**Exploratory Analysis of NFL Receiving Statistics and NFL Arrests Statistics**

**Team Members:**

**Josh Janda (joshlj2), Jack Massey (jmassey4), Ashley Arroyo (netid), Isais Lopez (netid)**

**Introduction**

For our project, we have chosen two NFL datasets which relate to statistics for receiving players from 2000 to 2016 by week and arrests statistics 2000 to 2017. Our interest in these datasets is to analyze how statistics differ between teams as well as years. These statistics include both playing statistics as well as arrest statistics. Some background information for the [first dataset](https://data.world/chrishayles/nfl-data/workspace/file?filename=receiving.csv)’s source is that it is from the website *Kaggle* which is a website that provides numerous datasets for data science competitions as well as data analysis. The [second dataset](https://www.kaggle.com/patrickmurphy/nfl-arrests) is gathered from the website *data.world*, which similarly to *Kaggle,* provides real world datasets for data analysis/data science competitions.

Overall, our main goal for this project is to analyze these datasets to answer questions regarding the most successful teams and players and if data on passing yards is in line with this belief, how risky a team is determined by number of arrests, what positions and side of the ball are arrested more frequently, and are better teams more likely to have more frequent arrests.

**Methods**

Our first dataset, mentioned above, includes the receiving statistics from 2000 until 2016 and has 67,761 observations of 14 variables. The variables and descriptions are:

* name = "Last, First Name of Player"
* team = "Team of Player"
* rec = "Total Number of Receives"
* yds = "Total Number of Yards"
* tgt = "Total Number of Targets"
* avg = "Average Yards per Receive"
* td = "Number of Touchdowns"
* fstdn = "Number of First Downs"
* pct = "Percentage of First Downs for each Receive"
* lng = "Longest Gain/Receive for Week"
* fum = "Number of Fumbles"
* fuml = "Number of Fumbles Lost"
* season = "Season Year"
* wk = "Week of Season"

Our second dataset, mentioned above, includes the arrests statistics of players from 2000 until 2017 and includes 850 observations of 8 variables. The variables and descriptions are:

* date = "Date of Incident"
* team = "Team of Player"
* name = "Name of Player"
* position = "Player's Position"
* case = "Incident Type"
* category = " Incident Crime Categories"
* description = "Description of Crime"
* outcome = "Incident outcome description"

For validating and cleaning the data, we are using guidelines that are of our knowledge of NFL statistics. Such as what the correct number of unique teams should be equal to, whether the statistics make sense or not, and analyzing for misspellings and mistyped values. We will be validating and cleaning the data through the use of multiple techniques, such as frequency tables, checking variable uniqueness, single variable analysis for missing/extreme values, and adding labels and appropriate formats to variables.

For starting, we first utilized the *nlevels* frequency table for both the arrests and receiving datasets to check for missing values and incorrect number of unique levels (**Table 1 and** 2). For our first issue, we noticed that the number of teams in our receiving dataset was not equal to the total current number of NFL teams (**Figure 1)**. This is due to the Saint Louis Rams moving to Los Angeles. This will be left as is due to it being a valid observation. Our second issue is that the names of the teams for the arrest and receiving dataset were not all the same throughout **(Figure 2 compared to Figure 1)**. This will be fixed by renaming the incorrectly abbreviated teams in the receiving dataset to avoid complications when merging later on. Our third issue was noticed when checking whether there were any players where receptions was greater than targets. This should not be possible, and we found one observation where this is the case (**Figure 3**). This was remedied by setting the target variable to the appropriate value which was obtained through Google. Our fourth issue was discovered when looking at the number of arrests per positions (**Figure 4**), in which the “DE” position had a “/” attached to the end. We know there is no “DE/” position, so this was remedied by setting the position to “DE”.

For cleaning, we have created labels for each variable in both datasets which the labels are included in the above descriptions of the data. We have also created a format for the “wk” variable in the receiving dataset, which changed the weeks from 101-117 to 1-17 to make the week numbers clearer as well as creating a separate arrests dataset which indicates whether a player is offense or defense.

Some additional data preparation we have performed is the use of subsetting, conditioning, and merging the datasets. For merging, we have merged the datasets based on team and included relevant statistics by team which are total yards and total number of arrests for all wide receivers (**Figure 5**). For subsetting, we have grouped the arrests data by offense or defense and then analyzed the total number of arrests by side of the ball (**Figure 6**). We have also grouped the arrests data by team to look at the frequency of arrests per team, and then lastly grouped by position to look at the frequency of arrests per position (**Figure 7 and 8**). Regarding the best players, we have grouped the players by year and found the player with the max number of yards (**Figure 9**). For the best teams, we have grouped the data by team and look at the total yardage for each team over all years (**Figure 10**). To confirm our belief that more yardage equates to a better team, we have looked at one team over one season (New England Patriots) and looked at the total yardage for this team for each game (**Figure 11**).

For analyzing variables, we have chosen to analyze them by group and through frequency tables. These tables can all be seen, and are mentioned, above.

Overall, we have read in and validated/cleaned the data and then utilized numerous summary/grouping/frequency techniques to answer our research questions.

**Results**