Analysis of the Effect of Stringency and HDI on COVID-19 Fatality Rate

Runtian 919013175 Group 10

2/4/2022

Content

Abstract

Since the first breakout in early 2020, COVID-19 has quickly spread to the entire world. It was declared as a pandemic by the World Health Organization (WHO) on March 11 2020 and more than 440 million people have been infected and about 6 millions deaths are recorded as of March 2022. In order to suppress the pandemic, many countries have mobilized resources to fight the pandemic and implemented policies like social distancing stay-at-home orders and migration restriction. However, despite pandemic control is a goal pursued globally, we have to admit that there is inescapable discrepancy in the power of countries which is resulted from many factors like regime, history and geographic locations. Our goal is to evaluate the significance of the effect of government interventions and the comprehensive background of the society on the impact of COVID-19.

To quantify the government intervention, we use Stringency Index which is defined by University of Oxford. The index records the strictness of 'lock-down style' policies that primarily restrict people's behavior. It is calculated using all ordinal containment and closure policy indicators, plus an indicator recording public information campaigns.

To quantify the social background of the country, we use Human Development Index (HDI) developed by Pakistani economist Mahbub ul Haq. As Figure 1 shows, this index is a comprehensive evaluation of life expectancy, education level and income level and it is adopted by United Nations Development Programme (UNDP)'s Human Development Report Office to measure the development of a country.

As for the variable to quantify the impact of COVID-19, we decide to use the fatality per million which is

$$FPM = \frac{\text{Number of Death Due to COVID-19}}{\text{Total Population (Million)}}$$

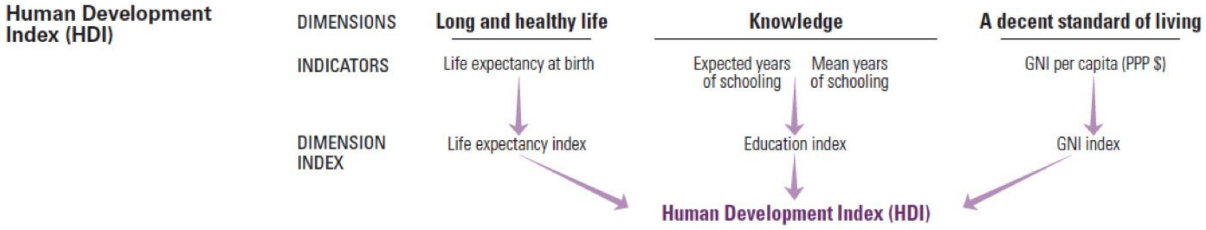


Figure 1: Figure 1.Decomposition of HDI

. This is because the absolute value is affected by the population of the region so it is more reasonable to measure the impact with a ratio.

We use k-Means clustering to categorize each country into 3 levels in terms of HDI and 4 levels in HDI and construct a Two-Way ANOVA model. The result shows that both factors (SI and HDI) are significant at 90% confidence level. This finding indicates that a strict and effective government intervention tends to reduce the FPM.

The datasets we used come from WHO and Our World in Data(OWID). Both datasets are publicly available.

Literature Review

A number of researchers have considered modeling the effect of government interventions and the society background of the country on the variables related to COVID-19. An intuitive understanding may be that the higher HDI, the less impact COVID-19 can have. This is confirmed by Georgiou (2021) who found that high HDI was helpful to reduce deaths caused by COVID-19. However Liu et al.(2020) concluded that there was an unexpected positive correlation between human development index and risk of infections and deaths of COVID-19 through research on Italy. He attributed this to the fact that people in high HDI regions were likely to have more than one chronic diseases. As for the stringency index, Ma et al.(2021) show early start of a high-level response is associated with early arrival of the peak number of daily new cases in their analysis on governments' Stringency Index in 148 countries.

Descriptive Analysis

We decide to only use the data in 2021 for following reasons:
(1) HDI is updated annually.
(2) The beginning date of COVID-19 transmission varies from country to country and the country took time to make responses which means data in 2020 may not fully reflect the true relationship between COVID-19 fatality and two explanatory variables.

Data Preprocessing

In WHO dataset, the data is classified by country and countries are encoded with ISO 2-letter code. However, in OWID dataset, the data is classified by location so it means some parts of a country are counted individually. Also, OWID uses 3-letter encoding system. Therefore, before analysis, we make some adjustments to two datasets to make them conformed. After filtering, there are 160 countries (cases). When it comes to Stringency Index, this index is updated from time to time so we decide to use the mean value for each country and those countries without a continuous Stringency Index record will be dropped.

Data Classification

We use k-Means clustering to classify Stringency Index into 4 groups from high(1) to low(4) and HDI into 3 groups from low(1) to high(3). The numbers of observations in each cell are displayed in the following table.

Numbers of Observations Table

Levels	1	2	3	4
1	12	18	13	2
2	1	23	25	11
3	4	11	16	20

EDA

Overview of COVID-19 in 2021

We observe that the average fatality per million is 381.2, the average stringency index is 53.78 and the average HDI is 0.74.

Descriptive Analysis Table

	Min	1st Qu.	Median	Mean	3rd Qu	Max
SI	7.28	44.74	53.78	52.37	62.06	77.09
FPM	0	63.65	381.2	643.01	891.32	3389.32
HDI	0.39	0.59	0.74	0.72	0.85	0.96

From Figure 2 we find that the FPM seems to increase as the SI decreases. Also, the HDI is likely to have impact on the average FPM.

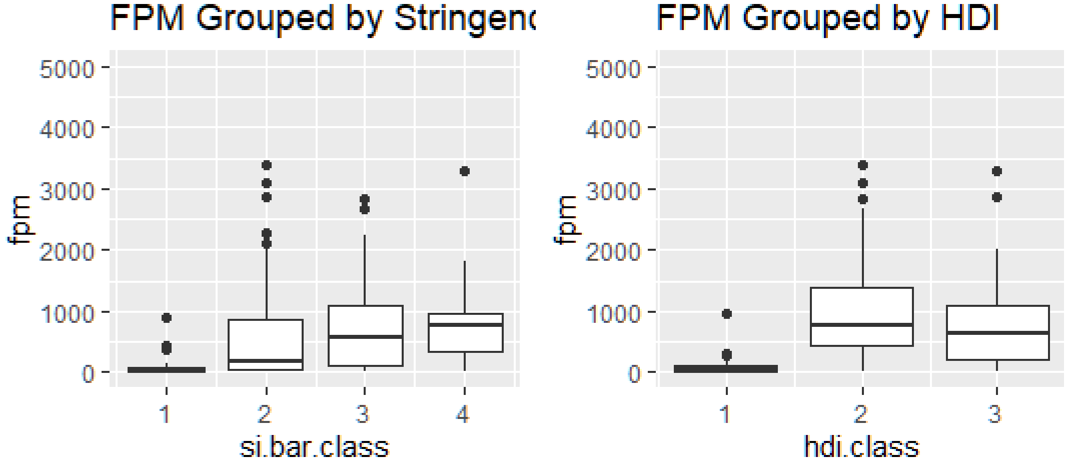


Figure 2: Figure 2.Boxplot of FPM by Group

The FPM seems to be strongly right-skewed, however, after logarithm transformation, it turns to be left-skewed. The distributions of SI and HDI don't show any specific pattern.

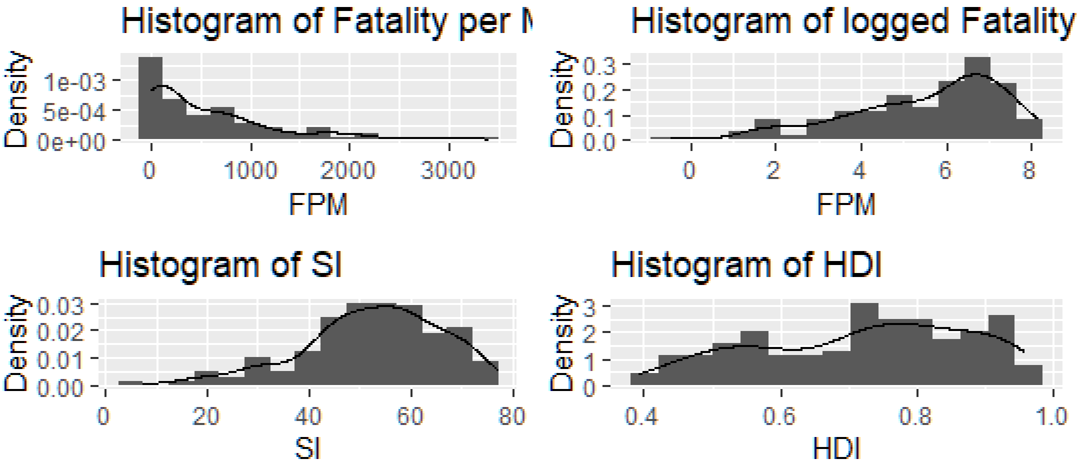


Figure 3: Figure 3.Histograms of Variables

Inferential Analysis

Introduction to Two-Way ANOVA Model

General Form

A general Two-Way ANOVA Model can be written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijk}$$

and

$$i = 1, ..., p, j = 1, ..., q, n_{ij}$$

Where Y_{ijk} is the k-th observation in (i,j) cell,
μ is grand mean effects,
α_i and β_j are effects of factor,
λ_{ij} is the interaction effects by two factors,
e_{ijk} is the random error components,
n_{ij} is the number of observations in cell (i, j).

In this case, both independent variables are supposed be random so we have
$\alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha^2)$
$beta_j \overset{iid}{\sim} N(0, \sigma_\beta^2)$
$e_{ijk} \overset{iid}{\sim} N(0, \sigma_e^2)$

Assumptions

Generally, there are 6 assumptions:
The dependent variable is continuous.
The independent variables should be independent and each variable consists of no less than 2 groups.
The observations should be independent.

No obvious outliers in each cell.
Residual of dependent variable's residuals should be approximately normal.
Variance of dependent variable's residuals should be heteroscedastic.

Test with Two-Way ANOVA Model

To examine if there is significant effect by independent variables on the dependent variable, we'd like to perform the following hypothesis testing.

$$I : H_0 : \alpha_1 = \alpha_2 = ... = \alpha_p = 0$$

$$II : H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0$$

$$III : H_0 : \lambda_1 = \lambda_2 = ... = \lambda_{pq} = 0$$

The corresponding test statistics are attained using Sum Squares of Error from full model and reduced model as follows:

$$F = \frac{(SSE_{reduced} - SSE_{full})/(df_{reduced} - df_{full})}{SSE_{full}/df_{full}}$$

Introduction to k-Means Clustering

k-Means clustering method aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The distance is usually gauged by Euclidean Distance which is

$$d_{ij}(2) = \{\sum_{t=1}^{n} |x_{it} - x_{jt}|^2\}^{1/2}$$

.

The algorithm proceeds by alternating between two steps with k initially assigned means

$$m_1^{(1)}, ..., m_k^{(1)}$$

.

Each time, the observation(x_p) is assigned to the cluster($S_i^{(t)}$) with the nearest mean, which is

$$S_i^{(t)} = \{x_p : |x_p - m_i^{(t)}|^2 \leq |x_p - m_j^{(t)}|^2, \forall\ 1 \leq j \leq k\}$$

.

Then, recalculate means (centroids) for observations assigned to each cluster by

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

.

Results

We perform k-Means clustering to transform Stringency Index (SI) and Human Development Index into categorical variables for easier interpretation. The following figure indicate that the optimal numbers of cluster in both variables are 3 and 4 respectively.

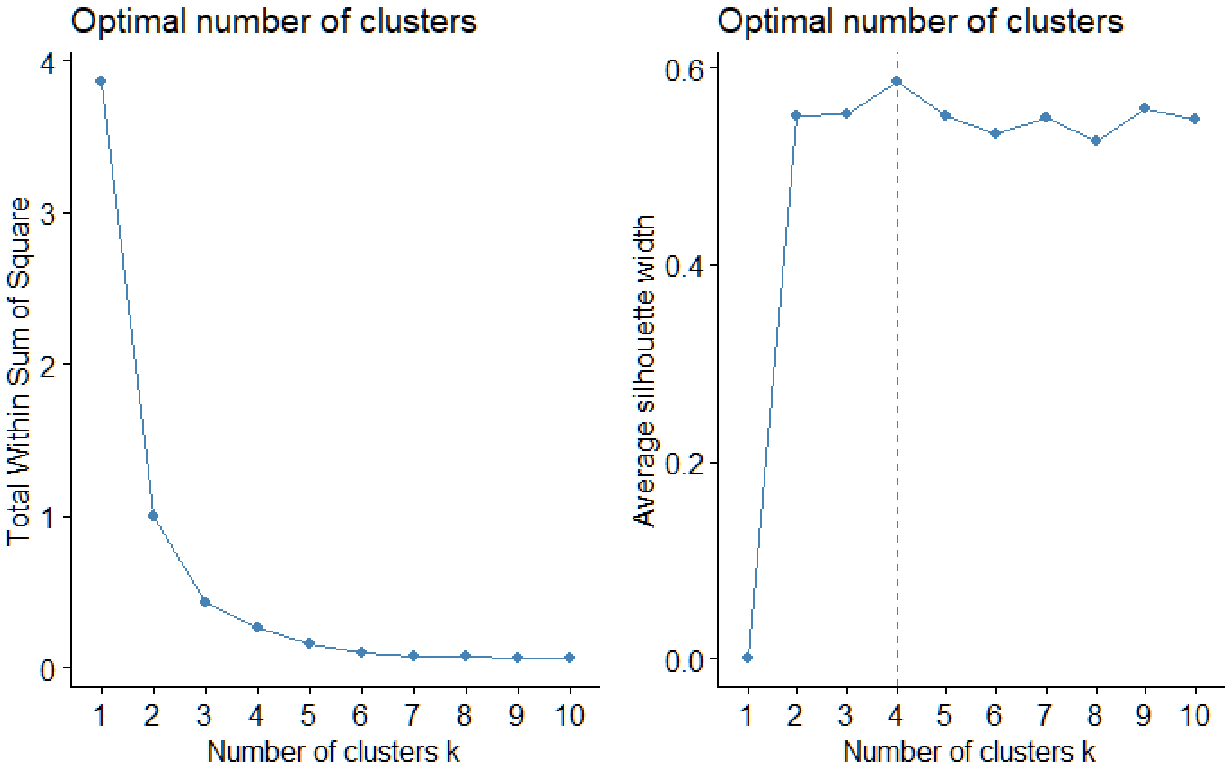


Figure 4: Figure 5. Optimal Number of Clusters

As for the Two-Way ANOVA model, we have the ANOVA table in which it is found that the p-values of two factors are less than 0.1.The p-value of interaction term is greater than 0.1. Therefore, we can conclude that two factors are all significantly effective on the FPM at 90% confidence level.

ANOVA Table

Source	Sum Sq	d.f.	F Statistic	p value
Intercept	109.08	1	54.70	0.000***
HDI	10.489	2	2.63	0.08
SI	28.811	3	3.48	0.02
HDI:SI	15.619	6	1.31	0.26
Residuals	287.152	144		

The difference of FPM under different treatments is displayed in the following tables. We observe that when SI=1, it has lower average logged FPM than other groups and when HDI = 1, it has lowest average logged FPM.

Difference under Effects of SI

Treatments	Difference	Lower	Upper
2-1	0.86	-0.16	1.89
3-1	1.26	0.24	2.28
4-1	1.20	0.10	2.29
3-2	0.40	-0.32	1.11
4-2	0.34	-0.48	1.15
4-3	-0.06	-0.87	0.75

Difference under Effects of HDI

Treatments	Difference	Lower	Upper
2-1	2.65	1.99	3.31
3-1	2.21	1.53	2.89
3-2	-0.44	-1.08	0.20

Sensitivity Analysis

To justify our model, we use Normal Q-Q plot of residuals and Scale-Location plot (Figure 6) to evaluate normality and constant variance assumptions.

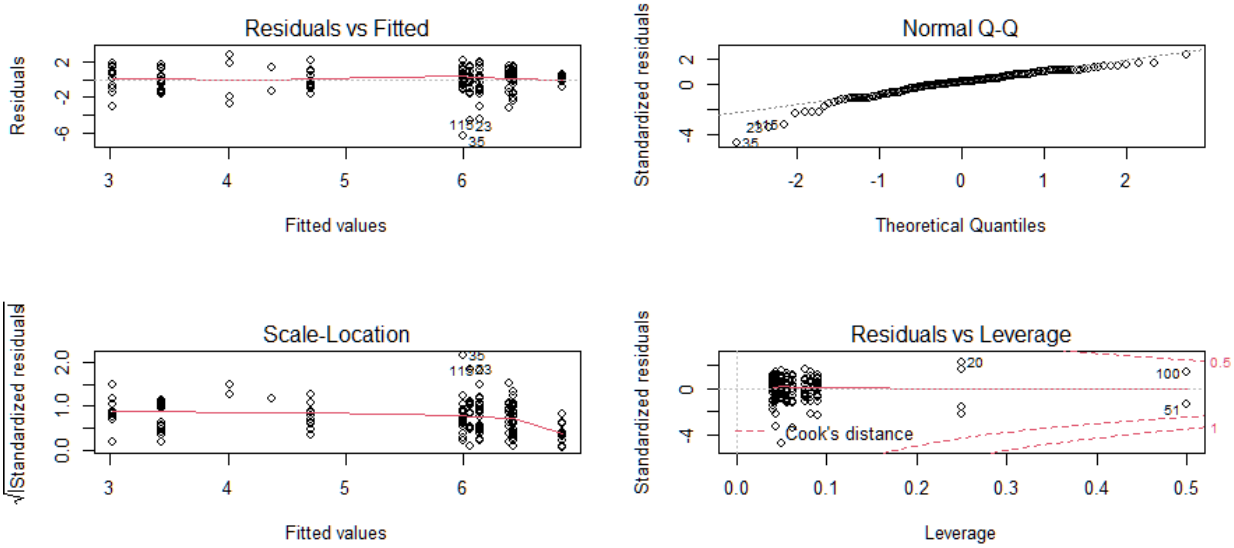


Figure 5: Figure 6.Diagnostic Plot

Theoretically, if the residuals follow a normal distribution, the scattered points should be closed to the line in the Q-Q plot. If the residuals show homoscedasticity, the scale of dispersion should be similar and there should be no significant fluctuation in a specific group.

In our model, we found that after neglecting several outliers, the deviations from the group means are likely to follow a normal distribution and the variability doesn't goes up and down subject to fitted values. Therefore, there is no significant violation on model assumptions and our model should be reasonable.

Discussion

This report reveals that both Human Development Index (HDI) and Stringency Index (SI) can be influential on the fatality per million due to COVID-19. Compared with other groups, countries with most stringent policies (1 in SI) tend to have lower average logged FPM than other groups by 0.86, 1.26 and 1.2. It indicates that strict government intervention is necessary to control the pandemic and save lives. Procedures like mask requirement, social distancing and restriction on migration is still necessary and very helpful. When it comes to HDI it is kind of surprising. We find that countries with least HDI tend to have lower logged FPM which is less than other groups by 2.64 and 2.21. One possible explanation is that the actual statistics are not fully and correctly measured in those less developed area.

As for the causality, we believe that it is not applicable for this topic. Because first of all, there are a lot of factors uncovered in this research such as vaccination rate, density of medical resources. The interaction between them is rather complicated and hard to depict. Also, we are using a Two-Way ANOVA model and propensity score, a typical way to evaluate estimate treatment effect, is usually generated by (1) Logistic Regression and (2) Classification and Regression Tree Analysis.

Generally speaking, our report just tried to explore a possible relationship of government intervention and society background. However, the way we classify the country is rather elementary and it is not easy to come up with a best indicator. Further analysis can be done with more well-constructed indicators and advanced models like linear mixed effect models in which propensity score applicable.

Acknowledgement

To finish this report, I consulted with Zhikuan Quan about causality and process of outliers in logarithm transformation. My teammates, Eric Chagnon, John Dinh and Xuanjun Gong also provided valuable suggestions on my drafts. Their efforts are appreciated here.

Additional Notes

The data analysis part of this report is performed based on local datasets for shorter loading and knitting time. However, all aforementioned datasets are publicly available through respective links. The code should be ran without break after commenting and uncommenting relative code. All relative materials can be found on this repository

References

(1) Kai Liu, Mu He, Zian Zhuang, Daihai He, Huaichen Li, Unexpected positive correlation between human development index and risk of infections and deaths of COVID-19 in Italy, One Health, Volume 10, 2020

(2) Georgiou, Miltiades N., Is Human Development Index a Shield against COVID-19?

(3) Ma, Y., Mishra, S.R., Han, XK. et al. The relationship between time to a high COVID-19 response level and timing of peak daily incidence: an analysis of governments' Stringency Index from 148 countries. Infect Dis Poverty 10, 96 (2021).

(4) Eze, F. and Nwankwo, E. (2016) Analysis of Variance in an Unbalanced Two-Way Mixed Effect Interactive Model. Open Jou nal of Statistics, 6, 310-319.

(5) Chen, S. (2022). Static Course Note of STA 207. Jupyter.

Appendix: R Code

```
library(tidyverse)
library(ggplot2)
library(countrycode) # iso countrycode conversion
library(cluster) # used for k-means clustering
library(factoextra) # used for clustering and clustering visualization
library(car) # unbalanced two-way anova
library(reshape2)
library(lme4) # mixed effect model fitting
# library(knitr)
library(easyGgplot2)
library(multcomp)
setwd("D:/UCD/Winter2022/STA207/FINAL PROJECT")
#
```

```
# Get live dataset online with *read_csv*
# covid <- read_csv("https://covid19.who.int/WHO-COVID-19-global-data.csv")
who = read.csv(file = 'WHO-COVID-19-global-data.csv')
dim(who) #Check
# Supplemantary dataset from OurWorldInData
# global = read_csv("https://covid.ourworldindata.org/data/owid-covid-data.csv")

# global = read.csv(file='owid-covid-data.csv')
# dim(global) # Check import

# Extract useful variables from OWID dataset to reduce loading time
# owid = global[,c(1,2,3,4,5,6,8,9,11,14,48,49,50,63)]
## write.csv(owid,'2ndData.csv',row.names = FALSE)
owid = read.csv(file = '2ndData.csv')
```

```
colnames(who)[1]='date'
# Convert data to the same format
```

```r
who.2021 = who[who$date >= '2021-01-01' & who$date <= '2021-12-31',]
owid$date = format(as.Date(owid$date,'%m/%d/%Y'), '%Y-%m-%d')
owid.2021 = owid[owid$date >= '2021-01-01' & owid$date <= '2021-12-31',]
```

```r
# Filter out non-country/continent cases
non.country = c('OWID_AFR','OWID_ASI','OWID_EUR','OWID_EUN','OWID_HIC','OWID_INT','OWID_LIC','OWID_LMC'

# owid.2 drops non-country cases
owid.2= owid.2021[!grepl(paste(non.country, collapse = '|'),owid.2021$iso_code),]

unique(owid.2$iso_code)
# Recode Kosov
owid.2$iso_code[owid.2$iso_code == 'OWID_KOS'] = 'KOS'

unique(owid.2$iso_code) # length 224

# Location coding transformation
owid.2$iso2 = countrycode(sourcevar = owid.2$iso_code , origin = 'iso3c',destination = 'iso2c')
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some val
```

```r
# Reorder the column
owid.2 = owid.2[,c(15,c(1:14))]
# Manually set iso.2 code for Kosov
owid.2$iso2[owid.2$iso_code=='KOS'] = 'XK'
unique(owid.2$iso2)

# remove Other in WHO dataset
who.2 = who.2021[!who.2021$Country == 'Other',]
# extract who iso.2 code
who.2.iso2 = unique(who.2$Country_code)
# there's na in who.2
who.3 = who.2[!is.na(who.2$Country_code),]
# remove na
who.3.iso2 = unique(who.3$Country_code)
who.3.iso2
# Some location in owid dataset is not country
code.diff = setdiff(owid.2$iso2, who.3$Country_code)
code.diff
non.country.iso2 = c('BQ','HK','MO','NA','TW')
# Remove cases not included in WHO dataset
owid.3= owid.2[!grepl(paste(non.country.iso2, collapse = '|'),owid.2$iso2),]

setdiff(who.3$Country_code,owid.3$iso2)
# Remove cases not included in OWID datset
rm.country.iso2 = setdiff(who.3$Country_code,owid.3$iso2)
# Final WHO dataset
who.4= who.3[!grepl(paste(rm.country.iso2, collapse = '|'),who.3$Country_code),]
# verify that the country number is equal
length(unique(who.4$Country))==length(unique(owid.3$iso2))
```

```
si.2021 = aggregate(owid.3$stringency_index,
                    list(owid.3$iso2), FUN = mean)
# some countries has no stringency index data
# extract countries with stringency index data
country.list = si.2021[!is.na(si.2021$x),]$Group.1
si.raw = si.2021[!is.na(si.2021$x),]
si.raw = as.data.frame(si.raw)
names(si.raw)[2]='si.bar'
# set country code list as row names
country.list = si.raw$Group.1

# select corresponding country in who dataset
who.5 = who.4[grepl(paste(country.list, collapse = '|'), who.4$Country_code),]

# get total death in 2021 from WHO dataset
death.2021 = aggregate(who.5$New_deaths, list(who.5$Country_code),FUN = sum)
colnames(death.2021) = c('iso2', 'death')
data = cbind(si.raw,death.2021)
rownames(data)=c(1:nrow(data))
## the death data is correct so far
```

```
# perform k-means clustering on HDI
owid.hdi = owid.3[owid.2021$date == '2021-12-31',]
hdi = as.data.frame(owid.hdi[,c(1,13,15)])
# select common countries
hdi = hdi[grepl(paste(country.list, collapse = '|'), hdi$iso2),]
# order cases in alphabetical order w.r.t. country name
hdi.ordered=hdi[order(hdi$iso2),]
# reorder rows
rownames(hdi.ordered)=c(1:nrow(hdi.ordered))
# there are duplicated cases for some reason, remove them
hdi.sorted = hdi.ordered[!duplicated(hdi.ordered$iso2),]
rownames(hdi.sorted) = c(1:nrow(hdi.sorted))
# check on difference b/w datasets
country.diff = setdiff(data$iso2,hdi.sorted$iso2)
data = data[!grepl(paste(country.diff, collapse = '|'),data$iso2),]
rownames(data) = c(1:nrow(data))
data = cbind(data,hdi.sorted)
# drop possible n/a values
data = na.omit(data)
rownames(data) = c(1:nrow(data))
# use this if necessary
data = data[,-c(3,5)]
names(data)[1]='iso2'
```

```
hdi.subset = data[5]
si.bar.subset = data[2]
# determine the optimal number of clusters of hdi
# using average silhoutte method
cluster.plot.1 = fviz_nbclust(hdi.subset,kmeans, method = 'wss')
# optimal cluster number is 3
# determine the optimal number of clusters of si.bar
cluster.plot.2 = fviz_nbclust(si.bar.subset, kmeans, method = 'silhouette')
```

```r
ggplot2.multiplot(cluster.plot.1,cluster.plot.2,cols = 2)
```

```r
# optimal cluster number is 4
hdi.cluster = kmeans(hdi.subset, centers = 3, nstart = 25)
si.bar.cluster = kmeans(si.bar.subset, centers = 4, nstart = 25)
# import country code and corresponding cluster
hdi.class = hdi.cluster$cluster
si.bar.class = si.bar.cluster$cluster
```

```r
# append columns
data = cbind(data, hdi.class)
data = cbind(data, si.bar.class)
# get fatality per million
data$fpm = data$death/(data$population/1000000)
# complete dataset for further analysis
names(data)[5] = 'hdi'
data$hdi.class.factor = as.factor(data$hdi.class)
data$si.bar.class.factor = as.factor(data$si.bar.class)
# dataset where FPM is logged
data.log = data
data.log$fpm = log(data.log$fpm)
# drop cases with -inf log(fpm)
data.log = data.log[-which(data.log$fpm < -1000),]
```

```r
# boxplots for FPM, grouped by SI
p1 = ggplot(data,
       aes(x = si.bar.class, y = fpm, group = si.bar.class))+
  geom_boxplot()+
  ylim(c(0,5000))+
  labs(
    xlab = 'Strigency Index Group',
    ylab = 'Fatality per Million in 2021',
    title = 'FPM Grouped by Stringency Index(SI)')

# boxplot for fpm, grouped by hdi
p2 = ggplot(data,
       aes(x = hdi.class, y = fpm, group = hdi.class))+
  geom_boxplot()+
  ylim(c(0,5000))+
  labs(
    xlab = 'HDI',
    ylab = 'Fatality per Million in 2021',
    title = 'FPM Grouped by HDI')

ggplot2.multiplot(p1,p2,cols=2)
```

```r
# histogram for FPM
p3 = ggplot(data,
       aes(x = fpm))+
  geom_histogram(aes(y = after_stat(density)),data = data, bins = 15)+
  geom_density()+
  labs(
    x = 'FPM',
    y = 'Density',
    title = 'Histogram of Fatality per Million(FPM)'
  )
# histogram for log(fpm)
p4 = ggplot(data.log,
       aes(x = fpm))+
  geom_histogram(aes(y = after_stat(density)),data = data.log, bins = 15)+
  geom_density()+
  labs(
    x = 'FPM',
    y = 'Density',
    title = 'Histogram of logged Fatality per Million(FPM)'
  )

# histogram for si
p5 = ggplot(data,
       aes(x = si.bar))+
  geom_histogram(aes(y = after_stat(density)),data = data, bins = 15)+
  geom_density()+
  labs(
    x = 'SI',
    y = 'Density',
    title = 'Histogram of SI'
  )
# histogram for HDI
p6 = ggplot(data,
       aes(x = hdi))+
  geom_histogram(aes(y = after_stat(density)),data = data, bins = 15)+
  geom_density()+
  labs(
    x = 'HDI',
    y = 'Density',
    title = 'Histogram of HDI'
  )
ggplot2.multiplot(p3,p4,p5,p6,cols = 2)
```

```
# covariance between independent variables (supposed to be low)
cov(data$hdi.class,data$si.bar.class)
```

```
# generate summary table for numerical variables
apply(data[,c(2,3,4,5,8)],2,summary)
```

```
# fit with 2-way anova model, w/ interaction term
anova.fit = aov(fpm~hdi.class*si.bar.class, data = data)
Anova(anova.fit, type = 'III')
```

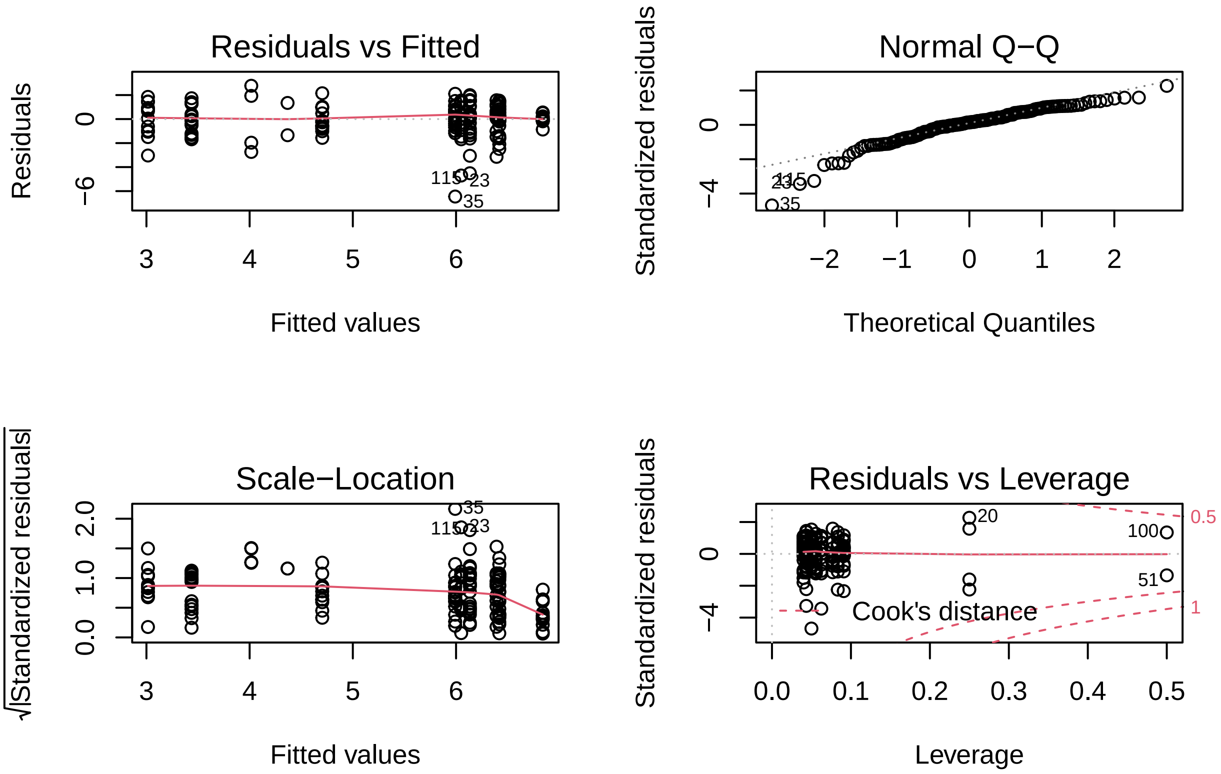
```
# anova.fit2 = aov(fpm~hdi.class + si.bar.class + hdi.class:si.bar.class, data = data.log)
# Anova(anova.fit2, type = 'III')
```

```
anova.fit3 = aov(fpm~hdi.class.factor*si.bar.class.factor, data = data.log)
Anova(anova.fit3, type = 'III')
```

```
# par(mfrow=c(2,2))
# plot(anova.fit2)
```

```
par(mfrow=c(2,2))
plot(anova.fit3)
```

```
## Warning: not plotting observations with leverage one:
##   25
```



```
# plot(anova.fit2, which = 2, add.smooth = TRUE)
# plot(anova.fit2, which = 3, add.smooth = FALSE)
```

```
ps.fit = glm(fpm~hdi.class+si.bar.class, data = data.log)
summary(ps.fit)
ps.df = data.frame(pr.score = predict(ps.fit, type = 'response'),
                   fpm = ps.fit$model$fpm)
ps.df
```

```
TukeyHSD(anova.fit3, which = 'hdi.class.factor')
TukeyHSD(anova.fit3,which = 'si.bar.class.factor')
```

```
shapiro.test(anova.fit3$residuals)
```

```
sessionInfo()
```