CSE 5544 - Lab 2

10/05/2020
Logan Frank

1. Dataset description.

In my lab, I used the following datasets:

- US Presidential Election Results from Harvard: this dataset contains information regarding how many votes each president elect received in each state for every election year.
- US State Data from NYTimes: this datasets contains information regarding the total number of cases and deaths in each state, tracking from late January to mid/late September.
- Population Data from US Census: this dataset contains the results of the 2010 US Census and estimates for the population in every state for 2011-2019.
- US Governors from Civil Services USA: this datasets tells us the current governors in every US state and their political affiliation.
- Cases Over Time in Each County from CDC: this dataset contains information regarding the total number of cases in each US county from late January to late September.
- Population Data by Age from US Census: this dataset contains information regarding the population of each age group for every state in the US.
- Data on Conditions Contributing to COVID-19 Deaths from CDC: this dataset contains information regarding whether any pre-existing conditions existed at the time of death for persons whose death was listed as 'COVID-19', also contains numbers for just COVID-19 deaths.
- Containment and Health Index Data from OWID: contains the health index for multiple countries from early January to present. The health index measures is a composite measure of 11 different response metrics.
- Worldometer Data from Kaggle: this dataset keeps track of multiple attributes for all countries around the world. These attributes include the total number of cases, total number of deaths, the number of active cases, and more.
- University information found using Google such as county and AU20 semester start date

---

2. Software description.

Python 3.8 through Miniconda and using Jupyter Notebook, Pandas for loading CSV files, Plotly for all visualizations

---

3. Analysis.

3.1) COVID-19 questions.

1. Do Democratic or Republican states have a higher percentage of COVID-19 cases based on their total population?

2. Did a country's early adoption of COVID-19 prevention measures contribute to them having a low

percent of active cases currently (compared to the total cases)? Did any continents adopt earlier than others? If so, are they doing better in terms of percent of active cases?

3. Did certain US regions have earlier peaks in daily increases in COVID-19 cases than others?

4. Does the percentage of total age 50+ citizens (by US state) correlate with the percentage of total US deaths?

5. Are all age groups dying of *just* COVID-19? Do certain age groups also tend to have complications with other conditions?

6. Did many US universities see increases in their county's 7 day averages of COVID-19 cases once the classes for AU20 semester began?

3.2) Dataset analysis.

Q1: This question involved using datasets "US Presidential Election Results from Harvard", "US Governors from Civil Services USA", "US State Data from NYTimes", and "Population Data from US Census" datasets. After consolidating the data into one dataset containing only the information needed for visualization, each visualization contained 3 variational components: the state name, the political affiliation of the state, and the percent of state population with a positive COVID-19 case. All datasets are of type Table with items and attributes as data types. The first two variational components are categorical and the last one (percent of state population with positive COVID-19 case) is quantitative.

Q2: This question involved using datasets "Containment and Health Index Data from OWID" and "Worldometer Data from Kaggle". Once the data had been cleaned and consolidated, the visualization contained 3 variational components: the country, the number of days it took a country to implement some COVID-19 prevention measures to get a health index score, and percent of active cases (corresponding to the total number of cases). Many of the variational components also depended on other attributes such as the date at which a health index score was given to a country, the total number of active cases, and the total number of cases. A continent attribute was also created but this depended on the country. All datasets are of type Table with items and attributes as data types. The country attribute is categorical, the second variational components is technical ordinal since it is a date, but in the format I have it, you can add and subtract the values thus making it quantitative. Then the percent of activate cases is quantitative.

Q3: This question involved using the dataset "Cases Over Time in Each County from CDC". After parsing the data to get the information desired for visualizing, there was 3 variational components: the county location (identified by the FIPS id), the largest day increase in positive COVID-19 cases, and the date at which that largest increase occurred. The final dataset would probably be best described as Geometrical since there is a position (the county FIPS id) and an Item (largest increase in daily cases and date) for each position. As mentioned, the data types are Position and Item(s). The position attribute is categorical, the date is ordinal, and the largest increase in daily cases is quantitative.

Q4: This question involved using datasets "Population Data by Age from US Census" and "US State Data from NYTimes". Once done cleaning the data, there was 3 variational components: the state name, the percent of total United States age 50+ citizens living in that state, and the percent of total United States COVID-19 deaths in that state. This dataset is of type Table with items and attributes as data types. The state name is categorical and the other two components are quantitative.

Q5: This question involved using dataset "Data on Conditions Contributing to COVID-19 Deaths from CDC". After cleaning the data, there was 3 variational components: age group, condition group, and scaled deaths. The dataset is of type table with items and attributes as data types. The age group and condition group are categorical and the scaled deaths is quantitative.

Q6: This question involved using dataset “Cases Over Time in Each County from CDC” and some self-collected data on 7 different universities, specifically their county location and AU20 semester start date. After consolidating and clean the data, there was 4 variational components: date, location (county & university name), the 7 day average in daily increase in positive COVID-19 cases, and the AU20 start date for each university. The dataset is of type tables with data types of item and attribute. The date is ordinal, location is used as categorical in this scenario, the 7 day average in daily increase in positive COVID-19 cases is quantitative, and the start date for each university is ordinal.

3.3) Describe how analysis will be done.

As mentioned, Pandas was used for all data loading and data preprocessing needs. Exact operations used will be described under their respective question.

Q1:

- Which state was Democratic or Republican was determined in two different ways: one by which president elect won the state in 2016 and one by what the political affiliation of their current governor is. Who won each state in 2016 was found by filtering the year in the “US Presidential Election Results from Harvard” dataset and the political affiliation of each governor was found using the “US Governors from Civil Services USA” dataset.
- The total number of cases each state has seen and reported thus far was determined by grouping a DataFrame containing the “US State Data from NYTimes” dataset by state and then getting the row with the maximum cases.
- The estimated 2019 population for each state was found using the “Population Data from US Census” dataset by simply extracting one column out of the DataFrame.
- The cases by population percentage was found by dividing the total number of cases found in item 2 by the total population found in item 3 then multiplying by 100 to make a percentage.
- With the above items, the states were grouped according to item 1 and then sorted by the percentage found in item 4. A mean for each political group was also found trivially.

Q2:

- The country was found simply by doing a groupby on the DataFrame when searching for the other variational components.
- The date was found by first filtering out all rows which had a health index of 0. Then a groupby operation was performed where the rows with the earliest date were selected. This got us both the country (which was already given, just needed to be reduced to one item per country) and the date. These were both found from the “Containment and Health Index Data from OWID” dataset.
- The percent active cases was found by filtering the “Worldometer Data from Kaggle” dataset to retrieve the total cases and total active cases for each country.
- Both sets of data were filtered to ensure they had the same country items and then were combined.

Q3:

- The daily change in COVID-19 cases was found similar to what was done in Lab 1 using the “Cases Over Time in Each County from CDC” dataset, then the largest daily change was found and that row was extracted. This row extraction was done so there was only one row for every county/state combination. The day for every row was compared to the largest examined day to determine how many days had occurred since each county’s ‘peak’. Thus we had a peak date for every county FIPS in the United States.

Q4:

- The percentage of US people aged 50 and above in each state was found by filtering the “Population Data by Age from US Census” dataset for age and then ensuring both sexes were accounted for (i.e., filtering out female-only and male-only data). The percentage was found by dividing each states 50+ population by the total in the United States.
- The percentage of US deaths in each state was found pulling the state and death data from the “US State Data from NYTimes” dataset and performing a similar operation to item 1 except age was not considered because we wanted to see if there was a correlation between total deaths and the percentage of older people.

Q5:

- The ‘scaled deaths’ information for this question was found using the “Data on Conditions Contributing to COVID-19 Deaths from CDC” dataset and removing two condition groups from the DataFrame: “Malignant neoplasms” as this appeared to be consistently low / insignificant amongst all age groups and “Intentional and unintentional injury, [etc.]” so we could limit our analysis to non-spontaneous conditions. The removal of these two condition groups made visualization much easier. The number of deaths per condition group per age group was then grouped together and scaled according to the sum of deaths across each age group. The result of this was used for analysis. In other words, the scaled deaths tell us what percentage of persons within each age group died of what conditions (e.g., 30% died from just COVID-19 and 70% died with complications of respiratory issues in the 40-50 age group - not an actual observation, just an example)

Q6:

- The daily increase in COVID-19 cases by county data was found using the “Cases Over Time in Each County from CDC” dataset. Only the counties corresponding to Ohio State University, Texas AM University, University of Alabama, University of North Carolina, Brigham Young University, University of Utah, and Penn State University were used in this visualization either for their student population, their start date (UNC had a much earlier start date), their prominence in the news (University of Alabama was in the news for their rapid increase in cases early in the semester), for being a large private university (BYU) or other. The daily increase in these counties was found similarly to Lab 1 (except by county instead of the entire United States).
- The 7 day average was computed using the information found in item 1.

3.4) Describe the data visualization design process.

Q1: This visualization uses lines for marks and vertical length, x-position, and color for channels. I am using size and color for visual variables. Size is used to represent which state contributes a larger percentage to the US total positive COVID-19 cases, sorted from smallest to largest within each political party group. Color is then used to easily differentiate between the political groups. By grouping by political party and sorting the states within each group, it is easier to determine which group has states with higher percentages, then by adding the mean line, it makes it even easier to draw conclusions from the visualization.

Q2: This visualization uses points for marks and x and y position and color for channels. I am using color for visual variables. The x axis encodes the delay in days from the first day of the year that a country took to receive a numerical health index value, the y axis encodes the current percentage of active cases (based on total cases), and the color encodes the continent of each state. By mapping the continent to color, a viewer can group together countries of the same continent.

Q3: This visualization technically uses points for marks since each county is just a point on the map and for this visualization, the area of the county carries no significance. Then this visualization uses color as a channel and visual variable (although the visual variable could be considered value). The color corresponds to the days since the peak in daily increase in positive COVID-19 cases, the lighter the color means it has been longer since the peak, darker obviously means the opposite - the peak has happened more recently.

Q4: This visualization uses lines for marks although there are points represented inside each line mark. The channels are both x and y position and color. X position encodes an ordering of states based on percentage of age 50+ people and y position encodes the actual value, then color encodes two different sets of data - percentage of age 50+ people and percentage of COVID-19 deaths. This visualization uses color as a visual variable, simply by assigning different data to different colors.

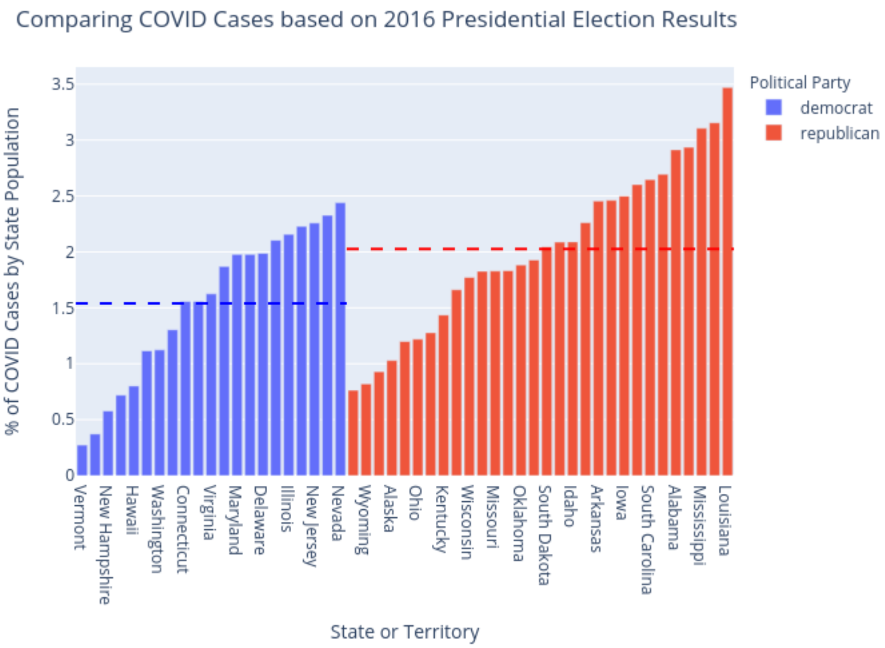
Q5: This visualization uses lines for marks although, similar to Q4, there are points represented inside each line mark. The channels are both x and y position and color. X position encodes the age group from youngest age group to oldest age group so we can see trends in condition groups as we increase the age and y position encodes the scaled deaths for each category in every age group. Color is used to differentiate between the condition categories. This visualization uses color as a visual variable and the x plane could be considered a visual variable. The different condition categories maps to the color and allows us to differentiate each condition and the x plane maps to the age groups thus we can single out individual age groups and analyze them individually.

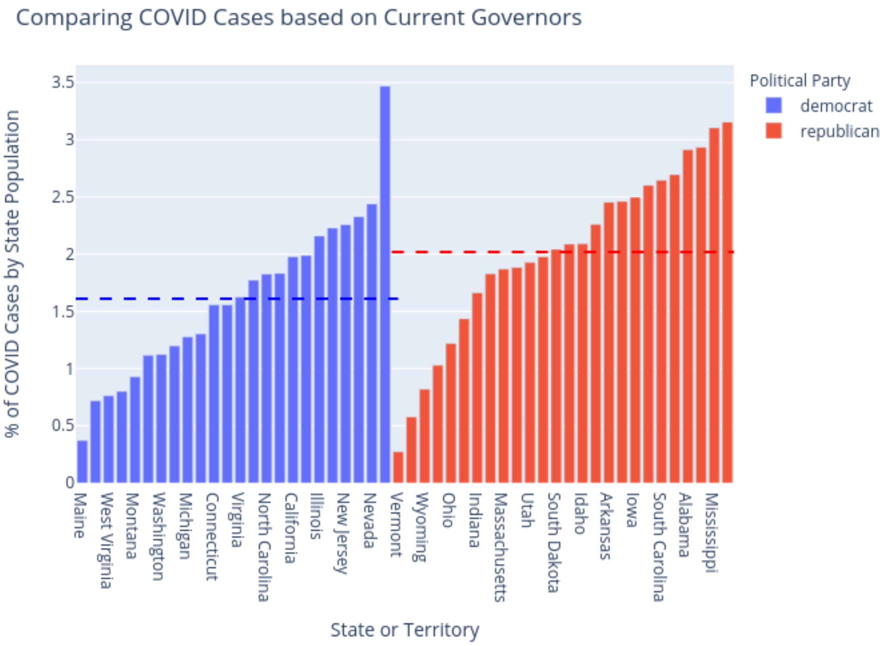
Q6: This visualization uses lines to show the change in 7 days averages over time and points to show when a university starts their AU20 semester. Color and x and y position are used as channels and color is a visual variable. Color maps to the different universities being analyzed and allows us to single out individual universities.

3.5) Justify your choice of data visualization.

Q1: Using a bar chart for this visualization seemed most appropriate because a viewer can easily differentiate the groups and since we are comparing a quantitative value, it seemed like the only right choice for visualization. It seemed even more correct since we wanted to add a mean line to the visualization to answer our question. The question can be answered quantitatively by looking at the mean lines and seeing the democratic line falls below the republican line. One may be able to determine the answer with selectivity by grouping the blue and red groups then determining the blue group looks smaller (height-wise) on average than the red group, but this is much more difficult than looking at the mean lines.

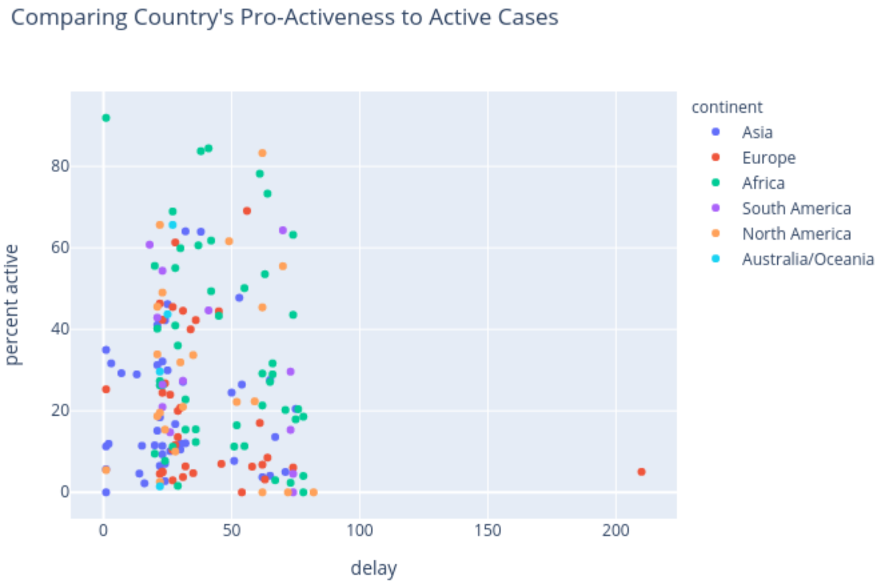
My questioned is answered by the mean line. In both visualizations - 2016 presidential election results and current governor political affiliation - the democratic states are lower than republican states.





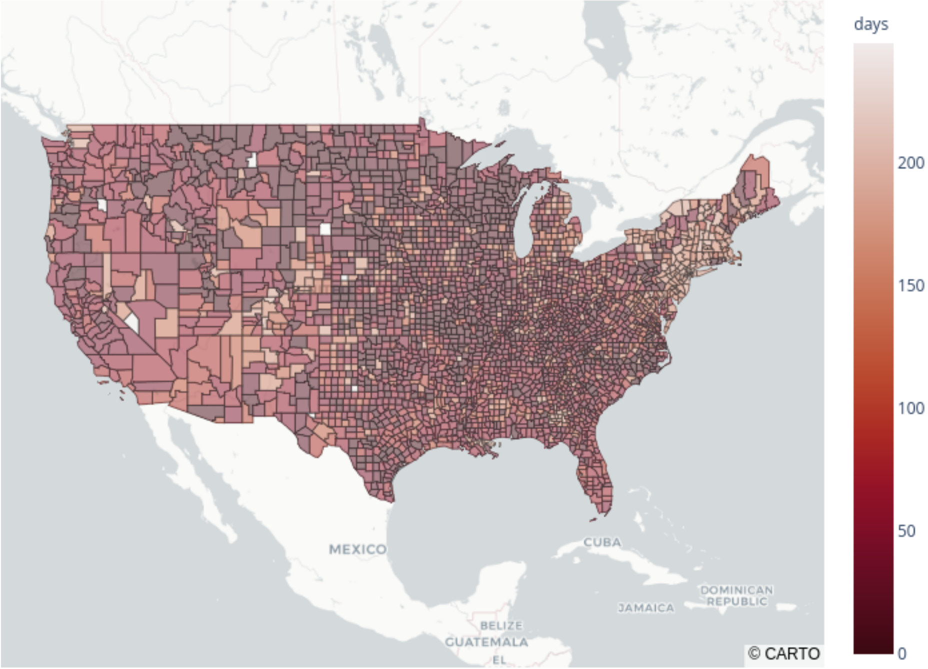
Q2: Using a scatter plot for this visualization seemed appropriate as there were many static items with no time varying data to track. The first question can be answered just by analyzing the general trend in the visualization possibly using some associativity and ignoring the color channel and the second question can be answered using selectivity by grouping countries of the same continent together.

The first question (regarding whether a country's early adoption contributes to a lower current active case rate) is answered by trying to identify trends in the data. If the question is correct, then we should see a linear trend, however we do not see that, so a country's early adoption of prevention methods does not seem to play a significant role in their current number of active cases. The second question is answered when we look at each continent individually. We can see that many countries in Asia adopted prevention methods earlier than North American and African countries and are also generally doing better in terms of the number of active cases.



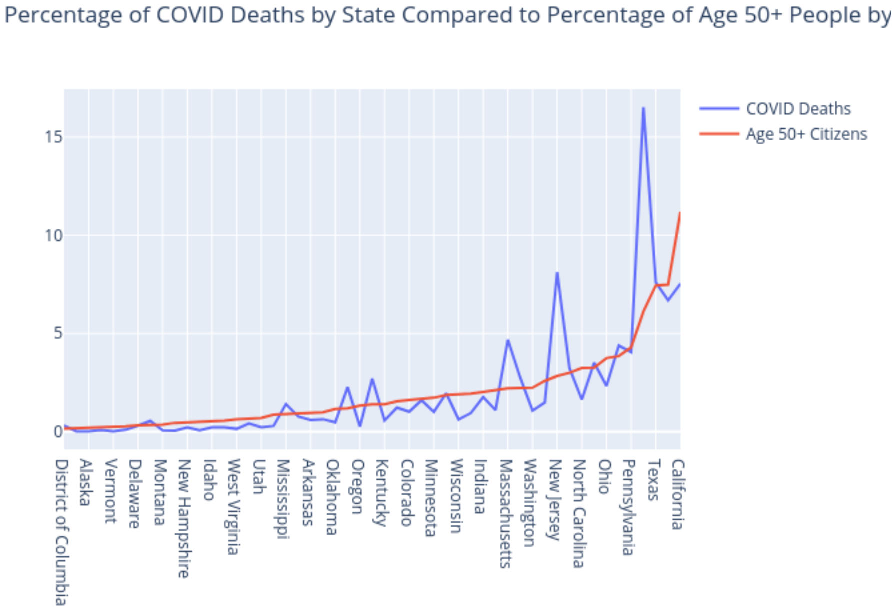
Q3: Using a 2D map for this visualization was most appropriate as we are trying to answer our question by looking at regions as a whole. The viewer can use selectivity based on color/value to answer the question. The question can be answered by looking at regions that have a large number of lighter counties.

My questioned is answered by looking for regions with substantially lighter counties. In this visualization, we can see that the northeast had the earliest peak but has stabilized, the west had their peak afterwards, and the ‘tornado alley’ areas had the most recent peaks.



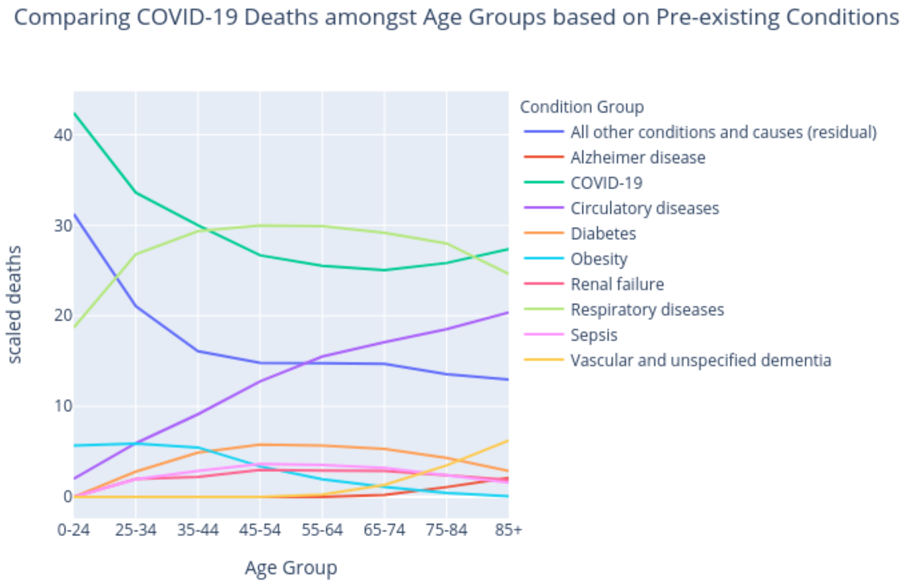
Q4: This question wanted to see if two sets of data align and correlated, i.e., the percentage of age 50+ citizens line should share a trend with the percentage of COVID-19 deaths line. Thus a line graph seemed most appropriate for answering this question. Associativity of color helped in solving this question as we were able to analyze the two lines together, regardless of color, to notice that there is a significant trend between the data.

My questioned was answered by disregarding the color of the lines and noticing that there is a significant trend between the lines, with the exception of some outliers.



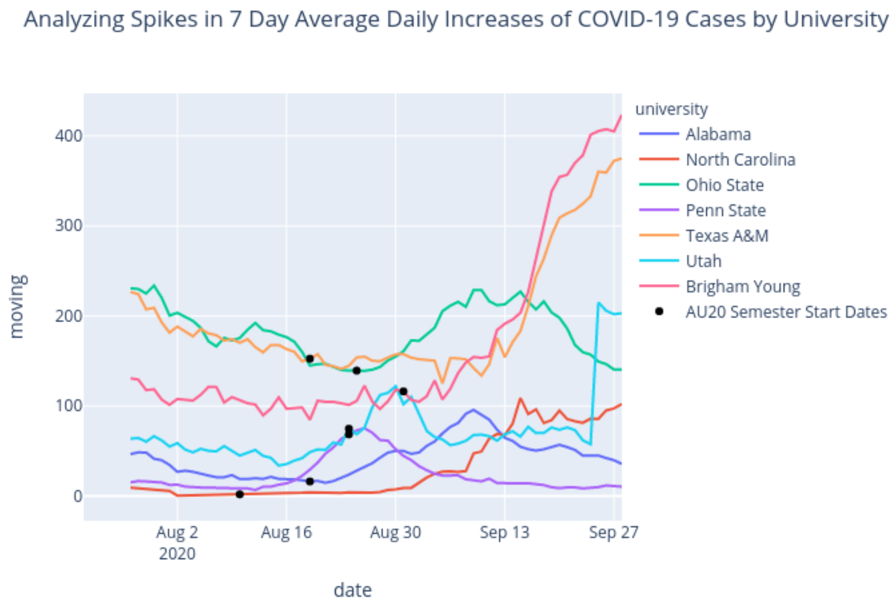
Q5: For this question we wanted to see what the trend of certain condition groups (in COVID-19 related deaths) were amongst the age groups. Thus a logical visualization for this was a line graph where the x axis was the age groups in ascending order and the y axis was a percentage that represented the scaled deaths (i.e., what percent of deaths in this age group contained just COVID-19 conditions or other conditions). Selectivity of the colors / condition groups allows for answering this question.

This question is answered by analyzing each individual condition group. For example, we can see circulatory disease becomes more prevalent as the person gets older, respiratory issues are most prevalent in the middle age groups, younger people tend to suffer from ‘all other conditions’ more than older people. We also see that more younger people are dying of just COVID-19, the number dips in the middle ages, then resurfaces in the older groups. One hypothesis that can be drawn here is that people who do live to 85+ generally live a very healthy lifestyle and do not suffer from many of the other conditions, thus COVID-19 is ultimately what causes their death.



Q6: Because we want to analyze the moving 7 day average in increases in COVID-19 cases for specific counties, a line graph was a trivial choice for this question. In conjunction with the start date points, the user can selectively pick the different universities / counties by color to see if there is an increase immediately after the university’s AU20 semester start date. This visualization is a little more difficult because all universities have different start dates.

For all universities there was a significant increase in the 7 day average of daily increases in COVID-19 cases. What is interesting is seeing that some counties had a more delayed response than others did. BYU was included specifically because it was a private university and its (better) public counterpart, University of Utah, served as a good comparison. Interestingly BYU saw one large spike and University of Utah has seen two different spikes. Also interesting is that UNC was one of the first universities to be in the new about how COVID-19 is spreading on their campus but they did not have a large spike like the other universities until about a month after their start date.


Analyzing Spikes in 7 Day Average Daily Increases of COVID-19 Cases by University C
moving
400
300
200
100
0
Aug 2
2020
Aug 16
Aug 30
Sep 13
Sep 27
date
university
Alabama
North Carolina
Ohio State
Penn State
Texas A&M
Utah
Brigham Young
● AU20 Semester Start Dates