# UCD Professional Academy – Certificate in Introductory Data Analytics –

# Gráinne Colreavy – October 2001 Intake – Group B

## Abstract

The URL for this projet is as follows:

<https://github.com/GCOLREAVY/UCDPA_CIDA_GCOLREAVY_CVD.git>

## Introduction

As an Employee of the Health Service Executive I have spent the last 22 months supporting our frontline colleagues in the delivery of health services to the population of this state under the most challenging of circumstances.

The Covid-19 pandemic has completely altered how as an organisation we plan, implement and deliver our services. Many of our support functions and admin support teams are now working remotely and the focus of our teams remains ensuring that our front line colleagues are facilitated and empowered to continue to provide their critical services to the population including Testing and Tracing, Vaccinations and Hospital and Community Services.

Gaining an understanding of some of the initial data in relation to Covid-19 has provided me with a sense of comfort around the performance of my organisation and indeed the state through this most difficult time.

## Dataset

I commenced the project by identifying the most appropriate ***real world datasets***. There were a number of iterations to this process as there are a significant number of similar datasets available on-line.

There were 3 areas of interest for me from a health service provider perspective. These were Total Covid Cases, Total Covid Deaths and Vaccination levels across the world. In order to examine the global perspective I selected to use a 2020 GDP per Capita dataset, which provides a list of countries of the world by gross domestic product (at purchasing power parity) per capita in order to broadly demonstrate a wealth ratio across the continents and countries of the world.

I proceeded to ***import the selected CSV files into 3 Pandas DataFrames***.

**Covid Data – Cases & Deaths – fluid dataset downloaded directly and updates automatically**

https://covid.ourworldindata.org/data/owid-covid-data.csv

**Covid Data – Vaccination Levels – fluid dataset downloaded directly and updates automatically**

https://covid19.who.int/who-data/vaccination-data.csv

**GDP per Capita – static dataset based on 2020 data – downloaded to C: drive and downloaded from there**

https://www.kaggle.com/daniboy370/world-data-by-country-2020?select=GDP+76per+capita.csv

## Implementation / Process

I selected the Jupyter Notebook application to develop code, visualizations and narrative text.

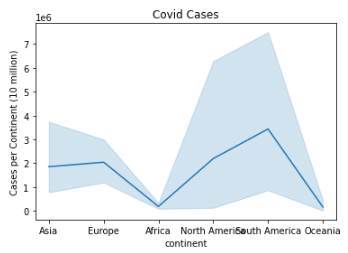
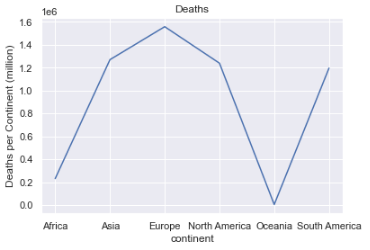
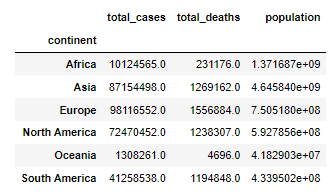
## Scipt, Functions, Visualisations and Insights

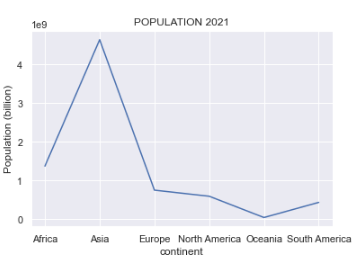
I initially designed a ***Custom Function*** to view the shape and headers of a dataframe or a table. This function allows me to quickly describe the shape, headers and datatypes relating to each DataFrame and subsidiary table.

There was significant data cleaning required across all datasets. This clean up included the following actions:

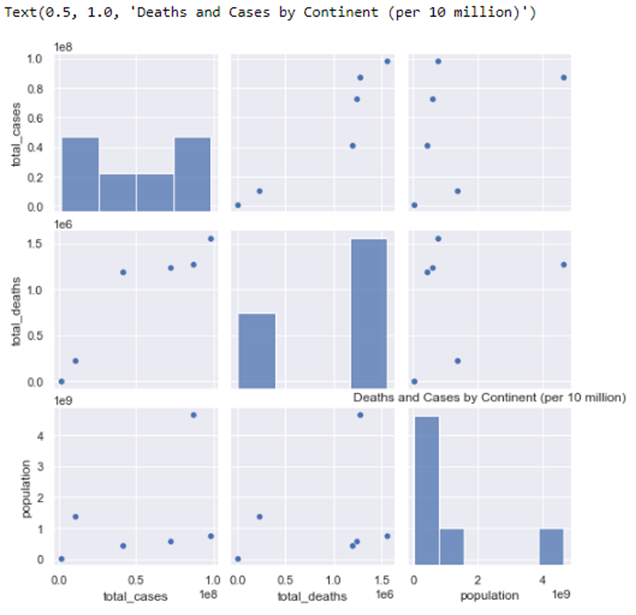
* **Drop duplicates**
* Drop NaN’s
* Drop outliers

Due to the large scale of each of these dataframes I used groupby functionality in order to examine the data using smaller and more manageable entities eg. by continent, by WHO\_Region and charted appropriately. I initially looked at the ‘total\_cases’, ‘total\_deaths’ and ‘population’ in the Cases & Deaths dataset and added some visualizations to have a quick look and sense check of the data. This quickly demonstrated that there is a correlation between deaths and cases but not population and cases or population and deaths. It appears that the statistics for the continent of Africa have provided some anomalies in the data. According to these figures less than 1 % of the population of Africa have been confirmed covid cases and only 0.02% of the population have died as a result of covid. The average across all other continents is 8.7% cases and 0.14% deaths.





When I had condensed the Covid Data Dataframe I also ran a pairplot to examine the data across all of the relevant data and the visualization below was the outcome. This pair plot demonstrates the anomalies across the continents in relation to population, covid cases and covid deaths.



When examining the GDP dataframe, I used the **loc** function and **index** function to check for and remove errors including NaN values from the GDP dataframe.

I included a line of script for each of the dataframes connected to the internet in order to notify when the data has been updated on-line.

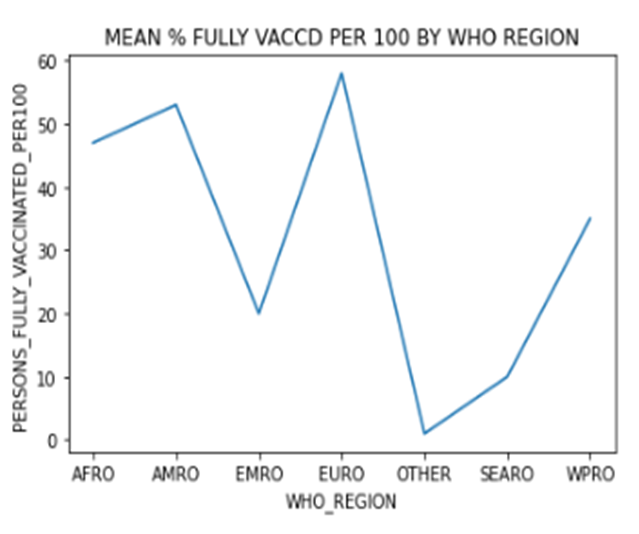
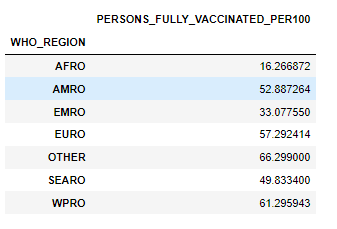
When examining the Vaccs dataframe a significant data cleaning was required and I again used a variety of functions to make the data more manageable.

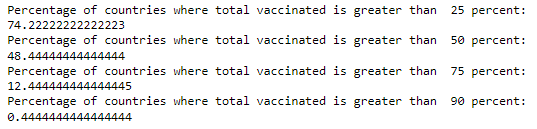
Throughout the script I have used numerous visualizations to describe the broad data of interest in each dataframe eg. ‘total\_deaths’, ‘total\_cases’,

I identified an outlier in the Vaccs table (PERSONS\_FULLY\_VACCINATED\_PER\_100) and deleted the related data using script.

I dropped all columns not required from each dataFrame and described the details pertaining to the reduced table using my custom function (viewTable).

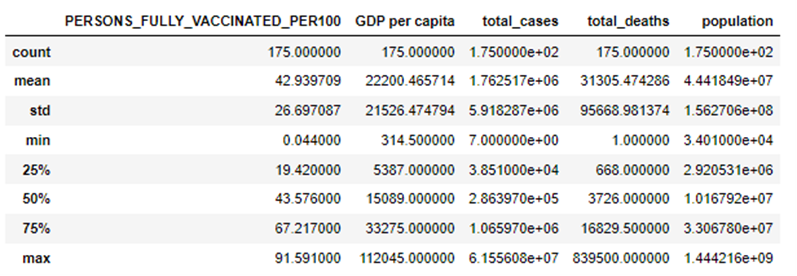
I used the groupby function to examine the % figures in relation to Persons\_FULLY\_VACCINATED\_PER\_100 in the Vaccs Table and obtained the mean by WHO\_REGION in relation to this data. I provided a visualization and summary table to describe this data.



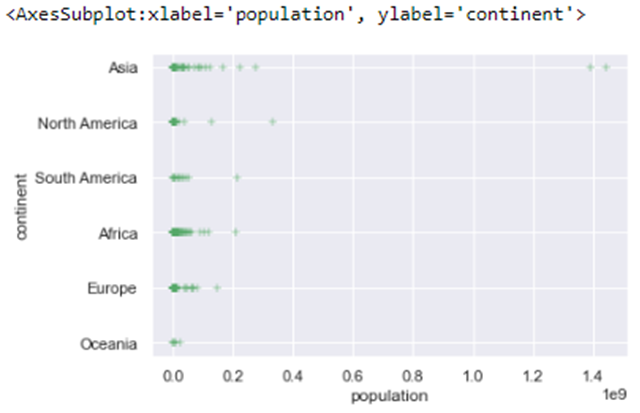
I developed a **looping** function to provide statistics on percentage of countries where total vaccinated is greater than 25%, 50%, 75%, 90% (see output):

I then proceeded to **merge** the 3 dataframes (GDP, Vaccs and Cases / Deaths) on the shared column in relation to the ISO code.

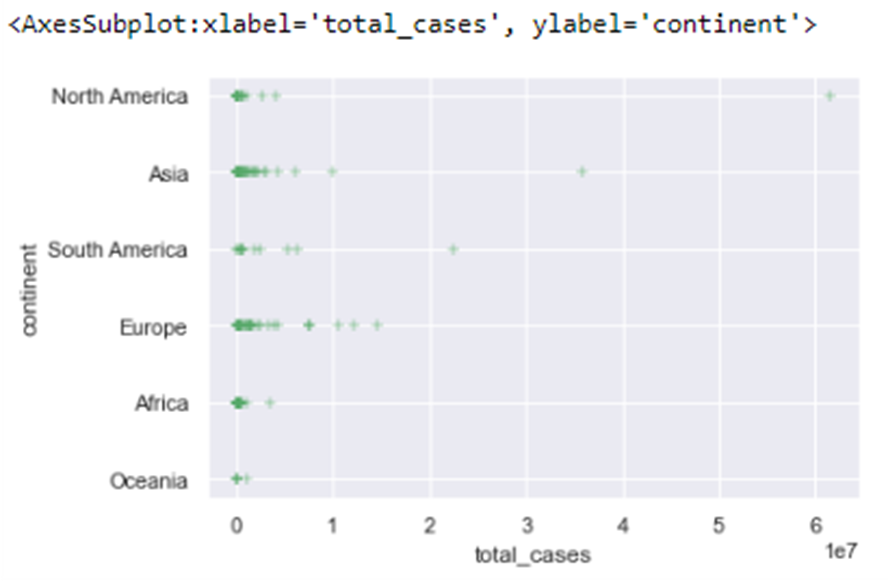
I provided a summary statistical table to describe all of the data in the merged dataframe (see table below).



I used a number of visualizations to attain insights in relation to this newly merged data comparing Population, Covid Deaths and Covid Cases across the continents.



This visualization indicates that there are significant variances in population levels in Asia. I validated this data against the header output when I used the sorting function on the merged dataframe. This indicated that China and India have the largest populations in the world according to these datasets. This visualization also illustrates that there is an accumulation of population levels in the countries with populations of less than 2 million across across all continents, with lower population numbers evident in Africa, Asia, North America and Europe.

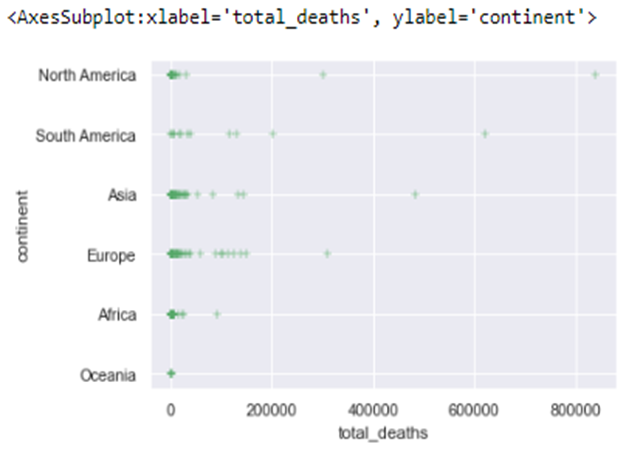


This visualization demonstrated that there was a significant variance in the numbers of covid cases across the continents. There are significantly larger numbers of deaths in:

* 1 country in North America
* 1 country in South America
* 1 country in Asia

It is also noted that there are a small number of countries in Europe who have significantly more covid cases than the majority of their European counterparts.

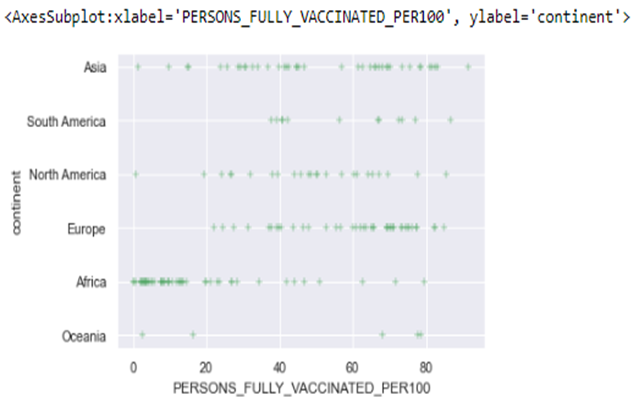
Again this visualization confirms that the level of covid case levels remained very low in Africa. We can discount Oceania as the population level is very small in this continent (demonstrated in the population visualization above).



This visualization demonstrated that there was a significant variance in the numbers of covid deaths across the continents. There are significantly larger numbers of deaths in:

* 2 of the countries in North America
* 1 country in South America
* 1 country in Asia
* 1 country in Europe

Again this visualization confirms that the level of covid death remained smallest in Africa. We can discount Oceania as the population level is very small in this continent (demonstrated in the population visualization above).



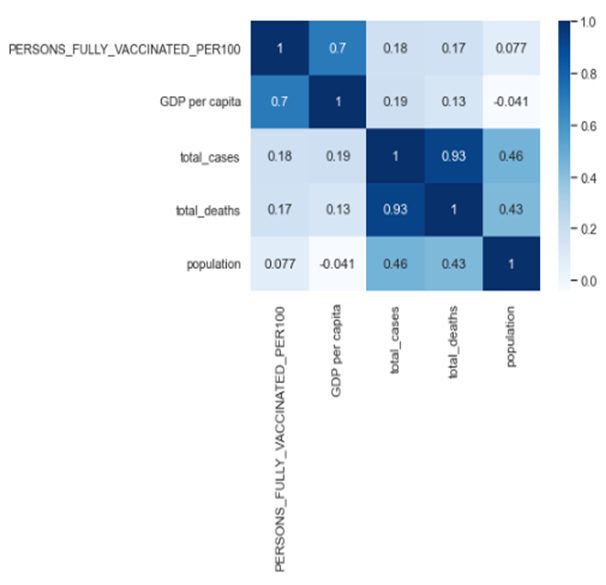
This visualization demonstrates that there is a broad spread of vaccination levels across Asia, North America and Europe. It is interesting to note that the vaccination levels across Africa are in the majority less than 20%. The vaccination levels in South America are in the majority in excess of 40%.

We can readily identify the following outliers in the above plot:

High Vacc levels: Brunei (Asia) at 91.59%, Chile (South America) at 86.7%, Cuba (North America) at 85.6%, Denmark (Europe) at 84.9%.

Lowest Vacc levels: Haiti (North America) at 0.6% and Yemen (Asia) at 1.2%.

I included a pairplot in these visualizations to ascertain whether there is actually correlation between GDP levels and Covid Deaths, Cases and Vaccination levels.



From all of this summary data we can note the following:

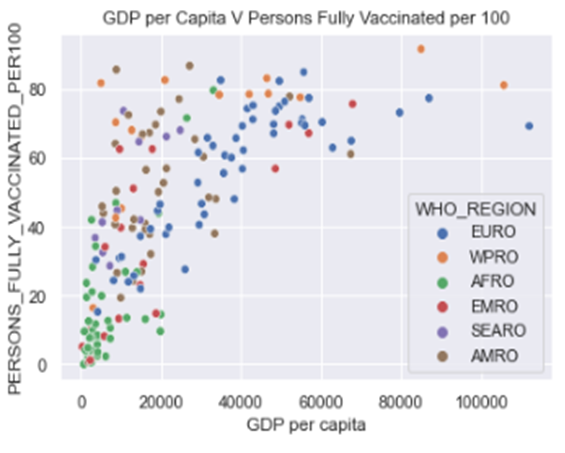
1. There is a significant correlation between Covid Cases and Covid Deaths (0.93)
2. There is a correlation between the wealth of a country and the levels of Vaccination ie. the wealthy countries have a higher vaccination levels than developing countries (0.7)
3. There is little correlation between population and GDP per capita
4. There is little correlation between population and Persons Fully Vaccinated per 100

Why is the covid case rate and death rate in Africa so low, while the vaccination levels remain low also?

* Was it early and decisive lockdown
* Was it the youthful African populations compared to many Western countries – only 2% of the population in Africa is over 70 years old
* Was it Africa’s more favourable weather a compelling factor
* Was it because 70% of the population of Africa live outside of City structures

I created a scatterplot to show meant vaccination levels across the WHO\_Regions.

There is a significant variance in mean vaccination levels across the WHO\_Regions with the majority of the AFRO region having very small mean vaccination levels. It also demonstrates that there are significant countries in EMRO and to a lesser extent EURO with mean vaccination levels of less than 40%.



I used a ***sorting*** function to establish the top and bottom 5 countries in the following categories (countries identified in red fall outside of the highest and lowest population categories):

* Population

(Highest - China 1.44 bn, India 1.39 bn, United States 333 mn, Indonesia 276 mn, Pakistan 225 mn)

(Lowest - San Marino 34 K, Saint Kitts and Nevis 54 K, Greenland 57 K, Dominica 72K, Antigua and Barbuda 98K)

* Total cases

(Highest – United States 61,556,085, India 35,875,790, Brazil 22,563,104, United Kingdom 14,651,468, France 12,251,701)

(Lowest – Vanuatu 7, Bhutan 2,802, Saint Kitts & Nevis 4,216, Antigua and Barbuda 5,058, Greenland 5,472)

* Total deaths
* (Highest – United States 839,500, Brazil 620,366, India 484,213, Russia 310,513, Mexico 300,334)

(Lowest – Vanuatu 1, Greenland 2, Bhutan 3, Saint Kitts & Nevis 28, Burundi 38)

* People fully vaccinated per 100

(Highest – Brunei 91.59%, Chile 86.71%, Cuba 85.6%, Denmark 84.9%, South Korea 83.1%)

(Lowest – Burundi 0.04%, Congo 0.15%, Chad 0.59%, Haiti 0.65%, Yemen 1.23%)

I finished the assessment by creating a dictionary of countries included in the dataframe and a second list of the Vaccination levels in each country.

I then used these lists to create a suite of statistical data using numpy. This suite of data includes minimum, maximum, range, variance and standard deviation in relation to vaccination levels across all countries.