MoneyBall Data Project

The 2002 Oakland Athletics'

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).

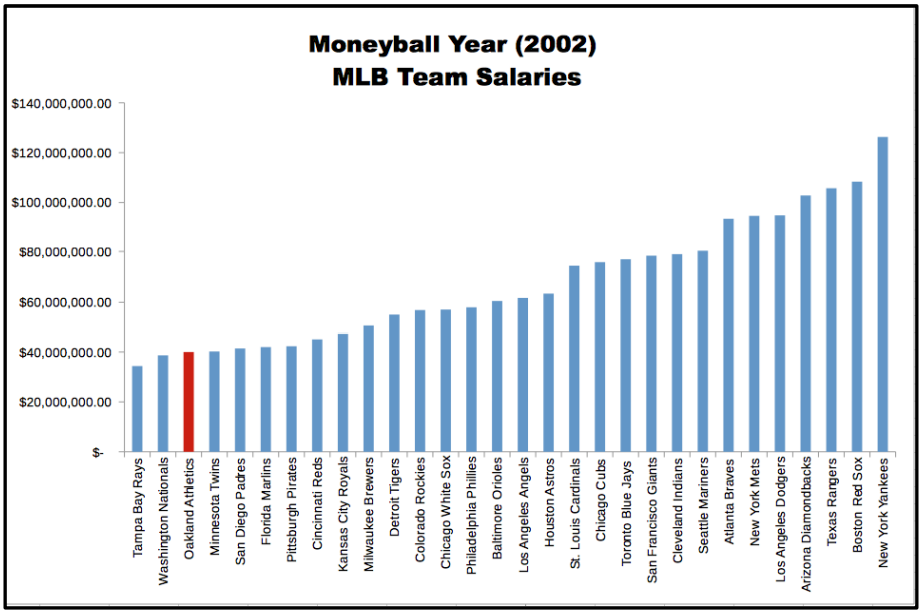
MoneyBall Data Project

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 Athletics 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 44 million dollars in salary, were competitive with larger market teams such as the New York Yankees, who spent over 125 million dollars 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 Athletics to the playoffs in 2002 and 2003.

MoneyBall Data Project

In this project I will 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.

I will be using data from Sean Lahaman's Website – this one is the 'Batting' table
For convenience, a translation of the table headings is attached (they are shown in acronyms).

playerID	yearID	stint	teamID	lgID	G	G_batting	AB
Player ID code	year	player's stint (order of appearances within a season)	Team	League	Games	Game as batter	At Bats
R	H	2B	3B	HR	RBI	SB	CS
Runs	Hits	Doubles	Triples	Homeruns	Runs Batted In	Stolen Bases	Caught Stealing
BB	SO	IBB	HBP	SH	SF	GIDP	G_Old
Base on Balls	Strikeouts	Intentional walks	Hit by pitch	Sacrifice hits	Sacrifice flies	Grounded into double plays	Old version of games (deprecated)

For are data I need to add three more statistics that were used in Moneyball:

- Batting Average
- On Base Percentage
- Slugging Percentage

MoneyBall Data Project

```
# Following this code, I adding these 3 stats to the dataframe

# Batting Average (BA)
# AVG = H (Hits) / AB (At Bats)
Batting$BA =round((Batting$H / Batting$AB),3)
print(tail(Batting$BA))

## [1] 0.000 0.123 0.275 0.147 0.275 0.214

# On Base Percentage (OBP)
Batting$OBP = round((Batting$H + Batting$BB + Batting$HBP)/ (Batting
$AB + Batting$BB + Batting$HBP + Batting$SF),3)
print(tail(Batting$OBP))

## [1] 0.000 0.134 0.344 0.147 0.354 0.290

# Slugging Percentage (SLG)
Batting$B1 = Batting$H - Batting$X2B - Batting$X3B - Batting$HR
Batting$SLG = round((Batting$B1 + (2*Batting$X2B) + (3*Batting$X3B)
+ (4*Batting$HR)) / Batting$AB,3)
print(tail(Batting$SLG))

## [1] 0.000 0.138 0.465 0.147 0.402 0.329
```

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! I have salary information (in the csv file 'Salaries.csv'.)

So now I going to marge these to data sets, but I notice that the minimum year in the batting data is `1871` and the salary data starts from `1985`, so I need to remove the batting data that occurred before `1985`.

```
print(summary(Batting))
```

```
##    playerID          yearID         stint          teamID
##  Length:97889       Min.   :1871   Min.   :1.000   Length:97889
##  Class :character   1st Qu.:1931   1st Qu.:1.000   Class :character
##  Mode  :character   Median :1970   Median :1.000   Mode  :character
##                     Mean   :1962   Mean   :1.077
##
```

```
print(summary(Salaries))
```

```
##     yearID         teamID              lgID              playerID
##  Min.   :1985   Length:23956       Length:23956       Length:23956
##  1st Qu.:1993   Class :character   Class :character   Class
```

```
Batting = subset(Batting, yearID > 1984)
```

Since we have players playing multiple years, we'll have repetitions of playerIDs for multiple years, meaning I need to merge on *both* players and years.

```
combo_data = merge(Batting,Salaries,by = c('playerID','yearID'))
```

MoneyBall Data Project

As previously mentioned, the Oakland Athletics lost 3 key players during the off-season. And I want to get their stats to see what we have to replace.
The players lost were:

1. First baseman 2000 AL MVP **Jason Giambi** (giambja01) to the New York Yankees.
2. Outfielder **Johnny Damon** (damonjo01) to the Boston Red Sox.
3. Infielder **Rainer Gustavo "Ray" Olmedo** ('saenzol01').

```
lost_players = subset(combo_data,playerID %in%
c('giambja01','damonjo01','saenzol01'))
```

Since all these players were lost in after 2001 in the offseason, I will only concern the data from 2001.

```
lost_players = subset(lost_players,yearID == 2001)
lost_players = select(lost_players,playerID,H,X2B,X3B,HR,OBP,SLG,BA,
AB)
print(lost_players)
```

```
##        playerID   H X2B X3B HR   OBP   SLG    BA  AB
## 5141  damonjo01 165  34   4  9 0.324 0.363 0.256 644
## 7878  giambja01 178  47   2 38 0.477 0.660 0.342 520
## 20114 saenzol01  67  21   1  9 0.291 0.384 0.220 305
```

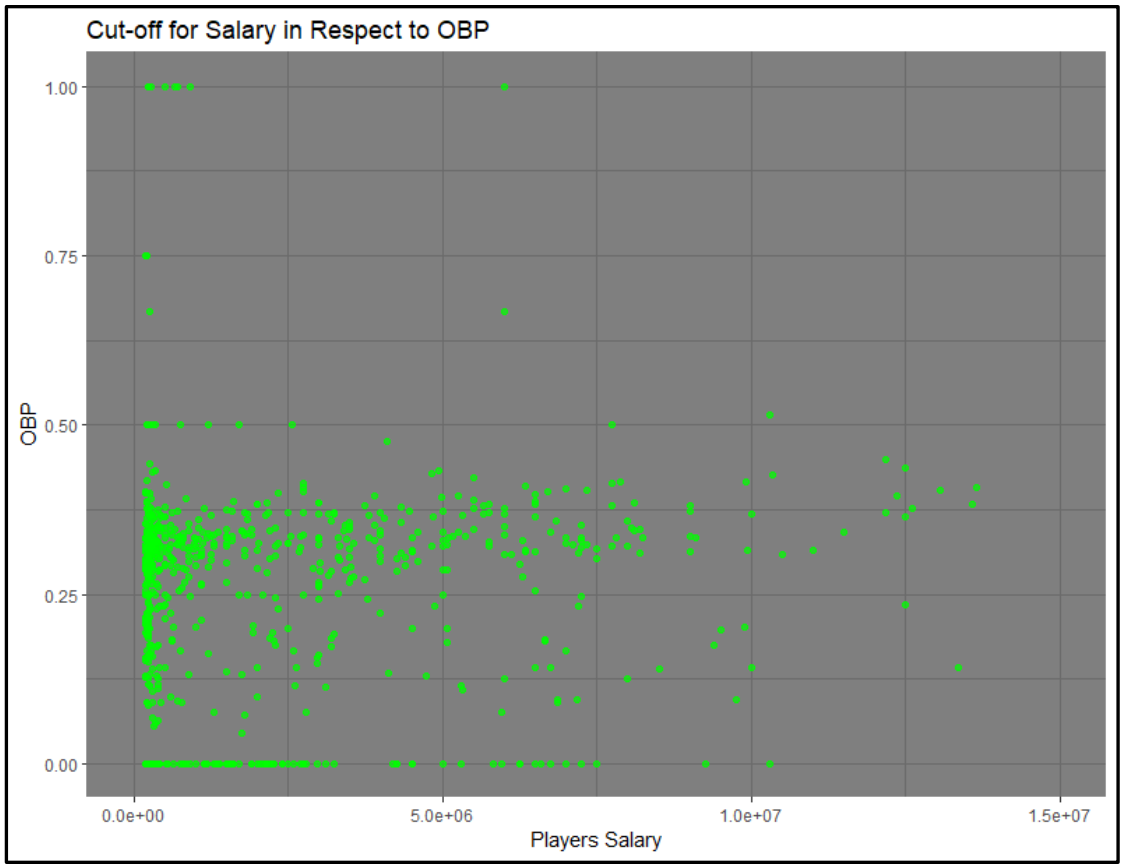
Now I have all the information I need, so now I will search replacement players for the key three players Oakland Athletics lost!

My search for interchanges will take into account the following factors:

- The total combined salary of the three players cannot exceed 15 million dollars.
- Their combined number of At Bats (AB) needs to be equal to or greater than the lost players. (1469)
- Their mean OBP had to equal to or greater than the mean OBP of the lost players. (0.364)

MoneyBall Data Project

First, I will look for the available players in 2001.



From this plot it looks like we don't need to pay more than 7 million $ (this number is from quick look on this plot) also we can see that there a lot of players with OBP == 0 I will remove them to.

The total AB of the lost players is 1469 (more o less 1500), it is mean I should probably cut off the AB from the **available players data 1500/3 = 500**

```
avl_players = filter(avl_players,salary < 7000000,OBP > 0,
AB >= 500,playerID %!in% c(' damonjo01','giambja01','saenzol01'))
```

And now I will arrange my data to find the top 10 available players with the highest OBP score and after it arrange again to find from this top 10 the cheapest players.

My recommendation for replacing the players"

	PlayerID	OBP	AB	Salary
	Pujalal01	0.403	590	200,000
	Berkmla01	0.43	577	305,000
	Gonzalu01	0.429	609	4,833,333
SUM		----------	**1776**	**5,338,333**
MEAN		**0.42**	----------	----------