The Most Valuable Baseball Players, Real World and Fantasy

By Griffin Miller

Summary

Real World Baseball:

1. What statistics are most correlated with WAR, an ambiguous stat which is often times used to represent how good a player is? Least correlated? Measure hitters and pitchers separately.

   For hitters, the statistic most correlated with WAR are runs scored, weighted on base average, on-base plus slugging percentage, slugging percentage, and on-base percentage. The statistics least correlated with hitter WAR are hit-by-pitch, caught stealing, fielding, and strikeouts.

   For Pitchers, the statistics most correlated with WAR are strikeouts, wins, innings pitched, games started, and strikeouts per nine innings. The statistics least correlated with pitcher WAR are fielding independent pitching, earned run average, walks and hits per inning, and walks per nine innings.

2. Is there a significant correlation between the number of innings pitched and a pitcher's earned run average? What about a correlation between losses and earned run average?

   There is no significant correlation between the projected innings pitched and the projected earned run average of a pitcher. However, there is a moderate positive correlation between a pitcher's earned run average and the number of losses they are projected to have in the coming season.

3. What teams are projected to perform the best? Do these teams have the best players in terms of WAR?

   The teams that are projected to have the most wins in the coming season are the Los Angeles Dodgers, New York Yankees, San Diego Padres, New York Mets, and the Atlanta Braves.

These teams do have the best players when it comes to WAR, and the WAR statistics has a very high positive correlation with the amount of wins a team is projected to have.

Fantasy Baseball:

Given a specific Fantasy League's scoring system:

1. Are batters or pitchers more valuable in fantasy baseball?

   Hitters score overwhelming more points than pitchers, and therefor can be viewed as more valuable to a certain extent.

2. What statistics are most correlated with scoring a lot of fantasy points? Least correlated? Measure hitters and pitchers separately.

   For hitters, the stats most correlated with scoring a lot of fantasy points are runs, runs batted in, plate appearances, home runs, and on-base plus slugging percentage. The statistics least correlated with hitters scoring a lot of fantasy points are fielding, hit-by-pitch, and caught stealing.

   For pitchers, the stats most correlated with scoring a lot of fantasy points are strikeouts, WAR, innings pitched, and wins. The statistics least correlated with pitchers scoring a lot of fantasy points are fielding independent pitching, earned run average, and walks and hits per inning.

3. Should you diversify your fantasy team (i.e., don't have multiple players from the same team)?

   Diversifying your roster is not important, there can be many high performing players from the same team.

Motivation and Background

Like many sports fans, my friends and I are really into fantasy sports. Of all fantasy sports, baseball is my favorite for a few reasons. Without going into too much detail, I believe fantasy baseball is the most realistic when it comes to playing the role of a team owner. Unlike most fantasy sports, in baseball, you keep your players from one year to the next. This is called 'Dynasty' format. This makes young, up-and-coming players just as important as current superstars because you have to plan for your future as well. For the research questions surrounding fantasy baseball, they will enable me to understand what types of players will give my team a higher chance of performing well. Like most fantasy leagues, there is money involved, so I think this insight could also be beneficial for my bank account.

As for baseball in general, it has become a sport that I have really fallen in love with since high school. I never played it growing up, but in the past six years, I have become a devoted fan and even found myself working for the Mariners. Through this process, I have become rather familiar with baseball statistics. I think there are a lot of takeaways that come from diving into statistics in sports. First, I want to look at what traits 'the best players' have using the WAR stat. Everyone has their own opinion on what defines a good player, but WAR is the typical indicator for this measure. WAR is defined as Wins Above Replacement, which essentially aims to measure a player's total contribution to his team. It represents the number of additional wins his team achieved above the number of expected wins if and average player took their place (the higher the better). I want to see what statistics are most correlated with WAR to try to better understand how the statistic is calculated, as it is pretty ambiguous. Next, I want to dive into the importance of rest for pitchers because there are several different opinions on it around the league. Finally, I want to see how important one player is to a team. This could help determine if it's better to sign one super-star to your roster versus multiple affordable, above average players.

Dataset

hitters: Link
Pitchers: Link
Team Standings: Link

These datasets contain the projected statistics for professional baseball players and teams for the upcoming 2021-2022 season. Each row in the hitter and pitcher datasets represent a single player.

The columns for the hitter dataset include the following:

Name, Team, Games played, Plate appearances, at-bats, hits, doubles, triples, homeruns, runs, RBIs, walks, strikeouts, hit-by-pitch, stolen bases, caught stealing, batting average, on base %, slugging %, on-base plus slugging %, wOBA, FID, BsR, Wins Above Replacement, and ADP.

Name	Team	G	PA	AB	H	2B	3B	HR	R	RBI	BB	SO	HBP	SB	CS	AVG	OBP	SLG	OPS	wOBA	Fld	BsR	WAR	ADP
Mike Trout	Angels	134	593	473	134	25	4	39	100	109	104	131	11	12	3	.283	.420	.600	1.020	.413	-2.0	0.7	6.8	6.0

The dataset for pitchers has columns representing these pitching statistics:

Name, Team, Wins, Losses, ERA, Games Started, Games played, Innings pitched, hits given, earned runs, home runs given, strikeouts, walks, walks and hits per inning, strikeouts/9 innings, walks/9 innings, FIP, Wins Above Replacement, ADP.

Name	Team	W	L	ERA	GS	G	IP	H	ER	HR	SO	BB	WHIP	K/9	BB/9	FIP	WAR	ADP
Lucas Giolito	White Sox	16	7	3.00	30	30	180.0	129	60	21	248	56	1.03	12.40	2.80	2.98	5.8	18.1

The dataset for teams has columns representing collective team statistics:
Team name, games, wins, losses, win %, Run Differential, Runs scored per game, and Runs against per game. This dataset will be joined with each player from the above datasets.

	2020 Year to Date							2021 Projected Rest of Season							2021 Projected Full Season					
Team	G	W	L	W%	RDif	RS/G	RA/G	G	W	L	W%	RDif	RS/G	RA/G	W	L	W%	RDif	RS/G	RA/G
Dodgers	60	43	17	.717	136	5.82	3.55	162	97	65	.598	161	5.30	4.30	97	65	.598	161	5.30	4.30

Methodology

Preparing the Data:

The first step of this project was scraping the data from the web. This required me to learn new python libraries: Beautiful Soup and Selenium. Each data set was scraped separately from webpages within FanGraphs.com.

I also needed to manually create a data frame that mapped team abbreviation (i.e., SEA) to team name (i.e., Mariners). This was needed so that I could join the team data set to both the hitter and pitcher data sets. In the team data set, it contained team names, whereas the player related datasets contained the team abbreviation.

For fantasy baseball questions:

First, I needed to create an additional column in the player related data frames. This column represented the total points a player is projected to score given my fantasy leagues scoring system. This allowed me to answer the research questions related to fantasy baseball.

- **Are batters or pitchers more valuable in fantasy?**
  To answer this, I will look at the average total points for hitters versus the average total points for pitchers. There is the potential for one type of player (pitcher or hitter) to be significantly more valuable than the other, so it will be important to highlight this difference in averages.
- **What stats are most correlated with scoring a lot of fantasy points?**
  To answer this, I will look at the correlation of several statistics with the newly created 'total points' column. I plan to do this both mathematically and visually. These findings will give insight into which players you should roster based on their projected stats (i.e., you may want a player who has a low batting average but hits a lot of home runs, or maybe a pitcher who gives up runs but throws a lot of strikeouts).
- **Should you have multiple players from the same team on your fantasy team?**
  To answer this, I will look at the top 100 hitters and 30 pitchers and count how many of these players are on each team. I plan to make bar charts showing the

number top players (pitchers and hitters seperate), based on total points, across the different MLB teams.

Real world baseball questions:

- **What statistics are most correlated with WAR? Least correlated?**
  To answer this, I will determine the top 200 hitters and 50 pitchers based on WAR. I will look at the correlation of several statistics with the WAR statistic. I plan to do this both mathematically and visually. The statistics that are identified to have a high correlation with WAR will conclude what stats may be most important in determining what makes a player great. Pitchers and hitters will have to be done separately because the stats that go into calculating their WAR is different.
- **Is there a significant correlation between the number of innings pitched and a pitcher's earned run average? What about a correlation between losses and earned run average?**
  To answer this, I will simply look at the correlation, mathematically and visually, of a pitcher ERA and number of innings pitched as well as the correlation between ERA and Losses. If the correlation is high, that may suggest that rest is very important in having pitchers perform well. If it is low, it would suggest that rest isn't as important as some may argue.
- **What teams will perform the best? Do these teams have the 'best' players?**
  To answer this, I will calculate the combined WAR of a team by summing the WAR of all its players within the top 300 hitters and 100 pitchers. I will then compare those values to the teams projected performance. I plan to do this both mathematically and visually. I also plan to make a bar chart showing the number of hitters in the top 300, based off of WAR, across the different MLB teams. I will do the same for pitchers – top 100, based off of WAR, across the different MLB teams.

Results

Real World Baseball:

1. What statistics are most correlated with WAR, an ambiguous stat which is often times used to represent how good a player is? Least correlated? Measure hitters and pitchers separately.

Hitters

Refer to Figure 1

The stats most correlated with WAR are runs scored, weighted on base average, on-base plus slugging percentage, slugging percentage, and on-base percentage. All of these stats have very significant correlation about 0.80.

Runs scored refers to the amount of time the individual player crosses home plate, scoring a run for their team. While this intuitively makes sense (the more runs you score, the more wins your team is likely to have), but I found it surprising that it had the highest correlation to WAR as opposed to the RBI stat, which represents the number of runs batted in (i.e., how many of your teammates cross home plate as a result of your plate appearance). My original thought would be that RBIs would be the most correlated with WAR because a players RBIs is typically higher than their runs. After thinking about it more and finding that runs actually have a higher correlation, I think this result makes sense. Here is why – the run statistic is more of an individual stat and is rarer to come by than that of an RBI which is highly dependent on your teammate's performance as well. Because WAR is also a more of a personal measurement, it makes sense the two are highly correlated.

The statistics least correlated with hitter WAR are hit-by-pitch, caught stealing, fielding, and strikeouts. All of these statistics have essentially no correlation with the WAR measurement. I was surprised that a statistic like strikeouts didn't have a correlation (0.12 coefficient) as this negatively affects your team's performance.

Pitchers

Refer to Figure 2

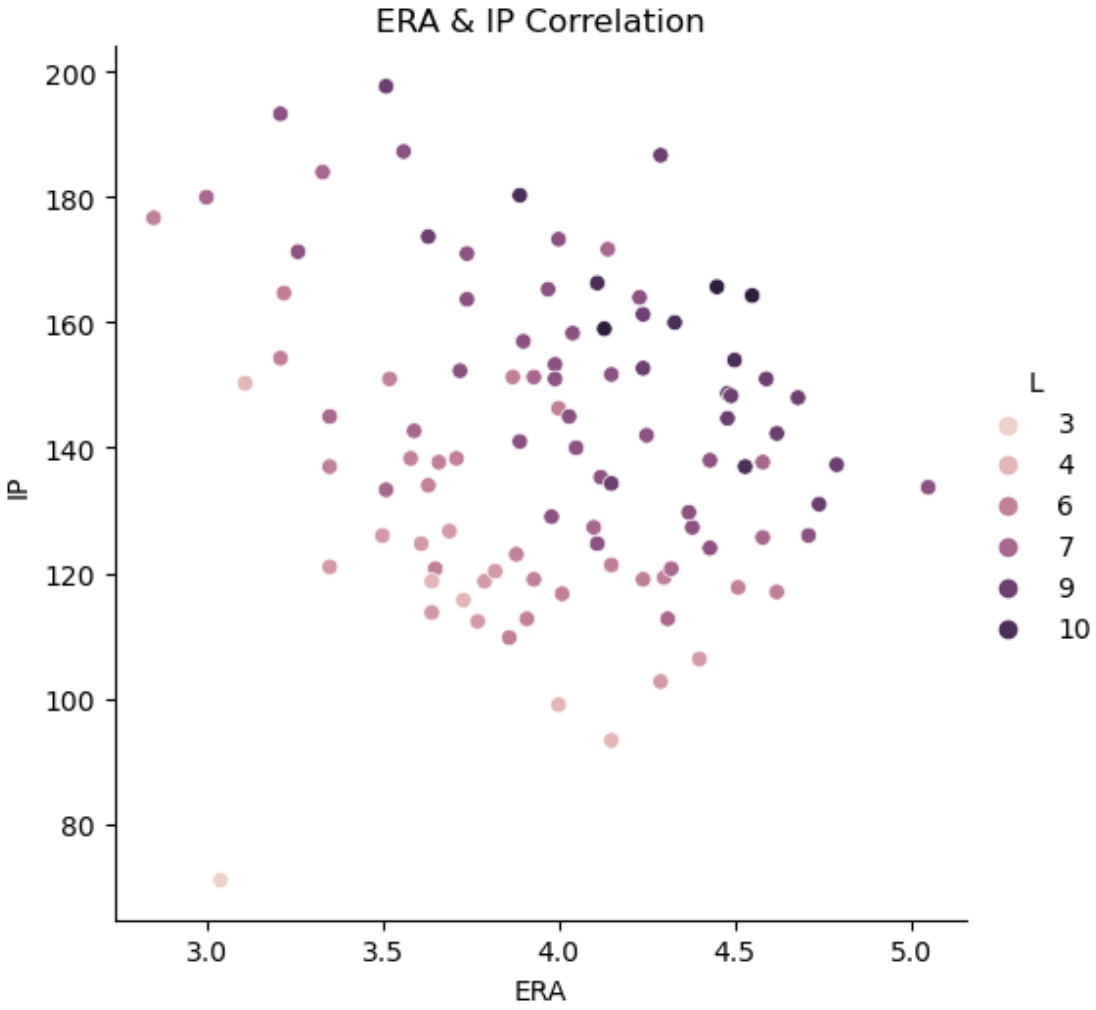
The statistics most correlated with pitcher WAR are strikeouts, wins, innings pitched, games started, and strikeouts per nine innings. Strikeouts has the highest correlation of all (0.82 coefficient). This is important because there are pitchers who are considered 'strikeout' type pitchers and some who are considered 'ground-out' type pitchers. I think it's interesting that the WAR statistic values a strikeout so much when its only one category of a type of out. Is it the strikeouts that make the pitcher great, or is it just the out itself? If I had more time, I would be interested to dive deeper into that question.

The statistics least correlated with pitcher WAR are fielding independent pitching (FIP), earned run average, walks and hits per inning (WHIP), and walks per nine innings. FIP actually has a very high negative correlation to WAR (-0.72), meaning the lower the FIP, the better the WAR. FIP measures what a player's ERA would look like over a given period of time if the pitcher were to have experienced league average results on balls in play and league average timing. This makes a lot of sense – the better a pitcher's ERA, the more wins they are likely to have and thus receive a higher WAR. Because of this, it's no surprise that ERA has a moderate negative correlation with WAR (0.63). Similarly, a low WHIP also makes instinctive sense to be negatively correlated with WAR.

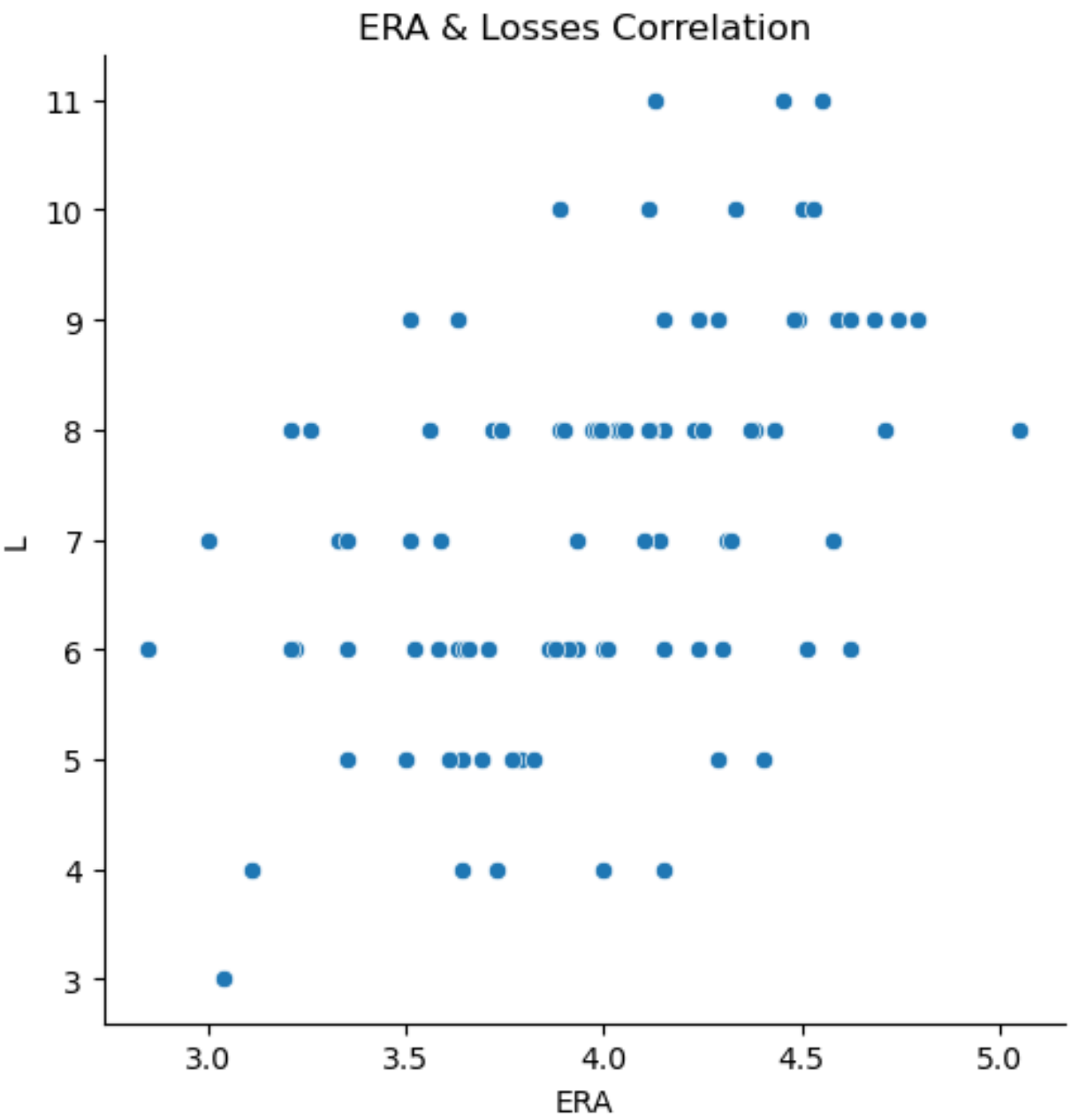
2. Is there a significant correlation between the number of innings pitched and a pitcher's earned run average? What about a correlation between losses and earned run average?

I found there to be no significant correlation between the projected innings pitched and the projected earned run average of a pitcher (-0.19 coefficient). I found this to be surprising because intuitively you would think that the more innings a pitcher plays, the more likely they are to give up runs. Because there is essentially no correlation, this could help support the argument that rest is not as important as some baseball coaches and players may think. Typically, pitchers play every 5 games, but there has been a desire for some of the best

pitchers in the league, such as Trevor Bauer, to pitch every 4 games instead. This analysis could help support players like Trevor Bauer's argument to pitch more often.

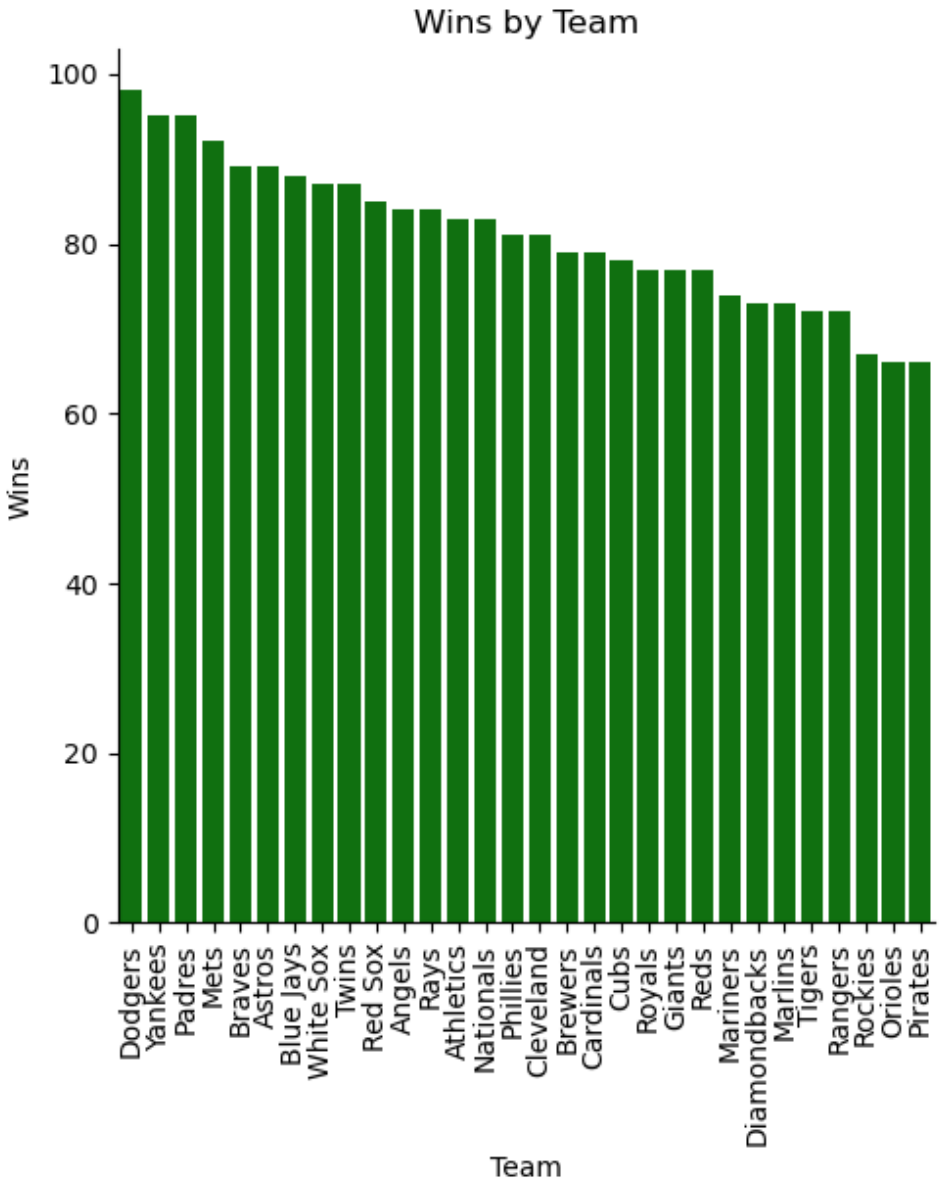


I also found that there is a moderate positive correlation between a pitcher's earned run average and the number of losses they are projected to have (0.47 coefficient). This means that the more runs they give up, the more losses they are likely to have. I was surprised that this correlation wasn't higher. A reason this may not be as high as I expected is because the result of a loss is also highly dependent on the offensive performance of a pitcher's team.

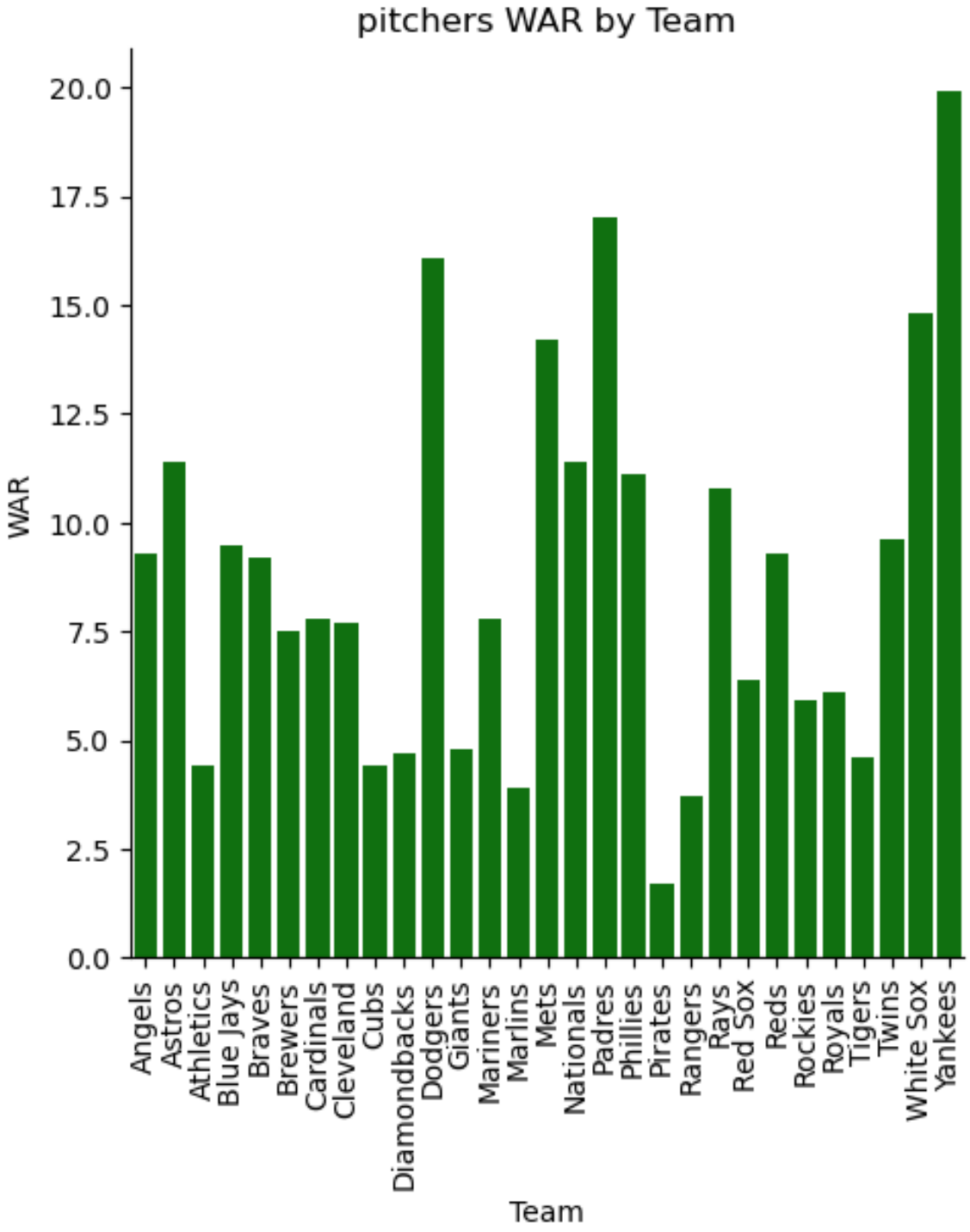

ERA & Losses Correlation
L
ERA
11
10
9
8
7
6
5
4
3
3.0
3.5
4.0
4.5
5.0

3. What teams are projected to perform the best? Do these teams have the best players in terms of WAR?

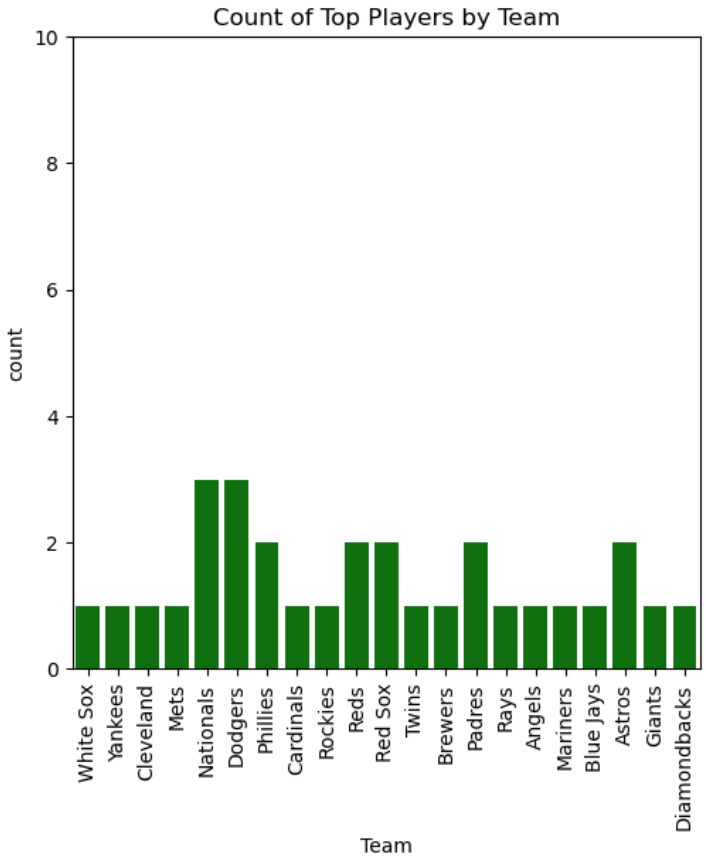
   The teams that are projected to have the four most win totals in the coming season are the Los Angeles Dodgers, New York Yankees, San Diego Padres, New York Mets.



If you look at the total pitcher WAR among the top 100 pitchers, you'll notice that the top four teams in this measurement are the same as the four teams projected to win the most games in the above graphic.

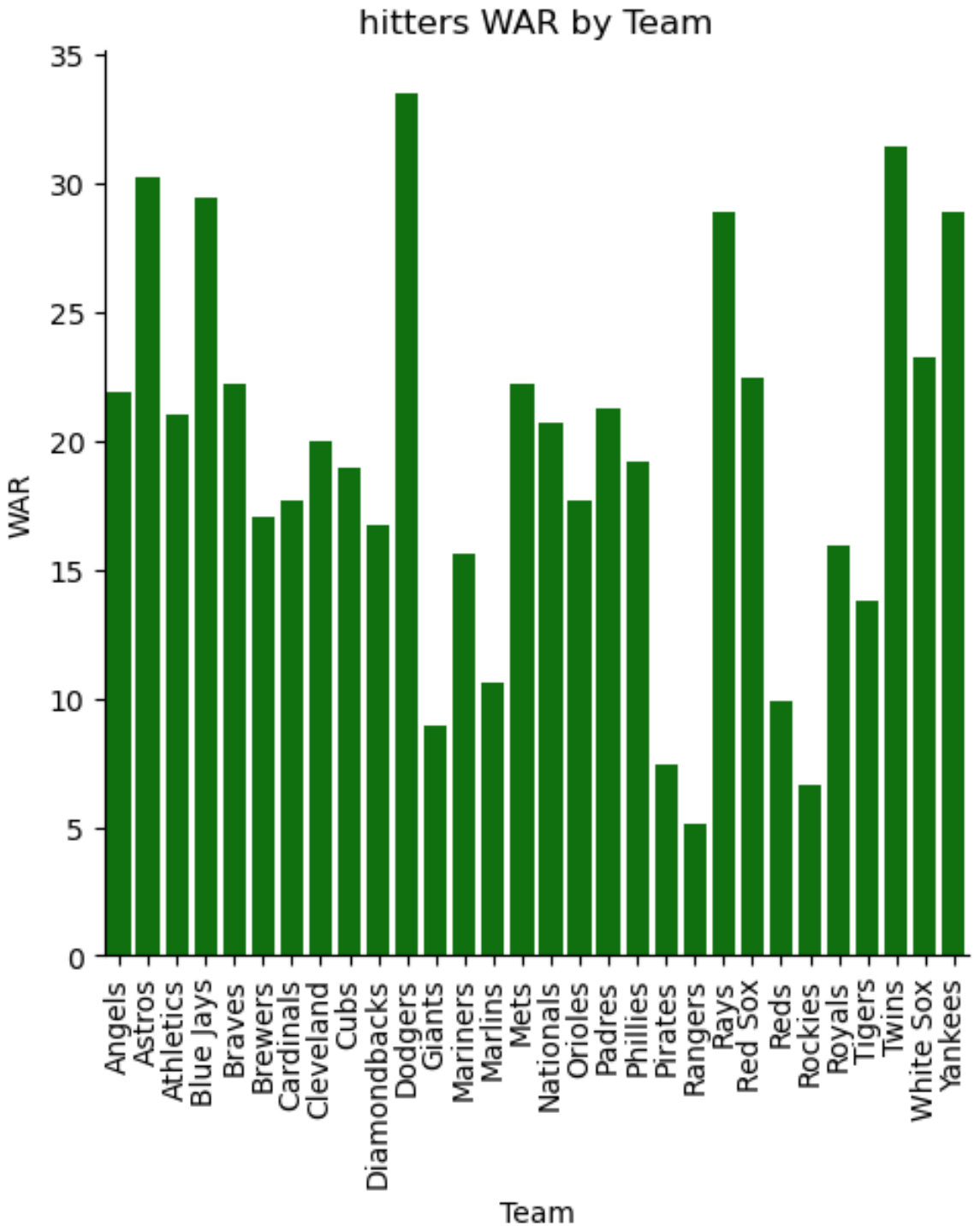


Here is a bar chart showing how many **pitchers** in the top 30, based on WAR, are on each team.

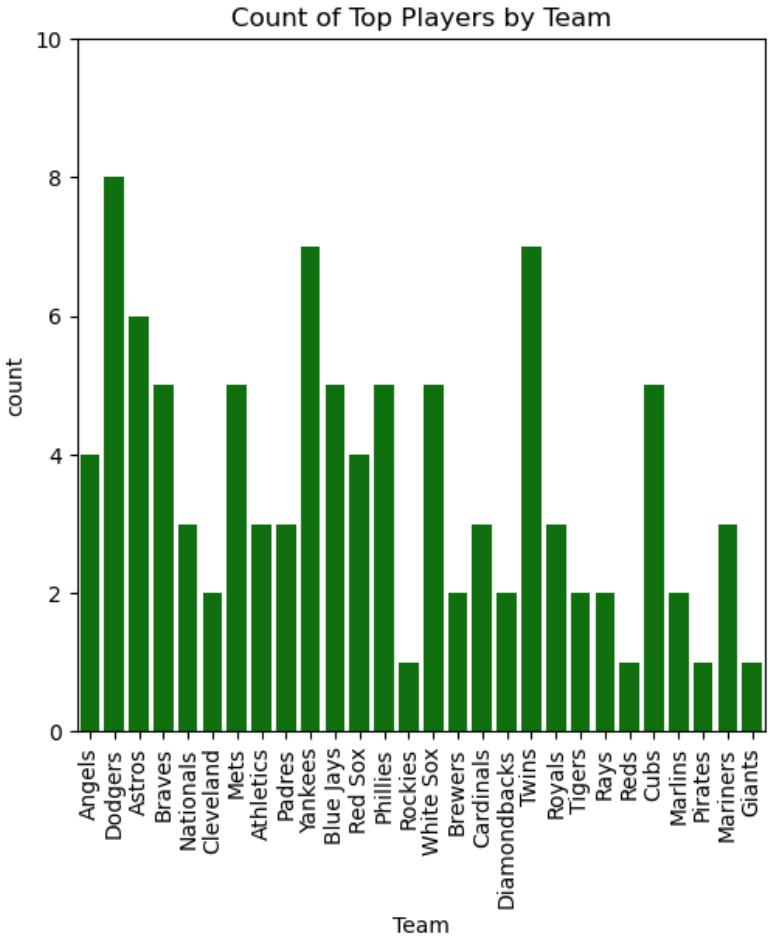


As you can, the Dodgers and Padres both perform well in the measurement too.

While not as prominent as the previous graphs, these same four teams also perform very well in total hitter WAR among the top 300 hitters in MLB.

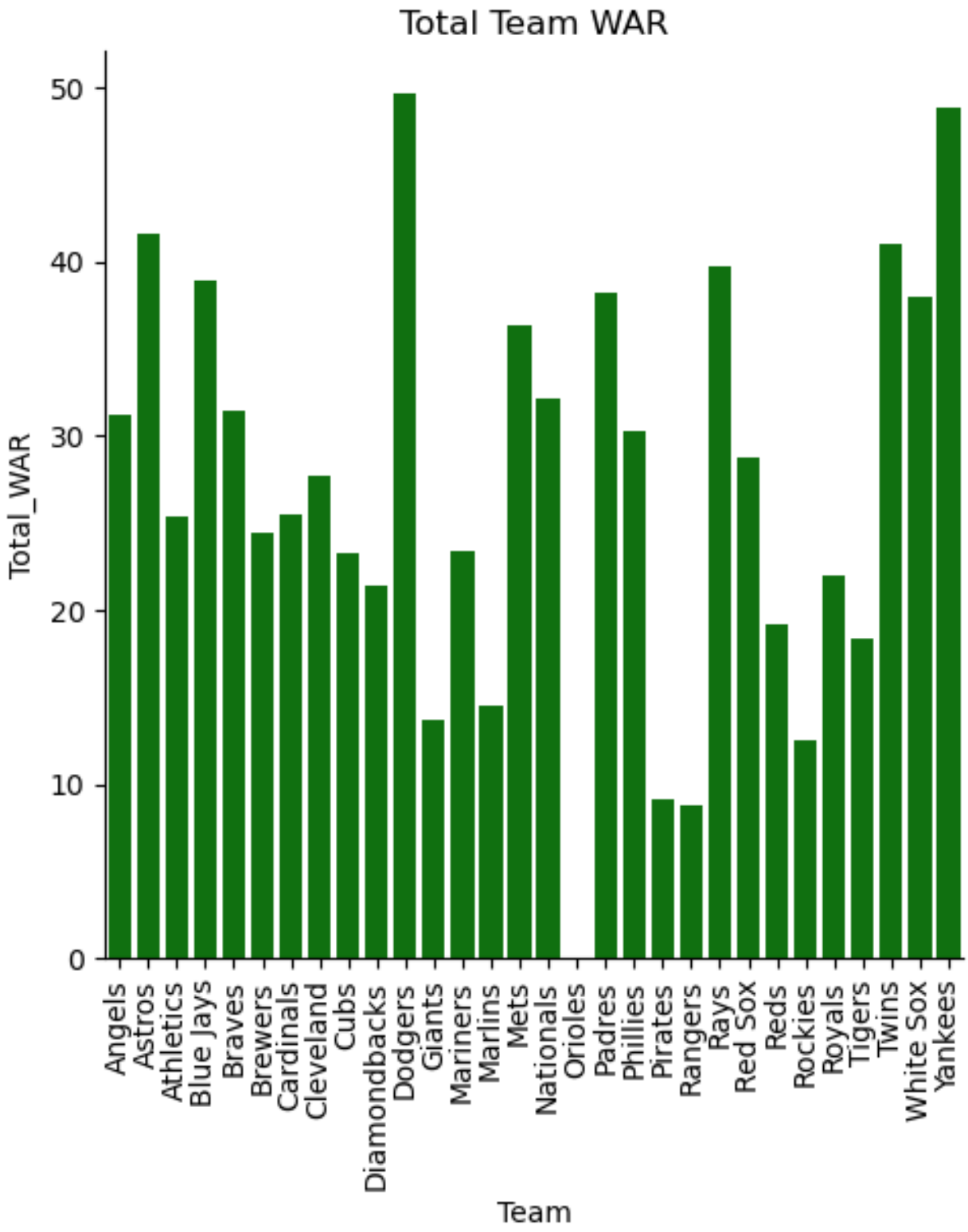


Here is bar chart showing how many **hitters** in the top 100, based on WAR, are on each team.

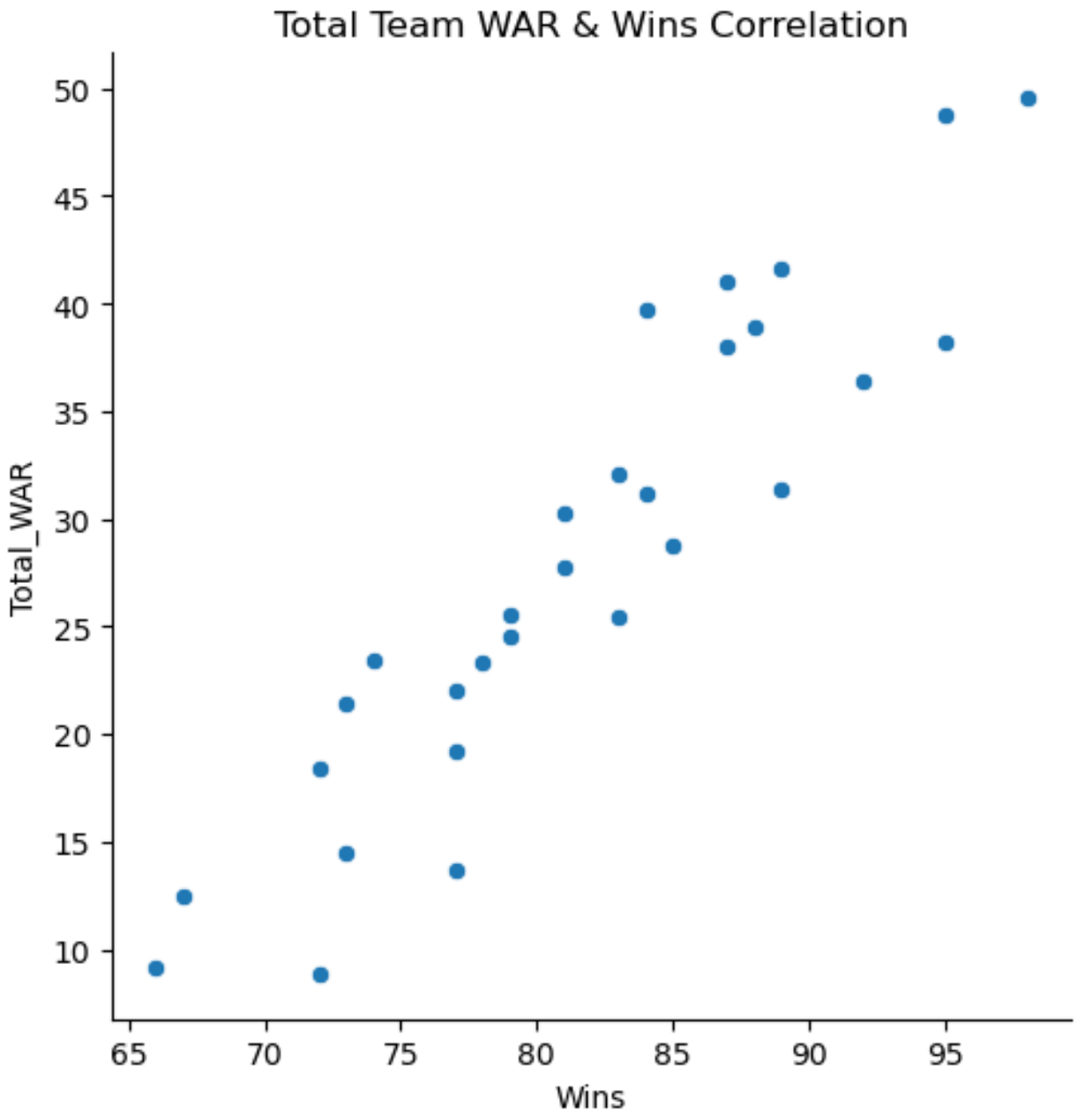


Again, the theme of the Dodgers, Yankees, and Mets exceeding the rest of the league continues.

Combining total pitcher WAR and total hitter WAR, the distribution of total team WAR is as shown below.



So yes, the teams that are projected to do the best in the upcoming season are teams with some of the highest team WAR. I calculated the correlation between total team WAR and projected wins, and the result was a coefficient of 0.92, indicating a very highly correlated relationship.



Fantasy Baseball:

1. Are batters or pitchers more valuable in fantasy baseball?

   I found that Hitters score overwhelming more points than pitchers. Top hitters on average score 1,268 points throughout the season, and top pitchers only score 708 points. With that being said, because you are required to roster pitchers on your team, it is still worth drafting the best pitchers early on. While pitchers may not score nearly as many points, by requiring the position to be rostered in fantasy leagues, there importance is still high. Furthermore, it is important to note that that pitchers typically only play one or two games a week, whereas hitters play six to seven games, ultimately leading to such a large difference in points scored. After doing this research, I still think it would be a good idea to increase the amount of points a pitcher can score you in order to increase the popularity of pitchers within fantasy baseball. If I had more time, I would have loved to analyze the distribution of pitchers' total points and hitters' total points to see if star pitchers are harder to come by then star hitters.

2. What statistics are most correlated with scoring a lot of fantasy points? Least correlated? Measure hitters and pitchers separately.

   Hitters

   Refer to Figure 3

   For hitters, the stats most correlated with scoring a lot of fantasy points are runs, RBIs, plate appearances, home runs, and on-base plus slugging percentage. This could help a lot when looking to construct a fantasy baseball roster. For example, because homeruns have such a high correlation (0.81 coefficient), you could conclude that 'power hitters' are very valuable in fantasy baseball and therefore should be drafted ahead of hitters who don't hit many homeruns. With that being said, because runs are most correlated with scoring a lot of fantasy points, some non-power hitters who still score a lot of runs may be even more valuable.

   The statistics least correlated with hitters scoring a lot of fantasy points are fielding, hit-by-pitch, and caught stealing. I was a bit surprised that hit-by-pitch isn't correlated with scoring

a lot of fantasy points (0.03 coefficient) because it is worth 4 points, which is more than a run is worth which was the highest correlated stat. One reason for this could be that hit-by pitches happen too rarely to play any significant role in the total amount of points scored across a season.

Pitchers

Refer to Figure 4

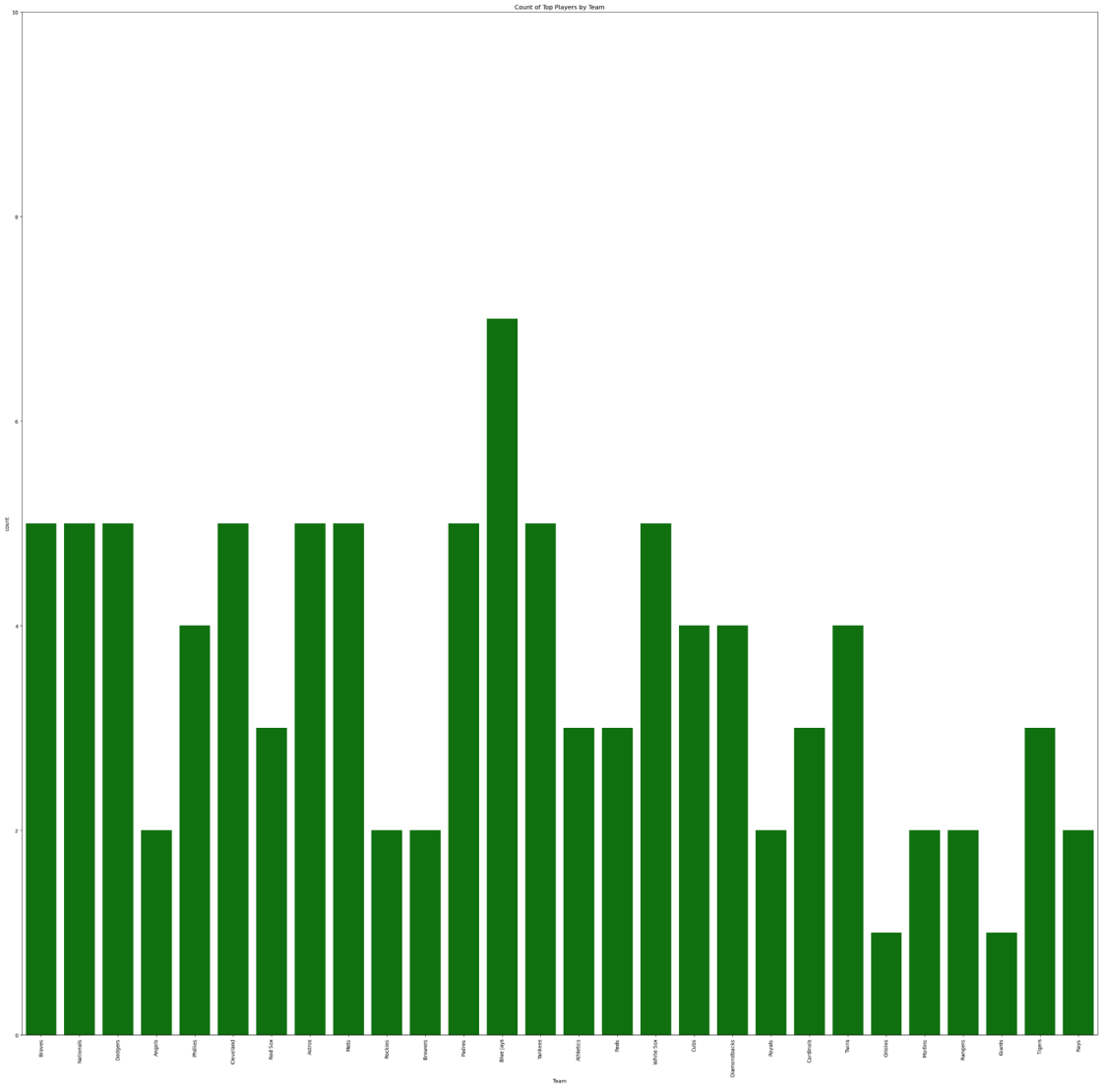
For pitchers, the stats most correlated with scoring a lot of fantasy points are strikeouts, WAR, innings pitched, and wins. Strikeouts appear to be the most important statistic for pitchers in fantasy baseball (0.98). While I think this makes sense intuitively, I am also concerned with how high of a correlation there is. This basically indicates that the only pitchers worth having are those that strikeout a lot of batters. In real baseball, while strikeouts are obviously great, every other out is important as well. This shows to me that there is a flaw when it comes to the scoring system of pitchers in my fantasy league. I think there is an excess of importance put on strikeouts as opposed to a stat like innings pitched, which essentially represents an out regardless of how it was made (strikeout, groundout, flyout, etc.).

The statistics least correlated with pitchers scoring a lot of fantasy points are fielding independent pitching, ERA, and WHIP. These stats actually all have a moderate negative correlation with scoring fantasy points, which makes sense. The higher the ERA and WHIP, the more batters the pitcher is allowing to reach base and score, which negatively effects their fantasy performance. This shows a good configuration of points taken away from pitchers in our league's scoring system.

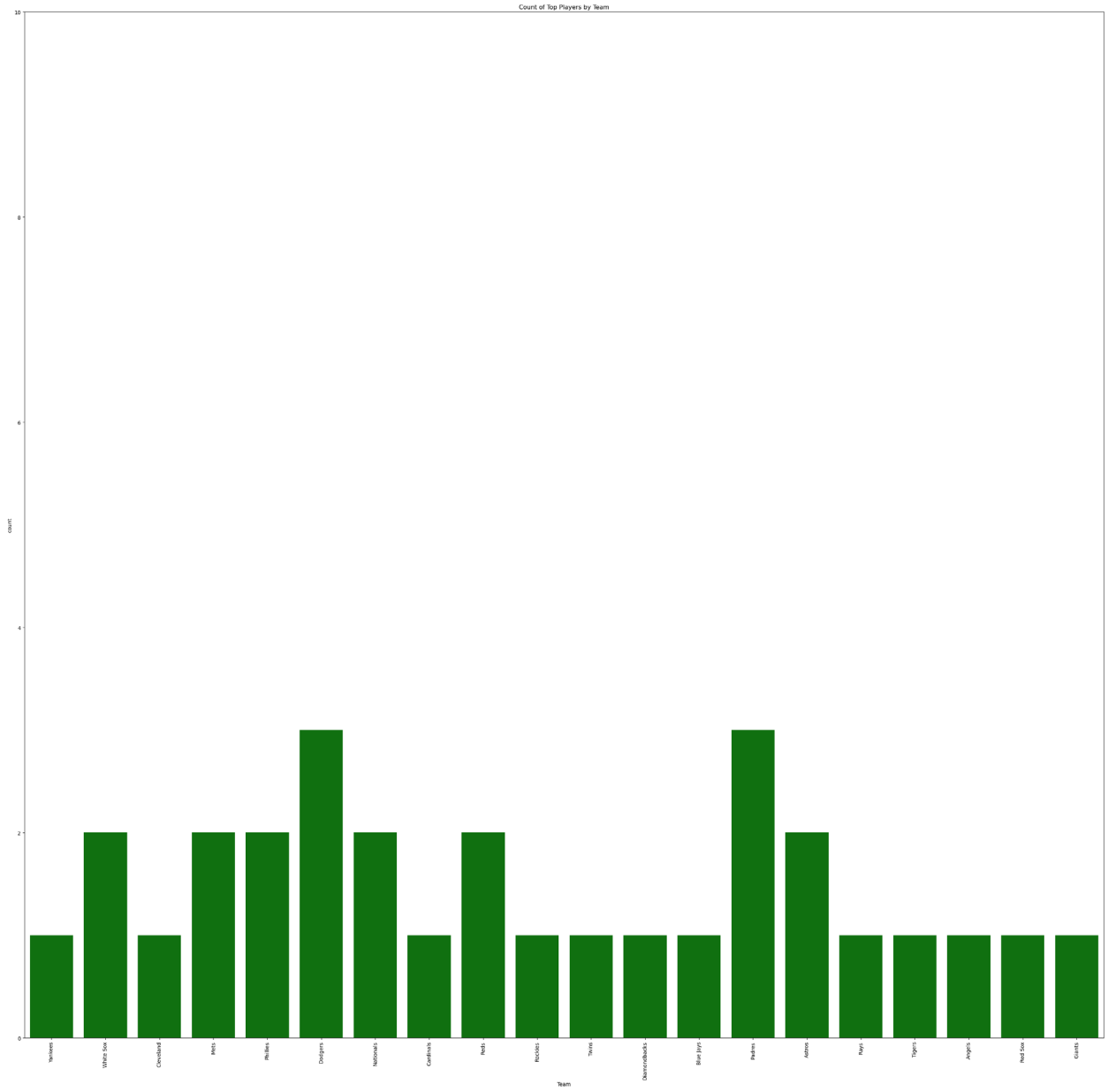
3. Should you diversify your fantasy team (i.e., don't have multiple players from the same team)?

If you're truly aiming to perform well in your fantasy league, I came to the conclusion that it doesn't matter what real-world team your players are from. Baseball is a sport where, while it is a team sport, players performances are for the most part independents of one another.

Below is a bar chart showing the distribution of the top 100 hitters, based on total fantasy points scored, across the different MLB teams.



Similarly, below is a bar chart showing the distribution of the top 30 pitchers, based on total fantasy points scored, across the different MLB teams.



These charts make it clear that one single team can have several great fantasy players. For examples, the first chart shows that the Toronto Blue Jays have seven of the top 100 fantasy hitters. Furthermore, the Dodgers and Padres each of 3 of the top 30 pitchers in fantasy baseball. There is no reason to avoid rostering good players because you fear you have too many players from the same team. It's simple, roster the best players you can, regardless of their real-world team.

Challenge Goals

1) **New library**
   I had to learn the BeautifulSoup and Selenium libraries to scrape these datasets from the web. I wanted to learn how to scrape the data myself, so I used Selenium to navigate the webpage to be able to collect all the rows of the data set. Each webpage only showed 30 players per page, so I needed to use selenium to scroll and click various buttons on the webpages in order to collect a sufficient amount of data. BeautifulSoup was used to scrape the data which I then converted to pandas data frames and CSV files.

2) **Multiple Datasets**
   I analyzed multiple datasets (individual pitching and hitting datasets) as well as combined a team projection dataset with the aforementioned player specific datasets to help answer my research questions. I also had to create a dataset of my own that mapped team abbreviation (i.e., SEA) to team name (i.e., Mariners). I had to join this dataset to the original team dataset before joining it with both the hitter and pitcher datasets.

Work Plan Evaluation

Work Plan from Part 1

Task One: Scrape Data (10 hours)
I have already began learning how to scrape data using beautiful soup. Additionally, because I will need to interact with the webpages I am scraping from, I have begun experimenting with the Selenium library to assist in scraping the data. So far, learning the libraries, implementing them, and talking with a TA to help debug has taken about 6 hours. I ran into a lot of issues because the webpage I am scraping from is sort of complicated. Because I believe I have a better understanding now, I think it should only take me another 4 hours of worktime to finish writing code to properly scrape all the data and then break the code up into organized functions.

Task Two: Fantasy baseball analysis (5 hours)
This will include calculating a new column, joining datasets, and doing several different mathematical and visual analysis. I see myself needing about 4-5 hours to explore all the different ways I could visualize this part of the research. I also see myself finding new ways of answering these questions once I start to actually work with the data.

Task Three: Real world analysis (4 hours)
This will include joining datasets and doing several different mathematical and visual analysis. I think this will be faster than task 2 because I don't believe I will have to create any additional columns, outside of joining datasets, to support my calculations. I also see myself finding new ways of answering these questions once I start to actually work with the data.

Task Four: Report write up (6 hours)

I'm not sure exactly what the write up entails, but I believe in order to write an in-depth paper that is organized well and contains neatly formatted plots and visuals will take a large chunk of time. Also, reflecting on my results and making conclusion will require a fair amount of thought and effort.

Evaluation

My predictions for Task One and Task Four were spot on, but Task Two and Task Three took my much longer than I anticipated. I spent a lot of time trying to reduce the redundancy of my code, something I didn't take into consideration when originally planning out my project. However, I was correct in my prediction that Task Three would take less time than Task Two. This was because I was able to utilize a lot of the functions that I had created in Task Two for my analysis in Task Three. I would say Task Two and Task Three required about 3 hours more each than expected.

Testing

load_data():

This method simply loaded in in data sets using the read csv function and returned them. Because of its simplicity, there was no precise way of testing this method. This method can be trusted to work properly because I was able to conduct analysis on the csv file that were imported without error.

create_team_key():

This method created a data frame that mapped team abbreviation (i.e., SEA) to team name (i.e., Mariners) and merged it with the scraped data set of team statistics. This method can be trusted to work properly because the analysis done specifically on teams ran without any error.

merge_player_and_team():

Similarly, this method was just responsible for merging the player data sets with the team data set. This method can be trusted to work properly because the analysis that included team information ran without any error.

stat_corr_hitters_visual():

This method was responsible for creating a correlation matrix for hitters based on the passed in statistic. To ensure this ran correctly, you can see that statistics that would intuitively have a high positive correlation with WAR (Runs, OPS, etc) do indeed have high positive correlations looking at the plot (**Figure 1**). Furthermore, in **Figure 3**, the stats that mathematically have a very high positive correlation with total points scored by hitters (i.e., runs) also show a very high correlation visually.

stat_corr_pitchers_visual():

This method was responsible for creating a correlation matrix for pitchers based on the passed in statistic. If you look at **Figure 2**, you will see that there are 50 points one

each plot, as the matrix was based upon the top 50 pitchers. Furthermore, looking at **Figure 4**, you will see that the stats that mathematically have a very high positive correlation with total points scored by pitchers (i.e., strikeouts) also show a very high correlation visually.

stat_corr_math():

This method was given a statistic and returned two data frames of the most and least correlated stats respectively. The results of this method can be trusted because the match up with the plots previously mentioned, **Figure 1, 2, 3,** and **4**.

era_to_ip_and_l():

To test this method, I created a smaller data set of only 3 players and purposefully made the correlation equal to one for both stats. I then used assert equals to confirm the method was executed correctly.

team_war_position():

This method created a plot showing the total WAR of each team based on a given position (pitcher or hitter). I can trust that these plots were correct because it makes sense that the teams with the higher total WAR in the plots (Dodgers, Yankees, Padres) were also the teams that were projected to perform the best based on the team data from FanGraphs.com. It makes sense that the teams who are expected to perform the best are also the teams with the highest team WAR.

team_war_total():

This method plots the teams' total war (hitter WAR and pitcher WAR combined). I can trust that this method worked correctly because if you combine the two bar charts created from **team_war_position()** (one for pitchers and one for hitters), the counts of the newly plotted bar chart are equal to their sum.

team_wins_plot():

This method plots the total projected wins for each team. It does not require any testing other than looking at the team dataset and making sure the appropriate number of wins are plotted on the bar chart, which they are. This was a pretty straightforward plot.

team_war_to_wins():

This plotted the correlation of team WAR to projected wins and also returned the correlation coefficient. I made two smaller dataset, one containing team WAR info and the other containing team win info, both consisting of only 3 teams. I purposefully made the correlation between WAR and wins equal to 1.0 and used assert equals to confirm the method was executed correctly.

average_points():

This method returns the average total points of the passed in data sets. I created two smaller data sets, one for pitchers and one for hitters, both consisting of 6 players in which I purposefully set their total points to average out to a certain value and used assert equals to confirm the method was executed correctly.

create_total_points_hitter() & create_total_points_pitcher():

These two methods are very similar in that they take in a dataset of players (hitters or pitchers) and calculate each player's total fantasy points based upon my league's scoring system and returns the new data frame with the new column. To ensure that this was done accurately, I manually calculated the total fantasy points for the first 3 hitters and first 3 pitchers, which resulted in the same values calculated in the new data frames.

Collaboration

This project was done completely by me. I did utilize some online resources, specifically a YouTube channel called 'Tech with Tim' where I learned how to use the BeautifulSoup and Selenium libraries. I also reference Seaborn and myplotlib documentation when creating plots.

Appendix

Figure 1: Hitter WAR Correlation Matrix

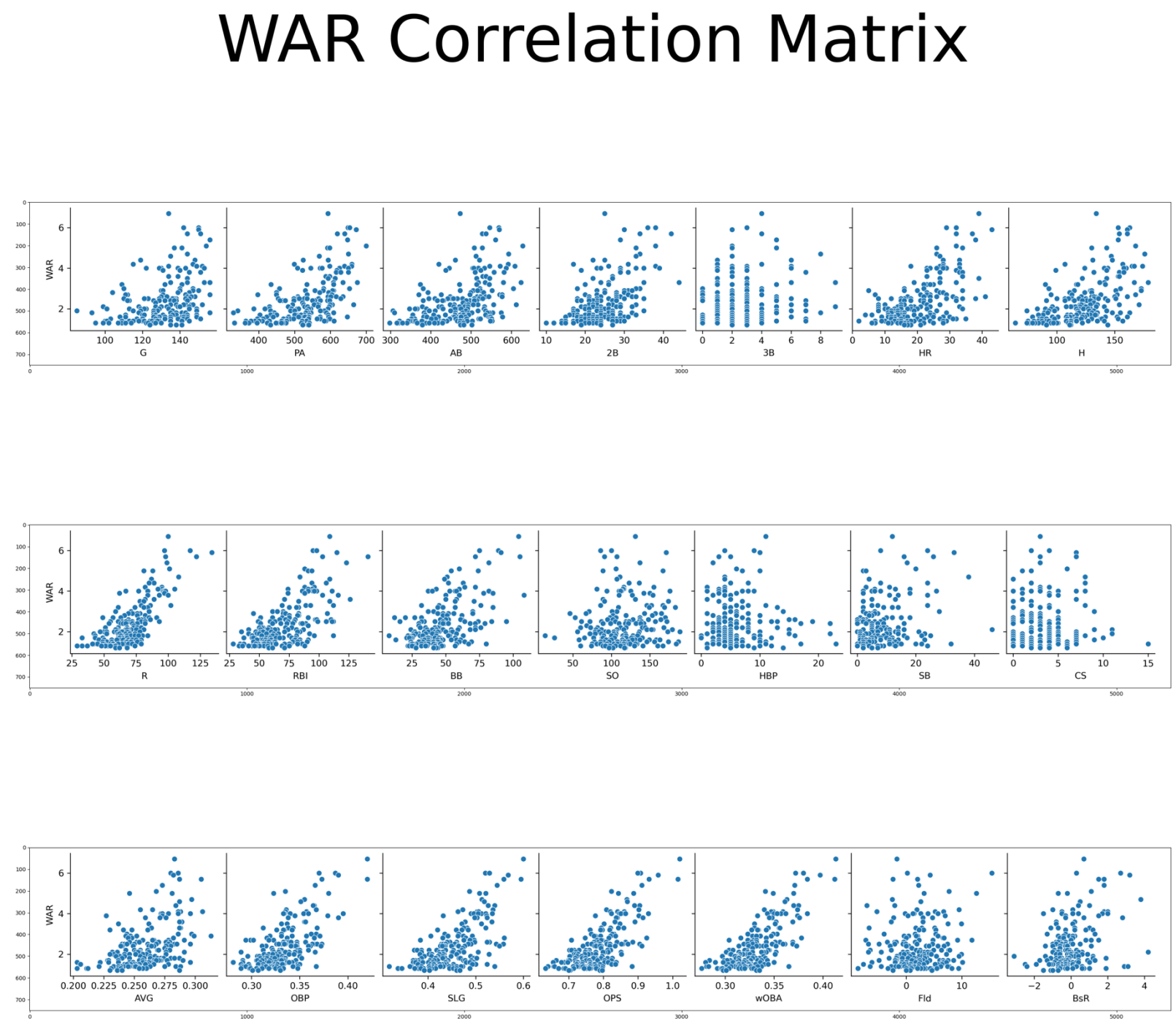


Figure 2: Pitcher WAR Correlation Matrix

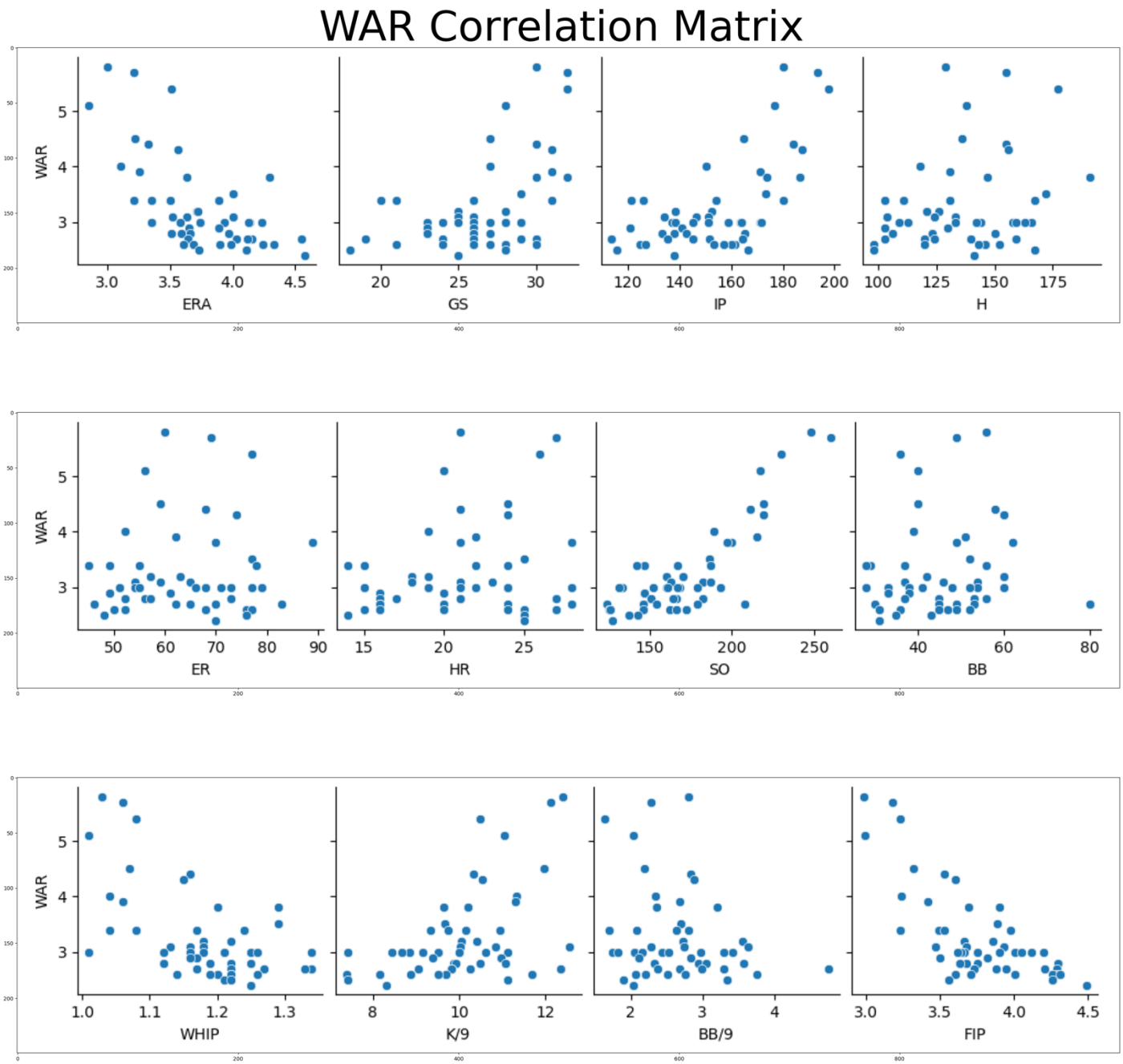


Figure 3: Hitter Total Points Correlation Matrix

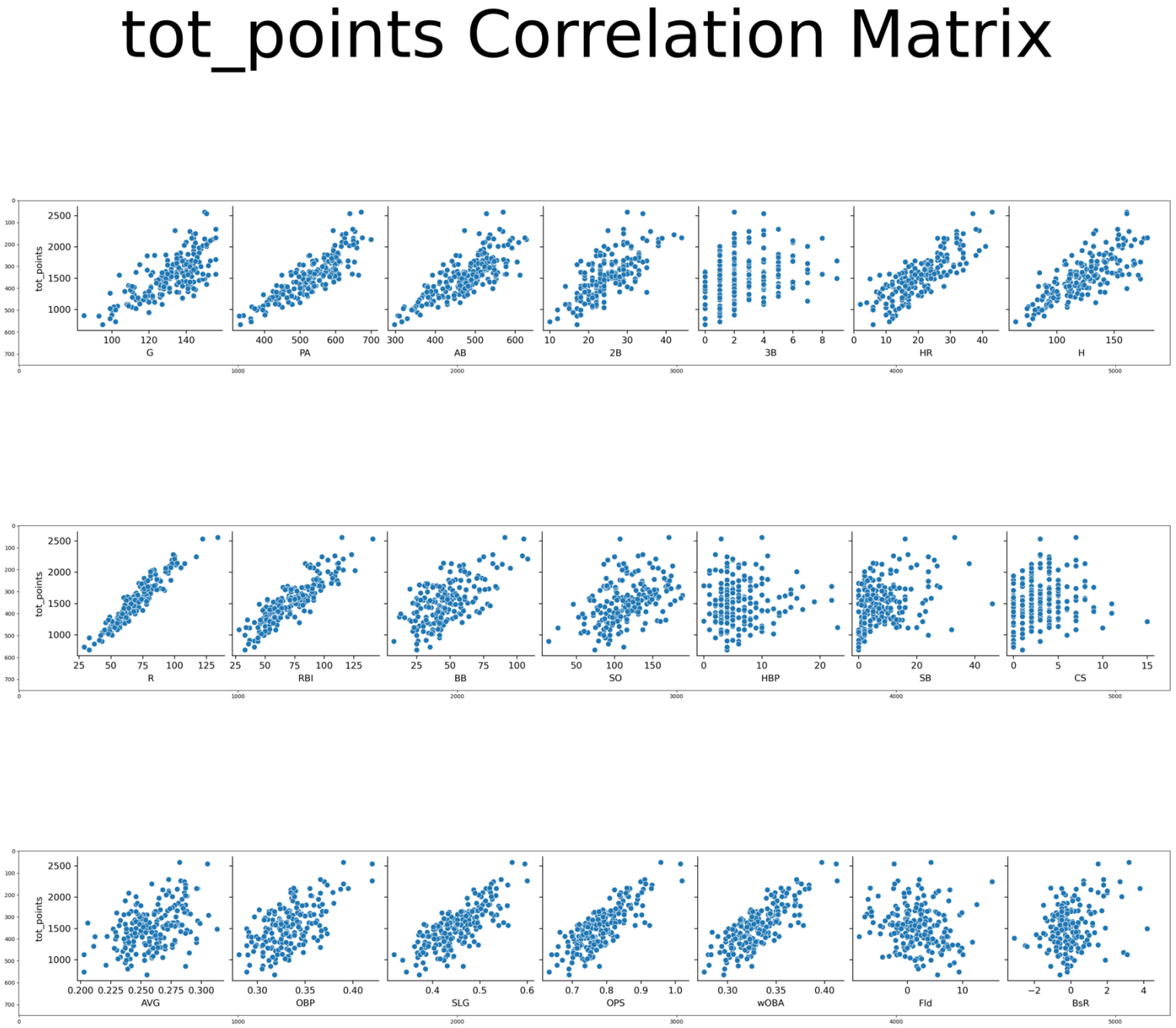


Figure 4: Pitcher Total Points Correlation Matrix