COVID19 Data

Mike Prodo

11/18/2021

The Data

The data being used for this project is COVID 19 cases in the US reported from Johns Hopkins. It is on their github site at https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series. This data set is time_series_covid19_confirmed_US.csv and it contains the number of reported COVID 19 cases in each county of each state dating back to March 22, 2020.

Libraries used in this project are ggplot2.

Reading in the data:

```
covid = read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/cs
```

For this project I will be looking at COVID 19 cases in the state of Virginia.

Cleaning the data:

```
# extract Virginia data
va = covid[covid[,"Province_State"] == "Virginia",]
```

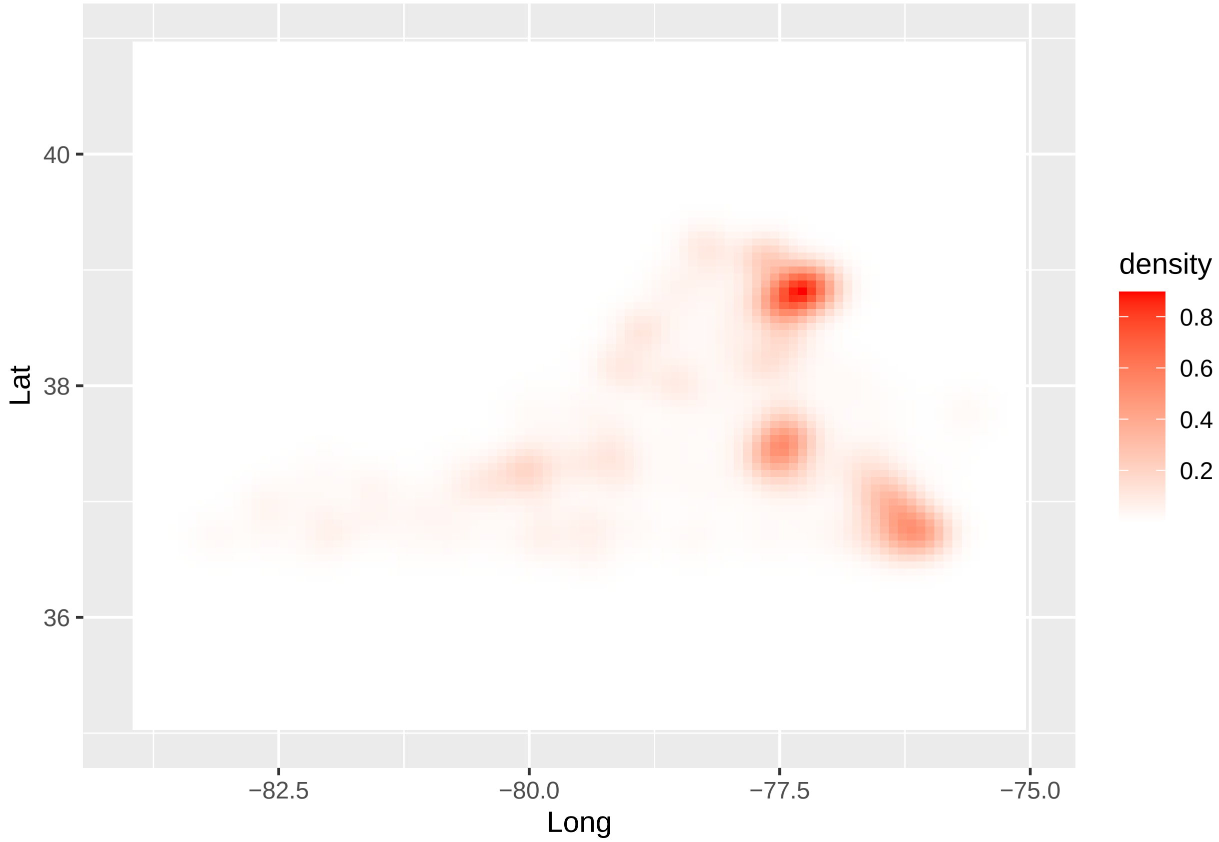
Virginia Heat Map

I will create a heat map for the number of cases in Virginia on November 19, 2021.

Creating a heat map:

```
# Data frame for heat map
va_yest = data.frame(matrix(ncol = 2))
colnames(va_yest) = c("Long","Lat")
va_yest = data.frame(Long = rep(va$Long_,va[,677]/100), Lat = rep(va$Lat,va[,677]/100))

# Heat map
ggplot() + stat_density2d(data = va_yest, aes(x = Long, y = Lat, fill = ..density..), geom = 'tile',
                          contour = F) + scale_fill_gradient(low = "white",high = "red") +
                          xlim(-84,-75) + ylim(35,41)
```



Looking at the heat map, most of the cases in Virginia are on the eastern border of the state, namely Northern Virginia as well as the Virginia Beach area. This makes sense as populations in those areas are much greater than the rest of the state.

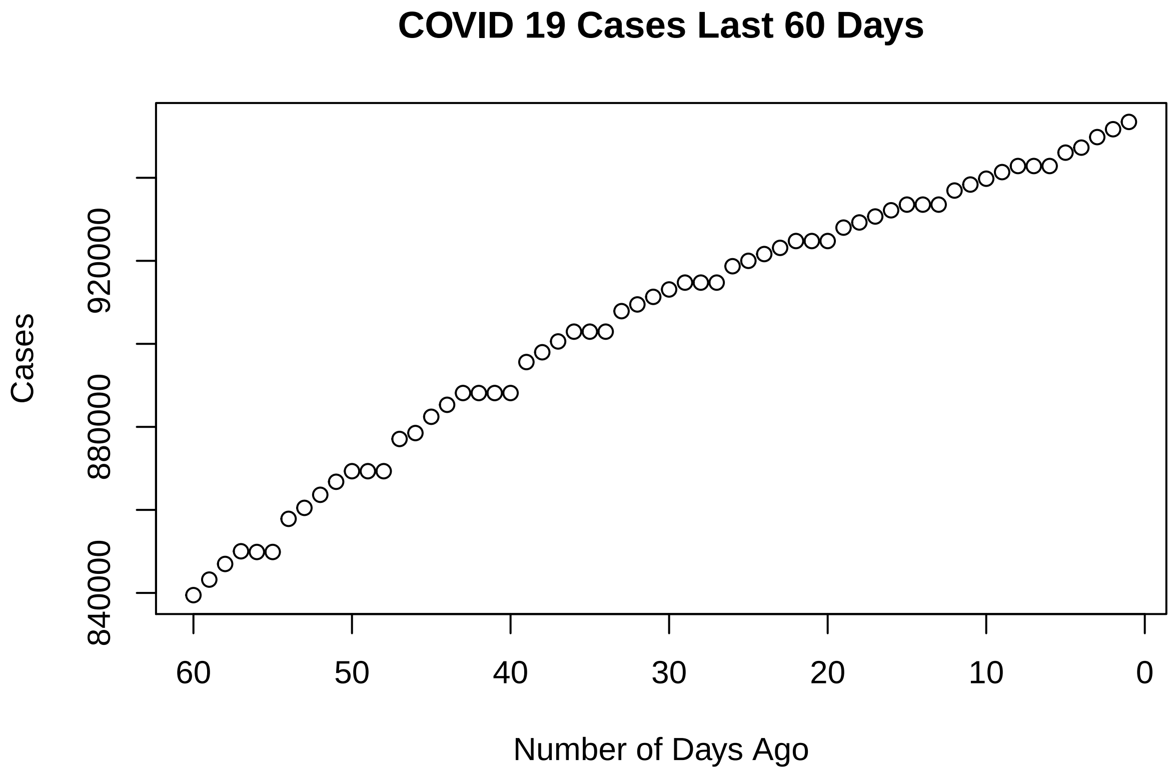
Recent COVID Trend

I want to look at trends in cases in Virginia over the past 60 days and try to fit a model to the data.

Plot for last two months:

```
# Data frame for last 60 days
days = 60
month = data.frame(matrix(ncol = 2, nrow = days))
colnames(month) = c("Date","Cases")
month$Date = days:1
month$Cases = colSums(va[,(ncol(va)-days+1):ncol(va)])

plot(x = month$Date, y = month$Cases, xlim = rev(range(month$Date)),
     main = "COVID 19 Cases Last 60 Days", xlab = "Number of Days Ago", ylab = "Cases")
```



Looking at the plot, it appears to be mostly linear, which means the number of new cases in Virginia are about the same every day.

Modeling the data:

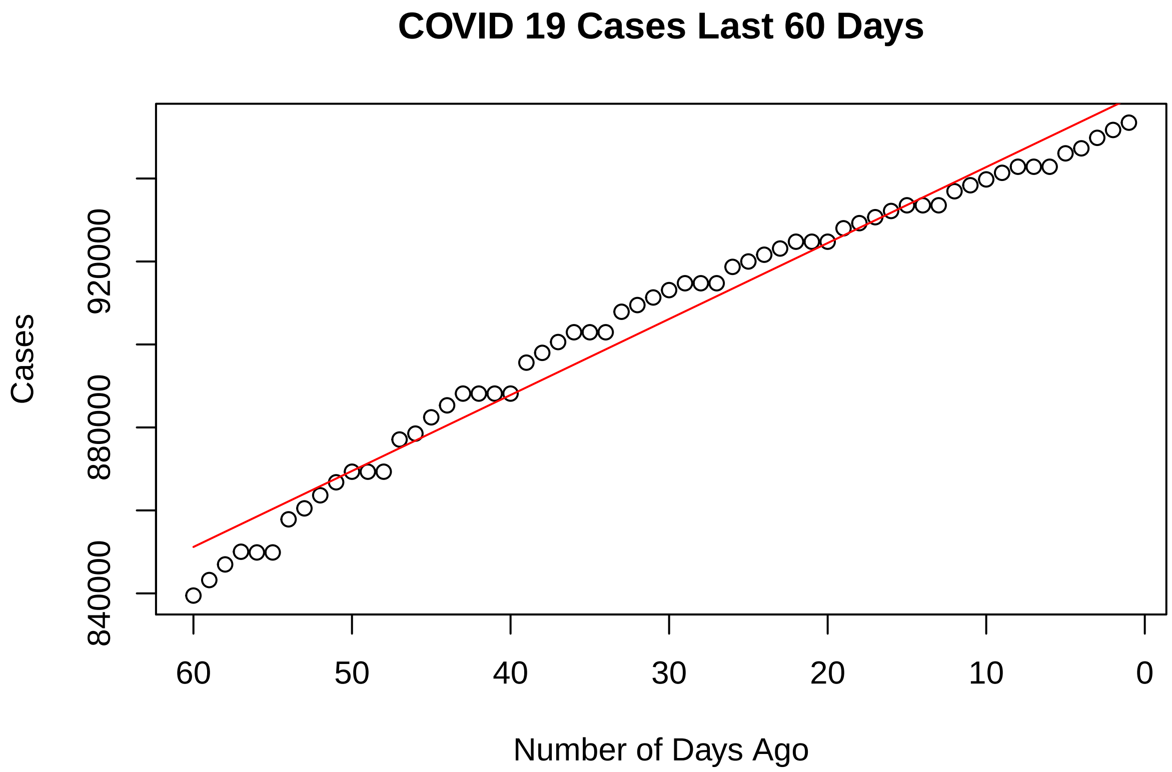
```
model = lm(month$Cases~month$Date)
summary(model)
```

```
##
## Call:
## lm(formula = month$Cases ~ month$Date)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11717  -3752    371   4560   7794
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 961072.66    1388.33  692.25   <2e-16 ***
## month$Date   -1831.35      39.58  -46.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5310 on 58 degrees of freedom
## Multiple R-squared:  0.9736, Adjusted R-squared:  0.9732
## F-statistic:  2141 on 1 and 58 DF,  p-value: < 2.2e-16
```

The linear model for predicting the number of new cases in Virginia each day is $961072.66 - 1831.35x$ where x is the number of days ago. The Adjusted R-squared value for this model is 0.9732, meaning it is a great fit for the data.

Fitting the model:

```
pred = predict(model, x = 100:1)
plot(x = month$Date, y = month$Cases, xlim = rev(range(month$Date)),
     main = "COVID 19 Cases Last 60 Days", xlab = "Number of Days Ago", ylab = "Cases")
lines(rev(pred), type = "l", col = "red")
```



Adding the model to the plot confirms the model is a good fit for the data.

Conclusion

In conclusion, most of the cases in Virginia are in Northern Virginia (outside of DC) and the Virginia Beach area. You can accurately predict the number of cumulative cases in Virginia in the last 60 days with the formula $961072.66 - 1831.35x$ where x is the number of days ago. This linear model has an Adjusted R-squared value of 0.9732, which means the model is a really good fit for the data. If you wanted to predict the number of total cases in the future you could use a negative value for x in the equation, i.e. 5 days in the future would be x=-5. If you wanted the number of new cases, subtract the number of previous cases from the output of the equation.

Possible bias in this project is that this model only applies to the state of Virginia as a whole; it can be very different if you were looking at another state or even one individual county in Virginia. Another bias could be that the model was using data from the past 60 days. Using data from the past year will likely yield a very different model that may not even be linear.