ETL Process Documentation (Group 3)

Datasets

* Health Insurance Historical Tables <https://www.census.gov/data/tables/time-series/demo/health-insurance/historical-series/hic.html>
* Historical Income Tables <https://www.census.gov/topics/income-poverty/income/data/tables.html>
* Gini Index <https://www.shsu.edu/eco_mwf/Frank_Gini_2018.xls>
* Median household income, by state: Selected years, 1990 through 2017 <https://nces.ed.gov/programs/digest/d18/tables/dt18_102.30.asp>
* 2020 Census Apportionment Results <https://www.census.gov/data/tables/2020/dec/2020-apportionment-data.html>
* Small Area Health Insurance Estimates (SAHIE) Program <https://www.census.gov/programs-surveys/sahie.html>
* SAHIE File Layout Overview: 2008 to 2019 <https://www2.census.gov/programs-surveys/sahie/technical-documentation/file-layouts/sahie-file-layout-2008-2019.pdf>

**Vahram Khachikyan**

Extract

* Downloaded 3 datasets (Health Insurance Historical Tables, Historical Income Tables, Gini Index)
* Most of them were csv files, and Gini Index was an xlsx file. Extracted data using Excel’s Get & Transform Data Feature

Transform

* Historical Income Tables
  + Removed irrelevant columns
  + Grouped Rows by State
  + Added indexes for each state
  + Moved index to be column 1
  + Removed “Puerto Rico” row
  + Removed “District of Columbia”
* Health Insurance Historical Tables
  + Removed irrelevant columns
  + Added indexes for each state
  + Moved index to be column 1
* Gini Index
  + Removed irrelevant columns
  + Created a pivot table to group by states
  + Indexed based on state
  + Removed Top rows
  + Promoted Headers
  + Added indexes for each state
  + Moved index to be column 1
* Merged these 3 tables using index

Load

* Clicked “Close&Load”
* Created Correlation Chart to see the relationship between the percent of uninsured people and income and income inequality.

**Isabella Caro**

1. Download from the US Census Bureau Website datasets on insurance rates by state & insurance type (American Community Survey), Small Area Health Insurance Estimates 2019 (SAHIE), median household income over time, state population, proportion of state population by race
2. For each dataset, remove rows showing aggregate data for the U.S. and data for Puerto Rico and District of Columbia
3. SAHIE

* Remove blank, null, and error values
* Filter “racecat” values to include only 1, 2, and 3
* Replace 1, 2, 3 with white (non-Hispanic), black (non-Hispanic), and Hispanic (any race), respectively
* Create a pivot table and create columns for each racecat with the average NUI as the values and state\_name as the rows

1. Insurance by Race Estimates

* Join on state the proportion of state population by race table with the cleaned and transformed SAHIE table
* Join on state the state population table with the previous table
* Add three custom columns by multiplying state population by the proportion of each race (White, Black, Hispanic) to get the population in each state by race
* Add three custom columns by dividing the NUI for each race by the population by race to get the percentage of the population by each race that is uninsured

1. Insurance and Income Over Time

* Remove blank, null, and error values from the insurance type dataset
* Unmerge all cells and remove first two rows
* Filter “Coverage” column to show only rows labeled “Uninsured”
* Remove all columns except for state those showing number of uninsured people in thousands for each year
* Remove blank, null, and error values from the median household income over time dataset
* Unmerge all cells and remove first four rows
* Remove all columns except for state and those showing the Median income for each year
* Join on state the insurance type table and the median household income table
* Using the quartile function, find the range of states included in each quartile of median income for 2018
* Create a pivot table with the states as rows, years as columns (filtering to include only 2010-2018), and average number of uninsured in thousands and median income as values
* Create a slicer for income quartiles

**Nick Verbus**

Uninsured People over Time

1. The HIC-4 dataset was used
2. The “Coverage” column was filtered down to uninsured
3. Then in the “United States” row the estimate columns with their respective year (in a new corresponding column) were copied and pasted to a new sheet for each year
4. Then a graph was made from the two condensed columns

Can we predict if someone is insure based on income and demographic factors?

1. The sahie\_2019 dataset was used for this
2. First selected are the rows with empty values under “county\_name”—this gives us state values rather than county values
3. Next all of the state names have to be renamed as they are represented as “Alabama          “ with many spaces instead of just “Alabama”. So, to join on state name, these spaces need to be eliminated
4. Following this join the Median\_Houshold\_Income\_by\_State to the sahie\_2019 table using state name as the join key. The “Income” is the only thing that we need.
5. Now column dropping is done. These columns are dropped: version, year, countyfips, Median\_Household\_Income\_by\_State.Income, geocat, nipr\_moe, nui\_moe, nic\_moe, pctui\_moe, pctic\_moe, pctelig\_moe, pctliic\_moe, state\_name, county\_name, statefips, PCTUI, PCTIC, PCTELIG, PCTLIIC, NIC
6. Note: The Median\_Household\_Income\_by\_State.Income, NIC can be kept depending on what wants to be measured.
7. After, we select the rows in which the values of these columns – racecat, agecat, sexcat, iprcat—correspond with the values that we want them to have, these columns can be dropped as they serve no more purpose.
8. Then we can run our regression.