Emmanuel Presley
Joshua Hale
Raheem Yusuff

The Analytical Truth of NFL Games

Introduction

For our project, we tried to find trends and correlations in NFL games such as game attendance, the gap of the score, start time of the game, and more to determine if we can predict the number of arrests that occur at a NFL stadium on a game day. These could be helpful data insights for NFL teams to make decisions on the required number of police present for each game or any other safety precautions which need to be taken. This project was inspired by our love for the sport, and on the frequency of NFL game disputes seen on social media. Are NFL games really that bad? In addition, during the project we discovered other insights within the data that we decided to take a closer look into. For example, home field advantage, and trends in attendance.

Datasets

We selected our datasets from Kaggle, see 'Reference' section for source links. We had three datasets titled: NFL Games, NFL Attendance, and NFL Arrests. NFL Games consisted of data from all NFL games played between 2000-2019 including fields such as the teams who played in each, the score, the day of the week, the date and time, the total turnovers of each team, and the total yards of each team. NFL Attendance dataset consisted of the fan attendance of every NFL game between 2000-2019. Lastly, the NFL Arrests dataset consisted of NFL game data between 2011-2015 including the teams who played, the score, if there was overtime played or not, and the number of arrests at the stadium on each game day. A huge limitation to the NFL Arrests dataset was that there was no data received for the following teams:

- Atlanta Falcons
- Buffalo Bills
- Cleveland Browns
- Detroit Lions
- Minnesota Vikings
- New Orleans Saints
- St. Louis Rams (now Los Angeles Rams)

Data Cleansing and Merging

Before we could find insights within our dataset, we had to clean and merge our datasets. We initially intended to merge the datasets on multiple like columns, however, the "Week" field in our NFL Games dataset consisted of string while the NFL Attendance and NFL Arrests "Week" column was an integer. We attempted to directly change the "Week" field in the NFL Games dataset, however, consistently received an error. We realized the dataset contained playoff games and the "Week" field included "Wildcard, "Division", "ConfChamp", and "Superbowl". As such, we removed the playoff games using a ".loc" function for each playoff game entry. *Refer to the exhibit below.*

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

Also, while we still believed we were going to merge the datasets on 3 columns we renamed the “team” field in the NFL Attendance dataset to “home_team_city” to match the other datasets. Additionally, we further cleaned our data by changing the “home city” of the New York Giants and New York Jets to “NYG” and “NYJ” in each dataset, so these two could be distinguished when we conducted the merge. *Refer to the exhibit below.*

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)
```

After cleaning the data, we attempted to use a “.merge” function on multiple columns between our datasets, but were not successful. To resolve the issue, we decided to concatenate the 3 columns we were attempting to merge on within each data set and use this common column to conduct the merge. (Season, Week, Home Team). *Refer to the exhibits below.*

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]:

e	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
1	30	27	374	1	425	1	Vikings	Minnesota	Bears	Chicago	20001Minnesota
1	27	14	386	2	280	1	Chiefs	Kansas City	Colts	Indianapolis	20001Kansas City
1	20	17	396	0	236	1	Redskins	Washington	Panthers	Carolina	20001Washington
1	36	28	359	1	339	1	Falcons	Atlanta	49ers	San Francisco	20001Atlanta
1	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

First, we first merged the NFL Attendance and NFL game datasets using the created concatenated columns.

```
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	

Then, we merged this newly created dataset with NFL arrests using the created concatenated columns.

```
# 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

Correlation Between Total Attendance and Total Arrest

After cleaning the data to the point we can conduct our analysis. Now we can work on answering our first question “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?” The first thing we did was to pull the tables that we needed from the large data.

In [61]:
```
combined_df = combined_df[["home_team_x", "total", "weekly_attendance", "arrests", "year_x"]]
combined_df
```

We collected the home team for the game, the total attendance for the game, the weekly attendance, arrest for the game and the NFL season the game was played.

In [62]:
```
#Corr between total attendance and giving number of arrest
#create a filtered dataframe with the information needed ( year, date, weekly attendance, )
attendance_arrest_df = combined_df.loc[:,['year_x','home_team_x','total','arrests']]

# call to see table
attendance_arrest_df.head(10)
```

Out[62]:

	year_x	home_team_x	total	arrests
0	2011	Green Bay Packers	1123023	8.0
1	2011	Baltimore Ravens	1083902	1.0
2	2011	Houston Texans	1059702	2.0
3	2011	Jacksonville Jaguars	1049655	4.0
4	2011	Chicago Bears	1053343	1.0
5	2011	Kansas City Chiefs	1107206	0.0
6	2011	Tampa Bay Buccaneers	1019250	0.0
7	2011	Arizona Cardinals	1001663	5.0
8	2011	San Diego Chargers	1064892	15.0
9	2011	San Francisco 49ers	1065296	3.0

After creating the new dataframe and checking for null values in our columns. We conducted our statistical analysis of the dataframe and created a table with the results. We did this in two ways,

first we used the data frame to show us on average what home team had the highest total attendance from 2011 - 2015. Then we used the same data frame to see what home team on average had the highest total arrest from 2011 - 2015. Graph both of them and see if there are any correlations between the total arrest and the total attendance.

First the average for the total attendance for the home teams from 2011 - 2015 and created a summary stat from the data frame.

In [64]:
```
mean = attendance_arrest_df.groupby(["home_team_x"])["total"].mean()
median = attendance_arrest_df.groupby(["home_team_x"])["total"].median()
var = attendance_arrest_df.groupby(["home_team_x"])["total"].var()
std = attendance_arrest_df.groupby(["home_team_x"])["total"].std()
sem = attendance_arrest_df.groupby(["home_team_x"])["total"].sem()

summary_stat = pd.DataFrame({"Mean Total Attendance":mean,
                             "Median Total Attendance":median,
                            "Total Attendance Variance":var,
                            "Total Attendance Std. Dev.":std,
                            "Total Attendance Std. Err.":sem})
```

We then check to see if the data frame had any null values in there. We then created a new dataset containing the home team on one column and the average total attendance on the other column. So we can create the graph.

In [69]:
```
Summary_Stat_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 25 entries, Arizona Cardinals to Washington Redskins
Data columns (total 5 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   Mean Total Attendance       25 non-null     float64
 1   Median Total Attendance     25 non-null     float64
 2   Total Attendance Variance   25 non-null     float64
 3   Total Attendance Std. Dev.  25 non-null     float64
 4   Total Attendance Std. Err.  25 non-null     float64
dtypes: float64(5)
memory usage: 1.2+ KB
```

In [70]:
```
mean_attendance = Summary_Stat_df["Mean Total Attendance"]
mean_attendance
```

Out[70]: home_team_x

After creating the first dataframe to see what team had the highest total attendance we can now use the same process to find out what team had the highest arrest.

In [78]:
```
#find mean, median, var, std, sem
mean = attendance_arrest_df.groupby(["home_team_x"])["arrests"].mean()
median = attendance_arrest_df.groupby(["home_team_x"])["arrests"].median()
var = attendance_arrest_df.groupby(["home_team_x"])["arrests"].var()
std = attendance_arrest_df.groupby(["home_team_x"])["arrests"].std()
sem = attendance_arrest_df.groupby(["home_team_x"])["arrests"].sem()

summary_stat2 = pd.DataFrame({"Mean Arrests":mean,
                              "Median Arrests":median,
                             "Arrests Variance":var,
                             "Arrests Std. Dev.":std,
                             "Arrests Std. Err.":sem})
```

In [83]:
```
Summary_Stat2_df.info()
```

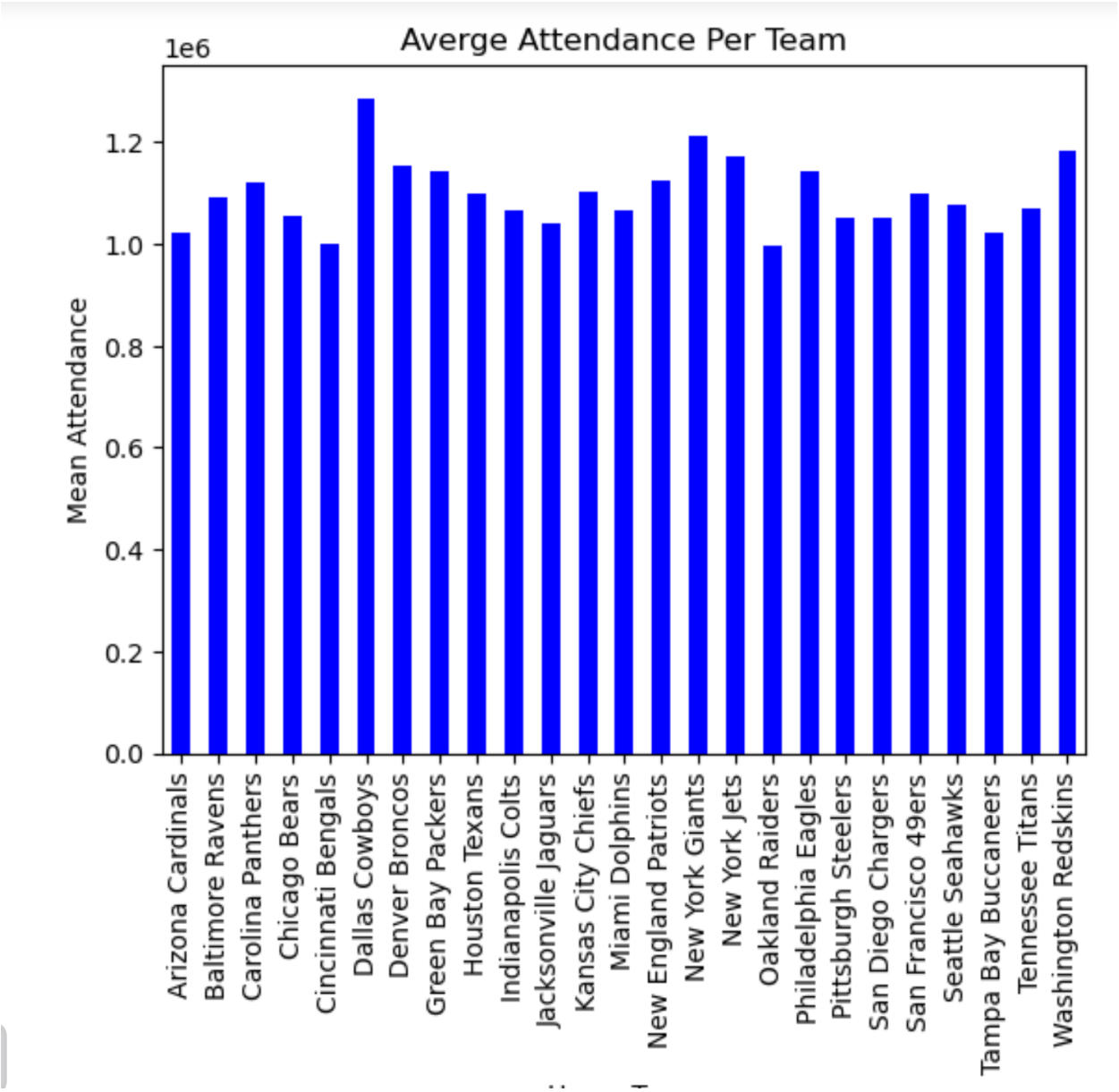
```
<class 'pandas.core.frame.DataFrame'>
Index: 25 entries, Arizona Cardinals to Washington Redskins
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Mean Arrests       25 non-null     float64
 1   Median Arrests     25 non-null     float64
 2   Arrests Variance   25 non-null     float64
 3   Arrests Std. Dev.  25 non-null     float64
 4   Arrests Std. Err.  25 non-null     float64
dtypes: float64(5)
memory usage: 1.2+ KB
```

In [84]:
```
mean_arrest = Summary_Stat2_df["Mean Arrests"]
mean_arrest
```

Out[84]:
```
home_team_x
Arizona Cardinals        4.150000
Baltimore Ravens         1.483871
Carolina Panthers        1.375000
Chicago Bears            0.812500
Cincinnati Bengals       1.425000
Dallas Cowboys           4.225000
Denver Broncos           2.625000
Green Bay Packers        7.200000
Houston Texans           1.000000
Indianapolis Colts       2.275000
Jacksonville Jaguars     1.729730
Kansas City Chiefs       1.820513
Miami Dolphins           2.387097
New England Patriots     4.700000
New York Giants         22.475000
New York Jets           21.717949
Oakland Raiders         17.783784
Philadelphia Eagles      3.150000
```

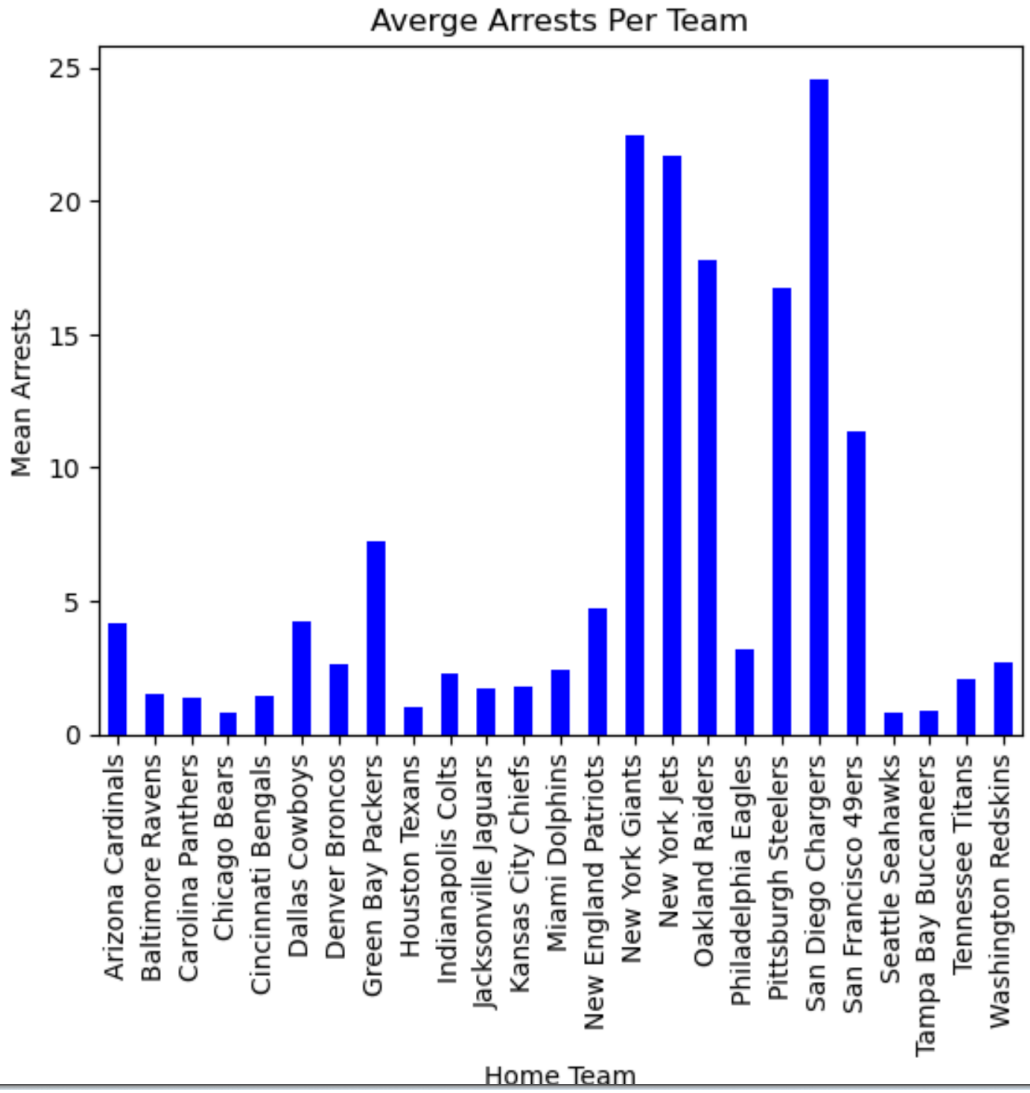
Now we can plot the two graphs and check to see if there were any correlations between total attendance and total arrest.

```
In [71]: plot_pandas = mean_attendance.plot.bar(color='b')
         # Set the xlabel, ylabel, and title using class methods
         plt.xlabel("Home Team")
         plt.ylabel("Mean Attendance")
         plt.title("Averge Attendance Per Team")
```



```
In [85]: plot_pandas = mean_arrest.plot.bar(color='b')
         # Set the xlabel, ylabel, and title using class methods
         plt.xlabel("Home Team")
         plt.ylabel("Mean Arrests")
         plt.title("Averge Arrests Per Team")
```

```
Out[85]: Text(0.5, 1.0, 'Averge Arrests Per Team')
```



We see that there was not a big correlation between the attendance and arrest. Most of the most attended NFL stadiums have low arrest rates. Which gave us a question for future analysis: why are people getting arrested in games in the states of New York and California? Since 5 out 8 stadiums with high arrest rates were in the two states.

This also answered our second question for the group. What stadium is the most dangerous to attend? Since our definition of danger in our project is the home team with the highest arrest is going to be the most dangerous stadium to go to. We find that the San Diego Chargers stadium is the most dangerous to attend with the highest mean arrest of 24.5. Just to be safe from arrest, avoid stadiums in New York or California.

Home Advantage and Score Gap-Arrests correlation

Before we could dive into answering a couple of our research questions regarding the strength of home field advantage, as well as, the correlation of score gap to arrests, we first had to filter out some columns in our data. This way it would make finding the necessary columns much easier when it came time to do our visualizations. To accomplish this we created a new subset data frame 'filtered_df' that showed only the dates, home and away teams, their respective scores, the winner of the specific date's game, the attendance of each individual game, the arrests made that day, as well as, a list of the total teams that we had data on.

```
# Create filtered dataframe with only pertinent information
filtered_df = combined_df[['year_x','date','home_team_x','away_team_x','home_score','away_score','winner','weekly_attenda
                           'arrests','team_name']]
# test for effect
filtered_df.head()
```

	year_x	date	home_team_x	away_team_x	home_score	away_score	winner	weekly_attendance	arrests	team_name
0	2011	September 8	Green Bay Packers	New Orleans Saints	42	34	Green Bay Packers	70555.0	8.0	Packers
1	2011	September 11	Baltimore Ravens	Pittsburgh Steelers	35	7	Baltimore Ravens	71434.0	1.0	Ravens
2	2011	September 11	Houston Texans	Indianapolis Colts	34	7	Houston Texans	71444.0	2.0	Texans
3	2011	September 11	Jacksonville Jaguars	Tennessee Titans	16	14	Jacksonville Jaguars	61619.0	4.0	Jaguars
4	2011	September 11	Chicago Bears	Atlanta Falcons	30	12	Chicago Bears	62115.0	1.0	Bears

Afterwards, we saw that the arrests columns had some null values where the data was not recorded; as this only affected a small number of games and the numbers of arrests at other games were relatively small, we chose to replace the null values with 0's. In this way, we could still have some data to work with rather than disposing of it. With the null values taken care of, we then created a new column that would show up the percent of attendants arrested at each game. We also created a column that stated whether the winner of each game had home field advantage; this would assist us later with writing code that would compare the winner to the home advantage column.

```
filtered_df.head()
```

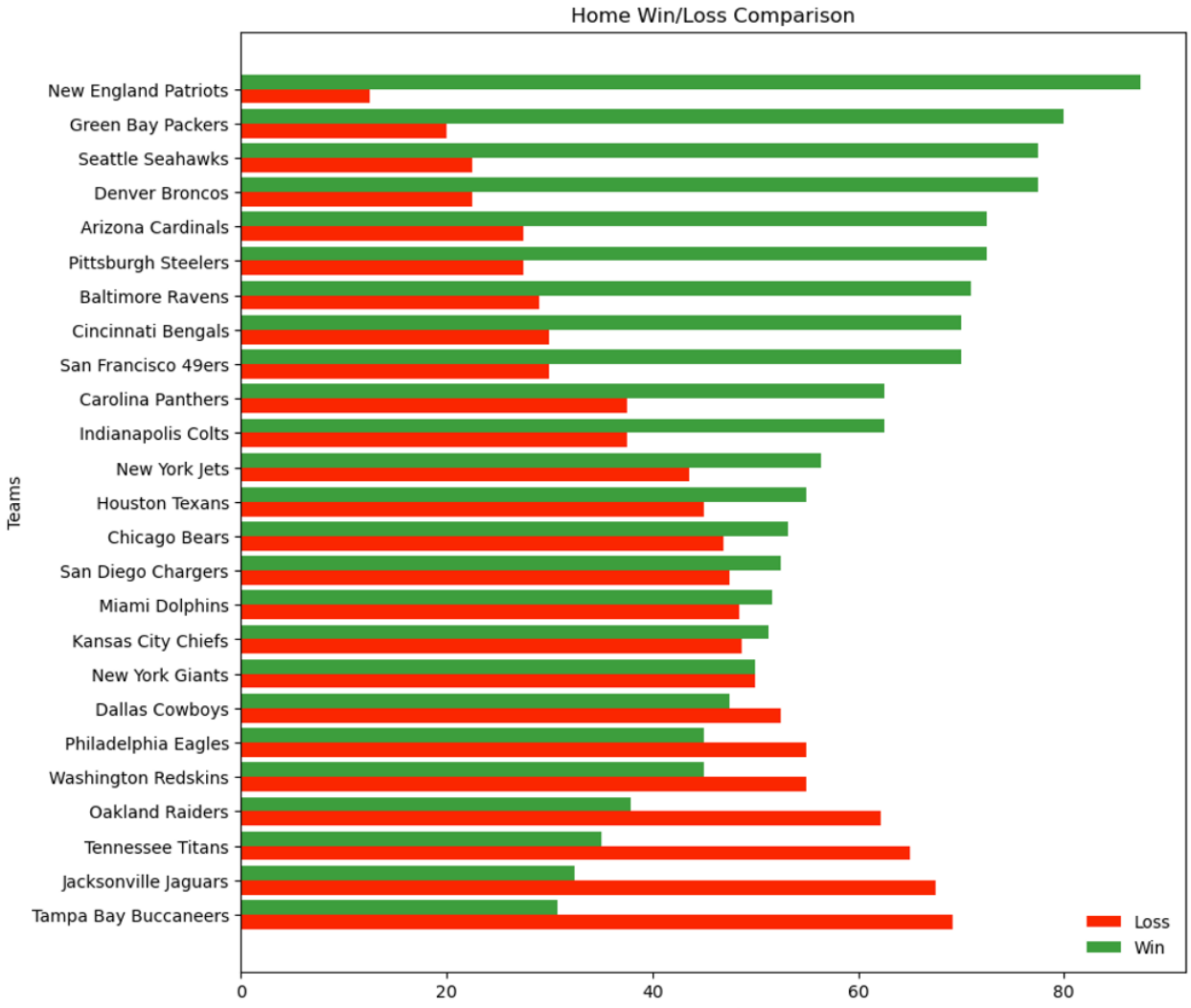
ear_x	date	home_team_x	away_team_x	home_score	away_score	score_gap	winner	home_advantage	weekly_attendance	arrests	%_arrested
2011	September 8	Green Bay Packers	New Orleans Saints	42	34	8	Green Bay Packers	Yes	70555.0	8.0	0.011339
2011	September 11	Baltimore Ravens	Pittsburgh Steelers	35	7	28	Baltimore Ravens	Yes	71434.0	1.0	0.001400
2011	September 11	Houston Texans	Indianapolis Colts	34	7	27	Houston Texans	Yes	71444.0	2.0	0.002799
2011	September 11	Jacksonville Jaguars	Tennessee Titans	16	14	2	Jacksonville Jaguars	Yes	61619.0	4.0	0.006492
2011	September 11	Chicago Bears	Atlanta Falcons	30	12	18	Chicago Bears	Yes	62115.0	1.0	0.001610

Then we created a new column that would show the score gap of the game, as well as reorganize the columns so that they could be a bit easier to read and aid in locating pertinent information. We then created 2 more columns that would allow us to easily write a code which gave percentage values to each team when they won and lost when they had the home field advantage.

```
# make and fill new columns for Home Wins/ Home Losses
filtered_df['home_wins'] = filtered_df.home_team_x == filtered_df.winner
filtered_df['home_loss'] = filtered_df.home_team_x != filtered_df.winner
win_perc = filtered_df.groupby('home_team_x').home_wins.sum()/ filtered_df.home_team_x.value_counts()*100
win_perc = win_perc.sort_values(ascending=False)
loss_perc = filtered_df.groupby('home_team_x').home_loss.sum()/ filtered_df.home_team_x.value_counts()*100
loss_perc = loss_perc.sort_values(ascending=True)
filtered_df.head(15)
```

	year_x	date	home_team_x	away_team_x	home_score	away_score	score_gap	winner	home_advantage	weekly_attendance	arrests	%_
0	2011	September 8	Green Bay Packers	New Orleans Saints	42	34	8	Green Bay Packers	Yes	70555.0	8.0	(
1	2011	September 11	Baltimore Ravens	Pittsburgh Steelers	35	7	28	Baltimore Ravens	Yes	71434.0	1.0	(
2	2011	September 11	Houston Texans	Indianapolis Colts	34	7	27	Houston Texans	Yes	71444.0	2.0	(
3	2011	September 11	Jacksonville Jaguars	Tennessee Titans	16	14	2	Jacksonville Jaguars	Yes	61619.0	4.0	(
4	2011	September 11	Chicago Bears	Atlanta Falcons	30	12	18	Chicago Bears	Yes	62115.0	1.0	(
5	2011	September 11	Kansas City Chiefs	Buffalo Bills	7	41	34	Buffalo Bills	No	68755.0	0.0	(
6	2011	September 11	Tampa Bay Buccaneers	Detroit Lions	20	27	7	Detroit Lions	No	51274.0	0.0	(

Once we had these variables and columns, we were to create a visual representation showing the percentage of home games that were won, and also lost. This visualization somewhat confirms a partial advantage when playing on the team's home turf, with 17 teams winning their home games. While our data was only limited to 25 teams, this demonstrated at least some type of home field advantage, statistically speaking.



There are a few factors to consider, however. Our data was not able to take into account environmental factors such as temperature, elevation, or humidity on each game day. These pieces of information could be crucial in better understanding what gave the winning teams a slight edge. For example, a team that is used to playing in colder temperatures may find themselves overheating at a game where the temperature is significantly higher than what they are used to.

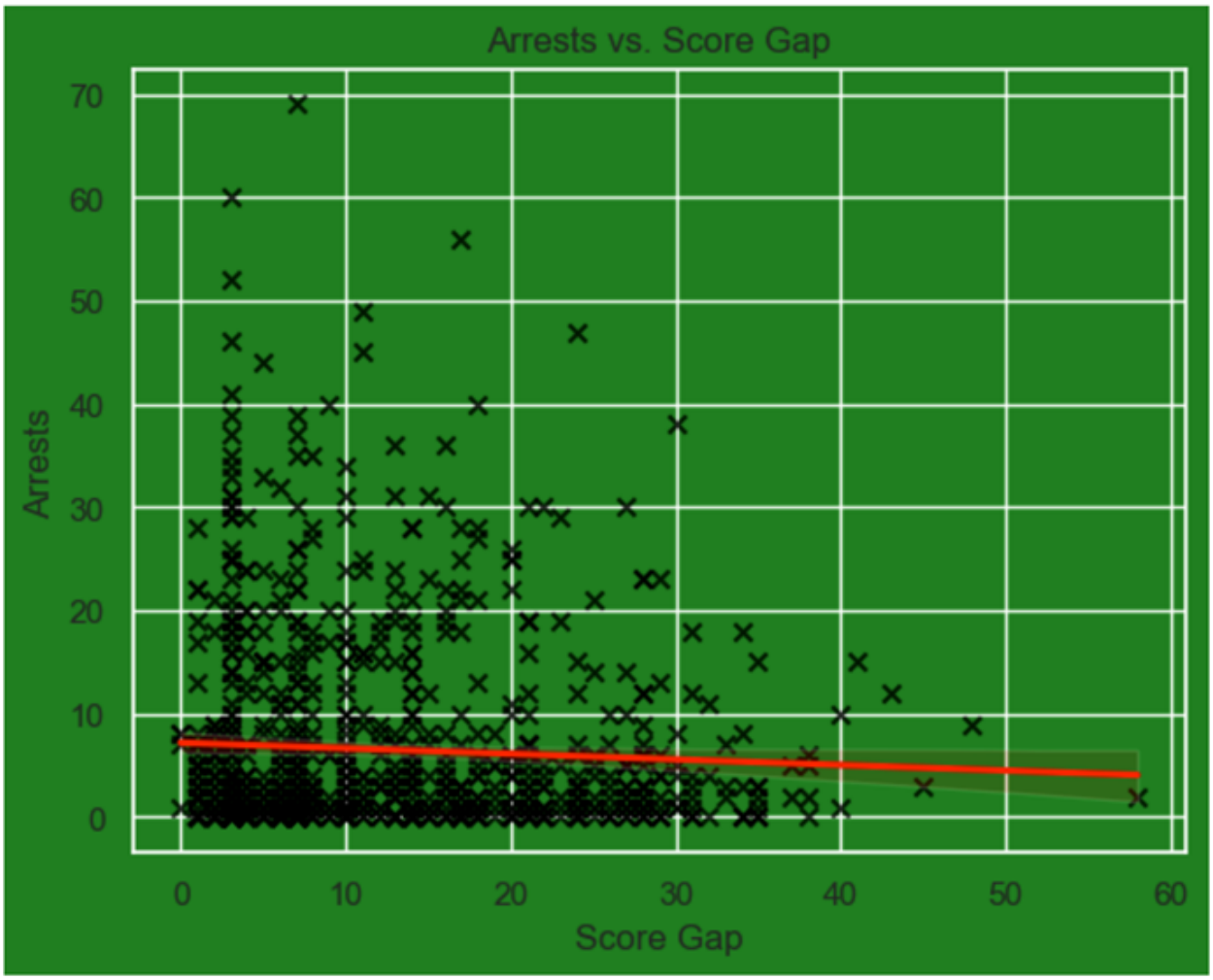
And finally, we took the rest of our filtered data, and created a plot graph that showed the relationship of arrests with score gaps. With some visual tweaking in the code, we were able to create a football styled graph that actually went against our initial opinions that a higher score gap meant more arrests.

```
# Are there any correlation between score gap and total number of arrests on that game day
sns.set(rc= {'axes.facecolor':'green','figure.facecolor':'green'})
sns.regplot(x="score_gap", y="arrests", data=filtered_df, marker='x',scatter_kws= {"color": "black"},line_kws={'color':
;

#add correlation coefficient to plot
filtered_df['score_gap'].corr(filtered_df['arrests'])

print(f"The correlation of score gap in a game to the number of arrests that day is {filtered_df['score_gap'].corr(filte
```

```
The correlation of score gap in a game to the number of arrests that day is -0.052864223082118994. There is very little corr
elaton to having a bigger gap in score and increased number of arrests. It can then be theorized that the closer a game is t
o a tie, the more likely it is that arrests will be made.
```



On the contrary, the closer the score gap was, the more likely people were to be arrested. This would suggest that when a score gap got wider, the fans perhaps lost hope and weren't as emotionally invested in the game. It also must be considered that this data didn't include information such as how much alcohol was sold, age and sex of the arrested person(s), or what the arrestable offense was.

Game Time-Arrests Correlation

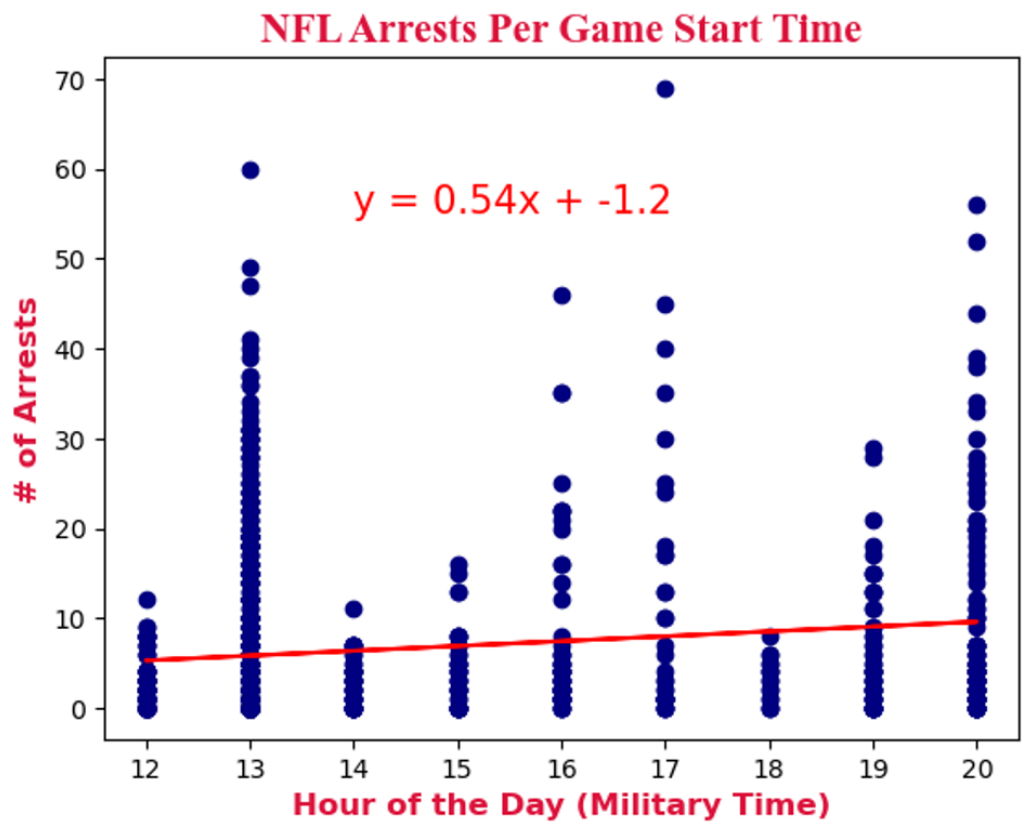
A trend we believed would be a good indicator of how many arrests would occur at an NFL game would be what time the game started. Our hypothesis before looking at the data for insights was that the later the game the more arrests that would occur. This hypothesis stemmed from the belief the longer people are tailgating throughout the day of the game and the more they are drinking alcohol the more wild they would be.

To begin looking at this insight we created a date-time data type column by concatenating the date, year and game time columns within the combined data set and then using a “to_datetime” function to change the string to a date-time data type for better data analysis. After creating this new DateTime column, we extracted the hour of the date-time value using the “dt.hour” function and created another column with the hour of the day for each game. *Refer to the exhibit below.*

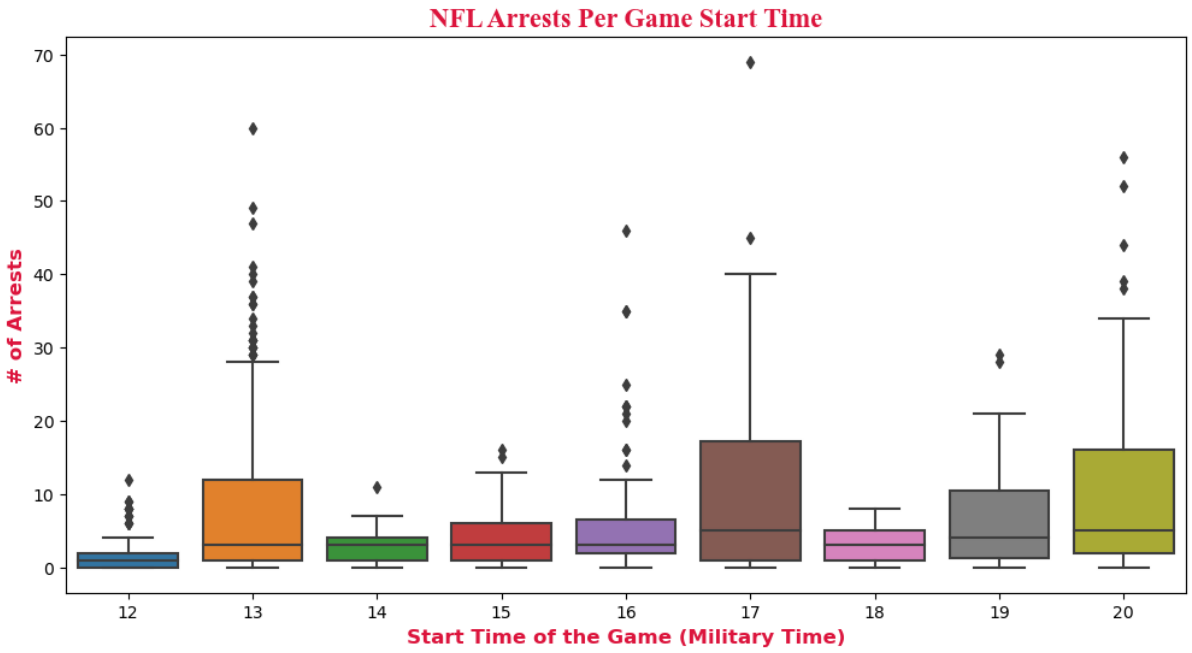
In [57]:
```
# create date and time coulmn of each game
combined_df["DateTime"] = combined_df.date + ", " + combined_df.year_x.astype(str) + " "+ combined_df.gametime_l
combined_df["DateTime"] = pd.to_datetime(combined_df.DateTime)
```

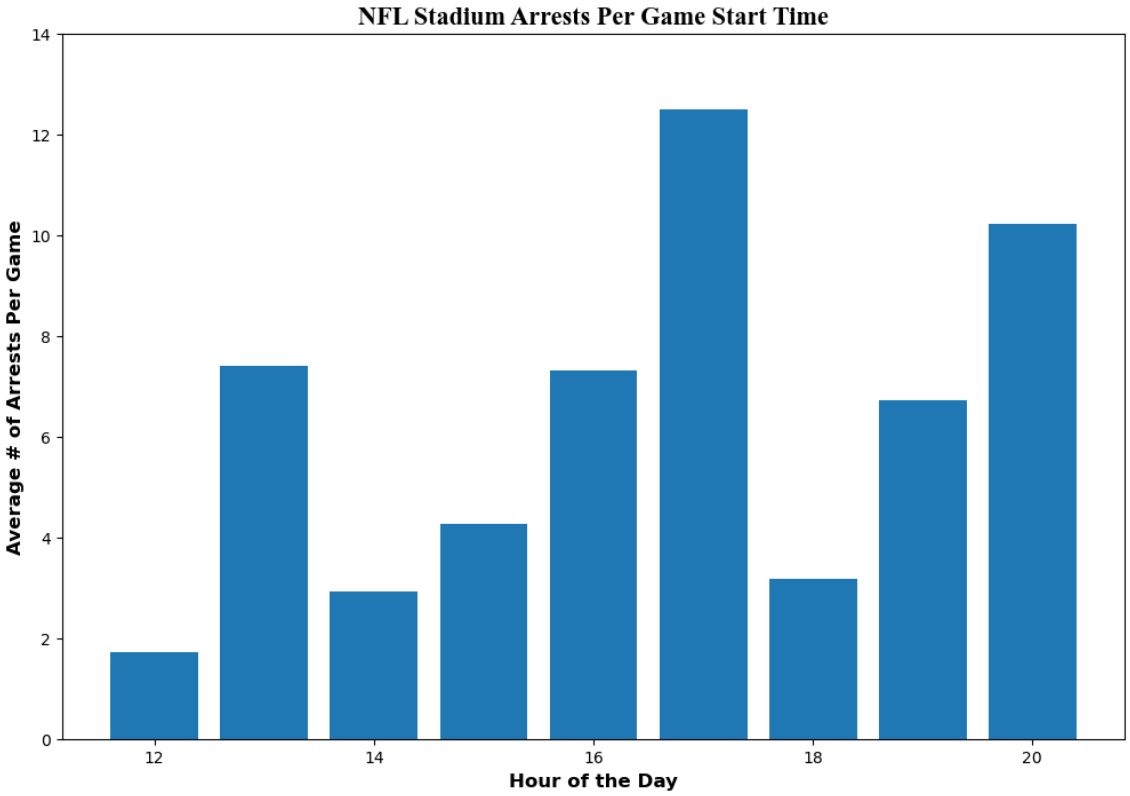
In [58]:
```
# create hour of day column
combined_df["hour"] = combined_df.DateTime.dt.hour
```

With this newly created column we created a scatter plot using the newly created ‘hour’ column along with the number of arrests that game to visualize the trend. Additionally, we found the correlation to see if there was a statistically proven trend and we got an r-value of about 0.147 which is a very weak correlation, and the regression line to see if we could predict the arrests in a game given the hour of the game start time with a formula. Unfortunately, our regression line is not a good predictor of how many arrests will be in a game. *Refer to the exhibit below.*


NFL Arrests Per Game Start Time
y = 0.54x + -1.2
# of Arrests
Hour of the Day (Military Time)
0
10
20
30
40
50
60
70
12
13
14
15
16
17
18
19
20

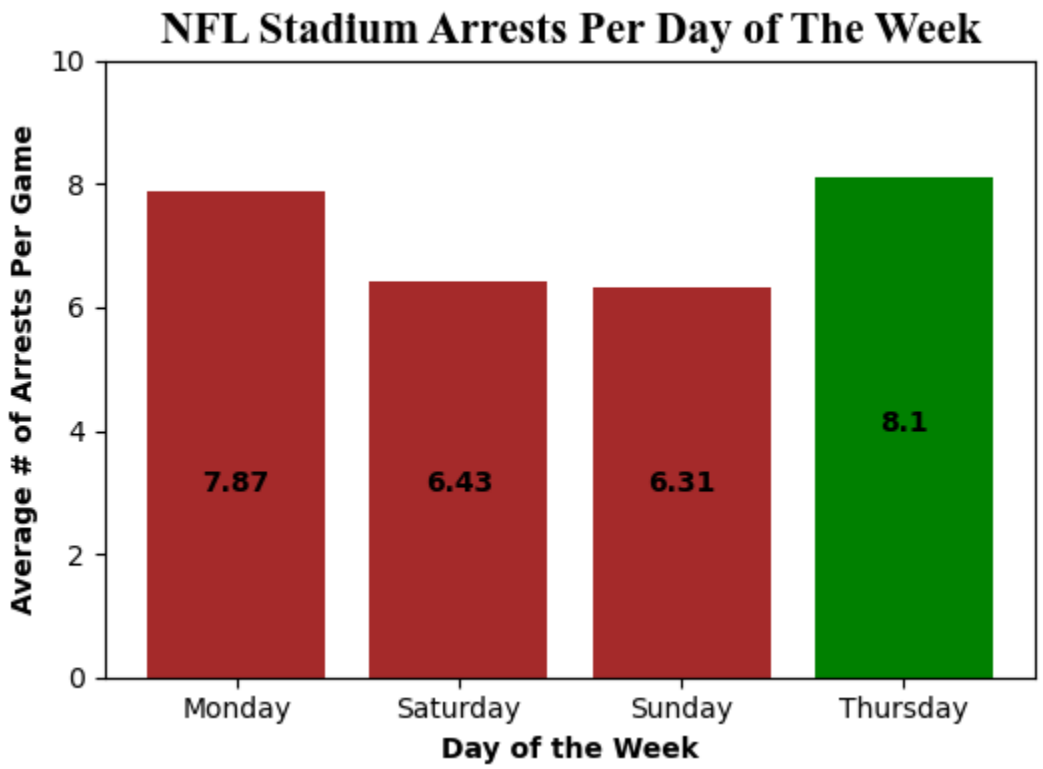
However, we additionally created a box plot and a bar graph to see what time of games had the most arrest. As you can see from the graphs created below, 17:00 or 5:00 pm had the most arrests with an average of 12.5 arrests. We believe this is the ‘sweet’ spot of the day when it’s not too early, but not too late so people are at their peak.

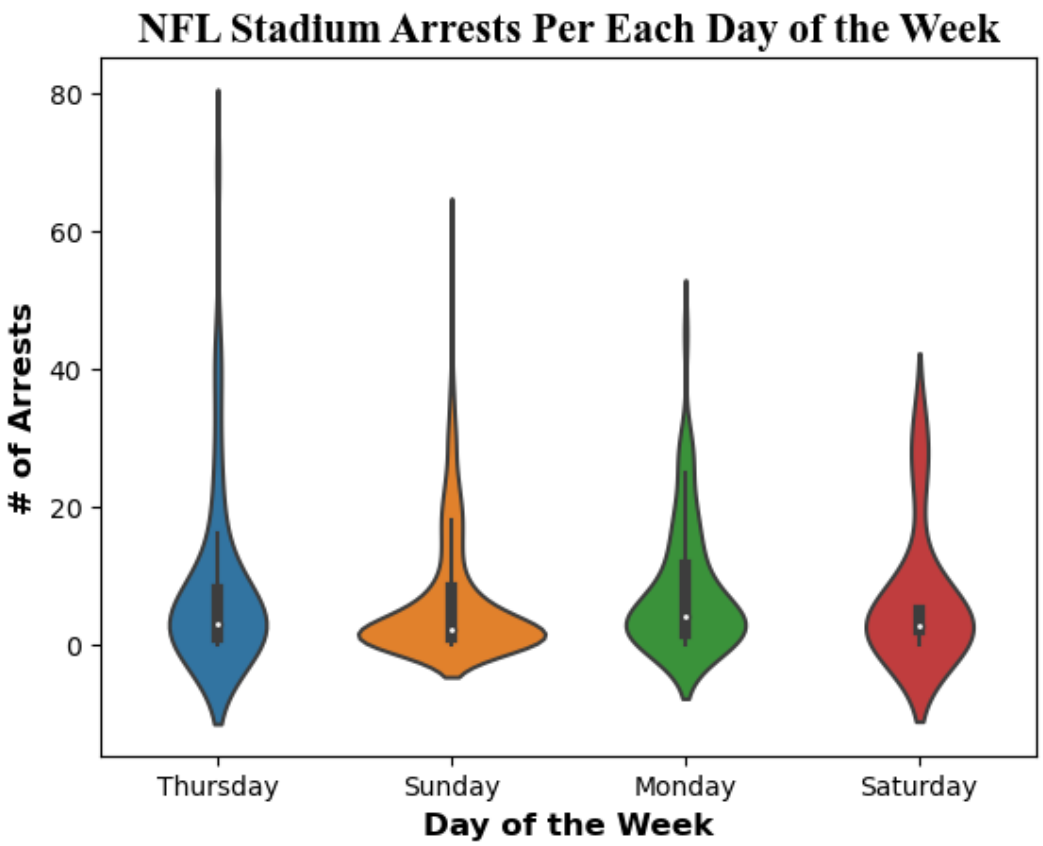




Game Day-Arrests Trend

In addition to game day time, we thought there may be a trend of arrests and the day the game took place. Our hypothesis, on average, there would be more arrests on weekend games (Saturdays and Sundays) as less people didn't have work that day, so maybe they were enjoying themselves more and would be behaving more reckless. In order to visualize this trend we created a bar chart and a violin plot of the average arrest per game for each day. As the graphs show, Thursday games had the highest average, however, not by very much. We had to remove one Wednesday game from the dataset that occurred in 2012. This game was originally scheduled as a Thursday night game, however, due to the democratic national convention being on that day the NFL rescheduled the game to the Wednesday night before (this game was between the NY Giants and Dallas Cowboys and had 39 arrests).





In order to statistically prove there was no significant difference between the day of the week the game took place and the average number of arrests, we conducted an ANOVA test. The p-value retired from the ANOVA test shows that there is no significant difference in the day of the game and the average number of arrests. It is to be considered that the majority of NFL games are played on Sunday, and, as such, our data set contained a significant difference in population of Sunday games versus any other day of the week. If redone, a sampling method of equal amount of data from each day of the week may be a preferred, less biased method. *Refer to the exhibit below.*

In [175]:
```
thurs = day_of_week_df.loc[day_of_week_df.day_of_week == "Thursday", "arrests"]
print(thurs.mean())
print(thurs.var())
```

```
8.10344827586207
171.21718088324258
```

In [176]:
```
sun = day_of_week_df.loc[day_of_week_df.day_of_week == "Sunday", "arrests"]
print(sun.mean())
print(sun.var())
```

```
6.3103030303030305
84.26524566048894
```

In [177]:
```
sat = day_of_week_df.loc[day_of_week_df.day_of_week == "Saturday", "arrests"]
print(sat.mean())
print(sat.var())
```

```
6.428571428571429
91.03296703296702
```

In [178]:
```
mon = day_of_week_df.loc[day_of_week_df.day_of_week == "Monday", "arrests"]
print(mon.mean())
print(mon.var())
```

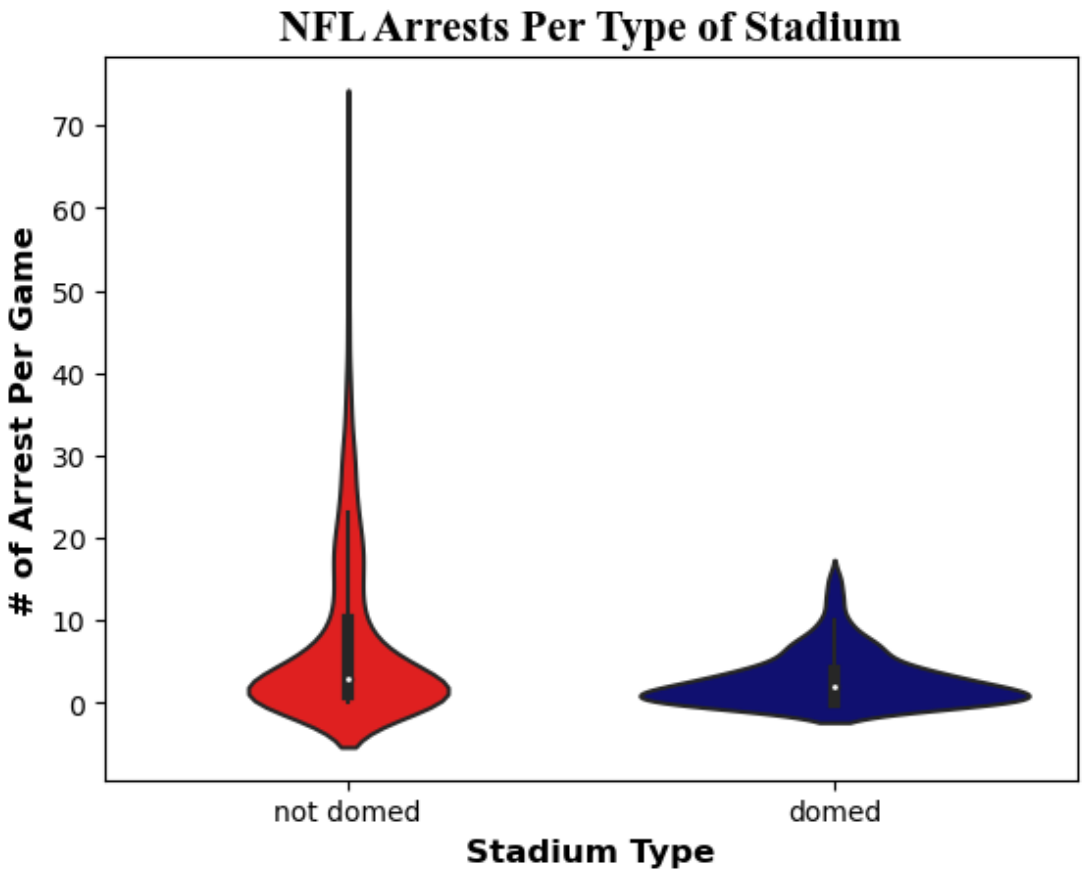
```
7.865671641791045
84.23925825418361
```

In [179]:
```
stats.f_oneway(thurs, sun, sat, mon)
```

Out[179]:
```
F_onewayResult(statistic=1.1285731867432665, pvalue=0.336456484377337)
```

Stadium Type-Arrests Trend

A new direction for the NFL as new stadiums are built for teams are more domed or retractable roof stadiums. In total there are 10 out of 32 total teams who currently have a domed or retractable roof stadium. However, due to the time of our data being gathered and our data set not including arrests data for 7 NFL teams, our dataset contained 5 teams with a domed or retractable roof stadium (refer to ‘Reference’ section for source link). For these 5 teams, we decided if we could find a significant difference in the number of arrests between an open NFL stadium and a stadium that has a roof. Our hypothesis was that when exposed to the elements people might be easier to upset and may cause more problems. To first visualize this we created a violin plot to visualize the difference of the two.



From the visualization, it can be seen that teams with domed stadiums, generally, have less arrests than stadiums that are open-roofed. The domed stadiums have about 3 arrests per game while open-roofed stadiums have about 7 arrests per game. We completed a T-test to prove the significant difference and found a p-value of about 5.04 e-22. This shows there is a significant difference in arrests between the two as the p-value is drastically less than 0.05. *Refer to the exhibit below.* Generally, domed stadiums are newer and recently built, and this raises the question: does the environment of the game such as the stadium amenities or comfortability of seats have an impact on fan behavior which will determine the number of arrests that occur during a game? For example, when at a 5 star restaurant you are bound to act differently than when you are at a sports bar. Should teams decide if a new stadium is worth it? Are there enough data driven trends that can prove it is a good investment?

```
In [220]:
1 domed_arr = combined_df.loc[combined_df.type_of_stadium == "domed", "arrests"]
2 print(domed_arr.mean())
3 print(domed_arr.var())
```

```
2.9125
10.44512578616351
```

```
In [221]:
1 notdomed_arr = combined_df.loc[combined_df.type_of_stadium == "not domed", "arrests"]
2 print(notdomed_arr.mean())
3 print(notdomed_arr.var())
```

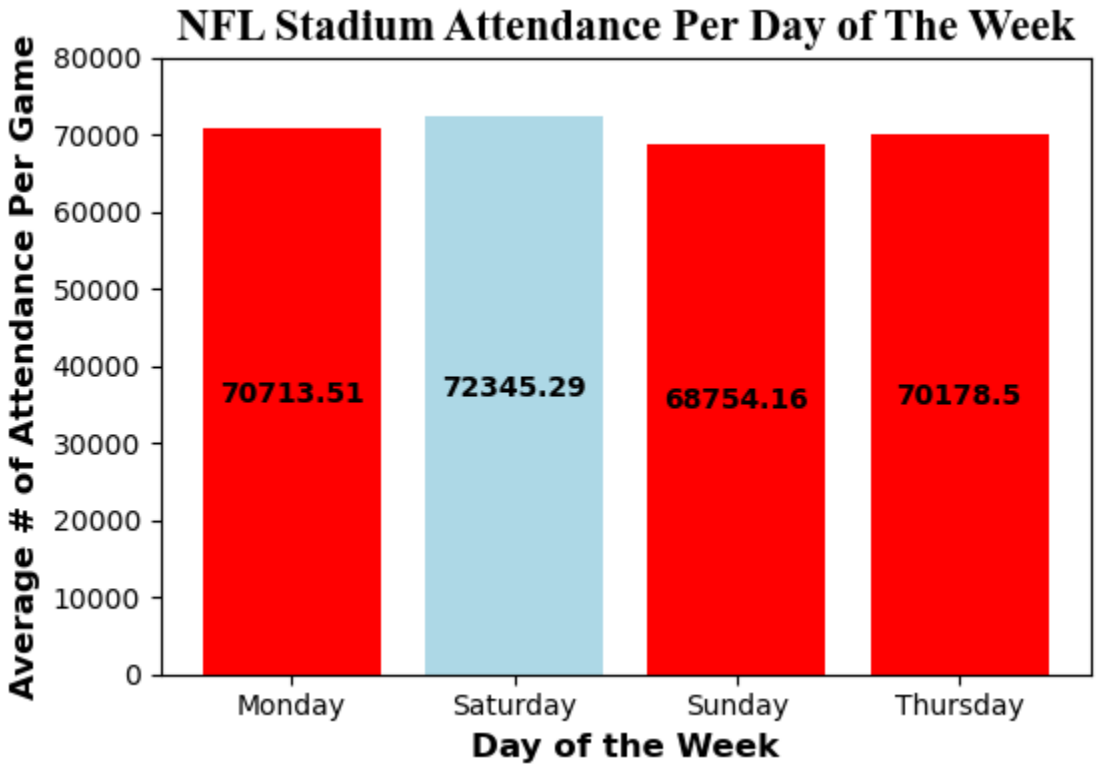
```
7.28695652173913
103.33173264114231
```

```
In [241]:
1 stats.ttest_ind(domed_arr, notdomed_arr, equal_var=False)
```

```
Out[241]: TtestResult(statistic=-9.940805850423384, pvalue=5.04482343911883e-22, df=792.8219330088324)
```

NFL Game Attendance by Day of the Week

As previously mentioned, there were other insights within our data that we found interesting and decided to take a look at. An additional insight we decided to look into was if the day of the week is a factor in the attendance of an NFL game. To visualize this we created a bar graph to view the average attendance of NFL games on each day of the week.



From the bar graph, it can be seen that Saturday had the highest average of attendance. However, it is to be noted that there is a very small population of Saturday NFL games when compared to the other days. When conducting a T-test between Saturday and Sunday to see if there was a significant difference and there was a p-value = 0.4998. As such, there is no significant difference between Saturday and Sunday.

```
In [201]: 1 stats.ttest_ind(mon_att, sat_att, equal_var=False)
Out[201]: TtestResult(statistic=-0.688426040243472, pvalue=0.4998006763724756, df=18.34026076548382)
```

However, we conducted a T-test between Monday and Sunday games and received a p-value of about 0.052. *Refer to the exhibit below.* As such, there is a significant difference in game attendance between Monday and Sunday games. As previously mentioned, the majority of NFL games are played on Sunday, so the data may have biased numbers that sway the data. This could be adjusted by using a sampling method to take a representative sample of games from each day of the week. Additionally, it may be better to look at this at a per team lens instead of the entire NFL because some teams have better attendance overall or a larger stadium capacity and it would be a more insightful conclusion to see the day to day comparison of attendance of each individual team.

```
In [200]: 1 # There is a significant difference between Sunday and Monday game attendance
          2 stats.ttest_ind(sun_att, mon_att, equal_var=False)
Out[200]: TtestResult(statistic=-1.969033735797458, pvalue=0.05245981282696772, df=78.90832934754886)
```

Limitations and Bias

Before any conclusions can be made from our work, there are few limitations and biases that need to be considered. The most influential limitation being that we were missing data for 7 NFL teams. Two of these teams missing included the Detroit Lions and Atlanta Falcons which due to the crime rates of the respective home cities we could expect there to be a “good” sum of arrests at these stadiums. A mistake we made was, for research questions not involving arrests, we could have created another dataset which only combined NFL Attendance and NFL Games. This would have allowed us to use a completed 32 team dataset along with a larger population (2000-2019), and include more recent data. Additionally, our data is a relatively small population of NFL data. We only looked at data from 2011-2015. The NFL has been around since about 1920. Additionally, our data is not up-to-date. Since 2015, 3 teams have changed home cities and there have been 4 new stadiums opened.

The strongest bias to our research question is that arrests at NFL stadiums have bias with each respective city’s crime rate. Some additional work that could be done in the future could be to look at a city's crime rate correlation to average arrests. For stadium attendance based research questions, there is a bias in stadium total capacity. Some stadiums are larger than others so total attendance may not be a fair parameter. For future work, I would recommend using stadium attendance percentage (total attendance/stadium capacity) for making data-driven conclusions on game attendance. For finding out who has the best home advantage, some teams are just better than others, so they win more home games because they just win more games in general. Lastly, there are more games played on Sunday at 1:00pm than any other day or time. For all “day of the week” research questions our conclusions will have to factor in that Sunday had a drastically larger population of games than any other day of the week. If redone I would suggest using a sampling method to have an equal representation of games that occur on each day. Similarly, for “time of day” research questions and 1:00pm.

Takeaways

A key takeaway from our project is 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. There probably is more correlation with city demographics or crime rate with the number of arrests. Additionally, we looked at it from a league-wide lens. There might be stronger correlations and more evident trends at a per team lens.

We found the San Diego Chargers, on average, had the most arrests with about 25 arrests per game. The Chargers relocated to Los Angeles in 2017. This data driven decision could have been one of the reasons for the move. It would be interesting work moving forward to try and find why the arrests for the San Diego Chargers was so high. In fact, the San Diego Chargers had the most arrests in a game within our dataset when they played the Oakland Raiders in San Diego in 2011 Week 10. There were a total of 69 arrests on this day!

Domed stadiums from our small sample show they, generally, have less arrests. This could be a helpful insight for NFL teams deciding if a new stadium is a good investment. A research question to take a deeper dive into is if the quality of the environment is a factor in fan behavior. If seats are more comfortable are fans more relaxed and less likely to cause trouble? When there are nicer amenities within the stadium does this have any effect? A good graph to visualize this may be a scatter plot of stadium opening year vs. average number of arrests during a game at that stadium.

Moving Forward

Within our data there were lots of research questions that could be answered and lots of data that could be added to create a better story. If the data is available, you could look for correlation of how many police were on site to the arrests per game. Are there more arrests because there are more police to catch people or are people more cautious when they see greater police presence? Additionally, we could look at alcohol sales for each game and see if there is a correlation with arrests per game. Are fans more reckless due to alcohol consumption?