Data Science Incubator
Project 1: Suyash Mhetre, Joseph Turcotte, Colby Wight

**Introduction**

Baseball has been referred to as the great american pastime. In this project our goal is to explore various aspects of baseball teams and players in Major League Baseball. The dataset we used has a lot of information about teams and players and their performance over the years. Our analysis focuses on certain teams performance over the years as well as team salaries and individual player salaries. We also explore the performance of different pitchers over the years. For analysis we used basic functions in the Pandas library and basic data manipulation techniques to subset the data based on certain features of interest. In our analysis we will learn the importance of team performance to game attendance which is directly related to and of the utmost importance to the financial success of a franchise. We will also see that this can lead to higher salaries and better players.
Slides link:
https://docs.google.com/presentation/d/1qR7ttFtoQaAXEwNGdBC2ehxG4nnN35sAiB7U0TSDn3g/edit?ts=5e1e40a4#slide=id.g2097b76b5493da3f_407

**Dataset**

Baseball is a popular team sport in the United States. In this sport one team takes the "field" and tries to get outs on the other team before they make it around the bases and score points. There are different positions in the field. Infield positions such as pitcher, short-stop, and second base, and outfield positions such as center fielder and left fielder. The other team picks players to try to hit the baseball.

The dataset can be found at http://www.seanlahman.com/baseball-archive/statistics. There are multiple tables in the form of comma separated values (csv). These tables deal with many of the statistics for both how players do in the field as well as hitting statistics, pitcher statistics and team and franchise statistics. These tables can be used together using different IDs such as player ID and team ID. The tables we used are the following: "Pitching.csv", "People.csv", "Teams.csv", and "Salaries.csv". The features include: "playerID", "nameFirst", "nameLast", "teamID", "ERA", "yearID". Variables for the team data include: "yearID", "teamID", "G" for the games in a year, "W" for wins in a year, "L" for the losses in a year, "name" the name of the team, "park" the name of the baseball stadium that year, and "attendance".

**Analysis Technique**

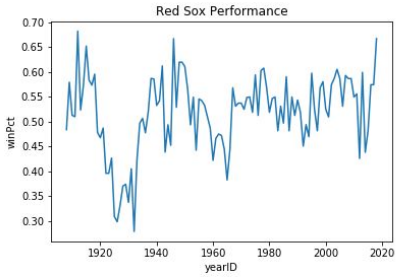
In order to get the information we wanted for the Red Sox. We first needed to read in the csv file using the Pandas read_csv function to read in the team data. There was more information here than we needed so we first removed the columns that we did not need and then we created a new dataframe that only contained the team Red Sox using subsetting techniques. In order to show the performance we decided on calculating the percentage of games a team won that seasonse. Since there were a different number of games played in seasons, we could not simply take the games won. We used the technique in Pandas to create a new column called win percent and calculated it by taking the rows games won value and

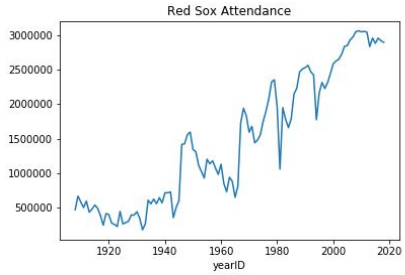
dividing it by the total games value. We then used the seaborn lineplot to easily plot both attendance over the years and performance over the years in the form of win percentage.

The analysis of pitchers used many of the same techniques. In addition, since two tables were used the Pandas merge function was used to get the players names of the pitchers by merging on playID. These simple techniques were sufficient for our purposes here of exploring the datasets. We were able to find meaningful insights just by subsetting the data and creating simple plots.

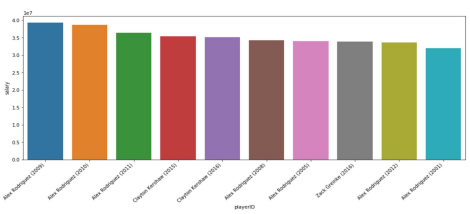
**Results**

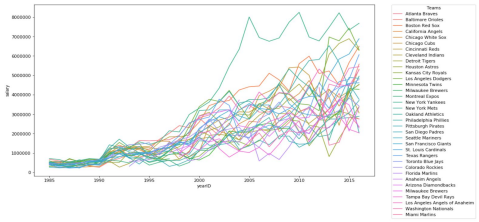
These were the plots we got from the Red Sox performance analysis:




Here we were able to see that in the 20's the Red Sox were not winning as many games. And in the early 30's they abruptly changed and started winning a lot more. What may have spurred this sudden change? Could it have been building a new stadium or the hiring of a new manager? Maybe it was the signing of a great player or an external factor. One thing is certain: that the performance and attendance are correlated. We see a general upward trend in attendance over the years, but years where performance was not as good as in the 20's were also years that attendance did not increase.





The charts above show an analysis regarding salaries. The one on the left shows a bar graph of the salaries of individual players for a given year. We see that only a few players show up here, indicating that they got paid the most multiple years. Are they really worth that big of a pay gap? The chart on the right shows a line graph with the total salaries a team paid its players by year. It's clear that there is one team that in recent decades has paid much more than the other teams. That team is the New York Yankees.