Johns Hopkins COVID Data Analysis

Bianca Verlangieri

2022-09-10

Setup

This is an R Markdown document describing the Johns Hopkins COVID Data Analysis. First we load in the appropriate libraries. You'll see that I suppressed the output from loading in these libraries.

```
library(RCurl)
library(tidyverse)
library(lubridate)
library(ggplot2)
```

Data Download

Now we download the data directly from the GitHub URL (rather than storing it locally). We also print out a summary of the data we just loaded. The data I've chosen to load first is the COVID death and case count which includes time series data. I chose to look at data just from the US first (rather than global data).

```
link_to_download = "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid
covid_data_us_deaths <- read_csv(link_to_download,show_col_types=FALSE)
covid_data_us_deaths
```

```
## # A tibble: 3,342 x 974
##         UID iso2  iso3  code3  FIPS Admin2   Provi~1 Count~2   Lat Long_ Combi~3
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>    <chr>   <chr>   <dbl> <dbl> <chr>
##  1 84001001 US    USA     840  1001 Autauga  Alabama US       32.5 -86.6 Autaug~
##  2 84001003 US    USA     840  1003 Baldwin  Alabama US       30.7 -87.7 Baldwi~
##  3 84001005 US    USA     840  1005 Barbour  Alabama US       31.9 -85.4 Barbou~
##  4 84001007 US    USA     840  1007 Bibb     Alabama US       33.0 -87.1 Bibb, ~
##  5 84001009 US    USA     840  1009 Blount   Alabama US       34.0 -86.6 Blount~
##  6 84001011 US    USA     840  1011 Bullock  Alabama US       32.1 -85.7 Bulloc~
##  7 84001013 US    USA     840  1013 Butler   Alabama US       31.8 -86.7 Butler~
##  8 84001015 US    USA     840  1015 Calhoun  Alabama US       33.8 -85.8 Calhou~
##  9 84001017 US    USA     840  1017 Chambers Alabama US       32.9 -85.4 Chambe~
## 10 84001019 US    USA     840  1019 Cherokee Alabama US       34.2 -85.6 Cherok~
## # ... with 3,332 more rows, 963 more variables: Population <dbl>,
## #   `1/22/20` <dbl>, `1/23/20` <dbl>, `1/24/20` <dbl>, `1/25/20` <dbl>,
## #   `1/26/20` <dbl>, `1/27/20` <dbl>, `1/28/20` <dbl>, `1/29/20` <dbl>,
## #   `1/30/20` <dbl>, `1/31/20` <dbl>, `2/1/20` <dbl>, `2/2/20` <dbl>,
## #   `2/3/20` <dbl>, `2/4/20` <dbl>, `2/5/20` <dbl>, `2/6/20` <dbl>,
## #   `2/7/20` <dbl>, `2/8/20` <dbl>, `2/9/20` <dbl>, `2/10/20` <dbl>,
## #   `2/11/20` <dbl>, `2/12/20` <dbl>, `2/13/20` <dbl>, `2/14/20` <dbl>, ...
```

```
link_to_download = "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid
covid_data_us_cases <- read_csv(link_to_download,show_col_types=FALSE)
covid_data_us_cases
```

```
## # A tibble: 3,342 x 973
##         UID iso2  iso3  code3  FIPS Admin2   Provi~1 Count~2   Lat Long_ Combi~3
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>    <chr>   <chr>   <dbl> <dbl> <chr>
##  1 84001001 US    USA     840  1001 Autauga  Alabama US       32.5 -86.6 Autaug~
##  2 84001003 US    USA     840  1003 Baldwin  Alabama US       30.7 -87.7 Baldwi~
##  3 84001005 US    USA     840  1005 Barbour  Alabama US       31.9 -85.4 Barbou~
##  4 84001007 US    USA     840  1007 Bibb     Alabama US       33.0 -87.1 Bibb, ~
##  5 84001009 US    USA     840  1009 Blount   Alabama US       34.0 -86.6 Blount~
##  6 84001011 US    USA     840  1011 Bullock  Alabama US       32.1 -85.7 Bulloc~
##  7 84001013 US    USA     840  1013 Butler   Alabama US       31.8 -86.7 Butler~
##  8 84001015 US    USA     840  1015 Calhoun  Alabama US       33.8 -85.8 Calhou~
##  9 84001017 US    USA     840  1017 Chambers Alabama US       32.9 -85.4 Chambe~
## 10 84001019 US    USA     840  1019 Cherokee Alabama US       34.2 -85.6 Cherok~
## # ... with 3,332 more rows, 962 more variables: `1/22/20` <dbl>,
## #   `1/23/20` <dbl>, `1/24/20` <dbl>, `1/25/20` <dbl>, `1/26/20` <dbl>,
## #   `1/27/20` <dbl>, `1/28/20` <dbl>, `1/29/20` <dbl>, `1/30/20` <dbl>,
## #   `1/31/20` <dbl>, `2/1/20` <dbl>, `2/2/20` <dbl>, `2/3/20` <dbl>,
## #   `2/4/20` <dbl>, `2/5/20` <dbl>, `2/6/20` <dbl>, `2/7/20` <dbl>,
## #   `2/8/20` <dbl>, `2/9/20` <dbl>, `2/10/20` <dbl>, `2/11/20` <dbl>,
## #   `2/12/20` <dbl>, `2/13/20` <dbl>, `2/14/20` <dbl>, `2/15/20` <dbl>, ...
```

Cleaning of the Data

Now that we have looked at a brief summary of the data, we can start cleaning it up. You'll notice it printed out a lot of information, and that's because the dates were stored as separate variables. We are going to follow the code that was shown in class, to consolidate and rename the fields we are interested in, like dates and numbers of deaths and cases. Let's look at a summary of the data after we've cleaned it up and combined into 1 data set.

```
covid_data_us_deaths <- covid_data_us_deaths %>% pivot_longer(cols = -(UID:Population), names_to = "dat

covid_data_us_cases <- covid_data_us_cases %>% pivot_longer(cols = -(UID:Combined_Key), names_to = "dat


covid_data <- covid_data_us_deaths %>%
  full_join(covid_data_us_cases)
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key",
## "date")
```

```
summary(covid_data)
```

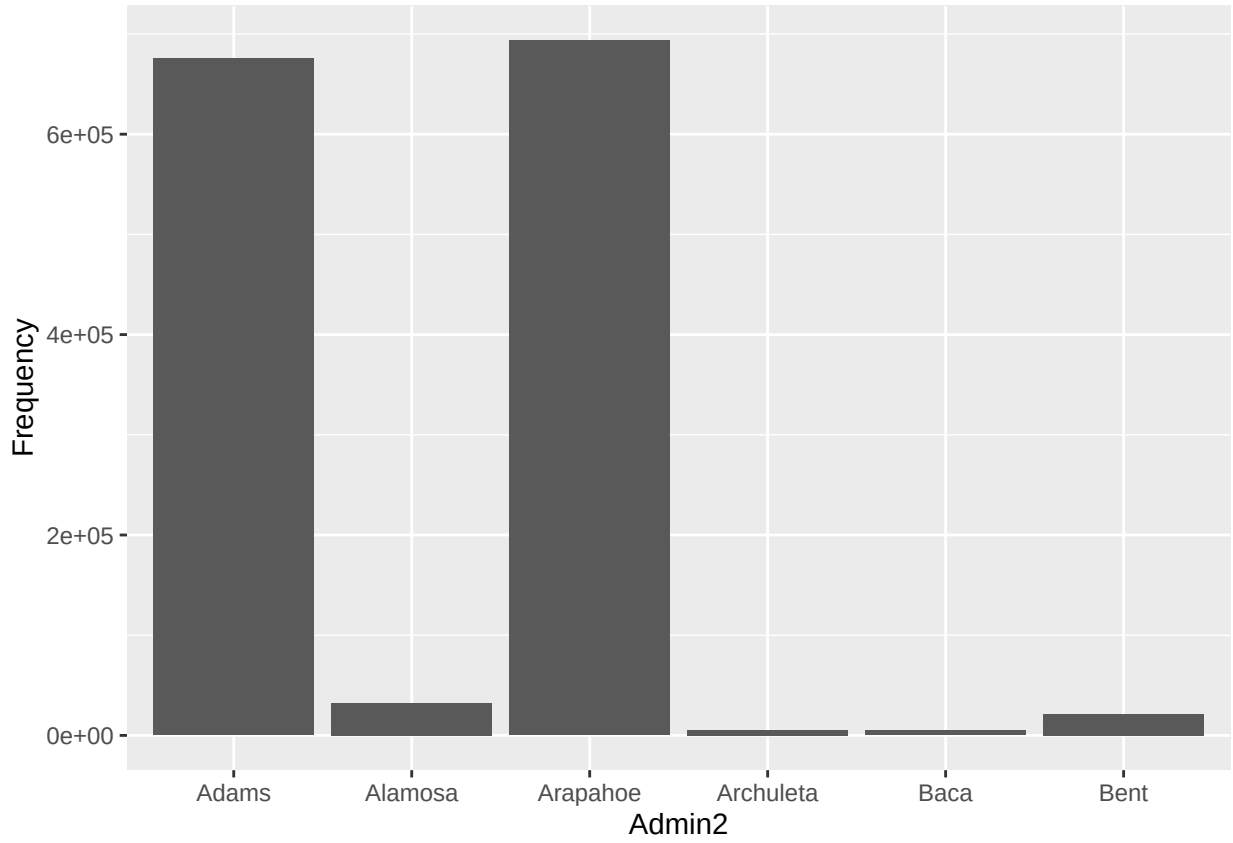
```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:3215004     Length:3215004     Length:3215004     Length:3215004
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##    Population            date                deaths            cases
##  Min.   :       0   Min.   :2020-01-22   Min.   :  -82.0   Min.   :  -3073
##  1st Qu.:    9917   1st Qu.:2020-09-18   1st Qu.:    2.0   1st Qu.:    192
##  Median :   24892   Median :2021-05-16   Median :   29.0   Median :   1685
##  Mean   :   99604   Mean   :2021-05-16   Mean   :  160.9   Mean   :  11133
##  3rd Qu.:   64979   3rd Qu.:2022-01-12   3rd Qu.:  101.0   3rd Qu.:   6285
##  Max.   :10039107   Max.   :2022-09-09   Max.   :33348.0   Max.   :3425863
```

Visualization

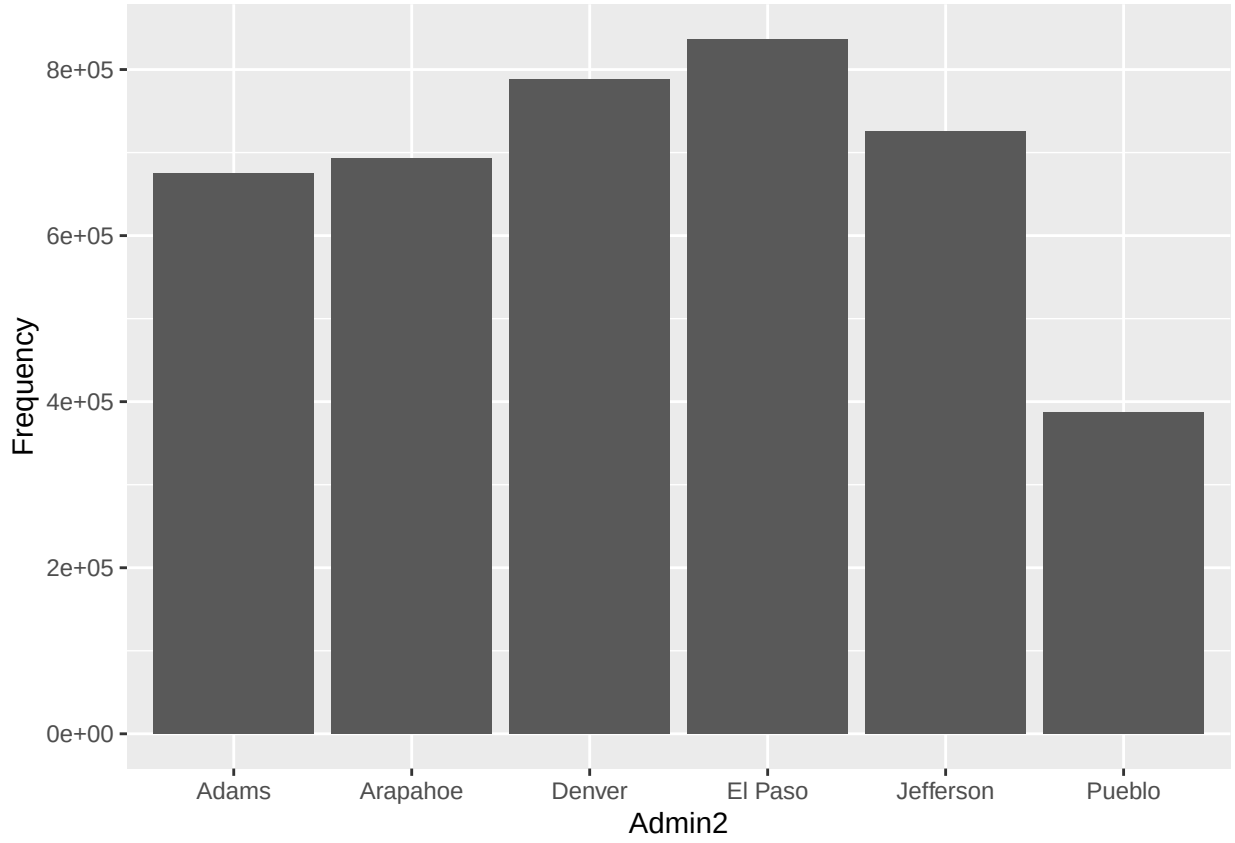
Now that we've cleaned up the data a bit let's visualize the data and see if we find anything interesting. We'll use ggplot to help us plot the data. First, we'll grab just the data from the state of Colorado. I want to look at how many COVID deaths there were per county.

```
colorado_data <- covid_data %>%
  filter(Province_State == "Colorado", deaths > 0, cases > 0, Population > 0) %>%
  group_by(date, Admin2)
co_counties_deaths <- colorado_data %>%
  group_by(Admin2) %>%
  summarise(Frequency = sum(deaths))
ggplot(co_counties_deaths[1:6,], aes(x=Admin2, y=Frequency)) + geom_bar(stat="identity")
```



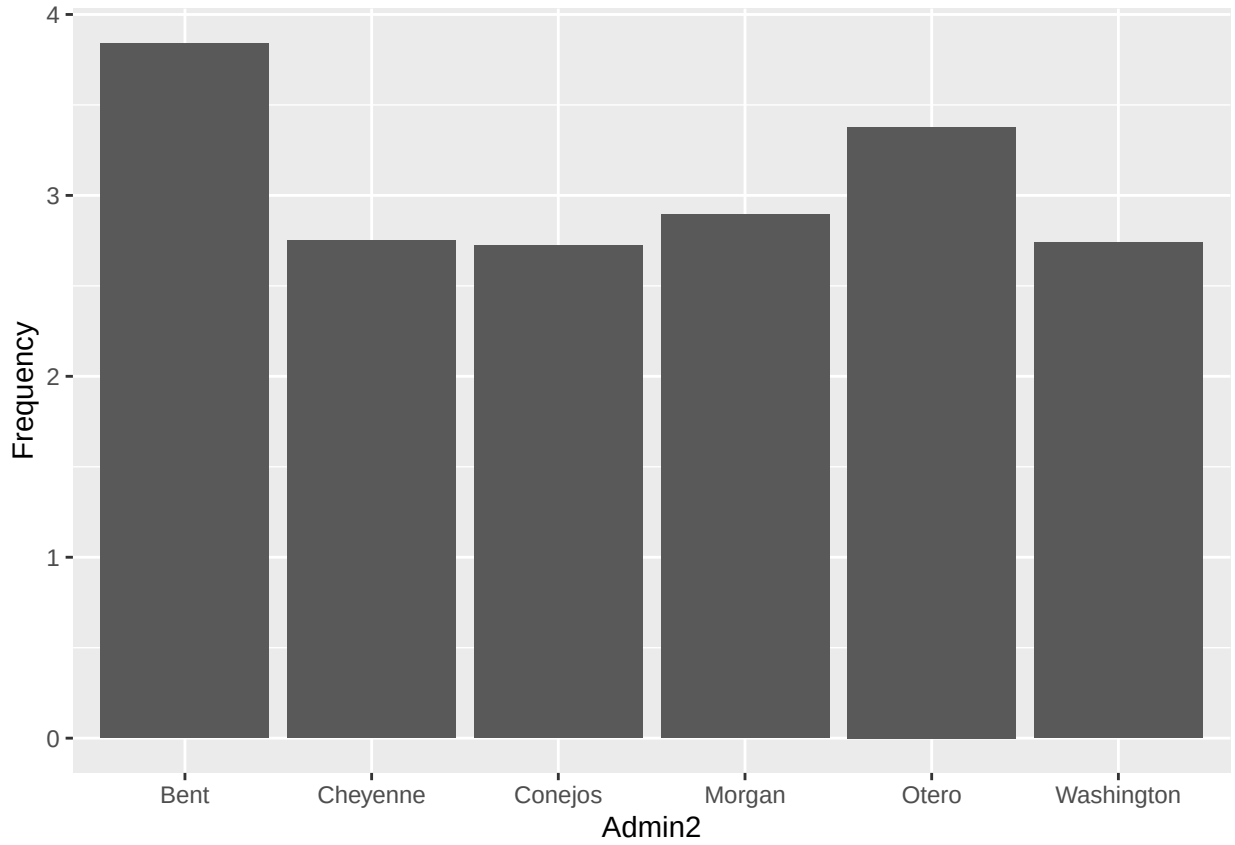
Now I just plotted the first 6 (which are in alphabetical order) because there are a lot of counties. Let's try to plot the top 6 counties with most COVID deaths.

```
co_counties_deaths = co_counties_deaths[order(-co_counties_deaths$Frequency),]
ggplot(co_counties_deaths[1:6,], aes(x=Admin2, y=Frequency)) + geom_bar(stat="identity")
```



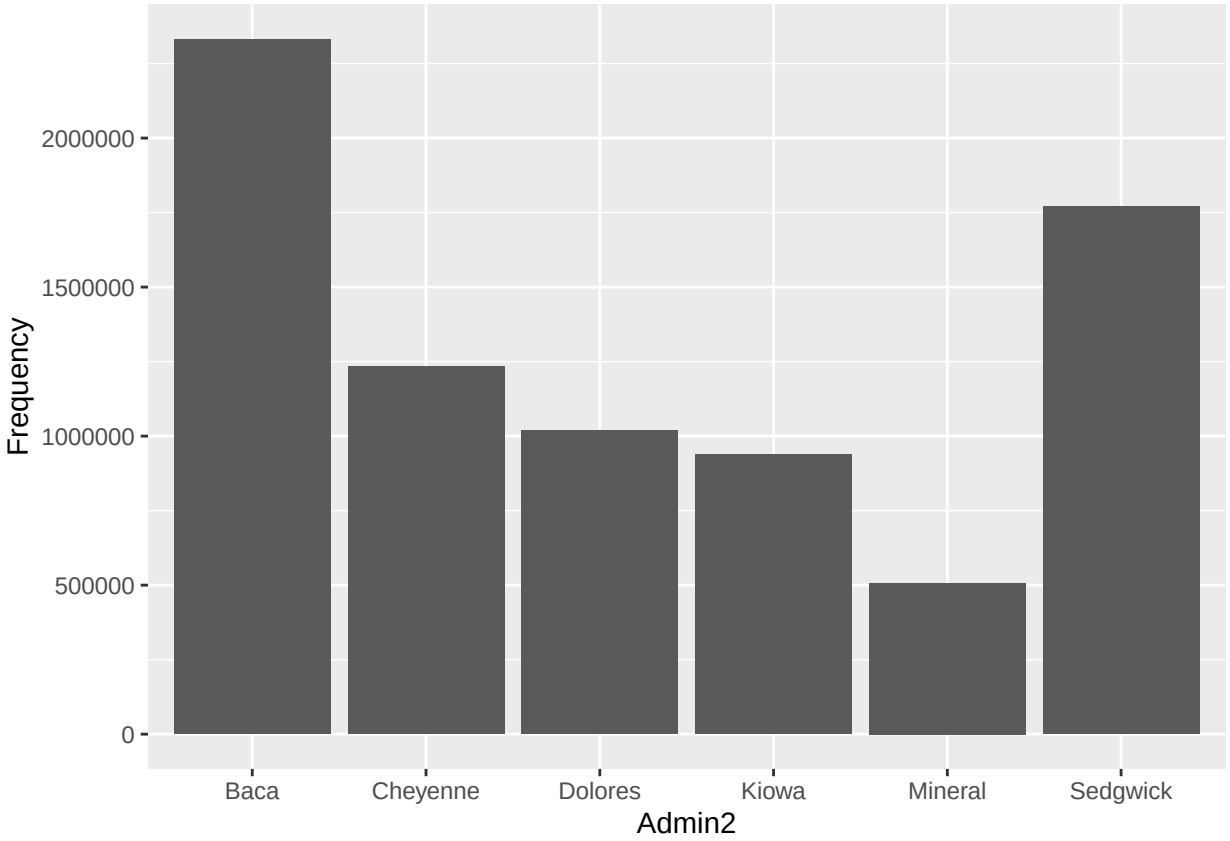
From these counts, it looks like El Paso, Denver, and Jefferson County had the most COVID deaths. This our some of the biggest counties in Colorado, so that makes sense. But how does this look if we normalize the counties by population?

```
co_counties_norm <- colorado_data %>%
  group_by(Admin2) %>%
  summarise(Frequency = sum(deaths/Population))
co_counties_norm = co_counties_norm[order(-co_counties_norm$Frequency),]
ggplot(co_counties_norm[1:6,], aes(x=Admin2, y=Frequency)) + geom_bar(stat="identity")
```



Looks like some fairly small counties took the cake on this one, let's check and make sure that these also had small populations.

```
co_counties_pop <- colorado_data %>%
  group_by(Admin2) %>%
  summarise(Frequency = sum(Population))
co_counties_pop = co_counties_pop[order(co_counties_pop$Frequency),]
ggplot(co_counties_pop[1:6,], aes(x=Admin2, y=Frequency)) + geom_bar(stat="identity")
```



Looks like Cheyenne was the only one in the smallest six counties, so looks like this might not be too biased by population when normalized.

Analysis

Let's see if there are any relationships in this data by looking at the Colorado county total cases and deaths normalized by population.

```
colorado_data <- colorado_data %>%
  group_by(Admin2) %>%
  summarize(deaths = max(deaths), cases = max(cases), Population = max(Population)) %>%
  mutate(cases_per_hundred = 100 * cases / Population, deaths_per_hundred = 100 * deaths / Population ) %>%
  select(Admin2, cases, deaths, Population, cases_per_hundred, deaths_per_hundred)

mod <- lm(deaths_per_hundred ~ cases_per_hundred, data = colorado_data)

summary(mod)
```

```
##
## Call:
## lm(formula = deaths_per_hundred ~ cases_per_hundred, data = colorado_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.32172 -0.14617 -0.03107  0.12360  0.47299
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.213139   0.092116   2.314   0.0242 *
## cases_per_hundred 0.003535   0.003239   1.091   0.2795
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1872 on 59 degrees of freedom
## Multiple R-squared:  0.01979,    Adjusted R-squared:  0.003175
## F-statistic: 1.191 on 1 and 59 DF,  p-value: 0.2795
```

The p value for cases per hundred is less than 0.05 which means that it could be significant. The p value for the model is 0.2795. Both need to be less than 0.05 for the linear model to be statistically significant.

Conclusions

We have looked at COVID-19 data from Johns Hopkins university in the US, and specifically in the state of Colorado. We looked at numbers of cases, deaths, and population size in different counties. We found El Paso and Denver county to have the most COVID-19 deaths, but not the most deaths by population in that county. We looked at a linear model for cases and deaths normalized by population, but did not find a statistically signficiant relationship. There could be bias in this data, for example, based on how the data was collected in each county. Some counties may have people reporting more cases and deaths than others (more cases gone unreported) which could largely impact the data, especially in smaller counties.