Data Curation Project

Samantha Berg
Hernan Razo
Christopher Anthony

Dataset Introduction

Datasets used:

- **Health insurance Coverage of the Total Population**
    - Provided by KFF, a third party non-profit organization focused on health policy analysis and health journalism
    - Gives number of people in the US that are insured or uninsured
    - Breaks down coverages by specific health plan and other health insurance sources

- **Coverage Numbers and Rates by Type of Health Insurance**
    - Provided by the United States Census Bureau
    - Gives number of people in the US that are insured or uninsured
    - Breaks down coverages by specific health plan
    - Also breaks down data by state

Domain

- Both datasets fall into the health insurance domain
- They both try to quantify the source of health insurance for the general US population

Methods of Acquirement

- The data from KFF is openly available on their website
- The data from the US Census Bureau is also open to the public on their website

Data Characteristics

- **KFF dataset**
  - composed of 58 rows and 8 columns
  - Has 8 attributes: Location (state) and various categories for insurance providers
    - Employer, medicaid, medicare, etc.
- **US Census Bureau dataset**
  - Composed of 24 rows and 10 columns
  - Has 10 attributes: various categories for insurance providers
    - Employer, medicaid, medicare, etc.

Data Quality Problems/Challenges

- The dataset from the US Census Bureau came in pdf format. This means that we had to manually input data values into csv format
- We also had to transform each dataset to only include the common insurance plans between both datasets:
    - Employer, Medicaid, Medicare, Military, Uninsured
- We filtered each dataset to only include the 50 US states (removed territories)
- It was hard to tell how each organization counted if someone was insured or not. Discrepancies that came up:
    - Being insured for only portions of the year counted as uninsured in some instances
    - What was considered to be uninsured
    - Discrepancies between ASEC vs. ACS reportings

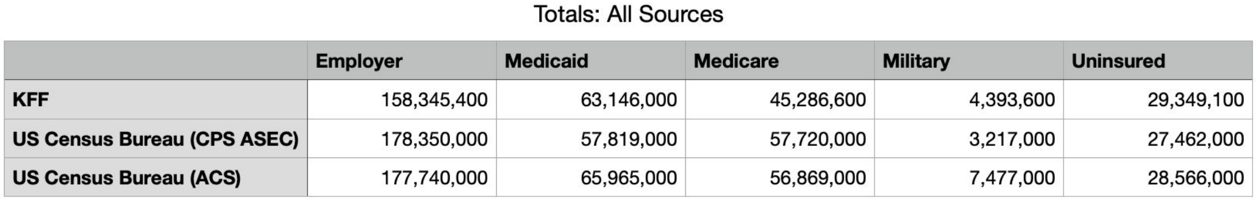
Solutions

- We used the Pandas python library to extract all necessary numbers
- We also used python to compute the sum total number of people that are insured by specific plans

Findings

- All three data sources had different numbers for all criterias
- The US Census Bureau reports that:
  - More people are insured through their employer and Medicare
- The KFF reports that:
  - More people are insured through Medicaid and the military
  - More uninsured people

Findings Continued



Totals: All Sources

	Employer	Medicaid	Medicare	Military	Uninsured
KFF	158,345,400	63,146,000	45,286,600	4,393,600	29,349,100
US Census Bureau (CPS ASEC)	178,350,000	57,819,000	57,720,000	3,217,000	27,462,000
US Census Bureau (ACS)	177,740,000	65,965,000	56,869,000	7,477,000	28,566,000

Why are there Differences in the Reported Numbers?

- The data from the US Census Bureau reported from two sources: CPS ASEC and ACS
- The **Current Population Survey (CPS)** is a monthly survey that can be aggregated to an annual report. It is the general survey used to extract various statistics from the population.
- The CPS sample size is about 60,000 occupied households

- The **Annual Social and Economic Supplement of the Current Population Survey (CPS ASEC)** is an extension of CPS that samples various census data at the state level. They only include data from 1,300 counties (out of 3,100 total counties) and 98,000 households.
- Also, CPS ASEC estimates each county number over a three-year weighted average centered around the current target year.

Differences in Reported Numbers Continued

- The methods of reporting from the KFF was less transparent
- They claim to do custom analysis of the American Community Survey (ACS) provided by the US Census Bureau
    - The ACS samples 1% of the total US population

- **Dual eligibles** - Individuals who report having both Medicare and Medicaid.
- Dual eligibles must be placed in EITHER the Medicare or Medicaid category, but NOT both
- It was not clear how the KFF decides which category to place these individuals. No valid information could be found on this.