Project #2: DATA EXPLORATION AND DESIGN

EN 605.662.SU20 Data Visualizations

Johns Hopkins University

11/9/2020

Abstract

Not only are effective data visualizations used to encode analyzed information for viewers to decode, but they ae also used to aid in data exploration. This project sought to explore a large dataset by examining key variable and their interactions visually.

Description of the Visualization

This project utilized the Center for Disease Control's 500 Cities Project data set (see Figure 1 for example of data visualization; CDC, 2016; Wang et al., 2018; Wang et al., 2017; Zhang et al., 2014). This project reports city and census tract-level data, obtained using small area estimation methods, for 27 chronic disease measures for the 500 largest American cities. To improve local population health, residents and healthcare provides will need to harness the power of current health status data and behavioral risk factors within their own communities. Here, data can be aggregated by State, City, or Census-tract and can be analyzed across 28 chronic disease measures are related to unhealthy behaviors (5), health outcomes (13), and use of preventive services (10). These measures include major risk behaviors that lead to illness, suffering, and early death related to chronic diseases and conditions, as well as the conditions and diseases that are the most common, costly, and preventable of all health problems.

As a former neuroscientist, understanding stroke risk factors is of a particular interest. And as a population health programs manager, I am also interested in access to healthcare. This analysis will examine stroke risk factors across U.S. states and Counties as well as the interaction between access to health care and stroke risk. Further analysis will focus solely on chronic disease risk factors in South Carolina, my home state.

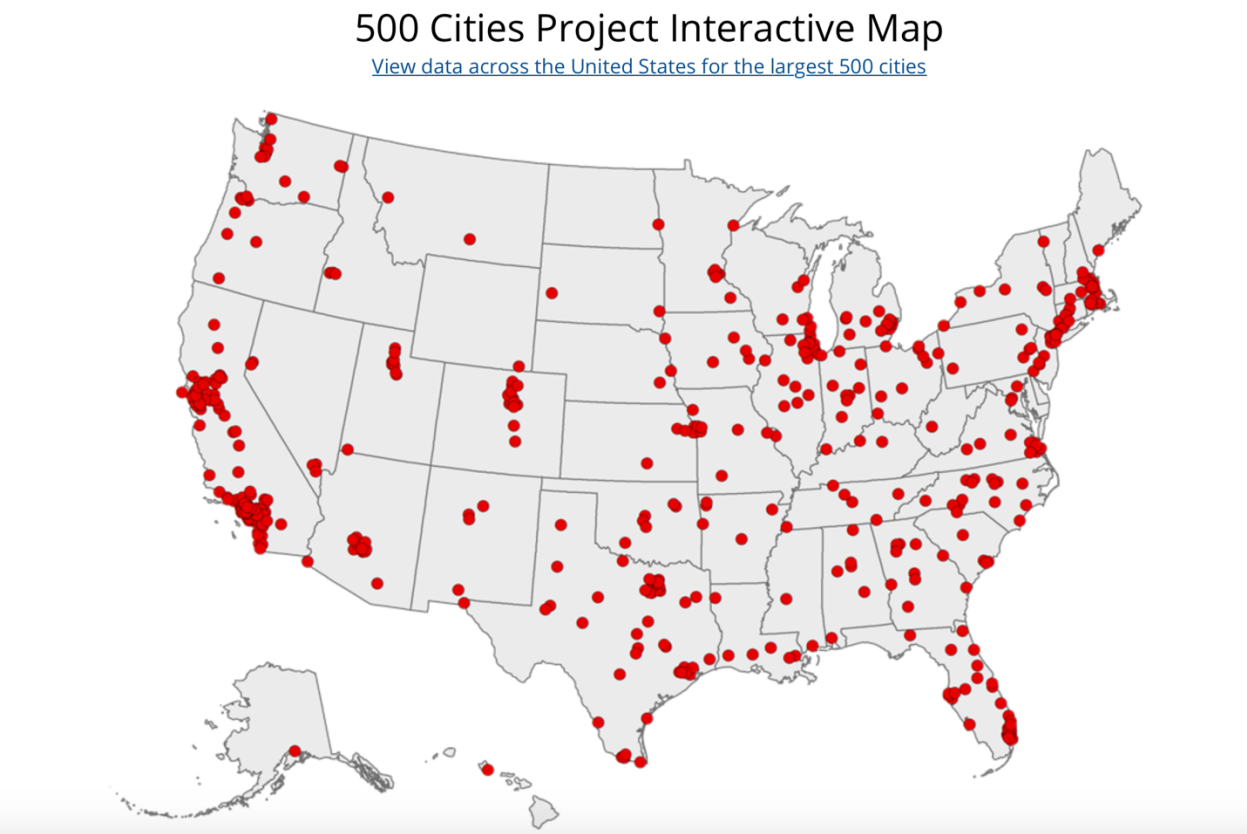

FIGURE 1. Center for Disease Control's 500 Cities Project interactive map.

Data

500 U.S. Cities

The 500 Cities Project includes data from 497 of the largest U.S. Cities, as well as data from 3 of the smallest: Burlington, Vermont; Charleston, West Virginia, and Cheyenne, Wyoming. The number of cities per state ranges from 1 to 121.Among these 500 cities, there are approximately 28,000 census tracts, for which data will be provided. The tracts range in population from less than 50 to 28,960, and in size from less than 1 square mile to more than 642 square miles. The number of tracts per city ranges from 8 to 2,140. The project includes a total population of 103,020,808, which represents 33.4% of the total United States population of 308,745,538.

28 Chronic Disease Risk Factors

The 500 Cities Project includes 28 chronic disease risk factors, which are separated into three categories: health outcomes, prevention, and unhealthy behavior. Each attribute within a category includes 3 measurements: raw-data, adjusted-data, and 95% confidence intervals for adjusted data. Most of the data is presented as a ratio of individuals who reported having a specific feature, divided by the total number of respondents. All respondents are aged ≥ 18 years.

There are 5 unhealthy behaviors included in the data set, which are all positively correlated with chronic disease: binge drinking among adults age ≥ 18 years, current smoking among adults ≥ 18 years, no leisure-time physical activity among adults ages ≥ 18 years, obesity among adults ages ≥ 18 years, and sleeping less than 7 hours among adults ≥ 18 years.

There are 13 health outcomes included in the data set, which are all positively correlated with chronic disease: current asthma among adults aged ≥ 18 years, high blood pressure among adults ages ≥ 18 years, cancer among adults aged ≥ 18 years, high cholesterol among adults aged ≥ 18 years who have been screened in the past 5 years, chronic kidney disease among adults aged ≥ 18 years, chronic obstructive pulmonary disease among adults aged ≥ 18 years, coronary heart disease among adults aged ≥ 18 years, diagnosed diabetes among adults ≥ 18 years, mental health not good for ≥ 14 days among adults aged ≥ 18 years, physical health not good for ≥ 14 days among adults aged ≥ 18 year, and stroke among adults aged ≥ 18 years.

There are 10 attributes indicating the Use of Preventative Services, which are all correlated with chronic disease: current lack of health insurance among adults aged 18-64 years, visits to the doctor for routine checkups within the past year among adults aged ≥ 18 years, visits to the dentist or dental clinic among adults aged ≥ 18 years, taking medicine for high blood

pressure control among adults aged ≥ 18 years, cholesterol screening among adults aged ≥ 18 years, mammography use among women aged 50-74 years, panicolaou smear testing among adult women aged 21-65 years, fecal occult blood test/sigmoidoscopy/colonoscopy among adults aged 50-75 years, older men aged ≥ 65 years who are up to date on a core set of clinical preventative services, and older women aged ≥ 65 who are up to date on a core set of clinical preventative services.

Data Exploration

Data analysis was conducted in R (and Python – but is only presented in R for this assignment). For purposes of this project, only the adjusted risk factor values ("_AdjPrev") were analyzed for each of the 28 attributes. Descriptive statistics were run for each variable in order to obtain: min, max, mean, median, mode, standard deviation, standard error, skew, kurtosis, sum, range, 1st quartile, 2nd quartile, and 3rd quartile (see Table 1 below).

Variable	Data Type	Variable Type	Min	Max	Mean	Median	Std	Description
Access2	Numeric Float	Quantitative Ratio Discrete	4.10	49.0	18.21	17.40	6.85	Lack of health insurance
Arthritis	Numeric Float	Quantitative Ratio Discrete	15.70	35.80	23.18	22.80	3.50	Diagnosed with Arthritis
Binge	Numeric Float	Quantitative Ratio Discrete	7.10	25.40	16.36	16.30	2.58	"Binge Drinking": 5+ (men) or 4+ (women) drinks in the past 30 days
BPHigh	Numeric Float	Quantitative Ratio Discrete	22.50	47.80	30.79	30.00	4.53	High Blood Pressure
BPMed	Numeric Float	Quantitative Ratio Discrete	48.50	72.10	58.21	59.20	5.48	Taking medicine for high blood pressure
Cancer	Numeric Float	Quantitative Ratio Discrete	4.10	6.9	5.852	5.9	0.49	Diagnosed with cancer (besides skin)

CAsthma	Numeric Float	Quantitative Ratio Discrete	6.6	14.7	9.332	9.10	1.34	Currently diagnosed with asthma
CHD	Numeric Float	Quantitative Ratio Discrete	3.9	8.60	5.956	6.0	0.93	Angina or Coronary Heart Disease
CheckUp	Numeric Float	Quantitative Ratio Discrete	54.80	80.98	67.97	67.30	5.03	Visits to a doctor for a routine checkup within the past year
CholScreen	Numeric Float	Quantitative Ratio Discrete	64.20	82.50	73.83	73.90	3.09	"Cholesterol screen": have had their cholesterol checked within the previous 5 years
ColonScreen	Numeric Float	Quantitative Ratio Discrete	44.90	76.90	62.63	63.15	5.54	Fecal occult blood (FOBT)test within the past year, sigmoidoscopy within the past 5 years and FOBT within the past 3 years, or colonoscopy within the past 10 years
COPD	Numeric Float	Quantitative Ratio Discrete	3.30	11.40	31.38	31.50	1.54	Chronic obstructive pulmonary disease (COPD), emphysema, or chronic bronchitis
CoreM	Numeric Float	Quantitative Ratio Discrete	17.70	49.20	31.85	31.90	5.14	"Core Men": Older men aged ≥ 65 years who are up to date on a core set of clinical

								preventative services
CoreW	Numeric Float	Quantitative Ratio Discrete	16.30	44.90	30.56	30.70	5.16	"Core Women": Older women ≥ 65 who are up to date on a core set of clinical preventative services
CSmoking	Numeric Float	Quantitative Ratio Discrete	7.90	31.40	18.14	17.90	4.37	Current smoker
Dental	Numeric Float	Quantitative Ratio Discrete	38.80	79.70	62.03	62.05	7.53	Visits to dentist or dental clinic
Diabetes	Numeric Float	Quantitative Ratio Discrete	5.40	18.20	10.34	10.15	2.42	Diabetes diagnosis
HighChol	Numeric Float	Quantitative Ratio Discrete	26.50	38.70	33.09	33.0	1.92	High Cholesterol
Kidney	Numeric Float	Quantitative Ratio Discrete	1.70	4.60	2.781	2.70	0.48	Chronic kidney disease
LPA	Numeric Float	Quantitative Ratio Discrete	11.60	41.60	24.05	23.95	5.48	No leisure-time physical activity
MammoUse	Numeric Float	Quantitative Ratio Discrete	62.80	88.90	77.80	78.55	3.96	Mammography use among women ≥ $50 - 74$
MHlth	Numeric Float	Quantitative Ratio Discrete	7.10	18.40	12.17	12.10	2.23	Mental health not good for ≥ 14 days
Obesity	Numeric Float	Quantitative Ratio Discrete	15.20	47.20	29.16	29.25	5.76	Currently obese (i.e., body mass index ≥ 65 30.0kg/m^2
PapTest	Numeric Float	Quantitative Ratio Discrete	69.30	89.40	80.64	80.80	3.50	Papanicolaou smear test among women ≥ 21-65 years
PHlth	Numeric Float	Quantitative Ratio Discrete	6.40	20.20	12.55	12.50	2.71	Physical health not good for ≥ 14 days
Sleep	Numeric Float	Quantitative Ratio Discrete	24.50	52.0	35.69	35.40	4.53	Sleeping less than 7 hours

Stroke	Numeric Float	Quantitative Ratio Discrete	1.70	5.80	3.001	2.90	0.68	Stroke
TeethLost	Numeric Float	Quantitative Ratio Discrete	5.10	30.40	14.51	14.10	4.94	All teeth lost

TABLE 1. Descriptive statistics for all rate adjusted chronic disease risk factors in the CDC 500 Cities Project Data Set

Interesting variables were also visualized individually, in order to examine distribution prior to further data processing. Figures 2 shows a histogram of the Stroke risk factor and Figure 3 shows a Normal QQ Plot of the Stroke risk factor. Although the mean stroke ratio was 3.001 and the median ratio was 2.90, visually, we see the histogram is slightly skewed left. Descriptive statistics indicate skew = 0.86 and kurtosis = 1.14, suggesting we may want to normalize this data before further analysis. Indeed, we see some slight tail deviations on the Normal QQ Plot, as well.

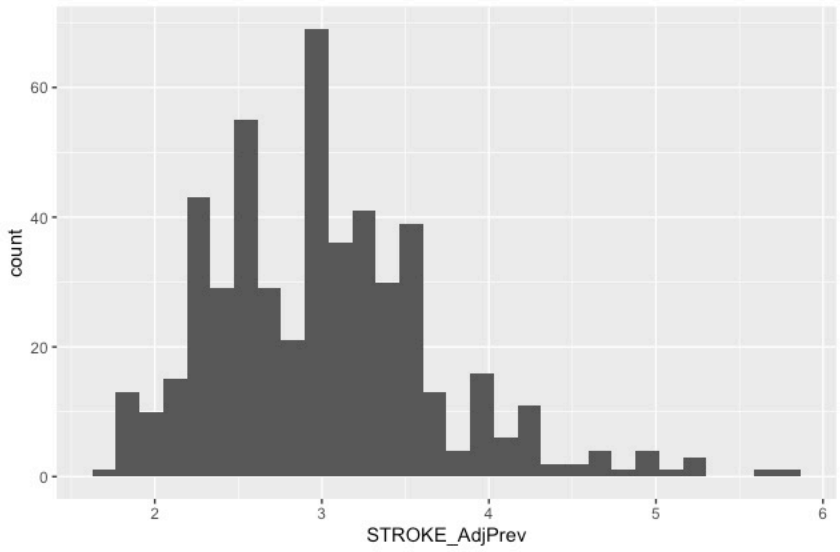

FIGURE 2. Histogram of the Stroke-adjusted risk factor.

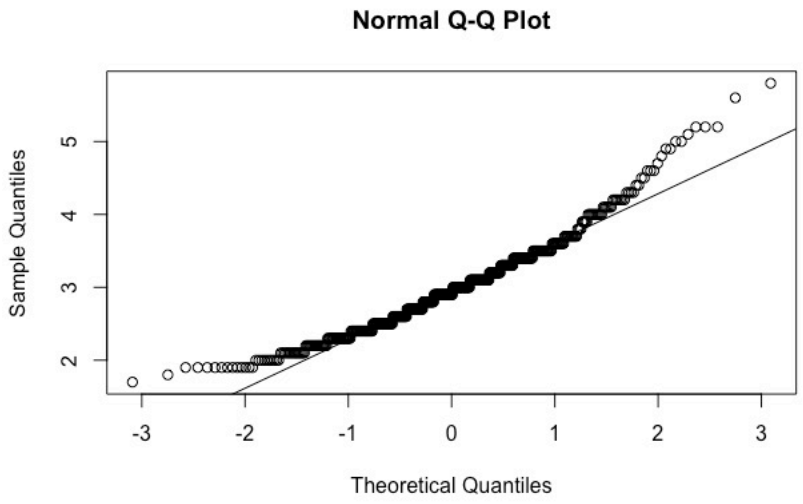

FIGURE 3. Normal QQ-Plot of the Stroke-adjusted risk factor.

Exploratory data visualizations were also performed on the Access risk factor. Figures 4 shows a histogram of the Access risk factor and Figure 5 shows a Normal QQ Plot of the Access risk factor. The mean Access ratio was 18.21, the median ratio was 17.40, and the standard deviation was 6.85. We confirm this visually by examining the histogram in Figure 4, which is skewed left. Descriptive statistics indicate skew = 0.99 and kurtosis = 1.65, suggesting we may want to normalize this data before further analysis. Indeed, we see some slight tail deviations on the Normal QQ Plot (Figure 5), as well.

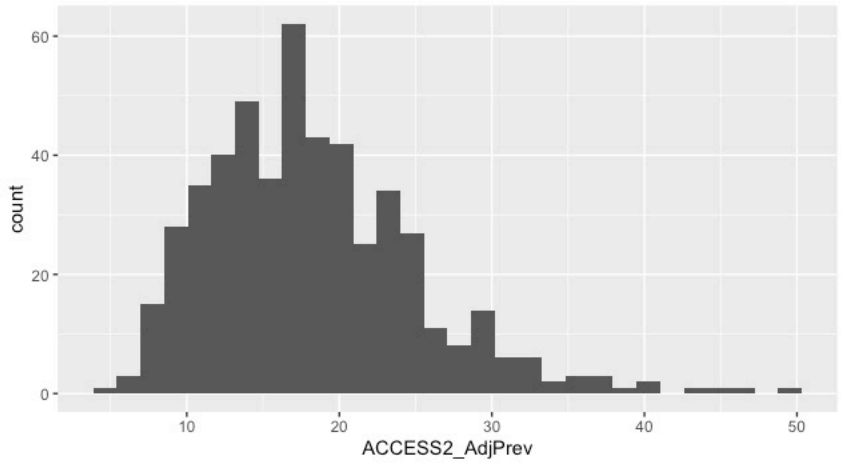

FIGURE 4. Histogram of the Access-adjusted risk factor.

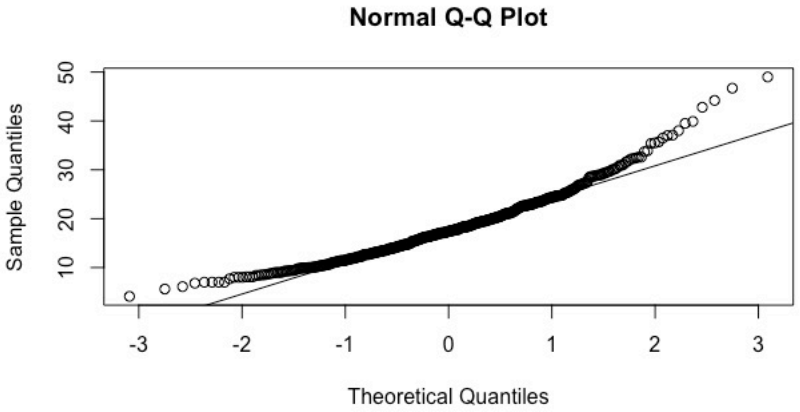

FIGURE 5. Normal QQ-Plot of the Access-adjusted risk factor.

Five Analytical Questions

Question 1

Do regions differ greatly in their stroke risk rate?

Question 2

What other risk factors are correlated with stroke risk?

Question 3

What is the relationship between stroke risk and access to health insurance?

Visualization Sketches

Do Regions Differ in Stroke Risk Rate?

Figure 6 shows a boxplot comparison of stroke risk rate by state. As a quantitative (ratio) variable type, stroke risk is well-suited for a boxplot. Boxplots visually display a variable's: minimum, 1st quartile, median, 3rd quartile, and maximum. The box indicates the interquartile range (Q1-Q3) and the median is located in the middle of the box. Whiskers extend out from the box to the minimum value (Q1-1.5*IQR) and to the maximum value on the opposite end (Q3+1.5*IQR). Outliers are displayed as dots located beyond the minimum or maximum values. Here, we can compare variability and central tendency across multiple states. Unfortunately, with so many states, comparing all boxplots in one visualization is not practical.

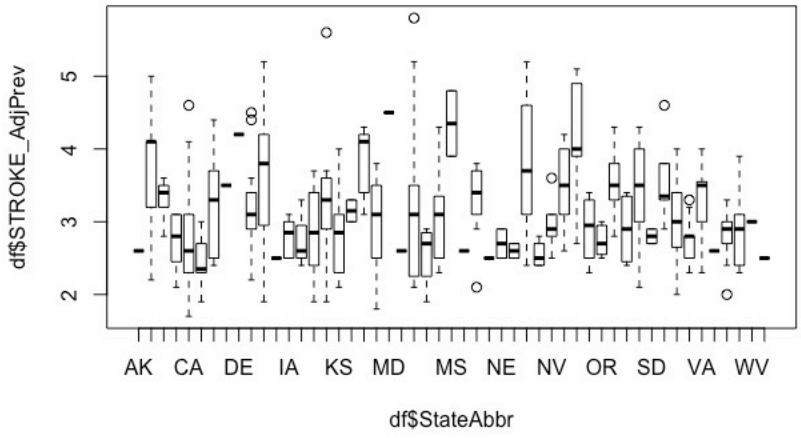

Figure 6. Boxplot of Stroke-adjusted risk rate by State.

Figure 7 shows the stroke-adjusted risk rate by state using a geo-local visualization. Here, a color gradient is used to indicate stroke-risk strength (yellow max, black min). Compared to the boxplot above, the stroke-risk rate is much easier to observe and compare across states. We can easily see the maximum risk located in the state of Maryland (4.5) and the minimum risk located

in the state of Colorado (2.43). In the bottom left-hand corner of the graph is the average stroke risk across all states (3.1). Overall, we can see a much higher stroke risk for eastern U.S. states.

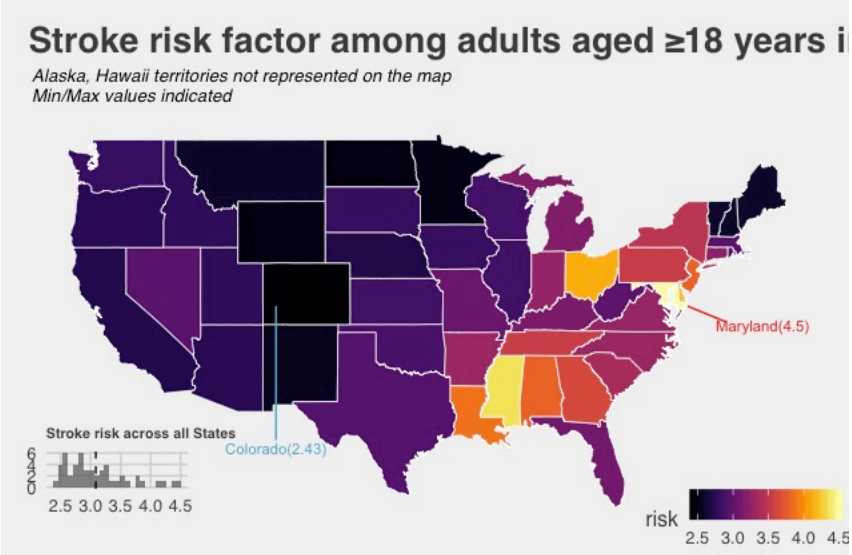

FIGURE 7. Color-gradient Geo-location Map of stroke-adjusted risk rate by State.

Similarly, Figure 8 shows the stroke-adjusted risk rate by city using a geo-local visualization. Here, a color gradient is used to indicate stroke-risk strength (yellow max, black min). The same overall pattern is observed: eastern U.S. cities have a higher stroke risk rate. The stroke risk rate across all U.S. cities is 3% (as shown in the bottom-left of the figure).

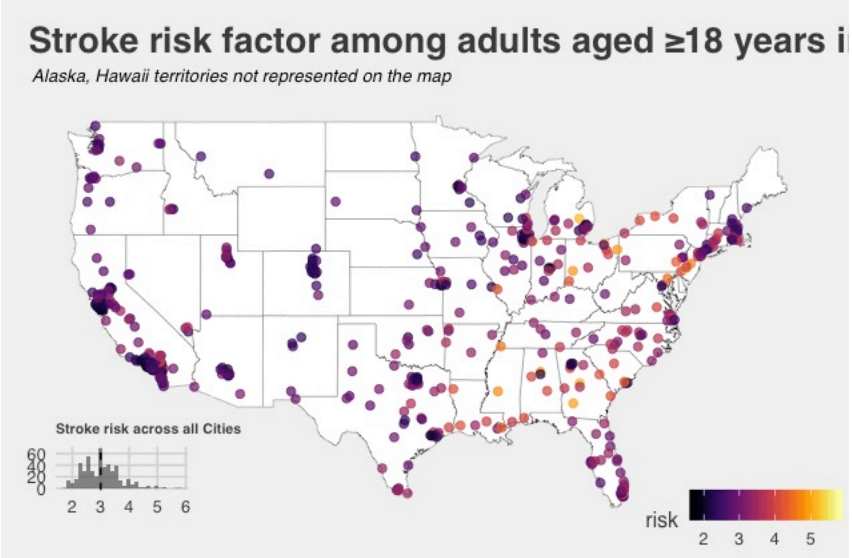

FIGURE 8. Color-gradient Geo-location Map of stroke-adjusted risk rate by State.

Other Risk Factors Correlated With Stroke Risk

Research indicates that there are many lift style risk factors which are positively correlated to stroke rates. These include: obesity, physical inactivity, binge drinking, high blood pressure, cigarette smoking, high cholesterol, and diabetes. As these have been researched previously, we can look for confirmation within this data set. Figure 9 shows a color-gradient correlation plot comparing all risk factor variables within this dataset. As there are a lot of variables to consider, this may not be the most effective means to relay information.

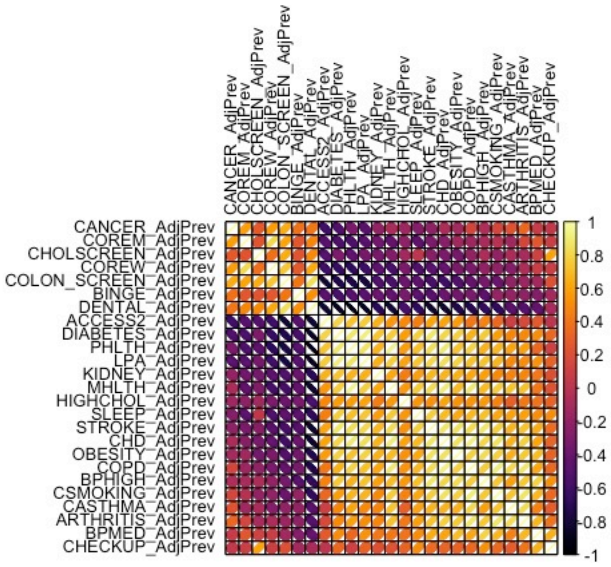

FIGURE 9. Correlation plot comparing each chronic disease risk factor in this data set.

Figure 10 illustrates all the negative correlations between stroke risk, including: binge drinking (r = -0.490250), cholesterol screening (r = -0.2489038), core men (r = -0.3141969), core women (r = -0.4742796), and dental (r = -0.8321158). We observe that the strongest negative correlation with the dental risk factor.

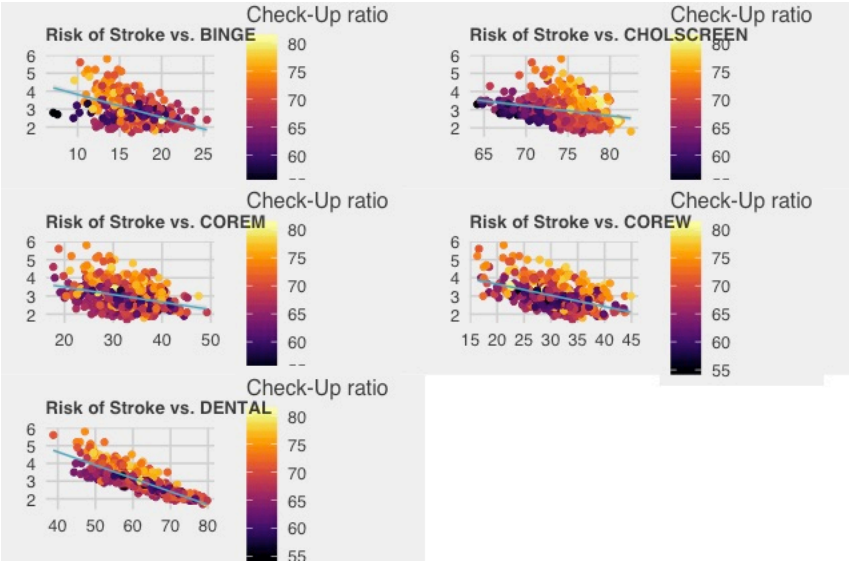

FIGURE 10. Negative correlations with stroke risk.

Figure 11 illustrates all the positive correlations between stroke risk, including: High blood pressure (r = 0.8785124), Coronary Heart Disease (r = 0.8718858), COPD (r = 0.8297992), Current Smoker (r = 0.7806072), Limited Physical Activity (r = 0.8135 r =452), Diabetes (r = 0.8780051), Obesity (r = 0.8520891), Poor physical health (r = 0.9423821), and Sleep (r =0.7783179). We observe that most of these factors have a strong positive correlation.

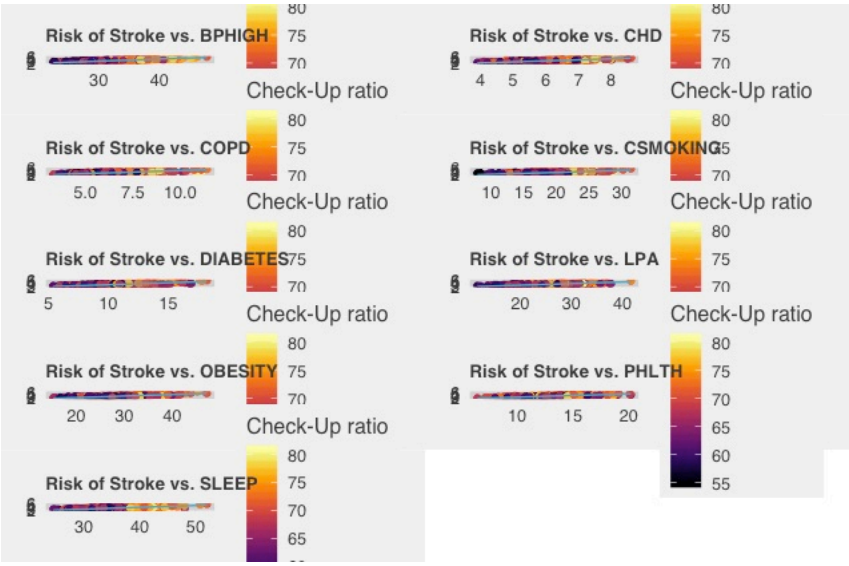

FIGURE 11. Positive correlations with stroke risk.

Relationship Between Stroke Risk and Access to Health Insurance

Although this topic may have changed after the Affordable Care Act, it's the one relationship that I don't see discussed much in research. Do individuals with health insurance have a lower stroke risk rate? Figure 12 shows the stroke risk rate as a function of lack of healthcare insurance throughout the U.S. The trend here looks virtually identical to the one we saw in Figure 7, suggesting there may not be a strong relationship between the two variables. Perhaps this is actually due to the fact that it is easier to access healthcare services in larger cities. To consider this possibility, we'll need to separate cities by size and investigate further.

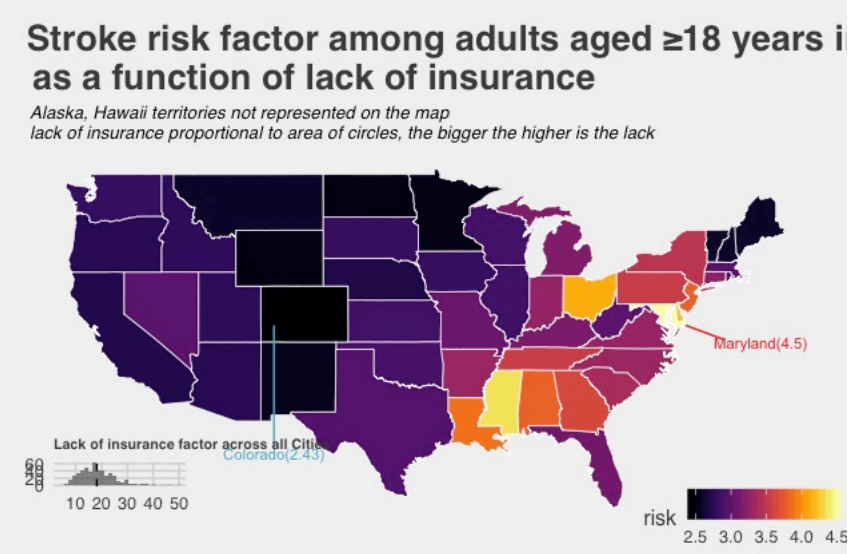

FIGURE 12. Stroke risk rate as a function of lack of insurance throughout the U.S

Figure 13 demonstrates the correlation between stroke risk rate and access risk (i.e., risk due to a lack of health insurance in highly populated cities and low populated cities. For both population sizes, we see the same trend: the lower the lack of health insurance, the lower the risk of stroke. However, for both population sizes, we see a cone-shaped trend, suggesting something else may be influencing the data.

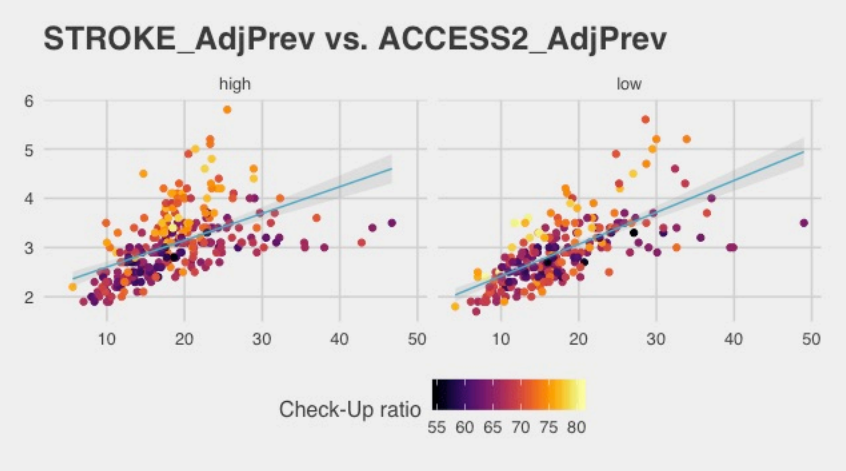

FIGURE 13. Stroke risk rate's correlation with Access (i.e., lack of insurance) risk rate in highly-populated cities and low-populated cities.

Finally, Figure 14 demonstrates a heatmap of features to illustrate the arrangement of clusters produced by hierarchical clustering.

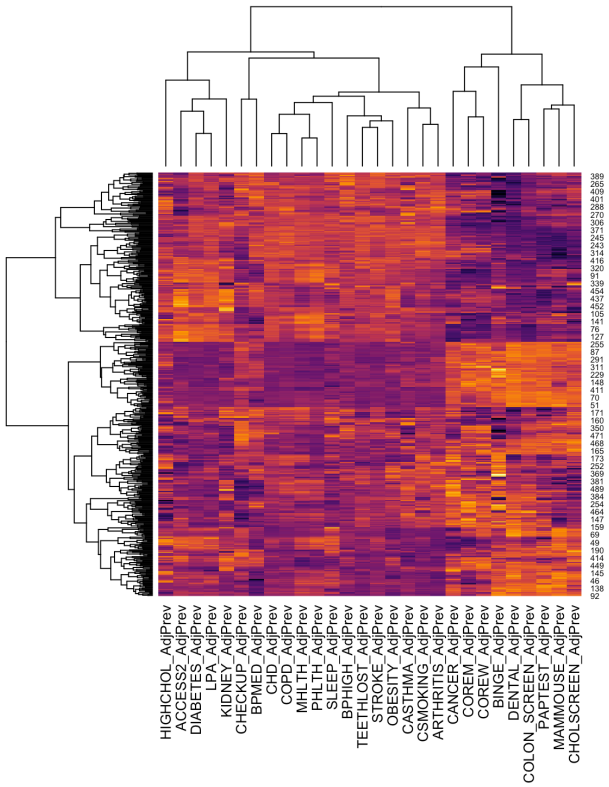

FIGURE 14. A heatmap of features to show hierarchical clustering of risk factors.

References

Centers for Disease Control., (2016, October 28). 500 Cities Project. Retrieved November 9, 2020, from https://chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2019-relea/6vp6-wxuq

Wang, Y., et al., (2017). Comparison of Methods for Estimating Prevalence of Chronic Diseases and Health Behaviors for Small Geographic Areas: Boston Validation Study. Preventing Chronic Disease, 14, 170-281.

Wang, Y. et al. (2018). Using 3 Health Surveys to Compare Multilevel Models for Small Area Estimation for Chronic Diseases and Health Behaviors. Preventing Chronic Disease, 15, 180-313.

Zhang, X., et al. (2014). Outcomes: A Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System. American Journal of Epidemiology, 179(8), 1025-1033.