Extracting Patterns and Feature Correlations from NFL Stadium Arrest Data

David Chung

Data Mining

Mike Weisman

11/19/2019

Contents

INTRODUCTION

American football is widely considered to be the most popular professional sport to watch in the United States. From the regular season months of September through December and concluding with the playoffs in January and February, the National Football League (NFL) draws attention from fans all across the country. Most fans watch their favorite teams compete from the comfort of their own homes or the local bar; however, a considerable number of them purchase tickets (or season passes) and go to NFL stadiums to watch their teams in person. On average, each NFL stadium welcomed 68,400 live spectators per game for the 2011-2015 regular season [1]. These consistently high attendance numbers not only attract passionate fans, but can also invite physical clashes between fans of opposing teams, rowdy game watchers, etc.

This study examined whether certain game attributes are correlated with the number of stadium arrests on a particular game day. The dataset contains game and stadium arrest data acquired from regular season games played during the 2011 to 2015 NFL regular seasons. The summary statistics (e.g. shape, center, spread) of the distribution of the arrest dataset is characterized before any analysis is conducted. Initial clustering algorithms were then used to determine similarities between NFL stadiums regarding the number of arrests and their home team's performance. Afterwards, a factor analysis of mixed data (FAMD) was conducted to determine which game attributes contain the most information regarding the number of arrests that could be expected at a particular venue. The results of the unsupervised learning and dimensionality reduction analyses are discussed, and recommendations for future studies are proposed.

DESCRIPTION OF FEATURES AND DATASET

The game and arrest data were acquired from public records requests submitted by the Washington Post to the local police departments that manage security at the 31 NFL stadiums [2]. Twenty-nine of the 31 police departments provided data (i.e. Cleveland and New Orleans did respond to the requests). However, five of the twenty-nine responders delivered partial data. The Buffalo police department omitted arrest numbers for several regular season games. St. Louis only gave year-by-year arrest data. Detroit, Minneapolis, and Atlanta exclusively provided arrest data inside the stadium (i.e. parking lot arrests were excluded). Due to the variation in data provided by each police department, the complete game and arrest data for twenty-five NFL teams were used for the analyses. It should be noted that the New York Giants and New York Jets both play their home games in MetLife Stadium.

There are four quantitative and five categorical features in this dataset. The features include numerous game attributes, such as the week of the NFL regular season (17 week-long NFL regular season), the day of the week that the game took place (Sunday, Monday, Thursday, Saturday), the local game time, the home and away teams of the game, the home and away team scores, an indicator variable that flagged if the game required overtime, and an indicator variable

that flagged if the game was a "rivalry" game (i.e. the two teams that played are in the same division) [2]. Although the data for seven NFL teams were incomplete, the remaining twenty-five NFL teams contributed game data for 966 total regular season games. The following section examines the distribution of the stadium arrests based on the available data.

DISTRIBUTION OF STADIUM ARRESTS

Before conducting any analyses to extract patterns from the data, the distribution of the number of stadium arrests was studied to characterize its shape, center, and spread. Figure 1 below plots a histogram of the number of stadium arrests reported on a single regular season game day:

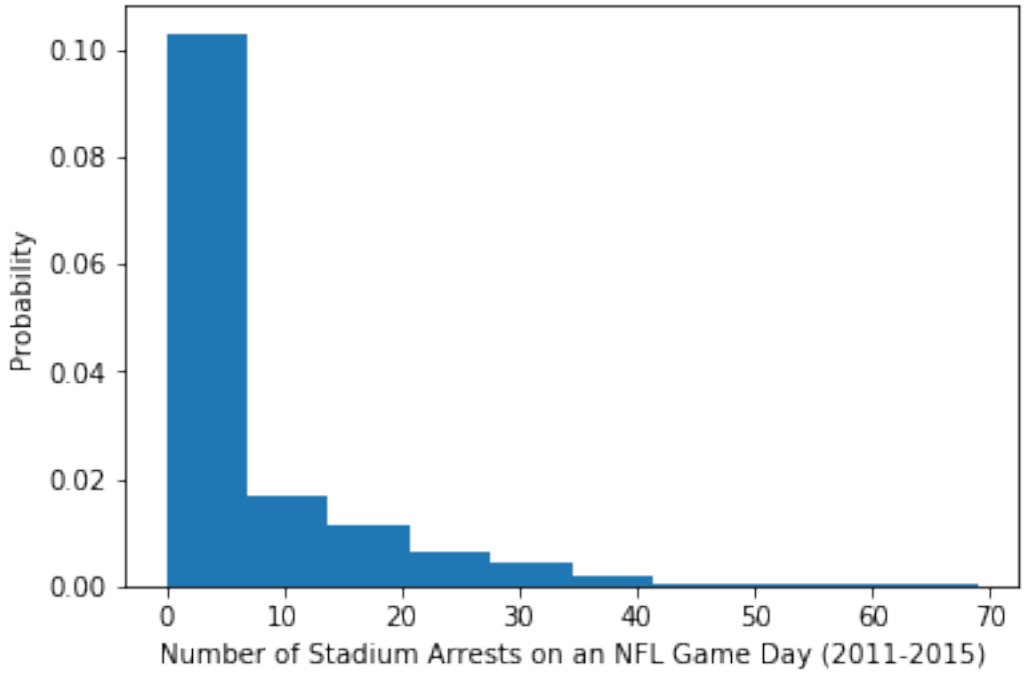

Figure 1: Histogram of numbers of stadium arrests on single NFL game days (2011-2015 regular season)

As evidenced above, the distribution of stadium arrests is heavily skewed to the right, which indicates that very few arrests are typically reported on the average regular season game day. This finding is further supported by noting that the mean of the distribution (7 arrests per game day) is greater than the median of the distribution (3 arrests per game day). In fact, the mode of the distribution is 0 arrests per game day, which was reported for 215 regular season games from the dataset. Thus, there are often no arrests reported at NFL stadium on most regular season game days (i.e. the observance of multiple arrests at an NFL stadium on game day is a moderately rare event).

While zero was the lowest number of arrests observed on a single game day, there was one police department that reported 69 stadium arrests at the game. The game was a Week 10 regular season game held at Qualcomm Stadium between the San Diego Chargers and the Oakland Raiders that was played on Thursday evening, November 10, 2011 [3]. This observation is fascinating when considering the generally great climate and temperament of the citizens living in San Diego.

After examining the marginal distribution of stadium arrests for the reported NFL stadiums, the arrest statistics for each reported stadium were scrutinized in further detail. Figure 2 plots the total number of reported arrests for each stadium during the 2011-2015 NFL regular seasons:

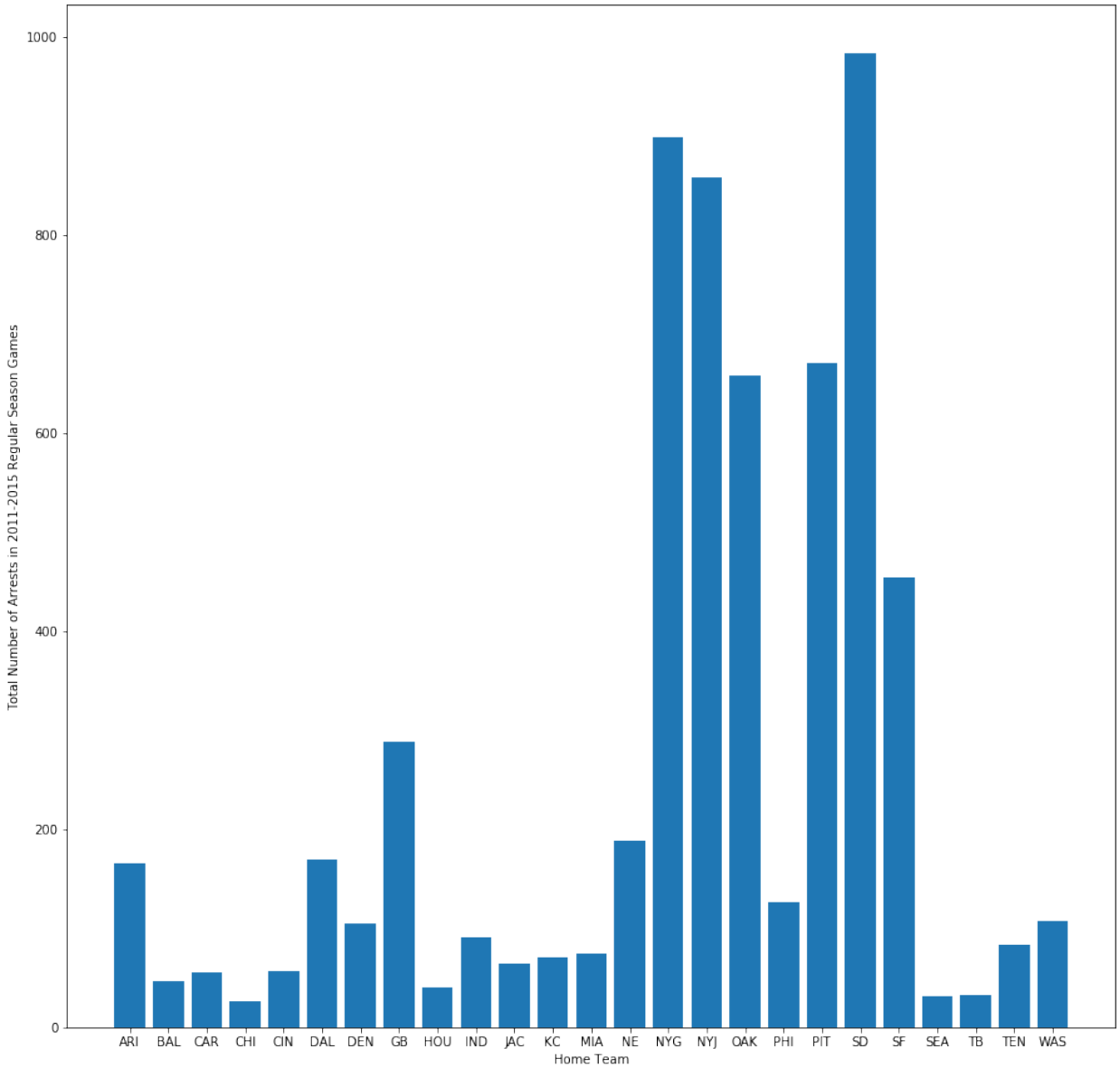

Figure 2: Total reported stadium arrests for 25 NFL teams (2011-2015 regular season)

The arrest totals convey that San Diego not only reported the largest number of stadium arrests on a single game day, but also the highest total of stadium arrests (983) for the 2011-2015 regular seasons. Both New York teams reported the second and third-highest arrest totals (899 and 858, respectively). Pittsburgh and Oakland, two cities with renowned football franchises and fanbases, also reported the fourth and fifth-highest arrest totals (670 and 658, respectively). San Francisco also reported considerably more arrests than other stadiums (454).

Once the total stadium arrests were plotted, the median numbers of stadium arrests were computed to determine the midpoint of the conditional distribution arrests for each reported stadium. Figure 3 plots the median numbers of arrests:

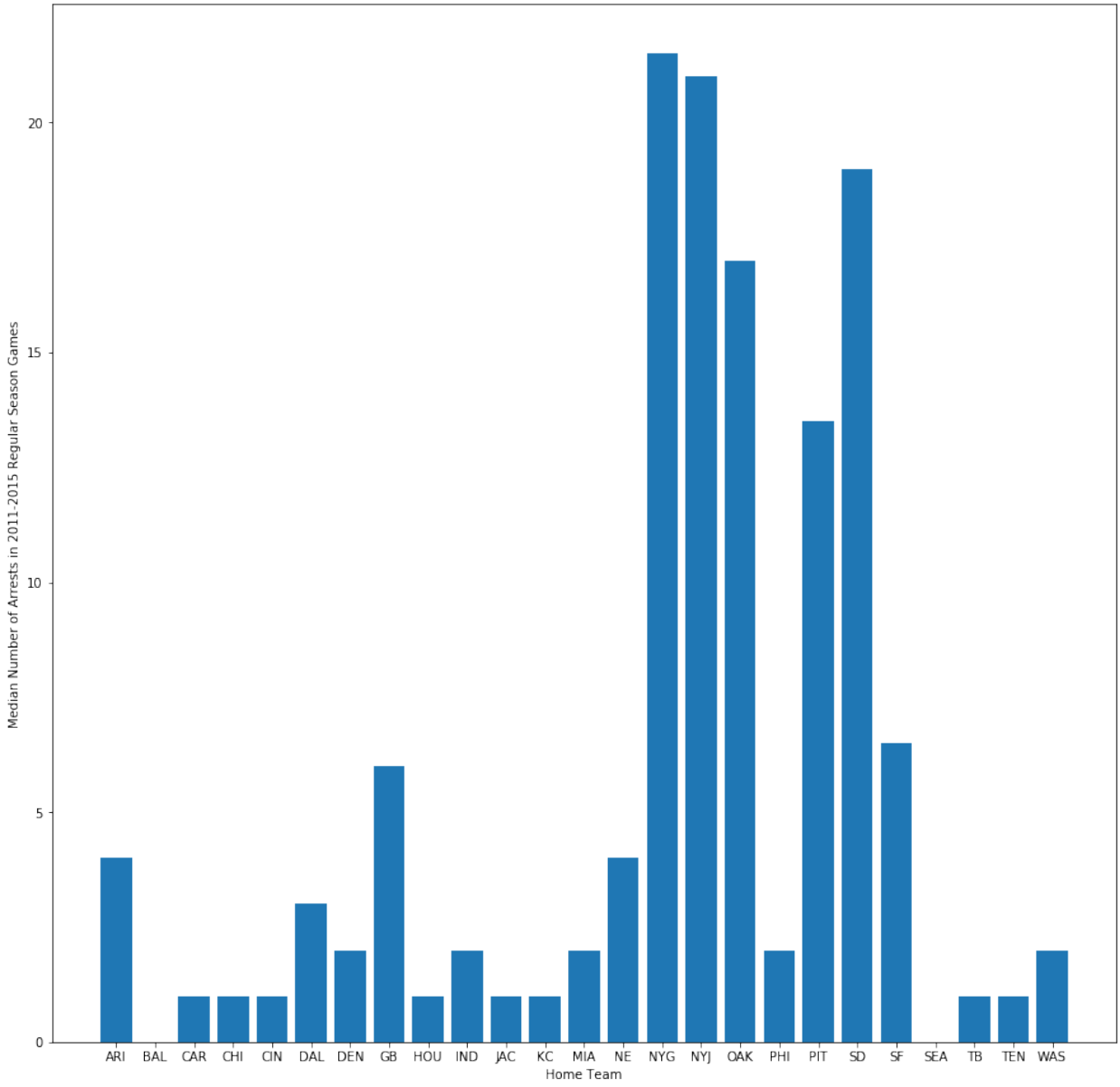

Figure 3: Median number of stadium arrests for 25 NFL teams (2011-2015 regular season)

Surprisingly, San Diego did not report the highest median number of arrests on a single game day. Instead, both New York teams reported the largest median numbers of arrests on single game days (21.5 and 21, respectively), while San Diego had the third-highest median number of stadium arrests (19). Oakland and Pittsburgh also reported the fourth and fifth-highest median numbers of arrests (17 and 13.5, respectively). Interestingly, San Francisco reported the sixth largest number of total arrests; however, its median number of single day stadium arrests (6.5) is comparable to the 19 remaining teams.

The deviations between the median and total arrest numbers observed for the San Diego and San Francisco data subsets suggest that the conditional distributions of stadium arrests for both teams are skewed. Figure 4 displays box-and-whisker plots of the single day stadium arrests for the 25 teams considered in the study:

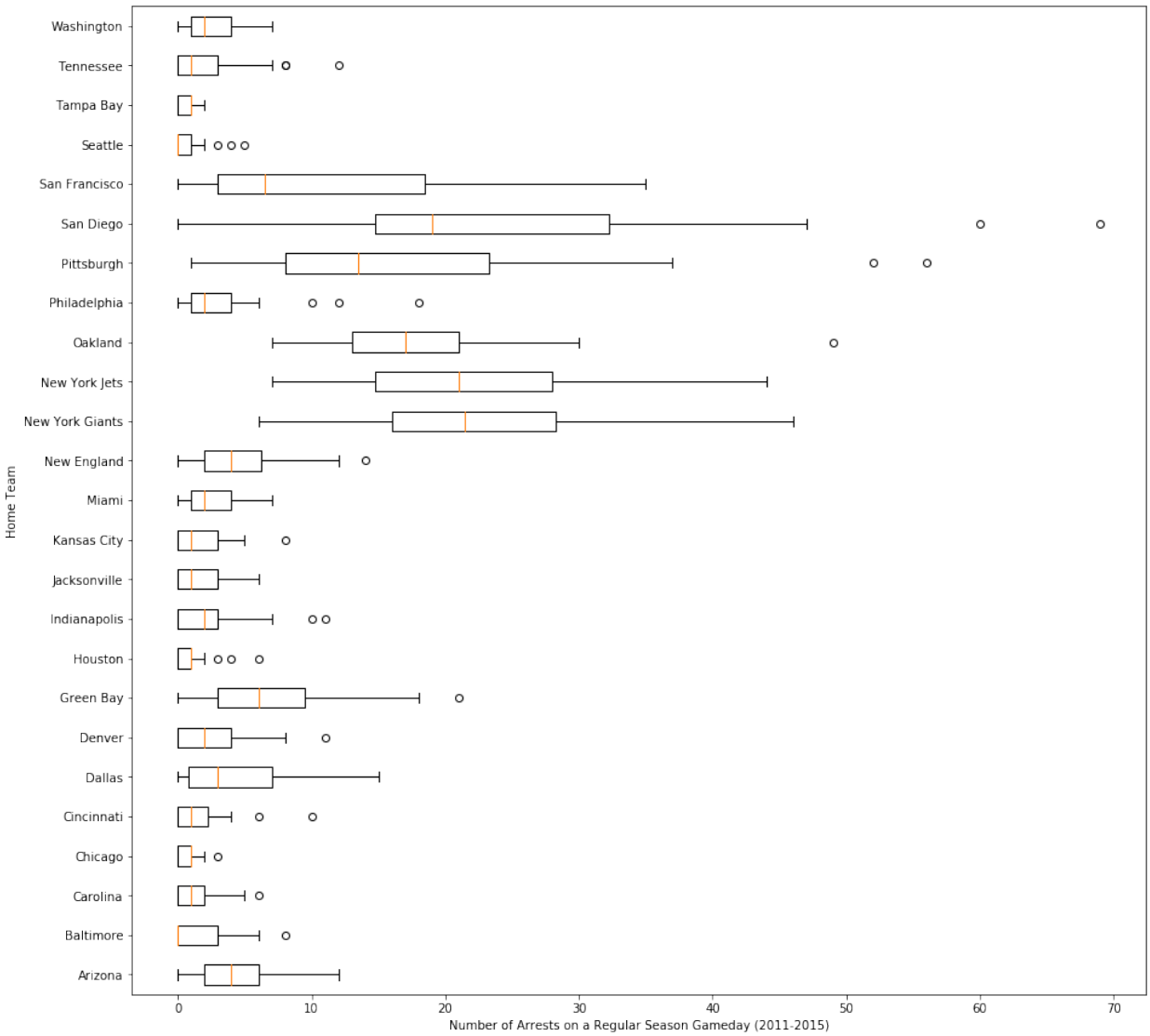

Figure 4: Box-and-whisker plots of single day stadium arrests for 25 NFL teams (2011-2015 regular season)

As evidenced above, the distribution of arrests for San Diego is slightly skewed to the right with two abnormally large outliers (60 and 69 arrests). On the other hand, the conditional distributions of stadium arrests for both New York teams appear to be virtually symmetric with a high median number of arrests. Thus, these results confirm that San Diego reported the highest total number of arrests, yet both New York teams had the highest median number of single day stadium arrests.

Similarly, the conditional distributions of single day arrests for Oakland and Pittsburgh also convey analogous patterns. Although Pittsburgh reported slightly higher arrest totals than Oakland over the five regular seasons, Oakland showed a higher median number of a single day arrests than Pittsburgh. The conditional distribution for Pittsburgh is slightly skewed to the right with two high outliers (52 and 56 arrests) while the conditional distribution for Oakland is symmetric with a lower median number of arrests and a single outlier (49 arrests).

Although the box-and-whisker plot for the single day arrests for San Francisco contains no outliers, the distribution is heavily skewed to the right. This also suggests that a few regular season games in San Francisco reported very large numbers of arrests that skewed the total (and thus, the mean) number of reported arrests to be misleadingly higher than the median number of arrests, which was pedestrian.

Generally speaking, the median numbers of single day stadium arrests for most NFL teams' home games fall between 0 to 6 arrests per day. The five outlier teams that were previously mentioned (San Diego, New York Giants, New York Jets, Oakland Raiders, Pittsburgh Steelers) exhibited significantly higher median numbers of single day arrests.

K-MEANS CLUSTERING – SIMILARITIES IN ARRESTS AND GAME OUTCOMES

After characterizing the conditional distribution of arrests for each NFL stadium, the next task was to implement a *k*-means clustering algorithm on the data to identify clusters of NFL teams that reported comparable arrest frequencies and team performance. Each teams' median performance was measured using its median margin of defeat (i.e. the result of the median home team score subtracted from the median away team score for each NFL team). This metric was selected because it quantifies each NFL team's typical level of success during the 2011-2015 regular seasons (i.e. a high median margin of defeat suggests that a team lost most of its home games, and vice versa). Additionally, the *k*-means clustering algorithm can achieve greater degrees of separation using features that are continuous instead of discrete features with few possible values [4].

The KMeans function from the Sci-kit Learn Cluster library was used to conduct *k*-means clustering on the data [5]. After the median numbers of single day arrests and margins of defeat were computed for the twenty-five NFL teams, the number of cluster centers K, the coordinates of the cluster centroids $\underline{\mu}_j = (\mu_{1j}, \mu_{2j}), j = 1, \ldots, K$, and the convergence tolerance δ were

initialized. In addition, the feature pairs for each NFL team were standardized by centering each feature to its mean value and scaled down to unit variance. Since the *k*-means clustering algorithm is isotropic, the computed clusters are circular and not elongated [4]. Thus, the features are scaled to ensure that the scales between the median numbers of arrests and margins of defeat are compatible.

For each NFL team with feature coordinate $\underline{X_i} = (X_{1i}, X_{2i}), i = 1, \dots, 25$, where X_{1i} and X_{2i} represents the median number of single day stadium arrests and margin of defeat for the *i*th team, the squared Euclidean distances between each feature coordinate and cluster centroid were computed [4]:

$$d_{ij} = \left\| \underline{X_i} - \underline{\mu_j} \right\|^2, i = 1, \dots, 25 \ \& \ j = 1, \dots, K$$

Each NFL team is assigned to the cluster whose centroid has the least squared Euclidean distance (i.e. $y_i = \{j: \left\| \underline{X_i} - \underline{\mu_j} \right\|^2 \leq \left\| \underline{X_i} - \underline{\mu_k} \right\|^2 \forall k, k = 1, \dots, K$). Once the cluster assignments are completed, the cluster centroid coordinates are updated:

$$\underline{\mu_j}^{(t+1)} = \frac{1}{n_j^{(t)}} \sum_{l=1}^{n_j^{(t)}} \underline{X_l}^{(t)}$$

where $n_j^{(t)}$ is the number of NFL teams assigned to the *j*th cluster after iteration *t* of the algorithm. The assignment and update steps are repeated until the change in the centroid coordinates is less than the specified convergence tolerance (i.e. convergence is reached when $\sum_{j=1}^{K} \left\| \underline{\mu_j}^{(t+1)} - \underline{\mu_j}^{(t)} \right\|^2 < \delta$) or when the maximum number of iterations is reached. A convergence tolerance of $\delta = 0.0001$ was selected, and a maximum of 300 iterations was enforced for this clustering algorithm.

A total of ten *k*-means clustering epochs were conducted with different initialization of the cluster center coordinates. After ten simulations were conducted, the centroid coordinates were selected that yielded the lowest inertia, which is defined as the sum of the squared Euclidean distances between each feature coordinate and its assigned cluster centroid coordinate [5]:

$$I = \sum_{i=1}^{25} \left\| \underline{X_i} - \underline{\mu_i} \right\|^2$$

where $\underline{\mu_i}$ is the cluster center that is assigned to feature coordinate $\underline{X_i}, i = 1, \dots, 25$. In other words, the inertia $I = \sum_{i=1}^{25} \left\| \underline{X_i} - \underline{\mu_i} \right\|^2$ is the loss function that will be minimized for this unsupervised learning problem.

After several epochs of the *k*-means algorithm with several numbers of cluster centers, it was determined that eight clusters captured the most informative groupings of NFL teams. Figure 5 plots the median margin of defeat and number arrests for the twenty-five NFL teams and their corresponding cluster assignments:

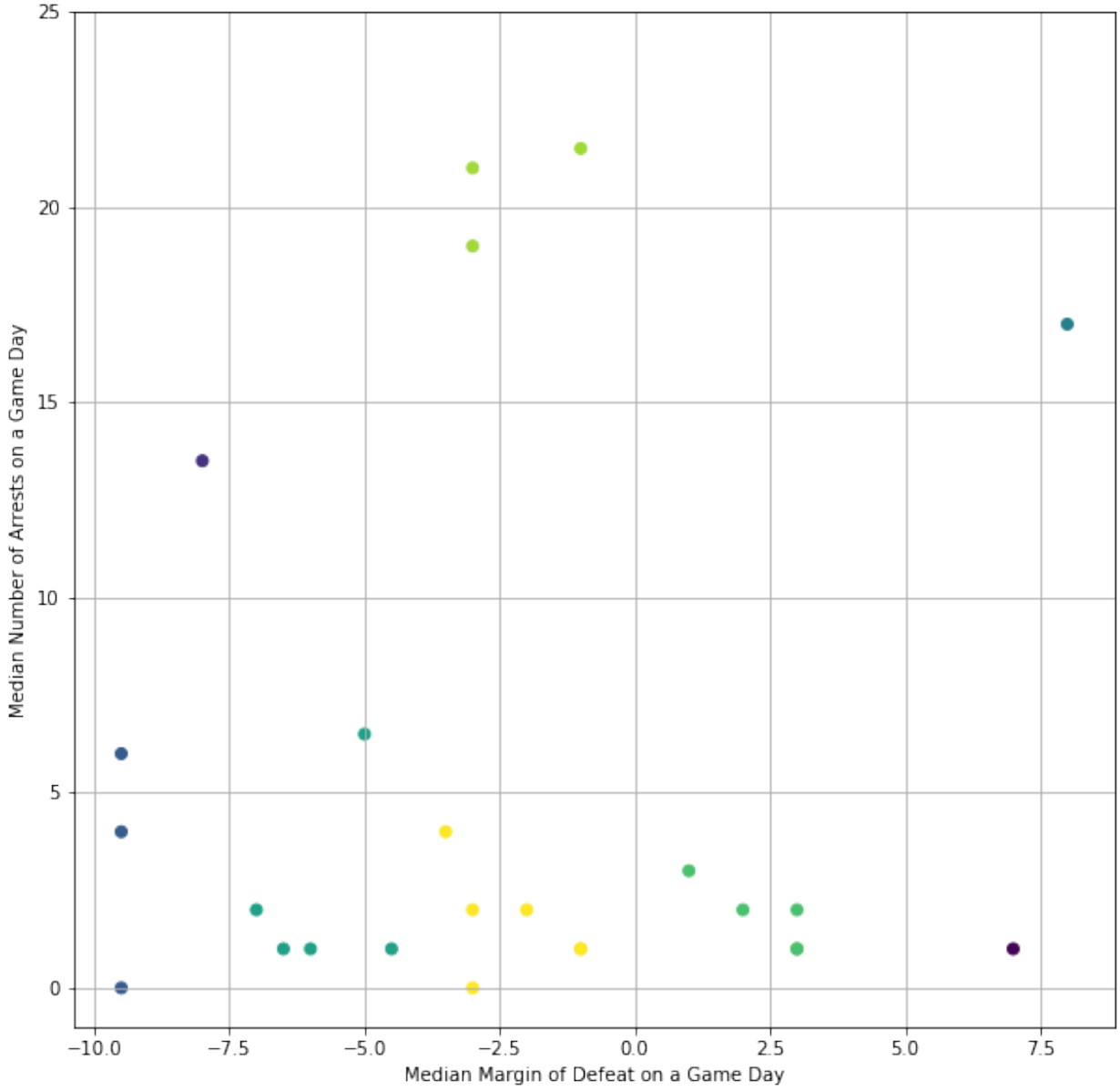

Figure 5: Median numbers of single day arrests and margins of defeat of and cluster assignments of 25 NFL teams

The simulation converged with a model inertia of 1.926, which resulted in eight distinct clusters that insightfully grouped all 25 NFL teams. The most apparent observations are the five NFL teams whose median numbers of arrests are convincingly higher than the remaining twenty

teams. The New York Giants, New York Jets, and San Diego Chargers were assigned to a single cluster with a centroid of 20.5 arrests per game and a -2.33 margin of defeat. This result suggests that these three teams managed to win mostly competitive home games and reported many arrests at the stadiums.

On the other hand, the Oakland Raiders and the Pittsburgh Steelers were assigned to their own clusters. Both of the teams reported large numbers of arrests, but their teams' performances during these regular seasons are polar opposites. The Oakland Raiders lost by at least one touchdown for the majority of it home games, while the Pittsburgh Steelers defeated their opponents by at least one touchdown.

The remaining twenty teams were grouped into five clusters. The Jacksonville Jaguars were assigned to its own cluster with a centroid of one arrest per game and a 7.0 margin of defeat. This finding is indicative of a franchise that performed poorly at their home venue (i.e. lost games by more than a touchdown) and recorded some of the lowest attendance figures among all NFL stadiums [6]. With few fans in attendance and many convincing defeats, the chances of observing an arrest are very likely to decrease substantially.

The Dallas Cowboys, Philadelphia Eagles, Tampa Bay Buccaneers, Tennessee Titans, and Washington Redskins were assigned to the same cluster centered at 1.8 arrests per game and a 2.4 margin of defeat. In contrast to the Jaguars, these five teams appeared to be competitive in their home games and lost by a small margin. Furthermore, they reported almost one more arrest per game than the Jaguars did.

The Arizona Cardinals, Baltimore Ravens, Chicago Bears, Indianapolis Colts, Kansas City Chiefs, and Miami Dolphins were assigned to a cluster with a centroid of 1.67 arrests per game and a -2.25 margin of defeat. These six teams played many competitive home games and barely managed to win. They also reported a similar number of arrests to the five aforementioned teams that lost numerous close home games.

The Carolina Panthers, Cincinnati Bengals, Denver Broncos, Houston Texans, and San Francisco 49ers were assigned to a cluster whose centroid coordinate is 2.3 arrests per game and a -5.8 margin of defeat. These five teams defeated most of their opponents at home by slightly less than a touchdown, and they also reported slightly more arrests per game than both the Jacksonville Jaguars and the NFL teams involved in close games.

Lastly, the Green Bay Packers, New England Patriots, and Seattle Seahawks were grouped into the final cluster. The centroid of this cluster is 3.33 arrests per game and a -9.5 margin of defeat. These three teams seem to epitomize the term "home-field advantage". They convincingly defeated many of their opponents at their home stadiums, and there appeared to be one more reported arrest per game than the five previously discussed teams.

Upon analysis of the remaining twenty teams, it appears that the variation in the median number of arrests increases as the median margin of defeat at the home stadium decreases. In other words, the likelihood of observing more arrests at a stadium appears to increase when the home team outscores their opponent by a greater margin. Despite the staggering median numbers of arrests reported by five NFL teams, these values are most likely anomalies. As a result, these five teams were isolated from the main dataset for the remainder of the analysis.

FACTOR ANALYSIS OF MIXED DATA – GAME FEATURE AND ARREST CORRELATIONS

By virtue of *k*-means clustering, a potential relationship between a game feature (i.e. the home team's margin of defeat) and the number of observed arrests was suggested. However, the dataset contains nine total features. The five categorical features and two quantitative features (local game time and week of NFL season) were ignored for the *k*-means clustering simulations due to the mechanism of the algorithm. In addition, *k*-means clustering was implemented to group similar NFL teams based on arrests and home game performance; however, this learning method would not be generalizable to identify correlations between features and arrests for an arbitrary NFL team (i.e. the computed clusters were applicable to the twenty-five NFL teams included in the training dataset).

On the other hand, nine features are possibly extraneous in predicting the number of arrests. Thus, dimensionality reduction was applied to the nine original features in order to select the most important ones and compute the percentages of the total variance that are captured. Factor analysis of mixed data (FAMD) is a dimensionality reduction tool that is applicable to datasets containing both quantitative and categorical features [7]. This hybridized method simultaneously conducts a principal components analysis (PCA) on the quantitative features and multiple correspondence analysis (MCA) on the categorical features of the training set.

Before implementing FAMD, the quantitative features were scaled to unit variance, and the categorical variables were organized into a disjunctive data table and also scaled such that each observation's categorical assignments were represented as the deviations from having independent categorical features. This preprocessing step ensures that both types of features are weighted equally when determining the factors.

Let $k = 1, \dots, 4$ be the number of quantitative features and $q = 1, \dots, 5$ be the number of categorical features in the training set. For each quantitative variable z, define $r(z, k)$ as the correlation coefficient between variables z and k and $\eta^2(z, q)$ as the squared correlation ratio between variables z and q (i.e. the ratio of the variance for values of variable z in category q and the variance for variable z across the entire training set) [7]. In PCA, the principal components are linear combinations of the original features that are collinear with the largest variances of the dataset [4]. They are determined such that the correlations between each principal component and the original quantitative features are maximized:

$$f(z_1, \dots, z_4) = \sum_k \sum_{i=1}^{4} r^2(z_i, k)$$

Analogously, MCA identifies new dimension of such that the correlation ratios between each category and variable *z* are maximized:

$$g(z_1, .., z_4) = \sum_q \sum_{i=1}^{4} \eta^2(z_i, q)$$

FAMD merges the objective functions of the PCA and MCA such that the resulting objective function distributes equal weights to both quantitative and categorical variables. Consequently, the computed factors maximize:

$$h(z_1, \dots, z_4) = \sum_k \sum_{i=1}^{4} r^2(z_i, k) + \sum_q \sum_{i=1}^{4} \eta^2(z_i, q)$$

In contrast to principal components, these factors are latent (unobserved) variables that also capture the variance of the original dataset with a potentially lower number of features [4]. The observed (original features) are linear combinations of these hidden variables. As mentioned in the results of the *k*-means clustering simulations, the game and arrest data for the five outlier NFL teams (New York Giants, New York Jets, San Diego Chargers, Oakland Raiders, Pittsburgh Steelers) were omitted from the FAMD.

It was determined that approximately 60% of the total variance could be represented by three latent factors. Table 1 below lists the correlations of the three factors and their explained variance ratios:

Table 1: Feature correlations and explained variance ratios of factors determined by FAMD

Factor	**1**	**2**	**3**
Game Feature			
OT Flag = 'OT'	-0.529	0	-0.409
OT Flag = 'RG'	0.529	0	0.409
Day of Week = 'Mon'	-0.470	-0.071	0.630
Day of Week = 'Sat'	-0.262	0.000	0.149
Day of Week = 'Sun'	0.766	-0.080	-0.894
Day of Week = 'Thu'	-0.491	0.168	0.571
Division Game = 'No'	0.455	-0.998	0.000
Division Game = 'Yes'	-0.455	0.998	0.000
Local Game Time	-0.506	0.010	0.624
Week Number	-0.106	0.223	0.000

Explained Variance Ratio	0.461	0.093	0.031

According to the FAMD results, the first factor represents 46.1% of the total variance. It is highly correlated with multiple categorical game features, such as the day of the week that the game was played, the local game time, and whether the game needed to continue into overtime. In contrast, the second factor accounts for 9.3% of the total variance; however, it is predominantly correlated with one categorical feature: whether the game was played between two teams in the same division. While the first factor recaptures the bulk of the original dataset's variance, the second factor primarily determines if a game was played between two rivals. Lastly, the third factor represents only 3.1 % of the total variance. It is correlated with the local game time and the day of the week that the game was played.

Interestingly, the FAMD determined that the home and away teams' scores are uncorrelated with the computed factors. This result suggests that the amount of points that the home team scores or surrenders is irrelevant in explaining the majority of the variance in the game data. In fact, the local game time and week number were the only two quantitative variables that exhibited some correlation with the latent factors, but both of these features' correlations were generally lower than those of the categorical features.

Therefore, the FAMD results suggest that game features that measure the game's level of competitiveness were the most meaningful variables of the original dataset (e.g. whether the game was a divisional matchup, needed to conclude in overtime, whether the game was played early or later in the regular season, whether the game was a prime-time game on Sunday, Monday, Thursday, or Saturday evening). This finding parallels the observed trend from the *k*-means clustering simulations; teams that generally perform well in their home stadiums exhibit greater variations in the number of reported stadium arrests than teams that are generally not competitive at their home stadium.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

This study analyzed NFL game and stadium arrest data for possible patterns or trends that could possibly explain why certain NFL stadiums report more arrests than others. Examining the marginal distribution of arrests in all NFL stadiums revealed that the typical NFL stadium reports very few arrests per game day. However, there are a few NFL stadiums that report significantly larger numbers of arrests at their stadiums, as evidenced by examining the conditional distributions of stadium arrests for each stadium. The *k*-means clustering algorithm was implemented to group NFL teams based on their median numbers of single day arrests and their median margins of defeat at their home stadiums. Ignoring the five stadiums that generally reported higher arrest numbers than the remaining game venues, the remaining NFL stadiums reported higher variations in the number of stadium arrests when their home teams played well. This outcome was further substantiated by a FAMD conducted on the nine original game

features of the training set. The latent variables were most correlated with the game features that described the level of competition between the two teams in a game.

Using the results obtained from this analysis, a predictive model could be constructed to estimate the expected number of arrests at a particular NFL stadium by using information about the game. The original game features can be re-expressed in terms of the latent variables determined from the FAMD, which will simplify the regression model or random forest ensemble that are constructed from the training data. In the event that a valid predictive model cannot be constructed, another *k*-means clustering simulation can also be conducted on the transformed game data to potentially uncover additional trends in stadium arrests based on the latent factors.

Works Cited

[1] S. Das, "Top 10 Most Popular Sports in America 2019 (TV Ratings)," Sports Show, August 10 2019. [Online]. Available: https://sportsshow.net/most-popular-sports-in-america/. [Accessed 17 October 2019].

[2] The Washington Post, "Kaggle," 2016. [Online]. Available: https://www.kaggle.com/washingtonpost/nfl-arrests. [Accessed 29 September 2019].

[3] A. Keatts, "Do the San Diego Chargers really have the NFL's worst behaved fans?," 20 December 2016. [Online]. Available: https://www.theguardian.com/sport/2016/dec/20/san-diego-chargers-most-arrested-fans. [Accessed 18 October 2019].

[4] R. O. Duda, Pattern Classification, Second Edition, Danvers, MA: John Wiley & Sons, Inc., 2001.

[5] Scikit-learn, "sklearn.cluster.KMeans," 2019. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. [Accessed 25 October 2019].

[6] ESPN, "NFL Attendance," 2019. [Online]. Available: http://www.espn.com/nfl/attendance/_/year/2015. [Accessed 3 November 2019].

[7] F. Husson, "FAMD: Factor Analysis for Mixed Data," 3 July 2019. [Online]. Available: https://rdrr.io/cran/FactoMineR/man/FAMD.html. [Accessed 23 October 2019].