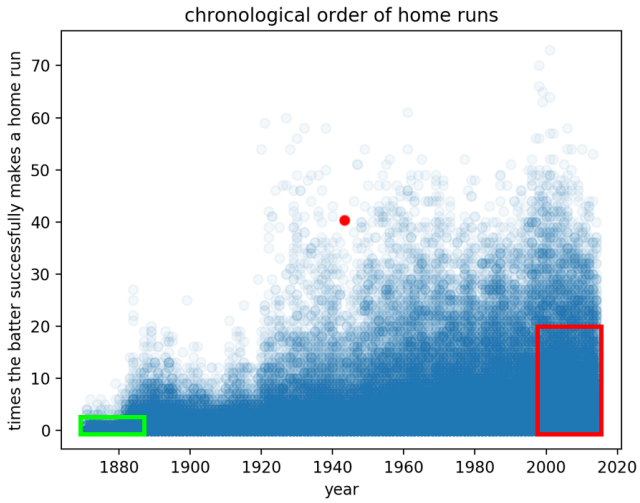
Baseball data analysis

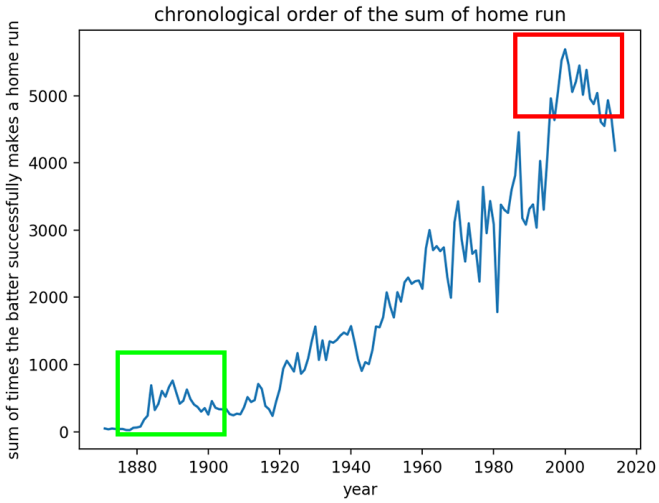
Jeffrey Wang, 5/24/2018

The dataset of baseball, including information of awards, batting, fielding, pitching and etc, concludes the status of each player, each team and each league(within the U.S.) from 1871 to 2014.

```
-- Get the max and the min year
SELECT MIN(yearID), MAX(yearID)
    FROM batting
    WHERE yearID != 0
```

The scale of the dataset enables analysts to conclude similarities and differences of the data between different time periods. By comparing those data, the tendency/pattern of the data is somewhat recognizable. One of the most recognizable pattern is the number of home runs a player successfully made in each year. I visualized the data with matplotlib, an open source plotting tool, and the programming language Python.





1-0-1 and 1-0-2

Graph 1-0-1 scattered the amount of home runs each player have made in a specific year. Each dot represents the home runs a player made in a specific year. For example, the red dot in the center of graph 1-0-1 symbolized a player made totally 42 home runs in 1941. There are two places in the graph that I found out interesting, and I marked them with boxes. Both boxes

are highly concentrated with dots. The green box indicates that around 1880, most player made 0-3 home runs per year. The highest record at 1880 was made by Bill Ahearn, who made 6 home runs in TRN. The red box, on the other hand, clearly suggests the vast difference of the baseball environment three decades after 1880. In "modern" era, baseball players usually make 0-20 home runs for a year. Even a regular baseball player in the twenty-first century can easily break Bill Ahearn's record, which the latter only made 6 home runs. In comparison, the record in 2010 was made by David Aardsma, who have made 54 home runs in SEA. David's data is 9 times greater than Bill's. The second graph was created with the same dataset but in a different approach: The broken line shows the summation of all home runs has been made in the statistic within a year. In this graph, the difference in terms of the number of home run is even more noticeable.

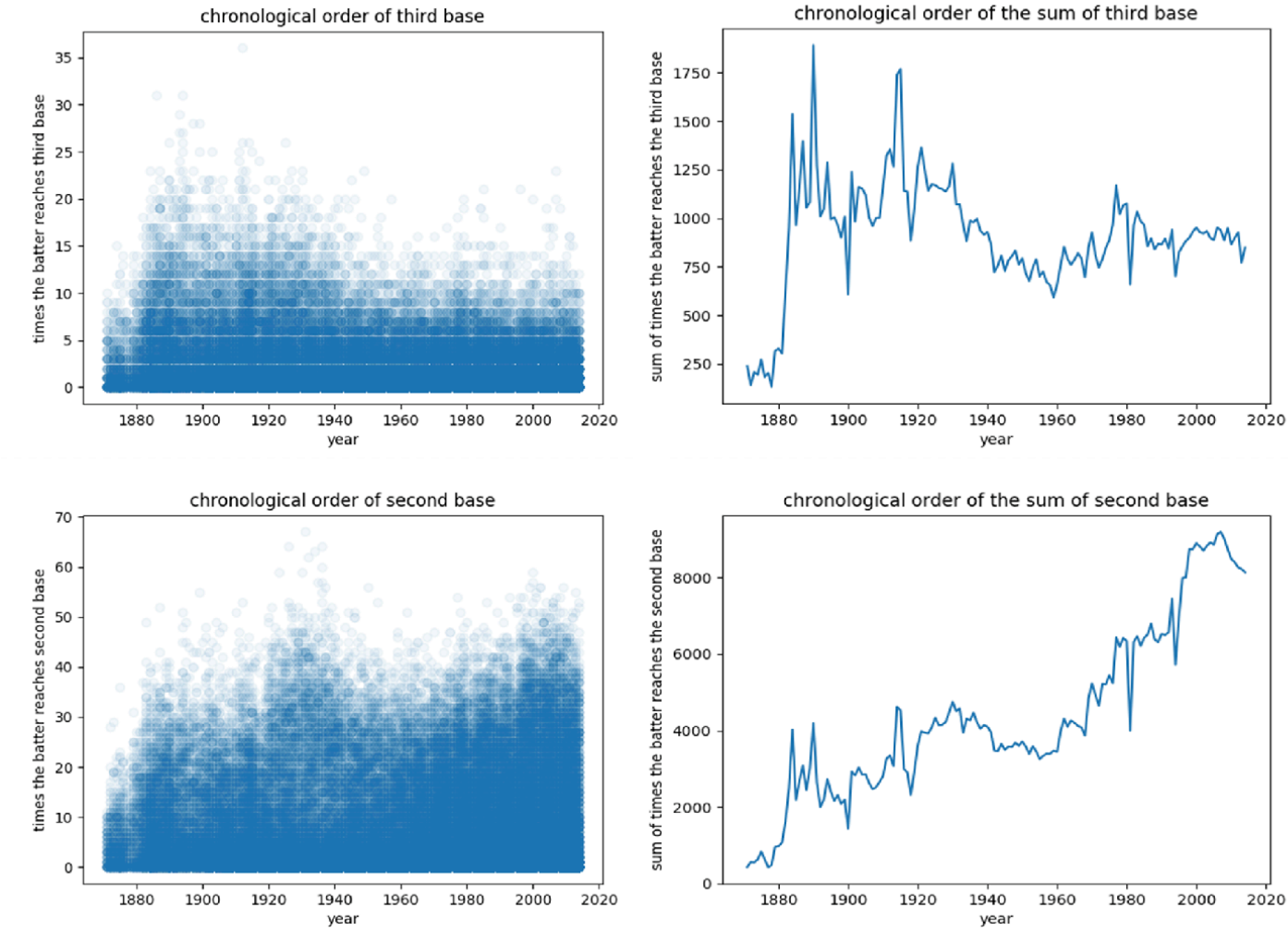
I gathered the data with the following SQL commands:

```
-- Get the number of home runs each player have made in each year
SELECT yearID, HR
    FROM batting
    WHERE yearID != 0
    ORDER BY yearID

-- Get the summation of home runs have made in each year
SELECT yearID, SUM(HR)
    FROM batting
    WHERE yearID != 0
    GROUP BY yearID

-- Get the highest record in term of home run in 1880 and 2010
SELECT batting.playerID, yearID, teamID, MAX(HR), nameFirst, nameLast
    FROM batting
    INNER JOIN `master` ON batting.playerID=`master`.playerID
    WHERE yearID=1880 OR yearID=2010
```

In order to find out if the similar pattern also exists in data other than the number of home runs, I also summarized the statistic for the total number of the batter who successfully reaches second base (2B) and third base(3B) for each year. I created graphs below with the same method I used to generate graph 1-0-1 and 1-0-2.



1-0-3(top-left), 1-0-4(top-right), 1-0-5(bottom-left) and 1-0-6(bottom-right)

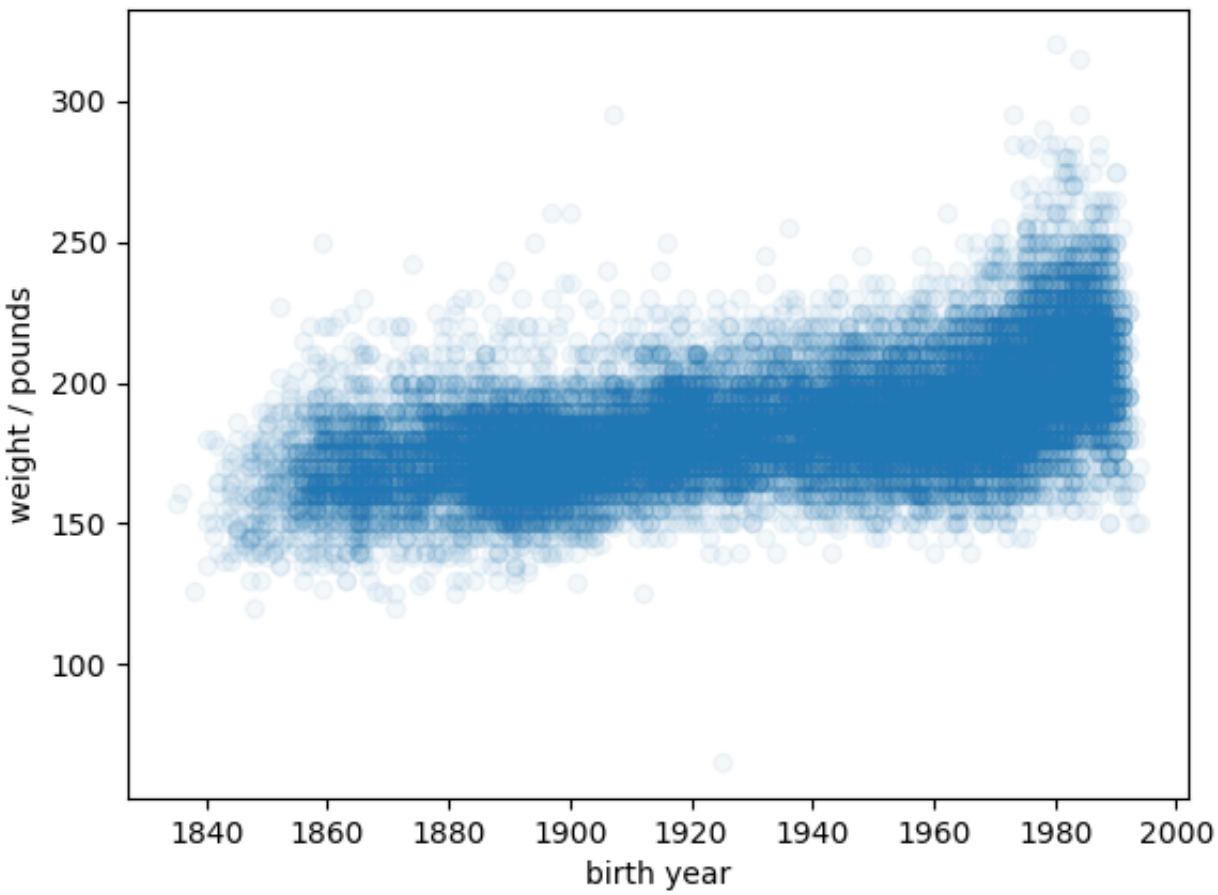
Graph 1-0-5 and 1-0-6, which visualized the change in the total number of the batter successfully makes to the second base, presents a surging trend. The trend corresponds to the tendency of the number of home runs, which increase dramatically in the past decades. On the other hand, the trend of the third base(graph 1-0-3 and 1-0-4) seems to be different. In the beginning of those figures, the total number of making to the third base increased dramatically, whereas after the growth, the trend decreased over time. Since the number of home runs, second base, and first base(I didn't put the graph in this document) all present the same characteristic, why does the characteristic of the third base becomes an outlier? Unfortunately, I did not find a strong evidence to explain this disorder. However, I believe that the following aspects might be the factor that shifts the trend:

1. The advancement of baseball strategies. Since players and coaches strengthened their understanding of baseball over time, they might develop some new strategies. For example, instead of taking the possibility of getting "knocked out" when running toward third base, players might just want to stay on the second base safely.

2. Although the summation of the number of reaching third base each year(graph 1-0-4) was decreasing, the highly concentrated area in graph 1-0-3 didn't change much. The highly concentrated area (dark area in 1-0-3) suggests the most common number of making to the third base in a year. In graphs with regular tendencies(e.g. the graph of the number home run), both light area and dark area changed. However, in the graph of making to the third base in a year, the highly concentrated area didn't modify a lot.

SQL commands for this section are quiet similar to the commands from the previous section. If you would like to see the code, please check out the codes section at the end of the document.

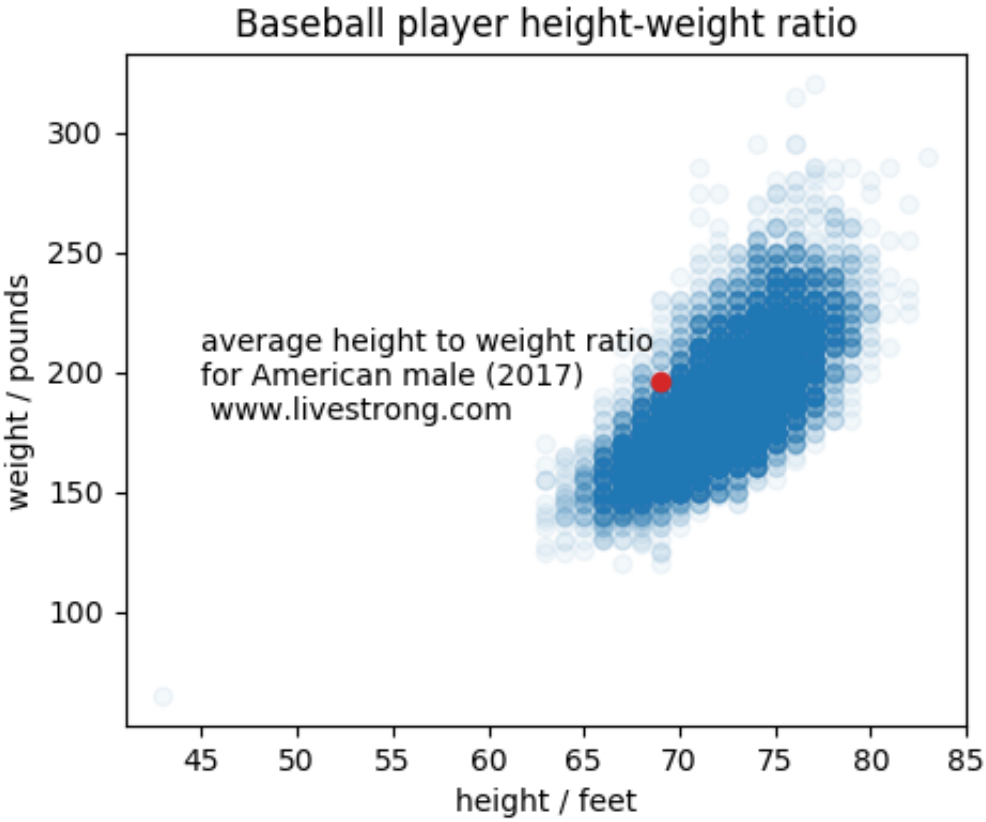
What causes the surge of the number of home runs, second base and so on? The first possible solution that poped up in my brain was the increase of the player's physical strength. I assumed that as players' weight increases over time, their strength would increase as well so that batting a home run would be more accessible.



1-0-7

The x-axis is the year that each player was born on, and the y-axis is the weight of the player when he is registered. From the graph, I observed that the average weight was gradually increasing from 1840 to 1960. After 1960, the average weight increased even faster. The phenomenon correlates with my assumption, but I think the evidence is not strong enough to prove the point, because the increasing rate(derivative) is not that high.

I also created a graph that scattered the height-to-weight ratio of baseball players in the dataset.



1-0-8

The x-axis is the height of the player in feet, and the y-axis is the weight of the player in pounds. In comparison, I marked the average weight and height of American male over age 20 on the graph(where the red dot located). According to FastStats, the average American man over age 20 weighs 195.7 pounds, and the average height is just over 5 feet 9 inches (about 69.2 inches) tall. Baseball athletes have an average weight at 185 pounds and height at 72.2 inches. The graph is beneficial for those who are planning to become a professional baseball

player. By calculating their own height-to-weight ratio, whether if their ratios fit with the majority of real baseball athletes is distinguishable.

I gathered the data with the following SQL commands:

```
-- Get the weight and year comparison
SELECT weight, birthYear
    FROM `master`

-- Get the weight-to-height ratio
SELECT weight, height
    FROM `master`

-- Get the average weight and height of all baseball player
SELECT AVG(weight), AVG(height)
    FROM `master`
```

Another interesting relationship I founded in the dataset is the salary. Within the same year, the wage difference between players is considerable. For example, In 2000, the highest income receiver gain 15M for the season, whereas the lowest only receive 80K. What makes the vast wage difference between baseball players? With the help of scikit-learn, an open source machine learning library, I'm able to find if the player's performance on the field has an impact on the salary. I also took the baseball team, league and also player's weight and height into account. Here is a list of data I extracted from the database:

- Data from salaries: year, league, team, salary
- Data from appearances: G_all, GS, G_batting, G_defense, G_p, G_c, G_1b, G_2b, G_3b, G_ss, G_lf, G_cf, G_rf, G_of
- Data from master: weight, height, bats, throws, firstName, lastName

I joined the results from different tables with the following SQL commands:

```
SELECT salaries.playerID, salaries.yearID, salaries.lgID, salaries.teamID, salary,
               G_all, GS, G_batting, G_defense, G_p, G_c, G_1b, G_2b, G_3b, G_ss,
G_lf, G_cf,
               G_rf, G_of,
               weight, height, bats, throws, nameFirst, nameLast
    FROM salaries
    INNER JOIN appearances ON salaries.playerID = appearances.playerID and
                                                  salaries.teamID = appearances.t
eamID and
                                                  salaries.lgID = appearances.lgI
```

```
D and
                                                    salaries.yearID = appearances.y
earID
    INNER JOIN `master` ON salaries.playerID = `master`.playerID
```

I assumed that a player's last year's performances would have an impact on his next year's salary. Therefore, I pushed the data of a player's appearances from year `x` to year `x+1`, in order to better fit the data. I then separate the data into training data and testing data with the ratio of 8:2. I trained the training data with a decision tree regressor. After several testings, the model has an average $r^2 score$ of `0.63`. Although the $r^2 score$ suggests that there's no substantial relationship between the salary and the labels(factors) I used to train the data, I still found following elements that are (might) decisive to the player's wage by looking deeply in the decision tree.

1. As the year goes up, the salary goes up. The average salary in 1960 was around 475K, and the average salary increased to 3.9M after a half-century. Since more and more people are putting their interests into sports for the past decades, more investors put their money into sports teams. With the technological advances, people can watch games worldwide which also increases the popularity. Furthermore, the inflation of currency can also be the factor that drives the salary up.
2. In certain baseball teams, the (average) salary is usually higher than other teams. For example, San Francisco Giants, Boston Red Sox, and New York Yankees usually give higher wage to the player.

For the implementation of modeling and graph drawing, please see main.py for detail.
For the testing result and accurate $r^2 score$, please see result.txt.

Conclusion

With the gigantic dataset, I found several relationships between different factors(in the dataset) by visualizing data with an open source library. Some visualizations are helpful for people to compare their own data to the average, and others are helpful for people to understand the tendency/trend of certain things. There (maybe) are problems and inaccurate info in my analysis, for example, perhaps I can get a higher $r^2 score$ by using a better way to train the model. However, I think that my investigation at least reveals my personal understanding of the provided dataset. Maybe in future, I can look back to my today's analysis and try to find a way to improve it.

Codes

If you would like to see the code, please click here.
Full SQL commands are also included in the repository.
If the above link doesn't work, please copy `https://github.com/JeffreyWang2864/Baseball-data-analysis` *and paste it into your browser.*