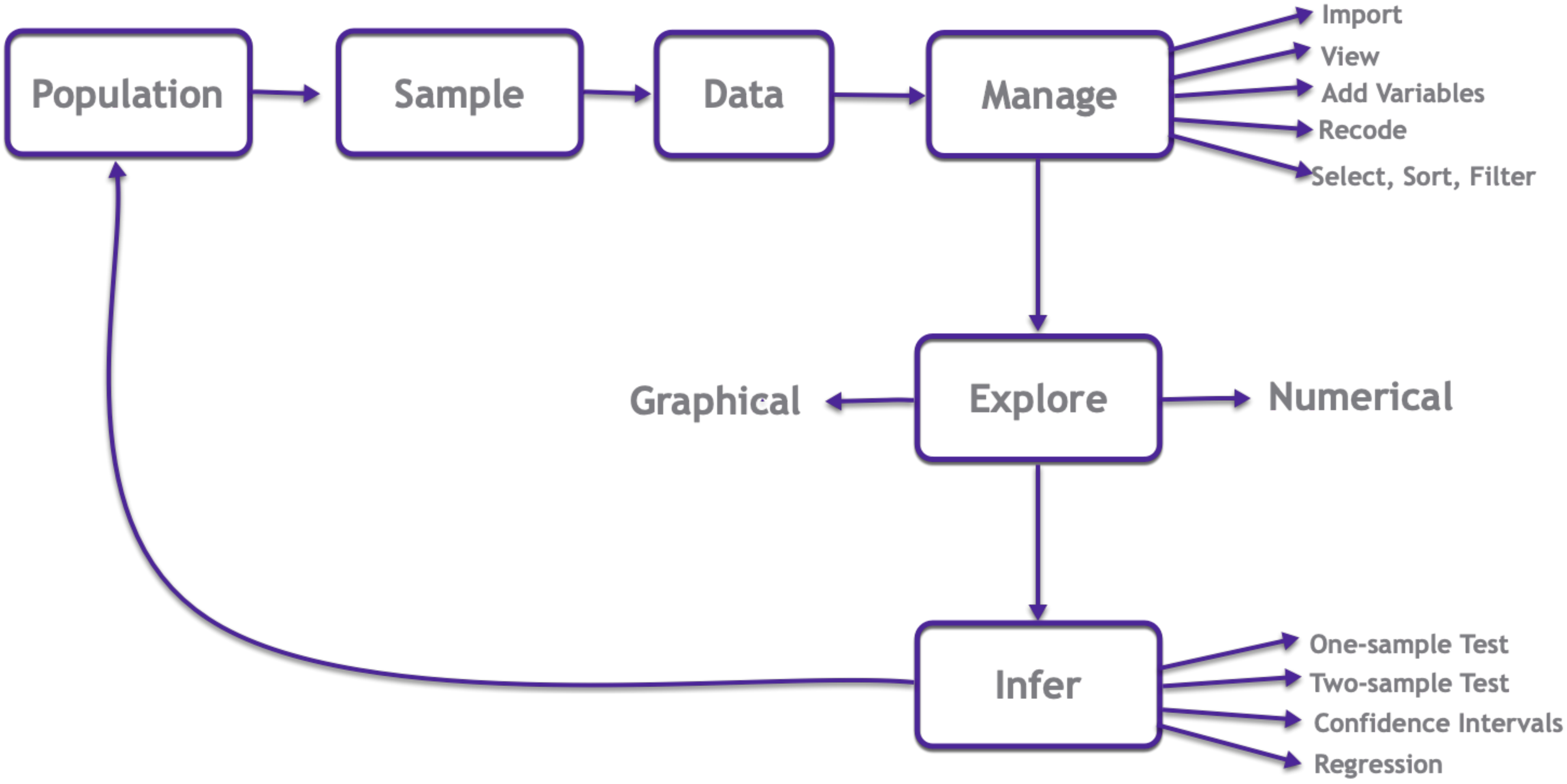
POLSCI 9590: Methods I

Data and Variables

Dave Armstrong

What are we up to in this course?



Videos

We covered a few different things in the videos:

1. Math and order of operations
2. Data and Variables
3. Levels of measurement
    - Qualitative, Categorical, Dichotomous/Polytomous, Discrete: nominal and ordinal
    - Quantitative, Numeric, Continuous: Interval and Ratio

Videos

We covered a few different things in the videos:

1. Math and order of operations
2. Data and Variables
3. Levels of measurement
   - Qualitative, Categorical, Dichotomous/Polytomous, Discrete: nominal and ordinal
   - Quantitative, Numeric, Continuous: Interval and Ratio

Questions?

R and RStudio

R and RStudio are different, but related pieces of software.

- *R* is the statistical software that does all of the "work".
- RStudio is an _I_ntegrated _D_evelopment _E_nvironment (IDE) for R - it is a nice window through which you can interact with R.
    - Knowing *R* is a marketable skill, knowing how to use *RStudio* is less so.

R and RStudio

R and RStudio are different, but related pieces of software.

- *R* is the statistical software that does all of the "work".
- RStudio is an _I_ntegrated _D_evelopment _E_nvironment (IDE) for R - it is a nice window through which you can interact with R.
    - Knowing *R* is a marketable skill, knowing how to use *RStudio* is less so.

We can do two different sorts of things to save our work.

- Write an R script file - this only contains R code and comments.
- Write an RMarkdown file - we can create reproducible reports with RMarkdown.

Packages

Packages are collections of add-on functions that enhance and expand R's base capabilities.

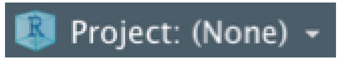
- These are almost entirely user-developed
    - relatively limited quality control.

Packages need to be ...

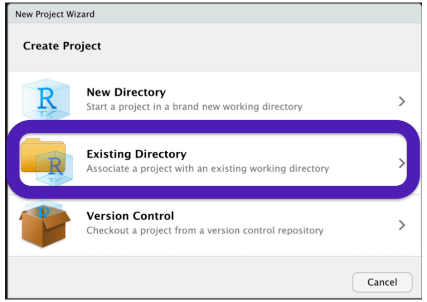
- Downloaded and installed once (per major R version)
- Loaded in every R session before you need to use them.
    - An R session starts when you open the R program and ends when you either quit or reset R.
    - An R session could last days or weeks if you're not someone who quits your open apps and shuts down your computer every night.
    - Resetting R will unload all of the packages you've loaded.

Downloading/Installing and Loading Packages

Create a Project for the Course

1. Somewhere on your computer, make a new folder where you want to save your work.
2. Open up RStudio.
3. In the upper right-hand corner, click on the project button: 

4. Choose "Existing directory":



New Project Wizard

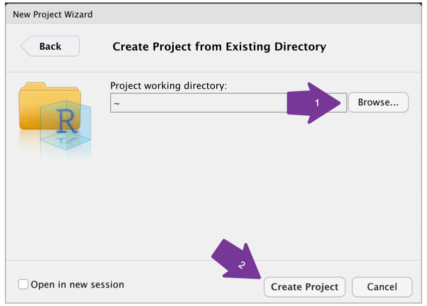
Create Project

New Directory
Start a project in a brand new working directory

Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

5. Click on the "Browse" button and browse to the folder that you just created then click "Create Project"



New Project Wizard

Create Project from Existing Directory

Project working directory:
~

□ Open in new session

Open an R Script

Once your project is open, you can open an R script file by:

- File → New File → R Script.

It will probably be most useful to have a different script file (or RMarkdown file) for every week or day of class.

- The # (hash tag) is the comment character, so anything on the same line after the hash will be disregarded by R.

Importing data

Data can be read in from other data formats - Stats, SPSS, Excel Workbook, CSV, ...

R Python Stata

- To do this, we need the `rio` package.
  - To install this, activate the "Packages" tab in the lower-left panel of RStudio.
  - Click on the "Install" button in the upper right-hand corner of that panel.
  - Type `rio` into the "Packages" text box.
  - Make sure the "Install dependencies" check box is checked.
  - Click "Install" at the bottom of the dialog box.

Import the CES dataset.

1. Download the CES 2019 data from the OWL site.
2. Save (or move) the data file to the folder that contains the project file you made earlier.
3. Import the data. For this to work:
    - **the `ces19.dta` file has to be in the same directory as the .R file you're working on**
    - Or you could give the full or relative path to the data file from the directory.

R   Python   Stata

```r
library(rio)
ces <- import("ces19.dta")
```

Looking at the data.

You can actually look at the data by clicking on the little grid next to the data set name in the "Environment" tab which is in the upper right-hand corner of my RStudio window.

R   Python   Stata

```
str(ces$agegrp)
```

```
##  num [1:2799] 3 1 2 3 1 2 2 2 2 3 ...
##  - attr(*, "label")= chr "Respondent age group"
##  - attr(*, "format.stata")= chr "%12.0g"
##  - attr(*, "labels")= Named num [1:3] 1 2 3
##   ..- attr(*, "names")= chr [1:3] "18-34" "35-54" "55+"
```

```
str(ces$leader_lib)
```

```
##  num [1:2799] 70 70 25 0 70 70 9 55 50 50 ...
##  - attr(*, "label")= chr "Liberal party leader feeling thermometer"
##  - attr(*, "format.stata")= chr "%10.0g"
```

Recognizing Categorical Variables

R   Python   Stata

- Generally things that are "strings" are categorical (ordinal or nominal) variables.
  - These can be made into factors with the `as.factor()` function.
- Variables that have a `factor` class are categorical.
- Variables that have `labels` attribute are often categorical.
  - If you read the data in with `rio`, these can be made into factors with the `factorize()` function.

Often variables with only a few unique values are intended to be categorical. A variable's level of measurement is not an immutable, objective attribute of the data.

- People can think differently about the same variable.

Loading data from R packages

R packages also often come with data. You can import those data into your R session with the `data()` function.

```
data("mtcars", package="datasets")
```

Note, you shouldn't/can't use `import()` and `data()` interchangeably.

- `import()` is for data external to R.
- `data()` is for data internal to R.

Levels of Measurement

Describe the levels of measurement in the following data framses from the `datasets` package:

- `ToothGrotwh`
- `mtcars`

Also, look at the other variables from the `ces19` dataset.

GSS Example

1. Download the `gss16_can.dta` file from the OWL site. This is the Canadian General Social Survey data from 2016.
2. Write code to import the data into your R session.
3. There are three variables - describe each of them.
4. What the level of measurement is each variable?

Mathematical Functions in R.

- `+`, `-`, `*` and `/` are the operators for add, subtract, multiply and divide, respectively.
- `log()` is the natural logarithm (with base e), `log10()` is the log base 10 function.
- `exp(x)` does e^x and the caret `^` is the "to-the-power-of" operator, so `4^2=16`.
- `sqrt()` is the square root function. More generally you can take any root r by raising a value to the power $\frac{1}{r}$.