MLB OBP (On-base Percentage) and SLG (Slugging Percentage) Analysis

Britton Green

ALY6010 Probability Theory and Introductory Statistics

Northeastern University

4/19/2025

ALY6010: R Practice 1 Assignment

Introduction

For the 2nd R Practice assignment, I chose to analyze MLB OBP (On-base Percentage) and SLG (Slugging Percentage) data for the 2023 – 2024 Boston Red Sox team (Baseball Reference, 2025). I loaded all of the required libraries I would need for this analysis, including tidyverse, ggplot2, janitor, psych, knitr, markdown, hrbrthemes and viridis in RStudio. I imported the data using the read.csv function and created a dataframe called “red_sox_hitters_all”. The raw imported file has 22 observations and 32 variables, such as Player, Age, BA (Batting Average), OBP (On-base Percentage), SLG (Slugging Percentage), among many others.

Data Cleaning:

Before any analysis, I had to perform minor cleaning functions. I created a new datafame called “red_sox_hitters_wk” to perform all of my cleaning functions on. Using the “janitor” package, I leveraged the “clean_names” function to flatten all of the column titles to lowercase for consistency. Next, I wanted to remove all non-imperative variables using the “select” function to further organize the data including “rk”, “awards”, and “player_additional”. The last requirement was to clean the “player” column. The majority of values under the “player” variable (player names) either had an * or # next to their name due to the nature of how the raw data was formatted. I needed to perform a “mutate” function leveraging “str_remove_all” to remove either of those characters from the cells under the “Player” column. (See below for how the * and # were removed from original “red_sox_hitters_all” dataframe to the new “red_sox_hitters_wk” dataframe).

```
> red_sox_hitters_all$Player
 [1] "Connor Wong"        "Dominic Smith*"     "Enmanuel Valdéz*"
 [7] "Jarren Duran*"      "Wilyer Abreu*"      "Masataka Yoshida*
[13] "Romy González"      "Reese McGuire*"     "Vaughn Grissom"
[19] "Nick Sogard#"       "Garrett Cooper"     "Pablo Reyes"
> red_sox_hitters_wk$player
 [1] "Connor Wong"       "Dominic Smith"     "Enmanuel Valdéz"   "
 [7] "Jarren Duran"      "Wilyer Abreu"      "Masataka Yoshida"  "
[13] "Romy González"     "Reese McGuire"     "Vaughn Grissom"    "
[19] "Nick Sogard"       "Garrett Cooper"    "Pablo Reyes"       "
```

Data Analysis

To begin my analysis of OBP and SLG, I used the “describe” function leveraging the “psych” package to calculate a comprehensive summary of the descriptive statistics of these 2 variables.

- N = 22 observations (players)
- Mean = 0.301/0.374, Median = 0.320/0.392
- SD = 0.046/0.098
- Min = 0.217/0.193, Max = 0.359/0.516

Given Slugging (SLG) is a statistic that takes into account all hits but does not weigh them equally like BA (Batting Average). While Batting Average is calculated by dividing the total number of hits by the total number of at-bats, the formula for slugging percentage is: (1B + 2Bx2 + 3Bx3 + HRx4)/AB. OBP only takes into account whether or not you reached a base but it does include situations such as walks and hit-by-pitch, which SLG does not (More detail on this to come). I stored these values in a dataframe called “hitter_stats”:

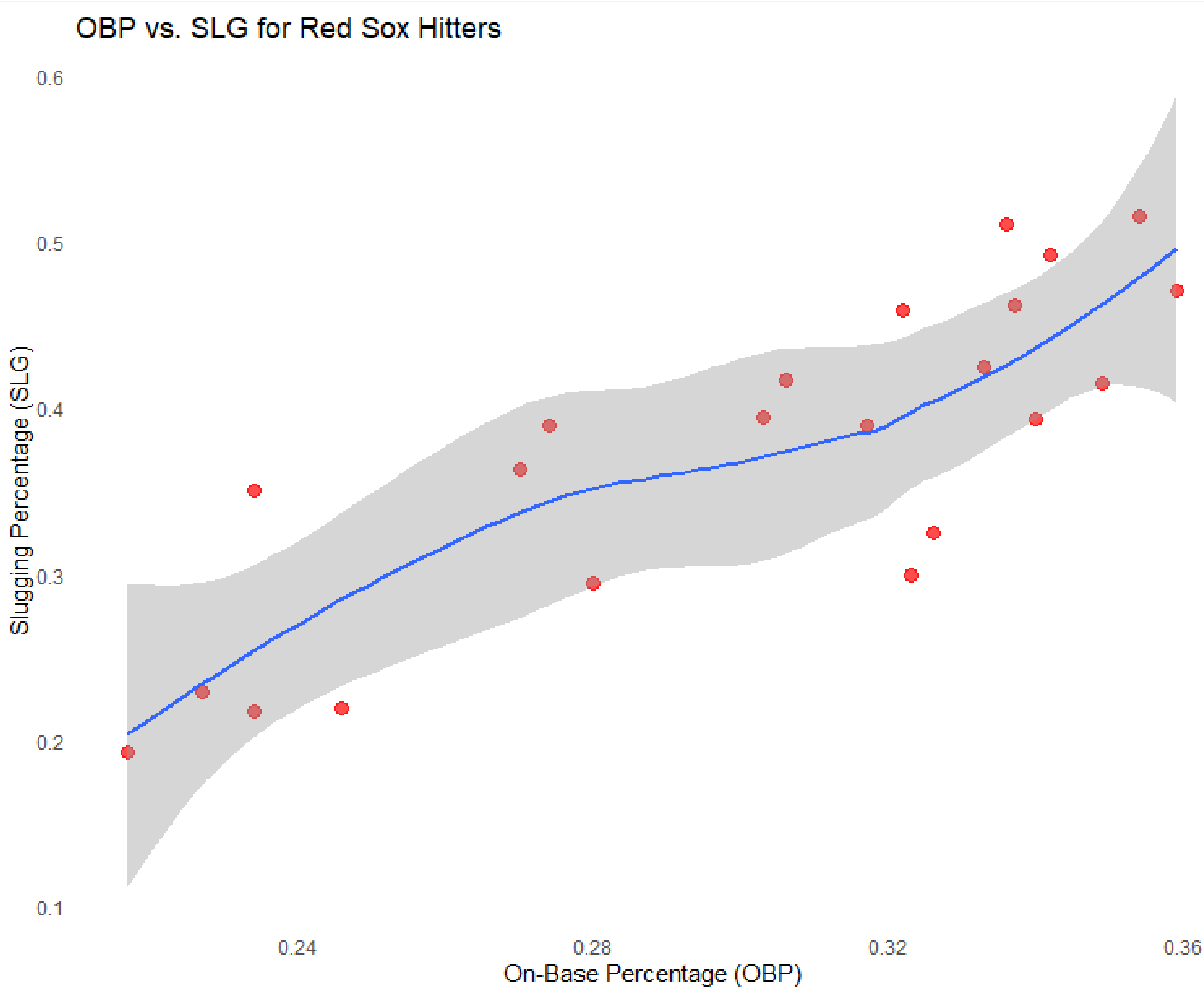
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
OBP	1	22	0.301	0.046	0.320	0.304	0.039	0.217	0.359	0.142	-0.537	-1.248	0.010
SLG	2	22	0.374	0.098	0.392	0.377	0.102	0.193	0.516	0.323	-0.386	-1.063	0.021

Additionally, I created a “kable” leveraging the “knitr” package to visualize the same data but in a different format:

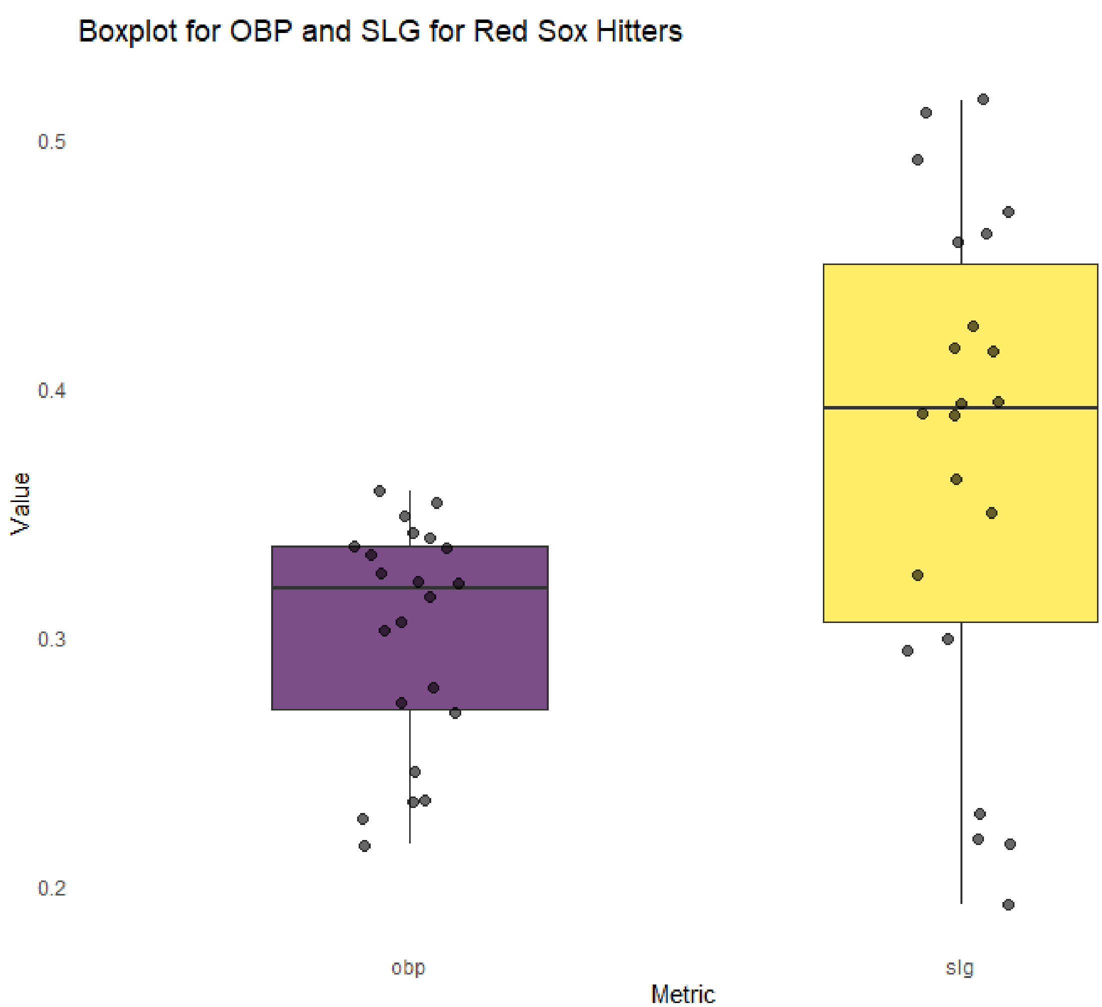
Table: Descriptive Statistics for OBP and SLG

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
OBP	1	22	0.301	0.046	0.320	0.304	0.039	0.217	0.359	0.142	-0.537	-1.248	0.010
SLG	2	22	0.374	0.098	0.392	0.377	0.102	0.193	0.516	0.323	-0.386	-1.063	0.021

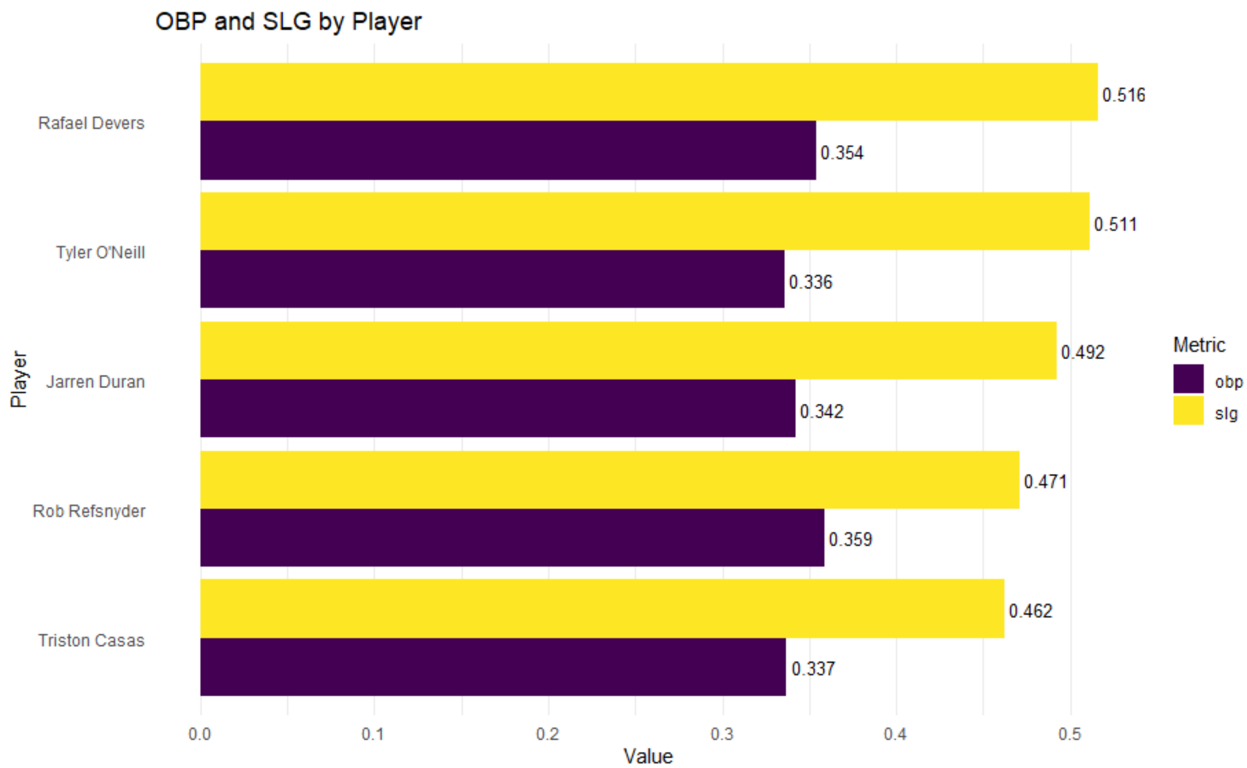
To analyze a potential correlation between OBP and SLG percentage, I created a scatterplot and included a "geom_smooth" function to track the plots. Reviewing the chart, there is a clear correlation showing that hitters who reach the base more often (OBP) also have a higher Slugging % (SLG). This seems like an obvious answer but it was still worth exploring to see how much of a correlation exists and what the variance is between the players on the Boston Red Sox.



In addition to the scatterplot analyzing correlation, I wanted to analyze the datapoints individually with a boxplot to identify potential outliers and how it may affect this data. Leveraging the "hrbrthemes" and "viridis" packages, I used geom_jitter to overlay the boxplots with the individual datapoints to determine the variance from the best to worst hitters. The result was extremely interesting, as shown below, the OBP data is much more compact (Mean = 0.301) than the SLG data and, as was already established, had a lower mean than SLG (0.374). Most interestingly, the "best" hitters (at least in terms of SLG) could potentially be hitting at nearly double their OBP. Given higher values are attributed to plays in which a hitter gets a double, triple or home run, this would explain the larger gap between the 2 variables.



Following up on the above findings, I want to finish by analyzing the players with the top 5 (out of 22) SLG % and compare them to their OBP % to visualize the gap. Per the outliers above, we established there could be a large discrepancy in SLG vs OBP.



Based on suspicions and previous analyses, the difference between SLG and OBP were not quite as large as "double" for the top hitters, however, they are significant. The outliers in our boxplot clearly show there are extremely high-performing hitters. Additionally, they also directly correspond to the highest OBP percentages which confirms there is not an indirect correlation to OBP (i.e. high SLG and low OBP or vice-versa). The same players who get extra bases on their hits (i.e. higher SLG) get on base more often (whether through straight hits or making the pitchers "work" for the out resulting higher-than-average walking %).

References

Baseball Reference. (2024). *Boston Red Sox*. Baseball-Reference.com. https://www.baseball-reference.com/teams/BOS/2024.shtml

Bluman, A. G. (2018). *Elementary statistics : a step by step approach*. New York, Ny: Mcgraw-Hill Education.

Holtz, Y. (n.d.-a). *Boxplot | the R Graph Gallery*. R-Graph-Gallery.com. https://r-graph-gallery.com/boxplot.html

Holtz, Y. (n.d.-b). *Histogram | the R Graph Gallery*. R-Graph-Gallery.com. https://r-graph-gallery.com/histogram.html

Holtz, Y. (n.d.-c). *Linear model and confidence interval in ggplot2*. Www.r-Graph-Gallery.com. https://r-graph-gallery.com/50-51-52-scatter-plot-with-ggplot2.html

Jittered points — geom_jitter. (n.d.). Ggplot2.Tidyverse.org. https://ggplot2.tidyverse.org/reference/geom_jitter.html

Kabacoff, R. I. (2015). *R in Action*. Simon and Schuster.

Northeastern University – Canvas – Module 2 learning materials provided by Professor Goulding

Posit Sotware, PBC. (2018, March 6). *Knit button not working in Rstudio*. Posit Community. https://forum.posit.co/t/knit-button-not-working-in-rstudio/5876