Spatial Variations of Health Conditions in the USA

Emily Zhou, Christina Chen, and Liam Smith

Who Cares?

The COVID-19 pandemic has highlighted systemic inequities in healthcare resource accessibility and health outcomes across the USA. It would be beneficial for policymakers to identify regions that are particularly vulnerable in order to reduce health disparities. For this purpose, our analysis seeks to illustrate how health outcomes vary across space due to a number of risk factors.

- **How do different factors affect health outcomes?**
- **Do health outcomes vary across space? By including geographic location in models that predict health outcomes, can we increase the reliability of our results?**
- **Are hospital beds accessible to all population across the country? How does hospital beds accessibility affect people's health?**
- **Are certain population more vulnerable than others under the pandemic?**

Data Description

Observations

- Counties

Variables

- Demographic variables include age & sex, race, income, education level, and total population
- Health variables include diabetes rate, obesity rate, uninsured rate, disease rates, mortality, unhealthy days per month, years of potential life lost, etc...
- Hospital variables include location, the number of hospital beds, and the type of hospital.

Data Sources

- Census (ACS 2019)
- Social Explorer
- Homeland Infrastructure Foundation

Health Indicator Categorizations

1 Personal Factors

- Diabetes
- Smokers
- Drinkers
- Physically Inactive
- Obesity

2 Environmental Factors

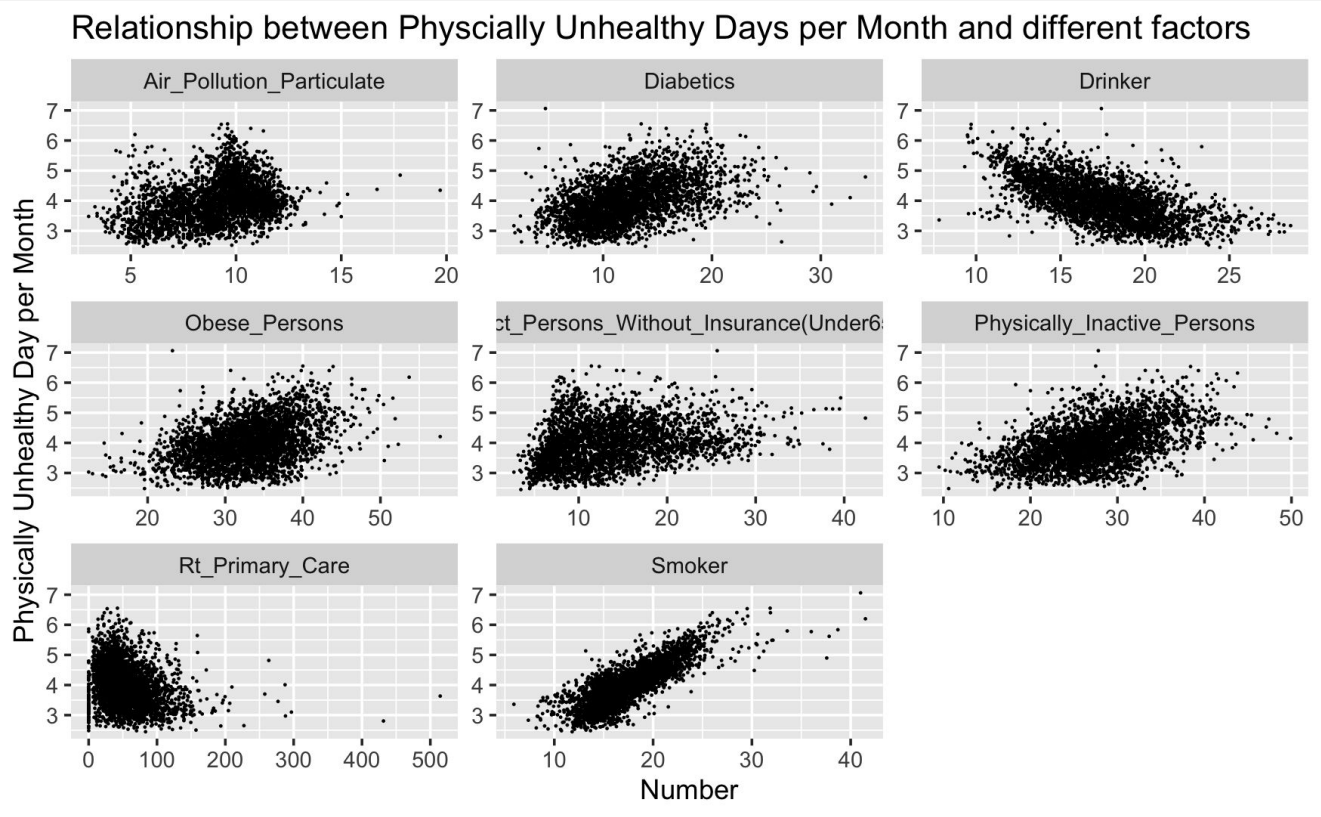
- Air pollution
- Drinking water violation

3 Structural Factors

- Primary care physicians
- Uninsured rate

Checking the validity of running a linear regression model

Based on the scatterplots, a positive linear relationship is observed among diabetics, obese persons, physically inactive persons, as well as smokers.



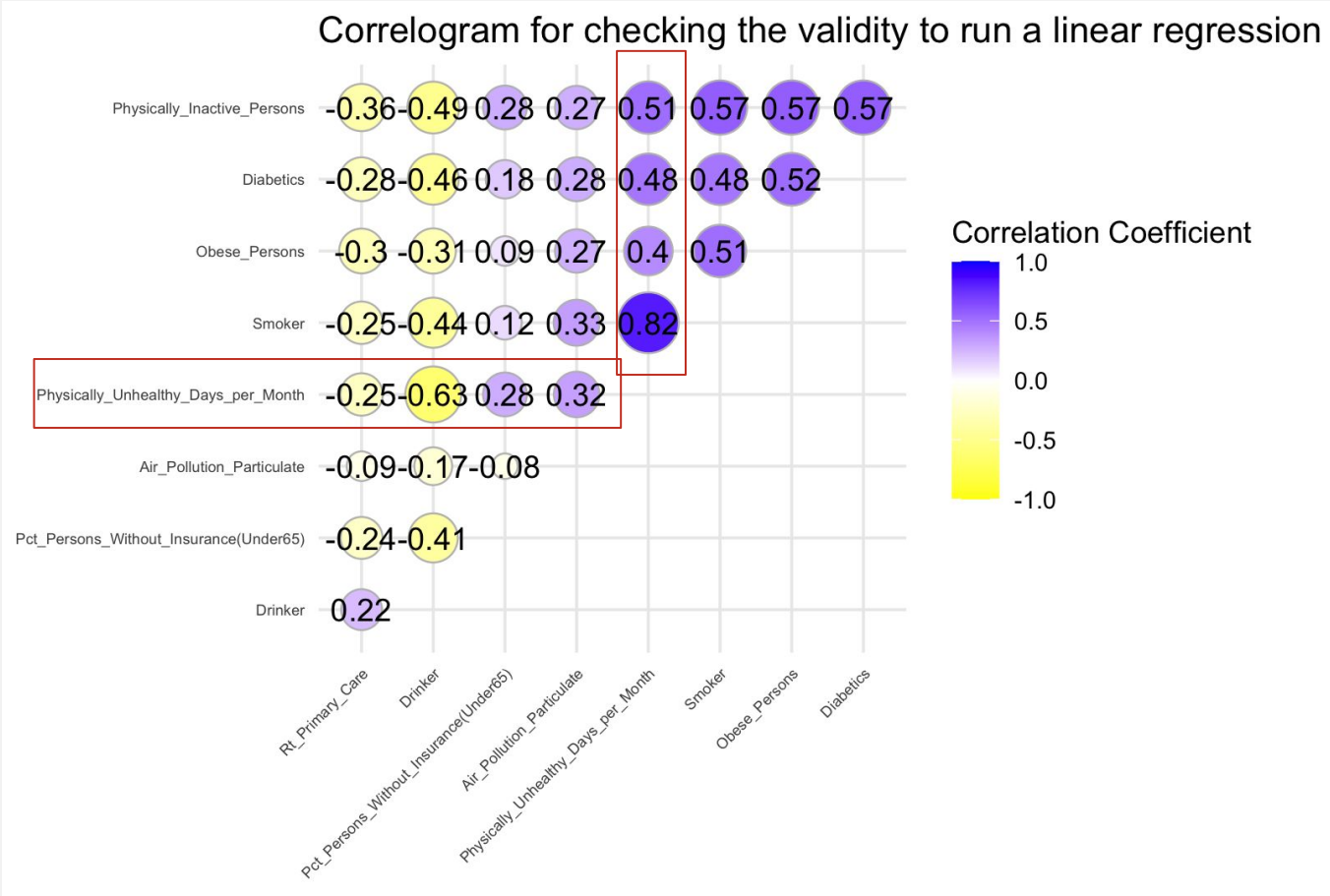

Correlogram for checking the validity to run a linear regression
Physically_Inactive_Persons
Diabetics
Obese_Persons
Smoker
Physically_Unhealthy_Days_per_Month
Air_Pollution_Particulate
Pct_Persons_Without_Insurance(Under65)
Drinker
-0.36 -0.49 0.28 0.27 0.51 0.57 0.57 0.57
-0.28 -0.46 0.18 0.28 0.48 0.48 0.52
-0.3 -0.31 0.09 0.27 0.4 0.51
-0.25 -0.44 0.12 0.33 0.82
-0.25 -0.63 0.28 0.32
-0.09 -0.17 -0.08
-0.24 -0.41
0.22
Rt_Primary_Care
Drinker
Pct_Persons_Without_Insurance(Under65)
Air_Pollution_Particulate
Physically_Unhealthy_Days_per_Month
Smoker
Obese_Persons
Diabetics
Correlation Coefficient
1.0
0.5
0.0
-0.5
-1.0

Table 1. Linear regression outcome predicted by the model.

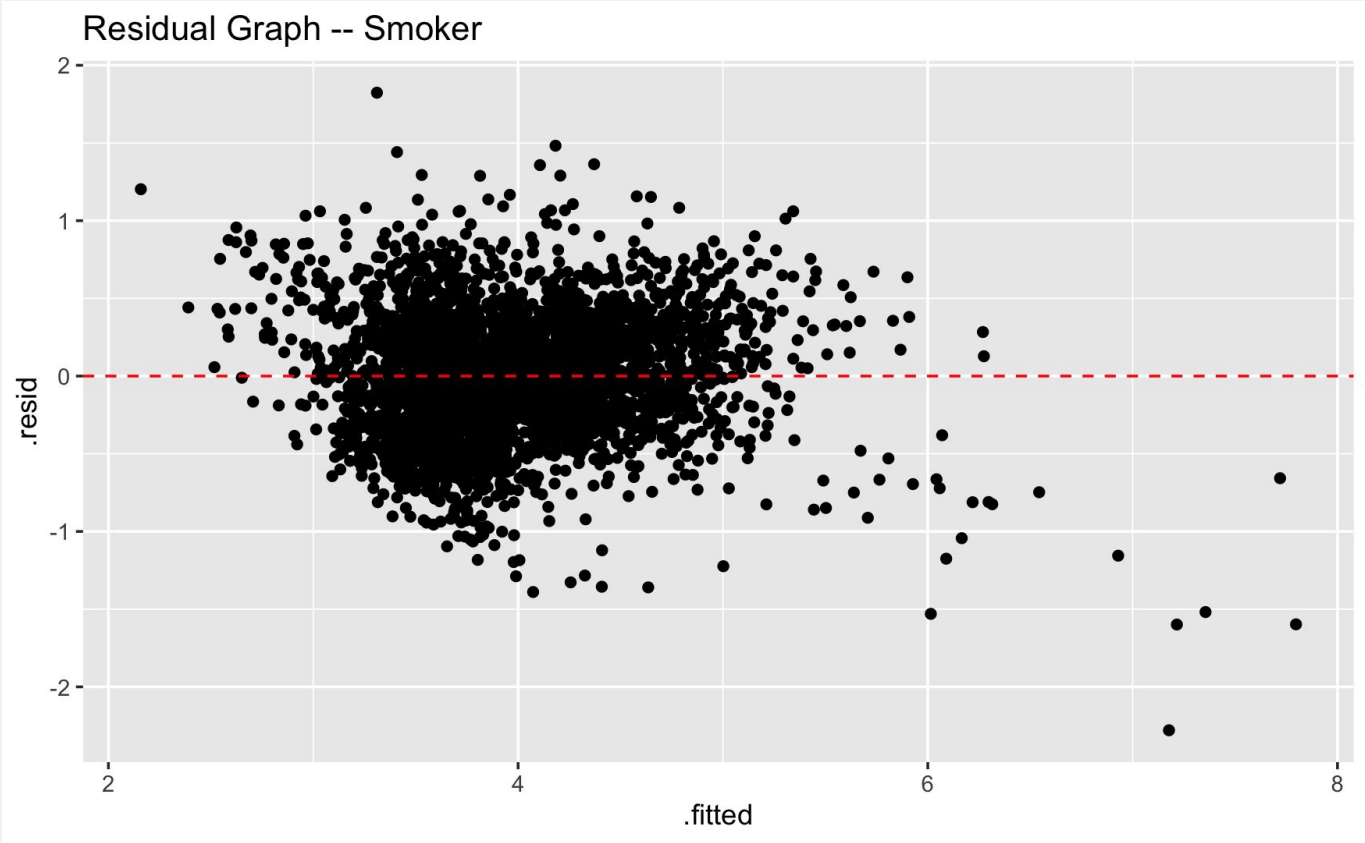
```
## # A tibble: 9 × 5
##   term                                    estimate std.error statistic   p.value
##   <chr>                                      <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                             2.76e+0  0.0874      31.6   1.88e-189
## 2 Rt_Primary_Care                        -1.79e-4  0.000196    -0.911 3.62e-  1
## 3 `Pct_Persons_Without_Insurance(Under65…  1.06e-2  0.00116      9.14  1.11e- 19
## 4 Diabetics                               5.31e-3  0.00205      2.60  9.47e-  3
## 5 Smoker                                  1.38e-1  0.00235     58.8   0
## 6 Drinker                                -6.63e-2  0.00253    -26.2   2.63e-136
## 7 Obese_Persons                          -7.06e-3  0.00151     -4.68  3.06e-  6
## 8 Physically_Inactive_Persons            -7.35e-3  0.00159     -4.63  3.89e-  6
## 9 Air_Pollution_Particulate               2.37e-2  0.00347      6.82  1.12e- 11
```

Table 2. R-Squared of the linear regression model.

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.766         0.765 0.337     1208.       0     8  -978. 1976. 2036.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

The residual plot between physically unhealthy day per month and smoker is displayed here.

It abides the Condition 3 Constant variability as the variability of residual locates around the 0 line.



How do the percentages of population with diabetes, HIV, and Chlamydia vary across space?

How do environmental conditions, such as air pollution and water quality, affect one's health condition?

What are the spatial distributions of uninsured rates and primary care physicians?

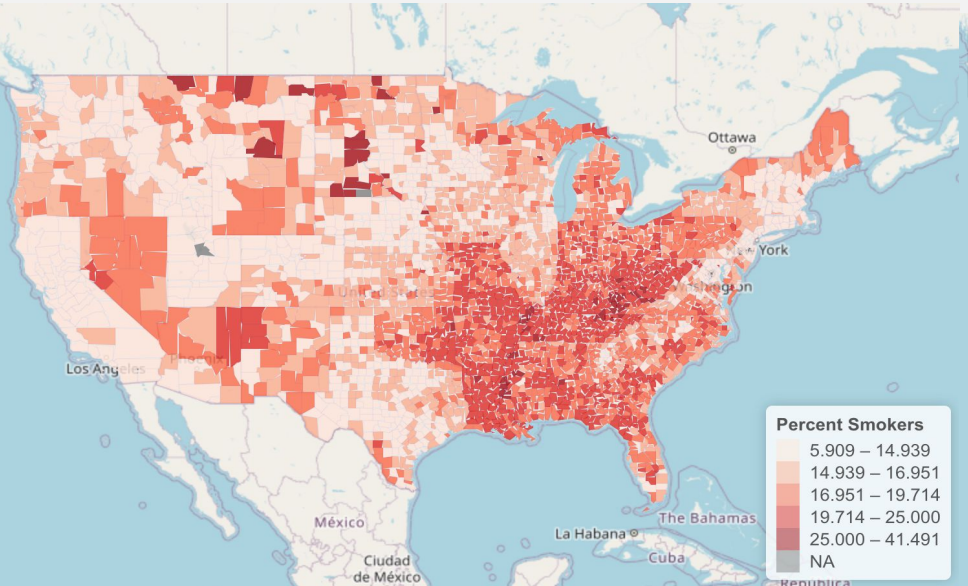
Which parts of the US are healthier?

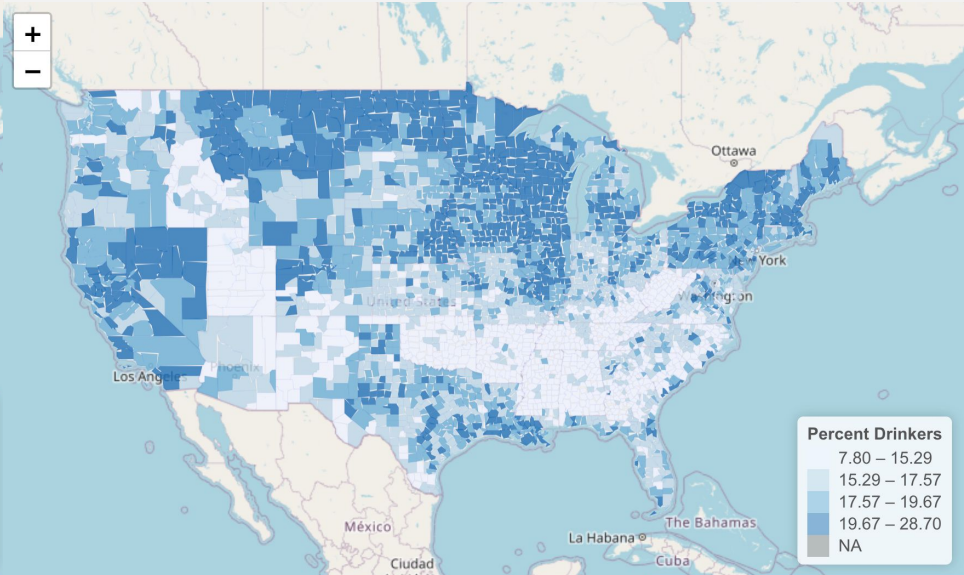
Spatial Distributions of Health Indicators

1 Personal Factors

- Heavy smokers are defined as those who smoke at least every day
- Heavy drinkers include binge drinkers and those who drink every day

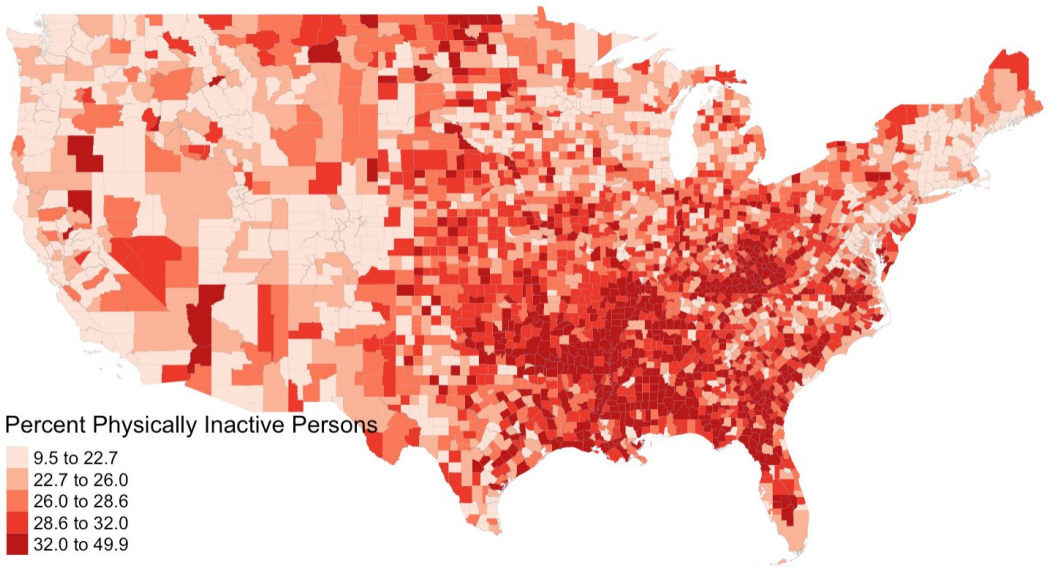
- Survey data
- General overview of geographic distribution of different health factors
- State by state distinctions
- Regions of prevalent smoking and prevalent drinking do not necessarily align

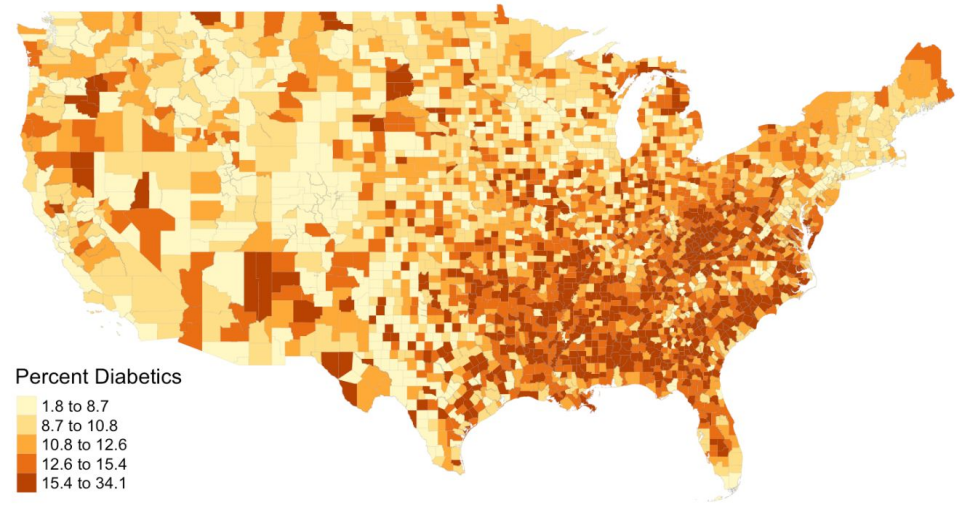


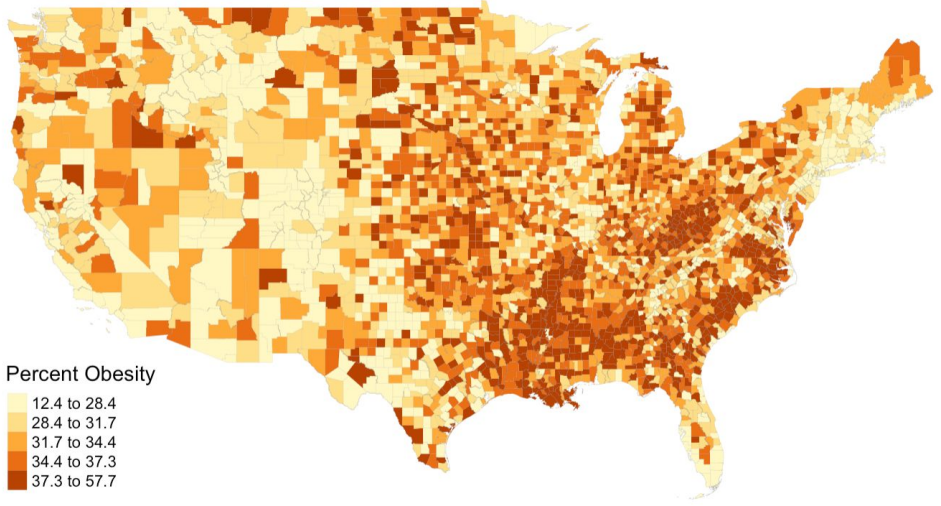


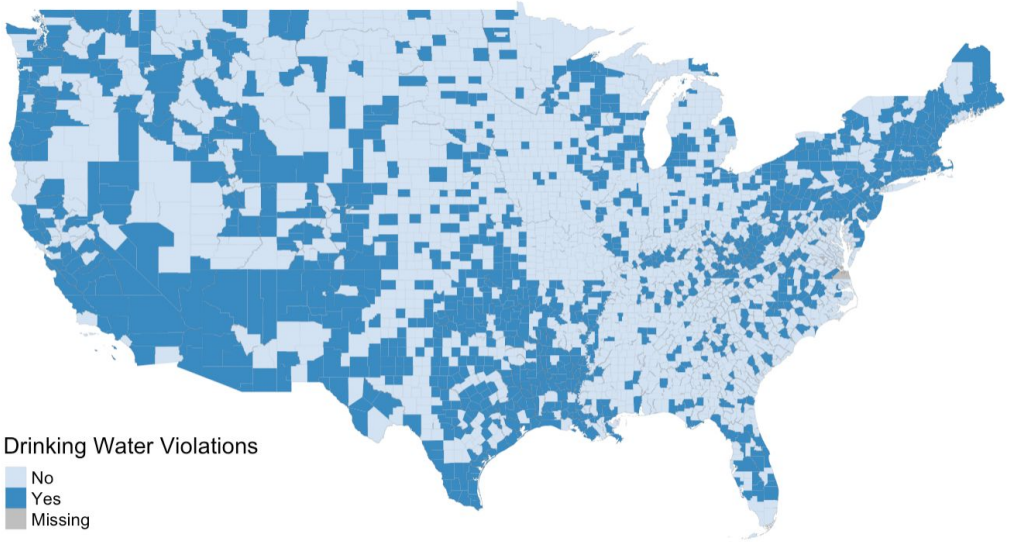
1 Personal Factors (Continued)

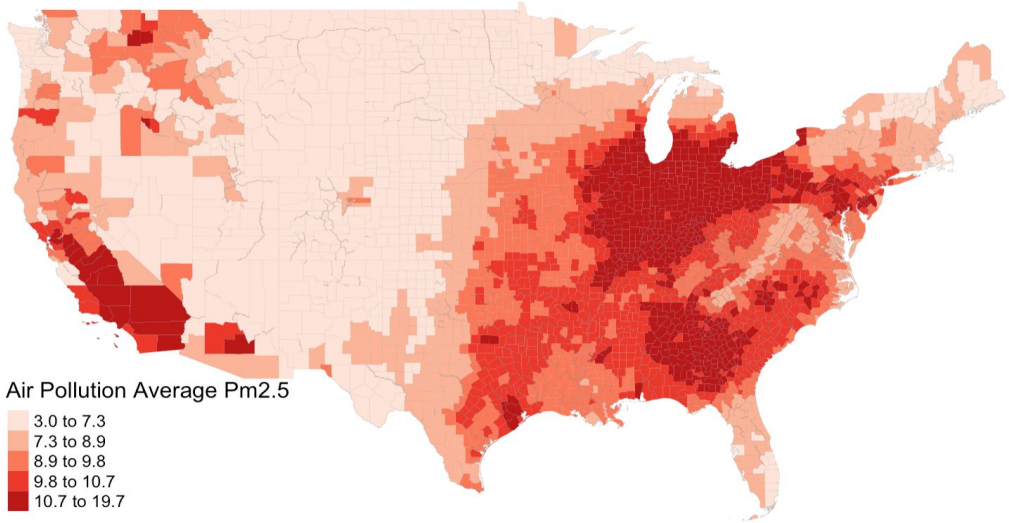
- Very clear clustering of these health risk factors in the Rust Belt/Southeast region
- Bins represent quantiles
- Survey data









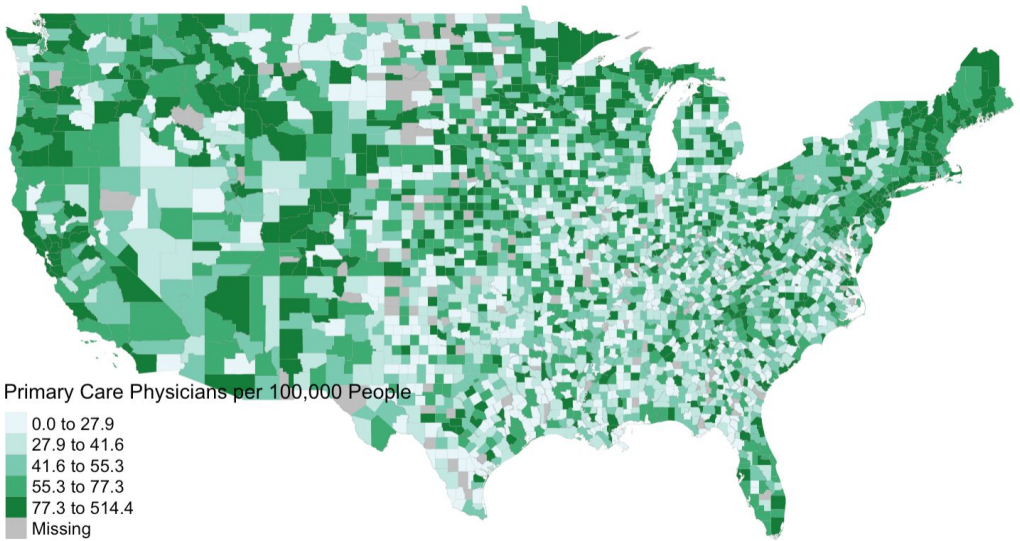


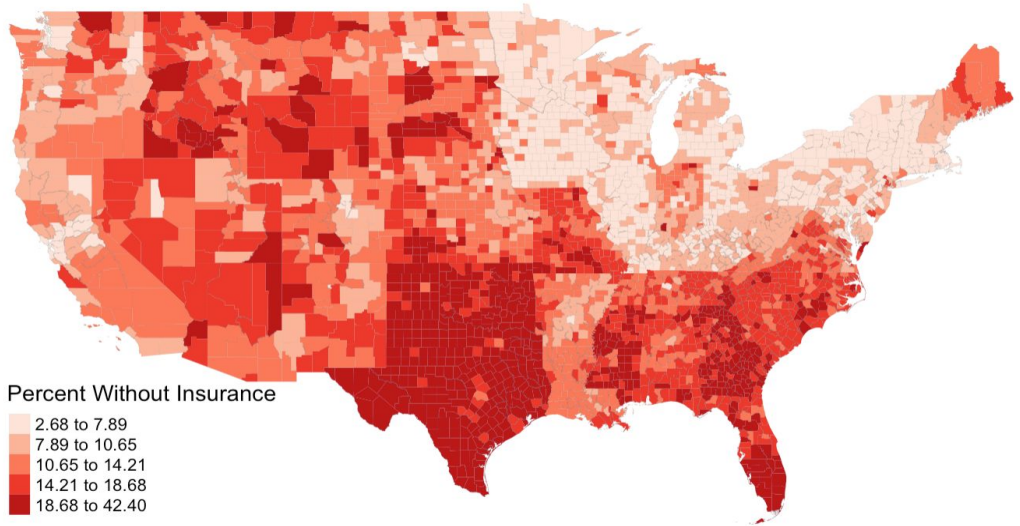
2 Environmental Factors

- Counties with drinking water violations are those which received safe drinking water violations (a few different tests)
- Air pollution levels measured by EPA
- Drinking water violations are prevalent in the southwest, west, and northeast
- Air pollution is most prevalent in the rust belt, southeast, and California
- Very clear region with clean air between Midwest and West
- Very different geographic distributions of environmental hazards

3 Structural Factors

- Primary care physicians are more readily available in the western US and in the northeast
- The adult uninsured rate varies greatly from state to state, probably due to differing state policies
- Note the magnitude of difference between different bins: there are substantial differences between different counties





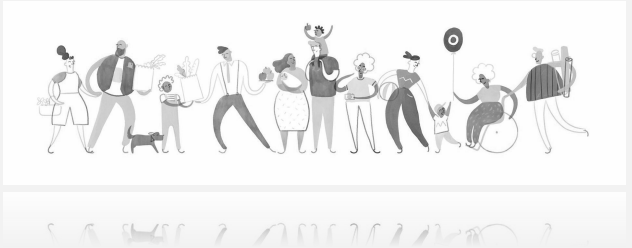
Spatial Autocorrelation and Geographically Weighted Regression (GWR)

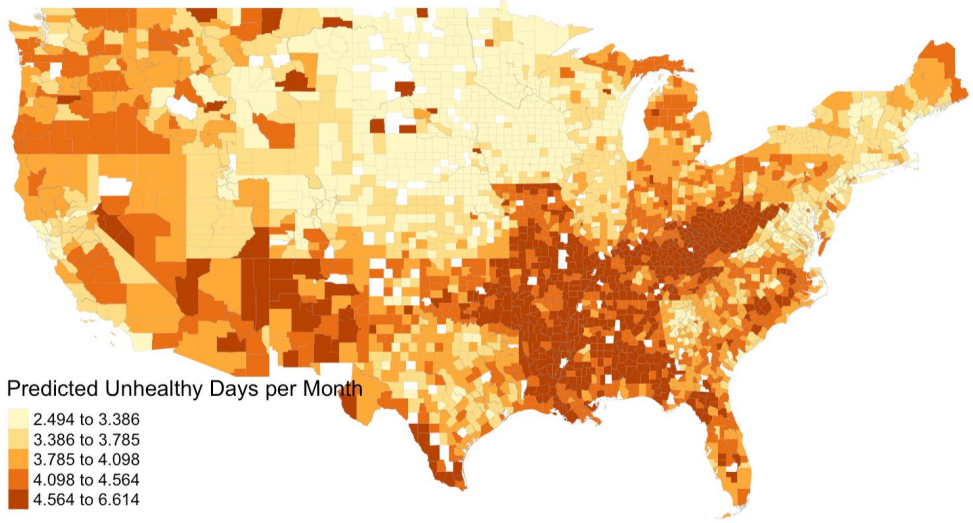
Linear Regression

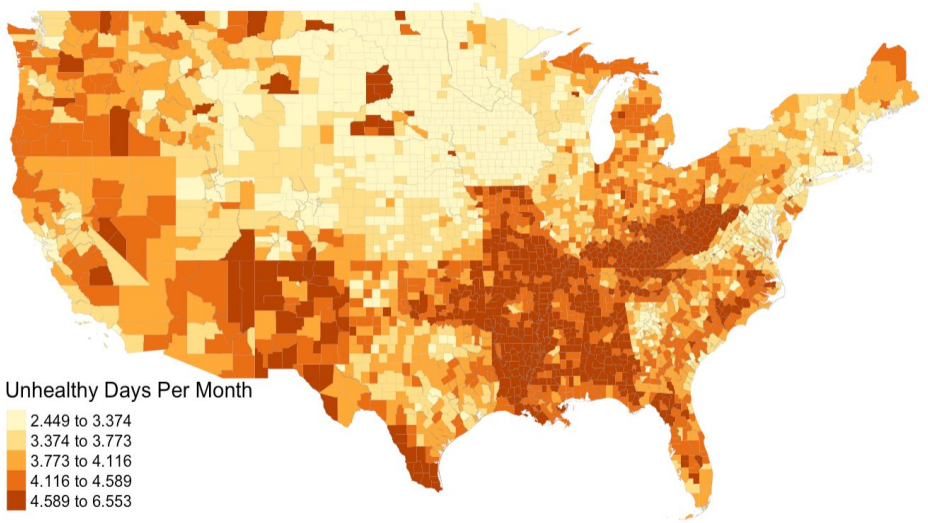
Models the relationships between variables, predicting outcomes of a response variable using multiple explanatory variables. Does not account for space.

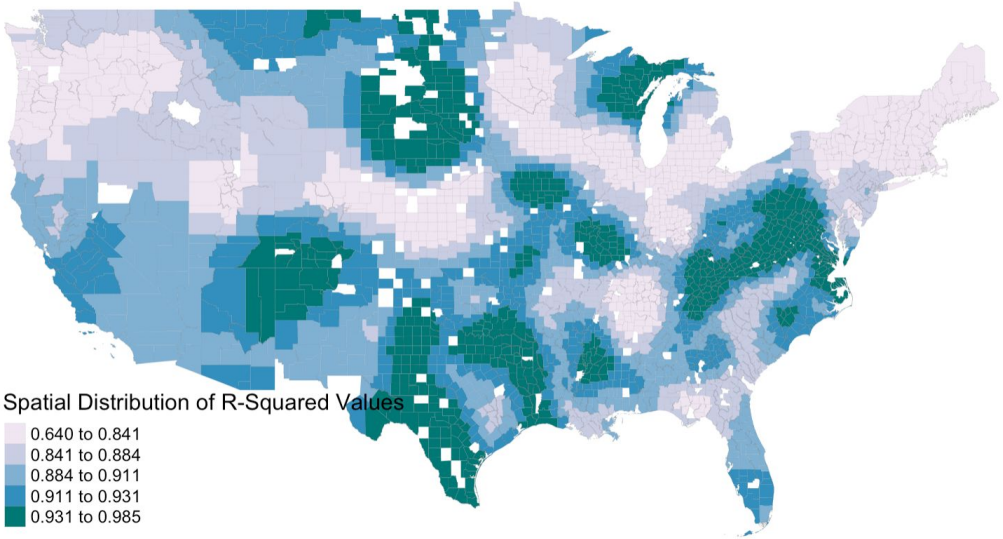
Geographically Weighted Regression

Considers spatial autocorrelation: that nearby observations are more closely related than faraway observations. Essentially performs linear regression for every observation, weighting nearby observations higher than faraway ones.









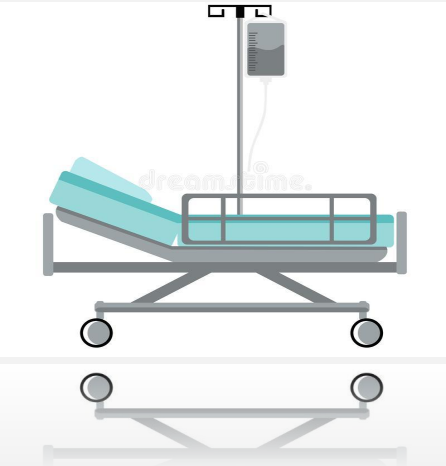
- Successfully predicts unhealthy populations in the Rust Belt/Southeast area of the USA
- Fails to predict unhealthy populations on the west coast
- Our model found more promising relationships between the predictors and the response variable in the vicinity of WV and in the western US
- Comparing the results with those of linear regression. Mean R-Squared: 0.885 vs 0.766

Are health care resources evenly distributed across space?

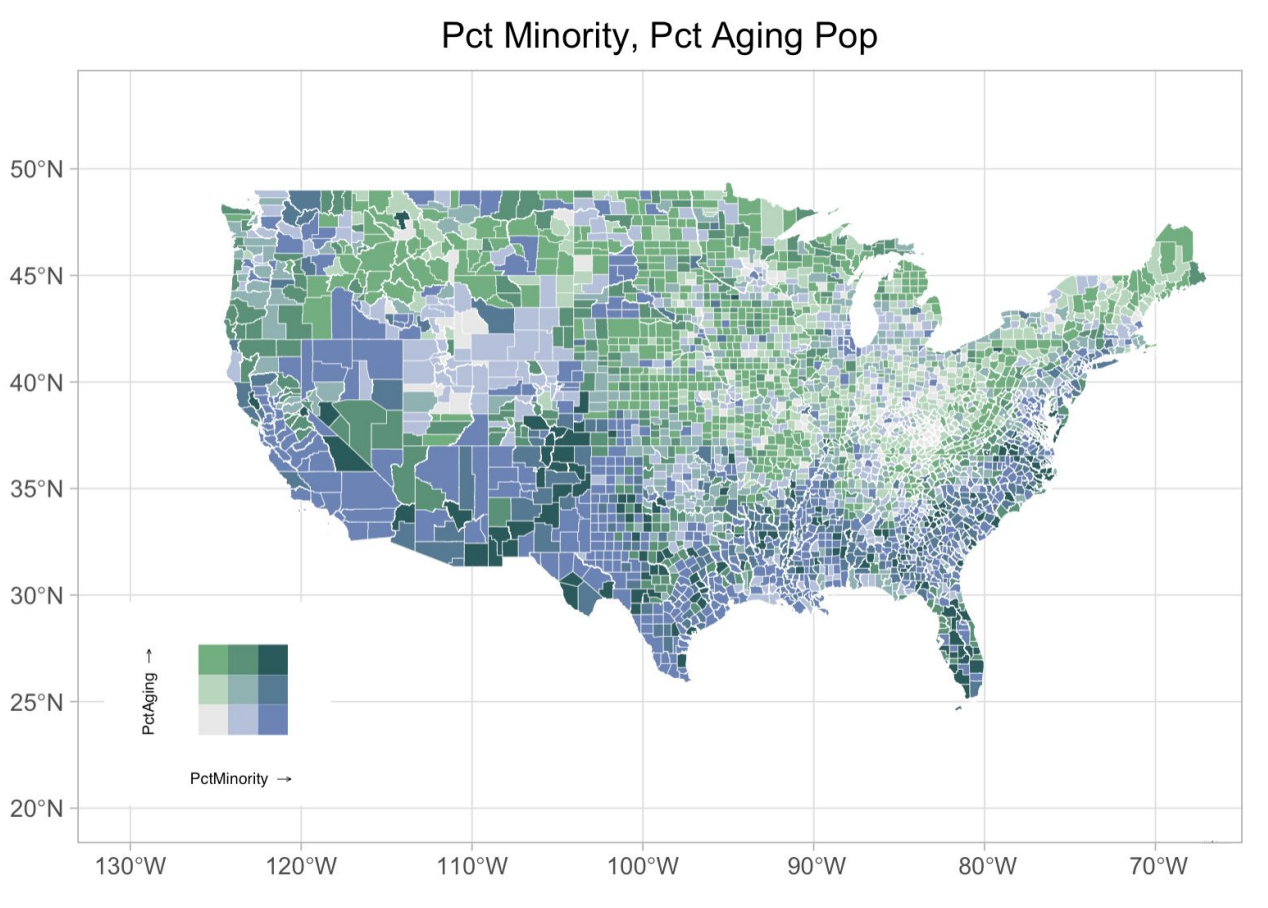
How accessible are those resources to people who are more vulnerable to the pandemic, for example?

What could be the consequences of even/uneven access to healthcare resources?

How could local authorities make use of this information to make healthcare resources more accessible to the general public?



Hospital Beds Spatial Accessibility



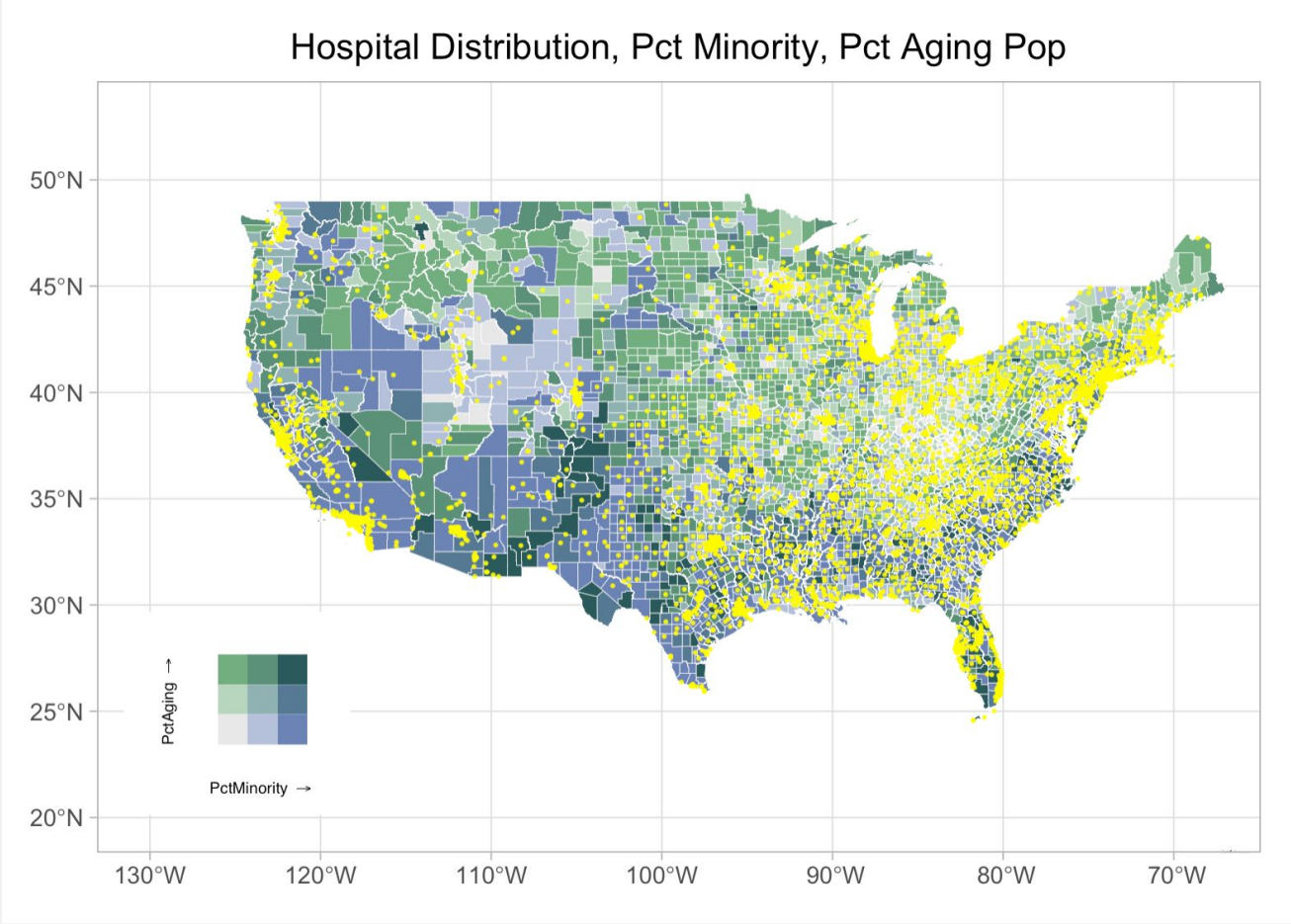
PctMinority: percentages of minority population in each county

PctAging: percentages of population over 70 in each county

Higher percentages of aging population in northeast, midwest.

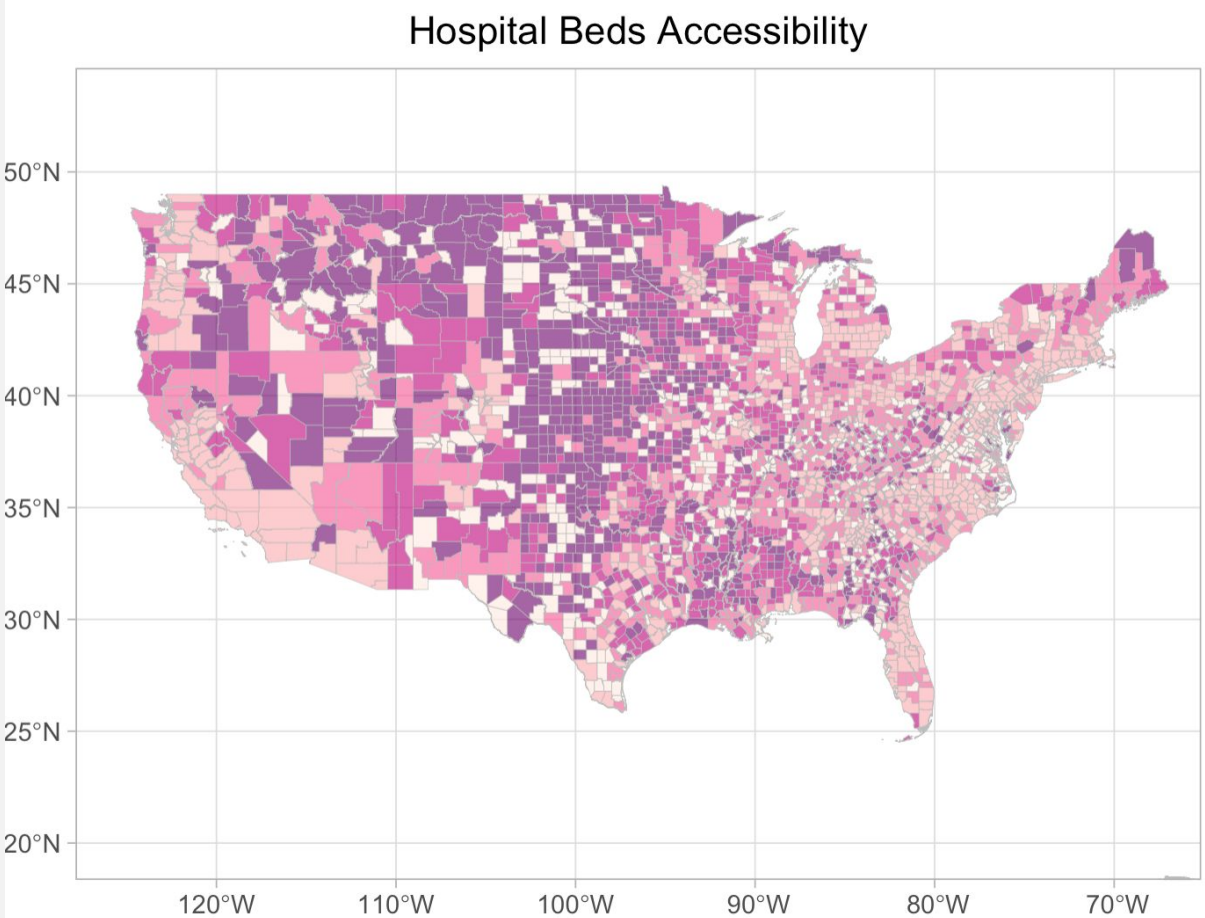
Higher percentages of minority populations in the south.

Many counties in the south contain more vulnerable populations: high percentages of aging and minority populations.



Hospitals are unevenly distributed across space, mostly concentrated along the east and west coast as well as in the south, which covers many counties that are designated as vulnerable.

Is mapping individual hospital point data helpful?



Number of hospital beds as a proxy for health care resources availability.

.

Normalize the data by:

- Spatially join counties point data into the geometry of each county (st_join) and summarize the total.
- Calculate the beds to population ratio for each county.

Now, we see that many counties that previously had a large number of hospital beds actually do not have that many hospital beds available for each 100,000 people.

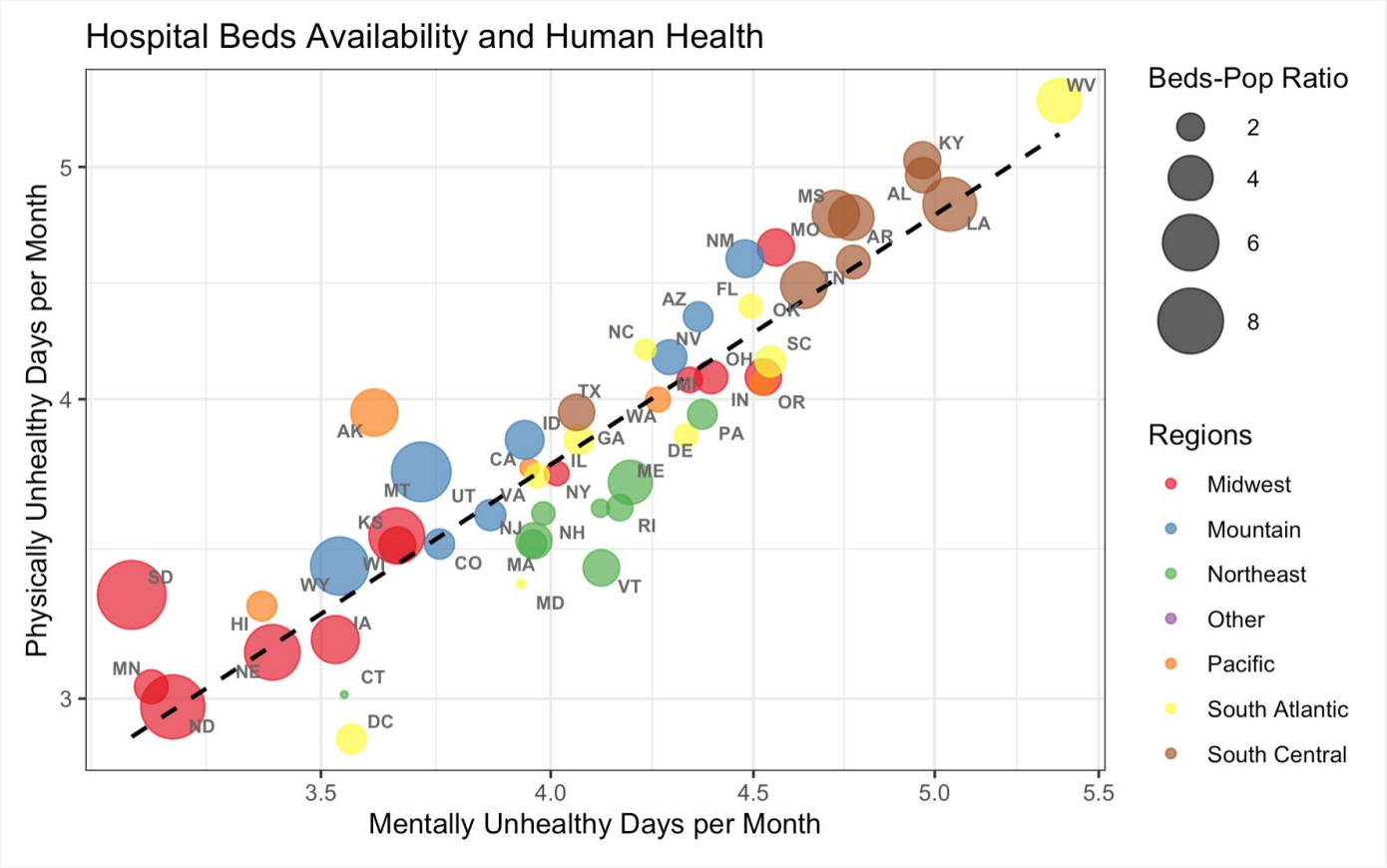
Size: beds to population ratio

Color: region

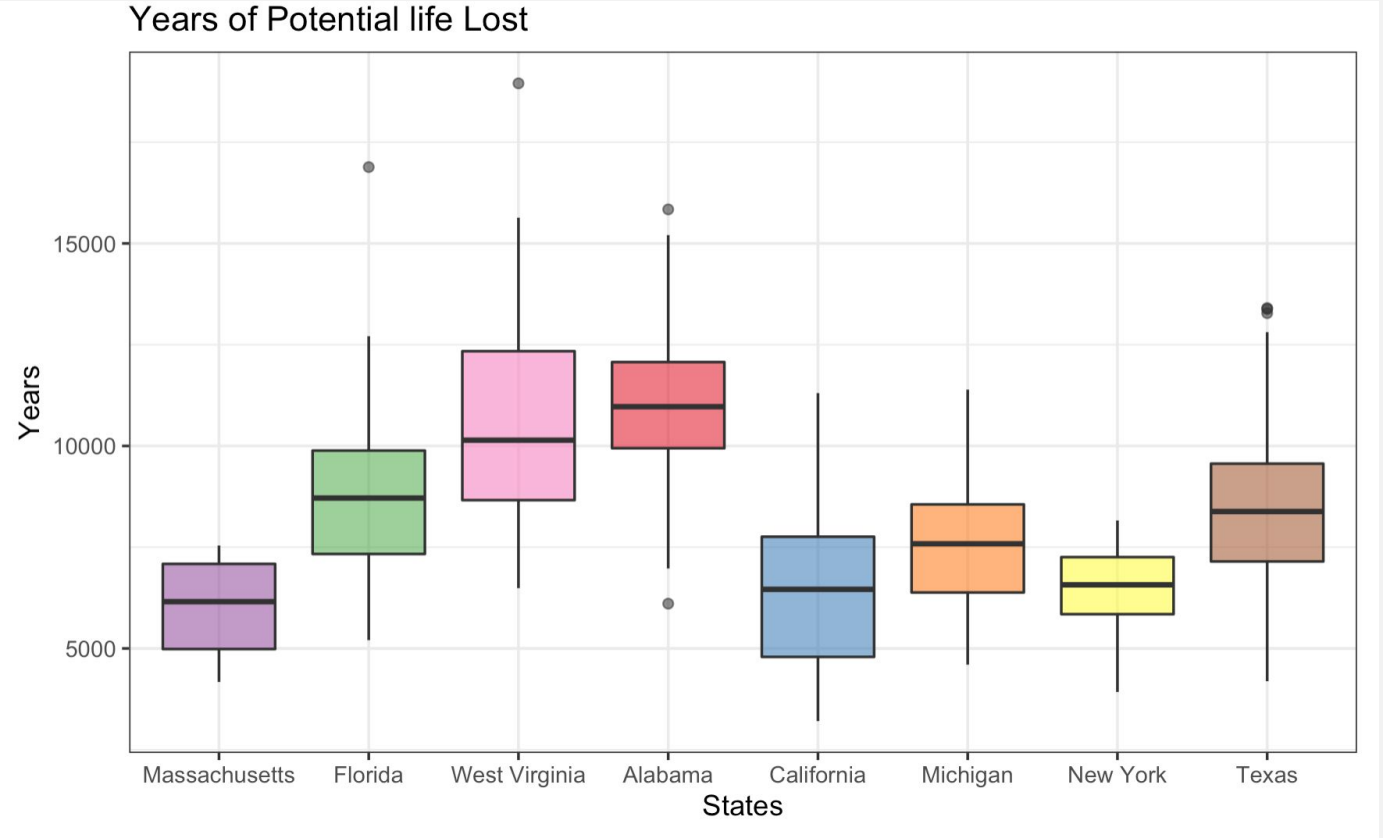
Axis: health indicators

The spatial distribution of hospital beds are on a national scale is just one way of visualizing spatial disparity.

Counties in the south are further along the healthy-unhealthy continuum, despite some of those states have a higher beds-pop ratio, which is in contrast to states in the Midwest



Zooming into several states of interests: given the demographic characteristics, general health conditions and hospital beds accessibility, let's also examine the years of potential life lost as an indicator of longevity of population in these states.



Results

Regression Analysis

A strong relationship between smokers and the number of unhealthy days per month is observed.

76.6% of the variation in number of unhealthy days per month could be explained by the variations of all the variables we ran in the model.

Spatial Distribution

Risk factors are more prevalent in the Rust Belt/Southeast.

Some factors vary greatly between states.

Spatial Autocorrelation

Geographically Weighted Regression produced higher R-Squared values, showing that more variability in our data can be explained when accounting for location.

This means that nearby health conditions are more closely related than faraway ones.

Resource Accessibility

Uneven distribution of hospital beds across space.

Minority population are particularly vulnerable for their inaccessibility to healthcare resources.

States with insufficient access to hospital beds have higher years of life lost.

Conclusions

Practiced cleaning/joining datasets from a variety of sources, some of them with spatial attributes.

Explored new R packages: sf, biscale, cowplot, spgwr, tmap

Tried different ways to effectively visualize our results: bubble plot, bivariate choropleth, tmap.

Explored new ways of modeling: Geographically Weighted Regression to account for spatial autocorrelation

And most importantly, applied the knowledge we learned in this class to our field of interest -> spatial analysis

Our Next Steps

Uncertainty Analysis: checking our results by assessing vulnerability with a different methodology.

Improving our model: Applying backwards stepwise regression to our linear regression model, perhaps looking for an analogous technique for Geographically Weighted Regression?

Examine how individual variables are related, zoom into specific states of interest, and include income as a factor in our analysis. If one could find the data, it would also be great to conduct the study on a census tract level in the future.