DATA100 Project: Exploring the Impact of Development on COVID-19 Statistics

Chris Chan

09/12/2021

Table of Contents

Introduction:

I learned about the human development index (HDI) in one of my other courses and decided that it would be an interesting topic. HDI is a relative measurement of a country's economic and social development, and is based on three areas of interest: health care, denoted by life expectancy; education, denoted by the average years of schooling; and standard of living, denoted by the country's GNI per capita. For the sake of this project, HDI will be viewed in a more general sense as a measure of the degree of a country's development - that is, countries with higher HDI values are more developed (what that specifically means is beyond this report) than those with lower HDI values.

This report includes the use of 12 data sets (the descriptions of which can be found below), 7 of which concern 4 potential factors of interest: educational, economical, health, and infrastructure. With these data sets, this report will showcase the exploration of the data in determining whether or not these factors are related to the level of observed COVID-related statistics at a given time and given region. In doing so, this report will also attempt to answer two more questions:

1. Is there a relationship between a countries' degree of development and the COVID-related statistics observed?
2. Is there a particular area of development - given by the concept of the HDI - that is strongly correlated with the amount, or lack there of, of COVID-related statistics?

This project will be particularly interested in the relationship on particular dates: March 15, 2021 and September 9, 2021. The first date was chosen because it is roughly a year after COVID-19 became a true global concern (I remember it was around this date in 2020 when schools were shut down in Ontario, which was already considerably late from a global perspective). It also stands as a point of time before the vaccine had been developed and distributed. The second date was chosen because it marked the start of the fall term for post-secondary students in Ontario. This is interesting because vaccines should have been largely distributed in preparation for the return to school.

On these dates, there is an expectation that health as a factor is most likely to affect the COVID deaths, tests, and vaccinations of a region. This also coincides with another factor to be explored: infrastructure; a country with better health infrastructure has a better chance of dealing with the virus. General infrastructure is also likely to coincide

with COVID-statistics. For example, I expect more urbanized countries to have more cases of infection due to proximity. This leaves education as the least likely factor to affect the numbers of COVID-19; the virus doesn't care how educated a population is, it infects everyone equally.

Data Set Descriptions

Note: Data set names are given as downloaded from the MLS page and from the provided links.

Health:

- **LIFEEXPECTANCYATBIRTH.csv**: Overall average life expectancy in years, as well as between gender demographics.
- **DEATHRATE.csv**: Death rate in the region per 1000 population. Also gives the countries' relative rank and the year in which the data is from.
- **HEALTHEXP.csv**: Annual health expenditure in the region, given by USD per capita from 2000 to 2020

Standard of Living:

- **INTERNETUSER.csv**: The annual percentage of a region's population that has access to the internet from 2000 to 2020
- **URBANIZATION.csv**: The annual percentage of the population living in urban areas from 2000 to 2020
- **WorldHappinessReport2021-Score.csv**: A number between 0 and 10 which measures the happiness of a region. Is a result of a survey in which people were asked to give a number between 0 and 10 which represents how they currently feel in life and concerns the years from 2018 to 2020.

Education:

- **EDUEXP.csv**: The percentage of a region's GDP spent on education. Also ranks the region's from highest to lowest, states which broader region it belongs to, and gives the year the percentage concerns.

COVID:

- **covid_complete.csv**: A complete set of data for each region, extending beyond COVID-related statistics into areas including population distribution and infrastructure. For this project, only a subset of the COVID-related statistics were taken into account. Records daily observations from the early days of 2020 to the data sets' last update.
- **covid_response.csv**: Daily observations of governments' responses to COVID. Concerns areas such as the status of school closures and stay-at-home policy.
- **covid_tests.csv**: Daily observations concerning a region's COVID testing totals and rates.
- **covid_vaccinations.csv**: Daily observations concerning a region's vaccination totals and rates.

Miscellaneous:

- **WorldRegions.csv**: Classifies regions as being part of the Global South or the Global North

Descriptive Statistics and Exploratory Analysis

Let's begin by looking at the human development index (HDI) of each country, given by the data set 'covid_complete.csv'.

```
## # A tibble: 237 x 3
##    Country             iso_code human_development_index
##    <chr>               <chr>                      <dbl>
##  1 Afghanistan         AFG                        0.511
##  2 Africa              OWID_AFR                  NA
##  3 Albania             ALB                        0.795
##  4 Algeria             DZA                        0.748
##  5 Andorra             AND                        0.868
##  6 Angola              AGO                        0.581
##  7 Anguilla            AIA                       NA
##  8 Antigua and Barbuda ATG                        0.778
##  9 Argentina           ARG                        0.845
## 10 Armenia             ARM                        0.776
## # ... with 227 more rows
```

For the purposes of this project, I will be exploring the extremes of this data: the most “developed” countries, and the least, because if there is any relationship to be seen between COVID-19 and how developed a country is, it will be the most evident at the extremes, between the most and least developed countries.

The first table lists the 20 countries/regions with the highest HDI rating. The second table lists the countries/regions with the lowest HDI rating.

```
## # A tibble: 20 x 3
##    Country        iso_code human_development_index
##    <chr>          <chr>                      <dbl>
##  1 Norway         NOR                        0.957
##  2 Ireland        IRL                        0.955
##  3 Switzerland    CHE                        0.955
##  4 Hong Kong      HKG                        0.949
##  5 Iceland        ISL                        0.949
##  6 Germany        DEU                        0.947
##  7 Sweden         SWE                        0.945
##  8 Australia      AUS                        0.944
##  9 Netherlands    NLD                        0.944
## 10 Denmark        DNK                        0.94
## 11 Finland        FIN                        0.938
## 12 Singapore      SGP                        0.938
## 13 United Kingdom GBR                        0.932
## 14 Belgium        BEL                        0.931
## 15 New Zealand    NZL                        0.931
## 16 Canada         CAN                        0.929
## 17 United States  USA                        0.926
## 18 Austria        AUT                        0.922
## 19 Israel         ISR                        0.919
## 20 Japan          JPN                        0.919
```

```
## # A tibble: 20 x 3
##    Country                      iso_code human_development_index
##    <chr>                        <chr>                      <dbl>
##  1 Niger                        NER                        0.394
##  2 Central African Republic     CAF                        0.397
##  3 Chad                         TCD                        0.398
##  4 Burundi                      BDI                        0.433
##  5 South Sudan                  SSD                        0.433
##  6 Mali                         MLI                        0.434
##  7 Burkina Faso                 BFA                        0.452
##  8 Sierra Leone                 SLE                        0.452
##  9 Mozambique                   MOZ                        0.456
## 10 Eritrea                      ERI                        0.459
## 11 Yemen                        YEM                        0.47
## 12 Guinea                       GIN                        0.477
## 13 Democratic Republic of Congo COD                        0.48
## 14 Guinea-Bissau                GNB                        0.48
## 15 Liberia                      LBR                        0.48
## 16 Malawi                       MWI                        0.483
## 17 Ethiopia                     ETH                        0.485
## 18 Gambia                       GMB                        0.496
## 19 Haiti                        HTI                        0.51
## 20 Sudan                        SDN                        0.51
```

To avoid any errors, I will need to determine which of these 40 countries actually have complete data. Using various ‘semi_joins()’, along with the combined data sets which represent the countries for which we have complete data, we obtain the following. For the purposes of this project, the analysis will center around the top four countries in the resulting tables.

On a side note, Unsurprisingly, only a fraction of the “least developed” countries have data, compared to the “most developed”, where all but one do.

```
## Joining, by = c("Country", "iso_code")
```

```
## # A tibble: 19 x 3
##    Country        iso_code human_development_index
##    <chr>          <chr>                      <dbl>
##  1 Norway         NOR                        0.957
##  2 Ireland        IRL                        0.955
##  3 Switzerland    CHE                        0.955
##  4 Iceland        ISL                        0.949
##  5 Germany        DEU                        0.947
##  6 Sweden         SWE                        0.945
##  7 Australia      AUS                        0.944
##  8 Netherlands    NLD                        0.944
##  9 Denmark        DNK                        0.94
## 10 Finland        FIN                        0.938
## 11 Singapore      SGP                        0.938
## 12 United Kingdom GBR                        0.932
## 13 Belgium        BEL                        0.931
## 14 New Zealand    NZL                        0.931
## 15 Canada         CAN                        0.929
## 16 United States  USA                        0.926
## 17 Austria        AUT                        0.922
## 18 Israel         ISR                        0.919
## 19 Japan          JPN                        0.919
```

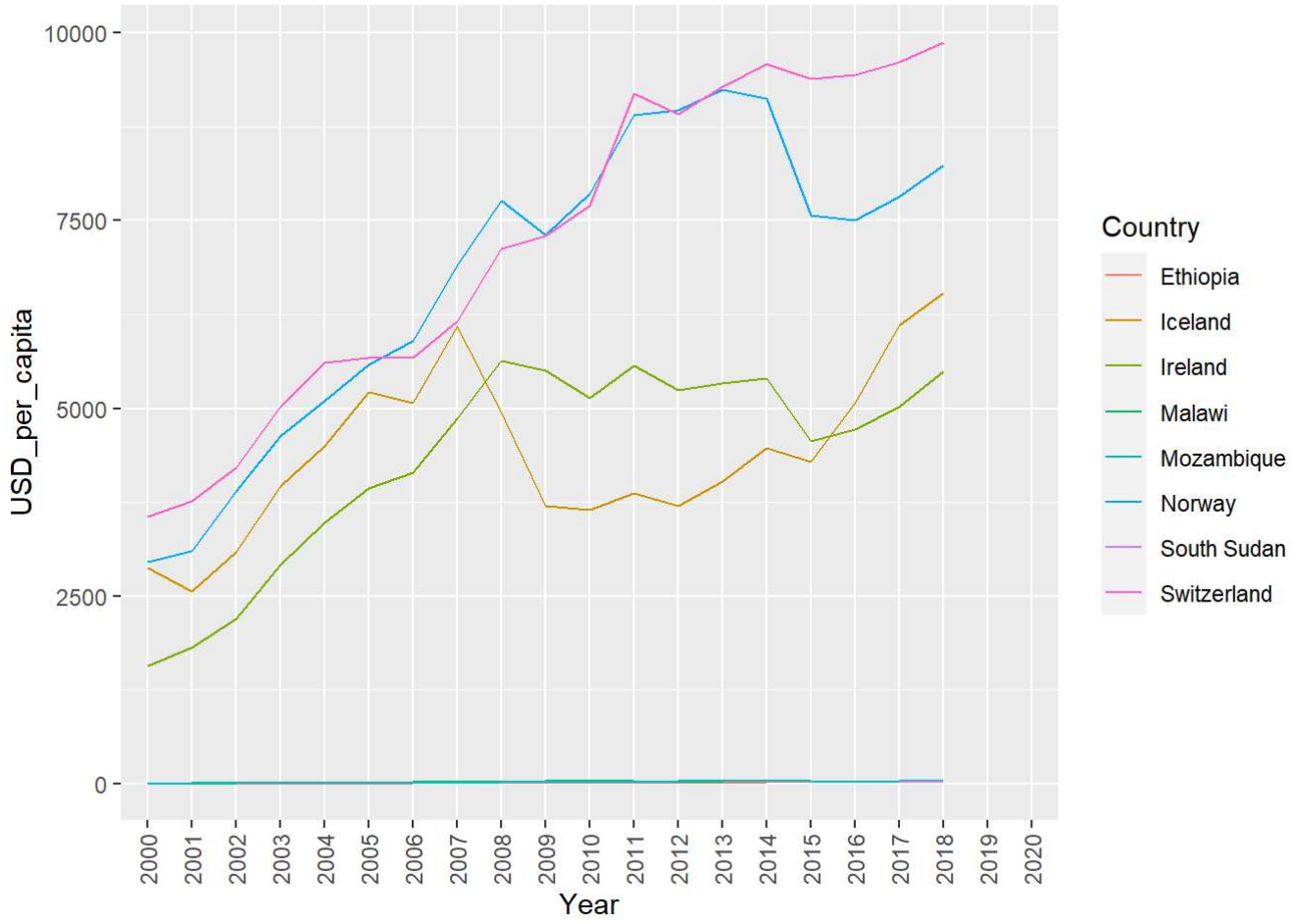
```
## Joining, by = c("Country", "iso_code")
```

```
## # A tibble: 4 x 3
##   Country     iso_code human_development_index
##   <chr>       <chr>                      <dbl>
## 1 South Sudan SSD                        0.433
## 2 Mozambique  MOZ                        0.456
## 3 Malawi      MWI                        0.483
## 4 Ethiopia    ETH                        0.485
```

```
## Joining, by = c("Country", "iso_code")
## Joining, by = c("Country", "iso_code")
```

To explore how COVID is impacting a countries' development, we first need to get a better understanding of how these countries have developed thus far. We are specifically going to look at health expenditure first, since COVID-19 is a global health crisis.

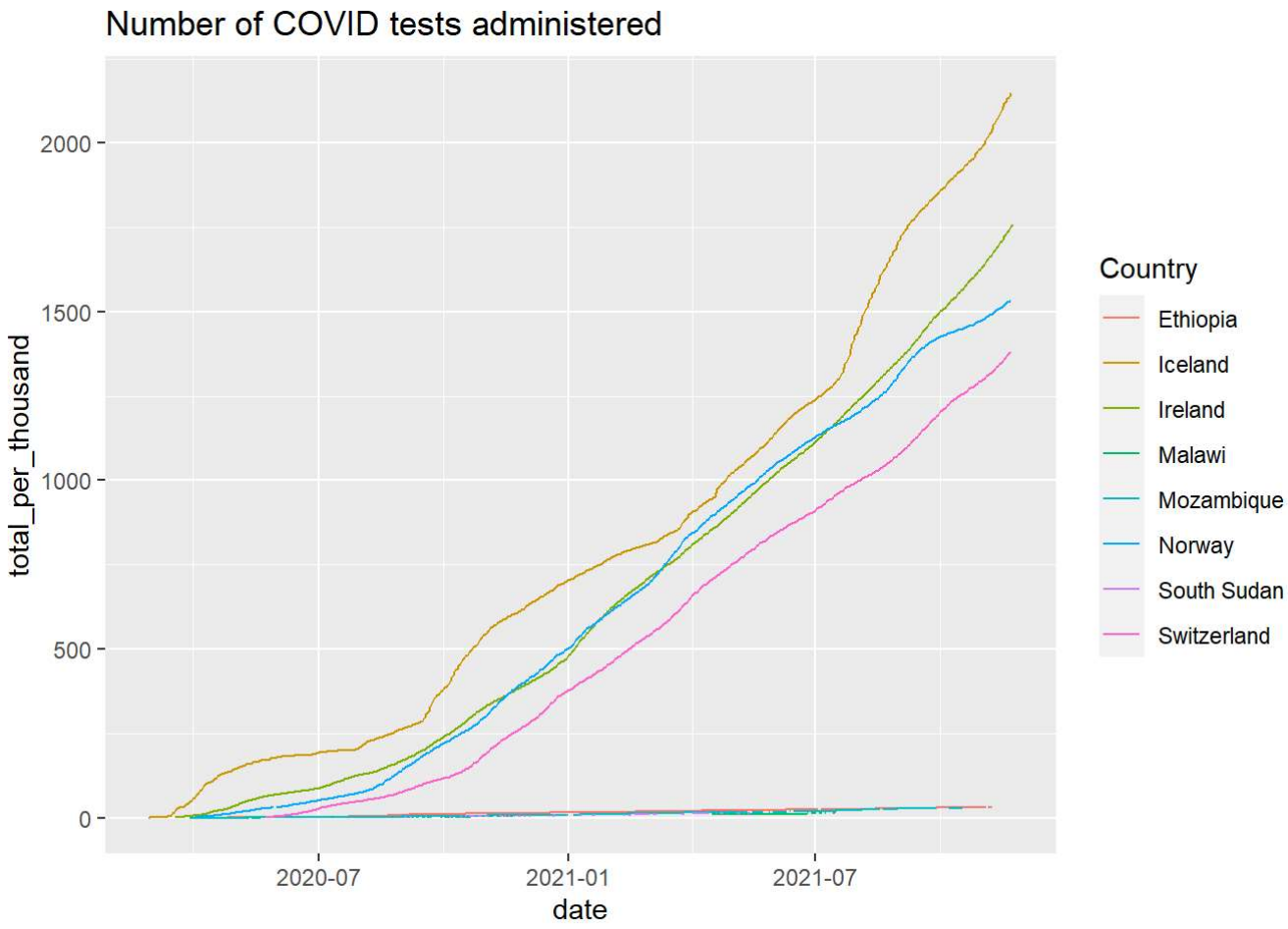
```
## Joining, by = c("Country", "iso_code")
```

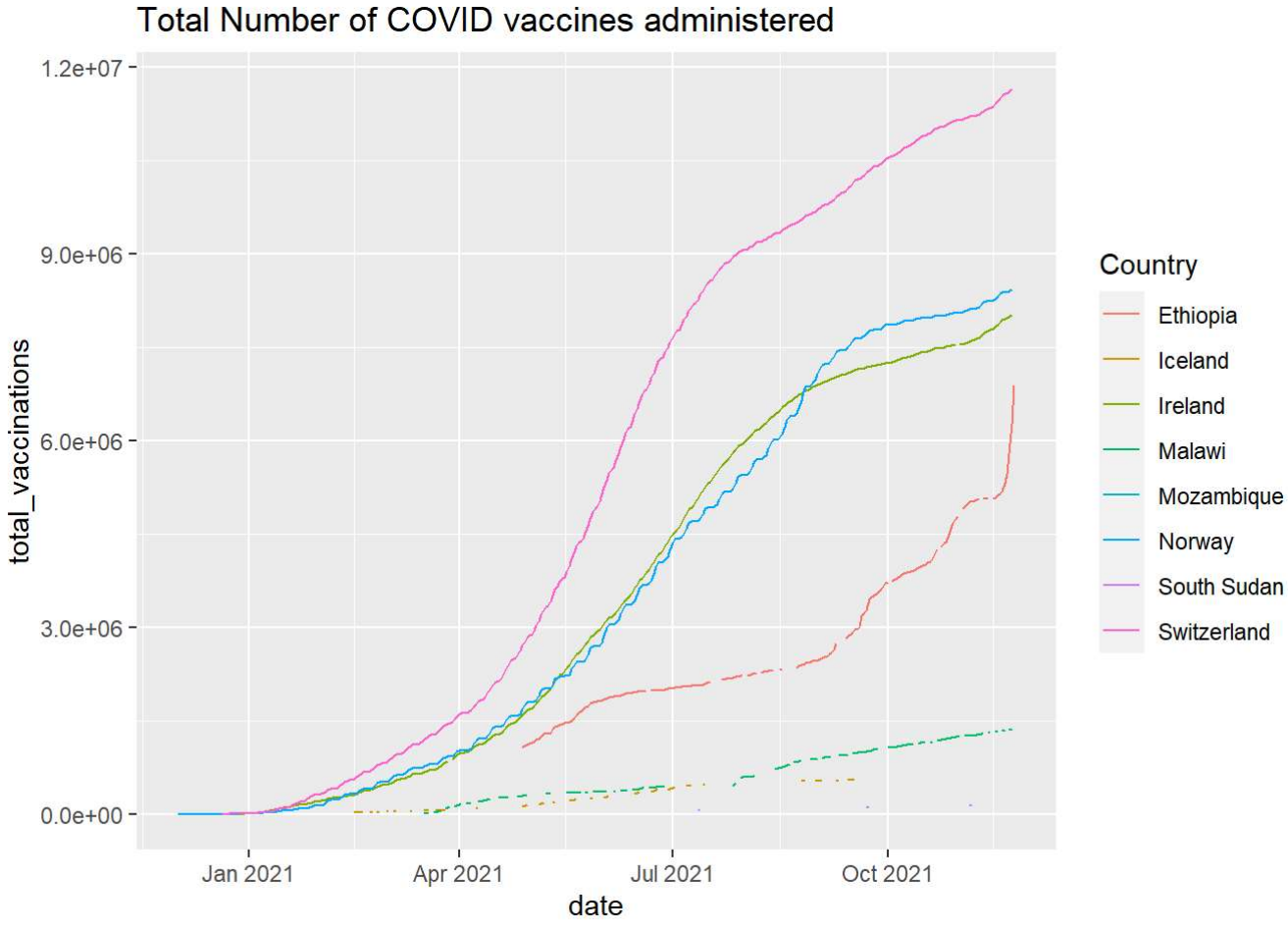


As you can see from the graph, only 4 lines are truly visible. The 4 that are not visible are the 4 countries with the lowest HDI; their health expenditure is marginal when compared to the most developed countries, even in 2020.

As stated earlier, countries with better overall health *should* see better rates for COVID testing as well as vaccinations. Let's explore that.

```
## Joining, by = c("Country", "iso_code")
```


Number of COVID tests administered
total_per_thousand
2000
1500
1000
500
0
2020-07
2021-01
2021-07
date
Country
Ethiopia
Iceland
Ireland
Malawi
Mozambique
Norway
South Sudan
Switzerland


Total Number of COVID vaccines administered
total_vaccinations
1.2e+07
9.0e+06
6.0e+06
3.0e+06
0.0e+00
Jan 2021
Apr 2021
Jul 2021
Oct 2021
date
Country
Ethiopia
Iceland
Ireland
Malawi
Mozambique
Norway
South Sudan
Switzerland

At a glance, we can see there is an upward trend all around; each of health expenditure, total COVID tests administered, and total vaccinations are all increasing over time.

Note: For most of the remaining analysis, linear regression models, and more specifically, the predictions that come from them will be used when plotting further data. Additionally, the red dots on the plots signify the data on our chosen dates.

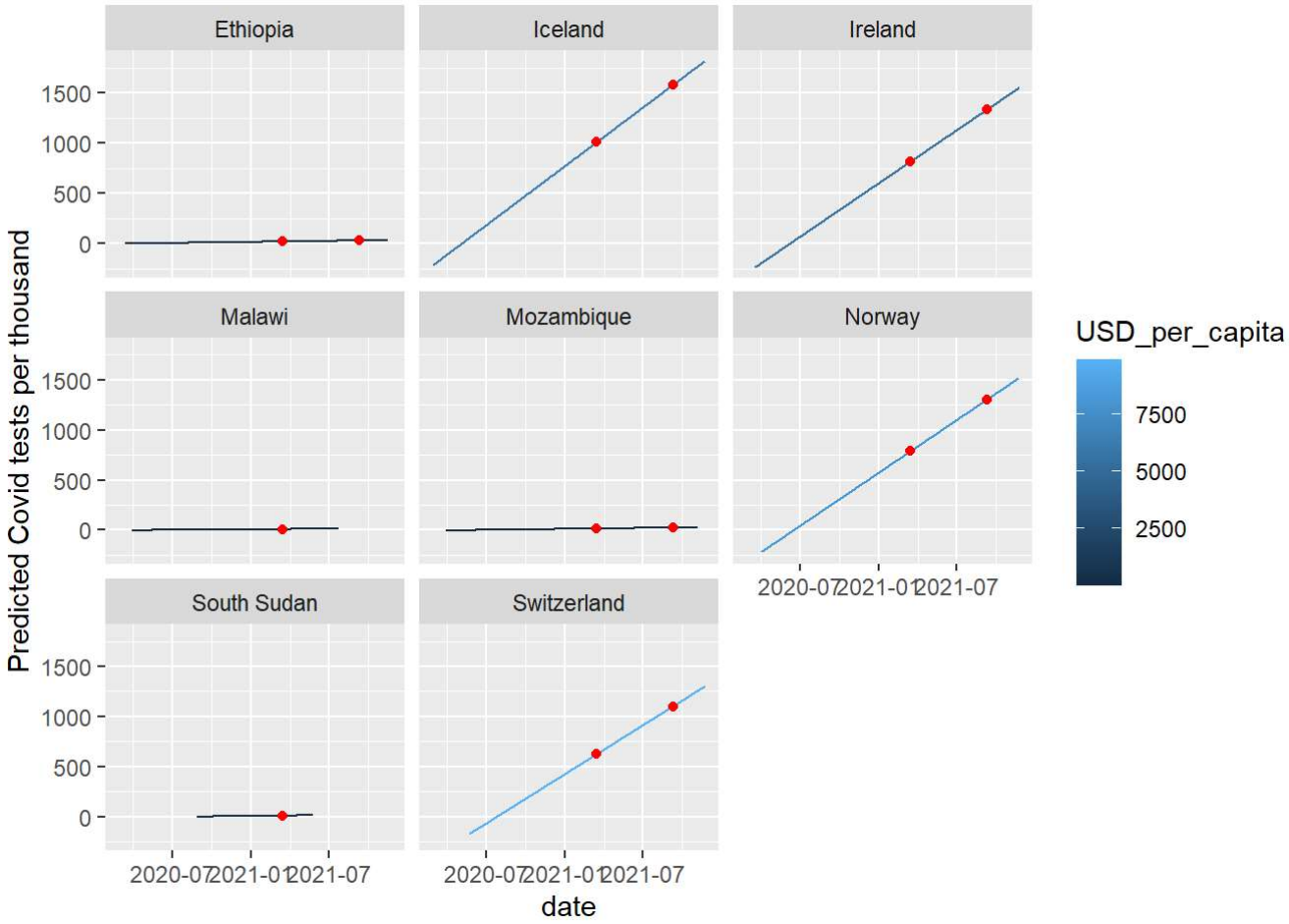
Is there a way in which we can see it all together?

Pictured below is a collection of graphs depicting the predicted number of COVID-tests per thousand for each of out 8 countries. Additionally, each line is coloured based on the countries' health expenditure, measured in USD per capita.

As we saw earlier, there is an overall positive trend with regards to the number of tests administered, and what stands out is the subset of countries with a significant positive trend: the 4 countries representing the highly developed countries and who spend the most on health.

```
## Adding missing grouping variables: `iso_code`
```

```
## Joining, by = c("Country", "iso_code")
## Joining, by = c("Country", "iso_code")
```



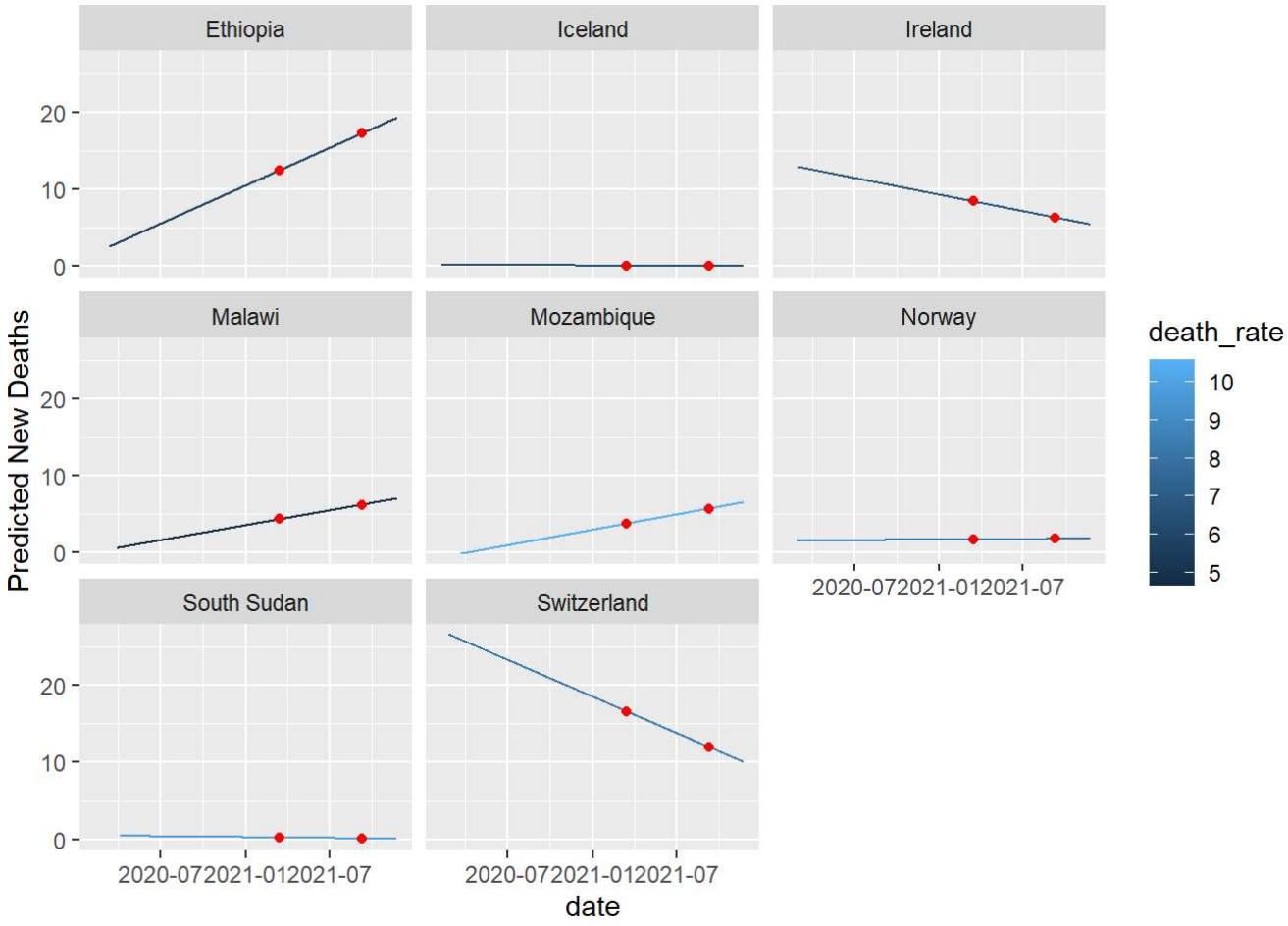
Moving onto the next representative of the health fact, we're going to take a look at the death rate of countries. In a similar fashion to the one prior, below is a collection of plots, one for each country, which plots the predicted new deaths from COVID-19.

This time around, there does not appear to be an overall trend; some countries are trending upwards, some downwards, and others remain relatively flat. Examine each country's death rate and South Sudan stands out as peculiar; its death rate is among the highest of the 8 and yet its daily COVID deaths show no significant trend. In fact, looking closely, it even appears to be trending downwards, albeit slightly. Compare this to Mozambique which shows a more expected result; its death rate is comparable to that of South Sudan's but its daily COVID deaths continue to rise. From this, it is clear that either there may not be a significant relationship between the two, or that the relationship is simply overshadowed by external factors.

```
## Adding missing grouping variables: `iso_code`, `continent`
```

```
## Joining, by = c("Country", "development_category")
```

```
## Joining, by = "Country"
```

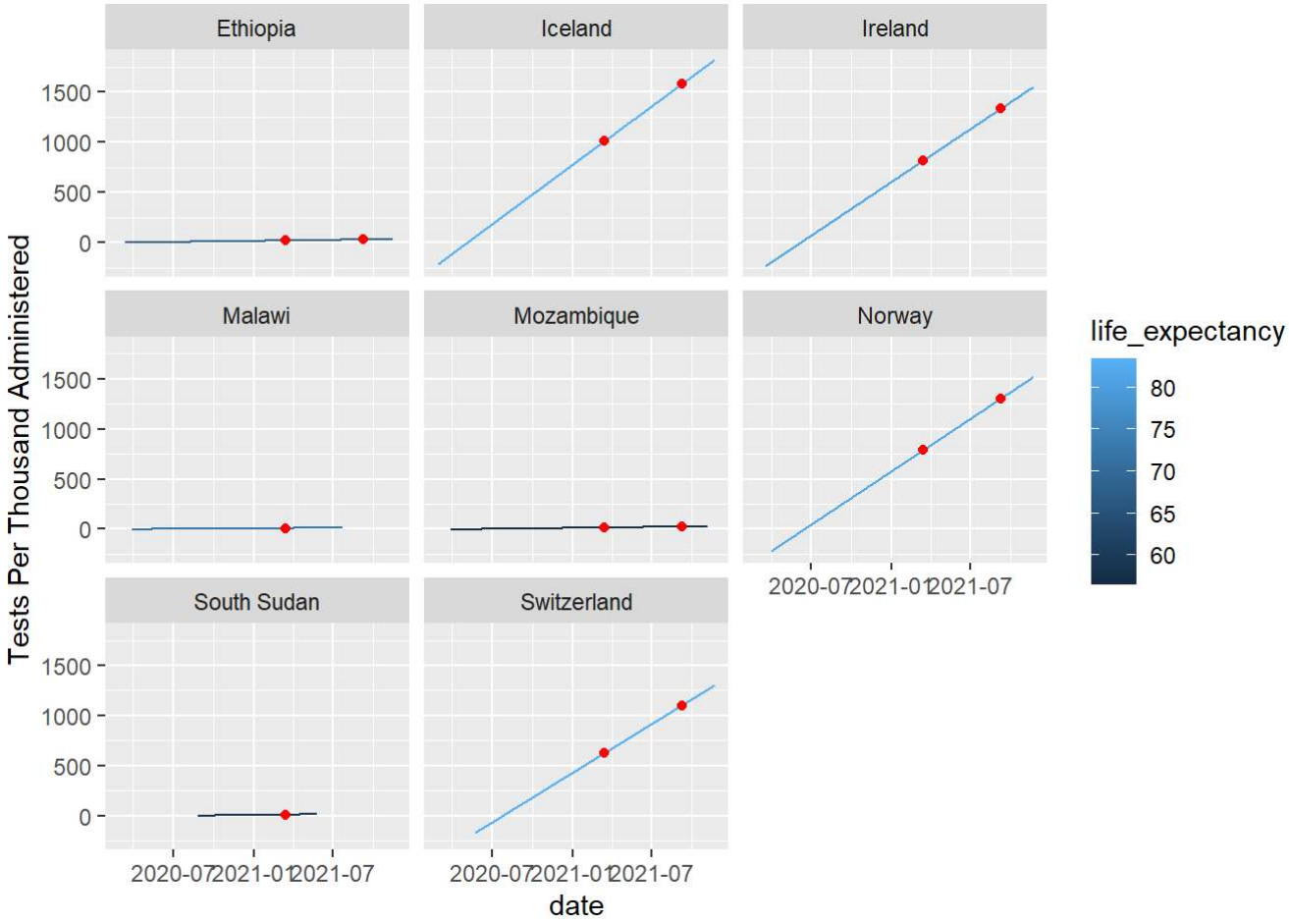


Next, we will look at a countries' life expectancy and its COVID testing per thousand.

There is nothing surprising here; countries whose life expectancy is higher are testing significantly more than those whose life expectancy is lower. More than likely, this stems from the fact that countries that live longer have the pre-existing infrastructure and systems that are capable. It may also be that in countries with a lower life expectancy, there are more inherent and systematic issues to deal with than in testing for COVID-19.

None of our 8 countries stand out in this aspect.

```
## Joining, by = c("Country", "iso_code")
## Joining, by = c("Country", "iso_code")
```



The following plots look at a country's happiness and its' governments' response to COVID-19. This is the only data that is not modelled as it features discrete values.

Before analyzing this, we must first learn what the values mean. According to the codebook associated with the data, values exist on a scale of 0 to 3 representing the status of "stay-at-home" measures, with each level meaning the following:

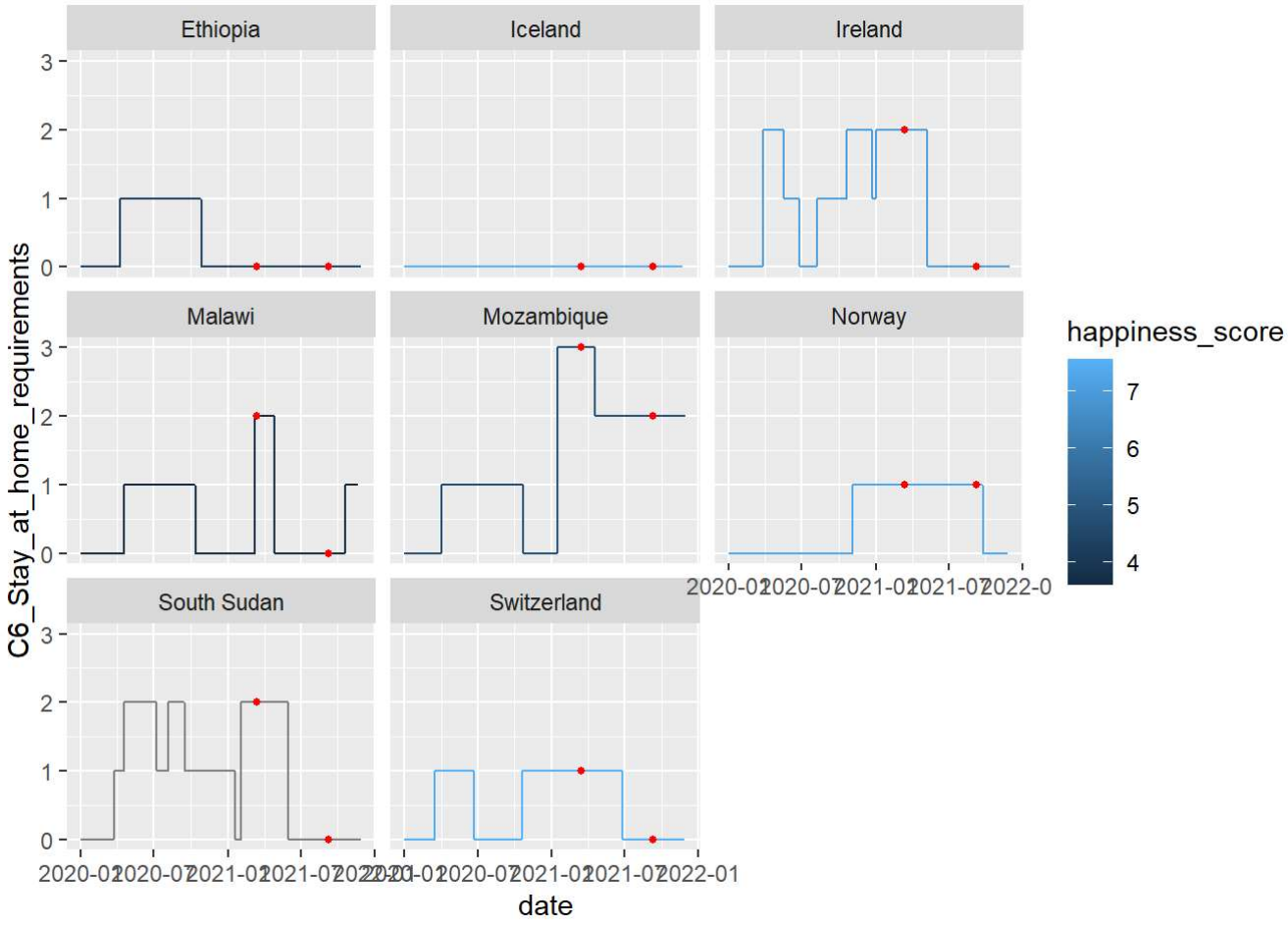
- 0: no measures in place
- 1: Recommended to not leave the house
- 2: Leaving the house for non-essential trips is prohibited
- 3: Leaving the house is prohibited with few exceptions

```
## Joining, by = c("Country", "iso_code")
```

```
## Joining, by = c("Country", "development_category")
```

```
## Joining, by = c("Country", "iso_code")
```

```
## Joining, by = "Country"
```



For the most part, there appears to be no connection between happiness and the governments' response. The happier countries vary greatly in their response to the pandemic; where Ireland saw multiple adjustments over the roughly 2 years, Iceland remained as normal with no measures in place.

Looking specifically at our chosen dates, as surprising as it may be, Ethiopia is one of two countries who had moved away from lock down by March 15, 2021, the other being Iceland. Despite having a happiness score on the lower end of the spectrum, it is arguably second in terms of how little it strayed from the norm that is regular life.

This furthers the notion that there is no relationship here.

If anything, it would be more interesting to see if geographical location is more of a factor in this scenario.

Next, we are going to take a look at urbanization and predicted COVID-infections.

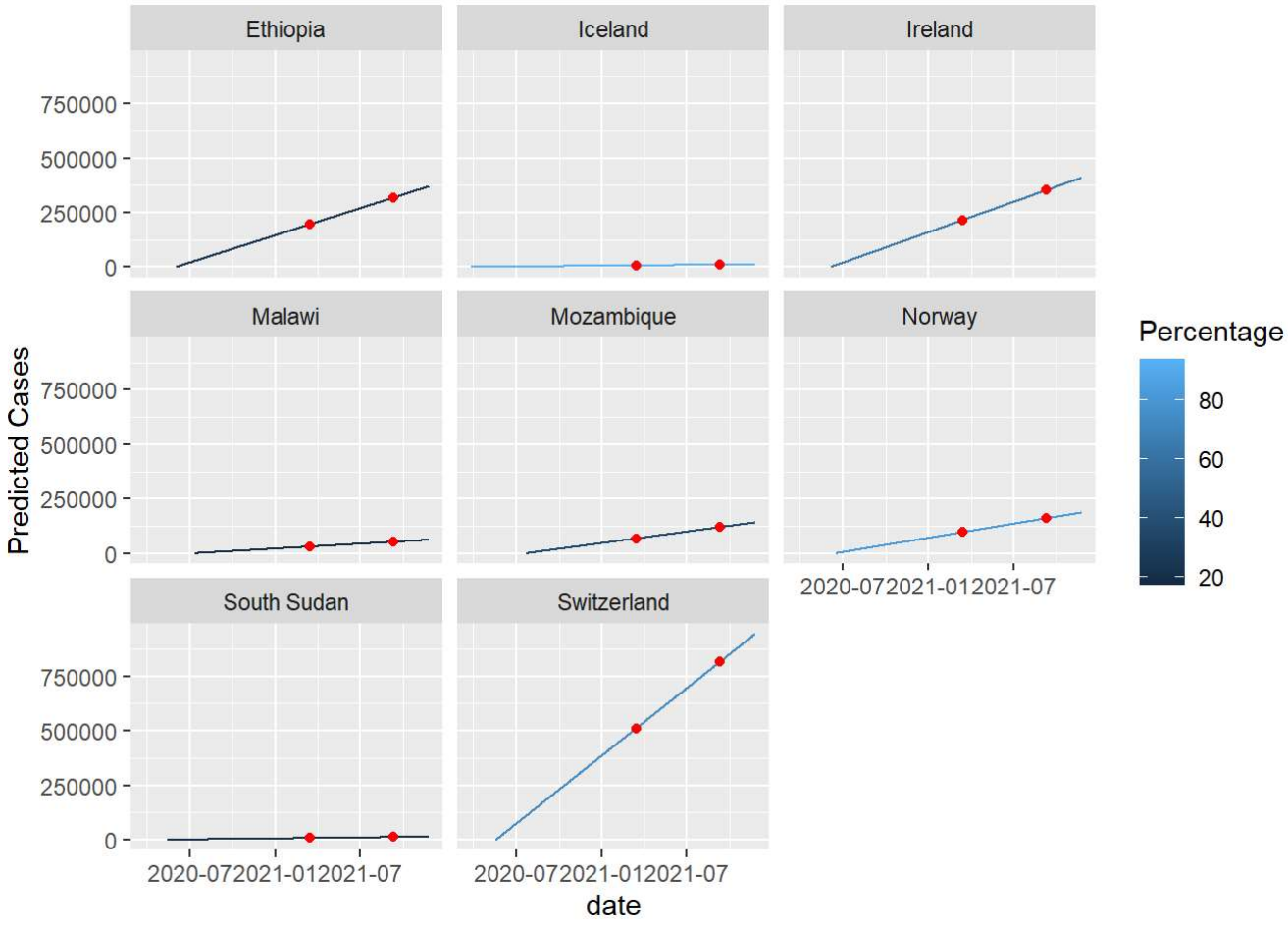
```
## Adding missing grouping variables: `iso_code`, `continent`
```

```
## Joining, by = c("Country", "iso_code")
```

```
## Adding missing grouping variables: `Year`
```

```
## Joining, by = "Country"
```

```
## Adding missing grouping variables: `Year`
```



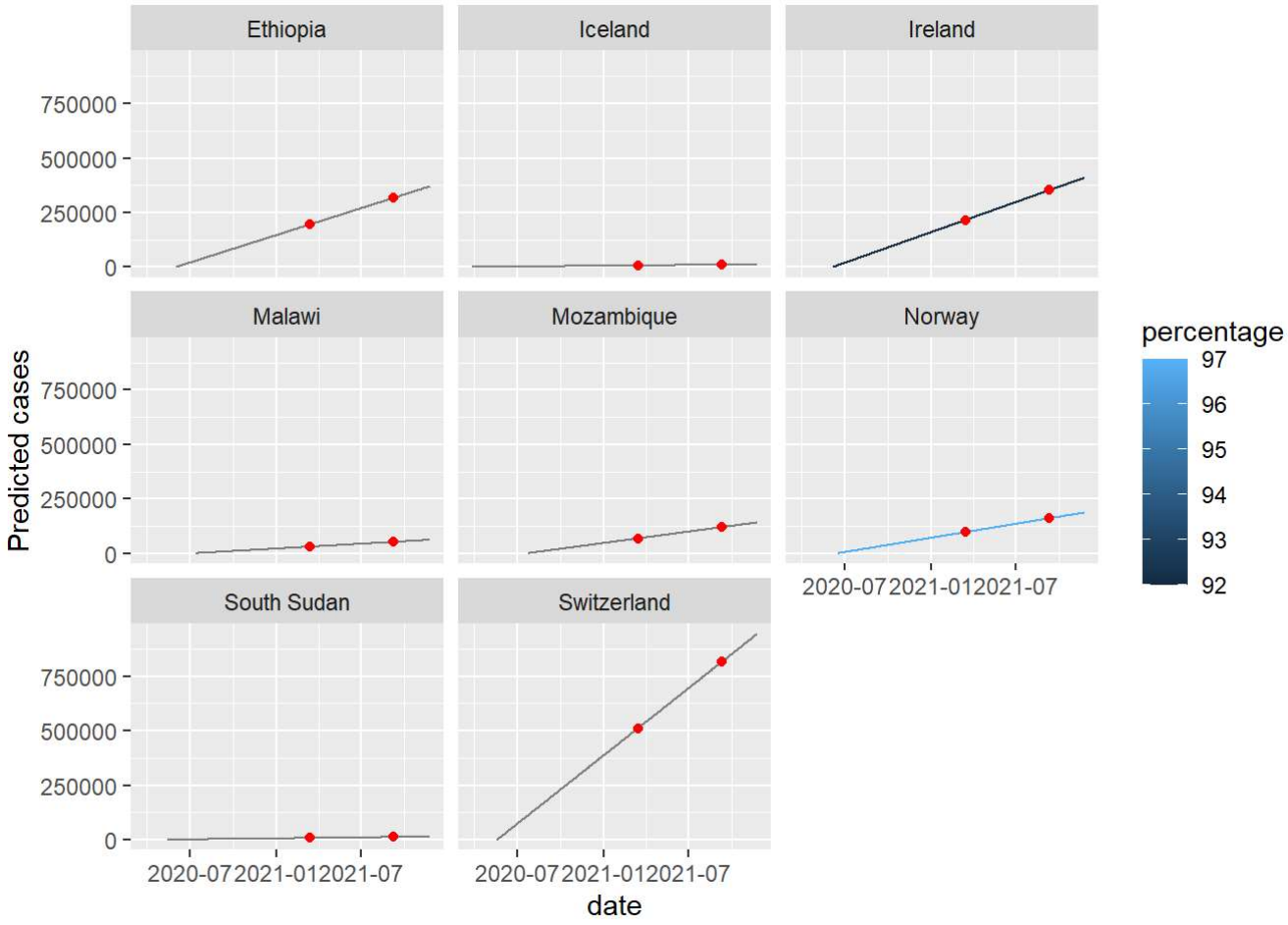
Looking at the percentage of urbanization, the two at the top appear to be Iceland and Norway. These two fair quite well compared to the other countries in terms of COVID cases. It must be noted however, that his may be skewed by the fact that these are modeled from the given data, and some of the countries may lack the data to create an accurate model.

Here, Switzerland stands out as the country with a high number of cases despite its status as one of the most developed countries. Perhaps then, urbanization, and thus infrastructure is actually a significant factor towards COVID infections; while two of the most developed countries in the world, Switzerland and Ireland are less urbanized than Iceland and Norway while also having more cases.

Keeping on the topic of infrastructure, lets investigate COVID cases in a country as compared to the amount of its internet users.

Note: South Sudan has no data for its internet usage.

```
## Joining, by = c("Country", "iso_code")
## Joining, by = c("Country", "iso_code")
```



There is nothing clear about the results here. Norway is clearly the leader in terms of the percentage of the population with access to the internet, but Iceland appears to be doing better in terms of cases, and Switzerland is doing worse. This may not be as significant a factor when considering the fact that with the way this plot has been set up, countries with lower populations will appear to do better in terms of cases, simply because there are less people to be infected.

Finally, let's take a look at our last factor: education. From the provided, data, education expenditure is the only education-related data set, which is something that must be kept in mind.

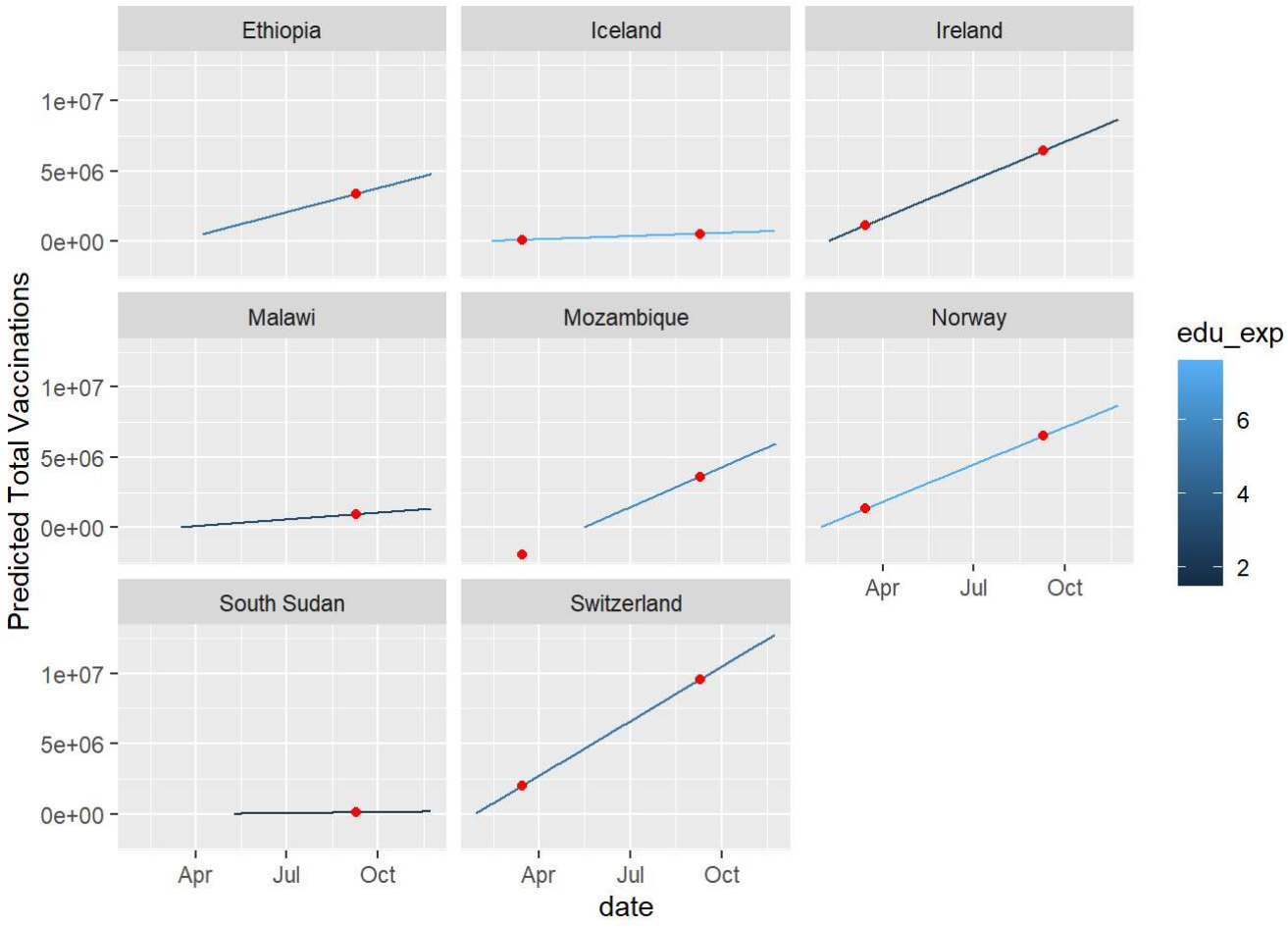
Note: Education expenditure is given in the most recent year, for example, South Sudan's is the oldest data, dating back to 2016, the rest falling somewhere in the middle.

```
## Adding missing grouping variables: `iso_code`
```

```
## Joining, by = "Country"
## Joining, by = "Country"
```

```
## # A tibble: 2,049 x 12
##    Country edu_exp_year edu_exp_rank edu_exp iso_code vaccination_data
##    <chr>          <dbl>        <dbl>   <dbl> <chr>    <list>
##  1 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  2 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  3 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  4 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  5 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  6 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  7 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  8 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
##  9 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
## 10 Iceland         2018           13     7.6 ISL      <tibble [329 x 4]>
## # ... with 2,039 more rows, and 6 more variables: total_model <list>,
## #   date <date>, total_vaccinations <dbl>, people_fully_vaccinated <dbl>,
## #   daily_people_vaccinated <dbl>, pred <dbl>
```

```
## Joining, by = "Country"
## Joining, by = "Country"
```



The thought process here was that a more highly educated population is more likely to get the vaccine due to a stronger belief in the science that supports it. The problem here is that education expenditure is not necessarily indicative of the education of the population, particularly when its measured as a percentage of the countries' GDP.

Switzerland and Ethiopia appear to have similar levels of education expenditure, but more than likely, Switzerland's GDP outclasses that of Ethiopia, lending to a significant difference in actual education expenditure.

As a result, there appears to be little in the way of any relationship between the two.

Conclusion

For the most part, the more developed countries saw better overall results with regards to COVID-statistics: cases and deaths were relatively lower while tests and vaccinations were higher than those of the less developed countries. There were some inconsistencies with the results compared to what was expected, and some of those were highlighted in the analysis. There are many possible unknown external factors that that may have affected the numbers beyond those examined in this report, In some cases, the choice of data also led to misleading results, from which the potential factor was known but no conclusion could be made regarding our own.

More specifically to the factors chosen for this project, it was assumed that health factors would contribute the most to the numbers of a global health crisis. As it turns out, that is not exactly the case. With the given data sets, while the health-related data proved to be more influential, it often implicated factors aside from health. Thus, from the exploration, it is difficult to establish a true correlation between health and the COVID-19 numbers. One major implication found from the health-data was an inferred presence of pre-established health infrastructure. As a result, it can be argued that infrastructure, and thus standard of living, is the aspect of the HDI which holds the most influence over the numbers from COVID-19.