Executive Summary

Nowadays, sports industry become bigger and bigger in terms of entertaining and filling people's spare time. Since people are willing to spend more money on watching professional sports competition, the prosperity of sports business emerge. And Major League Baseball (MLB) is one of the most successful sports leagues around the world. In order to keep the business stable or even boost the business, the impressiveness of competition should be remained on a higher level as possible. To achieve that, the manager of MLB should know the detail of the data generated by players, competitions and audience to know the internal insight of the business.

Data Source

In this project, I will use Lahman Baseball Database to analysis different aspect of player statistics. The Lahman Baseball Database contains pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2020. It includes data from the two current leagues (American and National). The URL link of this dataset: http://www.seanlahman.com/baseball-archive/statistics

Considered the data quality and computational expense issue, I slice the dataset and get the batting, people, team data after 2000.

Data Overview

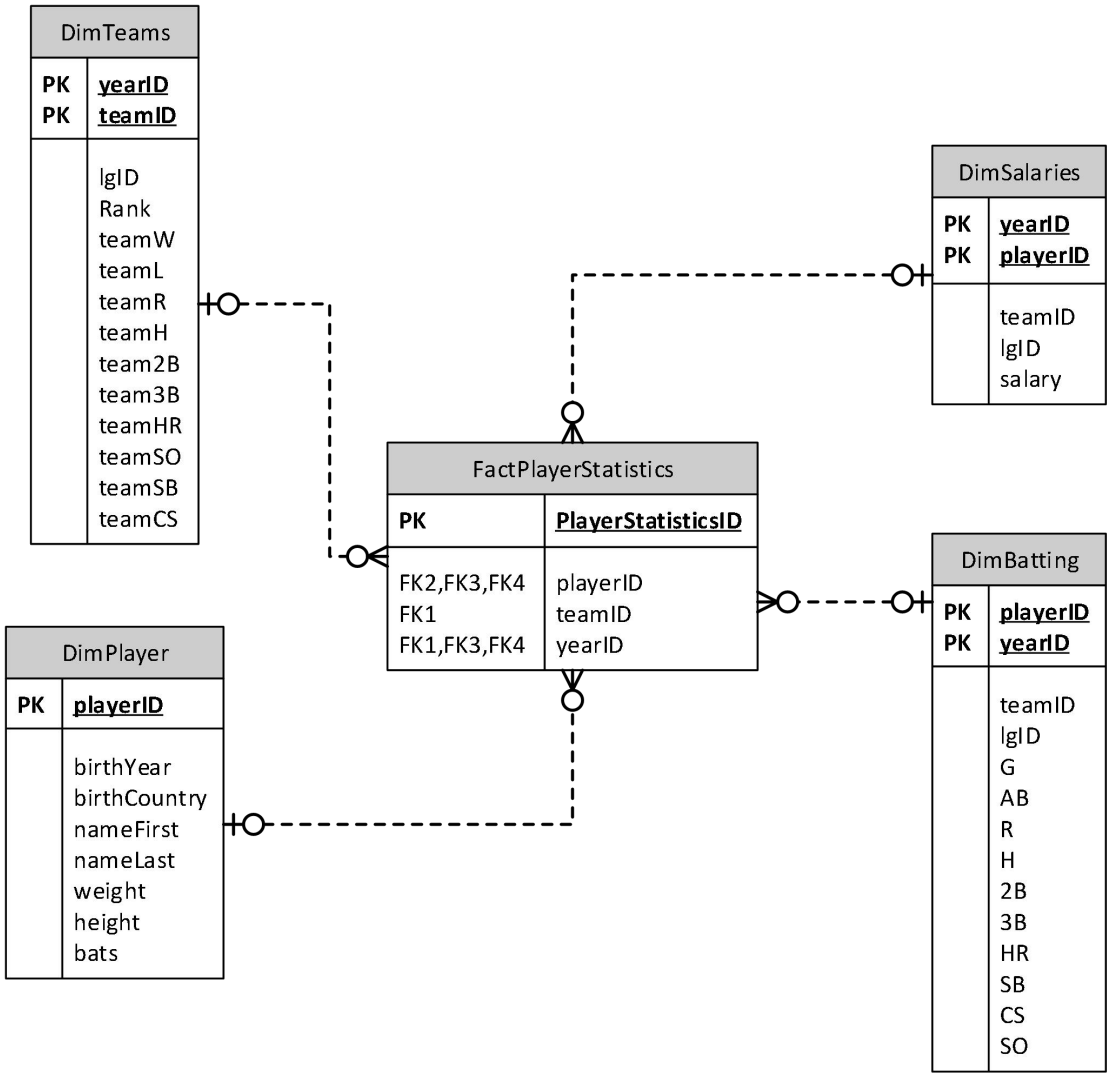
Just take a first glance of the salary data and do a data descriptive analysis as the figure shows below.

salary	
Mean	3116624.155
Standard Error	36114.91747
Median	1000000
Mode	1000000
Standard Deviation	4296005.678
Sample Variance	1.84557E+13
Kurtosis	6.07789824
Skewness	2.296792332
Range	32834426
Minimum	165574
Maximum	33000000
Sum	44100231787
Count	14150

The salary data contains each player's salary in each year (from 2000 to 2018). The maximum value is 33,000,000 dollar per year and the minimum value is 165,574 Dollar per year. You can see the maximum salary almost 10 times higher than average salaries, which make sense in sports industry because only top players attract the most attention and people are willing to pay their money just for top players. Also, the average salaries are much higher than the median which means the majority of player earn relatively low salaries in the league.

Data Warehousing Model

After data preprocessing (slice and truncate dataset and remove duplicate data in SQL management studio), I build a data warehousing model as figure shows below.



I split the data into 4 dimension tables: DimTeams, DimSalaries, DimPlayer, DimBatting. and 1 fact table: FactPlayerStatistics..

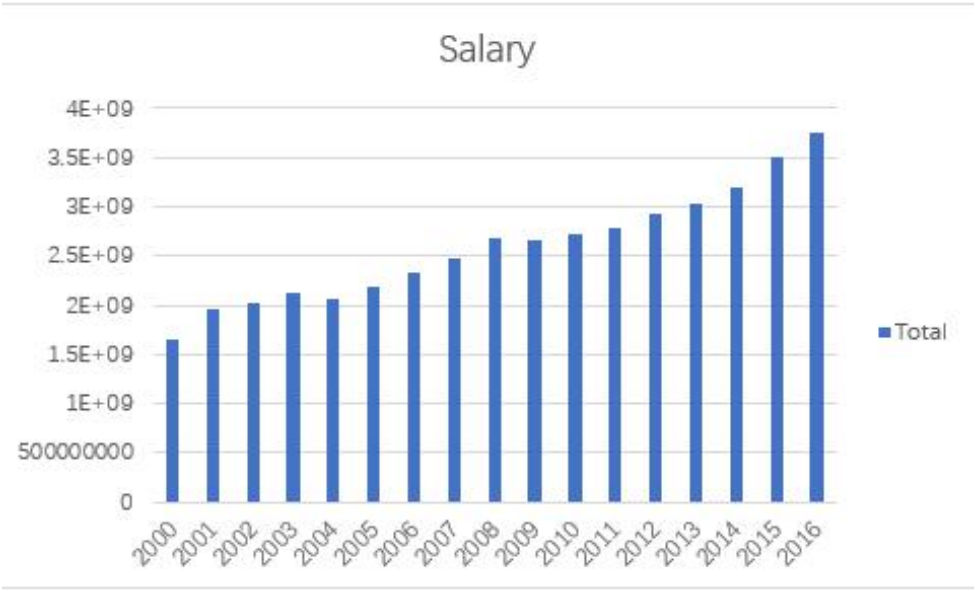
Column explanation: G means Games, AB means At Bats, R means Runs, H

means Hits, 2B means Doubles, 3B means Triples, HR means Homeruns, SB means Stolen Bases, CS means Caught Stealing, SO means Strikeouts.

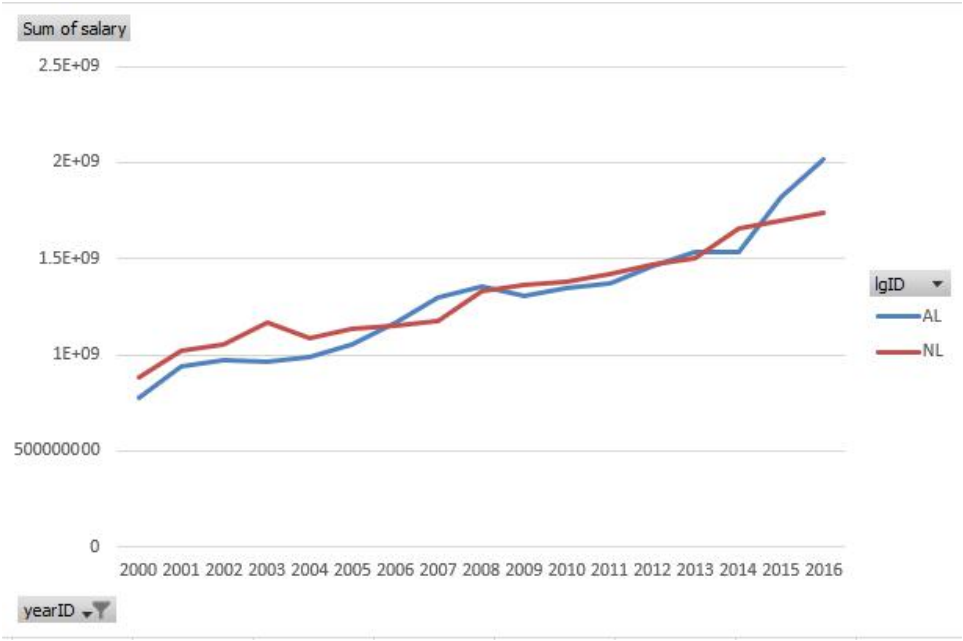
This data warehousing model is on the statistics level so each row in fact table does not represent a specific event, it shows a player statistic (summary of many event) instead.

Analysis

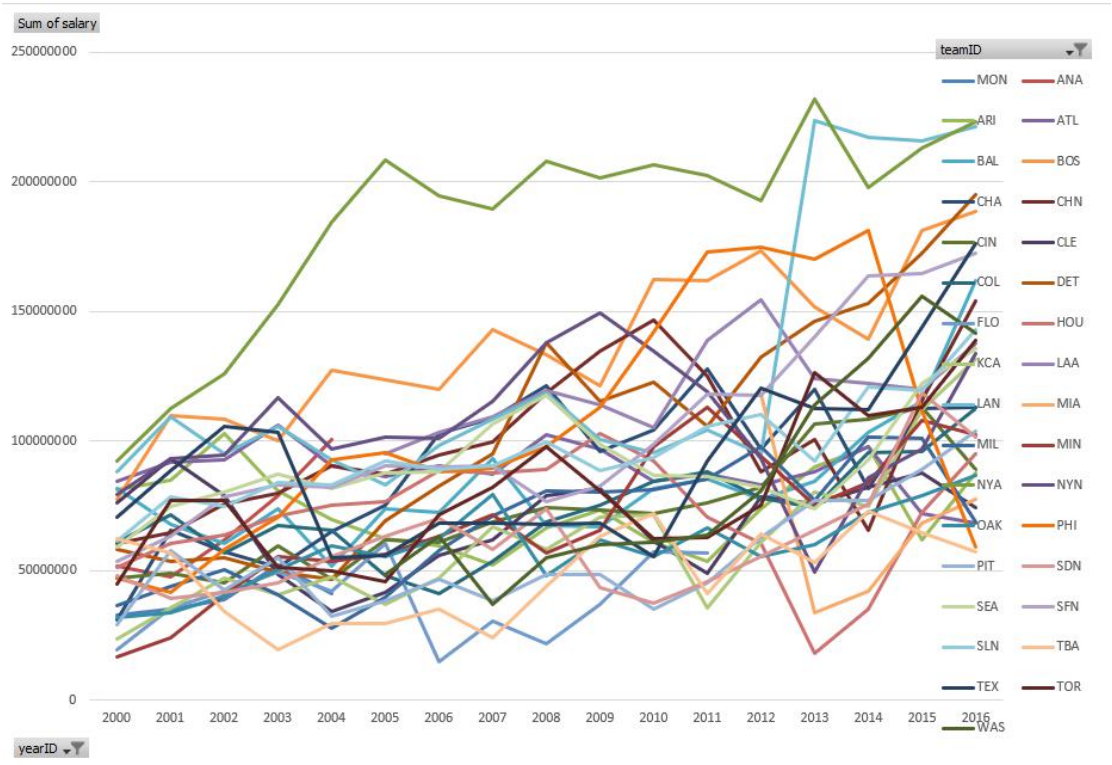
1. Salaries

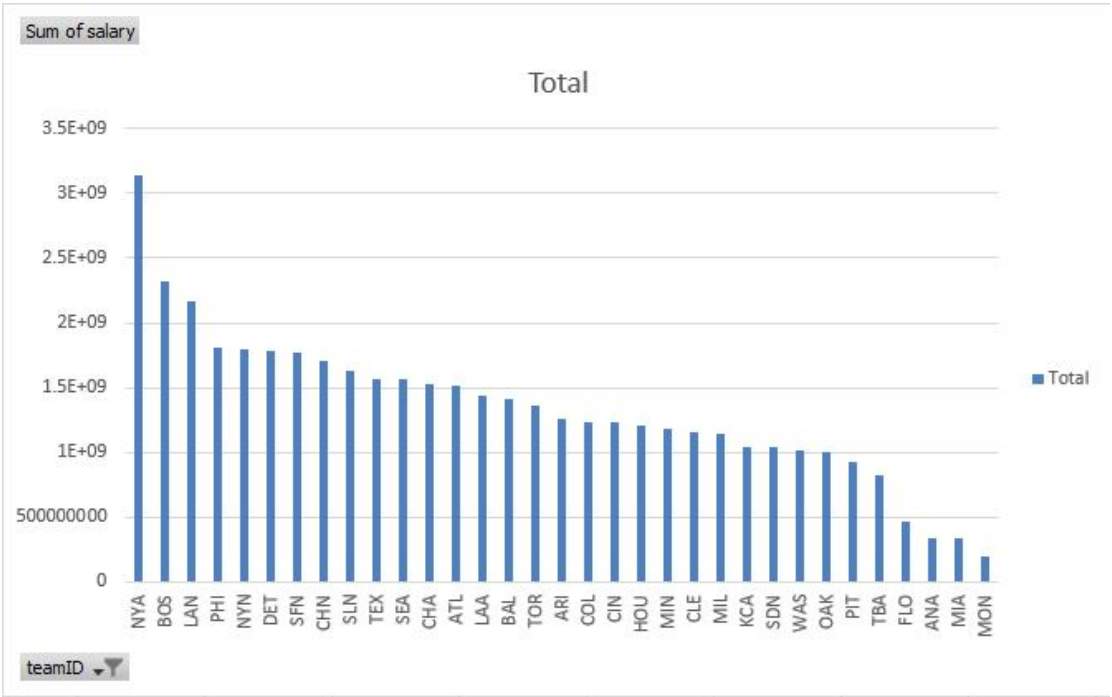


The database only contains the salaries data till 2016. As the figure shows above, the salaries gradually increase year by year indicate the operational success of business of MLB.



MLB is consisted of 15 teams in the National League (NL) and 15 in the American League (AL). The salaries of both leagues remain an increasing trend with the total salary, but AL have a greater slope after 2016.



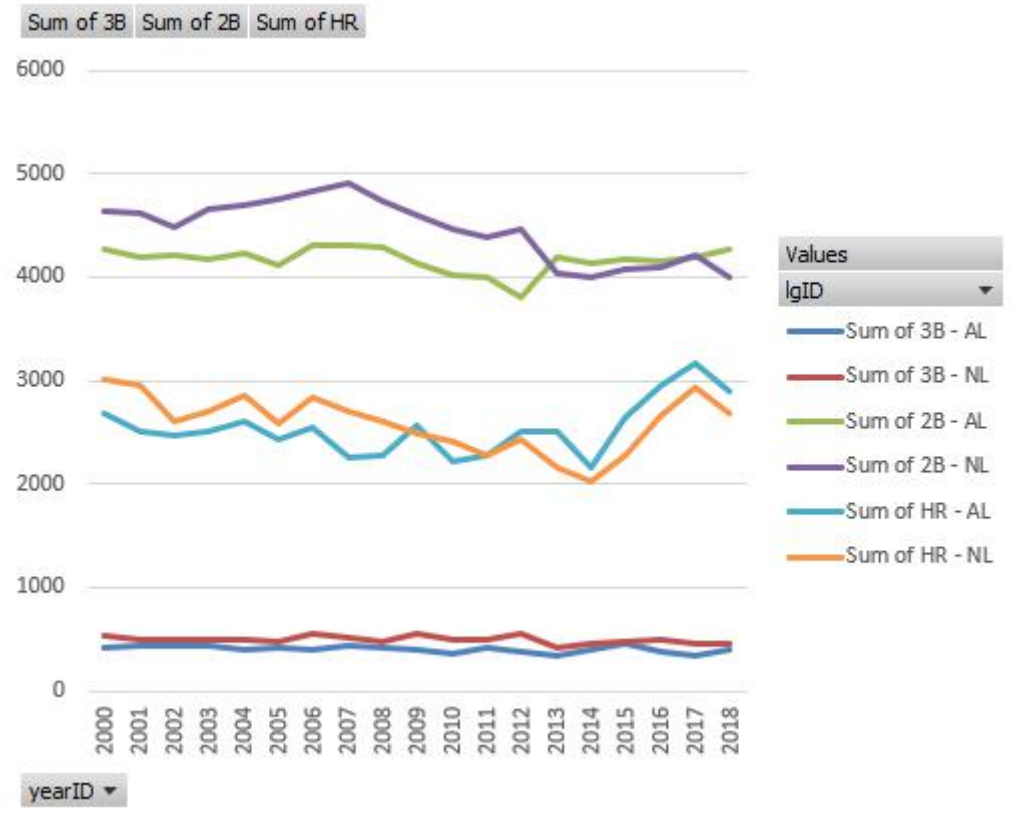


2 figures above show the relationship between each team and their total salaries paid. One interesting thing is that the trend of team salaries is not a simple increasing and it more like a fluctuation. This situation attribute to different reason such as strategy that team manager use, whether import star player to attract

audience and improve team competition level, whether the team decide to tank all season to get a higher draft pick. But overall, the most contribution to team's total salaries is the economy of the region (or maybe the richness of the team boss). In figure above, NYA (Yankees) and BOS (Red Sox) have the 2 highest salaries, also these 2 regions are the 2 biggest market in MLB.

Other than salaries, I want to get deep into the competition itself and understand the 20-year transition of MLB competition in terms of player performance.

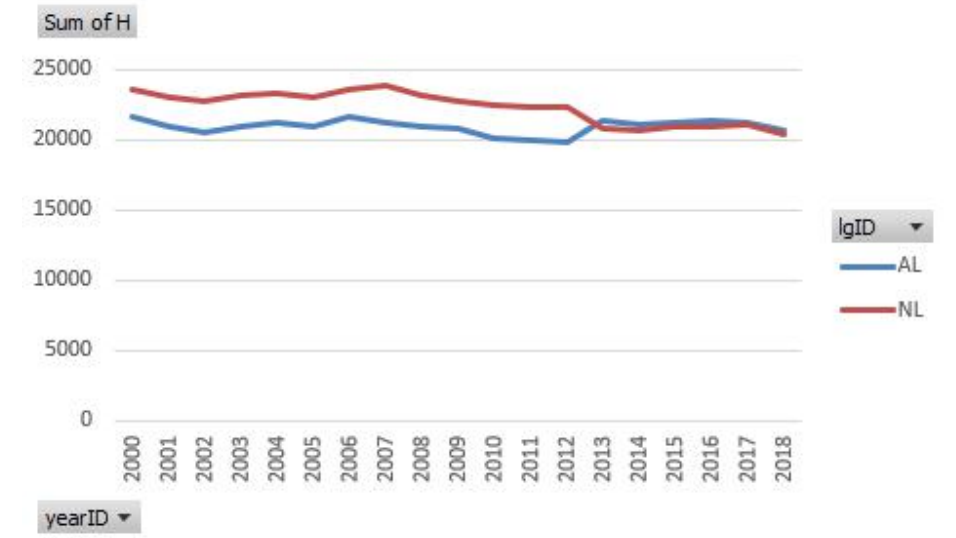
2. hitting

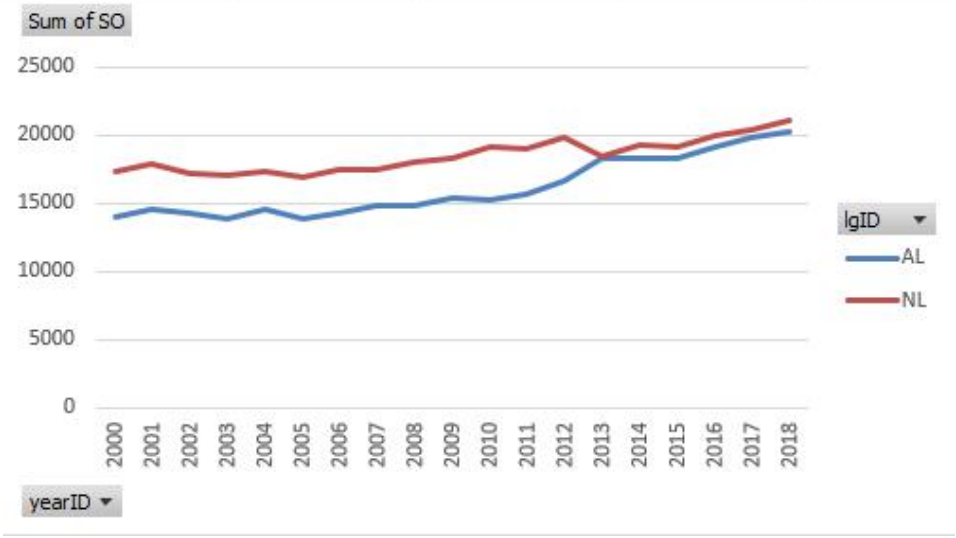


The hitting include 2 base hit and homerun has change a lot. The 2 base hit slightly decrease from 2007 to 2018. And the homerun increases a lot from 2014 to 2017. Seems like some 2 base hit transform to homerun. Let me add a total hit chart and strikeout chart to further explain why homerun increase in recent year.

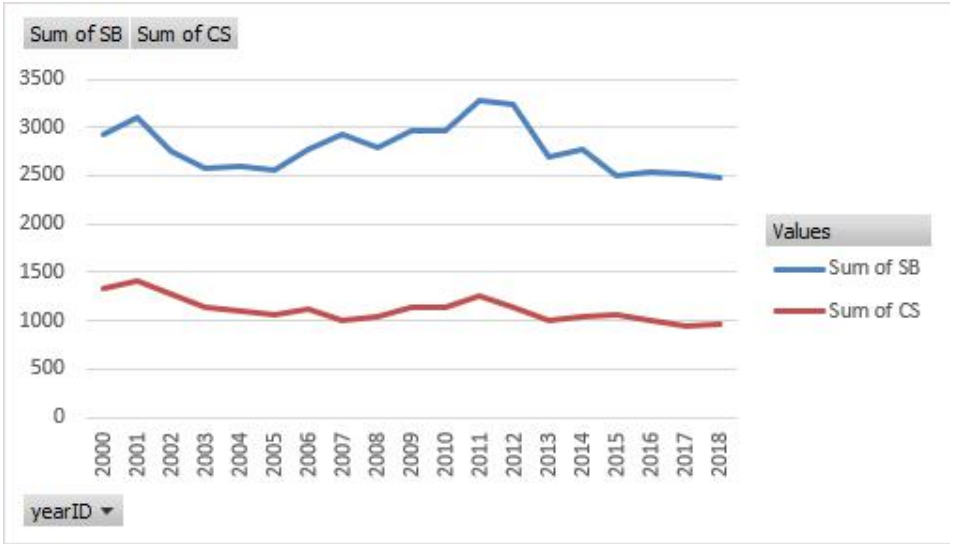
The total hits actually decrease small amount and strikeout increase drastically, which means players tend to take the risk of strikeout to hit a homerun rather than push one base in a hit with lower risk of strikeout.

Also, the National League (NL) have a higher strikeout number than American League (AL) because AL allows a pinch hitter to replace the pitcher in offensive round (normally pitcher do not train hitting a lot, they need to focus on pitching).



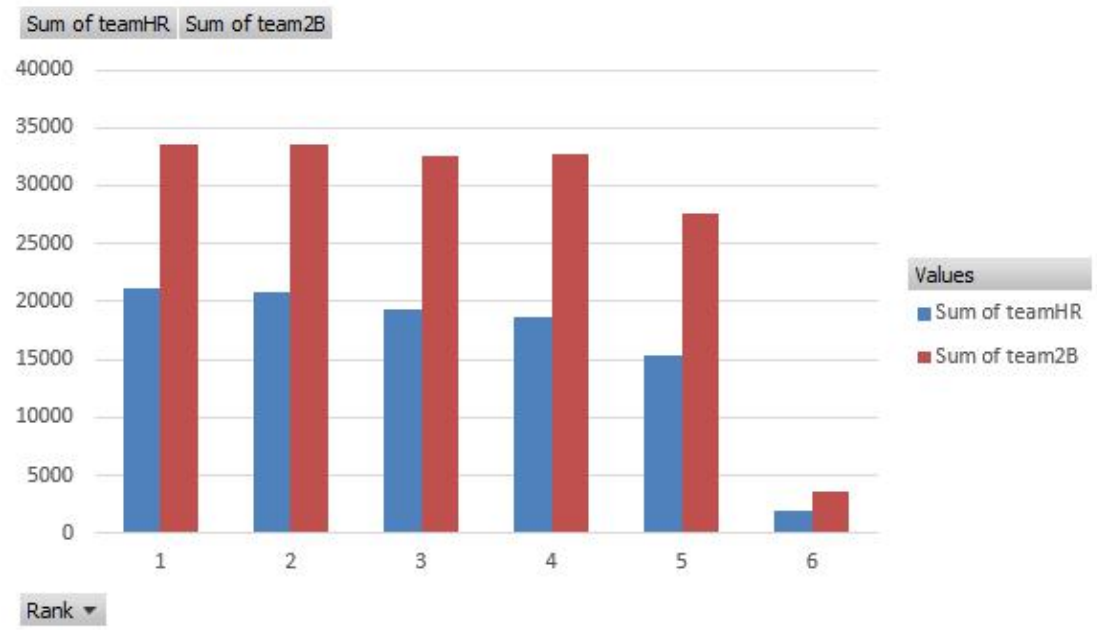


3. Stolen Base



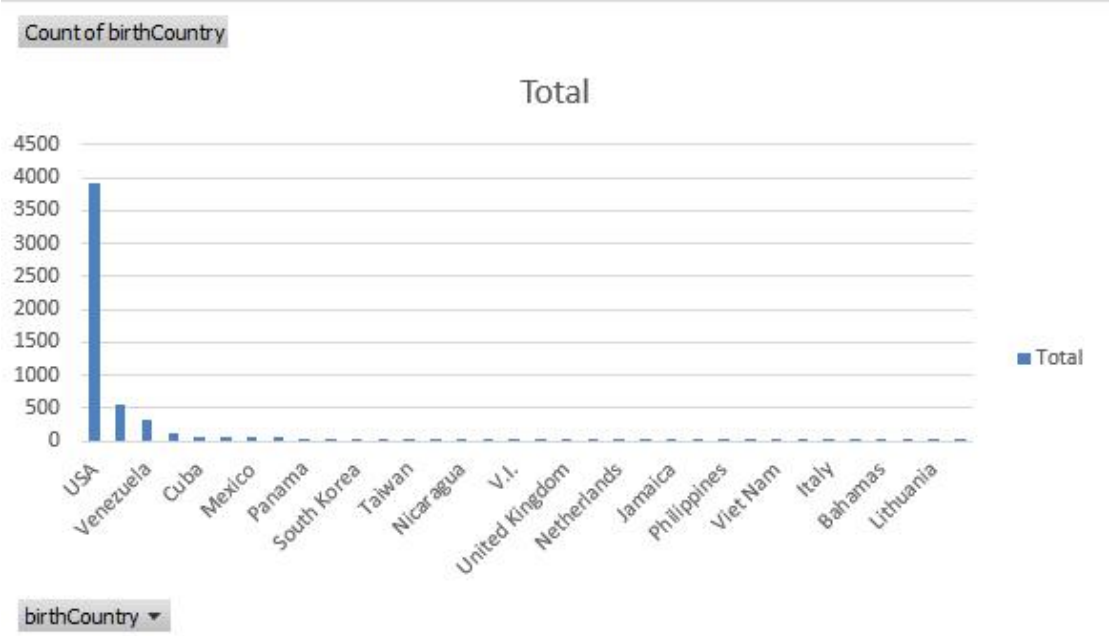
The stolen base and caught stolen have the similar line shape, and decrease from 2011 to 2018, which indirectly related to hitting habit change (prefer single base hit to prefer homerun so that less need of stolen base).

4. Team related analysis



Through those figures, it tells us the HR and 2B have close relationship with team rank, higher team HR and 2B obtain a higher rank (1 is the highest).

5. Player information



The majority of player comes from USA, and players from Venezuela, Cube, and Mexico also have their considerable quantities.

Conclusion

In this project, I mainly discuss the salaries and hitting style change from 2000 to 2018. With this specific visulization, MLB managers are able to deeply understand the competition style and business condition here.

Also, we can do more work towards the lahman baseball database such as fielding analysis, pitching analysis, difference between normal season and post season competition, player trade analysis.

However, only use lahman baseball database cannot meet the demand of analyzing the detailed data like a player's batting data in each game. If time allowed and no device limit, you are able to use Retrosheet (URL: https://www.retrosheet.org/) and Baseball-Reference website (https://www.baseball-reference.com/) to do a more comprehensive analysis.