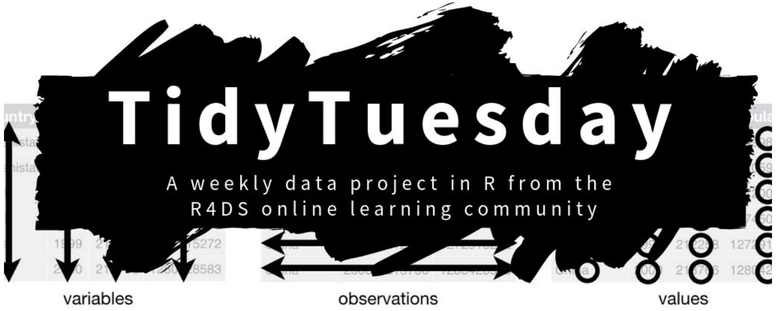
Fantastic data and where to find them

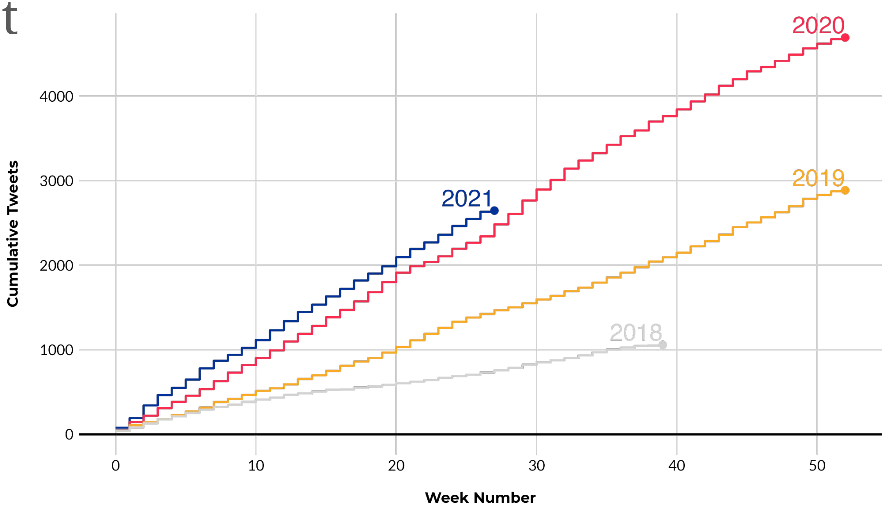
Tom Mock, RStudio

Lerman via @unsplash

What I know about data

- Managing the **#TidyTuesday** project for > 3 years
- Exclusively using public data sources
- Finding, cleaning, and sharing datasets for visualization and project based learning with R
- ~200 unique visitors/day to repo
- ~100 participants/week on Twitter


TidyTuesday
A weekly data project in R from the R4DS online learning community
variables
observations
values



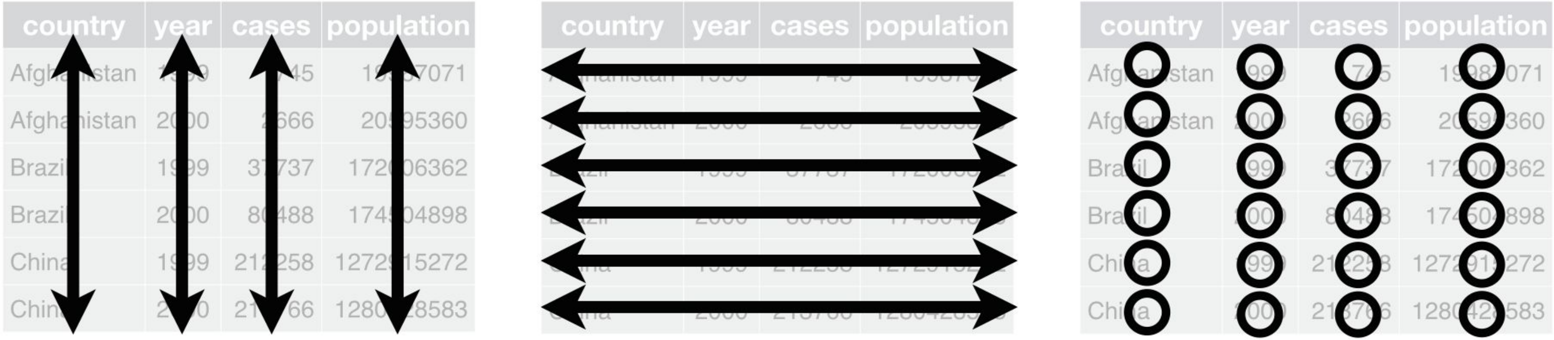
Cumulative tweets for #TidyTuesday by year

Note that Week 1 of 2018 started in April & tweets must contain: 'rstats, code, plot, graph, viz, data or tidyverse'

Data: rtweet | Plot: @thomas_mock

“Tidy” data

- Much of the data you can find publicly or on the web is not analysis ready
- “Tidy” data is “*structuring datasets to facilitate analysis*”
- The R for Data Science book covers this in greater depth



variables

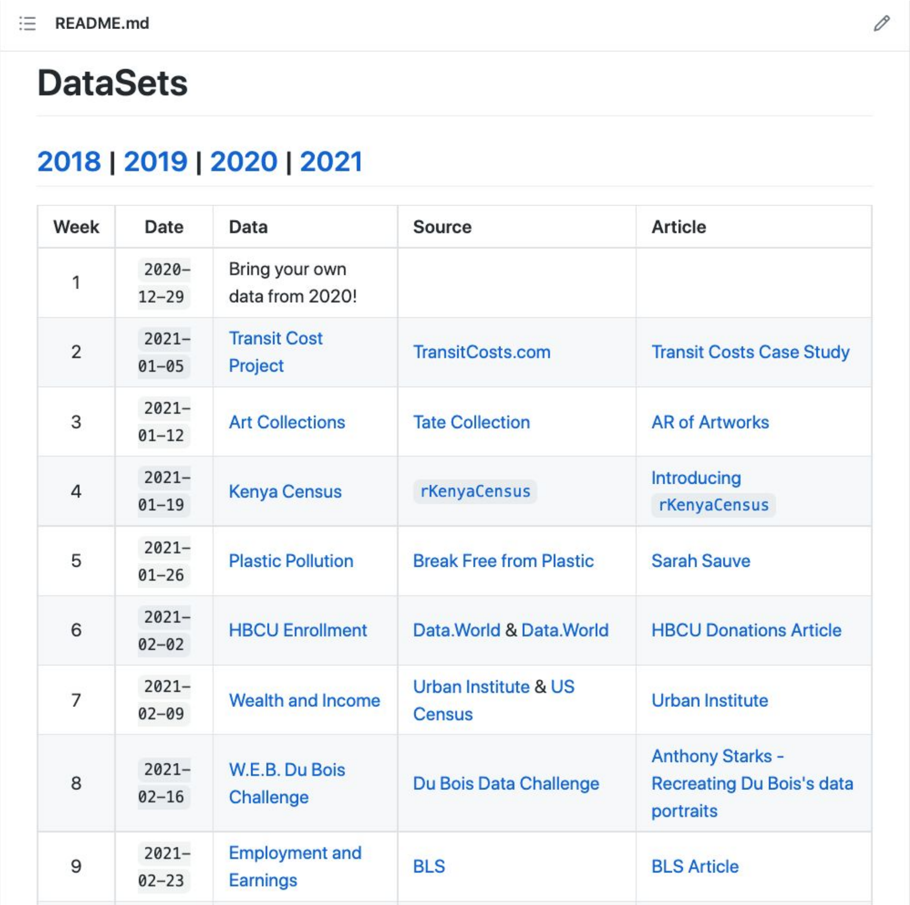
observations

values

What is #TidyTuesday?

A **scaffold**
for a **self-directed**
community of practice

#TidyTuesday



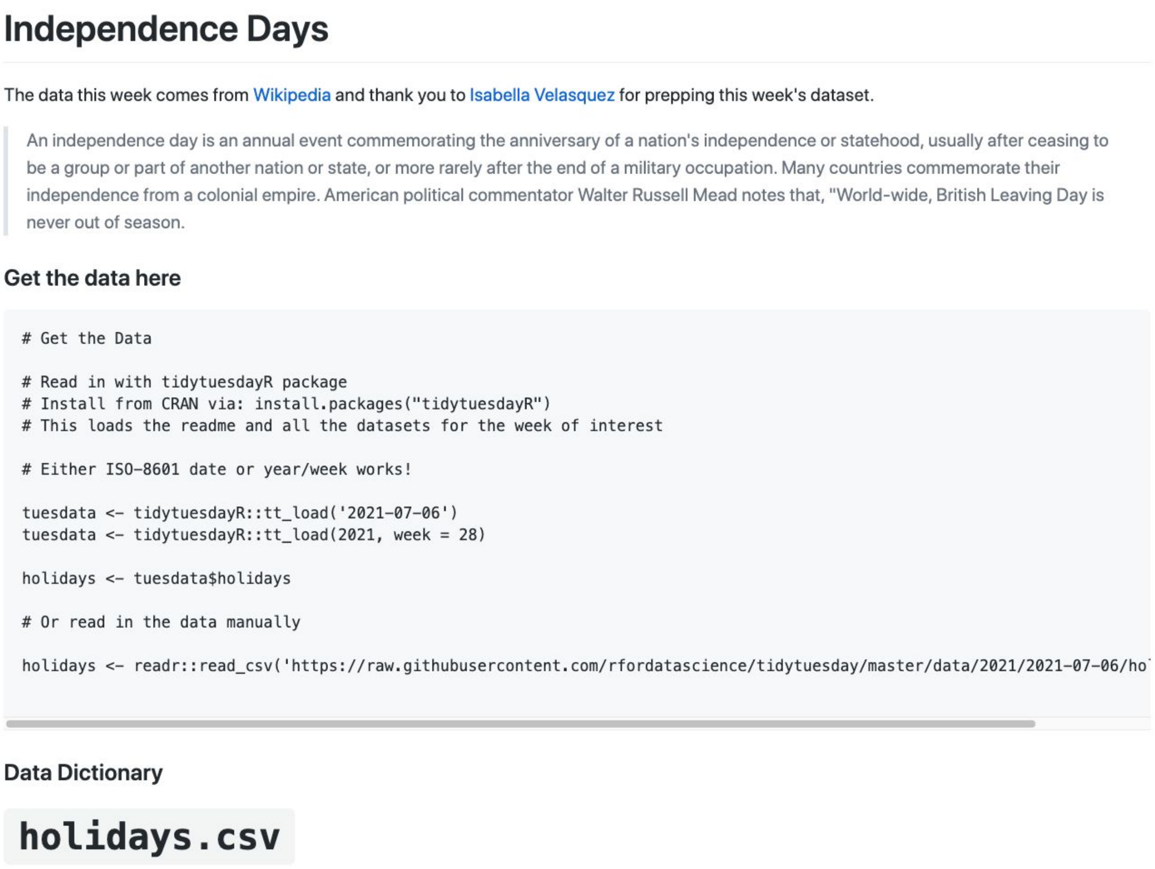
README.md

DataSets

2018 | 2019 | 2020 | 2021

Week	Date	Data	Source	Article
1	2020-12-29	Bring your own data from 2020!		
2	2021-01-05	Transit Cost Project	TransitCosts.com	Transit Costs Case Study
3	2021-01-12	Art Collections	Tate Collection	AR of Artworks
4	2021-01-19	Kenya Census	rKenyaCensus	Introducing rKenyaCensus
5	2021-01-26	Plastic Pollution	Break Free from Plastic	Sarah Sauve
6	2021-02-02	HBCU Enrollment	Data.World & Data.World	HBCU Donations Article
7	2021-02-09	Wealth and Income	Urban Institute & US Census	Urban Institute
8	2021-02-16	W.E.B. Du Bois Challenge	Du Bois Data Challenge	Anthony Starks - Recreating Du Bois's data portraits
9	2021-02-23	Employment and Earnings	BLS	BLS Article

#TidyTuesday



Independence Days

The data this week comes from Wikipedia and thank you to Isabella Velasquez for prepping this week's dataset.

> An independence day is an annual event commemorating the anniversary of a nation's independence or statehood, usually after ceasing to be a group or part of another nation or state, or more rarely after the end of a military occupation. Many countries commemorate their independence from a colonial empire. American political commentator Walter Russell Mead notes that, "World-wide, British Leaving Day is never out of season.

Get the data here

```
# Get the Data

# Read in with tidytuesdayR package
# Install from CRAN via: install.packages("tidytuesdayR")
# This loads the readme and all the datasets for the week of interest

# Either ISO-8601 date or year/week works!

tuesdata <- tidytuesdayR::tt_load('2021-07-06')
tuesdata <- tidytuesdayR::tt_load(2021, week = 28)

holidays <- tuesdata$holidays

# Or read in the data manually

holidays <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-07-06/ho
```

Data Dictionary

`holidays.csv`

So what is TidyTuesday?

It’s just data.

That can be read into R in a few seconds

With an article for context

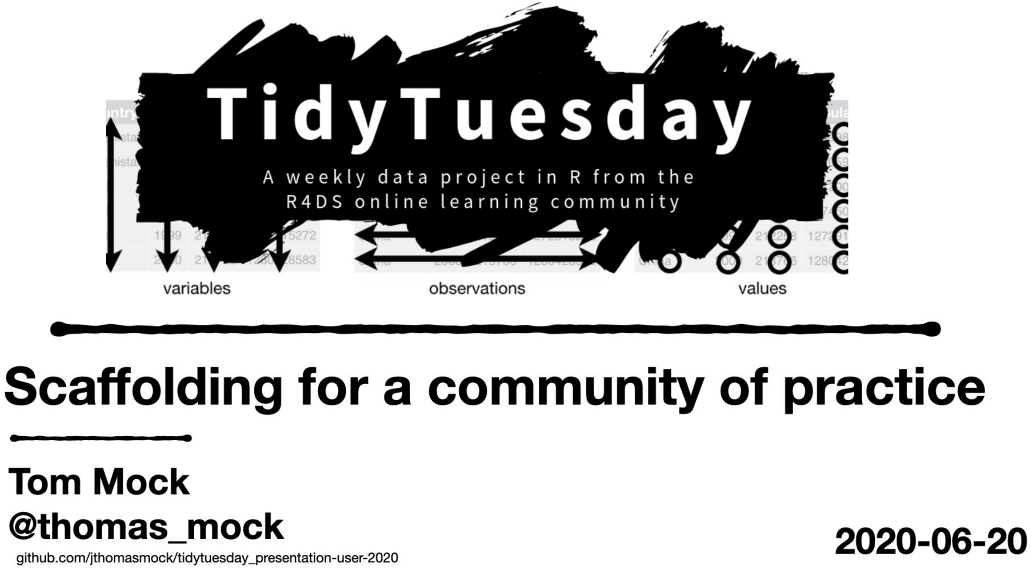
With a data dictionary

With dozens of example analyses WITH code very week

With a community of learners and mentors

bitly.com/tidyreadme

More about TidyTuesday


TidyTuesday
A weekly data project in R from the R4DS online learning community
variables
observations
values

Scaffolding for a community of practice

Tom Mock
@thomas_mock
github.com/jthomasmock/tidytuesday_presentation-user-2020

2020-06-20

https://www.youtube.com/watch?v=H8rhQOj7vDY

Government Data

- NHS efforts to support open data
- The UK open data portal data.gov.uk
- GovLab NHS Open Data portal
- US data via data.gov
- State-level data for US via data.texas.gov

Scientific Data

- Machine learning datasets via Papers with Code
- Public science data from Google Datasets
- Large datasets via Open Science Data Cloud
- Scientific data metacollections via Nature

Data Aggregators

- OurWorldInData.org - indicator data for most countries
- Machine Learning tasks with open data via Kaggle
- Data.World data catalogue
- Data-is-plural.com weekly newsletter
- Historical data from Wikipedia

Web scraping

- The `rvest` package allows for tables that can’t be “downloaded” to be imported directly into R
- The `polite` R package goes a step farther in asking permission to scrape
- JavaScript based web scraping from open API endpoints with `httr`

Your own data

- Your team/company/organization has data
- It probably needs to be visualized, to be explored, to be analyzed