Michael Puiszis
Data Science
Baseball Offensive Leaders Update

The topic of baseball statistics is an interesting one because of the vast amounts of data that exist, for every game, team, player and arena. Because of this, it has been heavily studied and analyzed for a number of years now, so much so that I almost feel like I'm beating a dead horse in trying to come up with my own for of analytics. But nonetheless, I will be attempting to show that there exists significant enough data to say with a large amount of confidence who, over a given period of time (in this case between 2010-2015) who the best offensive players are. Also, if possible, I would like to see if a certain position in the field has consistently better offensive players than the rest.

So far, the work that I have done includes getting a large data set that consisted of close to 1.7 million points of data (about 89,000 rows of data, with 20 fields each), over a time span that is closing in on 120 years. To begin to try and wrestle this data down to a workable level, I removed all of the data fields that were before the 2010, which chopped the data down by a factor of 10 (down to about 9000 rows with 20 rows each). From there I removed players that had less than 75 games played (G) and 150 at bats (AB) in a given season. Doing so brought the data down to just over 2,000 rows. Further pockets of player data were removed if they did not meet certain minimum levels including a batting average (BA - calculated by taking the total number of hits in a season divided by the number of AB) of at least .200, less than 10 Runs Batted In (RBI's – Total number of runs that scored as a result of an individuals actions while at the plate), and less than 33 total hits. Unfortunately, this only brought the data down by a few hundred rows, (now at

1796), and I did not want to make the criteria higher, in case I might lose significant pieces of data for players that might not have met a certain minimum, while excelling in other areas. As for the remaining data, there existed columns of data that were unnecessary for any of the calculations that I plan to perform, such as the number of times the player grounded into a double-play (GIDP), number of times caught stealing (CS), and the number of sacrifice bunts (SH) that did not result in an RBI.

There were also specific pieces of data that did not exist within the data set that I needed, and could be easily calculated given the data that I already had. These included slugging percentage {(1 * # of Singles + 2 * # of doubles + 3 * # of triples + 4 * # of Home Runs)/AB}, On-base percentage (OBP – # of hits + # of walks + # of times hit by pitch (HBP) all divided by AB), and OPS (On-base plus slugging).

Once these stats had been calculated and added to the data set, I began partitioning individual stats into manageable sizes that consisted of the top 50 individual performances for that particular stat (e.g. top 50 Home run seasons, that consisted of the player, their team, and their season total of home runs). Once this had been done I began reading in the .csv files into a Jupyter Notebook to begin working with and visualizing various data fields. Using the Seaborn library's strip-plot I was able to quickly make a few graphs showing the players with the highest number of home runs, runs batted in, total hits, and run scored in a given year. Within almost all of the visualizations, there exists multiple data points for certain players, which is simply the result of them having multiple years in which they were in the top 50 players in a specific category. It is with these players, the ones that

performed extremely well over the entire time span that I will be focusing my efforts going forward. Beyond the individuals in the top 50 of a given category, there are certainly other individuals that performed consistently well that may be just outside of this threshold, as such I plan to expand the specific data frames to see if there are other players that are not necessarily at the top of a given category, but are consistently near the top for a number of categories, in order to give proper consideration to these individuals as well. I am doing so because if playing the sport for over a decade and watching it for two have shown me anything, it is that consistency is the most important attribute that a baseball player can have.

Beyond the stats that I currently have, there are others out there that still may need to be calculated or gathered if calculations become impossible for one reason or another. Positional information will need to be gathered for the top performing players. Also, stats such as wins above replacement (WAR), Linear Weights, Runs Created, and Base Runs will need to be calculated or taken from various websites in order to get a more accurate set of data. These stats general take into consideration situational aspects of the game, such as number of outs, or number of people on base when the player is up to bat. Because all of this information would severely inflate the already large data set, they will have to be pulled individually as well.

I hope to be able to utilize certain multi-dimensional visualization tools as I get further into the project; so far in researching them I have come across Mayavi and SVG as possible avenues to try. Ideally I would like to find a library that would allow me to utilize a multi-dimensional star plot, which would have the end points

of the star be further from the center for players that are ranked highly in a specific

category. This would ultimately be the tool final tool that I would use to decide on

the best player.

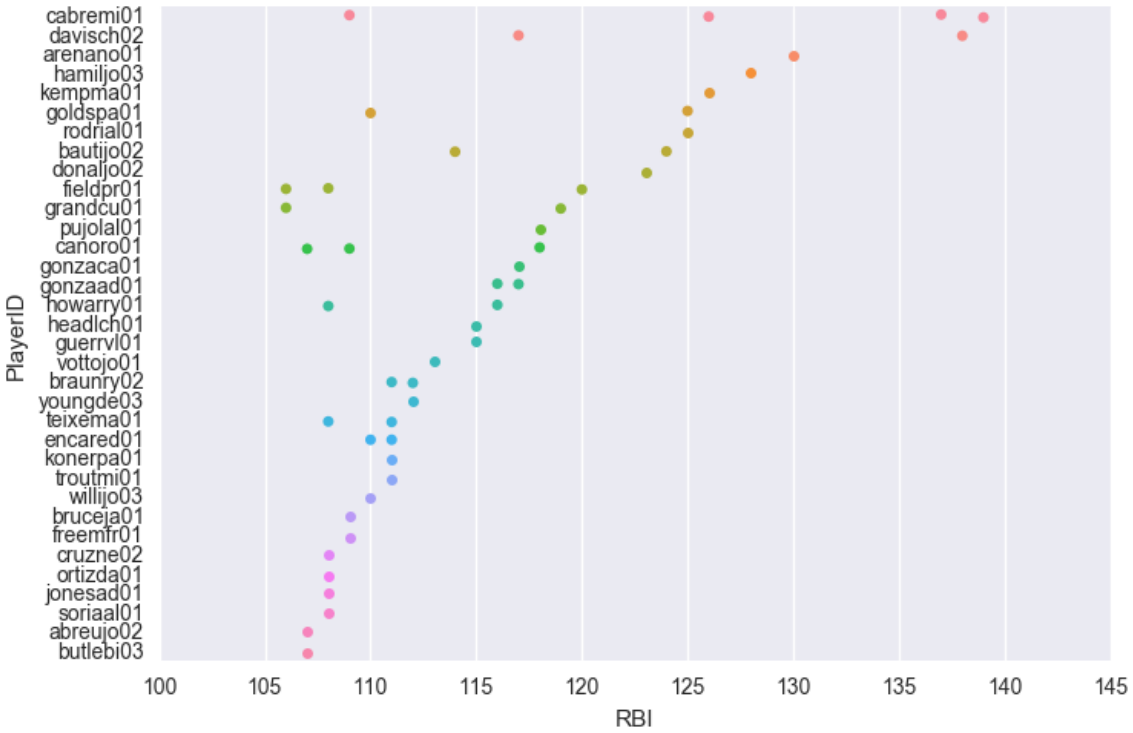

Figure 1: Top 50 RBI leaders over between 2010-2014. Multiple top 50 appearances have a corresponding number of points in a given row.

```
#Getting some more information on the main data set
df.describe()
```

G	AB	R	H	Doubles	Triples	HR	RBI	SB	SO	Walks HBP
1795.000000	1795.000000	1795.000000	1795.000000	1795.000000	1795.000000	1795.000000	1795.000000	1795.000000	1795.000000	1795.000000
123.222841	420.872981	55.205571	112.045682	22.132033	2.349861	12.814485	52.784401	7.996100	85.460724	45.869638
25.600381	130.492513	23.236930	40.428352	9.534240	2.438118	9.242648	24.984873	9.889776	34.534022	23.915723
76.000000	152.000000	10.000000	33.000000	1.000000	0.000000	0.000000	5.000000	0.000000	17.000000	4.000000
101.000000	315.000000	37.000000	80.000000	15.000000	1.000000	6.000000	33.000000	1.000000	59.000000	28.500000
126.000000	423.000000	53.000000	110.000000	21.000000	2.000000	11.000000	49.000000	4.000000	82.000000	41.000000
147.000000	534.500000	72.000000	144.000000	29.000000	3.000000	18.000000	69.500000	11.000000	106.000000	59.000000
162.000000	684.000000	136.000000	225.000000	55.000000	16.000000	54.000000	139.000000	68.000000	222.000000	163.000000

Figure 2: Summary Statistics for the main data set, other stats not shown include: AVG, OBP, OPS, SLG.