NFL Defensive Performance Through Principal Component Analysis and Regression

David Splane

Department of Mathematics

California Polytechnic State University

Spring 2023

APPROVAL PAGE

This report is being submitted as my senior project in the Department of Mathematics.

TITLE: NFL Defensive Performance Through Principal Component Analysis and Regression

AUTHOR: David Jamal Splane

ADVISOR: Dr. Joyce Lin

DATE SUBMITTED: June 13, 2023

Dr. Joyce Lin	
Senior Project Advisor	Signature
Ben Richert	
Mathematics Department Chair	Signature

TABLE OF CONTENTS

Introduction: 3

Methods: 5

Analysis and Results 12

Conclusion 20

References 21

INTRODUCTION: Preface

This project develops a new way to analyze defensive matchups in a more subjective way. Although we are in an analytic field where concrete details base our analyses, we analyze humans and their performance, so introducing a subjective lens is necessary. Our goal is to build a system in which we can fairly apply weights to situations based on players' efficiencies in different aspects of the game. Working as a part-timer at ProFootballFocus (PFF) introduced me to the processes in which players are given grades for their performance. For a system that garners so much traffic in the media, their grading system is quite simple. From their website, "Each player is given a grade of -2 to +2 in 0.5 increments on a given play with 0 generally being the average or 'expected' grade." As analysts in a field where concrete details are the foundation of our analyses, we recognize the importance of introducing a subjective lens when it comes to analyzing human performance. Our goal is to develop a system that can accurately and fairly apply weight to different situations based on a player's efficiency in various aspects of the game.

Through this project, we aim to take this grading system to the next level by incorporating additional data and data science techniques to gain a comprehensive understanding of a player's performance. By building upon PFF's system and introducing more sophisticated methods, we hope to develop a more nuanced and accurate system that can provide valuable insights for players, coaches, and analysts alike. Ultimately, our goal is to contribute to the ongoing conversation about how best to evaluate and assess the performance of NFL players.

We'll also take this in different directions, finding relationships between things like player height, draft measurements, and other player attributes and performance data.

Another important aspect of the project is investigating the pairwise relations between the various statistics that are available for NFL defensive players. By identifying these relationships, we'll be able to gain a deeper understanding of how these statistics and components relate to each other and how they contribute to a player's overall performance – and the variability between players, from an analytical standpoint. This will allow us to identify the key statistical attributes that are most important for a player to be “successful” in their position and to develop a more comprehensive understanding of the world of NFL defensive players.

Further, we wish to accomplish is to use our analyses to develop a ranking system that can identify the statistical attributes that most heavily contribute to a player's success on the field. When we say success on the field, we are mainly talking about having higher volume in the “better stats” and lower volume of the “worse stats,” which we also aim to establish analytically (it’s easy to tell which stats are more desirable for a player to have more of less of, but we hope to convey this judgment in our conclusion). By identifying these key attributes, we develop a more accurate and comprehensive evaluation of players' performance, which can inform decision-making for coaches, analysts, and team managers alike. Ultimately, that our rankings provide valuable insights into the characteristics that distinguish successful players from the rest, and contribute to a more data-driven approach to player evaluation in the NFL.

METHODS

We begin with a simple, small example and will extend this to the full dataset in Part 2.

Part 1:

Our first dataset we use includes each NFL player and their attributes of interest. Eleven attributes in particular were used: targets, completions allowed, passer rating allowed, yards per completion, yards per target, interceptions, touchdowns, DADOT (defensive average depth of target), air yards, yards after catch allowed, and blitzes.[22]

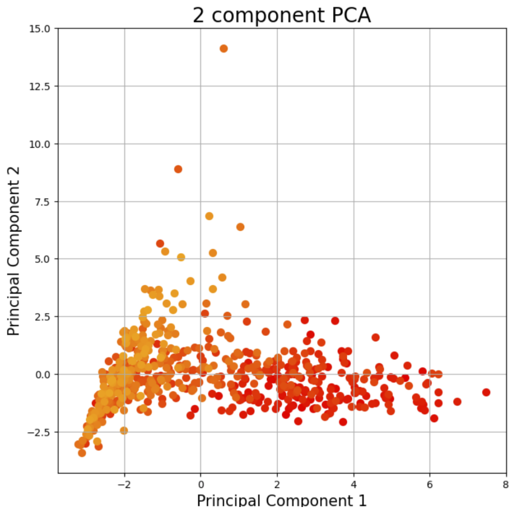


Figure 1: This graph shows the first (x-axis) and second (y-axis) principal components where each point is an observation given by the baseline dataset. Note the two directions that the graph tends to.

We perform principal component analysis on this dataset containing these performance statistics to determine the most critical factors. Running a PCA with 11 components (to match the number of attributes we are analyzing) and labeling principal components lets us create a dataframe the same length as the original, meaning each row or observation corresponds to a player. We concatenate this frame with the column of the initial dataset that holds each player's position (Cornerback, safety, linebacker, etc.). Plotting the first two components for each observation, we get the plot seen above (Figure 1). The scatter plot for the first principal component versus the second principal component seems to go in two directions, and we can illustrate these directions by actually plotting the eigenvectors next to each other to show direction shown to the left (Figure 2).

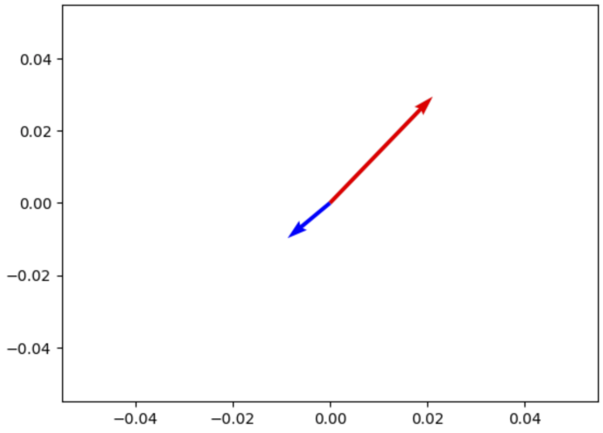


Figure 2: This graph shows the directions of the first two eigenvectors of the principal component analysis result on the baseline dataset.

In the analysis, we look at the precise weights or contributions from each component in the dataset. In preparing for the second part of this project, we extract data from Pro Football Reference– specifically, we use the heights, weights and approximate values of the players at hand.

Part 2:

Here we aim to construct a more concrete ranking and a new way to view the statistics.

We first rank the different data forms (i.e. grades vs. on-field stats vs. measurables) based on the results of principal component analysis (PCA). This ranking will provide us with insights into which data types contribute most significantly to our overall evaluation and decision-making process.

To do so, we use datasets from Pro Football Reference and ProFootballFocus for two of our sets, where we receive further measurements like height, weight and approximate value from Pro Football Reference. After manipulating the datasets appropriately for principal component analysis (PCA), we create a loose ranking based on variability captured by the first three components of the PCA result. We obtain three results corresponding to our three dataframes at hand.

To expedite this analysis, we manipulated and cleaned the dataframes during the initial phase. In part one of our process, we prepared the datasets, ensuring compatibility and consistency across hundreds of NFL players. Despite the variations in measurables within each frame, we were able to adjust the data, allowing for meaningful comparisons and analysis. We go through these outputs in the Analysis and Results

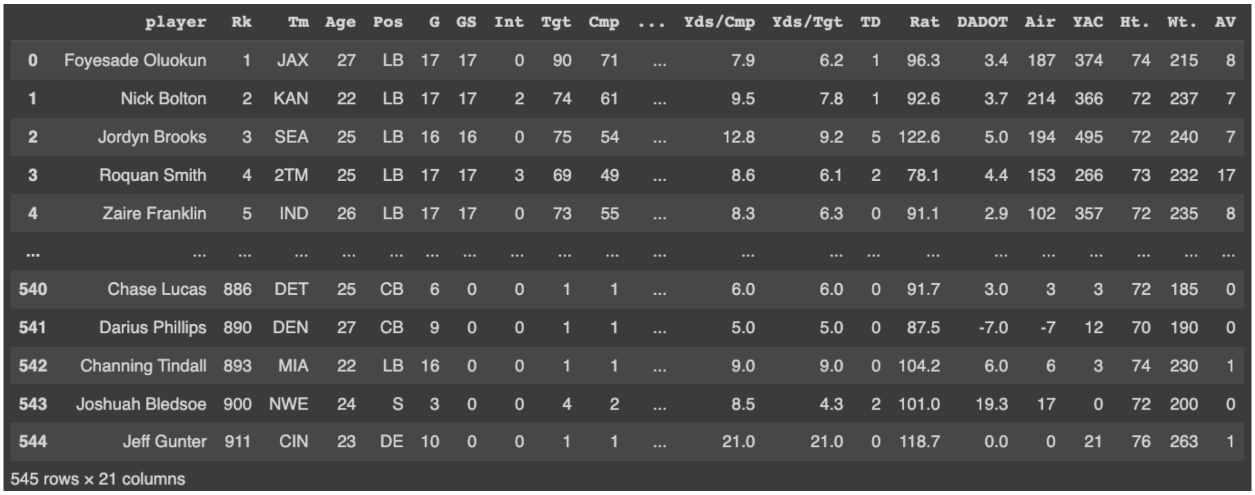
Through this thorough data manipulation and cleaning, we have established a foundation for the next steps. Prepped with comprehensive datasets and a solid understanding of the relative significance of different data forms, we are now well-equipped to go deeper into our evaluation and make informed decisions based on the insights gained from this process.

There were several steps taken to clean up and prepare the dataframes appropriately. One of the primary actions executed on the datasets involved eliminating columns that represented statistics contingent on other available data. For instance, if we already possessed information on a player's total yards and total catches given up, it became unnecessary to retain the column indicating yards per catch allowed for each player. By removing such redundant columns, we streamlined the data and minimized unnecessary duplication, bolstering the efficiency and clarity of our analysis.

A further transformation applied to the data involved converting the players' heights from apostrophe notation to a unified measurement system, such as total inches. This modification enabled standardized height representation and simplified consistent comparisons and computations. Alternatively, we could have chosen to record the height in total centimeters, meters, or any other universally accepted unit of measurement.

Now that we have prepped our datasets, we can move on with correlations, regressions, and making new statistic(s) and measurements that will help garner a refreshing ranking among NFL defensive players.

We start the next part of our process by plotting some pairplots for the datasets that show us visually how certain components behave given other components. We first run a pairplot of the characteristics Height, Weight, and Approximate Value (Table 1). We do this because we want to see a linear relationship between height and weight – something we should assume are highly correlated. Seeing the result that we expect in this small pairplot would give us confidence in using the pairplots to compare our datasets with a greater number of components (plotting pairplots is a method that has inefficient time complexity, so we don't want to keep guessing and checking whether what we're doing is useful as it would waste time).



	player	Rk	Tm	Age	Pos	G	GS	Int	Tgt	Cmp	...	Yds/Cmp	Yds/Tgt	TD	Rat	DADOT	Air	YAC	Ht.	Wt.	AV
0	Foyesade Oluokun	1	JAX	27	LB	17	17	0	90	71	...	7.9	6.2	1	96.3	3.4	187	374	74	215	8
1	Nick Bolton	2	KAN	22	LB	17	17	2	74	61	...	9.5	7.8	1	92.6	3.7	214	366	72	237	7
2	Jordyn Brooks	3	SEA	25	LB	16	16	0	75	54	...	12.8	9.2	5	122.6	5.0	194	495	72	240	7
3	Roquan Smith	4	2TM	25	LB	17	17	3	69	49	...	8.6	6.1	2	78.1	4.4	153	266	73	232	17
4	Zaire Franklin	5	IND	26	LB	17	17	0	73	55	...	8.3	6.3	0	91.1	2.9	102	357	72	235	8
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
540	Chase Lucas	886	DET	25	CB	6	0	0	1	1	...	6.0	6.0	0	91.7	3.0	3	3	72	185	0
541	Darius Phillips	890	DEN	27	CB	9	0	0	1	1	...	5.0	5.0	0	87.5	-7.0	-7	12	70	190	0
542	Channing Tindall	893	MIA	22	LB	16	0	0	1	1	...	9.0	9.0	0	104.2	6.0	6	3	74	230	1
543	Joshuah Bledsoe	900	NWE	24	S	3	0	0	4	2	...	8.5	4.3	2	101.0	19.3	17	0	72	200	0
544	Jeff Gunter	911	CIN	23	DE	10	0	0	1	1	...	21.0	21.0	0	118.7	0.0	0	21	76	263	1

545 rows × 21 columns

Table 1: This table shows the modified dataset with components from the baseline dataset accompanied by further advanced measurements like Approximate Value (AV), height and weight.

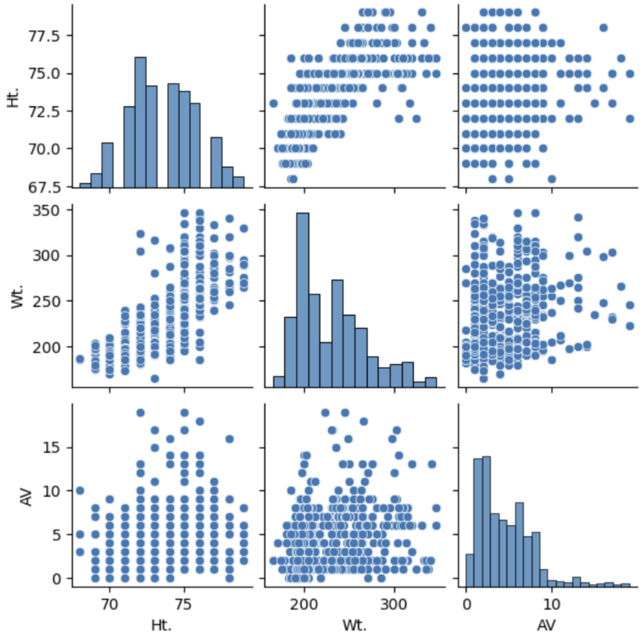


Figure 3: This figure shows the pairplots generated for pairs of attributes for height, weight, and AV. Note the linear patterns in the Ht. vs. Wt. and Wt. vs. Ht. plots.

As we see from the output, we have a linear relationship between height and weight, save for some expected variability due to the volume of our dataset (Figure 3). We maintain this newfound confidence to proceed with analyzing the pairplots of our two main datasets.

So we move on and implement this process once again, but with our dataset we've generated from ProFootballFocus. We keep this in mind when we dive into the analysis of our results.

These next few steps aim to give us some context for the correlation coefficients for each pair of components in order to decide which components would be the most appropriate or the most useful when creating a regression to predict the efficiency values.

First, we standardize each of the dataframes we are dealing with. Although standardization doesn't affect the correlation coefficients, it will help our regression construction process go more smoothly. For each dataframe at hand, we look at a heat map that maps to our correlation matrix; next, we choose a threshold for the correlation coefficients that would be considered to be "too correlated," so we eliminate said coefficients with a value of over 0.75. When building our predictive model, we include most of the components and then retrospectively eliminate the components that are highly correlated to another component. When choosing, we want to prioritize keeping the most "basic" statistic

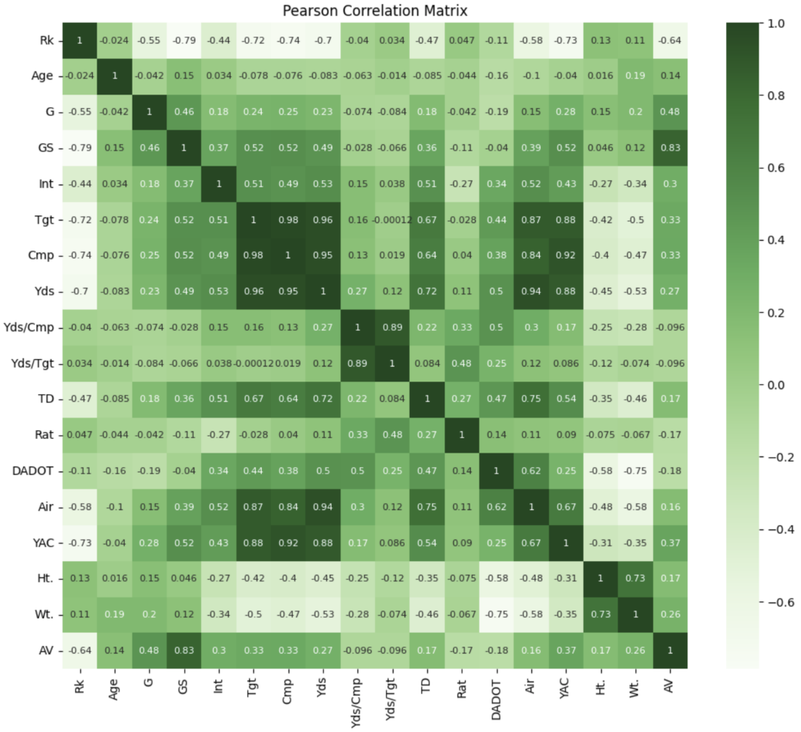


Figure 4: This heat map shows us each attribute of the dataset and how correlated they are by displaying each pair's correlation coefficient along with a color, where the color's saturation is given by the magnitude of the coefficient.

in that group of correlated components; in other words, the component Yards was highly correlated (> 0.75) with the components Targets, Completions, Air Yards, and Yards After Catch (Figure 4). In this case, we want to keep the Yards component for reasons that boil down to how the number of yards given up already include all of said highly correlated qualities:

- Yards (given up) could be a result of being the thrown at (targets, completions) at a high volume, and
- Yards given up, in essence, are Air Yards + Yards After Catch.

Wrapping this part of the process up, we run a regression on the smaller dataframe containing the PFF grading components. This model predicts the overall defensive grade of a player given his pass rush grade, run defense grade, pass coverage grade, and tackle grades. We obtain the p-values and coefficients that contribute to the regression, and achieve something that mimics this predictive equation:

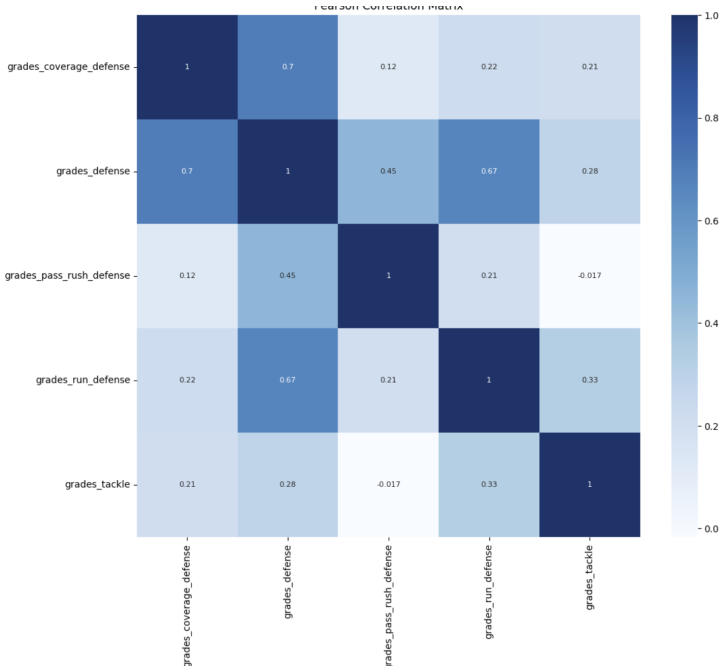


Figure 5: Same type of heat map as figure 4, but for the PFF dataset. Notice the table's symmetry along the diagonal.

```
predicted = [coverage defense]*0.5818 + [run defense]*0.4514
+[tackling]*0.007 + [pass
rush]*0.3283 - 21.9203
```

A new column is then added to the dataframe at hand that shows the predicted value based on the regression, and a second column is then added that shows the player's residual (actual value minus predicted). We have arrived at something that shows how a player has performed given his attributes, up to PFF's standards via their grading components. We plot the distribution of residuals which shows an approximately-normal distribution (Figure 6).

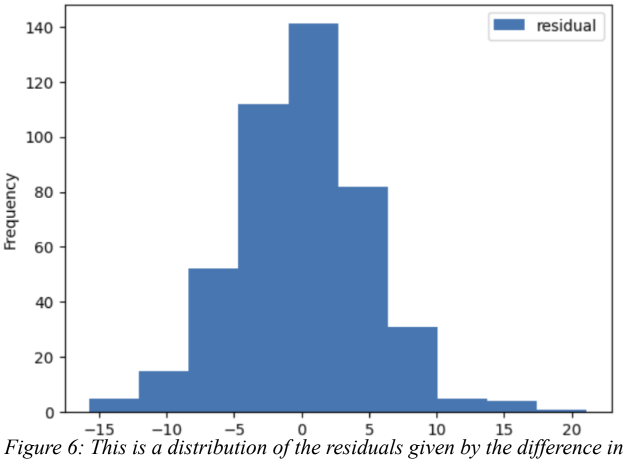
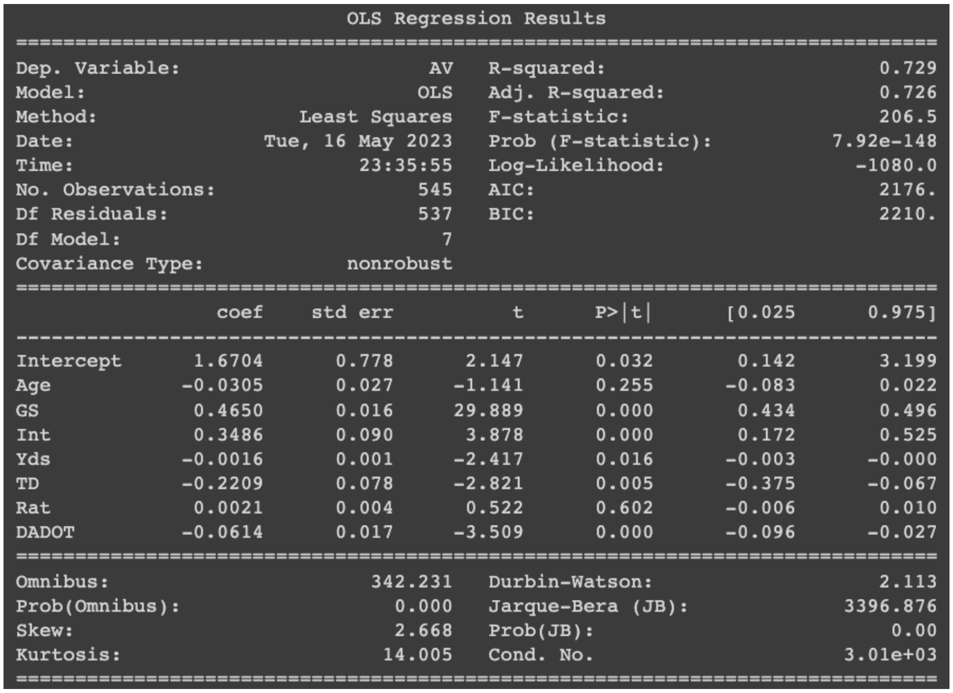


Figure 6: This is a distribution of the residuals given by the difference in each player's observed and predicted PFF grade.

We then run the same regression process on the bigger dataframe to predict Pro Football Reference’s Approximate Value parameter given a player’s attributes. Using the process described earlier regarding eliminating and keeping certain parameters, we aim to make a regression that predicts Approximate Value based on the player’s Age, Games Started, Interceptions, Yards, Touchdowns, Passer Rating (allowed), and DADOT (Table 2).



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     AV   R-squared:                       0.729
Model:                            OLS   Adj. R-squared:                  0.726
Method:                 Least Squares   F-statistic:                     206.5
Date:                Tue, 16 May 2023   Prob (F-statistic):          7.92e-148
Time:                        23:35:55   Log-Likelihood:                -1080.0
No. Observations:                 545   AIC:                             2176.
Df Residuals:                     537   BIC:                             2210.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.6704      0.778      2.147      0.032       0.142       3.199
Age           -0.0305      0.027     -1.141      0.255      -0.083       0.022
GS             0.4650      0.016     29.889      0.000       0.434       0.496
Int            0.3486      0.090      3.878      0.000       0.172       0.525
Yds           -0.0016      0.001     -2.417      0.016      -0.003      -0.000
TD            -0.2209      0.078     -2.821      0.005      -0.375      -0.067
Rat            0.0021      0.004      0.522      0.602      -0.006       0.010
DADOT         -0.0614      0.017     -3.509      0.000      -0.096      -0.027
==============================================================================
Omnibus:                      342.231   Durbin-Watson:                   2.113
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3396.876
Skew:                           2.668   Prob(JB):                         0.00
Kurtosis:                      14.005   Cond. No.                     3.01e+03
==============================================================================
```

Table 2: This is the output we get from running a regression to predict a player’s AV from the components Age, Games Started, Interceptions, Yards, Touchdowns, Passer Rating (Allowed), and DADOT.

We generate a regression that has the form:

```
'predicted' = [Age]*-0.0305 +[Games Started]*0.4650 +
[Interceptions]*0.3486 + [Yards]*-0.0016 + [Touchdowns]*-0.2209
+ [Passer Rating]*0.0021 + [DADOT]*-0.0614 + 1.6704
```

We then put our new columns into the dataframe and carry these results on over to the analysis section.

To finish off the methods we are describing, we run the same principal component analysis we ran in part one of the project and then analyze the variability given by the first three principal components.

ANALYSIS and RESULTS

Part 1:

The first part of our analysis includes a look at how each principal component weighs in investigating the variability found between observations, with an initial graph shown to the right (each bar represents a single principal component). There are discrepancies among researchers on what level of captured variability should be the benchmark for analyses, but in our case, let's look at components that give us over 80% of our variability. Taking the first component gives us 53.28% of our variability, the first two components give us 72.65% of our variability, and the first three components give us 82.89% of our variability, which is over our self-set threshold of 80% (Figure 7).

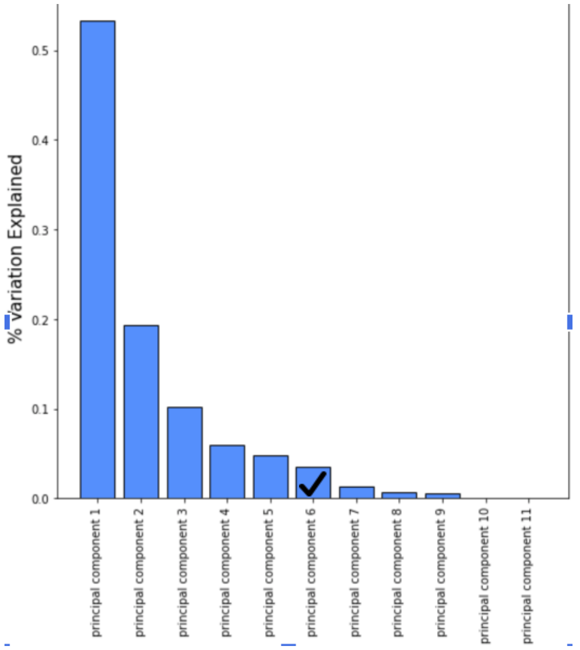


Figure 7: This bar graph shows the amount of variability captured by each principal component in our benchmark dataset.

Now we look at what goes into the vector given by the first principal component. We see that each of the vector's components contribute positively to the vector's direction and magnitude, although some more than others. A few of the heavier-weighted attributes include the calculated values for targets, completions allowed, yards allowed, air yards allowed, and yards after catch allowed. Recall that this vector gives us over fifty percent of our explained variance (Figure 8). Briefly, let's contrast this with the same mechanics shown by the second eigenvector, or the vector given by the second and third principal components.

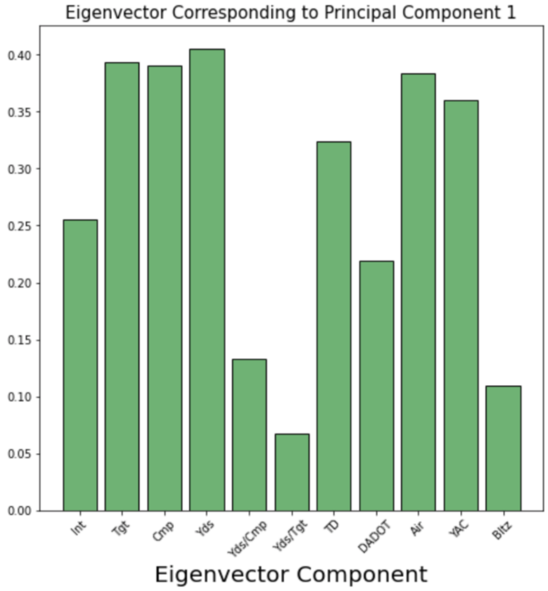


Figure 8: This bar graph shows the components of the dataset on the x-axis and how much they contribute to the overall eigenvector's direction and magnitude.

Upon examining principal component 2 and principal component 3, we see that there are several components that contribute to the direction and magnitude of each component (Figure 9). In principal component 2, we see that there are several

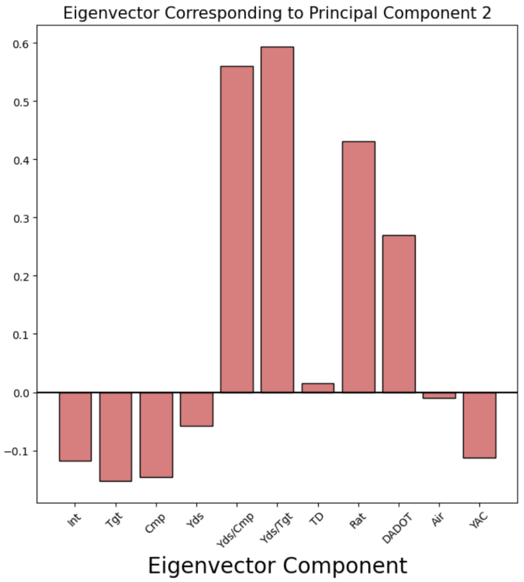


Figure 9: We have a bar graph similar to that of Figure 8, but this time we are examining the second principal component. Note that in contrast to the first, we have some attributes that contribute negatively towards the eigenvector's direction and magnitude.

components that contribute negatively, meaning that as these components increase, the score along principal component 2 decreases. Conversely, there are three components that contribute positively to principal component 2, indicating that as these factors increase, so does the score along principal component 2.

The three most-positively contributing factors in principal component 2 are yards per completion, yards per target, and DADOT (defensive average depth of target). These factors suggest that a team or player who performs well in these areas is likely to have a higher score along principal component 2. On the other hand, blitz attempts appear to have a negative impact on principal component 2.

Moving on to principal component 3, we see that the components have more uniform magnitudes, with the exception of blitz attempts, which contributes the most towards the direction and magnitude of principal component 3 (Figure 10). The other components in principal component 3 contribute in similar magnitudes, both positively and negatively. Notably, the components targets, completions, and yards seem to have little to no effect on the attributes of the eigenvector.

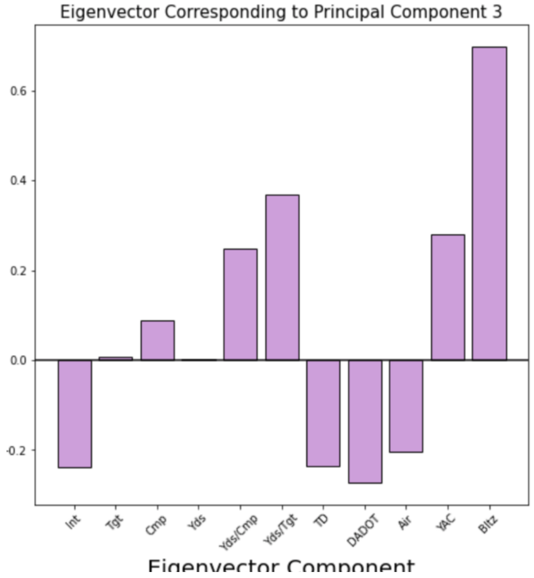
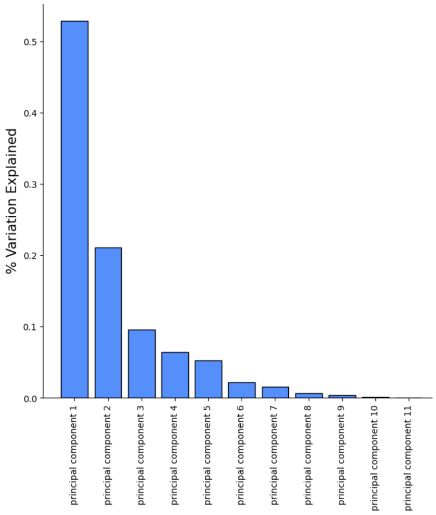


Figure 10: Same nature as Figures 8 and 9, and it seems that for the third principal component, the attributes have similar mazgnitudes to one another.

Part 2:

The second part of our analysis includes several steps with the goal of creating a ranking among our parameters of interest and creating a regression/fit-line. To arrive at this goal, we first want to analyze the eigenvectors of the principal component analysis of different dataframes and then create an informal ranking among the points of interest.

Figure 11: This bar graph shows the amount of variability captured by each principal component in our benchmark dataset.

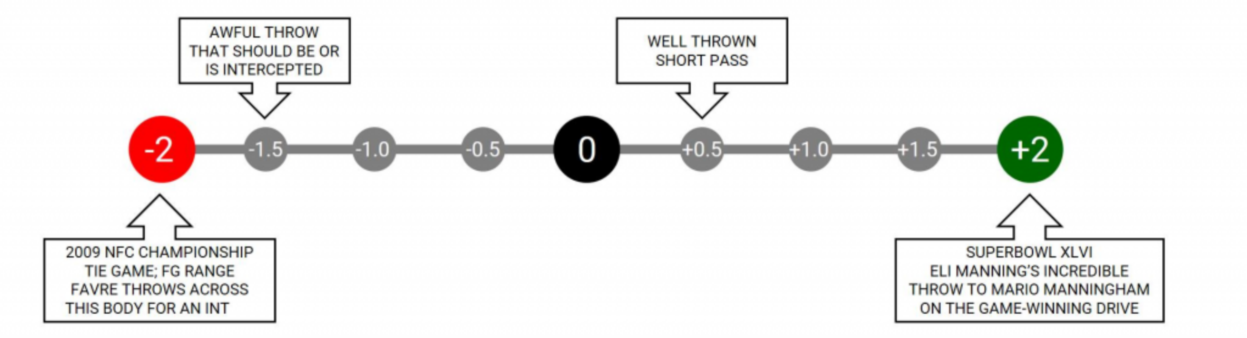


Starting with our baseline dataframe refined to contain defensive measurements such as interceptions, targets, yards, and others of the like. This dataframe is directly from the source Pro Football Reference, except for some attributes deemed useless (in terms of contributing to quantitative variation between units), such as the player's team and position. Running principal component analysis on this basic dataset gives what we have in part 1 (Figure 11).

The first three components return around 83.465% of our total variability captured. Here is where we want to capture more variability with the same number of components or less. So next, we investigate variability captured by a dataframe containing *graded* performance measurements via ProFootballFocus.

This second frame we are working with contains rows where each row represents a player, and the row's components consist of their defensive grade, defensive coverage grade, defensive pass rush grade, defensive run grade, and defensive tackling grade.

To explain this grading system that was referred to in the introduction, on every play, ProFootballFocus' (PFF) analysis team watches the play and assigns each player a number between -2 and 2 in increments of 0.5 based on their performance and impact on the play. A visual from PFF's official website:

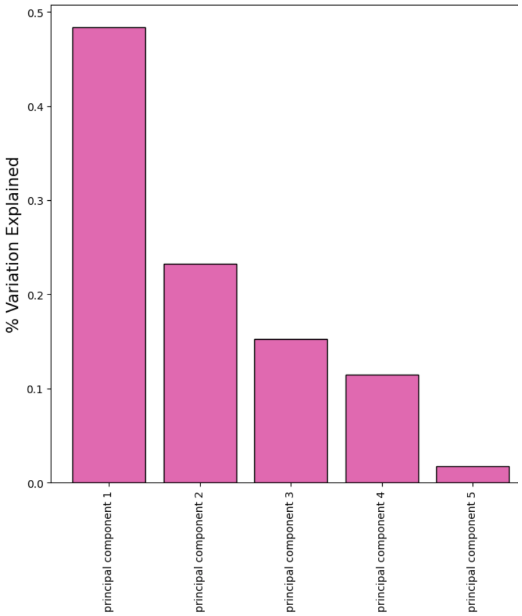

AWFUL THROW THAT SHOULD BE OR IS INTERCEPTED
WELL THROWN SHORT PASS
-2
-1.5
-1.0
-0.5
0
+0.5
+1.0
+1.5
+2
2009 NFC CHAMPIONSHIP TIE GAME; FG RANGE FAVRE THROWS ACROSS THIS BODY FOR AN INT
SUPERBOWL XLVI ELI MANNING'S INCREDIBLE THROW TO MARIO MANNINGHAM ON THE GAME-WINNING DRIVE

Moving forward with the construction of our second dataframe, we export a table from *ProFootballFocus'* official website. We obtain a relatively small dataframe where the components are player names, the team the play for, the position they play, their overall defense grade and four specific grades that go into that overall:

- Pass rush
- Tackling
- Coverage
- Run defense

Performing principal component analysis on this dataset, we capture about 86.854% variability in the first three components, which is a step up from the 83.465% captured in the baseline dataset (Figure 12).

Figure 12: This figure shows the five-component bar graph result for the PFF dataset, where each bar represents a component that contributes to the overall eigenvector.



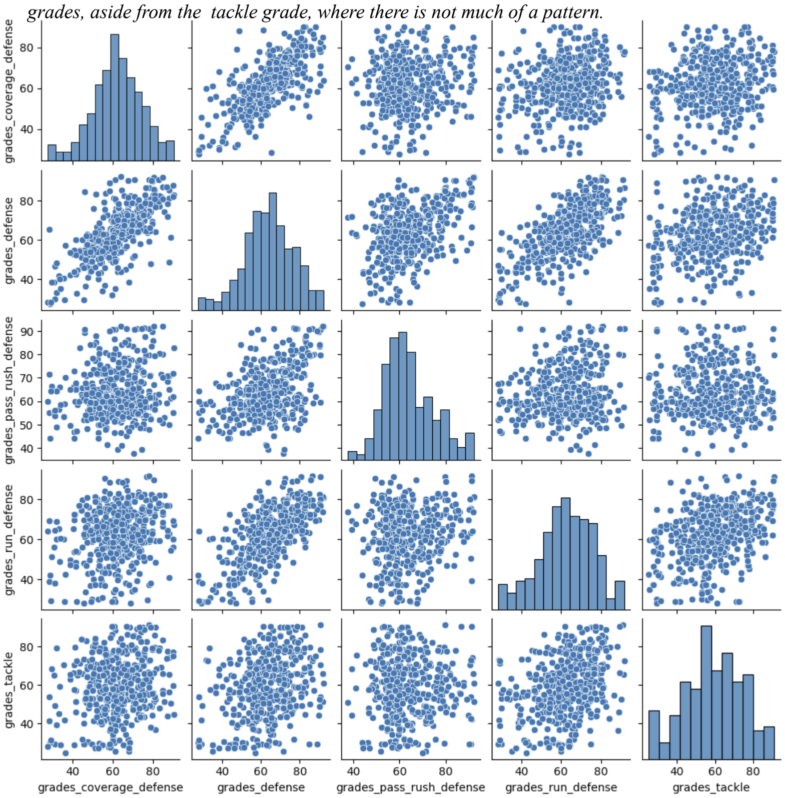
In principal component analysis, the amount of variability captured by each principal component is represented by a corresponding eigenvalue. The larger the eigenvalue, the more variability captured by the principal component. When the dataset has fewer components overall – such as this situation where we are using PCA on a data table with 5 components versus the previous table with 11 components – it is likely that each component captures a larger proportion of the total variability in the data. This means that the first few principal components may capture a significant amount of the overall variability, and the remaining components may capture relatively less variability.

This alone could be the cause of us capturing more variability with the first 3 components than the baseline dataset. However, this doesn't mean that this set of data can't capture more variability than the baseline set.

Now that we've created a skeleton for how to rank the explanatory variables, we delve into correlations and creating a new parameter via regression.

One of the first steps we are doing in this part of the process towards regression is analyzing the correlation matrices and heatmaps. First, we take a look at the pair plots we generated, which show an x-versus-y shape plot for each pair of components to get some insight into how each pair of components modify one another. In the smaller example given by our PFF dataframe, we see that in the plots with the overall grades, we have approximately linear relationships for each

Figure 13: This figure shows a 5x5 array of the behavior of pairs of components in the PFF dataframe. We see slight and strong linear correlations between pairs of attributes for the overall grades, aside from the tackle grade, where there is not much of a pattern.

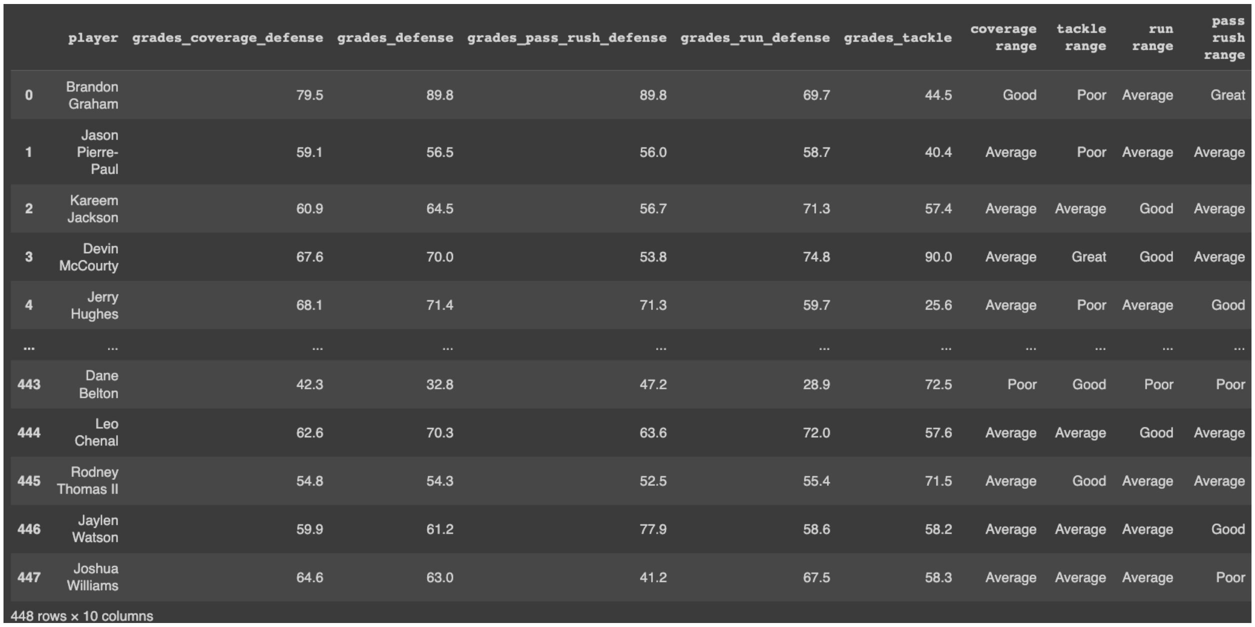


of the components versus overall grade, save for the tackle grade (Figure 13). It doesn't seem that the tackling grades for each player have a great bearing on the overall grade of the player at hand. Due to this, we can expect the coefficient for the tackling grade to be minimal paired with an insignificant p-value in our regression equation/output. This prediction boils down to significance, and a lack of significance from the tackling grade component would lead us to expect little impact or contribution in the regression model.

Since our dataframe for the concrete values including Approximate Value, Age, etc. is larger in shape, we're going to refrain from trying to analyze pair plots here as it would force a runtime error.

Carrying a similar approach over to our larger dataset, there are high correlation coefficients for some component pairs, as discussed in the *Methods* section. A lot of these highly correlated components are due to how some events on the football field record these components concurrently; i.e., a defender being responsible for allowing a 30-yard catch will be tacked with Yards Allowed, Air Yards Allowed, Yards After Catch Allowed, etc.

The regression process came with a handful of obstacles in manipulating the data in order to allow it to pass through a regression call. A regression model is then built to predict the PFF grade (overall defensive grade) using the component grades as predictors. One of the first steps done in an attempt to create a predictive function was to create bins, where grades were broken up into ranges of 0-50, 50-70, 70-80, 80-90, 90-100, and fitted to categorical bins "Poor," "Average," "Good," "Great," and "Outstanding," respectively (Table 3). Though it didn't prove to be fruitful in the regression aspect, it produced some insight into how we can break down the components going forward.



	player	grades_coverage_defense	grades_defense	grades_pass_rush_defense	grades_run_defense	grades_tackle	coverage range	tackle range	run range	pass rush range
0	Brandon Graham	79.5	89.8	89.8	69.7	44.5	Good	Poor	Average	Great
1	Jason Pierre-Paul	59.1	56.5	56.0	58.7	40.4	Average	Poor	Average	Average
2	Kareem Jackson	60.9	64.5	56.7	71.3	57.4	Average	Average	Good	Average
3	Devin McCourty	67.6	70.0	53.8	74.8	90.0	Average	Great	Good	Average
4	Jerry Hughes	68.1	71.4	71.3	59.7	25.6	Average	Poor	Average	Good
...	...	...	...	...	...	...	...	...	...	...
443	Dane Belton	42.3	32.8	47.2	28.9	72.5	Poor	Good	Poor	Poor
444	Leo Chenal	62.6	70.3	63.6	72.0	57.6	Average	Average	Good	Average
445	Rodney Thomas II	54.8	54.3	52.5	55.4	71.5	Average	Good	Average	Average
446	Jaylen Watson	59.9	61.2	77.9	58.6	58.2	Average	Average	Average	Good
447	Joshua Williams	64.6	63.0	41.2	67.5	58.3	Average	Average	Average	Poor

448 rows × 10 columns

Table 3: This table shows our efforts towards creating a regression where for each explanatory variable, each player is has a categorical variable rather than a quantitative variable.

As previously mentioned, we obtain the following predictive model:

```
predicted = [coverage defense]*0.5818 + [run defense]*0.4514
+[tackling]*0.007 + [pass rush]*0.3283 - 21.9203
```

The tackling grade component did not contribute much towards the regression, as was predicted due to the lack of modification of overall grade due to tackling grade and vice versa (Figure 5). After putting the residual value as a new column in the dataset, we analyze the "best" players, given by how much above expected they are.

We apply these same steps to the larger dataset to get a new insight on the players and how they compare using different parameters. We use our dataset from Pro Football Reference to predict AV using various advanced measurables and construct the following regression:

```
'predicted' = [Age]*-0.0305 + [Games Started]*0.4650 +
[Interceptions]*0.3486 + [Yards]*-0.0016 + [Touchdowns]*-0.2209
+ [Passer Rating]*0.0021 + [DADOT]*-0.0614 + 1.6704
```

Games started, interceptions, yards allowed, touchdowns allowed and defensive average depth of target were the main contributors towards predicting AV for the players in the dataset. Though some components in the regression aren't statistically significant, we can leave them since

a. it gives us a tiny bit more variability, and
b. the coefficients given by insignificant components are negligible.

Taking a look at the ten greatest residuals given by players' predicted PFF grades versus their actual PFF grades, we see players (in no order)

- Aaron Donald – 3x NFL Defensive Player of the Year, 7x First-Team All-Pro, 9x Pro-Bowler, [4]
- Za'Darius Smith – 3x Pro-Bowler, 1x Second-Team All-Pro, [21]
- Haason Reddick – 1x Pro-Bowler, 1x Second-Team All-Pro, [10]
- Bradley Chubb – 2x Pro-Bowler, [5]

and other up-and-coming players like Christian Barmore, Odafe Oweh, and Jaelan Phillips – all of whom were elected to PFWA's All-Rookie Team in 2021, and who are expected to rack up accolades throughout their careers.[7] [11] [17]

For our dataset from Pro Football Reference: taking a look at the ten greatest residuals given by players' predicted AV versus their actual AV, we see players (in no order)

- Micah Parsons – 2x First-Team All-Pro, 2x Pro-Bowler,[14]
- Nick Bosa – 1x NFL Defensive Player of the Year, 1x First-Team All-Pro, 3x Pro-Bowler, [15]
- Fred Warner – 2x First-Team All-Pro, 2x Pro-Bowler, [9]

along with players Roquan Smith, Demario Davis, Patrick Surtain II, Matt Milano, Marcus Jones, Chris Jones, and Quinnen Williams – all of whom have been nominated to All-Pro First-Teams *and* Pro Bowls (except for Marcus Jones, who hasn't been selected to a Pro Bowl in his young career where he was a rookie in this past season).[6] [8] [13] [14] [18] [19] [20]

We can use these results as somewhat self-confirming information in the sense that it gives us a grain of confidence that our regression and analyses are worthwhile. To wrap up this section, we quickly touch on the variability and principal component analysis of these frames.

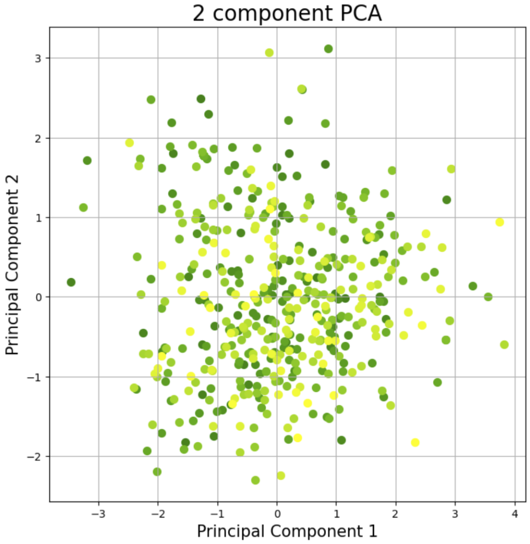


Figure 14: The nature of the first two dimensions of the principal component output have a birdshot-like spread, where we get almost 85% of the total variability in the first three principal components.

For the dataframe constructed from the PFF grades and their components, our first two principal components give us a birdshot spread and our first three principal components capture 84.94% of our total variability (Figure 14).

For the dataframe constructed from the Pro Football Reference statistics and their components, our first two principal components give us a similar spread as we got for our baseline data (Figure 15); our first three principal components capture only 74.16% of our total variability, so we need to add a fourth principal component to our sum to reach our desired threshold of 80%. Four components gives us 80.85% of our total variability.

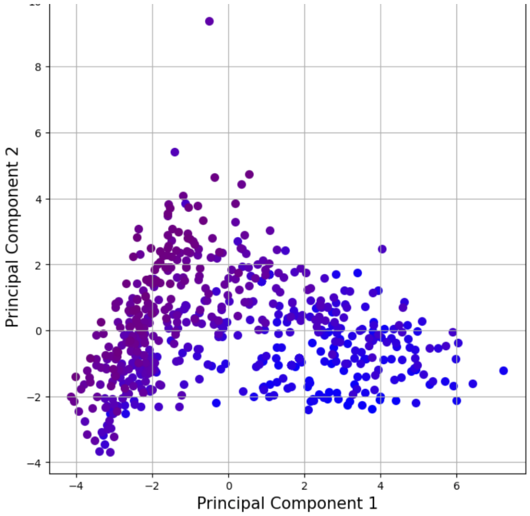


Figure 15: This graph for our final principal component analysis has a similar nature to that of our first principal component analysis, where we see the points going in two directions according to the first two eigenvectors.

CONCLUSION

As mentioned in the introduction, this project aimed to build a new way to analyze defensive matchups in a more subjective way. We try to take all of these media-driven analyses with a grain of salt and put our own effort in to investigate the validity of those analyses, and even compare our findings to see if there is any consistency among popular opinions, numbers, and our results. Again, while we're in a field obsessed with advanced analytics and having the right opinion where concrete details base our analyses, we analyze human beings and their on-field performance, so introducing a subjective lens felt necessary. The media gets caught up in that they forget that these football players are regular people and have tendencies, emotions, and lives outside of football. This sentiment is then absorbed by many fans without giving much thought to it. Our goal was to build a system in which we can fairly apply weight to situations based on players' efficiencies in different aspects of the game. Our goal to develop a system that accurately and fairly applies weight to different situations based on a player's efficiency in various aspects of the game was done through regression, principal component analysis, and data manipulation.

In conclusion, our journey towards our goal has been accompanied by challenges and hurdles that we had to get through. Throughout the process, we came across obstacles related to data manipulation, encountered dead-end processes that we had to abandon, and experienced passes of difficulty in generating motivation and creativity. Despite these setbacks, we drove forward and adapted, finding alternative solutions and embracing new approaches. The lessons learned from navigating these challenges have not only strengthened skills but also deepened understanding of the importance of resilience and flexibility in achieving success. As we reflect on our path up to this point, we can say that while it was not easy, the goal was certainly worth it. By confronting and conquering these obstacles, we have emerged stronger, more knowledgeable, and better equipped to face future endeavors with determination and confidence.

REFERENCES

1. Doss, Andrew. “The NFL Combine in Two Dimensions.” Medium, 23 Aug. 2021, innerjoin.bit.io/the-nfl-combine-in-two-dimensions-95ba79e3470d
2. Palazzolo, Steve. “PFF Player Grades.” PFF, www.pff.com/grades
3. “Premium Stats DataTable.” Premium Stats, premium.pff.com/nfl/positions/2022/REGPO/defense, Accessed 8 June 2023.
4. Wikipedia contributors. "Aaron Donald." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 30 Apr 2023. Web. 9 June 2023.
5. Wikipedia contributors. "Bradley Chubb." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 3 Feb 2023. Web. 9 June 2023.
6. Wikipedia contributors. "Chris Jones (defensive tackle, born 1994)." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 7 June 2023. Web. 9 June 2023.
7. Wikipedia contributors. "Christian Barmore." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 28 May 2023. Web. 9 June 2023.
8. Wikipedia contributors. "Demario Davis." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 14 May 2023. Web. 9 June 2023.
9. Wikipedia contributors. "Fred Warner (American football)." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 17 May. 2023. Web. 9 June 2023.
10. Wikipedia contributors. "Haason Reddick." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 17 May 2023. Web. 9 June 2023.
11. Wikipedia contributors. "Jaelan Phillips." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 30 May 2023. Web. 9 June 2023.
12. Wikipedia contributors. "Josh Uche." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 28 May 2023. Web. 9 June 2023.
13. Wikipedia contributors. "Marcus Jones (cornerback)." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 8 June 2023. Web. 9 June 2023.
14. Wikipedia contributors. "Matt Milano." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 17 May 2023. Web. 9 June 2023.
15. Wikipedia contributors. "Micah Parsons." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 4 June 2023. Web. 9 June 2023.

16. Wikipedia contributors. "Nick Bosa." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 7 June 2023. Web. 9 June 2023.
17. Wikipedia contributors. "Odafe Oweh." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 28 May 2023. Web. 9 June 2023.
18. Wikipedia contributors. "Patrick Surtain II." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 8 June 2023. Web. 9 June 2023.
19. Wikipedia contributors. "Quinnen Williams." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 5 June 2023. Web. 9 June 2023.
20. Wikipedia contributors. "Roquan Smith." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 25 May 2023. Web. 9 June 2023.
21. Wikipedia contributors. "Za'Darius Smith." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 7 June 2023. Web. 9 June 2023.
22. “2022 NFL Advanced Defense.” Pro Football Reference, www.pro-football-reference.com/years/2022/defense_advanced.html, Accessed 8 June 2023.