Reproducible Report on COVID19 Data

01/29/2023

Introduction

Novel Coronavirus (COVID-19) is a global pandemic that has affected countries of all sizes around the world. The Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) has provided a dataset of confirmed cases and fatalities linked to the virus which are available on their Github page. This analysis looks at the dataset provided by JHU CSSE to provide a better understanding of the severity of the outbreak, and how it has played out over time across different countries and regions.

Questions of Interest

1. How has the number of confirmed cases and fatalities changed over time?
2. How has the number of confirmed cases and fatalities changed over time in the United States?
3. How has the number of confirmed cases and fatalities changed over time in the United States compared to other countries?
4. What is the relationship between the number of confirmed cases and the number of fatalities?
5. Which countries have the most confirmed cases and fatalities?
6. Can we forecast the number of confirmed cases and fatalities in the future?

Used Packages

Note about additional packages used

The following packages are used within this document and they need to be installed first:

- prophet

```
library(tidyverse)
library(prophet)
```

Loading Data

```
library(tidyverse)

root <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid

files <- c("time_series_covid19_confirmed_global.csv", "time_series_covid19_deaths_global.csv")

items <- lapply(files, function(x) read_csv(paste0(root, x)))
```

```
## Rows: 289 Columns: 1112
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1110): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Rows: 289 Columns: 1112
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1110): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
confirmed <- items[[1]]
deaths <- items[[2]]
```

Tidying and Transforming Data

Let's tidy up the data a bit by melting the data into a long format, and then joining the confirmed and death data together.

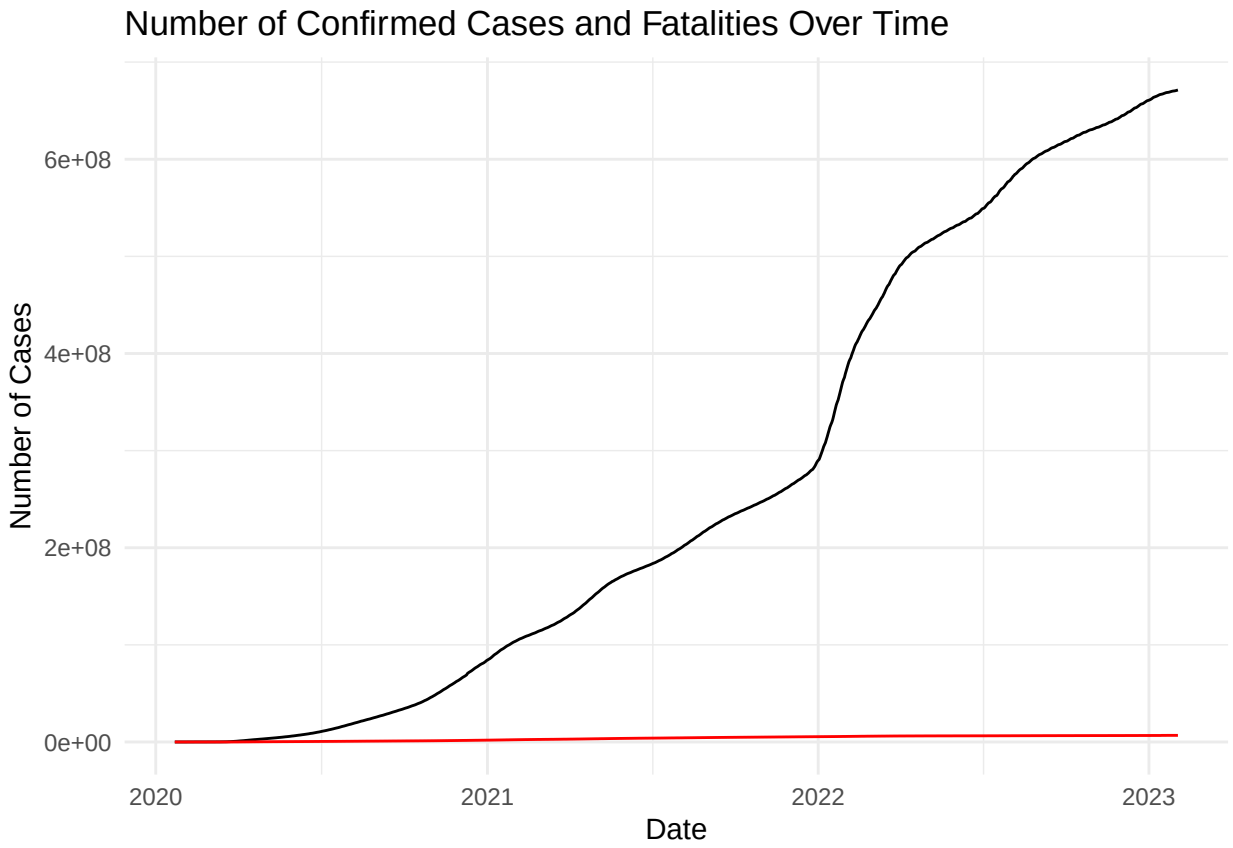
```
confirmed <- confirmed %>%
  select(-Lat, -Long) %>%
  pivot_longer(-c("Province/State", "Country/Region"), names_to = "date", values_to = "confirmed") %>%
  mutate(date = as.Date(date, format = "%m/%d/%y"))

deaths <- deaths %>%
  select(-Lat, -Long) %>%
  pivot_longer(-c("Province/State", "Country/Region"), names_to = "date", values_to = "deaths") %>%
  mutate(date = as.Date(date, format = "%m/%d/%y"))

df <- left_join(confirmed, deaths, by = c("Province/State", "Country/Region", "date"))

glimpse(df)
```

```
## Rows: 320,212
## Columns: 5
## $ `Province/State` <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ `Country/Region` <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghani~
## $ date             <date> 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-25, 2020~
## $ confirmed        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ deaths           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

Exploratory Data Analysis

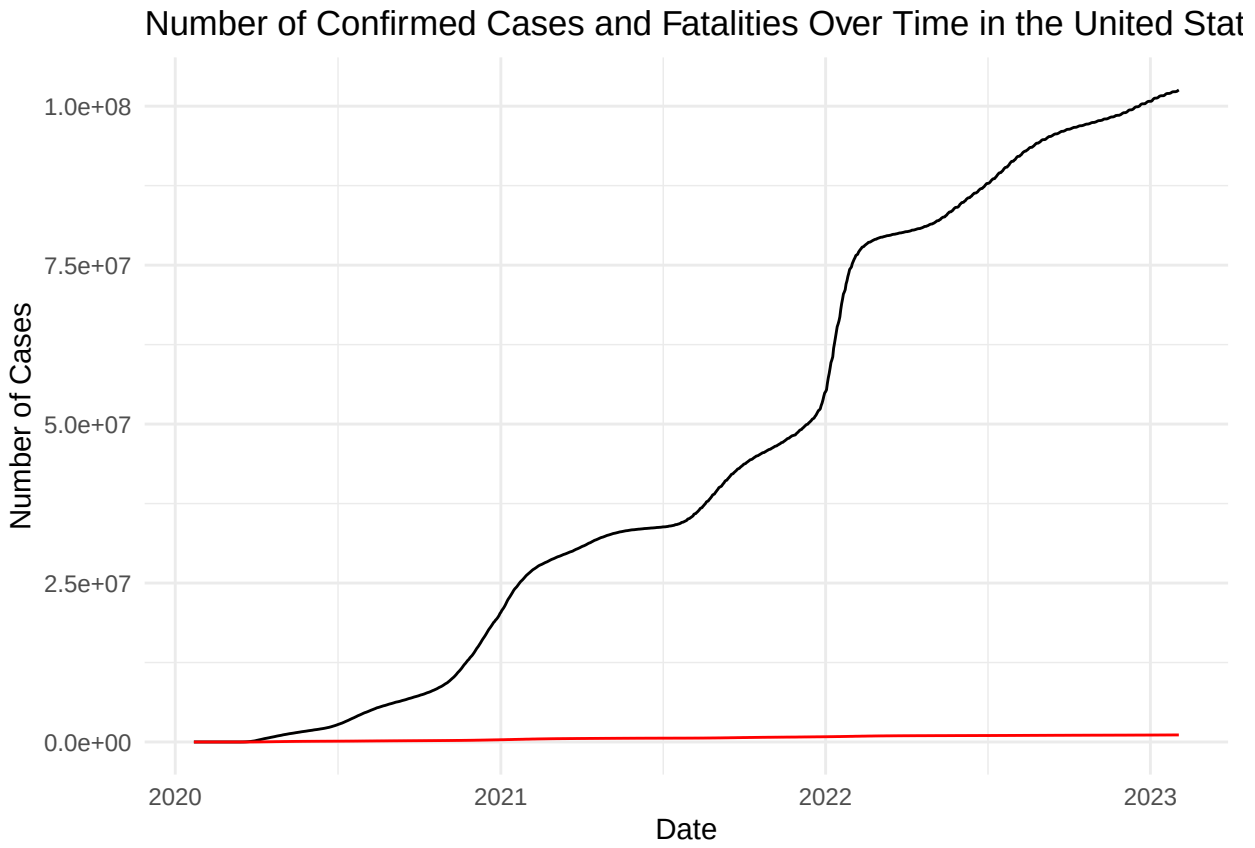
How has the number of confirmed cases and fatalities changed over time?

```
df %>%
  group_by(date) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths)) %>%
  ggplot(aes(x = date, y = confirmed)) +
  geom_line() +
  geom_line(aes(y = deaths), color = "red") +
  labs(title = "Number of Confirmed Cases and Fatalities Over Time",
       x = "Date",
       y = "Number of Cases") +
  theme_minimal()
```



How has the number of confirmed cases and fatalities changed over time in the United States?

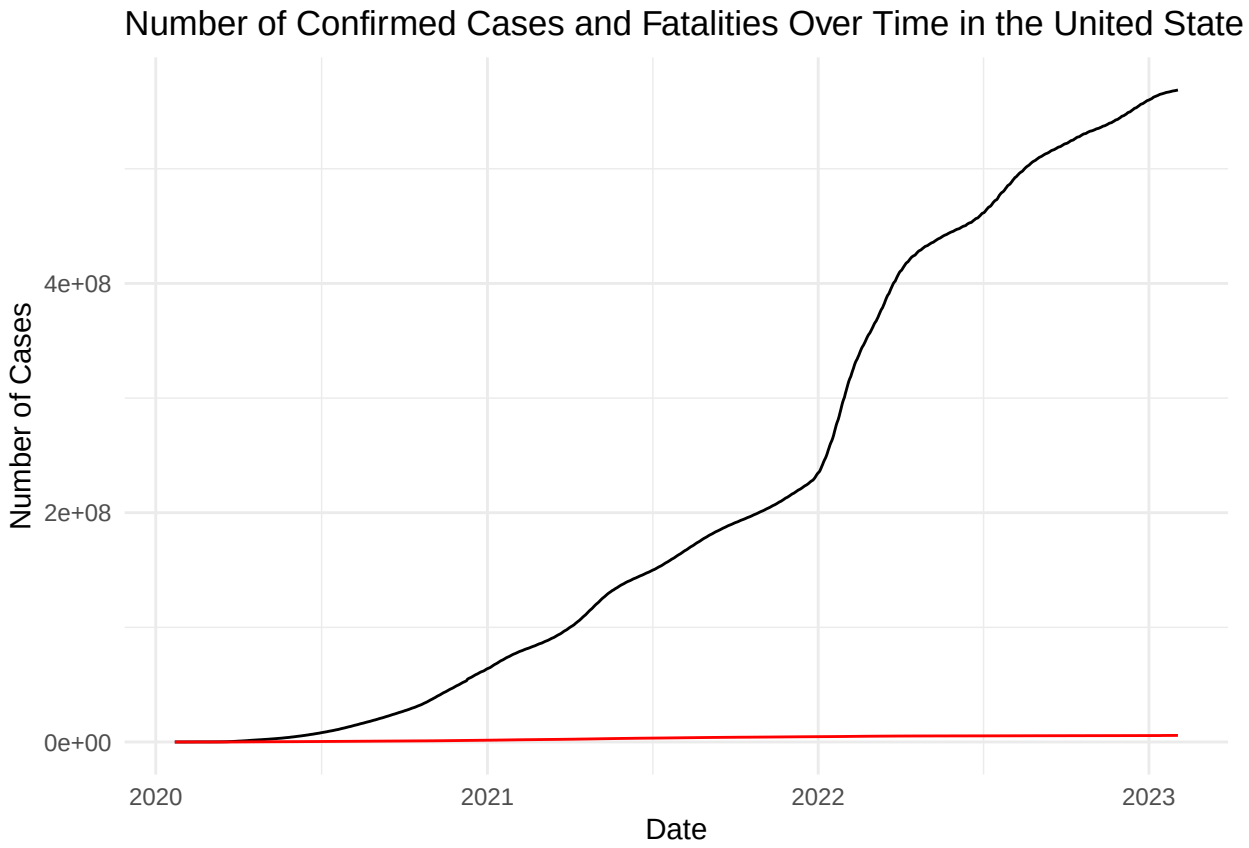
```
df %>%
  filter(`Country/Region` == "US") %>%
  group_by(date) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths)) %>%
  ggplot(aes(x = date, y = confirmed)) +
  geom_line() +
  geom_line(aes(y = deaths), color = "red") +
  labs(title = "Number of Confirmed Cases and Fatalities Over Time in the United States",
       x = "Date",
       y = "Number of Cases") +
  theme_minimal()
```



How has the number of confirmed cases and fatalities changed over time in the United States compared to other countries?

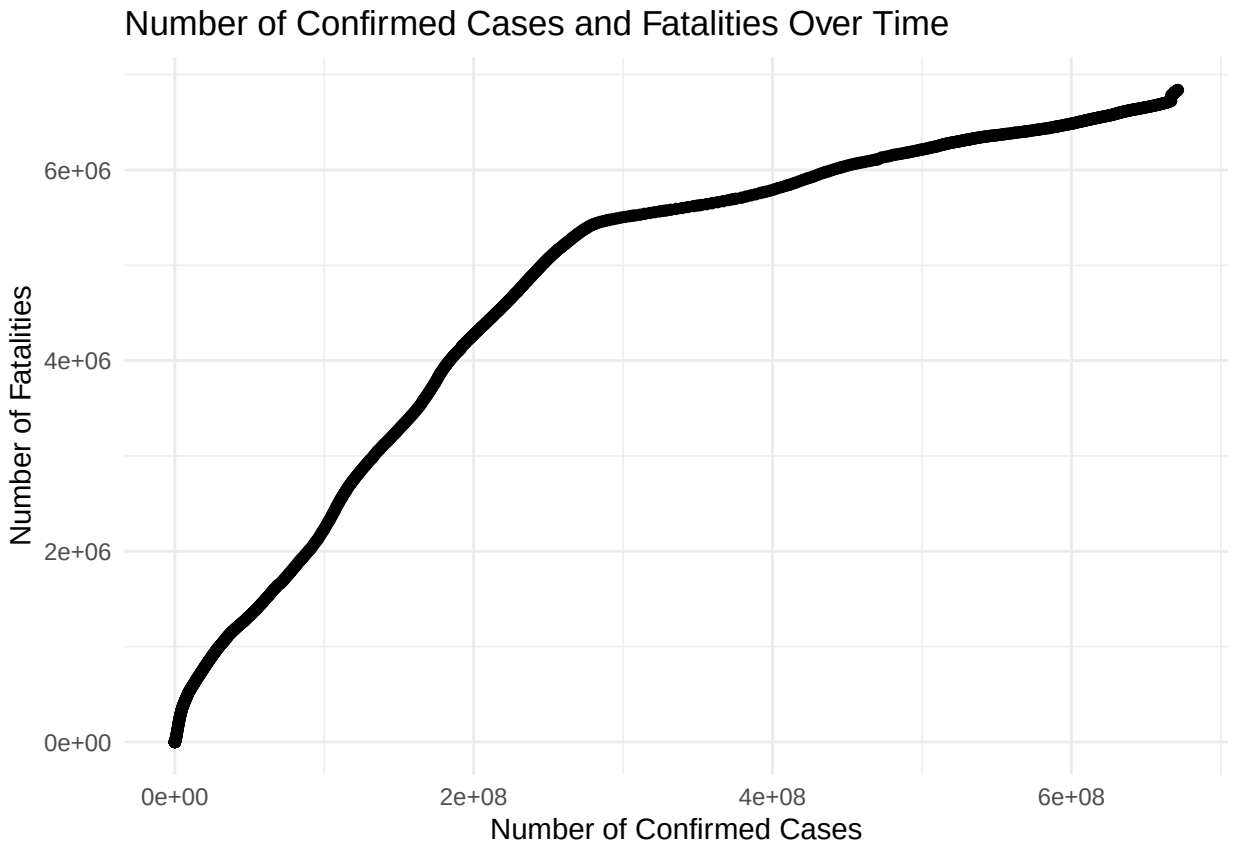
```
df %>%
  filter(`Country/Region` != "US") %>%
  group_by(date, `Country/Region`) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths)) %>%
  group_by(date) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths)) %>%
  ggplot(aes(x = date, y = confirmed)) +
  geom_line() +
  geom_line(aes(y = deaths), color = "red") +
  labs(title = "Number of Confirmed Cases and Fatalities Over Time in the United States Compared to Oth
       x = "Date",
       y = "Number of Cases") +
  theme_minimal()
```

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



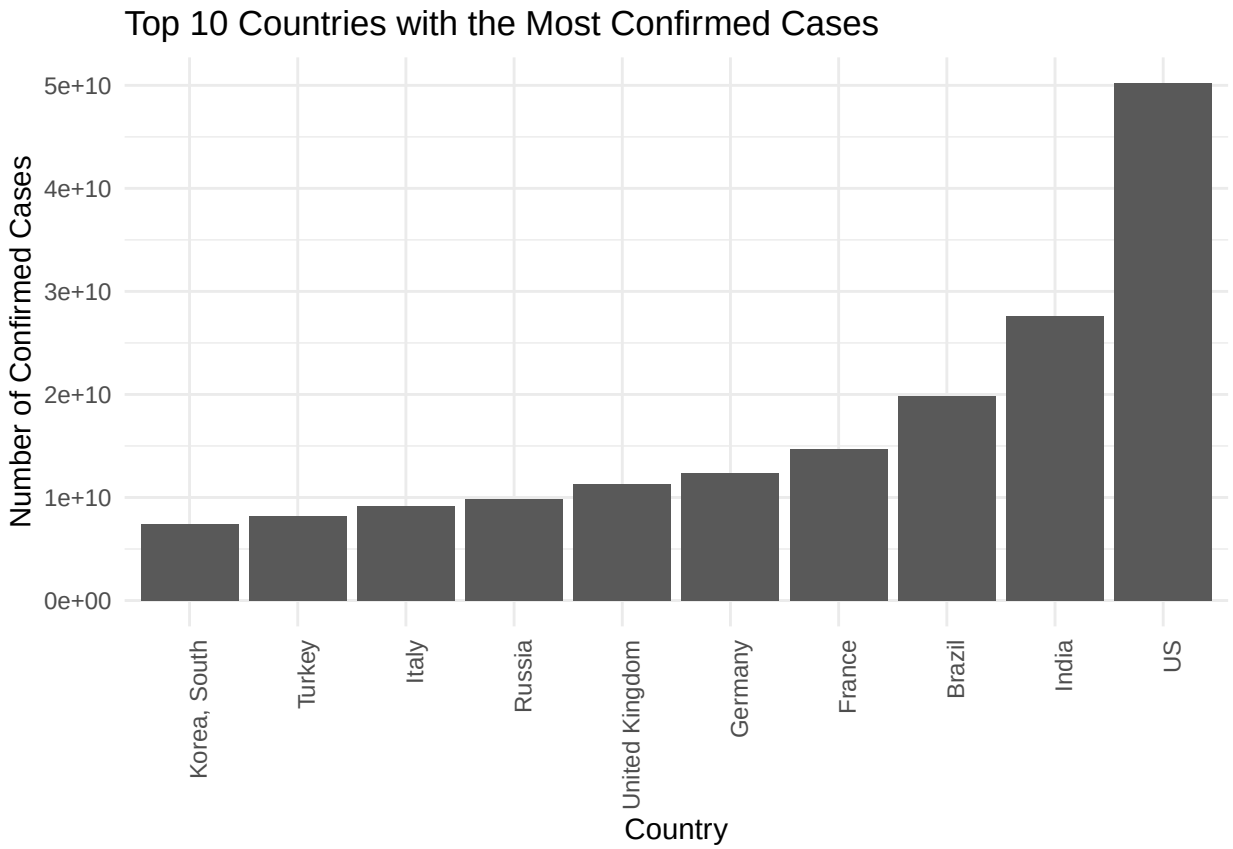
What is the relationship between the number of confirmed cases and the number of fatalities?

```
df %>%
  group_by(date) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths)) %>%
  ggplot(aes(x = confirmed, y = deaths)) +
  geom_point() +
  labs(title = "Number of Confirmed Cases and Fatalities Over Time",
       x = "Number of Confirmed Cases",
       y = "Number of Fatalities") +
  theme_minimal()
```

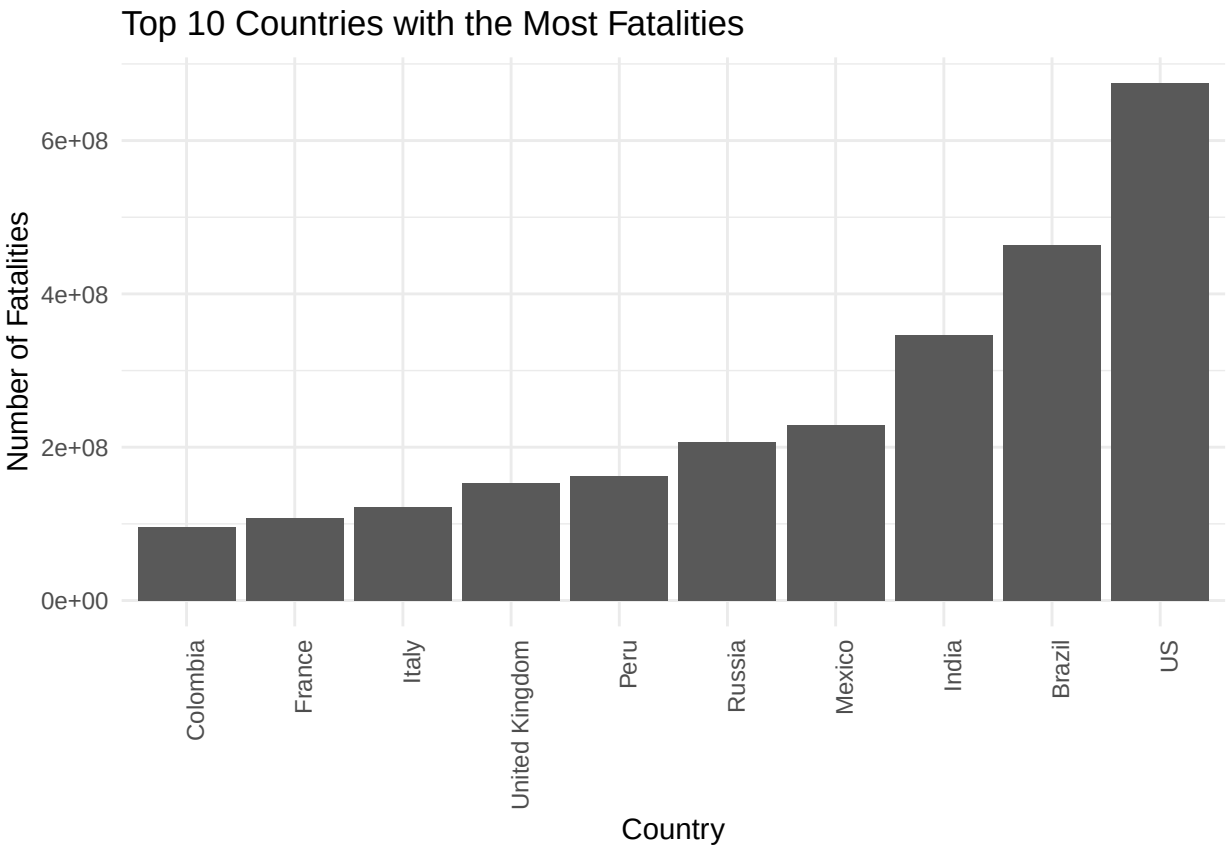


Where are the most confirmed cases and fatalities?

```
df %>%
  group_by(`Country/Region`) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths)) %>%
  arrange(desc(confirmed)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(`Country/Region`, confirmed), y = confirmed)) +
  geom_col() +
  labs(title = "Top 10 Countries with the Most Confirmed Cases",
       x = "Country",
       y = "Number of Confirmed Cases") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```




```
df %>%
  group_by(`Country/Region`) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths)) %>%
  arrange(desc(deaths)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(`Country/Region`, deaths), y = deaths)) +
  geom_col() +
  labs(title = "Top 10 Countries with the Most Fatalities",
       x = "Country",
       y = "Number of Fatalities") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Modeling: Time series forecasting

In this section, we will use the `forecast` package to forecast the number of confirmed cases and fatalities in the United States.

Forecasting the number of confirmed cases and fatalities in the United States

We will be using the `prophet` package to forecast the number of confirmed cases and fatalities in the United States. Also, we show the uncertainty interval of the forecast.

```
# Filter data for the US
df_us <- df %>%
  filter(`Country/Region` == "US") %>%
  group_by(date) %>%
  summarise(confirmed = sum(confirmed), deaths = sum(deaths))

# Create time series data for confirmed cases
df_confirmed <- df_us %>%
  select(date, confirmed) %>%
  rename(ds = date, y = confirmed) %>%
  mutate(ds = as.POSIXct(ds))

# Create time series data for fatalities
df_deaths <- df_us %>%
  select(date, deaths) %>%
  rename(ds = date, y = deaths) %>%
  mutate(ds = as.POSIXct(ds))
```

Let's fit the model to the data.

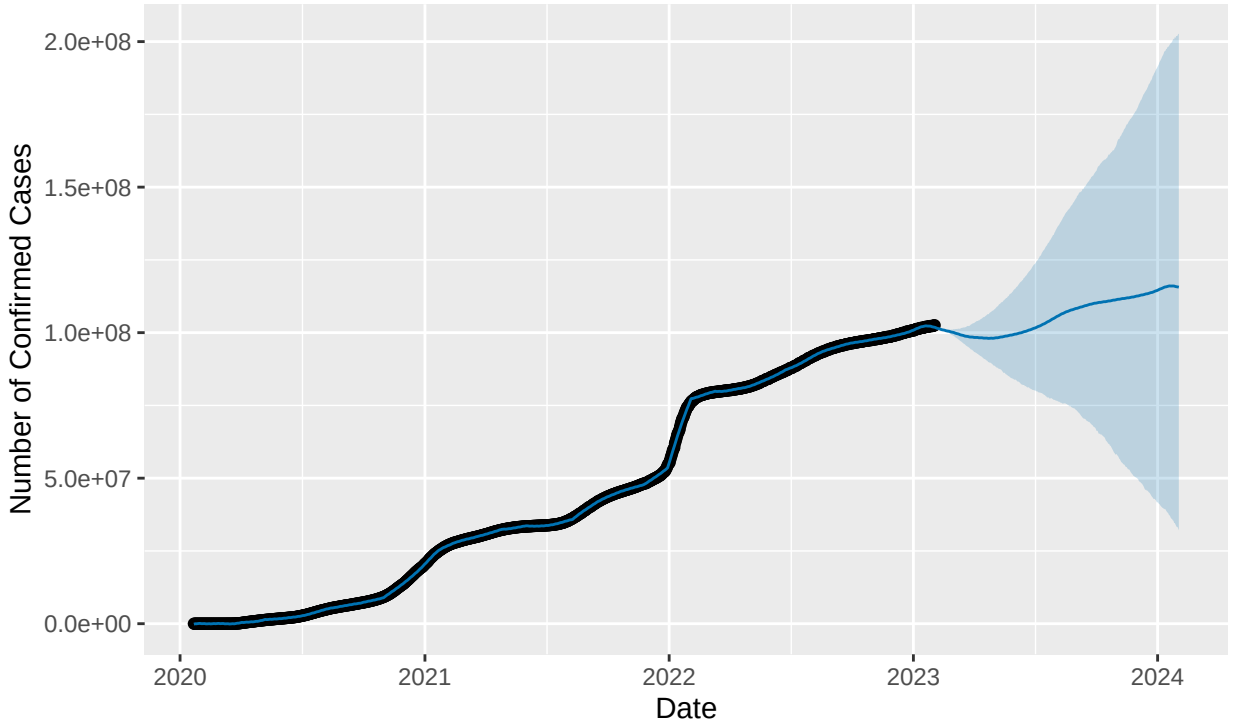
```
# Fit prophet models for confirmed cases and fatalities
prophet_confirmed <- prophet(df_confirmed, yearly.seasonality = TRUE, weekly.seasonality = TRUE, daily.
prophet_deaths <- prophet(df_deaths, yearly.seasonality = TRUE, weekly.seasonality = TRUE, daily.seasona
```

Let's forecast the number of confirmed cases and fatalities in the United States.

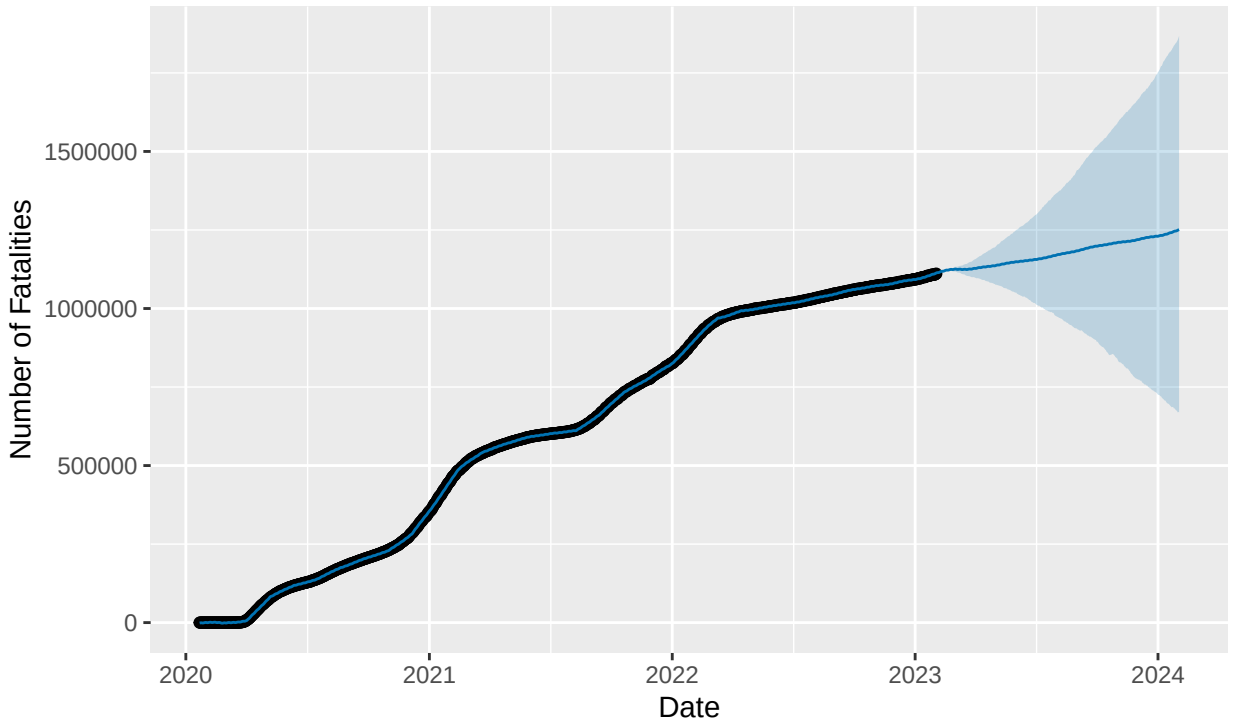
```
# Create forecast dataframes
future_confirmed <- make_future_dataframe(prophet_confirmed, periods = 365)
future_deaths <- make_future_dataframe(prophet_deaths, periods = 365)

# Generate forecasts
forecast_confirmed <- predict(prophet_confirmed, future_confirmed)
forecast_deaths <- predict(prophet_deaths, future_deaths)
```

Let's plot the forecast.

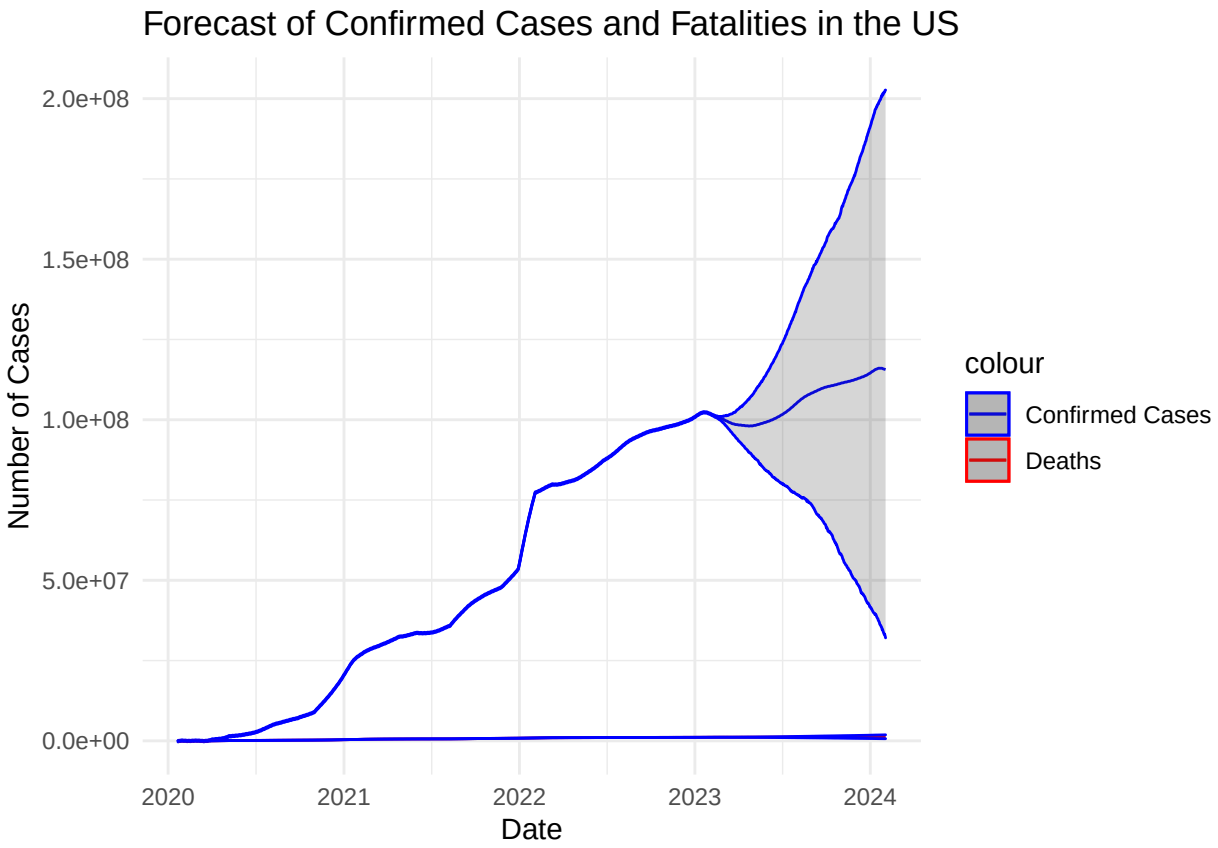
```
# Plotting the forecast for confirmed cases
plot(prophet_confirmed, forecast_confirmed, xlabel = "Date", ylabel = "Number of Confirmed Cases", main
```




```
# Plotting the forecast for fatalities
plot(prophet_deaths, forecast_deaths, xlabel = "Date", ylabel = "Number of Fatalities", main = "Forecas
```




```
# Combine the two forecasts into a single plot
ggplot(forecast_confirmed, aes(ds, yhat, color = "Confirmed Cases")) +
  geom_line() +
  geom_ribbon(aes(ymin = yhat_lower, ymax = yhat_upper), alpha = 0.2) +
  geom_line(data = forecast_deaths, aes(ds, yhat, color = "Deaths")) +
  geom_ribbon(data = forecast_deaths, aes(ymin = yhat_lower, ymax = yhat_upper), alpha = 0.2) +
  labs(title = "Forecast of Confirmed Cases and Fatalities in the US", x = "Date", y = "Number of Cases
  scale_color_manual(values = c("Confirmed Cases" = "blue", "Deaths" = "red")) +
  theme_minimal()
```



We can see that the number of confirmed cases and fatalities will continue to increase in the United States. The model's uncertainty interval is also increasing, which means that the model is less confident in its predictions as time goes on.

Conclusion

In this post, we have explored the COVID-19 dataset and performed some exploratory data analysis. We have also used the `forecast` package to forecast the number of confirmed cases and fatalities in the United States.

Possible source of bias

The data used in this post is from the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The data is updated daily, and the data for the United States is updated at 23:59 UTC. This means that the data for the United States is updated at 4:59 PM EST. This means that the data for the United States is not updated until the next day. This could be a source of bias in the data.