Covid 19 Analysis

Catherine Phan

2023-10-04

Introduction

The COVID-19 pandemic has made an unprecedented global impact, affecting both economies and public health systems worldwide. To gain insights into the virus's spread and to take measures for containment, data has been diligently gathered. My study is centered around identifying counties with the highest COVID-19 case counts and fatalities, within California. The dataset I will be working with is a comprehensive collection of four COVID-19 data extracted from the John Hopkins University GitHub repository. The data sets concern global cases, global deaths, US cases, and US deaths. However, for this analysis, only the US related date sets will be used.

Goals of the Analysis

The main goal of this analysis is to identify some counties within California with the highest COVID-19 case counts and fatalities. Following this identification, we will conduct an in-depth examination of those specific counties. By providing these insights, we aim to empower healthcare professionals with valuable information that can be used to optimize resource allocation effectively. I chose California as the main concentration in this study because of my strong connection to the State as I was born and raised here.

Libraries Needed for the Analysis

```
library(tidyverse)
library(lubridate)
library(dplyr)
library(tibble)
library(forecast)
library(knitr)
library(git2r)
library(gt)
```

Cleaning the Data

```
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", "time_seri
urls <- str_c(url_in, file_names)
US_cases <- read_csv(urls[1])
US_deaths <- read_csv(urls[2])



#transform the data from wide format to long format and remove columns not needed for the analysis.
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_))

#joined the US_deaths data set with US_cases using the full_join function
us <- US_cases %>%
  full_join(US_deaths)

#final step of cleaning, only look at cases greater than zero in the US
us <- us %>% filter(cases > 0)
```

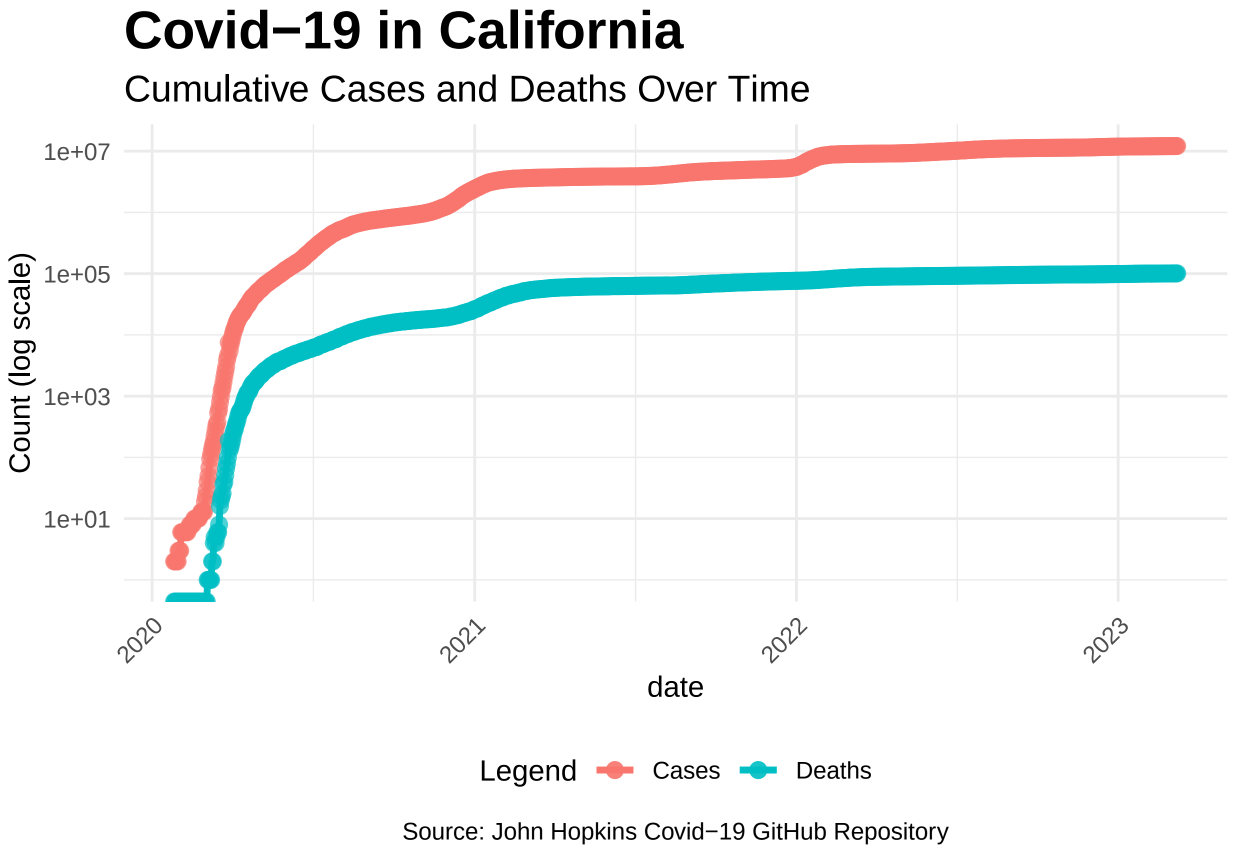
With the model our professor shared, I've gather the two US COVID-19 data sets from the John Hopkins GitHub. These data sets have been merged into the variable *us* containing data on COVID-19 cases and deaths in the United States.

Preparing the Data to Create Visuals

```
#into a new variable US_state, create a new column that calculates deaths per million
US_state <- us %>%
  group_by(Province_State, Country_Region, date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population)) %>%
  mutate(deaths_per_mil = deaths * 1000000 / Population) %>%
  select(Province_State, Country_Region, date, cases, deaths, deaths_per_mil, Population) %>%
  ungroup()
```

```
#Looking deeper into California filtering for cases greater than zero
state <- "California"

US_state %>%
  filter(Province_State == state, cases > 0) %>%
  ggplot(aes(x = date)) +
  geom_line(aes(y = cases, color = "Cases"), size = 1.2) +
  geom_point(aes(y = cases, color = "Cases"), size = 2.5, alpha =.7) +
  geom_line(aes(y = deaths, color = "Deaths"), size = 1.2) +
  geom_point(aes(y = deaths, color = "Deaths"), size = 2.5, alpha = .7, shape = 19) +
  scale_y_log10() +
  labs(title = paste("Covid-19 in", state),
       subtitle = "Cumulative Cases and Deaths Over Time",
       y = "Count (log scale)",
       color = "Legend",
       caption = "Source: John Hopkins Covid-19 GitHub Repository")+
    theme_minimal() +
  theme(
    legend.position = "bottom",
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(size = 20, face = "bold"),
    plot.subtitle = element_text(size = 14),
    plot.caption = element_text(hjust = 0.5)
  )
```


Covid−19 in California

Cumulative Cases and Deaths Over Time

Source: John Hopkins Covid−19 GitHub Repository

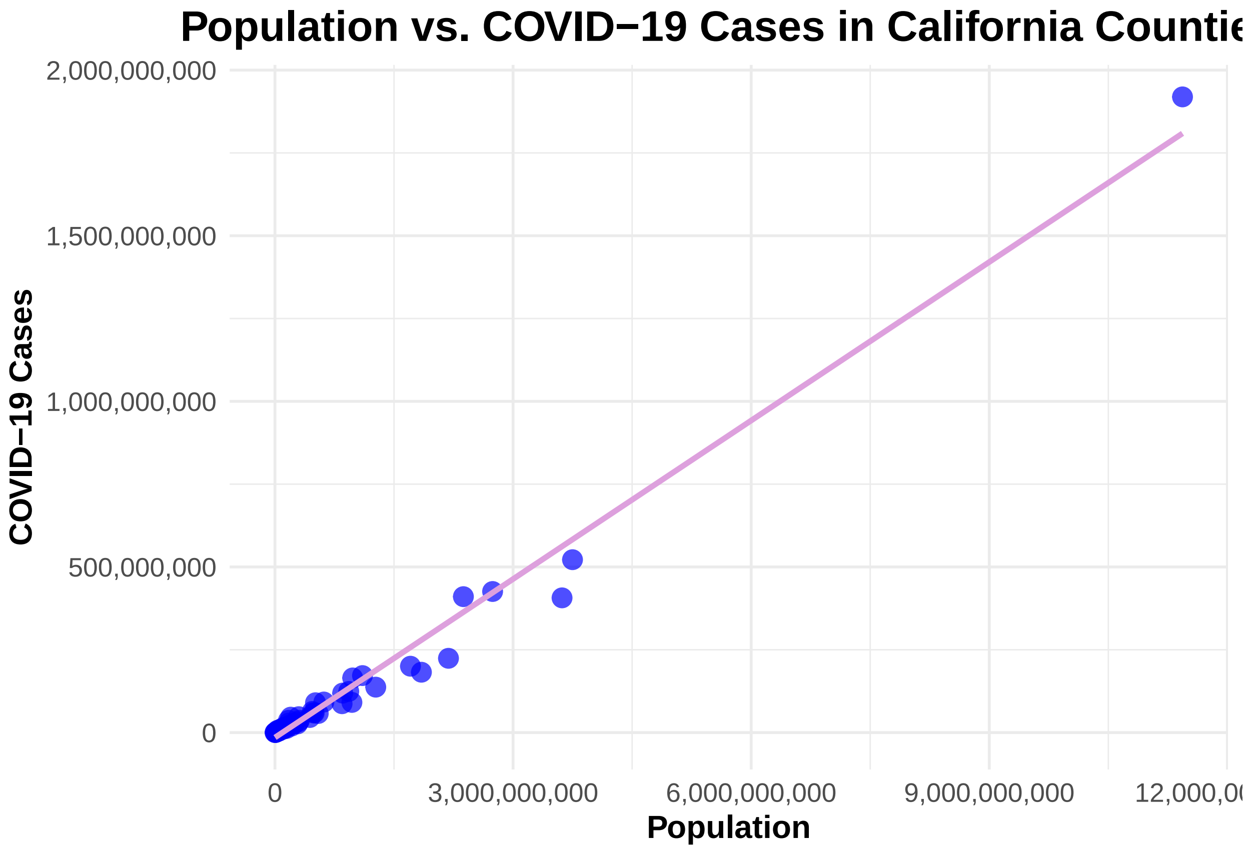
By analyzing the trajectory of Covid-19 in California, we uncover several encouraging trends. Over time, we notice a consistent pattern: the number of cases significantly outweighs the number of fatalities, a promising sign. Additionally, around the onset of 2022, there's a noticeable trend of both cases and deaths stabilizing. This could suggest that the spread of Covid-19 has been better managed, thanks to advances in technology and healthcare infrastructure. Healthcare facilities have likely gained a better understanding of how to treat Covid-19 patients effectively and implement measures to control its transmission. However, looking at California as a whole can be daunting. In the code chunks below, my goal is to narrow down to the top *7* counties in California that has the highest amount of deaths.

```
#All counties in California
all_cali <- us %>%
   filter(Province_State == "California", Admin2 != "Unassigned") %>%
  group_by(Admin2) %>%
  summarize(
    Population = sum(Population),
    deaths = sum(deaths),
    cases = sum(cases)) %>%
  mutate(deaths_per_thousands = deaths * 1000 / Population,
         cases_per_thousands = cases * 1000 / Population) %>%
  select(Admin2, Population, cases, deaths, deaths_per_thousands, cases_per_thousands)

#Renaming Admin2 to a more intuitive variable name, such as County
```

```
all_cali <- all_cali %>%
  rename("County" = Admin2)

#A scatter plot of population and number of covid-19 cases in California Counties
# Create a ggplot scatter plot with improved aesthetics
ggplot(all_cali, aes(x = Population, y = cases)) +
  geom_point(color = "blue", size = 3, alpha = 0.7) +  # Adjust alpha for transparency
  labs(
    title = "Population vs. COVID-19 Cases in California Counties",
    x = "Population",
    y = "COVID-19 Cases"
  ) +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Adjust title size and style
    axis.title.x = element_text(size = 12, face = "bold"),  # Adjust x-axis label size and style
    axis.title.y = element_text(size = 12, face = "bold"),  # Adjust y-axis label size and style
    axis.text = element_text(size = 10),  # Adjust axis tick label size
    legend.position = "none"  # Remove legend
  ) +
  geom_smooth(method = "lm", color = "plum", se = FALSE)
```



```
#Table of all the Covid Cases and Deaths across Counties of California
all_cali %>%
  gt() %>%
  fmt_number(
    columns = c(Population, cases, deaths),
    decimals = 0
  )
```

County	Population	cases	deaths	deaths_per_thousands	cases_per_thousands
Alameda	1,845,147,216	182,250,215	1,357,082	0.7354871	98.77272
Alpine	1,211,417	93,941	0	0.0000000	77.54638
Amador	43,091,168	5,999,897	58,876	1.3663125	139.23728
Butte	237,597,624	26,179,854	273,706	1.1519728	110.18567
Calaveras	50,265,975	4,692,416	77,047	1.5327863	93.35174
Colusa	23,206,119	3,090,957	15,745	0.6784848	133.19577
Contra Costa	1,270,032,126	137,155,292	930,862	0.7329437	107.99356
Del Norte	29,786,652	3,826,227	29,352	0.9854078	128.45442
El Dorado	209,234,655	19,610,886	138,169	0.6603543	93.72676
Fresno	1,103,007,504	172,083,673	1,902,474	1.7248060	156.01315
Glenn	30,579,261	4,388,805	29,696	0.9711157	143.52227
Humboldt	146,673,756	12,612,622	88,721	0.6048867	85.99099
Imperial	198,067,995	46,591,814	712,501	3.5972546	235.23141
Inyo	19,391,925	2,846,057	40,163	2.0711198	146.76506
Kern	979,419,776	164,977,425	1,599,756	1.6333711	168.44404
Kings	170,681,040	36,915,826	302,652	1.7732022	216.28545
Lake	68,764,248	7,321,778	86,537	1.2584592	106.47652
Lassen	32,804,829	6,899,164	37,396	1.1399541	210.30940
Los Angeles	11,434,542,873	1,919,132,962	24,114,001	2.1088732	167.83644
Madera	172,745,046	28,672,653	255,274	1.4777500	165.98249
Marin	283,932,122	26,218,381	228,723	0.8055552	92.34031
Mariposa	17,994,338	1,893,082	18,966	1.0539982	105.20431
Mendocino	94,209,414	10,088,242	80,541	0.8549146	107.08316
Merced	299,616,720	48,055,877	559,016	1.8657704	160.39117
Modoc	8,425,473	700,209	5,416	0.6428126	83.10619
Mono	15,657,296	1,998,577	4,386	0.2801250	127.64509
Monterey	471,824,307	63,975,642	503,798	1.0677661	135.59209
Napa	149,176,752	17,549,873	96,875	0.6493974	117.64483
Nevada	111,426,335	10,805,678	83,251	0.7471394	96.97598
Orange	3,617,113,188	406,745,946	5,099,694	1.4098796	112.45043
Placer	439,356,887	45,567,333	399,897	0.9101872	103.71371
Plumas	20,217,525	1,987,765	7,491	0.3705201	98.31891
Riverside	2,742,306,060	425,952,399	4,637,284	1.6910162	155.32635
Sacramento	1,707,263,800	200,316,159	2,148,685	1.2585548	117.33170
San Benito	69,025,992	9,387,650	69,900	1.0126620	136.00167
San Bernardino	2,374,112,565	410,418,540	5,187,328	2.1849545	172.87240
San Diego	3,748,944,590	521,934,721	3,823,439	1.0198708	139.22178
San Francisco	969,703,900	91,099,953	631,515	0.6512452	93.94616
San Joaquin	852,843,612	119,263,728	1,561,443	1.8308667	139.84244
San Luis Obispo	308,590,990	37,265,870	318,921	1.0334748	120.76137
San Mateo	845,530,019	87,141,865	547,775	0.6478481	103.06182
Santa Barbara	485,344,413	60,031,868	488,978	1.0074866	123.68921
Santa Clara	2,186,184,168	224,287,159	1,772,574	0.8108073	102.59298
Santa Cruz	299,987,874	32,736,936	191,563	0.6385691	109.12753
Shasta	197,367,680	23,765,934	343,600	1.7409132	120.41452
Sierra	3,068,105	189,107	1,704	0.5553917	61.63642

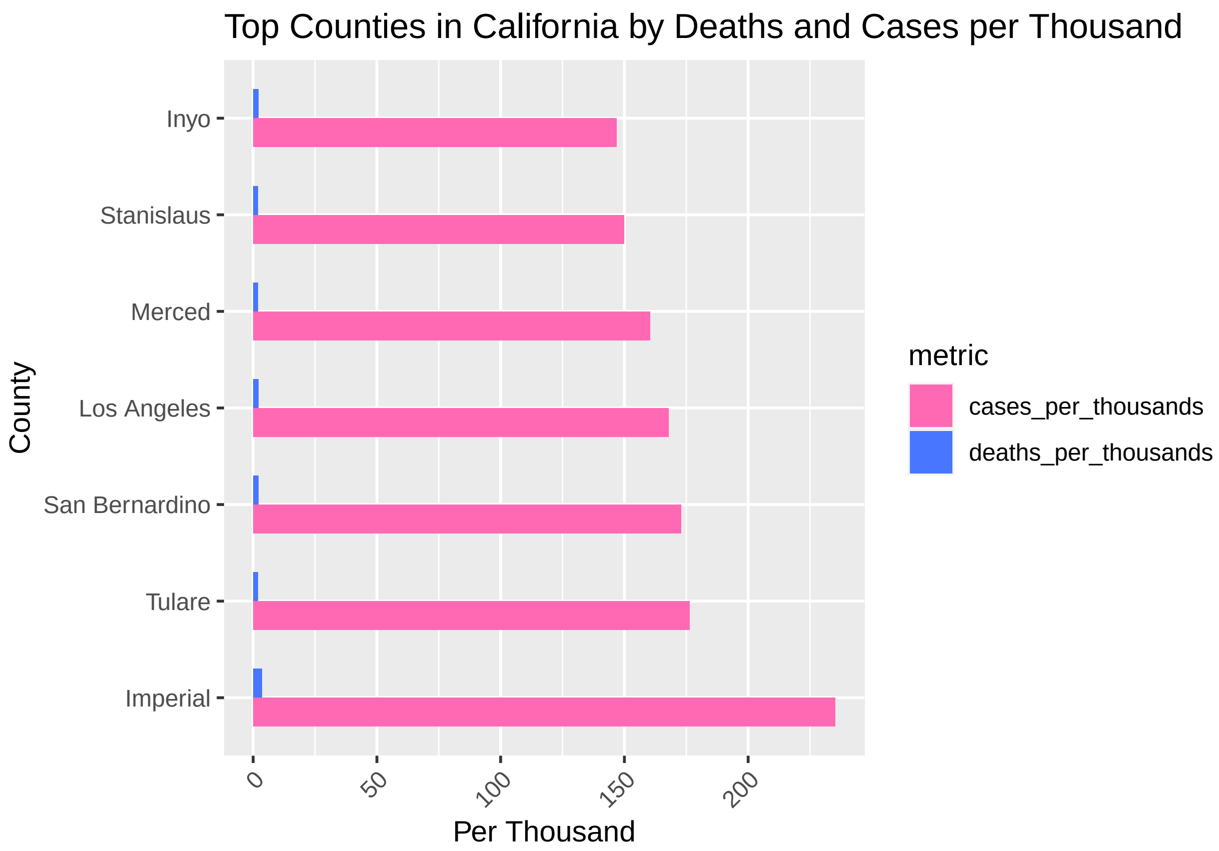
Siskiyou	47,152,737	4,492,782	46,429	0.9846512	95.28147
Solano	494,197,872	58,953,724	301,819	0.6107250	119.29174
Sonoma	545,252,608	57,615,292	358,923	0.6582692	105.66716
Stanislaus	616,188,540	92,291,103	1,184,638	1.9225252	149.77738
Sutter	104,728,680	15,106,240	147,389	1.4073413	144.24167
Tehama	69,770,048	9,459,379	112,494	1.6123538	135.57937
Trinity	12,714,975	941,277	11,264	0.8858846	74.02901
Tulare	510,949,720	90,121,244	1,002,885	1.9627861	176.37987
Tuolumne	58,345,938	8,448,037	104,812	1.7963890	144.79220
Ventura	929,760,594	124,193,332	1,041,284	1.1199485	133.57560
Yolo	242,329,500	26,581,890	239,888	0.9899249	109.69317
Yuba	85,276,112	11,156,686	75,310	0.8831313	130.83014

```r
#Looking at the top seven counties with the highest deaths per thousand statistic
cali <- us %>%
  filter(Province_State == "California", Admin2 != "Unassigned") %>%
  group_by(Admin2) %>%
  summarize(
    Population = sum(Population),
    deaths = sum(deaths),
    cases = sum(cases)) %>%
  mutate(deaths_per_thousands = deaths * 1000 / Population,
         cases_per_thousands = cases * 1000 / Population) %>%
  select(Admin2, Population, cases, deaths, deaths_per_thousands, cases_per_thousands) %>%
  slice_max(deaths_per_thousands, n = 7)

cali <- cali %>%
  rename("County" = Admin2)

#Creating a visual using ggplot of the deaths and cases per thousands count for these seven counties
cali_long <- cali %>%
  gather(metric, per_thousand, deaths_per_thousands, cases_per_thousands)

ggplot(cali_long, aes(x = reorder(County, -per_thousand), y = per_thousand, fill = metric)) +
  geom_bar(stat = "identity", position = position_dodge(), width = 0.6) +
  labs(
    title = "Top Counties in California by Deaths and Cases per Thousand",
    x = "County",
    y = "Per Thousand"
  ) +
  scale_fill_manual(values = c("deaths_per_thousands" = "royalblue1", "cases_per_thousands" = "hotpink"
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



```
#Create a table showing the top seven counties in California with the highest deaths per thousand
cali %>%
  gt() %>%
  fmt_number(
    columns = c(Population, cases, deaths),
    decimals = 0
  )
```

County	Population	cases	deaths	deaths_per_thousands	cases_per_thousands
Imperial	198, 067, 995	46, 591, 814	712, 501	3.597255	235.2314
San Bernardino	2, 374, 112, 565	410, 418, 540	5, 187, 328	2.184955	172.8724
Los Angeles	11, 434, 542, 873	1, 919, 132, 962	24, 114, 001	2.108873	167.8364
Inyo	19, 391, 925	2, 846, 057	40, 163	2.071120	146.7651
Tulare	510, 949, 720	90, 121, 244	1, 002, 885	1.962786	176.3799
Stanislaus	616, 188, 540	92, 291, 103	1, 184, 638	1.922525	149.7774
Merced	299, 616, 720	48, 055, 877	559, 016	1.865770	160.3912

When looking at California as a whole, we observe that as population in an area increases, there tend to be more covid-19 cases which makes intuitive sense. When we narrow down our counties, we learn that Imperial, Tulare, San Bernardino, Los Angeles, Merced, Stanislaus, and Inyo counties in California have consistently exhibited the highest incidence rates, both in terms of Covid-19 cases and related deaths, per one thousand residents. This critical insight underscores the significance of targeted interventions and resource allocation in these specific areas.

Model

```
model_1 <- lm(deaths ~ Population + cases + Population*cases, data = all_cali)
summary(model_1)
```

```
##
## Call:
## lm(formula = deaths ~ Population + cases + Population * cases,
##     data = all_cali)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1616150   -53505      488    27135  1242540
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.237e+03  5.182e+04  -0.062  0.95042
## Population       -5.333e-04  2.239e-04  -2.382  0.02077 *
## cases             1.354e-02  1.788e-03   7.572 4.85e-10 ***
## Population:cases  1.914e-13  5.653e-14   3.385  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 316700 on 54 degrees of freedom
## Multiple R-squared:  0.9913, Adjusted R-squared:  0.9908
## F-statistic:  2051 on 3 and 54 DF,  p-value: < 2.2e-16
```

In this analysis, a linear regression model was constructed to forecast the number of deaths in California counties by considering both their population and the incidence of COVID-19 cases. The presence of a small p-value associated with the predictor variables indicates their statistical significance in predicting the death toll within these counties. This outcome aligns with our expectations, as counties experiencing a higher number of COVID-19 cases tend to witness a correspondingly greater number of deaths. Furthermore, the statistical significance of the interaction term between 'Population' and 'Cases' underscores the complexity of the relationship between these variables.

Bias Identification and Conclusion

In this analysis, my primary objective was to identify the counties in California most profoundly impacted by Covid-19, focusing on both death and case counts. Despite my personal connection and biased towards San Francisco, a prominent Californian county, I made an effort to remain impartial and avoid and bias that could lead me to exclusively concentrate on my home city.

My analysis instead has a broader perspective, encompassing the entirety of California. I discovered that, in general, the number of reported cases consistently surpassed the number of deaths. Additionally, I learned that approximately two years after the onset of the pandemic, there was a notable trend towards the stabilization of both case and death rates. While there may be various reasons for this trend, one plausible explanation could be the result of improved healthcare infrastructure and vaccination efforts, resulting in reduced transmissions over time.

In addition to looking at California as a whole, I narrowed my focus to the seven counties that exhibited the highest rates of Covid-19 related death and case counts. These seven counties ended up being Imperial, Tulane, San Bernardino, Los Angeles, Merced, Stanislaud, and Inyo. I believe that by concentrating efforts into these particular areas, healthcare providers can better address the challenges posed by Covid-19 and provide a more timely and effective care to those who reside in those communities.

Appendix

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22621)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/Los_Angeles
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] gt_0.9.0        git2r_0.32.0    knitr_1.44      forecast_8.21.1
##  [5] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.0   dplyr_1.1.3
##  [9] purrr_1.0.2     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1
## [13] ggplot2_3.4.3   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.4      xfun_0.40         lattice_0.21-8    tzdb_0.4.0
##  [5] quadprog_1.5-8    vctrs_0.6.3       tools_4.3.1       generics_0.1.3
##  [9] curl_5.1.0        parallel_4.3.1    fansi_1.0.4       xts_0.13.1
## [13] pkgconfig_2.0.3   Matrix_1.5-4.1    lifecycle_1.0.3   compiler_4.3.1
## [17] farver_2.1.1      munsell_0.5.0     htmltools_0.5.6   yaml_2.3.7
## [21] pillar_1.9.0      crayon_1.5.2      nlme_3.1-162      fracdiff_1.5-2
## [25] tidyselect_1.2.0  digest_0.6.33     stringi_1.7.12    labeling_0.4.3
## [29] tseries_0.10-54   splines_4.3.1     fastmap_1.1.1     grid_4.3.1
## [33] colorspace_2.1-0  cli_3.6.1         magrittr_2.0.3    utf8_1.2.3
## [37] withr_2.5.1       scales_1.2.1      bit64_4.0.5       timechange_0.2.0
## [41] TTR_0.24.3        rmarkdown_2.25    quantmod_0.4.25   bit_4.0.5
## [45] nnet_7.3-19       timeDate_4022.108 zoo_1.8-12        hms_1.1.3
## [49] urca_1.3-3        evaluate_0.22     lmtest_0.9-40     mgcv_1.8-42
## [53] rlang_1.1.1       Rcpp_1.0.11       glue_1.6.2        xml2_1.3.5
## [57] rstudioapi_0.15.0 vroom_1.6.4       R6_2.5.1
```