Discovering our Community through Data

Where can I find data?

Provost Academy
Day 1, Morning
August 16, 2021

What topic are you interested in?

This slide deck includes links to various data repositories and sources.

Some are general and include data associated with many different fields and areas of study.

Others have data associated with a specific topic.

We will not go through each link. The links are provided to demonstrate just how much is available!

Census data via tidycensus

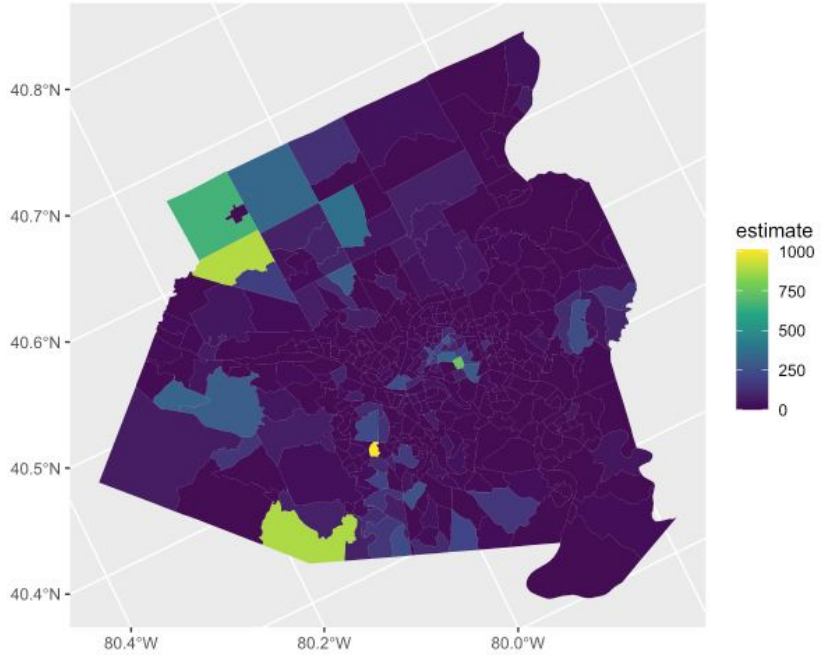
Link: [Load US Census Boundary and Attribute Data as tidyverse and sf-Ready Data Frames • tidycensus (walker-data.com)]

`tidycensus` provides Application Programming Interfaces (APIs) to data from the US Census Bureau census conducted every 10 years, and the American Community Survey (ACS) conducted every 5 years.

Example: population estimates within census tracts

Figure to the right shows the estimate of Indian expats living within Allegheny county.

Figure created as part of a final project in CMPINF 2130 Summer 2021.



Covid-19 related data

NY Times: nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. (github.com)

`tidycovid19`: Download, Tidy and Visualize Covid-19 Related Data • tidycovid19 (joachim-gassen.github.io)

- The tidycovid19 package provides access to numerous data sources associated with Covid-19.
- The documentation includes links to the various data sources.
- A web-based application for exploring the data overtime:
    - Explore the Spread of Covid-19 (shinyapps.io)

Text data - books/songs/speech

The tidytext book is a great resource if you want to learn about working with text data!

Analyzing the text in the Harry Potter book series:

- Text data source: bradleyboehmke/harrypotter: An R Package for the Harry Potter Book Series (github.com)
- Example: Text Mining: Creating Tidy Text · UC Business Analytics R Programming Guide (uc-r.github.io)

Project Gutenberg - available public domain texts: Free eBooks | Project Gutenberg

- R package: ropensci/gutenbergr: Search and download public domain texts from Project Gutenberg (github.com)

US Energy Information Agency (EIA)

Useful resource if you are interested to learn about the energy and especially electricity markets and distribution in the US.

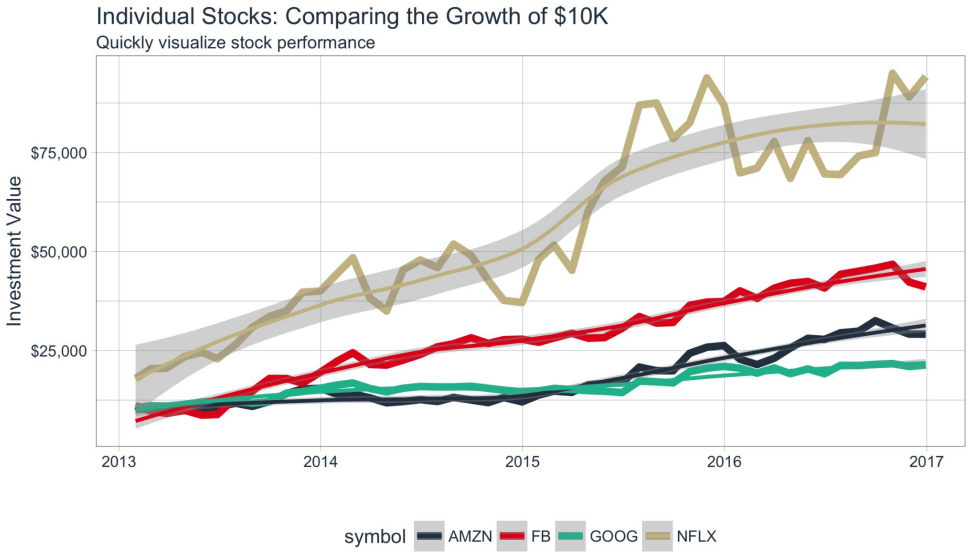
Website: Homepage - U.S. Energy Information Administration (EIA)

US Hourly electric grid dashboard: Real-time Operating Grid - U.S. Energy Information Administration (EIA)

Stocks and financial data

The `tidyquant` package provides tools for charting/visualizing stock prices, comparing companies, and comparing investment strategies!

Github page with links to examples, YouTube tutorials, and labs.



Sports related data

Baseball: Functions for acquiring and analyzing baseball data • baseballr (billpetti.github.io)

NFL: An R package to quickly obtain clean and tidy NFL play by play data • nflfastR

NHL: MoneyPuck.com -Download Datasets

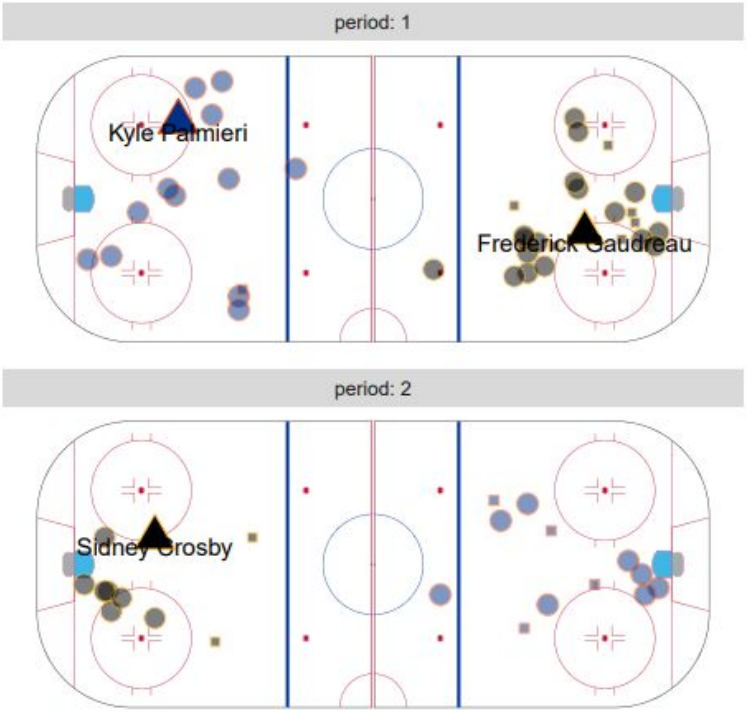
NBA: abresler/nbastatR: NBA Stats API Wrapper and more for R (github.com)

WNBA: An R package to quickly obtain clean and tidy women's basketball play by play data • wehoop (saiemgilani.github.io)

eSports:

- League of Legends (LoL): Cassiopeia Documentation — Cassiopeia 3.0.x documentation
- Dota2: OpenDota - Dota 2 Statistics

Example: NHL shot map from a Penguin's playoff game



TidyTuesday!

Link: rfordatascience/tidytuesday: Official repo for the #tidytuesday project (github.com)

Data are posted weekly. The community shares examples for how to explore, clean, visualize, and model the data! Neat way to learn new skills!

Example data sets:

- Ninja Warrior: tidytuesday/readme.md at master · rfordatascience/tidytuesday (github.com)
- Scooby Doo: tidytuesday/readme.md at master · rfordatascience/tidytuesday (github.com)
- Space Launches: tidytuesday/data/2019/2019-01-15 at master · rfordatascience/tidytuesday (github.com)

Other general data repositories:

UCI Machine Learning Repository: UCI Machine Learning Repository

FiveThirtyEight: Our Data | FiveThirtyEight

Awesome Public Datasets: awesomedata/awesome-public-datasets: A topic-centric list of HQ open datasets. (github.com)

Kaggle: Find Open Datasets and Machine Learning Projects | Kaggle

Data science competition show: SLICED

SLICED is a data science competition show that streams on Twitch.

See how professionals across various industries wrangle, explore, visualize, and ultimately model data in a live competition.

Our community of Western Pennsylvania!

Western Pennsylvania Regional Data Center (WPRDC):

WPRDC • The Region's Open Data at Your Fingertips

Link specifically to the available data: Datasets - WPRDC

Let’s look through an example together!

Allegheny County COVID-19 Tests, Cases and Deaths - Datasets - WPRDC