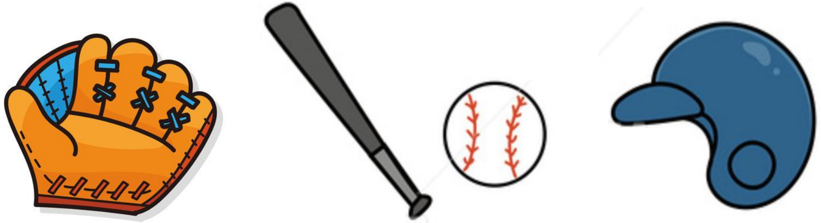
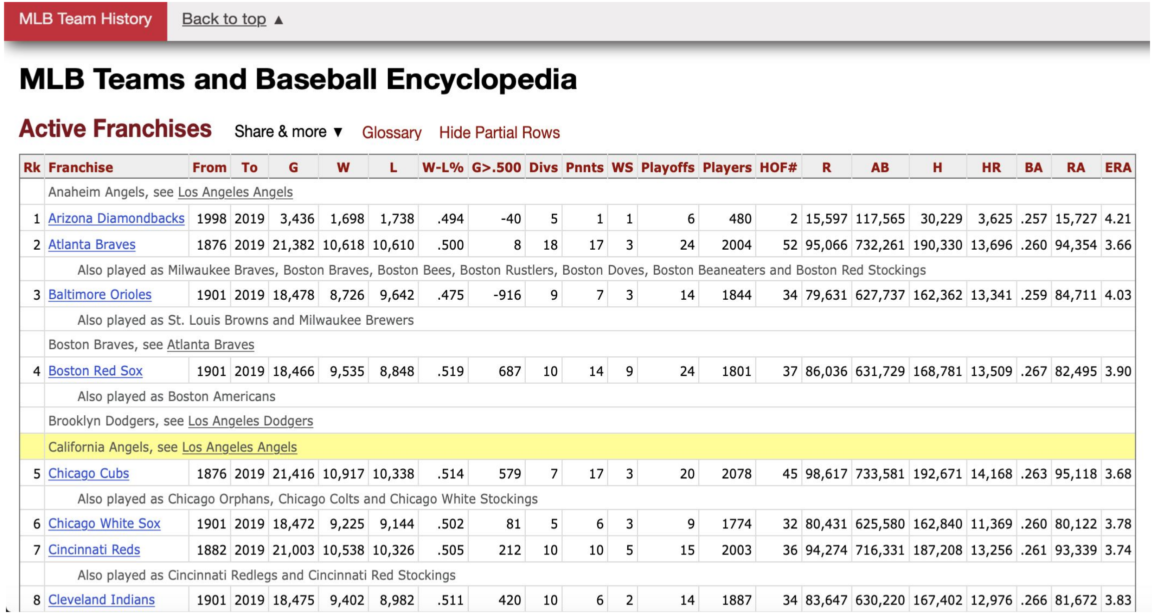
Analysis on MLB teams

Haorui Zhang | June 2019



I. Obtaining the Data

- Use rvest package in R.
- For each of the 30 Active Franchises, download the "Franchise History" table on each team's page and combine all of the tables into one.
- The resulted dataset has 2684 rows and 22 columns. Variables include wins, losses, ties, runs scored, runs allowed, batter's average age, pitcher's average age, etc.

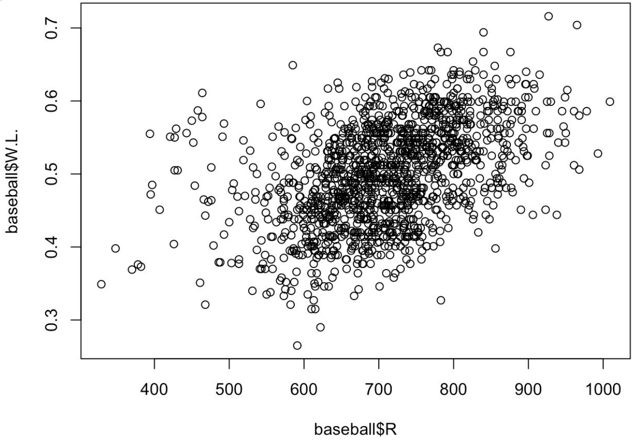


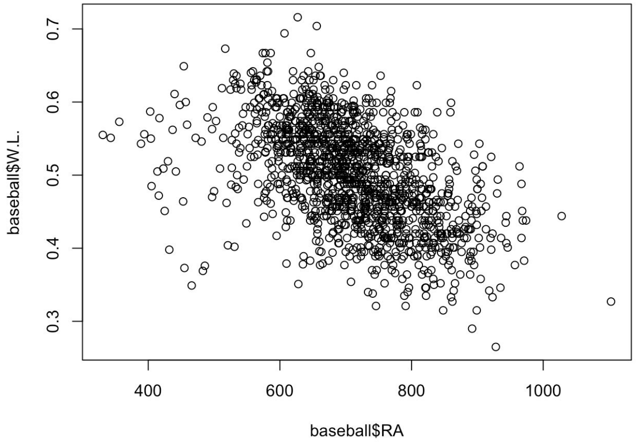
MLB Teams and Baseball Encyclopedia

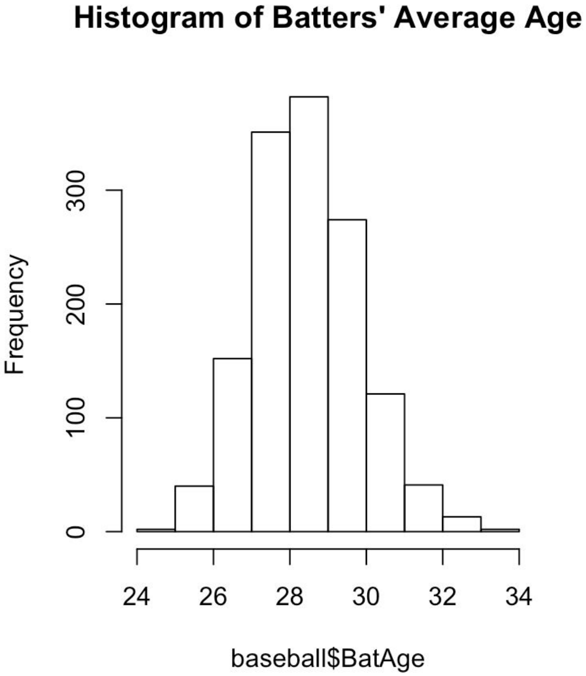
Active Franchises

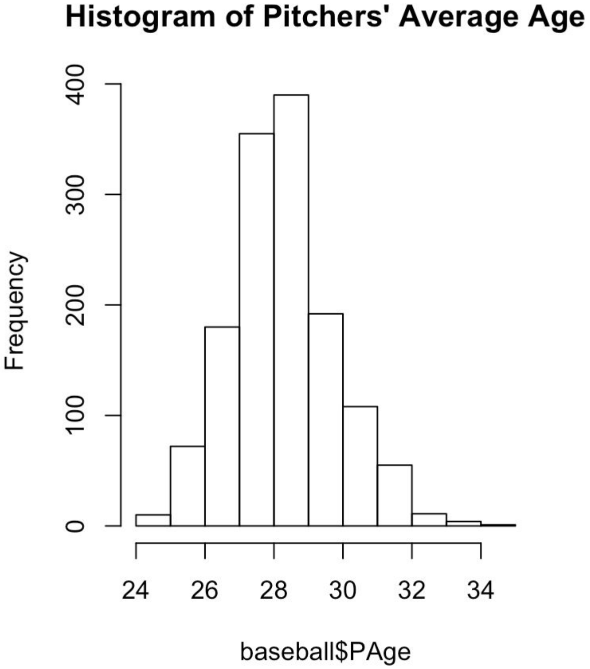
Rk	Franchise	From	To	G	W	L	W-L%	G>.500	Divs	Pnnts	WS	Playoffs	Players	HOF#	R	AB	H	HR	BA	RA	ERA
	Anaheim Angels, see Los Angeles Angels																				
1	Arizona Diamondbacks	1998	2019	3,436	1,698	1,738	.494	-40	5	1	1	6	480	2	15,597	117,565	30,229	3,625	.257	15,727	4.21
2	Atlanta Braves	1876	2019	21,382	10,618	10,610	.500	8	18	17	3	24	2004	52	95,066	732,261	190,330	13,696	.260	94,354	3.66
	Also played as Milwaukee Braves, Boston Braves, Boston Bees, Boston Rustlers, Boston Doves, Boston Beaneaters and Boston Red Stockings																				
3	Baltimore Orioles	1901	2019	18,478	8,726	9,642	.475	-916	9	7	3	14	1844	34	79,631	627,737	162,362	13,341	.259	84,711	4.03
	Also played as St. Louis Browns and Milwaukee Brewers																				
	Boston Braves, see Atlanta Braves																				
4	Boston Red Sox	1901	2019	18,466	9,535	8,848	.519	687	10	14	9	24	1801	37	86,036	631,729	168,781	13,509	.267	82,495	3.90
	Also played as Boston Americans																				
	Brooklyn Dodgers, see Los Angeles Dodgers																				
	California Angels, see Los Angeles Angels																				
5	Chicago Cubs	1876	2019	21,416	10,917	10,338	.514	579	7	17	3	20	2078	45	98,617	733,581	192,671	14,168	.263	95,118	3.68
	Also played as Chicago Orphans, Chicago Colts and Chicago White Stockings																				
6	Chicago White Sox	1901	2019	18,472	9,225	9,144	.502	81	5	6	3	9	1774	32	80,431	625,580	162,840	11,369	.260	80,122	3.78
7	Cincinnati Reds	1882	2019	21,003	10,538	10,326	.505	212	10	10	5	15	2003	36	94,274	716,331	187,208	13,256	.261	93,339	3.74
	Also played as Cincinnati Redlegs and Cincinnati Red Stockings																				
8	Cleveland Indians	1901	2019	18,475	9,402	8,982	.511	420	10	6	2	14	1887	34	83,647	630,220	167,402	12,976	.266	81,672	3.83

II. Exploratory Data Analysis









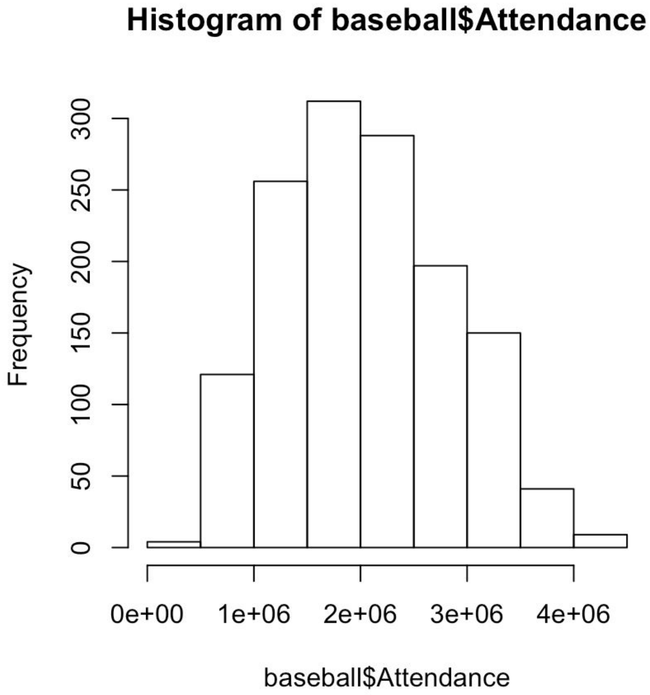
III. Feature Engineering

Outcome

- W.L.: Winning rate in decimals
  - ? Coefficients will be very small
  - 💡 Multiply by 100 to make it in percentage form

Predictor

- Attendance: number of tickets sold in home games
  - ? Very large values, coefficient will be very small
  - 💡 Sort it into four categories to indicate the popularity of a team
    - ➔ Resulted variable: Popularity which has four levels: Very unpopular, Unpopular, Popular, Very popular



IV. Modeling

- Multiple Linear Regression Model
- Started with the full model, then removed insignificant predictors.
- Our final model explains about 91% of the variations in the winning rates.

Predictors		Coefficient Estimate
LgAL East	League: AL East	0.601469 (base = NL Central)
LgAL West	League: AL West	0.106234
LgAL Central	League: AL Central	-0.086487
LgNL East	League: NL East	0.332738
LgNL West	League: NL West	0.083508
GB	Games back of league leader	-0.213373
R	Total runs scored	0.042761
RA	Total runs allowed	-0.042021
PopularityUnpopular	Popularity (based on number of tickets sold in home games)	0.700144 (base = Very unpopular)
PopularityPopular	Popularity (based on number of tickets sold in home games)	1.052674 (base = Very unpopular)
PopularityVery popular	Popularity (based on number of tickets sold in home games)	1.102880 (base = Very unpopular)
PAge	Pitchers' average age	-0.086416
X.Bat	Number of batters	-0.034715

Multiple R-squared	0.9101	Correlation	0.9546554

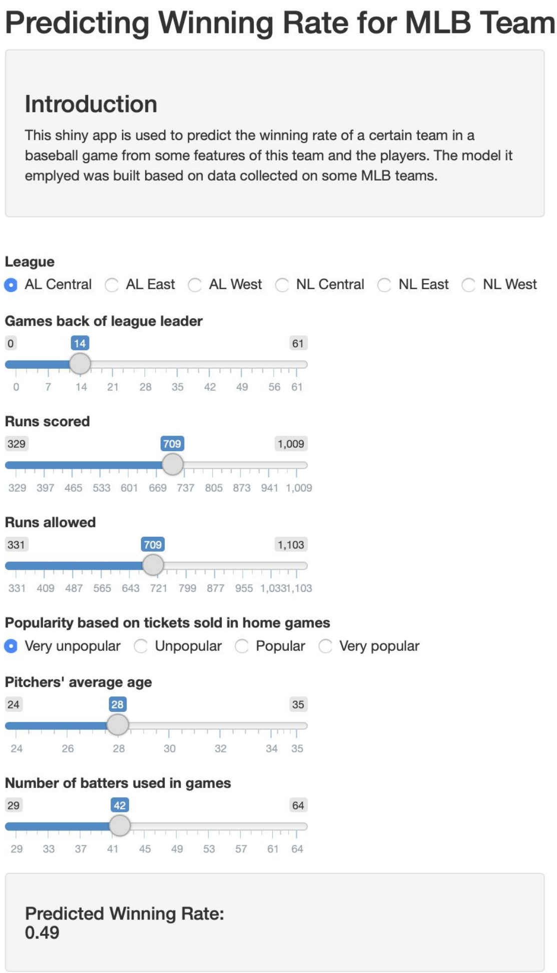
V. Deployment

- Using Shiny app

- Nice interface, get predictions simply by adjusting sliders and radio buttons

- Input: Desired team features
- Output: Predicted winning rate

- Can be deployed locally in terminal or simply through this url: https://mysticcc.shinyapps.io/Predicting_Winning_Rate_for_MLB_Teams/



Predicting Winning Rate for MLB Teams

Introduction

This shiny app is used to predict the winning rate of a certain team in a baseball game from some features of this team and the players. The model it emplyed was built based on data collected on some MLB teams.

League

● AL Central ○ AL East ○ AL West ○ NL Central ○ NL East ○ NL West

Games back of league leader: 14 (0–61)

Runs scored: 709 (329–1,009)

Runs allowed: 709 (331–1,103)

Popularity based on tickets sold in home games

● Very unpopular ○ Unpopular ○ Popular ○ Very popular

Pitchers' average age: 28 (24–35)

Number of batters used in games: 42 (29–64)

Predicted Winning Rate:
0.49