Capstone Data Project

MoneyBall Project

Source: Wikipedia

Background

The 2002 Oakland A's

ฤดูกาล 2002 ของ Oakland Athletics เป็นฤดูกาลที่ 35 ของทีมใน Oakland, California นอกจากนี้ยังเป็นฤดูกาลที่ 102 ในประวัติศาสตร์แฟรนไชส์อีกด้วย กรีฑาจบอันดับหนึ่งในลีกอเมริกันตะวันตกด้วยสถิติ 103-59

แคมเปญ The Athletics' ในปี 2545 ติดอันดับหนึ่งในแคมเปญที่มีชื่อเสียงที่สุดในประวัติศาสตร์แฟรนไชส์ หลังจากฤดูกาล 2544 โอกแลนด์เห็นการจากไปของผู้เล่นหลักสามคน Billy Beane ผู้จัดการทั่วไปของทีมตอบโต้ด้วยการเซ็นสัญญาตัวแทนฟรีภายใต้เรดาร์ การแข่งขันกรีฑาโฉมใหม่นี้แม้จะขาดดารานำ แต่ก็สร้างความประหลาดใจให้กับโลกเบสบอลด้วยการทำสถิติสูงสุดในฤดูกาลปกติของทีมในปี 2544 อย่างไรก็ตาม ทีมนี้มีชื่อเสียงที่สุดจากการชนะ 20 เกมติดต่อกันระหว่างวันที่ 13 สิงหาคมถึง 4 กันยายน พ.ศ. 2545 ฤดูกาลกรีฑาเป็นเรื่องของหนังสือ Moneyball ของ Michael Lewis ในปี 2546: ศิลปะแห่งการชนะเกมที่ไม่ยุติธรรม (เนื่องจาก Lewis ได้รับโอกาสให้ติดตามทีมตลอดทั้งฤดูกาลนั้น)

โปรเจ็กต์นี้สร้างจากหนังสือที่เขียนโดย Michael Lewis (ต่อมากลายเป็นภาพยนตร์)

Moneyball Book

หลักการสำคัญของหนังสือ*Moneyball*คือความรู้ร่วมกันของคนวงในเบสบอล (รวมถึงผู้เล่น ผู้จัดการ โค้ช แมวมอง และสำนักงานส่วนหน้า) ในช่วงศตวรรษที่ผ่านมานั้นเป็นเรื่องส่วนตัวและมักมีข้อบกพร่อง สถิติต่างๆ เช่น ฐานที่ถูกขโมย รันตีใน และค่าเฉลี่ยการตีโดยทั่วไปจะใช้เพื่อประเมินผู้เล่น เป็นโบราณวัตถุของมุมมองของเกมในศตวรรษที่ 19 และสถิติที่มีอยู่ในขณะนั้น หนังสือเล่มนี้ให้เหตุผลว่าสำนักงานส่วนหน้าของ Oakland A ใช้ประโยชน์จากมาตรวัดประสิทธิภาพของผู้เล่นในการวิเคราะห์มากขึ้นเพื่อให้ทีมสามารถแข่งขันกับคู่แข่งที่ร่ำรวยกว่าในเมเจอร์ลีกเบสบอล (MLB) ได้ดีขึ้น

การวิเคราะห์ทางสถิติอย่างเข้มงวดได้แสดงให้เห็นว่าเปอร์เซ็นต์บนฐานและเปอร์เซ็นต์การถลอกเป็นตัวบ่งชี้ความสำเร็จของการรุกได้ดีกว่า และฝ่าย A ก็เชื่อมั่นว่าคุณสมบัติเหล่านี้ถูกกว่าที่จะหาได้ในตลาดเปิดมากกว่าคุณสมบัติที่มีคุณค่าทางประวัติศาสตร์ เช่น ความเร็วและการสัมผัส ข้อสังเกตเหล่านี้มักขัดแย้งกับภูมิปัญญาดั้งเดิมของเบสบอลและความเชื่อของหน่วยสอดแนมและผู้บริหารทีมเบสบอลหลายคน

ด้วยการประเมินกลยุทธ์ที่สร้างชัยชนะในสนามอีกครั้ง การแข่งขันกรีฑาปี 2545 ซึ่งมีเงินเดือนประมาณ 44 ล้านดอลลาร์สหรัฐสามารถแข่งขันกับทีมในตลาดที่ใหญ่กว่า เช่น นิวยอร์กแยงกี้ส์ ซึ่งใช้จ่ายเงินเดือนกว่า 125 ล้านดอลลาร์สหรัฐในฤดูกาลเดียวกันนั้น

นื่องจากรายได้ที่น้อยลงของทีม Oakland จึงถูกบังคับให้หาผู้เล่นที่มีมูลค่าต่ำกว่าตลาด และระบบของพวกเขาสำหรับการค้นหามูลค่าของผู้เล่นที่มีมูลค่าต่ำได้พิสูจน์ตัวเองแล้ว วิธีการนี้ทำให้ A เข้าสู่รอบตัดเชือกในปี 2545 และ 2546

ในโครงการนี้ เราจะทำงานกับข้อมูลบางอย่างและมีเป้าหมายในการพยายามหาผู้เล่นทดแทนสำหรับผู้เล่นที่เสียไปในช่วงเริ่มนอกฤดูกาล - ในช่วงนอกฤดูกาล 2544–02 ทีมได้สูญเสียตัวแทนอิสระหลักสามคนในตลาดที่ใหญ่ขึ้น ทีม: 2000 AL MVP Jason Giambi กับ New York Yankees, จอห์นนี่ เดม่อน กองกลางของทีม Boston Red Sox และ Jason Isringhausen ที่ใกล้ชิดกับ St. Louis Cardinals

Get Data

เราจะใช้ข้อมูลจาก Sean Lahaman's Website ซึ่งเป็นแหล่งข้อมูลที่มีประโยชน์มากสำหรับสถิติเบสบอล

```
df <- read.csv('Batting.csv')
head(df)
```

→ Use str() to check the structure. Pay close attention to how columns that start with a number get an 'X' in front of them! You'll need to know this to call those columns!

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

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

→ Call the head of the doubles (X2B) column

```
head(df$AB)
head(df$X2B)
```

Feature Engineering

- Batting Average
- On Base Percentage
- 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:

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

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

→ On Base Percentage (OBP) and Slugging Percentage (SLG). (H): 1B = H-2B-3B-HR

- **Create an OBP Column**
- **Create an SLG Column**

```
# On Base Percentage
df$OBP <- (df$H + df$BB + df$HBP)/(df$AB + df$BB + df$HBP + df$SF)

# Creating X1B (Singles)
X1B <- df$H - df$X2B - df$X3B - df$HR

# Creating Slugging Average (SLG)
df$SLG <- (X1B + 2*df$X2B + 3*df$X3B + 4*df$HR)/df$AB
```

Merging Salary Data with Batting Data

→ Load the Salaries.csv file

```
sal <- read.csv('Salaries.csv')
```

Use summary to get a summary of the batting data frame and notice the minimum year in the yearID column. Our batting data goes back to 1871! Our salary data starts at 1985, meaning we need to remove the batting data that occured before 1985.

→ Use subset() to reassign batting to only contain data from 1985 and onwards

```
df <- subset(df, df$yearID >= 1985)
```

→ Now use summary again to make sure the subset reassignment worked, your yearID min should be 1985

```
summary(df)
```

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

→ Use summary to check the data

```
combo <- merge(df, sal, by = c('playerID', 'yearID'))
summary(combo)
```

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 subset() function to get a data frame called lost_players from the combo data frame consisting of those 3 players.

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

→ Use subset again to only grab the rows where the yearID was 2001.

```
lost_p2001 <- subset(lost_players, lost_players$yearID == 2001)
```

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

```
lost_players <- subset(lost_players, lost_players$yearID == 2001,
                       select = c(playerID,H,X2B,X3B,HR,OBP,SLG,BA,AB))
# summarize mean(OBP) , sum(AB) from lost_players
lost_s <- summarize(lost_players, mean_OBP = mean(OBP), sum_AB = sum(AB))
```

Replacement Players

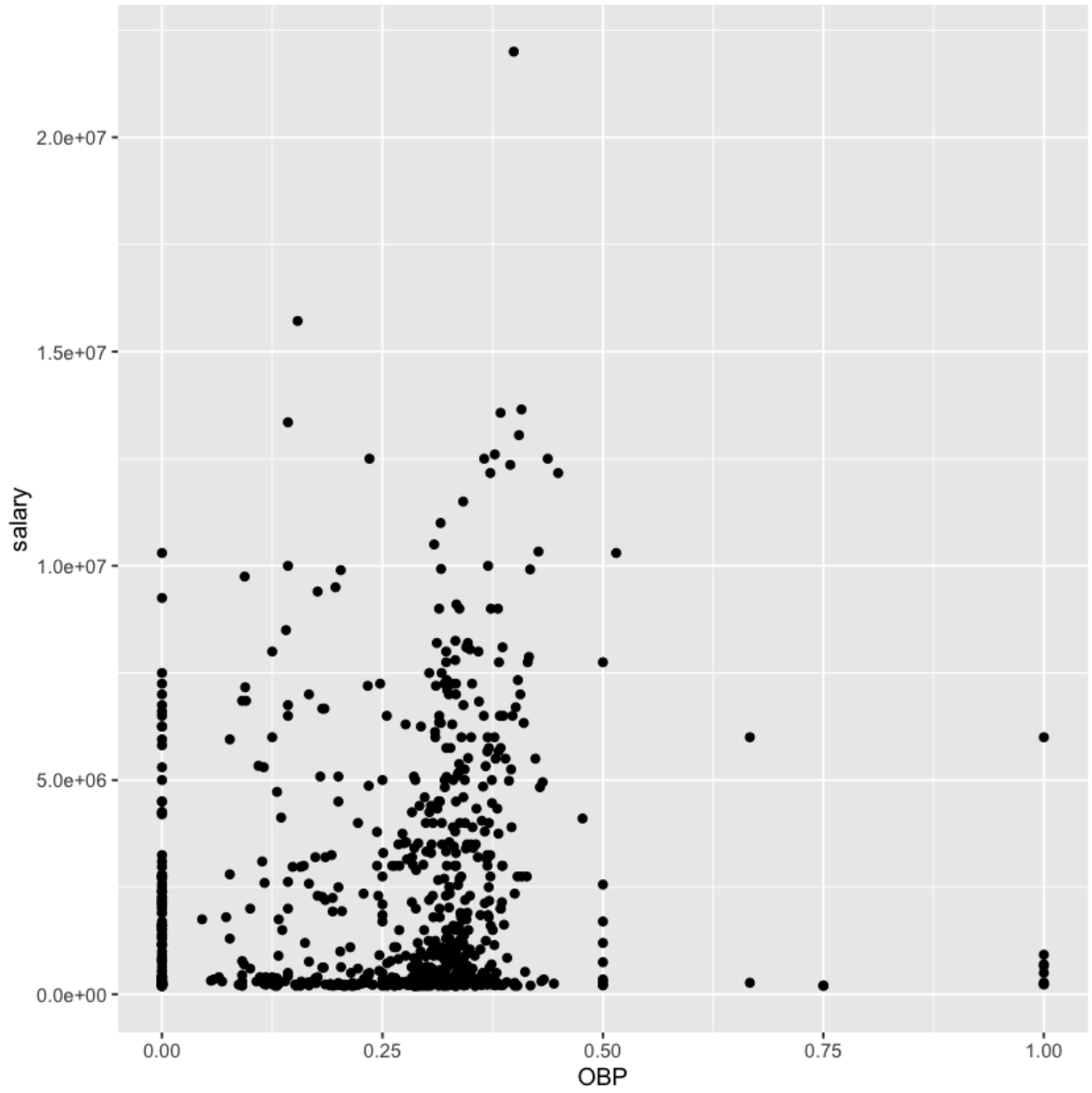
- The total combined salary of the three players can not exceed 15 million dollars.
- Their combined number of At Bats (AB) needs to be equal to or greater than the lost players.
- Their mean OBP had to equal to or greater than the mean OBP of the lost players

→ Use the combo dataframe you previously created as the source of information! Remember to just use the 2001 subset of that dataframe. There's lost of different ways you can do this, so be creative! It should be relatively simple to find 3 players that satisfy the requirements, note that there are many correct combinations available!

```
# Replacement Players
# clear lost_players in 2001
combo_s <- subset(combo, !combo$playerID %in%
                  c("giambja01", "damonjo01", "saenzol01") & combo$yearID == 2001)
# select columns 'playerID','OBP','AB','salary'
combo_s <- subset(combo_s, select = c('playerID','OBP','AB','salary'))
# clear missing value
combo_s <- subset(combo_s, ! is.na(combo_s$OBP))
# plot relation OBP , salary find insight
ggplot(combo_s,aes(x=OBP,y=salary)) + geom_point()

# filter clear high salary and OBP = 0
combo_s <- subset(combo_s, combo_s$salary < 8000000 & combo_s$OBP > 0)
# lost players sum_AB 1469 -> 1469/3 This is about 490
combo_s <- subset(combo_s, combo_s$AB >= 490)
# sort by OBP
combo_s <- arrange(combo_s, desc(OBP))
head(combo_s)
# Replacement Players == good_players
good_players <- combo_s[1:3,]
# summarize mean(OBP) , sum(AB) from good_players
good_ps <- summarise(good_players, mean_OBP = mean(OBP),
                     sum_AB = sum(AB), sum_salary = sum(salary))


# summary Replacement Players
cat("[Lost Players] \nmean_OBP:",lost_s$mean_OBP, "\nsum_AB:", lost_s$sum_AB,
    "\n[Good Players] \nmean_OBP:", good_ps$mean_OBP, "\nsum_AB:", good_ps$sum_AB, "\ntotal_salary:", good_ps$sum_salary)
```


2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
salary
0.00
0.25
0.50
0.75
1.00
OBP