

# Baseball Project

*Nicole Kadosh and Audrey Holloman*

*10/4/2018*

For this project we cleaned the dataset Teams by filtering it through dplyr. We filtered the new dataset, “data”, to show from year 2000 to 2016 and the five teams, which are Boston Red Socks, Cleveland Indians, Atlanta Braves, New York Yankees, and Chicago Cubs. We then added columns to the dataset to calculate the On-Base plus Slugging, OPS, and Batting Average on Balls in Play, BABIP, values for each team for each year.

```
if (!require('Lahman'))
{
  install.packages('Lahman');
  library(Lahman);
}

## Loading required package: Lahman

# install.packages(Lahman)
# library(Lahman)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(gapminder)
library(ggplot2)
data <- Teams %>% filter (yearID > 1999, franchID %in% c("BOS", "CLE", "ATL", "NYY", "CHC"))

# View(data)
attach(data)

TB <- (H + X2B + (2*X3B) + (3*HR))
data$OPS <- ((TB/AB) + ((H + BB + HBP) / (AB + BB + SF + HBP)))
data$BABIP <- ((H - HR) / (AB - SO - HR + SF))

ggplot(data, aes(x = yearID, y = OPS, color = franchID)) + geom_line()
```



The time series graph above is showing the OPS of each of the teams from 2000 through 2016. During this time frame all these teams are inconsistent since their time series oscillate. The biggest thing we noticed is that around 2013 all of the teams dropped.

```
ggplot(data, aes(x = yearID, y = BABIP, color = franchID)) + geom_line()
```



The time series graph above is showing the BABIP of each of the teams from 2000 through 2016. This is showing that the Boston Red Sox have the most consistent BABIP of the teams, as well as the highest peak. The Cleveland Indians have an inconsistent and oscillating time series.

The scatter plots below compare OPS and BABIP to Winning Percentage (WP) for different years in order to determine how important hitting is.

```

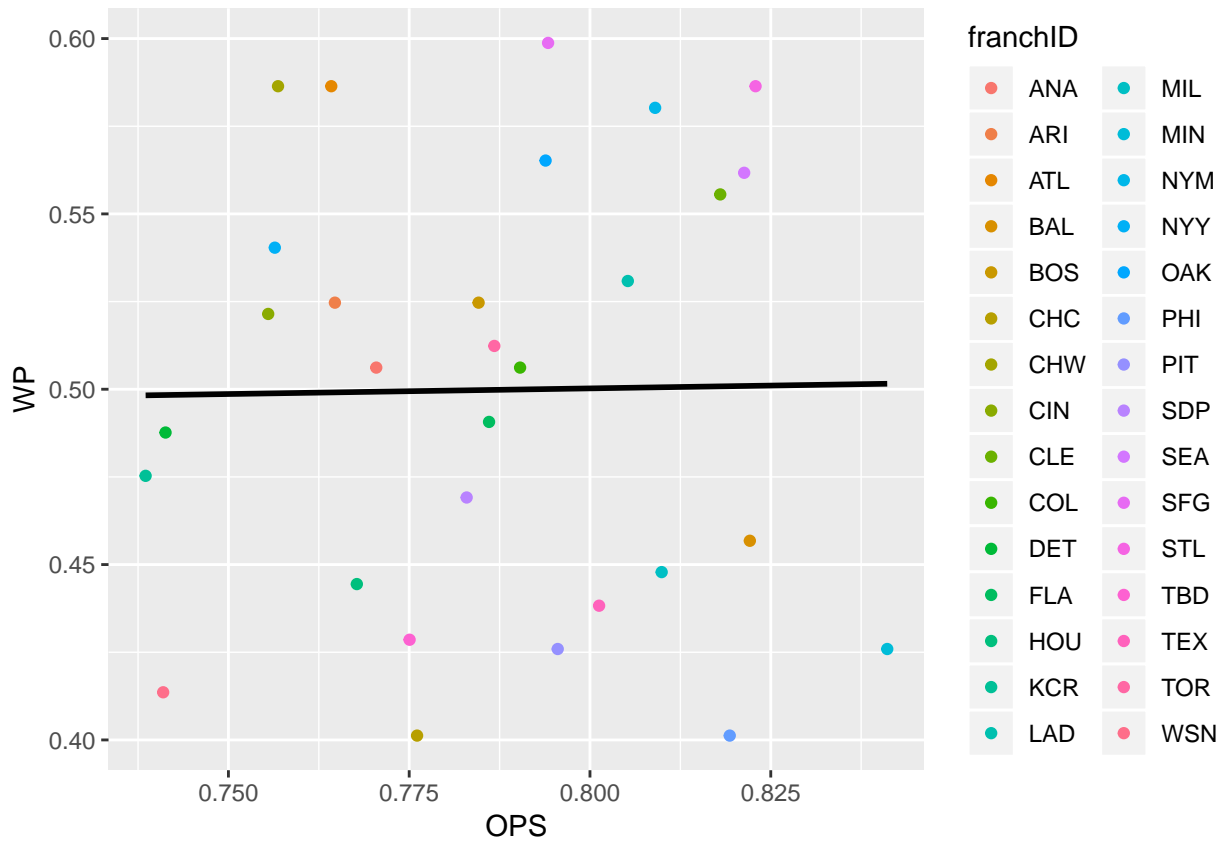
dataa <- Teams %>% filter(yearID>1999)
# View(dataa)
attach(dataa)

## The following objects are masked from data:
##
## AB, attendance, BB, BBA, BPF, CG, CS, divID, DivWin, DP, E,
## ER, ERA, FP, franchID, G, Ghome, H, HA, HBP, HR, HRA, IPouts,
## L, lgID, LgWin, name, park, PPF, R, RA, Rank, SB, SF, SHO, SO,
## SOA, SV, teamID, teamIDBR, teamIDlahman45, teamIDretro, W,
## WCWin, WSWin, X2B, X3B, yearID

dataa$OPS <- ((TB/AB) + ((H + BB + HBP) / (AB + BB + SF + HBP)))
dataa$BABIP <- ((H - HR) / (AB - SO - HR + SF))

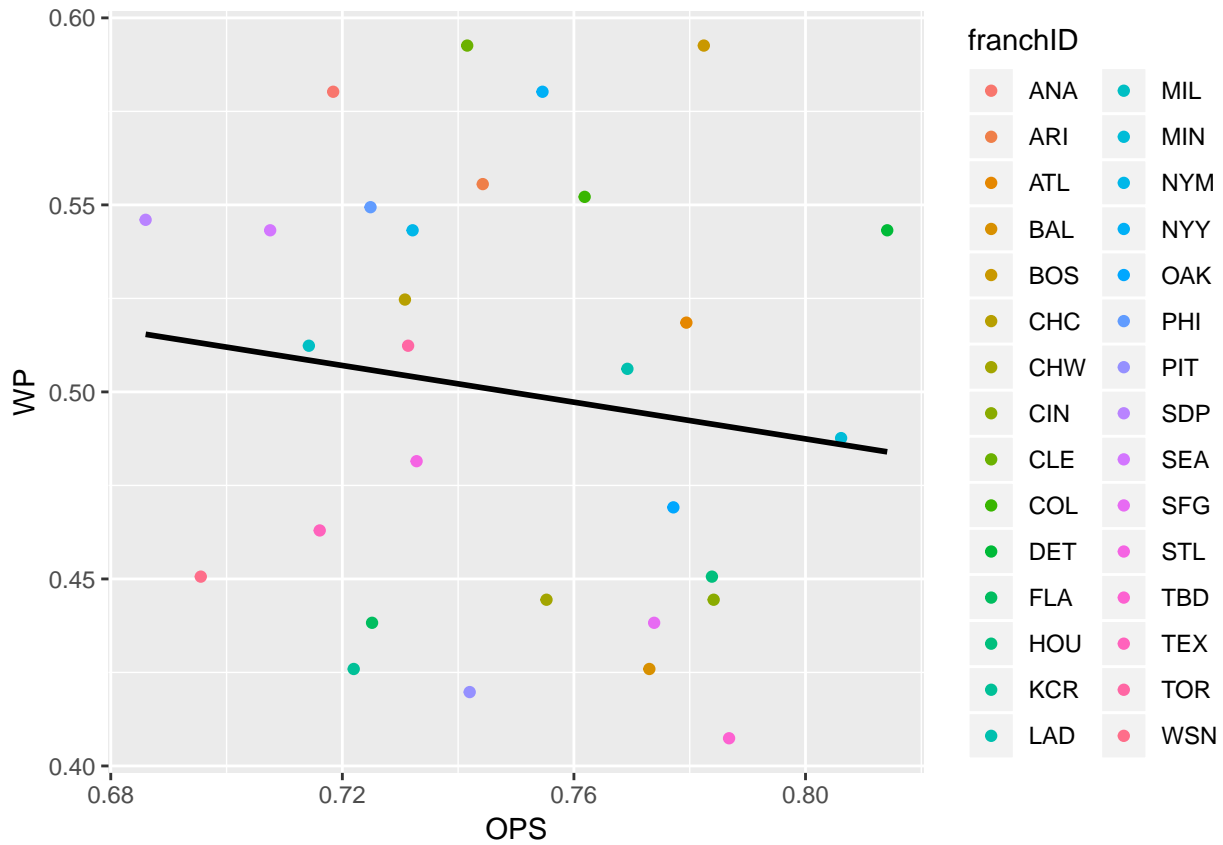
dataa$WP <- (W / G)
new <- dataa %>% filter(yearID == 2000)
ggplot(new, aes(x = OPS, y = WP, color = franchID)) + geom_point() + geom_smooth(method = "lm", colour =

```



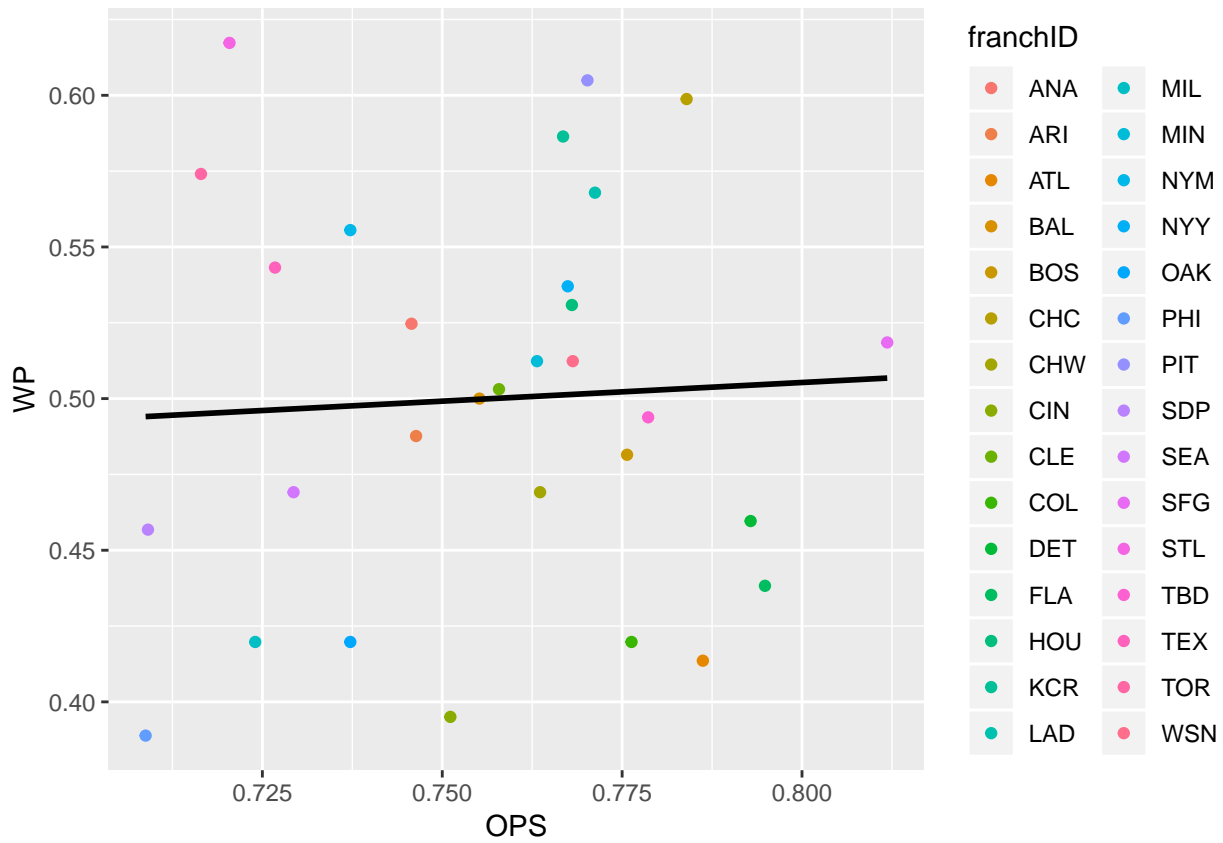
The scatter plots above is showing the OPS and WP for every team in year 2007.

```
new2 <- dataa %>% filter(yearID == 2007)
ggplot(new2, aes(x = OPS, y = WP, color = franchID)) + geom_point() + geom_smooth(method = "lm", colour = "black")
```



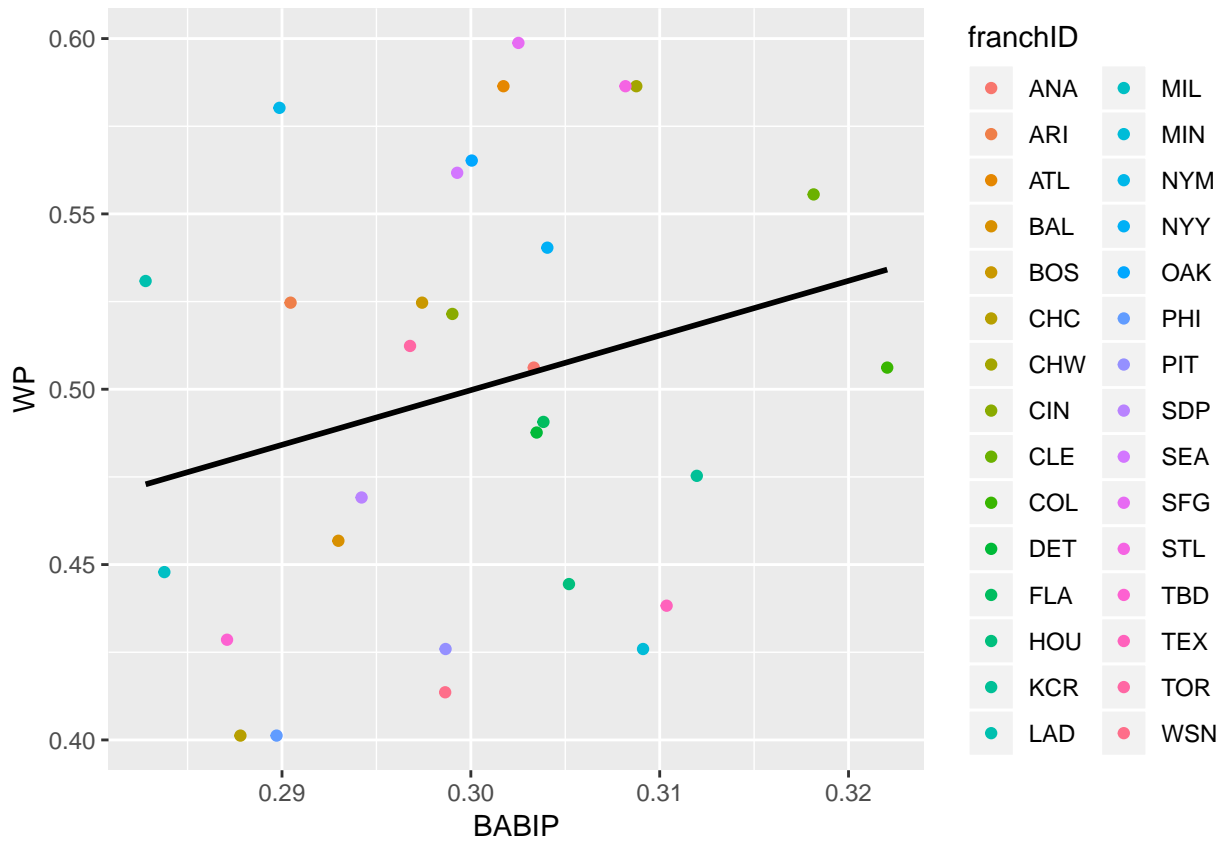
The scatter plots above is showing the OPS and WP for every team in year 2007.

```
new3 <- dataa %>% filter(yearID == 2015)
ggplot(new3, aes(x = OPS, y = WP, color = franchID)) + geom_point() + geom_smooth(method = "lm", colour
```



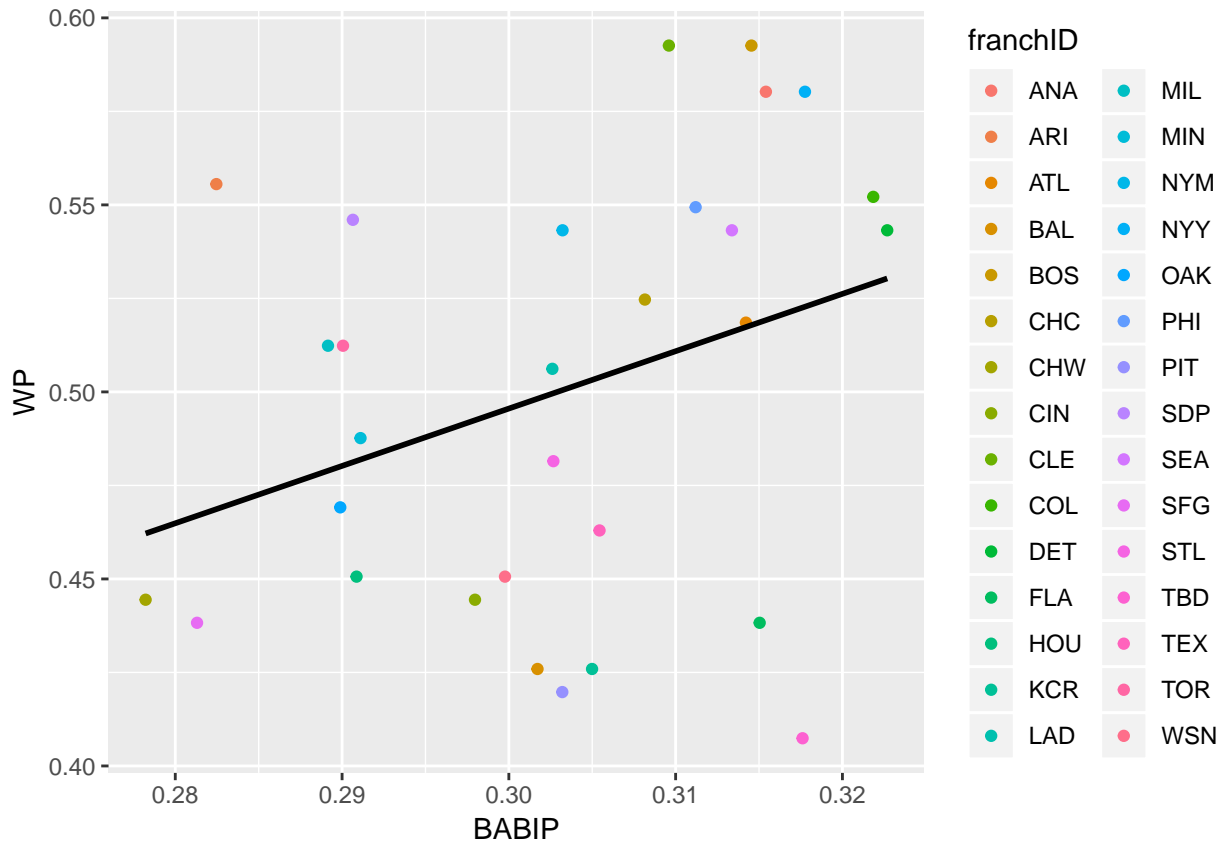
The scatter plots above is showing the OPS and WP for every team in year 2015.

```
new4 <- dataa %>% filter(yearID == 2000)
ggplot(new4, aes(x = BABIP, y = WP, color = franchID)) + geom_point() + geom_smooth(method = "lm", color = "black")
```



The scatter plots above is showing the BABIP and WP for every team in year 2007.

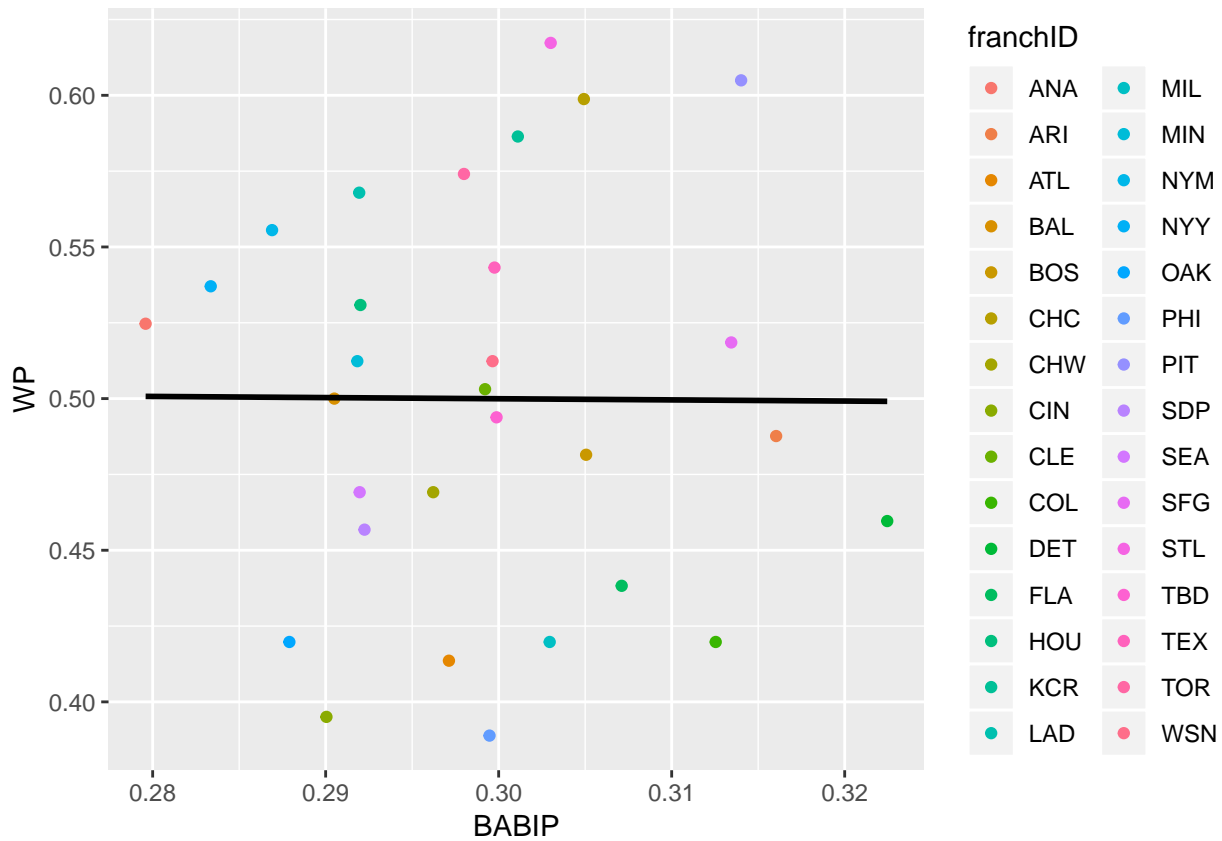
```
new5 <- dataa %>% filter(yearID == 2007)
ggplot(new5, aes(x = BABIP, y = WP, color = franchID)) + geom_point() + geom_smooth(method = "lm", color = "black")
```



The scatter plots above is showing the BABIP and WP for every team in year 2007.

```
new6 <- dataa %>% filter(yearID == 2015)
ggplot(new6, aes(x = BABIP, y = WP, color = franchID)) + geom_point() + geom_smooth(method = "lm", color
```





The scatter plots above is showing the BABIP and WP for every team in year 2015.

Overall, OPS is a better indicator than BABIP of WP. OPS is a good indicator of WP since teams with higher WP typically have higher OPS. However, there are some teams who had higher OPS values that gave lower WP values. BABIP is also a good indicator of WP except it has more inconsistencies. Therefore these scatter plots are showing that hitting is important, but is not the only factor in WP.