From Confusion to Confidence: Data wrangling and management

University of Essex - Department of Government

Lorenzo Crippa

Spring Term, 2021/2022

Your Q-Step GTA

Hi! My name is Lorenzo Crippa

Email: l.crippa@essex.ac.uk

Office: 6.020 (or home)

From Confusion to Confidence

What are we doing together?

Three videos to help you to think about next year's Capstone dissertation in advance

- Video 1: Getting started and producing effective research (Dr Howard Liu)
- Video 2: Data wrangling and management (Lorenzo Crippa)
- Video 3: Presenting research (Lorenzo Crippa)

What are we doing together?

Three videos to help you to think about next year's Capstone dissertation in advance

- Video 1: Getting started and producing effective research (Dr Howard Liu)
- Video 2: Data wrangling and management (Lorenzo Crippa)
- Video 3: Presenting research (Lorenzo Crippa)

Followed by two participative workshops (Q&A structure)

- How to define a research question and think about the research design
- How to find and clean data, and present research results scientifically

Today's topic: Data management

Typical workflow of data management:

1. Find data

Today's topic: Data management

Typical workflow of data management:

1. Find data $\longrightarrow$ 2. Clean data

Today's topic: Data management

Typical workflow of data management:

1. Find data $\longrightarrow$ 2. Clean data $\longrightarrow$ 3. Merge data

Find data

Where can we get data?

The answer really depends on your RQ and RD:

- If your RD is experimental, data will be generated in the implementation phase
- If your RD is observational, we will need to find data sources

Where can we get data?

The answer really depends on your RQ and RD:

- If your RD is experimental, data will be generated in the implementation phase
- If your RD is observational, we will need to find data sources

What should we keep in mind with observational studies?

General advices on gathering data

You should always consider five questions when gathering data:

1. What's the **unit of analysis** of your RD? *I.e.* the **rows** of your dataset. Clarify that before collecting needed information on those units.

General advices on gathering data

You should always consider five questions when gathering data:

1. What's the **unit of analysis** of your RD? *I.e.* the **rows** of your dataset. Clarify that before collecting needed information on those units.
2. What's the **outcome variable** of your design? *I.e.* the variable that you want to explain.

General advices on gathering data

You should always consider five questions when gathering data:

1. What's the **unit of analysis** of your RD? *I.e.* the **rows** of your dataset. Clarify that before collecting needed information on those units.
2. What's the **outcome variable** of your design? *I.e.* the variable that you want to explain.
3. What's the main **treatment variable** of your design? *I.e.* the explanation you provide of the outcome variable

General advices on gathering data

You should always consider five questions when gathering data:

1. What's the **unit of analysis** of your RD? *I.e.* the **rows** of your dataset. Clarify that before collecting needed information on those units.
2. What's the **outcome variable** of your design? *I.e.* the variable that you want to explain.
3. What's the main **treatment variable** of your design? *I.e.* the explanation you provide of the outcome variable
4. What **control variables** do you need information on?

General advices on gathering data

You should always consider five questions when gathering data:

1. What's the **unit of analysis** of your RD? *I.e.* the **rows** of your dataset. Clarify that before collecting needed information on those units.
2. What's the **outcome variable** of your design? *I.e.* the variable that you want to explain.
3. What's the main **treatment variable** of your design? *I.e.* the explanation you provide of the outcome variable
4. What **control variables** do you need information on?
5. Do you need data on these variables **over time**? This will be required by some RDs (*e.g.* panel data designs like FE, DiD...)

Where can I find data I need?

Various sources you can consider. Look at Google Dataset Search.
More specific political science sources include the following.

1. Governmental and Nongovernmental Organizations
    - IMF data
    - OECD
    - EU Open Data Portal
    - UK Data Service
    - US Data.gov
    - World Bank
    - Polity Project
    - Quality of Government
    - Centers for Disease Control and Prevention
    - Uppsala Conflict Data Program
    - AidData

Where can I find data I need? (continues)

2. Collections (replication data)
    - Harvard Dataverse
    - Mendeley Data
    - ICPSR at University of Michigan
    - UK Data Archive
    - Consortium of European Social Science Data Archives
    - Datahub
    - StatSci.org
    - UC Irvine
3. Or you can check the resources that are available through the Catalogue of the Essex library

Clean data

Cleaning dataset

When we clean data, our goal is to get a **tidy** dataset.

Cleaning dataset

When we clean data, our goal is to get a **tidy** dataset.

We want each row to represent a single observation (**depends on what your unit of analysis is!**).

Each column should represent a single variable.

Cleaning dataset

When we clean data, our goal is to get a **tidy** dataset.

We want each row to represent a single observation (**depends on what your unit of analysis is!**).

Each column should represent a single variable.

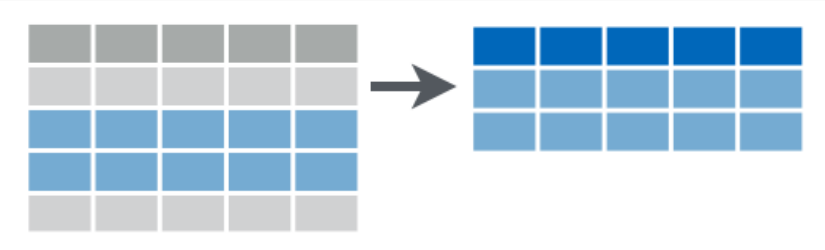
Let's introduce some operations using an ANES dataset (click here to download it)

```
ANES <- read.csv("ANES.csv")
head(ANES)
```

```
> head(ANES)
  year state      FTM     white        poor  turnout
1 1984    NH 73.61539 1.0000000 0.11764706 1.805556
2 1986    NH 76.73333 0.9444444 0.15384616 1.444444
3 1988    NH 66.31035 0.9354839 0.06666667 1.760000
4 1990    NH 72.63265 0.9591837 0.04761905 1.489796
5 1992    NH 50.61111 0.9142857 0.09090909 1.833333
6 1994    NH 52.66667 1.0000000 0.28571430 1.600000
```

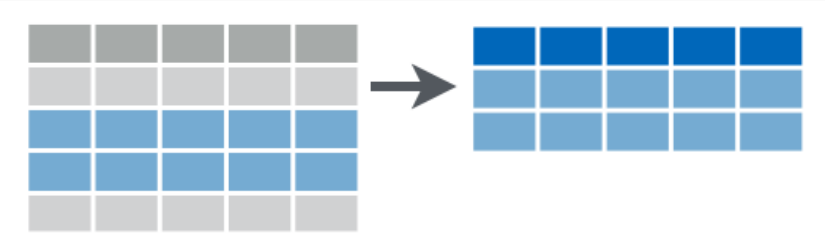
Data cleaning operations: Subsetting rows

We often need to select only some observations from our datasets, and remove observations we don't need:



Data cleaning operations: Subsetting rows

We often need to select only some observations from our datasets, and remove observations we don't need:



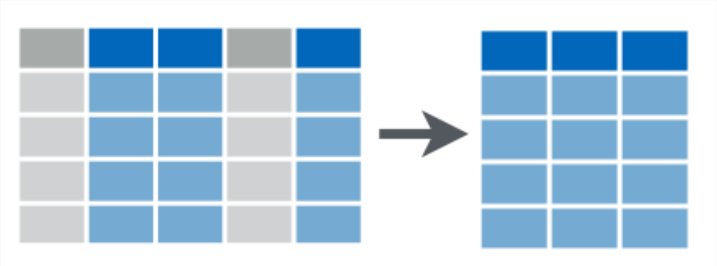
Suppose we wanted to keep only post-1990 observations:

```
# we can use base R:
my_new_data <- ANES[ANES$year > 1990,]

# or tidyverse:
my_new_data <- ANES %>% filter(year > 1990)

head(my_new_data, n = 3)
```

```
> head(my_new_data, n = 3)
  year state      FTM     white       poor  turnout
1 1992    NH 50.61111 0.9142857 0.09090909 1.833333
2 1994    NH 52.66667 1.0000000 0.28571430 1.600000
3 1996    NH 56.94444 1.0000000 0.17647059 1.823529
```

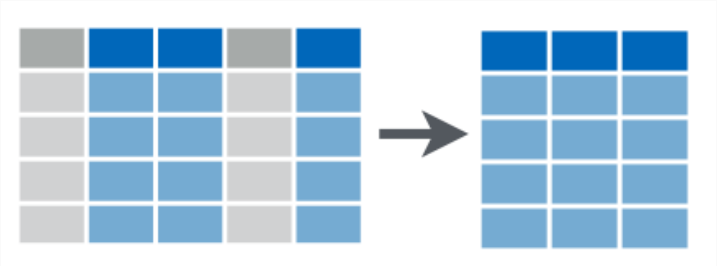
Data cleaning operations: Subsetting columns

Other times, we need to keep only some columns, and remove variables that are not necessary:



Data cleaning operations: Subsetting columns

Other times, we need to keep only some columns, and remove variables that are not necessary:



Suppose we wanted to keep only `year`, `state`, and `white`:

```r
# we can use base R:
my_new_data <- ANES[, c("year", "state", "white")]

# or tidyverse:
my_new_data <- ANES %>% select("year", "state", "white")

head(my_new_data, n = 3)
```

```
> head(my_new_data, n = 3)
  year state     white
1 1984    NH 1.0000000
2 1986    NH 0.9444444
3 1988    NH 0.9354839
```

Data cleaning operations: Creating new variables

We often need to create new variables:

Data cleaning operations: Creating new variables

We often need to create new variables:

Suppose we wanted to create a variable in the new dataset measuring the proportion of non-white population:

```
# we can use base R:
my_new_data$non_white <- 1- my_new_data$white

# or tidyverse:
my_new_data <- my_new_data %>% mutate(non_white = 1-white)

head(my_new_data, n = 5)
```

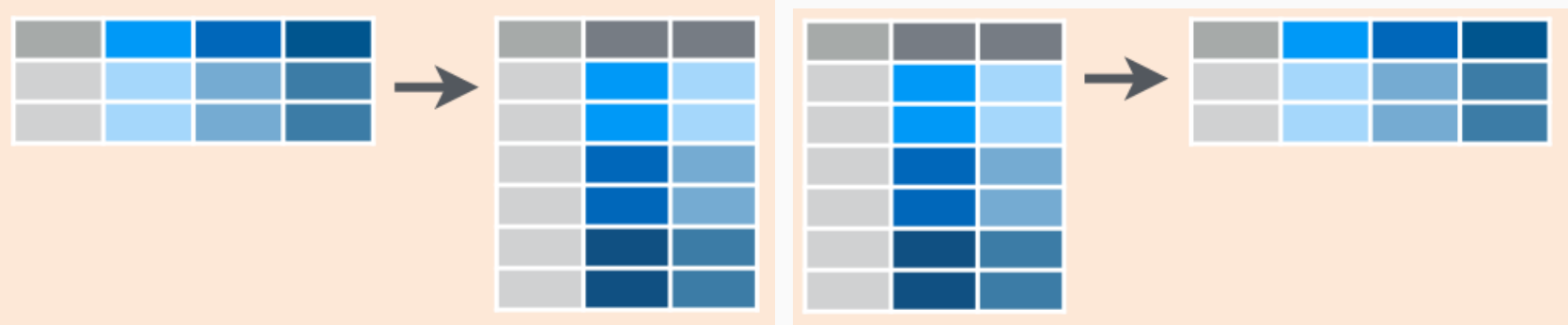
```
head(my_new_data, n = 5)
  year state     white  non_white
1 1984    NH 1.0000000 0.00000000
2 1986    NH 0.9444444 0.05555558
3 1988    NH 0.9354839 0.06451613
4 1990    NH 0.9591837 0.04081631
5 1992    NH 0.9142857 0.08571428
```

Data cleaning operations: Reshaping

With panel data we often need to change wide $\longleftrightarrow$ long panels.

Long: each row is a unit at a point in time.

Wide: each row is a unit, time points are in different columns

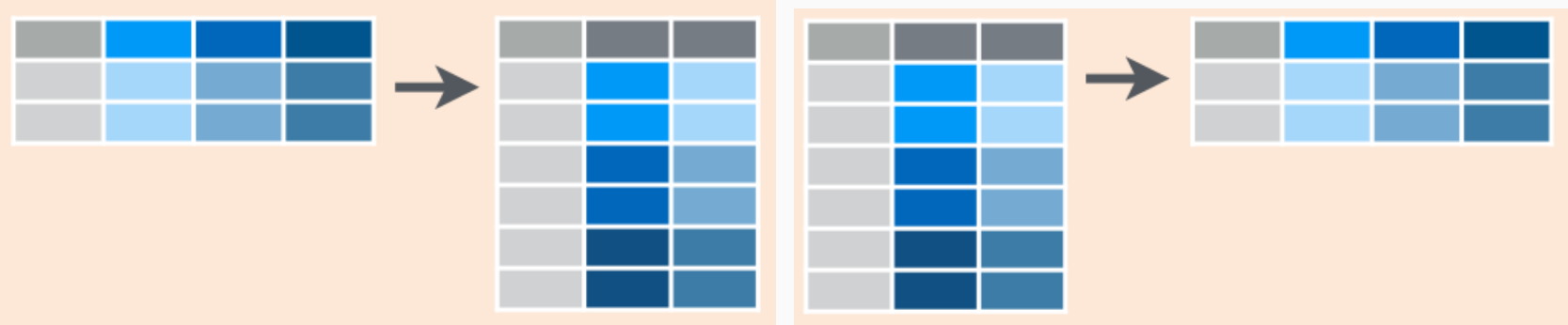


Data cleaning operations: Reshaping

With panel data we often need to change wide $\longleftrightarrow$ long panels.

Long: each row is a unit at a point in time.

Wide: each row is a unit, time points are in different columns



`ANES` is already long. Let's keep only the `year`, `state`, and `white` variables and turn this subset into wide. We'll then turn it back into long format

Reshape wide, then reshape long

Reshape long into wide:

```
# easier using tidyverse:
ANES_wide <- ANES %>%
  select("year", "state", "white") %>%
  pivot_wider(names_from = "year", values_from = "white")
head(ANES_wide[,1:10], n = 3)
```

```
> head(ANES_wide[,1:10], n = 3)
  state `1984` `1986` `1988` `1990` `1992` `1994` `1996` `1998` `2000`
1 NH     1      0.944  0.935  0.959  0.914  1      1      1      0.947
2 NY     0.780  0.707  0.75   0.861  0.766  0.648  0.598  0.719  0.728
3 OH     0.948  0.870  0.908  0.807  0.861  0.897  0.895  0.884  0.871
```

Reshape wide, then reshape long

Reshape long into wide:

```
# easier using tidyverse:
ANES_wide <- ANES %>%
  select("year", "state", "white") %>%
  pivot_wider(names_from = "year", values_from = "white")
head(ANES_wide[,1:10], n = 3)
```

```
> head(ANES_wide[,1:10], n = 3)
  state `1984` `1986` `1988` `1990` `1992` `1994` `1996` `1998` `2000`
1 NH      1      0.944  0.935  0.959  0.914  1      1      1      0.947
2 NY      0.780  0.707  0.75   0.861  0.766  0.648  0.598  0.719  0.728
3 OH      0.948  0.870  0.908  0.807  0.861  0.897  0.895  0.884  0.871
```

Reshape wide into long:

```
ANES_long <- ANES_wide %>%
  pivot_longer(cols = !state, # all columns but the "state" one!
               names_to = "year",
               values_to = "white")
head(ANES_long, n = 3)
```

```
> head(ANES_long, n = 3)
  state year  white
1 NH    1984  1
2 NH    1986  0.944
3 NH    1988  0.935
```

Merge data

Merging datasets

Once we have cleaned all our different data sources, we often want to join them in a single dataset.

In order to merge our data correctly, we'll need unique indicators across our data sources.

Merging datasets

Once we have cleaned all our different data sources, we often want to join them in a single dataset.
In order to merge our data correctly, we'll need unique indicators across our data sources.

For instance, suppose we wanted to join our `ANES` dataset with information on vote shares to Democrats by state-year, in the `ANES_extra` file (click here to download the file.)

Merging datasets: `left_join`

Import the dataset:

```
1 ANES_extra <- read.csv("ANES_extra.csv")
2 head(ANES_extra)
```

```
1 > head(ANES_extra)
2   year state   voteDem       dem
3 1 1984    NH 0.2307692 0.2051282
4 2 1986    NH 0.0000000 0.3333333
5 3 1988    NH 0.1612903 0.1935484
6 4 1990    NH 0.0000000 0.1836735
7 5 1992    NH 0.2777778 0.2500000
8 6 1994    NH 0.0000000 0.1333333
```

Merging datasets: `left_join`

Import the dataset:

```
ANES_extra <- read.csv("ANES_extra.csv")
head(ANES_extra)
```

```
> head(ANES_extra)
  year state   voteDem       dem
1 1984    NH 0.2307692 0.2051282
2 1986    NH 0.0000000 0.3333333
3 1988    NH 0.1612903 0.1935484
4 1990    NH 0.0000000 0.1836735
5 1992    NH 0.2777778 0.2500000
6 1994    NH 0.0000000 0.1333333
```

Merge `ANES` and `ANES_extra`:

```
merged <- ANES %>% left_join(ANES_extra, by = c("state", "year"))
head(merged)
```

```
> head(merged)
  year state      FTM     white       poor  turnout   voteDem       dem
1 1984    NH 73.61539 1.0000000 0.11764706 1.805556 0.2307692 0.2051282
2 1986    NH 76.73333 0.9444444 0.15384616 1.444444 0.0000000 0.3333333
3 1988    NH 66.31035 0.9354839 0.06666667 1.760000 0.1612903 0.1935484
4 1990    NH 72.63265 0.9591837 0.04761905 1.489796 0.0000000 0.1836735
5 1992    NH 50.61111 0.9142857 0.09090909 1.833333 0.2777778 0.2500000
6 1994    NH 52.66667 1.0000000 0.28571430 1.600000 0.0000000 0.1333333
```

Conclusion

To wrap up, when we manage data our workflow is:

1. Collect data
2. Clean them
3. Merge them

An additional useful resource for data management in R is this cheatsheet from `tidyr`.

Thanks for watching!

Thanks for watching this video!

Next video will provide examples and tips to present results.