Moneyball

By: Jack Whitman, James Teschner, Paul Kluckman, and Tyler Chaia

Process Book

Link to Published Storyboard:
https://public.tableau.com/profile/publish/MoneyballFinal/Story1#!/publish-confirm

Overview/Motivation

As baseball fans, we ultimately wanted to create a dashboard that could provide helpful insights about different players in baseball. At first, we were interested in home runs since that is one of the most exciting moments during a baseball game, however we determined that building the best lineup based on available cap space would be more useful for baseball managers. We decided to build a dashboard that could compare players based on their pay and performance. We focused on making our dashboard dynamic so that managers could focus on different positions or focus on players with the specific credentials that they are looking for. As analytics continues to play a role in baseball, we think that this visualization is something that could soon be used to compare different players performances and predict their future value to a team.

Questions

We had a few main questions in mind:

- How are salaries and batting performance correlated?
- What are the league average batting performance based on a particular range of salaries?
- What is the slugging percentage?
- What is the future value of a player to a team based on performance measures?
    - Predicted strikeouts?
    - Predicted runs created?
    - Probability of success on plate?
- What players should we pick over others based on their salaries and performance?
- If a manager is looking for a particular player, how do they compare to other players and the league averages in performance?

Data

To begin this project, our group had to dig through several websites in order to find data for baseball statistics. At first, we found data through the MLB website and Baseball references, but towards the end we found a new website that contained significantly more data in baseball

dating all the way back to 1985. This data source was found at SeanLahman.com, who is an award winning database journalist and author. The dataset included an updated version of batting, pitching, fielding, standings, team stats, managerial records, college statistics, and more.

From the original data we found, we came up with questions that looked into what factors go into a player hitting a home run, what month of the baseball season is associated with the most home runs, what teams have the most home runs and why. However, once we found another dataset we moved away from these questions and decided to examine the data on a managerial level. For example, we wanted to see salaries of players in a particular position, and then further look at what the expected performance of these players should be. With this, a manager could decide to obtain a batter that they are in need for based on a specified salary.

Data Cleaning, Integration, and Processing

Our analysis required a significant amount of cleaning, integration, and processing in order to provide meaningful insights. Since we downloaded a numerous data sets from various sources, it was critical that we established consistent identifiers between sets in order to create relationships. For example, when we joined a data table that listed players and their associated positions, that data set did not contain a “playerid” which our other main data set did. This required vlookup functions to target the players and assign the appropriate unique identifier. In addition, when we paired the data sets, we wanted to make sure all of our column headers (variable indicators) were consistent throughout.

After creating a unique ID that could be applied to each spreadsheet in Excel we used R to integrate all of our spreadsheets. The main datasets that we integrated included information about batting, salaries, and general player information. One problem that we encountered with our datasets was that they all contained a different amount of rows, and some of them included players that the other ones didn’t. Since our main focus was on each individual batter, we decided to merge all the data onto the batting spreadsheet using Payer ID and Year ID as the merging criteria. As a result, we were able to create a single spreadsheet containing approximately 34,000 rows and about 46 different columns with no missing values. The first few rows of our data are shown below.

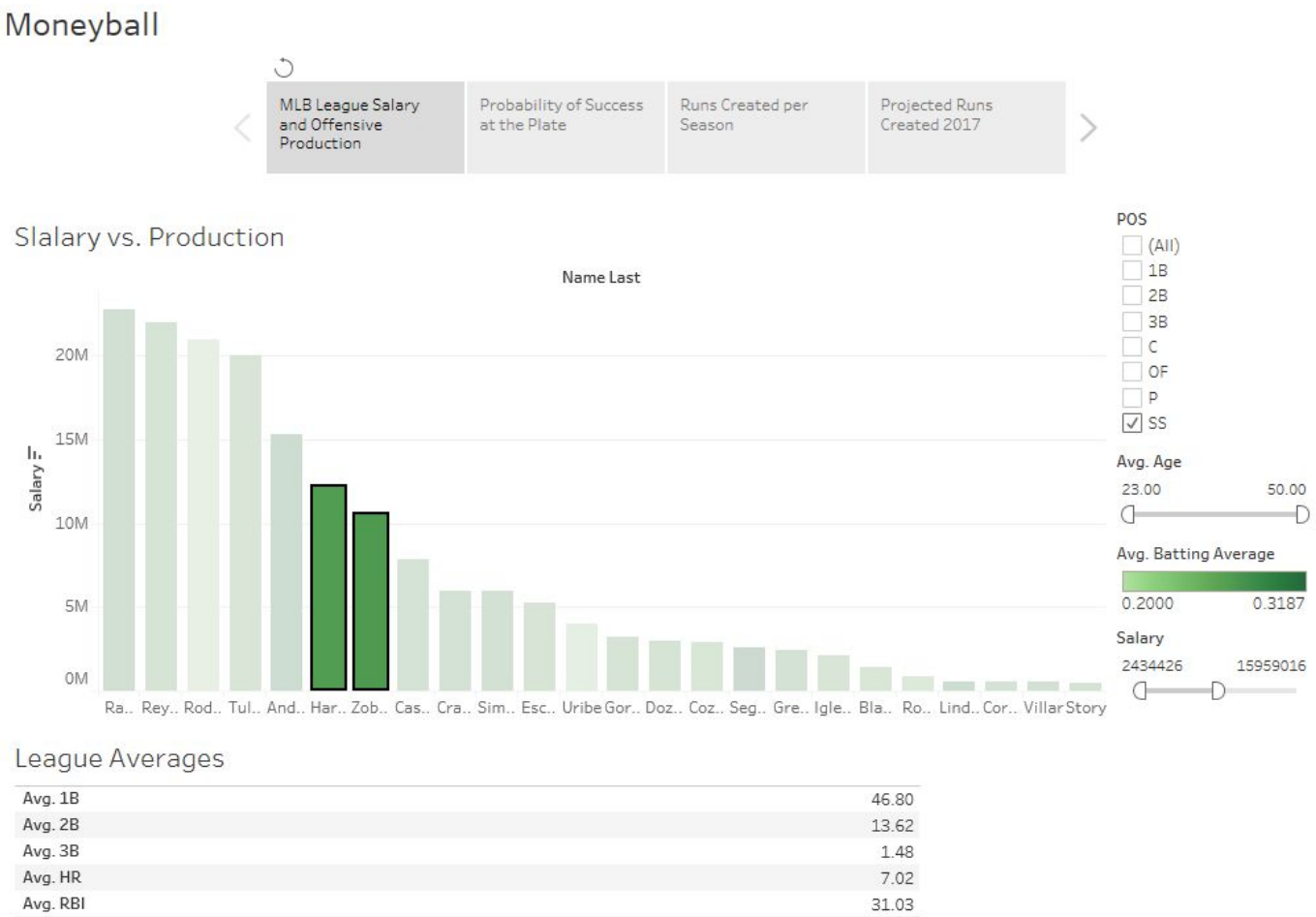
playerID	yearID	stint	teamID.x	lgID.x	G	AB	R	H	X2B	X3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP	Year	nameFirst	nameLast
aardsda01	2004	1	SFN	NL	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2004	David	Aardsma
aardsda01	2007	1	CHA	AL	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2007	David	Aardsma
aardsda01	2008	1	BOS	AL	47	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2008	David	Aardsma
aardsda01	2009	1	SEA	AL	73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2009	David	Aardsma
aardsda01	2010	1	SEA	AL	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2010	David	Aardsma
aardsda01	2012	1	NYA	AL	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2012	David	Aardsma
aasedo01	1986	1	BAL	AL	66	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1986	Don	Aase
aasedo01	1987	1	BAL	AL	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1987	Don	Aase
aasedo01	1988	1	BAL	AL	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1988	Don	Aase
aasedo01	1989	1	NYN	NL	49	5	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	1989	Don	Aase
abadan01	2006	1	CIN	NL	5	3	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2006	Andy	Abad
abadfe01	2011	1	HOU	NL	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2011	Fernando	Abad

However, our data still wasn’t ready to put put into Tableau. In order to provide the best insights possible, we decided to create additional columns that measure each batter's performance. Some of the advanced measures that we created include: slugging percentage, predicted strikeout percentage, and runs created. We will discuss these measures in further detail under the *Exploratory Data Analysis* section of the report.

The last step of the data cleaning process for us involved a measure that could used to predict a player's performance in 2017. In order to do this we need to use an equation that used the player's runs created from the last three years to project their runs created in the upcoming season (2017). One problem that we ran into with this process was the fact that our data set repeated players who played multiple seasons in the MLB. So, information that we needed from each player to project their runs created in 2017 was in all different rows. In order to fix this problem, we used filtering to remove duplicate players and created a vlookup table for each player. This way we could neatly project the 2017 runs created for each player and bring that number back into the data source that we were using. At this point, we were ready to bring our data into Tableau and begin building visualizations.

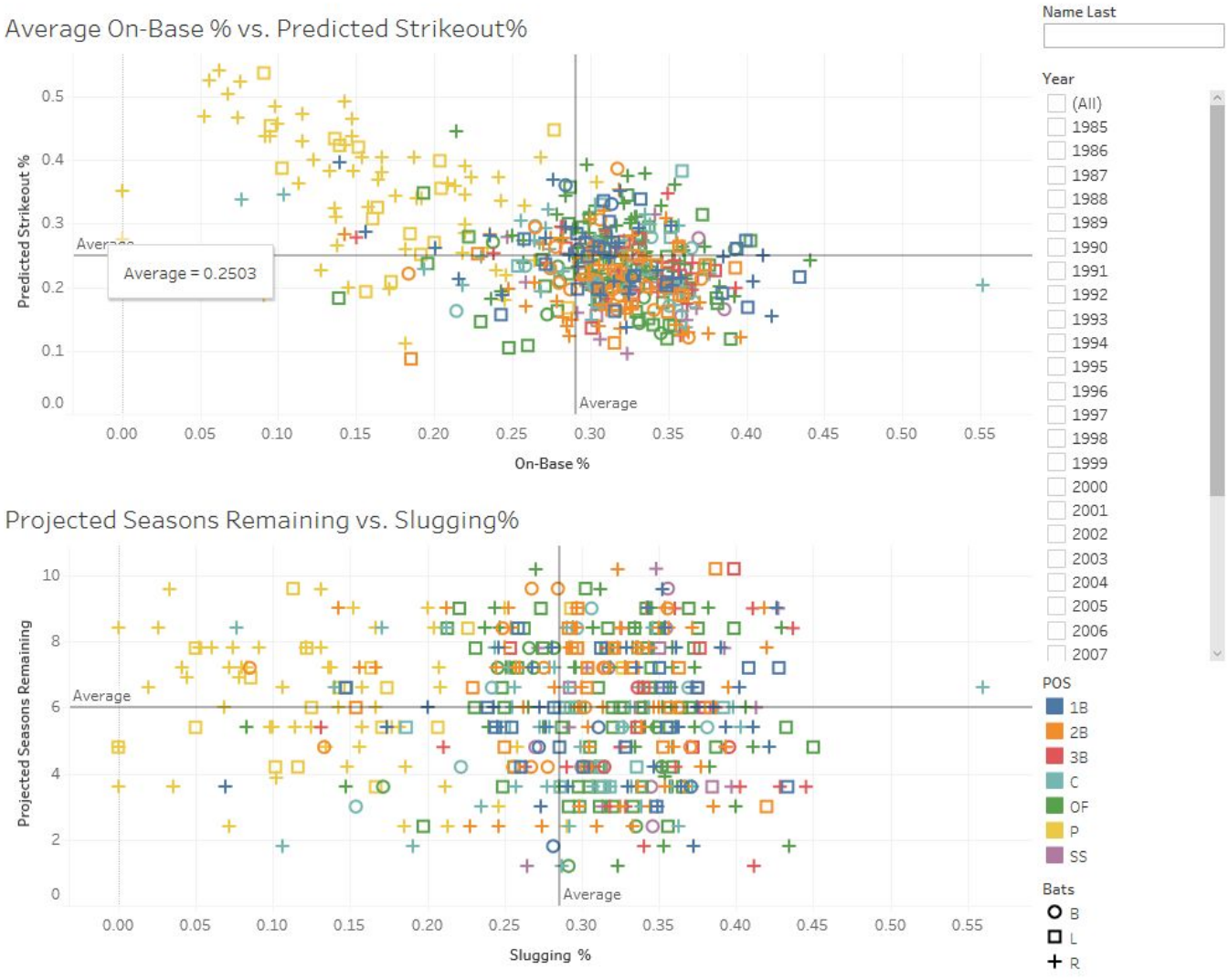
Exploratory Data Analysis

We thought that it would be important for our first dashboard to provide an insightful overview of each player based on salary level and their on-base percentage. We also wanted to make the first dashboard dynamic enough to allow us to compare similar players. In order to achieve this goal, we created a bar chart with multiple filters that allowed us to narrow down on player position, salary range, and age. As part of this visualization, we included a table that indicates the league averages for multiple batting statistics that allows us to see how players are doing relative to the performance standards of the MLB. The completed dashboard is pictured below.

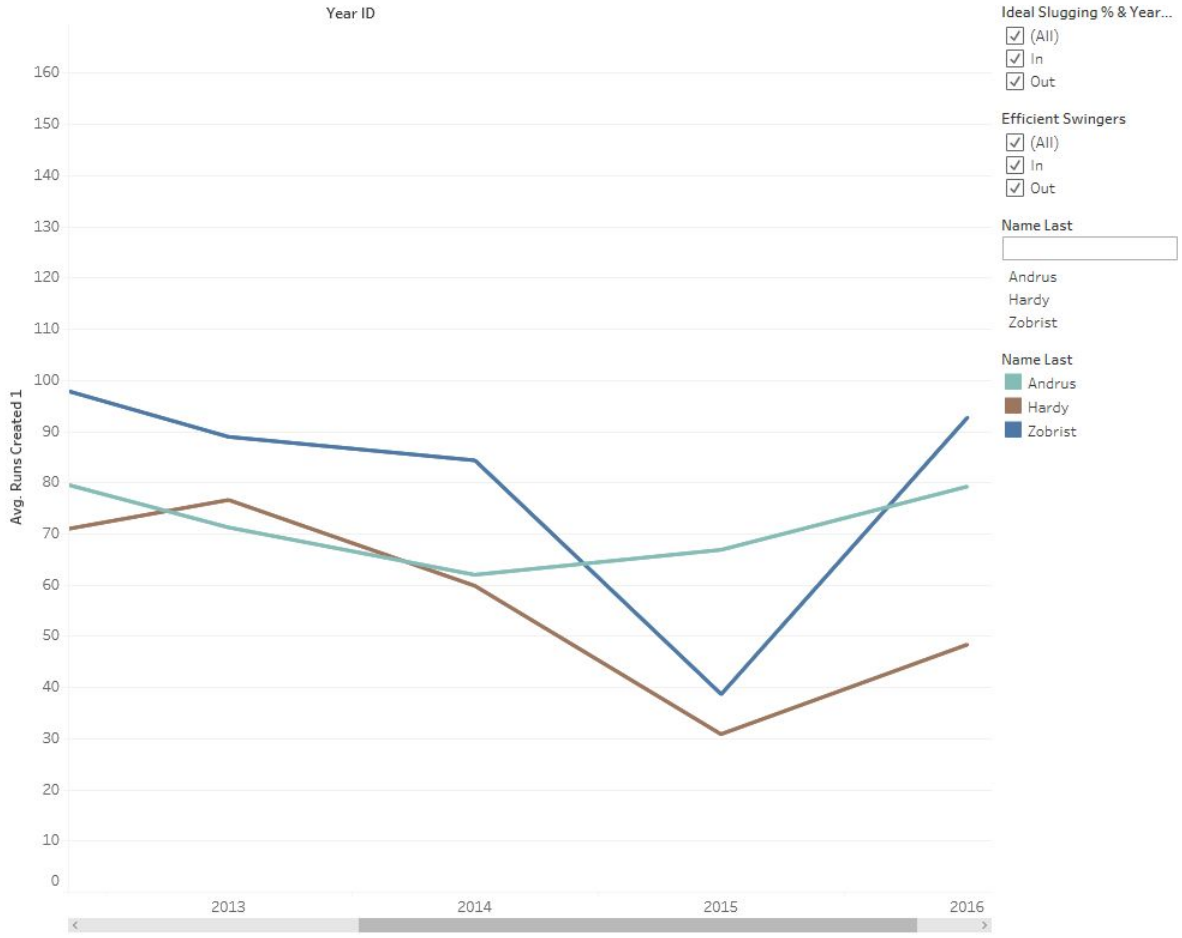


Our second dashboard is designed to focus on how successful a player may be when they are up at the plate. A batting average does not reflect when a batter gets walked or hit by a pitch, but most baseball managers would still consider it a successful plate appearance because the batter who was walked now has an opportunity to score a run. In addition, we think that it is important to measure strikeout percentage because players with a low strikeout percentage are most likely getting a lot of hits, or forcing the pitcher to throw them a strike. They have a "good-eye". Either way, a low strikeout percentage most likely is getting on base, which provides an opportunity to score a run.

Overall, we believe that when we measure these statistics together we can identify the most efficient swingers in the league. Batters who have a high on-base percentage may be getting walked a lot, which means that they aren't swinging at bad pitches. Batters who strike out a lot probably indicates that they swing at pitches outside the strike zone too often. The dashboard is pictured below.



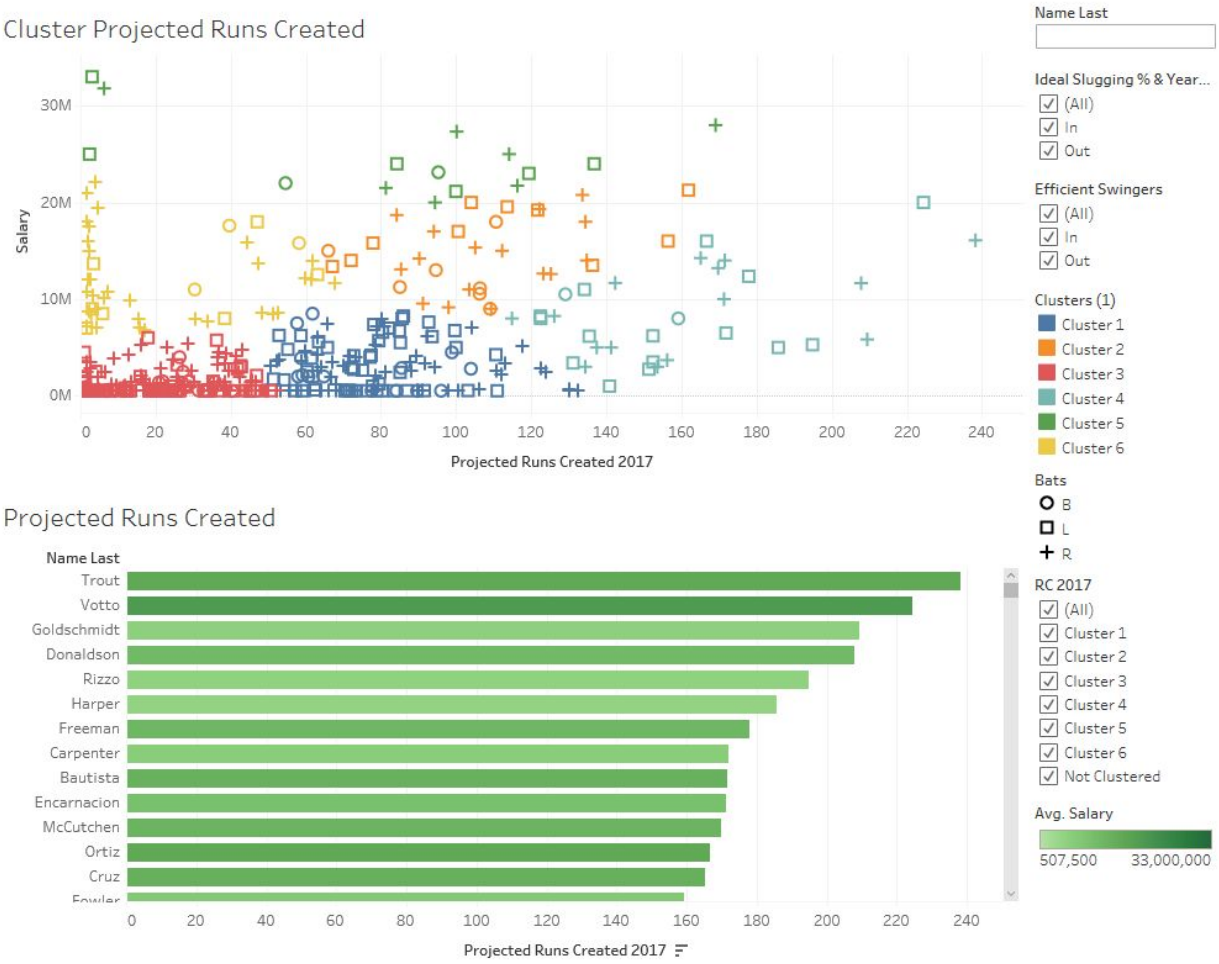
Our third dashboard depicts the runs created by players from the past five years. For example, if we chose to look at Andrus, Zobrist, and Hardy, whom are all shortstops, we can determine how they have been performing for the past few years. We choose this because it is another performance measure that could help us determine which player we would like to pick over the other based on their past. That being said, a player with an upward trend in the number of runs created in the past five years may be a better pick because they could potentially have another great season in the next year which would be helpful to our team as a whole.



The purpose of our last dashboard is to predict how valuable a player will be in the upcoming season. We believe that the best way to measure the value of a player is to consider the amount of runs they create versus their salary. It's obvious that teams who score a lot of runs have the highest probability of winning games. We believe that our final visualization helps identify the characteristics of players who contribute the most to the runs scored by their team during a game.

Our final visualization includes clusters that puts players into groups based on their projected runs created for 2017 and their current salary. Looking at the dashboard (pictured

below), we would consider cluster 4 the optimal cluster. These players have a mid-size to high salary, but they are projected to create a lot of runs. As baseball managers we would want to consider trading for these players or build a strategy for when we play against them. We have also included a bar chart in our final dashboard that clearly identifies the the top performing players. In the chart, a dark green bar indicates a high salary, and a light green bar represents a lower relative salary.



The performance measures in all the dashboards above were calculated in numerous of different ways within tableau and excel from all the data we had on the players:

- Slugging percentage - This was calculated within Tableau by adding singles, doubles, triples and home runs then dividing the result by at bats in order to factor slugging percentage into our analysis of home run hitting as a whole.
- Strikeout percentage - This was calculated in Tableau by dividing the amount of strikeouts by the at bats. The more strikeouts or higher strikeout percentage a

batter might have allowed us to determine less of a chance of making contact, therefore, less of a chance of hitting a home run.

- Projected Seasons Remaining

Age	Runs.Created.Next.Year.Proj	teamID.y	lgID.y	salary	Projected Seasons Remaining
36	0	SFN	NL	300000	=(24-(0.6*AD2))

- Runs Created per Season

$$RC = \frac{(H + BB - CS + HBP - GIDP) \times (TB + (.26 \times (BB - IBB + HBP)) + (.52 \times (SH + SF + SB)))}{AB + BB + HBP + SH + SF}$$

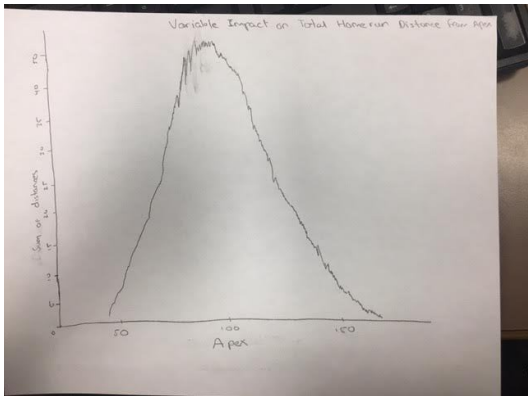
- Projected Runs Created 2017

playerID	2014	2015	2016	Projected RC
abreujo02	160.9185	135.9119	140.4445	=(1*B2)+(2*C2)+(3*D2)
abreuto01	0	0	0	0
accarje01	0	0	0	0

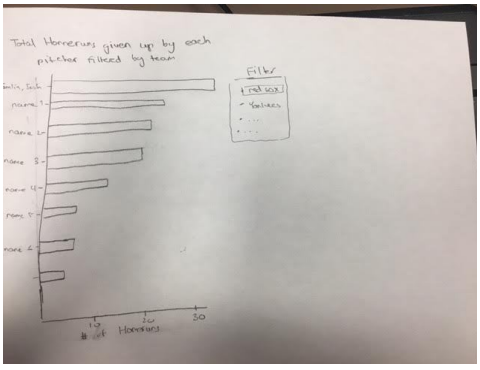
Design Evolution:

Originally, we wanted to create multiple dashboards that could provide helpful insights about different players in baseball. At first, we were interested in home runs since that is one of the most exciting moments during a baseball game. We wanted to explore how the apex and the distance the ball will travel based on that. Additionally, we wanted to further look into how many homeruns are given up by a particular pitcher by team to see which pitchers may be a bad selection for our team. The last visual we wanted to create was a heatmap of the home run location frequency which would tell us the best direction a ball should head towards if we want a player to produce a homerun. All of these original design evolutions are depicted below.

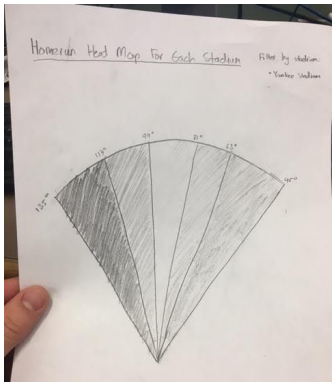
Variable Impact on Home Run Distance



Homeruns Given up by particular pitcher by team

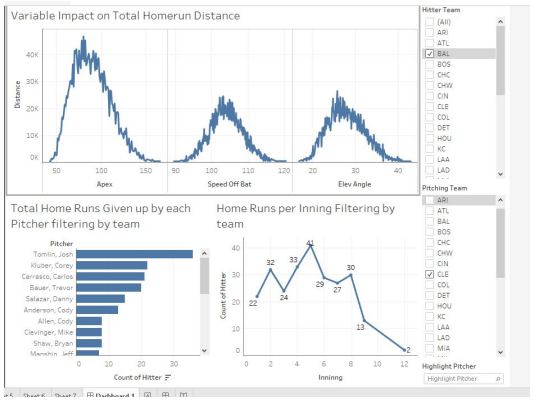


Heatmap of Homerun Location Frequency (filtered by field - darker color in this case will denote a higher frequency)



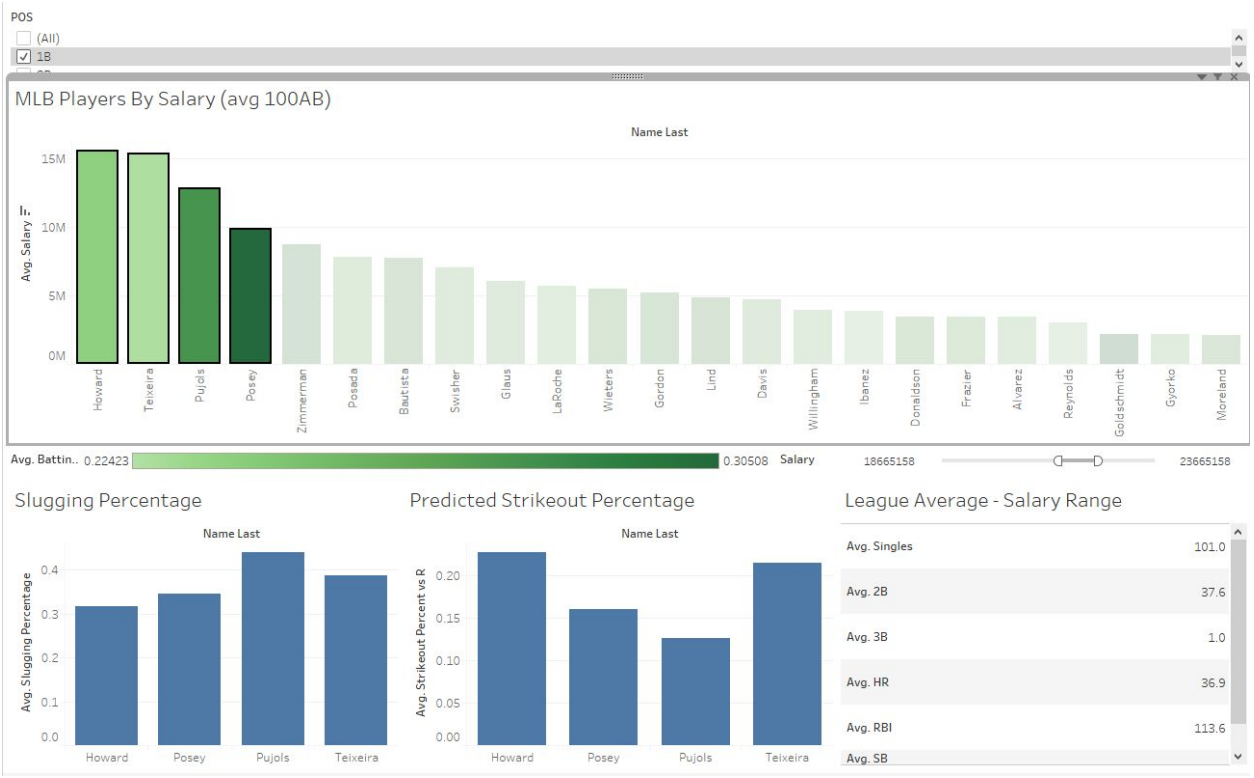
After deciding that our original visualizations would not be the best way to create a valuable conclusion about picking a particular player in baseball, we decided that building the best lineup based on available cap space would be more useful for baseball managers. We decided to build a dashboard that could compare players based on their pay and performance. We focused on making our dashboard dynamic so that managers could focus on different positions or focus on players with the specific credentials that they are looking for. As analytics continues to play a role in baseball, we think that this visualization is something that could soon be used to compare different players performances and predict their future value to a team. The visualizations below were part of our final submission for the original part of the project.

Original Dashboard:



The following were sketches and visualizations of the original dataset and questions that we established. As the semester progressed, we ended up switching our dataset and therefore changed many of our questions and our visualizations became much different.

Digital Visualizations - Part 1:



After further critique of these visualizations, we were able to dig even deeper into picking a particular player based on their performance measures, salaries, and even the future predictions for particular players.

Implementation

Implementing the data included adding several features in Tableau to make the visualization more dynamic and user friendly. For example, we added a search bar to each one of dashboard that allows users to search for specific players by their last name. We wanted our stage of our story to provide deeper analysis so we think that the search bar can help users focus on specific batters. We also implemented different sets in Tableau that can be applied to different dashboards. For example, we created a set by highlighting players who had an above average strikeout percentage and on-base percentage. This set can be applied to our final dashboard to see if the players who have a high amount of runs created are also efficient swingers. This same process was used to group players with an ideal combination of slugging percentage and projected seasons remaining. We also with dealt with grouping years so that the data showed the most up to date information of each player.

Evaluation

One of the biggest limitations of our visualization is that it does identify players who are free agents or will become free agents. If we included this in our analysis then we would be able to identify players that we could sign in the future.

Overall, we believe that our visualizations are effective in answering the questions that we were interested in. We were able to analyze the relationship between a player's historical performance, future performance, and salary level. We think that our dashboards highlight the most important statistics for assessing baseball hitters in the MLB. One way that we could potentially improve this analysis is by including statistics such as WAR. We also like for our visualizations to be able to use real time data so that it can still be used to assess how well a player is doing during the current season.