COVID-19 Analysis

Sybile Cherenfant, Megan Cogguillo, Brandon Lee, and Vasudha Nair

Objectives

Our project analyzed county-level data from Johns Hopkins on COVID-19 cases and deaths, assessing the relationship between certain demographic characteristics and COVID outcomes.

In particular, we sought to understand:

1. Do counties with older populations have higher case counts and/or death rates?
2. Do counties with higher poverty rates have higher case counts / death rates?
3. Do counties with limited health insurance coverage have higher cases counts / death rates?
4. Is there a relationship between political affiliation of each county (based on 2016 presidential election) and case counts / death rates?

Age (65 years old and over) - Process

Jupyter Notebook Overview:

- Import modules
- Read into COVID DATA
- Read into census dataset, Population by Age Group

Data Cleaning of census data:

- Only use age groups covering 65 years and over for male and female
- Create new dataframe for Male and new dataframe for Female
- Merge and sum male and female population per county based on age group 65 yrs and over
- Create new column “count_name_long” in new dataframe to match with COVID DATA set

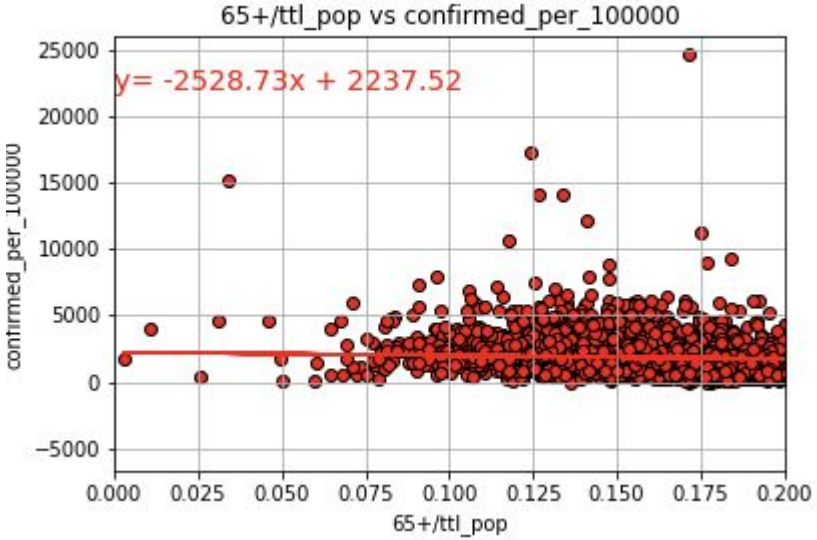
Merge:

- Merge Population cleaned up dataframe with COVID DATA

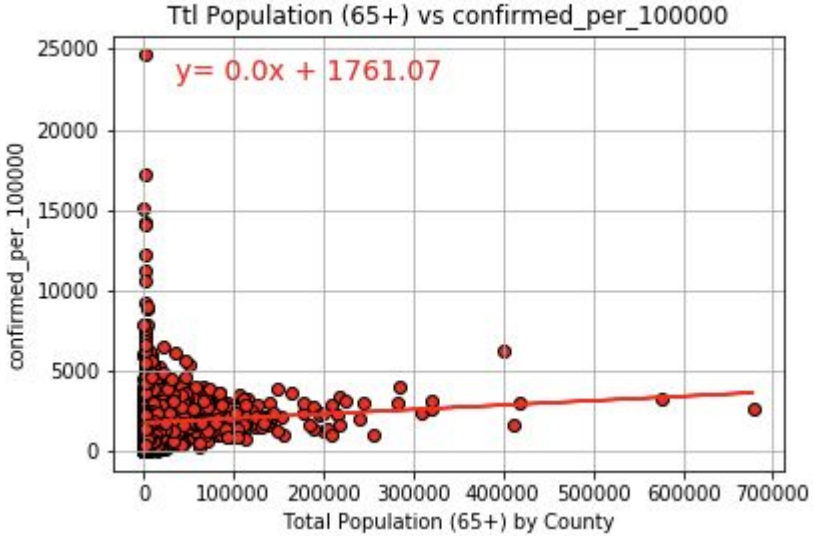
Scatter Plot with Regression Analysis:

- Plot to compare each county’s % of population 65+ to confirmed cases per 100,000 and confirmed deaths per 100,000 to see if counties with a higher % population of 65+ also had higher cases and deaths.
- Plot to compare each county’s total population 65+ to confirmed cases per 100,000 and confirmed deaths per 100,000 to see if counties with a higher % population of 65+ also had higher cases and deaths.

Age (65 years old and over) - Analysis



The r-value is -0.13303550281453858
The p-value is 2.7692254841305985e-13

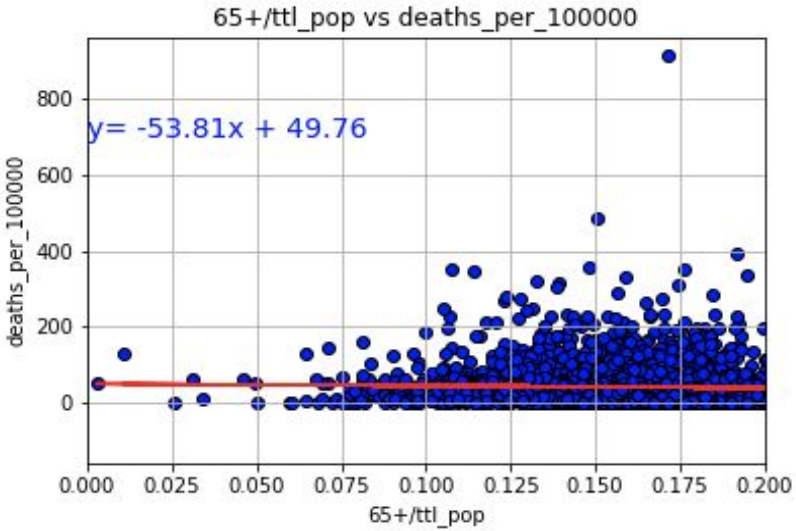


The r-value is 0.0651089821133206
The p-value is 0.00036649389292917143

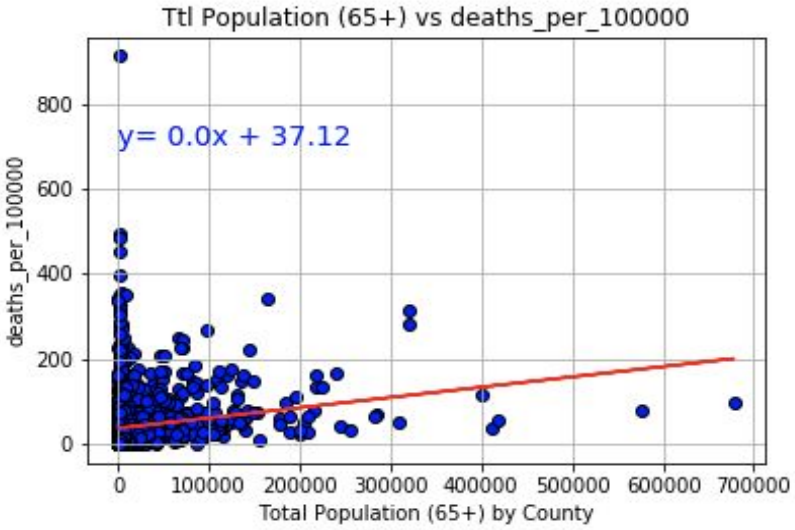
- Small r-value implies weak match to regression line
- Smaller p-value implies should reject null hypothesis

- Smaller r-value implies weaker match to regression line
- Small p-value implies should reject null hypothesis

Age (65 years old and over) - Analysis



The r-value is -0.07694155220794566
The p-value is 2.5264621638158352e-05



The r-value is 0.1553551224469933
The p-value is 1.2881236897800433e-17

- Smaller r-value implies weaker match to regression line
- Small p-value implies should reject null hypothesis

- Small r-value implies weak match to regression line
- Smaller p-value implies should reject null hypothesis

Age (65 years old and over) - Conclusion

Null Hypothesis:
There is no relationship between the 65 and over age group and COVID cases and COVID deaths.

Alternate Hypothesis:
The 65 year old and over age category has a correlated effect on confirmed COVID cases and COVID deaths.

Observations:
Percentage Population = percent of total population that is 65 years or older in each county
Total Population = total population by county

- In all cases, extremely low P-value suggests rejecting null hypothesis. Therefore we can say, age has a role in cases and deaths, but based on plots, age is not a significant influence.
- In all cases, we have well below 0.5 R-value, suggesting data does not fit regression line, illustrating there are many more factors associated with rising case and death rates. Particularly human behavior which is inherently hard to fit in a linear regression model.

Conclusion:
As there are too many factors that may influence rising cases and deaths, it would be more meaningful if we compared a group of factors. Alternatively, if we remained on age study, we could compare each age group to see which group was most affected by covid instead of the other way around.

Poverty Rates - Process

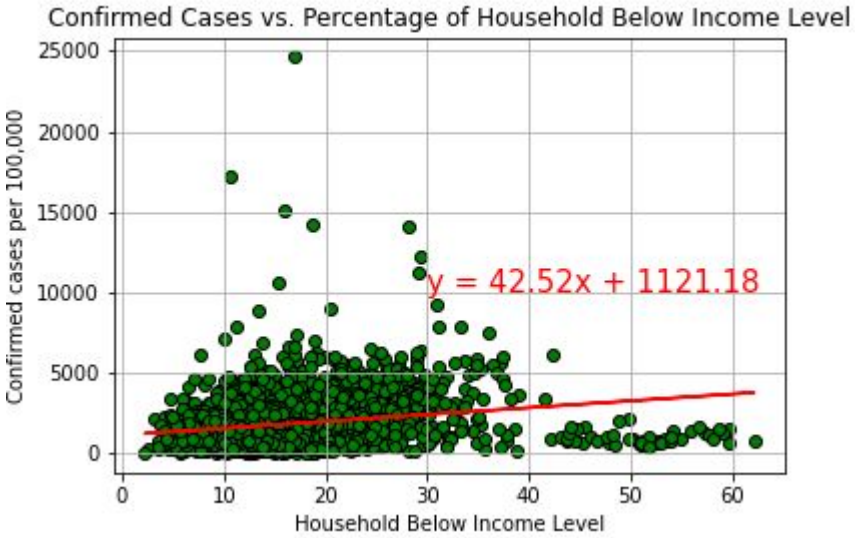
HYPOTHESES

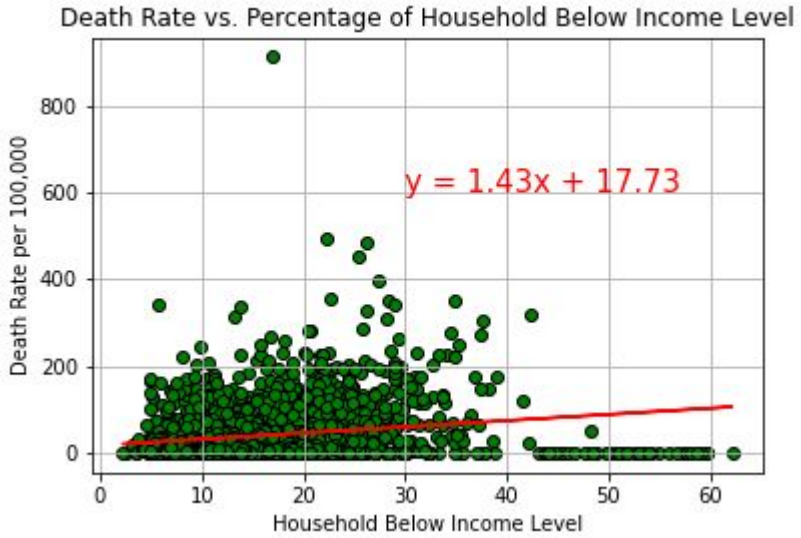
- **Null hypothesis** - There is no relationship between confirmed cases/death rates and poverty level
- **Alternative hypothesis** - the higher the poverty level, the more confirmed cases and death rates in the U.S.

PROCESS

- Jupyter Notebook
  - Dependencies (Pandas, OS, Matplotlib.pyplot, Numpy, Scipy.stats)
  - Files used (1_county_level_confirmed_cases.csv, Population_and_Poverty_-_Counties.csv)
  - Output file (Confirmed_cases-population_poverty.csv)
- Data Cleaning/Visualization
  - Restructured Population_and_Poverty_-_Counties.csv data
  - Selected necessary columns (county_name_long, HOUSEBELOWPOVP_CALC) before inner join merge
  - Scatter Plot & Linear Regression Model

Poverty Rates - Analysis





The p-value is: 1.1721639923375325e-33
The r-value is: 0.21922184609942164

The p-value is: 4.17425311193131e-28
The r-value is: 0.19971242720508164

Poverty Rates - Conclusion

Results

- P-value for both charts is < 0.05
  - There is a relationship between poverty level and rate of confirmed cases/ death
- R-value for both charts is about 0.2 (weak positive linear regression model)
  - Relationship between poverty level and rate of confirmed cases/ death is low

Conclusion

- Yes, there is a relationship between poverty level and rate of confirmed cases and death rate. However, poverty level alone does not impact them. there are other variables that play a role in determining the rate of confirmed cases and death.

Insurance Coverage - Process

HYPOTHESES:

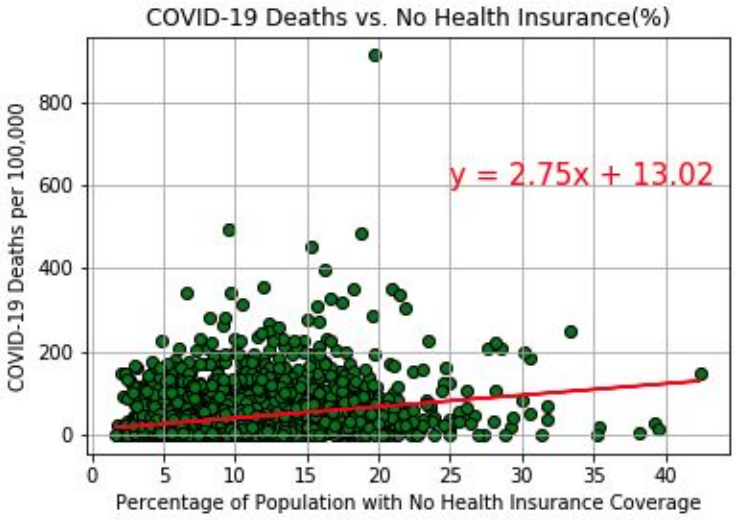
Null hypothesis：Population with no health insurance does not affect the case counts and/or death rates

Alternative hypothesis：More the percentage of population without health insurance, more the confirmed cases and/or death rates

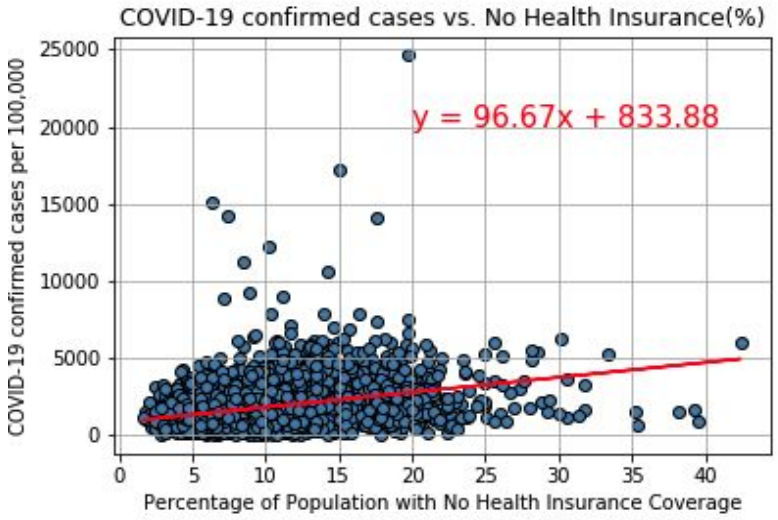
PROCESS:

- Data sets used for the study - John Hopkins data on COVID-19 cases and deaths/100,000 and the Health Insurance data from the Census dataset
- Cleaned up the data for both using the columns needed for analysis and merged them with an inner join. Dropna function was used to remove NaNs or missing rows/columns
- The scatter plots with linear regression analysis was carried out to see if there is any correlation between the COVID-19 confirmed cases/deaths per 100,000 and the percent of population with no health insurance

Insurance Coverage- Analysis



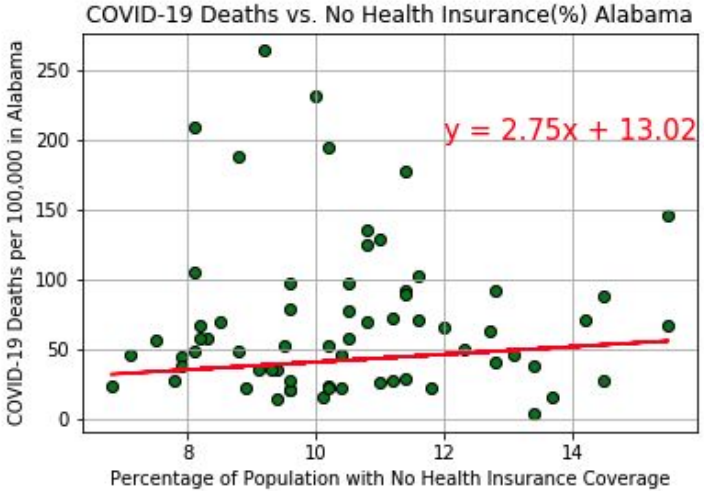
The p-value is: 9.82647770130023e-46
The r-value is: 0.2564281312395377



The p-value is: 3.081410482868751e-77
The r-value is: 0.3320297290567676

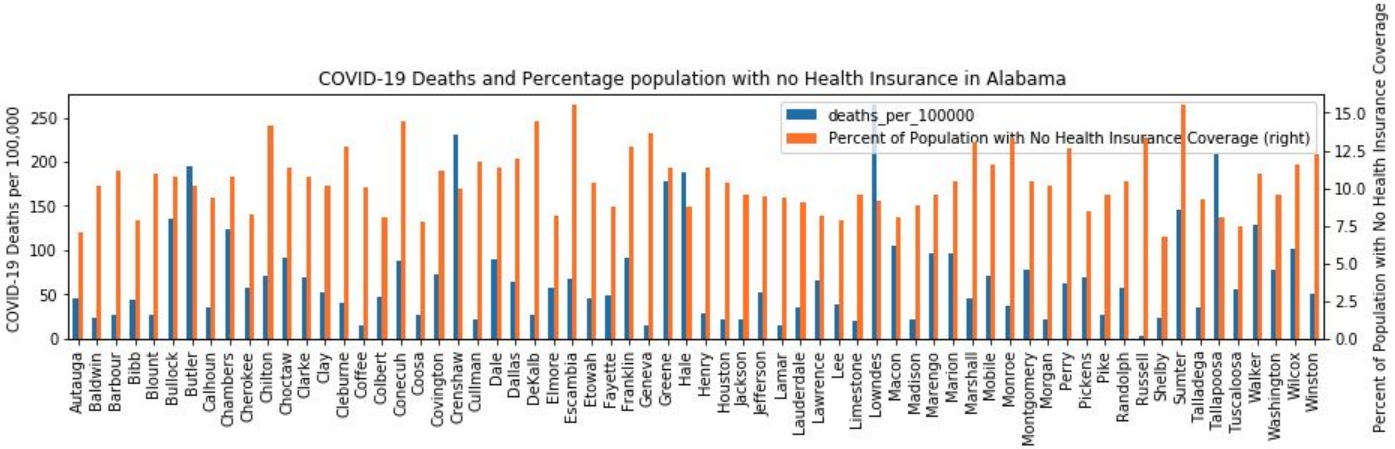
- Low p-value indicates that sample results are not consistent with a null hypothesis
- A small r-value indicates a weak correlation between the two. Here, r-value of 0.33 in case of COVID-19 confirmed cases has a slightly better correlation with percentage of population with no Health insurance coverage as compared to COVID-19 deaths(r-value of 0.26)

Insurance Coverage - Analysis

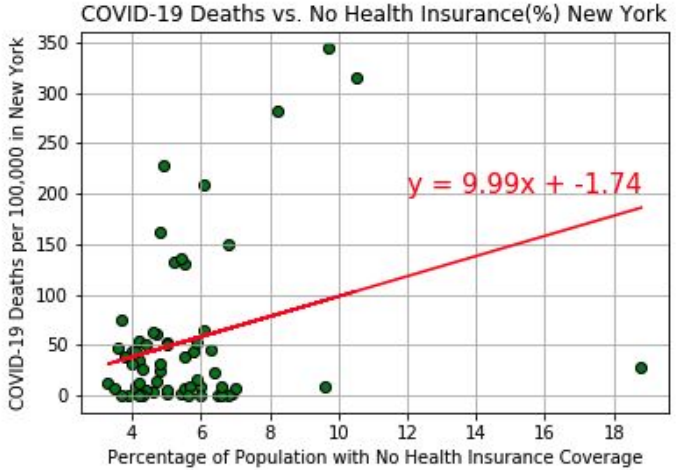


- Low p-value indicates that sample results are not consistent with a null hypothesis
- A low r-value shows a very weak correlation between the two

The p-value is: 9.82647770130023e-46
The r-value is: 0.2564281312395377

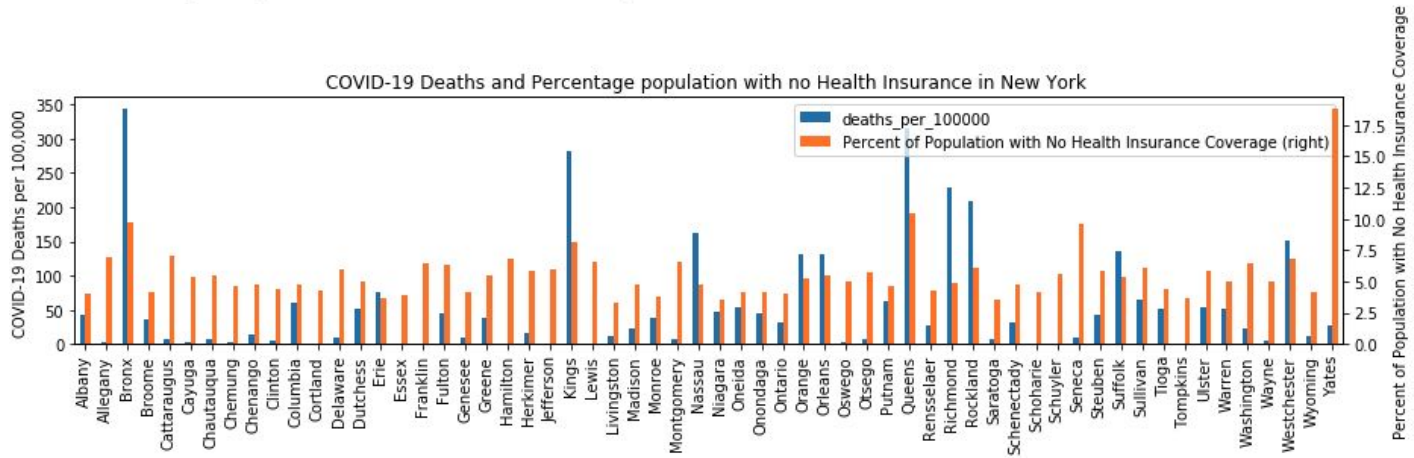


Insurance Coverage- Analysis



- Low p-value indicates that sample results are not consistent with a null hypothesis
- A low r-value shows a very weak correlation between the two data compared

The p-value is: 0.025348505561768778
The r-value is: 0.28857186730196116



Insurance Coverage- Conclusions

- Conclusions from the scatter plots and linear regression models indicate that COVID-19 deaths or confirmed cases in various counties are very weakly affected by the percentage of population without health insurance.
- The low p-values for all the plots implies that the null hypothesis (population without health insurance does not affect the case counts and/or death rates) should be rejected.
- The low r-values suggest that there is a weak correlation between population without health insurance and case counts and/or death rates.
- The values of r and p indicate that there might be other confounding factors that might contribute to a high positive correlation which these data sets were not able to help us analyse
- These data sets have sample data from a point in time rather than a time series(covering a few months)
- A much interesting data set would have been to look at the time series cases/deaths from the John Hopkins data. This would have helped us analyse the data for cases and deaths during the peak of the COVID-19 pandemic. There would be lot of factors affecting the infection rates and deaths that would be of statistical significance

Red Counties vs. Blue Counties - Process

HYPOTHESES

- Null hypothesis: there is no relationship between the margin by which Trump won/lost in a county and (1) case rates and (2) death rates
- Alternative hypothesis: “Redder” counties (based on margin of victory) have higher (1) case rates and (2) death rates than ”bluer” counties

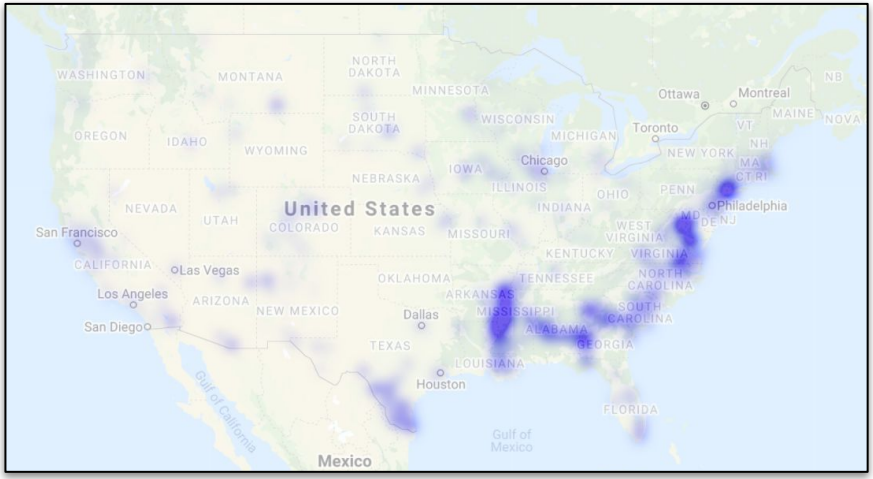
PROCESS

- Decided to categorize counties as “red” and “blue” based on who won in that county (Trump = red, Clinton = blue) in the 2016 presidential election
- Found MIT Election Lab dataset showing county level presidential results for 2000 - 2016
- Was only interested in most recent election, so limited dataset to 2016 only
- Cleaned / structured the data in such a way that it could be merged with the Johns Hopkins dataset on cases and deaths
- Created columns showing the winner in each county, as well as the margin by which Trump won or lost (margin is positive in counties where Trump won and negative in counties where Trump lost)
- Merged datasets and analyzed combined data to assess whether there is a correlation between Trump margin and cases/deaths per 100,000 residents

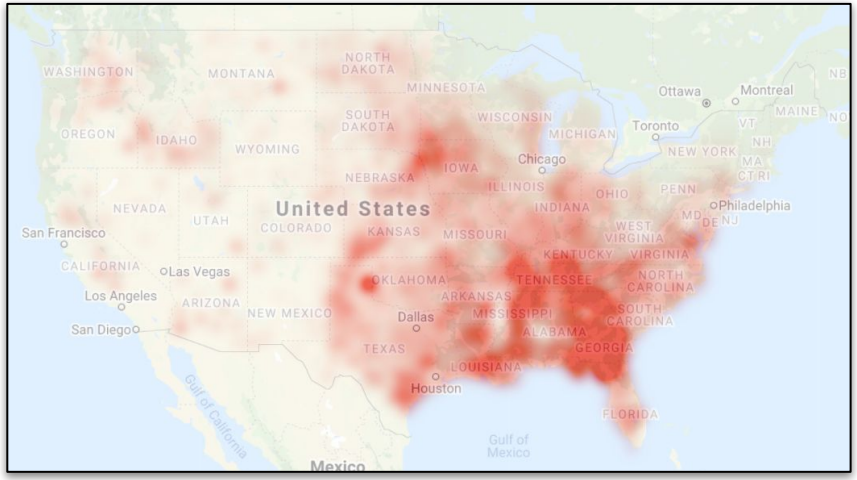
Red Counties vs. Blue Counties - Analysis

Creating a heatmap of the data (using the same intensity scale to ensure like-for-like comparison) initially seems to suggest that red counties have higher case rates than blue counties...

Blue Counties: Cases Per 100,000 Residents

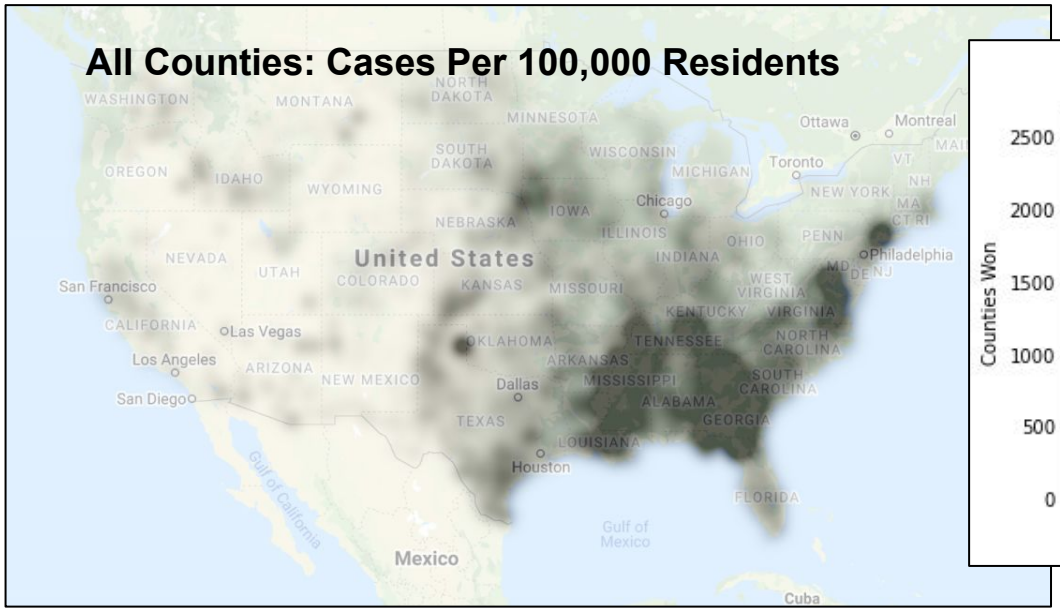


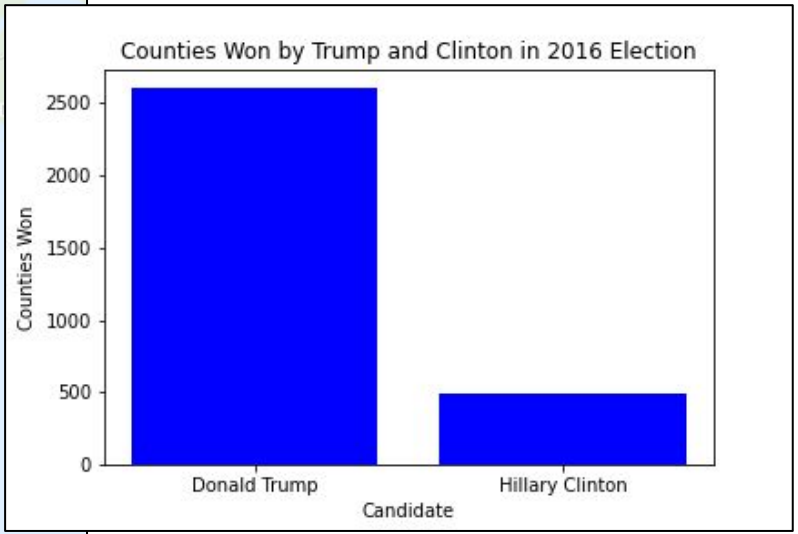
Red Counties: Cases Per 100,000 Residents



Red Counties vs. Blue Counties - Analysis

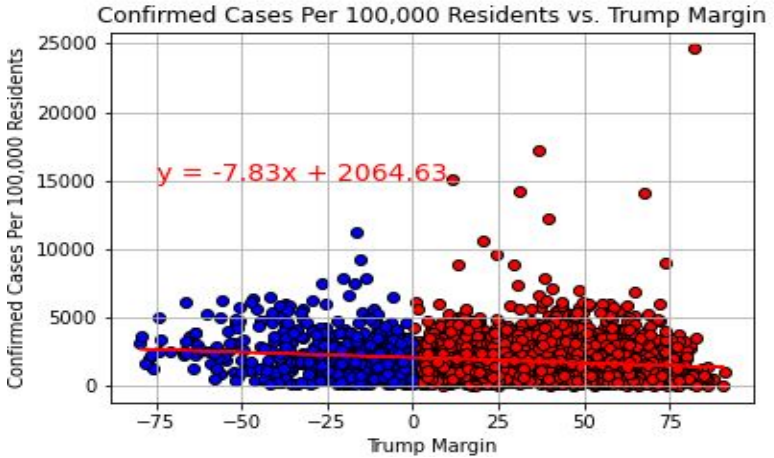
However, looking at the overall heatmap as well as data on the number of counties won by each candidate, it is possible that the heatmap is merely reflecting a higher number of red counties vs. blue counties overall...



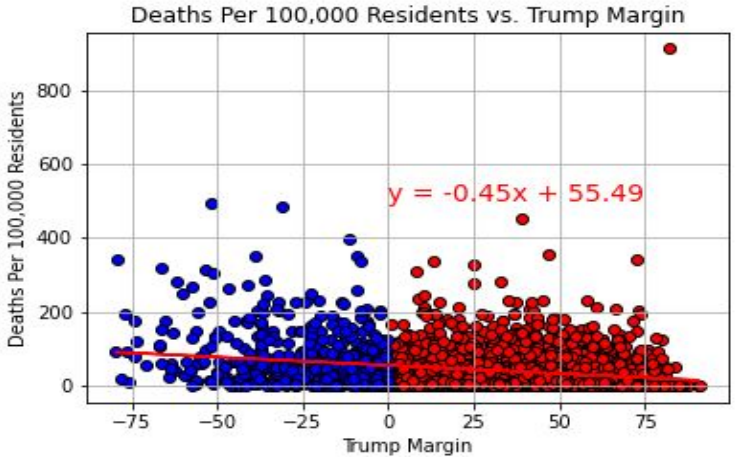


Red Counties vs. Blue Counties - Analysis

To further interrogate the data, we performed linear regressions of (1) cases per 100,000 residents and (2) deaths per 100,000 residents against the margin that Trump won/lost by. We see a statistically significant negative correlation between (1) Trump's margin and case rates and (2) Trump's margin and death rates...



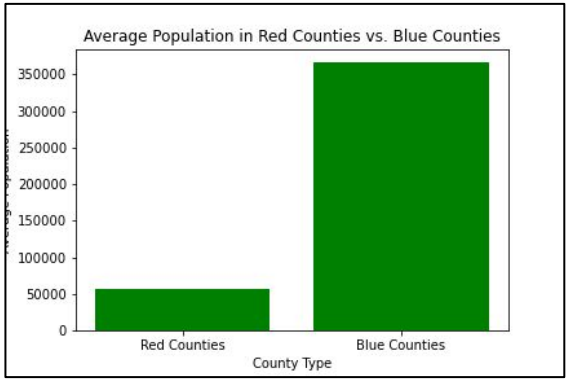
The r-value is -0.16458302217126783
The p-value is 2.6987853859125615e-20

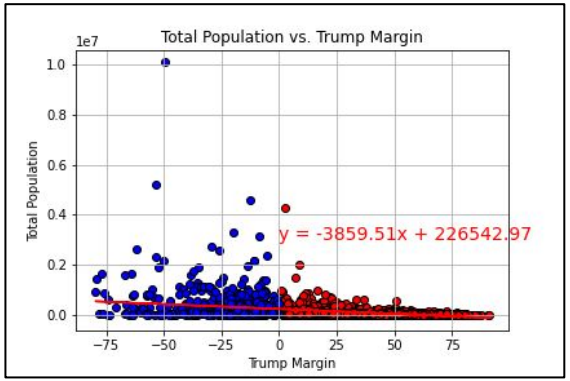


The r-value is -0.2551017821156189
The p-value is 2.5473821879783344e-47

Red Counties vs. Blue Counties - Conclusions

- While redder counties fared slightly better than bluer counties, there may be other factors driving this correlation. For instance, blue counties have higher average population size, and larger cities may have higher case rates and death rates than rural areas…
- Dataset shows cumulative cases and deaths at a point in time (data as of 9/23); looking at a time series dataset might produce more interesting results
    - News sources suggest that blue counties initially had higher case counts and death rates, but red counties have since caught up. Point in time data with cumulative totals might not be as valuable...
- Looking at red counties and blue counties does not give us information on actions each county has taken to address COVID, such as mask mandates, shutdown / reopening policies (restaurants, stores, schools, etc.), quarantine policies, and testing






```
The r-value is -0.3559666517082983
The p-value is 2.0384374429323915e-93
```

Overall Conclusions

- All the factors we examined had a statistically significant relationship with county COVID outcomes:
  - Percentage of population 65 and older
  - Percentage of population below the poverty level
  - Percentage of population without health insurance
  - Margin by which Trump won/lost in the county
- However, in all cases, the correlation was relatively weak. Intuitively, this makes sense: there are many factors that influence death rates and case counts, and no single factor is going to single-handedly determine the impact of COVID on a county
- Our analysis examined the relationship between demographic characteristics of each county and COVID outcomes. We did not have data on outcomes by group (i.e. our data does not actually tell us whether older people got infected more or died in larger numbers than younger people); examining this kind of data might tell us more about the relationship between the factors we looked at and COVID outcomes