US COVID

04/12/2021

INTRODUCTION

Hey welcome to my Covid 19 report and analysis. COVID-19 is a dangerous virus that has killed millions of people all over the world. This is a very serious pandemic and it is still going on right now. This report will be focusing on Covid-19 in the United States. The data that I will be using for my report can be found on github, the link is provided here : "https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series". This is a time series data set that has many attributes. This is a csv dataset that is updated daily regarding the number of cases, the number of deaths and the number of recovery. This data is from Johns Hopkins University. Thanks to them we are able to do our own reports and analysis.

What I will be focusing on in my reports is the correlation between the number of cases and the number of deaths regarding Covid-19 in the United States. I will also be looking in the state of Califorina to see how the cases are matched up against compared to the cases in New York.

DATA PREPERATION

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
confirmed_link <- c("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data
confirmed_df <- read_csv(confirmed_link)
```

```
## Rows: 3342 Columns: 693
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (687): 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.
```

```
tail(confirmed_df)
```

```
## # A tibble: 6 x 693
##        UID iso2  iso3  code3  FIPS Admin2    Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>     <chr>          <chr>          <dbl>
## 1 84056037 US    USA     840 56037 Sweetwat~ Wyoming        US              41.7
## 2 84056039 US    USA     840 56039 Teton     Wyoming        US              43.9
## 3 84056041 US    USA     840 56041 Uinta     Wyoming        US              41.3
## 4 84090056 US    USA     840 90056 Unassign~ Wyoming        US               0
## 5 84056043 US    USA     840 56043 Washakie  Wyoming        US              43.9
## 6 84056045 US    USA     840 56045 Weston    Wyoming        US              43.8
## # ... with 684 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>, 2/16/20 <dbl>,
## #   2/17/20 <dbl>, 2/18/20 <dbl>, 2/19/20 <dbl>, 2/20/20 <dbl>, ...
```

```
us_death <- c("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse
us_deathdf <- read_csv(us_death)
```

```
## Rows: 3342 Columns: 694
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (688): 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.
```

```
us_cases <- confirmed_df %>% pivot_longer(cols = -(UID:Combined_Key),names_to = "date",values_to = "cas
us_deaths <- us_deathdf %>% pivot_longer(cols = -(UID:Population),names_to = "date",values_to = "deaths
us_df<-us_cases %>% full_join(us_deaths)
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key", "date")
```

```
us_by_state <- us_df %>% group_by(Province_State,Country_Region,date) %>% summarize(cases = sum(cases),
```

```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can override using the `
```

```
us_totals <- us_by_state %>% group_by(Country_Region,date) %>% summarize(cases = sum(cases), deaths=sum
```

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

```
us_state_totals <- us_by_state %>% group_by(Province_State)%>% summarize(deaths=max(deaths),cases=max(ca
```

Graph Preperation

```
p1 <- ggplot(us_totals, aes(date)) +
geom_line(aes(y=cases), colour="red") +
ggtitle("Covid-19 in US")
state <- "California"
cali <- us_by_state %>% filter(Province_State== state) %>% ggplot(aes(x=date,y=cases))+geom_line(aes(col
state2 <- "New York"
ny <-  us_by_state %>% filter(Province_State== state2) %>% ggplot(aes(x=date,y=cases))+geom_line(aes(col
mod <- lm(deaths_per_thou ~ cases_per_thou,data=us_state_totals)
pred_ <- us_state_totals %>% mutate(pred=predict(mod))
model_graph <- pred_ %>% ggplot() +geom_point(aes(x=cases_per_thou,y=deaths_per_thou),color="green")+ g
```

GRAPHS

Here are some basic visualization from the state of California.

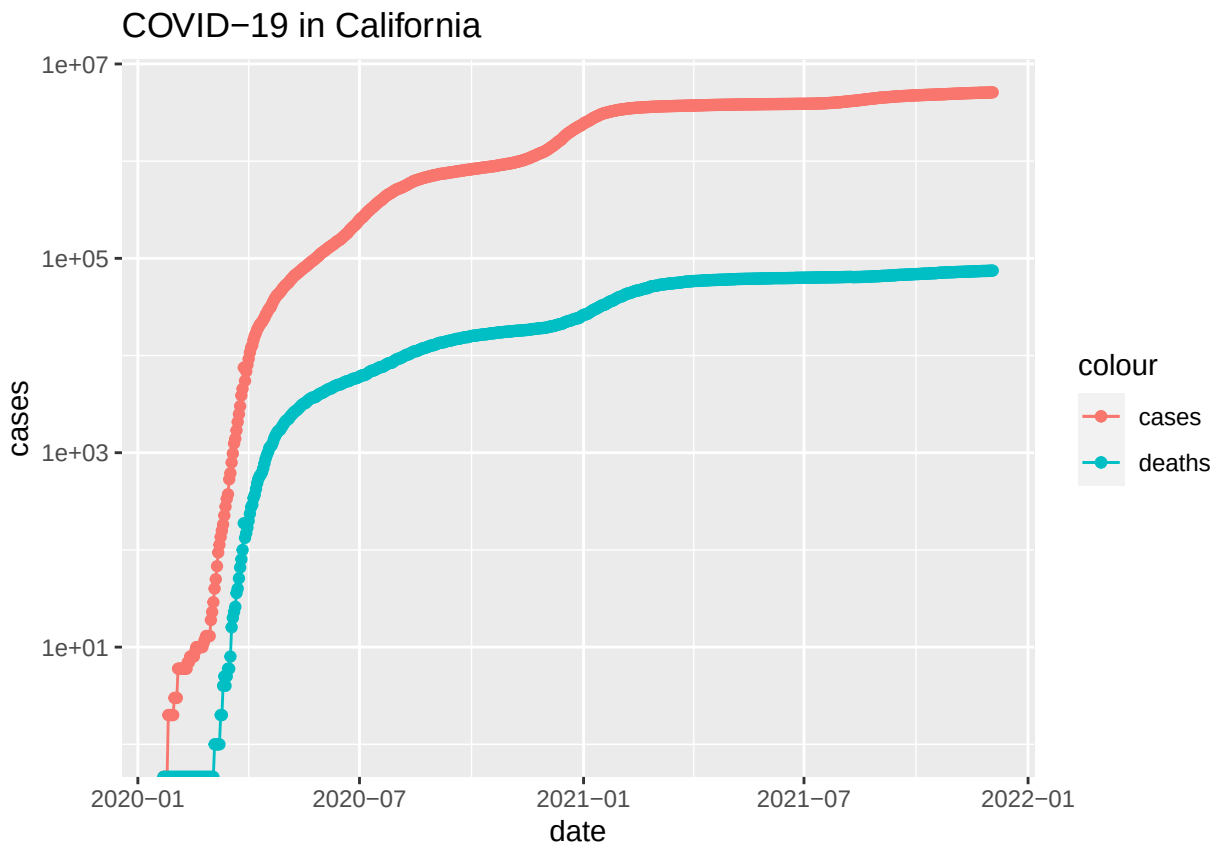
```
plot(cali)
```

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

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

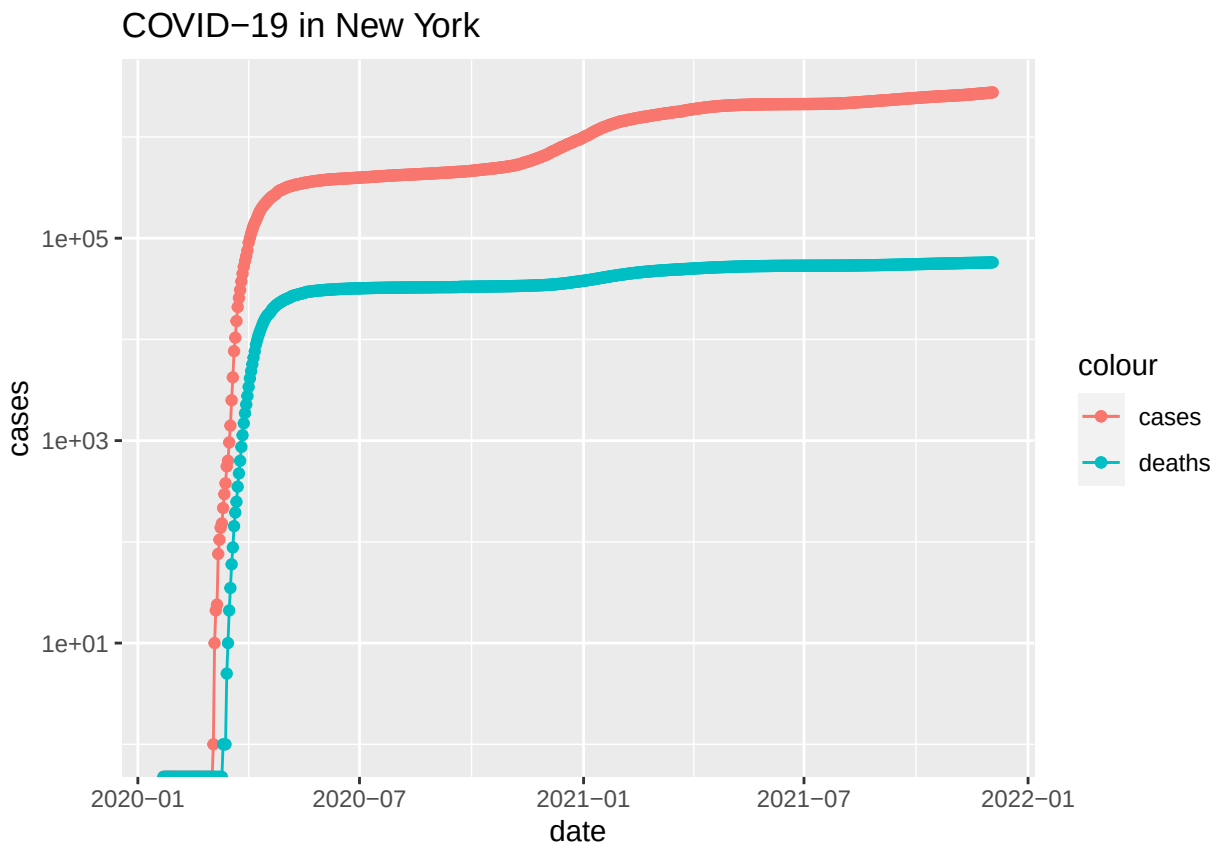
```
## Warning: Transformation introduced infinite values in continuous y-axis
```

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




```
plot(ny)
```

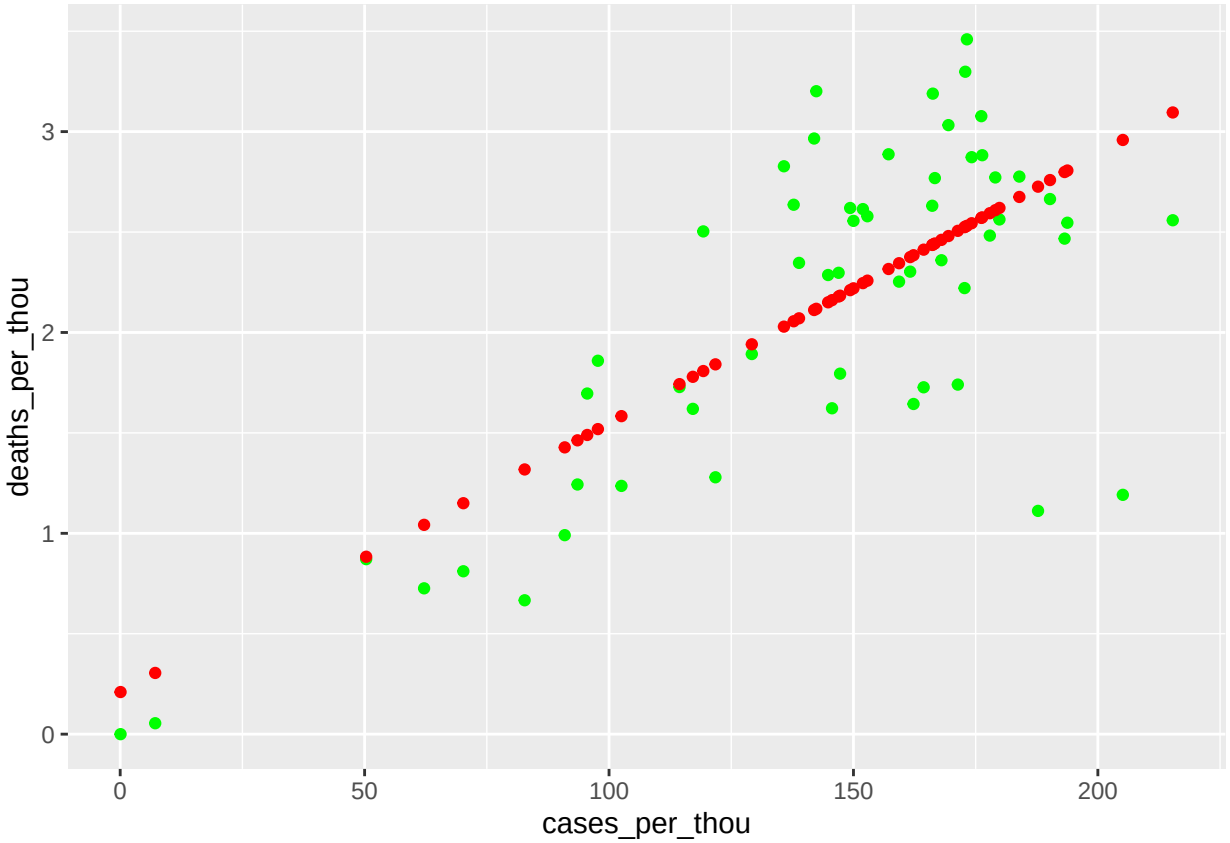
```
## Warning: Transformation introduced infinite values in continuous y-axis

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

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

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



MODELS

Here is a Linear regression model of cases per thousand versus deaths per thousands from all the states in United States.

```
plot(model_graph)
```



CONCLUSION AND BIAS

So from the visualization you can tell when cases goes up the death also goes up so there is a correlation between them both. Some bias that are applicable in this reports is that not all cases and deaths willbe reported. Not all cases are covid related but maybe a different type of virus and all the deaths are not covid deaths. Since covid is a new virus and there are not information or reaserach done on this virus its hard to tell if the cases and deaths were Coivd-19 related.