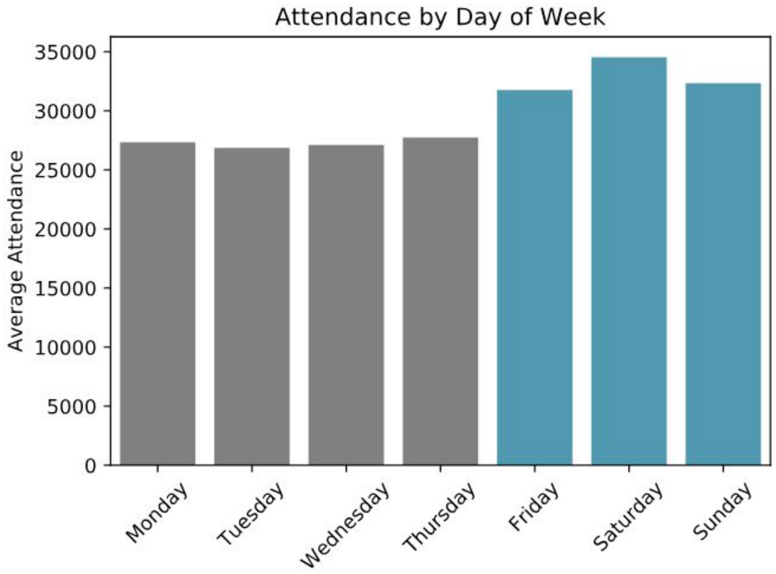
Forecasting MLB Game Attendance

Rob Pagano

Predicting Attendance - Why it Matters

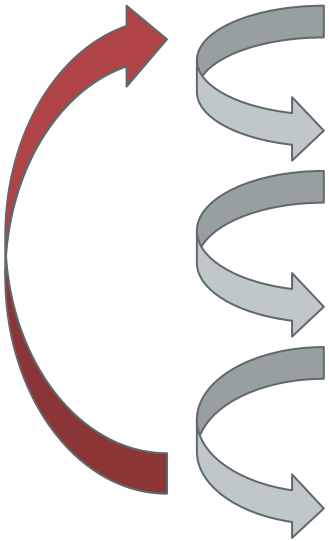
- Team marketing
- Local businesses
- Public Services

Approach



- What makes a *team* exciting
    - Baseball stats
    - Playoff contention
- What makes a *specific game* popular
    - Day of the week / Time of day
    - Opposing team
    - Promotions

Process



1. Scrape game data from 2014-2018
2. Clean data and transform features
3. Choose features for CV
4. Circle back to steps 1-3 until happy with model

Feature Editing

Original OLS

- Rank
- Games back
- Night or day
- Wins last 10
- Mean runs last 10
- Run differential

RMSE ≈ 9070.96

$R^2 \approx 0.16$

Feature Editing

Original OLS

- Rank
- Games back
- Night or day
- Wins last 10
- Mean runs last 10
- Run differential

RMSE ≈ 9070.96

$R^2 \approx 0.16$

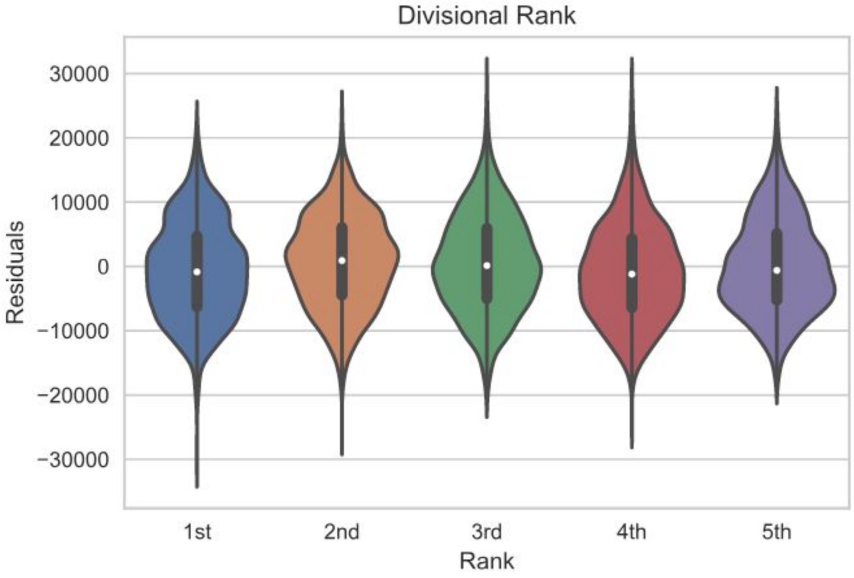
Lasso (Lamba ≈ 2.05, Standardized)

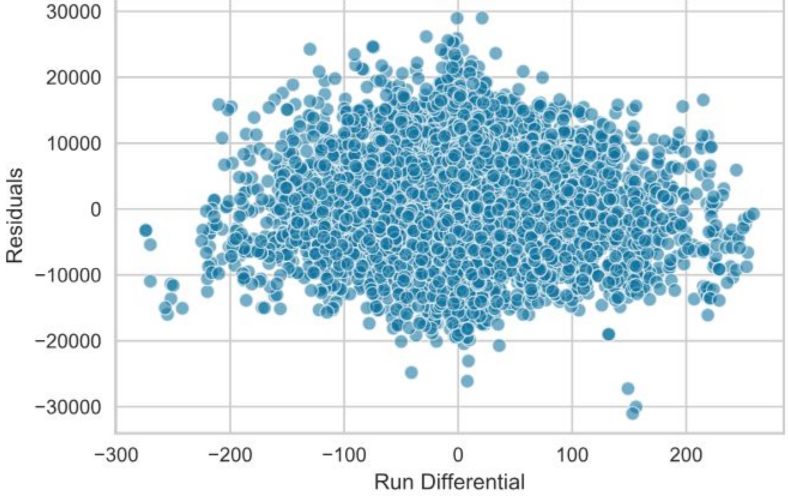
- Rank
- Games back
- Night or day
- Wins last 10
- Run differential
- Win differential
- Mean batter age
- # of current all-stars
- # of lifetime all-stars
- Team salary
- Day of week

RMSE ≈ 7434.25

$R^2 \approx 0.44$

Checking Residuals

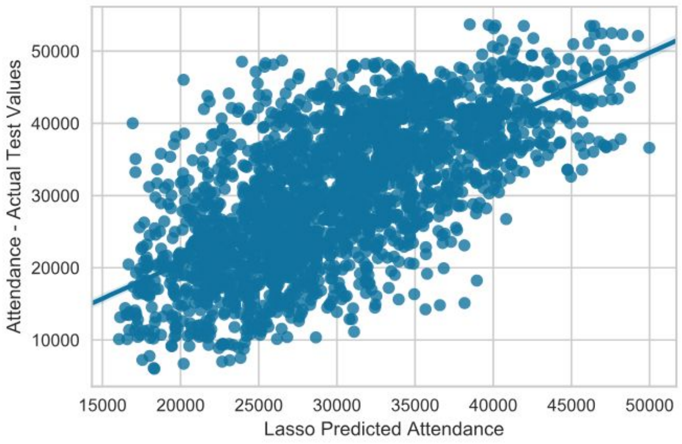


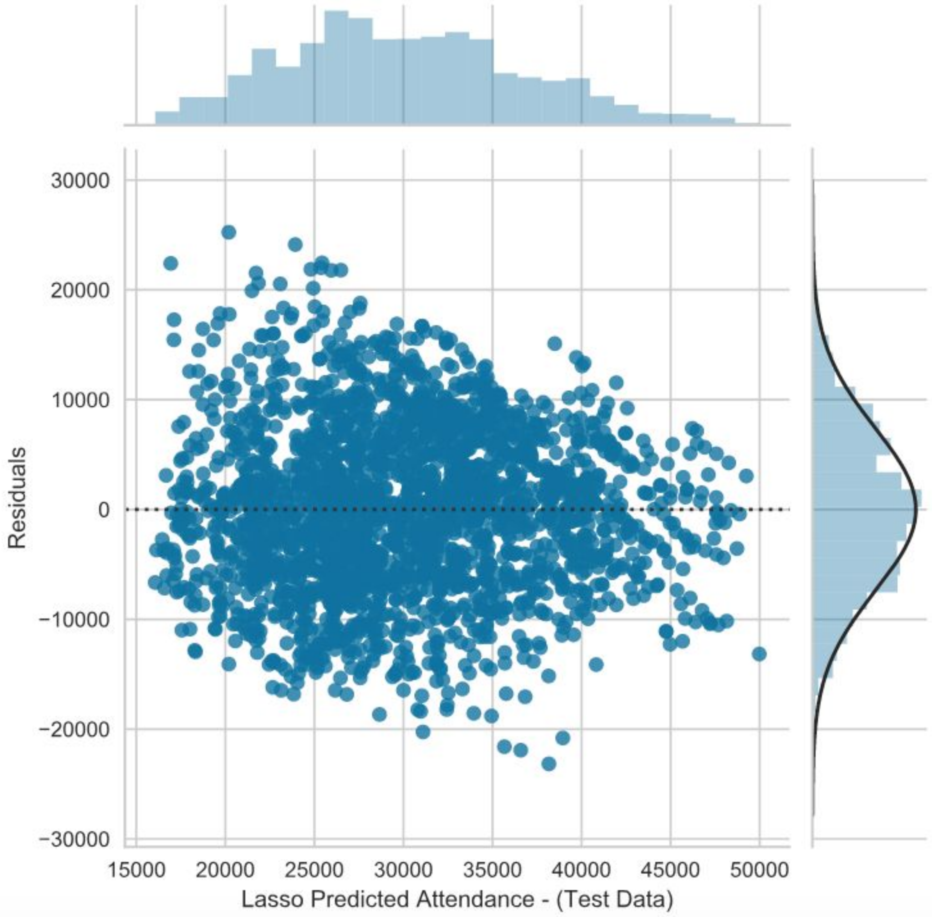


*Training Data

Outcome

- Lasso Model
  - RMSE ≈ 7124.2
  - R^2 ≈ 0.434



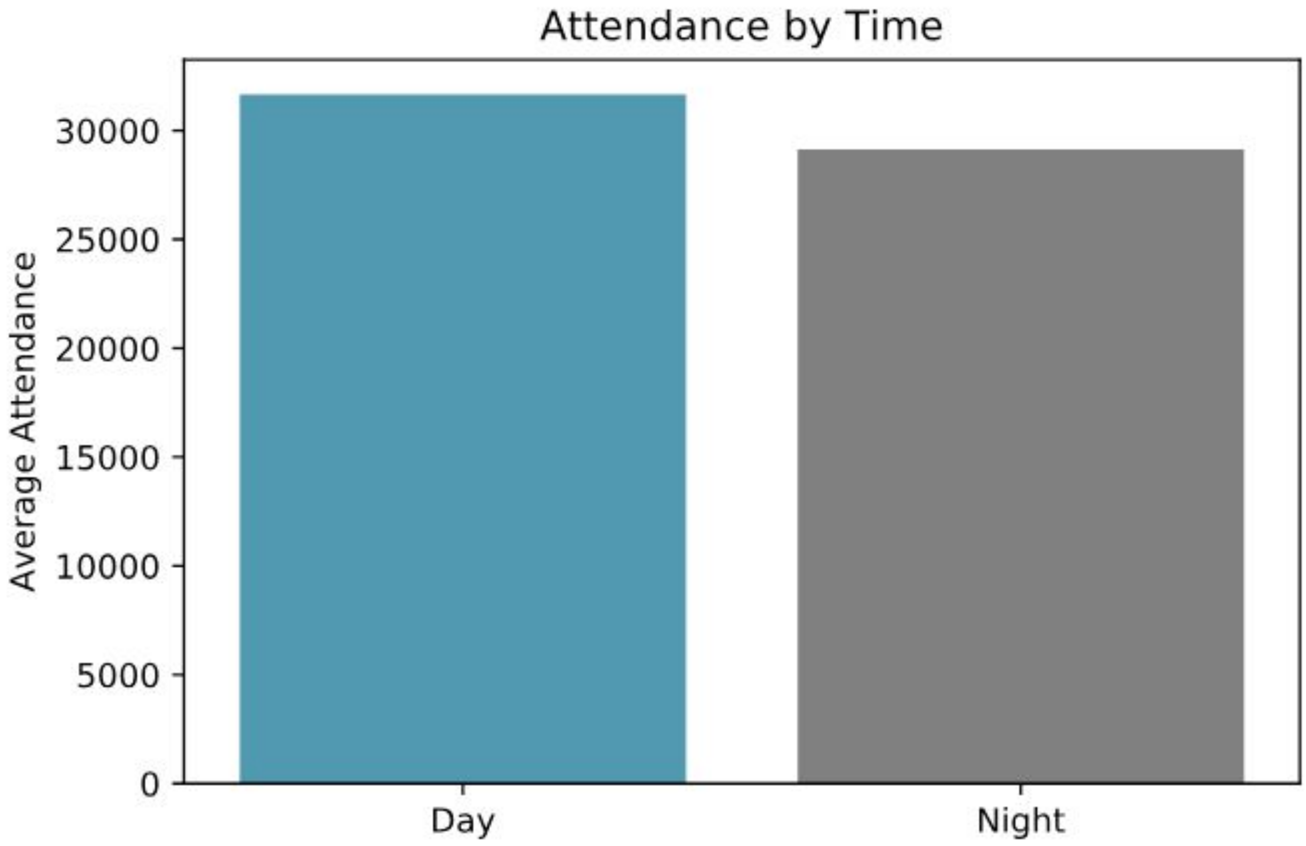


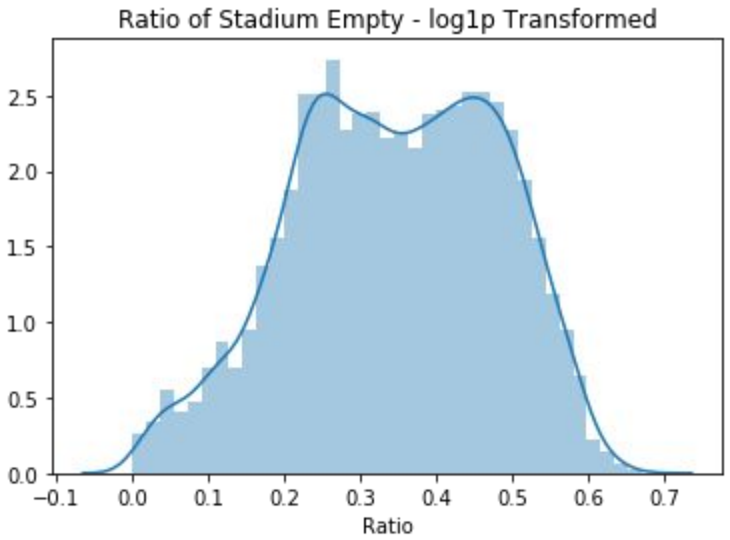
Future Work

- Transforming Y variable
- Adding more features
  - Weather
  - Digging in further on opponent features
  - Other sports leagues

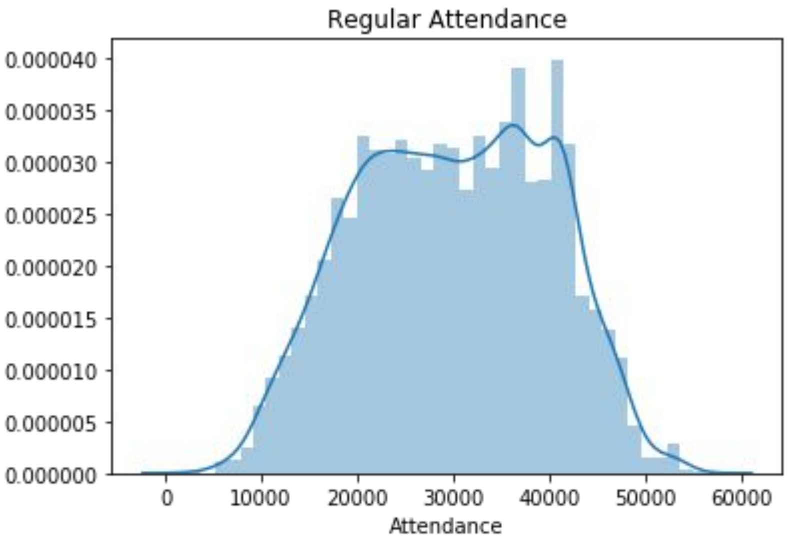
Questions?

Appendix

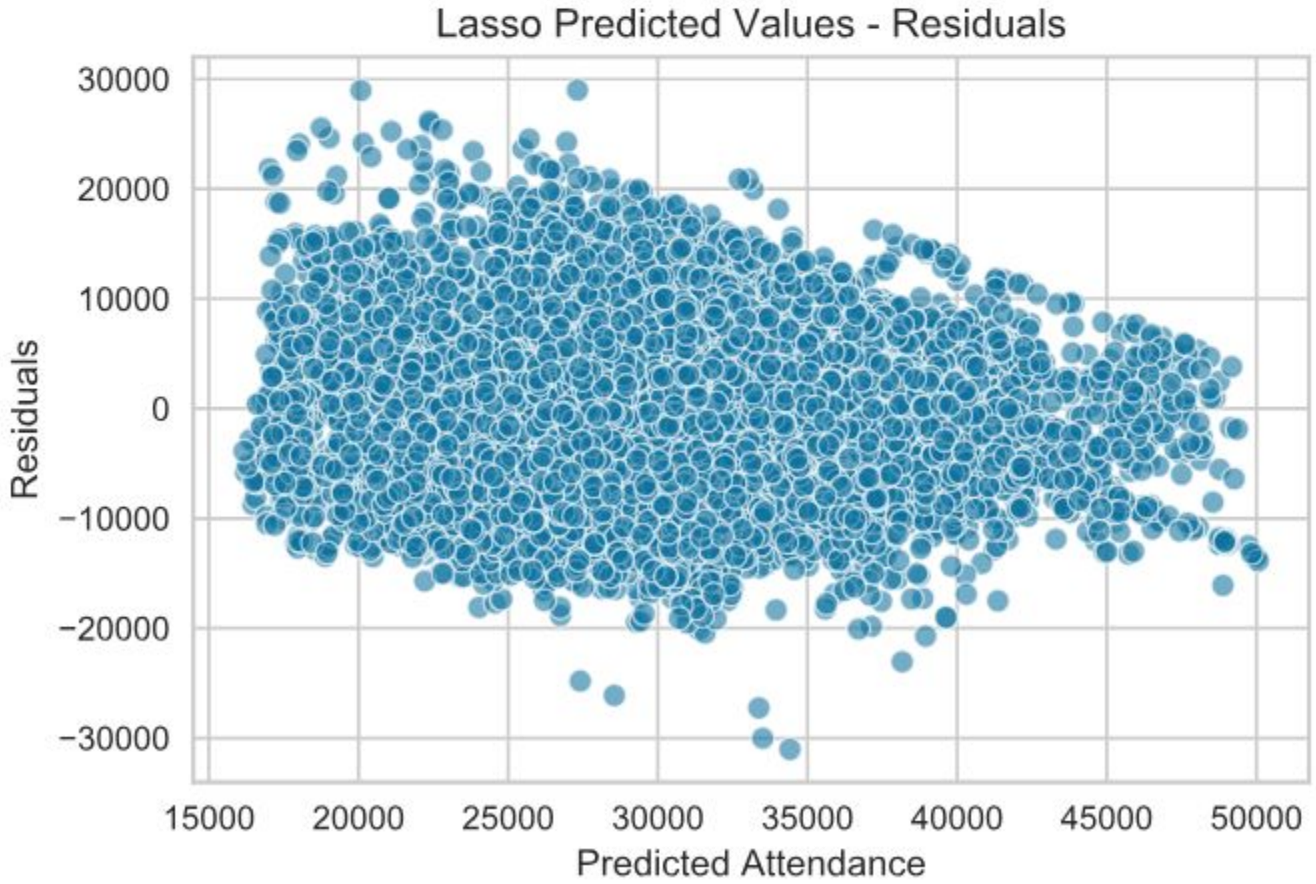

Attendance by Time
Average Attendance
30000
25000
20000
15000
10000
5000
0
Day
Night



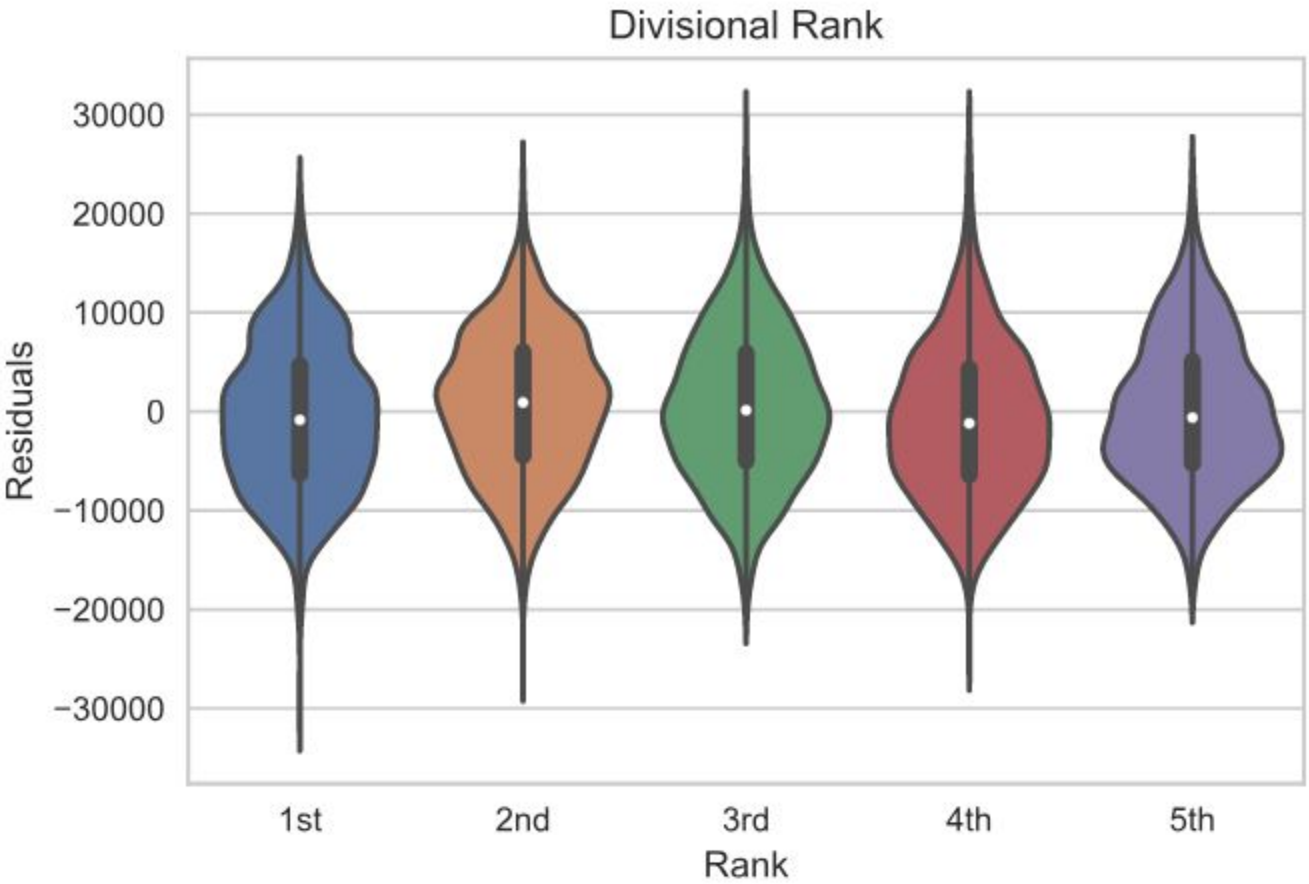
Ratio of Stadium Empty - log1p Transformed
2.5
2.0
1.5
1.0
0.5
0.0
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Ratio



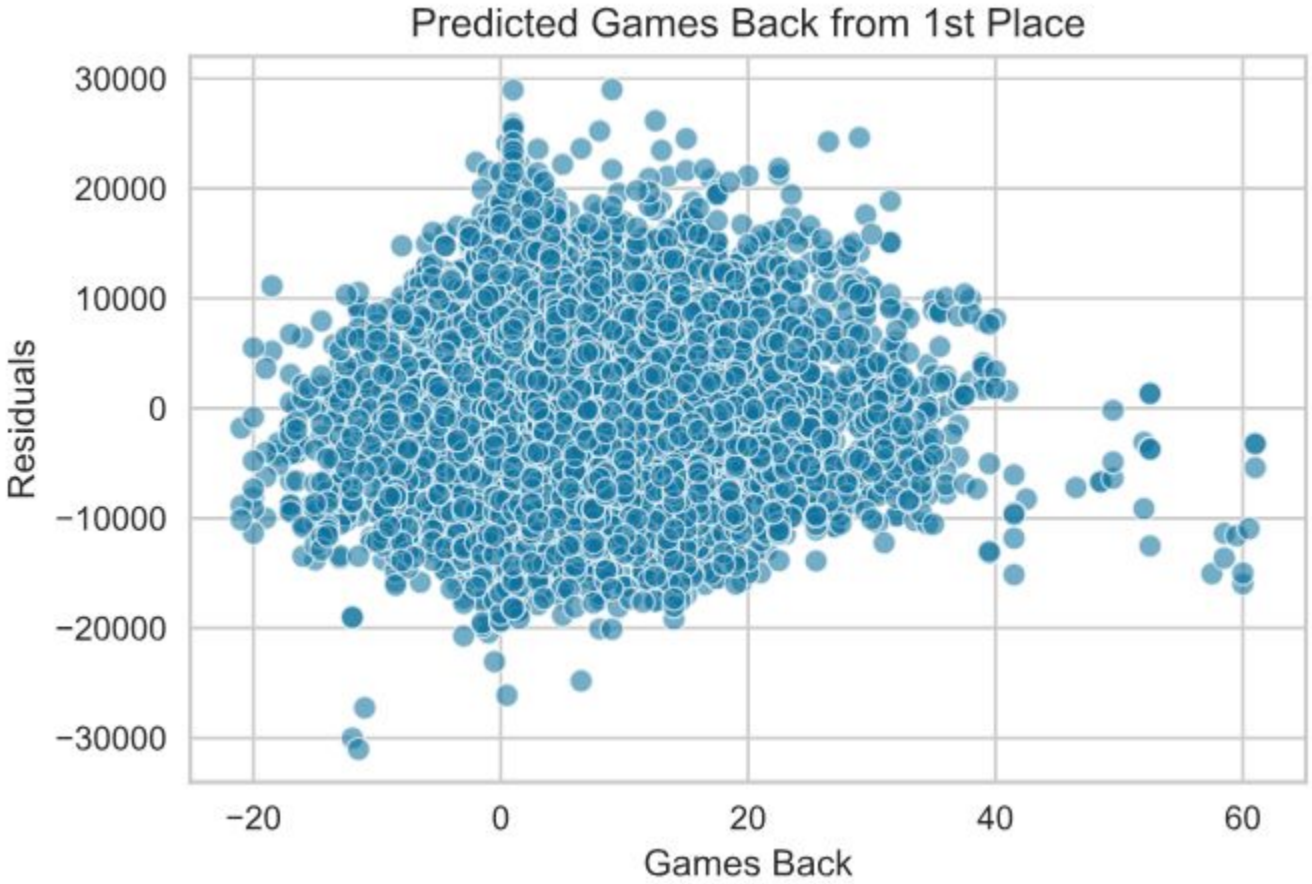
Regular Attendance
0.000040
0.000035
0.000030
0.000025
0.000020
0.000015
0.000010
0.000005
0.000000
0
10000
20000
30000
40000
50000
60000
Attendance



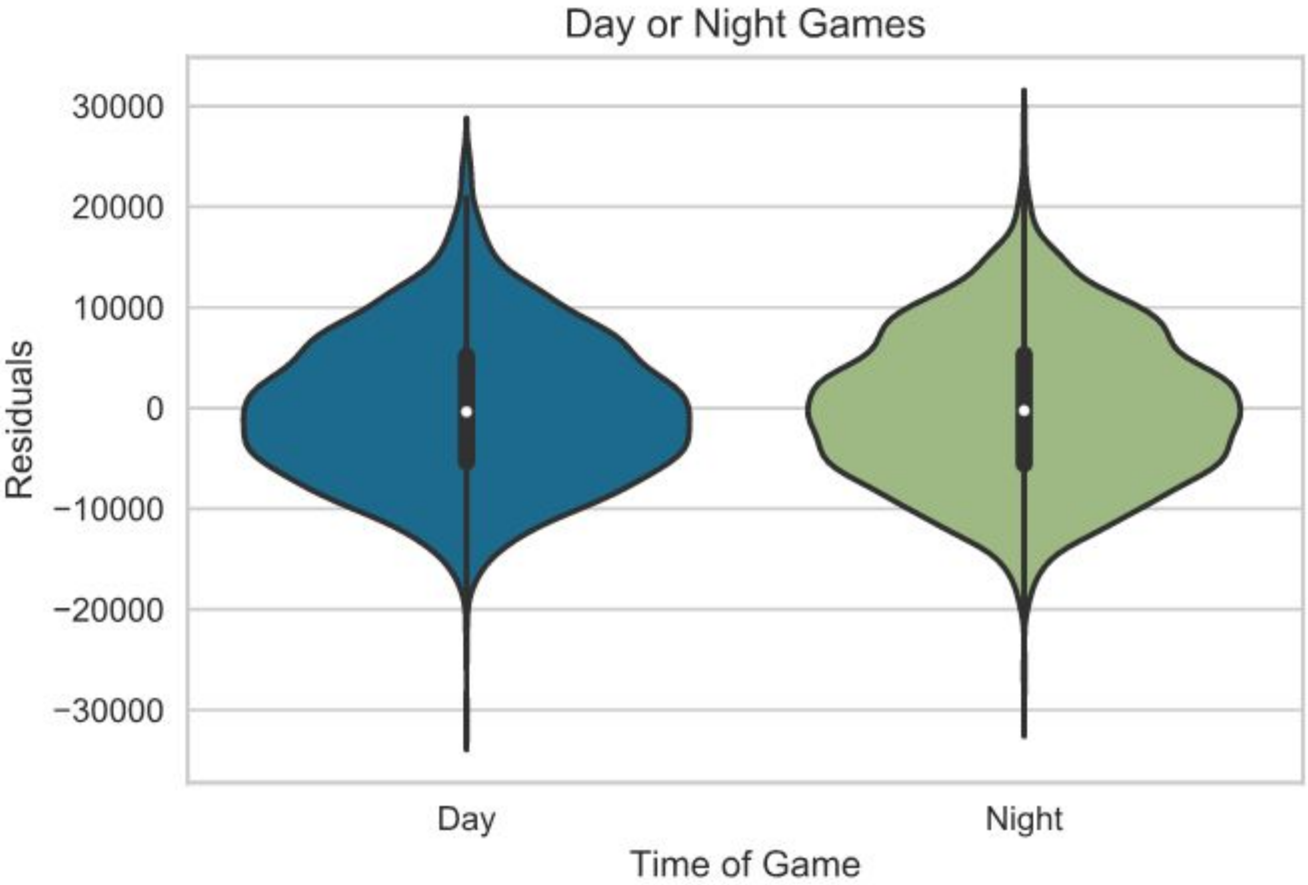
*Training Data



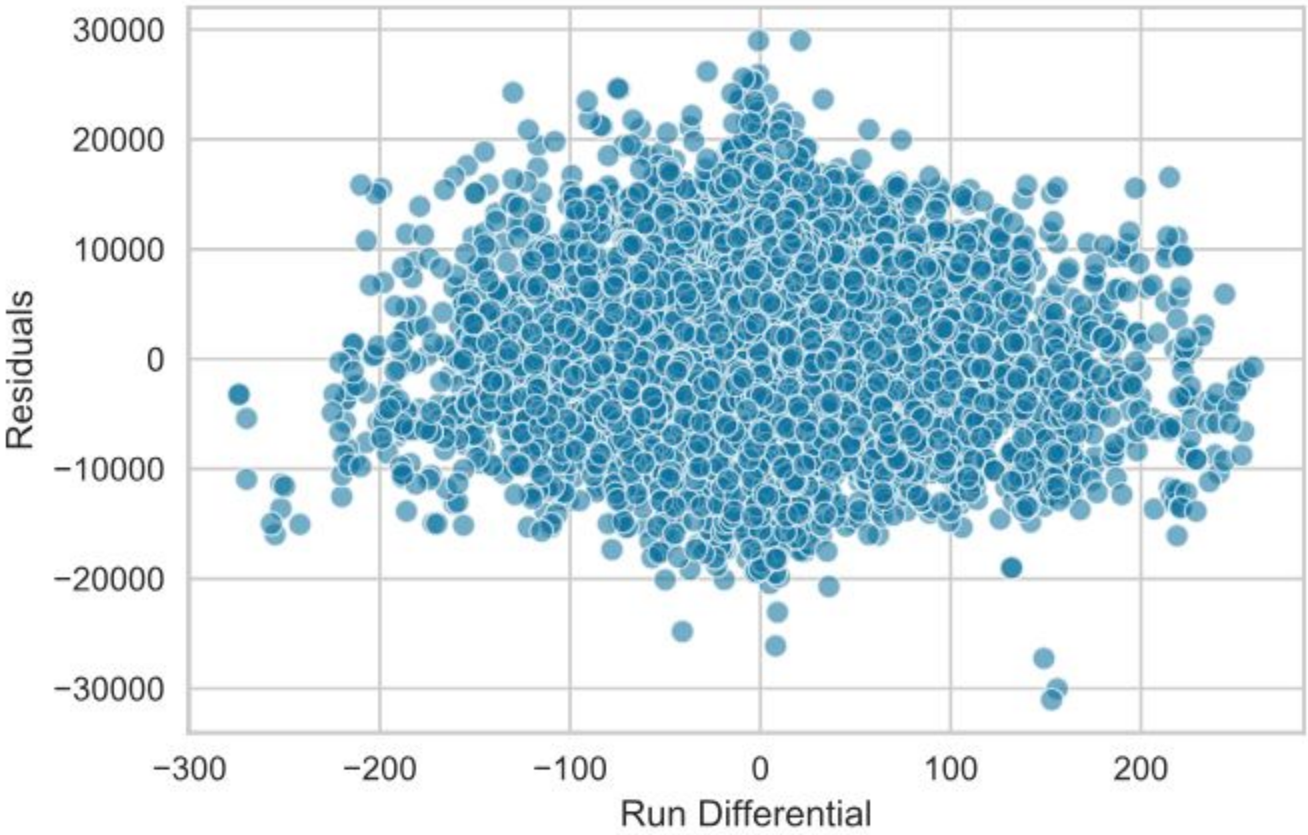
*Training Data



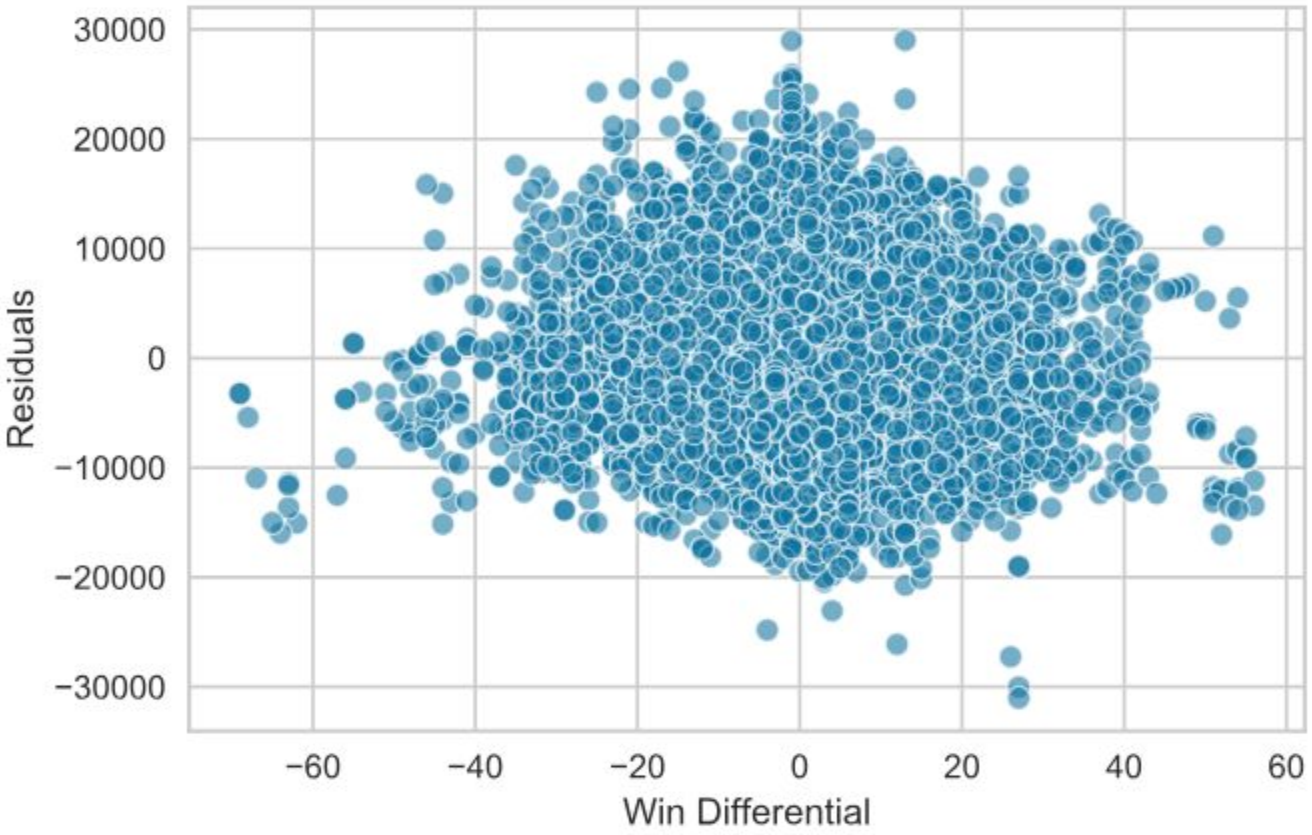
*Training Data



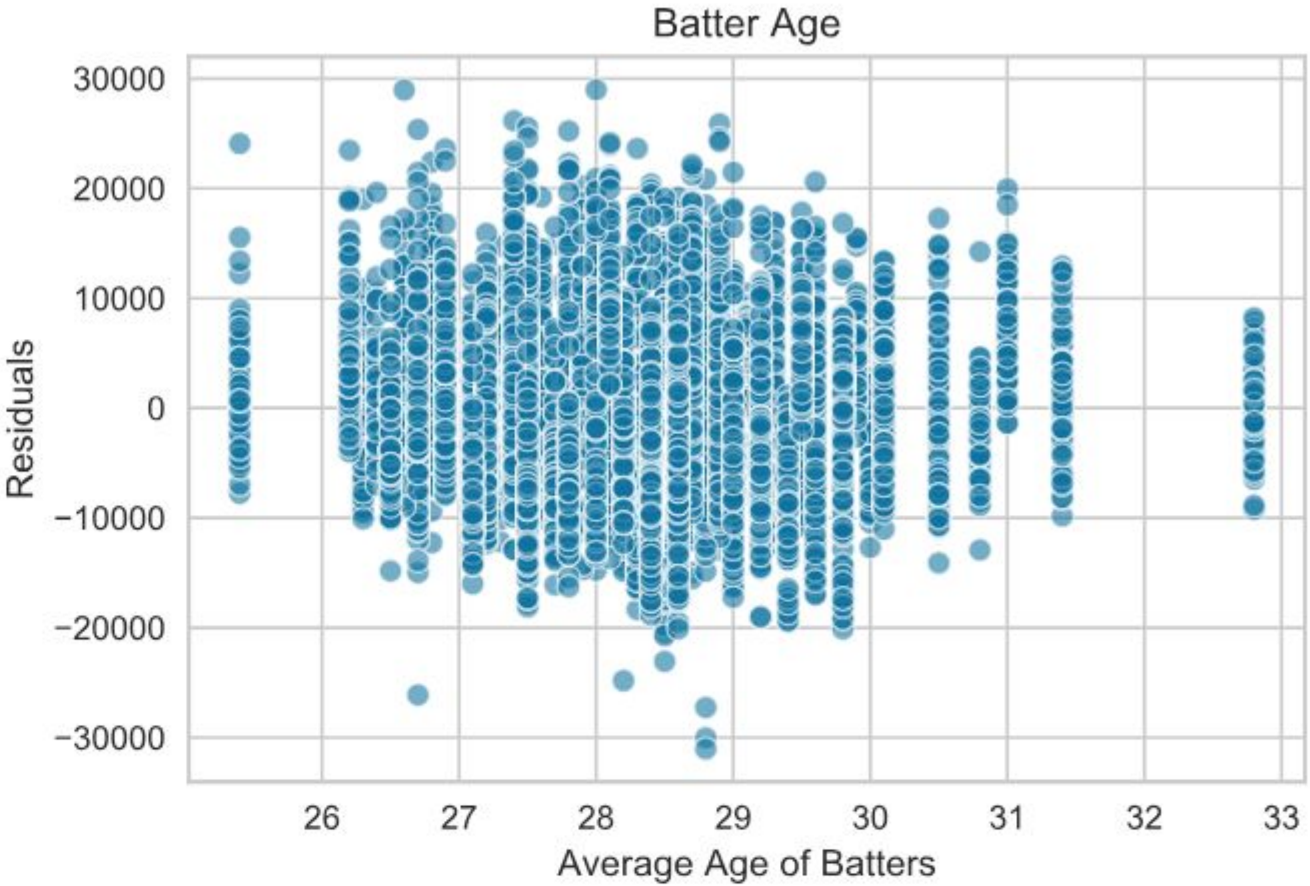
*Training Data



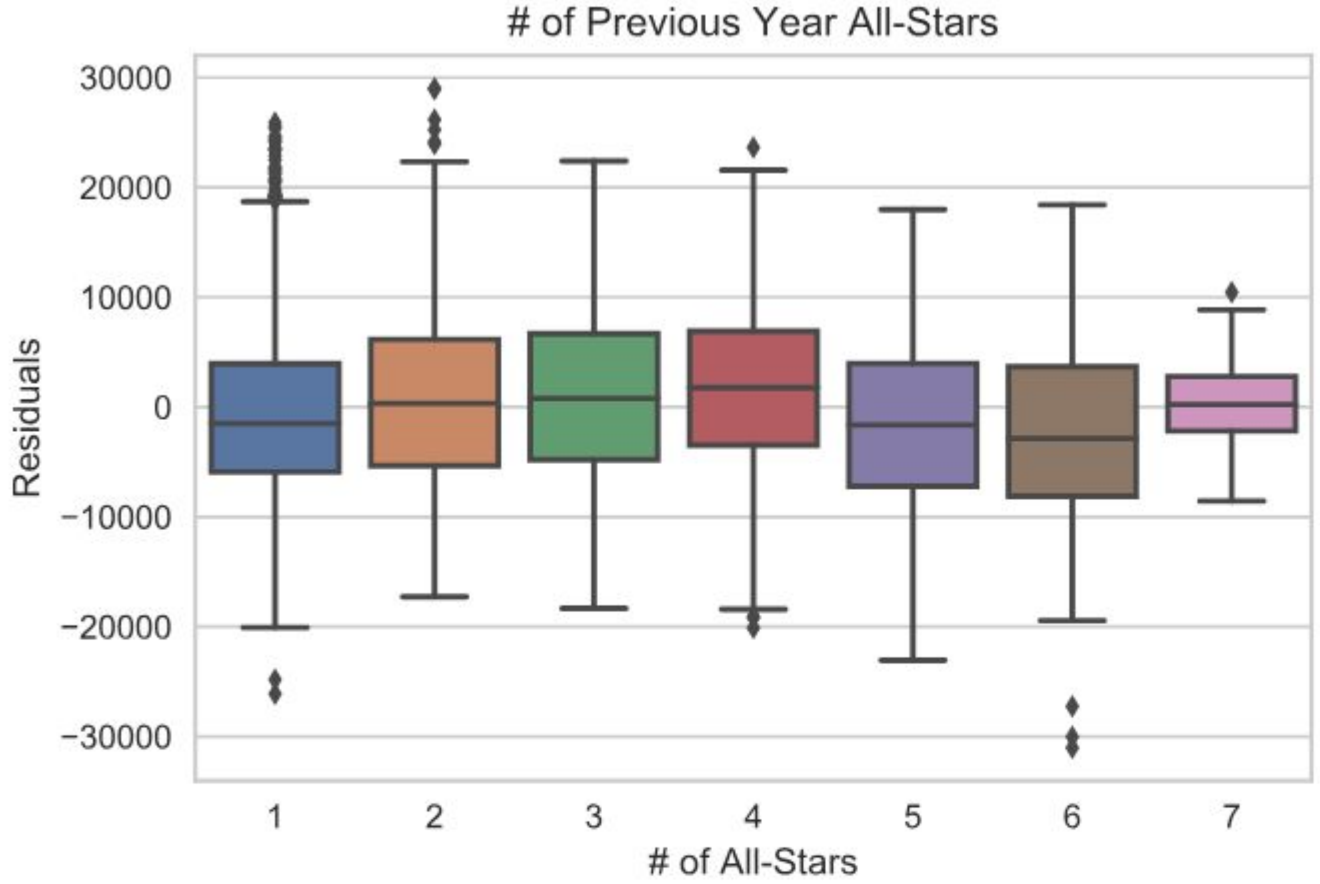
*Training Data



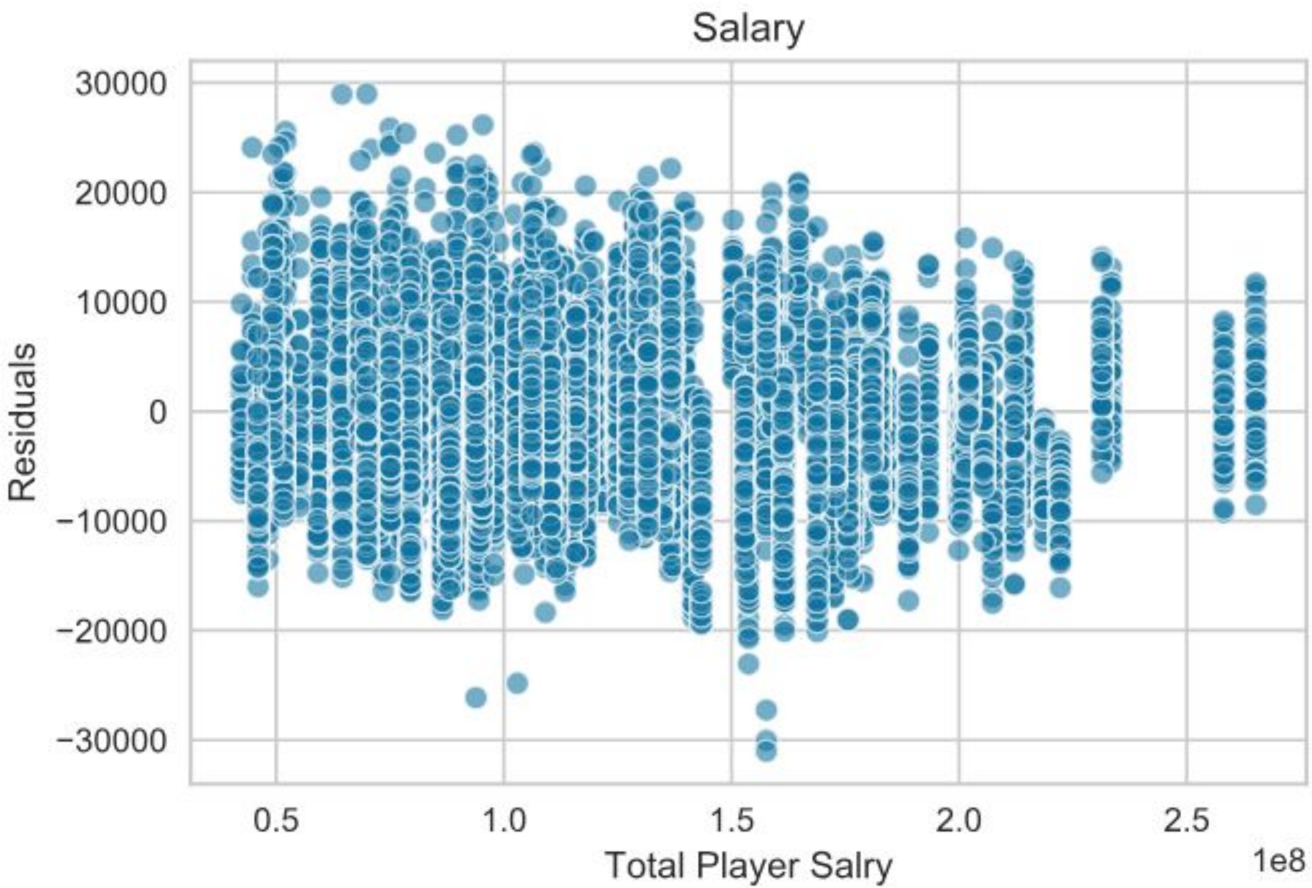
*Training Data



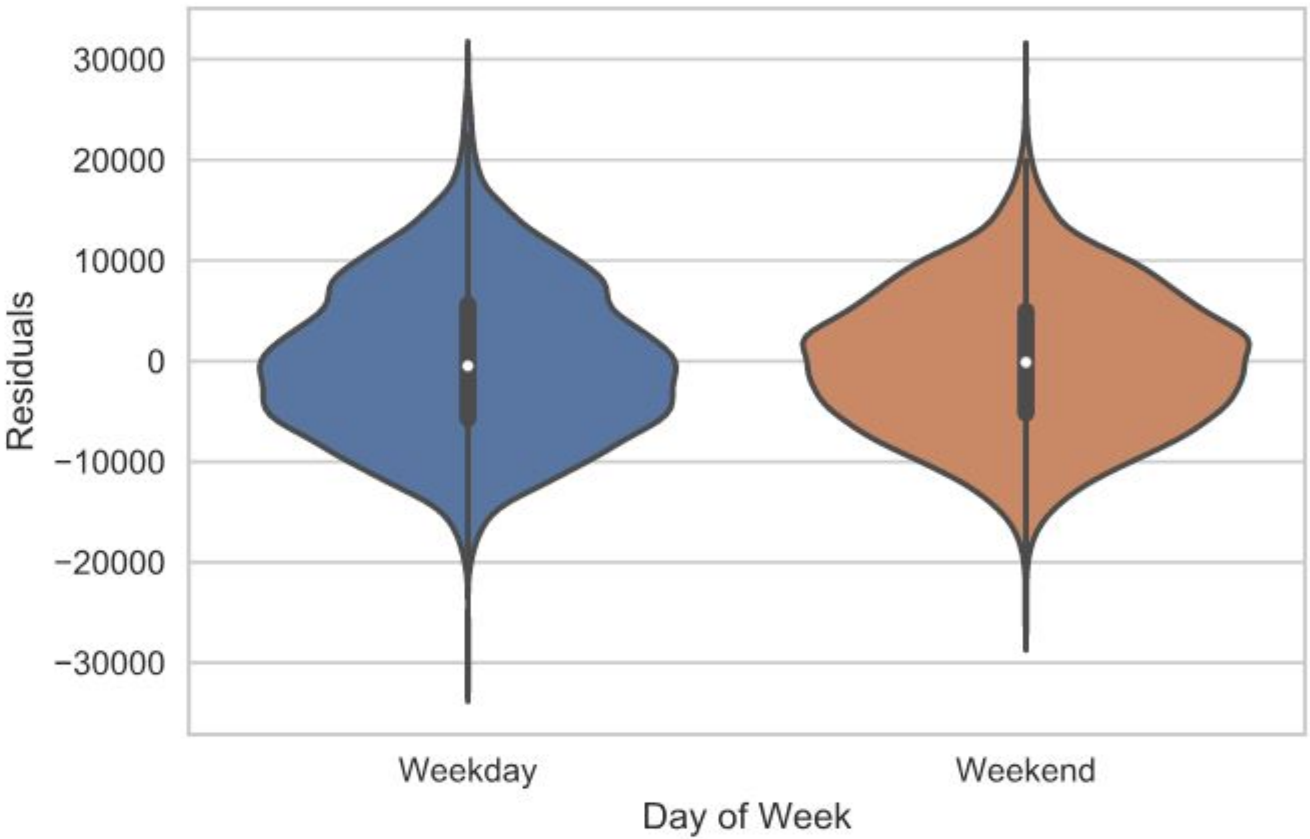
*Training Data



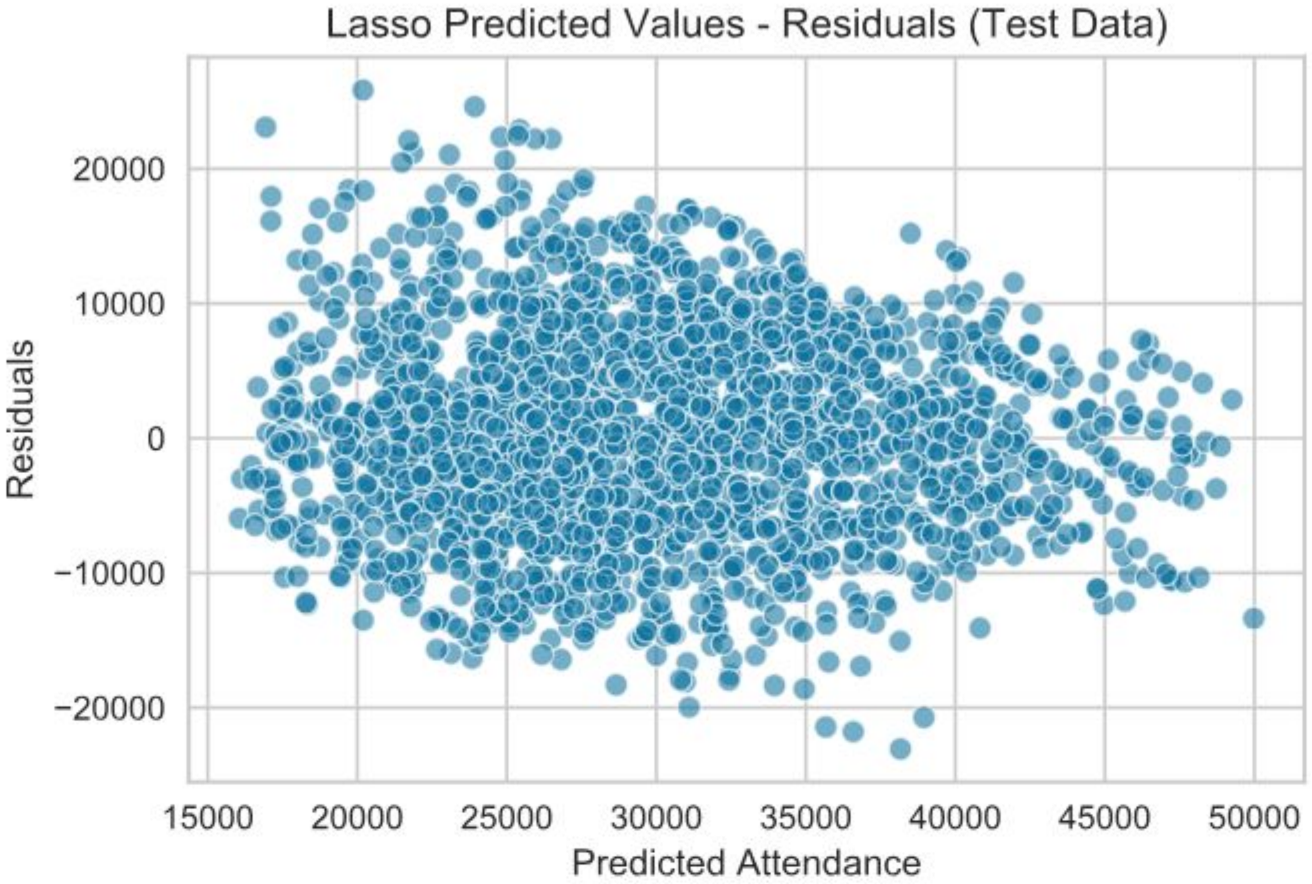
*Training Data



*Training Data



*Training Data



Lasso Predicted Values - Residuals (Test Data)
Residuals
20000
10000
0
−10000
−20000
15000
20000
25000
30000
35000
40000
45000
50000
Predicted Attendance

*Test Data

Opposing Team Features - Heat Map