README for replication of "Census Curated Data Enterprise Use Case Demonstration: Climate Resiliency of Skilled Nursing Facilities," Lancaster V, Shipp S, Keller S, Schroeder A, Mortveit H, Swarup S, Xie D, (Updated 2023)

Vicki A. Lancaster, November 2023

Overview

This README document is a guide to the data, documents, and code for replicating the Skilled Nursing Facility (SNF) Use Case demonstration that are in the GitHub repository. The structure and contents of the repository can be viewed here (https://github.com/uva-bi-sdad/census_cde_demo_2/tree/main.

Data Availability Statements

Statement about Rights

- I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

License for Data

The data created by this package from the raw source data are licensed under a Creative Commons/CC-BY license. The raw source data are in the public domain.

Summary of Availability

- The data sources that are publicly available and free to download are included in this. Two output data files generated by proprietary software are also included.

Details on each Data Source

All the data sources used, with the exception of two, are publicly available and free to download. Two of the data sources were generated by proprietary software. The output is available, but the code is not. The repository data folder contains three subfolders, community, proprietary data, and virginia_skilled_nursing_facility. The data sources are described below under these three topic headings.

Not all the data sources described below or in the repository were used to construct the final data tables, but all were evaluated for completeness and fitness-for-use.

Data Sources for Community

The community subfolder contains three sub subfolder topics, climate_change_risk, demographic, and resilience. A description of the data sources within the subfolders are described below.

Climate Change Risk
Department of Homeland Security, Federal Emergency Management Agency
National Risk Index (NRI)

The National Risk Index is a dataset that provides estimates of the risks [0 – 100] for 18 natural hazards at both the tract and county levels. The National Risk Index Technical Document (2023) describes the concepts and methodology used to develop the National Risk Index.

The data are in the public domain and are available at both the tract and county levels in the repository. The variable names and contents are explained in the data dictionary available in the repository.

The data file for the National Risk Index at the tract level is virginia_fema_nri_census_tract.csv.

Demographics
Census Bureau
American Community Survey (ACS)

The subfolder demographic does not contain data but rather information on how to download the data from the Census API using the functions `get_acs` from the **tidycensus** R package (version 1.3.2). Before you can use tidycensus, you must first acquire a free API Key from the Census Bureau. Click here to do so. The tidycensus package is built off of the Census API, so the functions in the package will not work until work until you install the API Key using the following R code:

```
census_api_key("YOUR KEY GOES HERE", install = TRUE).
```

Information on tidycensus package in th online document, *Analyzing US Census Data: Methods, Maps, and Models in R*, by Kyle Walker.

The demographic data used in the paper are from the American Community Survey (ACS) 5-YR (2016-2020). We constructed a community resilience index at both the track and county levels using the ACS tables:

- B18101 Sex by Age by Disability Status;
- B23025 Employment Status for the Population 16 Years and Over;
- B08201 Household Size by Vehicles Available; and
- S1501 Education Attainment.

The code that downloads and uses these data sources and the community resilience index at the census tract level is VA_Population_Resilience_Index_Census_Tract.R.

Resilience
Department of Homeland Security, Geospatial Management Office
Homeland Infrastructure Foundation Level Data (HILFD)

The HILFD provides foundation-level geospatial data in the public domain. The data sources used in this research are available in the repository. There is no data dictionary for the HILFD data, the file and variable names are self-explanatory.

American Red Cross Chapters

The data file for the American Red Cross Chapters is va_hifld_american_red_cross_chapter_facilities_2022.csv.

Emergency Medical Service Stations

The data file for the Emergency Medical Service Stations is va_hilfd_emergency_medical_service_stations_2022.csv.

Fire Stations

The data file for the Fire Stations is va_hilfd_fire_stations_2020.csv.

Hospitals

The data file for the Fire Stations is va_hilfd_hospitals_2022.csv.

National Shelter System Facilities

The data file for the National Shelter System Facilities is va_hilfd_national_shelter_system_facilities_2022.csv.

Urgent Care Facilities

The data file for Urgent Care Facilities is va_hilfd_urgent_care_facilities_2018.csv.

Department of Health & Human Services, Health Resources & Services Administration
Health Professional Shortage Areas (HPSA)

HPSA are geographic areas, population groups, or health care facilities that has been designated by the HRSA as having a shortage of health professionals, in this case primary care physicians. The variable names and contents are explained in the HAPSA data dictionary available in the repository. Additional information can be found in the User Documentation for the County Area Health Resources File (2020-2021).

The data file for the Health Professional Shortage Areas is va_hrsa_ahrf_2021.csv.

Data Sources Generated by Proprietary Software

In addition to the publicly available data are the data from proprietary (commercial and non-commercial) software used to construct the transportation routes to the SNFs. In this case the code cannot be shared. Proprietary software were used for this Use Case to amplify the mix of software types and governance complexities the CDE will need to address. They software packages are described below and links to the output data files are provided.

The BI NSSAC Building Database for the US (BDB-1.0). The description of the modeling and the data fields included in this product are in *NSSAC Building Knowledge Base Modeling and Implementation* (NSSAC Technical Report 2021-16). (Mortveit, Xie, & Marathe, 2023)

The full description of the HERE Premium Streets data (2021/Q1) is provided to customers in a proprietary document (HERE NAVSTREETS Reference Guide v15.1.pdf for their 2022 Q1 data), which, as part of their terms of use, we are unable share. Examples of fields extracted include link IDs, function class (hierarchy level), speed limits, number of lanes, bi-/uni-directionality, geometry, length, road access restrictions (e.g., bus lanes, HOV lanes), road link additional details (e.g., tunnel, bridge, ferry), and details regarding evacuation.

The output file from the proprietary software is va_snf_flood_risks.csv.

The output file from the proprietary software is va_snf_to_si_location_mapping.csv.

Data Sources for Virginia Skilled Nursing Facilities

The virginia_skilled_nursing_facilities subfolder contains four sub subfolder topics, facility, nursing_staff, owners, and residents. A descriptions of the data sources are provided below by topic.

Center for Medicare & Medicaid (CMS)

Nursing Homes Including Rehab Services Datasets

The Virginia skilled nursing facility subfolder contains four sub subfolder topics, facility, nursing_staff, owners, and residents. The variable names and contents are explained in a single data dictionary for all CMS SNF files (except payroll) and is available in the metadata folders. Additional technical information is provided in the Design for Care Compare Nursing Home Five-Star Rating System: Technology User's Guide also contained in the metadata folders. In most cases when downloading data from CMS it can by filtered by state. SMA description of the data sources within the subfolders are described below.

Facility

There are three data files in the facility sub subfolder, inspection dates, safety deficiencies received, and provider information used in the Five-Star Rating System. All are in the public domain. A list of nursing home *inspection dates* in the past three years, including health inspections, fire safety inspections, complaint inspections and infection control inspections.

The data file for the SNF inspection dates is us_cms_inspection_dates_2022-06.csv.

A list of *nursing home fire safety (nonhealth) deficiencies* in the last three years, including the associated inspection date, citation tag number and description, scope and severity, the current status of the citation and the correction date. Data are presented as one citation per row. A description of the SNF deficiency codes is available in the metadata folder.

The data file for the SNF nonhealth deficiencies is va_cms_fire_safety_deficiencies_2022-12.csv.

General information on currently active nursing homes, including number of certified beds, quality measure scores, staffing and other information used in the Five-Star Rating System. Data are presented as one row per nursing home.

The data file for the SNF providers is va_cms_provider_final_2022-07.csv.

Nursing Staff
There is one data file in the nursing_staff sub subfolder, it is contained in the public domain and free to download. The variable names and contents are explained in the Payroll Based Journal Daily Nursing Staffing Data Dictionary available in the metadata folder as is the technical document Payroll-Based Journal Public Use Files: Technical Specifications.

The Payroll Based Journal Employee Detail Public Use File detailed information on the hours worked by individual employees. These are based on data submitted by nursing homes to CMS through the Payroll Based Journal (PBJ) system.

The data file for the Nursing Staff is va_cms_pbj_puf_payroll_nursing_staff_2019-Q4.csv

Owners
There is one data file in the owners sub subfolder, it is contained in the public domain and free to download.

The data file for the Owners is va_cms_ownership_2022-08.csv

Residents
There is one data file in the residents sub subfolder, it is contained in the public domain and is free to download from Brown University Center for Gerontology and Healthcare Research after registering for an account at https://ltcfocus.org/data. The variable names and contents are explained in the data dictionary available in the metadata folder as is the technical document HRS Restricted Data File Companion: LTCFocus.

The data file for the Residents was not used for this project. The data from the CMS Provider was used for residents.

NOTE: Crosswalks

The challenge with locating shelter facilities and emergency service providers in a county or independent city was the use of different variables to identify their location (latitude and longitude, address, ZIP Code, FIPS code, and county/city name). In cases where the data source only had a ZIP or FIPS code, a HUD crosswalk was used to link the two codes; in other cases, a crosswalk that linked non-independent cities and towns to counties was used; and in others, a crosswalk that linked FIP codes to counties and independent cities. The crosswalks used are in the resilience subfolder, the crosswalks are listed below.

The data file for the Town to County Crosswalk is va_town_county_crosswalk.csv.

The data file for the Town to County Crosswalk is va_zipcode_town_county_crosswalk.csv.

Description of programs/code

The R code and the code output are described below by the topics, synthetic population and household living budget. The file names for the R code and output as well as the hyperlink to the repository are provided in the table below.

County Assets		
Purpose	**Program / Output**	**Notes**
Wrangles all the HIFLD data using crosswalks to create a data file with the locations of all the available shelter and emergency facilities.	County_Assets_Infrastructure.R / va_county_shelter_and_emergency_facility_resilience_index.csv	Creates the derived variable facilities_shelters_per_10000. The number of emergency and shelter facilities per 10,000 residents.
Contains the health professional shortage categories by county along with the geographies.	County_Assets_Workers.R / va_county_hpsa.csv	

A combined file with all the 55 community variables is: va_county_city_data_table.csv.
The data dictionary for the file is: va_county_city_data_dictionary.csv.

Skilled Nursing Facility Deficiency Index		
Purpose	**Program / Output**	**Notes**
Creates the deficiency indices for emergency preparedness and fire safety and combines them into a single index.	SNF_Deficiency_Index.R / va_snf_deficiency_indices.csv	Creates the derived variables k_def_index (fire safety), e_def_index (emergency preparedness), and ke_def_index (combined)

A combined file with all the derived SNF variables including latitude/longitude is: va_snf_data_table.csv.
The data dictionary for the file is: va_snf_data_dictionary.csv.

Resilience		
Purpose	**Program / Output**	**Notes**
Brings in the Census data to calculate a community resilience index for the tract level.	VA_Population_Resilience_Index_Census_Tract.R / va_census_tract_population_resilience.csv	Creates the derived variable pop_resilience (resilience) at the tract level.
Brings in the Census data to calculate a community resilience index for the county level.	VA_Population_Resilience_Index_County.R / va_county_population_resilience.csv	Creates the derived variable pop_resilience (resilience) at the county level.
Probability Nursing Staff Can Get to Work		
Purpose	**Program / Output**	**Notes**
Conducts a Monte Carlo simulation using the proprietary data to estimate the average number of nursing staff in an extreme flood event.	VA_Probability_of_Getting_to_SNF.R / va_snf_estimated_average_daily_nursing_staff_during_extreme_flood_event.csv	Creates all the variables used to construct Exhibit 3.
These data files were generated by proprietary software used to calculate the risk of getting to the SNF in the event of an extreme flood event.	va_snf_to_si_location_mapping.csv, va_snf_flood_risks.csv	These data were used to calculate the probability of the nursing reaching in the facility.

List of figures/code

Exhibit 1 was provided by the Census Bureau and Exhibits 2 and 9 were constructed by the authors in PowerPoint. Exhibit 4 is not included since it was constructed using proprietary software. R code reproduces Figures 3, 5, 6, 7, B-1, and B-3. The file names for the R code and figures as well as the hyperlinks to the repository are provided in the table below. The code for figures that were part of the exploratory data analysis process but were not included in the Technical Report are not included in the table. The code is available on GitHub in the discovery folder.

Exhibit	Program / Output	Notes
3	VA_Probability_of_Getting_to_SNF.R / snf_estimate_daily_nursing_staff_during_climate_event.pdf	
5	EDAshape.R / eda_emergency_preparedness_deficiency_index.pdf, eda_fire_life_safe_code_deficiency_index.pdf	EDAshape.R is an R function written by the author used to create the figures. The two .pdf figures were brought into PowerPoint to create a single figure.
6	VA_Population_Resilience_Index_Census_Tract.R / census_tract_population_resilience_index_choropleth.pdf	
7	County_Assets_Workers.R, County_Assets_Infrastructure.R / county_health_professional_shortage_area_choropleth.pdf, county_shelter_facilities_and_emergency_service_provider_choropleth.pdf	Two .pdf figures were brought into PowerPoint to create a single figure.
B-1	Richmond_Population_Resilience_Index.R / census_tract_richmond_population_resilience_index_choropleth.pdf	
B-3	Richmond_Isochrones_Map_All_Facilities.R / richmond_isochrone_map.jpeg	Requires an access token to Map Box.

Instructions to Replicators

The final product are two data tables that contain the variables from numerous data sources previously described and derived variables,

1. va_county_city_data_table.csv / va_county_city_data_dictionary.cvs and
2. va_snf_data_table.csv / va_snf_data_dictionary.

The data dictionaries for each table provide an explanation of the derived variables and a link to the R code used to construct them. To replicate the derived variables you just need to run the R code. For the other variables, a link to the data source is provided.