PO 399

Lecture 2

Ahyoung Cho

Department of Political Science

Summer 2024

TODAY

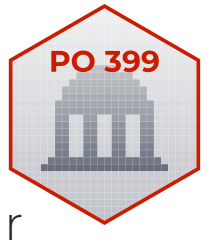
- Problem Set 0 Released (Due on Sunday, June 2 by 11:59 PM)
- Lab Session: Creating data frames
- Correlation
- Understanding File Paths
- Lab Session: Has voter turnout increased over time?
- Practice Problems

CORRELATIONS

WHICH OF THESE STATEMENTS DESCRIBE CORRELATIONS?

1. People who live to be 100 years old typically take vitamins.
2. Cities with more crime tend to hire more police officers.
3. Successful people have spent at least ten thousand hours honing their craft.
4. Most politicians facing a scandal win reelection.
5. Older people vote more than younger people.

CORRELATION



- Correlation tells us about the extent to which two features of the world tend to occur together or not.
- If two features of the world (or variables) tend to occur together, they are *positively* correlated.
- If the occurrence of one feature is unrelated to the other, they are *uncorrelated*.
- If when one feature occurs, the other tends not to occur, they are *negatively* correlated.

CORRELATION EXAMPLE WITH TWO BINARY VARIABLES

- Paradox of plenty: countries with an abundance of natural resources are often less economically developed and less democratic than those with fewer natural resources.
- Correlation between natural resources and economic/political system.
  - Natural Resources: Is country a major oil producer?
    - Major oil producer if a country exports more than 40,000 barrels per day per million people.
  - Political System: Is country considered an autocracy or a democracy?

CORRELATION EXAMPLE WITH TWO BINARY VARIABLES

	Not major oil producer	Major oil producer
Democracy	118	9
Autocracy	29	11

CORRELATION EXAMPLE WITH TWO BINARY VARIABLES

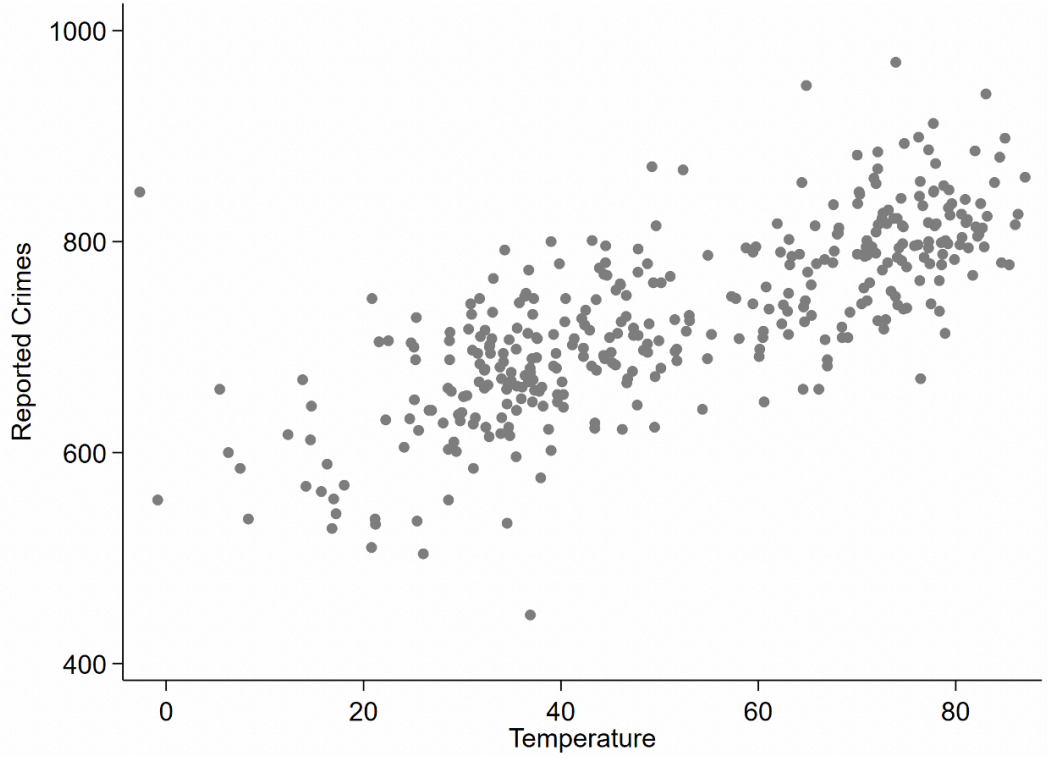
	Not major oil producer	Major oil producer
Democracy	118	9
Autocracy	29	11
Pr(Democracy)	118/(118+29) ≈ .802	9/(9+11) = .45

CORRELATION EXAMPLE WITH TWO BINARY VARIABLES

	Not major oil producer	Major oil producer	Pr(Oil)
Democracy	118	9	$9/(9+118) \approx .071$
Autocracy	29	11	$11/(11+29) = .275$

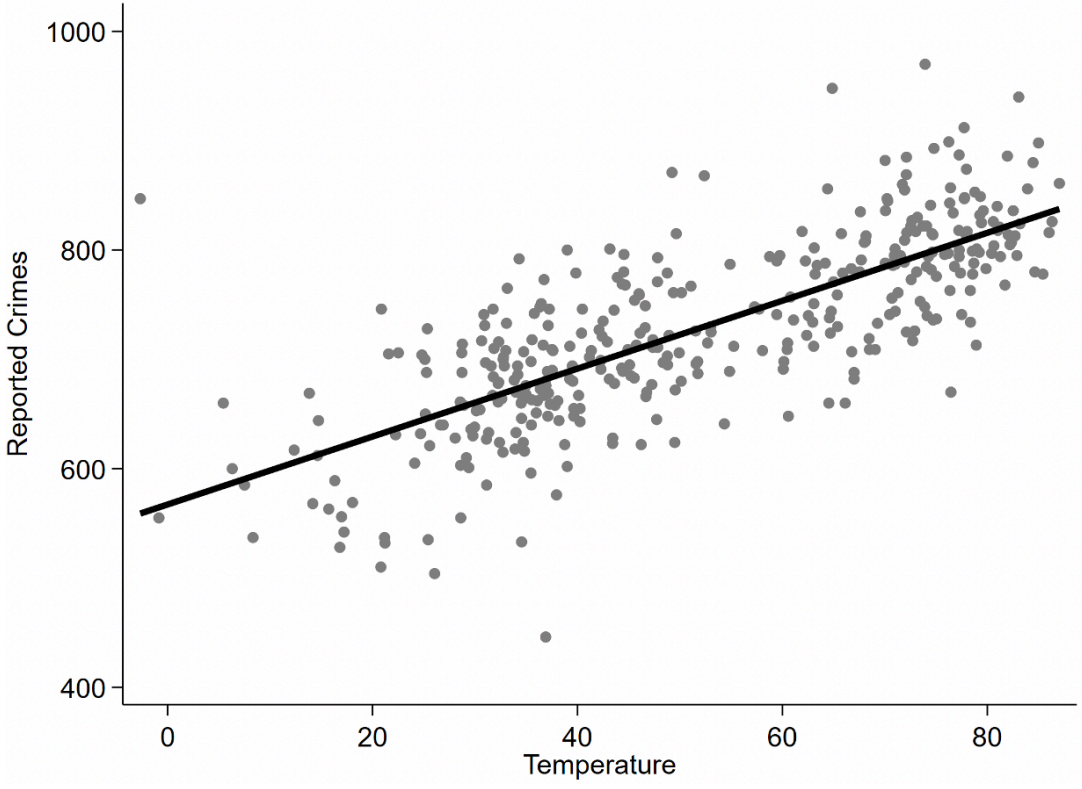
CORRELATION EXAMPLE WITH TWO CONTINOUS VARIABLES

Crime and Temperature in Chicago



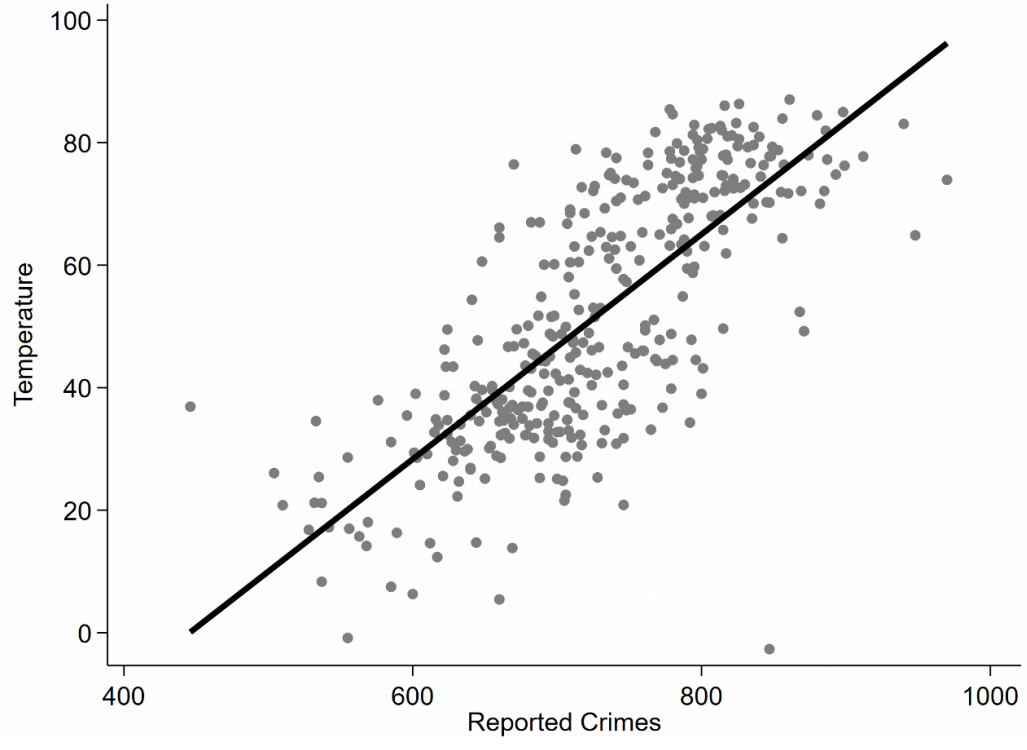
CORRELATION EXAMPLE WITH TWO CONTINOUS VARIABLES

Crime and Temperature in Chicago



CORRELATION EXAMPLE WITH TWO CONTINOUS VARIABLES

Crime and Temperature in Chicago



WHICH OF THESE STATEMENTS DESCRIBE CORRELATIONS?

1. People who live to be 100 years old typically take vitamins.
2. Cities with more crime tend to hire more police officers.
3. Successful people have spent at least ten thousand hours honing their craft.
4. Most politicians facing a scandal win reelection.
5. Older people vote more than younger people.

CORRELATION OR JUST DESCRIPTIVE FACT?

Most politicians facing a scandal win reelection

	No scandal	Scandal
Not reelected		8
Reelected		62

CORRELATION OR JUST DESCRIPTIVE FACT?

Most politicians facing a scandal win reelection

	No scandal	Scandal
Not reelected	101	8
Reelected	1,192	62

CORRELATION OR JUST DESCRIPTIVE FACT?

Most politicians facing a scandal win reelection

	No scandal	Scandal
Not reelected	101	8
Reelected	1,192	62
Pr(Reelected)	1,192/(1,192+101) ≈ .922	62/(62+8) ≈ .886

CORRELATION OR JUST DESCRIPTIVE FACT?

Most politicians facing a scandal win reelection

	No scandal	Scandal	Pr(Scandal)
Not reelected	101	8	$8/(8+101) \approx .073$
Reelected	1,192	62	$62/(62+1,192) \approx .049$

EXERCISE

The table below shows some data on which countries are major oil producers and which countries experienced a civil war between 1946 and 2004.

Are being a major oil producer and experiencing civil war positively correlated, negatively correlated, or uncorrelated? Explain your answer.

	Civil War	No Civil War
Oil Producer	7	12
Non-Oil Producer	55	94

POTENTIAL USES OF CORRELTION

Description

- Suppose we want to know if young people are underrepresented in elections?
- We might be interested in the correlation between age and voter turnout.
- No assumptions necessary (assuming we have good data).

POTENTIAL USES OF CORRELTION

Prediction/Forecasting

- Suppose we want to predict which restaurants are violating public health regulations.
- The correlation between negative Yelp reviews and hospitalizations from food-borne illness could be useful.
- We'd need to assume that the sample of restaurants for which we have data is representative of the population of restaurants for which we want to make predictions.
- We'd probably also want to think about linearity and avoid extrapolation (more on this later).

POTENTIAL USES OF CORRELTION

Causal inference

- Suppose we want to know if high school students would be more successful in life if they were forced to take calculus.
- The correlation between calculus and future success might be useful.
- But we'd have to assume that the students taking calculus are otherwise the same as everyone else in terms of their underlying chances of success.
- Aside from very special circumstances, this kind of assumption will be hard to defend.

EXERCISE

- Which of these statements describes a correlation?

    1. Most professional data analysts took a statistics course in college.

    2. Among Major League Baseball players, pitchers tend to have lower batting averages.

    3. Whichever presidential candidate wins Ohio tends to win the Electoral College.

- Consider the last statement about Ohio and presidential elections. Do you think it's useful for description? Forecasting? Causal inference? Why or why not?

UNDERSTANDING FILE PATHS

SETTING WORKING DIRECTORIES

- You can see what working directory you are in by running the function `getwd()`.

```
getwd()
```

- You can set your working directory with a function `setwd()`.

```
setwd("/Users/cho/PO 399 Summer 2024/Lectures/Lecture 2")
```

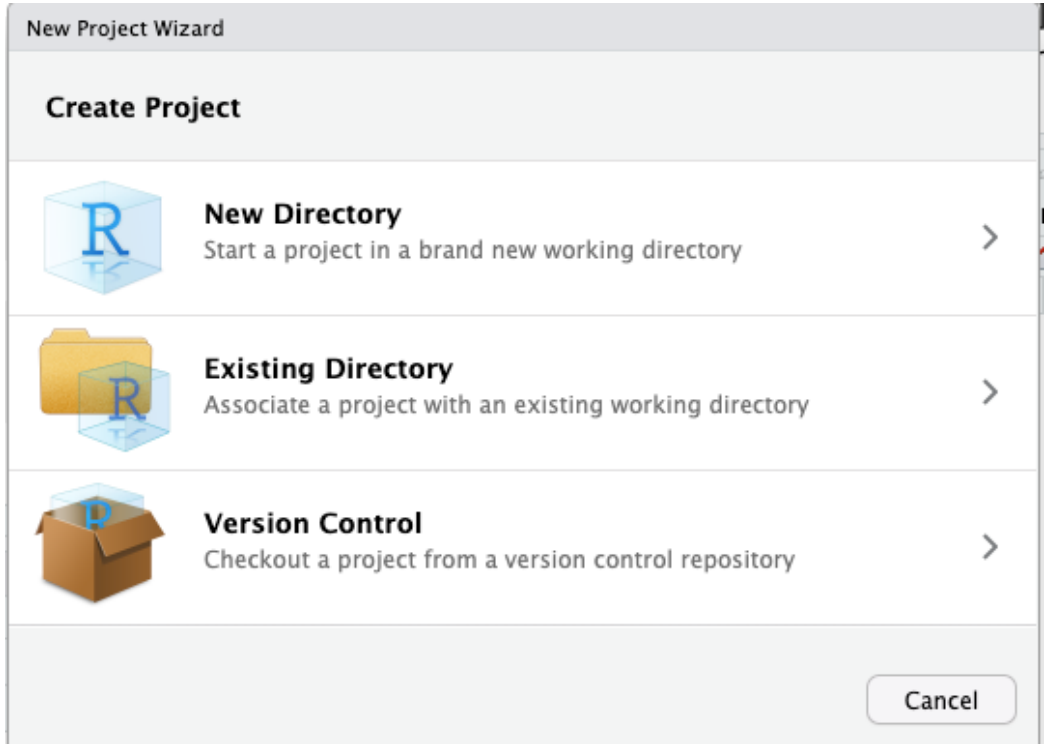
- `setwd()` uses an absolute file path which is specific to the computer you are working on.
- Very difficult to share your script with others.

RSTUDIO PROJECTS

- RStudio projects solve the problem by using the relative file paths.
- The RStudio project file (.Rproj) is located in the root directory.
- When you work on an RStudio session via the project file, the current working directory will be set to the root folder where the .Rproj file is stored.

CREATING RSTUDIO PROJECTS

- To create a project, open RStudio and select `File` -> `New Project...` from the menu.



HAS VOTER TURNOUT INCREASED OVER TIME?

HAS VOTER TURNOUT INCREASED OVER TIME?


```
library(tidyverse)

dat <- read_csv("voter_turnout.csv")
```

```
names(dat)
```

```
## [1] "year"              "turnout"           "vap"
## [4] "vep"               "registered_voters"
```

HAS VOTER TURNOUT INCREASED OVER TIME?


```
head(dat)
```

```
## # A tibble: 6 × 5
##    year   turnout       vap       vep registered_voters
##   <dbl>     <dbl>     <dbl>     <dbl>             <dbl>
## 1  1980  86496851 163945000 159635102         105135000
## 2  1984  92654861 173995000 167701904         116106000
## 3  1988  91586725 181956000 173579281         118598000
## 4  1992 104600366 189493000 179655523         126578000
## 5  1996  96389818 196789000 186347044         127661000
## 6  2000 105594024 209130000 194331436         129549000
```

HAS VOTER TURNOUT INCREASED OVER TIME?


```
glimpse(dat)
```

```
## Rows: 10
## Columns: 5
## $ year             <dbl> 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008, 2012…
## $ turnout          <dbl> 86496851, 92654861, 91586725, 104600366, 96389818, 1…
## $ vap              <dbl> 163945000, 173995000, 181956000, 189493000, 19678900…
## $ vep              <dbl> 159635102, 167701904, 173579281, 179655523, 18634704…
## $ registered_voters <dbl> 105135000, 116106000, 118598000, 126578000, 12766100…
```

HAS VOTER TURNOUT INCREASED OVER TIME?

- Create a new variable
  - `mutate` creates new variables.
  - Add a new variable for voter turnout percentage; overwrite the `dat` object (how do we define turnout?).

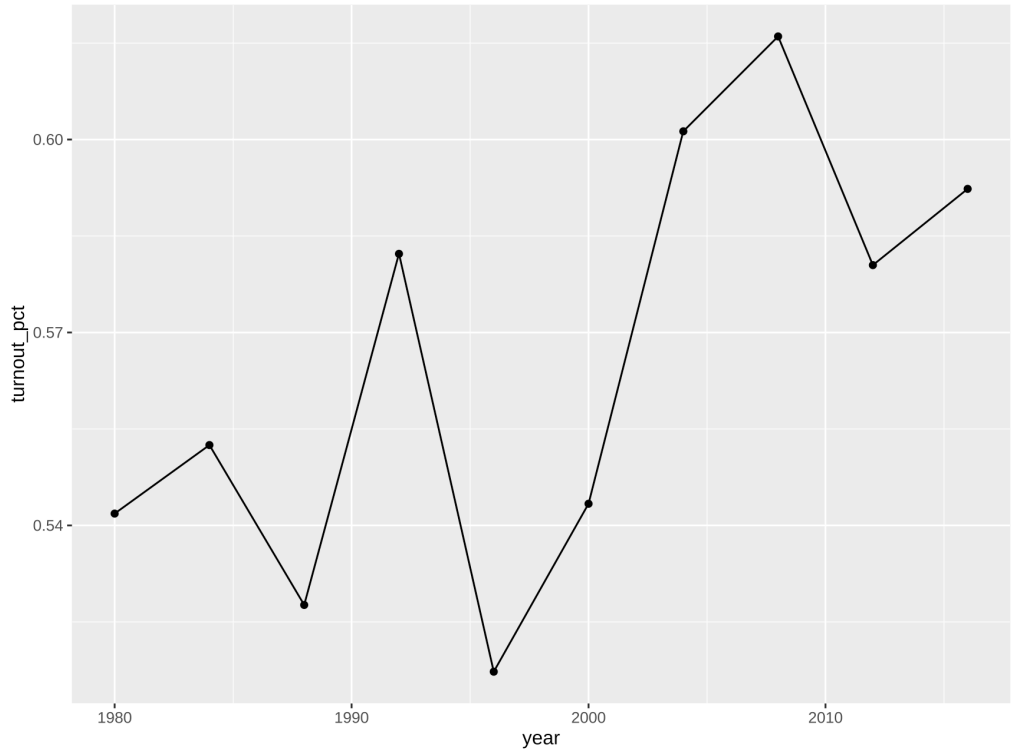
```
dat <- mutate(dat, turnout_pct = turnout/vep)
glimpse(dat)
```

```
## Rows: 10
## Columns: 6
## $ year             <dbl> 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008, 2012…
## $ turnout          <dbl> 86496851, 92654861, 91586725, 104600366, 96389818, 1…
## $ vap              <dbl> 163945000, 173995000, 181956000, 189493000, 19678900…
## $ vep              <dbl> 159635102, 167701904, 173579281, 179655523, 18634704…
## $ registered_voters <dbl> 105135000, 116106000, 118598000, 126578000, 12766100…
## $ turnout_pct      <dbl> 0.5418410, 0.5524974, 0.5276363, 0.5822274, 0.517259…
```

HAS VOTER TURNOUT INCREASED OVER TIME?

- Visualize the trend

```
ggplot(dat, aes(x=year, y=turnout_pct))
  geom_point() +
  geom_line()
```



HAS VOTER TURNOUT INCREASED OVER TIME?

- Measure the correlation
  - `cor()` is a new function, that takes two variables as arguments.
  - Reference a variable *within* a dataframe using `data_frame$variable_name`.

```
cor(dat$turnout_pct, dat$year)
```

```
## [1] 0.6458835
```

```
cor(dat$year, dat$turnout_pct)
```

```
## [1] 0.6458835
```

NEXT CLASS

- Causation