Team Members: 10/12/2023

Michael Chiaramonte

Cole Swartz

Matt Lantin

**COVID-19 Variant Exploration**

**Summary:**

The idea behind this project is to analyze how the overall picture of the COVID-19 virus can be better understood by analyzing its different variants. While most people can recognize that COVID-19 has variants through repetition of names such as Delta or Omicron, it turns out that there are many subsets of variants within these overall designations. Our goal was to drill down to some of the most prominent variants and see what we can discover about how they are evolving and propagating through different regions of the US.

**Research Questions:**

**Q1:** Which are the most dominant COVID variants over the last 3 years?

**A:** We first needed to collect the entire variant dataset and then trim it down to the relevant date range. When parsing the data, we discovered that there are 56 unique COVID variants recorded by the CDC. We decided to focus on the top 20 that have the highest percentage of overall COVID cases for a particular week. To determine this, we checked the max value for each of the variants (highest share % at any point in the trend).

A graph of a graph

Description automatically generated with medium confidence

As you can see from these variant trends, there are 5 very dominant variants over the course of the 3-year period we have trended:

* B.1.617.2
* BA.1.1
* BA.2
* BA.5
* XBB.1.5

Each of these variants have at one point a greater than 75% share of the overall COVID distribution for that week. We decided to focus on these specific variants to help see how COVID has been evolving throughout the last 3 years.

**Q2:** How are different COVID variants correlated to death count?

**A:** We now need to bring in the death data and see how this is evolving over time. Once the data is cleaned, the death count trend shows as follows:

A graph of a covid-19 virus

Description automatically generated

We can see that the highest death counts fall between July 2021 and April 2022. We need to check how this corresponds to which variants are most dominant at that time. Since both datasets are gathered over the same weekly time axis, we can overlay the plots together as shown below:

A graph of different colored lines

Description automatically generated

Here we are taking those most dominant variants and plotting along the death count trend. We can see that variants B.1.617.2 as well as BA.1.1 appear to be dominant when the most deaths were occurring in the previously mentioned period. Therefore, we can conclude that these are the more deadly strains of COVID than the subsequent variants that came after.

Now the question becomes, why would COVID evolve in such a way? Generally, the main goal of a virus is not to kill as many people as possible, but rather to infect as many people as possible. This is its way of continuing to survive. Based on this data, we can infer one possible reason that subsequent COVID variants mutated to become less deadly was the fact that killing the host was not allowing that host to infect more people. The virus needs the host alive to keep propagating Therefore, to become more efficient in spreading these strains of COVID needed to become less deadly.

**Q3:** Are different COVID variants prevalent in different regions?

**A:** We also wanted to explore if different COVID variants are more dominant in different regions of the United States. Our variant data fortunately has a column that breaks it down by different Health and Human Services (HHS) regions. While there are 10 regions total, to test our theory we will pick 3 for analysis (Region 2 – New York, Region 7 – Kansas City, and Region 9 – San Francisco).

A map of the united states

Description automatically generated

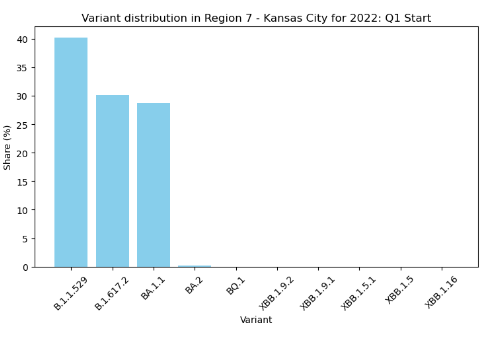
HHS Region Map

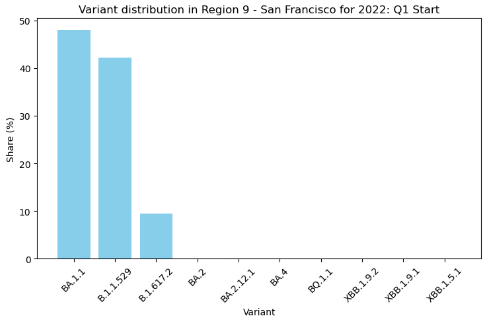
https://www.hhs.gov/about/agencies/iea/regional-offices/index.html

We broke it down by quarters of the year to better group the data (Q1/Q2/Q3/Q4) which correspond to the first data points from the following months respectively (January/April/July/October). Let’s examine Q1 of 2022 for these regions, which based on the graphs above is right between the transition between two of the biggest variants B.1.617.2 and BA.1.1.

A graph with blue bars

Description automatically generated



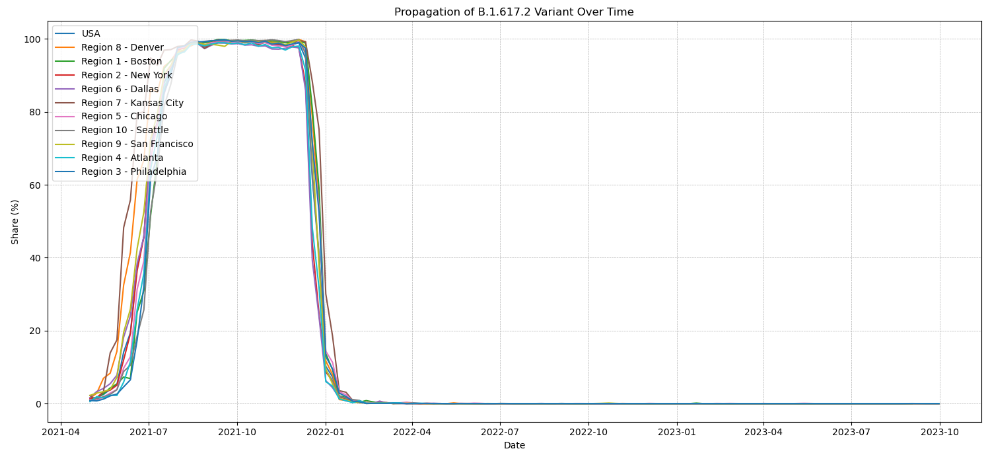


We can see from the above that at the same week at the beginning of Q1 2022 each of these three regions is having a different distribution of not only the two variants we know are some of the most dominant over the entire time period, but there is a third variant here (B.1.1.529) that is showing as a very high share % at this time as well. Analyzing these variant distributions per region during a ‘transition period’ we can infer a couple of things:

1. We know from our previous trends that BA.1.1 eventually becomes the largest share by a lot over the next few months after the beginning of Q1, so BA.1.1 is likely to be a sublineage (mutation) of B.1.1.529
2. There is a possible geographical component to how COVID variants are spreading throughout the US.

**Q4:** How do top variants propagate through regions over time? Did the variants originate from a specific region and if so then how did it spread?

**A:** Based on the previous charts, we know that COVID variants are evolving. However, we still need to determine whether these new variants show up uniformly within the regions as the disease evolves everywhere or do they originate somewhere and then make their way through the regions geographically. We plotted some of the top variants to try and answer this question.



A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a graph

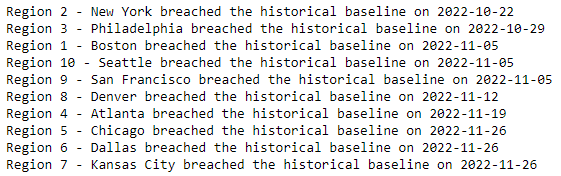
Description automatically generated

A graph with different colored lines

Description automatically generated

The idea here is to see how some of the top variants evolve within each region over time. Most of the variants look to be spiking up immediately in all regions, so since our data is weekly it becomes more difficult to see which region that variant may have started in. However, when we look at the XBB.1.5 trend we can see that the spikes are more spaced out over time even with the data being reported weekly. We can examine this variant to find out how it propagated throughout the entire country to maybe narrow down the area where it started.

To do this, what we would like to do is calculate a historical baseline for the variant per region, and then track the first week it deviated from this baseline. Using python code, this calculation becomes rather simple to perform for each region (please refer to region map above). The printout for XBB.1.5 reads as below:



Based on this printout, we can now see that it appears that the first instance is in the New York region on 10/22/2022. After that, we get the Philadelphia region next followed by Boston, Seattle, San Francisco, Denver, Atlanta, Chicago, Dallas, and finally Kansas City. This specific regional order tells me that it is possible that this variant originates outside the US and is coming in through the coastal cities. It then makes its way from both sides of the country and eventually shows up in the most central HHS region: Region 7 - Kansas City.

**Conclusion:**