The Untold Truth of NFL Games

Emmanuel Presley

Damarje Brown

Joshua Hale

Raheem Yusuff

What is our project?

Are there any trends to predict which games are more "dangerous" (have more fans arrested) than others?

Business question: **Can we predict what games we may need to have more police on site because we can expect that we will need to arrest more fans that game?**

Inspiration:

1. **Curious about a sport we all love**
2. **A lot of fan disputes and fights on social media, are NFL games really that bad? (Is it safe to take my kids?)**

Disclaimer – throughout the project we became interested in other insights in addition to NFL game arrests

Research Questions

Is there any correlation between the total attendance of an NFL game and the number of arrests on any given game day at that stadium?

Which stadium is the most dangerous (which stadium has the most arrests)?

Is there any correlation between score gap and total number of arrests on that game day?

How strong is home field advantage?

What game day of the week has the best attendance per stadium?

What time of the day of games has the most arrests?

Does stadium type (dome or no dome) have any relationship with arrests?

Our Datasets

NFL Arrests

- 2011-2015 Seasons
- Individual NFL Games
  - Teams
  - Score
  - Overtime
  - # of Arrests at the stadium on that gameday

NFL Games

- 2000-2019 Seasons
- Individual NFL Games
  - Teams
  - Score
  - Day of the week
  - Date and Time
  - Total turnovers of each team
  - Total yards of each team

NFL Attendance:

- 2000 – 2019 Seasons
- Grouped by Teams
- Weekly Attendance
- Season Total Attendance at all home games appended to each row
- Season Total Attendance at all away games appended to each row
- Season Total Attendance

Source: https://www.kaggle.com/datasets/washingtonpost/nfl-arrests/data
https://www.kaggle.com/datasets/sujaykapadnis/nfl-stadium-attendance-dataset

Data Engineering

Removed playoff games from Games dataset

```
games_df
```

	year	week	home_team	away_team	winner	tie	day	date	time	pts_win	pts_loss	yds_win	turnovers_win	yds_loss	turnovers_l
0	2000	1	Minnesota Vikings	Chicago Bears	Minnesota Vikings	NaN	Sun	September 3	1:00PM	30	27	374	1	425	
1	2000	1	Kansas City Chiefs	Indianapolis Colts	Indianapolis Colts	NaN	Sun	September 3	1:00PM	27	14	386	2	280	
2	2000	1	Washington Redskins	Carolina Panthers	Washington Redskins	NaN	Sun	September 3	1:01PM	20	17	396	0	236	
3	2000	1	Atlanta Falcons	San Francisco 49ers	Atlanta Falcons	NaN	Sun	September 3	1:02PM	36	28	359	1	339	
4	2000	1	Pittsburgh Steelers	Baltimore Ravens	Baltimore Ravens	NaN	Sun	September 3	1:02PM	16	0	336	0	223	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
5319	2019	Division	Kansas City Chiefs	Houston Texans	Kansas City Chiefs	NaN	Sun	January 12	3:05PM	51	31	434	1	442	
5320	2019	Division	Green Bay Packers	Seattle Seahawks	Green Bay Packers	NaN	Sun	January 12	6:40PM	28	23	344	0	375	
5321	2019	ConfChamp	Kansas City Chiefs	Tennessee Titans	Kansas City Chiefs	NaN	Sun	January 19	3:05PM	35	24	404	0	295	
5322	2019	ConfChamp	San Francisco 49ers	Green Bay Packers	San Francisco 49ers	NaN	Sun	January 19	6:40PM	37	20	354	0	358	
5323	2019	SuperBowl	Kansas City Chiefs	San Francisco 49ers	Kansas City Chiefs	NaN	Sun	February 2	6:30PM	31	20	397	2	351	

In [10]:

```
# Filter out playoff games
regular_season_games_df = games_df.loc[(games_df["week"] != "WildCard"),:]
regular_season_games_df = regular_season_games_df.loc[(regular_season_games_df["week"] != "Division"),:]
regular_season_games_df = regular_season_games_df.loc[(regular_season_games_df["week"] != "ConfChamp"),:]
regular_season_games_df = regular_season_games_df.loc[(regular_season_games_df["week"] != "SuperBowl"),:]
regular_season_games_df
```

Out[10]:

	year	week	home_team	away_team	winner	tie	day	date	time	pts_win	pts_loss	yds_win	turnovers_win	yds_loss	turnovers_loss
0	2000	1	Minnesota Vikings	Chicago Bears	Minnesota Vikings	NaN	Sun	September 3	1:00PM	30	27	374	1	425	1
1	2000	1	Kansas City Chiefs	Indianapolis Colts	Indianapolis Colts	NaN	Sun	September 3	1:00PM	27	14	386	2	280	1
2	2000	1	Washington Redskins	Carolina Panthers	Washington Redskins	NaN	Sun	September 3	1:01PM	20	17	396	0	236	1
3	2000	1	Atlanta Falcons	San Francisco 49ers	Atlanta Falcons	NaN	Sun	September 3	1:02PM	36	28	359	1	339	1
4	2000	1	Pittsburgh Steelers	Baltimore Ravens	Baltimore Ravens	NaN	Sun	September 3	1:02PM	16	0	336	0	223	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5308	2019	17	New York Giants	Philadelphia Eagles	Philadelphia Eagles	NaN	Sun	December 29	4:25PM	34	17	400	0	397	2
5309	2019	17	Dallas Cowboys	Washington Redskins	Dallas Cowboys	NaN	Sun	December 29	4:25PM	47	16	517	1	271	2
5310	2019	17	Baltimore Ravens	Pittsburgh Steelers	Baltimore Ravens	NaN	Sun	December 29	4:25PM	28	10	304	2	168	2
5311	2019	17	Los Angeles Rams	Arizona Cardinals	Los Angeles Rams	NaN	Sun	December 29	4:25PM	31	24	424	0	393	5
5312	2019	17	Seattle Seahawks	San Francisco 49ers	San Francisco 49ers	NaN	Sun	December 29	8:20PM	26	21	398	0	348	0

5104 rows × 19 columns

More Cleaning...

For all datasets, we changed the New York Giants and New York Jets ”home city” to NYG or NYJ, respectively.

In [18]:
```
# Change New York values in "home_team" and "away_team" columns to NYG for Giants and NYJ for Jets
# in regular_season_games_df dataset
arrests_df.loc[arrests_df['away_team'] == "New York Giants", 'away_team'] = "NYG"
arrests_df.loc[arrests_df['away_team'] == "New York Jets", 'away_team'] = "NYJ"
arrests_df.loc[arrests_df['home_team'] == "New York Giants", 'home_team'] = "NYG"
arrests_df.loc[arrests_df['home_team'] == "New York Jets", 'home_team'] = "NYJ"
arrests_df.home_team.unique()

```

Out[18]:
```
array(['Arizona', 'Baltimore', 'Carolina', 'Chicago', 'Cincinnati',
       'Dallas', 'Denver', 'Detroit', 'Green Bay', 'Houston',
       'Indianapolis', 'Jacksonville', 'Kansas City', 'Miami',
       'New England', 'NYG', 'NYJ', 'Oakland', 'Philadelphia',
       'Pittsburgh', 'San Diego', 'San Francisco', 'Seattle', 'Tampa Bay',
       'Tennessee', 'Washington'], dtype=object)
```

In [11]:
```
# Rename "team" column in attendance_df to "home_team_city" to match games_df
attendance_df = attendance_df.rename(columns={"team":"home_team_city"})

# Change New York values in renamed column to NYG for Giants and NYJ for Jets in attendance_df dataset
attendance_df.loc[attendance_df['team_name'] == "Giants", 'home_team_city'] = "NYG"
attendance_df.loc[attendance_df['team_name'] == "Jets", 'home_team_city'] = "NYJ"
attendance_df.home_team_city.unique()
```

Out[11]:
```
array(['Arizona', 'Atlanta', 'Baltimore', 'Buffalo', 'Carolina',
       'Chicago', 'Cincinnati', 'Cleveland', 'Dallas', 'Denver',
       'Detroit', 'Green Bay', 'Indianapolis', 'Jacksonville',
       'Kansas City', 'Miami', 'Minnesota', 'New England', 'New Orleans',
       'NYG', 'NYJ', 'Oakland', 'Philadelphia', 'Pittsburgh', 'San Diego',
       'San Francisco', 'Seattle', 'St. Louis', 'Tampa Bay', 'Tennessee',
       'Washington', 'Houston', 'Los Angeles'], dtype=object)
```

In [12]:
```
# Change New York values in "home_team" and "away_team" columns to NYG for Giants and NYJ for Jets
# in regular_season_games_df dataset
regular_season_games_df.loc[regular_season_games_df['home_team'] == "New York Giants", 'home_team_city'] = "NYG"
regular_season_games_df.loc[regular_season_games_df['home_team'] == "New York Jets", 'home_team_city'] = "NYJ"
regular_season_games_df.loc[regular_season_games_df['away_team'] == "New York Giants", 'away_team_city'] = "NYG"
regular_season_games_df.loc[regular_season_games_df['away_team'] == "New York Jets", 'away_team_city'] = "NYJ"
regular_season_games_df.away_team_city.unique()
```

Out[12]:
```
array(['Chicago', 'Indianapolis', 'Carolina', 'San Francisco',
       'Baltimore', 'Jacksonville', 'Tampa Bay', 'Detroit', 'Arizona',
       'Philadelphia', 'San Diego', 'Seattle', 'NYJ', 'Tennessee',
       'Denver', 'Miami', 'Cleveland', 'Oakland', 'Green Bay', 'NYG',
       'Kansas City', 'New Orleans', 'St. Louis', 'Washington', 'Atlanta',
       'Dallas', 'New England', 'Cincinnati', 'Pittsburgh', 'Buffalo',
       'Minnesota', 'Houston', 'Los Angeles'], dtype=object)
```

Merged Datasets (Concatenated Column)

Created a concatenated column for all 3 datasets to merge on (year, week, home team city)

In [13]:
```
# Assistance from Sherhone; Concantenate the columns we want to merge to one column in attendance_df dataset
attendance_df["Concat_Column"] = attendance_df["year"].astype(str)+attendance_df["week"].astype(str)+attendance_
attendance_df.head()
```

Out[13]:

	home_team_city	team_name	year	total	home	away	week	weekly_attendance	Concat_Column
0	Arizona	Cardinals	2000	893926	387475	506451	1	77434.0	20001Arizona
1	Arizona	Cardinals	2000	893926	387475	506451	2	66009.0	20002Arizona
2	Arizona	Cardinals	2000	893926	387475	506451	3	NaN	20003Arizona
3	Arizona	Cardinals	2000	893926	387475	506451	4	71801.0	20004Arizona
4	Arizona	Cardinals	2000	893926	387475	506451	5	66985.0	20005Arizona

In [15]:
```
# Concantenate the columns we want to merge to one column in regular_season_games_df dataset
regular_season_games_df["Concat_Column"]=regular_season_games_df["year"].astype(str) + regular_season_games_df["
regular_season_games_df.head()
```

Out[15]:

	pts_win	pts_loss	yds_win	turnovers_win	yds_loss	turnovers_loss	home_team_name	home_team_city	away_team_name	away_team_city	Concat_Column
M	30	27	374	1	425	1	Vikings	Minnesota	Bears	Chicago	20001Minnesota
M	27	14	386	2	280	1	Chiefs	Kansas City	Colts	Indianapolis	20001Kansas City
M	20	17	396	0	236	1	Redskins	Washington	Panthers	Carolina	20001Washington
M	36	28	359	1	339	1	Falcons	Atlanta	49ers	San Francisco	20001Atlanta
M	16	0	336	0	223	1	Steelers	Pittsburgh	Ravens	Baltimore	20001Pittsburgh

In [20]:
```
# Create concatted row for columns we want to merge arrests_df with attendance_and_games_df
arrests_df["Concat_Column"] = arrests_df["season"].astype(str) + arrests_df["week_num"].astype(str)+ arrests_df[
arrests_df.head()
```

Out[20]:

	season	week_num	day_of_week	gametime_local	home_team	away_team	home_score	away_score	OT_flag	arrests	division_game	Concat_Column
0	2011	1	Sunday	1:15:00 PM	Arizona	Carolina	28	21	NaN	5.0	n	20111Arizona
1	2011	4	Sunday	1:05:00 PM	Arizona	NYG	27	31	NaN	6.0	n	20114Arizona
2	2011	7	Sunday	1:05:00 PM	Arizona	Pittsburgh	20	32	NaN	9.0	n	20117Arizona
3	2011	9	Sunday	2:15:00 PM	Arizona	St. Louis	19	13	OT	6.0	y	20119Arizona
4	2011	13	Sunday	2:15:00 PM	Arizona	Dallas	19	13	OT	3.0	n	201113Arizona

1. Merged Attendance and Games Dataset first

```
attendance_and_games_df = pd.merge(regular_season_games_df, attendance_df, on = "Concat_Column")
attendance_and_games_df.head()
```

week_x	home_team	away_team	winner	tie	day	date	time	pts_win	...	away_team_city	Concat_Column	home_team_city_y	team_name	y
1	Minnesota Vikings	Chicago Bears	Minnesota Vikings	NaN	Sun	September 3	1:00PM	30	...	Chicago	20001Minnesota	Minnesota	Vikings	
1	Kansas City Chiefs	Indianapolis Colts	Indianapolis Colts	NaN	Sun	September 3	1:00PM	27	...	Indianapolis	20001Kansas City	Kansas City	Chiefs	
1	Washington Redskins	Carolina Panthers	Washington Redskins	NaN	Sun	September 3	1:01PM	20	...	Carolina	20001Washington	Washington	Redskins	
1	Atlanta Falcons	San Francisco 49ers	Atlanta Falcons	NaN	Sun	September 3	1:02PM	36	...	San Francisco	20001Atlanta	Atlanta	Falcons	
1	Pittsburgh Steelers	Baltimore Ravens	Baltimore Ravens	NaN	Sun	September 3	1:02PM	16	...	Baltimore	20001Pittsburgh	Pittsburgh	Steelers	

2. Merged first two combined datasets with Arrests dataset

```
# merge attendance_and_games_df with arrests_df on created concatenated column
combined_df = pd.merge(attendance_and_games_df, arrests_df, on = "Concat_Column")
combined_df.head()
```

	year_x	week_x	home_team_x	away_team_x	winner	tie	day	date	time	pts_win	...	week_num	day_of_week	gametime_local	home_tea
0	2011	1	Green Bay Packers	New Orleans Saints	Green Bay Packers	NaN	Thu	September 8	8:40PM	42	...	1	Thursday	7:30:00 PM	Green
1	2011	1	Baltimore Ravens	Pittsburgh Steelers	Baltimore Ravens	NaN	Sun	September 11	1:05PM	35	...	1	Sunday	1:05:00 PM	Baltin
2	2011	1	Houston Texans	Indianapolis Colts	Houston Texans	NaN	Sun	September 11	1:05PM	34	...	1	Sunday	12:00:00 PM	Hou
3	2011	1	Jacksonville Jaguars	Tennessee Titans	Jacksonville Jaguars	NaN	Sun	September 11	1:05PM	16	...	1	Sunday	1:00:00 PM	Jacksor
4	2011	1	Chicago Bears	Atlanta Falcons	Chicago Bears	NaN	Sun	September 11	1:06PM	30	...	1	Sunday	12:00:00 PM	Chic

Columns Created

Score Gap (difference of score for each game)

```
# create and fill new column for score gap
# credit for base code https://numpy.org/doc/stable/reference/generated/numpy.where.html
HA = filtered_df['home_score'] - filtered_df['away_score']
AH = filtered_df['away_score'] - filtered_df['home_score']
filtered_df['score_gap'] = np.where(HA < 0, AH, HA)

# reorganize columns
filtered_df = filtered_df.reindex(columns=['year_x', 'date', 'home_team_x', 'away_team_x','home_score','away_sco
                                           'score_gap','winner','home_advantage','weekly_attendance','arrests','
```

Hour (The hour of the day each game started)

```
# create date and time coulmn of each game
combined_df["DateTime"] = combined_df.date + ", " + combined_df.year_x.astype(str) + " "+ combined_df.gametime_l
combined_df["DateTime"] = pd.to_datetime(combined_df.DateTime)
```

```
# create hour of day column
combined_df["hour"] = combined_df.DateTime.dt.hour
```

Other Columns Created:

- Home/Away team winner
- Stadium Type (dome or not domed)

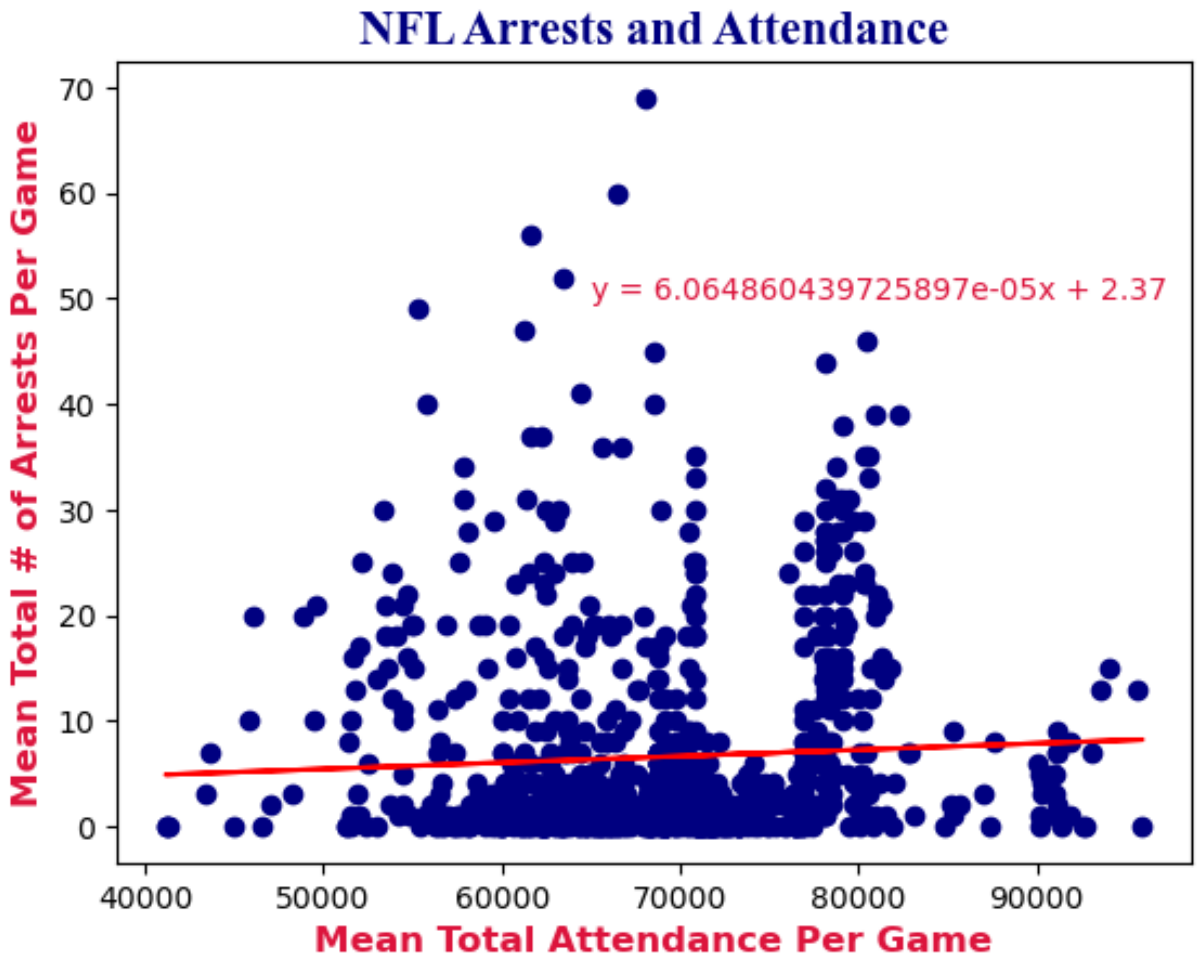
Dataset Limitations

- Arrests dataset did not include data for:
  1. Atlanta Falcons
  2. Buffalo Bills
  3. Cleveland Browns
  4. Detroit Lions
  5. Minnesota Vikings
  6. New Orleans Saints
  7. St. Louis Rams (now Los Angeles Rams)

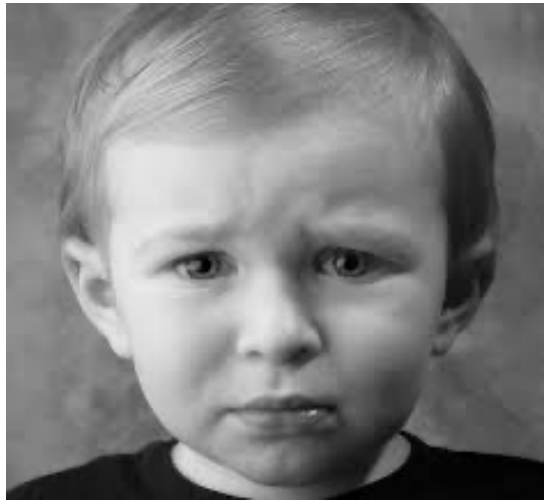
- Arrests data was only 2011-2015 seasons

- Attendance and Games data up to 2019 (3 seasons completed since)

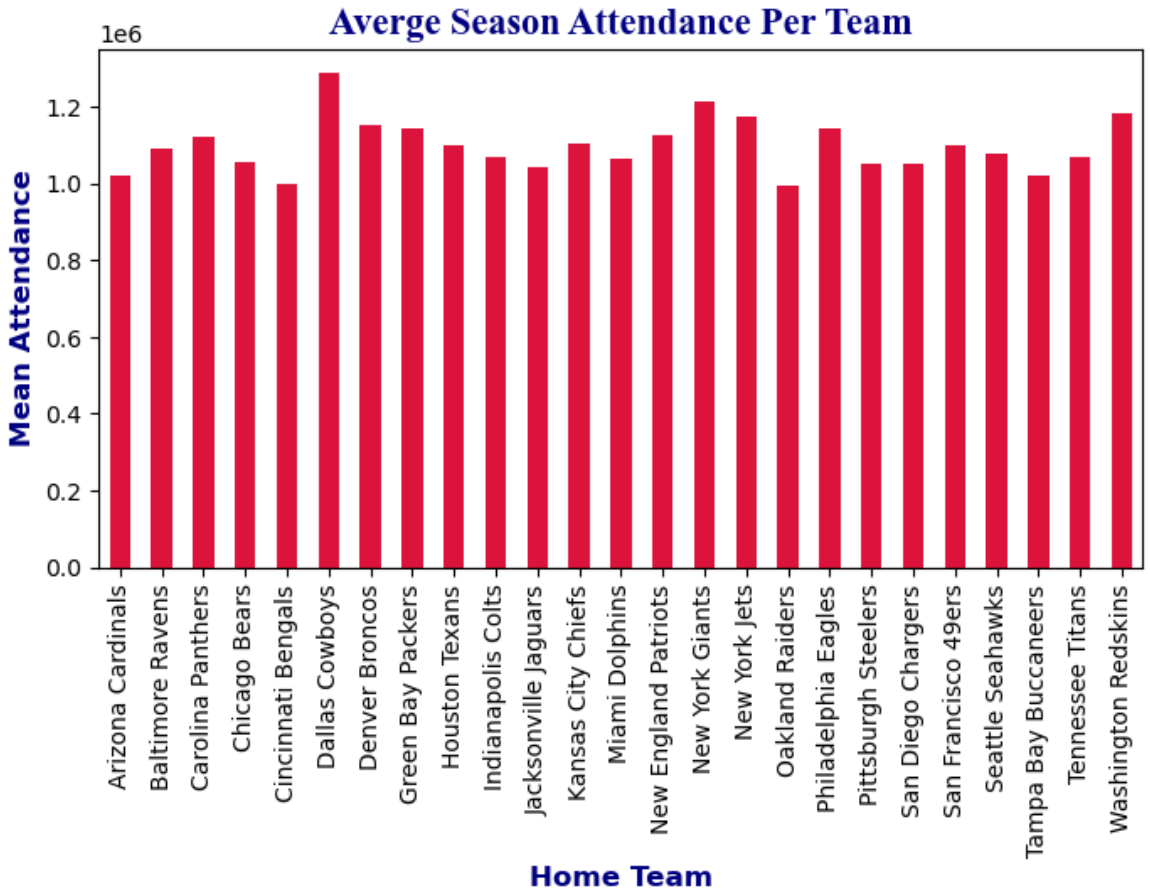
Correlation between Attendance and Arrests?

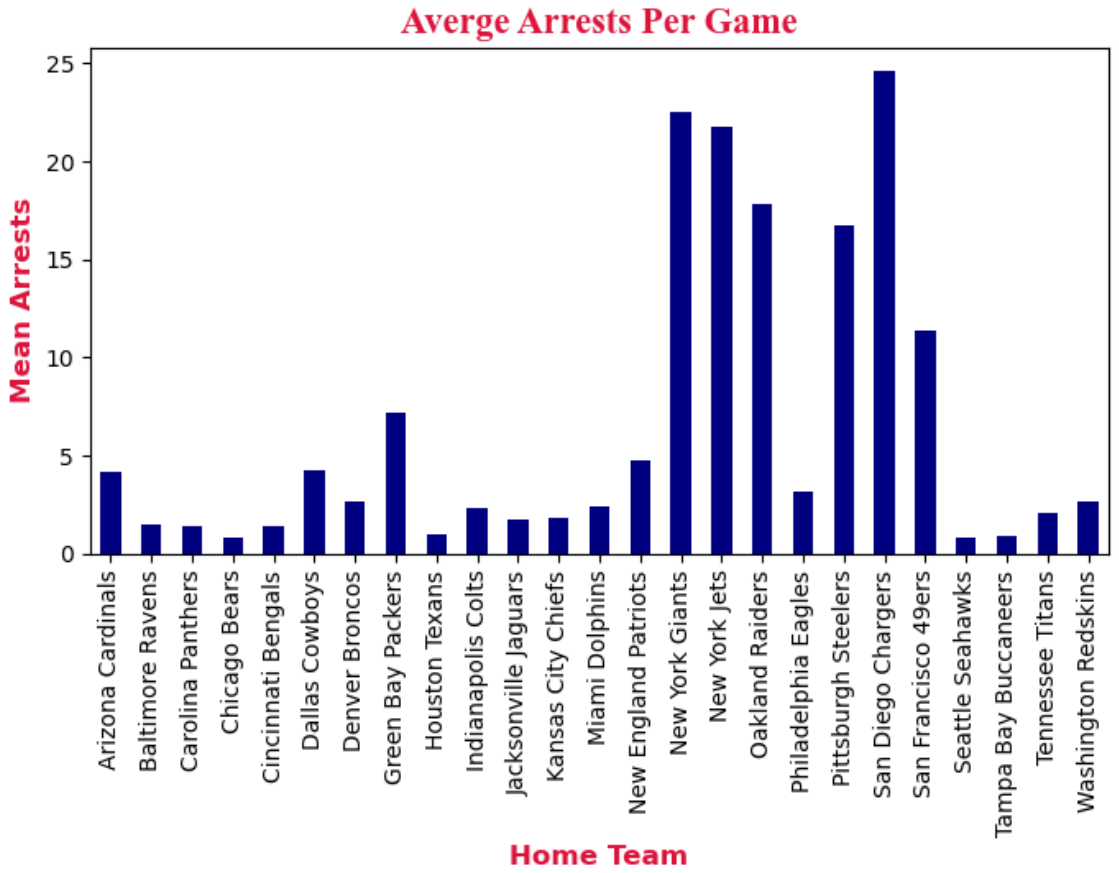


r = 0.053433 (No correlation)



But we found...

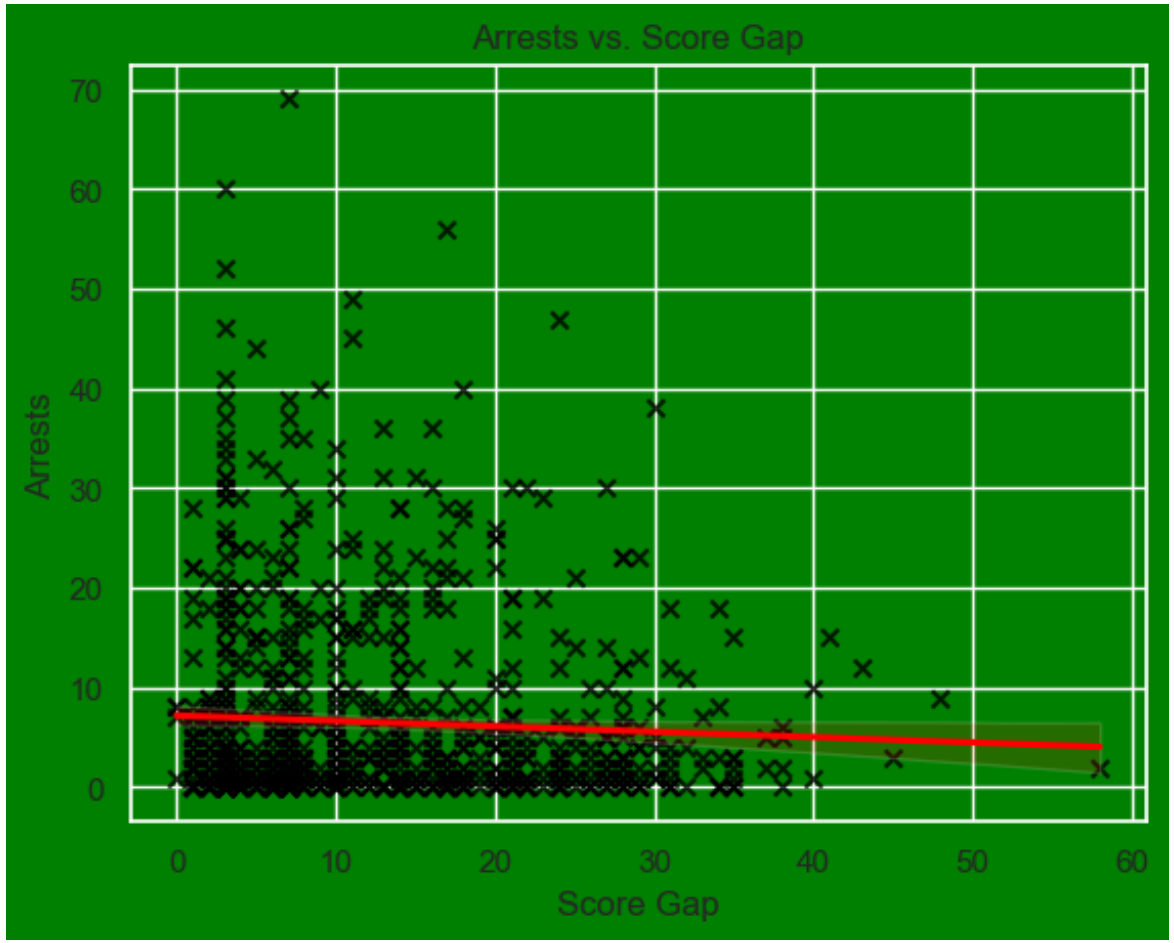




Cowboys have the highest average total attendance in a season (about 1,280,000 attendees).

San Diego Chargers had (moved to Los Angeles in 2017) the most dangerous stadium (about 25 arrests per game).

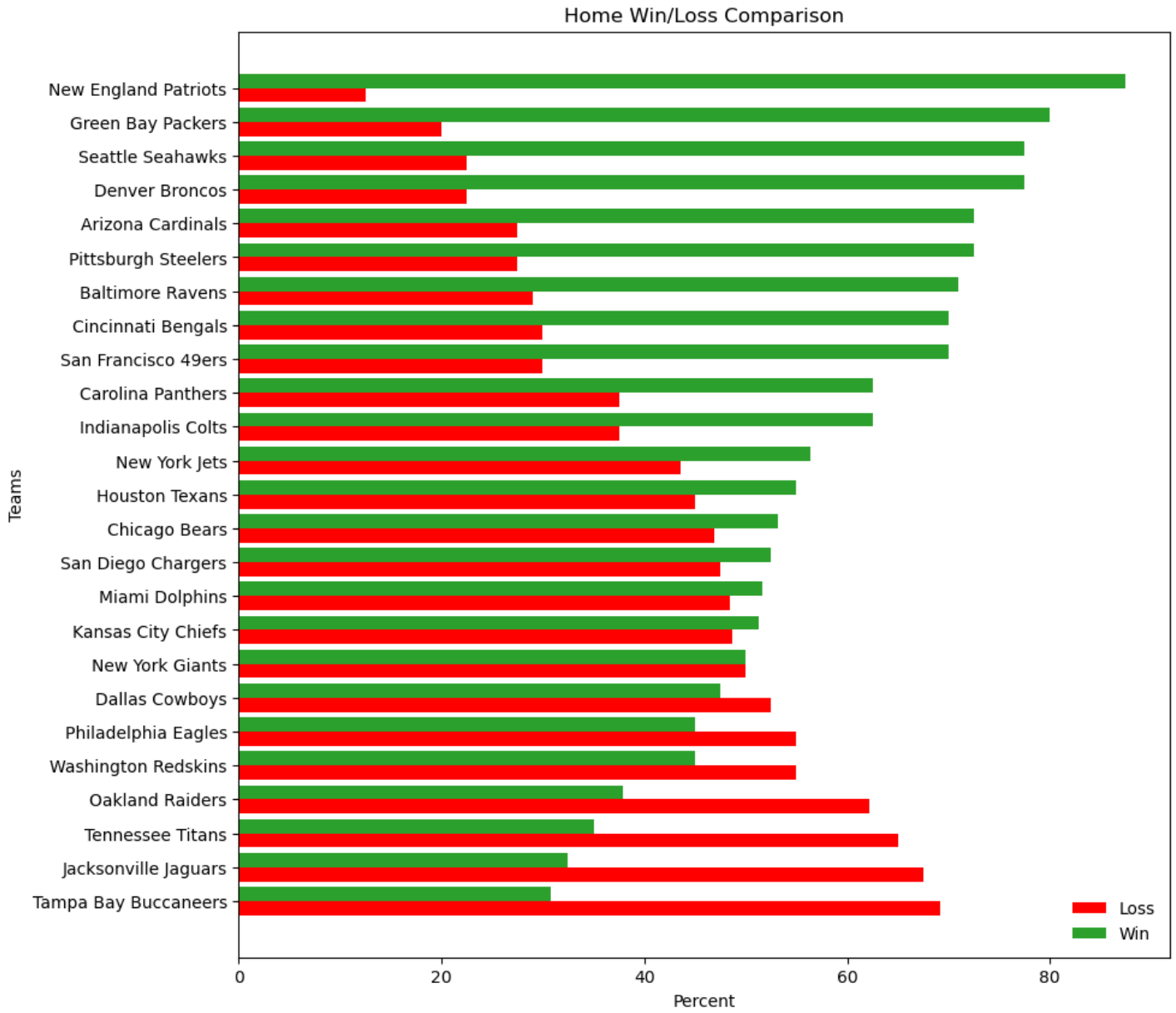
Correlation between Score Gap and Arrests?



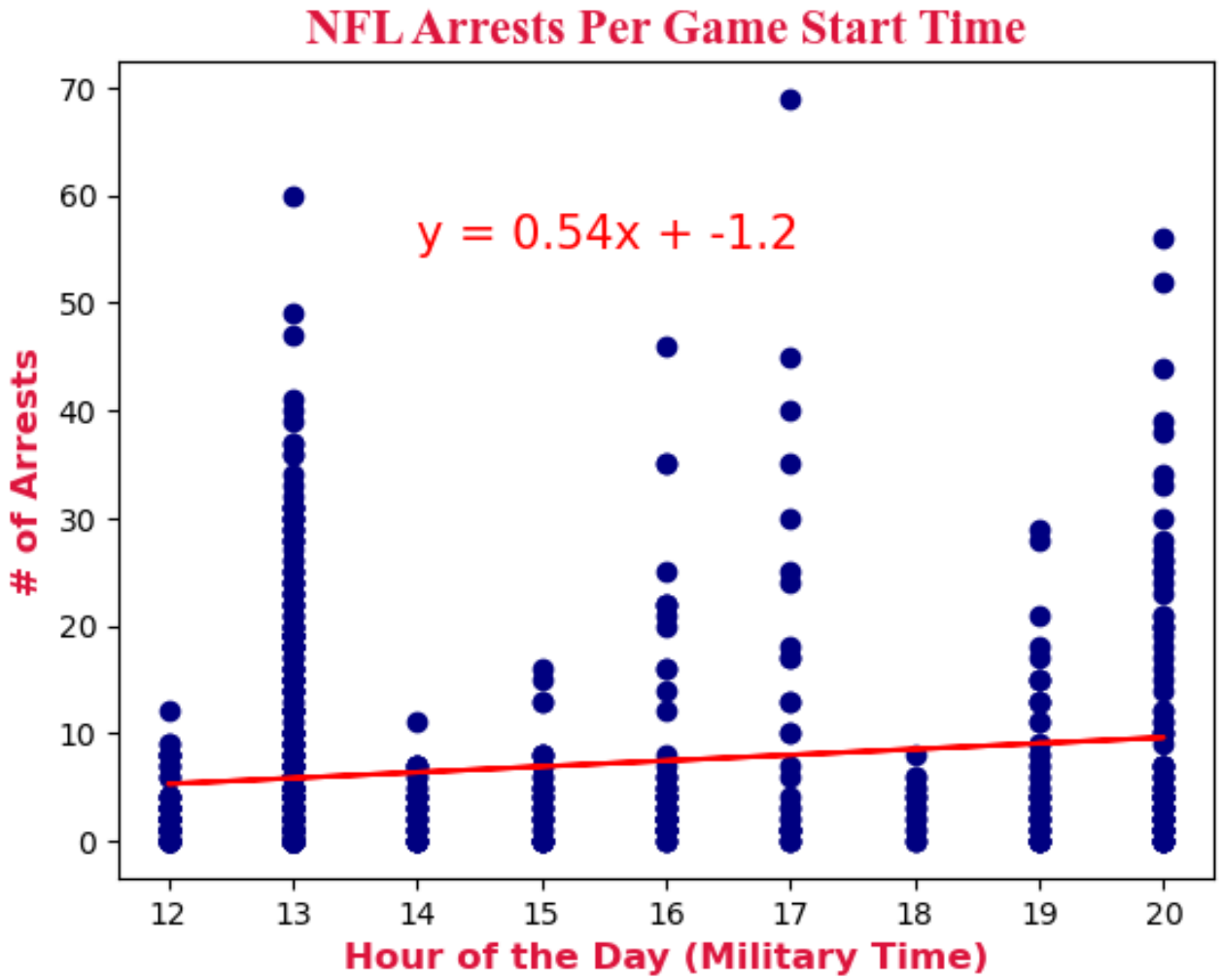
r = -0.05286 (No correlation)

However, the visual shows the closer the score of a game there may be more arrests.

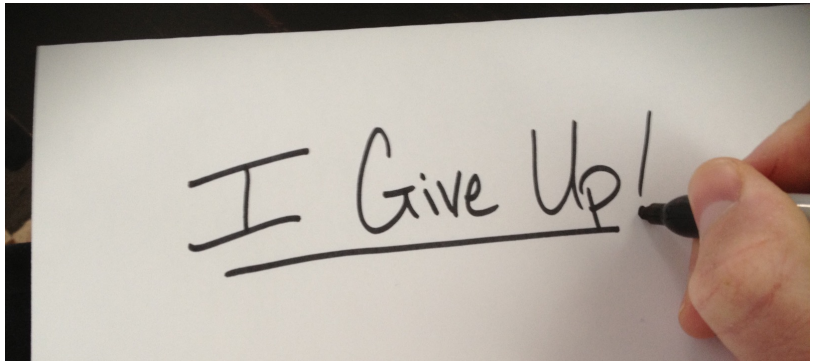
Home Advantage



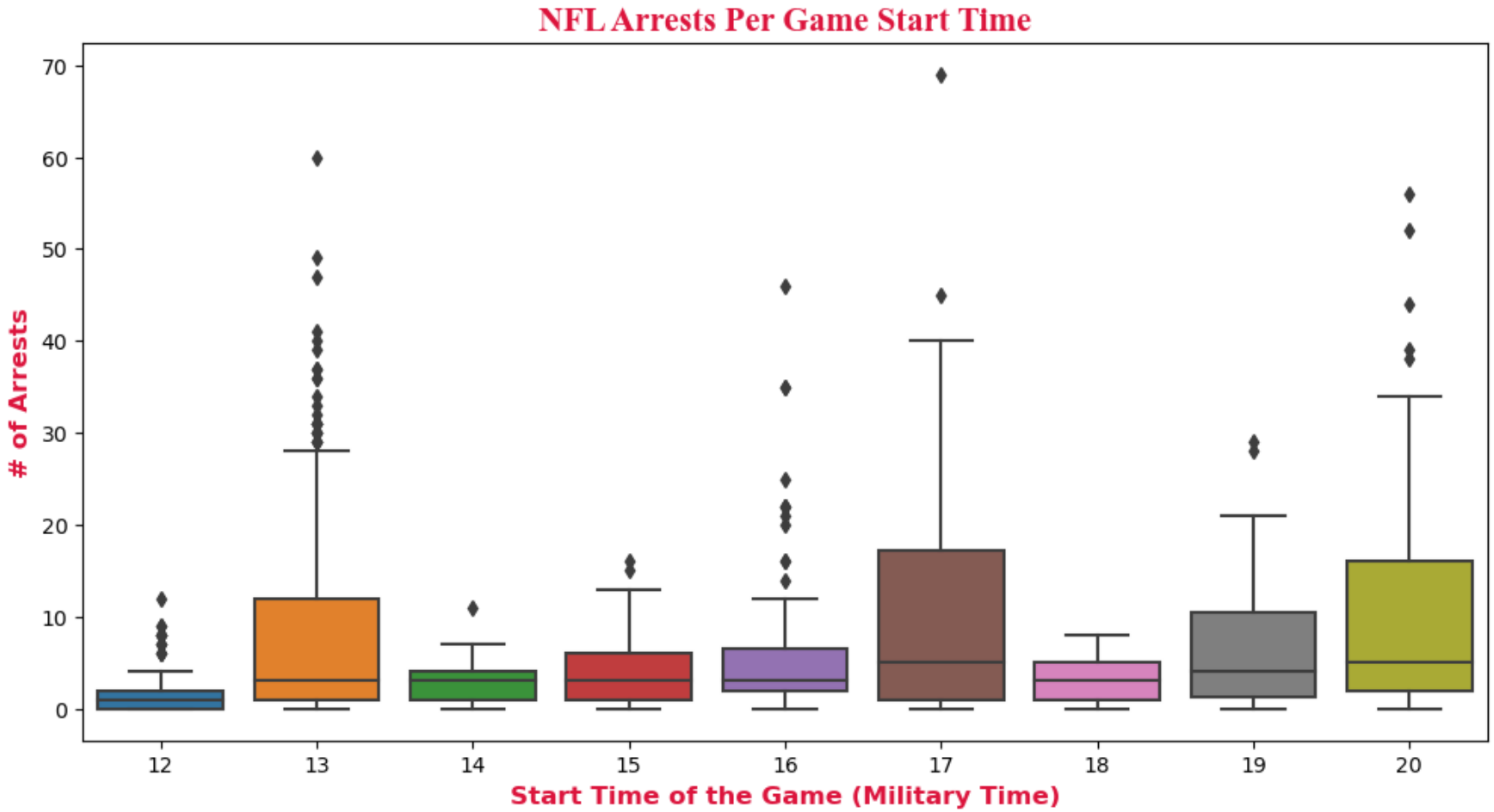
Correlation between Game Time and Arrests?



r = 0.146985 (Very weak Correlation)

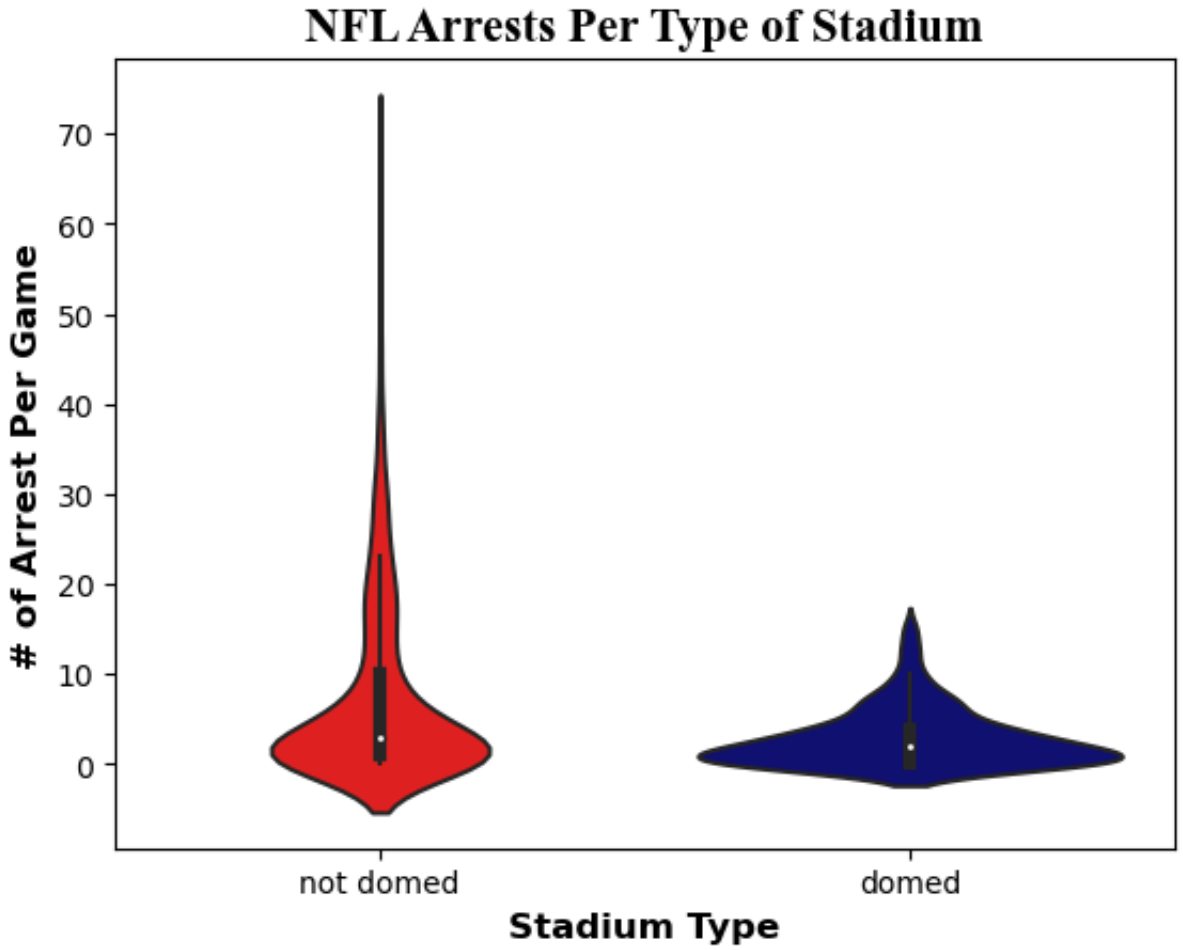


But we found...



games that start between 5:00 – 6:00 pm(respective local time), generally, have the most arrest (about 13 arrests per game). With 12:00 – 1:00 pm games the least (about 2 arrests per game)

Stadium Type and Arrests

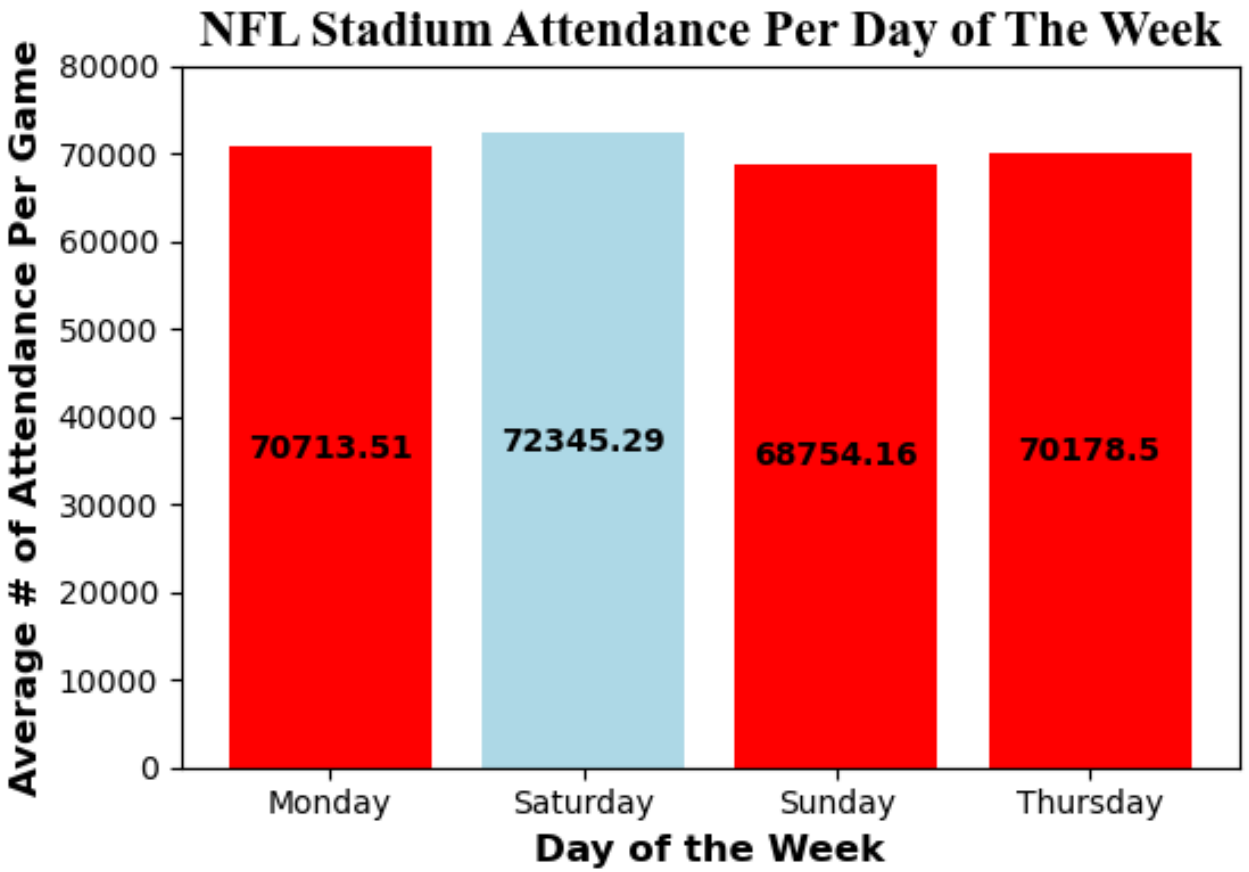


- Only 5 teams within our dataset with domed stadiums (as of today 10 teams):
  1. Arizona Cardinals
  2. Dallas Cowboys
  3. Houston Texans
  4. Indianapolis Colts
  5. New Orleans Saints

- T-test shows there is a significant difference
  - P-value = 5.04e-22

Source: https://www.profootballnetwork.com/list-of-indoor-and-outdoor-nfl-stadiums-how-many-nfl-stadiums-are-domed/

What day of the week has the best attendance?



- Saturday games have the best attendance by average (very small population of Saturday games)
- T-test showed there was no significant difference between Saturday and Sunday Games
  - P-value = 0.124
- T-test DID show there was a significant difference between Monday and Sunday Games
  - P-value = 0.0524

Wednesday was removed (there was one Wednesday game in 2012)

Takeaways

1. From a league-wide lens, there is very little correlation (that we found) between game attendance, score gap, and time of day with the number of arrests during an NFL game.
   - Probably more correlation with city demographics and crime rate
2. San Diego Chargers had the most arrests (average per game). This could have been a part of the decision to move to Los Angeles in 2017.
3. Domed stadiums from our small sample show they have less arrests. Should more teams move to domed arenas?
4. Fun Fact: In 2011 Week 10, the Chargers vs Raiders game in San Diego had the most arrest in our dataset with 69 total !

To be considered...

Limitations

- Missing Data
- Limited Population (2011-2015)
- Data not up-to-date

Bias

- Arrests at NFL stadiums have bias with each respective city's crime rate.
- Stadiums vary in total seats available, so total attendance may not be a fair parameter (use percentage instead).
- Some teams are just better than others, so they win more home games because they just win more games in general.
- There are more games played on Sunday at 1:00pm than any other day or time.

Moving Forward

- What we would do differently?
  - For research questions not involving arrests (ex. home field advantage), use only Attendance and Game datasets
  - Use percentage of stadium attendance (attendance/stadium capacity)

- What else would we do with more time?
  - Look for correlation of home city crime rate to average arrests per game
  - Look for correlation of how many police were on site to arrests per game
  - Look for correlation of alcohol sales to arrests per game
  - Look at correlations at a per team lens