MATH113/CAIS105: Intro to Data Science

Fall 2023

Homework 04

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

Learning Objectives:

- Practice data wrangling and analysis of 1 table in R

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

Answer the following questions using the `nycflights13` package:

1. How many planes have a missing date of manufacture?
2. What are the five most common manufacturers?
3. Has the distribution of manufacturer changed over time as reflected by the airplanes flying from NYC in 2013? (Hint: you may need to use `case_when()` to recode the manufacturer name and collapse rare vendors into a category called `Other`.)

Part 2

Install and load the Lahman package to answer the following questions. You will use the Batting, Pitching, and People tables in this package.

1. Name every player in baseball history who has accumulated at least 300 home runs (HR) and at least 300 stolen bases (SB). 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.
2. Similarly, name every pitcher in baseball history who has accumulated at least 300 wins (W) and at least 3,000 strikeouts (SO).
3. Identify the name and year of every player who has hit at least 50 home runs in a single season. Which player had the lowest batting average in that season?

Submission

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