Finding Fantastic Feats of Football in the NFL

Using Python and statistical libraries to mine unique relationships from NFL play-by-play data

Alex Plazas
Computer Science
University of Colorado-Boulder
Boulder, CO, USA
chpl8693@colorado.edu

Benjamin Price
Computer Science
University of Colorado-Boulder
Boulder, CO, USA
benjamin.price@colorado.edu

Raul Ramos
Computer Science
University of Colorado-Boulder
Boulder, CO, USA
Raul.Ramos@colorado.com

ABSTRACT

This project provides an in-depth look into mining play-by-play data from NFL games from 2009 - 2018 in order to gain knowledge by answering significant questions related to prediction and analysis. While focused on allowing the dataset to lead exploration, questions arose such as "Which aspects of a play have the greatest impact on play outcome?", "Can players be classified into levels of achieving tiers based on their performance in their plays?", and "Which players perform the best when faced with high-pressure situations, such as on 4th down or in the red zone?" Through the framework of feature selection, multiple linear regression, and clustering analysis, we found that more offensive behavior is related to better performance in a play. Clustering analysis through k-means discovered specific groups of players based on achievement levels which can highlight outliers for both good and bad performance. Lastly, based on clustering, it was found that groupings of high performance players in high-intensity situations can be defined. The following paper describes the inspiration for the project as well as related work, followed by an in-depth explanation of the data mining processes applied.

INTRODUCTION

Sports serve society by providing vivid examples of excellence. From the basketball court to center ice, athletes and teams strive for success in their respective disciplines. While there are many definitions of success, statistics generated from competition tend to come first when measuring performance.

In recent years, innovations in computing and information management have ushered in a new era of statistics in sports. Sports leagues and the teams therein commit vast sums of money to collect and process data that could give their side an edge in competition. Insights gained from larger bodies of statistical information from a sport enable teams to learn from the past, apply to the present, and potentially predict the future. And the sea-change in sentiment toward statistics is clear in the National Football League (NFL).

The NFL, including American football in a broader sense, has embraced statistics for the better of the sport. Each of the 32 teams of the league, including the league itself, has invested millions in dedicated offices for gathering and mining statistical data. And for good reason. At stake are billions of dollars invested by ownership, sponsors, and fans, with all of the above wanting their money's worth. Thus, each office works tirelessly to find the next big prospect or craft a guaranteed-winning strategy backed by data-driven methodologies that go beyond “gut instinct”.

Beyond the monetary incentives for collecting and processing vast sums of data, statistics offices for NFL teams are typically concerned with one problem: finding outliers. More specifically, positive outliers depend on the statistical category or measurement implied. Granular play-by-play data enables team front offices to detect outliers and assess performance in deeper detail. Teams emphasize finding players or situations that increase their probability of winning, and the players and situations involved are typically out-of-the-ordinary. Certain environments might also lend themselves to outlier performance such as specific training routines or game-time decisions, both of which can be modeled and tested using game statistics.

Our stated goal is to find such fantastic feats and study their impacts. We wish to understand the confluence of statistical factors that lead to outlier performances. We want to develop methods for grouping players based on their performance and identify outliers that might make the hall of fame. We also wish to build a model that incorporates data from a wide range of seasons to predict the outcomes of plays and games given the presence of outliers. In the end, the team wishes to emulate the work of an NFL front office in finding fantastic feats of football.

RELATED WORK

Many researchers have attempted to tackle the NFL outcome prediction problem by using a mix of historical and play-by-play data, with unique constraints and intentions. Some studies focus on player-specific metrics at certain positions such as quarterback. In a paper titled “Measuring Productivity of NFL Players”, Berri et al. sought to explore factors associated with the quarterback position and how quarterback productivity correlated with a team's offensive ability[3]. The study outlined several statistical attributes including the derived attribute of Quarterback Rating (QBR), and standard statistical categories such as yards, plays, and interceptions. Using regression analysis, the findings of the study indicated that there are measurable positive increases in the predicted point differential for each yard thrown for, and a negative impact on point differential for every play attempted. It was also shown that from 2000 to 2010, Peyton Manning had the highest QB rating and nearly a quarter of the top 40 seasons during the time period[3]. At the conclusion of the paper, Berri and Burke suggested shortcomings and alternatives to their regression model, such as calculating expected point values and “success rate.”

Other studies take a broader approach by analyzing team performance over time given certain conditions and historical outcomes. In “A Hybrid Prediction System for American NFL Results”, Uzoma et al. propose a method for predicting NFL games using particular models and features[5]. A hybrid linear regression and k-Nearest Neighbors model were designed to increase the prediction accuracy of already weighted features. Several data attributes of interest were also identified in the study including points scored by both teams, the number of turnovers, and offensive/defensive rating[5]. Using the hybrid model, the authors were able to predict the outcomes of games with nearly 80% accuracy during the 2013 regular season[5].

In a similar study titled "Predicting Margin of Victory in NFL Games...", Warner proposes using a machine learning model to accurately predict games better than Las Vegas bookmakers. Aside from using regular counting stats and attributes, the study also included novel features such as the location of the game, stadium conditions, and even climate/weather data for the particular game[4]. To generate predictions, a feature set was defined which minimized cross-validation error, and features were passed to a Gaussian process predictive model to yield a final prediction. Results of the study indicated that the model did not out-perform Las Vegas prediction models at a 95% confidence interval, despite the model including novel attributes[4].

On the topic of important attributes, a study published in 2010 titled "What Makes a Winner?" sought to identify attributes that strongly correlated to the expected outcome of a game. Gifford et al. defined a model which employed a decision tree coupled with binary logistic regression to find key attributes in NFL play-by-play data[6]. Several parameters were tested including passing yards, rushing yards, and turnovers to find which impacted the expected outcome the most. The results of the study indicated that offensive turnovers were the most important team statistic in determining the winner of an NFL game, having a strongly negative effect on the outcome for the offensive team, with the inverse being true for the defensive team[6].

DATASET

The data set used in this exploration is from the Kaggle website from user Max Horowitz[2]:

https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016?select=NFL+Play+by+Play+2009-2018+%28v5%29.csv

Detailed play-by-play statistics per game have been collected in this data set for National Football League games from 2009 through 2018. There are 253 attributes in the data set, with 449,371 data objects each identified by a combined key made up of both a Game Id and a Play Id.

Most data types are represented within this data set given that there are so many attributes provided. Several nominal attributes are included, such as Play Type, Home Team, Away Team, and Pass Location. Binary attributes are numerous within the data to detail the inclusion of certain types of plays, such as if there was a kickoff attempt or a punt attempt, and ordinal data is provided such as the quarter of the game in the Qtr attribute. There are plenty of numeric attributes as well, such as the ratio-scaled data of Yards Gained and Game Seconds Remaining.

The sheer voluminous nature of this data allows versatility in the knowledge discovery process and provides multiple windows of perspective through which to explore patterns in recent NFL seasons.

MAIN TECHNIQUES APPLIED

CLEANING TECHNIQUES

The project began with the found data as described in the previous section. The data was abundant and needed to be cleaned and aggregated in some meaningful ways to facilitate the types of exploration this project sought to perform without losing any information. Many data mining efforts struggle with how to deal with null or missing data. While our strategy could vary depending on the data type of a given attribute, many of our missing values were related to binary attributes and were safely set to 0. Similarly, null values in nominal attributes, such as Passer Player Name were included as “NA” because the field is an implied binary; “NA” means there was no passer.

Other forms of cleaning also needed to be performed to reduce the number of attributes to sift through in the data. Some of the attributes given for each play-by-play instance were similar in nature (i.e. forced fumbles, fumble forced, fumble not forced), and were combined to simplify the analysis process. Similarly, some attributes were implied from others and were safely removed, such as Game Half being implied from Qtr. Finally, several categorical attributes such as Play Type remained in the dataset with string values that were enumerated. By carefully reducing the number of attributes as well as enumerating categorical attributes in these ways, the data mining process was greatly sped up without the loss of information gain.

Several preprocessing steps also had to be performed once the data was cleaned. Our data after cleaning was broken up into many dimensions, with each object representing an individual play. While some exploration would look at data per play, full game data would also be required. A roll-up of the data per game was implemented which took place by totaling up binary data as well as the other numeric data and removing any attributes that would not apply at that level, such as seconds remaining or the Qtr attribute. With this higher-level view, comparisons could be made at a per-game level as well as the per play level provided by the original data.

REGRESSION TECHNIQUES

The most immediate obstacle before running a regression analysis was that the clean version of the dataset still occupied too much space and could not be easily ready by IDEs such as Visual Studio Code or JupyterHub. This was fixed by running a correlation matrix to find the top 10 attributes closer to EPA (expected points added) by sorting the list. Next, as a preprocessing step, Python’s NumPy and Pandas libraries were used to read the CSV file clean dataset, list its attributes, and filter the 10 attributes that correlated to EPA. This allowed for a CSV file that could be read fast.

For the actual regression analysis, a Jupyter notebook was used and Python libraries NumPy, Pandas, Matplotlib, Seaborn, and Sklearn were imported.

First, Python code was written to get a feeling for the dataset at hand. For example, reading the file, and printing the shape, information, and description helped verify that the rows and columns were in place and complete.

Second, using the Matplotlib and Seaborn libraries allowed for plotting columns and seeing the correlation between EPA (independent variable) against other attributes (dependent). Moreover, a heat map was used to visualize such correlations in case the scatterplots were not as effective. Both provide the reader with clear representations of the relationships among attributes.

Third, to perform a linear regression, four steps were followed: from the dataset, assign x and y variables, train and test the set, train the model, and evaluate the model. For the train and test step, the years had to be split into 10 segments (from 2009 to 2018)—9 for training and 1 for testing. Additionally, the Statsmodel library was imported for building and training the model.

Fourth, a scatterplot was drawn from the train and test step, and a regression line was drawn.

Fifth, Python code for residual analysis was implemented to measure and graph any error terms during the creation of the linear regression. Sixth, Python code for forming predictions on the test data or evaluating the model was also employed, which included the r-square of the data to measure the proportion of the variance between the independent and dependent attributes examined. This was indeed used to graph the test portion of the dataset to compare against the train.

CLUSTERING/CLASSIFICATION TECHNIQUES

After performing a regression analysis, the group sought to identify clusters and derive outliers from cluster analysis on the dataset. By creating clusters and classifying them, discrete groups arise based on pre-selected players or play attributes. Clusters also provide an easy way to view the data and make outliers more apparent when visualized. In the context of our analysis, certain clusters would be the baselines for classifying plays or players to be fantastic, qualifying them for the Hall of Fame.

Before performing the cluster analysis, a series of questions were defined which guided attribute and model selection. In professional football, the relationship between a quarterback and receiver is crucial, often defining game outcomes. A prolific passer paired with a prolific receiver can wreak havoc on opposing teams en route to a Super Bowl appearance. Thus, our first question: what do quarterback/receiver duos average for yards per completion versus the duo's number of attempts? Our next question involved running backs and the concept of 'big plays.' Big plays are in this instance defined as runs of 25 yards or more, making scoring on the play or subsequent play more likely. More veteran running backs might have higher average yards, while one-hit-wonders might have lower average yards, but more potential for creating explosive plays. In other words: What running backs have completed 'big plays' (runs of 25 yards or more) per average yards a rush? Finally, we wanted to understand how and what players performed when it mattered most: during the 4th downs and in the 'red zone.' Since a team only has 4 downs to travel 10 yards and gain a first down, 4th down moments are critical to a driver's success. Once the team is within 10 yards of a touchdown (goal-to-go or 'red zone'), scoring is the culminating event of the drive. Quarterbacks who are high-achieving in both situations would separate themselves from the pack once a cluster is applied. Thus: How do quarterbacks perform in the key situations of 4th down passes and 'red zone' passes?

With a series of questions defined, several clustering techniques were evaluated to determine which would best fit our data. It was determined that k-means clustering would have the greatest benefit in highlighting groups, and from an ease-of-implementation perspective. K-means is a clustering technique where data groupings are classified based on their proximity to a centroid shared among other data points in a cluster. Shown below is the formula describing

$$E(m_1, \ldots, m_M) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(x_i \in C_k) |x_i - m_k|^2$$

the k-means algorithm:

Some advantages of k-means clustering are that the technique is easy to implement from a code standpoint and that the technique runs in linear time, with a time complexity in big O notation of O(n). On the other hand, drawbacks of k-means include the random initialization of centroids which causes discrepancies between each iteration and the general lack of consistency.

Despite the shortcomings, k-means was chosen as the main clustering technique.

With questions and a technique defined, we moved to perform data segmentation and implement the k-means algorithm for each case. Starting with the question of “what do quarterback/receiver duos average for yards per completion versus the duo's number of attempts?” additional data processing was necessary. The original dataset was filtered on attributes including passer/receiver name, yards gained, and play type to create a new data frame. After narrowing down the key attributes, the data frame was iterated through and an additional dictionary was defined where the keys were a tuple of the passer player name and receiver player name, and the values were a subframe with each play for that duo. For each duo found in the filtered data frame, that row entry was saved to the duo dictionary at the duo tuple key. Once every duo was identified with plays saved to the dictionary, statistics were calculated for each duo, namely average yards per play and total plays. The stats calculated were then saved to the dictionary at the duo tuple key, overriding the previous play values. In the end, the dictionary consisted of the duo tuples as keys, and a list of the average yards per play and total plays as dictionary values. The augmented dictionary was then saved to a data frame which would be operated on by the k-means algorithm.

The k-means algorithm was implemented by first defining 3 groups to cluster the data. Random centroids were generated and used for initializing the algorithm. For each data point, the distance from each centroid was calculated and classified depending on which centroid was the closest. Once all distances were calculated, the mean of the group distances was determined and new centroids were set for that iteration. The same process was repeated until the centroids did not change much from the previous iteration, indicating that the centroid accurately defined a group cluster. After the clusters were generated, the original data was plotted and colored based on the cluster group. Finally, centroids were overlaid on the newly plotted data for visual reference.

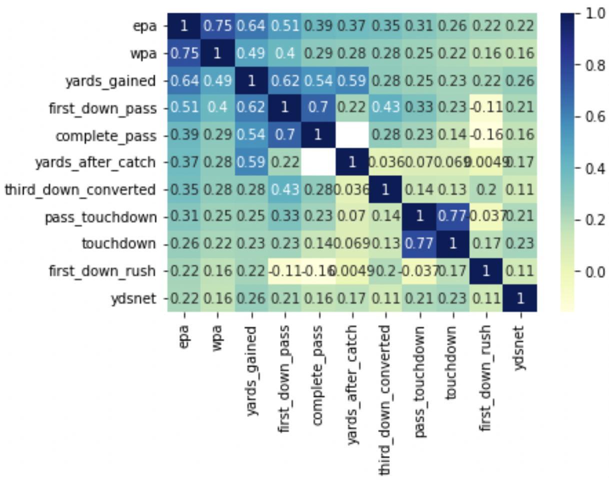
The next question, “what running backs have completed ‘big plays’ (runs of 25 yards or more) per average yards a rush?” followed a similar pattern. The original dataset was filtered along with attributes relevant to the question such as yards gained and running play type. The filtered data were grouped by running back specifically, resulting in a dictionary where keys were running back names, with values corresponding to each play that running back completed. The dictionary was converted to a data frame and used for the k-means clustering. The same k-means algorithm used to answer the first question was featured again, with few changes. 3 clusters of running backs were generated and plotted with centroids overlaid on the original data.

Finally, the question of “how do quarterbacks perform in the key situations of 4th down passes and ‘red zone’ passes?” was answered akin to the preceding questions. Data was filtered on attributes such as ‘goal to go’ and passing play type to create a new data frame. A dictionary was created with keys as quarterback name and values as each pass the player had thrown which met the filtering criteria. A success/fail ratio was computed for 4th down attempts and red zone attempts, overwriting the play values for the quarterback key in the dictionary. The dictionary was converted to a data frame and k-means were applied. The final clustering results were plotted with centroids presented for each cluster.

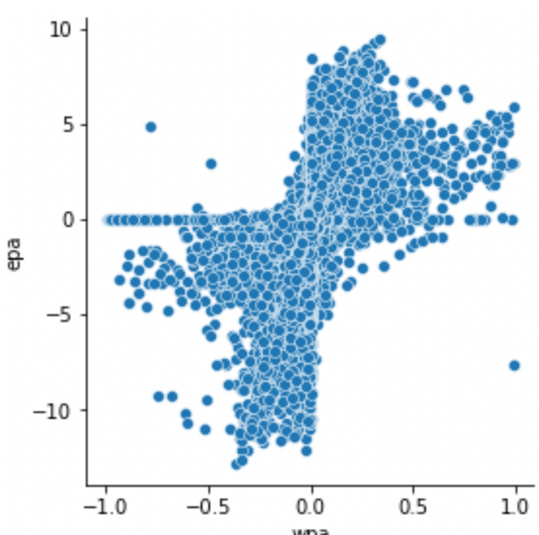
KEY RESULTS

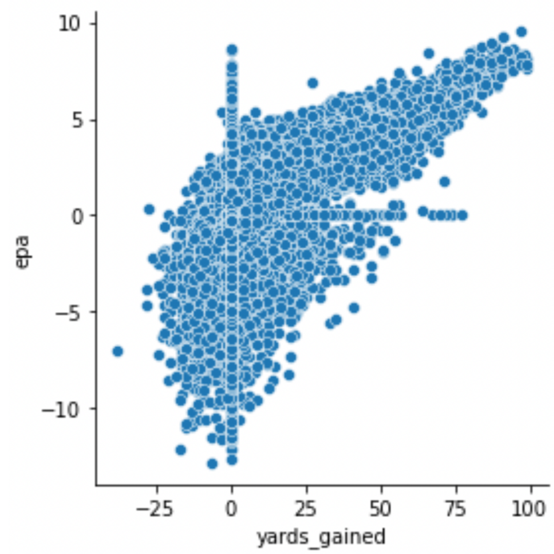
REGRESSION ANALYSIS

We have chosen to closely study the attribute EPA (expected points added), which is the difference between a team’s expected points at the beginning and the end of a play. The remaining attributes for comparison are shown on the visualizing a heatmap below (i.e., WPA, yards gained, first down pass).



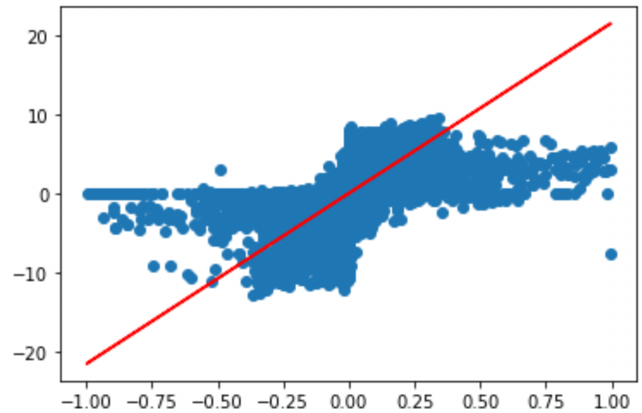
The second and third attributes above are win probability added (WPA), which is a measure of how a player’s contribution can change the outcome of a game and has a correlation to EPA of 0.75, and yards gained with a correlation of 0.64. We note from the correlation between WPA and EPA that the more talented the player, the better chances a team has at winning games, while the opposite is also true. Similarly, the more positive yards a team gains in any given play, the higher the probability of scoring, and vice versa.



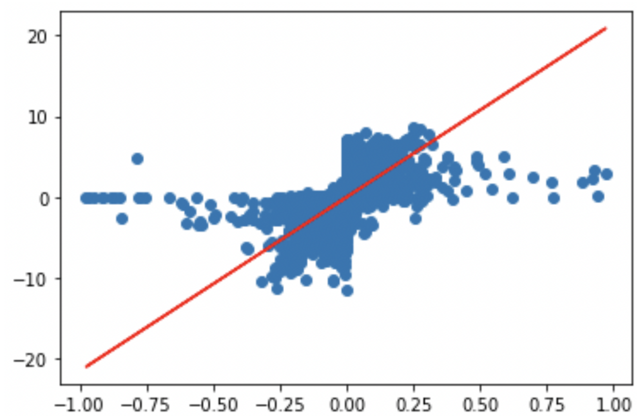


For our regression analysis, first, determine x and y. We must compare with EPA other highly correlated attributes, such as WPA and yards gained. Second, we train and test our sets by using a 9:1 ratio. Thus, we consider the following attributes in order: WPA, yards gained, first-down pass, complete pass, yards after catch, third-down converted, pass touchdown, touchdown, first down rush, yards net. Third, we build a linear regression model.

The linear regression below shows the relationship between EPA and the next 10 correlated attributes for the 9 out of 10 segments that were used to train the model:



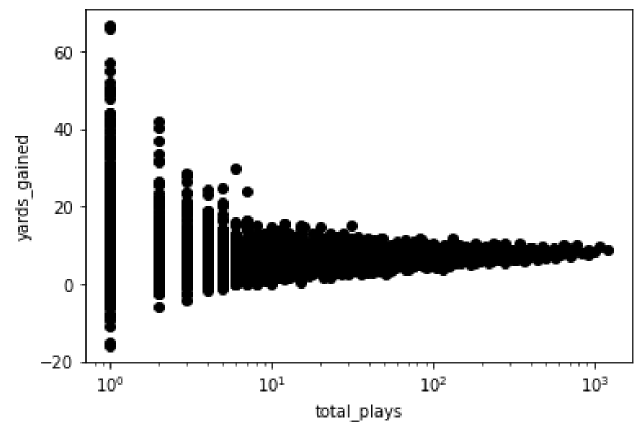
Similarly, the regression below shows the 1 out of 10 segments used to test the dataset:



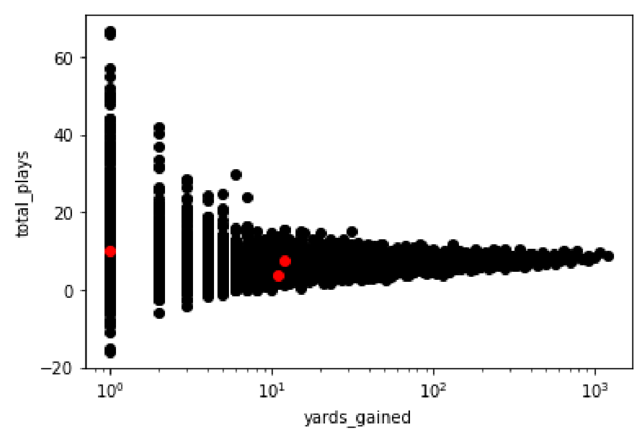
Both graphs demonstrate that talented players have a positive effect on their team’s outcome and that attacking plays are highly beneficial. In other words, attacking is the best form of defense. Knowing this can be beneficial, as teams can invest in highly talented players to win games. Nonetheless, the market could eventually level out if more teams are applying the same knowledge.

CLUSTERING/CLASSIFICATION

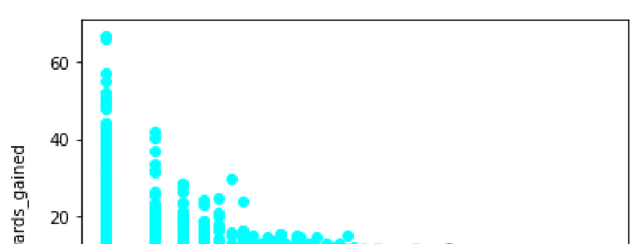
After performing a clustering analysis using k-means for the three questions described in the previous section, the results were dissected and contextualized to NFL player performance. For the first question, “what do quarterback/receiver duos average for yards per completion versus the duo's number of attempts?”, the results of the clustering indicated that there are discrete groups of quarterback and wide receiver duos. The unclustered data shows a clear distribution along the x-axis, with high variance along the y axis for x values closer to the origin, and lower variance as x increases. In this case, the x-axis corresponds to total plays for the duo, and the y-axis corresponds to average yards per play for the duo. Below is a graph of the pre-processed data, log-scaled:



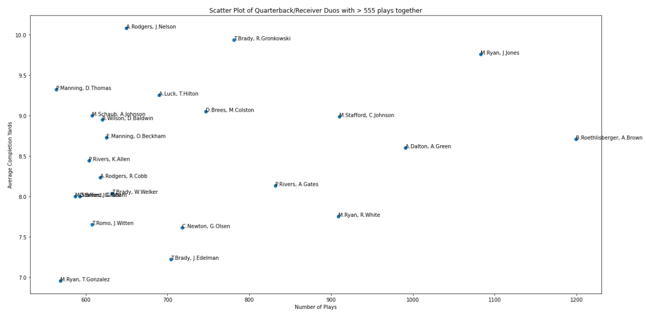
Because random centroids were required to initialize k-means properly, a new graph was generated which accurately displayed the randomized starting points:



After k-means was complete, clear clusters emerged and were colored accordingly:

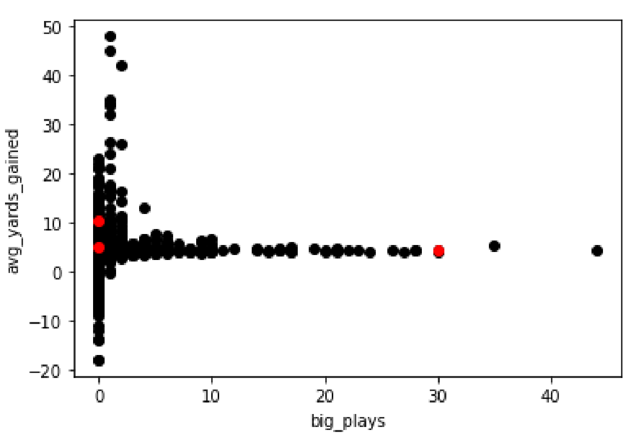


Focus was applied to the outermost cluster, which represented quarterback and wide receiver duos who were prolific over our time range.

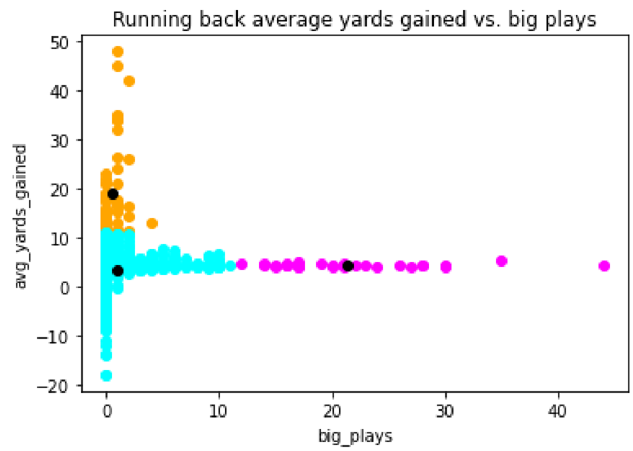


The results of the cluster analysis show that there is a clear grouping of all-star caliber talent for both quarterbacks and wide receivers. Duos like Ben Roethlisberger and Antonio Brown; Tom Brady and Rob Gronkowski; Matt Ryan and Julio Jones; each of these duos have one or more players who are almost guaranteed to go into the Hall-of-Fame. Another intriguing result from the cluster is the presence of players such as Andrew Luck, who was a prolific passer, but retired at a young age and thus handicapped his potential for the Hall-of-Fame. In essence, if you are present in this cluster, you are in a very esteemed company.

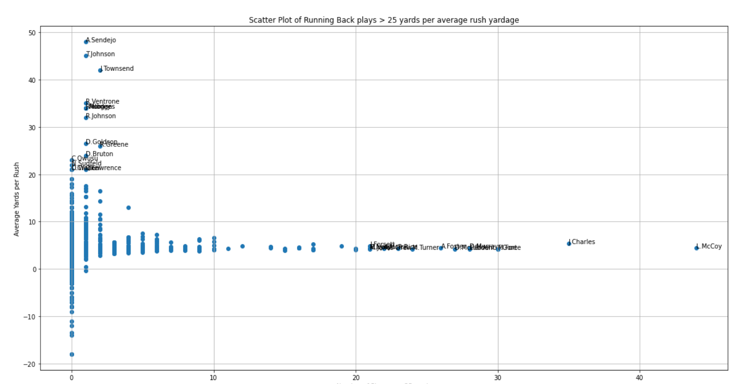
Our next question sought to explore “what running backs have completed ‘big plays’ (runs of 25 yards or more) per average yards a rush?” Each running back was plotted based on their average yards gained versus their total number of ‘big plays’, with initial random centroids overlaid as shown below:



After running the k-means algorithm using initially random centroids, clusters were achieved. The clusters were given unique colors as in the previous application, orange, cyan, and magenta. The resulting graph is as follows:



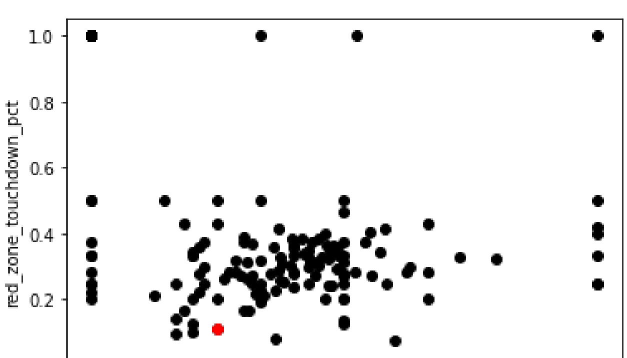
From the colored cluster plot, there exist 2 clear groupings around the origin and along the y-axis, with another grouping much further along the x-axis. In the context of our result, the cyan grouping represents players who were all-around players who were consistent in yards gained while having few big plays over their career. The orange grouping represents players who did not run often and who might’ve only had one running play over 25 yards, increasing their average yards gained substantially. The magenta grouping represents high-achieving running backs who were able to create explosive plays,



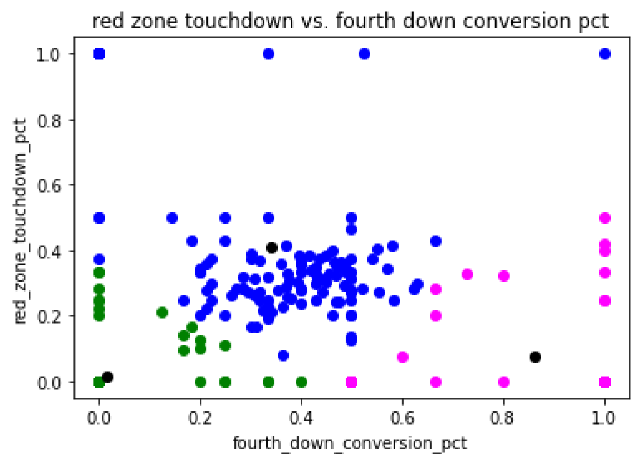
despite having lower average yards gained over the course of their career due to their volume of plays. Below is a graph with labels applied to the magenta and orange groups:

With the data labeled, it became clear that the player clusters were unique in their circumstances. Starting with the outermost cluster along the x-axis, Hall-of-Fame caliber running backs are present in the distribution. LeSean McCoy, Super Bowl champion and Pro Bowler, led all runners over this period with nearly 45 explosive plays. Jamaal Charles, while never winning a Super Bowl, was highly productive over our statistical time period, notching nearly 35 explosive plays and having one of the highest average yards per rush in the grouping. Closer to the origin, an intra-cluster grouping exists of other talented running backs such as Arian Foster and Demarco Murray. The results of the clustering confirm that if you have more than 20 explosive plays, it is likely that your career is among some of the best in the history of the league. When it comes to the orange cluster, a.k.a players who didn’t run often but were explosive when they did, we see a mix of players who do not fit the running back mold. In fact, there are several punters who fit inside of the orange cluster including Johnny Townsend, whose only run of his career was for 42 yards, skyrocketing his average yards per rush. The result? Punters can run too!

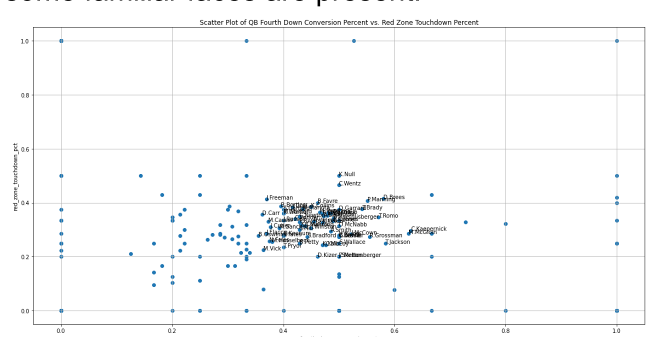
For our last cluster analysis result, we sought to answer the question “how do quarterbacks perform in the key situations of 4th down passes and ‘red zone’ passes?” The initial plot of the data appeared randomly scattered with random centroids selected:



After subsequent iterations of k-means, clear groupings became apparent from the data:



Our cluster results indicate that a clear grouping of quarterbacks exists between 50% red zone touchdowns and 50% fourth downs converted. When applying labels to the clustered data, some familiar faces are present:



A conclusion that can be drawn from this graph is that most successful quarterbacks will fall somewhere in this cluster. Names like Tom Brady and Aaron Rodgers are apparent in the grouping, but some names are rather interesting. Keith Null is just above current NFL quarterback Carson

Wentz in terms of end-zone touchdown percentage, despite only having thrown for 3 touchdowns in his entire career. Another name that sticks out is the one-and-only Colin Kaepernick. The player-turned-activist has one of the best fourth-down conversion percentages among players in the cluster and has a red zone touchdown percentage higher than some of the active quarterbacks in the cluster as well. As a result, Kaepernick could be considered one of the most reliable quarterbacks in crunch-time decisions. Can someone sign him already?

APPLICATIONS

The methodology and conclusions drawn from our analysis of fantastic football feats could be used in a variety of situations related to football, and sports in general. NFL coaches and scouts are primarily concerned with crafting gameplans that give their players and teams the best chances to win. In order to win, you must score more points than your opponent. Through our analysis, it was determined that one of the attributes most correlated to increasing the expected points total was first down passing. Coaches could add more first down passes to their gameplans, and increase their points added per play to gain an edge over the competition.

Another application avenue includes those who bet on sports, and football specifically. Having a model which operates at the granular level of play-by-play can boost expected outcomes from bets. With the advent of online and mobile gambling, some sportsbooks offer the ability to bet on plays as they happen. A bettor might wholly benefit from a model which factors in points scored per play, especially when evaluating odds that are more advantageous to teams who score more points. Bettors might also benefit from specific outlier analysis to inform them which players are outperforming their competitors in an individual stat category, optimizing their betting strategy to favor a certain statistic.

Sportswriters might benefit from aspects of our clustering and outlier analysis in work on stories about a player or a duo of players. Sportswriters are also voters for key distinctions given after a season, such as MVP. The same writers who vote on MVP are also voters for the NFL Hall-of-Fame, recognizing the greatest players in history. While sportswriters always have their finger on the pulse of what's happening in the league in any given season, having another tool for characterizing player success through our clustering analysis might prove to be invaluable.

Finally, anyone interested in sports could apply our methodology to other leagues, teams, and players to get a deeper sense of what contributes to success. Analysis methods used in this report such as linear regression and k-means clustering are applicable in many different contexts. Each method comes with a set of knobs and levers which can be augmented to fit the data in question. Hopefully, our project inspires others to find fantastic feats in their field of interest.

REFERENCES

[1] Kevin Clark. 2018. The NFL's Analytics Revolution Has Arrived. (Dec. 2018). Retrieved March 14, 2022 from https://www.theringer.com/nfl/2018/12/19/18148153/nfl-analytics-revolution

[2] Mark Horowitz. 2018. Detailed NFL Play-by-Play Data 2009-2018. (Dec. 2018). Retrieved March 14, 2022 from https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016?select=NFL+Play+by+Play+2009-2018+%28v5%29.csv

[3] Berri D.J., Burke B. 2012. Measuring Productivity of NFL Players. In: Quinn K. (eds) The Economics of the National Football League. Sports Economics, Management and Policy, vol 2. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-6290-4_8

[4] Warner, J. 2010. Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.8415&rep=rep1&type=pdf

[5] Uzoma et al., 2015 A Hybrid Prediction System for American NFL Results. International Journal of Computer Applications Technology and Research Volume 4– Issue 1, 42 - 47, 2015, ISSN:- 2319–8656 https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1038.1982&rep=rep1&type=pdf

[6] Gifford, M. and Bayrak, T., 2020. What Makes a Winner? Analyzing Team Statistics to Predict Wins in the NFL. https://core.ac.uk/download/pdf/326836349.pdf