Proposal: Identifying NFL Quarterback Archetypes via K-Means Clustering

Shubhan Tamhane

Undergraduate Data Science Student

University of Connecticut

February 19, 2026

Introduction The quarterback (QB) position is widely regarded as the most significant role in American football, yet conventional statistics used to evaluate QB performance, such as passer rating and completion percentage, collapse complex, multidimensional playing styles into a single value. I chose this topic because NFL tracking and play-by-play data currently offer rich, high-quality performance metrics that allow for a mathematical examination of whether meaningful groupings exist among quarterbacks beyond what traditional rankings reveal. Prior research has established that combinations of efficiency, mobility, and decision-making metrics are closely associated with team offensive success (Quealy and Carter, 2013), and studies in sports analytics have demonstrated that unsupervised clustering methods can uncover latent player archetypes that do not map cleanly onto conventional positional labels (Lutz, 2012). A data-driven approach to identifying quarterback archetypes could support scouting, player comparison, and defensive decision-making by offering a newer characterization of playing style.

Specific Aims The research question of this project is: **Can K-means clustering applied to career-level NFL quarterback performance data identify distinct, interpretable playing style archetypes?** This study plans to analyze patterns across passing efficiency, volume, mobility, and decision-making metrics to determine whether meaningful QB groupings emerge from the data without imposing labels before our clustering. This question is feasible given publicly available NFL play-by-play and season summary data and can be addressed using dimensionality reduction and unsupervised learning methods.

Data Description The dataset used in this study consists of career-aggregate and season-level performance statistics for NFL quarterbacks sourced from publicly available repositories, including Pro Football Reference. The sampling scheme is observational and includes all quarterbacks who have met a minimum threshold of career passing attempts, tentatively set at 100 and 20 rushing attempts, to ensure that the sample reflects players with meaningful NFL experience and to limit the influence of small-sample noise.

Key variables of interest include passing efficiency measures (completion percentage, passing yards, touchdown rate, interception rate), which are continuous variables. Mobility metrics such as rushing attempts per game, rushing yards per game, and longest rush will also be included as continuous variables. The primary dataset is publicly accessible, though certain advanced tracking metrics from proprietary systems such as Next Gen Stats may have restricted availability.

Research Design and Methods This project will use a cross-sectional, player-level design to model stylistic groupings among NFL quarterbacks using unsupervised learning. A feature set will first be constructed by selecting and engineering performance metrics that capture the key dimensions of quarterback play: efficiency, volume, mobility, and risk-taking. All features will be standardized using z-scores prior to clustering to ensure that variables with larger absolute magnitudes do not disproportionately influence the Euclidean distance calculations underlying K-means (Celebi et al., 2013). Highly correlated features will be examined where appropriate to reduce redundancy, and Principal Component Analysis (PCA) may be applied as a preprocessing and visualization step.

K-means clustering will then be applied with the k-means++ initialization strategy to improve solution quality and reduce sensitivity to initial centroid placement (Celebi et al., 2013). Solutions for k ranging from 2 to 10 will be evaluated using the elbow method, which examines within-cluster sum of squares as a function of k, and the silhouette criterion, which measures the separation of clusters (Hubáček et al., 2019). The value of k that produces a visible inflection in the elbow plot and a reasonably high average silhouette score will be selected as the final solution. Once a final solution is identified, each cluster will be profiled through descriptive statistics and visualized using radar plots and parallel coordinate plots, with representative players identified as those nearest to their cluster centroid.

Discussion Distinct quarterback archetypes, such as pocket passers, dual-threat quarterbacks, and game managers, are expected to emerge from the clustering analysis, and these groupings should be interpretable in terms of well-known stylistic differences that analysts and coaches recognize. The idea that unsupervised learning can recover meaningful structure from performance data would be strengthened if the resulting clusters map onto coherent statistical profiles and are populated by players who are widely understood to play the game in similar ways. Prior work in basketball and soccer has shown that data-driven clusters frequently reveal subtler distinctions than traditional labels capture (Lutz, 2012; Decroos et al., 2019), and similar nuance may emerge here.

If the clustering solution yields poorly separated or uninterpretable clusters, it may indicate that quarterback performance data lies on a more continuous spectrum than a discrete archetype model assumes, or that the selected features do not adequately capture the key dimensions of the predicted variation. Limitations of this study include the use of career-aggregate statistics, which obscure within-career stylistic evolution, the assumption of spherical clusters inherent to K-means, and the subjectivity involved in post-hoc archetype labeling.

Core elements of Data Science This project has important data science components. Career-level and season-level tracking data will be collected, cleaned, and modeled using programming in Python. Data management tasks include merging data from multiple sources, handling missing values for advanced metrics, applying eligibility filters based on career attempt thresholds, and constructing normalized feature vectors for each player. The data analysis process will draw on feature engineering, dimensionality reduction via PCA, unsupervised machine learning via K-means, and internal cluster validation using elbow and silhouette methods (Hubáček et al., 2019; Celebi et al., 2013). Cluster profiles and archetype comparisons will be communicated through radar plots, scatter plots of PCA-reduced data, and descriptive summary tables that coaches and analysts can interpret without statistical expertise. Data ethics will be addressed by using only publicly available data, refraining from overstating the precision or generalizability of cluster assignments, and being transparent about the limitations of post-hoc labeling.

Conclusion The goal of this research is to use career-level NFL performance data to build a data-driven framework for identifying quarterback playing style archetypes through K-means clustering. The study aims to determine whether meaningful and interpretable player groupings emerge from a curated set of passing, mobility, and efficiency metrics without imposing categories from prior research. The proposed statistical methods are well-suited to the available data and are feasible within the project timeline. The ultimate goal of this research is to contribute to the growing body of work on unsupervised learning in sports analytics while offering a practically useful tool for player evaluation, scouting, and strategic planning in professional football.

References

Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013), “A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm,” *Expert Systems with Applications*, 40, 200–210.

Decroos, T., Bransen, L., Van Haaren, J., and Davis, J. (2019), “Actions Speak Louder than Goals: Valuing Player Actions in Soccer,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, pp. 1851–1861.

Hubáček, O., Šourek, G., and Železný, F. (2019), “Incorporating Domain Knowledge in Machine Learning for Soccer Outcome Prediction,” *Machine Learning*, 108, 97–126.

Lutz, J. (2012), “Using Cluster Analysis to Derive a More Sophisticated Typology of NBA Player Roles,” *Journal of Quantitative Analysis in Sports*, 8.

Quealy, K. and Carter, S. (2013), “Quarterback Rankings: Beyond the Passer Rating,” The New York Times.