Michael Hanks

Project 1 Report

1/19/2024

Presentation Slides: Project 1 Baseball Dataset ;

https://docs.google.com/presentation/d/1dSeaFJR-Bnlv4Yr5-1cm4azWlF_L-crHpgHRxCxz00A/edit?usp=sharing

Project 1 Folder: https://github.com/MichaelPHanks/CS5830_project1

In baseball, a lot of teams work on signing players to contracts that fit the players ability and skills. All things considered, managers will have to decide whether to sign players to massive contracts or not. This database consists of data from American and National leagues, as well as the American Association, Union Association, Players League, Federal League, and finally the National Association. We discuss the benefits and causes for players making more money, and how teams can possibly decide whether to spend big money or not. We combine years of data including salaries, batting, pitching, fielding, and teams to gather intel about how salaries relate to success.

The datasets used include players and their batting, pitching, fielding, and salary statistics that tell the story about how they have made money in previous years. Another dataset used is the teams dataset, which gives key information per year per team. Keep in mind that because the salary dataset starts in 1985, the rest of the datasets are trimmed up to that year. Each of these datasets contain a player ID to distinguish players and a year ID to distinguish years. These datasets include (on a per year per player basis) hits, strikeouts, salary, putouts, assists, and errors.

The main goal of these datasets was to compare salaries with the number of hits, strikeouts, and fielding statistics. These stats are part of the game of baseball, and would have a large impact on a game. First, we pulled the salary dataset, and combined it with each of the other statistics, to give a per player per year table with the stats needed. This also includes the teams dataset, which contains the amount of money spent per team per year (since 1985). Because we got the data per player per year, with their salary, we can get a good representation of the difference it makes to statistics for that year. The more money they make, the better they do in statistical categories.

When it came to the results, they were for the most part as expected. Keep in mind that these results are based on specific stats and salaries, which vary between years and positions. Inflation makes it so teams pay a higher salary in years near the end, while some positions in fielding will hit the ball less. When it comes to hits and salary, there is a slight increase in hits when salary increases, with a few outliers, which is shown in figure 1. Strikeouts and salary have a very small but noticeable correlation (figure 2). The main source of my argument definitely comes from the fielding table though, as there is a clear and concise correlation between those stats and salary (figure 3). All of these results are based on the last dataset, which also shows a correlation in winning percentage and money spent per team (figure 4). I also included a smaller scope of data for this, from 2010 to 2016 (figure 5).

Given these results, team managers should really consider signing players to large contracts if need be, especially if they excel in hitting, strikeouts, or fielding percentage. This will correlate to winning games, which is what managers want, right?

To get the dataset ready for analysis, I had to trim all of the data for the pitching, batting, fielding, and teams to after 1984 (starting in 1985) because the salary dataset started in 1985. I had to merge each dataset based on the player id and year, so we could get a good idea per year how good they did in statistics and how much they made. To get team salaries, I just added each player's salary from each team per year, and put it in one table. In terms of the analysis, a line graph can show each players salary with their statistics to give a good idea if salary and statistics have a good or bad correlation.

At first I was thinking of only doing player statistics and winning, however, each player does not have their own winning percentage, so I had to go based off of team win percentage. This ended up being good as it is specific to managers and winning games as a team.

Figure 1:

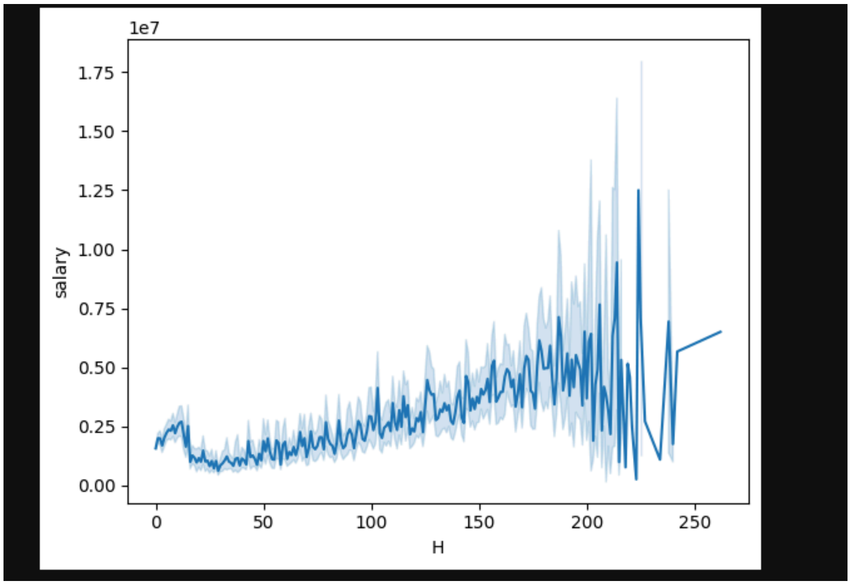


Figure 2:

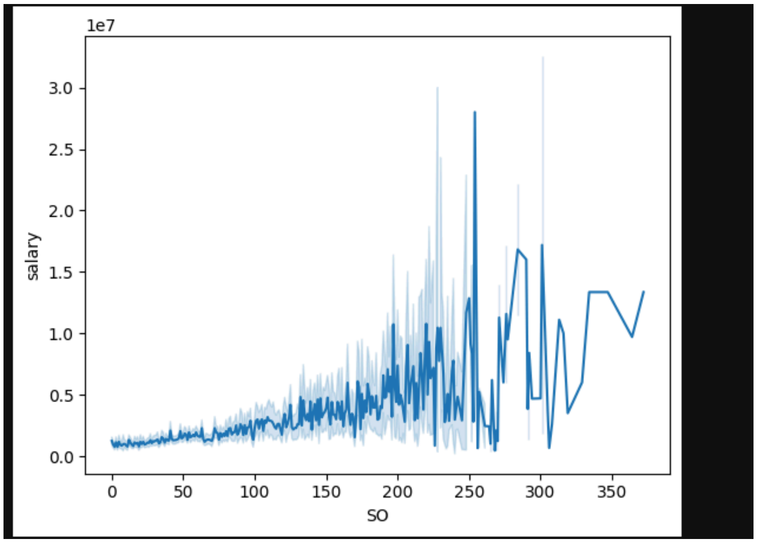


Figure 3:

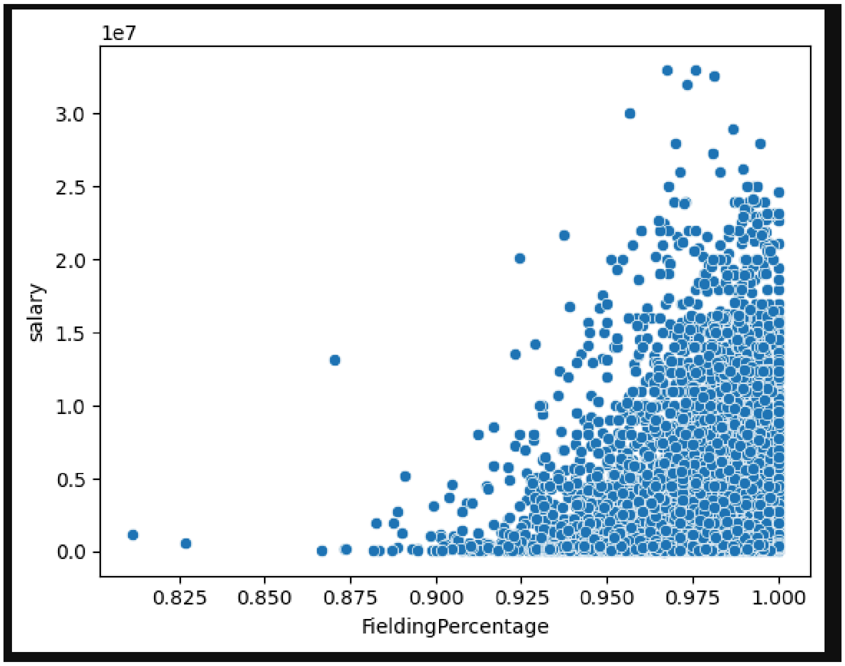


Figure 4:

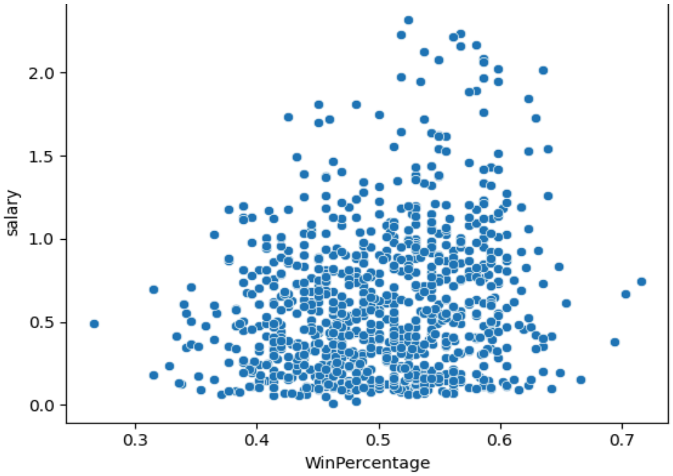
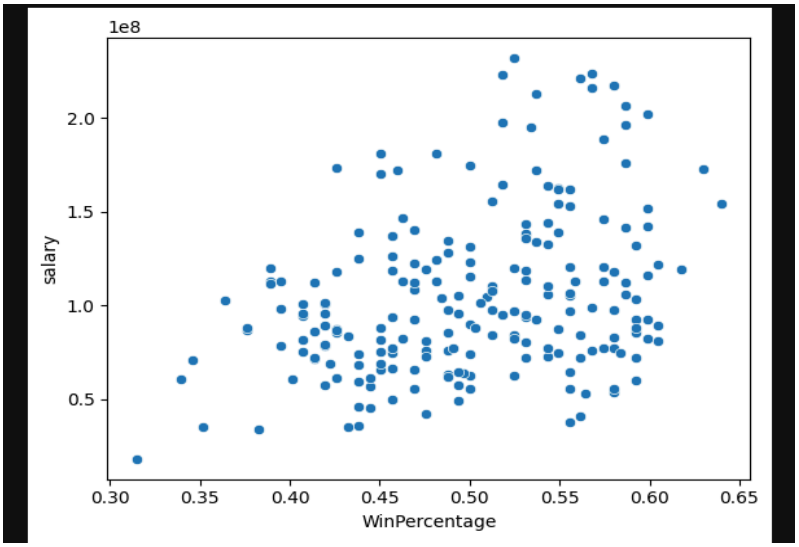


Figure 5:


1e8
salary
WinPercentage