Day 3 - Intro to Data

Introduction

In this section, we begin by defining the fundamental building blocks of statistics: **data.** The purpose of this activity is to introduce vocabulary terms that are essential to working with data. At the conclusion of this activity, you should understand the following terms:

- observational units
- quantitative/numeric variables
- categorical variables
- ordinal versus nominal
- discrete versus continuous
- response versus explanatory
- association/correlation and independence
- observational studies versus experiments
- confounding variables

We will use the `county` data set in the `usdata` package to guide our exploration. The following code will bring the data into R.

```
# library packages for data manipulation
# remember: you might have to install usdata!
library(tidyverse)
library(usdata)
library(ggthemes)

# load data
data(county)
```

If you choose to type your responses, do not forget to surround your answers in a callout box:

> Like this!

Observations, variables, and data frames

The following code chunk will print the first six rows of the county **data frame**, filtered to records from Vermont.

```
county %>%
  filter(state %in% "Vermont") |> # filter to VT
  head() # view first six rows
```

```
# A tibble: 6 x 15
  name            state pop2000 pop2010 pop2017 pop_change poverty homeownership
  <chr>           <fct>   <dbl>   <dbl>   <int>      <dbl>   <dbl>         <dbl>
1 Addison County  Verm~   35974   36821   36776      -0.13     7.8          75.6
2 Bennington Cou~ Verm~   36994   37125   35594      -2.61    12.8          72.6
3 Caledonia Coun~ Verm~   29702   31227   30164      -2.82    12.3          71.4
4 Chittenden Cou~ Verm~  146571  156545  162372       1.76    11.3          65.9
5 Essex County    Verm~    6459    6306    6230       0.6     15.5          83.9
6 Franklin County Verm~   45417   47746   49025       1.32     8            74.8
# i 7 more variables: multi_unit <dbl>, unemployment_rate <dbl>, metro <fct>,
#   median_edu <fct>, per_capita_income <dbl>, median_hh_income <int>,
#   smoking_ban <fct>
```

1. What are the **observational units** for these data?

   Counties in Vermont

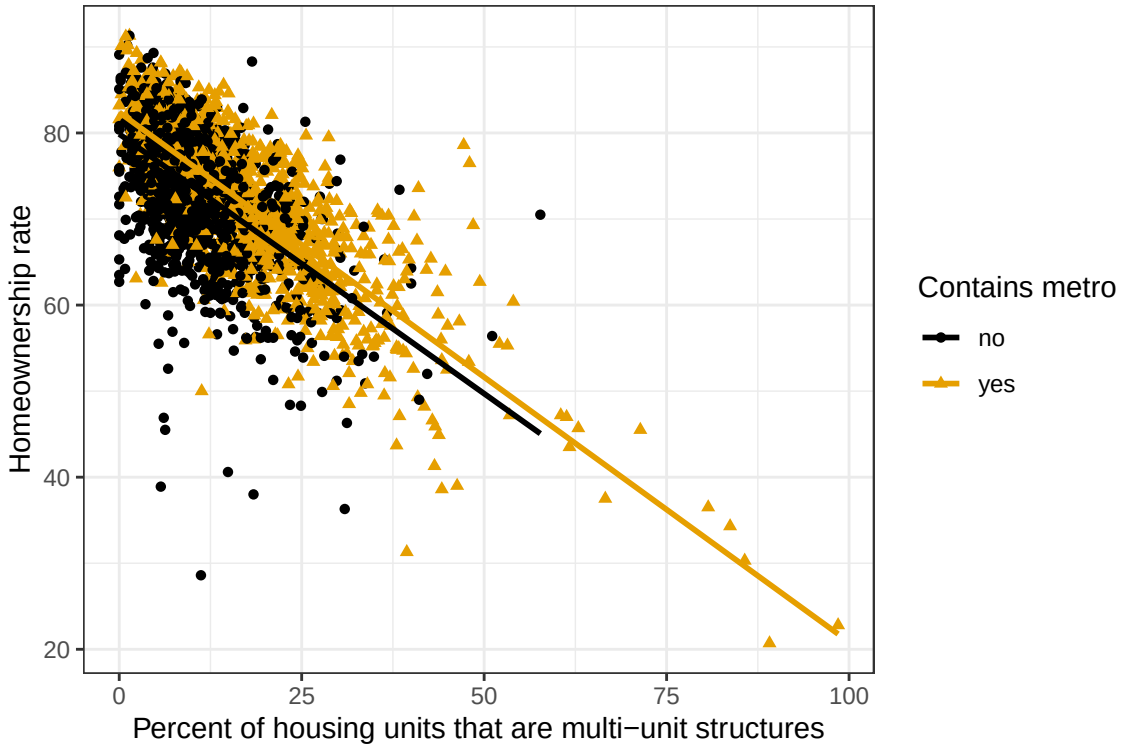
2. List each variable, and determine whether it is categorical or quantitative. For each categorical variable, determine whether it is nominal or ordinal. For each quantitative variable, determine whether it is discrete or continuous.

   - `name`: categorical, nominal
   - `state`: categorical, nominal
   - `pop2000`, `pop2010`, `pop2017`: quantitative, discrete
   - `pop_change`: quantitative, continuous
   - `poverty`: quantitative, continuous
   - `homeownership`: quantitative, continuous
   - `multi_unit`: quantitative, continuous
   - `unemployment_rate`: quantitative, continuous
   - `metro`: categorical, nominal
   - `median_edu`: categorical, ordinal
   - `per_capita_income`: quantitative, continuous
   - `median_hh_income`: quantitative, continuous
   - `smoking_ban`: categorical (`none`, `partial`).

Relationships between variables

Consider the following question and accompanying **scatterplot**:

> Is the homeownership rate associated with the percent of housing units in multi-unit structures? Does that relationship depend on whether the county contains a metropolitan area?



3. What variables are displayed on the graph. What type of variables are they?

   - `multi_unit` (x-axis): quantitative, continuous
   - `homeownership` (y-axis): quantitative, continuous
   - `metro` (color/shape): categorical, nominal

4. Which variable(s) is the response variable?

   Homeownership rate

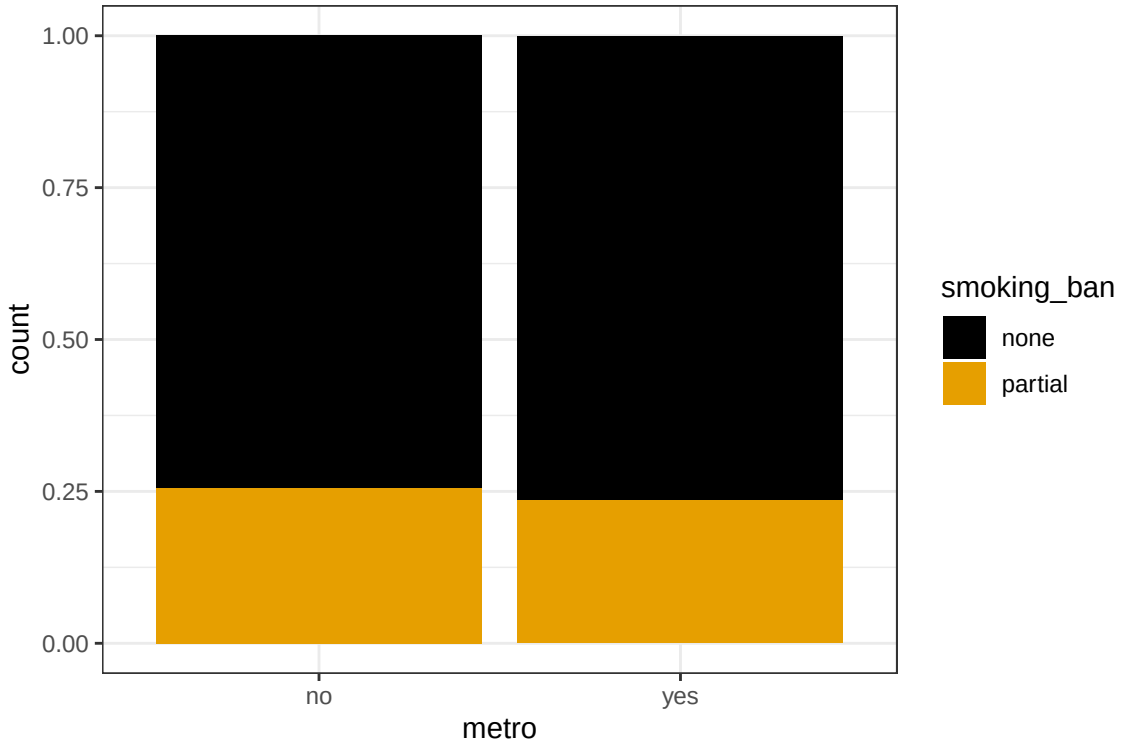
5. Which variable(s) is the explanatory variable?

   Percent of housing in multi-unit structures (`multi_unit`) and whether the county contains a metro area (`metro`).

6. Is homeownership rate **associated** with percent housing in multi-unit structures? Why or why not?

Yes. There is a clear, fairly strong, negative association: counties with a higher percentage of multi-unit housing tend to have lower homeownership rates. The relationshiop does not appear to depend much on metro, since the two lines are roughly parallel.

7. Is whether a county has a smoking ban **independent** of whether the county contains a metropolitan area? Why or why not?



They appear approximately independent. The proportion of counties with no smoking ban is about the same for non-metro counties and metro counties ($\approx 77\%$).

Observational studies and experiments

For each of the following scenarios, identify:

- whether it is an **observational study** or **randomized experiment**
- what the response variable and explanatory variables are
- whether we can conclude a causal relationship between the response variable and explanatory variable

For the observational studies, identify a potential confounding variable and explain why it may be a confounding variable.

8. Researchers collect survey data on people's daily coffee consumption and track whether they develop heart disease.

   - **Observational study**: researchers record coffee consumption but don't assign it.
   - **Response:** whether the person develops heart disease (categorical).
   - **Explanatory:** daily coffee consumption (quantitative).
   - **No causal conclusion.** We can only conclude an association, if there is one, because this is an observational study.
   - **Possible confounder: smoking.** Smoking is associated with heavier coffee drinking, and it independently raises heart disease risk. So any coffee-heart disease link could be driven partly or entirely by smoking.

9. A pharmaceutical company randomly assigns 500 patients with high blood pressure to either receive a new drug or a placebo, and measures their blood pressure after 6 months.

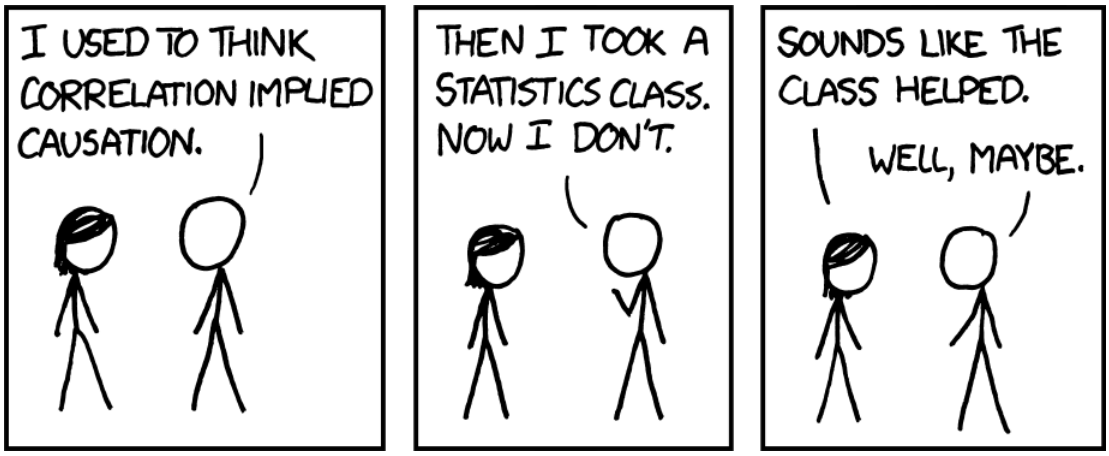
   - **Randomized experiment**: treatment is randomly assigned.
   - **Response:** blood pressure after 6 months (quantitative).
   - **Explanatory:** treatment group, drug vs. placebo (categorical).
   - **Yes, a causal conclusion is possible.** Random assignment means a difference in blood pressure between groups can be attributed to the drug.

10. A group of college students is randomly assigned to either play 30 minutes of a fast-paced video games or watch TV before taking a reaction-time test.

    - **Randomized experiment**: the activity is randomly assigned.
    - **Response:** reaction time (quantitative).
    - **Explanatory:** activity, video game vs. TV (categorical).
    - **Yes, a causal conclusion is possible.** Random assignment allows us to attribute differences in reaction time to the activity

! Correlation and causation

CORRELATION DOES NOT IMPLY CAUSATION. Two variables may be associated or correlated with one another despite no causal relationship existing.



xkcd comic on correlation and causation

11. Why does randomly assigning observational units allow us to draw causal inference?

    Random assignment will balance *all* other variables across the treatment groups in the long run (i.e., if we could keep repeating the experiment), such that the only source of the difference between the groups is the treatment we imposed.

Visit https://www.tylervigen.com/spurious-correlations to view a number of spurious correlations.

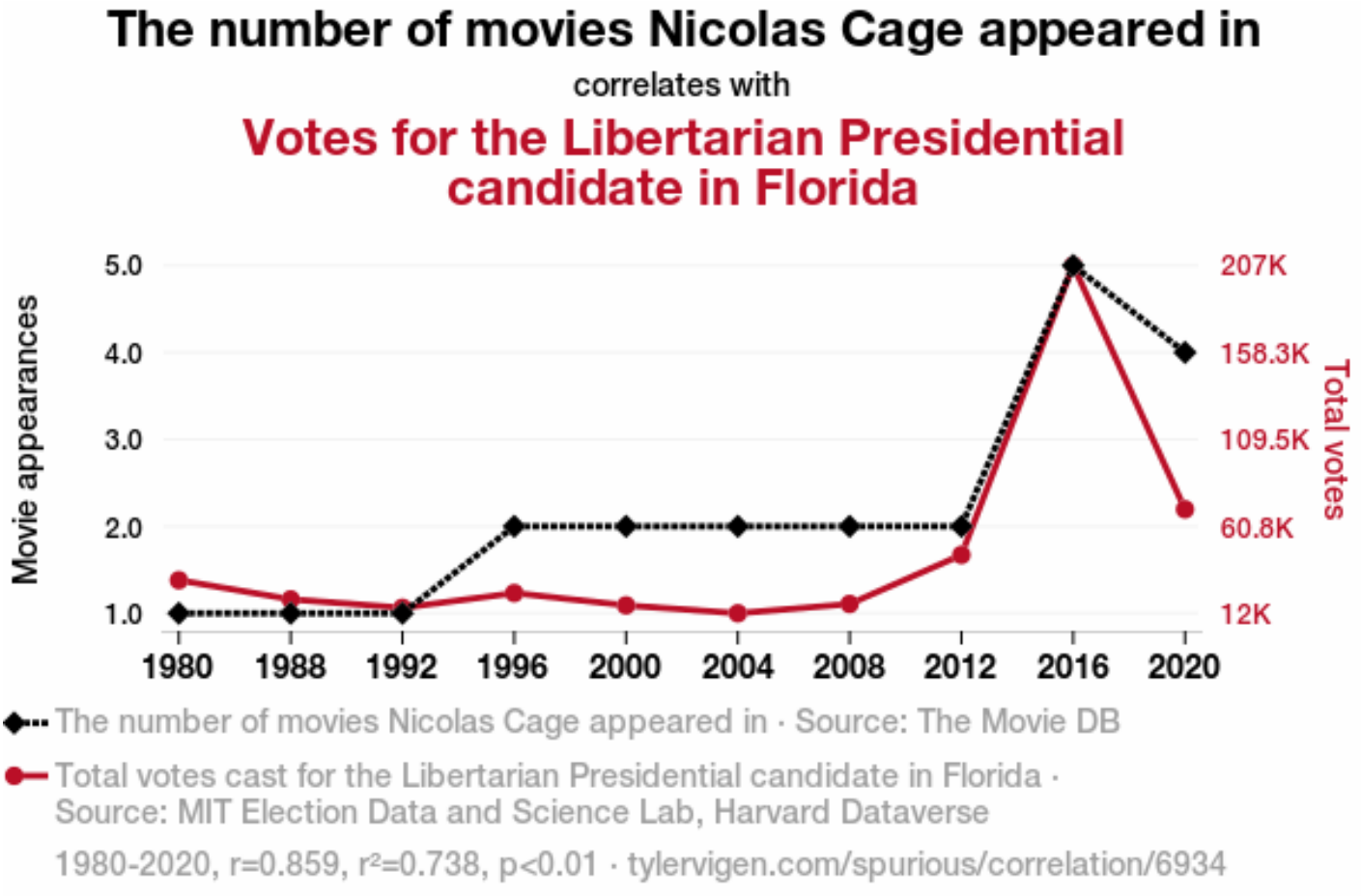


Image from https://www.tylervigen.com/spurious/correlation/6934