Data Science Incubator
Project 1

Chris & Hossein

January 15, 2020

Introduction

The data that we are perusing is historical baseball database. The database includes several relations, each of which explains various pieces of information including player salaries and players in teams. It covers almost every numerical datum that exists up to, and including, the 2018 baseball season. For example, in the data, we can see that Zack Greinke was the most payed player on the Arizona Diamondbacks in 2016. For our analysis, we combined several relations that we thought would be interesting to compare. Merging the data and plotting it allowed us to find several interesting observations. After preparing the data and completing initial steps, the results demonstrate that there is a positive association between number salary and H, HR, AB. Our analysis will find associations between player salaries and performance, allowing players a better estimation of what they deserve to be paid.

The presentation slides can be found at:
https://docs.google.com/presentation/d/1jPfyuiHATri6Wd1S1W0JnxsEmEVpZSgY1dw8SALjR2I

Dataset

For our analyses', we used the Lahman's Baseball Database to get complete batting and pitching statistics from 1871 to 2018, plus fielding statistics, standings, team stats, managerial records, post-season data, etc. The data is constrained to the two current major current baseball leagues (American and National), the four other 'major' leagues (American Association, Federal League, Players League, and Union Association), and the National Association of 1871-1875. The dataset was built by Sean Lahman in 1994 and has grown to become the largest and most accurate source for baseball statistics.

The data in contained in multiple CSV (comma-separated-value) files that correspond to a category. For example, there is a CSV pertaining to player salaries. Inside every CSV file, the foremost value in each row is a unique identifier that can be used to link together multiple files. This allows one to link together the 'People.csv' and 'Salaries.csv', for example. A blank space was used to indicate any missing values in the data.

We performed several merges between CSV files in order to link together the data that we wanted to work with. For this, we used inner joins based on player and team ids. We also eliminated anomalies to avoid issues with our analyses. For instance, we removed two players from a team's reported spendings as they both earned $0 for the season.

Analysis Technique

We captured the yearly salary variation of the data by plotting the average salary versus the year. We used the Pandas package to perform a 'groupby' function. This function clusters the years so that we can measure the average of each year in order to compare the overall yearly salary increase or decrease.

Secondly, we merged the salary table and batting table while projecting using the following columns: ['yearID', 'teamID', 'lgID', 'playerID', 'salary','AB', 'H', 'HR']. This time we used the 'groupby' function on 'playerID' while choosing the maximum 'H', 'HR', and 'AB'. We used scatter plots at this stage which resulted in a vague translation of the data. To clarify the results, we then used the 'groupby' function again in order to get the average salary on each value of 'H' (and similarly 'HR' and 'AB').

Lastly, we compared a team's budget versus their win-rate by merging the player salary data with the team record data. We then merged rows based on yearID and teamID so that we could account for team budgets greatly increasing or decreasing depending on the year. We then found every player that played on a team for a year and

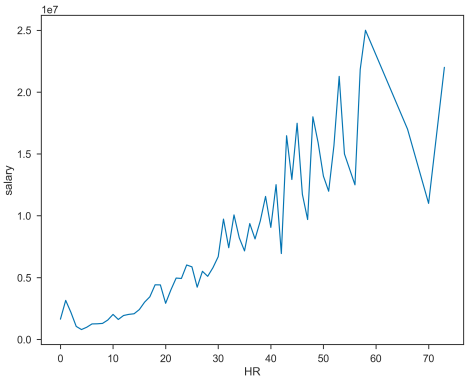
put them into a list corresponding to that team/year. Finally, we took the mean, standard deviation, and largest margin of difference between the salaried players for each team/year.

The 'merge' function was used to bring all the necessary information into one table. The 'groupby' function was used to reveal clear and observable trends in the data. The maximums were used for player evaluation since the performance of a player may degrade at the end of his career and their mean stats will not capture his/her 'glory' days. Line plots were chosen as they provide clearer visualization compared to scatter plots.

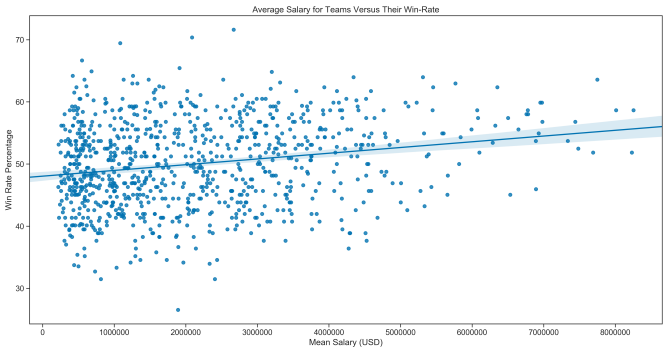
Results

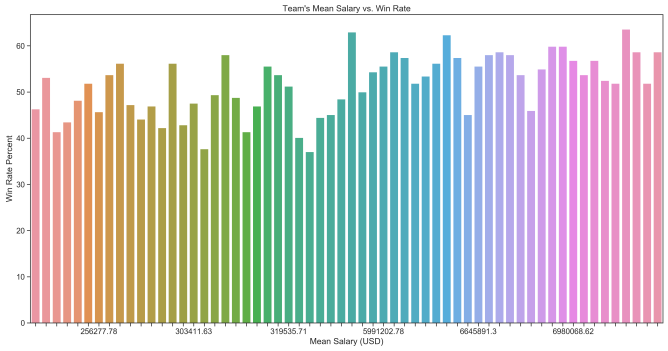
We developed our project to answer the following questions: 1) does the number of hits impact the salary of a player; 2) does the number of home runs impact the salary of a player; 3) does the number of at bats impact the salary of a player; and 4) does the average salary on a team correlate to a higher team winning ratio?

We found that the average value of salary has a not so smooth and clear positive association with the number of maximum hits in a player's lifetime. There are several anomalies between 0 and 25 hits. There are some players that did not have a high number of hits while having better salaries compared to players with better hits. These players may be just be good pitchers who will be replaced by designated pitchers so they would have a lower number of hits, home-runs, and at-bats. The same association can be observed between the salary and the number of maximum home-runs (and similarly between the salary and the number of maximum at-bats) in a player's lifetime. The salary of a player may increase as his number of hits and home runs increases. For players that may not hit often, other factors may come into play which needs further analyses.



For the average team salary versus win-rate, we found a small positive trend consistent across mean salary and smallest standard deviation of salaries. Additionally, we found teams that had a large pay gap between their top two most payed players would perform worse than the teams who had smaller gaps between top paid players. We cannot conclude from this that there exists a causation between the two, but from the data, we can see a clear, albeit small, trend that indicates having a well-payed team without large pay gaps leads to the highest win-rate ratio. There are several anomalies in the data, notably around the $2.5mil range.





The analyses used in this project was suitable (not comprehensive) because, first of all, it answered our research questions; and second, it used very simple plots that can be illustrative to everyone. The results we found can provide an estimation of worth for players determining what their salary should be. Also, our results can be used to argue that spending an entire team's budget on one or two players is not worth it as opposed to distributing the budget for an overall better team. In addition to our work, we could have alternatively looked at individual players to see how their individual performance and salary changed over the years as opposed to the league's average.