# COVID Data Analysis

S

2023-02-28

## Introduction and Summary

In this project, I analyzed COVID-19 case/death data provided by Johns Hopkins University. I imported and cleaned up both US and global data and looked at trends. For the visualization and analysis portions of the project, I focused just on the US data. In my visualizations, I looked at the cumulative aggregate statistics in the US by county. My plots and histograms showed that as a function of population, both case numbers and death numbers increased relatively linearly as expected, although the spread was wider for the deaths vs. population plot. For the analysis portion of the project, I explored the seasonable nature of COVID-19's spread by generating a polynomial regression model that related month of the year with the average latitude of county's with their worst day (as measured by number of new deaths). This showed that while winter months were overall correlated with more COVID spread, during the summer, there was a relative increase in 'bad' days in the south compared with the north. Such an insight could be useful to public health officials seeking to pre-allocate hospital supplies each month to the part of the country that will need it most.

## Import the Data

```
url_in = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names = c("time_series_covid19_confirmed_US.csv",  "time_series_covid19_confirmed_global.csv", "ti
urls = str_c(url_in,file_names)

global_cases = read_csv(urls[2])
```

```
## Rows: 289 Columns: 1137
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1135): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
global_deaths = read_csv(urls[4])
```

```
## Rows: 289 Columns: 1137
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
```

```
## dbl (1135): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
us_cases = read_csv(urls[1])
```

```
## Rows: 3342 Columns: 1144
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1138): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
us_deaths = read_csv(urls[3])
```

```
## Rows: 3342 Columns: 1145
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1139): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

### Tidy the Data and Generate Summary Statistics

```r
global_deaths = global_deaths %>% pivot_longer(cols = -c('Province/State','Country/Region', Lat, Long),

global_cases = global_cases %>% pivot_longer(cols = -c('Province/State','Country/Region', Lat, Long), na

us_cases = subset(us_cases, select = -c(UID,iso2,iso3,code3,FIPS))

us_deaths = subset(us_deaths, select = -c(UID,iso2,iso3,code3,FIPS))

us_cases = us_cases %>% pivot_longer(cols = -c('Province_State','Country_Region', Admin2, Combined_Key,

us_deaths = us_deaths %>% pivot_longer(cols = -c('Province_State','Country_Region', Admin2, Combined_Key

global = global_cases %>% full_join(global_deaths)
```

```
## Joining with `by = join_by(`Province/State`, `Country/Region`, Lat, Long,
## date)`
```

```r
us = us_cases %>% full_join(us_deaths)
```

```
## Joining with `by = join_by(Admin2, Province_State, Country_Region, Lat, Long_,
## Combined_Key, date)`
```

```r
global = global %>% mutate(global, date = mdy(date))


us = us %>% mutate(us, date = mdy(date))

global = global %>% filter(cases > 0)

us = us %>% filter(cases > 0)

summary(us)
```

```
##     Admin2          Province_State     Country_Region          Lat
##  Length:3441829     Length:3441829     Length:3441829     Min.   :-14.27
##  Class :character   Class :character   Class :character   1st Qu.: 34.12
##  Mode  :character   Mode  :character   Mode  :character   Median : 38.06
##                                                           Mean   : 37.46
##                                                           3rd Qu.: 41.67
##                                                           Max.   : 69.31
##      Long_         Combined_Key            date               cases
##  Min.   :-174.16   Length:3441829     Min.   :2020-01-22   Min.   :      1
##  1st Qu.: -97.66   Class :character   1st Qu.:2020-12-24   1st Qu.:    677
##  Median : -89.54   Mode  :character   Median :2021-09-15   Median :   2816
##  Mean   : -90.21                      Mean   :2021-09-14   Mean   :  15334
##  3rd Qu.: -82.63                      3rd Qu.:2022-06-07   3rd Qu.:   9242
##  Max.   : 145.67                      Max.   :2023-02-27   Max.   :3697797
##    Population            deaths
##  Min.   :       0   Min.   :    0.0
##  1st Qu.:   10953   1st Qu.:   10.0
##  Median :   26248   Median :   46.0
##  Mean   :  104523   Mean   :  203.8
##  3rd Qu.:   68098   3rd Qu.:  136.0
##  Max.   :10039107   Max.   :35366.0
```

```r
summary(global)
```

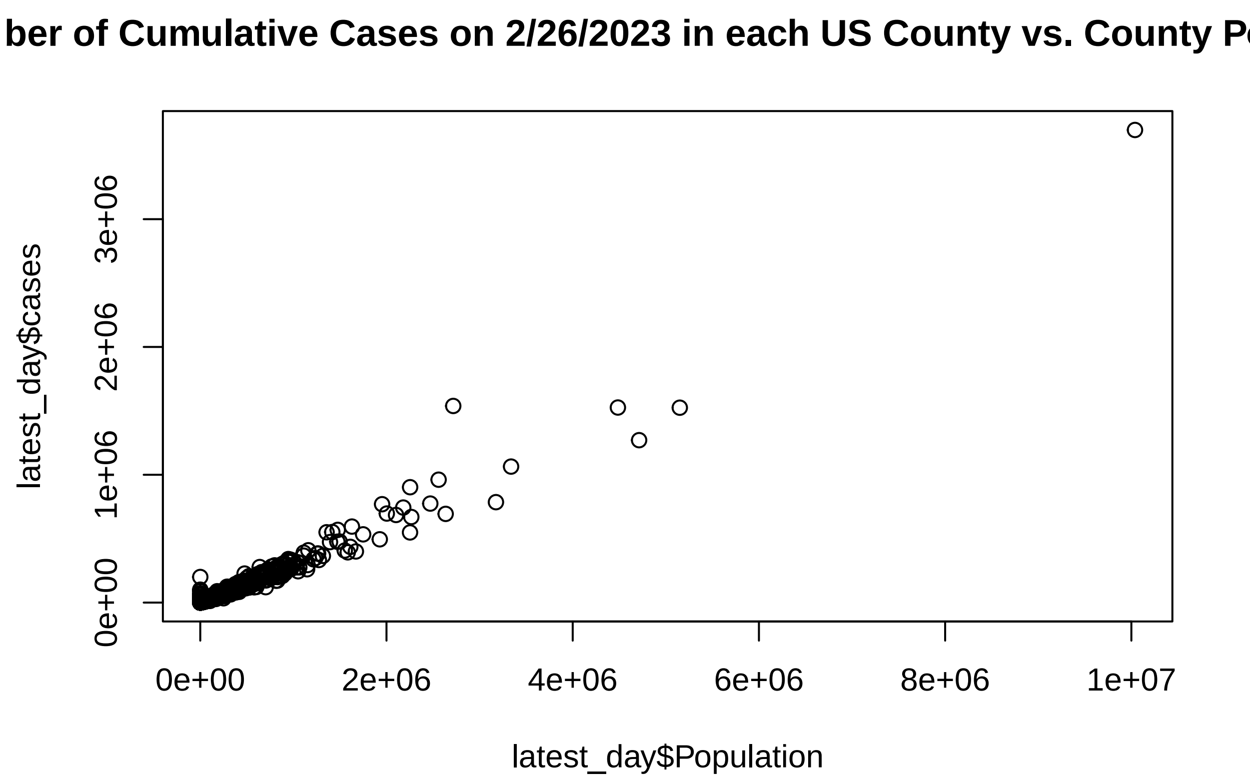
```
##  Province/State     Country/Region          Lat                Long
##  Length:303957      Length:303957      Min.   :-71.950   Min.   :-178.12
##  Class :character   Class :character   1st Qu.:  5.152   1st Qu.: -19.02
##  Mode  :character   Mode  :character   Median : 22.167   Median :  21.01
##                                        Mean   : 20.535   Mean   :  23.16
##                                        3rd Qu.: 41.113   3rd Qu.:  88.09
##                                        Max.   : 71.707   Max.   : 178.06
##                                        NA's   :1890      NA's   :1890
##       date                cases                deaths
##  Min.   :2020-01-22   Min.   :        1   Min.   :      0
##  1st Qu.:2020-12-10   1st Qu.:     1290   1st Qu.:      7
##  Median :2021-09-11   Median :    20049   Median :    212
##  Mean   :2021-09-06   Mean   :  1020376   Mean   :  14315
##  3rd Qu.:2022-06-08   3rd Qu.:   268070   3rd Qu.:   3630
##  Max.   :2023-02-27   Max.   :103389954   Max.   :1119560
##
```

## Generate Visualizations of the Data

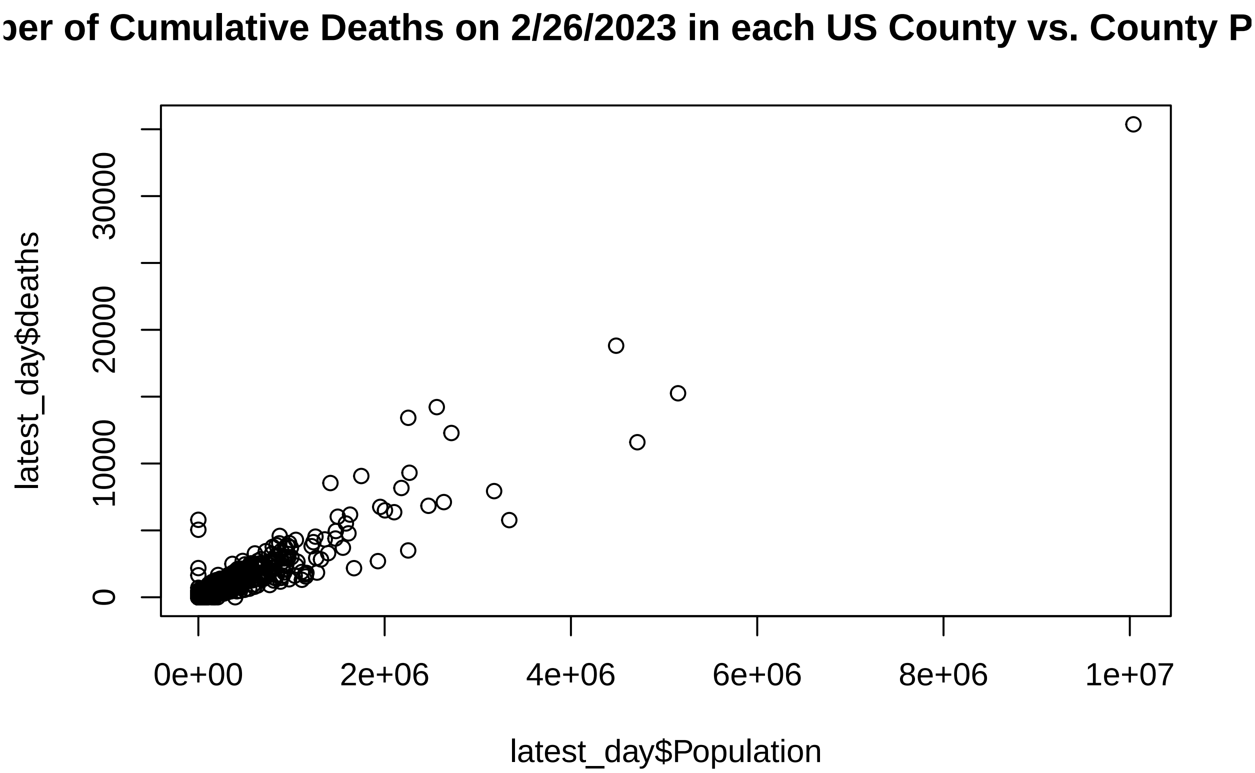
The first two plots below show the aggregate number of cases and deaths versus the population of each US county. Interestingly, the spread is wider for deaths vs. population (indicating variability in care and population susceptibility among other factors) than number of cases (which is relatively narrowly spread and linear).

The second two plots show histograms of case and population fatality rates for each US county. There's some degree of skewness, but they both look generally normal/bell curve shaped.

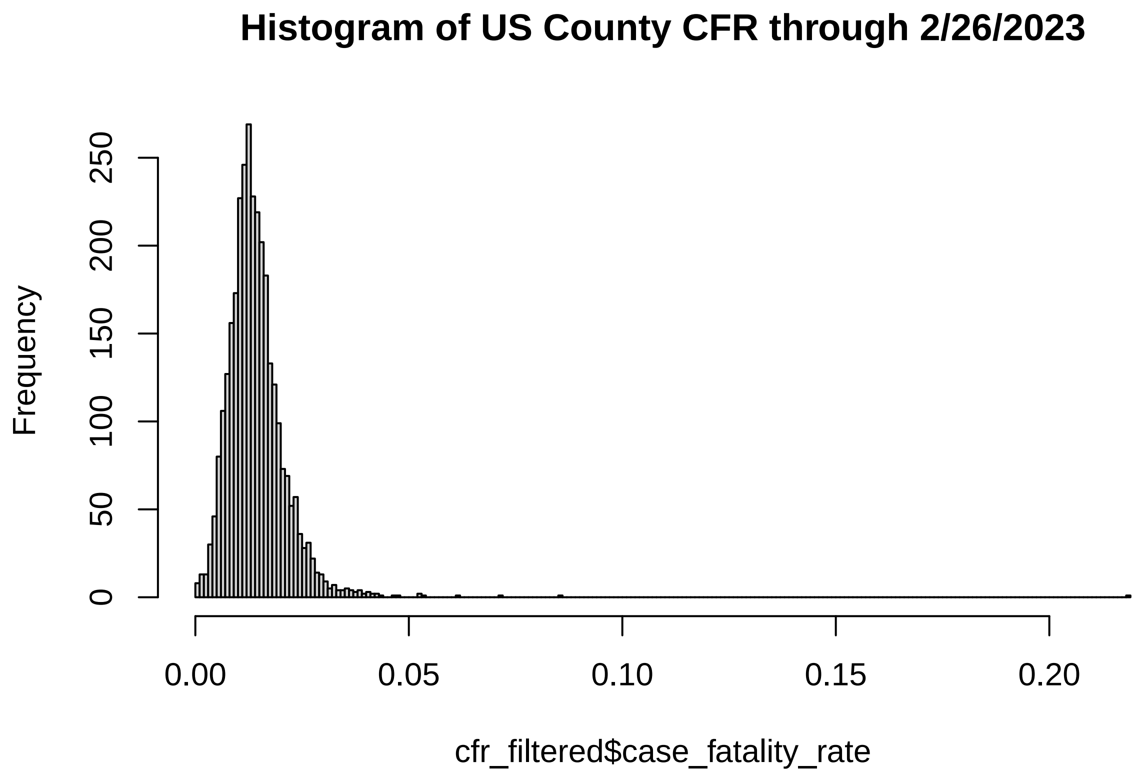
```
#filter to latest time
latest_day = us %>% filter(date == "2023-02-26")
frame()
plot(latest_day$Population,latest_day$cases)
title("Number of Cumulative Cases on 2/26/2023 in each US County vs. County Population")
```



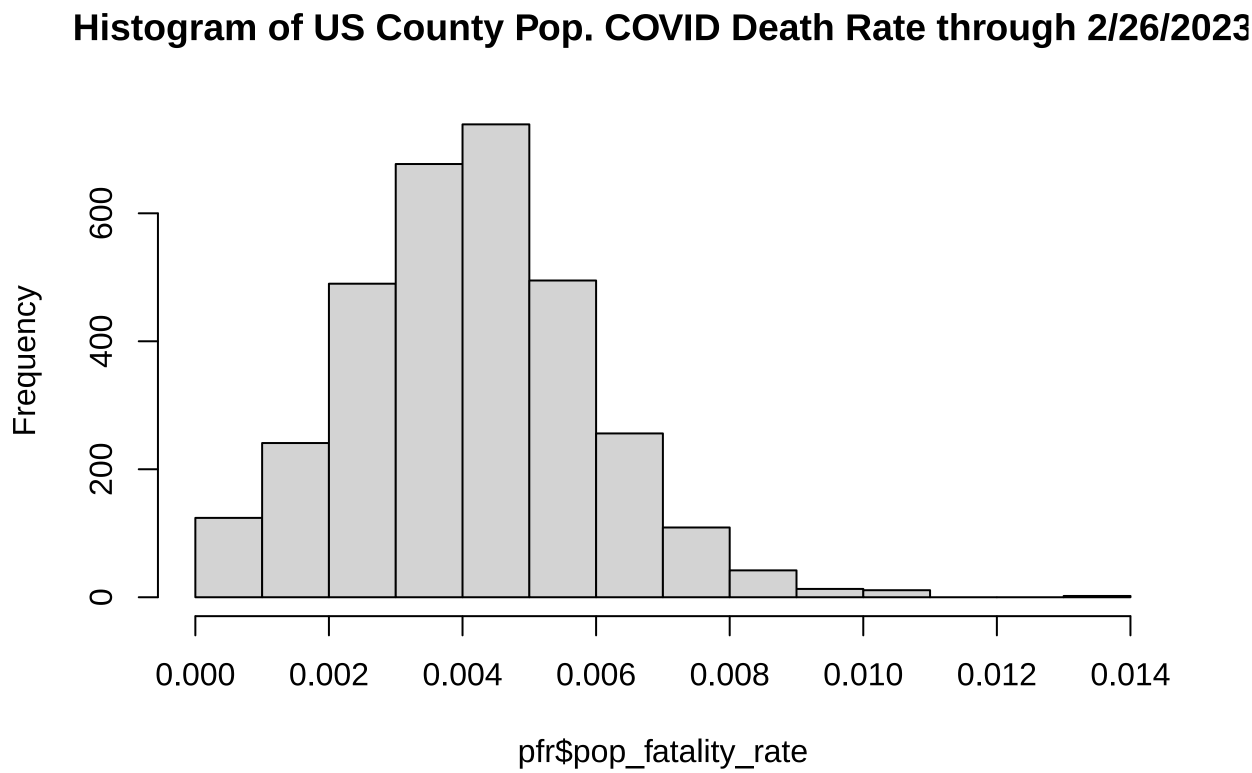

```
frame()
plot(latest_day$Population,latest_day$deaths)
title("Number of Cumulative Deaths on 2/26/2023 in each US County vs. County Population")
```




```
case_fatality_rate = latest_day$deaths/latest_day$cases
cfr = data.frame(case_fatality_rate)
cfr_filtered = filter(cfr, case_fatality_rate< .3, case_fatality_rate > 0) #filtering out a few countie
frame()
hist(cfr_filtered$case_fatality_rate, breaks = 200, main = "Histogram of US County CFR through 2/26/202
```




```
pop_fatality_rate = latest_day$deaths/latest_day$Population
pfr = data.frame(pop_fatality_rate)
frame()
hist(pfr$pop_fatality_rate, main = "Histogram of US County Pop. COVID Death Rate through 2/26/2023")
```




### Analysis

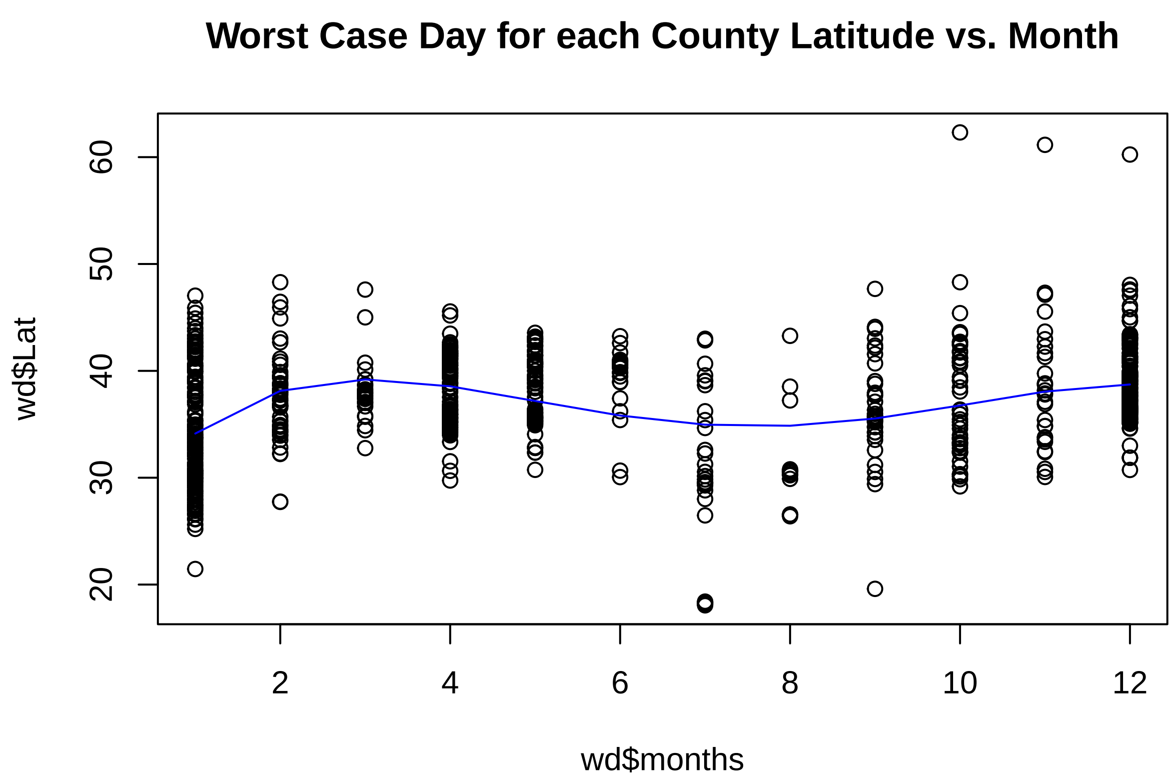
For my analysis, I generated a 4 degree polynomial regression model to correlate each US county's worst month (as measured by number of new deaths in a given day) with county latitude. More specifically, the model predicts the average latitude where the worst case days will occur based on an input month. The R^2 value of .1388 certainly indicates incompleteness of this correlation - as we know, COVID spread is highly multivariable. That said, it does show some degree of correlation that could public health planning. For example, there is a latitude dip during the summer months, which could be due to people spending more time inside in the south when the weather is hotter.

```
us = us %>% mutate(new_cases = cases - lag(cases),new_deaths = deaths - lag(deaths))
#Relationship between worst day and total population fatality rate

#relationship between day of the worst day and lat
usf = us
usf = usf %>% filter(new_deaths>10)
counties = factor(usf$Combined_Key)
usf$counties = counties

month = format(as.Date(usf$date, format = "%Y-%m-%d"), "%m")
months = factor(month)
months = as.numeric(months)
usf$months = months

wd = usf %>% group_by(Combined_Key) %>% slice_max(new_deaths)
```

```
wd = wd %>% filter(Lat > 0)

wd$months2 = (wd$months)^2
ms = wd$months
model = lm(wd$Lat~ poly(ms,4))

ms = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)

predict1 = predict(model, data.frame(months))
```

```
## Warning: 'newdata' had 14817 rows but variables found have 12 rows
```

```
frame()
plot(wd$months,wd$Lat)
lines(ms,predict1,col='blue')
title("Worst Case Day for each County Latitude vs. Month")
```




```
summary(model)
```

```
##
## Call:
## lm(formula = wd$Lat ~ poly(ms, 4))
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9030  -3.0109  -0.3534   2.9746  25.5520
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.0700     0.1553 238.636  < 2e-16 ***
## poly(ms, 4)1  29.8553     4.5898   6.505 1.31e-10 ***
## poly(ms, 4)2  -9.0085     4.5898  -1.963     0.05 *
## poly(ms, 4)3  38.6721     4.5898   8.426  < 2e-16 ***
## poly(ms, 4)4 -21.8674     4.5898  -4.764 2.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.59 on 868 degrees of freedom
## Multiple R-squared:  0.1388, Adjusted R-squared:  0.1348
## F-statistic: 34.96 on 4 and 868 DF,  p-value: < 2.2e-16
```

## Conclusion

In this project, I imported, cleaned, and visualized data on COVID cases and deaths within the United States and across the world. As an analysis, I built a polynomial regression model to predict latitude of counties with their highest death number days based on the month of the year. Public health officials could use such insights to pre-allocate hospital equipment in regions they anticipate will have higher death rates from COVID or other respiratory viruses.

A bias innate within all COVID data is the prevalence of testing to determine the number of cases as well as the criteria used to determine whether a death is caused by COVID or is due to another condition while the patient just happened test positive for COVID.

A personal bias I might have in this analysis is my belief that death numbers are more statistically important than case numbers. Especially as the pandemic progressed beyond the initial stages of uncertainty and effective vaccines became available (at least where I live in the US), I personally stopped thinking much about COVID. If I got it, I got it. I didn't think the public health measures were worth the societal costs they imposed for the most par; they only delayed the inevitable. This personal bias could have been a partial subconscious motivation to focus my analysis death rates by month rather than case rates. Taking a step back, case rates do have some relevance (even if not as much as death rates), so if I were to continue this project beyond the scope of the assignment, I would also generate a predictive model to correlate months with latitude of case spikes as well.