Covid-19 Analysis

KK

2024-06-20

```r
# Clear environment
rm(list = ls())
cat("\014")  # ctrl+L
```

Introduction

In this project, I analyze the COVID-19 data from 2020 to 2023. The COVID-19 dataset was obtained from the Johns Hopkins GitHub repository.

Descriptive Analysis: I start by analyzing the dataset for the USA where I answer the following questions:

- How did COVID-19 cases and deaths evolve over the years in the USA?
- How have COVID-19 cases and deaths varied across different states?

I then focus on analyzing COVID-19 in Kentucky, where I answer the following questions:

- How do COVID-19 new cases and deaths vary over time in Kentucky?
- How do the total numbers of new COVID-19 cases and deaths vary across counties in Kentucky?
- What is the mortality rate in Kentucky?

Predictive Analysis:

- Fit a simple linear regression model
- Interpret the results
- Conclusion and Bias

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(dplyr)
library(plotly)
```

COVID-19 Data Retrieval and Preparation

```
# All files begin with this string.
url_in <- ('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_co
# Vector containing four file names.
file_names <-
  c("time_series_covid19_confirmed_global.csv",
  "time_series_covid19_deaths_global.csv",
  "time_series_covid19_confirmed_US.csv",
  "time_series_covid19_deaths_US.csv")

urls <- str_c(url_in, file_names)
```

```
global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
US_cases <- read_csv(urls[3])
US_deaths <- read_csv(urls[4])
```

Data Transformation: Global COVID-19 Cases and Deaths

```
global_cases <- global_cases %>%
  pivot_longer(cols =
               -c('Province/State',
                    'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "cases") %>%
  select(-c(Lat, Long))
#head(global_cases,3)
```

```
global_deaths <- global_deaths %>%
  pivot_longer(cols =
                 -c('Province/State',
                    'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "deaths") %>%
  select(-c(Lat, Long))
#head(global_deaths,3)
```

```
 global <- global_cases %>%
  full_join(global_deaths) %>%
    rename(Country_Region = 'Country/Region',
       Province_State = 'Province/State') %>%
  mutate(date = mdy(date))
```

```
## Joining with ‘by = join_by(‘Province/State‘, ‘Country/Region‘, date)‘
```

Data Preparation: US COVID-19 Cases and Deaths

```
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_))
#head(US_cases,1)
```

```
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_))
#head(US_deaths,1)
```

```
US <- US_cases %>%
  full_join(US_deaths)
```

```
## Joining with ‘by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)‘
```

```
#head(US,1)
```

I notice here that cases and deaths has negative numbers which incorrect. So, I need to filter the dataframe

```
# Filter the US dataframe
US_by_state <- US %>%
  filter(cases > 0, deaths >= 0, Population > 0)

# Display the first few rows of the filtered dataframe to verify
#head(US_by_state,3)
```

```
US_by_state <- US_by_state %>% mutate(new_cases = cases - lag(cases),
                                      new_deaths = deaths - lag(deaths))
#head(US_by_state,3)
```

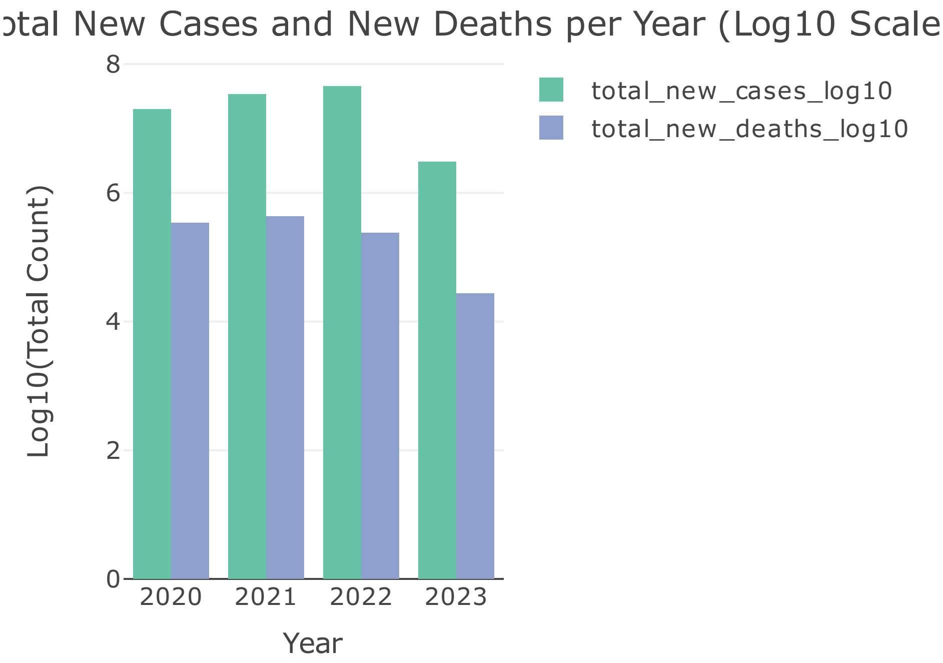
```
# Filter the US dataframe
US_by_state <- US_by_state  %>%
  filter(new_cases > 0, new_deaths >= 0, Population > 0)

# Display the first few rows of the filtered dataframe to verify
#head(US_by_state,3)
```

Descriptive Analysis for US COVID-19 Cases and Deaths: Question : How do COVID-19 cases and deaths evolved over the years in USA??

```
# Adding a 'year' column
data <- US_by_state %>%
  mutate(year = format(date, "%Y"))
# Group by 'year' and summarize new cases and new deaths
summary_by_year <- data %>%
  group_by(year) %>%
  summarise(
    total_new_cases = sum(new_cases, na.rm = TRUE),
    total_new_deaths = sum(new_deaths, na.rm = TRUE)
  ) %>%
  # Apply log10 transformation
  mutate(
    total_new_cases_log10 = log10(total_new_cases + 1),
    total_new_deaths_log10 = log10(total_new_deaths + 1)
  )

# Reshape the data for plotting
summary_melted <- summary_by_year %>%
  pivot_longer(cols = c(total_new_cases_log10, total_new_deaths_log10), names_to = "metric", values_to =

# Create the plot using plot_ly
bar_plot <- plot_ly(summary_melted, x = ~year, y = ~value, color = ~metric, type = 'bar') %>%
  layout(
    title = 'Total New Cases and New Deaths per Year (Log10 Scale)',
    xaxis = list(title = 'Year'),
    yaxis = list(title = 'Log10(Total Count)'),
    barmode = 'group'
  )


# Display the plot
bar_plot
```

```
## Google Chrome was not found. Try setting the ‘CHROMOTE_CHROME‘ environment variable to the executabl
```

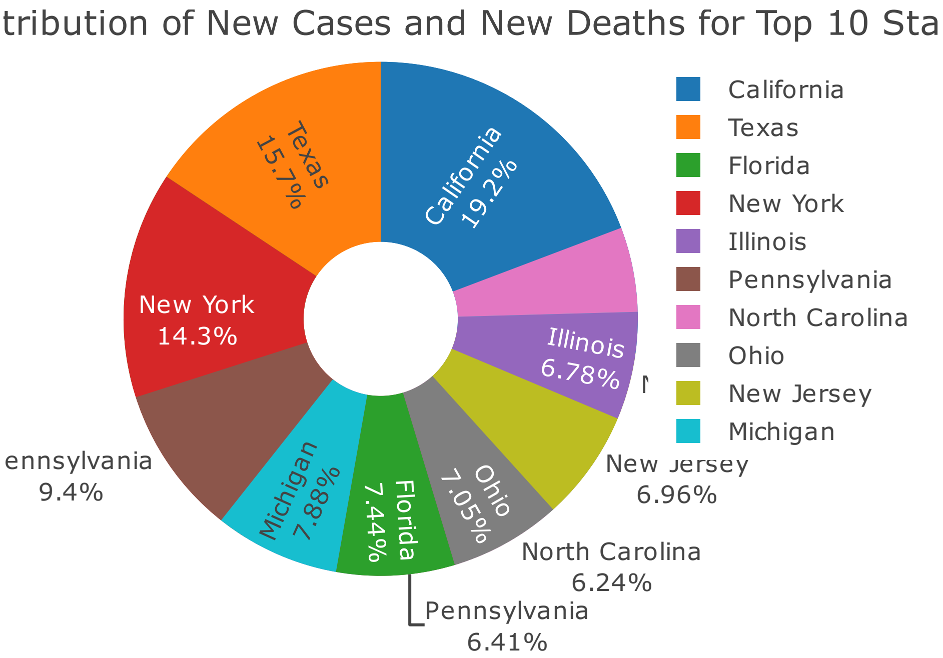
```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palett
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palett
```



The data shows that COVID-19 cases and deaths have been going up and down over the years. In 2020, there was a big increase in cases and deaths globally. This trend continued in 2021 and 2022, with more cases and deaths. But in 2023, there was a decrease in both. This might mean that the pandemic is changing. Using the logarithmic transformation helped us understand the data better, showing us how big the impact of the pandemic has been worldwide.

Question2 : How have COVID-19 cases and deaths varied across different states?

```
# Summarize data by Province_State for new cases and new deaths
summary_data <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(
    total_new_cases = sum(new_cases, na.rm = TRUE),
    total_new_deaths = sum(new_deaths, na.rm = TRUE)
  )

# Select the top 10 states by total new cases
top_10_states <- summary_data %>%
  arrange(desc(total_new_cases)) %>%
  slice(1:10)

# Reshape data to long format for plotting
top_10_long <- top_10_states %>%
  pivot_longer(cols = c(total_new_cases, total_new_deaths), names_to = "type", values_to = "count")

# Create a pie chart using Plotly for the top 10 states with new cases and new deaths
pie_plot <- plot_ly(top_10_long, labels = ~Province_State, values = ~count, type = 'pie', textinfo = 'la
                    name = ~type, hole = 0.3) %>%
  layout(
    title = 'Distribution of New Cases and New Deaths for Top 10 States',
    showlegend = TRUE
  )

# Display the pie plot
pie_plot
```



This data highlights the extensive and diverse impact of COVID-19 across various states. Populous states such as California, New York, Texas, and Florida exhibit notably high numbers of cases and deaths. Conversely, states like American Samoa report the lowest figures for both cases and deaths. These disparities likely reflect variations in population size, density, as well as differences in healthcare infrastructure and public health responses.

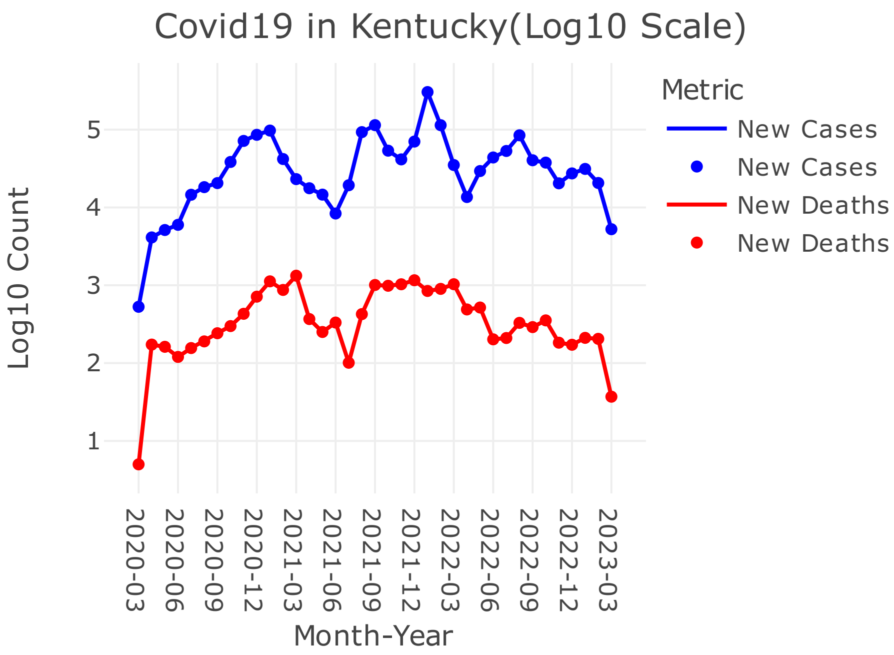
```
# Filter US dataset for only the rows where Province_State is Kentucky.
df_ky <- US_by_state%>%
  filter(Province_State == "Kentucky") %>%
  group_by(date, Admin2)
head(df_ky,3)
```

Descriptive Analysis for Kentucky State COVID-19:

```
## # A tibble: 3 x 10
## # Groups:   date, Admin2 [3]
##   Admin2 Province_State Country_Region Combined_Key  date       cases Population
##   <chr>  <chr>          <chr>          <chr>         <date>     <dbl>      <dbl>
## 1 Adair  Kentucky       US             Adair, Kentu~ 2020-04-09     6      19202
## 2 Adair  Kentucky       US             Adair, Kentu~ 2020-04-12    11      19202
## 3 Adair  Kentucky       US             Adair, Kentu~ 2020-04-14    43      19202
## # i 3 more variables: deaths <dbl>, new_cases <dbl>, new_deaths <dbl>
```

Question : How does COVID-19 new__cases and new deaths vary over time in Kentucky?

```
# Group by month-year and summarize new_cases and new_deaths
data_summary <- df_ky %>%
  mutate(month_year = format(date, "%Y-%m")) %>%
  group_by(month_year) %>%
  summarize(
    total_new_cases = sum(new_cases),
    total_new_deaths = sum(new_deaths)
  )

# Create the plot using plotly with log10 scale for y-axis
plot <- plot_ly(data_summary, x = ~month_year) %>%
  add_lines(y = ~log10(total_new_cases + 1), name = 'New Cases', line = list(color = 'blue')) %>%
  add_markers(y = ~log10(total_new_cases + 1), name = 'New Cases', marker = list(color = 'blue')) %>%
  add_lines(y = ~log10(total_new_deaths + 1), name = 'New Deaths', line = list(color = 'red')) %>%
  add_markers(y = ~log10(total_new_deaths + 1), name = 'New Deaths', marker = list(color = 'red')) %>%
  layout(
    title = 'Covid19 in Kentucky(Log10 Scale)',
    xaxis = list(title = 'Month-Year'),
    yaxis = list(title = 'Log10 Count'),
    legend = list(title = list(text = 'Metric'))
  )

# Display the plot
plot
```

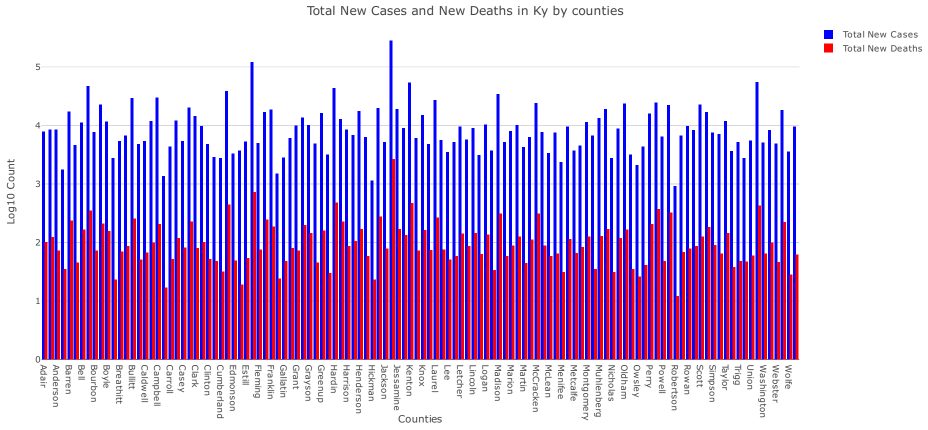


The data highlights notable trends in new COVID-19 cases and deaths over the analyzed period. Starting with 528 cases in March 2020, new cases peaked at 303,782 in January 2022, showing fluctuations that reflect changing infection rates. Similarly, new deaths ranged from 4 in March 2020 to a peak of 1,327 in March 2021, also varying over time. Both new cases and deaths generally increased from mid-2020 to early 2021, reflecting the pandemic's initial impact. However, there was a noticeable decline in both metrics from mid-2021 onward, although with occasional increases possibly influenced by changing infection rates and public health measures. In 2023, there was a significant reduction in both new cases and deaths compared to previous years, indicating potential progress in controlling COVID-19. Overall, these trends underscore the pandemic's dynamic nature, shaped by factors like public health measures, vaccination efforts, new virus variants, and community behaviors affecting transmission rates.

Question: How do the total numbers of new COVID-19 cases and deaths vary across counties in Kentucky?

```
# Assuming your data frame is named df
summary_data <- df_ky %>%
  mutate(Combined_Key = gsub(",.*", "", Combined_Key)) %>%  # Extracting only the first name from Combi
  group_by(Combined_Key) %>%
  summarize(
    total_new_cases = sum(new_cases, na.rm = TRUE),
    total_new_deaths = sum(new_deaths, na.rm = TRUE)
  )
```

```
# Plot using plotly with log10 scale for y-axis
plot <- plot_ly(summary_data, x = ~Combined_Key, width = 1300, height = 600) %>%
  add_bars(y = ~log10(total_new_cases + 1), name = 'Total New Cases', marker = list(color = 'blue')) %>
  add_bars(y = ~log10(total_new_deaths + 1), name = 'Total New Deaths', marker = list(color = 'red')) %
  layout(
    title = 'Total New Cases and New Deaths in Ky by counties',
    xaxis = list(title = 'Counties'),
    yaxis = list(title = 'Log10 Count'),
    barmode = 'group'
  )

# Display the plot
plot
```

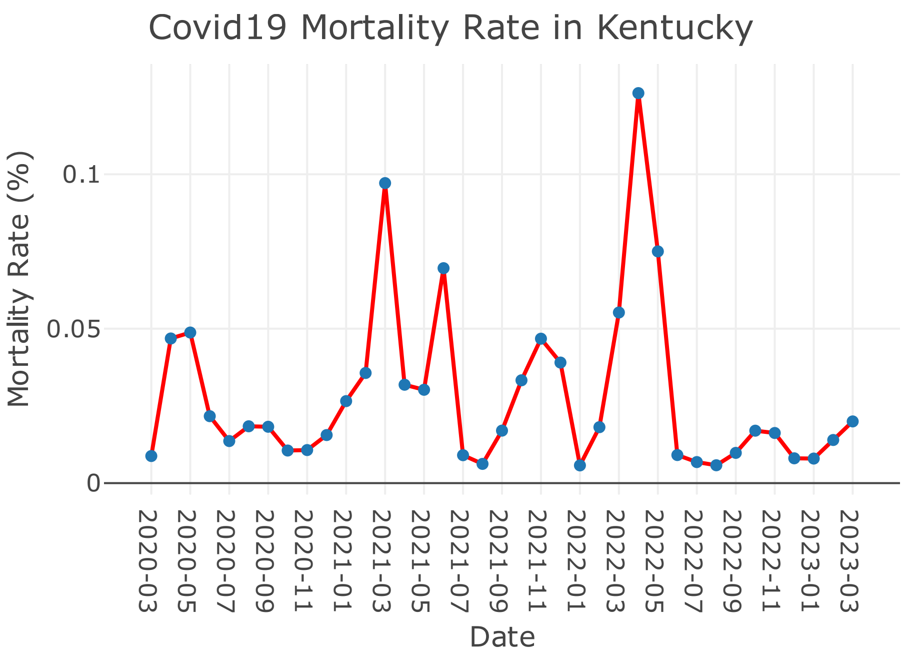


Jefferson County has the highest counts with 284,917 new cases and 2,650 new deaths, followed by Fayette County with 122,533 new cases and 722 new deaths. Other counties with high counts include Kenton, Boone, and Hardin. The lowest counts are found in Carlisle, Hickman, and Fulton counties. Most counties fall in the mid-range.Note that these results are from 2020 to 2023.

Question: HOw bout the mortality rate in Kentucky?

```
# Calculate mortality rate
df_ky <- df_ky %>%
  mutate(mortality_rate = new_deaths / new_cases)

# Handle infinite values or NaNs if any
df_ky <- df_ky %>%
  mutate(mortality_rate = ifelse(is.infinite(mortality_rate) | is.nan(mortality_rate), 0, mortality_rat

# Extract month/year and group by it
df_ky <- df_ky %>%
  mutate(month_year = format(date, "%Y-%m")) %>%
  group_by(month_year) %>%
  summarize(
    total_new_cases = sum(new_cases),
    total_new_deaths = sum(new_deaths),
    avg_mortality_rate = mean(mortality_rate, na.rm = TRUE)
  )
```

```
# Create the plot using plotly
plot <- plot_ly(df_ky, x = ~month_year, y = ~avg_mortality_rate, type = 'scatter', mode = 'lines+markers
  layout(
    title = 'Covid19 Mortality Rate in Kentucky ',
    xaxis = list(title = 'Date'),
    yaxis = list(title = 'Mortality Rate (%)'),
    legend = list(title = list(text = 'Metric'))
  )

# Display the plot
plot
```



Analysis reveals fluctuating case and death counts over time, with notable peaks in certain months like January 2021 and December 2021. Mortality rates vary widely, from below 1% to peaks around 12.6% in April 2022, indicating changing in COVID-19 severity.

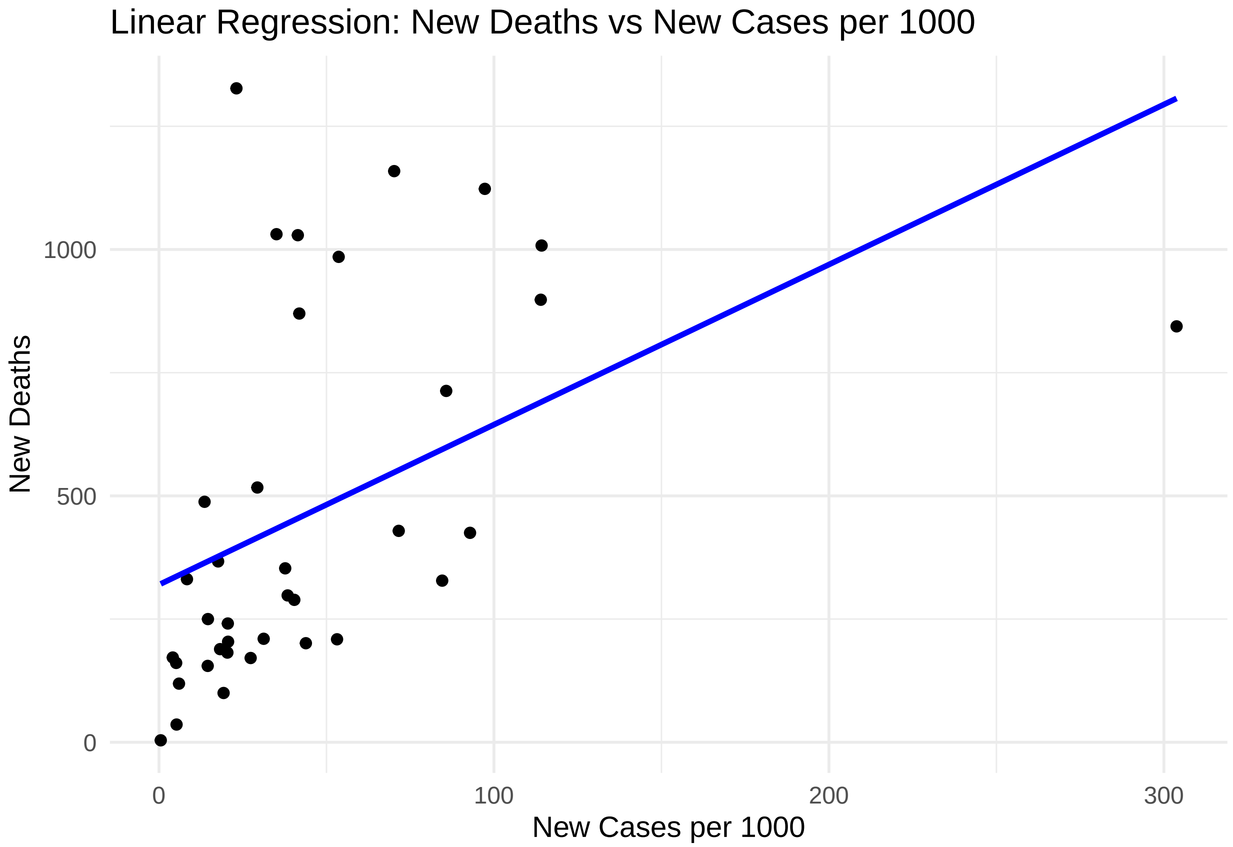
```

# Assuming your data frame is named df_ky

kentucky_model <- lm(total_new_deaths ~ total_new_cases  , data = df_ky )
summary(kentucky_model)
```

Data Modeling: Simple Regression Model:

```
##
## Call:
## lm(formula = total_new_deaths ~ total_new_cases, data = df_ky)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -462.5 -212.0 -146.5  124.1  932.2
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.197e+02  7.366e+01    4.34 0.000115 ***
## total_new_cases 3.248e-03  1.044e-03    3.11 0.003707 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 337 on 35 degrees of freedom
## Multiple R-squared:  0.2165, Adjusted R-squared:  0.1941
## F-statistic: 9.673 on 1 and 35 DF,  p-value: 0.003707
```

```
df_ky <- df_ky %>%
  mutate(new_cases_per_1000 = total_new_cases / 1000)

# Fit the linear regression model
kentucky_model <- lm(total_new_deaths ~ new_cases_per_1000, data = df_ky)

# Plot the data points and regression line using ggplot2
ggplot(df_ky, aes(x = new_cases_per_1000, y = total_new_deaths)) +
  geom_point() +  # Add scatter plot of data points
  geom_smooth(method = "lm", se = FALSE, color = "blue") +  # Add regression line
  labs(title = "Linear Regression: New Deaths vs New Cases per 1000",
       x = "New Cases per 1000",
       y = "New Deaths") +
  theme_minimal()
```

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



```
# Limit x-axis to [0, 2500]
```

Interpretation: In this linear regression model, the intercept (0.1702) signifies the estimated average number of new deaths when there are zero new cases reported. The coefficient for new_cases (0.003568) indicates that for each additional unit increase in new_cases, there is an estimated increase of 0.003568 units in new deaths. Both coefficients are highly significant (***), confirming that they reliably predict new deaths based on new cases. The R-squared value of 0.1007 reveals that new_cases explain approximately 10.07% of the variability observed in new deaths. The F-statistic (7423) with an exceedingly small p-value (< 2.2e-16) underscores the overall model's statistical significance, validating the relationship between new_cases and new_deaths.

Conclusion and Bias:

Despite the significance, we notice the model does not fit the data well due to the large variability which cannot be explained by a simple linear regression model. The modest R-squared value suggests that factors beyond new cases likely influence new deaths, urging consideration of additional variables to enhance predictive accuracy.I might have initial ideas about how COVID-19 affects different areas or groups, based on stereotypes or what I see in the media. But I've tackled these biases using data, peer review, and clear methods. Even though there are challenges, the information I've gathered can help make specific plans and policies to control COVID-19 and improve public health.