Analyzing Baseball Fans and Predicting Game Attendance

Troy Hepper
DC-DSI 4
May 25, 2017

AGENDA

- Introduction
- Exploring the Data
- Building Models
- Bayesian Inference
- Further Analysis

PROBLEM STATEMENT

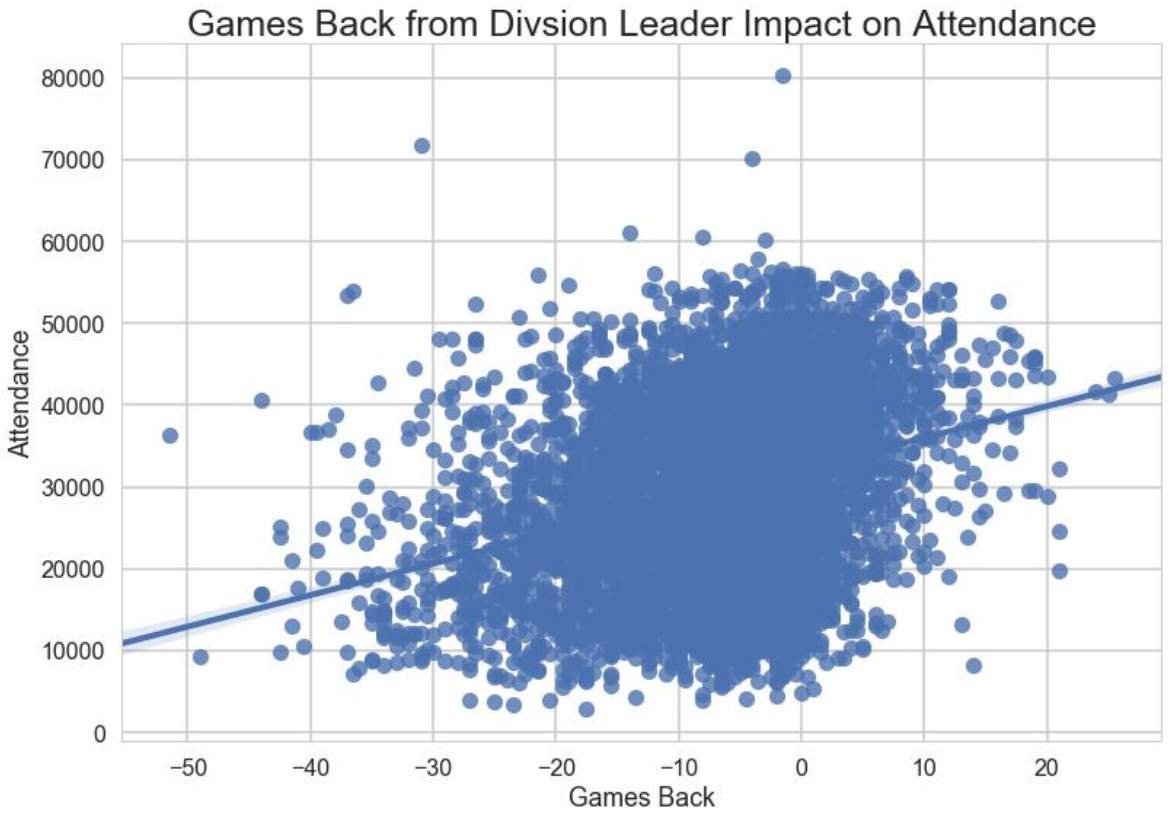
Predicting Major League Baseball game attendance based on historical data.

COLLECTING DATA

- Collected game data from baseball-reference.com for all MLB games from 1990 to 2016
- Engineered features that I thought might have an affect on attendance
- Also collected season summary information for each team

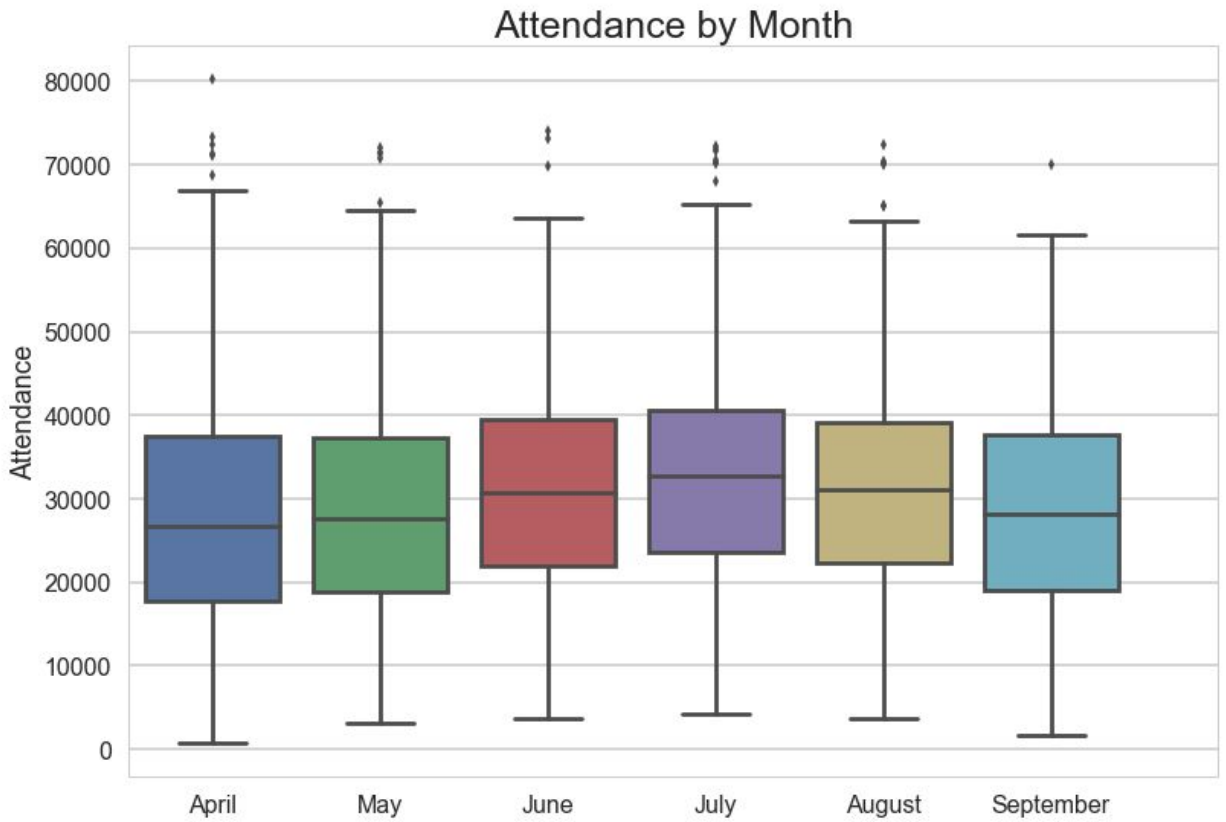
EXPLORING THE DATA

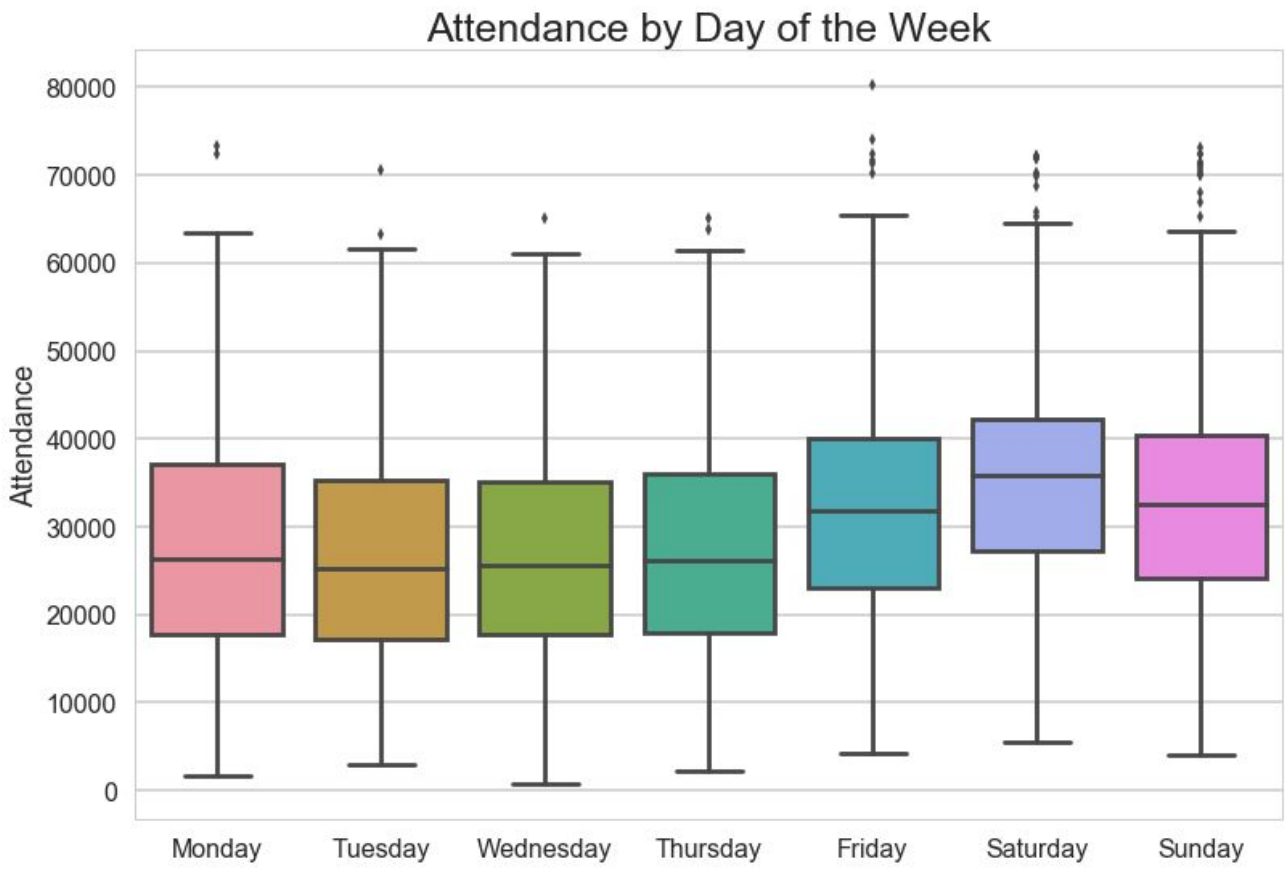
Team performance is directly correlated with fan attendance



EXPLORING THE DATA

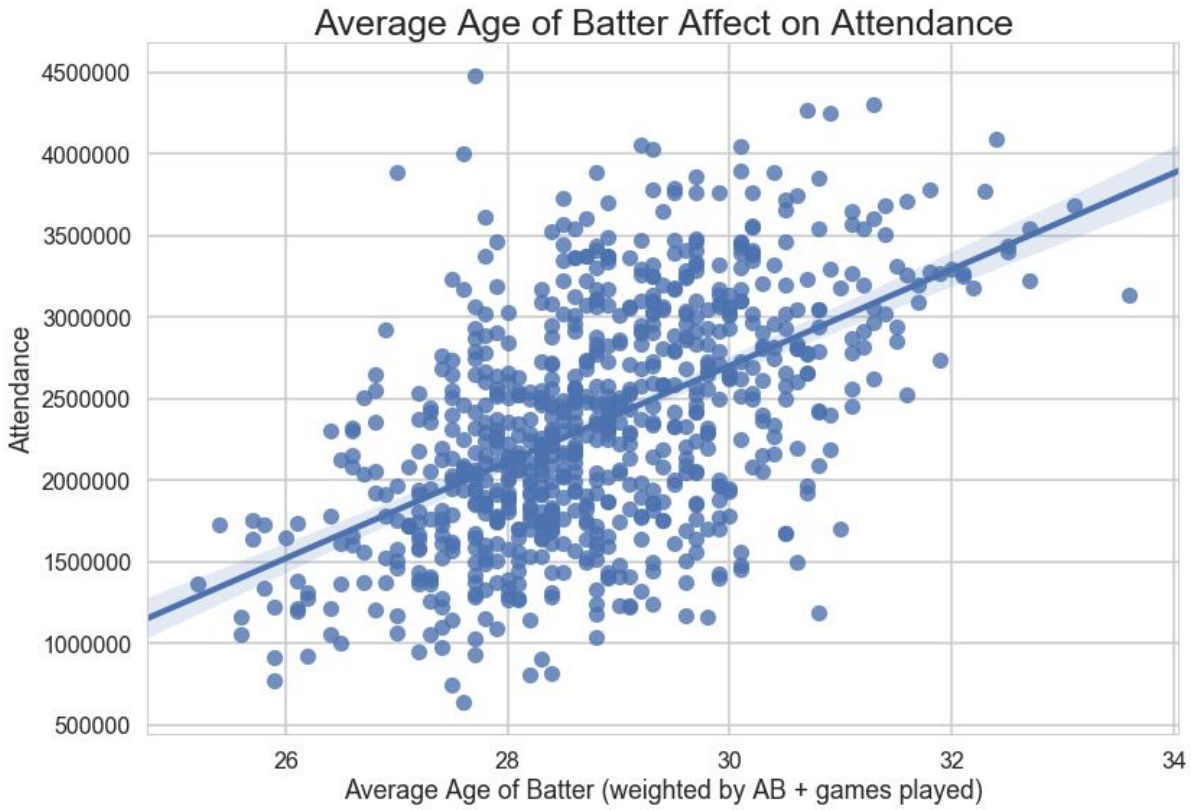
Time of year and day of the week certainly affects fan turnout




Attendance by Day of the Week
Attendance
80000
70000
60000
50000
40000
30000
20000
10000
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

LESS OBVIOUS FACTORS

The yearly summary data showed that the average age of the players on the team were highly correlated with attendance

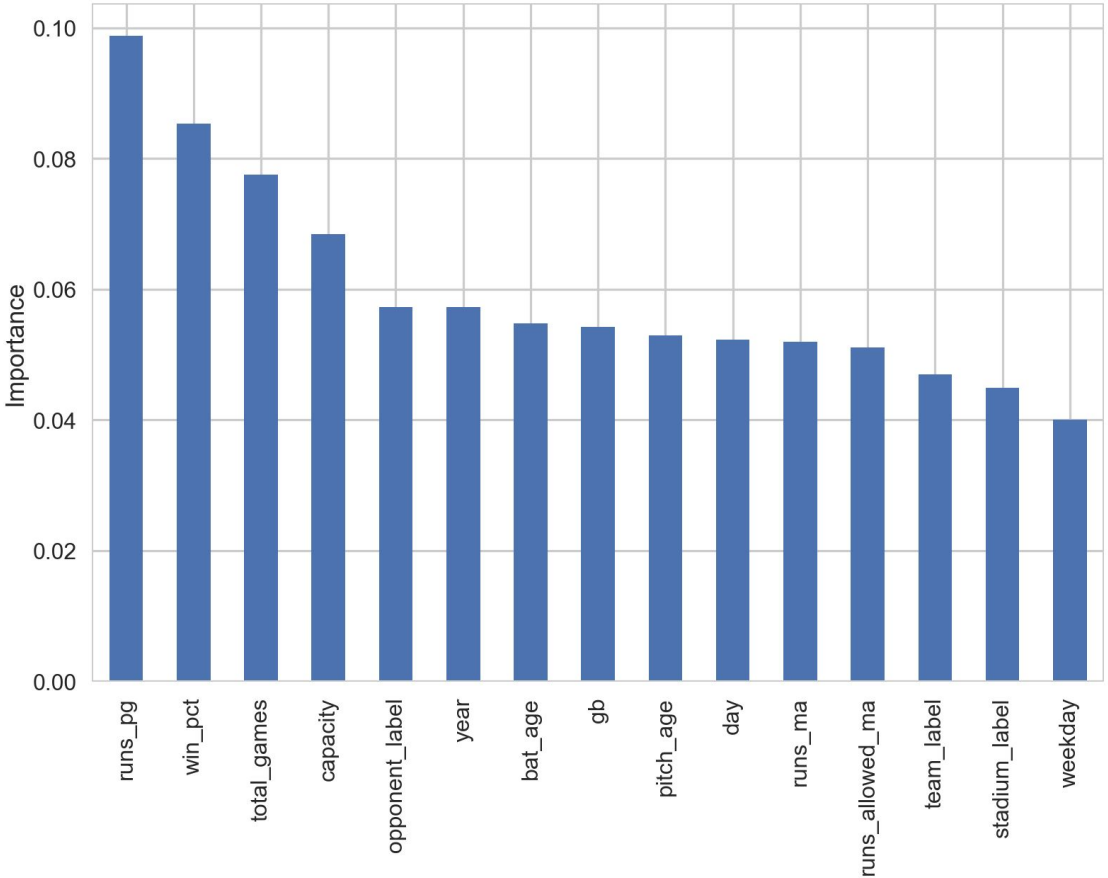


BUILDING MODELS

0.833

The R^2 value of my Gradient Boosting Regression model, which represents how close the actual attendance values are to the fitted regression line.

FEATURE IMPORTANCES



APPLYING THE RESULTS

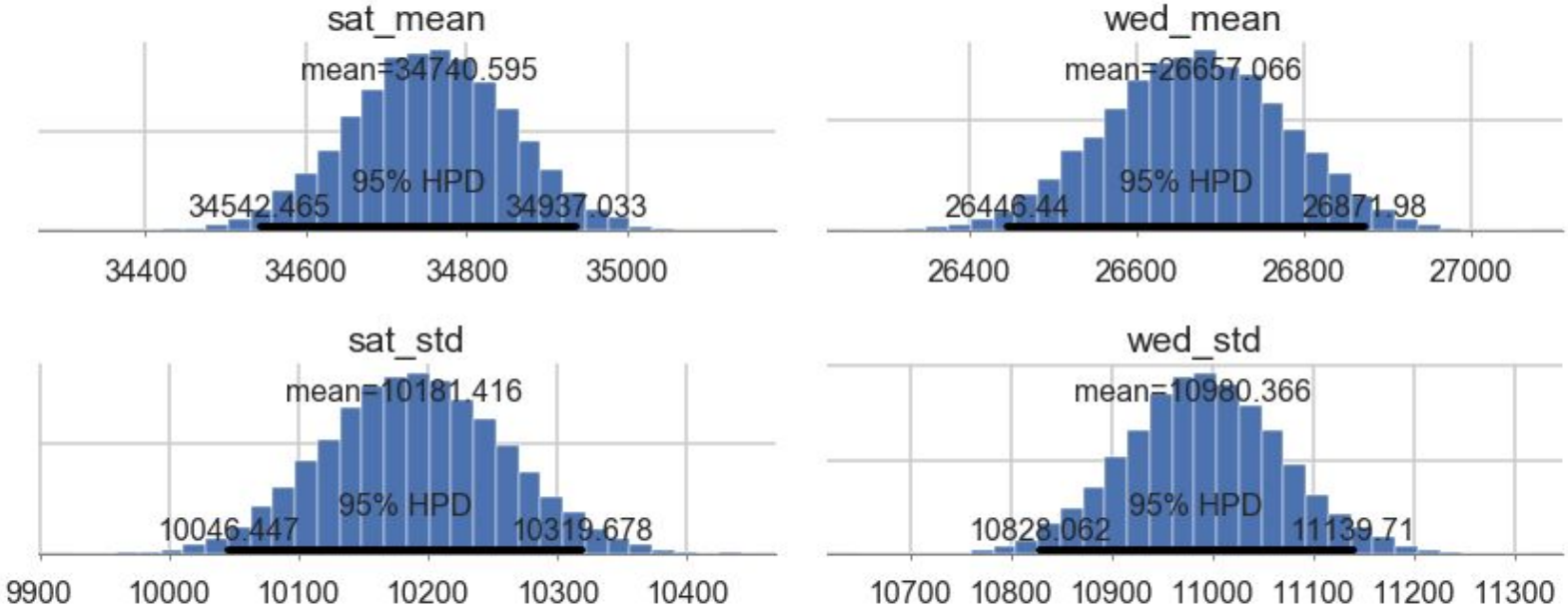
A team may find these results useful in order to develop effective marketing/promotional strategies

BAYESIAN INFERENCE

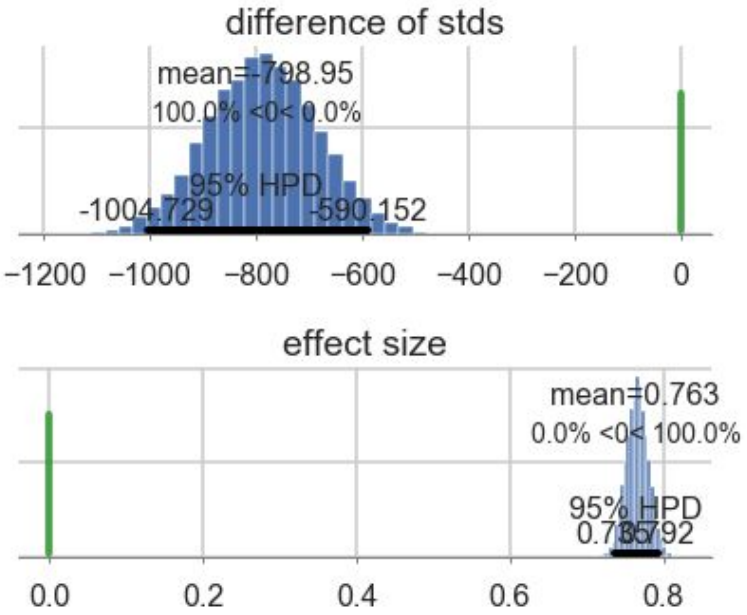
COMPARING TWO DIFFERENT DAYS OF THE WEEK

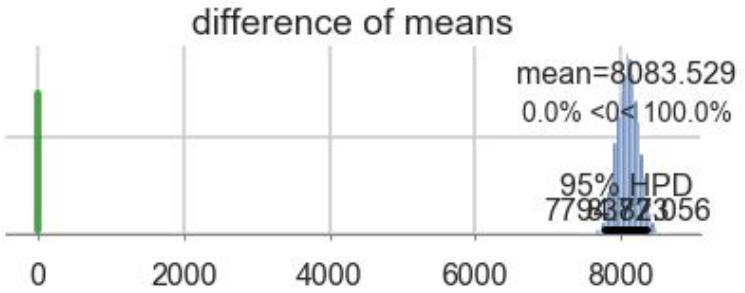
- I decided to compare the average attendance between Wednesday and Saturday games
- As a prior, I used all other days of the week

PLOTTING POSTERIOR DISTRIBUTIONS



PLOTTING POSTERIOR DISTRIBUTIONS





COMPARING TWO FAN GROUPS

Washington Nationals

Stadium: Nationals Park

Opened: 2008

Capacity: 41,500

Avg Win Pct: 0.498

New York Mets

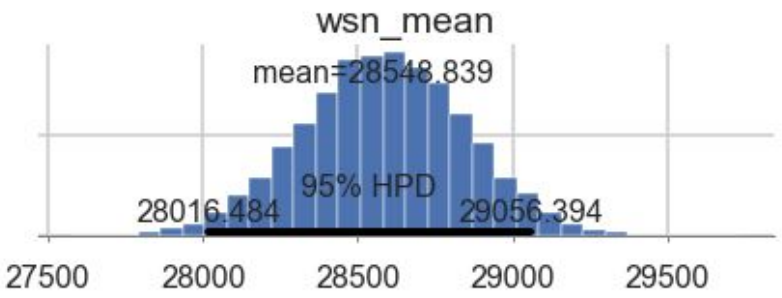
Stadium: Citi Field

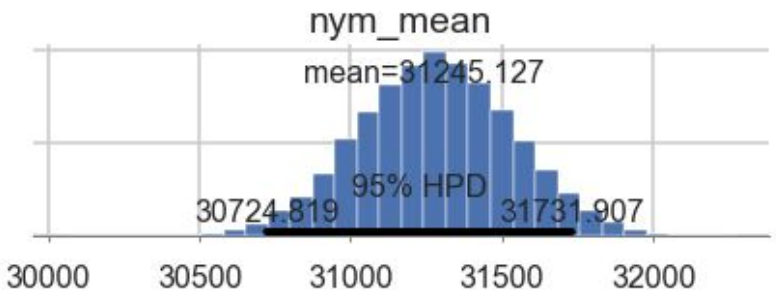
Opened: 2009

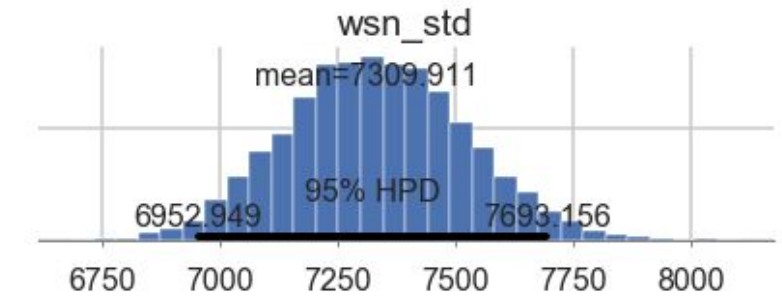
Capacity: 41,800

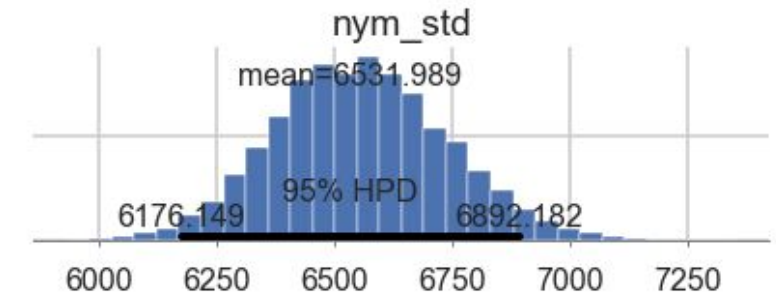
Avg Win Pct: 0.486

PLOTTING POSTERIOR DISTRIBUTIONS

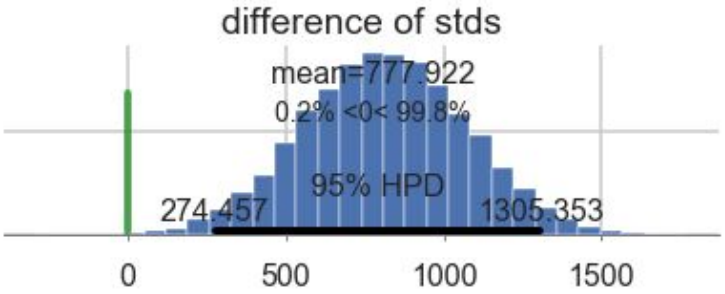


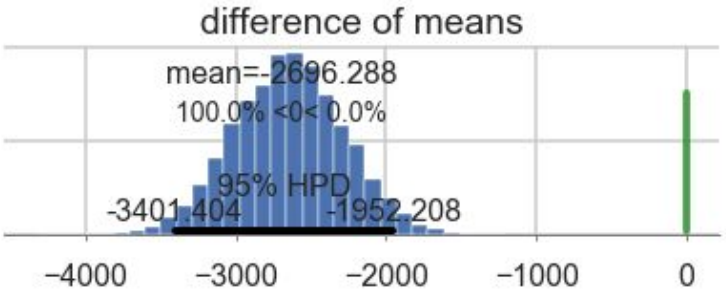


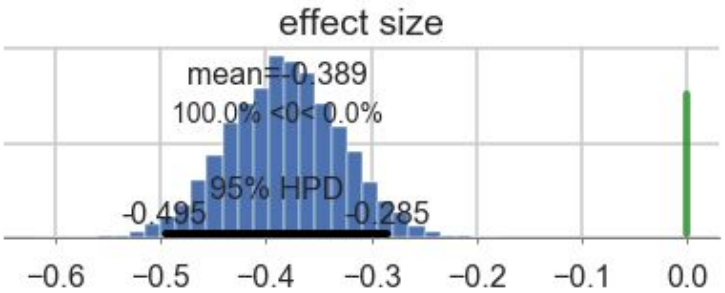




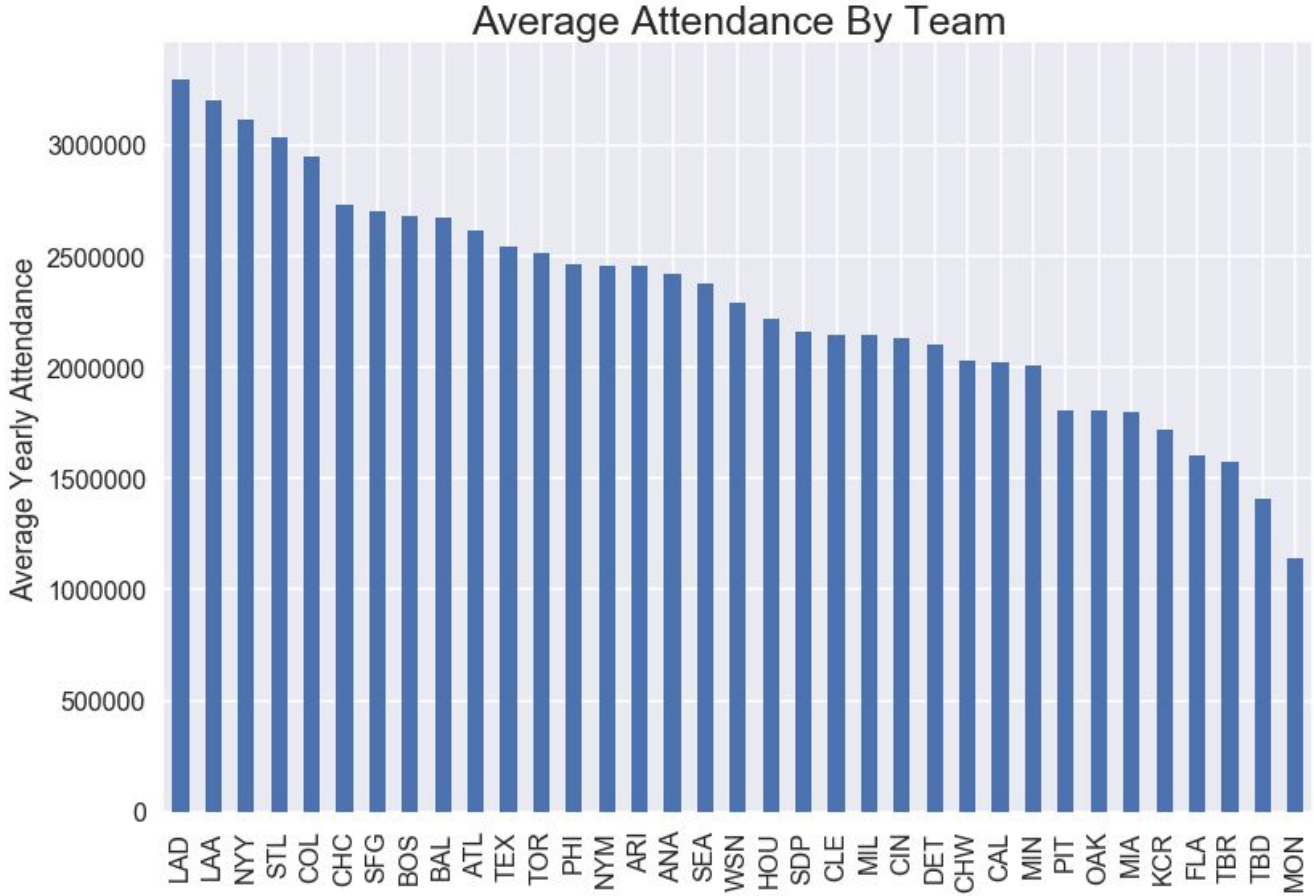
PLOTTING POSTERIOR DISTRIBUTIONS

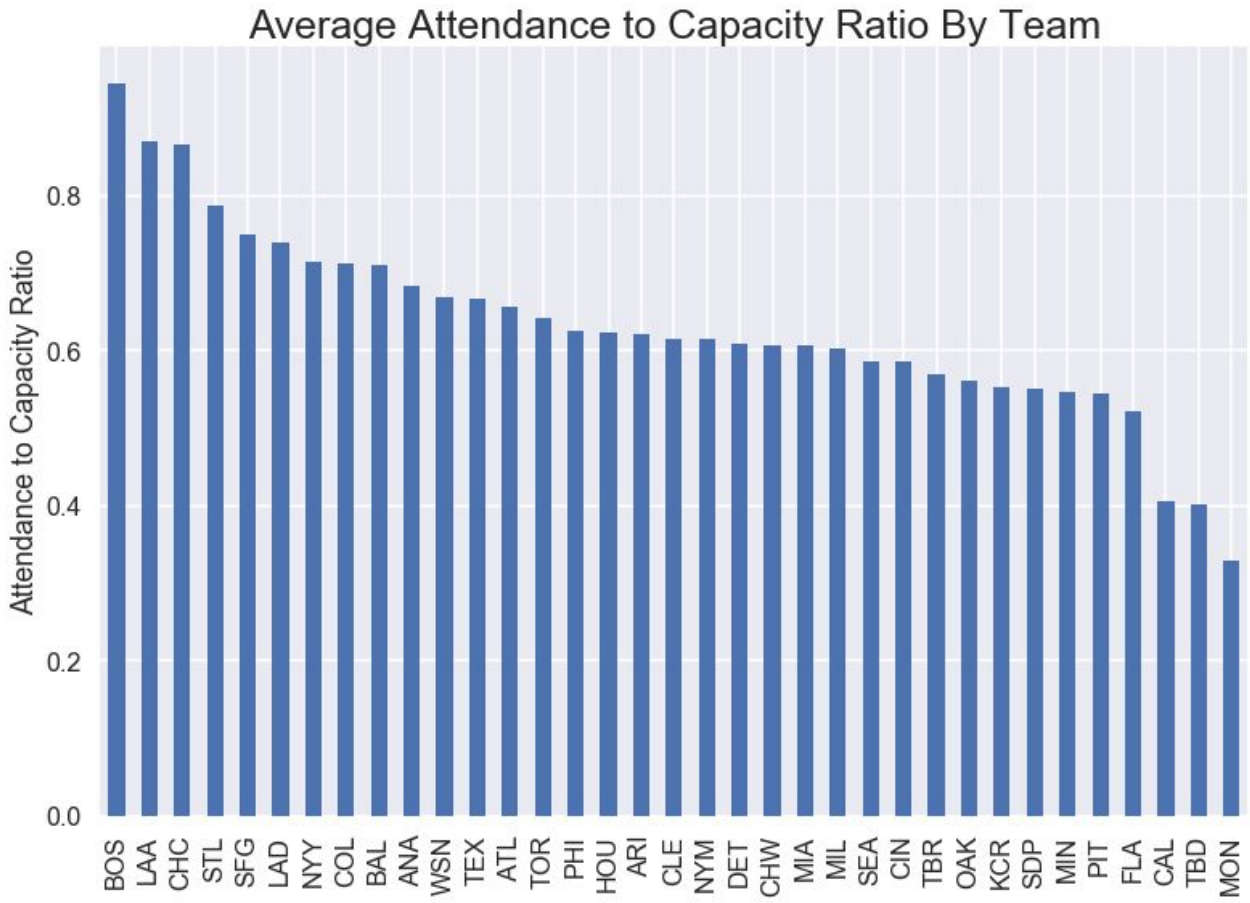


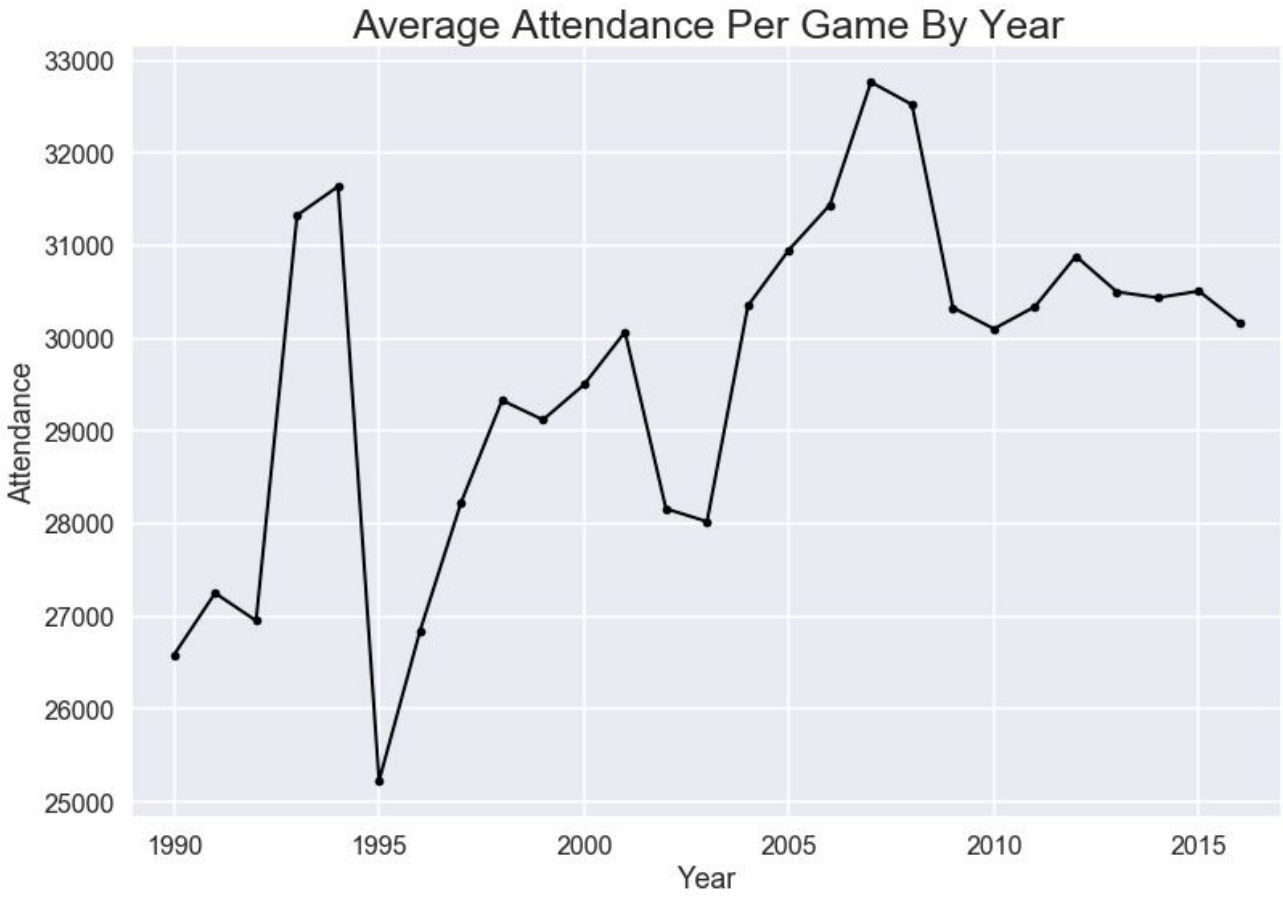


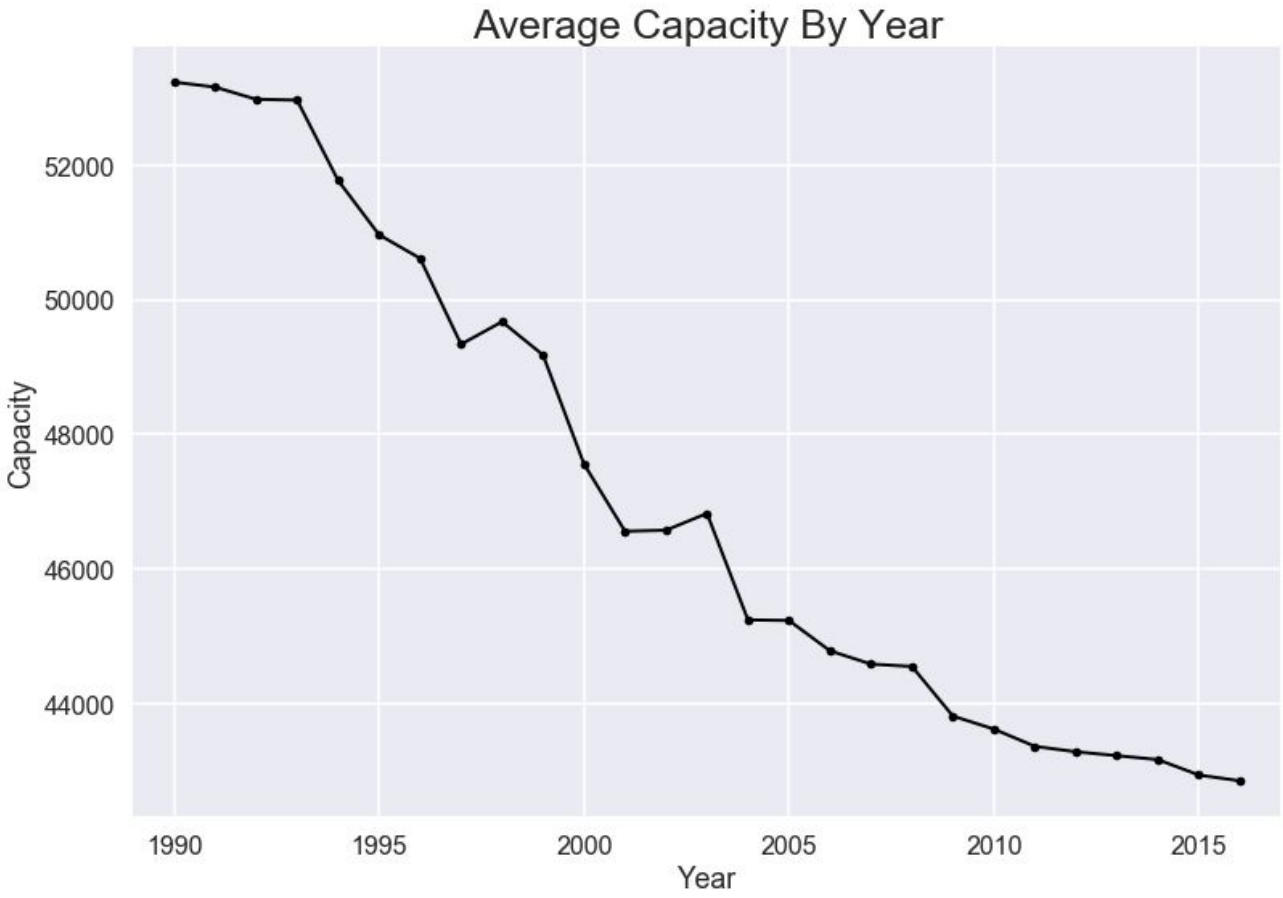


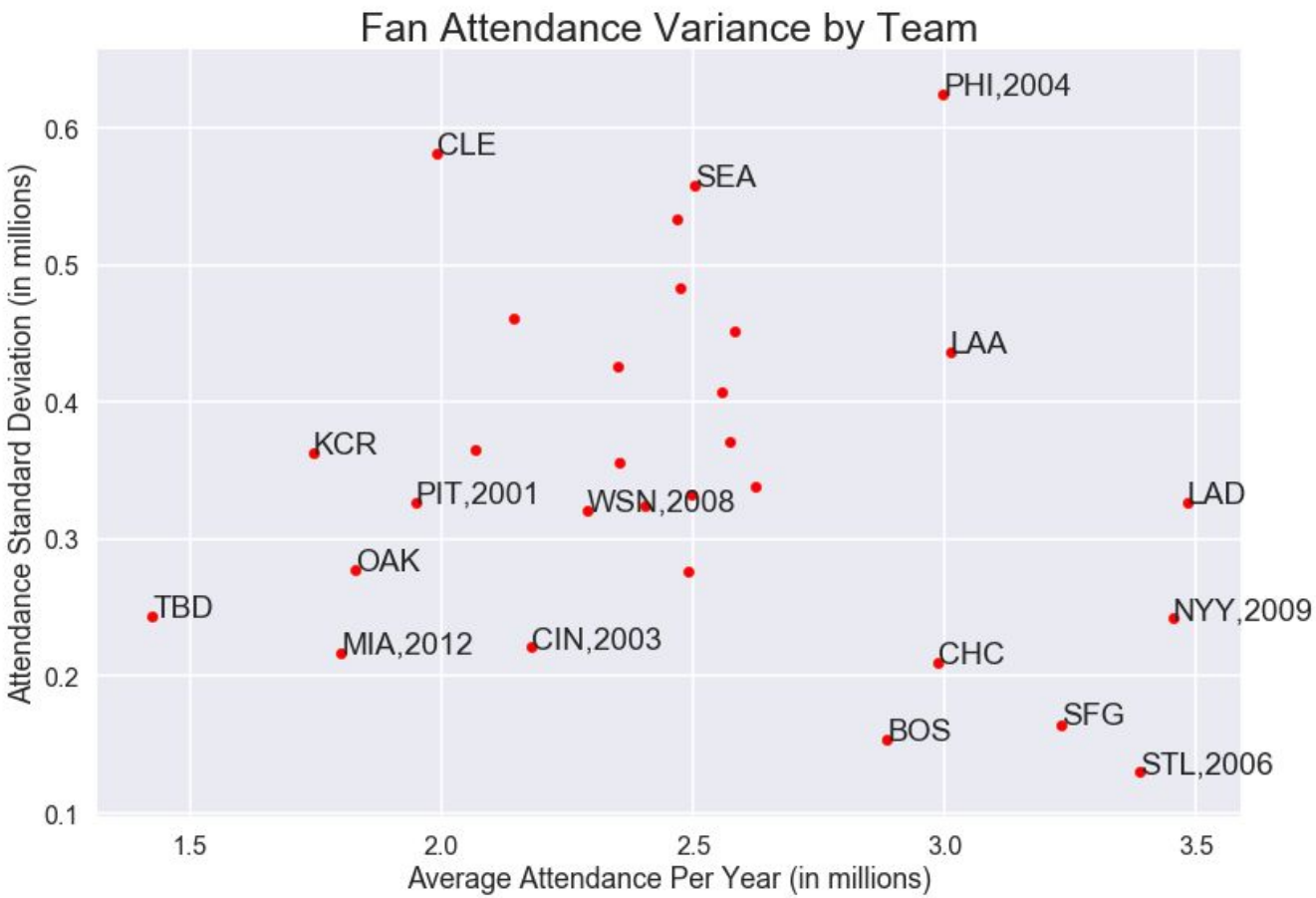
FURTHER ANALYSIS

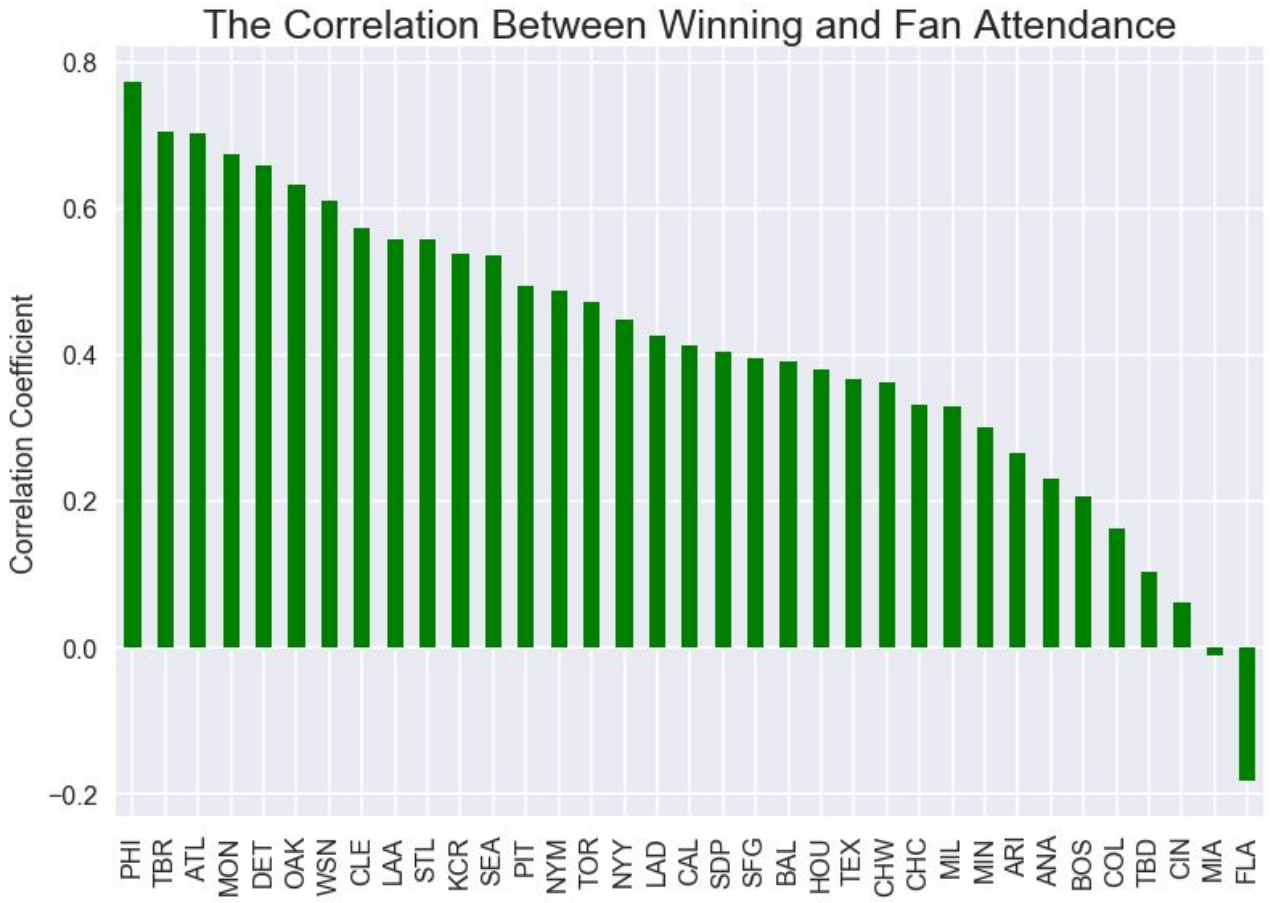

Average Attendance By Team
Average Yearly Attendance
3000000
2500000
2000000
1500000
1000000
500000
0
LAD
LAA
NYY
STL
COL
CHC
SFG
BOS
BAL
ATL
TEX
TOR
PHI
NYM
ARI
ANA
SEA
WSN
HOU
SDP
CLE
MIL
CIN
DET
CHW
CAL
MIN
PIT
OAK
MIA
KCR
FLA
TBR
TBD
MON


Average Attendance to Capacity Ratio By Team
Attendance to Capacity Ratio
0.0
0.2
0.4
0.6
0.8
BOS
LAA
CHC
STL
SFG
LAD
NYY
COL
BAL
ANA
WSN
TEX
ATL
TOR
PHI
HOU
ARI
CLE
NYM
DET
CHW
MIA
MIL
SEA
CIN
TBR
OAK
KCR
SDP
MIN
PIT
FLA
CAL
TBD
MON


Average Attendance Per Game By Year
33000
32000
31000
30000
29000
28000
27000
26000
25000
Attendance
1990
1995
2000
2005
2010
2015
Year


Average Capacity By Year
52000
50000
48000
46000
44000
Capacity
1990
1995
2000
2005
2010
2015
Year


Fan Attendance Variance by Team
PHI,2004
CLE
SEA
LAA
KCR
PIT,2001
WSN,2008
LAD
OAK
TBD
NYY,2009
MIA,2012
CIN,2003
CHC
SFG
BOS
STL,2006
0.6
0.5
0.4
0.3
0.2
0.1
1.5
2.0
2.5
3.0
3.5
Attendance Standard Deviation (in millions)
Average Attendance Per Year (in millions)


The Correlation Between Winning and Fan Attendance
Correlation Coefficient
0.8
0.6
0.4
0.2
0.0
−0.2
PHI
TBR
ATL
MON
DET
OAK
WSN
CLE
LAA
STL
KCR
SEA
PIT
NYM
TOR
NYY
LAD
CAL
SDP
SFG
BAL
HOU
TEX
CHW
CHC
MIL
MIN
ARI
ANA
BOS
COL
TBD
CIN
MIA
FLA

RECAP

QUESTIONS?