Michael Puiszis
Data Science
Project Topic Paper

Baseball has been and probably always will be a great interest to myself, stemming from playing and watching the sport for years. So selecting the topic of analyzing the statistics of the leagues top hitters and selecting an individual or group of individuals was a no brainer for me. Continuing the research and analysis started earlier in the semester, which consisted of simply mapping out and visualizing the data, has evolved into analysis of coexisting statistics with the goal of revealing some statistically significant insight into the various relations behind certain players offensive output.

After loading up the data set and gaining a solid statistical basis of the data, while also mapping and visualizing that data along side other relevant statistical fields, I began looking for correlations between certain statistical fields. To do this, I cut down the rather large data set such that only the top performers in each statistical category remained; so that the amount of data I was working with could give an accurate representation of the data set, while keeping the visualizations readable. After doing this I began simple regression analysis between pairs of fields, which yielded some interesting results.

The majority of investigation I did consisted of seeing how a player's batting average and slugging percentages affected other statistical season totals for the top performing players in each category. From initial analysis, there appears to be a significant relationship between a batter's slugging percentage and the number of RBI's that they got in a given season, with the added caveat that these results were only found among the top performing players (those with batting averages above

.320), meaning that one would be remised to apply these findings to the entire data set. Next I looked at how the number of strikeouts a player had in a given year affected their batting averages and slugging percentages. As one would expect, an increase in the number of strikeouts negatively impacted the player's batting average. A surprising finding was that the number of strikeouts appeared to have a positive impact on a player's slugging percentage, which seems to go against conventional logic. This can probably be attributed to a couple of things. Power hitters (individuals with higher slugging percentages) tend to strike out more, and the benefits for the slugging percentage from getting extra base hits (doubles, triples, and home runs) out weigh the negative impact that strikeouts have.

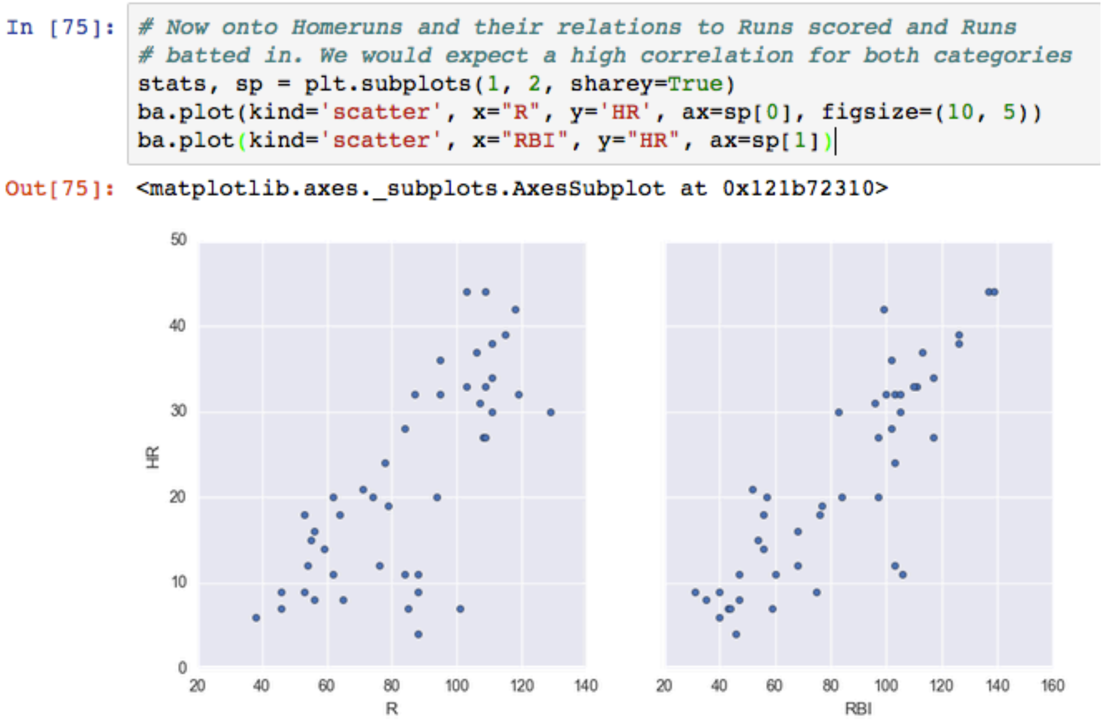
Continuing this analysis, it was found that as the number of home runs a player hit in a given season increased generally correlated to a larger number of runs that the player drove in (RBIs). A positive correlation was also found between the number of home runs that a player hit, and that player's season total for runs scored. These last too are not ground breaking findings, as the number of runs scored has a 1-1 relationship with the number of Home Runs that a player hits, and RBIs also have (at worst) a 1-1 relationship with Home Runs. Another not too surprising correlation could be found between the number of bases that a player stole in a season, and the number of runs that they scored. This makes sense because stealing bases increases the likelihood that a player will score during a inning sequence in a specific game.

Another surprising finding was the lack of a correlation between a player's On-base percentage and the number of runs they scored in a given season. One would think that the increasing the number of times a player gets on base (calculated by: hits + walks + HBP / Total ABs) would cause them to score more, but the least squared line that was plotted was practically flat. At this point, it should be reiterated that these findings only apply to this sample of players, and thus should not be applied to all the players in the league.

In the coming weeks, I plan on continuing the regression analysis to gain further understanding of the data set beyond the linear regression conducted recently. Other forms of regression would appear to fit better with other sub sets of the data field, and would most definitely provide a more accurate analysis of the data. On top of this other forms of analysis will be conducted such as applying SVMs to the data, before utilizing this analysis to pick the best performing players across the varying categories. Weights will be applied to certain categories, as certain fields are more beneficial to a team's success than others (e.g. triples are better than doubles). These weights will be applied into one number that shall be the final metric applied to each player in order to select the player with the best offensive season. Further, players who appear on the list of finalist multiple times will have their final metrics averaged, in order to see which players performed the best over the first five years of the 2010's.

Michael Puiszis
Data Science
Project Topic Paper

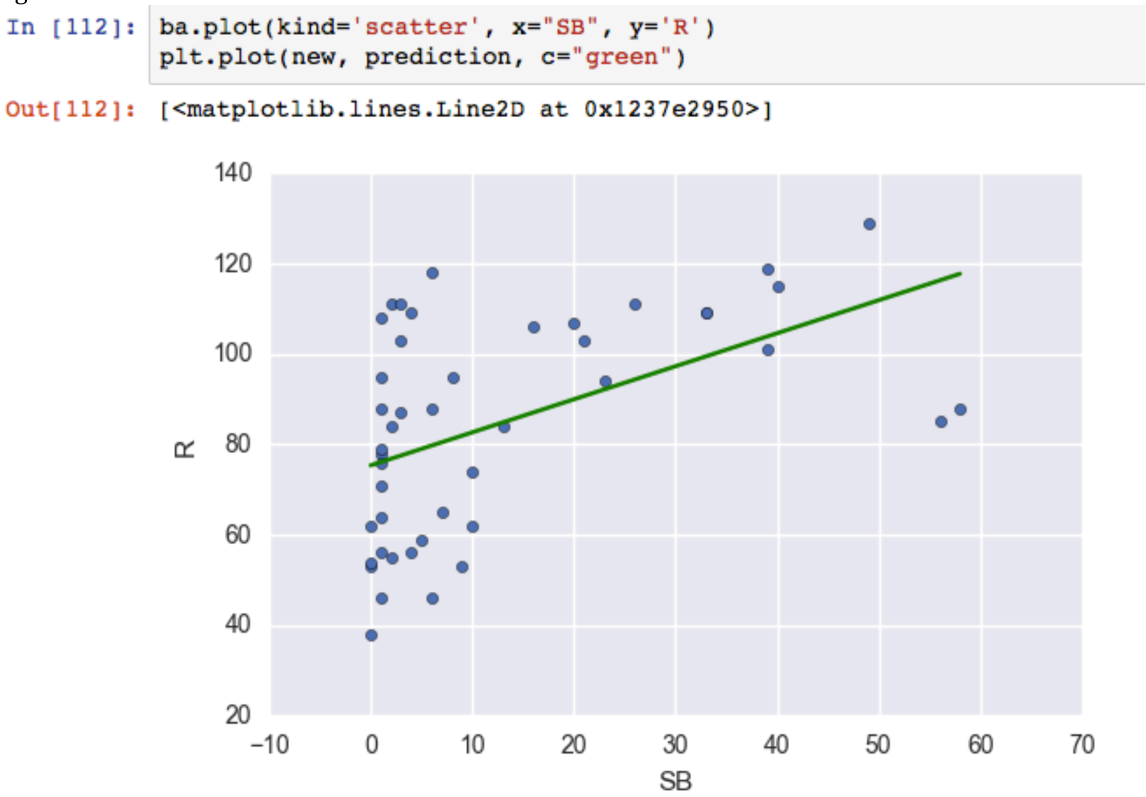
Figure 1: Homeruns vs. Runs/Runs batted in



```
In [75]: # Now onto Homeruns and their relations to Runs scored and Runs
         # batted in. We would expect a high correlation for both categories
         stats, sp = plt.subplots(1, 2, sharey=True)
         ba.plot(kind='scatter', x="R", y='HR', ax=sp[0], figsize=(10, 5))
         ba.plot(kind='scatter', x="RBI", y="HR", ax=sp[1])

Out[75]: <matplotlib.axes._subplots.AxesSubplot at 0x121b72310>
```

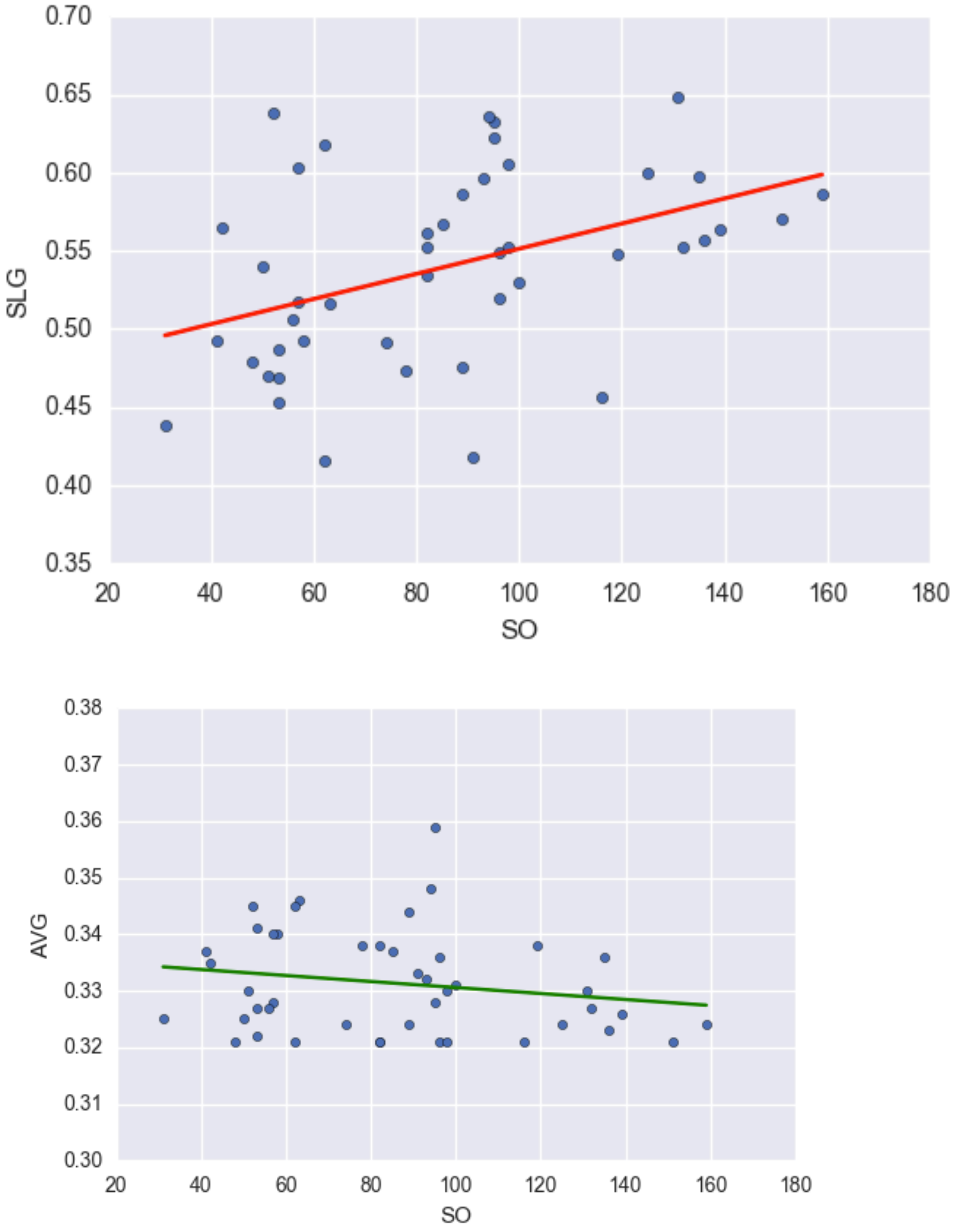
Figure 2: Stolen bases vs. Runs Scored



```
In [112]: ba.plot(kind='scatter', x="SB", y='R')
          plt.plot(new, prediction, c="green")

Out[112]: [<matplotlib.lines.Line2D at 0x1237e2950>]
```

Figure 3 and 4: Strikeouts vs. Slugging percentage/Batting Average


SLG
SO
AVG
SO