Data Analysis and Visualization in R

Homework Assignment for week 10

Gowri Anand, Jason Urias, Oliver Wiegel, Tom Blaber, Bibhash Nath

4/20/2021

Homework assignment

Group assignment

Attempt a preliminary answer to the following question using between one and five charts: **Over the course of the pandemic, were countries with lower life expectancies hit harder by COVID-19 than countries with higher life expectancies?** Submit your answer in an annotated RMarkdown file with code and explanation for each chart created and the summary of the findings of your study.

Do not use the rstudioapi library, as it will not work with RMarkdown. To create the charts use the provided datasets from: World Bank for life expectancy, Johns Hopkins for COVID-19, UN for country population and country code data. You may add further datasets if necessary, but the number of charts are limited. Only provide charts that help to answer the question directly.

One person from your group will present your findings during the first part of the next class. That person will be given 5 minutes to present and explain their results. ***Therefore you are not allowed to communicate findings or techniques between groups***.

Problem:

Over the course of the pandemic, were countries with lower life expectancies hit harder by COVID-19 than countries with higher life expectancies?

Hypothesis:

Countries with higher life expectancies were hit harder by COVID-19 than countries with lower life expectancies because of a lack of preparedness and preventive measures while also being hubs for global travel.

Methodology:

- We used COVID-19 deaths standardized by population and categorized them by either GDP (Gross Domestic Product) or average life expectancy quartiles to explain the progression of the virus and how life expectancy is related to COVID-19 deaths in different countries.
- We used GDP as a general proxy for countries that likely have extensive trade networks and a higher rate of international business and travel which would indicate a higher rate of transmission of the virus.
- We used the ISO (International Organization for Standardization) country codes to join files together as it was the most common column among the datasets.

- We used GDP data for each country as another potential factor to indicate life expectancy. In general, higher income nations have higher life expectancies.
- We included median-age data for each country to find how this relates to life expectancies in different countries.
- We used COVID-19 cases and deaths data for each country to show number of COVID cases and deaths for each country.
- We used average life expectancy data for each country to see if this has any relation to COVID-19 cases or deaths.
- We used total population data from 2020 for each country to standardize the COVID-19 cases and deaths.
- We included 65+ population percentage data for each country to find how this relates to life expectancies in different countries.
- We used Tuberculosis (TB) incidence data for each country as another potential factor to indicate the likelihood of contracting the virus, since risk factors for both TB and COVID-19 are similar, i.e, proximity, duration of contact, crowding, poor ventilation, etc.
- We joined all the dataframes to visualize the relationships between average life expectancies and COVID-19 cases and deaths in different countries.
- To look at country-wise COVID-19 data in finer detail the daily COVID-19 cases and deaths was also examined.

Results:

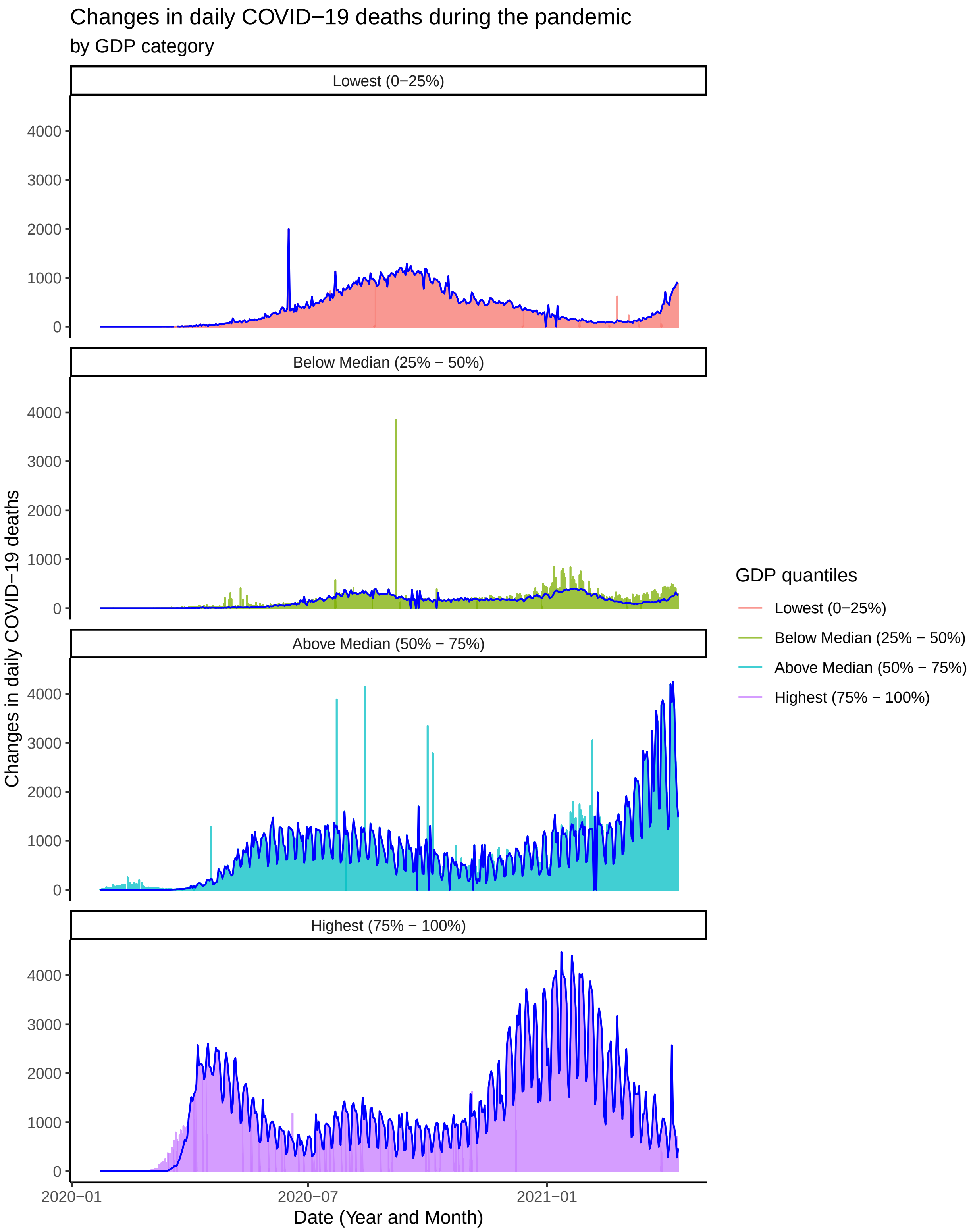
- We chose COVID-19 deaths standardized by total population to show how this pandemic relates to life expectancies in different countries and explained their relationships using GDP and TB incidence in those countries.

Plot_1: Daily COVID-19 deaths during the pandemic per GDP category

```
covid_deaths_Q1 %>%
  drop_na(GDP_quants) %>%
  ggplot(aes(Date, Daily_deaths, color=GDP_quants)) +
  geom_line(alpha = 0.75) +
  geom_line(data=cd_top4C, aes(Date, Daily_deaths), color="blue", show.legend = FALSE) +
  ylim(0,4500) +
  theme(legend.position="none") +
  labs(x = "Date (Year and Month)", y= "Changes in daily COVID-19 deaths", title = "Changes in daily CO
  theme_classic() +
  scale_color_discrete(name = "GDP quantiles") + #label legend title
  facet_wrap('GDP_quants', nrow = 4)
```

```
## Warning: Removed 174 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



The figure above shows the daily changes in COVID-19 deaths for all countries. The overlapping lines make it difficult to show every country. Each of the four figures has been separated and placed into quartiles by GDP per capita (USD). Though not shown here, there is a positive relationship between GDP per capita and average life expectancy of a country. It is based on this relationship that this data is included in our analysis.

The thick blue lines in each of the sub-plots above show a selected country from a given GDP quartile. The selections were: the United States from the 75%-100% quartile, Brazil from the 50%-75% quartile, Colombia from the 25%-50% quartile, and India from the 0-25% quartile. The curves for these different GDP quartiles seem to indicate that the peak of the first wave of the pandemic hit these countries at different times. The countries with the highest GDP peaked first, during March/April of 2020, followed by countries with above median GDP in May/June of 2020, then below median GDP countries in August of 2020, and lastly the countries with the lowest GDP during September of 2020.

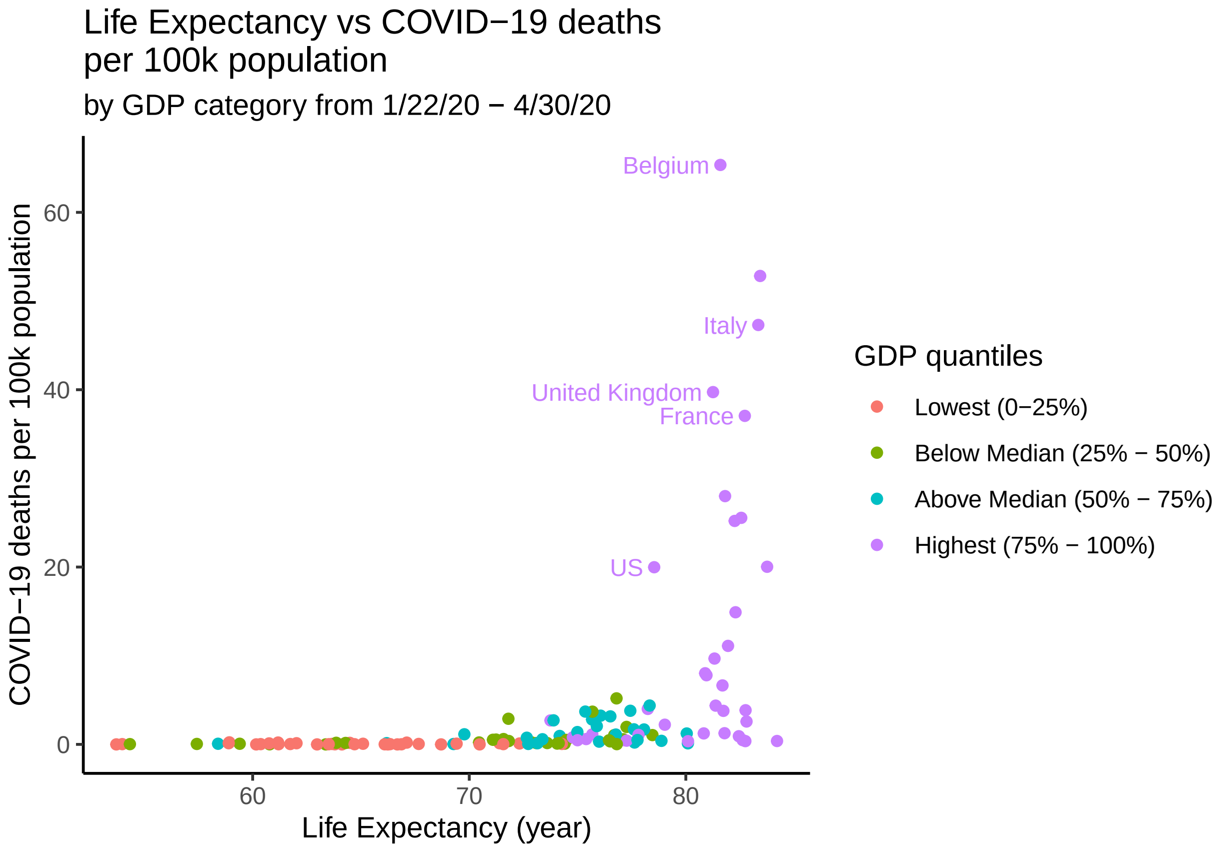
Plot_2: COVID-19 deaths during first wave (end of April) against life expectancy and GDP

```
country_deaths_Apr <- read_csv("time_series_covid19_deaths_global.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   ‘Province/State‘ = col_character(),
##   ‘Country/Region‘ = col_character()
## )
## i Use ‘spec()‘ for the full column specifications.
```

```
country_deaths_Apr <- country_deaths_Apr %>%
  group_by(`Country/Region`) %>%
  gather(key = Date, value = Deaths, "4/30/20") %>% #get deaths numbers from beginning of pandemic to e
  select(`Country/Region`, Deaths)%>%
  group_by(`Country/Region`) %>%
  summarise(AprDeaths = sum(Deaths)) %>% #Each country's April deaths
  rename(Country = `Country/Region`)

df6 <- df5 %>%
  left_join(country_deaths_Apr, by = c("Country")) %>% #join dataset that comprises of all 9 datasets w
  mutate(Deaths_100k = round((AprDeaths/Total_population*100000),2)) #get April deaths per 100k populat

top4c1 <- c("US", "Italy", "United Kingdom", "France", "Belgium")

df6a <- df6 %>% filter(Country %in% top4c1)

df6 %>%
  drop_na(GDP_quants) %>%
  ggplot(aes(Life_Expectancy, Deaths_100k, color=GDP_quants)) +
  geom_point() +
  geom_text(data = df6a, aes(label = Country), size = 3, nudge_x = -0.5, hjust = 1, show.legend = FALSE
  labs(x = "Life Expectancy (year)", y= "COVID-19 deaths per 100k population", title = "Life Expectancy
  theme_classic() +
  scale_color_discrete(name = "GDP quantiles") #label legend title
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



The figure above shows that during the earliest phases of the pandemic (January-April 2020), the countries with the highest life expectancy (typically >80 years) and highest GDP had the highest number of deaths based on population.

Plot_3: COVID-19 deaths during second wave (end of July) against life expectancy and GDP

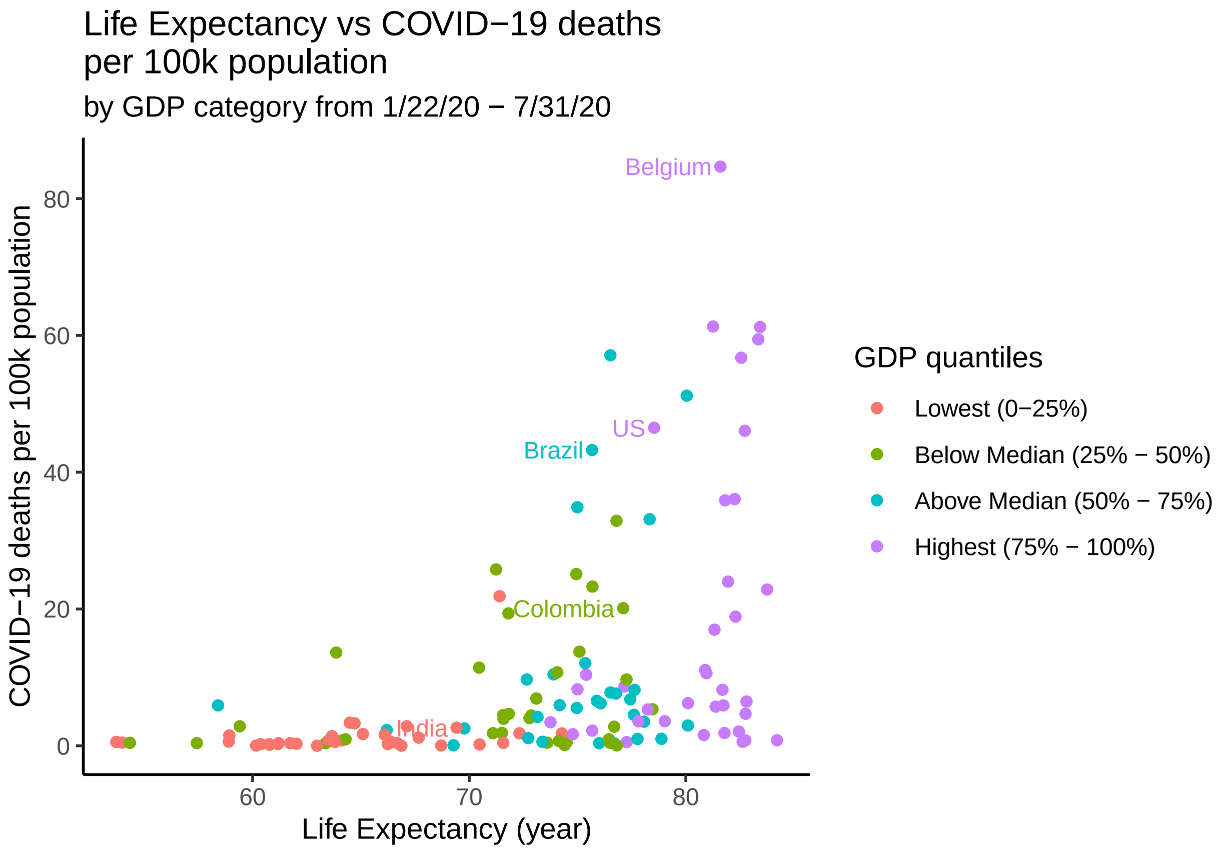
```
country_deaths_Jul <- read_csv("time_series_covid19_deaths_global.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   ‘Province/State‘ = col_character(),
##   ‘Country/Region‘ = col_character()
## )
## i Use ‘spec()‘ for the full column specifications.
```

```
country_deaths_Jul <- country_deaths_Jul %>%
  group_by(`Country/Region`) %>%
  gather(key = Date, value = Deaths, "7/31/20") %>%
  select(`Country/Region`, Deaths)%>%
  group_by(`Country/Region`) %>%
  summarise(JulDeaths = sum(Deaths)) %>%
  rename(Country = `Country/Region`)
```

```
df7 <- df5 %>%
  left_join(country_deaths_Jul, by = c("Country")) %>%
  mutate(Deaths_100k = round((JulDeaths/Total_population)*100000,2))

top4c2 <- c("India", "US", "Colombia", "Brazil", "Belgium")
df7a <- df7 %>% filter(Country %in% top4c2)

df7 %>% drop_na(GDP_quants) %>%
  ggplot(aes(Life_Expectancy, Deaths_100k, color=GDP_quants)) +
  geom_point() +
  geom_text(data = df7a, aes(label = Country), size = 3, nudge_x = -0.4, hjust = 1, show.legend = FALSE
  labs(x = "Life Expectancy (year)", y= "COVID-19 deaths per 100k population", title = "Life Expectancy
  theme_classic() +
  scale_color_discrete(name = "GDP quantiles")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



The figure above shows that with time (January-July 2020) the pandemic began to have a greater impact on countries with below and above median GDP (per capita in USD). The figure also shows how COVID-19 deaths had began increasing for both Brazil (Q2 GDP) and Colombia (Q3 GDP) while the death rate for India (Q1 GDP) remained low.

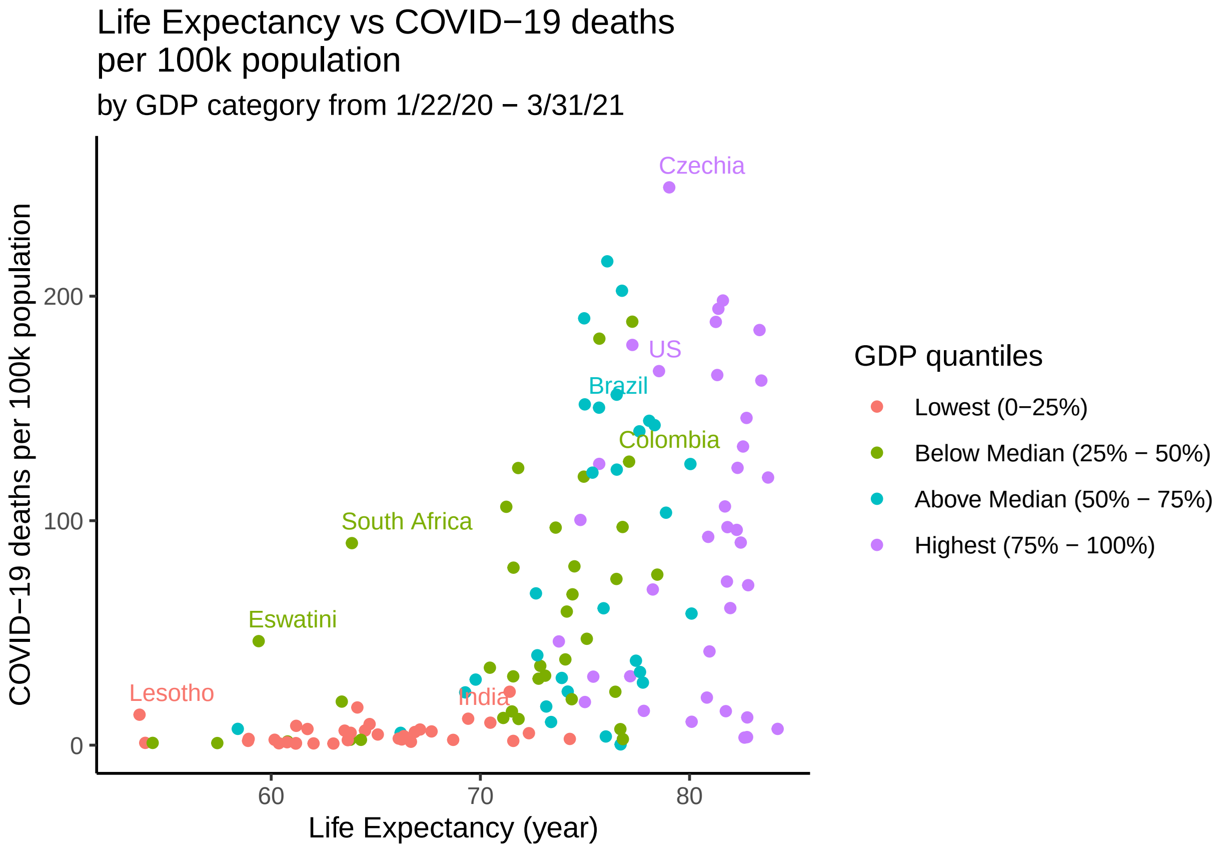
Plot_4: COVID-19 deaths at the end of March 2021 against life expectancy and GDP

```
country_deaths_Mar21 <- read_csv("time_series_covid19_deaths_global.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double(),
##   ‘Province/State‘ = col_character(),
##   ‘Country/Region‘ = col_character()
## )
## i Use ‘spec()‘ for the full column specifications.
```

```
country_deaths_Mar21 <- country_deaths_Mar21 %>%
  group_by(`Country/Region`) %>%
  gather(key = Date, value = Deaths, "3/31/21") %>%
  select(`Country/Region`, Deaths)%>%
  group_by(`Country/Region`) %>%
  summarise(MarDeaths = sum(Deaths)) %>%
  rename(Country = `Country/Region`)

df8 <- df5 %>%
  left_join(country_deaths_Mar21, by = c("Country")) %>%
  mutate(Deaths_100k = round((MarDeaths/Total_population)*100000,2))

top4c3 <- c("India", "US", "Colombia", "Brazil", "Czechia", "South Africa", "Eswatini", "Lesotho")
df8a <- df8 %>% filter(Country %in% top4c3)

df8 %>%
  drop_na(GDP_quants) %>%
  ggplot(aes(Life_Expectancy, Deaths_100k, color=GDP_quants)) +
  geom_point() +
  geom_text(data = df8a, aes(label = Country), size = 3, nudge_x = -0.5, nudge_y = 10, hjust = 0, show.l
  labs(x = "Life Expectancy (year)", y= "COVID-19 deaths per 100k population", title = "Life Expectancy
  theme_classic() +
  scale_color_discrete(name = "GDP quantiles")
```

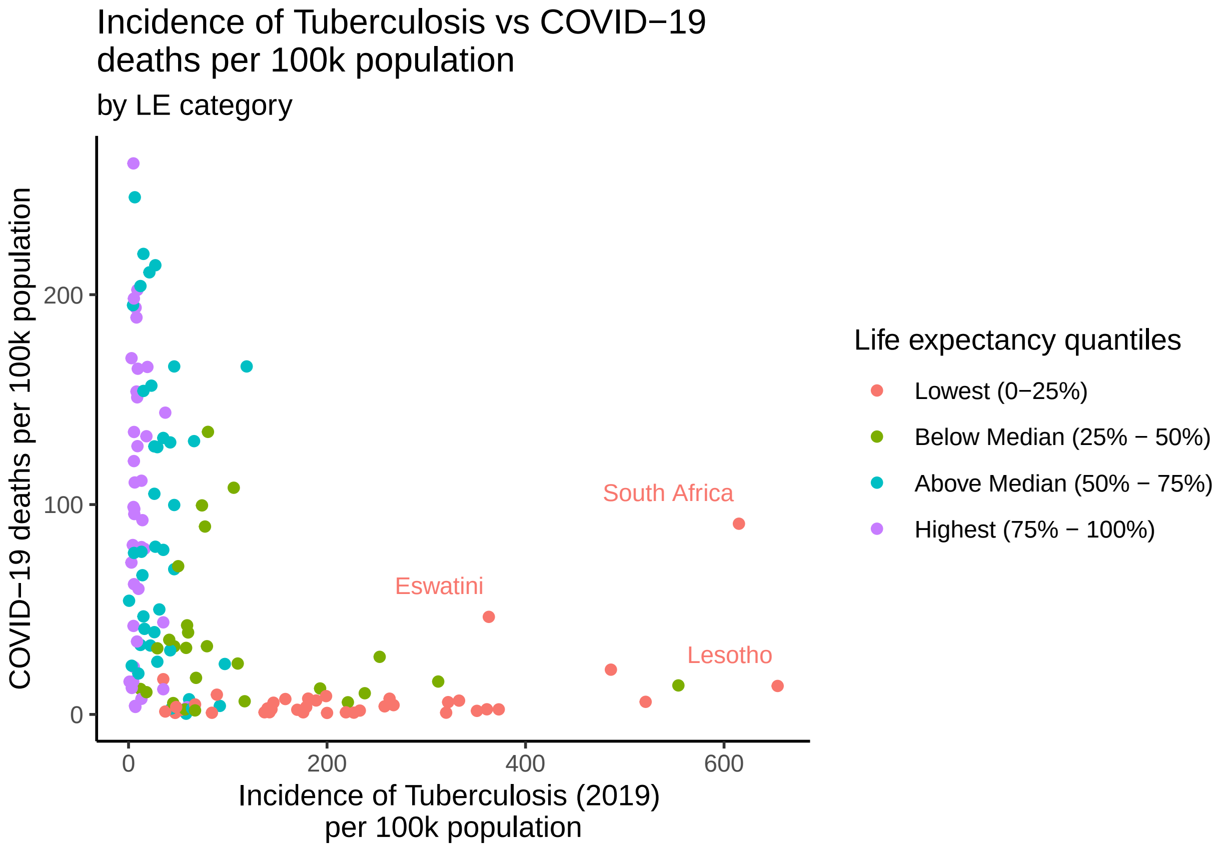
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



The figure above shows the world today (January 2020-March 2021). We can see that death rates in the US, Brazil and Colombia are comparable but the death rate in India has remained low. As can be seen from these figures, countries with higher life expectancies (>80 years) were the first to be impacted by Covid-19. However, as the pandemic progressed and to today, countries with life expectancy >72 years have started to encounter more COVID-19 deaths (per 100,000 people).

Plot_5: Tuberculosis (TB) cases against COVID-19 deaths to see if risk factors for TB is in relation to COVID-19 deaths

```
tbc3 <- c("South Africa", "Eswatini", "Lesotho")
df5tbc3 <- df5 %>% filter(Country %in% tbc3)

df5 %>%
  drop_na(LE_quants) %>%
  ggplot(aes(TB_per100k_pop_yr19, Deaths_per_100k, colour=LE_quants)) +
  geom_point() +
  labs(x = "Incidence of Tuberculosis (2019) \nper 100k population", y= "COVID-19 deaths per 100k popula
  geom_text(data = df5tbc3, aes(label = Country), size = 3, nudge_x = -5, nudge_y = 15, hjust = 1, show
  theme_classic() +
  scale_color_discrete(name = "Life expectancy quantiles")
```



In this figure we showed the relationship between the incidence of Tuberculosis (TB) and COVID-19 deaths in different countries. Since the risk factors for both COVID-19 and TB are similar, it may indicate a similar rate of transmission and deaths. Interestingly, the data shows that countries with low incidence of TB were more impacted by COVID-19 than countries with high incidences of TB. While there are likely several factors that cause this correlation, there are reports that there may be a link between the TB vaccine and protection from COVID-19. "TB Vaccine Linked to Lower Risk of Contracting COVID-19" ([https://www.cedars-sinai.org/newsroom/study-tb-vaccine-linked-to-lower-risk-of-contracting-covid-19/]).

Final remarks

As can be shown from the data above, there is a combination of several factors that may influence how and why COVID-19 spread the way that it did. Factors such as the virus host, community interaction, preventive measures by the individual or by the governments, intervention strategy by the government (including vaccination) and perhaps natural immunity as a result of the TB vaccine should all be considered when looking at the disparity of COVID-19 deaths between low- and high- life expectancy countries.

We found that the initial wave of COVID-19 impacted the countries with the highest GDP and highest life expectancy first and those peaks occurred with the lower GDP and lower life expectancy countries at later dates. While the death rate within some countries remains low, as a general trend all nations were impacted by high rates of COVID-19 deaths but at different points in time.

What is also clear is that disaster preparedness is paramount. As is indicated by the data many of the countries that first encountered COVID-19, despite being both wealthy and healthy, were unable to handle the spread of outbreak. As the virus spread and other countries had more time to prepare the instance rate of COVID-19 deaths began to subside. Clearly this is not the case in every country, and even today cases are spiking again in many areas, but the data shows that preventative measures are the key to stopping the spread of this virus.

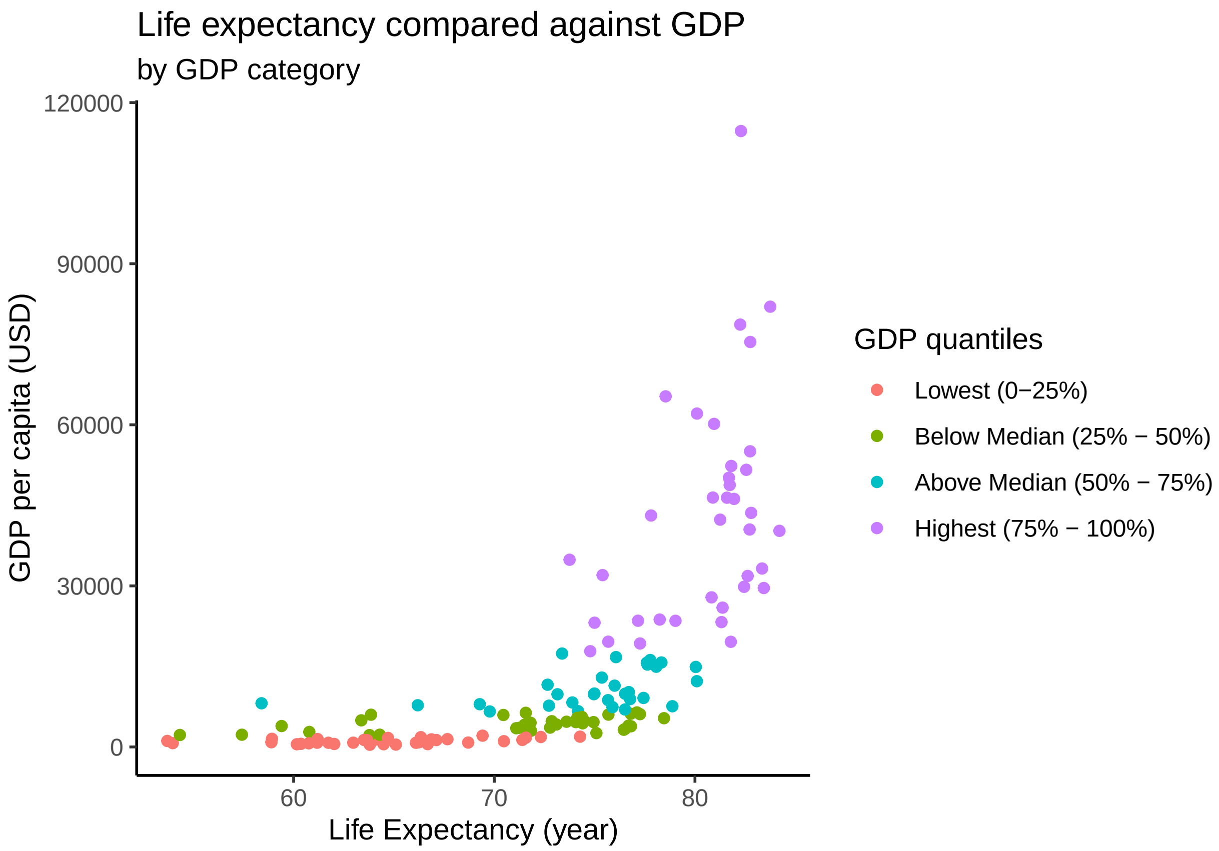
The final note is with regards to the flow of human movement within each country which can be a large factor as it relates to the rates of diffusion for the virus. In higher life expectancy and higher GDP countries there is more movement of individuals both nationally and internationally as compared to the low life expectancy and low GDP countries. This is reflected not only though recreational travel but, business, trade, etc. These higher GDP nations also usually have larger metropolitan cities, more industries, more air-conditioned and enclosed work spaces (indoor living) and many dense urban populations as compared to lower life expectancy and lower GDP countries.

Supplementary Figures (for viewing only):

Life expectancy vs GDP

```
df5 %>%
  drop_na(GDP_quants) %>%
  ggplot(aes(Life_Expectancy, GDP_PC, colour=GDP_quants)) +
  geom_point() +
  theme_classic() +
  labs(x = "Life Expectancy (year)", y= "GDP per capita (USD)", title = "Life expectancy compared agains
  theme_classic() +
  scale_color_discrete(name = "GDP quantiles") #label legend title
```

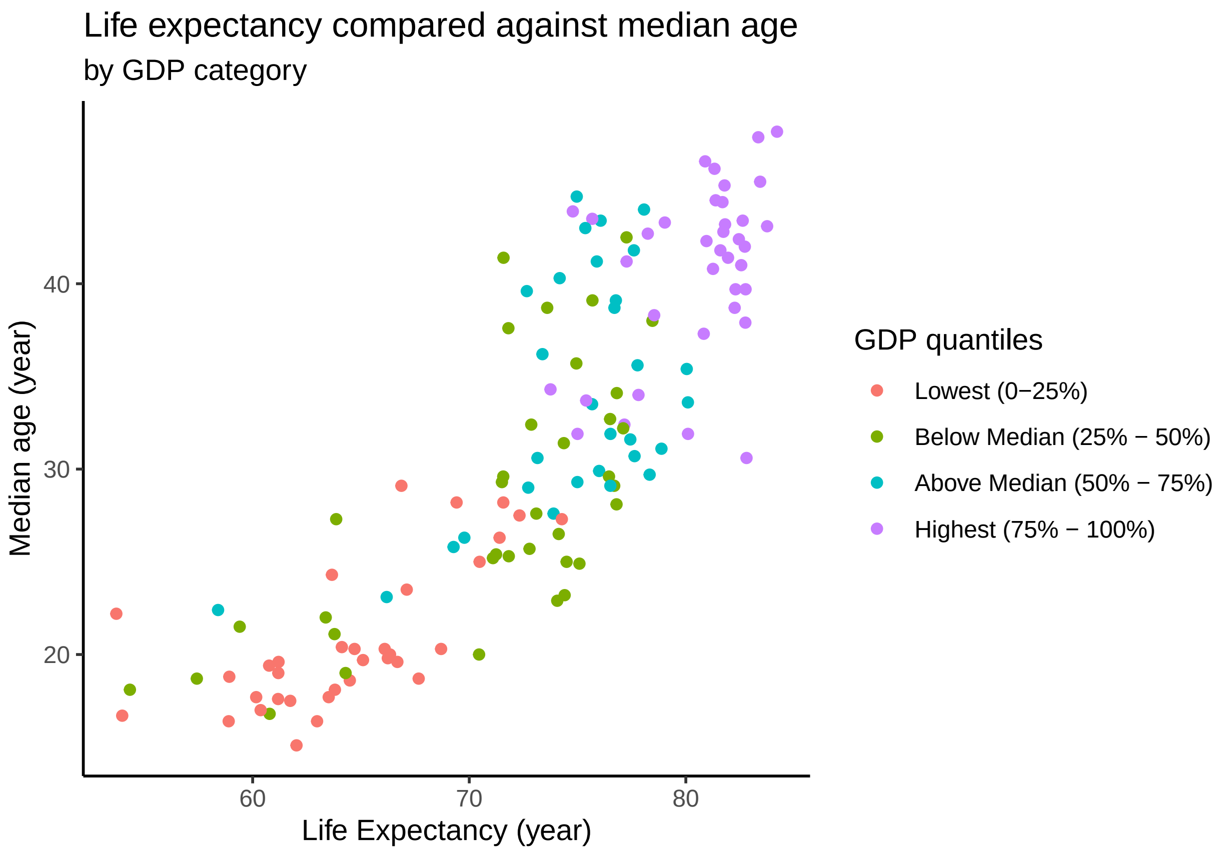
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Life expectancy vs median-age

```
df5 %>%
  drop_na(GDP_quants) %>%
  ggplot(aes(Life_Expectancy, Median_Age, colour=GDP_quants)) +
  geom_point() +
  theme_classic() +
  labs(x = "Life Expectancy (year)", y= "Median age (year)", title = "Life expectancy compared against m
  theme_classic() +
  scale_color_discrete(name = "GDP quantiles") #label legend title
```

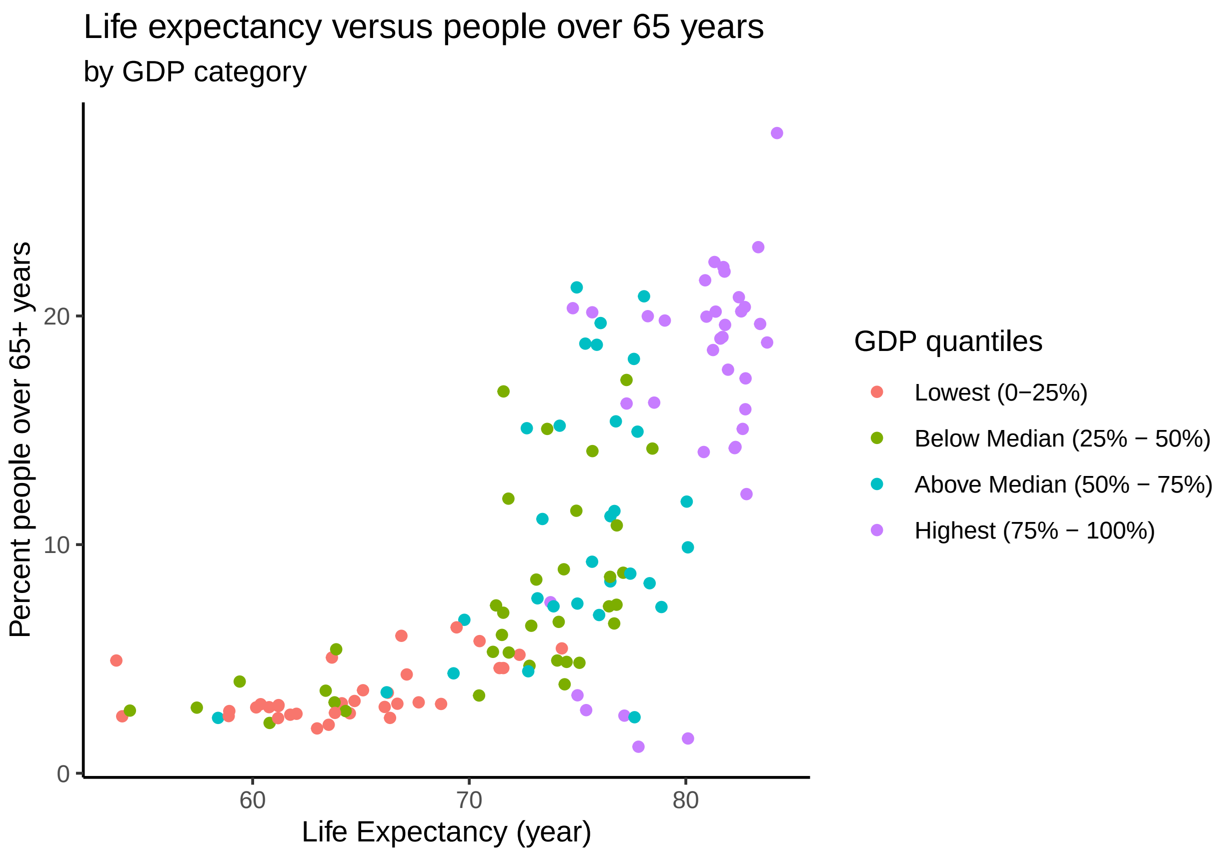
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Life expectancy vs 65+ people

```
df5 %>%
  drop_na(GDP_quants) %>%
  ggplot(aes(Life_Expectancy, pop65plus_perTotalpop, colour=GDP_quants)) +
  geom_point() +
  theme_classic() +
  labs(x = "Life Expectancy (year)", y= "Percent people over 65+ years", title = "Life expectancy versu
  theme_classic() +
  scale_color_discrete(name = "GDP quantiles") #label legend title
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```


Life expectancy versus people over 65 years
by GDP category
Percent people over 65+ years
0
10
20
60
70
80
Life Expectancy (year)
GDP quantiles
● Lowest (0−25%)
● Below Median (25% − 50%)
● Above Median (50% − 75%)
● Highest (75% − 100%)