Math 300 NTI Lesson 8

Importing Data

Professor Bradley Warner

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

Remember that you will be running this more like a lab than a lecture. You want them using `R` and answering questions. Have them open the notes rmd and work through it together.

Work through the learning checks LC4.1 - LC4.3.

- Although this chapter seems straightforward, it is not. Thinking about the form you want the data in means defining the observational unit. Spend time on the examples to help make the point. LC4.1 and LC4.2. In the background papers folder of the course materials, we have Wickham's paper on tidy data, this can give you more insight for class discussions.
- 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)

(LC4.1) What are common characteristics of "tidy" datasets?

Solution: Rows correspond to observations, while columns correspond to variables.

The data object `drinks_smaller` is tidy if the unit of observation is country. Thus each row is a country. The data object `drinks_smaller_tidy` is tidy if the unit of observation is alcoholic beverage consumption. The first case does not really help us in our analysis since the variables, columns, are all related to different beverages which we typically would want to compare.

LC 4.2 (Objective 2)

(LC4.2) What makes "tidy" datasets useful for organizing data?

Solution: Tidy datasets are an organized way of viewing data. This format is required for the `ggplot2` and `dplyr` packages for data visualization and wrangling.

Table 4.3 gives a good example of tidy data and uses the term *unique pieces of information.*

LC 4.3 (Objective 2)

(LC4.3) Take a look the `airline_safety` data frame included in the `fivethirtyeight` data. Run the following:

```
head(airline_safety)
```

```
## # A tibble: 6 x 9
##   airline     incl_reg_subsid~ avail_seat_km_p~ incidents_85_99 fatal_accidents~
##   <chr>       <lgl>                       <dbl>           <int>            <int>
```

```
## 1 Aer Lingus  FALSE                  320906734             2                0
## 2 Aeroflot    TRUE                  1197672318            76               14
## 3 Aerolineas~ FALSE                  385803648             6                0
## 4 Aeromexico  TRUE                   596871813             3                1
## 5 Air Canada  FALSE                 1865253802             2                0
## 6 Air France  FALSE                 3004002661            14               4
## # ... with 4 more variables: fatalities_85_99 <int>, incidents_00_14 <int>,
## #   fatal_accidents_00_14 <int>, fatalities_00_14 <int>
```

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:

The original data frame has an airline as the unit of observation. But we want an observation to be an airline in a time period.

This can been done using the `pivot_longer()` function from the `tidyr` package:

```
airline_safety_smaller_tidy <- airline_safety_smaller %>%
  pivot_longer(
    names_to = "fatalities_years",
    values_to = "count",
    cols = -airline
  )
```

```
head(airline_safety_smaller_tidy)
```

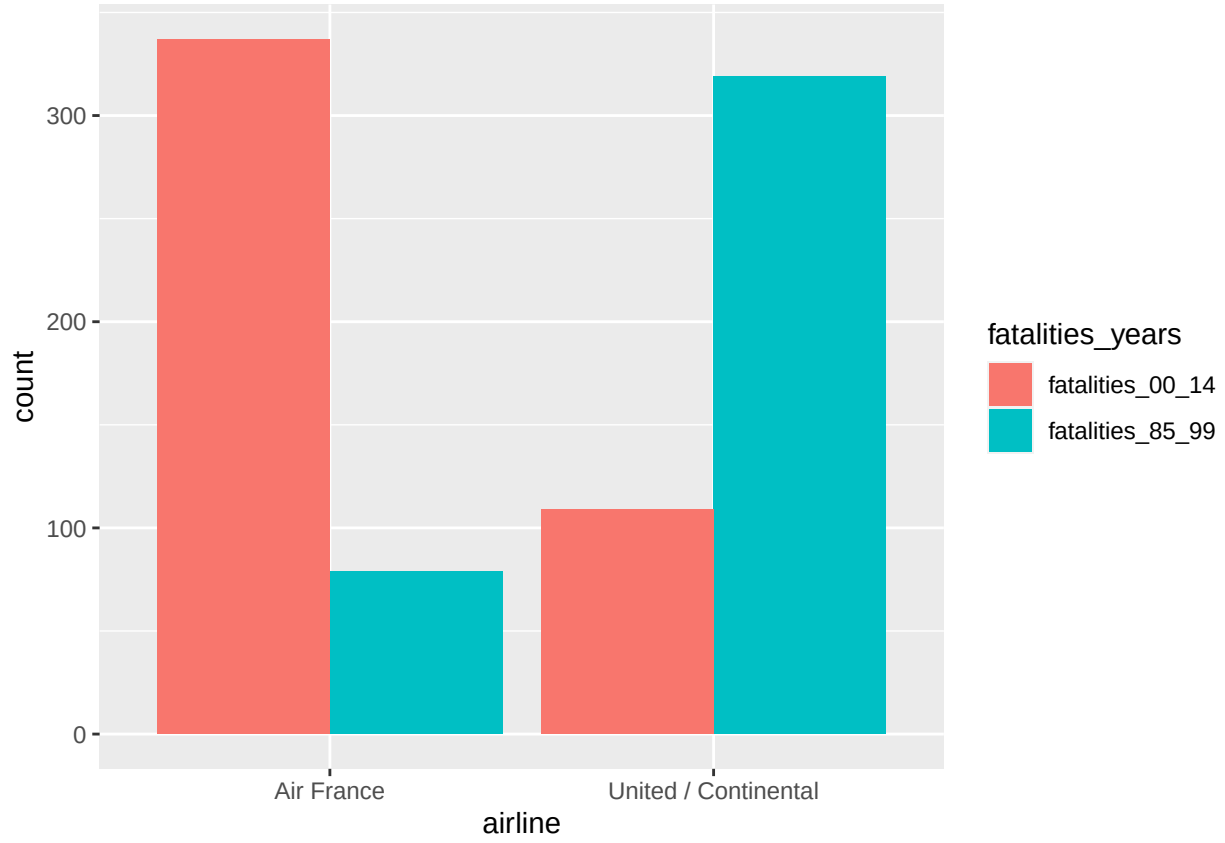
```
## # A tibble: 6 x 3
##   airline               fatalities_years count
##   <chr>                 <chr>            <int>
```

```
## 1 Aer Lingus            fatalities_85_99     0
## 2 Aer Lingus            fatalities_00_14     0
## 3 Aeroflot              fatalities_85_99   128
## 4 Aeroflot              fatalities_00_14    88
## 5 Aerolineas Argentinas fatalities_85_99     0
## 6 Aerolineas Argentinas fatalities_00_14     0
```

If you look at the resulting `airline_safety_smaller_tidy` data frame in the spreadsheet viewer, you'll see that the variable `fatalities_years` has 2 possible values: `"fatalities_85_99" and fatalities_00_14"` corresponding to the 2 columns of `airline_safety_smaller` we tidied.

Let's create plot of Air France and United Airlines.

```
airline_safety_smaller_tidy %>%
  filter(airline %in% c("United / Continental","Air France")) %>%
  ggplot(aes(x=airline,y=count,fill=fatalities_years)) +
  geom_col(position="dodge")
```



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