What Makes a Winning Baseball Team?

Project 2 Report - W200 Section 5

December 9, 2022

Team Members: Dave Zack, Gray Selby, and Jason Komorowski

Github Repository: https://github.com/UC-Berkeley-I-School/Project2_Komorowski_Selby_Zack

Primary Dataset: Major League Baseball 2017-2022 Team data - 1) hitting, 2) fielding, and 3) pitching

https://www.baseball-reference.com/leagues/majors/2022.shtml

Overview:

Data science is revolutionizing many industries, including professional sports. Gone are the days when a handful of basic statistics were used to explain, measure, and predict performance. Major League Baseball was one of the first sports to join the data revolution and the early adopters had a considerable advantage. This caused an avalanche of new analytical methods and statistics to be adopted and the collection of additional data to expand. Where there used to be a few simple statistics such as “errors” to measure fielding performance, now there is advanced data tracking the reaction time of a fielder and their likelihood of catching a ball based on the trajectory of the ball off the bat while adjusting for variables such as ballpark dimensions. With all of this new data, the question remains, what are the characteristics of a winning baseball team, which is interesting for many sports professionals and fans, ourselves included.

Utilizing the online database from Baseball Reference we were able to determine which team batting, pitching, and fielding performance metrics correlate to more wins, and therefore what statistics should be focused on when building a baseball team. The analysis supported the idea that batting and pitching are more important than fielding overall, and also that On-Base plus Slugging Plus (OPS+) and Earned Run Average (ERA) are of high importance within the subcategories. The results generally held year over year.

Questions:

The overall question is, what characteristics make a winning baseball team? This was broken down into three sub questions which were investigated.

1. Overall, is batting, fielding, or pitching more important to winning?
2. Deeper into each category, which subcategories and combinations of statistics within each area leads to more wins?
3. Are the characteristics for a winning team consistent year to year?

For this analysis, having more regular season wins and ultimately making the playoffs are what we define as a “winning” team.

Dataset:

The dataset used was from Baseball Reference and taken from 2017 through 2022. The ".csv" file for hitting, fielding, and pitching were downloaded and then read into the notebook for exploration. Table 1 (hitting), Table 2 (Fielding) and Table 3 (pitching) are the first 5 rows of 2022 full season team data from the three primary datasets analyzed as an example of the initial data exploration and validation.

Table 1: Team Hitting - 2022

	Tm	#Bat	BatAge	R/G	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	LOB
0	Arizona Diamondbacks	57	26.5	4.33	162	6027	5351	702	1232	262	24	173	658	104	29	531	1341	0.230	0.304	0.385	0.689	95	2061	97	60	31	50	14	1039
1	Atlanta Braves	53	27.5	4.87	162	6082	5509	789	1394	298	11	243	753	87	31	470	1498	0.253	0.317	0.443	0.761	111	2443	103	66	1	36	13	1030
2	Baltimore Orioles	58	27.0	4.16	162	6049	5429	674	1281	275	25	171	639	95	31	476	1390	0.236	0.305	0.390	0.695	97	2119	95	83	12	43	10	1095
3	Boston Red Sox	54	28.8	4.54	162	6144	5539	735	1427	352	12	155	704	52	20	478	1373	0.258	0.321	0.409	0.731	102	2268	131	63	12	50	23	1133
4	Chicago Cubs	64	27.9	4.06	162	6072	5425	657	1293	265	31	159	620	111	37	507	1448	0.238	0.311	0.387	0.698	96	2097	130	84	19	36	16	1100

Table 2: Team Fielding - 2022

	Tm	#Fld	RA/G	DefEff	G	GS	CG	Inn	Ch	PO	A	E	DP	Fld%	Rtot	Rtot/yr	Rdrs	Rdrs/yr	Rgood
0	Arizona Diamondbacks	57	4.57	0.704	162	1458	1097	12870	5746	4290	1370	86	134	0.985	34	3	55	-1	-3
1	Atlanta Braves	53	3.76	0.701	162	1458	1195	13032	5803	4344	1382	77	110	0.987	16	1	31	2	4
2	Baltimore Orioles	57	4.25	0.690	162	1458	1121	12900	5920	4300	1529	91	151	0.985	6	1	38	2	5
3	Boston Red Sox	53	4.86	0.683	162	1458	1105	12879	5825	4293	1447	85	134	0.985	-7	-1	-2	-2	-8
4	Chicago Cubs	64	4.51	0.697	162	1458	1086	12993	5880	4331	1453	96	139	0.984	1	0	2	0	-1

Table 3: Team Pitching - 2022

	Tmâ-²	#P	PAge	RA/G	W	L	W-L%	ERA	G	GS	GF	CG	tSho	cSho	SV	IP	H	R	ER	HR	BB	IBB	SO	HBP	BK	WP	BF	ERA+	FIP	WHIP	H9	HR9	BB9	SO9	SO/W	LOB
0	Arizona Diamondbacks	33	30.0	4.57	74	88	0.457	4.25	162	162	162	0	10	0	33	1430.0	1345	740	676	191	504	18	1216	59	3	51	6065	95	4.33	1.293	8.5	1.2	3.2	7.7	2.41	1051
1	Atlanta Braves	31	30.0	3.76	101	61	0.623	3.46	162	162	161	1	9	1	55	1448.0	1224	609	556	148	500	21	1554	62	4	55	6031	118	3.46	1.191	7.6	0.9	3.1	9.7	3.11	1101
2	Baltimore Orioles	35	27.7	4.25	83	79	0.512	3.97	162	162	160	2	15	1	46	1433.1	1406	688	632	171	443	8	1214	64	4	47	6058	102	4.03	1.290	8.8	1.1	2.8	7.6	2.74	1092
3	Boston Red Sox	32	30.2	4.86	78	84	0.481	4.53	162	162	157	5	10	2	39	1431.0	1411	787	721	185	526	17	1346	72	8	60	6167	93	4.17	1.354	8.9	1.2	3.3	8.5	2.56	1109
4	Chicago Cubs	42	29.5	4.51	74	88	0.457	4.00	162	162	162	0	11	0	44	1443.2	1342	731	642	207	540	19	1383	73	8	53	6162	103	4.33	1.304	8.4	1.3	3.4	8.6	2.56	1130

Data Validation and Preparation:

The initial dataset we looked at was just from 2022. Each table above (Table 1, Table 2, and Table 3) shows the number of games as 162 and lists all 30 teams which is what is expected. Additional spot checks were performed on the data to ensure it was clean, and no issues were found, which was expected. We used a very pristine and reliable dataset.

Next it was confirmed that the team's name is listed the same in each of the datasets, so the tables could be joined using the team name as the index. Before the tables were joined, the team names were abbreviated such that they would be easier to display on charts and figures. The other step that was completed before joining the tables was to add a suffix to each column name to represent if it came from batting, pitching, or fielding. This eliminated the confusion between variables in the tables after they were joined. The reason for this is that HR for batting refers to the amount of homeruns hit by an individual team while HR for pitching refers to the amount of homeruns given up by an individual team. The suffixes allowed us to easily distinguish between categories such as this and also allowed us to join the tables effectively.

Next, we created a list of playoff teams for each season we analyzed as the baseline for the definition of a "winning team". There were 10 playoff teams in 2017-2019, and 2021. There were 14 playoff teams in

2020 due to a one-time change based on the effects of the COVID pandemic. There were 12 playoff teams in 2022 due to a newly implemented rule increasing the number of playoff teams.

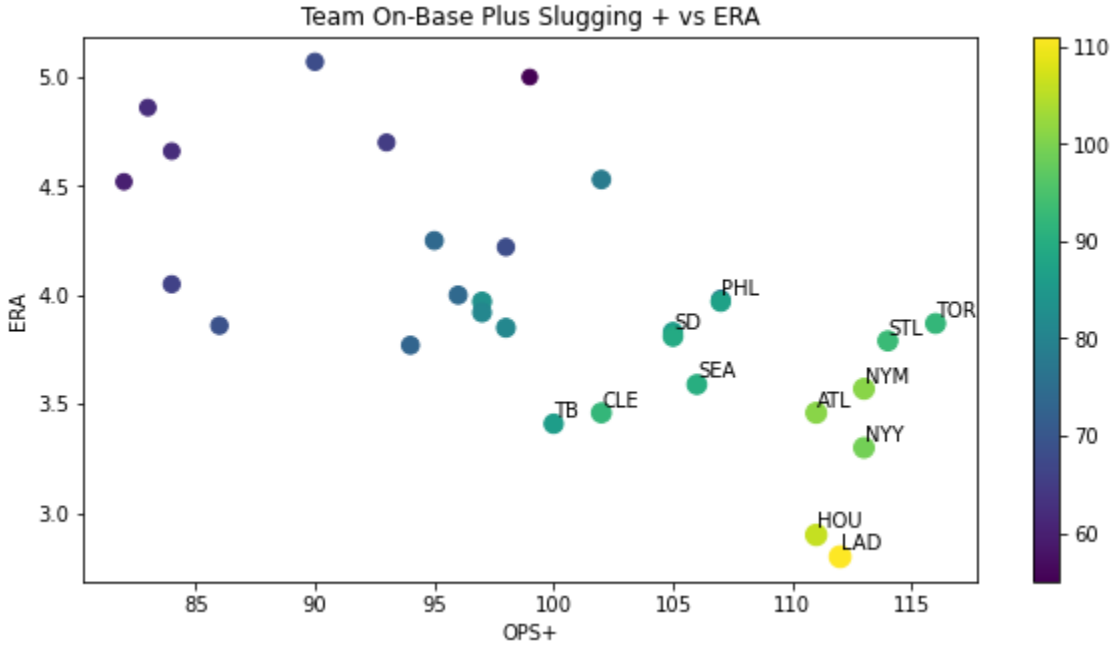
Exploration:

After the data was validated, formatted, and joined, we began our exploration. There are many variables listed that can be analyzed, so we explored the meaning of each one to make sure we fully understood each. We then made the decision to focus on the more basic statistics for our initial analysis, because the more complex statistics are often a combination of the simpler metrics. Some of the variables we focused on were:

- Batting: Tm = team, R/G = runs per game, BB = walks, BA = batting average, SO = strikeouts, HR = home runs, OBP = on base percentage, SLG = slugging, OPS = on base plus slugging, OPS+ = on base plus slugging plus
- Pitching: Tm = team, W = wins, L = losses, RA/G = runs against per game, ERA = earned run average, HR = home runs, BB = walks, SO = strikeouts, WHIP = (walks + hits) per inning pitched, SO/W = strikeouts per win
- Fielding: Tm = team, RA/G = runs against per game, DeffEff = defense efficiency, A = assists, E = errors, DP = double plays, Fld% = fielding percentage, Rdrs = defensive runs saved

Based on these choices of variables we began creating scatter plots with the playoff teams highlighted to look for patterns. Figure 1 below shows one of the more compelling graphs that shows that the combination of low pitching earned run average and high on-base plus slugging lead to making the playoffs.

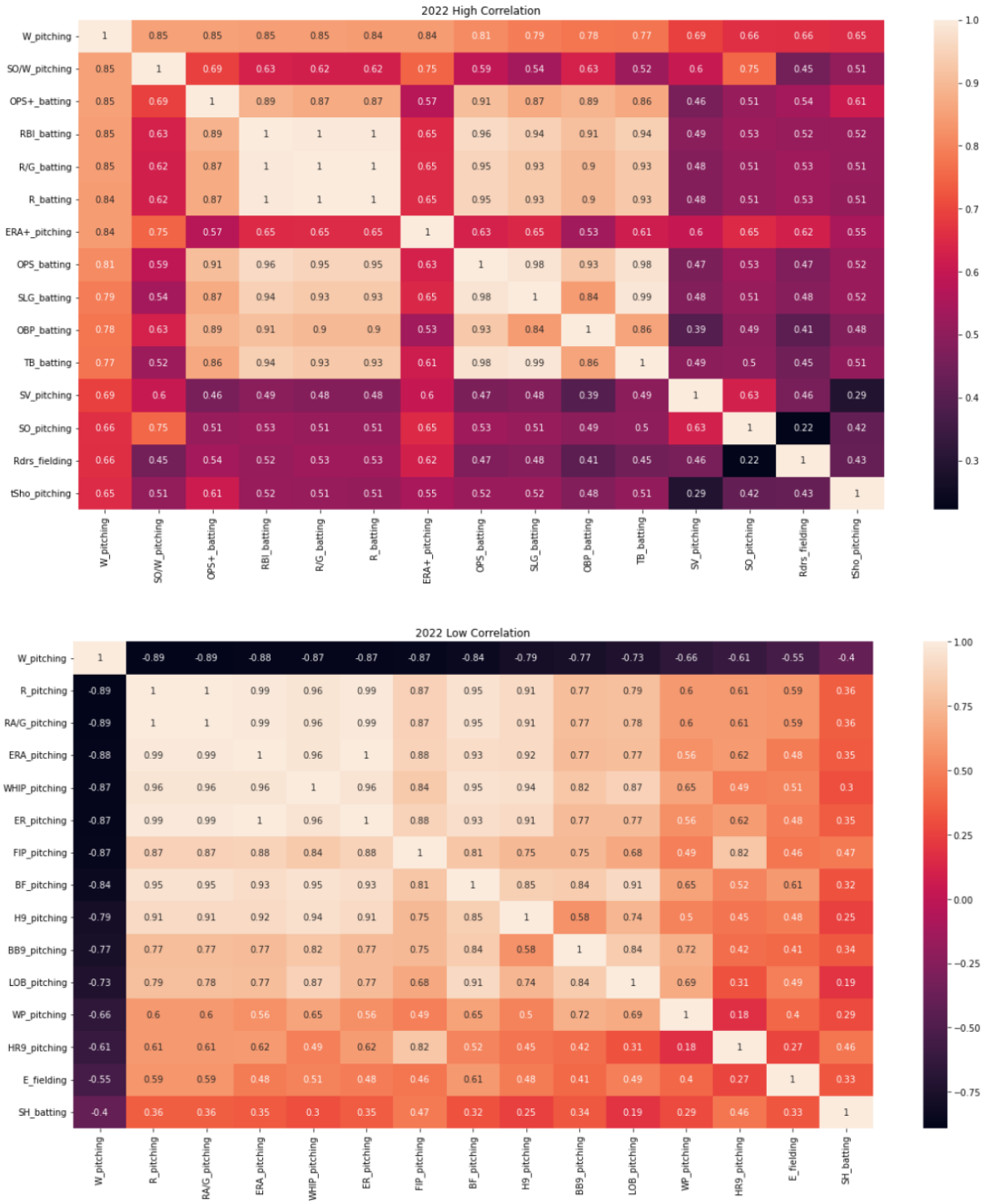
Figure 1: 2022 Team OPS+ vs ERA (colorbar corresponds to wins)



The next step was to create a heat map of the statistics and determine how well they correlate with wins. We created a high correlation and a low correlation chart due to some statistics being better as

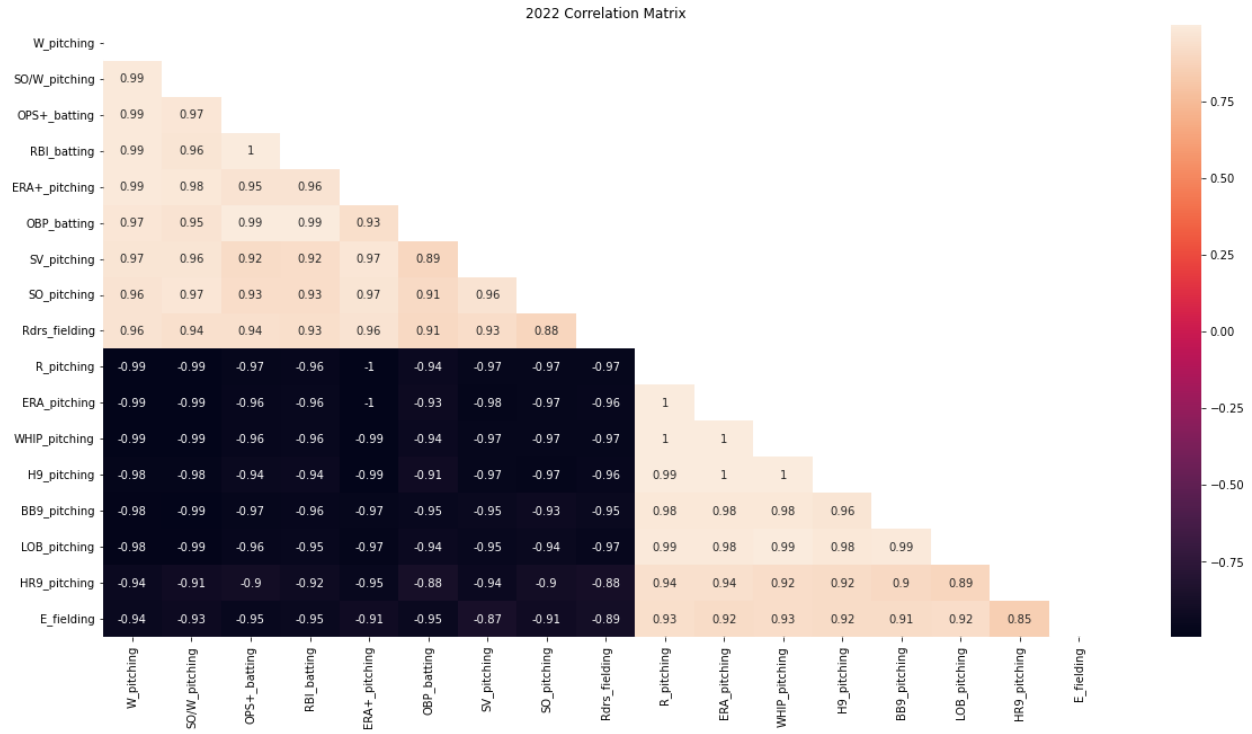
they increase and some that are better as they decrease. The top 15 variables for both positive and negative correlation were added to the heatmaps below as shown in figure 2.

Figure 2: 2022 High and Low Correlation Matrices (Focus on W_pitching)



We then looked more closely at the variables that showed high correlation and eliminated some from consideration because they were duplicate or similar variables with different names. For example, “RA/G_pitching”represents the runs per game allowed when pitching and “ERA_pitching” represents the earned run average per game, and “ER_pitching” represents earned runs given up per game. All three measure almost exactly the same thing, so only one of them needs to be focused on for further exploration. In addition, only one half of the heat map is needed as the other half is a mirror image. This process resulted in the following heat map with the redundant or non-important statistics removed, so it is a different top 15.

Figure 3: Updated Correlation Matrices with Updated Values



When we subsetted the correlation tables, the values changed, which we are not sure exactly how to explain. A next step would be to look further into this, and determine if the cause is something in the code or our understanding of how it is implemented. This does not affect our overall findings but would be interesting to understand.

The final data frame of just the important characteristics is shown below in table 4.

Table 4: Final Dataframe with important characteristics

Tm	W_pitching	SO/W_pitching	OPS+_batting	RBI_batting	ERA+_pitching	OBP_batting	SV_pitching	SO_pitching	Rdrs_fielding	R_pitching	ERA_pitching	WHIP_pitching	H9_pitching	BB9_pitching	LOB_pitching	HR9_pitching	E_fielding
AZ	74	2.41	95	658	95	0.304	33	1216	55	740	4.25	1.293	8.5	3.2	1051	1.2	86
ATL	101	3.11	111	753	118	0.317	55	1554	31	609	3.46	1.191	7.6	3.1	1101	0.9	77
BAL	83	2.74	97	639	102	0.305	46	1214	38	688	3.97	1.290	8.8	2.8	1092	1.1	91
BOS	78	2.56	102	704	93	0.321	39	1346	-2	787	4.53	1.354	8.9	3.3	1109	1.2	85
CHC	74	2.56	96	620	103	0.311	44	1383	2	731	4.00	1.304	8.4	3.4	1130	1.3	96

From the subsetted table we wanted to look at the variance of the remaining columns of interest. Variance, a measure of spread, is the average squared distance between the mean and each element. Formally:

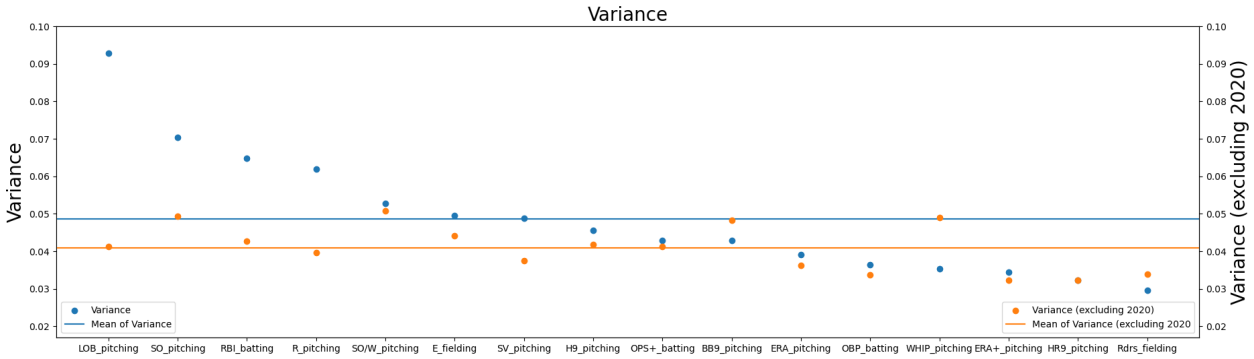
$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n - 1}$$

First we must normalize each column so that they share the same scale between zero and one. This is accomplished formally:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

The variance is calculated for the columns of interest across the 2017 through 2022 seasons. A high variance means that the data is relatively spread out. In order to understand how to distinguish a winning team from a losing team, we are interested in columns that separate teams. A column with nearly the same value for every team is not especially useful in differentiating winning and losing teams. We do the same operation except this time we exclude data from the 2020 season which we suspect to contain outliers due to the unprecedented nature of the period.

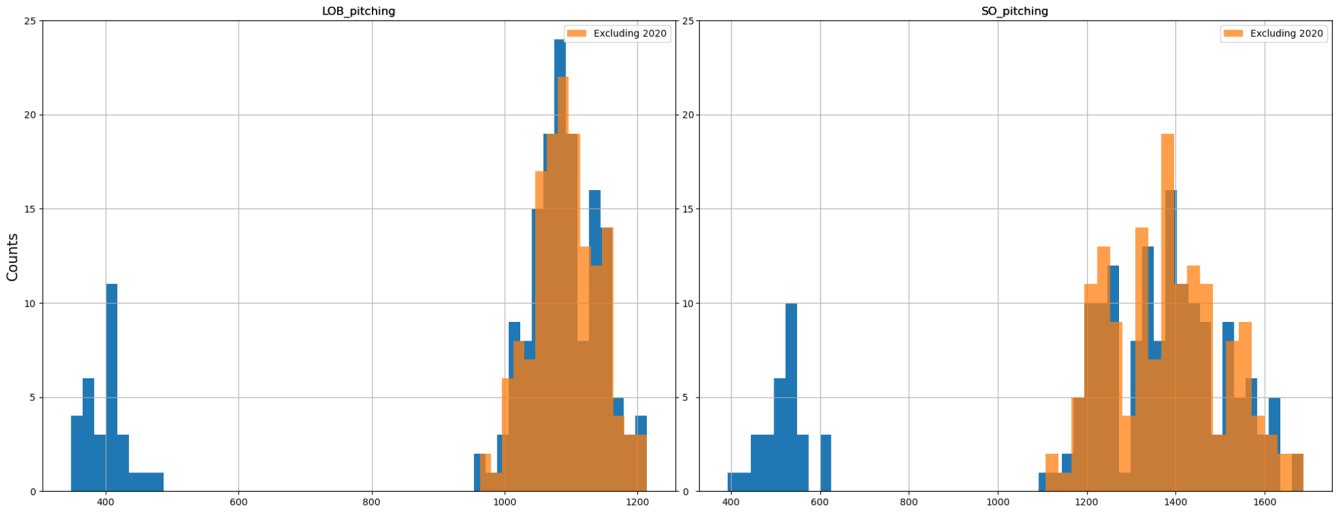
Figure 4: Variance of Key Variables with and without 2020 Included



We find that the 2020 season does appear to contain relative outliers as the mean variance is greater with the 2020 season. It appears that the 2020 season contains outliers especially for the columns: LOB_pitching, SO_pitching, RBI_batting, and R_pitching.

To investigate further into the differences between the 2020 season from the other previous 6 seasons, we examine the distribution of LOB_pitching and SO_pitching using histograms.

Figure 5: Histograms of Left on Base and Strikeouts (Pitching)

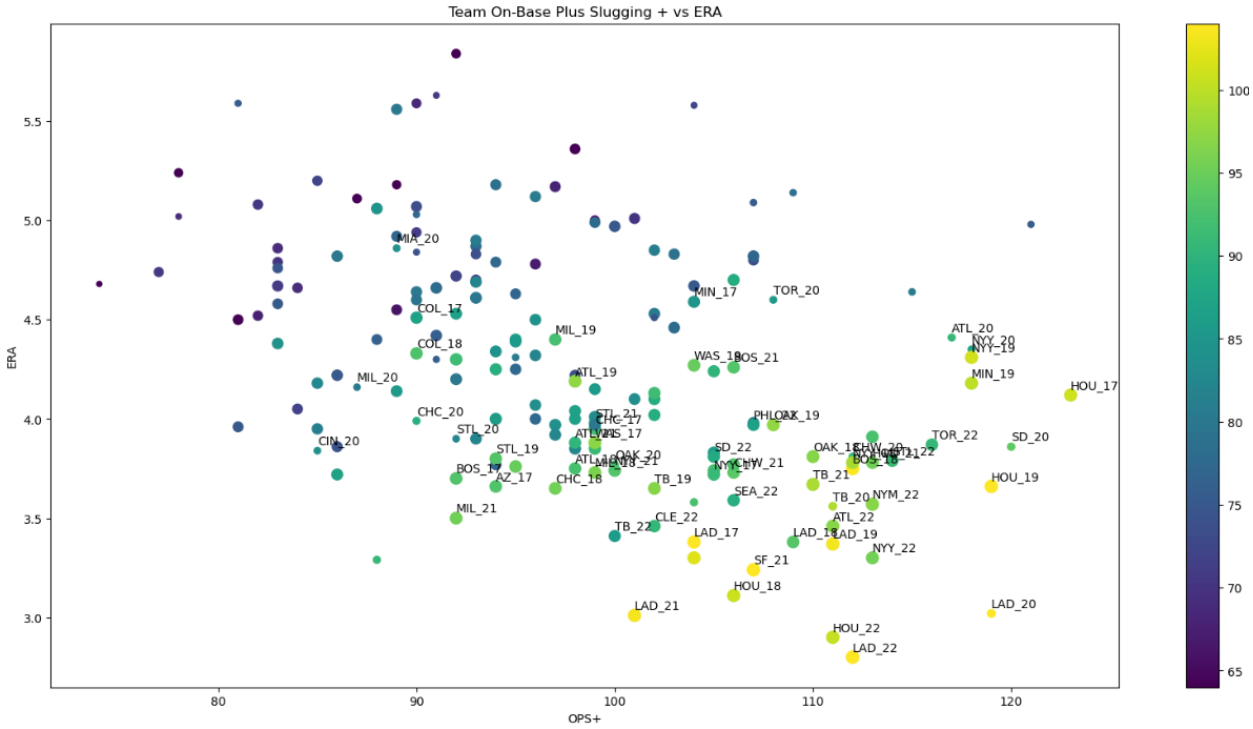


The 2020 season indeed contains significant outliers of lower LOB_pitching and SO_pitching. The same is true for RBI_batting and R_pitching but no completely separate group of data is found in any other column of interest. Of note, 2020 was the COVID year so the counting stats referenced above were not across 162 games but only 60. It makes sense to exclude this season because of that fact.

Next we did a sanity check of our results to ensure that the metrics make sense. Winning a baseball game, like many sports, breaks down to simply scoring more than the other team. ERA is a strong metric tied to reducing runs scored for the opposing team, while OPS+ is a good indicator of not just getting on base, but getting further around the base path, which leads to more runs scored. It is interesting to see that pitching was much more important than fielding, given they both are a part of stopping runs, and most outs are not from a strikeout, so the fielders make most of the outs. However, the pitcher does affect how well the hitter hits the ball, so a good pitcher can make it easier for the fielders to field the ball, which corresponds to pitching being more important. Reducing runs and adding runs are obvious characteristics. Further analysis could attempt to determine what are important traits on a player's skill, like hitting home runs, giving up strikeouts, etc. These are the characteristics that are more in control of management as they search for the right players to target.

Finally, the above explorations were repeated for years 2017-2021 to compare to 2022 to determine if the characteristics for a winning team are consistent year to year. Below in figure 6, it can be seen that year over year the winning teams share a high OPS+ and a low ERA, so the results of 2022 generally hold for the other years. There are a few outliers in 2020 , which was the COVID adjusted year where more teams were allowed into the playoffs than typical, so the threshold to make the playoffs was lowered. That being said, compared to the counting statistics shown above such as left on base and number of strikeouts, OPS+ and ERA are normalized statistics that should still represent a winning baseball team even in the shortened season.

Figure 6: OPS+ vs ERA Years 2017-2022 with Playoff Team Annotations (colorbar corresponds to wins)



As a final sanity check, we cross referenced the teams that made the playoffs with the ultimate World Series winner. We found that generally, the important characteristics in the regular season match those that are important in the playoffs. Outliers did not typically win the world series.

Our exploratory data analysis did not come without any hiccups though. We ran into merge issues with Jupyter notebooks and had to repeatedly save new versions to our github repository. To combat this a bit, we moved most of our functions into a .py file for reproducibility and ease of access.

Assumptions, Caveats, and Biases:

- **Assumptions:**
  - Data is accurate and complete
  - Removing variables that are deemed less important is appropriate
  - Winning is defined as having more wins and ultimately making the playoffs
- **Caveats:**
  - 2020 is the COVID year with only 60 games played so we removed it from some of our variance calculations and histogram plots
- **Biases:**
  - Jason and Dave are baseball fans and had preconceived notions about what constitutes a good baseball team
  - To combat this, we had Gray do some exploration on his own and come up with findings

Conclusion:

Through our exploration the data supported the idea that pitching and hitting are the most important characteristics when building a winning team. In addition, the subcategories of ERA and OPS+ closely correlated to winning teams. Finally, the characteristics that were important in the regular season were also important in the playoffs, and were likely to lead to a world series championship.

Appendix/Future Work

One might want to create a model that can separate teams that made it to the playoffs from teams that did not make it to the playoffs based on statistics without considering wins. The following columns are considered:

Table 5: Dataframe of features considered

Tm	OPS+_batting	RBI_batting	ERA+_pitching	OBP_batting	SV_pitching	SO_pitching	Rdrs_fielding	R_pitching	ERA_pitching	WHIP_pitching	H9_pitching	BB9_pitching	LOB_pitching	HR9_pitching	E_fielding
AZ	95	658	95	0.304	33	1216	55	740	4.25	1.293	8.5	3.2	1051	1.2	86
ATL	111	753	118	0.317	55	1554	31	609	3.46	1.191	7.6	3.1	1101	0.9	77
BAL	97	639	102	0.305	46	1214	38	688	3.97	1.290	8.8	2.8	1092	1.1	91
BOS	102	704	93	0.321	39	1346	-2	787	4.53	1.354	8.9	3.3	1109	1.2	85
CHC	96	620	103	0.311	44	1383	2	731	4.00	1.304	8.4	3.4	1130	1.3	96

There are many established techniques that we could utilize to create such a model. To stay with the theme of exploring the structure of the data itself, we explored leveraging the dimension (feature column) reduction technique known as Principal Component Analysis (PCA). PCA establishes a linear transformation of the data that best preserves the variance of the data. Put simply, PCA transforms the data optimally to separate the data with the fewest dimensions. Each resulting transformed dimension is called a principal component, with the first principal component enabling the greatest separation of the data and subsequent principal components being orthogonal to the first. We will focus on the first two principal components so that we can visualize how the higher dimensional data is clustered. PCA requires that the columns be z-score standardized so that each has a mean of zero and a standard deviation of 1 as seen in the equation below:

$$Z_{standardized} = \frac{x - \mu_{mean}}{\sigma_{std}}$$

As before, we will examine findings with data from seasons 2017-2022 as well as excluding 2020.

Figure 7: The variance accounted for per principal component

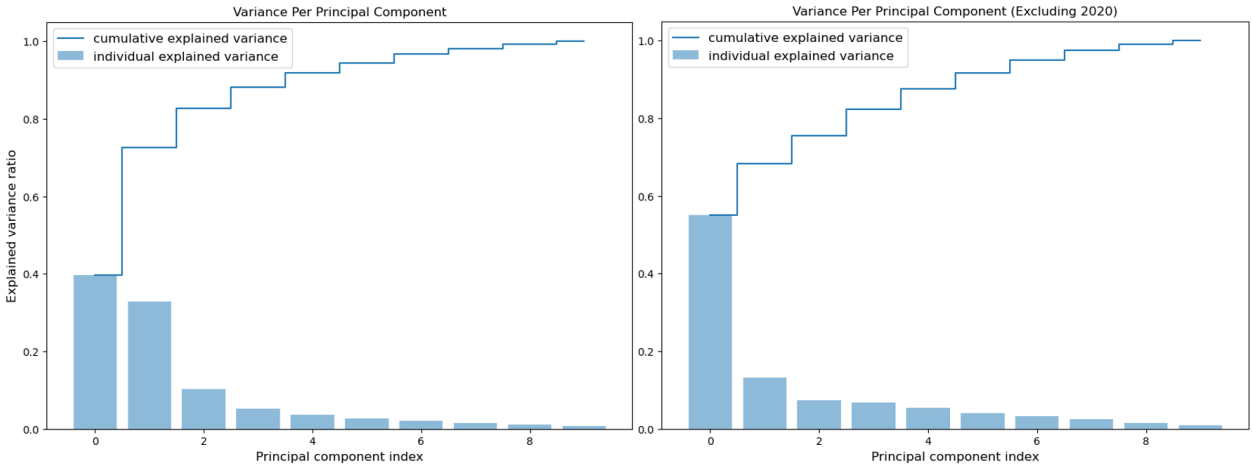
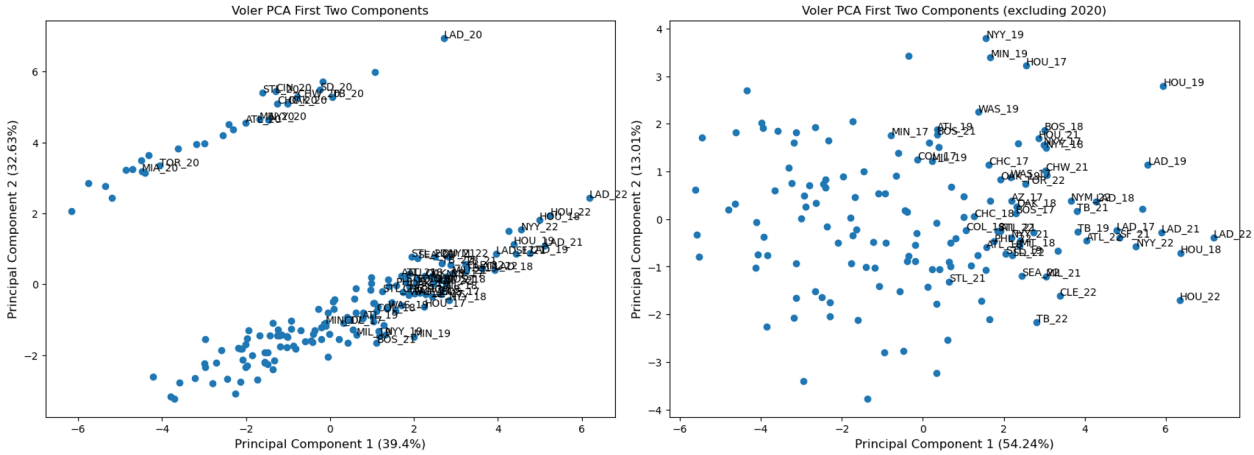


Figure 8: The first two principal components



Given our previous findings, the outliers presented by the 2020 season result in a more balanced variance for the first two principal components. Plotting the first two principal components, we find that they very effectively separate the 2020 teams from the rest, although a linear discriminator somewhat effectively distinguishes playoff and non playoff teams. Looking at the plots excluding the 2020 season, we see that the first principal component now accounts for more than half the variance in the transformed dataset. A more effective linear discriminator is now possible. This serves as one example of how a model could begin to be constructed using the structure of the data.

Citations

1. https://www.baseball-reference.com/leagues/majors/2022.shtml
2. https://statcorner.com/mlb/stats/ops-and-ops-plus
3. https://www.mlb.com/stats/