Project 3: Profiling a Data Set

Anthony Pagan

Monday, October 13, 2014

This data was retrieved from http://www.amstat.org/publications/jse/v6n2/datasets.watnik.html. It provides baseball statistcis for players including their salaries from 1991-1992 season.

Here is a summray of the data:

```
tf <- "C:\\Users\\Development\\Desktop\\CUNYSPS\\IS360\\bbproj.csv"
tb <- read.table(file = tf, header = TRUE, stringsAsFactors = FALSE, sep =
",")
bb <- data.frame(tb)
require(dplyr)
```

```
## Loading required package: 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
```

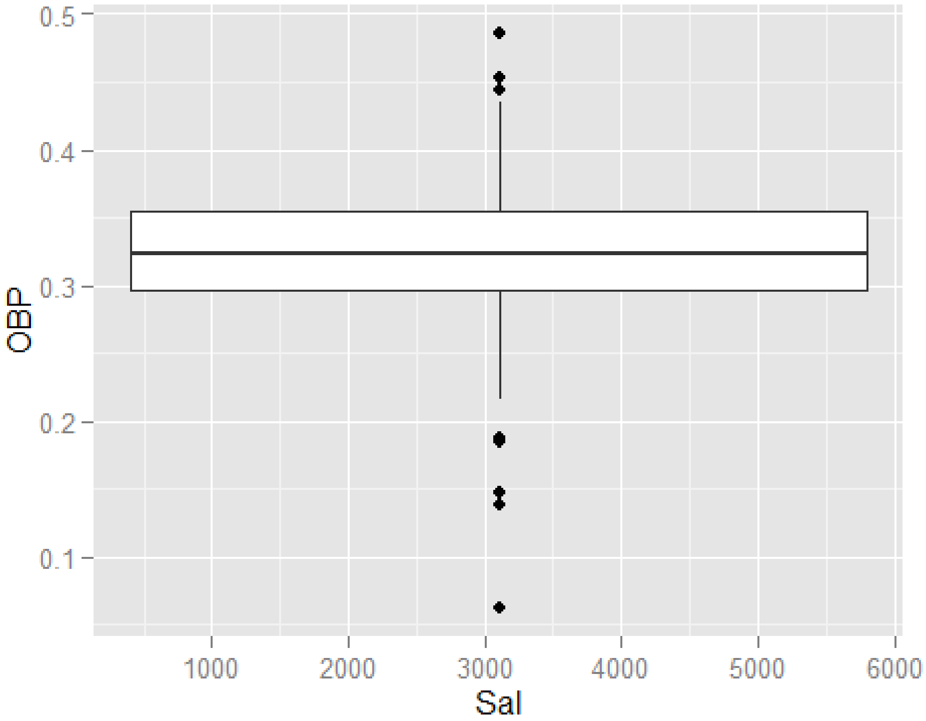
```
obp <- select(bb, Pname, Sal, BA, OBP)
summary(obp)
```

```
##     Pname                Sal               BA              OBP
##  Length:337         Min.   : 109   Min.   :0.063   Min.   :0.063
##  Class :character   1st Qu.: 230   1st Qu.:0.238   1st Qu.:0.297
##  Mode  :character   Median : 740   Median :0.260   Median :0.323
##                     Mean   :1249   Mean   :0.258   Mean   :0.324
##                     3rd Qu.:2150   3rd Qu.:0.281   3rd Qu.:0.354
##                     Max.   :6100   Max.   :0.457   Max.   :0.486
```

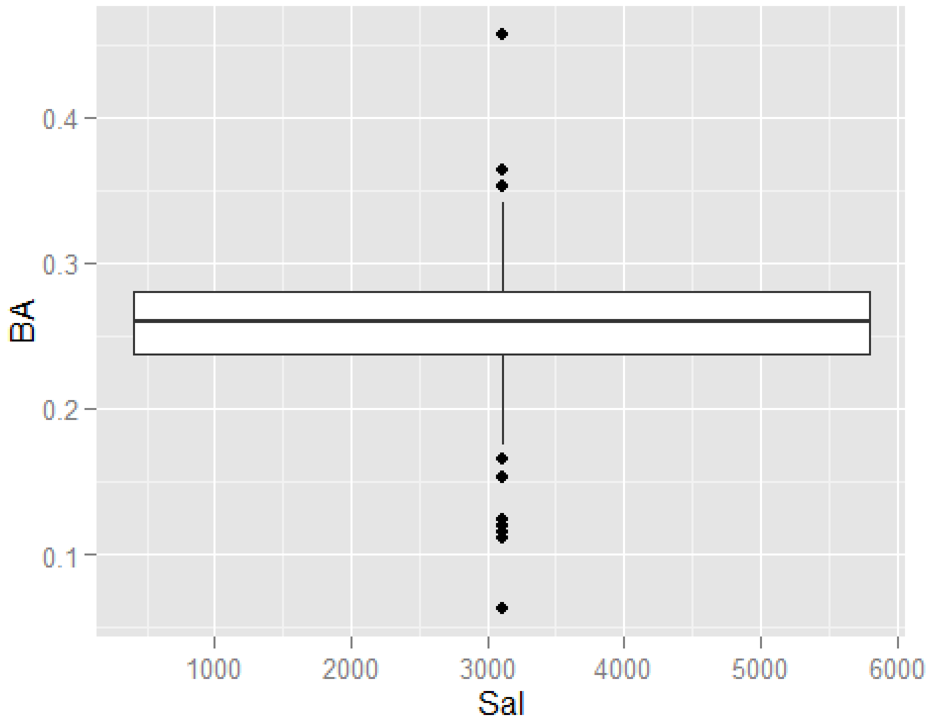
My objective for analysis was to see if player salaries correlated to "on base percentage" or "batting average".

Below shows 2 box plots. One box plot comapares OBP to salaries. The first box plot has no skewing, but does have several outliers. The second box plot compares batting average to salaries. This box plot has some skewing towoard a lower batting average and has several outliers.

```
library(ggplot2)
require(ggplot2)
ggplot(obp, aes(x = Sal, y = OBP)) + geom_boxplot()
```



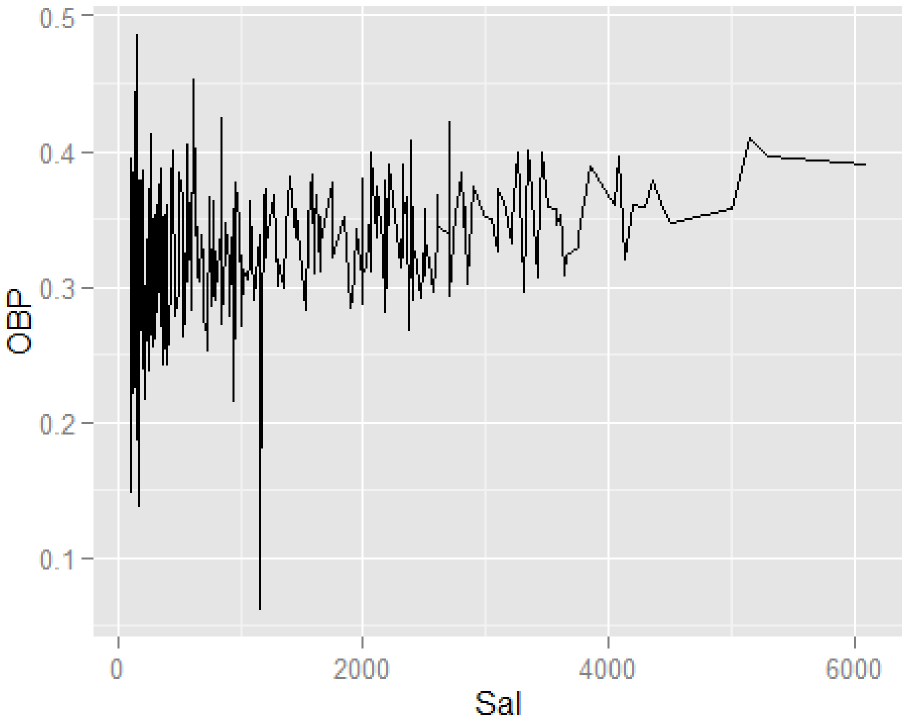
```
ggplot(obp, aes(x = Sal, y = BA)) + geom_boxplot()
```



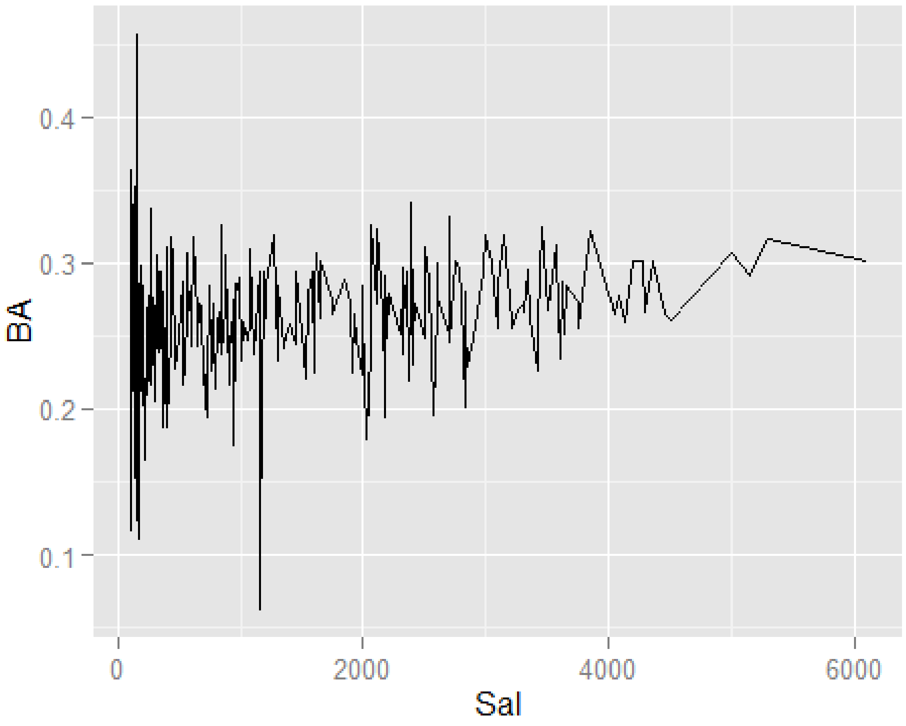
The next two charts are line charts. These also compare OBP to salaries and batting average to salaries. Both charts have an uptrend showing that higher salary has some correlation to higher batting average and OBP.

One argument to this analysis would be, the salary increase can be due to teams paying players more because of their OBP and BA performance or the BA and OBP could have increased because of the higher salary. My theory is that players are being paid more because of their performance. To investigate further we would need players statistics and salary increases from year to year throughout their careers and further analyze.

```
ggplot(obp, aes(x = Sal, y = OBP)) +geom_line()
```



```
ggplot(obp, aes(x = Sal, y = BA)) + geom_line()
```



x