# Homework 4

MTH 3270 Data Science

Due Fri., Feb. 25

| **Read These Chapters of the Book** | **Then Do These Exercises** |
|---|---|
| 4 | Problems 1-3 (**below**), Problem 6 (**Ch 4**), Problem 9 (**Ch 4**), Problem 14 (**Ch 4**)* |

* For **Problem 14**, instead of plotting the number of trips per *week* over the year, you may plot the number of trips per *month*.

**1** Use `filter()` with the `flights` data from the `"nycflights13"` package (and the logical operators '`&`', '`|`', and '`!`') to find all flights that:

a) Arrived more than two hours late but didn't leave late. Report your R command(s).

b) Were delayed by at least an hour, but made up over 30 minutes during flight. Report your R command(s).

**2** Use `arrange()` to sort the `flights` data (from the `"nycflights13"` package) to:

a) Find the fastest `flights` (i.e. the ones that spent the least time in the air). Report your R command(s).

b) Find the longest `flights` (i.e. the ones that spent the most time in the air). Report your R command(s).

**3** This problem uses the **nels88.txt** data set from the course website in Canvas.

**Data Set:** `nels88`

The National Educational Longitudinal Study data set is in the file **nels88.txt**. It is a nationally representative, longitudinal study of 8th graders in 1988 who were followed throughout secondary and postsecondary years.

It included surveys of students reporting on school, work, home experiences, educational resources and support, the role in education of parents and peers, neighborhood characteristics, educational and occupational aspirations, and other student perceptions.

Student assessments were made in reading, social studies, mathematics, and science (8th, 10th, and 12th grades).

The data are from the National Center for Education Statistics,

`https://nces.ed.gov/surveys/nels88/`

It contains 20 variables:

| | |
|---|---|
| `id` | Student identifier/student id (unique sample member id) |
| `sch_id` | School public release id |
| `heldback` | 8th grader ever held back a grade |
| `schtype` | Eighth grade school type |
| `race` | Race |
| `ses` | Socio-economic status |
| `female` | Sex: female |
| `minority` | Student is language minority |
| `asian` | Asian/Pacific Islander race |
| `hispanic` | Hispanic race |
| `black` | Black race |
| `white` | White race |
| `native` | Native American race |
| `catholic` | Catholic religion |
| `private` | Private schooled |
| `bymath` | Base year (1988) mathematics standardized score |
| `f1math` | First follow-up (1990) mathematics standardized score |
| `f2math` | Second follow-up (1990) mathematics standardized score |
| `f2dropout` | Second follow-up (1990) dropout status |

Save the **nels88.txt** data file from the course website, and read it into R using `read.table()`

or `read.csv()`.

a) Use `filter()` (from the `"dplyr"` package) to extract a subset of the rows of the `nels88` data. As examples, you could extract rows corresponding to students who attended a particular school or type of school or who are of a particular race or gender. Report your R command(s).

b) Use `summarize()` (from `"dplyr"`) to compute a summary statistic for each of at least three variables in the `nels88` data. Be careful because some summary statistics (e.g. the *mean* and *standard deviation*) don't make sense for **categorical** variables, but others do (e.g. the *proportion* of observations that fall in a given category). Report your R command(s).

c) Use `mutate()` or `transmute()` (from `"dplyr"`) to compute at least one new variable from existing variables in the `nels88` data. Be careful – computations on **categorical** variables coded as 0 and 1 might not make sense. Instead, consider computing the new variable from **numerical** variables (e.g. math scores). Report your R command(s).