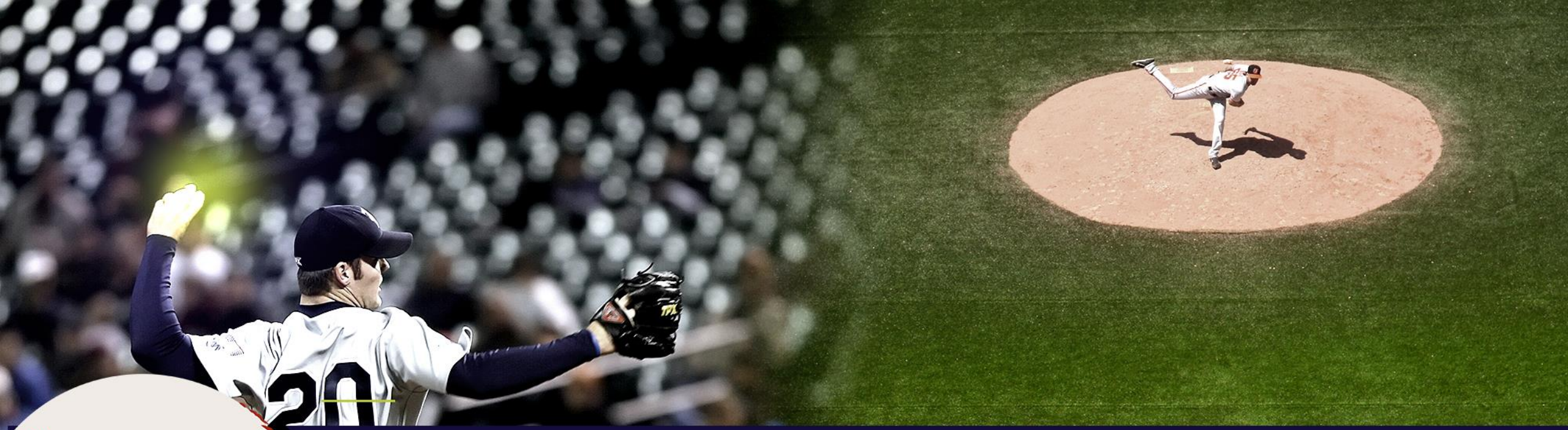
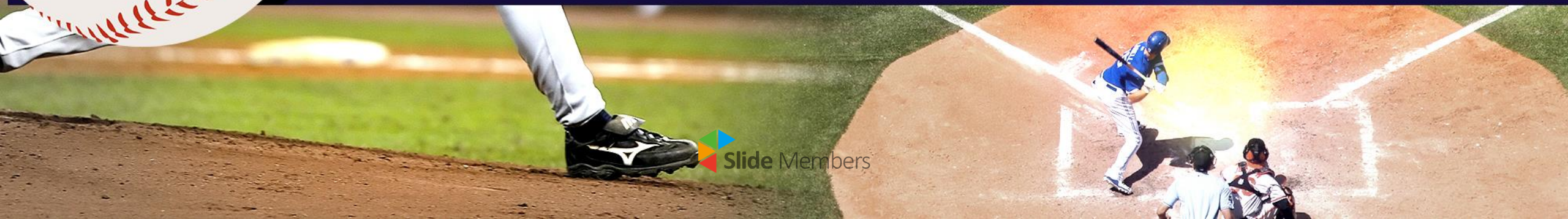


Major League Baseball

From 1901-2022



Problem Statement

Could we predict the game win/lose or team score based on the physical attributes of the game?

Could we project the team score, based on the ongoing game with the help of home team offensive stats, visiting team pitching and defensive stats?

Proposed Solution

Explore the physical attributes of the game and design a model to predict the team score or team win/lose.

Build a machine learning model to project the team score , based on the offensive , pitching and defensive measures from the given game logs .

IMPACT

To know the top features, that effects the team score and in turn effects the win/lose.

Overview of dataset

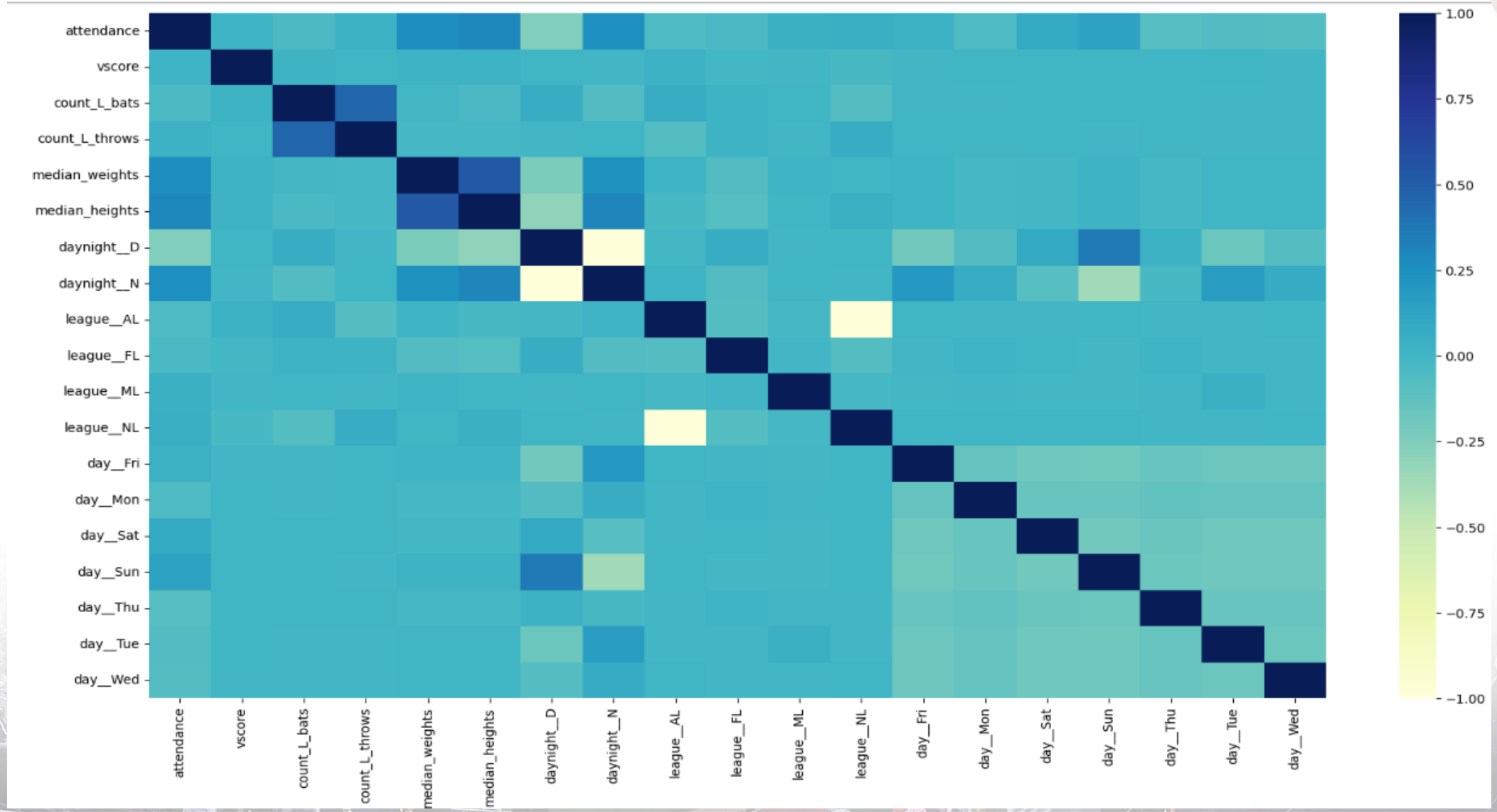
Offensive statistics
- *at-bats*
- *hits*
- *doubles*
- *triples*
- *homeruns*
- *RBI*
- *hit-by-pitch*
- *walks*
- *intentional walks*
- *strikeouts*
- *stolen bases*
- *caught stealing*
- *grounded into double plays*
- *catcher's interference*
- *left on base*

Pitching statistics
- *putouts*
- *assists*
- *errors*
- *passed balls*
- *double plays*
- *triple plays*

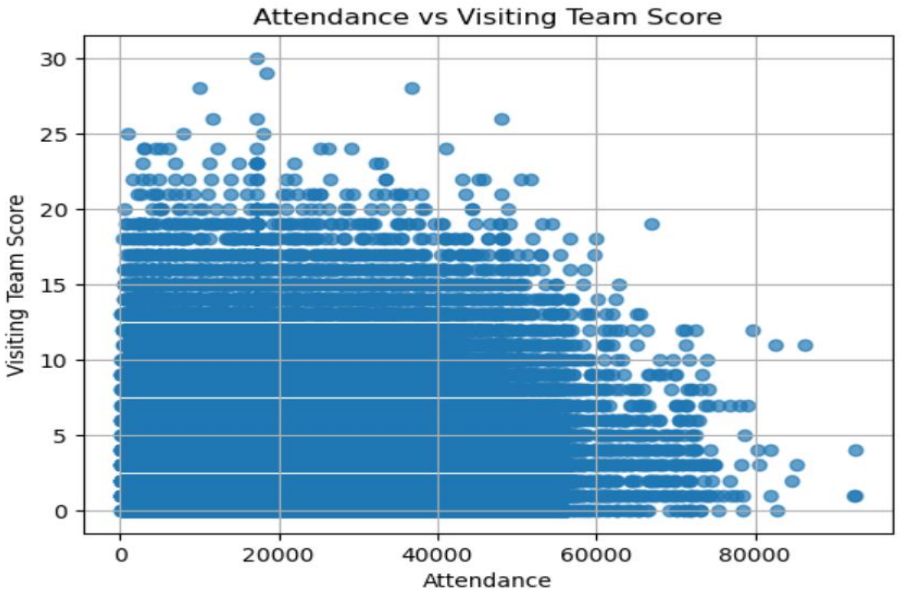
Defensive statistics
- *walks*
- *intentional walks*
- *strikeouts*
- *stolen bases*
- *caught stealing*
- *grounded into double plays*
- *catcher's interference*
- *left on base*

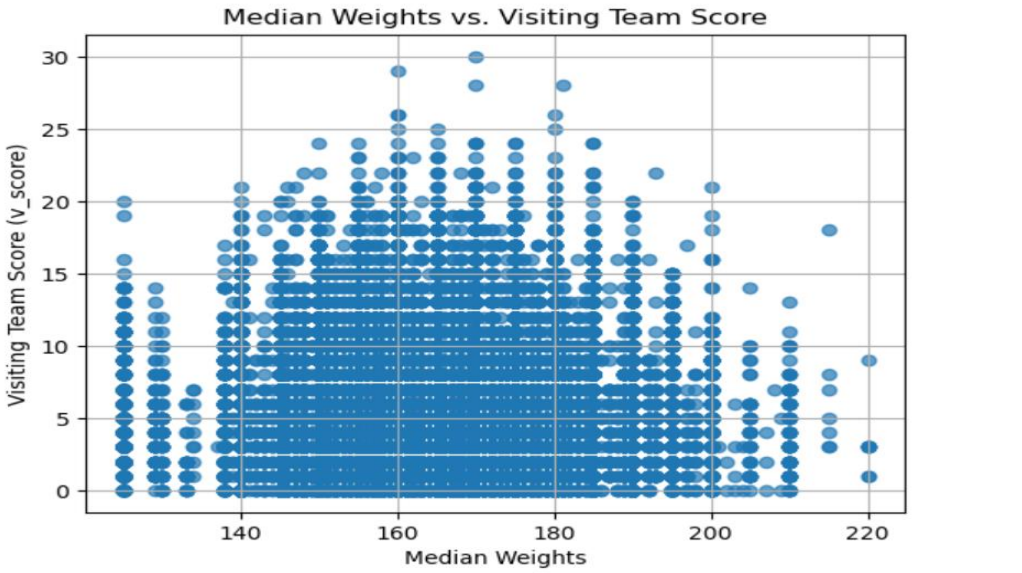
Physical attributes
- *attendance*
- *ball-park*
- *league*
- *week of the day*
- *day or night*
- *player weights*
- *player heights*
- *player throws*
- *player bats*

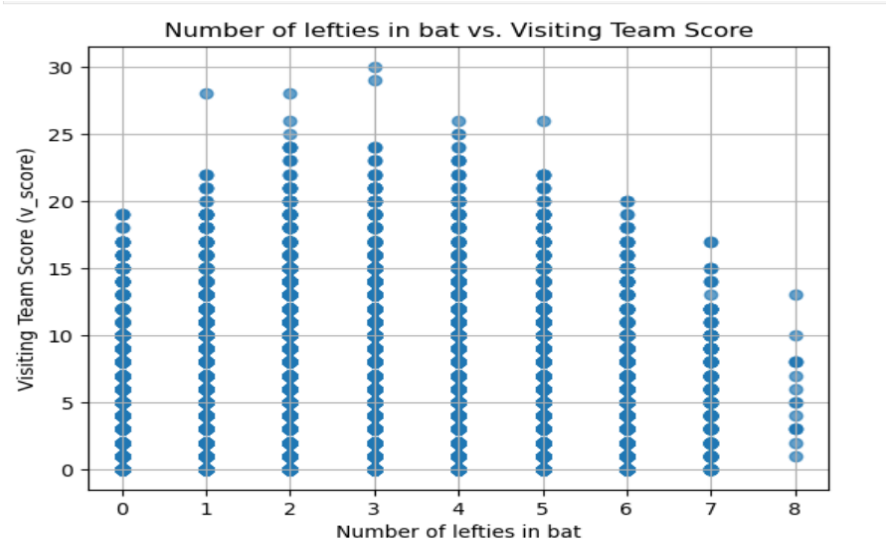
Physical features

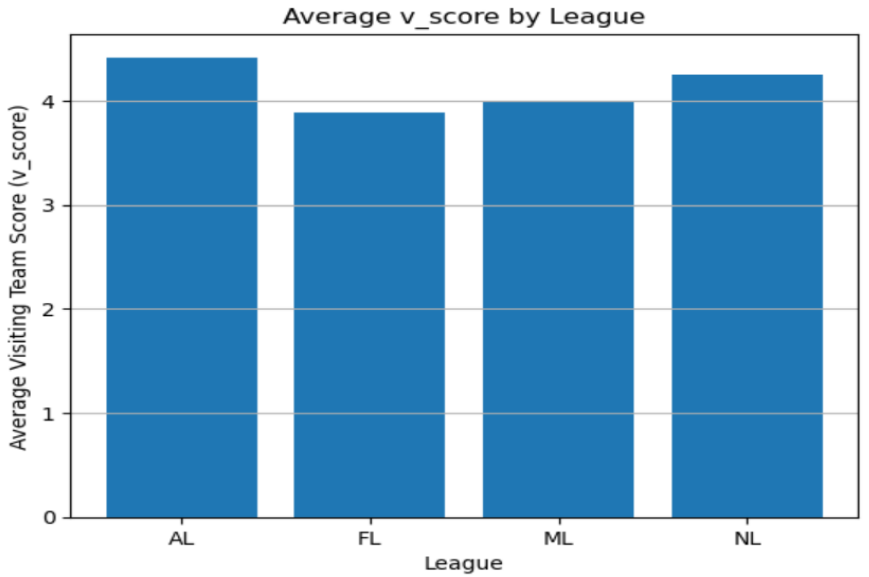


Feature variable Vs Target variable









Stats Model & Evaluation

Mean Squared Error: 9.95
R-squared (R^2): 0.02
Adjusted R-squared: 0.01

A non-zero MSE indicates there are prediction errors.

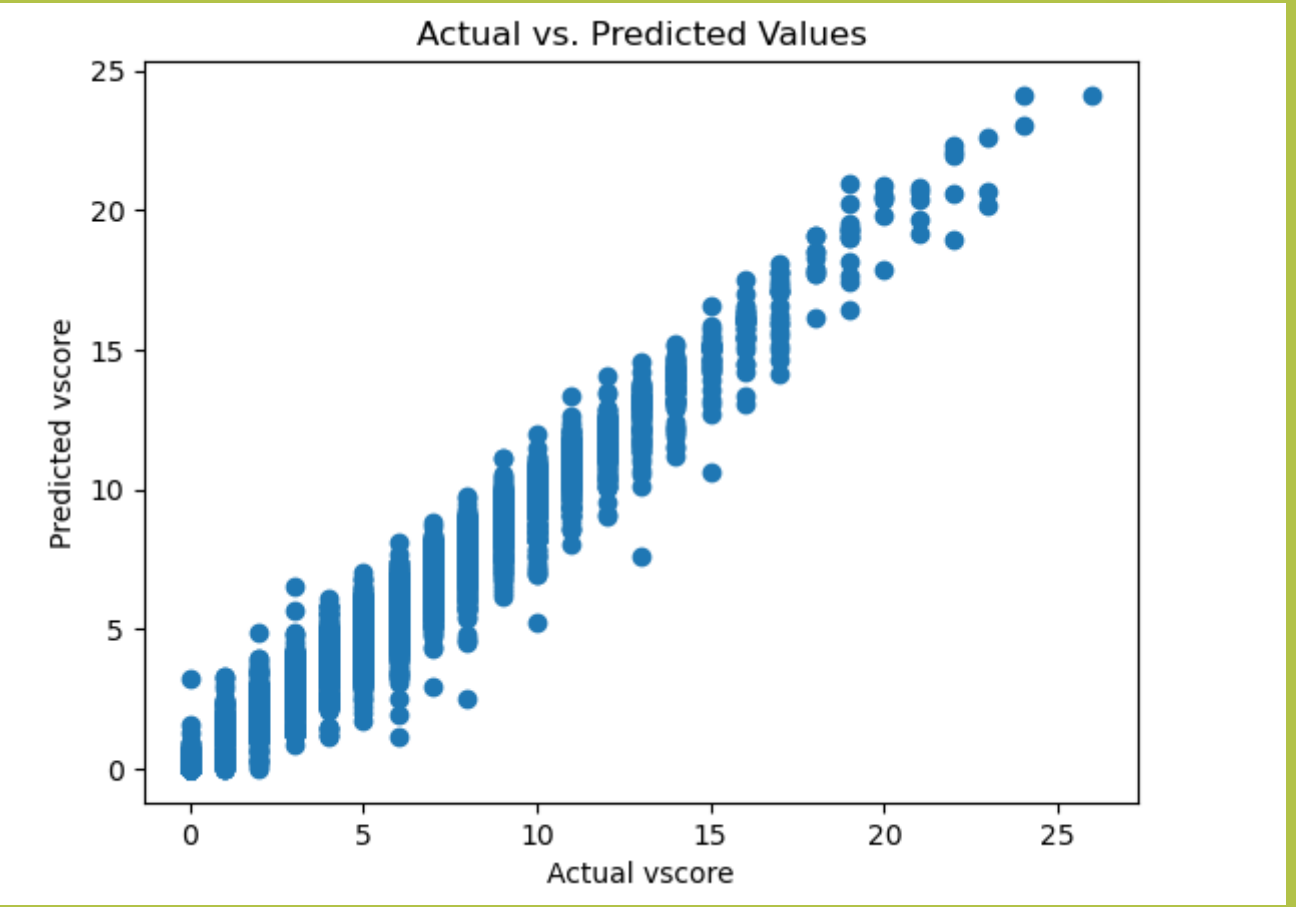
R-squared : A value close to 1 suggests a good fit

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.1207	0.111	28.235	0.000	2.904	3.337
attendance	2.537e-06	5.76e-07	4.407	0.000	1.41e-06	3.67e-06
count_L_bats	0.0464	0.005	8.812	0.000	0.036	0.057
median_weights	0.0062	0.001	9.613	0.000	0.005	0.008
daynight__D	0.0884	0.017	5.253	0.000	0.055	0.121
league__FL	-0.5112	0.091	-5.602	0.000	-0.690	-0.332
league__ML	-0.5299	0.329	-1.612	0.107	-1.174	0.115
league__NL	-0.1676	0.014	-12.051	0.000	-0.195	-0.140
day__Fri	-0.0071	0.025	-0.280	0.779	-0.057	0.043
day__Mon	0.0580	0.028	2.091	0.037	0.004	0.112
day__Sat	0.0329	0.025	1.303	0.193	-0.017	0.082
day__Sun	0.0200	0.027	0.750	0.453	-0.032	0.072
day__Thu	0.0562	0.027	2.077	0.038	0.003	0.109
day__Tue	0.0681	0.026	2.665	0.008	0.018	0.118

Baseline Model & Evaluation

Random Forest Regression Model
With Offensive, Pitching & Defensive Measures
From Game Logs

Mean Squared Error : 0.21
R-squared : 0.98



Top features in predicting the visiting team score

- Visiting team RBI
- Home team Errors
- Visiting Team Hits
- Home Team Individual Earned Runs
- Home Team Earned Runs
- Visiting Team Walks
- Visiting Team Left On Bases
- Number Of Innings
- Attendance
- Length Of The game

Future Actions.

Do PCA and Hyper tune the Random Forest Regression.

Find out the visiting team RBI's , best predictors.

Create some calculated fields to predict the RBI.

Do time-series analysis on how a particular team is doing, in terms of offensive, over the years and predict the next 5 years.

THANK YOU