BIOST 509: Homework 8 (Final Homework)

Instructor: Amy Willis, Biostatistics, UW

Due date: 5 p.m. Wednesday 4 December via Canvas

Instructions

This is the final "capstone" homework of this class. It is longer and more challenging than previous homeworks, so please start it early. Note that you have a bit more time than usual to complete it, though.

In this homework you will put into practice your knowledge on `for` loops, joining data, and function creation to accomplish a fairly sophisticated task.

Download the zipped folder "hw8" from "Files/Datasets" or "Module 8 Materials" on Canvas. Unzip the folder. This folder contains all the files that you will need for this assignment.

Suppose you work at the data coordinating center for a medical study. Every day, you receive new data from new patients.

The daily data consists of two .csv files. The first file contains the patients' personally identifiable information (PII). A generic file with the PII has the name "PII_MonthDay.csv"; for example "PII_Jan1.csv" contains all the PII for patients from January 1. The columns FirstName and LastName of these data files contain names recorded with different formats (names in upper/lower case) and there are different codes for what should be considered missing data (e.g., "NONAME", "" and ".").

The second set of files contains the variables that were collected for data analysis. A generic file with variables for statistical analysis has the name "AnalysisData_MonthDay.csv"; for example "AnalysisData_Jan1.csv" contains the variables for data analysis collected on the patients from January 1. Due to a glitch in the data collection, the variables `Gender`, `Age`, `Occupation`, and `Disease` contain the values -999 and -99 to denote missing data.

Questions

1. Read in the dataset "PII_Jan1.csv". Create a function to standardize a name column, so that all names are in lowercase (use `tolower`) and missing names are coded as `NA`. Use this function to standardize the `FirstName` and `LastName` columns of PII_Jan1. Convert the Postcodes to characters.
2. Read in the data "AnalysisData_Jan1.csv". Create a function to standardize a column, so that all codes for missing values are transformed to NA and so that the column is treated as a character string (or factor). Use this function standardize `Gender`, `Age`, `Occupation`, and `Disease` in AnalysisData_Jan1.
3. Merge the data files "PII_Jan1.csv" and "AnalysisData_Jan1.csv" according to patient ID.
4. *We now want to combine all the daily data into a single data frame containing all the patients of the month. We know that the set of patients is different each day.* Create a `for()` loop that goes over the 31 days, and for each day, automatically reads the corresponding PII and analysis data files, standardizes/cleans them, merges them by ID, and sequentially builds a data frame with all the data for the month. The final data frame containing the full month data should also contain a column that indicates the day corresponding to each patient.