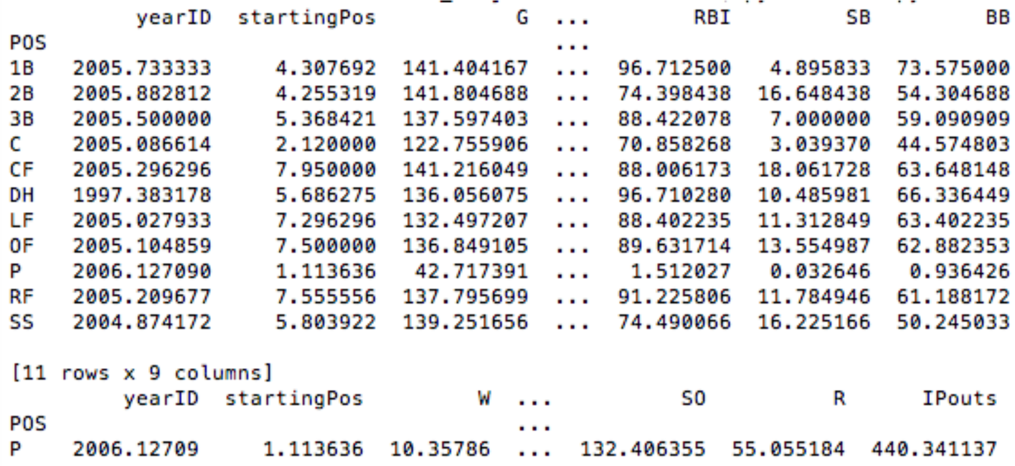
Meleake Wubbie
Professor Hunter Schafer
CSE 163
6 June 2022

MLB All-star Characteristics

By Meleake Wubbie

Summary of Questions and Results

1. What are the average yearly statistics for batters and starting pitchers who were all-stars by position from 1995-2015?



```
         yearID  startingPos           G  ...        RBI         SB         BB
POS                                       ...
1B  2005.733333     4.307692  141.404167  ...  96.712500   4.895833  73.575000
2B  2005.882812     4.255319  141.804688  ...  74.398438  16.648438  54.304688
3B  2005.500000     5.368421  137.597403  ...  88.422078   7.000000  59.090909
C   2005.086614     2.120000  122.755906  ...  70.858268   3.039370  44.574803
CF  2005.296296     7.950000  141.216049  ...  88.006173  18.061728  63.648148
DH  1997.383178     5.686275  136.056075  ...  96.710280  10.485981  66.336449
LF  2005.027933     7.296296  132.497207  ...  88.402235  11.312849  63.402235
OF  2005.104859     7.500000  136.849105  ...  89.631714  13.554987  62.882353
P   2006.127090     1.113636   42.717391  ...   1.512027   0.032646   0.936426
RF  2005.209677     7.555556  137.795699  ...  91.225806  11.784946  61.188172
SS  2004.874172     5.803922  139.251656  ...  74.490066  16.225166  50.245033

[11 rows x 9 columns]
        yearID  startingPos         W  ...          SO          R      IPouts
POS                                    ...
P   2006.12709     1.113636  10.35786  ...  132.406355  55.055184  440.341137
```

   a.

2. Create a model that can predict all-star status.
   a. I used a decision tree regression model to predict all-star status.
   b. I took an all-star from 2018, Nelson Cruz, who is outside of the testing dataset and he was accurately predicted as an all-star
3. What states produced the most all-star players from 1995-2015? How does that compare to birth states of all star players from 1933-1995? Plot the results.
   a. 1933-1995
      i. California
      ii. New York
      iii. Pennsylvania
      iv. Texas
      v. Illinois
   b. 1995-2015
      i. California
      ii. Florida
      iii. Texas
      iv. New York
      v. Georgia

Motivation

I want to explore these questions because I want to know what it means to be an MLB all-star player and make comparisons between different positions. I also want to know the strengths of different positions. Maybe DH's (designated hitters) are ranked high for hitting home runs but aren't ranked high for stealing bases. Maybe CF (center fielders) rank high for stealing bases. My hypothesis is that CF's are the best out of all the positions based on stealing bases. CF's are in the middle of the outfield and have to cover a lot of ground when playing defense. I think this leads to CF's being the fastest on the field. Speed is important when stealing bases and CF's seem the best at doing so. Another motivation for finding all-star characteristics is that some MLB players don't become all stars even though they deserve it. In addition, some MLB players become all stars even though they don't deserve it. If the player is a batter, their batting statistics may show that they are an all-star but they weren't chosen. The same can be said for pitchers who weren't chosen. Whether a player deserves to be an all-star or not is subjective. I want to verify that players who were all-stars deserved it. In addition, I want to see if players who weren't all-stars deserved to be one. Finding out who deserved to be an all-star is important because maybe outside factors influence who is an all-star. Maybe a player is more vocal, has a large social media influence, or has family within the MLB. This can lead to a judging bias which would wrongly award a player all-star status. Being an all-star is pretty important to players. They can lose or make money based on the selection. Their legacies can also be effected. I also want to predict all-star status. Predicting all-star status can be useful for fans and judges of the MLB all-star game. They can use my prediction model to find out who should be an all-star statistically. Fans can also insert a player's stats and see if they are likely to become an all-star. Another motivation for finding all-star characteristics is to see the birth states of all star players from different time periods.

Dataset

The data set I found is called "Baseball Databank" which is found on Kaggle.com. The dataset shows player performance data from 1871 to 2014.

https://www.kaggle.com/datasets/open-source-sports/baseball-databank

Method

Processing the data:

1. Create parent folder called 'CSE163_Project'
2. Create folders for plots and data
3. Download all csv folders from baseball-databank dataset from kaggle.com
4. Place the csv files in the data folder
5. Create functions for loading all-star data, batting data, pitching data, and fielding data (includes positions) using pandas library. **Lines 5-25 in mlb_data.py**

6. Select columns that relate to meaningful stat categories. **Lines 5-25 in mlb_data.py**
7. Join all-star player's data set with batting and pitching data sets to show a complete list of all-star pitchers and all-star batters. **Lines 27-36 in mlb_data.py**
8. Created final all-star batting and all-star pitching datasets with their stats and positions. Merged all-star, batting, and fielding datasets. Then merged all-star, pitching, and fielding datasets. Both data sets have a parameter where a certain year range can be specified. **Lines 39-69 in mlb_data.py**
   a. Having a year range specified will help me answer questions 1, 2, and 3.
9. Functions within these lines have all-star batters with positions dataset and all-star pitchers with positions dataset. 'Abf' dataframe is changed to include parameters to exclude or include positions from all-star batters with positions (abf) data frame. 'Apf' dataframe is changed to have parameters to include a minimum games started condition. **Lines 72-119 in mlb_data.py**
   a. For example, games started over 15 (based on 'GS' column) was used to determine a minimum start condition for question 1.
   b. I don't want to include pitchers who started less than 15 games. In my opinion a pitcher who started less than 15 games is not considered a starting pitcher
   c. I want to exclude pitchers from hitting statistics because in my opinion their hitting stats aren't meaningful
      i. Also having low batting statistics for pitchers might skew my model for question 2.

To answer question 1: What are the average yearly statistics for batters and starting pitchers who were all-stars by position from 1995-2015?

For batters I merged the datasets that contain a complete list of all-stars players, batting data, and fielding data (includes player positions). For pitchers I merged the datasets that contain a complete list of all-star players, pitching data, and fielding data (includes player positions). For pitchers I selected for columns such as 'playerID', 'yearID', 'W', 'L', 'GS', 'ERA', 'SO', 'R', 'IPouts'. For batters I selected for columns such as 'playerID', 'yearID', 'G', 'R', 'H', 'HR', 'RBI', 'SB', 'BB'. A year range can be specified for each function.

Pitching stat categories	**Definition**
'playerID'	Unique player ID
'yearID'	Year
'W'	Wins
'L'	Loses
'GS'	Games Started
'ERA'	Earned Run Average
'SO'	Strikeouts

‘R’	Runs allowed
‘IPouts’	Innings pitch outs
Batting stat categories	Definition
‘playerID’	Unique player ID
‘yearID’	Year
‘G’	Games
‘R’	Runs
‘H’	Hits
‘HR’	Home Runs
‘RBI’	Runs Batted In
‘SB’	Stolen Bases
‘BB’	Walks/Bases on Balls

To answer question 2: Create a model that can predict all-star status,

I decided to create a predictive model for determining all-star status. I decided to go with a decision tree regression model. First, I loaded batters with all-star status from 1995-2015 by merging the batting and fielding datasets. I added an ‘Is_Allstar’ column which is a Boolean value. It is calculated by determining if a player was an all-star or not by searching for each player in the all-star data set. Also, I excluded pitchers within the batting dataset. Before attempting to fit the model to the data, the data was split to an input dataframe and an output dataframe. For the input data frame, the player positions were one-hot encoded. For the input, ‘playerID’, ‘yearID’, and ‘Is_Allstar’ weren’t included. For the output, it is a one-hot encoding of the ‘Is_Allstar’ column.

The input and output data were split into a larger training set and a smaller testing set. It was 80/20 split. I fit the model to the training data using a decision tree regression. The error was calculated by comparing the actual results versus the expected result when making a prediction using a model on the test input data. The same data preparation method was used for the pitching model.

To answer question 3: What states produced the most all-star players from 1995-2015? How does that compare to birth states of all star players from 1933-1995? Plot the results.

I loaded the datasets that contain all the all-stars and all players with personal information including their birth state. I merged the two datasets together in order to associate the personal information of all players with all stars. From that I selected two different ranges of years. One for 1933-1995 and one for 1995-2015. The dataset was reduced to only include all-stars who were born in the United States. It wasn’t relevant to my question to include all-stars who were born outside the United States. I grouped the resulting dataset by the birth state of the all-star players and counted how many were in each group. After, I plotted the result and saved the resulting figure in a png file.

Results

Question 1:

Average yearly statistics for batters and starting pitchers who were all-stars by position from 1995-2015 plots and tables:

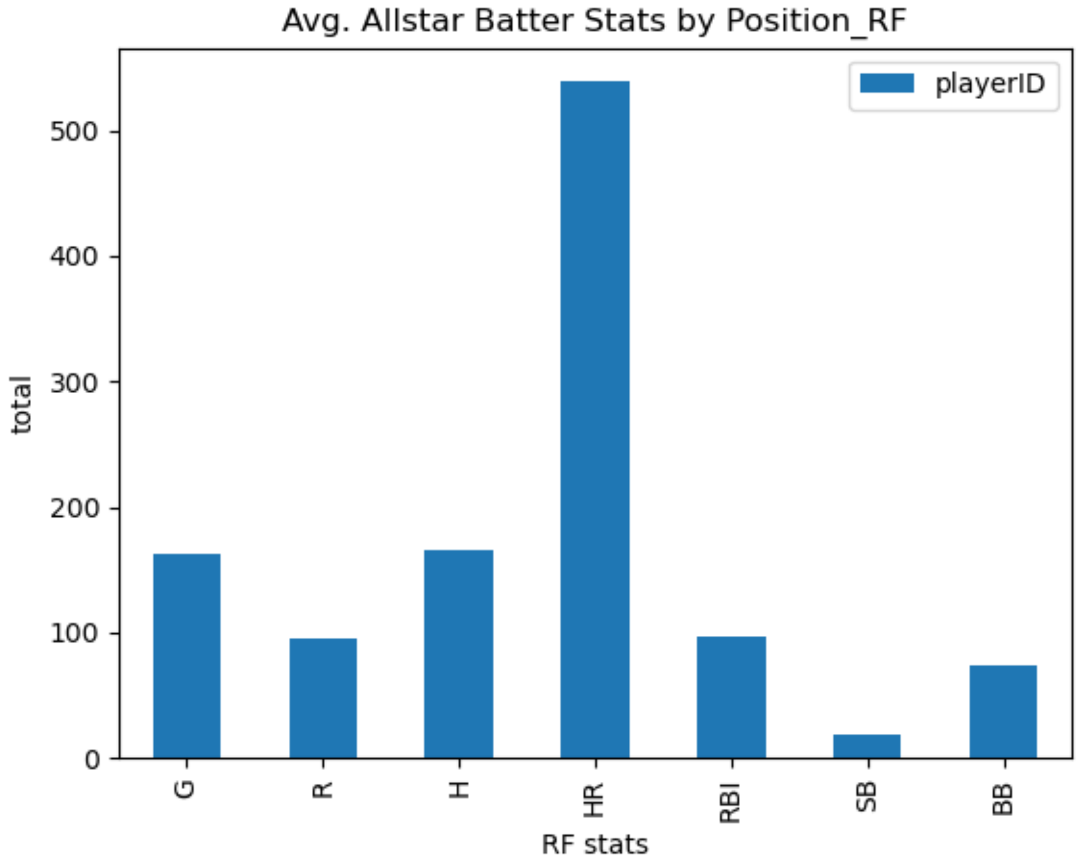
Yearly statistics for batters who were all-stars by position from 1995-2015

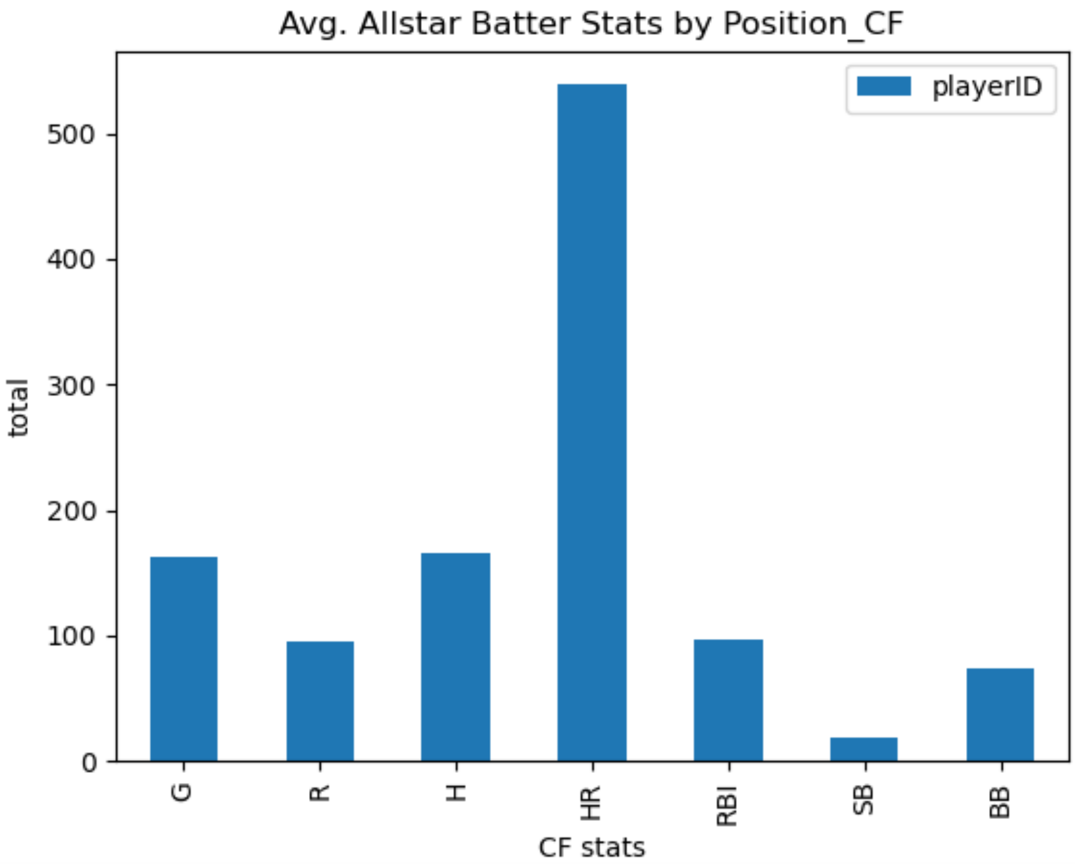
POS	G	R	H	HR	RBI	SB	BB
1B	141	87	156	28	97	5	74
2B	142	90	165	17	74	17	54
3B	138	83	151	24	88	7	59
C	123	62	129	18	71	3	45
CF	141	95	158	26	88	18	64
DH	136	93	157	28	97	10	66
LF	132	86	147	27	88	11	63
OF	137	89	154	27	90	14	63
P	43	1	4	0	2	0	1
RF	138	88	157	26	91	12	61
SS	139	87	162	18	74	16	50

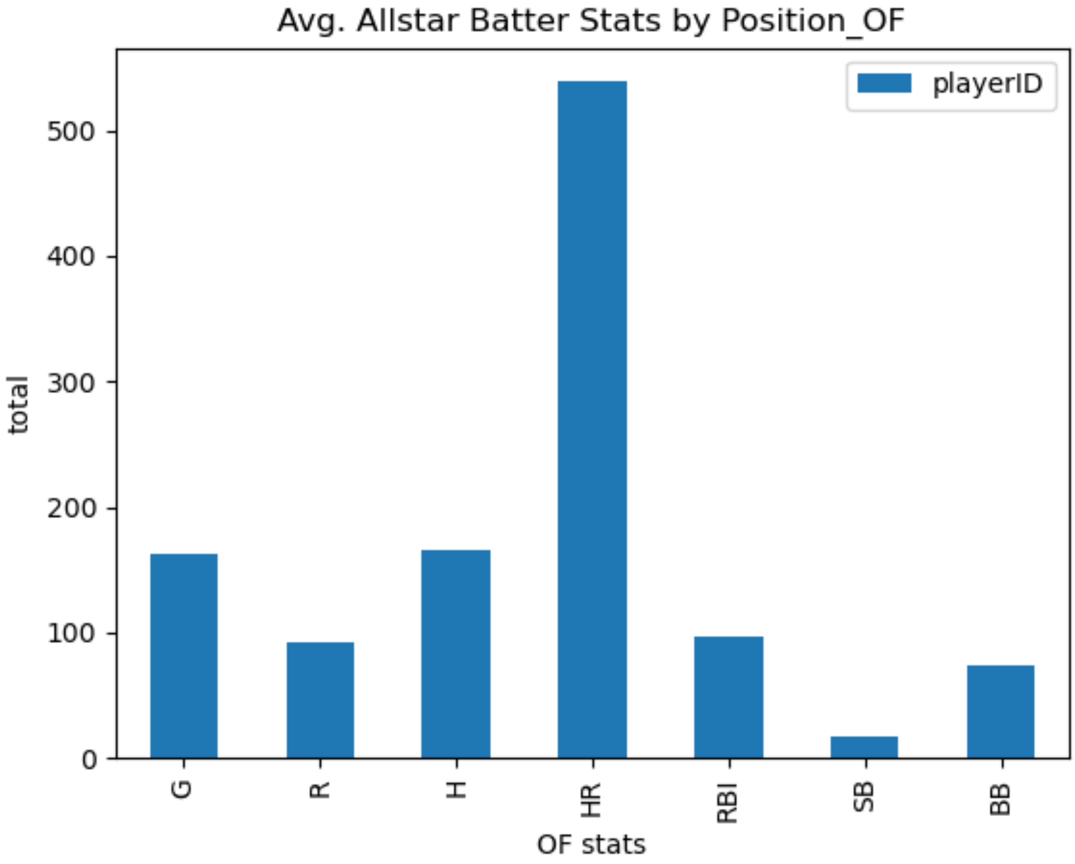
Yearly statistics for pitchers who were all-stars from 1995-2015

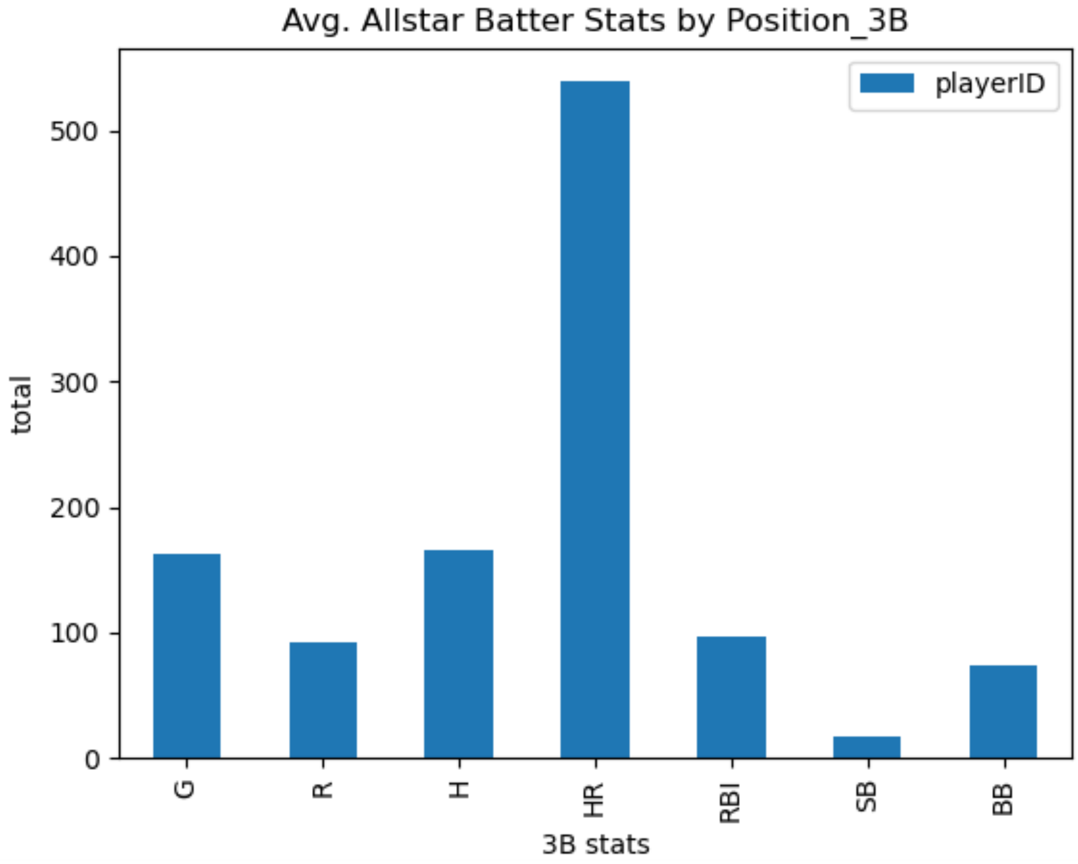
POS	W	L	GS	ERA	SO	R	IPouts
P	10.358	6.137	18.241	2.948	132.406	55.055	440.341

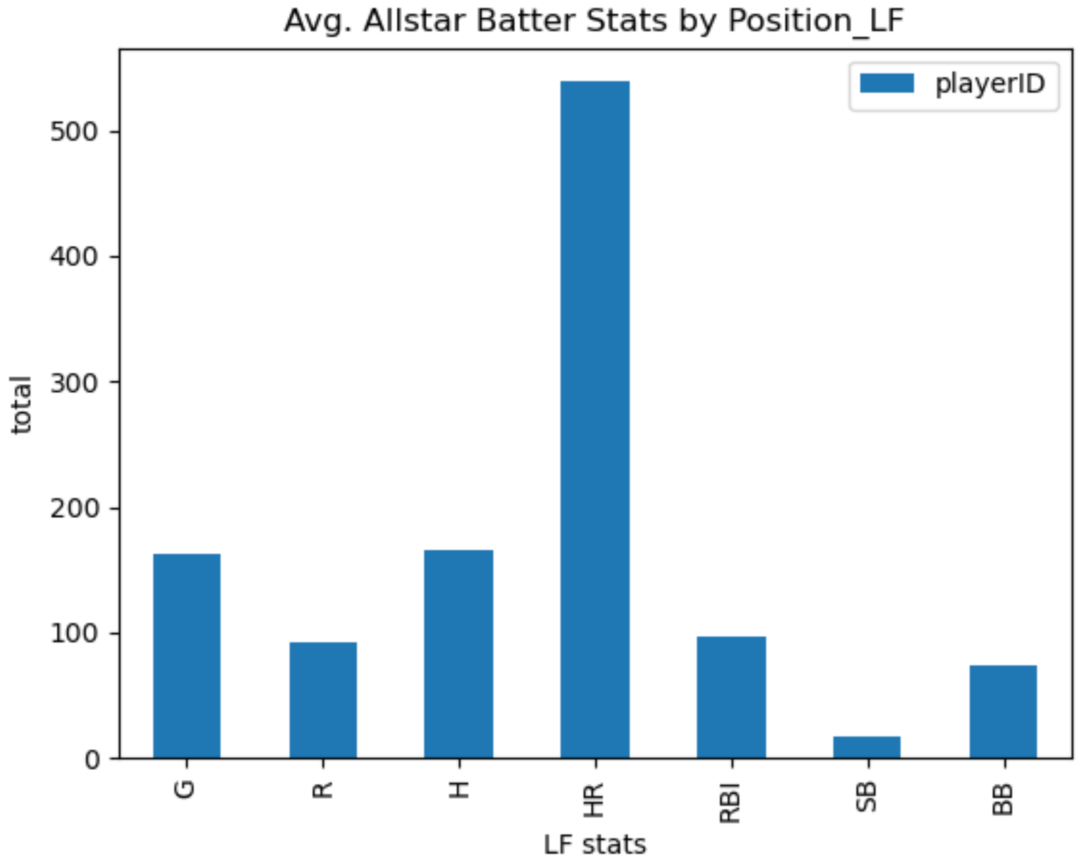
When looking at the batter table I was surprised to see the pitching stats so low. They couldn't even average a home run a year. I think the pitcher hitting data is skewed since there are so many pitchers. Pitchers aren't usually good hitters so having a high number of them in the dataset can bring down the groups average. My hypothesis that center fielders would have the most stolen bases was true. They had the most by a small margin. Center fielders also have the most runs per year for all stars which makes sense since they are usually the fastest on the field. Second basemen had the second most stolen bases. It seems like first basemen are the most patient hitters. They have the most bases on balls. Catchers are lagging in every category which makes sense. Catchers are mainly playing for their defensive capabilities and their ability to communicate with the pitcher. The pitching statistics were as expected. All-star pitchers usually win more than they lose. Now I have a general idea of what players should be all-stars.

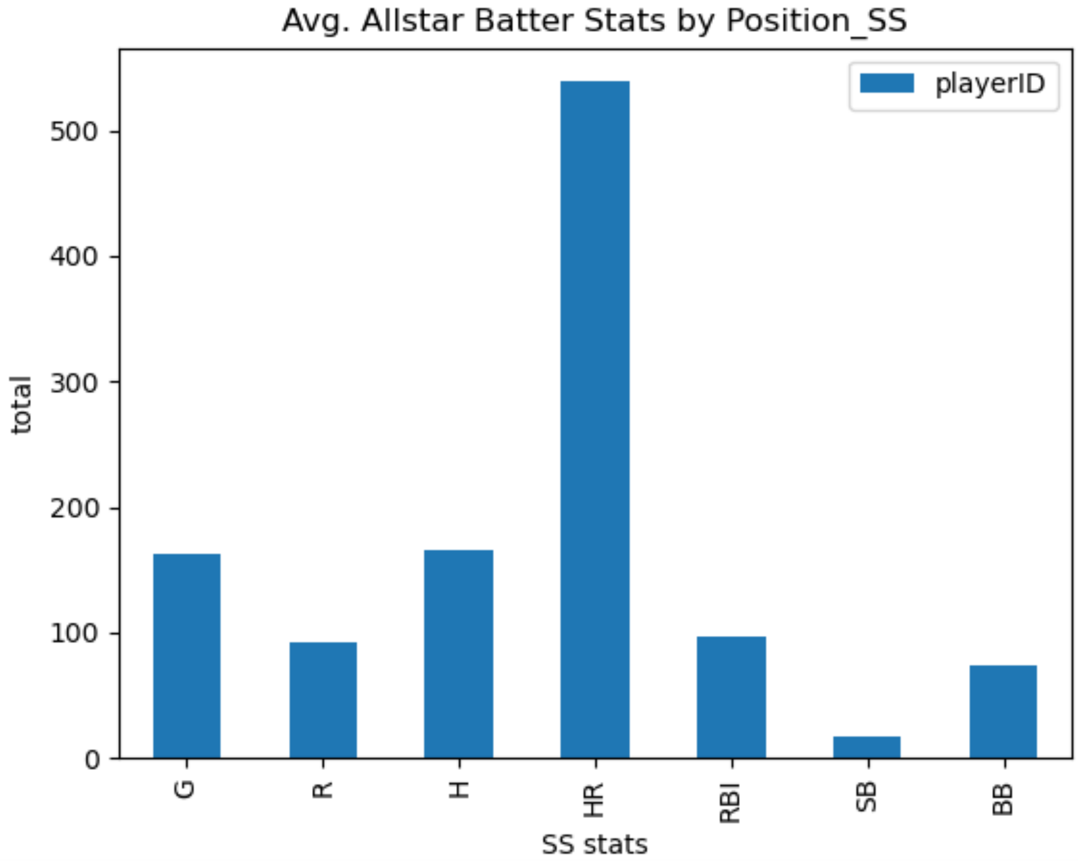

Avg. Allstar Batter Stats by Position_RF
playerID
total
0
100
200
300
400
500
G
R
H
HR
RBI
SB
BB
RF stats

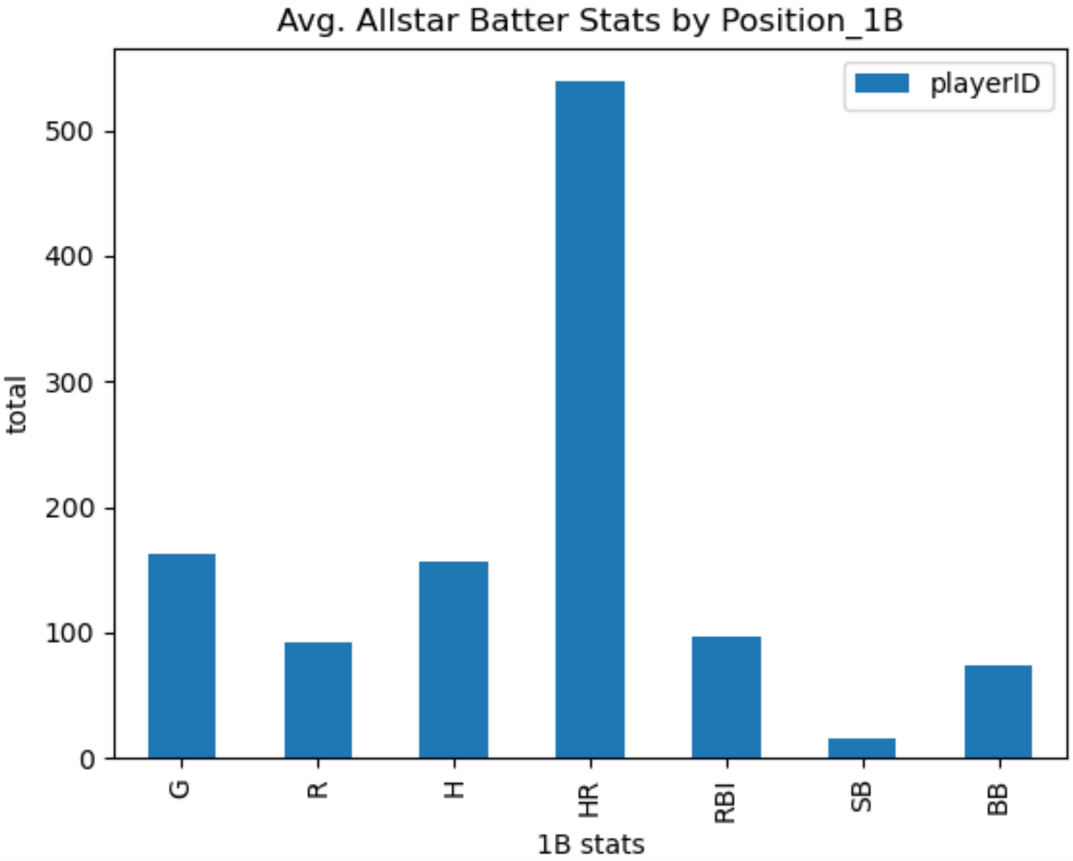

Avg. Allstar Batter Stats by Position_CF
playerID
total
500
400
300
200
100
0
G
R
H
HR
RBI
SB
BB
CF stats

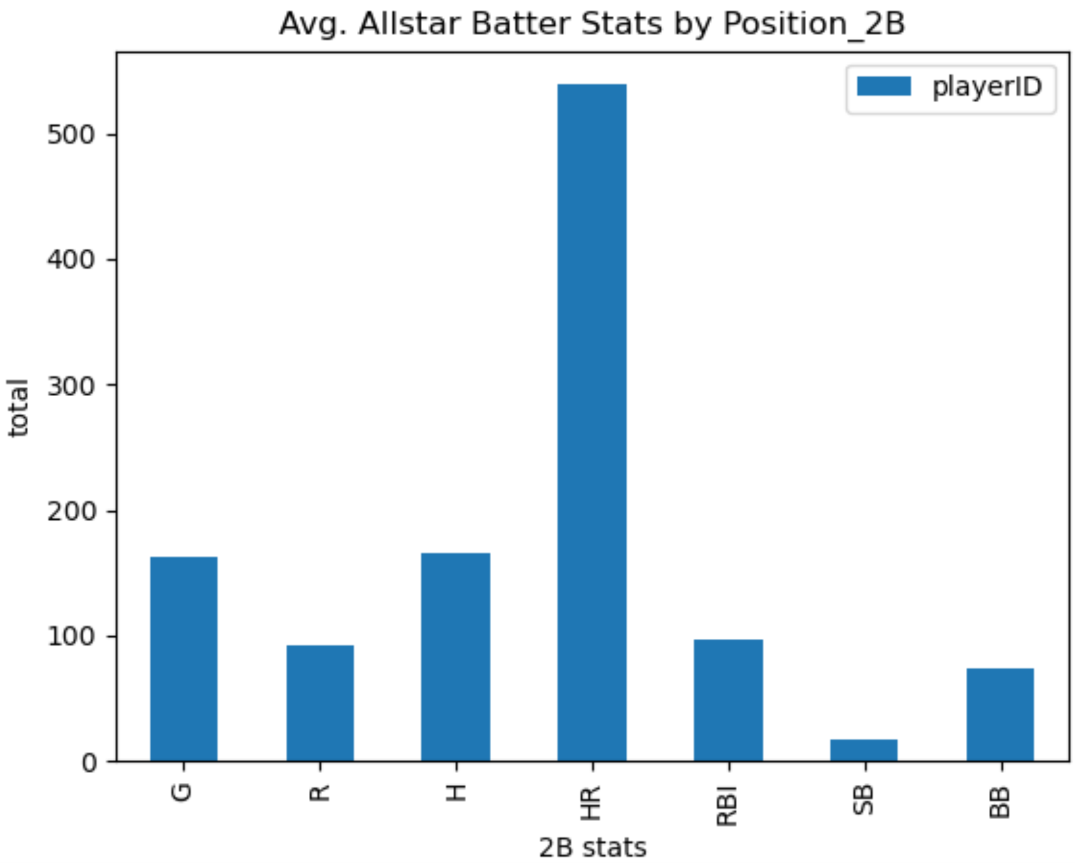

Avg. Allstar Batter Stats by Position_OF
playerID
total
0
100
200
300
400
500
G
R
H
HR
RBI
SB
BB
OF stats

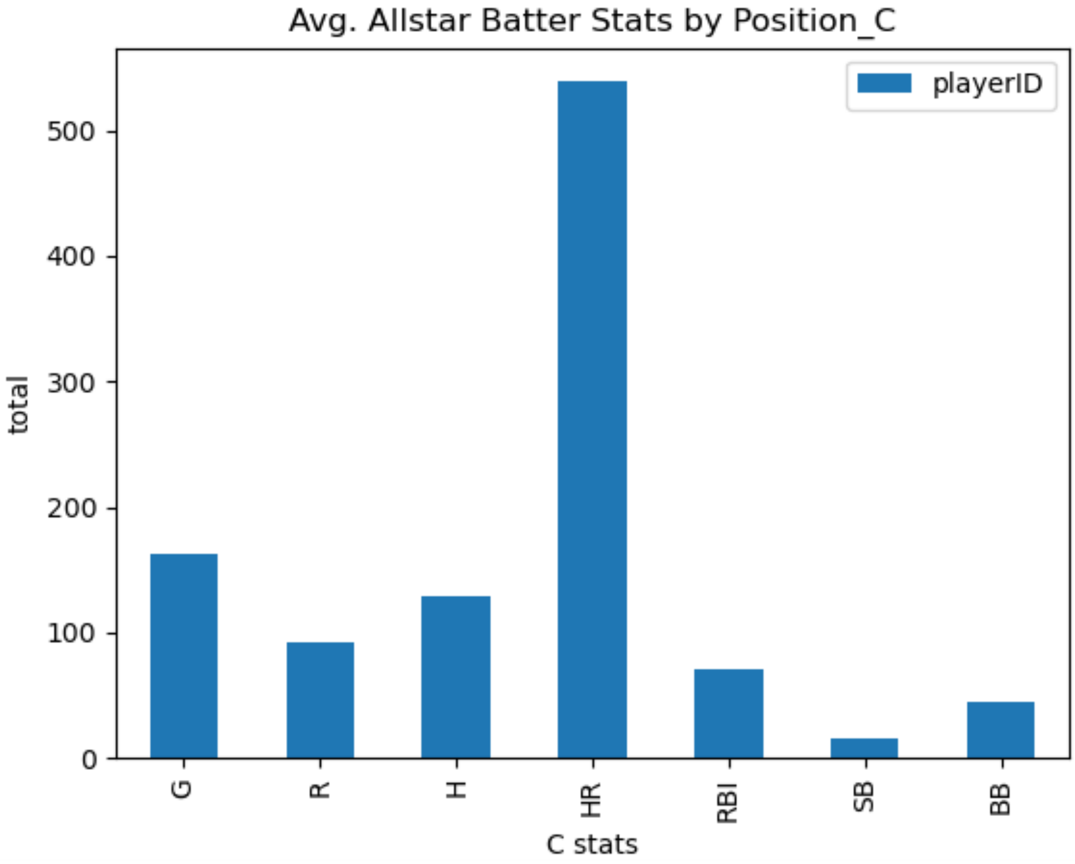

Avg. Allstar Batter Stats by Position_3B
playerID
total
500
400
300
200
100
0
G
R
H
HR
RBI
SB
BB
3B stats

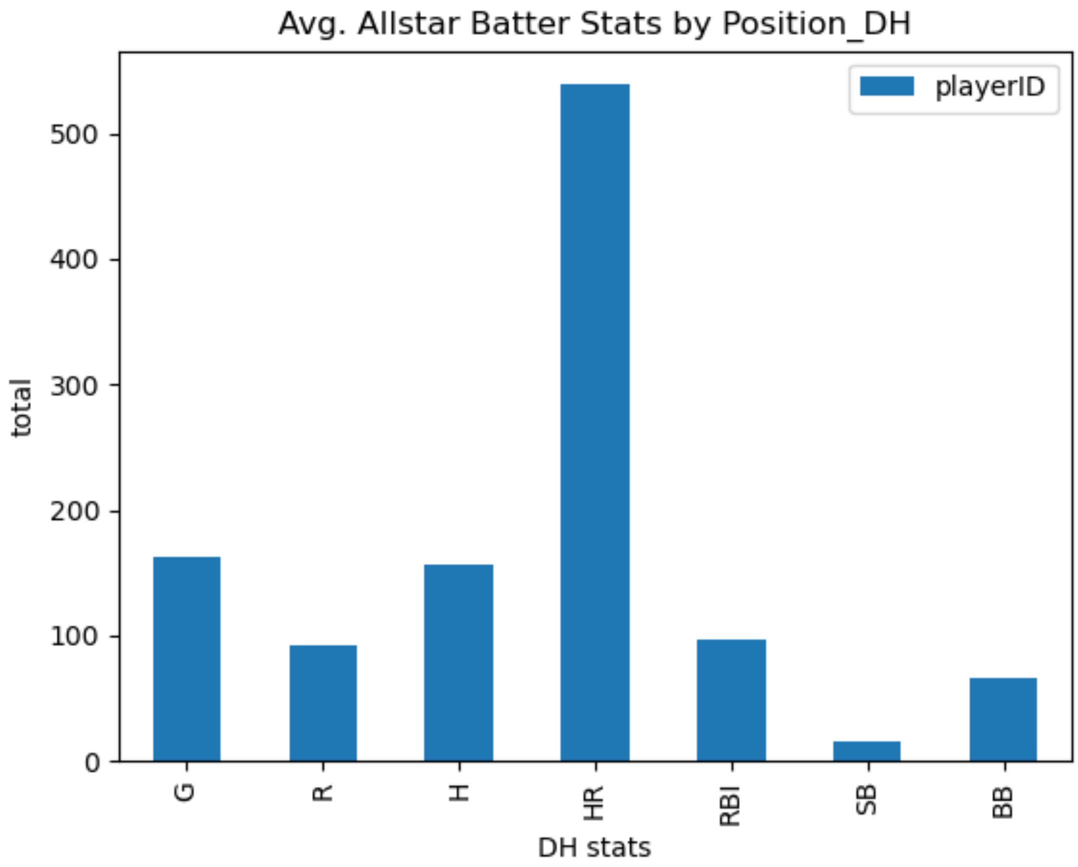

Avg. Allstar Batter Stats by Position_LF
playerID
total
0
100
200
300
400
500
G
R
H
HR
RBI
SB
BB
LF stats

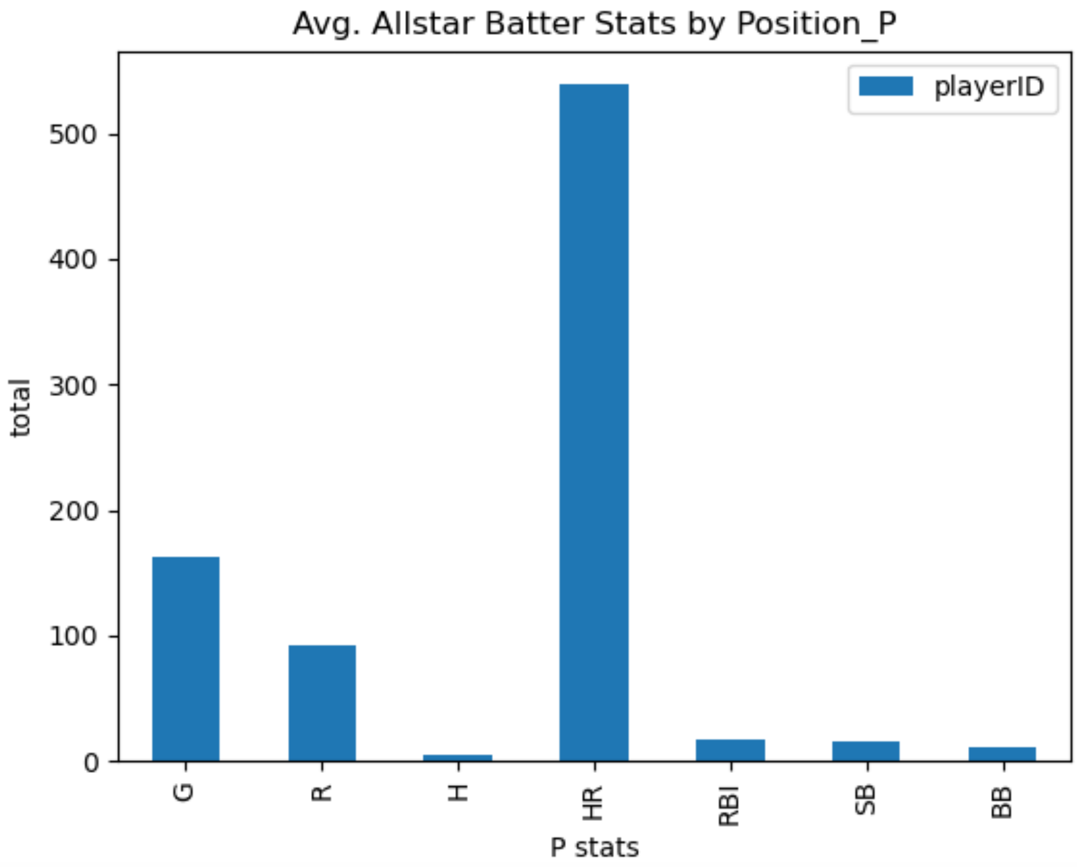

Avg. Allstar Batter Stats by Position_SS
playerID
total
0
100
200
300
400
500
G
R
H
HR
RBI
SB
BB
SS stats


Avg. Allstar Batter Stats by Position_1B
playerID
total
0
100
200
300
400
500
G
R
H
HR
RBI
SB
BB
1B stats


Avg. Allstar Batter Stats by Position_2B
playerID
total
500
400
300
200
100
0
G
R
H
HR
RBI
SB
BB
2B stats

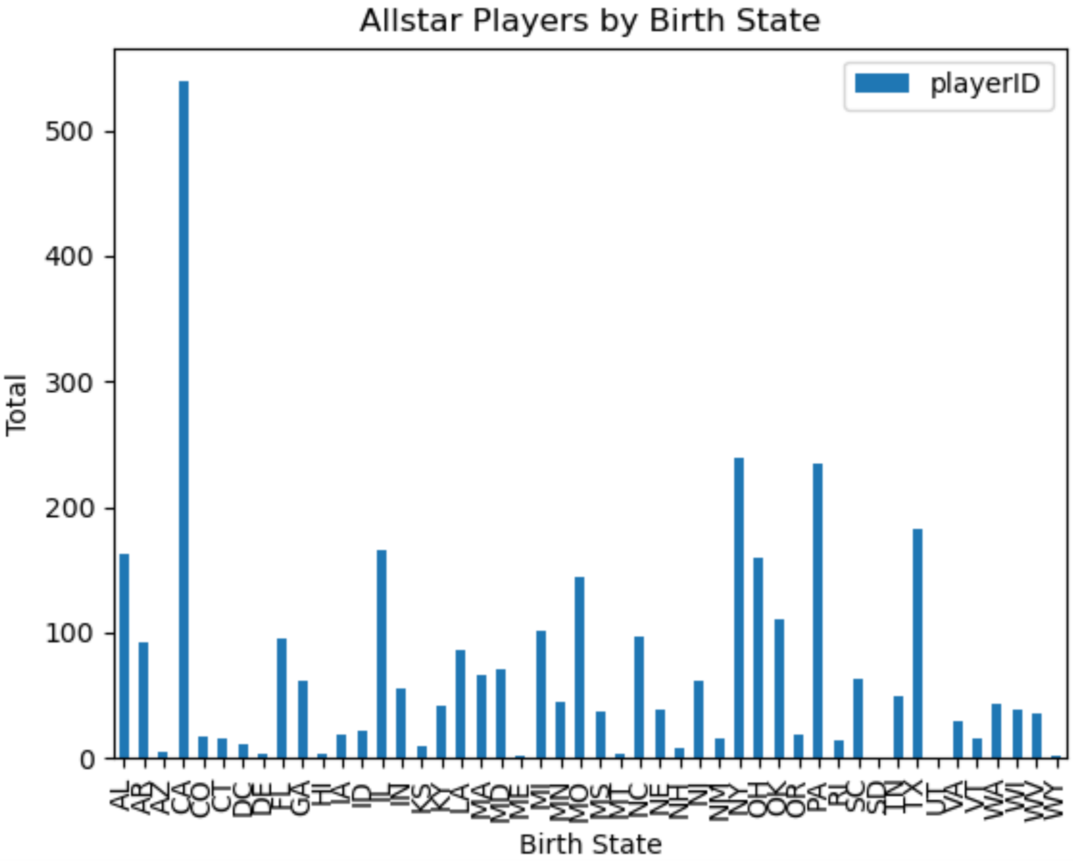

Avg. Allstar Batter Stats by Position_C
playerID
total
0
100
200
300
400
500
G
R
H
HR
RBI
SB
BB
C stats

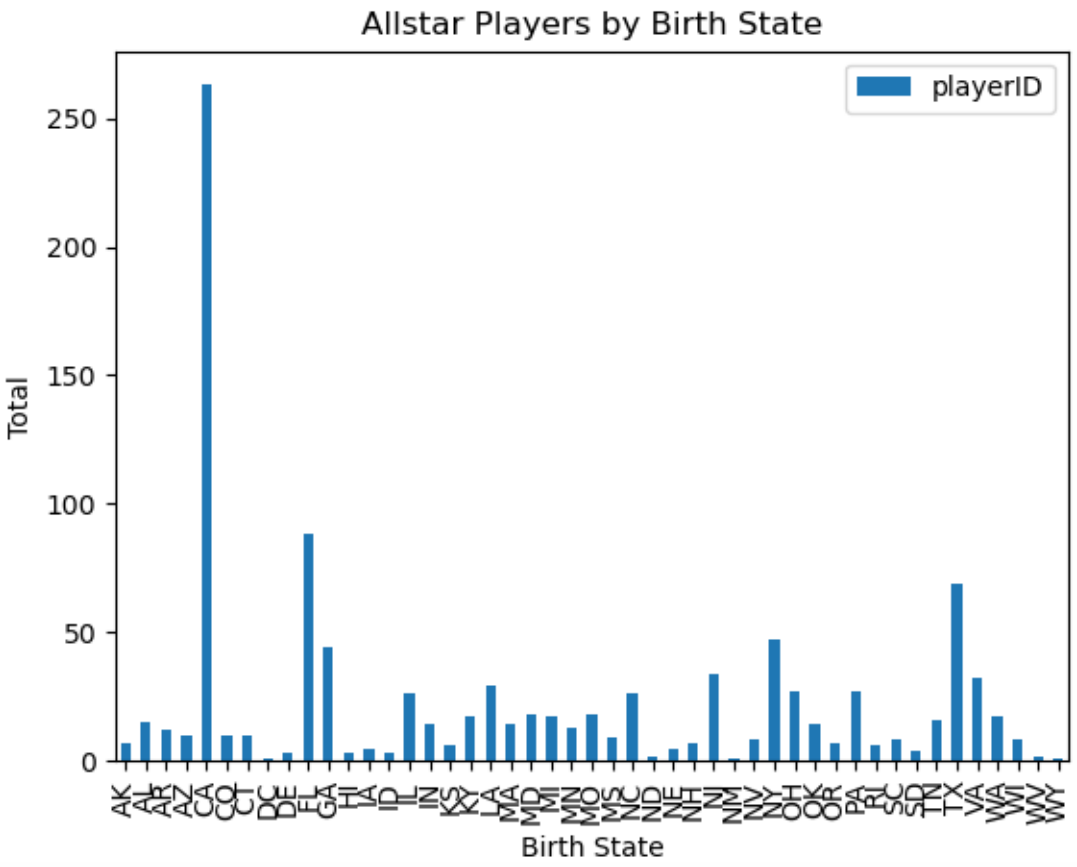

Avg. Allstar Batter Stats by Position_DH
playerID
total
0
100
200
300
400
500
G
R
H
HR
RBI
SB
BB
DH stats



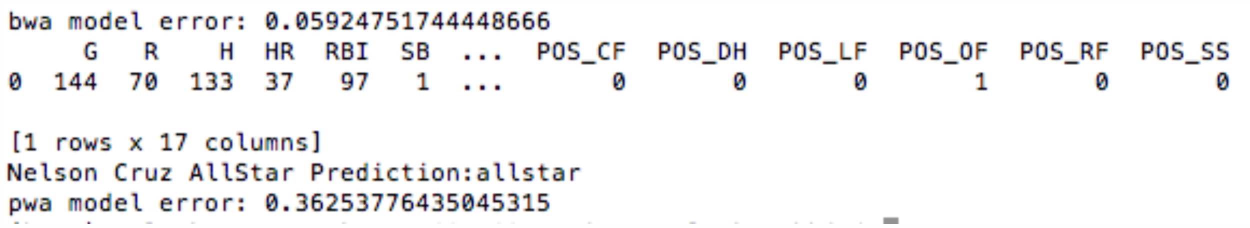
Question 3: What states produced the most all-star players from 1995-2015? How does that compare to birth states of all star players from 1933-1995? Plot the results.

California continues to dominate the production of all-star players. Florida wasn't in the top 5 for 1933-1995 all-star birth state players. In 1995-2015, Florida jumped up to 2nd. On the other hand, Pennsylvania was 3rd for 1933-1995 all-star birth state players. From 1995-2015, Pennsylvania wasn't in the top 10.


Allstar Players by Birth State
playerID
Total
0
100
200
300
400
500
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
NE
NH
NJ
NM
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY
Birth State



Question 2: Create a model that can predict all-star status.



```
bwa model error: 0.05924751744448666
     G   R    H  HR  RBI  SB  ...  POS_CF  POS_DH  POS_LF  POS_OF  POS_RF  POS_SS
0  144  70  133  37   97   1  ...       0       0       0       1       0       0

[1 rows x 17 columns]
Nelson Cruz AllStar Prediction:allstar
pwa model error: 0.36253776435045315
```

The all-star batting model I created can accurately predict whether a batter should be an all-star or not based on their hitting stats and their position. I tested the model by inputting a data frame with stats from a 2018 all-star, Nelson Cruz and it accurately predicted that he was an all-star. The model was just over 94% accurate. After training the all-star pitching model, it proved to be too inaccurate for use in prediction. The model was 64% accurate.

Impact and Limitations

One limitation of my analysis is not considering defensive stats for batters. I only account for hitting stats and not defending. Some players may be valued higher or lower based on their defensive capabilities. One implication of my results is that MLB players can look at the averages for all-star players for their position. They can keep these results in mind to know what it takes for them to be selected to an all-star game. There may be biases that effect my results

Challenge Goals

My project met the multiple datasets goal because I combined different datasets. I used different datasets within the baseball databank.

https://www.kaggle.com/datasets/open-source-sports/baseball-databank?select=Master.csv

I also met the machine learning goal because I performed one-hot encoding. I applied the all-star batter model to predict the all-star status of Nelson Cruz. My goal was to predict all-star status and that proved to be accurate when preforming a prediction on Nelson Cruz's 2018 stats.

I met the result validity because when I was training the model for batters I trained and tested the model accuracy by testing on a reserved subset of the data. Also, when I was training the model for pitchers it had a high error rate which showed that the model was not able to make accurate predictions. Therefore, demonstrating that the data was not able produce a valid prediction model.

Work Plan Evaluation

My work plan estimate was generally accurate. My work plan didn't account for assignments within the class other than the project like checkpoints and resubmissions. Also my work plan didn't consider work load in other classes. I had limited time to complete work for each week but it all ended up working out well.

Testing

I wrote tests for the utility functions. There's a test for preparing players for decision tree regression to ensure that data passed into the model was properly formatted. The test for test_select_by_year_range to make sure the function can be able to accurately select the rows for the years within the given range. There is a assert statement for test_select_by_year_range to make sure the resulting data frame matches the expected data frame. The test_players_for_dtr ensures that the actual input and output data frame and expected output and input data frame are equivalent.

Collaboration

No other people or resources consulted other than course staff