Covid-19 Trend Analysis and Visualization

https://github.com/xueying-F/ENV872-Project_XY_Feng

Xueying Feng

Contents

List of Tables

List of Figures

1 Rationale and Research Questions

China broke out a global disease, COVID-19, in December 2019. It has been confirmed that it is a new type of virus called 2019 new coronavirus (2019-nCoV) in Jan 2020. The World Health Organization (WHO) announced the official new name of the disease caused by nCoV2019 (2019 novel coronavirus) in Feb 2020. The CDC noted that the symptoms of the new coronavirus include "symptoms of fever and lower respiratory tract disease (eg. coughing, difficulty breathing). There is currently no vaccine, and because this is a virus, antibiotics will not work.

COVID-19 has spread worldwide in just four months, and more than 2.7 million people have been sick. It is a global pandemic disease announced by the World Health Organization (WHO). Also, it can easily infect humans, and can spread from person to person in a rapid and sustained manner. This research focus on the datasets from China, United States, and whole global countries to see the trend of confirmed cases, death cases and heal cases. Moreover, I will combine some news and report into policies in each country to analyze the trend patterns.

The research question is:
(1) Is there a same trend pattern on China and United States?
(2) What is trend pattern on global level?
(3) Why trend patterns look like that?

Downloading datasets from R package, nCov2019 (https://github.com/GuangchuangYu/nCov2019), which provides convenient access to epidemiological data on the coronavirus outbreak, which contains real-time data and historical data for each country. Please see the detail information in Dataset Information setion.

2 Dataset Information

2.1 Database Information

I access all Novel Coronavirus data from the R package, nCov2019, which includes detailed real-time statistics, historical data in all countries, and down to the city-level.

More information can be found here:https://github.com/GuangchuangYu/nCov2019.

2.2 Data Content Information

I pulled all relevent data from nCov2019 package. I named all pulled data as raw.csv and then saved into Data/Raw folder. Beside, I wrangled all raw and selected the relevent columns, and then saved into Data/Processed folder. Because data pulled from nCov2019 package are real-time data, I kept the update time on April 16th, 2020. Therefore, the following descriptions of my datasets are for processed data.

Table 1: Historical_Global_processed dataset

Column name	Description
time	Date
country	Country name
cum_confirm	Cumulative number of COVID-19 confirmed cases
cum_heal	Cumulative number of COVID-19 heal cases
cum_dead	Cumulative number of COVID-19 death cases

Table 2: Historical_China_processed dataset

Column name	Description
time	Date
country	Country name
province	Province name
city	City name
cum_confirm	Cumulative number of COVID-19 confirmed cases
cum_heal	Cumulative number of COVID-19 heal cases
cum_dead	Cumulative number of COVID-19 death cases

Table 3: Covid19_Global_processed dataset

Column name	Description
name	Country name
confirm	Number of COVID-19 confirmed cases
dead	Number of COVID-19 death cases

Column name	Description
deadRate	Total death number/ Cumulative total number of COVID-19 cases(%)
heal	number of COVID-19 heal cases
healRate	Total heal number/ Cumulative total number of COVID-19 cases(%)

Table 4: Covid19_China_processed dataset

Column name	Description
name	Province name
confirm	Number of COVID-19 confirmed cases
dead	Number of COVID-19 death cases
deadRate	Total death number/ Cumulative total number of COVID-19 cases(%)
heal	number of COVID-19 heal cases
healRate	Total heal number/ Cumulative total number of COVID-19 cases(%)

Table 5: Covid19_US_processed dataset

Column name	Description	NA
time	Date	NA
country	United States	NA
province	States name	NA
cum_confirm	Cumulative number of COVID-19 confirmed cases	NA
cum_heal	Cumulative number of COVID-19 heal cases	NA
cum_dead	Cumulative number of COVID-19 death cases	NA

3 Exploratory Analysis

Wrangling the raw data for thess datasets is to select the columns that are useful for this research. After wrangling all dataset, they are saved into processed folder. However, this section will show how processed data form, but processed data will be directly used to do analysis. During analysis, datasets will be wrangled again based on anlysis requirments.

```
#Covid19_China_raw <- read.csv("Covid19_China_raw.csv",
#                      header=TRUE,stringsAsFactors = FALSE, strip.white = TRUE,sep = '
#Covid19_China_processed <-
#  select(Covid19_China_raw, name, confirm, dead, deadRate, heal, healRate)


#Covid19_US_raw <- read.csv("Covid19_US_raw.csv", header=TRUE,
#                   stringsAsFactors = FALSE, strip.white = TRUE,sep = ',')
#Covid19_US_processed <-
#  select(Covid19_US_raw, time:cum_dead)


#Covid19_Global_raw <- read.csv("Covid19_Global_raw.csv",
#                       header=TRUE,stringsAsFactors = FALSE, strip.white = TRUE,sep =
#Covid19_Global_processed <-
#  select(Covid19_Global_raw, name, confirm, dead, deadRate,heal, healRate)


#historical_China_raw <- read.csv("historical_China_raw.csv",
#                         header=TRUE,stringsAsFactors = FALSE, strip.white = TRUE,sep
#Historical_China_processed <-
#  select(historical_China_raw, time:cum_dead)


#historical_Global_raw <- read.csv("historical_Global_raw.csv", header=TRUE,
#                                  stringsAsFactors = FALSE, strip.white = TRUE,sep =
#Historical_Global_processed <- historical_Global_raw
```

3.1 Explore China Dataset

```
Covid19_China_processed <- read.csv("./Data/Processed/Covid19_China_processed.csv")

# Check data frame
colnames(Covid19_China_processed)
head(Covid19_China_processed)
```

```
# scale_y_log10: transform the y-axis to make it easier to read
China.plot <-ggplot(Covid19_China_processed,aes(x = name,y = confirm)) +
```

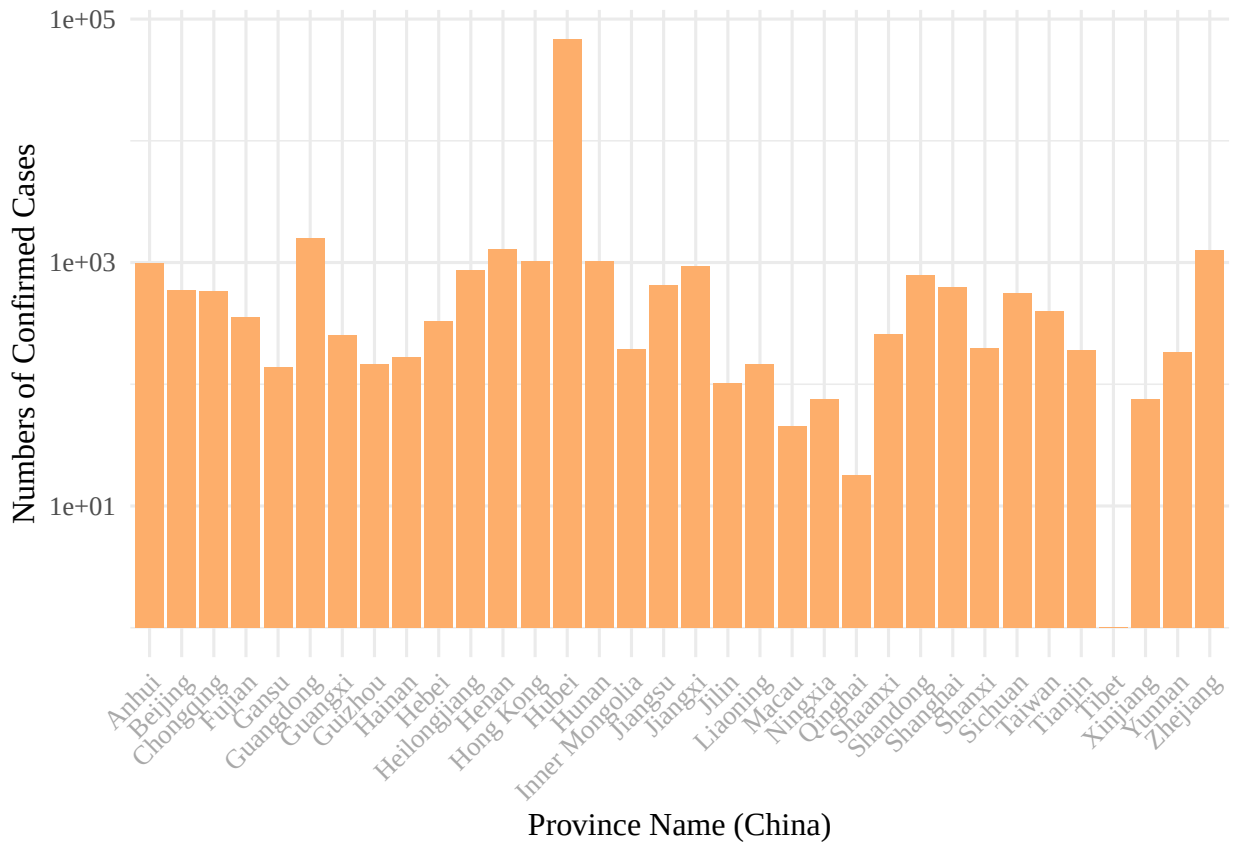

Figure 1: Compare total confimed number in each province in China

```
  geom_bar(size = 2, stat = "identity",position = "dodge", fill ="#fdae6b") +
  scale_y_log10() +
  labs(x= "Province Name (China)",
       y = "Numbers of Confirmed Cases") +
  mytheme +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(China.plot)
```

3.2 Explore Global Dataset

Top 15 countries

```
Covid19_Global_processed <- read.csv("./Data/Processed/Covid19_Global_processed.csv")

# Check data frame
head(Historical_Global_processed)
str(Historical_Global_processed)
```

```r
# Change date column to date
Historical_Global_processed$time <- as.Date(Historical_Global_processed$time,
                                            format = "%Y-%m-%d")
class(Historical_Global_processed$time)


# Filter out top 15 countries with the highest number of diagnoses
Fifty_countries <- Covid19_Global_processed %>%
  top_n(15, confirm) %>%
  arrange(desc(confirm))

# Save as csv
write.csv(Fifty_countries,
          file = "./Data/Processed/Top_Fifty_countries (cum).csv", row.names=FALSE)
```

```r
# Draw histogram plot
Global.plot <-ggplot(Fifty_countries,aes(x = name,y = confirm)) +
  geom_bar(size = 2, stat = "identity",position = "dodge", fill ="#fdae6b") +
  labs(x = " Country Name (Top 15)",
       y = "Numbers of Confirmed Cases") +
  mytheme +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

print(Global.plot)
```

3.3 Explore United States Dataset

Map of comfimed cases distribution in United State

```r
# Check data
head(Covid19_US_processed)
str(Covid19_US_processed)

# Change date column to date
Covid19_US_processed$time <- as.Date(Covid19_US_processed$time, format = "%m/%d/%y")
class(Covid19_US_processed$time)

# Check column names
colnames(Covid19_US_processed)

# Rename column where names is "province"
names(Covid19_US_processed)[names(Covid19_US_processed) == "province"] <- "state"
str(Covid19_US_processed)
```

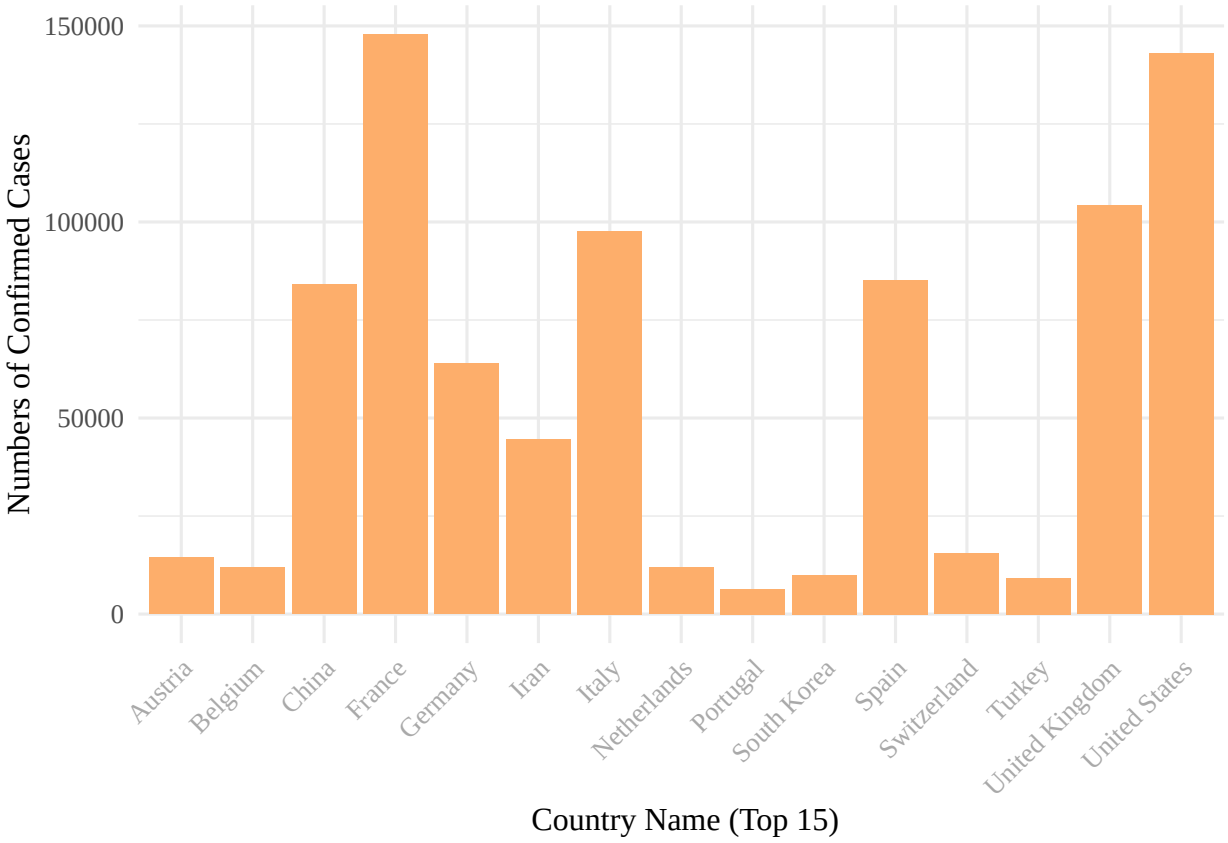

Figure 2: Compare total confimed number of each country

```r
# Select total confirmed cases in 51 States
US_States <- Covid19_US_processed %>%
  filter(time == as.Date("2020-04-16")) %>%
  filter(state!= "Unpublished sources") %>%
  top_n(51, cum_confirm)

# Convert state column to characters
US_States$state <- as.character(US_States$state)

# Rename specific cell name
US_States$state[US_States$state == 'New York state'] <- 'New York'
US_States$state[US_States$state == "Washington State"] <- "Washington"
US_States$state[US_States$state == 'the state of Wisconsin'] <- 'Wisconsin'
```

```r
#install.packages("usmap")
library(usmap)

USMap_Base <- plot_usmap(regions = "counties") +
  labs(title = "US Counties",
       subtitle = "This is a blank map of the counties of the United States.") +
  theme(panel.background = element_rect(color = "black", fill = "lightblue"))


### Draw US Case distribution on map
US.map <- plot_usmap(regions="state", data = US_States,
                                values = "cum_confirm", color = "black") +
  scale_fill_continuous(low = "#fee6ce", high = "#d95f0e",
                        name = "Total confirmed cases", label = scales::comma) +
  labs(title = "State Reporting Cases of Covid 19") +
  theme(legend.position = "right")

print(US.map)
```

Figure 1 shows the total confimed number of each province in China (update on April 16th, 2020). Due to this graph, I will seperate Hubei province and other provinces in China and see their trend patterns.

Figure 2 shows the total confimed number of each country (update on April 16th, 2020). Due to this graph, I will focus on the United States dataset and see its trend pattern.

Figure 3 shows confirmed cases distribution in each state by using map, which can make readers straighter forward to see which states are more serious.

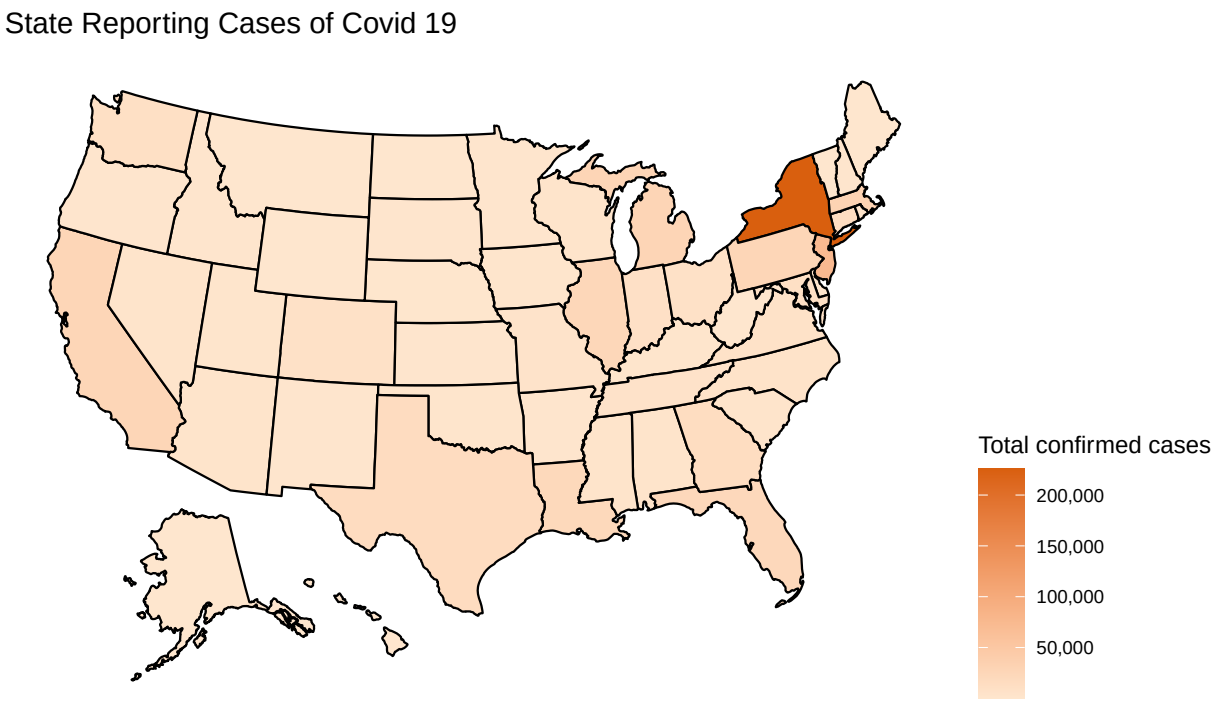

Figure 3: Map of comfimed cases didtribution in United State

4 Analysis

4.1 Anlysis trend of confirmed cases in China

4.1.1 Trend of Total confirmed cases in China

```
# Selsct data from China
str(Historical_Global_processed)
```

```
## 'data.frame':    8687 obs. of  5 variables:
##  $ time       : Date, format: "2019-12-01" "2019-12-02" ...
##  $ country    : Factor w/ 205 levels "Afghanistan",..: 39 39 39 39 39 39 39 39 39 39
##  $ cum_confirm: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cum_heal   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ cum_dead   : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
Historical_China = Historical_Global_processed[Historical_Global_processed$country == 'C
head(Historical_China)
```

```
##         time country cum_confirm cum_heal cum_dead
## 1 2019-12-01   China           1        0        0
## 2 2019-12-02   China           1        0        0
## 3 2019-12-03   China           1        0        0
## 4 2019-12-04   China           1        0        0
## 5 2019-12-05   China           1        0        0
## 6 2019-12-06   China           1        0        0
```

```
# Create a ggplot depicting cases incereasing over time
China_Total_Trend.plot <-
  ggplot(Historical_China, aes(x=time, y=cum_confirm)) +
  geom_point(colour = "#e6550d") +
  geom_line(colour = "#d95f0e") +
  labs(x = "Time",
       y = "Confirmed Cases") +
  scale_x_date(date_labels = "%Y-%m-%d")

print(China_Total_Trend.plot)
```

4.1.2 Trend of confirmed cases in each province in China

```
# Check data
head(Historical_China_processed)
str(Historical_China_processed)

# Change date column to date
Historical_China_processed$time <- as.Date(Historical_China_processed$time,
```

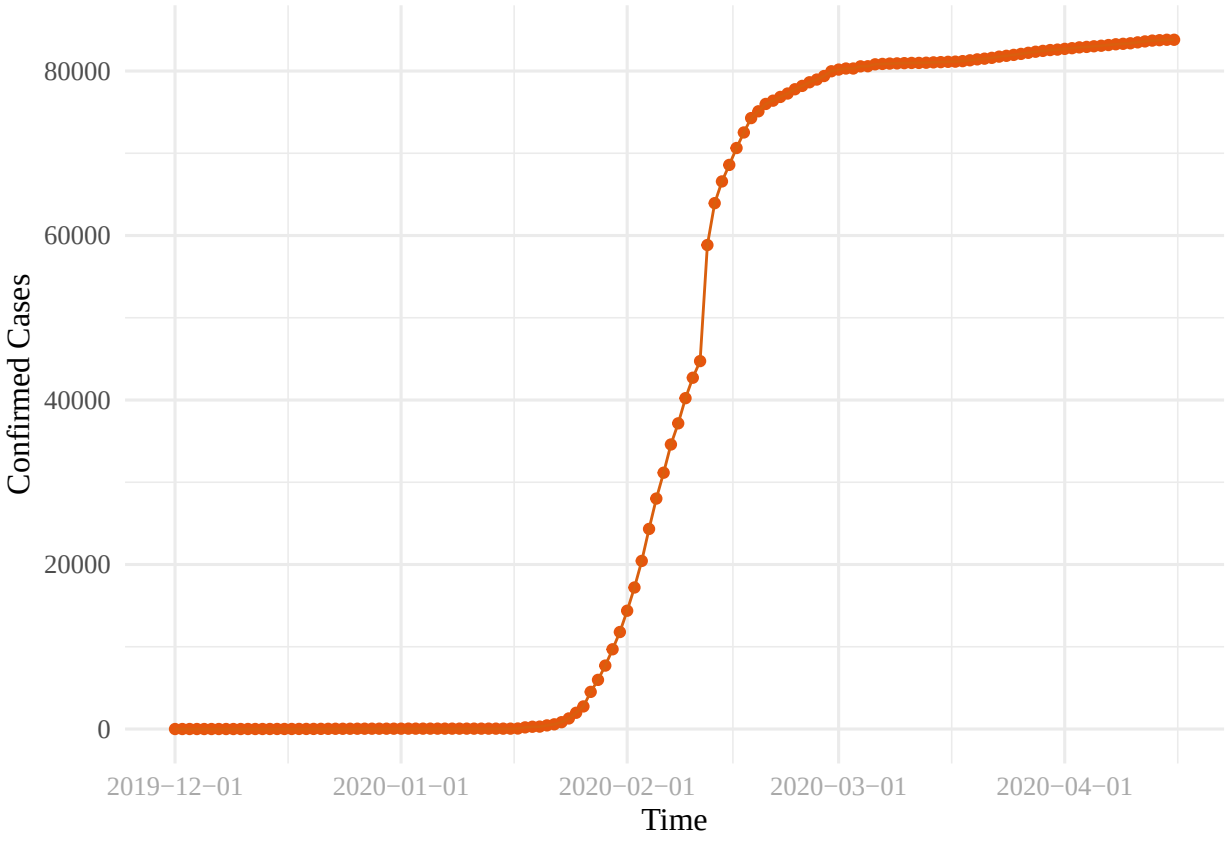

Figure 4: Trend of confirmed case in China

```
                                              format = "%m/%d/%y")
head(Historical_China_processed)
class(Historical_China_processed$time)
```

4.1.3 Trend of confirmed cases in Hubei, China

```
# Select Hubei province data
Historical_Hubei = Historical_China_processed[Historical_China_processed$province == 'Hu
head(Historical_Hubei)

# Group by Hubei cities
Historical_HubeiCityV2 <- Historical_Hubei %>%
  select(time, city:cum_dead)

# Save as csv
write.csv(Historical_HubeiCityV2,
          file = "./Data/Processed/Historical_HubeiCity.csv",row.names=FALSE)
```

```
# Create a ggplot depicting cases incereasing over time
HubeiCity_Trend.plot <- ggplot(Historical_HubeiCityV2,
                               aes(x=time, y=cum_confirm, color=city)) +
    geom_line(alpha = 0.95, size = 0.5) +
    geom_text_repel(aes(label=city),
                    function(Historical_HubeiCityV2)
                      Historical_HubeiCityV2[Historical_HubeiCityV2$time == as.Date("202
    mytheme +
    theme(legend.position = "none") +
    labs(x=expression(paste("Time"))) +
    labs(y=expression(paste("Comfimed Cases")))+
    labs(color="city") +
    scale_x_date(date_labels = "%Y-%m-%d")

print(HubeiCity_Trend.plot)
```

Due to the confimred cases in Hubei are almost 80% of total confimred cases in China. I draw the trend plots for total confimred cases in China and in each Hubei city, which shows the pattern are very similar between Wuhan trend and China Trend.

4.1.4 Trend of confirmed cases in Hubei cities, China (except Wuhan)

Most cases are from Wuhan, so I deleted data from Wuhan City, and see the if the other cities have same patterns.

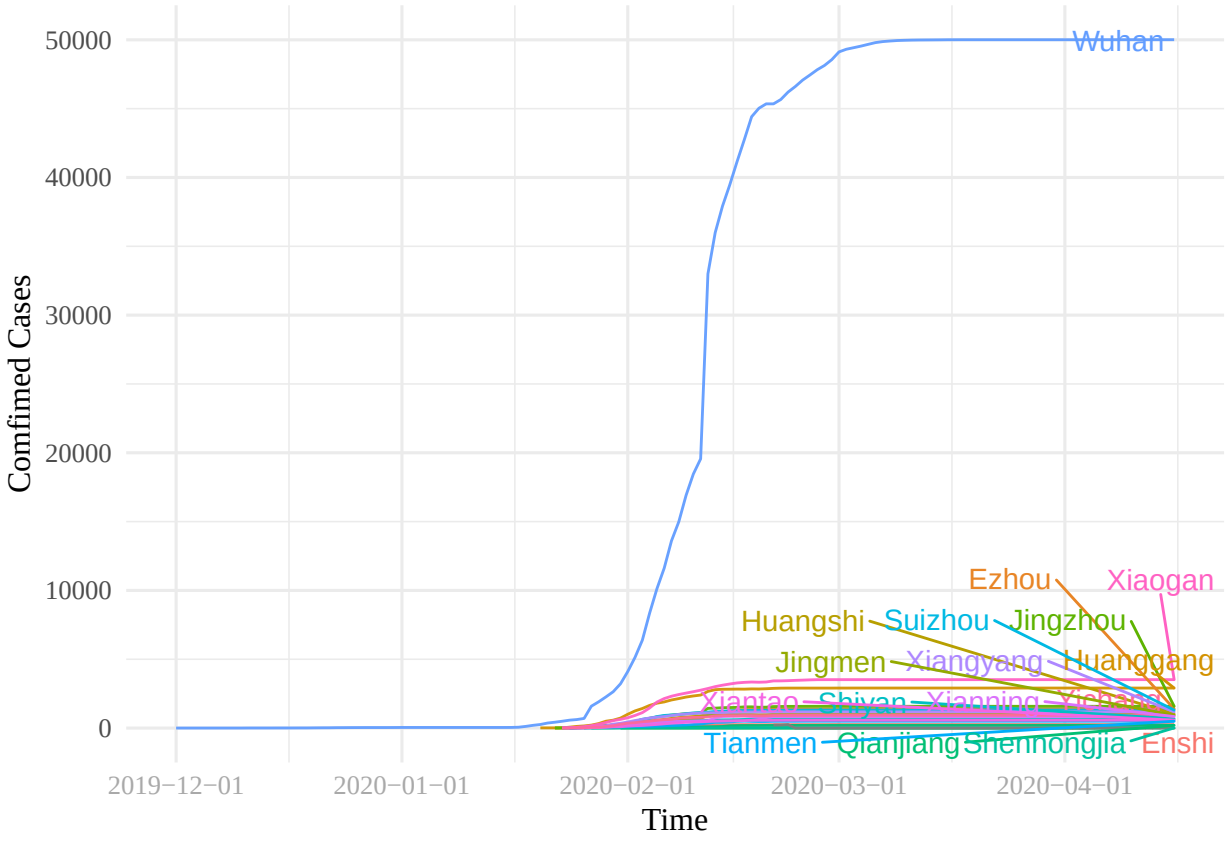

Figure 5: Trend of each city in Hubei province, China

```
# Group by Hubei cities
Historical_HubeiCityV3 <- Historical_HubeiCityV2 %>%
  filter(city!= "Location TBD" & city!= "Wuhan")

# Save as csv
write.csv(Historical_HubeiCityV3,
          file = "./Data/Processed/Historical_HubeiCity (except Wuhan).csv", row.names=F
```

```
# Create a ggplot depicting cases incereasing over time
No_Wuhan.plot <- ggplot(Historical_HubeiCityV3, aes(x=time, y=cum_confirm, color=city))
    geom_line(alpha = 0.95, size = 0.5) +
    geom_text_repel(aes(label=city),
                    function(Historical_HubeiCityV3)
                      Historical_HubeiCityV3[Historical_HubeiCityV3$time == as.Date("202
    mytheme +
    theme(legend.position = "none") +
    labs(x=expression(paste("Time"))) +
    labs(y=expression(paste("Comfimed Cases")))+
    labs(color="city") +
    scale_x_date(date_labels = "%Y-%m-%d")

print(No_Wuhan.plot)
```

Based on the information from Wikipedia, On 23 January 2020, the central government of China imposed a lockdown in Wuhan and other cities in Hubei in an effort to quarantine the center of an outbreak of coronavirus disease 2019 (COVID-19). To be noticed, Xiaogan and Huanggang become the worst places in the country outside Wuhan, because before the "lockdown" in Wuhan, the flow of people in Wuhan mainly flowed into Xiaogan and Huanggang. Hoever, based on this action, The confirmed cases has stabilized after one month.

4.1.5 Trend of confirmed cases in China provinces(except Hubei)

Hubei is not a typical changing pattern in China, so I deleted Hubei data, and see patterns of other provinces.

```
# Remove Hubei province data
Historical_ChinaProvince = Historical_China_processed[Historical_China_processed$provinc
head(Historical_ChinaProvince)

Historical_ChinaProvinceV2 <- Historical_ChinaProvince %>%
  group_by(time,province) %>%
  summarise(total_confirm = sum(cum_confirm),
            total_heal = sum(cum_heal),
            total_dead = sum(cum_dead))
```

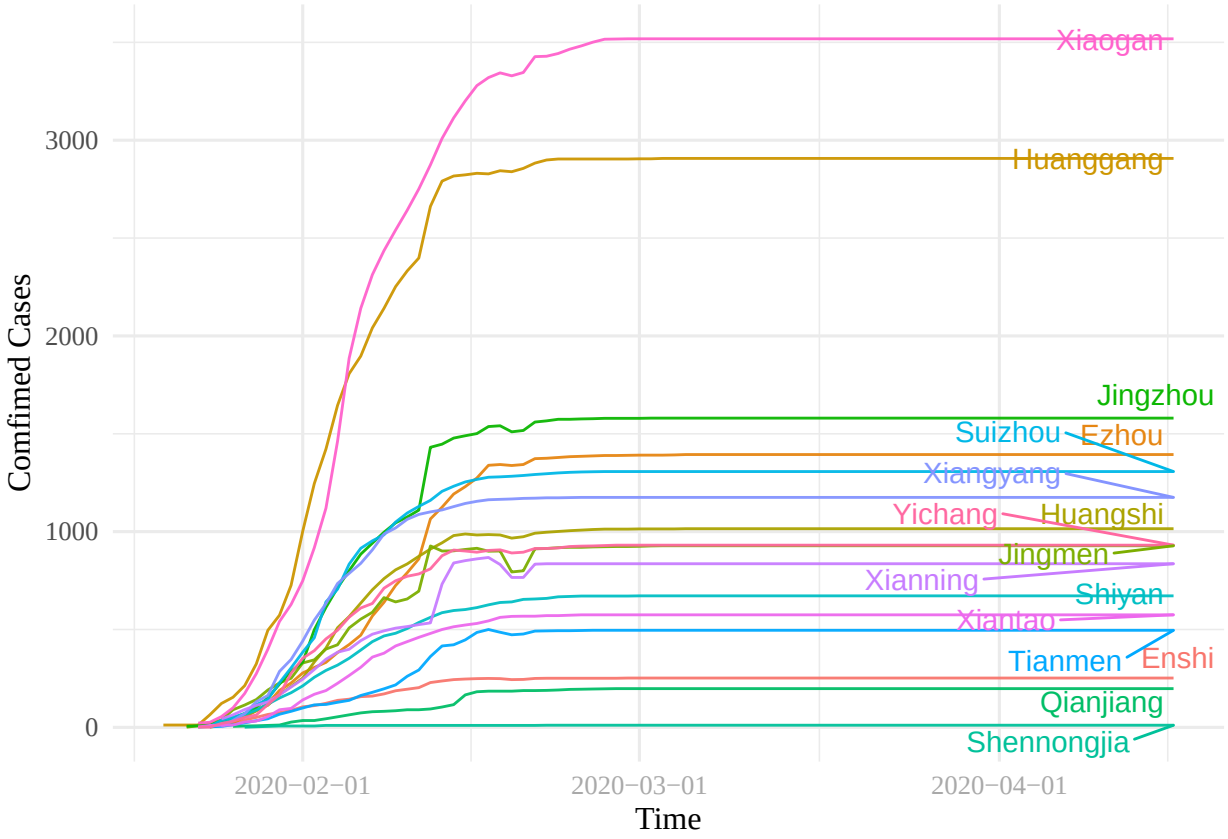

Figure 6: Trend of each city in Hubei province,China (except Wuhan)

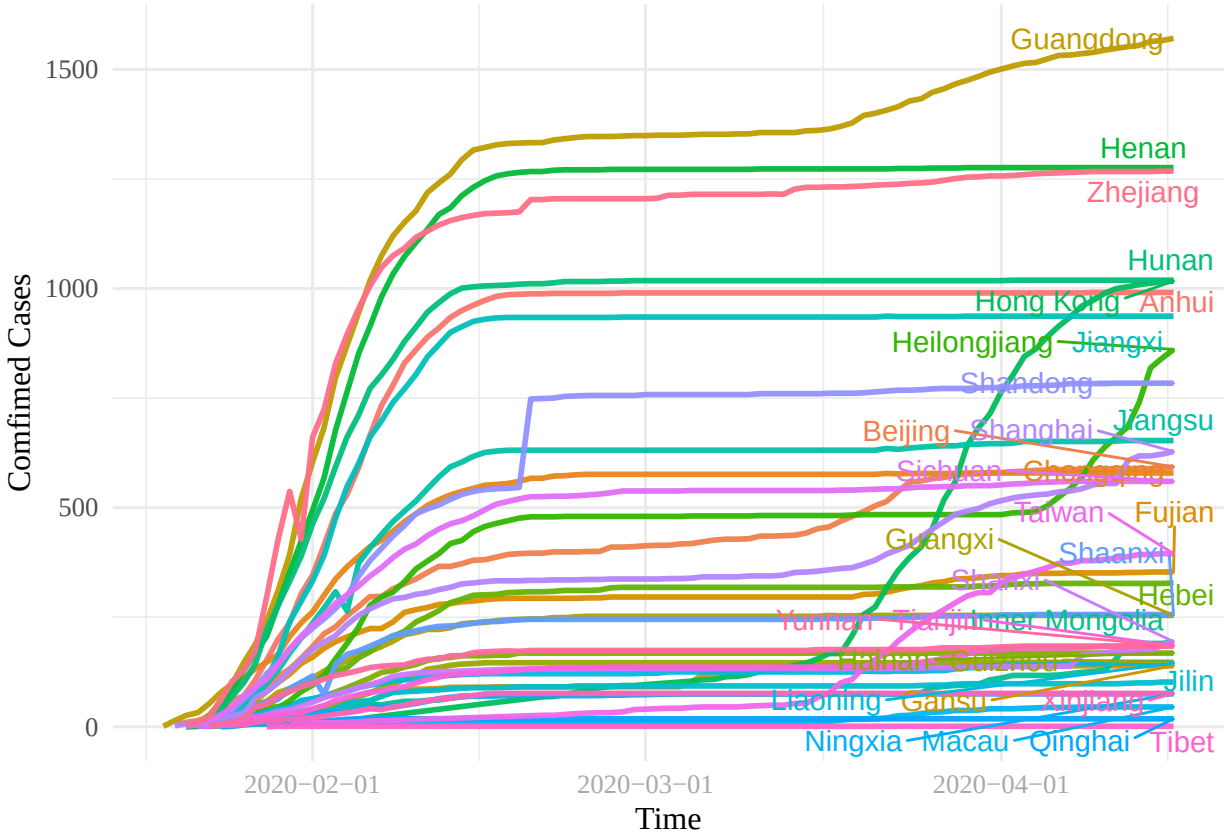

Figure 7: Trend of each province in China (except Hubei)

```
# Save as csv
write.csv(Historical_ChinaProvinceV2,
          file = "./Data/Processed/Historical_ChinaProvince (except Hubei).csv", row.nam
```

```
# Create a ggplot depicting cases incereasing over time
OtherProvinces.plot <- ggplot(Historical_ChinaProvinceV2, aes(x=time, y=total_confirm, c
    geom_line(alpha = 0.95, size = 1) +
    geom_text_repel(aes(label=province),
                    function(Historical_ChinaProvinceV2)
                      Historical_ChinaProvinceV2[Historical_ChinaProvinceV2$time == as.D
    mytheme +
    theme(legend.position = "none") +
    labs(x=expression(paste("Time"))) +
    labs(y=expression(paste("Comfimed Cases"))) +
    labs(color="province") +
    scale_x_date(date_labels = "%Y-%m-%d")

print(OtherProvinces.plot)
```

4.1.6 Trend of confirmed cases of top ten provinces in China (except Hubei)

Some province are far from Hubei, and has less cases, which is also not a typical trend pattern. Therfore, I selected top ten probince to see their trends.

```
# Filter out the top ten provinces with the highest number of diagnoses (except Hubei)
TopTen_provinces <- Historical_ChinaProvinceV2 %>%
  filter(time >= as.Date("2020-04-16")) %>%
  top_n(10, total_confirm) %>%
  arrange(desc(total_confirm))

TopTen_provinces <- pull(TopTen_provinces, province)

TenProvinces_China <- filter(Historical_ChinaProvinceV2, province %in% TopTen_provinces)
  arrange(desc(total_confirm))

head(TenProvinces_China)

# Save as csv
write.csv(TenProvinces_China,
          file = "./Data/Processed/TenProvinces_China (except Hubei).csv", row.names=FAL
```

```
# Draw plot
TenProvinces_trend.plot <- ggplot(TenProvinces_China, aes(x = time, y =total_confirm, co
    geom_line(alpha = 0.95, size = 1) +
    geom_text_repel(aes(label=province),
                    function(Historical_ChinaProvinceV2)
                      Historical_ChinaProvinceV2[Historical_ChinaProvinceV2$time == as.D
    mytheme +
    theme(legend.position = "none") +
    labs(x=expression(paste("Time"))) +
    labs(y=expression(paste("Comfimed Cases"))) +
    labs(color="province") +
    scale_x_date(date_labels = "%Y-%m-%d")

print(TenProvinces_trend.plot)
```

Heilongjiang, Hongkong, and Guangzhou provinces should be notied form TenProvinces_trend.plot.

According to the information I have obtained from the news, the epidemic situation of Russia 's COVID-19 has deteriorated rapidly in the past few days (mid-April). The number of cases of COVID-19 detected in Helongjiang from Russia has increased, which makes it the most "outside import" case in China.

According to the news I saw, the reason is that the new cases in early March were mainly from Hong Kong people who participated in two tours to India and

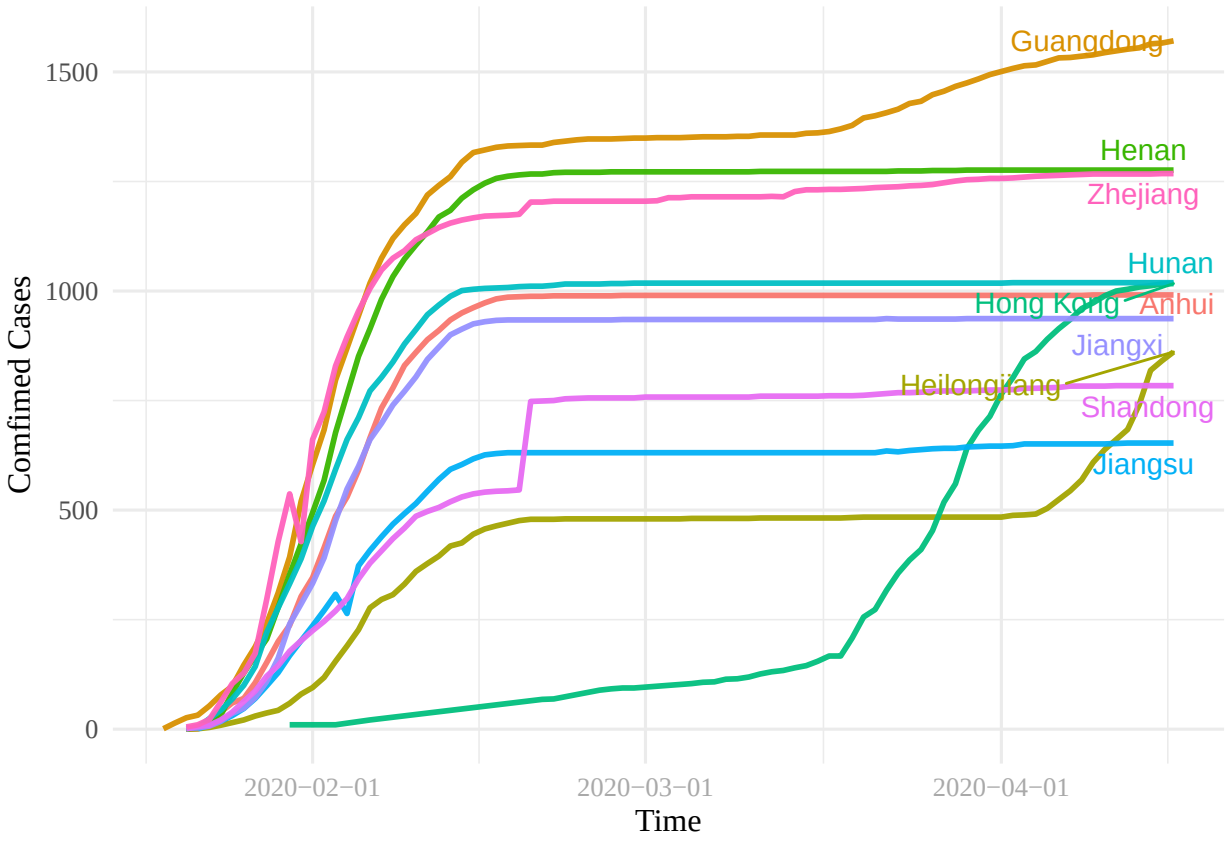

Figure 8: Trend of top ten provinces, China (except Hubei)

Egypt. The new cases increased rapidly in the second half, mainly from people returning from overseas International students and Hong Kong people who have settled in foreign countries, as well as the Languifang Bar infection group.

There are many people came back to Guangzhou from overseas coutries, and cause the new increase in late March.

4.2 Anlysis Global trend of confirmed cases of top ten countries (except China)

The breakout time of China is different from that of other countries, so I deleted China data and see other countries patterns. Also, I selected top ten countries to narrow down the data

```
# Remove China data
Historical_Global = Historical_Global_processed[Historical_Global_processed$country != '
str(Historical_Global)
```

```
## 'data.frame':    8549 obs. of  5 variables:
##  $ time       : Date, format: "2020-01-16" "2020-01-16" ...
##  $ country    : Factor w/ 205 levels "Afghanistan",..: 96 185 96 96 96 175 96 185 175
##  $ cum_confirm: int  1 1 1 1 1 1 1 2 1 1 ...
##  $ cum_heal   : int  1 0 1 1 1 0 1 0 0 1 ...
##  $ cum_dead   : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
# Filter out the top ten countries with the highest number of diagnoses (except China)
TopTen_countries <- Historical_Global %>%
  filter(time >= as.Date("2020-04-16")) %>%
  top_n(10, cum_confirm) %>%
  arrange(desc(cum_confirm))

# Selects a column in a data frame and transforms it into a vector
TopTen_countries <- pull(TopTen_countries, country)

Historical_TopTen <- filter(Historical_Global, country %in% TopTen_countries) %>%
  arrange(desc(cum_confirm))
head(Historical_TopTen)
```

```
##         time       country cum_confirm cum_heal cum_dead
## 1 2020-04-16 United States      650833    52739    32707
## 2 2020-04-15 United States      614726    38879    26126
## 3 2020-04-14 United States      587815    37315    23599
## 4 2020-04-13 United States      556569    32634    22063
## 5 2020-04-12 United States      529112    30548    20549
## 6 2020-04-11 United States      503177    29191    18777
```

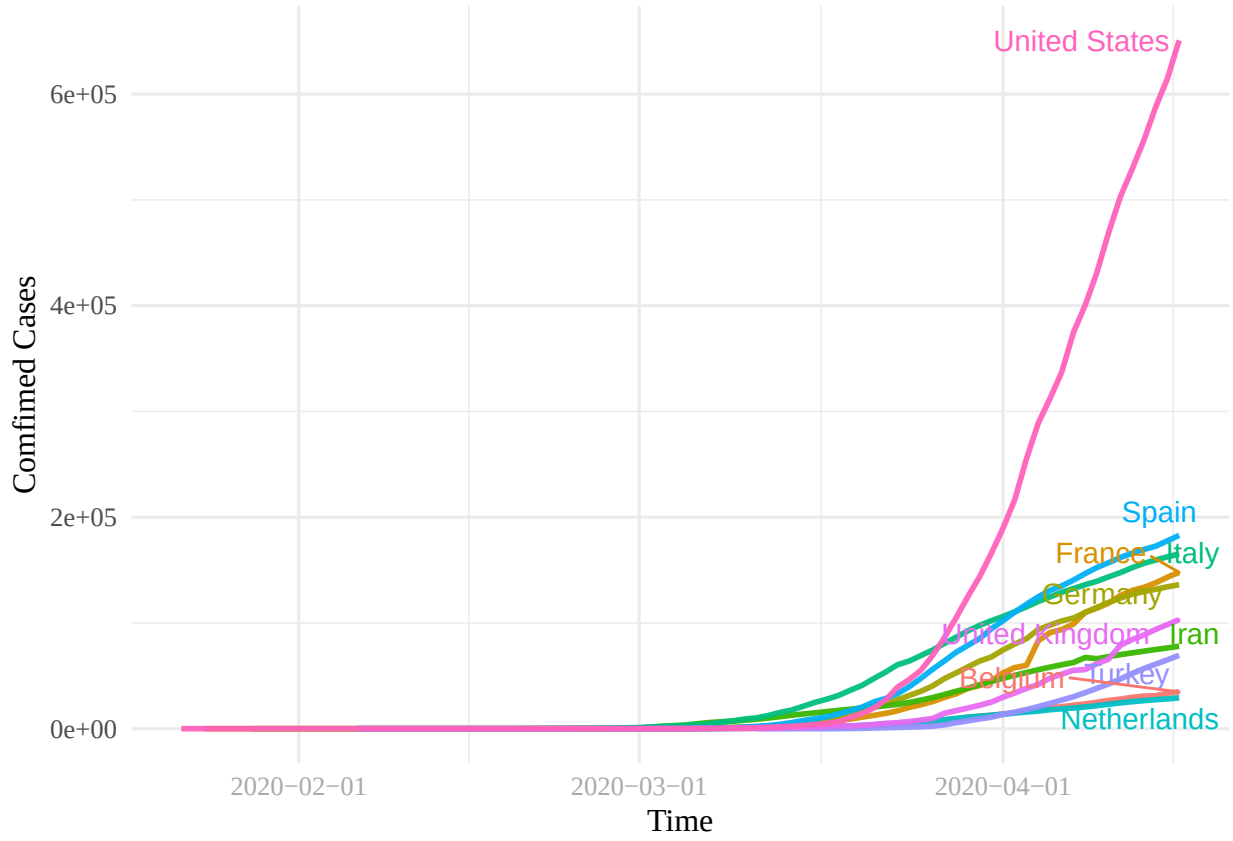

Figure 9: Trend of Top Ten Countries (except China)

```
# Save as csv
write.csv(Historical_TopTen,
          file = "./Data/Processed/TopTen_countries_trend(except China).csv", row.names=
```

```
# Draw plot
TopTenCountries.plot <- ggplot(Historical_TopTen, aes(x = time, y =cum_confirm, color=co
    geom_line(alpha = 0.95, size = 1) +
    geom_text_repel(aes(label=country),
                    function(Historical_TopTen)
                      Historical_TopTen[Historical_TopTen$time == as.Date("2020-04-16"),
    mytheme +
    theme(legend.position = "none") +
    labs(x=expression(paste("Time"))) +
    labs(y=expression(paste("Comfimed Cases"))) +
    labs(color="Country") +
    scale_x_date(date_labels = "%Y-%m-%d")

print(TopTenCountries.plot)
```

4.3 Analysis trend of confirmed cases in United States.

Based on last graph, obvisiouly, the United State has the highest confimred numbers, and I will do the analysis based on the U.S. dataset.

```
Historical_USTrend <- Historical_Global_processed[Historical_Global_processed$country ==
  na.omit(Historical_USTrend)

tail(Historical_USTrend)

# Save as csv
write.csv(Historical_USTrend,
          file = "./Data/Processed/Historical_USTrend.csv",row.names=FALSE)
```

4.3.1 Trend of total confirmed case in United States.

First I would like to see the entire confirmed case trend through the U.S.

```
Historical_USTrend <- read.csv("./Data/Processed/Historical_USTrend.csv")

head(Historical_USTrend)
str(Historical_USTrend)

Historical_USTrend$time <- as.Date(Historical_USTrend$time, format = "%Y-%m-%d")
class(Historical_USTrend$time)
```

```
US_Total_Trend.plot <-
  ggplot(Historical_USTrend, aes(x = time, y = cum_confirm)) +
  geom_point(colour = "#e6550d") +
  geom_line(colour = "#d95f0e") +
  labs(x = "Time",
       y = "Confirmed Cases") +
  scale_x_date(date_labels = "%Y-%m-%d")

print(US_Total_Trend.plot)
```

4.3.2 Trend of total confirmed case in all states in the U.S.

```
# Create a ggplot depicting cases incereasing over time in each state
US_State_Trend.plot <- ggplot(Covid19_US_processed, aes(x=time, y=cum_confirm, color=sta
    geom_line(alpha = 0.95, size = 0.5) +
    geom_text_repel(aes(label=state),
                    function(Covid19_US_processed)
                      Covid19_US_processed[Covid19_US_processed$time == as.Date("2020-04
    mytheme +
    theme(legend.position = "none") +
```

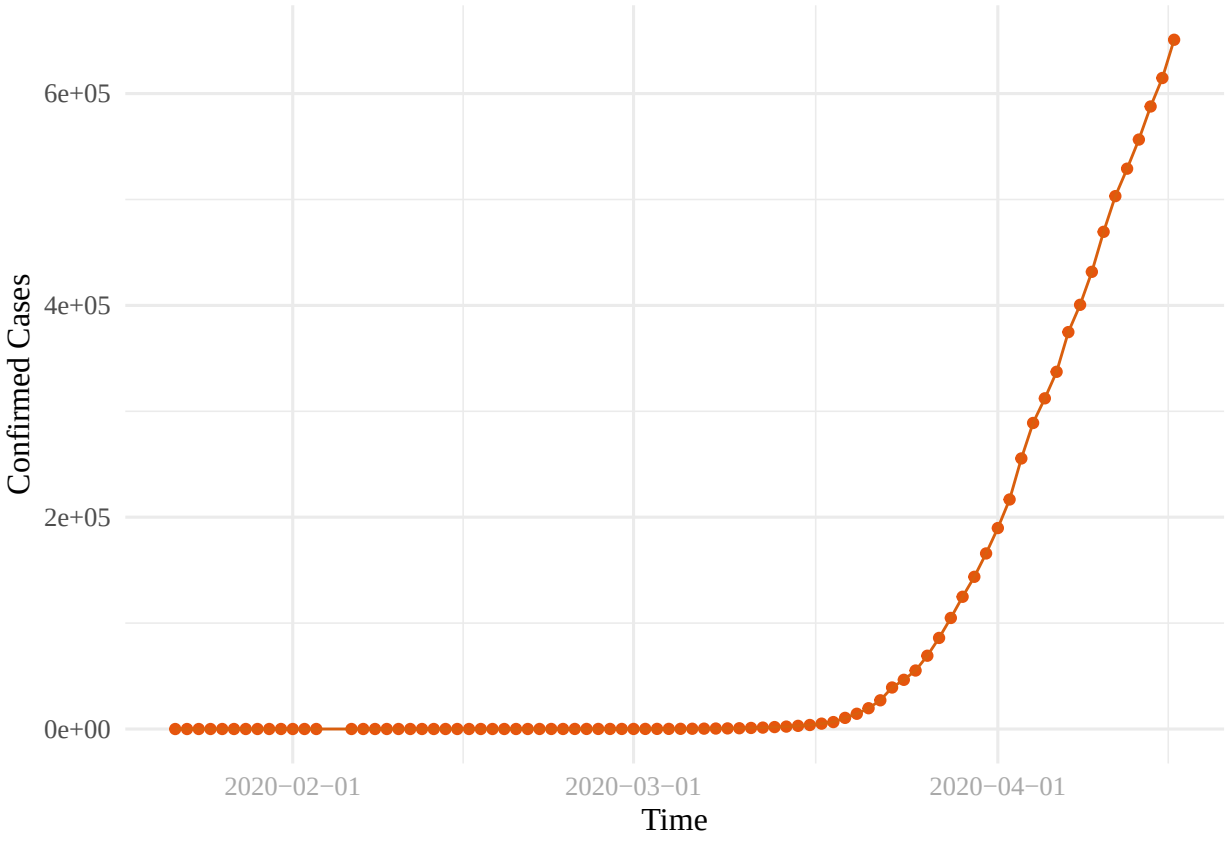

Figure 10: Trend of confirmed case in the U.S.

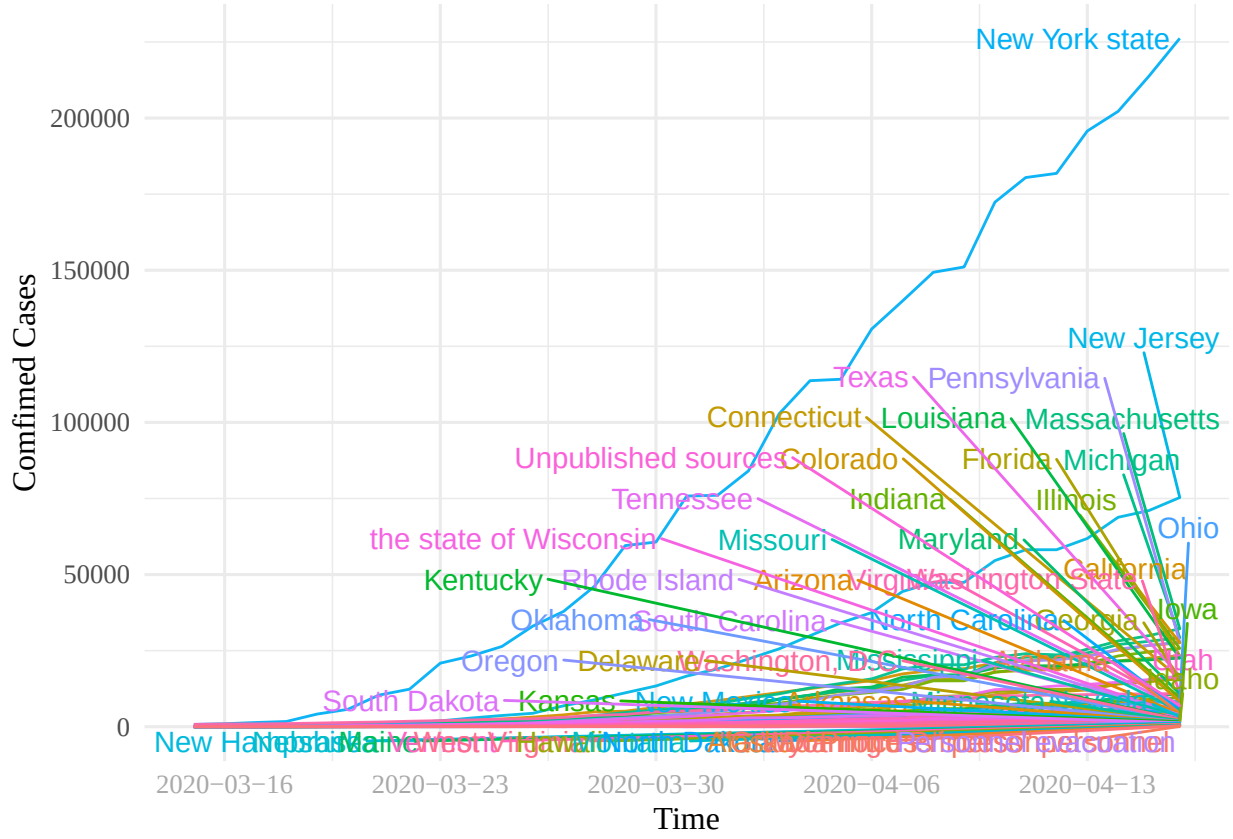

Figure 11: Trend of each states in United States

```
    labs(x=expression(paste("Time"))) +
    labs(y=expression(paste("Comfimed Cases")))+
    labs(color="Covid19_US_processed") +
    scale_x_date(date_labels = "%Y-%m-%d")

print(US_State_Trend.plot)
```

I found pattern of New York is simaliar to that of whole US pattern.

4.3.3 Trend of total confirmed case of top ten states in the U.S. (except NY and NJ)

New York and New Jersey have the almost 50% confirmed cases in US, so I seleted data fro these two states, and ckeck other states pattern. I also selected top ten states, which is more representative.

```
TopTen_States <- Covid19_US_processed %>%
  filter(time >= as.Date("2020-04-16")) %>%
  filter(state!= "New York state" & state!= "New Jersey" & state!= "Unpublishedsources")
```

```
  top_n(10, cum_confirm) %>%
  arrange(desc(cum_confirm))

TopTen_States <- pull(TopTen_States, state)

TenStates_US <- Covid19_US_processed %>%
  filter(state %in% TopTen_States)

head(TenStates_US)

# Save as csv
write.csv(TenStates_US,
          file = "./Data/Processed/TenStates_US (except NY and NJ).csv")
```

```
# Create a ggplot
TenStates_Trend.plot <- ggplot(TenStates_US, aes(x = time, y =cum_confirm, color=state))
  geom_line(alpha = 0.95, size = 1) +
  geom_text_repel(aes(label=state),
                      function(TenStates_US)
                      TenStates_US[TenStates_US$time == as.Date("2020-04-16"),])+
  mytheme +
  theme(legend.position = "none") +
  labs(x=expression(paste("Time"))) +
  labs(y=expression(paste("Comfimed Cases"))) +
  labs(color="state") +
  scale_x_date(date_labels = "%Y-%m-%d")


print(TenStates_Trend.plot)
```

Based on the top ten countries, California, Florida, Illinois, Texas and so on, they are largest state, and has more population, which can show it is an infectious disease.

4.4 Analysis trend of dead and heal cases number and rate

4.4.1 China

```
str(Covid19_China_processed)
head(Covid19_China_processed)

names(Covid19_China_processed)[1] <- "Province"
names(Covid19_China_processed)[4] <- "Dead Rate"
names(Covid19_China_processed)[6] <- "Heal Rate"
```

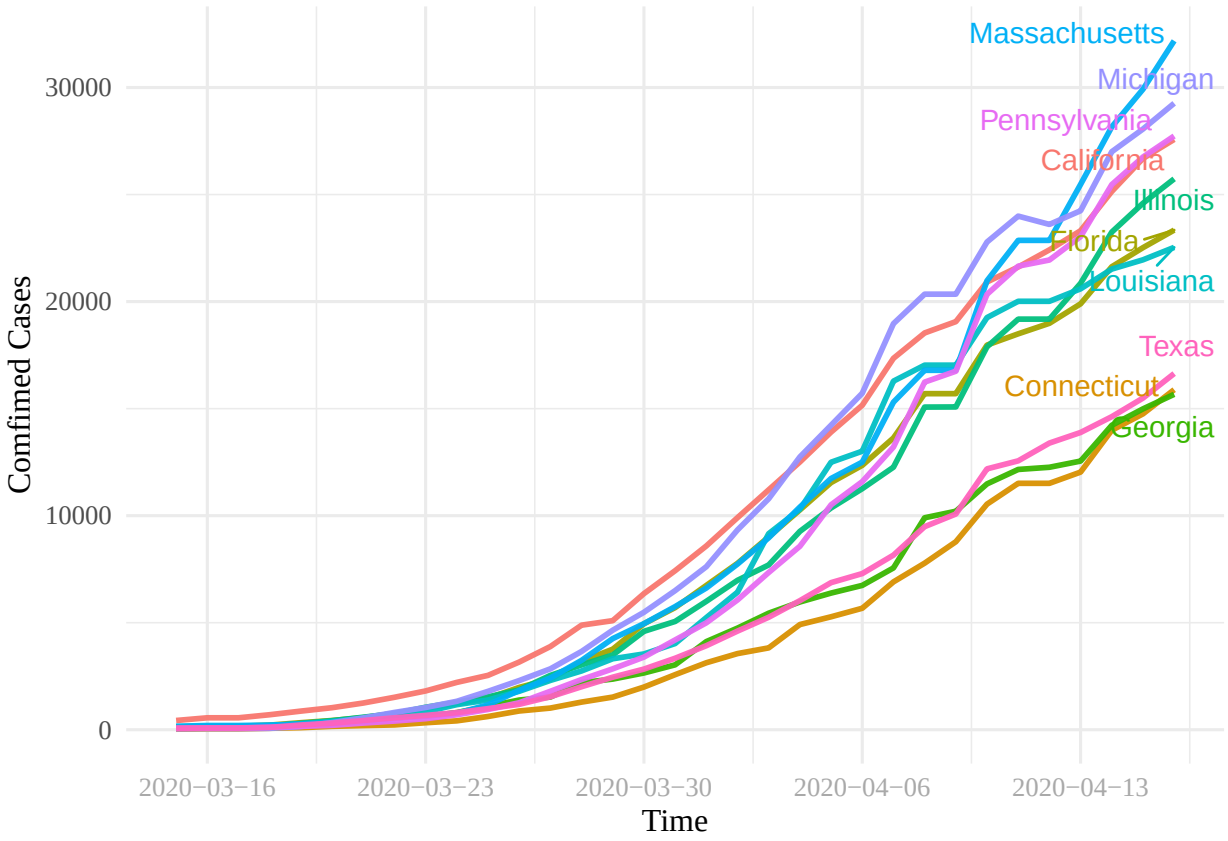

Figure 12: Trend of top ten states in United States (except NY and NJ)

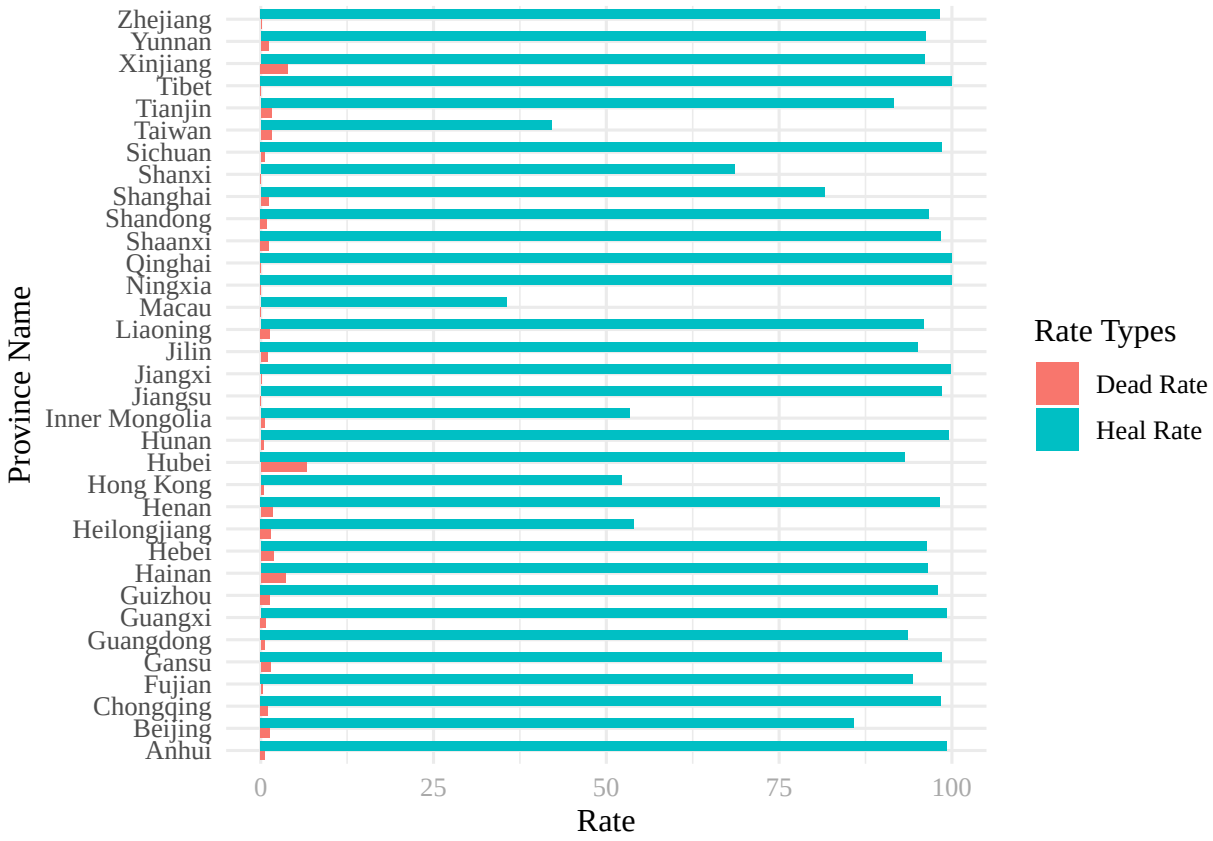

Figure 13: Compare total death rate and heal rate in each province (China)

```
ChinaTrend_gather <- tidyr::gather(Covid19_China_processed, "Type","Rate", 4,6)
head(ChinaTrend_gather)
```

```
China_Rate.plot <- ggplot(ChinaTrend_gather, aes(fill=Type, y=Rate, x=Province)) +
  geom_bar(stat = "identity",position=position_dodge()) +
  coord_flip() +
  labs(x=expression(paste("Province Name"))) +
  labs(y=expression(paste("Rate"))) +
  labs(fill = "Rate Types") +
  mytheme +
  theme(legend.position = "right")
print(China_Rate.plot)
```

In this Section, I will exeplore and analysis dead and heal cases number and rate. I used bar plot to see the heal rate and dead rate in each province in China (This dataset update on April 16th).

Trend of confirms, deaths,and heal number in China

```r
# Change Historical_China_processed dataset's Column names
str(Historical_China_processed)

Historical_China_processed2 <- Historical_China_processed %>%
  group_by(time,country) %>%
  summarise(total_confirm = sum(cum_confirm),
            total_heal = sum(cum_heal),
            total_dead = sum(cum_dead))
head(Historical_China_processed2)

names(Historical_China_processed2)[1] <- "Time"
names(Historical_China_processed2)[3] <- "Total Confirm"
names(Historical_China_processed2)[4] <- "Total Heal"
names(Historical_China_processed2)[5] <- "Total Dead"

Historical_China_gather <- tidyr::gather(Historical_China_processed2, "Type", "Number",
str(Historical_China_gather)
```

```r
China_Number_Trend.plot <- ggplot(Historical_China_gather, aes(Time, Number, color = Typ
  geom_point() +
  geom_line() +
  labs(x=expression(paste("Time"))) +
  labs(y=expression(paste("Total number")))+
  mytheme +
  theme(legend.position = "right") +
  scale_x_date(date_labels = "%Y-%m-%d")
print(China_Number_Trend.plot)
```

Trend of confirms, deaths,and heal number in China (12/1/2019-2/15/2020")

```r
China_Number_Trend_Limit.plot <- ggplot(Historical_China_gather, aes(Time, Number, color
  geom_point() +
  geom_line() +
  labs(x=expression(paste("Time"))) +
  labs(y=expression(paste("Total number")))+
  mytheme +
  theme(legend.position = "right") +
  scale_x_date(date_labels = "%Y-%m-%d",
    limits = c(as.Date("2019-12-1"), as.Date("2020-02-15"))) +
  ylim(c(0, 500))
print(China_Number_Trend_Limit.plot)
```

I enlarged the pattern of January and Febuary to see if there is difference between ealry period and late period. Comparing the trajectory of confirmed cases, heal cases, and death cases number may have some policy experience for other countries.

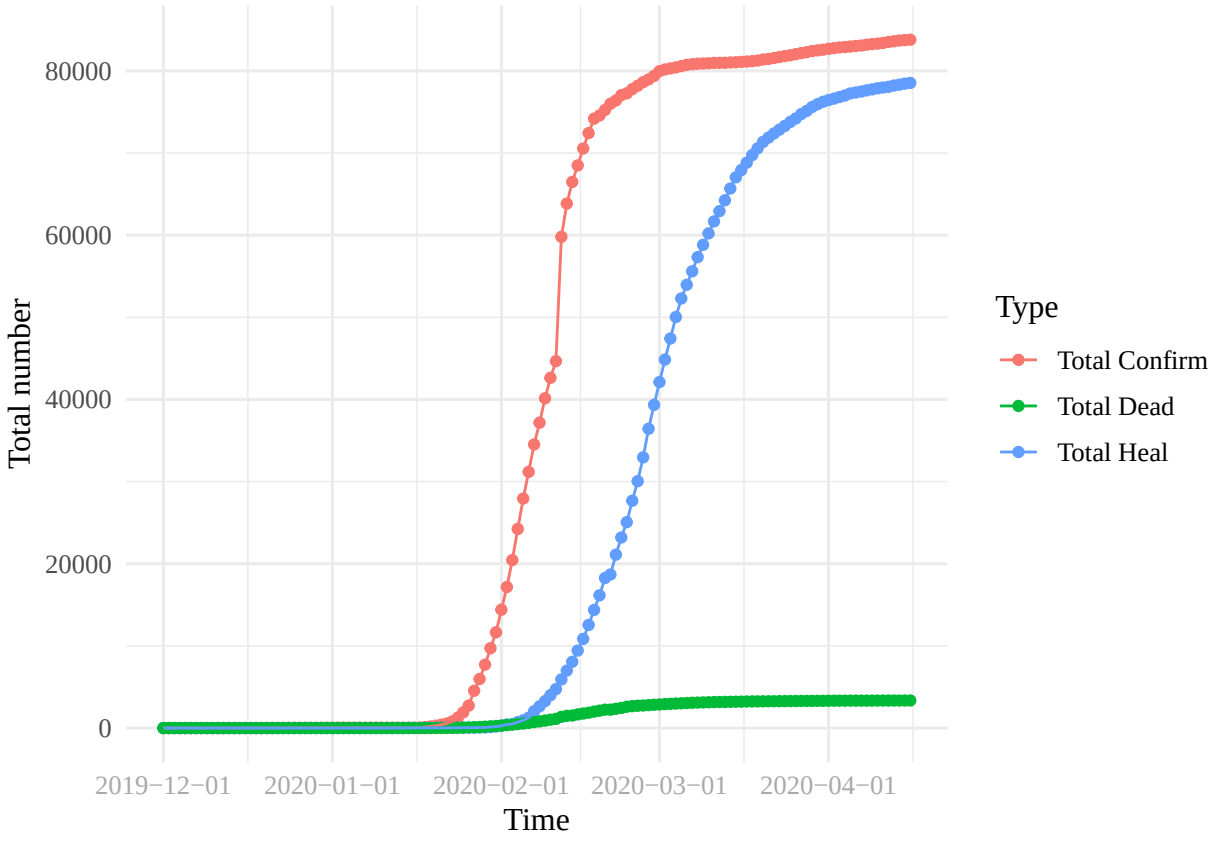

Figure 14: Trend of deaths, confirms, and heal number in China

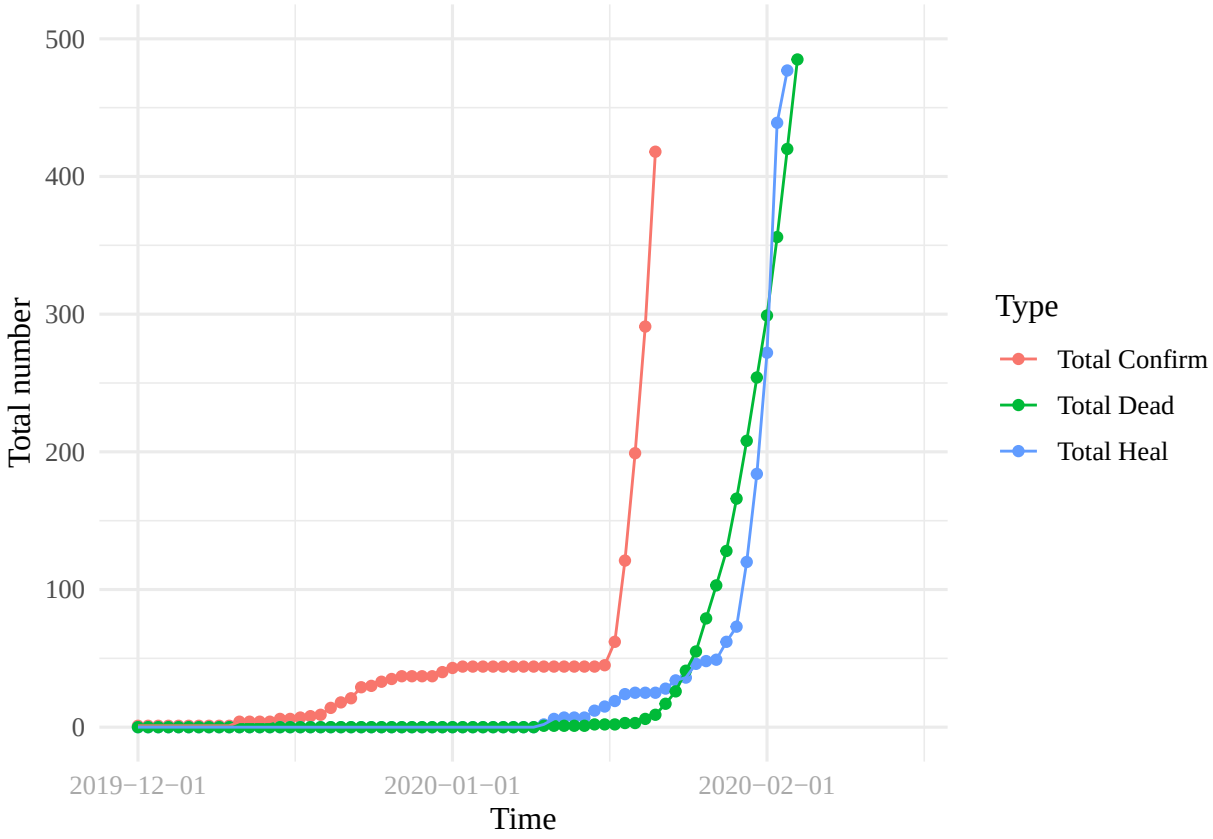

Figure 15: Trend of deaths, confirms, and heal number in China (12/1/2019-2/15/2019

During the late January, the confirmed cases and death numbers are increase rapidly. After lockdown the cities, the confirmed cases was gruadly stablized in mid-Febuary. Therefore, the social distancing wroks to reduece the infected cases.

Trend of confirms, deaths,and heal number top ten provines, China

```
head(TenProvinces_China)

names(TenProvinces_China)[1] <- "Time"
names(TenProvinces_China)[3] <- "Total Confirm"
names(TenProvinces_China)[4] <- "Total Heal"
names(TenProvinces_China)[5] <- "Total Dead"

TenProvinces_China_gather <- tidyr::gather(TenProvinces_China, "Type", "Number", 3:5)
str(TenProvinces_China_gather)
```

```
TenProvinces_Number_Trend.plot <-
  ggplot(TenProvinces_China_gather, aes(x = Time, y = Number, color = Type)) +
  geom_line() +
  geom_point() +
  labs(x=expression(paste("Time"))) +
  labs(y=expression(paste("Total number")))+
  mytheme +
  theme(legend.position = "right") +
  scale_x_date(date_labels = "%Y-%m-%d",
    limits = c(as.Date("2020-1-15"), as.Date("2020-02-15"))) +
  ylim(c(0, 500)) +
  facet_wrap(.~province)
print(TenProvinces_Number_Trend.plot)
```

```
## Warning: Removed 198 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 1909 rows containing missing values (geom_point).
```

To see the trends of top ten provinces (ecpect Hubei), Guangdong, Henan, Jiangxi, Zhejiang, and Hunan have smiliar trend. Because these five provinces have large population movements during the Spring Festival, and people in these province are more suspicious on COVID-19.

4.4.2 Hubei province, China

Trend of confirms, deaths,and heal number in top five cities, Hubei(12/1/2019-3/15/2020)

```
head(Historical_Hubei)

# Select top five cities
TopFive_Hubei <- Historical_Hubei %>%
```

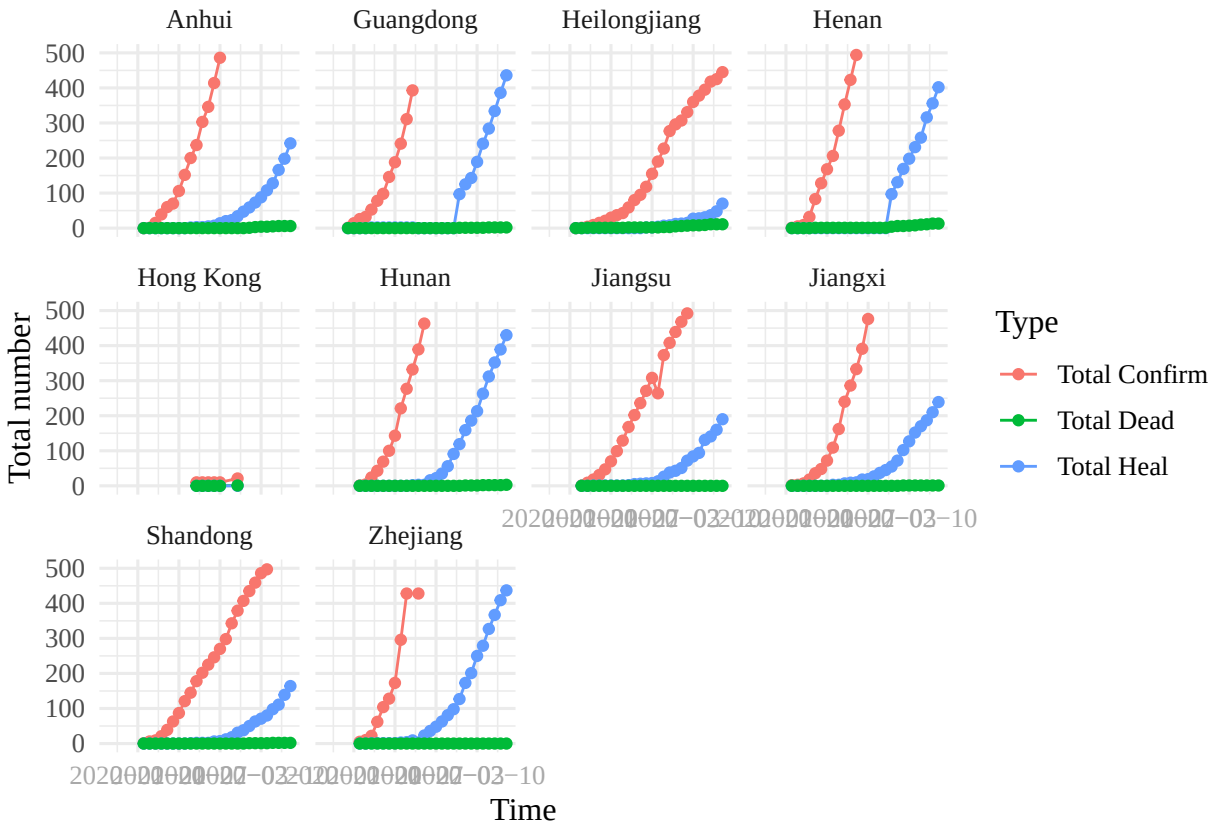

Figure 16: Trend of deaths, confirms, and heal number in top ten provine (expect Hubei), China

```
  filter(time >= as.Date("2020-04-16")) %>%
  top_n(5, cum_confirm) %>%
  arrange(desc(cum_confirm))

TopFive_Hubei <- pull(TopFive_Hubei, city)

TopFiveCity_Hubei <- filter(Historical_Hubei, city %in% TopFive_Hubei) %>%
  arrange(desc(cum_confirm))

head(TopFiveCity_Hubei)

# Change column names
names(TopFiveCity_Hubei)[1] <- "Time"
names(TopFiveCity_Hubei)[5] <- "Total Confirm"
names(TopFiveCity_Hubei)[6] <- "Total Heal"
names(TopFiveCity_Hubei)[7] <- "Total Dead"

Historical_Hubei_gather <- tidyr::gather(TopFiveCity_Hubei, "Type", "Number", 5:7)
head(Historical_Hubei_gather)
```

```
FiveCities_Number_Trend.plot <-
  ggplot(Historical_Hubei_gather, aes(x = Time, y = Number, color = Type)) +
  geom_line() +
  geom_point() +
  labs(x=expression(paste("Time"))) +
  labs(y=expression(paste("Total number")))+
  mytheme +
  theme(legend.position = "right") +
  scale_x_date(date_labels = "%Y-%m-%d",
    limits = c(as.Date("2019-12-1"), as.Date("2020-03-15"))) +
  ylim(c(0, 5000)) +
  facet_wrap(.~city)
print(FiveCities_Number_Trend.plot)
```

I also pulled the top five five cities in Hubei, and found excepy Wuhan, other cities have the same trend pattern.

4.4.3 United States

```
# Change Historical_USTrend dataset's Column names
str(Historical_USTrend)

Historical_USTrend2 <- Historical_USTrend
names(Historical_USTrend2)[1] <- "Time"
```

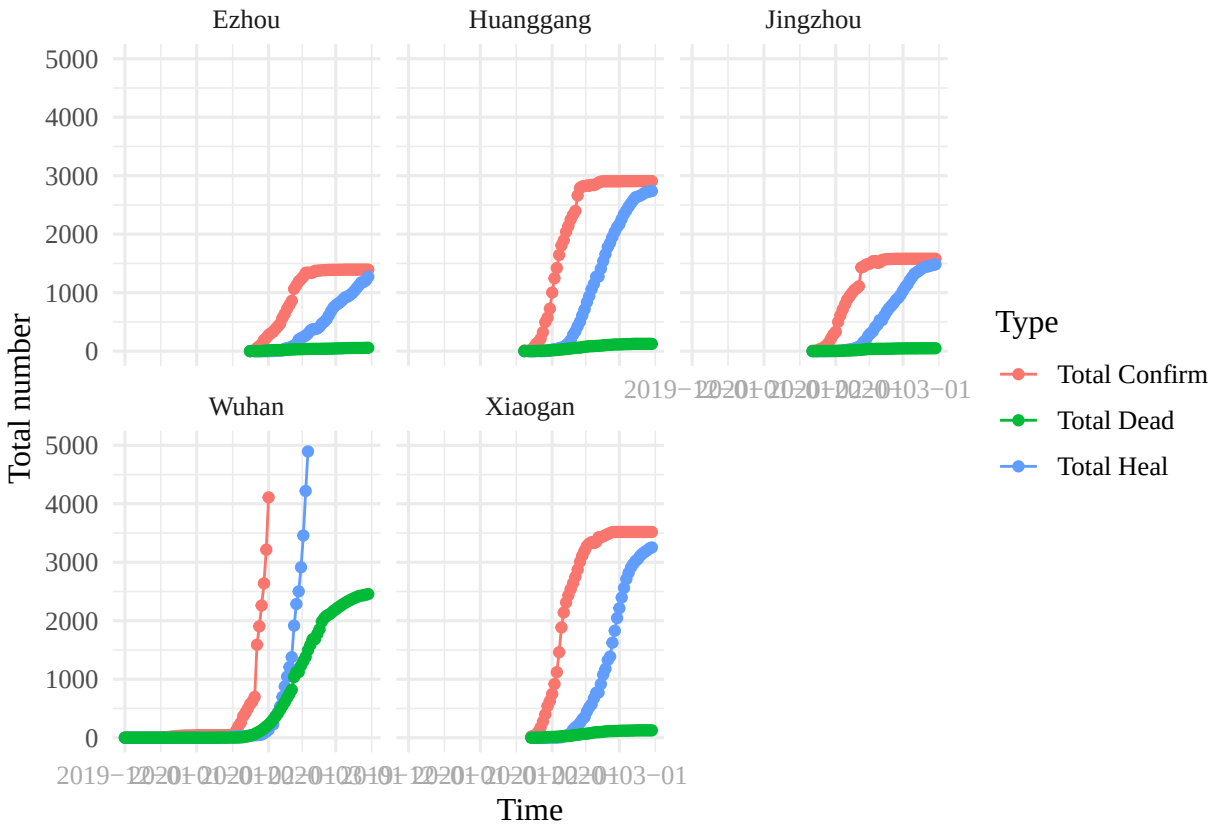

Figure 17: Trend of confirms, deaths,and heal number in top five cities, Hubei(12/1/2019-3/15/2020)

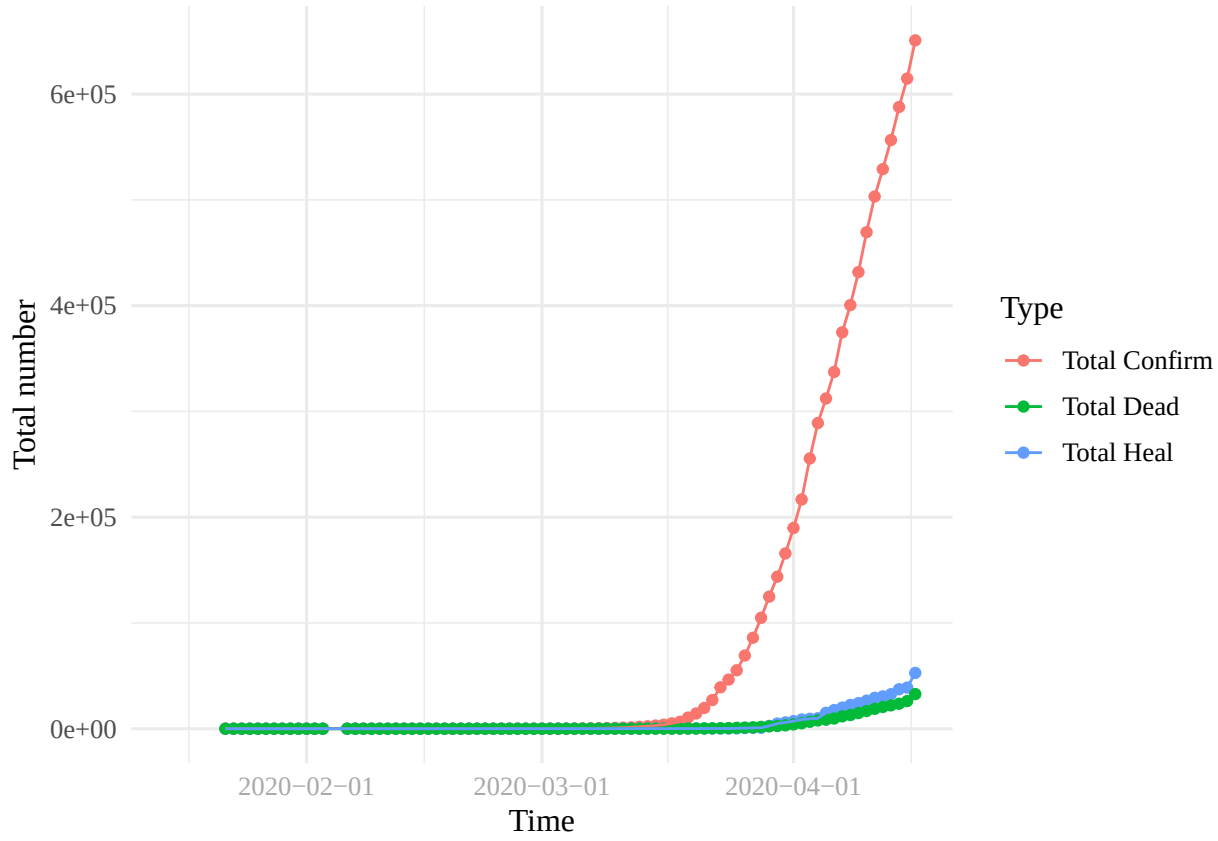

Figure 18: Trend of confirms, deaths,and heal number in the U.S.

```
names(Historical_USTrend2)[3] <- "Total Confirm"
names(Historical_USTrend2)[4] <- "Total Heal"
names(Historical_USTrend2)[5] <- "Total Dead"

USTrend_gather <- tidyr::gather(Historical_USTrend2, "Type", "Number", 3:5)
str(USTrend_gather)
```

```
US_Number_Trend.plot <- ggplot(USTrend_gather, aes(Time, Number, color = Type)) +
  geom_point() +
  geom_line() +
  labs(x=expression(paste("Time"))) +
  labs(y=expression(paste("Total number")))+
  mytheme +
  theme(legend.position = "right") +
  scale_x_date(date_labels = "%Y-%m-%d",
    limits = c(as.Date("2020-01-15"), as.Date("2020-04-16")))
print(US_Number_Trend.plot)
```

Heal and dead number plot for top ten states

```r
head(TenStates_US)

names(TenStates_US)[1] <- "Time"
names(TenStates_US)[4] <- "Total Confirm"
names(TenStates_US)[5] <- "Total Heal"
names(TenStates_US)[6] <- "Total Dead"

TenStates_US_gather <- tidyr::gather(TenStates_US, "Type", "Number", 4:6)
str(TenStates_US_gather)
```

```r
Ten_States_Number_Trend.plot <-
  ggplot(TenStates_US_gather, aes(x = Time, y = Number, color = Type)) +
  geom_line() +
  geom_point() +
  labs(x=expression(paste("Time"))) +
  labs(y=expression(paste("Total number")))+
  mytheme +
  theme(legend.position = "right") +
  facet_wrap(.~state)
print(Ten_States_Number_Trend.plot)
```

Compared tope ten states (except New York and New Jersey) to whole United States trends. The confimred cases lines are all increasing, but they are less steep in April than those in March.

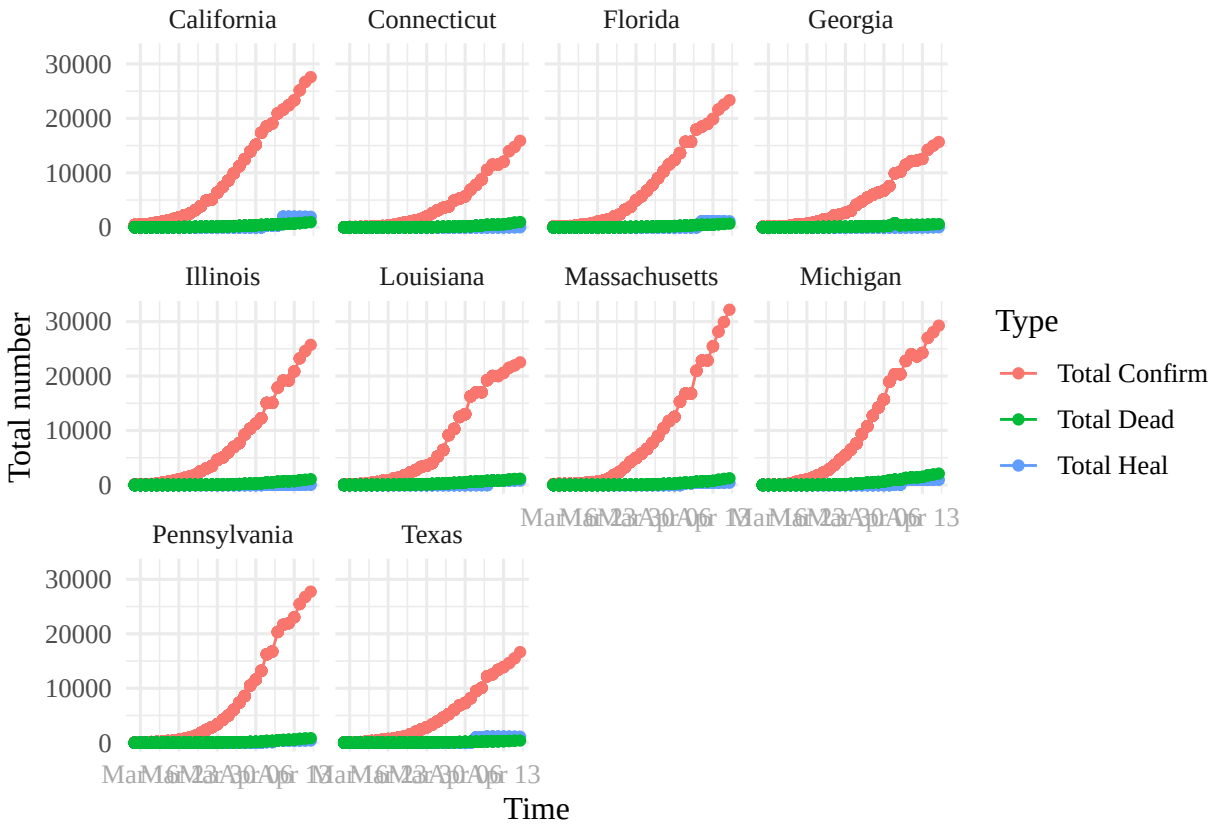

Figure 19: Trend of confirms, deaths,and heal number in top ten states

5 Summary and Conclusions

A "COVID-19" epidemic caused by a new coronavirus has spread worldwide in few months, which is an unprecedented rapid spread. China started to lockdown cities with all countries evacuating overseas, banned navigation, and many countries, such as United States, have closed their borders.

This Study uses visualization method to depict incresed confirmed cases trend on China, United States, and global level, and find their patterns. I tried to use time series test analysis to predict the trend of confirmeded case number, but it does not work due to short time period. However, Hubei Province, China is the first place was reported COVID-19. Therefore, I focus on ploting figures based on China and US data to see if there are same patterns and relevences.

By comparing the cases trend between provinces or states and country, contry and country, and global level, I found the intial trend pattern are very similar. Using the trajectory of the confirmed curve and death curve in China predicts the progress of the outbreak in the U.S. and other countries, which is a perspective way to set up some policies. However, the characteristics of the virus and the trend of the epidemic are still full of variables and unknowns. Every country's actions and every international organization's decision-making affect not only the safety of the group's lives, but also the political and economic development of each country, and are more likely to affect the global situation. Therefore, even if there are similarities in early data from various countries, one cannot assume that countries will follow the same trajectory.

China has closed off a city of more than 11 million people in an unprecedented effort to try to stop the spread of a deadly new virus. Also, Everyone has a health code after stoping lockdown the cities, and health code is an electronic voucher for individuals to pass in and out of the local area, using the location recorded by the mobile phone GPS to determine whether the other party is a close contact. Each state in the U.S. also operate social-distance to prevent virvus spread. However, China implemented restrictions earlier than other countries, and the health code may not appear in the United States for many reasons in western countries. Therefore, the trend of Covid-19 still need time to tell us.

In the furture research, I will explore the relationship of Covid-19 cases number with high-risk susceptible population and race. Besides, I would like to explore the relationship of Covid-19 with social, economic fields, such as states GDP and states crime rate, to have a whole overview on society and COVID-19.

6 References

Tianzhi Wu, Erqiang Hu, Xijin Ge, *Guangchuang Yu*. Open-source analytics tools for studying the COVID-19 coronavirus outbreak. medRxiv, 2020.02.25.20027433. doi: https://doi.org/10.1101/2020.02.25.20027433*