MoneyBall Project

2024-02-16

Intro: In this project we'll work with some data and with the goal of trying to find replacement players for the ones lost at the start of the off-season - During the 2001–02 offseason, the team lost three key free agents to larger market teams: 2000 AL MVP Jason Giambi to the New York Yankees, outfielder Johnny Damon to the Boston Red Sox, and closer Jason Isringhausen to the St. Louis Cardinals.

The main goal of this project is for you to feel comfortable working with R on real data to try and derive actionable insights!

Introduction of Features: "G"
"G_batting" "AB" = At bat
"R" = Runs
"H" = Hits
"2B" = Doubles
"3B" = Triples
"HR" = Home Runs
"RBI" = Runs Batted In
"SB" = Stolen Bases
"CS" = Caught Stealing
"BB" = Bases on Balls (Walks)
"SO" = Strikeouts
"IBB" = Intentional Baseson Balls(Walks)
"HBP" = Hit By Pitch
"SH" = Sacrifice Hits (Bunts)
"SF" = Sacrifice fly
"GIDP" = Ground into Double Plays
"G_old" = Metals

```r
library(data.table)
library(tidyr)
library(dplyr)
```

Goal: Help the Oakland A's recruit under-valued baseball players.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(readr)

batting <- read_csv('/Users/mac/Desktop/Capstone Project/Batting.csv')
```

```
## Rows: 97889 Columns: 24
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): playerID, teamID, lgID
## dbl (21): yearID, stint, G, G_batting, AB, R, H, 2B, 3B, HR, RBI, SB, CS, BB...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```


head(batting)
```

1. Take a glance into the dataset

```
## # A tibble: 6 x 24
##   playerID  yearID stint teamID lgID      G G_batting    AB     R     H ‘2B‘
##   <chr>      <dbl> <dbl> <chr>  <chr> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 aardsda01   2004     1 SFN    NL       11        11     0     0     0     0
## 2 aardsda01   2006     1 CHN    NL       45        43     2     0     0     0
## 3 aardsda01   2007     1 CHA    AL       25         2     0     0     0     0
## 4 aardsda01   2008     1 BOS    AL       47         5     1     0     0     0
## 5 aardsda01   2009     1 SEA    AL       73         3     0     0     0     0
## 6 aardsda01   2010     1 SEA    AL       53         4     0     0     0     0
## # i 13 more variables: ‘3B‘ <dbl>, HR <dbl>, RBI <dbl>, SB <dbl>, CS <dbl>,
## #   BB <dbl>, SO <dbl>, IBB <dbl>, HBP <dbl>, SH <dbl>, SF <dbl>, GIDP <dbl>,
## #   G_old <dbl>
```

```
str(batting)
```

```
## spc_tbl_ [97,889 x 24] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ playerID : chr [1:97889] "aardsda01" "aardsda01" "aardsda01" "aardsda01" ...
##  $ yearID   : num [1:97889] 2004 2006 2007 2008 2009 ...
##  $ stint    : num [1:97889] 1 1 1 1 1 1 1 1 1 1 ...
##  $ teamID   : chr [1:97889] "SFN" "CHN" "CHA" "BOS" ...
##  $ lgID     : chr [1:97889] "NL" "NL" "AL" "AL" ...
##  $ G        : num [1:97889] 11 45 25 47 73 53 1 122 153 153 ...
##  $ G_batting: num [1:97889] 11 43 2 5 3 4 NA 122 153 153 ...
##  $ AB       : num [1:97889] 0 2 0 1 0 0 NA 468 602 609 ...
##  $ R        : num [1:97889] 0 0 0 0 0 0 NA 58 105 106 ...
##  $ H        : num [1:97889] 0 0 0 0 0 0 NA 131 189 200 ...
##  $ 2B       : num [1:97889] 0 0 0 0 0 0 NA 27 37 34 ...
##  $ 3B       : num [1:97889] 0 0 0 0 0 0 NA 6 9 14 ...
##  $ HR       : num [1:97889] 0 0 0 0 0 0 NA 13 27 26 ...
##  $ RBI      : num [1:97889] 0 0 0 0 0 0 NA 69 106 92 ...
##  $ SB       : num [1:97889] 0 0 0 0 0 0 NA 2 3 2 ...
##  $ CS       : num [1:97889] 0 0 0 0 0 0 NA 2 1 4 ...
##  $ BB       : num [1:97889] 0 0 0 0 0 0 NA 28 49 37 ...
##  $ SO       : num [1:97889] 0 0 0 1 0 0 NA 39 61 54 ...
##  $ IBB      : num [1:97889] 0 0 0 0 0 0 NA NA 5 6 ...
##  $ HBP      : num [1:97889] 0 0 0 0 0 0 NA 3 3 2 ...
##  $ SH       : num [1:97889] 0 1 0 0 0 0 NA 6 7 5 ...
##  $ SF       : num [1:97889] 0 0 0 0 0 0 NA 4 4 7 ...
##  $ GIDP     : num [1:97889] 0 0 0 0 0 0 NA 13 20 21 ...
##  $ G_old    : num [1:97889] 11 45 2 5 NA NA NA 122 153 153 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   playerID = col_character(),
##   ..   yearID = col_double(),
##   ..   stint = col_double(),
##   ..   teamID = col_character(),
##   ..   lgID = col_character(),
##   ..   G = col_double(),
##   ..   G_batting = col_double(),
##   ..   AB = col_double(),
##   ..   R = col_double(),
##   ..   H = col_double(),
##   ..   ‘2B‘ = col_double(),
##   ..   ‘3B‘ = col_double(),
##   ..   HR = col_double(),
##   ..   RBI = col_double(),
##   ..   SB = col_double(),
##   ..   CS = col_double(),
##   ..   BB = col_double(),
##   ..   SO = col_double(),
##   ..   IBB = col_double(),
##   ..   HBP = col_double(),
##   ..   SH = col_double(),
##   ..   SF = col_double(),
##   ..   GIDP = col_double(),
##   ..   G_old = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
head(batting$AB,5)
```

```
## [1] 0 2 0 1 0
```

```r
head(batting[,'2B'])
```

```
## # A tibble: 6 x 1
##    ‘2B‘
##   <dbl>
## 1     0
## 2     0
## 3      0
## 4      0
## 5      0
## 6      0
```

####2. Feature Engineering

Firstly, we need to calculate 3 more statistics : (a) Batting Average : The measure of the performance of batter - AVG = the number of hits divided by at bats (b) On Base Percentage: The measure of how frequently a batter reach a base - OBP = (H+BB+HBP)/(AB+BB+HBP+SF) (c) Slugging Percentage: The measure of batting productivity of a hitter. -SLG = (1B + 2*2B + 3*3B + 4*HR)/AB

```
# (a)
batting$BA <- batting$H / batting$AB
# Alternative Way
mutate(batting, BA = H/AB)
```

```
## # A tibble: 97,889 x 25
##    playerID  yearID stint teamID lgID      G G_batting    AB     R     H  ‘2B‘
##    <chr>      <dbl> <dbl> <chr>  <chr> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 aardsda01   2004     1 SFN    NL       11        11     0     0     0     0
##  2 aardsda01   2006     1 CHN    NL       45        43     2     0     0     0
##  3 aardsda01   2007     1 CHA    AL       25         2     0     0     0     0
##  4 aardsda01   2008     1 BOS    AL       47         5     1     0     0     0
##  5 aardsda01   2009     1 SEA    AL       73         3     0     0     0     0
##  6 aardsda01   2010     1 SEA    AL       53         4     0     0     0     0
##  7 aardsda01   2012     1 NYA    AL        1        NA    NA    NA    NA    NA
##  8 aaronha01   1954     1 ML1    NL      122       122   468    58   131    27
##  9 aaronha01   1955     1 ML1    NL      153       153   602   105   189    37
## 10 aaronha01   1956     1 ML1    NL      153       153   609   106   200    34
## # i 97,879 more rows
## # i 14 more variables: ‘3B‘ <dbl>, HR <dbl>, RBI <dbl>, SB <dbl>, CS <dbl>,
## #   BB <dbl>, SO <dbl>, IBB <dbl>, HBP <dbl>, SH <dbl>, SF <dbl>, GIDP <dbl>,
## #   G_old <dbl>, BA <dbl>
```

```
tail(batting$BA, 5)
```

```
## [1] 0.1230769 0.2746479 0.1470588 0.2745098 0.2138728
```

```
# (b)
batting$OBP <- (batting$H + batting$BB + batting$HBP) / (batting$AB + batting$BB + batting$HBP + battin

#(c)
#1B = H-2B-3B-HR
batting$'1B' <- batting$H - batting$'2B' - batting$'3B' - batting$HR
batting$SLG <- (batting$'1B' + 2*batting$'2B' + 3*batting$'3B' + 4*batting$HR)/batting$AB

str(batting)
```

```
## spc_tbl_ [97,889 x 28] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ playerID : chr [1:97889] "aardsda01" "aardsda01" "aardsda01" "aardsda01" ...
##  $ yearID   : num [1:97889] 2004 2006 2007 2008 2009 ...
##  $ stint    : num [1:97889] 1 1 1 1 1 1 1 1 1 1 ...
##  $ teamID   : chr [1:97889] "SFN" "CHN" "CHA" "BOS" ...
##  $ lgID     : chr [1:97889] "NL" "NL" "AL" "AL" ...
##  $ G        : num [1:97889] 11 45 25 47 73 53 1 122 153 153 ...
##  $ G_batting: num [1:97889] 11 43 2 5 3 4 NA 122 153 153 ...
##  $ AB       : num [1:97889] 0 2 0 1 0 0 NA 468 602 609 ...
##  $ R        : num [1:97889] 0 0 0 0 0 0 NA 58 105 106 ...
##  $ H        : num [1:97889] 0 0 0 0 0 0 NA 131 189 200 ...
##  $ 2B       : num [1:97889] 0 0 0 0 0 0 NA 27 37 34 ...
##  $ 3B       : num [1:97889] 0 0 0 0 0 0 NA 6 9 14 ...
##  $ HR       : num [1:97889] 0 0 0 0 0 0 NA 13 27 26 ...
##  $ RBI      : num [1:97889] 0 0 0 0 0 0 NA 69 106 92 ...
##  $ SB       : num [1:97889] 0 0 0 0 0 0 NA 2 3 2 ...
##  $ CS       : num [1:97889] 0 0 0 0 0 0 NA 2 1 4 ...
##  $ BB       : num [1:97889] 0 0 0 0 0 0 NA 28 49 37 ...
##  $ SO       : num [1:97889] 0 0 0 1 0 0 NA 39 61 54 ...
##  $ IBB      : num [1:97889] 0 0 0 0 0 0 NA NA 5 6 ...
##  $ HBP      : num [1:97889] 0 0 0 0 0 0 NA 3 3 2 ...
##  $ SH       : num [1:97889] 0 1 0 0 0 0 NA 6 7 5 ...
##  $ SF       : num [1:97889] 0 0 0 0 0 0 NA 4 4 7 ...
##  $ GIDP     : num [1:97889] 0 0 0 0 0 0 NA 13 20 21 ...
##  $ G_old    : num [1:97889] 11 45 2 5 NA NA NA 122 153 153 ...
##  $ BA       : num [1:97889] NaN 0 NaN 0 NaN ...
##  $ OBP      : num [1:97889] NaN 0 NaN 0 NaN ...
##  $ 1B       : num [1:97889] 0 0 0 0 0 0 NA 85 116 126 ...
##  $ SLG      : num [1:97889] NaN 0 NaN 0 NaN ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   playerID = col_character(),
##   ..   yearID = col_double(),
##   ..   stint = col_double(),
##   ..   teamID = col_character(),
##   ..   lgID = col_character(),
##   ..   G = col_double(),
##   ..   G_batting = col_double(),
##   ..   AB = col_double(),
##   ..   R = col_double(),
##   ..   H = col_double(),
##   ..   `2B` = col_double(),
##   ..   `3B` = col_double(),
##   ..   HR = col_double(),
##   ..   RBI = col_double(),
##   ..   SB = col_double(),
##   ..   CS = col_double(),
##   ..   BB = col_double(),
##   ..   SO = col_double(),
##   ..   IBB = col_double(),
##   ..   HBP = col_double(),
##   ..   SH = col_double(),
##   ..   SF = col_double(),
##   ..   GIDP = col_double(),
##   ..   G_old = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

3. Merger batting dataframe with salary.csv We want to find the most undervalue player, thus, it is worth to look into salary dataset

```
salary <- read_csv('/Users/mac/Desktop/Capstone Project/Salaries.csv')
```

```
## Rows: 23956 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): teamID, lgID, playerID
## dbl (2): yearID, salary
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(salary)
```

```
## # A tibble: 6 x 5
##   yearID teamID lgID  playerID   salary
##    <dbl> <chr>  <chr> <chr>       <dbl>
## 1   1985 BAL    AL    murraed02 1472819
## 2   1985 BAL    AL    lynnfr01  1090000
## 3   1985 BAL    AL    ripkeca01  800000
## 4   1985 BAL    AL    lacyle01   725000
## 5   1985 BAL    AL    flanami01  641667
## 6   1985 BAL    AL    boddimi01  625000
```

```
arrange(salary, yearID)
```

```
## # A tibble: 23,956 x 5
##    yearID teamID lgID  playerID   salary
##     <dbl> <chr>  <chr> <chr>       <dbl>
##  1   1985 BAL    AL    murraed02 1472819
##  2   1985 BAL    AL    lynnfr01  1090000
##  3   1985 BAL    AL    ripkeca01  800000
##  4   1985 BAL    AL    lacyle01   725000
##  5   1985 BAL    AL    flanami01  641667
##  6   1985 BAL    AL    boddimi01  625000
##  7   1985 BAL    AL    stewasa01  581250
##  8   1985 BAL    AL    martide01  560000
##  9   1985 BAL    AL    roeniga01  558333
## 10   1985 BAL    AL    mcgresc01  547143
## # i 23,946 more rows
```

```
head(batting)
```

```
## # A tibble: 6 x 28
##   playerID  yearID stint teamID lgID      G G_batting    AB     R     H `2B`
##   <chr>      <dbl> <dbl> <chr>  <chr> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 aardsda01   2004     1 SFN    NL       11        11     0     0     0     0
## 2 aardsda01   2006     1 CHN    NL       45        43     2     0     0     0
## 3 aardsda01   2007     1 CHA    AL       25         2     0     0     0     0
## 4 aardsda01   2008     1 BOS    AL       47          5     1     0     0     0
## 5 aardsda01   2009     1 SEA    AL       73          3     0     0     0     0
## 6 aardsda01   2010     1 SEA    AL       53          4     0     0     0     0
## # i 17 more variables: '3B' <dbl>, HR <dbl>, RBI <dbl>, SB <dbl>, CS <dbl>,
## #   BB <dbl>, SO <dbl>, IBB <dbl>, HBP <dbl>, SH <dbl>, SF <dbl>, GIDP <dbl>,
## #   G_old <dbl>, BA <dbl>, OBP <dbl>, '1B' <dbl>, SLG <dbl>
```

```
arrange(batting, yearID)
```

```
## # A tibble: 97,889 x 28
##    playerID  yearID stint teamID lgID      G G_batting    AB     R     H  '2B'
##    <chr>      <dbl> <dbl> <chr>  <chr> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 abercda01   1871     1 TRO    <NA>      1         1     4     0     0     0
##  2 addybo01    1871     1 RC1    <NA>     25        25   118    30    32     6
##  3 allisar01   1871     1 CL1    <NA>     29        29   137    28    40     4
##  4 allisdo01   1871     1 WS3    <NA>     27        27   133    28    44    10
##  5 ansonca01   1871     1 RC1    <NA>     25        25   120    29    39    11
##  6 armstbo01   1871     1 FW1    <NA>     12        12    49     9    11     2
##  7 barkeal01   1871     1 RC1    <NA>      1         1     4     0     1     0
##  8 barnero01   1871     1 BS1    <NA>     31        31   157    66    63    10
##  9 barrebi01   1871     1 FW1    <NA>      1         1     5     1     1     1
## 10 barrofr01   1871     1 BS1    <NA>     18        18    86    13    13     2
## # i 97,879 more rows
## # i 17 more variables: '3B' <dbl>, HR <dbl>, RBI <dbl>, SB <dbl>, CS <dbl>,
## #   BB <dbl>, SO <dbl>, IBB <dbl>, HBP <dbl>, SH <dbl>, SF <dbl>, GIDP <dbl>,
## #   G_old <dbl>, BA <dbl>, OBP <dbl>, '1B' <dbl>, SLG <dbl>
```

NOTICE: Salaries dataset start at 1985, but batting dataset goes back to 1871 We need to remove the rows with yearID prior to 1985 from batting

```
#Method 1
batting <- subset(batting, yearID >= 1985)

#Method 2 :batting <- filter(batting, yearID >= 1985)
arrange(batting, yearID)
```

```
## # A tibble: 35,652 x 28
##    playerID  yearID stint teamID lgID      G G_batting    AB     R     H  '2B'
##    <chr>      <dbl> <dbl> <chr>  <chr> <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 aasedo01    1985     1 BAL    AL       54         0    NA    NA    NA    NA
##  2 abregjo01   1985     1 CHN    NL        6         6     9     0     0     0
##  3 ackerji01   1985     1 TOR    AL       61         0    NA    NA    NA    NA
##  4 adamsri02   1985     1 SFN    NL       54        54   121    12    23     3
##  5 agostju01   1985     1 CHA    AL       54         4     0     0     0     0
##  6 aguaylu01   1985     1 PHI    NL       91        91   165    27    46     7
##  7 aguilri01   1985     1 NYN    NL       22        22    36     1    10     2
##  8 aikenwi01   1985     1 TOR    AL       12        12    20     2     4     1
##  9 alexado01   1985     1 TOR    AL       36         0    NA    NA    NA    NA
## 10 allenga01   1985     1 TOR    AL       14        14    34     2     4     1
## # i 35,642 more rows
## # i 17 more variables: '3B' <dbl>, HR <dbl>, RBI <dbl>, SB <dbl>, CS <dbl>,
## #   BB <dbl>, SO <dbl>, IBB <dbl>, HBP <dbl>, SH <dbl>, SF <dbl>, GIDP <dbl>,
## #   G_old <dbl>, BA <dbl>, OBP <dbl>, '1B' <dbl>, SLG <dbl>
```

```r
#Merge
combo <- merge(batting, salary, by = c('playerID','yearID'))
head(combo)
```

```
##    playerID yearID stint teamID.x lgID.x  G G_batting AB  R  H 2B 3B HR RBI SB
## 1 aardsda01   2004     1      SFN     NL 11        11  0  0  0  0  0  0   0  0
## 2 aardsda01   2007     1      CHA     AL 25         2  0  0  0  0  0  0   0  0
## 3 aardsda01   2008     1      BOS     AL 47         5  1  0  0  0  0  0   0  0
## 4 aardsda01   2009     1      SEA     AL 73         3  0  0  0  0  0  0   0  0
## 5 aardsda01   2010     1      SEA     AL 53         4  0  0  0  0  0  0   0  0
## 6 aardsda01   2012     1      NYA     AL  1        NA NA NA NA NA NA NA  NA NA
##   CS BB SO IBB HBP SH SF GIDP G_old  BA OBP 1B SLG teamID.y lgID.y  salary
## 1  0  0  0   0   0  0  0    0    11 NaN NaN  0 NaN      SFN     NL  300000
## 2  0  0  0   0   0  0  0    0     2 NaN NaN  0 NaN      CHA     AL  387500
## 3  0  0  1   0   0  0  0    0     5   0   0  0   0      BOS     AL  403250
## 4  0  0  0   0   0  0  0    0    NA NaN NaN  0 NaN      SEA     AL  419000
## 5  0  0  0   0   0  0  0    0    NA NaN NaN  0 NaN      SEA     AL 2750000
## 6 NA NA NA  NA  NA NA NA   NA    NA  NA  NA NA  NA      NYA     AL  500000
```

```r
summary(combo)
```

```
##    playerID             yearID         stint         teamID.x
##  Length:25397       Min.   :1985   Min.   :1.000   Length:25397
##  Class :character   1st Qu.:1993   1st Qu.:1.000   Class :character
##  Mode  :character   Median :1999   Median :1.000   Mode  :character
##                     Mean   :1999   Mean   :1.098
##                     3rd Qu.:2006   3rd Qu.:1.000
##                     Max.   :2013   Max.   :4.000
##
##     lgID.x               G           G_batting            AB
##  Length:25397       Min.   :  1.00   Min.   :  0.00   Min.   :  0.0
##  Class :character   1st Qu.: 26.00   1st Qu.:  8.00   1st Qu.:  5.0
##  Mode  :character   Median : 50.00   Median : 42.00   Median : 85.0
##                     Mean   : 64.06   Mean   : 57.58   Mean   :182.4
##                     3rd Qu.:101.00   3rd Qu.:101.00   3rd Qu.:336.0
##                     Max.   :163.00   Max.   :163.00   Max.   :716.0
##                                      NA’s   :906      NA’s   :2661
##        R                H                 2B               3B
##  Min.   :  0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:  0.00   1st Qu.:  1.00   1st Qu.: 0.000   1st Qu.: 0.000
##  Median :  9.00   Median : 19.00   Median : 3.000   Median : 0.000
##  Mean   : 24.71   Mean   : 48.18   Mean   : 9.276   Mean   : 1.033
##  3rd Qu.: 43.00   3rd Qu.: 87.25   3rd Qu.:16.000   3rd Qu.: 1.000
##  Max.   :152.00   Max.   :262.00   Max.   :59.000   Max.   :23.000
##  NA’s   :2661     NA’s   :2661     NA’s   :2661     NA’s   :2661
##         HR             RBI               SB                CS
##  Min.   : 0.000   Min.   :  0.00   Min.   :  0.000   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.: 0.00
##  Median : 1.000   Median :  8.00   Median :  0.000   Median : 0.00
##  Mean   : 5.369   Mean   : 23.56   Mean   :  3.568   Mean   : 1.54
##  3rd Qu.: 7.000   3rd Qu.: 39.00   3rd Qu.:  3.000   3rd Qu.: 2.00
##  Max.   :73.000   Max.   :165.00   Max.   :110.000   Max.   :29.00
##  NA’s   :2661     NA’s   :2661     NA’s   :2661      NA’s   :2661
##        BB               SO              IBB              HBP
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.000   Min.   : 0.000
##  1st Qu.:  0.00   1st Qu.:  2.00   1st Qu.:  0.000   1st Qu.: 0.000
##  Median :  6.00   Median : 20.00   Median :  0.000   Median : 0.000
##  Mean   : 17.98   Mean   : 33.52   Mean   :  1.533   Mean   : 1.614
##  3rd Qu.: 29.00   3rd Qu.: 55.00   3rd Qu.:  2.000   3rd Qu.: 2.000
##  Max.   :232.00   Max.   :223.00   Max.   :120.000   Max.   :35.000
##  NA's   :2661     NA's   :2661     NA's   :2662      NA's   :2670
##        SH               SF              GIDP            G_old
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :  0.00
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 20.00
##  Median : 0.000   Median : 0.000   Median : 2.000   Median : 47.00
##  Mean   : 1.786   Mean   : 1.554   Mean   : 4.127   Mean   : 61.43
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 7.000   3rd Qu.:101.00
##  Max.   :39.000   Max.   :17.000   Max.   :35.000   Max.   :163.00
##  NA's   :2661     NA's   :2662     NA's   :2661     NA's   :3414
##        BA             OBP              1B             SLG
##  Min.   :0.000   Min.   :0.000   Min.   :  0.0   Min.   :0.000
##  1st Qu.:0.160   1st Qu.:0.208   1st Qu.:  0.0   1st Qu.:0.200
##  Median :0.242   Median :0.305   Median : 13.0   Median :0.351
##  Mean   :0.212   Mean   :0.270   Mean   : 32.5   Mean   :0.317
##  3rd Qu.:0.276   3rd Qu.:0.346   3rd Qu.: 59.0   3rd Qu.:0.432
##  Max.   :1.000   Max.   :1.000   Max.   :225.0   Max.   :4.000
##  NA's   :5618    NA's   :5562    NA's   :2661    NA's   :5618
##    teamID.y            lgID.y              salary
##  Length:25397       Length:25397       Min.   :       0
##  Class :character   Class :character   1st Qu.:  255000
##  Mode  :character   Mode  :character   Median :  550000
##                                        Mean   : 1879256
##                                        3rd Qu.: 2150000
##                                        Max.   :33000000
##
```

4.Extract lost players As previously mentioned, the Oakland A's lost 3 key players during the off-season. We'll want to get their stats to see what we have to replace the players lost were: first baseman 2000 AL MVP Jason Giambi (giambja01) to the New York Yankees, outfielder Johnny Damon (damonjo01) to the Boston Red Sox and infielder Rainer Gustavo "Ray" Olmedo ('saenzol01').

```
lost_players <- subset(combo, playerID %in% c('giambja01','damonjo01', 'saenzol01'))
head(lost_players)
```

```
##       playerID yearID stint teamID.x lgID.x   G G_batting  AB   R   H 2B 3B HR
## 5135 damonjo01   1995     1      KCA     AL  47        47 188  32  53 11  5  3
## 5136 damonjo01   1996     1      KCA     AL 145       145 517  61 140 22  5  6
## 5137 damonjo01   1997     1      KCA     AL 146       146 472  70 130 12  8  8
## 5138 damonjo01   1998     1      KCA     AL 161       161 642 104 178 30 10 18
## 5139 damonjo01   1999     1      KCA     AL 145       145 583 101 179 39  9 14
## 5140 damonjo01   2000     1      KCA     AL 159       159 655 136 214 42 10 16
##      RBI SB CS BB SO IBB HBP SH SF GIDP G_old        BA       OBP  1B       SLG
## 5135  23  7  0 12 22   0   1  2  3    2    47 0.2819149 0.3235294  34 0.4414894
## 5136  50 25  5 31 64   3   3 10  5    4   145 0.2707930 0.3129496 107 0.3675048
## 5137  48 16 10 42 70   2   3  6  1    3   146 0.2754237 0.3378378 102 0.3855932
## 5138  66 26 12 58 84   4   4  3  3    4   161 0.2772586 0.3394625 120 0.4392523
## 5139  77 36  6 67 50   5   3  3  4   13   145 0.3070326 0.3789954 117 0.4768439
## 5140  88 46  9 65 60   4   1  8 12    7   159 0.3267176 0.3819918 146 0.4946565
##      teamID.y lgID.y  salary
## 5135      KCA     AL  109000
## 5136      KCA     AL  180000
## 5137      KCA     AL  240000
## 5138      KCA     AL  460000
## 5139      KCA     AL 2100000
## 5140      KCA     AL 4000000
```

```
#Since all these players were lost in after 2001 in the offseason,

lost_players <- subset(lost_players,yearID == 2001)
lost_players <- lost_players[,c('playerID','H','2B','3B','HR','OBP','SLG','BA','AB')]
head(lost_players)
```

```
##        playerID   H 2B 3B HR       OBP       SLG        BA  AB
## 5141  damonjo01 165 34  4  9 0.3235294 0.3633540 0.2562112 644
## 7878  giambja01 178 47  2 38 0.4769001 0.6596154 0.3423077 520
## 20114 saenzol01  67 21  1  9 0.2911765 0.3836066 0.2196721 305
```

```
summary(lost_players)
```

```
##   playerID               H               2B              3B
##  Length:3           Min.   : 67.0   Min.   :21.0   Min.   :1.000
##  Class :character   1st Qu.:116.0   1st Qu.:27.5   1st Qu.:1.500
##  Mode  :character   Median :165.0   Median :34.0   Median :2.000
##                     Mean   :136.7   Mean   :34.0   Mean   :2.333
##                     3rd Qu.:171.5   3rd Qu.:40.5   3rd Qu.:3.000
##                     Max.   :178.0   Max.   :47.0   Max.   :4.000
##        HR             OBP              SLG               BA
##  Min.   : 9.00   Min.   :0.2912   Min.   :0.3634   Min.   :0.2197
##  1st Qu.: 9.00   1st Qu.:0.3074   1st Qu.:0.3735   1st Qu.:0.2379
##  Median : 9.00   Median :0.3235   Median :0.3836   Median :0.2562
##  Mean   :18.67   Mean   :0.3639   Mean   :0.4689   Mean   :0.2727
##  3rd Qu.:23.50   3rd Qu.:0.4002   3rd Qu.:0.5216   3rd Qu.:0.2993
##  Max.   :38.00   Max.   :0.4769   Max.   :0.6596   Max.   :0.3423
##        AB
##  Min.   :305.0
##  1st Qu.:412.5
##  Median :520.0
##  Mean   :489.7
##  3rd Qu.:582.0
##  Max.   :644.0
```

####5. Find Replacement Players for the key three players we lost.

constraints:

(1). The total combined salary of the three players can not exceed 15 million dollars. (2). Their combined number of At Bats (AB) needs to be equal to or greater than the lost players. (3). Their mean OBP had to equal to or greater than the mean OBP of the lost players

```r
lost_players <- subset(combo, playerID %in% c('giambja01','damonjo01', 'saenzol01'))
head(lost_players)
```

```
##       playerID yearID stint teamID.x lgID.x   G G_batting  AB   R   H 2B 3B HR
## 5135 damonjo01   1995     1      KCA     AL  47        47 188  32  53 11  5  3
## 5136 damonjo01   1996     1      KCA     AL 145       145 517  61 140 22  5  6
## 5137 damonjo01   1997     1      KCA     AL 146       146 472  70 130 12  8  8
## 5138 damonjo01   1998     1      KCA     AL 161       161 642 104 178 30 10 18
## 5139 damonjo01   1999     1      KCA     AL 145       145 583 101 179 39  9 14
## 5140 damonjo01   2000     1      KCA     AL 159       159 655 136 214 42 10 16
##      RBI SB CS BB SO IBB HBP SH SF GIDP G_old        BA       OBP  1B       SLG
## 5135  23  7  0 12 22   0   1  2  3    2    47 0.2819149 0.3235294  34 0.4414894
## 5136  50 25  5 31 64   3   3 10  5    4   145 0.2707930 0.3129496 107 0.3675048
## 5137  48 16 10 42 70   2   3  6  1    3   146 0.2754237 0.3378378 102 0.3855932
## 5138  66 26 12 58 84   4   4  3  3    4   161 0.2772586 0.3394625 120 0.4392523
## 5139  77 36  6 67 50   5   3  3  4   13   145 0.3070326 0.3789954 117 0.4768439
## 5140  88 46  9 65 60   4   1  8 12    7   159 0.3267176 0.3819918 146 0.4946565
##      teamID.y lgID.y  salary
## 5135      KCA     AL  109000
## 5136      KCA     AL  180000
## 5137      KCA     AL  240000
## 5138      KCA     AL  460000
## 5139      KCA     AL 2100000
## 5140      KCA     AL 4000000
```

```r
#Since all these players were lost in after 2001 in the offseason,

lost_players <- subset(lost_players,yearID == 2001)
lost_players <- lost_players[,c('playerID','H','2B','3B','HR','OBP','SLG','BA','AB')]
head(lost_players)
```

```
##        playerID   H 2B 3B HR       OBP       SLG        BA  AB
## 5141  damonjo01 165 34  4  9 0.3235294 0.3633540 0.2562112 644
## 7878  giambja01 178 47  2 38 0.4769001 0.6596154 0.3423077 520
## 20114 saenzol01  67 21  1  9 0.2911765 0.3836066 0.2196721 305
```

```r
summary(lost_players)
```

```
##    playerID               H               2B              3B       
##  Length:3           Min.   : 67.0   Min.   :21.0   Min.   :1.000  
##  Class :character   1st Qu.:116.0   1st Qu.:27.5   1st Qu.:1.500  
##  Mode  :character   Median :165.0   Median :34.0   Median :2.000  
##                     Mean   :136.7   Mean   :34.0   Mean   :2.333  
##                     3rd Qu.:171.5   3rd Qu.:40.5   3rd Qu.:3.000  
##                     Max.   :178.0   Max.   :47.0   Max.   :4.000  
##        HR             OBP              SLG               BA        
##  Min.   : 9.00   Min.   :0.2912   Min.   :0.3634   Min.   :0.2197  
##  1st Qu.: 9.00   1st Qu.:0.3074   1st Qu.:0.3735   1st Qu.:0.2379  
##  Median : 9.00   Median :0.3235   Median :0.3836   Median :0.2562  
##  Mean   :18.67   Mean   :0.3639   Mean   :0.4689   Mean   :0.2727  
##  3rd Qu.:23.50   3rd Qu.:0.4002   3rd Qu.:0.5216   3rd Qu.:0.2993  
##  Max.   :38.00   Max.   :0.4769   Max.   :0.6596   Max.   :0.3423  
##         AB
##  Min.   :305.0
##  1st Qu.:412.5
##  Median :520.0
##  Mean   :489.7
##  3rd Qu.:582.0
##  Max.   :644.0
```

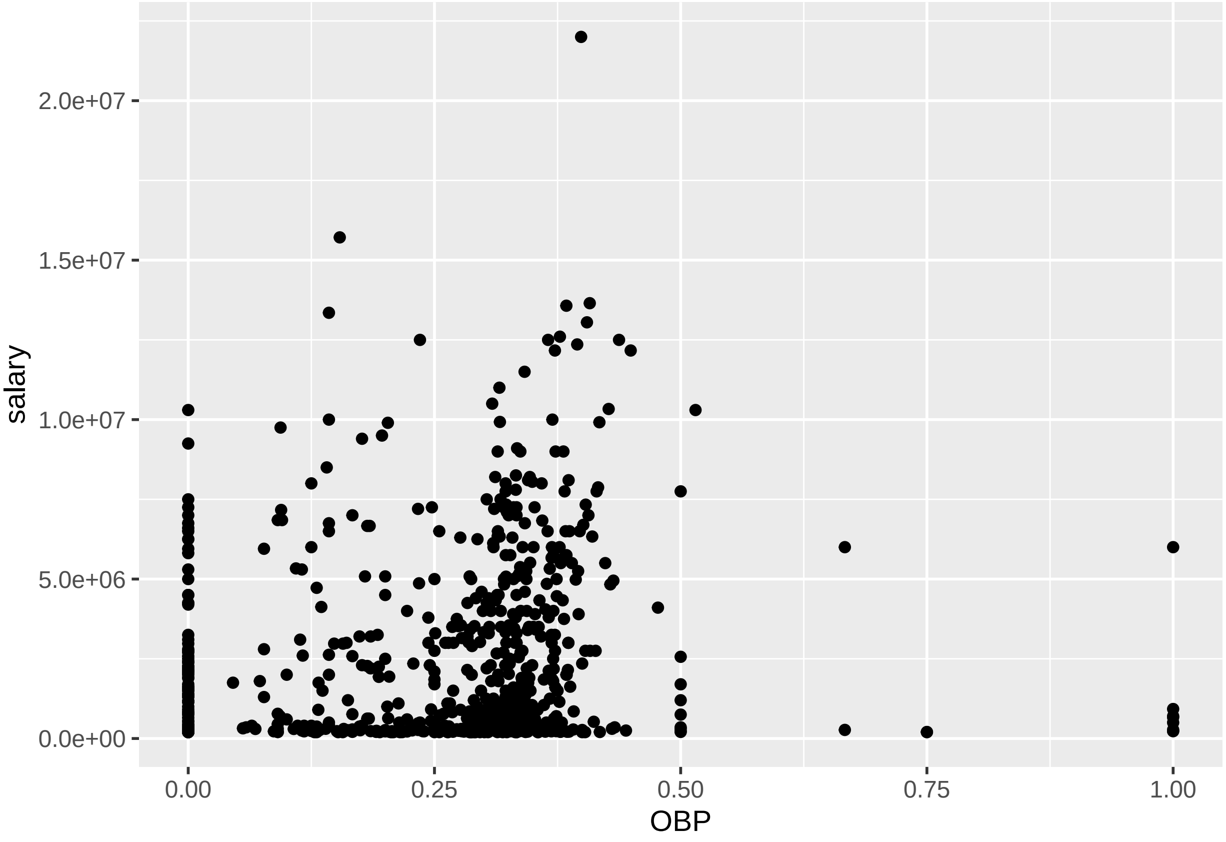
```
#grab available players after 2001
avail.players <- filter(combo,yearID==2001)

#The mean OBP of lost player is 0.3639
summary(lost_players$OBP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2912  0.3074  0.3235  0.3639  0.4002  0.4769
```

```
library(ggplot2)
ggplot(avail.players,aes(x=OBP,y=salary)) + geom_point()
```

```
## Warning: Removed 168 rows containing missing values (`geom_point()`).
```



```
# No one has salary 3 million, thus we could pick any of 3 players
```

```
avail.players <- filter(avail.players,salary<8000000,OBP>0.3639)

# The sum of AB of 3 lost players is 644+520+305 = 1469, thus, each of
#the AB of replace player should not less than 500

avail.players <- filter(avail.players,AB >= 500)

possible <- head(arrange(avail.players,desc(OBP)),10)
possible <- possible[,c('playerID','OBP','AB','salary')]
head(possible)
```

```
##    playerID       OBP  AB  salary
## 1 giambja01 0.4769001 520 4103333
## 2 heltoto01 0.4316547 587 4950000
## 3 berkmla01 0.4302326 577  305000
## 4 gonzalu01 0.4285714 609 4833333
## 5 thomeji01 0.4161491 526 7875000
## 6 alomaro01 0.4146707 575 7750000
```