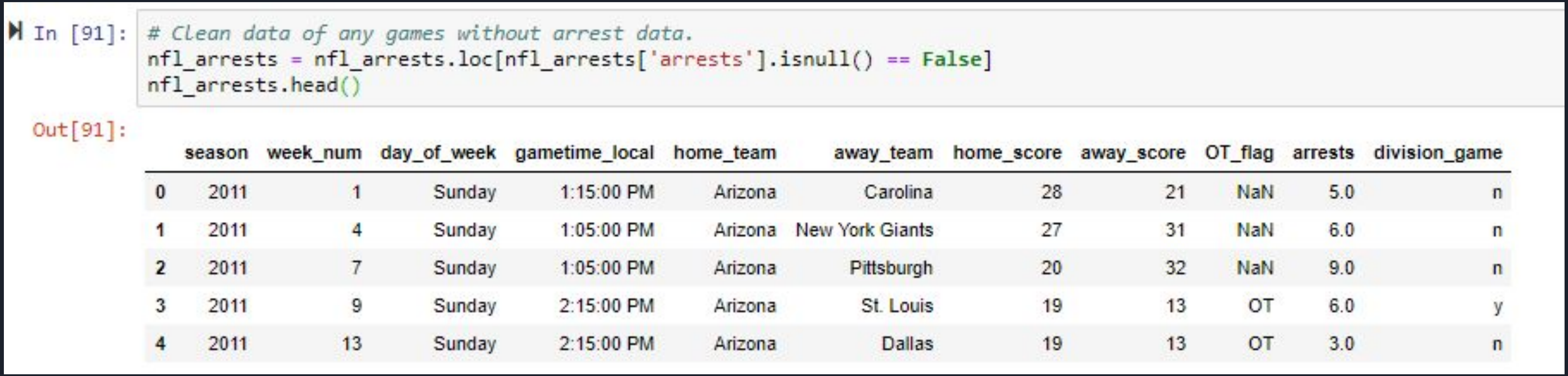
NFL Stadiums Arresting Fans

Eric Tuscanes, Isaac Muck, Mike Toriello
Group 7

About this dataset

- This data has the recorded number of fans arrested for every NFL stadium from 2011-2015.
  - 7 teams did not report any arrests and therefore will not be considered in this analysis.
  - (St. Louis, Cleveland, Buffalo, Atlanta, New Orleans, Minnesota, Detroit*)
    - * Detroit games were in the set, but no arrest data was provided. Deleted those rows.



```
In [91]: # Clean data of any games without arrest data.
         nfl_arrests = nfl_arrests.loc[nfl_arrests['arrests'].isnull() == False]
         nfl_arrests.head()
```

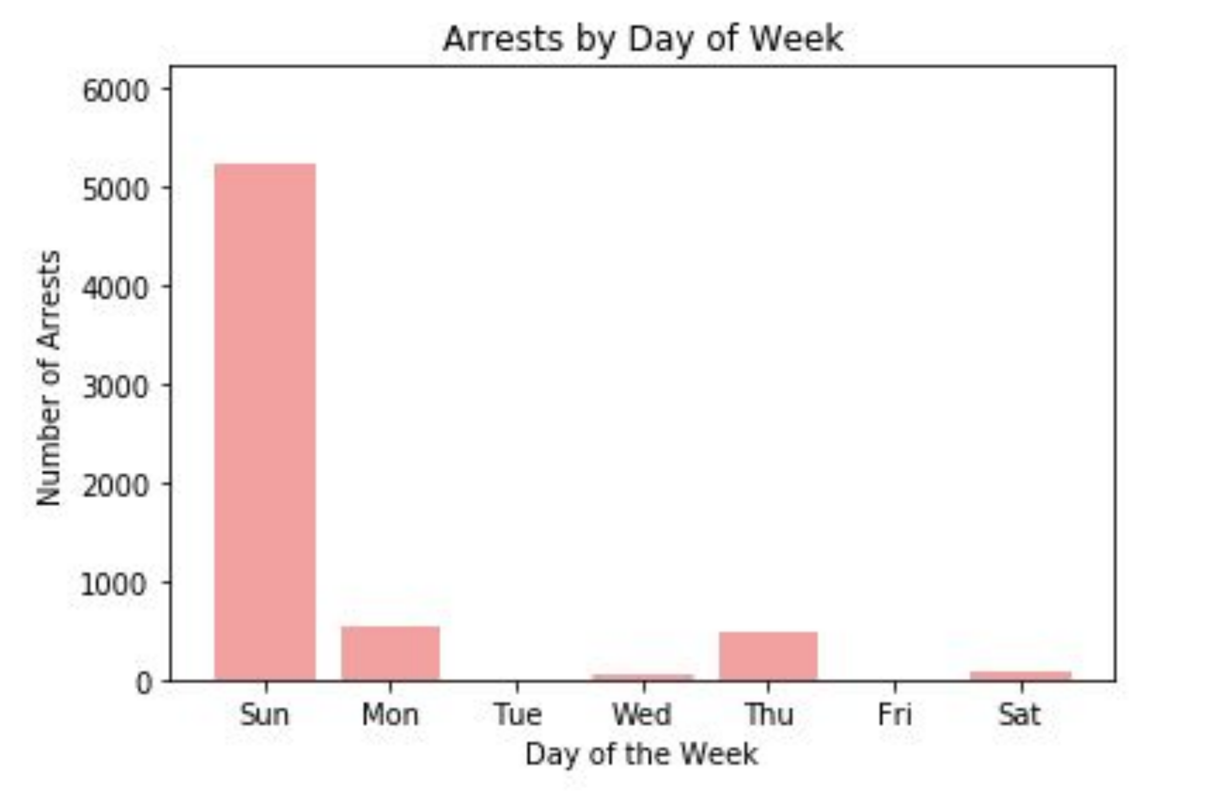
Out[91]:

	season	week_num	day_of_week	gametime_local	home_team	away_team	home_score	away_score	OT_flag	arrests	division_game
0	2011	1	Sunday	1:15:00 PM	Arizona	Carolina	28	21	NaN	5.0	n
1	2011	4	Sunday	1:05:00 PM	Arizona	New York Giants	27	31	NaN	6.0	n
2	2011	7	Sunday	1:05:00 PM	Arizona	Pittsburgh	20	32	NaN	9.0	n
3	2011	9	Sunday	2:15:00 PM	Arizona	St. Louis	19	13	OT	6.0	y
4	2011	13	Sunday	2:15:00 PM	Arizona	Dallas	19	13	OT	3.0	n

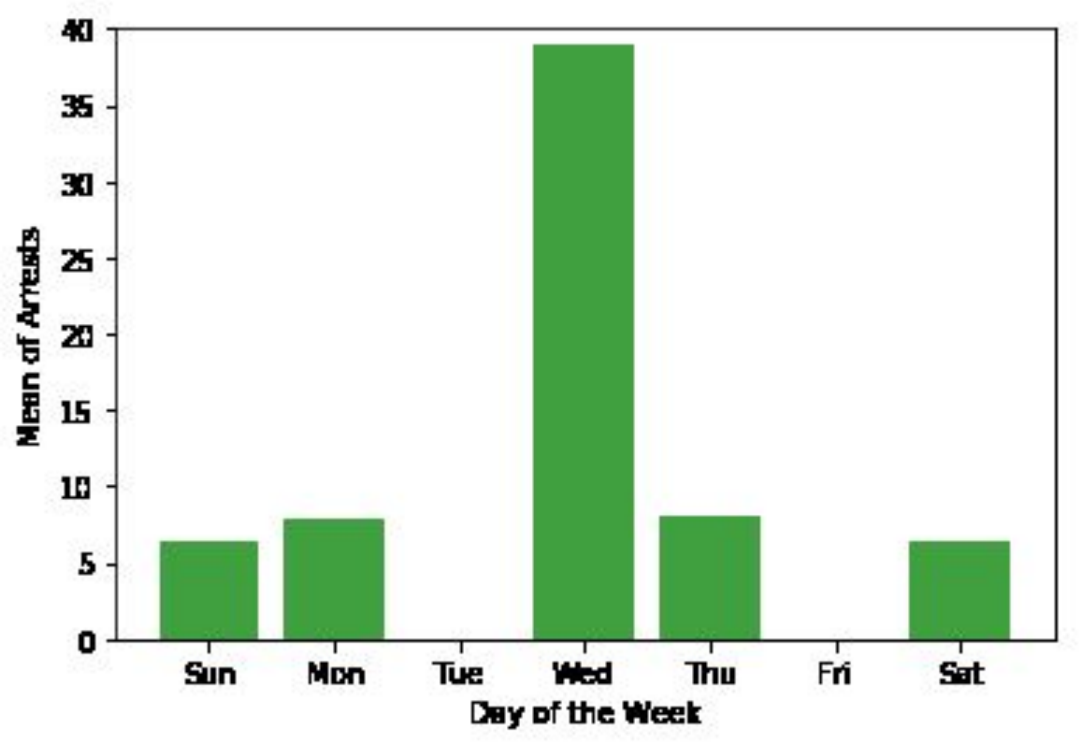
Day of the Week vs. Arrests | Score vs. Arrests

- Does the day of the week have any correlation with the number of arrests per stadium?
  - Sunday games are the most common, but there are some Monday and Thursday games.
- If the game is a close score, are there more likely to be arrests at the stadium?
  - To do this, we took the absolute value of the difference in the scores, and binned them according to 10 point intervals.

Day of the Week vs Sum of Arrests



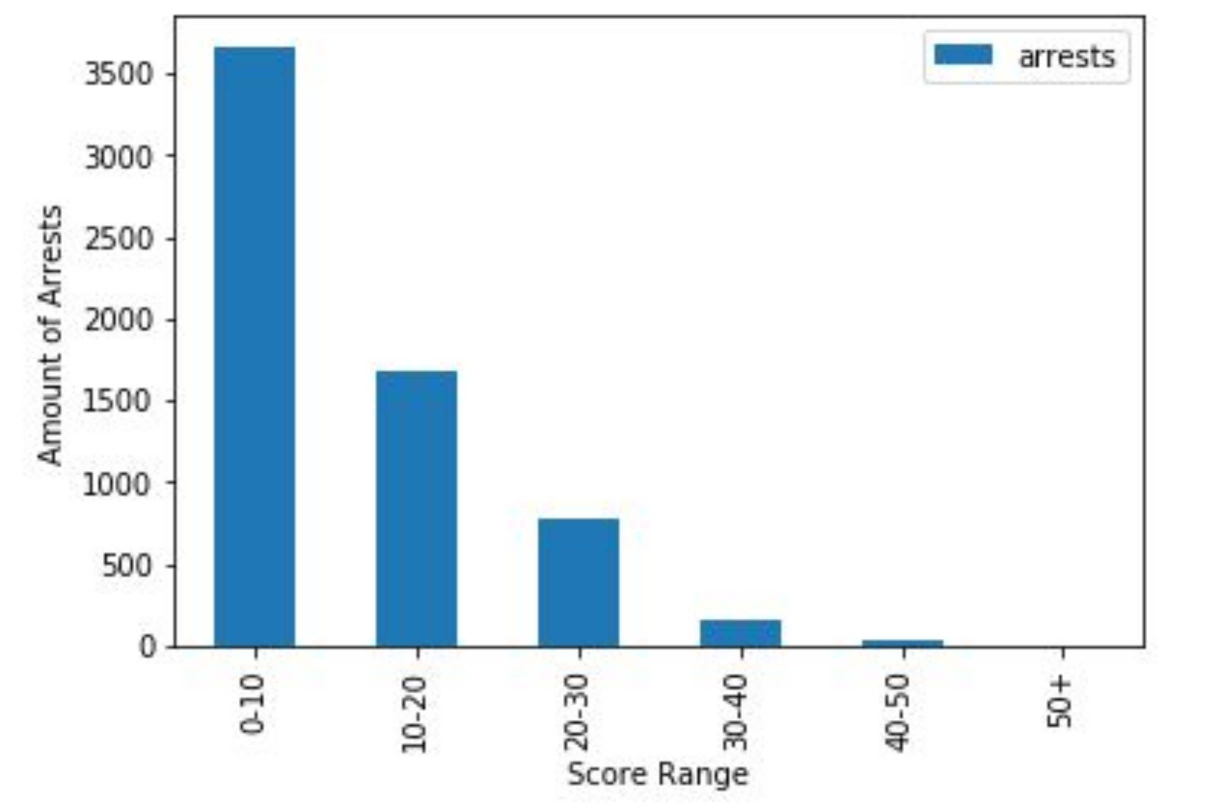
Day of the Week vs. Mean of Arrests



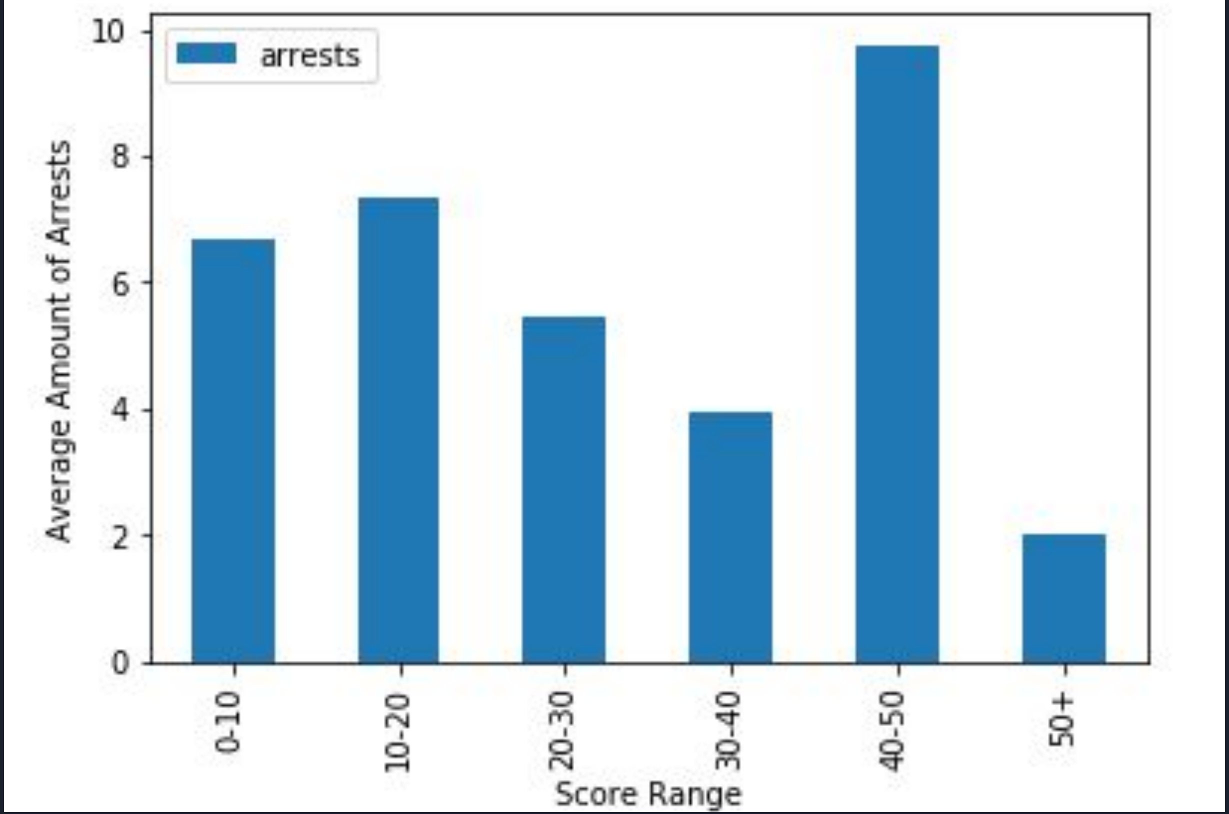
Weekday and Arrest Observations

- Sunday obviously has the most in terms of value count.
  - Sunday also has the most when arrests are counted--not taking the final sum into consideration.
- The average for Wednesday is highest because in 2012, the first game had a total of 39 arrests. There was only one Wednesday game in the dataset, giving it the same average and mean.
  - Therefore, the best conclusion we can make is that there more arrests on Sundays because there are more games on Sundays.

Score Range vs Amount of Arrests



Score Range vs Avg Number of Arrests



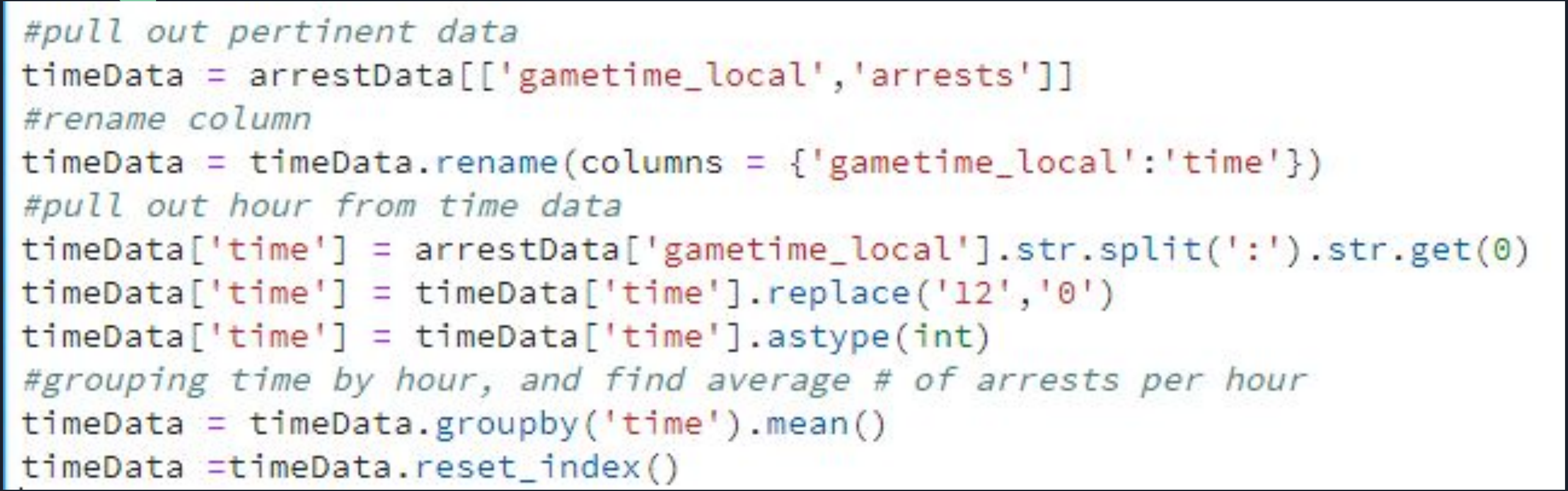
Score range and arrest observations

- The data descends based on the low amount of score.
  - When there is a difference of ten points or less (3,666), there are 1,987 less arrests than there would be in a range of 10-20 (1,679).
  - The maximum amount of arrests per score range is 69, which also falls in the 0-10 range.
- The average is higher for the 40-50 range because 40-50 has a smaller amount of arrests overall (39) and a value count of four overall.
- The best conclusion we can draw is that a smaller difference in score is more likely to lead to a higher arrest rate.

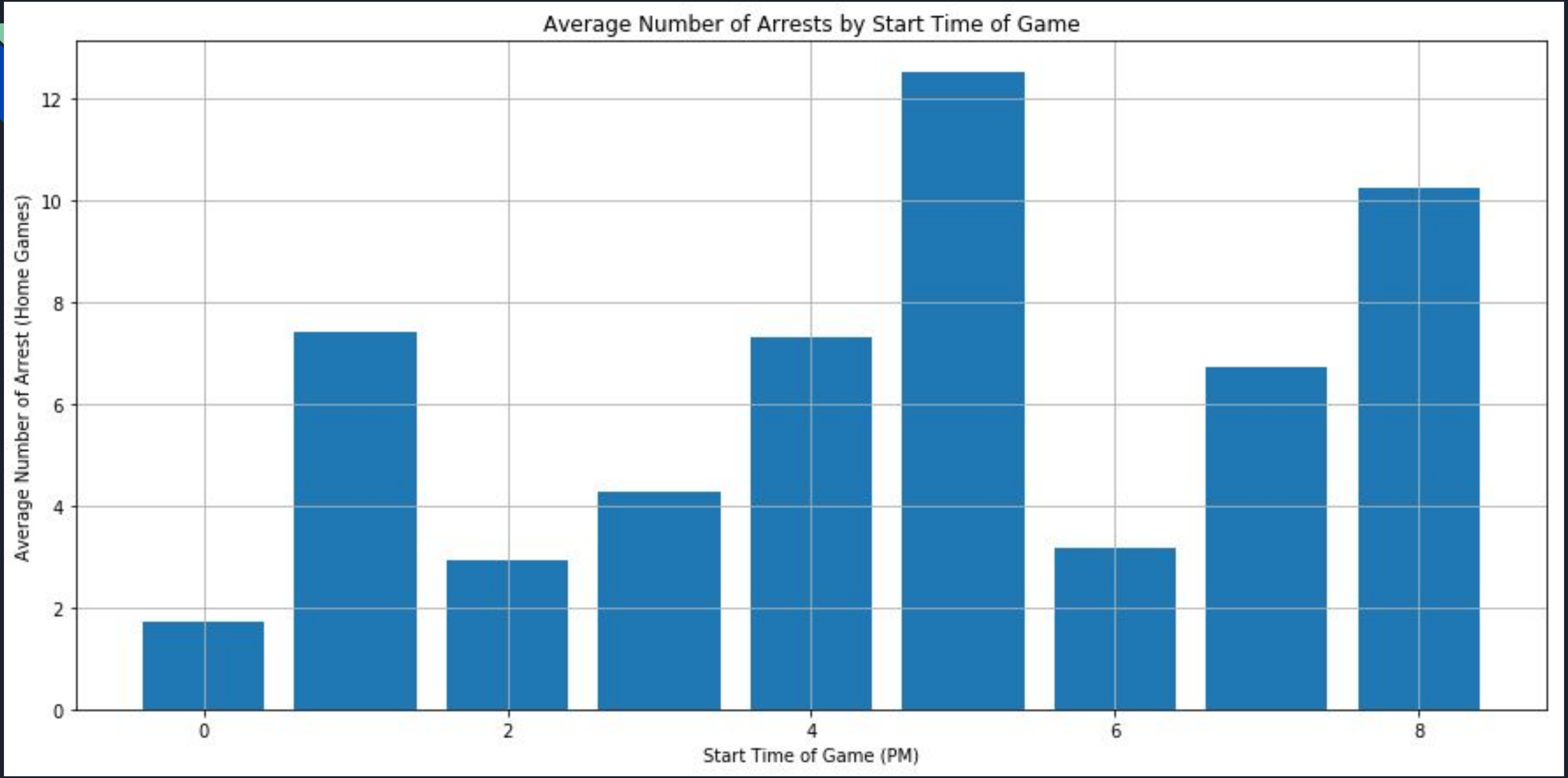
Start Time of Games vs Avg Number of Arrests

- Does the start time of a game correlate to the number of arrests?

- Time Format: xx:xx:xx PM
  - Stripped the hour from the data field
- Binned the start time by hours, and calculated the average number of arrests by start time



```
#pull out pertinent data
timeData = arrestData[['gametime_local','arrests']]
#rename column
timeData = timeData.rename(columns = {'gametime_local':'time'})
#pull out hour from time data
timeData['time'] = arrestData['gametime_local'].str.split(':').str.get(0)
timeData['time'] = timeData['time'].replace('12','0')
timeData['time'] = timeData['time'].astype(int)
#grouping time by hour, and find average # of arrests per hour
timeData = timeData.groupby('time').mean()
timeData =timeData.reset_index()
```


Average Number of Arrests by Start Time of Game
Average Number of Arrest (Home Games)
0
2
4
6
8
10
12
0
2
4
6
8
Start Time of Game (PM)

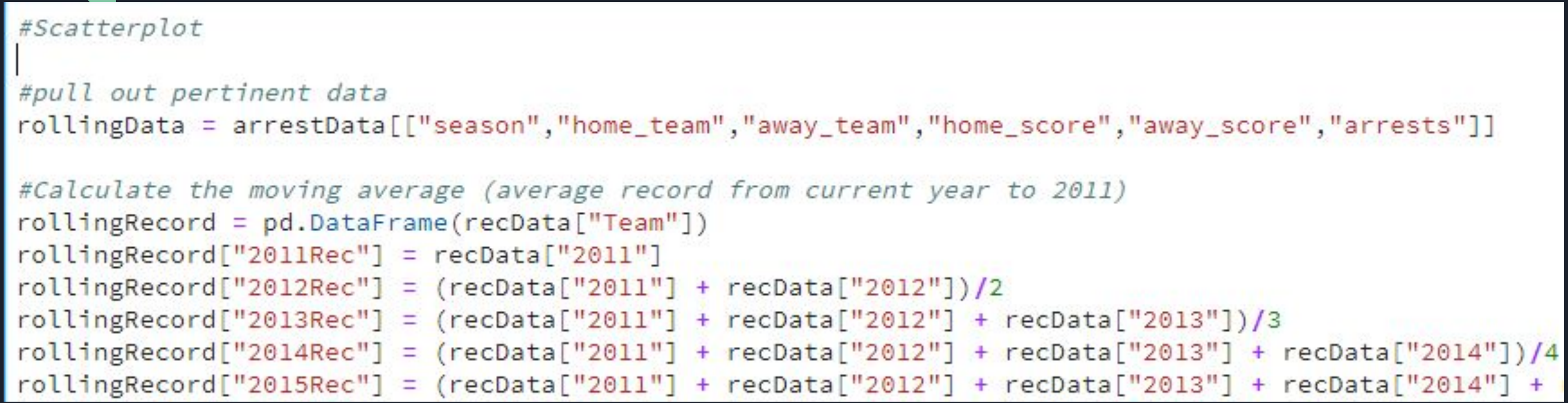
Game Start Time Observations

- Unfortunately, there doesn’t seem to be a correlation between the start time of a game and the # of arrests.

Team Record vs Average Number of Arrests

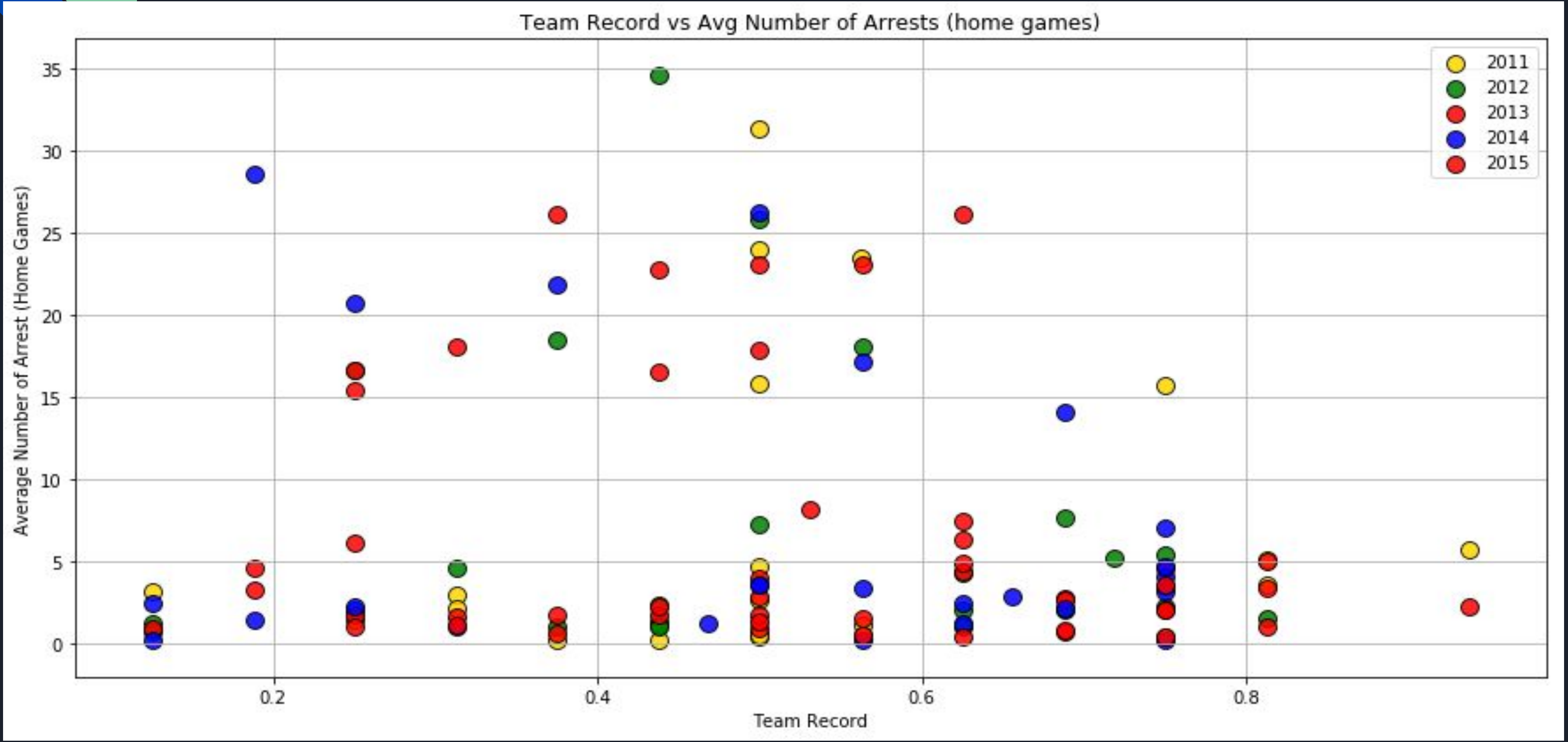
- Merged two data sets
  - NFL Arrests per Game
  - NFL Team Records
    - Curated our own Data Set (NFL makes you pay for most data)
- Looked at this in two ways:
  - Team Record vs Average Arrests
  - Moving Team Record vs Average Arrests
    - Moving Team Record is the average team record up to that year
      - Idea is to see if team’s previous year’s performances play a part in the arrests that year
- Does the team’s record come into play in regards to the # of arrests?
- Does a team’s previous performances (years prior) have any effect on the # of arrests?

Rolling Team Record Calculation

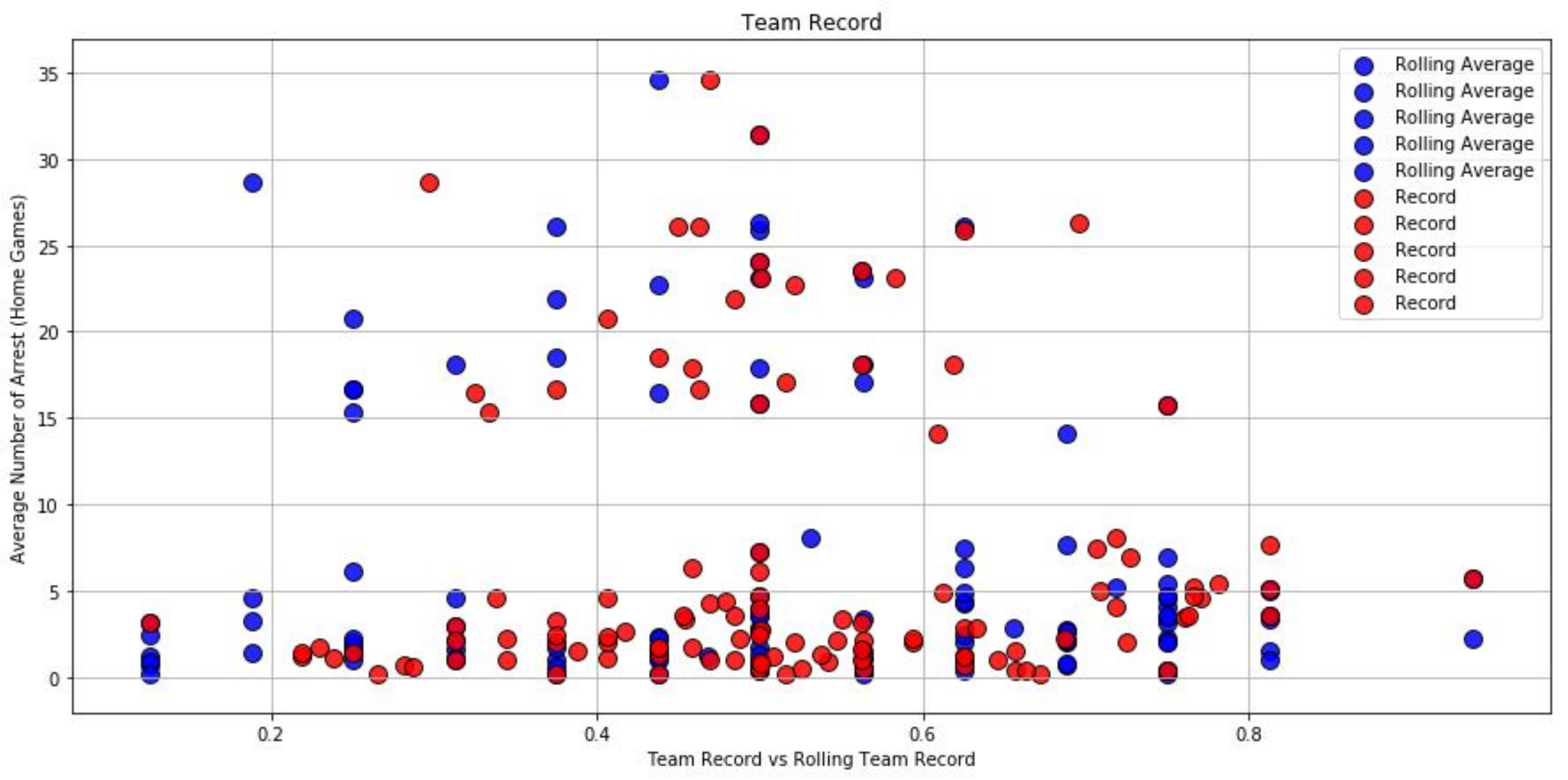


```
#Scatterplot

#pull out pertinent data
rollingData = arrestData[["season","home_team","away_team","home_score","away_score","arrests"]]

#Calculate the moving average (average record from current year to 2011)
rollingRecord = pd.DataFrame(recData["Team"])
rollingRecord["2011Rec"] = recData["2011"]
rollingRecord["2012Rec"] = (recData["2011"] + recData["2012"])/2
rollingRecord["2013Rec"] = (recData["2011"] + recData["2012"] + recData["2013"])/3
rollingRecord["2014Rec"] = (recData["2011"] + recData["2012"] + recData["2013"] + recData["2014"])/4
rollingRecord["2015Rec"] = (recData["2011"] + recData["2012"] + recData["2013"] + recData["2014"] +
```

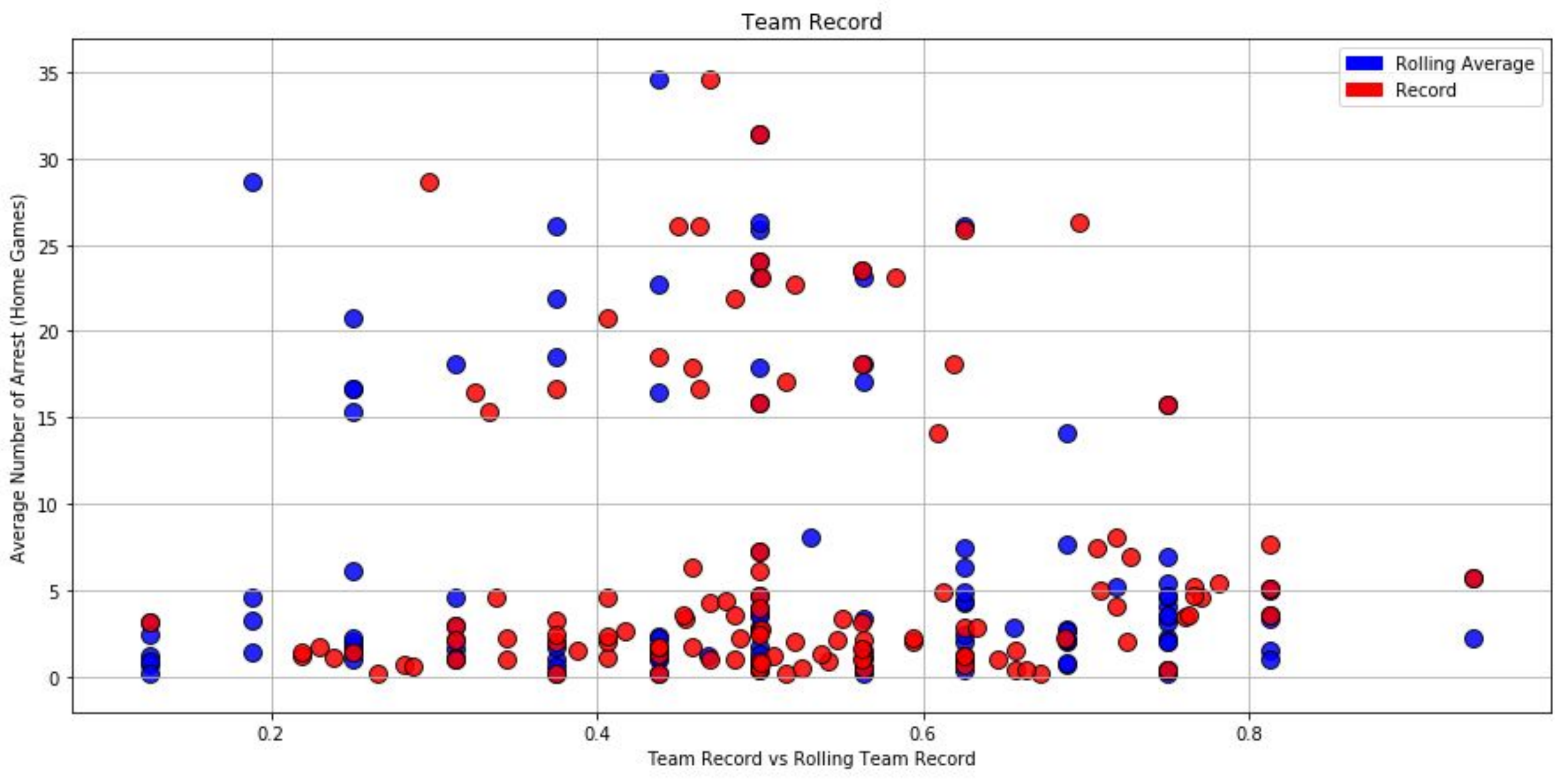
Team Record vs Average Number of Arrests



Rolling Team Record vs Average Number of



Rolling Team Record vs Team Record



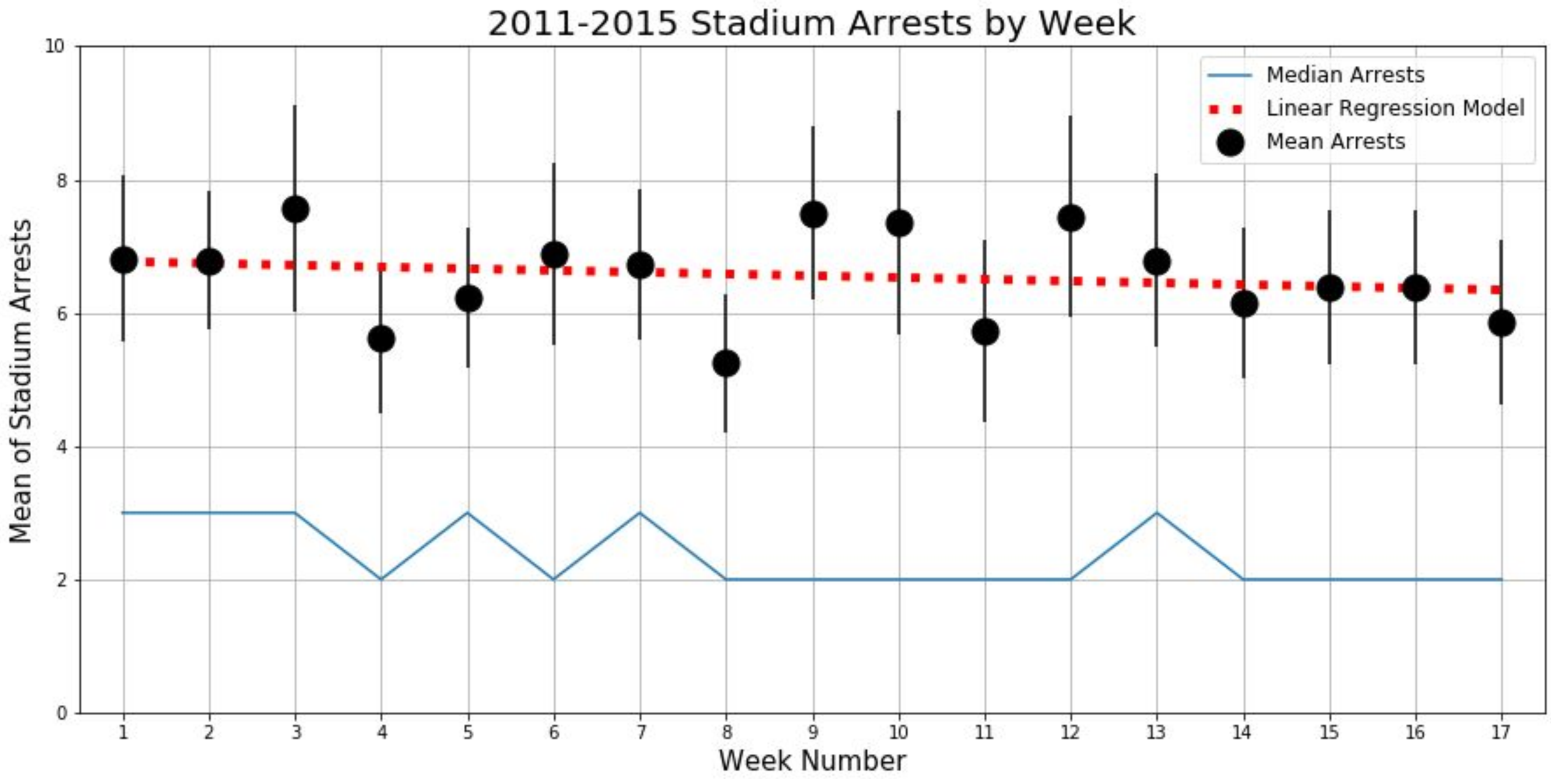
Team Record Observations

- Both team average and the moving (rolling) record average exhibit very similar behaviors
- Both show spikes around the 0.5 mark for the most number of arrest occurrences
- Similar behavior could be due to our data set.
    - Only 5 years of data so the team’s record may not be changing much year to year, causing the team’s rolling record average for that year to not differ greatly from the team’s record that year
    - A set with a larger # of years may give a better picture of the behavior

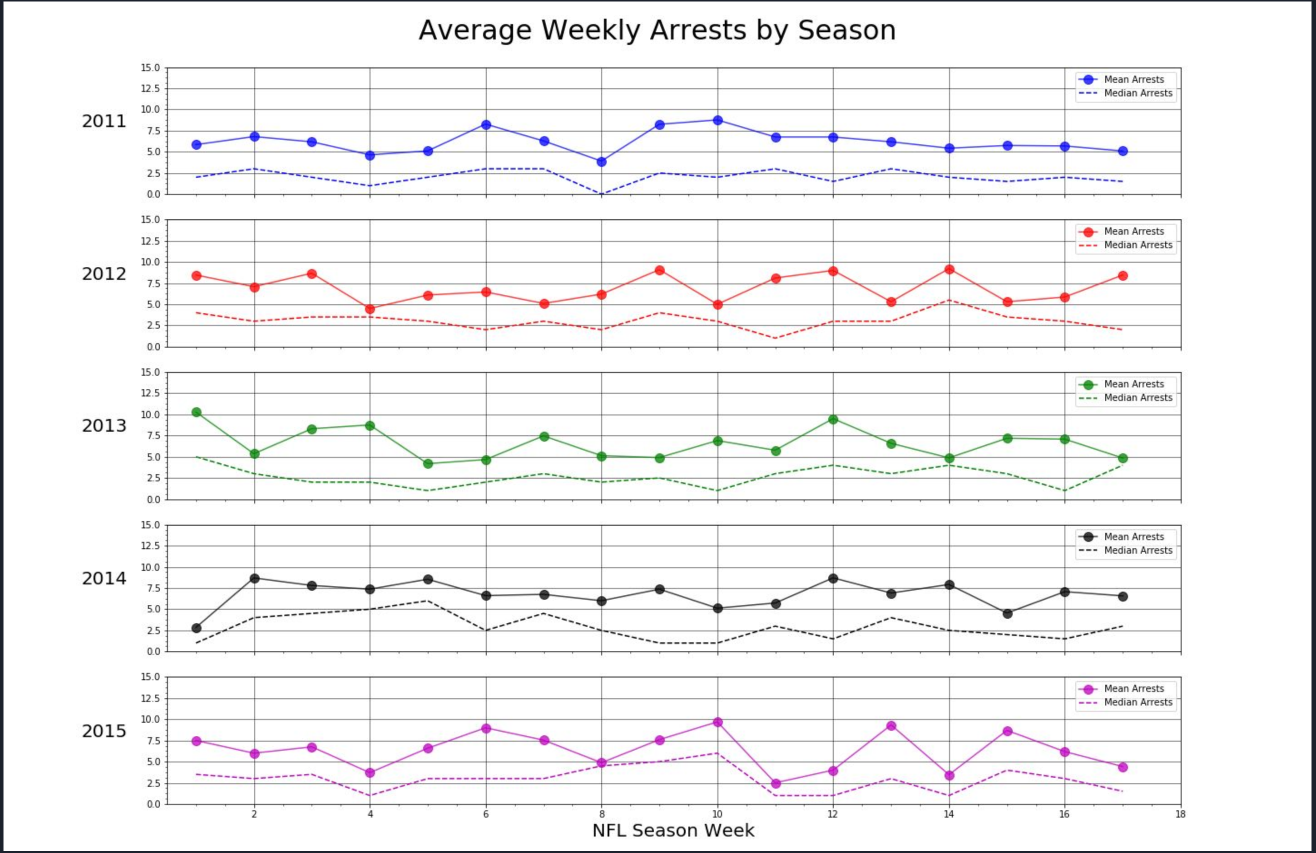
Week in the Season vs Average Number of Arrests

- Amongst all seasons, is there a correlation in number of arrests around the league, and the week a game is played on?
    - FIrst analyzed the combination of all seasons to see if there were weekly patterns.
    - Next checked each of the 5 seasons individually to see if there are any seasons are outliers.
- Potential reasons we could see results are:
    - Does the weather play a part in the number of arrests at a stadium?
        - Later weeks will be much colder compared to earlier weeks
    - Do more arrests occur as the season comes closer towards playoffs?

Week Number vs Average Arrests Among All

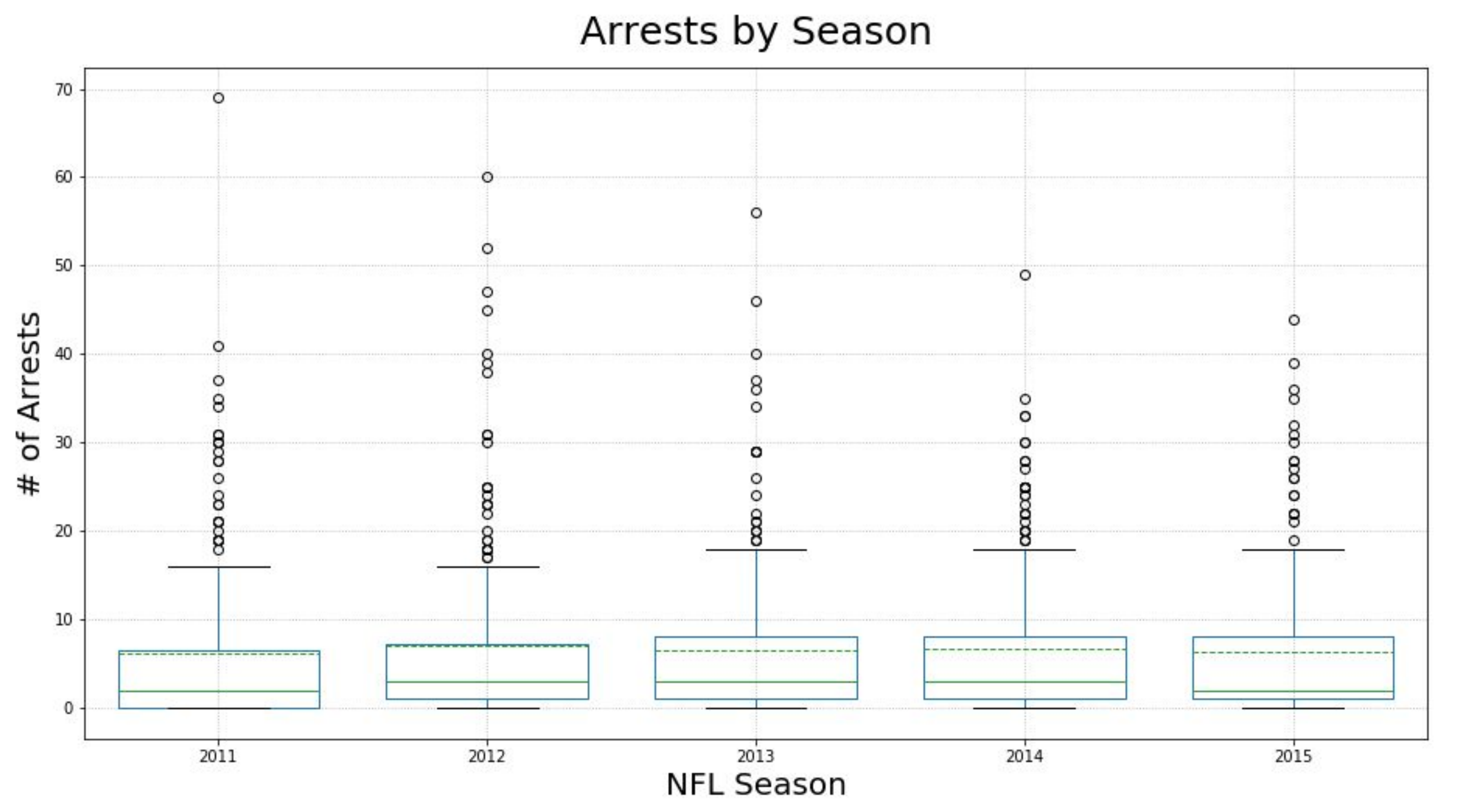


Week Number vs Average Arrests By Season



Weekly Observations:

- Did not achieve the correlation we were hoping for… However:
  - Weeks 4, 8, and 11 tend to have much lower average arrests.
    - This can be observed as well on the seasonal plots.
  - The linear regression plot looks almost flat, meaning the number of arrests does not seem to trend in any direction as the season progresses.
- The medians are <u>much lower</u> than the means.
  - The mean for the weekly number of arrests are often in the 3rd quartile of the samples.
  - This means a majority of the data rests below the mean
  - There must be some major outliers above the mean to balance the calculation.


Arrests by Season
# of Arrests
NFL Season
70
60
50
40
30
20
10
0
2011
2012
2013
2014
2015

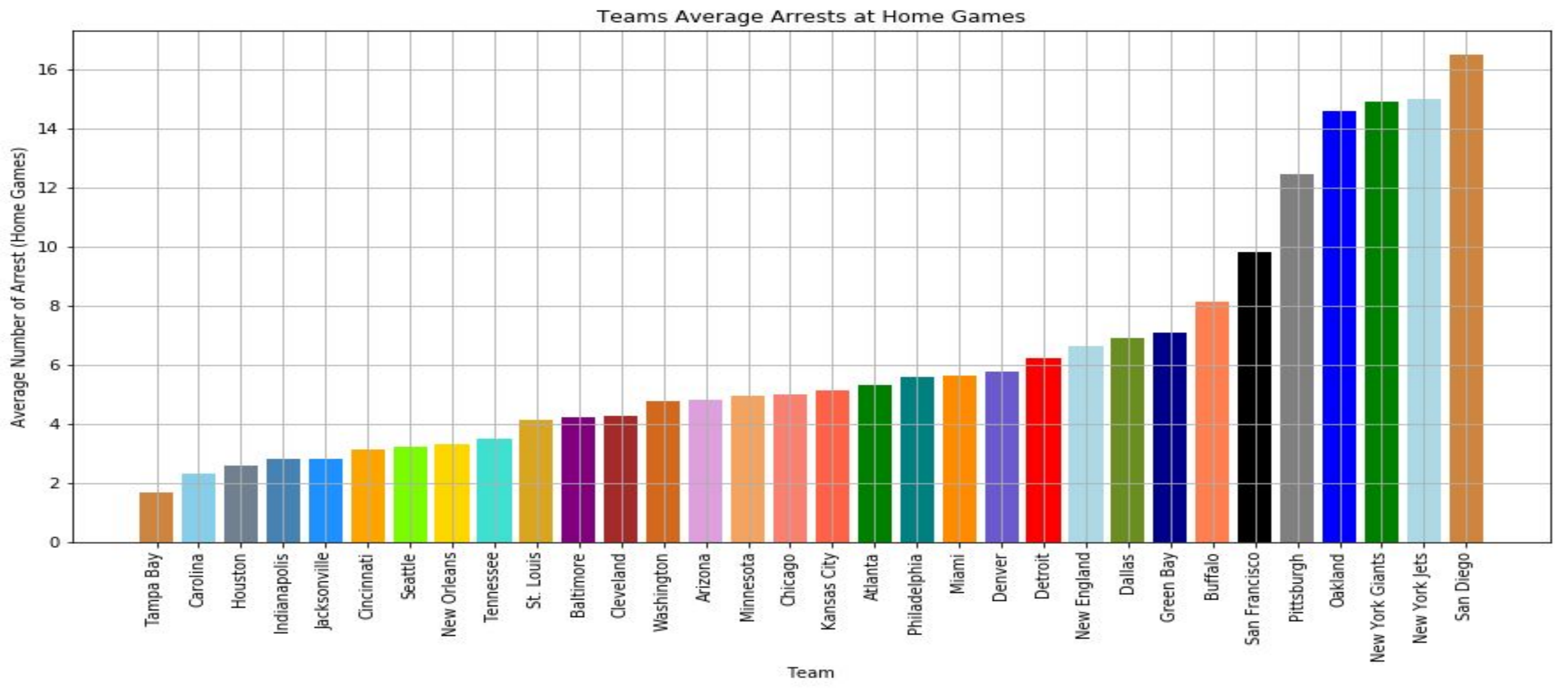
Outlier definitions for NFL Arrests

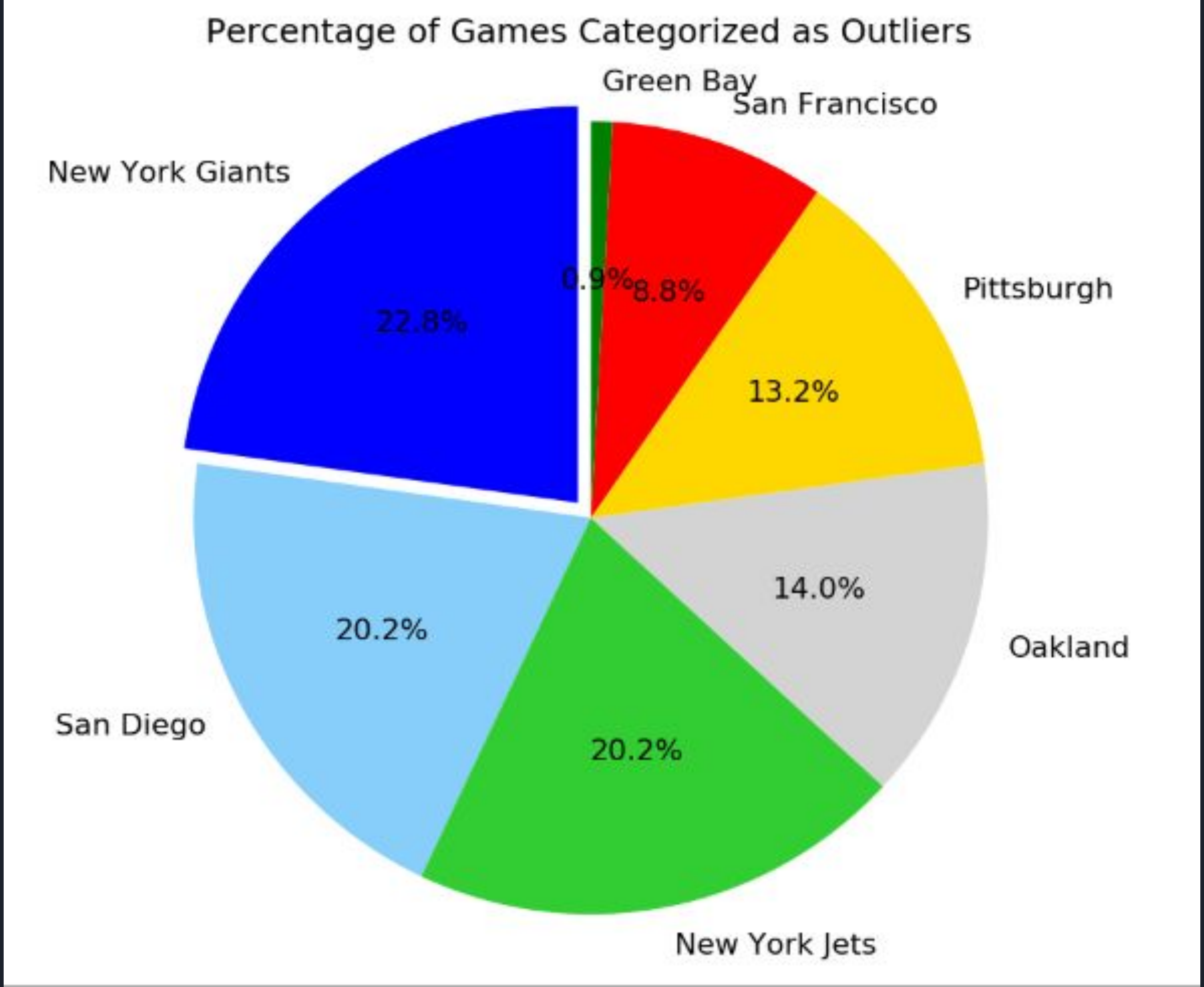
- How are the outliers defined?
  - IQR = Q3 - Q1
  - Outlier > (1.5 * IQR) + Q3
  - Major Outlier > (3 * IQR) + Q3
- What are the statistics in this case?
  - Median = **3** arrests
  - IQR = **7** arrests
  - **Outlier** > **18** arrests : Consists of **11.8%** of the data
  - **Major Outlier** > **29** arrests: Consists of **4.04%** of the data
- Which begs the question.... Who are these teams with angry and violent fanbases?
  - The first teams/cities that come to mind are:
    - Oakland, Philadelphia. (These fanbases are notorious for aggressive behaviour)

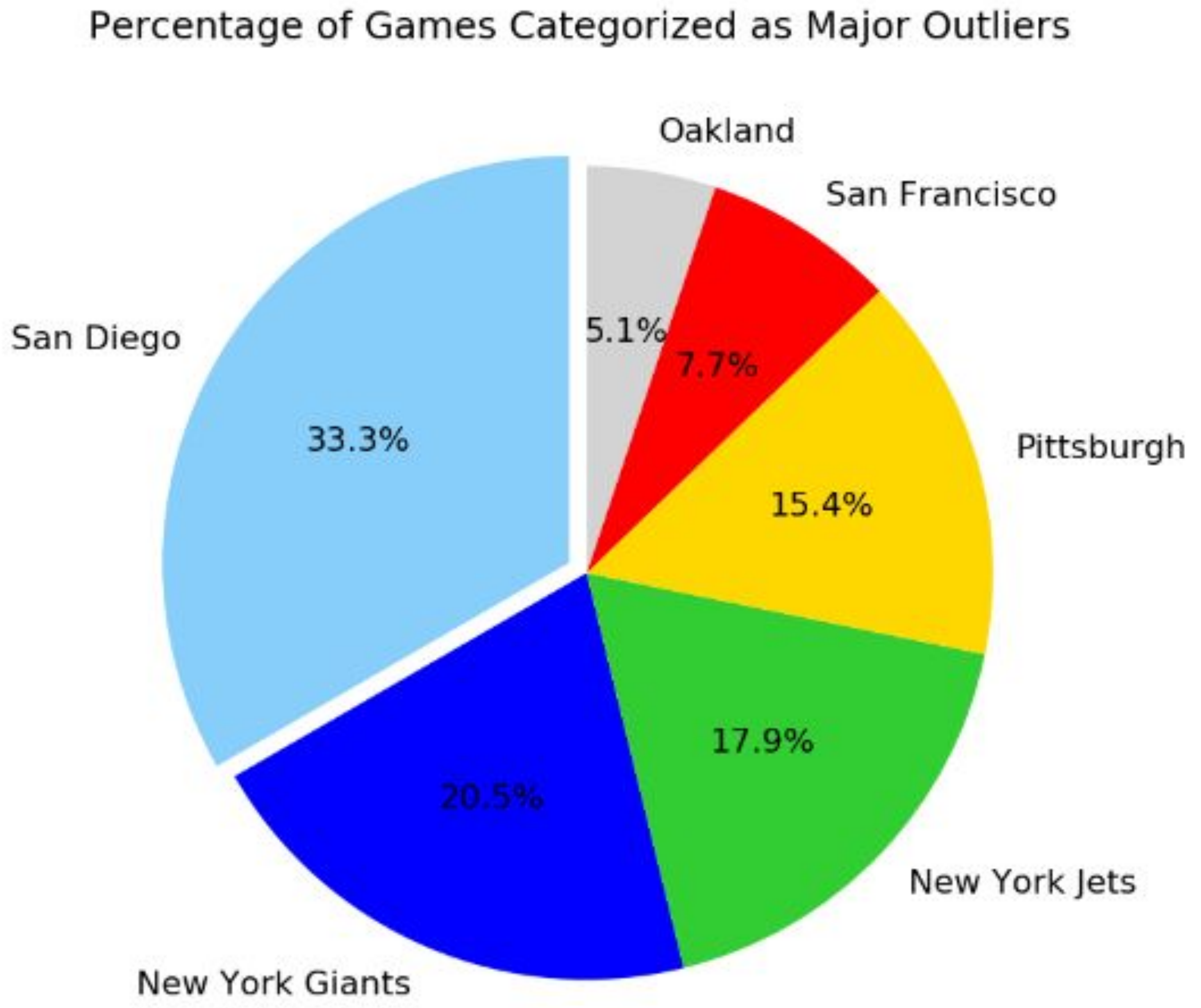
What teams have an Angry and Violent Fanbase?

- The teams that first come to mind are:
  - Oakland Raiders
  - Philadelphia Eagles
- Lets see if the assumptions are accurate:

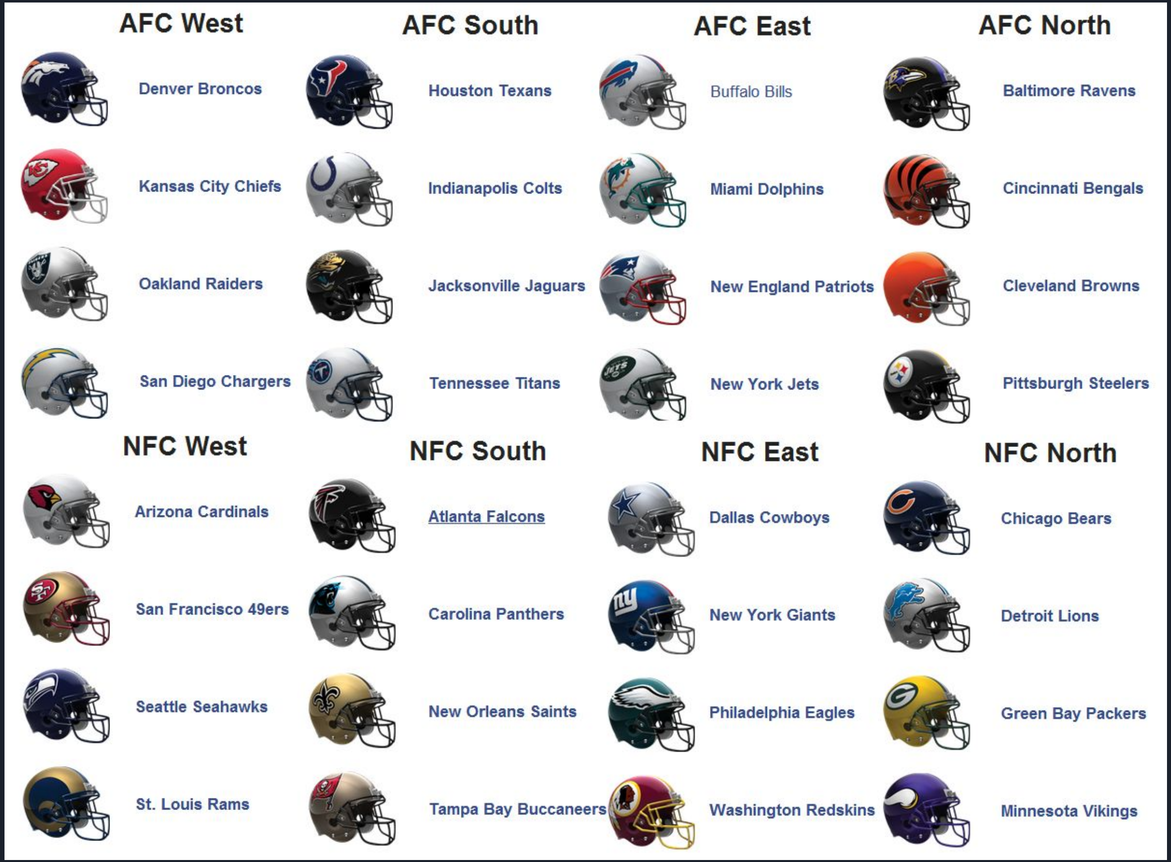
Team’s Average Number of Arrests




Percentage of Games Categorized as Outliers
Green Bay
San Francisco
New York Giants
Pittsburgh
Oakland
San Diego
New York Jets
0.9%
8.8%
22.8%
13.2%
14.0%
20.2%
20.2%


Percentage of Games Categorized as Major Outliers
Oakland
5.1%
San Francisco
7.7%
Pittsburgh
15.4%
New York Jets
17.9%
New York Giants
20.5%
San Diego
33.3%

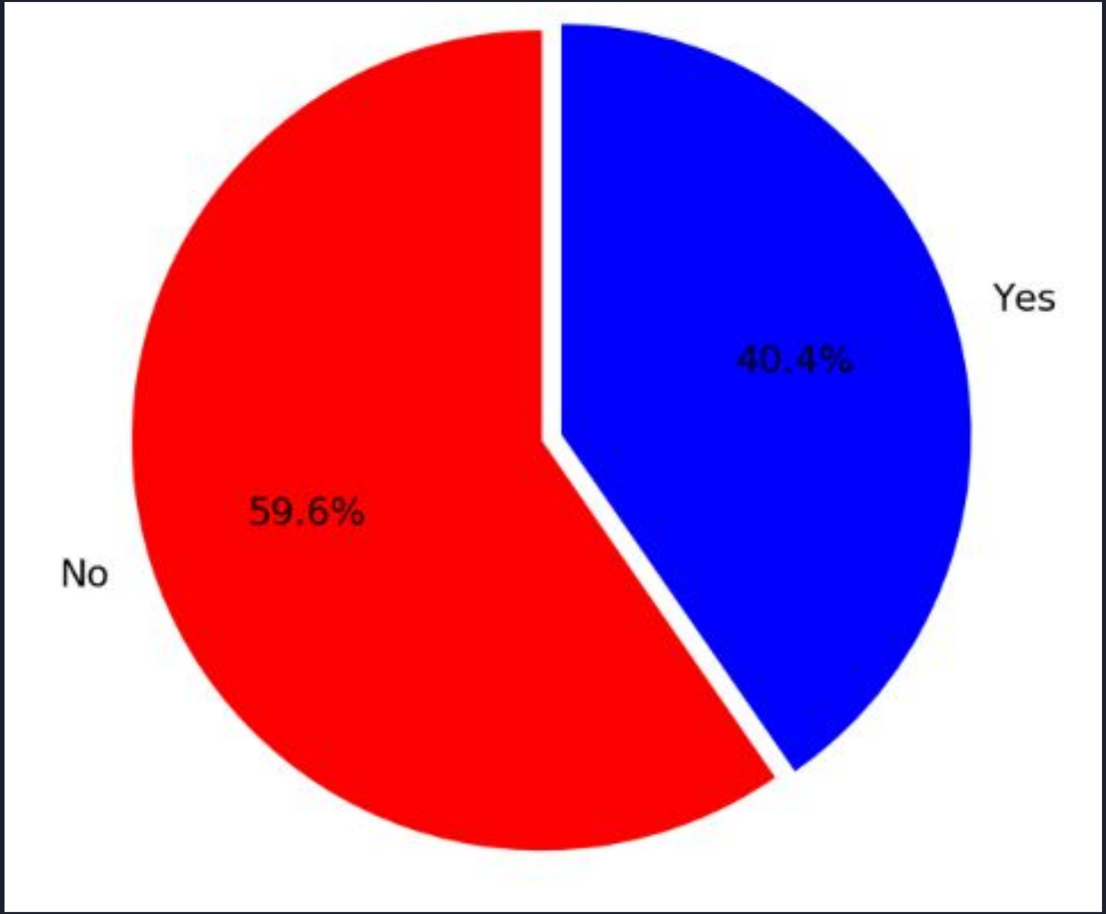
Division games



AFC West	AFC South	AFC East	AFC North
Denver Broncos	Houston Texans	Buffalo Bills	Baltimore Ravens
Kansas City Chiefs	Indianapolis Colts	Miami Dolphins	Cincinnati Bengals
Oakland Raiders	Jacksonville Jaguars	New England Patriots	Cleveland Browns
San Diego Chargers	Tennessee Titans	New York Jets	Pittsburgh Steelers

NFC West	NFC South	NFC East	NFC North
Arizona Cardinals	Atlanta Falcons	Dallas Cowboys	Chicago Bears
San Francisco 49ers	Carolina Panthers	New York Giants	Detroit Lions
Seattle Seahawks	New Orleans Saints	Philadelphia Eagles	Green Bay Packers
St. Louis Rams	Tampa Bay Buccaneers	Washington Redskins	Minnesota Vikings

Are Outliers Divisional Games?

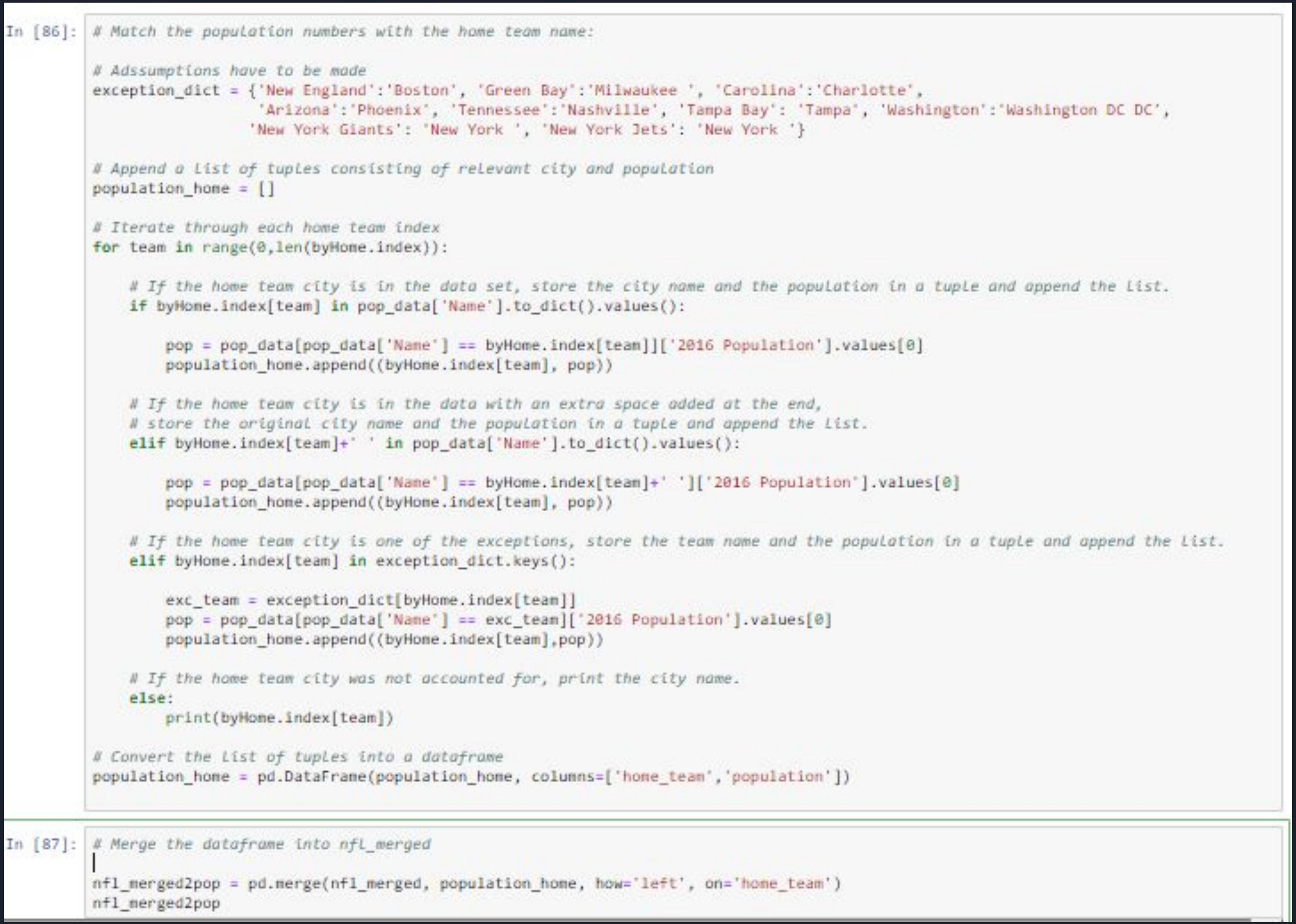


Observations:

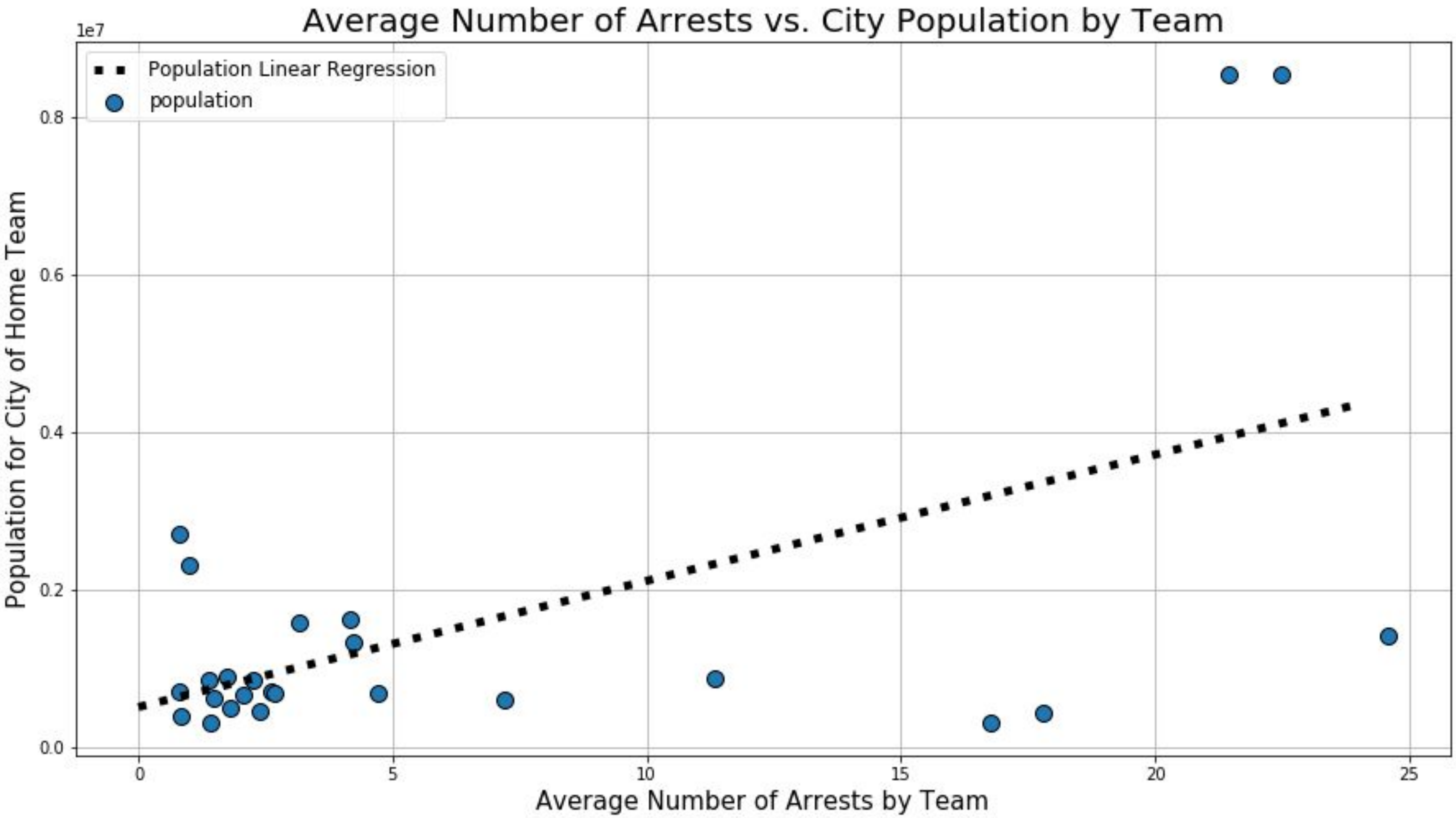
- In the five year span, San Diego, New York Jets, New York Giants, Pittsburgh, and Oakland were the highest offenders:
    - However, New York Giants have the most number of games considered outliers.
    - San Diego Chargers have the most number of games considered major outliers.
- Do Divisional games have a correlation:
    - 40.4% of the outlier games are divisional games.
    - However, a team has 6 out of 16 Divisional matchups per season (37.5% of games),
    - It cannot be inferred divisional games are more likely to result in high volume of arrests.
- The maximum arrests at a single game occured at a Chargers home game against the Raiders with 69 total arrests… Curious?
    - https://www.sbnation.com/2013/1/3/3832536/fans-arrested-chargers--fight

Number of Arrests vs. City Population:

- Some of the findings may suggest there could be a relationship with the number of stadium arrests, and the metro city population
  - Used a public dataset from http://worldpopulationreview.com/us-cities/
    - Consists of the population of the top 100 cities in the US.
  - Initially attempted to use www.census.gov and their API protocol
    - Format was inconvenient, and the above public dataset was much more obtainable.
- Are the more populated cities more likely to have arrests at their stadium?
  - We used 2016 data for the most relevant city according to the home team.
  - Note: There are some exceptions and judgement calls to this development.
    - Ex. New England Patriots play in Foxborough, MA, but for the sake of argument, we used Boston as the home city.



```
In [86]: # Match the population numbers with the home team name:

         # Adssumptions have to be made
         exception_dict = {'New England':'Boston', 'Green Bay':'Milwaukee ', 'Carolina':'Charlotte',
                           'Arizona':'Phoenix', 'Tennessee':'Nashville', 'Tampa Bay': 'Tampa', 'Washington':'Washington DC DC',
                          'New York Giants': 'New York ', 'New York Jets': 'New York '}

         # Append a list of tuples consisting of relevant city and population
         population_home = []

         # Iterate through each home team index
         for team in range(0,len(byHome.index)):

             # If the home team city is in the data set, store the city name and the population in a tuple and append the list.
             if byHome.index[team] in pop_data['Name'].to_dict().values():

                 pop = pop_data[pop_data['Name'] == byHome.index[team]]['2016 Population'].values[0]
                 population_home.append((byHome.index[team], pop))

             # If the home team city is in the data with an extra space added at the end,
             # store the original city name and the population in a tuple and append the list.
             elif byHome.index[team]+' ' in pop_data['Name'].to_dict().values():

                 pop = pop_data[pop_data['Name'] == byHome.index[team]+' ']['2016 Population'].values[0]
                 population_home.append((byHome.index[team], pop))

             # If the home team city is one of the exceptions, store the team name and the population in a tuple and append the list.
             elif byHome.index[team] in exception_dict.keys():

                 exc_team = exception_dict[byHome.index[team]]
                 pop = pop_data[pop_data['Name'] == exc_team]['2016 Population'].values[0]
                 population_home.append((byHome.index[team],pop))

             # If the home team city was not accounted for, print the city name.
             else:
                 print(byHome.index[team])

         # Convert the list of tuples into a dataframe
         population_home = pd.DataFrame(population_home, columns=['home_team','population'])

In [87]: # Merge the dataframe into nfl_merged

         nfl_merged2pop = pd.merge(nfl_merged, population_home, how='left', on='home_team')
         nfl_merged2pop
```


Average Number of Arrests vs. City Population by Team
1e7
Population Linear Regression
population
Population for City of Home Team
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
Average Number of Arrests by Team

Population Observations:

- Results do not show too much of a relationship between the 2 factors.
  - Despite the linear regression model showing a modest increase in arrests for increase in population, the results are inconclusive due to the weight New York data has on the dataset.
  - There may be a relationship if the population is under 2,000,000 with the regression model.
- Perhaps a better dataset for this sort of observation would be crime rates per city in the US.

Limitations

- We do not know why these people were arrested
  - Disorderly conduct? Public intoxication? Selling counterfeit tickets?
    - Regardless of the factors we could guess, with no data we can’t know for sure
    - Because there is no reason for arrest, we cannot run a further analysis of what type of arrests are likely to occur.
- This data only goes to 2015.
  - The last two years--and current weeks of 2018--are not accounted for
    - The data from the past two years could produce conflicting results
- Some rows did not have full datasets, in which case the N/A values were dropped.