MLB Analysis - Project Draft 1

Code

Group 6: Nick Keohan, Harrison Flynn, and Shaun Harrington

March 28, 2021

Data from this analysis is thanks to Sean Lahman. The R package ‘Lahman’ is used to load this data. For more information see the documentation at http://www.seanlahman.com/files/database/readme2017.txt (http://www.seanlahman.com/files/database/readme2017.txt).

The data will be formatted and visualized by means of the ‘tidyverse’ and ‘ggplot2’ R packages and Tableau.

Code

Top Team Managers

We’ll take a look at the most successful team managers since 1884. This will be done by looking at most wins and most World Series Championships.

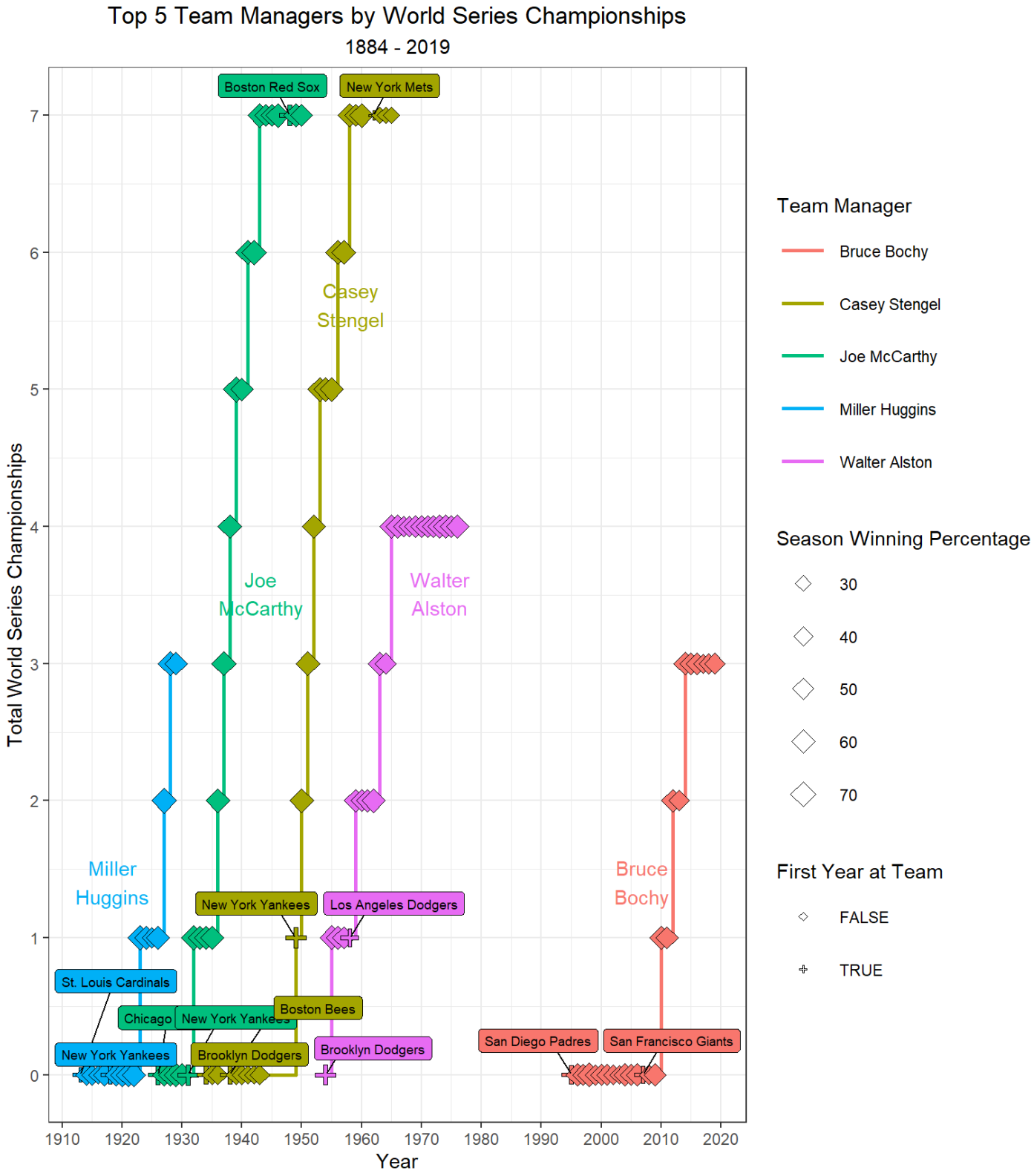
By manipulating the Managers and Teams tables in the ‘Lahman’ package, we can get our dataset started.

Code

Determining the performance of managers can be from numerous perspectives. We'll begin with the metric used to determine the best team each year: World Series Championships.

Below are the top 5 coaches by World Series Championships. Showing the cumulative championships by manager, irrespective of the team that they managed.

Code

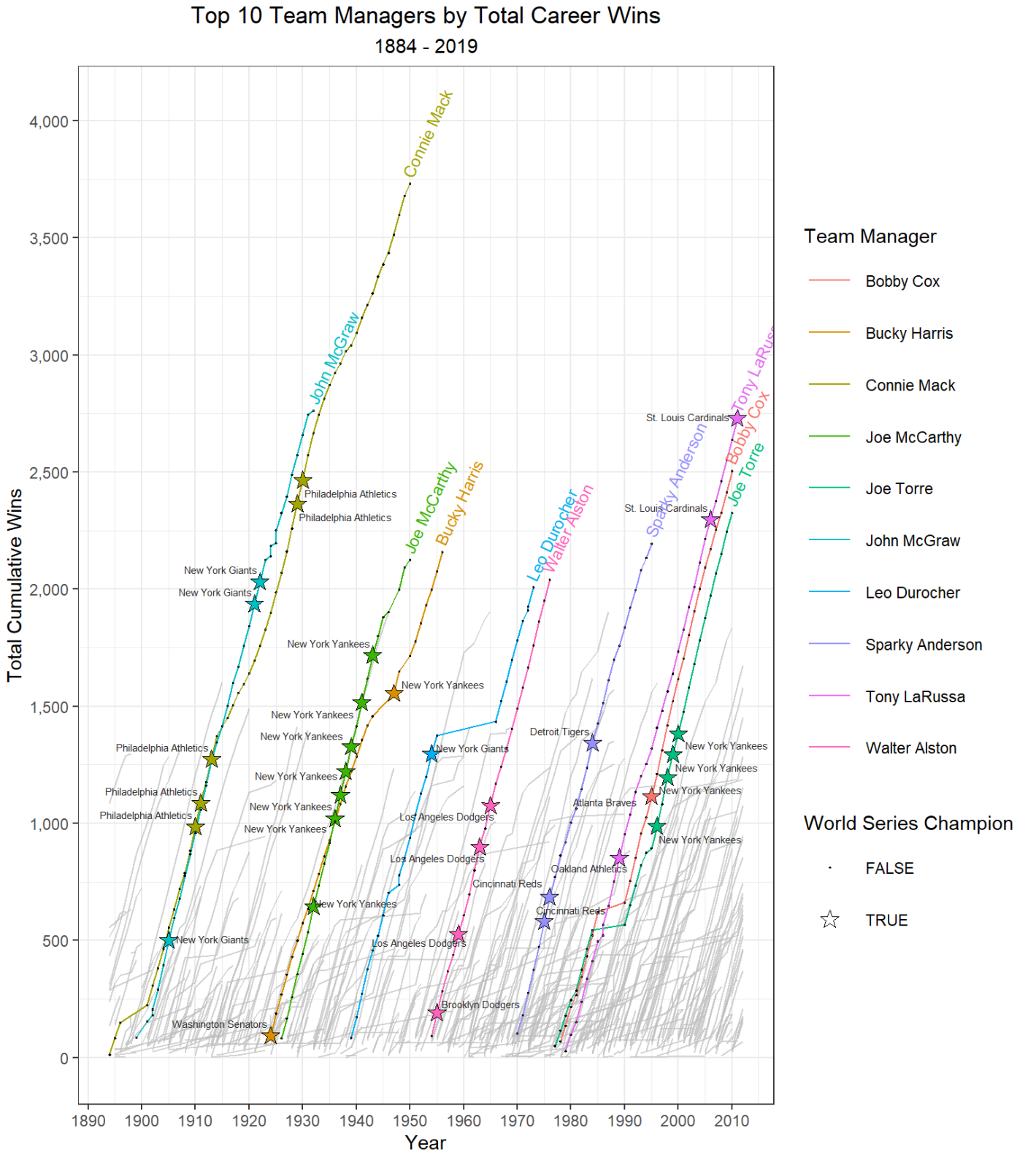


Another perspective is to consider the total number of wins a manager is able to lead their teams to.

Here is the code to prep this dataset:

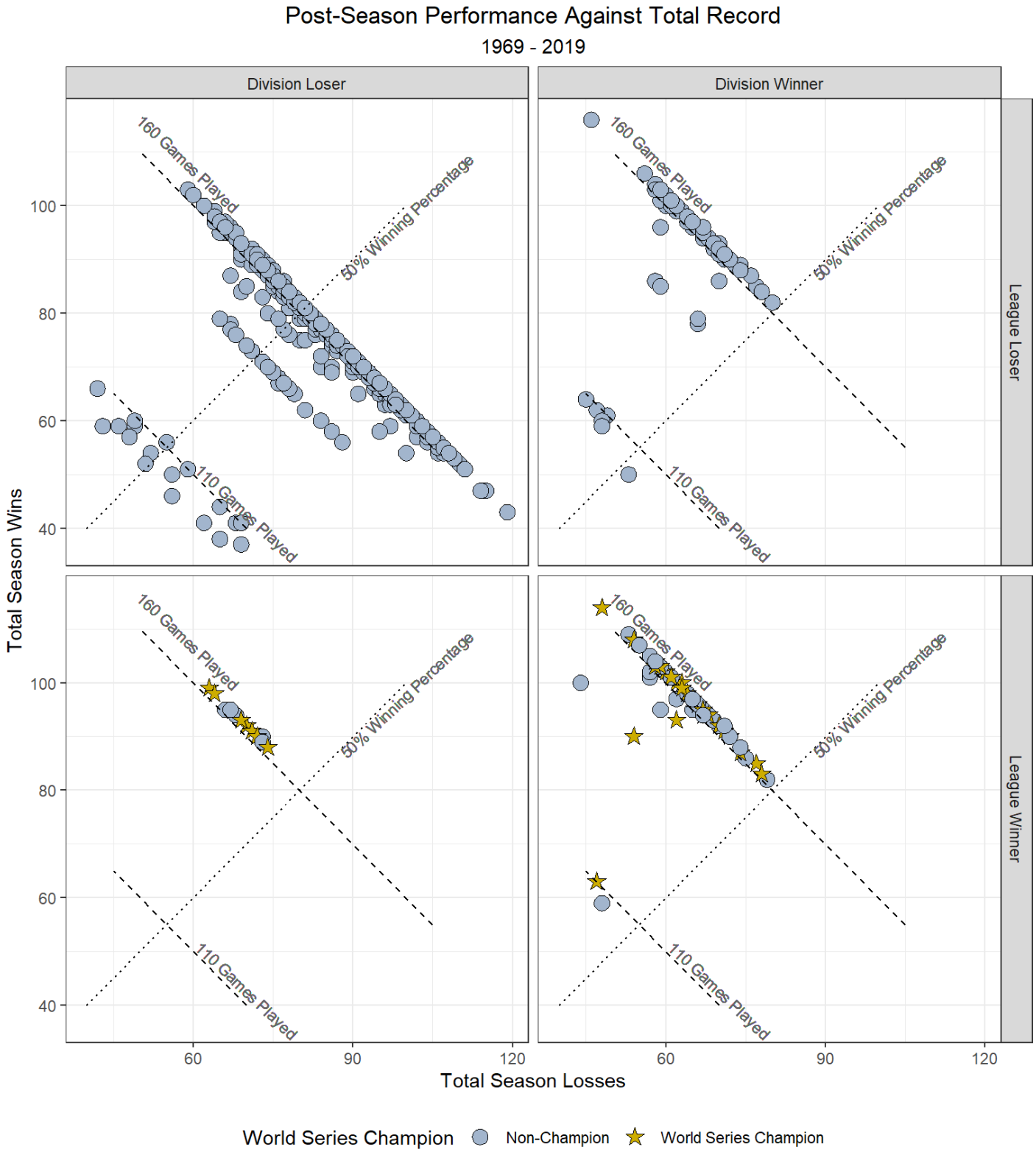
Code

Code



Regular vs. Post Season Performance

Code



Nick Keohan

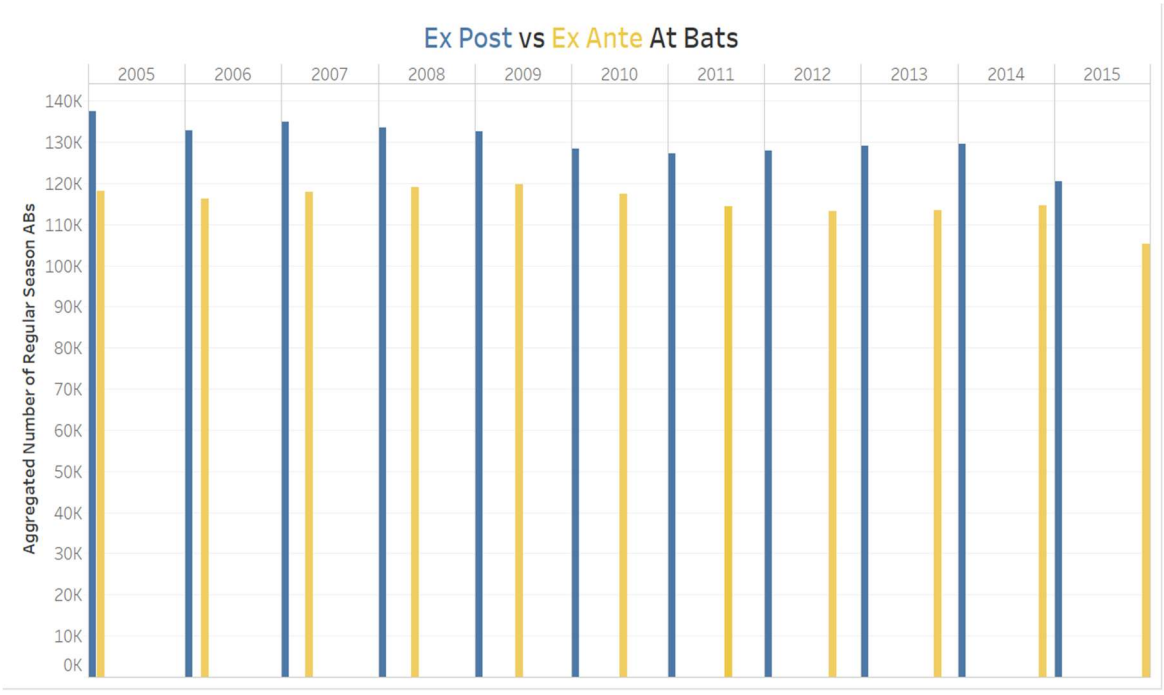
Project Draft: Outline of Ex Post vs Ex Ante Batting Metrics and

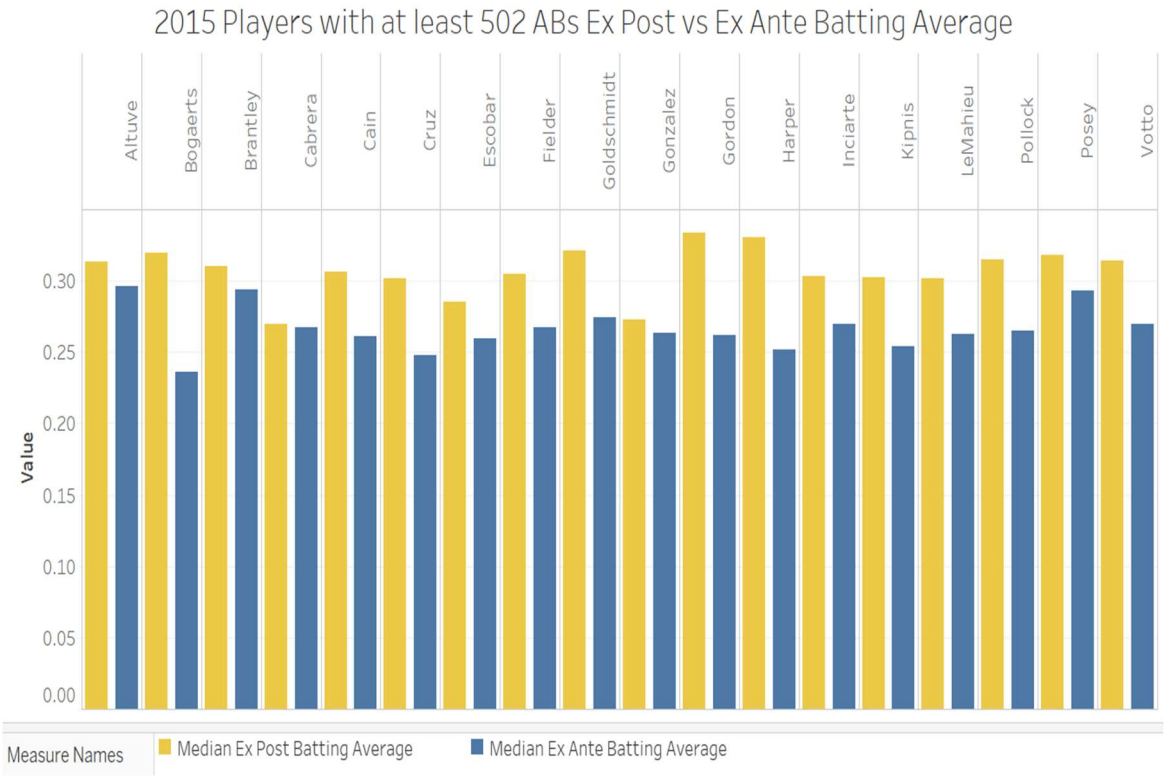
The Impact on Defensive Metrics of the Left Handed Batter Shift

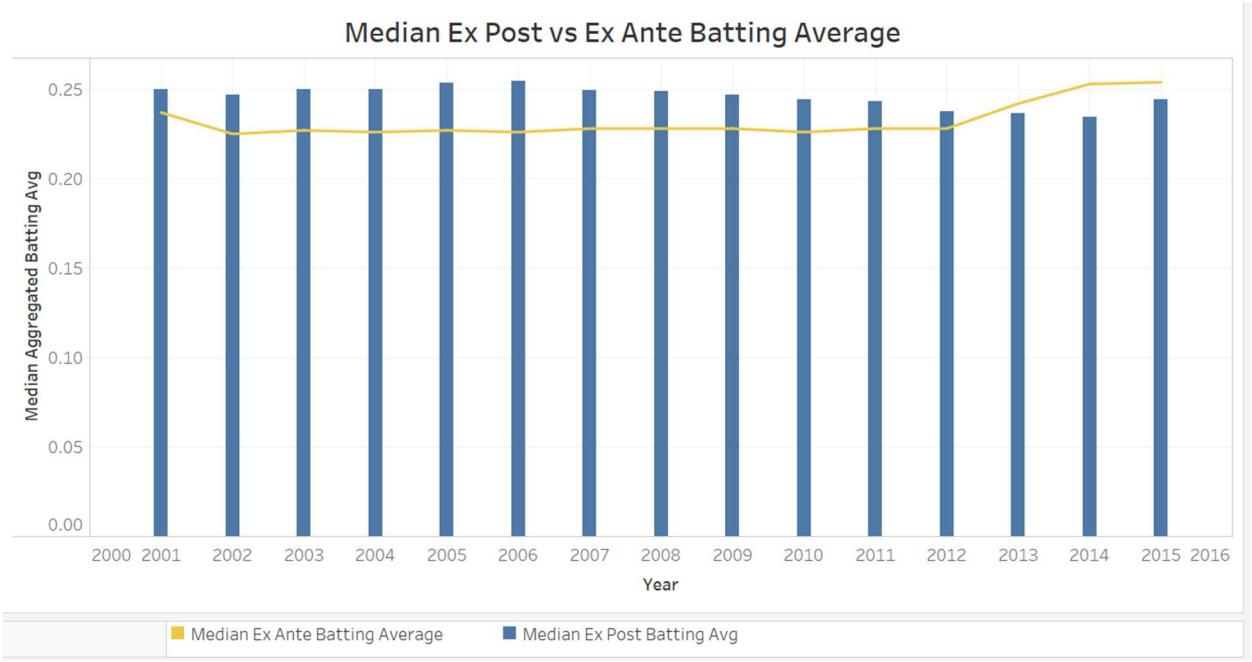
Data Overview:

The projected or ex ante batting statistics were obtained from baseballguru.com and span from 2005 through 2020. The data sets contains player level data with the variables of interest mAB (Forecasted at bats), mAVG (forecasted batting average), mH (forecasted hits), mOBP(forecasted on base percentage), and mSLG (forecasted slugging percentage). The data sets from baseballguru.com also contains the same player identification as in Sean Lahman's datasets allowing for an easy merge of data. These variables will be compared to the player level data obtained from the Sean Lahman data set. The Lahman data set contains the variables of interest AB (at bats), H (hits), 1B (singles), 2B (doubles), 3B (triples), HR (homeruns), HBP (hit by pitch), SF (sacrifice fly) and BB (walks). These extra variables are required from the Lahman data set because slugging percentage and on base percentage are not explicitly supplied. Slugging Percentage = (1B + 2B*2 + 3B*3 + HR*4)/AB and On Base Percentage = (H + BB + HBP)/(AB + BB + HBP + SF).

In order to adeptly analyze ex post verse ex ante baseball offensive statistics, special attention will be placed on analyzing the at bats (AB) statistics. The metric AB is considered one of the most difficult of metric to predict in baseball due to its multitude of confounding factors. This metric is also a key component of all forecasted batting statistics making its impact on the overall batting predictions extensive. By observing the accuracy of these ex ante verse ex post AB statistics, an estimation of the overall accuracy of forecasted batting statistics can be obtained.







Based on the preliminary bird's eye view results for some of the variables of interest, it appears that forecasts league wide, have historically under-estimated batting averages. This coincides with the fact that aggregated forecasts of At Bats have under shot the observed aggregation of

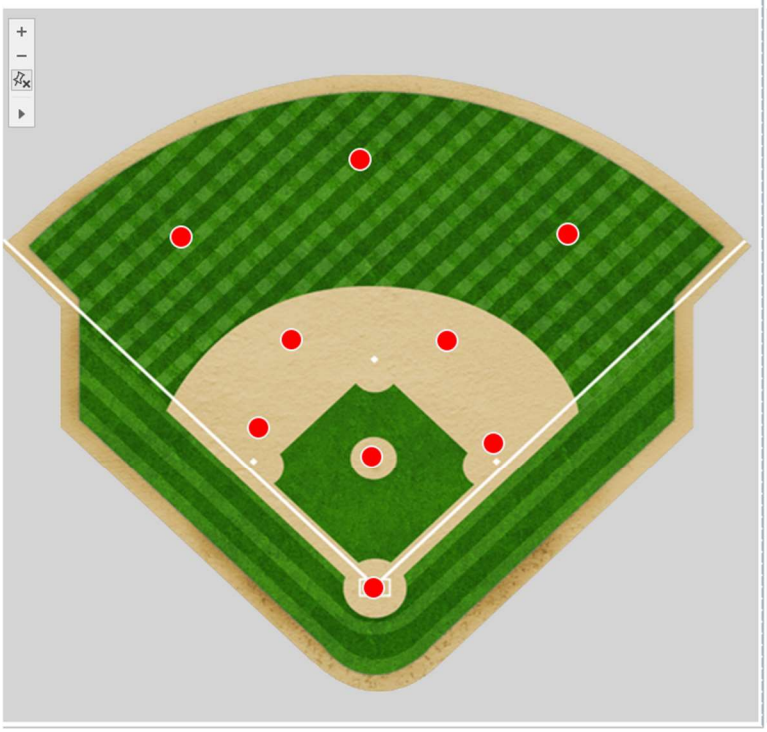
At Bats. Further research into the number of players these forecasts cover as well as the number of player the Lahman data set contains are required to draw any significant conclusions from these visuals.

In order to study the impacts on defensive metrics from baseball teams implementing a shift, the variables PO (putouts or when a defensive player records an out), A (assists or when a fielder touches the ball before a putout is recorded by another fielder), and E (errors or when a fielder fails to convert on a play that an average fielder should have made) are of significance. These variables will be observed from 2000 to 2019 with special attention given to the Baltimore Orioles, Tampa Bay Rays, and Houston Astros. These teams are considered the perennial leaders in implementing defensive shifts, with the vast majority of these shifts occurring with left handed batters.

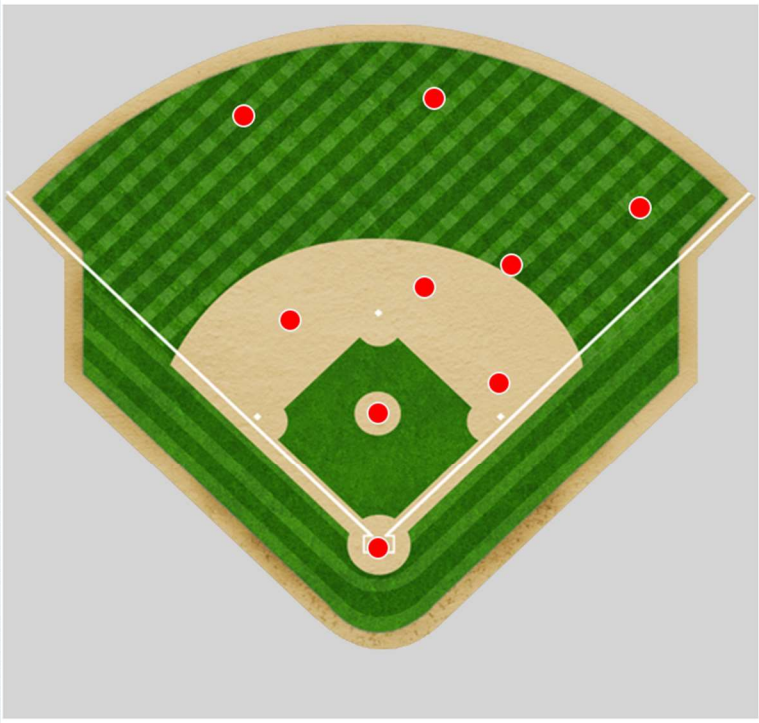
While the shift has been around since the early 1920s, its prevalence has drastically increased in recent decades. Prior to the Rays popularized use of it in 2006, the shift was only used for elite powerful left-handed hitters that had a strong tendency to pull the ball. In 2016, the Houston Astros performed a defensive shift on 52% of the left handed batters the team faced. In the same year, the Rays implemented the shift for 35% of the batters they faced. In 2019, the Houston Astros preformed defensive shifts on 77.% of At Bats, the Orioles 62.3% of At Bats, and the Rays 44.1% of At Bats against lefties.

The reference lines on each of the team graphs indicates the year in which the respective teams began significantly implementing the leftie defensive shift into their playbooks. Based on a preliminary analysis of defensive metrics for the three forementioned teams, it appears that each team experienced a decrease in the number of errors they committed throughout a season. This trend is most apparent with the Houston Astros. The number of assists also appears to make a steady decline after each teams' implementation. This could be attributed to more line drive or fly balls being caught due to the shift. This decline in assists also coincides with a slight increase in putouts for each team, backing up the claim that the shift has led to an increase in the number of fly balls or line drives being caught.

Standard Defensive Positioning

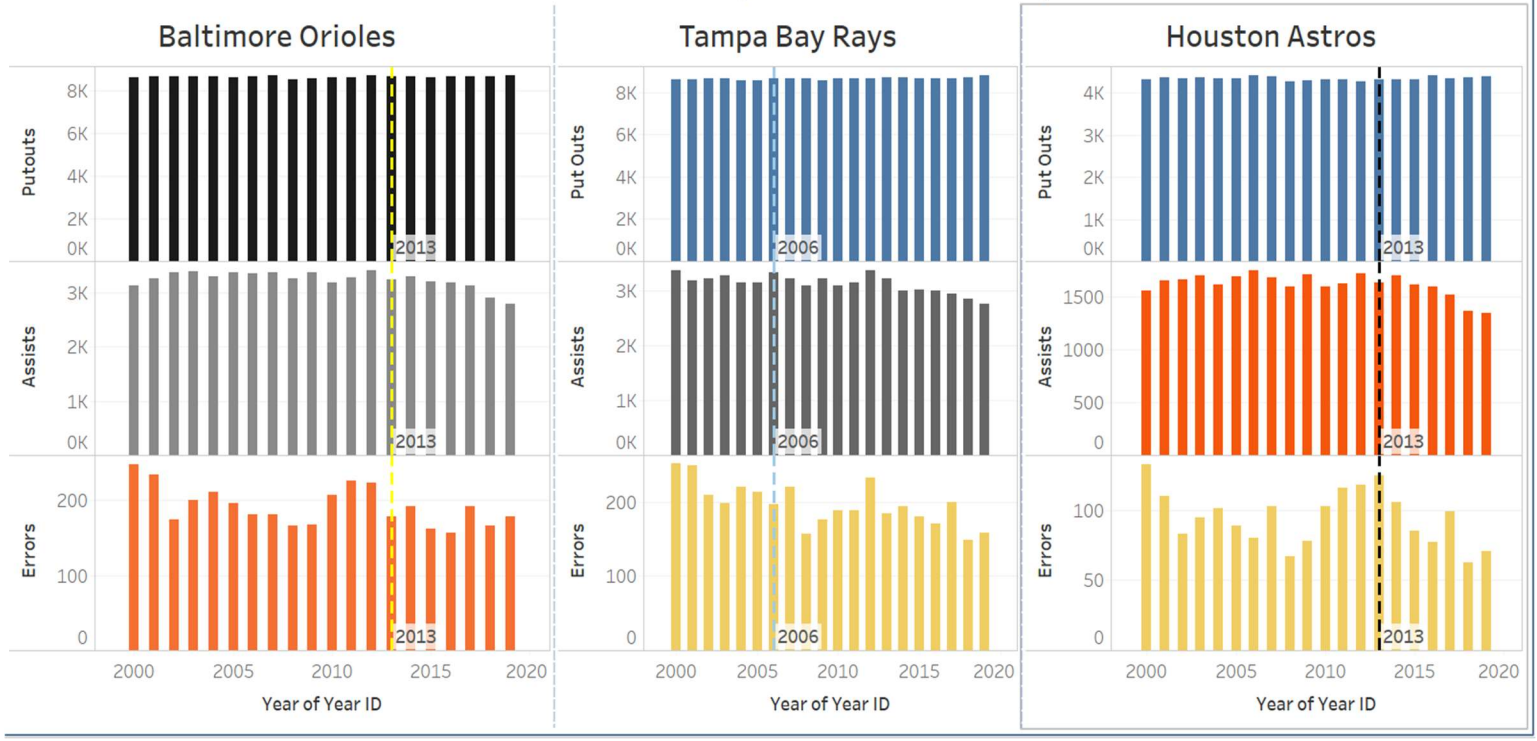


Left Handed Batter Shift

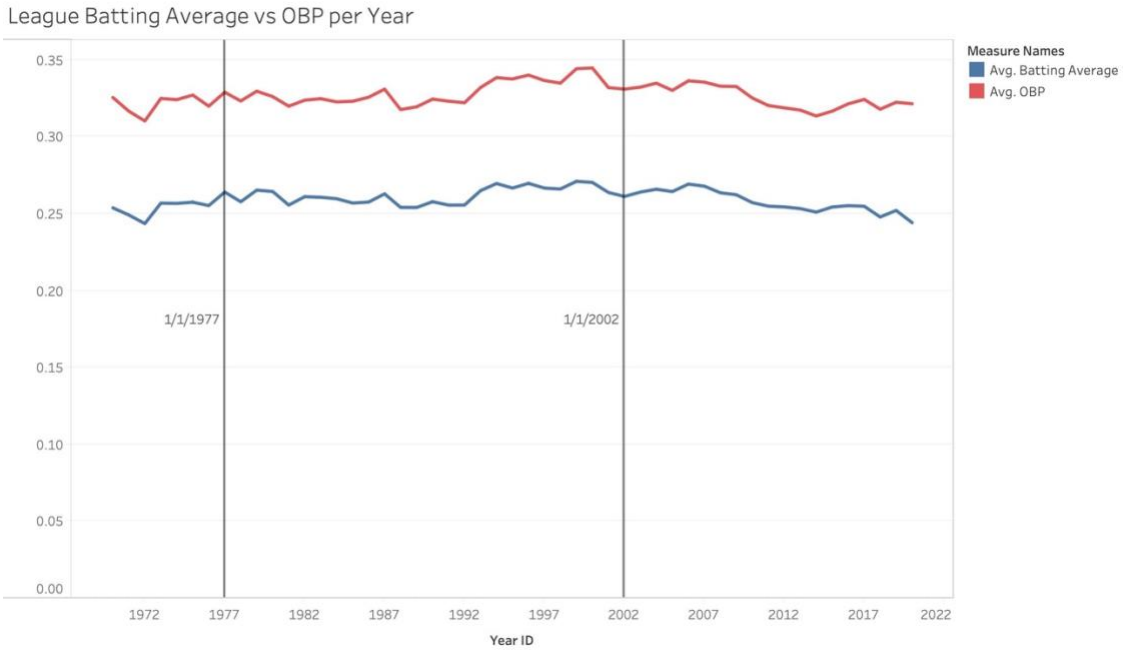


Fielding Metrics For the Top 3 Adopters of the Defensive Shift

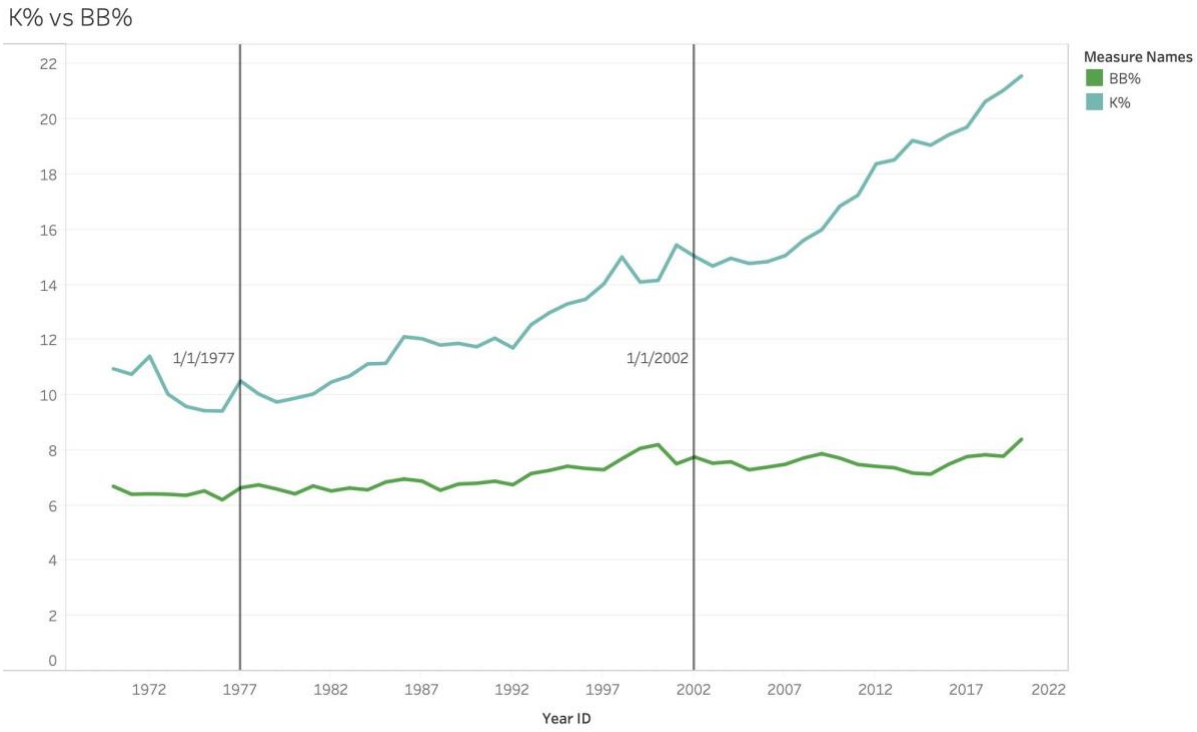
reference lines Indicate the year the leftie shift was adopted



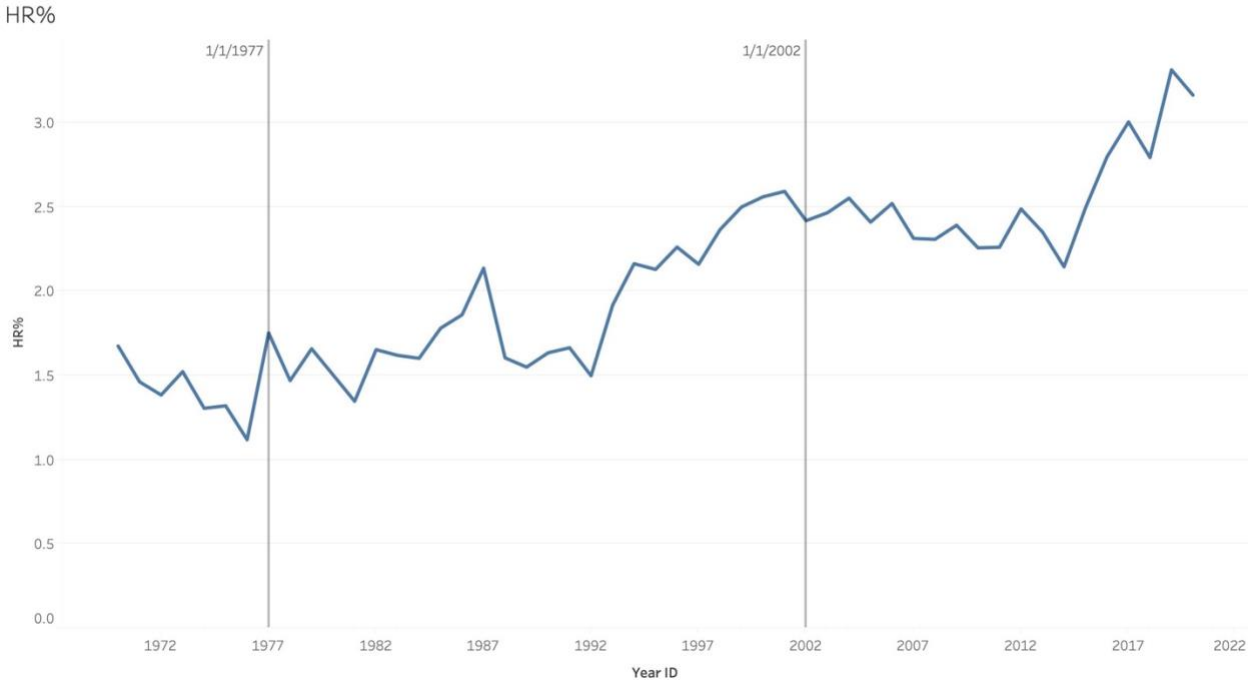
The following charts show league statistical trends from 1970-2020. The years 1977 and 2002 are marked off to indicate the years Bill James introduced his Baseball Abstract in 1977, which started the sabermetrics revolution in baseball, and the year the Oakland Athletics gained notoriety for their Moneyball tactics.



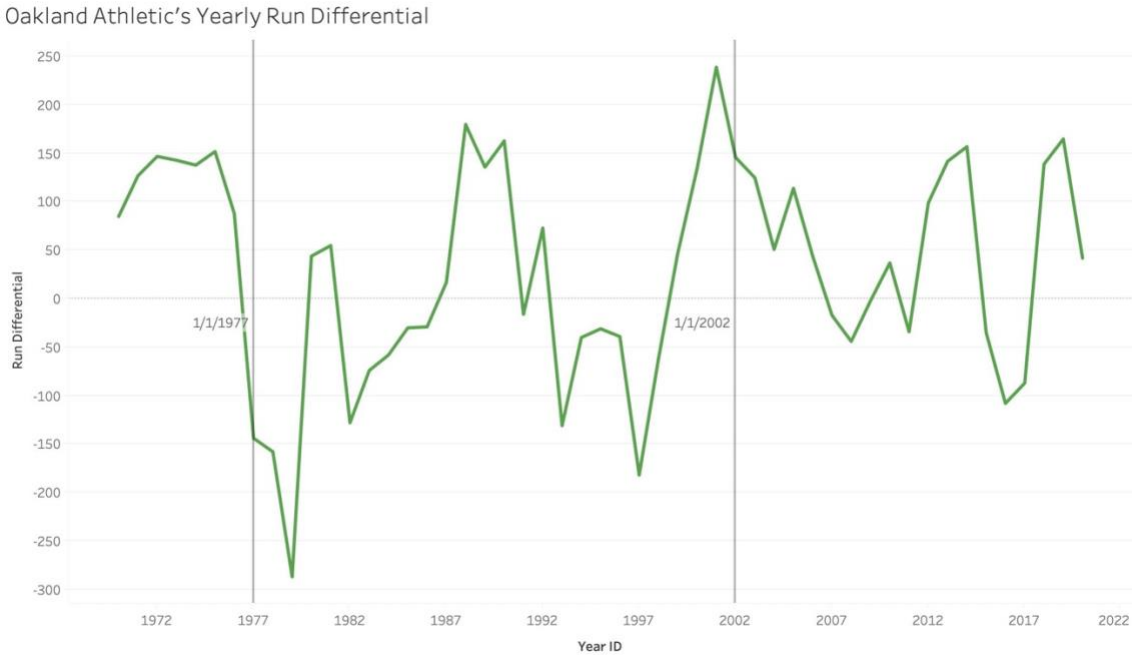
League batting average and on base percentage trends have not changed much in the past 50 years. There are some small up ticks in each in the years following 1977 and 2002, but the percentages have remained relatively stable over the years.



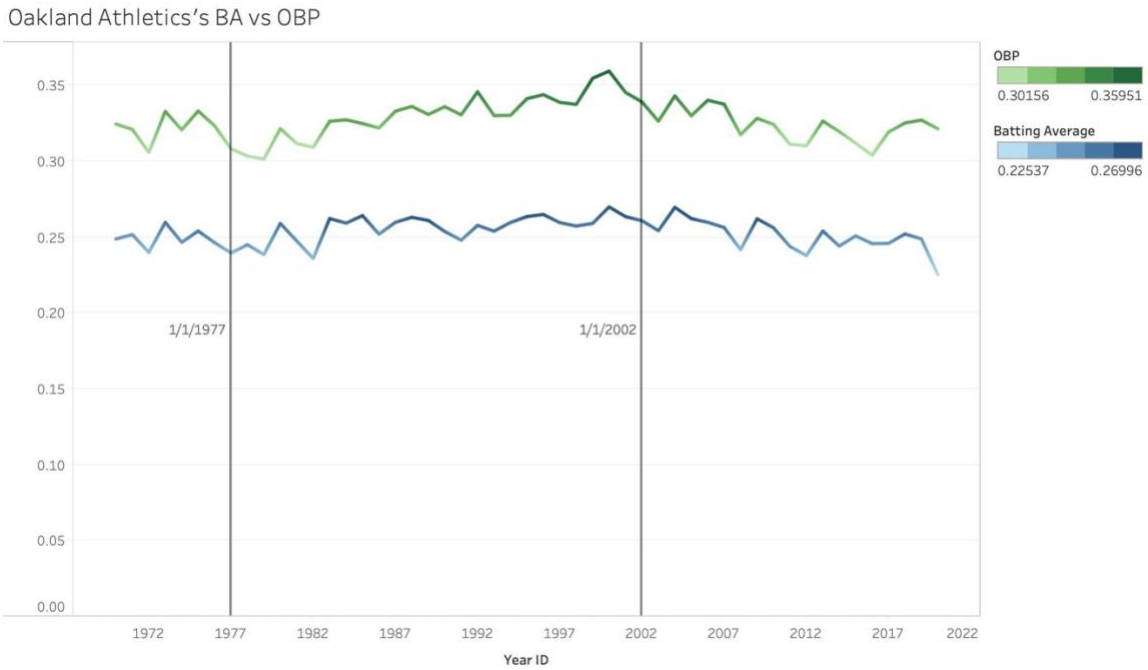
For league strikeout percentage (K%) and walk rate (BB%) there is a clear increase in K% following Bill James' Baseball Abstract and Moneyball. James spoke about the efficiency of committing to the 3 true outcomes in baseball (walk, strikeout, and homerun) and it is clear that K% has risen substantially.



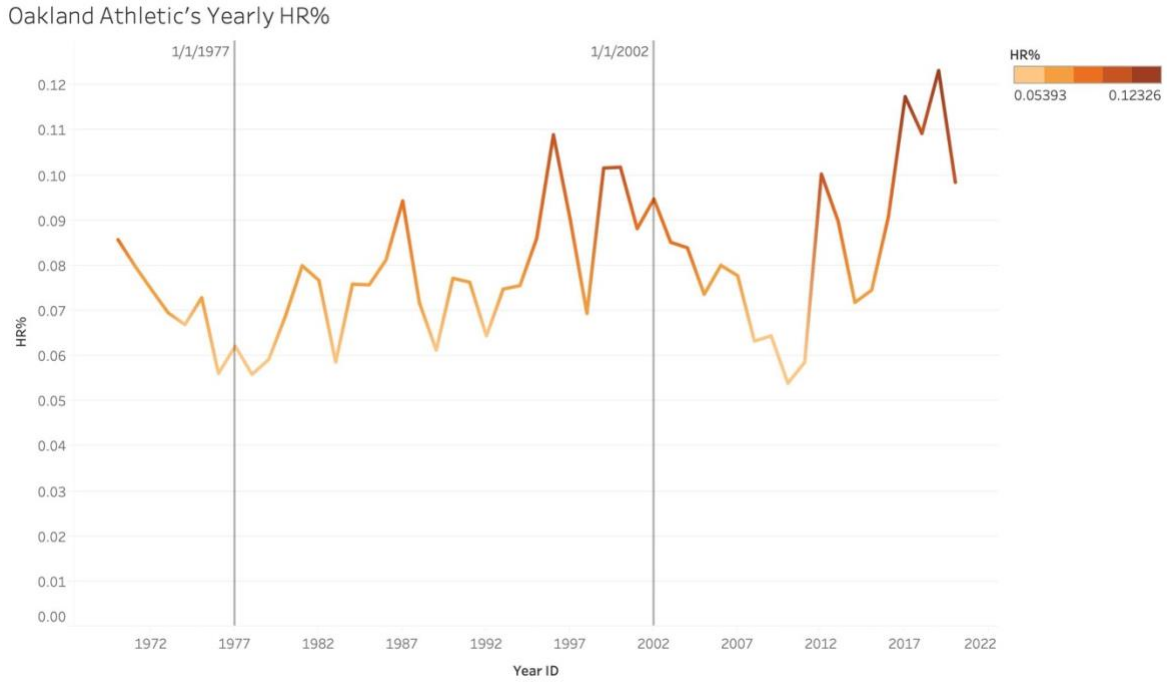
Home run rate (HR%) has steadily increased overall over the past 50 years. It rose steadily following Bill James's Baseball Abstract in 1977 and then began to slowly decrease following the Oakland A's Moneyball notoriety. The reasoning for this could be because the A's stressed the importance of putting the ball in play over hitting homeruns and the chart reflects that.



The above chart shows the Run Differential (Runs scored minus runs allowed) over the last 50 years for the A's.

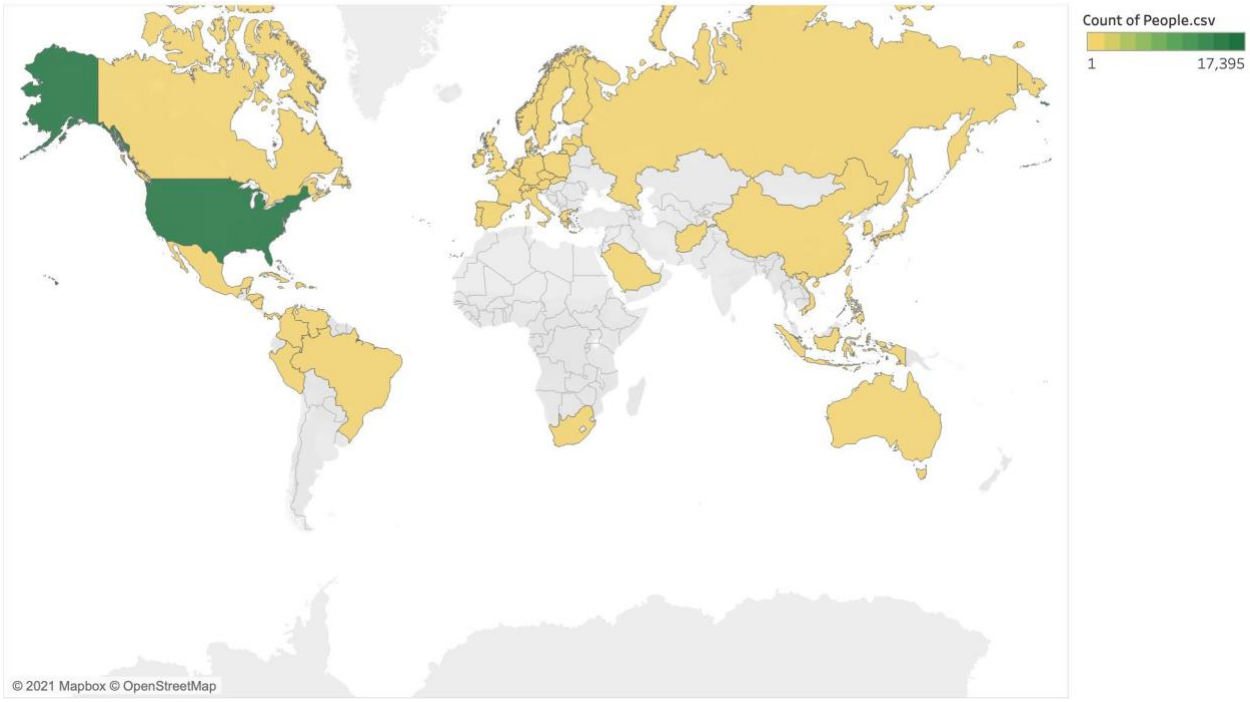


Much like the league trends for Batting Average and On Base Percentage, those rates haven't changed much over the last 50 years and have remained relatively stable. The team batting average has started to slowly decrease since their Moneyball era teams and On Base Percentage being dependent on batting average has decreased slightly as well.



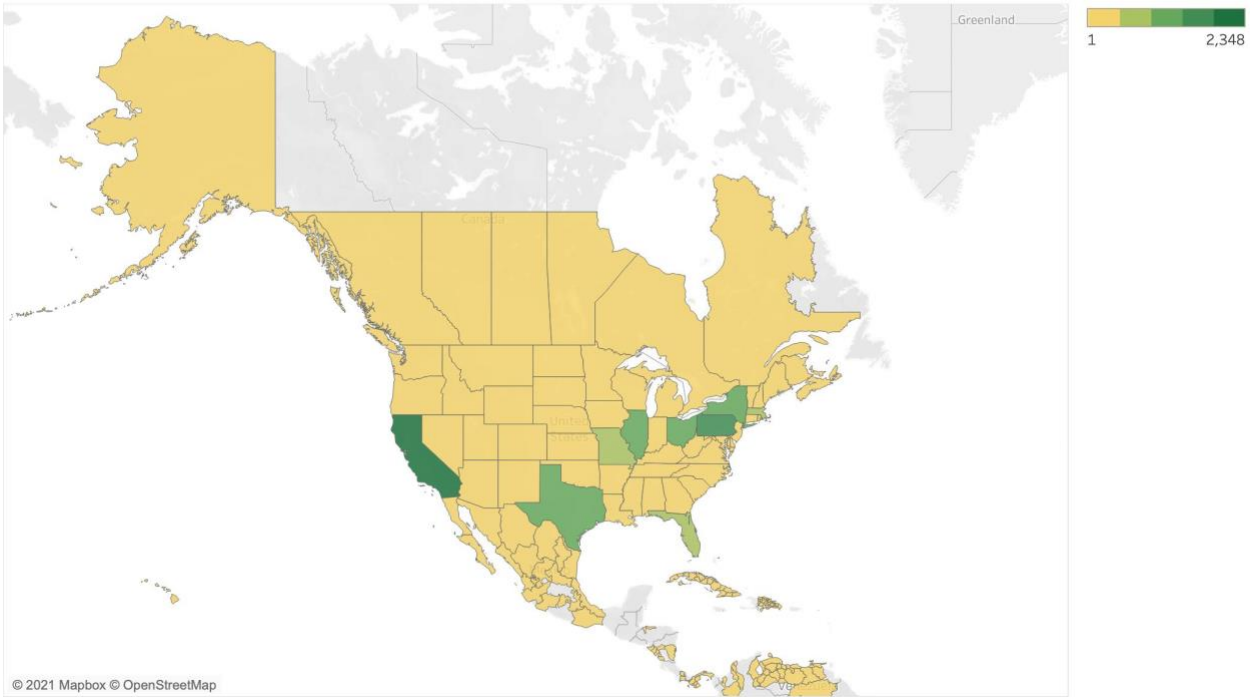
Following Bill James's Baseball Abstract, the A's steadily began to increase their HR% up until they began to implement their Moneyball strategy. Once they did so in 2002, there was a clear de-emphasis put on hitting homeruns for the next 10 or so years.

Player Demographics - Country



The above map shows the birthplace of every MLB player of all time. Due to factors like segregation and the league only beginning to be globalized over the past few decades, the number of players born in the United States dwarfs every other country.

Player Demographics - States



The North American map shows the birth state of every player in MLB history in North America. The states that produced the most players have been California, Pennsylvania, New York, Illinois, and Ohio.