Identifying Defensive Playstyles in the National Football League (NFL)

University of California, Davis

STA 160: Practice in Statistical Data Science

Prepared for:

Dr. Fu-Shing Hsieh

Prepared by:

Ethan Park

1.0 Introduction

In football, defense can be the difference between a good team and a great one. Although offensive explosiveness typically dominates headlines, it is oftentimes the defensive side that determines whether a team can consistently compete at a championship level. Defensive strategies vary widely across the NFL and are heavily influenced by coaching philosophies, player personnel, and opponent tendencies. Some teams take aggressive approaches by blitzing often and aiming for high-impact plays like sacks and turnovers, while others prefer conservative zone defenses that prioritize stopping large gains. Understanding and categorizing these playstyles is not just for hypothetical understanding—it's crucial for scouting, game-planning, and team identity formation.

In a league where even slight advantages can change entire outcomes, identifying and comparing defensive nuances across teams offers practical insight. It allows teams to prepare better for opponents, evaluate player fit, and understand league-wide trends. The typical classifications, such as "4-3 vs. 3-4 defenses," are too simple and do not reflect the nuances of modern NFL schemes, which are becoming increasingly meshed and situational. A data-driven approach allows us to move beyond labels and discover natural groupings of team behavior.

This project's central research question asks: Is it possible to identify distinct defensive playstyles among NFL teams using data-driven clustering techniques? By aggregating play-level data and focusing on eight critical features, we apply K-means clustering to group teams based on their defensive tendencies. This allows us to build an interpretable classification of NFL defenses that reflects how teams actually play on the field.

2.0 Data Overview and Exploratory Data Analysis (EDA)

Our analysis uses two primary datasets from the 2025 Big Data Bowl: plays.csv and player_play.csv. The plays.csv file contains 16,124 plays across 50 variables, offering valuable information for each snap like down, distance, possession, and defensive alignment. The

player_play.csv dataset provides a much more detailed view, with 354,727 player-level entries and 50 variables, showcasing individual player actions like pressures, coverage type, and yards gained or allowed. These datasets collectively offer a comprehensive coverage of large game context and individual player contributions, combining to form holistic team-level outcomes. We identify defensive plays using the defensiveTeam column in plays.csv, and limit our scope to plays that were valid and not nullified by penalties (playNullifiedByPenalty != 'Y'). This ensures that the metrics reflect actual strategic execution rather than interruptions. Key defensive metrics—such as causedPressure and sackYardsAsDefense—are extracted from player_play.csv, then aggregated at the team and game level to construct the following features for clustering analysis:

– **Pressure Rate**: The proportion of defensive plays in which any player was credited with causing pressure on the quarterback. This is a key marker of how aggressively a team tries to disrupt the opposing offense.

– **Man Coverage Percentage**: The proportion of plays where the team was recorded using man coverage, derived from pff_manZone. This variable reveals tendencies in coverage ideology.

– **Zone Coverage Percentage**: The complement to man coverage for interpretability and validation, although often redundant.

– **Average EPA per Play:** Mean expected points added (EPA) given up per defensive play. This metric estimates the overall impact of a defense on opponent scoring efficiency—more negative values indicate better performance.

– Average Yardage per Play: Measures the average number of yards allowed by a defense per play. It is a straightforward gauge of how much field position is conceded on each down.

– **Quarterback Hit Rate:** Percentage of plays in which a defender made contact with the quarterback (excluding sacks). This captures the physical toll exerted on quarterbacks even when a sack isn't recorded.

– **Sack Yardage Average per Play:** The average number of yards lost due to sacks on a per-play basis. This combines frequency and impact of sacks to reflect how damaging a defense can be when it gets home.

– **Interception Rate per Play:** The frequency at which a defense records an interception, normalized by total plays. It reflects ball-hawking ability and coverage effectiveness in creating turnovers.

Initial exploratory plots of a select set of variables revealed meaningful variance across teams:

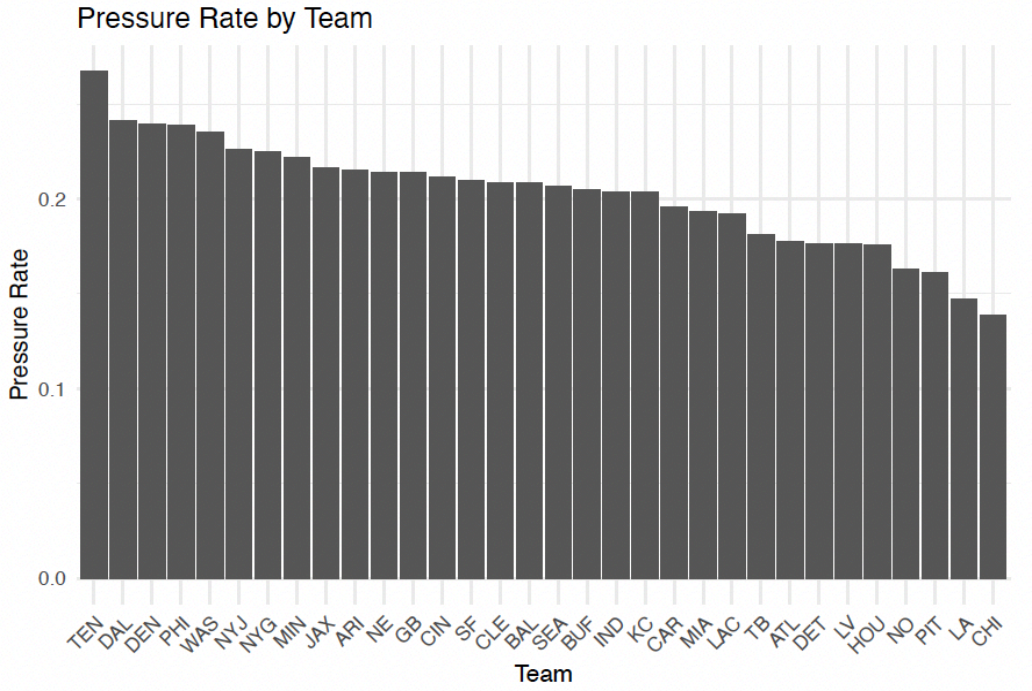

Figure 1. Pressure Rate by Team (in descending order)

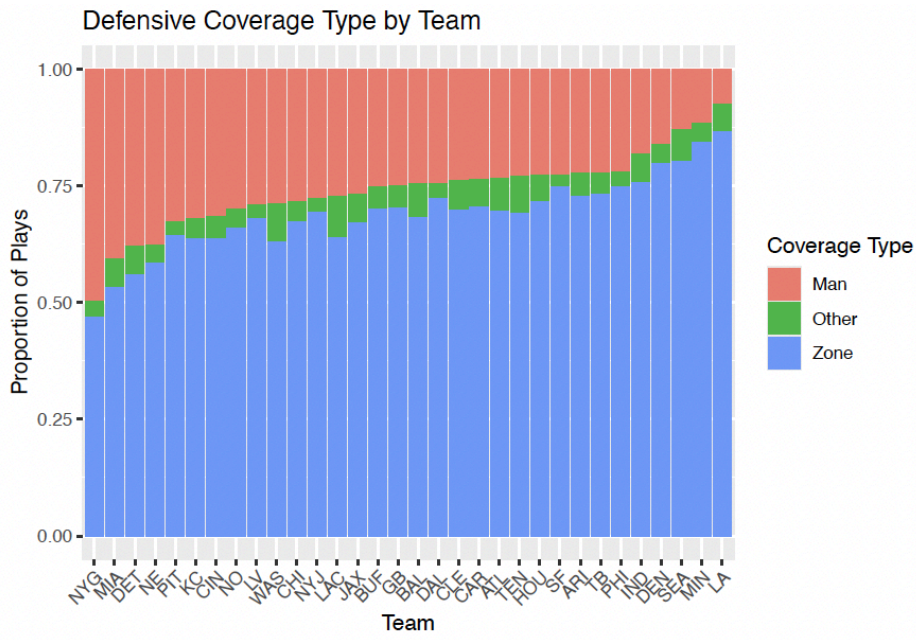

Figure 2. Defensive Coverage Distribution for Each Team

These visuals show that defenses like the Tennessee Titans (TEN) and the Dallas Cowboys (DAL) excel in pressure, while teams like the LA Rams (LA) use man coverage sparingly. Some defenses employed man coverage on over 60% of snaps, while others dropped below 20%. Pressure rate ranged from under 20% to over 40%.

To support the clustering analysis, we selected the above features based on both game importance and their suitability for unsupervised learning:

2.1 Theoretical Justification:

Each variable encapsulates a unique and strategically meaningful dimension of defensive play. Together, they reflect both tactical intent and execution quality:

– **Average EPA and Average Yards Allowed:** These are aggregate indicators of how effective a defense is at limiting opponent advancement and scoring probability. EPA, in particular, accounts for context (down, distance, field position), making it a powerful performance summary. Yards per play complements this by offering a more straightforward, raw measure of play success.

– **Man and Zone Coverage Percentages:** These features speak to the structural identity of the defense. High man coverage usage often signals a trust in individual matchups and an aggressive, blitz-compatible posture. In contrast, high zone usage indicates a preference for discipline and coverage over disruption. Together, they form a variety of defensive philosophies.

– **Pressure Rate and Quarterback Hit Rate:** These metrics are essential for understanding how a defense influences the quarterback. Pressure rate quantifies overall disruption, while hit rate adds a layer of physicality and risk imposition, even on plays that don't result in sacks. These variables often separate passive contain defenses from attack-minded fronts.

– **Sack Yardage per Play:** This hybrid metric combines the frequency and severity of sacks. It reflects how damaging a team's pressure is in terms of field position loss, and serves as a proxy for high-leverage defensive success.

– **Interception Rate:** The only turnover-related feature, this captures a defense's ability to generate takeaways, disrupt passing lanes, and shift momentum. It connects the secondary's performance with overall schematic success in confusing and out-smarting quarterbacks.

Altogether, these eight features provide a comprehensive profile of a defense's priorities, risk tolerance, and execution efficiency. Teams may emphasize some areas (e.g., pressure or coverage) while sacrificing others (e.g., yardage allowed), and these trade-offs form a basis for strategic clustering.

2.2 Analytical Sustainability:

From a modeling point of view, the selected features meet key criteria for effective clustering. All are numeric and standardized to ensure equal influence within the distance-based K-means algorithm. While some features are related (e.g., pressure rate and sack yardage), exploratory correlation analysis (Figure 3) will show sufficient uniqueness to justify inclusion. In addition, these variables are measured at the team level and aggregated across a full season, reducing the influence of outliers or short-term noise. The balance of performance outcomes (like EPA) with style indicators (like coverage rates) ensures that the clusters reflect *how*

defenses operate—not just how well they perform. Finally, these features are interpretable and actionable. Coaches and analysts understand and use these metrics regularly, increasing the utility of the clustering output. The goal is not simply segmentation, but providing meaningful categories that can drive future evaluation frameworks.

2.3. Correlation Matrix

To evaluate the interrelationships among the eight selected features, a correlation matrix is computed. This provides insight into which variables move together and helps assess potential redundancy before clustering.

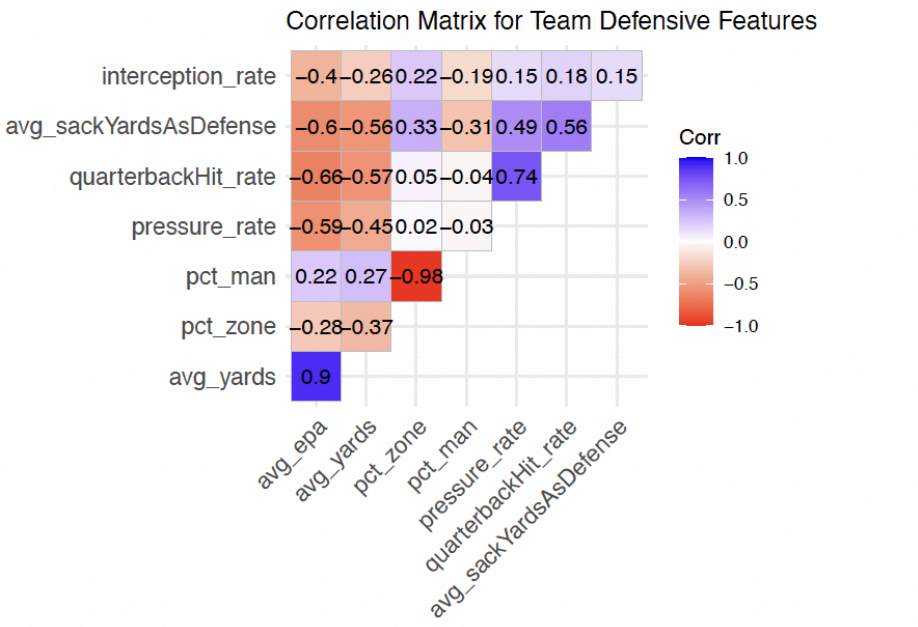

Figure 3. Correlation matrix for team-level defensive features

To discover potential redundancy and variable relationships, all eight defensive features were analyzed in this correlation matrix. Average EPA and average yards allowed were highly correlated (r = 0.90), reflecting their shared role as overall performance indicators. Man and zone coverage percentages were almost perfectly inversely correlated (r = -0.98), so only man coverage was retained for interpretability. Pressure rate, quarterback hit rate, and sack yardage

showed moderate to strong positive correlations (r = 0.49–0.74), forming a coherent group of pass-rush metrics. Interception rate was moderately negatively correlated with both EPA and yards allowed, suggesting a link between takeaways and suppressing opponent efficiency. Overall, the features offered distinct insights that supported their joint inclusion in clustering.

3.0 Methodology

K-means clustering has been selected in this project due to its interpretability and effectiveness for identifying groupings based on numeric features. It is well-suited for discovering natural clusters in structured, continuous variables like pressure rate and coverage frequency. Before clustering, both variables were standardized (z-score normalization) to ensure they contributed equally to the Euclidean distance metric used by K-means. We tested values of K ranging from 1 to 10 using the elbow method, examining the within-cluster sum of squares (WCSS) as a function of k.

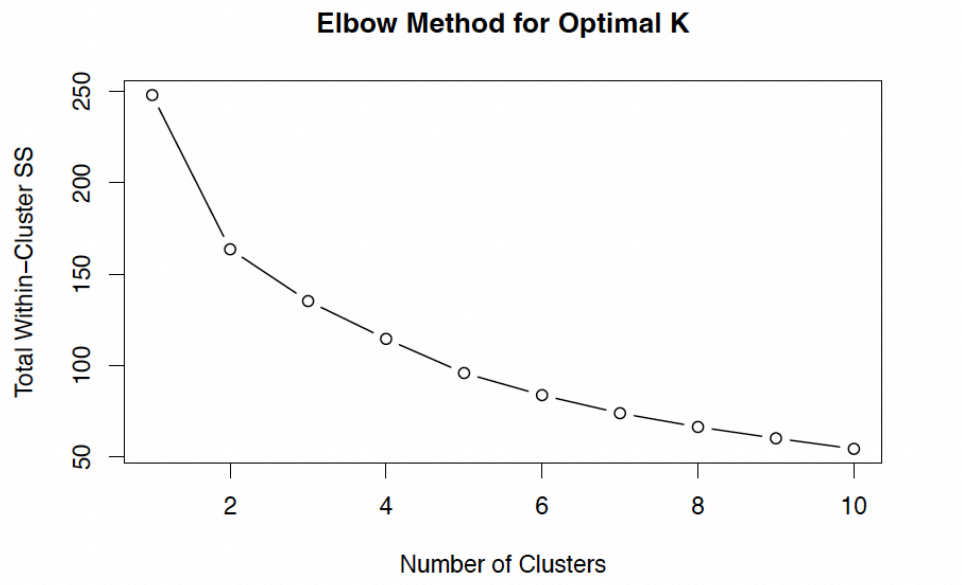

Figure 4. Elbow Method Chart for choosing an optimal K value

The plot exhibits a clear elbow at K = 3, indicating a point of diminishing returns in reducing WCSS with more clusters. This was further supported by silhouette analysis, which measured how similar each team was to its own cluster versus other clusters. A peak in average

Silhouette score at k = 3 provided empirical backing for our choice. Each NFL team was treated as a single observation in an eight-dimensional feature space. The K-means algorithm initialized centroids randomly (with reproducibility via a seed) and iteratively updated them to minimize intra-cluster variance. Convergence typically occurred within 10 iterations, with stability checked across multiple runs. While K-means assumes spherical clusters and equal variance, which may not always hold perfectly in real-world football data, our eight-feature setup and pre-standardization mitigated these concerns. The result is a compact, interpretable segmentation of the NFL landscape based on defensive tendencies.

4.0 Results

Using K-means clustering with $k = 3$, teams were grouped into three defensively distinct clusters based on eight standardized variables. A cluster profile plot (Figure 5) visualizes the mean values of each metric by cluster, providing a comprehensive view of playstyle differences.

Cluster 1

Teams: NYG, CAR, MIA, ATL, DET, LV, HOU, PIT, CHI

These defenses allowed the most average yards per play (5.91) and the highest average EPA (0.05), indicating less overall efficiency in stopping opponents. However, they exhibited moderate man coverage use (0.32), balanced zone usage, and mid-range values for pressure-related metrics. Their interception rate (0.0101) and sack yardage (0.15) were the lowest across all clusters.

These teams often give up yardage and scoring efficiency, potentially due to softer coverage shells or lower disruption. They represent a style that bends frequently and occasionally breaks.

Cluster 2

Teams: TEN, DAL, DEN, PHI, WAS, NYJ, NE, BUF

This group produced the lowest EPA per play (-0.119) and allowed the fewest yards (5.05), reflecting elite defensive effectiveness. They also led in pressure rate (0.233), quarterback hit rate (0.107), sack yardage per play (0.292), and interception rate (0.0163). Coverage was mostly balanced with 25.6% man and 69.6% zone, leaning toward zone structures. These defenses disrupt passing plays, finish sacks, and create turnovers.

Cluster 2 teams display high-impact, efficient defense through sustained disruption and sound structure, achieving both pressure and turnover creation.

Cluster 3

Teams: MIN, JAX, ARI, GB, CIN, SF, CLE, BAL, SEA, IND, KC, LAC, TB, NO, LA

These defenses had moderate yardage allowed (5.41) and near-average EPA (-0.03), with the highest zone usage (72%) and lowest man coverage (22.6%). They showed middle-tier pressure (0.20) and sack metrics (0.24), with interception rates similar to Cluster 1 (0.0107).

These defenses take a bend-don't-break approach with soft zone shells, limiting explosive plays but producing fewer disruptive outcomes. They are more reactive than aggressive.

The cluster features can also be visually compared side-by-side for further insight:

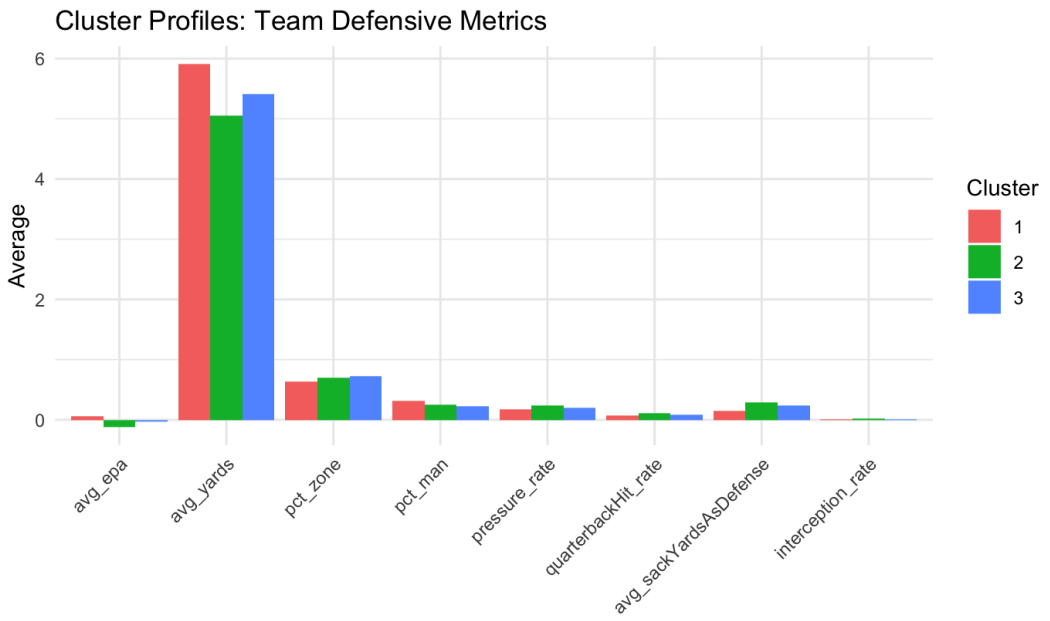

Figure 5. Cluster Profiles: Bar chart displaying average values of all eight metrics across clusters.

This cluster configuration reveals that defensive success (low EPA and yards) can coexist with both zone-heavy and pressure-heavy schemes (Cluster 2), while less successful defenses (Cluster 1) may suffer from inadequate disruption and limited ball-hawking ability. Cluster 3 teams appear structurally disciplined but lack the game-changing traits observed in Cluster 2.

5.0 Discussion and Reflection

This clustering analysis highlights that NFL defenses can succeed through diverse strategies—some dominate with pressure and takeaways, while others rely on structural discipline and limiting big plays. Effective playstyles are not one-size-fits-all. Cluster 2 defenses, marked by disruptive metrics, showed the strongest performance, while Cluster 1 teams struggled due to lower pressure and turnover generation. These insights suggest that blending disruption with sound coverage is a more consistent path to defensive efficiency. Ultimately, data-driven characterization offers valuable tools for comparing defensive philosophies and identifying areas for strategic refinement.

Several challenges arose during the execution of this analysis. First, interpreting defensive style from numeric summaries required simplification of more nuanced football concepts. For example, coverage shells like Cover 6 or match-zone hybrid schemes are not captured in binary man-zone splits. Similarly, pressure rate does not distinguish between simulated pressure and true blitzing. Another challenge was ensuring variable balance during clustering. Strong correlations (e.g., between man and zone coverage, or between pressure metrics) required careful preprocessing and interpretation to avoid redundancy while still capturing meaningful variation. We chose to retain some correlated variables (e.g., hit rate and sack yardage) to reflect different expressions of pass-rush success.

From a modeling perspective, K-means assumes spherical clusters and equal variance, which may not hold in this football context. Some teams might occupy transitional or hybrid identities that are better represented with soft clustering methods or time-aware techniques. For instance, a team may shift identity mid-season due to injuries or strategic reevaluation—K-means treats them as fixed. Finally, the project demanded significant iteration in feature engineering, data cleaning, and exploratory analysis. Extracting and transforming variables from play-level data (e.g., from player_play.csv) into usable team-level aggregates involved filtering for valid plays, interpreting null entries, and dealing with column inconsistencies like ambiguous naming. Despite these limitations, the clustering analysis remains valuable. It offers an interpretable structure that captures major axes of defensive behavior in the NFL. With future refinements—such as incorporating tracking data or temporal variation—this framework could develop into a more dynamic and predictive system of defensive profiling.

6.0 Conclusion

This project applied K-means clustering to identify distinct defensive playstyles among NFL teams using a multidimensional dataset of eight performance and strategic metrics. The analysis revealed three coherent clusters, each representing unique tactical identities. Despite

limitations such as simplified coverage labeling and static season-long aggregation, the model provides a meaningful framework for understanding how NFL defenses differentiate themselves. By combining statistical computation with football relevance, this work displays the value of data-driven approaches in characterizing team strategy and performance.

Bibliography

NFL Big Data Bowl 2025. (2025). Kaggle.
[https://www.kaggle.com/competitions/nfl-big-data-bowl-2025/overview]