# Covid_19 Data Analysis Report

MSDS_VB

2023-07-11

## 1. IMPORT COVID_19 DATASET

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
## [1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [3] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [4] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
```

```
## Rows: 289 Columns: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): 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: 1147
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): 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: 3342 Columns: 1154
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): 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.
## Rows: 3342 Columns: 1155
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): 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.
```

```
## [1]  289 1147
```

```
## [1]  289 1147
```

```
## [1] 3342 1154
```

```
## [1] 3342 1155
```

# 2.TIDYING THE DATASET

### Working on global cases data

```
head(global_cases)
```

```
## # A tibble: 6 x 1,147
##   ‘Province/State‘ ‘Country/Region‘   Lat  Long ‘1/22/20‘ ‘1/23/20‘ ‘1/24/20‘
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: ‘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>,
## #   ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>, ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ...
```

### Converting the data from wide format to a long format and tidying by removing lat and long

```r
library(lubridate)
global_cases <- global_cases %>%
    pivot_longer(cols = -c('Province/State',
                           'Country/Region', Lat, Long),
                 names_to = "date",
                 values_to = "cases") %>%
    select(-c(Lat, Long))

head(global_cases)
```

```
## # A tibble: 6 x 4
##   ‘Province/State‘ ‘Country/Region‘ date    cases
##   <chr>            <chr>            <chr>   <dbl>
## 1 <NA>             Afghanistan      1/22/20     0
## 2 <NA>             Afghanistan      1/23/20     0
## 3 <NA>             Afghanistan      1/24/20     0
## 4 <NA>             Afghanistan      1/25/20     0
## 5 <NA>             Afghanistan      1/26/20     0
## 6 <NA>             Afghanistan      1/27/20     0
```

## Checking to see if there are any negative values in cases

```r
global_cases %>% filter(cases < 0)
```

```
## # A tibble: 0 x 4
## # i 4 variables: Province/State <chr>, Country/Region <chr>, date <chr>,
## #   cases <dbl>
```

**There are no negative values in global cases data**

## Working on global deaths data

```r
head(global_deaths)
```

```
## # A tibble: 6 x 1,147
##   ‘Province/State‘ ‘Country/Region‘   Lat  Long ‘1/22/20‘ ‘1/23/20‘ ‘1/24/20‘
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: ‘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>,
## #   ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>, ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ...
```

### Converting the data from wide format to a long format and tidying by removing lat and long

```
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c('Province/State',
                         'Country/Region', Lat, Long),
               names_to = "date",
               values_to = "deaths") %>%
  select(-c(Lat, Long))

head(global_deaths)
```

```
## # A tibble: 6 x 4
##   ‘Province/State‘ ‘Country/Region‘ date    deaths
##   <chr>            <chr>            <chr>    <dbl>
## 1 <NA>             Afghanistan      1/22/20      0
## 2 <NA>             Afghanistan      1/23/20      0
## 3 <NA>             Afghanistan      1/24/20      0
## 4 <NA>             Afghanistan      1/25/20      0
## 5 <NA>             Afghanistan      1/26/20      0
## 6 <NA>             Afghanistan      1/27/20      0
```

### Checking to see if there are any negative values in deaths

```
global_deaths %>% filter(deaths < 0)
```

```
## # A tibble: 0 x 4
## # i 4 variables: Province/State <chr>, Country/Region <chr>, date <chr>,
## #   deaths <dbl>
```

#### There are no negative values in global deaths data

### Joining global_cases and global_deaths into one global data for data exploration analysis

```
global <- global_cases %>%
    full_join(global_deaths) %>%
    rename(`Country_Region` = `Country/Region`) %>%
    mutate(date = mdy(date))
```

```
## Joining with ‘by = join_by(‘Province/State‘, ‘Country/Region‘, date)‘
```

```
global
```

```
## # A tibble: 330,327 x 5
##    ‘Province/State‘ Country_Region date        cases deaths
##    <chr>            <chr>          <date>     <dbl>  <dbl>
##  1 <NA>             Afghanistan    2020-01-22     0      0
##  2 <NA>             Afghanistan    2020-01-23     0      0
##  3 <NA>             Afghanistan    2020-01-24     0      0
##  4 <NA>             Afghanistan    2020-01-25     0      0
##  5 <NA>             Afghanistan    2020-01-26     0      0
##  6 <NA>             Afghanistan    2020-01-27     0      0
##  7 <NA>             Afghanistan    2020-01-28     0      0
##  8 <NA>             Afghanistan    2020-01-29     0      0
##  9 <NA>             Afghanistan    2020-01-30     0      0
## 10 <NA>             Afghanistan    2020-01-31     0      0
## # i 330,317 more rows
```

```r
unique(global$Country_Region)
```

```
##   [1] "Afghanistan"                      "Albania"
##   [3] "Algeria"                          "Andorra"
##   [5] "Angola"                           "Antarctica"
##   [7] "Antigua and Barbuda"              "Argentina"
##   [9] "Armenia"                          "Australia"
##  [11] "Austria"                          "Azerbaijan"
##  [13] "Bahamas"                          "Bahrain"
##  [15] "Bangladesh"                       "Barbados"
##  [17] "Belarus"                          "Belgium"
##  [19] "Belize"                           "Benin"
##  [21] "Bhutan"                           "Bolivia"
##  [23] "Bosnia and Herzegovina"           "Botswana"
##  [25] "Brazil"                           "Brunei"
##  [27] "Bulgaria"                         "Burkina Faso"
##  [29] "Burma"                            "Burundi"
##  [31] "Cabo Verde"                       "Cambodia"
##  [33] "Cameroon"                         "Canada"
##  [35] "Central African Republic"         "Chad"
##  [37] "Chile"                            "China"
##  [39] "Colombia"                         "Comoros"
##  [41] "Congo (Brazzaville)"              "Congo (Kinshasa)"
##  [43] "Costa Rica"                       "Cote d’Ivoire"
##  [45] "Croatia"                          "Cuba"
##  [47] "Cyprus"                           "Czechia"
##  [49] "Denmark"                          "Diamond Princess"
##  [51] "Djibouti"                         "Dominica"
##  [53] "Dominican Republic"               "Ecuador"
##  [55] "Egypt"                            "El Salvador"
##  [57] "Equatorial Guinea"                "Eritrea"
##  [59] "Estonia"                          "Eswatini"
##  [61] "Ethiopia"                         "Fiji"
##  [63] "Finland"                          "France"
##  [65] "Gabon"                            "Gambia"
##  [67] "Georgia"                          "Germany"
##  [69] "Ghana"                            "Greece"
##  [71] "Grenada"                          "Guatemala"
##  [73] "Guinea"                           "Guinea-Bissau"
##  [75] "Guyana"                           "Haiti"
##  [77] "Holy See"                         "Honduras"
##  [79] "Hungary"                          "Iceland"
##  [81] "India"                            "Indonesia"
##  [83] "Iran"                             "Iraq"
##  [85] "Ireland"                          "Israel"
##  [87] "Italy"                            "Jamaica"
##  [89] "Japan"                            "Jordan"
##  [91] "Kazakhstan"                       "Kenya"
##  [93] "Kiribati"                         "Korea, North"
##  [95] "Korea, South"                     "Kosovo"
##  [97] "Kuwait"                           "Kyrgyzstan"
##  [99] "Laos"                             "Latvia"
## [101] "Lebanon"                          "Lesotho"
## [103] "Liberia"                          "Libya"
## [105] "Liechtenstein"                    "Lithuania"
## [107] "Luxembourg"                       "MS Zaandam"
## [109] "Madagascar"                       "Malawi"
## [111] "Malaysia"                         "Maldives"
## [113] "Mali"                             "Malta"
## [115] "Marshall Islands"                 "Mauritania"
## [117] "Mauritius"                        "Mexico"
## [119] "Micronesia"                       "Moldova"
## [121] "Monaco"                           "Mongolia"
## [123] "Montenegro"                       "Morocco"
## [125] "Mozambique"                       "Namibia"
## [127] "Nauru"                            "Nepal"
## [129] "Netherlands"                      "New Zealand"
## [131] "Nicaragua"                        "Niger"
## [133] "Nigeria"                          "North Macedonia"
## [135] "Norway"                           "Oman"
## [137] "Pakistan"                         "Palau"
## [139] "Panama"                           "Papua New Guinea"
## [141] "Paraguay"                         "Peru"
## [143] "Philippines"                      "Poland"
## [145] "Portugal"                         "Qatar"
## [147] "Romania"                          "Russia"
## [149] "Rwanda"                           "Saint Kitts and Nevis"
## [151] "Saint Lucia"                      "Saint Vincent and the Grenadines"
## [153] "Samoa"                            "San Marino"
## [155] "Sao Tome and Principe"            "Saudi Arabia"
## [157] "Senegal"                          "Serbia"
## [159] "Seychelles"                       "Sierra Leone"
## [161] "Singapore"                        "Slovakia"
## [163] "Slovenia"                         "Solomon Islands"
## [165] "Somalia"                          "South Africa"
## [167] "South Sudan"                      "Spain"
## [169] "Sri Lanka"                        "Sudan"
## [171] "Summer Olympics 2020"             "Suriname"
## [173] "Sweden"                           "Switzerland"
## [175] "Syria"                            "Taiwan*"
## [177] "Tajikistan"                       "Tanzania"
## [179] "Thailand"                         "Timor-Leste"
## [181] "Togo"                             "Tonga"
## [183] "Trinidad and Tobago"              "Tunisia"
## [185] "Turkey"                           "Tuvalu"
## [187] "US"                               "Uganda"
## [189] "Ukraine"                          "United Arab Emirates"
## [191] "United Kingdom"                   "Uruguay"
## [193] "Uzbekistan"                       "Vanuatu"
## [195] "Venezuela"                        "Vietnam"
## [197] "West Bank and Gaza"               "Winter Olympics 2022"
## [199] "Yemen"                            "Zambia"
## [201] "Zimbabwe"
```

**Out of 201 unique countries list,I found two names (Summer Olympics 2020,Winter Olympics 2022) which doesn't make sense as Country or Region. So I removed it from the dataframe.**

```
global <- global %>%
  filter(!(Country_Region == 'Summer Olympics 2020') & !(Country_Region == 'Winter Olympics 2022'))

global
```

```
## # A tibble: 328,041 x 5
##    ‘Province/State‘ Country_Region date       cases deaths
##    <chr>            <chr>          <date>     <dbl>  <dbl>
##  1 <NA>             Afghanistan    2020-01-22     0      0
##  2 <NA>             Afghanistan    2020-01-23     0      0
##  3 <NA>             Afghanistan    2020-01-24     0      0
##  4 <NA>             Afghanistan    2020-01-25     0      0
##  5 <NA>             Afghanistan    2020-01-26     0      0
##  6 <NA>             Afghanistan    2020-01-27     0      0
##  7 <NA>             Afghanistan    2020-01-28     0      0
##  8 <NA>             Afghanistan    2020-01-29     0      0
##  9 <NA>             Afghanistan    2020-01-30     0      0
## 10 <NA>             Afghanistan    2020-01-31     0      0
## # i 328,031 more rows
```

## Global data after filtering the cases to more than 1

```
global <- global %>% filter(cases>0)
head(global)
```

```
## # A tibble: 6 x 5
##   ‘Province/State‘ Country_Region date       cases deaths
##   <chr>            <chr>          <date>     <dbl>  <dbl>
## 1 <NA>             Afghanistan    2020-02-24     5      0
## 2 <NA>             Afghanistan    2020-02-25     5      0
## 3 <NA>             Afghanistan    2020-02-26     5      0
## 4 <NA>             Afghanistan    2020-02-27     5      0
## 5 <NA>             Afghanistan    2020-02-28     5      0
## 6 <NA>             Afghanistan    2020-02-29     5      0
```

### Country wise cases and deaths count

```
new_global <-global %>% group_by(Country_Region) %>%
  summarise(total_cases= sum(cases),
            total_deaths = sum(deaths))
new_global
```

```
## # A tibble: 199 x 3
##    Country_Region      total_cases total_deaths
##    <chr>                     <dbl>        <dbl>
##  1 Afghanistan           129988469      5421435
##  2 Albania               185562654      2485380
##  3 Algeria               182741650      4901275
##  4 Andorra                24547525       127190
##  5 Angola                 60025203      1231834
##  6 Antarctica                 4961            0
##  7 Antigua and Barbuda     4310255        80291
##  8 Argentina            5625482921     91037145
##  9 Armenia               285491323      5705393
## 10 Australia            3508864881      5590832
## # i 189 more rows
```

### Countries with high number of deaths

```
new_global1 <- new_global %>% arrange(desc(total_deaths))
new_global1
```

```
## # A tibble: 199 x 3
##    Country_Region total_cases total_deaths
##    <chr>                <dbl>        <dbl>
##  1 US             53813184406    713877215
##  2 Brazil         21182690594    488181000
##  3 India          29131119694    364921237
##  4 Mexico          3944108014    241085189
##  5 Russia         10578569842    220983590
##  6 Peru            2499413018    170749849
##  7 United Kingdom 12118271679    160836676
##  8 Italy          10083161678    127936784
##  9 France         16105911886    113410357
## 10 Colombia        4214829115    100671637
## # i 189 more rows
```

**US has highest number of deaths recorded followed by Brazil.**

### Countries with high number of cases

```
new_global2 <- new_global %>% arrange(desc(total_cases))
new_global2
```

```
## # A tibble: 199 x 3
##    Country_Region total_cases total_deaths
##    <chr>                <dbl>        <dbl>
##  1 US             53813184406    713877215
##  2 India          29131119694    364921237
##  3 Brazil         21182690594    488181000
##  4 France         16105911886    113410357
##  5 Germany        13686043720     96058800
##  6 United Kingdom 12118271679    160836676
##  7 Russia         10578569842    220983590
##  8 Italy          10083161678    127936784
##  9 Turkey          8840742699     62808714
## 10 Korea, South    8467888968     11220890
## # i 189 more rows
```

**US has high number of cases recorded followed by India.**

## Countries with low number of deaths

```
new_global3 <- new_global %>% arrange(total_deaths)
new_global3
```

```
## # A tibble: 199 x 3
##    Country_Region   total_cases total_deaths
##    <chr>                  <dbl>        <dbl>
##  1 Antarctica              4961            0
##  2 Holy See               26807            0
##  3 Tuvalu                322901            0
##  4 Nauru                1184912          251
##  5 Korea, North             300         1800
##  6 MS Zaandam              9665         2146
##  7 Palau                2074263         2648
##  8 Marshall Islands     3135141         3463
##  9 Tonga                4975228         4140
## 10 Vanuatu              3782631         4939
## # i 189 more rows
```

**There are no deaths recorded in countries like Antartica, Holy See and Tuvalu. This may be either due to not being reported or counted as covid related deaths which may include bias in the data.**

## Countries with low number of cases

```
new_global4 <- new_global %>% arrange(total_cases)
new_global4
```

```
## # A tibble: 199 x 3
##    Country_Region        total_cases total_deaths
##    <chr>                       <dbl>        <dbl>
##   1 Korea, North                 300         1800
##   2 Antarctica                  4961            0
##   3 MS Zaandam                  9665         2146
##   4 Holy See                   26807            0
##   5 Tuvalu                    322901            0
##   6 Diamond Princess          796020        14189
##   7 Nauru                    1184912          251
##   8 Kiribati                 1396540         5290
##   9 Palau                    2074263         2648
## 10 Saint Kitts and Nevis     2981130        21522
## # i 189 more rows
```

**North Korea has least cases recorded, but deaths are way more than cases which clearly indicates improper data acquisition. In ideal cases cases should be more than deaths.**

### Yearwise global cases and deaths of 199 countries

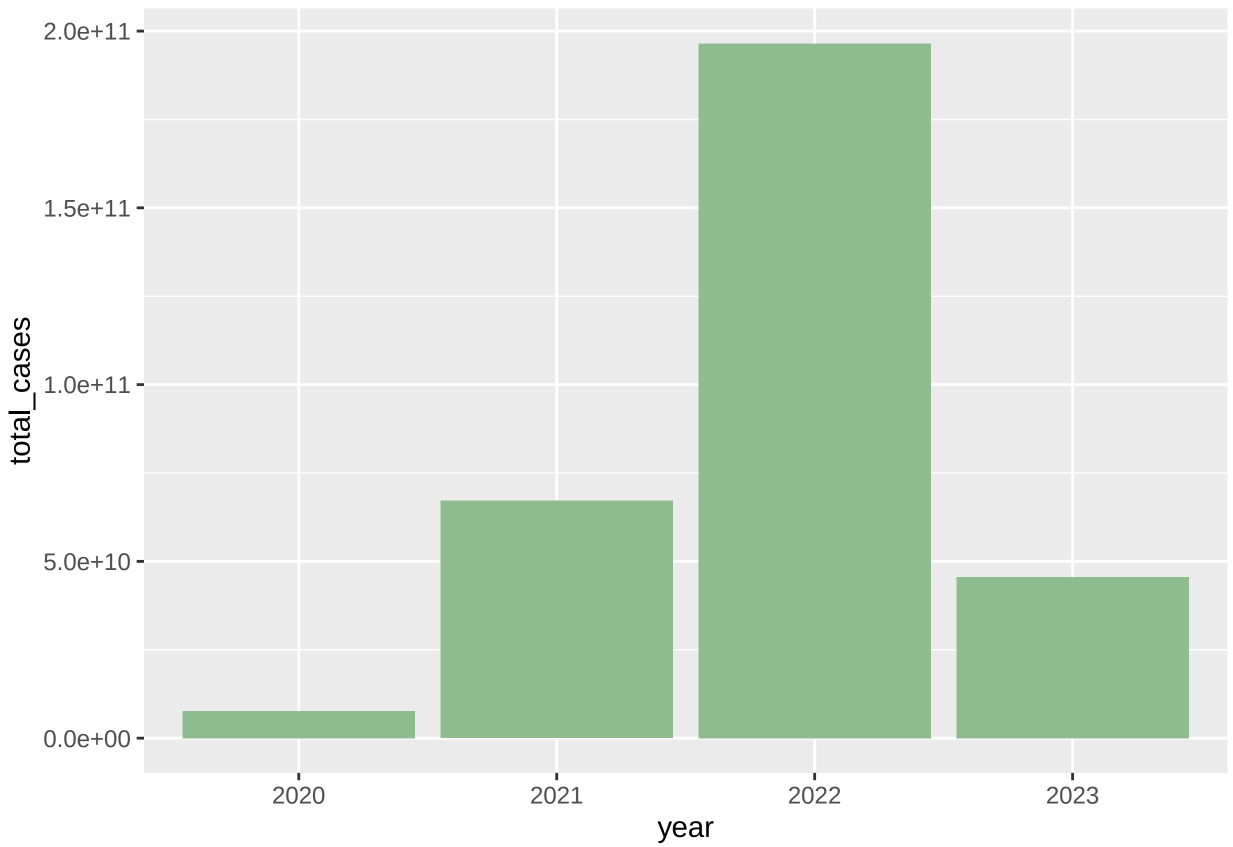
```r
global_year <- global %>%
  mutate(year = format(date,"%Y")) %>%
  group_by(year) %>%
  summarise(total_cases = sum(cases) , total_deaths = sum(deaths))

global_year
```

```
## # A tibble: 4 x 3
##   year   total_cases total_deaths
##   <chr>        <dbl>        <dbl>
## 1 2020    7642565602    237467004
## 2 2021   67131593849   1417654187
## 3 2022  196528815359   2301003485
## 4 2023   45606608359    463691160
```

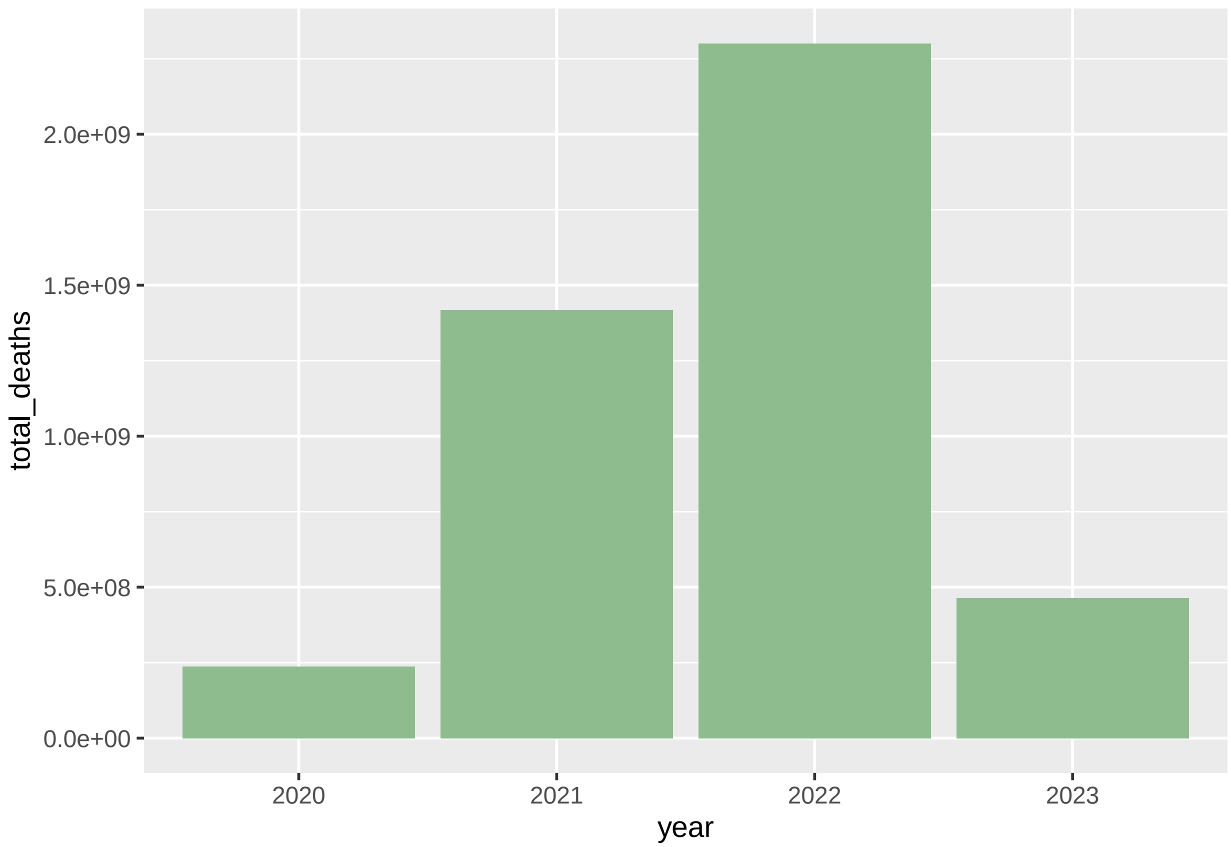
### Visualize barplot for total cases over the 4 years on global data

```r
global_year %>%
  ggplot(aes(x= year, y = total_cases)) +
  geom_bar(fill="darkseagreen",stat="identity")
```



### Visualize barplot for total deaths over the 4 years on global data

```
global_year %>%
  ggplot(aes(x= year, y = total_deaths)) +
  geom_bar(fill="darkseagreen",stat="identity")
```



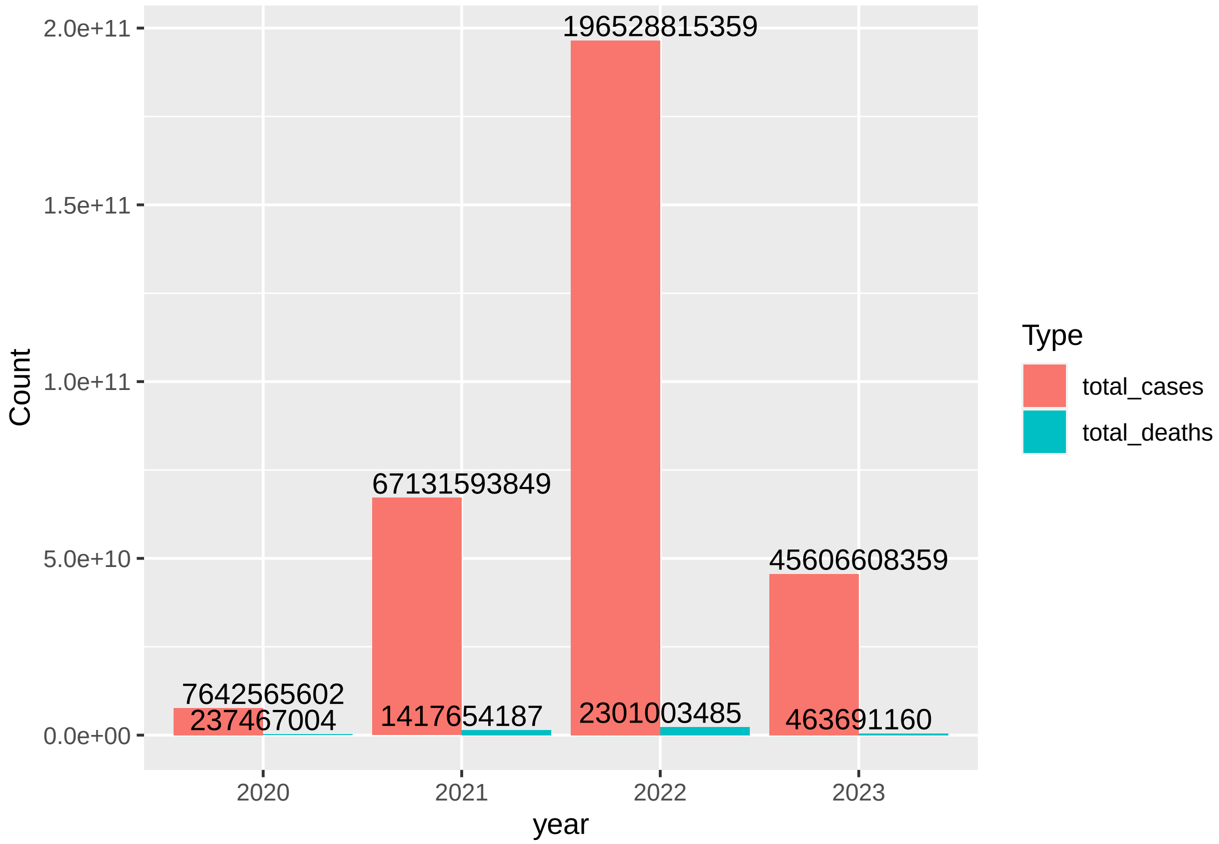
## Visualize barplot for total cases and deaths together over the 4 years on global data

```
df_long <- global_year %>%
  pivot_longer(cols=-year, names_to ="Type", values_to = "Count")

df_long
```

```
## # A tibble: 8 x 3
##   year  Type                Count
##   <chr> <chr>               <dbl>
## 1 2020  total_cases    7642565602
## 2 2020  total_deaths    237467004
## 3 2021  total_cases   67131593849
## 4 2021  total_deaths   1417654187
## 5 2022  total_cases  196528815359
## 6 2022  total_deaths   2301003485
## 7 2023  total_cases   45606608359
## 8 2023  total_deaths    463691160
```

```
df_long %>%
  ggplot(aes(x=year, y= Count , fill= Type))+
  geom_col(position="dodge")+
  geom_text(aes(label = Count), vjust = -0.2)
```



**From the plot, we can see year 2022 has highest number of cases recorded globally due to dangerous variants like Delta and Omicron.**

## Populatin of each country

```
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U


uid <- read_csv(uid_lookup_url) %>%
      select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))
```

```
## 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.
```

```
uid
```

```
## # A tibble: 4,321 x 5
##      UID FIPS  Province_State Country_Region      Population
##    <dbl> <chr> <chr>          <chr>                    <dbl>
##  1     4 <NA>  <NA>           Afghanistan           38928341
##  2     8 <NA>  <NA>           Albania                2877800
##  3    10 <NA>  <NA>           Antarctica                  NA
##  4    12 <NA>  <NA>           Algeria               43851043
##  5    20 <NA>  <NA>           Andorra                  77265
##  6    24 <NA>  <NA>           Angola                32866268
##  7    28 <NA>  <NA>           Antigua and Barbuda      97928
##  8    32 <NA>  <NA>           Argentina             45195777
##  9    51 <NA>  <NA>           Armenia                2963234
## 10    40 <NA>  <NA>           Austria                9006400
## # i 4,311 more rows
```

### Removing missing entries in the Population column

```
uid <- uid %>%
  filter(!is.na(Population))

uid
```

```
## # A tibble: 4,170 x 5
##      UID FIPS  Province_State Country_Region      Population
##    <dbl> <chr> <chr>          <chr>                    <dbl>
##  1     4 <NA>  <NA>           Afghanistan           38928341
##  2     8 <NA>  <NA>           Albania                2877800
##  3    12 <NA>  <NA>           Algeria               43851043
##  4    20 <NA>  <NA>           Andorra                  77265
##  5    24 <NA>  <NA>           Angola                32866268
##  6    28 <NA>  <NA>           Antigua and Barbuda      97928
##  7    32 <NA>  <NA>           Argentina             45195777
##  8    51 <NA>  <NA>           Armenia                2963234
##  9    40 <NA>  <NA>           Austria                9006400
## 10    31 <NA>  <NA>           Azerbaijan            10139175
## # i 4,160 more rows
```

### Removing the first 3 columns from the Population dataset to simplify.

```
uid <- uid %>%
  select(-c(Province_State,UID,FIPS))

uid
```

```
## # A tibble: 4,170 x 2
##   Country_Region      Population
##   <chr>                    <dbl>
## 1 Afghanistan           38928341
## 2 Albania                2877800
## 3 Algeria               43851043
##   4 Andorra                  77265
##   5 Angola                32866268
##   6 Antigua and Barbuda      97928
##   7 Argentina             45195777
##   8 Armenia                2963234
##   9 Austria                9006400
##  10 Azerbaijan            10139175
## # i 4,160 more rows
```

### Group by Country to get total population of each Country

```
new_uid <-uid %>%
  group_by(Country_Region) %>%
  summarise(TotalPopulation= sum(Population))

new_uid
```

```
## # A tibble: 197 x 2
##    Country_Region      TotalPopulation
##    <chr>                         <dbl>
##  1 Afghanistan                38928341
##  2 Albania                     2877800
##  3 Algeria                    43851043
##  4 Andorra                       77265
##  5 Angola                     32866268
##  6 Antigua and Barbuda           97928
##  7 Argentina                  45195777
##  8 Armenia                     2963234
##  9 Australia                  50919400
## 10 Austria                     9006400
## # i 187 more rows
```

### Joining the population column from the new_uid to the new_global data by each country

```
new_global <- new_global %>%
  left_join(new_uid, by = c("Country_Region")) %>%
  select( Country_Region,
         total_cases, total_deaths,TotalPopulation)

new_global
```

```
## # A tibble: 199 x 4
##    Country_Region      total_cases total_deaths TotalPopulation
##    <chr>                     <dbl>        <dbl>           <dbl>
##  1 Afghanistan           129988469      5421435        38928341
##  2 Albania               185562654      2485380         2877800
##  3 Algeria               182741650      4901275        43851043
##  4 Andorra                24547525       127190           77265
##  5 Angola                 60025203      1231834        32866268
##  6 Antarctica                 4961            0              NA
##  7 Antigua and Barbuda     4310255        80291           97928
##  8 Argentina            5625482921     91037145        45195777
##  9 Armenia               285491323      5705393         2963234
## 10 Australia            3508864881      5590832        50919400
## # i 189 more rows
```

**Here population of Antartica and Diamond Princess are not recorded. Also I found another entry listed as "MS Zaandam" which is a cruise ship. So I chose to remove the 3 entries.**

## Remove missing entries from the total population column

```
new_global <- new_global %>%
  filter(!is.na(TotalPopulation))

new_global
```

```
## # A tibble: 196 x 4
##    Country_Region      total_cases total_deaths TotalPopulation
##    <chr>                     <dbl>        <dbl>           <dbl>
##  1 Afghanistan           129988469      5421435        38928341
##  2 Albania               185562654      2485380         2877800
##  3 Algeria               182741650      4901275        43851043
##  4 Andorra                24547525       127190           77265
##  5 Angola                 60025203      1231834        32866268
##  6 Antigua and Barbuda     4310255        80291           97928
##  7 Argentina            5625482921     91037145        45195777
##  8 Armenia               285491323      5705393         2963234
##  9 Australia            3508864881      5590832        50919400
## 10 Austria              2210457634     13732468         9006400
## # i 186 more rows
```

## Deaths per million for each Country(Mortality Rate)

```
global_by_country <- new_global %>%
  mutate(deaths_per_mill = (total_deaths* 1000000) / TotalPopulation) %>%
  select(Country_Region,
         total_cases, total_deaths, TotalPopulation, deaths_per_mill) %>%
  ungroup()

global_by_country %>% arrange(desc(deaths_per_mill))
```

```
## # A tibble: 196 x 5
##    Country_Region       total_cases total_deaths TotalPopulation deaths_per_mill
##    <chr>                      <dbl>        <dbl>           <dbl>           <dbl>
##  1 Bulgaria               683611436     22892110         6948445        3294566.
##  2 Bosnia and Herzegov~   247573190     10346576         3280815        3153660.
##  3 Hungary               1142912051     30193695         9660350        3125528.
##   4 North Macedonia        202398382      6163536          2083380         2958431.
##   5 Montenegro             153358754      1808081           628062         2878826.
##   6 San Marino              10185486        93636            33938         2759031.
##   7 Czechia               2439147631     28258087         10708982         2638728.
##   8 Peru                  2499413018    170749849         65597846         2602980.
##   9 Croatia                647706645     10063965          4105268         2451476.
##  10 Georgia                898731351      9615342          3989175         2410359.
## # i 186 more rows
```

**Bulgaria has high covid mortality rate.**

### Rate of deaths per cases (Case Fatality Rate)

```
new_global1 <- new_global %>%
  mutate(death_rate = (total_deaths*100) / total_cases) %>%
  select(Country_Region,
          TotalPopulation, total_cases, total_deaths, death_rate) %>%
  ungroup()

new_global1 %>% arrange(desc(death_rate))
```

```
## # A tibble: 196 x 5
##    Country_Region         TotalPopulation total_cases total_deaths death_rate
##    <chr>                            <dbl>       <dbl>        <dbl>      <dbl>
##  1 Korea, North                  25778815         300         1800     600
##  2 Yemen                         29825968     7879435      1515446      19.2
##  3 Sudan                         43849269    42936981      3180915       7.41
##  4 Peru                          65597846  2499413018    170749849       6.83
##  5 Mexico                       255584572  3944108014    241085189       6.11
##  6 Syria                         17500657    35209217      2062701       5.86
##  7 Egypt                        102334403   334600873     17248941       5.16
##  8 Somalia                       15893219    17864013       897718       5.03
##  9 Ecuador                       17643060   584150381     26441796       4.53
## 10 Bosnia and Herzegovina         3280815   247573190     10346576       4.18
## # i 186 more rows
```

**Clearly North Korea is an outlier in this case.**

### WORKING ON US CASES AND DEATHS

```
head(US_cases)
```

```
## # A tibble: 6 x 1,154
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # i 1,145 more variables: Long_ <dbl>, Combined_Key <chr>, ‘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>, ...
```

## Converting the data from wide format to a long format

```
US_cases <- US_cases %>%
       pivot_longer(cols = -c(UID:Combined_Key),
              names_to = "date",
              values_to = "cases") %>%
      select(Admin2: cases) %>%
      select(-c(Lat,Long_))

head(US_cases)
```

```
## # A tibble: 6 x 6
##   Admin2  Province_State Country_Region Combined_Key         date    cases
##   <chr>   <chr>          <chr>          <chr>                <chr>   <dbl>
## 1 Autauga Alabama        US             Autauga, Alabama, US 1/22/20     0
## 2 Autauga Alabama        US             Autauga, Alabama, US 1/23/20     0
## 3 Autauga Alabama        US             Autauga, Alabama, US 1/24/20     0
## 4 Autauga Alabama        US             Autauga, Alabama, US 1/25/20     0
## 5 Autauga Alabama        US             Autauga, Alabama, US 1/26/20     0
## 6 Autauga Alabama        US             Autauga, Alabama, US 1/27/20     0
```

## Checking to see if there are any negative values in cases

```
## # A tibble: 3 x 6
##   Admin2     Province_State Country_Region Combined_Key              date  cases
##   <chr>      <chr>          <chr>          <chr>                     <chr> <dbl>
## 1 Unassigned North Carolina US             Unassigned, North Caroli~ 11/9~   -34
## 2 Unassigned South Carolina US             Unassigned, South Caroli~ 5/5/~ -3073
## 3 Unassigned South Carolina US             Unassigned, South Caroli~ 5/6/~ -3073
```

**There are 3 rows with nagative cases. I chose to remove them from the dataset by filtering cases to greater than 0.**

## Further tidying up US_cases

```
US_cases <- US_cases %>%
  filter(cases > 0) %>%
  select(-c(Combined_Key))

US_cases
```

```
## # A tibble: 3,474,292 x 5
##    Admin2  Province_State Country_Region date    cases
##    <chr>   <chr>          <chr>          <chr>   <dbl>
##  1 Autauga Alabama        US             3/24/20     1
##  2 Autauga Alabama        US             3/25/20     5
##  3 Autauga Alabama        US             3/26/20     6
##  4 Autauga Alabama        US             3/27/20     6
##  5 Autauga Alabama        US             3/28/20     6
##  6 Autauga Alabama        US             3/29/20     6
##  7 Autauga Alabama        US             3/30/20     8
##  8 Autauga Alabama        US             3/31/20     8
##  9 Autauga Alabama        US             4/1/20     10
## 10 Autauga Alabama        US             4/2/20     12
## # i 3,474,282 more rows
```

## Cases per state

```r
US_cases_state <- US_cases %>%
  group_by(Province_State) %>%
  summarise(total_cases= sum(cases)) %>%
  select(Province_State,total_cases)

US_cases_state
```

```
## # A tibble: 58 x 2
##    Province_State   total_cases
##    <chr>                  <dbl>
##  1 Alabama            872756073
##  2 Alaska             153011898
##  3 American Samoa       2608837
##  4 Arizona           1330372436
##  5 Arkansas           549955573
##  6 California        6166190335
##  7 Colorado           922394521
##  8 Connecticut        507631287
##  9 Delaware           171886464
## 10 Diamond Princess       53306
## # i 48 more rows
```

## Working on US deaths

```
## # A tibble: 6 x 1,155
##        UID iso2  iso3  code3  FIPS Admin2  Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>          <chr>          <dbl>
## 1 84001001 US    USA     840  1001 Autauga Alabama        US              32.5
## 2 84001003 US    USA     840  1003 Baldwin Alabama        US              30.7
## 3 84001005 US    USA     840  1005 Barbour Alabama        US              31.9
## 4 84001007 US    USA     840  1007 Bibb    Alabama        US              33.0
## 5 84001009 US    USA     840  1009 Blount  Alabama        US              34.0
## 6 84001011 US    USA     840  1011 Bullock Alabama        US              32.1
## # i 1,146 more variables: Long_ <dbl>, Combined_Key <chr>, 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>, ...
```

## Converting the data from wide format to a long format

```
## # A tibble: 6 x 7
##   Admin2  Province_State Country_Region Combined_Key     Population date  deaths
##   <chr>   <chr>          <chr>          <chr>                 <dbl> <chr>  <dbl>
## 1 Autauga Alabama        US             Autauga, Alabam~      55869 1/22~      0
## 2 Autauga Alabama        US             Autauga, Alabam~      55869 1/23~      0
## 3 Autauga Alabama        US             Autauga, Alabam~      55869 1/24~      0
## 4 Autauga Alabama        US             Autauga, Alabam~      55869 1/25~      0
## 5 Autauga Alabama        US             Autauga, Alabam~      55869 1/26~      0
## 6 Autauga Alabama        US             Autauga, Alabam~      55869 1/27~      0
```

## Checking to see if there are any negative values in cases

```
## # A tibble: 3 x 7
##   Admin2     Province_State Country_Region Combined_Key  Population date  deaths
##   <chr>      <chr>          <chr>          <chr>              <dbl> <chr>  <dbl>
## 1 Unassigned North Carolina US             Unassigned, ~          0 11/9~     -6
## 2 Unassigned South Carolina US             Unassigned, ~          0 5/5/~    -82
## 3 Unassigned South Carolina US             Unassigned, ~          0 5/6/~    -82
```

**There are 3 rows with negative entries, so I chose to filter the data by deaths greater than or equal to 0.**

## Further tidying up US_deaths

```
US_deaths <- US_deaths %>%
  filter(deaths >= 0) %>%
  select(-c(Combined_Key))
```

```
US_deaths
```

```
## # A tibble: 3,819,903 x 6
##    Admin2  Province_State Country_Region Population date    deaths
##    <chr>   <chr>          <chr>               <dbl> <chr>    <dbl>
##  1 Autauga Alabama        US                  55869 1/22/20      0
##  2 Autauga Alabama        US                  55869 1/23/20      0
##  3 Autauga Alabama        US                  55869 1/24/20      0
##  4 Autauga Alabama        US                  55869 1/25/20      0
##  5 Autauga Alabama        US                  55869 1/26/20      0
##  6 Autauga Alabama        US                  55869 1/27/20      0
##  7 Autauga Alabama        US                  55869 1/28/20      0
##  8 Autauga Alabama        US                  55869 1/29/20      0
##  9 Autauga Alabama         US                  55869 1/30/20      0
## 10 Autauga Alabama         US                  55869 1/31/20      0
## # i 3,819,893 more rows
```

### Lets count population of each state

```
pop_state <- US_deaths %>%
  select(Admin2, Province_State, Population)

pop_state <- distinct(pop_state)
pop_state
```

```
## # A tibble: 3,342 x 3
##    Admin2   Province_State Population
##    <chr>    <chr>               <dbl>
##  1 Autauga  Alabama             55869
##  2 Baldwin  Alabama            223234
##  3 Barbour  Alabama             24686
##  4 Bibb     Alabama             22394
##  5 Blount   Alabama             57826
##  6 Bullock  Alabama             10101
##  7 Butler   Alabama             19448
##  8 Calhoun  Alabama            113605
##  9 Chambers Alabama             33254
## 10 Cherokee Alabama             26196
## # i 3,332 more rows
```

### Total Population of each state and filtering out rows with missing population total entries

```
pop_state <- pop_state %>%
  group_by(Province_State) %>%
  summarise(Total_Population = sum(Population)) %>%
  select(Province_State, Total_Population)

pop_state <- pop_state %>% filter(Total_Population >0)
pop_state
```

```
## # A tibble: 56 x 2
##    Province_State       Total_Population
##    <chr>                           <dbl>
##  1 Alabama                       4903185
##  2 Alaska                         740995
##  3 American Samoa                  55641
##  4 Arizona                       7278717
##  5 Arkansas                      3017804
##  6 California                   39512223
##  7 Colorado                      5758736
##  8 Connecticut                   3565287
##  9 Delaware                            973764
## 10 District of Columbia                705749
## # i 46 more rows
```

### Deaths per state

```
US_deaths_state <- US_deaths %>%
  group_by(Province_State) %>%
  summarise(total_deaths= sum(deaths))

US_deaths_state
```

```
## # A tibble: 58 x 2
##    Province_State   total_deaths
##    <chr>                   <dbl>
##  1 Alabama              13398261
##  2 Alaska                 751555
##  3 American Samoa          10804
##  4 Arizona              20789702
##  5 Arkansas              7721989
##  6 California           65490302
##  7 Colorado              8942186
##  8 Connecticut           8911110
##  9 Delaware              2089142
## 10 Diamond Princess            0
## # i 48 more rows
```

### Join US_cases_state and US_deaths_state

```
US <- US_cases_state %>%
  full_join(US_deaths_state)
```

```
## Joining with ‘by = join_by(Province_State)‘
```

```
US
```

```
## # A tibble: 58 x 3
##    Province_State   total_cases total_deaths
##    <chr>                  <dbl>        <dbl>
##  1 Alabama            872756073     13398261
##  2 Alaska             153011898       751555
##  3 American Samoa       2608837        10804
##  4 Arizona           1330372436     20789702
##  5 Arkansas           549955573      7721989
##  6 California        6166190335     65490302
##  7 Colorado           922394521      8942186
##  8 Connecticut        507631287      8911110
##  9 Delaware           171886464      2089142
## 10 Diamond Princess       53306            0
## # i 48 more rows
```

## Add Population column

```
US <- US %>%
  left_join(pop_state, by = c("Province_State")) %>%
  select( Province_State,
         total_cases, total_deaths,Total_Population)

US
```

```
## # A tibble: 58 x 4
##    Province_State   total_cases total_deaths Total_Population
##    <chr>                  <dbl>        <dbl>            <dbl>
##  1 Alabama            872756073     13398261          4903185
##  2 Alaska             153011898       751555           740995
##  3 American Samoa       2608837        10804            55641
##  4 Arizona           1330372436     20789702          7278717
##  5 Arkansas           549955573      7721989          3017804
##  6 California        6166190335     65490302         39512223
##  7 Colorado           922394521      8942186          5758736
##  8 Connecticut        507631287      8911110          3565287
##  9 Delaware           171886464      2089142           973764
## 10 Diamond Princess       53306            0               NA
## # i 48 more rows
```

**I found two names under Province_state column(Diamond Princess,Grand Princess). They are cruise ships and not states, So I removed it from the dataframe.**

```
## # A tibble: 56 x 4
##    Province_State       total_cases total_deaths Total_Population
##    <chr>                      <dbl>        <dbl>            <dbl>
##  1 Alabama                872756073     13398261          4903185
##  2 Alaska                 153011898       751555           740995
##  3 American Samoa           2608837        10804            55641
##  4 Arizona               1330372436     20789702          7278717
##  5 Arkansas               549955573      7721989          3017804
##  6 California            6166190335     65490302         39512223
##  7 Colorado               922394521      8942186          5758736
##  8 Connecticut            507631287      8911110          3565287
##  9 Delaware               171886464      2089142           973764
## 10 District of Columbia    90279276      1140001           705749
## # i 46 more rows
```

## States with high deaths

```
## # A tibble: 56 x 4
##    Province_State total_cases total_deaths Total_Population
##    <chr>                <dbl>        <dbl>            <dbl>
##  1 California      6166190335     65490302         39512223
##  2 Texas           4566537657     61302166         28995881
##  3 New York        3392006819     58121236         19453561
##  4 Florida         3978357707     51475342         21477737
##  5 Pennsylvania    1836846159     31912144         12801989
##  6 Illinois        2122240785     28240376         12671821
##  7 New Jersey      1536872925     28101090          8882190
##  8 Georgia         1698658727     26228841         10617423
##  9 Ohio            1765525036     26072614         11689100
## 10 Michigan        1561076712     25546398          9986857
## # i 46 more rows
```

### States with high cases

```
## # A tibble: 56 x 4
##    Province_State total_cases total_deaths Total_Population
##    <chr>                <dbl>        <dbl>            <dbl>
##  1 California      6166190335     65490302         39512223
##  2 Texas           4566537657     61302166         28995881
##  3 Florida         3978357707     51475342         21477737
##  4 New York        3392006819     58121236         19453561
##  5 Illinois        2122240785     28240376         12671821
##  6 Pennsylvania    1836846159     31912144         12801989
##  7 Ohio            1765525036     26072614         11689100
##  8 North Carolina  1726912486     16746953         10488084
##  9 Georgia         1698658727     26228841         10617423
## 10 Michigan        1561076712     25546398          9986857
## # i 46 more rows
```

### States with less deaths

```
## # A tibble: 56 x 4
##    Province_State           total_cases total_deaths Total_Population
##    <chr>                          <dbl>        <dbl>            <dbl>
##  1 American Samoa               2608837        10804            55641
##  2 Northern Mariana Islands     5153291        16895            55144
##  3 Virgin Islands              10749871        71105           107268
##  4 Guam                        27172745       232819           164229
##  5 Vermont                     68003350       421227           623989
##  6 Alaska                     153011898       751555           740995
##  7 Hawaii                     153864444       922359          1415872
##  8 Wyoming                    101470234      1136735           578759
##  9 District of Columbia        90279276      1140001           705749
## 10 Maine                      143770501      1420548          1344212
## # i 46 more rows
```

### States with less cases

```
## # A tibble: 56 x 4
##    Province_State           total_cases total_deaths Total_Population
##    <chr>                          <dbl>        <dbl>            <dbl>
##  1 American Samoa               2608837        10804            55641
##  2 Northern Mariana Islands     5153291        16895            55144
##  3 Virgin Islands              10749871        71105           107268
##  4 Guam                        27172745       232819           164229
##  5 Vermont                     68003350       421227           623989
##   6 District of Columbia       90279276      1140001           705749
##   7 Wyoming                   101470234      1136735           578759
##   8 Maine                     143770501      1420548          1344212
##   9 Alaska                    153011898       751555           740995
##  10 Hawaii                    153864444       922359          1415872
## # i 46 more rows
```

### Joining the actual US_cases and US_deaths for further analysis and modeling

```
## Joining with ‘by = join_by(Admin2, Province_State, Country_Region, date)‘
```

```
## # A tibble: 3,819,903 x 7
##    Admin2  Province_State Country_Region date       cases Population deaths
##    <chr>   <chr>          <chr>          <date>     <dbl>      <dbl>  <dbl>
##  1 Autauga Alabama        US             2020-03-24     1      55869      0
##  2 Autauga Alabama        US             2020-03-25     5      55869      0
##  3 Autauga Alabama        US             2020-03-26     6      55869      0
##  4 Autauga Alabama        US             2020-03-27     6      55869      0
##  5 Autauga Alabama        US             2020-03-28     6      55869      0
##  6 Autauga Alabama        US             2020-03-29     6      55869      0
##  7 Autauga Alabama        US             2020-03-30     8      55869      0
##  8 Autauga Alabama        US             2020-03-31     8      55869      0
##  9 Autauga Alabama        US             2020-04-01    10      55869      0
## 10 Autauga Alabama        US             2020-04-02    12      55869      0
## # i 3,819,893 more rows
```

```r
US_year <- US_Total %>%
  select(date,cases,deaths)

US_year
```

```
## # A tibble: 3,819,903 x 3
##    date       cases deaths
##    <date>     <dbl>  <dbl>
##  1 2020-03-24     1      0
##  2 2020-03-25     5      0
##  3 2020-03-26     6      0
##  4 2020-03-27     6      0
##  5 2020-03-28     6      0
##  6 2020-03-29     6      0
##  7 2020-03-30     8      0
##  8 2020-03-31     8      0
##  9 2020-04-01    10      0
## 10 2020-04-02    12      0
## # i 3,819,893 more rows
```

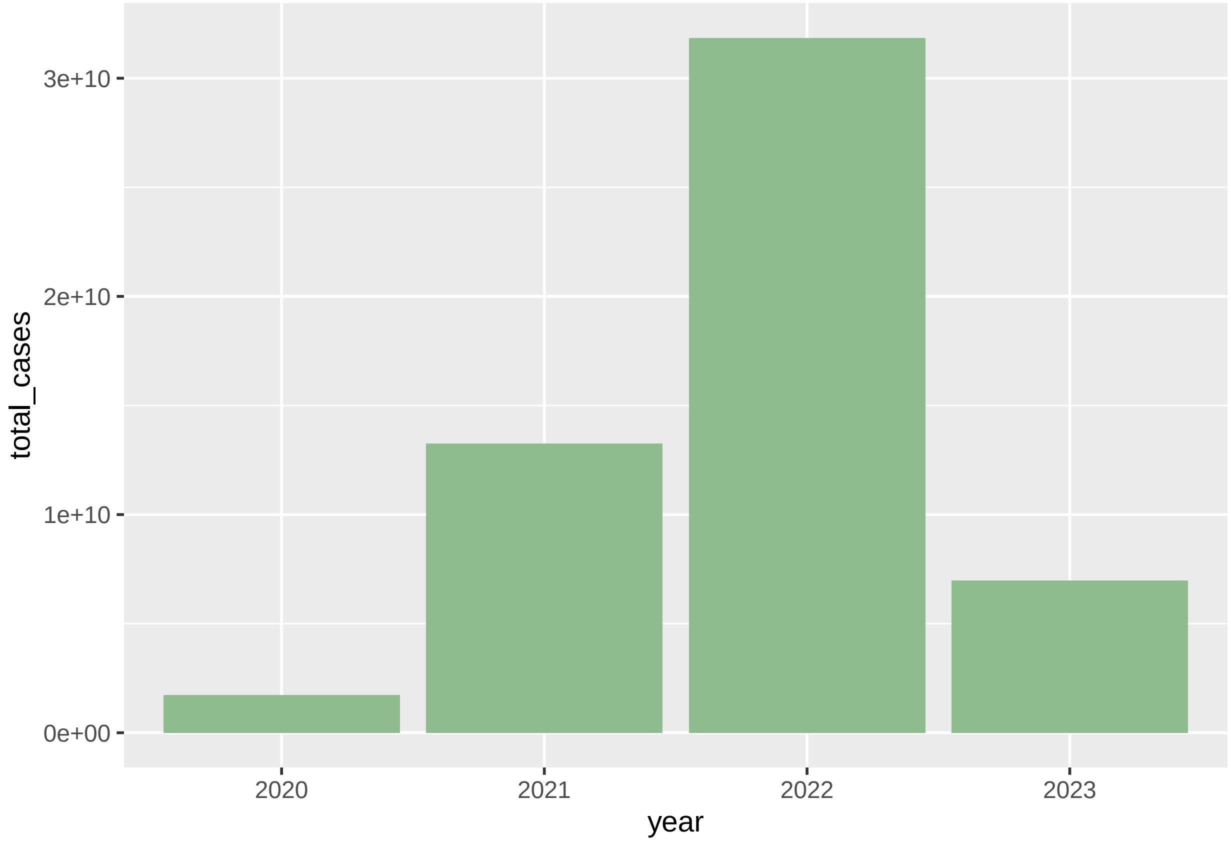
```r
US_year <- US_year %>%
  mutate(year = format(date,"%Y")) %>%
  group_by(year) %>%
  summarise(total_cases = sum(cases, na.rm=T) , total_deaths = sum(deaths))

US_year
```

```
## # A tibble: 4 x 3
##   year  total_cases total_deaths
##   <chr>       <dbl>        <dbl>
## 1 2020   1729023025     46810979
## 2 2021  13268290820    223468200
## 3 2022  31846841612    368125203
## 4 2023   6969198716     75471019
```

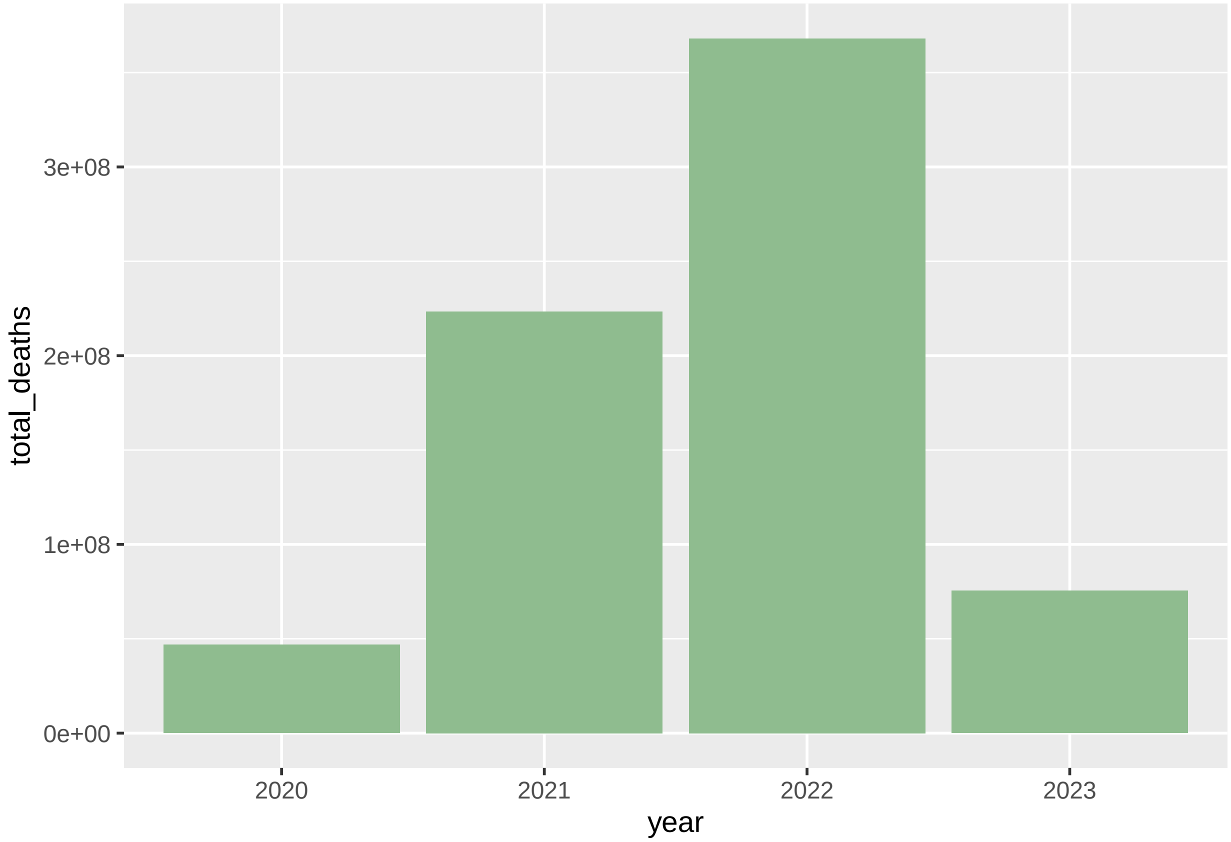
### Visualize barplot for total cases over the 4 years on global data

```
US_year %>%
  ggplot(aes(x= year, y = total_cases)) +
  geom_bar(fill="darkseagreen",stat="identity")
```



### Visualize barplot for total deaths over the 4 years on global data

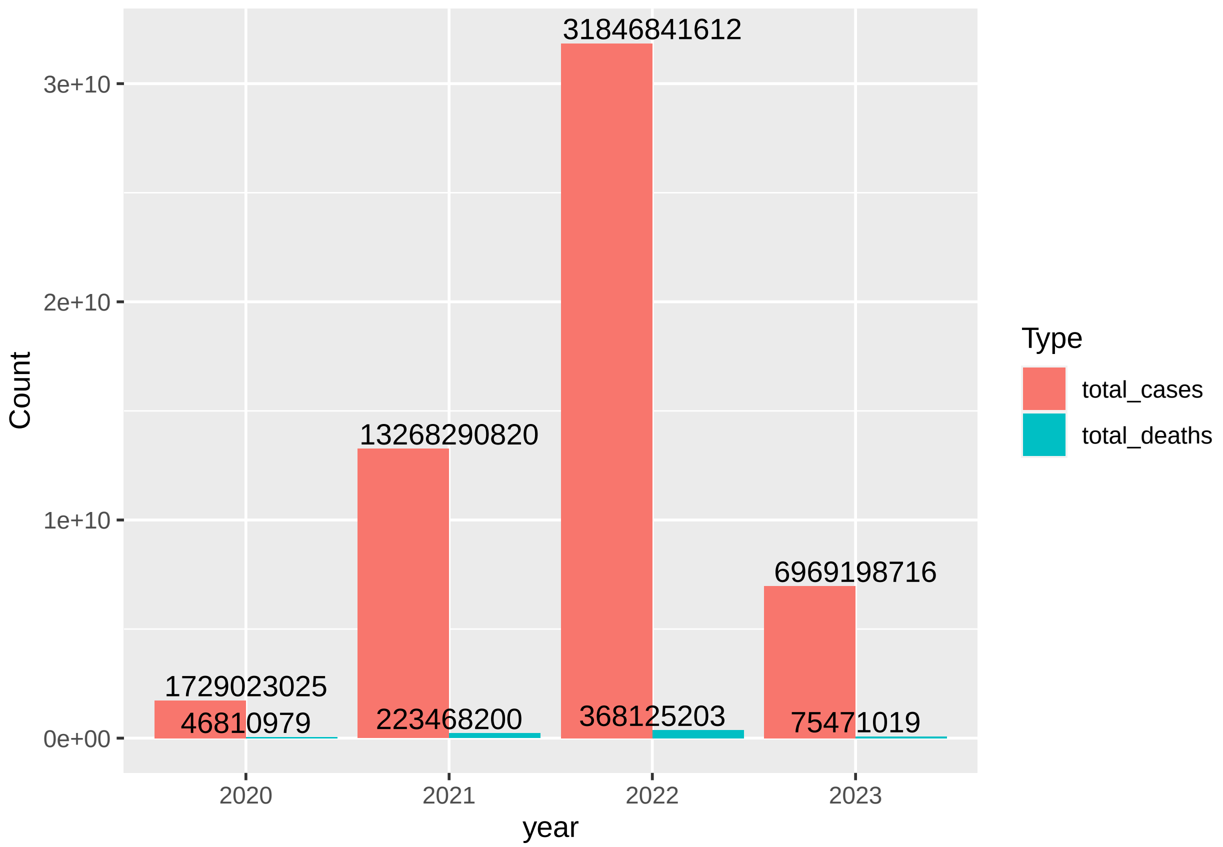
```
US_year %>%
  ggplot(aes(x= year, y = total_deaths)) +
  geom_bar(fill="darkseagreen",stat="identity")
```



## Visualize barplot for total cases and deaths together over the 4 years on global data

```
df_US <- US_year %>%
  pivot_longer(cols=-year, names_to ="Type", values_to = "Count")


df_US %>%
  ggplot(aes(x=year, y= Count , fill= Type))+
  geom_col(position="dodge")+
  geom_text(aes(label = Count), vjust = -0.2)
```



**From the plot, we can see year 2022 has highest number of cases recorded in the US.**

```
US_Total
```

```
## # A tibble: 3,819,903 x 7
##    Admin2  Province_State Country_Region date       cases Population deaths
##    <chr>   <chr>          <chr>          <date>     <dbl>      <dbl>  <dbl>
##  1 Autauga Alabama        US             2020-03-24     1      55869      0
##  2 Autauga Alabama        US             2020-03-25     5      55869      0
##  3 Autauga Alabama        US             2020-03-26     6      55869      0
##  4 Autauga Alabama        US             2020-03-27     6      55869      0
##  5 Autauga Alabama        US             2020-03-28     6      55869      0
##  6 Autauga Alabama        US             2020-03-29     6      55869      0
##  7 Autauga Alabama        US             2020-03-30     8      55869      0
##  8 Autauga Alabama        US             2020-03-31     8      55869      0
##  9 Autauga Alabama        US             2020-04-01    10      55869      0
## 10 Autauga Alabama        US             2020-04-02    12      55869      0
## # i 3,819,893 more rows
```

## Deaths per million by state

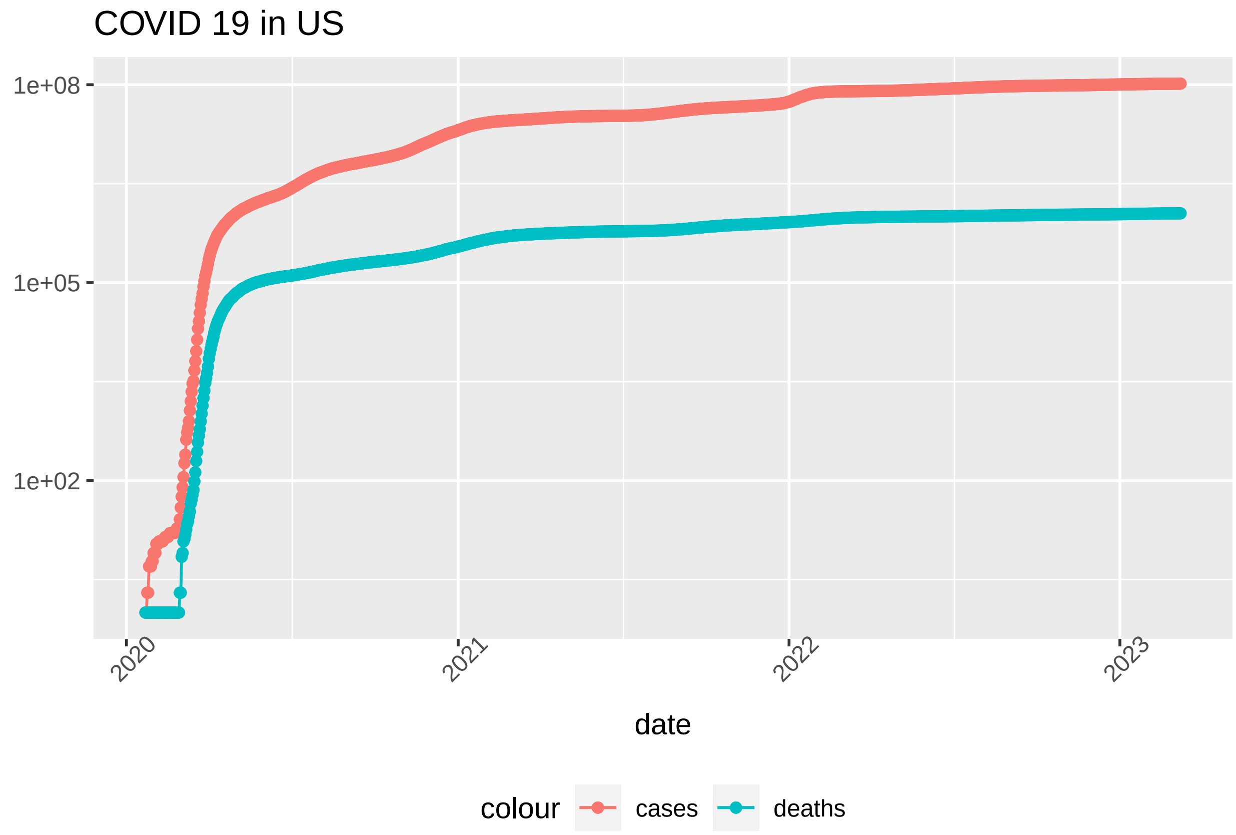
```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can
## override using the `.groups` argument.
## # A tibble: 66,294 x 7
##    Province_State Country_Region date       cases deaths deaths_per_mill
##    <chr>          <chr>          <date>     <dbl>  <dbl>           <dbl>
##  1 Alabama        US             2020-01-22     0      0               0
##  2 Alabama        US             2020-01-23     0      0               0
##  3 Alabama        US             2020-01-24     0      0               0
##  4 Alabama        US             2020-01-25     0      0               0
##  5 Alabama        US             2020-01-26     0      0               0
##  6 Alabama        US             2020-01-27     0      0               0
##  7 Alabama        US             2020-01-28     0      0               0
##  8 Alabama        US             2020-01-29     0      0               0
##  9 Alabama        US             2020-01-30     0      0               0
## 10 Alabama        US             2020-01-31     0      0               0
## # i 66,284 more rows
## # i 1 more variable: Population <dbl>
```

## US totals

```
## ‘summarise()‘ has grouped output by ’Country_Region’. You can override using
## the ‘.groups‘ argument.
```

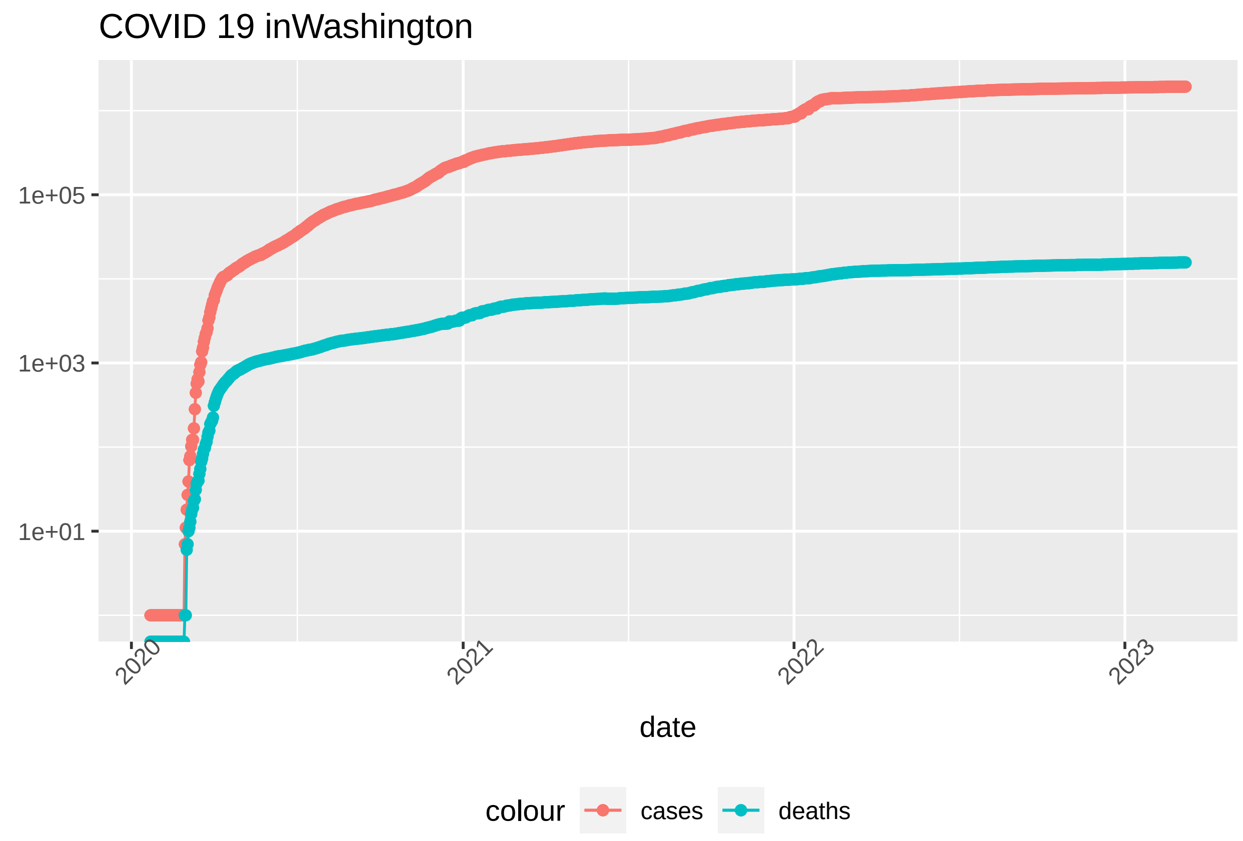
```
## # A tibble: 1,143 x 6
##    Country_Region date       cases deaths deaths_per_mill Population
##    <chr>          <date>     <dbl>  <dbl>           <dbl>      <dbl>
##  1 US             2020-01-22     1      1         0.00300  332875137
##  2 US             2020-01-23     1      1         0.00300  332875137
##  3 US             2020-01-24     2      1         0.00300  332875137
##  4 US             2020-01-25     2      1         0.00300  332875137
##  5 US             2020-01-26     5      1         0.00300  332875137
##  6 US             2020-01-27     5      1         0.00300  332875137
##  7 US             2020-01-28     5      1         0.00300  332875137
##  8 US             2020-01-29     6      1         0.00300  332875137
##  9 US             2020-01-30     6      1         0.00300  332875137
## 10 US             2020-01-31     8      1         0.00300  332875137
## # i 1,133 more rows
```

## Visualize the time seriesgraph of covid data in US



### Visualize Washington state covid data

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```



### Adding new variables for analysis

```
## # A tibble: 6 x 8
##   new_cases new_deaths Country_Region date          cases deaths deaths_per_mill
##       <dbl>      <dbl> <chr>          <date>        <dbl>  <dbl>           <dbl>
## 1      2147          7 US             2023-03-04   1.04e8 1.12e6           3371.
## 2     -3862        -38 US             2023-03-05   1.04e8 1.12e6           3371.
## 3      8564         47 US             2023-03-06   1.04e8 1.12e6           3371.
## 4     35371        335 US             2023-03-07   1.04e8 1.12e6           3372.
## 5     64861        730 US             2023-03-08   1.04e8 1.12e6           3374.
## 6     46931        590 US             2023-03-09   1.04e8 1.12e6           3376.
## # i 1 more variable: Population <dbl>
```

### 10 states with less deaths per thousand

```
## # A tibble: 10 x 6
##    Province_State        deaths   cases population cases_per_thou deaths_per_thou
##    <chr>                  <dbl>   <dbl>      <dbl>          <dbl>           <dbl>
##  1 American Samoa            34 8.32e3      55641           150.            0.611
##  2 Northern Mariana Isl~     41 1.37e4      55144           248.            0.744
##  3 Virgin Islands           130 2.48e4     107268           231.            1.21
##  4 Hawaii                  1841 3.81e5    1415872           269.            1.30
##  5 Vermont                  929 1.53e5     623989           245.            1.49
##  6 Puerto Rico             5823 1.10e6    3754939           293.            1.55
##  7 Utah                    5298 1.09e6    3205958           340.           1.65
##  8 Alaska                  1486 3.08e5     740995           415.           2.01
##  9 District of Columbia    1432 1.78e5     705749           252.           2.03
## 10 Washington             15683 1.93e6    7614893           253.           2.06
```

### 10 states with highest deaths per thousand

```
## # A tibble: 10 x 6
##    Province_State deaths   cases population cases_per_thou deaths_per_thou
##    <chr>           <dbl>   <dbl>      <dbl>          <dbl>           <dbl>
##  1 Arizona         33102 2443514    7278717           336.            4.55
##  2 Oklahoma        17972 1290929    3956971           326.            4.54
##  3 Mississippi     13370  990756    2976149           333.            4.49
##  4 West Virginia    7960  642760    1792147           359.            4.44
##  5 New Mexico       9061  670929    2096829           320.            4.32
##  6 Arkansas        13020 1006883    3017804           334.            4.31
##  7 Alabama         21032 1644533    4903185           335.            4.29
##  8 Tennessee       29263 2515130    6829174           368.            4.28
##  9 Michigan        42205 3064125    9986857           307.            4.23
## 10 Kentucky        18130 1718471    4467673           385.            4.06
```

# MODELLING

## Linear Model

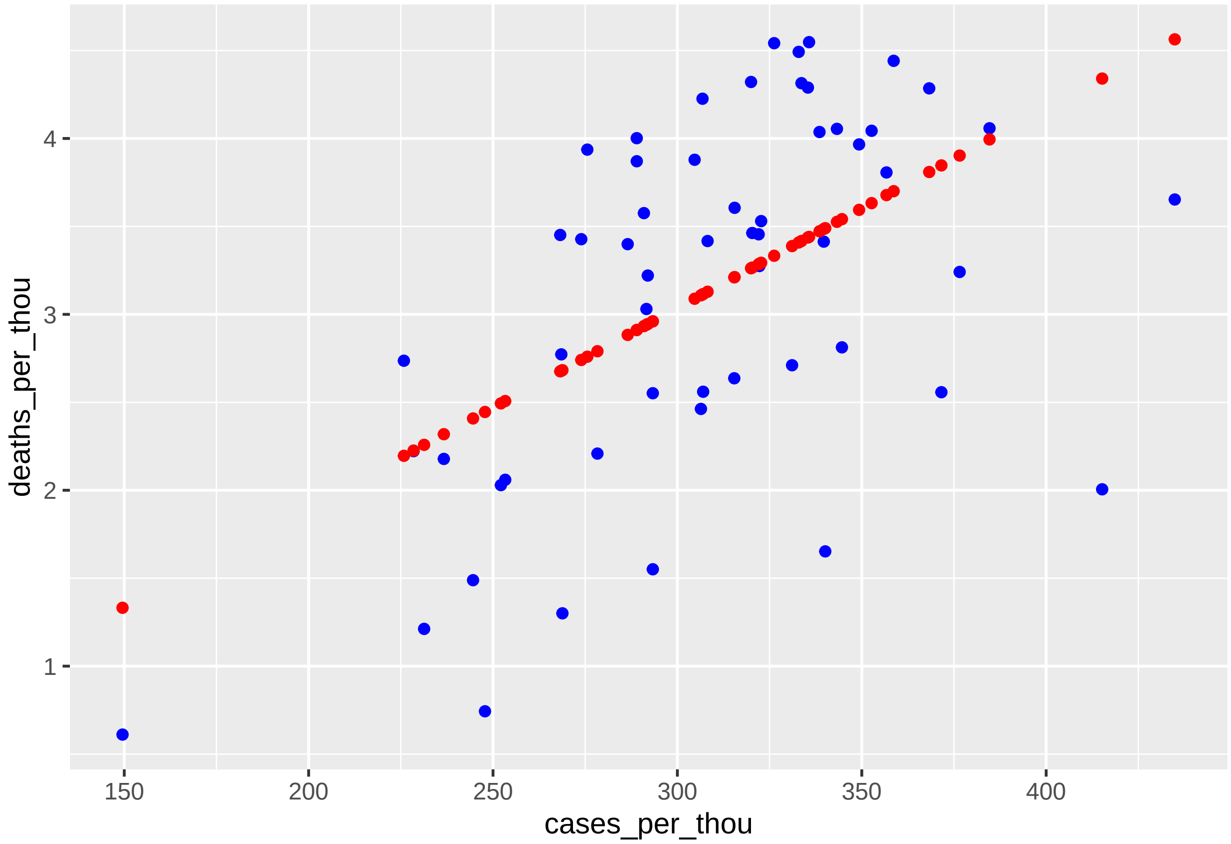
```
## 
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.3352 -0.5978  0.1491  0.6535  1.2086 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -0.36167    0.72480  -0.499     0.62    
## cases_per_thou  0.01133    0.00232   4.881 9.76e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.8615 on 54 degrees of freedom
## Multiple R-squared:  0.3061, Adjusted R-squared:  0.2933 
## F-statistic: 23.82 on 1 and 54 DF,  p-value: 9.763e-06
```

```
## # A tibble: 56 x 7
##    Province_State  deaths   cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>   <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          21032 1.64e6    4903185           335.           4.29   3.44
##  2 Alaska            1486 3.08e5     740995           415.           2.01   4.34
##  3 American Samoa      34 8.32e3      55641           150.           0.611  1.33
##  4 Arizona          33102 2.44e6    7278717           336.           4.55   3.44
##  5 Arkansas         13020 1.01e6    3017804           334.           4.31   3.42
##  6 California      101159 1.21e7   39512223           307.           2.56   3.12
##  7 Colorado         14181 1.76e6    5758736           306.           2.46   3.11
##  8 Connecticut      12220 9.77e5    3565287           274.           3.43   2.74
##  9 Delaware          3324 3.31e5     973764           340.           3.41   3.49
## 10 District of Co~   1432 1.78e5     705749           252.           2.03   2.49
## # i 46 more rows
```

### US totals with prediction

```
## # A tibble: 56 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          21032 1.64e6    4903185           335.           4.29   3.44
##  2 Alaska            1486 3.08e5     740995           415.           2.01   4.34
##  3 American Samoa      34 8.32e3      55641           150.           0.611  1.33
##  4 Arizona          33102 2.44e6    7278717           336.           4.55   3.44
##  5 Arkansas         13020 1.01e6    3017804           334.           4.31   3.42
##  6 California      101159 1.21e7   39512223           307.           2.56   3.12
##  7 Colorado         14181 1.76e6    5758736           306.           2.46   3.11
##  8 Connecticut      12220 9.77e5    3565287           274.           3.43   2.74
##  9 Delaware          3324 3.31e5     973764           340.           3.41   3.49
## 10 District of Co~   1432 1.78e5     705749           252.           2.03   2.49
## # i 46 more rows
```

## Model Plot



## Analysis

**The Model linearly predicted the deaths per the number of cases which is statistically significant.**