Math 300 Lesson 8 Notes

Importing Data

YOUR NAME HERE

June, 2022

Contents

Objectives

1. Import csv and Excel data files into `R`.
2. Explain and use appropriately the concept of tidy data.
3. Create a tidy data frame using the appropriate functions in `R`.

Reading

Chapter 4 - 4.2

Lesson

Work through the learning checks LC4.1 - LC4.3. Complete the code as necessary.

- Although this chapter seems straightforward, it is not. Thinking about the form you want the data in means defining the observational unit.
- From Wickham's paper: Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. You must think about your data before analyzing it.
- The book makes it seem that long data frames are tidy and wide ones are not. Be careful, this can be too simplistic.
- The `pivot_longer()` function is difficult when you first use it. The function arguments can be confusing. The `names_to` and `values_to` are really just asking for the names of columns when done. The `names_to` takes the column names and creates a variable with the assigned name. The `values_to` takes the values in the selected columns and makes them a variable. The `cols` is subtle and can be done in a variety of ways. Practice, run `?pivot_longer` or go to the tidyverse for more examples.

Setup

```
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)
library(nycflights13)
library(fivethirtyeight)
```

Import data 4.1.2 (Objective 1)

Repeat the import of `dem_score.xlsx` into `R`. Experiment with the options in the GUI. Also import `https://moderndive.com/data/dem_score.csv` using the `Import Dataset` icon under the **Environment** tab.

LC 4.1 (Objective 2)

(lc 4.1) What are common characteristics of "tidy" datasets?

Solution:

LC 4.2 (Objective 2)

(LC 4.2) What makes "tidy" datasets useful for organizing data?

Solution:

LC 4.3 (Objective 2)

(LC 4.3) Take a look the `airline_safety` data frame included in the `fivethirtyeight` data. Run the following:

```
head(airline_safety)
```

After reading the help file by running `?airline_safety`, we see that `airline_safety` is a data frame containing information on different airlines companies' safety records. This data was originally reported on the data journalism website FiveThirtyEight.com in Nate Silver's article "Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?". Let's ignore the `incl_reg_subsidiaries` and `avail_seat_km_per_week` variables for simplicity:

```
airline_safety_smaller <- airline_safety %>%
  select(airline, starts_with("fatalities"))
```

```
head(airline_safety_smaller)
```

```
## # A tibble: 6 x 3
##   airline              fatalities_85_99 fatalities_00_14
##   <chr>                           <int>            <int>
## 1 Aer Lingus                          0                0
```

```
## 2 Aeroflot                          128               88
## 3 Aerolineas Argentinas               0                0
## 4 Aeromexico                         64                0
## 5 Air Canada                          0                0
## 6 Air France                         79              337
```

This data frame is not in "tidy" format. How would you convert this data frame to be in "tidy" format, in particular so that it has a variable `fatalities_years` indicating the incident type/year and a variable `count` of the counts?

Solution:

Documenting software

- File creation date: 2022-06-16
- R version 4.1.3 (2022-03-10)
- `ggplot2` package version: 3.3.6
- `tidyr` package version: 1.2.0
- `readr` package version: 2.1.2

- `dplyr` package version: 1.0.9
- `nycflights13` package version: 1.0.2
- `fivethirtyeight` package version: 0.6.2