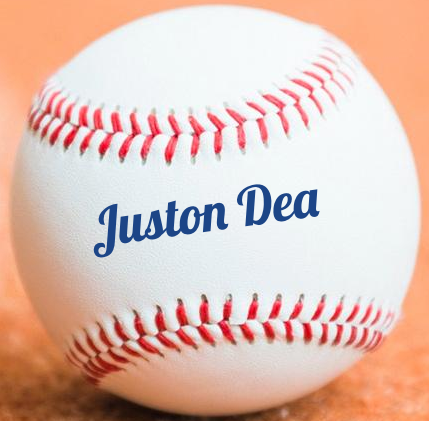
MLB ODDS PREDICTOR



Frustrated picking wrong MLB bets? No worries, we got you “covered”

Contents

Problem Statement

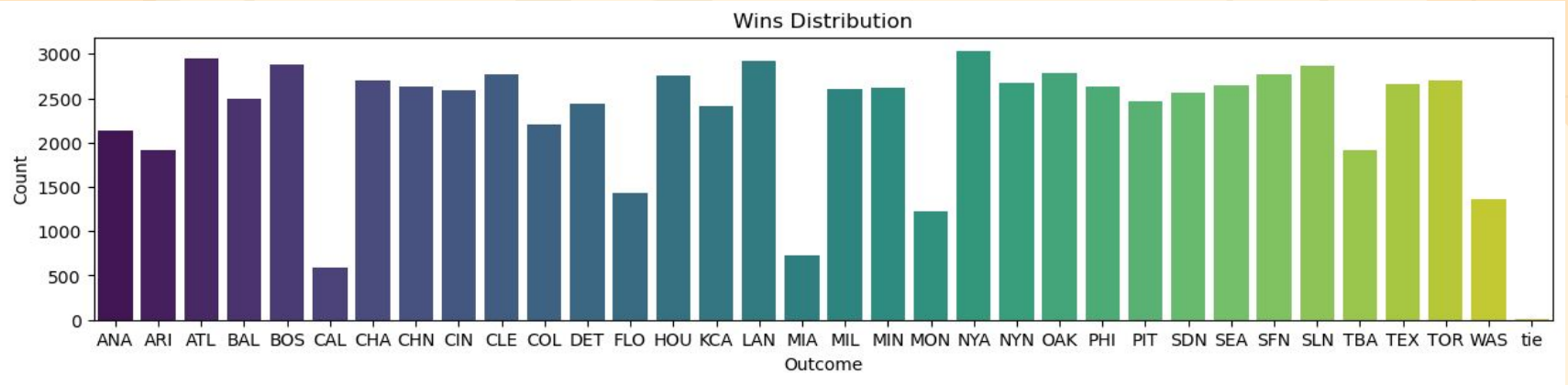
I want to build a baseball betting consulting app that can guide sports bettors on educated betting decisions. I downloaded data from retrosheet.org that has the stats for each game with the respective home and visiting teams game statistics.

Data Collection and Cleaning

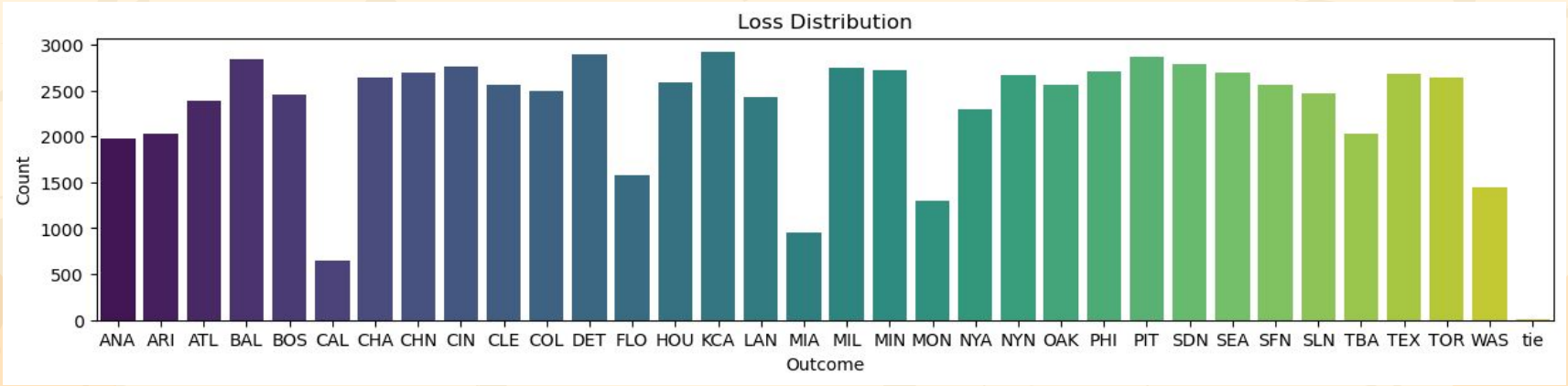
- Struggling with finding usable data, asked my friend Adam for some guidance and referred me to Retrosheet game data.
- Took the last 33 years of games starting from 1989 (earthquake World Series)
- Relabeled column names from the data dictionary given by Retrosheet
- Created y variable columns (run_diff, total, h_W/L, v_W/L)
- Filled missings, deleted columns that were repetitive, too many missings, binarized boolean columns
- Used pandas rolling sum function on features to be engineered to make data “historical” (162)
- Engineer features, taking ratios of given features to make averages
- Split data frame at 2020 to use for testing data

EDA

BA	OBP	TBS	SLG	OPS
Batting Average	**On-Base %**	**TBS**	**Slugging**	**On-Base + Slugging**
Hits / At bats	(Hits + Walks + Hit by Pitch) / At bats	1(1B) + 2(2B) + 3(3B) +4(HR)	TBS / At bats	OBP + SLG
Hit production	On base production NOT from error	Base traffic ONLY from hits	Power, XBH	Overall hitter production
.255	.368	14.08	.407	.772



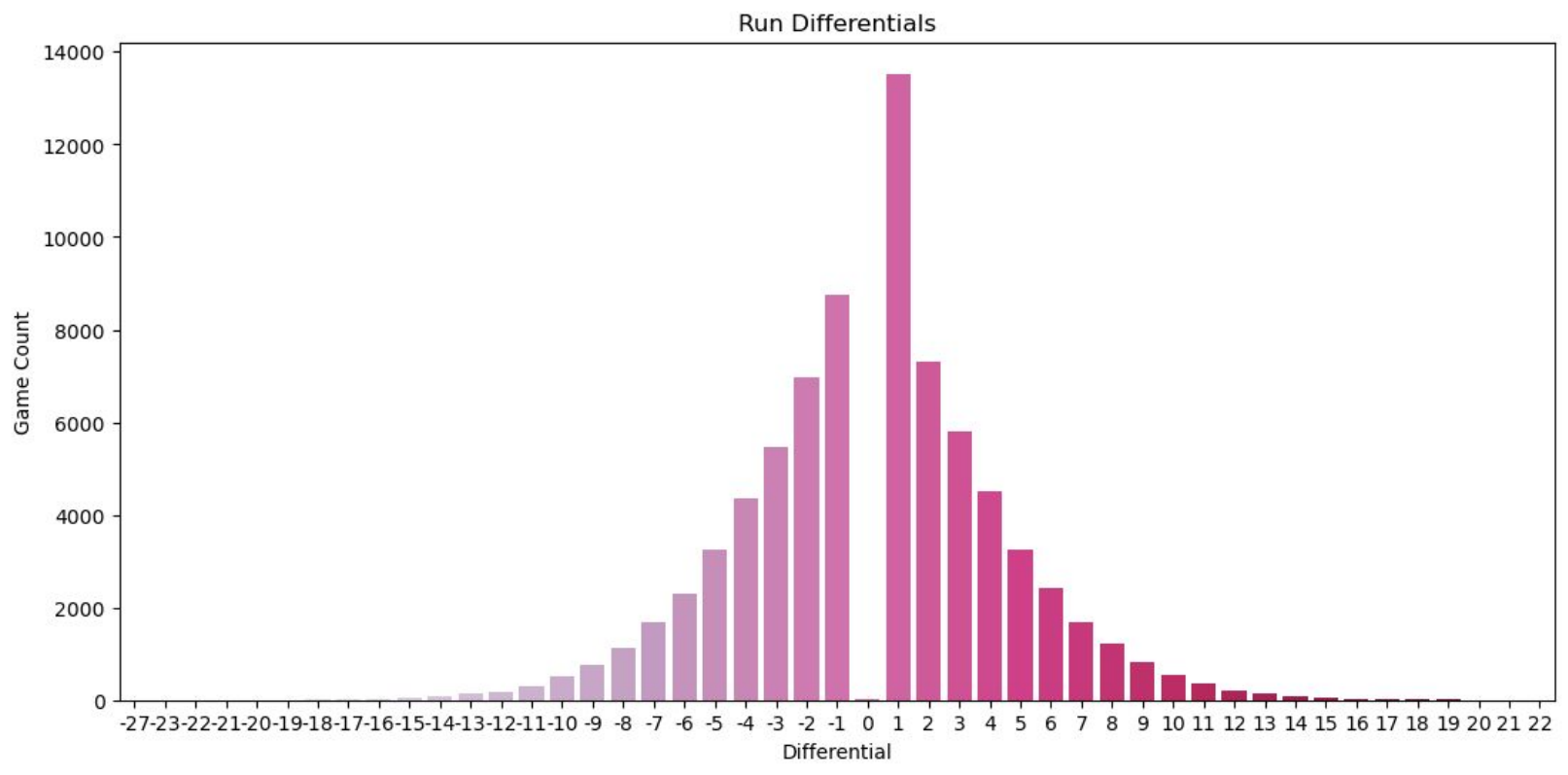
SFN 2775 - 2564, 52% Win Percentage



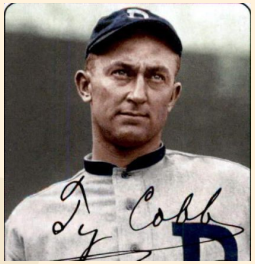
TIES?

	date	visitor_team	home_team	total_outs	day_night	park_id	attndnce	game_mins	h_runs	v_runs	W
614	19890528	ATL	SLN	57	0	STL09	35832.0	170	3	3	tie
707	19890605	PIT	PHI	45	1	PHI12	17696.0	160	3	3	tie
1885	19890913	PIT	SLN	31	1	STL09	28561.0	78	0	0	tie
9065	19930528	MON	CHN	30	0	CHI11	28523.0	102	2	2	tie
10715	19940406	SLN	CIN	31	0	CIN08	20179.0	128	8	8	tie
13318	19950717	CHA	NYA	39	1	NYC16	22707.0	136	1	1	tie
15249	19960617	TEX	BAL	34	1	BAL12	45581.0	90	1	1	tie
19172	19980426	TOR	CHA	36	0	CHI12	13705.0	130	5	5	tie
20796	19980824	SLN	PIT	39	1	PIT07	28435.0	138	5	5	tie
23688	19991003	MIN	CHA	39	0	CHI12	18694.0	101	1	1	tie
23704	20000403	MIL	CIN	31	0	CIN08	55596.0	111	3	3	tie
28457	20010930	BAL	NYA	90	0	NYC16	55351.0	301	1	1	tie
30348	20020815	SFN	ATL	55	1	ATL02	35340.0	193	3	3	tie
33261	20030918	NYA	BAL	30	1	BAL12	29093.0	87	1	1	tie
37005	20050630	HOU	CIN	42	1	CIN09	19903.0	148	2	2	tie
64948	20160929	CHN	PIT	31	1	PIT08	19991.0	75	1	1	tie

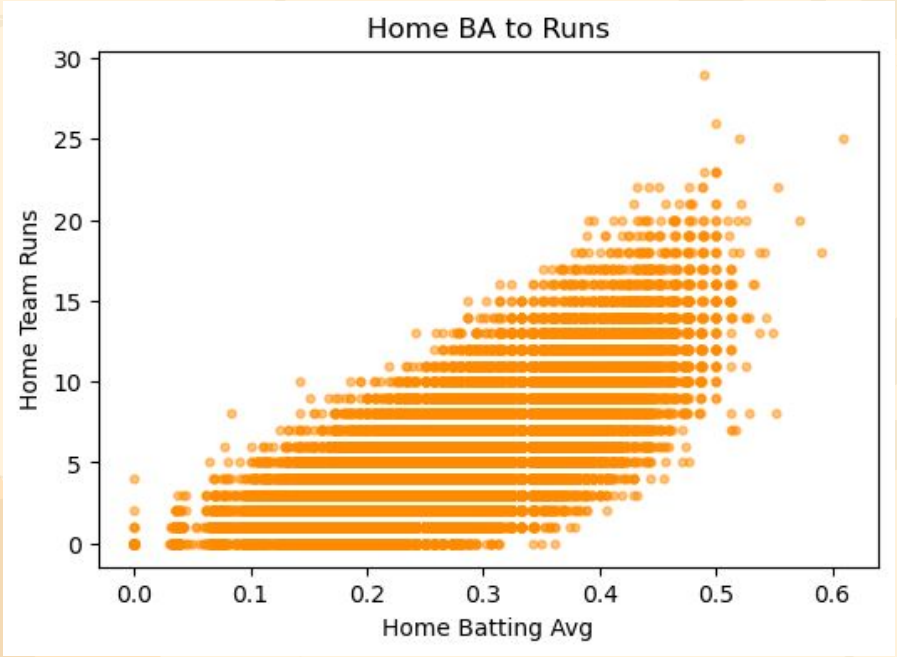
Run Differential

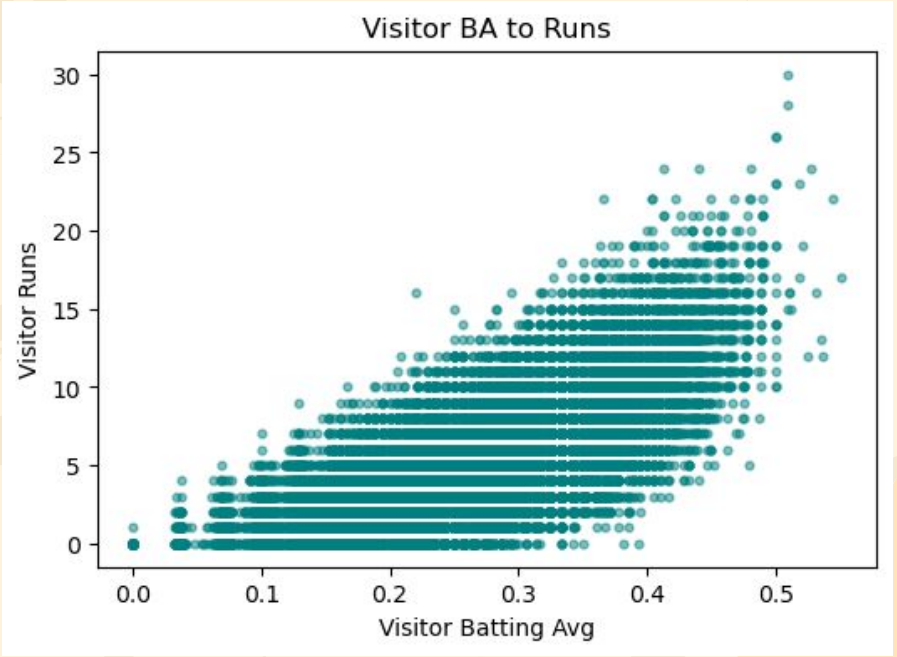


BA to Runs



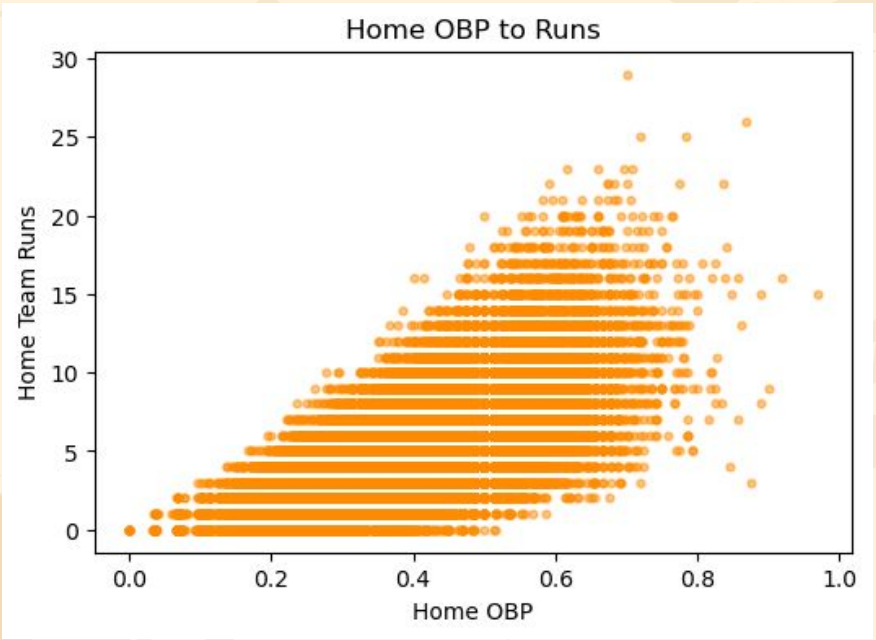
Ty Cobb
BA .366

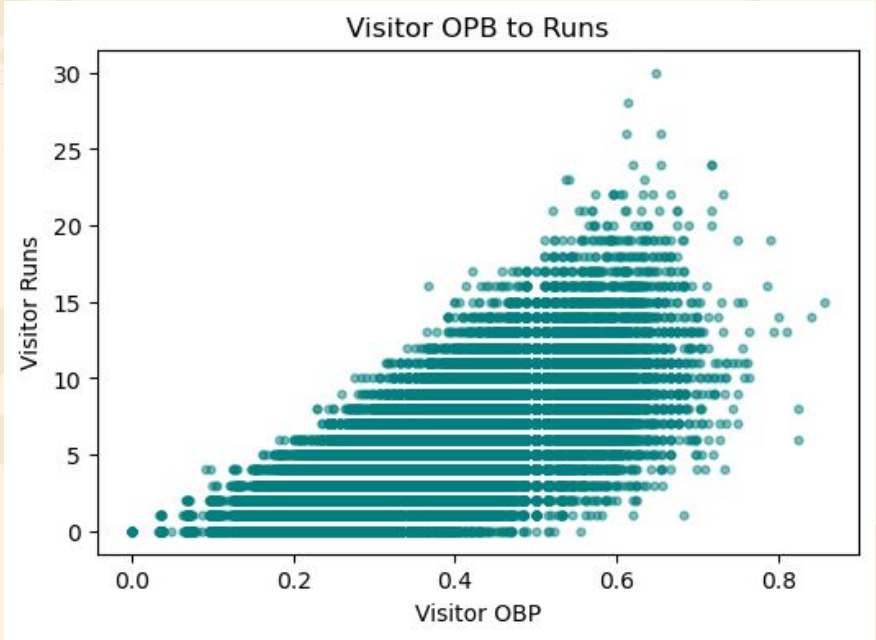




OBP to Runs

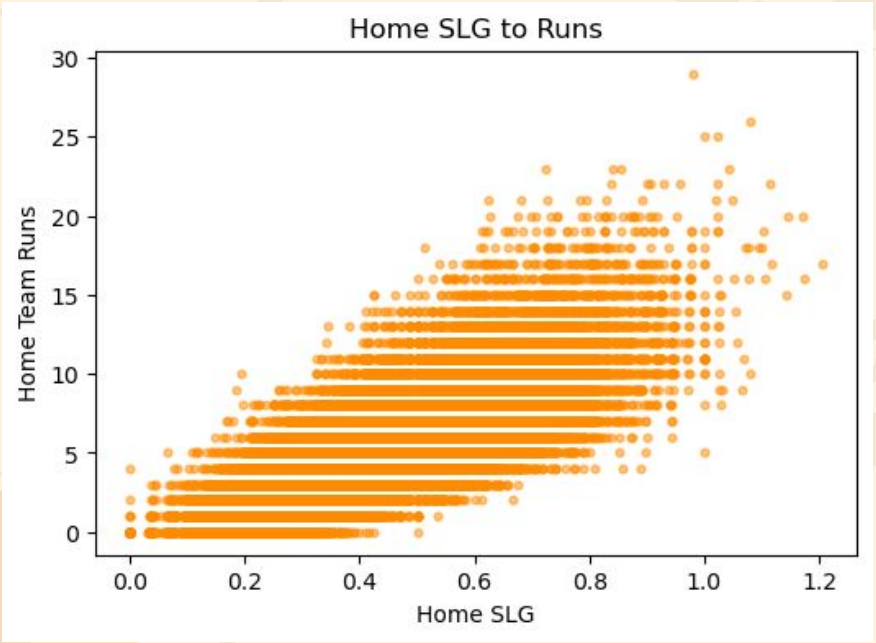
Ted Williams
OBP .482

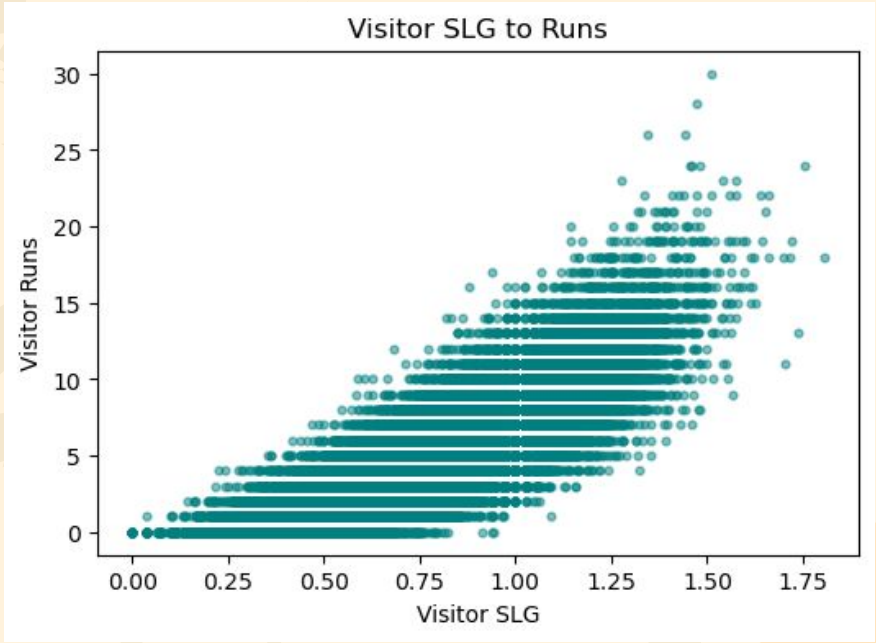




SLG to Runs

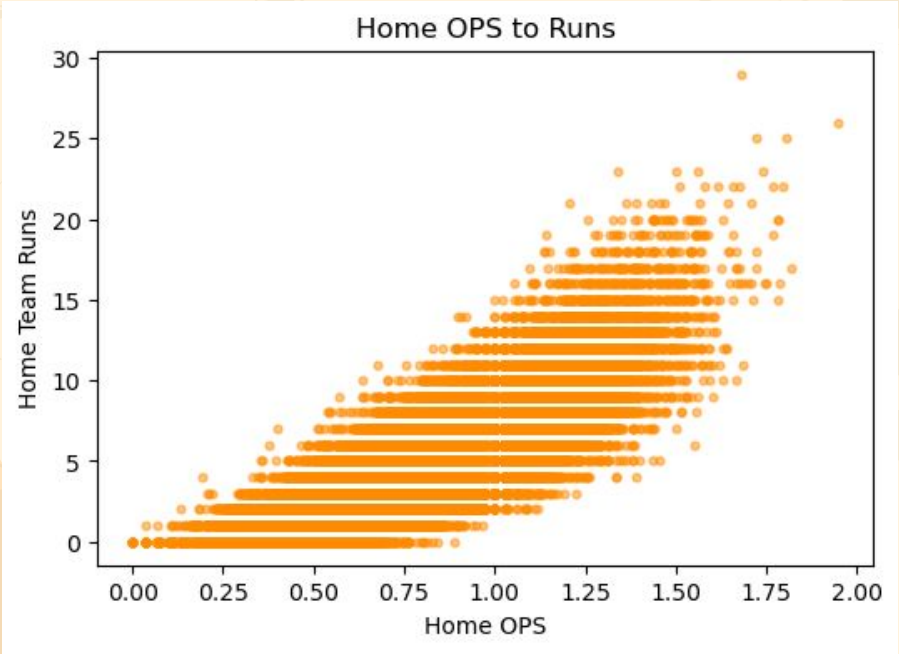
Babe Ruth
SLG .6897

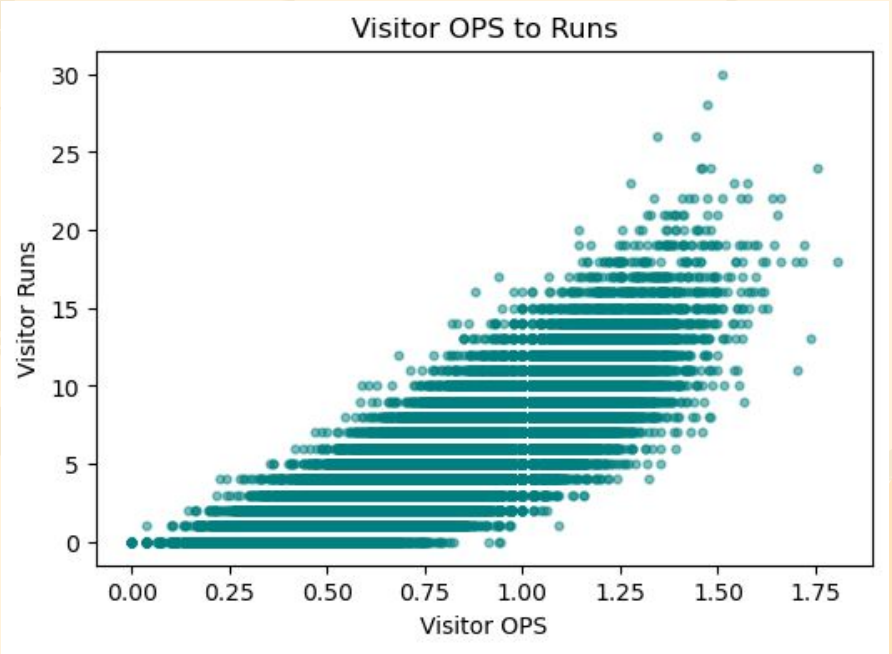




OPS to Runs

Babe Ruth
SLG 1.164



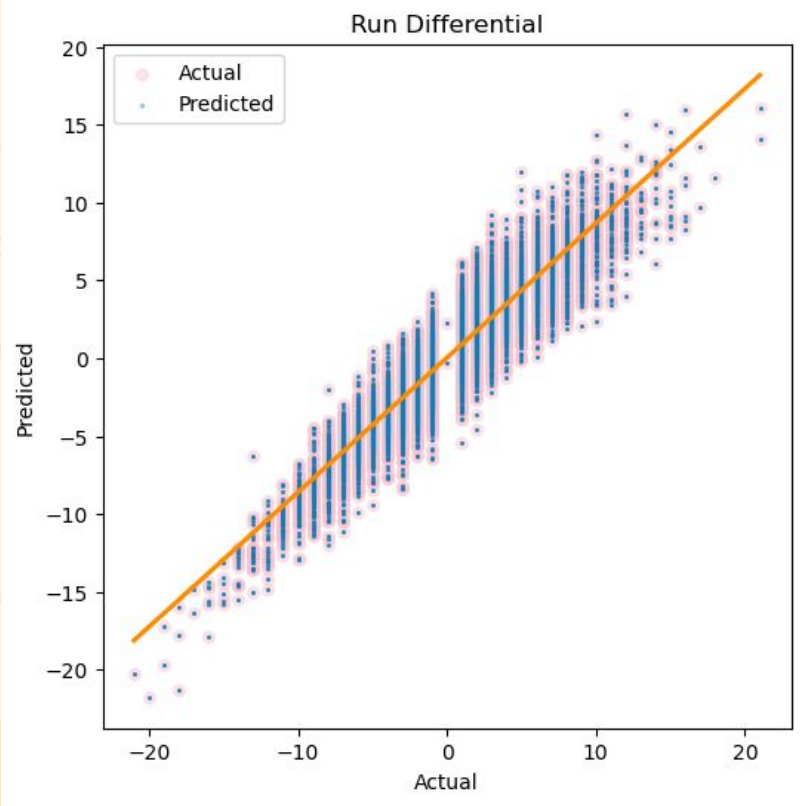


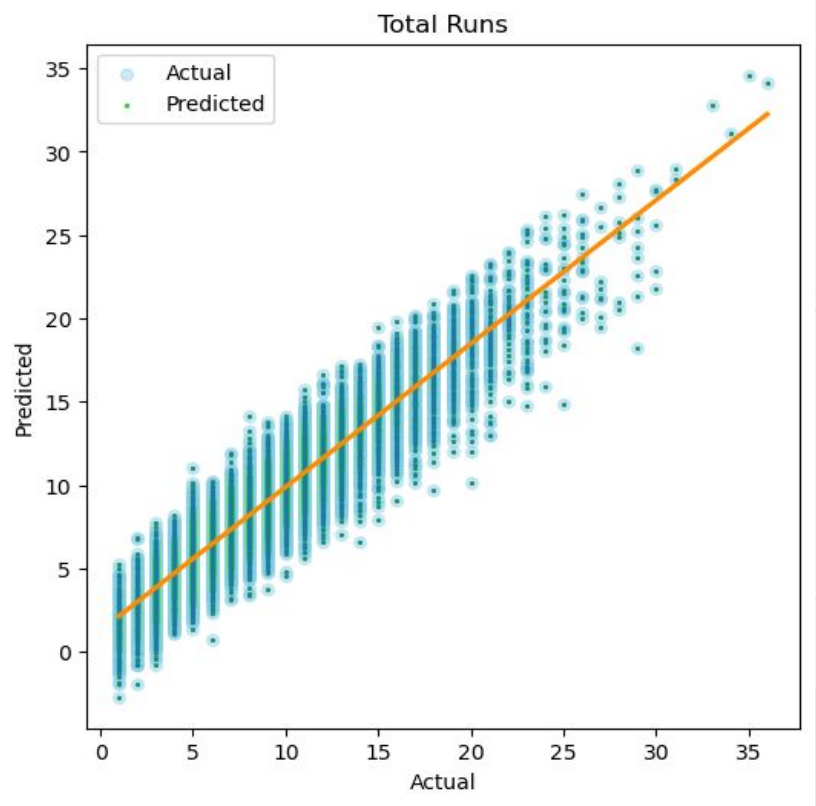
Modeling Process

Outcome Feature	Model	R^2 / Accuracy
Run Differential	Random Forest Regressor	86.3%
Total Runs	Linear Regression	86.4%
Win Loss	Random Forest Classifier	91.8%

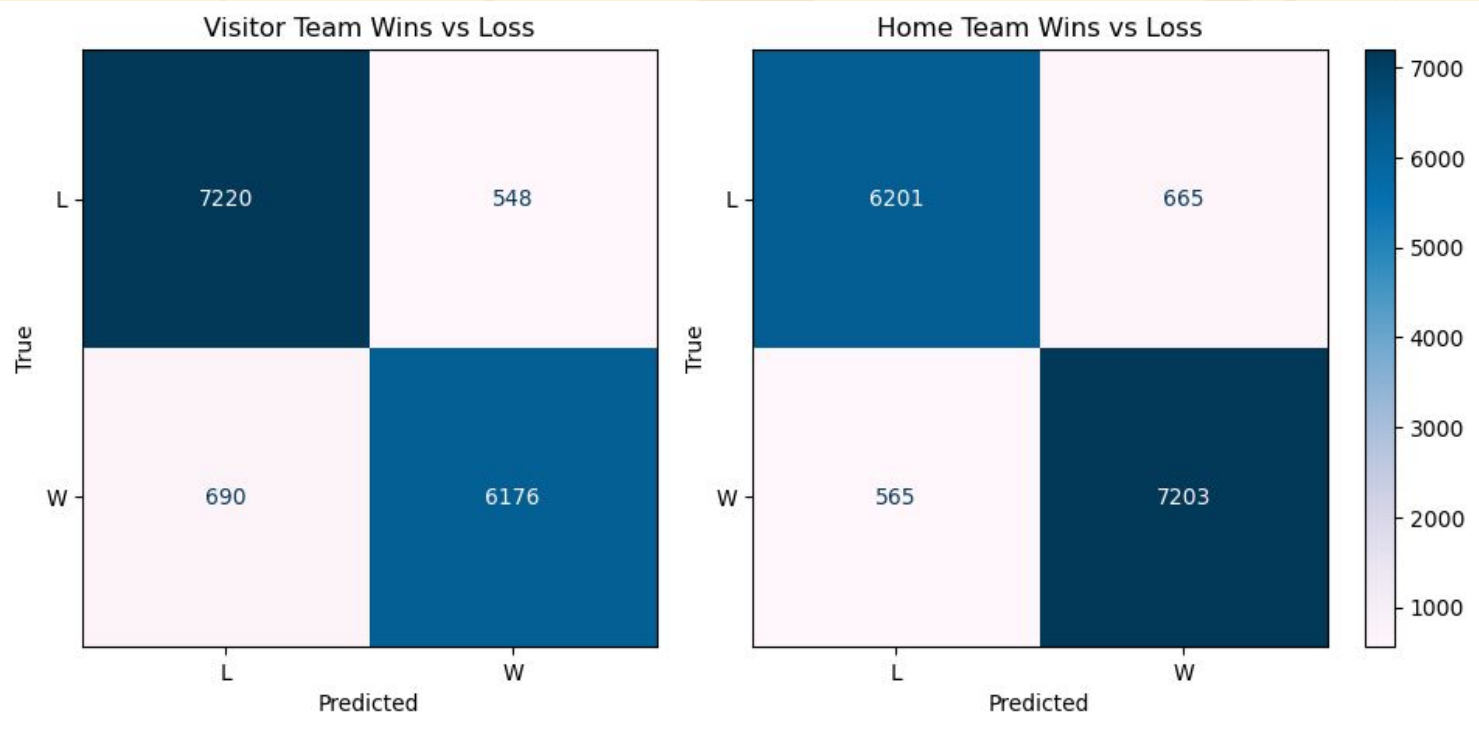
- Included other features like **RBIS, HRs, Stolen base success rate, runners left on base, strikeout ratio, walk ratio**.
- Some are from defense and pitching stats but very bland, **earned runs, errors, double plays**.
- Train data games from 1989-2000
- Test data games from 2001-2002
- Used One Hot Encoder for categorical columns. **Ballpark, Manager**. Didn't make much of a difference, focused on numerical data.
- Took out ties for classifier, only want Win or Loss.

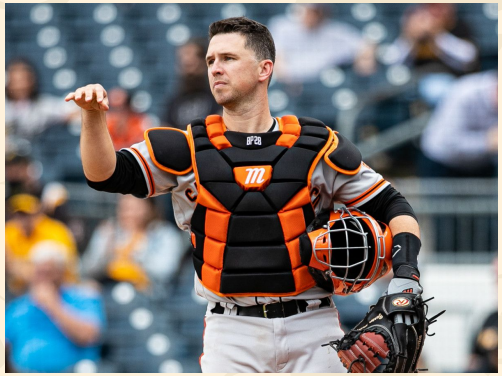
Model Evaluation





Win Loss Classifier





Streamlit!

Conclusion / Recommendations

- Confident models, but could tinker with more feature engineering to see if more complex ratios would make a bigger impact
- Engineer more pitching stats, WHIP, ERA, K/9, HR/9
- Incorporate individual player strength and how it changes outcomes
- Explore different models and preprocessors
- Keep updating the game data to have the most uptodate stats

Mariano Rivera
652 Career Saves

Sources / Shoutouts

- Data
  - Retrosheet www.retrosheet.org
  - Baseball Reference www.baseball-reference.com
  - MLB www.mlb.com/stats/
  - ESPN https://www.espn.com/
- Streamlit
- Chat GPT
- Shoutouts
  - Adam Boyd
  - DSI 0911
  - Rowan and Tim

Bruce Bochy
Legendary Manager

Thank You!