In [7]:
```
from IPython.display import Image
```

Baseball Data: Relational Database Project ¶

By Jay Dickson and Jill Reiner

Introduction:

Sean Lahman is the manager of the largest repository of baseball data which is available to the public. An avid baseball fan, he has compiled data in multiple tables such that the public can easily access and view the data. Originally produced in 1995, "The Lahman Baseball Database" is updated yearly to account for data produced with each additional season. It covers all of baseball from 1871 and on. This data is organized in multiple tables, what is called a relational database model. Essentially this means that each team, player, and season's statistics reside in different, interconnected tables. There is an incredible amount of data compiled in this database, but due to this size, it can be rather difficult to work with. Further information about the data can be found here: Lahman Database (https://docs.google.com/document/d/1Q6Mn2g_QT6u0_tcwRbXYp51eEhdzQceiochoI2GATEA/edit).

For this project, we used SQL and wrote queries in order to access the data. These queries would specify which data table we wanted and which elements should be accessed. Through different SQL techniques, we were able to group certain columns and establish various cutoffs, writing functions that abstracted queries and made them reproducable. SQL was used for much of the data wrangling and manipulation while Python's pandas library enabled us to write the resultant tables into CSV's. Finally, graphics were produced using Tableau. These graphics aimed to answer questions which we will discuss in the next section.

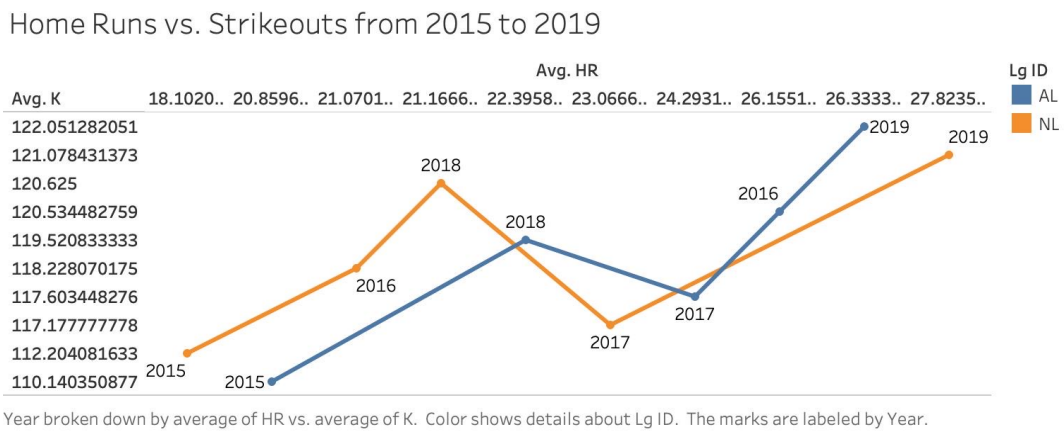
Since we were a two person group, we wrote four different SQL queries for four different questions. **Question 1** looked at the relationship between strikeouts and home runs. We looked at this from 2015-2019. For **Question 2**, we asked if there is a position which produces the most RBIs. **Question 3** looked at players who have lead the league in home runs and asked if a player lead the league in home runs for one season, how likely was it for them to be in the Hall of Fame? Finally, **Question 4** looked at how batting averages per team have changed from the 1919 season to the 2019 season, over a century. We were able to solve these questions using a variety of data wrangling techniques and producing a number of interesting charts below.

Question 1:

For Question 1, we were interested in seeing if there was a relationship between the number of strikeouts a player had in a single season versus the number of homeruns hit in a single season. The home run and the strikeout are arguably two of the most defining characteristics of the game of baseball. There have been many studies analyzing this relationship, specifically as of late because the 2019 season saw records for both statistics (https://www.boston.com/sports/mlb/2019/05/02/mlb-home-runs-strikeouts-record). This is due to the increase in skill in both pitchers and batters as well, and some credit this to a philosophical change in baseball where striking out isn't as bad as it used to be back in the day. Many players have talked about how their swings have evolved, and many have followed this "all or nothing" type of trend, which can be seen in our graph.

In [8]:
```
Image("HRvSO.jpg")
```

Out[8]:

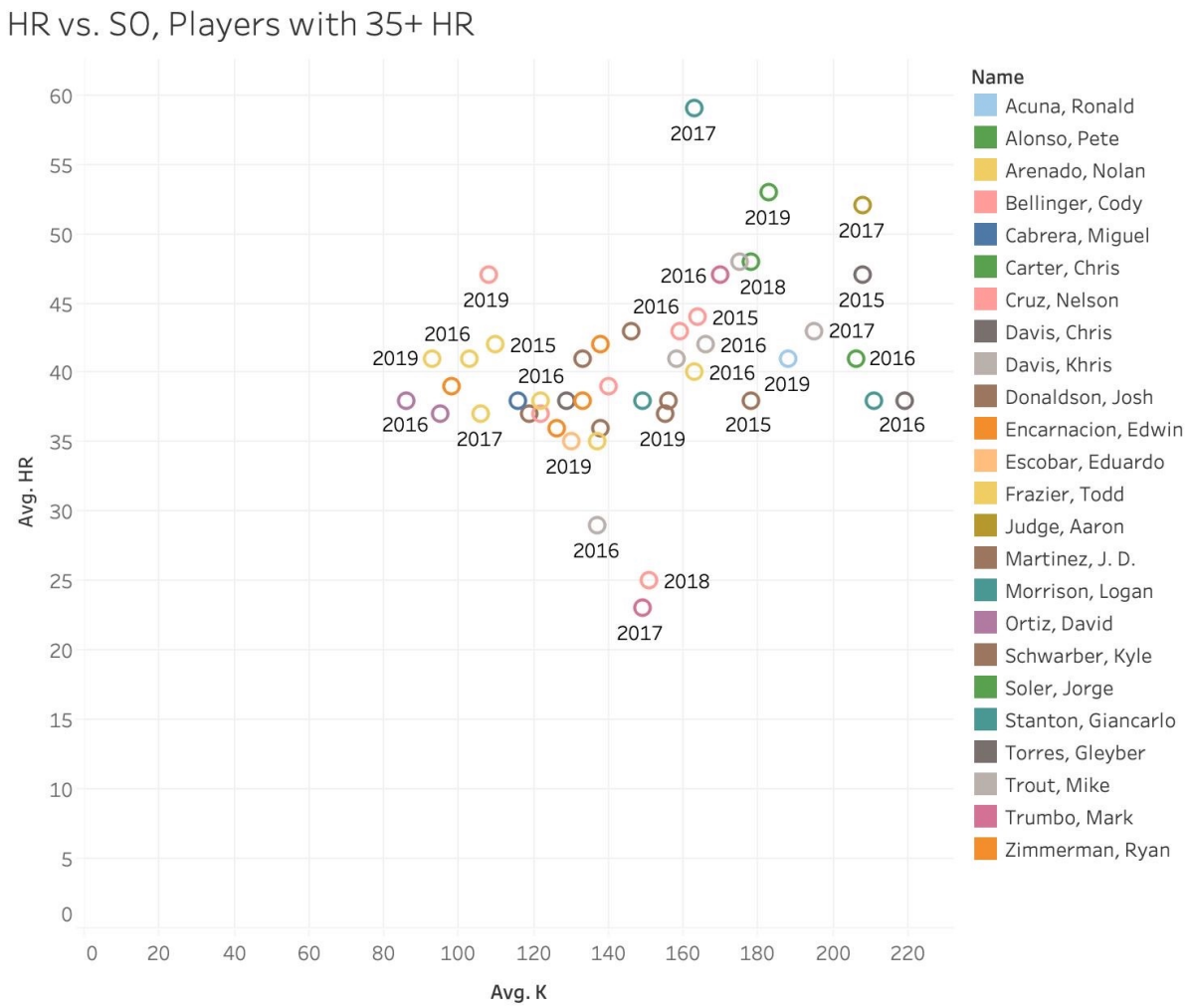


We looked at statistics from the last five MLB seasons because we were interested in the recent trend of increased home runs and strikeouts that has taken place. As seen in the graph, the average amount of strikeouts and home runs per league have generally increased, which is in line with many other studies analyzing this relationship. The 2019 season did see record highs for home runs and strikeouts. Specifically looking at the American League, the average number of strikeouts per player is slightly higher than that of the National League and the average amount of home runs hit by National League players seemed to be about one home run higher than AL players.

Additionally looking at our second graph for this question, we looked at individual player home run and strikeout statistics for the same time span, this time limiting our players to those who have hit 35 or more home runs in a single season.

In [9]:
```
Image("HRvSO 35+.jpg", width=600, height=300)
```

Out[9]:



Average of K vs. average of HR. Color shows details about Name. The marks are labeled by Year. The context is filtered on Name, which keeps 24 of 233 members.

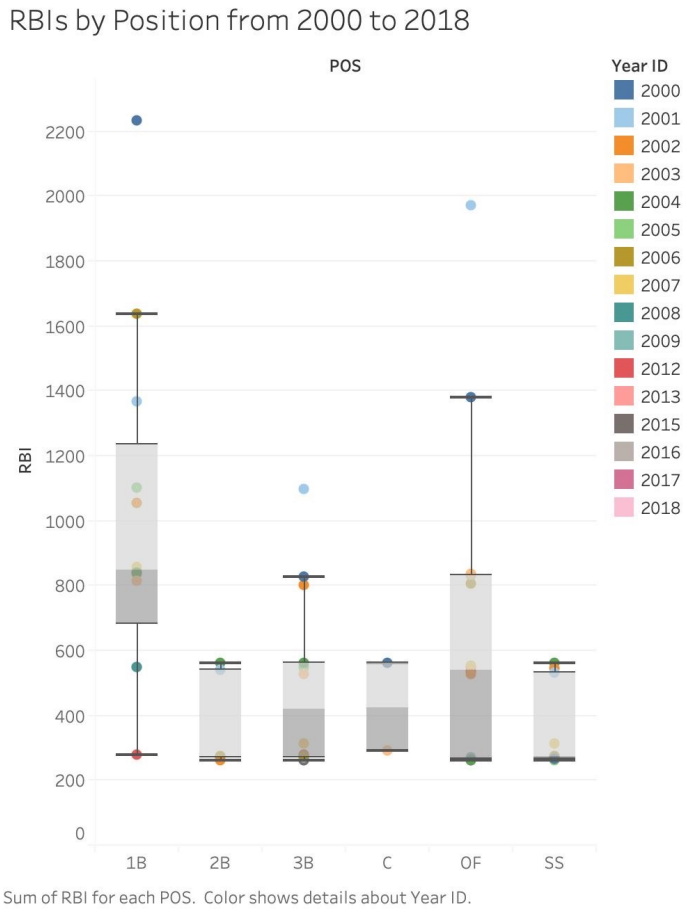
This graph also shows that there is a strong relationship between home runs and strikeouts. As the average number of strikeouts increases, the average number of home runs hit generally increases as well. This can be seen on the graph for players like Aaron Judge of the New York Yankees and Pete Alonso of the New York Mets (the gold and light green points, respectively). These two players are two of the best players in the MLB right now and also have been Home Run Derby champions in the All-Star game, so it is clear that these two have the "all or nothing" mentality. However, there are more efficient players that can be seen on our graph, like the light pink 2019 dot, which is Cody Bellinger of the Los Angeles Dodgers. Bellinger's average strikeouts in 2019 were around 110 and his homeruns were around 47, which seems to be one of the more efficient players during this span in terms of home runs and strikeouts. It will be interesting to see whether this all or nothing strategy will continue in the upcoming seasons as players keep swinging for the fences.

Question 2:

In Question 2, we wanted to discover if there was a position which hit more RBIs than other positions. We wanted to look at the most recent years of baseball in order to try and uncover this trend so we looked at seasons played from 2000-2018. With this in my mind, we created a CSV mapping from Year, Team, Player Name to League, Position and RBIs. Then we worked in Tableau to create our visualizations. Using these most recent years eliminated possibly faulty data recorded from baseball's early days and allowed us to account for stronger and more results driven athletes. Ultimately, we produced the following boxplot to show how RBI's varied by position. One important note is that we made the decision to exclude both Designated Hitters and Pitchers. We did this because half the teams in the MLB use Designated Hitters while the other half uses Pitchers in at bats. This has to do with the rules governing baseball.

In [11]:
```
Image("RBIbyPOS.jpg", width=400)
```

Out[11]:



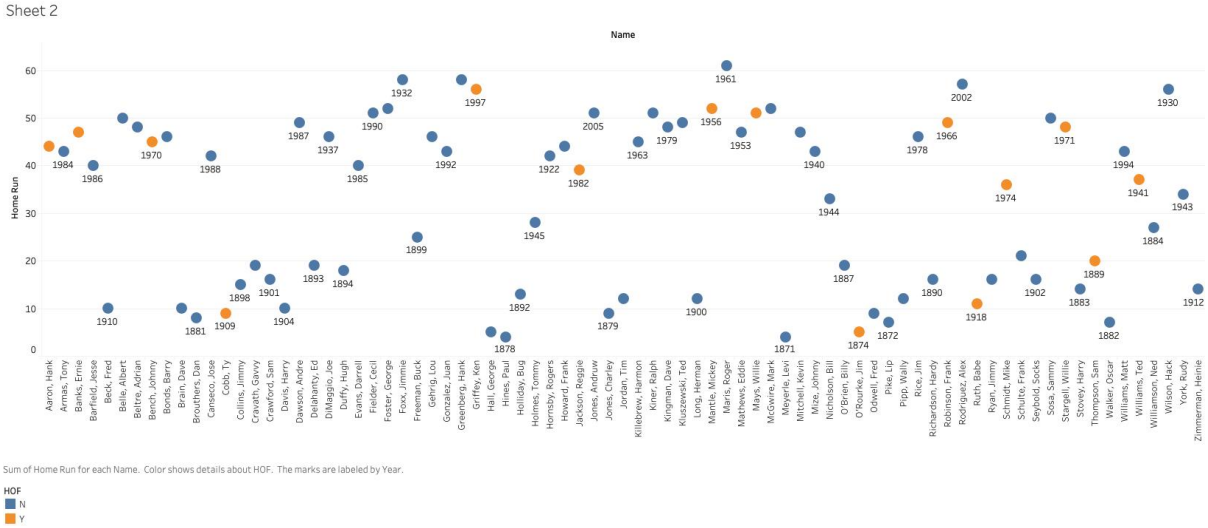
From this boxplot, it seems that first baseman tend to bat in the most RBIs out of all other position players. The main reason for this seems to be that first baseman usually are the worst fielders for a team. They are given a position in the lineup so their batting prowess can be used. It is the least mobile position in the field and first baseman have the heaviest bats (hit the farthest). The opposite can be said of short stops. Short stops may have the least RBIs because they are employed for their fielding abilities. They are the most agile fielders and are an important piece to a team's defense. Lastly, outfielders seems to have a flaw. Since there are three outfielders per team, it would be beneficial if if it was labelled by right fielder, left fielder and center fielder. So here, we have the aggregation of three different positions, creating a possible fault in the data. From this, we have concluded that first basemen most likely bat in the most RBI's, while shortstops and catchers bat in the least RBI's.

Question 3:

Question 3 was concerned with the likelihood that if someone lead the league in Home Runs for at least one year, how likely was it that they would be inducted to the Hall of Fame? One of the main complications of this was Hall of Fame eligibility. This is because in order to be considered eligible, a player must have been retired for at least five years. To combat this, we attempted to filter players out, however this created incosistencies in the data as players who did not lead the league in Home Runs were often considered for the Hall of Fame. This did not work in the data and we were forced to consider every year, unretired players were simply counted as not in the Hall of Fame. A functional dependency was created, mapping from year and maximum home runs to player name and Hall of Fame status. We generated the following table to illustrate this effect.

In [12]:
```
Image("HR_HOF_Tableau.jpg")
```

Out[12]:



From this, it was difficult to tell how big of a role leading the league in home runs would have on a players chance at the Hall of Fame. It appeared that about one-third of the players who lead the league in home runs would wind up making it into the Hall of Fame. However, when we look at the names on this list, it seems that there are some really important figures left out of the Hall of Fame. Both Barry Bonds and Mark McGwire have lead the league in Home Runs and are Hall of Fame eligible, but due to scandals regarding Performance Enhancing Drugs (PED's) have been left out of the Hall of Fame. It is quite possible there are a number of people who have lead the league in Home Run's, however their name has been given an asterisk when it comes to Hall of Fame induction. We can certainly conclude that there is a lot more that determines a player's Hall of Fame status than Home Runs hit over one year. This graph shows that people are elected based upon their entire career and one season cannot make or break a players Hall of Fame status. Additionally, it is likely that the very limited spots in the Hall of Fame have been reserved for people that have pushed the game forward rather than being caught up in irreversible scandals (like this one from the Astros 2017 Championship season: https://www.nytimes.com/article/astros-cheating.html (https://www.nytimes.com/article/astros-cheating.html)).

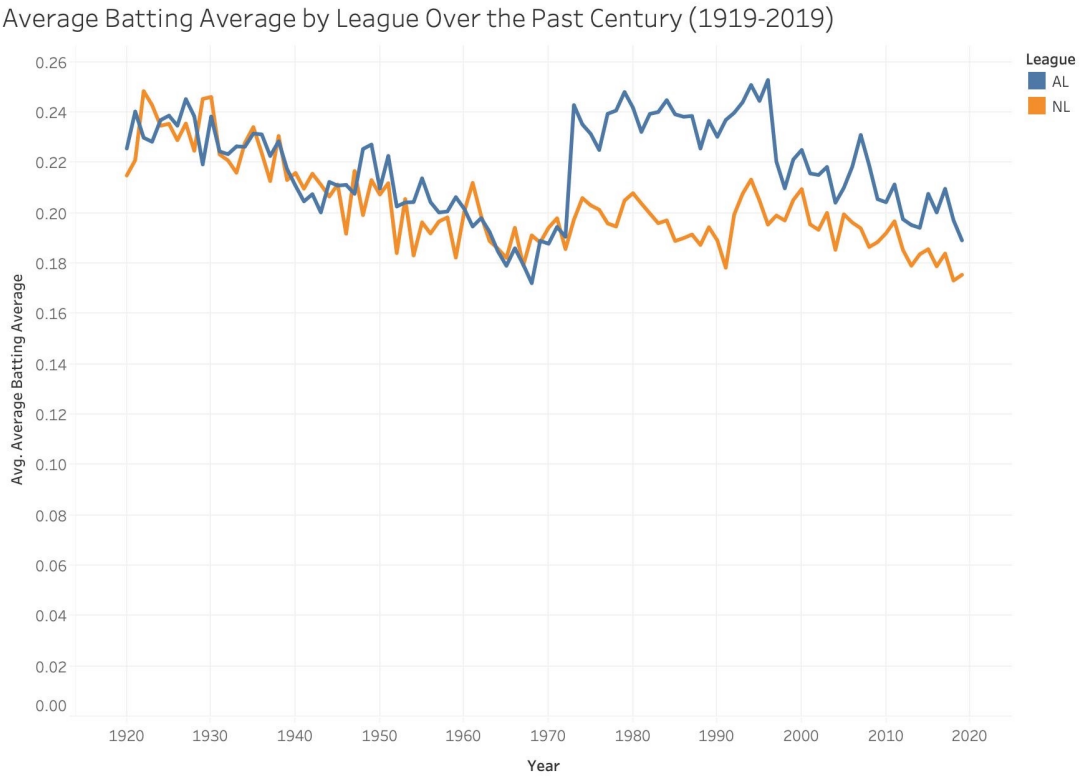
Many factors determine a player's spot in the Baseball Hall of Fame, so we should not be surprised when one season is not the best indicator for Hall of Fame induction. As the game of baseball continues, we hope to get more data such that we can draw better conclusions about how to predict a player's Hall of Fame status.

Question 4:

Finally, we wanted to see how batting averages per league have changed over a century from 1919 to 2019. Going into this, we believed that as hitters became more capable and data began taking over baseball, we would expect to see batting averages rise. But also, we were aware that the American League has a Designated Hitter and in theory should have higher averages than the National League. The graph we produced shows how pitchers have grown and adpated to the hitters against how hitters have grown and adapted the pitchers. Additionally, it does show that the AL has higher batting averages than the NL.

In [13]:
```
Image("AVG_BA.jpg", height="200")
```

Out[13]:



When looking at this graph, we see that batting average per team in both the National League and American League from 1919 to around 1970 for the most part decreased, and there wasn't a large difference between the two leagues. In 1970, batting average per team spikes for the American League and doesn't change all that much for the National League. Then from 1970 to around 1995, there is a considerable margin between the two leagues. The American League hits its peak around 1995 as well. From 1995 to present day, for both leagues, batting average per team has mostly decreased. This is likely due to the American League's adoption of the Designated Hitter. In 1973, pitchers in the AL stopped batting while the NL continued using pitchers. This shows the importance of a strong Designated Hitter to a team.

Batting average in baseball now means less to players than it used to in the 1900s. Back then, it was one of the more evaluative statistics, but since analytics have become more prevalent, there are even more statistics that can measure a player's performance better than batting average can. Batting average alone does not tell you a whole lot about a hitter. With all of these new statistics, you can find out a lot more about a hitter. But batting average isn't meaningless (https://www.espn.com/mlb/story/_/id/26757283/is-300-hitter-thing-past). It can help tell you what kind of a hitter a batter is, if not how effective.

Though batting average per team has generally decreased over time, players' offensive production has not slowed one bit as seen by some of our other plots with home runs and RBIs.

Conclusion:

In conclusion, the Lahman Baseball Database has allowed us to answer many interesting questions about baseball and how it has changed over time. Overall, we have discovered that there is a strong relationship between a batter's average number of strikeouts and average number of home runs. In addition to this, we learned that first basemen generally have the most RBIs out of any position and shortstops generally have the least RBIs. Our third question showed that there is a lot more that determines a player's Hall of Fame status than the amount of home runs hit over one year. Finally, our last question showed that batting average per team has decreased over time.