



# *Working with a Data Librarian*

# What's a Data Librarian?

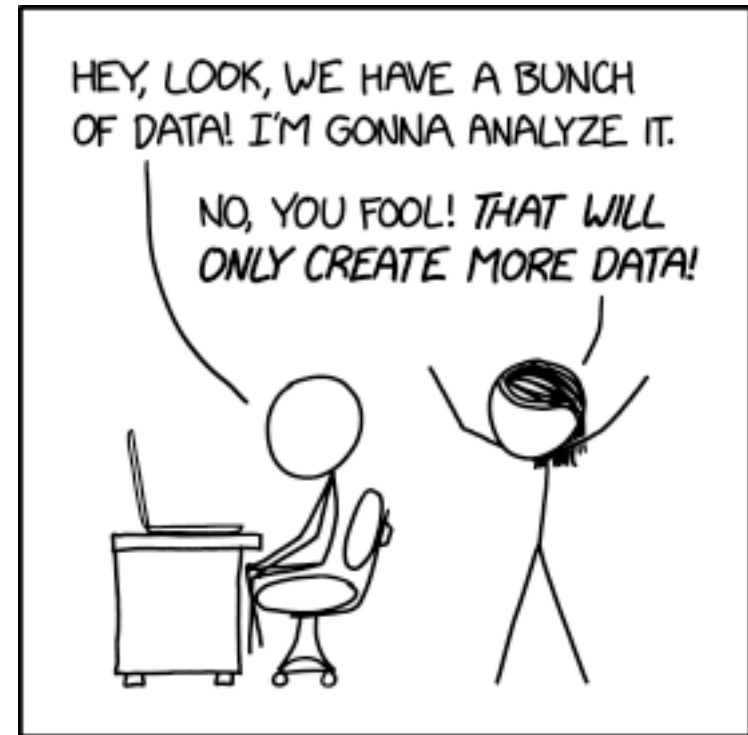
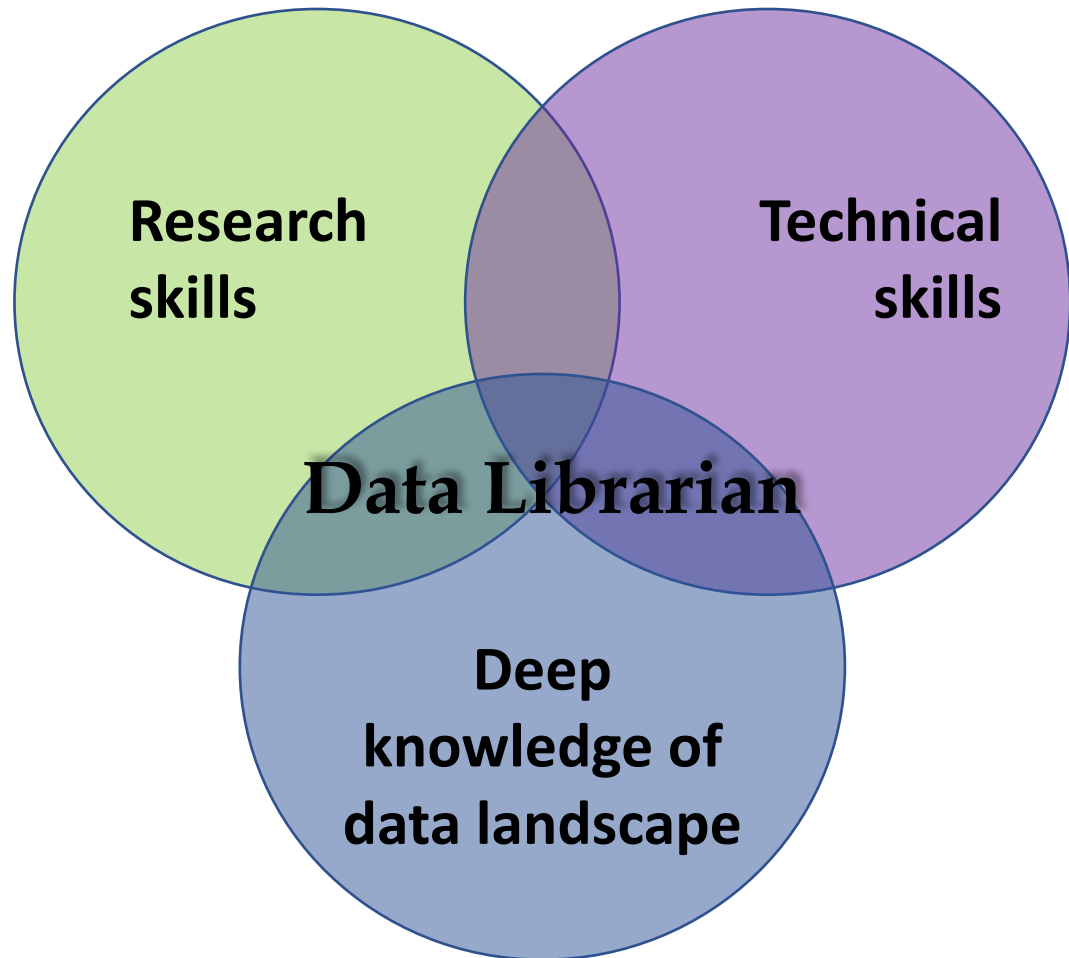
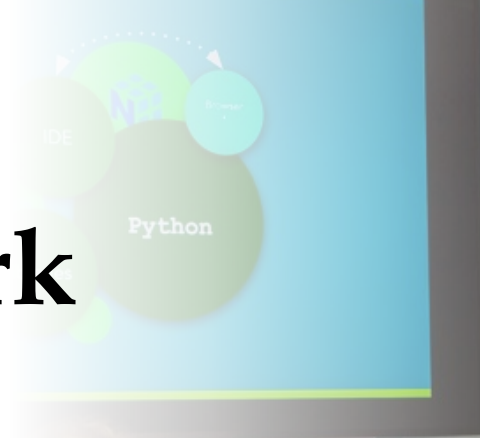


Image via [xkcd](#)

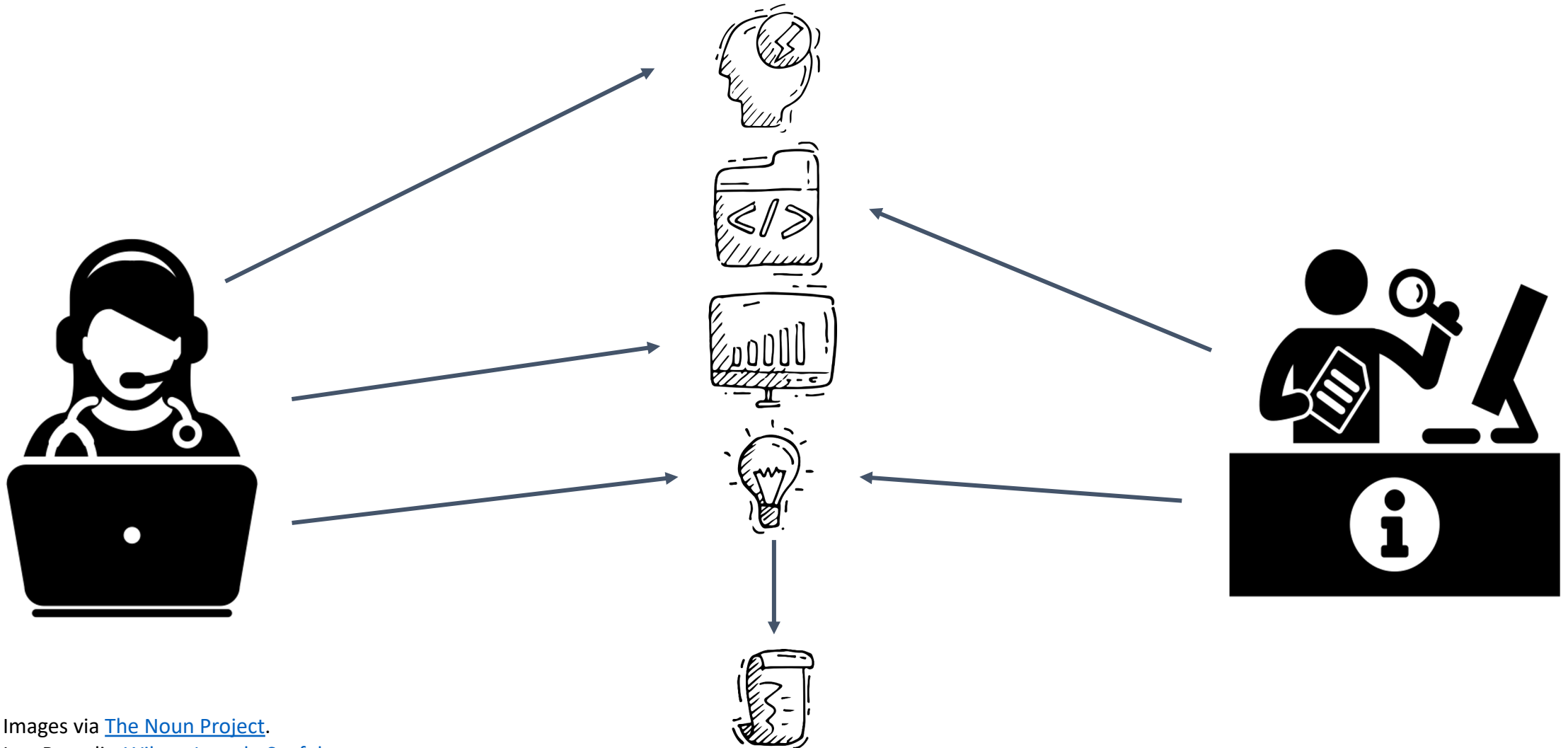
# Current Work

- Supporting research
- Teaching
- Advising on data policy and infrastructure
  - *For example:* NIH Data Management and Sharing Policy
- Conducting research

Image by Terry Dagradi and courtesy of Cushing/Whitney Medical Library



# Clinician-Librarian Collaboration



Example of  
original,  
collaborative  
research

Original Investigation | Geriatrics



January 6, 2023

## Association of Nursing Home Exposure to Hurricane-Related Inundation With Emergency Preparedness<sup>1</sup>

Natalia Festa, MD, MBA<sup>1,2</sup>; Kaitlin F. Throgmorton, MLIS<sup>3</sup>; Nora Heaphy, BA<sup>4</sup>; Maureen Canavan, PhD, MPH<sup>5</sup>; Thomas M. Gill, MD<sup>1</sup>

» [Author Affiliations](#) | [Article Information](#)

*JAMA Netw Open.* 2023;6(1):e2249937. doi:10.1001/jamanetworkopen.2022.49937

### Key Points

**Question** Are nursing homes exposed to potential hurricane-related inundation more likely to meet Centers for Medicare & Medicaid Services criteria for adequate emergency preparedness?

**Findings** In this cross-sectional study of 5914 nursing homes, a higher prevalence of emergency preparedness deficiencies among nursing homes exposed to hurricane-related inundation in the Mid-Atlantic region was observed. Exposure status remained positively associated with the presence and number of emergency preparedness deficiencies after adjustment for facility characteristics, with the converse for facilities within the Western Gulf Coast.

**Meaning** These findings suggest opportunities to reduce regional heterogeneity and improve the alignment of nursing home emergency preparedness with surrounding environmental risks.

# Challenges Along the Way

- Access
- Management
- Processing
- Methodology



# Approaches

- Interpreting and assessing data retrieval methods, documentation and quality
- Tracking data sourcing and citing data appropriately

▼ **2023 Annual files** <sup>2</sup> [Download all 2023 annual files](#)

---

[nursing\\_homes\\_including\\_rehab\\_services\\_01\\_2023.zip](#) 01 / 25 / 2023 • ZIP • 39 MB

---

► **2022 Annual files** [Download all 2022 annual files](#)

---

► **2021 Annual files** [Download all 2021 annual files](#)

---

► **2020 Annual files** [Download all 2020 annual files](#)

---

► **2019 Annual files** [Download all 2019 annual files](#)

---

► **2018 Annual files** <sup>3</sup> [Download all 2018 annual files](#)

---

► **2017 Annual files** [Download all 2017 annual files](#)

---

► **2016 Annual files** [Download all 2016 annual files](#)

---

CMS is required to maintain data from the last 7 years for this provider and make it available to download here. If you're looking to access older data and if it is available, visit one of our archived data websites.

- [Data 2015 to current year](#)
- [Data 2014 and older years](#) (Refer to the [Frequently Asked Questions](#) for instructions if a redirect error occurs.)

[^ Back to top](#)

# Approaches

- Making data dictionaries and missingness tables
- Shifting from tools like Stata to Python
- Converting hard-to-document Excel processes to replicable code
- Documenting work in multiple modalities with Jupyter Notebooks

Column	Non-Null Count	Missing Value Counts	Pandas Dtype	Conventional Type	Variable Description
federalprovidernum	16569	0	object	string	exact match to data dictionary
providername	16502	67	object	string	exact match to data dictionary
provideraddress	16502	67	object	string	exact match to data dictionary
providercity	16502	67	object	string	exact match to data dictionary
providerstate	16502	67	object	string	exact match to data dictionary
providerzipcode	16502	67	object	string	exact match to data dictionary
providercountynam	16502	67	object	string	exact match to data dictionary
ownershiptype	16502	67	object	string	exact match to data dictionary
numberofcertifiedb	16502	67	float64	decimal	exact match to data dictionary
averagenumberofre	16395	174	float64	decimal	exact match to data dictionary
providerresidesin	16502	67	object	string	exact match to data dictionary
legalbusinessname	16502	67	object	string	exact match to data dictionary
mostrecenthealthin	16502	67	object	string	exact match to data dictionary
automaticsprinklers	16502	67	object	string	exact match to data dictionary
overallrating	16378	191	float64	decimal	exact match to data dictionary

The screenshot shows a Jupyter Notebook titled "data\_prep\_for\_storm\_surge\_20220729". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, a "Run" button, and a "Markdown" dropdown. The main content area displays a custom function definition:

```
In [4]: 1 def fix_missing_chars(column, dataframe, num_chars, missing_char='0'):
2         """This function fixes missing characters, such as leading zeroes, in pandas dataframe columns.
3         **Args:
4         column = column in pandas dataframe where characters are missing
5         dataframe = name of pandas dataframe
6         num_chars = number of total characters desired
7         missing_char='0' (default)
```

Below the code is a "Summary" section:

**Summary**

This notebook contains code and notes used to prepare and process a dataset for statistical analysis -- as part of a series of projects using public nursing home and long term care facility data (from the Centers for Medicaid & Medicare Services, CMS, among other sources) alongside various environmental exposure data to identify risks to older populations resulting from climate disasters like hurricanes and wildfires.

In a nutshell, the code below:

- imports and filters data from various sources
- performs light data cleaning
- transforms, subsets, and groups data (using [split-apply-combine Pythonic pandas methodology](#)) as needed for various analyses
- creates and computes new data fields
- combines data from various sources, as well as new groupings and computed columns, into one merged dataset for analysis, though for this set of analyses, that results in three different final datasets:

[Redacted]

- calculates missing data, and culls missing data based on pre-determined criteria
- summarizes basics findings and visualizes some information for manuscript tables

# Approaches

- Reviewing data literature and methodology
- Implementing data processing methodologies, such as:
  - tidy data<sup>4</sup>
  - split-apply-combine<sup>5 6</sup>

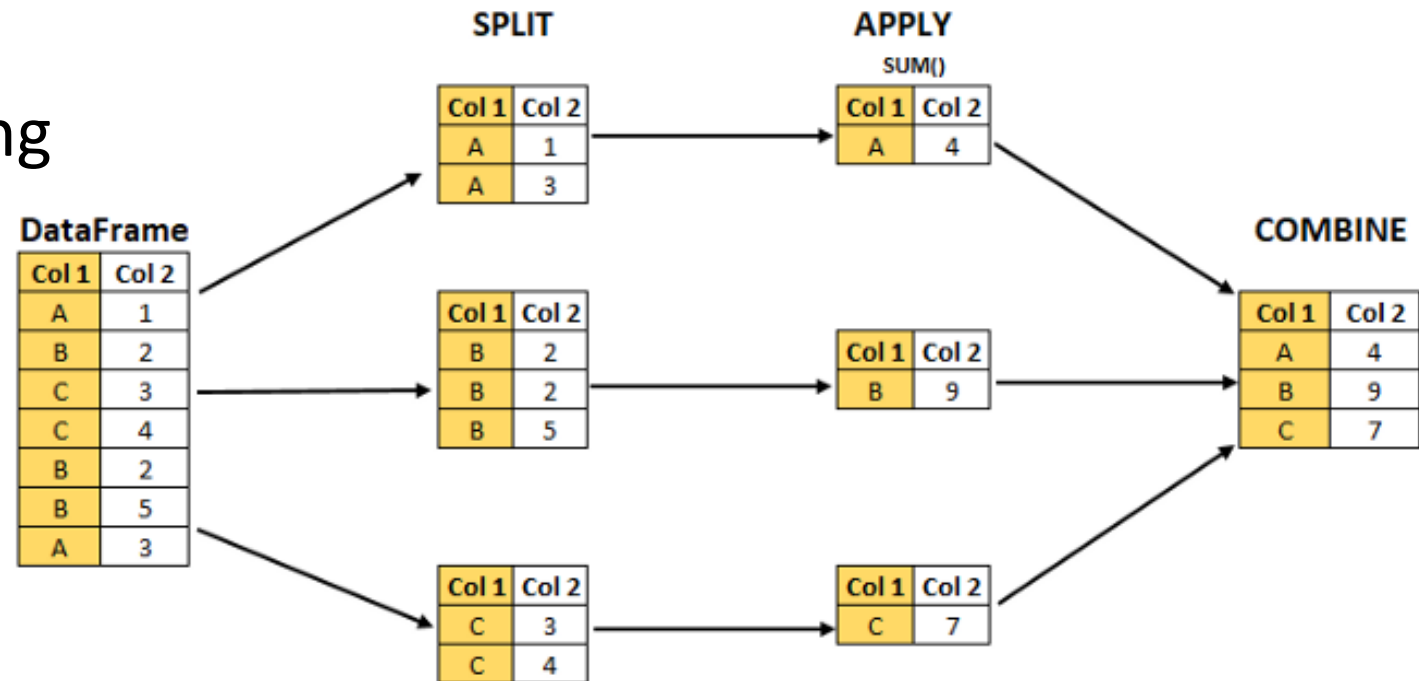


Image via [Analytics Vidhya](#)

# What we've learned from each other

## *Clinician perspective:*

- The utility of publicly available data to answer questions with clinical and health services relevance
- Public data can present many challenges to cleaning and merging across multiple data sources
  - Unstructured data
  - Data with inconsistent identifiers across years and data sources (healthcare facility identification numbers, for example)
  - Merging data with inconsistent or error-laden identifiers presents many challenges

# What we've learned from each other

## *Clinician perspective:*

- Many data sources can be more amenable to cleaning and organizing in languages other than those in which clinician researchers are trained
- Accurately logging procedures used in data pre-processing is essential to replicability
- Data Librarians have rich expertise in addressing each of the challenges above

# What we've learned from each other

## *Librarian perspective:*

- Growing clinician ambition to take on complex, multi-disciplinary data projects
- More support needed, especially in early clinician-researcher career, for data training
- Opportunity to:
  - Fully participate in research data lifecycle
  - See the power of open data in action

# How to succeed as a clinician-librarian team

- Communicate early and often
- Grow project portfolio over time
- Teach and trade skills
- Seek expertise elsewhere when needed

# Advice to clinicians on involving librarians

- Learn what data support is offered by your library
- Learn and practice principles that your data librarian encourages as foundational to data science integrity and reproducibility
- Establish your team and expectations regarding roles and responsibilities during early stages of the project
- Ensure that there are opportunities for learning and professional development for all team members



# Summing Up

**DISCOVER**

**data support services**

**USE**

**open data**

**TAKE**

**a Team Science  
approach**

# Resources

- **Data management planning guidance for NIH grants and beyond:**
  - [Ten simple rules for maximizing the recommendations of the NIH data management and sharing plan](#) | *PLOS Computational Biology*
  - [The FAIR Guiding Principles for scientific data management and stewardship](#) | *Scientific Data*
  - [Ten simple rules for the care and feeding of scientific data](#) | *PLOS Computational Biology*
- **Excellent data organization and cleaning principles:**
  - [Data organization in spreadsheets](#) | *The American Statistician*
- **For those interested in Python:**
  - [Python for Non-Programmers](#) | python.org

# References

1. Festa, N., Throgmorton, K. F., Heaphy, N. M., Canavan, M., & Gill, T. M. (2023). Association of nursing home exposure to hurricane-related inundation with emergency preparedness. *JAMA Network Open* 6(1):e2249937. <https://doi.org/10.1001/jamanetworkopen.2022.49937>
2. Centers for Medicare and Medicaid. Provider Information. <https://data.cms.gov/provider-data/dataset/4pq5-n9py>
3. LTCFocus Public Use Data sponsored by the National Institute on Aging (P01 AG027296) through a cooperative agreement with the Brown University School of Public Health. Available at [www.ltcfocus.org](http://www.ltcfocus.org). <https://doi.org/10.26300/h9a2-2c26>
4. Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), 1–23. <https://doi.org/10.18637/jss.v059.i10>
5. Pandey, A. (2018). Split-Apply-Combine Strategy for Data Mining. *Medium*. <https://medium.com/analytics-vidhya/split-apply-combine-strategy-for-data-mining-4fd6e2a0cc99>
6. Pandas. (n.d.) Group by: split-apply-combine. [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/groupby.html](https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html)