Phase 2 - Info 2950 Final Project

Aryana Thapliyal, Ishneet Sachar, Avni Kulshrestha

Data description. This should be inspired by the format presented in Gebru et al, 2018. Answer the following questions:

1. **What are the observations (rows) and the attributes (columns)?**
    - **a.** For income: The observations are all of the counties listed under each state. The states are listed in alphabetical order, with all the counties underneath the respective state. The attributes are Per Capita Personal Income (showing the dollar amounts over the course of 3 years), and including columns showing the percent changes in income. It also shows us the rank of the state in terms of per capita income.
    - b. For AQI: The observations are the states with all the counties listed next to them. Like the income dataset, the states are listed in alphabetical order, with all the counties listed as well in order. The attributes are the state code, air quality index (AQI), category the AQI falls under, the defining parameter, defining site, and number of sites.
    - c. For population estimates: The observations are the states with all the counties listed next to them. It includes population estimates for each county in all states. The attributes are the last census data for population size, and the population estimates for each county since 2010 to 2018. It also includes the estimated number of births and deaths, and birth and death rate for each year.
    - d. For our dataset: We have combined the observations and attributes from the above datasets for our project (focusing on 2018 data). This combined dataset includes observations: the state name and all the counties within that state (from the AQI dataset we averaged the county values so each county is only listed once), population estimates for each county by including the Census for each year for each county (from the population dataset). The attributes we are pulling from the income dataset are the Per Capita Personal Income (showing the dollar amounts over the course of 3 years), and including columns showing the percent changes in income. From the AQI dataset, we are pulling the average AQI index for each county (which was taken from average), and the condition that AQI falls under as it may be helpful in the future for our understanding.
2. **Why was this dataset created?**
    - a. Why each dataset was created: Each dataset was created to provide detailed county data. Oftentimes, county data is overlooked and only state data is analyzed. Although, all counties are not similar within a state and the differences should be analyzed. The income dataset gives us the income per capita for all the counties, the AQI dataset allows us to see differences in the air quality levels between counties, the population dataset is good for us to see how many people are estimated to be in each county.

b. Our dataset: Our dataset was created for us to see some type of relationship between air quality, a population size, and income across different counties in the U.S. We want to see how income and population size may affect the AQI index. We want to see if there is any significant outcome between these individual datasets, for example, we will compare different counties that have a higher population size (cities, more urban areas) with those in more rural parts. There are of course more relationships that we will look for, as we study our data more.

3. **Who funded the creation of the dataset?**
    a. The income dataset collection was funded by the U.S. Bureau of Economic Analysis, the air quality dataset collection was funded by the U.S. Environmental Protection Agency, and the population dataset collection was funded by the U.S. Department of Agriculture.
4. **What processes might have influenced what data was observed and recorded and what was not?**
    a. Income data: Used reports for salaries and wages etc from places of work → data is reported by industry in the state and the county where the establishment is located. Also use tax filing data sources and place of production for farms. For rental income it is the place of residence
        i. To calculate the per capita income, BEA used Census Bureau's annual midyear population estimates → using area's population might affect the data recorded because it doesn't take into account institutional populations (such as colleges or prisons) and doesn't account for unusual conditions or money remittances. Population changes can also significantly impact income
        ii. Population is also measured mid-year and income flows through the year so changes in population around mid-year can change year on year estimates of per capita income (https://www.bea.gov/system/files/methodologies/LAPI2018.pdf)
    b. Population Estimates: https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2019/natstcopr-methv2.pdf
        i. The USDA ERS procured population estimates from the U.S Census Bureau → each time series is recorded using the latest administrative data (geography boundaries, methodology etc). So across different time series, data might not be consistent across different geography and characteristics → this affects which data was observed and recorded
        ii. The U.S Census is a questionnaire. The way the questionnaire is structured, along with the distribution of the surveys affect the response rate and the quality of responses received. Distributed electronically and via mail or phone calls with good marketing efforts beforehand and followed up with those who either didn't receive the questionnaire or didn't fill it out

(https://www.census.gov/history/www/through_the_decades/overview/2010_overview_1.html)

iii. The way the population estimates are calculated (using an equation that includes measures for other variables) can affect the accuracy of the population estimate (https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2019/natstcopr-methv2.pdf)

iv. For county data, they include domestic migration. State estimates are just the sum of county estimates (https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2019/natstcopr-methv2.pdf)

c. AQI Data: Collects major pollutant information using monitors across locations. Other pollutants might be excluded by the monitors, which would influence the AQI data (https://www3.epa.gov/airnow/aqi_brochure_02_14.pdf)

i. Not real time data

ii. Not every site measures the same pollutants

iii. Measured continuously and intermittently (this has longer durations of measurement and time gaps between each measurement) (https://aqs.epa.gov/aqsweb/documents/about_aqs_data.html#_introduction)

5. **What preprocessing was done, and how did the data come to be in the form that you are using?**

a. Income data: Measured income and divided it by the area's population (mid-year) https://www.bea.gov/system/files/methodologies/LAPI2018.pdf

b. AQI Data: EPA has monitors in different locations that measure major pollutant levels in those locations. The raw measurements are processed into an AQI score for each pollutant measured using EPA-set formulas. The highest AQI detected is the overall AQI for that day (https://www3.epa.gov/airnow/aqi_brochure_02_14.pdf)

c. Population Estimates:

i. The USDA ERS procured population estimates from the U.S Census Bureau → each time series is recorded using the latest administrative data (geography boundaries, methodology etc). So across different time series, data might not be consistent across different geography and characteristics → this might affect the preprocessing done

1. The U.S Census is conducted once every 10 years. The Bureau then makes uses an equation consisting of the population base, births, deaths and migrations to produce a population estimate for each year

ii. The ERS dataset also included some influence and continuum codes: these are classification schemes that assign codes to counties based on characteristics such as metro and non-metro categories etc → this pre-processing has already happened and the codes are included in the

dataset (https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/)

d. Our dataset: We first processed each of the raw datasets individually to account for NaN values and unnamed column headings
    i. Income data: the headings were re-named according to the raw data excel file, and NaN values that resulted from the data recorders leaving a row blank after each state, were removed. Inc_2018 is a pandas dataframe containing personal income per capita in 2018, the data we need for our dataset. Each row contains state, county, and income.
    ii. Population data: The headings were adjusted to reflect the raw data-set's, and we removed any NaN rows (the only NaN row was the last row of the data-frame that seemed to have been included into the data-frame from when the CSV file was read in). Additionally, we added a dataframe that had the state with their abbreviations and were able to merge this with the population data so that when merging the final dataframes we would be able to merge on the state names. The counties in this dataset also contained extra words after the county name such as "County, Borough. And Census City", so we used regex to get rid of these extra words so we could merge more easily. We then extracted the data we needed, which was population estimates of 2018, or pop_2018. Each row contains the state, abbreviation, county, and population.
    iii. For AQI, we saw that under each state name, the same county was listed multiple times. We only wanted a county's data to be counted once, so we averaged the numbers and received one datapoint for each column for a county. We did this by creating a pivot table so that each county's information would be in one row and then averaging each row. We extracted the average AQI index as well as the state and county.
    iv. For our dataset: After cleaning each dataset to show the attributes we wanted, we merged the three to make our own dataset so it will be easier to analyze. We merged them using inner joins so that each county would have the three data points we need (population, AQI, and income).

6. **If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?**
    a. Income data: yes reported by people or places/institutions and they knew that their data was going to be used for these purposes
    b. AQI Data:
        i. No people were involved in the collection of the data apart from the researchers
    c. Population Estimates:
        i. For Census data: yes they were aware since it was self-reported data and participants are aware that their data will be used to portray national statistics → users know their responses will be confidential and safe

ii. The urban-rural continuum codes and influence codes are created internally so there is no direct data collection from participants since the researchers themselves break down counties into codes

7. **Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted in a Cornell Google Drive or Cornell Box)**

Income - https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas

Data on AQI - https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI

Population estimates - https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

The raw data files can be found on our public github repository.
https://github.com/isachar98/info2950-final

The raw data can also be found in this folder on Google Drive
https://drive.google.com/drive/folders/1AqKeMpjKk1VXLdC44LPg4dh5vBdflewG?usp=sharing

Additionally, our combined dataframe can be found here:
https://drive.google.com/open?id=1FuzHaidEkkRPY6EjX1VMrIkenzXAQ7tB