Milestone 2

Group Team Name: Group 27 AKA COVID KILLING ANALYZERS

Team Members: Natalie Mosqueda, Stephanie Younes, Eric Por

1. **Briefly restate your data analysis goals. Note that if you have substantially changed your project ideas since submitting milestone 1, please also include a short description of the reason for the change and discuss your new project ideas using the same format as required in milestone.**

Our goal for this data analysis is to analyze the efficiency of the COVID-19 vaccine in many different methods. One main method we will be using is to compare country's vaccination data to number of day-to-day cases data from March 2021 to October 2021 to analyze how the vaccine impacts covid cases. With this, we can see the difference the vaccination has made on countries with more access to vaccines compared to the ones with less access. Our second method, we will be comparing the vaccination count to the virus' mortality rate of these countries. Lastly, we plan on seeing the impact the vaccine has on the hospitalization rate of countries.

2. **Describe the data you have**

We got our data from the Our World in Data official website (https://ourworldindata.org/coronavirus-source-data) The general format the data we found looked as such with a couple more columns:

new_vaccina	stringency_ir	population	population_c	median_age	aged_65_old	aged_70_old	gdp_per_cap	extreme_pov	cardiovasc_d	diabetes_pre	female_smo	male_smoke	handwashing	hospital_bed	life_expectan	human_deve
34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
40	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
45	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
50	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
56	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
61	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
66	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
72	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
72	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
72	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511
[illegible]	30.56	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59			37.746	0.5	64.83	0.511

new_vaccina	total_vaccina	people_vacci	people_fully	total_booste	new_vaccina	stringency_ir	population	population_c	median_age	aged_65_old	aged_70_old	gdp_per_cap	extreme_pov	cardiovasc_d	diabetes_pre
1367					34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
1367					34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
1367					34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
1367					34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
1367	0.02	0.02			34	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
1580					40	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
1794					45	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
2008					50	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
2221					56	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
2435					61	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59
2649					66	12.04	39835428	54.422	18.6	2.581	1.337	1803.987		597.029	9.59

OLD DATA

But, after pre-processing, we have taken out the data that is not necessary and the data now includes, **iso_code, continent, location, data, total_cases, new_cases, total_deaths, new deaths, new_vaccinations_smoothed, population, .** Here is a snippet of our data now:

iso_code	continent	location	date	total_cases	new_cases	total_deaths	new_deaths	total_cases_	new_cases_	total_deaths	new_deaths_
AFG	Asia	Afghanistan	2/24/21	55664	18	2436	1	1397.349	0.452	61.152	0.025
AFG	Asia	Afghanistan	2/25/21	55680	16	2438	2	1397.751	0.402	61.202	0.05
AFG	Asia	Afghanistan	2/26/21	55696	16	2442	4	1398.152	0.402	61.302	0.1
AFG	Asia	Afghanistan	2/27/21	55707	11	2443	1	1398.429	0.276	61.327	0.025
AFG	Asia	Afghanistan	2/28/21	55714	7	2443	0	1398.604	0.176	61.327	0
AFG	Asia	Afghanistan	3/1/21	55733	19	2444	1	1399.081	0.477	61.352	0.025
AFG	Asia	Afghanistan	3/2/21	55759	26	2446	2	1399.734	0.653	61.403	0.05
AFG	Asia	Afghanistan	3/3/21	55770	11	2446	0	1400.01	0.276	61.403	0
AFG	Asia	Afghanistan	3/4/21	55775	5	2446	0	1400.136	0.126	61.403	0
AFG	Asia	Afghanistan	3/5/21	55827	52	2449	3	1401.441	1.305	61.478	0.075
AFG	Asia	Afghanistan	3/6/21	55840	13	2449	0	1401.767	0.326	61.478	0
AFG	Asia	Afghanistan	3/7/21	55847	7	2449	0	1401.943	0.176	61.478	0
AFG	Asia	Afghanistan	3/8/21	55876	29	2451	2	1402.671	0.728	61.528	0.05
AFG	Asia	Afghanistan	3/9/21	55876	0	2451	0	1402.671	0	61.528	0
AFG	Asia	Afghanistan	3/10/21	55894	18	2451	0	1403.123	0.452	61.528	0
AFG	Asia	Afghanistan	3/11/21	55917	23	2451	0	1403.7	0.577	61.528	0

NEW DATA

icu_patients	icu_patients_	hosp_patient	hosp_patient	weekly_icu_a	weekly_icu_a	weekly_hosp	weekly_hosp	new_vaccina	population	median_age	aged_65_old	aged_70_old	cardiovasc_death_rate
								1367	39835428	18.6	2.581	1.337	597.029
								1367	39835428	18.6	2.581	1.337	597.029
								1367	39835428	18.6	2.581	1.337	597.029
								1367	39835428	18.6	2.581	1.337	597.029
								1367	39835428	18.6	2.581	1.337	597.029
								1580	39835428	18.6	2.581	1.337	597.029
								1794	39835428	18.6	2.581	1.337	597.029
								2008	39835428	18.6	2.581	1.337	597.029
								2221	39835428	18.6	2.581	1.337	597.029
								2435	39835428	18.6	2.581	1.337	597.029
								2649	39835428	18.6	2.581	1.337	597.029
								2862	39835428	18.6	2.581	1.337	597.029

We will of course save a copy of the original data in the case we need to refer to deleted columns. The size of the data is about 55,815 rows and at the moment we have all the data we need but we are open to researching for more data if necessary.

3. **Show and discuss your preliminary analysis. For example, this could include:**

The data we plan on using is split into three different methods we want to use.

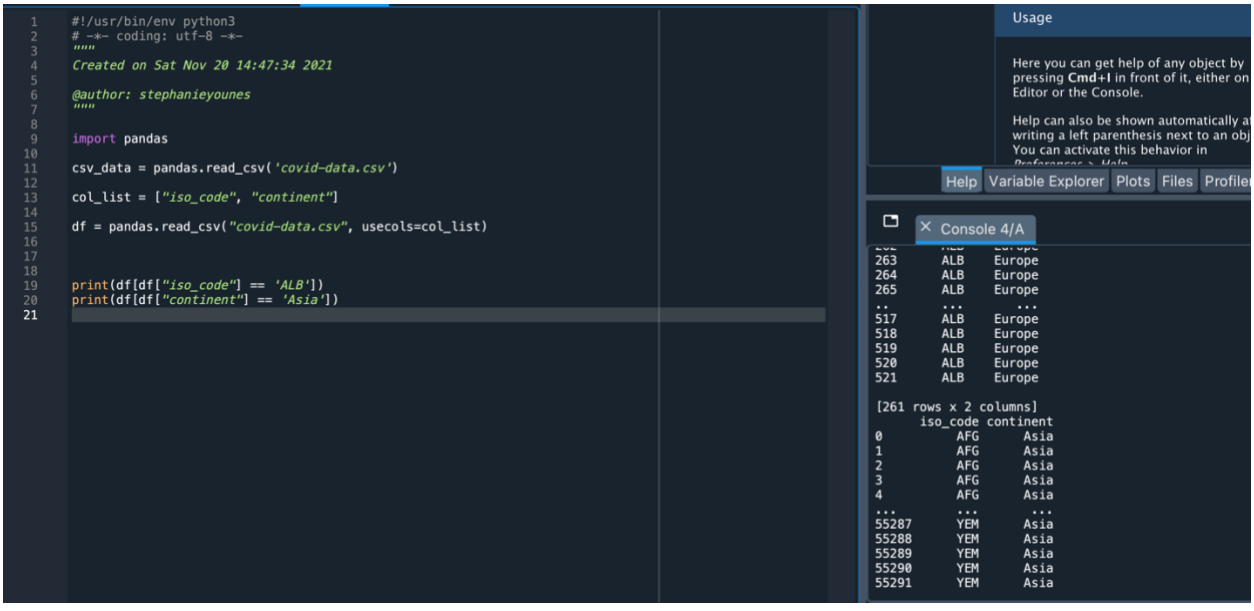
The first method being we want to compare country's vaccination data to number of day-to-day cases data from March 2021 to October 2021. For this method the data needed is location, date, new vaccinations, and total cases per million.

The second method being we want to compare vaccination count to the virus' mortality rate of the countries. For this method the data needed is also location, date, total vaccination, and total deaths per million.

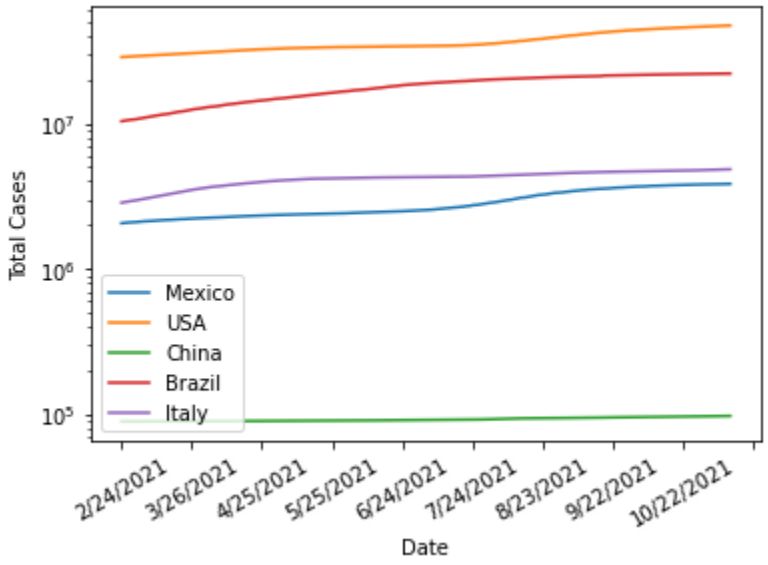
For our last method, we plan on seeing the impact the vaccine has on the hospitalization rate of countries. For this last method, we would need location, date, total vaccination, and icu patients.

To achieve our goal, we do not plan on transforming the data, but we plan on parsing the data by columns. With this, we will be able to read each column to test our methods out.

Below is an example of how we plan on parsing the columns and the result of doing so in order to manipulate the data:



We first sort out the desired data from the file by specifying the ISO_CODE from the datasets to obtain the total cases and dates for each country. Currently we're working on switching the whole process to have each country's data parsed without specifying the ISO_CODE. Below, is an example of the code we implemented alongside the graph




```
import numpy as np
import pandas as pd
from pandas import Grouper
import matplotlib.pyplot as plt
import matplotlib.dates as mdates

df = pd.read_csv('Covid Data final.csv', header=[0],parse_dates=True)



mex_cases = df.loc[df["iso_code"] == 'MEX',"total_cases"]
usa_cases = df.loc[df["iso_code"] == 'USA',"total_cases"]
chn_cases = df.loc[df["iso_code"] == 'CHN',"total_cases"]
bra_cases = df.loc[df["iso_code"] == 'BRA',"total_cases"]
ita_cases = df.loc[df["iso_code"] == 'ITA',"total_cases"]

mex_date = df.loc[df["iso_code"] == 'MEX', "date"]
usa_date = df.loc[df["iso_code"] == 'USA', "date"]
chn_date = df.loc[df["iso_code"] == 'CHN', "date"]
bra_date = df.loc[df["iso_code"] == 'BRA', "date"]
ita_date = df.loc[df["iso_code"] == 'ITA', "date"]

countries = ['Mexico','USA','China','Brazil','Italy']

fig = plt.figure()
plt.xlabel("Date")
plt.ylabel("Total Cases")

plt.plot(mex_date,mex_cases)
plt.plot(usa_date,usa_cases)
plt.plot(chn_date,chn_cases)
plt.plot(bra_date,bra_cases)
plt.plot(ita_date,ita_cases)

plt.semilogy()
plt.legend(countries)
plt.gca().xaxis.set_major_locator(mdates.DayLocator(interval=30))
plt.gca().xaxis.set_tick_params(rotation = 30)

plt.show()

```