# First Steps in R

7 October 2020
EARNConversations

Ben Zipperer
Economic Policy Institute

bzipperer@epi.org
@benzipperer

# Materials for this class

*https://economic.github.io/data_bootcamp/*

# Big picture outline for today

1. R/RStudio basics
2. Analyze simple data
   - national wage percentiles, by race
3. Analyze complex data
   - ACS microdata
   - calculate demographic profile of low-wage workers in Virginia
4. Basic programming in R

# R/RStudio basics: tasks

We will learn

- the layout of R/Rstudio
- some very basic R commands and functions
- how to store results in R

# R basics: review

- R is essentially a very fancy calculator
- R uses functions (commands). Functions
    - have a name
    - have inputs (arguments) in parentheses
    - have an output (object)
    - can be nested
    - are described in help files: `?function`
- We store objects with assignment arrow: `<-`

# Analyze simple data: tasks

- Let's look at national hourly wage percentiles over time, by race
    - easily accessible from EPI:
    - provided to you as .csv file: `epi_wage_percentiles.csv`
- We will use R to load and manipulate this data

# Analyze simple data: review

Workflow: load data, manipulate it, and save output

`read_csv(”filename.csv”)` loads csv file

`select(data, column1, column2, ...)` keeps *column1, column2, ...*

`filter(data, condition)` keeps rows satisfying *condition*

`arrange(data, column1, column2, ...)` sorts rows according to *column1, column2, ...*

`mutate(data, column = ...)` change or create *column* according to the rule ...

`write_csv(”filename.csv”)` save resulting data as csv file

# Analyze complex data: tasks

- Let’s calculate the share of workers who earn low wages in Virginia
- We will need microdata with wage and state information
- A good candidate for this is the American Community Survey
  - easily accessible via IPUMS: *`https://usa.ipums.org/`*
  - 2018 ACS provided to you in Stata format: *`acs_2018.dta`*

# Analyze complex data: review

`haven::read_dta("filename.dta")` loads Stata data file

`count(data, var1, var2, ...)` tabulates *var1, var2, ...*

`if_else(condition, true, false)` provides value *true* and *false* according to *condition*

`summarize(data, function)` provides summary statistic outputted by *function*

`mean(var)` and `weighted.mean(var, w = weight)` calculate means of *var*

# Basic R programming: what and why?

- We just learned how to do data analysis in R *interactively*
- In general you should write and run R scripts
- An R script will
    - provide a fully documented record of your work
    - allow you to tweak or extend your analysis more easily
    - aid replication by others (and yourself!)

# Review

Today we learned to

1. Load and use R/RStudio
2. Analyze simple data: national wage percentiles, by race
3. Analyze complex data: profile of low-wage workers in Virginia
4. Code in R
    - always write and run R scripts
    - add comments to document your work
    - write better R code with the pipe: `%>%`
    - use packages

# Preview of next week's R session

Using R effectively: Tuesday, October 13, 3:00 pm - 4:30 pm

- reshape
- combine data (bind and join)
- directory/project management