ETC5512: Introduction to Data Collection

Table of contents

Learning Objectives

- Identify whether data are experimental or observational

- Delineate the data collection methods
- Logically suggest the population that a sample represents

> ⚠ Warning
>
> This tutorial is on the long side. Don't expect to finish all the exercises in your tutorial. You will likely need to follow up in your own time.

Before your tutorial

Work through the following startR modules on Tidy Data Analysis:

- Reading data from files
- Tidy data and restructuring
- Data manipulation

These should take you ~ 1 hour and 30 minutes.

Package Installation

Ensure you have the packages installed from Week 1's tutorial.

Also install

```
install.packages("tibble")
install.packages("maps")
install.packages("ggthemes")
```

Exercise 1

This question relates to the Tidy Tuesday Data on locations of alternative fuel recharging stations. Have a read through this site, and also visit the link to the data providers, DOT.

a. About the data

Read the details about the data at DOT. How is this data collected, do you think?

b. Data type

What type of data is this? (observational, experimental, survey, census)

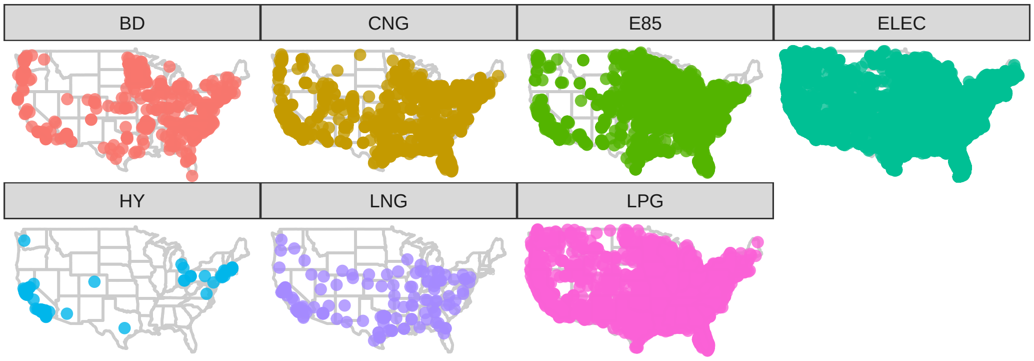
c. Population vs sample

Describe the population, and what is the sample.

d. Download and plot the data

Download the data and plot the fueling locations on a map, coloured by fuel type.

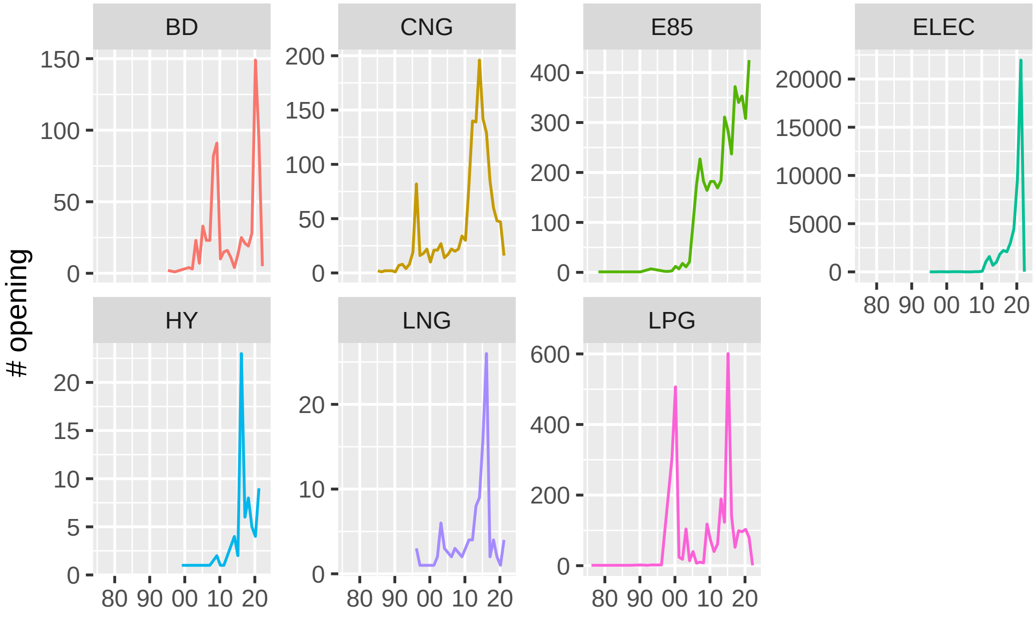
```r
library(tidyverse)
library(ggthemes)
library(maps)

# Note you can read data directly from a website
stations <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/da

# Get the map data for the USA
usa <- map_data("state")

# Filter to continental USA using map boundary
stations <- stations |>
  filter(between(LONGITUDE, min(usa$long)-1, max(usa$long)+1),
         between(LATITUDE, min(usa$lat)-1, max(usa$lat)+1))

# Plot the sites on a map
# Create a different map for each fuel type

ggplot() +
  geom_path(data=usa, aes(x=long, y=lat, group=group), colour="grey80") +
  geom_point(data=stations, aes(x=LONGITUDE,
                                y=LATITUDE,
                                colour=FUEL_TYPE_CODE),
             alpha=0.8) +
  facet_wrap(~FUEL_TYPE_CODE, ncol=4) +
  coord_map() +
  theme_map() +
  theme(legend.position = "none")
```



e. New Fuel Stations

Count the number of new stations by month, and make a time series plot by fuel type.

```
# Time line of opening
stations |>
  mutate(OPEN_DATE = as.Date(OPEN_DATE)) |>
  filter(!is.na(OPEN_DATE)) |>
  mutate(m = month(OPEN_DATE),
         yr = year(OPEN_DATE)) |>
  mutate(open_yrmth = as.Date(paste(yr, m, "01", sep="-"), "%Y-%M-%d")) |>
  group_by(open_yrmth, FUEL_TYPE_CODE) |>
  summarise(nopen = n(), .groups = "drop") |>
ggplot(aes(x=open_yrmth,
           y=nopen,
           colour=FUEL_TYPE_CODE)) +
  geom_line() +
  facet_wrap(~FUEL_TYPE_CODE, ncol=4, scales="free_y") +
  ylab("# opening") +
  scale_x_date("", date_labels="%y") +
  theme(legend.position = "none")
```



g. Fuel growth

If the question to answer is "which alternative fuel vehicle is the fastest growing?" what is the explanatory (independent, predictor) variable and what is the response variable?

Exercise 2

Here we will look at the Chocolate bar ratings. Details (brief) of how the data was collected are provided here and more about the data itself is here.

```
chocolate <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/
```

a. Type of data

What type of data is this? (observational, experimental, survey, census)

b. Data collection

How is the data collected?

c. Population vs sample

Describe the population.

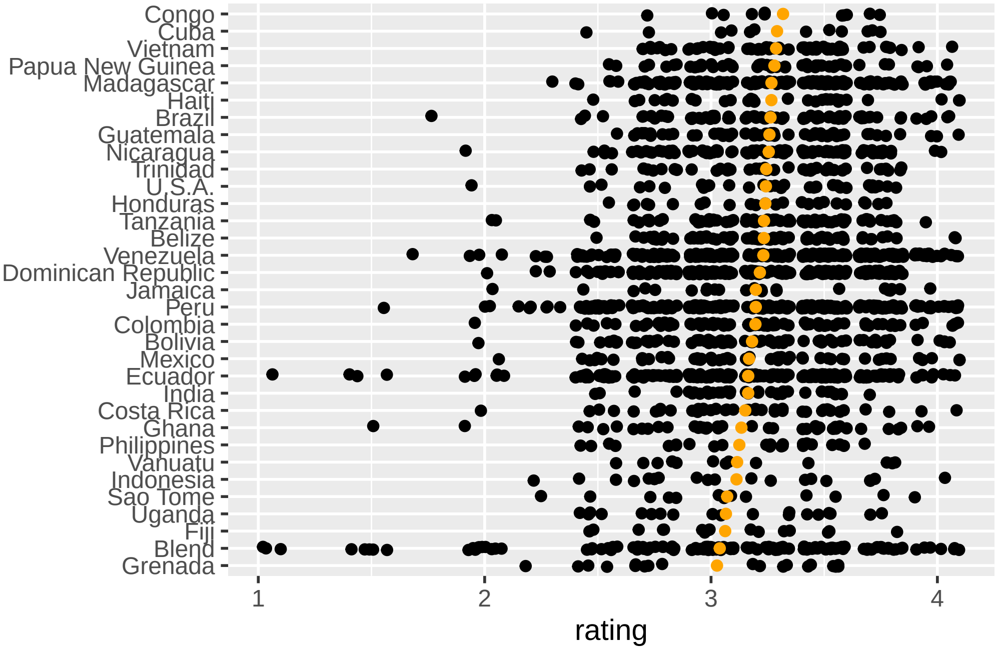
d. Response and predictor variables

For the question “Which country of origin of the bean obtains the best rating?” state the response and predictor variables.

e. Visualise your data

Make a plot to answer the previous question. (Only use countries with more than 10 records.)

```r
# Which countries have more than 10
keep <- chocolate |>
  count(country_of_bean_origin, sort = TRUE) |>
  filter(n>10) |>
  pull(country_of_bean_origin)

# Filter to those countries
chocolate_countries = chocolate |>
  filter(country_of_bean_origin %in% keep)

# Plot the results
  ggplot(data = chocolate_countries, aes(x=fct_reorder(country_of_bean_origin, rating, mean)
             y=rating)) +
    geom_jitter(width=0.1) +
    stat_summary(fun = mean, fun.min = median, fun.max = median,
                 geom = "point", colour = "orange") +
    xlab("") +
    coord_flip()
```



In your own time

Read the description of the study titled "Clearing the Fog: Is Hydroxychloroquine Effective in Reducing COVID-19 Progression (COVID-19)".

a. Data type

What type of data is this? (observational, experimental, survey, census)

b. Study Participants

How many subjects participated in the study at the start, and to completion?

c. Study details

What are the:

- experimental units?
- factor?
- blocking factors?
- response variable (outcome measures)?

d. Study desing

How are subjects assigned to treatment groups?

e. Study results

What were the results of the study?

f. Analyse results

Construct the data from the results reported. Compute the proportion of subjects with progression of COVID after 5 days, for the two treatments. Include the standard error of the estimate.

```r
hcq <- tibble(trt = c("standard", "standard", "hcq", "hcq"),
              progression = c("all", "yes", "all", "yes"),
              count = c(151, 5, 349, 11))
hcq |>
  pivot_wider(names_from = "progression", values_from = "count") |>
  mutate(p = yes/all) |>
  mutate(se = sqrt(p*(1-p)/all))
```

```
# A tibble: 2 x 5
  trt        all   yes      p      se
  <chr>    <dbl> <dbl>  <dbl>   <dbl>
1 standard   151     5 0.0331 0.0146
2 hcq        349    11 0.0315 0.00935
```

g. Study concultsions

Based on the proportions and their standard errors, why would the result of the study be that HCQ does NOT improve the outcomes of COVID patients?

h. Population vs sample

What is the population for this experiment?