MATH113/CAIS105: Intro to Data Science

Fall 2023

Homework 05

Homework is DUE before class on the day indicated on the course schedule.

Learning Objectives:

- Practice making Tidy data

Overview

Complete this assignment in an R Markdown file.

When you answer the questions below, be sure to include your code *and* a written answer in your R Markdown file. For example, if I were answering the question: "What were the most popular baby names in the 1990s", my R Markdown report would look something like:

```
babynames %>%
  filter(year >= 1990 & year < 2000) %>%
  group_by(name) %>%
  summarize(num_births = sum(n)) %>%
  arrange(desc(num_births))
```

```
## # A tibble: 45,928 x 2
##    name        num_births
##    <chr>            <int>
##  1 Michael         464249
##  2 Christopher     361251
##  3 Matthew         352341
##  4 Joshua          330046
##  5 Jessica         303854
##  6 Ashley          303125
##  7 Jacob           298926
##  8 Nicholas        275906
##  9 Andrew          273515
## 10 Daniel          273347
## # … with 45,918 more rows
```

The most popular baby names from the 1990s were Michael, Christopher, and Matthew.

Part 1

The `HELPfull` data within the `mosaicData` package contains information about the Health Evaluation and Linkage to Primary Care (HELP) randomized trial in *tall* format.

a. Generate a table of the data for subjects (`ID`) 1, 2, and 3 that includes the `ID` variable, the `TIME` variable, and the `DRUGRISK` and `SEXRISK` variables (measures of drug and sex risk-taking behaviors, respectively).

b. The HELP trial was designed to collect information at 0, 6, 12, 18, and 24 month intervals. At which timepoints were measurements available on the `*RISK` variables for subject 3?

c. Let's restrict our attention to the data from the baseline (`TIME = 0`) and 6-month data. Create a table that looks like the following:

```
# A tibble: 3 × 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
```

Part 2

a) Consider the number of home runs hit (HR) and home runs allowed (HRA) for the Chicago Cubs (CHN) baseball team. Reshape the `Teams` data from the `Lahman` package into "long" format and plot a time series conditioned on whether the HRs that involved the Cubs were hit by them or allowed by them.

b) Write a function called `count_seasons` that, when given a `teamID`, will count the number of seasons the team played in the `Teams` data frame from the `Lahman` package

Submission

Knit your R Markdown file to a PDF and submit through PLATO.