CSC/ST 442 (Fall 2021): Assignment 3

Instruction

This assignment consists of 2 problems. The assignment is due on **Monday, September 23** at 11:59pm EDT. Please submit your assignment electronically through **Moodle**. You are encouraged (but not required) to use RMarkdown to write up your homework solution. To start using Rmarkdown read

- Section 40.2 of Introduction to Data Science
- the RStudio tutorial
- the Rmarkdown cheatsheet.

Problem 1

Install the dslabs library to get access to the **admissions** data frame described below.

```
library("dslabs") ## Do a install.packages("dslabs") if the dslabs library is not yet installed.
admissions
```

```
## # A tibble: 12 x 4
##   major gender admitted applicants
##   <chr> <chr>     <dbl>      <dbl>
## 1 A     men          62        825
## 2 B     men          63        560
## 3 C     men          37        325
## 4 D     men          33        417
## 5 E     men          28        191
## 6 F     men           6        373
## # ... with 6 more rows
```

This data frame describe the number of people who applied and who was admitted into several undergraduate majors. Transform this data frame into the following form.

```
## # A tibble: 6 x 5
##   major admitted_men applicants_men admitted_women applicants_women
##   <chr>        <dbl>          <dbl>          <dbl>            <dbl>
## 1 A               62            825             82              108
## 2 B               63            560             68               25
## 3 C               37            325             34              593
## 4 D               33            417             35              375
## 5 E               28            191             24              393
## 6 F                6            373              7              341
```

Hint You can proceed via the following steps

- First use **pivot_longer** from the tidyr library to transform the **admissions** data frame into the following form

```
## # A tibble: 24 x 4
##   major gender name       value
##   <chr> <chr>  <chr>      <dbl>
## 1 A     men    admitted      62
```

```
## 2 A     men    applicants   825
## 3 B     men    admitted      63
## 4 B     men    applicants   560
## 5 C     men    admitted      37
## 6 C     men    applicants   325
## # ... with 18 more rows
```

- Next use **unite** (also from the tidyr library) to combine the columns for `name` and `gender` into one column.
- Finally use **pivot_wider** from the **tidyr** library to get the required data frame.

Problem 2

This problem is from Chapter 5 of the book Modern Data Science with R. The problem uses the `Batting`, `Pitching`, and `Master` data frames in the Lahman package.

```
library(Lahman) ## install.packages("Lahman") if the library is not yet installed.
Batting
```

```
## # A tibble: 108,789 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
## # ... with 108,785 more rows, and 10 more variables: RBI <int>, SB <int>,
## #   CS <int>, BB <int>, SO <int>, IBB <int>, HBP <int>, SH <int>, SF <int>,
## #   GIDP <int>
```

```
Pitching
```

```
## # A tibble: 48,399 x 30
##   playerID  yearID stint teamID lgID      W     L     G    GS    CG   SHO    SV
##   <chr>      <int> <int> <fct>  <fct> <int> <int> <int> <int> <int> <int> <int>
## 1 bechtge01   1871     1 PH1    NA        1     2     3     3     2     0     0
## 2 brainas01   1871     1 WS3    NA       12    15    30    30    30     0     0
## 3 fergubo01   1871     1 NY2    NA        0     0     1     0     0     0     0
## 4 fishech01   1871     1 RC1    NA        4    16    24    24    22     1     0
## # ... with 48,395 more rows, and 18 more variables: IPouts <int>, H <int>,
## #   ER <int>, HR <int>, BB <int>, SO <int>, BAOpp <dbl>, ERA <dbl>, IBB <int>,
## #   WP <int>, HBP <int>, BK <int>, BFP <int>, GF <int>, R <int>, SH <int>,
## #   SF <int>, GIDP <int>
```

```
Master
```

```
## # A tibble: 20,093 x 26
##   playerID  birthYear birthMonth birthDay birthCountry birthState birthCity
##   <chr>         <int>      <int>    <int> <chr>        <chr>      <chr>
## 1 aardsda01      1981         12       27 USA          CO         Denver
## 2 aaronha01      1934          2        5 USA          AL         Mobile
## 3 aaronto01      1939          8        5 USA          AL         Mobile
## 4 aasedo01       1954          9        8 USA          CA         Orange
## # ... with 20,089 more rows, and 19 more variables: deathYear <int>,
## #   deathMonth <int>, deathDay <int>, deathCountry <chr>, deathState <chr>,
## #   deathCity <chr>, nameFirst <chr>, nameLast <chr>, nameGiven <chr>,
## #   weight <int>, height <int>, bats <fct>, throws <fct>, debut <chr>,
```

```
## #   finalGame <chr>, retroID <chr>, bbrefID <chr>, deathDate <date>,
## #   birthDate <date>
```

Using the above data frames, answer the following questions.

- Name every player in baseball history who has accumulated at least 300 home runs (`HR` column) and at least 300 stolen bases (`SB` column). You can find the first and last name of the player in the `Master` data frame. Join this to your result along with the total home runs and total bases stolen for each of these elite players.
- Similarly, name every pitcher in baseball history who has accumulated at least 300 wins (`W` column) and at least $3,000$ strikeouts (`SO` column).
- Identify the name and year of every player who has hit at least 50 home runs in a single season. Let `table1` refer to the data frame that contains this information. For each season that appeared in the data frame `table1`, find the player that has the lowest batting average that season. Hint: Use a `semi_join`