Guided Capstone Project Proposal

I plan to analyze Covid-19 data from the CDC. There are approximately 30 million rows (individual level) with death, icu, and hospitalization variables (binary), demographic data such as age, sex, and race; geographical data (state and county) and underlying conditions, among other variables. Potentially identifying data have been removed and set to missing for those fields where the values narrow the number of identifiable possible individuals down to 10 or less, per census or publicly available data.

The primary idea is to set death, icu status, or hospitalization status as a target variable, and explore the significance of the other variables as explanatory or causal variables. If a variable would be more naturally conceived of as an outcome, I would be less inclined to include it in a model as an explanatory variable (i.e., correlation, not causation). As for geography, including hundreds of counties as categorical variable values is neither promising nor interesting, so we would consider bringing in a geographically linked variable such as median household income (at the county level) assuming such information exists in the public domain (census, wikipedia,etc).

The type of model could be logistic regression, and such approaches as random forest, neural network or decision trees could be considered.

The proposed data source can be found here:
https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4

Because of the size of the dataset, it is possible that google colaboratory will need to be used in order to speed up the handling of the data.