Hyperiondev

TASK

Exploratory Data Analysis on the Coronavirus Data Set

Visit our website

Introduction

I will be performing an EDA on the worldwide data for the novel coronavirus. This data was sourced from https://ourworldindata.org/. The data is time series data until today.

DATA CLEANING

Before I began any data cleaning, I first looked at the csv file to see what kind of data I had to work with. I then loaded the csv into a data frame and set the display columns to max. I did this so that when I called the head function, I could see all the columns and get a good idea of all the data I had to work with. Next, I called the dtypes function on the data frame to see what kind of data types all the columns where. Upon inspection I could see that the 'date' column was in object format. So, I parsed that column back into itself with the correct format. I then called the dtypes function again and the column was now in datetime format. Next, I called the isnull and sum function on the data frame to see which columns have missing values. I found out that 'iso_code' had 64 missing values. This column is the column I will be using as the key. So, I made the data frame equal to the data frame where 'iso_code' was not null. Therefore, removing all the rows with no data in this column.

After this had been taken care of, I extracted the countries I wanted to look at from the data frame. I then created a separate data frame per country based on the extracted values. The countries I extracted from the data where South Africa, South Korea, USA and New Zealand.

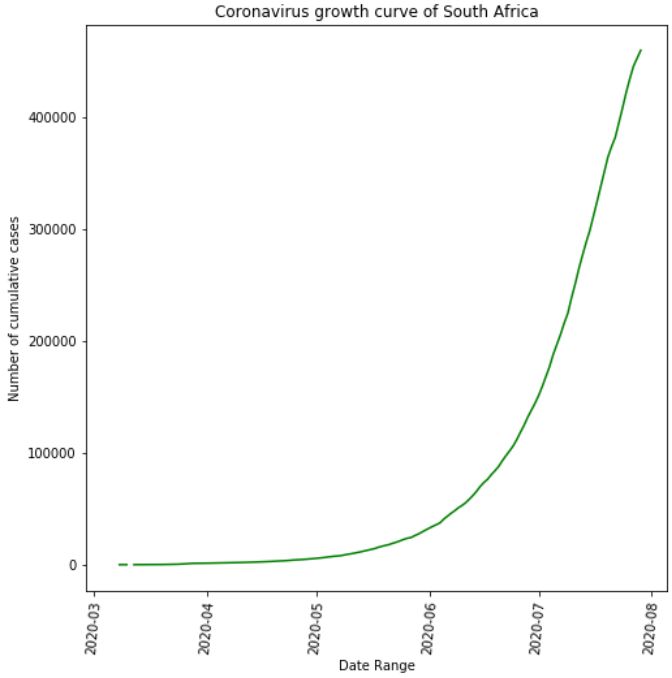
MISSING DATA

For this section, I could see that there were lots of null values in the data. However, upon inspection I deemed this to be fine as some of the null values per column where actually null. As in if there are no cases of the virus on that date there are no cases. Same goes for testing so there was no need to replace remove these rows. Nor was there a need to preform imputation. As the average would only skew the findings.

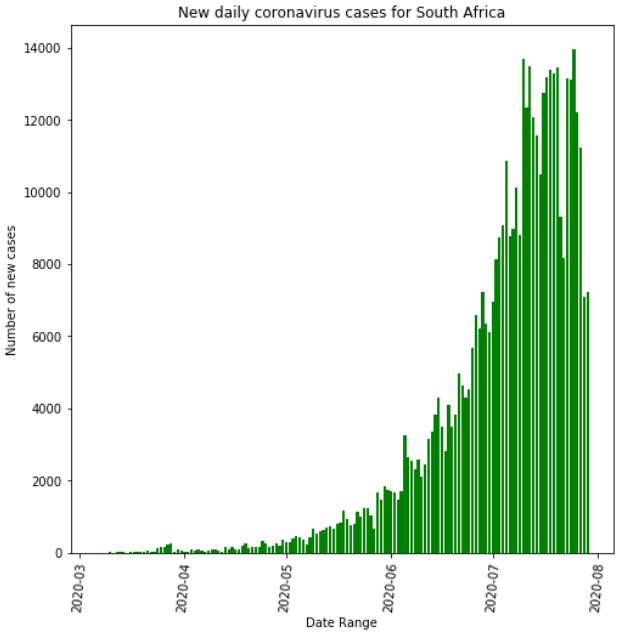
DATA STORIES AND VISUALIZATIONS

For my data stories and visualisations, I am going to group each visualisation for every country together and discuss the findings.

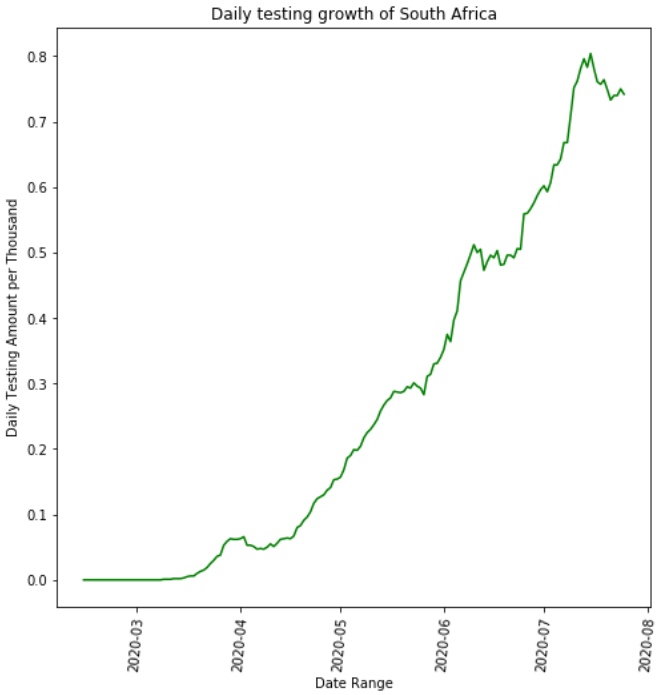
First up, we have South Africa.



So, lets unpack the graph above. This is the cumulative cases over time for South Africa of the novel coronavirus. As we can see South Africa has failed to flatten the curve. The data clearly shows that the virus has continued to spread throughout South Africa. At an accelerating rate.

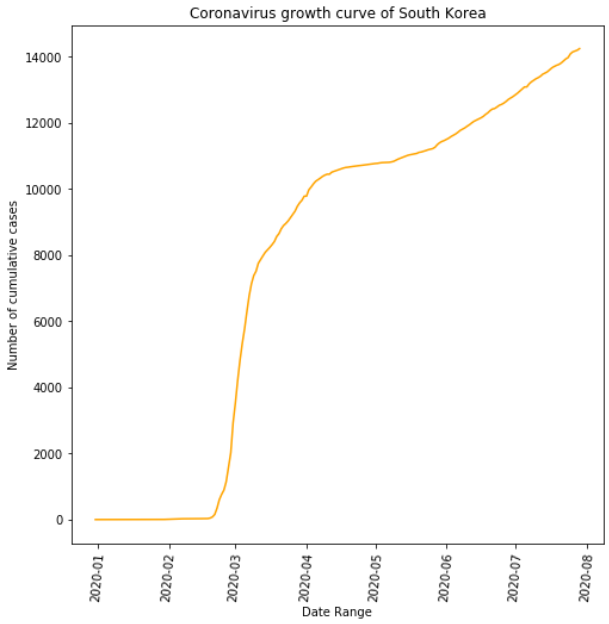


Next, we have a bar graph which gives us better insight into how many new cases of the virus are found on each day. In this figure we can see that the most cases found on a single day is around fourteen thousand. The graph does show that there is a sharp drop in new cases for two days around the 27th of July to the 29th of July. This is promising as it may mean that the number new cases reported each day may be beginning to go down. Which in turn would show on the first graph over time as the cumulative cases curve may begin to flatten. This will only hold true if the new daily cases continue with this trend.



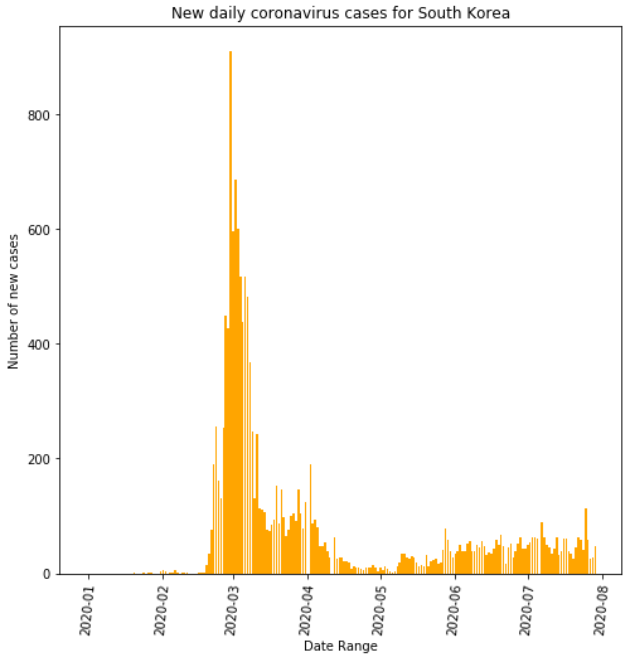
The last graph we will analyse for South Africa is the growth in testing over time. As illustrated in the graph we can see the daily number of tests has been increasing steadily. More testing means we better understand where the virus is spreading rapidly. The necessary steps can then be taken in order to minimise the spread.

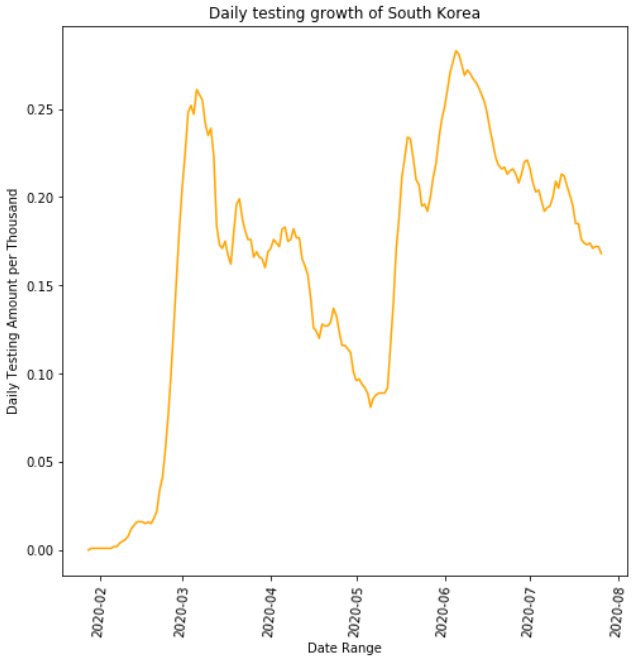
The next country we will have a look at is South Korea. They have been praised over their handling of the pandemic.



This is their cumulative cases over timeline plot. As we can see when we compare this figure to the growth curve there is a clear bend in the curve. Now what does mean? Basically, it means that they were able to slow

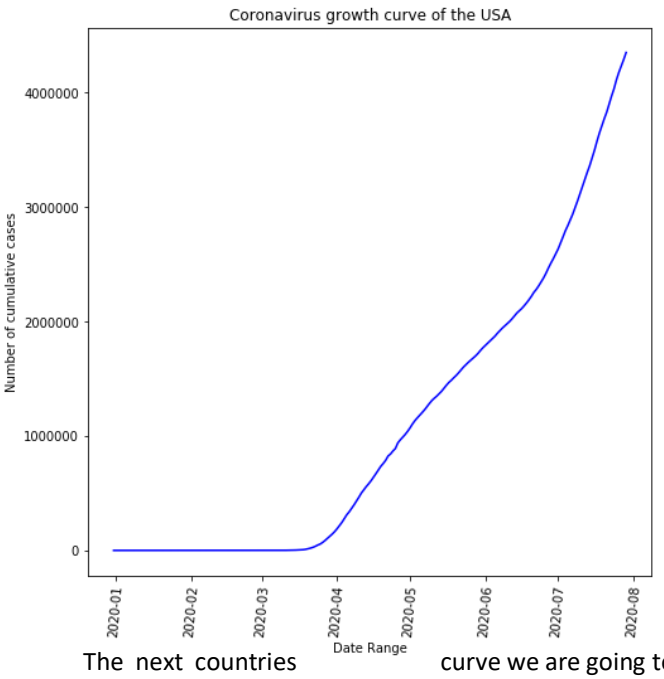
the rate of transmission. In doing so the curve no long carries on its original trajectory. This change in the curve can be see around April. Cases still increase however at a slower rate.

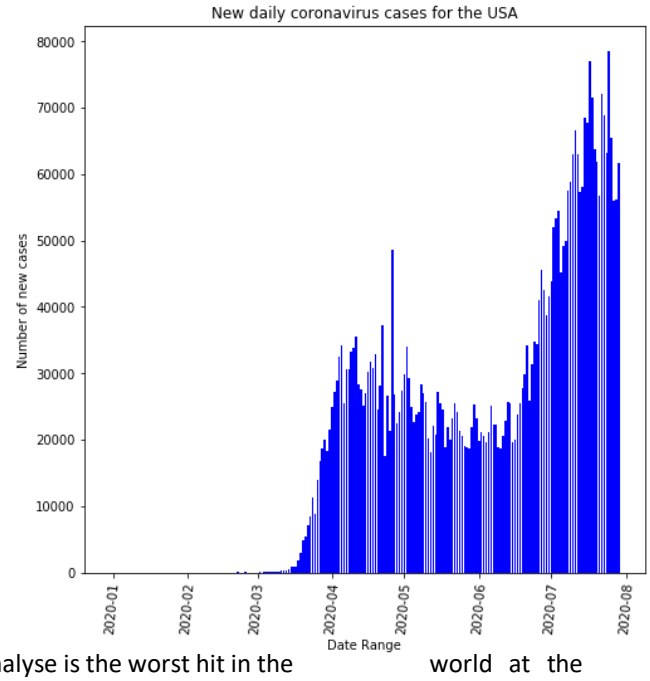




I have placed these two graphs together as I believe they paint an interesting story. Firstly, we can see that Korea test more than double the amount that South Africa do per thousand people. I believe this far greater testing amount to be the reason that South Korea have been able to slow the spread of the virus.

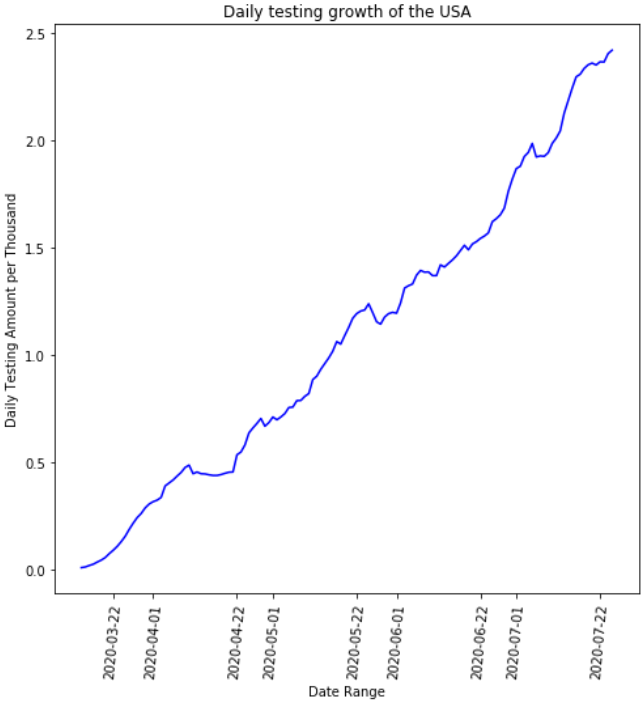
The data also tells us another story. South Korea are very reactive with their testing. Meaning, when the curve began to flatten around April so did the increase in daily testing. However, when cases started to go up again so did their testing.



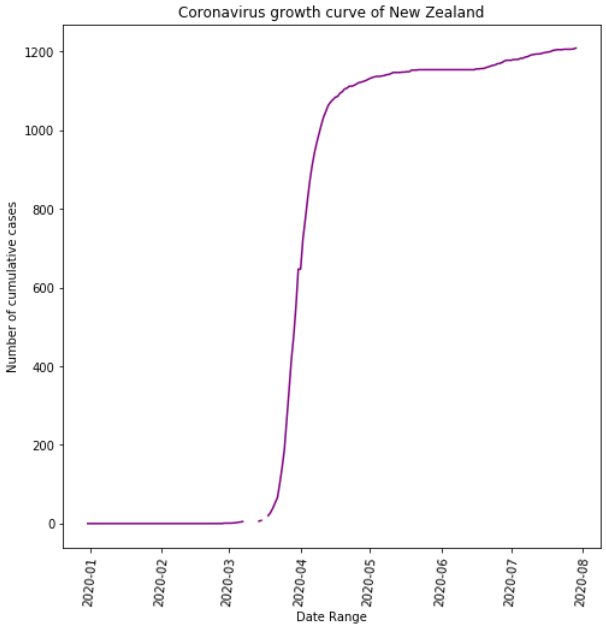


The next countries curve we are going to analyse is the worst hit in the world at the time of writing the USA. As we can see from the cumulative cases graph, they currently have more than 4 million cases. They have done a good job at ramping up testing, but it taken them a long time. Around May the

USA's growth in testing was around 0.7 per thousand people. Contrast that with New Zealand whose testing was growing around 0.9 per thousand and their country has a far smaller population.



The last country's data that will be discussed is New Zealand. I chose to include them as they arguably have had the best response to the virus. With only 1200 cases at the time of writing. Another reason I chose to include them is unlike the other countries in this dataset they are an island. I wanted to see if this had an effect on the spread.

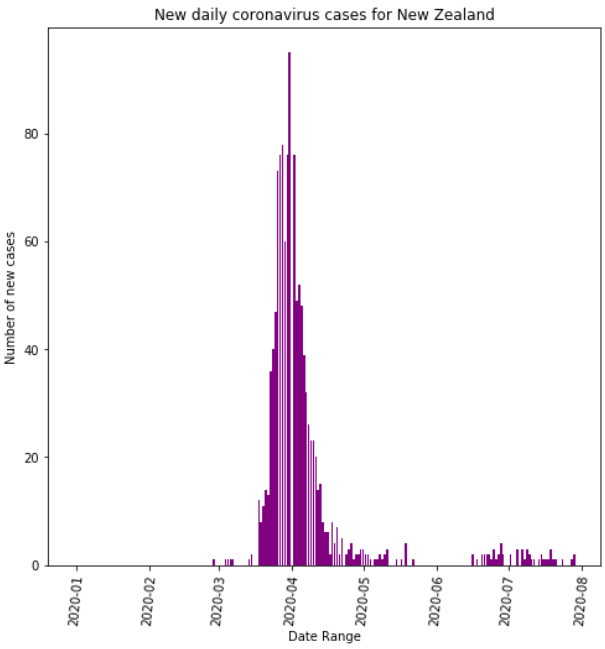


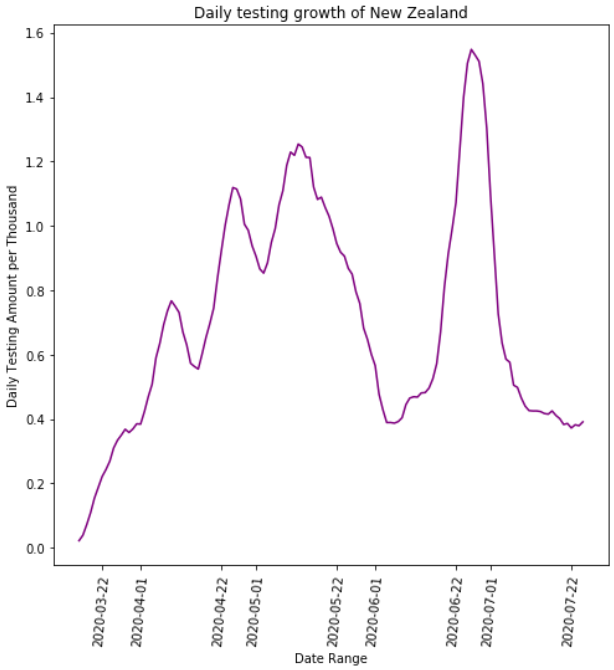
With that being said let's have a look at the cumulative cases overtime.

We can see that their cumulative cases overtime curve looks remarkably similar to South Koreas. Although South Korea has far more cases.

This means like South Korea; New Zealand were able to slow the spread and change the trajectory of their curve.

Was this due to testing? Well they didn't and still don't have a very large amount of cases. So, there is no need to do a large amount of testing. However, they are still preforming far more testing than South Africa who have been hit by the virus way harder.





So, to conclude the data shows out of the four countries in the data the two countries that ramped up testing quickly New Zealand and South Korea where able to contain their outbreaks. At the very least slow the rate of infection. When you compare the data for America who when factoring in the size of their population are yet to achieve sufficient testing numbers. And South Africa who are far off where they need to be when again factoring in their population size. These countries were unable to ramp up testing to an effective level. Consequently, their curves are yet to show any sign on slowing down. South Africa has shown some merit in the last 2 days (at time of writing). As the number of infections reported has decreased however, only time will tell if this will develop into a trend or if it is just a temporary dip.

THIS REPORT WAS WRITTEN BY : Matthew van der Westhuizen