DA Lab4_HW

Na SeungChan

2023-10-02

Exercise 1

```
manny <- Batting %>% filter(playerID == "ramirma02")
head(manny)
```

```
##    playerID yearID stint teamID lgID   G  AB   R   H X2B X3B HR RBI SB CS BB
## 1 ramirma02   1993     1    CLE   AL  22  53   5   9   1   0  2   5  0  0  2
## 2 ramirma02   1994     1    CLE   AL  91 290  51  78  22   0 17  60  4  2 42
## 3 ramirma02   1995     1    CLE   AL 137 484  85 149  26   1 31 107  6  6 75
## 4 ramirma02   1996     1    CLE   AL 152 550  94 170  45   3 33 112  8  5 85
## 5 ramirma02   1997     1    CLE   AL 150 561  99 184  40   0 26  88  2  3 79
## 6 ramirma02   1998     1    CLE   AL 150 571 108 168  35   2 45 145  5  3 76
##    SO IBB HBP SH SF GIDP
## 1   8   0   0  0  0    3
## 2  72   4   0  0  4    6
## 3 112   6   5  2  5   13
## 4 104   8   3  0  9   18
## 5 115   5   7  0  4   19
## 6 121   6   6  0 10   18
```

```
manny %>% summarize(
  span = paste(min(yearID), max(yearID), sep = "-"),
  numYears = n_distinct(yearID), numTeams = n_distinct(teamID),
  BA = sum(H)/sum(AB), tH = sum(H), tHR = sum(HR), tRBI = sum(RBI)
  )
```

```
##        span numYears numTeams        BA   tH tHR tRBI
## 1 1993-2011       19        5 0.3122271 2574 555 1831
```

1-1

```
manny_q1 <- manny %>%
  group_by(teamID) %>%
  summarise(span = paste(min(yearID), max(yearID), sep = "-"), totalHit = sum(H), totalHR = sum(HR), BA
manny_q1
```

```
## # A tibble: 5 x 5
##   teamID span      totalHit totalHR     BA
##   <fct>  <chr>        <int>   <int>  <dbl>
## 1 BOS    2001-2008     1232     274 0.312
## 2 CHA    2010-2010       18       1 0.261
## 3 CLE    1993-2000     1086     236 0.313
## 4 LAN    2008-2010      237      44 0.322
## 5 TBA    2011-2011        1       0 0.0588
```

어떤 기록을 표시해야 하는가? 문제에서 정확히 정의되어 있지 않지만… 우선 각 팀별 안타(totalHit), 각 팀별 홈런(totalHR), 각 팀별 타율(BA), 각 팀별 기간을 복원하는 것이 적절해 보인다. 이에 각 팀별 기록을 적절히 복원하였다. 그런데… teamID의 알파벳 순으로 정렬되었고, 이는 일반적으로 선수의 기록에 기대하는 시간 순 정렬과 부합하지 않는다. 따라서 시간 순으로 정렬되도록 arrange()함수를 추가로 사용하였다. 다행히도 이 함수는 시간에 대해 별도의 처리를 하지 않아도 잘 동작한다.

```
arrange(manny_q1, span)
```

```
## # A tibble: 5 x 5
##   teamID span      totalHit totalHR     BA
##   <fct>  <chr>        <int>   <int>  <dbl>
## 1 CLE    1993-2000     1086     236 0.313
## 2 BOS    2001-2008     1232     274 0.312
## 3 LAN    2008-2010      237      44 0.322
## 4 CHA    2010-2010       18       1 0.261
## 5 TBA    2011-2011        1       0 0.0588
```

1-2

```
table_q2 <- full_join(Batting, People, by = "playerID") %>%
  mutate(fullName = paste(nameGiven, nameLast))

table_q2 %>%
  group_by(fullName) %>%
  summarise(totalHR = sum(HR)) %>%
  arrange(desc(totalHR)) %>%
  slice_head(n = 20)
```

```
## # A tibble: 20 x 2
##    fullName                    totalHR
##    <chr>                         <int>
##  1 George Kenneth Griffey          782
##  2 Barry Lamar Bonds               762
##  3 Henry Louis Aaron               755
##  4 George Herman Ruth              714
##  5 Jose Alberto Pujols             703
##  6 Alexander Enmanuel Rodriguez    696
##  7 Willie Howard Mays              660
##  8 James Howard Thome              612
##  9 Samuel Peralta Sosa             609
## 10 Frank Robinson                  586
```

```
## 11 Mark David McGwire                583
## 12 Harmon Clayton Killebrew          573
## 13 Rafael Palmeiro                   569
## 14 Reginald Martinez Jackson         563
## 15 Manuel Aristides Ramirez          555
## 16 Vladimir Guerrero                 553
## 17 Michael Jack Schmidt              548
## 18 David Americo Ortiz               541
## 19 Mickey Charles Mantle             536
## 20 James Emory Foxx                  534
```

```
table_q2a <- Batting %>%
  group_by(playerID) %>%
  summarise(totalHR = sum(HR)) %>%
  left_join(People, by = "playerID") %>%
  mutate(fullName = paste(nameGiven, nameLast)) %>%
  select(fullName, totalHR) %>%
  arrange(desc(totalHR)) %>%
  slice_head(n = 20)

table_q2a
```

```
## # A tibble: 20 x 2
##    fullName                     totalHR
##    <chr>                          <int>
##  1 Barry Lamar Bonds                762
##  2 Henry Louis Aaron                755
##  3 George Herman Ruth               714
##  4 Jose Alberto Pujols              703
##  5 Alexander Enmanuel Rodriguez     696
##  6 Willie Howard Mays               660
##  7 George Kenneth Griffey           630
##  8 James Howard Thome               612
##  9 Samuel Peralta Sosa              609
## 10 Frank Robinson                   586
## 11 Mark David McGwire               583
## 12 Harmon Clayton Killebrew         573
## 13 Rafael Palmeiro                  569
## 14 Reginald Martinez Jackson        563
## 15 Manuel Aristides Ramirez         555
## 16 Michael Jack Schmidt             548
## 17 David Americo Ortiz              541
## 18 Mickey Charles Mantle            536
## 19 James Emory Foxx                 534
## 20 Willie Lee McCovey               521
```

우선, fullname을 정의하기 위해 `nameGiven` + `nameLast`를 `paste()`함수를 통해 합쳐 하나의 문자열로 만들고 `mutate()`를 통해 새 변수로 생성했다.

한편, key는 playerID여야 한다. full name을 key로 쓰는 경우, George Kenneth Griffey가 홈런 1위가 되는 광경을 목격할 수 있다. 부자가 같은 이름을 쓰면서 'senior'와 'junior'로 구분되었다고 하는데, 앞서의 과정을 통해 정의된 `fullName`에서는 구분이 되지 않는다.

이에, `playerID`를 key로 사용하여 `group_by()`함수로 묶어 총 홈런 개수 `totalHR`를 구하였다. 이 경우 `group_by()` 함수를 사용할 때 key에 `playerID`를 대응시킨 경우 미리 원 데이터에 만들어 둔 `fullName`을

가져오기 난해한 측면이 있어 우선 `Batting` Dataset에서 `totalHR`을 구한 뒤 `left_join()`함수를 사용해 playerID 대신 fullName을 대응시키는 과정을 수행하였다. join을 먼저 하는 방법은 없는지 현재 조사 중이다.

```
pluck(table_q2a['fullName'], 1)[1]
```

```
## [1] "Barry Lamar Bonds"
```

```
pluck(table_q2a['fullName'], 1)[15]
```

```
## [1] "Manuel Aristides Ramirez"
```

Manuel Aristides Ramirez가 15번째 줄에 존재하며, Barry Lamar Bonds가 1번째 줄에 존재한다.

1-3

```
table_q3 <- Pitching %>%
  group_by(playerID) %>%
  summarise(totalW = sum(W), totalSO = sum(SO)) %>%
  left_join(People, by = 'playerID') %>%
  mutate(fullName = paste(nameGiven, nameLast)) %>%
  filter(totalW >= 300&totalSO >= 3000) %>%
  select(fullName, totalW, totalSO)
table_q3
```

```
## # A tibble: 10 x 3
##    fullName              totalW totalSO
##    <chr>                  <int>   <int>
##  1 Steven Norman Carlton    329    4136
##  2 William Roger Clemens    354    4672
##  3 Randall David Johnson    303    4875
##  4 Walter Perry Johnson     417    3509
##  5 Gregory Alan Maddux      355    3371
##  6 Philip Henry Niekro      318    3342
##  7 Gaylord Jackson Perry    314    3534
##  8 Lynn Nolan Ryan          324    5714
##  9 George Thomas Seaver     311    3640
## 10 Donald Howard Sutton     324    3574
```

1-4

```
df <- AwardsPlayers %>% filter(awardID == "World Series MVP")
table_q4 <- left_join(df, People, by = "playerID") %>%
  select(playerID, awardID, yearID, nameFirst:nameGiven, birthYear:birthDay) %>%
  arrange(desc(yearID)) %>%
  slice_head(n = 10) %>%
  mutate(age_at_award = ifelse(birthMonth <= 10, yearID - birthYear , yearID - birthYear + 1))
```

df는 월드시리즈 MVP들의 리스트이고, key로 playerID를 공유하므로 left_join()함수를 통해 People dataset과 합치면 적절하다. 그러나, People dataset은 나이를 직접 밝히지 않으므로 '수상 당시의 나이'는 직접 계산하여 넣어야 한다. 이 과정에서 mutate() 함수가 사용되었다.

사실 나이를 엄밀하게 계산하려면 '월드시리즈 MVP'가 확정된 날짜가 각 시즌마다 유동적이므로 월드시리즈 MVP가 확정된 날짜를 각 시즌별로 조사하고, 각 시즌별로 해당 선수의 birthDate와 해당 날짜까지 비교해야 정확하지만… 찾아보니 그런 사람은 없었다! 그러나, 2013년의 월드 시리즈 MVP ortizda01의 경우 월드 시리즈 MVP를 확정지은 이후 생일을 맞았고, 나머지 9인의 플레이어는 그 해 생일을 맞은 후 월드 시리즈 MVP를 확정지었다. 위의 나이는 이와 같은 요소를 고려하여 진행한 분석이다.

Exercise 2

2-1

(a)

```
table_q2 <- HELPfull %>%
  select(ID, TIME, DRUGRISK, SEXRISK)
```

(b)

```
table_q2b <- table_q2 %>%
  filter(ID == 3)
unique(table_q2b$TIME)
```

```
## [1]  0  6 24
```

0, 6, 24시에 측정되었다.

(c)

```
table_q2ca <- table_q2 %>%
  filter(TIME == 0|TIME == 6, ID >= 1 & ID <= 3) %>%
  select(-SEXRISK) %>%
  pivot_wider(names_from = TIME, values_from = DRUGRISK)

colnames(table_q2ca) <- c('ID', 'DRUGRISK_0', 'DRUGRISK_6')

table_q2cb <- table_q2 %>%
  filter(TIME == 0|TIME == 6, ID >= 1 & ID <= 3) %>%
  select(-DRUGRISK) %>%
  pivot_wider(names_from = TIME, values_from = SEXRISK)

colnames(table_q2cb) <- c('ID', 'SEXRISK_0', 'SEXRISK_6')

table_q2c <- full_join(table_q2ca, table_q2cb, join_by('ID'))
table_q2c
```

```
## # A tibble: 3 x 5
##      ID DRUGRISK_0 DRUGRISK_6 SEXRISK_0 SEXRISK_6
##   <int>      <int>      <int>     <int>     <int>
## 1     1          0          0         4         1
## 2     2          0          0         7         0
## 3     3         20         13         2         4
```

이렇게 ID가 key로 잘 작동하는 방법을 이용해서 쪼갰다 붙이는 거 말고 한 번에 처리하는 방법이 있을 것 같은데…

(d)

```
table_q2d <- table_q2 %>%
  filter(TIME == 0|TIME == 6) %>%
  pivot_wider(names_from = TIME, values_from = c(DRUGRISK, SEXRISK))

cor(table_q2d, use = "complete.obs")
```

```
##                       ID   DRUGRISK_0  DRUGRISK_6    SEXRISK_0    SEXRISK_6
## ID          1.000000000 -0.065954189  0.04820307 -0.060054220  0.009676324
## DRUGRISK_0 -0.065954189  1.000000000  0.59911459 -0.004570644 -0.032417415
## DRUGRISK_6  0.048203072  0.599114586  1.00000000 -0.026745382  0.116530917
## SEXRISK_0  -0.060054220 -0.004570644 -0.02674538  1.000000000  0.504800397
## SEXRISK_6   0.009676324 -0.032417415  0.11653092  0.504800397  1.000000000
```

그냥 c()로 묶어 주면 되는군요! DRUGRISK에 대해서는 0.5991이고, SEXRISK에 대해서는 0.5048이다.

Exercise 3

3-1

```
Macbeth <- Macbeth_raw %>%
  str_split("\r\n") %>%
  pluck(1)#
```

(a)

```
Macbeth %>%
  str_subset("[A-z]+-[A-z]+") %>%
  str_extract("[A-z]+-[A-z]+")
```

```
##  [1] "Gutenberg-tm"      "GUTENBERG-tm"      "AS-IS"
##  [4] "self-comparisons"  "rump-fed"          "tempest-toss"
##  [7] "theme-I"           "all-hailed"        "top-full"
## [10] "all-hail"          "temple-haunting"   "be-all"
## [13] "even-handed"       "trumpet-tongued"   "taking-off"
```

```
## [16] "new-born"          "sticking-place"    "men-children"
## [19] "heat-oppressed"    "half-world"        "firm-set"
## [22] "Re-enter"          "devil-porter"      "nose-painting"
## [25] "Re-enter"          "Re-enter"          "leave-taking"
## [28] "horses-a"          "prophet-like"      "Re-enter"
## [31] "demi-wolves"       "shard-borne"       "cut-throats"
## [34] "not-Are"           "air-drawn"         "Re-enter"
## [37] "self-abuse"        "hedge-pig"         "thirty-one"
## [40] "blind-worm"        "hell-broth"        "salt-sea"
## [43] "birth-strangled"   "Ditch-deliver"     "pale-hearted"
## [46] "lion-mettled"      "earth-bound"       "high-placed"
## [49] "gold-bound"        "blood-bolter"      "shag-ear"
## [52] "leave-taking"      "ill-composed"      "more-having"
## [55] "summer-seeming"    "king-becoming"     "bloody-scepter"
## [58] "over-credulous"    "here-approach"     "here-remain"
## [61] "strangely-visited" "fee-grief"         "hell-kite"
## [64] "faith-breach"      "cream-faced"       "over-red"
## [67] "lily-liver"        "whey-face"         "Seyton-I"
## [70] "mouth-honor"       "thick-coming"      "night-shriek"
## [73] "Re-enter"          "bear-like"         "Re-enter"
## [76] "fiend-like"
```

(b)

```
Macbeth %>%
  str_subset("[A-z]+(more|less) ") %>%
  str_extract("[A-z]+(more|less)")
```

```
##  [1] "harmless"    "merciless"   "careless"    "peerless"    "sightless"
##  [6] "sightless"   "boneless"    "measureless" "bless"       "bless"
## [11] "Thriftless"  "dauntless"   "fruitless"   "reckless"    "restless"
## [16] "Unless"      "Bless"       "confineless" "Boundless"   "stanchless"
```

'Bless'는 형용사가 아닌데, 확인해 보면 'God Bless Us'에서 튀어나온 말이다. 이런 words를 처리하려면 안타깝지만 정규 표현식만으로는 답이 없다. 일일이 예외 처리를 해 주든지 자연어 처리를 도입하든지 해야 한다. 둘 다 정규 표현식 문제에서 할 일은 아닌 것 같다…

(c)

```
Macbeth %>%
  str_subset("(Exit|Exeunt)")
```

```
##  [1] "    Hover through the fog and filthy air.                Exeunt."
##  [2] "                                        Exit Sergeant, attended."
##  [3] "                                                         Exeunt."
##  [4] "  MACBETH. Till then, enough. Come, friends.             Exeunt."
##  [5] "    Which the eye fears, when it is done, to see.          Exit."
##  [6] "    It is a peerless kinsman.                  Flourish. Exeunt."
##  [7] "    He brings great news.                        Exit Messenger."
```

```
##  [8] "    Leave all the rest to me.                                Exeunt."
##  [9] "    By your leave, hostess.                                  Exeunt."
## [10] "                                                             Exeunt."
## [11] "                                       Exeunt Banquo. and Fleance."
## [12] "    She strike upon the bell. Get thee to bed.     Exit Servant."
## [13] "    That summons thee to heaven, or to hell.                 Exit."
## [14] "    For it must seem their guilt.         Exit. Knocking within."
## [15] "                                                           Exeunt."
## [16] "    For 'tis my limited service.                             Exit."
## [17] "                                     Exeunt Macbeth and Lennox."
## [18] "                             Exeunt all but Malcolm and Donalbain."
## [19] "                                                           Exeunt."
## [20] "                                                           Exeunt."
## [21] "    Farewell.                                       Exit Banquo."
## [22] "                         Exeunt all but Macbeth and an Attendant."
## [23] "  MACBETH. Bring them before us.                  Exit Attendant."
## [24] "                                                  Exit Attendant."
## [25] "                                                Exeunt Murtherers."
## [26] "    If it find heaven, must find it out tonight.             Exit."
## [27] "  SERVANT. Madam, I will.                                    Exit."
## [28] "    So, prithee, go with me.                               Exeunt."
## [29] "                                                           Exeunt."
## [30] "                                                 Exit Murtherer."
## [31] "    Shall be the maws of kites.                      Exit Ghost."
## [32] "    Unreal mockery, hence!                           Exit Ghost."
## [33] "                           Exeunt all but Macbeth and Lady Macbeth."
## [34] "    We are yet but young in deed.                        Exeunt."
## [35] "    Sits in a foggy cloud and stays for me.                 Exit."
## [36] "                                                         Exeunt."
## [37] "                                                         Exeunt."
## [38] "    Come, bring me where they are.                       Exeunt."
## [39] "    I take my leave at once.                               Exit."
## [40] "    I dare abide no longer.                                Exit."
## [41] "                              Exit Lady Macduff, crying \"Murther!\""
## [42] "                                 Exeunt Murtherers, following her."
## [43] "  MALCOLM. I thank you, Doctor.                     Exit Doctor."
## [44] "    The night is long that never finds the day.          Exeunt."
## [45] "Exit."
## [46] "                                                         Exeunt."
## [47] "    Make we our march towards Birnam.           Exeunt marching."
## [48] "  MACBETH. Take thy face hence.                    Exit Servant."
## [49] "    Profit again should hardly draw me here.             Exeunt."
## [50] "                                                 Exeunt Marching."
## [51] "  SEYTON. It is the cry of women, my good lord.            Exit."
## [52] "    At least we'll die with harness on our back.         Exeunt."
## [53] "                                                         Exeunt."
## [54] "    Brandish'd by man that's of a woman born.              Exit."
## [55] "    And more I beg not.                           Exit. Alarums."
## [56] "                                                 Exeunt. Alarum."
## [57] "                                       Exeunt fighting. Alarums."
## [58] "                                               Flourish. Exeunt."
```

3-2

(a)

```
babynames %>%
  filter(sex == 'M', str_detect(babynames$name, "(a|e|i|o|u)$") == TRUE) %>%
  group_by(name, sex) %>%
  summarise(count = sum(n)) %>%
  ungroup() %>%
  slice_max(count, n = 10)
```

```
## `summarise()` has grouped output by 'name'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 10 x 3
##    name     sex     count
##    <chr>    <chr>   <int>
##  1 George   M     1464186
##  2 Joshua   M     1202454
##  3 Jose     M      560679
##  4 Kyle     M      477768
##  5 Lawrence M      456773
##  6 Joe      M      450780
##  7 Willie   M      448702
##  8 Jesse    M      416530
##  9 Bruce    M      382257
## 10 Eugene   M      378539
```

(b)

```
babynames %>%
  filter(str_detect(babynames$name, "(joe|jo|Joe|Jo)$") == TRUE) %>%
  group_by(name) %>%
  summarise(count = sum(n)) %>%
  ungroup() %>%
  slice_max(count, n = 10)
```

```
## # A tibble: 10 x 2
##    name      count
##    <chr>     <int>
##  1 Joe      462099
##  2 Jo       180579
##  3 Maryjo     7017
##  4 Billiejo   1455
##  5 Marijo     1280
##  6 Bobbijo    1237
##  7 Bobbiejo   1009
##  8 Alejo       794
##  9 Bettyjo     764
## 10 Amyjo       486
```