In [1]:
```
# libraries
import numpy as np
import pandas as pd
import altair as alt
from sklearn.decomposition import PCA
```

PSTAT 100 Project plan report

Group information

Group members: Tyler Barton, Kayla Benitez, Patrick Chen, Victoria Christensen

Contributions:

1. Kayla Benitez studied the topic and prepared the background.
2. Tyler Barton prepared the data description and explored the data semantics and structure.
3. Patrick Chen studied the basic properties of the dataset and prepared an exploratory plot.
4. Victoria Christensen proposed a few methods to explore quesitons about the dataset and compiled our work into one file.
5. We all worked on tidying the dataset together and proposing interesting questions to explore.

0. Background

We will be analyzing COVID-19 data for various countries to see the global impact the pandemic has had. According to the World Health Organization, Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. Where most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Vaccines, testing and other variables aim to decrease the transmission of the COVID-19 virus and reduce cases and the severity of cases that lead to ICU & hospitalizations or even death. Our dataset comes from data maintained by Our World in Data (https://github.com/owid/covid-19-data/blob/master/public/data/README.md#confirmed-cases) (OWID) that has been updated daily throughout the duration of the COVID-19 pandemic. The dataset includes multiple countries and has various variables mentioned below.

The metrics of our dataset includes:

- Vaccinations
- Tests & Positivity
- Hospital & ICU
- Confirmed Cases
- Confirmed Deaths
- Reproduction Rate
- Policy Responses
- Other Variables of Interest

To explore the data, we can keep track of:

- Daily number of confirmed cases
- Cumulative number of confirmed cases
- Weekly/Bi-weekly cases
- Global Comparison

The COVID-19 pandemic is still ongoing and with this data, we wish to explore the impact of the pandemic and how countries have been able to mitigate cases. We will analyze for any trends, or relationships, such as whether there are positive or negative correlations between variables. We wish to observe the effect that tests, vaccinations, policy responses, reproduction rate and other variables of interest have on hospital & ICU, confirmed cases, and confirmed deaths.

1. Data description

Basic information

The citation is: Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus (https://ourworldindata.org/coronavirus)' [Online Resource].

All 207 countries sampled received updated data every day. The data is received with the help of thousands of researchers and the information of covid tests worldwide. The covid vaccination/cases data is collected with the help of health clinics/organizations. The actual number of people infected is higher than the number of cases displayed due to the limit of tests.

The relevant population is the whole world as it is a sample of 207 countries as they wanted to get a wide range of data to represent everyone. The sampling mechanism is through the use of the SDG-Tracker which is paired with Our World in Data and thousands of health organizations to gather the data. Due to the large amount of data and it being spread across 207 countries the scope of inference is the whole world as it involves a range of different countries.

Data semantics and structure

In [2]:
```
dataTable = pd.DataFrame(
    data = { 'Name':  ['date','continent', 'location','total_cases','total_deaths', 'human_development_index'
,'people_fully_vaccinated','total_vaccinations_per_hundred','total_vaccinations','median_age','gdp_per_capit
a','hdi_fac'],
             'Variable description':  ['date of observation','continent of the geographical location', 'geogr
aphical location','total number of cases','total number of deaths', 'index of average achievement of human de
velopment','number of people fully vaccinated','total vaccination doses per 100 people','total vaccination do
ses','mediam age of those surveyed','gross domestic product divided by population','Human development index l
evel'],
             'Type':  ['String','String', 'String','Numeric','Numeric', 'Numeric','Numeric','Numeric','Numeri
c','Numeric','Numeric','String'],
             'Units of measurement':  ['Calendar Year','Continent', 'Country','Positive covid tests','Mortali
ty number from covid', 'Average of health, education, and income','Fully vaccination count','Vaccination per
 100 people count','Vaccination count','Calendar year','Local current currency','Level of development']
            })
dataTable
```

Out[2]:

	Name	Variable description	Type	Units of measurement
0	date	date of observation	String	Calendar Year
1	continent	continent of the geographical location	String	Continent
2	location	geographical location	String	Country
3	total_cases	total number of cases	Numeric	Positive covid tests
4	total_deaths	total number of deaths	Numeric	Mortality number from covid
5	human_development_index	index of average achievement of human development	Numeric	Average of health, education, and income
6	people_fully_vaccinated	number of people fully vaccinated	Numeric	Fully vaccination count
7	total_vaccinations_per_hundred	total vaccination doses per 100 people	Numeric	Vaccination per 100 people count
8	total_vaccinations	total vaccination doses	Numeric	Vaccination count
9	median_age	mediam age of those surveyed	Numeric	Calendar year
10	gdp_per_capita	gross domestic product divided by population	Numeric	Local current currency
11	hdi_fac	Human development index level	String	Level of development

In [3]:
```
# load tidied data and print rows
covid = pd.read_csv('tidy-covid_data.csv')
covid.head()
```

Out[3]:

	date	continent	location	population	population_density	total_cases	total_deaths	new_deaths	aged_65_older	gdp_per_capita	...	total_v
0	2020-03-01	Asia	Afghanistan	39835428.0	54.422	5.0	NaN	NaN	2.581	1803.987	...	
1	2020-03-01	Asia	Macao	658391.0	20546.766	10.0	NaN	NaN	9.798	104861.851	...	
2	2020-03-01	Europe	Luxembourg	634814.0	231.447	1.0	NaN	NaN	14.312	94277.965	...	
3	2020-03-01	Europe	Lithuania	2689862.0	45.135	1.0	NaN	NaN	19.002	29524.265	...	
4	2020-03-01	Europe	Portugal	10167923.0	112.371	NaN	NaN	NaN	21.502	27936.896	...	

5 rows × 34 columns

2. Initial explorations

Basic properties of the dataset

Dimensions: We were thinking about comparing different countries' Covid-19 statistics among different variables. We choose the first day of each month for each country from March 2020 to February 2022. We have 4979 rows and 34 columns after we cleaned this dataset.

Missing values: There are missing values because many countries did not update the values every single day or some countries did not record those values at all.

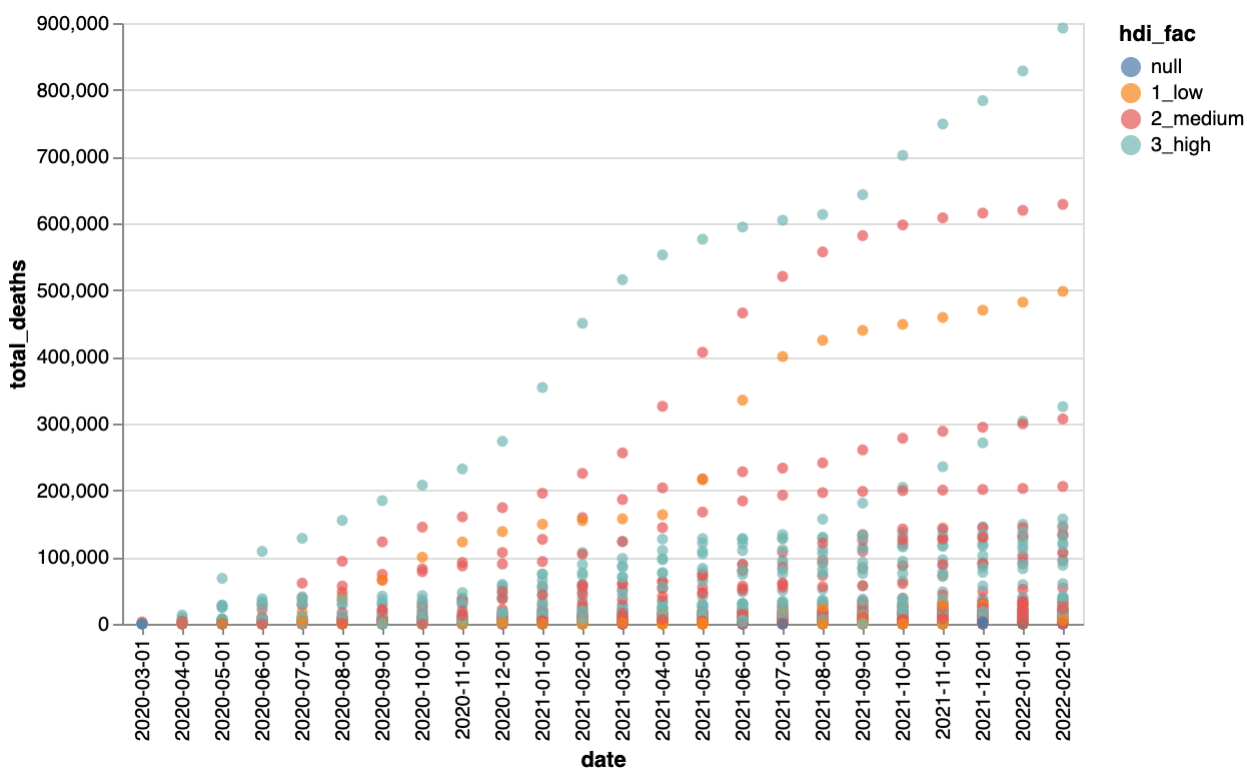
Variable summaries:

Name	Variable summary
date	Date of observation
continent	Continent of the geographical location
location	Geographical location
population	Population (latest available values). See https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv (https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_latest.csv) for full list of sources
population_density	Number of people divided by land area, measured in square kilometers, most recent year available
aged_65_older	Share of the population that is 65 years and older, most recent year available
gdp_per_capita	Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
extreme_poverty	Share of the population living in extreme poverty, most recent year available since 2010
cardiovasc_death_rate	Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)
diabetes_prevalence	Diabetes prevalence (% of population aged 20 to 79) in 2017
female_smokers	Share of women who smoke, most recent year available
male_smokers	Share of men who smoke, most recent year available
life_expectancy	Life expectancy at birth in 2019
human_development_index	A composite index measuring average achievement in three basic dimensions of human development—a long and healthy life, knowledge and a decent standard of living. Values for 2019, imported from http://hdr.undp.org/en/indicators/137506 (http://hdr.undp.org/en/indicators/137506)
total_tests	Total tests for COVID-19
new_tests	New tests for COVID-19 (only calculated for consecutive days)
total_tests_per_thousand	Total tests for COVID-19 per 1,000 people
new_tests_per_thousand	New tests for COVID-19 per 1,000 people
positive_rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average (this is the inverse of tests_per_case)
tests_per_case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average (this is the inverse of positive_rate)
tests_units	Units used by the location to report its testing data
total_vaccinations	Total number of COVID-19 vaccination doses administered
people_vaccinated	Total number of people who received at least one vaccine dose
people_fully_vaccinated	Total number of people who received all doses prescribed by the initial vaccination protocol
total_boosters	Total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol)
new_vaccinations	New COVID-19 vaccination doses administered (only calculated for consecutive days)
total_vaccinations_per_hundred	Total number of COVID-19 vaccination doses administered per 100 people in the total population
people_vaccinated_per_hundred	Total number of people who received at least one vaccine dose per 100 people in the total population
people_fully_vaccinated_per_hundred	Total number of people who received all doses prescribed by the initial vaccination protocol per 100 people in the total population
total_boosters_per_hundred	Total number of COVID-19 vaccination booster doses administered per 100 people in the total population
total_cases	Total confirmed cases of COVID-19.
total_deaths	Total deaths attributed to COVID-19
new_deaths	New deaths attributed to COVID-19

Exploratory analysis

In [4]:
```
scatter = alt.Chart(covid).mark_circle(opacity =0.7).encode(
    x = alt.X('date', scale = alt.Scale(zero=False)),
    y = alt.Y('total_deaths', scale = alt.Scale(zero=False)),
    color = 'hdi_fac'
)
scatter
```

Out[4]:



While we still need to address the null values and possibly remove them from the data set, we can see there are a few prominent lines. It might prove useful to identify which country these prominent lines come from and compare these specific countries.

---

3. Planned work

Questions

1. *How do the death rates among more developed countries compare to the death rates among less developed countries over time?*
2. *Is there an association between the total amount of cases relative to the different population densities around the world?*

Proposed approaches

Question 1 Potential Approaches:

1. *We could approach this question by doing something like we did in lab 6 where we create scatter plots colored by their level of HDI and then fit a simple linear model regressing death rates on time. Then we could layer our scatter plot with a line plot of the fitted values.*
2. *We can potentially isolate the countries that we have the most data for and see the comparisons we can make between those specific countries. For example, from looking at some of our exploratory plots, the United States seems to have fewer missing values than other countries. After finding several countries we could use some of the techniques we used in lab 4 and create multiple kernel density estimates to compare their distributions.*

Question 2 Potential Approach:

1. *We could use some of the techniques we used in lab 3 and create faceted scatterplots that are faceted by hdi_fac or continents and place total cases on the y-axis and population densities on the x-axis. We could then add loess curves like we did in lab 4 to better visualize the trends we see in the scatterplots.*

Submission Checklist

1. Save file to confirm all changes are on disk
2. Run *Kernel* > *Restart & Run All* to execute all code from top to bottom
3. Save file again to write any new output to disk
4. Select *File* > *Download as* > *HTML*.
5. Open in Google Chrome and print to PDF on A3 paper in portrait orientation.
6. Submit to Gradescope