SQL_script.R

mathe

Sat Jan 12 01:54:36 2019

```
# December 8

library(DBI)
```

```
## Warning: package 'DBI' was built under R version 3.4.4
```

```
library(RSQLite)
drv <- dbDriver("SQLite")
con <- dbConnect(drv, dbname="baseball.db")

dbListTables(con)
```

```
##  [1] "AllstarFull"         "Appearances"         "AwardsManagers"
##  [4] "AwardsPlayers"       "AwardsShareManagers" "AwardsSharePlayers"
##  [7] "Batting"             "BattingPost"         "Fielding"
## [10] "FieldingOF"          "FieldingPost"        "HallOfFame"
## [13] "Managers"            "ManagersHalf"        "Master"
## [16] "Pitching"            "PitchingPost"        "Salaries"
## [19] "Schools"             "SchoolsPlayers"      "SeriesPost"
## [22] "Teams"               "TeamsFranchises"     "TeamsHalf"
## [25] "sqlite_sequence"     "xref_stats"
```

```
dbListFields(con, "Batting")
```

```
##  [1] "playerID"  "yearID"    "stint"     "teamID"    "lgID"
##  [6] "G"         "G_batting" "AB"        "R"         "H"
## [11] "2B"        "3B"        "HR"        "RBI"       "SB"
## [16] "CS"        "BB"        "SO"        "IBB"       "HBP"
## [21] "SH"        "SF"        "GIDP"      "G_old"
```

```
dbListFields(con, "Pitching")
```

```
##  [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"     "W"
##  [7] "L"        "G"        "GS"       "CG"       "SHO"      "SV"
## [13] "IPouts"   "H"        "ER"       "HR"       "BB"       "SO"
## [19] "BAOpp"    "ERA"      "IBB"      "WP"       "HBP"      "BK"
## [25] "BFP"      "GF"       "R"
```

```
batting <- dbReadTable(con, "Batting")
class(batting)
```

```
## [1] "data.frame"
```

```
dim(batting)
```

```
## [1] 93955    24
```

```
# Check Yourself
library(plyr)

salaries <- dbReadTable(con, "Salaries")
```

```
my.sum.func <- function(team.yr.df) {
  return(sum(team.yr.df$salary))
}

payroll <- ddply(salaries, .(yearID, teamID), my.sum.func)
payroll <- payroll[payroll$yearID == 2010, ]
payroll <- payroll[order(payroll$V1, decreasing = TRUE), ]
payroll[1:3, ]
```

```
##     yearID teamID        V1
## 733   2010    NYA 206333389
## 719   2010    BOS 162447333
## 721   2010    CHN 146609000
```

```
payroll[nrow(payroll), ]
```

```
##     yearID teamID       V1
## 737   2010    PIT 34943000
```

```
# Querying

query <- paste("SELECT playerID, yearID, AB, H, HR",
               "FROM Batting LIMIT 10")
query
```

```
## [1] "SELECT playerID, yearID, AB, H, HR FROM Batting LIMIT 10"
```

```
dbGetQuery(con, query)
```

```
##     playerID yearID  AB   H HR
## 1  aardsda01   2004   0   0  0
## 2  aardsda01   2006   2   0  0
## 3  aardsda01   2007   0   0  0
## 4  aardsda01   2008   1   0  0
## 5  aardsda01   2009   0   0  0
## 6  aaronha01   1954 468 131 13
## 7  aaronha01   1955 602 189 27
## 8  aaronha01   1956 609 200 26
## 9  aaronha01   1957 615 198 44
## 10 aaronha01   1958 601 196 30
```

```
batting[1:10, c("playerID", "yearID", "AB", "H", "HR")]
```

```
##     playerID yearID  AB   H HR
## 1  aardsda01   2004   0   0  0
## 2  aardsda01   2006   2   0  0
## 3  aardsda01   2007   0   0  0
## 4  aardsda01   2008   1   0  0
## 5  aardsda01   2009   0   0  0
## 6  aaronha01   1954 468 131 13
## 7  aaronha01   1955 602 189 27
## 8  aaronha01   1956 609 200 26
## 9  aaronha01   1957 615 198 44
## 10 aaronha01   1958 601 196 30
```

```
query <- paste("SELECT playerID, yearID, AB, H, HR",
               "FROM Batting",
```

```
               "ORDER BY HR DESC",
               "LIMIT 10")
query
```

```
## [1] "SELECT playerID, yearID, AB, H, HR FROM Batting ORDER BY HR DESC LIMIT 10"
```

```
dbGetQuery(con, query)
```

```
##     playerID yearID  AB   H HR
## 1  bondsba01   2001 476 156 73
## 2  mcgwima01   1998 509 152 70
## 3   sosasa01   1998 643 198 66
## 4  mcgwima01   1999 521 145 65
## 5   sosasa01   2001 577 189 64
## 6   sosasa01   1999 625 180 63
## 7  marisro01   1961 590 159 61
## 8   ruthba01   1927 540 192 60
## 9   ruthba01   1921 540 204 59
## 10  foxxji01   1932 585 213 58
```

```
# Check Yourself
query1 <- paste("SELECT playerID, yearID, AB, H, HR",
               "FROM Batting",
               "WHERE yearID >= 1990 AND yearID <= 2000",
               "ORDER BY HR DESC",
               "LIMIT 10")
dbGetQuery(con, query1)
```

```
##     playerID yearID  AB   H HR
## 1  mcgwima01   1998 509 152 70
## 2   sosasa01   1998 643 198 66
## 3  mcgwima01   1999 521 145 65
## 4   sosasa01   1999 625 180 63
## 5  griffke02   1997 608 185 56
## 6  griffke02   1998 633 180 56
## 7  mcgwima01   1996 423 132 52
## 8  fieldce01   1990 573 159 51
## 9  anderbr01   1996 579 172 50
## 10 belleal01   1995 546 173 50
```

```
bat.ord <- batting[order(batting$HR, decreasing = TRUE), ]
subset  <- bat.ord$yearID >= 1990 & bat.ord$yearID <= 2000
columns <- c("playerID", "yearID", "AB", "H", "HR")

head(bat.ord[subset, columns], 10)
```

```
##        playerID yearID  AB   H HR
## 54613 mcgwima01   1998 509 152 70
## 78578  sosasa01   1998 643 198 66
## 54614 mcgwima01   1999 521 145 65
## 78579  sosasa01   1999 625 180 63
## 31877 griffke02   1997 608 185 56
## 31878 griffke02   1998 633 180 56
## 54610 mcgwima01   1996 423 132 52
## 25517 fieldce01   1990 573 159 51
## 1575  anderbr01   1996 579 172 50
```

```
## 5124  belleal01   1995 546 173 50
```

```
query2 <- paste("SELECT playerID, yearID, MAX(HR)",
                "FROM Batting")
dbGetQuery(con, query2)
```

```
##    playerID yearID MAX(HR)
## 1 bondsba01   2001      73
```

```
batting[which.max(batting$HR), c("playerID", "yearID", "HR")]
```

```
##       playerID yearID HR
## 7514 bondsba01   2001 73
```

```
# Computations
query <- paste("SELECT AVG(HR), AVG(H)",
                "FROM Batting")
dbGetQuery(con, query)
```

```
##    AVG(HR)   AVG(H)
## 1 2.970549 40.67684
```

```
mean(batting$HR, na.rm = TRUE)
```

```
## [1] 2.970549
```

```
query <- paste("SELECT playerID, AVG(HR)",
               "FROM Batting",
               "GROUP BY playerID",
               "ORDER BY AVG(HR) DESC",
               "LIMIT 5")
dbGetQuery(con, query)
```

```
##    playerID  AVG(HR)
## 1 pujolal01 40.80000
## 2 howarry01 36.14286
## 3 rodrial01 36.05882
## 4 bondsba01 34.63636
## 5 mcgwima01 34.29412
```

```
query <- paste("SELECT playerID, AVG(HR)",
               "FROM Batting",
               "WHERE yearID >= 1990",
               "GROUP BY playerID",
               "ORDER BY AVG(HR) DESC",
               "LIMIT 5")
dbGetQuery(con, query)
```

```
##    playerID  AVG(HR)
## 1 pujolal01 40.80000
## 2 bondsba01 37.66667
## 3 howarry01 36.14286
## 4 rodrial01 36.05882
## 5 mcgwima01 35.84615
```

```
# Check YourSelf

query <- paste("SELECT teamID, AVG(HR)",
               "FROM Batting",
```

```
               "WHERE yearID >= 1990",
               "GROUP BY teamID",
               "ORDER BY AVG(HR) DESC",
               "LIMIT 5")
dbGetQuery(con, query)
```

```
##   teamID  AVG(HR)
## 1    CHA 6.164251
## 2    NYA 5.986486
## 3    TOR 5.760937
## 4    CAL 5.625731
## 5    TEX 5.563961
```

```
bat.sub <- batting[batting$yearID >= 1990, ]

my.mean.func <- function(team.df) {
  return(mean(team.df$HR, na.rm = TRUE))
}
avg.hrs <- daply(bat.sub, .(teamID), my.mean.func)
avg.hrs <- sort(avg.hrs, decreasing = TRUE)
head(avg.hrs, 5)
```

```
##      CHA      NYA      TOR      CAL      TEX
## 6.164251 5.986486 5.760937 5.625731 5.563961
```

```
query <- paste("SELECT teamID, AVG(HR) as avgHR",
               "FROM Batting",
               "WHERE yearID >= 1990",
               "GROUP BY teamID",
               "ORDER BY AVG(HR) DESC",
               "LIMIT 5")
dbGetQuery(con, query)
```

```
##   teamID    avgHR
## 1    CHA 6.164251
## 2    NYA 5.986486
## 3    TOR 5.760937
## 4    CAL 5.625731
## 5    TEX 5.563961
```

```
query <- paste("SELECT teamID, AVG(HR) as avgHR",
               "FROM Batting",
               "WHERE yearID >= 1990",
               "GROUP BY teamID",
               "HAVING avgHR >= 4.5",
               "ORDER BY avgHR DESC")
dbGetQuery(con, query)
```

```
##    teamID    avgHR
## 1     CHA 6.164251
## 2     NYA 5.986486
## 3     TOR 5.760937
## 4     CAL 5.625731
## 5     TEX 5.563961
## 6     DET 5.531437
## 7     CLE 5.370262
```

```
## 8     BAL 5.152174
## 9     BOS 5.126227
## 10    SEA 5.027299
## 11    OAK 5.023677
## 12    ML4 4.834146
## 13    ANA 4.678445
```

```
# Check Yourself

query <- paste("SELECT teamID, SUM(salary) as SUMsal",
               "FROM Salaries",
               "WHERE yearID == 2010",
               "GROUP BY teamID",
               "ORDER BY SUMsal DESC",
               "LIMIT 3")
dbGetQuery(con, query)
```

```
##   teamID    SUMsal
## 1    NYA 206333389
## 2    BOS 162447333
## 3    CHN 146609000
```

```
# JOIN
query <- paste("SELECT *",
               "FROM Salaries",
               "ORDER BY PlayerID",
               "LIMIT 8")
dbGetQuery(con, query)
```

```
##   yearID teamID lgID  playerID  salary
## 1   2004    SFN   NL aardsda01  300000
## 2   2007    CHA   AL aardsda01  387500
## 3   2008    BOS   AL aardsda01  403250
## 4   2009    SEA   AL aardsda01  419000
## 5   2010    SEA   AL aardsda01 2750000
## 6   1986    BAL   AL  aasedo01  600000
## 7   1987    BAL   AL  aasedo01  625000
## 8   1988    BAL   AL  aasedo01  675000
```

```
query <- paste("SELECT yearID, teamID, lgID, playerID, HR",
               "FROM Batting",
               "ORDER BY PlayerID",
               "LIMIT 8")
dbGetQuery(con, query)
```

```
##   yearID teamID lgID  playerID HR
## 1   2004    SFN   NL aardsda01  0
## 2   2006    CHN   NL aardsda01  0
## 3   2007    CHA   AL aardsda01  0
## 4   2008    BOS   AL aardsda01  0
## 5   2009    SEA   AL aardsda01  0
## 6   2010    SEA   AL aardsda01  0
## 7   1954    ML1   NL aaronha01 13
## 8   1955    ML1   NL aaronha01 27
```

```
query <- paste("SELECT yearID, playerID, salary, HR",
               "FROM Batting JOIN Salaries USING(yearID, playerID)",
```

```
               "ORDER BY PlayerID",
               "LIMIT 8")
dbGetQuery(con, query)
```

```
##   yearID  playerID  salary HR
## 1   2004 aardsda01  300000  0
## 2   2007 aardsda01  387500  0
## 3   2008 aardsda01  403250  0
## 4   2009 aardsda01  419000  0
## 5   2010 aardsda01 2750000  0
## 6   1986  aasedo01  600000 NA
## 7   1987  aasedo01  625000 NA
## 8   1988  aasedo01  675000 NA
```

```
merged <- merge(x = batting, y = salaries,
                by.x = c("yearID", "playerID"), by.y = c("yearID", "playerID"))
names <- c("yearID", "playerID", "salary", "HR")
merged[order(merged$playerID), names][1:8, ]
```

```
##       yearID  playerID  salary HR
## 16708   2004 aardsda01  300000  0
## 19378   2007 aardsda01  387500  0
## 20277   2008 aardsda01  403250  0
## 21164   2009 aardsda01  419000  0
## 21990   2010 aardsda01 2750000  0
## 585     1986  aasedo01  600000 NA
## 1360    1987  aasedo01  625000 NA
## 2033    1988  aasedo01  675000 NA
```

```
query <- paste("SELECT yearID, playerID, salary, HR",
               "FROM Batting LEFT JOIN Salaries USING(yearID, playerID)",
               "ORDER BY PlayerID",
               "LIMIT 8")
dbGetQuery(con, query)
```

```
##   yearID  playerID  salary HR
## 1   2004 aardsda01  300000  0
## 2   2006 aardsda01      NA  0
## 3   2007 aardsda01  387500  0
## 4   2008 aardsda01  403250  0
## 5   2009 aardsda01  419000  0
## 6   2010 aardsda01 2750000  0
## 7   1954 aaronha01      NA 13
## 8   1955 aaronha01      NA 27
```

```
query <- paste("SELECT playerID, AVG(salary), AVG(HR)",
               "FROM Batting JOIN Salaries USING(yearID, playerID)",
               "GROUP BY playerID",
               "ORDER BY AVG(HR) DESC",
               "LIMIT 10")
dbGetQuery(con, query)
```

```
##     playerID AVG(salary)  AVG(HR)
## 1  howarry01   9051000.0 45.80000
## 2  pujolal01   8953204.1 40.80000
## 3  fieldpr01   3882900.0 38.00000
```

```
## 4  rodrial01  15553897.2 36.05882
## 5  reynoma01    550777.7 34.66667
## 6  bondsba01   8556605.5 34.63636
## 7  mcgwima01   4814020.8 34.29412
## 8  gonzaca01    406000.0 34.00000
## 9   dunnad01   6969500.0 33.50000
## 10 kingmda01    908750.0 32.50000
```

```
# Check Yourself


query <- paste("SELECT playerID, yearID, E",
               "FROM Fielding",
               "WHERE yearID >= 1990",
               "ORDER BY E DESC",
               "LIMIT 10")
dbGetQuery(con, query)
```

```
##     playerID yearID  E
## 1  offerjo01   1992 42
## 2  offerjo01   1993 37
## 3  valenjo03   1996 37
## 4  valenjo03   2000 36
## 5  carusmi01   1998 35
## 6  offerjo01   1995 35
## 7  reynoma01   2008 34
## 8  desmoia01   2010 34
## 9  cordewi01   1993 33
## 10 glaustr01   2000 33
```

```
query <- paste("SELECT playerID, yearID, E, salary",
               "FROM Fielding LEFT JOIN Salaries USING(yearID, playerID)",
               "WHERE yearID >= 1990",
               "ORDER BY E DESC",
               "LIMIT 10")
dbGetQuery(con, query)
```

```
##     playerID yearID  E  salary
## 1  offerjo01   1992 42  135000
## 2  offerjo01   1993 37  300000
## 3  valenjo03   1996 37  300000
## 4  valenjo03   2000 36 1320000
## 5  carusmi01   1998 35  170000
## 6  offerjo01   1995 35 1600000
## 7  reynoma01   2008 34  396500
## 8  desmoia01   2010 34  400000
## 9  cordewi01   1993 33  126500
## 10 glaustr01   2000 33  275000
```