Pierce Renio

STP 429

Schneider

23 October 2022

Lab 2

Abstract/Executive Summary

The Major League Baseball organization is an extremely competitive league where millions of dollars are involved. Using the baseball data from Baseball Reference, we are able to look into what could be the main predictors towards the number of wins the San Francisco Giants obtain in a season. Most of the data was skewed left as less games were played near the start of the organization in the late 19th century and early 20th century. Many outliers existed near the skew because of this. Looking at the correlation between the independent variables and the dependent variable, the top 4 variables were PA, RBI, H, and OPS. A regression analysis was performed to identify which variables of these 4 were the best at making a model for the number of wins. Of those 4 variables, the best 2 variables for creating a model to predict the number of wins were PA and OPS. An interactive term model and a second order term model were compared to the best model where they were found to be not as efficient at predicting wins relative to the original. The best model found was:

Predicted wins = -33.82248 + 0.00939(PA) + 84.50119(OPS)

This report details each step of the exploratory data analysis, the statistical analysis, and the regression analysis. The model built will help baseball teams obtain more wins in their season.

Data

The data consisted of yearly statistics for the San Francisco Giants baseball team for the years 2022 to 1883, with the exception of the year 2020 as COVID-19 ended the season prematurely. There are 139 observations in the data and 27 variables were used in the analysis (Figure 1). Some variables are in the data set but were not considered when looking at relationships with our dependent variable: wins. The variable L (Losses) was not considered because they are directly related to wins; a team either wins or loses so it is not a good measure to use. The variable Finish was not used because where a team ranks in their division is based on the amount of wins they have; teams cannot predict wins based on rankings that cannot be determined until after wins are determined. Additionally, the variable G (Games played or pitched) was not used as the number has stayed essentially the same for the last 20 plus years; all teams will play the same amount of games so it is not a good predictor for wins. The remainder of the variables were considered in the data analysis. R/G (Runs scored per game) was used because the more runs a team scores then the more likely that they will end up winning. PA (Plate appearances) and AB (At bats) were considered because the more opportunities a team has at bat, the more likely they are to have hits and runs. R (Runs scored/allowed) was considered as runs are the metric to determine which team wins a game. H (Hits/Hits Allowed), 2B (Doubles Hit/Allowed), 3B (Triples Hit/Allowed), HR (Home Runs HIt/Allowed), and BA (Hits/At bat) were all considered because all these types of hits are the main way to score runs in a game. RBI (Runs batted in) was considered as this is a method of scoring runs. SB (Stolen bases) and CS (Caught stealing) were considered because taking bases is a way for a team to get closer to scoring a run and how often they are caught will affect a teams ability to score. BB (Bases on balls/walks) were considered as this is also a way for a team to get closer to scoring runs. SO

(Strikeouts) was considered as the number of failed attempts at bat via strikes can determine if a team scores many runs. OBP (On base percentage), SLG (Total bases/at bat), and OPS (On-base + Slugging Percentage) were all considered as they are all ways of determining how well a team is doing regarding being on base/getting closer to score. E (Errors committed) was included as mistakes on the field will determine if the opposing team scores runs. Most of these high values are from earlier years. DP (Double plays turned) and Fld% (Fielding %) was considered as good defensive plays that will allow a team to keep their opposition from scoring runs. The last variable, BatAge (Batters average age), was considered as it is a way to determine if newer or older players win more games.

Methodology

For this data analysis, we will use exploratory data analysis first to analyze each of the variable distributions and look for outliers. The PROC UNIVARIATE statement and HISTOGRAM statement were used for the exploratory data analysis as they gave summary statistics and histograms for each variable to analyze. Finding which variables correlated the most with wins was done by using the PROC CORR statement, and this allowed the dependent variable to be correlated with all of the independent variables so that the 4 main independent variables could be chosen for model creation. The statement PROC SGPLOT provided scatterplots for the dependent variable with the 4 main independent variables chosen. Linear regression was able to be performed by using the statement PROC REG as it created the models with the dependent variable and combinations of each of the 4 independent variables chosen. To determine if an interaction term was needed for the model, a PROC GLM and PROC PLM model was used to create a contour plot and a fit plot.

Results

The SAS software was used to complete this project. To be able to find a solution to the research question "Which model of 2 variables is best at predicting wins for the San Francisco Giants baseball team?", we must first look at each individual independent variable. Many of the distributions are skewed left. These low values are due to earlier years not having as many games played in a season thus not as high of values compared to later years. The distribution of R/G appears normal however very slightly skewed right (Figure 2). The distributions for PA and AB are skewed left with some outliers in the 3000's/4000's range (Figure 3 and Figure 4). This is due to the number of games played in that season being low. The distribution for R appears normal (Figure 5). The distributions for H and 2B appear normal but H is slightly skewed left (Figure 6 and Figure 7). The distribution for 3B is skewed right, but no apparent outliers (Figure 8). The distribution for HR appears normal with a spike on the left part of the histogram (Figure 9). The distribution for BA also is skewed right with a few outliers near 0.3 and above (Figure 22). The distribution for RBI appears normal but skewed left (Figure 10). This mostly due to seasons with fewer games played. The distributions for SB and CS are skewed right, which differ from most of the data as many of the higher values are due to earlier seasons (Figure 11 and Figure 12). The distribution for BB appears normal (Figure 13). The distribution for SO appears not normal, with 2 maxes being near the median and Q3 (Figure 14). The distributions for these 3 variables all appear normal with some outliers on the left part of the histograms (Figures 15, 16, 17). The distribution for E is skewed right (Figure 18). The distribution for DP appears normal but skewed left (Figure 19). The distribution for Fld% is greatly skewed left, with many outliers near 0.9 and below (Figure 20). The distribution for BatAge appears to be normal but is slightly skewed right (Figure 21). Next, we had to find which variables are best correlated with the dependent variable

W (Wins). There were several variables that moderately correlated with W including AB (0.42562), R (0.44849), 2B (0.35943), BB (0.38252), OBP (0.42280), and SLG (0.39455) but the top 4 most correlated variables were PA (0.48449), H (0.51680), RBI (0.48124), and OPS (0.45408). These 4 independent variables will be used to create the models. The next step was to continue to look at the relationship each independent variable has with the number of wins through scatterplots. Each of these scatterplots all show a moderately strong positive linear relationship between the independent variables and wins (Figures 23, 24, 25, 26). After confirming the relationships visually, model building through multiple linear regression was done.There exists 6 possible combinations of 2 pairs for the 4 variables thus 6 models can be created. All the models had the F-Value p-test less than 0.001 which means they all were lower than the critical value of 0.05. For the model with PA and H, the F- value was 27.95, the R^2 adjusted value was 0.2809, and the parameter coefficients had p-values of less than the critical value of 0.05 (Figure 27). For the model with PA and RBI, the F- value was 28.77, the R^2 adjusted value was 0.2870, and the parameter coefficients had p-values of less than the critical value of 0.05 (Figure 28). For the model with PA and OPS, the F- value was 31.44, the R^2 adjusted value was 0.3061, and the parameter coefficients had p-values of less than the critical value of 0.05 (Figure 29). For the model with H and RBI, the F- value was 26.20, the R^2 adjusted value was 0.2675, and only the parameter coefficient for H had a p-value of less than the critical value of 0.05. The parameter coefficient p-value for RBI had a p-value of 0.1510 (Figure 30). For the model with H and OPS, the F- value was 26.55, the R^2 adjusted value was 0.2702, and only the parameter coefficient for H had a p-value of less than the critical value of 0.05. The parameter coefficient p-value for OPS was 0.1095 (Figure 31). Lastly for the model with RBI and OPS, the F- value was 20.60, the R^2 adjusted value was 0.2212, and only the parameter

coefficient for RBI had a p-value of less than the critical value of 0.05. The parameter coefficient p-value for OPS was 0.6881 (Figure 32). The best model was the one that had PA and OPS as the independent variables because it had the greatest F-value, the greatest R^2 adjusted value, and it is the only model that has its parameters coefficient p-values less than 0.001. The best model among the 6 was:

Predicted wins = -33.82248 + 0.00939(PA) + 84.50119(OPS)

With this model, we also need to look if an interaction term or second order term is needed for the model. A contour plot was created in SAS to identify the slicing needed for the fit plot. The slicing was 0.60, 0.70, and 0.80 (Figure 33). The fit plot displays the lines all intersecting near the origin indicating that there might be some interaction between OPS and PA (Figure 34). A model was created with an interaction term. For the model with PA, OPS, and PA*OPS, the F- value was 21.07, the R^2 adjusted value was 0.3038, and the parameter coefficient p-value for the interactive term is 0.4592 which is greater than the critical value of 0.05 (Figure 35). The model for the interaction term is not better than the original model with the 2 independent variables so it will not be used. Next, we will look at if a second order term is needed for the model. For the model with PA, OPS, and PA^2, the F- value was 20.97, the R^2 adjusted value was 0.3027, and the parameter coefficient p-value for the second order term is 0.5561 which is greater than the critical value of 0.05 (Figure 36). For the model with PA, OPS, and OPS^2, the F- value was 20.93, the R^2 adjusted value was 0.3023, and the parameter coefficient p-value for the second order term is 0.6130 which is greater than the critical value of 0.05 (Figure 37). Both of the models for the second order term were not as good as the original

model as the F-values and the R^2 adjusted values are smaller than the values for the original model. The final model is:

Predicted wins = -33.82248 + 0.00939(PA) + 84.50119(OPS)

Final Conclusions and Next Steps

It was unknown prior to the analysis which variables were correlated the best with the number of wins and it was unknown which 2 variables were best at making a model to predict the number of wins. There were many variables that had moderate correlation to the number of wins but only 4 variables (PA, H, RBI, OPS) had the strongest correlation with the dependent variable. With those 4 variables, the 2 that were most effective at predicting the number of wins are PA and OPS, and those variables did not need an interaction term nor a second order term in the model to make the model better. If this analysis was completed again, another team would be used to determine if these 2 variables would again be the best at predicting wins for the team. Being able to look at different teams in this analysis would be useful at determining the overall best predicting variables for all the teams in the MLB.

Appendix

The CONTENTS Procedure

Data Set Name	WORK.BASEBALL	Observations	139
Member Type	DATA	Variables	27
Engine	V9	Indexes	0
Created	10/21/2022 16:31:19	Observation Length	216
Last Modified	10/21/2022 16:31:19	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Figure 1

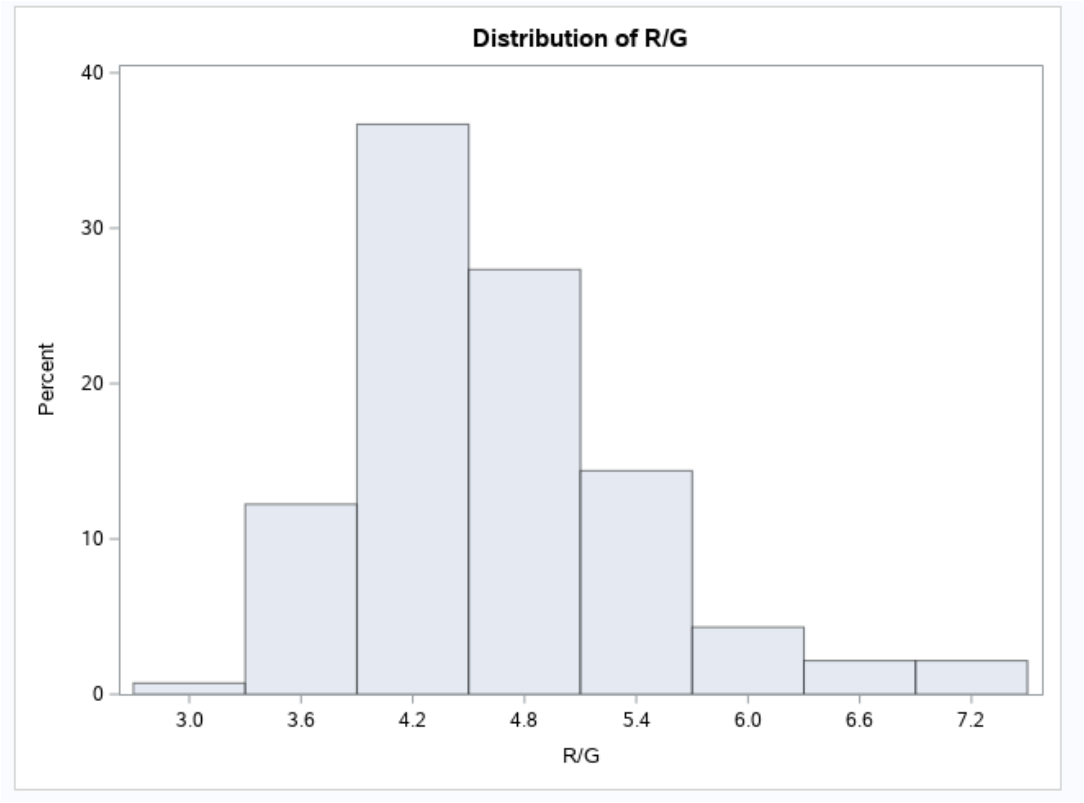


Figure 2

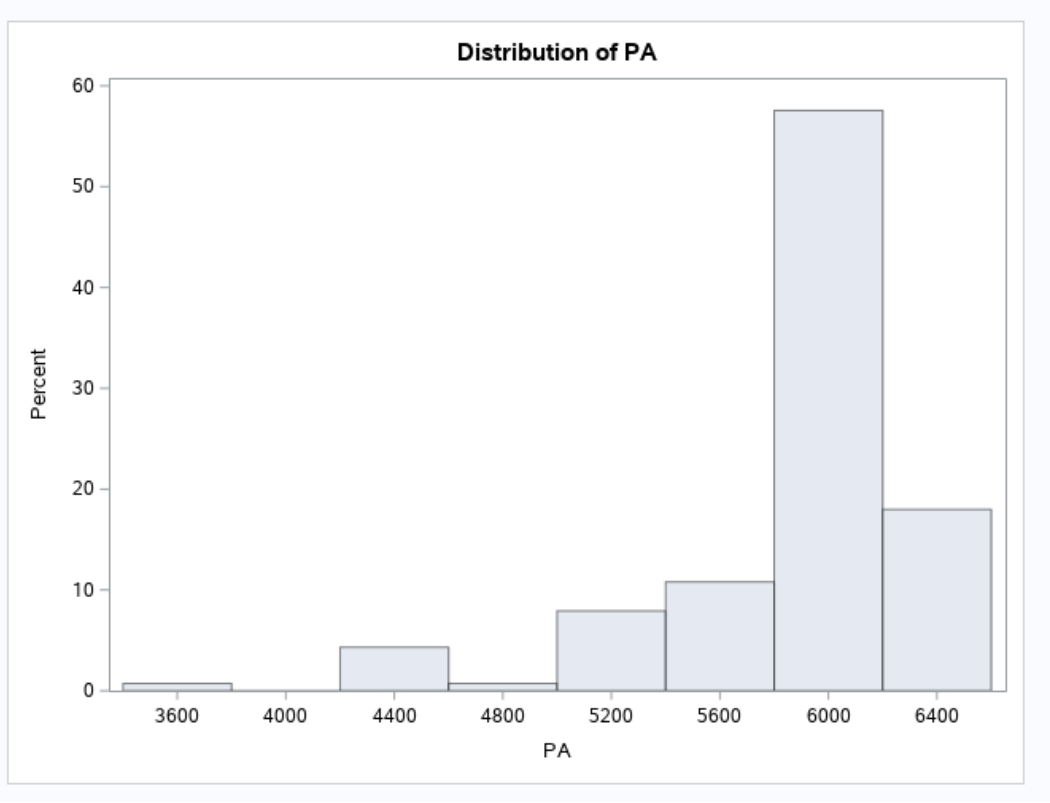

Distribution of PA
Percent
60
50
40
30
20
10
0
3600
4000
4400
4800
5200
5600
6000
6400
PA

Figure 3

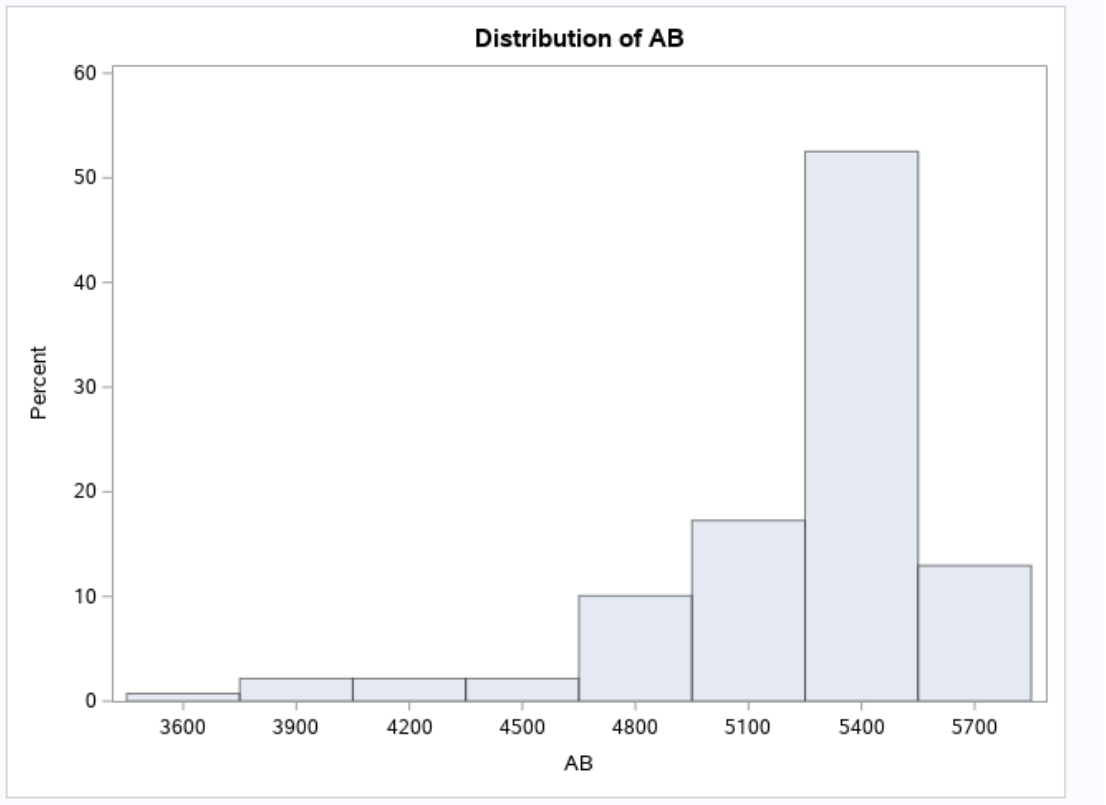

Distribution of AB
Percent
60
50
40
30
20
10
0
3600
3900
4200
4500
4800
5100
5400
5700
AB

Figure 4

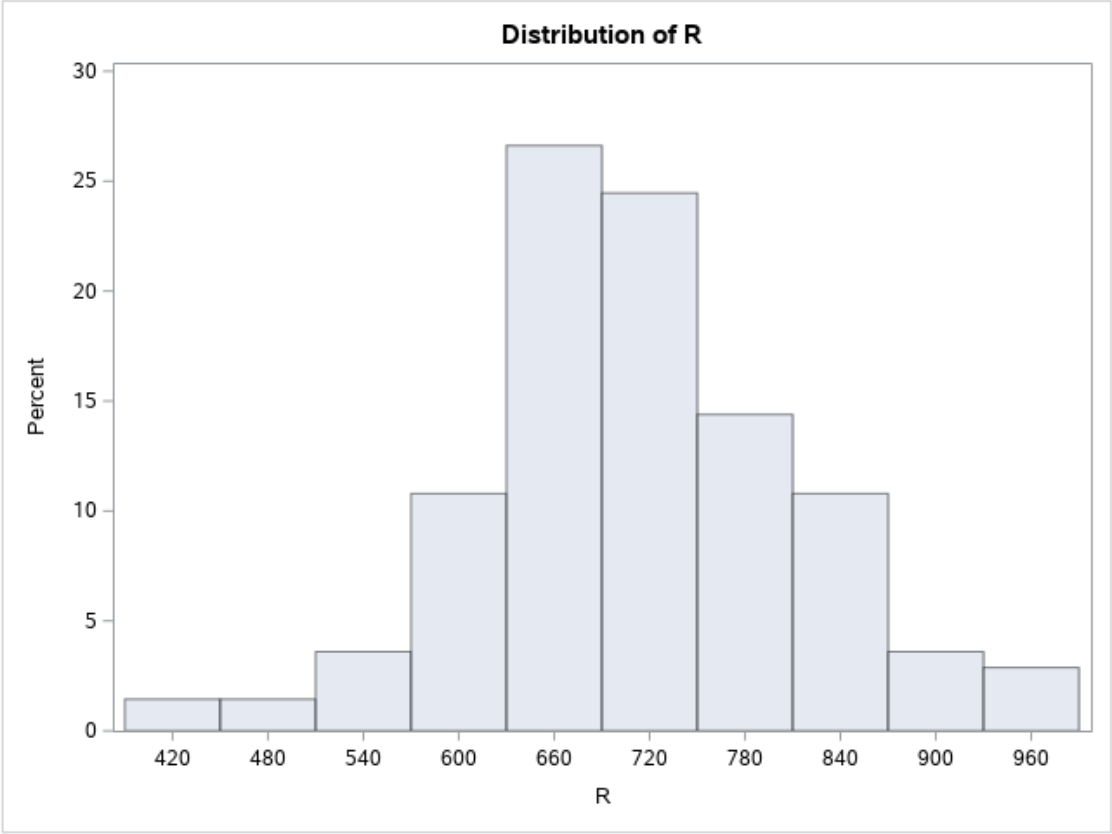
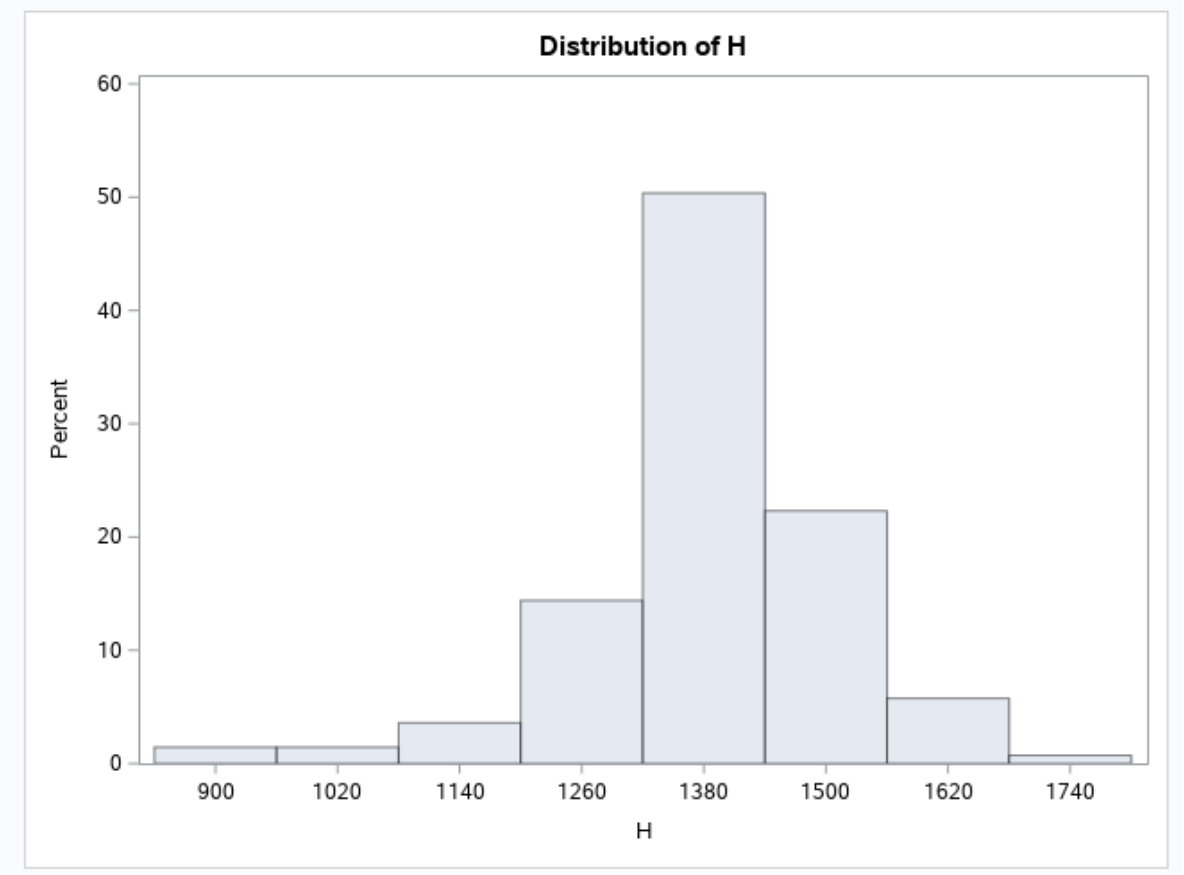
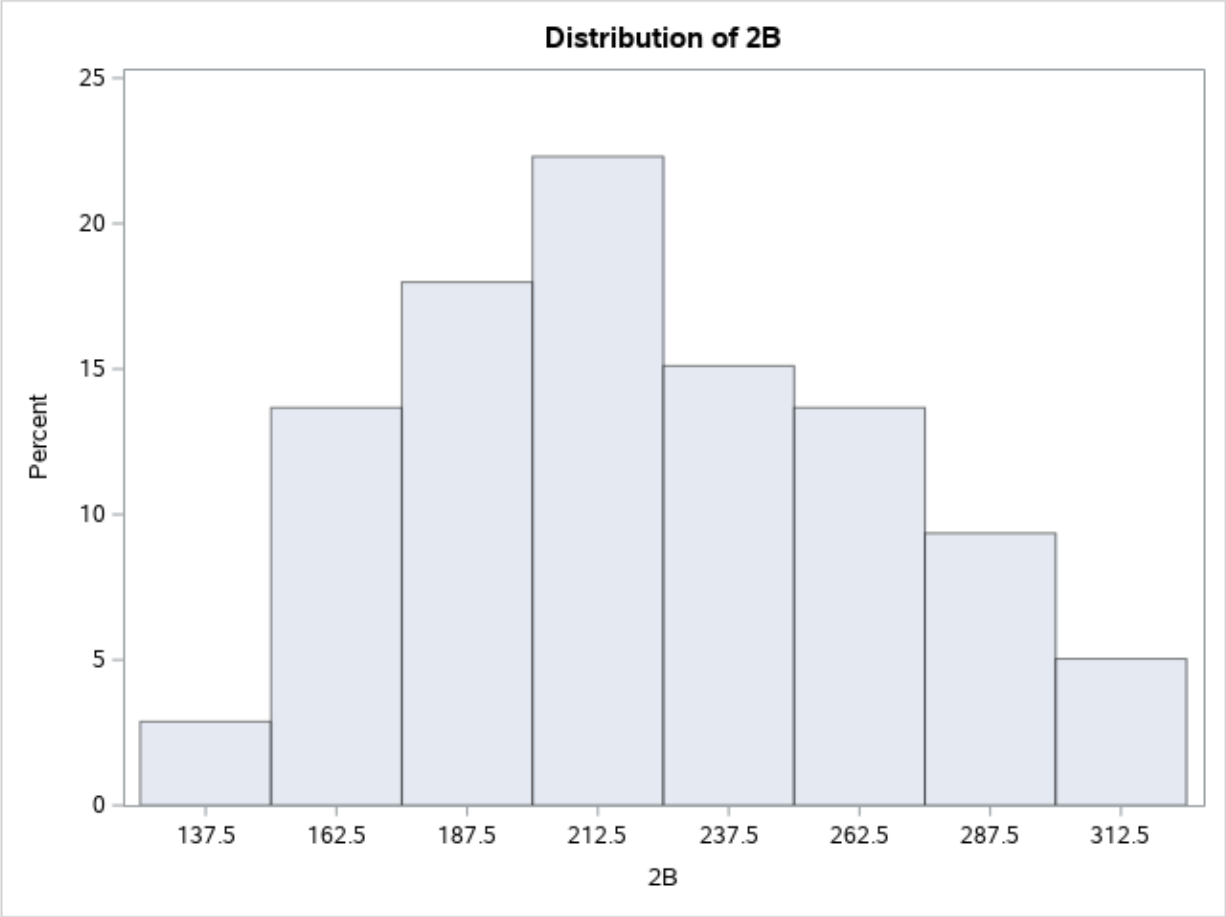
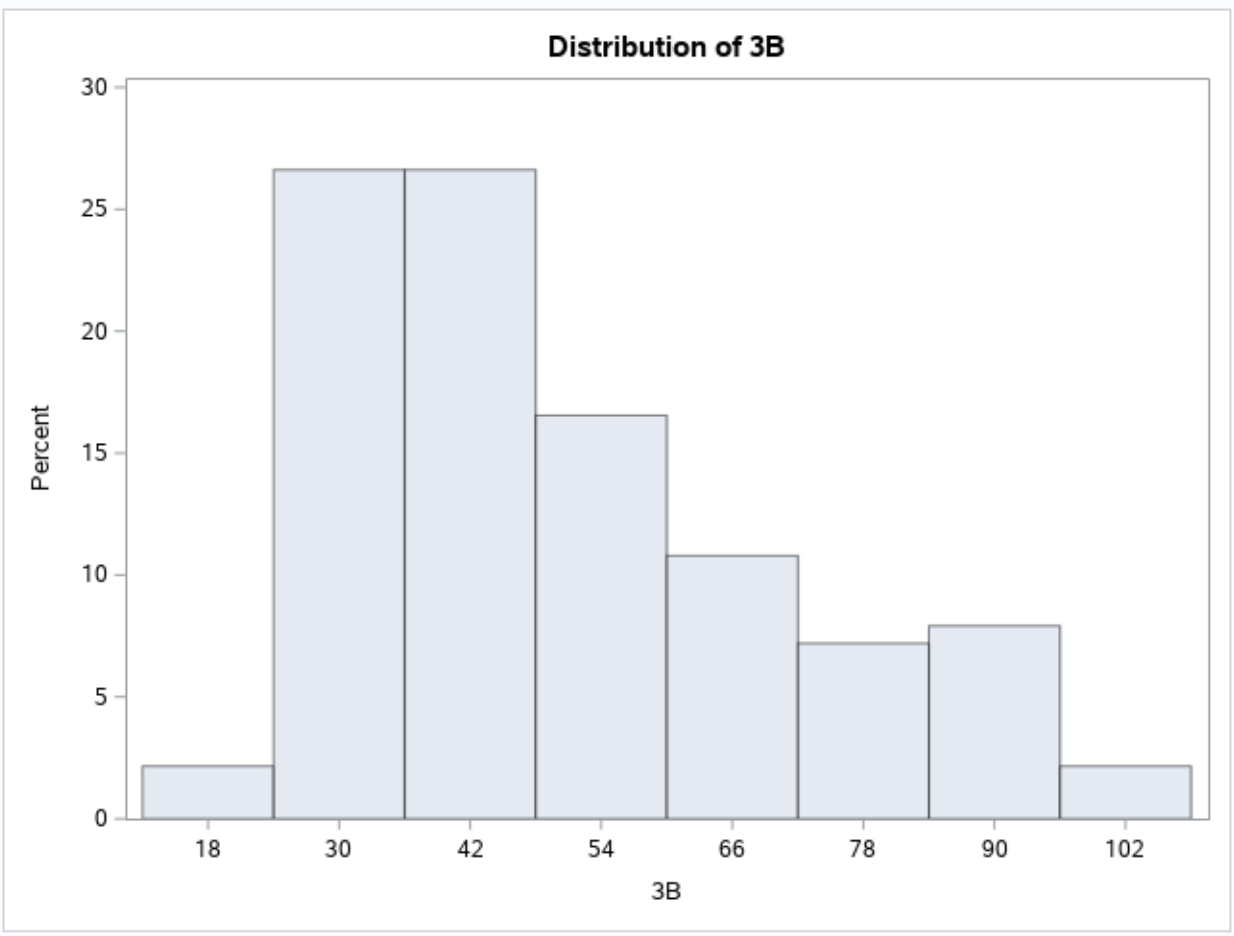


Figure 5


Distribution of H
Percent
60
50
40
30
20
10
0
900
1020
1140
1260
1380
1500
1620
1740
H
Figure 6


Distribution of 2B
Percent
25
20
15
10
5
0
137.5
162.5
187.5
212.5
237.5
262.5
287.5
312.5
2B
Figure 7


Distribution of 3B
Percent
30
25
20
15
10
5
0
18
30
42
54
66
78
90
102
3B
Figure 8

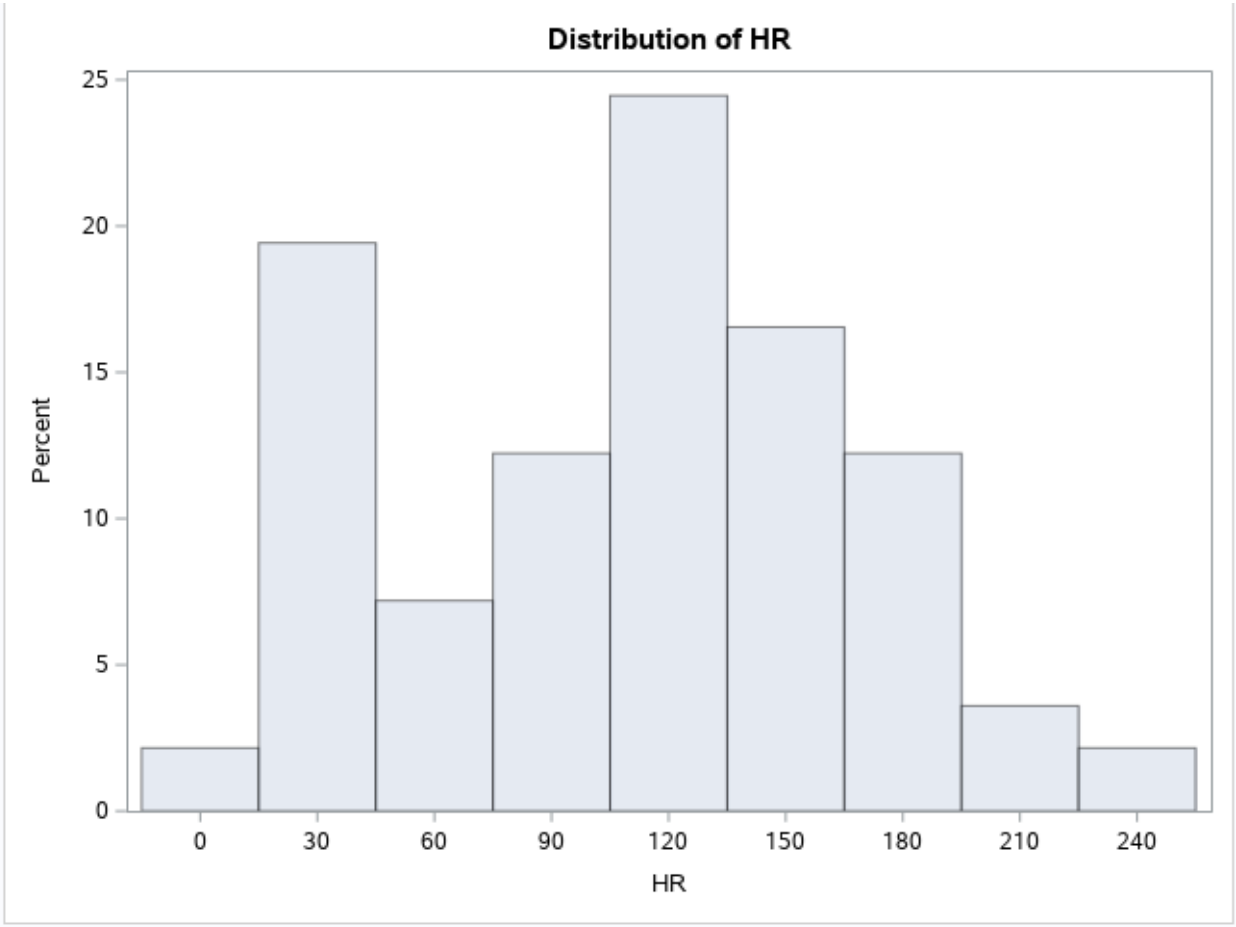

Distribution of HR
Percent
25
20
15
10
5
0
0
30
60
90
120
150
180
210
240
HR

Figure 9

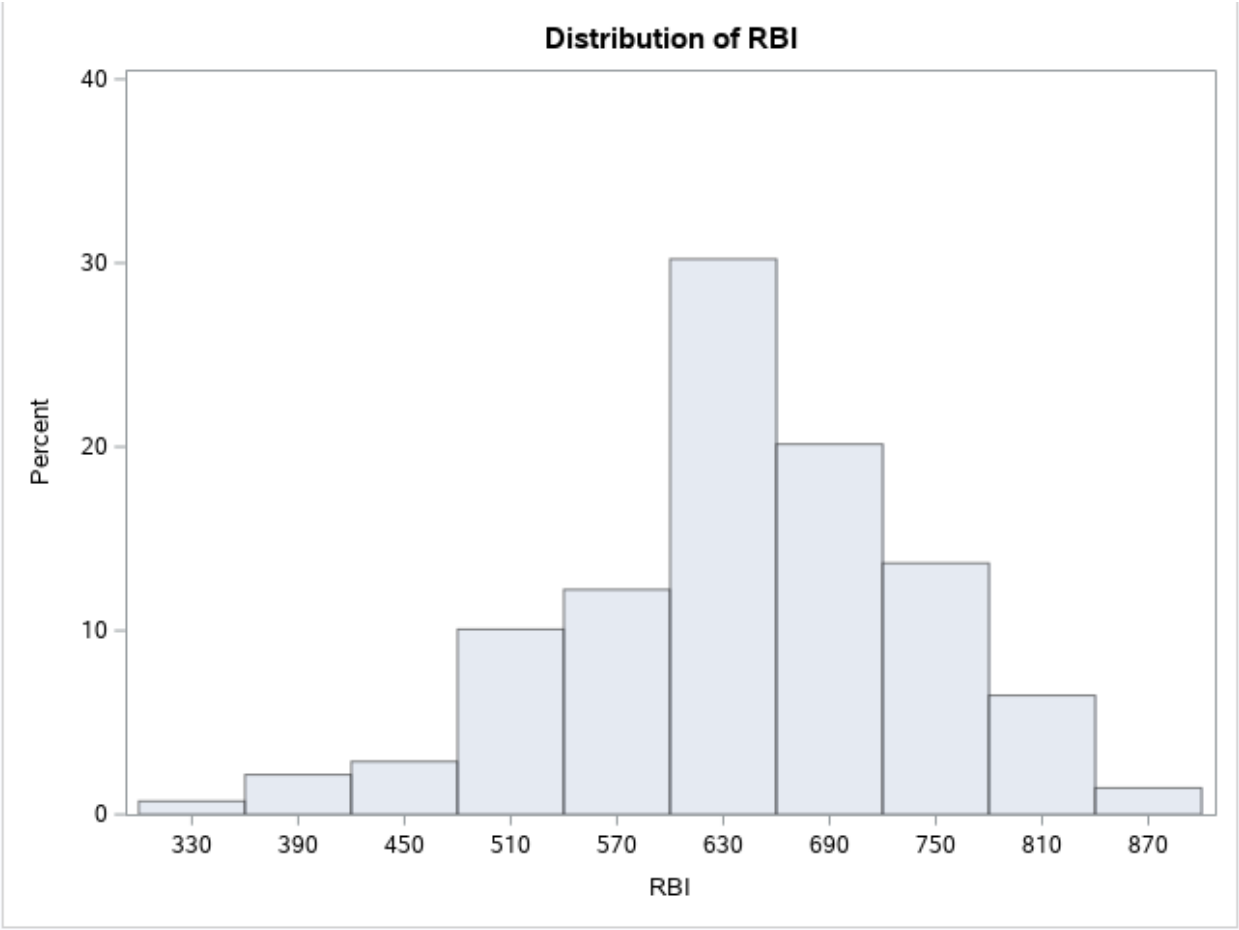

Distribution of RBI
Percent
40
30
20
10
0
330
390
450
510
570
630
690
750
810
870
RBI

Figure 10

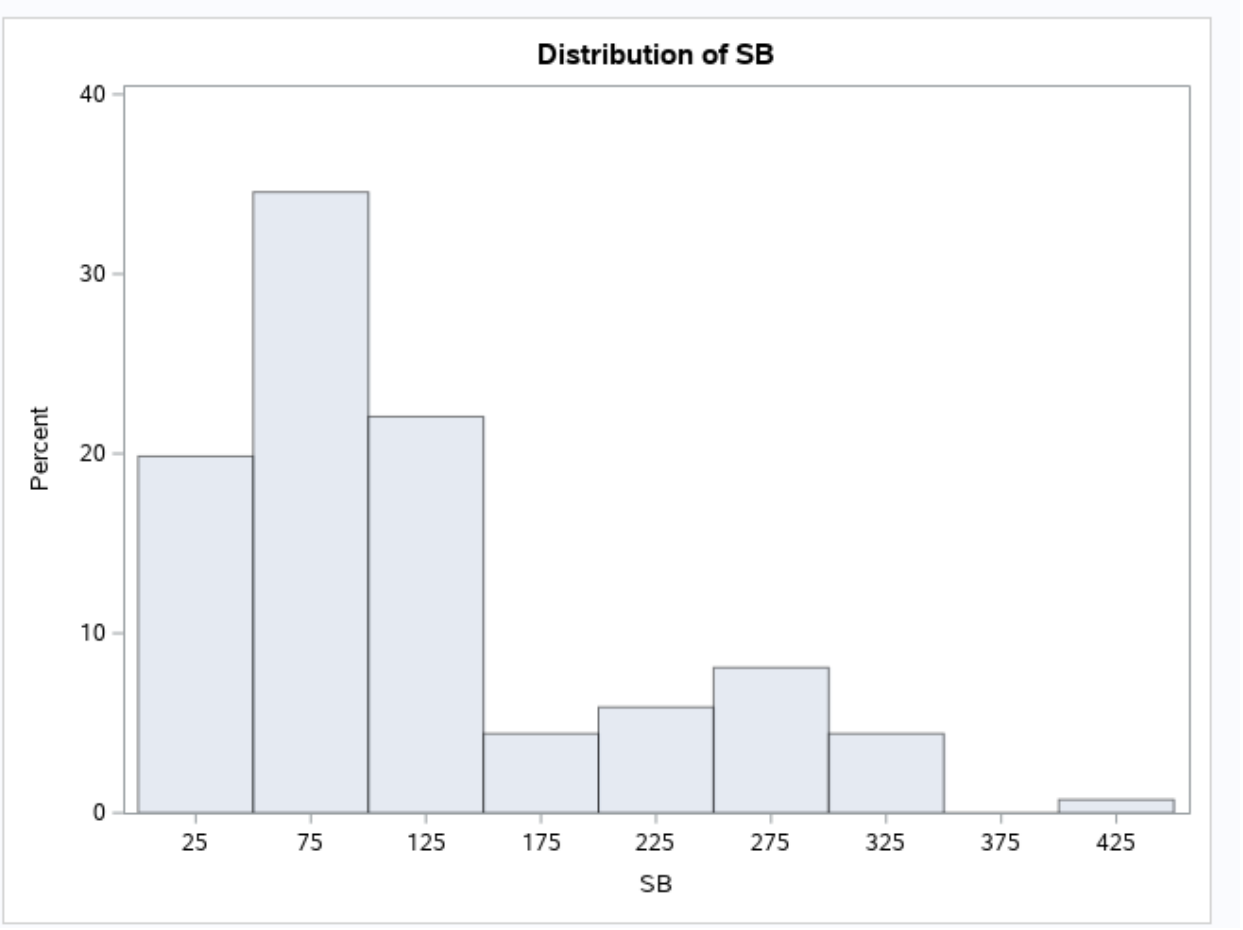
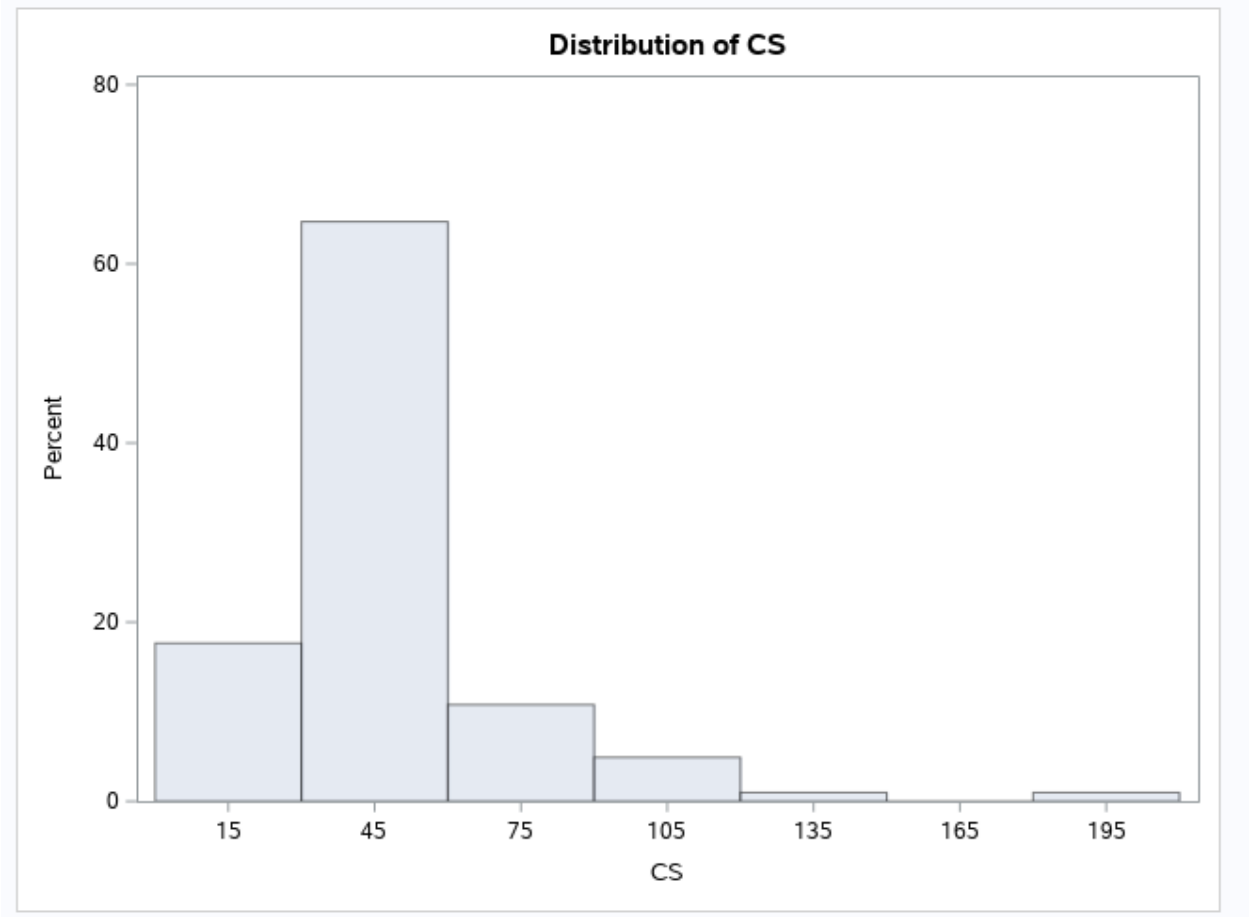
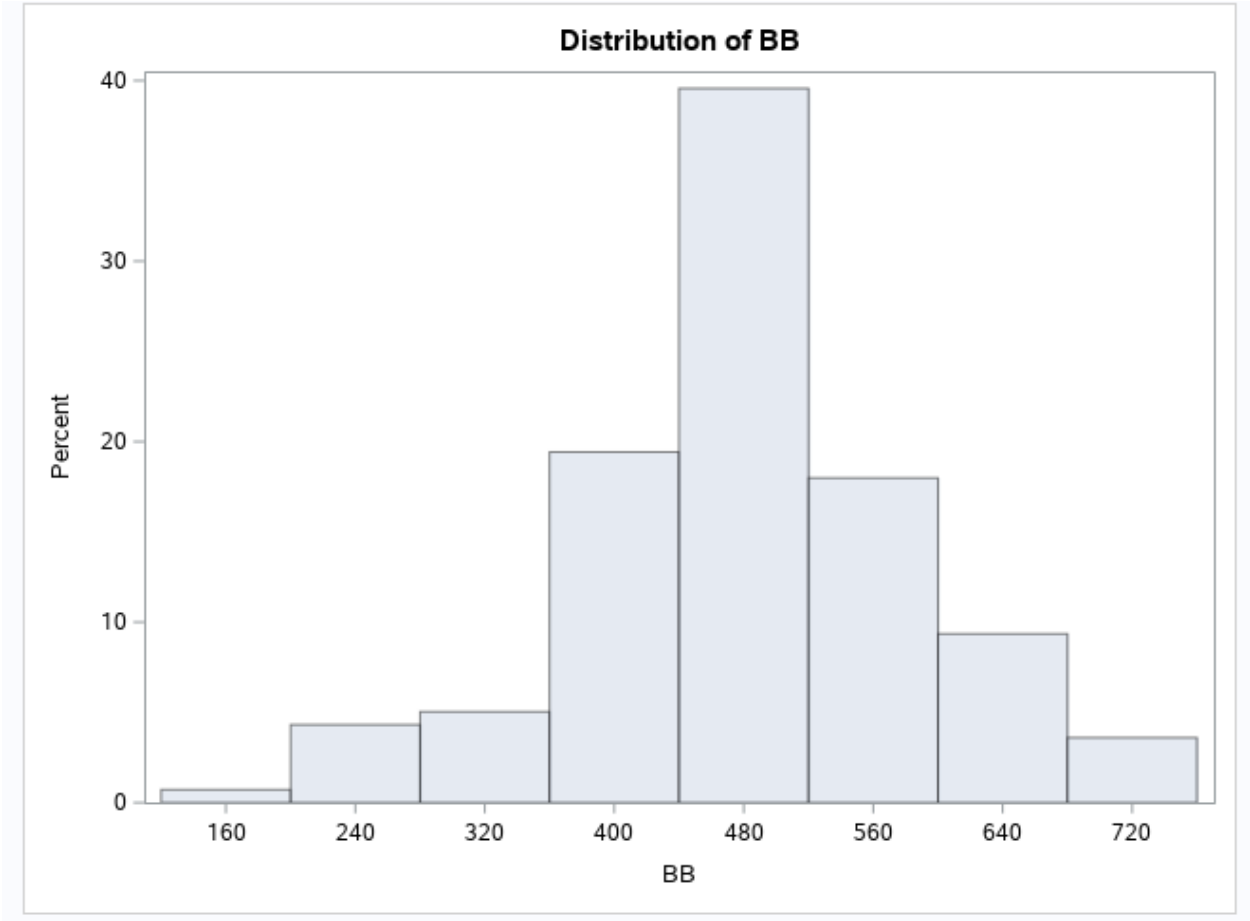

Distribution of SB
Percent
40
30
20
10
0
25
75
125
175
225
275
325
375
425
SB

Figure 11


Distribution of CS
Percent
80
60
40
20
0
15
45
75
105
135
165
195
CS
Figure 12


Distribution of BB
Percent
40
30
20
10
0
160
240
320
400
480
560
640
720
BB
Figure 13

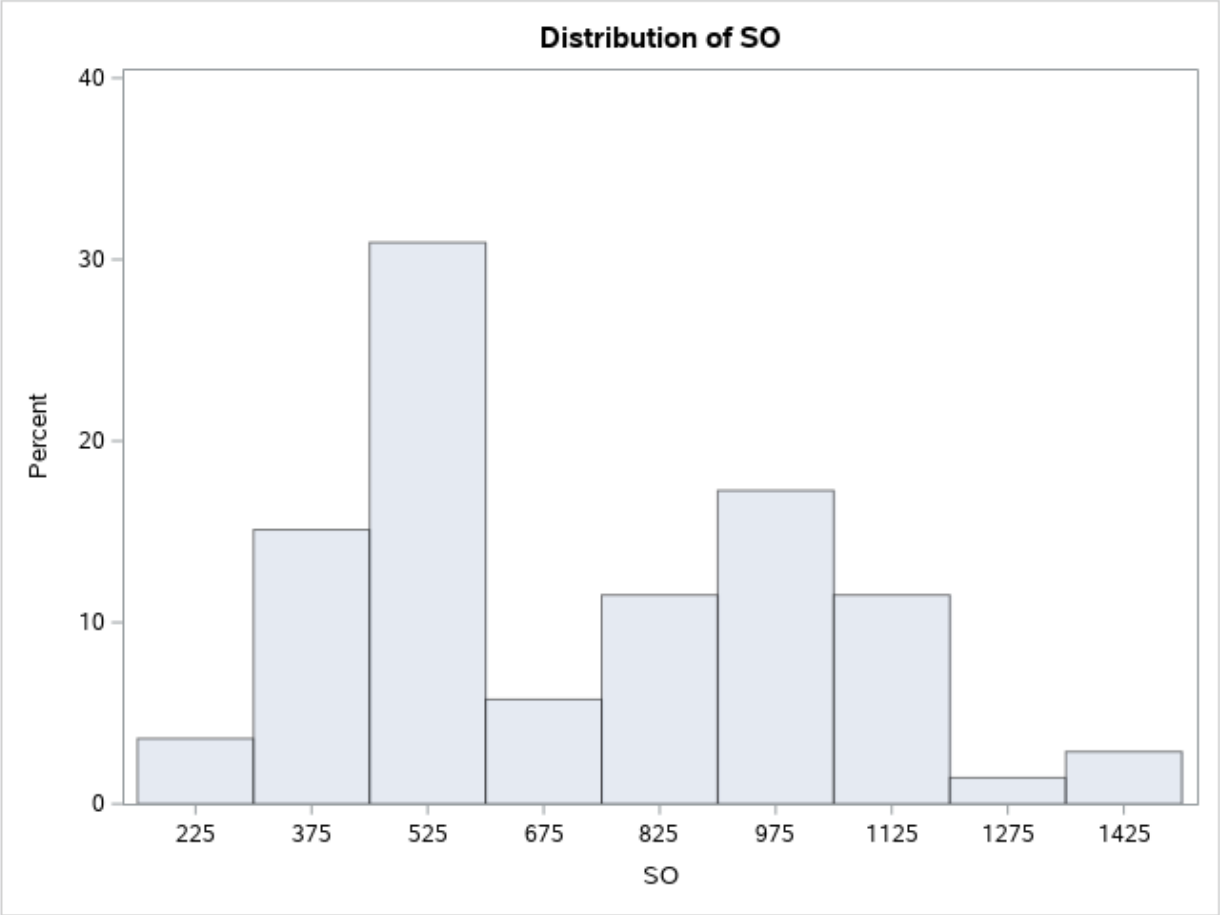

Distribution of SO
Percent
40
30
20
10
0
225
375
525
675
825
975
1125
1275
1425
SO

Figure 14

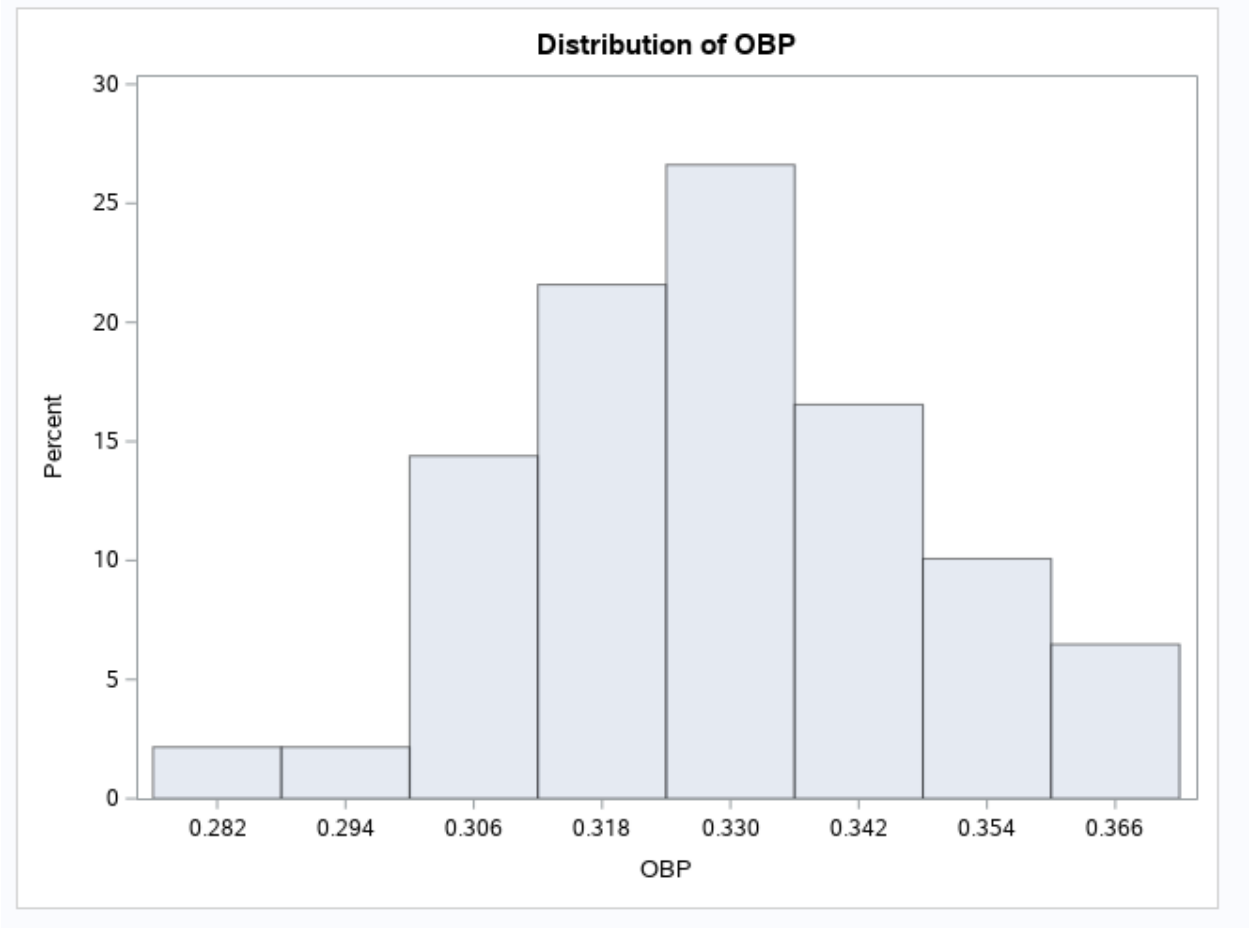

Distribution of OBP
Percent
30
25
20
15
10
5
0
0.282
0.294
0.306
0.318
0.330
0.342
0.354
0.366
OBP

Figure 15

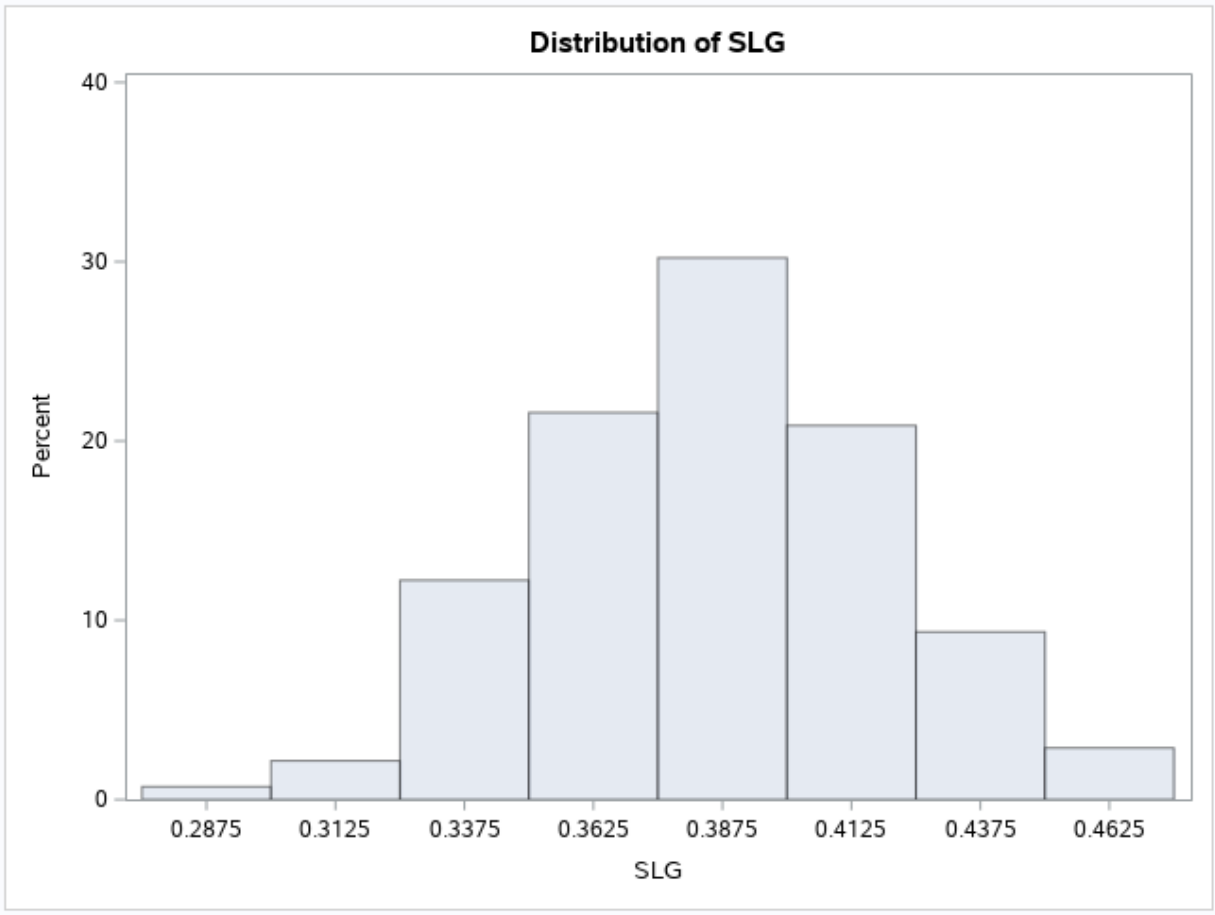

Distribution of SLG
Percent
40
30
20
10
0
0.2875
0.3125
0.3375
0.3625
0.3875
0.4125
0.4375
0.4625
SLG

Figure 16

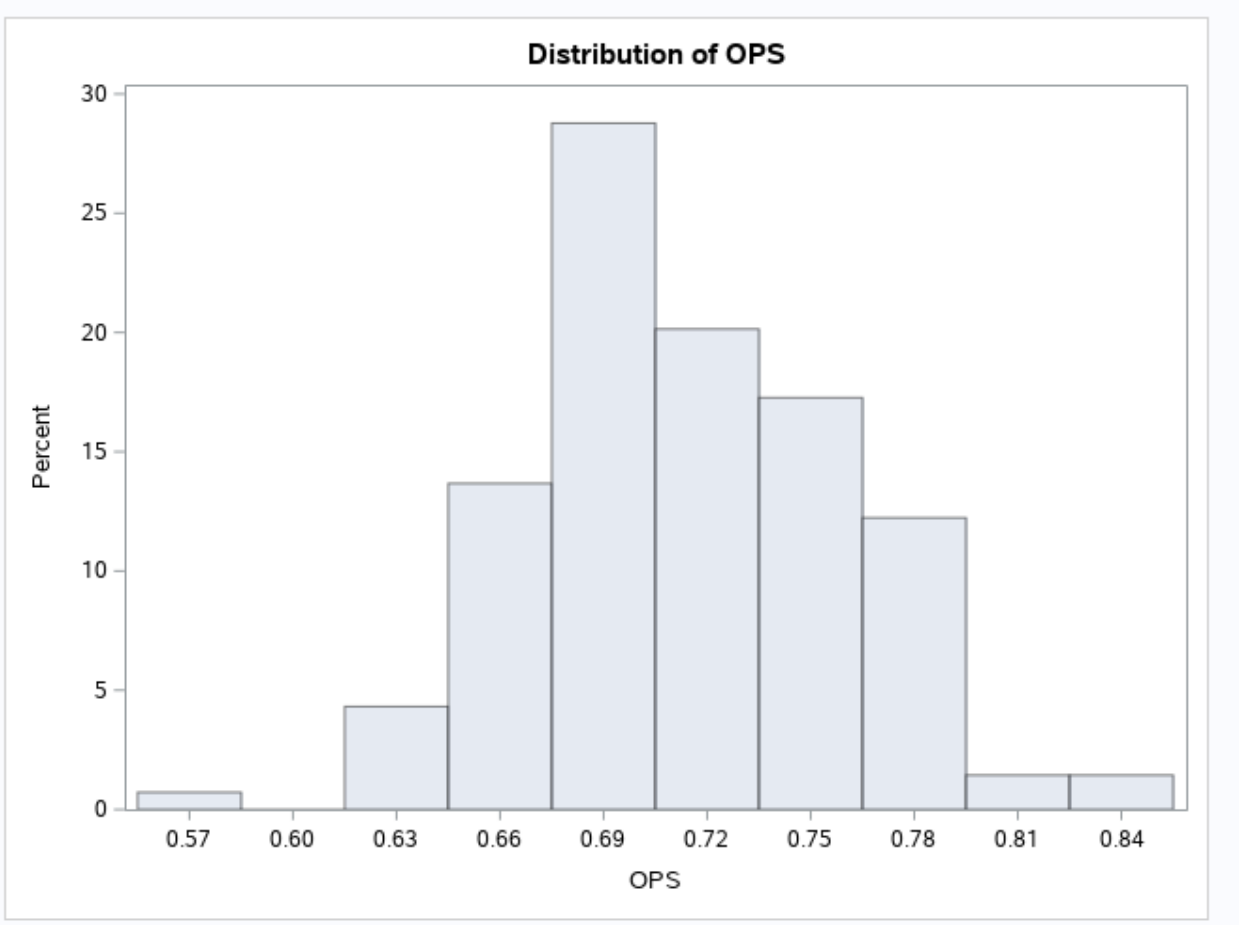
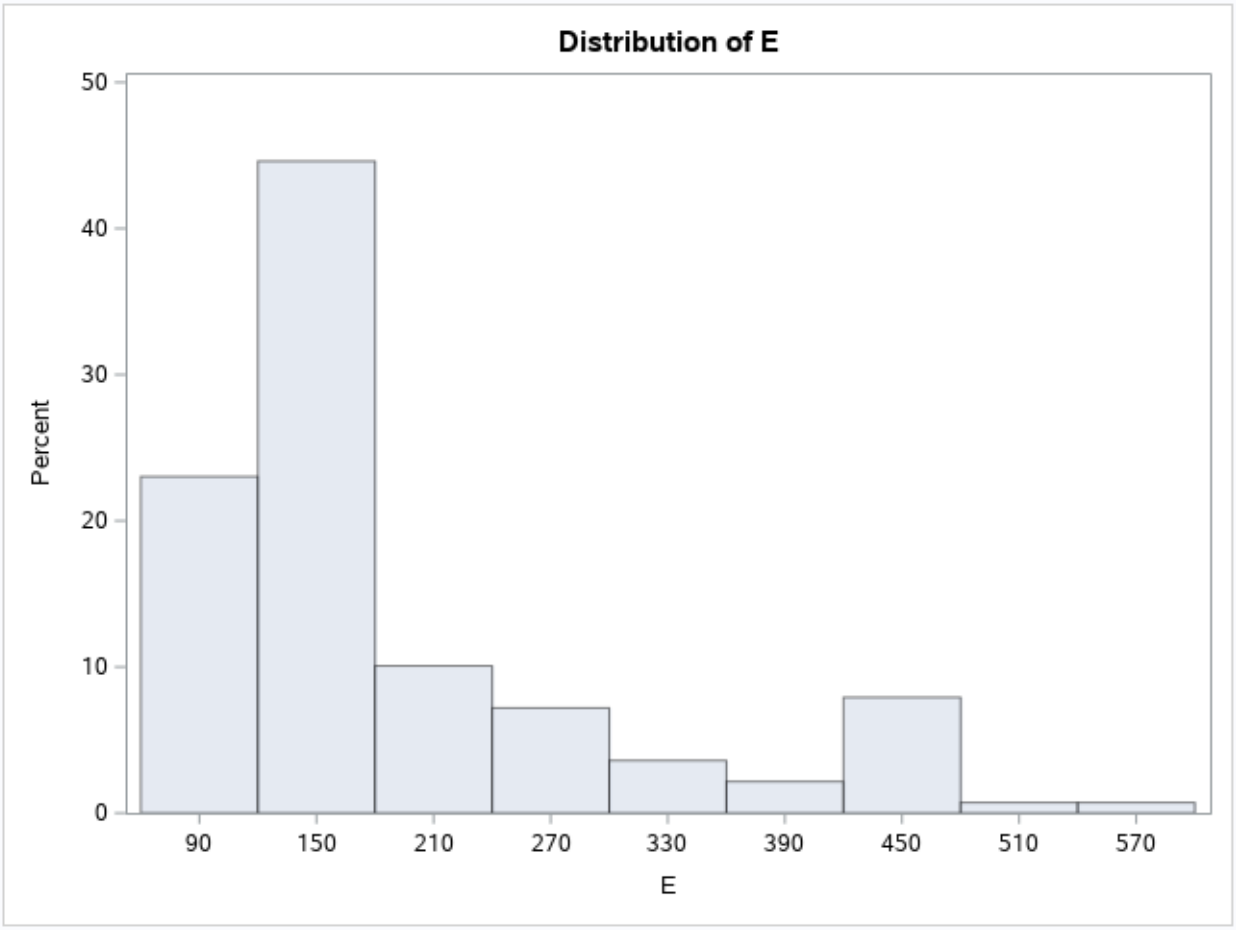

Distribution of OPS
Percent
30
25
20
15
10
5
0
0.57
0.60
0.63
0.66
0.69
0.72
0.75
0.78
0.81
0.84
OPS

Figure 17


Distribution of E
Percent
50
40
30
20
10
0
90
150
210
270
330
390
450
510
570
E
Figure 18

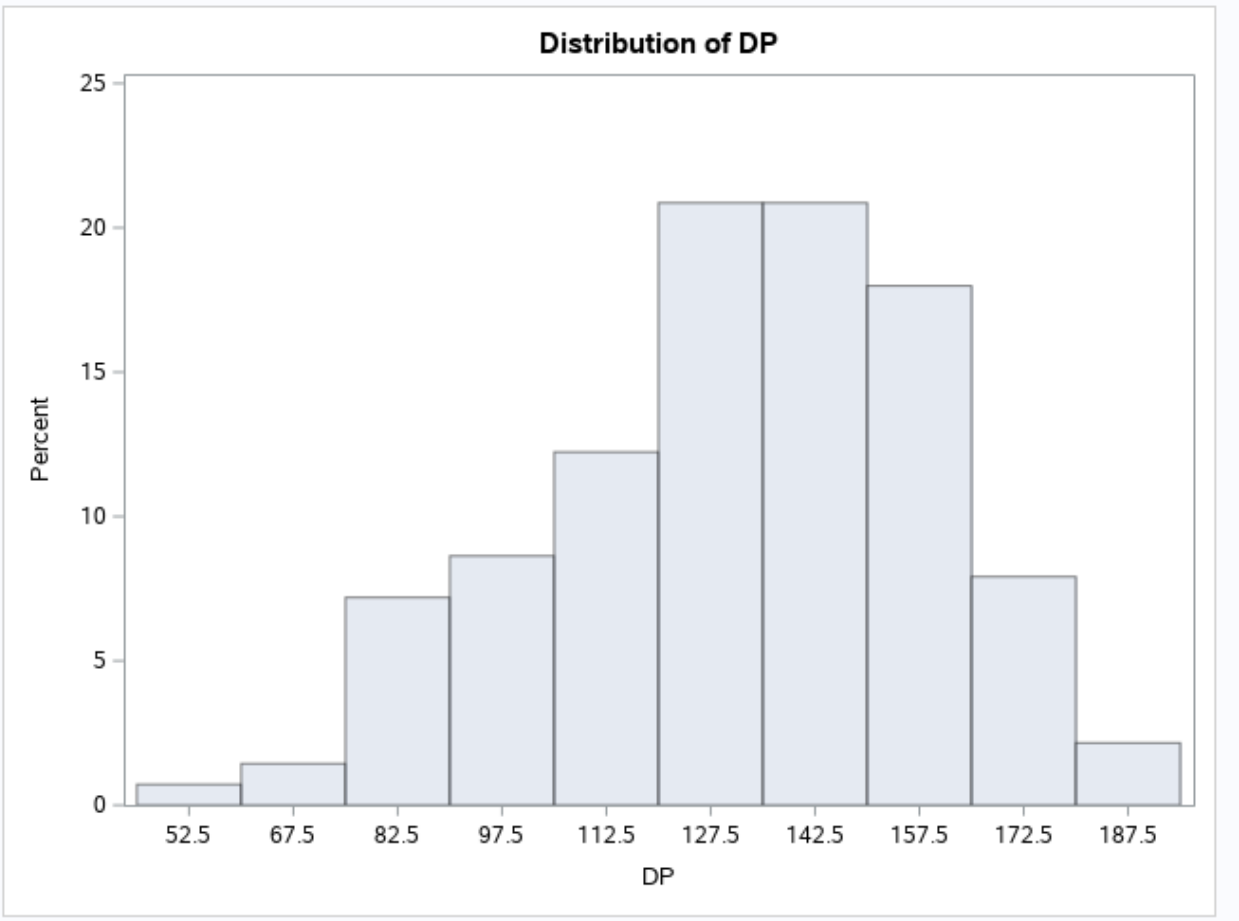

Distribution of DP
Percent
25
20
15
10
5
0
52.5
67.5
82.5
97.5
112.5
127.5
142.5
157.5
172.5
187.5
DP

Figure 19

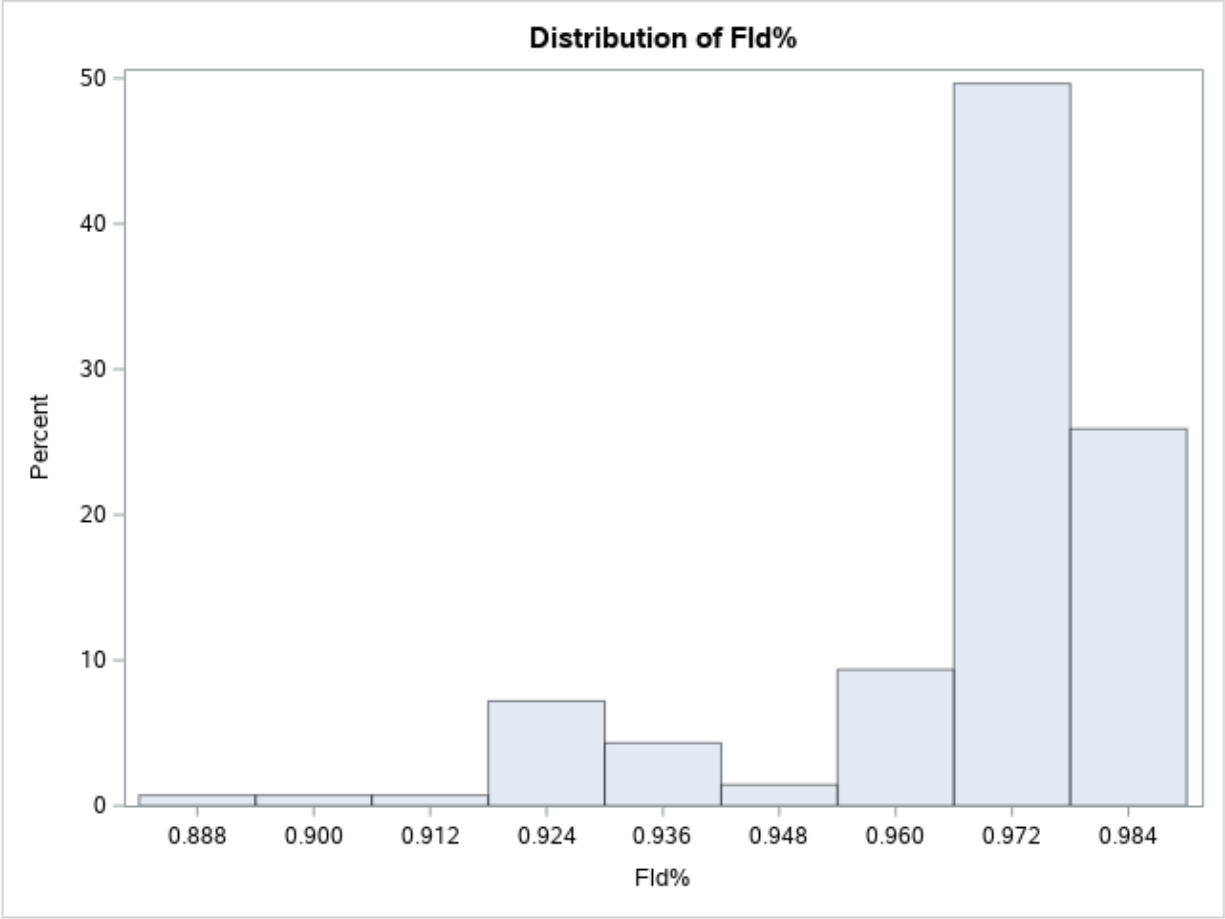
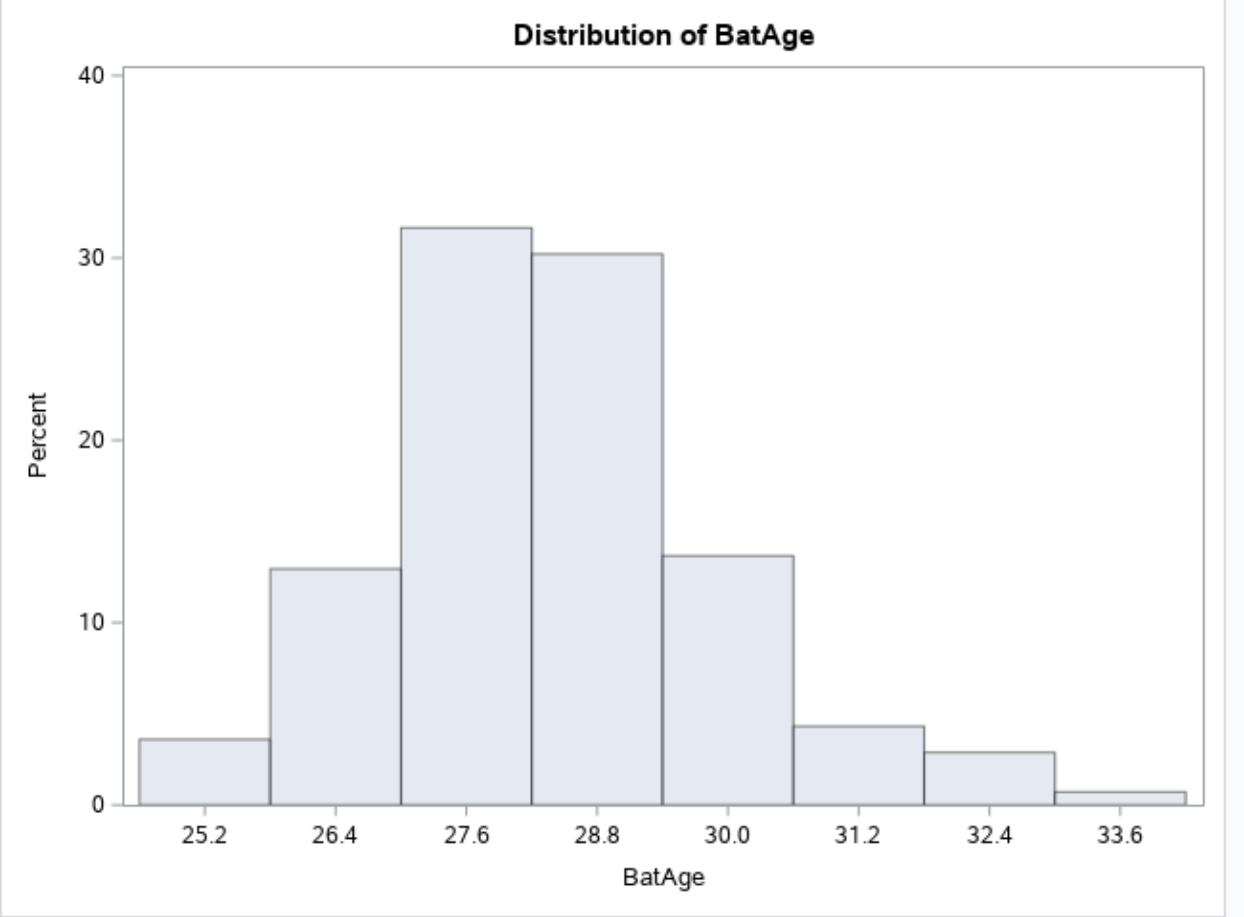

Distribution of Fld%
Percent
50
40
30
20
10
0
0.888
0.900
0.912
0.924
0.936
0.948
0.960
0.972
0.984
Fld%

Figure 20


Distribution of BatAge
Percent
40
30
20
10
0
25.2
26.4
27.6
28.8
30.0
31.2
32.4
33.6
BatAge
Figure 21

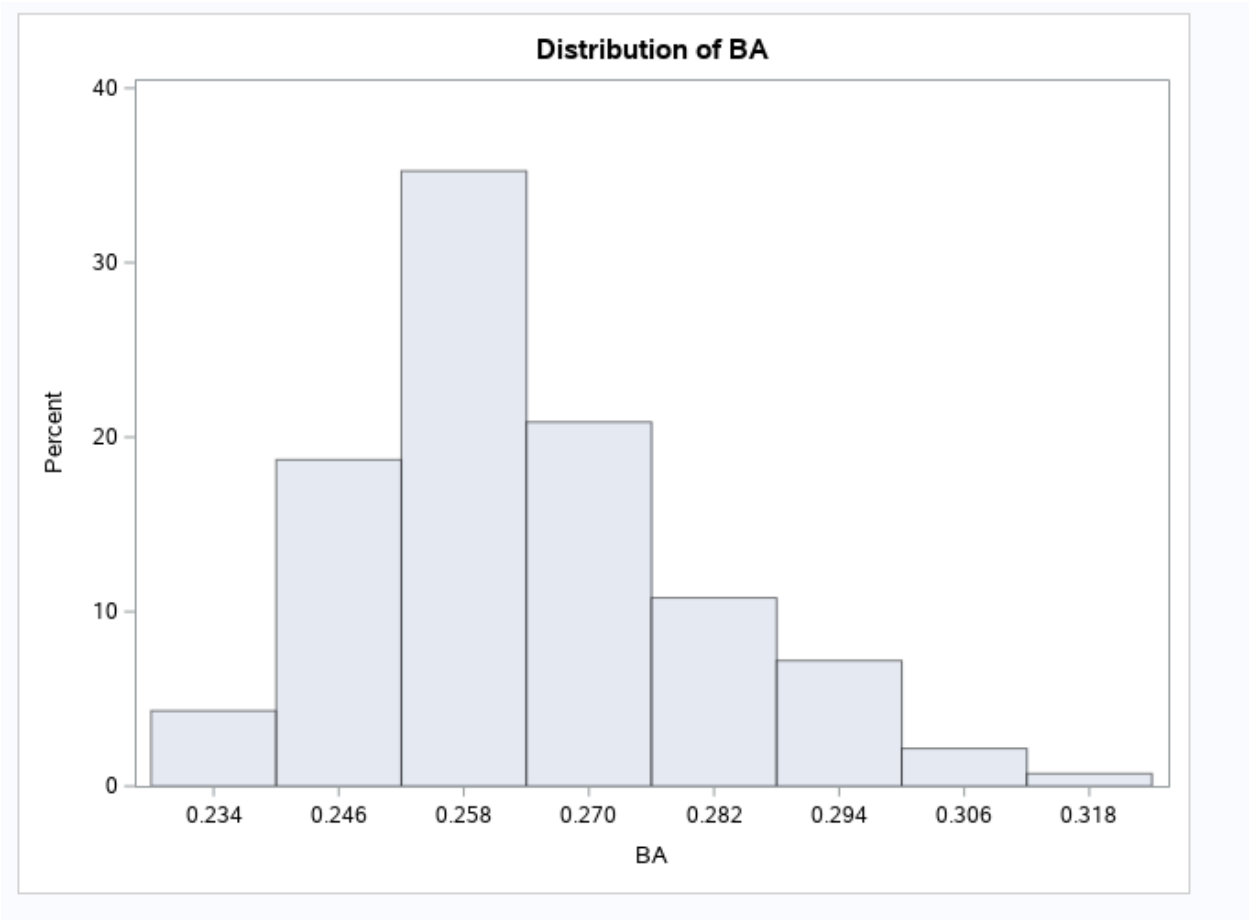

Distribution of BA
Percent
40
30
20
10
0
0.234
0.246
0.258
0.270
0.282
0.294
0.306
0.318
BA

Figure 22

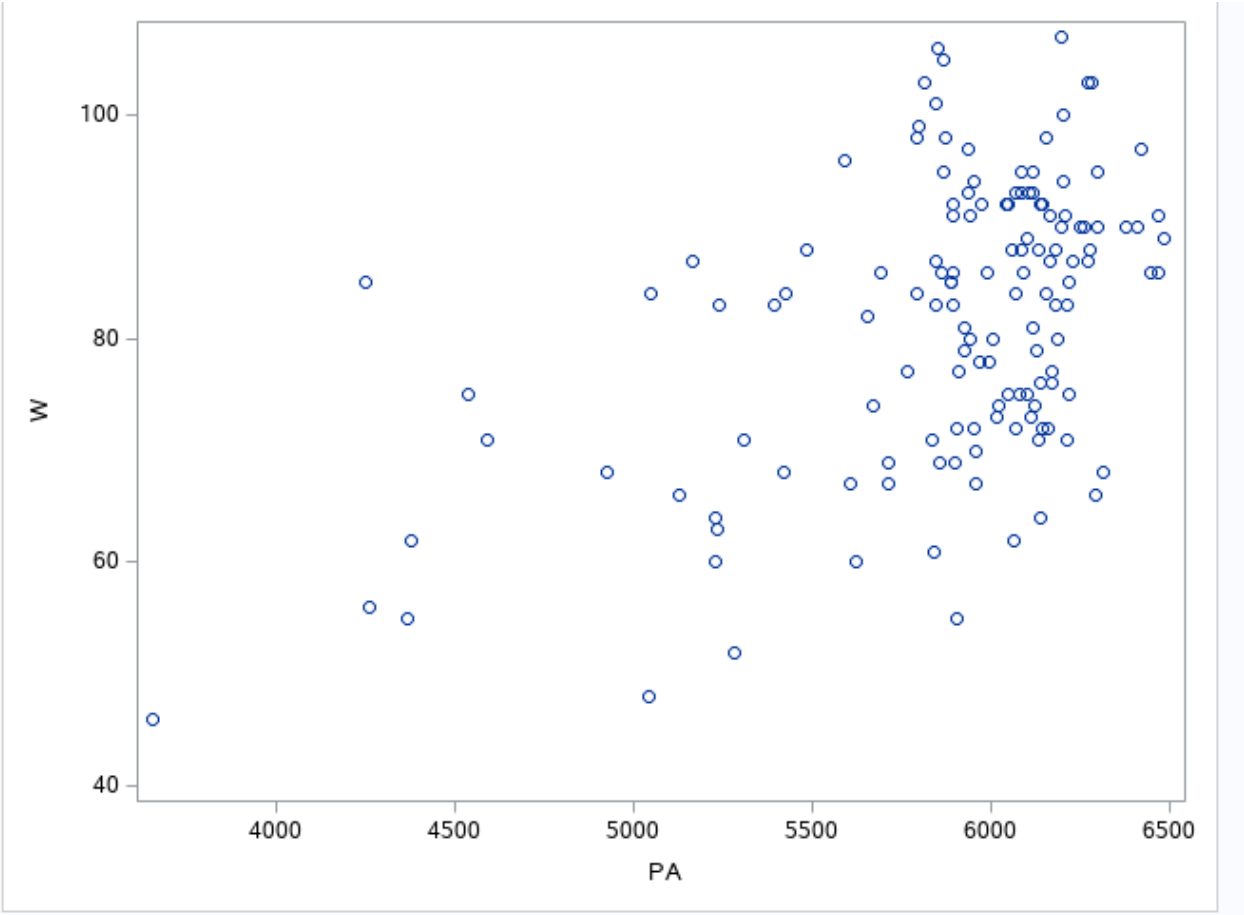


Figure 23

Figure 24

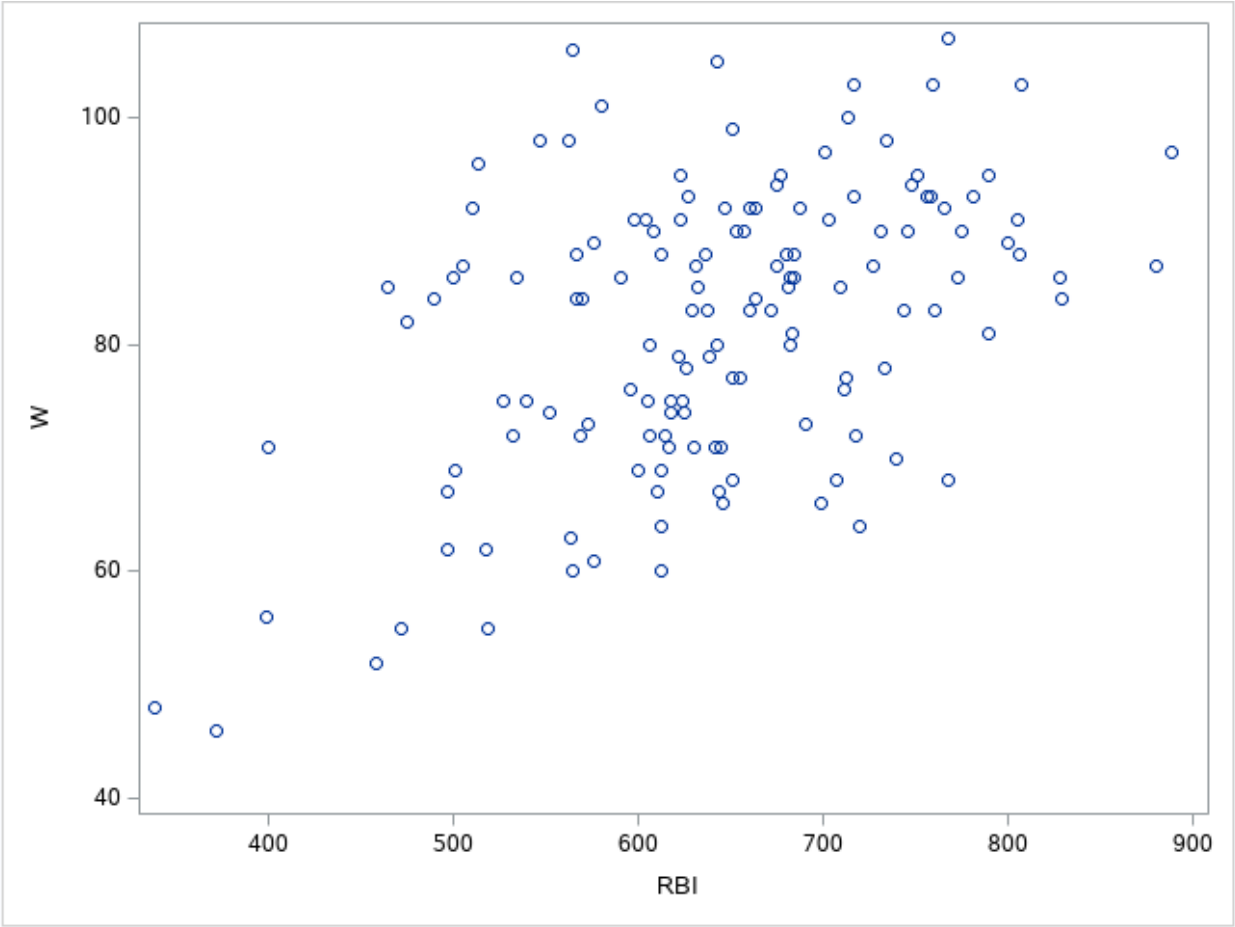


Figure 25

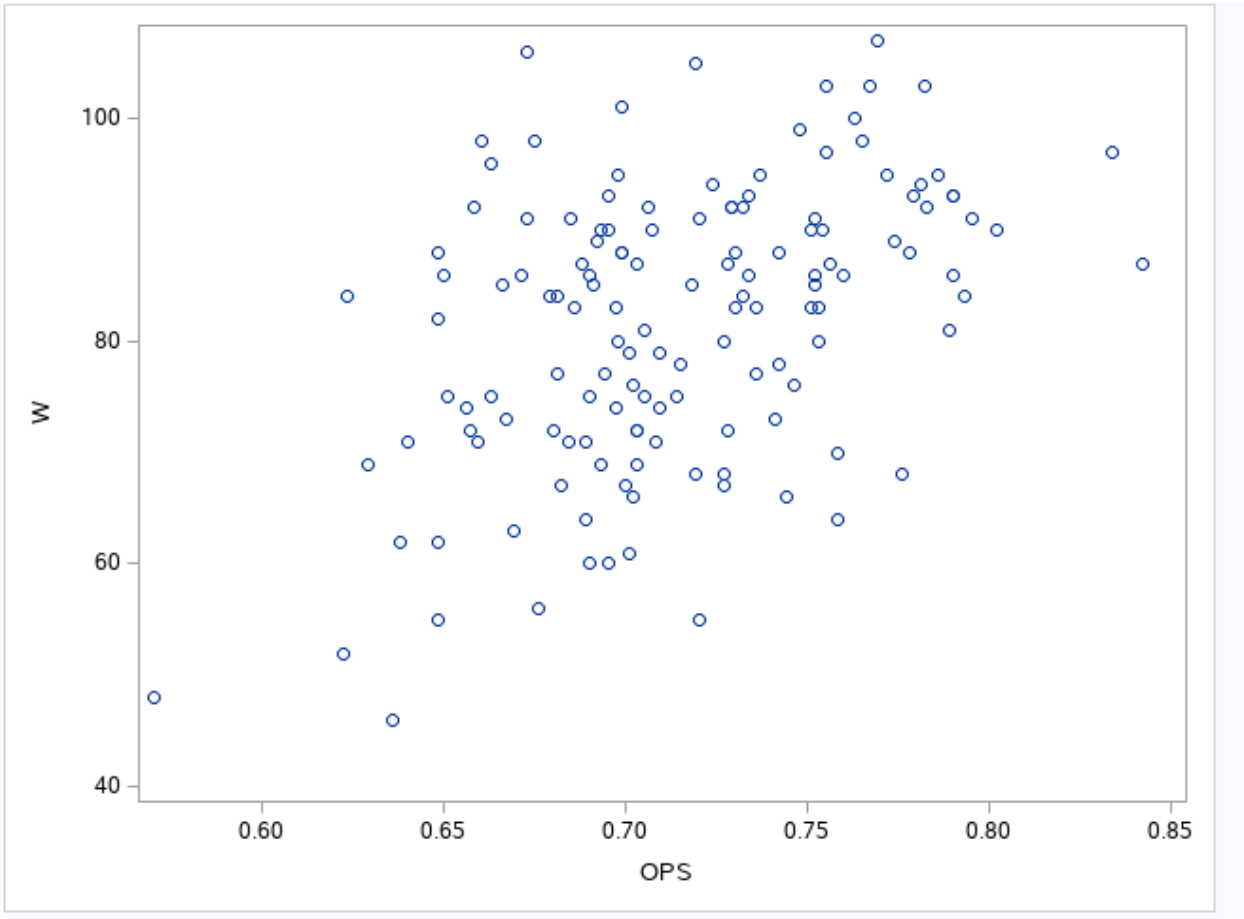


Figure 26

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6421.46660	3210.73330	27.95	<.0001
Error	136	15621	114.86004		
Corrected Total	138	22042			

Root MSE	10.71728	R-Square	0.2913
Dependent Mean	81.67626	Adj R-Sq	0.2809
Coeff Var	13.12166		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.04737	11.29821	0.00	0.9967
PA	1	0.00594	0.00276	2.16	0.0328
H	1	0.03384	0.01027	3.30	0.0013

Figure 27

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6553.31216	3276.65608	28.77	<.0001
Error	136	15489	113.89058		
Corrected Total	138	22042			

Root MSE	10.67195	R-Square	0.2973
Dependent Mean	81.67626	Adj R-Sq	0.2870
Coeff Var	13.06616		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.68028	11.02259	0.88	0.3814
PA	1	0.00810	0.00227	3.57	0.0005
RBI	1	0.03808	0.01094	3.48	0.0007

Figure 28

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6968.55534	3484.27767	31.44	<.0001
Error	136	15074	110.83733		
Corrected Total	138	22042			

Root MSE	10.52793	R-Square	0.3161
Dependent Mean	81.67626	Adj R-Sq	0.3061
Coeff Var	12.88983		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-33.82248	14.98574	-2.26	0.0256
PA	1	0.00939	0.00201	4.68	<.0001
OPS	1	84.50119	20.99997	4.02	<.0001

Figure 29

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6131.13628	3065.56814	26.20	<.0001
Error	136	15911	116.99482		
Corrected Total	138	22042			

Root MSE	10.81641	R-Square	0.2782
Dependent Mean	81.67626	Adj R-Sq	0.2675
Coeff Var	13.24303		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.65869	10.40889	1.70	0.0921
H	1	0.03585	0.01211	2.96	0.0036
RBI	1	0.02260	0.01565	1.44	0.1510

Figure 30

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6189.62268	3094.81134	26.55	<.0001
Error	136	15853	116.56477		
Corrected Total	138	22042			

Root MSE	10.79652	R-Square	0.2808
Dependent Mean	81.67626	Adj R-Sq	0.2702
Coeff Var	13.21867		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.98301	14.16906	-0.28	0.7791
H	1	0.03823	0.01018	3.76	0.0003
OPS	1	46.00898	28.56272	1.61	0.1095

Figure 31

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5125.02720	2562.51360	20.60	<.0001
Error	136	16917	124.39268		
Corrected Total	138	22042			

Root MSE	11.15315	R-Square	0.2325
Dependent Mean	81.67626	Adj R-Sq	0.2212
Coeff Var	13.65531		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.78700	23.70673	1.43	0.1564
RBI	1	0.05145	0.02383	2.16	0.0326
OPS	1	20.81176	51.72814	0.40	0.6881

Figure 32

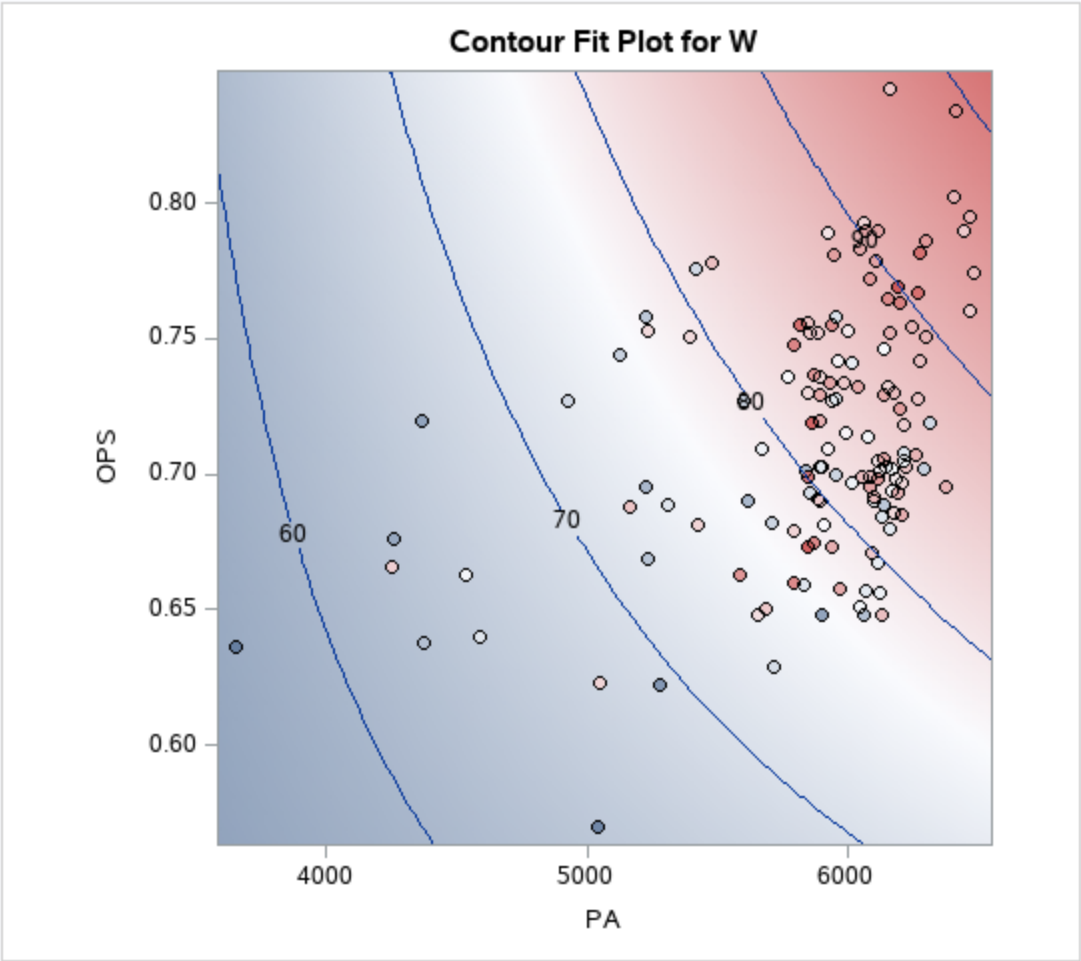
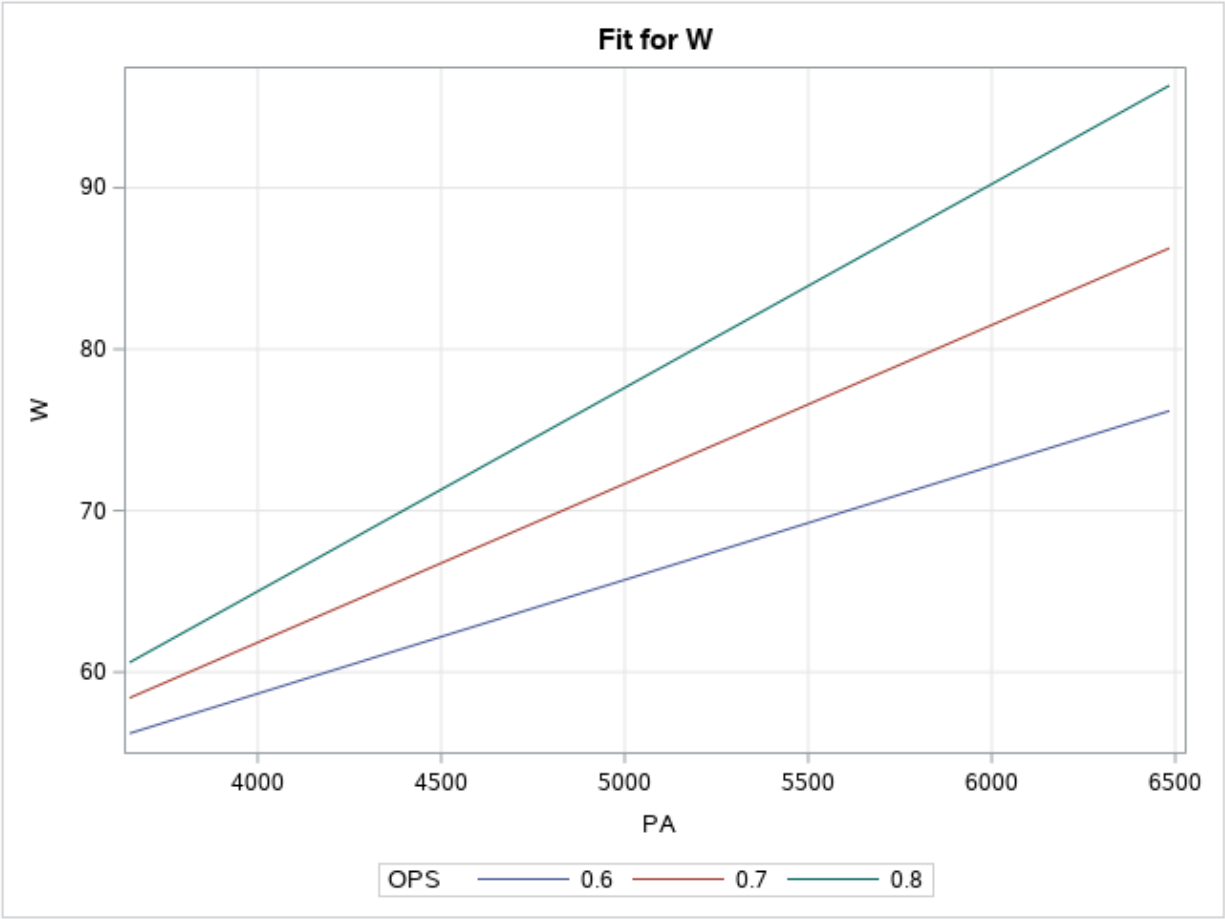

Contour Fit Plot for W
0.80
0.75
0.70
0.65
0.60
OPS
60
70
80
4000
5000
6000
PA

Figure 33


Fit for W
90
80
70
60
W
4000
4500
5000
5500
6000
6500
PA
OPS
0.6
0.7
0.8
Figure 34

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6575.00135	2191.66712	19.13	<.0001
Error	135	15467	114.57356		
Corrected Total	138	22042			

Root MSE	10.70390	R-Square	0.2983
Dependent Mean	81.67626	Adj R-Sq	0.2827
Coeff Var	13.10528		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.14452	46.08192	0.63	0.5282
PA	1	0.00473	0.00807	0.59	0.5582
RBI	1	0.00133	0.08516	0.02	0.9876
RBI_PA	1	0.00000629	0.00001446	0.44	0.6642

Figure 35

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7007.34591	2335.78197	20.97	<.0001
Error	135	15035	111.37101		
Corrected Total	138	22042			

Root MSE	10.55325	R-Square	0.3179
Dependent Mean	81.67626	Adj R-Sq	0.3027
Coeff Var	12.92082		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-75.32574	71.91075	-1.05	0.2967
PA	1	0.02474	0.02609	0.95	0.3446
OPS	1	86.04725	21.21285	4.06	<.0001
PA2	1	-0.00000143	0.00000243	-0.59	0.5561

Figure 36

The REG Procedure
Model: MODEL1
Dependent Variable: W

Number of Observations Read	139
Number of Observations Used	139

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6997.20494	2332.40165	20.93	<.0001
Error	135	15045	111.44612		
Corrected Total	138	22042			

Root MSE	10.55680	R-Square	0.3174
Dependent Mean	81.67626	Adj R-Sq	0.3023
Coeff Var	12.92518		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-110.80245	152.56948	-0.73	0.4689
PA	1	0.00921	0.00204	4.51	<.0001
OPS	1	302.71023	430.88877	0.70	0.4836
OPS2	1	-151.91827	299.62857	-0.51	0.6130

Figure 37