**COVID-19 Case Count Overview**

Vipul Sapra, Brian Peterson, and Nick Snow

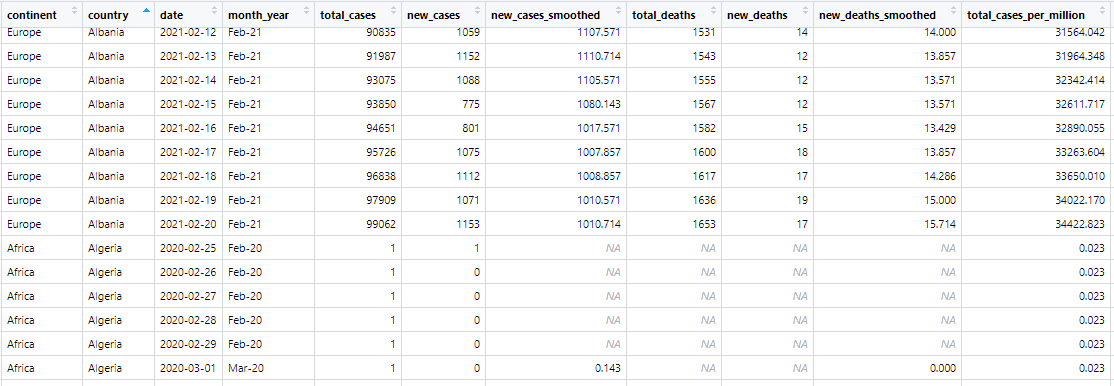
University of Cincinnati

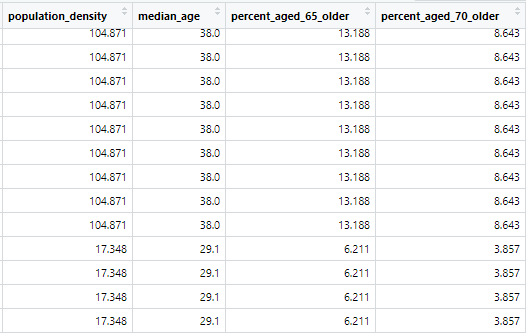
BANA 6043 Statistical Computing

February 27th, 2021

The COVID-19 pandemic has had a profound impact on people’s way of life. It has reached every country, with the exception of remote island nations. However, different countries have had different approaches to handling the pandemic, resulting in varying amounts of cases and deaths. There is a lot of data around the pandemic that can be examined. Understanding the data is important for health leaders and politicians to enact the proper safety measures for the people. Further, examining how different countries have fared in handling the pandemic can give us valuable insight about steps to take when similar events occur in the future.

We will examine a dataset provided by ourworldindata.org. Here is a sample from the dataset:

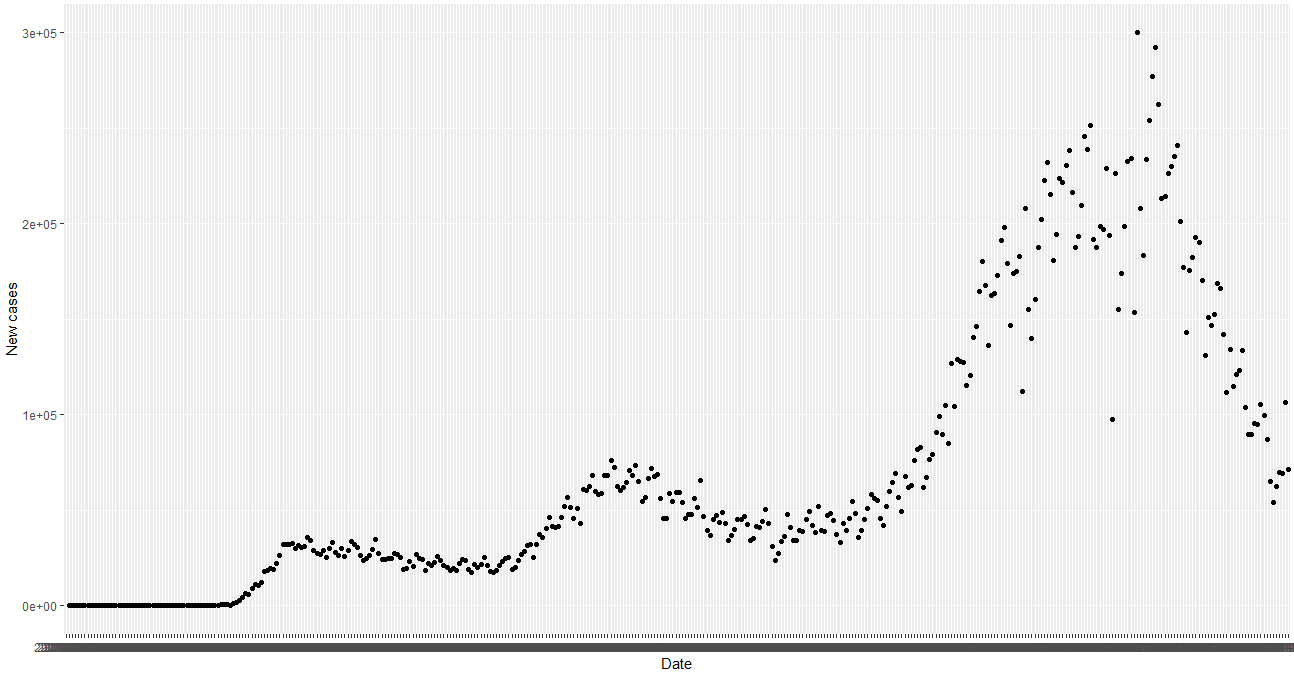
Continent and Country determine the geographical location. The date is the individual day, from early 2020 to the present day. Total\_cases are the total number of COVID-19 cases at the given date. Total\_deaths are the total number of deaths from COVID-19.

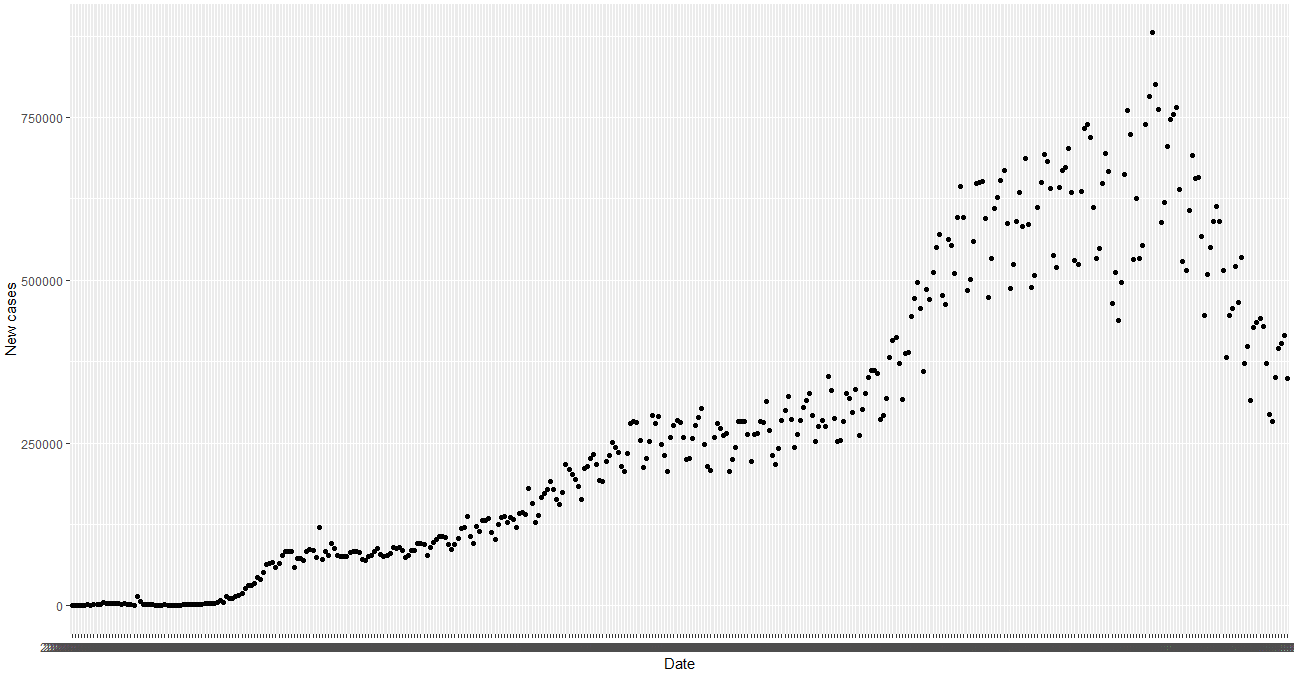


Population\_density is the number of people per square kilometer, while Percent\_aged\_65\_older is the percentage of the population that is age 65 or above.

First, we’ll look at some graphs that show the number of new daily case counts over time.

US:



World:

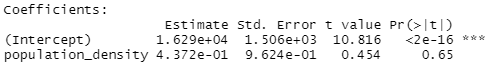
The US experienced larger spikes in cases than the world overall, meaning the case count in other countries was not increasing as much as the US. The new case counts have notably dropped for both the US and the world overall in the past few weeks.

There are multiple variables that be compared for each country. First, we’ll look at each country’s population density, and if there is a notable relation to the number of total cases per million.



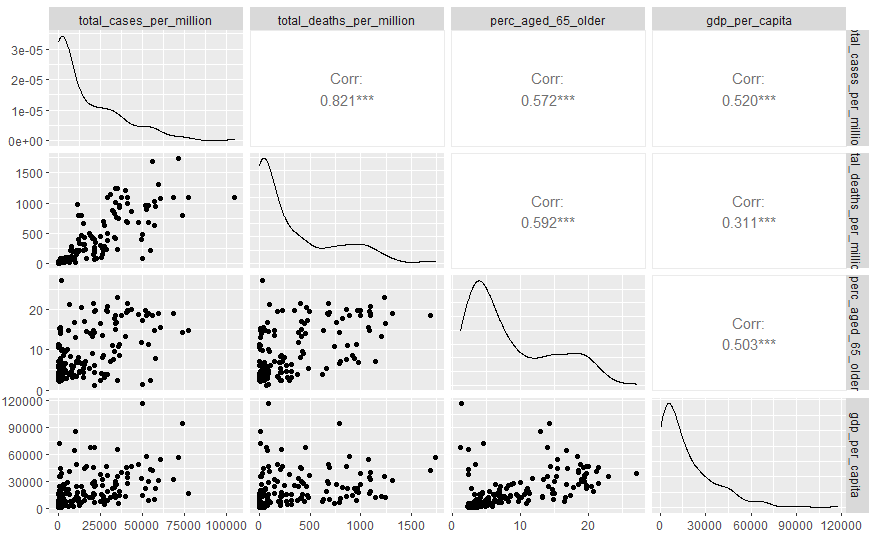
We see that there are many countries with a lower population density and lower total case count. There are a few countries in Asia with a very high population density, with differing amounts of total case count per resident. There are countries in Europe, Asia, and South America that have higher total cases per million despite having lower population density.

This is confirmed by a linear regression model of total cases per million vs. Population density which is shown below:

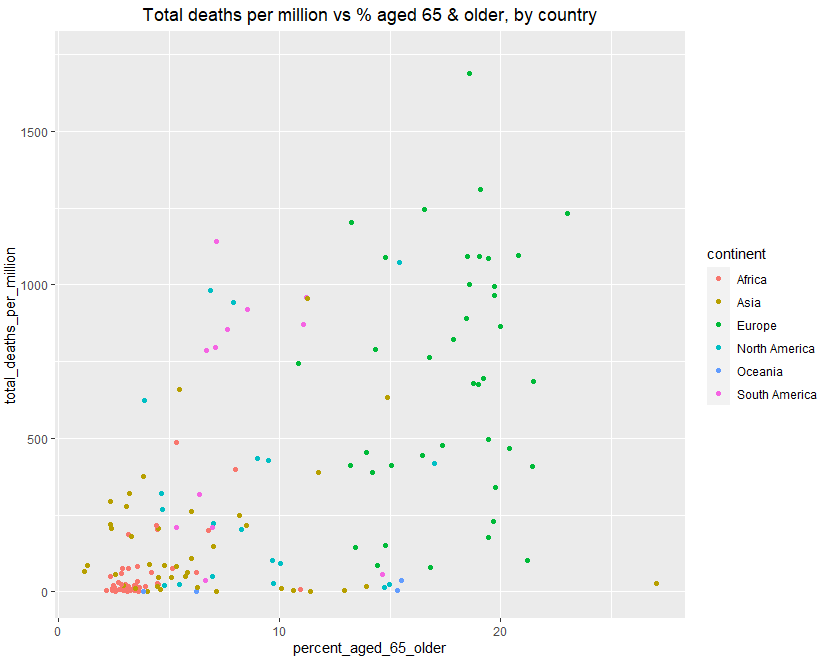


Indeed, the value of |t| above 0.05 shows there is not a great deal of correlation between population density and total cases per million.

Next, a correlation matrix of 4 dataset variables is presented. An interesting note is that gdp\_per\_capita and total\_deaths\_per\_million has relatively the lowest correlation factor of all the variables. This may suggest that health outcomes may not have a strong direct link to a country’s wealth. Something else noteworthy is that total\_cases\_per\_million and total\_deaths\_per\_million have a strong positive correlation of 0.821. Some may be surprised that this value isn’t closer to 1 since a death cannot occur without a recorded case. A correlation of 0.592 for total\_deaths\_per\_million and perc\_aged\_65\_and\_older shows that there is a positive relationship between the two variables. Age over 65 is a relatively positive predictor of covid deaths for a given country.



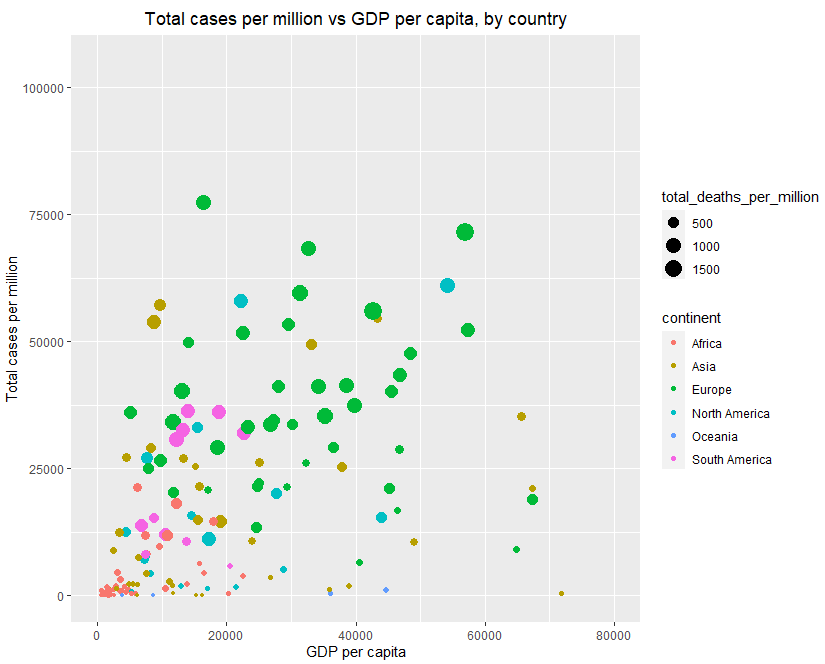
Next, we’ll look at the number of deaths compared with the percentage of the population that is over the age of 65.



Here, we see that European countries have the highest percentage of older people. Some of these countries have lower deaths per million, and some have higher. African countries have much smaller older populations and total deaths. Several countries in South America have lower percentages of older people than the European countries, but have death counts as high as many of these countries.

Another variable that we can examine is Gross Domestic Product. GDP is the abbreviation for Gross Domestic Product and will be used hereafter.

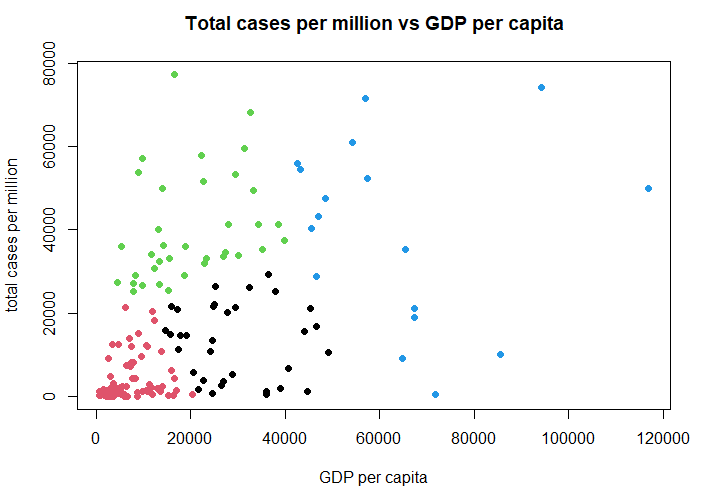
We examine each country’s GDP per capita and total cases per million. The size of the bubble represents the total deaths per million.



We see that countries in Europe, Asia, and North America have the highest GDP per capita. Some have lower total cases per million, and some are higher. We see that there are a wide variety of GDP per capita and case count combinations, meaning that how rich or poor a country is does not automatically determine how they will fare with the pandemic; other factors are at play.

There actually appears to be a small amount of positive correlation between GDP per capita and total cases per million. What could be the reason for this? Countries with higher GDP per capita are likely to have increased infrastructure and ability to travel throughout the country, which causes the virus to spread. In addition, more GDP per capita could result in higher amounts of testing, which would result in higher amounts of confirmed cases. There are people that have contracted COVID-19 that were not tested and would not appear in the final numbers.

Looking at these graphs, if we were asked to sort the countries into different groups, what would that look like? To help with this, we can use k-means clustering to determine clusters of similar data points. Running the data for total cases per million versus GDP per capita with four clusters, we get the following result:



What do the cluster groups represent? As before, each point represents a country. The red points are countries that have both a low GDP per capita and low total cases per million. The black cases have a middle range for GDP per capita, and also have lower total cases per million. The green points are countries that have low to middle GDP per capita, and have higher total cases per million. Finally, the blue dots are countries that have the highest GDP per capita, with some having lower total cases per million and others higher.

So which group is doing the best and which is doing the worst? Strictly in terms of total case count per million residents, the green countries and some of the blue are doing the worst. If we were to include GDP per capita in our determination of which countries are the “best”, the black countries are doing better than the red countries because GDP per capita is higher.

Some of the blue countries are doing better in total case count compared to poorer countries. Other blue countries, however, are doing worse. We also do not see much correlation between GDP per capita and total cases per million. This suggests that there are other factors at play besides a country’s wealth that determine how well each country has done in controlling the spread of the pandemic.

The COVID-19 pandemic is a learning experience for how the response to a pandemic should be tackled on a country by country basis and how the collective of all countries should work together. There are findings that countries can share with each other and we continue to learn more each day.

**References**

Ritchie, H. (2021, February 21). Coronavirus Source Data. <https://ourworldindata.org/coronavirus-source-data>

Ritchie, H. (2018). *GDP per capita*. Our World in Data. <https://ourworldindata.org/grapher/gdp-per-capita-worldbank>

*R: K-Means Clustering*. (2021). ETH Zürich: Seminar for Statistics. <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html>