Data Science Fall 2020
Assignment 4
17 December, 2020

Angelina Martineau
Andrew Updegrove
Vaishnavi Neema

1.

a)

The goal of this project is to determine which areas in the United States are not prepared for heart disease. More specifically, the goal is to discover which counties in the United States have relatively high rates of heart disease and low rates of insurance coverage and hospital bed availability. Heart Disease, also known as cardiovascular disease, is the leading cause of death in the United States, therefore areas with greater counts of heart disease need to be prepared with the proper resources to combat the disease. This project aims to support the third goal of sustainable development: "Ensure healthy lives and promote well-being for all at all ages," (un.org, 2020). Analysing the areas where heart disease is highly prevalent and the resources to treat it are low will help determine which areas need more funding, which will in turn aid in the treatment of heart disease and promote healthier living for those suffering from it.

Five datasets were used for this project. They were found by searching Google for datasets containing information relevant to medical care and heart disease. The first dataset focuses on heart disease prevalence broken down by county, and was retrieved from https://chronicdata.cdc.gov. The second dataset includes information about hospital bed availability in the United States, and was retrieved from https://opendata.dc.gov. Third, a dataset containing information about health insurance coverage, broken down by insurance type, was retrieved from https://covid19.census.gov. These three datasets are all broken down by county, so the fourth and fifth datasets focus on county data. The fourth dataset was taken from https://www2.census.gov, and contains population data for each county in the United States. Finally, a fifth dataset was needed to help organize the county data. This dataset was retrieved from https://www.ssa.gov, and contains the standard state abbreviation codes. The datasets were all downloaded from their respective websites, and were added to the 'original_datasets' folder in the GitHub repository for this project.

b)

Each dataset was able to be downloaded in CSV format, and the CSV files were then stored in the 'original_datasets' folder in the GitHub repository. The metadata for each of the files were written into the websites they came from (i.e, not an extractable file), so to properly include all of this data, a datasets.txt file was created containing all links to the download sites for the datasets. One dataset, the county population dataset, included a PDF file with metadata information including attribute descriptions. This PDF file was downloaded and added to the GitHub repository to preserve the metadata information that came with the dataset.

The methods of discovery used focused on using the Google search engine to find relevant datasets. When a dataset was found and it was determined that it would aid in the development of this project, it was downloaded from its given website. Each website has

an easy-to-use, intuitive interface, so downloading the different data files was not difficult. This made it easy to obtain the appropriate data in a common, usable format.

2.

a)

Two questions and hypotheses were developed for this project. The questions being investigated are listed below:

1. Where in the United States (divided by county) is health insurance lower on average with heart disease higher on average?
2. Where in the United States (divided by county) is hospital bed availability lower on average with heart disease higher on average?

For the first question, the team hypothesized that counties closer to major cities would have lower levels of healthcare coverage and higher levels of heart disease. For the second question, the team hypothesized that counties in more rural areas would have less hospital bed availability, and that these areas would potentially have higher levels of heart disease.

A data analysis plan was developed to perform a formal investigation on the datasets, focusing on the two questions presented above. The following image describes the data analysis process used in this project:

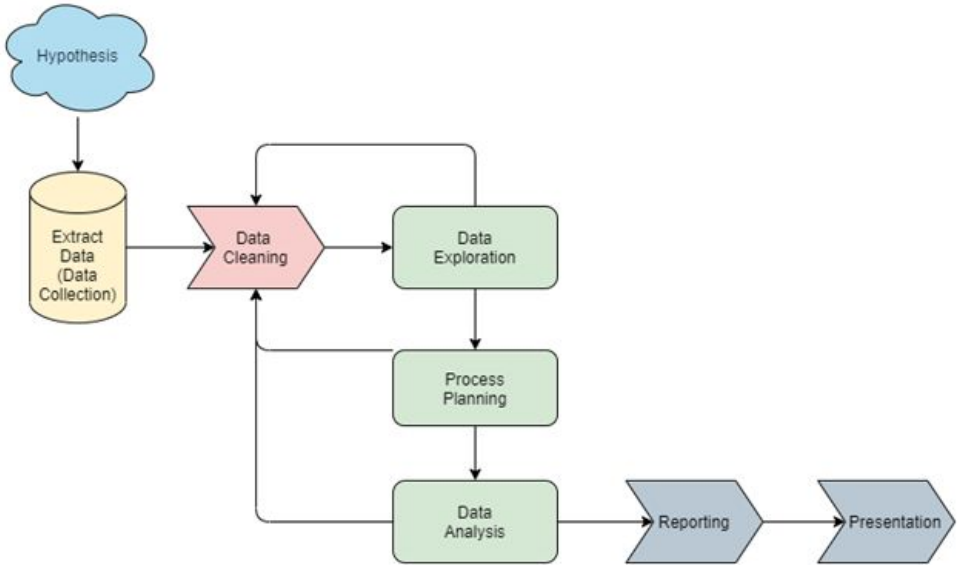

Image 1: Project Workflow for the 'Are We Prepared For Heart Disease?' Project.

The beginning of the project focused on developing a topic to research. Once the topic was decided upon, questions and hypotheses were developed. Using the hypotheses, datasets were found on the web and were downloaded and saved to a project repository. Next, data cleaning was performed on each of the retrieved data files. Data cleaning was a cyclical process. As the data was cleaned, it was explored, and planning for the data analysis phase was done. Once the data was properly cleaned and data analysis could be

done with it, the project moved into the data analysis phase. Analysis for question one was done using the heart disease dataset, the insurance dataset, and the population dataset, and analysis for question two was done using the heart disease dataset, the hospital bed dataset, and the population dataset. Both analyses made use of the state code dataset for organizational purposes. Finally, after the analysis had been done, reporting of the findings could be done. This was done through a report PDF and a poster, which included visualizations of the project's findings.

b)

The two main tools used for this project were GitHub and RStudio. GitHub was used as a means of managing different versions of the project and as a platform where all group members could access the project. In addition to group members, anyone in the public can access this project on GitHub, they just won't be allowed to edit it. RStudio was chosen as it is a well-established tool for data analysis, and all group members are proficient in the R programming language.

One R script was written for analysis. This script is 'question_2_analysis.R', and is stored in the team GitHub repository. The script cleans each of the datasets, and uses multi-linear regression to determine any patterns present relating to heart disease and health care/coverage in the counties. The data produced from this script was saved to a folder titled 'analysis_dataset' in the GitHub repository. This folder also includes the metadata that goes along with the newly created data, in the form of a JSON document.

c)

The data analysis process used for this project can be replicated, meaning the results can be validated by other data scientists. First, the team made sure to include the design analysis process, so that others will be able to follow it. Second, the original datasets are provided on the GitHub repository, as well as the links to the original download sites, which allow the users to download the raw data on their own and view the metadata embedded in the website. Third, the team made sure to document the R script and readme.md file to clearly explain how the project and analysis were performed. Finally, visualizations were made in Tableau to explain the results.

In order for others to replicate the work done in this project, they will need to download the datasets, either from the GitHub repository or from the original links to the data. Then they will need to develop their own R script to perform linear regression on the data, and then they can compare their results to the results from this project, either through their own visualization techniques or the ones used in this project.

3.

a)

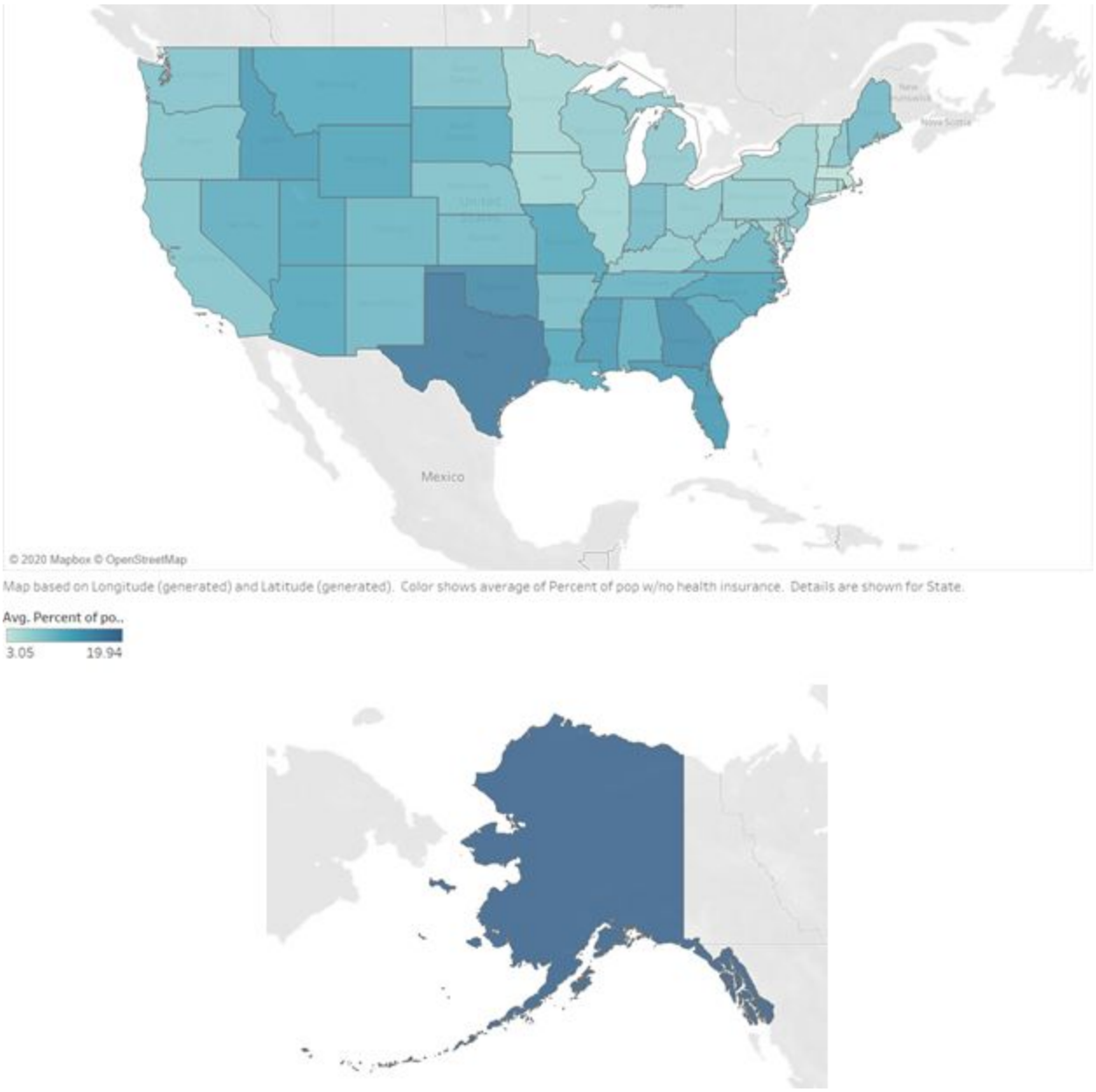

Image 2: Average Population for each state with no health insurance

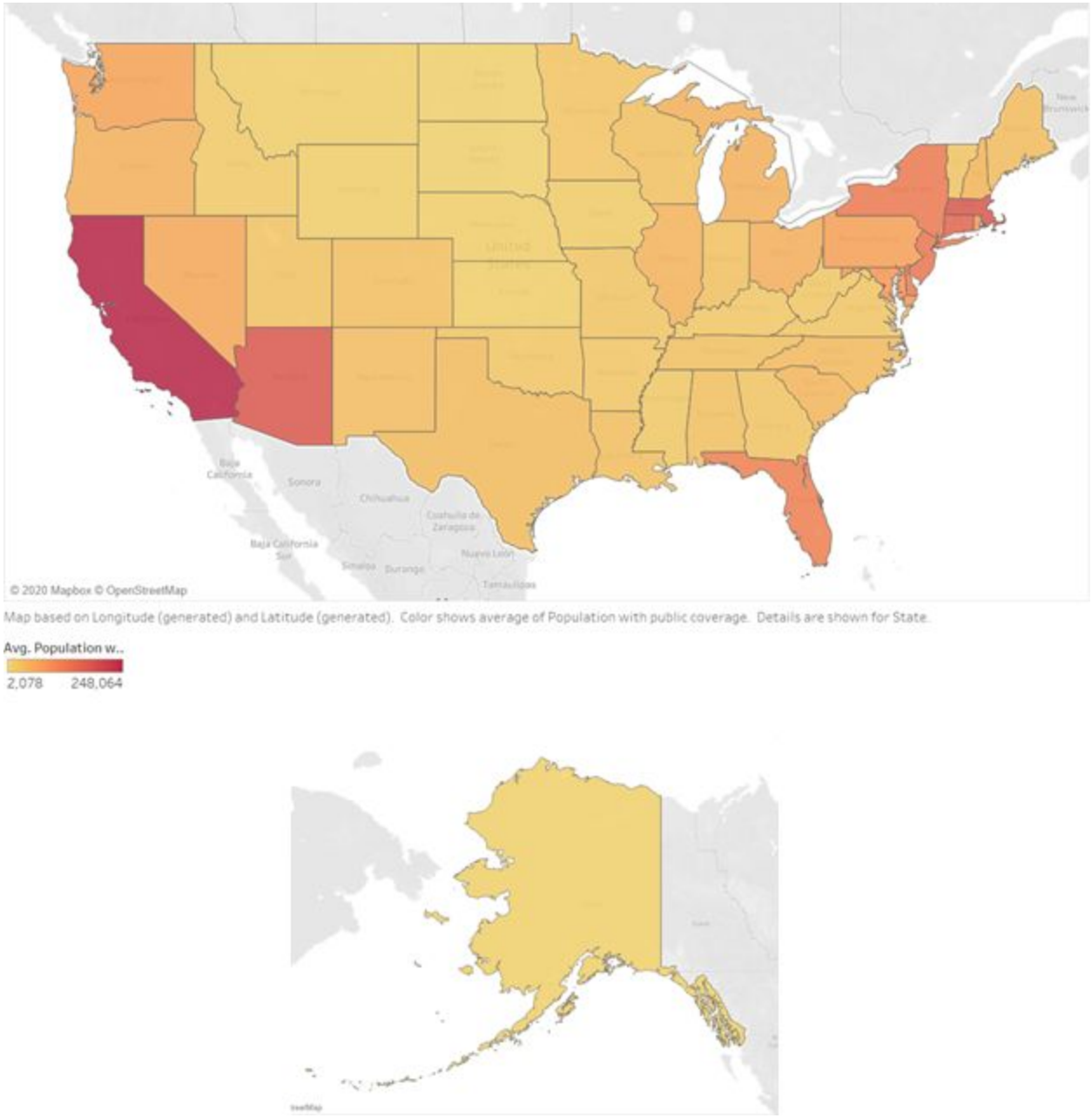

Image 3: Average Population with public coverage

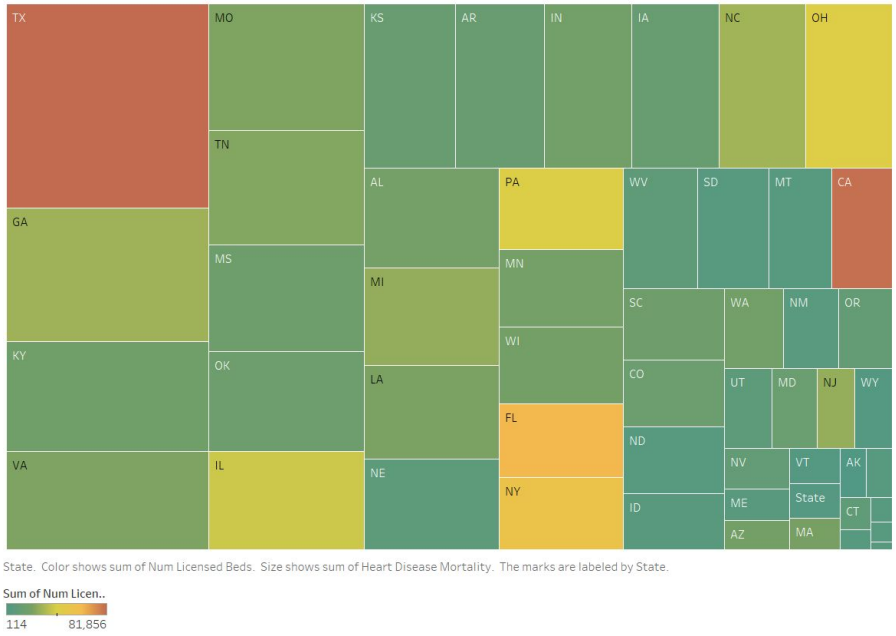

Image 4: Comparison between number of licensed beds and heart disease mortality rate per state. Here, the size of the block represents the heart disease mortality rate and the color represents the number of licensed beds.

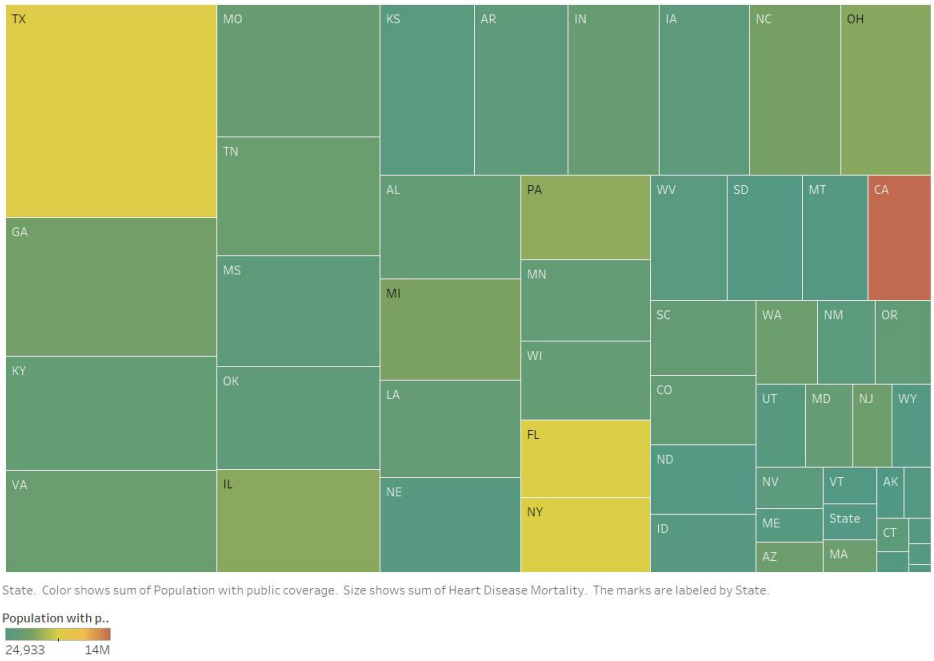

Image 5: Comparison between population with public health coverage and heart disease mortality rate per state. Here, the size of the block represents the heart disease mortality rate and the color represents the number of licensed beds.

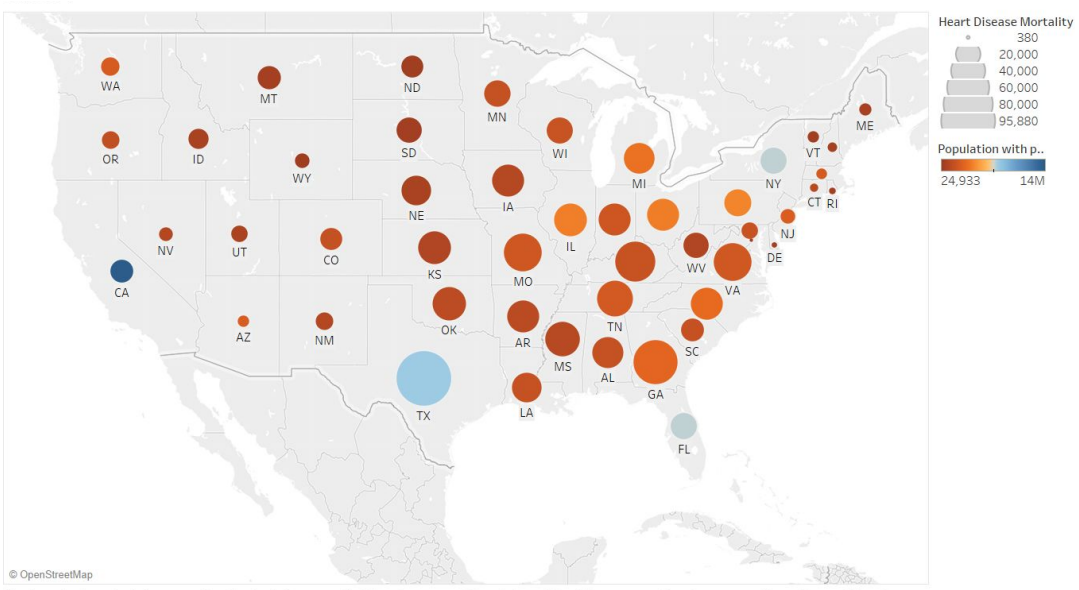

Image 6: Comparison between number of licensed beds and heart disease mortality rate per state. Here, the size of the block represents the heart disease mortality rate and the color represents the number of licensed beds.

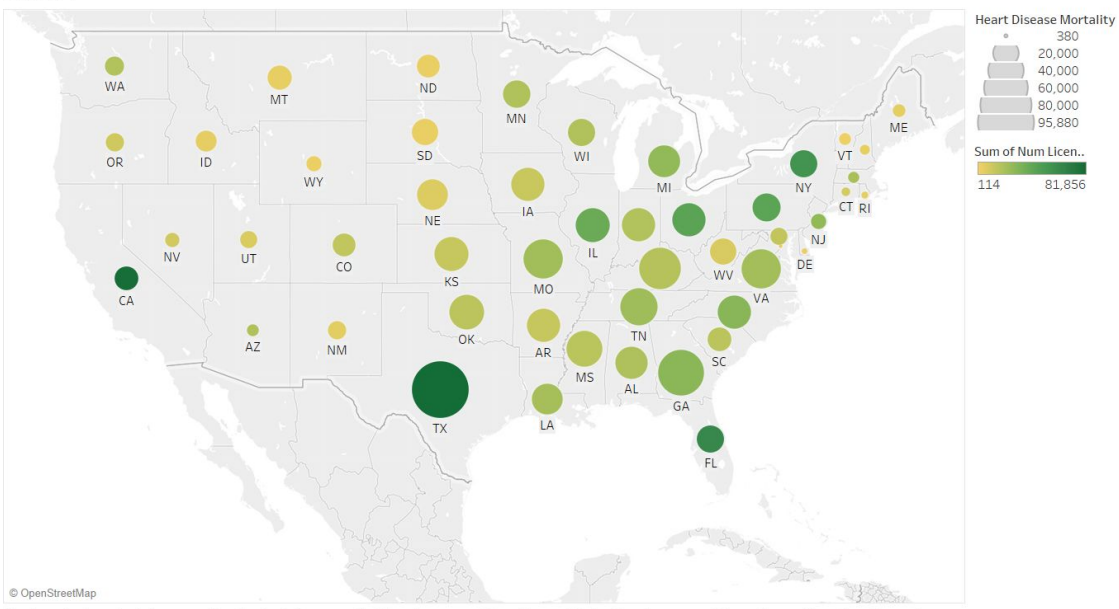

Image 7: Comparison between number of licensed beds and heart disease mortality rate per state. Here, the size of the block represents the heart disease mortality rate and the color represents the number of licensed beds.

Before starting to analyse the data, the data was first cleaned and normalised to reduce redundancies. This resulted in a better and filtered dataset. Also, since the data was at a county level, most of the fields had to be aggregated, averaged or divided by the population for that state in order to visualize data on a state level.

The initial data analysis was done with the population fields being plotted against states and counties to see the initial distribution of population. This helped in determining the states with the highest populations and looking closer into the data for those states. For example, Image 2 and Image 3 visualise the populations with and without health insurance for each state. Ideally the relationship between the ratio of population with health insurance and heart disease mortality rates should be linear.

Building on these analyses and the hypothesis described in 2(a), the next step was to see if these hypotheses actually hold true with the actual numbers in the data. Image 4 and 5 visualize the relationships defined in 2 (a). The size of the block represents the heart disease mortality rate and the color represents the number of licensed beds and population with health insurance coverage for Image 4 and 5 respectively.

Similarly, Image 5 and 6 visualize the same relationships on a USA map in order to give a more intuitive idea to the audience to see which states are actually doing poorly with their health insurance and/or healthcare.

b)

We used multiple platforms to facilitate the sharing and coordination of work within our team. The major platforms were Google Drive, Github and Slack. We used Slack for all communications within the team. Google Drive was used to work on the poster together and make suggestions, editing and viewing the progress of our work. The folder was shared privately with the team members and no one else was authorised to access the drive folder.

We stored both the poster and the associated plots, datasets and metadata in our course's GitHub repository. The plots have a detailed description at the bottom and the names of the flies describe the plots as well. All the images have legends that explain the colors and sizes used in the plots.

GitHub repo ensures that the data, visualizations, metadata and analysis are archived for future use for anyone to replicate the work that we did during the course of this project. The readme file in the repository explains what the dataset comprises and how to go about it. The repo url is https://github.com/ITWSDataScience/Team2_2020.

c)

Our hypotheses defined earlier were analysed using visualisations (Image 2 - 7). The initial visualisations helped us to see if there were any relationships (ideal or non ideal) between healthcare, health insurance and heart diseases. The visualisations gave us all the information that we needed to prove that our hypotheses were true for a lot of states in the USA. This was the Data exploration part in our data workflow (Image 1) and it helped us to get a better insight into the health scenario surrounding heart diseases. These visualisations also helped us to better plan our processes and move on to the Data analysis part where these relationships could be further examined. Our presentation would highlight the areas where heart disease is more prevalent and health insurance or healthcare is not as prevalent.

4.

There are three main logical collections for this project, and they can be viewed on the GitHub repository. First, the entire project is contained in one directory. Here, there are two folders and an R script. The second and third collections are the two folders: original_datasets and analysis_dataset. The original_datasets folder contains the CSV files for each of the five original datasets, the datasets.txt, and the PDF with metadata for the population data. The other folder, analysis_dataset contains the newly generated dataset from the R script and the metadata that goes with it.

For the physical data handling, CSV files were downloaded from five different sites. Each of the CSV files were added to the GitHub repository along with their download links. The newly created data was produced from the R script and is kept in a separate folder on the GitHub repository. Metadata for the new data is also added to this folder.

This project provides interoperability support, as the entire project is available online. Also, the data files are in CSV format, which is a commonly used format and can be opened and read using nearly all programming tools. The project can be downloaded and run/viewed on many different systems, and does not contain anything that is specific to only one system.

The security support for this project is handled through GitHub. Those with access to the project are able to edit and add to it, but those without access are only permitted to view it. Any security issues that arise will be found by GitHub, and an automated email will be sent out to the developers of the project. This is unlikely though, as the project does not contain any dependencies.

The original datasets are owned by their respective organizations, which can be viewed at their download sites. This project makes no claim to these datasets. The data generated by the R script belongs to Rensselaer Polytechnic Institute.

Metadata for each of the original datasets was embedded in the website they were downloaded from. Therefore, to properly include all metadata, the links to the download sites are provided in datasets.txt. The links will take users directly to the webpage where the data can be downloaded, and they will be able to see and navigate through the available metadata on the webpage. One dataset, the population dataset, had a PDF file alongside it that contained metadata

about its attributes. This PDF file was downloaded and included with the datasets in the 'original_datasets' folder. The metadata for the newly created data is in JSON format, and is included in the 'analysis_dataset' folder with its corresponding dataset.

All the work related to this project is uploaded on the Github Repository that was initially created for the team members to coordinate their work. This platform would help in persistence of the data, metadata, visualisations and analyses that can be used to replicate the work done by the team on this project. The repository is private and is currently shared within the team and the Professor.

As discussed, since the data is available on Github, it would be discoverable by anyone who has access to it. The access to this repository is limited and can be given in case someone needs to access anything related to this project.

This project will be presented in the Data Science poster session at Rensselaer Polytechnic Institute on December 17th, 11.30 a.m.

Works Cited

"Definitive Healthcare: USA Hospital Beds." *Open Data DC*, 13 Nov. 2020, https://opendata.dc.gov/datasets/1044bb19da8d4dbfb6a96eb1b4ebf629_0.

"Health – United Nations Sustainable Development." *United Nations* , 2020, https://www.un.org/sustainabledevelopment/health/#:~:text=Goal%203%3A%20Ensure%20healthy%20lives,for%20all%20at%20all%20ages&text=Ensuring%20healthy%20lives%20and%20promoting,is%20essential%20to%20sustainable%20development.

"Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County – 2016-2018 | Chronic Disease and Health Promotion Data & Indicators." *Open Data | Centers for Disease Control and Prevention | Chronic Disease and Health Promotion Data & Indicators*, 27 Apr. 2020, https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/6x7h-usvx.

Martineau, Angelina, et al. "ITWS Data Science Team 2 2020." *GitHub*, 2020, https://github.com/ITWSDataScience/Team2_2020.

"RStudio | Open Source & Professional Software for Data Science Teams - RStudio." *RStudio | Open Source & Professional Software for Data Science Teams - RStudio*, 2020, https://rstudio.com/.

"US Census Bureau COVID-19 Site." *US Census Bureau COVID-19 Site*, 6 June 2020, https://covid19.census.gov/datasets/b69c7076eaa8433eab8f7fa077bac3b6/data.

Census, United State Census Bureau, 18 Apr. 2019, https://www2.census.gov/programs-surveys/popest/datasets/2010-2018/counties/totals/.