**Project Status Update**

Our **objective** is to provide our client with a data-driven recommendation of the most optimal country to roll out their company’s COVID-19 vaccine. Below is our team’s approach:

* Using the **current country vaccination dataset** provided by the client, we will clean and query the data to:
  + Insert **day numbers** (day 1, day 2..) for each country from the first day they received their vaccinations
  + **Clean the daily vaccinations and total vaccination columns** to observe which countries have increased their **daily vaccination rates** the fastest. Later we can use this information to predict which of the new countries will have the fastest daily rates
  + Create a column for the **percentage of the population that has been vaccinated** so far in a given day. We can use this to observe the countries with the fastest rates that get vaccinated.
  + Some countries in the dataset, England, Northern Ireland, Scotland, and Wales are all part of the United Kingdom but we do not currently have any external data for these countries. There is already data for the United Kingdom, so we may **drop these countries (England, Northern Ireland, Scotland, and Wales)** from the dataset since the data for the UK combines vaccination data from all the individual UK countries
* Use a complete **COVID-19 dataset** from ourworldindata.org (<https://ourworldindata.org/covid-cases?country=IND~USA~GBR~CAN~DEU~FRA>). It contains data such as confirmed COVID-19 cases, vaccinations, deaths, testing, as well as demographics of each country such as population, amount of smokers, amount of people aged 65 or older, median age, gdp, and much more.
  + There are a lot of attributes in this dataset, so we will condense it and focus on the ones that seem most important. The data we will consider looking into are: **total cases and new cases of COVID-19 (both are available by date), total and new deaths from COVID-19, population, median age, gdp\_per\_capita**. If time permits, we could also look into other available fields in the dataset and add them into our dataset
  + **Population dataset** will be kept in a separate table. This table will contain population and population density for the year 2020 for each country. The original data comes from the United Nations World Population Prospects. (<https://github.com/owid/covid-19-data/blob/master/scripts/input/un/population_2020.csv>)
* Other possible data attributes suggested by the team include **urbanization, health care expenditure, and size of country** (data can be collected from worldbank.org)
* Compare each **country’s distance to the United States**. We calculate the distances in meters by using the coordinates (latitude, longitude) of each country to the coordinates of the United States. These information will be kept in a separate table.
* **Divide all countries into two separate tables**: one table for countries that have already have the COVID-19 vaccines already rolled out to, and another table for the newer countries that don’t have the vaccines yet (countries in this table will be a list of possible countries we will recommend to the client)