CS5830 Group 15

Project 1 Report

Introduction

The data analysis described in this report covers just a small variety of the data that can be found in the Lahman's Baseball Database. There are more than 25 different data sets included within the database, some of which include information about players, batting, fielding, teams, salaries, All Stars, management, and much more.

Our analysis focuses on salaries, base stealing stats, All Stars, and retirement all of which contain some interesting and potentially useful information. Through graphing various relationships between data points we were able to gather that teams with a high contribution to the All Star teams also had the highest salaries. Additionally, a large majority of the attempts made to steal bases are successful.

Dataset

In the first analysis, we used the Salaries.csv table to find out the average salary for a baseball player on a given team. The table columns of "salary" and "teamID" were used to determine the statistics. Our second analysis depicts the relation between number of bases successfully stolen and the number of times caught stealing bases for each player. This information was acquired using the "playerID", "SB" (stolen bases), and "CS" (caught stealing) columns from the Batting.csv file. For the third analysis, we used the AllstarFull.csv file. Each year there are 64 major league players chosen to play in an all-star game. We used the columns of "yearID" and "teamID" to find the percentage of players from each team that were selected for the all-stars teams in 20 year sections. For our final analysis, we found the year that each player retired based on the "finalGame" column of the People.csv table.

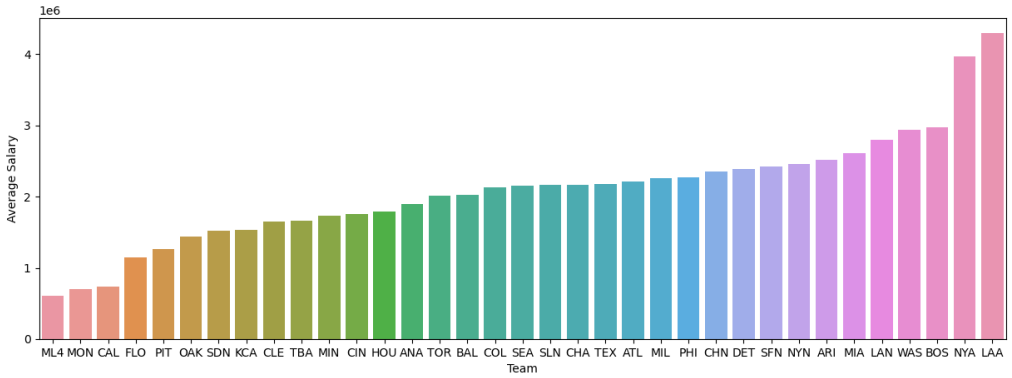
Analysis Technique

Graphing analytics were used to display relationships between the data. When there is a significant amount of data to be analyzed, like in cases such as ours, graphs provide a great platform to both visualize and analyze such data. With graphs we were able to be as general or as specific as necessary with our analyses. We included various graphs such as categorical bar charts, scatter plots, and line plots that were all chosen as we saw fit to represent each problem/question appropriately.

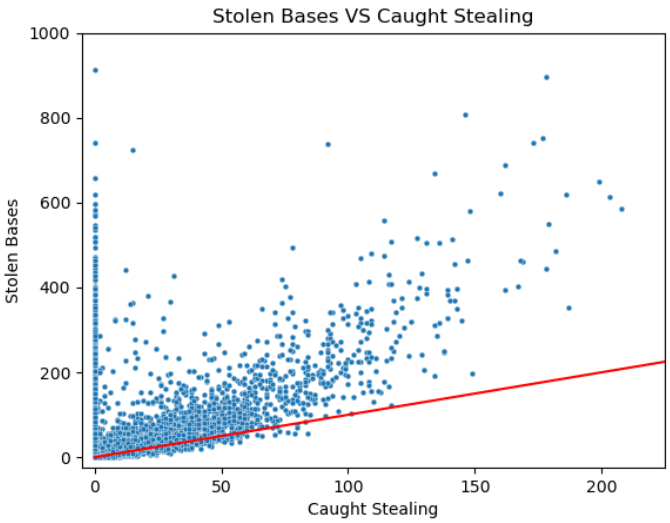
Results

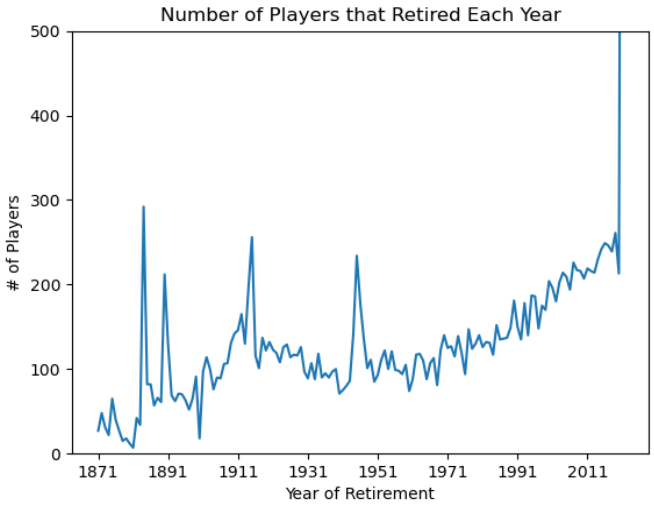
The first analysis (below) shows that the teams with the highest average salaries are LAA, NYA, and BOS. The third analysis (seen in 6 figures at start of next page)

shows that, for each 20 year segment between 1920 and 2019, NYA had more players play in the all-stars games than any other team. In 2020 and later, BOS had more all-stars than any other team. There appears to be a correlation that indicates that when a team has more players that play in the all-stars, the average salary of their players is higher. This may indicate that the all-stars get higher salaries.

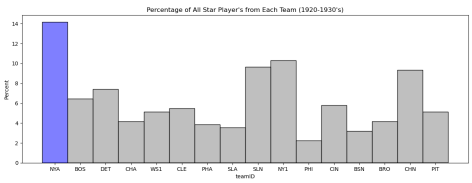


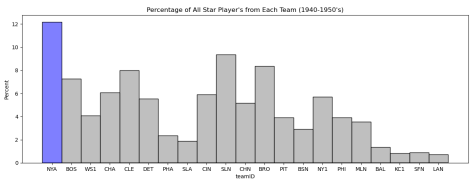
The second analysis (right) depicts the ratio of times a player stole a base to the times they were caught stealing. The red line indicates a 1 to 1 ratio. Dots above the line indicate that the player stole more bases than they were caught. There is also a vertical line of data points around x=0 indicating that there are players that have successfully stolen bases without ever being caught. This may lead a coach to encourage their players to steal more bases.

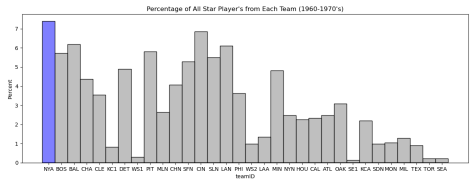


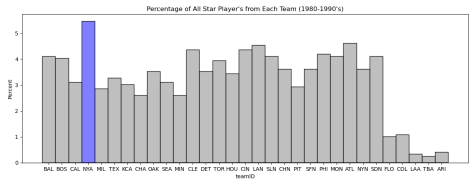


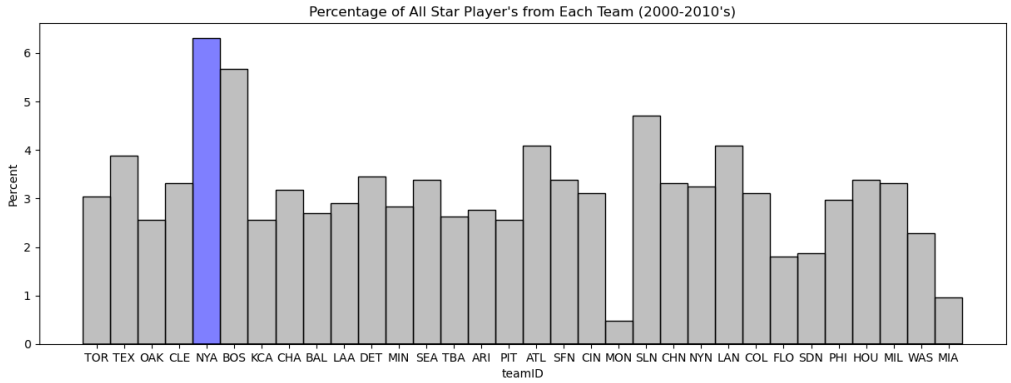
Our final analysis (left) shows how many players retired each year. There are four main spikes that correlate to WWI, WWII, the first year pitchers were allowed to pitch overhand, and the year a league folded.

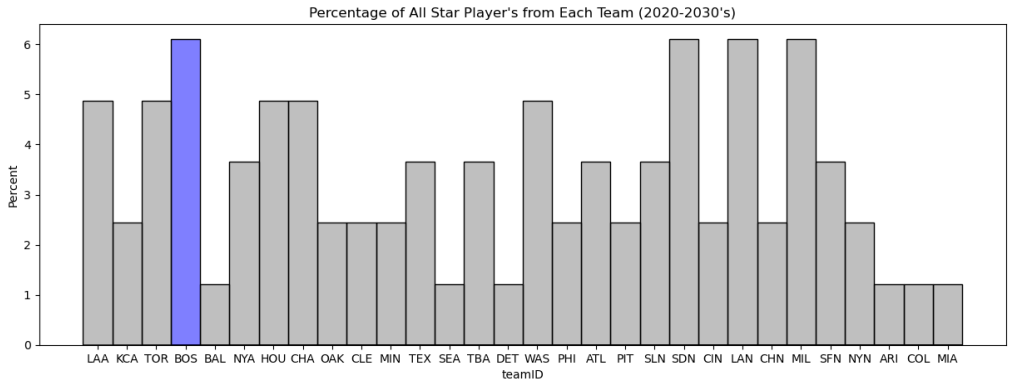












Technical

In preparation for each analysis we read only the necessary data into a dataframe object. From there it was up to us to properly group, aggregate, and/or perform any other actions on the data fields we were looking at before plotting them against each other. Most of our graphs provided us with a simple way to interpret the data and come up with explanations for our findings. Additionally, by comparing them to one another we came up with new findings.

In one instance, we struggled to interpret one of the graphs. However, by adding a line of equality to the “Stolen Bases VS Caught Stealing” scatter plot it was easy to see that the majority of players have great success when stealing bases.

Project Links

Github Repository - https://github.com/jakepper/cs5830-group15-project01

Presentation - https://docs.google.com/presentation/d/1D-vwbYdlld7CzmLD_uONItDPreHvuOU2t_ZH9R-p_hI/edit#slide=id.p