COVID-19 Dataset Analysis

Sepideh Forouzi

Introduction

The COVID-19 pandemic has had profound impacts on public health, economies, and societies worldwide. Understanding the dynamics of COVID-19 spread and mortality within a specific region is crucial for guiding public health responses and policy decisions. California, being the most populous state in the United States, offers a significant case study for exploring these dynamics.

In this report, I used a combination of data preprocessing, aggregation, visualization, and simple linear regression modeling to analyze COVID-19 data for California. I began by importing two datasets containing cumulative cases and deaths, then cleaned and merged them by aggregating county-level data to the state level.

Challenges encountered included the cumulative nature of the data (necessitating differencing to estimate daily changes) and weekly reporting artifacts (requiring smoothing via rolling averages). To address the first challenge, I computed the daily new cases and deaths by differencing consecutive cumulative values. For the second challenge, I applied a 7-day rolling average to smooth the time series data, minimizing weekly reporting fluctuations and highlighting underlying trends. This methodology allowed for a more accurate understanding of the spread and mortality trends of COVID-19 in California, overcoming data irregularities and enabling clearer interpretation of pandemic dynamics.

1. Data Preprocessing

1.1 Import Datasets

```
cases_data <- read_csv("time_series_covid19_confirmed_US.csv", show_col_types = FALSE)
deaths_data <- read_csv("time_series_covid19_deaths_US.csv", show_col_types = FALSE)
```

I imported two datasets containing cumulative COVID-19 cases and deaths for U.S. counties.

1.2 Combining Datasets

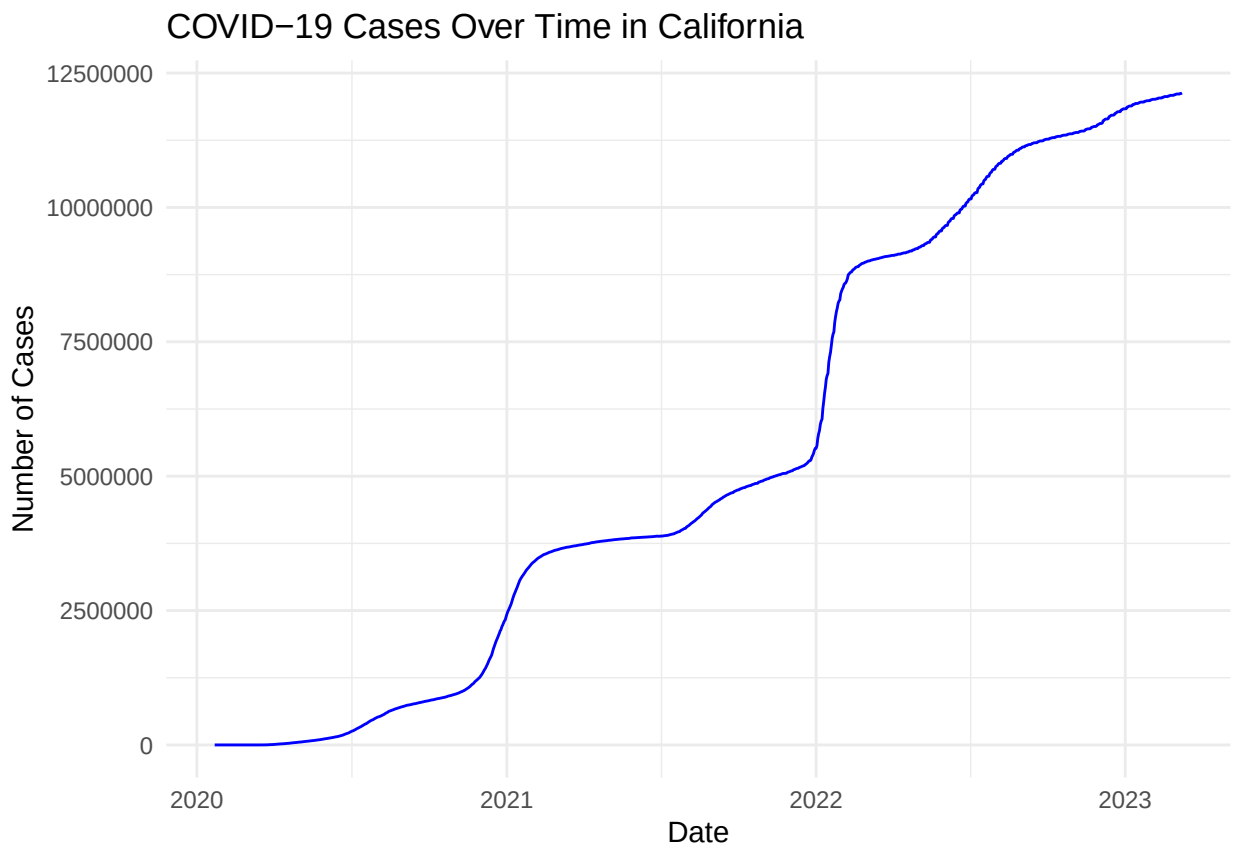
```
cases_california <- cases_data %>% filter(Province_State == "California")
deaths_california <- deaths_data %>% filter(Province_State == "California")

cases_dates <- names(cases_california)[12:ncol(cases_california)]
deaths_dates <- names(deaths_california)[13:ncol(deaths_california)]

total_cases <- colSums(cases_california[, cases_dates])
total_deaths <- colSums(deaths_california[, deaths_dates])
total_population <- sum(deaths_california$Population)

# Create the unified California dataset
df_california <- tibble(
  date = mdy(cases_dates),
```

```
  cases = total_cases,
  deaths = total_deaths,
  population = total_population
)
```

2. Exploratory Data Analysis (EDA)

2.1 Cases Over Time

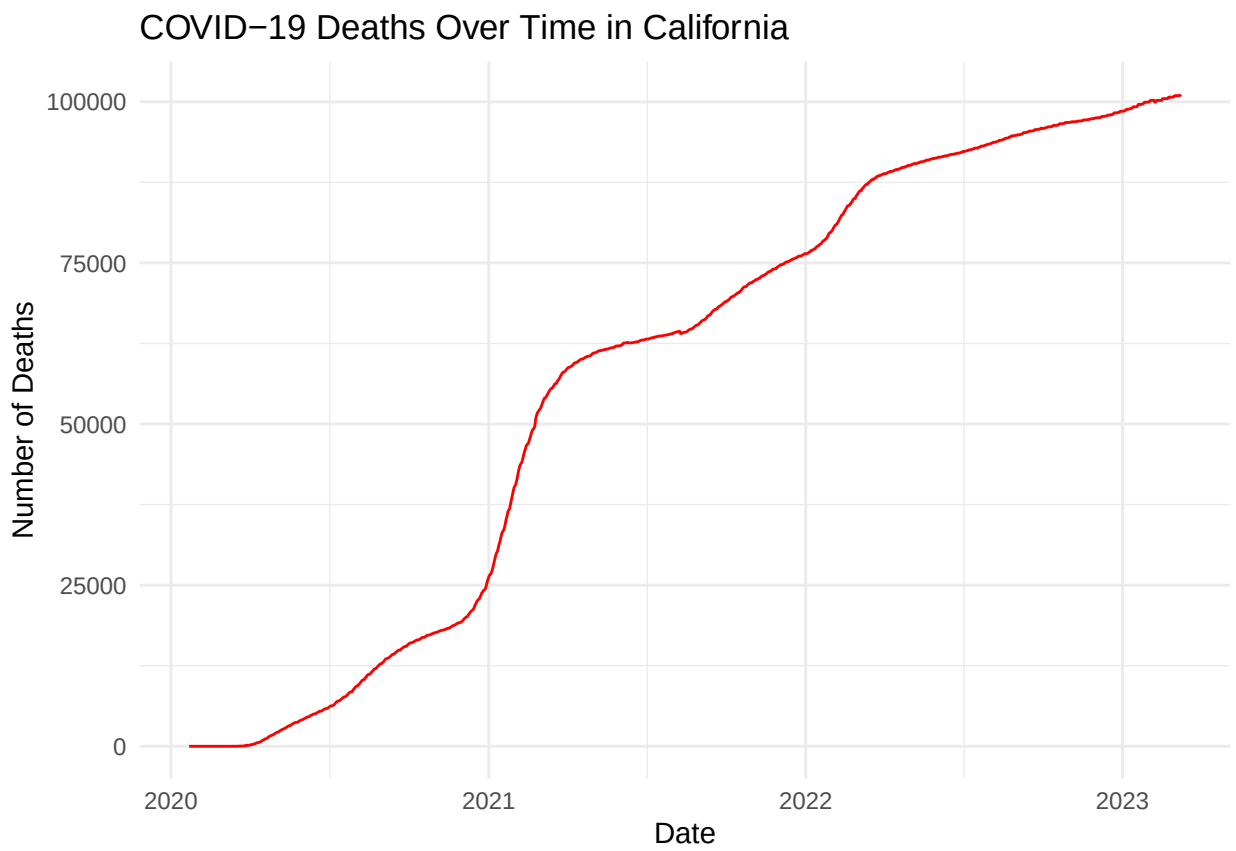
```
ggplot(df_california, aes(x = date, y = cases)) +
  geom_line(color = "blue") +
  labs(title = "COVID-19 Cases Over Time in California", x = "Date", y = "Number of Cases") +
  theme_minimal()
```



Analysis: The case growth reveals exponential phases, showing key pandemic waves over time. The case growth emphasizes not only major infection waves but also periods of relative stability, suggesting the impact of public health interventions such as lockdowns and vaccination campaigns.

2.2 Deaths Over Time

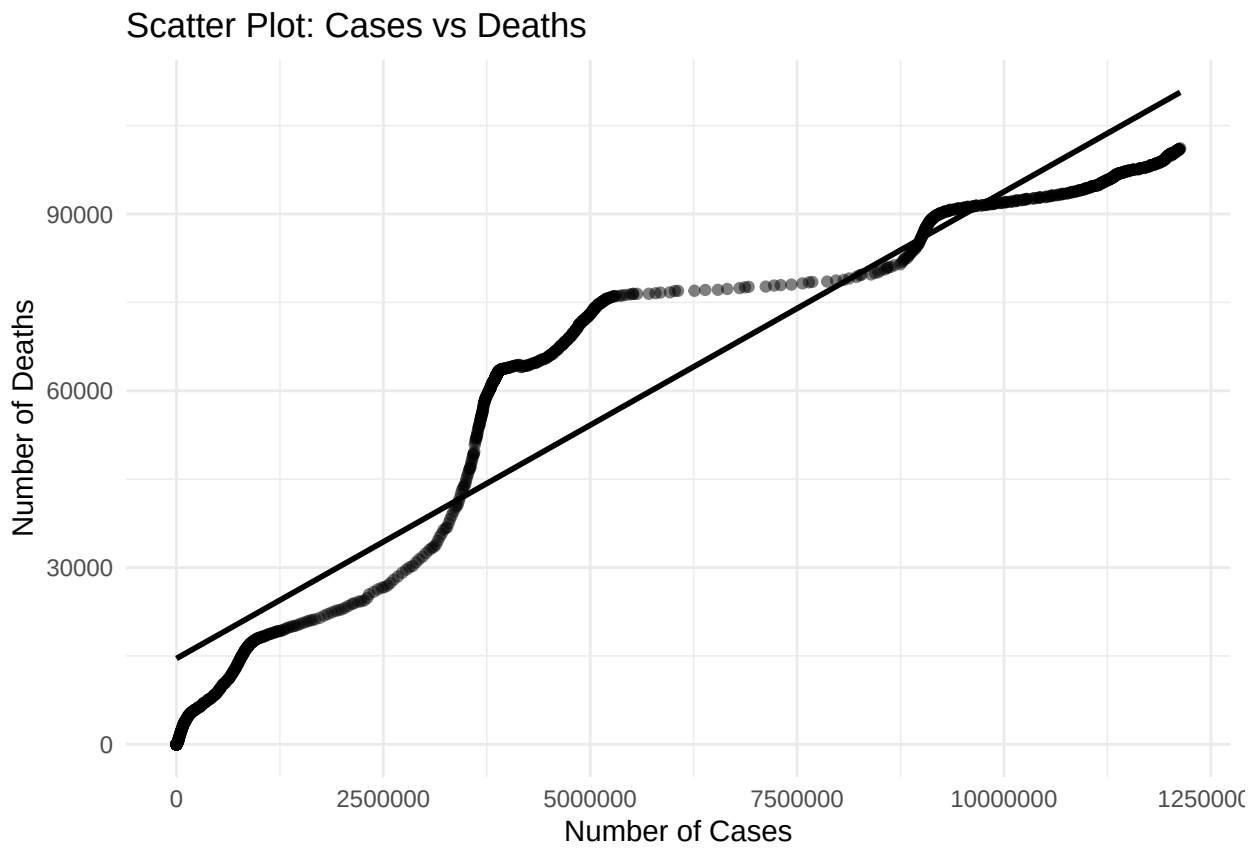
```
ggplot(df_california, aes(x = date, y = deaths)) +
  geom_line(color = "red") +
  labs(title = "COVID-19 Deaths Over Time in California", x = "Date", y = "Number of Deaths") +
  theme_minimal()
```



Analysis: Deaths trend closely follows cases with a slight lag, matching typical infection-to-mortality patterns. The time lag between cases and deaths is evident, underlining the disease progression timeline. Peaks in death counts provide insights into the critical burden faced by healthcare facilities.

2.3 Cases vs Deaths Scatter Plot

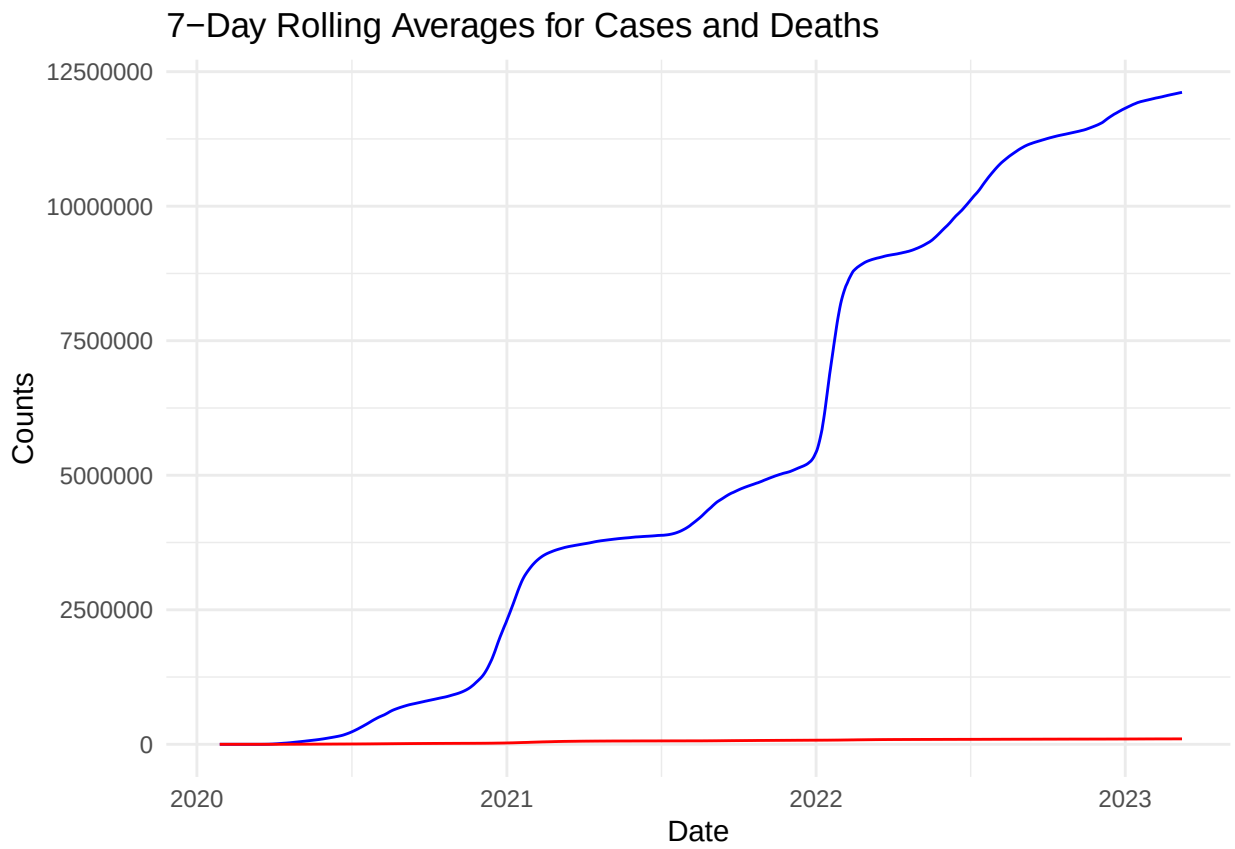
```
ggplot(df_california, aes(x = cases, y = deaths)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(title = "Scatter Plot: Cases vs Deaths", x = "Number of Cases", y = "Number of Deaths") +
  theme_minimal()
```



Analysis: A strong linear relationship appears between cases and deaths. The linear pattern in the scatter plot reinforces the proportional relationship between infection rates and mortality outcomes, although slight deviations could suggest improvements in medical treatments over time.

2.4 Rolling Averages

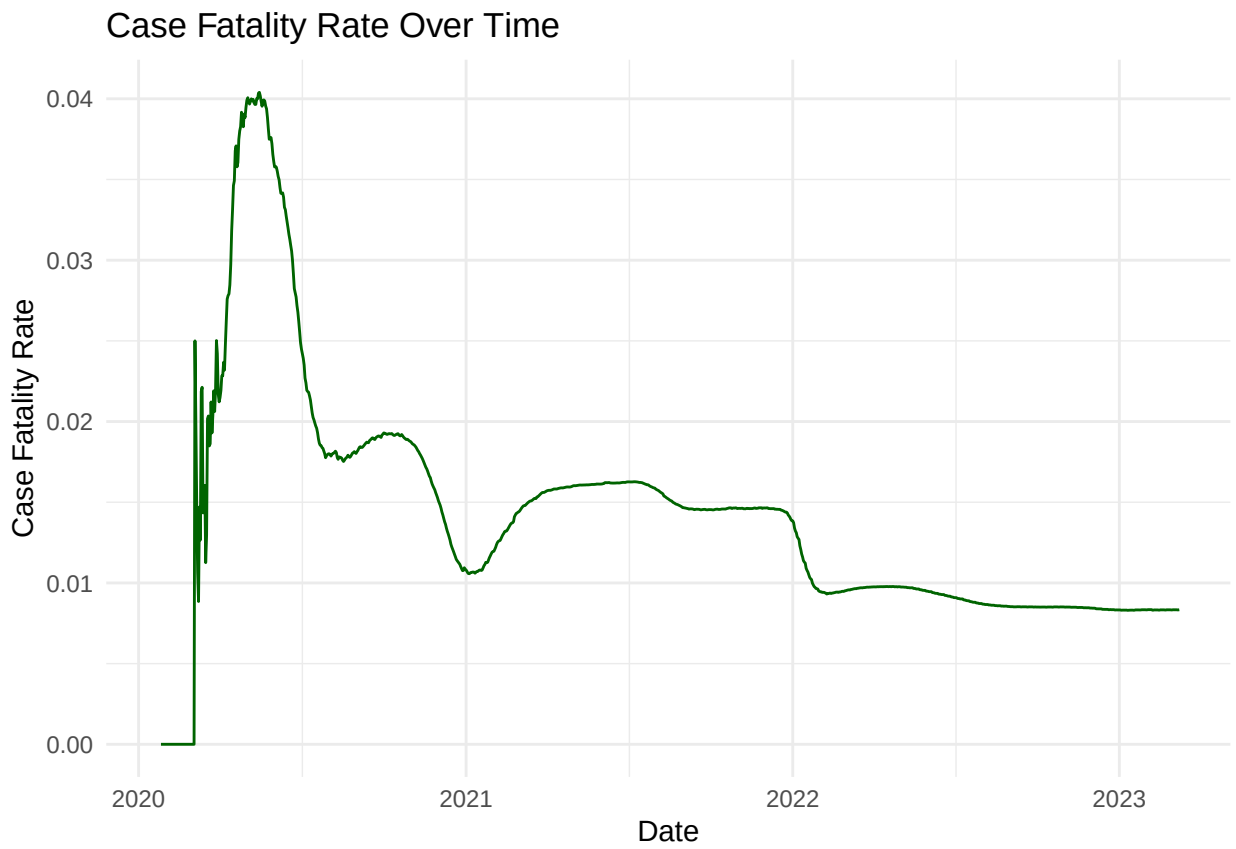
```
df_california <- df_california %>%
  mutate(
    cases_7d_avg = zoo::rollmean(cases, 7, fill = NA, align = "right"),
    deaths_7d_avg = zoo::rollmean(deaths, 7, fill = NA, align = "right")
  )

ggplot(df_california, aes(x = date)) +
  geom_line(aes(y = cases_7d_avg), color = "blue") +
  geom_line(aes(y = deaths_7d_avg), color = "red") +
  labs(title = "7-Day Rolling Averages for Cases and Deaths", x = "Date", y = "Counts") +
  theme_minimal()
```



Analysis: Smoothing helps detect real pandemic waves by removing weekly reporting noise.

2.5 Case Fatality Rate Over Time

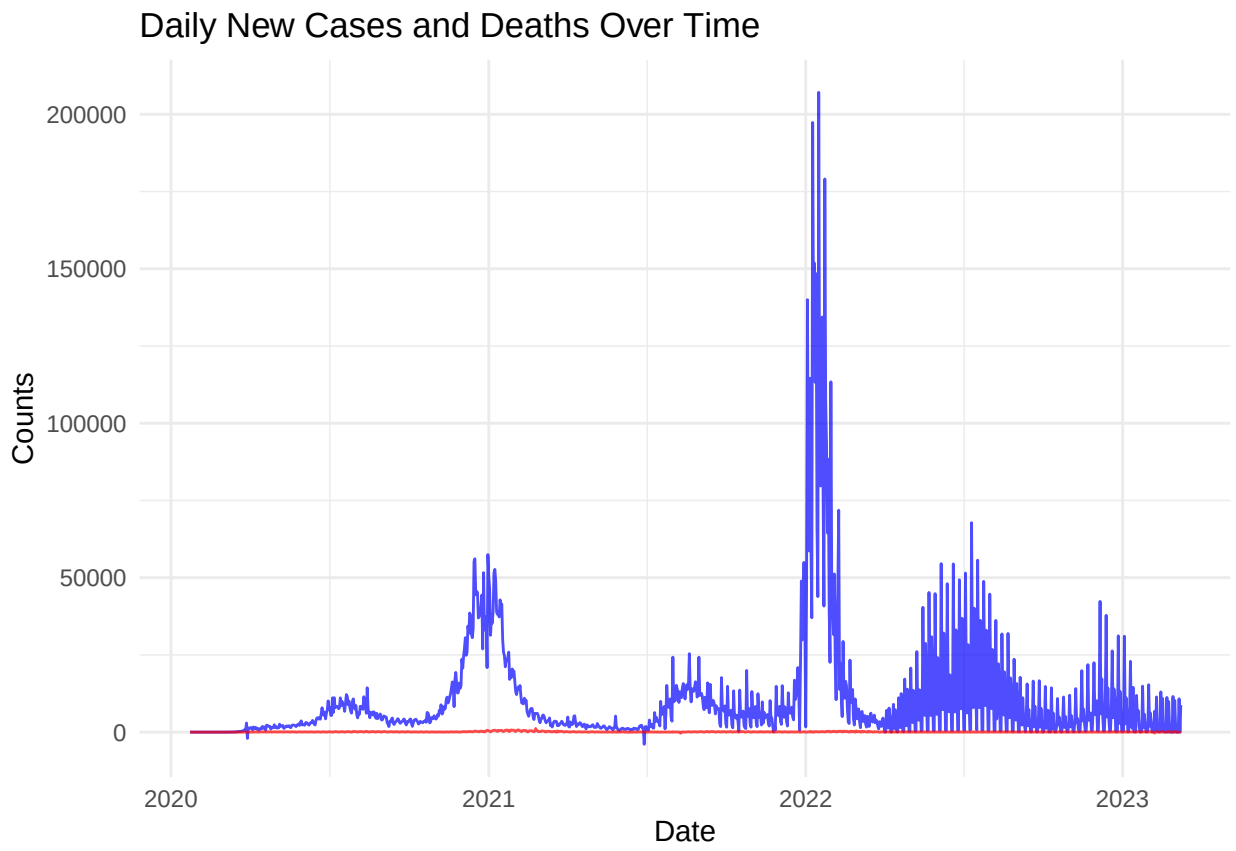
```
df_california <- df_california %>%
  mutate(case_fatality_rate = deaths / cases)

ggplot(df_california, aes(x = date, y = case_fatality_rate)) +
  geom_line(color = "darkgreen") +
  labs(title = "Case Fatality Rate Over Time", x = "Date", y = "Case Fatality Rate") +
  theme_minimal()
```



Analysis: The case fatality rate fluctuated over time, possibly due to treatment improvements or variant changes.

2.6 Daily New Cases and Deaths Over Time

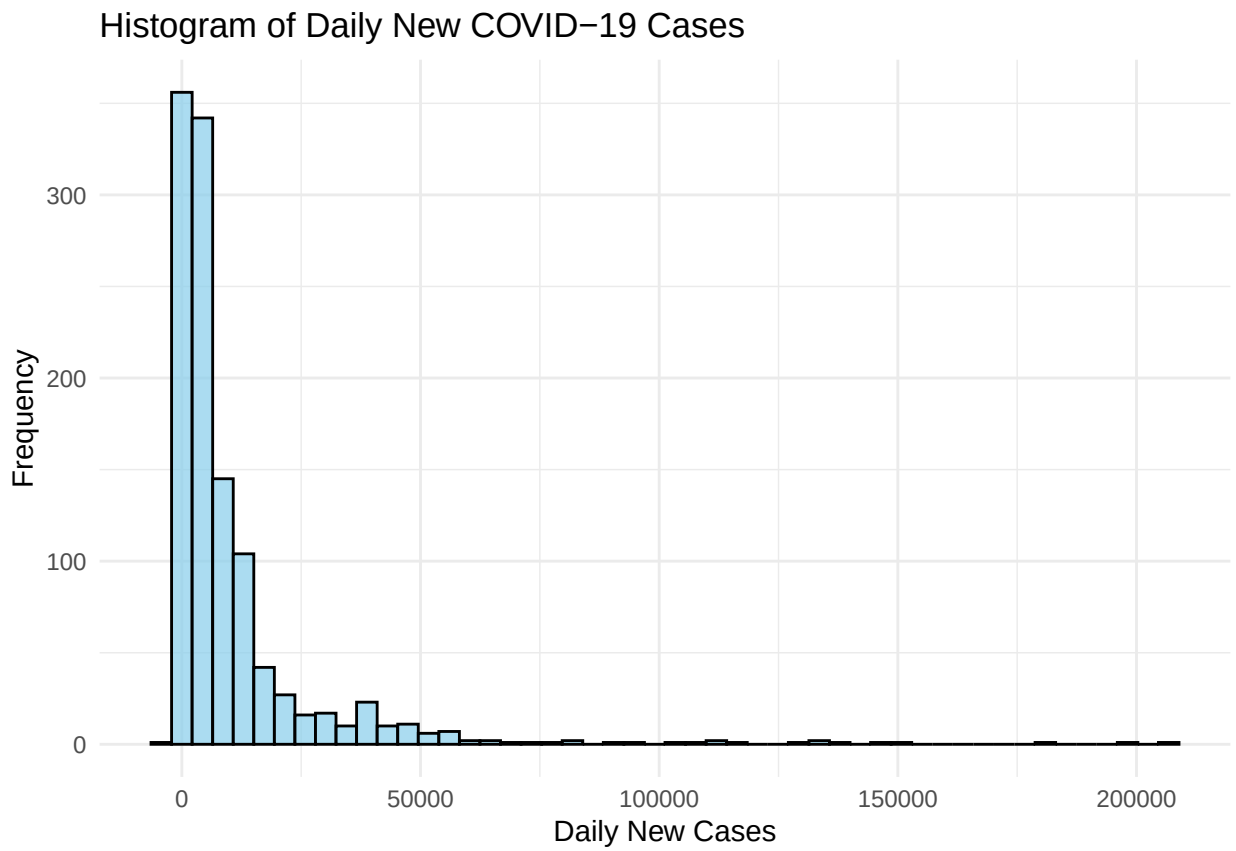
```
df_california <- df_california %>%
  mutate(
    daily_new_cases = c(NA, diff(cases)),
    daily_new_deaths = c(NA, diff(deaths))
  )

ggplot(df_california, aes(x = date)) +
  geom_line(aes(y = daily_new_cases), color = "blue", alpha = 0.7) +
  geom_line(aes(y = daily_new_deaths), color = "red", alpha = 0.7) +
  labs(title = "Daily New Cases and Deaths Over Time", x = "Date", y = "Counts") +
  theme_minimal()
```



Analysis: Day-to-day variations reveal weekly patterns and surges during specific periods.

2.7 Histogram of Daily New Cases

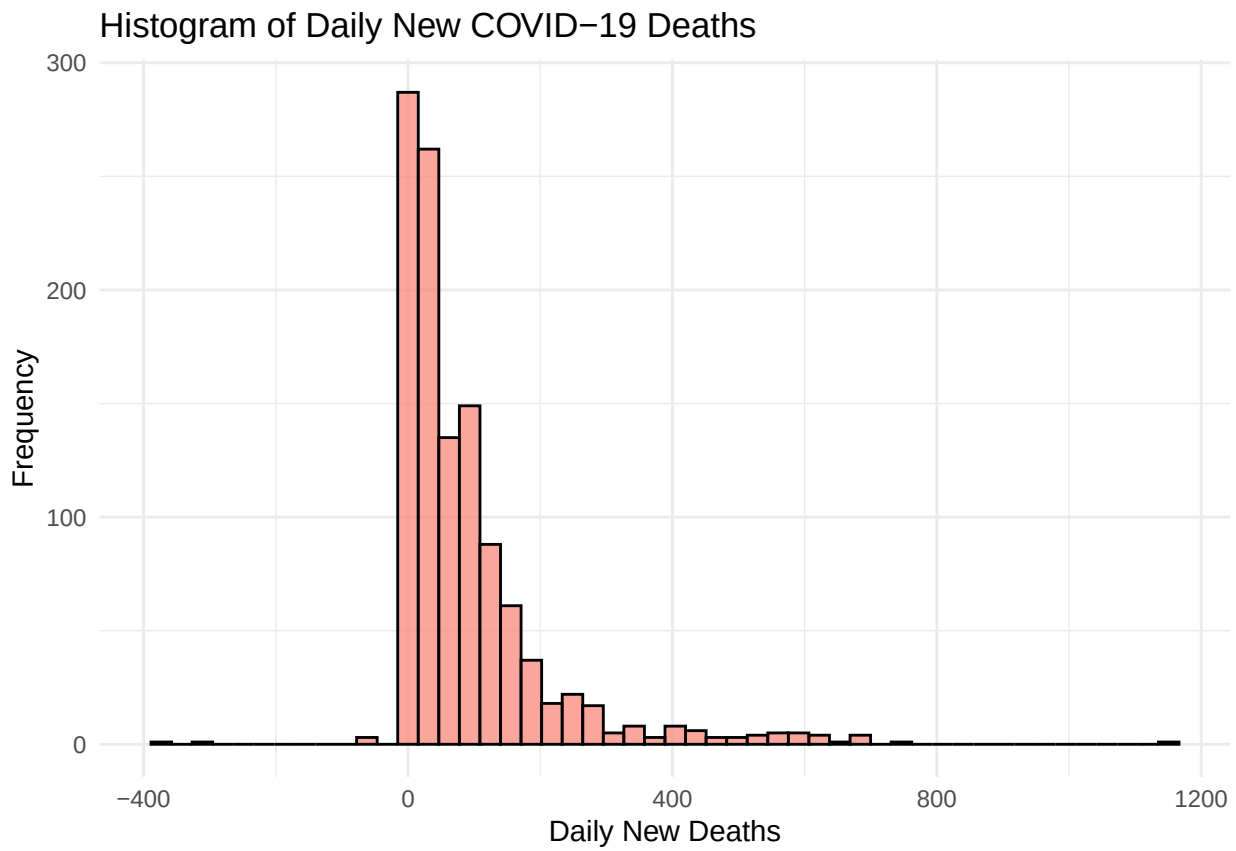
```
ggplot(df_california, aes(x = daily_new_cases)) +
  geom_histogram(bins = 50, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Daily New COVID-19 Cases", x = "Daily New Cases", y = "Frequency") +
  theme_minimal()
```



Analysis: Most days had moderate numbers of new cases, but extreme outbreak days are visible as a long right tail. The histogram's shape indicates that while most days experienced moderate case counts, the presence of rare but intense spikes reveals the system's vulnerability to super-spreader events or testing surges.

2.8 Histogram of Daily New Deaths

```
ggplot(df_california, aes(x = daily_new_deaths)) +
  geom_histogram(bins = 50, fill = "salmon", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Daily New COVID-19 Deaths", x = "Daily New Deaths", y = "Frequency") +
  theme_minimal()
```



Analysis: Death distributions show a sharp skew with some major death spikes during severe waves. The pronounced skewness in daily deaths shows that although mortality was manageable most days, some dates witnessed exceptional death tolls, likely corresponding to overwhelmed healthcare systems or reporting backlogs.

3. Modeling

3.1 Linear Regression

```
## Summary of model:

##
## Call:
## lm(formula = deaths ~ cases, data = df_california)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14562  -9072  -4625  14038  19911
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.456e+04  5.446e+02   26.73   <2e-16 ***
## cases       7.922e-03  7.935e-05   99.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 11390 on 1141 degrees of freedom
## Multiple R-squared:  0.8973, Adjusted R-squared:  0.8972
## F-statistic:  9967 on 1 and 1141 DF,  p-value: < 2.2e-16
```

```
##
##
## Tidy model output:
```

```
## # A tibble: 2 x 5
##   term           estimate   std.error statistic   p.value
##   <chr>             <dbl>       <dbl>     <dbl>     <dbl>
## 1 (Intercept) 14561.      545.             26.7 1.17e-122
## 2 cases           0.00792   0.0000793      99.8 0
```

3.2 Model Interpretation

```
## Regression Equation:
```

```
## Predicted Deaths = 14561.30 + 0.00792 × Cases
```

Analysis: The simple linear regression suggests that every 1000 new cases predict approximately 8 additional deaths.

Conclusion

Through cumulative trends, daily variations, fatality rates, and regression analysis, I deeply examined COVID-19's impacts in California. I found strong evidence that case surges directly translated into mortality burdens, albeit with timing lags and policy influences.

This type of COVID-19 data analysis is essential for public health planning. Health authorities can use such analyses to anticipate resource needs, forecast hospital demand, evaluate the effects of public health measures, and guide vaccination strategies. Data-driven decisions are critical for saving lives and preventing healthcare system collapse.

Moreover, by understanding the dynamics between cases and deaths, we can help identify when health systems might become overwhelmed, when intervention strategies should be intensified, and when resources like ICU beds and ventilators should be allocated more proactively.

This analysis also highlights how public health strategies such as early testing, vaccination, and masking can alter the trajectory of cases and deaths. If more sophisticated models are developed using these foundational analyses, predictive tools could be built that help optimize emergency responses for future pandemic waves or different infectious diseases.

From an epidemiological standpoint, this work underscores the importance of monitoring not just cumulative numbers but also dynamic indicators like daily new cases, death rates, and case fatality trends, ensuring timely and evidence-based actions are taken.