Math 300 NTI Lesson 1

Data with R

Professor Bradley Warner

June, 2022

Contents

Objectives

1. Install, load, and use R packages.
2. Explore data sets with R functions to include `glimpse()`, `View()`, `kable()` and `$`.
3. Identify and justify whether a variable is used for identification or measurement and, if measurement, whether it is categorical or quantitative.
4. Understand and practice the tips on learning to code to include using help functions and reading error messages.

Reading

Chapter 1

Lesson

Remember that you will be running this more like a lab than a lecture. You want to get them using `R` and answering questions. The focus will be on answering the learning checks. This will require us to use our book and to use RStudio. Those should be our primary teaching tools.

- Have a copy of the textbook open in a browser tab, the NTI in a browser tab, and the student notes Rmd open. Work through the Rmd with them using the answers in the NTI. You can go back to the book to bring clarity to specific points.
- Work through the learning checks LC1.1 - LC1.7. You may not be able to get through all of them. You can pick and choose which you want to emphasize in class. Refer to the book to get the question and appropriate reference material.

- Have them knit their work prior to leaving class. You may want to show how to insert a code chunk using Ctrl-Alt-I. At some point in the semester, you can discuss options in code chunks.
- To comment out code in a code chunk use the # symbol.
- We have an additional html document that they can read through that gives more information on `R`. It is *Intro_to_R_RStudio_desktop.html*.

LC 1.1 (Objective 1)

(LC1.1) Repeat the package installation steps from the text, but for the `dplyr`, `nycflights13`, and `knitr` packages. This will install the earlier mentioned `dplyr` package, the `nycflights13` package containing data on all domestic flights leaving a NYC airport in 2013, and the `knitr` package for writing reports in R.

- These may be already be installed. If so, show them how to load package using the Packages tab in RStudio. *

LC 1.2 (Objective 1)

(LC1.2) "Load" the `dplyr`, `nycflights13`, and `knitr` packages as well by using the `library()` function.

Solution: If the following code runs with no errors, you've succeeded!

```
library(dplyr)
library(nycflights13)
library(knitr)
```

Order does matter if multiple packages have functions or datasets of the same name. The most recently loaded package will overwrite existing objects of the same name.

LC 1.3 (Objective 2)

(LC1.3) What does any *ONE* row in the `flights` dataset refer to? Remind them that the `flights` dataset was loaded when we loaded the `nycflights13` package.

- A. Data on an airline
- B. Data on a flight
- C. Data on an airport
- D. Data on multiple flights

Solution: This is data on a single flight, and not a flight path! Example:

- a flight path would be the United 1545 to Houston
- a single flight would be the United 1545 to Houston at a specific date/time. For example: 2013/1/1 at 5:15am.

LC 1.4 (Objective 3, 4)

(LC1.4) What are some examples in this dataset of **categorical** variables? What makes them different than **quantitative** variables?

Solution: Hint: Type `?flights` in the console to see what all the variables mean!

- Categorical:
  - `carrier` the company
  - `dest` the destination
  - `flight` the flight number. Even though this is a number, its simply a label. Example United 1545 is not less than United 1714
- Quantitative:
  - `distance` the distance in miles
  - `time_hour` time

Go to R and walk through the help. They struggle with using help

LC 1.5 (Objective 3)

(LC1.5) What properties of the observational unit do each of `lat`, `lon`, `alt`, `tz`, `dst`, and `tzone` describe for the `airports` data frame? Note that you may want to use `?airports` to get more information.

Solution: `lat long` represent the airport geographic coordinates, `alt` is the altitude above sea level of the airport (Run `airports %>% filter(faa == "DEN")` to see the altitude of Denver International Airport), `tz` is the time zone difference with respect to GMT in London UK, `dst` is the daylight savings time zone, and `tzone` is the time zone label.

LC 1.6 (Objective 3)

(LC1.6) Create your own data frame. First, provide the names of at least three variables, one of which is an identification variable and the other two are not. Next, create your own tidy dataset that matches these conditions.

Solution:

```
LC6 <- tibble(id=c(1,2,3),gpa=c(3.4,2.7,3.6),pea=c(2.7,3.0,3.3))
```

```
glimpse(LC6)
```

```
## Rows: 3
## Columns: 3
## $ id  <dbl> 1, 2, 3
## $ gpa <dbl> 3.4, 2.7, 3.6
## $ pea <dbl> 2.7, 3.0, 3.3
```

- In the example,`id` is an identification variable as it identifies the observation in question.
- Anything else pertains to measurements: `gpa` and `pea`.

We can also look at the `weather` data object.

```
glimpse(weather)
```

```
## Rows: 26,115
## Columns: 15
## $ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EW~
## $ year       <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,~
## $ month      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ hour       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, ~
## $ temp       <dbl> 39.02, 39.02, 39.02, 39.92, 39.02, 37.94, 39.02, 39.92, 39.~
## $ dewp       <dbl> 26.06, 26.96, 28.04, 28.04, 28.04, 28.04, 28.04, 28.04, 28.~
## $ humid      <dbl> 59.37, 61.63, 64.43, 62.21, 64.43, 67.21, 64.43, 62.21, 62.~
## $ wind_dir   <dbl> 270, 250, 240, 250, 260, 240, 240, 250, 260, 260, 260, 330,~
## $ wind_speed <dbl> 10.35702, 8.05546, 11.50780, 12.65858, 12.65858, 11.50780, ~
## $ wind_gust  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 20.~
## $ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ pressure   <dbl> 1012.0, 1012.3, 1012.5, 1012.2, 1011.9, 1012.4, 1012.2, 101~
## $ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,~
## $ time_hour  <dttm> 2013-01-01 01:00:00, 2013-01-01 02:00:00, 2013-01-01 03:00~
```

- The combination of `origin`, `year`, `month`, `day`, `hour` are identification variables as they identify the observation in question.
- Anything else pertains to measurements: `temp`, `humid`, `wind_speed`, etc.

LC 1.7 (Objective 2, 4)

(LC1.7) Look at the help file for the `airports` data frame. Revise your earlier guesses about what the variables `lat`, `lon`, `alt`, `tz`, `dst`, and `tzone` each describe.

Solution:

Type `?airports` in the console to see what all the variables mean.

Documenting software

- File creation date: 2022-06-14
- R version 4.1.1 (2021-08-10)
- `dplyr` package version: 1.0.7
- `nycflights13` package version: 1.0.2