**Data Challenge Methods**

1. *Data Acquisition*

In pursuing our research question, we first created a list of priority data to obtain, which included demographics, unemployment data, health care coverage, hospital location and capacities, voting outcomes by county across the US, flight data, comorbidities, and air quality data. County level COVID - 19 cases and deaths, as well as county level shapefiles were provided by the course instructor. All other data was found via google search using keywords such as: “Voting”, “Unemployment”, “Health care”, “Transportation”, “State Funding”, “COVID-19 Impacts”, and “Flights”. Data that were reported at the county level were unified using FIPS codes. If the data did not have FIPS codes upon download, they were found and added to files. We created a data registry to keep track of all acquired data along with key characteristics of the data seen below.

1. *Data Pre-processing*

To make the data analysis ready, data were processed in R Studio (R Core Team, 2020) by choosing key variables from each dataset, converting it to “long” format, and merged all datasets with a cross-walk of state and county names and FIPS codes to ensure spatial consistency among variables for incorporation into models and visualizations. All variables were labeled as either predictor or outcome variables and all possible combinations of the two variable types were run through simple linear regression models to plot pairwise scatterplots. From this, correlations were explored to determine which variables would be of greatest significance to our logistic regression model and spatial autocorrelation analysis.  Predictor variables used for this pre-analysis were: Number of hospitals and beds, median AQI (air quality), community resilience, non-pharmaceutical response measures implemented by state governments, demographic data from the Census Bureau’s American Community Survey (ACS) about race, income, age, use of public transportation, and portion of the population with health insurance, comorbidity, and percent of votes earned by the democratic candidate by county in 2016. Outcome variables used in the simple linear regression were: change in unemployment at the state and county level, COVID- 19 reported cases per capita, and COVID-19 reported deaths per capita.

1. *Data analysis/model building and implementation*

Several preliminary linear models were run to test for differences between red and blue states in the 2016 election and COVID-19 cases, death counts per capita, and unemployment rate. Redness/blueness was assessed by the percentage of total votes for democratic candidates (Hillary Clinton) in a given county/state. However, these models were poor at describing relationships with the greatest adjusted R2 of 0.5972.

We next created correlation plots to assess the relationships between all dependent and independent variables at the county and state levels. At the county level, this included: air quality, community resilience, and hospital data (number of beds, etc.). At the state level, relevant predictor variables included: hospitals per state, number of governmental measures in response to COVID-19, median income, and fractions of the population who identify as black, pacific islander, “other” races, under 18, over 65, and those who are insured.  Several significant correlations were identified. Next, we constructed linear and/or generalized linear models to identify predictors of cases per capita, deaths per capita, and change in unemployment rate. The initial models considered all numeric and categorical variables as predictors. After running the first linear model, we removed all of the numeric predictors that were not significant at the 5% level. Variance inflation factors (VIFs) were determined to test for multicollinearity among predictors; variables with VIF values greater than 5 were removed from their associated model. AIC tests were applied to determine optimal predictors for each response variable, respectively. Individual predictor variables that were included in the optimal models that resulted from the AIC tests were then transformed by exploring typical transformations for each variable, i.e. log(variable), sqrt(variable), (variable^2), (1/variable), etc. Of these, the transformations that most significantly improved the fit of the model (if any) were kept. This led to an improvement of fit based on the plots (qqplot, histogram, residuals vs. fits, etc.) as well as the significance of each variable and adjusted R^2 value. However, these models did not account for spatial autocorrelation in the data and were thus not used as the final model.

Other preliminary models run were quasi-binomial logistic regressions, beta logistic regressions, and beta log-log regressions using the same variables that were included in the optimal linear model for each response variable. Each model was run thrice, once for cases per capita, deaths per capita, and change in unemployment rate, respectively. Of all approaches explored, the beta models outperformed all other models in nearly every case, but the R^2 was still extremely low.

The final model we used to answer our question was a random forest model. Random Forest Regression was used to identify features that contribute most to COVID-19 Case Counts, COVID-19 Death Counts, and Unemployment at county scale. We chose a random forest model because it is an ensemble model that is robust, and because there was a pretty large dimension to the data given. Initial analysis included Moran’s I spatial autocorrelation analysis, and in all three cases (COVID-19 Cases, COVID-19 Deaths, and Unemployment) spatial autocorrelation was significant. The model was trained using a 70/30 split and validated with the full data set.

1. *Data Visualization*

We used R to visualize our data and create figures that could be embedded on our website. Our code for these figures can be found in the Github repository in the master branch in the visualization folder. We created figures for both our input data (such as unemployment by state over time or COVID-19 rates at the county level) and our analysis output. We created non-interactive figures in R then exported to a *png* or *jpg* file. For interactive figures, we created them in R then exported them as html widgets which can also be embedded in our website.

We used github pages with Jekyll to summarize and present our findings and figures. We chose this medium due to its shareability and interactivity. Once we have finalized our analysis, the page can be made public and will succinctly display our analysis with links to our methods and github page. The page is located in the gh-pages branch of the main repository. The format for the page is a modified version of the Minimalist theme from Jekyll. These formatting modifications were made by Garret Miller (PhD Student in NC State’s Center for Geospatial Analytics, <https://gcmillar.github.io/>) and our group.

Table Data registry of all datasets collected

