SCHARE

Tutorial Think-a-Thons

National Institutes of Health

SCHARE

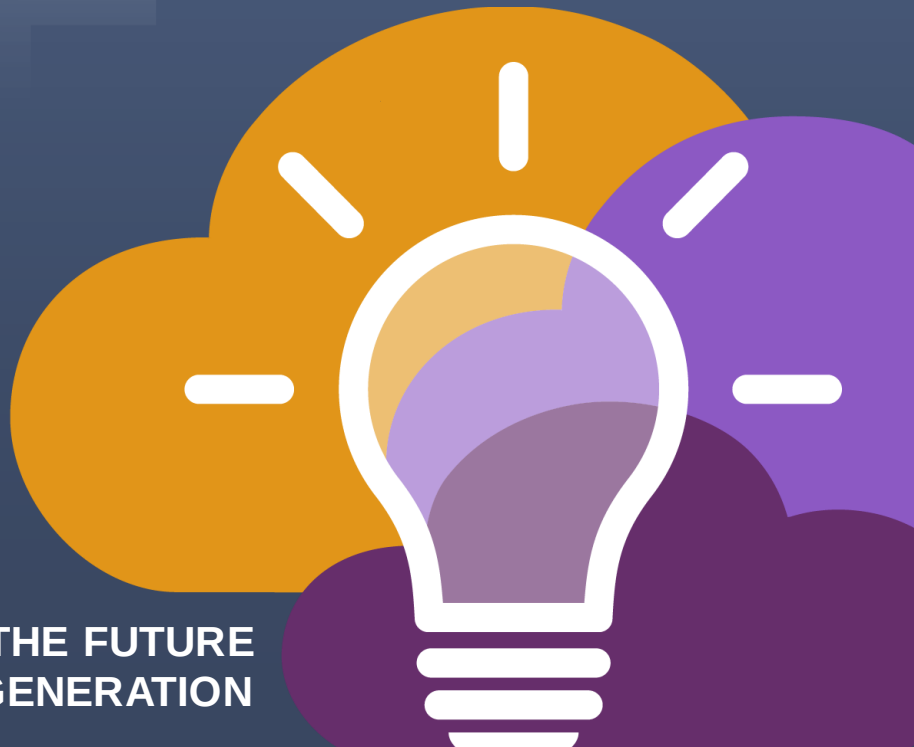
Health Outcome Research Paradigm Shift: Understanding How Big Data Expands Knowledge

April 16, 2025

Deborah Duran, PhD · NIMHD
Elif Dede Yildirim, PhD · NIMHD
Mark Aronson, PhD · NIMHD

SCHARE

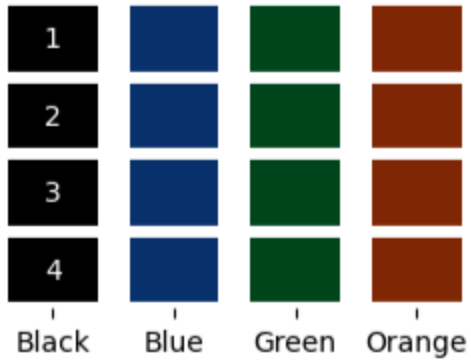
Merging datasets



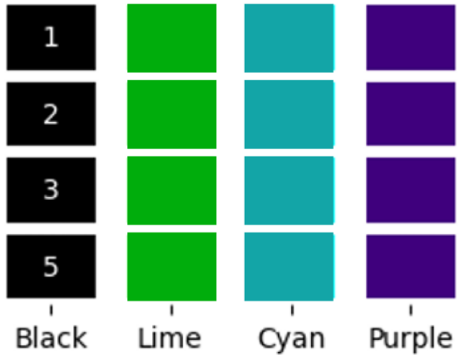
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Ways to merge datasets with shared respondents

Dataset 1

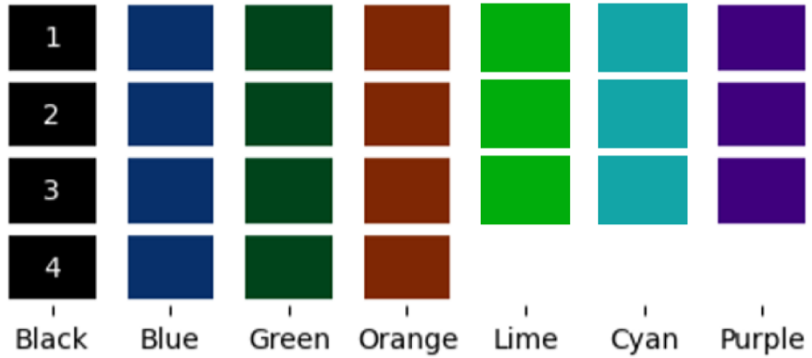


Dataset 2



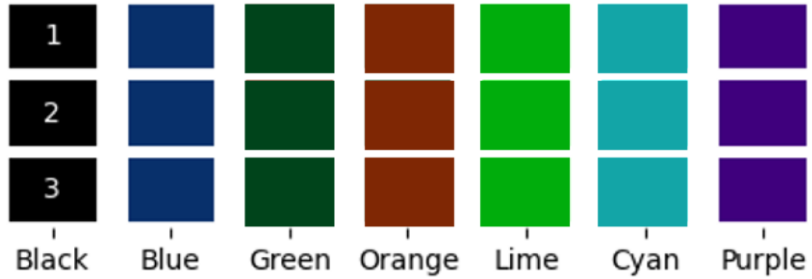
Left Merge

Keep all entries in Dataset 1



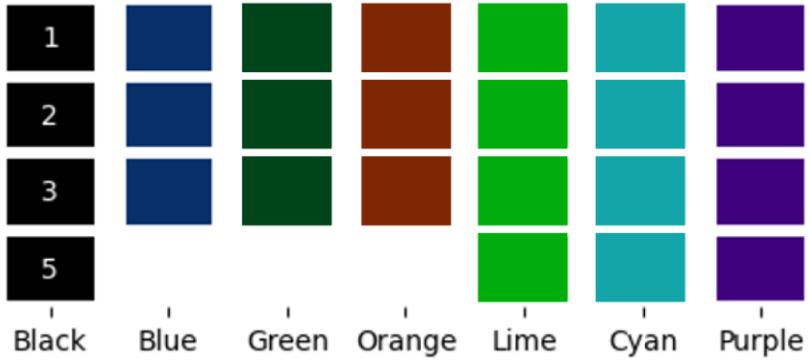
Inner Merge

Keep all entries in both Datasets



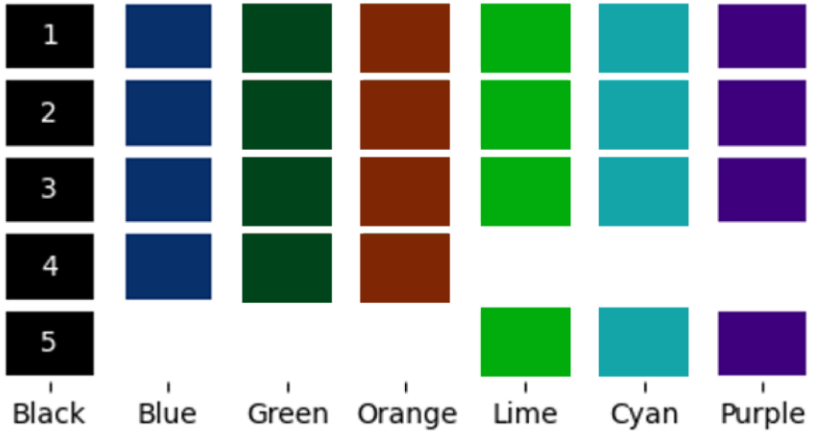
Right Merge

Keep all entries in Dataset 2



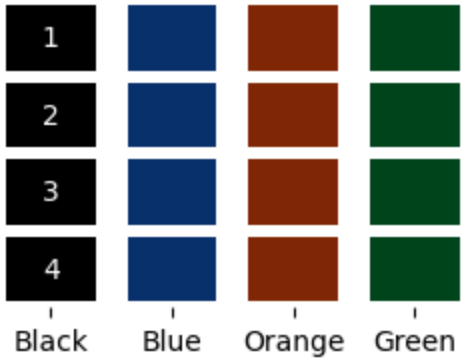
Outer Merge

Keep all entries in either Dataset

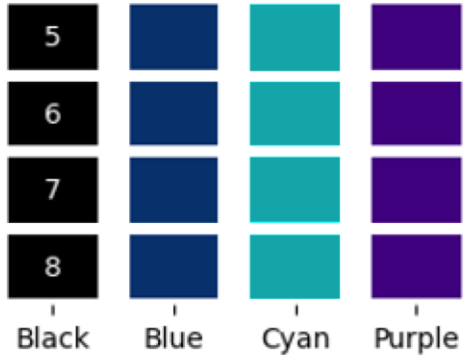


Ways to combine datasets with different respondents

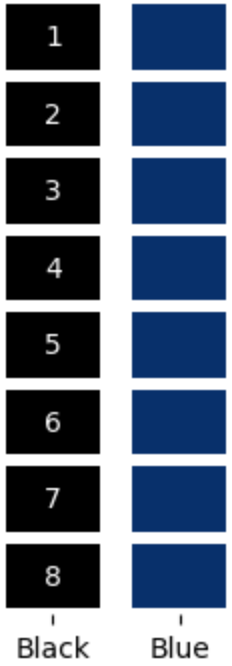
Dataset 1



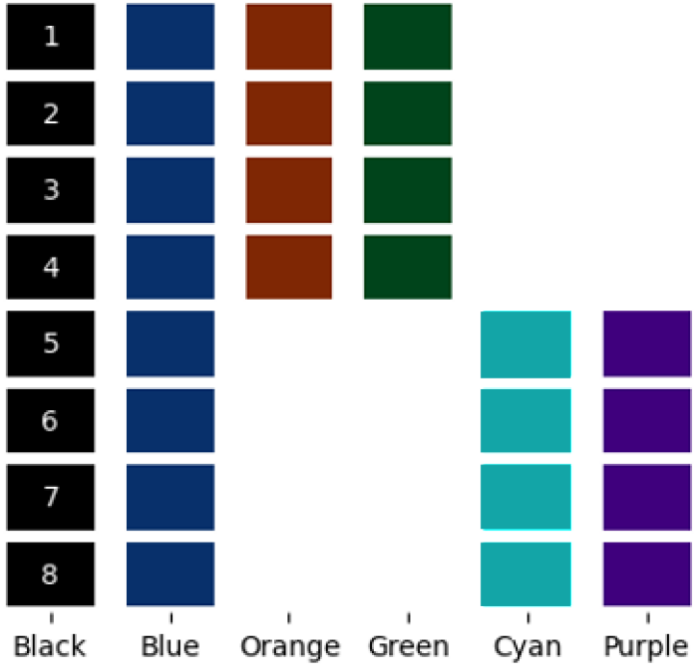
Dataset 2



Inner Merge
Keep shared variables



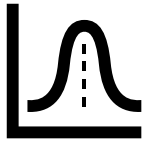
Outer Merge
Keep all variables



Bringing it all together

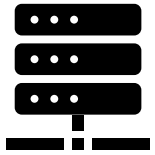
Picking an Analytical Approach

Statistical Models



or

Machine Learning

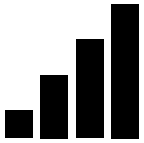


The Research Question

The Data

Categories of Data

Data Format

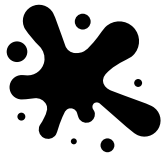


Quantitative

Qualitative

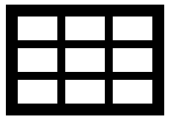
Data Structure

Structured



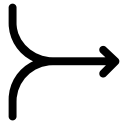
Unstructured

Data Types



Ways of combining datasets

Merging



Concatenating

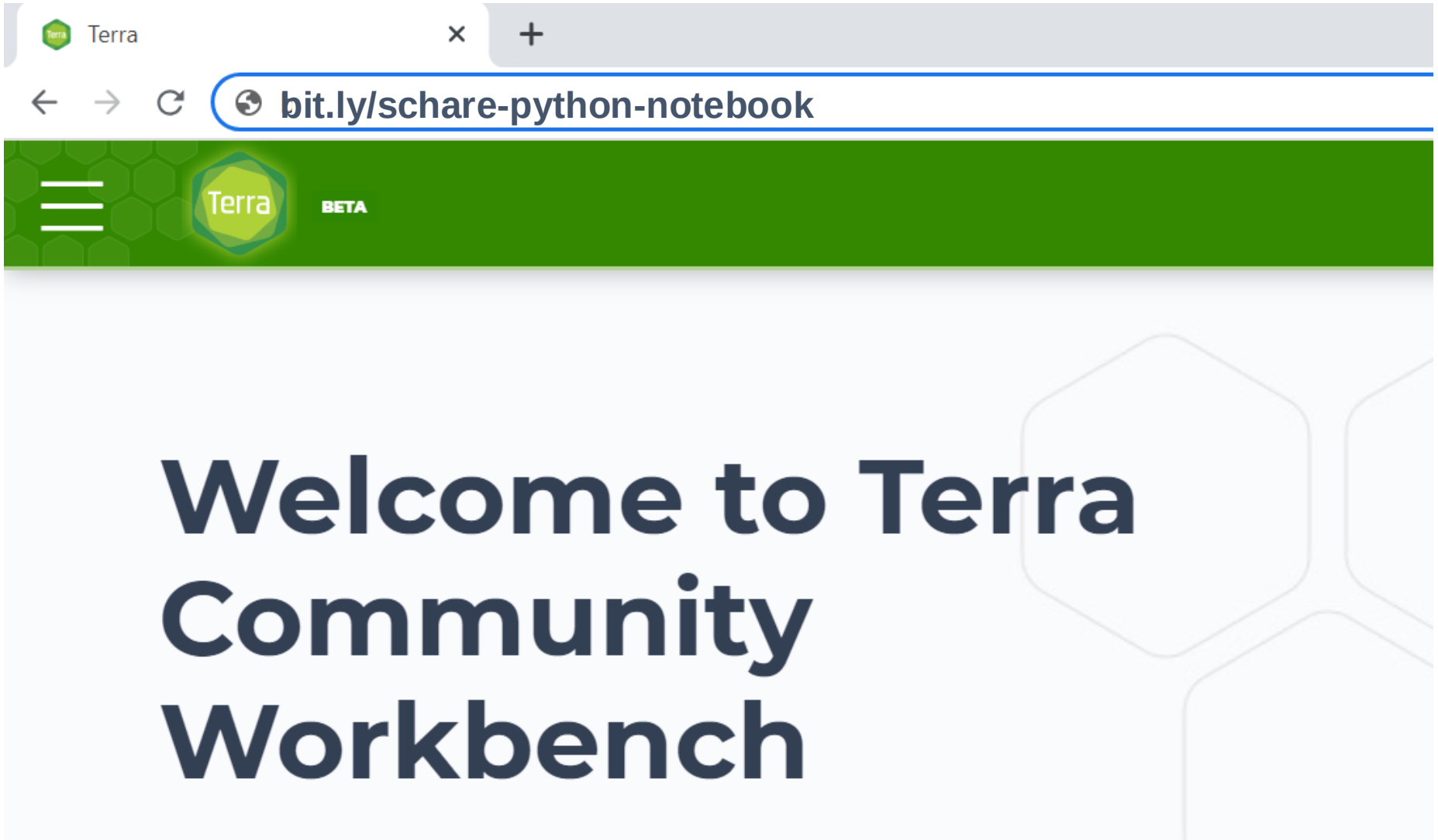
How to go about finding datasets?

Slido Poll

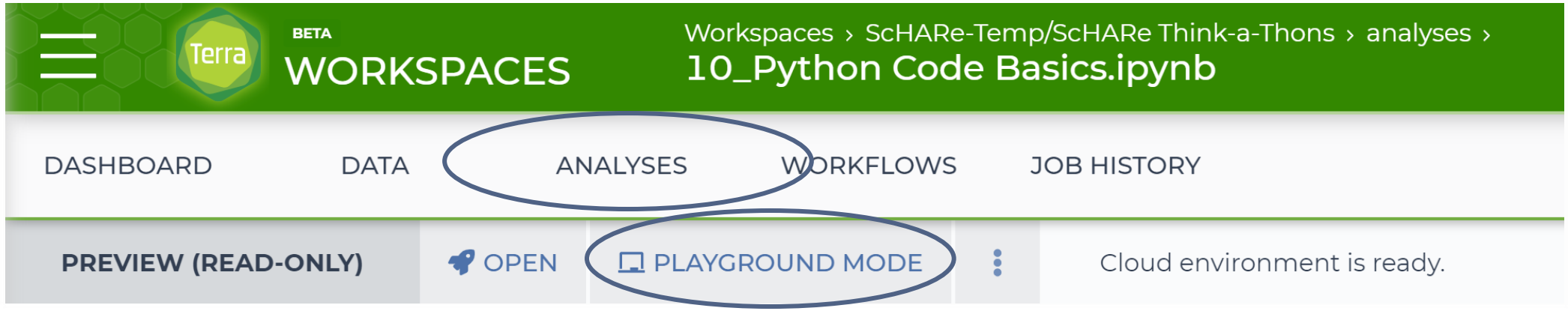
Which of the following correctly describes common methods for merging two datasets in data science?

a) Concatenation appends columns from two datasets regardless of index alignment.
b) An inner join includes only rows with keys that are present in both datasets.
c) A left join returns only the rows that are unique to the second dataset.
d) A full outer join excludes any rows with missing values in either dataset.

Paste this address in your browser: bit.ly/schare-python



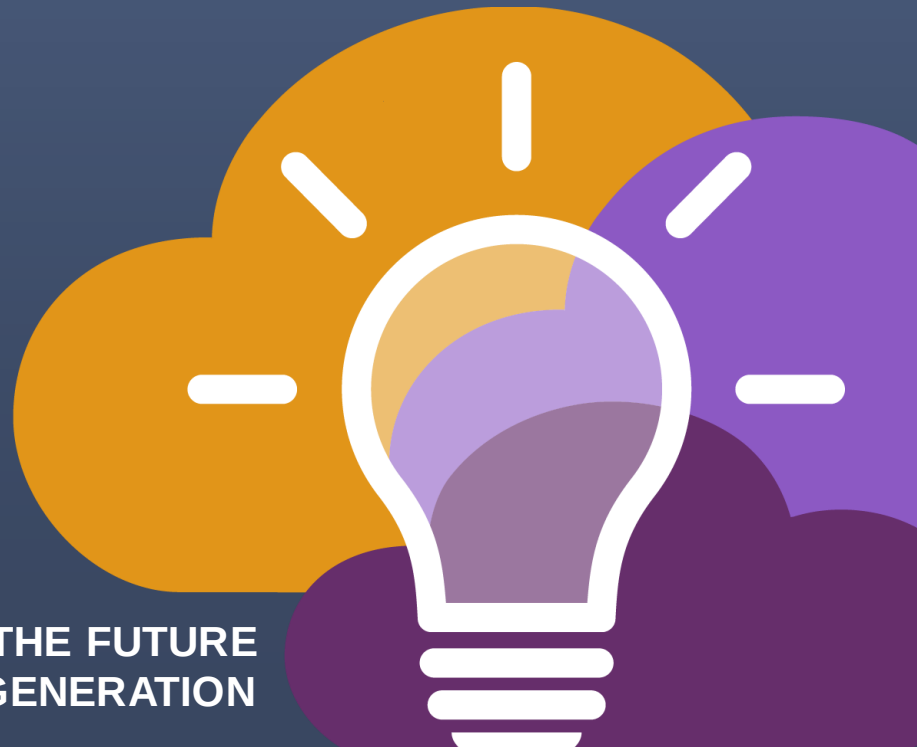
Do you see a Playground mode button?



If yes, click on it to start your virtual computer. You are done!

SCHARE

SCHARE Datasets



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

SCHARE Ecosystem

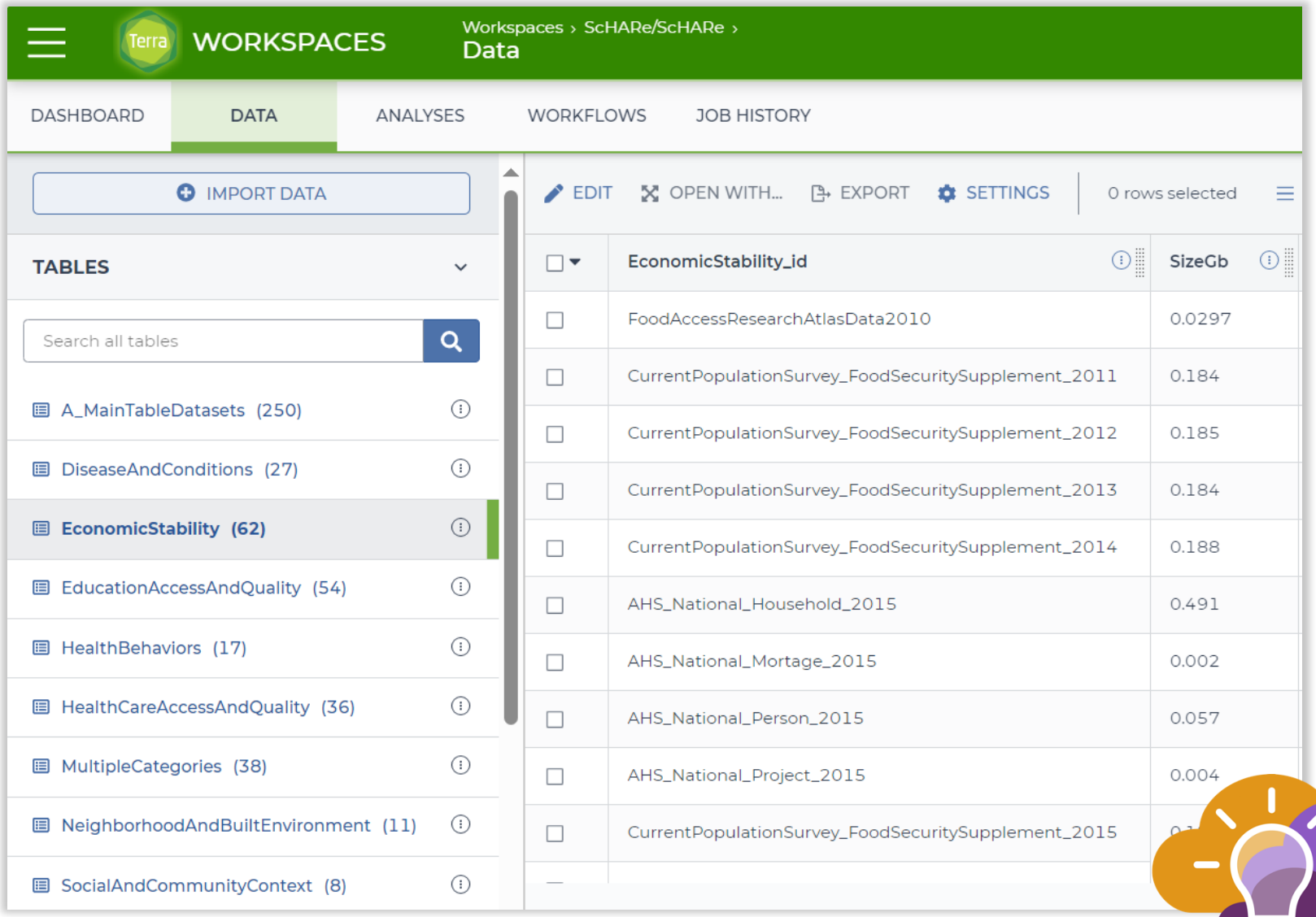
OVER 300 DATA SETS CENTRALIZED

Datasets are categorized by content based on the CDC **Social Determinants of Health categories**:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

with the addition of:

- **Health Behaviors**
- **Diseases and Conditions**



SCHARE Ecosystem

Examples of datasets for each category include:

Education access and quality

Data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.

Examples:

- **EDFacts Data Files** (U.S. Dept. of Education) - Graduation rates and participation/proficiency assessment
- **NHES - National Household Education Surveys Program** (U.S. Dept. of Education) – Educational activities

SCHARE Ecosystem

Health care access and quality

Data on health literacy, use of health IT, emergency room waiting times, preventive healthcare, health screenings, treatment of substance use disorders, family planning services, access to a primary care provider and high quality care, access to telehealth and electronic exchange of health information, access to health insurance, adequate oral care, adequate prenatal care, STD prevention measures, etc.

Example:

- **MEPS - Medical Expenditure Panel Survey** (AHRQ) - Cost and use of healthcare and health insurance coverage
- **Dartmouth Atlas Data -** Selected Primary Care Access and Quality Measures - Measures of primary care utilization, quality of care for diabetes, mammography, leg amputation and preventable hospitalizations

SCHARE Ecosystem

Neighborhood and built environment

Data on access to broadband internet, access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, deaths from motor vehicle crashes, asthma and COPD cases and hospitalizations, noise exposure, smoking, mass transit use, etc.

Examples:

- **National Environmental Public Health Tracking Network** (CDC) - Environmental indicators and health, exposure, and hazard data
- **LATCH - Local Area Transportation Characteristics for Households** (U.S. Dept. of Transportation) – Local transportation characteristics for households

SCHARE Ecosystem

Social and community context

Data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc.

Example:

- **Hate crime statistics** (FBI) - Data on crimes motivated by bias against race, gender identity, religion, disability, sexual orientation, or ethnicity
- **General Social Survey** (GSS) - Data on a wide range of characteristics, attitudes, and behaviors of Americans.

SCHARE Ecosystem

Economic stability

Data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.

Examples:

- **Current Population Survey (CPS) Annual Social and Economic Supplement** (U.S. Bureau of Labor Statistics) - Labor force statistics: annual work activity, income, health insurance, and health
- **Food Access Research Atlas** (U.S. Dept. of Agriculture) – Food access indicators for low-income and other census tracts

SCHARE Ecosystem

Health behaviors

Data on health-related practices that can directly affect health outcomes.

Examples:

- **BRFSS - Behavioral Risk Factor Surveillance System** (CDC) - State-level data on health-related risk behaviors, chronic health conditions, and use of preventive services
- **YRBSS - Youth Risk Behavior Surveillance System** (CDC) – Health behaviors that contribute to the leading causes of death, disability, and social problems among youth and adults

SCHARE Ecosystem

Diseases and conditions

Data on incidence and prevalence of specific diseases and health conditions.

Examples:

- **U.S. CDI - Chronic Disease Indicators** (CDC) - 124 chronic disease indicators important to public health practice
- **UNOS - United Network of Organ Sharing** (Health Resources and Services Administration) – Organ transplantation: cadaveric and living donor characteristics, survival rates, waiting lists and organ disposition

SCHARE Ecosystem: Public datasets

Examples of interesting datasets include:

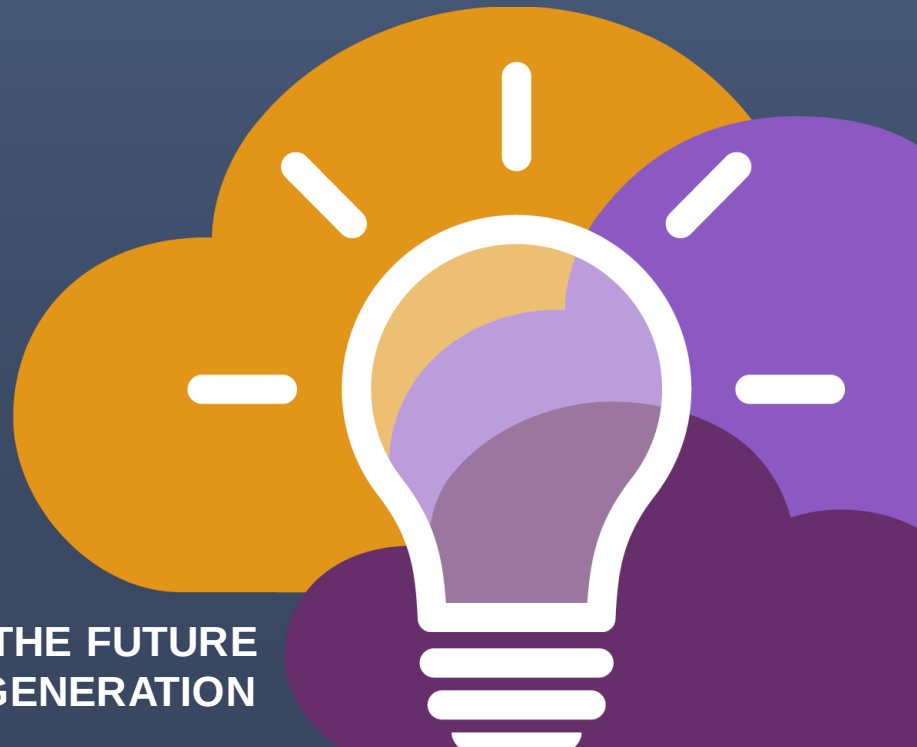
- **American Community Survey** (U.S. Census Bureau)
- **US Census Data** (U.S. Census Bureau)
- **Area Deprivation Index** (BroadStreet)
- **GDP and Income by County** (Bureau of Economic Analysis)
- **US Inflation and Unemployment** (U.S. Bureau of Labor Statistics)
- **Quarterly Census of Employment and Wages** (U.S. Bureau of Labor Statistics)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **Low Income Housing Tax Credit Program** (U.S. Dept. of Housing and Urban Development)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **Medicare** (U.S. Dept. of Health & Human Services)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **COVID-19 Data Repository by CSSE at JHU** (Johns Hopkins University)
- **COVID-19 Mobility Impact** (Geotab)
- **COVID-19 Open Data** (Google BigQuery Public Datasets Program)
- **COVID-19 Vaccination Access** (Google BigQuery Public Datasets Program)

Slido

When you see a data set, what questions about the data pop into your head?

SCHARE

How to Use PySCHARE Package



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Search

This widget allows you to search for variables and descriptions across multiple datasets. Follow these instructions to effectively use the search functionality:

- **If you want to search within a specific dataset:** Use the "Datasets" dropdown menu to select the dataset you wish to search. Scroll down to see the list of available datasets and select your choice.
- **If you want to search across all datasets:**Leave the "Datasets" dropdown set to "None". This is the default option.

In the text box, type the word or phrase you want to search for and click the "Search" button. Note: You need to enter at least 3 characters for the search to function.

Search Results: The widget will display a table below the search box, showing the search results.

- **If a specific dataset was selected:** The table will show the variables and descriptions from that dataset that match your search terms.
- **If "None" was selected:** The table will show results from all datasets that match your search terms, including the dataset name, variable name, and description.

Save Table: If you want to save the search results as an HTML file, click the "Save Table" button.

- **If a specific dataset was selected:** The file will be named using the dataset name and the search terms (e.g., "Food Security Data 2021_searchterm.html").
- **If "None" was selected:** The file will be named using "Datasets" and the search terms (e.g., "Datasets_with_searchterm.html").

A confirmation message will appear below the "Save Table" button, indicating the file name and location.

Search Variables

Datasets

None
BRFSS Phone Survey 2012
BRFSS Phone Survey 2013
BRFSS Phone Survey 2014
BRFSS Phone Survey 2015
BRFSS Phone Survey 2016
BRFSS Phone Survey 2017
BRFSS Phone Survey 2018
BRFSS Phone Survey 2019

Search

depress

Search

Save Table

Dataset	Variables	Description
BRFSS Phone Survey 2012	ACEDEPRS	'LIVE WITH ANYONE DEPRESSED, MENTALLY ILL, OR SUICIDAL?'
BRFSS Phone Survey 2012	ADDEPEV2	'EVER TOLD YOU HAD A DEPRESSIVE DISORDER'
BRFSS Phone Survey 2012	MISDEPRD	'HOW OFTEN FEEL DEPRESSED PAST 30 DAYS'
BRFSS Phone Survey 2012	QLMENTL2	'HOW MANY DAYS DEPRESSED IN PAST 30 DAYS'
BRFSS Phone Survey 2012	VHDRPTSD	'DOCTOR DIAGNOSED DEPRESSION, ANXIETY, OR POST TRAUMATIC STRESS DISORDER (PTSD)?'

Data Explorer

This widget allows you to explore and manipulate datasets. Follow the steps below to work with the data:

1. **Selecting a Dataset:**

   Use the "Select Dataset" dropdown menu to choose the dataset you want to work with. Click on the dropdown to see a list of available datasets and select your choice.

2. **Selecting Variables:**

   After selecting a dataset, the "Select Variable" dropdown menu will populate with a list of variables available within that dataset. Choose the variables from the "Select Variable" dropdown you want to analyze. You can select multiple variables from this dropdown. (*Note:* If you do not select any variables, actions will be applied to all variables).

3. **Viewing Data:**

   To view the first few rows of the selected dataset or the selected variables, click the "Show Data" button. The results will be displayed in the output area below the widget.

4. **Describing Data:**

   To view summary statistics (like mean, median, standard deviation) for the selected variables, click the "Describe Data" button. The summary statistics will be shown in the output area below the widget. If you haven't selected any variables, statistics for all variables in the dataset will be displayed.

5. **Saving Data:**

   To save the displayed data (either the entire dataset or the subset of selected variables), click the "Save Data" button. A confirmation message, including the location where the data is saved, will be displayed in the output area below the buttons.

   Note: Ensure you have selected a dataset and, if applicable, variables before clicking "Save Data."

6. **Clearing Output:**

   To clear both the confirmation message (from saving) and the displayed data table or statistics in the output area, click the "Clear Output" button.

Data Explorer

Select Dataset

Select Variable

2021FoodSecurityData
2021FullYearConsolidatedData
2021JobsFileData
2021MedicalConditionsData
2021PersonRoundPlanPublicUseData
2022FoodSecurityData
2022FullCharacteristicsData
2022FullYearConsolidatedData
2022JobsData

Clear Output

Show Data

Describe Data

Save Data

Interactive Plots

This widget allows you to create various types of plots using your selected dataset. Follow the steps below to build your visualization:

1. **Choose a Dataset:**

   Begin by selecting a dataset from the "Select Dataset" dropdown menu. Click the dropdown to see the list of available datasets and choose the one you want to use.

2. **Select a Plot Type:**

   Next, choose the type of plot you want to create from the "Select Plot" dropdown menu. Common plot types include:

   - **Bar Plots, Count Plots, Box Plots, Boxen Plots, Strip Plots, Swarm Plots, and Violin Plots:** These are typically used to show relationships between categories. They usually require a categorical variable for the X-axis (or "Hue") and a numeric variable for the Y-axis. (See the Categorical Tutorial for more details).
   - **Scatter Plots and Line Plots:** These are used to show relationships between two numeric variables. For example, you might plot time versus measurement. (See the Relational Tutorial for more details).
   - **Histograms:** These are used to show the distribution of a single numeric variable. (See the Distributions Tutorial for more details).

3. **Configure Plot Parameters:**

   - **X-Axis and Y-Axis:** Use the "Select X" and "Select Y" dropdown menus to choose the variables you want to plot on the X and Y axes. The available options will depend on the dataset you selected.
   - **Hue:** Use the "Select Hue" dropdown to color-code your data points based on categories. This helps to visualize how different categories are distributed.
   - **Style:** Use the "Select Style" dropdown to vary the markers or lines in your plot, based on categories.
   - **Size:** Use the "Select Size" dropdown to scale the size of the markers based on another variable.
   - **Column and Row:** Use the "Select Column" and "Select Row" dropdowns to create subplots (facets). This allows you to compare different categories across multiple plots.
   - **Layer (Multiple):** Use the "Select Layer" dropdown to manage how overlapping data points are displayed. Options like "Dodge," "Stack," and "Fill" are available.

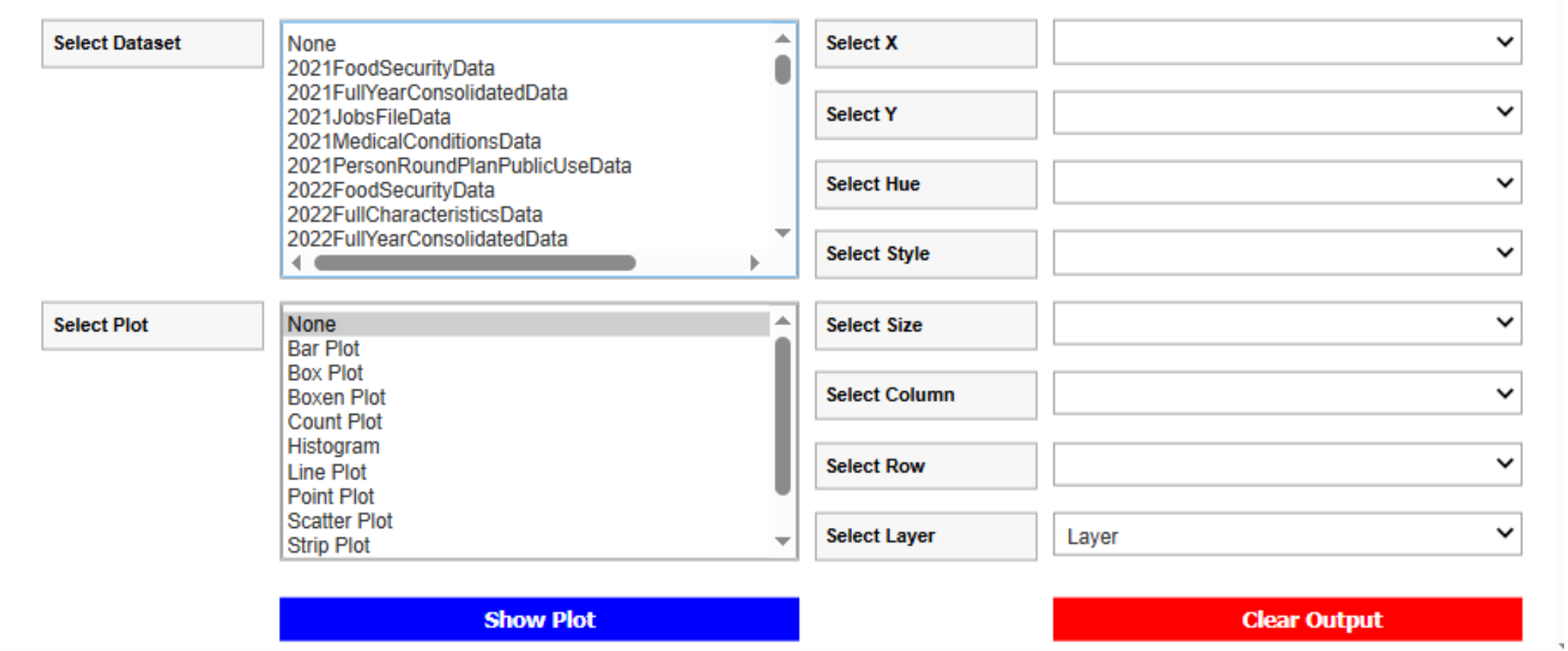
4. **View Your Plot:**

   Once you have selected your plot type and configured the parameters, click the "Show Plot" button. Your plot will be displayed below the widget.

5. **Clear Output:**

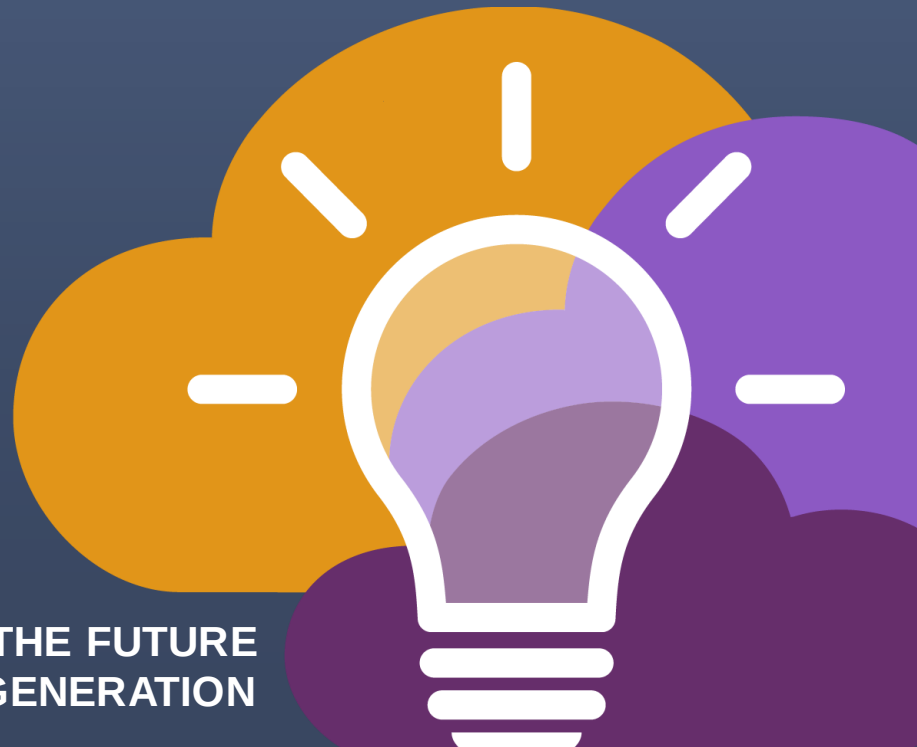
   To clear the displayed plot, click the "Clear Output" button.

Interactive Plots



SCHARE

How to Upload Datasets to Your Workspace



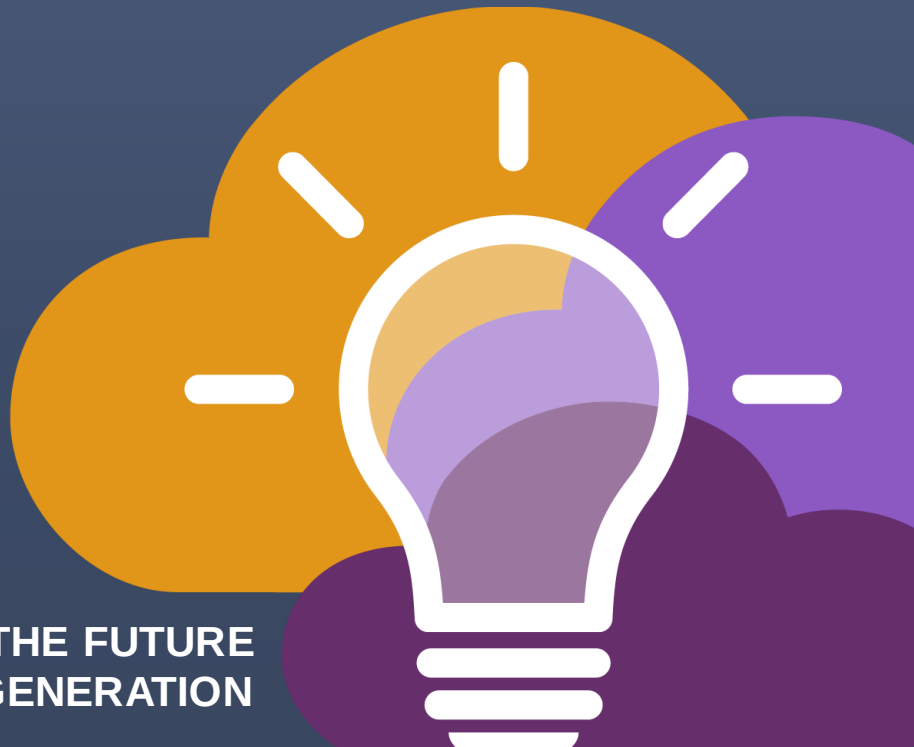
BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

SCHARE Notebooks

Jupyter	b. 06_How to access plot and save data from public BigQuery datasets using Python 3.ipynb	Dec 10, 2024	
Jupyter	b. 07_How to access plot and save data from ScHARe hosted datasets using Python 3.ipynb	Dec 10, 2024	
Jupyter	b. 08_How to upload access plot and save data stored locally using Python 3.ipynb	Dec 10, 2024	
Jupyter	c. 05_How to access plot and save data from public BigQuery datasets using R.ipynb	Mar 11, 2025	

SCHARE

How to Upload Datasets from SCHARE Data Repository



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Public Collections

Biological

Example NMHSS Analysis LIVE

This contains data from the 2018 National Mental Health Services Survey (N-MHSS) and links to Minority Healt...

a year ago

Health Care Systems and Clinical Care

Example NMHSS Analysis LIVE

This contains data from the 2018 National Mental Health Services Survey (N-MHSS) and links to Minority Healt...

a year ago

Minority Health SVI LIVE

The Centers for Disease Control and Prevention (CDC) and U.S. Department of Health and Human Services (HHS)...

a year ago

Sociocultural Environment

ScHARe Think-a-Thon Demonstration Nov 20 LIVE

November 20 Think-a-Thon Demo Files

4 months ago

Minority Health SVI LIVE

The Centers for Disease Control and Prevention (CDC) and U.S. Department of Health and Human Services (HHS)...

a year ago

Public Collections

This contains data from the 2018 National Mental Health Services Survey (N-MHSS) and links to Minority Health SVI data, also from 2018.

Levels of Influence	Domains of Influence
Community	*Biological* *Health Care Systems and Clinical Care*

> Links and Documents

> Data Items

v Data Access

Jupyter (Terra) | Python | Other Tools

To access data from this collection, copy the following cells into your Jupyter notebook:

```
!pip install pypigeon
```

```
import pypigeon
client = pypigeon.login('test-schare.nimhd.nih.gov')
```

```
collection = client.get_collection_by_name('Example NMHSS Analysis')

item = collection['nmhss-puf-2018-csv.csv']  # Replace with your desired item name

### Retrieve an item into a Pandas DataFrame:
df = item.table()

### Or download its raw contents:
#item_data = item.open('rb').read()
```

Analysis Readiness

Ready >

CDE Compliance - ScHARe

0 / 17 CDEs assigned

Tags

Topics tagged in this collection

Health Care Delivery

Public Collections

DASHBOARD DATA ANALYSES WORKFLOWS JOB HISTORY

PREVIEW (READ-ONLY) Open

Rate: $0.02 per hour

In [1]:
```
## Installation of the pypigeon library, do this once
#
# import sys
# !{sys.executable} -m pip install pypigeon
```

Documentation for the PyPigeon client can be found here:

https://bioteam.github.io/project-pigeon/pypigeon_api.html

In [2]:
```
from pypigeon import login
client = login('test-schare.nimhd.nih.gov')
```

```
To activate your session, visit the URL below:
   https://test-schare.nimhd.nih.gov/login/activate/1w4HqehJJkHX0jHQW7DHRw.jnRx5pzandz47A7C4liKeUHDQ3g

Waiting for session activation...
```

In [3]:
```
collection = client.get_collection_by_name('Example NMHSS Analysis')
```

In [4]:
```
nmhss = collection.get_table('nmhss-cbt-facilities')
```

```
Loading nmhss-cbt-facilities: 0it [00:00, ?it/s]
```

In [5]:
```
nmhss
```

Out[5]:

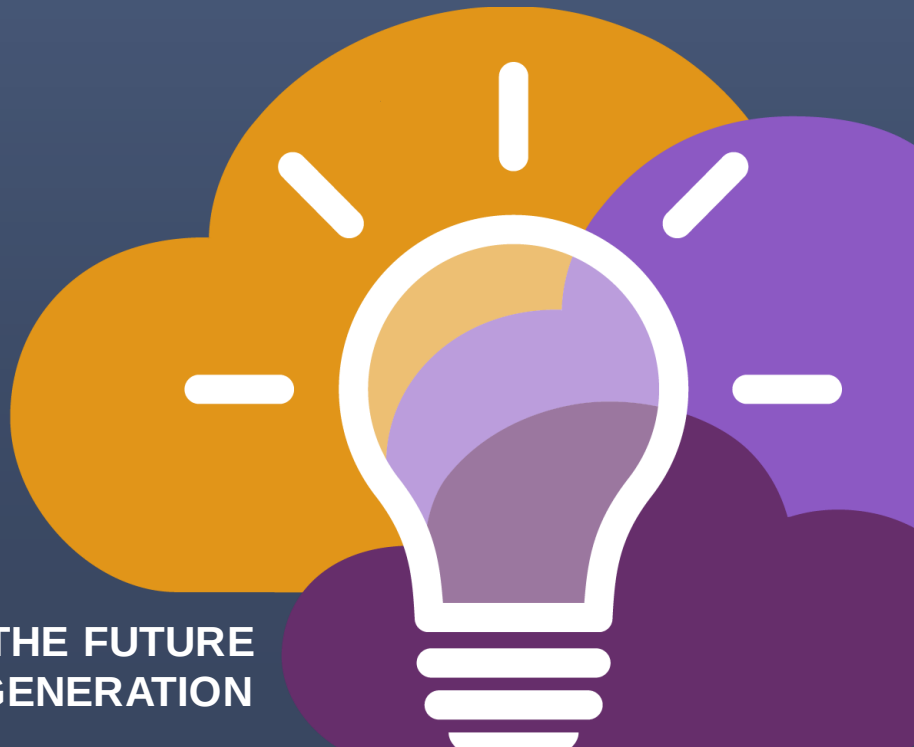
	CASEID	LST	MHINTAKE	OWNERSHP	PUBLICAGENCY	TREATCOGTHRPY	SENIORS	ALZHDEMENTIA	STATE	E_TOTPOP	...	E_HH	E_POV	E_UNEMP	E_PCI	E_NOHSDP
0	201800025	AK	1	2.0	-2.0	1.0	1.0	0.0	ALASKA	738516.0	...	253462.0	77865.0	28067.0	32531.206897	34760.0
1	201800093	AL	1	2.0	-2.0	1.0	1.0	1.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
2	201800099	AL	1	1.0	-2.0	1.0	1.0	1.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
3	201800104	AL	1	1.0	-2.0	1.0	1.0	1.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
4	201800109	AL	1	2.0	-2.0	1.0	1.0	0.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
779	201809433	PR	1	1.0	-2.0	1.0	0.0	0.0	None	NaN	...	NaN	NaN	NaN	NaN	NaN
780	201809435	PR	1	1.0	-2.0	1.0	1.0	1.0	None	NaN	...	NaN	NaN	NaN	NaN	NaN

Slido

What tools or features would make it easier for you to clean and prepare your data?

SCHARE

Live Demo



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

SCHARE

Thank you

Think-a-Thon poll

1. **Rate how useful this session was:**

- ☐ Very useful
- ☐ Useful
- ☐ Somewhat useful
- ☐ Not at all useful

Think-a-Thon poll

2. Rate the pace of the instruction for yourself:

- ☐ Too fast
- ☐ Adequate for me
- ☐ Too slow

Think-a-Thon poll

3. How likely will you participate in the next Think-a-Thon?

- ☐ Very interested, will definitely attend
- ☐ Interested, likely will attend
- ☐ Interested, but not available
- ☐ Not interested in attending any others

SCHARE

Next Think-a-Thons:

bit.ly/think-a-thons

Register for SCHARE:

https://bit.ly/registerschare

schare@mail.nih.gov