MoneyBall Project:

The 2002 Oakland A's

The Oakland Athletics' 2002 season was the team's 35th in Oakland, California. It was also the 102nd season in franchise history. The Athletics finished first in the American League West with a record of 103-59.

The Athletics' 2002 campaign ranks among the most famous in franchise history. Following the 2001 season, Oakland saw the departure of three key players (the lost boys). Billy Beane, the team's general manager, responded with a series of under-the-radar free agent signings. The new-look Athletics, despite a comparative lack of star power, surprised the baseball world by besting the 2001 team's regular season record. The team is most famous, however, for winning 20 consecutive games between August 13 and September 4, 2002. The Athletics' season was the subject of Michael Lewis' 2003 book Moneyball: The Art of Winning an Unfair Game (as Lewis was given the opportunity to follow the team around throughout that season)

This project is based off the book written by Michael Lewis (later turned into a movie).

Moneyball Book

The central premise of book Moneyball is that the collective wisdom of baseball insiders (including players, managers, coaches, scouts, and the front office) over the past century is subjective and often flawed. Statistics such as stolen bases, runs batted in, and batting average, typically used to gauge players, are relics of a 19th-century view of the game and the statistics available at that time. The book argues that the Oakland A's' front office took advantage of more analytical gauges of player performance to field a team that could better compete against richer competitors in Major League Baseball (MLB).

Rigorous statistical analysis had demonstrated that on-base percentage and slugging percentage are better indicators of offensive success, and the A's became convinced that these qualities were cheaper to obtain on the open market than more historically valued qualities such as speed and contact. These observations often flew in the face of conventional baseball wisdom and the beliefs of many baseball scouts and executives.

By re-evaluating the strategies that produce wins on the field, the 2002 Athletics, with approximately US 44 million dollars in salary, were competitive with larger market teams such as the New York Yankees, who spent over US$125 million in payroll that same season. Because of the team's smaller revenues, Oakland is forced to find players undervalued by the market, and their system for finding value in undervalued players has proven itself thus far. This approach brought the A's to the playoffs in 2002 and 2003.

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!

Let's get started!

Required Libraries

In [1]:

```
library("ggplot2")
library("dplyr")
library("tidyr")
library("Hmisc")
library("plotly")
```

```
Warning message:
"package 'dplyr' was built under R version 3.6.3"

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


Warning message:
"package 'tidyr' was built under R version 3.6.3"
Warning message:
"package 'Hmisc' was built under R version 3.6.3"
Loading required package: lattice

Loading required package: survival

Warning message:
"package 'survival' was built under R version 3.6.3"
Loading required package: Formula

Warning message:
"package 'Formula' was built under R version 3.6.3"

Attaching package: 'Hmisc'


The following objects are masked from 'package:dplyr':

    src, summarize


The following objects are masked from 'package:base':

    format.pval, units



Attaching package: 'plotly'


The following object is masked from 'package:Hmisc':

    subplot
```

```
The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

Data

We'll be using data from Sean Lahaman's Website a very useful source for baseball statistics. The documentation for the csv files is located in the readme2013.txt file. You may need to reference this to understand what acronyms stand for.

Use R to open the Batting.csv file and assign it to a dataframe called batting using read.csv

Use R to open the Batting.csv file and assign it to a dataframe called batting using read.csv

In [2]:

```
bats=read.csv("Batting.csv")
```

Use head() to check out the batting

In [3]:

```
head(bats)
```

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	...	SB	CS	BB	SO	IBB	H
aardsda01	2004	1	SFN	NL	11	11	0	0	0	...	0	0	0	0	0	
aardsda01	2006	1	CHN	NL	45	43	2	0	0	...	0	0	0	0	0	
aardsda01	2007	1	CHA	AL	25	2	0	0	0	...	0	0	0	0	0	
aardsda01	2008	1	BOS	AL	47	5	1	0	0	...	0	0	0	1	0	
aardsda01	2009	1	SEA	AL	73	3	0	0	0	...	0	0	0	0	0	
aardsda01	2010	1	SEA	AL	53	4	0	0	0	...	0	0	0	0	0	

Use colnames() to check out column

In [4]:

```
colnames(bats)
```

'playerID' 'yearID' 'stint' 'teamID' 'lgID' 'G' 'G_batting' 'AB' 'R' 'H' 'X2B' 'X3B'
'HR' 'RBI' 'SB' 'CS' 'BB' 'SO' 'IBB' 'HBP' 'SH' 'SF' 'GIDP' 'G_old'

Use str() to check the structure.

In [5]:

```
str(bats)
```

```
'data.frame':   97889 obs. of  24 variables:
 $ playerID : Factor w/ 18107 levels "aardsda01","aaronha01",..: 1 1 1 1 1 1
1 2 2 2 ...
 $ yearID   : int  2004 2006 2007 2008 2009 2010 2012 1954 1955 1956 ...
 $ stint    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ teamID   : Factor w/ 149 levels "ALT","ANA","ARI",..: 117 35 33 16 116 11
6 93 80 80 80 ...
 $ lgID     : Factor w/ 6 levels "AA","AL","FL",..: 4 4 2 2 2 2 2 4 4 4 ...
 $ G        : int  11 45 25 47 73 53 1 122 153 153 ...
 $ G_batting: int  11 43 2 5 3 4 NA 122 153 153 ...
 $ AB       : int  0 2 0 1 0 0 NA 468 602 609 ...
 $ R        : int  0 0 0 0 0 0 NA 58 105 106 ...
 $ H        : int  0 0 0 0 0 0 NA 131 189 200 ...
 $ X2B      : int  0 0 0 0 0 0 NA 27 37 34 ...
 $ X3B      : int  0 0 0 0 0 0 NA 6 9 14 ...
 $ HR       : int  0 0 0 0 0 0 NA 13 27 26 ...
 $ RBI      : int  0 0 0 0 0 0 NA 69 106 92 ...
 $ SB       : int  0 0 0 0 0 0 NA 2 3 2 ...
 $ CS       : int  0 0 0 0 0 0 NA 2 1 4 ...
 $ BB       : int  0 0 0 0 0 0 NA 28 49 37 ...
 $ SO       : int  0 0 0 1 0 0 NA 39 61 54 ...
 $ IBB      : int  0 0 0 0 0 0 NA NA 5 6 ...
 $ HBP      : int  0 0 0 0 0 0 NA 3 3 2 ...
 $ SH       : int  0 1 0 0 0 0 NA 6 7 5 ...
 $ SF       : int  0 0 0 0 0 0 NA 4 4 7 ...
 $ GIDP     : int  0 0 0 0 0 0 NA 13 20 21 ...
 $ G_old    : int  11 45 2 5 NA NA NA 122 153 153 ...
```

Make sure you understand how to call the columns by using the $ symbol.

Call the head() of the first sx rows of AB (At Bats) column

In [6]:

```
head(bats$AB)
```

0 2 0 1 0 0

Call the head of the doubles (X2B) column

In [7]:

```
head(bats$X2B)
```

0 0 0 0 0 0

Use Summary() function to displaying summary statistics of bats data:

In [8]:

```
summary(bats)
```

```
     playerID         yearID         stint           teamID        lgID       
 mcguide01:   31   Min.   :1871   Min.   :1.000   CHN    : 4720   AA  : 1890  
 henderi01:   29   1st Qu.:1931   1st Qu.:1.000   PHI    : 4621   AL  :44369  
 newsobo01:   29   Median :1970   Median :1.000   PIT    : 4575   FL  :  470  
 johnto01 :   28   Mean   :1962   Mean   :1.077   SLN    : 4535   NL  :49944  
 kaatji01 :   28   3rd Qu.:1995   3rd Qu.:1.000   CIN    : 4393   PL  :  147  
 ansonca01:   27   Max.   :2013   Max.   :5.000   CLE    : 4318   UA  :  332  
 (Other)  :97717                                  (Other):70727   NA's:  737  
       G            G_batting           AB              R         
 Min.   :  1.00   Min.   :  0.00   Min.   :  0.0   Min.   :  0.00  
 1st Qu.: 13.00   1st Qu.:  7.00   1st Qu.:  9.0   1st Qu.:  0.00  
 Median : 35.00   Median : 32.00   Median : 61.0   Median :  5.00  
 Mean   : 51.65   Mean   : 49.13   Mean   :154.1   Mean   : 20.47  
 3rd Qu.: 81.00   3rd Qu.: 81.00   3rd Qu.:260.0   3rd Qu.: 31.00  
 Max.   :165.00   Max.   :165.00   Max.   :716.0   Max.   :192.00  
                  NA's   :1406     NA's   :6413    NA's   :6413    
       H              X2B             X3B              HR        
 Min.   :  0.00   Min.   : 0.0   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:  1.00   1st Qu.: 0.0   1st Qu.: 0.000   1st Qu.: 0.000  
 Median : 12.00   Median : 2.0   Median : 0.000   Median : 0.000  
 Mean   : 40.37   Mean   : 6.8   Mean   : 1.424   Mean   : 3.002  
 3rd Qu.: 66.00   3rd Qu.:10.0   3rd Qu.: 2.000   3rd Qu.: 3.000  
 Max.   :262.00   Max.   :67.0   Max.   :36.000   Max.   :73.000  
 NA's   :6413     NA's   :6413   NA's   :6413     NA's   :6413    
      RBI               SB               CS               BB        
 Min.   :  0.00   Min.   :  0.000   Min.   : 0.000   Min.   :  0.00  
 1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.: 0.000   1st Qu.:  0.00  
 Median :  5.00   Median :  0.000   Median : 0.000   Median :  4.00  
 Mean   : 18.47   Mean   :  3.265   Mean   : 1.385   Mean   : 14.21  
 3rd Qu.: 28.00   3rd Qu.:  2.000   3rd Qu.: 1.000   3rd Qu.: 21.00  
 Max.   :191.00   Max.   :138.000   Max.   :42.000   Max.   :232.00  
 NA's   :6837     NA's   :7713      NA's   :29867    NA's   :6413    
       SO              IBB              HBP              SH        
 Min.   :  0.00   Min.   :  0.00   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:  2.00   1st Qu.:  0.00   1st Qu.: 0.000   1st Qu.: 0.000  
 Median : 11.00   Median :  0.00   Median : 0.000   Median : 1.000  
 Mean   : 21.95   Mean   :  1.28   Mean   : 1.136   Mean   : 2.564  
 3rd Qu.: 31.00   3rd Qu.:  1.00   3rd Qu.: 1.000   3rd Qu.: 3.000  
 Max.   :223.00   Max.   :120.00   Max.   :51.000   Max.   :67.000  
 NA's   :14251    NA's   :42977    NA's   :9233     NA's   :12751   
       SF              GIDP            G_old       
 Min.   : 0.0    Min.   : 0.00   Min.   :  0.00  
 1st Qu.: 0.0    1st Qu.: 0.00   1st Qu.: 11.00  
 Median : 0.0    Median : 1.00   Median : 34.00  
 Mean   : 1.2    Mean   : 3.33   Mean   : 50.99  
 3rd Qu.: 2.0    3rd Qu.: 5.00   3rd Qu.: 82.00  
 Max.   :19.0    Max.   :36.00   Max.   :165.00  
 NA's   :42446   NA's   :32521   NA's   :5189    
```

Feature Engg.

We need to add three more statistics that were used in Moneyball! These are:

1.Batting Average

2.On Base Percentage

3.Slugging Percentage

Click on the links provided and search the wikipedia page for the formula for creating the new statistic! For example, for Batting Average, you'll need to scroll down until you see:

AVG=H/AB

Which means that the Batting Average is equal to H (Hits) divided by AB (At Base). So we'll do the following to create a new column called BA and add it to our data frame:

Matrics

Batting Average, On Base Percentage, and Sluggish percentage.

In [9]:

```
# Batting Average ---> BA

bats$BA<-bats$H/bats$AB
```

After doing this operation, check the first 6 entries of the BA column of your data frame and it should look like this:

In [10]:

```
head(bats$BA)
```

NaN 0 NaN 0 NaN NaN

After doing this operation, check the last 6 entries of the BA column of your data frame and it should look like this:

In [11]:

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

0.123076923076923 0.274647887323944 0.147058823529412 0.274509803921569
0.213872832369942

Now do the same for some new columns! On Base Percentage (OBP) and Slugging Percentage (SLG). Hint: For SLG, you need 1B (Singles), this isn't in your data frame. However you can calculate it by subtracting doubles,triples, and home runs from total hits (H): 1B = H-2B-3B-HR

.Create an OBP Column

.Create an SLG Column

In [12]:

```
#  On Base Percentage-----OBP

bats$OBP<-(bats$H + bats$BB+bats$HBP)/(bats$AB+bats$BB+bats$HBP+bats$SF)
head(bats$OBP)
```

NaN 0 NaN 0 NaN NaN

In [13]:

```
# Sluggish percent

# creating X1B (single)
bats$X1B<-bats$H-bats$X2B-bats$X3B-bats$HR

#creating  Sluggish Average
bats$SLG<-((1*bats$X1B)+(2*bats$X2B)+(3*bats$X3B)+(4*bats$HR))/bats$AB
```

In [14]:

```
head(bats,10)
```

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	...	IBB	HBP	SH	SF
aardsda01	2004	1	SFN	NL	11	11	0	0	0	...	0	0	0	0
aardsda01	2006	1	CHN	NL	45	43	2	0	0	...	0	0	1	0
aardsda01	2007	1	CHA	AL	25	2	0	0	0	...	0	0	0	0
aardsda01	2008	1	BOS	AL	47	5	1	0	0	...	0	0	0	0
aardsda01	2009	1	SEA	AL	73	3	0	0	0	...	0	0	0	0
aardsda01	2010	1	SEA	AL	53	4	0	0	0	...	0	0	0	0
aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA	...	NA	NA	NA	NA
aaronha01	1954	1	ML1	NL	122	122	468	58	131	...	NA	3	6	4
aaronha01	1955	1	ML1	NL	153	153	602	105	189	...	5	3	7	4
aaronha01	1956	1	ML1	NL	153	153	609	106	200	...	6	2	5	7

Merging Salary Data with Batting Data

We know we don't just want the best players, we want the most undervalued players, meaning we will also need to know current salary information! We have salary information in the csv file 'Salaries.csv'.

Load the Salaries.csv file into a dataframe called sal using read.csv

In [15]:

```
Sal=read.csv("Salaries.csv")
```

Displaying Salary:

In [16]:

```
head(Sal,10)
```

yearID	teamID	lgID	playerID	salary
1985	BAL	AL	murraed02	1472819
1985	BAL	AL	lynnfr01	1090000
1985	BAL	AL	ripkeca01	800000
1985	BAL	AL	lacyle01	725000
1985	BAL	AL	flanami01	641667
1985	BAL	AL	boddimi01	625000
1985	BAL	AL	stewasa01	581250
1985	BAL	AL	martide01	560000
1985	BAL	AL	roeniga01	558333
1985	BAL	AL	mcgresc01	547143

Use Summary() function to displaying summary statistics of salary data:

In [17]:

```
summary(Sal)
```

```
     yearID          teamID      lgID           playerID         salary
 Min.   :1985   CLE    :  867   AL:11744   moyerja01:   25   Min.   :
0
 1st Qu.:1993   LAN    :  861   NL:12212   vizquom01:   24   1st Qu.:  25000
0
 Median :1999   PHI    :  861              glavito02:   23   Median :  50795
0
 Mean   :1999   SLN    :  858              bondsba01:   22   Mean   : 186435
7
 3rd Qu.:2006   BAL    :  855              griffke02:   22   3rd Qu.: 210000
0
 Max.   :2013   NYA    :  855              thomeji01:   22   Max.   :3300000
0
                (Other):18799              (Other)  :23818
```

Available players since 1985

In [18]:

```
avail_player=bats %>% filter(yearID>=1985)
```

In [19]:

```
head(avail_player,10)
```

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	...	IBB	HBP	SH	SF	GI
aardsda01	2004	1	SFN	NL	11	11	0	0	0	...	0	0	0	0	
aardsda01	2006	1	CHN	NL	45	43	2	0	0	...	0	0	1	0	
aardsda01	2007	1	CHA	AL	25	2	0	0	0	...	0	0	0	0	
aardsda01	2008	1	BOS	AL	47	5	1	0	0	...	0	0	0	0	
aardsda01	2009	1	SEA	AL	73	3	0	0	0	...	0	0	0	0	
aardsda01	2010	1	SEA	AL	53	4	0	0	0	...	0	0	0	0	
aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA	...	NA	NA	NA	NA	
aasedo01	1985	1	BAL	AL	54	0	NA	NA	NA	...	NA	NA	NA	NA	
aasedo01	1986	1	BAL	AL	66	0	NA	NA	NA	...	NA	NA	NA	NA	
aasedo01	1987	1	BAL	AL	7	0	NA	NA	NA	...	NA	NA	NA	NA	

In [20]:

```
# filter salary by yearID==2001
s1=Sal %>% filter(yearID==2001)
head(s1,10)
```

yearID	teamID	lgID	playerID	salary
2001	ANA	AL	vaughmo01	13166667
2001	ANA	AL	salmoti01	6500000
2001	ANA	AL	anderga01	4500000
2001	ANA	AL	erstada01	3450000
2001	ANA	AL	percitr01	3400000
2001	ANA	AL	valdeis01	2500000
2001	ANA	AL	rapppa01	2000000
2001	ANA	AL	hasegsh01	1500000
2001	ANA	AL	hillgl01	1500000
2001	ANA	AL	glaustr01	1250000

filter batting by yearID==2001

In [21]:

```
s2=bats %>% filter(yearID==2001)
head(s2,10)
```

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	...	IBB	HBP	SH	SF
abadan01	2001	1	OAK	AL	1	1	1	0	0	...	0	0	0	0
abbotje01	2001	1	FLO	NL	28	28	42	5	11	...	0	1	0	0
abbotku01	2001	1	ATL	NL	6	6	9	0	2	...	0	0	0	0
abbotpa01	2001	1	SEA	AL	28	2	4	0	1	...	0	0	1	0
abernbr01	2001	1	TBA	AL	79	79	304	43	82	...	1	0	3	1
abreubo01	2001	1	PHI	NL	162	162	588	118	170	...	11	1	0	9
acevejo01	2001	1	CIN	NL	18	18	34	1	4	...	0	0	2	0
aceveju01	2001	1	COL	NL	39	35	0	0	0	...	0	0	0	0
aceveju01	2001	2	FLO	NL	20	20	3	1	1	...	0	0	1	0
adamste01	2001	1	LAN	NL	43	41	39	2	2	...	0	0	5	1

Use the merge() function to merge the batting and sal data frames by c('playerID','yearID'). Call the new data frame s3

In [22]:

```
# merge both data
s3=merge(s1,s2,by=c('playerID', 'yearID'))
head(s3,10)
```

playerID	yearID	teamID.x	lgID.x	salary	stint	teamID.y	lgID.y	G	G_batting	...	IBB
abbotje01	2001	FLO	NL	300000	1	FLO	NL	28	28	...	0
abbotku01	2001	ATL	NL	600000	1	ATL	NL	6	6	...	0
abbotpa01	2001	SEA	AL	1700000	1	SEA	AL	28	2	...	0
abreubo01	2001	PHI	NL	4983000	1	PHI	NL	162	162	...	11
adamste01	2001	LAN	NL	2600000	1	LAN	NL	43	41	...	0
agbaybe01	2001	NYN	NL	260000	1	NYN	NL	91	91	...	0
alfonan01	2001	FLO	NL	2450000	1	FLO	NL	58	54	...	0
alfoned01	2001	NYN	NL	5750000	1	NYN	NL	124	124	...	0
alicelu01	2001	KCA	AL	800000	1	KCA	AL	113	113	...	0
allench01	2001	MIN	AL	240000	1	MIN	AL	57	57	...	1

Analyzing the 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').

Use the filter() function to get a data frame called lost_players from the combo data frame consisting of those 3 players. Hint: Try to figure out how to use %in% to avoid a bunch of or statements!

In [23]:

```
lost_player= bats %>% filter(playerID %in% c("giambja01","damonjo01","saenzol01"))
```

In [24]:

```
lost_player
```

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	...	IBB	HBP	SH	SF
damonjo01	1995	1	KCA	AL	47	47	188	32	53	...	0	1	2	3
damonjo01	1996	1	KCA	AL	145	145	517	61	140	...	3	3	10	5
damonjo01	1997	1	KCA	AL	146	146	472	70	130	...	2	3	6	1
damonjo01	1998	1	KCA	AL	161	161	642	104	178	...	4	4	3	3
damonjo01	1999	1	KCA	AL	145	145	583	101	179	...	5	3	3	4
damonjo01	2000	1	KCA	AL	159	159	655	136	214	...	4	1	8	12
damonjo01	2001	1	OAK	AL	155	155	644	108	165	...	1	5	5	4
damonjo01	2002	1	BOS	AL	154	154	623	118	178	...	5	6	3	5
damonjo01	2003	1	BOS	AL	145	145	608	103	166	...	4	2	6	6
damonjo01	2004	1	BOS	AL	150	150	621	123	189	...	1	2	0	3
damonjo01	2005	1	BOS	AL	148	148	624	117	197	...	3	2	0	9
damonjo01	2006	1	NYA	AL	149	149	593	115	169	...	1	4	2	5
damonjo01	2007	1	NYA	AL	141	141	533	93	144	...	1	2	1	3
damonjo01	2008	1	NYA	AL	143	143	555	95	168	...	0	1	2	1
damonjo01	2009	1	NYA	AL	143	143	550	107	155	...	1	2	2	1
damonjo01	2010	1	DET	AL	145	145	539	81	146	...	2	2	2	1
damonjo01	2011	1	TBA	AL	150	150	582	79	152	...	1	7	2	5
damonjo01	2012	1	CLE	AL	64	NA	207	25	46	...	0	0	0	0
giambja01	1995	1	OAK	AL	54	54	176	27	45	...	0	3	1	2
giambja01	1996	1	OAK	AL	140	140	536	84	156	...	3	5	1	5
giambja01	1997	1	OAK	AL	142	142	519	66	152	...	3	6	0	8
giambja01	1998	1	OAK	AL	153	153	562	92	166	...	7	5	0	9
giambja01	1999	1	OAK	AL	158	158	575	115	181	...	6	7	0	8
giambja01	2000	1	OAK	AL	152	152	510	108	170	...	6	9	0	8
giambja01	2001	1	OAK	AL	154	154	520	109	178	...	24	13	0	9
giambja01	2002	1	NYA	AL	155	155	560	120	176	...	4	15	0	5
giambja01	2003	1	NYA	AL	156	156	535	97	134	...	9	21	0	5
giambja01	2004	1	NYA	AL	80	80	264	33	55	...	1	8	0	3
giambja01	2005	1	NYA	AL	139	139	417	74	113	...	5	19	0	1
giambja01	2006	1	NYA	AL	139	139	446	92	113	...	12	16	0	7
giambja01	2007	1	NYA	AL	83	83	254	31	60	...	2	8	0	1
giambja01	2008	1	NYA	AL	145	145	458	68	113	...	5	22	0	9
giambja01	2009	1	OAK	AL	83	83	269	39	52	...	1	7	0	2
giambja01	2009	2	COL	NL	19	19	24	4	7	...	0	0	0	0
giambja01	2010	1	COL	NL	87	87	176	17	43	...	5	6	0	5
giambja01	2011	1	COL	NL	64	64	131	20	34	...	0	3	0	1

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	...	IBB	HBP	SH	SF
giambja01	2012	1	COL	NL	60	NA	89	7	20	...	2	2	0	2
saenzol01	1994	1	CHA	AL	5	5	14	2	2	...	0	0	1	0
saenzol01	1999	1	OAK	AL	97	97	255	41	70	...	1	15	0	3
saenzol01	2000	1	OAK	AL	76	76	214	40	67	...	2	7	0	1
saenzol01	2001	1	OAK	AL	106	106	305	33	67	...	1	13	1	3
saenzol01	2002	1	OAK	AL	68	68	156	15	43	...	1	7	0	2
saenzol01	2004	1	LAN	NL	77	77	111	17	31	...	1	2	0	3
saenzol01	2005	1	LAN	NL	109	109	319	39	84	...	1	3	0	2
saenzol01	2006	1	LAN	NL	103	103	179	30	53	...	1	7	0	4
saenzol01	2007	1	LAN	NL	92	92	110	9	21	...	0	2	0	4
giambja01	2013	1	CLE	AL	71	71	186	21	34	...	0	4	0	3

Since all these players were lost in after 2001 in the offseason, let's only concern ourselves with the data from 2001.

Use filter() function again to only filter the rows where the yearID was 2001.

In [25]:

```
lost_player=lost_player %>% filter(yearID=='2001')
```

In [26]:

```
lost_player
```

playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H	...	IBB	HBP	SH	SF
damonjo01	2001	1	OAK	AL	155	155	644	108	165	...	1	5	5	4
giambja01	2001	1	OAK	AL	154	154	520	109	178	...	24	13	0	9
saenzol01	2001	1	OAK	AL	106	106	305	33	67	...	1	13	1	3

Reduce the lost_players data frame to the following columns: playerID,H,X2B,X3B,HR,OBP,SLG,BA,AB

In [27]:

```
lost_player <- lost_player[,c('playerID','H','X2B','X3B','HR','OBP','SLG','BA','AB')]
```

In [28]:

```
lost_player
```

playerID	H	X2B	X3B	HR	OBP	SLG	BA	AB
damonjo01	165	34	4	9	0.3235294	0.3633540	0.2562112	644
giambja01	178	47	2	38	0.4769001	0.6596154	0.3423077	520
saenzol01	67	21	1	9	0.2911765	0.3836066	0.2196721	305

Use Summary() function to displaying summary statistics of lost_player:

In [29]:

```
summary(lost_player)
```

```
     playerID       H               X2B             X3B              HR       
 damonjo01:1   Min.   : 67.0   Min.   :21.0   Min.   :1.000   Min.   : 9.00  
 giambja01:1   1st Qu.:116.0   1st Qu.:27.5   1st Qu.:1.500   1st Qu.: 9.00  
 saenzol01:1   Median :165.0   Median :34.0   Median :2.000   Median : 9.00  
 aardsda01:0   Mean   :136.7   Mean   :34.0   Mean   :2.333   Mean   :18.67  
 aaronha01:0   3rd Qu.:171.5   3rd Qu.:40.5   3rd Qu.:3.000   3rd Qu.:23.50  
 aaronto01:0   Max.   :178.0   Max.   :47.0   Max.   :4.000   Max.   :38.00  
 (Other)  :0                                                                 
      OBP              SLG              BA              AB       
 Min.   :0.2912   Min.   :0.3634   Min.   :0.2197   Min.   :305.0  
 1st Qu.:0.3074   1st Qu.:0.3735   1st Qu.:0.2379   1st Qu.:412.5  
 Median :0.3235   Median :0.3836   Median :0.2562   Median :520.0  
 Mean   :0.3639   Mean   :0.4689   Mean   :0.2727   Mean   :489.7  
 3rd Qu.:0.4002   3rd Qu.:0.5216   3rd Qu.:0.2993   3rd Qu.:582.0  
 Max.   :0.4769   Max.   :0.6596   Max.   :0.3423   Max.   :644.0  
```

Replacement Players

The total combined salary of the three players cannot be greater than 15 million dollars. Their combined number of At Bats needs to be greater than or equal to the lost players. Their mean OBP must be greater than or equal to the mean OBP of the lost players.

In [30]:

```
avail_player1=s3 %>% filter(yearID=='2001')
avail_player1
```

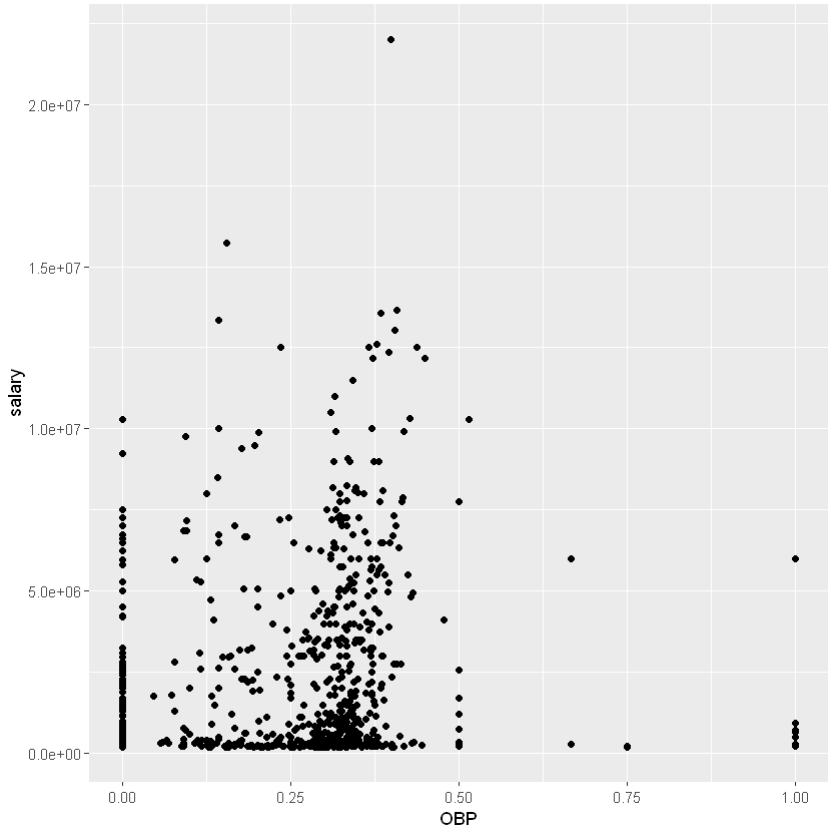
playerID	yearID	teamID.x	lgID.x	salary	stint	teamID.y	lgID.y	G	G_batting	...	IBB
abbotje01	2001	FLO	NL	300000	1	FLO	NL	28	28	...	0
abbotku01	2001	ATL	NL	600000	1	ATL	NL	6	6	...	0
abbotpa01	2001	SEA	AL	1700000	1	SEA	AL	28	2	...	0
abreubo01	2001	PHI	NL	4983000	1	PHI	NL	162	162	...	11
adamste01	2001	LAN	NL	2600000	1	LAN	NL	43	41	...	0
agbaybe01	2001	NYN	NL	260000	1	NYN	NL	91	91	...	0
alfonan01	2001	FLO	NL	2450000	1	FLO	NL	58	54	...	0
alfoned01	2001	NYN	NL	5750000	1	NYN	NL	124	124	...	0
alicelu01	2001	KCA	AL	800000	1	KCA	AL	113	113	...	0
allench01	2001	MIN	AL	240000	1	MIN	AL	57	57	...	1
almanar01	2001	FLO	NL	225000	1	FLO	NL	52	49	...	0
almanca01	2001	NYA	AL	270000	1	NYA	AL	10	1	...	0
alomaro01	2001	CLE	AL	7750000	1	CLE	AL	157	157	...	5
alomasa02	2001	CHA	AL	2900000	1	CHA	AL	70	70	...	1
aloumo01	2001	HOU	NL	5250000	1	HOU	NL	136	136	...	14
anderbr01	2001	BAL	AL	7200000	1	BAL	AL	131	131	...	4
anderbr02	2001	ARI	NL	4125000	1	ARI	NL	33	31	...	0
anderga01	2001	ANA	AL	4500000	1	ANA	AL	161	161	...	4
anderji02	2001	PIT	NL	285000	1	PIT	NL	34	32	...	0
anderma01	2001	DET	AL	355000	1	DET	AL	62	4	...	0
anderma02	2001	PHI	NL	280000	1	PHI	NL	147	147	...	5
ankieri01	2001	SLN	NL	400000	1	SLN	NL	6	6	...	0
appieke01	2001	NYN	NL	8500000	1	NYN	NL	33	31	...	0
ariasal01	2001	SDN	NL	550000	1	SDN	NL	70	70	...	1
armasto02	2001	MON	NL	230000	1	MON	NL	34	32	...	0
arrojro01	2001	BOS	AL	1625000	1	BOS	AL	41	3	...	0
arroybr01	2001	PIT	NL	225000	1	PIT	NL	24	23	...	0
ashbyan01	2001	LAN	NL	6000000	1	LAN	NL	3	3	...	0
astacpe01	2001	COL	NL	6850000	1	COL	NL	22	20	...	0
astacpe01	2001	COL	NL	6850000	2	HOU	NL	4	4	...	0
...	...	...	...	...	...	...	...	...	...		...
wilsokr01	2001	KCA	AL	205000	1	KCA	AL	29	1	...	0
wilsopa02	2001	TBA	AL	350000	1	TBA	AL	37	4	...	0
wilsopr01	2001	FLO	NL	1000000	1	FLO	NL	123	123	...	2
winchsc01	2001	CIN	NL	202000	1	CIN	NL	12	11	...	0

playerID	yearID	teamID.x	lgID.x	salary	stint	teamID.y	lgID.y	G	G_batting	...	IBB
winnra01	2001	TBA	AL	270000	1	TBA	AL	128	128	...	0
wisema01	2001	ANA	AL	207500	1	ANA	AL	11	0	...	NA
witasja01	2001	SDN	NL	800000	1	SDN	NL	31	30	...	0
witasja01	2001	SDN	NL	800000	2	NYA	AL	32	3	...	0
wittbo01	2001	ARI	NL	500000	1	ARI	NL	14	14	...	0
wohlema01	2001	CIN	NL	500000	2	NYA	AL	31	5	...	0
wohlema01	2001	CIN	NL	500000	1	CIN	NL	30	29	...	0
wolfra02	2001	PHI	NL	365000	1	PHI	NL	28	25	...	0
womacto01	2001	ARI	NL	4000000	1	ARI	NL	125	125	...	2
woodast01	2001	CLE	AL	1325000	1	CLE	AL	29	3	...	0
woodke02	2001	CHN	NL	1940000	1	CHN	NL	29	28	...	0
woodwch01	2001	TOR	AL	223000	1	TOR	AL	37	37	...	0
wootesh01	2001	ANA	AL	200500	1	ANA	AL	79	79	...	0
worreti01	2001	SFN	NL	1300000	1	SFN	NL	73	72	...	0
wrighja01	2001	MIL	NL	2350000	1	MIL	NL	34	33	...	0
wrighja02	2001	CLE	AL	2562500	1	CLE	AL	7	1	...	0
wunscke01	2001	CHA	AL	260000	1	CHA	AL	33	2	...	0
yanes01	2001	TBA	AL	650000	1	TBA	AL	54	2	...	0
youngdm01	2001	CIN	NL	3500000	1	CIN	NL	142	142	...	10
younger01	2001	CHN	NL	4500000	1	CHN	NL	149	149	...	1
youngke01	2001	PIT	NL	6125000	1	PIT	NL	142	142	...	3
zaungr01	2001	KCA	AL	1150000	1	KCA	AL	39	39	...	0
zeileto01	2001	NYN	NL	6833333	1	NYN	NL	151	151	...	3
zimmeje02	2001	TEX	AL	307500	1	TEX	AL	66	3	...	0
zitoba01	2001	OAK	AL	240000	1	OAK	AL	35	2	...	0
zuletju01	2001	CHN	NL	200000	1	CHN	NL	49	49	...	1

In [31]:

```
ggplot(avail_player1, aes(x=OBP, y=salary)) + geom_point()
```

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



players with salary less than 8 million and OBP greater than 0

In [38]:

```
avail_player2 <- filter(avail_player1, salary<8000000, OBP > 0)
avail_player2
```

playerID	yearID	teamID.x	lgID.x	salary	stint	teamID.y	lgID.y	G	G_batting	...	IBB
abbotje01	2001	FLO	NL	300000	1	FLO	NL	28	28	...	0
abbotku01	2001	ATL	NL	600000	1	ATL	NL	6	6	...	0
abbotpa01	2001	SEA	AL	1700000	1	SEA	AL	28	2	...	0
abreubo01	2001	PHI	NL	4983000	1	PHI	NL	162	162	...	11
adamste01	2001	LAN	NL	2600000	1	LAN	NL	43	41	...	0
agbaybe01	2001	NYN	NL	260000	1	NYN	NL	91	91	...	0
alfoned01	2001	NYN	NL	5750000	1	NYN	NL	124	124	...	0
alicelu01	2001	KCA	AL	800000	1	KCA	AL	113	113	...	0
allench01	2001	MIN	AL	240000	1	MIN	AL	57	57	...	1
alomaro01	2001	CLE	AL	7750000	1	CLE	AL	157	157	...	5
alomasa02	2001	CHA	AL	2900000	1	CHA	AL	70	70	...	1
aloumo01	2001	HOU	NL	5250000	1	HOU	NL	136	136	...	14
anderbr01	2001	BAL	AL	7200000	1	BAL	AL	131	131	...	4
anderbr02	2001	ARI	NL	4125000	1	ARI	NL	33	31	...	0
anderga01	2001	ANA	AL	4500000	1	ANA	AL	161	161	...	4
anderji02	2001	PIT	NL	285000	1	PIT	NL	34	32	...	0
anderma02	2001	PHI	NL	280000	1	PHI	NL	147	147	...	5
ankieri01	2001	SLN	NL	400000	1	SLN	NL	6	6	...	0
ariasal01	2001	SDN	NL	550000	1	SDN	NL	70	70	...	1
armasto02	2001	MON	NL	230000	1	MON	NL	34	32	...	0
arroybr01	2001	PIT	NL	225000	1	PIT	NL	24	23	...	0
ashbyan01	2001	LAN	NL	6000000	1	LAN	NL	3	3	...	0
astacpe01	2001	COL	NL	6850000	1	COL	NL	22	20	...	0
astacpe01	2001	COL	NL	6850000	2	HOU	NL	4	4	...	0
aurilri01	2001	SFN	NL	3250000	1	SFN	NL	156	156	...	2
ausmubr01	2001	HOU	NL	4250000	1	HOU	NL	128	128	...	6
aybarma01	2001	CHN	NL	665000	1	CHN	NL	17	17	...	0
bagweje01	2001	HOU	NL	6500000	1	HOU	NL	161	161	...	5
baineha01	2001	CHA	AL	1000000	1	CHA	AL	32	32	...	0
bakopa01	2001	ATL	NL	450000	1	ATL	NL	61	61	...	2
...	...	...	...	...	...	...	...	...	...		...
washbja01	2001	ANA	AL	270000	1	ANA	AL	30	2	...	0
weathda01	2001	MIL	NL	1200000	1	MIL	NL	52	50	...	0
wehnejo01	2001	PIT	NL	375000	1	PIT	NL	43	43	...	0
whitede03	2001	MIL	NL	5000000	1	MIL	NL	126	126	...	1

playerID	yearID	teamID.x	lgID.x	salary	stint	teamID.y	lgID.y	G	G_batting	...	IBB
whitero02	2001	CHN	NL	4000000	1	CHN	NL	95	95	...	4
willige02	2001	TBA	AL	3000000	2	NYA	AL	38	38	...	0
willige02	2001	TBA	AL	3000000	1	TBA	AL	62	62	...	0
williwo02	2001	SDN	NL	5083333	2	SLN	NL	11	11	...	0
williwo02	2001	SDN	NL	5083333	1	SDN	NL	26	26	...	0
wilsoda01	2001	SEA	AL	4400000	1	SEA	AL	123	123	...	0
wilsoen01	2001	PIT	NL	635000	1	PIT	NL	46	46	...	0
wilsoen01	2001	PIT	NL	635000	2	NYA	AL	48	48	...	0
wilsoja02	2001	PIT	NL	200000	1	PIT	NL	108	108	...	2
wilsokr01	2001	KCA	AL	205000	1	KCA	AL	29	1	...	0
wilsopr01	2001	FLO	NL	1000000	1	FLO	NL	123	123	...	2
winnra01	2001	TBA	AL	270000	1	TBA	AL	128	128	...	0
wittbo01	2001	ARI	NL	500000	1	ARI	NL	14	14	...	0
wolfra02	2001	PHI	NL	365000	1	PHI	NL	28	25	...	0
womacto01	2001	ARI	NL	4000000	1	ARI	NL	125	125	...	2
woodke02	2001	CHN	NL	1940000	1	CHN	NL	29	28	...	0
woodwch01	2001	TOR	AL	223000	1	TOR	AL	37	37	...	0
wootesh01	2001	ANA	AL	200500	1	ANA	AL	79	79	...	0
wrighja01	2001	MIL	NL	2350000	1	MIL	NL	34	33	...	0
wrighja02	2001	CLE	AL	2562500	1	CLE	AL	7	1	...	0
youngdm01	2001	CIN	NL	3500000	1	CIN	NL	142	142	...	10
younger01	2001	CHN	NL	4500000	1	CHN	NL	149	149	...	1
youngke01	2001	PIT	NL	6125000	1	PIT	NL	142	142	...	3
zaungr01	2001	KCA	AL	1150000	1	KCA	AL	39	39	...	0
zeileto01	2001	NYN	NL	6833333	1	NYN	NL	151	151	...	3
zuletju01	2001	CHN	NL	200000	1	CHN	NL	49	49	...	1

find players with AB higher than or equal to 500

In [33]:

```
avail_player2 <- filter(avail_player2, AB >= 489.7 )
```

sort by OBP in descending order

In [34]:

```
possible <- head(arrange(avail_player2, desc(OBP)), 10)
possible <- possible[,c('playerID', 'OBP', 'AB', 'salary')]
possible
```

playerID	OBP	AB	salary
giambja01	0.4769001	520	4103333
heltoto01	0.4316547	587	4950000
berkmla01	0.4302326	577	305000
gonzalu01	0.4285714	609	4833333
thomeji01	0.4161491	526	7875000
alomaro01	0.4146707	575	7750000
edmonji01	0.4102142	500	6333333
gilesbr02	0.4035608	576	7333333
pujolal01	0.4029630	590	200000
olerujo01	0.4011799	572	6700000

In [35]:

```
p_1=possible[2:4,]
p_1
```

	playerID	OBP	AB	salary
2	heltoto01	0.4316547	587	4950000
3	berkmla01	0.4302326	577	305000
4	gonzalu01	0.4285714	609	4833333

In [36]:

```
sum(p_1$salary)
```

10088333

CONCLUSION:

The sum of combined salary of the three players are less than 15 million dollars.

In [ ]:

```

```