STAT 534: Homework 2

Fall 2021

Due: Wednesday, September 22, 12:30 pm

1. Use the *nycflights13* package to answer the following questions:
   (a) (Excerpted from Problem 4.9 in textbook) Use the *flights* table. What month had the highest proportion of cancelled flights (flights with missing departure/arrival delay time)? What month had the lowest? Interpret any seasonal patterns.
   (b) Challenge: Use the *weather* table. On how many days was there precipitation in the New York area for each month in 2013? What do you observe in combine with the results from part (a)? (Hint: the `distinct` function in *dplyr* can be useful.)
   (c) (Excerpted from Problem 5.4 in textbook) Use the *flights* and *plane* tables. What is the oldest plane (specified by the tailnum variable) that flew from New York City airports in 2013?
2. Use the *mosaicData* package to answer the following questions:
   (a) (Excerpted from Problem 6.6 in textbook) The *HELPfull* data contains information about the Health Evaluation and Linkage to Primary Care (HELP) randomized trial in tall format. Create a table that each row displays the DRUGRISK and SEXRISK scores at the baseline and 6 months for a subject ID. (Hint: See the textbook for breakdown steps.)
   (b) (Excerpted from Problem 7.3 in textbook) Use the `purrr::map()` function and the *HELPrct* data frame to fit a regression model predicting cesd as a function of age separately for each of the levels of the substance variable. Generate a list of results (estimates and confidence intervals) for the slope parameter for each level of the grouping variable. (Hint: The `tidy()` function with the option `conf.int = T` computes confidence intervals for a `lm()` object.)
3. Health Care Coverage Data
   (a) Read the dataset from CSVs hosted on GitHub using `read_csv()`. (Hint: you want to skip the first two lines with the `skip` option and read up to the 52nd state with the `n_max` option.)
   (b) Convert all year-based columns to integer using `mutate(across(···))`. (Hint: Read Section 7.2 in the textbook for the `across()` function.)
   (c) Further tidy the dataset and convert it to a long data format as shown below.

```
## # A tibble: 1,456 × 4
##    Location       year type         tot_coverage
##    <chr>         <int> <chr>              <int>
##  1 United States  2013 Employer        155696900
##  2 United States  2013 Non-Group        13816000
##  3 United States  2013 Medicaid         54919100
##  4 United States  2013 Medicare         40876300
##  5 United States  2013 Other Public      6295400
##  6 United States  2013 Uninsured        41795100
##  7 United States  2013 Total           313401200
##  8 United States  2014 Employer        154347500
##  9 United States  2014 Non-Group        19313000
## 10 United States  2014 Medicaid         61650400
## # … with 1,446 more rows
```

4. Challenge: The *Violations* data set in the *mdsr* package contains information regarding the outcome of health inspections of restaurants in New York City. The *ViolationCodes* data set includes violation

description and classification of critical violations (violations that are most likely to contribute to food-borne illness). See original data source for more information: NYC Open Data.
Use these data to calculate, by zip code, the average number of inspections per restaurant and the rate of critical violations. What pattern do you see between the average number of inspections per restaurant and the rate of critical violations?
(Hint: Note that an inspection can appear across several rows in the *Violations* data set if multiple violations were associated with the inspection. The rate of critical violations should be defined as the rate of inspections that result in at least one critical violation.)