COVID-19 Project: Mass. Spread of COVID-19

This project uses the packages tidyverse, which now contains lubridate. Please make sure it is installed before trying to knit this project.

COVID-19 was tracked on a global scale likely with more detail than any pandemics of the past. We are going to look at the different counties of Massachusetts and plot both the number of cases and deaths over time. Additionally, we will try to model and predict the number of cases and deaths in the state overall.

Question of interest: Can we model and predict relationship between the cases and deaths over time in Massachusetts?

This project is broken down into 4 steps.

1. Get the data
2. Convert the data into something useful
3. Create a model and present the data
4. Identify possible biases

Step 1: Import data in a way that's reproducible

- Install all the packages used in this project and load the corresponding libraries
- Use the tidyverse package to read the csv directly from the source
- The data source is provided by Johns Hopkins University on GitHub at: `https://github.com/CSSEGISandData/COVID-`
- The exact directory for the download is currently: `https://raw.githubusercontent.com/CSSEGISandData/COVID-19/` and the individual files are: time_series_covid19_confirmed_US.csv, and time_series_covid19_deaths_US.csv.

A. Install and load the libraries

```
##
## The downloaded binary packages are in
##  /var/folders/h4/0yj9g1kx5bvgbvsm3b1209tr0000gn/T//RtmpD2P0Mu/downloaded_packages
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

B. - D. Use the Tidyverse package to Read the CSV directly from the data sources.
https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_
https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_

```
## Rows: 3342 Columns: 1154
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): 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.
## Rows: 3342 Columns: 1155
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): 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.
```

Step 2. Convert the Data into something usefull.

Preview the data.

```
## # A tibble: 6 x 1,154
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # i 1,145 more variables: Long_ <dbl>, Combined_Key <chr>, `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>, ...

## # A tibble: 6 x 1,155
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # i 1,146 more variables: Long_ <dbl>, Combined_Key <chr>, 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>, ...
```

```
cases <- confirmed_us %>%
  pivot_longer(cols = -c(UID:Combined_Key), names_to = "date", values_to = "Cases")%>%
  select(-c(iso2, iso3, code3, FIPS, UID, Country_Region))%>%
  mutate(date = mdy(date))

summary(cases)
```

Pivot the data so each day isn't a separate column.

```
##     Admin2          Province_State          Lat             Long_
##  Length:3819906     Length:3819906     Min.   :-14.27   Min.   :-174.16
##  Class :character   Class :character   1st Qu.: 33.90   1st Qu.: -97.81
##  Mode  :character   Mode  :character   Median : 38.01   Median : -89.49
##                                        Mean   : 36.72   Mean   : -88.64
##                                        3rd Qu.: 41.58   3rd Qu.: -82.31
##                                        Max.   : 69.31   Max.   : 145.67
##  Combined_Key            date                Cases
##  Length:3819906     Min.   :2020-01-22   Min.   :  -3073
##  Class :character   1st Qu.:2020-11-02   1st Qu.:    330
##  Mode  :character   Median :2021-08-15   Median :   2272
##                     Mean   :2021-08-15   Mean   :  14088
##                     3rd Qu.:2022-05-28   3rd Qu.:   8159
##                     Max.   :2023-03-09   Max.   :3710586
```

```
deaths <- deaths_us %>%
  pivot_longer(cols = -c(UID:Population), names_to = "date", values_to = "deaths")%>%
  select(-c(iso2, iso3, code3, FIPS, UID, Country_Region))%>%
  mutate(date = mdy(date))

summary(deaths)
```

```
##     Admin2          Province_State          Lat             Long_
##  Length:3819906     Length:3819906     Min.   :-14.27   Min.   :-174.16
##  Class :character   Class :character   1st Qu.: 33.90   1st Qu.: -97.81
##  Mode  :character   Mode  :character   Median : 38.01   Median : -89.49
##                                        Mean   : 36.72   Mean   : -88.64
##                                        3rd Qu.: 41.58   3rd Qu.: -82.31
##                                        Max.   : 69.31   Max.   : 145.67
##  Combined_Key         Population           date                deaths
##  Length:3819906     Min.   :       0   Min.   :2020-01-22   Min.   :  -82.0
##  Class :character   1st Qu.:    9917   1st Qu.:2020-11-02   1st Qu.:    4.0
##  Mode  :character   Median :   24892   Median :2021-08-15   Median :   37.0
##                     Mean   :   99604   Mean   :2021-08-15   Mean   :  186.9
##                     3rd Qu.:   64979   3rd Qu.:2022-05-28   3rd Qu.:  122.0
##                     Max.   :10039107   Max.   :2023-03-09   Max.   :35545.0
```

```
#summary(mass_density)
```

Initial Filtering and Joins Filter the other states' data out, then join the data for Cases and Deaths in Massachusetts.

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

More Mutation and Filtering Because of the population correlation, mutate the Mass rows to add deaths per 1000 and cases per 1000. While we're at it, create a couple other date formats and remove the ... from Long. Additionally, we will filter for 0 Population as this will cause divide by zero errors.

```
##     Admin2          Province_State          Lat            Long_
##  Length:16002       Length:16002       Min.   :41.29   Min.   :-73.21
##  Class :character   Class :character   1st Qu.:41.79   1st Qu.:-72.59
##  Mode  :character   Mode  :character   Median :42.24   Median :-71.16
##                                        Mean   :42.11   Mean   :-71.47
##                                        3rd Qu.:42.37   3rd Qu.:-70.81
##                                        Max.   :42.67   Max.   :-70.09
##  Combined_Key            date                Cases          Population
##  Length:16002       Min.   :2020-01-22   Min.   :     0   Min.   :  11399
##  Class :character   1st Qu.:2020-11-02   1st Qu.:  1475   1st Qu.: 124944
##  Mode  :character   Median :2021-08-15   Median : 23196   Median : 493787
##                     Mean   :2021-08-15   Mean   : 65074   Mean   : 492322
##                     3rd Qu.:2022-05-28   3rd Qu.:104131   3rd Qu.: 789034
##                     Max.   :2023-03-09   Max.   :437431   Max.   :1611699
##      deaths      deaths_per_k     cases_per_k       month_year
##  Min.   :   0   Min.   :0.0000   Min.   :  0.000   Length:16002
##  1st Qu.:  77   1st Qu.:0.7267   1st Qu.:  7.139   Class :character
##  Median : 683   Median :1.8719   Median : 75.010   Mode  :character
##  Mean   :1028   Mean   :1.7397   Mean   :106.850
##  3rd Qu.:1794   3rd Qu.:2.6503   3rd Qu.:203.338
##  Max.   :4822   Max.   :4.5843   Max.   :368.944
##       Lng              month
##  Min.   :-73.21   Min.   : 1.000
##  1st Qu.:-72.59   1st Qu.: 3.000
##  Median :-71.16   Median : 6.000
##  Mean   :-71.47   Mean   : 6.335
##  3rd Qu.:-70.81   3rd Qu.: 9.000
##  Max.   :-70.09   Max.   :12.000
```

```
## # A tibble: 6 x 14
##   Admin2     Province_State   Lat Long_ Combined_Key date       Cases Population
##   <chr>      <chr>          <dbl> <dbl> <chr>        <date>     <dbl>      <dbl>
## 1 Barnstable Massachusetts   41.7 -70.3 Barnstable,~ 2020-01-22     0     212990
## 2 Barnstable Massachusetts   41.7 -70.3 Barnstable,~ 2020-01-23     0     212990
## 3 Barnstable Massachusetts   41.7 -70.3 Barnstable,~ 2020-01-24     0     212990
## 4 Barnstable Massachusetts   41.7 -70.3 Barnstable,~ 2020-01-25     0     212990
## 5 Barnstable Massachusetts   41.7 -70.3 Barnstable,~ 2020-01-26     0     212990
## 6 Barnstable Massachusetts   41.7 -70.3 Barnstable,~ 2020-01-27     0     212990
## # i 6 more variables: deaths <dbl>, deaths_per_k <dbl>, cases_per_k <dbl>,
## #   month_year <chr>, Lng <dbl>, month <dbl>
```

Step 3. Analyze the data, create a model and present everything

Quick Correlation Check To begin with, we do a quick correlation analysis to try to get a better sense of the relationship between the columns of data. I'm looking for correlations between the deaths, cases, and the population.

```
## [1] "Deaths & Population: "
```

```
## [1] 0.7976191
```

```
## [1] "Cases & Population: "
```

```
## [1] 0.6327786
```

```
## [1] "Cases & Deaths: "
```

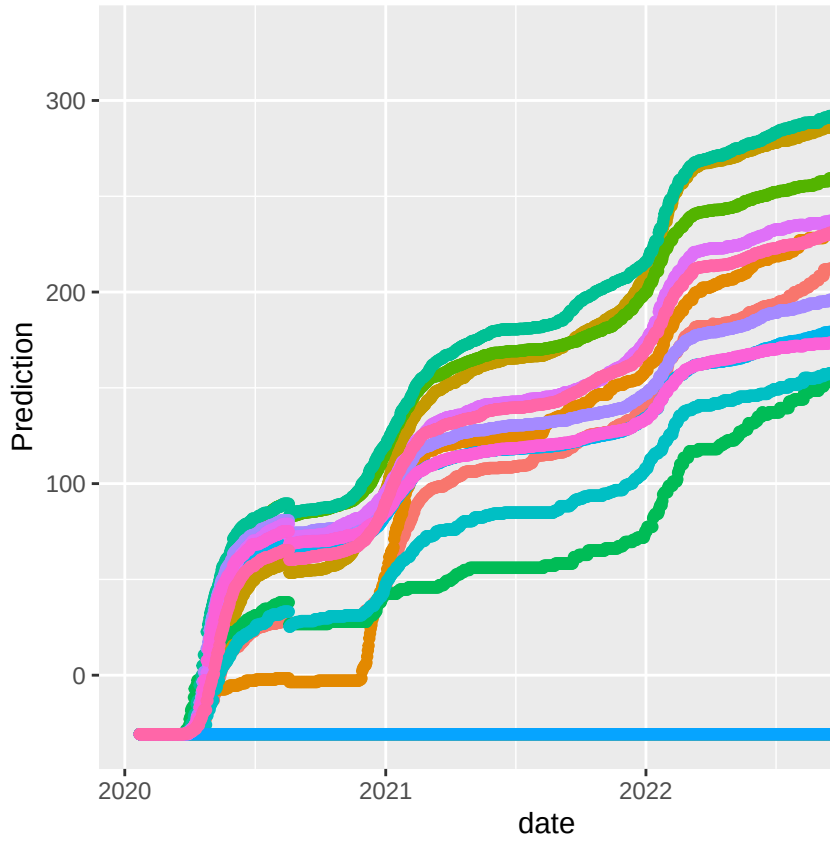
```
## [1] 0.924681
```

```
## [1] "Cases/1000 & Deaths/1000: "
```

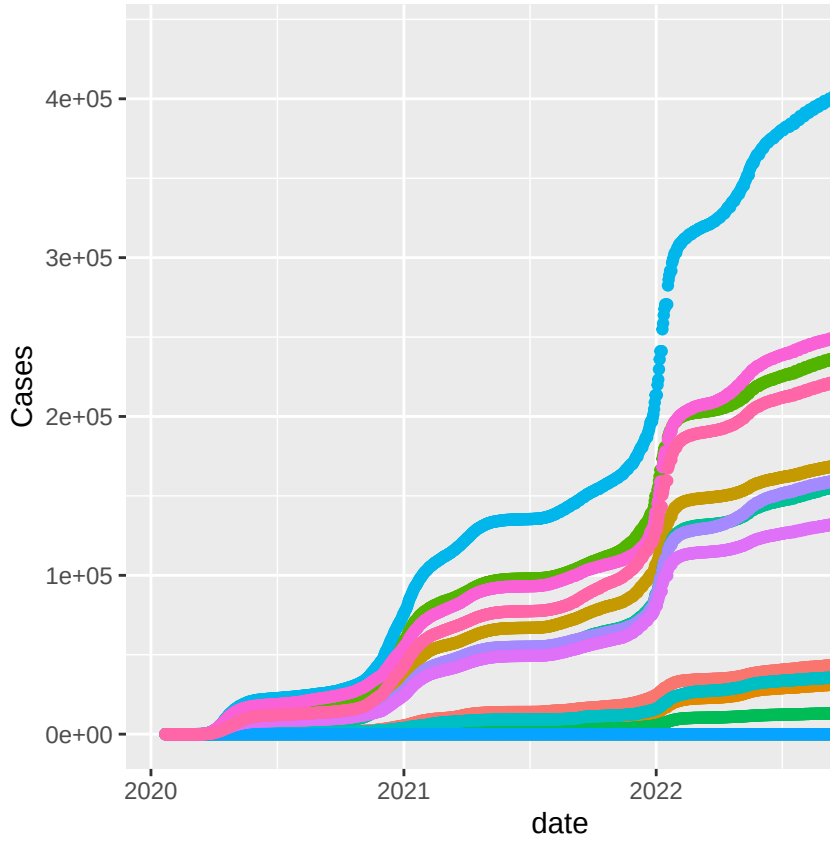
```
## [1] 0.91824
```

Build a model. We will build a model based on cases per 1000 and deaths per 1000, output the summary, then add the predictions to the Mass. county data.

```
##
## Call:
## lm(formula = cases_per_k ~ deaths_per_k, data = Mass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.130 -38.676   6.675  30.785 146.494
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -30.7845     0.5764  -53.41   <2e-16 ***
## deaths_per_k  79.1149     0.2698  293.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.33 on 16000 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8432
## F-statistic: 8.602e+04 on 1 and 16000 DF,  p-value: < 2.2e-16
```



Plot Predictions for the individual Counties.



Plot the Actual County Data for comparison

Group the Mass. data by county.

```
##     Admin2            Max_Deaths      Total_Deaths       Max_Cases
##  Length:14          Min.   :   0.0   Min.   :      0   Min.   :     0
##  Class :character   1st Qu.: 459.8   1st Qu.: 276102   1st Qu.: 36522
##  Mode  :character   Median :2050.0   Median :1382960   Median :157826
##                     Mean   :1735.9   Mean   :1175195   Mean   :143892
##                     3rd Qu.:2498.8   3rd Qu.:1665867   3rd Qu.:226117
##                     Max.   :4822.0   Max.   :3378924   Max.   :437431
##   Total_Cases          Population
##  Min.   :        0   Min.   :  11399
##  1st Qu.: 15948862   1st Qu.: 133916
##  Median : 81161396   Median : 493787
##  Mean   : 74379791   Mean   : 492322
##  3rd Qu.:118428435   3rd Qu.: 768469
##  Max.   :220834357   Max.   :1611699
```

```
## # A tibble: 6 x 6
##   Admin2     Max_Deaths Total_Deaths Max_Cases Total_Cases Population
##   <chr>           <dbl>        <dbl>     <dbl>       <dbl>      <dbl>
## 1 Barnstable        785       447242     49617    23514236     212990
## 2 Berkshire         480       276606     35456    15223406     124944
## 3 Bristol          2555      1619263    182344    98339486     565217
## 4 Dukes               0            0         0           0      17332
## 5 Essex            3272      2235421    256987   140031284     789034
## 6 Franklin          198       108143     14736     6453660      70180
```

Conclusion

Our prediction managed to capture the low correlation of Nantucket County, and followed the general pattern of increase over time. Unfortunately, it didn't entirely reflect the extreme increase of Middlesex County. It would be worth investigating what caused that county to stand out.

Step 4: Add Bias Identification

Data Bias

Massachusetts is one state in the United States, out of the entire planet may not be an accurate sample. It is difficult to say how accurate the data itself is or how consistent it is from county to county. Furthermore, as more was known about the COVID-19 virus, methods of accurately identifying cases and deaths are likely to improve. This may skew the data.

Personal Bias

On a personal note, I chose to examine only Massachusetts. I did so based on the basis of my perception of Massachusetts as a place with cutting edge medicine and unbiased data, with both rural and very urban areas. This may be have be completely wrong.