STA 3100 Programming with Data: Assignment 040

(100 points)

Reading the Tidy Data Paper

The purpose of this assignment is to test your reading and understanding of the Tidy Data paper (Wickham 2014), hence forth referred to as "the paper". Since the paper was written, packages such as plyr and reshape2 have been superceded by the tidyverse meta-package, and in particular by the dplyr and tidyr packages, respectively. In this assignment you will answer questions about sections 2 and 3 of the paper and you will reproduce some of the examples **using functions from the modern tidyverse package**. (If you do this correctly, you will see that the modern tidyverse functions are generally easier to use and understand than the functions and methods used in the R script that accompanies the paper.)

You may find it helpful to consult the vignettes from the tidyr package, but you should think about each question and try to solve it on your own before looking at the vignettes for help. The exercises here will generally require something slightly different from what you will find in the vignettes (and in the paper), so be sure to read the questions carefully.

Note: Throughout this assignment, "data frame" should be taken to be synonymous with "tibble". For the purposes of this assignment, you will be working with tibbles.

Section 2: Defining Tidy Data

1. (4 pts) According to the paper, values in a dataset are organized in two ways: "Every value belongs to a *variable* and an *observation*. What definitions are given for the terms *variable* and *observation*? (Format your solution as a bullet list.)
2. (6 pts) List the 3 defining characteristics of tidy data?
3. (5 pts) Section 2 of the paper contains an hypothetical example in which there are 3 types of observational unit. Briefly describe the context (in 20 words or less) and then list the 3 types of observational unit in the example.

Section 3: Tidying Messy Datasets

4. (5 pts) List the five most common problems with messy datasets as given in the paper.

5. The Pew Forum dataset in Table 4 of the paper is included in the tidyr package with the name `relig_income`.

   (a) (5 pts) Which of the five most common problems with messy datasets is/are illustrated by the Pew dataset?

   (b) (10 pts) Convert `relig_income` to a data frame `pew` in the tidy form illustrated in Table 6 of the paper, but with the following differences:

   - Make `income` an ordered factor by *first* replacing "Don't know/refused" with `NA` and *then* converting `income` and ordered factor with levels `<$10k`, `$10-20k`, `$20-30k`, `$30-40k`, `$40-50k`, `$50-75k`, `$75-100k`, `$100-150k`, and `>150k`.
   - Arrange the rows in the data frame in alphabetical order of `religion` and in ascending order of `income` within `religion`.

   Display the first 10 rows of `pew`.

6. (5 pts) The Billboard dataset in Table 7 of the paper is included in the tidyr package with the name `billboard`. The `billboard` dataset does not include the year or the time columns shown in Table 7, so we will not use those columns.

   (a) (5 pts) Which of the five most common problems with messy datasets is/are illustrated by the Billboard dataset?

   (b) (10 pts) Using tidyverse and lubridate commands, create a data frame `bb` with the tidy form illustrated in Table 8 of the paper. You should end up with columns `artist`, `track`, `date`, `week`, and `rank`, in that order. The `week` column should be of type integer or double and the `date` column should of type date. Use `filter()` to remove those rows for which `rank` is missing (`NA`). Display the first 15 rows of `bb`.

7. A version of tuberculosis (TB) dataset in Table 9 of the paper is included in the tidyr package with the name `who`. However, the columns of the `who` dataset are named slightly differently from the columns of Table 9. The meaning of the column names is explained in the R help file for `who`. The values in the last 56 columns (`new_sp_m014:newrel_f65`) are case counts.

   (a) (5 pts) Which of the five most common problems with messy datasets is/are illustrated by the TB dataset?

   (b) (10 pts) One way in which the `who` dataset differs from the dataset in Table 9 is that the `who` dataset has three columns representing the variable country (`country`, `iso2`, and `iso3`). As discussed in the last paragraph of section 3.4, this kind of redundancy is often required (or at least very convenient) for data analysis, but it arguably violates one of the defining characteristics of tidy data by combining two types of observational unit (country and country-year) in a single table. Create a data frame `who_countries` with columns `country`, `iso2`, and `iso3` and no duplicated rows. Arrange the rows of `who_countries` in alphabetical order of `country`. Display the first 6 rows of `who_countries`. *Hint*: The dplyr function `distinct()` is useful here.

   (c) (10 pts) Using tidyverse functions, create a data frame `who_tb` similar to Table 10(b) of the paper and satisfying the following criteria:

   - the columns are `iso2`, `year`, `diagnosis` `sex`, `age`, and `cases` in that order.
   - `age` should be an ordered factor with levels `0-14`, `15-24`, `25-34`, `35-44`, `45-54`, `55-64`, and `65+`.
   - The rows for which `cases` is missing (`NA`) have been removed.
   - The rows are sorted
     - by `iso2`,
     - by `year` within `iso2`,
     - by `diagnosis` within `iso2` and `year`,
     - by `sex` within `iso2`, `year`, and `diagnosis`, and
     - by `age` within `iso2`, `year`, `diagnosis`, and `sex`.

   Display the first 15 rows of the `who_tb` data frame.

8. The file `weather.csv` contains a version of the weather dataset in Table 11 of the paper. In addition to minimum and maxumum temperatures, this version also includes measurements of precipitation (`prcp`).

   (a) (5 pts) Which of the five most common problems with messy datasets is/are illustrated by the weather dataset?

   (b) (15 pts) Read the data into R and convert them to the form seen in Table 12(b) of the paper, execpt with an additional column `prcp`. Arrange the rows of the the data frame in ascending order of date. Display the first 10 rows of the data frame. *Hint*: You may find that both `pivot_longer()` and `pivot_wider()` are useful here. The lubridate function `make_date()` maybe also be useful, but you should to filter out rows with missing dates (notice what happens if you run `make_date(year = 2010, month = 2, day = 31)` in R).

References

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (1): 1–23. https://www.jstatsoft.org/index.php/jss/article/view/v059i10.