DTSA Final

October 14, 2023

1 I. Data Cleaning and EDA

[4]:
```
library(lubridate)
library(tidyverse)
```

```
  Attaching core tidyverse packages                tidyverse
2.0.0
  dplyr   1.1.2       readr   2.1.4
  forcats 1.0.0       stringr 1.5.0
  ggplot2 3.4.2       tibble  3.2.1
  purrr   1.0.1       tidyr   1.3.0
  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
  Use the conflicted package
(<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

[6]:
```
# Load data, change from wide to long format and get basic summary
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/
 ↪csse_covid_19_data/csse_covid_19_time_series/"
file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")
urls <- str_c(url_in, file_names)
US_cases <- read_csv(urls[1])
US_deaths <- read_csv(urls[2])
US_cases <- US_cases %>% pivot_longer(cols = -(UID:Combined_Key),
                                      names_to = "date", values_to = "cases")␣
 ↪%>%
                                      select(Admin2:cases) %>%
                                      mutate(date = mdy(date)) %>%
                                      select(-c(Lat, Long_))
US_deaths <- US_deaths %>% pivot_longer(cols = -(UID:Population),
                                        names_to = "date", values_to =␣
 ↪"deaths") %>%
                                        select(Admin2:deaths) %>% mutate(date =␣
 ↪mdy(date)) %>%
```

```
                                        select(-c(Lat, Long_))
US <- US_cases %>% full_join(US_deaths)
summary(US)
```

```
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…

  Use `spec()` to retrieve the full column specification for this
data.
  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…

  Use `spec()` to retrieve the full column specification for this
data.
  Specify the column types or set `show_col_types = FALSE` to quiet
this message.
Joining with `by = join_by(Admin2, Province_State, Country_Region,
Combined_Key, date)`
```

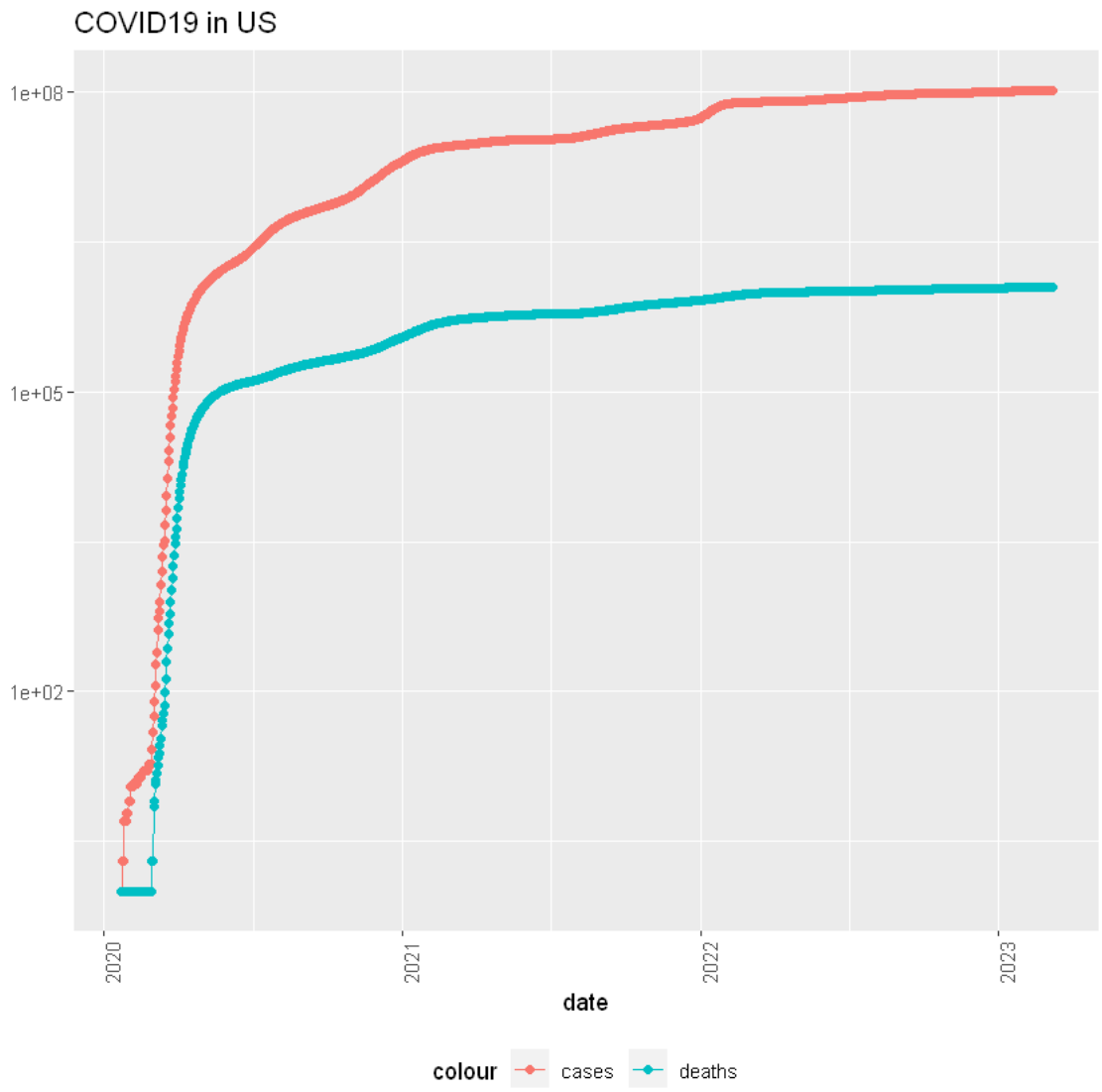
```
    Admin2          Province_State     Country_Region     Combined_Key
 Length:3819906     Length:3819906     Length:3819906     Length:3819906
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



      date                cases             Population          deaths
 Min.   :2020-01-22   Min.   : -3073   Min.   :       0   Min.   :  -82.0
 1st Qu.:2020-11-02   1st Qu.:   330   1st Qu.:    9917   1st Qu.:    4.0
 Median :2021-08-15   Median :  2272   Median :   24892   Median :   37.0
 Mean   :2021-08-15   Mean   : 14088   Mean   :   99604   Mean   :  186.9
 3rd Qu.:2022-05-28   3rd Qu.:  8159   3rd Qu.:   64979   3rd Qu.:  122.0
 Max.   :2023-03-09   Max.   :3710586   Max.   :10039107   Max.   :35545.0
```

This would be useful for anyone looking to analyze the progression of COVID-19 in the US over

time. By fetching the data directly from a trusted source (Johns Hopkins University) and processing it into a usable format, researchers and analysts can gain insights into the trends and patterns of the pandemic in the US.

[7]:
```
# The data is segmented by county so we can modify the data to get totals by␣
 ↪state and the whole USA
US_by_state <- US %>%
               group_by(Province_State, Country_Region, date) %>%
               summarize(cases = sum(cases), deaths = sum(deaths), Population =␣
 ↪sum(Population)) %>%
               mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
               select(Province_State, Country_Region, date, cases, deaths,␣
 ↪deaths_per_mill, Population) %>% ungroup()

US_totals <- US_by_state %>%
             group_by(Country_Region, date) %>%
             summarize(cases = sum(cases), deaths = sum(deaths), Population =␣
 ↪sum(Population)) %>%
             mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
             select(Country_Region, date, cases, deaths, deaths_per_mill,␣
 ↪Population) %>%
             ungroup()

US_totals %>% filter(cases > 0) %>%
    ggplot(aes(x = date, y = cases)) +
    geom_line(aes(color = "cases")) +
    geom_point(aes(color = "cases")) +
    geom_line(aes(y = deaths, color = "deaths")) +
    geom_point(aes(y = deaths, color = "deaths")) +
    scale_y_log10() +
    theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
    labs(title = "COVID19 in US", y = NULL)
```

```
`summarise()` has grouped output by 'Province_State', 'Country_Region'.
You can
override using the `.groups` argument.
`summarise()` has grouped output by 'Country_Region'. You can override
using
the `.groups` argument.
```

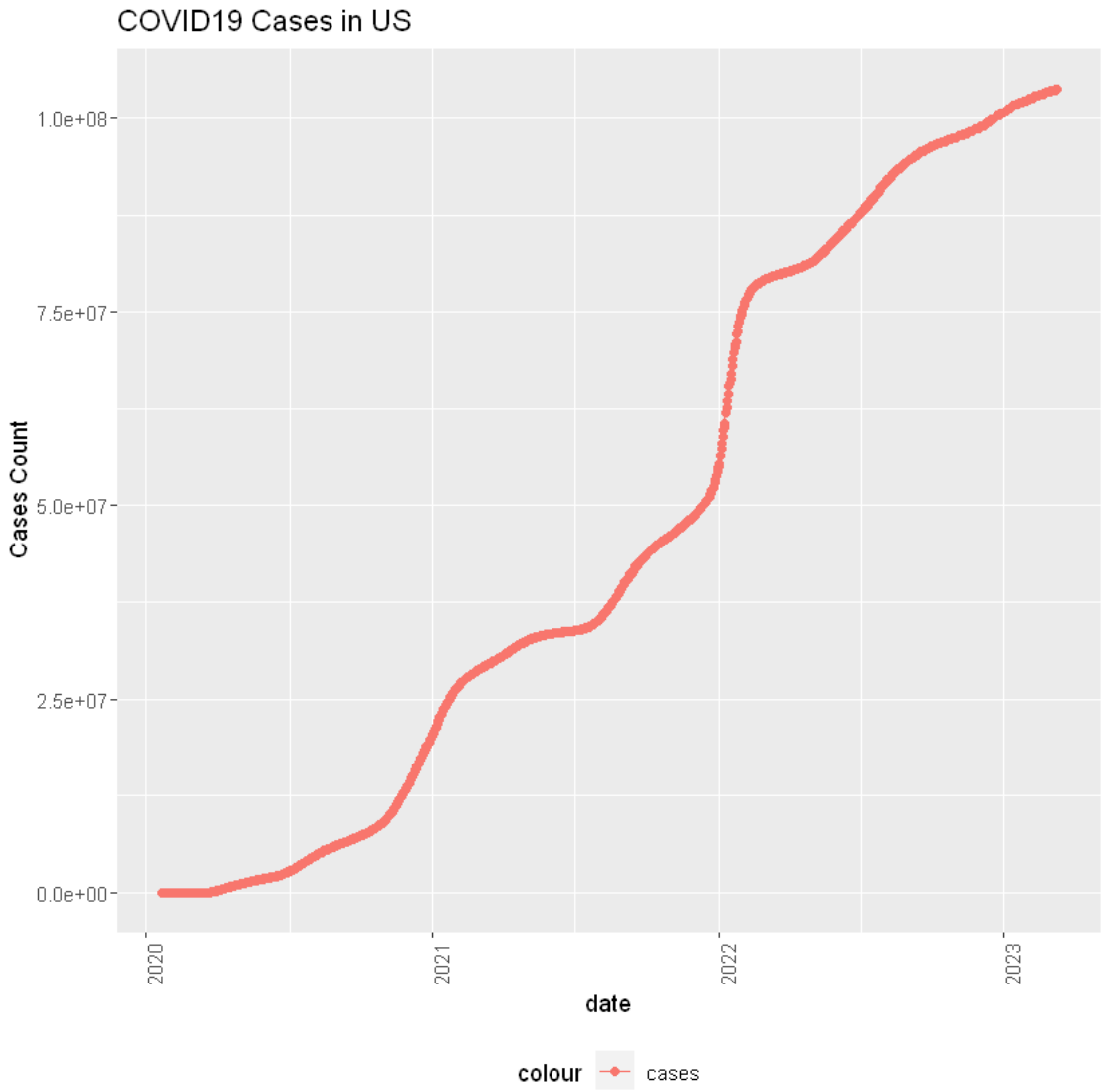


This provides an aggregated view of the COVID-19 situation in the USA, both at the state and national levels. By visualizing the data, stakeholders can make informed decisions, assess the efficacy of interventions, and track the progress of the pandemic.

[10]:
```
US_by_state <- US %>%
               group_by(Province_State, Country_Region, date) %>%
               summarize(cases = sum(cases), deaths = sum(deaths), Population =␣
 ↪sum(Population)) %>%
               mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
               select(Province_State, Country_Region, date, cases, deaths,␣
 ↪deaths_per_mill, Population) %>% ungroup()

US_totals <- US_by_state %>%
```

```
            group_by(Country_Region, date) %>%
            summarize(cases = sum(cases), deaths = sum(deaths), Population =␣
 ↪sum(Population)) %>%
            mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
            select(Country_Region, date, cases, deaths, deaths_per_mill,␣
 ↪Population) %>%
            ungroup()

US_totals %>% filter(cases > 0) %>%
    ggplot(aes(x = date, y = cases)) +
    geom_line(aes(color = "cases")) +
    geom_point(aes(color = "cases")) +
    theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
    labs(title = "COVID19 Cases in US", y = "Cases Count")
```

```
`summarise()` has grouped output by 'Province_State', 'Country_Region'.
You can
override using the `.groups` argument.
`summarise()` has grouped output by 'Country_Region'. You can override
using
the `.groups` argument.
```

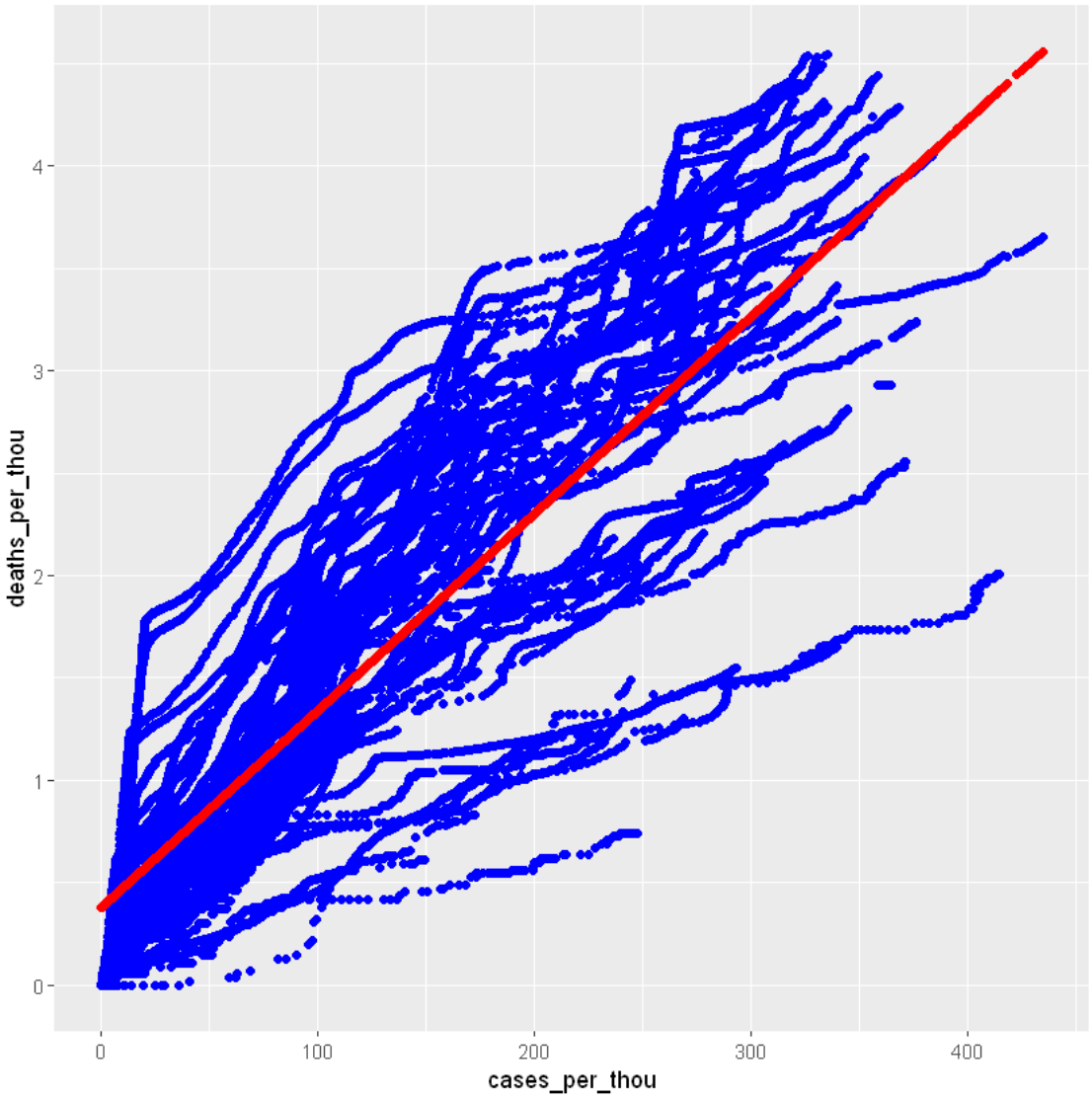


This provides a focused visualization of the cumulative confirmed cases of COVID-19 in the United States over time. By isolating and displaying only the cases metric, we gain a singular focus on simply cases. This facilitates a clearer interpretation of the data, enabling stakeholders, ranging from policymakers to the general public, to better gauge the situation in a narrower view.

[8]:
```
US_by_state <- US_by_state %>%
    mutate(
        new_cases = cases - lag(cases),
        new_deaths = deaths - lag(deaths),
        cases_per_thou = cases * 1000 / Population,
        deaths_per_thou = deaths * 1000 / Population
    ) %>%
    filter(cases > 0, Population > 0)
```

```r
mod <- lm(deaths_per_thou ~ cases_per_thou, data = US_by_state)
summary(mod)

US_w_pred <- US_by_state %>% mutate(pred = predict(mod))
US_w_pred %>% ggplot() +
    geom_point(aes(x = cases_per_thou, y = deaths_per_thou), color = "blue") +
    geom_point(aes(x = cases_per_thou, y = pred), color = "red")
```

```
Call:
lm(formula = deaths_per_thou ~ cases_per_thou, data = US_by_state)

Residuals:
     Min       1Q   Median       3Q      Max
-2.40763 -0.35462 -0.04173  0.45783  1.49623

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.793e-01  4.159e-03   91.19   <2e-16 ***
cases_per_thou 9.611e-03  2.257e-05  425.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6349 on 61037 degrees of freedom
Multiple R-squared:  0.7481,        Adjusted R-squared:  0.7481
F-statistic: 1.813e+05 on 1 and 61037 DF,  p-value: < 2.2e-16
```



Here we aim to understand the relationship between the number of confirmed cases and deaths due to COVID-19. By creating a regression model, we can quantify how the number of cases (or the case rate) is related to the death rate. This might provide insights into the severity of the disease, the quality of medical care, or other factors affecting mortality. The visualization helps in comparing the actual death rates with the predicted ones, which can be useful for validation or to spot any anomalies or trends.

2 II. Conclusion

In our analysis of COVID-19's progression in the US, we employed data aggregation and visualization to highlight trends at state and national levels. By examining the relationship between confirmed cases and deaths, we gained insights into disease severity over time.