Analysis of COVID-19 cases and deaths data

Detailed analysis on worst affected countries in the world and states in United States.

Rajesh Rao

Contents

Description of Datasets

1. The primary COVID-19 data is collected and maintained by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE)
2. Four main time series based datasets used for the analysis
    - Confirmed US cases
    - Confirmed Global cases
    - Confirmed US deaths
    - Confirmed Global deaths
3. Confirmed US cases: Provides a detailed time series of COVID-19 cases for each county in the United States since the start of pandemic. One record represents one county and historical number of confirmed COVID-19 cases in that county.
4. Confirmed Global cases: Provides a detailed time series of COVID-19 cases for each country in the world since the start of pandemic. One record represents one country and historical number of confirmed COVID-19 cases in that country
5. Confirmed US deaths: Provides a detailed time series of COVID-19 related death for each county in the United States since the start of pandemic. One record represents one county and historical number of confirmed COVID-19 deaths in that county.
6. Confirmed Global cases: Provides a detailed time series of COVID-19 deaths for each country in the world since the start of pandemic. One record represents one country and historical number of confirmed COVID-19 deaths in that country
7. US State Population: Provides the population for states in USA by the US census bureau. One record represents a state and it's population history on yearly basis since 2010
8. Global Population: Provides population estimate for countries around the world. Compiled and maintained by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). One record represents the population of a state or province or country in the world and related data.

Objective of Analysis

1. Determine the worst states with number of cases per 100,000 population in United States
2. Analyze which months has most number of new cases in United States
3. Analyze top 15 countries with worst death per 100,000 population
4. Build a analytical model to determine the relationship between new cases and new deaths in India

Pre-processing and preparation

Setting up of R options

```
knitr::opts_chunk$set(echo = TRUE)
```

Initializing Session Information and Loading R packages

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22000)
##
## 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
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.31   lifecycle_1.0.3 magrittr_2.0.3  evaluate_0.19
##  [5] rlang_1.0.6     stringi_1.7.8   cli_3.5.0       rstudioapi_0.14
##  [9] vctrs_0.5.1     rmarkdown_2.19  tools_4.2.2     stringr_1.5.0
## [13] glue_1.6.2      xfun_0.36       yaml_2.3.6      fastmap_1.1.0
## [17] compiler_4.2.2  htmltools_0.5.4 knitr_1.41
```

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

Data Import

Importing historical data from respective sources. Read CSV commands from the tidyverse are used to read the data from web URL link. Description of these datasets are provided earlier.

```
url_base <- c('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse
us_cases <- read_csv(paste(url_base,'time_series_covid19_confirmed_US.csv',sep = ""))
```

```
## Rows: 3342 Columns: 1138
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1132): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20...
##
## 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.
```

```
us_deaths <- read_csv(paste(url_base,'time_series_covid19_deaths_US.csv',sep = ""))
```

```
## Rows: 3342 Columns: 1139
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1133): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## 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.
```

```
global_cases <- read_csv(paste(url_base,'time_series_covid19_confirmed_global.csv',sep = ""))
```

```
## Rows: 289 Columns: 1131
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1129): 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.
```

```
global_deaths <- read_csv(paste(url_base,'time_series_covid19_deaths_global.csv',sep = ""))
```

```
## Rows: 289 Columns: 1131
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1129): 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.
```

```
us_pop <- read.csv("https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/state/totals/nst
global_population <- read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_co
```

```
## Rows: 4321 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population
##
## 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.
```

Data Exploration and Data Tidying

Exploratory Analysis

Investigating sample data to determine the data tidying process

```
head(us_cases, n=5)
```

```
## # A tibble: 5 x 1,138
##        UID iso2  iso3  code3  FIPS Admin2  Provinc~1 Count~2   Lat Long_ Combi~3
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>     <chr>   <dbl> <dbl> <chr>
## 1 84001001 US    USA     840  1001 Autauga Alabama   US       32.5 -86.6 Autaug~
## 2 84001003 US    USA     840  1003 Baldwin Alabama   US       30.7 -87.7 Baldwi~
## 3 84001005 US    USA     840  1005 Barbour Alabama   US       31.9 -85.4 Barbou~
## 4 84001007 US    USA     840  1007 Bibb    Alabama   US       33.0 -87.1 Bibb, ~
## 5 84001009 US    USA     840  1009 Blount  Alabama   US       34.0 -86.6 Blount~
```

```
## # ... with 1,127 more variables: '1/22/20' <dbl>, '1/23/20' <dbl>,
## #   '1/24/20' <dbl>, '1/25/20' <dbl>, '1/26/20' <dbl>, '1/27/20' <dbl>,
## #   '1/28/20' <dbl>, '1/29/20' <dbl>, '1/30/20' <dbl>, '1/31/20' <dbl>,
## #   '2/1/20' <dbl>, '2/2/20' <dbl>, '2/3/20' <dbl>, '2/4/20' <dbl>,
## #   '2/5/20' <dbl>, '2/6/20' <dbl>, '2/7/20' <dbl>, '2/8/20' <dbl>,
## #   '2/9/20' <dbl>, '2/10/20' <dbl>, '2/11/20' <dbl>, '2/12/20' <dbl>,
## #   '2/13/20' <dbl>, '2/14/20' <dbl>, '2/15/20' <dbl>, '2/16/20' <dbl>, ...
```

```
head(global_cases, n=5)
```

```
## # A tibble: 5 x 1,131
##   Province~1 Count~2   Lat  Long 1/22/~3 1/23/~4 1/24/~5 1/25/~6 1/26/~7 1/27/~8
##   <chr>      <chr>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 <NA>       Afghan~  33.9 67.7        0       0       0       0       0       0
## 2 <NA>       Albania  41.2 20.2        0       0       0       0       0       0
## 3 <NA>       Algeria  28.0  1.66       0       0       0       0       0       0
## 4 <NA>       Andorra  42.5  1.52       0       0       0       0       0       0
## 5 <NA>       Angola  -11.2 17.9        0       0       0       0       0       0
## # ... with 1,121 more variables: '1/28/20' <dbl>, '1/29/20' <dbl>,
## #   '1/30/20' <dbl>, '1/31/20' <dbl>, '2/1/20' <dbl>, '2/2/20' <dbl>,
## #   '2/3/20' <dbl>, '2/4/20' <dbl>, '2/5/20' <dbl>, '2/6/20' <dbl>,
## #   '2/7/20' <dbl>, '2/8/20' <dbl>, '2/9/20' <dbl>, '2/10/20' <dbl>,
## #   '2/11/20' <dbl>, '2/12/20' <dbl>, '2/13/20' <dbl>, '2/14/20' <dbl>,
## #   '2/15/20' <dbl>, '2/16/20' <dbl>, '2/17/20' <dbl>, '2/18/20' <dbl>,
## #   '2/19/20' <dbl>, '2/20/20' <dbl>, '2/21/20' <dbl>, '2/22/20' <dbl>, ...
```

```
head(us_deaths, n=5)
```

```
## # A tibble: 5 x 1,139
##         UID iso2  iso3  code3  FIPS Admin2  Provinc~1 Count~2   Lat Long_ Combi~3
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>     <chr>   <dbl> <dbl> <chr>
## 1 84001001 US    USA     840  1001 Autauga Alabama   US       32.5 -86.6 Autaug~
## 2 84001003 US    USA     840  1003 Baldwin Alabama   US       30.7 -87.7 Baldwi~
## 3 84001005 US    USA     840  1005 Barbour Alabama   US       31.9 -85.4 Barbou~
## 4 84001007 US    USA     840  1007 Bibb    Alabama   US       33.0 -87.1 Bibb, ~
## 5 84001009 US    USA     840  1009 Blount  Alabama   US       34.0 -86.6 Blount~
## # ... with 1,128 more variables: Population <dbl>, '1/22/20' <dbl>,
## #   '1/23/20' <dbl>, '1/24/20' <dbl>, '1/25/20' <dbl>, '1/26/20' <dbl>,
## #   '1/27/20' <dbl>, '1/28/20' <dbl>, '1/29/20' <dbl>, '1/30/20' <dbl>,
## #   '1/31/20' <dbl>, '2/1/20' <dbl>, '2/2/20' <dbl>, '2/3/20' <dbl>,
## #   '2/4/20' <dbl>, '2/5/20' <dbl>, '2/6/20' <dbl>, '2/7/20' <dbl>,
## #   '2/8/20' <dbl>, '2/9/20' <dbl>, '2/10/20' <dbl>, '2/11/20' <dbl>,
## #   '2/12/20' <dbl>, '2/13/20' <dbl>, '2/14/20' <dbl>, '2/15/20' <dbl>, ...
```

```
head(global_deaths, n=5)
```

```
## # A tibble: 5 x 1,131
##   Province~1 Count~2   Lat  Long 1/22/~3 1/23/~4 1/24/~5 1/25/~6 1/26/~7 1/27/~8
##   <chr>      <chr>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 <NA>       Afghan~  33.9 67.7        0       0       0       0       0       0
## 2 <NA>       Albania  41.2 20.2        0       0       0       0       0       0
## 3 <NA>       Algeria  28.0  1.66       0       0       0       0       0       0
```

```
## 4 <NA>       Andorra  42.5  1.52       0       0       0       0       0       0
## 5 <NA>       Angola  -11.2 17.9        0       0       0       0       0       0
## # ... with 1,121 more variables: '1/28/20' <dbl>, '1/29/20' <dbl>,
## #   '1/30/20' <dbl>, '1/31/20' <dbl>, '2/1/20' <dbl>, '2/2/20' <dbl>,
## #   '2/3/20' <dbl>, '2/4/20' <dbl>, '2/5/20' <dbl>, '2/6/20' <dbl>,
## #   '2/7/20' <dbl>, '2/8/20' <dbl>, '2/9/20' <dbl>, '2/10/20' <dbl>,
## #   '2/11/20' <dbl>, '2/12/20' <dbl>, '2/13/20' <dbl>, '2/14/20' <dbl>,
## #   '2/15/20' <dbl>, '2/16/20' <dbl>, '2/17/20' <dbl>, '2/18/20' <dbl>,
## #   '2/19/20' <dbl>, '2/20/20' <dbl>, '2/21/20' <dbl>, '2/22/20' <dbl>, ...
```

Data Tidying

1. Since the cases and deaths data set are in time series format, we will be converting all the date columns into individual rows to meaningfully analyze and aggregate the data and confirm to tidy format
2. Additionally, date columns will be changed to Date from character format.
3. Columns not required for our analysis like UID, ISO etc. will be removed from the data set to focus on critical attributes

```
cases_long <- us_cases %>%
  pivot_longer(-c("UID", "iso2", "iso3", "code3", "FIPS", "Admin2", "Province_State", "Country_Region",
                  "Lat", "Long_", "Combined_Key"),
               names_to = "Date", values_to = "Confirmed") %>%
  mutate(Date = as.Date(Date, format ="%m/%d/%y")) %>%
  select (Admin2:Confirmed) %>% select(-c(Lat, Long_, Combined_Key))
deaths_long <- us_deaths %>%
  pivot_longer(-c("UID", "iso2", "iso3", "code3", "FIPS", "Admin2", "Province_State", "Country_Region",
                  "Lat", "Long_", "Combined_Key"),
               names_to = "Date", values_to = "Deaths") %>%
  mutate(Date = as.Date(Date, format ="%m/%d/%y")) %>%
  select (Admin2:Deaths) %>% select(-c(Lat, Long_, Combined_Key))
global_cases_long <- global_cases %>%
  pivot_longer(-c("Province/State", "Lat",
                  "Long","Country/Region"),
               names_to = "Date", values_to = "Confirmed") %>%   select("Country/Region", "Date","Confi
global_deaths_long <- global_deaths %>%
  pivot_longer(-c("Province/State", "Lat",
                  "Long","Country/Region"),
               names_to = "Date", values_to = "Deaths") %>%  select("Country/Region", "Date","Deaths")
```

Data Analysis and Transformations

Objective 1: Determining the states in USA with worst cases per 100,000 population.

1. To aid this analysis, both US cases and deaths will be combined to get a complete dataset
2. Combined cases and deaths datasets will be joined with US population dataset to get the population figures.
3. Using the dataset, data will be aggregated at state level and maximum number of cases will be determined
4. Cases per 100K population will be determined for each state.

```
us_covid <- cases_long %>%
  left_join(deaths_long, by = c("Admin2", "Province_State", "Country_Region","Date"))
us_pop_tidy <- us_pop %>%
  select(c(NAME, POPESTIMATE2020)) %>%
  rename(Province_State = NAME, Population = POPESTIMATE2020)
us_covid_pop <- us_covid %>%
  left_join(us_pop_tidy, by = c("Province_State"))
cases_long_max <- cases_long %>%
  group_by(Province_State) %>%
  summarise(max_cases = max(Confirmed))
us_covid_max <- cases_long_max %>%
  left_join(us_pop_tidy, by = c("Province_State"))
cases_long_max <- cases_long %>%
  group_by(Province_State) %>%
  summarise(max_cases = max(Confirmed))
us_covid_max_per100k <- us_covid_max %>%
  mutate(cases_per_100k = max_cases/Population * 100000) %>% arrange(desc(cases_per_100k)) %>%
  slice_max(cases_per_100k, n=25)
```

Objective 2: Analyze the months which had worst Covid-19 cases

1. US cases data set will be used as base and data will be aggregated based on month of the year

```
us_covid_pop_new_cases <- us_covid_pop  %>%
  mutate(new_cases = Confirmed - lag(Confirmed)) %>%
  mutate(Date = as.Date(Date,"%m/%d/%y"))
us_covid_pop_new_cases_worst <- us_covid_pop_new_cases %>%
  mutate(yearmon = floor_date(Date, unit = "month")) %>%
  group_by(yearmon) %>%
  summarise(sum = sum(new_cases, default=0)) %>%
  filter(sum > 0)
```

Objective 3: Determine the countries with worst deaths per 100K population

1. To aid this analysis, both global cases and deaths will be combined to get a complete dataset
2. Combined cases and deaths datasets will be joined with Global population dataset to get the population figures.
3. Using the dataset, data will be aggregated at country level and maximum number of deaths per 100K populatiom

```
global_covid <- global_cases_long %>% left_join(global_deaths_long, by = c("Country/Region","Date"))
global_population_tidy <- global_population %>%
  filter(is.na(Province_State)) %>%
  select("Country_Region","Population") %>%
  rename("Country/Region" = Country_Region)
global_covid_pop <- global_covid %>%
  left_join(global_population_tidy, by =c("Country/Region"))
global_covid_pop_maxdeaths <- global_covid_pop %>%
  rename(Country=  "Country/Region") %>%
  group_by(Country) %>%
  summarise(max_deaths = max(Deaths)) %>%
```

```
  rename(`Country/Region`= Country) %>%
  left_join(global_population_tidy, by = "Country/Region") %>%
  mutate(deaths_per_100k = max_deaths/Population * 100000) %>%
  slice_max(deaths_per_100k, n=25)
```

Objective 4: Determine the relationship between India's new death rate and new cases

1. Cases from global dataset will be filtered to contain India specific data
2. New cases and deaths for India will be determined on daily basis

```
india_cases <- global_covid_pop %>%
  filter(str_detect(`Country/Region`, 'India'))
india_new_cases_deaths <- india_cases %>%
  mutate(new_cases = Confirmed - lag(Confirmed)) %>%
  mutate(new_deaths = Deaths - lag(Deaths)) %>% drop_na(new_cases) %>% drop_na(new_deaths) %>%
  mutate(Date = as.Date(Date,"%m/%d/%y"))
india_model <- lm(new_deaths ~ new_cases, data= india_new_cases_deaths)
india_new_cases_deaths_pred <- india_new_cases_deaths %>% mutate(pred = predict(india_model))
```

Data Visualization

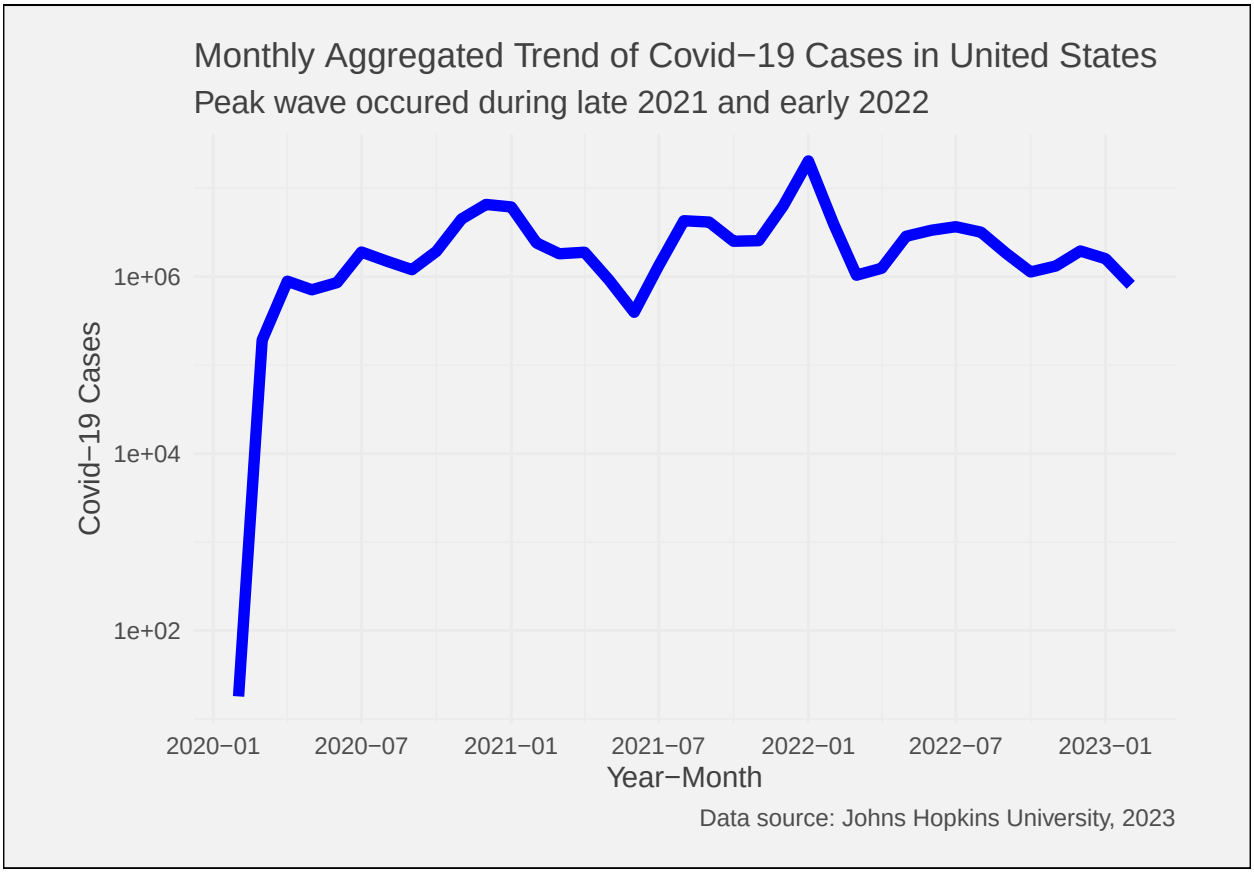
Setting the theme for visualization

```
theme_shooting <- function() {
    theme_minimal() +
theme(
    text = element_text(color = "gray25"),
    plot.subtitle = element_text(size = 12),
    plot.caption = element_text(color = "gray30"),
    plot.background = element_rect(fill = "gray95"),
    plot.margin = unit(c(5, 10, 5, 10), units = "mm")
  )
}
```

Visualization 1: Year-Month in USA with worst cases per 100,000 population

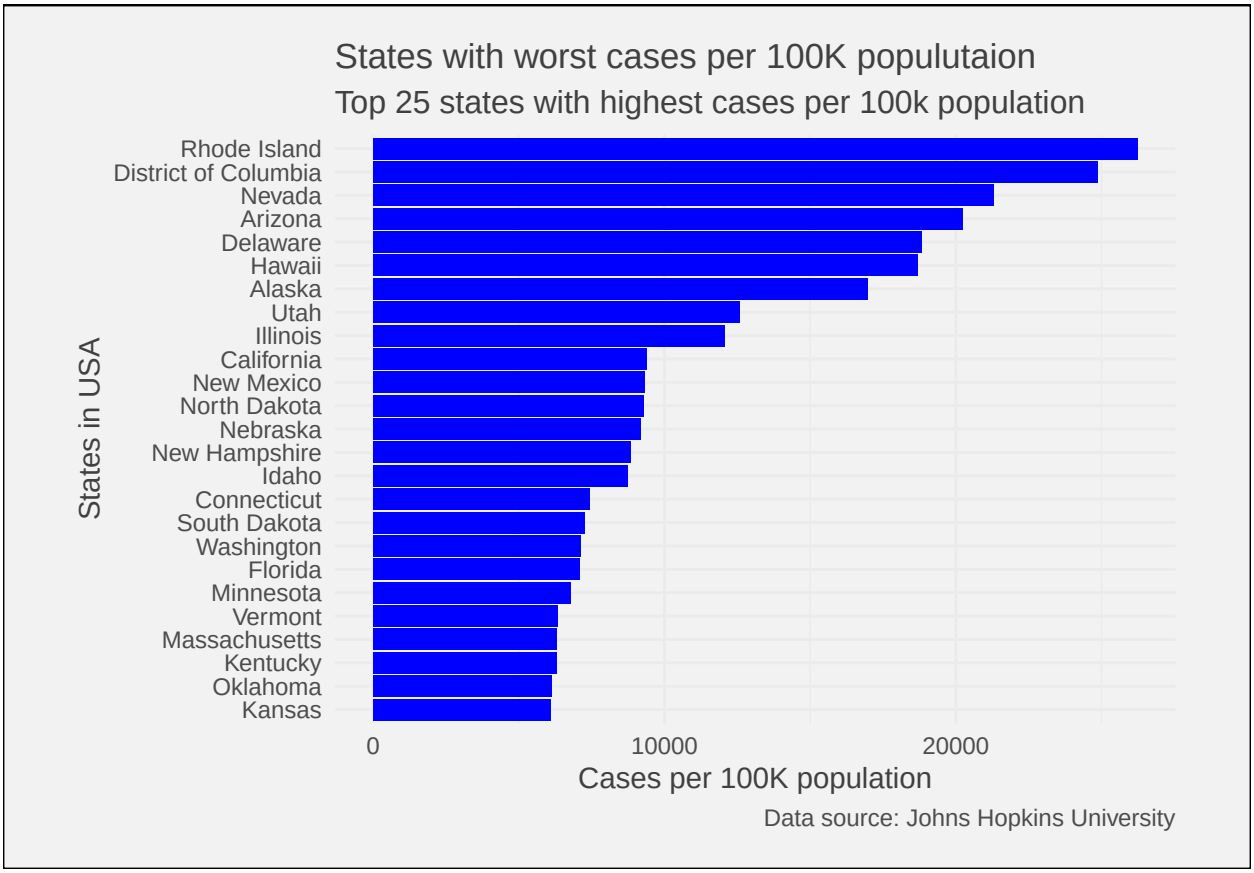
```
ggplot(data = us_covid_pop_new_cases_worst) +
  geom_line(aes(x = yearmon, y = sum), size = 2, color = "Blue") +
  scale_y_log10() +
  scale_x_date(date_labels = "%Y-%m",
               breaks = seq(from = as.Date("2020-01-01"),
                            to = as.Date("2023-02-01"), by = "6 months")) +
  labs(
    x = "Year-Month",
    y = "Covid-19 Cases",
    title = "Monthly Aggregated Trend of Covid-19 Cases in United States",
    subtitle = "Peak wave occured during late 2021 and early 2022",
    caption = "Data source: Johns Hopkins University, 2023"
  ) + theme_shooting()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```



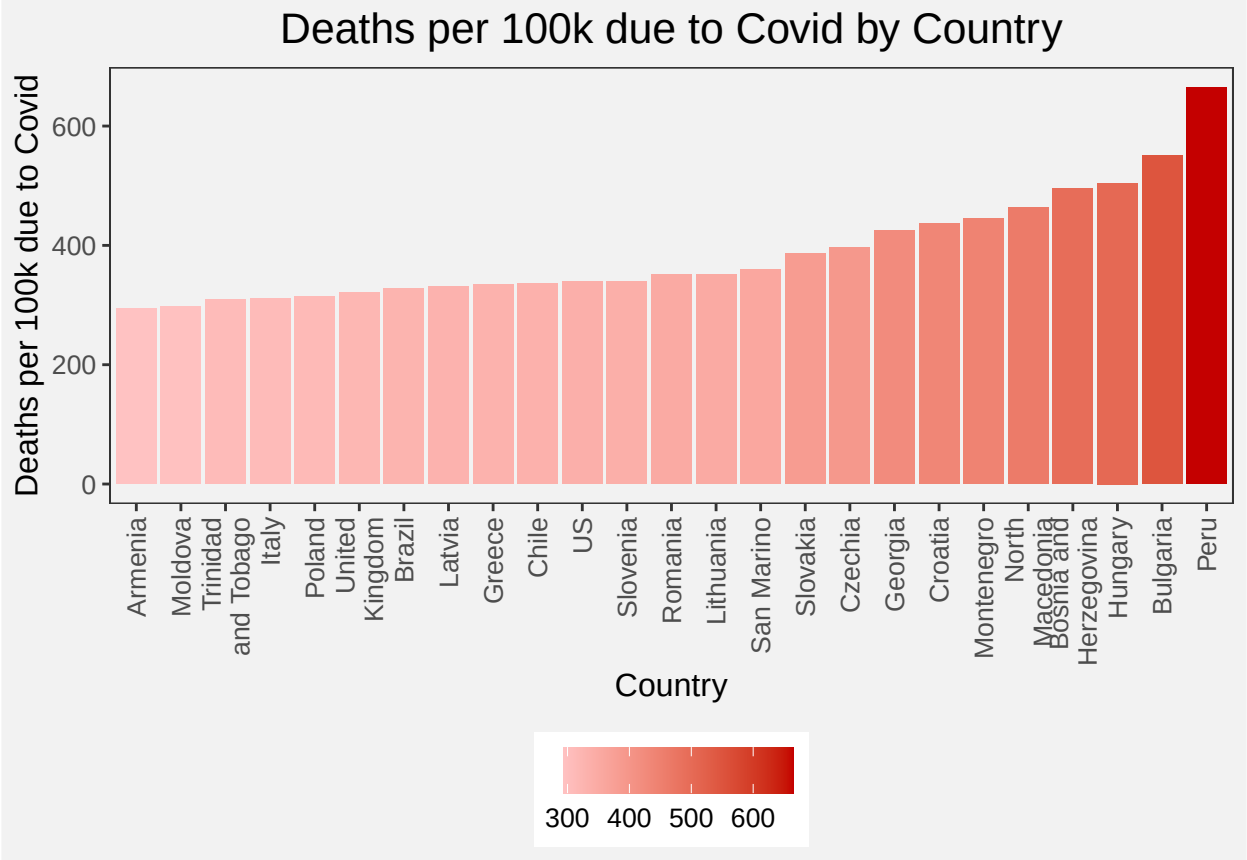
Visualization 2: States with worst cases per 100K population

```
ggplot(data = us_covid_max_per100k,
       aes(x = reorder(Province_State, cases_per_100k),
           y = cases_per_100k)) +
  geom_bar(stat = "identity", fill = "blue") + labs(
    x = "States in USA",
    y = "Cases per 100K population",
    title = "States with worst cases per 100K populutaion",
    subtitle = "Top 25 states with highest cases per 100k population",
    caption = "Data source: Johns Hopkins University"
  ) + coord_flip() + theme_shooting()
```



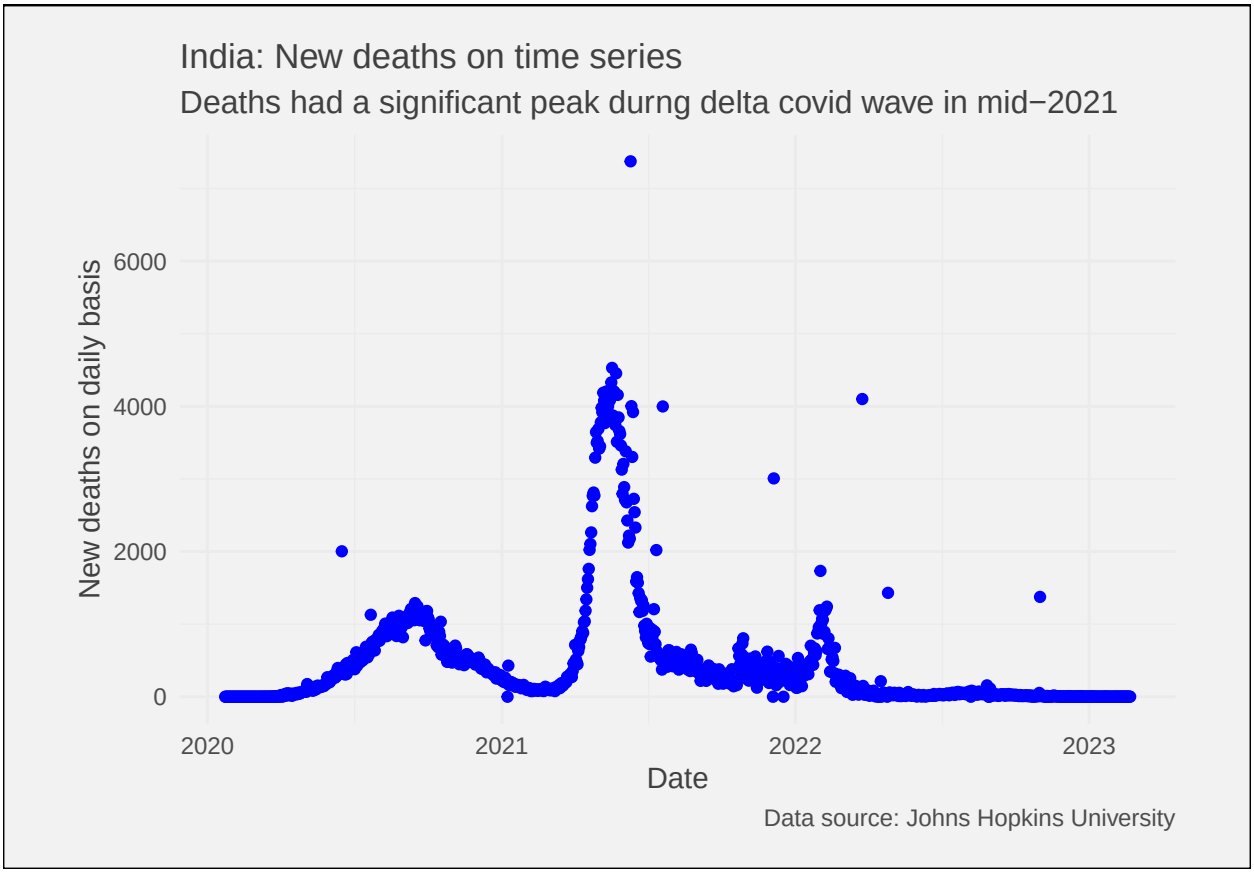
Visualization 3: Countries with worst s per 100K population

```
library(stringr)
ggplot(global_covid_pop_maxdeaths, aes(x = reorder(`Country/Region`, deaths_per_100k) ,
                                       y = deaths_per_100k, fill = deaths_per_100k)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Country", y = "Deaths per 100k due to Covid",
       title = "Deaths per 100k due to Covid by Country") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, size = 16),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        axis.text.x = element_text(size = 10, angle = 90, hjust = 1, vjust = 0.5),
        axis.text.y = element_text(size = 10),
        legend.title = element_blank(),
        legend.position = "bottom",
        legend.text = element_text(size = 10),
        panel.background = element_rect(fill = "grey95"),
        panel.grid = element_blank(),
        plot.background = element_rect(fill = "gray95")) +
  scale_fill_gradient(low = "#FFC2C2", high = "#C40000") +
  scale_x_discrete(labels = function(x) str_wrap(x, width = 10))
```



Visualization 4: India's new deaths on time series

```
ggplot(data = india_new_cases_deaths ) +
  geom_point(aes(x = Date, y = new_deaths), color="Blue") + labs(
    x = "Date",
    y = "New deaths on daily basis",
    title = "India: New deaths on time series",
    subtitle = "Deaths had a significant peak durng delta covid wave in mid-2021",
    caption = "Data source: Johns Hopkins University"
  )  + theme_shooting()
```



Data Model: Analyzing the relationship between new deaths and new cases in India

Creating a linear model

1. A linear model will be created to predict relationship between new deaths and new cases in India

```
india_model <- lm(new_deaths ~ new_cases, data= india_new_cases_deaths)
```

2. Showing the linear model for it's co-efficients

```
india_model
```

```
##
## Call:
## lm(formula = new_deaths ~ new_cases, data = india_new_cases_deaths)
##
## Coefficients:
## (Intercept)    new_cases
##   1.315e+02    8.565e-03
```

3. Summarizing the performance of linear model

```
summary(india_model)
```

```
##
## Call:
## lm(formula = new_deaths ~ new_cases, data = india_new_cases_deaths)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2535.8  -133.1   -99.4    80.0  6452.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.315e+02  1.861e+01   7.063 2.85e-12 ***
## new_cases   8.564e-03  2.265e-04  37.818  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 547 on 1124 degrees of freedom
## Multiple R-squared:  0.5599, Adjusted R-squared:  0.5595
## F-statistic:  1430 on 1 and 1124 DF,  p-value: < 2.2e-16
```

4. Discussion on the model

The p-value of the model is significantly less than .05 indicating that model should be able to accurately predict the count of new deaths based on new cases. Based on above, there is a significant relationship between new deaths and new cases in India.

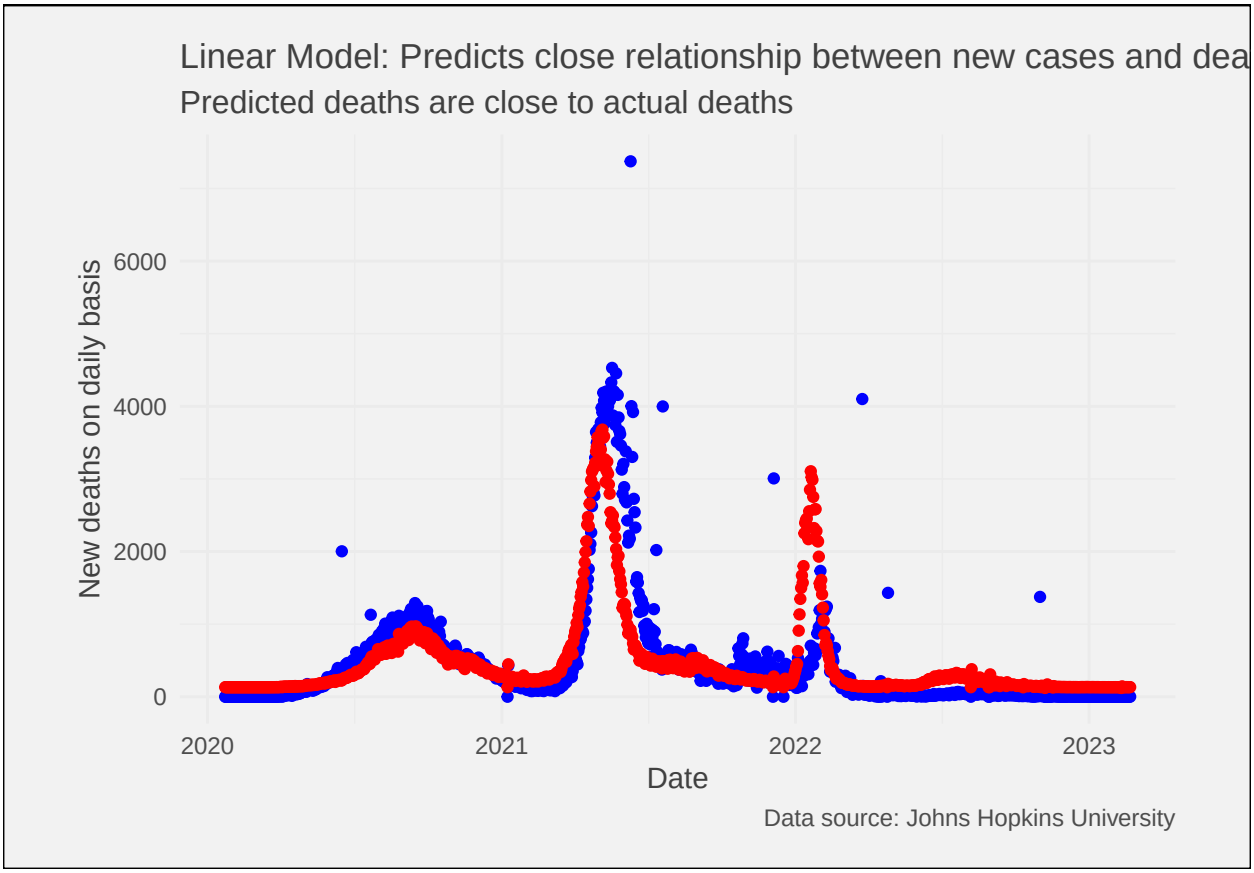
Creating a new dataset to compare linear model vs actual shooting incident

```
india_new_cases_deaths_pred <- india_new_cases_deaths %>% mutate(pred = predict(india_model))
```

6. Model is able to closely follow the relationship between new cases and deaths and it is statistically significant.

Plotting the model performance visually

```
ggplot(data = india_new_cases_deaths_pred ) +
  geom_point(aes(x = Date, y = new_deaths),
             color="Blue") +
  geom_point(aes(x = Date, y = pred), color="Red") +
  labs(
    x = "Date",
    y = "New deaths on daily basis",
    title = "Linear Model: Predicts close relationship between new cases and deaths",
    subtitle = "Predicted deaths are close to actual deaths",
    caption = "Data source: Johns Hopkins University"
  )  + theme_shooting()
```



Data bias and limitations of the Study

1. Data collection bias: The data may be biased if it is collected from certain sources, such as hospitals or testing centers, and does not include all cases or deaths. Additionally, different countries or states may have different testing policies or reporting requirements, which can affect the accuracy and comparability of the data.
2. Confounding bias: There may be other factors that are associated with both the COVID-19 cases/deaths and the independent variable being analyzed (such as population, GDP, or weather), which can make it difficult to establish causality.
3. Bias in testing: The number of confirmed cases and deaths could be influenced by variations in testing rates and policies across different states, leading to under- or over-reporting of cases and deaths.
4. Bias in data reporting: There could be inconsistencies or errors in the way that cases and deaths are reported, which could lead to under- or over-counting of COVID-19 cases and deaths.

Conclusion

1. We can conclude the new deaths is strongly co-related to new cases in India.

- The correlation coefficient measures the strength of the linear relationship between two variables. A high correlation coefficient between new deaths and new cases suggests that the two variables are closely related.

- The number of new cases and new deaths follow a similar pattern over time, it can suggest that there is a relationship between the two variables.
- The relationship between new deaths and new cases holds over an extended period

2. USA witnessed two peaks in COVID cases between December 2020 to January 2021 and December 2021 to January 2022.

- The number of daily new cases and deaths in the USA reached their highest levels during these periods, which suggests that the pandemic was at its peak during that time.
- This conclusion is also supported by news reports and public health data from that time period, which highlighted the strain on hospitals, shortages of medical supplies, and the implementation of stricter measures such as lockdowns and travel restrictions