EM_Project7

Elizabeth McGuckin

October 25, 2018

Question 1 The Battling.csv file contains the data for home runs and stolen bases. Since we want to see the names of the players who hit 40 home runs and stole 40 bases in one season we set criteria to look for the names of those who hit 40 or more home runs and stole 40 or more bases within this file. Using 'head' allowed us to see the first 5 rows of data for the different categories.

```
myDF = read.csv("/home/emcgucki/Downloads/baseballdatabank-master/core/Batting.csv")
head(myDF)
```

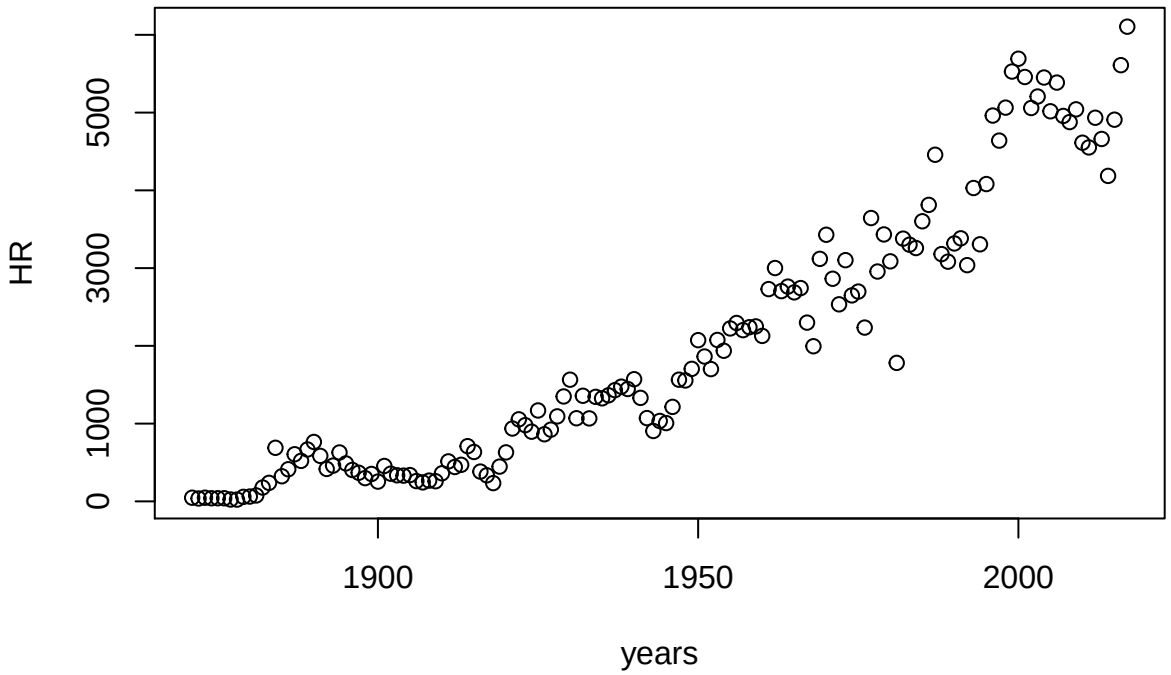
```
##    playerID yearID stint teamID lgID  G  AB  R  H X2B X3B HR RBI SB CS BB
## 1 abercda01   1871     1    TRO <NA>  1   4  0  0   0   0  0   0  0  0  0
## 2  addybo01   1871     1    RC1 <NA> 25 118 30 32   6   0  0  13  8  1  4
## 3 allisar01   1871     1    CL1 <NA> 29 137 28 40   4   5  0  19  3  1  2
## 4 allisdo01   1871     1    WS3 <NA> 27 133 28 44  10   2  2  27  1  1  0
## 5 ansonca01   1871     1    RC1 <NA> 25 120 29 39  11   3  0  16  6  2  2
## 6 armstbo01   1871     1    FW1 <NA> 12  49  9 11   2   1  0   5  0  1  0
##   SO IBB HBP SH SF GIDP
## 1  0  NA  NA NA NA    0
## 2  0  NA  NA NA NA    0
## 3  5  NA  NA NA NA    1
## 4  2  NA  NA NA NA    0
## 5  1  NA  NA NA NA    0
## 6  1  NA  NA NA NA    0
```

```
myDF[myDF$HR>=40 & myDF$SB>=40, ]
```

```
##        playerID yearID stint teamID lgID   G  AB   R   H X2B X3B HR RBI SB
## 65469 cansejo01   1988     1    OAK   AL 158 610 120 187  34   0 42 124 40
## 74270 bondsba01   1996     1    SFN   NL 158 517 122 159  27   3 42 129 40
## 77656 rodrial01   1998     1    SEA   AL 161 686 123 213  35   5 42 124 46
## 88497 soriaal01   2006     1    WAS   NL 159 647 119 179  41   2 46  95 41
##       CS  BB  SO IBB HBP SH SF GIDP
## 65469 16  78 128  10  10  1  6   15
## 74270  7 151  76  30   1  0  6   11
## 77656 13  45 121   0  10  3  4   12
## 88497 17  67 160  16   9  2  3    3
```

Question 2 In question 2 we are looking to make a graph which shows the total number of home runs per year. We use the tapply function to find the total number of home runs per year. We also use 'sum' since we are looking for the toal. To make the graph we plot the tapply function and set the lables for the axis as 'year' and 'HR'.

```
v <- tapply(myDF$HR, myDF$yearID, sum)
plot(names(v),v, xlab="years", ylab="HR")
```



Question 3 In question 3 we want to be able to see full names instead of player ID. We merge databanks myBatting and myPeople in order to see the relationship between the two. By using the 'paste' function we are able to see the player's first and last name. In order to apply this to the names of players who have both hit at least 40 home runs and stolen at least 40 bases we tile the function as 'x' and incorporate it in the previously used code in question 1.

```
myBatting = read.csv("/home/emcgucki/Downloads/baseballdatabank-master/core/Batting.csv")
myPeople = read.csv("/home/emcgucki/Downloads/baseballdatabank-master/core/People.csv")
myQ3 <- merge(myBatting, myPeople, by="playerID")

x <- paste(myQ3$nameFirst, myQ3$nameLast)

myQ3$names <- x
myQ3[myQ3$HR>=40 & myQ3$SB>=40, ]
```

```
##        playerID yearID stint teamID lgID   G  AB   R   H X2B X3B HR RBI SB
## 8520  bondsba01   1996     1    SFN   NL 158 517 122 159  27   3 42 129 40
## 14008 cansejo01   1988     1    OAK   AL 158 610 120 187  34   0 42 124 40
## 80212 rodrial01   1998     1    SEA   AL 161 686 123 213  35   5 42 124 46
## 88502 soriaal01   2006     1    WAS   NL 159 647 119 179  41   2 46  95 41
##       CS  BB  SO IBB HBP SH SF GIDP birthYear birthMonth birthDay
## 8520   7 151  76  30   1  0  6   11      1964          7       24
## 14008 16  78 128  10  10  1  6   15      1964          7        2
## 80212 13  45 121   0  10  3  4   12      1975          7       27
## 88502 17  67 160  16   9  2  3    3      1976          1        7
##       birthCountry           birthState            birthCity deathYear
## 8520           USA                   CA            Riverside        NA
## 14008         Cuba            La Habana            La Habana        NA
## 80212          USA                   NY             New York        NA
## 88502         D.R. San Pedro de Macoris San Pedro de Macoris        NA
##       deathMonth deathDay deathCountry deathState deathCity nameFirst
## 8520          NA       NA                                      Barry
## 14008         NA       NA                                       Jose
```

```
## 80212         NA       NA                                      Alex
## 88502         NA       NA                                   Alfonso
##        nameLast          nameGiven weight height bats throws      debut
## 8520      Bonds        Barry Lamar    185     73    L      L 1986-05-30
## 14008   Canseco               Jose    240     76    R      R 1985-09-02
## 80212 Rodriguez Alexander Enmanuel    230     75    R      R 1994-07-08
## 88502   Soriano  Alfonso Guilleard    195     73    R      R 1999-09-14
##        finalGame  retroID   bbrefID           names
## 8520  2007-09-26 bondb001 bondsba01     Barry Bonds
## 14008 2001-10-06 cansj001 cansejo01    Jose Canseco
## 80212 2016-08-12 rodra001 rodrial01  Alex Rodriguez
## 88502 2014-07-05 soria001 soriaal01 Alfonso Soriano
```