Covid Data Analysis

Swapnil Bhatta

2022-12-01

Data Retrieval and Description

JHU CSSE COVID-19 Dataset is present in their github repo -> https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

The data contains daily case reports for COVID data both throughout the world and specifically for US states. All time stamps are in UTC (GMT+0) and the data is updated daily. A detailed description of the columns present are given in the repo itself along with any flags and data collecting methodologies.

The URLs will be read into data frames one at a time, and then pivoted to tidy up the fields.

```
US_confirmed <- read_csv(urls[3]) %>%
    pivot_longer(cols = -(UID:Combined_Key), names_to = "Date", values_to = "Confirmed_cases") %>%
    select(Admin2:Confirmed_cases) %>%
    mutate(Date = mdy(Date))
```

```
## Rows: 3342 Columns: 1056
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1050): 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.
```

```
US_deaths <- read_csv(urls[4]) %>%
    pivot_longer(cols = -(UID:Population), names_to = "Date", values_to ="Deaths") %>%
    select(Admin2:Deaths) %>%
    mutate(Date = mdy(Date))
```

```
## Rows: 3342 Columns: 1057
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1051): 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.
```

We now join the cases and deaths in the US

```
US <-  US_deaths %>%
    full_join(US_confirmed,
              by = c("Combined_Key", "Date",
                   "Admin2", "Province_State",
                   "Country_Region")) %>%
```

```
    rename(Long = Long_.x, Lat = Lat.x)  %>%
    select(Admin2, Province_State, Country_Region,
           Lat, Long, Population, Date, Confirmed_cases, Deaths)
```

Statement of Interest

We wish to look at the number of cases and deaths in Colorado, US. We also want to show the top 10 counties in the state with the highest number of confirmed cases to date.

```
# Filter out Colorado Data
colorado <- US %>% filter(Province_State == "Colorado") %>%
  select(Admin2, Lat, Long, Province_State, Date, Confirmed_cases, Deaths)
```

```
summary(colorado)
```

```
##     Admin2               Lat             Long        Province_State
##  Length:68970       Min.   : 0.00   Min.   :-108.6   Length:68970
##  Class :character   1st Qu.:37.90   1st Qu.:-106.9   Class :character
##  Mode  :character   Median :38.87   Median :-105.4   Mode  :character
##                     Mean   :37.76   Mean   :-102.3
##                     3rd Qu.:39.86   3rd Qu.:-103.8
##                     Max.   :40.88   Max.   :   0.0
##       Date            Confirmed_cases      Deaths
##  Min.   :2020-01-22   Min.   :     0   Min.   :   0.0
##  1st Qu.:2020-10-09   1st Qu.:   117   1st Qu.:   1.0
##  Median :2021-06-27   Median :  1015   Median :  12.0
##  Mean   :2021-06-27   Mean   : 10899   Mean   : 109.8
##  3rd Qu.:2022-03-15   3rd Qu.:  4166   3rd Qu.:  54.0
##  Max.   :2022-12-01   Max.   :230067   Max.   :1845.0
```

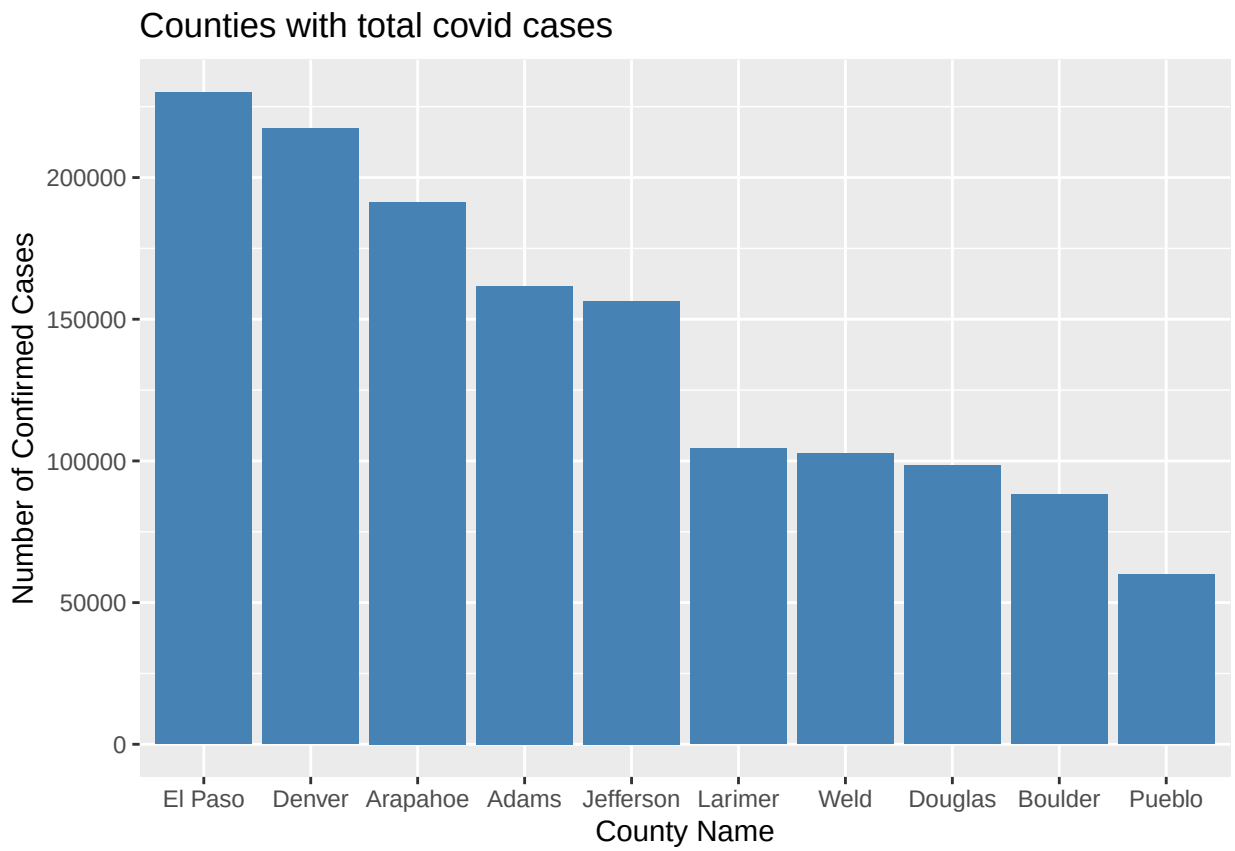
We see that the minimum and maximum Lat and Long values are wrong. A lat and long of (0, 0) would not be in Colorado, US. These will be filtered out.

```
colorado <- colorado %>% filter(Lat != 0 | Long != 0)
```

Top Counties with Deaths

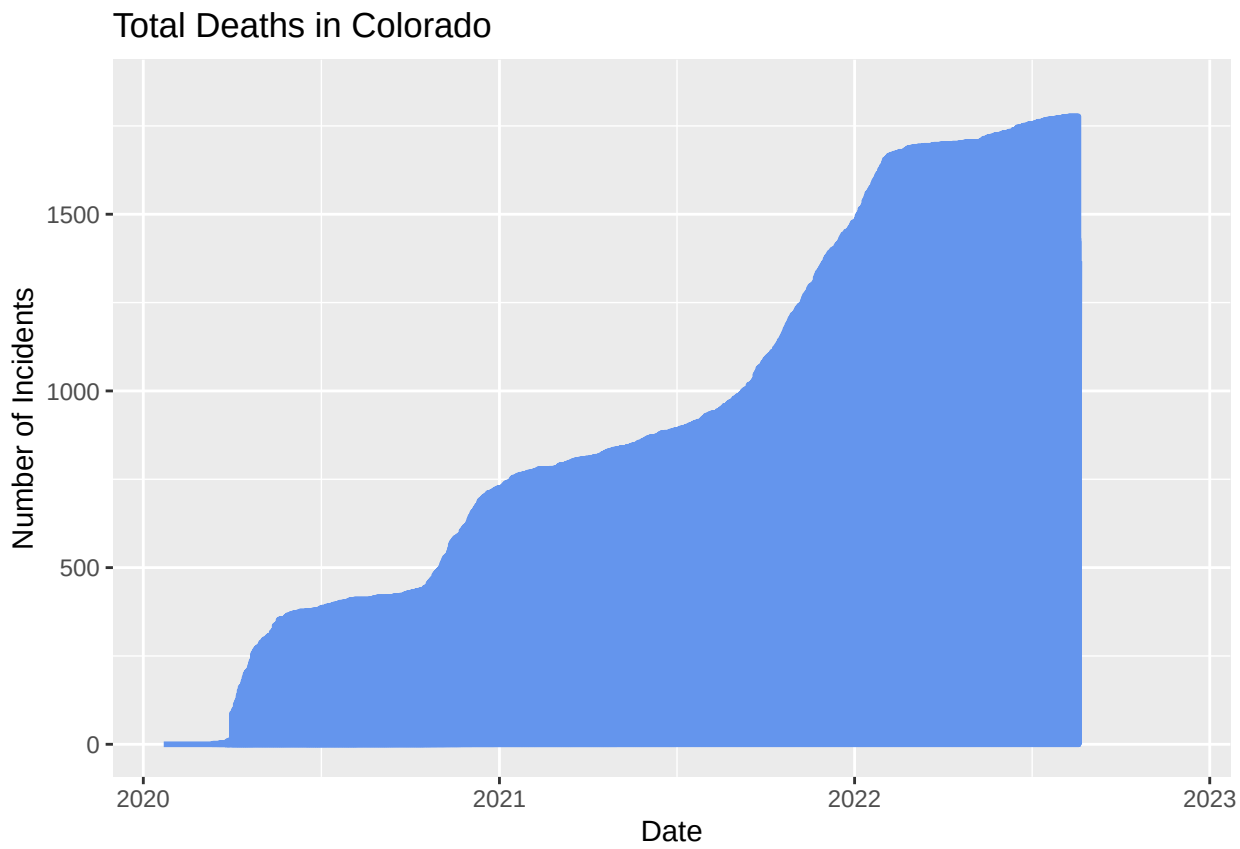
```
colorado_top <- colorado %>% group_by(Admin2) %>%
            summarize(Confirmed_cases = max(Confirmed_cases),
                      Deaths = max(Deaths)) %>%
            ungroup() %>%
            arrange(desc(Confirmed_cases)) %>%
            slice(1:10)

ggplot(data=colorado_top, aes(x=reorder(Admin2, -Confirmed_cases), y=Confirmed_cases)) +
     geom_bar(stat="identity", fill="steelblue") +
     labs(x = "County Name", y = "Number of Confirmed Cases", title='Counties with total covid cases')
```



We now look at the total number of deaths for the state

```
ggplot(colorado, aes(x=Date, y=Deaths)) +
            geom_line(color = "cornflowerblue", linewidth=1) +
            labs(x = "Date", y = "Number of Incidents", title='Total Deaths in Colorado')
```

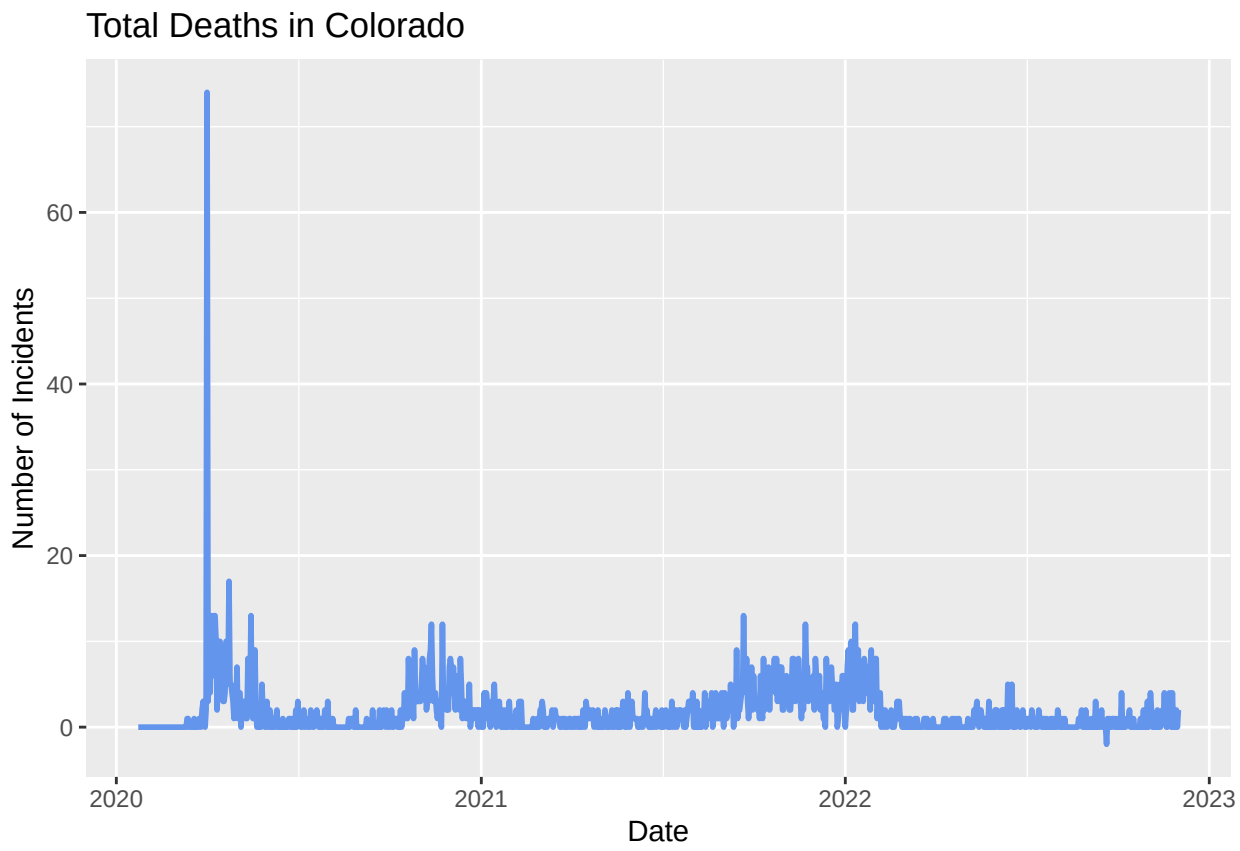


We can use a lag difference to generate just the new cases (both confirmed cases and death values)

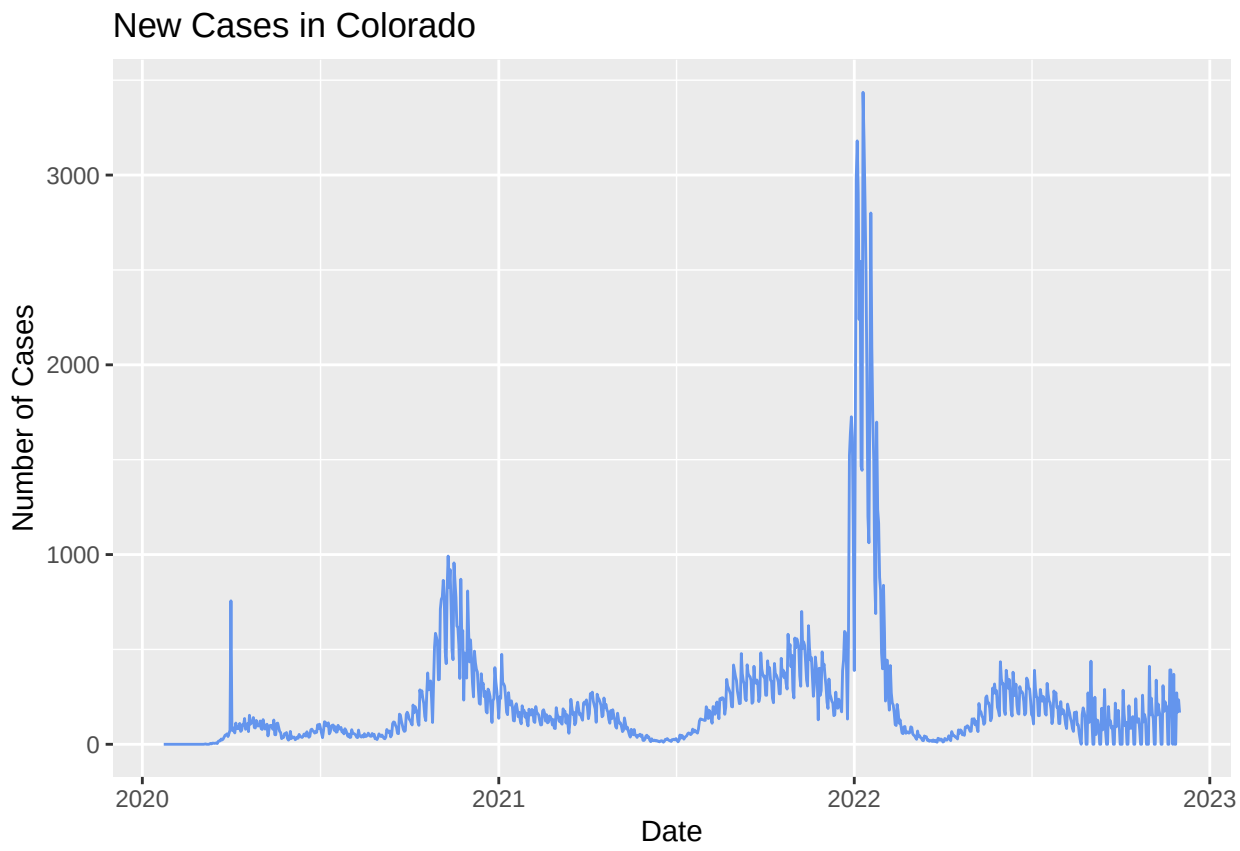
```
colorado_aggregated <- colorado %>% group_by(Date) %>%
                        summarize(Confirmed_cases = max(Confirmed_cases),
                                  Deaths = max(Deaths)) %>%
                        ungroup() %>%
                        mutate(new_cases = Confirmed_cases - lag(Confirmed_cases)) %>%
                        mutate(new_deaths = Deaths - lag(Deaths))

# First day lag is NA
colorado_aggregated <-  colorado_aggregated[-1,]
```

```
#New Deaths
ggplot(colorado_aggregated, aes(x=Date, y=new_deaths)) +
            geom_line(color = "cornflowerblue", linewidth=1) +
            labs(x = "Date", y = "Number of Incidents", title='Total Deaths in Colorado')
```



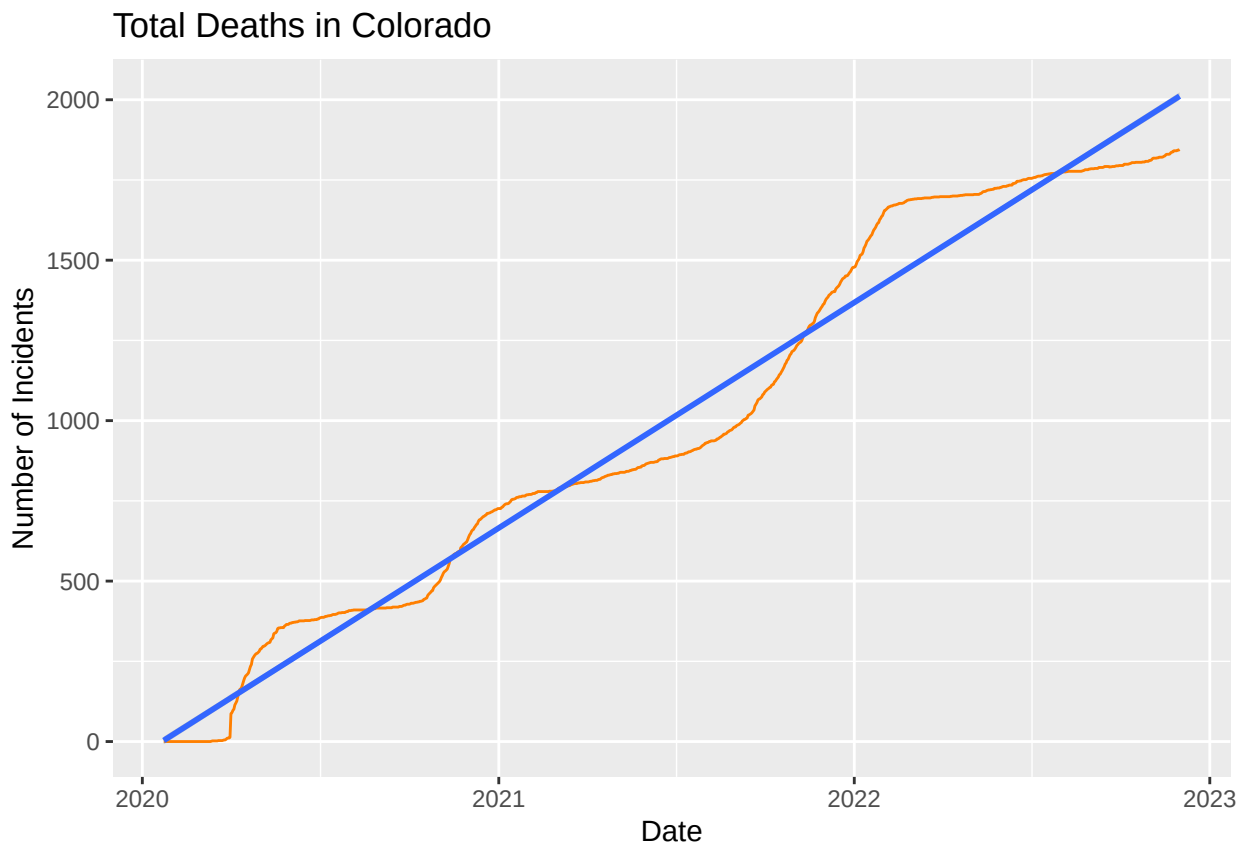

```
#New Cases
ggplot(colorado_aggregated, aes(x=Date, y=new_cases)) +
            geom_line(color = "cornflowerblue") +
            labs(x = "Date", y = "Number of Cases", title='New Cases in Colorado')
```



Fitting a linear model to the number of deaths

```
linear_deaths = lm(Deaths~Date, data=colorado_aggregated)
ggplot(colorado_aggregated, aes(x=Date,
                                y=Deaths, group = 1)) +
                                geom_line(color = "darkorange1") +
                                geom_smooth(method='lm', se = TRUE) +
                    labs(x = "Date", y = "Number of Incidents", title='Total Deaths in Colorado')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
summary(linear_deaths)
```

```
##
## Call:
## lm(formula = Deaths ~ Date, data = colorado_aggregated)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166.17  -76.90   -3.44   66.26  230.46
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.519e+04  1.908e+02  -184.4   <2e-16 ***
## Date         1.925e+00  1.015e-02   189.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.8 on 1042 degrees of freedom
## Multiple R-squared:  0.9719, Adjusted R-squared:  0.9718
## F-statistic: 3.599e+04 on 1 and 1042 DF,  p-value: < 2.2e-16
```

The linear model suggests the number of deaths are increasing linearly. The R value however suggests a poor fit, which is expected given the increase in the number of deaths is not linear.

Looking at only new cases

Conclusion, Sources of bias

In conclusion, we were able to identify the top five counties and visualize the overall count of covid cases in the state.

At the beginning of the pandemic the data, testing was limited to people with severe symptoms. This affects the number of new cases being counted everyday during the early stages of the pandemic. The overall reporting methodology by the counties also added bias as some reported data aggregated over the last week rather than a daily update.