Covid Data Notebook

Goal

The purpose of this notebook is to show the analysis of Covid data from the Virginia Department of Health

```
library(ggplot2)
```

Preparing the data

We read in raw data from a CSV as a dataframe in R. This data is provided on the Virginia Department of Health website:

https://www.vdh.virginia.gov/coronavirus/

```
rawData <- read.csv("VDH-COVID-19-PublicUseDataset-EventDate.csv", stringsAsFactors = FALSE)

# Fix the date field
rawData$FixedDate <- as.Date(rawData$Event.Date,format="%m/%d/%Y")
```

Now we build two new data frames: one with confirmed cases and one with probable cases and then merge them together. Before we merge them we change the column headings of the 2 dataframes. We also need to replace all of the NA values with 0's.

```
# Build a data frame with just the confirmed cases
confirmedCases <- subset(rawData, Case.Status == "Confirmed")

# Build a data frame with just the probable cases
probableCases <- subset(rawData, Case.Status == "Probable")

# Change the column headings to reflect the two data sets
col_headings          <- c('Event.Date',
                           'Health.Planning.Region',
                           'Case.Status - Confirmed',
                           'Cases-confirmed',
                           'Hospitalizations-confirmed',
                           'Deaths-confirmed',
                           'FixedDate')
col_headings2         <- c('Event.Date',
                           'Health.Planning.Region',
                           'Case.Status - Probable',
                           'Cases-probable',
                           'Hospitalizations-probable',
                           'Deaths-probable',
                           'FixedDate')
names(confirmedCases) <- col_headings
names(probableCases)  <- col_headings2

# Merge the two dataframes and build the new columns combining the data
mergedData            <- merge(confirmedCases,
                               probableCases,
```

```
                                by = c("FixedDate","Health.Planning.Region"),
                                all.x = TRUE,
                                all.y = TRUE)
mergedData[is.na(mergedData)] <- 0 # Replace all of the NA's with 0
```

Next we build a new column with the total of cases, deaths and hospitalizations.

```
mergedData$totalCases         <- mergedData$'Cases-confirmed' +
  mergedData$'Cases-probable'
mergedData$totalHospitalized  <- mergedData$`Hospitalizations-confirmed` +
  mergedData$`Hospitalizations-probable`
mergedData$totalDeaths        <- mergedData$`Deaths-confirmed` +
  mergedData$`Deaths-probable`
```

Build 5 new dataframes for the different regions in Virginia:

```
casesEastern   <- subset(mergedData, mergedData$Health.Planning.Region == "Eastern")
casesCentral   <- subset(mergedData, mergedData$Health.Planning.Region == "Central")
casesNorthern  <- subset(mergedData, mergedData$Health.Planning.Region == "Northern")
casesNorthwest  <- subset(mergedData, mergedData$Health.Planning.Region == "Northwest")
casesSouthwest  <- subset(mergedData, mergedData$Health.Planning.Region == "Southwest")
```

Calculating the Totals for the State

In order to get an idea of overall statistics for the state we will total the numbers in the raw data set.

```
totalCases            = sum(rawData$'Number.of.Cases')
totalHospitalizations = sum(rawData$'Number.of.Hospitalizations')
totalDeaths           = sum(rawData$'Number.of.Deaths')
```

Totals for the State:

- Cases: 67964
- Hospitalizations: 6223
- Deaths: 1896

Calculate heard immunity rate of Eastern Virginia

The CDC has released the infection fatility rate (IFR) as 0.26%. Using the total number of deaths in eastern Virginia we can can calculate the total number of people who have had COVID in this region.

```
numberHadit = sum(casesEastern$'totalDeaths') / .0026
```

Individuals who have had Covid: 89230.77

Using the number of individuals who have had COVID divided by the total population of the Eastern region, we can calculate the heard immunity threshold of the region.

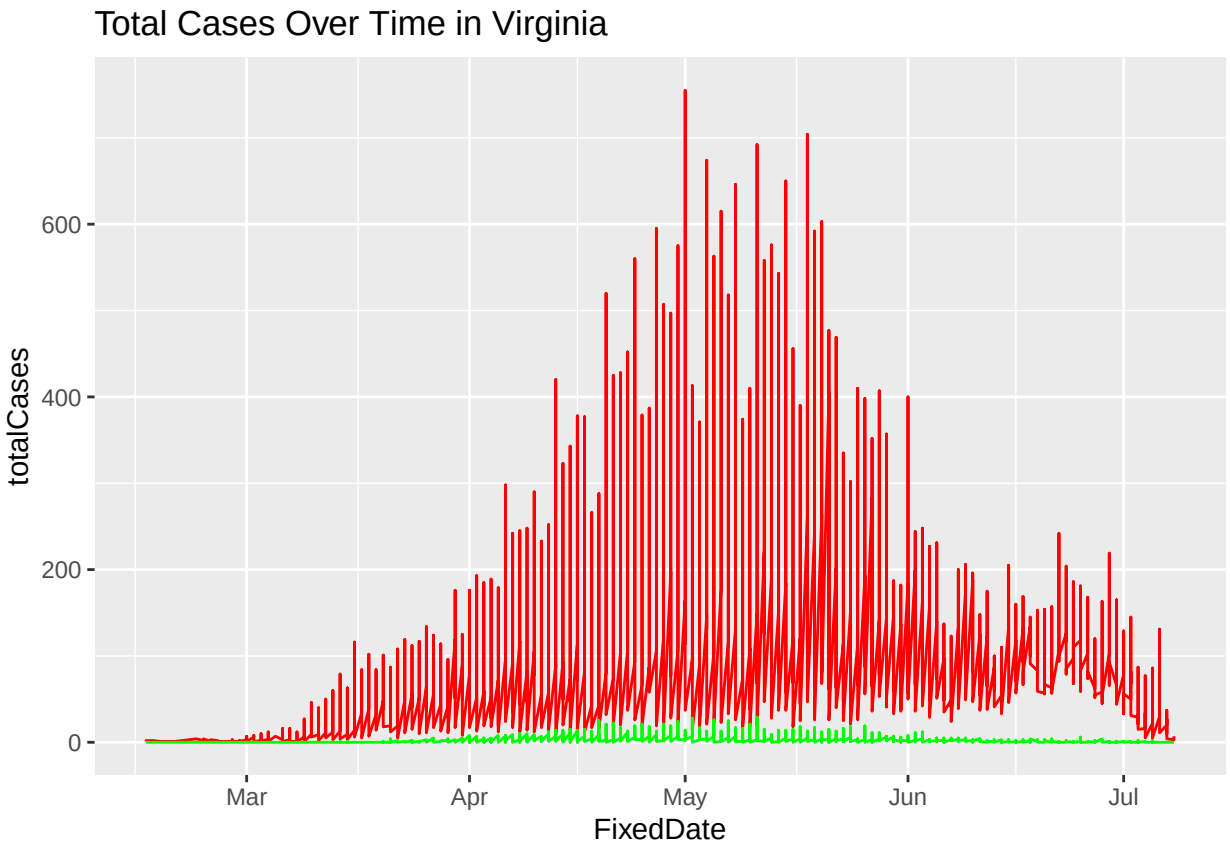
Total Population in the Eastern Planning Region (as of 2016): 1,854,806. Percentage of individuals who have had Covid: 4.8107872%

It looks like we are a couple of percentage points below the threshold needed for herd immunity.

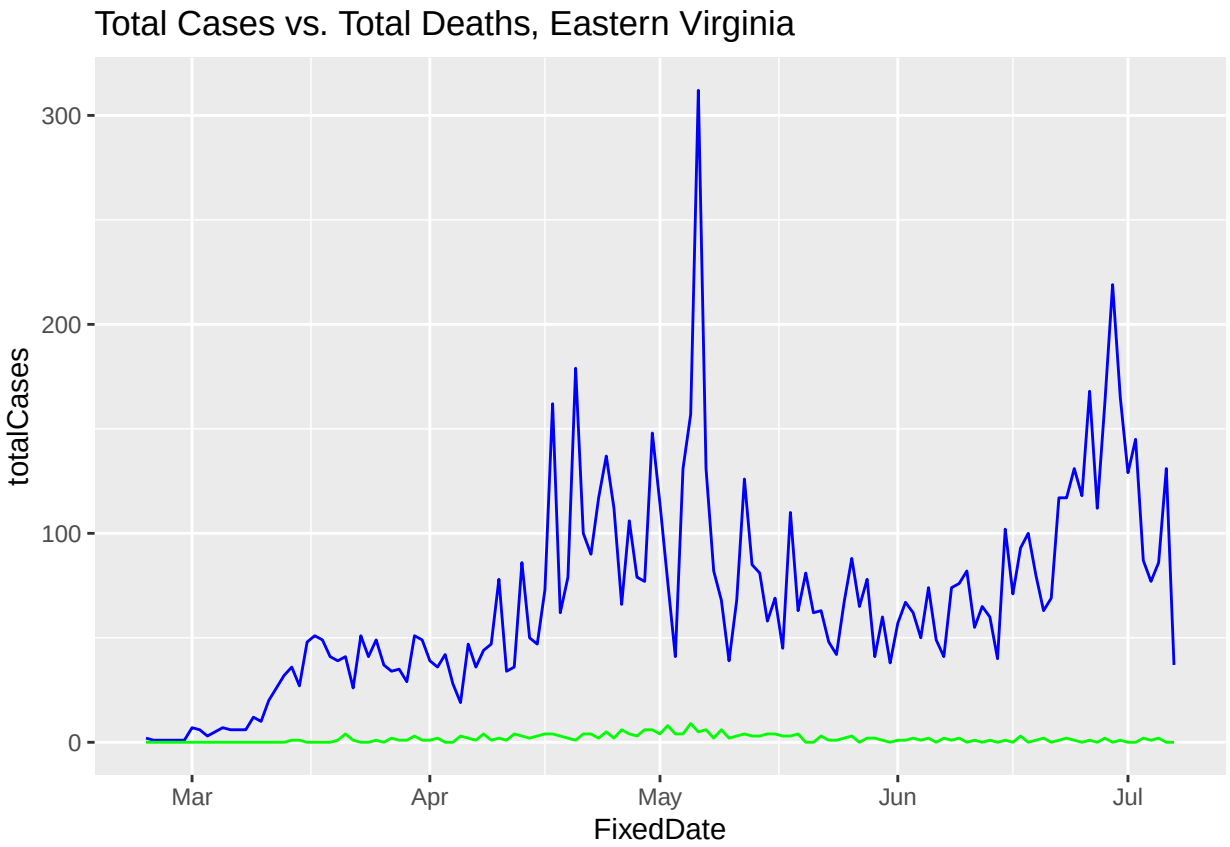
Total Cases Over Time

Some additional data gleaned from the data set:

```
totalCases_vs_Deaths_Plot = ggplot() + geom_line(data = mergedData, aes(x = FixedDate, y = totalCases),
print(totalCases_vs_Deaths_Plot + ggtitle("Total Cases Over Time in Virginia"))
```

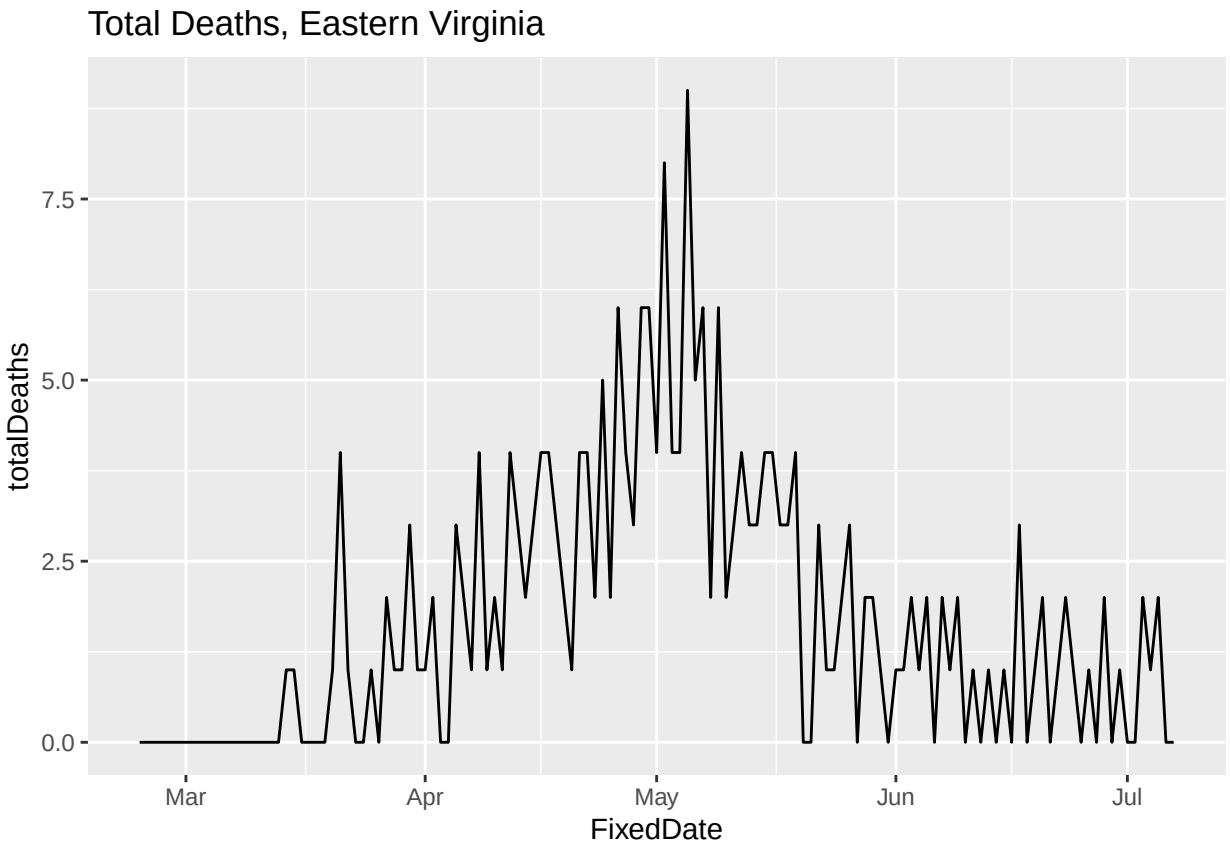



```
totalCases_Eastern_Plot <- ggplot() + geom_line(data = casesEastern, aes(x = FixedDate, y = totalCases)
  geom_line(data = casesEastern, aes(x = FixedDate, y = totalDeaths), color = 'green')
print(totalCases_Eastern_Plot + ggtitle("Total Cases vs. Total Deaths, Eastern Virginia"))
```



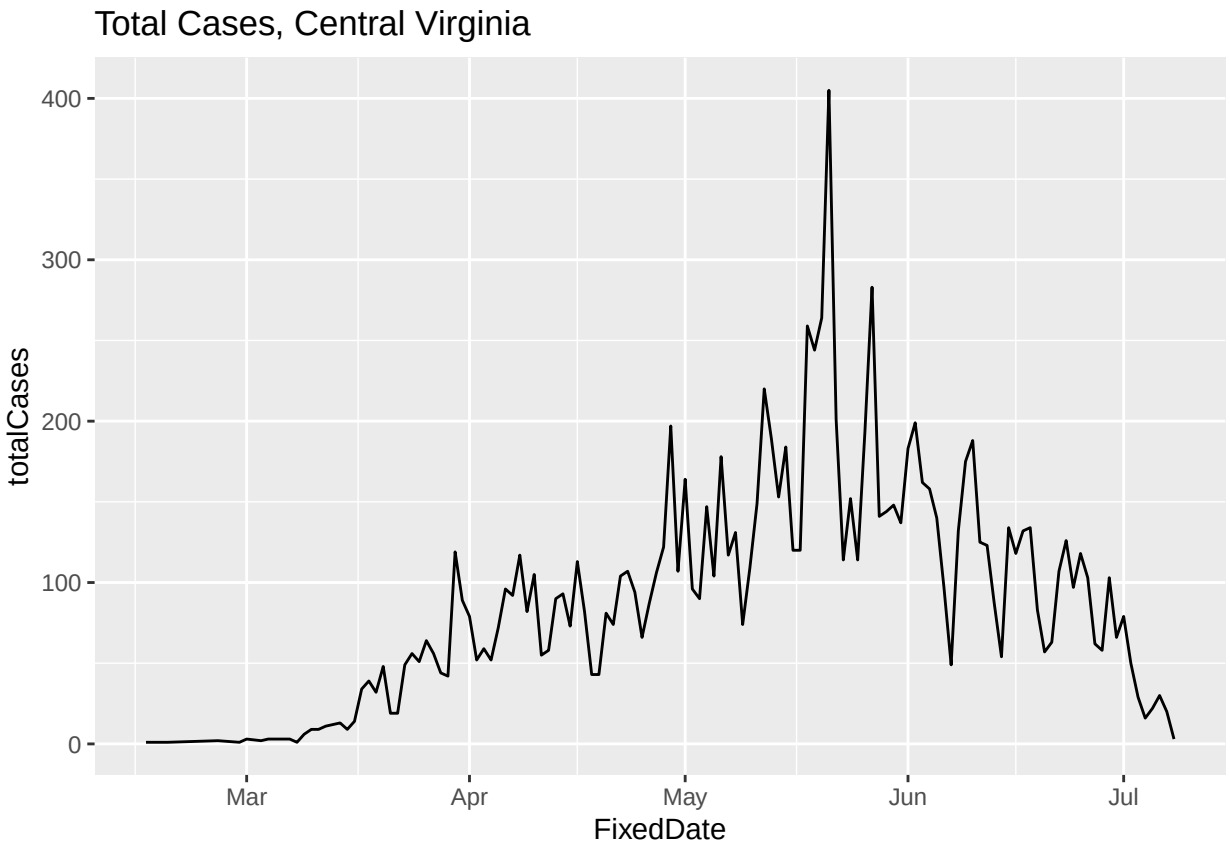
We can also seperate out the graph of deaths related to COVID to more closely examine the trend:

```
totalDeaths_Eastern_Plot = ggplot(data = casesEastern, aes(x = FixedDate, y=totalDeaths)) + geom_line()
print(totalDeaths_Eastern_Plot + ggtitle("Total Deaths, Eastern Virginia"))
```



Some additional plots

```
totalCases_Central_Plot = ggplot(data = casesCentral, aes(x = FixedDate, y = totalCases)) + geom_line()
print(totalCases_Central_Plot + ggtitle("Total Cases, Central Virginia"))
```


Total Cases, Central Virginia
totalCases
400
300
200
100
0
Mar
Apr
May
Jun
Jul
FixedDate