New York COVID-19 Statistical Analysis

Marco Lopez

09/13/2023

Background

New York is a state in the northeastern United States located on the East coast. New York is the fourth most populous state in the U.S. New York ranks fourth in the U.S. in all-time COVID-19 cases, with over 6.7 million cases to date. New York also ranks fourth in the U.S. in all-time COVID-19 deaths, with over 77,000 deaths to date.

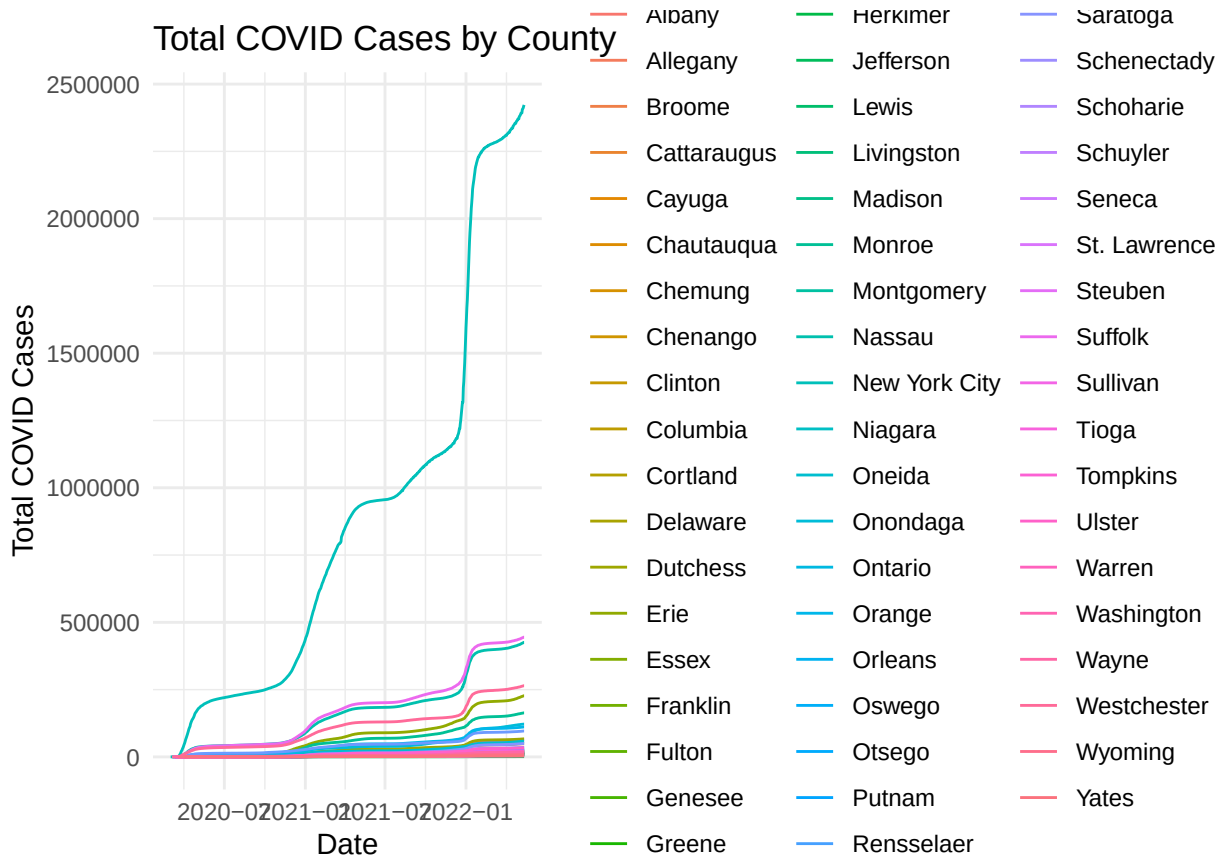
Overview and goals

This statistical research project will primarily focus on looking at the correlation, if any, between New York's COVID-19 data (such as cases and deaths) and various demographic data. I have assembled data regarding race, education levels, presidential voting, unemployment, median household income, and population. To provide statistical basis to these various correlations, I will employ statistical tools such as correlation coefficients, p-values, and regression models.

Note

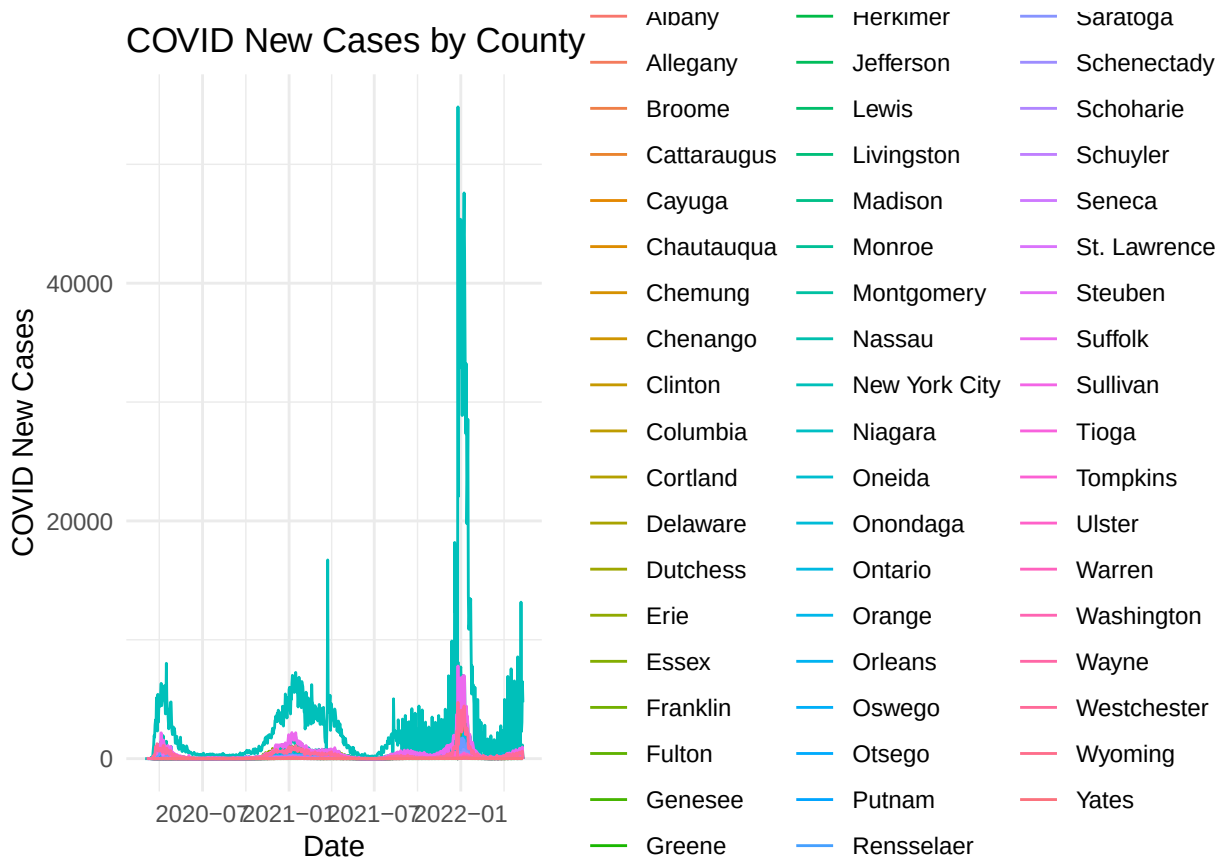
For this analysis, the "New York City" county consists of five different counties in New York (the five boroughs of New York City): Bronx county (The Bronx), Kings county (Brooklyn), New York county (Manhattan), Queens county (Queens), and Richmond county (Staten Island).

Progression of total COVID-19 cases by county



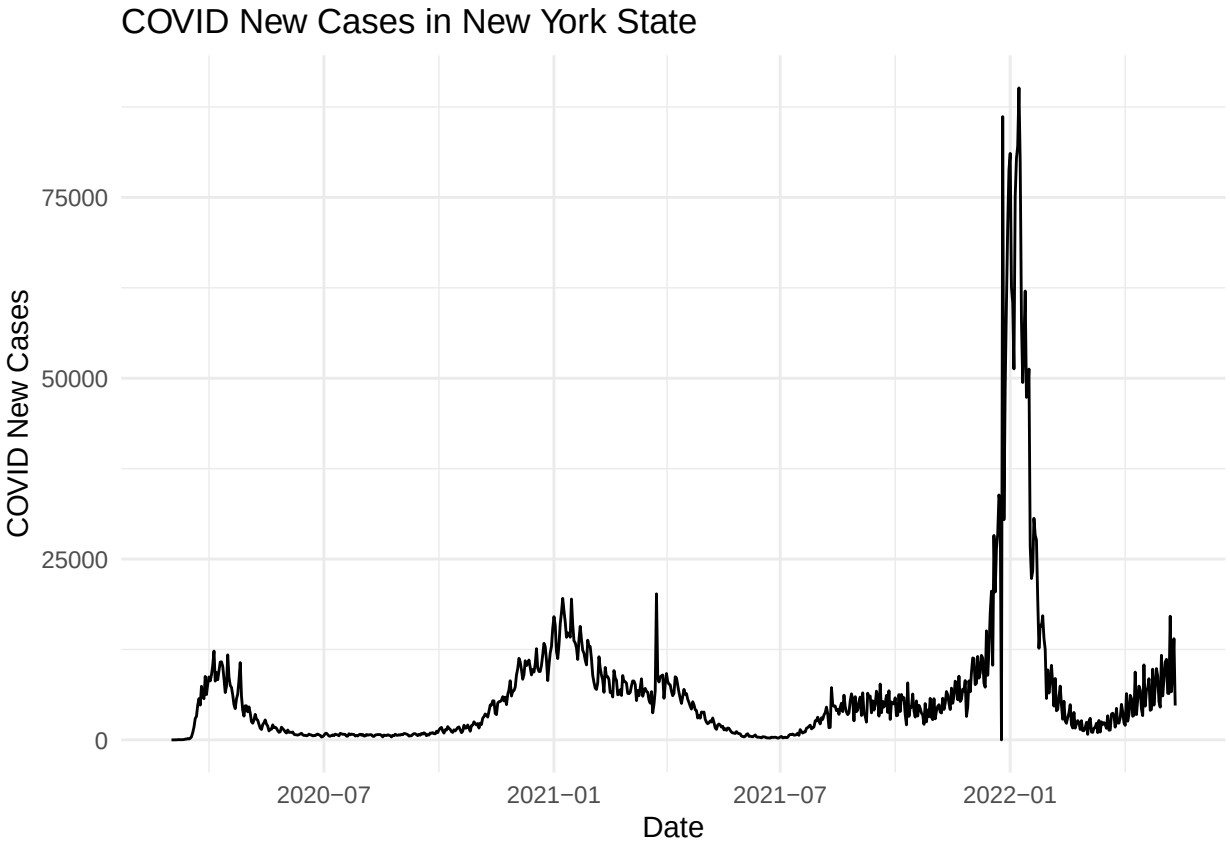
New York City, Suffolk, and Nassau counties were top three in COVID-19 cases.

Progression of COVID-19 new cases



New York City, Suffolk, and Nassau were also top three in new cases during various peaks of the pandemic.

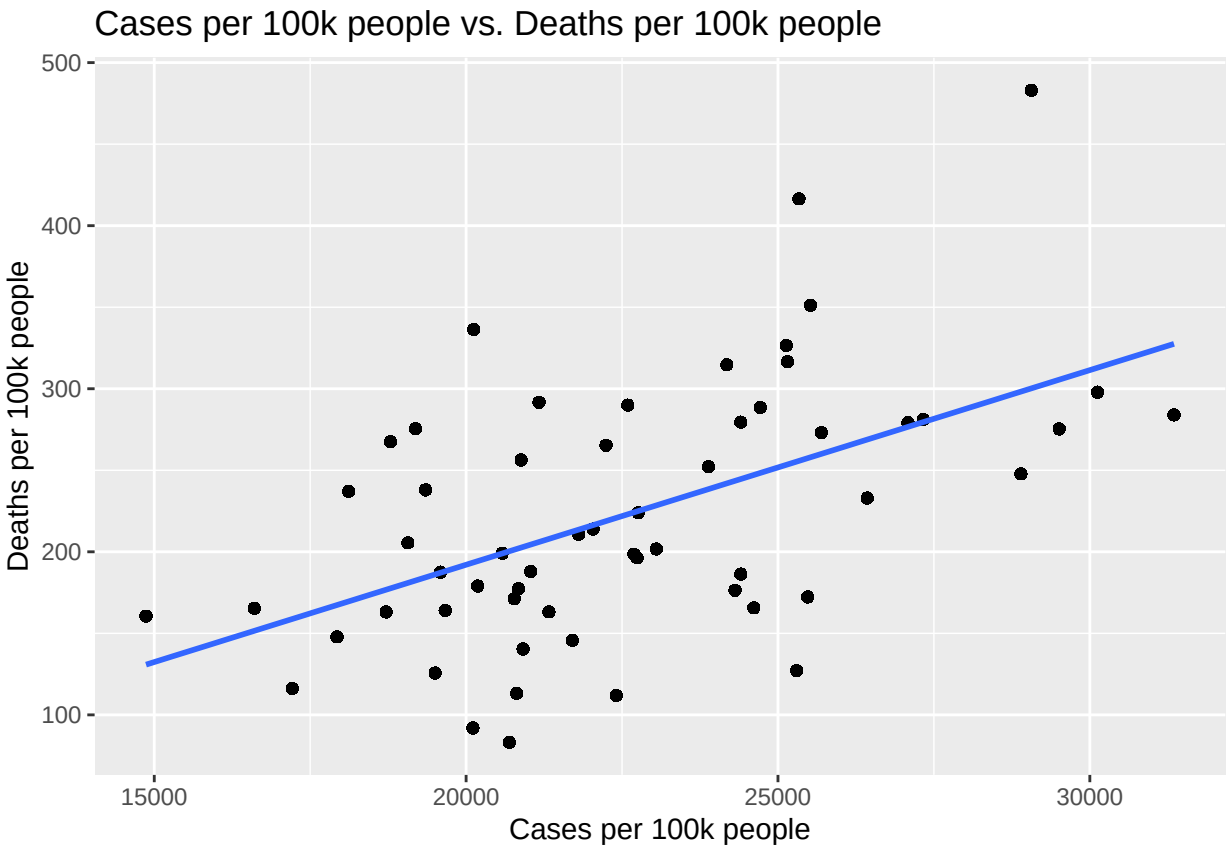
Progression of COVID-19 new cases in New York State



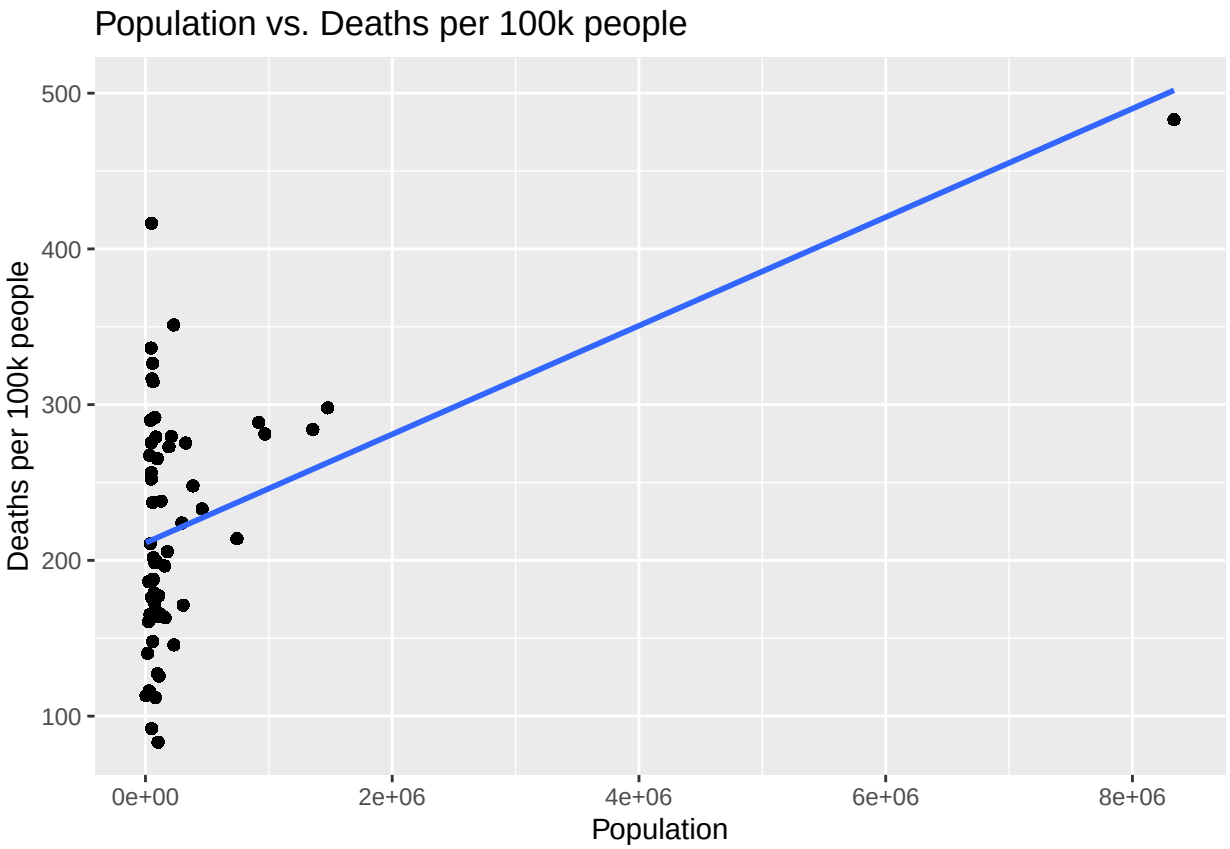
The above chart shows the overall progression of COVID cases in New York state for all counties.

Correlation charts for deaths per 100k

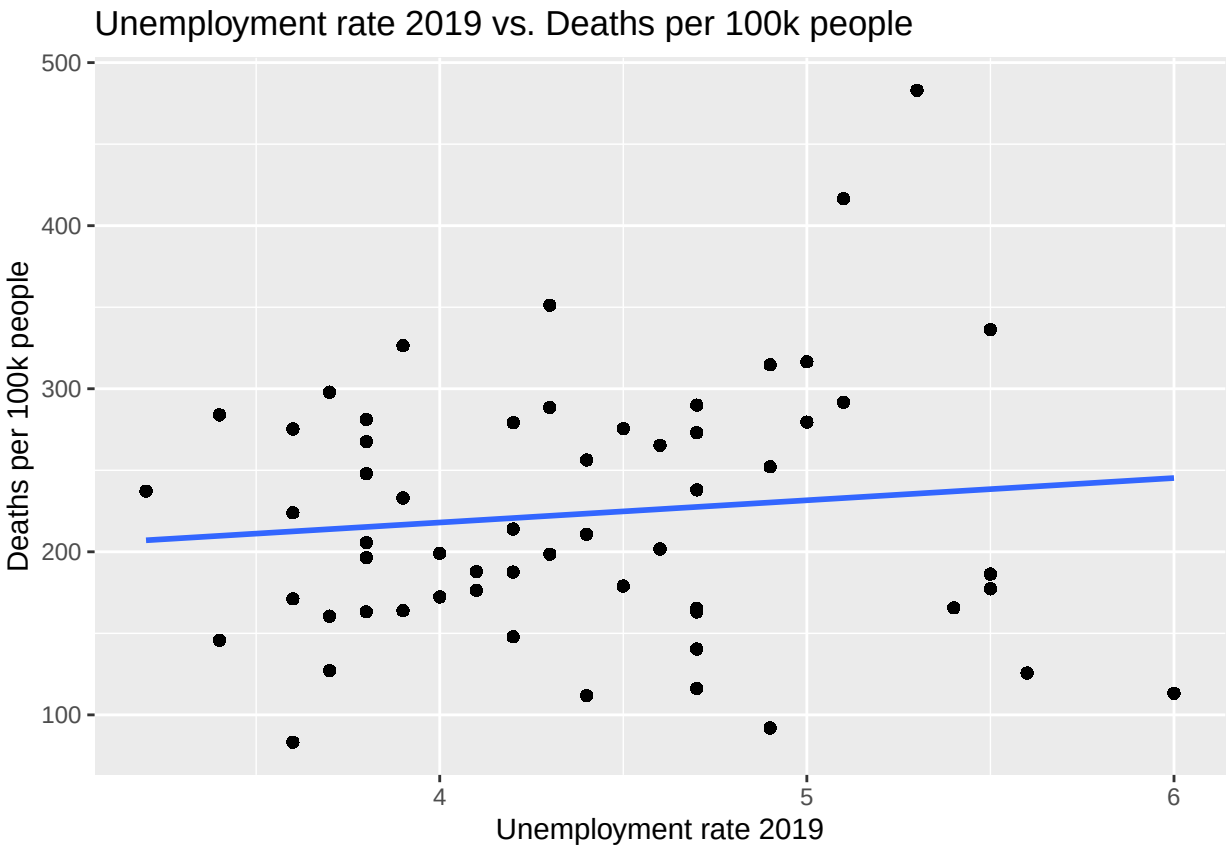
```
## `geom_smooth()` using formula = 'y ~ x'
```



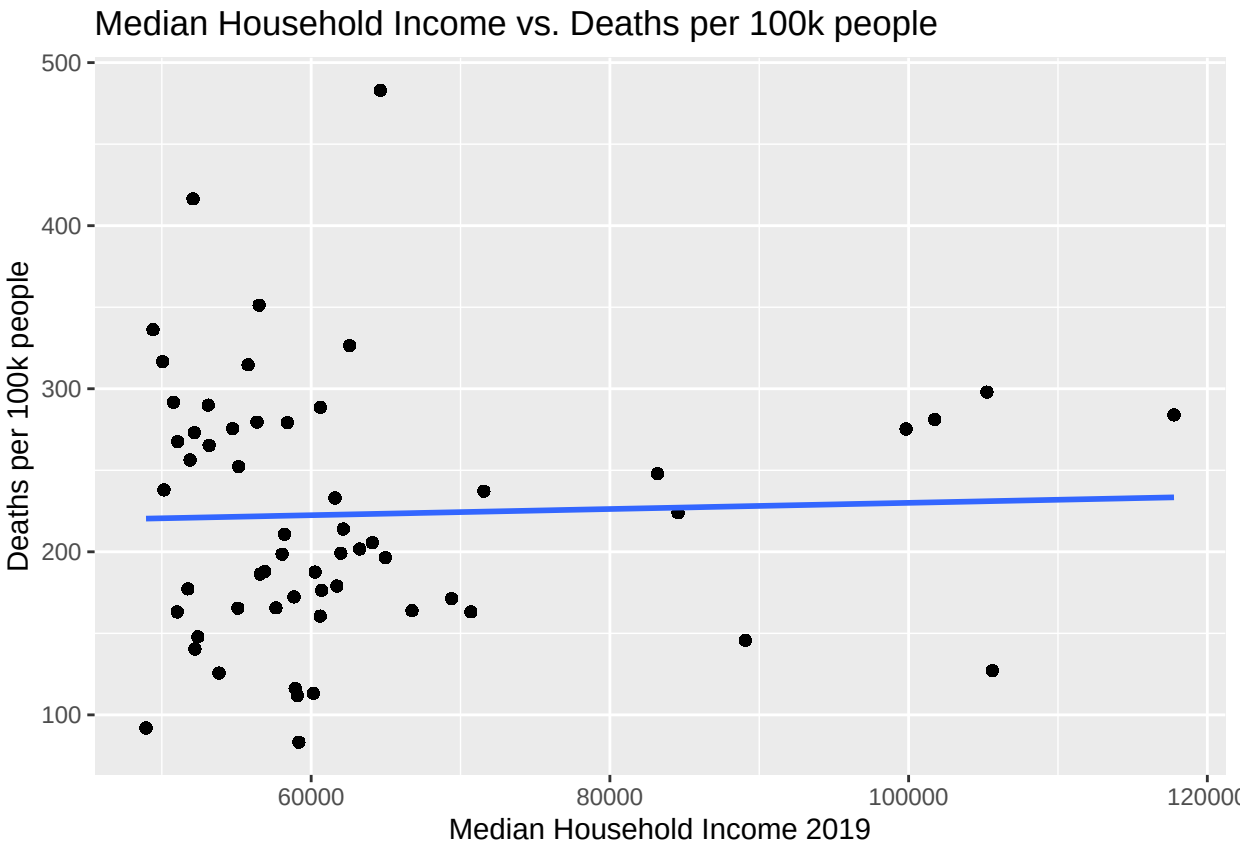

```
## `geom_smooth()` using formula = 'y ~ x'
```



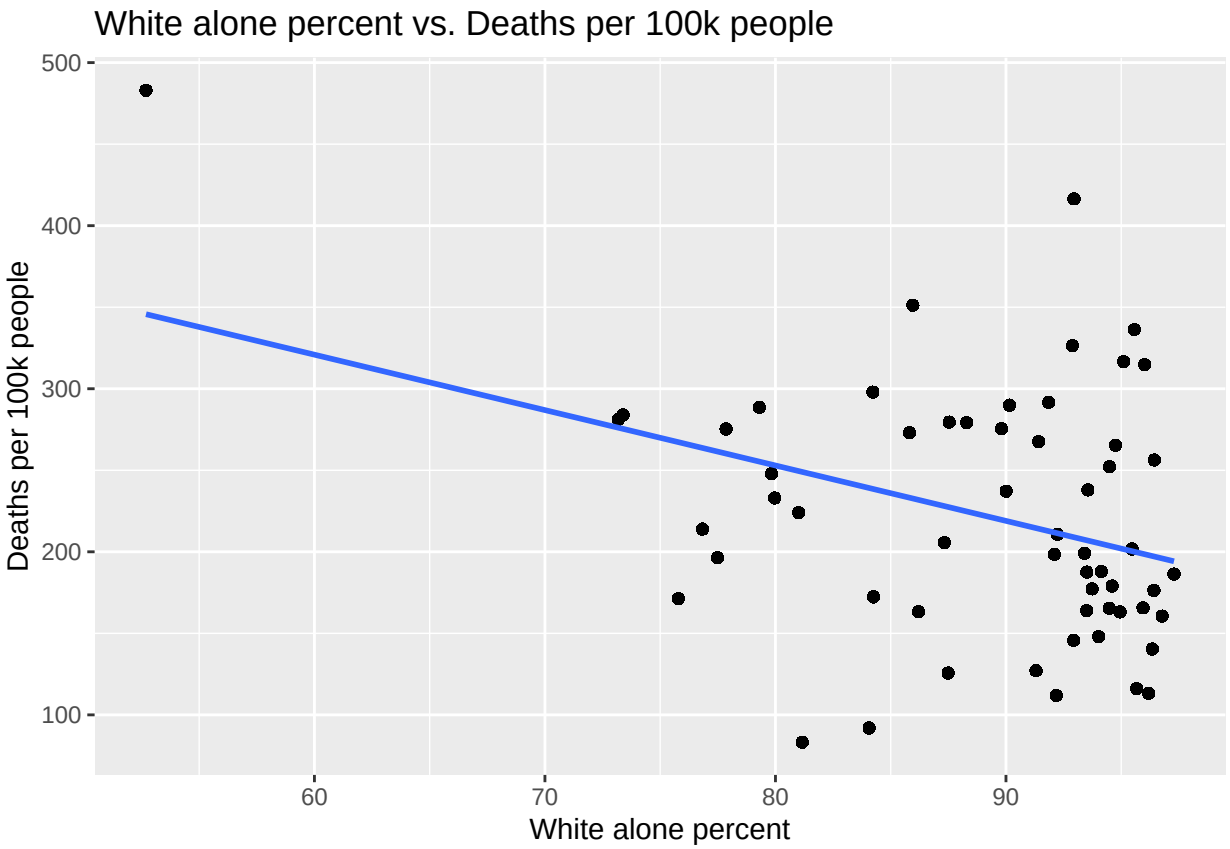

```
## `geom_smooth()` using formula = 'y ~ x'
```



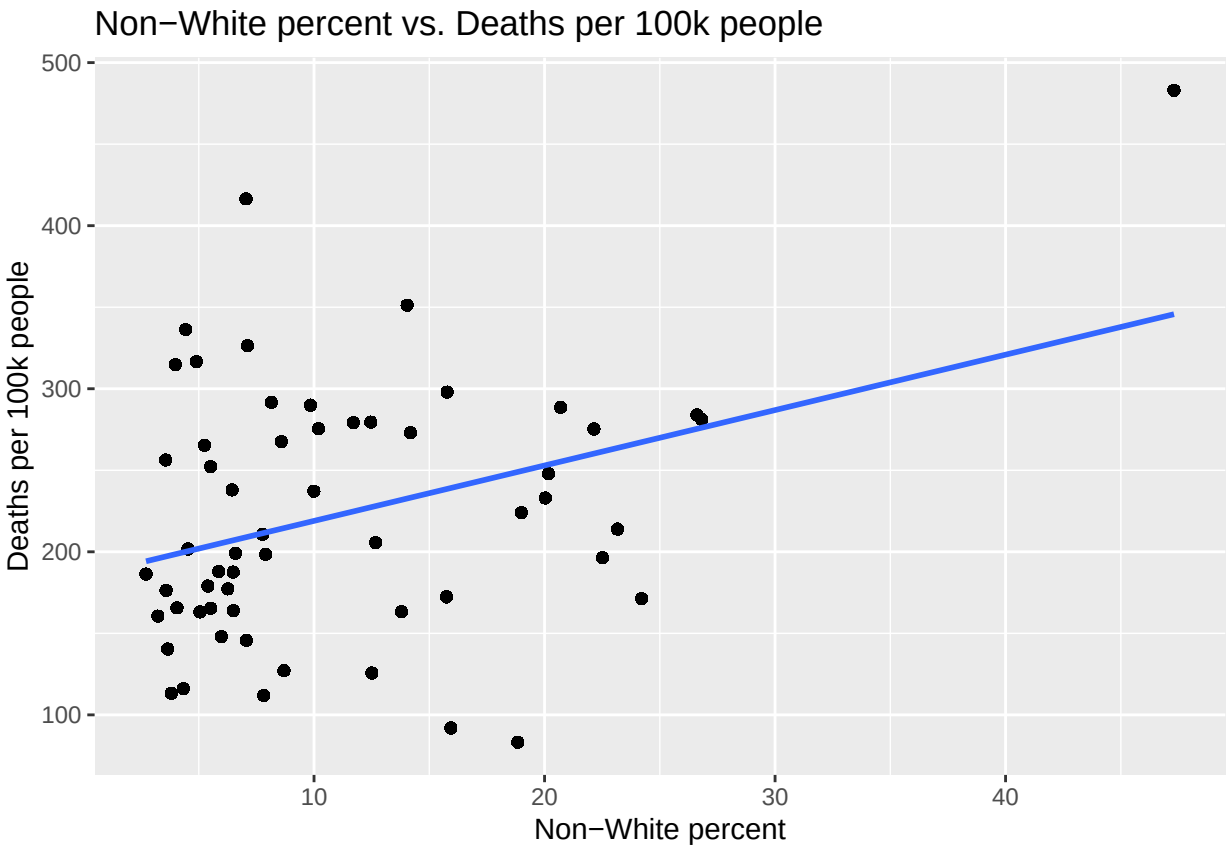

```
## `geom_smooth()` using formula = 'y ~ x'
```



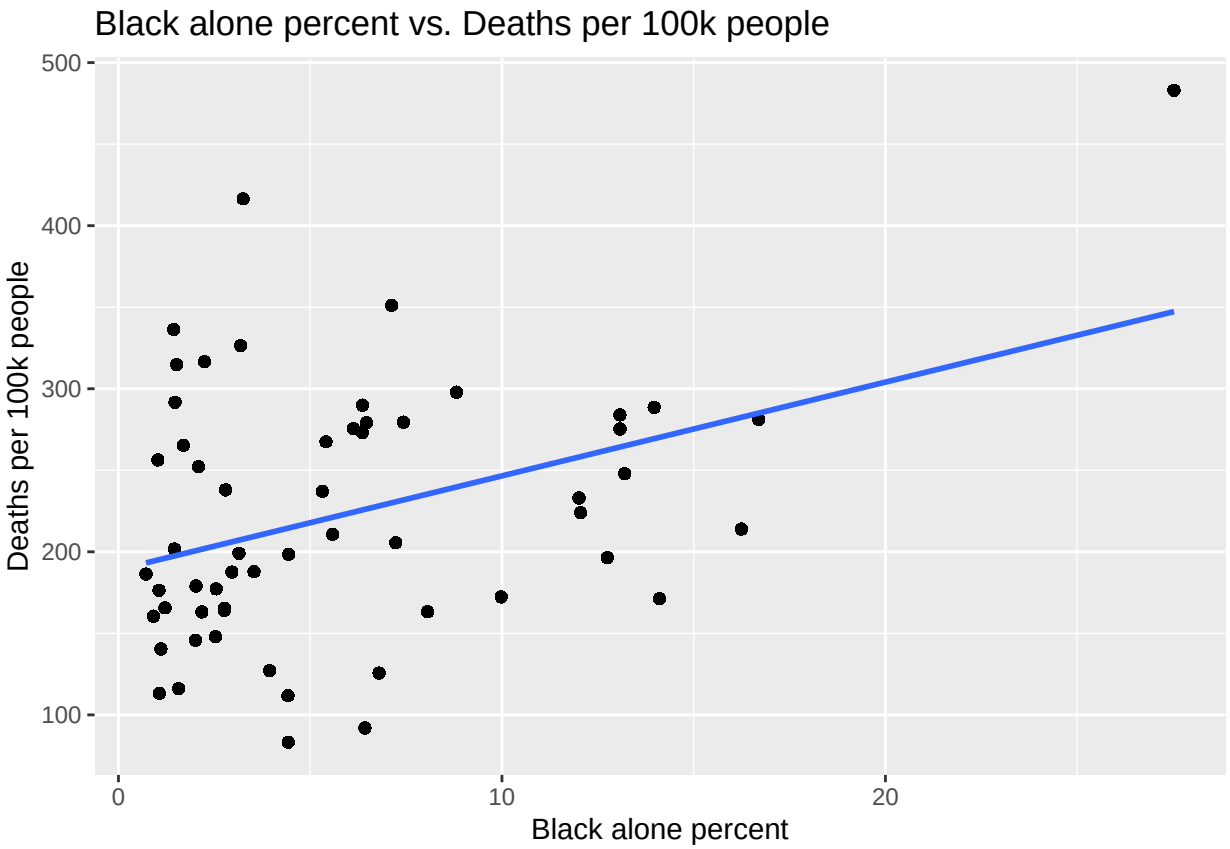

```
## `geom_smooth()` using formula = 'y ~ x'
```



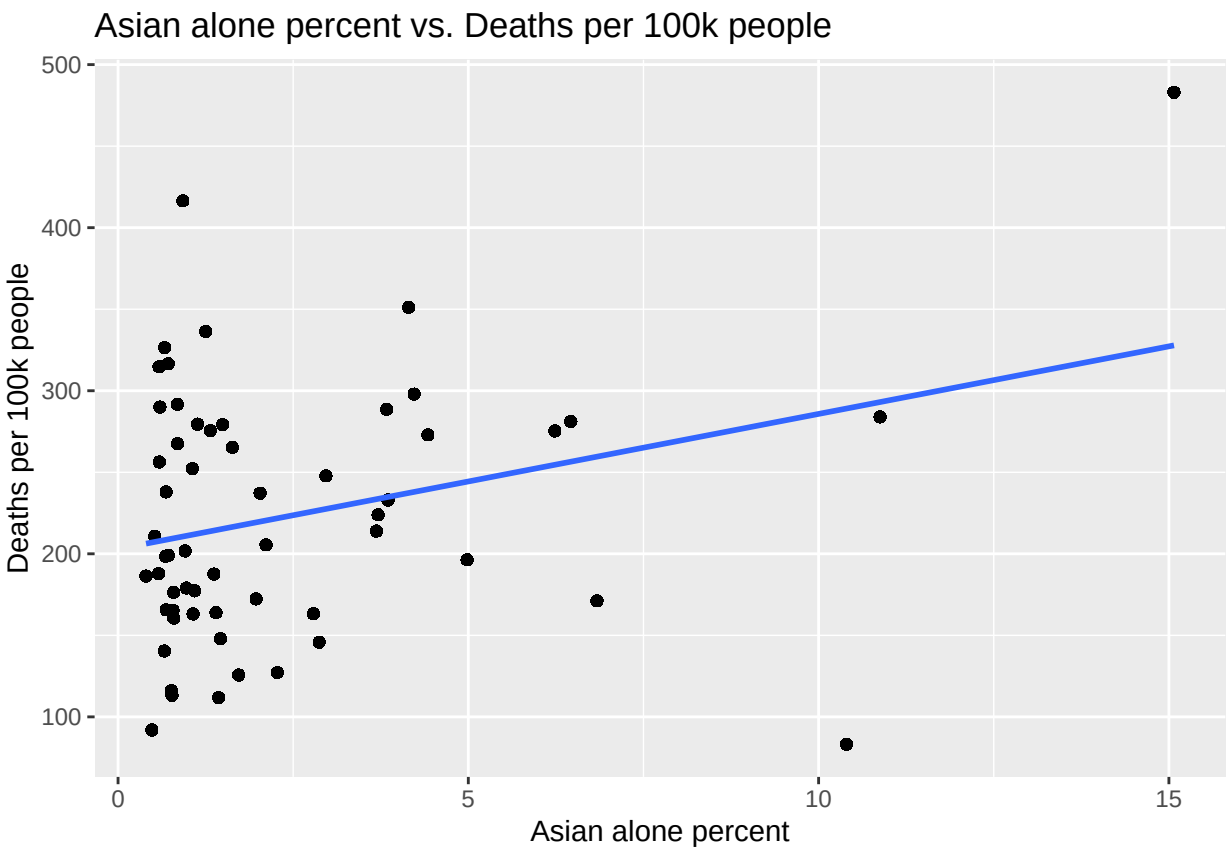

```
## `geom_smooth()` using formula = 'y ~ x'
```



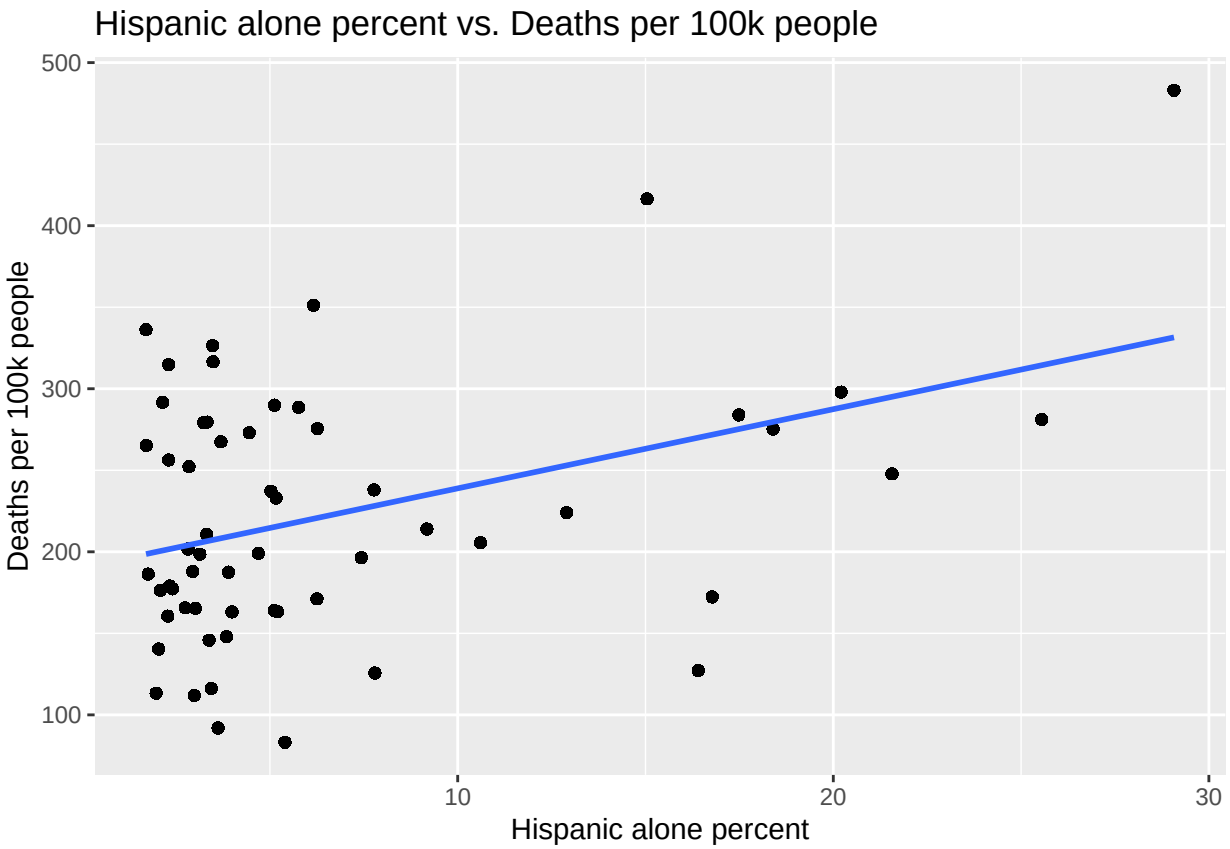

```
## `geom_smooth()` using formula = 'y ~ x'
```



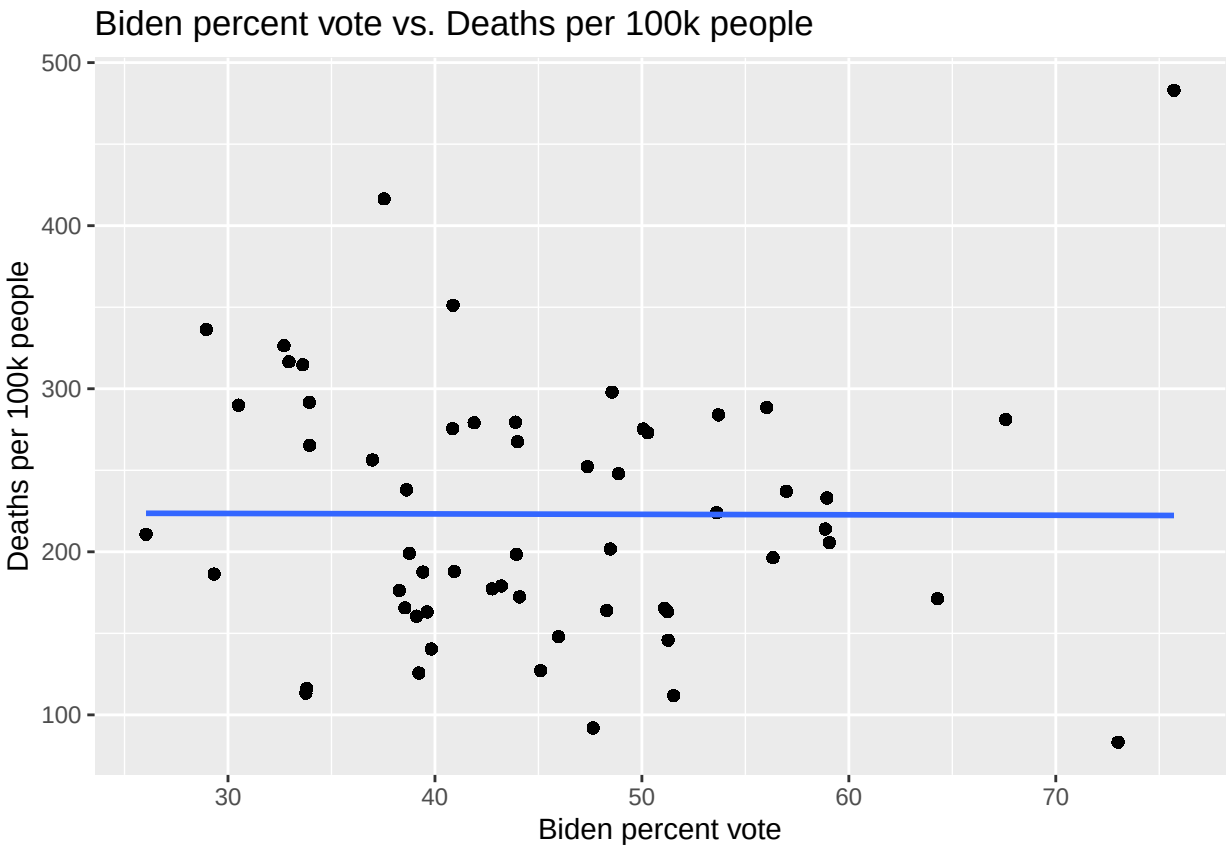

```
## `geom_smooth()` using formula = 'y ~ x'
```



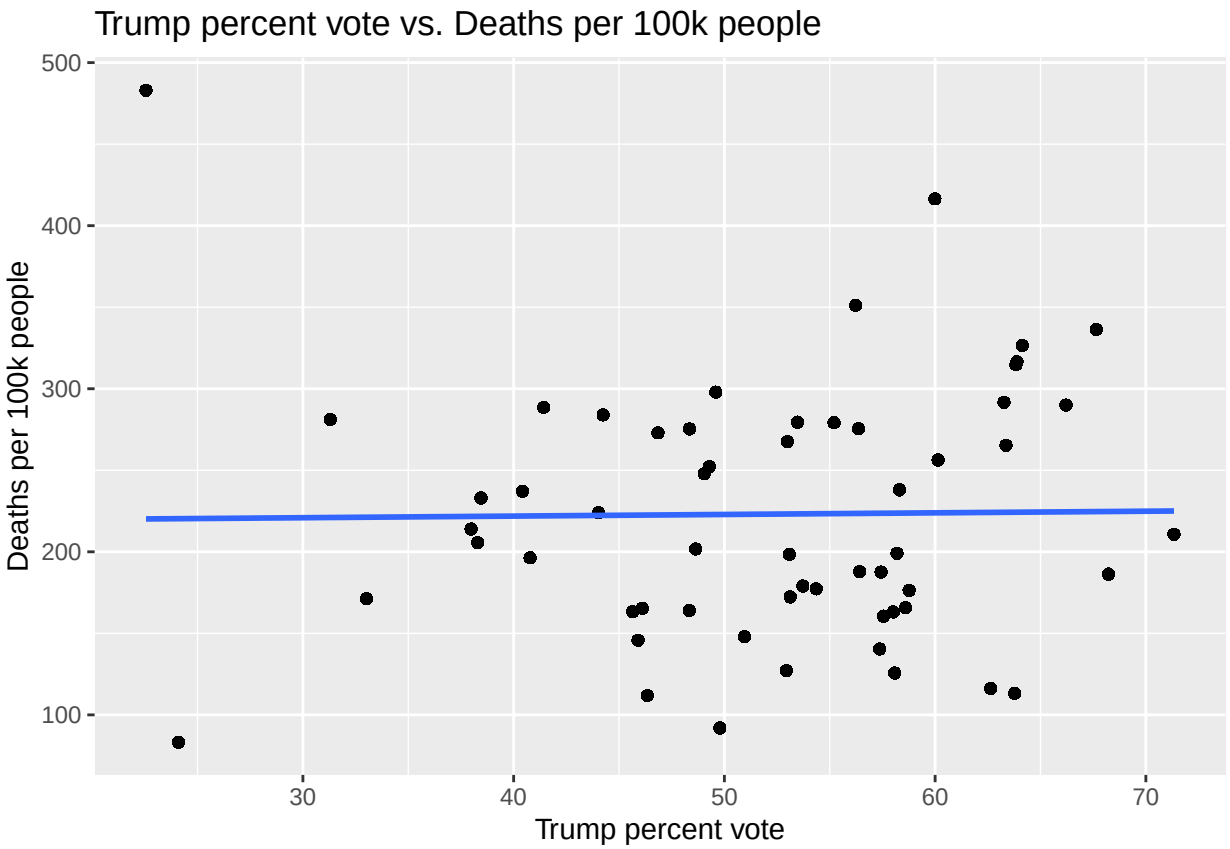

```
## `geom_smooth()` using formula = 'y ~ x'
```



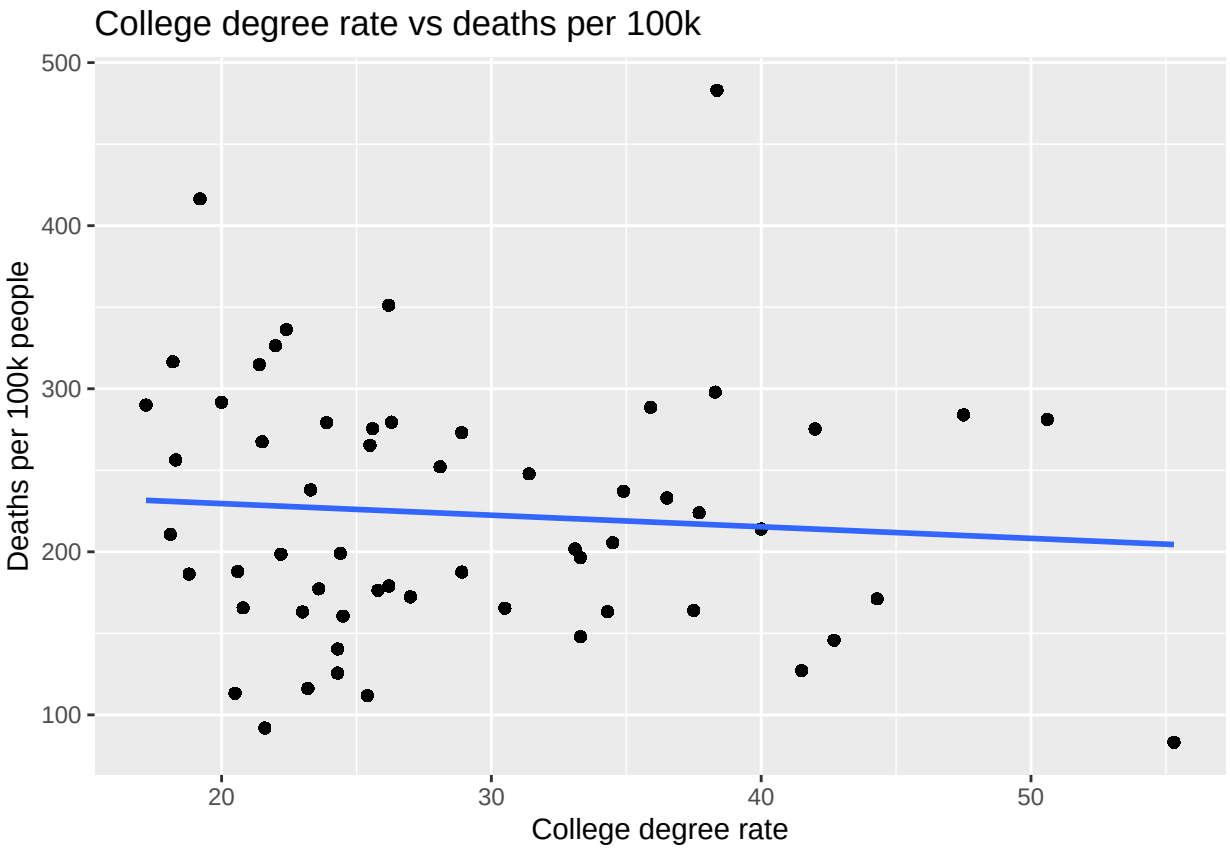

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



Single Regression models

Cases vs. Deaths

```
## # A tibble: 2 x 7
##   term                 estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept             -46.6       2.05     -22.7       0  -50.6    -42.6
## 2 total_cases_per_100k    0.012     0        133.        0    0.012    0.012
```

We first examine the relationship between cases and deaths, to see if the total cases per 100,000 people is correlated with total deaths per 100,000 people.

According to the results of the regression model, the equation is -46.630 + 0.012 cases per 100k. The intercept is the expected number of deaths per 100k when all other variables are not present. This means that in the model, there is expected to be -46.630 deaths when there are 0 cases per 100k.

The estimate of 0.012 means that for every one unit increase in cases per 100k the expected number of deaths per 100k will increase by 0.012.

Both the P-values of the intercept and the P-value of cases per 100k are significant (P-Value of 0). However, it should be noted that the estimate value of 0.012 is quite small.

```
## # A tibble: 2 x 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           211.      0.335      630.       0     211.     212.
## 2 total_population      0       0          121.       0       0        0
```

```
## # A tibble: 2 x 7
##   term                   estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                 163.      2.54       64.5       0    158.     168.
## 2 Unemployment_rate_2019     13.6     0.573      23.8       0     12.5     14.8
```

```
## # A tibble: 2 x 7
##   term                         estimate std_er~1 stati~2 p_value lower~3 upper~4
##   <chr>                           <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 intercept                        211.     1.54  137.         0    208.    214.
## 2 Median_Household_Income_2019       0      0       8.01       0      0       0
## # ... with abbreviated variable names 1: std_error, 2: statistic, 3: lower_ci,
## #   4: upper_ci
```

```
## # A tibble: 2 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept       525.       3.73      141.        0   517.     532.
## 2 white_percent    -3.40     0.042     -81.3       0    -3.48    -3.32
```

```
## # A tibble: 2 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           185.       0.582     318.        0   184.     186.
## 2 non_white_percent     3.40     0.042      81.3       0     3.32     3.48
```

```
## # A tibble: 2 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept       189.       0.512     369.        0   188.     190.
## 2 black_percent     5.75     0.065      88.9       0     5.62     5.88
```

```
## # A tibble: 2 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept       203.       0.465     437.        0   202.     204.
## 2 asian_percent     8.28     0.125      66.5       0     8.04     8.53
```

```
## # A tibble: 2 x 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept          190.       0.493     386.        0   189.     191.
## 2 hispanic_percent     4.86     0.053      91.5       0     4.75     4.96
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    218.      1.87     116.      0      214.     222.
## 2 trumpvote      0.1     0.035      2.82    0.005    0.031    0.169
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept  224.        1.61    139.      0      221.     227.
## 2 bidenvote   -0.027     0.035    -0.785   0.432   -0.095    0.041
```

```
## # A tibble: 2 x 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
```

```
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept         244.        1.26      194.        0  241.      246.
## 2 degree_rate_2010   -0.711     0.041     -17.2       0   -0.792    -0.63
```

Multiple regression model

We will consider a multiple regression model to see how COVID-19 deaths are influenced by a variety of factors. We have a wide variety of independent variables we can consider, including data on: cases, population, unemployment, median household income, white-alone percent, non-white percent, black-alone percent, asian-alone percent, hispanic-alone percent, Biden percent vote, Trump percent vote, and college degree rate.

However, in a multiple regression model we must carefully select our variables to avoid impact of confounding variables. We will consider core predictors, demographic information, race, and political opinions.

The model included demographic factors like total cases per 100k, socioeconomic factors like unemployment rate and degree rate, race factors like non-white percent, and political factors like percent voted for Trump.

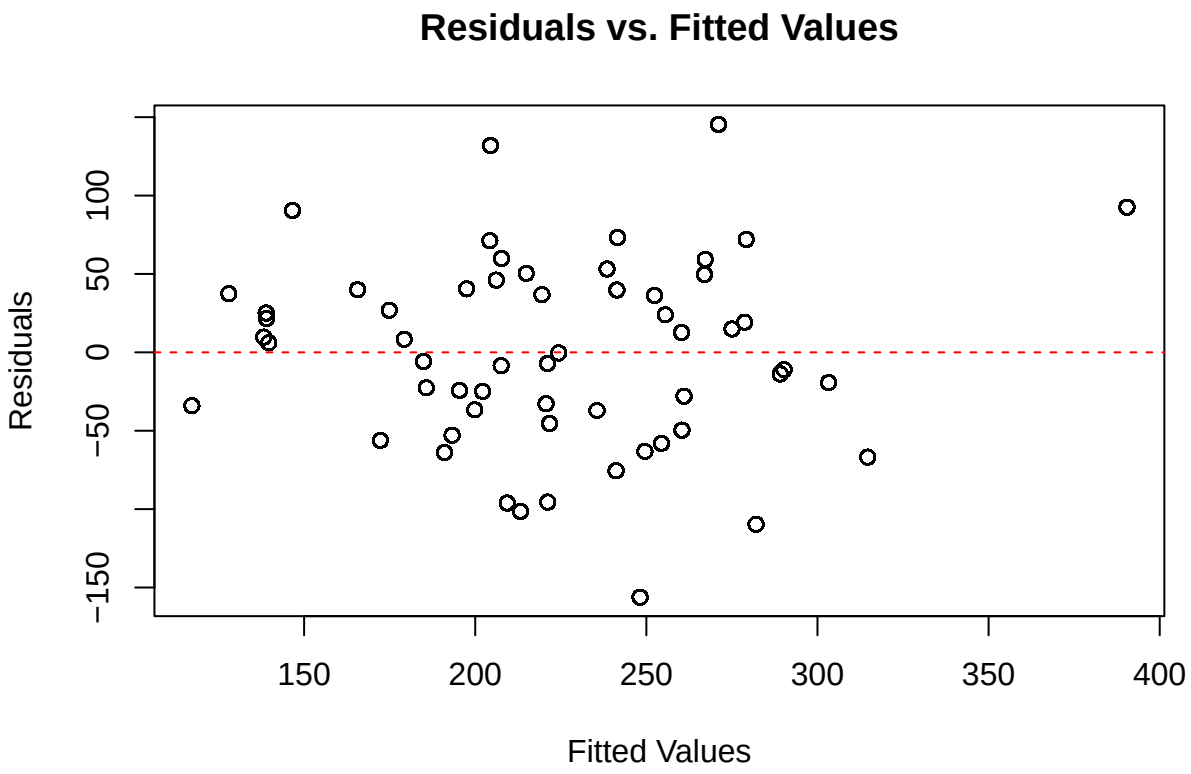
These factors were selected since they had among the highest correlation to deaths in our prior analysis. We avoided including a lot of factors that overlapped with each other. The multiple regression model allows us to consider all of these factors simultaneously.

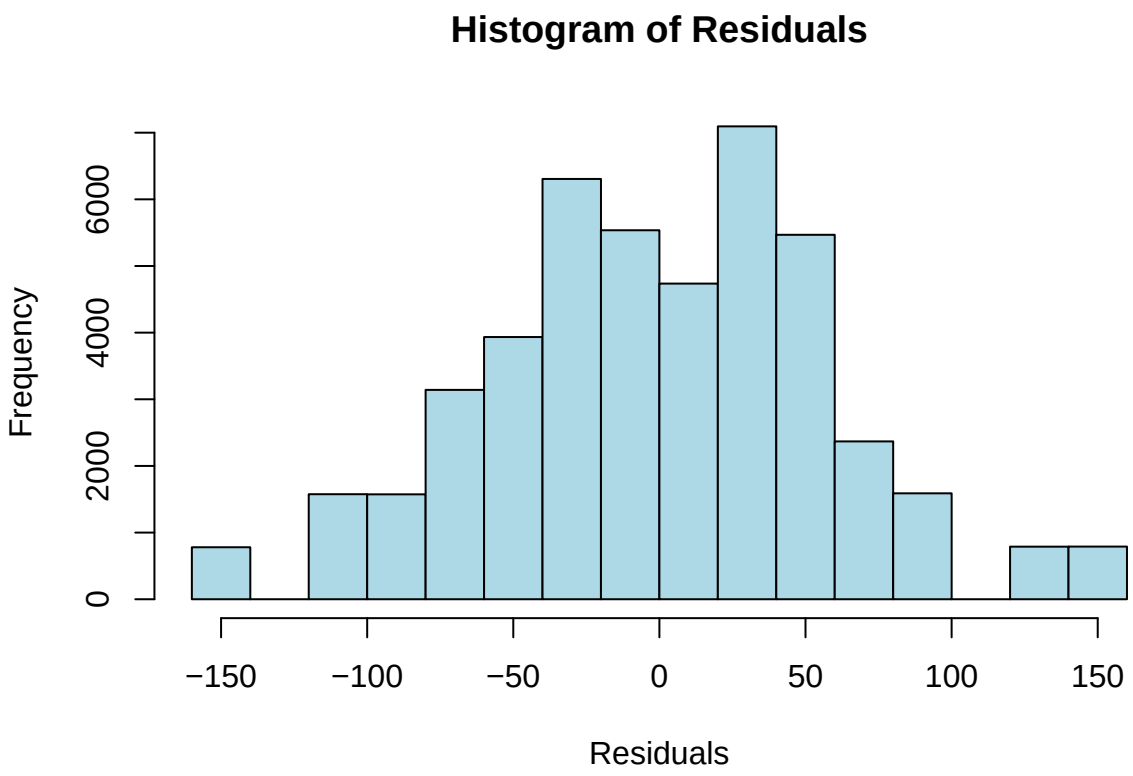
The R-squared value and root mean square error (RMSE) indicate how well the model fits the data. The R-squared value of 0.429 suggests that the model explains 42.9% of the variability in COVID-19 deaths, which indicates some predictive power. The higher a R-squared value, the more predictive power it has, if the R-squared value was 1 then it would explain 100% of the variability in COVID-19 deaths. The RMSE of 59.52 measures the model's accuracy in predicting death counts, and lower RMSE values (closer to 0) are are more desirable.

The histogram of residuals resembles a bell-shape curve exhibiting normality, supporting validity of the model. Residuals vs. fitted values are randomly scattered with no clear pattern, indicating it captures the relationship between predictors and response variable well, with constant variance and no clear patterns.

```
## # A tibble: 6 x 7
##   term                   estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                     <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                38.2       5.64       6.78       0   27.2     49.3
## 2 total_cases_per_100k      0.009     0         86.3        0    0.009    0.009
## 3 Unemployment_rate_2019   -3.63      0.567     -6.41       0   -4.74    -2.52
## 4 degree_rate_2010         -3.85      0.067    -57.5        0   -3.98    -3.72
## 5 non_white_percent         4.89      0.066     74.0        0    4.76     5.02
## 6 trumpvote                 0.991     0.064     15.5        0    0.866    1.12
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.429         0.429 3543.  59.5  59.5     6864.       0     5 45682
```


Residuals vs. Fitted Values
Residuals
100
50
0
−50
−150
150
200
250
300
350
400
Fitted Values


Histogram of Residuals
Frequency
0
2000
4000
6000
−150
−100
−50
0
50
100
150
Residuals