Into the Tidyverse

2023-06-06

```
Batting <- as_tibble(Batting)
# tibble = Tidyverse version of data frame
```

```
dim(Batting)
```

```
## [1] 112184     22
```

```
head(Batting, 10)
```

```
## # A tibble: 10 x 22
##    playerID  yearID stint teamID lgID      G    AB     R     H   X2B   X3B    HR
##    <chr>      <int> <int> <fct>  <fct> <int> <int> <int> <int> <int> <int> <int>
##  1 abercda01   1871     1 TRO    NA        1     4     0     0     0     0     0
##  2 addybo01    1871     1 RC1    NA       25   118    30    32     6     0     0
##  3 allisar01   1871     1 CL1    NA       29   137    28    40     4     5     0
##  4 allisdo01   1871     1 WS3    NA       27   133    28    44    10     2     2
##  5 ansonca01   1871     1 RC1    NA       25   120    29    39    11     3     0
##  6 armstbo01   1871     1 FW1    NA       12    49     9    11     2     1     0
##  7 barkeal01   1871     1 RC1    NA        1     4     0     1     0     0     0
##  8 barnero01   1871     1 BS1    NA       31   157    66    63    10     9     0
##  9 barrebi01   1871     1 FW1    NA        1     5     1     1     1     0     0
## 10 barrofr01   1871     1 BS1    NA       18    86    13    13     2     1     0
## # i 10 more variables: RBI <int>, SB <int>, CS <int>, BB <int>, SO <int>,
## #   IBB <int>, HBP <int>, SH <int>, SF <int>, GIDP <int>
```

```
names(Batting)
```

```
##  [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"     "G"
##  [7] "AB"       "R"        "H"        "X2B"      "X3B"      "HR"
## [13] "RBI"      "SB"       "CS"       "BB"       "SO"       "IBB"
## [19] "HBP"      "SH"       "SF"       "GIDP"
```

```
str(Batting)
```

```
## tibble [112,184 x 22] (S3: tbl_df/tbl/data.frame)
##  $ playerID: chr [1:112184] "abercda01" "addybo01" "allisar01" "allisdo01" ...
##  $ yearID  : int [1:112184] 1871 1871 1871 1871 1871 1871 1871 1871 1871 1871 ...
##  $ stint   : int [1:112184] 1 1 1 1 1 1 1 1 1 1 ...
##  $ teamID  : Factor w/ 149 levels "ALT","ANA","ARI",..: 136 111 39 142 111 56 111 24 56 24 ...
##  $ lgID    : Factor w/ 7 levels "AA","AL","FL",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ G       : int [1:112184] 1 25 29 27 25 12 1 31 1 18 ...
##  $ AB      : int [1:112184] 4 118 137 133 120 49 4 157 5 86 ...
##  $ R       : int [1:112184] 0 30 28 28 29 9 0 66 1 13 ...
##  $ H       : int [1:112184] 0 32 40 44 39 11 1 63 1 13 ...
##  $ X2B     : int [1:112184] 0 6 4 10 11 2 0 10 1 2 ...
##  $ X3B     : int [1:112184] 0 0 5 2 3 1 0 9 0 1 ...
##  $ HR      : int [1:112184] 0 0 0 2 0 0 0 0 0 0 ...
```

```
##  $ RBI     : int [1:112184] 0 13 19 27 16 5 2 34 1 11 ...
##  $ SB      : int [1:112184] 0 8 3 1 6 0 0 11 0 1 ...
##  $ CS      : int [1:112184] 0 1 1 1 2 1 0 6 0 0 ...
##  $ BB      : int [1:112184] 0 4 2 0 2 0 1 13 0 0 ...
##  $ SO      : int [1:112184] 0 0 5 2 1 1 0 1 0 0 ...
##  $ IBB     : int [1:112184] NA NA NA NA NA NA NA NA NA NA ...
##  $ HBP     : int [1:112184] NA NA NA NA NA NA NA NA NA NA ...
##  $ SH      : int [1:112184] NA NA NA NA NA NA NA NA NA NA ...
##  $ SF      : int [1:112184] NA NA NA NA NA NA NA NA NA NA ...
##  $ GIDP    : int [1:112184] 0 0 1 0 0 0 0 1 0 0 ...
```

```
summary(Batting$yearID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1871    1938    1978    1969    2003    2022
```

```
table("Leagues" = Batting$lgID)
```

```
## Leagues
##    AA    AL    FL    NA    NL    PL    UA
##  1893 51799   472   737 56800   149   334
```

```
#table(Batting$lgID, Batting$teamID)
```

```
mlb_batting <- Batting %>%
  filter(lgID %in% c("AL", "NL"))
```

```
sel_batting <- Batting %>%
  select(yearID, G, AB, R, H)
```

```
hr_batting <- Batting %>%
  arrange(desc(HR))
```

```
summarize(Batting, max(stint), median(AB))
```

```
## # A tibble: 1 x 2
##   `max(stint)` `median(AB)`
##          <int>        <dbl>
## 1            5           45
```

```
Batting %>%
  arrange(desc(stint))
```

```
## # A tibble: 112,184 x 22
##    playerID  yearID stint teamID lgID      G    AB     R     H   X2B   X3B    HR
##    <chr>      <int> <int> <fct>  <fct> <int> <int> <int> <int> <int> <int> <int>
##  1 dowseto01   1892     5 WAS    NL        6    23     5     6     1     0     0
##  2 huelsfr01   1904     5 WS1    AL       84   303    21    75    19     4     2
##  3 chouife01   1914     5 BRF    FL       26    70     4    15     0     2     0
##  4 drakeol01   2018     5 MIN    AL       19     0     0     0     0     0     0
##  5 striege01   1884     4 CL2    NL        8    29     2     7     2     0     0
##  6 wheelha01   1884     4 BLU    UA       17    69     3    18     2     0     0
##  7 dowseto01   1892     4 PHI    NL       16    54     3    10     0     0     0
##  8 kuehnbi01   1892     4 SLN    NL        1     4     0     0     0     0     0
##  9 ohageha01   1902     4 NY1    NL       22    73     5    11     2     1     0
## 10 huelsfr01   1904     4 SLA    AL       20    68     6    15     2     1     0
## # i 112,174 more rows
## # i 10 more variables: RBI <int>, SB <int>, CS <int>, BB <int>, SO <int>,
```

```
## #   IBB <int>, HBP <int>, SH <int>, SF <int>, GIDP <int>
```

```
new_batting <- Batting %>%
  mutate(batting_avg = H/AB)

head(new_batting)
```

```
## # A tibble: 6 x 23
##   playerID  yearID stint teamID lgID      G    AB     R     H   X2B   X3B    HR
##   <chr>      <int> <int> <fct>  <fct> <int> <int> <int> <int> <int> <int> <int>
## 1 abercda01   1871     1 TRO    NA        1     4     0     0     0     0     0
## 2 addybo01    1871     1 RC1    NA       25   118    30    32     6     0     0
## 3 allisar01   1871     1 CL1    NA       29   137    28    40     4     5     0
## 4 allisdo01   1871     1 WS3    NA       27   133    28    44    10     2     2
## 5 ansonca01   1871     1 RC1    NA       25   120    29    39    11     3     0
## 6 armstbo01   1871     1 FW1    NA       12    49     9    11     2     1     0
## # i 11 more variables: RBI <int>, SB <int>, CS <int>, BB <int>, SO <int>,
## #   IBB <int>, HBP <int>, SH <int>, SF <int>, GIDP <int>, batting_avg <dbl>
```

```
## command + shift + m for shortcut for pipe
new_batting %>%
  select(batting_avg, playerID) %>%
  arrange(desc(batting_avg)) %>%
  head()
```

```
## # A tibble: 6 x 2
##   batting_avg playerID
##         <dbl> <chr>
## 1           1 snowch01
## 2           1 baldwki01
## 3           1 mccafsp01
## 4           1 gumbebi01
## 5           1 oconnfr01
## 6           1 brownpe01
```

```
Batting %>%
  filter(lgID %in% c("AL", "NL"),
         AB > 300) %>%
  mutate(BA = H/AB) %>%
  arrange(desc(BA)) %>%
  select(playerID, yearID, BA) %>%
  head(n = 5)
```

```
## # A tibble: 5 x 3
##   playerID  yearID    BA
##   <chr>      <int> <dbl>
## 1 duffyhu01   1894 0.440
## 2 barnero01   1876 0.429
## 3 lajoina01   1901 0.426
## 4 keelewi01   1897 0.424
## 5 hornsro01   1924 0.424
```

```
Batting %>%
  filter(lgID %in% c("AL", "NL"),
         AB > 300) %>%
  mutate(so_to_bb = SO/BB) %>%
  arrange(desc(so_to_bb)) %>%
```

```
  select(playerID, yearID, so_to_bb) %>%
  slice(c(1,2,10,100))
```

```
## # A tibble: 4 x 3
##   playerID  yearID so_to_bb
##   <chr>      <int>    <dbl>
## 1 galvipu01   1883    26.3
## 2 flintsi01   1882    25
## 3 meinkfr01   1884    14.8
## 4 odorro01    2016     7.11
```

```
Batting %>%
  filter(lgID %in% c("AL", "NL")) %>%
  group_by(yearID) %>%
  summarize(tot_hr = sum(HR), tot_so = sum(SO), tot_bb = sum(BB)) %>%
  arrange(desc(tot_hr)) %>%
  slice(1:5)
```

```
## # A tibble: 5 x 4
##   yearID tot_hr tot_so tot_bb
##    <int>  <int>  <int>  <int>
## 1   2019   6776  42823  15895
## 2   2017   6105  40104  15829
## 3   2021   5944  42145  15794
## 4   2000   5693  31356  18237
## 5   2016   5610  38982  15088
```

```
year_batting_summary <- Batting %>%
  filter(lgID %in% c("AL", "NL")) %>%
  group_by(yearID) %>%
  summarize(total_hits = sum(H, na.rm = TRUE), #removes missing values
            total_hr = sum(HR, na.rm = TRUE),
            total_so = sum(SO, na.rm = TRUE),
            total_walks = sum(BB, na.rm = TRUE),
            total_at_bats = sum(AB, na.rm = TRUE)) %>%
  mutate(overall_batting_avg = total_hits/total_at_bats)

head(year_batting_summary)
```

```
## # A tibble: 6 x 7
##   yearID total_hits total_hr total_so total_walks total_at_bats
##    <int>      <int>    <int>    <int>       <int>         <int>
## 1   1876       5338       40      589         336         20121
## 2   1877       3705       24      726         345         13667
## 3   1878       3539       23     1081         364         13644
## 4   1879       6171       58     1843         508         24155
## 5   1880       5946       62     1993         740         24301
## 6   1881       6339       76     1784        1033         24377
## # i 1 more variable: overall_batting_avg <dbl>
```

```
year_batting_summary %>%
  arrange(desc(total_hr)) %>%
  slice(1:3)
```

```
## # A tibble: 3 x 7
##   yearID total_hits total_hr total_so total_walks total_at_bats
```

```
##    <int>      <int>    <int>    <int>       <int>         <int>
## 1   2019      42039     6776    42823       15895        166651
## 2   2017      42215     6105    40104       15829        165567
## 3   2021      39484     5944    42145       15794        161941
## # i 1 more variable: overall_batting_avg <dbl>
```

```
year_batting_summary %>%
  select(yearID, overall_batting_avg) %>%
  rename(Year = yearID, `Overall Batting Average` = overall_batting_avg) %>%
  slice(c(1, n())) %>%   #n() gives you last row in data frame
  gt()
```

Year	Overall Batting Average
1876	0.2652950
2022	0.2427125