# STAT2011 Project II
# Exploratory Data Analysis on the Novel Coronavirus

Name: King Yeung CHAN
SID: 1155119394

07/05/2020

## Introduction

At the late end of 2019, the novel coronavirus, also known as COVID-19, has become an epidemic and communicable diseases around the world. It, originally, has been discovered in Wuhan, Hubei Province, China. The number of confirmed cases and deaths are growing exponentially. Facing the pandemic of the novel coronavirus, classes are suspended and many companies recommend their employees to work from home. Staying at home, we concern the current situation of the novel coronavirus. Unlike others, the novel coronavirus is asymptomatic and many people do not know they are infected. To have a wider view of the novel coronavirus, thus, we make use of the **Novel Corona Virus 2019 Dataset** from Kaggle. The data set consists of daily information about the novel coronavirus, which including the number of confirmed, death and recovered cases. As said, the data set is updated on a daily basis, we treat it as a time series data set. For each data, it represents in terms of cumulative on any given day. All the recorded start on 22 January 2020.

The dataset was found from Kaggle, which is well-known for many competitions about Machine Learning and Data Science. Data are collected from some well-known international organisations, such as the World Health Organization (WHO). We believe the given data are reliable with such a reason. Kaggle is referred from the Statistics Department website page under the section of **Datasets for Data Mining and Data Science, listed by KDnuggets**.

## Data Import

We import the dataset “covid_19_data.csv” from a defined working directory.

```
setwd("/Users/jackchan/Documents/I go to school by bus/03 The Chinese University of H
ong Kong/Year 2/Semester 2/STAT2011 Workshop on Data Exploration and Technical Writin
g/Project II/novel-corona-virus-2019-dataset")

data <- read.csv("covid_19_data.csv", header = TRUE)
```

After importing the data set, here, we inspect a few observation from the dataset to have a quick view of the data set.

```
head(data)
```

```
##   SNo ObservationDate Province.State Country.Region     Last.Update Confirmed
## 1   1      01/22/2020          Anhui Mainland China 1/22/2020 17:00         1
## 2   2      01/22/2020        Beijing Mainland China 1/22/2020 17:00        14
## 3   3      01/22/2020      Chongqing Mainland China 1/22/2020 17:00         6
## 4   4      01/22/2020         Fujian Mainland China 1/22/2020 17:00         1
## 5   5      01/22/2020          Gansu Mainland China 1/22/2020 17:00         0
## 6   6      01/22/2020      Guangdong Mainland China 1/22/2020 17:00        26
##   Deaths Recovered
## 1      0         0
## 2      0         0
## 3      0         0
## 4      0         0
## 5      0         0
## 6      0         0
```

# Data Cleaning

As mentioned previously, the **Novel Corona Virus 2019 Dataset** is organised every day. For the consistency of the EDA report, we only focus on the data up to the end of March 2020. Also, we attempt to discard some variables, such as “SNo” and “Last.Update”, that are not interested in the EDA and further studies.

```
Date <- as.Date(data$ObservationDate, format = "%m/%d/%Y")

data <- cbind(data, Date)

data <- data[Date <= as.Date("03/31/2020", format = "%m/%d/%Y"),]

data <- data[, c(9, 3:4, 6:8)]
```

After the modification, we examine some earlier cases to obtain an insight into the data information.

```
head(data)
```

```
##         Date Province.State Country.Region Confirmed Deaths Recovered
## 1 2020-01-22          Anhui Mainland China         1      0         0
## 2 2020-01-22        Beijing Mainland China        14      0         0
## 3 2020-01-22      Chongqing Mainland China         6      0         0
## 4 2020-01-22         Fujian Mainland China         1      0         0
## 5 2020-01-22          Gansu Mainland China         0      0         0
## 6 2020-01-22      Guangdong Mainland China        26      0         0
```

# Missing Data and Duplicate Data

To ensure the upcoming analysis, missing data and duplicate data are undesired. As a result, we try to discover whether those data exist in the data frame.

```
all(is.na(data$Country.Region), is.na(data$Confirmed), is.na(data$Deaths), is.na(data
$Recovered))
```

```
## [1] FALSE
```

```
all(duplicated(data))
```

```
## [1] FALSE
```

Since both of the above outputs return “FALSE”, it shows that the given dataset does not contain any missing data or duplicate data.

# Description of the Dataset

To grab more information from the data frame, we are going to study each of the variables one by one.

```
names(data)
```

```
## [1] "Date"           "Province.State" "Country.Region" "Confirmed"
## [5] "Deaths"         "Recovered"
```

```
dim(data)
```

```
## [1] 10671     6
```

There are 6 variables in total and altogether 10671 observations are recorded from 22/01/2020 to 31/03/2020.

```
class(data$Date)
```

```
## [1] "Date"
```

“Date” represents the date of observation being recorded for a particular region. It is formatted as mm/dd/yyyy and “Date” in variable type.

```
class(data$Province.State)
```

```
## [1] "factor"
```

```
class(data$Country.Region)
```

```
## [1] "factor"
```

```
head(unique(data$Province.State))
```

```
## [1] Anhui     Beijing   Chongqing Fujian    Gansu     Guangdong
## 295 Levels:   Montreal, QC  Norfolk County, MA Alabama ... Zhejiang
```

```
head(unique(data$Country.Region))
```

```
## [1] Mainland China Hong Kong      Macau          Taiwan         US
## [6] Japan
## 216 Levels:  Azerbaijan ('St. Martin',) Afghanistan Albania Algeria ... Zimbabwe
```

“Province.State” and “Country.Region” specify a particular state and region of observation. They are in “factor” type. Up to 31/03/2020, there are already 216 countries and regions suffer from the novel coronavirus.

```
class(data$Confirmed)
```

```
## [1] "numeric"
```

```
class(data$Deaths)
```

```
## [1] "numeric"
```

```
class(data$Recovered)
```

```
## [1] "numeric"
```

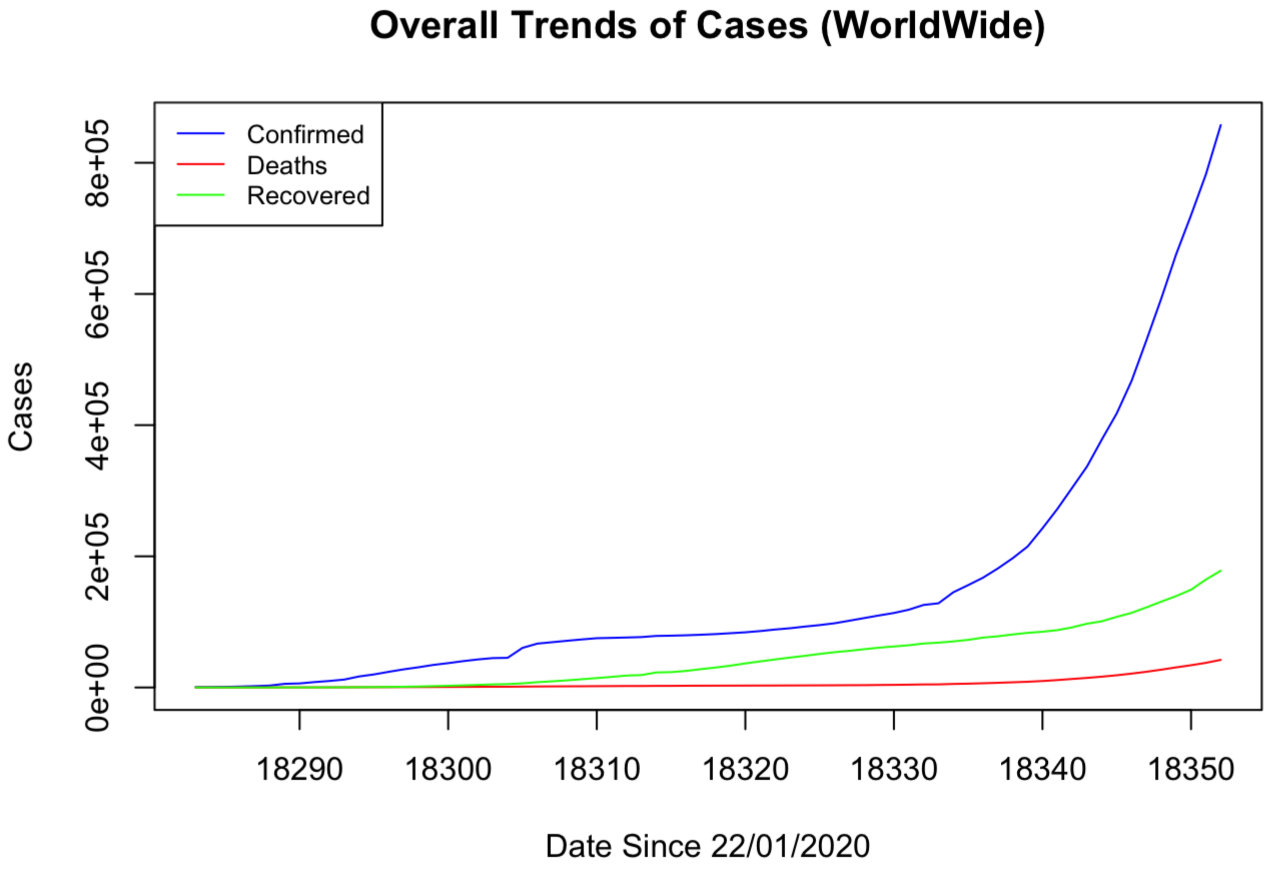
“Confirmed” corresponds to the number of confirmed cases. “Deaths” denotes the number of people who lose their lives from the novel coronavirus. “Recovered” describes the number of patients recovered. Their data type matches our intuition, which is all in the “numeric” type.

# Visualisation for the Overview of Number of Confirmed, Deaths and Recovered Cases (Worldwide)

We use a line plot here to describe the trends of the number of confirmed, deaths and recovered cases.

```
# Pre-preparation for the line plot
tempCum <- data.frame(as.character(data$Date), data$Confirmed, data$Deaths, data$Reco
vered)
names(tempCum) <- c("Date", "Confirmed", "Deaths", "Recovered")

cumulativeConfirmed <- aggregate(tempCum$Confirmed, by = list(tempCum$Date), sum)
names(cumulativeConfirmed) <- c("Date", "Confirmed")

cumulativeDeaths <- aggregate(tempCum$Deaths, by = list(tempCum$Date), sum)
names(cumulativeDeaths) <- c("Date", "Deaths")

cumulativeRecovered <- aggregate(tempCum$Recovered, by = list(tempCum$Date), sum)
names(cumulativeRecovered) <- c("Date", "Recovered")

cumulativeCases <- cbind(cumulativeConfirmed, cumulativeDeaths$Deaths, cumulativeReco
vered$Recovered)
```

```
matplot(as.Date(cumulativeConfirmed$Date), cumulativeCases[,-1], xlab = "Date Since 2
2/01/2020", ylab = "Cases", type = "l", col = c("blue", "red", "green"), lty = 1, mai
n = "Overall Trends of Cases (WorldWide)")
legend("topleft", legend = c("Confirmed", "Deaths", "Recovered"), col = c("blue", "re
d", "green"), lty = 1, cex = 0.8)
```

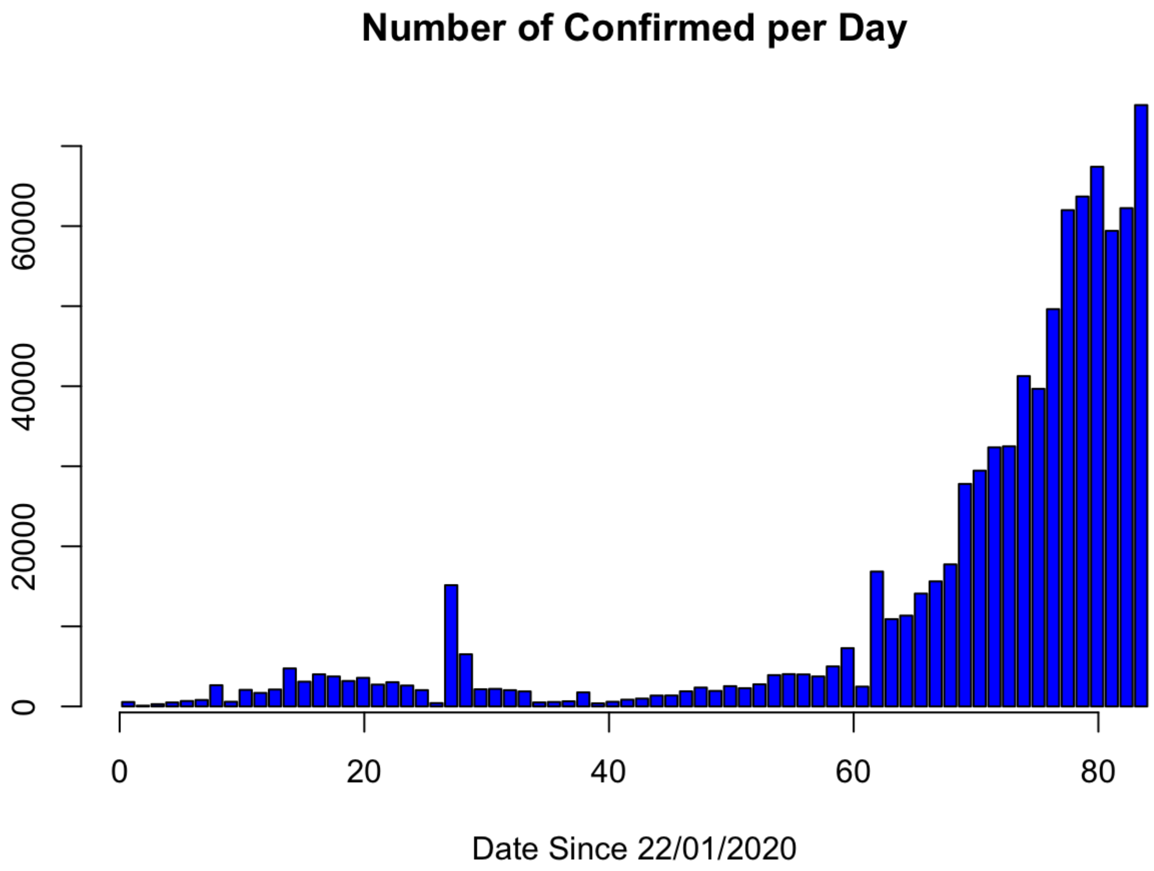


We can see that the number of confirmed cases is incredibly increasing in the end. A piece of good news is that the number of recovered patients is also increasing, but the rate of increment is not as fast as the increment in confirmed cases. Thankfully, the growth of the number of deaths is not very significant at the moment in 31/03/2020.

To grasp more details on the data, we are going to use bar charts to illustrate the number of confirmed, deaths and recovered cases on each day.

```
# Pre-preparation for the confirmed cases
confirmedPerDay = c()
confirmedPerDay[1] = cumulativeConfirmed$Confirmed[1]

for(i in 2:length(cumulativeConfirmed$Confirmed)) {
  confirmedPerDay = c(confirmedPerDay, cumulativeConfirmed$Confirmed[i] - cumulativeC
onfirmed$Confirmed[i - 1])
}
```

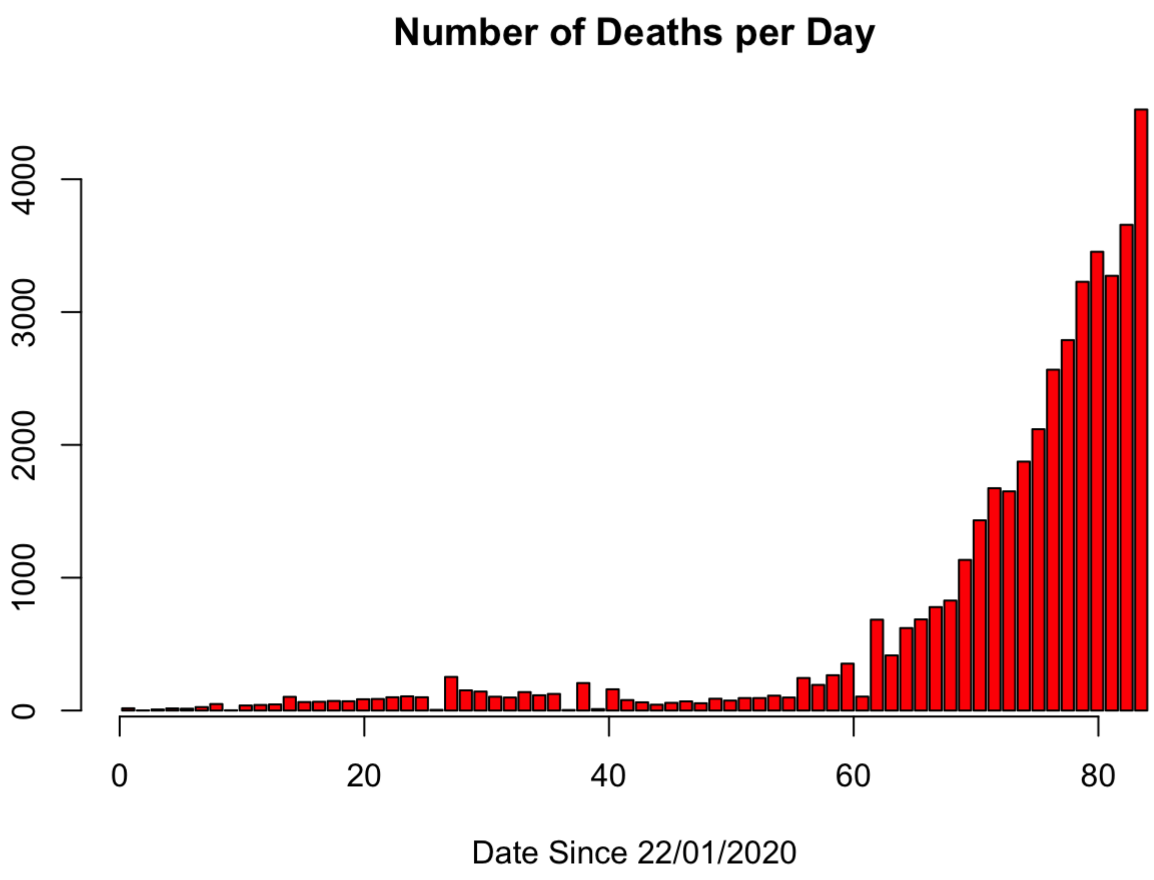
```
barplot(confirmedPerDay, xlab = "Date Since 22/01/2020", col = "blue", main = "Number
of Confirmed per Day")
axis(cumulativeConfirmed$Date)
```



In reviewing the bar chart, it gives us an insight into a sudden increment in the number of confirmed case is all most at the right tail of the graph. We may interpret such a fact that in the next half month in March, the novel coronavirus became a threat to human beings. More and more people are suffering from the novel coronavirus, such as getting cough, fever and tiredness.

```
# Pre-preparation for the deaths cases
deathsPerDay = c()
deathsPerDay[1] = cumulativeDeaths$Deaths[1]

for(i in 2:length(cumulativeDeaths$Deaths)) {
  deathsPerDay = c(deathsPerDay, cumulativeDeaths$Deaths[i] - cumulativeDeaths$Deaths
[i - 1])
}
```

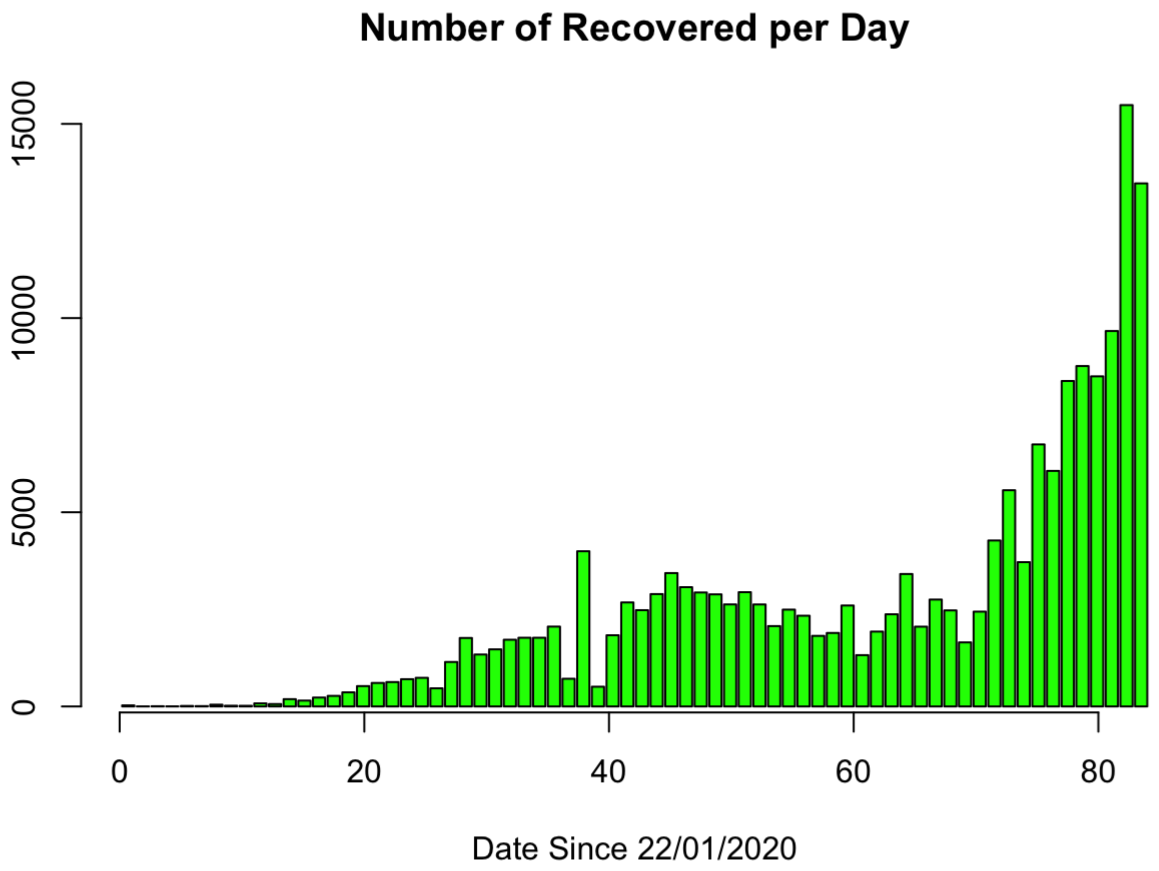
```
barplot(deathsPerDay, xlab = "Date Since 22/01/2020", col = "red", main = "Number of
 Deaths per Day")
axis(cumulativeDeaths$Date)
```



A piece of sad news is that the number of people losing their lives in also increasing every day. Especially in the peak of the number of confirmed cases in the last bar chart, the number of deaths also goes to the peak. Yet, we do not know if it is really a peak such data since the data are only updated to the end of March. One thing we can sure is that the novel coronavirus is intimidating our lives.

```
# Pre-preparation for the recovered cases
recoveredPerDay = c()
recoveredPerDay[1] = cumulativeRecovered$Recovered[1]

for(i in 2:length(cumulativeRecovered$Recovered)) {
  recoveredPerDay = c(recoveredPerDay, cumulativeRecovered$Recovered[i] - cumulativeRecovered$Recovered[i - 1])
}
```

```
barplot(recoveredPerDay, xlab = "Date Since 22/01/2020", col = "green", main = "Number of Recovered per Day")
axis(cumulativeRecovered$Date)
```

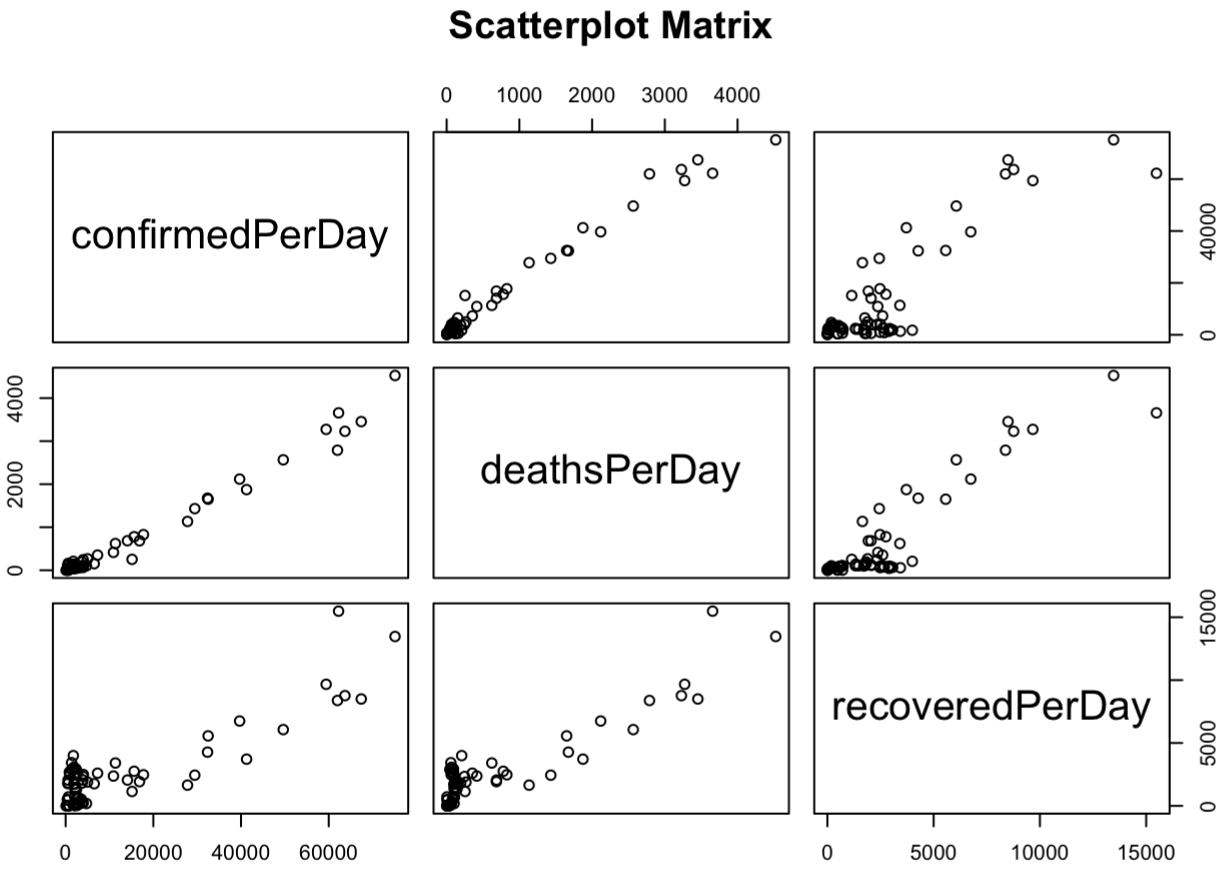


We should not dwell on but look forward to a better place. Patients are getting recovered as well at the end of March. We may notice that there is a little triangle in the middle on the graph, in about the period of February. Ideally, the number of recovered cases would decrease due to the controlled situation. However, from the previous graphs, we understand that the number of confirmed and deaths are heightening in March. One explanation on the little triangle is the novel coronavirus has been under controlled in February, but it is out of control again during the next half of March.

# Correlations between the Number of Confirmed, Deaths and Recovered Cases (Worldwide)

The correlations among 3 types of cases are good representations of the current situation of the novel coronavirus. Instead of computing the correlations, we are going to visualise them with the use of a scatterplot matrix.

```
pairs(~confirmedPerDay+deathsPerDay+recoveredPerDay, main="Scatterplot Matrix")
```



From the above scatterplot, we can see that the association between the number of confirmed, deaths and recovered cases is positively correlated. We may interpret such phenomenons is that the number of deaths and recovered cases increases as the rise in the number of confirmed cases. We may expect such relationships because the previous graphs have given us some hints on it, that is the number of cases is increasing simultaneously as time passes. Under the fear of the novel coronavirus, the above diagrams would clarify the actual condition around the world.

# The Severest Countries and Regions

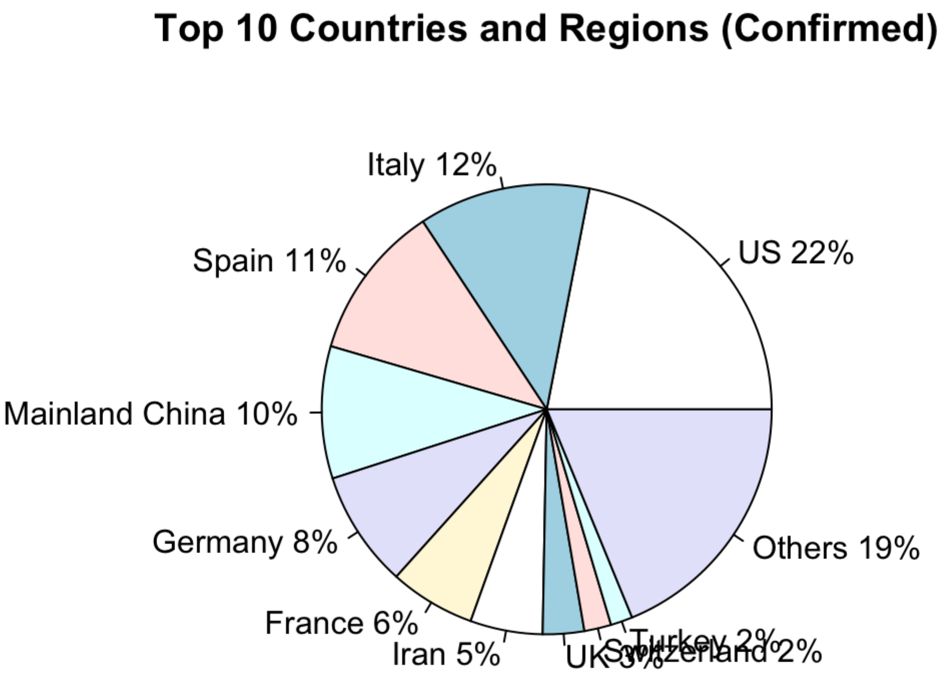
One of the most concern must be the severest countries and regions. Having such an idea, we are going to find out which countries and regions are suffering severely from the novel coronavirus. As we have mentioned previously, the last observation is only up to the end of March. The investigation of the following analysis is biased to the date of 31/03/2020.

```
lastCases <- data[data$Date == "2020-03-31", c(1, 3:6)]
```

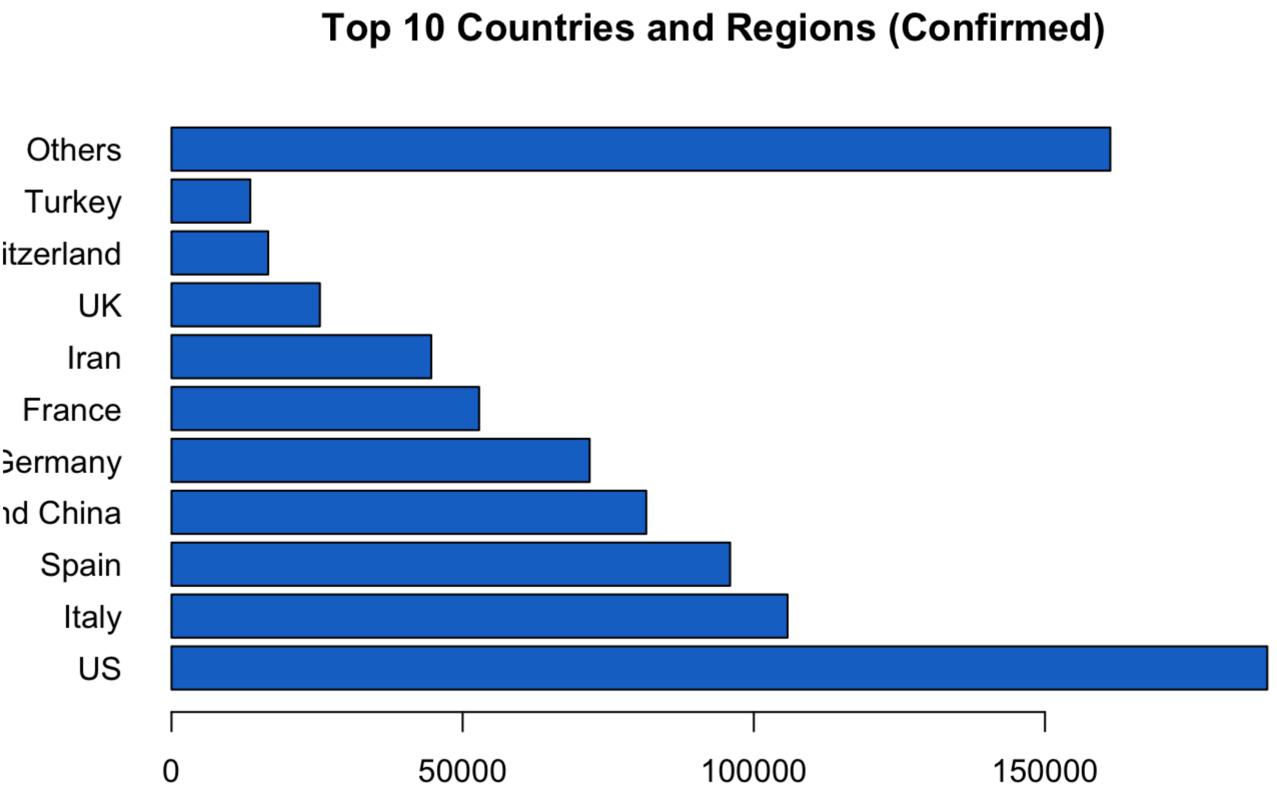
```
# Pre-preparation for the severest countries and region (confirmed cases)
lastCasesConfirmed <- aggregate(lastCases$Confirmed, by = list(lastCases$Country.Regi
on), sum)
names(lastCasesConfirmed) <- c("Region", "Confirmed")

severestRegionConfirmed <- lastCasesConfirmed[order(lastCasesConfirmed$Confirmed, dec
reasing = TRUE), ]

severestRegionTopTenConfirmed <- c(severestRegionConfirmed$Confirmed[1:10], sum(sever
estRegionConfirmed$Confirmed[11:length(severestRegionConfirmed$Confirmed)]))
```

```
severestRegionTopTenConfirmedRegion <- c(as.character(severestRegionConfirmed$Region[
1:10]), "Others")
```

```
pct <- round(severestRegionTopTenConfirmed / sum(severestRegionTopTenConfirmed) * 100
)
lblsConfirmed <- paste(paste(severestRegionTopTenConfirmedRegion, pct), "%", sep = ""
)
pie(severestRegionTopTenConfirmed, labels = lblsConfirmed, main = "Top 10 Countries a
nd Regions (Confirmed)")
```



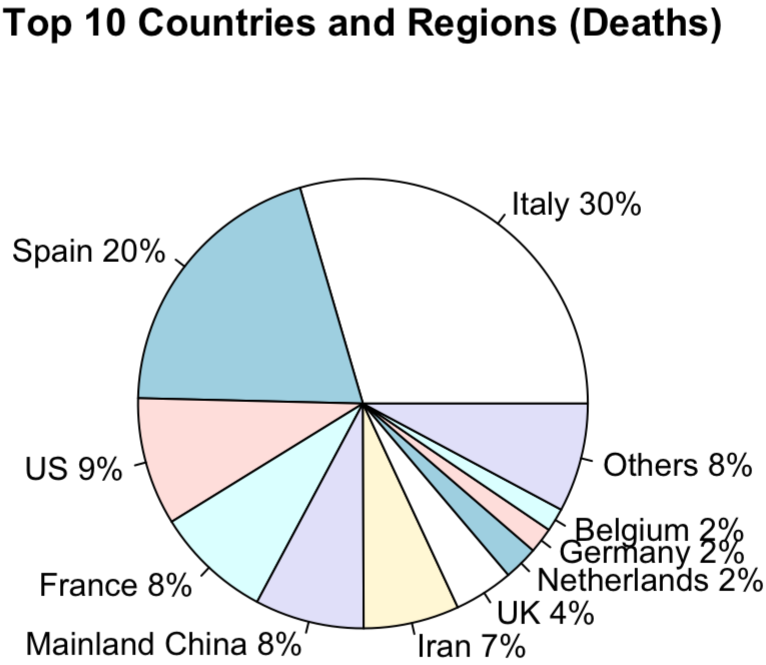
```
barplot(severestRegionTopTenConfirmed, main = "Top 10 Countries and Regions (Confirme
d)", horiz = TRUE, col = "dodgerblue3", names.arg = severestRegionTopTenConfirmedRegi
on, las = 1)
```



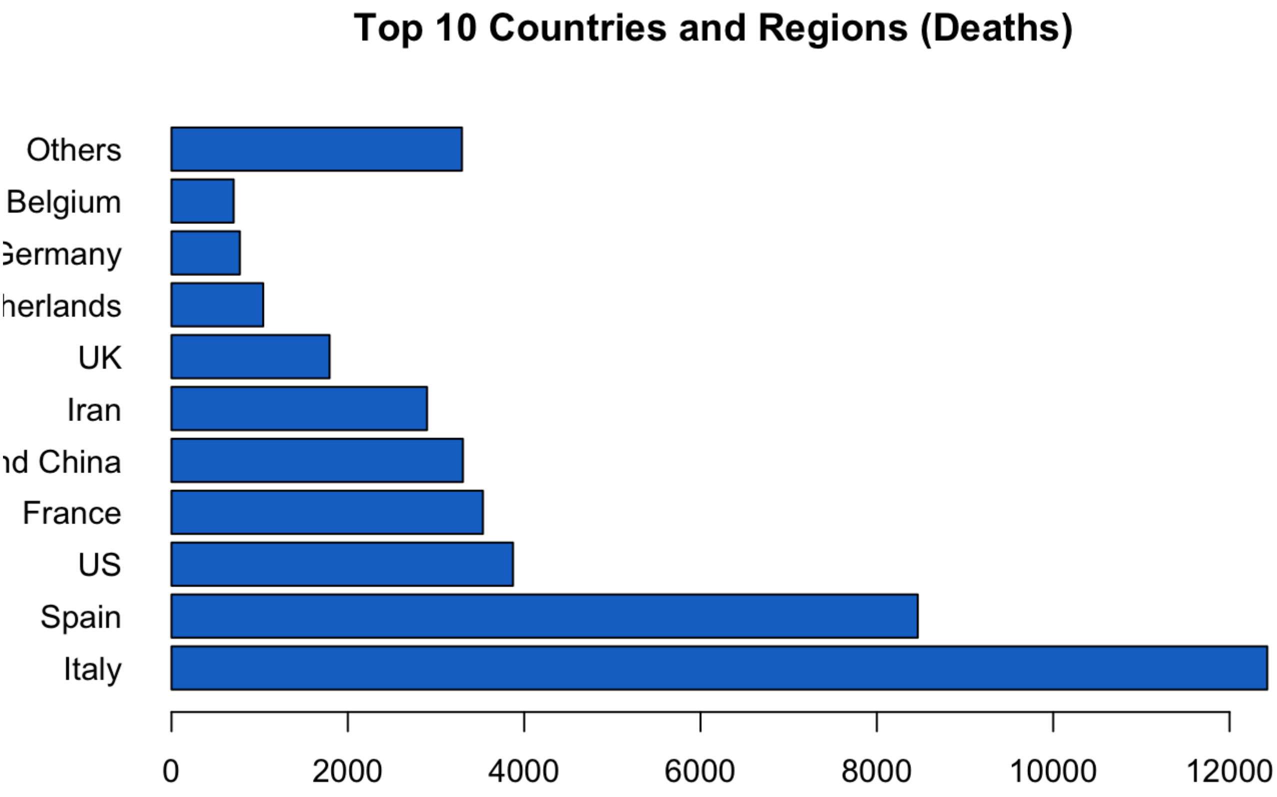
The top 3 severest countries and regions suffer from the novel coronavirus are “US”, “Italy” and “Spain”, 22%, 12%, and 11% respectively. One may raise a question here. The novel coronavirus originated from “Mainland China” but it only ranks in number 4 from other countries and regions. In retrieving the recent news from foreign countries, they recommend citizens not to wear a mask while hanging out. The lack of consensus on the content of public health may cause such a result.

```
# Pre-preparation for the severest countries and region (deaths cases)
lastCasesDeaths <- aggregate(lastCases$Deaths, by = list(lastCases$Country.Region), s
um)
names(lastCasesDeaths) <- c("Region", "Deaths")

severestRegionDeaths <- lastCasesDeaths[order(lastCasesDeaths$Deaths, decreasing = TR
UE), ]

severestRegionTopTenDeaths <- c(severestRegionDeaths$Deaths[1:10], sum(severestRegion
Deaths$Deaths[11:length(severestRegionDeaths$Deaths)]))
severestRegionTopTenDeathsRegion <- c(as.character(severestRegionDeaths$Region[1:10
]), "Others")
```

```
pct <- round(severestRegionTopTenDeaths / sum(severestRegionTopTenDeaths) * 100)
lblsDeaths <- paste(paste(severestRegionTopTenDeathsRegion, pct), "%", sep = "")
pie(severestRegionTopTenDeaths, labels = lblsDeaths, main = "Top 10 Countries and Reg
ions (Deaths)")
```



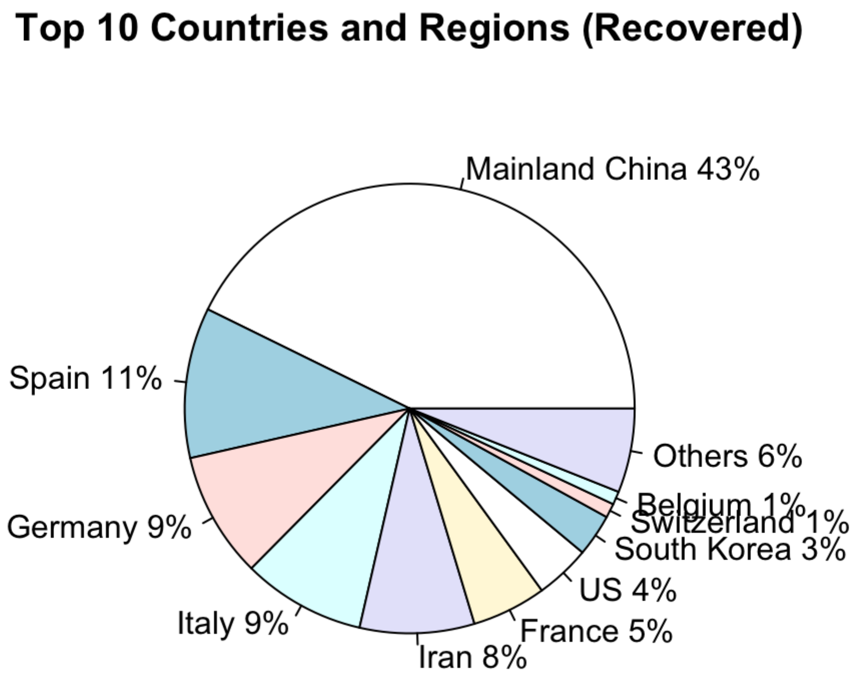
```r
barplot(severestRegionTopTenDeaths, main = "Top 10 Countries and Regions (Deaths)", horiz = TRUE, col = "dodgerblue3", names.arg = severestRegionTopTenDeathsRegion, las = 1)
```



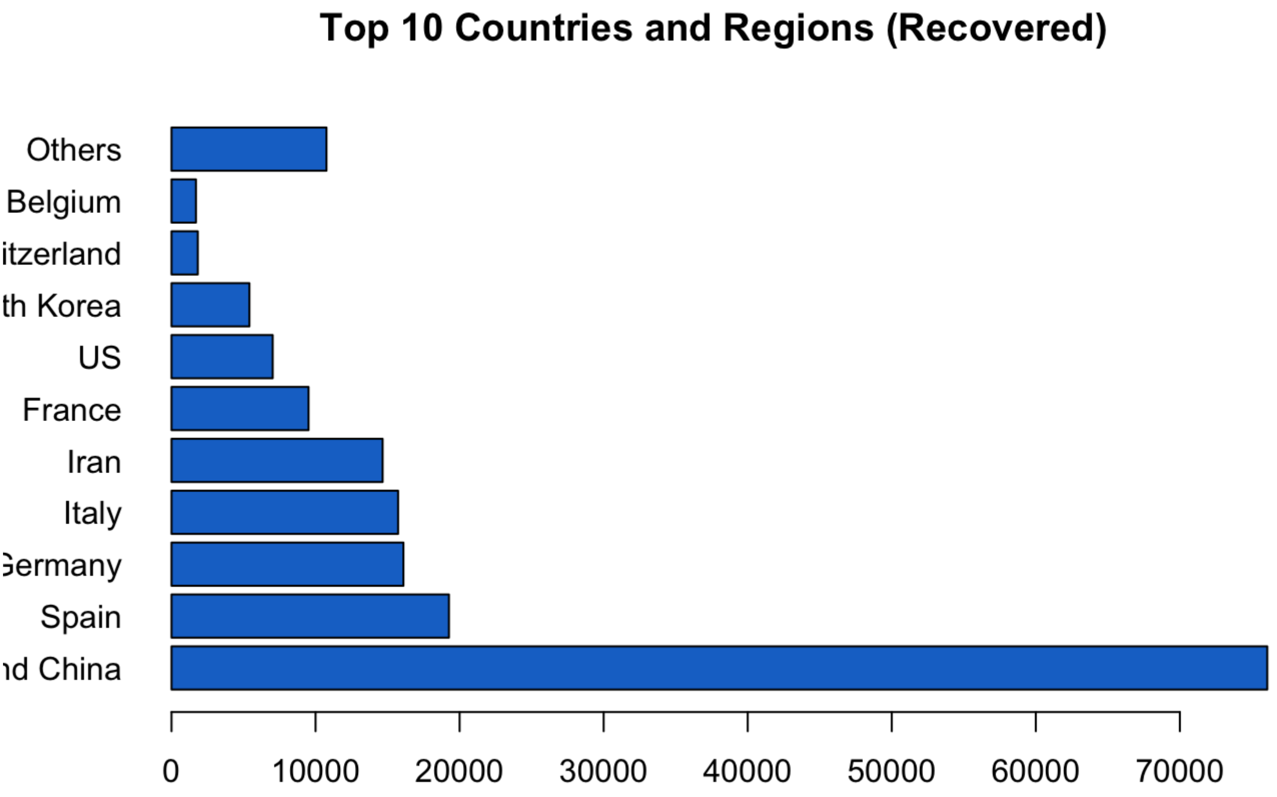
“Italy” has the highest losses during the fight with the novel coronavirus. “Spain” and “US” are also ranked in front of other countries and regions. Such ranking seems reasonable as they have been placed in the 3 top in the previous ranking. Having a large base case of confirmed patients, more deaths cases is prospective. Even we without a belief in sharing the same probability of death by the novel coronavirus, the treatments they have are mainly resulting in such a phenomenon. The hospital in foreign countries are packed, and they did not have good separation between patients. Such information from the news may explain why foreign countries have a higher rate of deaths.

```
# Pre-preparation for the severest countries and region (recovered cases)
lastCasesRecovered <- aggregate(lastCases$Recovered, by = list(lastCases$Country.Regi
on), sum)
names(lastCasesRecovered) <- c("Region", "Recovered")

severestRegionRecovered <- lastCasesRecovered[order(lastCasesRecovered$Recovered, dec
reasing = TRUE), ]

severestRegionTopTenRecovered <- c(severestRegionRecovered$Recovered[1:10], sum(sever
estRegionRecovered$Recovered[11:length(severestRegionRecovered$Recovered)]))
severestRegionTopTenRecoveredRegion <- c(as.character(severestRegionRecovered$Region[
1:10]), "Others")
```

```
pct <- round(severestRegionTopTenRecovered / sum(severestRegionTopTenRecovered) * 100
)
lblsRecovered <- paste(paste(severestRegionTopTenRecoveredRegion, pct), "%", sep = ""
)
pie(severestRegionTopTenRecovered, labels = lblsRecovered, main = "Top 10 Countries a
nd Regions (Recovered)")
```



```r
barplot(severestRegionTopTenRecovered, main = "Top 10 Countries and Regions (Recovered)", horiz = TRUE, col = "dodgerblue3", names.arg = severestRegionTopTenRecoveredRegion, las = 1)
```



The novel coronavirus has been affecting the world for at least 4 months. The number of people who inflected by it is also getting recovered by time. “Mainland China” has the largest number of recovered patients around the world. Yet, “Mainland China” is not arranged in the top 3 countries and regions of confirmed cases, but it still has a surprising statistic on it. Even some professionals from the WHO would say they want to have treatment in China if they got infected. What a joke. Without a doubt, the above graphs are biased on the given data. Here, we assume the data are dependable. Thus, we could only say “Mainland China” is outstanding in helping patients to recover from the novel coronavirus.

# Mortality Percentage and Recovery Percentage

If we got inflected by the novel coronavirus accidentally, we may worry about the likelihood of death or recovery based on historical data. It is a common concern for us to understand the percentage for both the mortality and the recovery in different countries and regions. To have a better understanding of them, we construct a table to display the corresponding mortality and the recovery rate for the top 10 countries and regions and, of course, Hong Kong. The formula of the mortality percentage and the recovery percentage is simply to divide the number of deaths or recovered cases by the number of confirmed cases, such that $deaths_i/confirmed_i$ and $recovered_i/confirmed_i$.

```
# Pre-preparation for the mortality percentage and the recovery percentage
mortalityPercentage <- paste(round(lastCasesDeaths$Deaths / lastCasesConfirmed$Confirmed * 100, 2), "%", sep = "")
recoveryPercentage <- paste(round(lastCasesRecovered$Recovered / lastCasesConfirmed$Confirmed * 100, 2), "%", sep = "")
overallPercentage = cbind(lastCasesConfirmed, lastCasesDeaths$Deaths, lastCasesRecovered$Recovered, mortalityPercentage, recoveryPercentage)
names(overallPercentage) <- c("Countries or Regions", "Confimred", "Deaths", "Recovered", "Mortality Percentage", "Recovery Percentage")
```

```
overallPercentage <- overallPercentage[order(overallPercentage$Confimred, decreasing
 = TRUE), ]

overallPercentage <- rbind(overallPercentage[1:10, ], overallPercentage[overallPercen
tage$"Countries or Regions" == "Hong Kong", ])
```

```
# Reference "Create stylish tables in R using formattable" from https://www.littlemis
sdata.com/blog/prettytables
library(formattable)

badGreen <- "#DeF7E9"
goodGreen <- "#71CA97"
badRed <- "#ff7f7f"
goodRed <- "#ffb3b3"

formattable(overallPercentage, align = c("l", "c" ,"c" ,"c" ,"r", "r"), list("Region"
= formatter("span", style = ~ style(color = "grey",font.weight = "bold")), "Mortality
Percentage"= color_tile(goodRed, badRed), "Recovery Percentage"= color_tile(badGreen,
goodGreen)))
```

| | **Countries or Regions** | **Confimred** | **Deaths** | **Recovered** | **Mortality Percentage** | **Recovery Percentage** |
|---|---|---|---|---|---|---|
| 176 | US | 188172 | 3873 | 7024 | 2.06% | 3.73% |
| 84 | Italy | 105792 | 12428 | 15729 | 11.75% | 14.87% |
| 156 | Spain | 95923 | 8464 | 19259 | 8.82% | 20.08% |
| 104 | Mainland China | 81524 | 3305 | 76068 | 4.05% | 93.31% |
| 64 | Germany | 71808 | 775 | 16100 | 1.08% | 22.42% |
| 60 | France | 52827 | 3532 | 9513 | 6.69% | 18.01% |
| 80 | Iran | 44605 | 2898 | 14656 | 6.5% | 32.86% |
| 172 | UK | 25481 | 1793 | 179 | 7.04% | 0.7% |
| 161 | Switzerland | 16605 | 433 | 1823 | 2.61% | 10.98% |
| 170 | Turkey | 13531 | 214 | 243 | 1.58% | 1.8% |
| 75 | Hong Kong | 714 | 4 | 128 | 0.56% | 17.93% |

The above table shows the observation of top 10 confirmed countries and region, also Hong Kong. The table includes the number of confirmed, deaths and recovered cases along with the mortality percentage and the recovery percentage. Those percentages are coloured based on the performance on the mortality and the recovery. “Mainland China”, for example, has the highest recovery percentage among 11 countries and regions. As a result, the cell is coloured in deep green. “UK” has the lowest recovery percentage, so it is in light green. It behaves similarly in the mortality percentage. Again, the above data only reflect the condition on 31/03/2020.

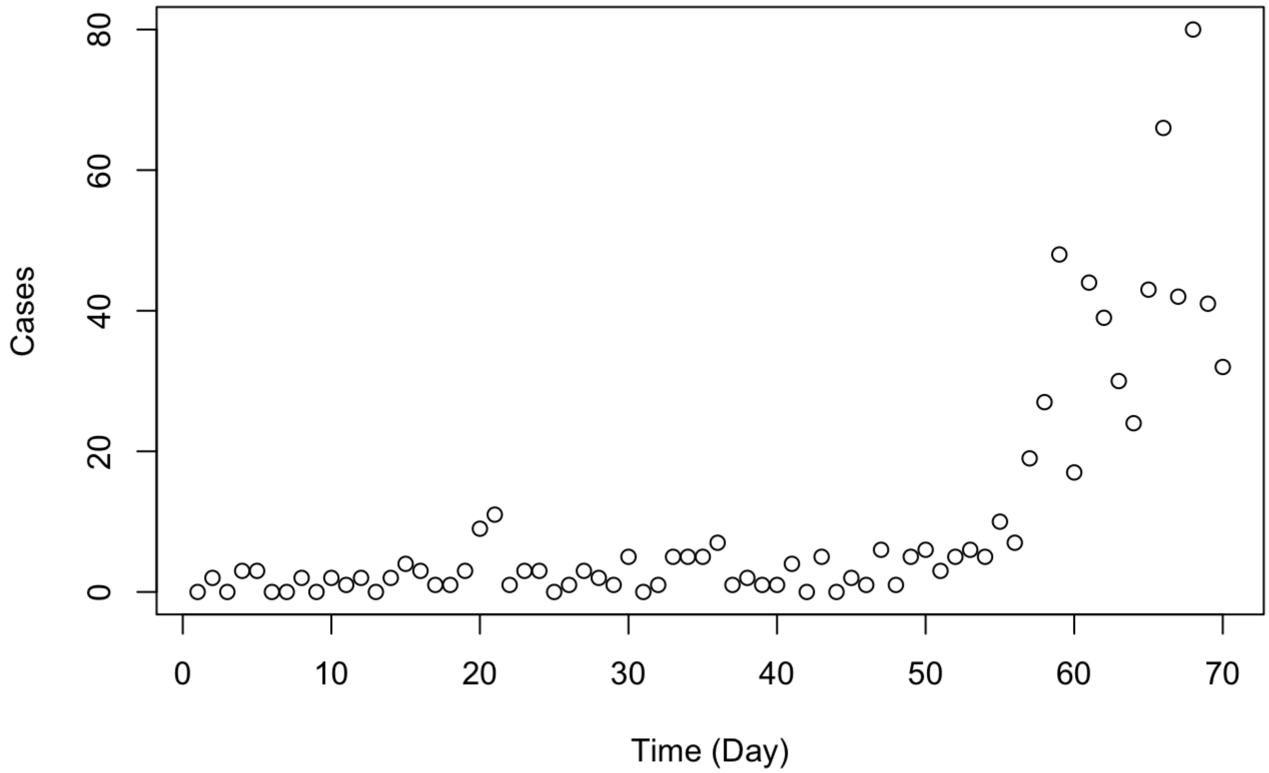
# Hong Kong Data

As a Hong Kong local citizen, the most concern one must be the circumstance in Hong Kong. If we hang out in the current status, the majority of people are wearing masks and using a lot of instant hand sanitiser. Having experience in fighting with the Severe Acute Respiratory Syndrome in 2003, we understood the importance of public health. Hong Kong is a bullet place unquestionably. If we do not protect ourselves in such a way, we believe the situation in Hong Kong would be similar to Italy. Hence, we want to know the challenges that Hong Kong is facing. In order to obtain an overview of the effect of the novel coronavirus for the newly confirmed cases on each day, here we suggest using a regression model (the number of confirmed cases against time) to fit the data as a naive visualisation.

```r
# Pre-preparation for daily confirmed cases in Hong Kong
dataHongKong <- data[data$Country.Region == "Hong Kong", ]

confirmedPerDayHongKong <- c()
confirmedPerDayHongKong[1] <- dataHongKong$Confirmed[1]

deathsPerDayHongKong <- c()
deathsPerDayHongKong[1] <- dataHongKong$Deaths[1]

recoveredPerDayHongKong <- c()
recoveredPerDayHongKong[1] <- dataHongKong$Recovered[1]

for(i in 2:length(dataHongKong$Confirmed)) {
  confirmedPerDayHongKong <- c(confirmedPerDayHongKong, dataHongKong$Confirmed[i] - d
ataHongKong$Confirmed[i - 1])
  deathsPerDayHongKong <- c(deathsPerDayHongKong, dataHongKong$Deaths[i] - dataHongKo
ng$Deaths[i - 1])
  recoveredPerDayHongKong <- c(recoveredPerDayHongKong, dataHongKong$Recovered[i] - d
ataHongKong$Recovered[i - 1])
}
```

```r
# Pre-preparation for the simple linear regression against time
dayHongKong <- 1:length(confirmedPerDayHongKong)

regHongKong <- lm(confirmedPerDayHongKong~dayHongKong)
```

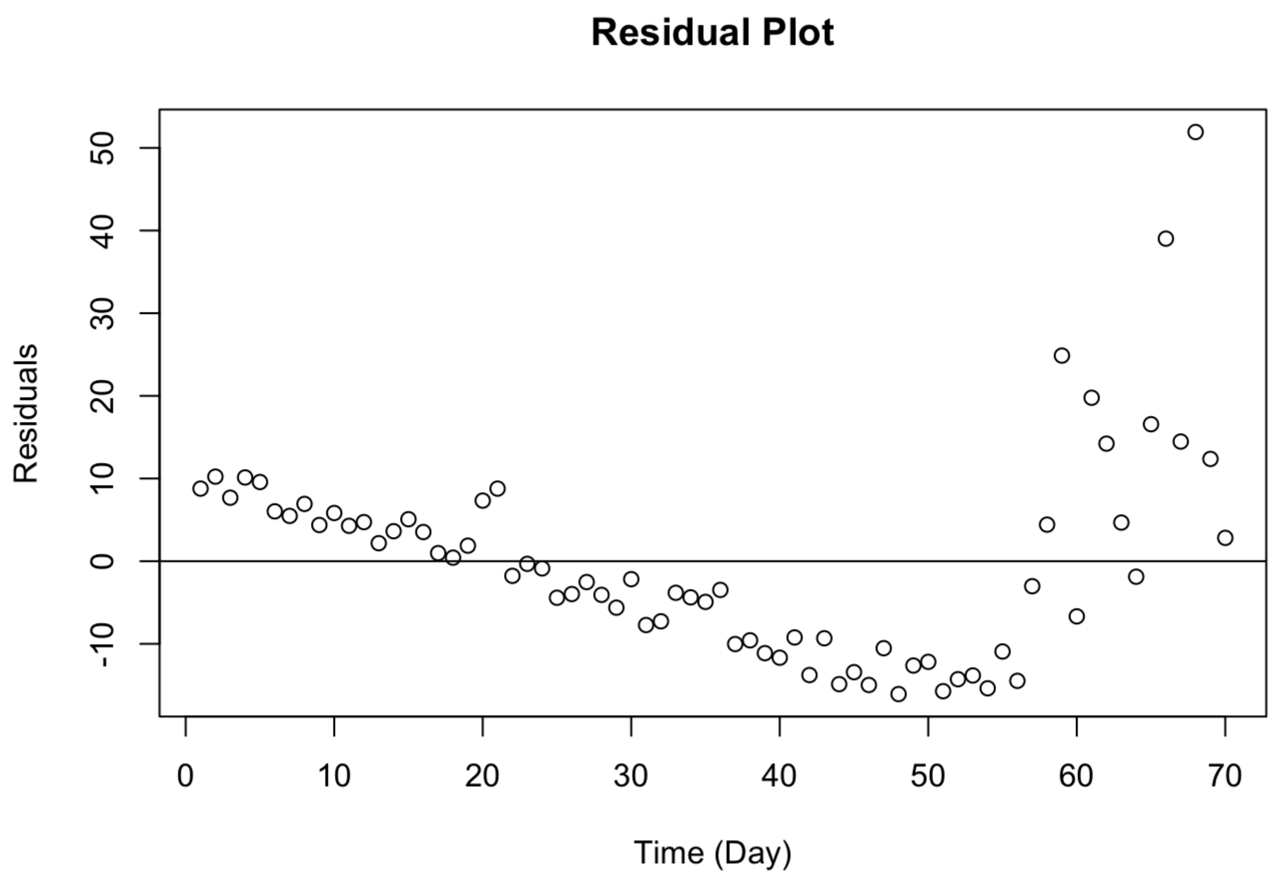
```r
# scatterplot for the number of confirmed cases against time
plot(dayHongKong, confirmedPerDayHongKong, xlab = "Time (Day)", ylab = "Cases")
```



The x-axis refers to the date from 22 January to 30 March 2020, whereas the y-axis indicates the number of newly confirmed cases. From the scatterplot, we can see that the association between Cases and Time is positively correlated. That is the number of newly confirmed cases increases as time passes. From the time of 0 to about 55, the number of newly confirmed cases is very stable. However, things are getting worse after it. About 20 to 80 cases are newly confirmed on a daily basis afterword. According to the recent news, most of the patients are those who came back from foreign countries. During the time in such period, we are very helplessness. We have no choice but to accept the impact made from that selfishness.

We may want to predict the number of cases afterwards, and a time series model (or a simple linear regression model against time) is a good method to predict the future number of newly confirmed cases. Yet, it is obvious that the number of newly confirmed cases is not linear. Simple linear regression is definitely not a good choice to adopt for studying the relationship between Cases and Time. To verify a simple linear regression model is not suitable for the given data, we construct a residual plot for the regression model $confirmedCases_i = \beta_0 + \beta_1 \times time_i + e_i$ .

```
resHongKong <- resid(regHongKong)

plot(dayHongKong, resHongKong, ylab = "Residuals", xlab = "Time (Day)", main = "Resid
ual Plot")
abline(0,0)
```

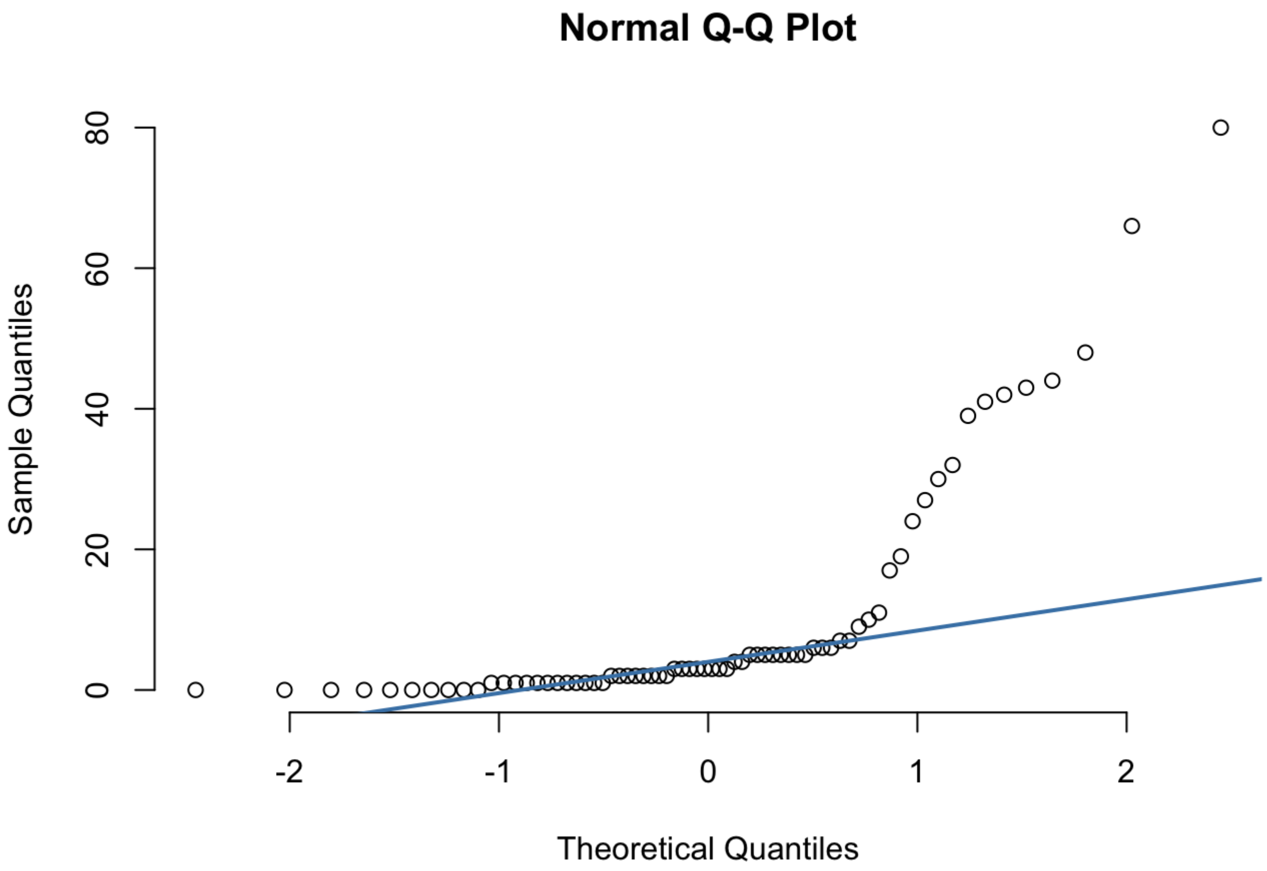


The assumption of linear regression is $E(e) = 0$, $Var(e) = \sigma^2 I_n$ and $e_i$ are uncorrelated. If the regression is well fitted, a null plot is expected. Having said that, the above residual plot does not look like a null plot at all. From the residual plot, we notice that the residuals follow a pattern known as positive autocorrelation. Such statistical phenomenon occurs when the residuals follow by others with the same sign. For instance, positive residuals followed by a positive one, whereas a negative one followed by a negative residual. A formal hypothesis testing using Durbin-Watson is recommended in the report of data analysis. The issue of autocorrelation would underestimate the mean square of error, which may lead to incorrect decisions in hypothesis testing. In other words, a simple linear regression model indeed is misspecified for the given set of data.

Still, there can be remedial solutions for the given set of data to investigate the future number of newly confirmed cases. For non-linear data, a polynomial regression with higher-order may help to explain more variation of data. Moreover, transformation is also favourable to such a situation. For example, modified power transformation $\Psi_M(y, \lambda)$ converts the response variable (the number of newly confirmed cases) with the same unit after transformation, such that data will approximate to a linear association. The choice of $\lambda$ can be further studied in the data analysis report by choosing with the minimum $RSS_\lambda$.

In the later works, we may be interested in studying the statistical inference based on the given data frame. For example, a hypothesis testing on the number of newly confirmed cases, such as $H_0 : \mu = c_0$ against $H_1 : \mu > c_0$, where $c_0$ is arbitrary. Those composited hypothesises require the sample data are normally distributed. As a result, we use a Q-Q plot to check whether the normality assumption of the daily confirmed cases in Hong Kong is satisfied.

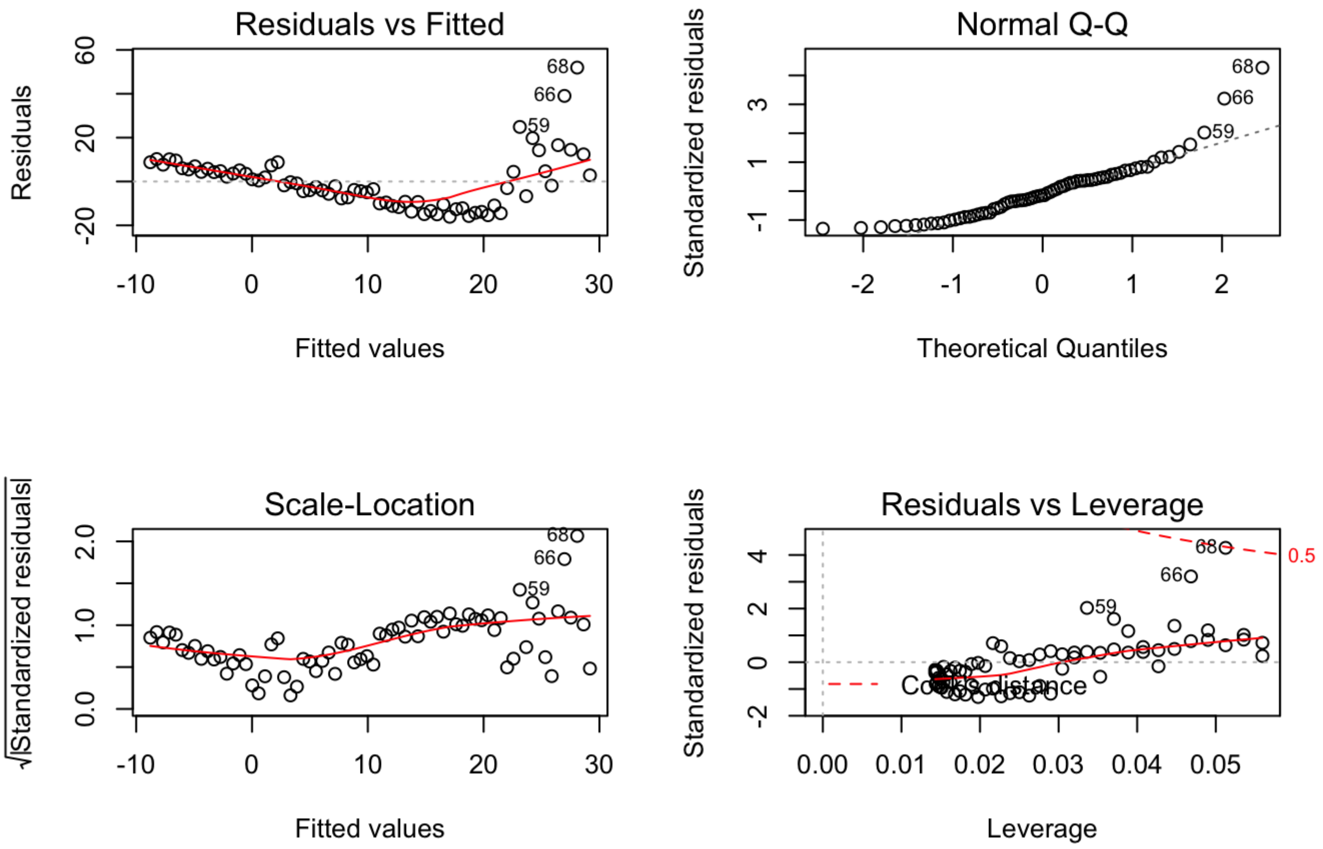
```
qqnorm(confirmedPerDayHongKong, pch = 1, frame = FALSE)
qqline(confirmedPerDayHongKong, col = "steelblue", lwd = 2)
```



The data are not concentrated with a 45-degree straight line. Thus, the given data is crystal clear that not normally distributed. If we still want to make a hypothesis testing from those data, the large sample size is desired such that we can adopt central limited theorem (CLT) to obtain an approximated normal distribution. However, in some sense of humanity, we do not want any more on those data. Besides, we do not know whether the number of confirmed cases is a random variable. Linking with the recent news, the novel coronavirus is asymptomatic which means data cannot reflect the true information. The given data did not record the patients who did not to the hospital.

As the simple linear regression is not being favoured in the given set of data, by saying that, we want to diagnose the nature of the data. That is retrieving any outliers, leverage points, and influential points in the Hong Kong data.

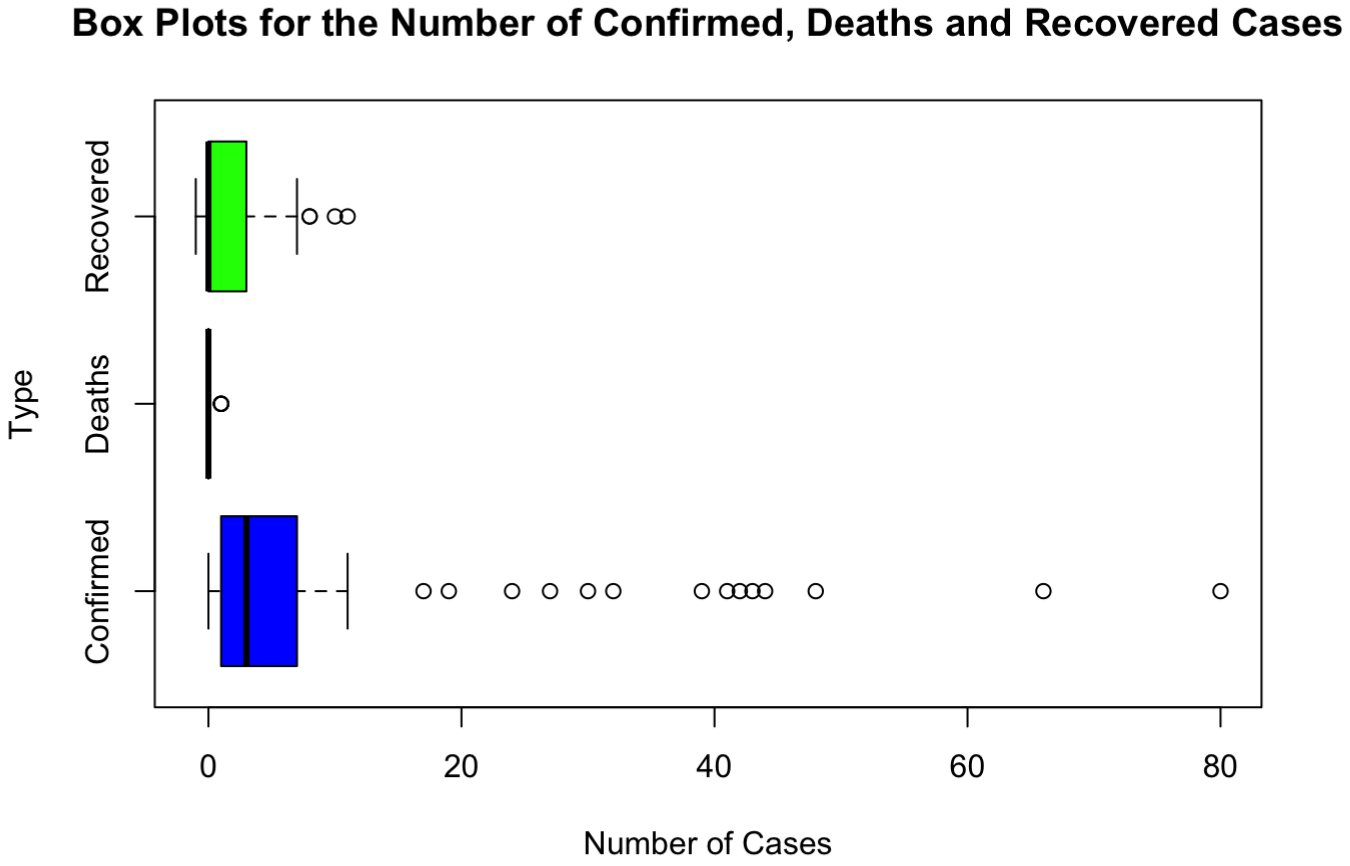
```
par(mfrow = c(2, 2))
plot(regHongKong)
```



The above influence diagnostics graphs are biased under the fit of simple linear regression. From the scale-location graph (the second-row first graph), it is obvious that there are a few points above the level of 1.645 in the y-axis, also known as the critical region. Those data points are said to be outliers in the dataset. Moreover, we can easily to notice that there are indeed influential points in the data from the residuals and leverage graph (the second-row second graph). The influential points are data points that fall inside the region of Cook's distance. Further numerical studies on discovering the nature of data are recommended. The Studentised residual, leverage, the difference in fits, the difference in betas, and Cook's distance are common approaches to diagnose outlier, leverage point, and influential point. We advise researchers to use the function "influence.measure()" for a better understanding of the nature of data along with different fitting on the regression models.

Box plot is an alternative way to check the distribution of a given dataset. To obtain a general view on the data, we plot the number of confirmed, deaths and recovered cases on the same graph.

```
maxCasesHongKong <- data.frame(c(confirmedPerDayHongKong, deathsPerDayHongKong, recov
eredPerDayHongKong), c(rep("Confirmed", length(confirmedPerDayHongKong)),rep("Deaths"
, length(deathsPerDayHongKong)),rep("Recovered", length(recoveredPerDayHongKong))))
names(maxCasesHongKong) <- c("Cases", "Types")

boxplot(maxCasesHongKong$Cases~maxCasesHongKong$Types, horizontal = TRUE, col = c("bl
ue", "red", "green"), xlab = "Number of Cases", ylab = "Type", main = "Box Plots for
 the Number of Confirmed, Deaths and Recovered Cases")
```



Outliers frequently appear in the above box plot. The distribution of each variable is known as right-skewed. In the view of statistical analysis, the dataset is not adorable to use since data do not fulfil the assumptions of normality. Regardless, we can make use of non-parametric tests to inspect the data for the presumption free of data. During the data analysis, we advise researchers to focus on non-parametric methods for hypothesis testing. For an illustration, the median can be tested using the sign test. Such hypothesis testing does not require the knowledge of sample distribution, which is more feasible for the given set of data.

# Possible use of the Dataset

Due to the limitation on the number of datasets can be used in the EDA report, further analysis can be done when more variables are available. Some suggestions on the future studies are the statistical dependency between other variables. For example, the correlation between the novel coronavirus and financial stability is a hot topic for sure. With a huge impact of the novel coronavirus, it has turned the economy down and many companies have decided to cut resources in this year. This motivates the research in the association between the unemployment rate and the novel coronavirus confirmed rate. In particular, the novel coronavirus is damaging many industries, such as the industry of airline service. “Lockdown” is a common approach for many countries and regions to avoid inputting the novel coronavirus and many airlines went out of service. By the theory of demand and supply, the oil price should decline in some sense. The study of the relationship between the number of confirmed cases and the oil price is also an interesting matter. Upon the conclusion, there are so many attractive topics can be studied in the future while more variables are available.

# Conclusion

In the EDA, it carries out an overview of the data of how the behaviour of the number of confirmed, deaths and recovery cases from 22/01/2020 to 31/03/2020. We used many different graphics to describe the overall situation and finding the severest countries and regions. The percentage of both mortality and recovery also give us a quick review of the situation in different countries and regions. From the given dataset, there are indeed some findings that do not match with our intuition, but we try to believe them to be true on the EDA report here. Although we do raise a lot of question about it, such as the recovery rate in Mainland China, we have no choice but to assume them to be correct and without holding out with us.

Moreover, we have gone through the details on the data of Hong Kong. The findings show that the data is not as perfect as we desired to construct statistical inference, given the fact that data are not normally distributed. In statistic courses, we learnt CLT is a reasonable method to approximate a normal distribution when the large sample size is available. Be honest, if we still have our heart of sympathy which emphasis by Mencius, a famous Chinese Confucian philosopher, we do not want any more data on it.

The major reason for the novel coronavirus has been a hot topic around the world is not only the fault of Mainland China but also the responsibility of the WHO. The WHO ridiculously underestimate the seriousness of the novel coronavirus. Yet, with different education also lead to such a heavy situation. For example, the foreign countries alarm their citizen not to wear a mask, even the workers in the hospital. The severe circumstance indeed made by ourselves. Because of the novel coronavirus, some of us have already lost their works, their relationships, and even their lives. Here we are not going to blame some of us, but we want to make a kindly remind all of us to stay safe. It is a hard time for all of us, and we believe we can pass it through the difficulties together.

# Reference

Donald G. (2020) *Restrictions Are Slowing Coronavirus Infections, New Data Suggest*. The NewYork Times. 30 March 2020. https://www.nytimes.com/2020/03/30/health/coronavirus-restrictions-fevers.html retrieved

Elizabeth C. and Karen Z. (2020) *Coronavirus: Hong Kong records 25 new cases, including two-month-old baby and clinic staff member; tally at 960*. South China Morning Post. 8 April 202. https://www.scmp.com/news/hong-kong/health-environment/article/3079010/coronavirus-hong-kong-records-25-new-cases retrieved

Ellis L. (2018) *Create stylish tables in R using formattable*. Littlemissdata. 24 September 2018. https://www.littlemissdata.com/blog/prettytables

Katherine H.C. (2020) *How people are spreading Covid-19 without symptoms*. Vox. 22 April 2020. https://www.vox.com/2020/4/22/21230301/coronavirus-symptom-asymptomatic-carrier-spread retrieved

SRK (2020) *Novel Corona Virus 2019 Dataset*. Kaggle. https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Tufekci Z. and Howard J. (2020) *The Real Reason to Wear a Mask*. The Atlantic. 22 April 2020. https://www.theatlantic.com/health/archive/2020/04/dont-wear-mask-yourself/610336/ retrieved

---

A work by Jack_KYCHAN

*1155119394@link.cuhk.edu.hk*