Explain the 'big question':

Our ‘big question’ for this project is to determine which team statistics have the highest correlation with wins in Major League Baseball (regular season). As a secondary task, we will also be identifying the regression equation and the overall quality of our model.

Baseball, like most team sports, is a zero-sum game. Every game played must have one winning team and one losing team. Of course, every team (players, management, ownership) and their respective fan bases are hoping to win as many games as possible. But professional sports are a multi-billion dollar industry with many, many stakeholders. With the legalization and expansion of online sports betting, people with limited interest in sports have also become engaged in statistical analysis.

Understanding which statistics significantly influence outcomes of games will help teams to:

- Identify team performance issues
- Make smart, data-driven financial decisions during player acquisitions and drafts, and ultimately devise more effective strategies. And those involved in gambling can also adjust their own strategies to better predict future games.

Introduce the dataset:

Origin of the dataset (From where did you obtain the data?):

The data we are using is an amalgamation of all publicly available quantitative data (from multiple spreadsheets) taken from Baseball-Reference.com, an online sports database that provides comprehensive statistics, records, and historical data for baseball players, teams, and games. Below is a partial sample of our data set.

Column1	Season	W	L	Payroll	Runs	Runs	Stre	Luck	#Bat	batAge	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	TB
Arizona Diamondbacks	2021	52	110	$ 91,632,929.00	4.2	5.5	0.2	-9	64	28.9	1297	308	31	144	644	43	16	537	1465	0.236	0.309	0.382	0.692	209
Atlanta Braves	2021	88	73	$ 152,750,691.00	4.9	4.1	-0.1	-6	56	28.2	1307	269	20	239	762	59	19	549	1453	0.244	0.319	0.435	0.754	233
Baltimore Orioles	2021	52	110	$ 42,421,870.00	4.1	5.9	0.3	-2	62	26.7	1296	266	15	195	632	54	23	451	1454	0.239	0.304	0.402	0.705	217
Boston Red Sox	2021	92	70	$ 187,100,784.00	5.1	4.6	0.1	4	56	28	1434	330	23	219	783	40	21	512	1386	0.261	0.328	0.449	0.777	246
Chicago Cubs	2021	71	91	$ 144,037,170.00	4.4	5.2	-0.1	3	69	29.1	1255	225	26	210	672	86	37	502	1596	0.237	0.312	0.407	0.719	216
Chicago White Sox	2021	93	69	$ 140,926,169.00	4.9	3.9	-0.2	-4	47	28	1373	275	22	190	757	57	20	586	1389	0.256	0.336	0.422	0.758	226
Cincinnati Reds	2021	83	79	$ 126,587,447.00	4.9	4.7	-0.2	0	55	28.9	1352	295	13	222	756	36	24	553	1425	0.249	0.328	0.431	0.759	233
Cleveland Indians	2021	80	82	$ 50,670,534.00	4.4	4.5	-0.1	0	48	26.7	1269	248	22	203	686	109	17	453	1387	0.238	0.303	0.407	0.71	217
Colorado Rockies	2021	74	87	$ 116,408,966.00	4.6	4.9	0.1	-1	45	28.1	1338	275	34	182	709	76	23	491	1356	0.249	0.317	0.414	0.731	222
Detroit Tigers	2021	77	85	$ 86,348,945.00	4.3	4.7	-0.1	2	49	28.1	1299	236	37	179	675	88	25	490	1514	0.242	0.308	0.399	0.707	214
Houston Astros	2021	95	67	$ 194,222,042.00	5.3	4.1	-0.1	-6	52	28.9	1496	299	14	221	834	53	16	569	1222	0.267	0.339	0.444	0.783	248
Kansas City Royals	2021	74	88	$ 91,595,545.00	4.2	4.9	0	3	48	29.3	1349	251	29	163	647	124	33	421	1258	0.249	0.306	0.396	0.702	214
Los Angeles Angels	2021	77	85	$ 183,849,560.00	4.5	5	0.2	4	64	29.2	1331	265	23	190	691	79	26	464	1394	0.245	0.31	0.407	0.717	221
Los Angeles Dodgers	2021	106	56	$ 266,020,809.00	5.1	3.5	-0.1	-3	61	29.2	1330	247	24	237	799	65	17	613	1408	0.244	0.33	0.429	0.759	233
Miami Marlins	2021	67	95	$ 58,157,900.00	3.8	4.3	0	-5	61	28.2	1244	226	23	158	594	106	29	450	1553	0.233	0.298	0.372	0.671	199
Milwaukee Brewers	2021	95	67	$ 99,377,415.00	4.6	3.8	-0.3	2	61	28.7	1251	255	18	194	700	82	21	586	1465	0.233	0.317	0.396	0.713	212
Minnesota Twins	2021	73	89	$ 120,084,606.00	4.5	5.1	0	2	57	28.3	1311	271	17	228	690	54	15	525	1405	0.241	0.314	0.423	0.738	230
New York Mets	2021	77	85	$ 201,189,189.00	3.9	4.1	0	0	64	28.2	1243	228	18	176	604	54	26	495	1392	0.239	0.315	0.391	0.705	203
New York Yankees	2021	92	70	$ 205,669,863.00	4.4	4.1	0.1	6	59	29.3	1266	213	12	222	666	63	18	621	1482	0.237	0.322	0.407	0.729	216
Oakland Athletics	2021	86	76	$ 90,400,598.00	4.6	4.2	0.1	-1	50	30.1	1284	271	19	199	698	88	20	545	1349	0.238	0.317	0.406	0.723	219
Philadelphia Phillies	2021	82	80	$ 197,263,223.00	4.5	4.6	0	2	55	29.1	1288	262	24	198	700	77	19	564	1402	0.24	0.318	0.408	0.726	219
Pittsburgh Pirates	2021	61	101	$ 54,356,609.00	3.8	5.1	0	3	64	27.5	1261	240	35	124	570	60	30	529	1328	0.236	0.309	0.364	0.673	194
San Diego Padres	2021	79	83	$ 179,764,272.00	4.5	4.4	0.1	-4	54	28	1305	273	21	180	695	110	39	586	1324	0.242	0.321	0.401	0.722	216
San Francisco Giants	2021	107	55	$ 171,890,308.00	5	3.7	-0.1	4	63	27	1209	233	11	199	673	64	24	535	1492	0.226	0.303	0.385	0.688	206
Seattle Mariners	2021	90	72	$ 83,822,113.00	4.3	4.6	0.1	14	54	30.6	1360	271	25	241	768	66	14	602	1461	0.249	0.329	0.44	0.769	240
St. Louis Cardinals	2021	90	72	$ 151,469,994.00	4.4	4.1	-0.2	5	51	28.5	1303	261	22	198	678	89	22	478	1341	0.244	0.313	0.412	0.725	220
Tampa Bay Rays	2021	100	62	$ 70,836,327.00	5.3	4	0	-1	61	27.7	1336	288	36	222	810	88	42	585	1542	0.243	0.321	0.429	0.75	236
Texas Rangers	2021	60	102	$ 95,788,819.00	3.9	5	0.2	-2	54	26.8	1254	225	24	167	598	106	29	433	1381	0.232	0.294	0.375	0.67	202
Toronto Blue Jays	2021	91	71	$ 150,140,253.00	5.2	4.1	0	-8	62	26.8	1455	285	13	262	816	81	20	496	1218	0.266	0.33	0.466	0.797	255
Washington Nationals	2021	65	97	$ 144,415,187.00	4.5	5.1	0	-7	60	28.7	1388	272	20	182	686	56	26	573	1303	0.258	0.337	0.417	0.754	224
Arizona Diamondbacks	2022	74	88	$ 85,964,090.00	4.3	4.6	0	-3	57	26.5	1232	262	24	173	658	104	29	531	1341	0.23	0.304	0.385	0.689	206
Atlanta Braves	2022	101	61	$ 183,438,888.00	4.9	3.8	-0.1	1	53	27.5	1394	298	11	243	753	87	31	470	1498	0.253	0.317	0.443	0.761	244
Baltimore Orioles	2022	83	79	$ 44,888,388.00	4.2	4.2	0.2	4	58	27	1281	275	25	171	639	95	31	476	1390	0.236	0.305	0.39	0.695	211
Boston Red Sox	2022	78	84	$ 211,812,131.00	4.5	4.9	0.2	2	54	28.8	1427	352	12	155	704	52	20	478	1373	0.258	0.321	0.409	0.731	226
Chicago Cubs	2022	74	88	$ 151,054,737.00	4.1	4.5	-0.1	1	64	27.9	1293	265	31	159	620	111	37	507	1448	0.238	0.311	0.387	0.698	209
Chicago White Sox	2022	81	81	$ 203,205,326.00	4.2	4.4	-0.1	3	44	29.3	1435	272	9	149	654	58	10	388	1269	0.256	0.31	0.387	0.698	217
Cincinnati Reds	2022	62	100	$ 115,467,321.00	4	5	-0.1	-2	66	29.4	1264	235	18	156	618	58	33	452	1430	0.235	0.304	0.372	0.676	200

The full table is too large (61R x 56C including headers) to be readable, thus is included in the appendix (in addition to what each acronym means).

What is represented by the data :

This data represents the average statistical performance (per team, per year) during the 2021 and 2022 MLB regular season. There are 30 teams in the league, and each team plays 162 games per season, for a total of 2,430 cumulative games per season

W (wins, column C) is the most suitable to be our dependent variable. Weighted statistics (such as win% vs winning/losing teams) is unsuitable as the dependent variable because every team plays the same number of games, and all games are weighted equally for the purposes of advancing into playoffs. We used all other quantitative variables as the independent variables, but eliminated the ones below our threshold of 0.05.

Size of the data (e.g., sample size, attributes):

The dataset used in this analysis is 39kb and has 56 columns, but the first two are used to represent the team and the year (which collectively acts as a primary key). The next two columns represent wins and losses, meaning 52 columns represent the analysis statistics.

All numbers are seasonal team average stats during the 2021 and 2022 MLB regular season. The MLB season consists of 30 teams, 162 games per team, and 2430 total games per season. For the purposes of our analysis, we will not be combining the two years, meaning we have 60 rows of data.

Which software (e.g., SAS enterprise guide version 8.3 or SAS enterprise miner workstation 15.2) OR programming language (e.g., R) will you use?

We will be using SAS Enterprise Guide Version 8.3 and Enterprise Miner 15.2 to run our analysis.

Names of the statistical tests/analyses that you will perform:

We will run multiple linear regression using backwards elimination, a decision tree model and analyze its associated StatExplore page, a cluster analysis, and an analysis of variance.

Rationale of performing each statistical test:

Multiple Linear Regression Analysis

Performing a Multiple Linear Regression analysis with backward elimination can help us understand the relationship between various team statistics and the win rate of Major League Baseball (MLB). This analysis aims to identify which specific team statistic has the highest correlation with the win rate, and the factors that contribute most significantly to a team's success in the league. Backward elimination removes statistically unimportant variables from the regression model, leaving just the most important predictors. By applying this method, the analysis ensures that only those team statistics that have a meaningful impact on the win rate are retained in the final model, reducing the risk of overfitting and enhancing the model's predictive accuracy. An analysis of variance gives us deeper insight into the model.

Cluster Analysis

K-means cluster works by partitioning the data into clusters in such a way that each observation belongs to the cluster with the nearest mean, it allows the formation of clusters that share common characteristics. Performing a cluster analysis will allow us to confirm our findings from the multiple linear regression analysis.

Decision Tree

By creating a decision tree, we will be able to break down metrics and analyze specific constraints that constitute a winning team. This will be helpful when we made our final conclusions and recommendations.

Statistical Test Interpretation

Multiple Linear Regression

We performed a multiple linear regression analysis, with wins as our target variable and other metrics comprising our explanatory variables. At a confidence interval of 95%, we used backward elimination to remove non-significant predictors and reduce overfitting in our dataset. Of the 51 original explanatory variables in the database, following the completion of backward elimination, we are left with 19 variables that are statistically significant. We also ran a correlation matrix on the 19 variables. We determined that highly correlated pairs would have a threshold of 0.7. The correlation matrix showed us that there were no highly correlated pairs. This makes sense, as the backward elimination has already removed variables based on statistical significance.

The regression equation is as follows: Wins = -163.78 + 18.52*(Runs) - 14.72*(Runs Against) + 0.84*(Luck) + 0.28*(BatAge) - 0.10*(H) - 0.10*(2B) - 0.22*(3B) - 0.33*(HR) - 0.03*(BB) - 0.003*(SO Batting) + 628.55*(SLG) - 0.05*(HBP) - 0.04*(SH) + 0.03*(LOB Batting) - 0.03*(Hits Pitching) - 0.03*(HR Pitching) - 0.03*(BB Pitching) + 0.05*(BF) - 0.03*(LOB Pitching)

To display our understanding of the regression equation, we will explain the relationship of runs and runs against. For every unit increase in "Runs," the predicted number of wins is expected to increase by 18.52. On the other hand, for every unit increase in "Runs Against," the predicted number of wins is expected to decrease by 14.72. From the backward elimination, in analyzing the standardized estimate value, our most significant variables (based on the threshold of 0.5) are Runs (0.58), Runs Against (-0.58), Hits (-0.51), Slugging Percentage (1.11), and Home Runs (-0.79). These findings are significant for MLB decision-makers, which is explained in the recommendations and conclusions section.

Decision Tree Model

Furthermore, the game of baseball is split clearly into offensive situations and defensive situations. When a team is playing defense on the field, they are not capable of playing any offense at the same time. So we have decided to split the two sides of the game (offensive and defensive) into Runs and Runs Against.

We used this to create a decision tree, and found some rules that are related to Runs and Runs Against and how they correlate to the amount of wins a team should be expected to have:

Node = 4
----
if Runs_Against < 4.25
AND Runs < 4.55 or MISSING
then
Tree Node Identifier = 4
Number of Observations = 10
Predicted: W = 84.4

Node = 5
---
if Runs_Against < 4.25
AND Runs >= 4.55
then
Tree Node Identifier = 5
Number of Observations = 10
Predicted: W = 97

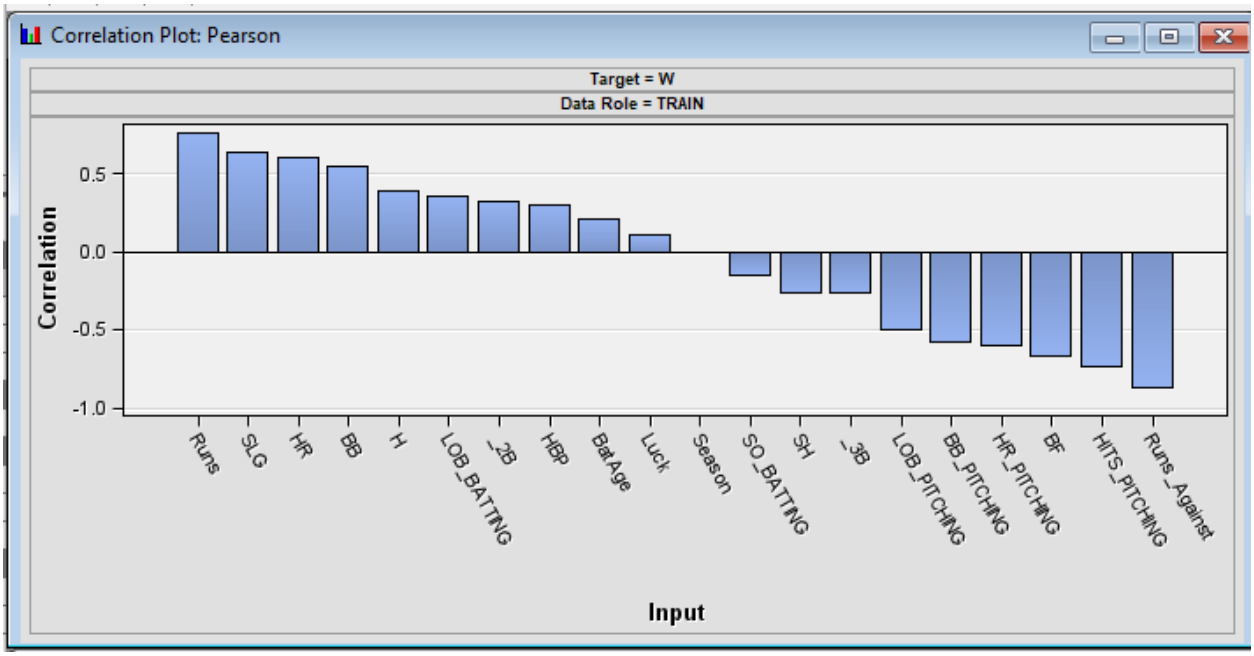
Node = 6
---
if Runs_Against < 4.8 AND Runs_Against >= 4.25 or MISSING
then
Tree Node Identifier = 6
Number of Observations = 12
Predicted: W = 78.083333333

Node = 7
---
if Runs_Against >= 4.8
then
Tree Node Identifier = 7
Number of Observations = 10
Predicted: W = 66.9

So, a team should want to average less than 4.25 Runs Against per game, and have more than 4.55 Runs scored per game. A team with these metrics would expect to get around 97 wins, which would put them near the top of the standings and into the playoffs.

StatExplore

For further analysis, we have broken down all the stats into either batting stats (which would affect Runs) and pitching stats (which would affect Runs Against). Our team used StatExplore to find the worth of each variable to add onto our Multiple Linear Regression results.



For simplicity, we have decided to take the three stats on both the left and right sides that are not Runs or Runs Against. Broken down into offensive and defensive stats, we have determined the most valuable statistics when it comes to winning baseball games is:

Offensive Stats (Runs For): .SLG, Home Runs scored, Base on balls

Defensive Stats (Runs Against): Hits allowed, Home runs allowed, Batters Faced

So, using our decision rule tree, any team that is not scoring at least 4.55 runs per game should concentrate on increasing their .SLG, Home Runs scored, or the amount of Bases on balls. If they are allowing more than 4.25 Runs Against per game, they should acquire players that would increase their Hits, Home Runs, and Batters Faced allowed.

Cluster Analysis

The results from our cluster analysis are consistent with the findings in the multiple linear regression. Through the cluster analysis, we grouped similar data points based on various metrics, and it was evident that clusters with higher win totals tended to exhibit more positive values in the significant metrics identified during regression analysis. This agreement between the two analytical approaches strengthens our understanding that specific performance metrics play a crucial role in a team's success. Teams belonging to clusters with higher win totals (as seen in Cluster 1 with the 100 win team) are likely to excel in areas such as scoring runs, preventing runs against, and demonstrating better offensive and defensive performances. This coherence across different methods of analysis enhances our confidence in identifying key factors influencing team success, offering valuable insights for decision-making in sports management and related fields.

Analysis of Variance

Our regression model has a F-value of 920.9, and a p-value < 0.0001, meaning that the model is very statistically significant. Our relative error (SSE/SST) is 27.92565/12295 = 0.0022713, or 0.22713%, and our adjusted R2 is 0.\9967, which is exceptional. .

Linear Regression Results

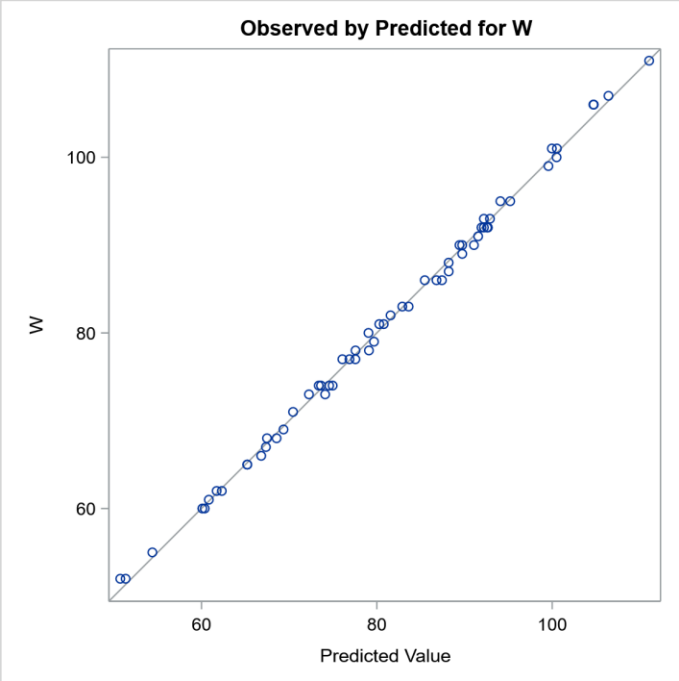
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: W

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr >
Model	19	12295	647.10830	926.90	<.000
Error	40	27.92565	0.69814		
Corrected Total	59	12323			

Root MSE	0.83555	R-Square	0.9977
Dependent Mean	80.98333	Adj R-Sq	0.9967
Coeff Var	1.03175		



This can be visually observed in the best-fit line, which has no real outliers to speak of. The primary reason for this is because our dataset consists only of averages, meaning outliers simply do not get an opportunity to skew some metrics one way or another.

To test this in real-life scenarios, we created a basic spreadsheet

Intercept	Runs	Runs Against	Luck	Bat Age	H	2B	3B	HR	BB	SO BATTING	SLG	HBP	SH	LOB BATTING	HITS PITCHING	HR PITCHING	BB PITCHING	BF	LOB PITCHING
-163.78035	-18.5173	14.7239	0.84464	0.27957	-0.10146	-0.10376	-0.20223	-0.30232	-0.00234	-0.00038	62.85483	-0.00648	-0.00687	0.00329	-0.00431	-0.00411	-0.00383	0.04747	-0.00329
Oakland Athletics 2021 W=86	4.6	4.2	-1	30.1	1284	271	19	199	545	13.49	0.406	98	17	10.84	13.62	19.1	43.9	60.60	10.89
86.79842398	-78.6004	-61.84038	-0.84464	8.41505	-13.02	-28.11	-4.127	-66.08	-18.54	-5.126	-5.190	-4.709	-0.660	31.46	-42.27	-5.560	-14.85	28.523	-31.77

	48				7464	896	37	989	09	2	825	88	79	2	648	01	137		702
Oakland Athletics 2022 W=60	3.5	4.8	1	28.3	1147	249	15	137	433	1389	0.346	59	22	969	1394	195	503	6121	1087
60.36393118	-98.9695 55	-70.67472	0.84464	7.911831	-116.37462	-25.8362 4	-3.245	-45.4907	-14.7306	-5.278 2	21 7.4778 952	-2.8354	-0.8551 4	28.1300 7	-43.2697 6	-5.601 76	-17.6445 9	28 7.9305	-31.71866
106.4145696	-71.1935	-54.47843	3.37856	7.54839	-12 2.6 6514	-24.17 608	-2.389 53	-66.08 989	-18.20 07	-5.669 6	24 1.991 2996	-3.460 32	-0.349 83	30.09 1912	-42.02 4	-5.740 67	-16.75 5	28 7.287 3	-31.07 67

2021-2022 MLB Standings t-test

Conclusion and Recommendation

Following the results of our statistical tests, we have identified several metrics that are significant toward winning. A main takeaway from our multiple linear regression is that our most significant variables (based on the threshold of 0.5 from the standardized estimate value) are Runs (0.58), Runs Against (-0.58), Hits (-0.51), Slugging Percentage (1.11), and Home Runs (-0.79), with Slugging Percentage and Home Runs as the top two variables. Our cluster analysis confirmed our results, as clusters portraying a higher number of wins also had positive results in the metrics we deemed as most valuable.

A main takeaway from our classification tree is that any team not scoring at least 4.55 runs per game should concentrate on increasing their .SLG, Home Runs scored, or the amount of Bases

on balls. If they are allowing more than 4.25 Runs Against per game, they should acquire players that would increase their Hits, Home Runs, and Batters Faced allowed. Teams with Runs Against under 4.25, and Runs over or equal to 4.55 have a predicted 97 wins, which is typically enough to earn a playoff berth.

Considering the insights derived from our statistical outputs, we would like to make a recommendation for players to target for competitive MLB teams in the 2023 offseason, in preparation for the 2024 season. As shown through the classification tree, teams with runs equal to or over 4.55 and with runs against under 4.25 have a predicted wins count of 97. Any team that is not scoring at least 4.55 runs per game should concentrate on increasing their .SLG, Home Runs scored, or the amount of Bases on balls. If they are allowing more than 4.25 Runs Against per game, they should acquire players that would increase their Hits, Home Runs, and Batters Faced allowed. For teams struggling to score runs, 2024 free agents like Shohei Ohtani or J.D. Martinez would be favourable signings, as they rank first and sixth respectively in .SLG for the current season (Baseball America, 2023). For teams allowing more than 4.25 Runs Against per game, 2024 free agents like Kyle Gibson and Lucas Giolito, the two free agents with the highest BF total, will certainly add value to a competitive team. With these metrics in mind, MLB executives can develop an optimal player acquisition strategy.

References

Baseball America (n.d.). Baseballamerica.com

https://www.baseballamerica.com

Baseball Reference (n.d.). Baseball-Reference.com.

https://www.baseball-reference.com/

Appendix

Linear Regression results

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: W

Number of Observations Read	60
Number of Observations Used	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	12295	647.10830	926.90	<.0001
Error	40	27.92565	0.69814		
Corrected Total	59	12323			

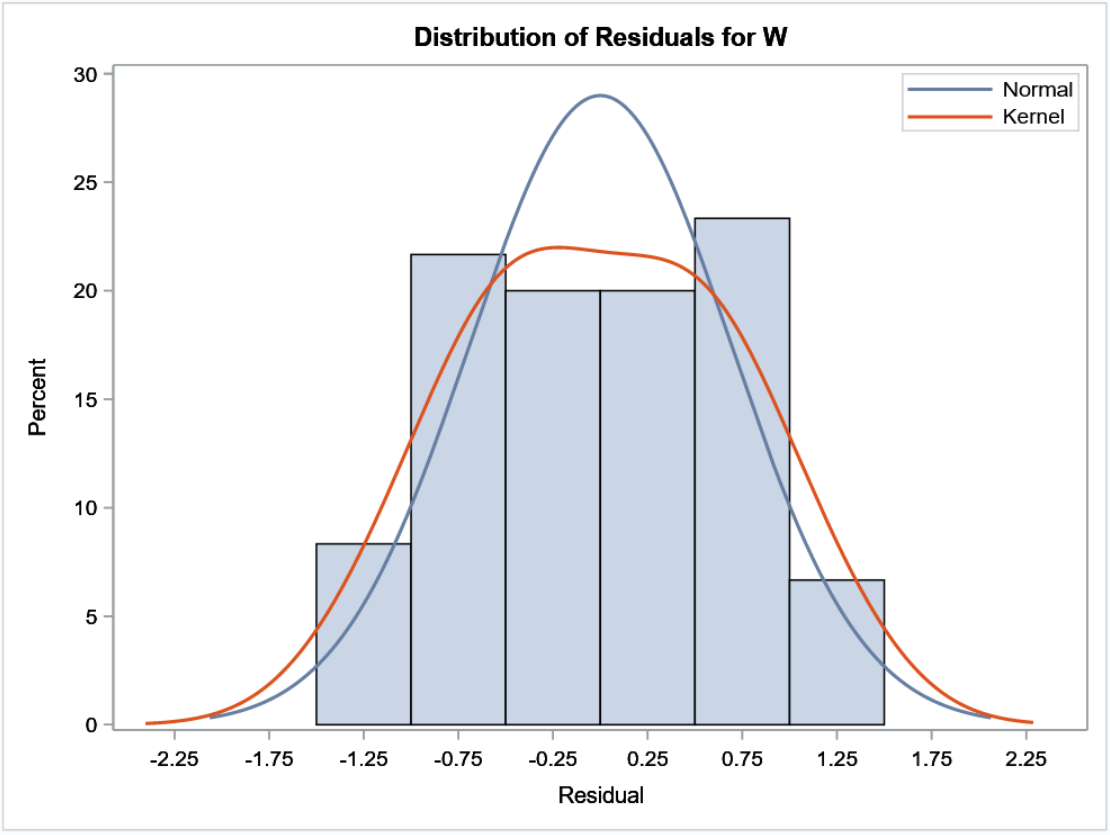
Root MSE	0.83555	R-Square	0.9977
Dependent Mean	80.98333	Adj R-Sq	0.9967
Coeff Var	1.03175		

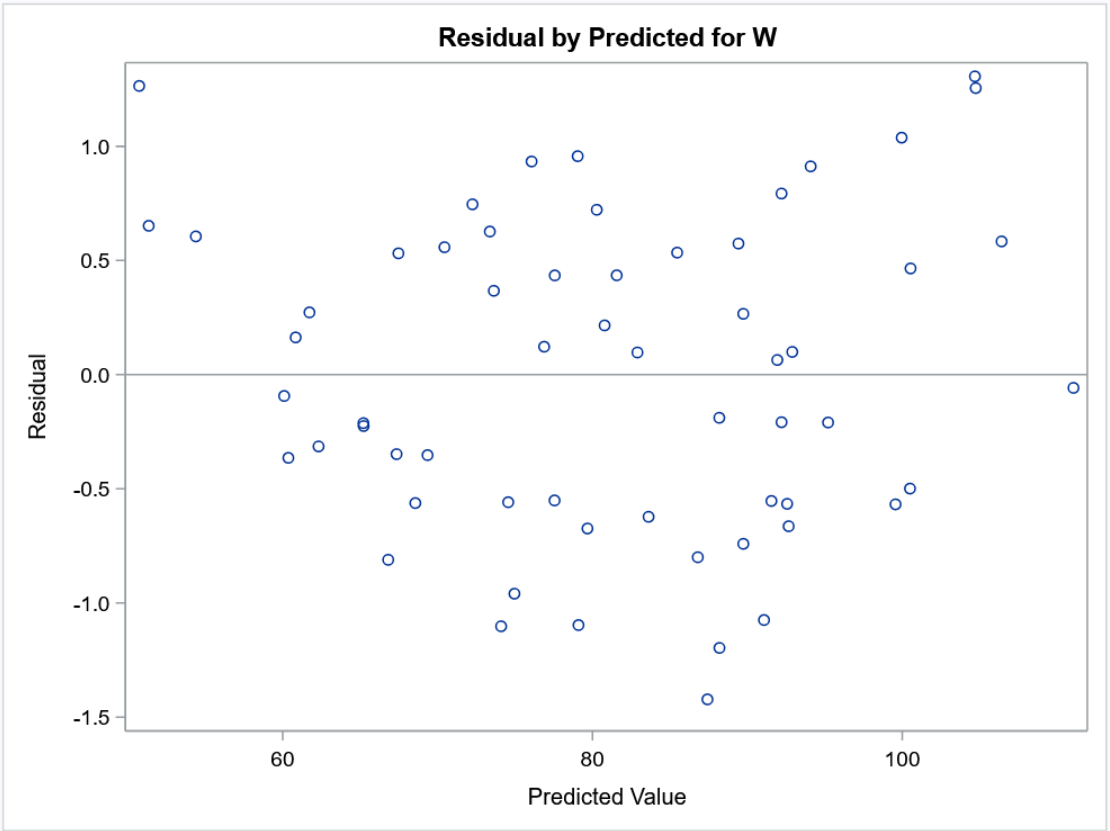
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-163.78035	40.00666	-4.09	0.0002	0
Runs	1	18.51737	1.01460	18.25	<.0001	0.58488
Runs Against	1	-14.72390	0.92595	-15.90	<.0001	-0.58329
Luck	1	0.84464	0.03748	22.54	<.0001	0.23969
BatAge	1	0.27957	0.13426	2.08	0.0438	0.02051
H	1	-0.10146	0.01787	-5.68	<.0001	-0.51253
2B	1	-0.10376	0.01975	-5.25	<.0001	-0.20902
3B	1	-0.21723	0.04011	-5.42	<.0001	-0.11261
HR	1	-0.33211	0.05809	-5.72	<.0001	-0.79746
BB	1	-0.03402	0.00846	-4.02	0.0002	-0.14238
SO BATTING	1	-0.00380	0.00179	-2.12	0.0402	-0.02589
SLG	1	628.54883	107.83866	5.83	<.0001	1.10987
HBP	1	-0.04806	0.01128	-4.26	0.0001	-0.05296
SH	1	-0.03887	0.01666	-2.33	0.0247	-0.03209
LOB BATTING	1	0.02903	0.00833	3.48	0.0012	0.10101
HITS PITCHING	1	-0.03104	0.00563	-5.51	<.0001	-0.20370
HR PITCHING	1	-0.02911	0.00962	-3.02	0.0043	-0.05622
BB PITCHING	1	-0.03383	0.00697	-4.85	<.0001	-0.13523
BF	1	0.04705	0.00757	6.21	<.0001	0.35598
LOB PITCHING	1	-0.02918	0.00787	-3.71	0.0006	-0.09460

Linear Regression Results

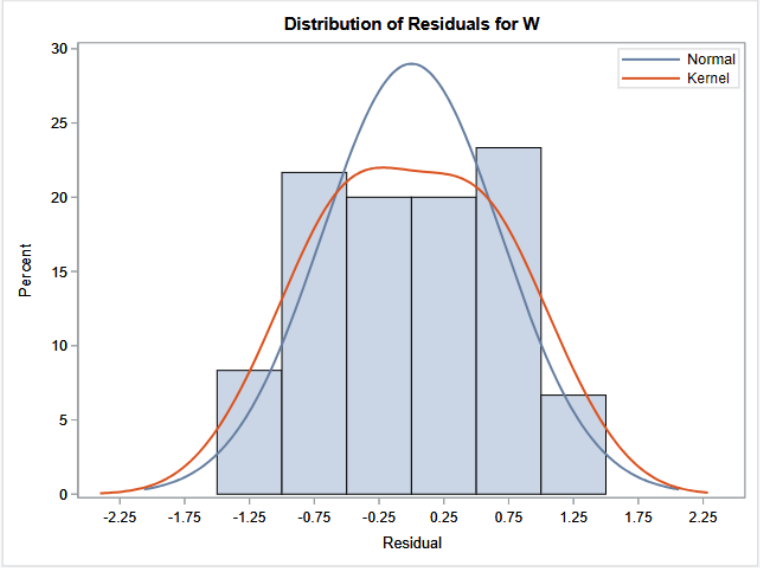
The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: W

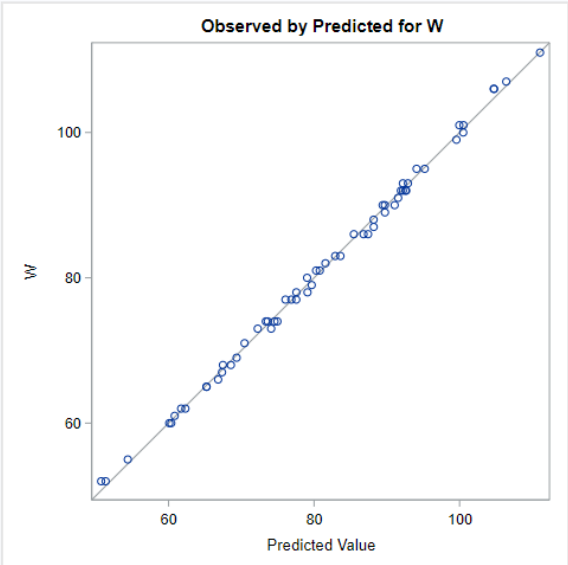


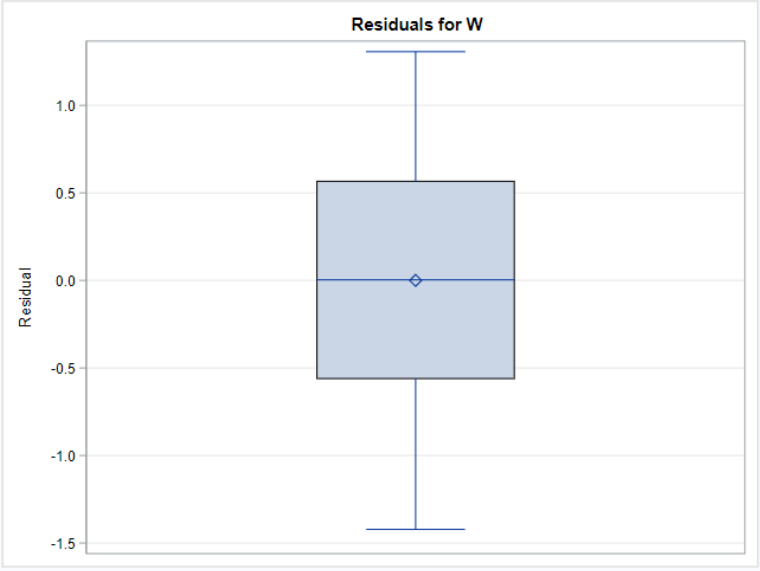


Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: W







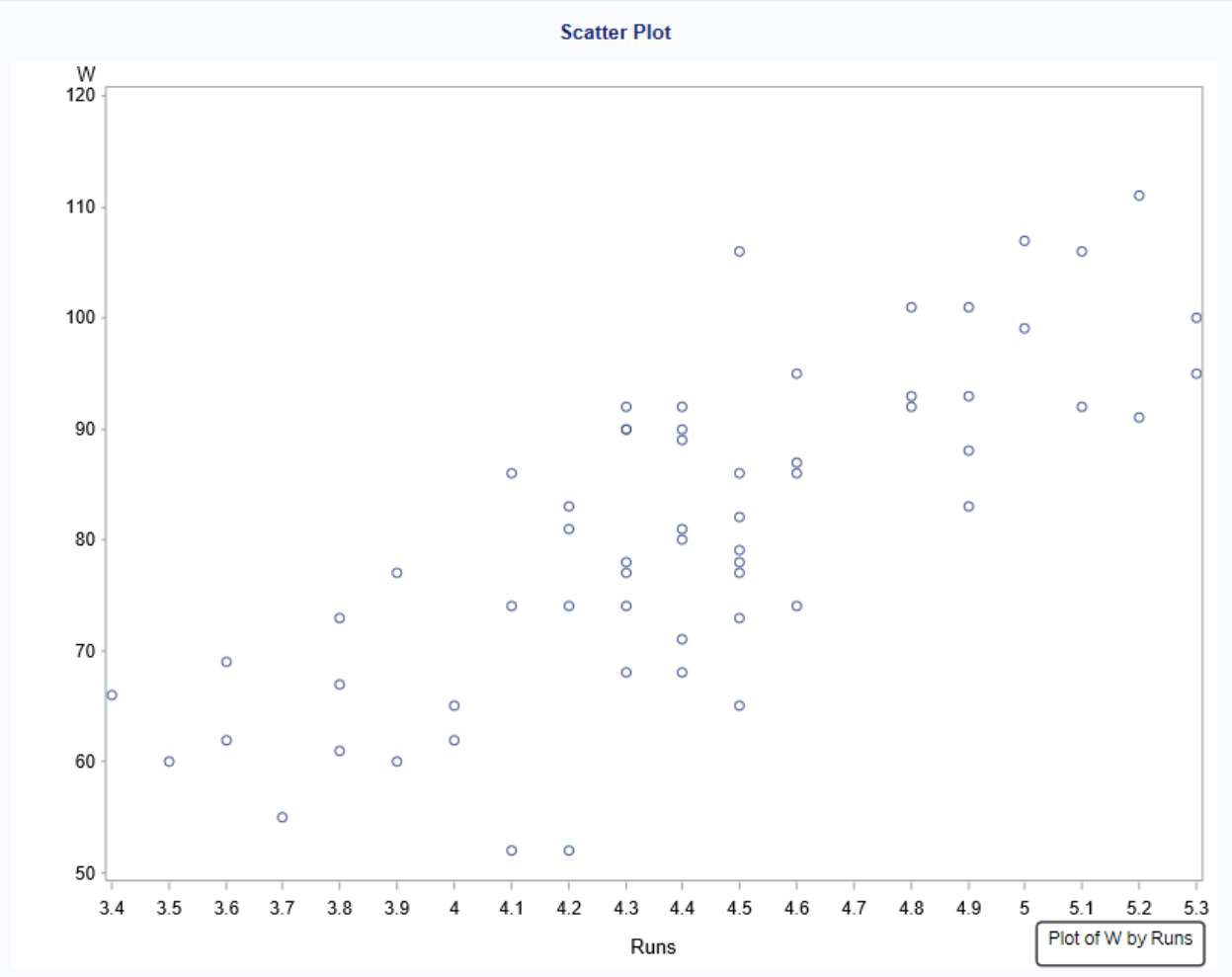
Cluster Analysis Results (2 parts)

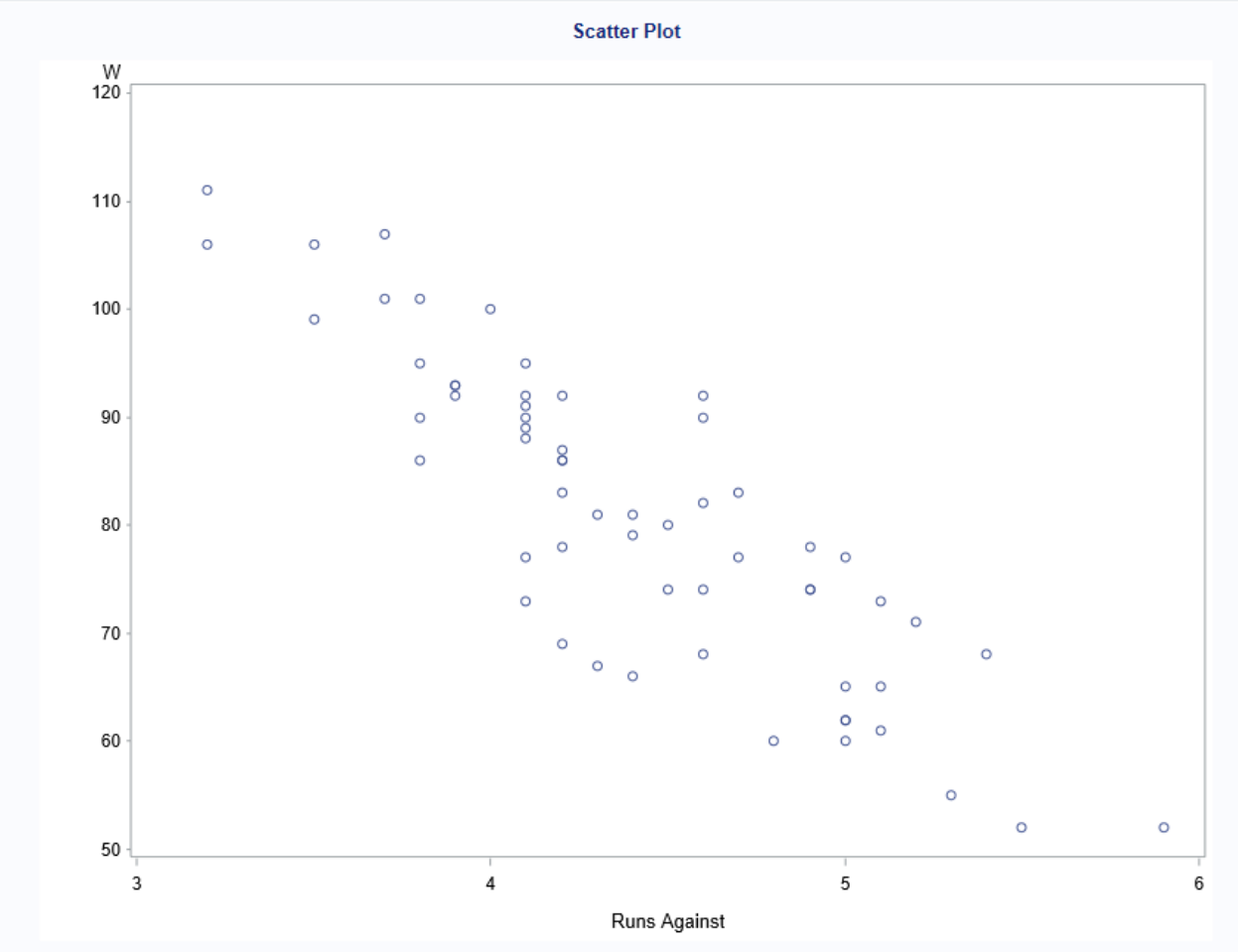
Cluster Means

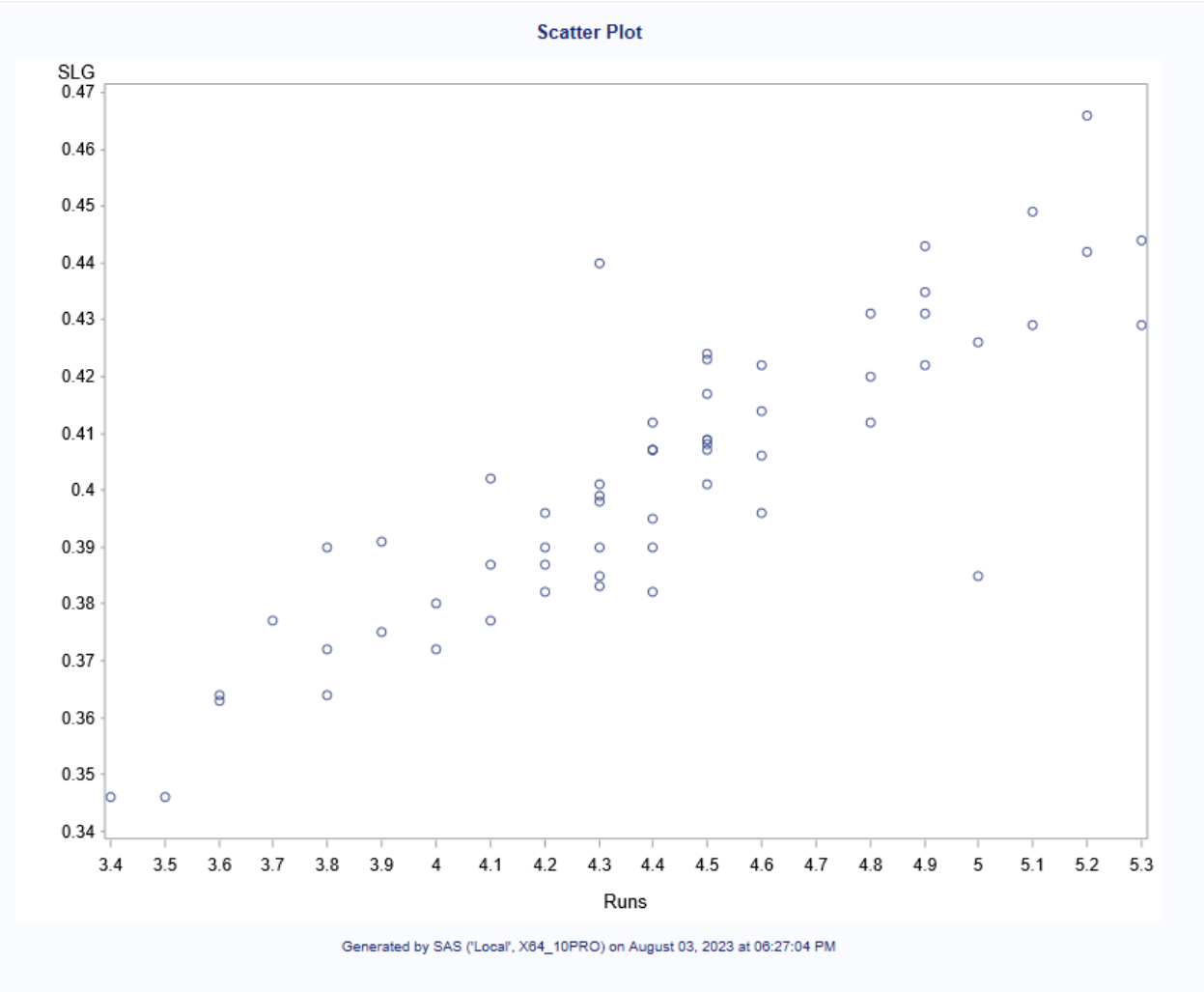
Cluster	Season	W	L	Payroll	Runs	Runs Against	Strength of Schedule	Luck	#Bat	BatAge	H	2B	3B	HR	RBI	SB	CS	BB	SO BATTING	BA	OBP	SLG	OPS	TB	GDP	HBP	SH	SF	IBB BATTING	LOB BATTING	#P
1	2021.8	100.8	61.2	260427207.6	4.9	3.6	-0.1	-2.6	56.6	29.4	1374.0	264.8	23.8	215.8	765.8	86.4	23.6	565.6	1350.6	0.3	0.3	0.4	0.8	2333.8	108.0	78.8	15.0	45.4	26.8	1130.8	33.8
2	2021.4	75.7	86.2	124875152.3	4.4	4.6	-0.0	-1.2	56.4	28.4	1315.6	261.9	22.6	181.8	672.0	75.9	25.8	503.6	1386.1	0.2	0.3	0.4	0.7	2168.2	114.6	71.5	21.2	39.2	17.4	1083.3	33.8
3	2021.5	74.4	87.6	63963586.3	4.1	4.6	0.0	1.5	57.8	27.6	1274.0	254.1	25.9	171.7	641.5	90.6	27.0	489.4	1415.9	0.2	0.3	0.4	0.7	2095.0	103.7	63.8	20.3	38.5	18.9	1059.8	34.1
4	2021.5	87.9	74.1	188440131.6	4.6	4.2	0.0	0.8	55.6	28.4	1343.4	272.3	17.2	194.3	700.7	70.1	24.0	521.5	1361.4	0.2	0.3	0.4	0.7	2232.9	116.1	67.4	16.9	38.8	21.4	1099.9	32.9

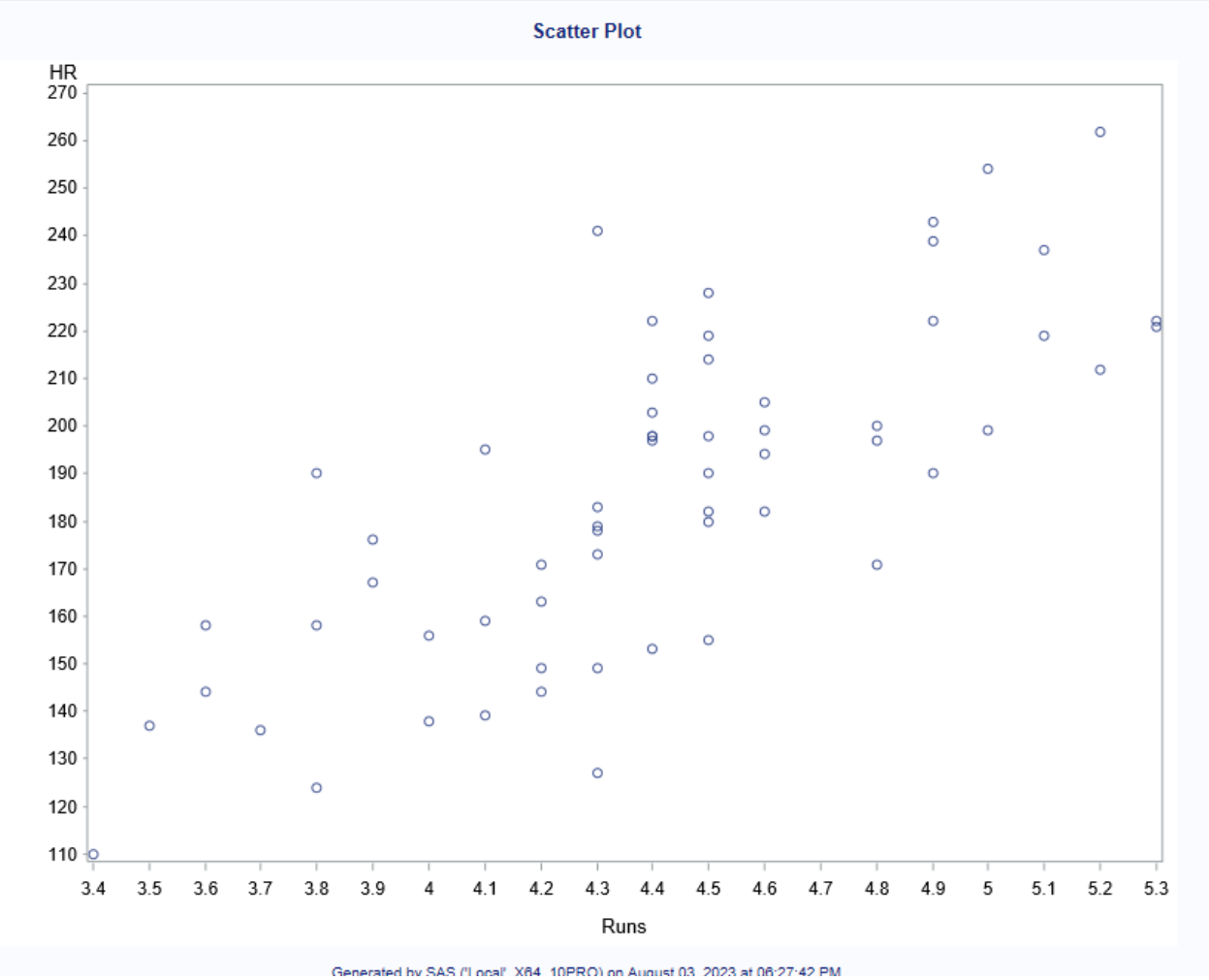
PAge	ERA	CG	tSho	cSho	SV	HITS PITCHING	HR PITCHING	BB PITCHING	IBB PITCHING	SO PITCHING	HBP PITCHING	BK	WP	BF	ERA+	FIP	WHIP	H9	HR9	BB9	SO9	SO/W	LOB PITCHING
29.4	3.3	1.2	16.4	0.6	45.8	1200.4	157.8	445.6	19.0	1502.2	68.4	4.4	42.4	5941.4	122.2	3.5	1.1	7.5	1.0	2.8	9.4	3.4	1043.6
28.8	4.3	1.3	9.8	0.8	38.1	1345.4	195.0	528.5	19.5	1352.6	69.6	4.6	60.8	6091.8	98.8	4.3	1.3	8.5	1.3	3.3	8.6	2.6	1098.0
27.8	4.2	1.2	9.4	0.5	38.5	1346.8	188.4	509.6	20.7	1322.3	69.8	4.6	58.9	6084.8	98.8	4.2	1.3	8.5	1.2	3.2	8.3	2.6	1089.0
29.2	3.9	1.9	11.4	1.0	43.8	1293.5	177.9	502.9	19.0	1438.2	68.6	4.7	56.6	6043.6	105.5	3.9	1.3	8.1	1.1	3.2	9.0	2.9	1080.9

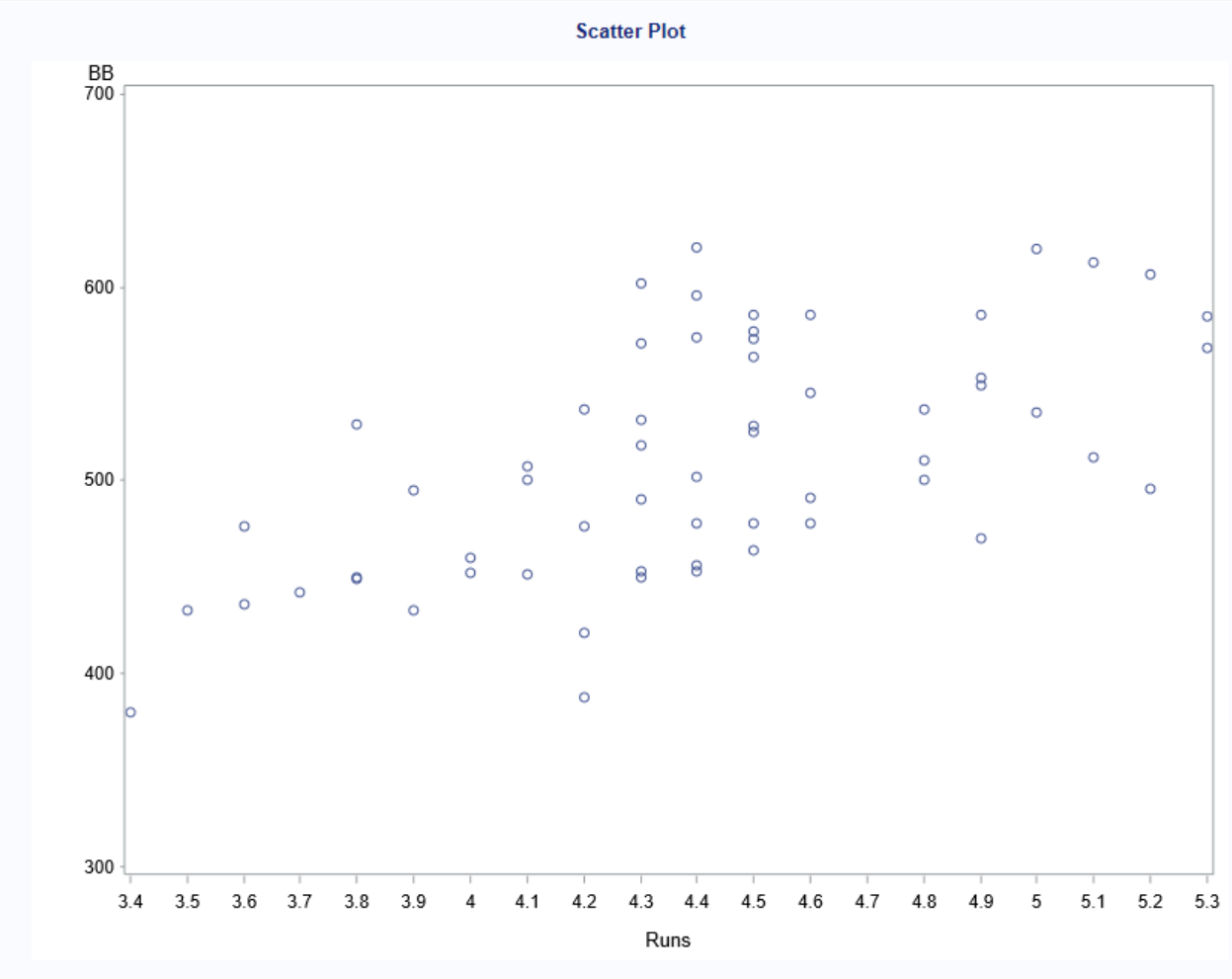
Various Scatterplots

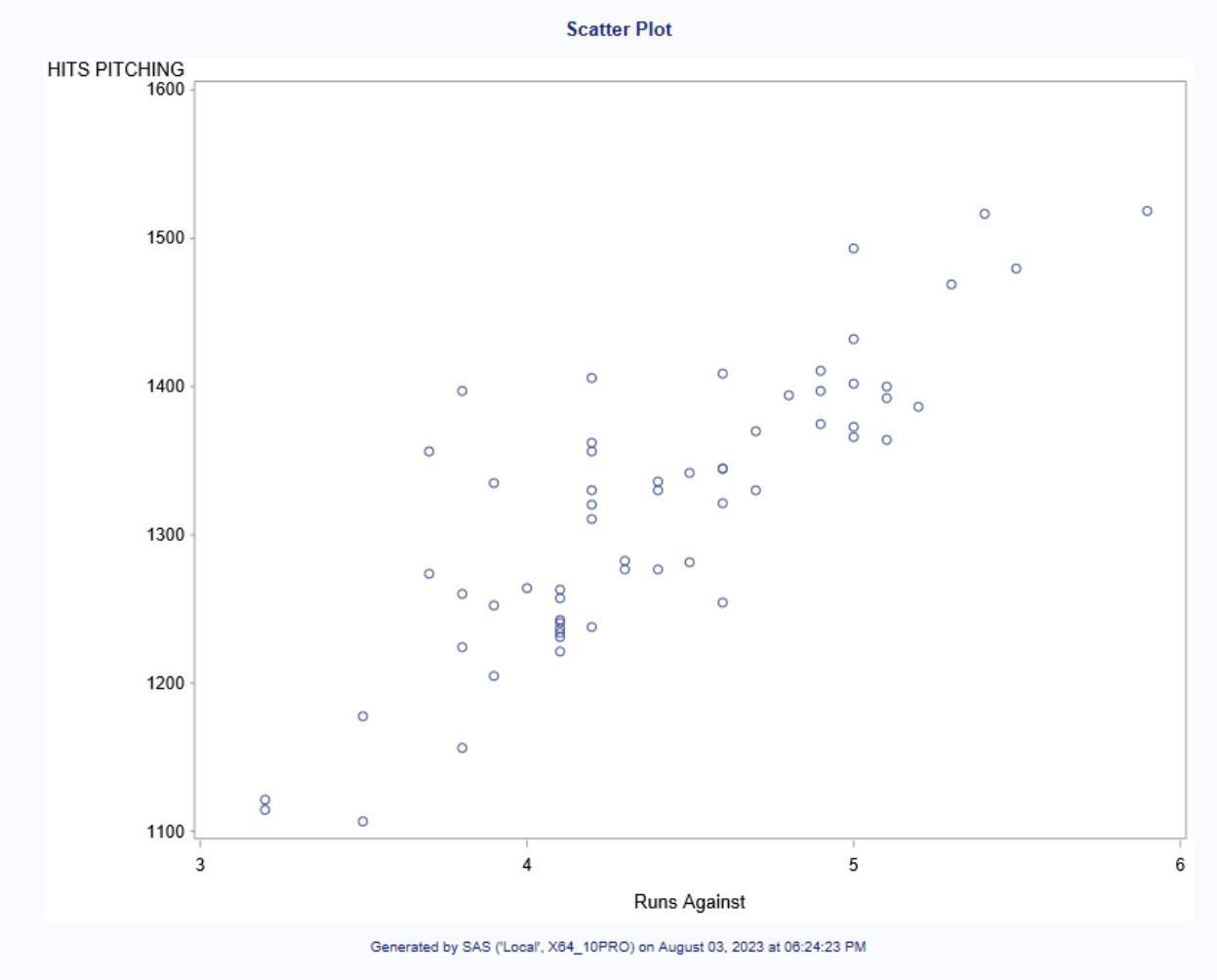


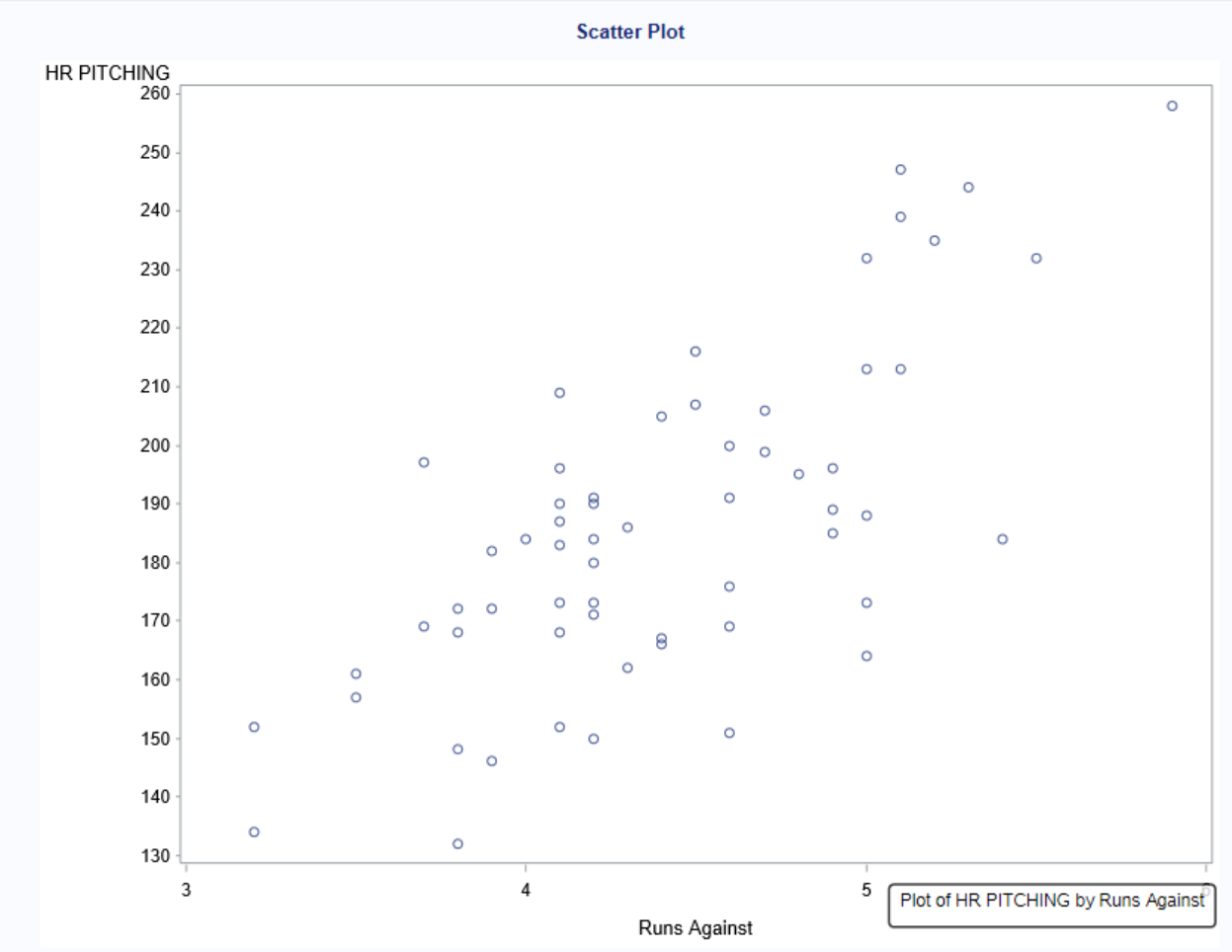

Scatter Plot
W
120
110
100
90
80
70
60
50
3
4
5
6
Runs Against

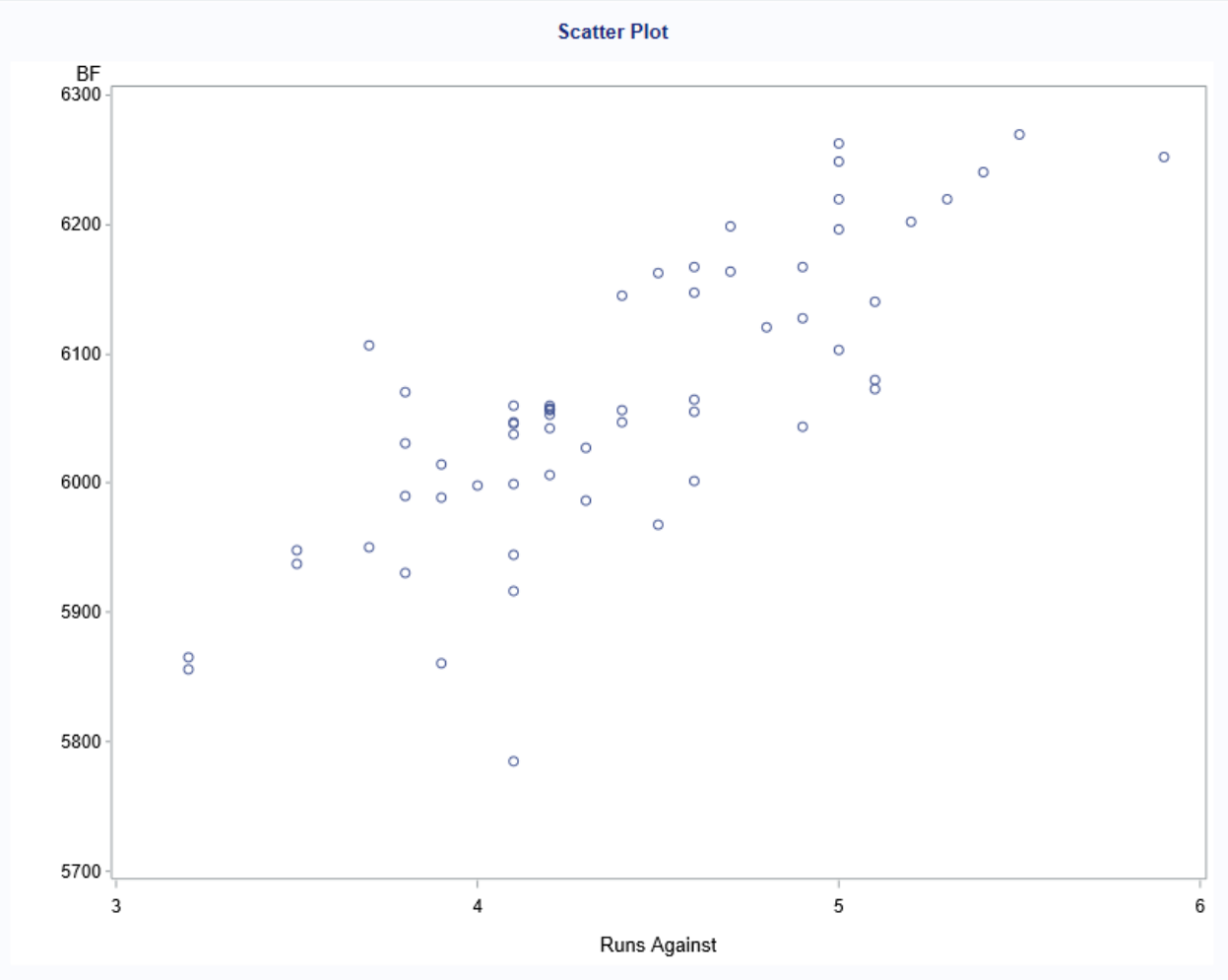

Scatter Plot
SLG
0.47
0.46
0.45
0.44
0.43
0.42
0.41
0.4
0.39
0.38
0.37
0.36
0.35
0.34
3.4
3.5
3.6
3.7
3.8
3.9
4
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
5
5.1
5.2
5.3
Runs
Generated by SAS ('Local', X64_10PRO) on August 03, 2023 at 06:27:04 PM

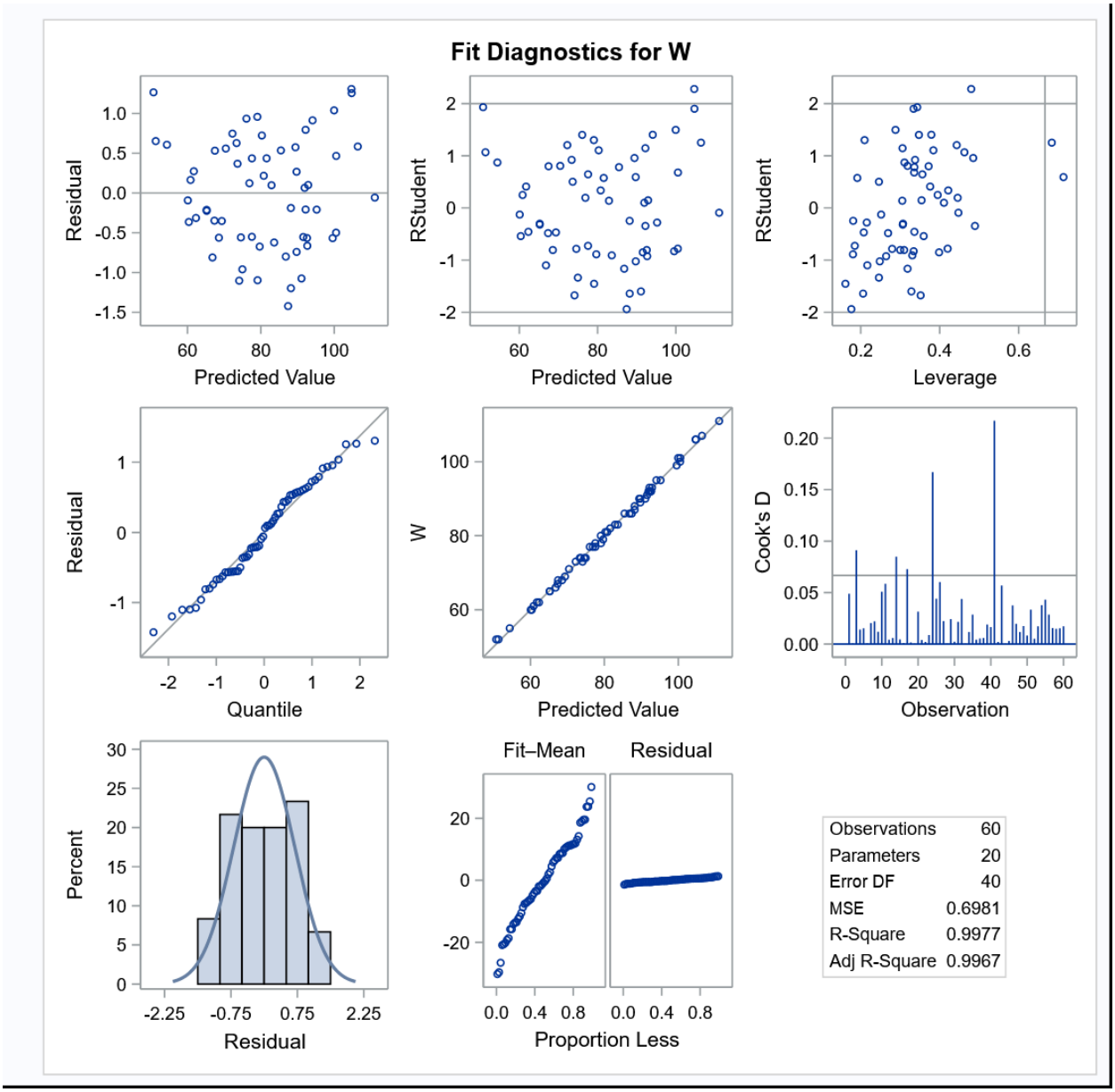

Scatter Plot
HR
270
260
250
240
230
220
210
200
190
180
170
160
150
140
130
120
110
3.4
3.5
3.6
3.7
3.8
3.9
4
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
5
5.1
5.2
5.3
Runs
Generated by SAS (Local, X64_10PRO) on August 03, 2023 at 06:27:42 PM

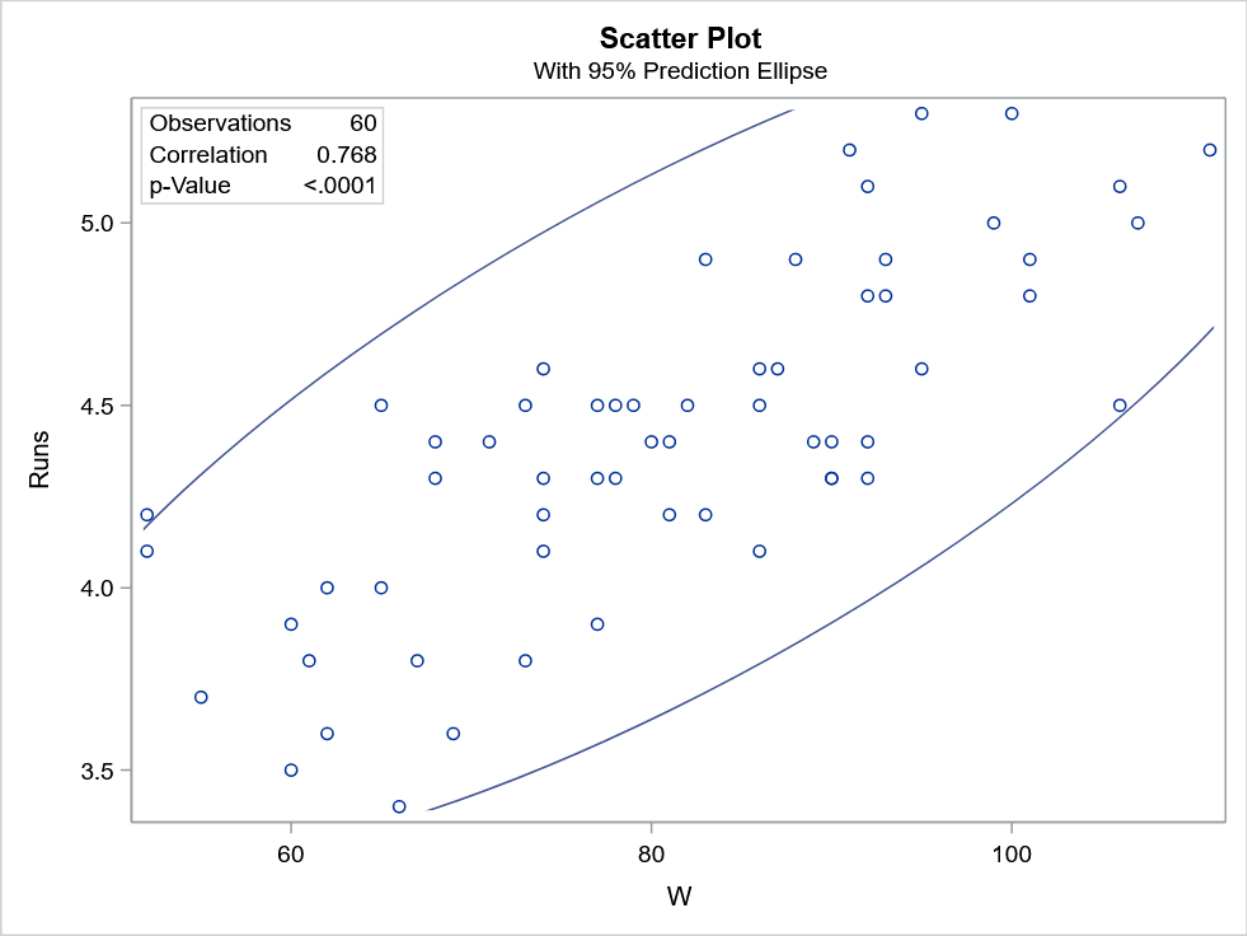

Scatter Plot
BB
700
600
500
400
300
3.4
3.5
3.6
3.7
3.8
3.9
4
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
5
5.1
5.2
5.3
Runs

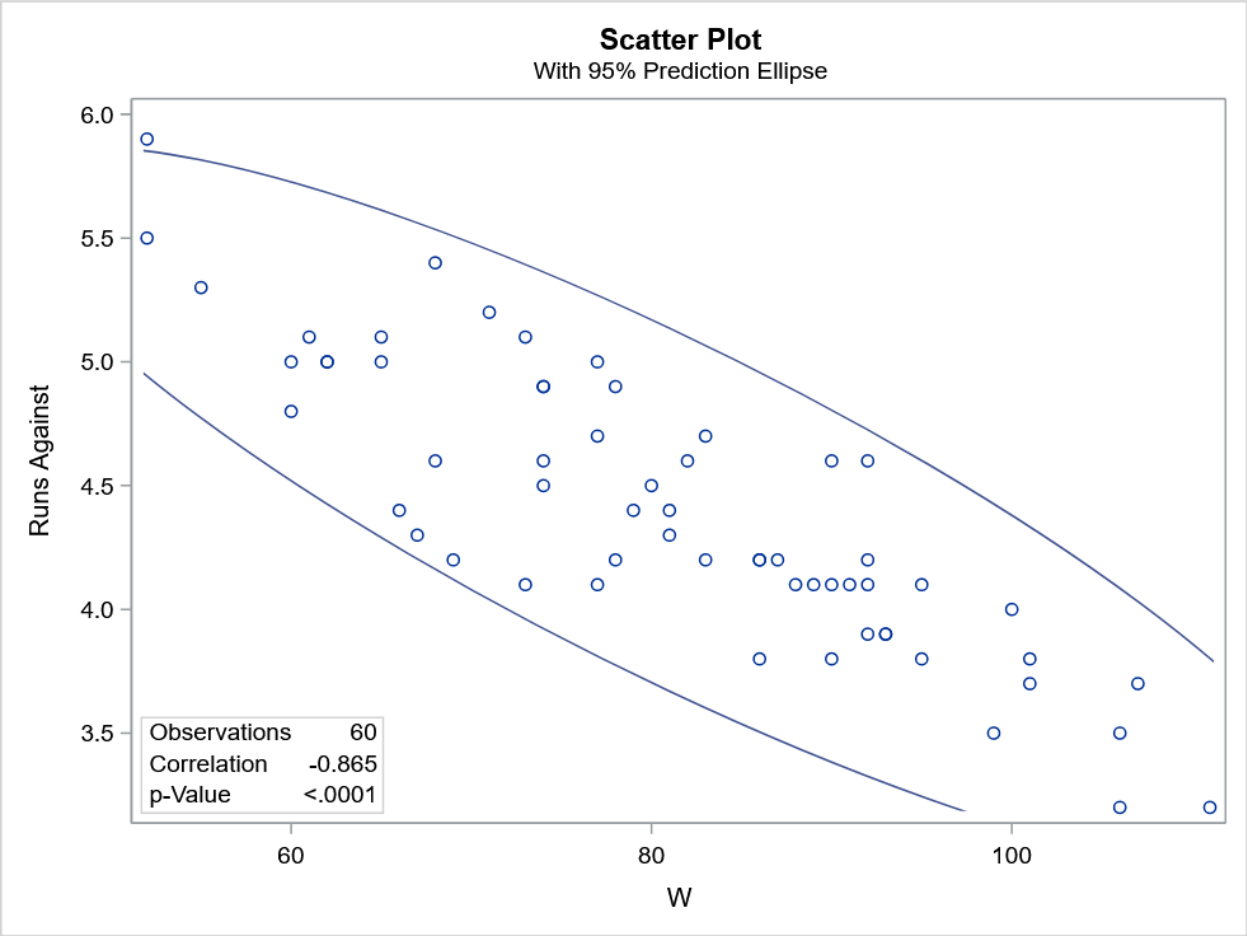

Scatter Plot
HITS PITCHING
1600
1500
1400
1300
1200
1100
3
4
5
6
Runs Against
Generated by SAS ('Local', X64_10PRO) on August 03, 2023 at 06:24:23 PM

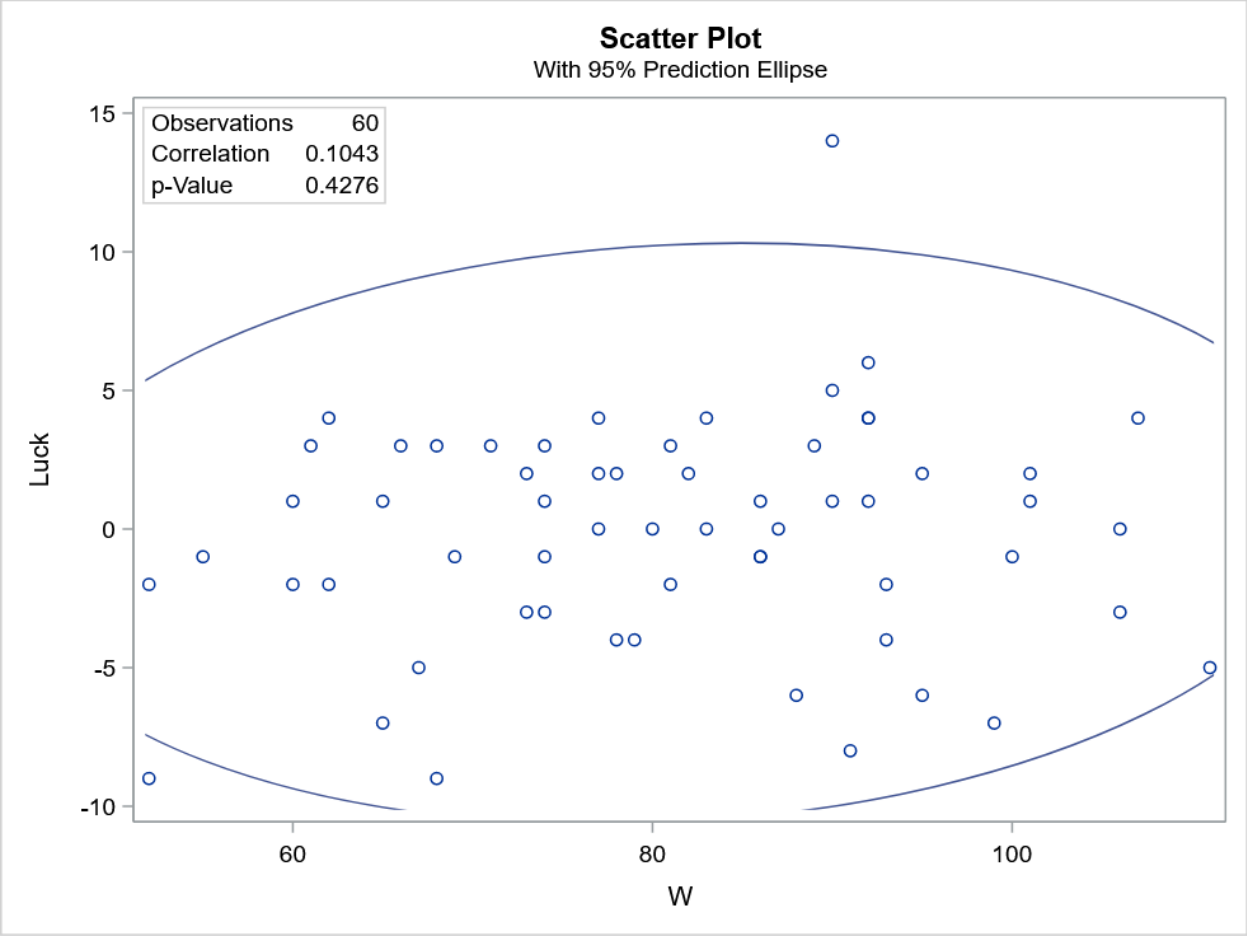

Scatter Plot
HR PITCHING
260
250
240
230
220
210
200
190
180
170
160
150
140
130
3
4
5
Runs Against
Plot of HR PITCHING by Runs Against

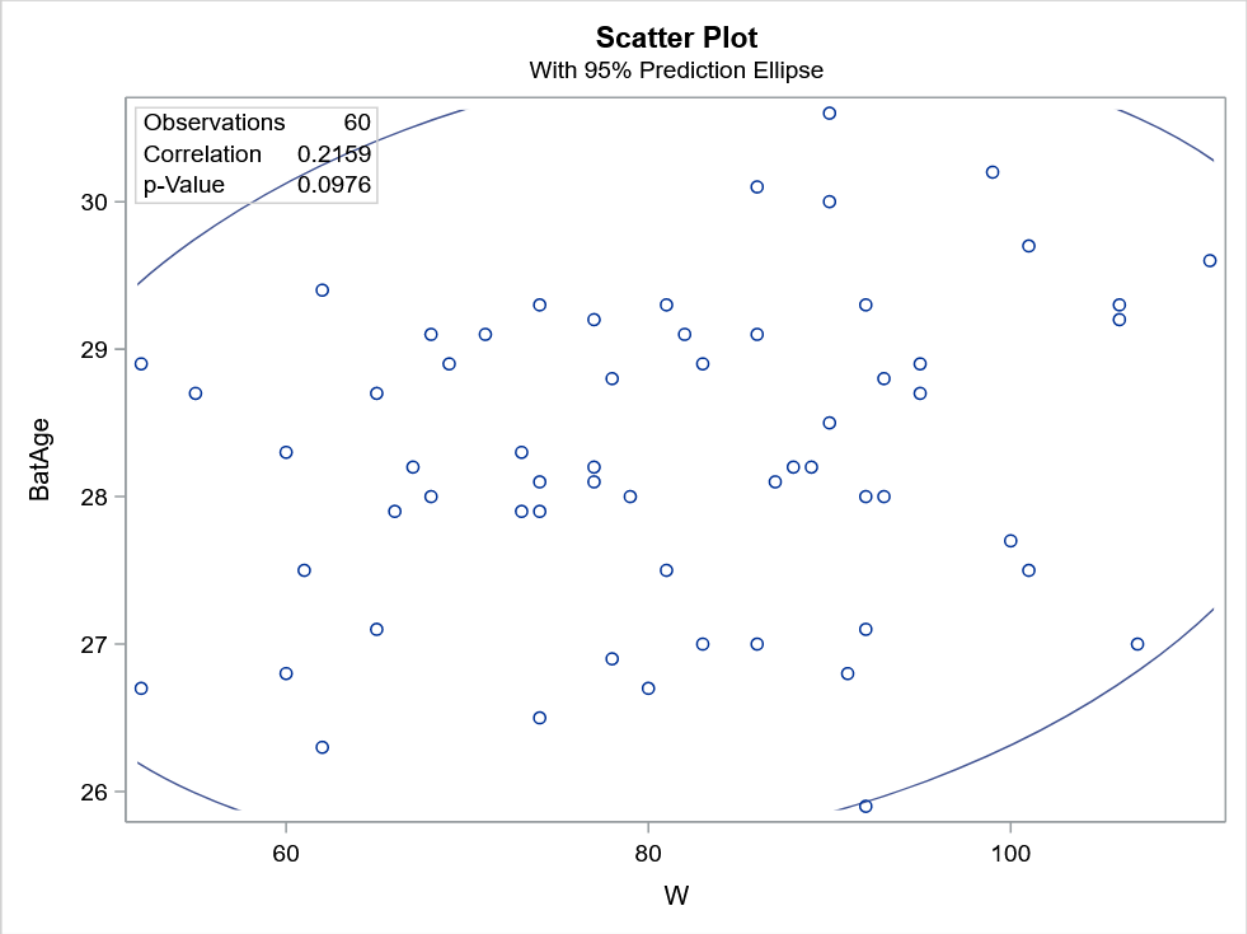

Scatter Plot
BF
6300
6200
6100
6000
5900
5800
5700
3
4
5
6
Runs Against

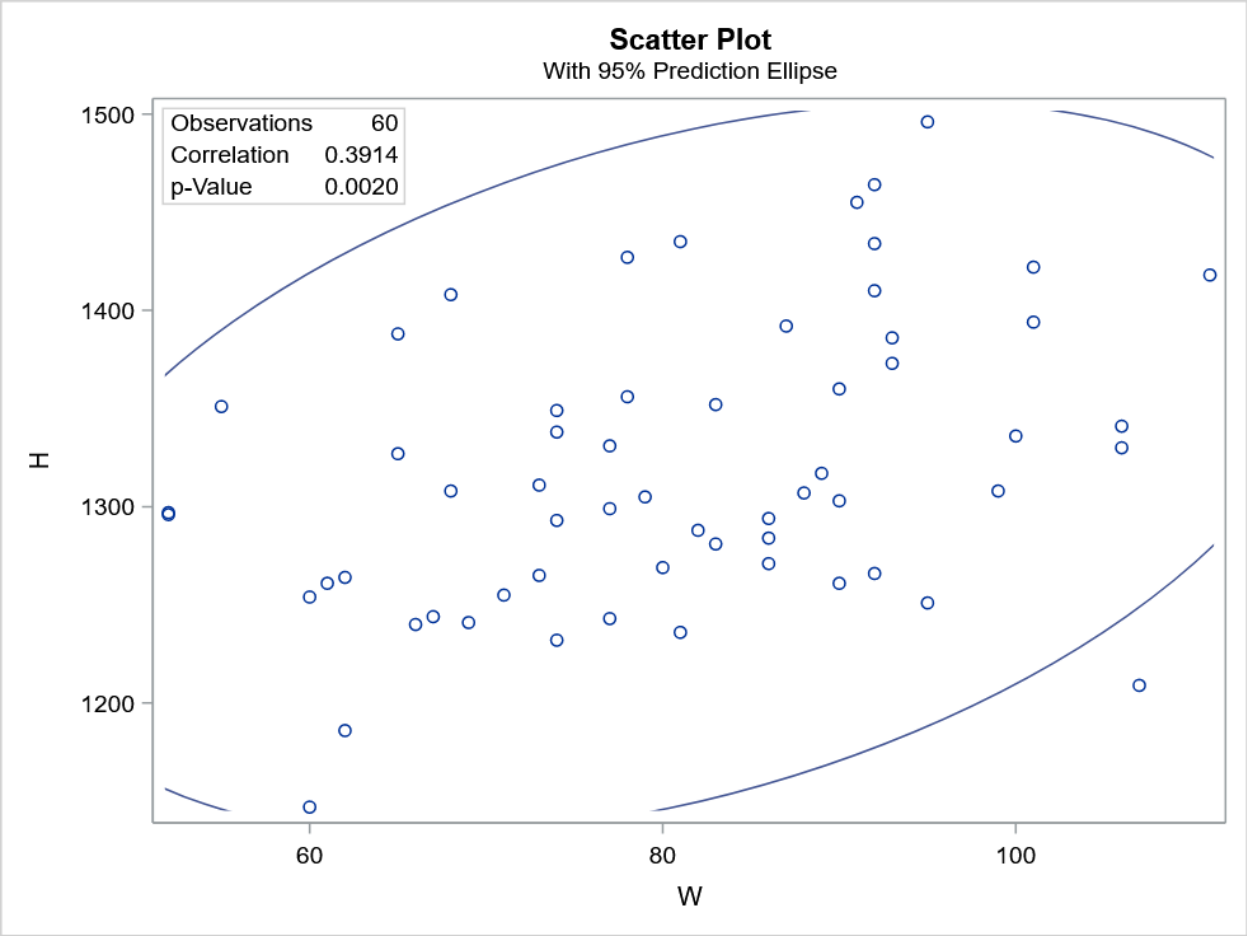

Fit Diagnostics for W
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
W
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 60
Parameters 20
Error DF 40
MSE 0.6981
R-Square 0.9977
Adj R-Square 0.9967


Scatter Plot
With 95% Prediction Ellipse
Observations 60
Correlation 0.768
p-Value <.0001
Runs
5.0
4.5
4.0
3.5
60
80
100
W


Scatter Plot
With 95% Prediction Ellipse
Runs Against
6.0
5.5
5.0
4.5
4.0
3.5
60
80
100
W
Observations 60
Correlation -0.865
p-Value <.0001


Scatter Plot
With 95% Prediction Ellipse
Observations 60
Correlation 0.1043
p-Value 0.4276
Luck
W
15
10
5
0
-5
-10
60
80
100


Scatter Plot
With 95% Prediction Ellipse
Observations 60
Correlation 0.2159
p-Value 0.0976
BatAge
30
29
28
27
26
60
80
100
W



Glossary of Metrics

Stat	Description	Stat	Description	Stat	Description
#Bat	Players used in Games	#P	Pitchers used in Games	BatAge	Batters' average age
H	Hits/Hits Allowed	ERA	9 * ER / IP	2B	Doubles Hit/Allowed
3B	Triples Hit/Allowed	HR	Home Runs Hit/Allowed	RBI	Runs Batted In
SB	Stolen Bases	CS	Caught Stealing	BB	Bases on Balls/Walks
SO BAT	Batting Strikeouts	BA	Hits/At Bats	OBP	(H + BB + HBP)/(At Bats + BB + HBP + SF)
SLG	Total Bases/At Bats	OPS	On-Base + Slugging Percentages	OPS+	OPS+
TB	Total Bases	GDP	Double Plays Grounded Into	HBP	Times Hit by a Pitch
SH	Sacrifice Hits (Sacrifice Bunts)	SF	Sacrifice Flies	IBB BAT	Intentional Bases on Balls
LOB BAT	Runners Left On Base (Batting)	#P	Pitchers used in Games	PAge	Pitchers' average age
HITS PITCH	Hits/Hits Allowed	ERA	9 * ER / IP	CG	Complete Game
tSho	Shutouts by a team	cSho	Shutouts	SV	Saves
IP	Innings Pitched	HR PITCH	Home Runs Hit/Allowed	BB PITCH	Bases on Balls/Walks
SO PITCH	Pitching Strikeouts	IBB PITCH	Intentional Bases on Balls	HBP PITCH	Times Hit by a Pitch
BK	Balks	WP	Wild Pitches	BF	Batters Faced
ERA+	ERA+	FIP	Fielding Independent Pitching	WHIP	(BB + H)/IP
H9	9 x H / IP	HR9	9 x HR / IP	BB9	9 x BB / IP
SO9	9 x SO / IP	SO/W	SO/W or SO/BB	LOB PITCH	Runners Left On Base (Pitching)

Column1	Season	W	L	Payroll	Runs	Runs Against	Strength of Schedule	Luck	#Bat
Arizona Diamondbacks	2021	52	110	$ 91,632,929.00	4.2	5.5	0.2	-9	64
Atlanta Braves	2021	88	73	$ 152,750,691.00	4.9	4.1	-0.1	-6	56
Baltimore Orioles	2021	52	110	$ 42,421,870.00	4.1	5.9	0.3	-2	62
Boston Red Sox	2021	92	70	$ 187,100,784.00	5.1	4.6	0.1	4	56
Chicago Cubs	2021	71	91	$ 144,037,170.00	4.4	5.2	-0.1	3	69
Chicago White Sox	2021	93	69	$ 140,926,169.00	4.9	3.9	-0.2	-4	47
Cincinnati Reds	2021	83	79	$ 126,587,447.00	4.9	4.7	-0.2	0	55
Cleveland Indians	2021	80	82	$ 50,670,534.00	4.4	4.5	-0.1	0	48
Colorado Rockies	2021	74	87	$ 116,408,966.00	4.6	4.9	0.1	-1	45
Detroit Tigers	2021	77	85	$ 86,348,945.00	4.3	4.7	-0.1	2	49
Houston Astros	2021	95	67	$ 194,222,042.00	5.3	4.1	-0.1	-6	52
Kansas City Royals	2021	74	88	$ 91,595,545.00	4.2	4.9	0	3	48
Los Angeles Angels	2021	77	85	$ 183,849,560.00	4.5	5	0.2	4	64
Los Angeles Dodgers	2021	106	56	$ 266,020,809.00	5.1	3.5	-0.1	-3	61
Miami Marlins	2021	67	95	$ 58,157,900.00	3.8	4.3	0	-5	61
Milwaukee Brewers	2021	95	67	$ 99,377,415.00	4.6	3.8	-0.3	2	61
Minnesota Twins	2021	73	89	$ 120,084,606.00	4.5	5.1	0	2	57
New York Mets	2021	77	85	$ 201,189,189.00	3.9	4.1	0	0	64
New York Yankees	2021	92	70	$ 205,669,863.00	4.4	4.1	0.1	6	59
Oakland Athletics	2021	86	76	$ 90,400,598.00	4.6	4.2	0.1	-1	50
Philadelphia Phillies	2021	82	80	$ 197,263,223.00	4.5	4.6	0	2	55
Pittsburgh Pirates	2021	61	101	$ 54,356,609.00	3.8	5.1	0	3	64
San Diego Padres	2021	79	83	$ 179,764,272.00	4.5	4.4	0.1	-4	54
San Francisco Giants	2021	107	55	$ 171,890,308.00	5	3.7	-0.1	4	63
Seattle Mariners	2021	90	72	$ 83,822,113.00	4.3	4.6	0.1	14	54
St. Louis Cardinals	2021	90	72	$ 151,469,994.00	4.4	4.1	-0.2	5	51

The full dataset is attached separately for formatting reasons.