Investigation on World Happiness, Covid Deaths and Vaccination

fa21-prj-shiyuan8-sw20-zihanz12-zl32

12/8/2021

Project Outline

1. Visualize the life satisfaction (happiness index) across the globe.
2. Visualize the trend of Covid-19 cases and mortality rates, and combine the world happiness with total Covid-19 cases and deaths.
3. Combine the world happiness with vaccination rate, and visualize the correlation between Covid-19 vaccination rates and world happiness.
4. Use Anova analysis to figure the impacts of Covid-19 cases, deaths, vaccination rates and different factors across fields, such as public policy, public health, economics etc. on the happiness index.
5. Explore as many R packages as we can.

Data Source

- Covid-19 daily cases and excess deaths (https://github.com/TheEconomist/covid-19-the-economist-global-excess-deaths-model/blob/main/source-data/country_daily_excess_deaths.csv)
- Covid-19 daily vaccination rates (https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations)
- World happiness (https://github.com/owid/owid-datasets/blob/master/datasets/World%20Happiness%20Report%20(2021)/World%20Happiness%20Report%20(2021).csv)

For above datasets, we remove all the missing values in the dataset of Covid-19 daily cases, excess deaths, and vaccination rates. Then we calculate the total number of Covid-19 daily cases, excess deaths, and vaccination rates in 2020. After that, we merge them with the world happiness data, on the column of country name.

Analysis on [world-happiness] dataset

Include libraries for data cleaning, data visuilazation

```
# include libraries for data cleaning, data visuilazation
library(data.table)
library(curl)
```

```
## Using libcurl 7.68.0 with GnuTLS/3.6.13
```

```
library(tidyverse)  # package for data manipulation
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.1.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ─────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::between()    masks data.table::between()
## x dplyr::filter()     masks stats::filter()
## x dplyr::first()      masks data.table::first()
## x dplyr::lag()        masks stats::lag()
## x dplyr::last()       masks data.table::last()
## x readr::parse_date() masks curl::parse_date()
## x purrr::transpose()  masks data.table::transpose()
```

```
library(tidyr)
library(ggrepel)   # packages for plotting
library(sf)# package for working with spatial data; sf has functions compatible with ggplot
```

```
## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1
```

```
library(rnaturalearth)  # package for loading world map
library(rnaturalearthdata) #  same as above
library(countrycode)
library(dplyr)
library(ggplot2)
```

Implementing data cleaning and data wrangling

```
happiness <- fread("World Happiness Report (2021).csv")
```

Using Package(countrycode) to assign each country to according region, and add the column to data.table(happiness)

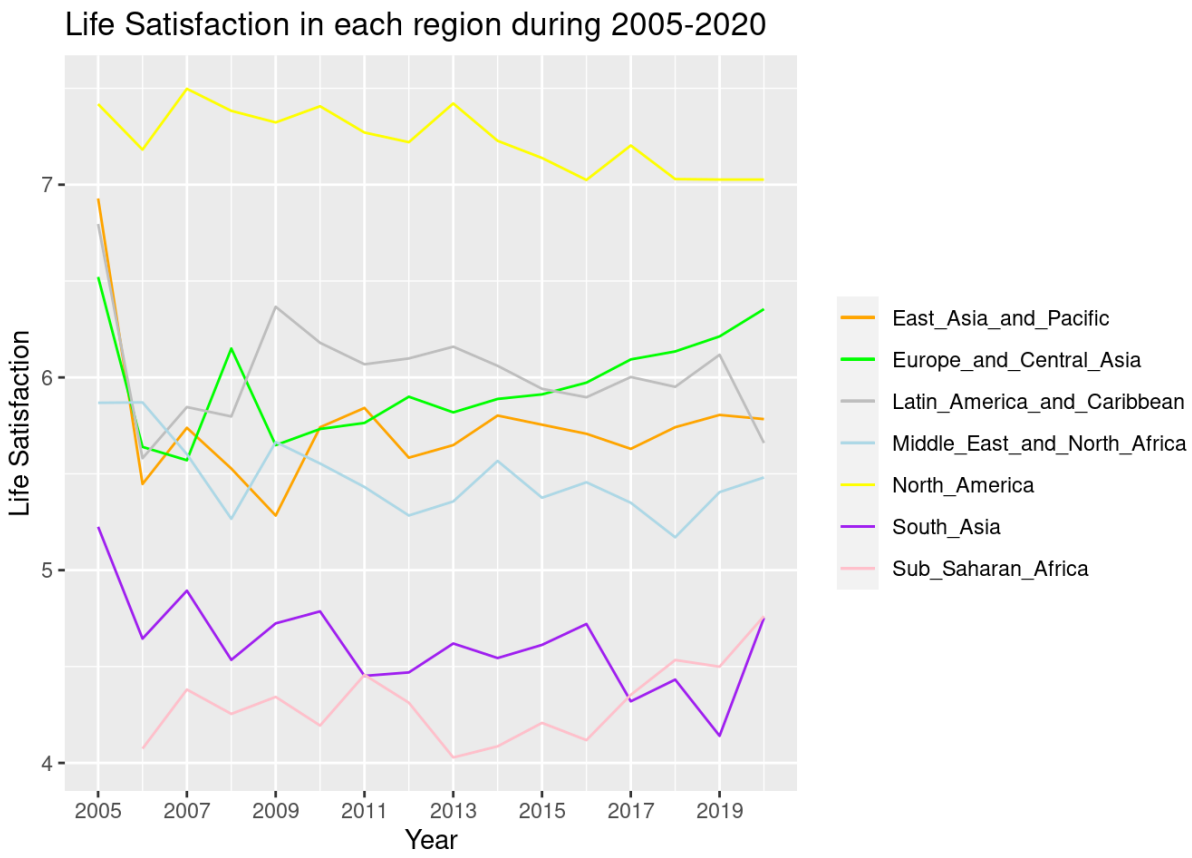
```
happiness_ <- as_tibble(data.table::fread("World Happiness Report (2021).csv"))
source<- pull(happiness_,Entity)
country_code <- countrycode(source, origin = "country.name",destination = "region")
happiness$countrycode_dt <- as.data.table(country_code)
names(happiness)[names(happiness) == colnames(happiness)[3] ] <- "life_satisfaction"
names(happiness)[names(happiness) == colnames(happiness)[4] ] <- "region"
```

Implement ‘dcast’ method to create desired data.table, and change the columnnames to avoid “&” and blank space.

```
happiness_1<- dcast(happiness, Year~region, value.var = "life_satisfaction", fun= list(mean), drop= FALSE)
colnames(happiness_1)[2] = "East_Asia_and_Pacific"
colnames(happiness_1)[3] = "Europe_and_Central_Asia"
colnames(happiness_1)[4] = "Latin_America_and_Caribbean"
colnames(happiness_1)[5] = "Middle_East_and_North_Africa"
colnames(happiness_1)[6] = "North_America"
colnames(happiness_1)[7] = "South_Asia"
colnames(happiness_1)[8] = "Sub_Saharan_Africa"
```

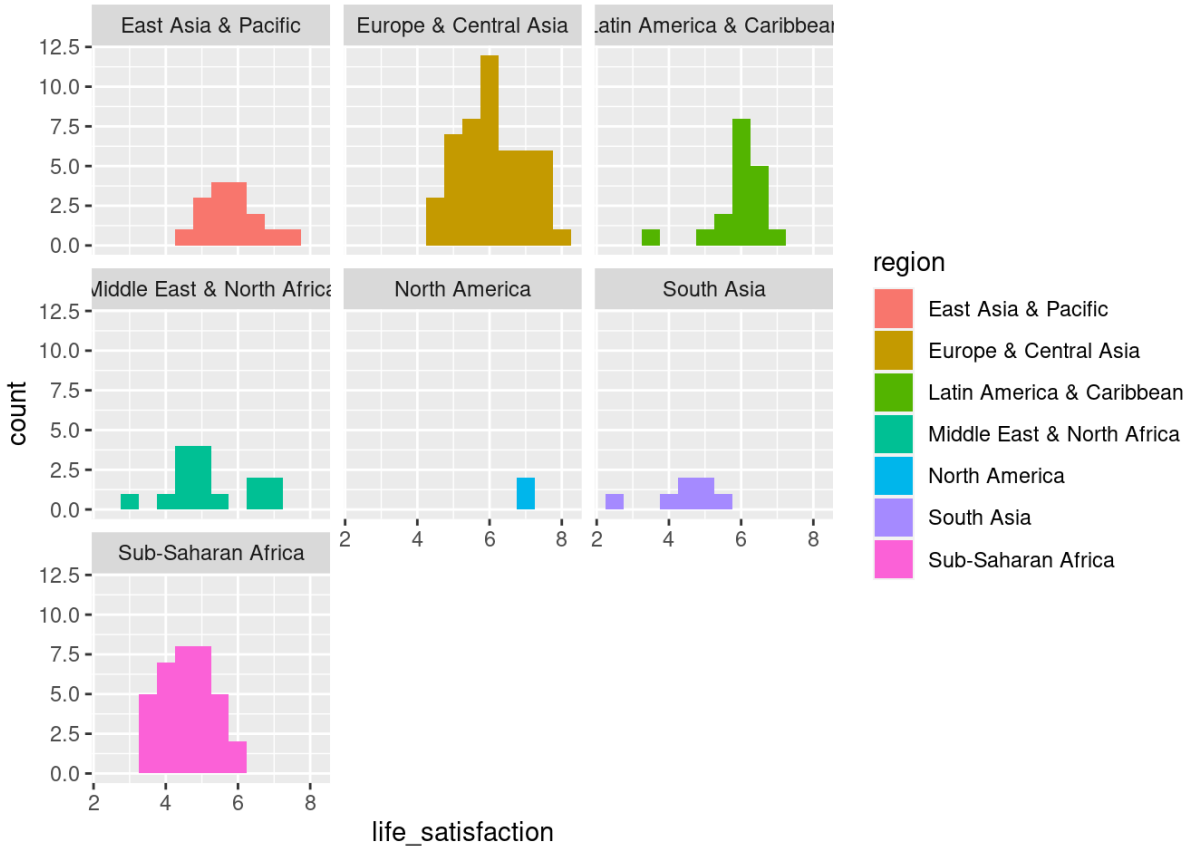
Data visualization

```
ggplot(happiness_1,aes(x=Year)) +
  geom_line(aes( y= East_Asia_and_Pacific,colour= "East_Asia_and_Pacific"))+
  geom_line(aes( y= Europe_and_Central_Asia,colour= "Europe_and_Central_Asia"))+
  geom_line(aes( y= Latin_America_and_Caribbean,colour= "Latin_America_and_Caribbean"))+
  geom_line(aes( y= Middle_East_and_North_Africa,colour= "Middle_East_and_North_Africa"))+
  geom_line(aes( y= North_America,colour= "North_America"))+
  geom_line(aes( y= South_Asia,colour= "South_Asia"))+
  geom_line(aes( y= Sub_Saharan_Africa,colour= "Sub_Saharan_Africa")) +
  scale_colour_manual("", values = c("East_Asia_and_Pacific" ="orange", "Europe_and_Central_Asia" ="green",
    "Latin_America_and_Caribbean"="grey",  "Middle_East_and_North_Africa"="lightblue",
    "North_America"="yellow", "South_Asia" ="purple", "Sub_Saharan_Africa" ="pink"))+
  scale_x_continuous("Year",breaks = seq(2005,2020,by=2))+ scale_y_continuous("Life Satisfaction",breaks =
 seq(4,8)) +
  labs(title = "Life Satisfaction in each region during 2005-2020")
```



From this graph, we can see that: Among all these 7 regions, North_America has the highest level of life-satisfaction, and sub_saharan_africa has the lowest level of life-satisfaction for almost each year except year 2017-2019. There is also an decrease for each region(except sub_saharan_africa) during 2005-2020.

```
ggplot(happiness[Year==2018], aes(x=life_satisfaction))+geom_histogram(aes(fill=region),binwidth = 0.5)+
  facet_wrap( ~region) + labs("Life Satisfaction in 2018")
```



From the histogram graph above, we can see that in year 2018, most countries in Europe&central asia, Lation America& Caribbran, and East Asia&Pacific has life_satisfaction over 5, while most of coutries in Sub-Saharan Africa and South Asia has life_satisfaction less than 5.There exits huge disparities worldwide.

```
h1 <- happiness[Year==2018]
world <- ne_countries(scale = "medium", returnclass = "sf")
```

Merge dataset “world” and dataset “happiness”, and transform the output from data.frame to sf using method ‘st_as_sf’

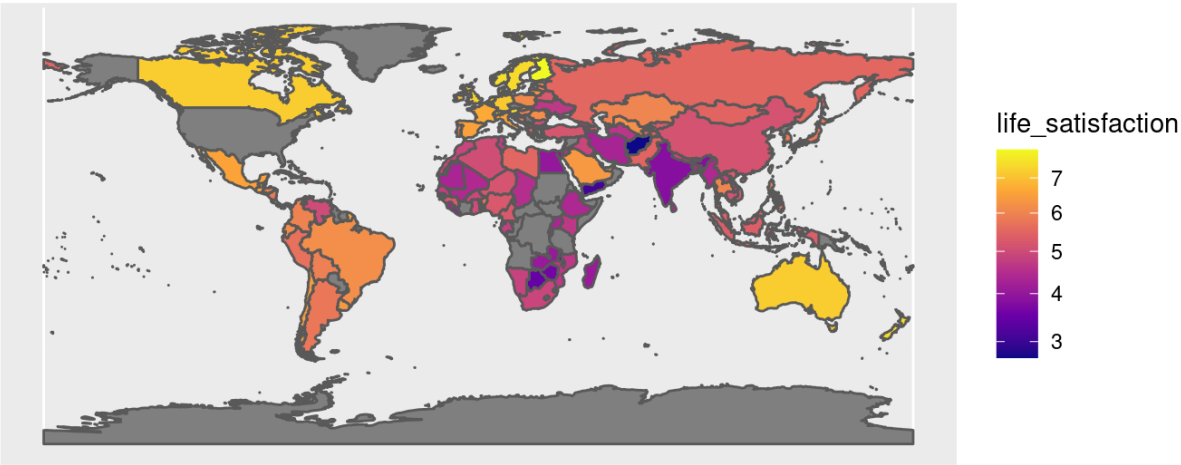
```
colnames(world)[9] = "Entity"
merge_data <- merge(h1, world, by="Entity",all.y=TRUE)
world_ <- st_as_sf(merge_data)
```

```
ggplot(data = world_) + geom_sf()
```

```
ggplot(data = world_) +geom_sf() + xlab("Longitude") + ylab("Latitude") +
ggtitle("World map", subtitle = paste0("(", length(unique(world$name)), " countries)"))
```

```
ggplot(data = world_) + geom_sf(color = "black", fill = "lightgreen")
```

```
ggplot(data = world_) + geom_sf(aes(fill = life_satisfaction)) +
scale_fill_viridis_c(option = "plasma", trans = "sqrt")
```



From the map above, we can see that: grey parts stand for countries without data recorded in year 2018; and for other parts in this world map,from colour yellow to blue, the darker the colour is for each country, the lower life_satisfaction score the coutry has. We can see that Canada and Australia and most parts of Europe has high level of life_satisfaction, while most parts of Africa has low level of life_satisfaction.

Analysis on the trend of Covid-19 daily cases and excess deaths.

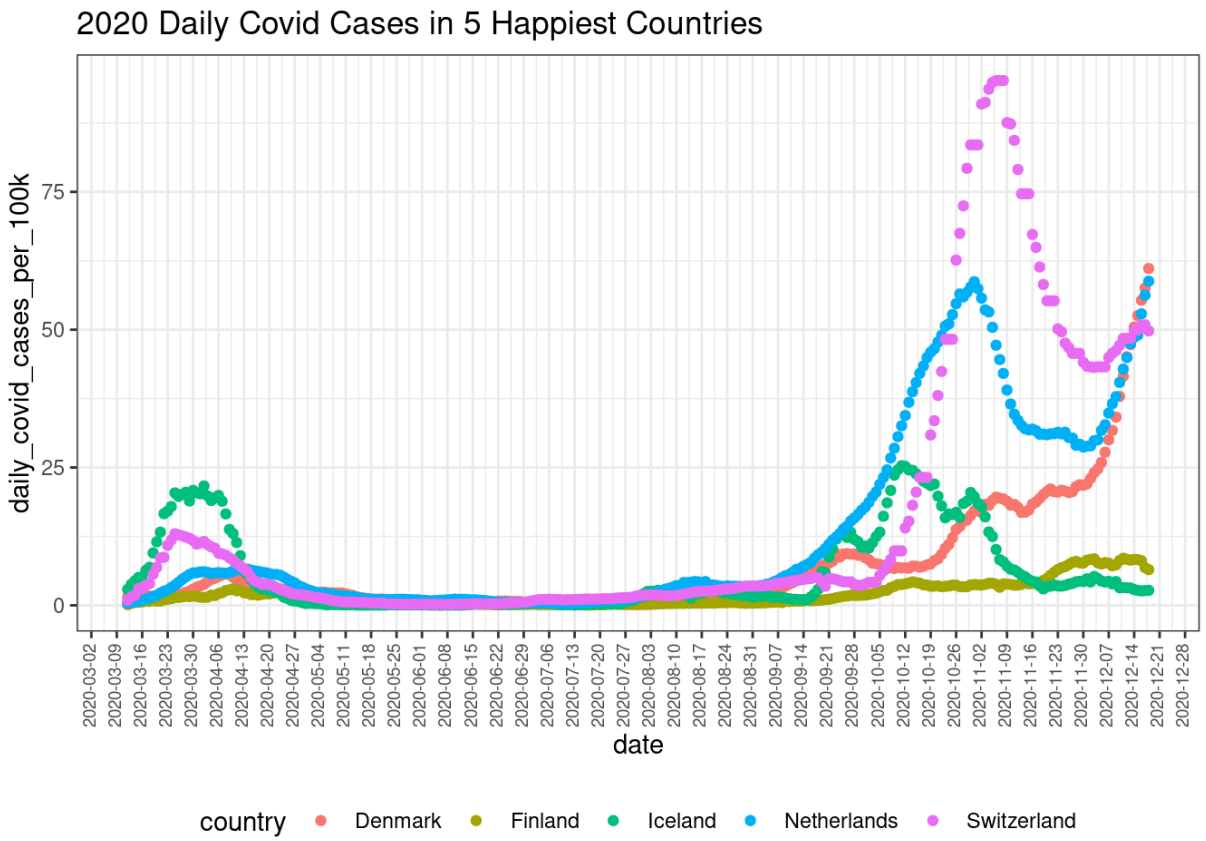
First, we only want to analyze the countries which have a happiness index, so we did the sub-setting as follow.

```
covid = read.csv("covid_country_daily_excess_deaths.csv")
happiness_2020 = read.csv("world-happiness-report-2020.csv")

# Match the countries
country = unique(happiness_2020$Country)
covid1 = covid[covid$country %in% country,]
```

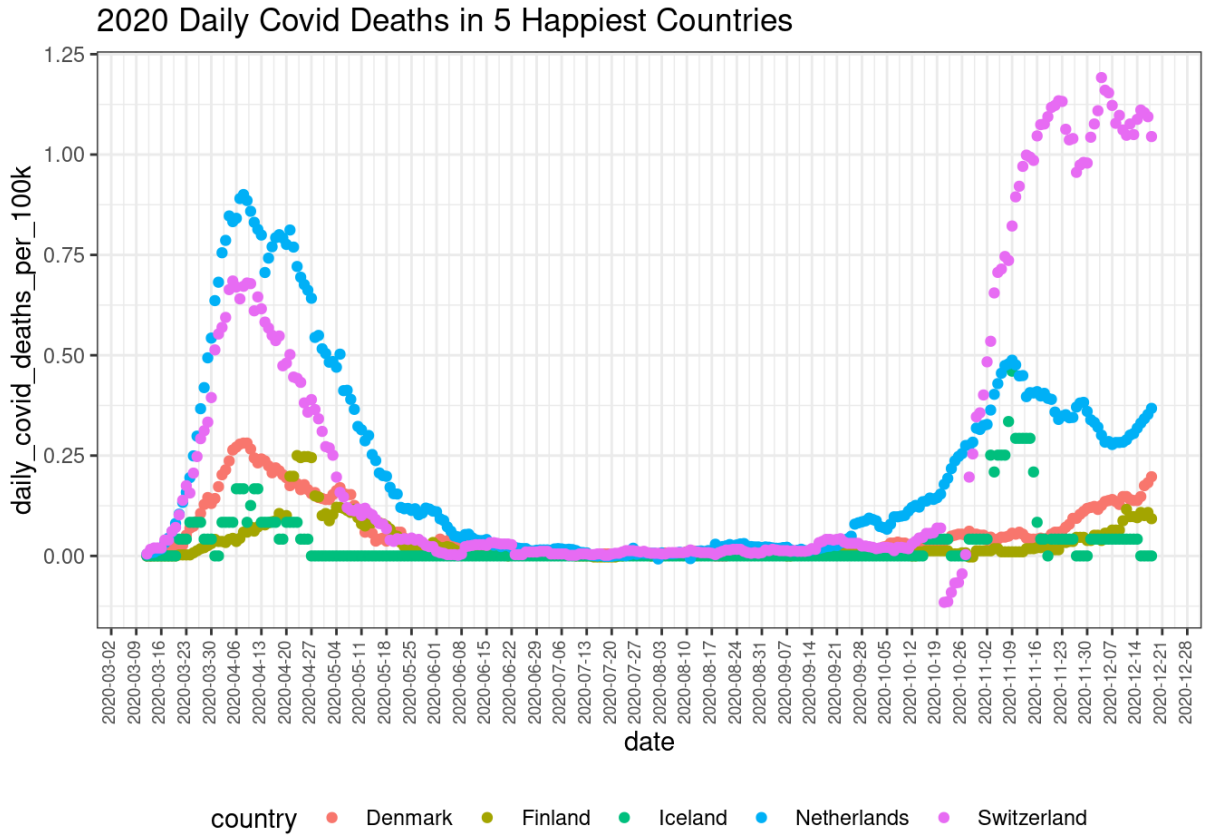
Then we want to visualize the Covid-19 cases and deaths in the happiest and least happiest country in the world, and try to see whether there exists a relationship between them.

```
# Top and bottom happiness (2020)
happiness_2020 = happiness_2020[,c("Country.name","Ladder.score")]
names(happiness_2020) <- c("Country", "Score")

## split date into year,month,day
covid1 = separate(data = covid1, col = date, into = c("Year", "Month","day"), sep = "-")
covid_2020 = covid1[covid1$Year==2020,]
Top_2020 = happiness_2020[order(happiness_2020$Score,decreasing = TRUE),]$Country[1:5]
Bottom_2020 = happiness_2020[order(happiness_2020$Score),]$Country[1:5]
covid_2020_bottom = covid_2020[covid_2020$country %in% Bottom_2020,]
covid_2020_bottom$country = as.factor(covid_2020_bottom$country)
covid_2020_bottom$date = as.Date(paste("2020","-",covid_2020_bottom$Month,"-",covid_2020_bottom$day,sep=''
))
covid_2020_top = covid_2020[covid_2020$country %in% Top_2020,]
covid_2020_top$country = as.factor(covid_2020_top$country)
covid_2020_top$date = as.Date(paste("2020","-",covid_2020_top$Month,"-",covid_2020_top$day,sep=''))
```

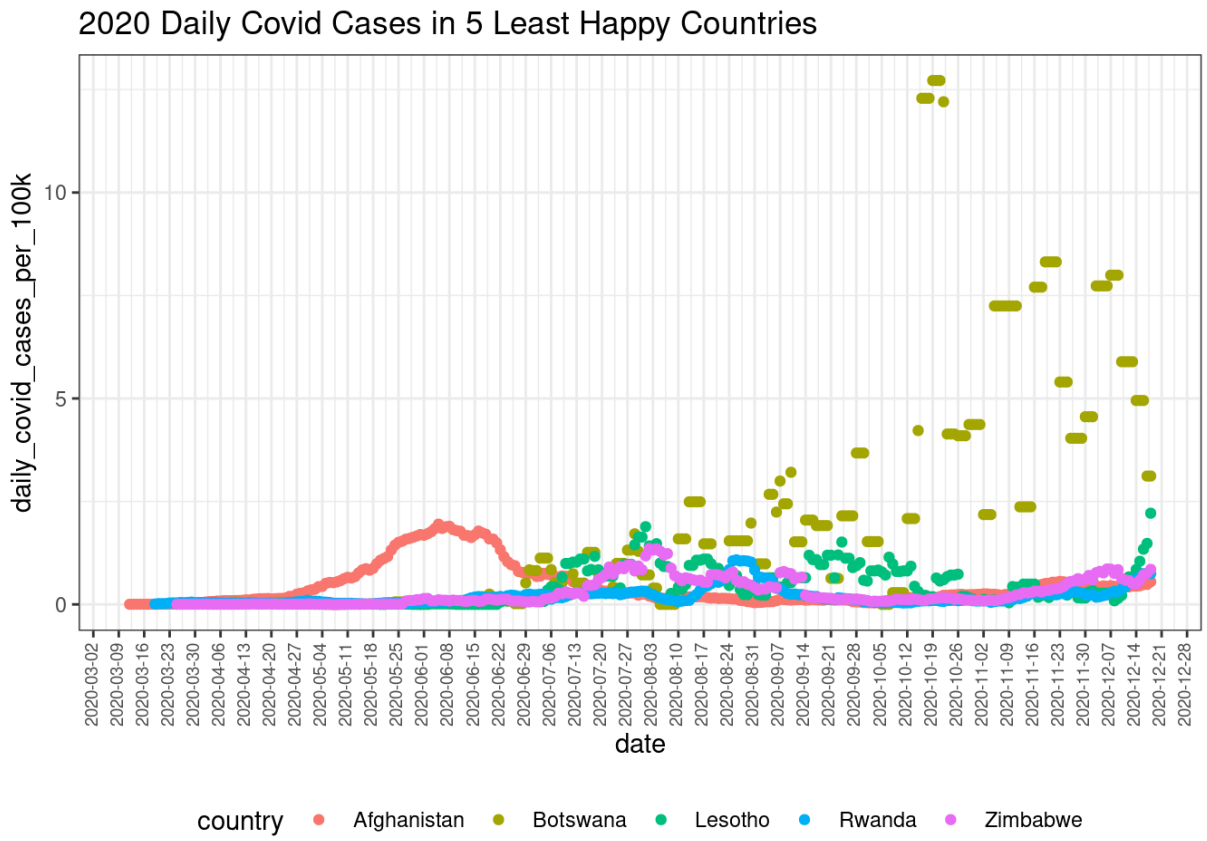
```
ggplot(covid_2020_top,aes(x=date,y=daily_covid_cases_per_100k,color=country))+
  geom_point()+
  labs(title = '2020 Daily Covid Cases in 5 Happiest Countries')+
  theme_bw()+
  scale_x_date(date_breaks = '1 week',limits=c(as.Date("2020-03-12"),as.Date("2020-12-18")))+
  theme(axis.text.x = element_text(size = 7,angle = 90, hjust = 0.3, vjust = 0.3),legend.position="bottom")
```



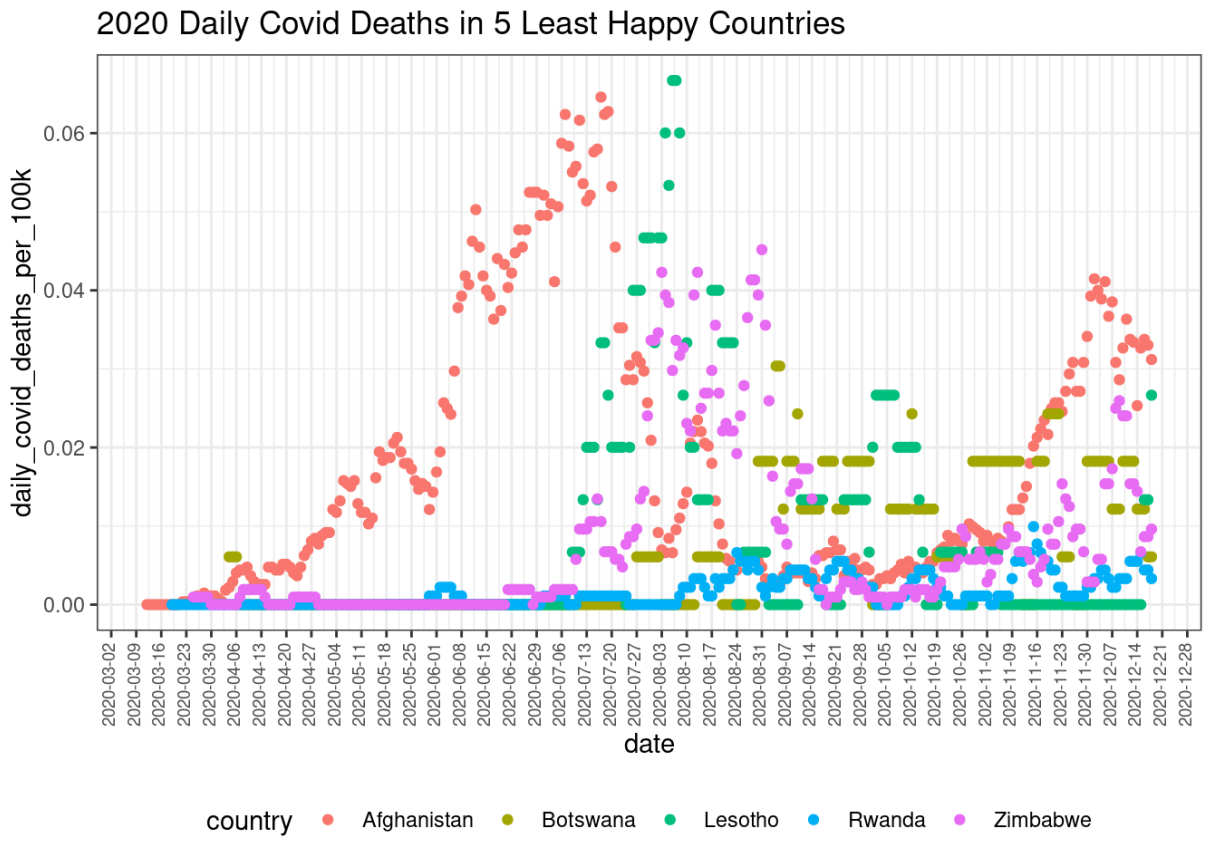
```
ggplot(covid_2020_top,aes(x=date,y=daily_covid_deaths_per_100k,color=country))+
  geom_point()+
  labs(title = '2020 Daily Covid Deaths in 5 Happiest Countries')+
  theme_bw()+
  scale_x_date(date_breaks = '1 week',limits=c(as.Date("2020-03-12"),as.Date("2020-12-18")))+
  theme(axis.text.x = element_text(size = 7,angle = 90, hjust = 0.3, vjust = 0.3),legend.position="bottom")
```




```
ggplot(covid_2020_bottom,aes(x=date,y=daily_covid_cases_per_100k,color=country))+
  geom_point()+
  labs(title = '2020 Daily Covid Cases in 5 Least Happy Countries')+
  theme_bw()+
  scale_x_date(date_breaks = '1 week',limits=c(as.Date("2020-03-12"),as.Date("2020-12-18")))+
  theme(axis.text.x = element_text(size = 7,angle = 90, hjust = 0.3, vjust = 0.3),legend.position="bottom")
```




```
ggplot(covid_2020_bottom,aes(x=date,y=daily_covid_deaths_per_100k,color=country))+
  geom_point()+
  labs(title = '2020 Daily Covid Deaths in 5 Least Happy Countries')+
  theme_bw()+
  scale_x_date(date_breaks = '1 week',limits=c(as.Date("2020-03-12"),as.Date("2020-12-18")))+
  theme(axis.text.x = element_text(size = 7,angle = 90, hjust = 0.3, vjust = 0.3),legend.position="bottom")
```

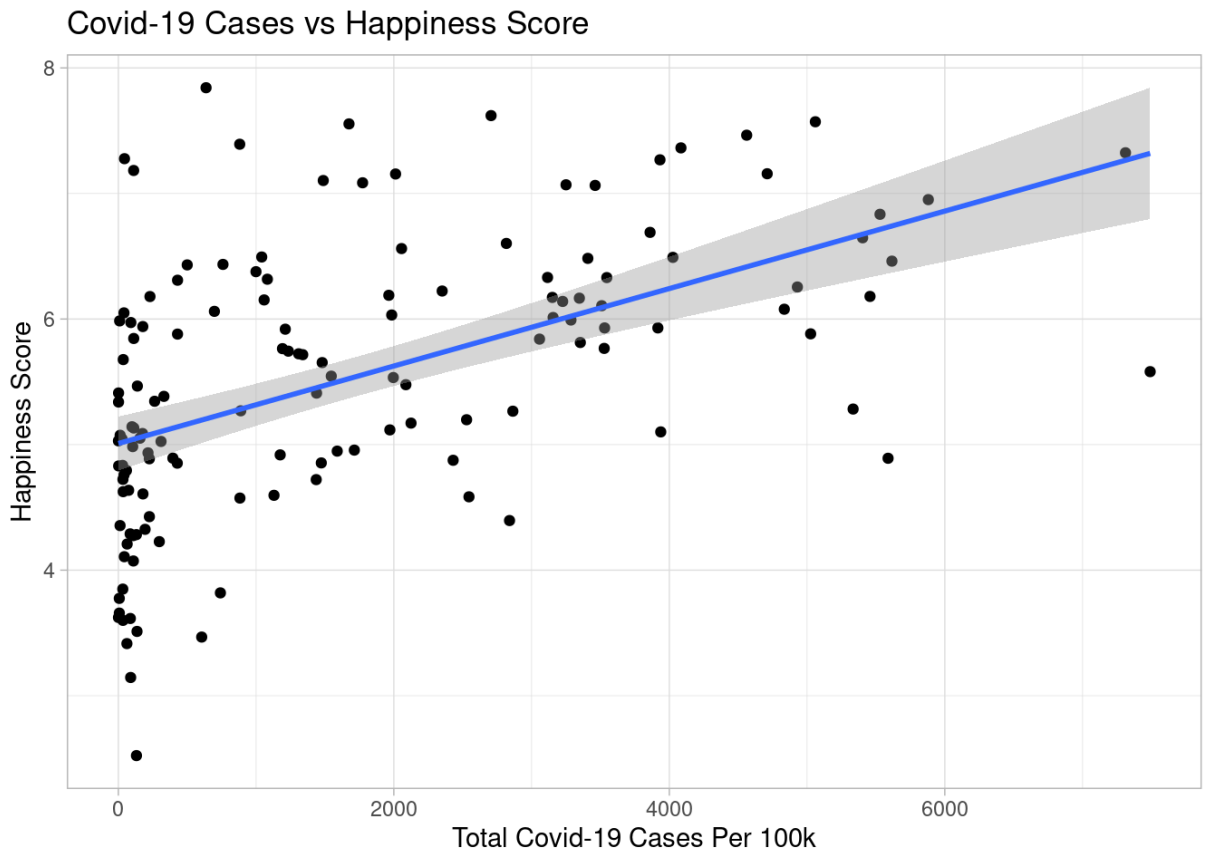


From those plots we can see that, the happiest countries had two spikes both in Covid-19 cases and deaths, one during the beginning of March, and one around November; while the least happiest countries witnessed rises in cases and deaths in the middle and the end of 2020. Overall, the happiest countries had more cases and excess deaths than the least happiest ones.

It's rather counter-intuitive that the countries with more Covid-19 deaths are happier, and thus we want to see the relationship between Covid-19 cases, deaths and happiness scores.

```
co_ha = read.csv("covid_happiness.csv")
ggplot(co_ha, aes(x = total_cases_per_100k, y = Ladder.score)) +
    geom_point() +
    geom_smooth(method = "lm") +
    ggtitle("Covid-19 Cases vs Happiness Score") +
    xlab("Total Covid-19 Cases Per 100k") +
    ylab("Happiness Score") +
    theme_light()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(co_ha, aes(x = total_deaths_per_100k, y = Ladder.score)) +
    geom_point() +
    geom_smooth(method = "lm") +
    ggtitle("Covid-19 Cases vs Happiness Score") +
    xlab("Total Covid-19 Deaths Per 100k") +
    ylab("Happiness Score") +
    theme_light()
```

```
## `geom_smooth()` using formula 'y ~ x'
```

From this plot, we surprisingly found that the Covid-19 cases and deaths seem have a positive relationship with the happiness score.

After that, we want total values for each country, so we combine happiness data with covid data.

In this step, we calculate the death_case_ratio, which the ratio of the number of deaths and the number of cases. The lower the ratio means there would be fewer people dying from the covid when they are infected, and intuitively, we would assume a country with a lower ratio would have a higher happiness index.

```
covid_2020 = data.table(covid_2020)

covid_2020_total = covid_2020[,.(total_cases = sum(daily_covid_cases,na.rm = TRUE),total_deaths = sum(daily
_covid_deaths,na.rm = TRUE),
                   total_cases_per_100k = sum(daily_covid_cases_per_100k,na.rm = TRUE),total_deaths_per_100
k = sum(daily_covid_deaths_per_100k,na.rm = TRUE)),
                   by = country][order(total_deaths,total_cases,decreasing = TRUE)]
covid_2020_total$death_case_ratio = covid_2020_total$total_deaths / covid_2020_total$total_cases

# write.csv(x = covid_2020_total,file = "covid_2020_total.csv")
```

Analysis on the relationship between each country's happiness score and vaccination rate.

We are surprised to find that there is a strong positive relationship between the two.

First we do some data cleaning job. Since the raw data is a daily reported dataset, we get the latest vaccination rate by groupby each country and select the last rows from each group.

```
row_data = read.csv("https://raw.githubusercontent.com/Hush-Baby-Hush/covid-19-data/master/public/data/vacc
inations/vaccinations.csv")
```

Get 2021 data by select from last row:

```
world_vacci_data_2021 = row_data %>%
  select(location, date, people_vaccinated, people_fully_vaccinated, people_vaccinated_per_hundred, people_
fully_vaccinated_per_hundred) %>%
  group_by(location) %>%
  filter(row_number()==n()) %>%
  select(-date)
```

Get hapiness data

```
happiness_2020 = read.csv("https://raw.githubusercontent.com/illinois-stat447/fa21-prj-shiyuan8-sw20-zihanz
12-zl32/master/data/world-happiness-report-2020.csv?token=ANGGHITN6HBZTJCVIN66RETBXLPVK")
happiness_2020 = happiness_2020 %>%
  select(Country.name, Ladder.score)
```

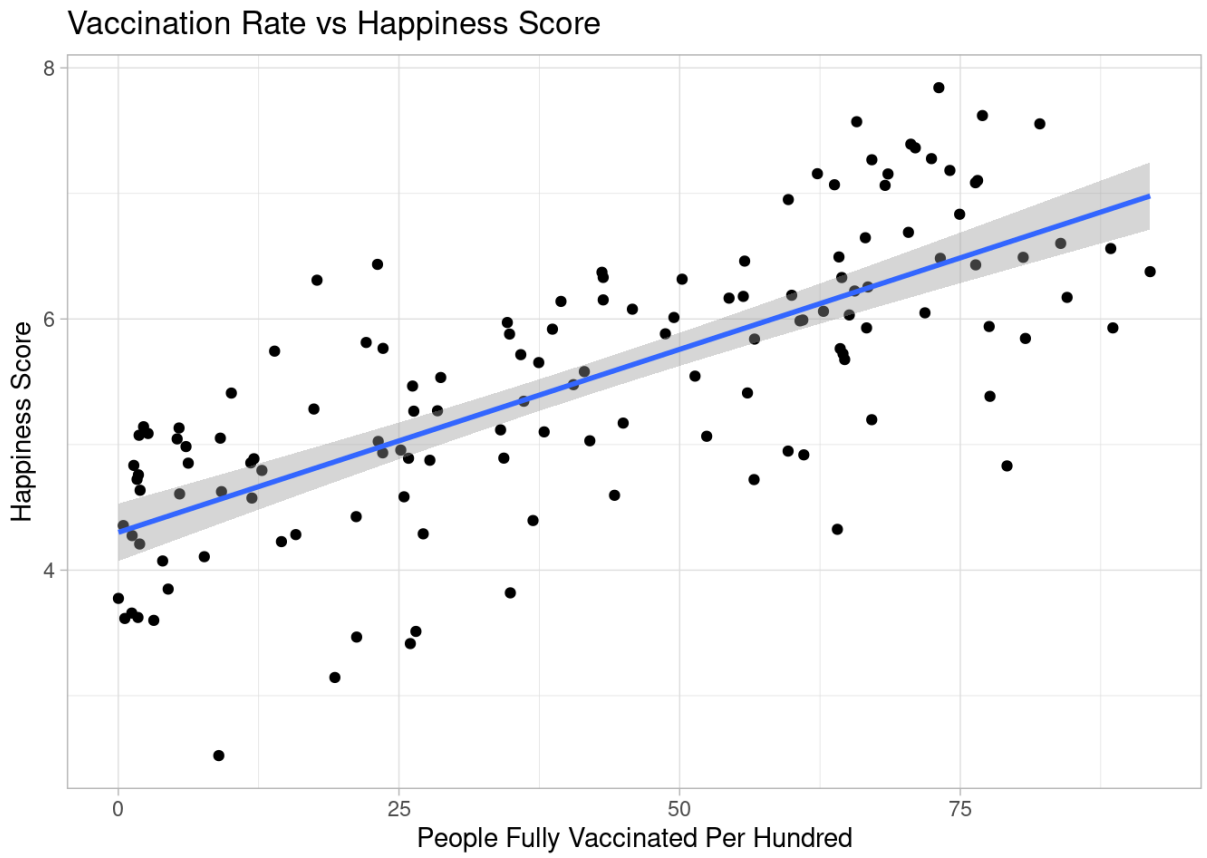
Merge two dataset:

```
vacci_happiness = left_join(world_vacci_data_2021, happiness_2020, by = c("location"="Country.name"))
vacci_happiness = vacci_happiness %>%
  select(location, people_fully_vaccinated_per_hundred, Ladder.score) %>%
  drop_na()
```

Plot scatter plot and see correlation between vaccination rate and happiness score

```
ggplot(vacci_happiness, aes(x = people_fully_vaccinated_per_hundred, y = Ladder.score)) +
    geom_point() +
    geom_smooth(method = "lm") +
    ggtitle("Vaccination Rate vs Happiness Score") +
    xlab("People Fully Vaccinated Per Hundred") +
    ylab("Happiness Score") +
    theme_light()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



This plot reveals the relationship between country vaccination rate and country happiness scores. We can see from the plot that the are positive related, and countries with higher vaccination rate generally have higher happiness score.

Next, we prepare data for world map vaccination rate:

```
row_data = read.csv("https://raw.githubusercontent.com/Hush-Baby-Hush/covid-19-data/master/public/data/vaccinations/vaccinations.csv")
```

Vaccination is not available in some countries 2020. Get location where vaccination available in 2020

```
location_2020 = row_data %>%
  select(location, date, people_vaccinated, people_fully_vaccinated, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred) %>%
  filter(date<"2021-01-01") %>%
  drop_na(people_vaccinated) %>%
  group_by(location) %>%
  filter(row_number()==n())
```

Get location where vaccination not available in 2020

```
location_2021 = row_data %>%
  select(location, date, people_vaccinated, people_fully_vaccinated, people_vaccinated_per_hundred, people_fully_vaccinated_per_hundred) %>%
  group_by(location) %>%
  filter(row_number()==1) %>%
  filter(date>"2020-12-31")
```

Create 2020 zero data for these country

```
country2020_zero = location_2021 %>%
  select(location) %>%
  mutate(date = "2020-12-31", people_vaccinated = 0, people_fully_vaccinated = 0, people_vaccinated_per_hun
dred = 0, people_fully_vaccinated_per_hundred = 0)

country2020_zero
```

location <chr>	date <chr>	people_vaccinated <dbl>	people_fully_vaccinated <dbl>
Afghanistan	2020-12-31	0	0
Africa	2020-12-31	0	0
Albania	2020-12-31	0	0
Algeria	2020-12-31	0	0
Andorra	2020-12-31	0	0
Angola	2020-12-31	0	0
Anguilla	2020-12-31	0	0
Antigua and Barbuda	2020-12-31	0	0
Armenia	2020-12-31	0	0
Aruba	2020-12-31	0	0

1-10 of 186 rows | 1-4 of 6 columns Previous **1** 2 3 4 5 6 … 19 Next

Combine the two datasets into world_vacci_data_2020

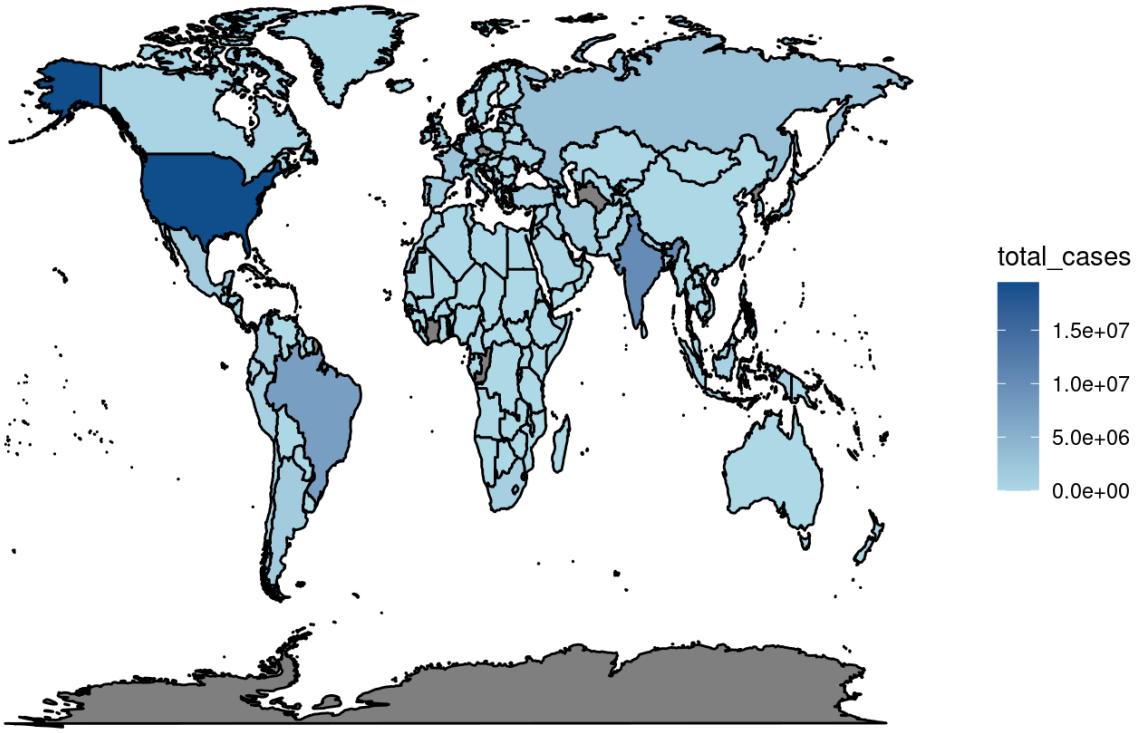
```
world_vacci_data_2020 = rbind(location_2020, country2020_zero)
world_vacci_data_2020 = world_vacci_data_2020 %>%
  select(-date)
```

Now, we have finished data cleaning and preparation for world maps vaccination rate.

World maps for covid-related data and analysis

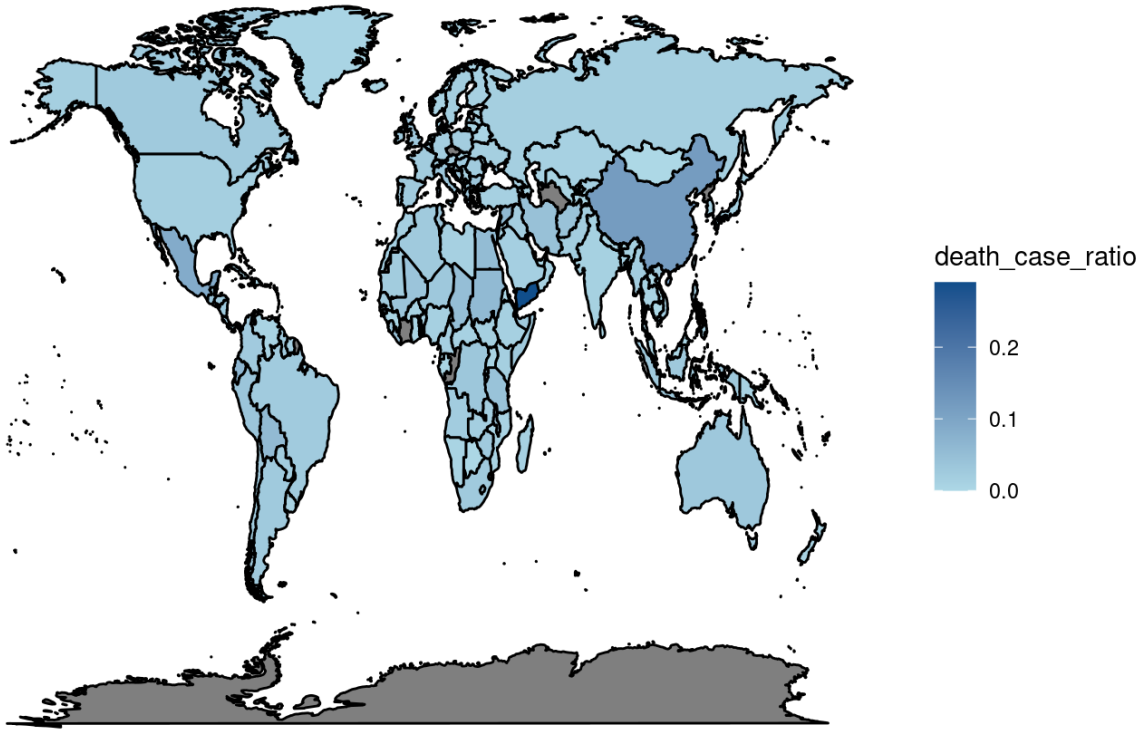
```
covid_data_2020 = readr::read_csv("https://raw.githubusercontent.com/illinois-stat447/fa21-prj-shiyuan8-sw20-zihanz12-zl32/master/data/covid_2020_total.csv?token=AKH4K33CEG4UJXXTNV7LLFLBXJTT2", show_col_types = FALSE)
#remove first index column

covid_data_2020 <- covid_data_2020[ -c(1) ]
# View(covid_data_2020)
covid_data_2020$region = covid_data_2020$`country`


mapdata <- map_data("world") ##ggplot2
# View(mapdata)

new_mapdata = mapdata |>
  mutate(region = replace(region, region == 'USA', 'United States'))

new_mapdata = new_mapdata |>
  mutate(region = replace(region, region == 'Democratic Republic of the Congo', 'Democratic Republic of Congo'))

new_mapdata = new_mapdata |>
  mutate(region = replace(region, region == 'UK', 'United Kingdom'))

new_mapdata = new_mapdata |>
  mutate(region = replace(region, region == 'Greenland', 'Denmark'))


# unique(new_mapdata[c("region")])

mapdata_and_covid_2020 <- left_join(new_mapdata, covid_data_2020, by="region")
# View(mapdata_and_covid_2020)



# View(unique(mapdata_and_covid_2020[c("region")]))

#remove na
# mapdata_and_covid_2020<-mapdata_and_covid_2020 |> filter(!is.na(mapdata_and_covid_2020$total_cases))
# View(mapdata_and_covid_2020)

map3<-ggplot(mapdata_and_covid_2020, aes( x = long, y = lat, group=group)) +
  geom_polygon(aes(fill = total_cases), color = "black") +
  ggtitle("total number of covid cases world map")

map4 <- map3 + scale_fill_gradient(name = "total_cases", low = "light blue", high =  "dodgerblue4", na.value = "grey50")+
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        rect = element_blank())
map4
```

total number of covid cases world map




```
map5<-ggplot(mapdata_and_covid_2020, aes( x = long, y = lat, group=group)) +
  geom_polygon(aes(fill = death_case_ratio), color = "black")+
  ggtitle("covid death ratio world map")

map6 <- map5 + scale_fill_gradient(name = "death_case_ratio", low = "light blue", high =  "dodgerblue4", n
a.value = "grey50")+
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        rect = element_blank())
map6
```

covid death ratio world map



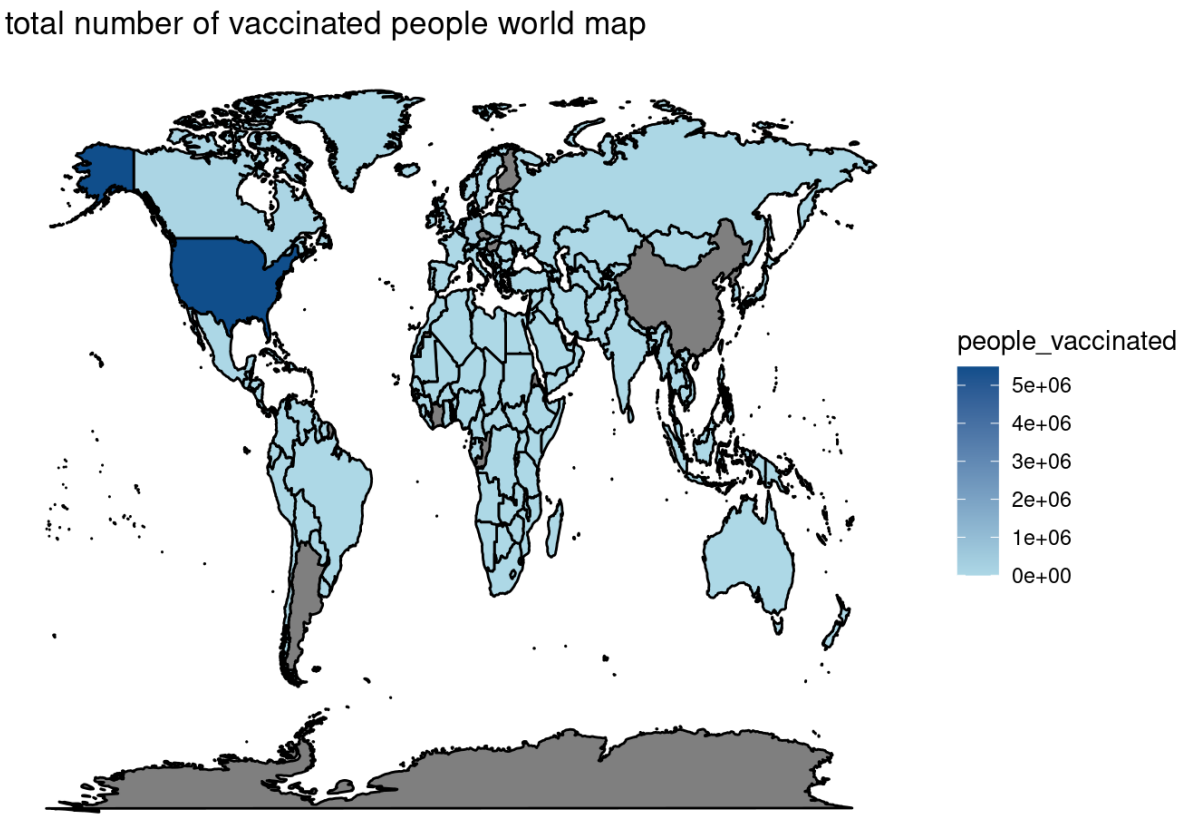

```
# require(gridExtra)
#
# grid.arrange(map6, map5, widths = c(10,10), ncol=2)
```

From the **total number of covid cases world map**, we can observe that countries such as United States, India and Brazil have very high number of covid infection rate. On the other hand, countries like China have lower covid infection rate.

However, from the **covid death ratio world map**, we can observe that countries like United States, India and Brazil, which have very high number of covid infection rate, now have a relatively lower death ratio. This interesting phenomenon could be due to effective vaccination or unreported/misrecorded death rate.

World maps for vaccination-related data and analysis

```
vaccination_data_2020 <- readr::read_csv("https://raw.githubusercontent.com/illinois-stat447/fa21-prj-shiyu
an8-sw20-zihanz12-zl32/master/vaccination/data/world_vacci_data_2020.csv?token=AKH4K37S6CBXCEW544XX6K3BXJ4C
K",show_col_types = FALSE)
vaccination_data_2020$region = vaccination_data_2020$`location`

mapdata_and_vaccination_2020 <- left_join(new_mapdata, vaccination_data_2020, by="region")

map5<-ggplot(mapdata_and_vaccination_2020, aes( x = long, y = lat, group=group)) +
  geom_polygon(aes(fill = people_vaccinated), color = "black")+
  ggtitle("total number of vaccinated people world map")

map6 <- map5 + scale_fill_gradient(name = "people_vaccinated", low = "light blue", high =  "dodgerblue4", n
a.value = "grey50")+
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks = element_blank(),
        axis.title.y=element_blank(),
        axis.title.x=element_blank(),
        rect = element_blank())
map6
```



From the **total number of vaccinated people world map**, we can observe that countries such as United States have very high vaccination rate. This justifies my previous assumption about United State's low death rate ratio. On the other hand, vaccination information is not provided for countries like China in the raw data we find.

The low vaccination rate in countries like India, Brazil and Russia could be one of the key factors that contributed to their high covid-related death incidents.

ANOVA Analysis and Determine significant factors

To examine the whether the differences between the covid data, vaccination data and world happiness data are statistically significant, we wish to conduct ANOVA.

Now we would like to merge covid data, vaccination data and world happiness data together to continue our ANOVA analysis.

```
world_happiness <- readr::read_csv("https://raw.githubusercontent.com/illinois-stat447/fa21-prj-shiyuan8-sw
20-zihanz12-zl32/master/data/world-happiness-report-2020.csv?token=AKH4K3ZTOOQABV5NEUJJGQLBXJOCQ",show_col_
types = FALSE)

world_happiness$region = world_happiness$`Country name`

world_happiness = world_happiness |>
  mutate(region = replace(region, region == 'USA', 'United States'))

world_happiness = world_happiness |>
  mutate(region = replace(region, region == 'Democratic Republic of the Congo', 'Democratic Republic of Con
go'))

world_happiness = world_happiness |>
  mutate(region = replace(region, region == 'UK', 'United Kingdom'))

world_happiness = world_happiness |>
  mutate(region = replace(region, region == 'Greenland', 'Denmark'))


covid_data_2020 = readr::read_csv("https://raw.githubusercontent.com/illinois-stat447/fa21-prj-shiyuan8-sw2
0-zihanz12-zl32/master/data/covid_2020_total.csv?token=AKH4K33CEG4UJXXTNV7LLFLBXJTT2",show_col_types = FALS
E)
```

```
## New names:
## * `` -> ...1
```

```
#remove first index column
covid_data_2020 <- covid_data_2020[ -c(1) ]
# View(covid_data_2020)
covid_data_2020$region = covid_data_2020$`country`


vaccination_data_2020 <- readr::read_csv("https://raw.githubusercontent.com/illinois-stat447/fa21-prj-shiyu
an8-sw20-zihanz12-zl32/master/vaccination/data/world_vacci_data_2020.csv?token=AKH4K37S6CBXCEW544XX6K3BXJ4C
K",show_col_types = FALSE)
vaccination_data_2020$region = vaccination_data_2020$`location`


covid_world_happiness <- left_join(world_happiness, covid_data_2020, by="region")
vaccination_covid_world_happiness_2020 <- left_join(covid_world_happiness, vaccination_data_2020, by="regio
n")
vaccination_covid_world_happiness_2020
```

Country name <chr>	**Regional indicator** <chr>	**Ladder score** <dbl>
Finland	Western Europe	7.842
Denmark	Western Europe	7.620
Switzerland	Western Europe	7.571

Country name <chr>	Regional indicator <chr>	Ladder score <dbl>
Iceland	Western Europe	7.554
Netherlands	Western Europe	7.464
Norway	Western Europe	7.392
Sweden	Western Europe	7.363
Luxembourg	Western Europe	7.324
New Zealand	North America and ANZ	7.277
Austria	Western Europe	7.268

1-10 of 149 rows | 1-3 of 32 columns Previous **1** 2 3 4 5 6 … 15 Next

```
# write.csv(vaccination_covid_world_happiness_2020,"/cloud/project/vaccination_covid_world_happiness_2020.csv")
```

In this analysis we consider all the data and want to examine happiness variables, covid-related variables and vaccination-related variables, explaining the happiness ladder score.

```
# head(vaccination_covid_world_happiness_2020)

fit1 <- aov(`Ladder score` ~ `Generosity`+ `Perceptions of corruption`+ `total_cases` +`total_deaths` +`people_fully_vaccinated_per_hundred`+`total_cases_per_100k`+ `total_deaths_per_100k`+ `death_case_ratio` + `people_vaccinated` + `people_fully_vaccinated` + `people_vaccinated_per_hundred`+ `Freedom to make life choices` +   + `Logged GDP per capita`+ `Social support`+ `Healthy life expectancy`, vaccination_covid_world_happiness_2020)

anova(fit1)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Generosity	1	0.1860204	0.1860204	0.6467329	4.232497e-01
`Perceptions of corruption`	1	26.4689799	26.4689799	92.0241133	1.015261e-15
total_cases	1	2.5999653	2.5999653	9.0392416	3.364534e-03
total_deaths	1	3.8032436	3.8032436	13.2226523	4.450872e-04
people_fully_vaccinated_per_hundred	1	0.6554219	0.6554219	2.2786908	1.344131e-01
total_cases_per_100k	1	20.9247853	20.9247853	72.7487355	1.988698e-13
total_deaths_per_100k	1	0.4887979	0.4887979	1.6993929	1.954529e-01
death_case_ratio	1	1.7859926	1.7859926	6.2093207	1.440442e-02
people_vaccinated	1	2.5986568	2.5986568	9.0346922	3.372193e-03
people_fully_vaccinated	1	0.6114300	0.6114300	2.1257452	1.480725e-01

1-10 of 16 rows Previous **1** 2 Next

For this data, we will take p-value = 0.05 as a bench mark and we will run our analysis on significant factors based on this p-value.

According to the Analysis of Variance Table, we can conclude the following:

1. Generosity is not a significant predictor of happiness ladder score as its Pr > F value (0.4232497) is larger than 0.05. Therefore, we have fail to reject the null hypothesis that intervals is a non-significant predictor.
2. Perceptions of corruption is a significant predictor of happiness ladder score as its Pr > F value (1.015e-15) is smaller than 0.05. Therefore, we have to reject the null hypothesis that year is a non-significant predictor.
3. total_cases is a significant predictor of happiness ladder score as its Pr > F value (0.0033645) is smaller than 0.05. Therefore, we have to reject the null hypothesis that year is a non-significant predictor.
4. total_deaths is a significant predictor of happiness ladder score as its Pr > F value (0.0004451) is smaller than 0.05. Therefore, we have to reject the null hypothesis that species is a non-significant predictor.
5. people_fully_vaccinated_per_hundred is not a significant predictor of happiness ladder score as its Pr > F value (0.1344131) is larger than 0.05. Therefore, we have fail to reject the null hypothesis that intervals is a non-significant predictor.
6. total_cases_per_100k is a significant predictor of happiness ladder score as its Pr > F value (1.989e-13) is smaller than 0.05. Therefore, we have to reject the null hypothesis that species is a non-significant predictor.
7. total_deaths_per_100k is not a significant predictor of happiness ladder score as its Pr > F value (0.1954529) is larger than 0.05. Therefore, we have fail to reject the null hypothesis that intervals is a non-significant predictor.
8. death_case_ratio is a significant predictor of happiness ladder score as its Pr > F value (0.0144044) is smaller than 0.05. Therefore, we have to reject the null hypothesis that species is a non-significant predictor.
9. people_vaccinated is a significant predictor of happiness ladder score as its Pr > F value (0.0033722) is smaller than 0.05. Therefore, we have to reject the null hypothesis that species is a non-significant predictor.
10. people_fully_vaccinated is not a significant predictor of happiness ladder score as its Pr > F value (0.1480725) is larger than 0.05. Therefore, we have fail to reject the null hypothesis that intervals is a non-significant predictor.
11. people_vaccinated_per_hundred is not a significant predictor of happiness ladder score as its Pr > F value (0.4607941) is larger than 0.05. Therefore, we have fail to reject the null hypothesis that intervals is a non-significant predictor.
12. Freedom to make life choices is a significant predictor of happiness ladder score as its Pr > F value (1.301e-12) is smaller than 0.05. Therefore, we have to reject the null hypothesis that species is a non-significant predictor.
13. Logged GDP per capita is a significant predictor of happiness ladder score as its Pr > F value (4.285e-12) is smaller than 0.05. Therefore, we have to reject the null hypothesis that species is a non-significant predictor.
14. Social support is a significant predictor of happiness ladder score as its Pr > F value (0.0001988) is smaller than 0.05. Therefore, we have to reject the null hypothesis that species is a non-significant predictor.
15. Healthy life expectancy is not a significant predictor of happiness ladder score as its Pr > F value (0.0725201) is larger than 0.05. Therefore, we have fail to reject the null hypothesis that intervals is a non-significant predictor.

In conclusion, we can conclude that perceptions of corruption, total covid cases, total covid deaths, total covid cases per 100k, covid death case ratio, number of vaccinated people, freedom to make life choices, logged GDP per capita and social support are significant predictor of happiness of a country's citizens.