Sebastian Cerna
[redacted]

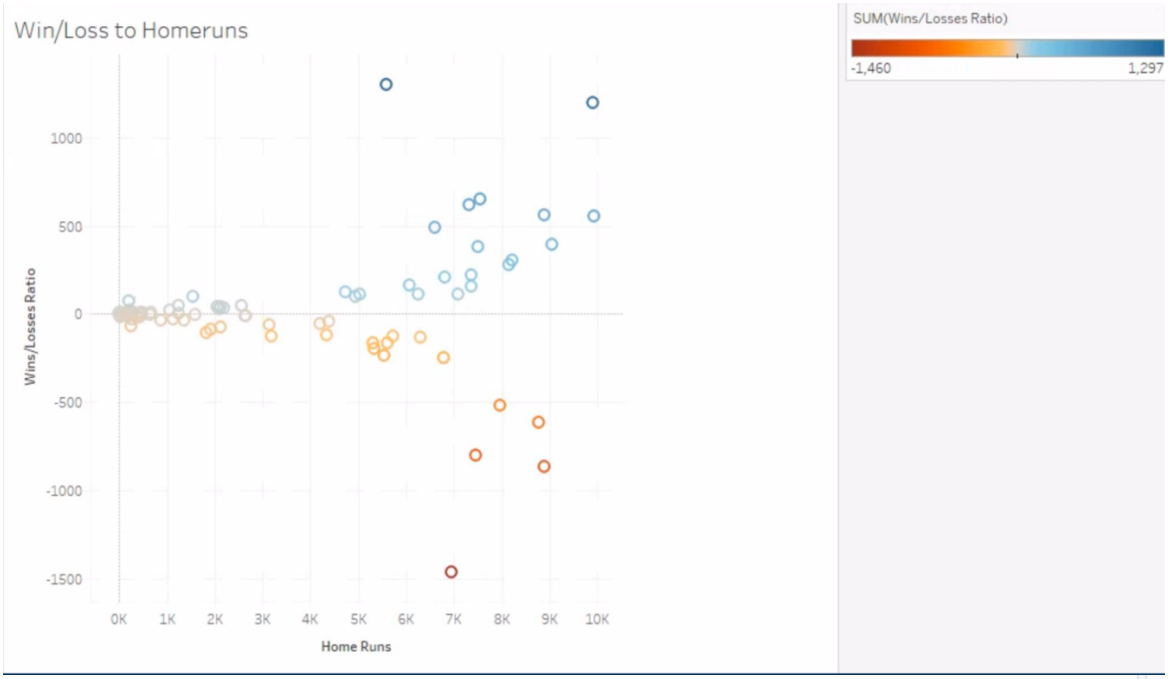
Business Understanding:

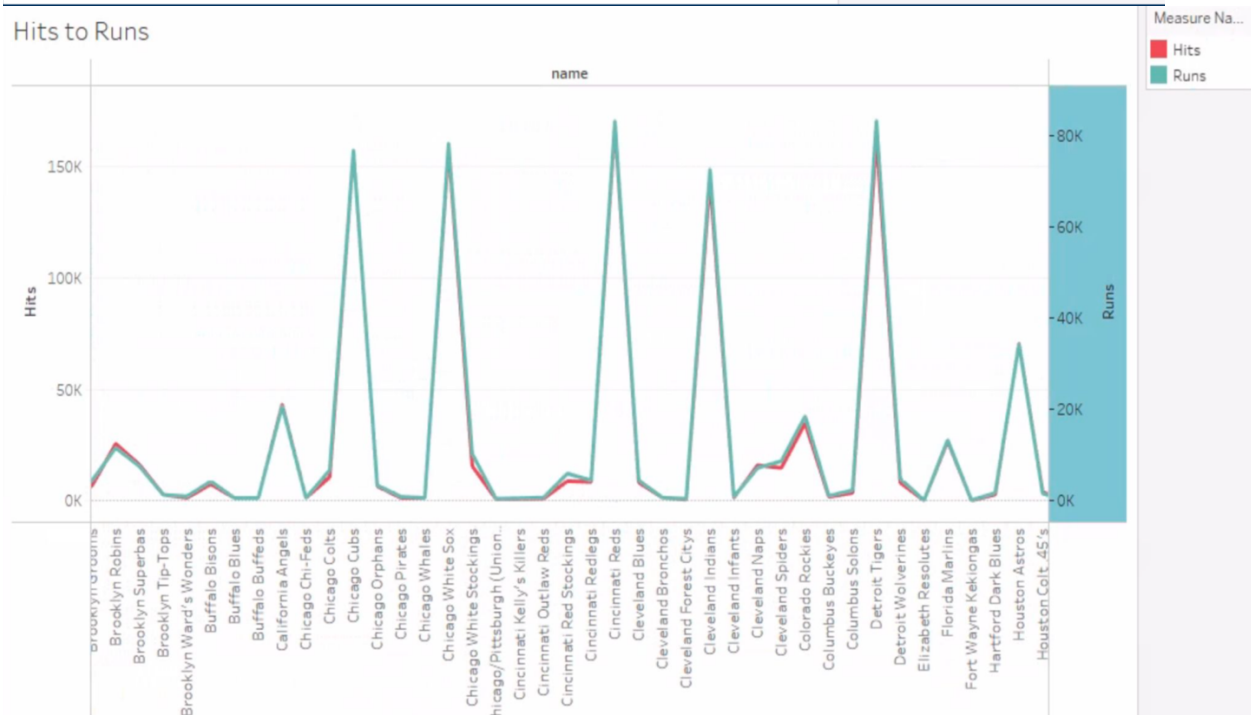
Numerous questions can be answered by our dataset, such as: Which team statistics have the strongest correlation with wins and losses? What impact do statistics like batting averages, home runs, or pitches have on a team's overall performance? Which player qualities are most important for winning games? This dataset's significance is in its comprehensive analysis of team performance, which can help players and coaches make strategic decisions both on and off the field. By focusing on power hitters and strong pitchers, finding metrics which produce the best results, and improving recruitment strategies, it assists major league baseball organizations in improving their team structure. Coaches can better allocate resources, prioritize training programs, and adjust game plans for the most wins by examining the connections and trends within each team.

Data Understanding:

Using K-means clustering we were able to determine the most important attributes that a team needs to increase their win percentage. In our K-means clustering we created four different clusters using the number of all-star players on a team, the win percentage, runs per season, and strikeouts per season. We noticed that the most important attributes to a teams win percentage was how many Allstars they had. Team with only one all-star had the lowest win percentage, while teams with 5 Allstars had the highest win percentage. The mid-tier teams average 3 Allstars. From the clustering we can also see that the worse performing teams had significantly less runs and strikeouts than the higher performing teams. The mid-tier performing teams were either good at pitching or good at batting while the top teams weren't the best at pitching or hitting, but they had good performance in both. Wins and losses, home runs, pitching metrics, hitting statistics, and all-star players were among the crucial fields we used. A scatter plot illustrating the correlation between home runs and victories is our initial visualization. This graphic shows how home runs may win games on average, but it also illustrates how it varies depending on how many home runs the sides score. A line chart that contrasts a team's hits and runs is the second visualization. The graph makes it clear that there are more runs than hits. The third graph helped us visualize our finding that the amount of Allstars a team has directly correlates to their win percentage. You can see an increase in a teams win percentage every time they gain an all star and their win percentage dips when they lose one. There are some outliers

where the win percentage doesn't directly follow the all-star count, but majority of the time it does.

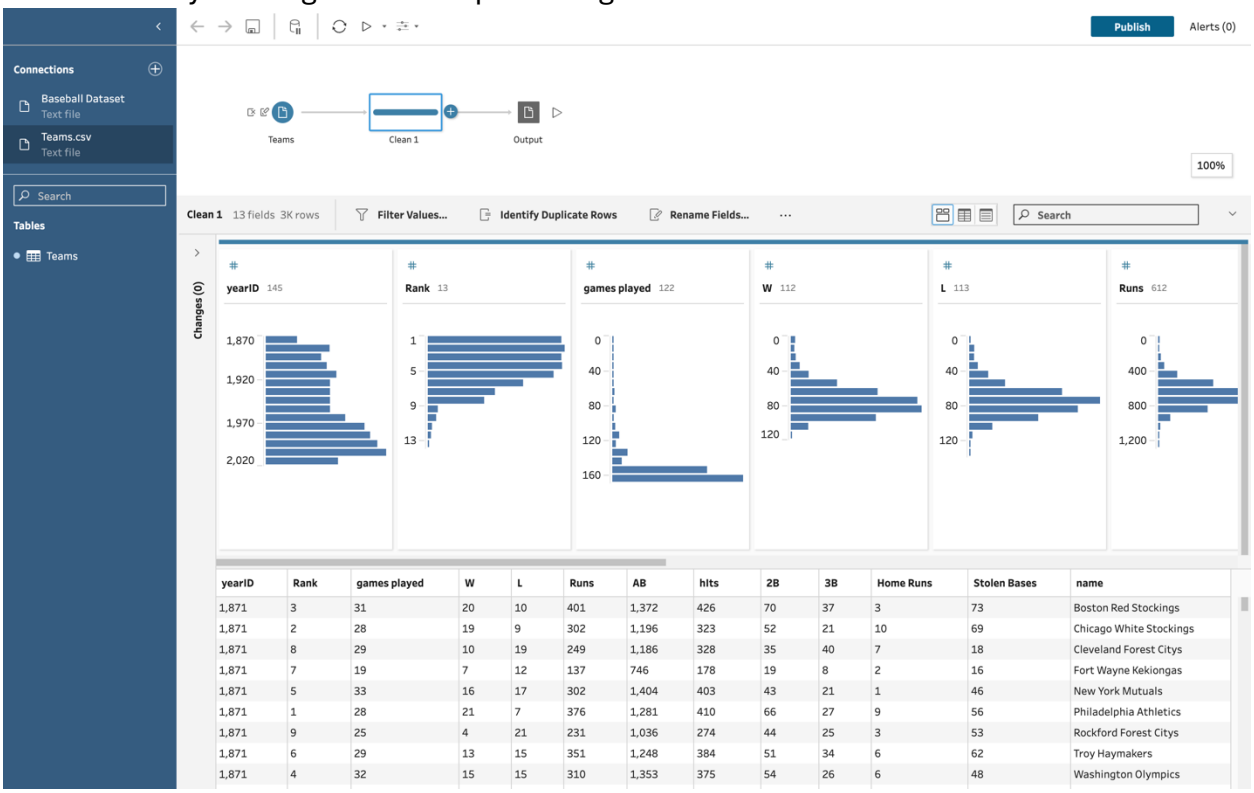




Data Preparation:

There were more than 60 fields in a single area of the Kaggle dataset, which included 12 distinct datasets in total. We didn't use the management, recruiting, etc.; we just used the

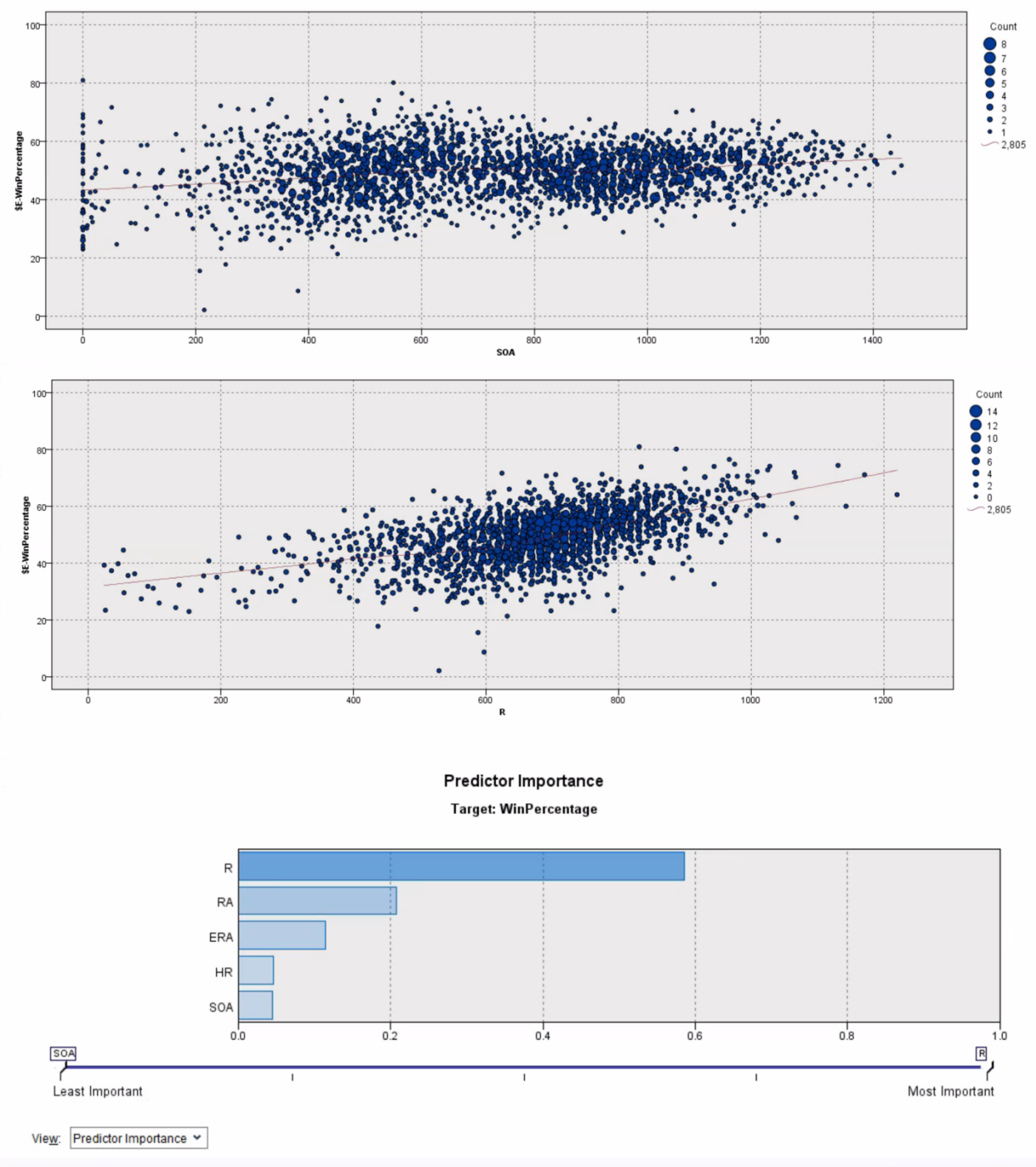
data from each team. Numerous null values, duplicate entries, and data inconsistencies had to be eliminated. We already had the name of the baseball league, so I deleted the team IDs that we didn't need. In our initial visualization, my group identified outliers in the home runs that were won and lost. For the clustering and the predictive models we had to use spss to merge tables from the database together. We also got rid of any data that happened before the allstar game was started. We also created a new field called win percentage that took the amount of wins a team had and divided them by the total games and times it by 100 to get their win percentage.



yearID	Rank	games played	W	L	Runs	AB	hits	2B	3B	Home Runs	Stolen Bases	name
1,871	3	31	20	10	401	1,372	426	70	37	3	73	Boston Red Stockings
1,871	2	28	19	9	302	1,196	323	52	21	10	69	Chicago White Stockings
1,871	8	29	10	19	249	1,186	328	35	40	7	18	Cleveland Forest Citys
1,871	7	19	7	12	137	746	178	19	8	2	16	Fort Wayne Kekiongas
1,871	5	33	16	17	302	1,404	403	43	21	1	46	New York Mutuals
1,871	1	28	21	7	376	1,281	410	66	27	9	56	Philadelphia Athletics
1,871	9	25	4	21	231	1,036	274	44	25	3	53	Rockford Forest Citys
1,871	6	29	13	15	351	1,248	384	51	34	6	62	Troy Haymakers
1,871	4	32	15	15	310	1,353	375	54	26	6	48	Washington Olympics

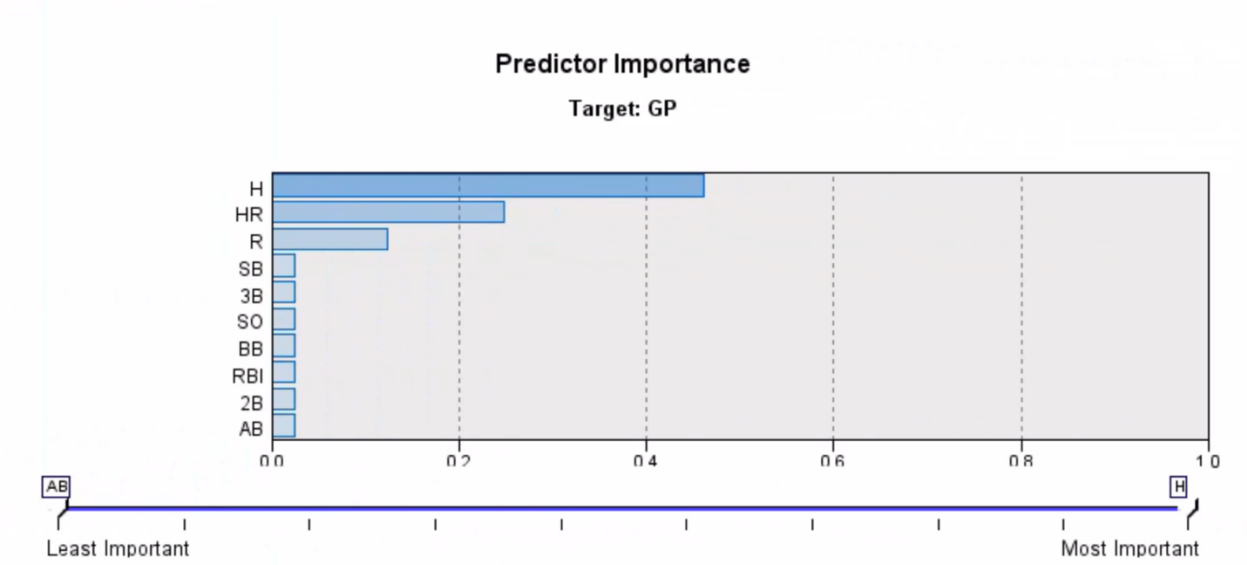
Model Building:

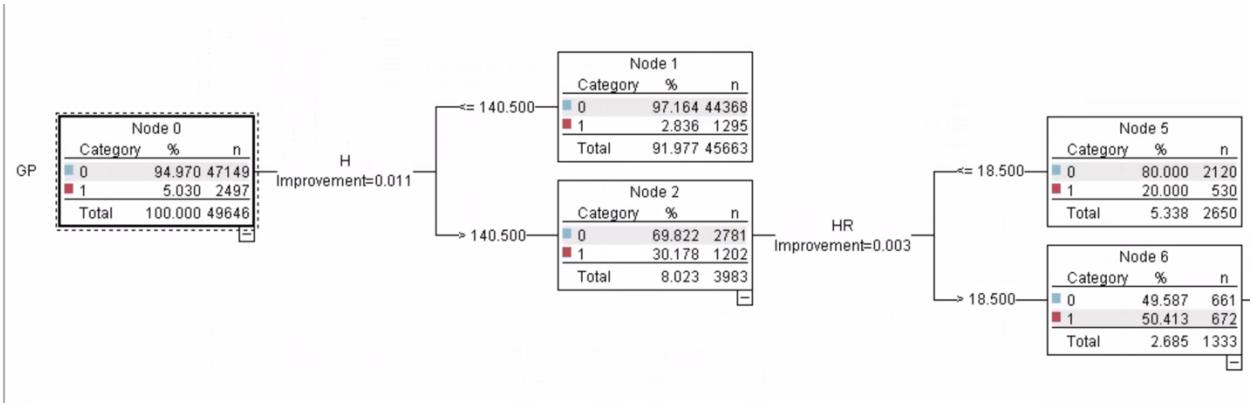
In this analysis, a regression model was built to predict a team's win-loss percentage (W-L%) using several key factors such as runs scored (R), home runs (HR), runs allowed (RA), earned run average (ERA), and strikeouts by pitchers (SOA). The goal of the model was to assess the impact of these variables on a team's overall performance and forecast the win percentage. According to the model runs has the greatest impact on win percentage while strikeouts by pitchers (SOA) had the least impact on our expected win percentage. From the graphs the visualize our regression model we can see the win percentage to strikeout slope is not as big as the slope for win percentage and runs. It is also important to notice that HR didn’t have as big of an impact on win percentage as we previously thought. For teams it might be more important to focus on getting on base instead of trying to hit homeruns.

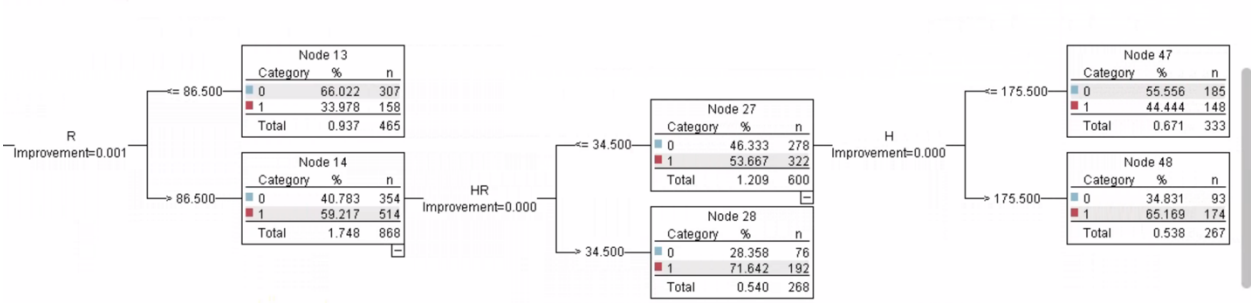


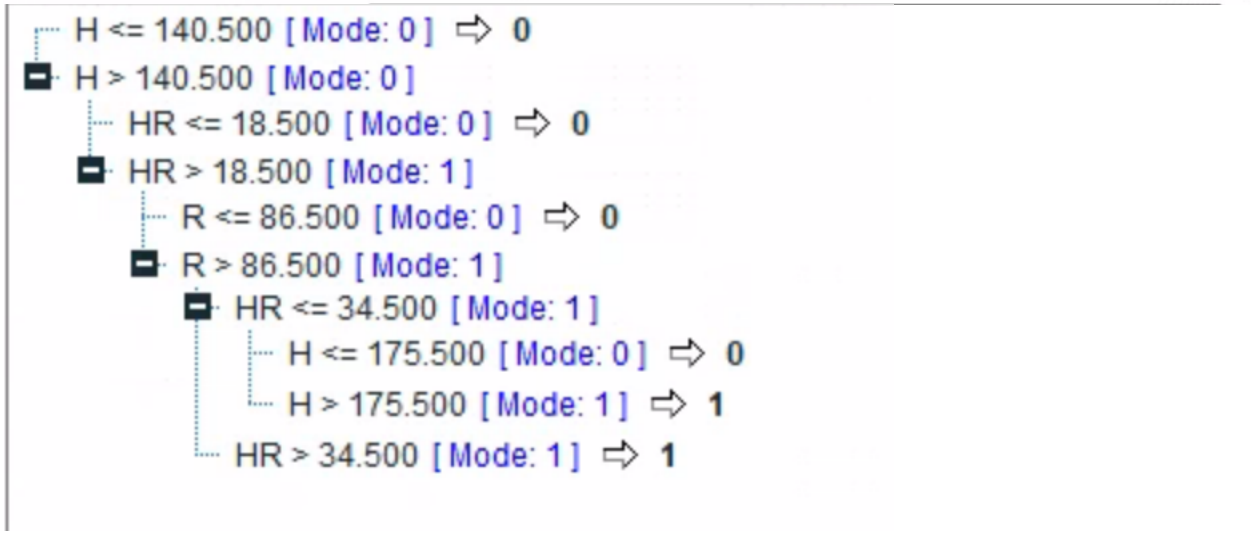
Our second model is a decision tree which we wanted to use to predict which batters would become Allstars based on their season stats. We did this by merging the batters table with the all-star table. We used an outer join which made it so any players that didn't make the all-star game they had a null value in all-star game played and then turned all the null values to 0 using a filler node. We wanted to include all our batting stats into our decision tree and let it decide what the most important stats were. As we can see the most important stats that determine if a player is an all-star is hits, then HR, then runs.

According to our decision tree If a player has fewer than 140.5 hits, the model immediately predicts they will not be an All-Star. Players with more than 140.5 hits and fewer than 18.5 home runs are also predicted to not be All-Stars. Players who hit more than 18.5 home runs and have more than 86.5 runs might be candidates for the All-Star game, depending on their total home runs and hits. We noticed the model was better at predicting if a player wasn't an allstar as opposed to predicting if he was an allstar. We believe this might be because overall in history there have been very little amount of allstars compared to the total player population.






R
Improvement=0.001
<= 86.500
> 86.500
Node 13
Category % n
0 66.022 307
1 33.978 158
Total 0.937 465
Node 14
Category % n
0 40.783 354
1 59.217 514
Total 1.748 868
HR
Improvement=0.000
<= 34.500
> 34.500
Node 27
Category % n
0 46.333 278
1 53.667 322
Total 1.209 600
Node 28
Category % n
0 28.358 76
1 71.642 192
Total 0.540 268
H
Improvement=0.000
<= 175.500
> 175.500
Node 47
Category % n
0 55.556 185
1 44.444 148
Total 0.671 333
Node 48
Category % n
0 34.831 93
1 65.169 174
Total 0.538 267


H <= 140.500 [Mode: 0] ⇨ 0
H > 140.500 [Mode: 0]
HR <= 18.500 [Mode: 0] ⇨ 0
HR > 18.500 [Mode: 1]
R <= 86.500 [Mode: 0] ⇨ 0
R > 86.500 [Mode: 1]
HR <= 34.500 [Mode: 1]
H <= 175.500 [Mode: 0] ⇨ 0
H > 175.500 [Mode: 1] ⇨ 1
HR > 34.500 [Mode: 1] ⇨ 1