Exploratory Analysis of NFL Receiving Statistics and NFL Arrests Statistics

Team Members:
Josh Janda (joshlj2), Jack Massey (jmassey4), Ashley Arroyo (aarroy30), Isaias Lopez (ielopez2)

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'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 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

In Table 1, we see that there are three missing values from "Average Yards per Receiver", "Percentage of First Downs for each Receiver", and "caught/targets". We also note that while there are 33 teams, there are only 32 since one team changed city.

In Table 2, we see a missing value for "Incident outcome description".

In Figure 1, we see the distribution of players who are receiving among the teams. There should only be 32 teams, but the dataset includes STL which is now LA because the team changed cities.

In Figure 2, we are using the arrested dataset. There are is a 35 total of "Team Identifier at time of incident". Thirty-five are listed because the data covers the time between January 2000 to March 2017 and we discovered that some teams moved cities. This figure also includes distribution of arrested players by team.

In the "Incident Type" table, we see the frequency and percentage specific degrees of sentences and consequences due to player's crime for "case 1". The category with the most frequency is "arrested" (731), with players being "charged" as second. Players detained, or have died, or were summoned is only has a frequency of 1.

In Figure 3, we set our target to 1 after finding that it had a 0, an error.

In Figure 7, group by processing through a query was used to sort teams by the number of player arrests. The team with the highest player arrests is MIN with 49 arrests. The lowest is LAC with 1. We thought that there might be some connection between the arrest statistics of the teams and the actual cities where they played but there didn't seem to be a connection.

In Figure 4 and 8, the number of arrests were grouped by position. The position with highest number of arrests is WR with 136 arrests. The position with the lowest is OL with 1 arrest. The O-line is regarded as a position that requires more leadership and decision making than WR. This might be why the arrest numbers are so different.

In Figure 6, we use broader categories of positions either on Offense or Defense. Positions on defense have more player arrests than positions on offense with 450. Positions on offense have 400 player arrests. This is a pretty large sample size and discrepancy and would be an interesting research question as to why defenders are arrested more often.

In Figure 9, players with the most yards in a game per season is displayed. Calvin Johnson ran 329 yards in one game in the 2013 season, the most yards receiving in a game per season. He is universally agreed upon to be one of the best receivers of all-time, so this statistic would be a good way to get an idea of who the best players are.

In Figure 10, total yards per team over the course of the time period is displayed on a table. The team with the highest total number of yards is NO with 76701 yards. It turns out the number of receiving yards is a good indicator of success because the top five teams in total yards also had the most playoff appearances over the time period . LA has the lowest yards of any team, with 3313 yards receiving. This could be because recently they were once STL in St. Louis.

In Figure 11, looking at the total number of yards per week by the New England Patriots in season 2016, we find that the week with their highest recorded yards receiving is weeks 5 and 14 with 406. New England Patriots currently has a high attendance in the Super Bowl. NE has a high 71081 total yards run, which supports that teams with more yards receiving tend to be better.
We also looked at actual game results for each of their games during the 2016 season and noted that in each game they had more receiving word they won the game. So, while we saw that receiving yards are an indicator of long term sustained success they are also an indicator of single game success.

In Figure 5, the player arrests and receiving player datasets were inner joined. The table displayed shows total yards and total arrests grouped by teams. Teams that have more yards such as NE (71081 yards) have a low number of arrests (4 arrests). The team with the highest arrests number of arrests is TEN (12 arrests) a total of 60299 yards. While TEN has a lower number of total yards run then NE, the number can be considered large, suggesting that TEN is a "good" team. Looking at the lowest total yards, HOU has a total of 55171 yards, but has one the lowest counts of arrests (3). There is no reasonable evidence to suggest that better teams have more arrests. Our stated belief was incorrect.

After analyzing the receiving and arrests dataset, we were able to answer the questions regarding the most successful teams and players. Initially, we believed that better teams (teams with more passing yards) had more arrests. However, after reviewing figure five, we were able to conclude that better teams do not necessarily have more arrests.

Table 1

The FREQ Procedure

Number of Variable Levels

Variable	Label	Levels	Missing Levels	Nonmissing Levels
name	Last, First Name of Player	2353	0	2353
team	Team of Player	33	0	33
rec	Total Number of Receives	21	0	21
yds	Total Number of Yards	259	0	259
tgt	Total Number of Targets	26	0	26
avg	Average Yards per Receive	445	1	444
td	Number of Touchdowns	5	0	5
fstdn	Number of First Downs	15	0	15
pct	Percentage of First Downs for each Receive	67	1	66
lng	Longest Gain/Receive for Week	116	0	116
fum	Number of Fumbles	4	0	4
fuml	Number of Fumbles Lost	3	0	3
season	Season Year	17	0	17
wk	Week of Season	17	0	17
catch_pct	caught/targets	79	1	78
performance	how well the player did that week	3	0	3

Table 2

The FREQ Procedure

Number of Variable Levels

Variable	Label	Levels	Missing Levels	Nonmissing Levels
date	Date of Incident	783	0	783
team	Team of Player	35	0	35
name	Name of Player	640	0	640
position	Player's Position	18	0	18
case1	Incident Type	10	0	10
category	Incident Crime Categories	107	0	107
description	Description of Crime	842	0	842
outcome	Incident outcome description	50	1	49

Figure 1

The FREQ Procedure

Team of Player				
team	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ARZ	2207	3.26	2207	3.26
ATL	2095	3.09	4302	6.35
BAL	2141	3.16	6443	9.51
BUF	2019	2.98	8462	12.49
CAR	2062	3.04	10524	15.53
CHI	2049	3.02	12573	18.55
CIN	2119	3.13	14692	21.68
CLE	2173	3.21	16865	24.89
DAL	2069	3.05	18934	27.94
DEN	2101	3.10	21035	31.04
DET	2201	3.25	23236	34.29
GB	2255	3.33	25491	37.62
HOU	1802	2.66	27293	40.28
IND	1967	2.90	29260	43.18
JAX	2186	3.23	31446	46.41
KC	2087	3.08	33533	49.49
LA	123	0.18	33656	49.67
MIA	2117	3.12	35773	52.79
MIN	2171	3.20	37944	56.00
NE	2121	3.13	40065	59.13
NO	2247	3.32	42312	62.44
NYG	2071	3.06	44383	65.50
NYJ	2042	3.01	46425	68.51
OAK	2236	3.30	48661	71.81
PHI	2205	3.25	50866	75.07
PIT	2063	3.04	52929	78.11
SD	2004	2.96	54933	81.07
SEA	2188	3.23	57121	84.30
SF	2088	3.08	59209	87.38
STL	2051	3.03	61260	90.41
TB	2156	3.18	63416	93.59
TEN	2156	3.18	65572	96.77
WSH	2189	3.23	67761	100.00

Figure 2

The FREQ Procedure

Number of Variable Levels		
Variable	Label	Levels
team	Team of Player	35

Team of Player				
team	Frequency	Percent	Cumulative Frequency	Cumulative Percent
ARI	21	2.47	21	2.47
ATL	20	2.35	41	4.82
BAL	27	3.18	68	8.00
BUF	19	2.24	87	10.24
CAR	21	2.47	108	12.71
CHI	32	3.76	140	16.47
CIN	44	5.18	184	21.65
CLE	33	3.88	217	25.53
DAL	17	2.00	234	27.53
DEN	47	5.53	281	33.06
DET	18	2.12	299	35.18
Fre	3	0.35	302	35.53
GB	23	2.71	325	38.24
HOU	13	1.53	338	39.76
IND	35	4.12	373	43.88
JAC	34	4.00	407	47.88
KC	32	3.76	439	51.65
LAC	1	0.12	440	51.76
LAR	5	0.59	445	52.35
MIA	31	3.65	476	56.00
MIN	49	5.76	525	61.76
NE	20	2.35	545	64.12
NO	25	2.94	570	67.06
NYG	16	1.88	586	68.94
NYJ	20	2.35	606	71.29
OAK	21	2.47	627	73.76
PHI	18	2.12	645	75.88
PIT	24	2.82	669	78.71
SD	26	3.06	695	81.76
SEA	26	3.06	721	84.82
SF	24	2.82	745	87.65
STL	12	1.41	757	89.06
TB	36	4.24	793	93.29
TEN	36	4.24	829	97.53
WAS	21	2.47	850	100.00

Figure 3

Obs	team	name	rec	tgt
51808	PHI	Peters, Jason	1	0

Figure 4

Player's Position	count
C	5
CB	112
DB	4
DE	68
DE/	1
DT	77
FB	15
K	12
LB	118
OG	25
OL	1
OT	49
P	3
QB	18
RB	97
S	70
TE	39
WR	136

Figure 5

Team of Player	sumarrests	totalyds
NO	4	76701
IND	2	73155
GB	1	71558
NE	4	71081
PHI	2	68355
DEN	8	67733
SD	5	67114
DET	1	66436
NYG	2	65997
PIT	7	65791
DAL	5	65058
ATL	2	63537
MIN	7	62271
KC	5	61961
CIN	10	61780
SEA	2	61665
OAK	2	61139
TB	5	61047
TEN	12	60299
BAL	3	59780
MIA	4	59607
CAR	5	58988
CHI	6	57725
BUF	4	57559
NYJ	3	57501
CLE	6	57232
SF	3	57124
HOU	3	55171

Figure 6

off_def	count
D	450
O	400

Figure 7

Team of Player	NumCases
MIN	49
DEN	47
CIN	44
TB	36
TEN	36
IND	35
JAC	34
CLE	33
CHI	32
KC	32
MIA	31
BAL	27
SD	26
SEA	26
NO	25
SF	24
PIT	24
GB	23
OAK	21
CAR	21
ARI	21
WAS	21
NYJ	20
NE	20
ATL	20
BUF	19
PHI	18
DET	18
DAL	17
NYG	16
HOU	13
STL	12
LAR	5
Fre	3
LAC	1

Figure 8

Player's Position	count
C	5
CB	112
DB	4
DE	69
DT	77
FB	15
K	12
LB	118
OG	25
OL	1
OT	49
P	3
QB	18
RB	97
S	70
TE	39
WR	136

Figure 9

Last, First Name of Player	Season Year	MaxYds
Smith, Jimmy	2000	291
Gardner, Rod	2001	208
Burress, Plaxico	2002	253
Boldin, Anquan	2003	217
Bennett, Drew	2004	233
Chambers, Chris	2005	238
Evans, Lee	2006	265
Curtis, Kevin	2007	221
Owens, Terrell	2008	213
Austin, Miles	2009	250
Britt, Kenny	2010	225
Johnson, Calvin	2011	244
Johnson, Andre	2012	273
Johnson, Calvin	2013	329
Jones, Julio	2014	259
Brown, Antonio	2015	284
Jones, Julio	2016	300

Figure 10

Team of Player	TotYds
NO	76701
IND	73155
GB	71558
NE	71081
PHI	68355
DEN	67733
SD	67114
ARZ	66647
DET	66436
NYG	65997
PIT	65791
DAL	65058
ATL	63537
WSH	62957
STL	62815
MIN	62271
KC	61961
CIN	61780
SEA	61665
OAK	61139
TB	61047
TEN	60299
BAL	59780
JAX	59665
MIA	59607
CAR	58988
CHI	57725
BUF	57559
NYJ	57501
CLE	57232
SF	57124
HOU	55171
LA	3313

Figure 11

Obs	wk	total_yds
1	1	264
2	2	324
3	3	103
4	4	205
5	5	406
6	6	376
7	7	222
8	8	315
9	10	316
10	11	280
11	12	286
12	13	269
13	14	406
14	15	188
15	16	220
16	17	276