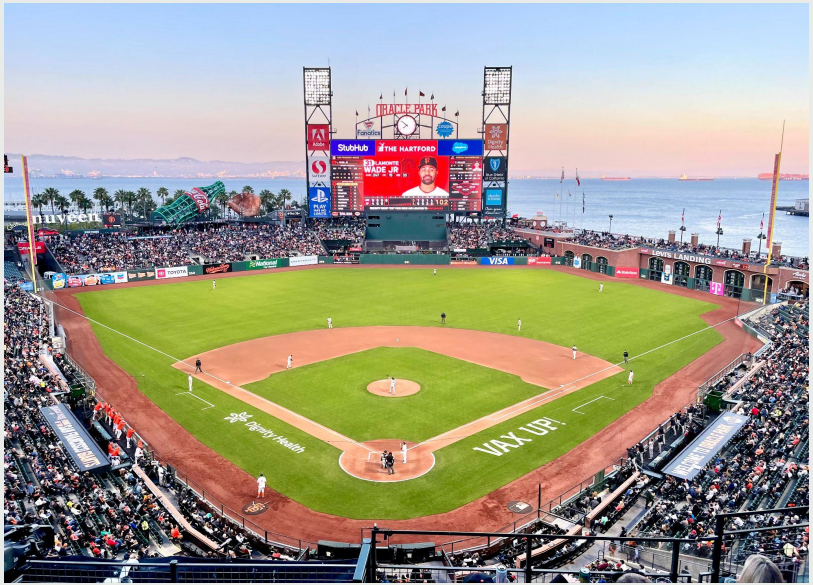
Can We Predict a MLB Team's Attendance Using Team Performance Stats?



Made by Kyle Matsui

Introduction: Topic and Motivation

MLB attendance is an important metric for teams, reflecting both fan engagement and financial health, as a lot of the teams' revenue comes from attendance. I've been playing baseball since I was five, and I've always wondered if a team's performance affect their attendance, and this project allows me to answer that question.

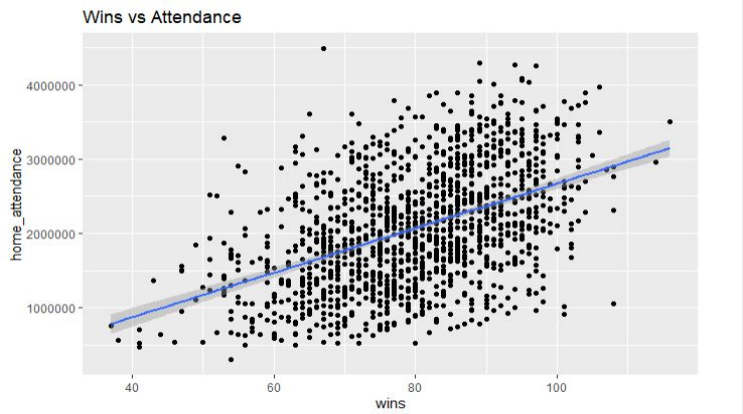
Introduction: The Dataset

This dataset contains season-level stats for MLB teams starting from 1876. Each row represents one team's performance in a given season and it includes variables such as wins, losses, runs scored, walks, homeruns, and more. After removing all the NA values, the dataset has 1384 observations across 43 variables.

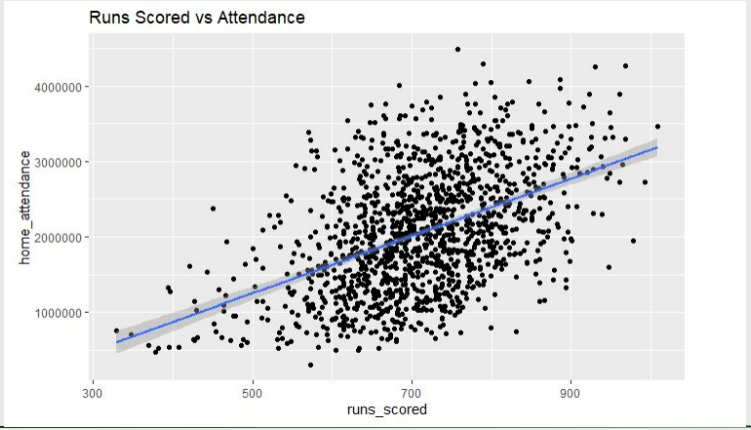
10	8	4221	1233	5	510	477	309	120	0.952	Cincinnati	Palace of	424643
15	3	4083	1212	9	348	568	278	110	0.957	Cleveland	League Pa	354627
17	12	4261	1254	16	359	528	276	87	0.959	Detroit Tig	Bennett Pa	490490
17	15	4322	1248	28	397	735	307	99	0.954	New York	Polo Grou	783700
18	8	4051	1223	21	422	597	330	94	0.948	New York	Hilltop Pa	501000
27	3	4134	1069	9	386	728	245	92	0.961	Philadelpl	Shibe Parl	674915
17	6	4173	1190	23	472	612	241	97	0.962	Philadelpl	Baker Bow	303177
21	11	4205	1174	12	320	490	228	100	0.964	Pittsburgh	Exposition	534950
21	4	4064	1287	16	383	620	267	107	0.958	St. Louis B	Sportsmai	366274
5	4	4139	1368	22	483	435	322	90	0.95	St. Louis C	Robison F	299982
11	2	4124	1288	12	424	653	280	100	0.957	Washingt	American	205199
12	6	4290	1236	30	414	670	309	80	0.954	Boston Re	Huntingto	584619
15	5	4261	1331	17	545	555	235	125	0.964	Brooklyn S	Washingt	279321
12	9	4171	1328	36	599	440	305	137	0.954	Boston Dc	South End	149027
23	7	4263	1130	16	381	785	314	100	0.954	Chicago W	South Side	552084
25	13	4136	1171	18	474	609	230	110	0.963	Chicago C	West Side	526152
16	11	4160	1334	27	528	497	291	103	0.955	Cincinnati	Palace of	380622
13	5	4401	1392	10	488	617	248	112	0.964	Cleveland	League Pa	293456
17	5	4141	1257	34	460	532	288	79	0.956	Detroit Tig	Bennett Pa	391288
9	10	4175	1290	30	397	717	291	117	0.955	New York	Polo Grou	511785
14	8	4197	1238	16	364	654	286	95	0.956	New York	Hilltop Pa	355857

Exploratory Data Analysis

Through EDA, I saw that attendance tends to increase with team success. Multiple scatterplots showed positive relationships between attendance and variables like wins and runs scored. However, the r-squared value were relatively low, suggesting that other factors are playing a significant role in attendance.



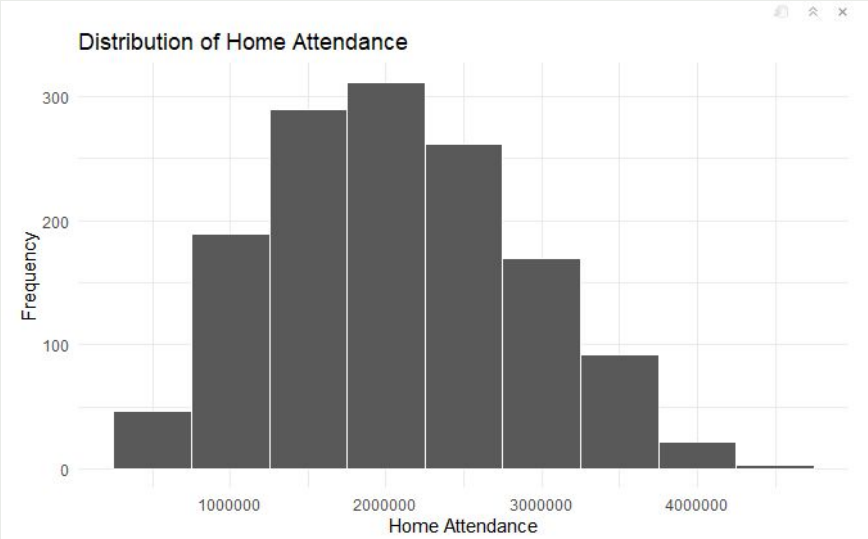
Multiple R-squared: 0.2181, Adjusted R-squared: 0.2176



Multiple R-squared: 0.2211, Adjusted R-squared: 0.2205

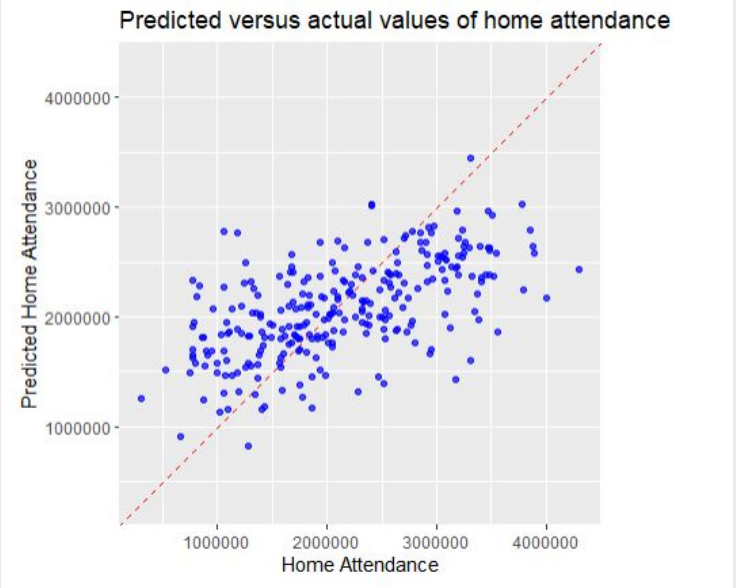
Exploratory Data Analysis

I also created a histogram of the home attendance variable to see the distribution of the values. The histogram showed a slightly right-skewed distribution, with most teams hovering around the 2,000,000 value. This shows us that while attendance is usually consistent, some teams draw more fans while some draw less fans.



Exploratory Data Analysis

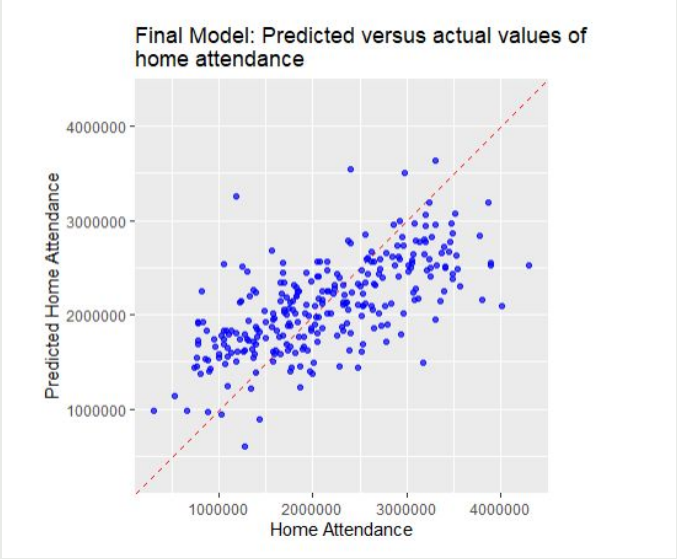
Additionally, I implemented cross-validation to assess the accuracy of the linear regression model comparing attendance to wins, homeruns, runs scored, and hits, variables that I believed were the best predictors of attendance. As both the MAE and RMSE values are high while the R-squared value was pretty low, the model is still only moderately effective at predicting attendance.



.metric <chr>	.estimator <chr>	.estimate <dbl>
rmse	standard	663592.3041564
rsq	standard	0.3584348
mae	standard	525000.4165822

Final Model

To find the most accurate predictors of team stats on attendance, I used a Lasso Regression model on 16 variables. Lasso regression is a linear model that adds an L1 penalty, shrinking the less important coefficients towards zero, sometimes setting them to exactly zero. According to the Lasso model, the most accurate predictors were **wins**, **runs per game**, **hits**, **homeruns**, and **strikeouts by pitchers**. I then used these variables and created my final linear regression model with cross-validation. The RMSE and MAE values decreased a tiny bit while the R-squared value increased a little bit but not enough to deem a strong predictive model.



.metric <chr>	.estimator <chr>	.estimate <dbl>
rmse	standard	619372.4178201
rsq	standard	0.4389988
mae	standard	496003.6700120

Interesting Findings

Despite the improvements to the final model, the overall model performance was still not strong enough to be considered highly predictive. This indicates that while these team stats help explain some variation in attendance, other external factors such as weather, promotions, market size, and more likely play a significant role in fan turnout.

Conclusion

While this model provides a reasonable starting point for predicting attendance, it is clear that on-field performance cannot explain all the variation in fan turnout. Future work could include using the external factors and using a different model, such as random forests.