**Disparities in Health Insurance Coverage**

Data Wrangling: Spring 2024

University of Utah, Department of Bioinformatics

Project repository: <https://github.com/r-buch/social-covid>

**Group Members:**

|  |  |  |  |
| --- | --- | --- | --- |
| Chase Maughan  [U0962361@umail.utah.edu](mailto:U0962361@umail.utah.edu)  u0962361 | Kalpana Simhadri  [u1473696@umail.utah.edu](mailto:u1473696@umail.utah.edu)  u1473696 | Rachel Buchleiter  [U1414409@umail.utah.edu](mailto:U1414409@umail.utah.edu)  u1414409 | Xincen Xi  [Xincen.xi@utah.edu](mailto:Xincen.xi@utah.edu)  u1475541 |

**Research questions**: Create a dataset with Race and Health insurance as variables and try to examine the following research questions (or similar):

1. How do health insurance coverage rates vary across different racial and ethnic groups within Los Angeles County?
2. Are there any significant correlations between race/ethnicity and lack of health

insurance coverage in specific census tracts of Los Angeles County?

**Background and Motivation**

Social Determinants of Health (SDOH) are recognized as important factors influencing health outcomes of the population. SDOH is primarily measured in five domains as defined by the CDC. These are economic stability, education access and quality, health care access and quality, neighborhood and built environment, social and community context (Healthy People 2030, n.d.) It can be challenging to measure these insecurities directly. While there are assessments that can be used when patients have a medical encounter, this is not standardized nationwide and are inconsistently used. Since this priority area has high impact but can be elusive to identify, more effort is needed to understand related data. One proxy correlated with factors of with these insecurities is census tract data. It is a priority for national health organizations and funding. The flexibility and numerous research questions that could leverage this data set, or similar social determinants of health sources, made it an appealing learning opportunity for the group. We decided to narrow our focus by considering the addition of health insurance data to explore possible differences within a census tract. Lack of health insurance coverage is known to be a barrier to accessing quality healthcare and subsequently positive health outcomes. Given this linkage, we aim to explore what characteristics of a population are more highly correlated with health insurance coverage.

**Project Objectives**

The purpose of our report is to investigate health insurance coverage rates across different racial and ethnic groups within Los Angeles County. We hypothesize that there are significant correlations between race/ethnicity and lack of health insurance coverage in specific census tracts of Los Angeles County. We will identify these correlations by reviewing publicly available data concerning SDOH. The primary source of these data come from the United States Census Bureau.

Overarching Research Questions:

1. How do health insurance coverage rates vary across different racial and ethnic groups within Los Angeles County?
2. Are there any significant correlations between race/ethnicity and lack of health insurance coverage in specific census tracts of Los Angeles County?

Benefits to using this data:

* Identify disparities in health coverage to guide community health interventions
* Provide insight into barriers to health coverage among racial and ethnic groups in a community
* Support health policy legislation

Learning Objectives:

* Identify publicly available data can be used in analyses focused on social determinants of health
* Clean publicly available data associated with domains of social determinants of health
* Identify and merge relevant open data into a data set suitable for analysis identifying social determinants of health. Use the FACETS project as an example.

**Data**

The primary data set used in this report can be accessed on the US Census Bureau website, specifically on the Census Bureau Data page. The refining tools given on the website can be used to focus the data specifically on Los Angeles County, California. With the filtered data pertaining to age, race/ethnicity, and health insurance status. In the project the next steps involve downloading these datasets and compiling them to be able to easily visualize and compare the data.

Other Possible data set:

<https://www.cdc.gov/nchs/nhis/2022nhis.htm> (national data set, could be interesting to compare Los Angeles to National Population)

**Data Processing**

Some key points that can be considered for data processing are:

* **Data cleanup**: identifying and handling missing data, standardizing formats, checking the data consistency. Make sure all data of interest to the report is not lost, but is properly identified
* **Data integration**: merging the datasets, this will be particularly important if we decide to compare multiple census years. ensuring proper alignment of different geographic regions. Also ensuring that the variables from different sources are harmonized to be directly comparable, involving adjustments to variable scales, definitions, or methodologies.
* **Develop scripts**: The coding for this project will be completed using Python, allowing us to automate as much data cleaning and transformation process as possible.
* **Version control**: Use version control for the code and data processing steps to track changes and ensure reproducibility
* **Documentation**: thoroughly document all data cleaning and processing steps, including decisions made and their rationale

**Design**

We plan to use visual representations to display data that best answers our project objectives. Here are the methods we have chosen to represent the results for each of our objective questions:

* Bar charts to show the number of insured and uninsured people across different age groups in Los Angeles County
* Bar charts to show the number of insured and uninsured people among different races
* Pie charts to show the proportion of insured and uninsured people within different categories
* If we include the Margin of Error, we can also add error bars on top of the bar charts to display the range of uncertainty for each estimated value

**Must-Have Features**

Unique Identifier with the following harmonized columns:

* Race (standard NIH)
* Ethnicity (standard NIH)
* Age Label (Grouping)
  + Total, Under 19 years
  + 19 to 64 years
  + 65 years and over
* Insurance Label (Grouping)
  + With health insurance coverage
  + Partial insurance coverage
  + No health insurance coverage
* Los Angeles County, California
* Census Track ID

These features will allow us to complete the project objectives which can help us to determine the distribution of health insurance coverage across different age groups, racial and ethnic groups in Los Angeles County. With this information we can also glean insight into if race, ethnicity and lifespan (age) correlate with health insurance coverage.

**Optional Features**

* Margin of Error
  + For each census track
  + For each race and ethnicity group
  + For each age group
* Include additional time periods (prior to 2022) to explore if there are trends over time in insurance coverage within the population of interest

Margin of Error for each grouping will allow us to assess the reliability of the data and understand the range of the error. Additionally, by including data from different census tract numbers, we can compare whether there are differences between various sub groups.

**Project Schedule**

|  |  |  |
| --- | --- | --- |
| **Schedule** | **Deliverable** | **Deliverable date** |
| Week 1 | Project Proposal | 2/12 |
| Week 2 | Meet with instructor | 2/19 |
| Week 3 | Project Update in Github | 3/1 |
| *Spring Break* |  | *3/3-3/10* |
| Week 4 | Intermediate work presentation preparation |  |
| **Week 5** | **Intermediate Work Presentation** | **3/18** |
| Week 6 | Updates in Github | 3/25 |
| Week 7 | Meet with instructor | 4/1-4/8 |
| **Week 8** | **Intermediate work presentation** | **4/8** |
| Week 9 | QA, presentation preparation |  |
| **Week 10** | **Final Project presentation** | **4/29** |
| Week 11 | Final Project Submission | 5/6 |
|  |  |  |

**Work allocation among the team**

|  |  |
| --- | --- |
| Chase Maughan | Data Processing Lead |
| Kalpana Simhadri | Data, Github lead, QA, Project Manager |
| Rachel Buchleiter | Documentation/Presentation Lead (ensuring consistency, references),QA |
| Xincen Xi | Features, QA |

**References**

FACETS: using open data to measure community social determinants of health. doi: 10.1093/jamia/ocx117 Healthy People 2030, n.d. “Social Determinants of Health.”

<https://health.gov/healthypeople/priority-areas/social-determinants-health>

<https://www.kff.org/racial-equity-and-health-policy/issue-brief/health-coverage-by-race-and-ethnicity/>

**Data discovery notes**

* The datasets required for answering the above-mentioned questions are available on the US Census Bureau website, see [Census Bureau Data](https://data.census.gov/) page
* This data set was filtered California, Los Angeles county, All census tracts withing LA county, Race and ethnicity, 2022, health (health insurance) 5 year estimate data. [Census Bureau Search](https://data.census.gov/all?t=Health%20Insurance:Race%20and%20Ethnicity&g=040XX00US06_050XX00US06037,06037$1400000&y=2022&d=ACS%205-Year%20Estimates%20Detailed%20Tables)
* Image for reference:

A screenshot of a computer

Description automatically generated

* All the datasets are available separately for each race/ethnicity (as shows in the screenshot above)
* We should download them and compile (with census track number as unique identifier)
* Then wrangle/clean the data as per the requirement.