DSC 465 Data Visualization

Final report

Outbreak

James Valles

Kriti Srivastava

Dufang Qu

Frank DeRango

Hima Spandana Barla

Shadhana Palaniswami

Introduction

In early 2020, the COVID-19 virus, a coronavirus type disease previously-unknown, spread at an exponential rate globally. The virus, which appears to have originated in Wuhan, a Chinese city with a population over 11 million, has led to widespread and unprecedented disruption affecting not only the world's economy, but the lives of billions of people worldwide. As of March 15, 2020, most covid-19 confirmed cases are in China. Globally, according to BNO News, 157,305 cases have been confirmed along with 5,836 deaths - these numbers continue to rise. The mortality rate is estimated between 1% - 7%.

Our project is focused on creating visualizations that illustrate the evolution and spread of COVID-19 and its impact in China and on the U.S. economy during the first two months of the outbreak. Additionally, we used a variety of datasets to compare the latest coronavirus outbreak to others, such as SARS, MERS, Ebola, H1N1 (Swine Flu), all within the last decade. The stories we hope to convey to our audience through our visualizations include: the comparison between the spreading patterns of each outbreak, the COVID-19 death toll, recovery rate, and peak. We hope to give our audience, our classmates who are well-versed in visualization techniques, a look at how COVID-19 is evolving, now, while providing insight into past outbreaks. We hope the information will leave our audience with a greater sense of understanding outbreaks.

For this project, we used a number of datasets for our exploratory and explanatory visualizations, as outlined in our previous milestones. We had a limited number of COVID-19 datasets, due to it being a novel virus. As a result, we were forced to create our own. For the sake of keeping this final report below seven pages, we have only outlined Datasets used in the final visualizations, which include the following:

COVID-19 Death: This time series dataset provides a daily count for all patients who have died by location. It includes province/state, country/region, lat, long. Link to dataset: https://bit.ly/2TSNeXo

World Covid-19 Cases: (2020.1.22 - 2020.2.28, daily)

SARS: Source for SARS is the official website of WHO(https://www.who.int/csr/sars/country/2003_07_11/en/). It is the cumulative number of reported probable cases of SARS. The main focus was on the countries affected, cases count and reported date.

MERS: (January 2012 - Dec 2014). This dataset contains countries affected by Mers. Has many variables to describe the seriousness of the virus.The dataset is obtained from https://www.nature.com/articles/s41597-019-0330-0

EBOLA: (August 2014 - March 2016 monthly)The data set contains the details of confirmed cases and deaths.The data set was obtained from https://data.world/hdx/0d089fa0-3567-4b01-9c03-39d340ff34e3.

H1N1:(2009 - weekly**)** This dataset comes from the report of WHO official website. The dataset contains the countries that were affected by HiNi in 2009 along with the confirmed cases and death counts on the reported date. The data was reported weekly and it is froin the following link:
www.kaggle.com/de5d5fe61fcaa6ad7a66/pandemic-2009-h1n1-swine-flu-influenza-a-dataset

China PMI: (2019.1 - 2020.2, monthly) This dataset is obtained from the official website of the National Bureau of Statistics of China. PMI stands for purchasing manager's index, which is a diffusion index of the prevailing direction of economic trends. This dataset contains date variable and 25 indices, covering the manufacturing and service sectors.

China Air Passenger Traffic: (2019.1 - 2020.1, monthly) This dataset is obtained from CEIC website. It contains China air passenger traffic year over year growth for a monthly basis.

Wuhan Air Quality Index: (2019.1 - 2020.2, daily) Wuhan is the city having the most confirmed cases, and it encountered a strict shutdown for almost two months. This dataset is obtained from www.aqistudy.cn, containing a date variable and air quality index, and variable air quality level is created based on air quality index.

NASDAQ, Dow, S&P 500, SSE Indexes: (Dec 31, 2020 - March 10, 2020) Created four datasets from scratch, using stock data from Yahoo News. Variables used in visualization: Date and Close (closing price) .

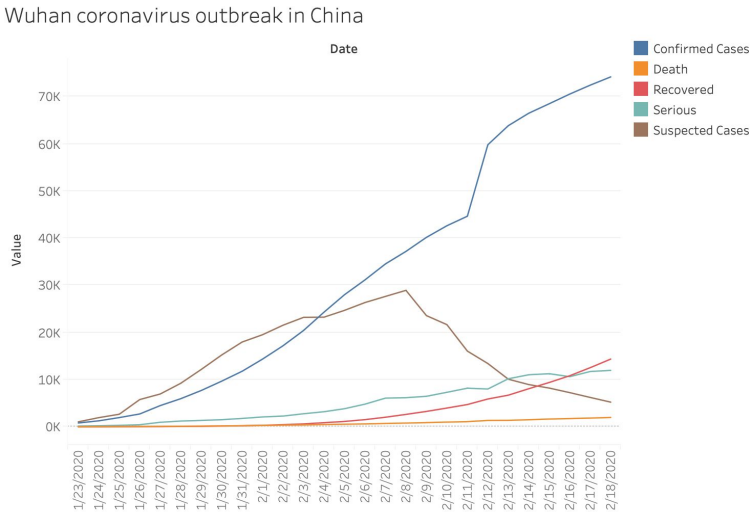
The variables we used in our visualizations include countries, recorded dates, the number of confirmed cases for each disease, number of patients who have recovered or died for each disease, stock market indices including SSE, NASDAQ, Dow Jones Industrial Average, and S&P 500. Furthermore, we used the China Purchasing Manager Index, China air passenger traffic data (year over year growth rate), and Wuhan Air Quality Index.

Coronavirus Tests Per One Million People: This dataset is obtained from newyorktimes https://www.nytimes.com/interactive/2020/03/17/us/coronavirus-testing-data.html

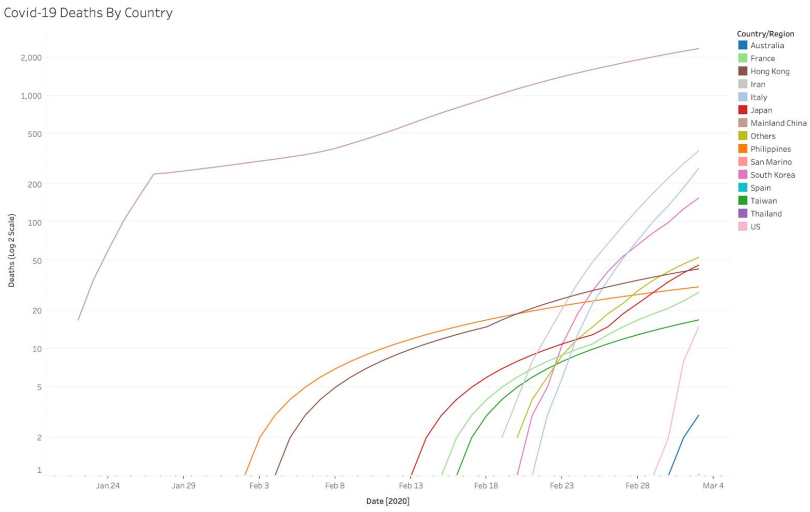
Exploratory Analysis

During data exploration, each of our group members chose a topic which he/she thought is worth studying. We then conducted the exploratory analysis. For the sake of keeping this report to a minimum, we have only attached seven exploratory visualizations. The folder containing R/Tableau workbooks and submitted milestones include dozens more.

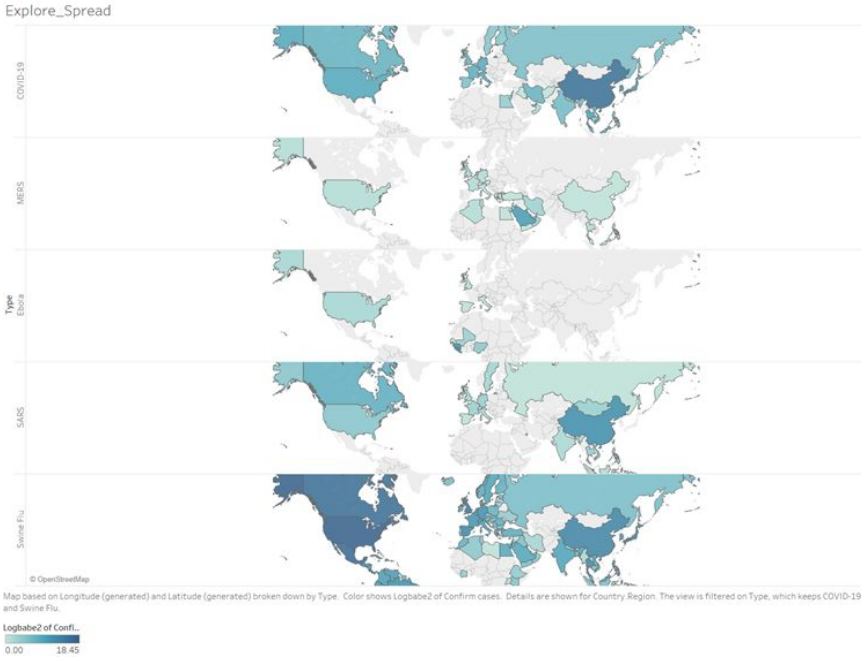
Wuhan Coronavirus Outbreak: This time-series line graph plots the number of confirmed, suspected, and serious COVID-19 cases in Wuhan, China along with the number of patients who have recovered by date. We started gathering exploratory data at the beginning of the outbreak beginning on 1/23/2020. It is very interesting to note the number of suspect cases dropped around Feb. 8th. This was the time the Chinese government changed the definition of a suspected case.



COVID-19 Deaths: The following is a line graph that shows COVID-19 deaths by country as separated using a 20-color palette in Tableau. Used a log scale to add China deaths to the graph, which are significantly greater than other countries. This graph illustrates when each country starts to record deaths and its general trend. Color is used to separate each country. Date is plotted on x-axis, while number of deaths on the y-axis. I found that this visualization was hard to decode and therefore opted to go with a bar chart and interactive dashboard. There are too many lines. But, it does show when countries started to report deaths which is nice.

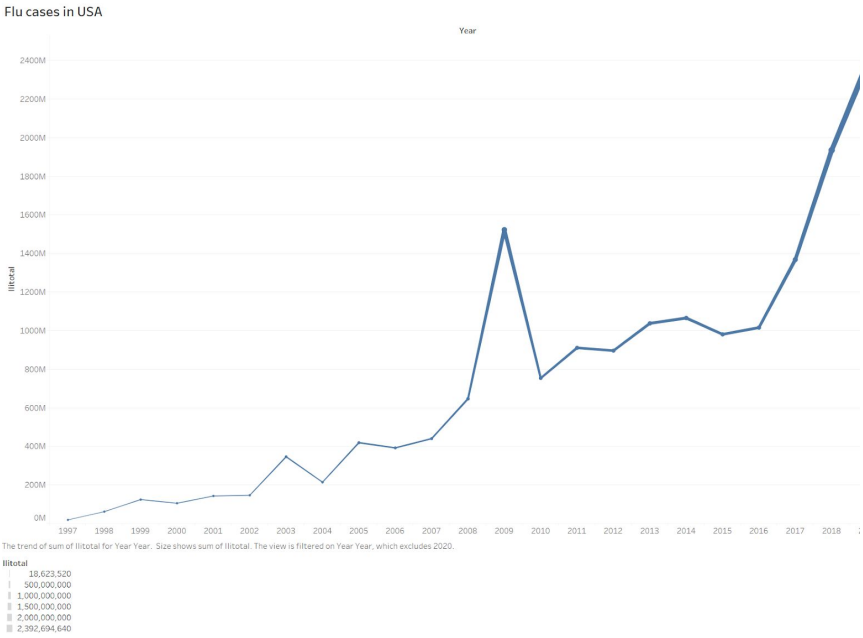


Explore Spread: This graph is the exploratory graph to show countries that were affected by each of the five viruses: SARS, MERS, Ebola, H1N1 (Swine Flu), COVID-19. Each map in a row shows the spread of one of the virus.The continuous colour palette is used to represent the number of confirmed cases by country. Darker the colour, more the confirmed cases.

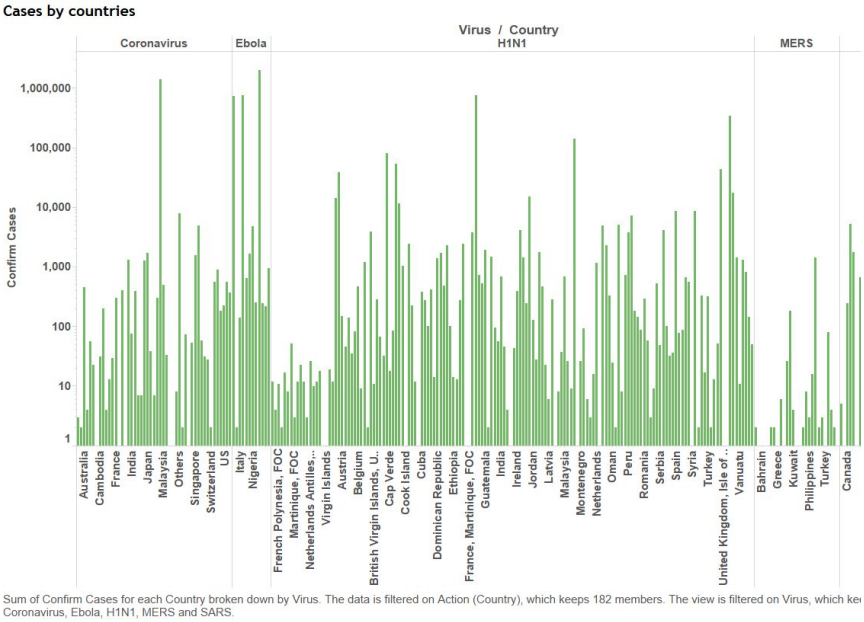


Seasonal Flu Cases in the USA: This line graph is created to represent the number of flu cases in the USA from 1997 to 7th week of 2020. This graph was created to look for some interesting pattern or information. The y axis shows the total flu like cases which includes all the cases caused by any TypeA or TypeB viruses.

The x-axis shows the years from 1997 to 7th week of 2020. The width of the line increases as the number of cases increases. We can see the increasing trend in the number of cases through time. We can also see a prominent spike in 2009. This spike is due to the increased cases due to the H1N1 Swine flu virus.

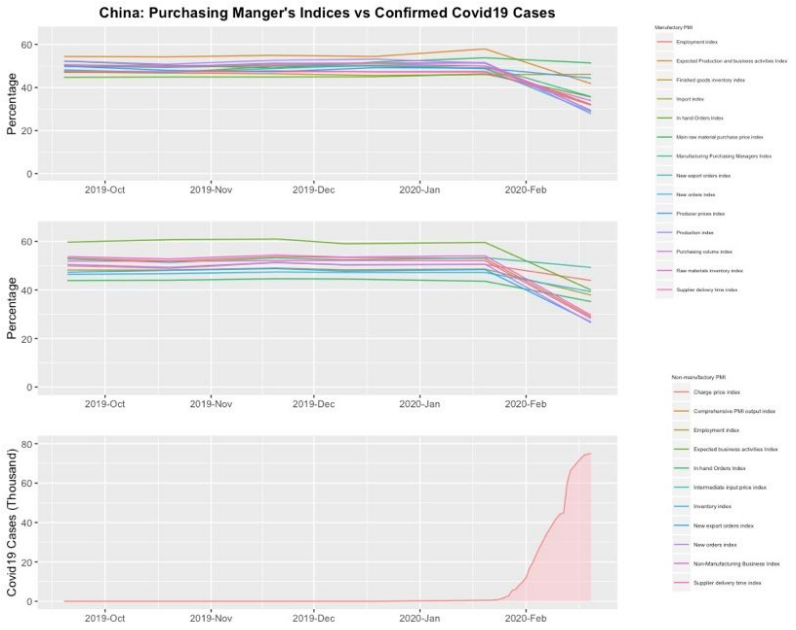


Confirmed cases by countries: The bar graph is to show the number of cases of each virus(Coronavirus, Ebola, H1N1, MERS and SARS) by countries. The x-axis shows the categorical value "Countries" and the y-axis shows the total confirmed cases of a virus. Since there was a huge difference between the minimum number of cases to the maximum number of the cases, which makes it difficult to show a bar for the countries with very few cases on normal y-axis scaling, therefore exponential scaling on the y-axis.



Covid-19 Impact in China: We produced line graphs for all 24 indices. Although lines cluttered together, they proved there was a strong relationship between the trend and the outbreak of Covid-19.

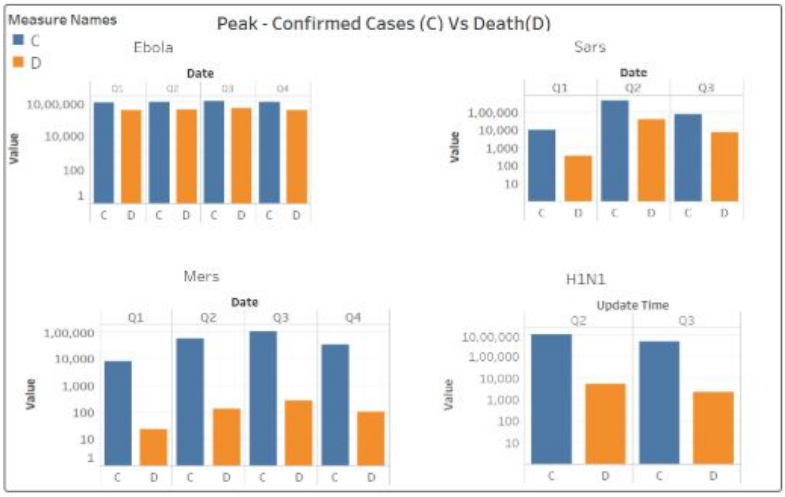
The original datasets contain more than thirty indices. Only one key index for each group was kept in the final visualization.

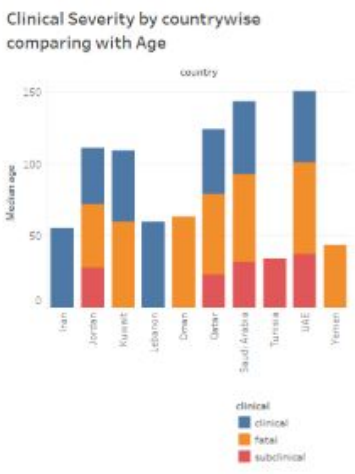


Peak of Confirmed and Death Cases reported by H1N1, Ebola, Mers and Sars

This slide explains the number of confirmed cases and deaths reported by all the viruses. The death cases are reported by D and confirmed cases reported are represented by C

The peak is shown by plotting a bar graph using the 4 different datasets of Ebola, H1N1, Mers and Sars. When comparing the data the number of confirmed cases are more than the deaths reported in each quarter of the respective year.



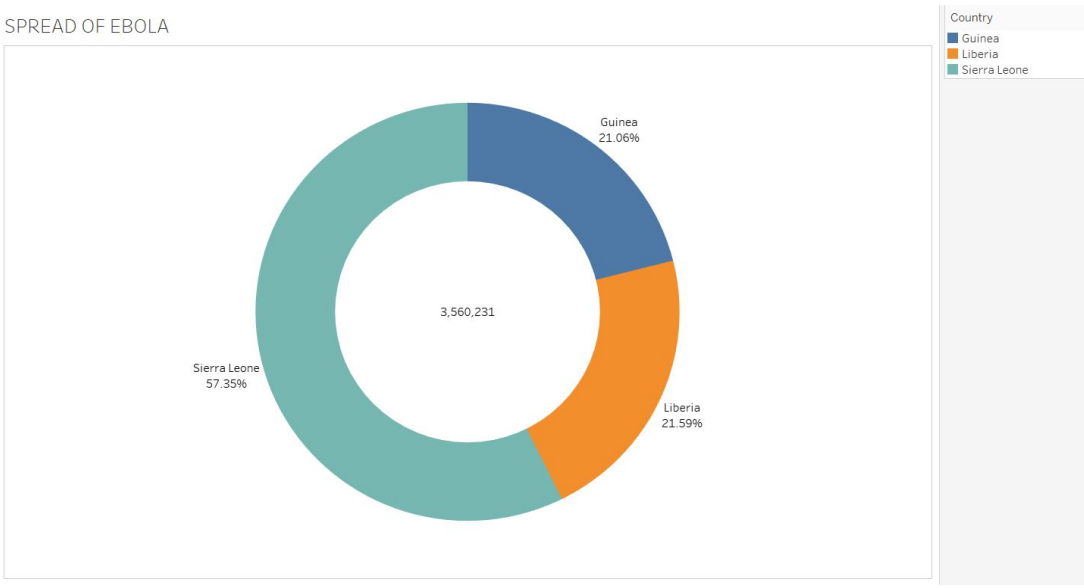


Mers Dataset and its explanatory visualization

The above plot has three different sets of information, plotted using Mers dataset.Plotted a geographical plot to show the Gulf countries are affected by Middle East Respiratory Syndrome.Clinical severity graph plotted explains the level of severity measured in all GCC countries. Saudi Arabia

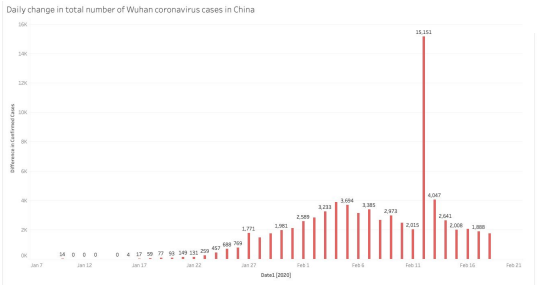
and UAE being the most affected countries and Tunisia being the least.Third graph plotted in between animal contact and Death rate. This shows that TRUE death cases reported by animal contact are less when compared to FALSE death cases, that is it is without animal contact from the year 2012 to next two consecutive years.

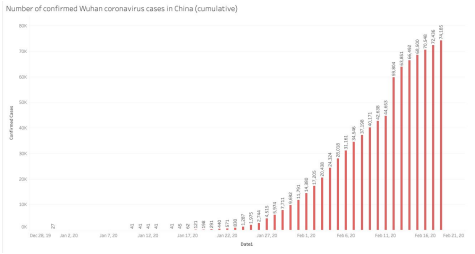
EBOLA SPREAD:

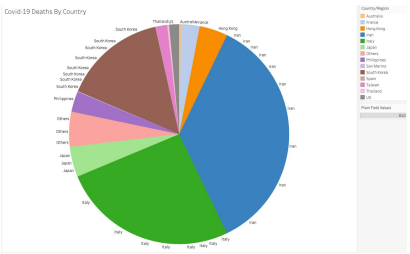


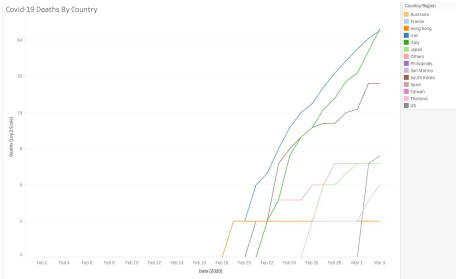
The spread of ebola in three different regions Guinea,Liberia,Sierra Leone was visualized using the donut visualization .Here we can notice that the Sierra Leone region is the most affected region which accounts for 57.35 % of the total cases reported .Liberia and Guinea appears to have an equal share of the number of confirmed ebola cases each accounting for almost 21% of the confirmed cases .This epidemic occurred around end of 2014 to the mid of 2016 .This spread and effect of the ebola was studied to compare its effect and severity with the present pandemic Corona.

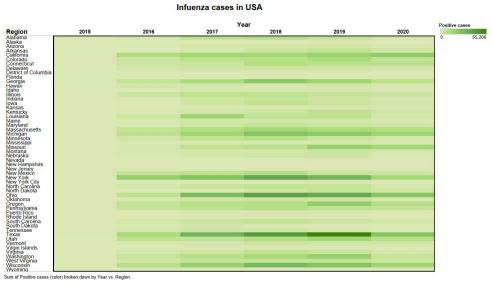
In addition to the samples featured above, we also tried several other exploratory visualization techniques throughout this process to make better sense of our data. Here are a few more samples. See our extra explanatory or exploratory graphs in the appendix section.







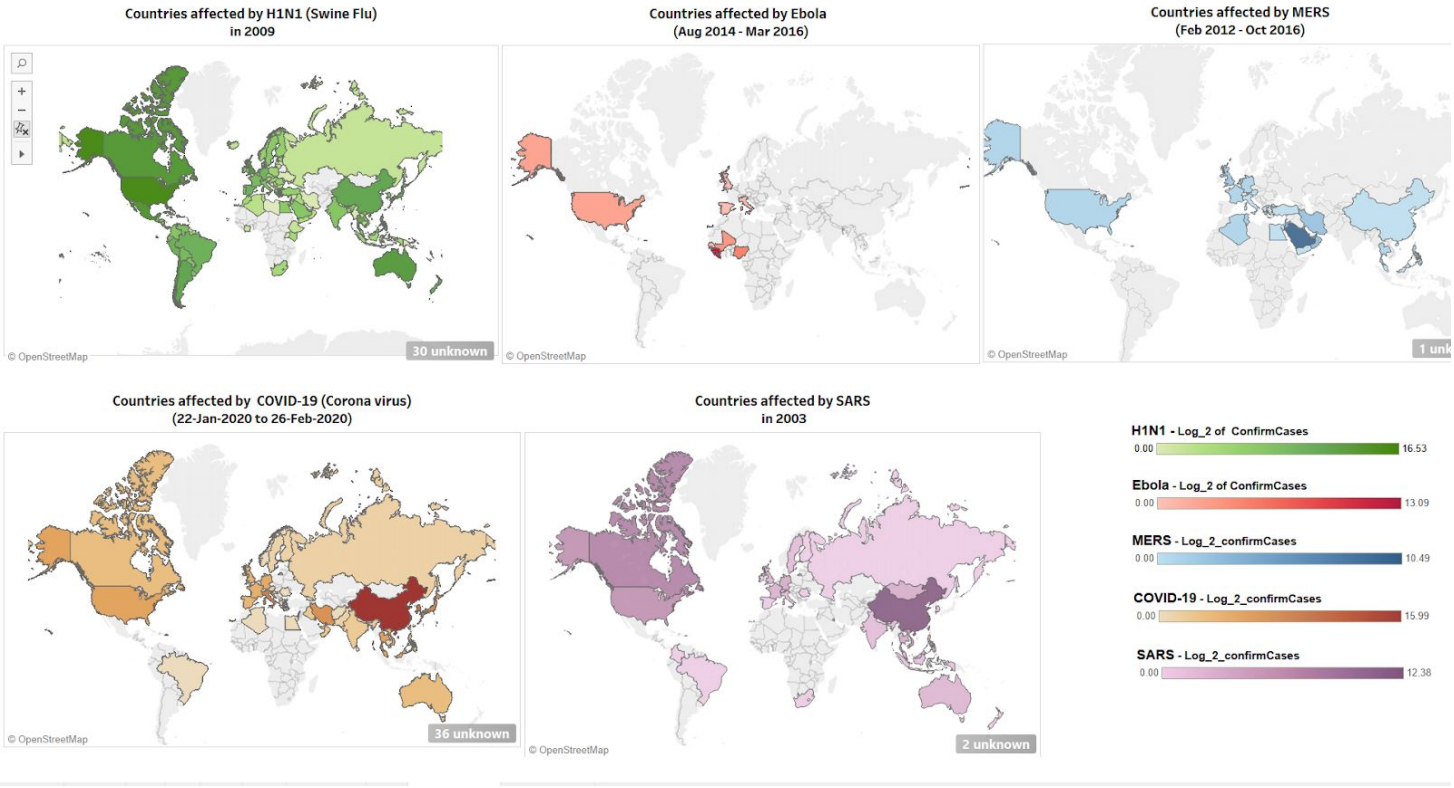




Explanatory: Visualizations and Analysis

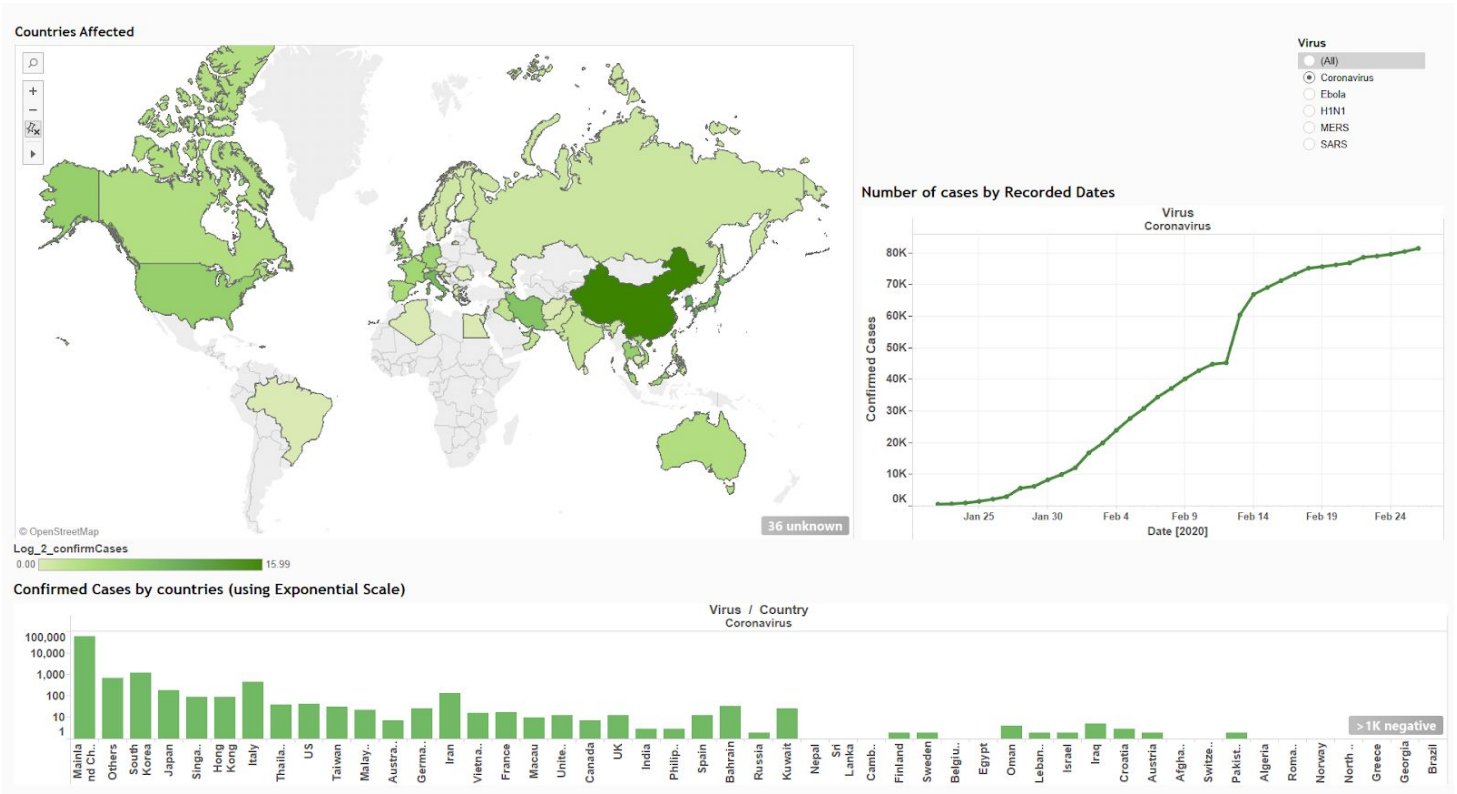
We have included the following 7 visualizations for grading.

Visualization 1: Geographical Spread



The above image shows the countries affected by the virus. The image has five world maps, in five different colors, each representing the countries affected by each virus. The colors get darker as the cases increase.
The map with a continuous palette of green shows the geographic spread of the H1N1 virus in 2009.
The map with a continuous palette of pink-red shows the geographic spread of Ebola virus from Aug-2014 to March-2016.
The map with a continuous palette blue shows the geographic spread of MERS virus from Feb 2012 to Oct 2016.
The map with a continuous palette of orange-red shows the geographic spread of COVID-19 (coronavirus) virus from 22 Jan 2020 to 26 Feb 2020.
The map with a continuous palette of purple shows the geographic spread of SARS in the year of 2003.
Since, the color distribution is linear but our data is skewed so most of the data was in a smaller color range. Therefore, I intentionally choose Log2 of Confirmed Cases of their corresponding viruses in all of the maps to show more variations within the color hue.
Observations: If we look at the image, we can see that the USA was most affected by H1N1, then Mexico and Canada. Likewise, Sierra Leone was the most affected country followed by Liberia and Guinea. Saudi Arabia was severely affected by MERS. Mainland China is the most affected country from COVID-19 and SARS.We can also observe that the USA is affected by all five viruses followed by China with four. This can be because people travel the most from these countries as compared to all other countries.

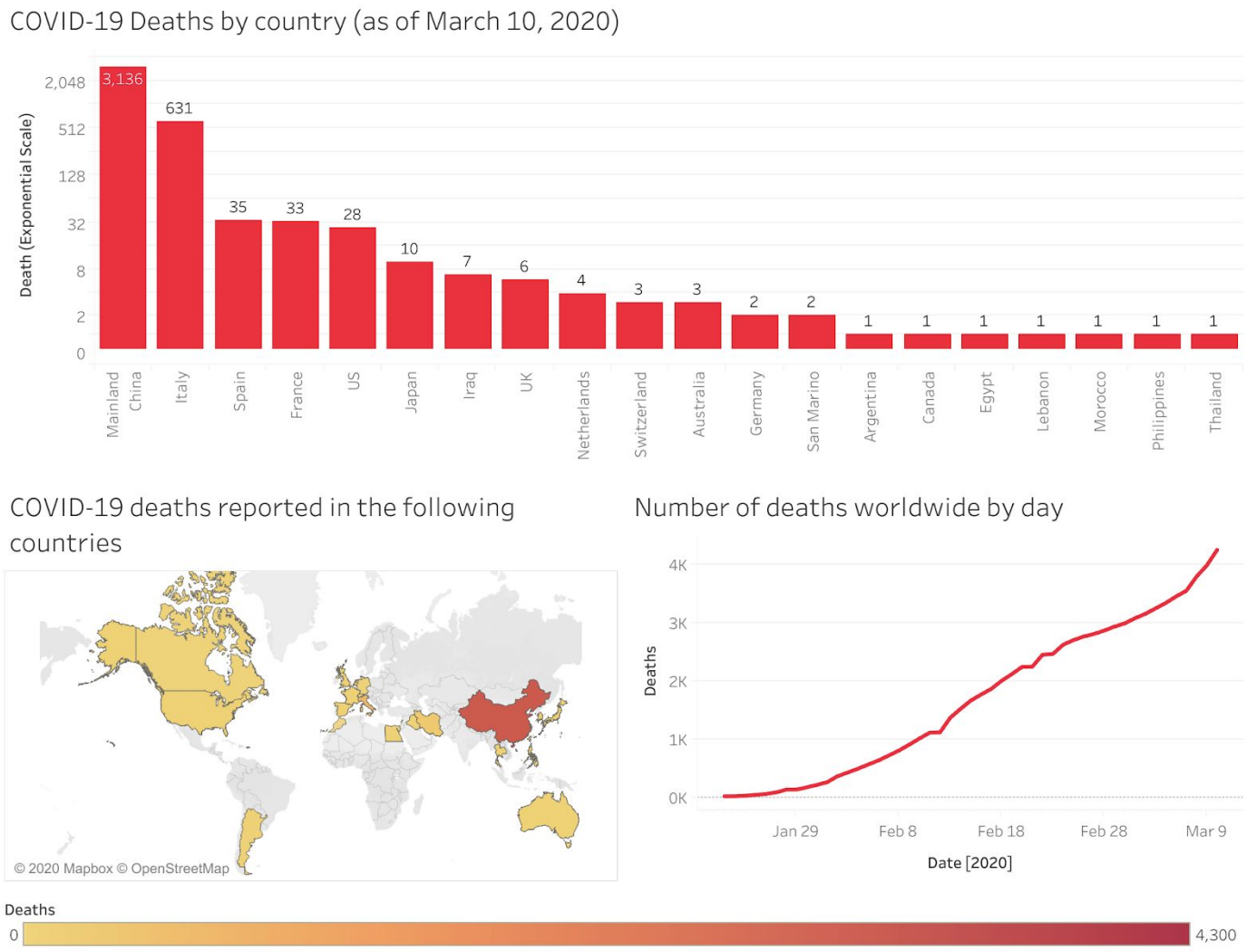
Visualization 2: Dashboard Snapshot - Virus Evolution



https://public.tableau.com/profile/kriti6381#!/vizhome/Outbreak_15843179001170/Dashboard2?publish=yes

The dashboard is giving the information about the selected virus from the aspect of geographic, time and count. The radio button is used to filter all three graphs by selected virus which aims to focus the interest on one virus and look for details by selected virus. The geographic map is used to give the visual representation of the virus spread and the continuous color palette gives a sense of the magnitude of confirmed cases by countries. Again, Log2 of Confirmed Cases of viruses is used to show more variations within the color. The line graph is to show the change in the confirmed cases over the time for the selected virus. The x-axis is the date on which the confirmed case was recorded and the y-axis shows the number of confirmed cases. The bar graph shows the confirmed cases with respect to the countries in decreasing order for the selected virus.The x-axis shows affected countries. It is need to be noted that the y-axis has an exponential scale. It is done to make the bar line visible even for countries with very low count because the data at the lower end of the scale were not visible when the normal scale was used.

Visualization 3: Dashboard Snapshot - COVID-19 Deaths by Country

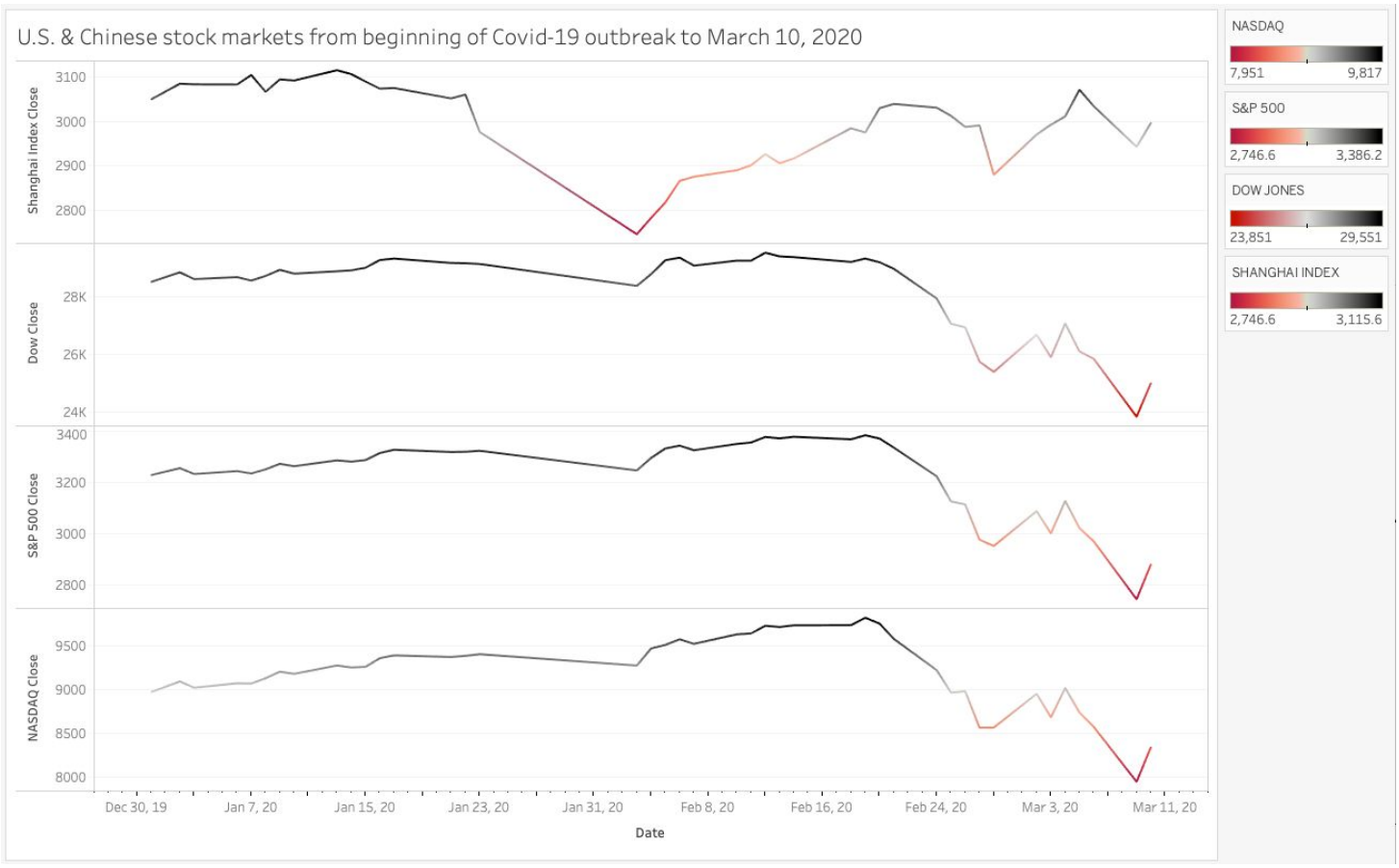


https://public.tableau.com/profile/james.valles#!/vizhome/COVID-19DeathsByCountasofMarch102020/Dashboard1?publish=yes

This dashboard, created in Tableau, provides an interactive view of the number of lives lost globally due to COVID-19. It features three charts. Among them is a bar chart of countries sorted in order from highest to the lowest number of deaths. The y-axis is on an exponential scale, because of China's significant death toll. This allows countries below 35 deaths to be represented. The color red represents death and the sense of urgency in the time of an emergency. This visualization was created using Tableau. I had to pivot my dataset to make this work. The bar chart uses length encoding; therefore, it is necessary 0 was included. Overall, the bar chart makes it easy to decode which country has the highest number of deaths. As of March 10, 2020, China had the highest number of reported deaths, followed by Italy, Spain, France, and the U.S. respectively. The second graph in the interactive dashboard is the Choropleth map, which shows which countries have reported at least one death as a result of COVID-19. A sequential color scale (orange and red) is used to distinguish which countries have reported the highest number of deaths (red represents the most while orange represents the least). It is clear in this map that China has the highest number of deaths. When you zoom in you are able to see Italy and Iran also have a high number of deaths. While the rest of the world has the same color as they have a death toll less than

30 people. Making this chart interactive allows the user to obtain the actual death toll for each country. Similarly, the line graph, without interactivity, illustrates the world's COVID-19 death rate has grown exponentially since the beginning. By using a time-series line graph, a pattern is established as to the rate of growth. Interactivity allows the user to drill the actual total death toll by day.

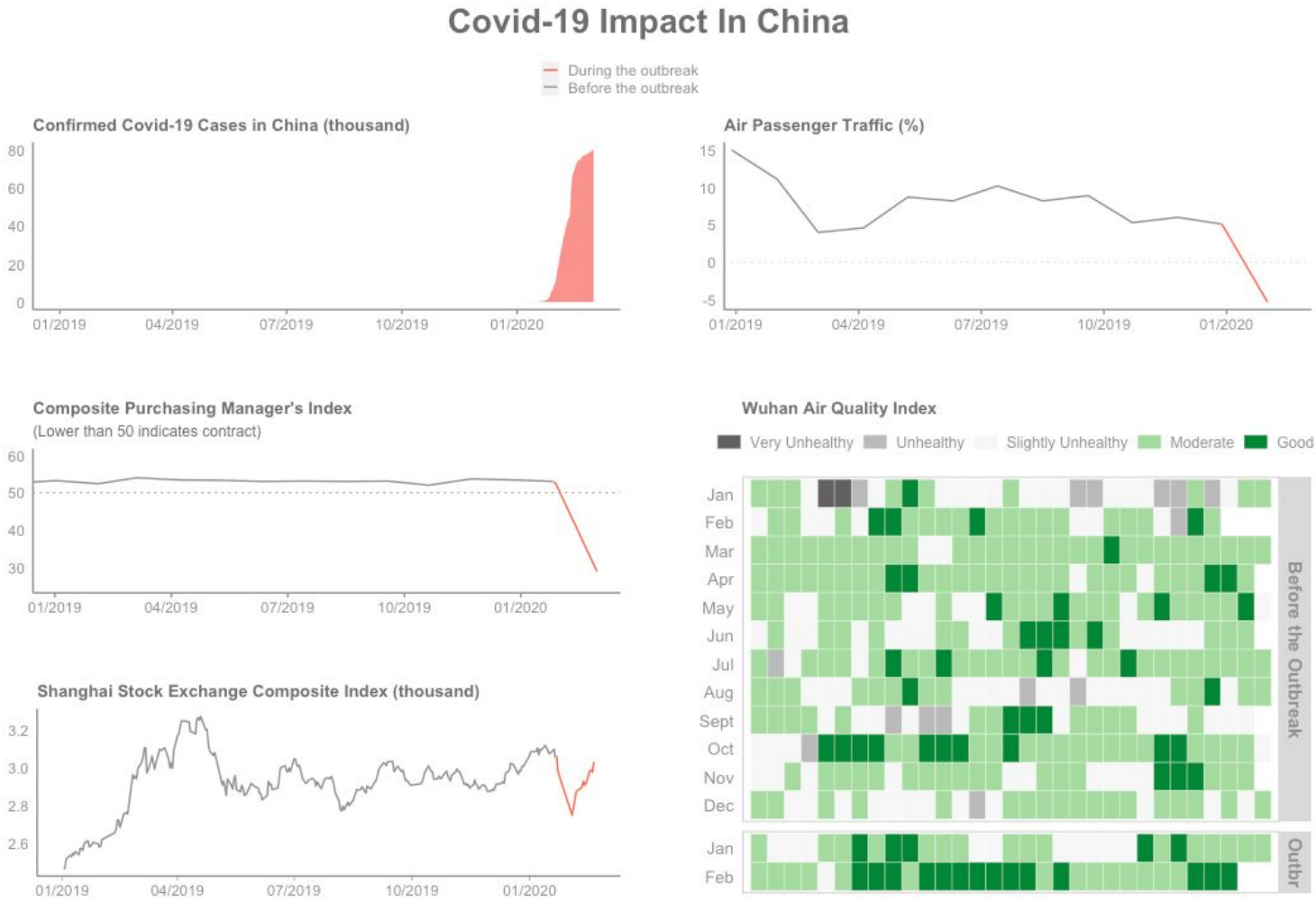
Visualization 4: COVID-19 Impact on U.S. & Chinese Stock Markets



The time-series line graph, created in Tableau, shows the impact COVID-19 has had on the U.S. & Chinese Stock Markets. The x-axis is the exact date. The y-axis is the closing price for NASDAQ, S&P 500, Dow Jones Industrial Average, and the Shanghai Composite Index. The aligned base line in this graph is the date. The date ranges from Dec 30, 2019 through March 11, 2020. Dec 30th is when the COVID-19 outbreak began to take shape in much of China, specifically in the city of Wuhan. A diverging color scale was used to show losses. Originally, I had selected green and red to show % gains and losses respectively, but per feedback, I toned it down and changed it to showing the actual close value instead of percent gains or losses. As per feedback, I didn't want there to be a double encoding issue. The red color encoding allows our audience to easily decode where the markets showed losses. Around January 23, the Shanghai Composite index takes a steep drop that lasts until Feb 1st before it begins to make gains. This is right around the time cases in China became more prominent and lockdowns began taking place. In the U.S., you see as the number of cases in China increase spreading begins globally the U.S. markets begin to drop. On Feb 24, all three U.S. indexes, NASDAQ, S&P 500, and Dow began to post losses. Interestingly enough, while the U.S. market plunges, the Shanghai index appears to show some recovery. Around this time in March, China was reporting that the number of cases had slowed, and

that businesses were beginning to reopen. This chart makes it easy to compare all four indexes at any given time. It also reveals if there are any correlations between the U.S, & Chinese markets.

Visualization 5: Covid-19 Impact in China



This visualization is built in R, covering five datasets: China covid-19 confirmed cases, PMI data, SSE data, Air Traffic data and Wuhan AQI data, and including a graph panel and a heatmap. The graph panel, consisting of an area graph and three line graphs with the same date scale on the x axis, relates the increase of the confirmed Covid-19 cases to the trends of the other three indices. As for the heatmap, days and months in the Wuhan AQI dataset are extracted as scales on its x and y axes respectively, and air quality levels are coded into colors filling each day.

Color is carefully chosen. For all the line graphs, red color is used for data during the Covid-19 outbreak, distinguishing from the gray line for data of dates before the outbreak. For the heatmap, diverging colors from dark gray to green are used to encode five air quality levels from very unhealthy to good.

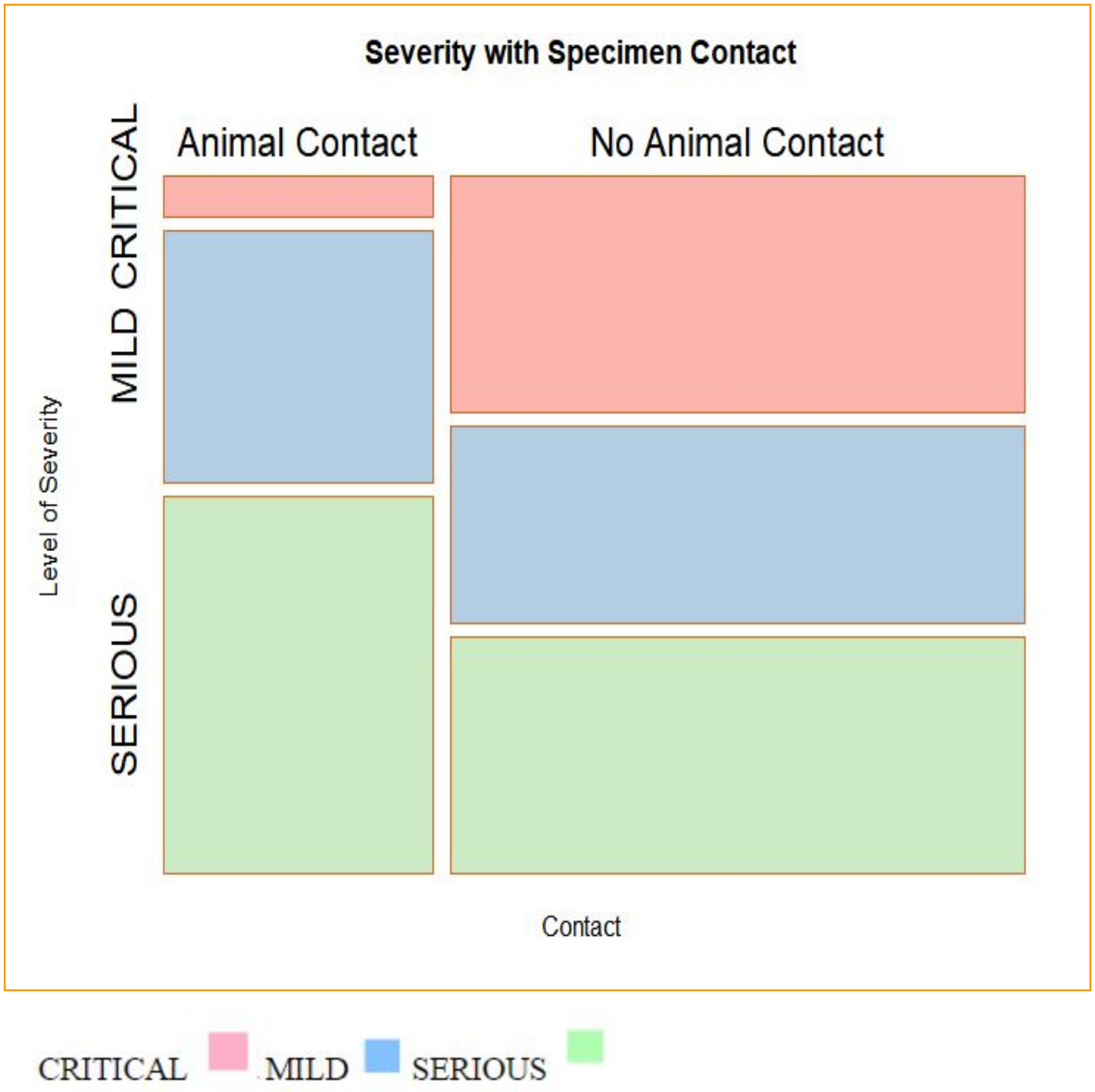
The range of scale is tricky for this dataset. The outbreak started at the Spring Festival. Air traffic, stock market and air quality could possibly show different patterns during this time even without the outbreak. A date scale from Jan, 2019 to Feb, 2020 was chosen so that we can compare the data during the outbreak not only with the data from the previous several months, but also with the data from the previous Spring Festival. Although the line graph of Covid-19 shows 0 cases throughout 2019 but a big rise in 2020, along with other graphs in the

panel, how things have changed when the outbreak started is obvious. Besides that, in the heatmap, months on y axis are grouped into two sections, before outbreak and outbreak for a better comparison.

This visualization gives a comparatively overall view about Covid-19 impact for a country. Purchasing Manager's Index is a commonly-used economic indicator, usually released at the beginning of a month. PMI values lower than 50 indicate contract. The visualization shows a sharp plunge of PMI in Feb, 2020, from a previous consistent expanding rate about 53 to 29. Similarly, SSE index and Air Passenger Traffic also dropped greatly in Jan and Feb, 2020. While the heatmap shows the epicenter, Wuhan, had the most days with good air quality in Feb, 2020. Along with the graph showing the growth of confirmed Covid-19 cases, we can see all these dramatic changes happened during the Covid-19 outbreak.

This visualization fits our analysis perfectly. The outbreaks outside have just started while China has almost gone through a full epidemic cycle. The damaging impact in China we've already observed probably will be an epitome of the Covid-19 impact in the whole world.

Visualization 6: Virus Severity with or without Specimen Contact



Health care providers contacted the local/state health department to notify the number of patients with fever and lower respiratory illness who they suspect may have got the viruses. Clinical specimens were collected for testing respiratory pathogens at either clinical or public health labs.

The above mosaic plot showing the virus severity with specimen contact is built in R. There are certain parts of the world, where the virus may have been spread through handling and consumption of wild animal meat or hunted wild animals infected with Virus. The visualization plotted here is a cumulative data obtained for all the viruses.The viruses which we included in our project as mentioned in the introduction are MERS, SARS, Ebola, Corona and H1N1.

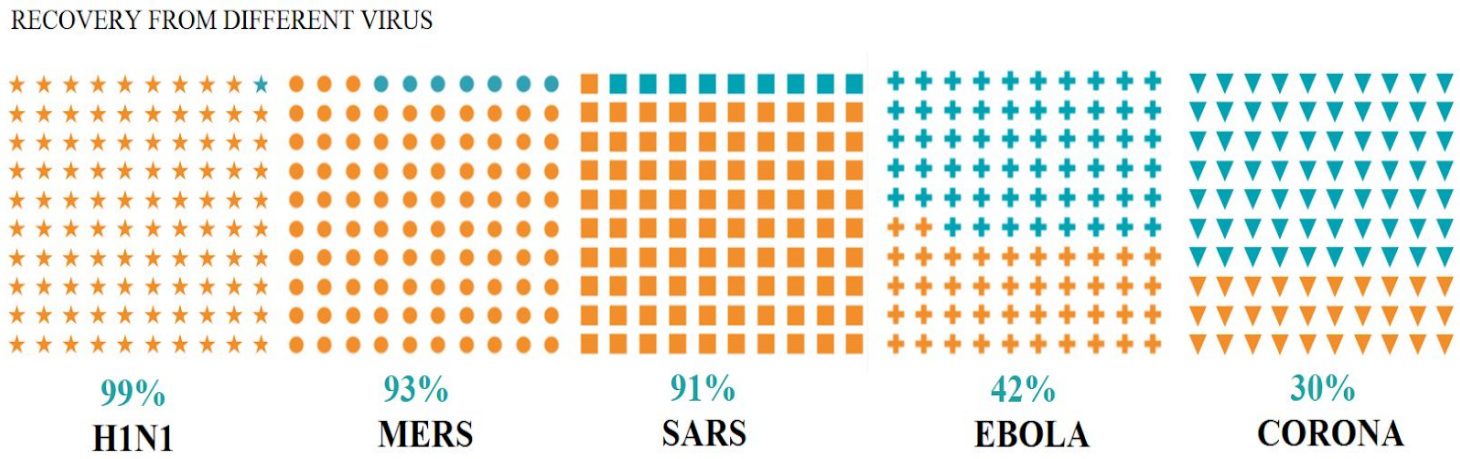
The width of the boxes are proportional to the percentage of animal contact or no animal contact, respectively. In fact, 12% of cases were reported as Critical, 13% were reported as Mild cases and 12% cases were Serious with No Animal Contact.

The height of the boxes are proportional to percentage of level of severity cases admitted.In fact, 1% were reported as Critical cases, 3% are Mild cases and 9% of cases were reported as Severe with Animal Contact.This seems to show a large contrast in the level of severity.

To make the plot easier to interpret, the boxes for level of severity are colored with peech, baby blue and mint respectively. The severity level is arranged in the alphabetical order. The code used to plot the mosaic plot is in appendices.

I tried arranging the severity level in an order, but R was not executing the order syntax.

Visualization 7: Recovery of Different Viruses



Recovery is the main factor of concern when it comes to check the severity of a disease .Higher the recovery rate lesser the severity of the disease .Here waffle plot was used to show the recovery from different viruses .A Waffle chart is essentially a squarified pie. Since it does not involve angles, it's easier for the reader to compare accurately as well as evoke emotional comparisons.A Waffle chart is a square divided by 10×10 cells. The value is displayed as a percentage, so you can clearly see the difference down to 1%.This helps us to convey the message to the audience in a more easy manner .A reader can quickly and easily be informed that a single filled square equals 1% and that a filled row/column equals 10%. By squarifying and gridding pie charts, significant reading accuracy can be gained, and none of the simplicity, accessibility, or scalability of the traditional pie is

lost.Waffle charts can be used in lieu of traditional bar charts. They're particularly effective when comparing numbers that are highly variant, which makes them easily tolerant of outliers.Tableau was used here for building a waffle plot to represent the rate of recovery for different viruses The data set containing the details of confirmed deaths and cases was used to build this plot .From the visualization above we can see that H1N1 has the highest recovery rate when compared to the viruses ,this means that this virus is the least severe when compared to the other viruses .We can notice an increase in severity of the viruses as we move to the other waffle plot .Corona being the most severe with the recovery rate as low as 30%.This as of 17 march 2020.The virus causing ebola being the second most severe of all with as recovery rate of 42%.Also different shapes was used to represent different viruses which makes the visualization more clear and the audience will not have to spend so much time understanding what the visualization is trying to communicate the message we intend for the audience to acknowledge.Hence from the above plot we can say that Corona has been the most severe disease so far and no wonder it has been declared a pandemic by the World Health Organization.

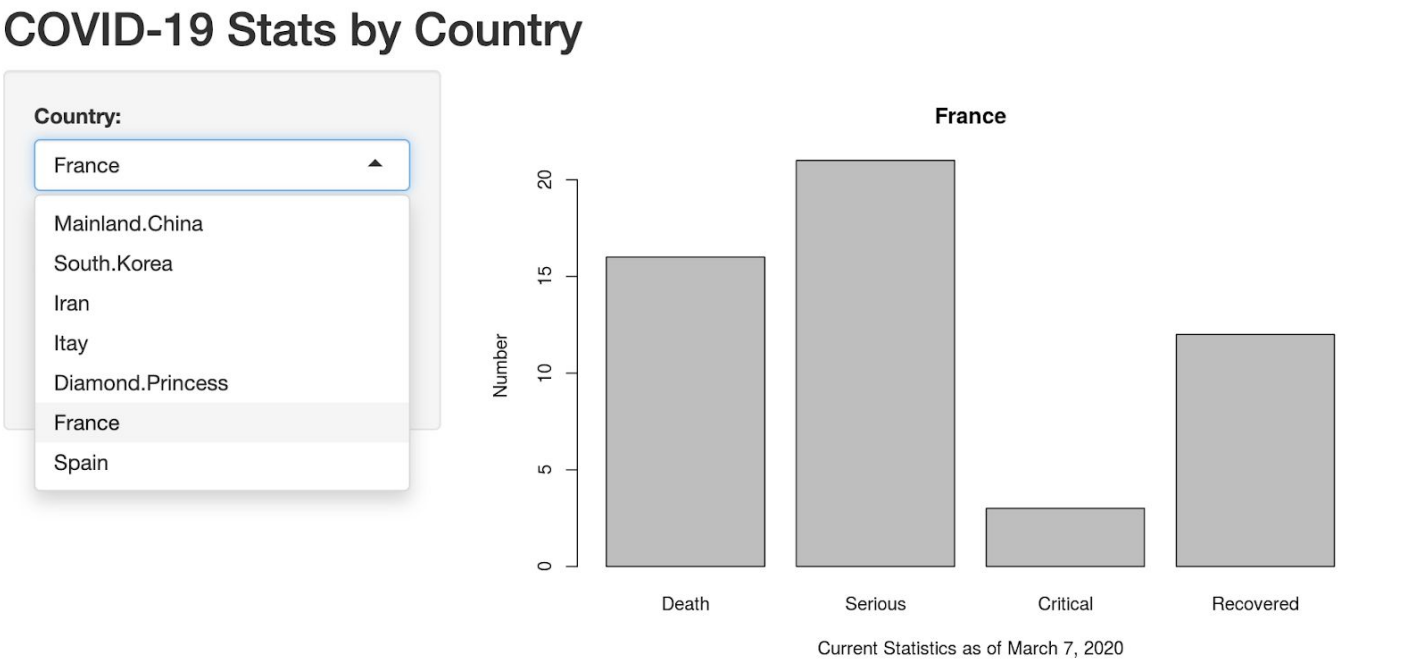
Conclusion

As the COVID-19 virus continues to evolve its full effects are unknown. Based on our visualizations, we can conclude a few things - it's much more infectious than SARS, MERS, Ebola, H1N1 (Swine Flu). So far, our geographical spread visualization illustrates how COVID-19 has affected almost all countries on every continent of the globe within its first 30 days. Furthermore, the number of deaths appear to be growing exponentially with several countries outside of China reporting deaths. We will need more time to truly understand its death rate. However, the impacts it is having on the economy is unprecedented. Stock markets have continued to take a beating after March10th. Our graphs began to show this downward trend. As a group, we had an awesome time practicing many of the techniques we learned in class all while exploring a topic that is altering our lives in unprecedented ways.

As COVID-19 continues to spread outside of China, there are a number of other stories our team would like to explore such as the relationship between the spreading patterns and geographic characteristics, including city locations, city population densities, local average temperatures, etc. along with countermeasures and policies governments enforced to contain the virus.

Extra Credit #1 - James Valles

Interactive Barchart using Shiny/R Studio



https://jamesvalles.shinyapps.io/Corona-19Stats/

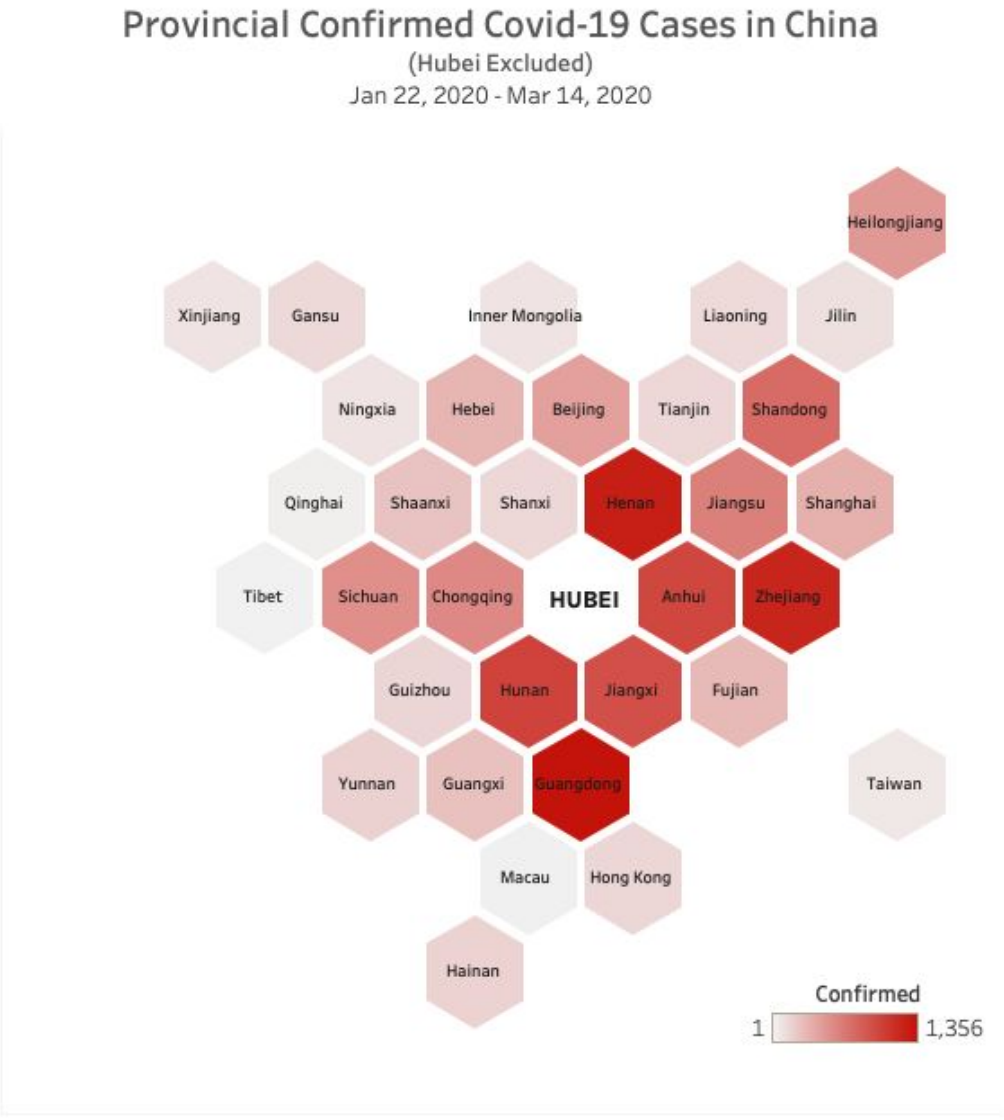
This interactive bar chart is a Shiny application using RStudio which details COVID-19 stats for all countries where cases have been reported as of March 7, 2020. It shows the breakdown of deaths, recoveries, and series/critical cases. There are visualizations for over 100 countries.This is a nice way to provide our audience with a quick way to get stats in the form of a visualization. While Shiny was mentioned in class, it was nice to run through the entire process of creating the app, altering the data, and deploying the app to a server.

Extra Credit #2 - Dufang Qu

Cartogram (built in Tableau)

This tile cartogram is built in Tableau with two datasets: a manually-made China province coordinates dataset, and Covid-19 outbreak dataset (Jan 22, 2020 - March 12, 2020). A new variable death rate by province is calculated based on aggregated deaths and confirmed cases.

The purpose of this visualization is to discover geographic features of the Covid-19 outbreak in China. Hubei Province, as many have already known, is the epicenter of this outbreak as well as has the most confirmed cases and death toll. Its confirmed cases account for 83% of all confirmed covid-19 cases in China. Therefore, Hubei was filtered out of this tile graph to avoid the differences between the encoded sequential colors of other provinces becoming too small to be seen.



From the cartogram, we can clearly see that geographic contiguity and the direction of population migration are two possible factors affecting the spreading of the virus. The adjacent provinces all have more confirmed cases compared to provinces in the same direction but further away from Hubei. The provinces to the northeast, east and south of Hubei

province have the most confirmed cases among the rest of the provinces. While this is not true for the provinces to the west and northwest of Hubei. This probably is because people in Hubei travel more frequently towards the coastal area.

Appendices

Individual report (James Valles)

For this project, I served as my group's team leader. I was the one who submitted all of the milestones and final project and served as communication between the professor and team. I set up the Slack chatroom, Google Docs/Drive for team collaboration. **Final project's explanatory visualizations include**: I worked solely on the COVID-19 Death Dashboard, line graph of COVID-19 on U.S. and China Stock Markets, Shiny Application, which provides COVID-19 stats for over 100 countries detailing the number of deaths, critical/serious cases, along with recoveries. I also created and consolidated four data sets on SSE, Dow, NASDAQ, S&P 500 stock index. I create three additional explanatory visualizations: Overall cases/death by country (Donut graph), Interactive U.S. COVID-19 case map. **Exploratory visualizations include:** I created 11 exploratory visualizations, and for the Daily Change in Total COVID-19 Cases, Confirmed Cases Wuhan, COVID-19 Death in China, COVID-19 Suspected Cases, COVID-19 Recovered Cases in Wuhan, Tracking Confirmed Cases, Death, Serious, Recovered on one chart, Worldwide Suspected COVID-19 cases, COVID-19 Death by Country, SSE Stock Index.

This topic is not only timely, but it also made me more passionate about exploring and visualizing data. There couldn't be a better time. I have learned many valuable lessons along the way including the importance of consistency, proper labeling, making sure my visualizations are clear and free of clutter, the importance of using color correctly, and the different ways I can encode data to convey the story. Most importantly, I learned each visualization begins with data, then understanding my audience, and finally crafting my message through visualization iteratively. While we had less than 7 weeks to put this project together on top of other obligations and assignments, it gave me much practice in working with the different tools such as Tableau and R. While it hasn't always been easy, what I know for sure is I am much more comfortable with the process of creating a visualization and have learned there are many things I must consider along the way. Before starting this class, I thought visualization was much like art, but now I understand there is much more to crafting that picture.

I have learned it has much to do with mapping and perceptions. The many different ways to encode data symbolically. And, the effective ways to encode so that a human can decode; otherwise, the visualization fails. There are many ways to encode data, whether it's through position, length, color, size, or shape, and the list goes on. Some are more effective than others. In this project, I carefully had to choose my colors and had to consider using either a sequential or diverging color scale. I had to decide whether it was better to create a cooler, much more complex graph, or settle on one that appears more simplistic because it was more effective at telling the story. More complex and fancy isn't always better. From contingency plots, to violin plot, and on, there are so many visualizations to choose from, getting lost is easy. Knowing what to use for the task you're trying to complete is key. When it came to bar charts, which are best for categories (in this case countries) not only did I have to ensure that length could be decoded by including 0 to avoid misleading (Weber's Law), but I also had to consider order. For the COVID-19 effects on U.S. and China Stock Markets visualization, having an aligned based line, allowed me to compare four different stock markets by date allowing my audience to find patterns. These were just some of the things I considered - there were many more. I learned a great deal about time series in this class, how line graphs are continuous and how their connections have meaning.

With each visualization, I tried carefully to avoid adding extra ink (data to ink ratio) and definitely

avoided 3D visualizations and chart junk. Bad news! While there have been many moments I wondered if I was doing this correctly, I have come to understand that data visualization is a skill that gets better with practice. The more I do, the better I will become. I can say I have come a long way since starting this class and I am thankful for the project and homework. Aside from creating explanatory and exploratory visualizations, this project has also taught me a great deal about working within a group. Communication is so important. Juggling schedules, selecting the best visualizations to submit without hurting feelings were some challenges we all had to face because everyone truly gave it their all.

I am proud of my work and what our team has accomplished. We learned a great deal. I am thankful for the opportunity. Thanks for a great quarter. I look forward to adding these skills to my toolbox as I continue on my data science journey.

Individual report (Kriti Srivastava)

Personal contribution:

As a team member, my contribution and responsibilities for the success of the project are:

- I had the responsibility to research and find the data for swine flu and seasonal flu.
- Make sure to communicate, share and explain the significance of the data to the team and provide relevant graphs for it.
- Created the consolidated excel sheet 'outbreak_confirm.csv' and 'outbreak_death' once the data collected by the rest of my team members from different sources for all the viruses related to our were available.
    - The 'outbreak_comfirm' file has 4 variables – country, virus, confirmed cases, date (reported date of the case).
    - The 'outbreak_death' which has 4 variables. country, virus, deaths, date
- Also worked on some exploratory graphs as mentioned above in exploratory analysis to understand the data and see the type of graph which will be suitable for this data set.

Worked on the following Visualizations:

Exploratory: Explore_Spread (Sheet1 in Geographical spread.twb), Flu Cases in the USA (Sheet2 in Age line graph.twb), Time Series of flu cases in the USA (Sheet- Flue time series in SeasonalFlu.twb), Number of confirmed cases by countries (bar_graph in WordSpread2.twb) and Influenza and Pneumonia Mortality (Sheet- Mortality in Mortality rate.twb).

Explanatory: Geographical spread (Dashboard1 in WordSpread2.twb) , Dashboard Snapshot - Virus Evolution(Dashboard2 in WordSpread2.twb)

The details of the visualization are next to them in the report.

Learnings:

From a learning standpoint, below are the key points that I value the most from this exercise.

- Data research, cleansing, and analysis techniques to extract and preprocess the appropriate data set suitable for creating a visualization.
- Understanding the data set, exploring and analyzing various graphs that can be utilized to represent the data.

- Evaluating representation techniques that are user-friendly and self-explanatory while answering the key questions by keeping in mind the three keywords: “Data, Message and Audience”.
- I get to learn about different features of Tableau including dashboards and animation. Also, the guidelines to keep in mind while creating a good visualization like proper use of colours, the line graph for showing the time-series, a bar graph looks simple but is easy to understand and keep an eye on-axis marks and scaling.
- I realized that the success of any visualization does not depend upon simplicity or complexity but on how correctly, effectively and efficiently it conveys the message to the desired audience.
- I have also come to an understanding that data represented appropriately can show several correlation between events which we cannot understand easily. For example, economic impact of natural calamity, pandemics, sudden socio-economic changes etc. Data visualization provides a powerful mechanism to tell a story in a simple way that could be understood by anyone and hence can be used as a tool to achieve the desired result.

Individual report (Dufang Qu)

I am very interested in the topic our team chose for many reasons. First of all, it is a very timely topic and involves different kinds of variables such as numerical data, temporal data, spatial data and categorical data. Second, the major pandemic, Covid-19, started spreading from my home country, even though its origin is still unclear and remained for scientists to discover. More importantly, I believe that, unfortunately, the Covid-19 will affect or has been affecting not only the 2020 election in the US, but also the bilateral relations between countries and the international division of labor and so forth.

With all the above-mentioned motivations and a strong work ethic, I actively participated in our group project, and have contributed in the following aspects:

- Participated group meetings and discussions timely in person or online;
- Provided opinions and suggestions actively about our project;
- Devoted considerable time to look for good datasets including contagious disease datasets, economic indices such as stock market indices, consumer price indices etc., and other indices which may reflect the impact of Covid-19 such as air traffic indices, road traffic indices and air quality etc.;
- Contribute to the five milestones, including part of visualization and partial write-ups;
- Contribute to the presentation, including partial slides write-ups, producing visualization and presenting Covid-19 impact in China in the final presentation video;
- Contribute to the final project, including two visualizations for grading, paper structure building, writing of introduction and conclusion, and write-ups of other parts of the report.

The final visualization I’ve contributed to the report:

- Exploratory visualization: Impact of covid-19 in China (built in R studio)
- Explanatory visualization: Visualization 5 (built in R studio), Covid-19 Impact in China, which is a panel graph, containing an area graph, three line graphs and a heatmap
- Explanatory visualization: (for extra credit) Visualization 2 (built in Tableau), tile cartogram Provincial Confirmed Cased in China

This project was my first project focused on visualization. I’ve benefited a lot from it.

First of all, I gained a lot through the use of R and Tableau. I've mastered a lot of techniques about using R language to preprocess data and visualize it. My major visualizations were built in R, but I only had very limited experience about R before. I encountered many difficulties either while cleaning and preprocessing those datasets or perfecting the visualizations. I've tried many techniques, tricks and packages I learnt from the reference book or online materials, and now I feel much more confident and comfortable with it, and fall in love with it because of its flexibility. Meanwhile, I also played with Tableau a lot and discovered an easy way to produce tile cartograms with it.

Besides, I've learnt a lot about data visualization theories in the class. This project gave me the opportunity to apply those theories in practice, including the considerations about audience, message and data, how to choose the best or better type of visualization to pass the message, how to choose colors, how to choose scales, when we can have a non-zero baseline, what common mistakes need to be avoid, what elements are necessary and what are not, and so forth. Also, we had the opportunity to get feedback from peers and the instructors, which is the most effective way to learn.

The topic of this project is intriguing, due to our schedule and the length limit of the report, there are some many intriguing aspects I haven't discovered. I definitely will continue visualizing data about this topic in the Spring Break.

Individual report (Frank DeRango)

Personal contribution:

As a team member, my contribution and responsibilities for the success of the project are:

1. Searched for datasets related to confirm cases and deaths of the coronavirus
2. Used and distributed the dataset so my teammates can understand my approach to my graph being built.
3. Communicated with all the team members on their graphs and gave input.
4. Ask for feedback on my approach taking my visuluation.

Learnings:

From what I learned there are a few things for this project I took away from.

- Search for datasets that are as most accurate to the graph being built.
- Understand what Data, Message, Audience is and how to communicate to the goal I'm trying to reach and accomplish it.
- Was able to use a software tool Tableau for the first time and really understand how the tool can express what is built to the audience.

Individual report (Hima Spandana Barla)

Personal Contribution:

- As a group we have assigned individual tasks for each group member in the team and my area of research was MERS - Middle East Respiratory Syndrome.

- Tried to analyse all the variables that are present in the dataset and discuss with the group members to plot a graph using numeric or categorical data.
- Compared the confirmed death and confirmed cases using bar graphs for all viruses we have included in the project.Used Tableau to plot the bar graphs, created a dashboard to reflect all the topics selected and selected suitable colors to represent the peak of confirmed death vs cases on the dashboard
- Analysed each dataset available to check the categorical data to plot a Mosaic plot examining the severity level with animal contact using R and calculated the severity level. The level of severity was mentioned using critical, mild and serious.
- Attended all the group meetings and timely shared all my work on the shared document. Completed the work assigned to me on the milestones
- With the available dataset made exploratory and explanatory graphs and shared the dataset to the respective team members who were working on a each cumulative data

I worked on the following exploratory visualizations on this report:

Using R

- Created a mosaic plot to show the severity of all the viruses with specimen contact using R

Using Tableau

- Coronavirus Tests Per One Million People
- Worked on the peak of confirmed death vs confirmed
- Mers dataset and its explanatory Visualization

I have learnt the following:

- Data Handling: I collected data that is available on health websites and CAD, and decided the most valuable variables, and labeled each branch of information clearly to make it easy to separate, analyze, and decipher. And made sure the data is easily accessible to all other group members.
- Message for the audience: While working and researching on our dataset I kept audience message as my top priority and collated the work to ensure visuals and level of detail meet the desired needs
- Objective: We analysed the data with all the viruses and with the variables involved, all our group members planned a group meeting and established a clear cut of aims and goals, prior to visualizing the data. Here I chose my topic and did complete research on the data and cleaning the data
- Chart Type: Keeping in mind "DATA, AUDIENCE, MESSAGE" I presented my exploratory graphs using Tableau and R in the most effective way with the available data.For Instance: If you are showing the range of a particular entity with more than a small handful of insights, a horizontal bar graph is an effective means of visualization.Moreover, Bar graphs gives ascending and descending order.
- Color Composition: Selecting the right color played a very important role in data visualization.Color theory helped me enhance my efforts significantly.I tried to use consistent color schemes to distinguish between the elements.
- Use ordering, layout and hierarchy to prioritize: After I categorized the data and broke it down to the branches of information, I assigned each dataset a visualization model or chart type that will showcase it to the best of its ability.
- Comparison: When I was plotting my graphs to present the information, I included as many tangible comparisons as possible. The charts show contrasting versions of the same information
- Visualization Tools:Used task-specific, interactive online dashboard or tool offers a digestible, intuitive, comprehensive, and interactive means of collecting, collating, arranging, and presenting data with ease – ensuring the techniques have the most possible impact while taking up a minimal amount of your time.

RCode:

Individual report (Shadhana Palaniswami):

Individual contribution:

- Collected the time series data for ebola outbreak ,the data set for ebola outbreak before 2014 ,ebola spread in different regions of the world ,and the confirmed and death cases for different subtypes of Ebola virus.
- Analysed the ebola data set and explained the relevant work and analysis to my teammates.
- Submitted my milestone works on time in the shared folder.
- Attended all the in person and online meetings
- Asked for feedback from the teammates and incorporated the same in my works of visualization.
- Contributed to the final project by adding one explanatory graph and four other exploratory graphs

Analysis was done using many visualizations out of which I have included five of them .

Exploratory graphs:

- Donut visualization to analyse the spread of Ebola in regions of the world .
- Bar graph to compare the severity of each Ebola virus subtype in different regions.
- Line graphs to compare the rate of recovery from corona to that of the other diseases considered
- Rate of Corona recovery in different regions of the world

Explanatory Graph:

- A waffle or squared pie plot to compare the recovery from different diseases.

Learnings:

This is the first time that I'm working on a data visualization project .This project helped explore the different aspects in data visualization using R/Tableau .The main takeaways from the projects are as follows:

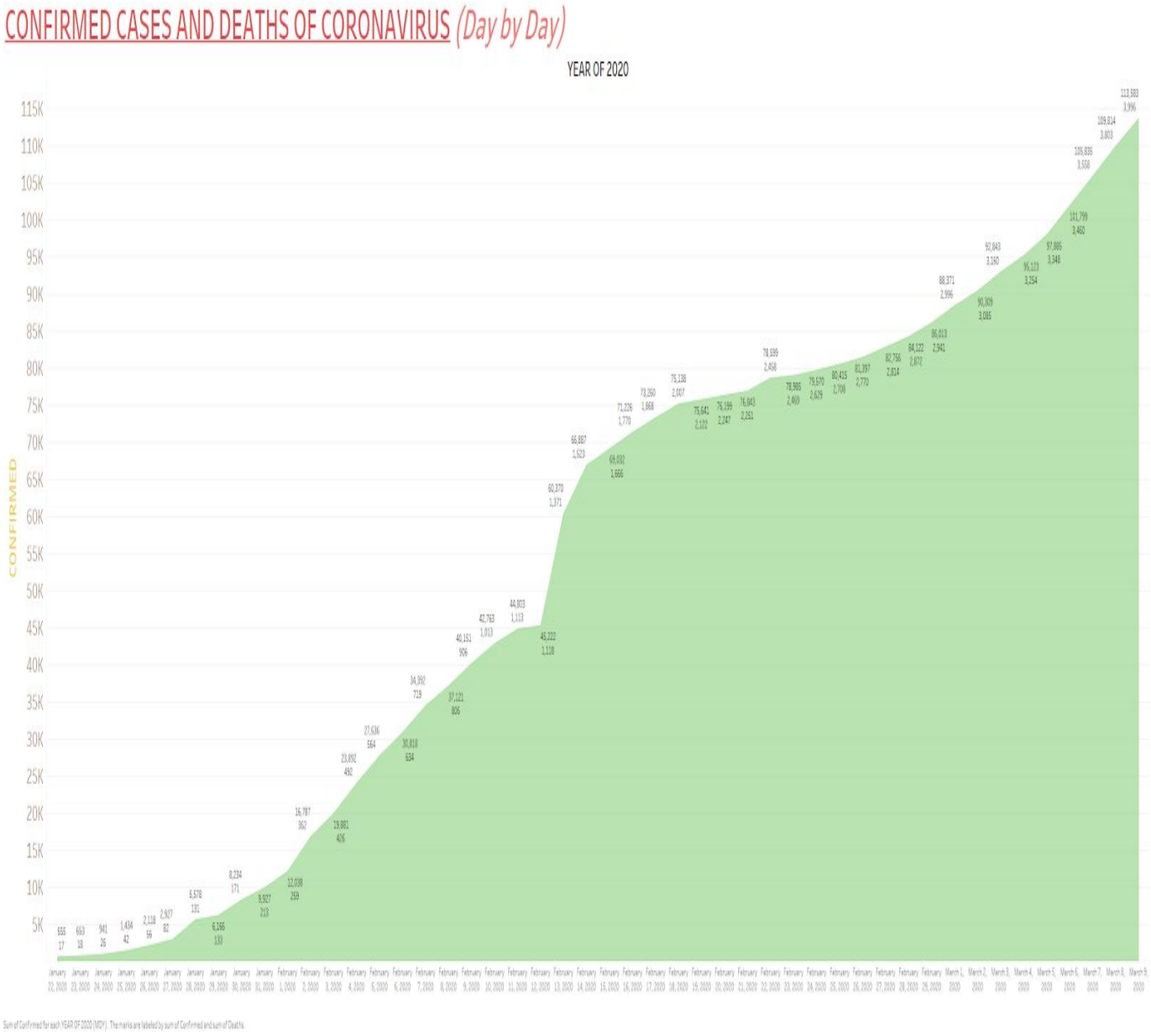
- Data Handling plays a crucial role in analysis and visualization of the data ,Always be sure to take a data that is clean or perform data preprocessing to clearly communicate the idea you using the right visualization
- Always keep in the mind the data ,message and the audience .Be sure to choose appropriate chart type to display your data .
- Also the color plays a very important role in the success of communicating your message through data visualization .Eyes beat memory so make sure to choose the right color to the chart type .
- Use necessary scaling techniques to visualize the data in the best manner
- When trying to compare graphs always make sure to use the same scale for comparison

Working on the assignment on this projects helped me understanding how important visualization is important in the data world.When we discover or mine data it is also necessary to show the results in the best way possible which is visually appealing as well as simple method.Hence the success of analysis also lies in how we are going to visualize the data and communicate the findings .

Extra exploratory graphs

Below you will find some additional explanatory visualizations our group attempted.

Day by Day of the CoronaVirus in the Year 2



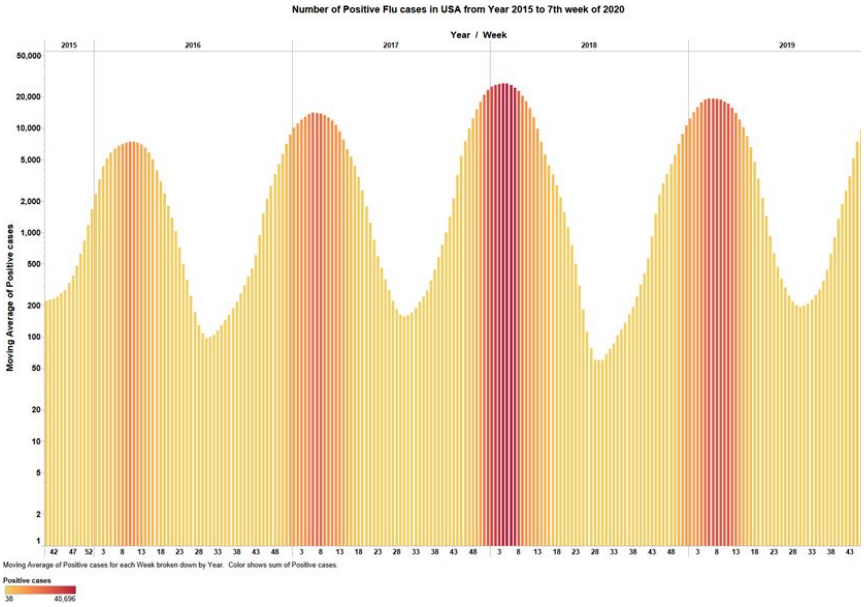
(Frank Derango)

In this graph which is a timeline over a period of a few months this is showing a gradual increase from the time the disease was reported on up until the current day. The x-axis of the graph shows the day by day of the confirmed cases and deaths around the world from the start of when the data was collected. Data started getting collected on January 22nd and went to the current date of March 9th. On the y-axis the number range is created as the cases started to gradual increase over time. You can see a spike in cases around February 13 th and February 14th as well as deaths increase probably around a third. I will probably say a lot more reporting was being conducted the first few weeks. There were a few reasons why I thought this was a good graph as visual to

the audience that I was presenting too for the simple fact it sticks out as soon as you look at it. The title of the graph is eye catching because this virus is all over the world and people would know how dangerous it is. With the incline of the line it shows over time how much it has spread over the world in the little time it has been reported on. Since its a day by day this would be looked at by individuals a lot more than if it was graphed by month. This way it gives you the most current information which for something people will need and want. I don't think there could have been another graph that I could have built where the audience and message was more captured by their attention.

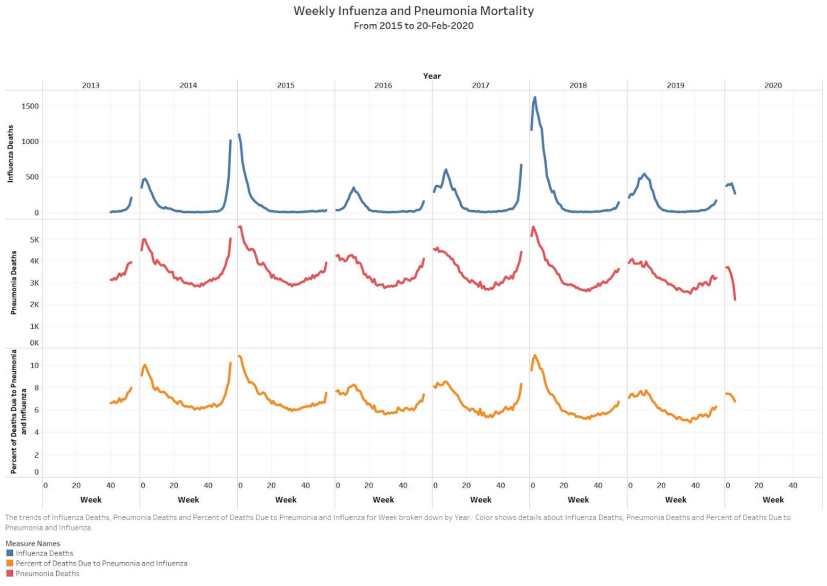
Time Series of flu cases in the USA: This graph is to show the flu cycle and weakly cases of flu in the time series panel plot. We can see that we have more cases in the 1st to 13th week of the year with 3 to 8 weeks with the largest number of cases and then at the end of the year near about 48th to 52nd week of the year (shown in the shades of dark red in color.)

For the panel plot, log_10 to show the tick marks on the y-axis to show the count of positive cases and used moving average for smoothing.



Kriti Srivastava

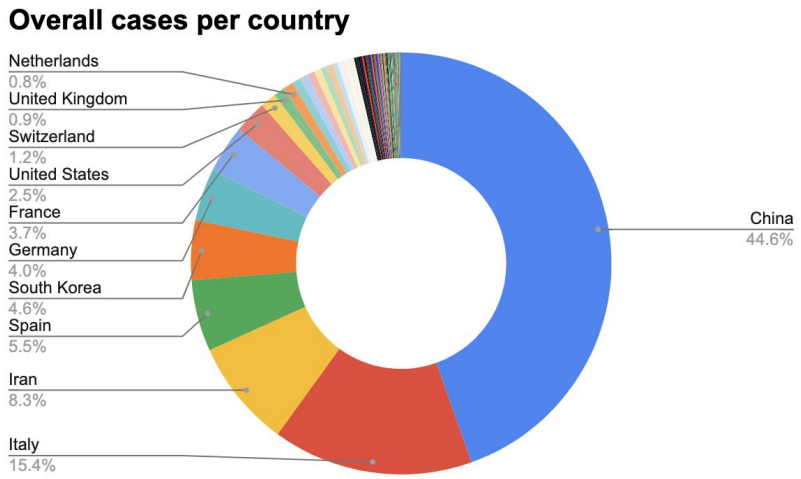
Mortality: This graph is to see the weekly Influenza and Pneumonia mortality rate in the USA and to see the death counts of Influenza and Pneumonia separately from 40th week of 2013 to 6th week of 2020. We can see that in three seperate line graphs represented in one sheet. The purpose for creating the mortality graph was if we can see how mortality of seasonal flu differs from other virus outbreak for future.



Kriti Srivastava

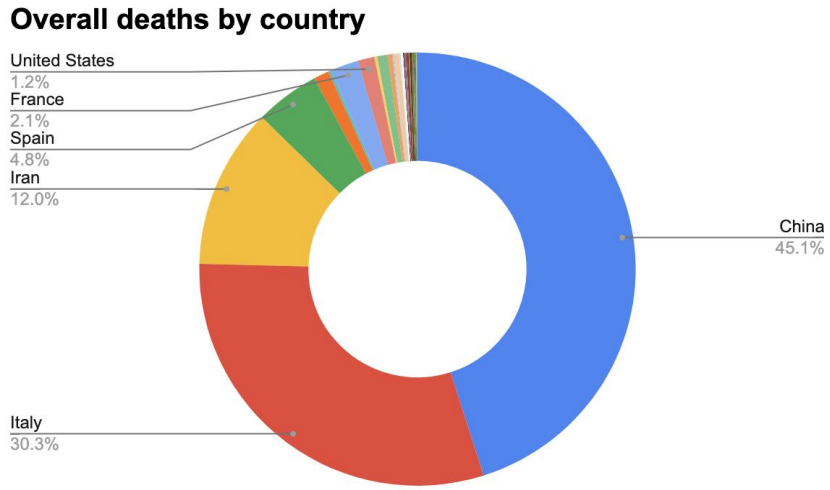
Extra explanatory graphs - James Valles

Overall cases per country: Used a doughnut chart using Google Docs to create this visualization. Segments each country by proportion of overall cases of COVID-19 per country. China had 44.6% of all confirmed cases, followed by Italy at 15.4%, and Iran, 8.3%.



James Valles

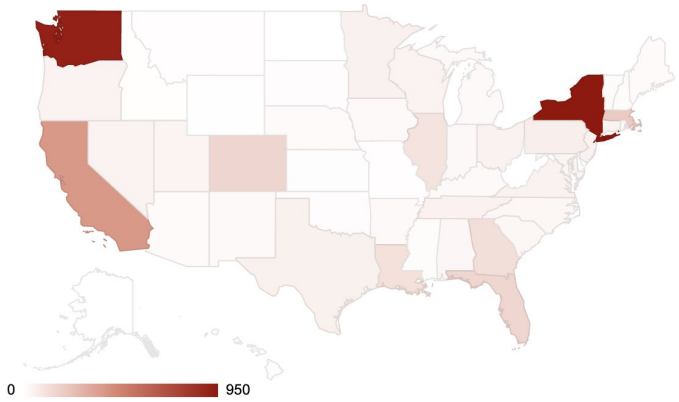
Overall deaths per country: Used a doughnut chart using Google Docs to create this visualization. Segments each country by proportion of overall deaths of COVID-19 per country. China had 45.1% of all deaths, followed by Italy at 30.3%, and Iran, 12.0%.



James Valles

Interactive U.S. Map of confirmed COVID-19 cases: Neat feature I found in Google Docs, is this tool that allows you to create an interactive map based on your data. I plotted my countries as columns and the confirmed cases as rows.

Visit Interactive Map Link:
tiny.cc/covid19chart



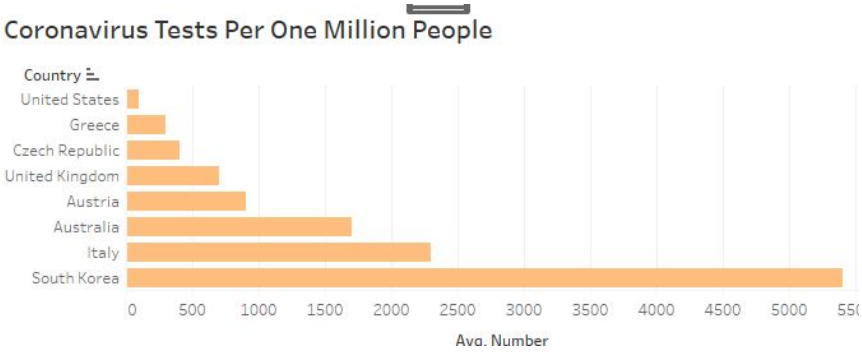
James Valles

Coronavirus Tests Per One Million People

The above data shows that about 125 people per million have been tested in the United States — far fewer than most other countries where data is available.

Through intensive testing and monitoring, South Korea has managed to slow the growth of new cases. Health officials there tracked down people with symptoms and even set up drive-through testing, allowing at least 10,000 people to be tested per day.

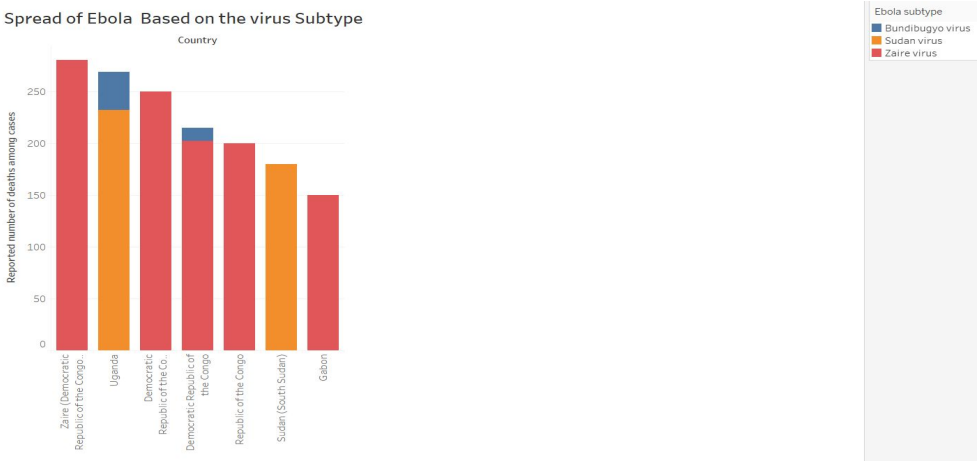
Italy has also tested aggressively for the virus, which could help explain why its total confirmed cases are higher than every other country in Europe.



Hima Spandana Barla

Spread of Ebola virus subtype:

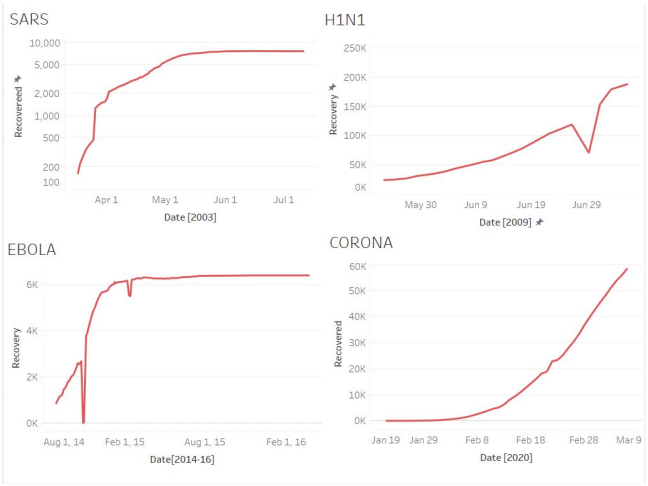
The spread of different subtype of the ebola virus in different regions is represented using the Bar chart .Different colors were applied to represent an ebola virus subtype.



Shadhana Palaniswami

Comparison of recovery from different diseases :

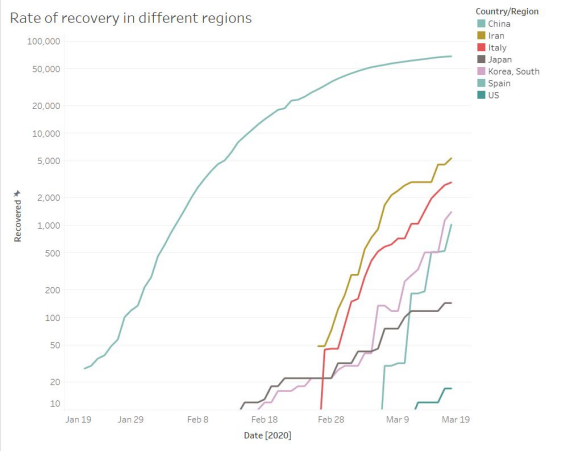
I used a line graph initially to compare the rate of recovery from different diseases using a line graph .We can see there is a increase in the recovery in all the diseases but the time taken to increase the number of recovery differs in all the diseases .The time provided here are the time period when it was a epidemic/pandemic



Shadhana Palaniswami

Corona recovery rate in different regions of the world:

I used a line plot to show the rate of recovery in different parts of the world .We can see that Only China has taken a significant amount of time to show a significant rise in recovery rate .This can can also show us the time around which the disease started spreading in different regions of the world .



Shadhana Palaniswami