Intro to the Tidyverse: A Collection of R Packages

Amit Talapatra

https://cran.r-project.org/

Setup

1. Download and Install R: https://www.r-project.org/
2. Download and Install RStudio: https://rstudio.com/
3. Run these lines in RStudio to install the packages we'll be using:

```
install.packages("unvotes")
install.packages("WDI")
install.packages("tidyverse")
install.packages("ggraph")
install.packages("igraph")
install.packages("tidygraph")
install.packages("widyr")
```

4. Download this Rmarkdown file: https://github.com/atalapatra/Intro-to-the-Tidyverse/blob/master/Intro%20to%20the%20Tidyverse.Rmd

Tidyverse Resources

- Tidyverse Website
- David Robinson's 'Teach the Tidyverse to Beginners'
  - I attended David's workshop at the DC R 2019 Conference. I'll be using some of his materials and examples for this talk.
- Hadley Wickam's Paper "Tidy Data"
  - This paper explains the founding principles on Tidy Data. Hadley Wickam is the primary developer of packages in the Tidyverse.

https://twitter.com/W_R_Chase/status/1155215530997035014

What is the tidyverse?

> *"The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures."*
>
> - From https://www.tidyverse.org/

The tidyverse is based on the concept of "Tidy Data", which follows these principles:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

This is a version of Codd's Third Normal Form

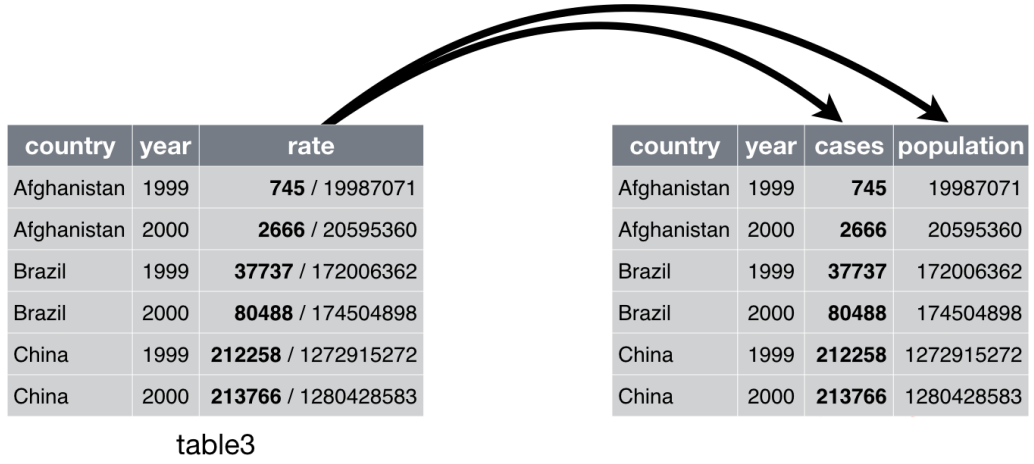
https://en.wikipedia.org/wiki/Third_normal_form#Definition_of_third_normal_form
https://codeblab.com/wp-content/uploads/2009/12/rmdb-codd.pdf

What does tidy data look like?

Here are some examples of transformations that make data tidy.

Principles of “Tidy Data”:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.



country	year	rate
Afghanistan	1999	**745** / 19987071
Afghanistan	2000	**2666** / 20595360
Brazil	1999	**37737** / 172006362
Brazil	2000	**80488** / 174504898
China	1999	**212258** / 1272915272
China	2000	**213766** / 1280428583

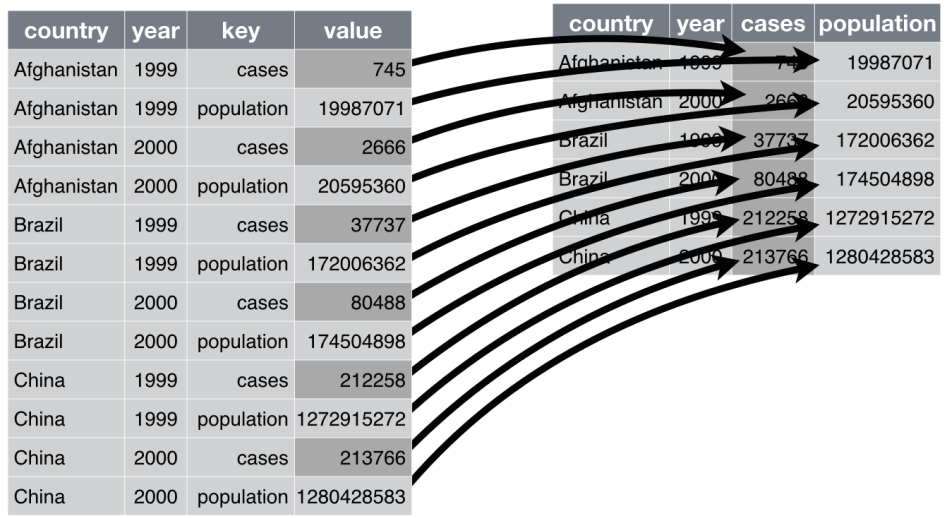
country	year	cases	population
Afghanistan	1999	**745**	19987071
Afghanistan	2000	**2666**	20595360
Brazil	1999	**37737**	172006362
Brazil	2000	**80488**	174504898
China	1999	**212258**	1272915272
China	2000	**213766**	1280428583

table3

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4



country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table2

https://r4ds.had.co.nz/tidy-data.html

Why learn about the Tidyverse?

- Tidy data introduces concepts that are useful to anyone working with data.

Why learn about the Tidyverse?

- Tidy data introduces concepts that are useful to anyone working with data.
- Some people love it.

https://twitter.com/rstatsmemes

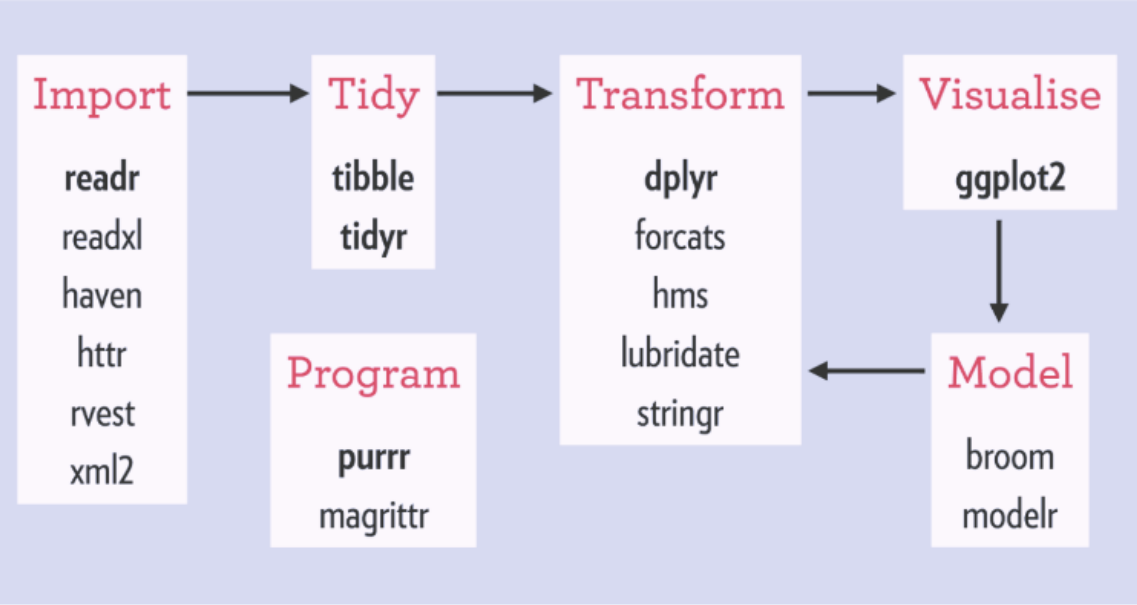
Why learn about the Tidyverse?

- Tidy data introduces concepts that are useful to anyone working with data.
- Some people love it.
- You might love it.

https://twitter.com/rstatsmemes

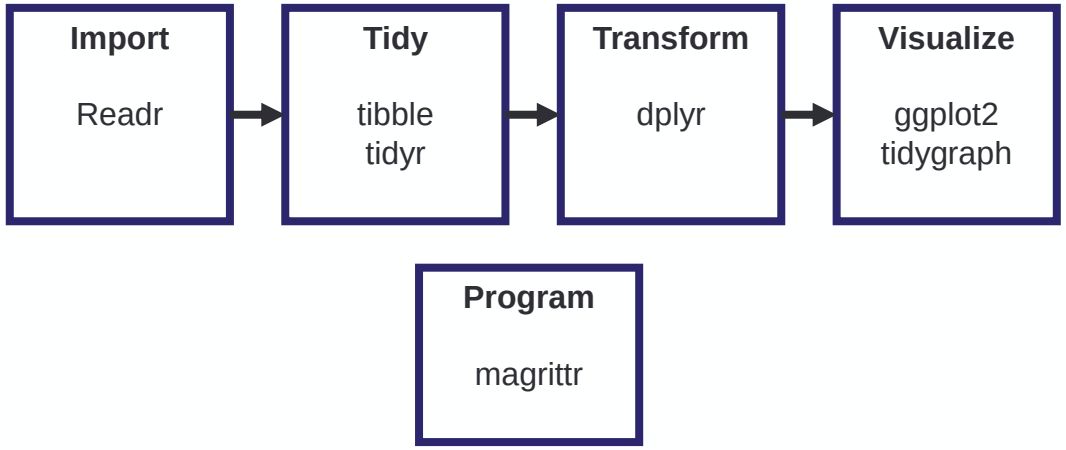
What we'll cover

Where tidyverse packages fit into the data analysis process (from David Robinson):



https://github.com/rstudio/rstudio-conf/blob/master/2018/TeachTidyverse-DavidRobinson/TeachTidyverse-DavidRobinson.pdf

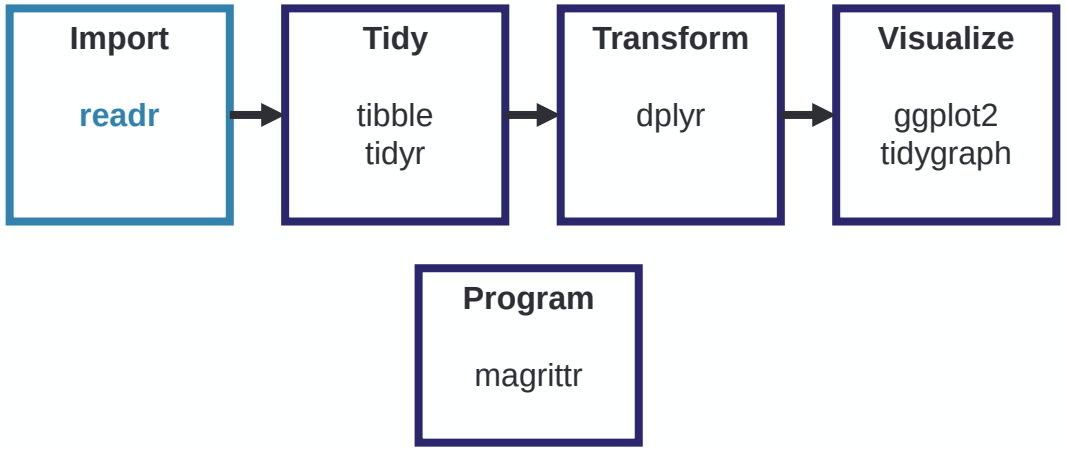
What we'll cover:

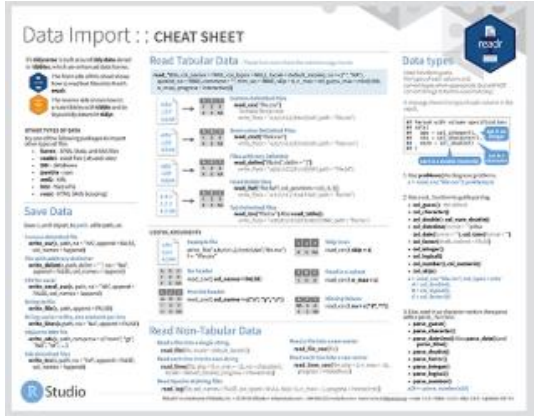


https://twitter.com/tslumley/status/1008799349998903299

readr

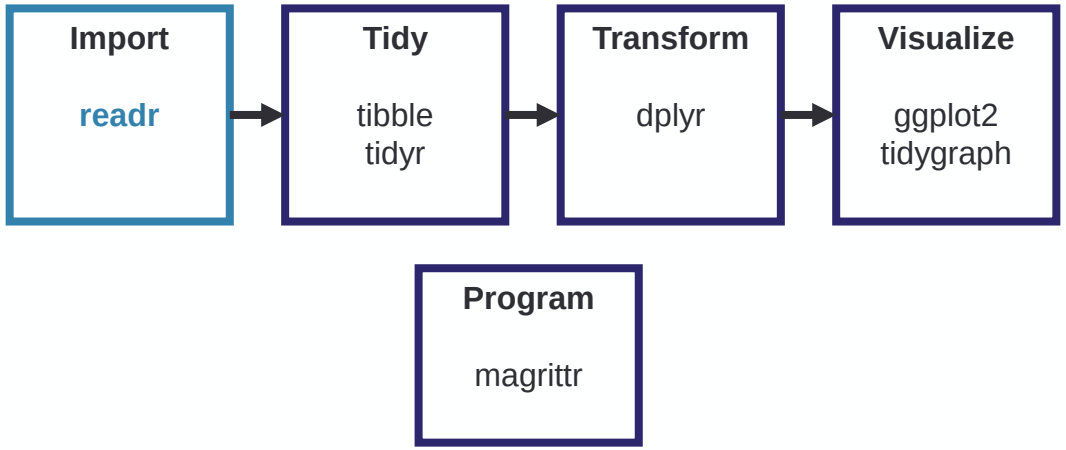
- A fast way to read rectangular data into tibbles – a type of enhanced data frame designed for tidy data.
- readr can also guess column data types and parse columns appropriately (date, int, double, etc.).





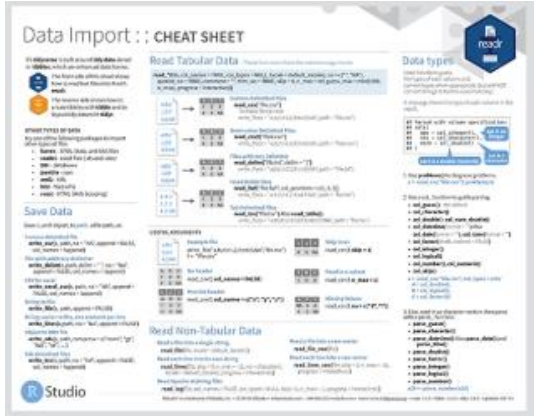
readr

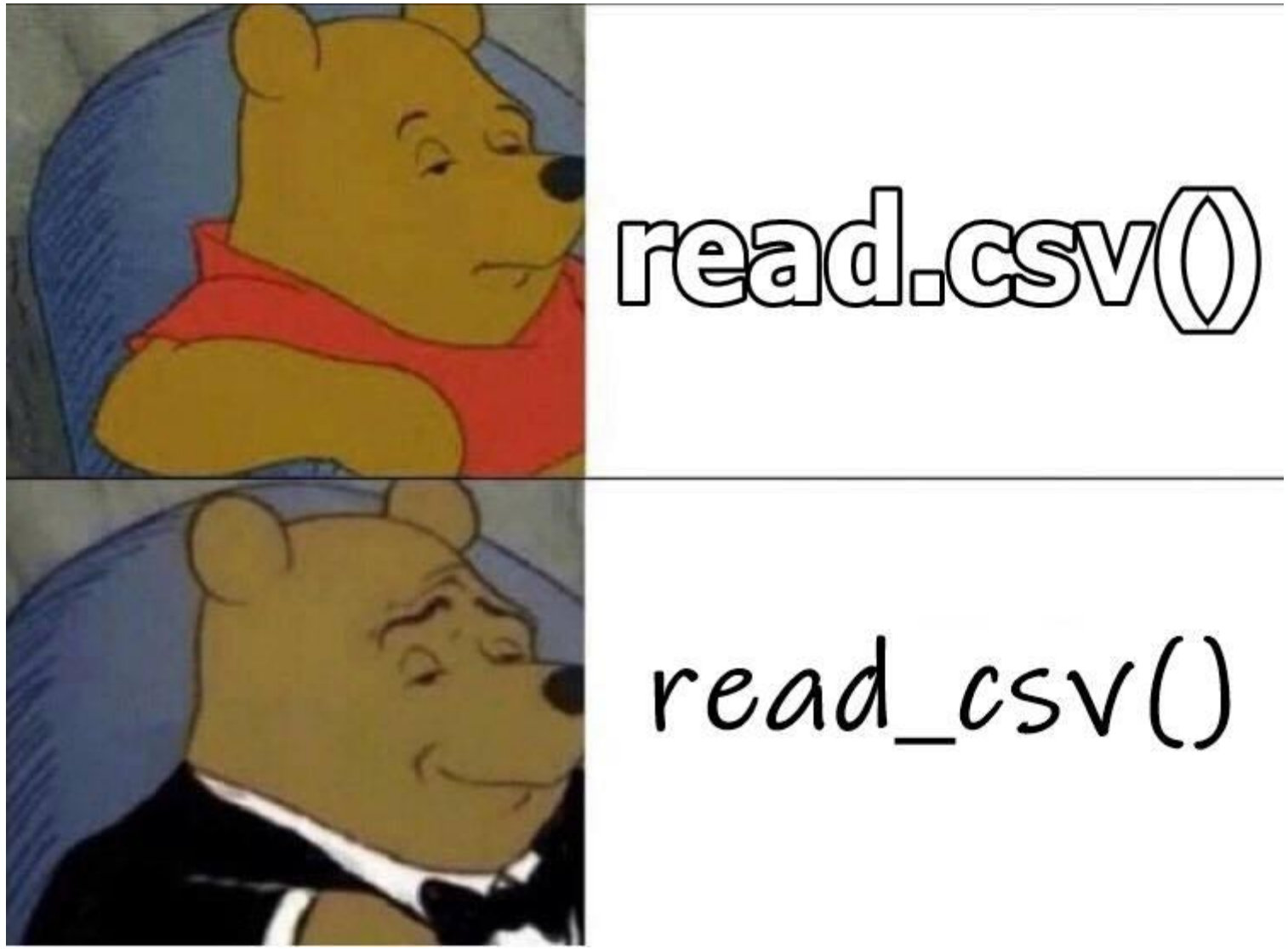
- A fast way to read rectangular data into tibbles – a type of enhanced data frame designed for tidy data.
- readr can also guess column data types and parse columns appropriately (date, int, double, etc.).



Principles of “Tidy Data”:

1. **Each variable forms a column.**
2. Each observation forms a row.
3. Each type of observational unit forms a table.

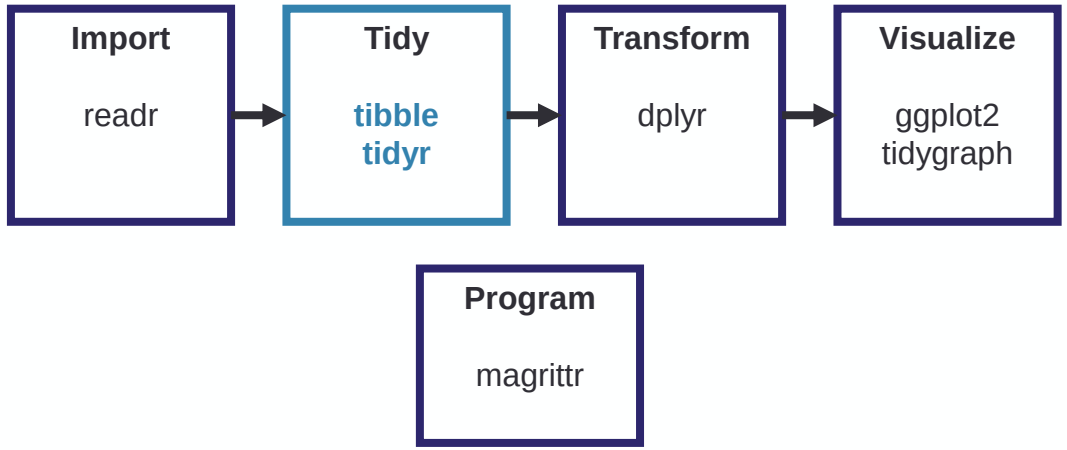




https://twitter.com/favstats/status/1112060496977510400

tibble and tidyr

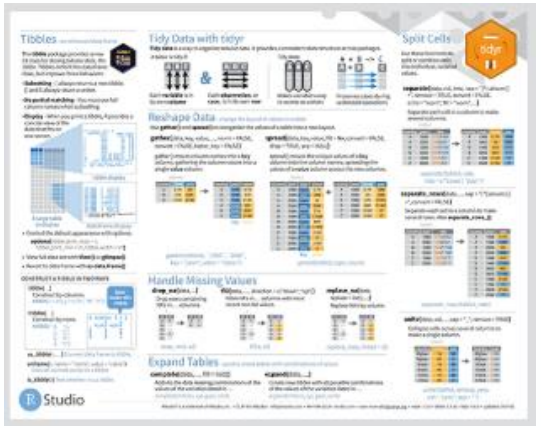
- The tibble package gives us the tibble object. Tibbles are data frames with improvements to subsetting, viewing data, and defining data types.
- tidyr provides functions for data transformations that convert messy data into tidy data.



country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table2



https://r4ds.had.co.nz/tidy-data.html

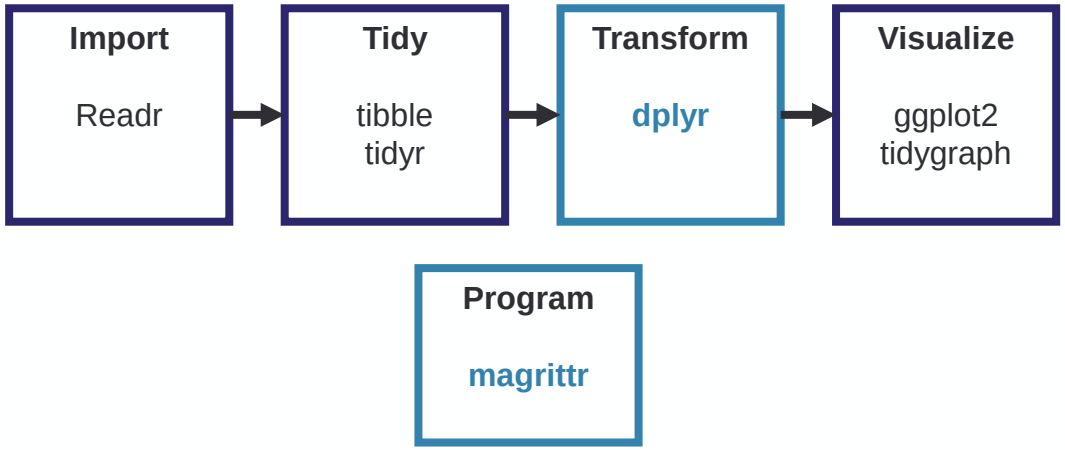
magrittr and dplyr

magrittr

- Pipes (%>%): https://uc-r.github.io/pipe

dplyr:

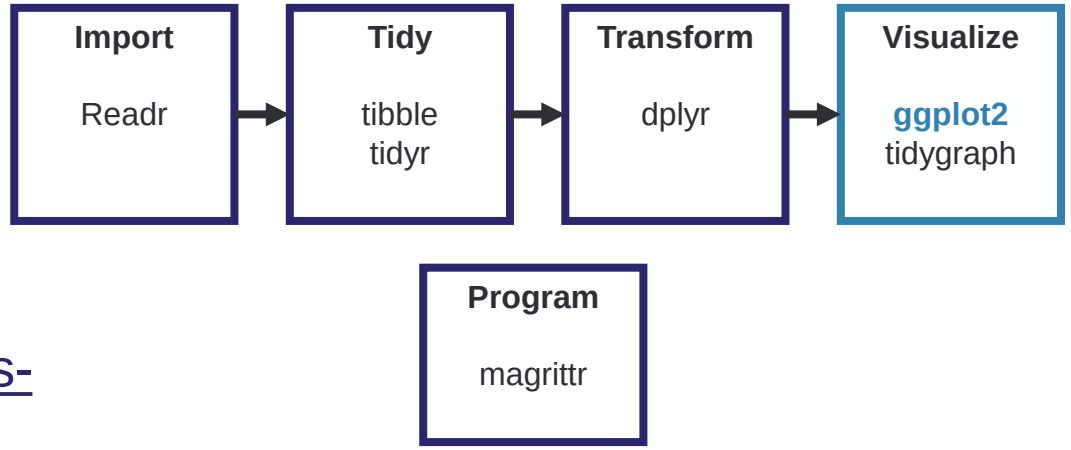
- Functions (https://dplyr.tidyverse.org/):
  - mutate() adds new variables that are functions of existing variables
  - select() picks variables based on their names.
  - filter() picks cases based on their values.
  - summarise() reduces multiple values down to a single summary.
  - arrange() changes the ordering of the rows.

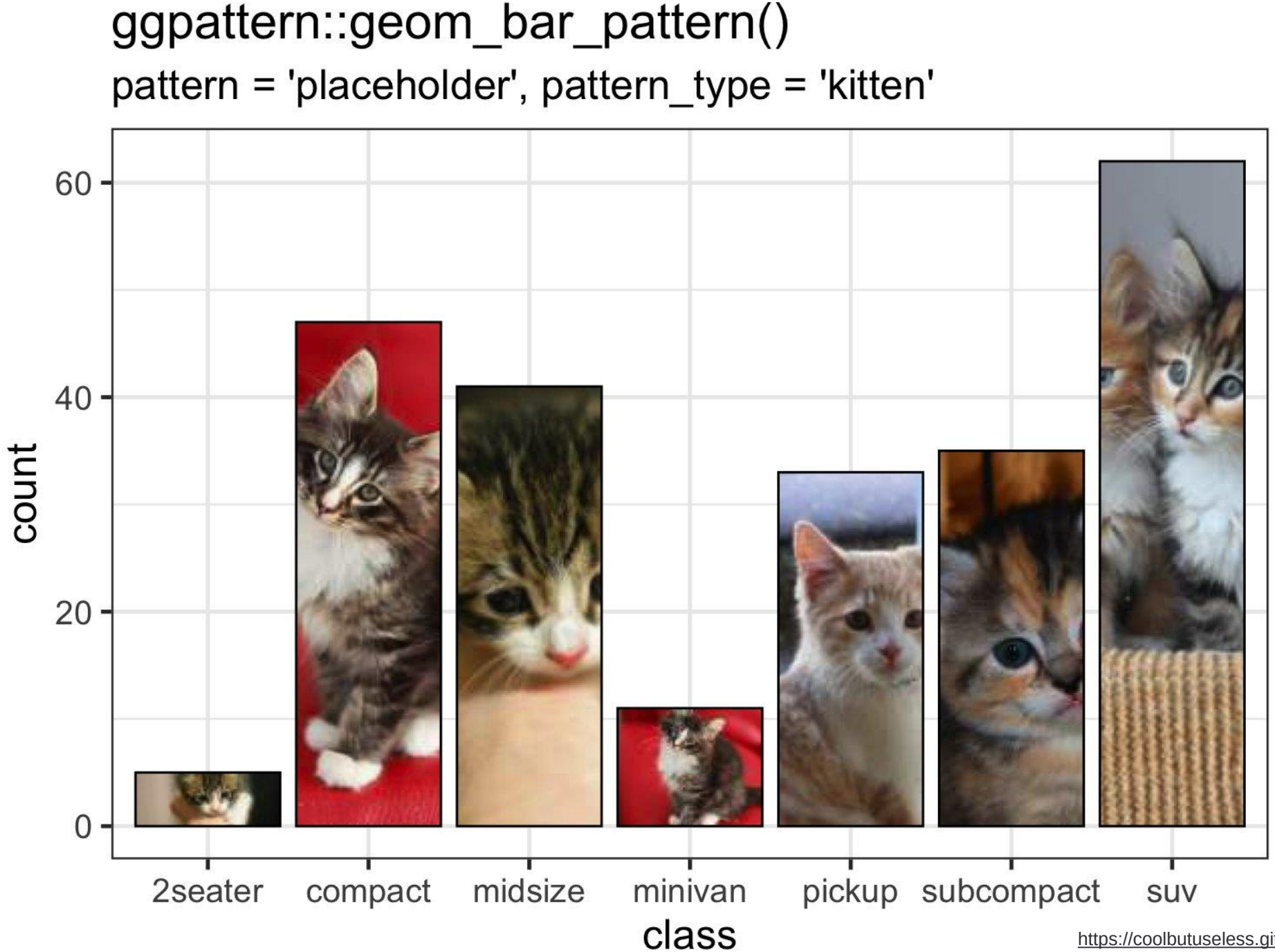


https://twitter.com/rstatsmemes

ggplot2

- General information on ggplot2: https://ggplot2.tidyverse.org/reference/ggplot.html
- ggplot2 cheat sheet: https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf
- Leland Wilkinson's The Grammar of Graphics https://www.amazon.com/Grammar-Graphics-Statistics-Computing/dp/0387245448
- Hadley Wickam's A Layered Grammar of Graphics https://vita.had.co.nz/papers/layered-grammar.html
- ggplot2 incorporates a wide range of data visualization options: http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

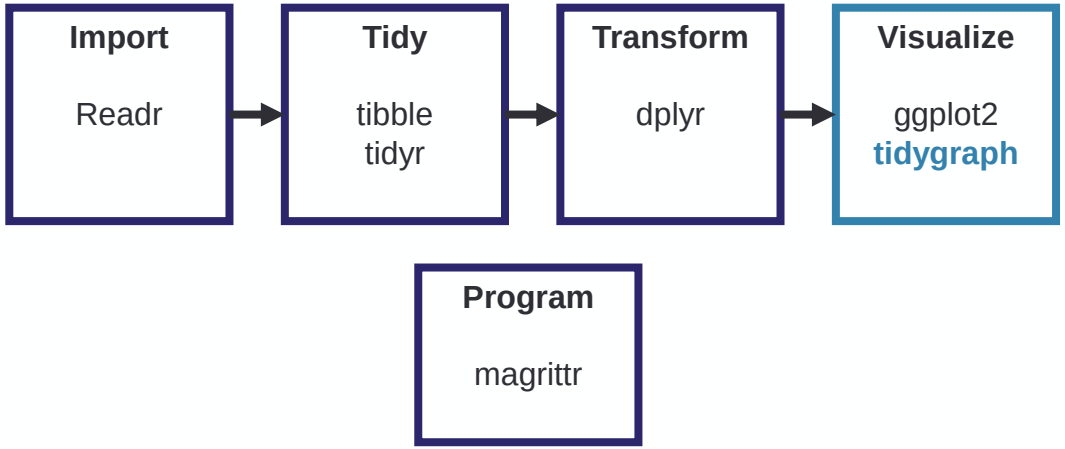


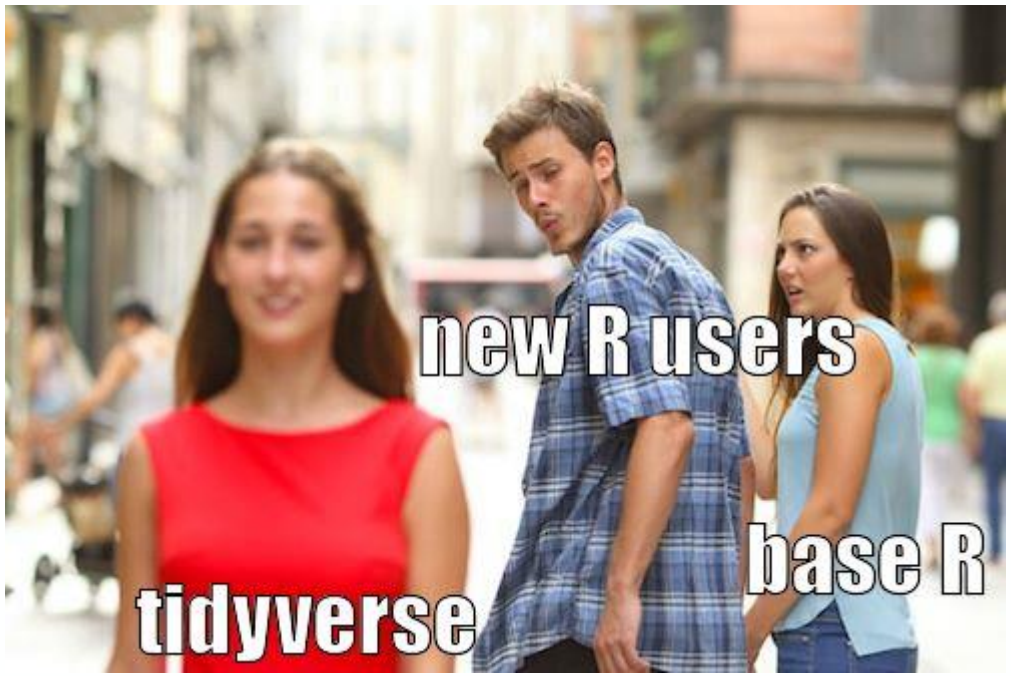


https://coolbutuseless.github.io/package/ggpattern/articles/geom-gallery-array.html#geom-bar-pattern-

tidygraph

- Network graph data is stored as edges and nodes, which is not a traditional format for “tidy” data.
- Tidygraph enables users to manipulate graph data using dplyr. Graph data can then be plotted using traditional network graph packages (e.g. ggraph, igraph).
- More info on tidygraph: https://www.data-imaginist.com/2017/introducing-tidygraph/





https://twitter.com/girsanov/status/1202389699958050818

Additional Sources

- R code examples used for this talk make use of David Robinson's 'unvotes' package, which uses data from the following source:
  - Erik Voeten "Data and Analyses of Voting in the UN General Assembly" Routledge Handbook of International Organization, edited by Bob Reinalda (published May 27, 2013)
- R memes courtesy of: https://twitter.com/rstatsmemes