COVID-19 DataViz Challenge

Participant Name: Han Liang Wee, Eric
Email ID: e0363460@u.nus.edu
URL of D3 Visualization:

1. **Q1:** https://eric-han.com/covid-19/q1.html
2. **Q2:** https://eric-han.com/covid-19/q2.html
3. **Q3:** https://eric-han.com/covid-19/q3.html
4. **Q4:** https://eric-han.com/covid-19/q4.html
5. **Q5:** https://eric-han.com/covid-19/q5.html

Introduction

In this paper, we will analyse the data for COVID-19 to inform the public about the recent COVID-19 health crisis and present some actionable, striking insights. In our work, we will use the COVID-19 dataset[1] as a basis to conduct our analysis, along with a few other supplementary datasets to enhance our analysis. Our aim is to provide accurate, informative and actionable insights to the public so that individuals and businesses can protect themselves adequately.

Dataset

We will be using part of the COVID-19 dataset, more specifically:

1. **covid_19_data.csv**: Contains the number of cases (confirmed, deaths and recovered) quantitative attributes for every country (identity, categorical) since 22 Jan 2020. We understand that patient zero was infected before 22 Jan 2020, but it will have minimal impact on the analysis as the number of cases before 22 Jan 2020 is minimal.
2. **COVID19_line_list_data.csv**: Contains an archive of individual cases that happen throughout the world, that includes attributes such as age (ordinal) and death/survived (categorical).
3. **COVID19_open_line_list.csv**: Also contains another archive of individual cases. We note that this dataset is considerably larger than **COVID19_line_list_data.csv**, however, the data quality is poorer. It includes many free-formed descriptions in fields that were supposed to be categorical.

Governance Indicator Dataset

Source: https://info.worldbank.org/governance/wgi/

[1] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

We are using part of the dataset, only looking at Government Effectiveness (GOVEFF). In the GOVEFF, government effectiveness have been measured for over 200 countries over the period 1996 - 2018. In our project we will be only looking at the 2018 value for government effectiveness. In the GOVEFF study, the metric used is composed of several measures of governance such as transparency, political stability. We will only be interested in the latest GOVEFF mean value in the year 2018. A detailed description of the derivation of the metric is given here[2]. The metric is a quantitative value with range from -2.5 to 2.5. -2.5 would mean that the government is not effective, whereas a value of 2.5 would mean that the government is very effective. The reasons for why we chose this dataset would be apparant later.

Other Datasets Considered

We would also like to highlight the other datasets that we considered but not used in the final analysis. We considered the following datasets:

1. **CPIA transparency, accountability, and corruption in the public sector rating[3]** was considered but was not included as governance indicator takes into account this metric(1-6 quantitative value), which considers other metrics and is wholesome.
2. **Current health expenditure per capita (current US$)[4]** was considered but in our analysis it was a poor indicator to analyse how a country perform in this crisis.

We will omit the graphs concerning these datasets in the interest of space and focus on the datasets aforementioned.

Pre-Processing

The preprocessing was done in Python and is specific each analytic question (ie. merging 2 datasets on a key). In this section, we will discuss some of the pre-processing done across the datasets. We will discuss any pre-processing that is specific in the next section.

Country Names

The country names across datasets had huge variation. For instance, the city-state of Hong Kong can be referred to in one datset as Hong Kong, China and in another as Hong Kong SAR, China and yet another: Hong Kong. These differences in country names are due to political or geographical differences in the organizations that collect them. We normalize the country names across datasets by writing adaptors and/or using the standardized 3-digit country code. We regonize that country is the identity channel for a lot of our analysis questions and need to be dealt with effectively.

---

[2] https://info.worldbank.org/governance/wgi/Home/Documents
[3] https://data.worldbank.org/indicator/IQ.CPA.TRAN.XQ
[4] https://data.worldbank.org/indicator/SH.XPD.CHEX.PC.CD

Inconsistent Data

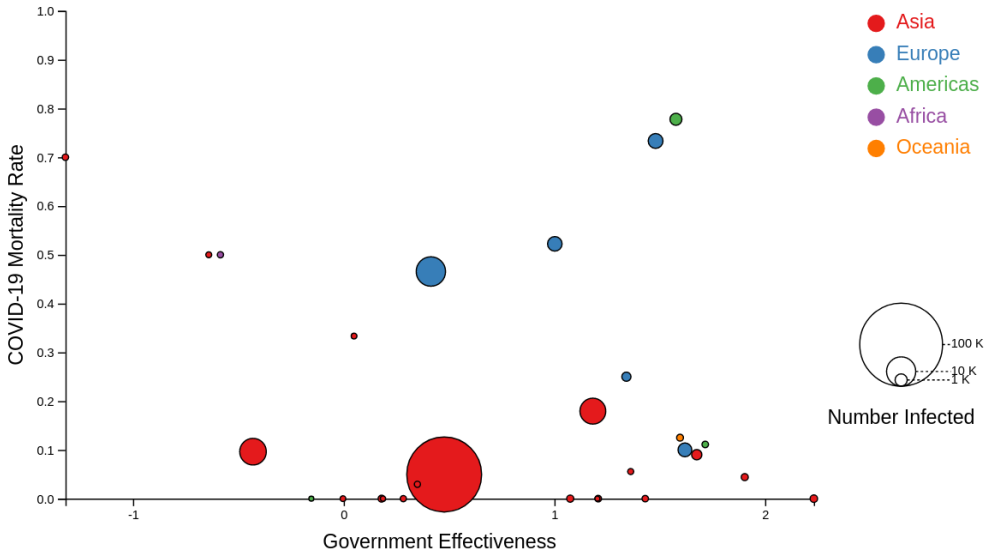
There were many issues with the datasets, most notably on missing values or inconsistent data. The technique used to deal with such data is to normalize the data. For instance, in the **COVID19_line_list_data.csv**, there were various ways to represent data that is not avaliable. One of the ways used would be simply to leave the data cell empty, while there are others: N/A, NA, na. We simply wrote an adaptor to normaize the data to using the same symbol to represent the same thing. In this case, we used an empty data cell to represent throughout the dataset consistently.

Visualizations

In this section, we will describe the analytic question we pursued, the insights we gained and also visual encoding of the visuzliations.

Q1: Does Government Efficiency have correlation to death?

Direct Link: https://eric-han.com/covid-19/q1.html



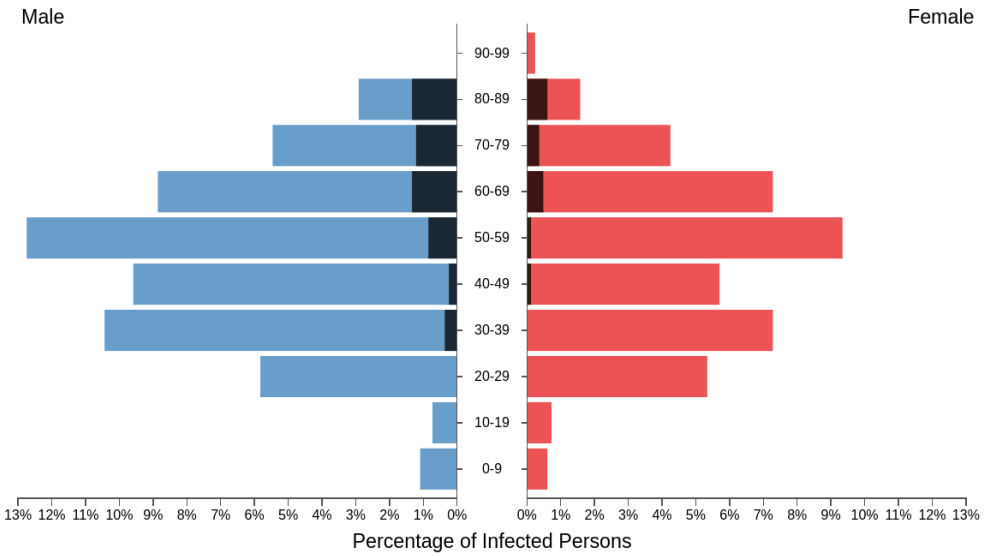
Category: Statistical aspects - correlation
Visual Encoding:

1. **Magnitude, Position (y)**: Mortitality Rate - quantitative, also known as death rate
2. **Magnitude, Position (x)**: Government Effectiveness - quantitative between -2.5 and 2.5, with 2.5 representing most effective.
3. **Magnitude, Area (2D)**: Number of people infected - quantitative, the larger the circle the more infected.
4. **Identiy, Color Hue**: Continent with differing colors - categorical
5. **Identiy, (motion) tooltip**: Country with the number of days since local patient zero.
6. **Mark - Circle**: Each cicle represents the country - categorical. It is chosen so it can be easily identifed and there is meaning to the size of the circles.

Additional Comments: We compute mortality rate to be (Number of Death)/(Number of Death + Number of Recovered). Due to the metric used, we need to be careful of countries which just started the outbreak, hence we also show the number of days since the first local infection for each country in the tool tip. Also, we merged the Government effectiveness dataset and the covid_19_data on the key - Country Code. We also used data, accurate to 03/10/2020.
Insights: Perhaps a suprising finding that government effectiveness is one of the indicators determining the mortality rate. In a sense, it is not the only metric that would determine death rate due to COVID-19, but the higher the government effectiveness, the chances are better. When the government is not effective, more things need to be right in order for the death rate to be low. When the government is effective, less things need to be right as the government is effective. For instance, there are 2 clear outliers - USA and France, these 2 countries took an nonchalant attitude in the eariler days of infection and that attitude is affecting morality rate.
Conclusion: Any persons living in a country such as Singapore with good governance would have a much greater chance of surviving a pandemic such as COVID-19.

Q2: How much better does particular age group and sex cope with COVID-19?

Direct Link: https://eric-han.com/covid-19/q2.html



Category: Quantitative aspects
Visual Encoding:

1. **Identiy, Position (y)**: Age groups channel, ordinal between 0-9 to 90-99
2. **Magnitude, Position (x), unaligned**: Percentage of the total infected persons, quantitative
3. **Identiy, Color Hue**: Gender and death - Red for female, blue for male and black for patients who died from COVID-19.
4. **Identiy, (motion) tooltip**: Displays the exact numbers for that age,sex combination and number of deaths, quantitative.
5. **Mark - Bars**: Chosen as each category is not *strictly* continous over the next, hence each age category is represent separetly as individual bars.

Insights: Firstly, the number of infections for children is abnormally low. One possible explaination is that younger people's immune systems deal better with COVID-19 infection and they are not even reported as they get well possibly before reporting. Another possible explaination is that more precautions are taken over the young and that might be a reflection of the effective social distancing measures taken by the respective governments. Next, the death rate is higher for older folks. One possible explaintion is that older folks have poorer immune, which is hotly discussed in the news now. Curiously, the number of females infected is markedly

lower compared with the males. It might be that females bodies handle the virus slightly differently compared with males.
Conclusion: Younger people is much less likely to die from COVID-19, also females are less likely to get infected from COVID-19.

Q3: What are some of the likely common symptoms of COVID-19?

Direct Link: https://eric-han.com/covid-19/q3.html

Common Symptoms of COVID-19

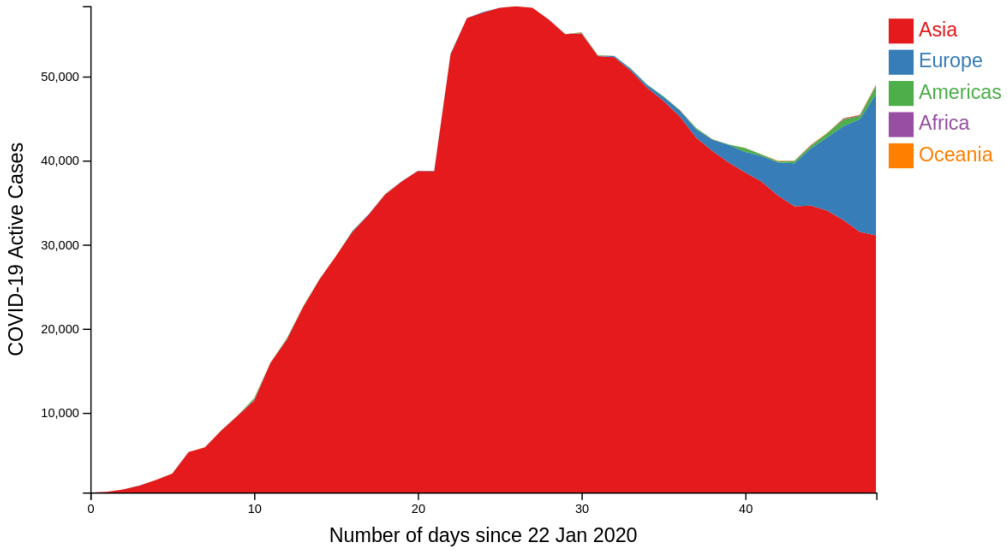
Category: Statistical aspects - clusters
Visual Encoding:

1. **Magnitude, Color saturation**: Number of observations of the symptoms, quantitative
2. **Magnitude, Area (Font Size)**: Number of observations of the symptoms, quantitative
3. **Mark - Identity, Text**: The symptom, categorical, chosen so one can identify the symptom.

Insights: The most common symptom is fever. However, just merely looking at only fever is meaningless as many illness would have fever as one of the symptoms. Hence, plotting the various symptoms would be helpful to diagnose COVID-19. The world cloud would show the intensity of the symptoms across the population as reported. Suprisingly pneumonia is not one of the most (top 5) reported symptom.
Conclusion: If you have fever and cough, you should check if you have other symptoms from the list here.

Q4: How has the situation improved over time in different parts of the world?

Direct Link: https://eric-han.com/covid-19/q4.html



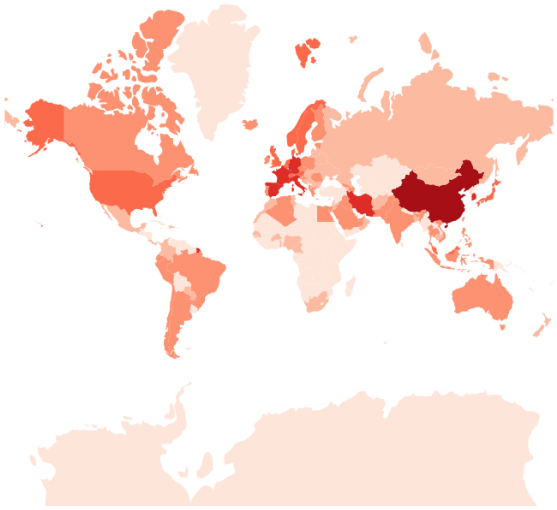
Category: Temporal aspects
Visual Encoding:

1. **Magnitude, Position (y)**: Active cases of COVID-19 Infection - quantitative
2. **Magnitude, Position (x)**: Number of days since 22 Jan 2020 - quantitative
3. **Identity, Color Hue**: Continent - categorical
4. **Magnitude, Area**: Improvement of the particular continent - quantitative
5. **Mark - Line**: Tracing the active cases for a given region as the data is temporal - continuous

Insights: The peak seemed to have passed for Asia, with the number of active cases coming down considerably and consistently for 20 days. However, Europe has started and will be climbing even higher. Americas is not affected much as of now, but if it is poorly managed, it would experience a huge number of COVID-19 infection. Oceania might have escaped almost without harm even though it is so near to asia.
Conclusion: Good if you are in Asia as the peak of infection seemed to be over and should refrain from travelling to Europe.

Q5: How different parts of the world are getting affected?

Direct Link: https://eric-han.com/covid-19/q5.html



Category: Spatial aspects
Visual Encoding:

1. **GeoSpatial Position (x,y)**: World map position of country in the world, would be able to tell how that region around the country is affected
2. **Magnitude, Color Saturation**: Number of cases - quantitative
3. **Identity, Mark - Country Shape**: Country - categorical

Insights: The most number of cases still reside inside China and the countries around China. Hoever, Europe is becoming the new hotspot for the virus. There seems to be another hotspot developing in the middle east from Iran. Americas and Africa still have lower numbers as of now
Conclusion: Stay away from China, Europe and Middle East and the areas connected to them. Best if you defer all your travel plans towards these regions effective immediately.

References

Our D3.js Graphs are derived from the following resources. Our code contains snippets from the following resources.

1. https://www.d3-graph-gallery.com/bubble.html
2. https://www.d3-graph-gallery.com/graph/choropleth_basic.html
3. https://www.d3-graph-gallery.com/graph/connectionmap_csv.html
4. https://www.d3-graph-gallery.com/graph/wordcloud_size.html
5. http://jsbin.com/jalex/1/edit?js,output
6. https://www.d3-graph-gallery.com/graph/stackedarea_template.html