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 Since the paper was written, packages such as plyr and reshape2 have been superceded by the tidyverse meta-package, and in particular by the dplyr and tidyr packages, respectively.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.)

- A variable contains all values that measure the same underlying attribute (like height, temperature, duration) across units
- An observation contains all values measured on the same unit (like a person, or a day, or a race) across attributes.

2. (6 pts) List the 3 defining characteristics of tidy data?

   In tidy data:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit is a table.

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.

The hypothetical example in the paper is a trial of a new allergy medication. The three types of observational units listed in the example are:

- Demographic data collected from each person(age, sex, race).
- Medical data collected from each person on each day (number of sneezes, redness of eyes).
- Meteorological data collected on each day (temperate, pollen count).

Section 3: Tidying Messy Datasets

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

The five most common problems with messy datasets given in the paper are: - Column headers are values, not variable names. - Multiple variables are stored in one column. - Variables are stored in both rows and columns. - Multiple types of observational units are stored in the same table. - A single observational unit is stored in multiple tables.

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?

The common problem illustrated in the Pew dataset is that the column headers are values and not variables names.

```
(b) (10 pts) Convert ‘relig_income‘ to a data frame ‘pew‘ in the tidy form illustrated in Table 6 of the

        - Make ‘income‘ an ordered factor by *first* replacing "Don’t know/refused" with ‘NA‘ and *then*
        - Arrange the rows in the data frame in alphabetical order of ‘religion‘ and in ascending order o

    Display the first 10 rows of ‘pew‘.
```

```
library(tidyr)
library(dplyr)
pew <- relig_income %>%
  rename("NA" = "Don't know/refused") %>%
  pivot_longer(!religion, names_to = "income", values_to = "freq") %>%
  arrange(religion)

pew %>% slice_head(n = 10)
```

```
## # A tibble: 10 x 3
##    religion income     freq
##    <chr>    <chr>     <dbl>
##  1 Agnostic <$10k        27
##  2 Agnostic $10-20k      34
##  3 Agnostic $20-30k      60
##  4 Agnostic $30-40k      81
##  5 Agnostic $40-50k      76
##  6 Agnostic $50-75k     137
##  7 Agnostic $75-100k    122
##  8 Agnostic $100-150k   109
##  9 Agnostic >150k        84
## 10 Agnostic NA           96
```

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?

The common problem illustrated by the Billboard dataset is that multiple types of observational units are stored in the same table, column headers are values and not variables, muliple variables are in one column.

```
(b) (10 pts) Using tidyverse and lubridate commands, create a data frame ‘bb‘ with the tidy form illust
```

```
library(tidyverse)
```

```
## Warning: package ’stringr’ was built under R version 4.2.2
```

```
library(lubridate)
bb <- billboard %>%
  pivot_longer(cols = starts_with("wk"),
               names_to = "week",
               names_prefix = "wk",
               values_to = "rank") %>%
  rename("date" = "date.entered") %>%
  filter(!is.na(rank)) %>%
  mutate(week = as.double(week),
         date = as.Date(date) + 7*(week-1))
bb %>% slice_head(n=15)
```

```
## # A tibble: 15 x 5
##    artist       track                   date        week  rank
##    <chr>        <chr>                   <date>     <dbl> <dbl>
##  1 2 Pac        Baby Don't Cry (Keep... 2000-02-26     1    87
##  2 2 Pac        Baby Don't Cry (Keep... 2000-03-04     2    82
##  3 2 Pac        Baby Don't Cry (Keep... 2000-03-11     3    72
##  4 2 Pac        Baby Don't Cry (Keep... 2000-03-18     4    77
##  5 2 Pac        Baby Don't Cry (Keep... 2000-03-25     5    87
##  6 2 Pac        Baby Don't Cry (Keep... 2000-04-01     6    94
##  7 2 Pac        Baby Don't Cry (Keep... 2000-04-08     7    99
##  8 2Ge+her      The Hardest Part Of ... 2000-09-02     1    91
##  9 2Ge+her      The Hardest Part Of ... 2000-09-09     2    87
## 10 2Ge+her      The Hardest Part Of ... 2000-09-16     3    92
## 11 3 Doors Down Kryptonite              2000-04-08     1    81
## 12 3 Doors Down Kryptonite              2000-04-15     2    70
## 13 3 Doors Down Kryptonite              2000-04-22     3    68
## 14 3 Doors Down Kryptonite              2000-04-29     4    67
## 15 3 Doors Down Kryptonite              2000-05-06     5    66
```

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?

The problem with the TB dataset is that there are multiple variables stored in one column, and the column headers are values and not variables.

```
(b) (10 pts) One way in which the ‘who‘ dataset differs from the dataset in Table 9 is that the ‘who‘ da
```

```
who_countries <- who %>%
  select("country", "iso2", "iso3") %>%
  distinct() %>%
  arrange(country)

who_countries %>% slice_head(n = 6)
```

```
## # A tibble: 6 x 3
##   country        iso2  iso3 
##   <chr>          <chr> <chr>
## 1 Afghanistan    AF    AFG  
## 2 Albania        AL    ALB  
## 3 Algeria        DZ    DZA  
## 4 American Samoa AS    ASM  
## 5 Andorra        AD    AND  
## 6 Angola         AO    AGO
```

```
(c) (10 pts) Using tidyverse functions, create a data frame ‘who_tb‘ similar to Table 10(b) of the pape

      - 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‘, ‘55-6
      - 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.
```

```
who_tb <- who %>%
  pivot_longer(cols = new_sp_m014:newrel_f65,
    names_to= c("diagnosis", "sex", "age"),
    names_pattern = "new_?(.*)_(.)(.*)",
    values_to = "cases") %>%
  select(-c(country, iso3)) %>%
  filter(!is.na(cases))
```

```
num <- "([0-9]{1, })([0-9]{2})"

who_tb$age[which(who_tb$age != "65")] <-
  sub(num, "\\1-\\2", who_tb$age[who_tb$age != "65"])

who_tb$age[which(who_tb$age== "65")] <-
  sub("65", "65\\+", who_tb$age[who_tb$age == "65"])

who_tb$age <- ordered(who_tb$age, levels = c("0-14","15-24","25-34","35-44","45-54","55-64","65+"))

who_tb %>% slice_head(n=15)
```

```
## # A tibble: 15 x 6
##    iso2   year diagnosis sex   age   cases
##    <chr> <int> <chr>     <chr> <ord> <int>
##  1 AF     1997 sp        m     0-14      0
##  2 AF     1997 sp        m     15-24    10
##  3 AF     1997 sp        m     25-34     6
##  4 AF     1997 sp        m     35-44     3
##  5 AF     1997 sp        m     45-54     5
##  6 AF     1997 sp        m     55-64     2
##  7 AF     1997 sp        m     65+       0
##  8 AF     1997 sp        f     0-14      5
##  9 AF     1997 sp        f     15-24    38
## 10 AF     1997 sp        f     25-34    36
## 11 AF     1997 sp        f     35-44    14
## 12 AF     1997 sp        f     45-54     8
## 13 AF     1997 sp        f     55-64     0
## 14 AF     1997 sp        f     65+       1
## 15 AF     1998 sp        m     0-14     30
```

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?

The problem with the weather dataset is that variables are stored in both rows and columns.

```
(b) (15 pts) Read the data into R and convert them to the form seen in Table 12(b) of the paper, execpt
```

```
weather <- read_csv("C:/Users/camid/Documents/UF/SOPHOMORE YR/FALL2022/STA3100/tidy data/weather.csv")

weather <- weather %>%
  pivot_longer(cols = starts_with("d"),
               names_to = "day",
               values_to = "value")

weather$day <- gsub("d", "",as.character(weather$day))

weather <- weather %>%
  mutate(date = make_date(year = weather$year,
                          month =weather$month,
                          day = weather$day),
         .after = id) %>%
  select(-c("year", "month", "day"))

weather <- weather %>%
  pivot_wider(names_from = element,
              values_from = value) %>%
  mutate(tmax = as.character(tmax),
         tmin = as.character(tmin),
         prcp = as.character(prcp)) %>%
  arrange(date)
```

```
## Warning: Values from ‘value‘ are not uniquely identified; output will contain list-cols.
## * Use ‘values_fn = list‘ to suppress this warning.
## * Use ‘values_fn = {summary_fun}‘ to summarise duplicates.
## * Use the following dplyr code to identify duplicates.
##   {data} %>%
##     dplyr::group_by(id, date, element) %>%
##     dplyr::summarise(n = dplyr::n(), .groups = "drop") %>%
##     dplyr::filter(n > 1L)
```

```
slice_head(weather, n = 10)
```

```
## # A tibble: 10 x 5
##    id      date       tmax  tmin  prcp
##    <chr>   <date>     <chr> <chr> <chr>
##  1 MX17004 1955-04-01 31    15    0
##  2 MX17004 1955-04-02 31    15    0
##  3 MX17004 1955-04-03 31    16    0
##  4 MX17004 1955-04-04 32    15    0
##  5 MX17004 1955-04-05 33    16    0
##  6 MX17004 1955-04-06 32    16    0
##  7 MX17004 1955-04-07 32    16    0
##  8 MX17004 1955-04-08 33    16    0
##  9 MX17004 1955-04-09 33    16    0
## 10 MX17004 1955-04-10 33    17    0
```

References

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