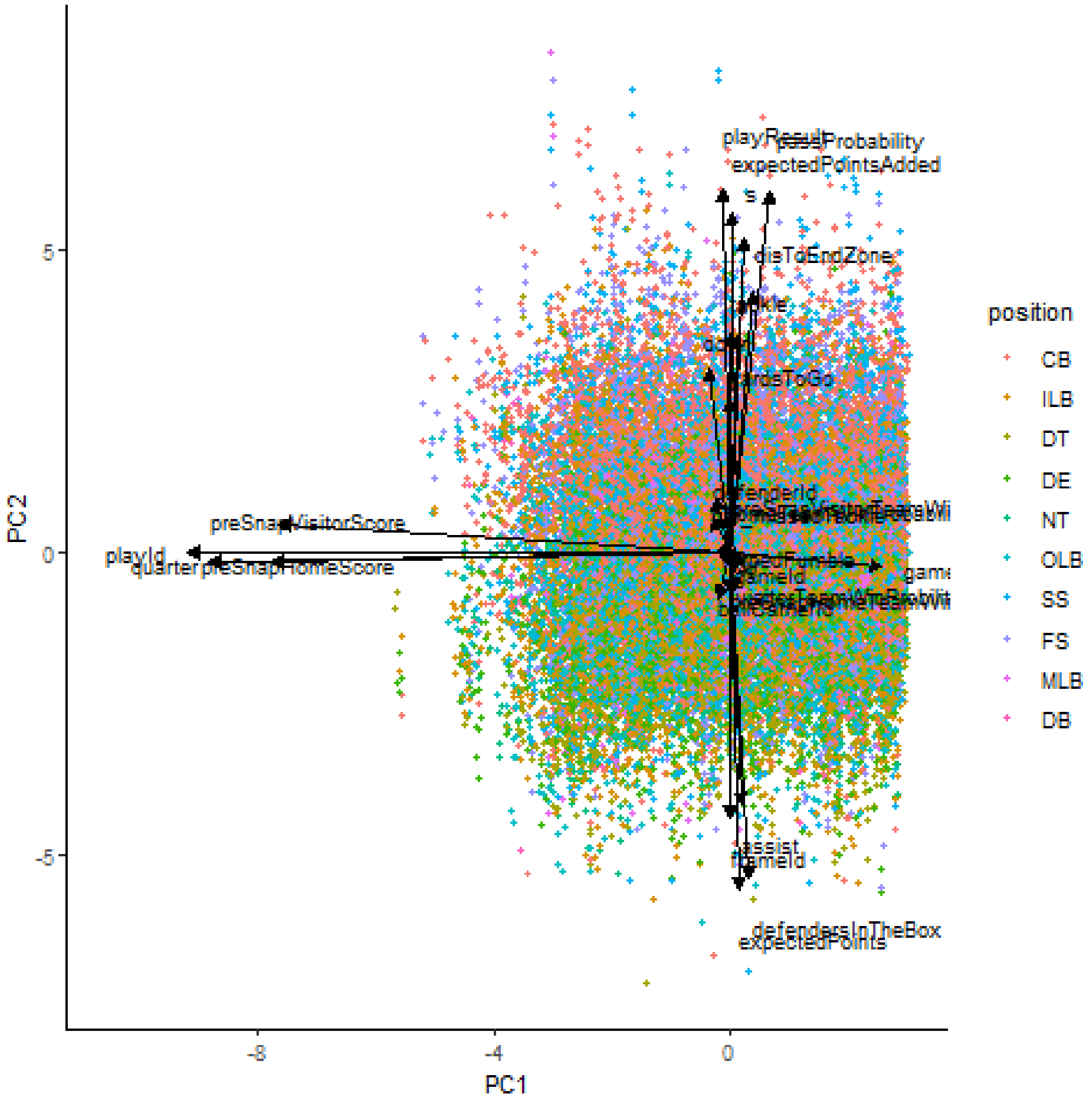


Figure 3: Figure 3

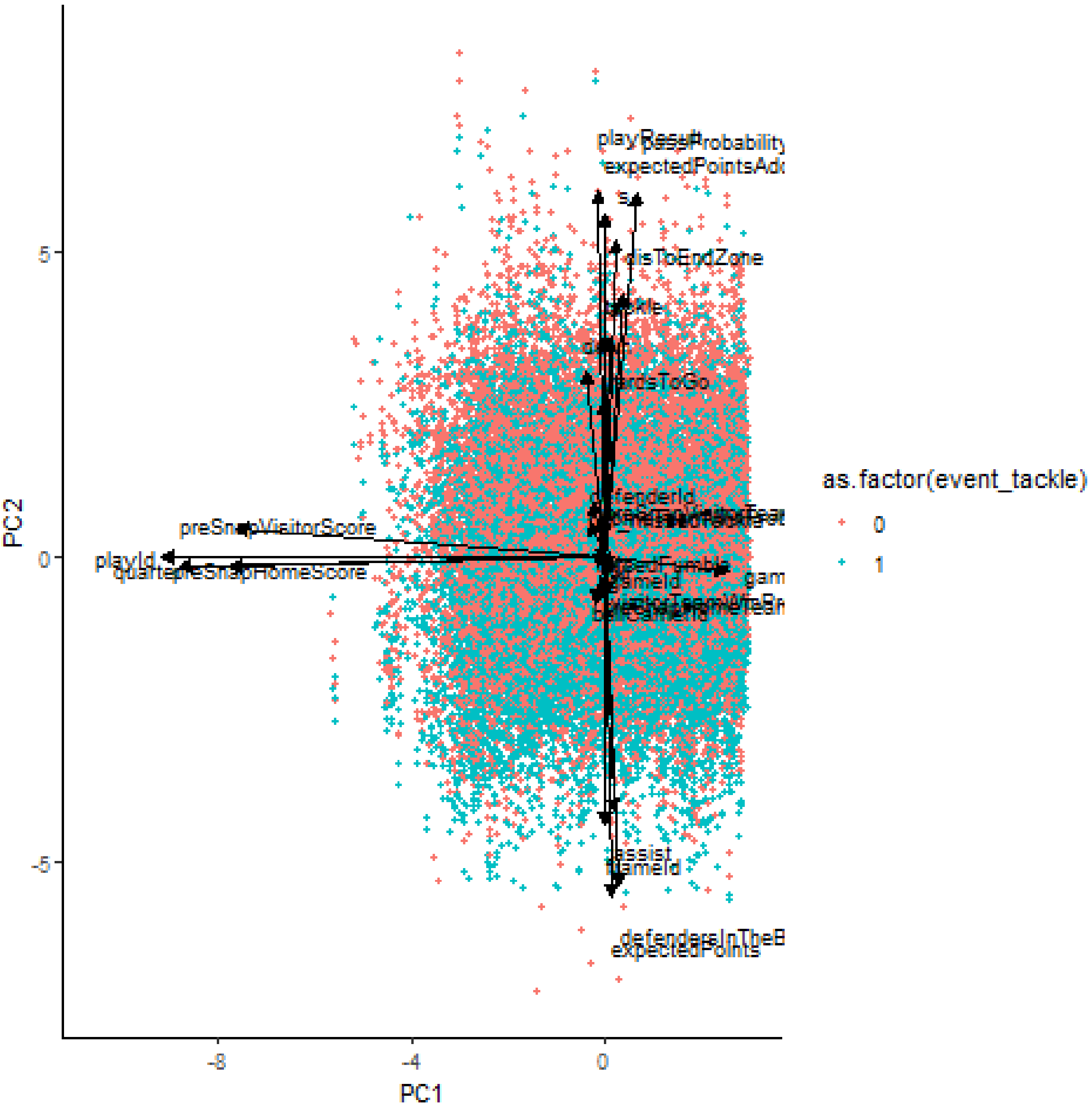


Figure 4: Figure 4

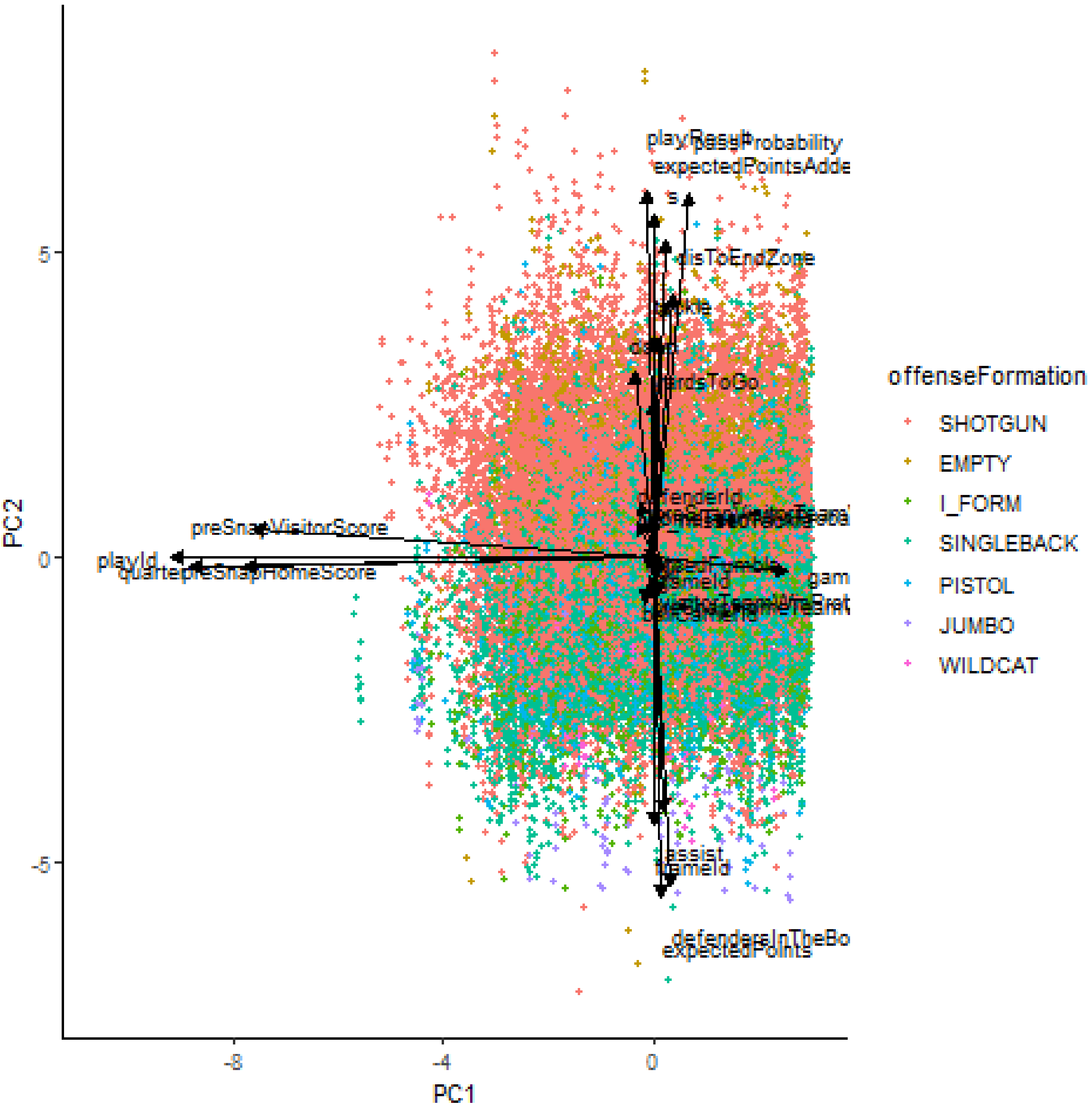


Figure 5: Figure 5

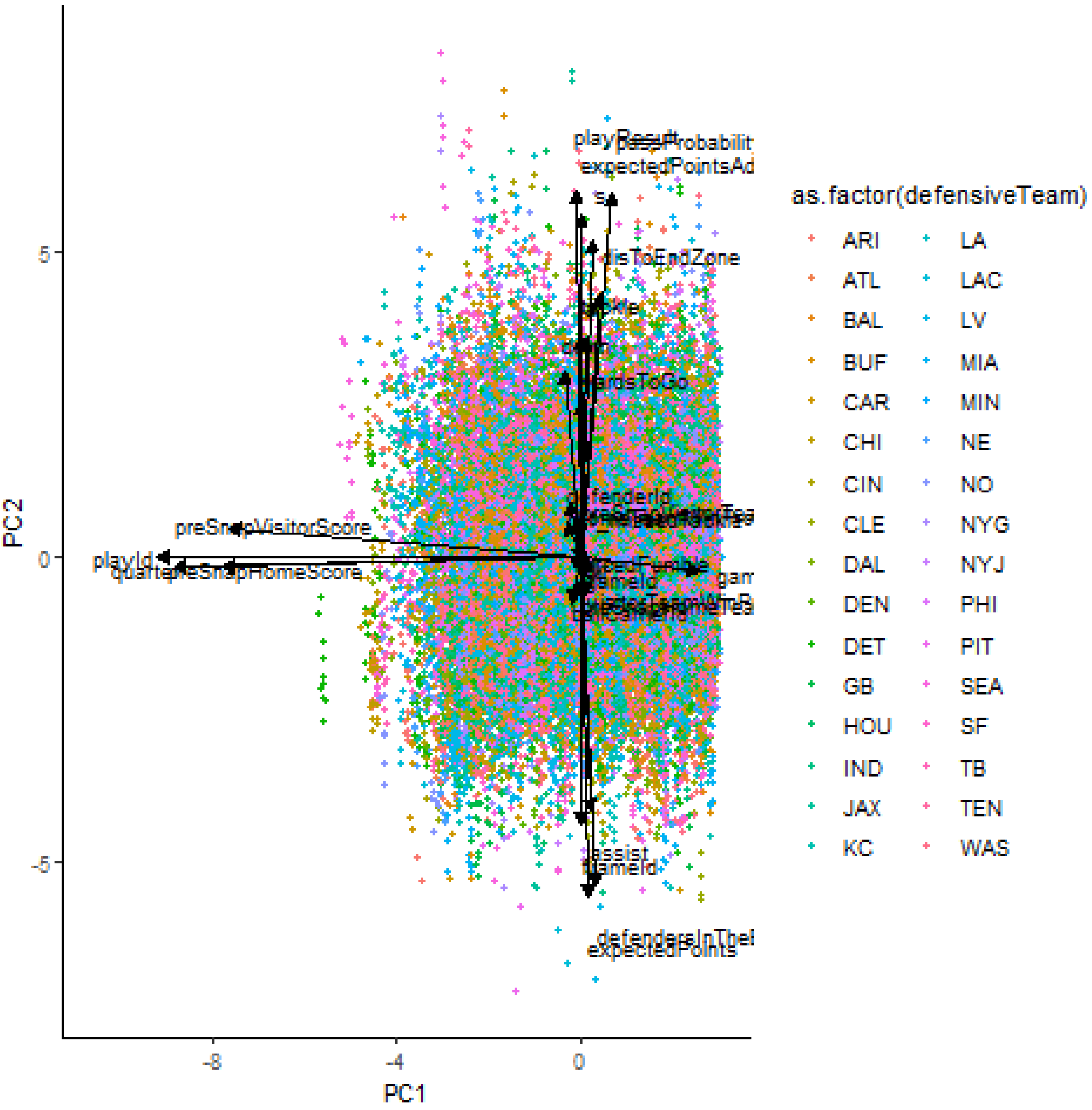


Figure 6: Figure 6

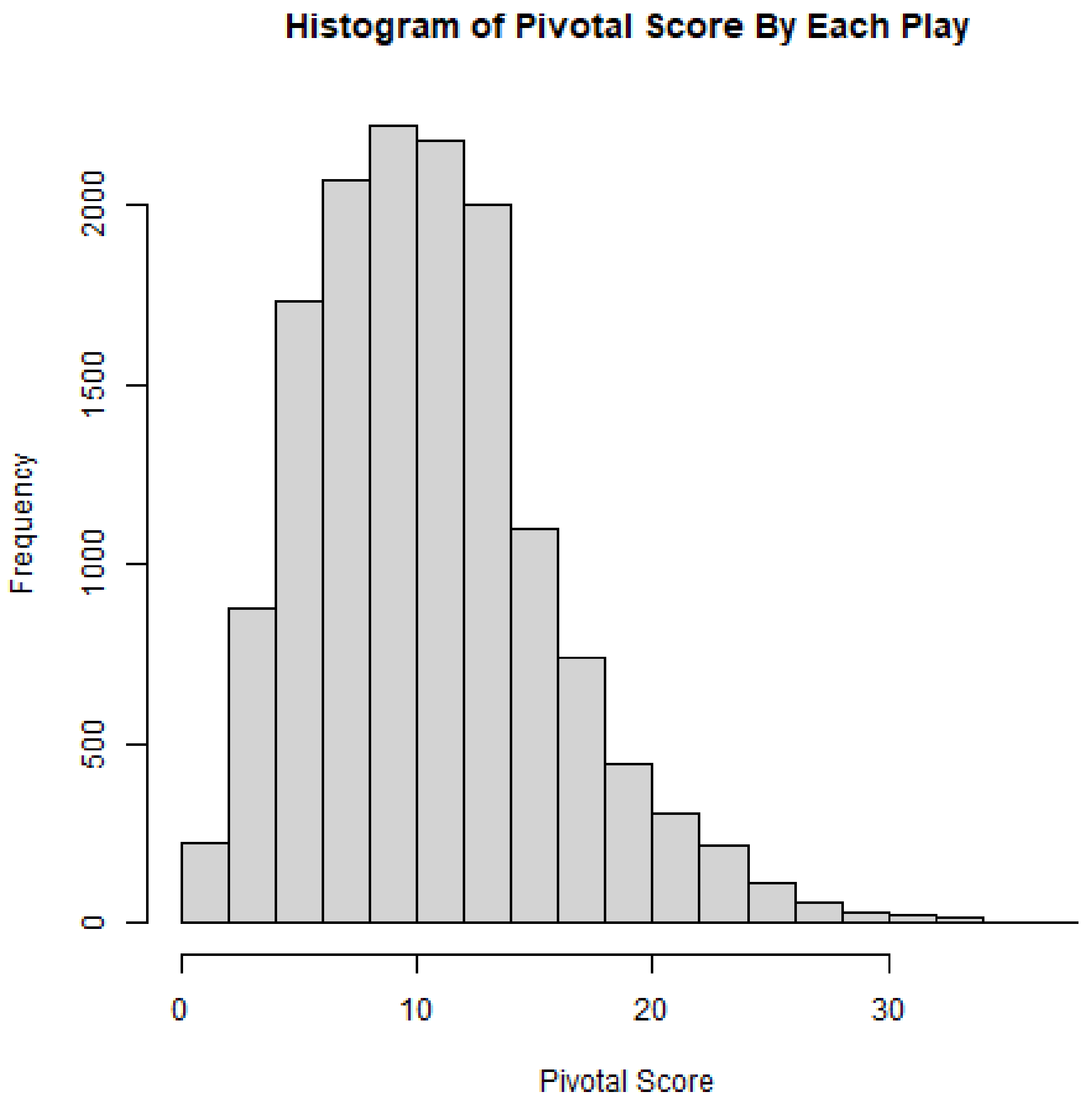


Figure 7: Figure 7

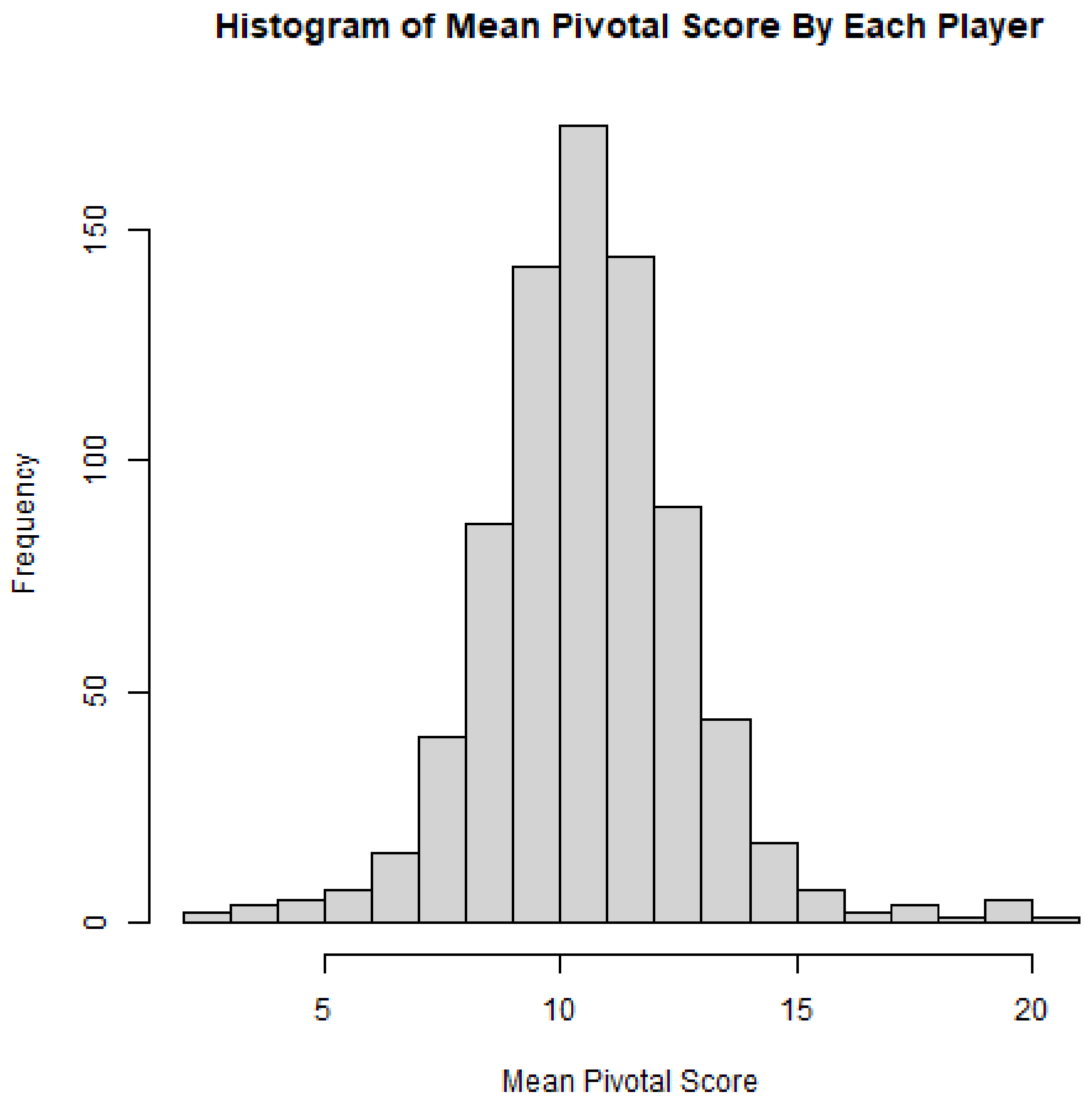


Figure 8: Figure 8

Appendix - Tables

```
# print Table 1
load("output/tab1.rdata")
print(tab4)
```

```
   defenderId         defenderName     ballCarrierName quarter down yardsToGo
1       52637 James Smith-Williams        Kirk Cousins       4    3         6
2       44976          Matt Milano        J.K. Dobbins       4    2         1
3       46082          Daron Payne         Dalvin Cook       4    1         2
4       43373       Kendall Fuller         Dalvin Cook       4    2         4
5       52498    Jonathan Greenard     Jonathan Taylor       5    2         2
6       52624          Kamren Curl         Dalvin Cook       4    2         4
7       47872        Bobby Okereke       Melvin Gordon       5    3         2
8       43296     DeForest Buckner       Melvin Gordon       5    3         2
9       47799          Brian Burns       Rondale Moore       1    4         1
10      46846          Joshua Kalu JuJu Smith-Schuster       5    4         1
   gameClock playResult disToEndZone defensiveTeam               event position
1       3600         -4            6           WAS                 run       DE
2      20460         -3            1           BUF             handoff      OLB
3       6720         -2            2           WAS             handoff       DT
4       6480         -2            4           WAS             handoff       CB
5      10380         -3           16           HOU             handoff       DE
6       6480         -2            4           WAS             handoff       SS
7      11520          1            6           IND             handoff      ILB
8      11520          1            6           IND             handoff       DT
9      27060         -4           10           CAR             handoff      OLB
10     18780          2           13           TEN pass_outcome_caught       FS
   offenseFormation dist_tackle  pivotal tackle_number
1        SINGLEBACK    8.249339 37.31801            17
2             JUMBO    7.260062 35.83616            40
3        SINGLEBACK    3.582583 34.06930            32
4        SINGLEBACK    4.238585 33.89574            28
5           SHOTGUN    8.416579 33.87634             7
6        SINGLEBACK    3.771154 33.84900            35
7           SHOTGUN    6.279252 33.80849            66
8           SHOTGUN    4.580764 33.63864            31
9        SINGLEBACK    8.091508 33.11137            34
10          SHOTGUN    8.712807 32.58613            17
```

```r
# print Table 2
load("output/tab2.rdata")
print(pc_tab)
```

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     1.7940 1.72732 1.54881 1.42153 1.39458 1.31030 1.20293
Proportion of Variance 0.1038 0.09625 0.07738 0.06519 0.06274 0.05538 0.04668
Cumulative Proportion  0.1038 0.20007 0.27745 0.34263 0.40537 0.46075 0.50743
                           PC8     PC9   PC10    PC11    PC12    PC13    PC14
Standard deviation     1.16715 1.15889 1.0854 1.03793 1.03052 1.00697 0.99574
Proportion of Variance 0.04394 0.04332 0.0380 0.03475 0.03426 0.03271 0.03198
Cumulative Proportion  0.55137 0.59470 0.6327 0.66745 0.70171 0.73442 0.76640
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.98794 0.98214 0.96963 0.95537 0.91077 0.79848 0.72039
Proportion of Variance 0.03148 0.03112 0.03033 0.02944 0.02676 0.02057 0.01674
Cumulative Proportion  0.79789 0.82900 0.85933 0.88877 0.91553 0.93610 0.95284
                          PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.70832 0.58864 0.54628 0.41158 0.32908 0.13821 0.10410
Proportion of Variance 0.01618 0.01118 0.00963 0.00546 0.00349 0.00062 0.00035
Cumulative Proportion  0.96902 0.98020 0.98983 0.99529 0.99879 0.99940 0.99975
                          PC29      PC30      PC31
Standard deviation     0.08789 2.866e-15 7.967e-16
Proportion of Variance 0.00025 0.000e+00 0.000e+00
Cumulative Proportion  1.00000 1.000e+00 1.000e+00
```

```r
# print Table 3
load("output/tab3.rdata")
print(tab3)
```

```
                                        Dim.1        Dim.2        Dim.3
gameId                           2.160049e-03 2.918905e-02 1.718708e-02
playId                           2.918664e+01 9.297647e-05 8.337409e-03
frameId                          1.515534e-05 6.631928e+00 3.023905e-03
x                                1.099690e-02 1.441111e-02 4.906424e-02
y                                4.658977e-04 7.678995e-05 5.316865e-02
s                                2.046940e-02 9.281567e+00 4.591657e-03
a                                3.686028e-03 4.393064e+00 6.378280e-04
o                                8.292433e-05 3.645307e-04 6.780855e-03
```

```
dir                              3.230032e-03 6.768203e-03 6.094307e-05
tackle                           5.901342e-05 4.472036e+00 5.705505e-02
assist                           1.172193e-02 5.979275e+00 1.725317e-02
forcedFumble                     7.583233e-03 1.393880e-03 2.908677e-03
pff_missedTackle                 2.776680e-02 1.163542e-01 3.703918e-02
ballCarrierId                    1.179095e-02 1.955355e-01 9.880198e-04
quarter                          2.693226e+01 7.522861e-03 1.423448e-02
down                             4.656213e-02 3.189576e+00 1.069499e-01
yardsToGo                        1.661490e-04 2.216921e+00 2.878889e-01
gameClock                        2.283260e+00 1.848981e-02 9.316183e-03
preSnapHomeScore                 2.095024e+01 9.363466e-03 8.104831e+00
preSnapVisitorScore              2.021565e+01 7.974619e-02 8.422218e+00
playResult                       5.124674e-03 1.243494e+01 9.163246e-02
defendersInTheBox                3.640371e-02 9.980517e+00 2.169100e-02
passProbability                  1.599986e-01 1.221715e+01 2.265045e-02
preSnapHomeTeamWinProbability    7.692050e-05 1.498794e-01 4.034149e+01
preSnapVisitorTeamWinProbability 7.692050e-05 1.498794e-01 4.034149e+01
homeTeamWinProbabilityAdded      2.360632e-03 1.298806e-01 1.841103e-01
visitorTeamWinProbilityAdded     2.360632e-03 1.298806e-01 1.841103e-01
expectedPoints                   7.066214e-03 1.065338e+01 7.179059e-01
expectedPointsAdded              1.099006e-04 1.085563e+01 2.688238e-02
defenderId                       2.119982e-02 2.788713e-01 4.773878e-02
disToEndZone                     5.041733e-02 6.376304e+00 8.167538e-01
                                        Dim.4        Dim.5        Dim.6
gameId                           1.424737e-01 5.967747e-03 7.950300e-02
playId                           5.114878e-03 1.530775e-03 3.250074e-03
frameId                          1.930775e+00 3.013098e+00 3.712754e-01
x                                1.284406e-02 4.416189e-02 3.192506e-03
y                                5.979632e-02 6.071603e-03 5.833270e-03
s                                1.497089e+00 2.585604e+00 4.920847e-02
a                                7.062282e-01 1.027595e+00 2.654489e-02
o                                2.155948e-02 2.873992e-04 2.001919e-03
dir                              5.109758e-02 2.716343e-04 2.628850e-03
tackle                           5.744630e+00 1.659294e+01 1.339776e+01
assist                           3.646413e+00 1.208004e+01 2.269280e+00
forcedFumble                     1.387690e-01 1.442679e-01 1.468010e+00
pff_missedTackle                 9.250328e-01 1.621937e+00 1.294260e+01
ballCarrierId                    1.288424e-01 3.114233e-02 6.518374e-02
quarter                          8.994906e-04 1.051662e-02 9.482206e-03
down                             5.650868e-01 4.895485e-01 1.786684e+00
yardsToGo                        1.241667e+00 3.933154e+00 4.410003e+00
gameClock                        7.829846e-02 3.753351e-01 1.348606e-01
preSnapHomeScore                 6.127096e-04 5.126126e-02 2.452586e-05
```

```
preSnapVisitorScore              7.979143e-03 2.084381e-01 4.545728e-02
playResult                       3.091641e-01 3.684127e-01 1.773152e+01
defendersInTheBox                2.463543e-06 5.633876e-01 3.040807e-02
passProbability                  8.904694e-02 3.801909e-02 1.029322e-01
preSnapHomeTeamWinProbability    2.927051e-04 7.781544e-01 4.160727e-02
preSnapVisitorTeamWinProbability 2.927051e-04 7.781544e-01 4.160727e-02
homeTeamWinProbabilityAdded      3.695662e+01 1.228425e+01 7.823958e-02
visitorTeamWinProbilityAdded     3.695662e+01 1.228425e+01 7.823958e-02
expectedPoints                   3.797109e+00 1.452830e+01 9.555432e+00
expectedPointsAdded              1.510079e-01 4.455401e-02 2.315303e+01
defenderId                       5.184904e-02 9.067698e-02 1.555939e-01
disToEndZone                     4.782795e+00 1.601868e+01 1.195861e+01
                                        Dim.7        Dim.8        Dim.9
gameId                           2.918537e-03 2.490361e-02 1.733455e-02
playId                           1.842150e-02 1.441938e-03 7.578114e-04
frameId                          4.195756e+00 8.766838e-02 8.437577e-01
x                                9.090582e-02 2.356736e-01 3.247122e-02
y                                5.182828e-02 6.582142e-02 2.167514e-02
s                                1.553591e-01 5.894432e-02 1.488767e+00
a                                1.850715e-01 2.064484e-02 9.520263e-01
o                                1.409455e-02 4.968739e+01 1.722513e-02
dir                              2.346828e-02 4.938708e+01 2.964110e-02
tackle                           7.059447e+00 3.371014e-02 9.311004e-01
assist                           1.643833e+01 9.795885e-02 1.819384e-01
forcedFumble                     1.866355e+00 1.769513e-02 1.549343e-01
pff_missedTackle                 3.652716e+00 3.521173e-02 8.145207e-01
ballCarrierId                    4.293937e-01 5.375930e-02 6.209668e-01
quarter                          2.379882e-02 1.571457e-04 6.206696e-02
down                             1.606304e+01 1.902578e-02 3.152962e+01
yardsToGo                        6.474651e-01 3.730934e-02 3.911174e+01
gameClock                        1.905541e+00 4.150773e-02 1.025666e+00
preSnapHomeScore                 2.357420e-06 1.998670e-04 1.343335e-04
preSnapVisitorScore              8.876661e-02 1.253549e-03 4.818313e-03
playResult                       7.274440e+00 1.350897e-03 4.642547e-02
defendersInTheBox                1.238388e+01 1.360455e-02 5.389222e+00
passProbability                  1.779223e+01 1.192499e-03 1.731002e+00
preSnapHomeTeamWinProbability    1.097543e-01 1.930210e-03 1.070532e-02
preSnapVisitorTeamWinProbability 1.097543e-01 1.930210e-03 1.070532e-02
homeTeamWinProbabilityAdded      1.891776e-01 3.185753e-02 5.665065e-04
visitorTeamWinProbilityAdded     1.891776e-01 3.185753e-02 5.665065e-04
expectedPoints                   2.393921e-01 1.540527e-03 8.032778e+00
expectedPointsAdded              5.173631e+00 1.977906e-03 7.832273e-01
defenderId                       4.600152e-04 1.284811e-04 1.181792e+00
```

```
disToEndZone                     3.625426e+00 5.275041e-03 4.971851e+00
                                       Dim.10       Dim.11       Dim.12
gameId                           1.651672e-01 4.929185e+01 5.299489e-01
playId                           6.662900e-04 3.019026e-02 2.582164e-02
frameId                          1.122691e+01 3.305182e-01 9.274677e+00
x                                2.013900e-01 3.615648e+00 1.949588e-03
y                                1.277613e-04 3.959765e-01 7.009866e+00
s                                1.286535e+01 4.240599e-01 9.500306e+00
a                                6.730427e+00 4.515615e-01 3.291526e+00
o                                1.430472e-01 4.765968e-02 1.517696e-02
dir                              9.414581e-02 8.560643e-02 4.255667e-03
tackle                           1.089957e+00 3.650316e-01 3.096834e+00
assist                           1.740664e+01 1.903612e-03 9.133586e-01
forcedFumble                     7.974650e-02 1.075645e+00 3.012237e+01
pff_missedTackle                 2.217669e+01 1.081005e+00 1.874102e+01
ballCarrierId                    6.498413e-01 3.286699e+01 6.960446e-01
quarter                          2.141593e-01 1.054967e-01 2.277918e-02
down                             2.612085e+00 5.957488e-01 6.033624e-02
yardsToGo                        3.387871e-01 8.486884e-02 8.704592e-01
gameClock                        3.316434e+00 3.803931e-01 4.655242e-05
preSnapHomeScore                 3.968062e-02 5.710236e-02 5.237330e-04
preSnapVisitorScore              5.184714e-02 2.786665e-05 2.649475e-03
playResult                       1.843711e+00 5.105688e-02 6.002774e-01
defendersInTheBox                8.489787e+00 4.468553e-03 2.806588e+00
passProbability                  7.242712e+00 2.294850e-01 1.546686e+00
preSnapHomeTeamWinProbability    1.758467e-02 1.375307e-03 3.059246e-03
preSnapVisitorTeamWinProbability 1.758467e-02 1.375307e-03 3.059246e-03
homeTeamWinProbabilityAdded      5.011970e-03 5.081746e-02 3.284239e-05
visitorTeamWinProbilityAdded     5.011970e-03 5.081746e-02 3.284239e-05
expectedPoints                   4.901855e-03 1.173558e-01 4.997954e-01
expectedPointsAdded              1.801596e+00 3.525747e-03 2.745254e+00
defenderId                       7.389550e-01 8.184373e+00 6.840275e+00
disToEndZone                     4.300463e-01 1.806073e-02 7.749969e-01
                                       Dim.13       Dim.14       Dim.15
gameId                           8.829001e-01 5.039139e-01 1.158998e+00
playId                           8.716071e-03 1.561554e-05 1.371185e-02
frameId                          1.091926e+00 7.090263e-01 5.157542e+00
x                                5.362892e+01 7.177391e+00 2.665537e+01
y                                1.168809e+01 7.717395e+01 4.477356e-02
s                                2.439074e-01 6.877236e-01 4.041735e+00
a                                1.010184e+00 7.595473e-01 1.453889e+00
o                                2.083017e-02 2.377844e-02 1.270938e-02
dir                              3.650973e-01 1.348575e-01 1.007966e-01
```

```
tackle                           9.098416e-04 9.475113e-02 1.452453e-01
assist                           3.421880e-06 3.194286e-02 8.638538e-03
forcedFumble                     1.733424e-01 7.369527e+00 1.892471e+00
pff_missedTackle                 1.103966e-03 6.115906e-01 1.969077e-01
ballCarrierId                    3.515311e+00 7.076628e-01 1.544168e+01
quarter                          2.747060e-01 5.753797e-03 7.153638e-02
down                             1.746207e-02 3.446639e-02 1.286005e-01
yardsToGo                        2.816986e-02 1.646830e-01 1.257190e-04
gameClock                        3.831111e+00 6.995751e-02 2.710430e+00
preSnapHomeScore                 1.192433e-02 7.346600e-03 4.458703e-05
preSnapVisitorScore              8.657465e-04 3.584546e-05 9.351582e-03
playResult                       1.191342e-02 7.350975e-02 2.750002e-02
defendersInTheBox                3.333170e-01 9.000588e-02 1.128501e-01
passProbability                  2.024239e-01 3.266620e-01 2.699390e-01
preSnapHomeTeamWinProbability    5.472172e-03 2.893745e-02 2.028189e-02
preSnapVisitorTeamWinProbability 5.472172e-03 2.893745e-02 2.028189e-02
homeTeamWinProbabilityAdded      8.512509e-04 3.318479e-02 1.166420e-02
visitorTeamWinProbilityAdded     8.512509e-04 3.318479e-02 1.166420e-02
expectedPoints                   2.104993e-01 9.380660e-03 6.739313e-02
expectedPointsAdded              2.975282e-06 9.979476e-03 1.539204e-02
defenderId                       2.215906e+01 3.040136e+00 4.008306e+01
disToEndZone                     2.746555e-01 5.816351e-02 1.154197e-01
```

```
# print Table 4
load("output/tab4.rdata")
print(tab5)
```

```
         defenderName position  pivotal tackle_number
1   Jason Pierre-Paul       DE 17.85443             5
2        Byron Cowart       DT 17.62271             5
3   Jonathan Greenard       DE 17.10609             7
4         Krys Barnes      ILB 15.35040             7
5   Marshon Lattimore       CB 15.26449             6
6  Jonathan Ledbetter       DE 14.86540             6
7        John Jenkins       NT 14.59066             8
8      Dre'Mont Jones       DE 14.18291            28
9   Kayvon Thibodeaux      OLB 14.09371            11
10      Demone Harris       DE 13.98518             5
```

Appendix - Code

Preprocess and data wrangle

```r
library(tidyverse)
library(readr)
library(factoextra)
library(lubridate)
```

```r
# Read in data
tracking_week1 <- read_csv("Data/tracking_week_1.csv")
tracking_week2 <- read_csv("Data/tracking_week_2.csv")
tracking_week3 <- read_csv("Data/tracking_week_3.csv")
tracking_week4 <- read_csv("Data/tracking_week_4.csv")
tracking_week5 <- read_csv("Data/tracking_week_5.csv")
tracking_week6 <- read_csv("Data/tracking_week_6.csv")
tracking_week7 <- read_csv("Data/tracking_week_7.csv")
tracking_week8 <- read_csv("Data/tracking_week_8.csv")
tracking_week9 <- read_csv("Data/tracking_week_9.csv")
games <- read_csv("Data/games.csv")
plays <- read_csv("Data/plays.csv")
players <- read_csv("Data/players.csv")
tackles <- read_csv("Data/tackles.csv")
```

```r
# Combine tracking data to play data, but need to preprocess each week's tracking data fir
# before combing them all, because the combined data is too large to process

# Add in player names to tackles data
tackles <- tackles |> left_join(players, by = c("nflId" = "nflId")) |> select(-c(height, w

# There are many different kinds of events, but we are only interested in the ones that ar
# to tackles (handoff, pass_outcome_caught, run, tackle)
tracking_week1 <- tracking_week1 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  # add in tackles data to tracking data
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  # add plays data to include ball carrier
  left_join(plays, by = c("gameId", "playId")) |>
  # only keep tracking data for the ball carrier and tackler
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week2 <- tracking_week2 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week3 <- tracking_week3 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week4 <- tracking_week4 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week5 <- tracking_week5 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week6 <- tracking_week6 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week7 <- tracking_week7 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week8 <- tracking_week8 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)

tracking_week9 <- tracking_week9 |> group_by(gameId, playId) |>
  filter(event %in% c("pass_outcome_caught", "handoff", "run", "tackle")) |>
  filter(displayName != "football") |>
  select(gameId, playId, displayName, frameId, time, x, y, s, a, o, dir, event) |>
  left_join(tackles, by = c("gameId", "playId", "displayName")) |>
  left_join(plays, by = c("gameId", "playId")) |>
  filter(!is.na(nflId) | displayName==ballCarrierDisplayName)
```

```
# Load pre-saved environment when running on local machine, in order to save computation t
if (TRUE){
  combined_variables <- read_rds("Data/combined_variables.rds")
  tracking_combined <- combined_variables$tracking_combined
  tracking_defense <- combined_variables$tracking_defense
  tracking_defense_new <- combined_variables$tracking_defense_new
  tracking_offense <- combined_variables$tracking_offense
  pca_defense <- combined_variables$pca_defense
  players <- combined_variables$players
  plays <- combined_variables$plays
  games <- combined_variables$games
} else{
  # Combine 9 weeks tracking data and summarize to mutate several features
  tracking_combined <- bind_rows(tracking_week1, tracking_week2, tracking_week3, tracking_
                                 tracking_week5, tracking_week6, tracking_week7, tracking_
                                 tracking_week9) # combine all weeks of tracking data
  tracking_combined <- tracking_combined |>
    mutate(defenderId = nflId,
           event = as.factor(event),
           position = as.factor(position),
           offenseFormation = as.factor(offenseFormation)) |>
    mutate(disToEndZone = ifelse(
      possessionTeam == yardlineSide, 100 - yardlineNumber, yardlineNumber))

  # split combined data into defense and offense players
  tracking_defense <- tracking_combined |>
    select(-c(displayName, nflId, ballCarrierDisplayName, playDescription, yardlineSide,
              yardlineNumber, passResult, passLength, penaltyYards, prePenaltyPlayResult,
              playNullifiedByPenalty, absoluteYardlineNumber, foulName1, foulName2, foulNF
              foulNFLId2)) |>
    filter(!is.na(tackle))

  tracking_offense <- tracking_combined |>
    # anti_join(tracking_defense)
    select(-c(displayName, nflId, ballCarrierDisplayName, playDescription, yardlineSide,
              yardlineNumber, passResult, passLength, penaltyYards, prePenaltyPlayResult,
              playNullifiedByPenalty, absoluteYardlineNumber, foulName1, foulName2, foulNF
              foulNFLId2)) |>
    select(-c(tackle, assist, forcedFumble, pff_missedTackle, position)) |>
    # mutate(defenderId = nflId) |>
    filter(is.na(defenderId))
}
```

PCA on combined tracking data

```
# PCA implementation for defense players
if (!exists("pca_defense")){
  tracking_defense_new <- tracking_defense |> select(-c(time, possessionTeam, defensiveTea
                                                        position, offenseFormation)) |>
  mutate(gameClock = as.numeric(gameClock)) |>
  na.omit()
  pca_defense = prcomp(tracking_defense_new, center = TRUE, scale. = TRUE)
}
pc_tab <- summary(pca_defense)
pca_var_p <- qplot(1:(ncol(tracking_defense_new)), pca_defense$sdev) +
  geom_hline(yintercept = 1, linetype = 2) +
  xlab("Principal Component") +
  ylab("Standard deviation") +
  ggtitle("Scree plot")
```

```
Warning: `qplot()` was deprecated in ggplot2 3.4.0.
```

```
# Use first 15 PCs for explaining nearly 80% of variance and arrive at standard deviation
```

```
#####################################################################################
# Save the working environment
#####################################################################################

# temporarily save environment
# save pca_defense, tracking_combined
# Create a list containing the variables
variables_list <- list(pca_defense = pca_defense, tracking_combined = tracking_combined,
                       tracking_defense = tracking_defense, tracking_offense = tracking_of
                       tracking_defense_new = tracking_defense_new,
                       players = players, plays = plays, games = games)

# Save the list to a single RDS file
saveRDS(variables_list, "Data/combined_variables.rds")
```

Biplot

```
tracking_defense_new$PC1 <- pca_defense$x[,1]
tracking_defense_new$PC2 <- pca_defense$x[,2]

# Get arrow end point locations (loadings*17 for effect)
l.x <- pca_defense$rotation[,1]*17
l.y <- pca_defense$rotation[,2]*17

# Get label positions (%15 further than end of arrows)
l.posx <- l.x*1.15
l.posy <- l.y*1.15

# Get labels for plot (variable names)
```

```
l.labels <- row.names(pca_defense$rotation)

# Plot the biplot with superimposed feature importance (loadings)
bip1 <- ggplot() +
  geom_point(data = tracking_defense_new, aes(PC1, PC2, color = as.factor(tackle)), size =
  geom_segment(aes(x=0, y=0, xend = l.x, yend = l.y),
               arrow = arrow(length = unit(0.2, "cm"), type = "closed"), color = "darkoran
  geom_text(aes(x = l.posx, y = l.posy, label = l.labels),
            size = 3, hjust = 0, color = "darkorange4") + # labels
  theme_classic()

# Get variable contributions to PCs
tab3 <- get_pca_var(pca_defense)$contrib[,1:15]
```

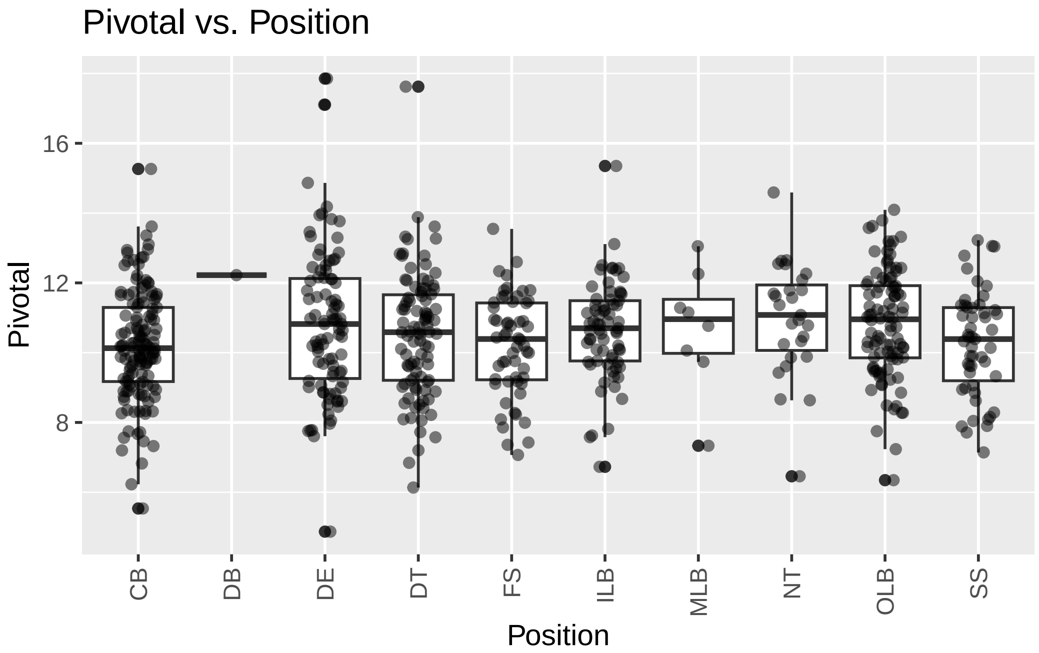
```
# Add categorical variable back to pca dataset and analyze biplot by different categorical
tracking_defense_new <- tracking_defense_new |>
  left_join(tracking_defense %>% select(gameId, playId, defenderId, frameId, possessionTea
                                        defensiveTeam, event, position, offenseFormation),
            by = c("gameId", "playId", "defenderId", "frameId"))


## biplot by position
bip2 <- ggplot() +
  geom_point(data = tracking_defense_new, aes(PC1, PC2, color = position), size = 1) +
  geom_segment(aes(x=0, y=0, xend = l.x, yend = l.y), arrow = arrow(length = unit(0.2, "cm
  geom_text(aes(x = l.posx, y = l.posy, label = l.labels), size = 3, hjust = 0) +
  theme_classic()

## biplot by event (tackle or not)
tracking_defense_new <- tracking_defense_new |>
  mutate(event_tackle = ifelse(event == "tackle", 1, 0))

bip3 <- ggplot() +
  geom_point(data = tracking_defense_new, aes(PC1, PC2, color = as.factor(event_tackle)),
  geom_segment(aes(x=0, y=0, xend = l.x, yend = l.y), arrow = arrow(length = unit(0.2, "cm
  geom_text(aes(x = l.posx, y = l.posy, label = l.labels), size = 3, hjust = 0) +
  theme_classic()

## biplot by offenseFormation
bip4 <- ggplot() +
  geom_point(data = tracking_defense_new, aes(PC1, PC2, color = offenseFormation), size =
  geom_segment(aes(x=0, y=0, xend = l.x, yend = l.y), arrow = arrow(length = unit(0.2, "cm
  geom_text(aes(x = l.posx, y = l.posy, label = l.labels), size = 3, hjust = 0) +
  theme_classic()

## biplot by defensiveTeam
bip5 <- ggplot() +
  geom_point(data = tracking_defense_new, aes(PC1, PC2, color = as.factor(defensiveTeam)),
  geom_segment(aes(x=0, y=0, xend = l.x, yend = l.y), arrow = arrow(length = unit(0.2, "cm
  geom_text(aes(x = l.posx, y = l.posy, label = l.labels), size = 3, hjust = 0) +
  theme_classic()
```

Creating metrics for defensive player's tackling contribution

```
# Get the distance between the tackle spot and the spot of previous event
tracking_defense_new <- tracking_defense_new |>
  mutate(x_tackle = ifelse(event == "tackle", x, NA),
         y_tackle = ifelse(event == "tackle", y, NA))
tackle_position <- tracking_defense_new |>
  group_by(gameId, playId, defenderId) |>
  summarise(x_tackle = max(x_tackle, na.rm = TRUE),
            y_tackle = max(y_tackle, na.rm = TRUE))
tracking_defense_new <- tracking_defense_new |>
  select(-c(x_tackle, y_tackle)) |>
  left_join(tackle_position, by = c("gameId", "playId", "defenderId")) |>
  mutate(dist_tackle = ifelse(event!="tackle", sqrt((x_tackle - x)^2 + (y_tackle - y)^2),
  filter(dist_tackle != Inf)

# Construct pivot factor dataframe
pivot <- tracking_defense_new |>
  # gameClock credit, 0.5 for each minute, 7.5 for max credit in 4th quarter, 10 for overt
  mutate(credit_gameClock = case_when(
    quarter == 4 ~ (54000-gameClock)/(3600*2),
    quarter == 5 ~ 10,
    TRUE ~ 0
  )) |>
  # score differential credit, max credit is 10 for 0 score differential
  mutate(credit_scoreDiff = case_when(
    abs(preSnapHomeScore - preSnapVisitorScore) <= 14 ~ 10*exp(-abs(preSnapHomeScore - pre
    TRUE ~ 0
  )) |>
  # down credit, max credit is 4 for a 4th down tackle
  mutate(credit_down = case_when(
    down == 4 ~ 4,
    down == 3 ~ 3,
    down == 2 ~ 2,
    down == 1 ~ 1,
    TRUE ~ 0
  )) |>
  # play result credit, max credit is 4*log(100)=16 for a tackle for loss of 100 yards, on
  mutate(credit_playResult = ifelse(playResult < 0, 4*log(-playResult), 0)) |>
  # distance to endzone credit, max credit is 2*exp(39/20)=14 for a tackle at 1 yardline
  mutate(credit_disToEndZone = ifelse(disToEndZone <= 40, 2*exp((40-disToEndZone)/20), 0))
  # distance to 1st down credit, max credit is exp(5/2.5)=7 for a tackle resulting inches
  mutate(credit_disTo1stDown = ifelse(
    (yardsToGo-playResult) > 0 & yardsToGo <=5 & down != 1 & down != 2,
    exp((5-yardsToGo)/3), 0)) |>
  # distance to tackle spot credit, max credit is 2*exp(39/20)=14 for a tackle at 1 yardli
  mutate(credit_distTackle = dist_tackle/10) |>
  mutate(pivotal = credit_gameClock + credit_scoreDiff + credit_down + credit_playResult +
           credit_disToEndZone + credit_disTo1stDown + credit_distTackle)
```

Analyze the metric validity

```
# a <- pivot |>
#   group_by(defenderId) |>
#   mutate(tackle_number = n())
# a[a$defenderId == unique(a$defenderId),]
# a[1,]

pivot <- pivot |>
  group_by(defenderId) |>
  mutate(tackle_number = n())
# |>
#   ungroup() |>
#   select(gameId, playId, tackle, assit, forcedFumble, pff_missedTackle, ballCarrierId, q

pivot_perTackle <- pivot |>
  left_join(players |> select(nflId, displayName), by = c("defenderId" = "nflId")) |>
  mutate(defenderName = displayName) |> select(-displayName) |>
  left_join(players |> select(nflId, displayName), by = c("ballCarrierId" = "nflId")) |>
  mutate(ballCarrierName = displayName) |> select(-displayName) |>
  select(defenderName, ballCarrierName, quarter, down, yardsToGo, gameClock, playResult, d

pivot_perPlayer <- pivot |>
  group_by(defenderId) |>
  summarise(pivotal = mean(pivotal), tackle_number = n()) |>
  left_join(players |> select(nflId, displayName, position), by = c("defenderId" = "nflId"
  mutate(defenderName = displayName) |> select(-displayName, -defenderId) |>
  select(defenderName, position, pivotal, tackle_number)

tab4 <- pivot_perTackle |> filter(tackle_number >=5) |> select(-defenderId)|> arrange(desc
tab5 <- pivot_perPlayer |> filter(tackle_number >=5) |> arrange(desc(pivotal)) |> head(10)

# plot player position vs. pivotal
pivot_perPlayer |> filter(tackle_number >=5) |>
  ggplot(aes(x = position, y = pivotal)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, height = 0, alpha = 0.5) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  labs(title = "Pivotal vs. Position", x = "Position", y = "Pivotal")
```



```
# save the plots

# png("output/fig1.png")
# pca_var_p
# dev.off()

png("output/fig2.png")
bip1
dev.off()

png("output/fig3.png")
bip2
dev.off()

png("output/fig4.png")
bip3
dev.off()

png("output/fig5.png")
bip4
dev.off()

png("output/fig6.png")
bip5
dev.off()

# plot histograms of pivotal factor
png("output/fig7.png")
hist(pivot$pivotal, breaks = 20, main = "Histogram of Pivotal Score By Each Play", xlab =
dev.off()

png("output/fig8.png")
hist(pivot |> group_by(defenderId) |> summarise(mean_pivotal = mean(pivotal)) |> pull(mean
dev.off()


# save the data into rdata
save(tab4, file = "output/tab1.rdata")
save(tab5, file = "output/tab4.rdata")
save(pc_tab, file = "output/tab2.rdata")
save(tab3, file = "output/tab3.rdata")
```