

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 <chr>	state <fct>	pop2000 <dbl>	pop2010 <dbl>	pop2017 <int>	pop_change <dbl>	poverty <dbl>	homeownership <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?

First box is free - pay attention to the spacing!

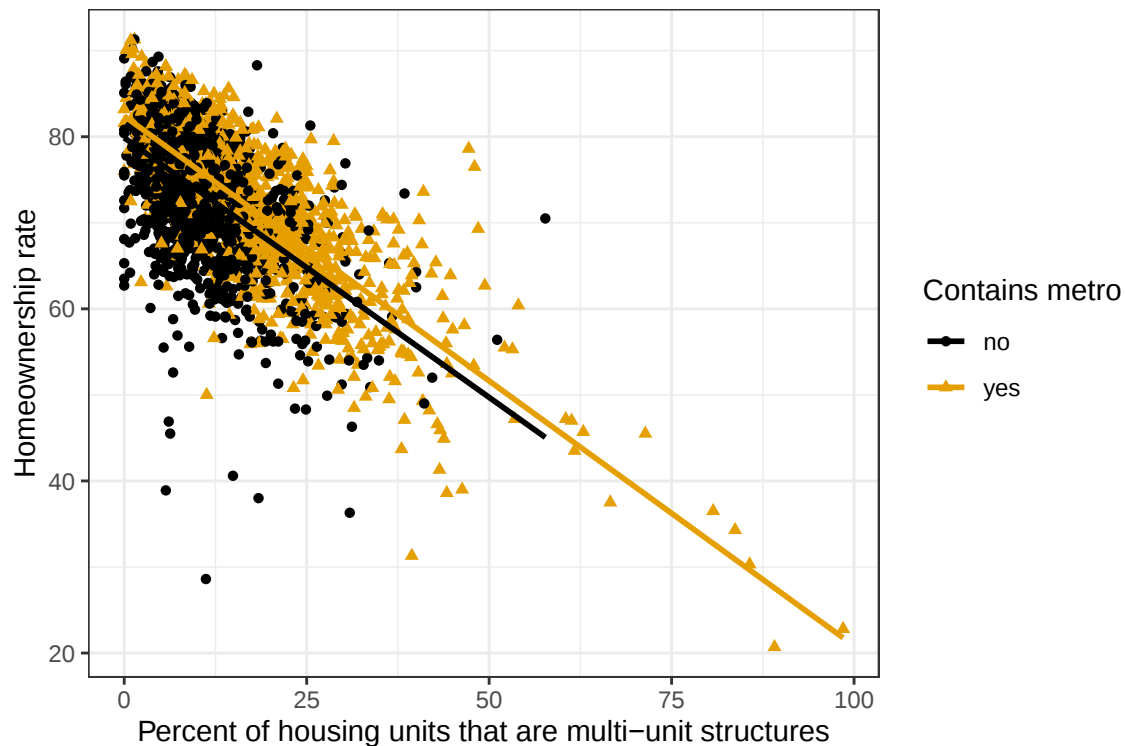
2. List each variable, and determine whether it is categorical or quantitative. For each categorical variable, determine whether it is nominal or ordinal. For each quantitative variable, determine whether it is discrete or continuous.

- list 1
- list 2

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?



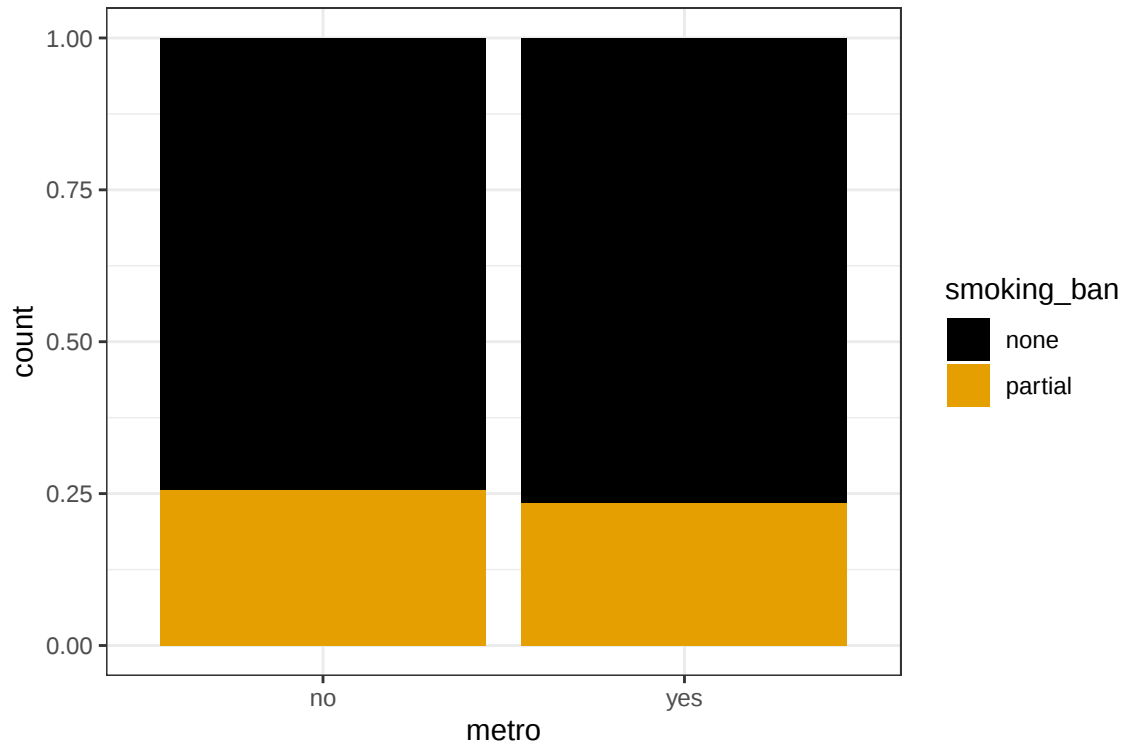
3. What variables are displayed on the graph. What type of variables are they?

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

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

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

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



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.

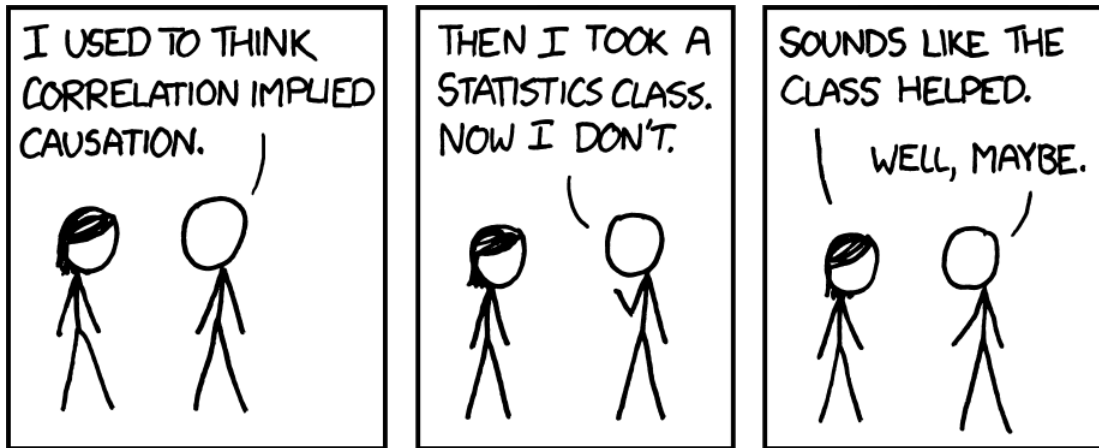
8. Researchers collect survey data on people's daily coffee consumption and track whether they develop heart disease.

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.

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.

! Correlation and causation

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



on correlation and causation

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

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

The number of movies Nicolas Cage appeared in correlates with **Votes for the Libertarian Presidential candidate in Florida**

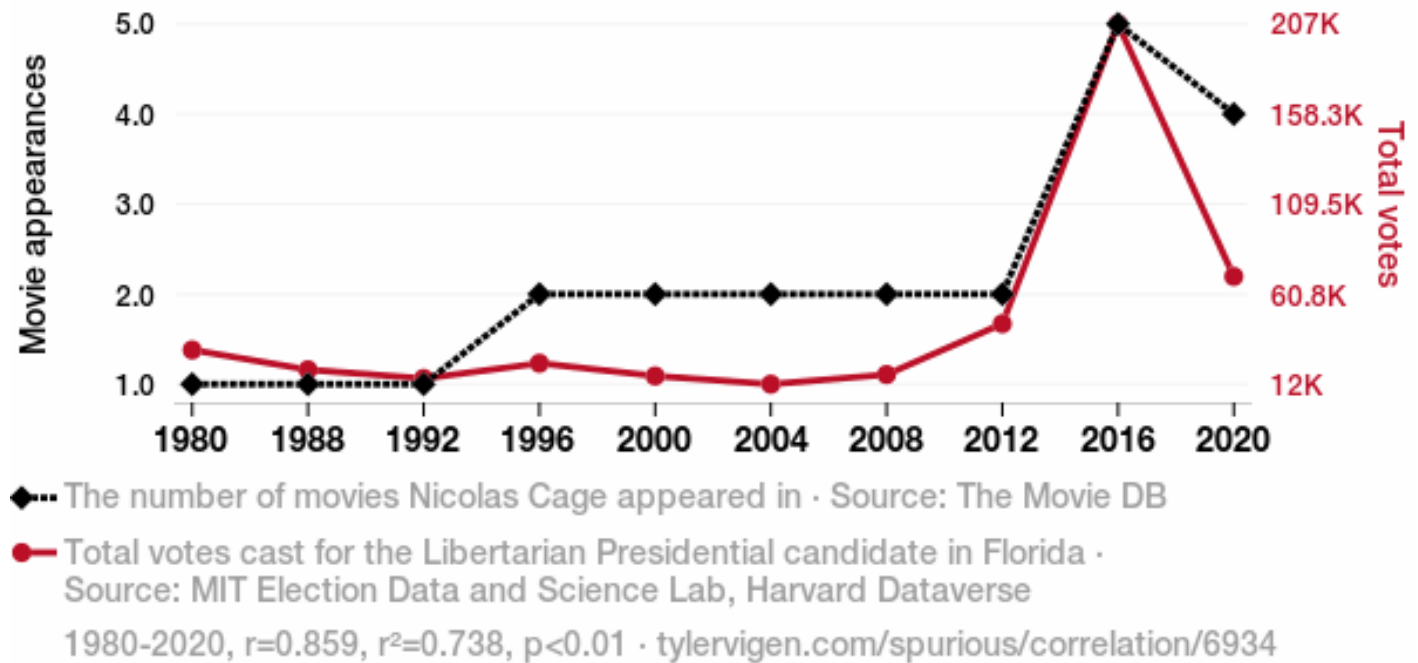


Image from <https://www.tylervigen.com/spurious/correlation/6934>