DATA607 Final Project

Muyan Cheng | 120429798

December 06, 2024

1 Project Overview

For this project, I used the COVID-19 Data Hub[1] as the object of analysis to investigate the development of the COVID-19 epidemic in the United States and the impact of policy on the epidemic.
I first analyzed the whole dataset to identify the COVID-19 epidemic development in the United States as the primary object of study. I then conducted an exploratory data analysis to analyze the curve of change in the U.S. COVID-19 epidemic and the impact of policy on new confirmed during 2020.

2 Data Preprocessing

2.1 Dataset

The dataset I used is the COVID-19 Data Hub[1].This dataset collects COVID-19 fine-grained case data from around the world. This dataset contains multiple dimensions, with epidemiologically relevant variables such as the number of cases, deaths, and recoveries, as well as data related to policy implementation and population. The scope of the data covers almost all countries and regions and contains data from the national to the city level. The dataset is updated daily.
I chose the state-level COVID-19 data for analysis, which includes over 860,000 data instances with 47 variables.

2.2 Data Inspection

First, I inspected the data and it can be seen that the dataset contains 862,520 instances of data and 47 variables. Because this is a huge and realistic dataset, it will have many missing values that need to be dealt with. Therefore, I have plotted the missing value situation of this dataset and can find from Figure 1 that most variables have more or

less missing values except for a few variables such as ID, date and so on which have no missing values.

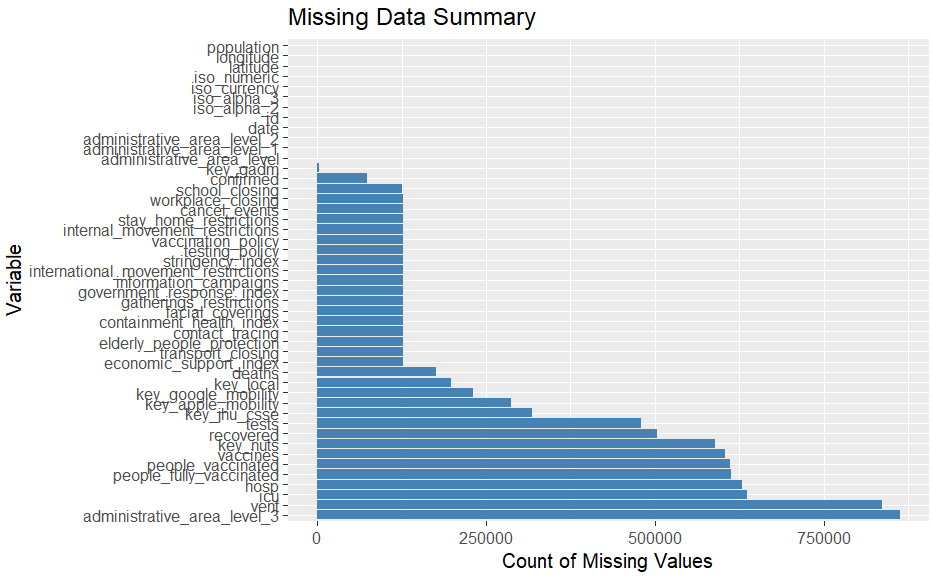


Figure 1: Missing Data Summary

Since the date variable has no missing values, I considered using it as a basis for observing the distribution of the data, so I plotted Figure 2 to observe the changes in the data instances from day to day.

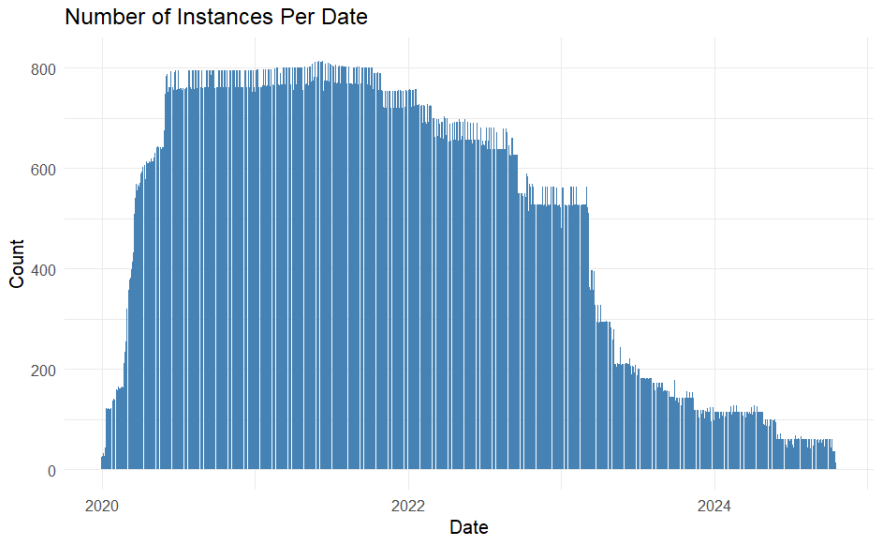


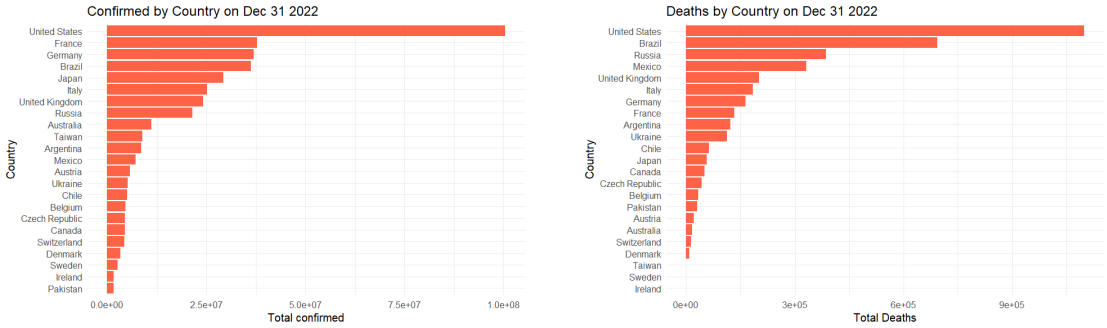
Figure 2: Number of Instances Per Date

The daily data volume in the dataset shows a relatively clear downward trend after 2023, indicating that the COVID-19 epidemic is gradually ending from 2023 onward.

Therefore, I selected December 31, 2022 as an important time point to observe the status of COVID-19 epidemics in each country.

2.3 Data Filitering

Next, I plotted the rankings of cumulative confirmed cases and cumulative deaths for each country on December 31, 2022, as shown in Figures 3, and it can be seen that the United States is ranked first in both rankings. Therefore, I chose the COVID-19 epidemic in the United States as the main object of study.



(a) Confirmed by Country on Dec 31 2022

(b) Deaths by Country on Dec 31 2022

Figure 3: Confirmed and Deaths by Country on Dec 31 2022

2.4 Data Cleaning

After selecting the data range, I started the data cleaning process. I plotted the summary of missing variables for the U.S. COVID-19 data, and I can see that some of the more important cumulative variables, such as confirmed, deaths and recovered, have missing values, and that there are also missing values for the related variables involving policy, such as school_closing and workplace_closing. Therefore, I need to deal with them.

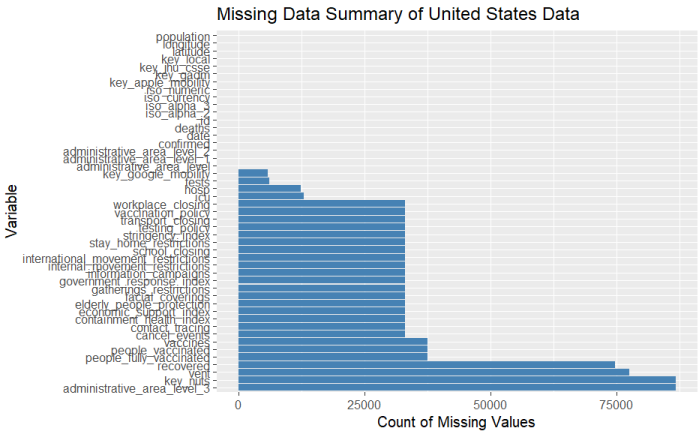


Figure 4: Missing Data Summary of United States Data

2.4.1 Missing Value Processing

First, I work with the cumulative value variables. I start by grouping the data by state and sorting by date, followed by two steps of missing value filling:

- If the first value of the variable in the group is missing, it is filled with zeros.
- The remaining missing values are filled using Last Observation Carried Forward (LOCF), which means that the last known value preceding the missing value is used to fill the missing value.

Next, I deal with missing values for policy-related variables. I first grouped the data by state and sorted it by date, then took the following steps to fill in the missing values:

- If the first value of a variable in the group is missing, it is filled with a zero, meaning "No measures".
- For missing values in each group, the nearest known value in the sequence is used to fill in the missing value, which means that I first used the nearest previous non-missing value to cover the current missing value and then used the nearest next non-missing value to cover the current missing value.
- The remaining missing values are filled in using the mode of the group.

2.4.2 Outliers detection and fixing

For cumulative value variables in the data, such as confirmed, deaths, etc., I have found that there are many outliers where the cumulative prior value is higher than the subsequent value. Normally, cumulative values should increase monotonically, so they need to be corrected. I took the following steps:

1. Group the data by state and sort by date.

2. If the current value is less than the value of the previous day, adjust the current value to the value of the previous day.

3. If the current value is larger than the value of the day after, adjust the current value to the value of the day after.

4. Iterate over each value to ensure that the maximum value is taken up to the current date and there is no decrementing.

2.4.3 Categorical Variables Processing

In the original data, numbers have been used instead of specific labels for policy-related variables such as school_closing and workplace_closing. However, these variables use positive integers to indicate policies that apply to the entire administrative region, and negative integers to identify policies that represent the best guess of an effective policy, which may not represent the given region's true situation. The negative sign is only used to distinguish between the two cases and is not a true negative value. For the uniformity of the policy data, I need to absolute the negative integers as the current policy situation in the region.

2.4.4 Feature Selection and New Features

Finally, I chose to keep the epidemiological variables, policy measure variables, and ID, date, and area information related variables for subsequent analyses, including:

- Epidemiological variables: confirmed, deaths, covered, tests, vaccines, people_vaccinated, people_fully_vaccinated

- Policy measure variables: school_closing, workplace_closing, cancel_events, gatherings_restrictions, transport_closing, stay_home_restrictions, internal_movement_restrictions, international_movement_restrictions, information_campaigns, testing_policy, contact_tracing, facial_coverings, vaccination_policy, elderly_people_protection

- Others: id, date, administrative_area_level_1, administrative_area_level_2, population

In addition, I added some new epidemiological variables for subsequent analyses:

- new_cases: number of new confirmed cases per day.

- cases_per_million: number of new confirmed cases per million people per day.

- new_vaccines: number of new vaccine doses per day.

- new_deaths: new deaths per day.
- deaths_per_million: new deaths per million people per day.

3 Exploratory Data Analysis

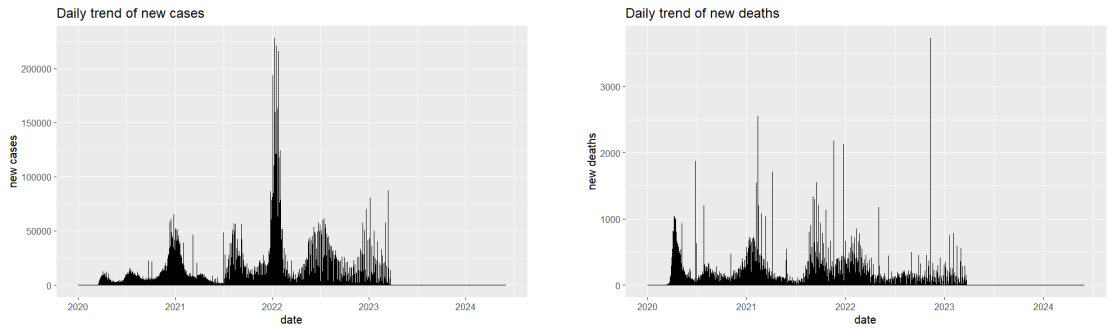
3.1 Inference Task Definition

In this project, I defined two inference tasks, including:

1. The development of the COVID-19 epidemic in the U.S.: explore the development of the COVID-19 epidemic in the U.S. by investigating the changes in confirmed and death cases.
2. The effect of policy on the development of the epidemic: investigate whether the isolation policies can control the spread of the virus by studying the relationship between policy and new confirmed cases.

3.2 The development of the COVID-19 epidemic in the U.S.

To analyze the development of the COVID-19 epidemics in the U.S., I plotted the daily changes in the number of new confirmed and new deaths as shown in Figure 5:



(a) Daily trend of new cases in the United States (b) Daily trend of new deaths in the United States

Figure 5: Daily Trend of New Confirmed and Deaths in the United States

It can be seen that there were not many new confirmed in the early stages of the epidemic, but there were large-scale infections in late 2021 to early 2022 and in mid-2022. In contrast, there were more new deaths in the early stages of the epidemic and the change in new deaths afterward was largely consistent with the change in new confirmed. This is probably because a large number of patients died in the early stages of the COVID-19 epidemic due to a lack of medical resources. Subsequent mass infections

were caused by several major variant strains of COVID-19, including the Delta variant and Omicron, which were more infectious and led to a significant increase in confirmed cases[2].

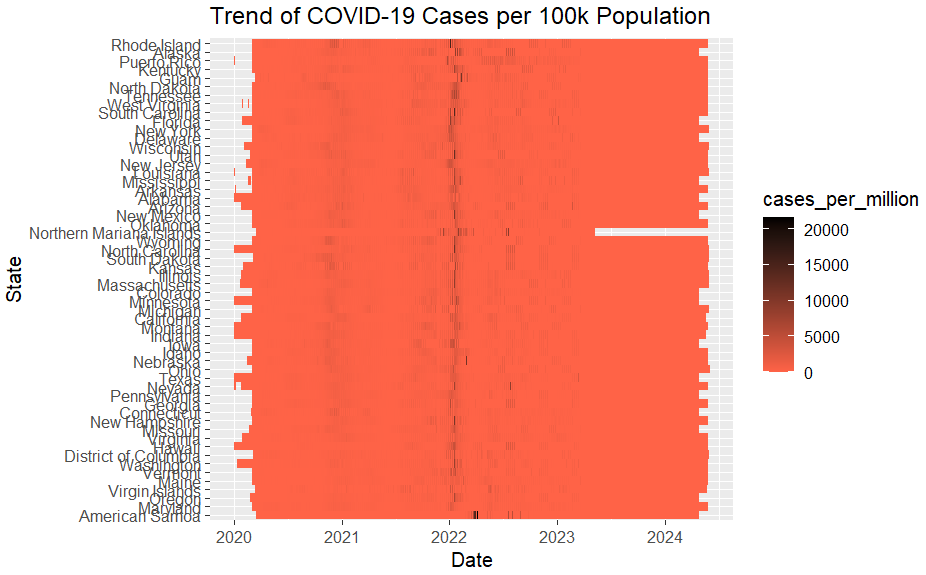


Figure 6: Trend of COVID-19 Cases per 100k Population

Next, I analyzed the variation in new confirmed across states. Considering that the total population of each state is different, I used new confirmed per million people on a single day as an observational variable to see how it changes with date. As shown in Figure 6, it can be seen that the change in new confirmed per million people per single day tends to be consistent across states, but varies slightly over the 2020 period.
In order to better observe the variation, I restricted the data to the 2020 period and plotted Figure 7. It can be seen that before October 2020, new confirmed per million people per day were not consistent across states, but all increased after October.

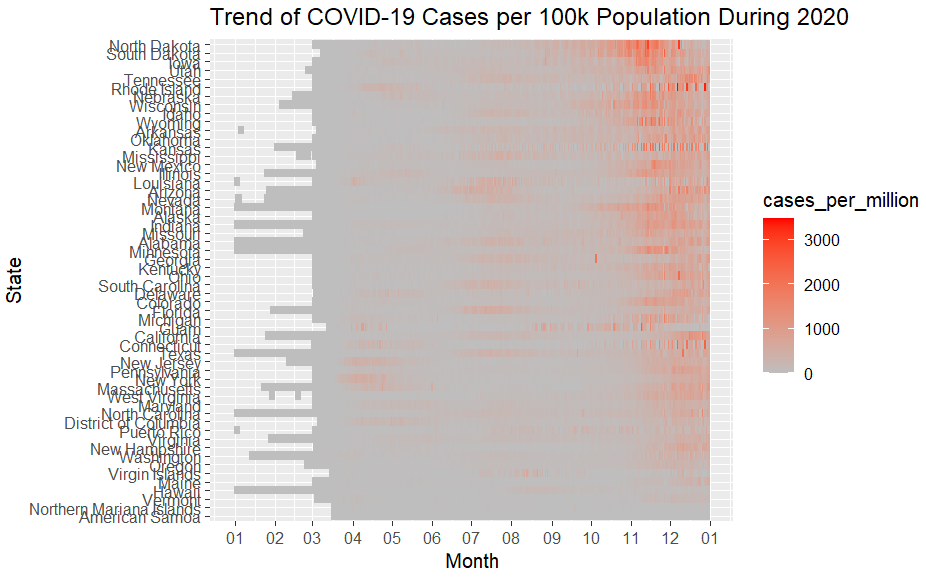


Figure 7: Trend of COVID-19 Cases per 100k Population During 2020

Based on Figure 7, I hypothesize that the difference in the change in new confirmed per million people between states may be due to different policies. According to the survey, at the beginning of the COVID-19 epidemic, states implemented different policies to control the further spread of the COVID-19 virus[3]. Next, I will explore whether different policies had an impact on the control of the epidemic during 2020.

3.3 The effect of policy on the epidemic

First of all, I conduct a regression analysis using new confirmed per million people as the dependent variable and the other policy measure variables as independent variables. The results of the regression analysis I obtained are shown in Figure 8.

```
Call:
lm(formula = cases_per_million ~ vaccination_policy + school_closing +
    workplace_closing + cancel_events + gatherings_restrictions +
    transport_closing + stay_home_restrictions + internal_movement_restrictions +
    international_movement_restrictions + information_campaigns +
    testing_policy + contact_tracing + facial_coverings + elderly_people_protection,
    data = df_usa_2020)

Residuals:
   Min     1Q Median     3Q    Max
-706.1 -121.0  -51.1   48.6 3198.7

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         58.877      5.500  10.704  < 2e-16 ***
vaccination_policy                 142.577      5.092  28.000  < 2e-16 ***
school_closing                     -79.042      3.210 -24.623  < 2e-16 ***
workplace_closing                  -46.554      3.937 -11.825  < 2e-16 ***
cancel_events                      -15.965      4.683  -3.409 0.000654 ***
gatherings_restrictions             14.067      1.779   7.909 2.75e-15 ***
transport_closing                  -17.810      3.118  -5.713 1.13e-08 ***
stay_home_restrictions              53.101      4.679  11.349  < 2e-16 ***
internal_movement_restrictions     -13.099      4.557  -2.874 0.004053 **
international_movement_restrictions -25.197      4.393  -5.736 9.88e-09 ***
information_campaigns              124.249      7.869  15.789  < 2e-16 ***
testing_policy                      54.276      3.183  17.052  < 2e-16 ***
contact_tracing                    -70.496      4.077 -17.293  < 2e-16 ***
facial_coverings                    56.140      2.222  25.266  < 2e-16 ***
elderly_people_protection           17.982      2.822   6.372 1.91e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 233.8 on 17914 degrees of freedom
Multiple R-squared:  0.2738,    Adjusted R-squared:  0.2733
F-statistic: 482.6 on 14 and 17914 DF,  p-value: < 2.2e-16
```

Figure 8: Regression Analysis Result

The $p-value < 0.001$ can be seen for most independent variables, indicating that these variables play a significant role in explaining the variation in cases_per_million.

Among them, the coefficients of vaccination_policy and information_campaigns are 142.577 and 124.249 respectively, which have higher estimated values, indicating that they may have a large positive impact on the number of cases.

This is because vaccination policies are usually introduced more aggressively when the epidemic is advancing. At the same time, the promotion of vaccination is often accompanied by increased detection capacity, which reveals more hidden cases, thus making the number of new cases appear to rise. Besides, the effects of vaccination policies usually take time to show up. Therefore, when vaccination policies are first implemented, the number of new cases may continue to remain high because the vaccine has not yet fully taken effect. Also, information campaigns increase awareness of testing and case reporting, which leads to an increase in the proportion of confirmed cases. In addition, when the outbreak is very severe, the government conducts information campaigns more intensively.

The coefficients on school_closing and contact_tracing are -79.042 and -70.496 respectively, with negative estimates, suggesting that these policies may be associated with a decrease in cases.

This may be because closing schools reduces the movement of people and people contact, stopping the spread of the virus. And contact tracing helps potentially infected people to be isolated as quickly as possible, again stopping the spread of the virus.

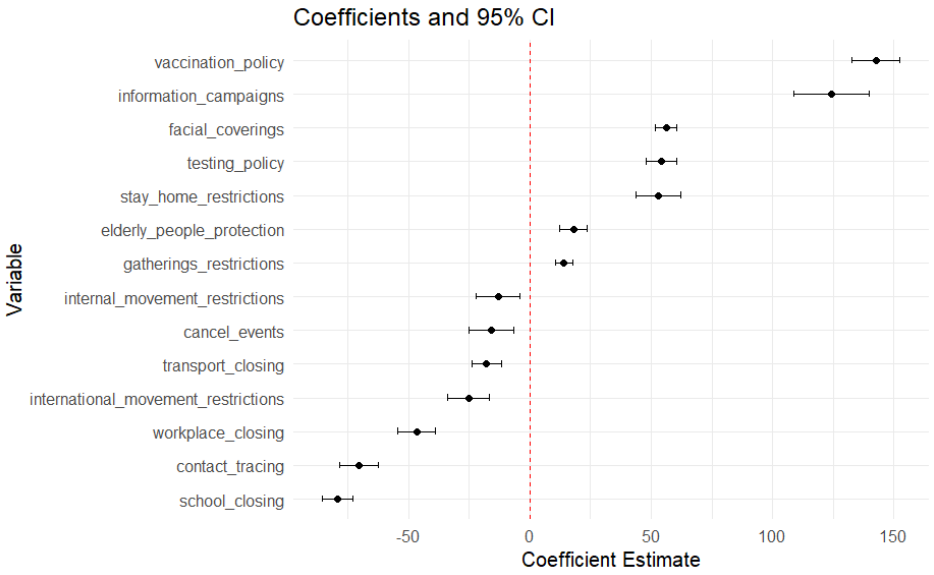
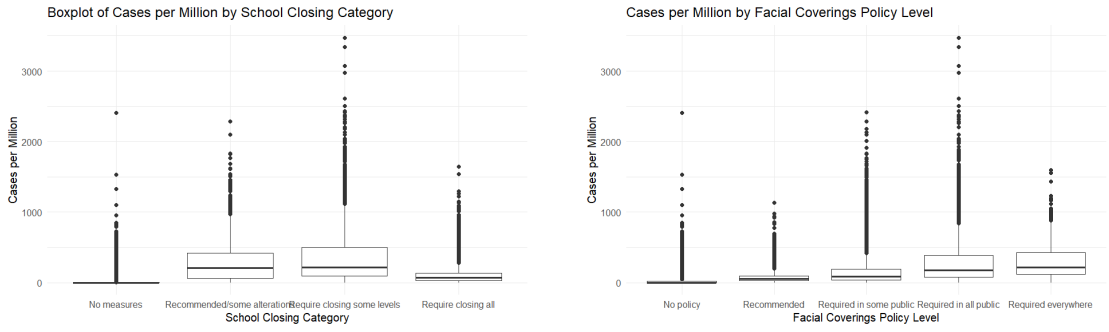


Figure 9: Coefficients and 95% CI

Then, I plotted the coefficient estimates and their confidence intervals for each variable, showing the regression coefficients and their 95% confidence intervals for each independent variable. It can be seen that the lower and upper confidence intervals for each coefficient estimate have the same sign and neither crosses 0. This means that at the 95% confidence level, the effect of each of these variables on the new confirmed per million people is statistically significant.

Because the policy measure variables are all categorical, I use box plots to analyze the effect of school closures and mask-wearing policies on new confirmed per million people. It can be seen that better control of virus spread is achieved when schools are completely closed. In contrast, the stricter the policy of wearing masks, the more new confirmed there are. Probably because masks are strictly required by the government when the outbreak is very severe.



(a) Cases per Million by School Closing

(b) Cases per Million by Facial Coverings

Figure 10: Boxplot of of Cases per Million by Policy

4 Conclusion

Combining the results of the above analysis, we can conclude that during the COVID-19 epidemic, the United States did not have a large number of confirmed cases in the pre-epidemic period, but the lack of medical resources still resulted in a high number of deaths. And the subsequent variants of the COVID-19 virus caused very significant harm. During 2020, the government took many policy measures to control the spread of the virus, and while most of these measures had an impact on new cases, not all of them had a positive impact.

References

[1] Emanuele Guidotti and David Ardia. "COVID-19 Data Hub". In: *Journal of Open Source Software* 5.51 (2020), p. 2376. DOI: 10.21105/joss.02376.

[2] Wikipedia contributors. *COVID-19 pandemic in the United States*. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_the_United_States. Accessed: 2024-12-06. 2024.

[3] American Legislative Exchange Council. *COVID-19 Executive Orders Tracker for the 50 States*. Accessed: 2024-12-06. 2020. URL: https://alec.org/article/covid-19-executive-orders-tracker-for-the-50-states/.

A Code Listings

A.1 Data Preprocessing

```
install.packages("COVID19", repos = "https://cloud.r-project.org/")
library("COVID19")
df <- covid19(level = 2)
summary(df)
head(df)
library(dplyr)
library(ggplot2)
library(tidyr)
missing_summary <- df %>%
  summarise_all(¬sum(is.na(.))) %>%
  pivot_longer(cols = everything(), names_to = "Variable", ...
     values_to = "MissingCount")

ggplot(missing_summary, aes(x = reorder(Variable, -MissingCount), y ...
    = MissingCount)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Missing Data Summary", x = "Variable", y = "Count ...
     of Missing Values")
date_counts <- df %>%
  group_by(date) %>%
  summarise(Count = n(), .groups = "drop")

ggplot(date_counts, aes(x = date, y = Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Number of Instances Per Date",
       x = "Date",
       y = "Count") +
  theme_minimal()
library(tidyr)
library(zoo)

df_processed <- df %>%
  arrange(administrative_area_level_2, date) %>%
  group_by(administrative_area_level_2) %>%
  mutate(
    deaths = ifelse(is.na(deaths) & date == min(date), 0, deaths),
    confirmed = ifelse(is.na(confirmed) & date == min(date), 0, ...
       confirmed)
  ) %>%
  mutate(
```

```
    deaths = zoo::na.locf(deaths, na.rm = FALSE),
    confirmed = zoo::na.locf(confirmed, na.rm = FALSE)
  ) %>%
  ungroup()
df_now<- df_processed %>%
  filter(date == "2022-12-31")
country_deaths <- df_now %>%
  group_by(administrative_area_level_1) %>%
  summarise(TotalDeaths = sum(deaths, na.rm = TRUE), .groups = "drop")

country_deaths <- country_deaths %>%
  arrange(desc(TotalDeaths))

ggplot(country_deaths, aes(x = reorder(administrative_area_level_1, ...
    TotalDeaths), y = TotalDeaths)) +
  geom_bar(stat = "identity", fill = "tomato") +
  coord_flip() +
  labs(title = "Deaths by Country on Dec 31 2022",
       x = "Country",
       y = "Total Deaths") +
  theme_minimal()
country_comfirmed <- df_now %>%
  group_by(administrative_area_level_1) %>%
  summarise(Totalconfirmed = sum(confirmed, na.rm = TRUE), .groups ...
      = "drop")

country_comfirmed <- country_comfirmed %>%
  arrange(desc(Totalconfirmed))

ggplot(country_comfirmed , aes(x = ...
    reorder(administrative_area_level_1, Totalconfirmed ), y = ...
    Totalconfirmed )) +
  geom_bar(stat = "identity", fill = "tomato") +
  coord_flip() +
  labs(title = "Confirmed by Country on Dec 31 2022",
       x = "Country",
       y = "Total confirmed") +
  theme_minimal()
df_usa<- df_processed %>%
  filter(administrative_area_level_1 == "United States")
head(df_usa)
missing_summary_usa <- df_usa %>%
  summarise_all(¬sum(is.na(.))) %>%
  pivot_longer(cols = everything(), names_to = "Variable", ...
      values_to = "MissingCount")

```

```
ggplot(missing_summary_usa, aes(x = reorder(Variable, ...
    -MissingCount), y = MissingCount)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Missing Data Summary of United States Data", x = ...
      "Variable", y = "Count of Missing Values")
# Missing value processing
cumulative_vars <- c("recovered", "tests", "vaccines", ...
    "people_vaccinated", "people_fully_vaccinated")

df_usa_clean <- df_usa %>%
  arrange(administrative_area_level_2, date) %>%
  group_by(administrative_area_level_2) %>%
  mutate(across(all_of(cumulative_vars), ¬ifelse(is.na(.) & ...
      row_number() == 1, 0, .))) %>%
  mutate(across(all_of(cumulative_vars), ¬zoo::na.locf(., na.rm = ...
      FALSE))) %>%
  ungroup()

policy_vars <- c(
  "school_closing", "workplace_closing", "cancel_events", ...
      "gatherings_restrictions", "transport_closing", ...
      "stay_home_restrictions", "internal_movement_restrictions",
  "international_movement_restrictions", "information_campaigns", ...
      "testing_policy",
  "contact_tracing", "facial_coverings", "vaccination_policy", ...
      "elderly_people_protection"
)
get_mode <- function(x) {
  if (length(x) == 0 || all(is.na(x))) return(NA)
  uniq_x <- unique(na.omit(x))
  uniq_x[which.max(tabulate(match(x, uniq_x)))]
}

safe_na_locf <- function(x) {
  if (all(is.na(x))) return(x)
  na.locf(na.locf(x, na.rm = FALSE), fromLast = TRUE)
}

df_usa_clean <- df_usa_clean %>%
  group_by(administrative_area_level_2) %>%
  mutate(across(all_of(policy_vars), ¬ifelse(is.na(.) & ...
      row_number() == 1, 0, .))) %>%
  mutate(across(all_of(policy_vars), ¬safe_na_locf(.))) %>%
  mutate(across(all_of(policy_vars), ¬ifelse(is.na(.), ...
      get_mode(.[!is.na(.)]), .))) %>%
  ungroup()
```

```r
cumulative_features<- c("confirmed","deaths","recovered", "tests", ...
    "vaccines", "people_vaccinated", "people_fully_vaccinated")
df_usa_clean_pro <- df_usa_clean %>%
  arrange(administrative_area_level_2, date) %>%
  group_by(administrative_area_level_2) %>%
  mutate(across(all_of(cumulative_features), ¬pmax(., lag(., ...
      default = 0)))) %>%
  mutate(across(all_of(cumulative_features), ¬pmin(., lead(., ...
      default = tail(., 1))))) %>%
  mutate(across(all_of(cumulative_features), ¬cummax(.))) %>%
  ungroup()
df_usa_clean_pro %>%
  arrange(administrative_area_level_2, date) %>%
  group_by(administrative_area_level_2) %>%
  summarise(across(all_of(cumulative_features), ¬sum(. < lag(.), ...
      na.rm = TRUE)))
df_usa_clean_pro <- df_usa_clean_pro %>%
  mutate(across(all_of(policy_vars), ¬abs(.)))
df_usa_processed <- df_usa_clean_pro%>%
  select(id, date,deaths, ...
      confirmed,all_of(cumulative_vars),all_of(policy_vars),
administrative_area_level_1,administrative_area_level_2,population)

df_usa_processed <- df_usa_processed %>%
  arrange(date) %>%
  group_by(id) %>%
  mutate(new_cases = confirmed - lag(confirmed, default = 0))
df_usa_processed <- df_usa_processed %>%
  mutate(cases_per_million = (new_cases / population) * 1e6)
df_usa_processed <- df_usa_processed %>%
  arrange(date) %>%
  group_by(id) %>%
  mutate(new_vaccines = vaccines - lag(vaccines, default = 0))
df_usa_processed <- df_usa_processed %>%
  arrange(date) %>%
  group_by(id) %>%
  mutate(new_deaths = deaths - lag(deaths, default = 0))
df_usa_processed <- df_usa_processed %>%
  mutate(deaths_per_million = (new_deaths / population) * 1e6)

head(df_usa_processed)
summary(df_usa_processed)
```

A.2 Exploratory Data Analysis

```
ggplot(df_usa_processed, aes(x = date, y = new_cases)) +
  geom_line() +
  labs(title = "Daily trend of new cases", x = "date", y = "new cases")
ggplot(df_usa_processed, aes(x = date, y = new_deaths)) +
  geom_line() +
  labs(title = "Daily trend of new deaths", x = "date", y = "new ...
     deaths")
df_usa_processed %>%
  group_by(administrative_area_level_2, date) %>%
  summarise(cases_per_million = sum(cases_per_million, na.rm = ...
     TRUE), .groups = "drop") %>%
  ggplot(aes(x = date, y = reorder(administrative_area_level_2, ...
     cases_per_million), fill = cases_per_million)) +
  geom_tile() +
  scale_fill_gradient(low = "tomato", high = "black") +
  labs(title = "Trend of COVID-19 Cases per 100k Population", x = ...
     "Date", y = "State")
df_usa_2020 <- df_usa_processed %>%
  filter(format(date, "%Y") == "2020")
head(df_usa_2020)
df_usa_2020 %>%
  group_by(administrative_area_level_2, date) %>%
  summarise(cases_per_million = sum(cases_per_million, na.rm = ...
     TRUE), .groups = "drop") %>%
  ggplot(aes(x = date, y = reorder(administrative_area_level_2, ...
     cases_per_million), fill = cases_per_million)) +
  geom_tile() +
  scale_fill_gradient(low = "grey", high = "red") +
  scale_x_date(date_labels = "%m", date_breaks = "1 month") +
  labs(title = "Trend of COVID-19 Cases per 100k Population During ...
     2020", x = "Month", y = "State")
library(dplyr)
library(ggplot2)

model <- lm(cases_per_million ¬ vaccination_policy + school_closing ...
    + workplace_closing +
              cancel_events + gatherings_restrictions + ...
                 transport_closing +
              stay_home_restrictions + ...
                 internal_movement_restrictions +
              international_movement_restrictions + ...
                 information_campaigns +
              testing_policy + contact_tracing + facial_coverings + ...
                 elderly_people_protection,
            data = df_usa_2020)

```

```r
summary(model)
confint(model, level=0.95)
library(broom)
library(dplyr)
model_coef <- tidy(model, conf.int=TRUE) %>%
  filter(term != "(Intercept)")

ggplot(model_coef, aes(x = estimate, y = reorder(term, estimate))) +
  geom_point() +
  geom_errorbarh(aes(xmin=conf.low, xmax=conf.high), height=0.2) +
  geom_vline(xintercept=0, linetype="dashed", color="red") +
  labs(x = "Coefficient Estimate", y = "Variable",
       title = "Coefficients and 95% CI") +
  theme_minimal()
library(ggplot2)

df_usa_2020$school_closing <- factor(df_usa_2020$school_closing,
                                     levels = c(0,1,2,3),
                                     labels = c("No measures",
                                                "Recommended/some ...
                                                   alterations",
                                                "Require closing ...
                                                   some levels",
                                                "Require closing all"))

ggplot(df_usa_2020, aes(x = school_closing, y = cases_per_million)) +
  geom_boxplot() +
  labs(x = "School Closing Category", y = "Cases per Million",
       title = "Boxplot of Cases per Million by School Closing ...
          Category") +
  theme_minimal()
df_usa_2020$facial_coverings <- factor(df_usa_2020$facial_coverings,
                                       levels = c(0, 1, 2, 3, 4),
                                       labels = c("No policy",
                                                  "Recommended",
                                                  "Required in some ...
                                                     public",
                                                  "Required in all ...
                                                     public",
                                                  "Required ...
                                                     everywhere"))
ggplot(df_usa_2020, aes(x =  facial_coverings, y = ...
    cases_per_million)) +
  geom_boxplot() +
  labs(x = "Facial Coverings Policy Level", y = "Cases per Million",
       title = "Cases per Million by Facial Coverings Policy ...
          Level") + theme_minimal()
```