**NFL Fan Arrests—Causes and Analysis**

**Executive Summary**

The National Football League boasts 32 unique teams spanning across the greatest of American Cities. From Seattle to New York, and many more in between, these teams come prepackaged with their own histories and reputations. Among these kinds of stereotypes is the perception of the general behavior of fans of each team, with some teams cultivating a more hoolaginistic vibe than others (which is strangely relevant considering the impact the “12th man” can have on the outcome of a game). Speaking from personal experience, for example, I know of many individuals in Las Vegas who are legitimately concerned with the arrival of the Raiders, primarily due to their perception of the quality of Raiders’s fans as opposed to the housing of the team itself.

Focusing in on this particular issue, a natural question arises: “are fan behavior stereotypes valid, and if so, can that behavior be understood, modeled, or predicted?”

To answer this question, a dataset from the Washington Post was found and analyzed, showing the number of arrests by game in the NFL (presumably made up primarily of audience members and fans). There were gaps in the data due to the Washington Post’s collection method and the cities’ inclination to release these statistics (as well as their competence in doing so), resulting in only 25 teams being represented in the analysis, and requiring data imputation in some cases.

Additional sources of data were found and merged with the initial dataset in order to demonstrate possible relationships to fan arrests: a dataset from USA Today showing a history of all NFL player arrests from 2000 to 2020, rates of violent crime for each represented city taken from neighborhoodscout.com, zip code numbers for all stadiums of the represented teams, and missing game data from Wikipedia.com.

These data sources were cleaned and their “city” columns standardized so that, when imported into Tableau, they could become related tables and shown in tandem.

Ultimately the results of the visualizations were, to the author of the project, unintuitive—with very little apparent relationship between fan arrests and factors such as NFL player arrests or city crime rates. However, the fan arrest totals did generally reinforce some preconceived notions, namely that the Raiders, Steelers, Packers, New York teams, and California teams had the most ill-behaved (or more charitably, “passionate”) fans overall.

**Data Preparation**

The primary dataset, “data-nfl-arrests” from the Washington Post, contained most normal-season NFL games from 2011 to 2015, with the following columns:

* season
* week\_num
* day\_of\_week (not used in this study)
* gametime\_local (not used in this study)
* home\_team
* away\_team
* home\_score
* away\_score
* OT-flag (blank if the game did not go into overtime, and “OT” if it did)
* arrests
* division\_game (“n” if it was not a division game, and “y” if it was—this data was used in some capacity during the creation of visualizations but no visualizations kept it by the end)

The first step in data cleaning was to update the OT\_flag and division\_game variables to numeric categorical variables.

data.fillna({'OT\_flag':0}, inplace=True)

data['OT\_flag'] = data['OT\_flag'].replace(['OT'],1)

data["OT\_flag"]=pd.to\_numeric(data["OT\_flag"])

data['division\_game'] = data['division\_game'].replace(['n'],0)

data['division\_game'] = data['division\_game'].replace(['y'],1)

data["division\_game"]=pd.to\_numeric(data["division\_game"])

data.head()

Once completed, the data was checked for missing games. As a team in a normal season has eight home games, it was fairly simple to figure out which games, among the teams represented in the dataset, were missing. Most of these were due to overseas matches, as a few games each year at the time were played in London. Not wanting to skew the results of the study by reporting on fewer games in the season for some teams compared to others, these values were imputed.

A function was created in Python to help automate the imputation process. By being passed the relevant statistics of the missing game (gathered from Wikipedia), the function calculates the mean of the arrests from the other games in that season, and then creates a new line in the dataframe with that information.

def imputeLondon(year, home, away, homescore, awayscore, OT, division):

new = data[(data['home\_team'] == home) & (data['season'] == year )]["arrests"].mean()

data.loc[len(data)] = np.array([year,0,0,0,home,away,homescore,awayscore,OT,new,division])

return;

imputeLondon(2013, "Arizona", "Houston", 30, 9, 0, 0)

imputeLondon(2013, "Jacksonville", "San Francisco", 10, 42, 0, 0)

imputeLondon(2014, "Jacksonville", "Dallas", 17, 32, 0, 0)

imputeLondon(2015, "Jacksonville", "Buffalo", 34, 31, 0, 0)

imputeLondon(2015, "Kansas City", "Detroit", 45, 10, 0, 0)

imputeLondon(2015, "Miami", "New York Jets", 14, 27, 0, 1)

imputeLondon(2014, "Oakland", "Miami", 14, 38, 0, 0)

imputeLondon(2014, "Oakland", "Denver", 17, 41, 0, 1)

imputeLondon(2014, "Oakland", "Kansas City", 24, 20, 0, 0)

imputeLondon(2011, "Tampa Bay", "Chicago", 18, 24, 0, 0)

Unfortunately there were three cases where a city which was participating in the Washington Post’s efforts had one of their five years’ data missing. Not wanting to drop three additional teams from the analysis, the missing year’s worth of data was imputed as well, using a slightly modified function that imputed the new arrests number by taking the mean of the other four years of data for that team.

def imputeYear(year, home, away, homescore, awayscore, OT, division):

new = data[(data['home\_team'] == home)]["arrests"].mean()

data.loc[len(data)] = np.array([year,0,0,0,home,away,homescore,awayscore,OT,pd.to\_numeric(new),division])

data["arrests"]=pd.to\_numeric(data["arrests"]) #kept getting type errors without brute-forcing it

return;

imputeYear(2012, "Baltimore", "Cincinnati", 44, 13, 0, 1)

imputeYear(2012, "Baltimore", "New England", 31, 30, 0, 0)

imputeYear(2012, "Baltimore", "Cleveland", 23, 16, 0, 1)

imputeYear(2012, "Baltimore", "Dallas", 31, 29, 0, 1)

imputeYear(2012, "Baltimore", "Oakland", 55, 20, 0, 0)

imputeYear(2012, "Baltimore", "Pittsburgh", 20, 23, 0, 1)

imputeYear(2012, "Baltimore", "Denver", 17, 34, 0, 0)

imputeYear(2012, "Baltimore", "New York Giants", 33, 14, 0, 0)

imputeYear(2015, "Chicago", "Green Bay", 23, 31, 0, 1)

imputeYear(2015, "Chicago", "Arizona", 23, 48, 0, 0)

imputeYear(2015, "Chicago", "Oakland", 22, 20, 0, 0)

imputeYear(2015, "Chicago", "Minnesota", 20, 23, 0, 1)

imputeYear(2015, "Chicago", "Denver", 15, 17, 0, 0)

imputeYear(2015, "Chicago", "San Francisco", 20, 26, 1, 0)

imputeYear(2015, "Chicago", "Washington", 21, 24, 0, 0)

imputeYear(2015, "Chicago", "Detroit", 20, 24, 0, 1)

imputeYear(2011, "Miami", "New England", 24, 38, 0, 1)

imputeYear(2011, "Miami", "Houston", 13, 23, 0, 0)

imputeYear(2011, "Miami", "Denver", 15, 18, 1, 0)

imputeYear(2011, "Miami", "Washington", 2, 9, 0, 0)

imputeYear(2011, "Miami", "Buffalo", 35, 8, 0, 1)

imputeYear(2011, "Miami", "Oakland", 34, 14, 0, 0)

imputeYear(2011, "Miami", "Philadelphia", 10, 26, 0, 0)

imputeYear(2011, "Miami", "New York Jets", 19, 17, 0, 1)

(As a side note, it would have been ideal to create a linear regression model to predict the values requiring imputation rather than taking a blanket mean in this way, but that was beyond the scope of this project.)

As a final step, the data was sorted by home\_team, season, then week\_num and then an index was applied, to help facilitate visualizations in Tableau.

The second dataset to enter the picture was the NFL Player Arrests table from usatoday.com. Most of the data contained in this table was not used in this study, as the year of the arrest and the number of rows ascribed to the team name were all that was needed. Unfortunately the formatting of the team name was different from the previous dataset’s formatting, in this case using a three-letter abbreviation for each team. Wanting to keep things consistent between datasets, a small table was created with each relevant team and the two versions of the team names, as follows:

|  |  |
| --- | --- |
| Team\_Name | Team\_City |
| ARI | Arizona |
| BAL | Baltimore |
| CAR | Carolina |
| CHI | Chicago |
| CIN | Cincinnati |
| DAL | Dallas |
| DEN | Denver |
| GB | Green Bay |
| HOU | Houston |
| IND | Indianapolis |
| JAC | Jacksonville |
| KC | Kansas City |
| MIA | Miami |
| NE | New England |
| NYG | New York Giants |
| NYJ | New York Jets |
| OAK | Oakland |
| PHI | Philadelphia |
| PIT | Pittsburgh |
| SD | San Diego |
| SF | San Francisco |
| SEA | Seattle |
| TB | Tampa Bay |
| TEN | Tennessee |
| WAS | Washington |

This table was imported into Python and a few For loops were used to update the team names. (NFL Player Arrests was imported as data2 and the above table of name conversions was imported as data3.)

data2["Found"] = 0

for meh in range(0,len(data3)):

teamname = data3["Team\_Name"][meh]

for bleh in range(0,len(data2)):

if data2.iloc[bleh]['TEAM'] == teamname:

data2.iat[bleh,8] = 1

data2.iat[bleh,1] = data3.iloc[meh]["Team\_City"]

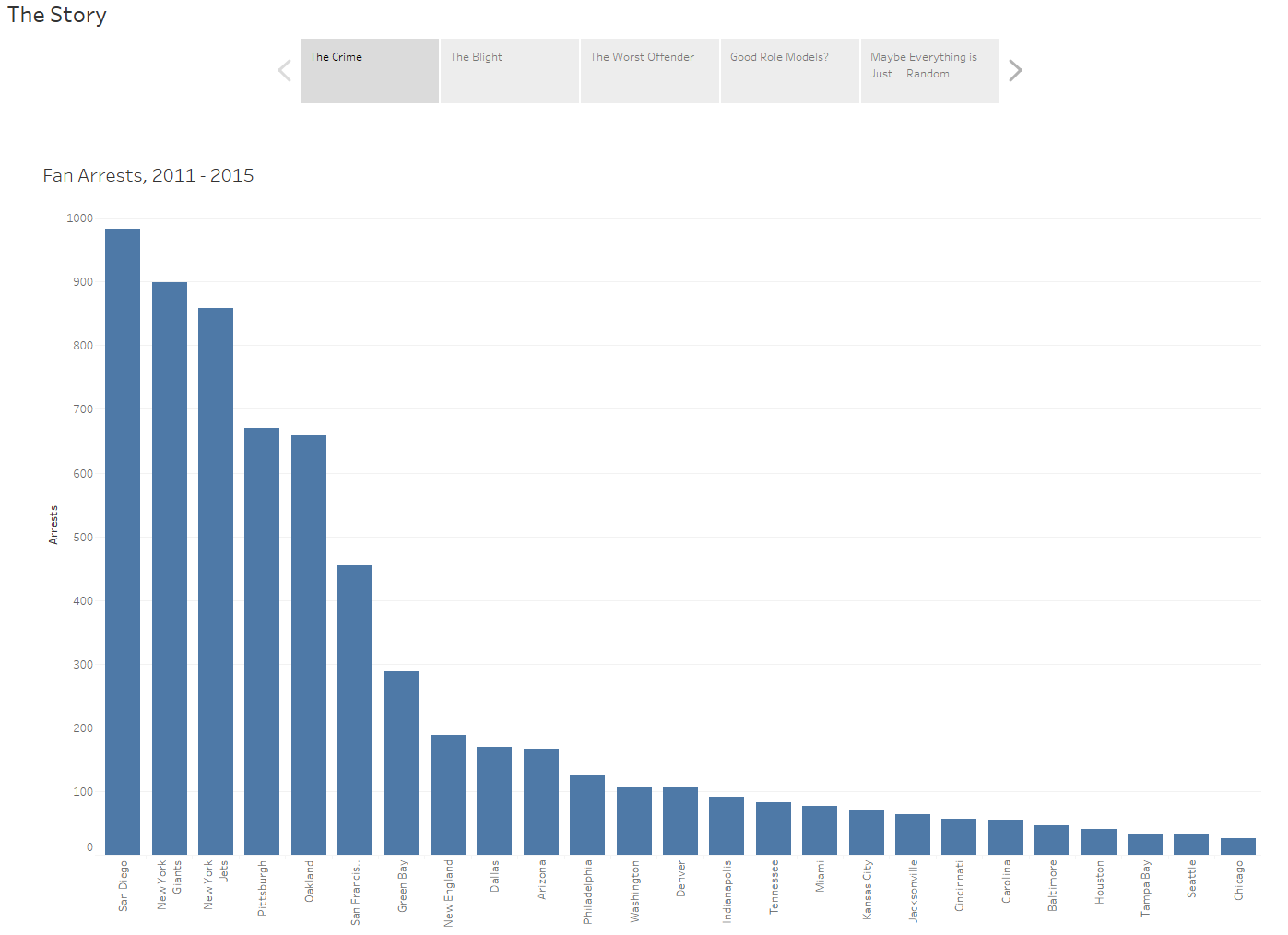
data2.drop(data2[data2['Found'] == 0].index, inplace = True)

data2.head(20)

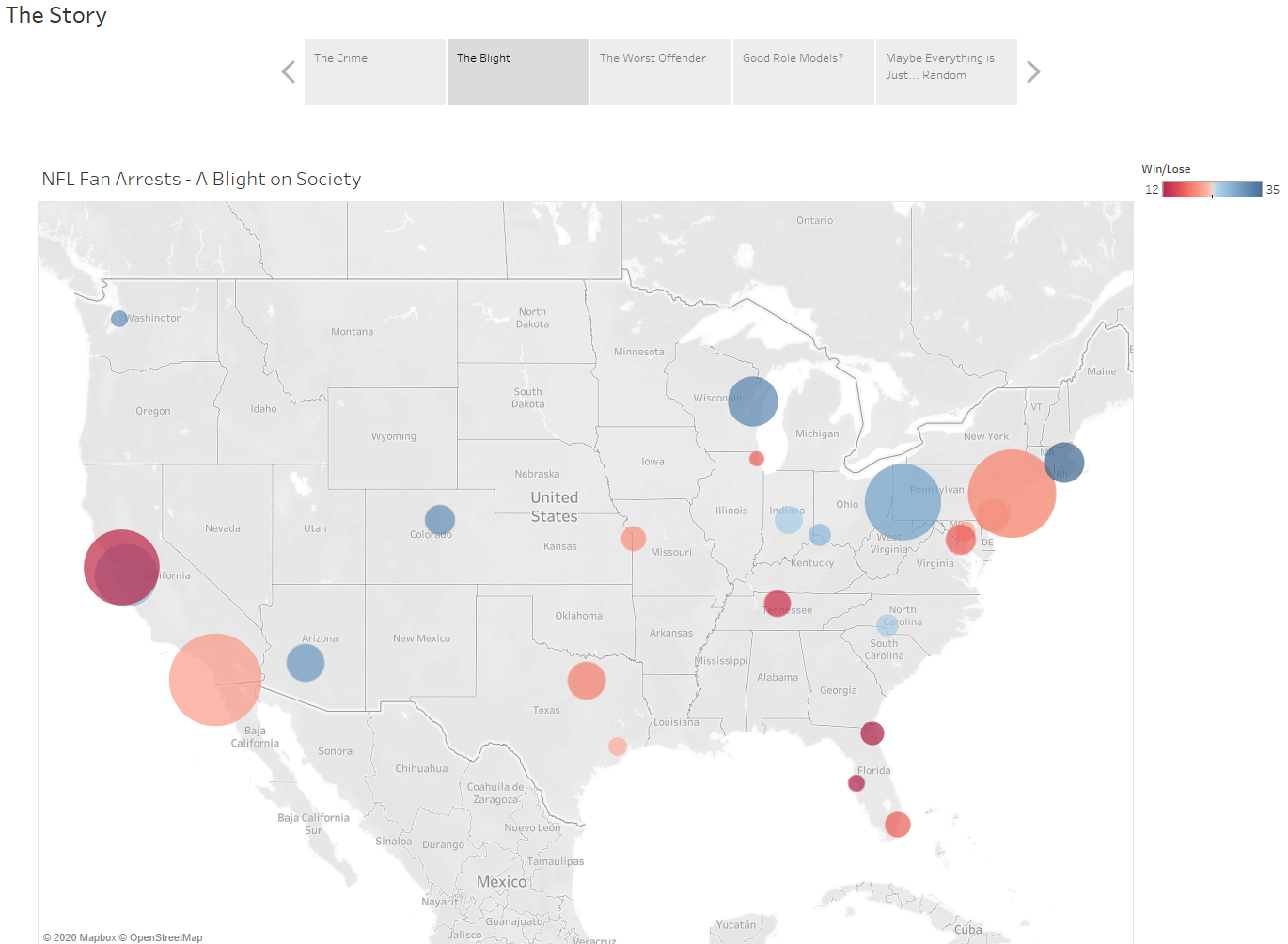
A small complication came up in that the Number of Player Arrests table contained all possible NFL teams instead of the ones present in the Fan Arrests table. To solve this, a new column titled “Found” (default value 0) was added to data2 and then, as the loops went about their business, each item they found and updated had their corresponding Found value set to the number 1. After the loop process, items with value 0 were dropped from the dataset.

With these changes, the two main datasets for the project were ready to be imported to Tableau. Alongside them, some manually created csv files showing crime rates by city and zip code by city were also imported to Tableau and linked via relationships to the primary dataset.

**Visualizations**

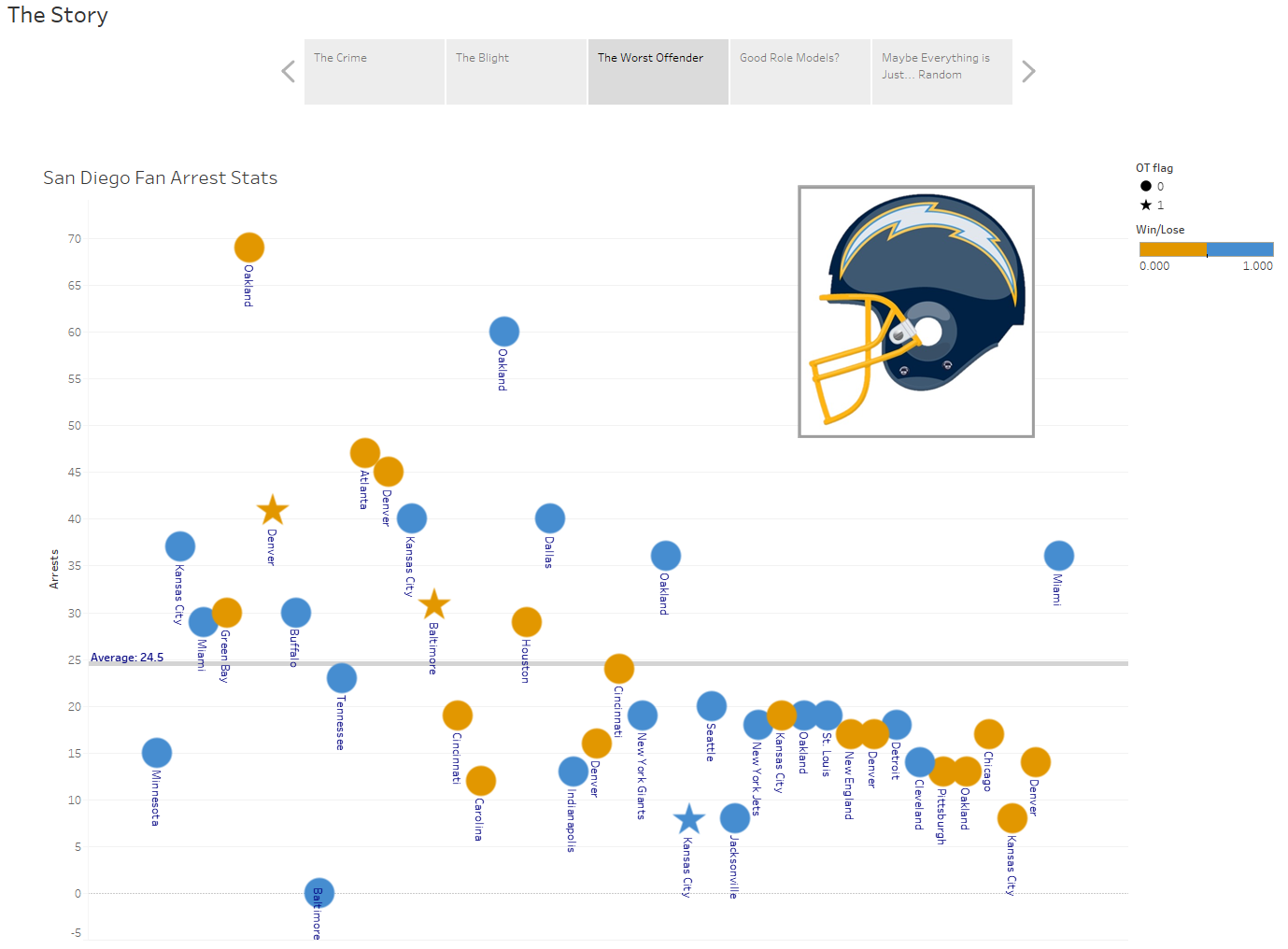


The first and most directly obvious visualization was a simple bar chart showing each represented team and their total number of fan arrests throughout the available timespan of 2011 – 2015. The distribution is not surprising in general ordering (it is safe to say many would have guessed the California teams, New York teams, and the Raiders would top the charts), but an interesting takeaway is the sheer magnititude of difference between the top teams and the bottom. The distribution almost appears to follow a Pareto distribution, which I, personally, would not have expected. But what factors might be contributing to these large differences?



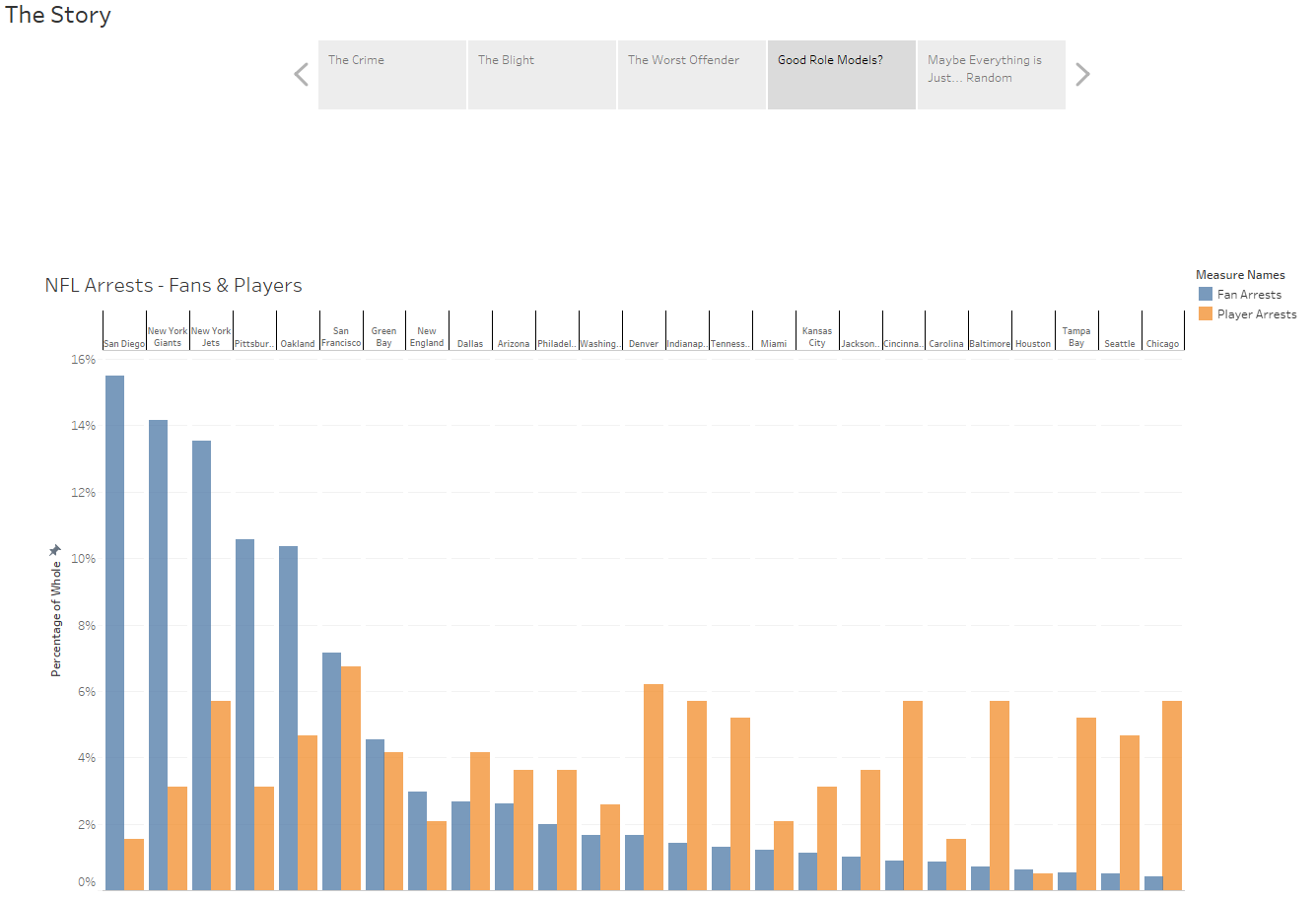
Here we can take a geographical look at the represented teams and their disproportionate rates of fan arrests, with an additional metric added in: wins. The more blue the circle the higher the total number of wins throughout the season, with the color red representing the opposite. It appears in general that most of the very large circles are red, but there are enough “exceptions to the rule” that no very reliable conclusions can be drawn quite yet. (“Exceptions” referring to the multitude of small red circles in the SouthEast, and the overlap of similarly sized blue and red circles in California, among others.)

Perhaps a more in-depth analysis of the country’s largest offender, the San Diego Chargers, would be illuminating?



Here a scatterplot of every normal game in the 2011 to 2015 seasons are shown, and color coded (according to the Charger’s official colors) to represent wins (blue) vs losses (gold), with the shape of each observation modified into a star for games that went into overtime (the logic being that overtime games may put the audience into a more tense mood, as well as increase the length of time over which arrests could be made).

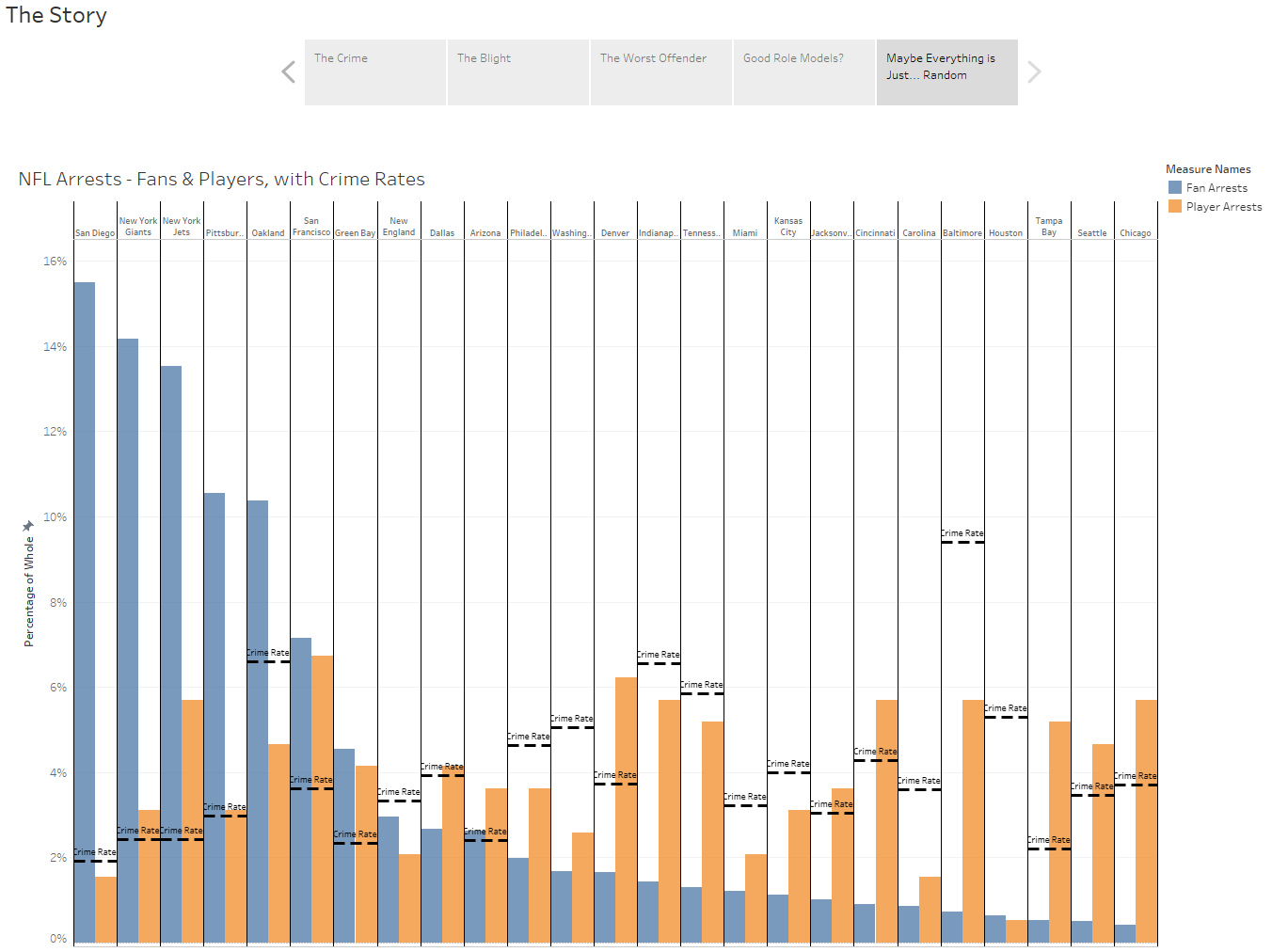
Unfortunately, as per the previous example, there does not seem to be any significant difference between winning and losing games. The only very obvious item that jumps out is the top two observations in terms of arrests made were both against the Oakland Raiders—indeed, three out of the four home games played against the Raiders have arrest results far above average. (The top example being a loss the Raiders—who in the audience could stand something like that?)



Now we come to the primary hypothesis of the study, a comparison between fan arrests and player arrests among NFL teams over the course of the same seasons. Surely there must be some relation? But again reality is not so clean as the idea, and we see not only very little agreement between the two sets of data, but if any pattern could be gleaned it would conclude an *inverse* relationship!

(Also please note that this bar chart has converted the fan arrests and player arrests numbers into percentages of the whole for the categories they are representing—this is to prevent fan arrests, which far outnumber player arrests, to dominate the axis.)

Hoping to salvage the study with some sort of concrete takeaway, one final question can be posed: if fan arrests and player arrests are mostly unrelated to each other, which would their corresponding city’s crime rate mirror (if any)?



Here there are three different metrics being reported for each team and the chart starts to become cluttered. However, with the crime rate overlay added on top, it is fairly easy to conclude that between fan arrests and player arrests, the crime rate more closely follows the pattern of player arrests.

**Conclusion**

Some previously held stereotypes of teams’ fandoms may have proven true through the sheer number of fan arrests reported for each team, but so far attempts to understand or predict those numbers have proven unfruitful. It may be beyond the scope of simple data analytics to understand the decades of tradition and backstory that no doubt influence fan behavior.

It may also be the case, considering the gaps in the data gathered by the Washington Post to begin with, that the data itself contained some large flaws. A significant possibility is that the cities which reported the statistics collect or classify “fan arrests” in different ways, leading to some significant over/under reporting for some teams. If the NFL or another organization wished to understand the issue in more detail, the first step no doubt would be to standardize the gathering of such data.

**Appendix #1 – Data Sources**

Fan arrests: <https://github.com/washingtonpost/data-nfl-arrests/blob/master/nfl_arrests_2011-2015.csv>

Player arrests: <https://www.usatoday.com/sports/nfl/arrests/>

City crime rates: <https://www.neighborhoodscout.com/>

ZIP Codes of NFL Stadiums: <https://www.stadiumsofprofootball.com/>

Missing Data of NFL games: <https://en.wikipedia.org/wiki/National_Football_League>

**Appendix #2 – Python Code**

NFL Arrests Visualization Project

Allen Butt

Dataset 1

import pandas as pd

import re

import numpy as np

data = pd.read\_csv("nfl\_arrests\_2011-2015.csv", encoding = 'unicode\_escape')

#Fix missing data in OT\_flag and turn it into a numeric variable

data.fillna({'OT\_flag':0}, inplace=True)

data['OT\_flag'] = data['OT\_flag'].replace(['OT'],1)

data["OT\_flag"]=pd.to\_numeric(data["OT\_flag"])

#Update "division\_game" into numeric as well

data['division\_game'] = data['division\_game'].replace(['n'],0)

data['division\_game'] = data['division\_game'].replace(['y'],1)

data["division\_game"]=pd.to\_numeric(data["division\_game"])

data.head()

#Some observations have missing data--they should be dropped from the dataframe.

data = data[data['arrests'].notna()]

#Some games were played in London and so have missing data.

#We can impute the missing values taking the mean of the arrests of the same year for that team.

#Create a function to help with this process

def imputeLondon(year, home, away, homescore, awayscore, OT, division):

new = data[(data['home\_team'] == home) & (data['season'] == year )]["arrests"].mean()

data.loc[len(data)] = np.array([year,0,0,0,home,away,homescore,awayscore,OT,new,division])

return;

#Use the function to fill in the missing data with imputed values.

imputeLondon(2013, "Arizona", "Houston", 30, 9, 0, 0)

imputeLondon(2013, "Jacksonville", "San Francisco", 10, 42, 0, 0)

imputeLondon(2014, "Jacksonville", "Dallas", 17, 32, 0, 0)

imputeLondon(2015, "Jacksonville", "Buffalo", 34, 31, 0, 0)

imputeLondon(2015, "Kansas City", "Detroit", 45, 10, 0, 0)

imputeLondon(2015, "Miami", "New York Jets", 14, 27, 0, 1)

imputeLondon(2014, "Oakland", "Miami", 14, 38, 0, 0)

imputeLondon(2014, "Oakland", "Denver", 17, 41, 0, 1)

imputeLondon(2014, "Oakland", "Kansas City", 24, 20, 0, 0)

imputeLondon(2011, "Tampa Bay", "Chicago", 18, 24, 0, 0)

#Three teams had a missing year of data--we can impute this data by taking the mean of the existing years.

def imputeYear(year, home, away, homescore, awayscore, OT, division):

new = data[(data['home\_team'] == home)]["arrests"].mean()

data.loc[len(data)] = np.array([year,0,0,0,home,away,homescore,awayscore,OT,pd.to\_numeric(new),division])

data["arrests"]=pd.to\_numeric(data["arrests"]) #kept getting type errors without brute-forcing it

return;

imputeYear(2012, "Baltimore", "Cincinnati", 44, 13, 0, 1)

imputeYear(2012, "Baltimore", "New England", 31, 30, 0, 0)

imputeYear(2012, "Baltimore", "Cleveland", 23, 16, 0, 1)

imputeYear(2012, "Baltimore", "Dallas", 31, 29, 0, 1)

imputeYear(2012, "Baltimore", "Oakland", 55, 20, 0, 0)

imputeYear(2012, "Baltimore", "Pittsburgh", 20, 23, 0, 1)

imputeYear(2012, "Baltimore", "Denver", 17, 34, 0, 0)

imputeYear(2012, "Baltimore", "New York Giants", 33, 14, 0, 0)

imputeYear(2015, "Chicago", "Green Bay", 23, 31, 0, 1)

imputeYear(2015, "Chicago", "Arizona", 23, 48, 0, 0)

imputeYear(2015, "Chicago", "Oakland", 22, 20, 0, 0)

imputeYear(2015, "Chicago", "Minnesota", 20, 23, 0, 1)

imputeYear(2015, "Chicago", "Denver", 15, 17, 0, 0)

imputeYear(2015, "Chicago", "San Francisco", 20, 26, 1, 0)

imputeYear(2015, "Chicago", "Washington", 21, 24, 0, 0)

imputeYear(2015, "Chicago", "Detroit", 20, 24, 0, 1)

imputeYear(2011, "Miami", "New England", 24, 38, 0, 1)

imputeYear(2011, "Miami", "Houston", 13, 23, 0, 0)

imputeYear(2011, "Miami", "Denver", 15, 18, 1, 0)

imputeYear(2011, "Miami", "Washington", 2, 9, 0, 0)

imputeYear(2011, "Miami", "Buffalo", 35, 8, 0, 1)

imputeYear(2011, "Miami", "Oakland", 34, 14, 0, 0)

imputeYear(2011, "Miami", "Philadelphia", 10, 26, 0, 0)

imputeYear(2011, "Miami", "New York Jets", 19, 17, 0, 1)

#For one of the visualizations, we need to sort the dataset according to team, then year, then week--then an index will need

#to be added to keep things properly sorted.

data = data.sort\_values(by = ["home\_team","season", "week\_num"])

data["Index\_num"] = 0

for snuh in range(0,len(data)):

data.iat[snuh,11] = snuh

#Export Dataframe to CSV

data.to\_csv(r'nfl\_arrests.csv', index = False)

Dataset 2

#New Dataset, NFL Player Arrests

data2 = pd.read\_csv("nfl\_player\_arrests.csv", encoding = 'unicode\_escape')

#Check out the data, look for Missing Data

data2.head()

#We need to standardize Team Names--importing a new csv file with two columns to help ease the transition

data3 = pd.read\_csv("nfl\_names\_conversion.csv", encoding = 'unicode\_escape')

data3.head()

#Loop through each row in this small dataset, and change obervations in data2 that match "Team Name" to "Team City".

#Also add a new column to data3 that selects 1 for items that were matched. This will allow us to delete all observations

#with teams outside of our dataset easily.

data2["Found"] = 0

for meh in range(0,len(data3)):

teamname = data3["Team\_Name"][meh]

for bleh in range(0,len(data2)):

if data2.iloc[bleh]['TEAM'] == teamname:

data2.iat[bleh,8] = 1

data2.iat[bleh,1] = data3.iloc[meh]["Team\_City"]

data2.drop(data2[data2['Found'] == 0].index, inplace = True)

data2.head(20)

#Export Dataframe to CSV

data2.to\_csv(r'nfl\_players.csv', index = False)