Data Visualization

Outbreak

Kriti Srivastava
March 1, 2020

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%.

The study is focused on creating visualizations that illustrate the evolution and spread of COVID-19 and its impact in China and on the U.S. 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 tried to convey to our audience through our visualizations include: the comparison between the spreading patterns of each outbreak, the COVID-19 death toll.

For this project, a number of datasets for our exploratory and explanatory visualizations. At this point only number of COVID-19 datasets available, due to it being a novel virus.

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)

EBOLA: (August 2014 - March 2016 monthly)

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) .

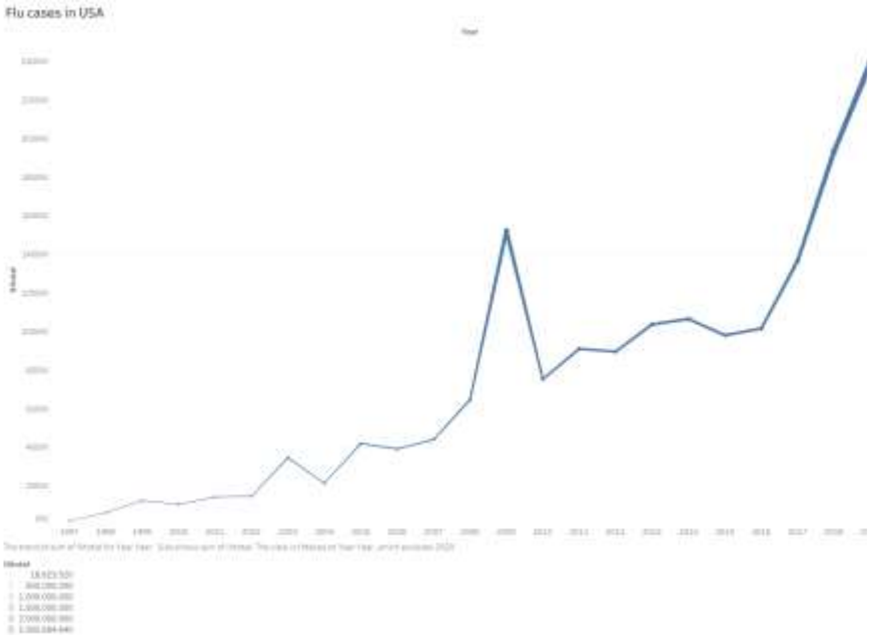
The variables we used in our visualizations include countries, recorded dates, the number of confirmed cases for each disease, number of patients who have recovered or died for each disease, stock market indices including SSE, NASDAQ, Dow Jones Industrial Average, and S&P 500. Furthermore, we used the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

Exploratory Analysis

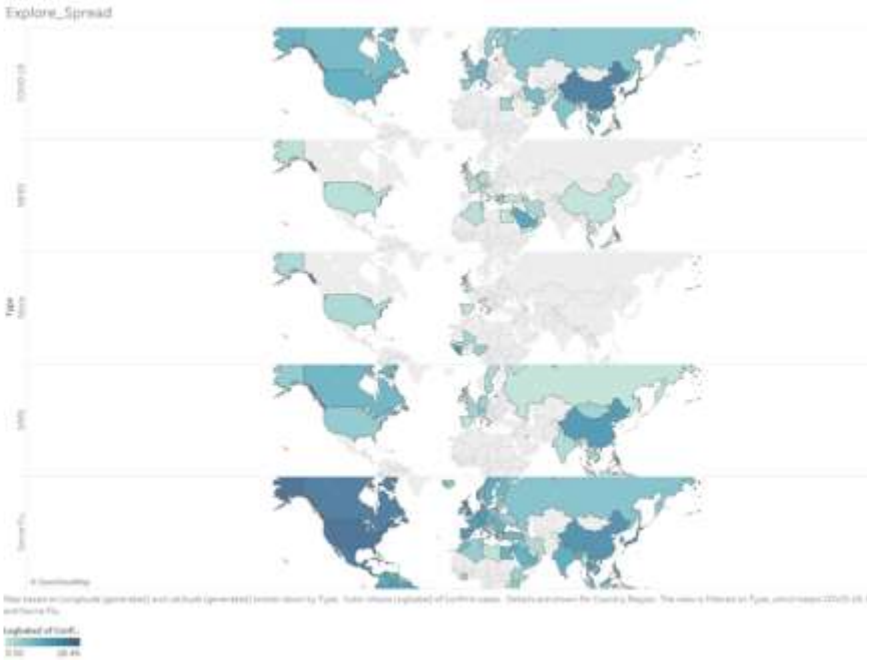
During data exploration, each of our group members chose a topic which he/she thought is worth studying. We then conducted the exploratory analysis. For the sake of keeping this report to a minimum, we have only attached seven exploratory visualizations. The folder containing R/Tableau workbooks and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

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.

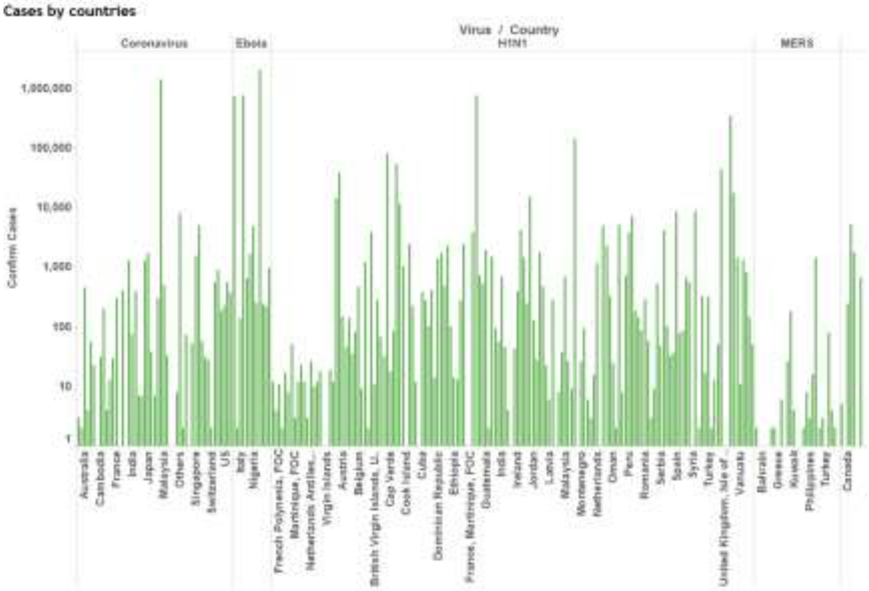
The x-axis shows the years from 1997 to 7th week of 2020. The width of the line increases as the number of cases increases. We can see the increasing trend in the number of cases through time. We can also see a prominent spike in 2009. This spike is due to the increased cases due to the H1N1 Swine flu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



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.



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.

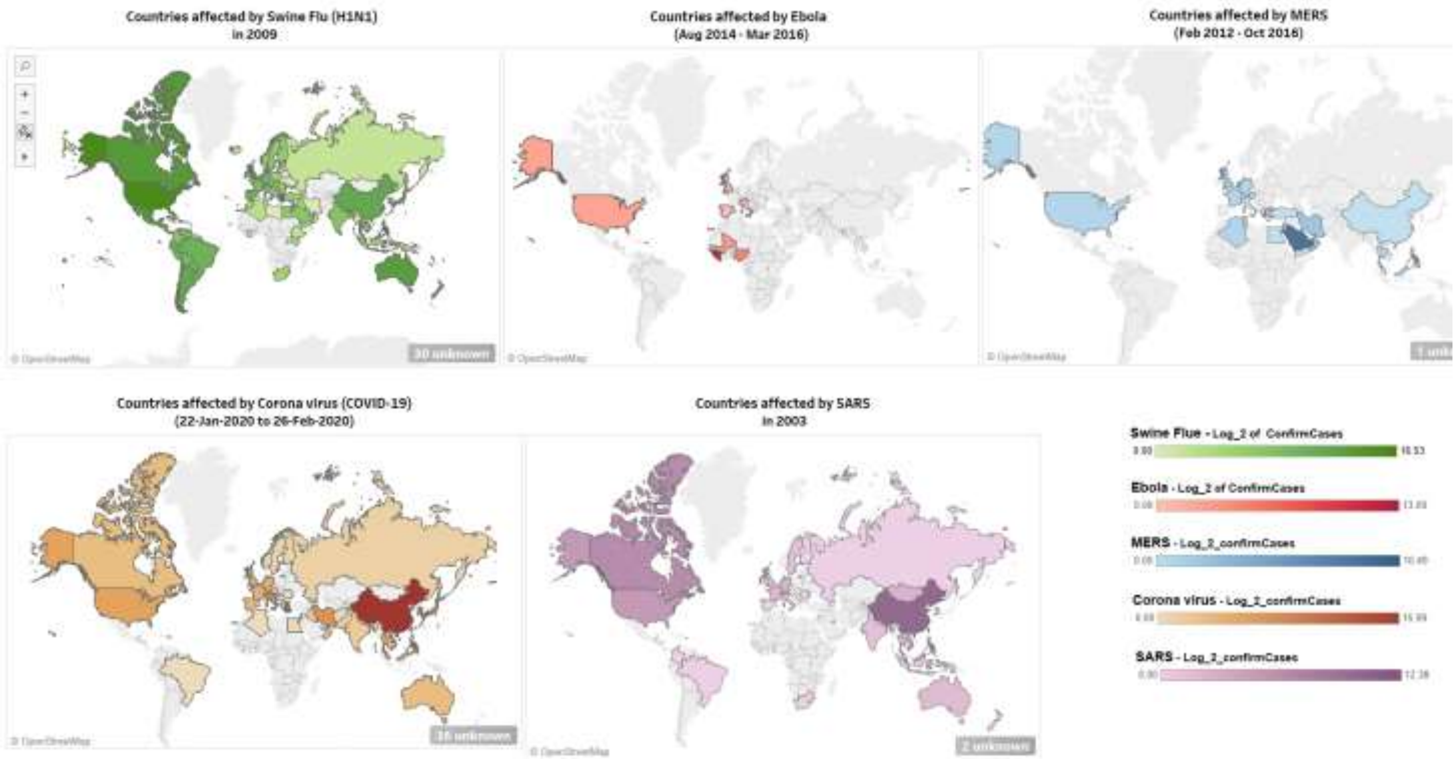


Sum of Confirm Cases for each Country broken down by Virus. The data is filtered on Action (Country), which keeps 182 members. The view is filtered on Virus, which kee Coronavirus, Ebola, H1N1, MERS and SARS.

Explanatory: Visualizations and Analysis

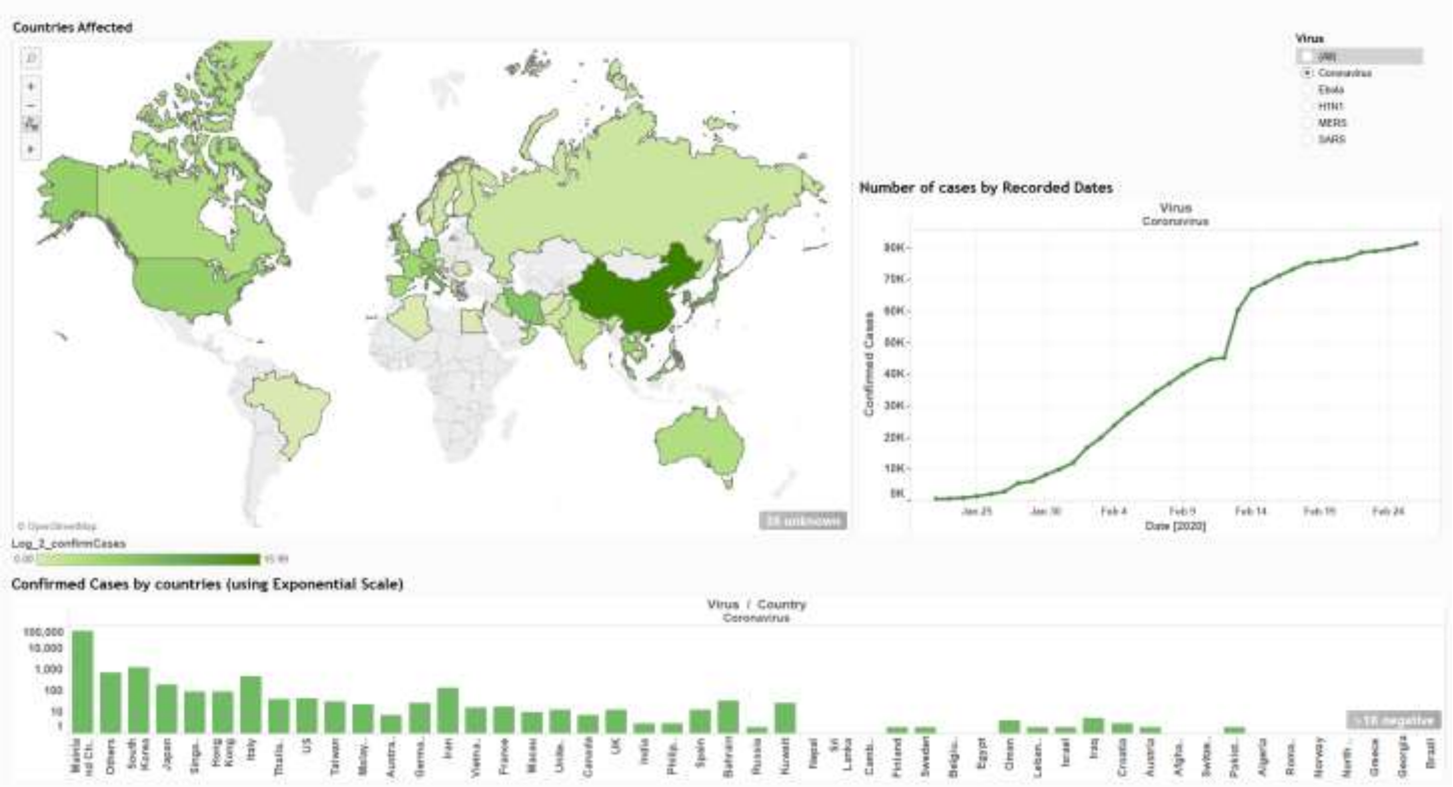
We have included the following 7 visualizations for grading.

Visualization 1: Geographical Spread


Countries affected by Swine Flu (H1N1)
in 2009
Countries affected by Ebola
(Aug 2014 - Mar 2016)
Countries affected by MERS
(Feb 2012 - Oct 2016)
Countries affected by Corona virus (COVID-19)
(22-Jan-2020 to 26-Feb-2020)
Countries affected by SARS
in 2003
30 unknown
16 unknown
2 unknown
© OpenStreetMap
Swine Flue - Log_2 of ConfirmCases
0.00
16.53
Ebola - Log_2 of ConfirmCases
0.00
13.83
MERS - Log_2_confirmCases
0.00
10.40
Corona virus - Log_2_confirmCases
0.00
15.99
SARS - Log_2_confirmCases
0.00
12.38

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.

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, the 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. Needed more time to truly understand its death rate.

Appendices

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 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)

Learnings:

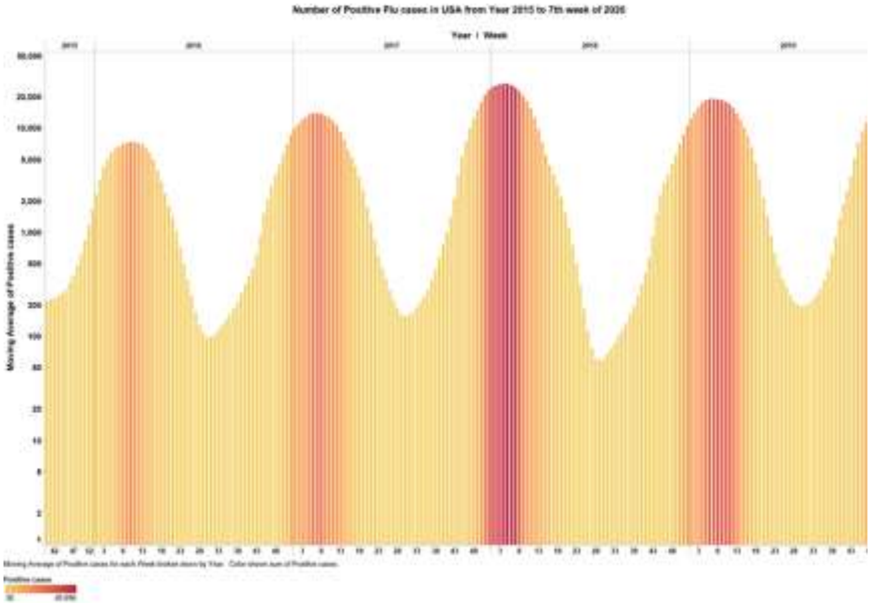
From a learning standpoint, below are the key points that I value the most from this exercise.

- 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.

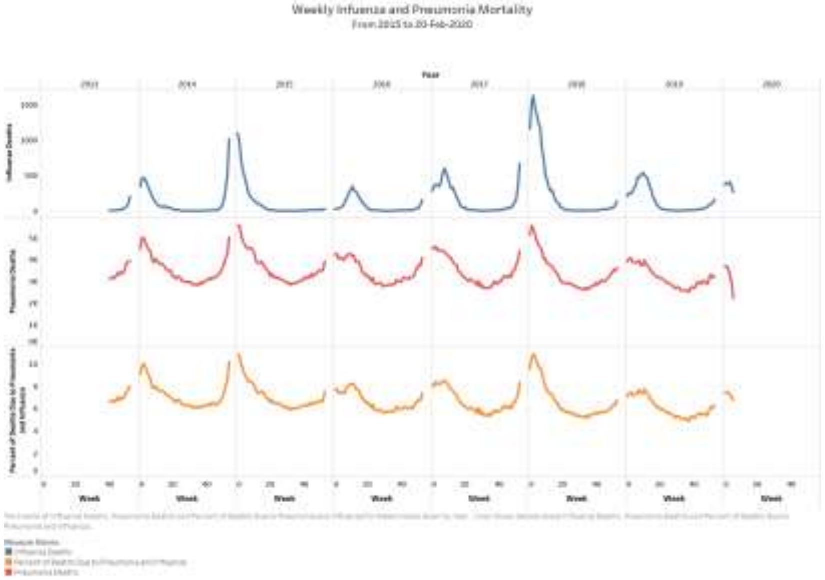
.

Extra exploratory graphs

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.



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.


Weekly Influenza and Pneumonia Mortality
From 2015 to 20-Feb-2020
Year
Influenza Deaths
Pneumonia Deaths
Percent of Deaths Due to Pneumonia and Influenza
Week
Measure Names
Influenza Deaths
Percent of Death Due to Pneumonia and Influenza
Pneumonia Deaths