COVID-19 Tableau Project Document

By:

Diego Amenabar

Ming Ki Toby Cheng

Alexander M Ilyin

Wenqiao Xu

Ziyuan Yan

I. Introduction

On December 31, 2019, the World Health Organization's (WHO's) China Country Office received information that a pneumonia of unknown cause had infected patients in the city of Wuhan. Less than two months later, on February 11, the WHO named this disease COVID-19. This disease was identified as a novel coronavirus, and at that point, had spread through many provinces of China while taking a heavy toll on its economy. It took another month before COVID-19 had taken its foothold in the United States, with California becoming the first state in the US to put a shelter in place order into effect on March 20. COVID-19 is something that has affected the lives of most of us on some degree, and will continue to for some time. With the advances and increased access to data science/visualization tools, many data practitioners have used the tools at their disposal to create insightful visualizations and predictions about COVID-19. Using some popular visualizations as inspiration, our group sought Tableau to create an interactive tool to help individuals better understand the progression of this disease, as well as take the necessary precautions.

After examining the data, we created a plan for the creation of our dashboards. Since there were two types of data (time series, daily summary), we felt that the best course of action was to create two separate dashboards: one for United States' cases and one for worldwide cases. After creating the dashboard, we decided we would aggregate the data by continent, and display the progression of COVID-19 by countries within each continent. This dashboard can be used to learn from the data that has been collected and minimize the damage the COVID-19 can potentially cause.

II. Data

The COVID-19 dataset we used can be found on the Tableau coronavirus portal, and is updated daily to reflect the constant flow of new cases that are reported worldwide. The dataset itself contains both deaths and confirmed cases per country and state/province if applicable. Every time the dataset is updated, a new entry is created for each country rather than deleting the previous entries. For this reason, aggregation is necessary to get the total numbers for the visualizations. Finally, the data was split into two types: time series and daily summary. Entries tagged as "time series" represented worldwide data, while "daily summary" data only concerns cases in the United States. The first time series data was collected on January 22, while the daily summary table was first created on March 24 as the researchers were hoping to create an aggregated view of United States COVID-19 cases.

III. Data Preparation

While we hoped to create two separate dashboards, the time series and daily summary data was contained within the same dataset. Before splitting the data, some cleaning steps were necessary since the two table types share the same fields. First, some extra fields such as Prep_Flow_Runtime, FIPS, Combined_Key, and Admin2 were removed. These attributes would not be necessary for our visualizations and removing them would improve the efficiency of our prep flow/visualization. At this point, we could split the dataset by the Table_Name attribute, and two tables were created. While the United States table was ready, our worldwide dashboard would require some external data. The worldwide data uses both Table_Name attributes as from March 23rd, US has not aggregated data in the time series, this way we can aggregate the US data for the US worldwide view..

For the worldwide COVID-19 views, we hoped to factor in per-capita cases. In order to make this calculation, we needed external population data for each country in our original dataset. To do this, we used the UN population data/estimates. This introduced a couple of new problems, since the UN data contained 2019 data as well as population projections through 2050, and some of the country names had different naming conventions than our original dataset. Unfortunately, this part of the data cleaning process required a bit of manual work as we had to replace the formal names of each country with their shorter/abbreviated names to join the two datasets. Since the UN data was collected in 2019, the data for 2020 was a prediction. For this reason we used 2019 population data for our per capita calculation, as we felt actual data would provide us with a reliable result compared to a prediction. For this reason, we excluded all population data not from the year 2019 and joined the population and COVID-19 by country name. Finally, we divided the population counts by 1000 since the data was in the 1000s. This is because calculating per capita infections is more meaningful at the actual scale of population. At this point, both of our datasets were ready and we could begin creating our dashboards.

IV. Dashboard Creation

Our US centric dashboard contains a total of four views: two maps, a table and a trend line. In the top left corner of the dashboard are two counters that show the total US confirmed cases and deaths as of April 1st. These numbers give the user a general overview of the severity of the virus as of April 1st. The first map is a country-wide view of US COVID-19 cases. In this map, each state has a blue color, with its shade determined by the number of total cases in said state (the blue gets progressively darker with the amount of cases). If we click on a state, we activate the right side map. This map provides a state and county wide view of confirmed cases, with each circle within the state corresponding to a county (if that county has any confirmed

cases). The size of the county circle is defined by the amount of cases: counties with a large amount of cases will have larger circles.

On the bottom left hand corner is a summary table of both confirmed cases and deaths by state. On the last row of the table is a row corresponding to the grand total cases in the US. Finally, the trend line in the bottom right corner shows the progression in confirmed cases and deaths since the start of US data collection on March 24. For users from the United States, this dashboard gives an in-depth view of the COVID-19 crisis as it spreads throughout many states and counties.

As mentioned previously, the second dashboard is a view of worldwide COVID-19 cases (including the US). The top half of this dashboard is reserved for a worldwide map with two scales: total cases and cases per capita. Similarly to the US map, each country is colored in a different shade of blue depending on its number of confirmed cases. However, each country also has a circle within its borders representing the number of cases per capita. The size and color of the circle is determined by the severity of cases per capita. On the bottom left corner of the dashboard is a trend line for cumulative confirmed cases and deaths. Finally, the bottom right corner contains a line plot representing infection growth for countries, with the X-axis defined as the number of days since the country first recorded 1000 cases. This plot and the rationale for choosing 1000 cases will be explained in the story creation description.

V. Story Creation

While the dataset contained both country names and states/provinces, we felt there was an additional level of aggregation that was missing. In order to create an effective story visualization, we decided to aggregate each country by its continent. We were forced to do this manually by plotting each country and using the lasso selection tool to create groups. Each group received the correct label (since North and South America don't have many countries with >1000, we aggregated these continents). In the story visualization itself, there is a tab for each continent, and each continent tab contains a recreation of the line plot from the worldwide dashboard limited to the countries of that continent. We chose the number of days since 1000 infections as the X-axis since any number less than 1000 produced a cluttered visualization.

VI. Conclusion

With the advent of data visualization tools in recent years, the potential for creating visualizations to drive the insight/understanding of this complex situation has greatly increased. We hope that these dashboards and story visualization can be helpful for individuals and decision makers alike, as COVID-19 continues to take its toll on the world. While many countries have

taken a great sacrifice, others should use the mistakes/lessons learned from previous outbreaks to better prepare themselves, and there is no better way to drive this learning process than with data.