Reproducible report on COVID-19 data analysis

2025-09-02

1. Project Overview

This project analyzes trends in COVID-19-related deaths in New York and Kentucky using publicly available data. Its primary goal is to provide hands-on practice with data management and analysis, focusing on data cleaning, transformation, and visualization. The project will interpret how these death counts have evolved over time.

2. Data Source

This analysis utilizes publicly available COVID-19 data collected and maintained by the Johns Hopkins University Center for Systems Science and Engineering (CSSE). The dataset is accessible via the GitHub repository, with specific links provided in the data importing section below. Only the portion of the data pertaining to the United States was used for this project.

3. Data management

3.1. Importing data

```
# Import tidyverse library
library(tidyverse)

# Concatenate partial URL and file names to make a list of complete URLs
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi

file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")
urls <- str_c(url_in, file_names)

# Read data into R
us_cases = read_csv(urls[1])
us_deaths = read_csv(urls[2])
us_cases; us_deaths
```

```
## # A tibble: 3,342 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
##  7 84001013 US    USA     840  1013 Butler   Alabama        US              31.8
##  8 84001015 US    USA     840  1015 Calhoun  Alabama        US              33.8
##  9 84001017 US    USA     840  1017 Chambers Alabama        US              32.9
## 10 84001019 US    USA     840  1019 Cherokee Alabama        US              34.2
```

```
## # i 3,332 more rows
## # 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>, ...
```

```
## # A tibble: 3,342 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
##  7 84001013 US    USA     840  1013 Butler   Alabama        US              31.8
##  8 84001015 US    USA     840  1015 Calhoun  Alabama        US              33.8
##  9 84001017 US    USA     840  1017 Chambers Alabama        US              32.9
## 10 84001019 US    USA     840  1019 Cherokee Alabama        US              34.2
## # i 3,332 more rows
## # 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>, ...
```

The R output shows daily data stored across many columns. For easier analysis, we'll reshape this data into two new columns: date, cases/deaths.

3.2. Tidying and transforming data

- Transformed the "wider" US case and death files into "longer" files.
- Created columns for **date** and **cases** for the case data, and **date** and **deaths** for the death data.
- Set the **date** column as date object.
- Removed the columns not needed for analysis.
- Combined the two files into a single **US** file.

```
# Tidying and transforming US cases file
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_))

# Tidying and transforming US deaths file
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_))

# Combining cases and deaths files
US <- us_cases %>%  full_join(us_deaths)
US
```

```
## # A tibble: 3,819,906 x 8
##    Admin2 Province_State Country_Region Combined_Key date       cases Population
##    <chr>  <chr>          <chr>          <chr>        <date>     <dbl>      <dbl>
##  1 Autau~ Alabama        US             Autauga, Al~ 2020-01-22     0      55869
##  2 Autau~ Alabama        US             Autauga, Al~ 2020-01-23     0      55869
##  3 Autau~ Alabama        US             Autauga, Al~ 2020-01-24     0      55869
##  4 Autau~ Alabama        US             Autauga, Al~ 2020-01-25     0      55869
##  5 Autau~ Alabama        US             Autauga, Al~ 2020-01-26     0      55869
##  6 Autau~ Alabama        US             Autauga, Al~ 2020-01-27     0      55869
##  7 Autau~ Alabama        US             Autauga, Al~ 2020-01-28     0      55869
##  8 Autau~ Alabama        US             Autauga, Al~ 2020-01-29     0      55869
##  9 Autau~ Alabama        US             Autauga, Al~ 2020-01-30     0      55869
## 10 Autau~ Alabama        US             Autauga, Al~ 2020-01-31     0      55869
## # i 3,819,896 more rows
## # i 1 more variable: deaths <dbl>
```

The R output shows that the "US" data contains 3,819,906 rows and 8 columns. The **date** variable/column is now in the **date** format. But there are some rows with 0 case, which may be not very useful.

3.3. Data aggregation by states

- Calculate the total nubmer of **cases**, **deaths** and **Population** by states.
- Remove rows where either case count or population size is 0.
- Create new columns for **cases_per_mill**, **deaths_per_mill**.
- Remove unnecessary columns

```
US_by_State <- US %>%
               group_by(Province_State, Country_Region, date) %>%
               summarize(cases = sum(cases),
                         deaths = sum(deaths),
                         Population = sum(Population)) %>%
               filter(cases > 0, Population > 0) %>%
               mutate(cases_per_mill = cases * 1000000 / Population,
                      deaths_per_mill = deaths * 1000000 / Population) %>%
               select(Province_State, Country_Region, date,
                      cases, deaths, cases_per_mill,
                      deaths_per_mill, Population) %>%
               ungroup()
```

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

```
US_by_State
```

```
## # A tibble: 61,039 x 8
##    Province_State Country_Region date       cases deaths cases_per_mill
##    <chr>          <chr>          <date>     <dbl>  <dbl>          <dbl>
##  1 Alabama        US             2020-03-11     3      0          0.612
##  2 Alabama        US             2020-03-12     4      0          0.816
##  3 Alabama        US             2020-03-13     8      0          1.63
```

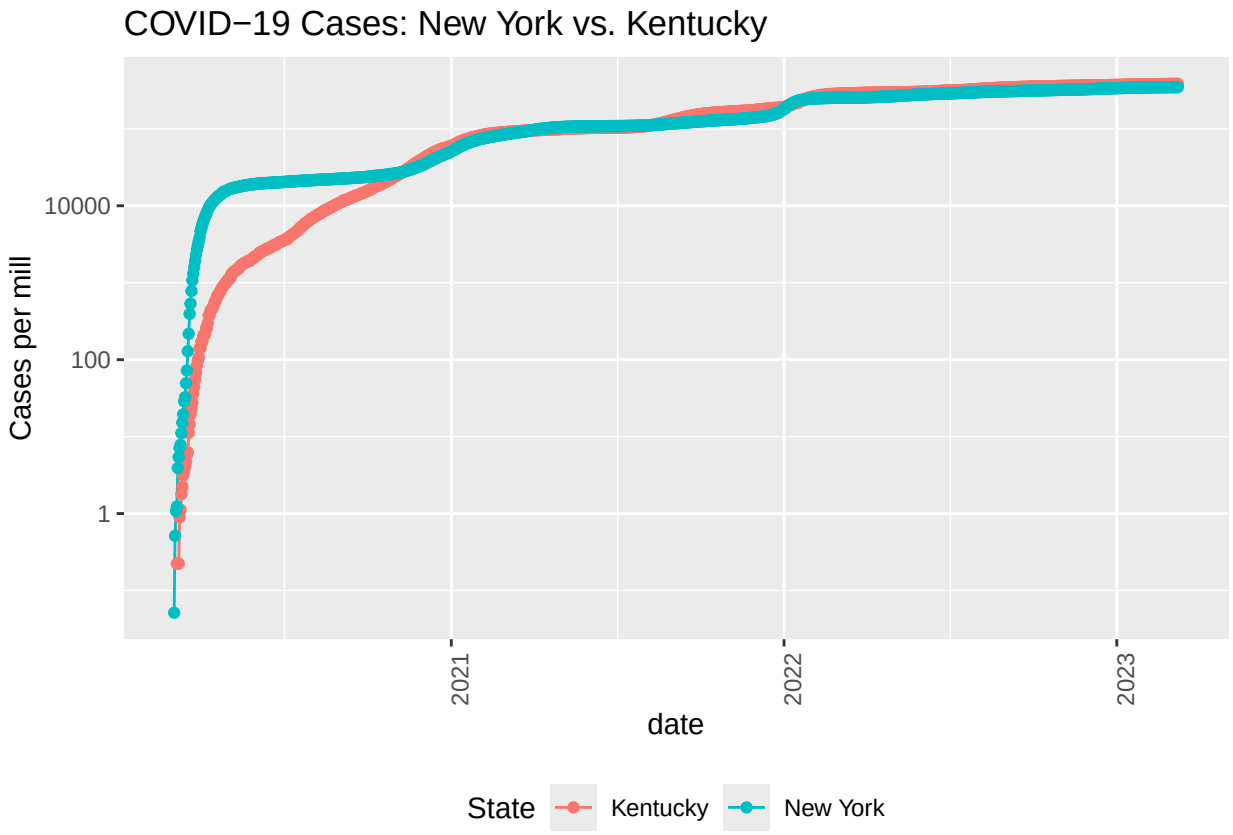
```
##  4 Alabama        US             2020-03-14    15      0          3.06
##  5 Alabama        US             2020-03-15    28      0          5.71
##  6 Alabama        US             2020-03-16    36      0          7.34
##  7 Alabama        US             2020-03-17    51      0         10.4
##  8 Alabama        US             2020-03-18    61      0         12.4
##  9 Alabama        US             2020-03-19    88      0         17.9
## 10 Alabama        US             2020-03-20   115      0         23.5
## # i 61,029 more rows
## # i 2 more variables: deaths_per_mill <dbl>, Population <dbl>
```

After aggregation, there are 61,039 rows and 8 columns. There are no more rows with 0 case.

4. Analysis and visualization

4.1. Compare the COVID-19 case trends in two states

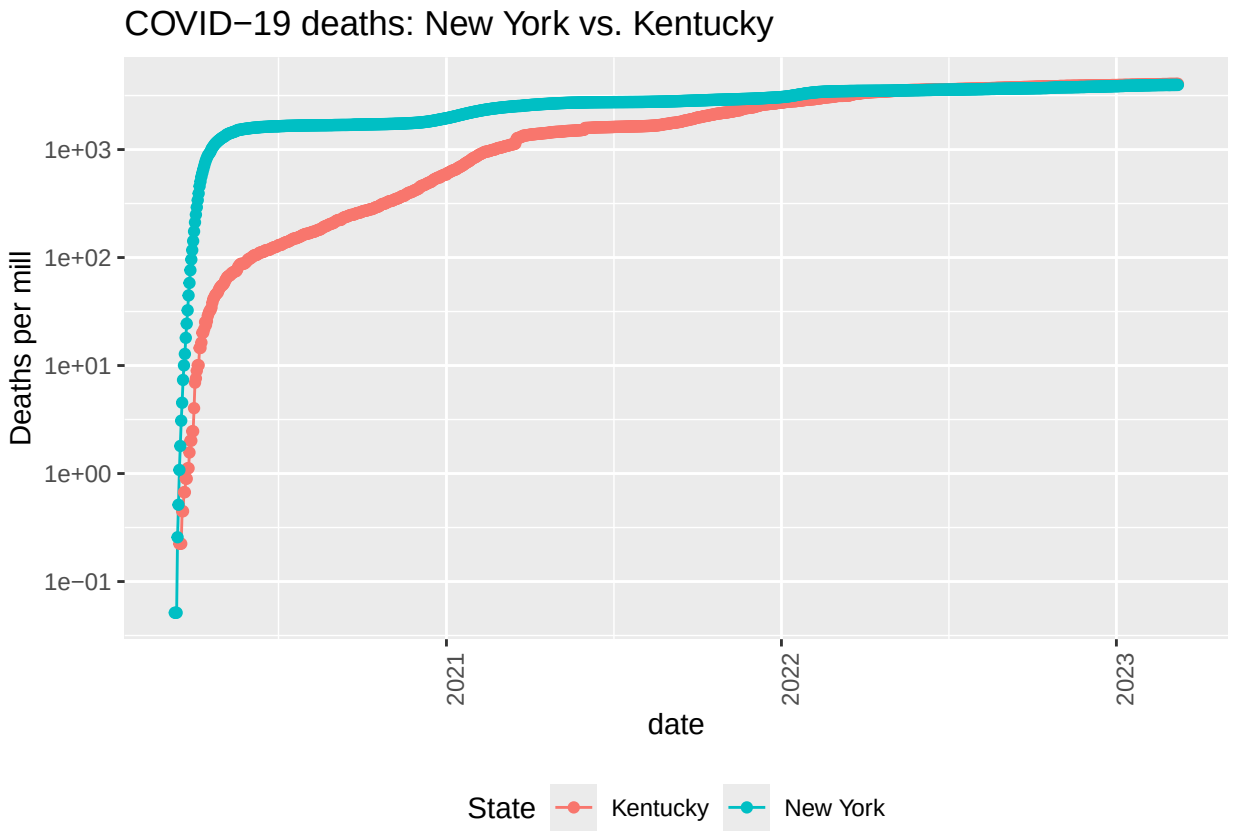
```
# Choose which two states to compare
states_to_compare <- c("New York", "Kentucky")

# Compare two states on the nubmer of cases weighted by population sizes
US_by_State %>%
    filter(Province_State %in% states_to_compare) %>%
    ggplot(aes(x = date, y = cases_per_mill, color = Province_State)) +
    geom_line() +
    geom_point() +
    scale_y_log10() +
    theme(legend.position = "bottom",
          axis.text.x = element_text(angle = 90)) +
    labs(title = "COVID-19 Cases: New York vs. Kentucky",
          y = "Cases per mill", color = "State")
```



To compare the two states with different population sizes, we calculated cases per million. The graph shows that New York experienced a rapid and steep initial increase in cases, with its curve quickly rising above Kentucky's. While Kentucky's curve also increased, it was at a slower rate. Over time, both states' cumulative case curves have flattened, indicating a slowdown in the rate of new infections. Since late 2020, the two states have maintained a relatively constant difference in their cumulative case counts per million.

4.2 Compare the COVID-19-related death trends in two states

```
US_by_State %>%
    filter(Province_State %in% states_to_compare) %>%
    filter(deaths > 0) %>%
    ggplot(aes(x = date, y = deaths_per_mill, color = Province_State)) +
    geom_line() +
    geom_point() +
    scale_y_log10() +
    theme(legend.position = "bottom",
          axis.text.x = element_text(angle = 90)) +
    labs(title = "COVID-19 deaths: New York vs. Kentucky",
          y = "Deaths per mill", color = "State")
```



Just like with the case counts, the trends for cumulative COVID-19 deaths per million in New York and Kentucky show a similar pattern. New York experienced a rapid and steep rise in deaths early in the pandemic, with its curve quickly rising far above Kentucky's. While Kentucky's death count also increased, it was at a much slower rate. Over time, both states saw their curves flatten out, indicating a slowdown in the rate of new deaths. However, a significant gap remains, with New York consistently having a substantially higher cumulative death count per million than Kentucky throughout the period shown, particularly before 2022.

4.3. Modeling the trend of COVID-19-related deaths using a simple linear regression model

```
# Filter the data for each state and fit a separate linear model
mod_ny <- US_by_State %>%
  filter(Province_State == "New York") %>%
  lm(deaths_per_mill ~ cases_per_mill, data = .)
summary(mod_ny)
```

```
##
## Call:
## lm(formula = deaths_per_mill ~ cases_per_mill, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1539.93  -155.75   -15.86   312.29   404.92
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.540e+03  1.814e+01   84.91   <2e-16 ***
## cases_per_mill 7.403e-03  9.164e-05   80.78   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 361.6 on 1100 degrees of freedom
## Multiple R-squared:  0.8557, Adjusted R-squared:  0.8556
## F-statistic:  6526 on 1 and 1100 DF,  p-value: < 2.2e-16
```

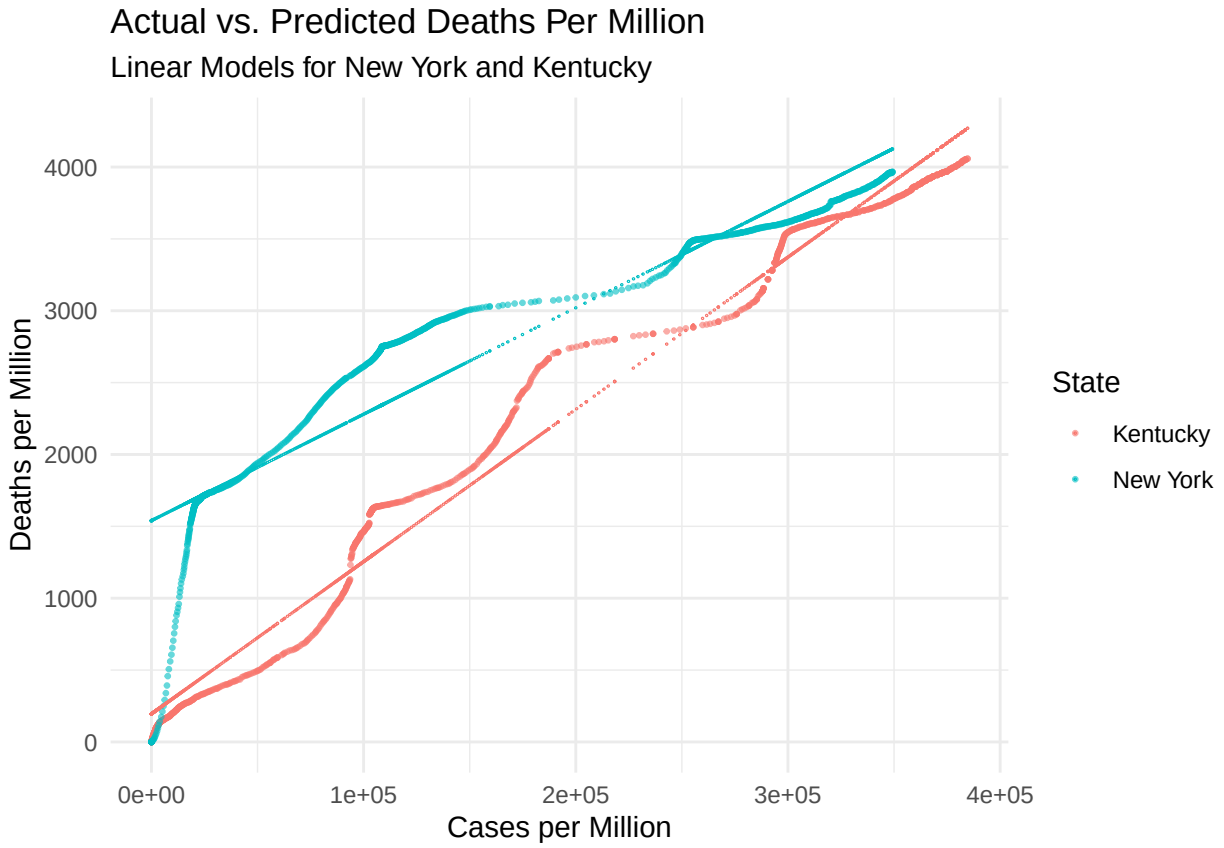
```
mod_ky <- US_by_State %>%
  filter(Province_State == "Kentucky") %>%
  lm(deaths_per_mill ~ cases_per_mill, data = .)
summary(mod_ky)
```

```
##
## Call:
## lm(formula = deaths_per_mill ~ cases_per_mill, data = .)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -271.9 -148.8 -100.6  161.1  500.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.953e+02  9.512e+00   20.53   <2e-16 ***
## cases_per_mill 1.059e-02  4.343e-05  243.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195.1 on 1097 degrees of freedom
## Multiple R-squared:  0.9819, Adjusted R-squared:  0.9819
## F-statistic: 5.949e+04 on 1 and 1097 DF,  p-value: < 2.2e-16
```

The linear regression analysis shows a statistically significant relationship between COVID-19-related deaths and case counts in both New York and Kentucky. However, the adjusted R-squared values reveal a key difference: while 85.6% of the variation in death count is explained by case count in New York, that figure rises to 98.2% in Kentucky. This indicates that the simple linear regression model is a much better fit for the data from Kentucky.

```
# Generate predictions for each state's data using its respective model
US_by_State_w_pred <- US_by_State %>%
  filter(Province_State %in% c("New York", "Kentucky")) %>%
  filter(cases > 0, Population > 0) %>%
  mutate(
    # Create a new column 'pred' with predictions based on the state
    pred = case_when(
      Province_State == "New York" ~ predict(mod_ny, newdata = .),
      Province_State == "Kentucky" ~ predict(mod_ky, newdata = .)
    )
  )


US_by_State_w_pred %>%
  ggplot(aes(x = cases_per_mill, color = Province_State)) +
  geom_point(aes(y = deaths_per_mill), size = 0.5, alpha = 0.6) +
  geom_point(aes(y = pred), shape = 1, size = 0.2, stroke = 0.3) +
  labs(title = "Actual vs. Predicted Deaths Per Million",
       subtitle = "Linear Models for New York and Kentucky",
       x = "Cases per Million",
```

```
      y = "Deaths per Million",
      color = "State") +
theme_minimal()
```



In this figure, the curved lines represent the actual trends of COVID-19-related deaths per million as a function of cases per million, while the straight lines show the fit from the simple linear regression model.

It's clear from the graph that the model provides a much better fit for Kentucky's data. The red straight line closely aligns with the actual data points. Conversely, the model is a poor fit for New York's data, as the blue curved line has a pronounced concave pattern. This indicates that the relationship between cases and deaths in New York is non-linear, and a quadratic regression model may be a more appropriate fit.

5. Conclusion

- Initial Trends: Both New York and Kentucky showed similar overall trends in cumulative COVID-19 cases and deaths per million, with a rapid initial increase followed by a flattening of the curves. New York consistently had a higher cumulative case and death count per million.
- Linear Regression Model Fit: A statistically significant relationship between case and death counts exists in both states. However, the simple linear regression model is a much better fit for Kentucky's data (98.2% explained variation) than for New York's (85.6% explained variation).
- Model Limitations: The poor fit of the simple linear regression model for New York's data, as evidenced by a concave pattern, suggests that the relationship between cases and deaths in that state is non-linear and may be better captured by a quadratic regression model.

6. Possible biases

- New York (Urban Density Bias): The early and rapid spread of the virus in New York City, a global epicenter with high population density and extensive public transit, likely led to a significant number of early cases and deaths going uncounted due to limited testing capacity.
- Kentucky (Rural and Socioeconomic Bias): Kentucky, a more rural state, may have had different data biases. Due to less population density, the spread might have been slower, but access to testing and healthcare could have been a greater challenge in remote areas. Likely some cases and deaths were not counted.