Harper Messer
Math 2820L
4/22/2023

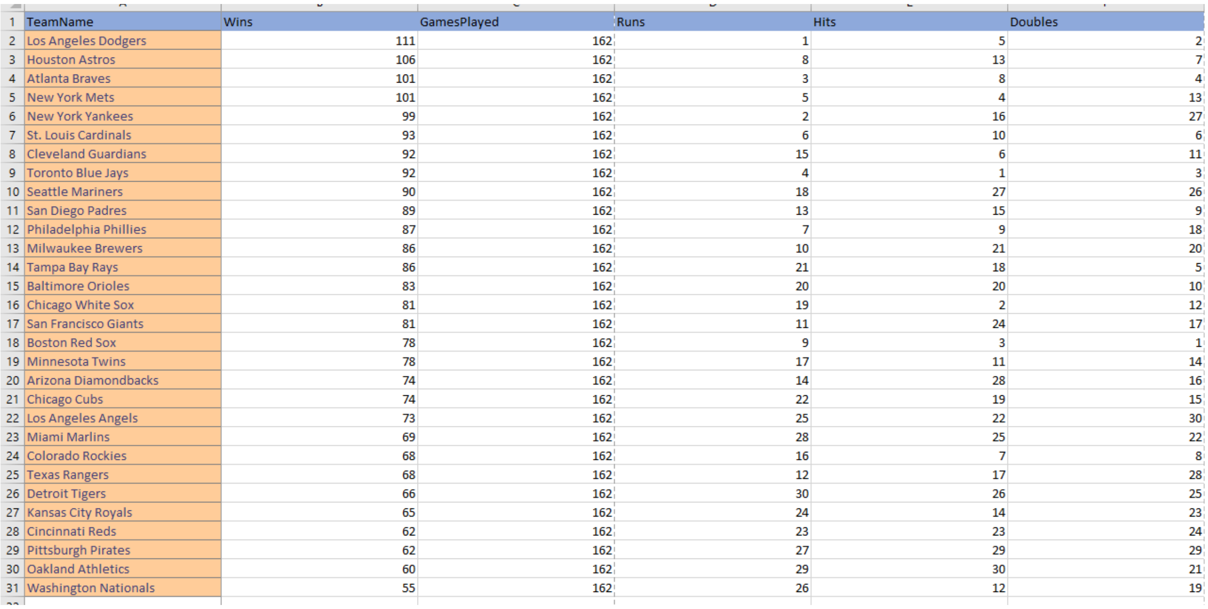
Math 2820L Project Write Up

Introduction:

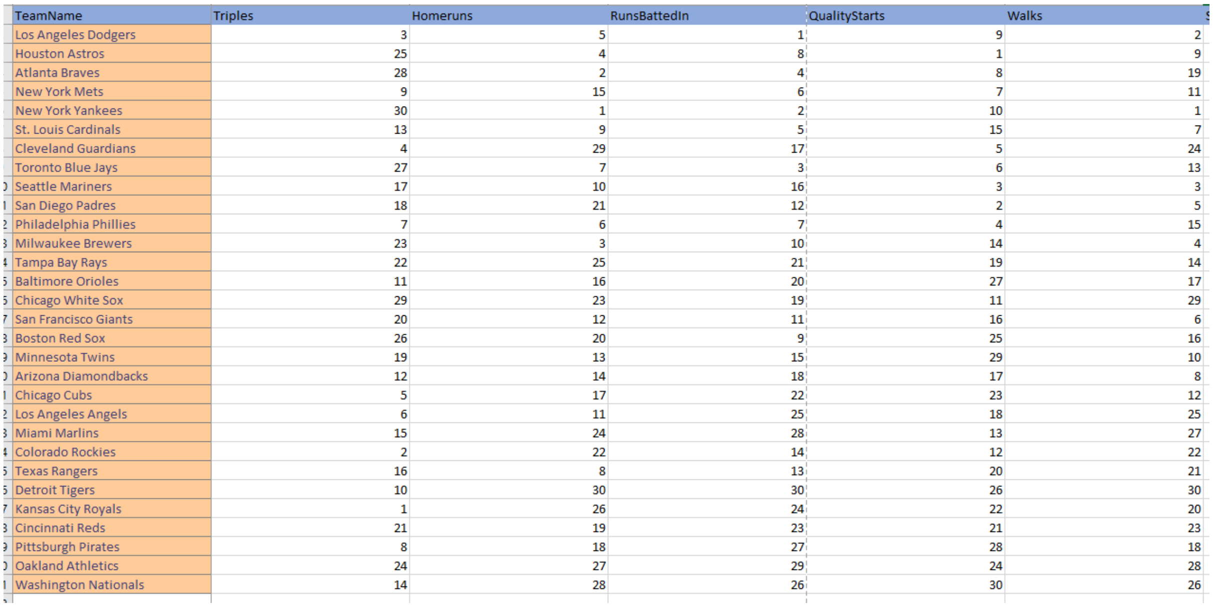
For this project, I will solve the problem of: *"What statistics correlate strongest with winning"*. In order to solve this problem, I will be analyzing MLB team's pitching, batting, and fielding statistics from the 2022 season bycreating graphs and using regression and confidence intervals to determine what statistics correlate the strongest with winning. At the conclusion of this analysis, I will be able to inform MLB teams and others interested in what stats they should focus on improving the most and least in order to hypothetically win more games. This project is inspired by the movie Moneyball and my love for baseball, analytics and sports betting. The data I extracted is from ESPN and Covers and then I scraped and cleaned it all up into an Excel and then of having the actual team statistics, I used Excel to put the team's respective rank in that category instead to make my analysis better. For example, the Los Angeles Dodgers had the best pitching ERA in 2022 so that's why the cell corresponding to it is "1". Some popular stats I left out were On Base Percentage + Slugging (OPS), Walks and Hits Per Inning (WHIP), and Complete Games. I left OPS because we already have On Base Percentage in the dataset as well as Slugging just in their own columns. This applies the same for WHIP because we already have columns for hits and walks. I separated them because I thought it would be better to analyze these statistics separately because maybe teams that allow more walks win less over allowing more hits. I also decided to not include Complete Games because Complete Games are so rare nowadays as pitchers are throwing harder and more injury prone and such plus usually when a pitcher throws a complete game it is a shutout and we already have shutouts being tracked.

**All stats courtesy of:*
Pitching Stats: https://www.espn.com/mlb/stats/team/_/view/pitching/season/2022/seasontype/2
Hold Stats: https://www.covers.com/sport/baseball/mlb/statistics/team-bullpenera/2022
Batting Stats: https://www.espn.com/mlb/stats/team/_/season/2022/seasontype/2
Fielding Stats: https://www.espn.com/mlb/stats/team/_/view/fielding/season/2022/seasontype/2

Harper Messer
Math 2820L
4/22/2023

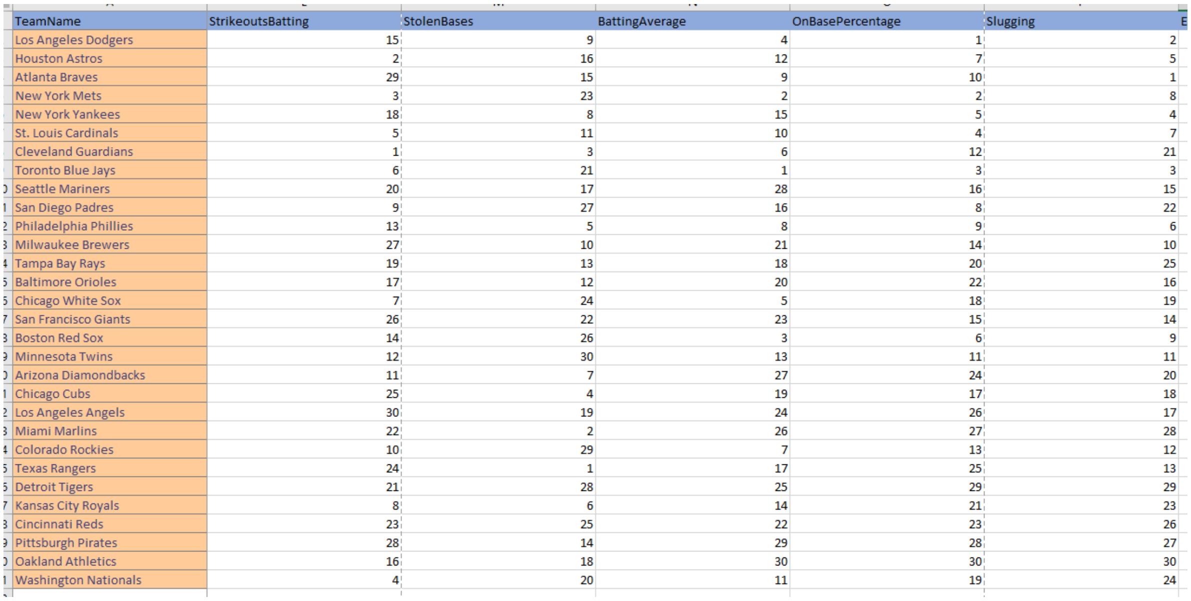


	TeamName	Wins	GamesPlayed	Runs	Hits	Doubles
2	Los Angeles Dodgers	111	162	1	5	2
3	Houston Astros	106	162	8	13	7
4	Atlanta Braves	101	162	3	8	4
5	New York Mets	101	162	5	4	13
6	New York Yankees	99	162	2	16	27
7	St. Louis Cardinals	93	162	6	10	6
8	Cleveland Guardians	92	162	15	6	11
9	Toronto Blue Jays	92	162	4	1	3
10	Seattle Mariners	90	162	18	27	26
11	San Diego Padres	89	162	13	15	9
12	Philadelphia Phillies	87	162	7	9	18
13	Milwaukee Brewers	86	162	10	21	20
14	Tampa Bay Rays	86	162	21	18	5
15	Baltimore Orioles	83	162	20	20	10
16	Chicago White Sox	81	162	19	2	12
17	San Francisco Giants	81	162	11	24	17
18	Boston Red Sox	78	162	9	3	1
19	Minnesota Twins	78	162	17	11	14
20	Arizona Diamondbacks	74	162	14	28	16
21	Chicago Cubs	74	162	22	19	15
22	Los Angeles Angels	73	162	25	22	30
23	Miami Marlins	69	162	28	25	22
24	Colorado Rockies	68	162	16	7	8
25	Texas Rangers	68	162	12	17	28
26	Detroit Tigers	66	162	30	26	25
27	Kansas City Royals	65	162	24	14	23
28	Cincinnati Reds	62	162	23	23	24
29	Pittsburgh Pirates	62	162	27	29	29
30	Oakland Athletics	60	162	29	30	21
31	Washington Nationals	55	162	26	12	19

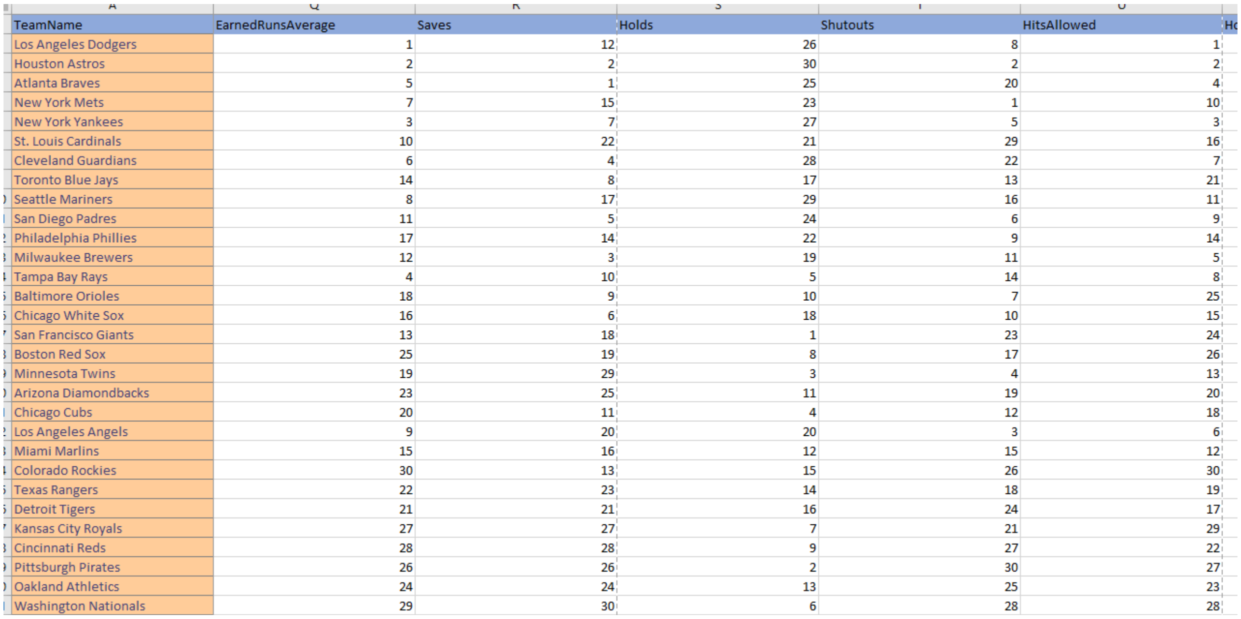


TeamName	Triples	Homeruns	RunsBattedIn	QualityStarts	Walks
Los Angeles Dodgers	3	5	1	9	2
Houston Astros	25	4	8	1	9
Atlanta Braves	28	2	4	8	19
New York Mets	9	15	6	7	11
New York Yankees	30	1	2	10	1
St. Louis Cardinals	13	9	5	15	7
Cleveland Guardians	4	29	17	5	24
Toronto Blue Jays	27	7	3	6	13
Seattle Mariners	17	10	16	3	3
San Diego Padres	18	21	12	2	5
Philadelphia Phillies	7	6	7	4	15
Milwaukee Brewers	23	3	10	14	4
Tampa Bay Rays	22	25	21	19	14
Baltimore Orioles	11	16	20	27	17
Chicago White Sox	29	23	19	11	29
San Francisco Giants	20	12	11	16	6
Boston Red Sox	26	20	9	25	16
Minnesota Twins	19	13	15	29	10
Arizona Diamondbacks	12	14	18	17	8
Chicago Cubs	5	17	22	23	12
Los Angeles Angels	6	11	25	18	25
Miami Marlins	15	24	28	13	27
Colorado Rockies	2	22	14	12	22
Texas Rangers	16	8	13	20	21
Detroit Tigers	10	30	30	26	30
Kansas City Royals	1	26	24	22	20
Cincinnati Reds	21	19	23	21	23
Pittsburgh Pirates	8	18	27	28	18
Oakland Athletics	24	27	29	24	28
Washington Nationals	14	28	26	30	26

Harper Messer
Math 2820L
4/22/2023

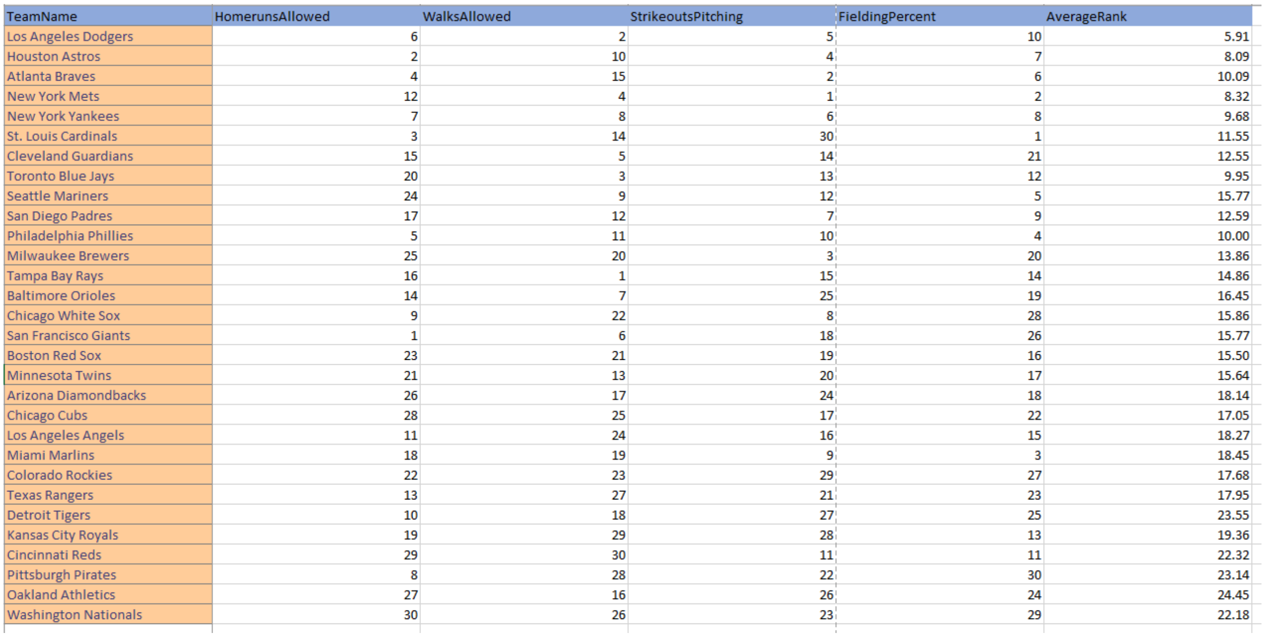


TeamName	StrikeoutsBatting	StolenBases	BattingAverage	OnBasePercentage	Slugging	E
Los Angeles Dodgers	15	9	4	1	2	
Houston Astros	2	16	12	7	5	
Atlanta Braves	29	15	9	10	1	
New York Mets	3	23	2	2	8	
New York Yankees	18	8	15	5	4	
St. Louis Cardinals	5	11	10	4	7	
Cleveland Guardians	1	3	6	12	21	
Toronto Blue Jays	6	21	1	3	3	
Seattle Mariners	20	17	28	16	15	
San Diego Padres	9	27	16	8	22	
Philadelphia Phillies	13	5	8	9	6	
Milwaukee Brewers	27	10	21	14	10	
Tampa Bay Rays	19	13	18	20	25	
Baltimore Orioles	17	12	20	22	16	
Chicago White Sox	7	24	5	18	19	
San Francisco Giants	26	22	23	15	14	
Boston Red Sox	14	26	3	6	9	
Minnesota Twins	12	30	13	11	11	
Arizona Diamondbacks	11	7	27	24	20	
Chicago Cubs	25	4	19	17	18	
Los Angeles Angels	30	19	24	26	17	
Miami Marlins	22	2	26	27	28	
Colorado Rockies	10	29	7	13	12	
Texas Rangers	24	1	17	25	13	
Detroit Tigers	21	28	25	29	29	
Kansas City Royals	8	6	14	21	23	
Cincinnati Reds	23	25	22	23	26	
Pittsburgh Pirates	28	14	29	28	27	
Oakland Athletics	16	18	30	30	30	
Washington Nationals	4	20	11	19	24	



TeamName	EarnedRunsAverage	Saves	Holds	Shutouts	HitsAllowed	Ho
Los Angeles Dodgers	1	12	26	8	1	
Houston Astros	2	2	30	2	2	
Atlanta Braves	5	1	25	20	4	
New York Mets	7	15	23	1	10	
New York Yankees	3	7	27	5	3	
St. Louis Cardinals	10	22	21	29	16	
Cleveland Guardians	6	4	28	22	7	
Toronto Blue Jays	14	8	17	13	21	
Seattle Mariners	8	17	29	16	11	
San Diego Padres	11	5	24	6	9	
Philadelphia Phillies	17	14	22	9	14	
Milwaukee Brewers	12	3	19	11	5	
Tampa Bay Rays	4	10	5	14	8	
Baltimore Orioles	18	9	10	7	25	
Chicago White Sox	16	6	18	10	15	
San Francisco Giants	13	18	1	23	24	
Boston Red Sox	25	19	8	17	26	
Minnesota Twins	19	29	3	4	13	
Arizona Diamondbacks	23	25	11	19	20	
Chicago Cubs	20	11	4	12	18	
Los Angeles Angels	9	20	20	3	6	
Miami Marlins	15	16	12	15	12	
Colorado Rockies	30	13	15	26	30	
Texas Rangers	22	23	14	18	19	
Detroit Tigers	21	21	16	24	17	
Kansas City Royals	27	27	7	21	29	
Cincinnati Reds	28	28	9	27	22	
Pittsburgh Pirates	26	26	2	30	27	
Oakland Athletics	24	24	13	25	23	
Washington Nationals	29	30	6	28	28	

Harper Messer
Math 2820L
4/22/2023



TeamName	HomerunsAllowed	WalksAllowed	StrikeoutsPitching	FieldingPercent	AverageRank
Los Angeles Dodgers	6	2	5	10	5.91
Houston Astros	2	10	4	7	8.09
Atlanta Braves	4	15	2	6	10.09
New York Mets	12	4	1	2	8.32
New York Yankees	7	8	6	8	9.68
St. Louis Cardinals	3	14	30	1	11.55
Cleveland Guardians	15	5	14	21	12.55
Toronto Blue Jays	20	3	13	12	9.95
Seattle Mariners	24	9	12	5	15.77
San Diego Padres	17	12	7	9	12.59
Philadelphia Phillies	5	11	10	4	10.00
Milwaukee Brewers	25	20	3	20	13.86
Tampa Bay Rays	16	1	15	14	14.86
Baltimore Orioles	14	7	25	19	16.45
Chicago White Sox	9	22	8	28	15.86
San Francisco Giants	1	6	18	26	15.77
Boston Red Sox	23	21	19	16	15.50
Minnesota Twins	21	13	20	17	15.64
Arizona Diamondbacks	26	17	24	18	18.14
Chicago Cubs	28	25	17	22	17.05
Los Angeles Angels	11	24	16	15	18.27
Miami Marlins	18	19	9	3	18.45
Colorado Rockies	22	23	29	27	17.68
Texas Rangers	13	27	21	23	17.95
Detroit Tigers	10	18	27	25	23.55
Kansas City Royals	19	29	28	13	19.36
Cincinnati Reds	29	30	11	11	22.32
Pittsburgh Pirates	8	28	22	30	23.14
Oakland Athletics	27	16	26	24	24.45
Washington Nationals	30	26	23	29	22.18

Harper Messer
Math 2820L
4/22/2023

Explanatory Analysis:

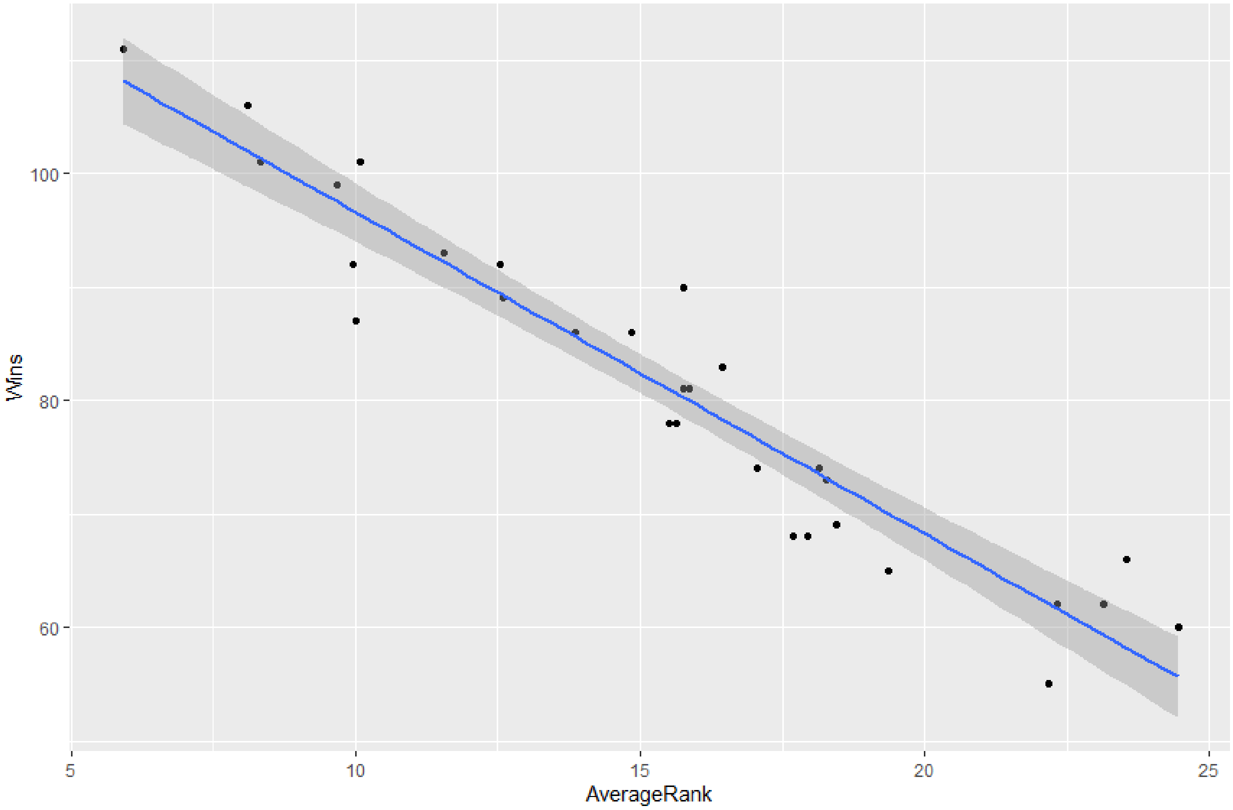
Looking at the dataset I can already see statistics that I believe will have strong correlation to winning. The top three I found were: Earned Runs Average, Runs and Runs Batted In. I know I will need to place an emphasis on these statistics yields because they are common statistics that contemporary baseball teams and analysis use to rate players and teams as baseball is a game of scoring and these statistics are the most direct correlation to scoring their is for baseball. I can already observe that the top 10 teams based on wins 7/10 are top 10 in runs and runs batted in and 8/10 are top in earned runs average. For the teams that had top 10 wins but didnt have top 10 rankings in runs/runs batted in, they had top 10 in earned runs average and vice versa for not having a top 10 in earned runs average. This is basically a balancing act for these winning teams. Looking at a team like the Los Angeles Angels who have the 9th best ERA in 2022, but was the 21st ranked team out of 30 and this is because their runs and runs batted in ranked 25th. Another interesting metric is the Texas Rangers who were #8 in homeruns and we all know that homeruns are guaranteed 1 or more runs yet the Rangers were the 24th best team in the league and I noticed that their On Base Percentage is 25th in the league meaning a good proportion of these homeruns they hit were solo homeruns, hence only scoring 1 run and on top of that their team ERA ranked 22nd. This believes me to also believe that On Base Percentage will generate some strongly positive correlated winning results as only 2 teams out of the top 10 winning teams have a not top 10 ranking in On Base Percentage. Some rather various metrics are triples and stolen bases where some of the best teams rank the worse in these categories and vice versa and this leads me to think they may have a neutral dependency on winning.

Harper Messer
Math 2820L
4/22/2023

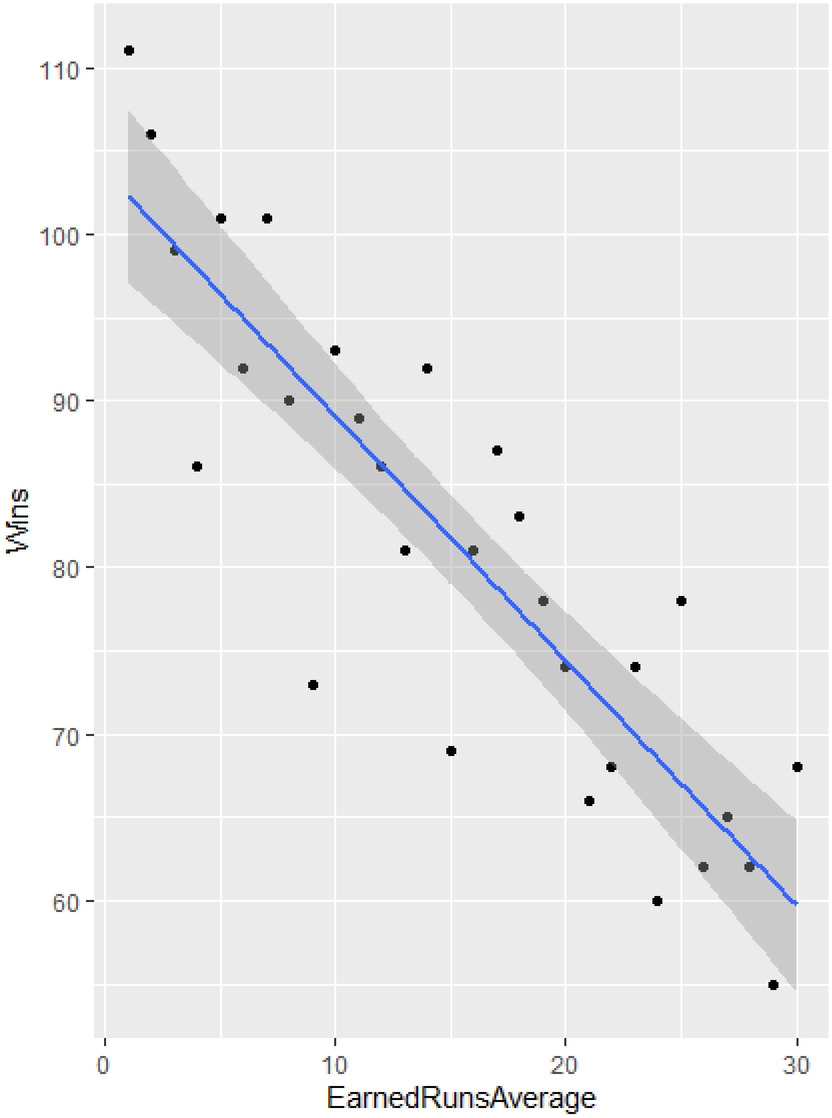
Model:

For my analysis, I constructed linear regression graphs and 95% confidence intervals for each statistic to determine the equation of the regression line for each statistic and its linear dependency. I decided to do separate graphs for each because for my analysis I am trying to determine which statistics correlate strongest/weakest with winning so having multiple stats together wouldn't help me produce conclusive graphs and results. Each graph has Wins out of 162 (not as a rank 1-30) on the y axis and then the x axis contains the respective statistic rank out of 30. For example, the most top left dot in the Earned Runs Average vs Wins graph is the Los Angeles Dodgers because they ranked #1 in Earned Runs Average and had the most wins in the 2022 MLB season with 111. After constructing each individual graph and calculating the confidence interval, I put the equation for each statistic's regression line and confidence interval into an spreadsheet so it was easier to visualize the equations of the lines and make comparisons when they are all in one spreadsheet to analyze. I also color coded it so that Green Statistics correlate to negative linear dependency meaning strong positive correlation to winning, Yellow Statistics correlate with a neutral linear dependency meaning the confidence interval contains 0 and linear dependence can't be concluded, and then Red Statistics correlate with a positive linear dependency meaning strong negative correlation to winning. I also added in a statistic for Wins based on Average Rank in all statistics. I thought this would be interesting to include because hypothetically speaking if a team is the best at everything they should win every game, but in the conclusion I will dive into this more because it is a flawed point of view. With all of this being said, my analysis of the strongest correlated statistics for winning is driven by these regression lines and the respective confidence intervals. The higher the initial intercept and the steeper the slope is in the downward direction correlates with the statistics that strongest correlate with winning the most.

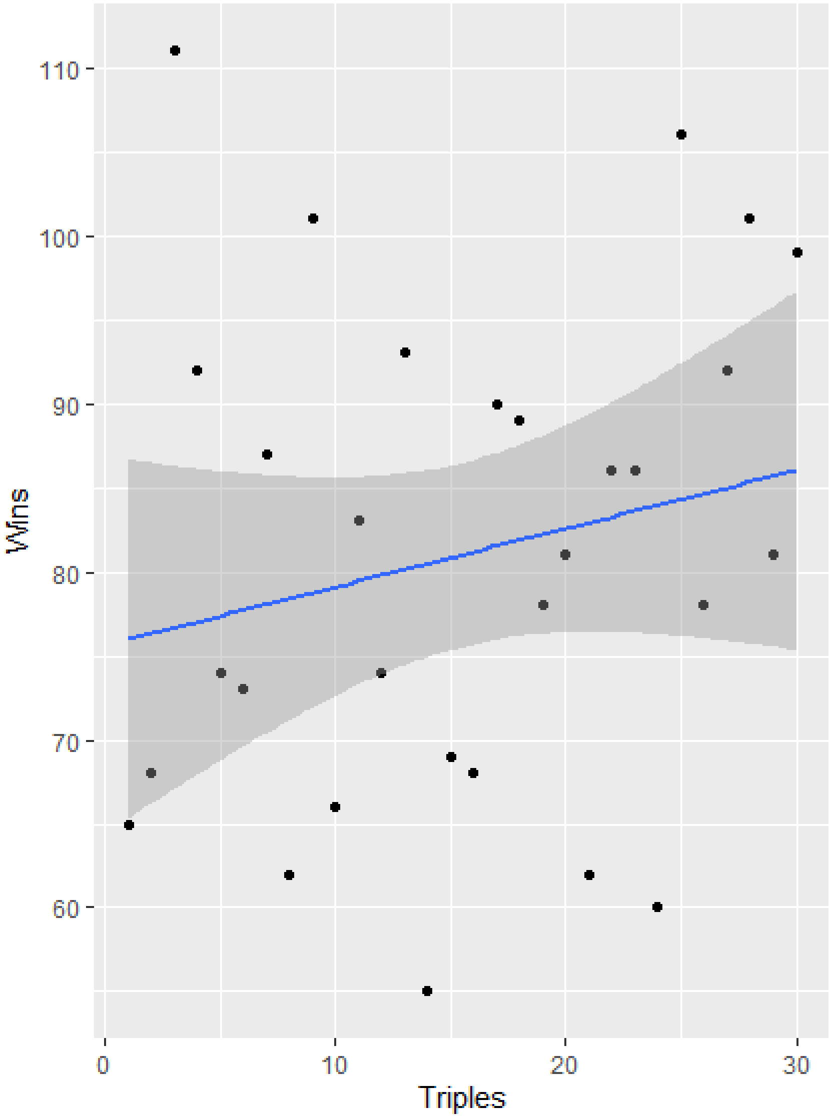
Harper Messer
Math 2820L
4/22/2023



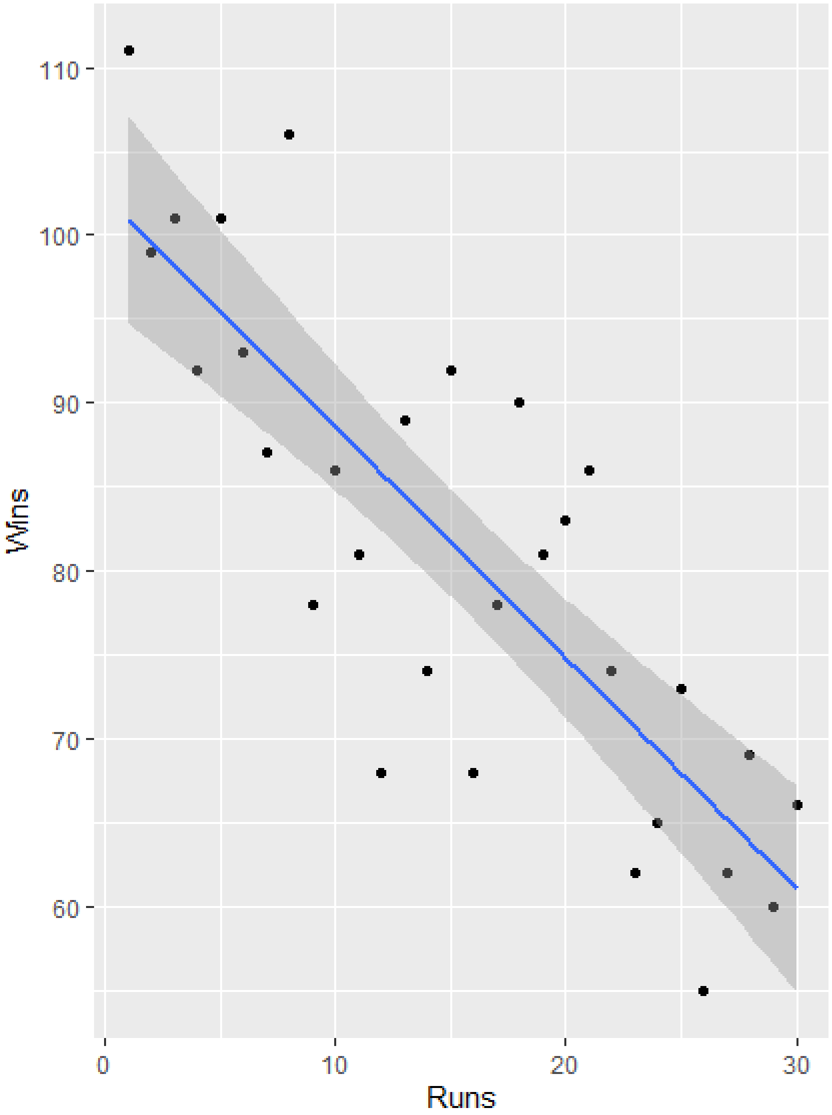
Harper Messer
Math 2820L
4/22/2023



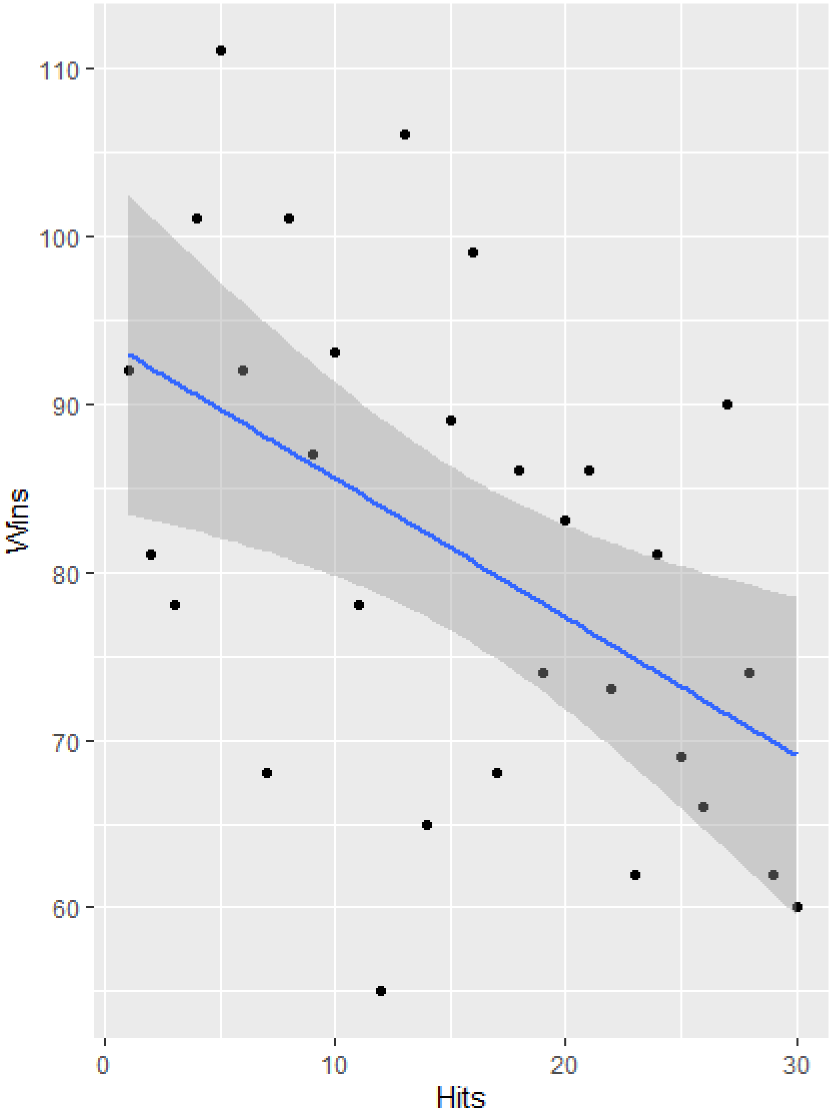
Harper Messer
Math 2820L
4/22/2023



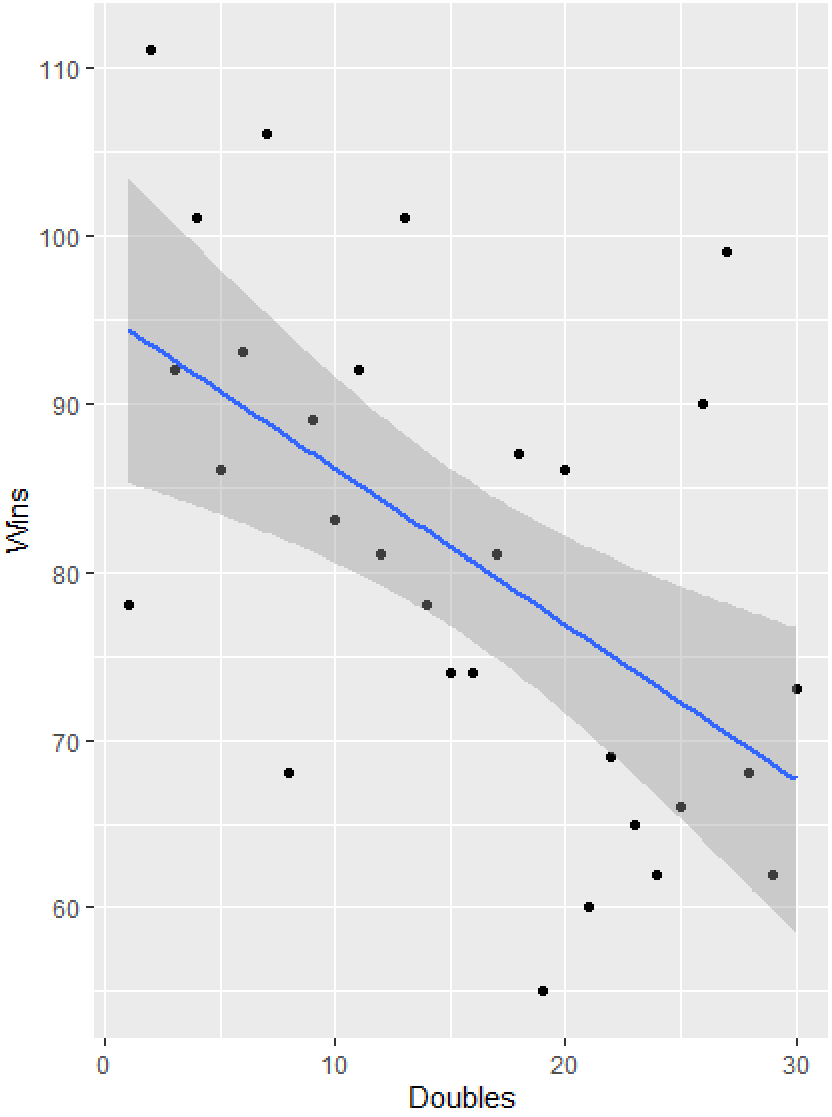
Harper Messer
Math 2820L
4/22/2023



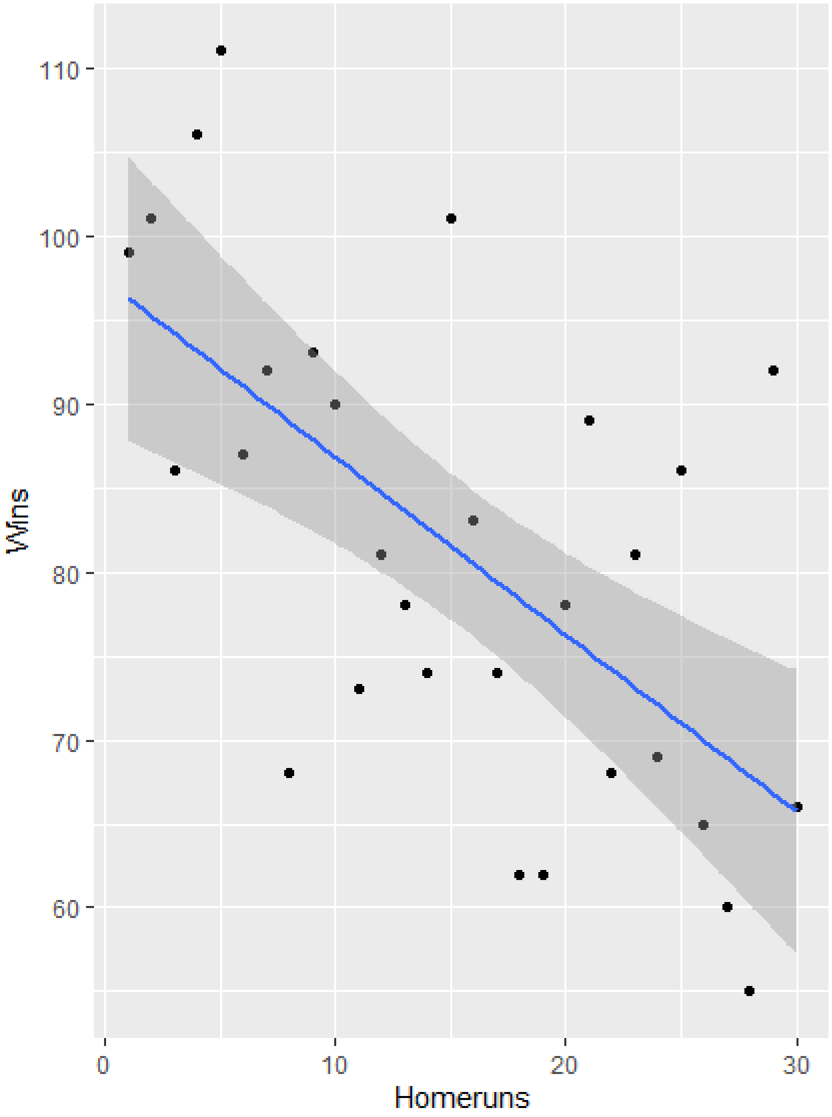
Harper Messer
Math 2820L
4/22/2023



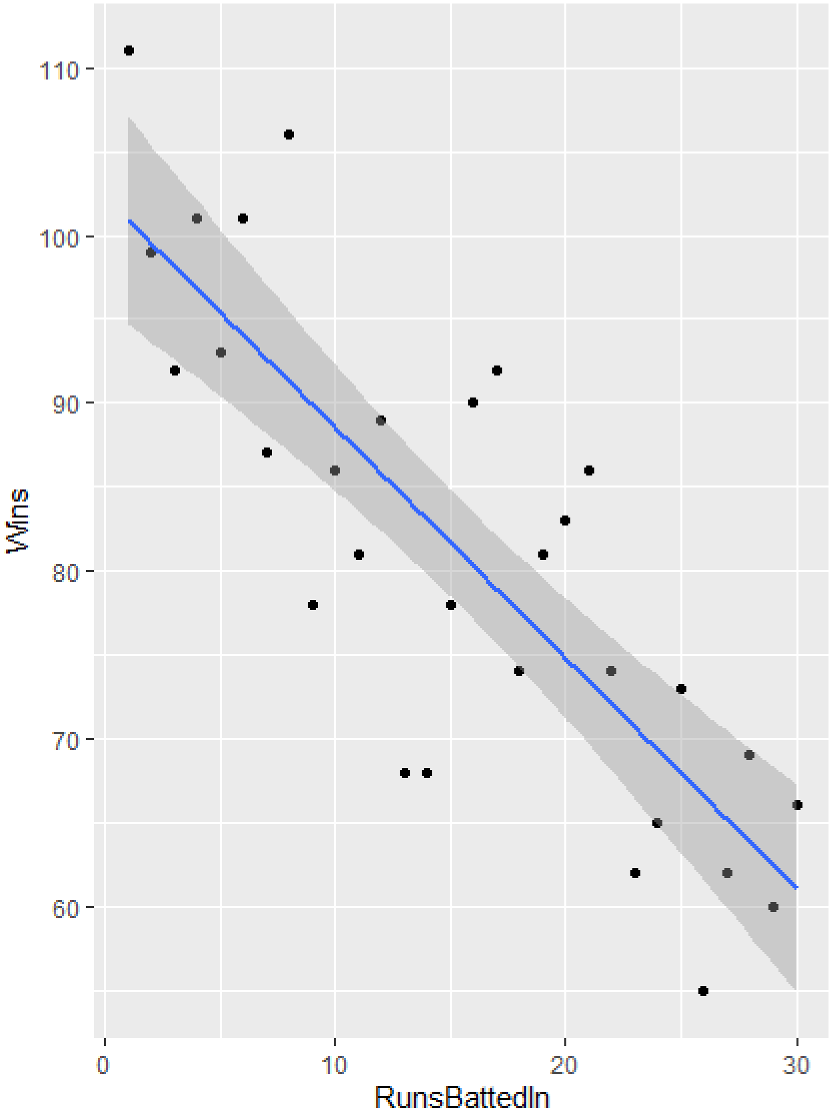
Harper Messer
Math 2820L
4/22/2023



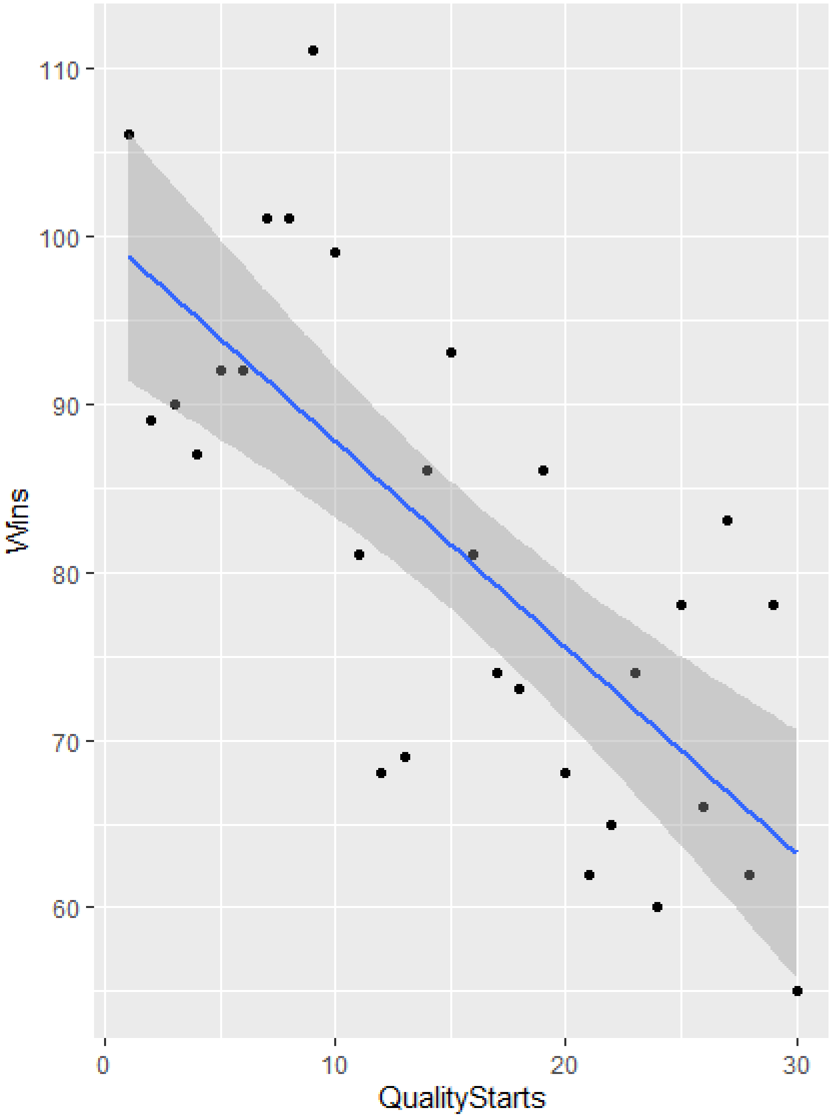
Harper Messer
Math 2820L
4/22/2023



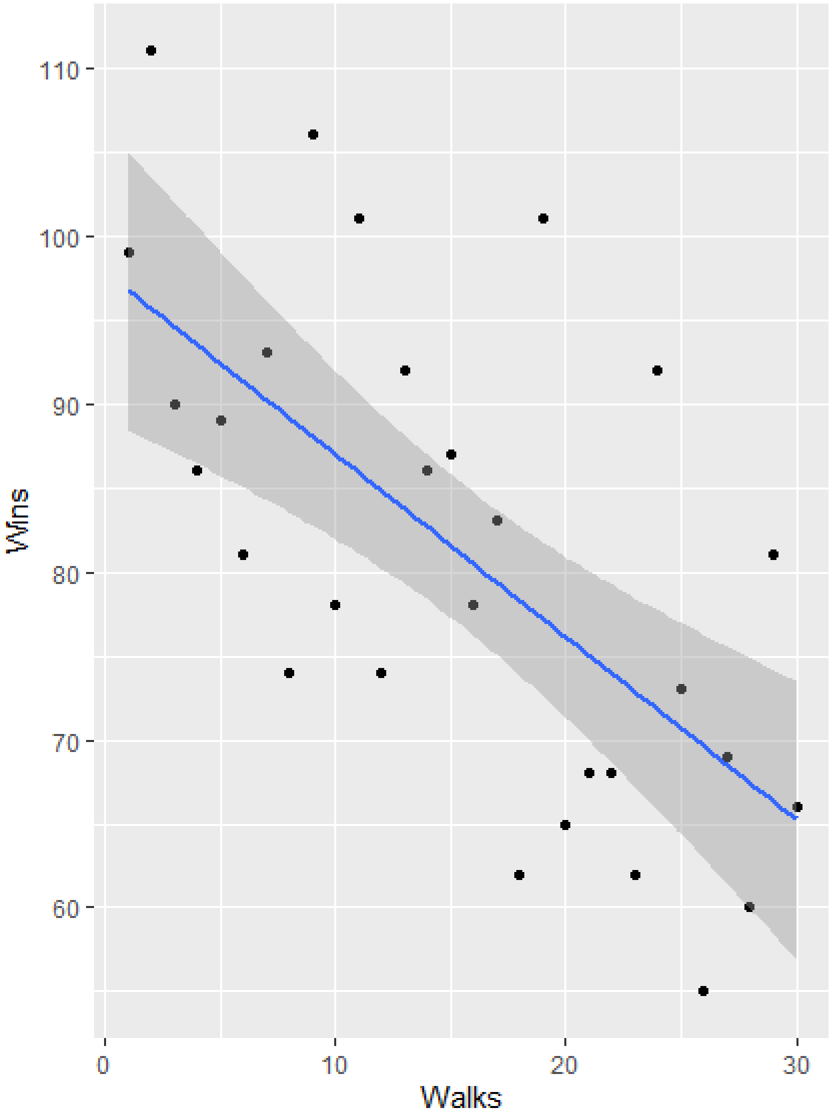
Harper Messer
Math 2820L
4/22/2023



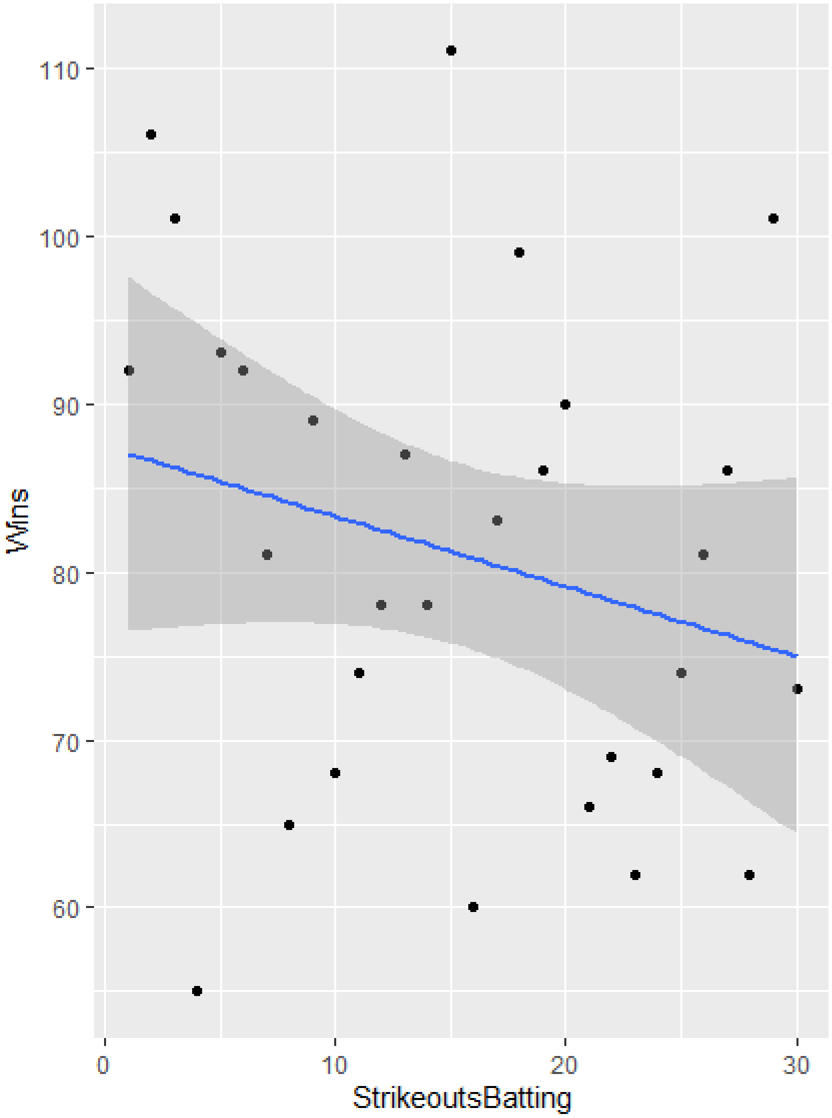
Harper Messer
Math 2820L
4/22/2023



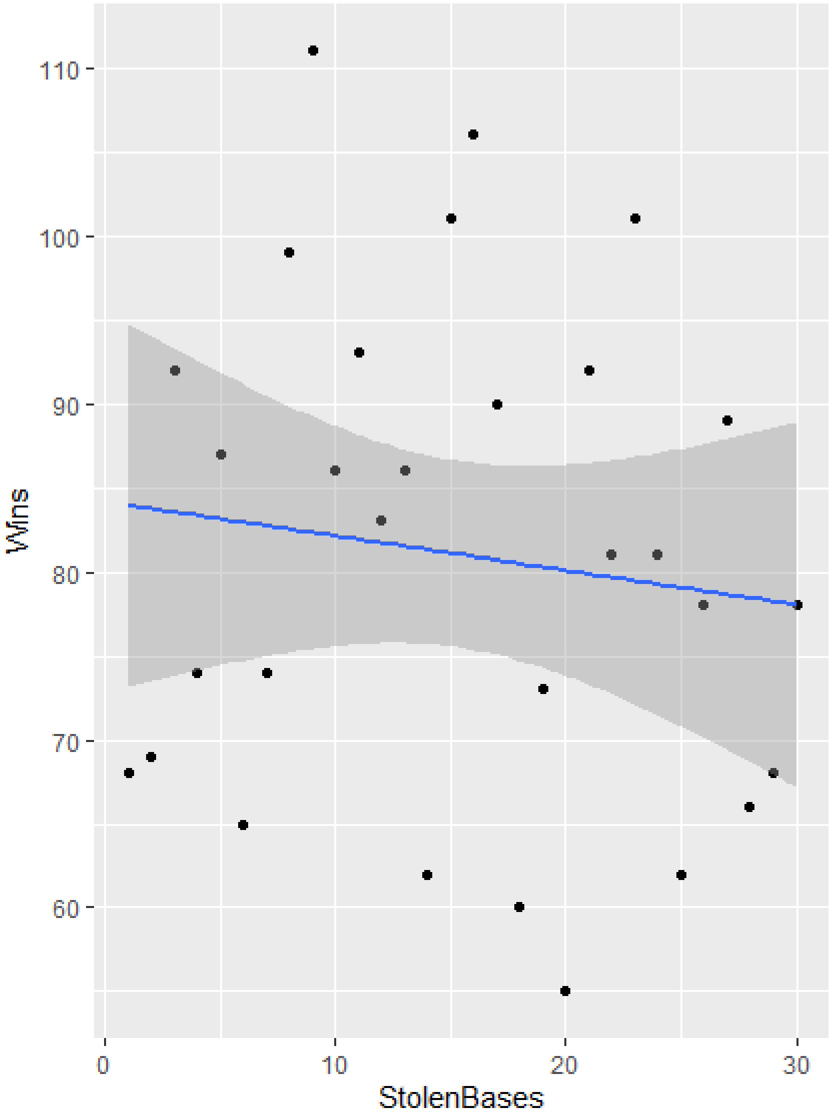
Harper Messer
Math 2820L
4/22/2023



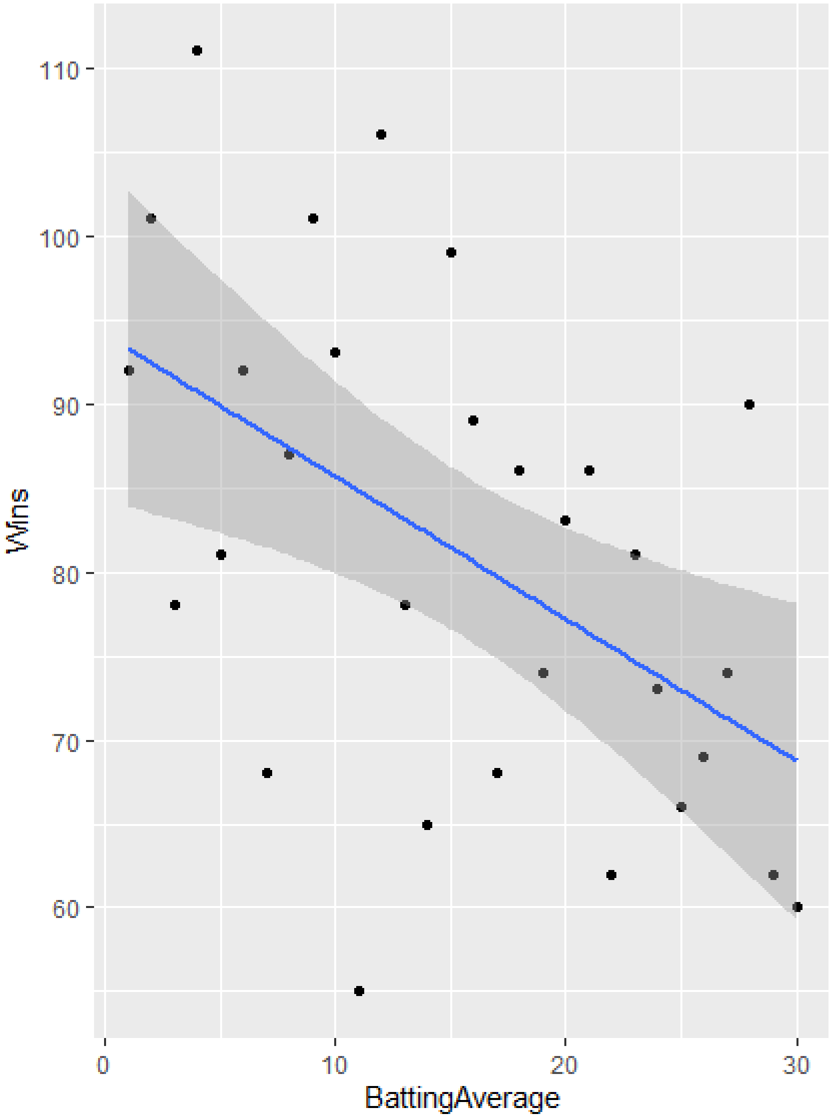
Harper Messer
Math 2820L
4/22/2023



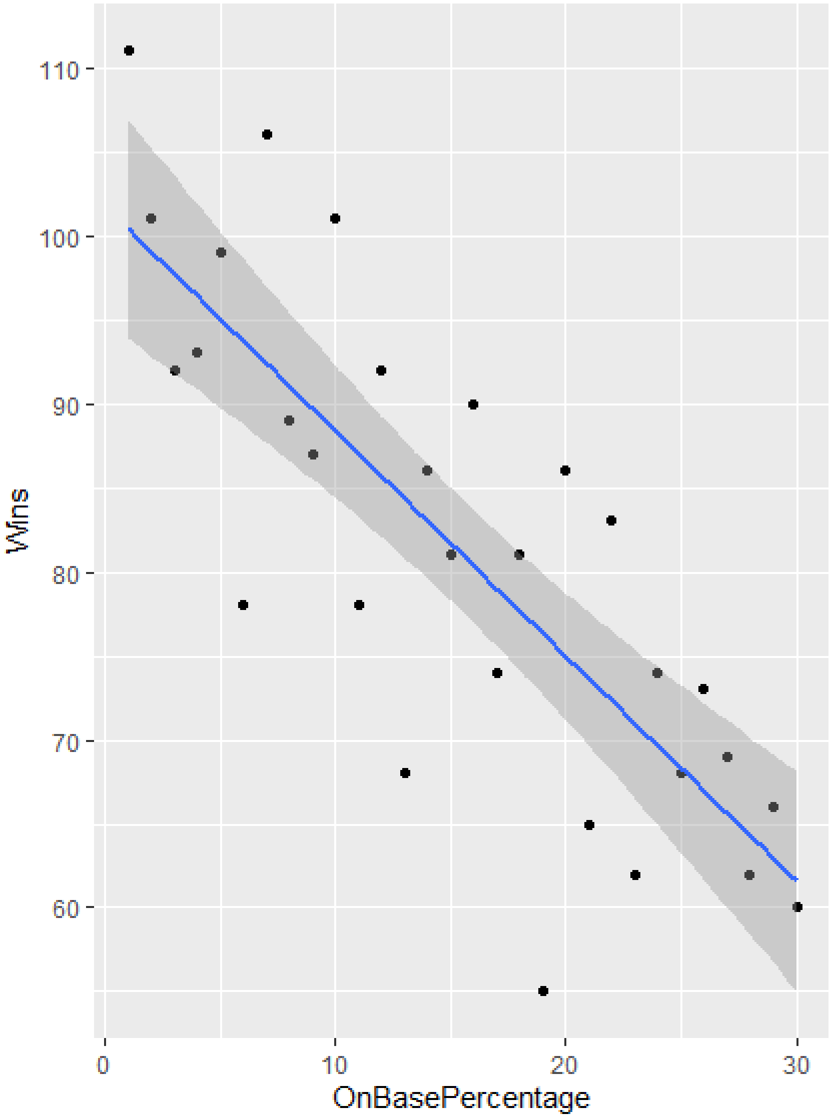
Harper Messer
Math 2820L
4/22/2023



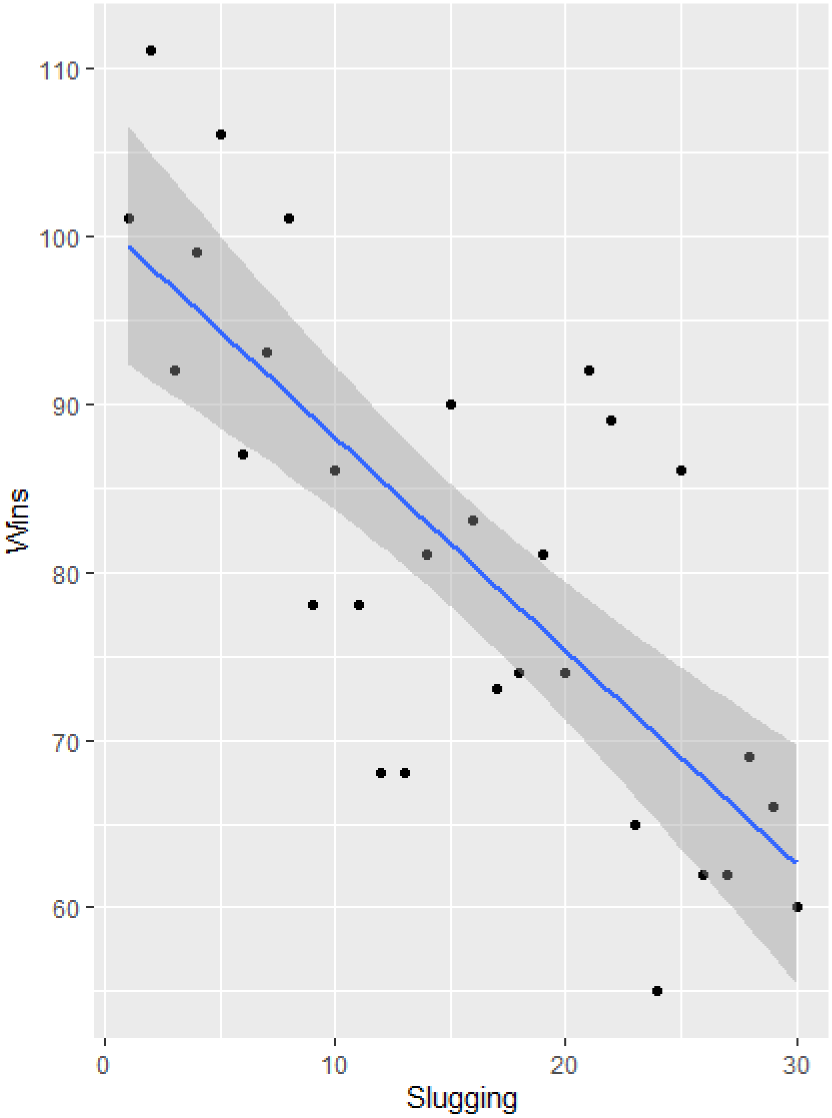
Harper Messer
Math 2820L
4/22/2023



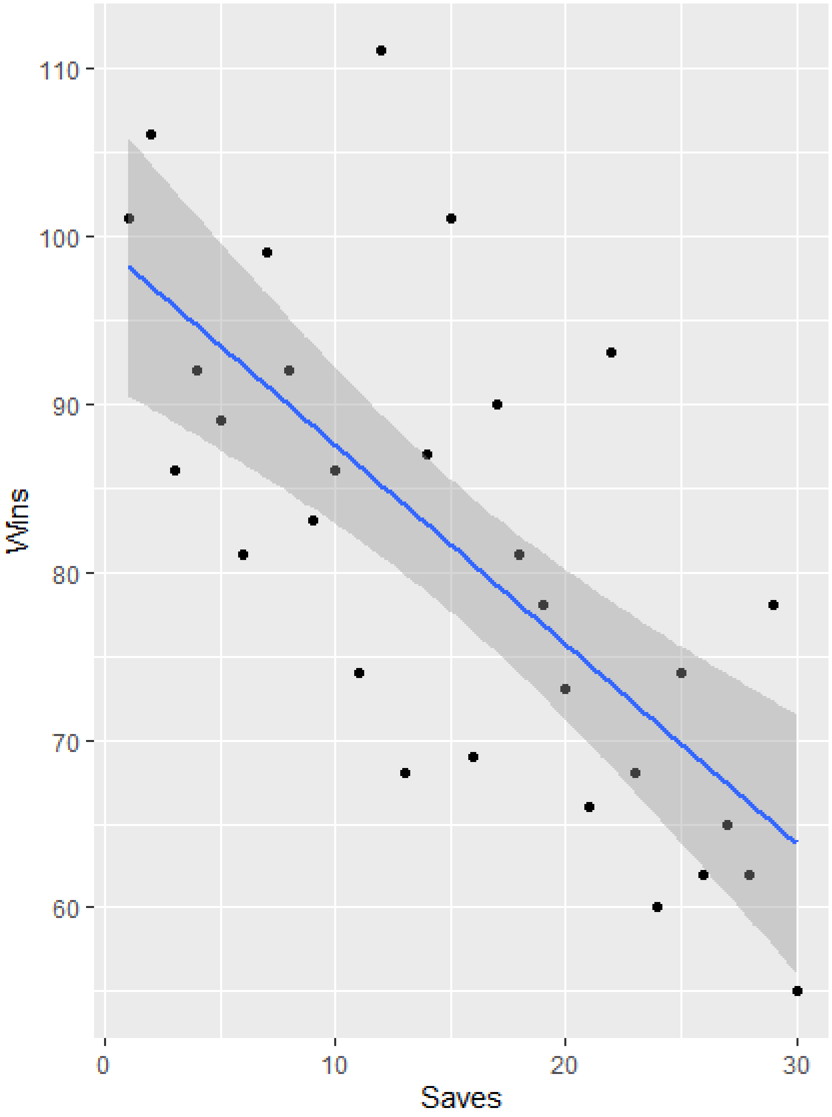
Harper Messer
Math 2820L
4/22/2023



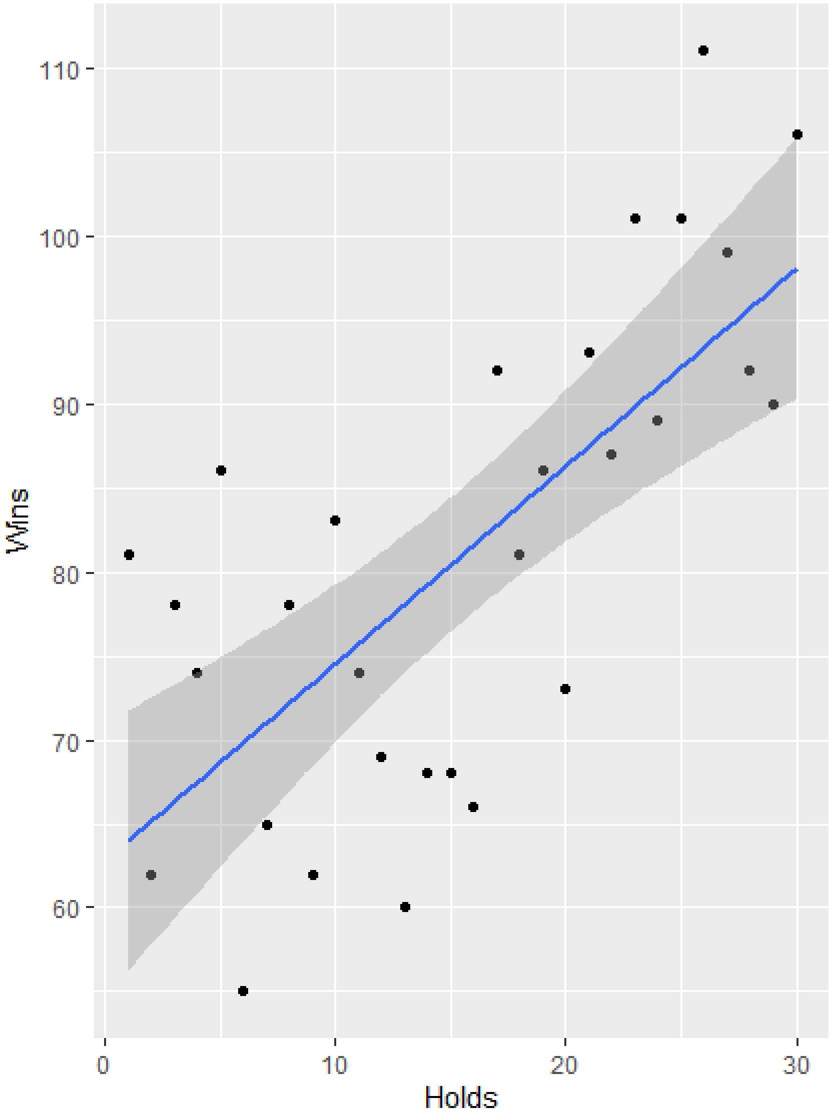
Harper Messer
Math 2820L
4/22/2023



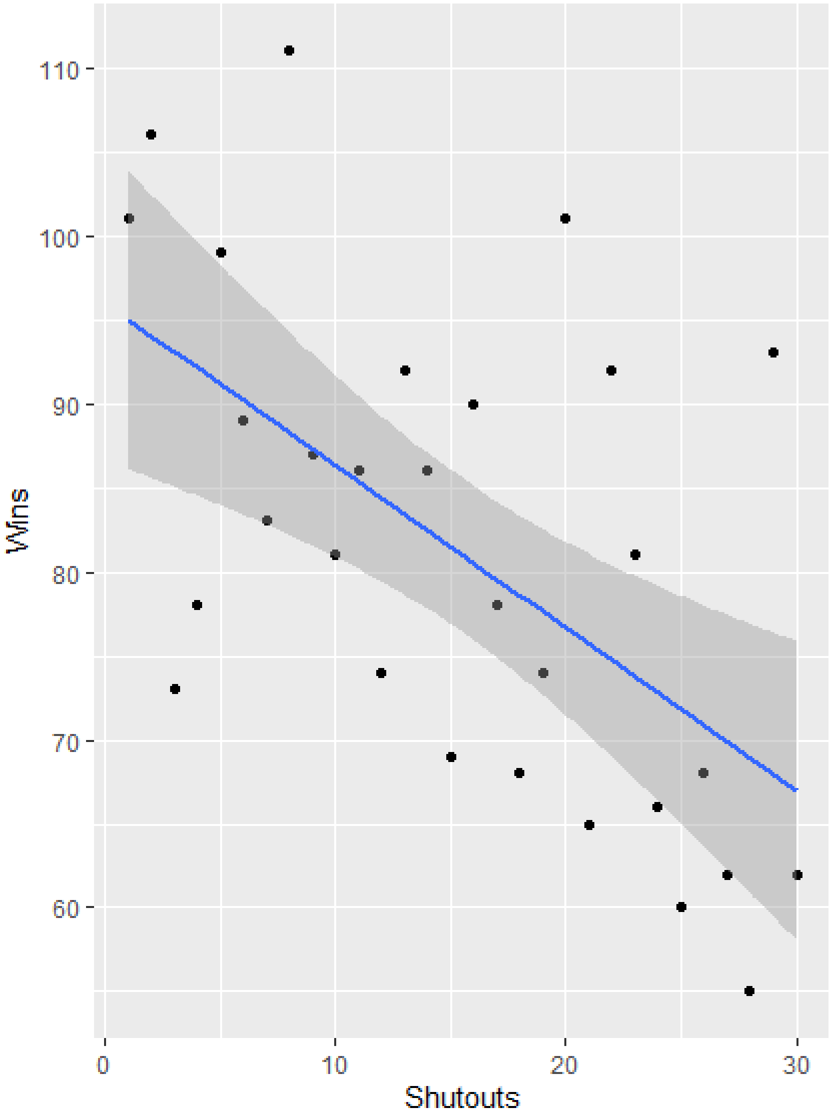
Harper Messer
Math 2820L
4/22/2023



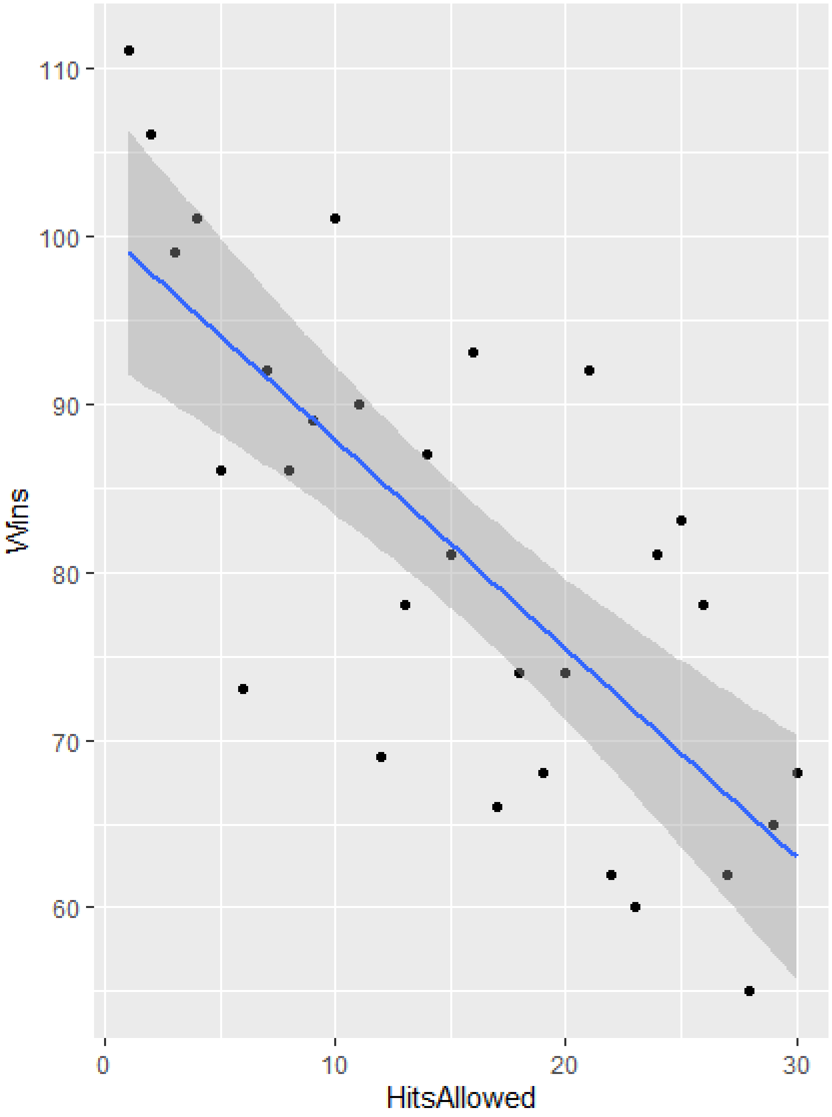
Harper Messer
Math 2820L
4/22/2023



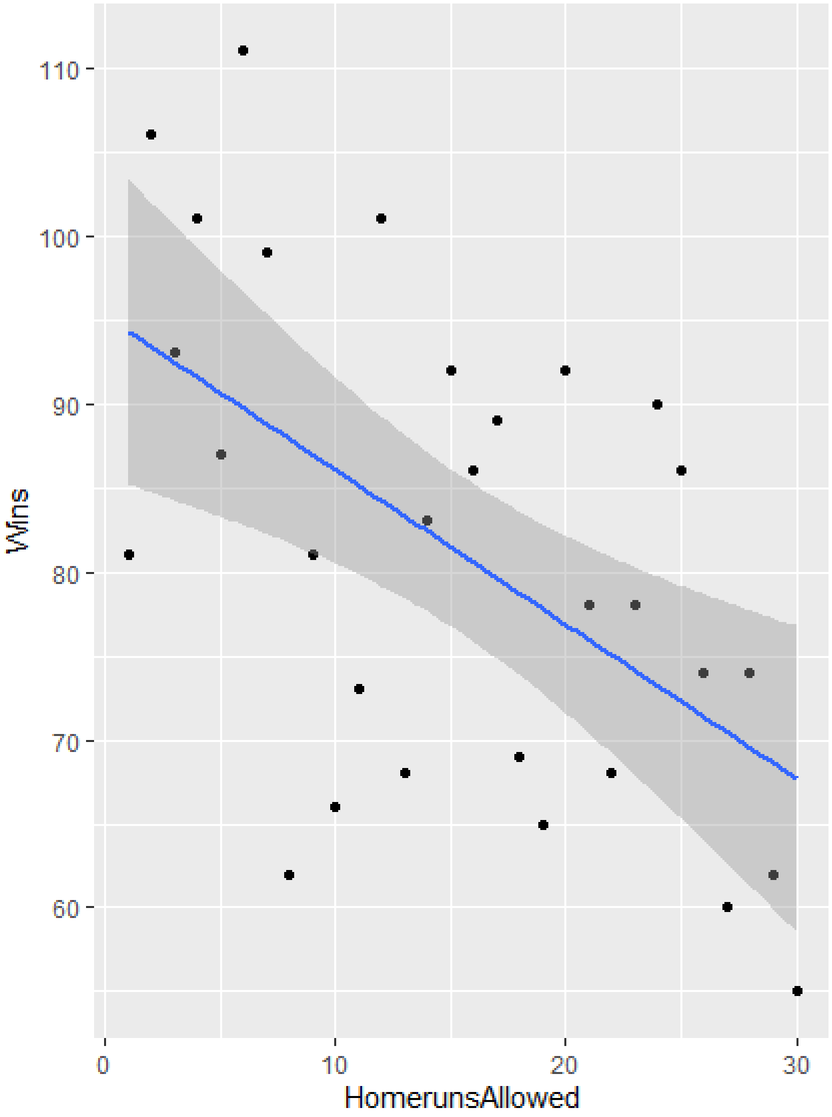
Harper Messer
Math 2820L
4/22/2023



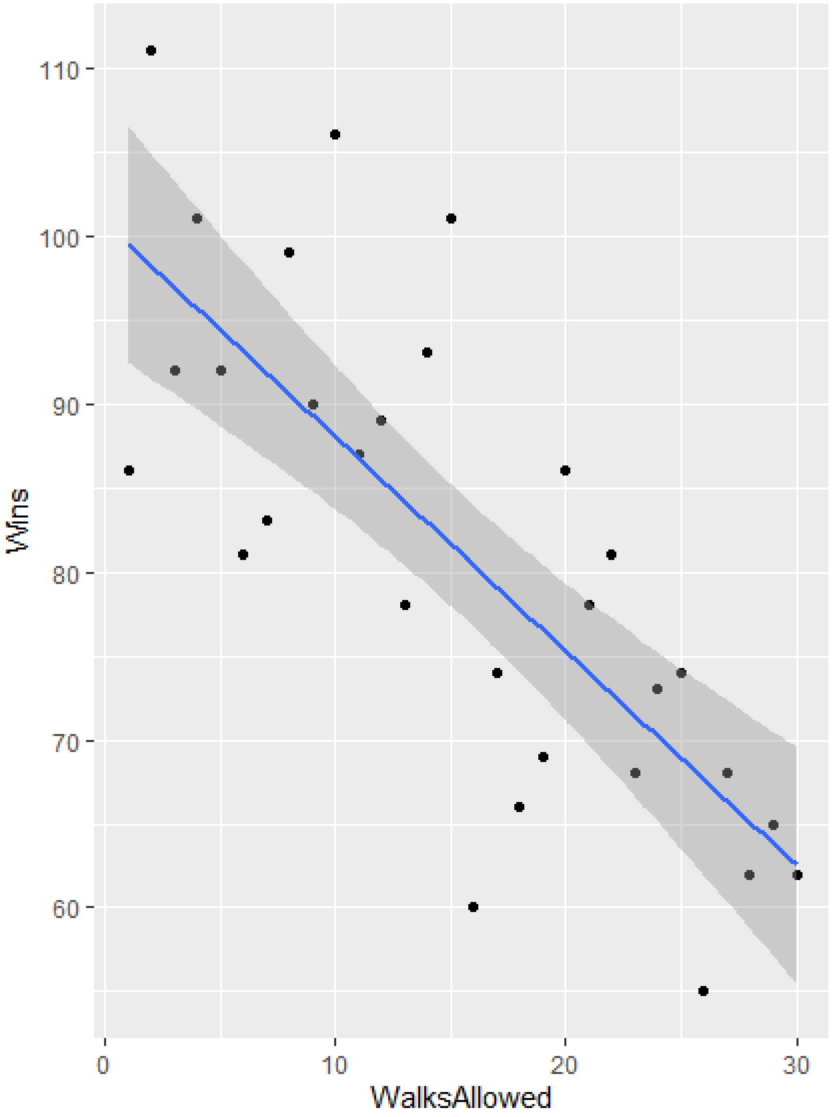
Harper Messer
Math 2820L
4/22/2023



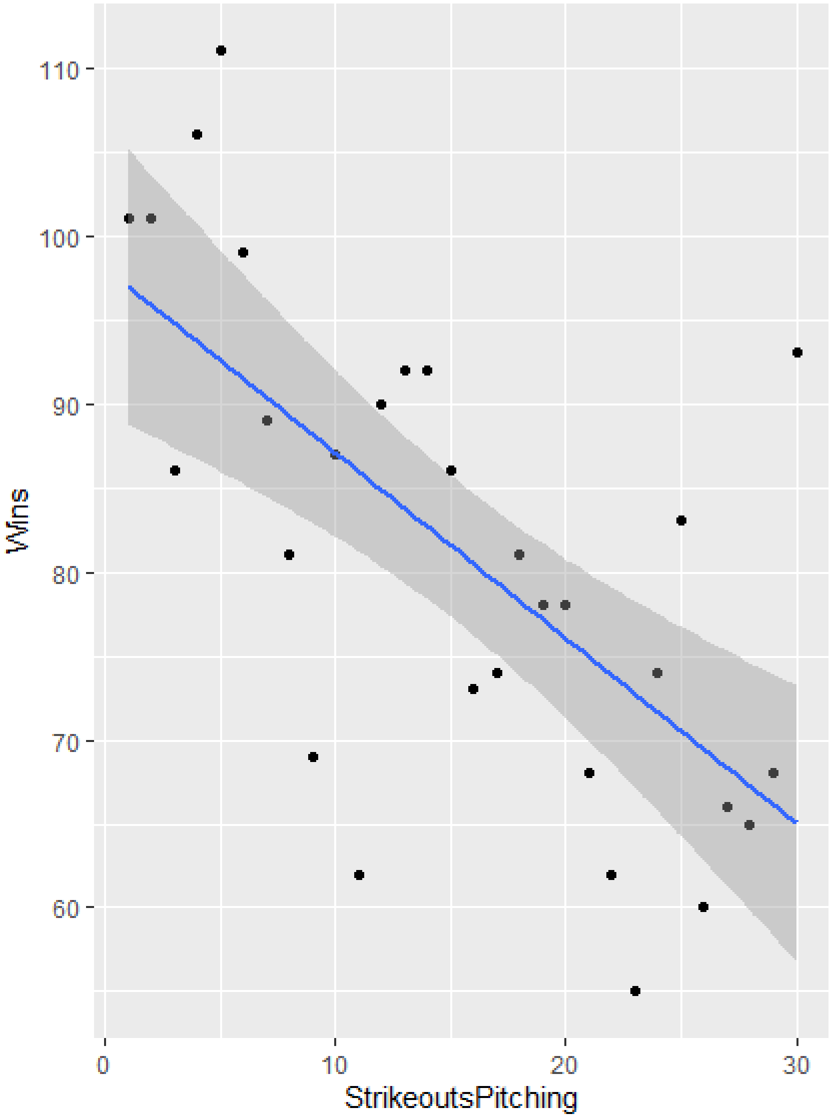
Harper Messer
Math 2820L
4/22/2023



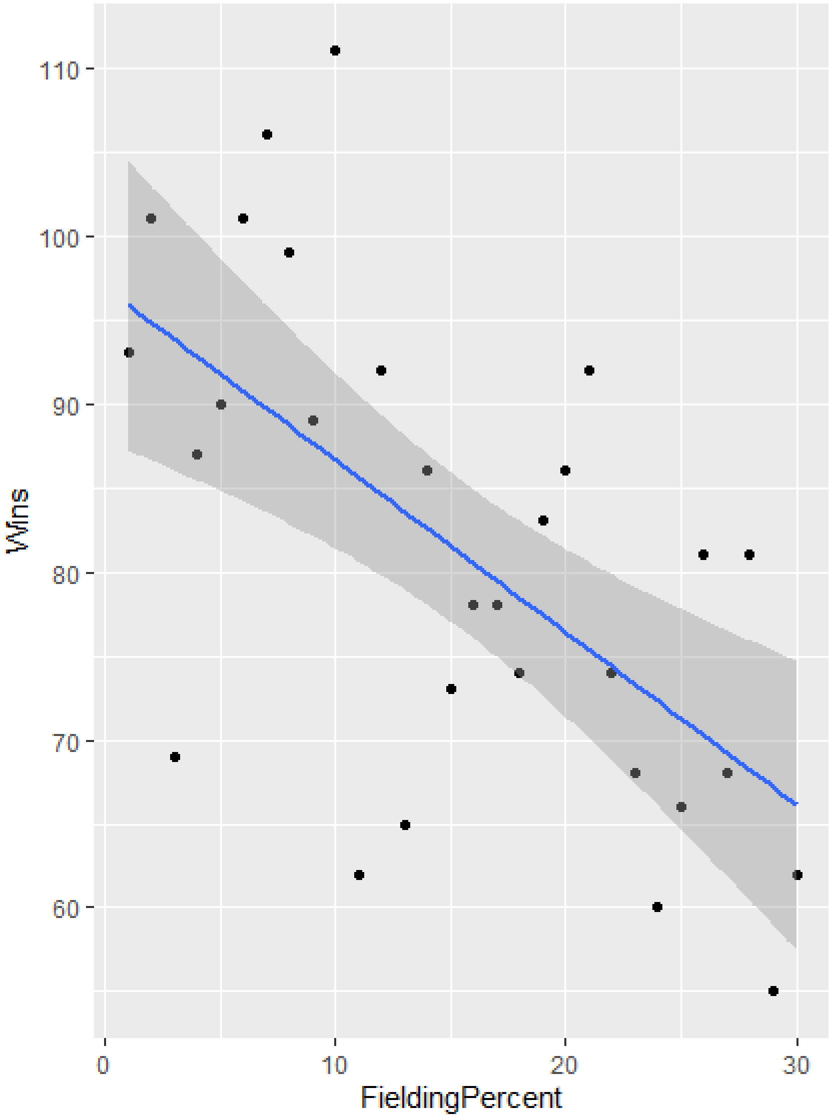
Harper Messer
Math 2820L
4/22/2023



Harper Messer
Math 2820L
4/22/2023

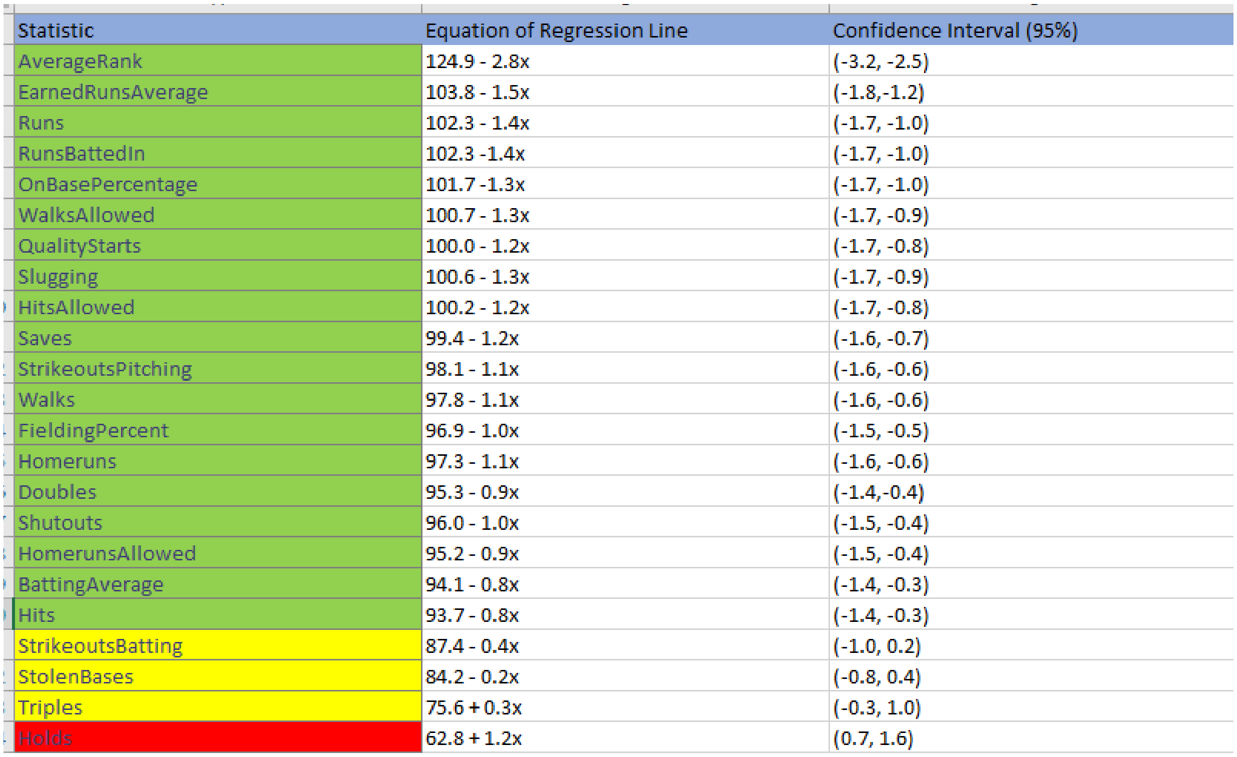


Harper Messer
Math 2820L
4/22/2023



Harper Messer
Math 2820L
4/22/2023

Conclusion



Statistic	Equation of Regression Line	Confidence Interval (95%)
AverageRank	124.9 - 2.8x	(-3.2, -2.5)
EarnedRunsAverage	103.8 - 1.5x	(-1.8,-1.2)
Runs	102.3 - 1.4x	(-1.7, -1.0)
RunsBattedIn	102.3 -1.4x	(-1.7, -1.0)
OnBasePercentage	101.7 -1.3x	(-1.7, -1.0)
WalksAllowed	100.7 - 1.3x	(-1.7, -0.9)
QualityStarts	100.0 - 1.2x	(-1.7, -0.8)
Slugging	100.6 - 1.3x	(-1.7, -0.9)
HitsAllowed	100.2 - 1.2x	(-1.7, -0.8)
Saves	99.4 - 1.2x	(-1.6, -0.7)
StrikeoutsPitching	98.1 - 1.1x	(-1.6, -0.6)
Walks	97.8 - 1.1x	(-1.6, -0.6)
FieldingPercent	96.9 - 1.0x	(-1.5, -0.5)
Homeruns	97.3 - 1.1x	(-1.6, -0.6)
Doubles	95.3 - 0.9x	(-1.4,-0.4)
Shutouts	96.0 - 1.0x	(-1.5, -0.4)
HomerunsAllowed	95.2 - 0.9x	(-1.5, -0.4)
BattingAverage	94.1 - 0.8x	(-1.4, -0.3)
Hits	93.7 - 0.8x	(-1.4, -0.3)
StrikeoutsBatting	87.4 - 0.4x	(-1.0, 0.2)
StolenBases	84.2 - 0.2x	(-0.8, 0.4)
Triples	75.6 + 0.3x	(-0.3, 1.0)
Holds	62.8 + 1.2x	(0.7, 1.6)

This spreadsheet above is a complete compilation of the statistic's regression lines and confidence intervals. As you can see, Earned Runs Average is the strongest correlated statistic with winning followed by Runs/Runs Batted In, On Base Percentage and Walks Allowed. So MLB teams need to focus on lowering their team ERA and walks allowed, while raising their Quality Starts, Runs, Runs Batted In and On Base Percentage in order to win more games. In a game where scoring is the main objective, I am not surprised that these statistics are the ones most correlated with winning. I was surprised about Walks Allowed being so high and being over popular statistics like Slugging and Hits Allowed by a narrow margin, but in baseball pitchers can only throw so many pitches and when a pitcher walks a batter that is at minimum 4 pitches thrown creating more fatigue for pitchers and therefore making a pitcher more prone to giving up more hits, walks and runs. Something that is interesting in the results is how Hits, Walks, and Batting Average all aren't top 10 winning statistics, but On Base Percentage is #4 when On Base Percentage is mostly a combination of waks and hits with sprinkle of getting on base due to hit by pitch, yet the marginal difference between the intercepts is 3.9 for Walks, 7.6

for BattingAverage and a whopping 8 points for Hits. Mainly here we can observe the difference between On Base Percentage compared to Batting Average. These two stats are always looked at as the same by many casual fans, but analyst and I now know that On Base Percentage is a way better metric for creating a winning baseball team.

Next looking at our 1 red statistic, which is Holds, we must deduce why this statistic is negatively correlated with winning. A hold in baseball is when a relief pitcher (non starter) enters a baseball game when his team has a lead of 3 or less or if the game tying run is on deck, at the plate or on the bases and then in order to receive the hold, the pitcher must record at least 1 out while maintaining his team's lead. Although this seems like a statistic a team would want the most in, it actually isn't and relizing this is that team's with the most holds are always exhibiting close games and the more holds means the more pitchers brought into a game because their starting pitcher didn't pitch that many innings. Putting in more pitchers to a game will create fatigue for them in the next games, due to relievers throwing on shorter rest than starters and also since holds are only for small leads of 3 and under, these team with a bunch of holds are playing very close games meaning either their bats are cold or their pitching is bad. So based on this I can deduce that MLB Team's need to create less hold opportunities for their teams by having starter pitch longer and give up less runs and by increasing their run production so they can potentially have bigger lead differentials against opponents.

For our 3 yellow statistics of Triples, Strikeouts from Batting and Stolen Bases, I can see why Strikeouts and Triples are on there, but not really for Stolen Bases. Striking out in baseball is one of the worse things to do when hitting. It basically allows for a team's defense to get away with not having to field the ball and you potentially getting on base and as we can see, fielding percentage does have a positive correlation to winning. I was expecting strikeouts to be in the red, but then I remembered that the league's best hitters actually strike out more often than worse hitters and this is because these guys are trying to rip the ball out the park, totally opposite of just trying to put the ball in play. (In 2022 7 of the top 10 RBI batters had over 100 strikeouts)

Triples on the other hand are a great statistic and a very exciting aspect of baseball to see, but they are so rare which is why Triples can't exhibit either a positive or negative. In fact, no player hit over 9 triples in 2022 and no team hit over 38 meaning that the best triple hitting teams only saw a triple once every 4 games basically which is why their is a neutral correlation between triples and winning. Then for Stolen Bases, this was more surprising than Strikeouts, but this elicited the same analysis as Triples. The league leader in stolen bases in 2022 was the Texas Rangers with 128 meaning they didn't even steal a base once a game showing how little stolen bases actually appear in modern baseball. So it can be concluded that due to stolen bases and triples little appearance in day to day games along with the variance in top teams numbers in these categories, that a correlation to winning cannot be deduced.

Something that I included just for fun was the Average Rank graph. This is an analysis where I added up every team's rank in every category and divided it by total statistics to find a teams average statistical ranking and of course this projection yielded the strongest regression line and strongest correlation with winning, but it still only had a line of 124.9- 2.8x. This means that in this analysis with all of these statistics, if a team led in every one of these categories they would win 122 games which is a 75% win rate, which goes against the motto of if your team is the best at everything then no one should be able to beat you, but in sports at the professional level, their is so much variance to the point where some sports analysis even consider weather or time of the day or day of the week when thinking if a team will win or not. Also we know we had Holds, Strikeouts, Triples, and Stolen Bases, all which had a neutral or negative correlation to winning, which is definitely driving down the intercept for the Average Rank regression line. I do wonder what the line would be if we only used our top 5 green statistics for it. This would be an opportunity for further analysis.

Some other opportunities for further analysis would be to include different statistics. I was at the gym the other day and someone told me that an interesting statistic to look at would be a team's batting average with runners in scoring position with 2 outs. This is a very specific statistic, but very important. I didn't include statistics like these because it takes a while to scrape the internet for this data and then on top of this, I only know beginner to intermediate baseball statistics, some of these more advanced metrics would've been harder for me to discuss. I also instead of holds wanted to find hold percentage and for saves find save percentage, which i believe are better metrics for analysis and better for winning in my opinion, hence maybe why holds is a red, but maybe hold percentage could be green because if a team only has a few holds on a season, but is 90% successful in these scenarios, their bullpen is strong and therefore they can hold their team's lead.