Designing a Prototype Environmental Health Data Dashboard for North Carolina

Data + Team: Jerry Fu, Leah Roffman, and Anna Zolotor
Project Manager: Melyssa Minto
Project Leads: Mike Dolan Fliss, Kim Gaetz

Abstract

Health and wellbeing are largely dependent on both the natural and built environment. More research is published every day which identifies how pollutants in our air or water can affect a number of health outcomes. In recognition of this, the Centers for Disease Control and Prevention (CDC) created the Environmental Public Health Tracking (EPHT) program to visualize population health measures alongside environmental health measures. One of the major goals of the CDC EPHT program is to help states provide a user-friendly portal through which community members, public health professionals, and policy makers can access environmental health data. Currently, North Carolina does not have its own EPHT program (only half of all states in the U.S. have funded EPHT programs). In this project, the team developed a prototype for an EPHT tool for North Carolina in the form of an Environmental Health Data Dashboard (EHDD) which can later be used to apply for a CDC grant for a fully-funded EPHT program.

This pilot includes the integration of data such as air quality, emergency department visits, and demographics data processed in RStudio and visualized as a dashboard in a Tableau (Fig 1).
The three main aspects of this project were 1) Development of metadata, 2) Processing of raw datasets, and 3) Visualization in Tableau. In the course of the project, the team filled out descriptive metadata for over 150 required and non-required measures, processed and harmonized 12 raw datasets, and designed six interactive dashboard pages. An analysis of various case studies run on the prototype demonstrated the potential of the EPHT tool.

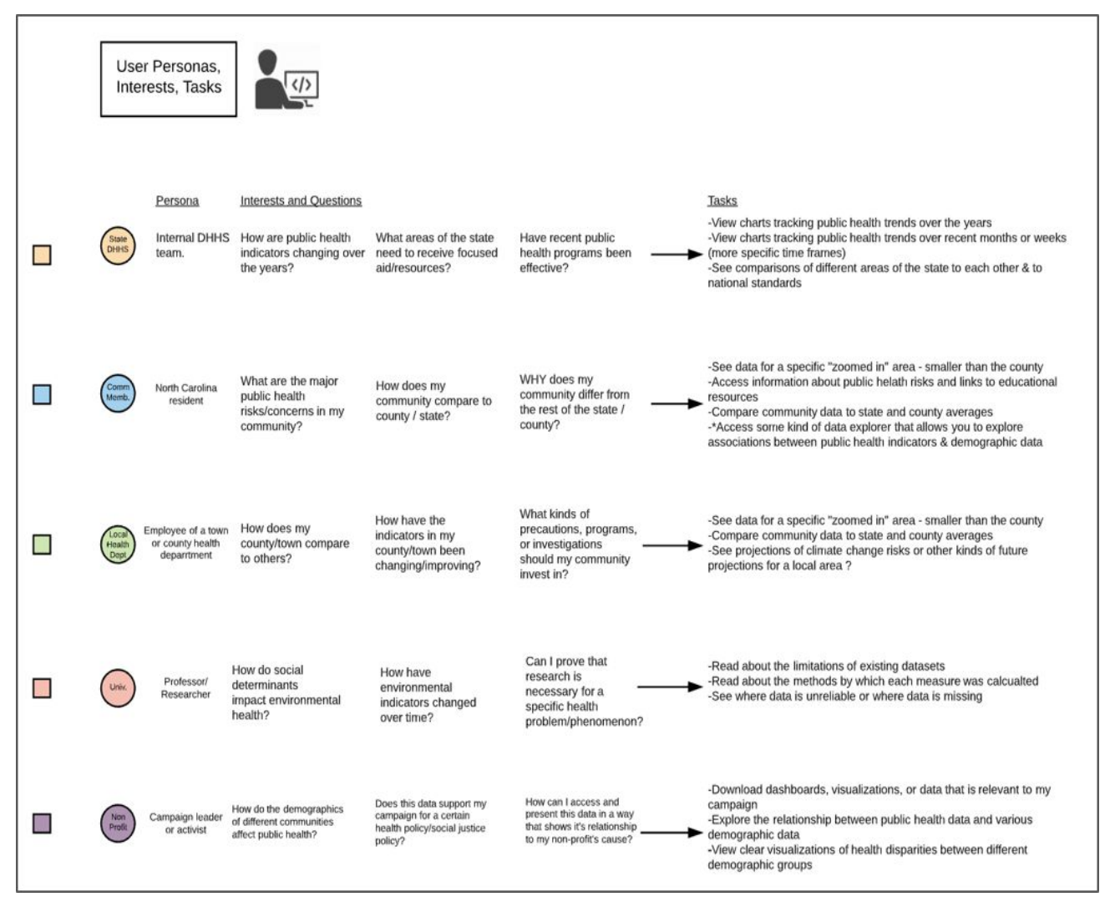
Fig 1: These icons are representations of the software used throughout the project. a) RStudio was used for data processing, b) Tableau was used for the development of the environmental health data dashboard and the metadata dashboard, and c) git was used for version control management.

Initial goals

The major goals of this project were to create a prototype data dashboard, in the form of a Tableau story, a metadata document to house various information about each measure, and a package of R scripts to document our data processing workflow and aid future developers. The team developed goals for the prototype dashboard by envisioning what dashboard users would want to be able to get from the dashboard. Accessibility and users' situational needs were considered by compiling user stories as shown in Fig. 2.

Based on these user stories, the team concluded that our dashboard prototype should be able to:

- Show yearly data for each measure in a specific county or tract
- Show comparisons of data trends for two or more specific locations
- Visualize relationships between demographic and socioeconomic data and environmental health data.



Persona	Interests and Questions			Tasks
Internal DHHS team.	How are public health indicators changing over the years?	What areas of the state need to receive focused aid/resources?	Have recent public health programs been effective?	-View charts tracking public health trends over the years -View charts tracking public health trends over recent months or weeks (more specific time frames) -See comparisons of different areas of the state to each other & to national standards
North Carolina resident	What are the major public health risks/concerns in my community?	How does my community compare to county / state?	WHY does my community differ from the rest of the state / county?	-See data for a specific "zoomed in" area - smaller than the county -Access information about public helath risks and links to educational resources -Compare community data to state and county averages -*Access some kind of data explorer that allows you to explore associations between public health indicators & demographic data
Employee of a town or county health deparrtment	How does my county/town compare to others?	How have the indicators in my county/town been changing/improving?	What kinds of precautions, programs, or investigations should my community invest in?	-See data for a specific "zoomed in" area - smaller than the county -Compare community data to state and county averages -See projections of climate change risks or other kinds of future projections for a local area ?
Professor/ Researcher	How do social determinants impact environmental health?	How have environmental indicators changed over time?	Can I prove that research is necessary for a specific health problem/phenomenon?	-Read about the limitations of existing datasets -Read about the methods by which each measure was calcualted -See where data is unreliable or where data is missing
Campaign leader or activist	How do the demographics of different communities affect public health?	Does this data support my campaign for a certain health policy/social justice policy?	How can I access and present this data in a way that shows it's relationship to my non-profit's cause?	-Download dashboards, visualizations, or data that is relevant to my campaign -Explore the relationship between public health data and various demographic data -View clear visualizations of health disparities between different demographic groups

Fig 2: The five user stories were designed to reflect the needs of five different potential users of the EHDD. The interests of each user were translated into "tasks" the user might seek to complete, which were used to brainstorm features to be included in the dashboard.

Data Overview

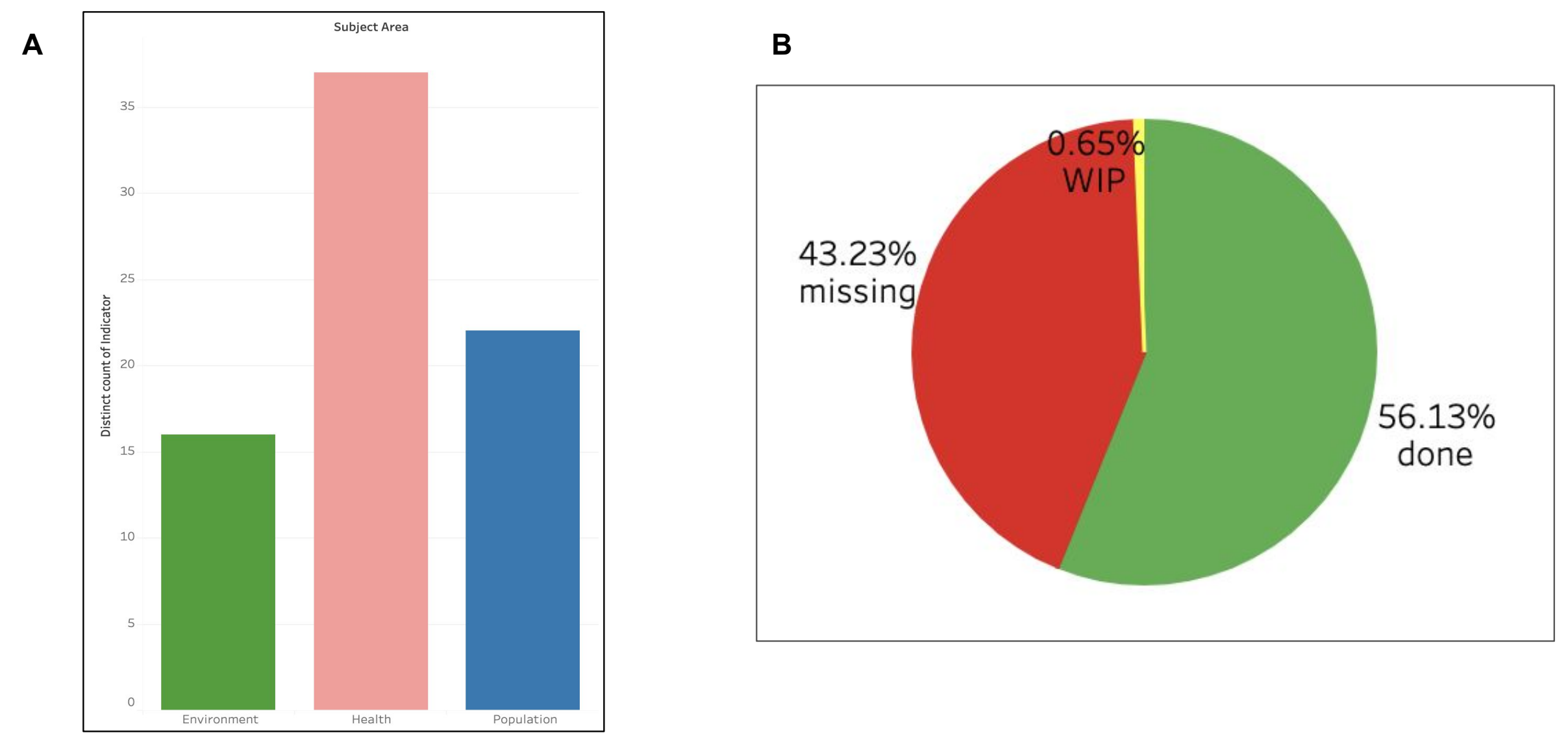
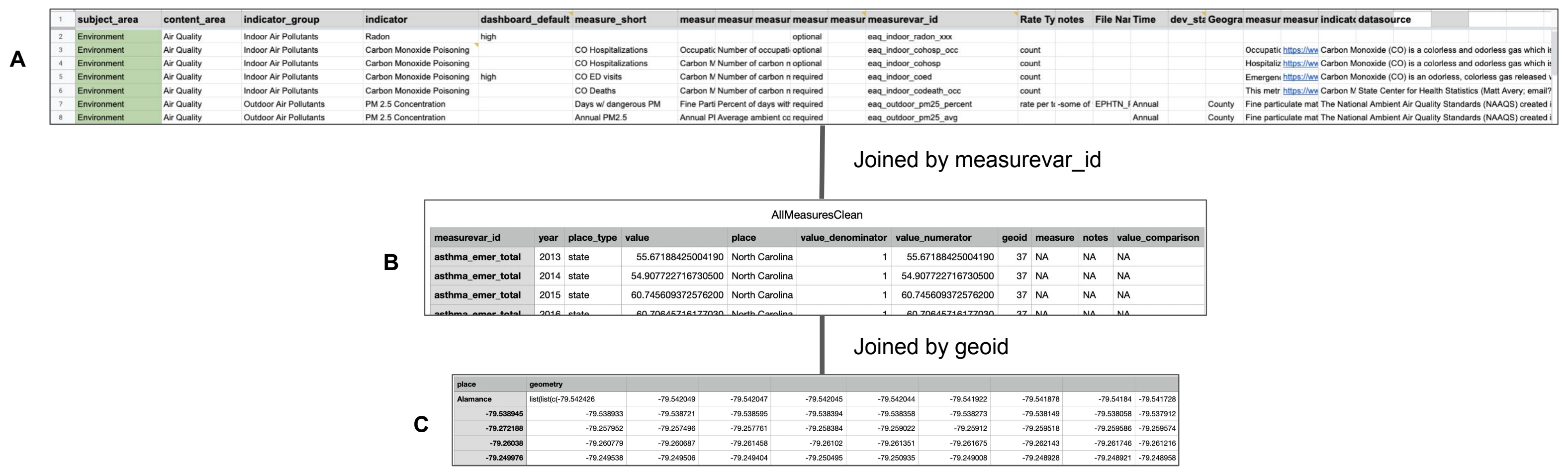


Fig 3: These figures from the metadata dashboard show the distribution of the measures in our metadata a) by subject area and b) by development status. 56% of the measures in the metadata are finished being processed while others are missing.

Our data comes from a variety of sources, including the CDC and NOAA, and contains environment, health, and population measures. The organizational hierarchy of the data is subject area > content area > indicator group > indicator > measure > measure variant. At a higher level, we have 16 environment indicators, 37 health indicators, and 22 population indicators listed in our metadata. At the smallest level, there may be several measure variants for a single measure, ex) Rate of emergency department visits for COPD by age groups 25-44, 45-64, 65-84, and 85 and older.

Methods and Workflow

All data processing centered around standardizing data to match the "data skeleton," a set of 11 columns that each set of measure data was standardized to. Those columns include three that can be used to uniquely identify any observation: year, geoid (a numeric code that indicates the place the observation is from), and measure variant identification (a code that refers to the specific way the observation was measured, as well as what was measured).
The metadata includes columns for information about every measure variant, including measure descriptions, development status, whether or not the measures is required by the CDC for the prototype, and who provided the data. Each row represents a unique measure variant. This spreadsheet is joined into the EHDD by measure id and is also used in a separate metadata visualizations dashboard.



1	subject_area	content_area	indicator_group	indicator	dashboard_default	measure_short	measur	measur	measur	measur	measur	measurevar_id	Rate Ty	notes	File Nai	Time	dev_sta	Geogra	measur	measur	indicatc	datasource
2	Environment	Air Quality	Indoor Air Pollutants	Radon	high						optional	eaq_indoor_radon_xxx										
3	Environment	Air Quality	Indoor Air Pollutants	Carbon Monoxide Poisoning		CO Hospitalizations	Occupatic	Number of occupati			optional	eaq_indoor_cohosp_occ	count						Occupatic	https://ww	Carbon Monoxide (CO) is a colorless and odorless gas which is	
4	Environment	Air Quality	Indoor Air Pollutants	Carbon Monoxide Poisoning		CO Hospitalizations	Carbon M	Number of carbon n			optional	eaq_indoor_cohosp	count						Hospitaliz	https://ww	Carbon Monoxide (CO) is a colorless and odorless gas which is	
5	Environment	Air Quality	Indoor Air Pollutants	Carbon Monoxide Poisoning	high	CO ED visits	Carbon M	Number of carbon n			required	eaq_indoor_coed	count						Emergen	https://ww	Carbon Monoxide (CO) is an odorless, colorless gas released v	
6	Environment	Air Quality	Indoor Air Pollutants	Carbon Monoxide Poisoning		CO Deaths	Carbon M	Number of carbon n			required	eaq_indoor_codeath_occ	count						This metr	https://ww	Carbon M	State Center for Health Statistics (Matt Avery; email?
7	Environment	Air Quality	Outdoor Air Pollutants	PM 2.5 Concentration		Days w/ dangerous PM	Fine Parti	Percent of days witl			required	eaq_outdoor_pm25_percent	rate per t	-some of	EPHTN_F	Annual		County	Fine particulate mat		The National Ambient Air Quality Standards (NAAQS) created i	
8	Environment	Air Quality	Outdoor Air Pollutants	PM 2.5 Concentration		Annual PM2.5	Annual Pl	Average ambient cc			required	eaq_outdoor_pm25_avg				Annual		County	Fine particulate mat		The National Ambient Air Quality Standards (NAAQS) created i	

AllMeasuresClean

measurevar_id	year	place_type	value	place	value_denominator	value_numerator	geoid	measure	notes	value_comparison
asthma_emer_total	2013	state	55.67188425004190	North Carolina	1	55.67188425004190	37	NA	NA	NA
asthma_emer_total	2014	state	54.907722716730500	North Carolina	1	54.907722716730500	37	NA	NA	NA
asthma_emer_total	2015	state	60.745609372576200	North Carolina	1	60.745609372576200	37	NA	NA	NA
asthma_emer_total	2016	state	60.70645716177030	North Carolina	1	60.70645716177030	37	NA	NA	NA

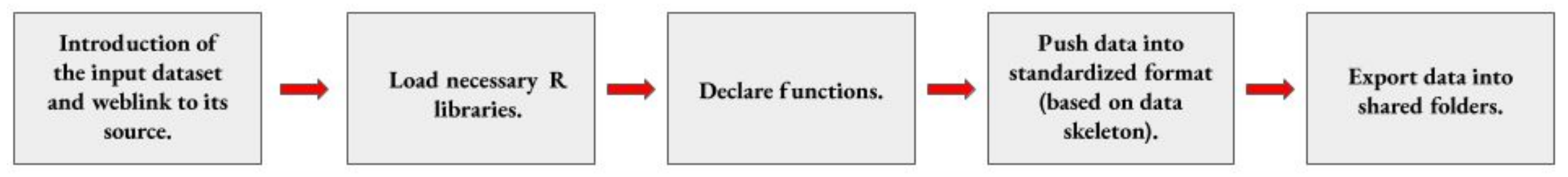
place	geometry								
Alamance	list(list(c(-79.542426	-79.542049	-79.542047	-79.542045	-79.542044	-79.541922	-79.541878	-79.54184	-79.541728
-79.538945	-79.538933	-79.538721	-79.538595	-79.538394	-79.538358	-79.538273	-79.538149	-79.538058	-79.537912
-79.272188	-79.257952	-79.257496	-79.257761	-79.258384	-79.259022	-79.25912	-79.259518	-79.259586	-79.259574
-79.26038	-79.260779	-79.260687	-79.261458	-79.26102	-79.261351	-79.261675	-79.262143	-79.261746	-79.261216
-79.249976	-79.249538	-79.249506	-79.249404	-79.250495	-79.250935	-79.249008	-79.248928	-79.248921	-79.248958

Fig 5: The data that powers the dashboard relies on joins between a) The metadata, created as a google sheet, b) The measure data, which contains vertically joined cleaned data from various data sources, and c) The multipolygon data, which is displayed here as a CSV but is actually a shapefile used for visualizing the data on a map.

Data processing

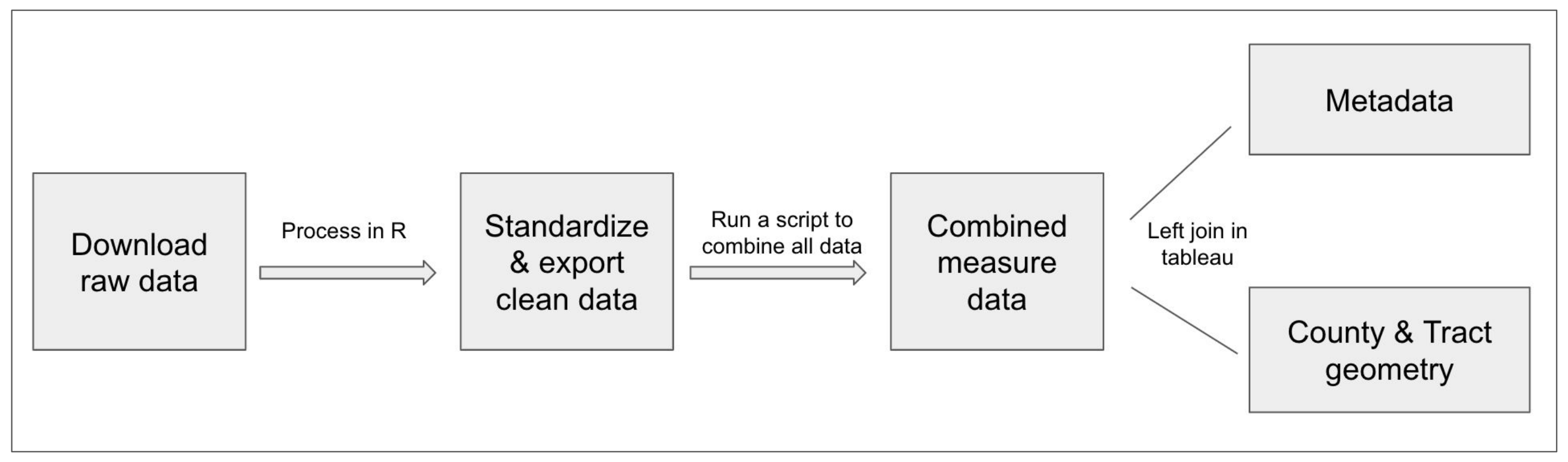
All data processing was done in RStudio. As each new dataset was received, the goal was to push it into a standardized form based on the "data skeleton" (as discussed above).

The general order and content of each script are as follows:



Introduction of the input dataset and weblink to its source. → Load necessary R libraries. → Declare functions. → Push data into standardized format (based on data skeleton). → Export data into shared folders.

Key data processing challenges included variations in the way places were named (for example, some county names were recorded in all capital letters, so converting them into lowercase changed the name of McDowell county to Mcdowell, which resulted in join failures), differences in format from year to year, and the codes for American Community Survey variables changing from year to year.



Download raw data → Process in R → Standardize & export clean data → Run a script to combine all data → Combined measure data → Left join in tableau → Metadata; County & Tract geometry

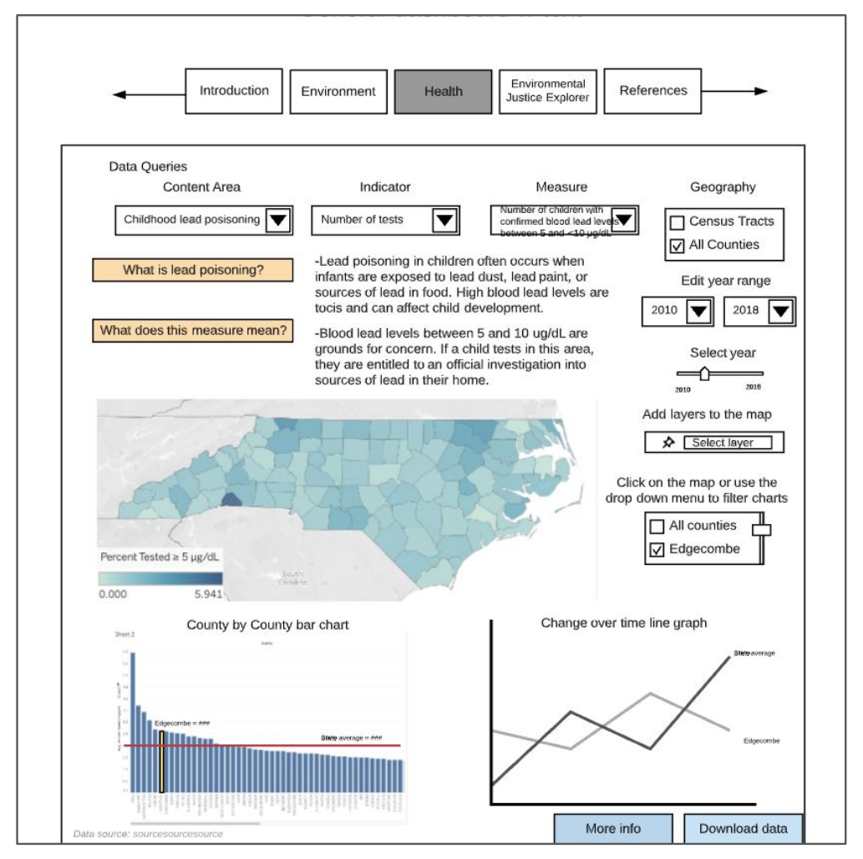
Fig 6: This flowchart shows the overarching data processing workflow. The data is standardized, harmonized, and exported using RStudio and then joined with the metadata in Tableau.

Dashboard Design

Although the EHDD prototype is referred to as a dashboard, in actuality it consists of six separate Tableau dashboards displayed as tabs in a Tableau story. The process of designing each dashboard page began in week one by reviewing states' existing tools and wireframing pages in Lucidchart. In order to translate our wireframes into Tableau, several parameters, calculated fields and sets were created to help filter the data.

Several challenges were involved in designing the dashboards:

- Creating the map legends required making dynamic value bins to reflect the range of values for each measure.
- Displaying the units for various types of measures in the dashboard required using calculations to selectively display both rate and raw count values for certain measures
- The existence of duplicate census tract names in North Carolina required creating several calculated fields to display each place name uniquely

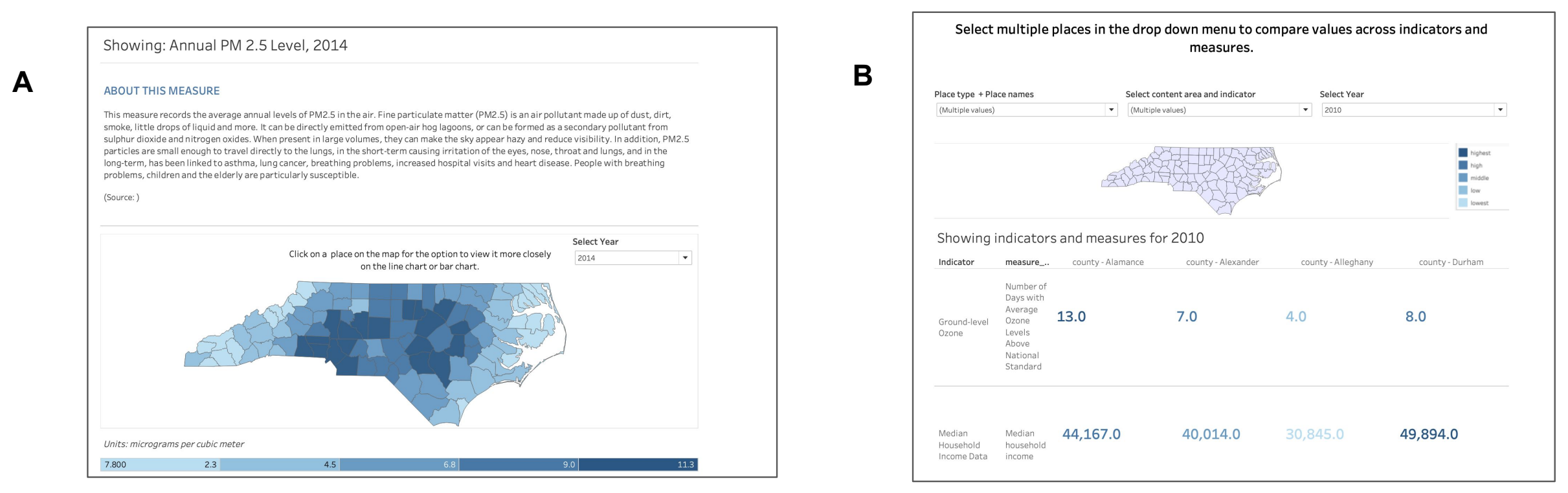


-Lead poisoning in children often occurs when infants are exposed to lead dust, lead paint, or sources of lead in food. High blood lead levels are tocis and can affect child development.

-Blood lead levels between 5 and 10 ug/dL are grounds for concern. If a child tests in this area, they are entitled to an official investigation into sources of lead in their home.

Fig 7: In the initial wireframes completed in Lucidchart, the map dashboard included map layers such as hospitals, flood zones, and highways.

Results and Deliverables


A
Showing: Annual PM 2.5 Level, 2014
ABOUT THIS MEASURE
(Source:)
Click on a place on the map for the option to view it more closely on the line chart or bar chart.
Select Year
2014
Units: micrograms per cubic meter
7.800
2.3
4.5
6.8
9.0
11.3
B
Select multiple places in the drop down menu to compare values across indicators and measures.
Place type + Place names
(Multiple values)
Select content area and indicator
(Multiple values)
Select Year
2010
highest
high
middle
low
lowest
Showing indicators and measures for 2010

This measure records the average annual levels of PM2.5 in the air. Fine particulate matter (PM2.5) is an air pollutant made up of dust, dirt, smoke, little drops of liquid and more. It can be directly emitted from open-air hog lagoons, or can be formed as a secondary pollutant from sulphur dioxide and nitrogen oxides. When present in large volumes, they can make the sky appear hazy and reduce visibility. In addition, PM2.5 particles are small enough to travel directly to the lungs, in the short-term causing irritation of the eyes, nose, throat and lungs, and in the long-term, has been linked to asthma, lung cancer, breathing problems, increased hospital visits and heart disease. People with breathing problems, children and the elderly are particularly susceptible.

Indicator	measure_..	county - Alamance	county - Alexander	county - Alleghany	county - Durham
Ground-level Ozone	Number of Days with Average Ozone Levels Above National Standard	13.0	7.0	4.0	8.0
Median Household Income Data	Median household income	44,167.0	40,014.0	30,845.0	49,894.0

Fig 8: These figures show a) a partial screenshot of the prototyped Map Dashboard in Tableau, and b) a partial screenshot of the prototyped Comparison Dashboard in Tableau.

Our final dashboard prototype contained six dashboard pages in total: a welcome page, a map dashboard, a content area explorer, an indicator zoom, a fact sheet, and a comparison dashboard. Each page allows for user interactivity through drop down menus that will filter visualizations by content area, indicator and explorer. In addition, the team produced a metadata dashboard to help with data management, and a write-up paper to document challenges and suggestions for future development.

Analysis and Limitations

In order to test the usefulness of our prototype, the team ran several test cases on it by attempting to use the dashboards to answer questions about environmental health concerns in North Carolina.

- One case study aimed to detect disparities in asthma emergency department visits around low-income housing stock like mobile home parks. The prototype was helpful in identifying counties in which there were noticeably high rates of asthma ED visits, as well as counties with a high percent of the population spending more than 30% of income on housing. However, neither of these measures were available at a census tract level of aggregation.
- In this test case and others, one of the greatest limitations of the prototype is the lack of tract-level and neighborhood-level data.

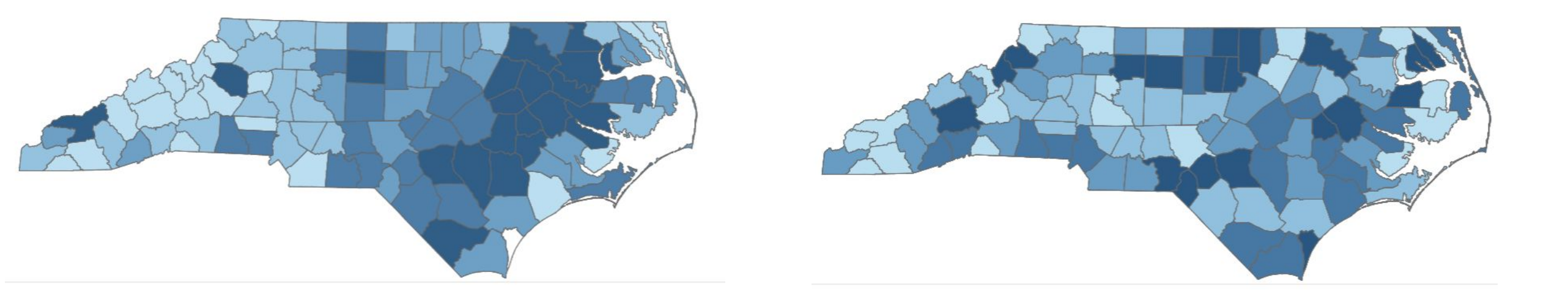


Fig. 9 and 10, a) Rate of asthma ED visits in 2019, b) Percent of population spending 30 percent or more of income on housing, 2018

Directions for Further Development

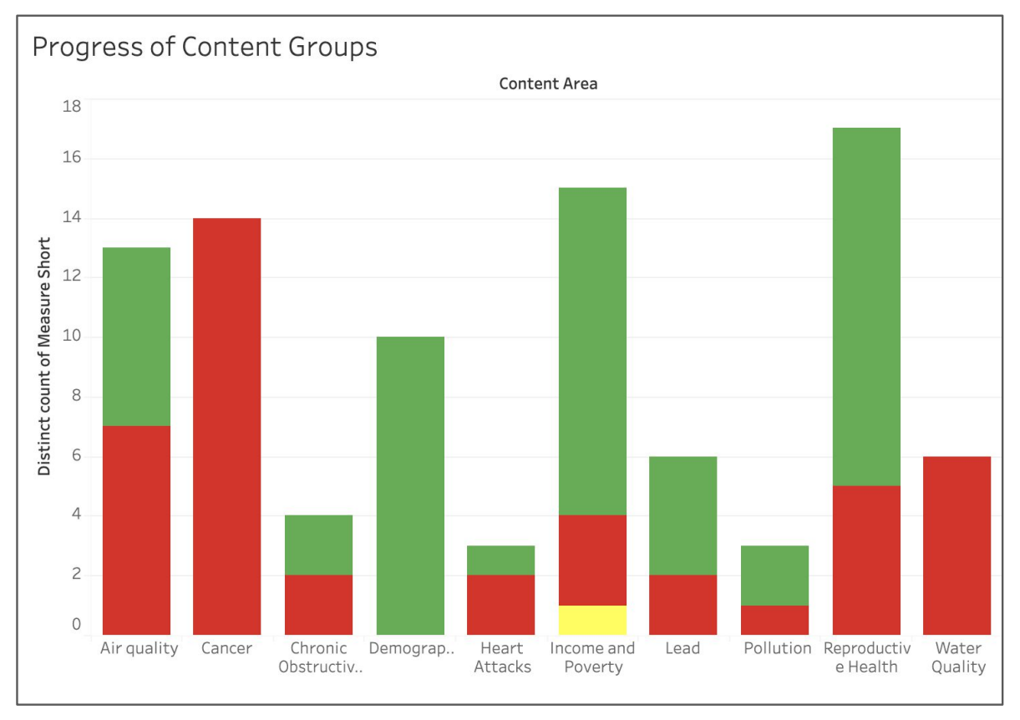


Fig 11: In this figure from the metadata dashboard, red bars indicate the number of measures within each content area that have yet to be processed.

One of the first and most important steps for future developers will be to acquire and process the rest of the CDC-required indicators, including hospitalizations, cancer, and pesticide prevalence.

There are also several other areas for development:

- Creating a vulnerability index specific to environmental health.
- Continuing to perfect language of measure descriptions and improve accessibility.
- Acquiring data at more granular geographic levels, including community data if possible

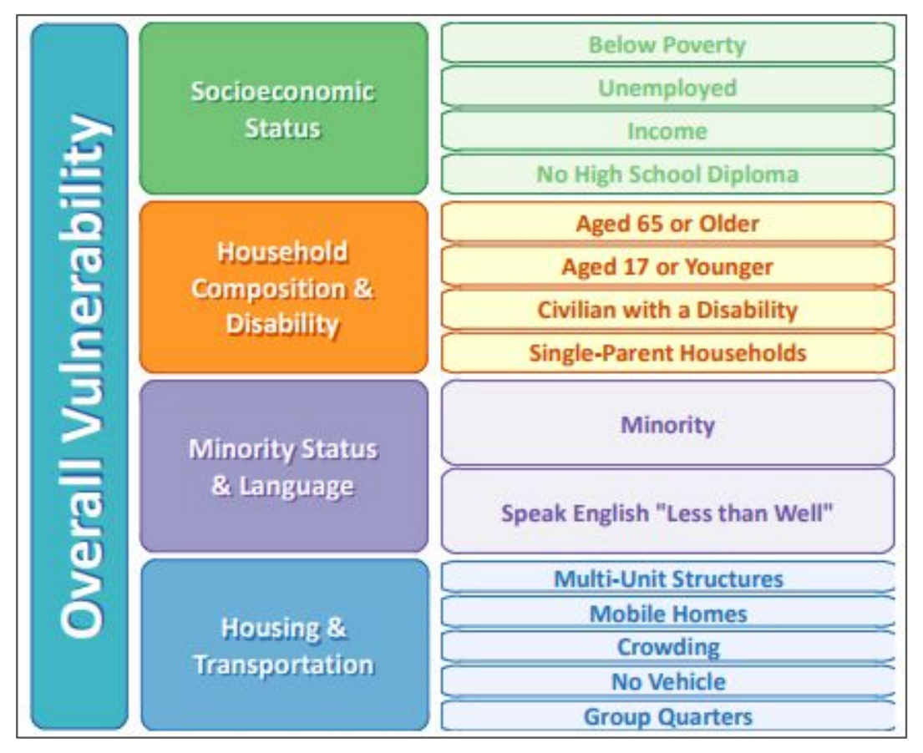


Fig 12: The CDC currently structures their social vulnerability index by equally weighing four different categories of vulnerability measures.

Acknowledgements

The datasets for this project were supplied by several people from North Carolina's Department of Health and Human Services.

Thank you to our project mentors, Dr. Mike Dolan Fliss and Dr. Kim Gaetz, and our project manager Melyssa Minto for all their help over the course of this project