Coronavirus (COVID-19) Data Analysis in the United States

Keith Lee

Introduction

Coronavirus (COVID-19) pandemic is the most significant health disaster that has affected the world since its outbreak in early 2020. Since it is a contagious disease by the virus, the spread is entirely dependent on the human interactions. Reducing the chances of infection by wearing a mask, avoiding close distance from others, practicing good hygiene, and others have been a practice of life.

In this project, the data of COVID-19 pandemic in USA from January 2020 till present has been analyzed to examine the statistical distributions. The historical data of the coronavirus cases and deaths for each geography was included in the data. By analyzing it, the specific times and locations for the most incidences were examined. This analysis can also show which locations in the US are more prone to contagious disasters.

Packages Required

Packages with useful collection of functions were loaded in order to reproduce the code and results. They were standard packages used for tidying and analyzing the data. Packages for formatting the data visualization were also loaded.

Data Preparation

Loading Data

The data included incidences in the US from the start till present, including separate files for each year, entire country, and states.

```
## [1] "Daily Incidences in US"
```

```
## [1] 1086    3
```

```
## [1] "Daily Incidences in each State in US"
```

```
## [1] 57910     5
```

```
## [1] "Daily Incidences in each County in US during 2020"
```

```
## [1] 884737      6
```

```
## [1] "Daily Incidences in each County in US during 2021"
## [1] 1185373       6
```

```
## [1] "Daily Incidences in each County in US during 2022"
```

```
## [1] 1188042       6
```

```
## [1] "Daily Incidences in each County in US during 2023"
```

```
## [1] 32533     6
```

Tidying Data

Separate objects of datasets selected by groups of dates were created, including months which was extracted from the date, and by states. The datasets for the cases and deaths were accumulated for different groups: - incidences by year - incidences by month and year - incidences in each county - incidences in each state by month - incidences in each state by year

```
# cleaning and reformating date for all tables
us$date <- as.Date(us$date)
us$yr <- format(as.Date(us$date), "%Y")
us$mnt <- format(as.Date(us$date), "%m")
us$dt <- format(as.Date(us$date), "%d")
us$year_month <- format(as.Date(us$date), "%Y-%m")

usstates$date <- as.Date(usstates$date)
usstates$yr <- format(as.Date(usstates$date), "%Y")
usstates$mnt <- format(as.Date(usstates$date), "%m")
usstates$dt <- format(as.Date(usstates$date), "%d")
usstates$year_month <- format(as.Date(usstates$date), "%Y-%m")

us2020$date <- as.Date(us2020$date)
us2020$yr <- format(as.Date(us2020$date), "%Y")
us2020$mnt <- format(as.Date(us2020$date), "%m")
us2020$dt <- format(as.Date(us2020$date), "%d")
us2020$year_month <- format(as.Date(us2020$date), "%Y-%m")

us2021$date <- as.Date(us2021$date)
us2021$yr <- format(as.Date(us2021$date), "%Y")
us2021$mnt <- format(as.Date(us2021$date), "%m")
us2021$dt <- format(as.Date(us2021$date), "%d")
us2021$year_month <- format(as.Date(us2021$date), "%Y-%m")

us2022$date <- as.Date(us2022$date)
us2022$yr <- format(as.Date(us2022$date), "%Y")
us2022$mnt <- format(as.Date(us2022$date), "%m")
us2022$dt <- format(as.Date(us2022$date), "%d")
us2022$year_month <- format(as.Date(us2022$date), "%Y-%m")

us2023$date <- as.Date(us2023$date)
us2023$yr <- format(as.Date(us2023$date), "%Y")
us2023$mnt <- format(as.Date(us2023$date), "%m")
us2023$dt <- format(as.Date(us2023$date), "%d")
```

```r
us2023$year_month <- format(as.Date(us2023$date), "%Y-%m")

# Data by month and year
us_monthly <- us %>%
  group_by(year_month) %>%
  dplyr::summarise(cases = sum(cases),
            deaths = sum(deaths))
print("Data by month and year")
```

```
## [1] "Data by month and year"
```

```r
dim(us_monthly)
```

```
## [1] 37  3
```

```r
# Data by year
us_yearly <- us %>%
  group_by(yr) %>%
  dplyr::summarise(cases = sum(cases),
            deaths = sum(deaths))
print("Data by year")
```

```
## [1] "Data by year"
```

```r
dim(us_yearly)
```

```
## [1] 4 3
```

```r
# combine all years and extract date elements
us_incidences <- us2020 %>%
  rbind(us2021) %>%
  rbind(us2022) %>%
  rbind(us2023)
print("Total incidences")
```

```
## [1] "Total incidences"
```

```r
dim(us_incidences)
```

```
## [1] 3290685      10
```

```r
# Data by month, year, state
us_monthly_state <- us_incidences %>%
  group_by(year_month, state) %>%
  dplyr::summarise(cases = sum(cases),
            deaths = sum(deaths))
```

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

```
print("Monthly by each state")
```

```
## [1] "Monthly by each state"
```

```
dim(us_monthly_state)
```

```
## [1] 1956    4
```

```
# Data by year, state
us_yearly_state <- us_incidences %>%
  group_by(yr, state) %>%
  dplyr::summarise(cases = sum(cases),
            deaths = sum(deaths))
```

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

```
print("Yearly by each state")
```

```
## [1] "Yearly by each state"
```

```
dim(us_yearly_state)
```

```
## [1] 223   4
```

```
# Cases by the month of each year
us2020_sum <- us2020 %>% group_by(mnt) %>% dplyr::summarise(cases = sum(cases))
us2021_sum <- us2021 %>% group_by(mnt) %>% summarise(cases = sum(cases))
us2022_sum <- us2022 %>% group_by(mnt) %>% dplyr::summarise(cases = sum(cases))
us2023_sum <- us2023 %>% group_by(mnt) %>% dplyr::summarise(cases = sum(cases))
us_incidences_tb <- us2020_sum %>%
  full_join(us2021_sum, by=c("mnt"), suffix=c("_2020","_2021")) %>%
  full_join(us2022_sum, by=c("mnt"), suffix=c("_2020","_2022")) %>%
  full_join(us2023_sum, by=c("mnt"), suffix=c("_2020","_2023"))
print("Total incidences Year Comparison")
```

```
## [1] "Total incidences Year Comparison"
```

```
dim(us_incidences_tb)
```

```
## [1] 12  5
```

```
us_incidences_tb
```

```
## # A tibble: 12 x 5
##    mnt   cases_2020 cases_2021 cases_2020_2020 cases_2023
##    <chr>      <int>      <int>           <dbl>      <int>
##  1 01            41  729984096      2037666117 1008361520
##   2 02           736  773617710      2173725838         NA
##   3 03       1095533  916857453      2465286532         NA
##   4 04      19611708  944997168      2416616245         NA
##   5 05      45452114 1020501942      2562588439         NA
##   6 06      65288844 1003878600      2575997226         NA
##   7 07     111626136 1059339596      2769305619         NA
##   8 08     166758528 1148978762      2879851473         NA
##   9 09     199758786 1247666211      2859276626         NA
## 10 10     252794114 1390097002      2997886391         NA
## 11 11     338932078 1417533896      2936481621         NA
## 12 12     525970605 1575677425      3085650622         NA
```

```
# us_monthly_state
# us_yearly_state
```

Then, the accumulated datasets were filtered to the state of New York to analyze its distributions.

```
us_monthly_state_NY <- us_monthly_state %>%
  filter(state == "New York")
# dim(us_monthly_state_NY)

us_yearly_state_NY <- us_yearly_state %>%
  filter(state == "New York")
# dim(us_yearly_state_NY)

us_NY <- us_incidences %>%
  filter(state == "New York")
print("Filtered to NY")
```

```
## [1] "Filtered to NY"
```

```
dim(us_NY)
```

```
## [1] 59728    10
```

```
us_NY_counties <- us_NY %>%
  group_by(county) %>%
  dplyr::summarise(cases = sum(cases),
            deaths = sum(deaths))
print("Incidences by County")
```

```
## [1] "Incidences by County"
```

```
dim(us_NY_counties)
```

```
## [1] 59  3
```

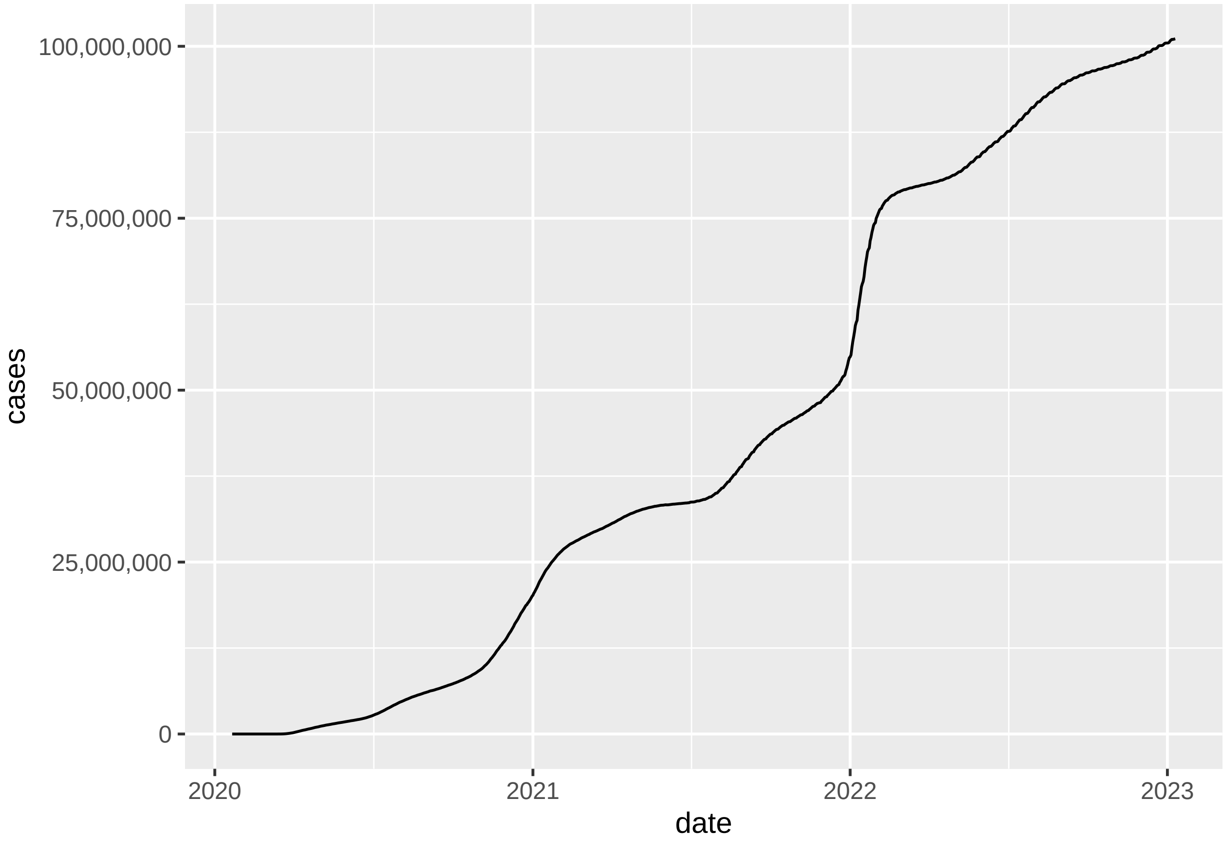
```
us_NY_counties
```

```
## # A tibble: 59 x 3
##    county         cases deaths
##    <chr>          <int>  <int>
##  1 Albany      35190317 365511
##  2 Allegany     5044626  96030
##  3 Broome      26542338 357555
##  4 Cattaraugus  8641641 124948
##  5 Cayuga       9209962  94541
##  6 Chautauqua  13336983 173552
##  7 Chemung     12032234 145648
##  8 Chenango     5225899  75885
##  9 Clinton      8943469  48873
## 10 Columbia     5970468 104193
## # ... with 49 more rows
```

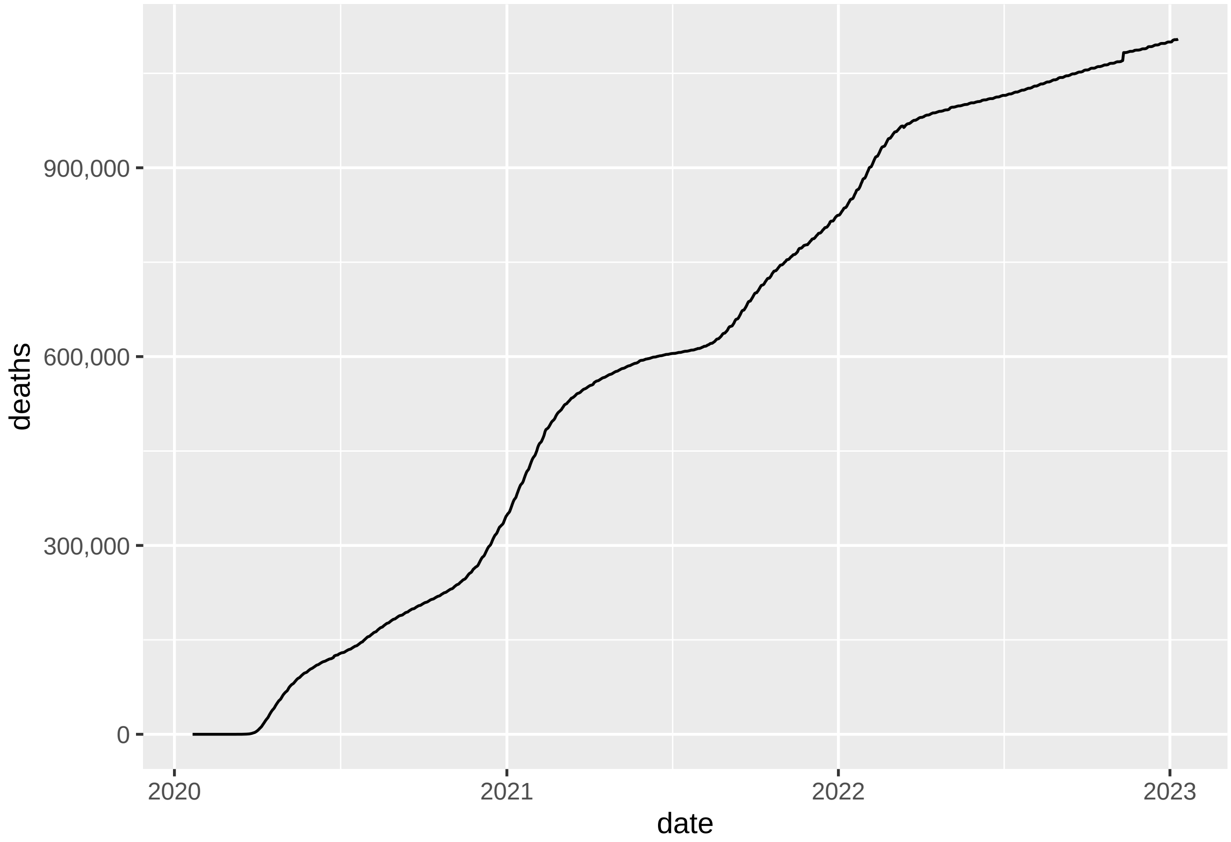
Exploratory Data Analysis

The pre-processed data was analyzed to calculate the statistical distribution of the cases and deaths by the groups. Firstly, the cases and deaths in the US from the beginning were visualized.

```
ggplot(us, aes(x=date, y=cases)) +
  geom_line() +
  scale_y_continuous(labels = comma)
```



```
ggplot(us, aes(x=date, y=deaths)) +
  geom_line() +
  scale_y_continuous(labels = comma)
```



The descriptive statistics was calculated for the tables.

```
print("Total daily cases in US")
```

```
## [1] "Total daily cases in US"
```

```
summary(us)
```

```
##       date                cases               deaths              yr           
##  Min.   :2020-01-21   Min.   :        1   Min.   :      0   Length:1086       
##  1st Qu.:2020-10-18   1st Qu.:  8230724   1st Qu.: 219670   Class :character  
##  Median :2021-07-16   Median : 34102714   Median : 608502   Mode  :character  
##  Mean   :2021-07-16   Mean   : 43945776   Mean   : 597470                     
##  3rd Qu.:2022-04-13   3rd Qu.: 80451740   3rd Qu.: 986294                     
##  Max.   :2023-01-10   Max.   :101091495   Max.   :1104459                     
##      mnt                 dt             year_month       
##  Length:1086        Length:1086        Length:1086       
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
## 
## 
## 
```

```
print("Total daily cases in 2020")
```

```
## [1] "Total daily cases in 2020"
```

```
summary(us2020)
```

```
##       date               county             state                fips      
##  Min.   :2020-01-21   Length:884737      Length:884737      Min.   : 1001  
##  1st Qu.:2020-06-08   Class  :character  Class  :character  1st Qu.:18183  
##  Median :2020-08-17   Mode  :character   Mode  :character   Median :29215  
##  Mean   :2020-08-15                                         Mean   :31262  
##  3rd Qu.:2020-10-24                                         3rd Qu.:46099  
##  Max.   :2020-12-31                                         Max.   :78030  
##                                                             NA's   :8266   
##      cases            deaths             yr                mnt           
##  Min.   :     0   Min.   :    0.0   Length:884737      Length:884737     
##  1st Qu.:    36   1st Qu.:    0.0   Class  :character  Class  :character 
##  Median :   228   Median :    4.0   Mode  :character   Mode  :character  
##  Mean   :  1952   Mean   :   53.6                                        
##  3rd Qu.:   993   3rd Qu.:   21.0                                        
##  Max.   :770915   Max.   :25144.0                                        
##                   NA's   :18761                                          
##       dt             year_month       
##  Length:884737      Length:884737     
##  Class  :character  Class  :character 
##  Mode  :character   Mode  :character  
##                                       
##                                       
##                                       
##                                       
```

```
print("Total daily cases in 2021")
```

```
## [1] "Total daily cases in 2021"
```

```
summary(us2021)
```

```
##       date               county             state                fips      
##  Min.   :2021-01-01   Length:1185373     Length:1185373     Min.   : 1001  
##  1st Qu.:2021-04-02   Class  :character  Class  :character  1st Qu.:19035  
##  Median :2021-07-02   Mode  :character   Mode  :character   Median :30026  
##  Mean   :2021-07-02                                         Mean   :31472  
##  3rd Qu.:2021-10-01                                         3rd Qu.:46119  
##  Max.   :2021-12-31                                         Max.   :78030  
##                                                             NA's   :10803  
##      cases             deaths             yr                mnt           
##  Min.   :      0   Min.   :    0.0   Length:1185373     Length:1185373    
##  1st Qu.:   1136   1st Qu.:   20.0   Class  :character  Class  :character 
##  Median :   2778   Median :   52.0   Mode  :character   Mode  :character  
##  Mean   :  11160   Mean   :  193.6                                        
##  3rd Qu.:   7340   3rd Qu.:  125.0                                        
##  Max.   :1697286   Max.   :35382.0
##                    NA's   :28470
##       dt             year_month
##  Length:1185373     Length:1185373
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
print("Total daily cases in 2022")
```

```
## [1] "Total daily cases in 2022"
```

```
summary(us2022)
```

```
##       date               county             state                fips
##  Min.   :2022-01-01   Length:1188042     Length:1188042     Min.   : 1001
##  1st Qu.:2022-04-02   Class :character   Class :character   1st Qu.:19035
##  Median :2022-07-02   Mode  :character   Mode  :character   Median :30027
##  Mean   :2022-07-02                                         Mean   :31483
##  3rd Qu.:2022-10-01                                         3rd Qu.:46121
##  Max.   :2022-12-31                                         Max.   :78030
##                                                             NA's   :13101
##      cases             deaths              yr                 mnt
##  Min.   :      0   Min.   :    0.0   Length:1188042     Length:1188042
##  1st Qu.:   2707   1st Qu.:   41.0   Class :character   Class :character
##  Median :   6723   Median :   99.0   Mode  :character   Mode  :character
##  Mean   :  26733   Mean   :  316.9
##  3rd Qu.:  17622   3rd Qu.:  235.0
##  Max.   :3632440   Max.   :43935.0
##                    NA's   :28470
##       dt             year_month
##  Length:1188042     Length:1188042
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
print("Total daily cases in 2023")
```

```
## [1] "Total daily cases in 2023"
```

```
summary(us2023)
```

```
##       date               county             state                fips
##  Min.   :2023-01-01   Length:32533       Length:32533       Min.   : 1001
##  1st Qu.:2023-01-03   Class :character   Class :character   1st Qu.:19035
##  Median :2023-01-05   Mode  :character   Mode  :character   Median :30027
##  Mean   :2023-01-05                                         Mean   :31486
##  3rd Qu.:2023-01-08                                         3rd Qu.:46121
##  Max.   :2023-01-10                                         Max.   :78030
##                                                             NA's   :338
##      cases             deaths             yr                 mnt
##  Min.   :      0   Min.   :    0.0   Length:32533       Length:32533
##  1st Qu.:   3112   1st Qu.:   47.0   Class :character   Class :character
##  Median :   7835   Median :  110.0   Mode  :character   Mode  :character
##  Mean   :  30995   Mean   :  347.2
##  3rd Qu.:  20486   3rd Qu.:  257.0
##  Max.   :3654167   Max.   :44178.0
##                    NA's   :780
##       dt              year_month
##  Length:32533       Length:32533
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
print("Total daily cases difference by the year")
```

```
## [1] "Total daily cases difference by the year"
```

```
summary(us_incidences_tb)
```

```
##      mnt               cases_2020          cases_2021        cases_2020_2020
##  Length:12          Min.   :       41   Min.   :7.300e+08   Min.   :2.038e+09
##  Class :character   1st Qu.: 14982664   1st Qu.:9.380e+08   1st Qu.:2.453e+09
##  Mode  :character   Median : 88457490   Median :1.040e+09   Median :2.673e+09
##                     Mean   :143940769   Mean   :1.102e+09   Mean   :2.647e+09
##                     3rd Qu.:213017618   3rd Qu.:1.283e+09   3rd Qu.:2.894e+09
##                     Max.   :525970605   Max.   :1.576e+09   Max.   :3.086e+09
##
##    cases_2023
##  Min.   :1.008e+09
##  1st Qu.:1.008e+09
##  Median :1.008e+09
##  Mean   :1.008e+09
##  3rd Qu.:1.008e+09
##  Max.   :1.008e+09
##  NA's   :11
```

```
print("Distribution of county cases")
```

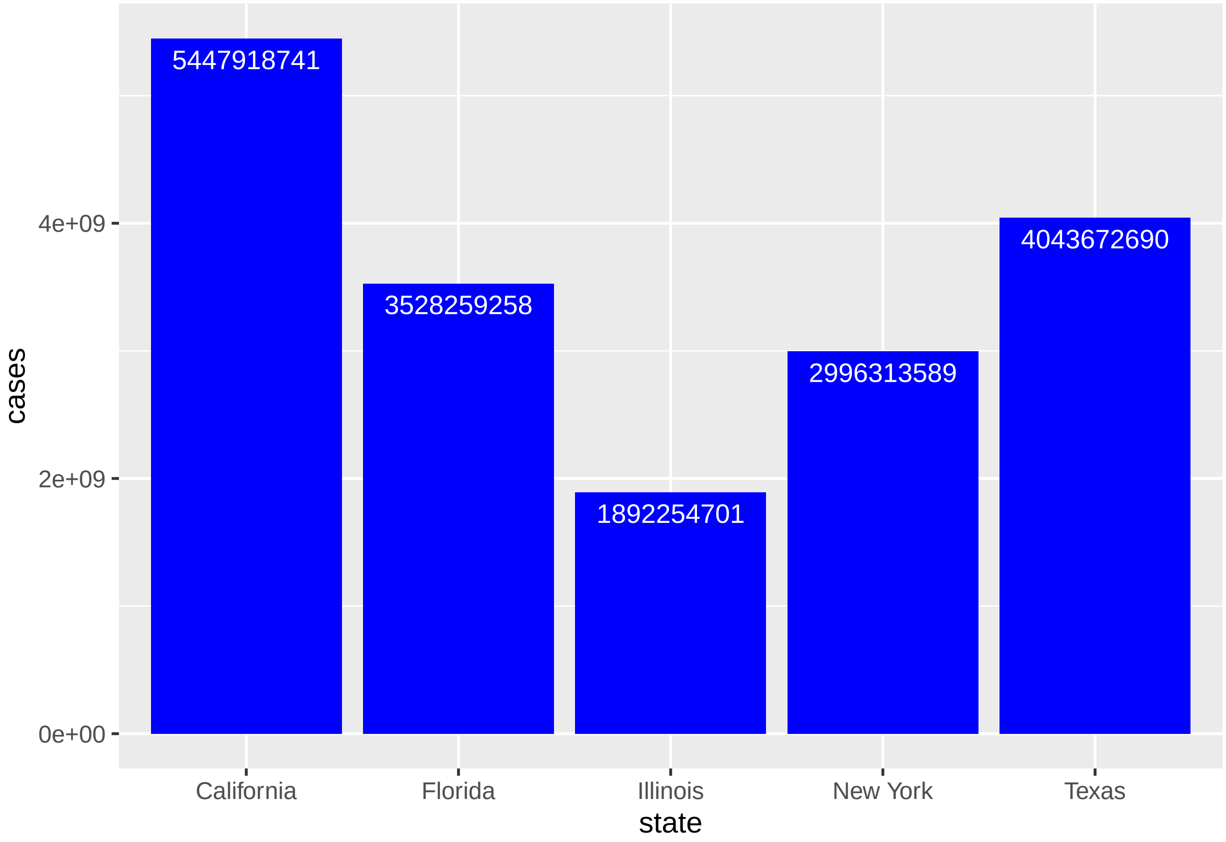
```
## [1] "Distribution of county cases"
```

```
summary(us_NY_counties)
```

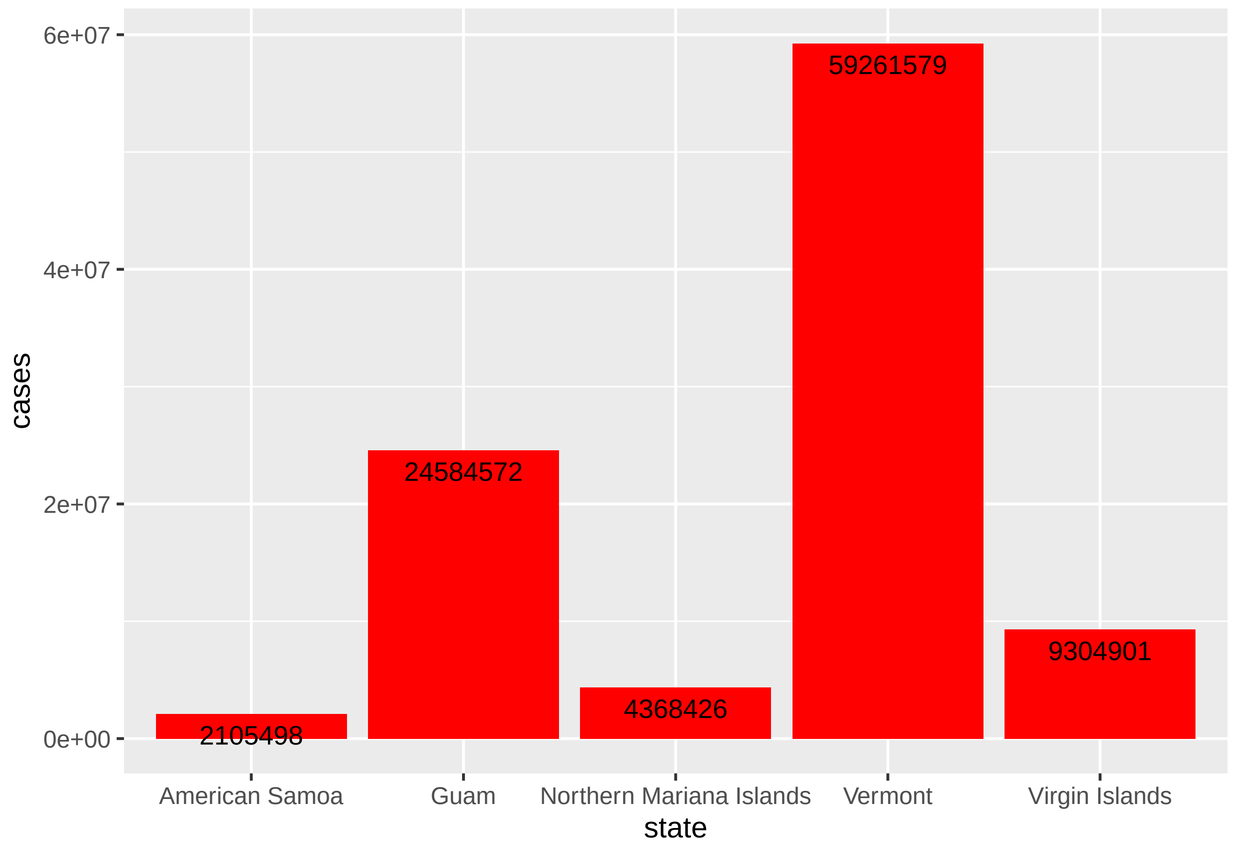
```
##      county              cases              deaths
##  Length:59          Min.   :0.000e+00   Min.   :    2822
##  Class :character   1st Qu.:5.398e+06   1st Qu.:   67544
##  Mode  :character   Median :9.210e+06   Median :  104193
##                     Mean   :5.078e+07   Mean   :  904379
##                     3rd Qu.:2.255e+07   3rd Qu.:  228599
##                     Max.   :1.389e+09   Max.   :33138240
```

The states in the US were ranked according to the sum of cases and deaths. Likewise, the New York counties were then ranked according to the sums of cases and deaths.

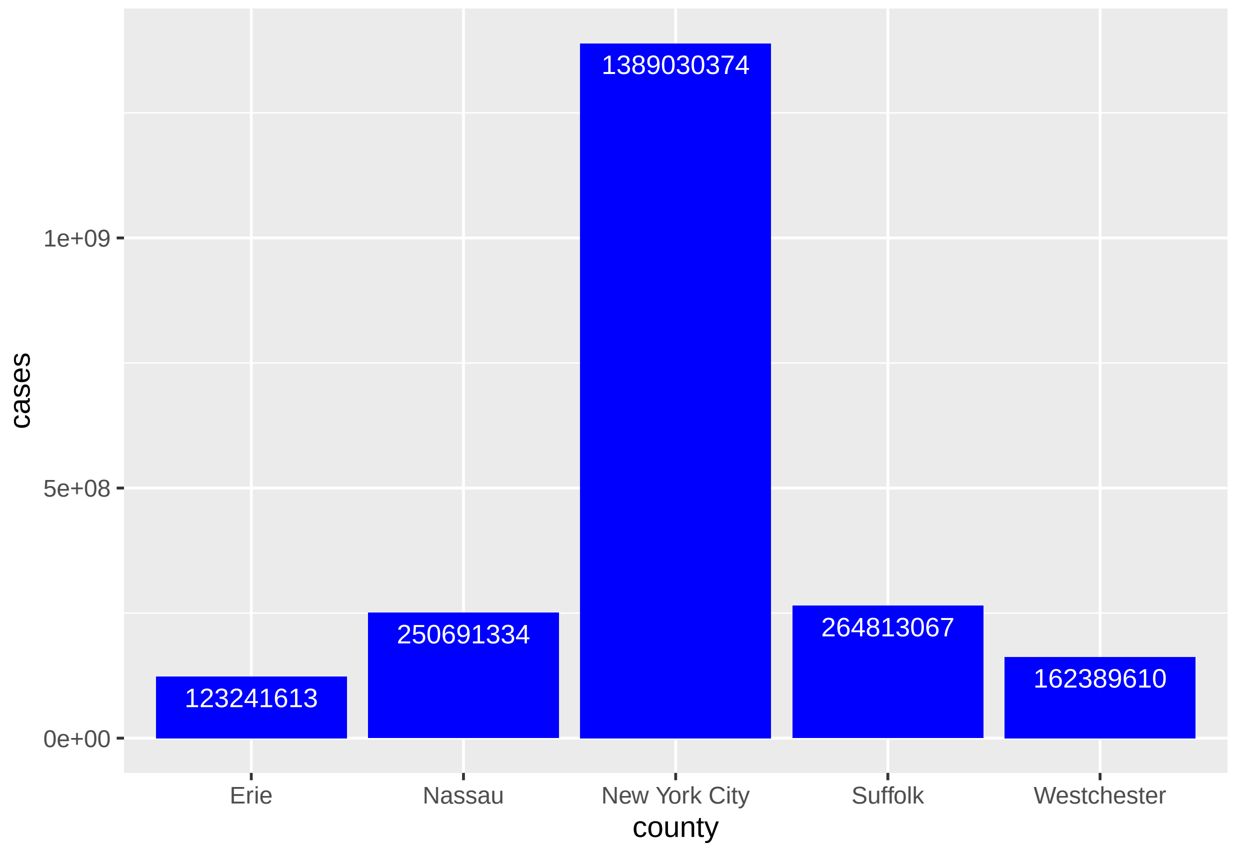
```
# top 5 counties according to number of cases
us_states_top5 <-  us_yearly_state %>%
  group_by(state) %>%
  dplyr::summarise(cases = sum(cases)) %>%
  arrange(desc(cases)) %>%
  slice(1:5)
ggplot(us_states_top5, aes(x=state, y=cases)) +  geom_bar(stat="identity", fill="blue") + geom_text(aes
```



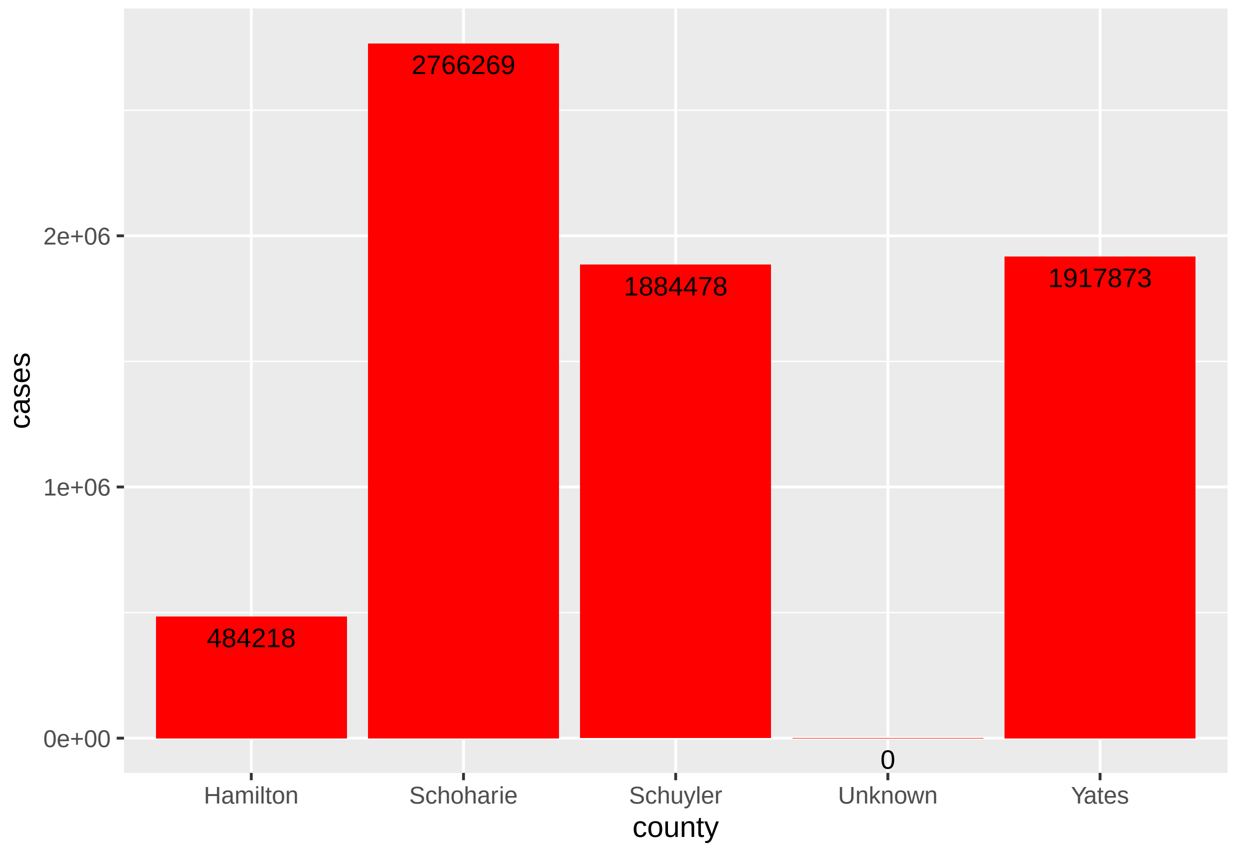
```
# last 5 counties according to number of cases
us_states_last5 <-  us_yearly_state %>%
  group_by(state) %>%
  dplyr::summarise(cases = sum(cases)) %>%
  arrange(cases) %>%
  slice(1:5)
ggplot(us_states_last5, aes(x=state, y=cases)) +  geom_bar(stat="identity", fill="red") + geom_text(aes
```



```
# top 5 counties according to number of cases
us_NY_counties_top5 <-  us_NY_counties %>%
  arrange(desc(cases)) %>%
  slice(1:5)
# us_NY_counties_top5
ggplot(us_NY_counties_top5, aes(x=county, y=cases)) +  geom_bar(stat="identity", fill="blue") + geom_te
```



```
# last 5 counties according to number of cases
us_NY_counties_last5 <-  us_NY_counties %>%
  arrange(cases) %>%
  slice(1:5)
# us_NY_counties_last5
ggplot(us_NY_counties_last5, aes(x=county, y=cases)) +  geom_bar(stat="identity", fill="red") + geom_te
```



Summary

After analyzing the COVID-19 pandemic data for the US, more details about the cases and deaths was understood. Firstly, the trend was examined over time and to see what are the frequent times of the year for most incidences. The line graph clearly showed that the highest jumps were during the new year time. The holiday season during the end of the year appeared to have the highest number of cases within the timeline. Looking back at the table comparing the number of cases within each month between the years, the largest difference occurred during New Year's time of 2021. The New Year's of 2020 was the second largest and it was only half. The trends become more linear after the holiday season. This last holiday season appeared better than previous. The descriptive statistics was also calculated and it showed that the cases and deaths increased exponentially over the years. The states in the US with the highest sum of cases were the following ranked from the top: California, Texas, Florida, New York, and Illinois. Starting from the least sums, they were American Samoa, Northern Marinara Islands, Virgun Islands, Guam, and Vermont. The data was then filtered to New York state and the 5 counties with the most cases and least cases were examined. Counties with the most cases were New York City, Suffolk, Nassau, Westchester, and Erie, ranked descending. The least cases occurred in Hamilton, Schuyler, Yates, and Schoharie, ranked ascending.

Thank You!