DATA MODERNIZATION: AUTOMATED DATA PIPELINES, ADVANCED ANALYTICS & DASHBOARDING

Pima County, Health Department

Hieu Nguyen, Data Engineer Fellow | *University of Chicago, MS in Computational Analysis and Public Policy*

Lidia Ghebreamlak, Data Engineer Fellow | *Iowa State University, PhD in Computer Science*

Keywords:

pipeline automation, data processing, indicator management, public health data, dashboarding

Summary:

Hieu and Lidia worked on several health-related projects. Utilizing Python and R, we streamlined data geoprocessing in ArcGIS and respiratory illness dashboard updates in Qlik to create end-to-end automation. We also created and edited key health indicators in an indicator management system, helping with comprehensive data communication to the public. We had the chance to process and analyze a massive, sparse dataset with modern statistical tools to identify the demographics and non-opioid diagnoses for patients with a history of of opioid overdose.

DATA MODERNIZATION: AUTOMATED DATA PIPELINES, ADVANCED ANALYTICS & DASHBOARDING

Pima County, AZ – Health Department

Supervisor: Amanda Lam,
Data Modernization and Informatics Strategist

coding it forward >

PIMA COUNTY
HEALTH DEPARTMENT

HIEU NGUYEN
University of Chicago
MS in Computational Analysis and Public Policy

LIDIA GHEBREAMLAK
Iowa State University
PhD in Computer Science

We worked on many interesting projects that involve **streamlining data pipelines, analyzing hospital discharge data,** and **managing health indicators**.

cif>

Healthy Pima Indicators

cif>

Healthy Pima Indicators

- We helped with uploading and editing an indicator management system that is helpful for communicating data with the public
- The new dashboard offers quick and comprehensive visualizations of key health indicators

Public and Community Health

	VALUE	COMPARED TO:					
		AZ Counties	U.S. Counties	AZ Value (13.1%)	US Value (13.5%)	Prior Value (10.9%)	Trend
Food Insecurity Rate County: Pima	14.0% (2022)						
Child Food Insecurity Rate County: Pima	19.7% (2022)			AZ Value (18.5%)	US Value (18.5%)	Prior Value (14.7%)	
DTaP Vaccination Rate: Kindergarten Students County: Pima	93.7% (2022)			AZ Value (89.6%)	US Value (92.7%)		
Polio Vaccination Rate: Kindergarten Students County: Pima	94.2% (2022)			AZ Value (90.3%)	US Value (93.1%)		

Healthy Pima Indicators

- The platform automatically graphs trends and can display the data by various breakout groups.
- Each indicator is linked to related data, “promising practices”, and funding opportunities.

Related Content for: Food Insecurity Rate

Indicators MORE →

- Projected Food Insecurity Rate
- Child Food Insecurity Rate
- Projected Child Food Insecurity Rate
- Students Eligible for the Free Lunch Program
- Households Receiving SNAP with Children

Promising Practices MORE →

- Supplemental Nutrition Assistance Program (SNAP) Participation and Health Care Expenditures Among Low-Income Adults
- Portland Fruit Tree Project
- BackPack Program
- Cooking Matters at the Store
- SNAP 2 It!

Funding Opportunities

- Charles Stewart Mott Foundation
- Ford Foundation
- Hearst Foundations
- Michael & Susan Dell Foundation
- W.K. Kellogg Foundation

SocioNeeds Index® Suite

- Food Insecurity Index
- Health Equity Index
- Mental Health Index

cif>

Respiratory Illness Dashboard

cif>

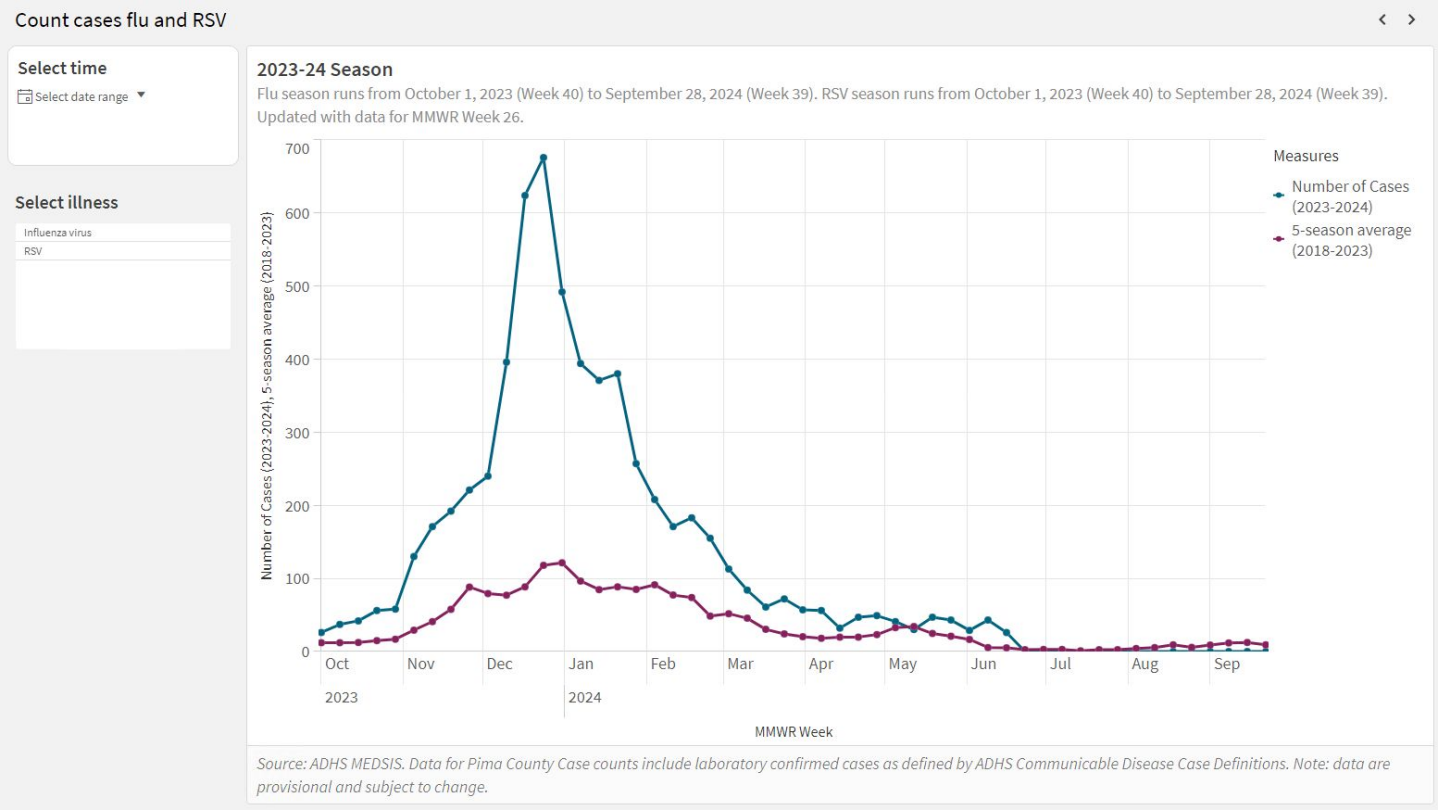
Respiratory Illness Dashboard

- Implemented some R code to create dataframes suitable for Qlik.
- Created the Qlik dashboard with interactive features and automated data updates from SharePoint.
- I had the chance to practice reading other's code alongside learning about standards for public health data visualization.

cif>

Respiratory Illness Dashboard

An example dashboard sheet



cif>

Respiratory Illness Dashboard

R code

Added some R code at the end

Tyler: cleaning, manipulating, and processing data from raw sources

Hieu: creating an Excel file containing processed dataframes from Tyler's code

Qlik connects to SharePoint and updates data daily

```
├──main_R_script.Rmd
├──other_files_and_folders
├──dataframes_upload_qlik        # folder
│  │  all_dataframes_qlik.xlsx        # most updated file for Qlik
│  │
│  └──archive                    # folder
│       all_dataframes_qlik_2024-08-07_12-50-16.xlsx
│       all_dataframes_qlik_2024-08-08_16-53-41.xlsx
│       all_dataframes_qlik_2024-08-08_16-54-06.xlsx
```

cif>

Opioid Overdose Study

cif>

Opioid Overdose Study

- Teamed up with a medical student to process a massive raw hospital discharge dataset
- Aimed to identify characteristics of patients with/without history of opioid overdose
- Literature review on the opioid relationships and data cleaning methods
- Future: continue developing/evaluating more causal models, refine feature selection techniques, and collaborate with healthcare professionals

cif>

Opioid Overdose Study

Process the raw data files from this (conceptually)...

id	sex	race	diagnosis_1	diagnosis_2	diagnosis_3	…	ecode_1	ecode_2	…	procedure_1	procedure_2	…
1	M	5	S123				E123			P123	J321	
2	F	3	Z123	S123	T123		F456			J321		
3	F	2	H123	S123			F456	E123		P123		

cif>

Opioid Overdose Study

... to this

id	sex	race	D_S123	D_Z123	D_H123	D_T123	E_E123	E_F456	P_P123	P_J321	opiod_hx	…
1	0	5	1	0	0	0	1	0	1	1	1	
2	1	3	1	1	0	1	0	1	0	1	0	
3	1	2	1	0	1	1	1	1	1	0	1	

cif>

Opioid Overdose Study

Evaluate the raw data:

- 300,000+ rows per file, each file represents a year from 2016 to 2023
- Too many ICD-10 codes
- Sparse diagnosis, ecode, and procedure columns

Process (each data file):

- Convert ICD-10 into CCSR codes (broader)
- Standardize, frequency-encode categorical variables
- Convert code values into binary columns
- Reduce memory use of dataframes based on data types

cif>

Opioid Overdose Study

After cleaning the data:

- Match 1:1 for opioid vs non-opioid groups based on propensity score and nearest neighbor
- Final, matched data has 1,048,576 observations and 3000+ variables
- Analyzing these binary, sparse columns:
  - quick computation, less memory usage
  - but complex operations on sparse matrices, difficult to interpret

cif>

Opioid Overdose Study

Current step: Variable selection

- Problem: high dimensionality

  => how to select a subset of useful variables?
- Tried several linear testing methods to assess individual significance
- Found some demographics and diagnosis codes (non-opioid) positively correlated with a patient's history of opioid overdose
- But what about non-linear relationships or combinations of weak variables? Future!

cif>

Tribal Data Tool

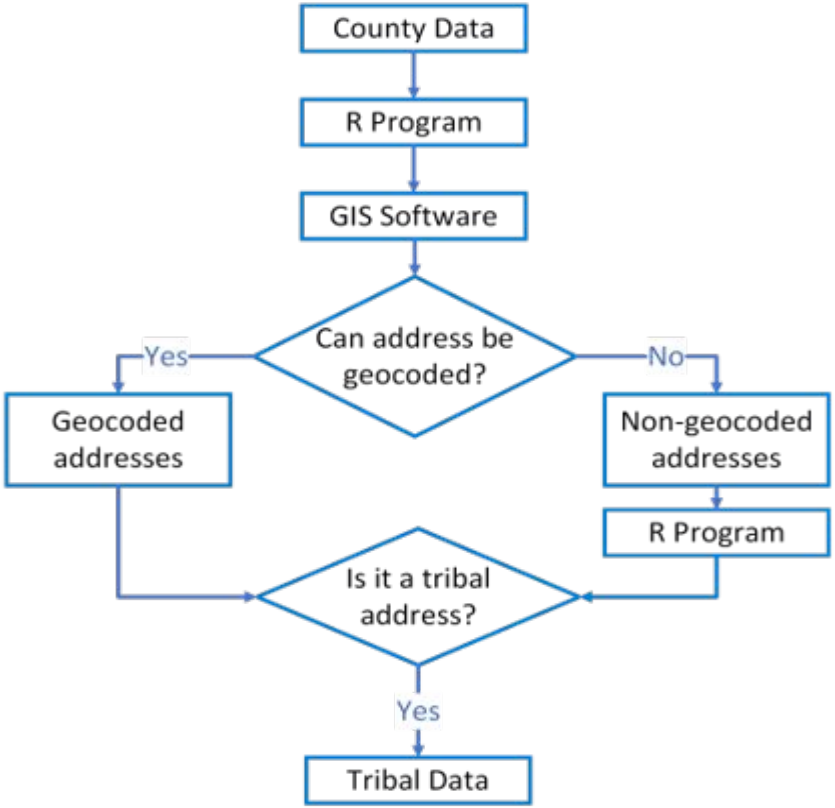
cif>

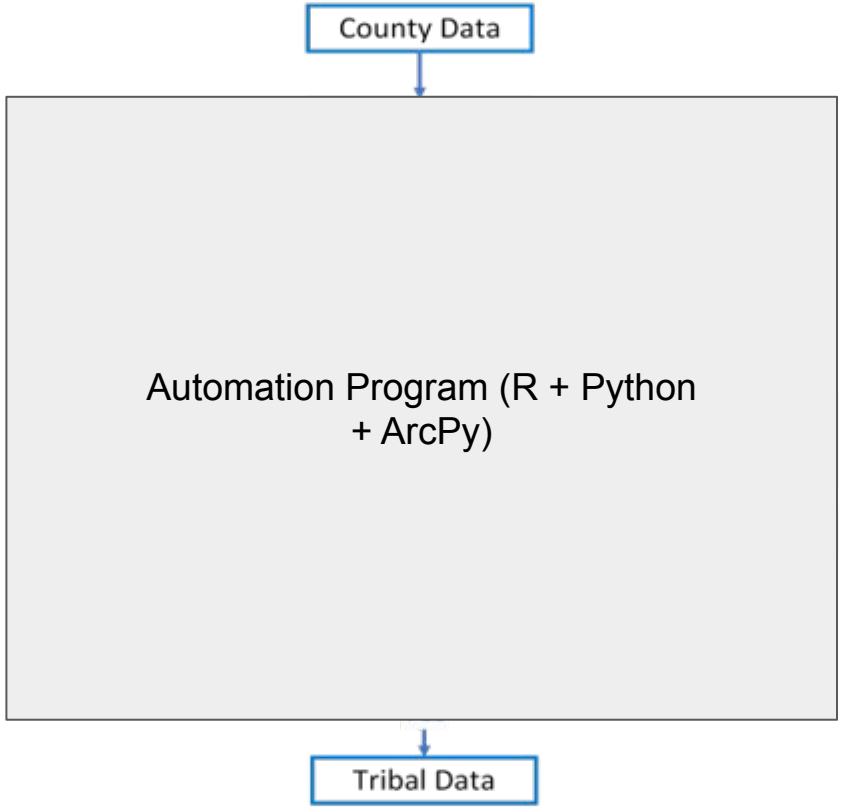
Tribal Data Tool

- Automating the existing manual data geoprocessing in ArcGIS
  - Geocoding
  - Spatial Join
  - Export to Excel
- Integrate python code with R code to get end-to-end automation
- Logging quality check information

cif>

Tribal Data Tool





cif>

Tribal Data Tool

- Log.txt

```
############################

2024_08_14_12_55_01:  Initiating [redacted] program

############################
Check for duplicates----
No duplicates found
############################
Check that reservation name is still [redacted]

############################

2024_08_14_12_55_01:  Initiating geocoding program

############################
```

```
Geocode addresses log messages:  Start Time: Wednesday, August 14, 2024 12:55:36 PM
Executing Geocode Addresses...
60243 Matched (95.75%)
2638 Unmatched (4.19%)
38 Tied (0.06%)
Average speed: 5329609 (records/hour)
Succeeded at Wednesday, August 14, 2024 12:56:19 PM (Elapsed Time: 42.79 seconds)

############################

Spatial Join log messages:  Start Time: Wednesday, August 14, 2024 12:56:20 PM
Succeeded at Wednesday, August 14, 2024 12:56:21 PM (Elapsed Time: 0.94 seconds)

############################

Excel output log messages:  Start Time: Wednesday, August 14, 2024 12:56:21 PM
Succeeded at Wednesday, August 14, 2024 12:56:46 PM (Elapsed Time: 25.03 seconds)

############################
Number of records not previously sent to [redacted] for manual review ---
0
############################
Previous record count to compare with current ---
693
############################
Current record count to compare with previous ---
704
```

Tribal Data Tool

- Directory structure

input_data	8/13/2024 4:25 PM	File folder
[redacted]output_data	8/14/2024 12:55 PM	File folder
latest_[redacted]	8/13/2024 4:52 PM	R Project
PRISM Pre-Geocoding and Post-Geocodi...	8/14/2024 12:58 PM	RMD File
[redacted]Functions	8/13/2024 9:25 AM	R Source File
[redacted]geocoding_python	8/14/2024 7:21 AM	Python File
README.txt	8/14/2024 12:52 PM	Text Document

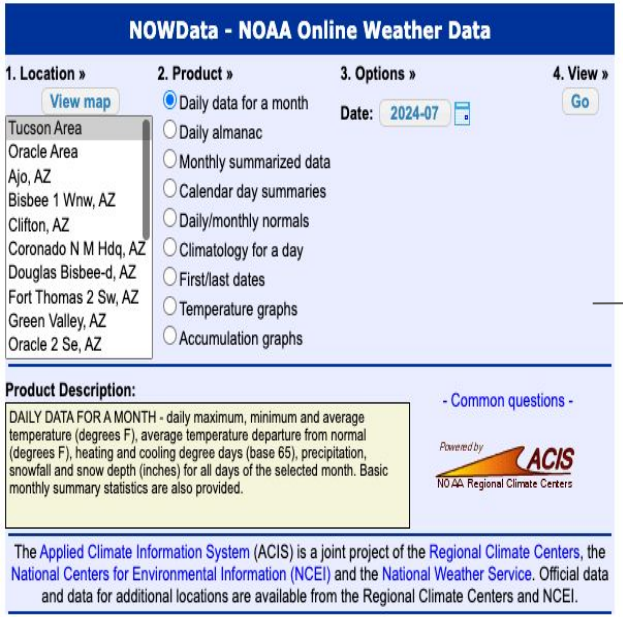
GeocodedData.gdb
prism 2024_08_14_12_55_01

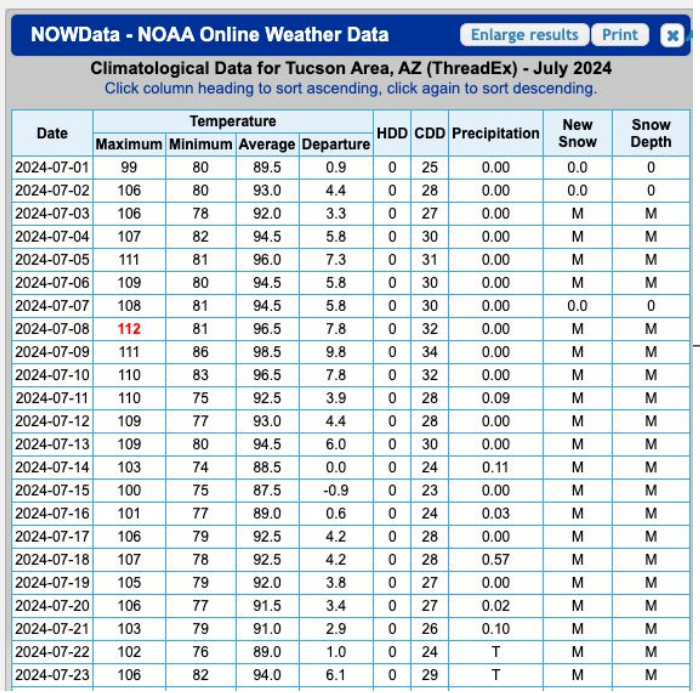
log
PRISM cases for geocoding 2024_08_15_08_51_55
PRISM cases for post geocoding 2024_08_15_08...
PRISM full extract for [redacted]2024_08_15_08_51_55
PRISM records for manual review 2024_08_15_0...

cif>

Extra – web scraping data collection

- Manual downloading

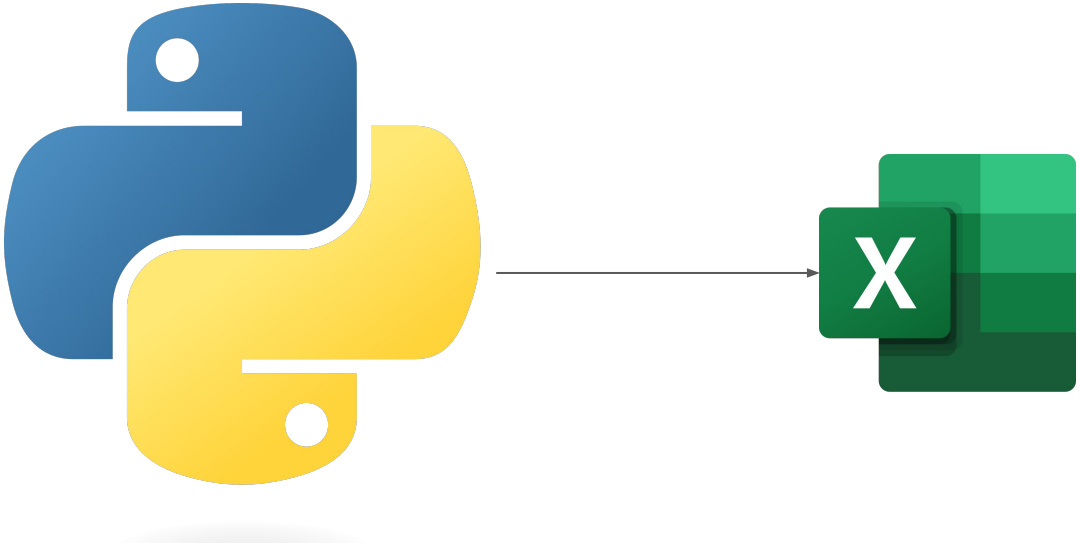




Date	Temperature				HDD	CDD	Precipitation	New Snow	Snow Depth
	Maximum	Minimum	Average	Departure					
2024-07-01	99	80	89.5	0.9	0	25	0.00	0.0	0
2024-07-02	106	80	93.0	4.4	0	28	0.00	0.0	0
2024-07-03	106	78	92.0	3.3	0	27	0.00	M	M
2024-07-04	107	82	94.5	5.8	0	30	0.00	M	M
2024-07-05	111	81	96.0	7.3	0	31	0.00	M	M
2024-07-06	109	80	94.5	5.8	0	30	0.00	M	M
2024-07-07	108	81	94.5	5.8	0	30	0.00	0.0	0
2024-07-08	112	81	96.5	7.8	0	32	0.00	M	M
2024-07-09	111	86	98.5	9.8	0	34	0.00	M	M
2024-07-10	110	83	96.5	7.8	0	32	0.00	M	M
2024-07-11	110	75	92.5	3.9	0	28	0.09	M	M
2024-07-12	109	77	93.0	4.4	0	28	0.00	M	M
2024-07-13	109	80	94.5	6.0	0	30	0.00	M	M
2024-07-14	103	74	88.5	0.0	0	24	0.11	M	M
2024-07-15	100	75	87.5	-0.9	0	23	0.00	M	M
2024-07-16	101	77	89.0	0.6	0	24	0.03	M	M
2024-07-17	106	79	92.5	4.2	0	28	0.00	M	M
2024-07-18	107	78	92.5	4.2	0	28	0.57	M	M
2024-07-19	105	79	92.0	3.8	0	27	0.00	M	M
2024-07-20	106	77	91.5	3.4	0	27	0.02	M	M
2024-07-21	103	79	91.0	2.9	0	26	0.10	M	M
2024-07-22	102	76	89.0	1.0	0	24	T	M	M
2024-07-23	106	82	94.0	6.1	0	29	T	M	M

Extra – web scraping data collection

- Automated process



- Run web_scraping.py

cif>

What we learned

cif>

What we learned

Hieu

- Worked on many data sources and with many people who are doing impactful and meaningful work
- Raw data are messy, requiring flexible and consistent manipulation and integration
- Inspired to work on future health-related projects
- Learned and practiced Snowflake and PySpark basics

cif>

What we learned

Lidia

- Tribal data sensitivity and challenges
- ArcGIS Pro software and Python libraries (ArcGIS Python API, ArcPy)
- Public health data through the indicators project and general data systems in PCHD
- Important departments and the valuable work done in the PCHD department

cif>

Future Recommendations

cif>

Future Recommendations

- Snowflake holds great potential to enhance data integration and enable scalable, secure storage for large health datasets from different sources.
- This would help with Epidemiology team's public health surveillance reporting and monitoring as well as research into causal inferences.
- Utilizing Snowflake, Qlik and/or ArcGIS to maintain data systems that can replace third party systems like MySidewalk.

cif>

Acknowledgments:
Amanda Lam and all the people at PCHD

Thank you for hosting us this summer!

Stay connected:

Hieu - doanquanghieu.nguyen@gmail.com
Lidia - lidiat0408@gmail.com

cif>