A comparative study of COVID-19, H1N1 and the Ebola virus.

1. Data set description, including where the data are from and any data cleaning, integration or transformation conducted on the data. Please include the URLs of your data sets

https://www.kaggle.com/de5d5fe61fcaa6ad7a66/pandemic-2009-h1n1-swine-flu-influenza-a-dataset

https://www.kaggle.com/imdevskp/ebola-outbreak-20142016-complete-dataset

https://covid19.who.int/table

I sourced the datasets for H1N1 (swine flu virus) and the Ebola virus from Kaggle. I had to make some assumptions when using the data. The Covid-19 (Novel Corona virus) data set was found directly from the WHO website. This one was the most extensive dataset, most probably due to the fact that it is the most widespread one, with the most accurate data and testing. For the H1N1 data set, I had to clean up the formatting for dates, which was not consistent. Additionally, I had to assume that the 'Cases' and 'Deaths' were cumulative totals, and not daily counts. I reached this conclusion by combing the data and figuring out that cases/deaths always increased. This would not meet the reality where cases initially increased, but soon decreased.

2. Application scenario description

The coronavirus outbreak has affected lives all over the globe. It has led to a global lockdown, causing massive economic losses to Countries and individuals. The enormous scale of this virus, and the way it has affected different countries differently is just one piece of the puzzle that is Covid-19. I believe that the data I am presenting puts the sheer scale of this virus into perspective, and also dives deeper into how differently the virus has spread as compared to other outbreaks in the past.

My main task was to compare different metrics for Ebola, H1N1 and Covid-19.

I was able to discover and compare the following for Ebola, H1N1 and Covid-19.

- Number of total cases
- Number of total deaths
- Mortality rate
- Number of infected countries
- Time taken to spread

I also focused more on the Covid-19 dataset, and compared:

- Mortality in countries.
- Cases/Deaths per region.

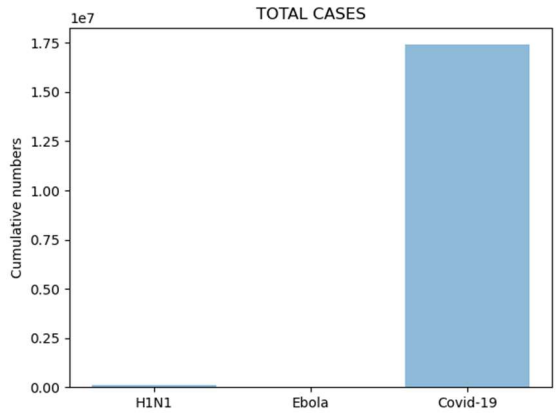
I believe my report helps us to compare viruses from the past and how we dealt with them before as compared to now. It also shows how contagious Covid-19 is as compared to other viruses.

3. Description of your method and implementation

I utilized python to code this project. I made the most of its graphing capabilities and created lots of visualizations of the data. I used from LinearRegression from the sklearn toolkit, to show compare the speed of the spread between Ebola, H1N1 and Covid-19. I used a heatmap tool to visually depict how the three viruses spread amongst countries, and how many deaths were in each of these countries.

I used pandas, numpy and excel extensively to complete this project.

4. Data mining result presentation and visualization



The first visualisation I created, was of the total reported cases for each of the viruses.

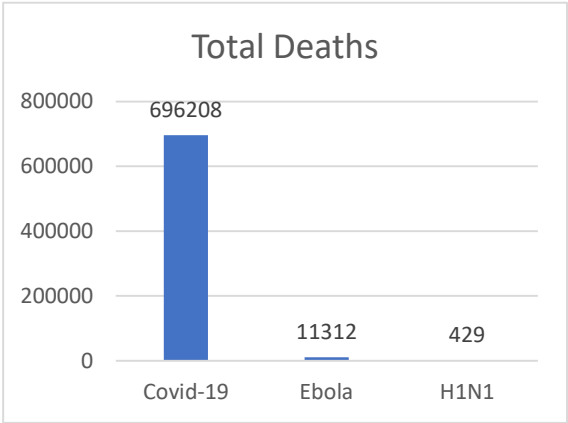
Covid-19: 17,411,592

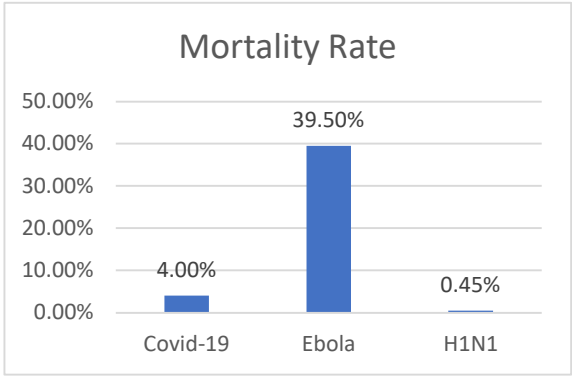
Ebola: 28638

H1N1: 94512

It is clear that the gulf between the three viruses is huge. This is also due to the data being incomplete or cases be underreported for both Ebola and H1N1.

The total deaths showed a similar story. However, the mortality rate was what truly expressed how lethal each of the viruses were. A simple bar chart puts into perspective just how much Covid-19 has spread as compared to other viruses. The total death of toll of Corona is also far more than the other two viruses, according to the data sets that I acquired, however this may be due to the sheer volume of cases present for covid-19.



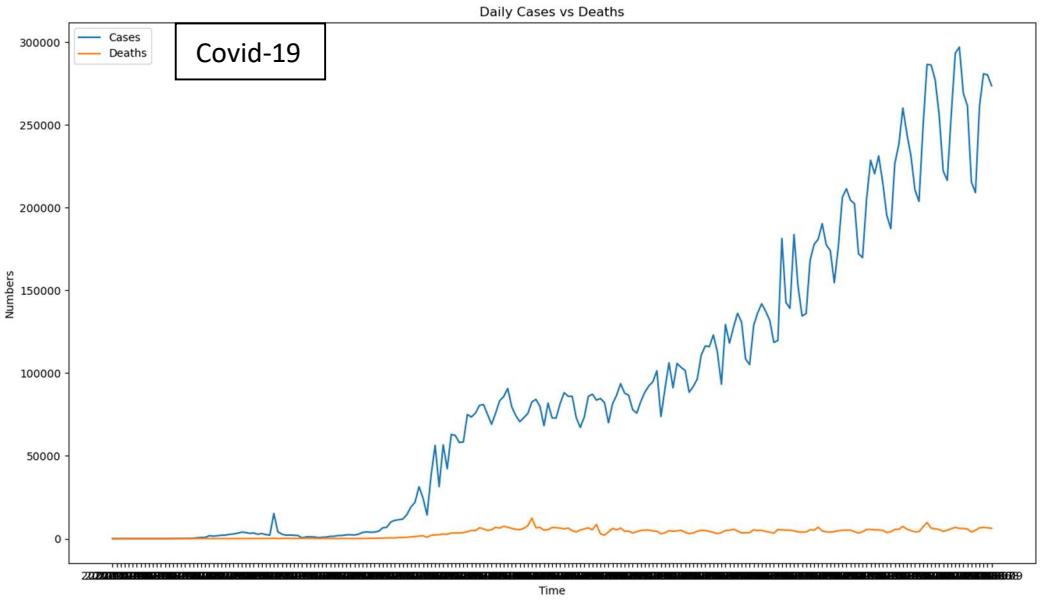


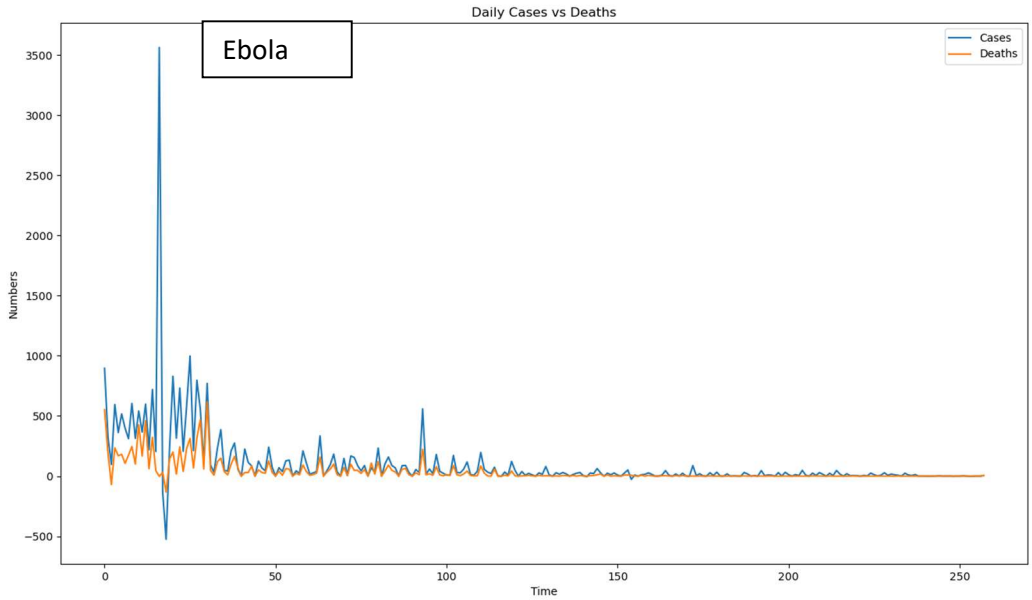
Taking a deeper dive into how dangerous each virus is. I found that the mortality rate of Ebola was far greater than that of H1N1 and Covid-19. This supported my earlier theory that the death toll for covid-19 is only much greater due to the number of people effected.

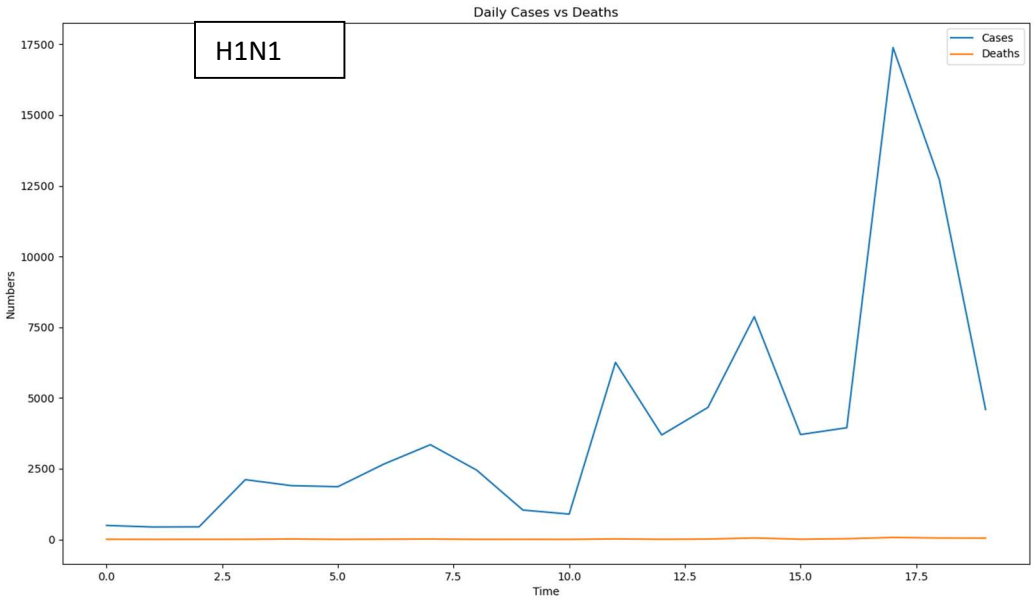
Although, Covid-19 is known to have been more lethal for higher age groups. I was not able to test this hypothesis, since age groups were not included in the dataset.

The Swine flu (H1N1) virus according to the data, was the least lethal virus of the three, with only a 0.45% mortality rate.

Next, I ran a comparison of daily/weekly deaths vs daily/weekly cases. I ran this for all three viruses, and it gave some interesting results. The covid-19 dataset was the most interesting in particular.





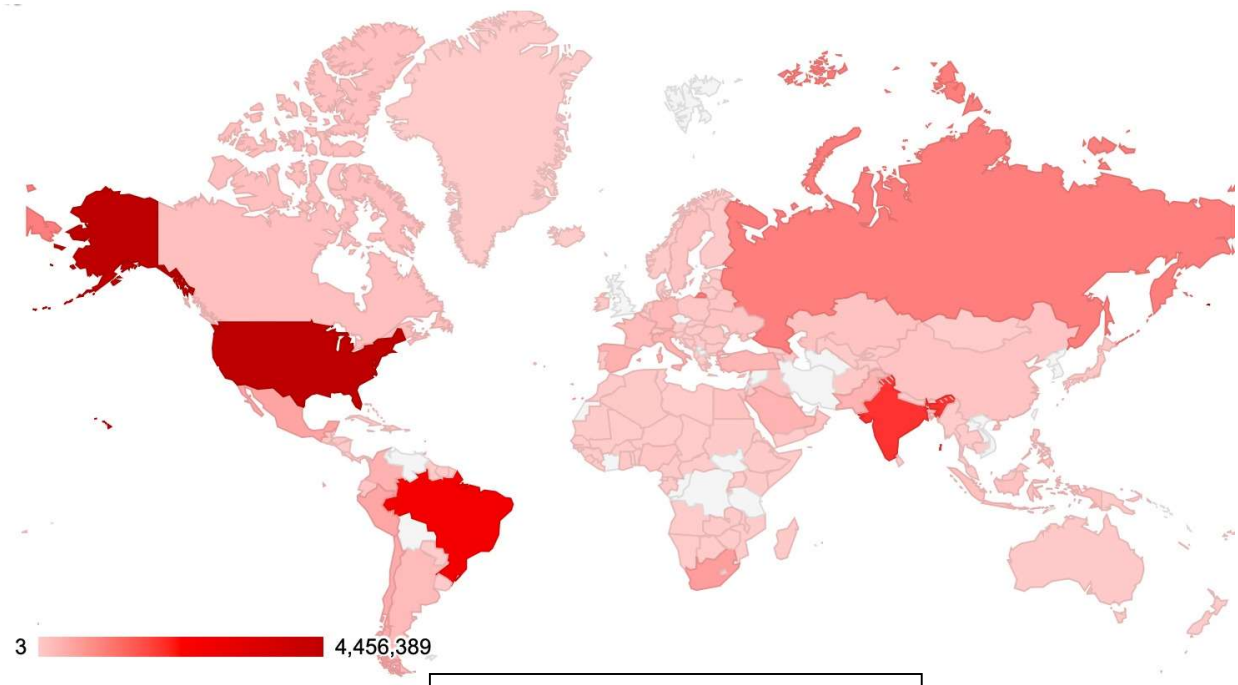


We can see from the data that although cases are increasing, the death tolls seem to remain the same in case of H1N1 and Covid-19. However, with Ebola, the number deaths seem to follow the trend of the number of cases. The data mined from the Ebola dataset returned a negative value for one of the weeks, this outlier may have been due to the fact that there was false reporting that was rectified at this date.

The spread of the viruses is clear when comparing the number of affected countries. Ebola virus although it did not affect even a 10^{th} of H1N1, it had significantly a higher death toll.

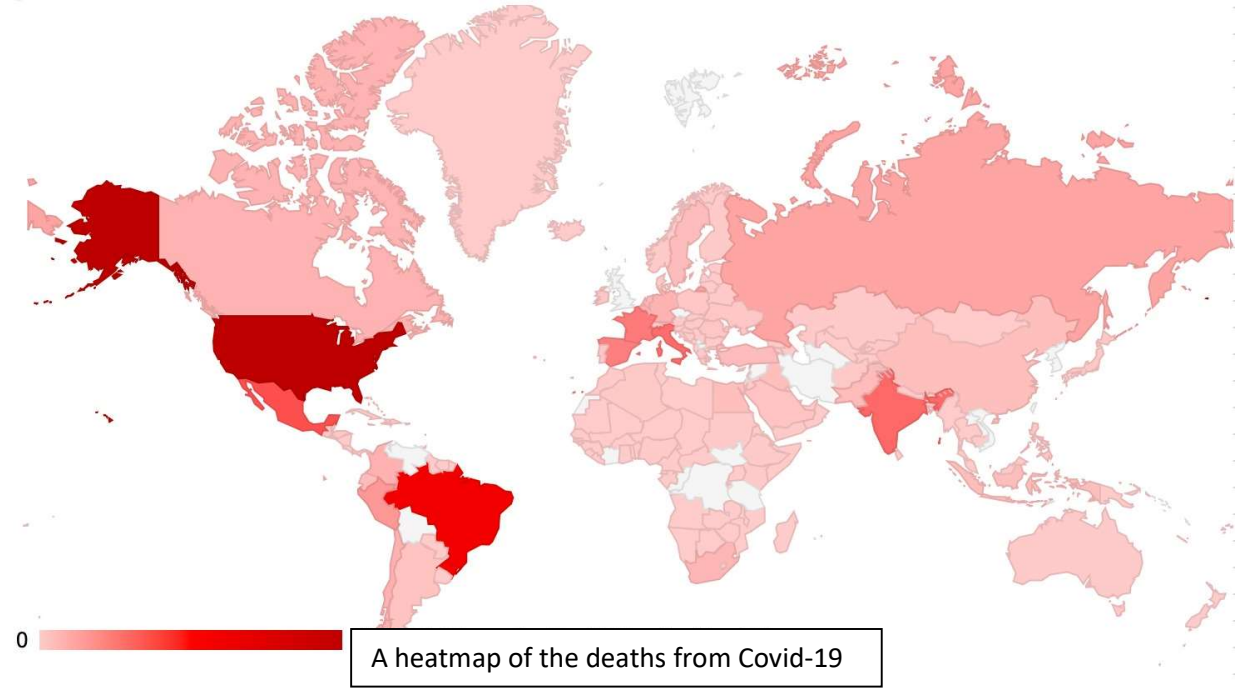
Virus	Covid-19	Ebola	H1N1
Countries affected	216	10	152

Going deeper into the analysis of Covid-19.



A heatmap of the cases of Covid-19

I constructed a heatmap to depict how different regions were doing in terms of cases of Covid-19. Unsurprisingly to most, USA, Brazil and India seem to be the lowest performing countries.



A heatmap of the deaths from Covid-19

On the other hand, the heatmap for deaths per country tells a different story. USA is doing extremely terribly in this metric, along with Brazil. And although India still seems to not being doing that well, it is relatively doing much better than the other two.

But by looking at the mortality rate for those three countries, we can see Brazil doing the worst, with India following, and lastly USA. Below is are the numbers for the 3 worst performing countries in terms of mortality (Yemen, France and The UK) as well as the 3 most infected countries.

I have attached a csv file with data for all countries available.

usa	Brazil	India	Yemen	France	The United Kingdom
1.704545	3.361146	2.014807	28.48418	16.29377	15.03258901

Lastly, I showed distribution of covid-19 by region.

1 African Region (AFRO)

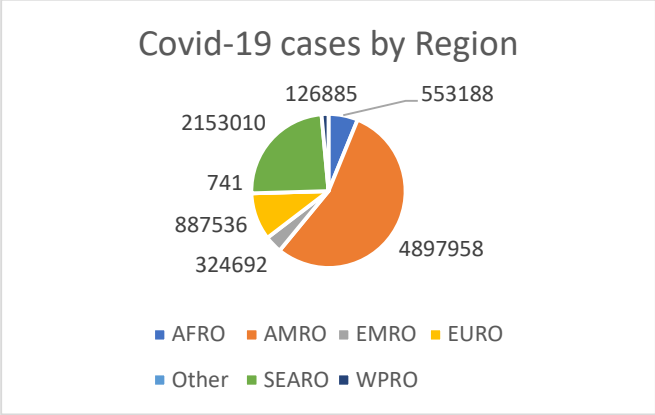
2 Region of the Americas (AMRO)

3 South-East Asia Region (SEARO)

4 European Region (EURO)

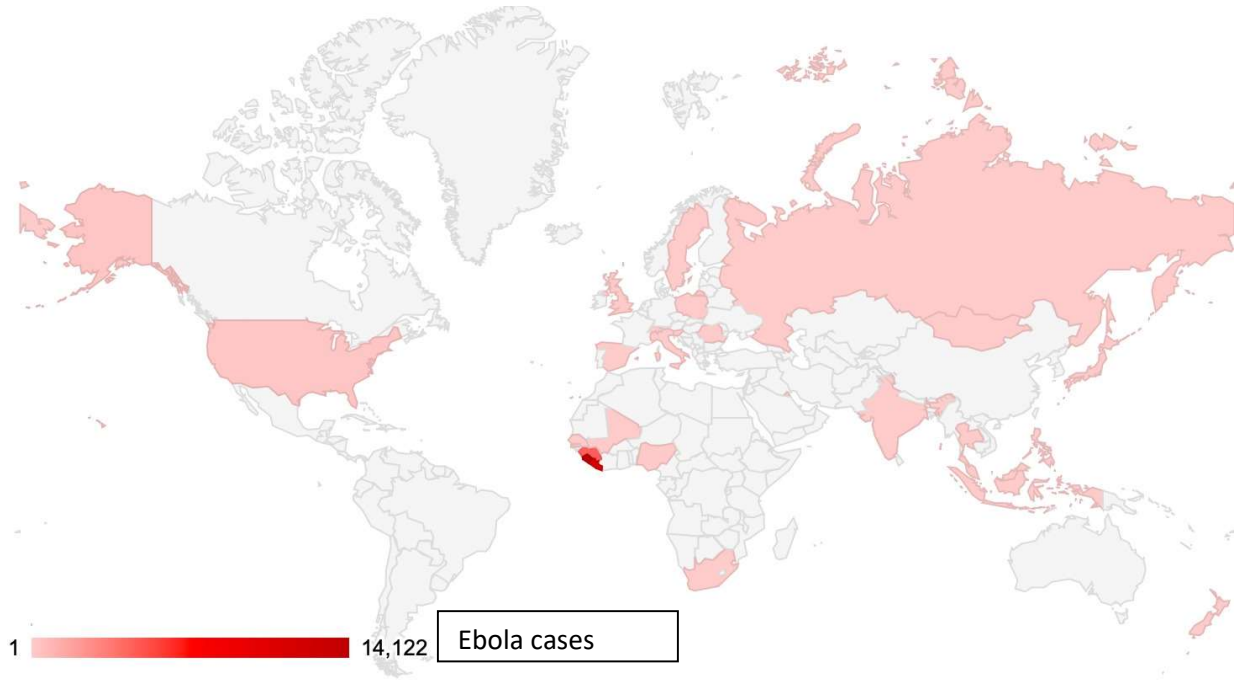
5 Eastern Mediterranean Region (EMRO)

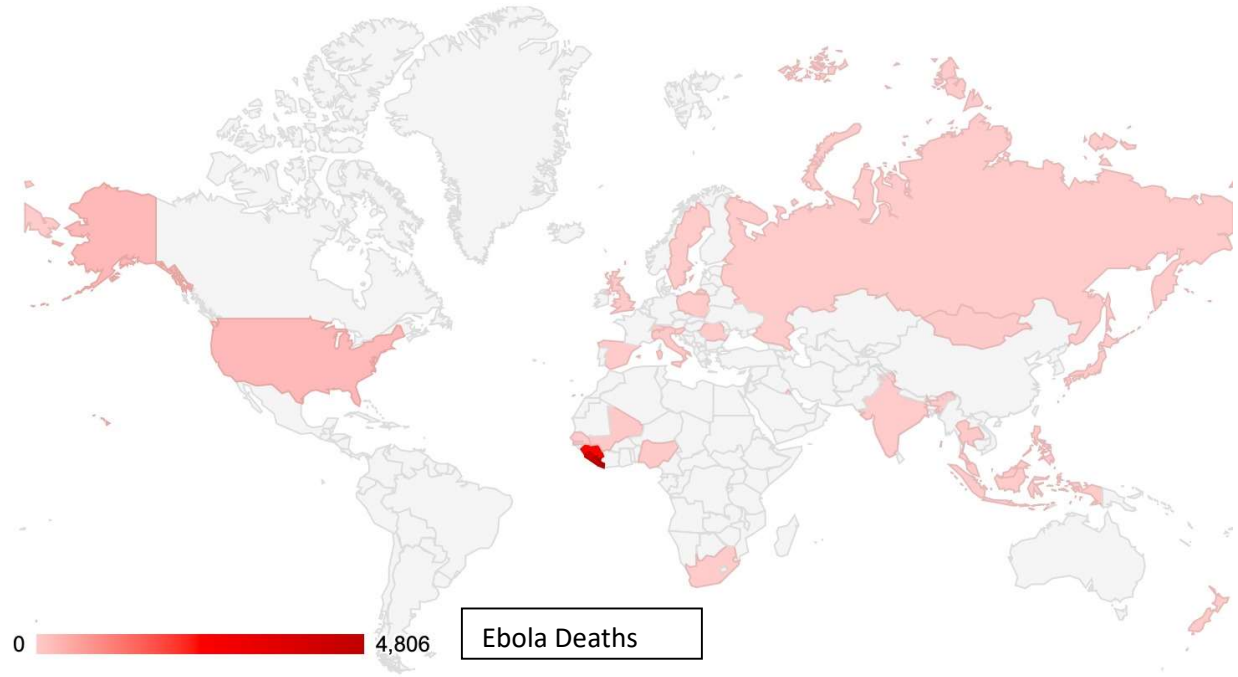
6 Western Pacific Region (WPRO)

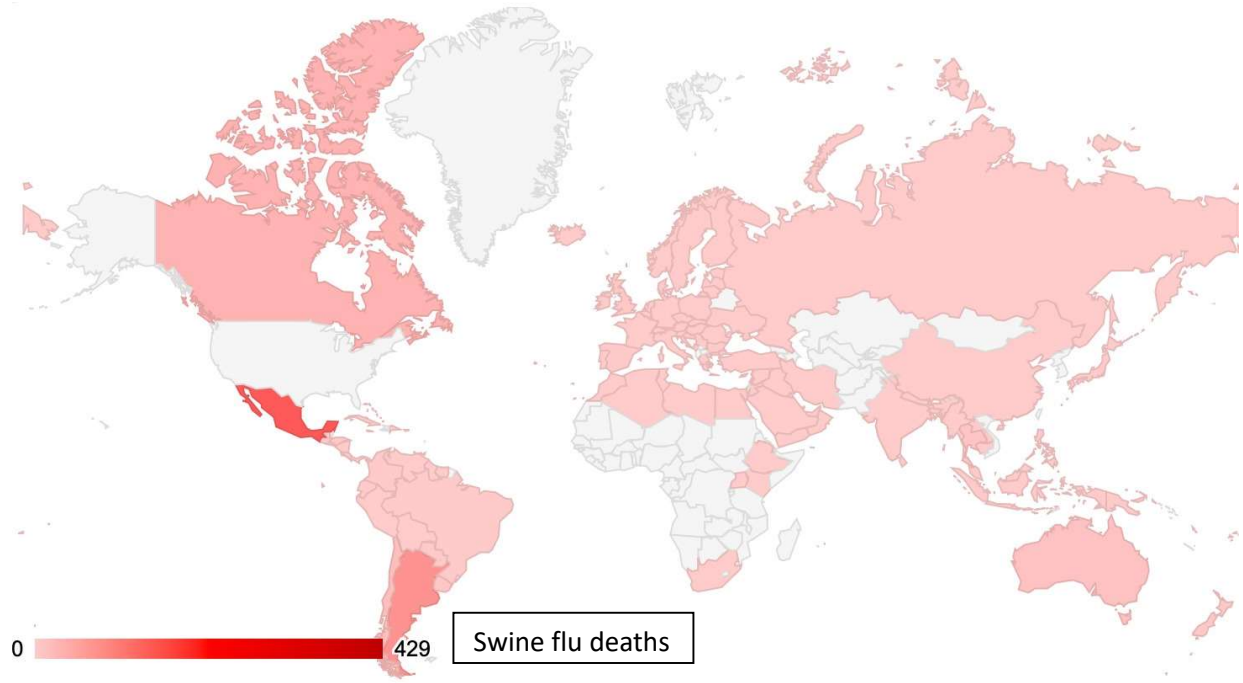


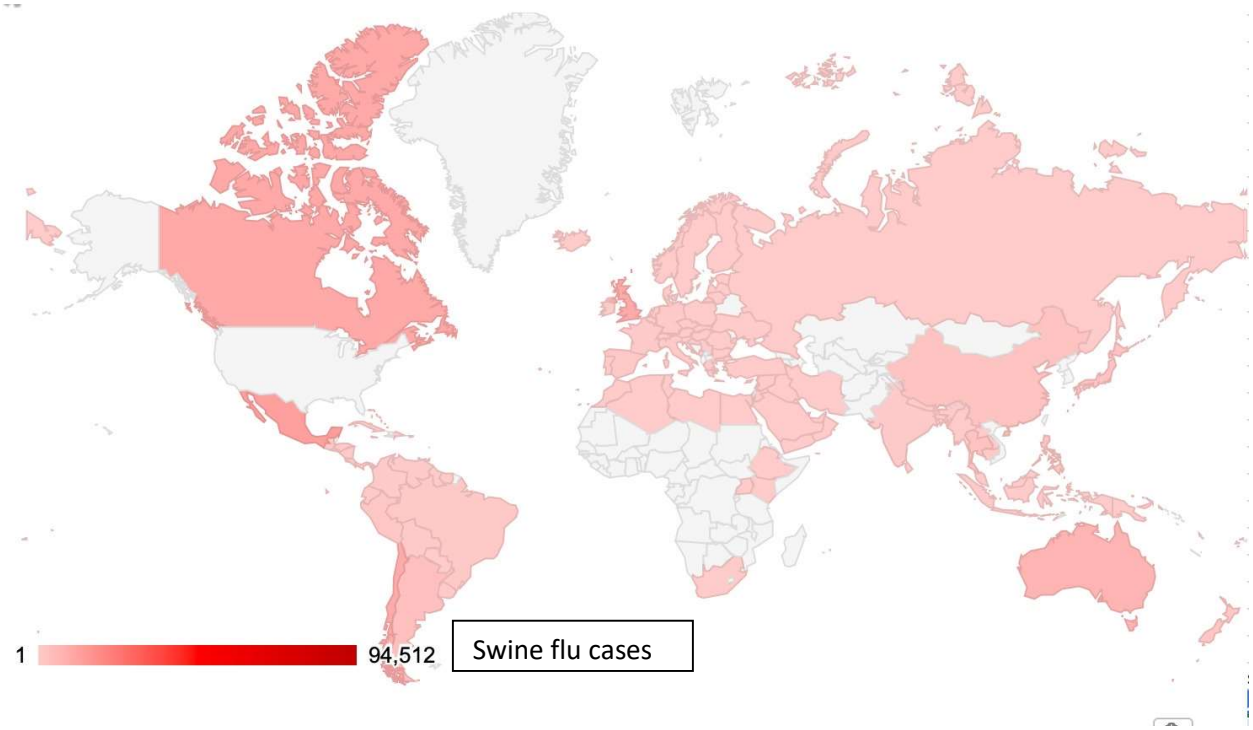
The pie-chart helps to show just how many cases belong to the Americas regions, as compared to the rest of the world. This would be significantly from the US, and Brazil.

I also made heatmaps for Swine flu and Ebola, to show their spread around the globe.


1
14,122
Ebola cases


0
4,806
Ebola Deaths


0
429
Swine flu deaths


1
94,512
Swine flu cases