REPORT

We had large amount of datas, and our goal was to find the answer to our question. In order to do this, i selected a question, and tried to find an answer to my question using pandas module and i did my work on jupyter notebook. Since the data was too big i had to select a very key question in order to do mu calculations. My question was simple. Since there is two different leagues in the history of baseball (learned after searching the datas), -al and nl leagues- they had their own all-star games. So my question was which league has the more players that played in that all-star games ? I used the data named AllstarFull.csv. In the data, if the player played in all-star game, it showed as 1.0, and if not it was 0.0. Also there are several columns in my database that should be explanied. My column explanitaion is like following; #playerID : Player ID code yearID : Year

gameNum : Game number (zero if only one All-Star game played that season)

gameID : Retrosheet ID for the game idea

TeamID : Team name

lgID : League Name

GP : Games played (1.0 if played in the game)

startingPos : If the player was started, its position

Means, i want to deal with lgID and GP in order the get the values that i wanted (which league had the most players played in the all-star game). I calculated whole of the players that has been played in all-star game, without including their own league (whether the player is in nl league or al league). I get the result

```
1.0    3990
0.0    1139
dtype: int64
```

Meaning, there are 3990 players that played (1.0) in all-star game, and 1139 players that not played(0.0) in all-star game.

```
al_played = []
al_notplayed = []
```

i created two empty lists here for al league in order to get the total amount of players whether they played or not in the all star game. I also did the same procedure in nl league.

```
nl_played = []
nl_notplayed = []
```

And this loop, helped me to go through every element and returning the result that i want (adding the values into a list regarding whether they played or not).

```
for i in al.GP:
    if i == 1.0:
        al_played.append(i)
    if i == 0.0:
        al_notplayed.append(i)

for i in nl.GP:
    if i == 1.0:
        nl_played.append(i)
    if i == 0.0:
        nl_notplayed.append(i)
```

So bacically, i add my values into lists so if i just print len(list) i get the total amount of players regarding the list (whether they played or not corresponding to its own league.)

```
al_played = []
al_notplayed = []
nl_played = []
nl_notplayed = []
```

A basically created 4 lists here in order to get the results that i want. If i basically print len of these graphs i will get the total players correspondingly whether they played or not, or in which league that they played. In order to do this, i used these for loops:

```
for i in nl.GP:
    if i == 1.0:
        nl_played.append(i)
    if i == 0.0:
        nl_notplayed.append(i)

for i in al.GP:
    if i == 1.0:
        al_played.append(i)
    if i == 0.0:
        al_notplayed.append(i)
```

After this for loop i got lots of results inside those lists but i wont deal with them since i only want len of those. In order to do this, i also created another variables for doing my calculations in much more easier way;

```
alPlayed = len(al_played)
alNotPlayed = len(al_notplayed)
nlPlayed = len(nl_played) , nlNotPlayed = len(nl_notplayed)
```

So if i print them one by one, i get the results that i want;

#Number of players that played in AL league all-star game:

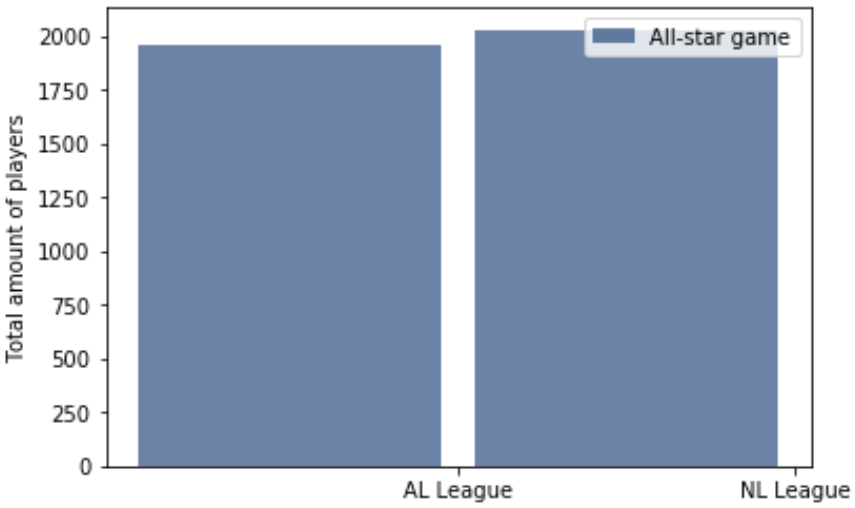
Print len(al_played)

This will print the players played in the allstar game.

print len(al_notplayed)

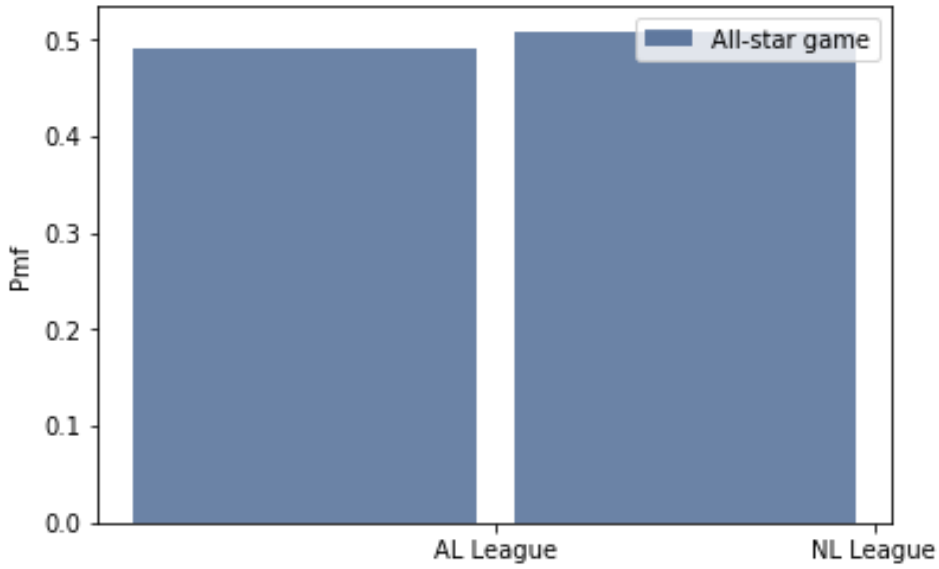
This will also print the count of the players that not played in the all star game

To visualize it, i used Histogram, PDF and CDF. Its Histogram is like in the below:



As it can easily be seen, NL League slightly more players that played in all-star games. Allthough the difference between them is not very much, i found answer to my question with using this graph easily.

Also there is another way to plot my results/data. Here is the PDF plot of my data:

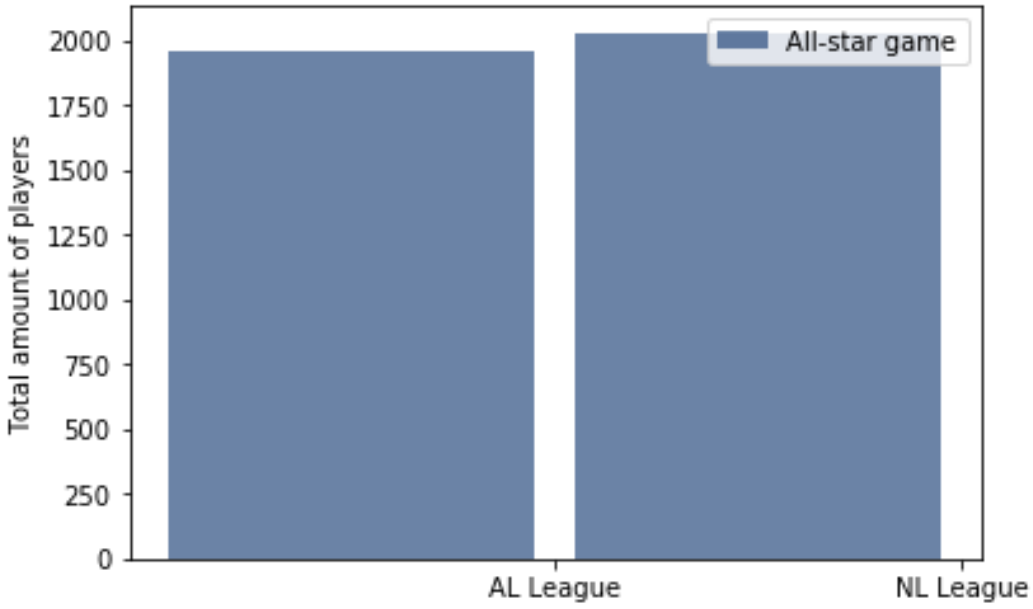


Since we are dealing with probabilties in here, the values of ylabel (PMF) will be show difference, but overall, as a graph there wont be any difference in looking.

Conclusion

As conclusion, i tried to find answer to my question. My question was Which league has more players played in its own all star game ?

So after plotting my histogram for this situation;



As can be easily seen from here, i can say that NL league has more players played in its own league. But for making things more clear i also want to see the results from my code. If i print al_played and nl_played one by one, regarding to this histogram, nl_played must be greater than al_played.

print len(al_played)

>>>1961

print len(nl_played)

>>>2029

As can be seen easily from here. My histogram came up with correct results and overall, with looking to my histogram and code i can say that everything is accurate.