Major League Baseball Analysis

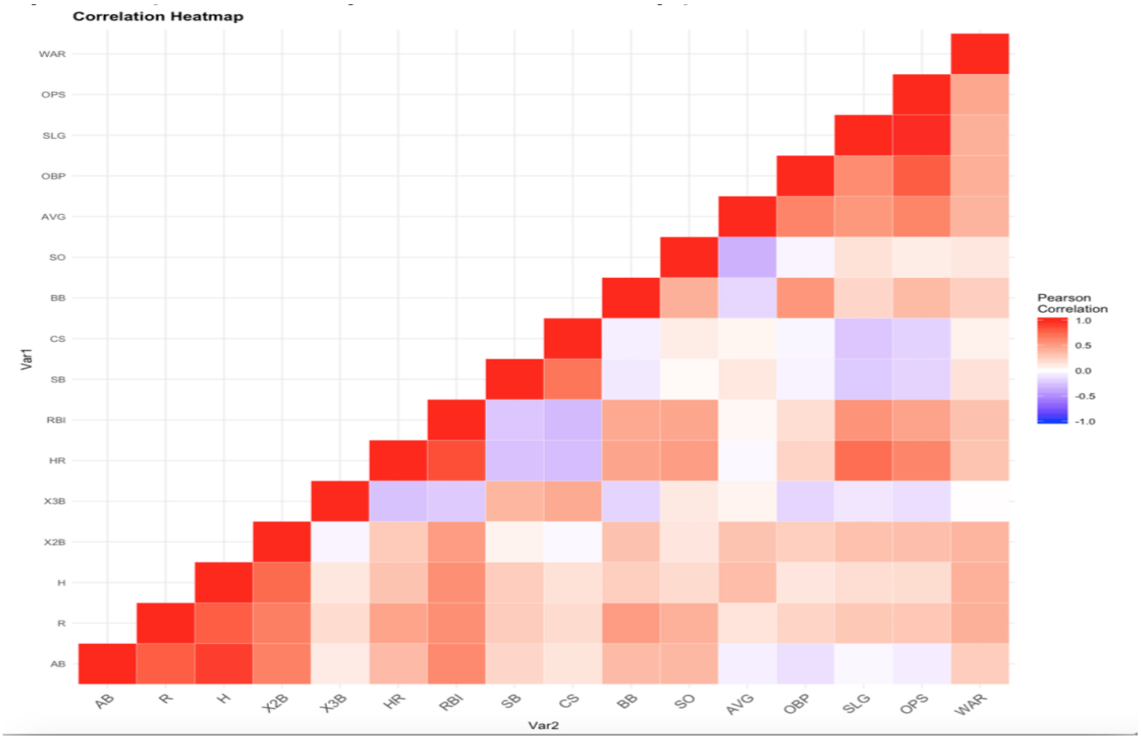
Abram Beyer
Chris Jin
Jonggoo Kang
Matt Winkler
Sungmin Kim

1. Introduction

Our group analyzed Major League Baseball data. We used three core datasets: MLB Homeruns 2016, MLB Team Stats 2013-2016 and 2016 Top batters by WAR. The datasets come from Major League Baseball stats sites such as Yahoo Sports and ESPN, however, we obtained the data directly from statcrunch.com and copied them into Excel. We used a combination of R, Tableau and D3 to create the visualization. WAR (Wins Above Replacement) is a key metric to understand for the Kim / Jin / Kang subgroup. This is a measure of the “value added” by each player in the league. This subgroup did some interesting work creating plots which speak to the drivers of WAR, and what skills make some players truly exceptional. The Beyer / Winkler subgroup focused more on home runs, showing the relationship between home runs and WAR and some more details about home runs in themselves.

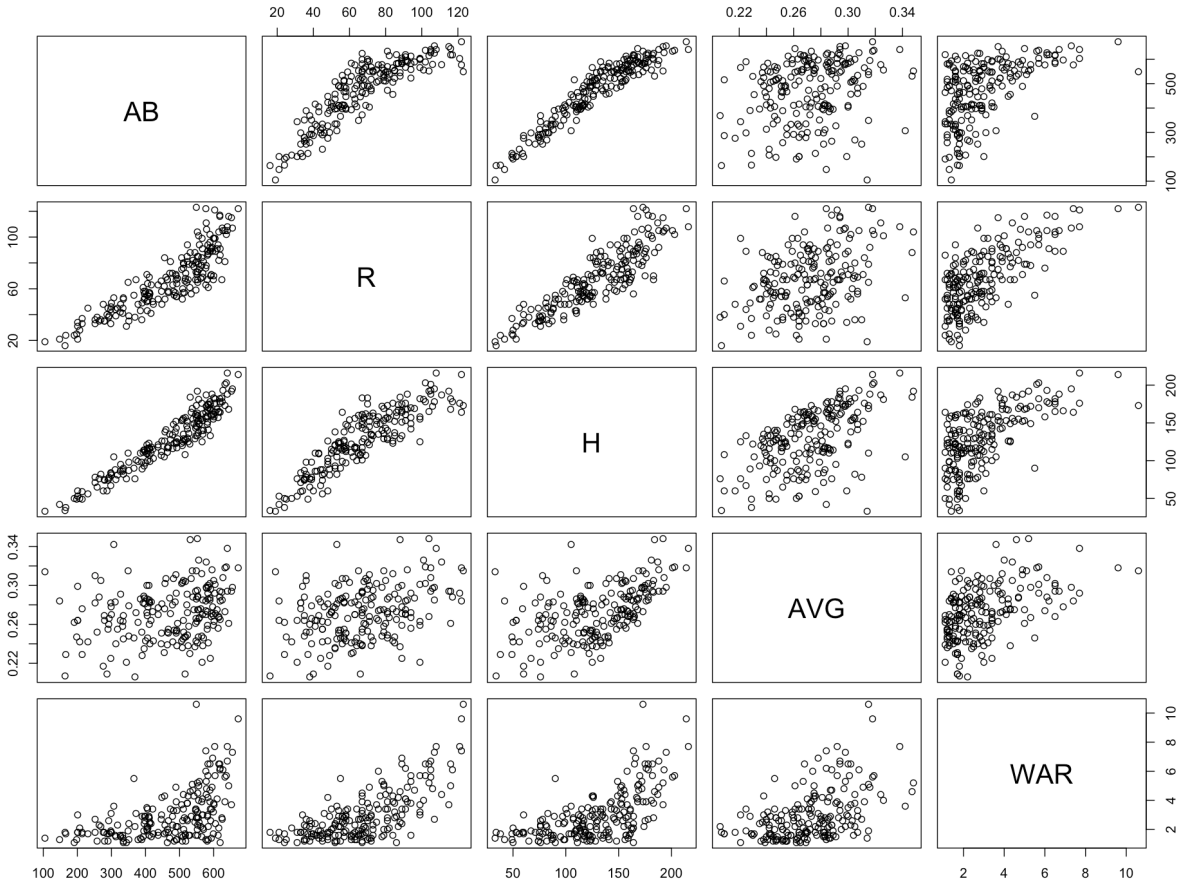
2. Exploratory Analysis

(1) Correlation Analysis



- As a basic step, we conducted a correlation analysis to see which attributes are significant to determining ‘WAR’. According to the correlation heatmap, the number of third-base hit has no correlation to ‘WAR’. One supportive reason for it is that it may be true because not a lot of players made many third-base hits. Simply looking at the correlation is not helpful to discover such informative insight.

(2) Feature Selection




```
Call:
lm(formula = WAR ~ AB + R + H + AVG, data = teams)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4091 -0.6437 -0.1051  0.5409  3.6223

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.465763   2.612960   2.857  0.00474 **
AB           -0.027928   0.005518  -5.061 9.61e-07 ***
R             0.059943   0.006843   8.760 9.45e-16 ***
H             0.089617   0.020468   4.378 1.95e-05 ***
AVG         -25.499541   9.688140  -2.632  0.00917 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.06 on 195 degrees of freedom
Multiple R-squared:  0.6255,    Adjusted R-squared:  0.6178
F-statistic: 81.43 on 4 and 195 DF,  p-value: < 2.2e-16
```

- To see which attributes are most closely related to WAR, we ran correlation analysis using R, through backward step. After removing less-correlated attributes, we decided that the most effective correlations can be found among AB, R, H and AVG. The correlation graph above shows how they are all interrelated to one another. One interesting factor we noticed is that all four attributes have exponential relationships to WAR. This led us to believe that exponentially higher values for these four attributes are keys to differentiate players with higher WAR values.

3. Visualizations

D3.js Viz: "2016 MLB Winnings Above Replacement vs. Other Variables"

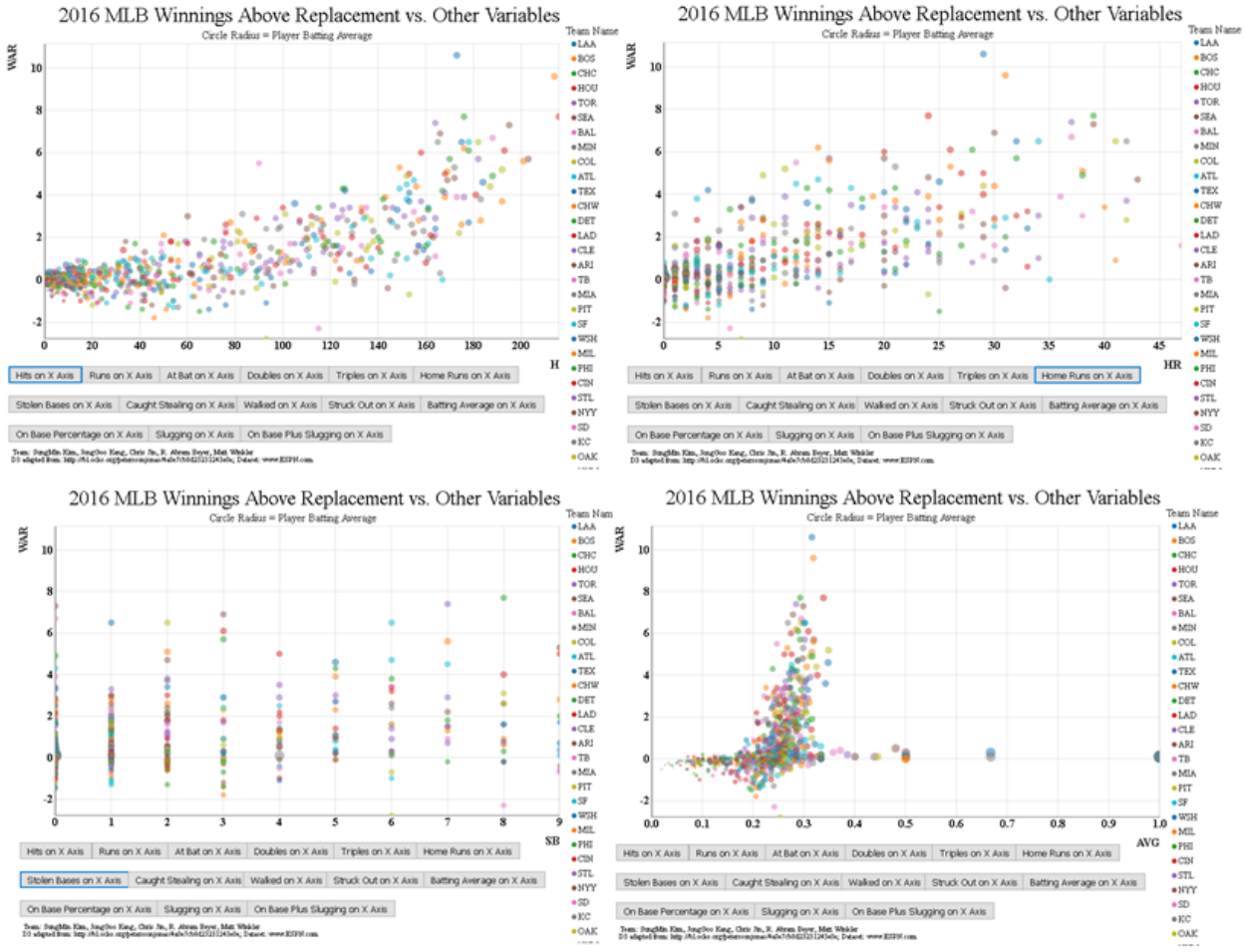
The D3.js visualization titled "2016 MLB Winnings Above Replacement vs. Other Variables" is a scatter plot. This is an interactive, zoomable scatter plot adapted from Jonas Petersson's blog: http://bl.ocks.org/peterssonjonas/4a0e7cb8d23231243e0e. To create this visualization, I copied Jonas' example code and modified it to accept our dataset's variables, added a title, x and y axis labels, modified font size and boldness. In order to make the scatterplot accept more x axis variables and toggle between multiple variables, I added an input button in the html file and an if/else if control in the javascript file function.

The purpose of this scatter plot is to explore the association strength and effect of the several baseball stat variables on winning above replacement. Since winning above replacement is an aggregate value made up of home runs, hits, runs, rbis, batting avg., etc., we wanted to visually explore which variable(s) were most strongly associated with winning above replacement. Screenshots of the graph can be found in Figure 1.
This graph was refined over time. I first modified Jonas' code in order to simply make it work with a few of our dataset's variables. Then, I modified it to include a title, and x and y axis labels. Once we got the scatter plot to work with our data, we modified the scatter plot to include many more legend color boxes to show all baseball teams. Finally, we modified the x-axis toggle control to accept all the variables, not just hits and homeruns. The toggling x-axis is critical to this data visualization. It allows the user to quickly compare all the associations between variables and W.A.R. This fits our analysis because our main goal was to learn about W.A.R. It fits that we should be able to quickly compare the association between all the variables and W.A.R. side-by-side in a clear and easy way.

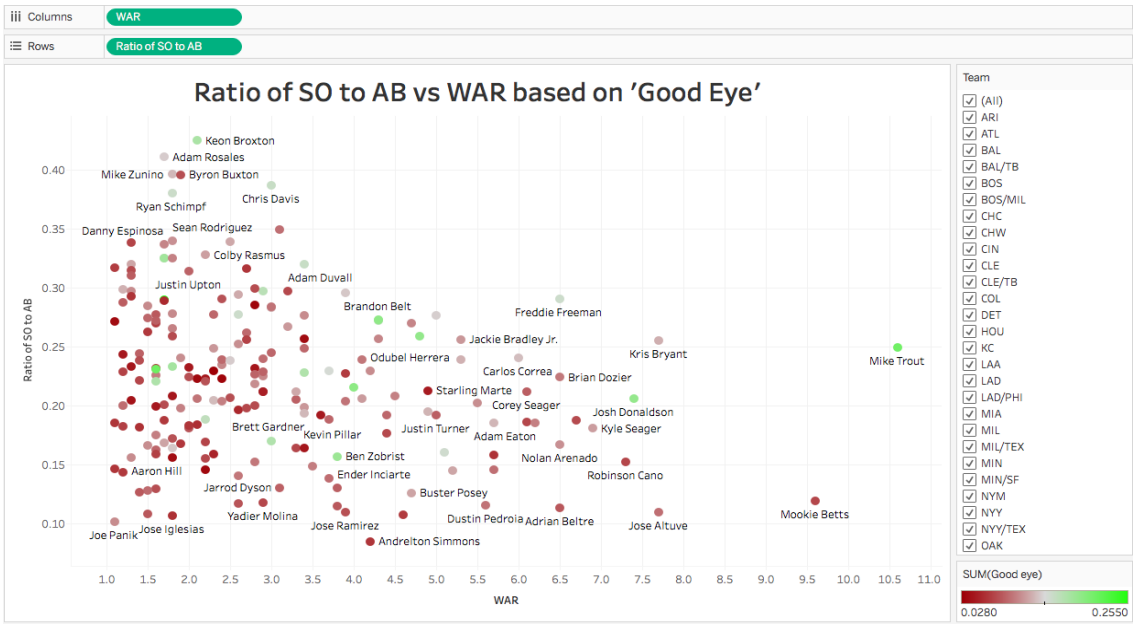
Variable Mapping:

- Y-axis: Winning Above Replacement
- Circle Radius: Batting Average
- Circle Color: MLB Team
- X-Axis: Home Runs, Hits, RBIs, Doubles, Triples, At Bats, Stolen Bases, Caught Stealing, Walks, Struck Out, Batting Average, On Base Percentage, Slugging Percentage, On Base Plus Slugging. (These variables can be toggled using input buttons at the bottom of the graph)

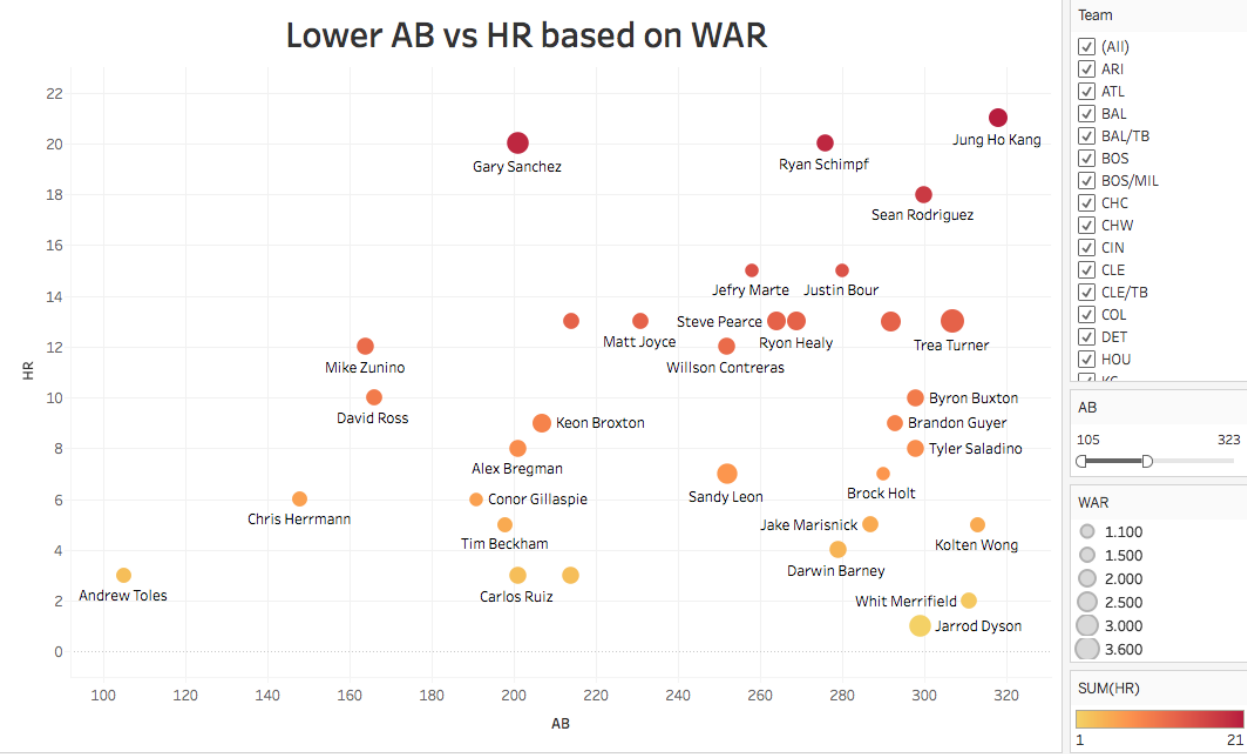
“2016 MLB Home Runs vs. Other Variables” D3.js



-Figure 1 Shows the association between four main variables on he x axis against W.A.R. The buttons below the graph allow the user to change the variable on the x axis. Circle diameter indicates player batting average and circle color indicates player team.

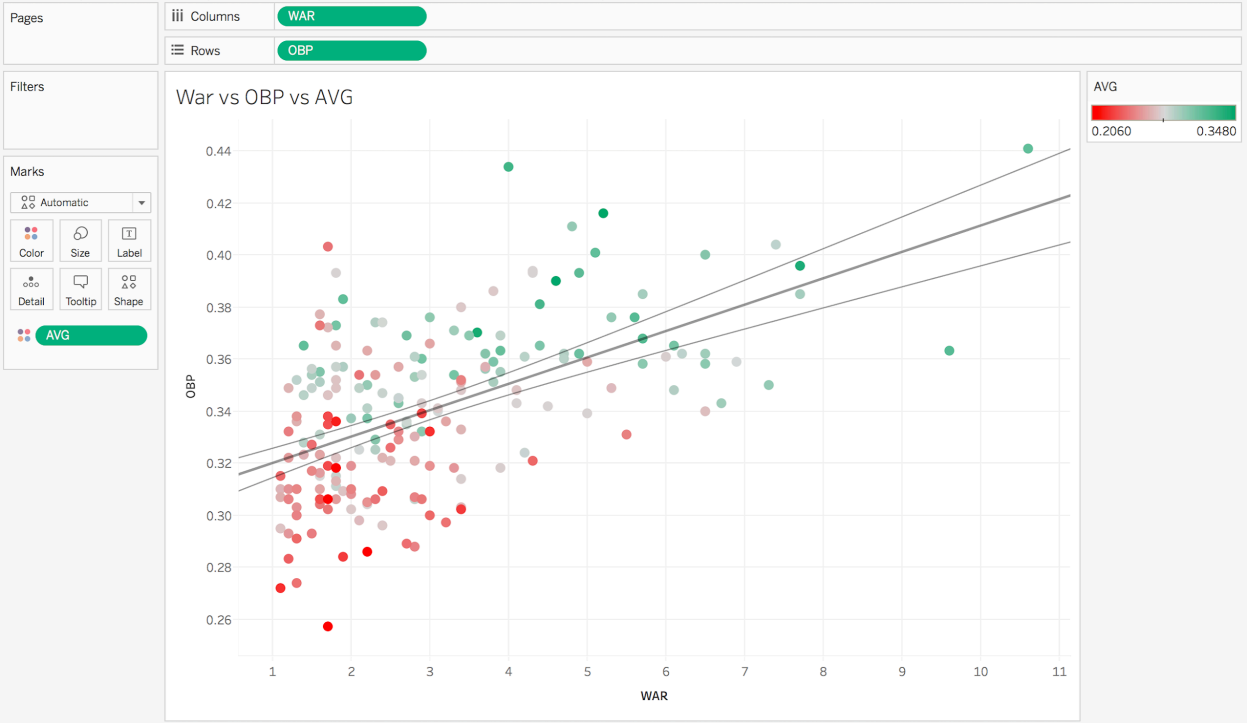


- Figure2 shows a correlation between WAR and Ratio of SO to AB. We created two variables (for more information refer to the appendix 1). Based on the plot above, there exists a clear decreasing trend, meaning that the lower the ratio of SO to AB implies the higher WAR score. In addition, we included a 'Good Eye' variable as an indicator of whether a batter earned lots of walks throughout the entire season. Obviously, the number of walks batters earned does not guarantee the higher WAR score. (As a matter of fact there is an outlier, Mike Trout, who earned lots of base on balls as well as the highest WAR score).



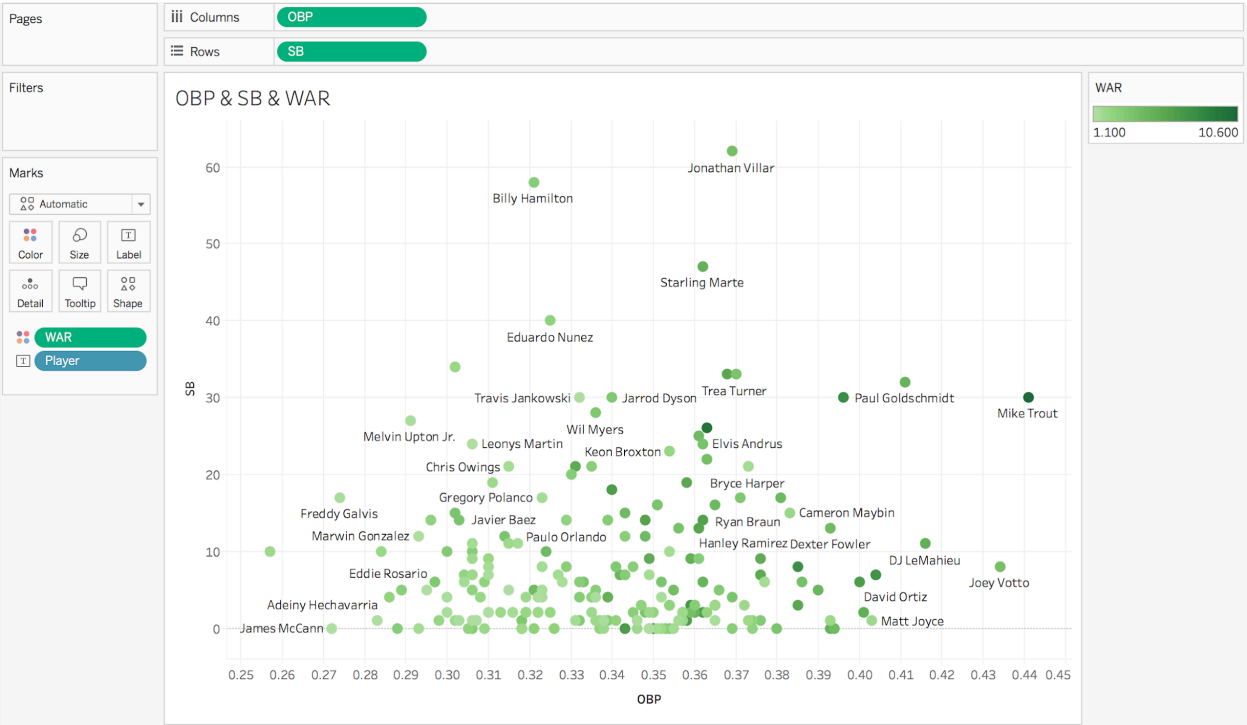
This visualization consists of x axis representing AB (At-Bat) scores and y axis representing HR (Home Run) scores. The size of circles shows WAR scores and the color range from yellow to red reflects numbers of home-run from 1 to 21. We wanted to see good players even though they went to at-bat less than 320 times. Through this visualization, we can see that a batter with higher WAR made more home-run, but went to less at-bat. Gary Sanchez is an ideal example in this case. However, a batter with lower WAR made less home-run, but went to more at-bat. Jarrod Dyson is a good example based on our data.

[2] WAR, OBP, AVG



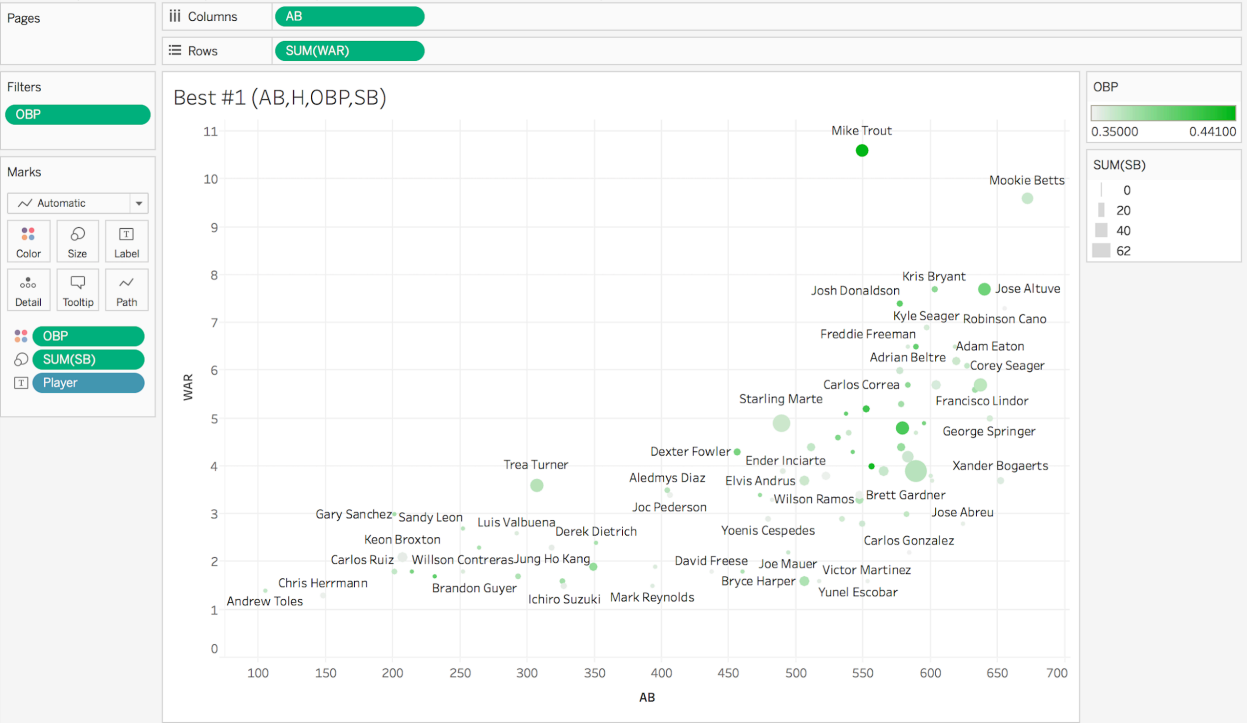
We wanted to apply these attributes into visualization using Tableau. First, we wanted to see how ‘contact batters’ are scattered within scatterplot with WAR, OBP and AVG. My assumption is that a good contact batters should have higher AVG (also known as average – batter on average by hitting) and OBP (also known as on-baseball percentage – includes hits, base-on- balls, hit-by- pitch, and due to various error) and WAR above average trend. I wanted to visualize if players with higher WAR are shown with indeed higher AVG and OBP. Visualization shown does reflect exactly my initial assumption. Most players above the least linear regression were hitting much higher AVG rate, and their OBP were generally higher for players with higher WAR. To address different types of batters, I simply divided into two groups – quick runners and sluggers.

[3] OBP, SB, WAR



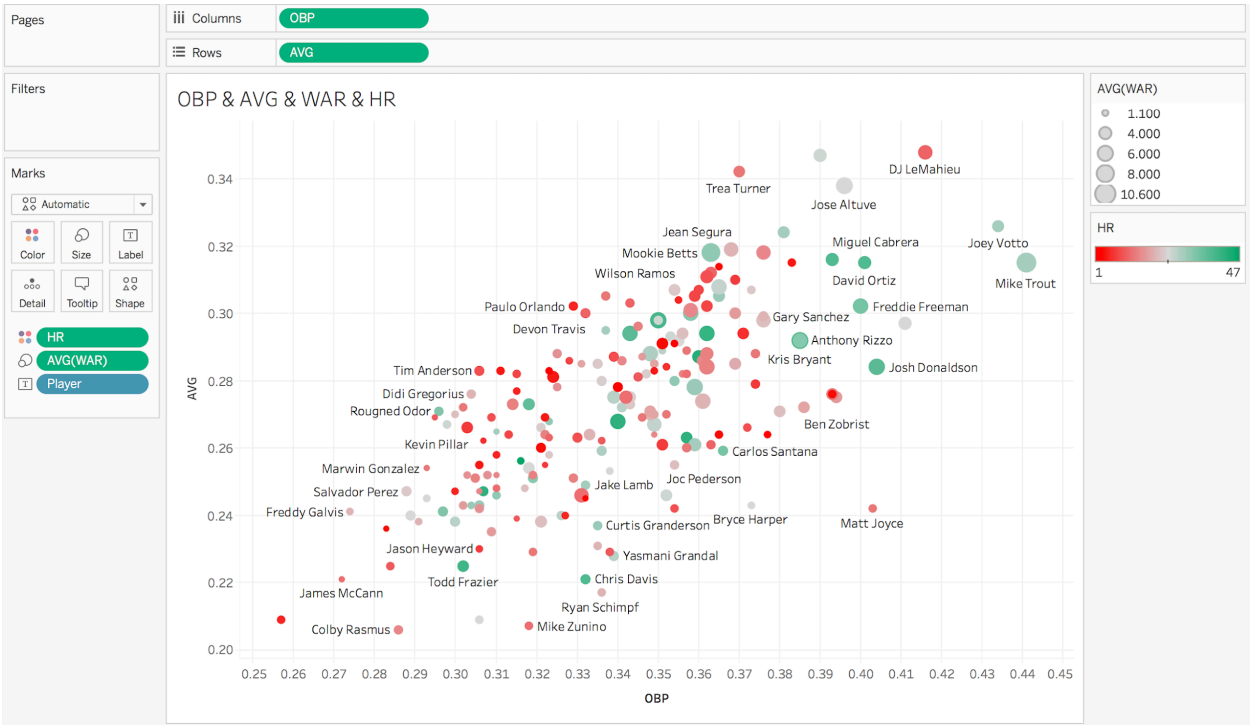
For quick runners, assumingly top-of- lineup-batters, We wanted to see whether WAR is related to batter's SB (also known as stealing base counts – good barometer of how quick and effective the runner is) and OBP. Sabermetrics highly values players with higher OBP because this means batter has higher chance of getting into the base, meaning higher chance to score. A simple correlation between OBP and SB with WAR, however, shows that there is not a strong correlation amongst players with higher OBP and SB leading to higher WAR, but players with higher WAR seems to be generally coming from players with higher OBP, and less so with players with higher SB counts.

[4] Best #1 (AB, H, OBP, SB)



To see more correlation, We added players' AB counts (also known as at bat – number of batters in the batting box) against WAR with OBP and SB counts. In the beginning, we filtered so that the scatterplot only reflects players with OBP higher than 0.35. At bat and hit counts can visualize players with fewer games played. By including these attributes, we could easily visualize player counts with higher OBP and SB as opposed to batters' AB counts. Although some players only played half the numbers of AB, they were able to produce as much, if not more, SB. The graph shows that most players with higher WAR were strongly correlated based on numbers of AB, less so with SB and OBP.

[5] OBP, AVG, WAR, HR



To dissect attributes relating to sluggers, we chose players' OBP, AVG, WAR and HR (homerun counts). The scatterplot can show that while OBP and AVG are visualized, players' WAR as well as HR counts are also presented. By running this scatterplot, we was able to see power batters with their AVG and OBP. Our assumption was that players' AVG may not show strong correlation to homerun counts, there would be some sort of positive correlation between AVG and OBP, and players with higher HR counts are likely to be related with higher WAR. The scatterplot shows that there is mostly a linear relationship between OBP and AVG, and most players with higher WAR tend to have higher HR counts.

Home Runs Infographic

Our initial idea was to analyze the contribution of home runs to WAR. After starting to visualize the home runs data, we became more interested in some of the numbers surrounding home runs themselves. Since home runs are one of the most engaging and popular aspects of baseball for fans to watch, we developed an infographic which summarizes some interesting information about home runs. The general points of interest around home runs are who hits them and when they're hit. So, most of our charts speak to that theme. Digging a little deeper, we were able to reveal some more aspects about the flow of the game of baseball and which types of players tend to hit the most homers.

Home Runs Visualization Development

We used a combination of Tableau and R to create the visualizations in the infographic, and then imported them into Adobe Illustrator to create the final product. There was a substantial amount of effort devoted to organizing and re-coloring the images once we imported them to Illustrator. Overall, Illustrator seemed to be a good tool for aggregating many different plots, but it is extremely oriented towards editing fine details in the graphics. It's really more of a general purpose visual tool than something built specifically to visualize data. So, if the use case were to create a single visual for a project, we wouldn't recommend its use. However, it was good in the sense that it did what we needed and offered a great deal of flexibility for adjusting the results.

In general, our strategy for the infographic was to use a dark background with bright foreground colors to make the plots pop as much as possible. All axes and summary metrics are white to be consistent throughout the piece. Any lines within the plots themselves appear orange. We tried to mix up the fill colors according to the type of data displayed in order to reinforce that the plots are displaying different aspects of the home runs data. The "# Hit by Each Team" and "# Given up by Pitching Team" charts are exceptions to this rule, because those charts are plotting a similar statistic with the same groupings. Here we used different shapes (circles vs. bars) as a secondary way of illustrating the difference in the data.

4. Analysis and Discussion

Home Runs Insights

Reviewing the information in the charts themselves shows several interesting aspects of home runs. The summary numbers, especially ~188 home runs per team and the 47 home runs by Mark Trumbo (tops in the league) are good benchmarks for the rest of the visual. The "# Hit by Each Team" plot shows that the Orioles (Trumbo's team) are also the top hitting team, and that in general (sorting visually by the league colors) the American League hits more home runs than the National League. The time plot shows relatively few home runs in the early season, suggesting that it takes some time for hitters to warm up. "Home Runs by Inning" is interesting, especially how it shows the spikes for innings 1, 4, 5, and 6, and a large dip for inning 9. The first and middle innings are when the best hitters tend to bat, so those increases make sense. The dip for inning 9 is likely caused by games ending before the last half of the 9th inning when the home team is ahead. "# Given up by Pitching Team" shows that the American League also gives up more home runs, which isn't terribly surprising. The distance distribution shows that the average homer travels around 400 feet, and that home run distance actually follows a surprisingly normal pattern. Finally, "Distributions by Age Group" shows that younger players tend to hit fewer home runs than older players, which is very interesting and different from most sports performance metrics, where younger players tend to dominate.