Random Forest:

- Query data to represent all available independent variables and dependent variables
  - Both party vote percentages were included as separate variables
- Join this data to the county shapefile
- Perform random forest with OOB errors
  - A trained model was tested
  - Tests of trained model had higher error than OOB
- Identify important features
- Test spatial autocorrelation of independent variables and residuals

The results of the random forest were very similar to the linear model, and capture 53% of the variation for cases and 41% for deaths. The exception is that all of the features were used, including republican and democratic vote percentages, and the most important feature was the percentage of population that is black. The other predictors, population that is white, can be considered because of a high reduction in MSE, but their node purity did not look very high in comparison to population that is black.

Spatial autocorrelation results showed that there is high spatial autocorrelation among counties for all three variables. The LISA cluster map is available in the Git Repository for comparison. The random forest regression did not account for spatial variance by performing spatial autocorrelation on the residuals.

The conclusion we can come to based on these models is that the data we had accessible was not enough to identify what causes Covid19's spatial variation. There is some evidence in the linear model that democratic counties are contributing to higher Covid cases, but this can also be because New York City is included in this analysis. As stated above, the next approach is to identify other features to include in the model, consider interactions, and begin to explore spatial statistical models.