The Effects of COVID-19, Unemployment, and Education Level per County in the U.S.

Jessica Lynch, Jenna Castañeda, David Wiley, Azeezat Adekanmbi

Introduction

The purpose of this project was to select and merge datasets that are more valuable together than individually, analyze the data, and construct data visualizations. The datasets chosen include COVID-19 data, unemployment and labor force data, and demographic education data. All datasets contain information for US states and counties, including a specific county identifier known as a FIPS code. In theory, the combination of this data should allow us to observe how the COVID-19 pandemic has affected the labor force per county, specifically based on education level. We assume that those with a four-year degree or higher would be less affected by the COVID-19 pandemic. Therefore, we expect counties with a greater percentage of people with higher education to have lower rates of COVID-19 and lower unemployment rates. Additionally, the data will allow for the examination of COVID-19 incidence and severity in relation to geographic region and whether or not labor force (town size) acts as an indicator for COVID-19 prevalence. We assume regions with larger labor forces, and therefore a larger population, will have increased numbers of COVID-19 cases and deaths.

The potential significance of the relationships between the datasets is important to investigate, as these issues could be key to controlling COVID-19 and any future pandemics. For example, regions with a larger population and larger labor force will likely need stricter measures to contain the virus. These regions will likely require more resources to combat the virus, and struggle to do so due to pandemic related downturns in business production. In the future, resource allocation can be directed towards the more densely populated areas in an effort to control the virus, return people to work, and return business production to a normal state.

Our original and primary COVID-19 dataset was obtained from Kaggle and is sourced by the New York Times. In order to provide the most accurate information, the rest of our datasets were obtained from government-affiliated websites.

Cleaning and Merging Raw Data

In the process of cleaning the COVID-19 data, we discovered FIPS data was missing for Kansas City, Joplin, and New York City. These three cities are spread out over more than one county and therefore, were not allocated to a specific county. The greater majority of both Kansas City and Joplin are in one county. Due to this, both of these cities were assigned to the corresponding county containing the majority of the county. The dataset was then grouped to sum the counties

data for each month. New York City is comprised of 5 counties (the 5 boroughs), each with a vastly different percentage of the total population. To avoid bias, a new COVID-19 dataset was found for New York City. The NYC dataset was not cumulative, so each month had to be summed, and each month had to be added to the previous to obtain total counts at the end of each month. The county identifier was then assigned to each county. The data output was intentionally structured the same as the rest of the COVID-19 data to allow the data frames to be concatenated together.

Additionally, when cleaning the data, we found the unemployment data for Washington D.C. was missing a state. We did not want to remove DC's data, as it is a part of the continental United States, so we assigned the state equal to "DC". We also removed the Puerto Rico data, as this area is not one of the US states or in the continental US. While unemployment data lacked a full FIPS identifier, it had the FIPS split into two columns by the state and county portion of the FIPS code. To merge the datasets, we needed to concatenate these two columns to create a full fips identifier. All 2019 unemployment data was dropped from the data frame, as there were no recorded cases of COVID-19 in the US at that time. Additionally, July 2020 unemployment data is considered preliminary, and a flag column was created to identify this.

To clean the education dataset, many columns were dropped. We chose to focus on the most recently recorded education data to set a standard for education level per county. The most recent data available was from 2014-2018.

After thoroughly cleaning the data frames, we merged them together. First, we concatenated the similarly structured COVID-19 data frames together. We chose to use the unemployment data frame as our primary data frame, and to left join the others on to it. In doing so, we would start with 2020 monthly unemployment data for all counties in the US. Once a COVID-19 case is recorded in a county, it appears in the COVID-19 data frame. By left joining the unemployment and COVID-19 data frames together using the FIPS identifier and date, we can use the fillna function to populate all counties COVID-19 counts to 0 when there are no recorded cases. We then left joined the education dataset to these two datasets using the FIPS identifier.

Data Analysis

A correlation matrix was created to find significant correlations (> 0.15) between variables. Table 1 provides the correlation matrix for our data. The duplicated data is labeled in gray. The data labeled in blue show the 14 pairs that are significantly correlated.

	Labor_Force	Unemployment_Rate	Preliminary_Unemp_Data	cases	deaths	Less_HS_Dip_pct	HS_Dipl_pct	Some_col_pct	Bach_plus_pct
Labor_Force	1	0.07913337435	0.0003663609635	0.5575670229	0.4855215754	-0.06320267336	-0.3260244866	-0.1048172061	0.3488226793
Unemployment_Rat	0.07913337435	1	0.03488607837	0.1427718075	0.1723946589	0.08674020592	0.05492459129	-0.1029787569	-0.04327272781
Preliminary_Unemp	0.000366360963	0.03488607837	1	0.1288648695	0.08232480867	-1.91E-17	-1.44E-17	1.18E-18	-2.19E-17
cases	0.5575670229	0.1427718075	0.1288648695	1	0.8432292316	-0.008226508454	-0.153812449	-0.08084637071	0.1674106727
deaths	0.4855215754	0.1723946589	0.08232480867	0.8432292316	1	-0.0300967975	-0.1432947861	-0.1052002051	0.1875328735
Less_HS_Dip_pct	-0.06320267336	0.08674020592	-1.91E-17	-0.00822650845	-0.0300967975	1	0.2501161919	-0.479834241	-0.597758997
HS_Dipl_pct	-0.3260244866	0.05492459129	-1.44E-17	-0.153812449	-0.1432947861	0.2501161919	1	-0.2818754193	-0.7747350772
Some_col_pct	-0.1048172061	-0.1029787569	1.18E-18	-0.08084637071	-0.1052002051	-0.479834241	-0.2818754193	1	-0.01435654894
Bach_plus_pct	0.3488226793	-0.04327272781	-2.19E-17	0.1674106727	0.1875328735	-0.597758997	-0.7747350772	-0.01435654894	1

Table 1 - Correlation Matrix

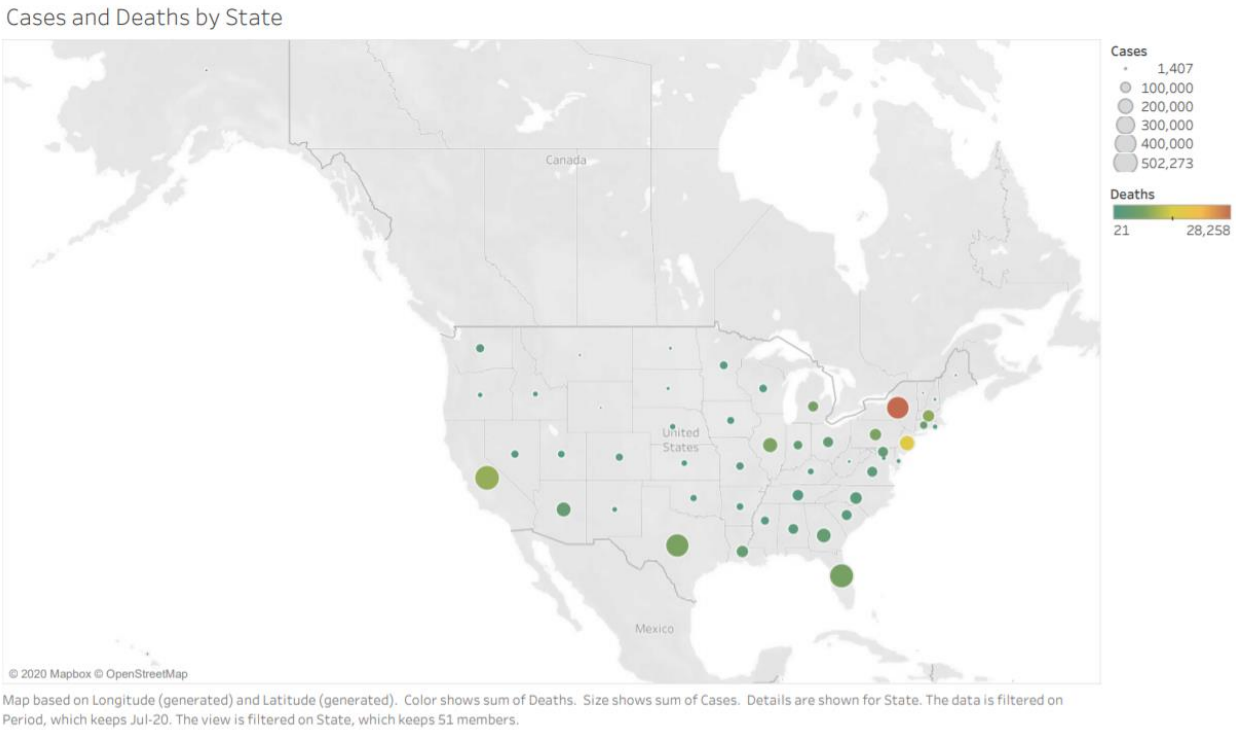
As seen in the table, labor force is strongly positively correlated with COVID-19 cases and deaths. This means that the larger the labor force, the higher the number of cases and related deaths. This agrees with our assumption that regions with higher labor forces will have an increased number of cases and deaths due to COVID-19.

The correlation matrix also demonstrates a negative correlation between cases and the percentage of people with a high school diploma. It also depicts a and a positive correlation between cases and the percentage of people with advanced degrees. The correlation matrix shows that as the number of people with high school diplomas increases, COVID-19 cases decrease. Likewise, as the number of people with advanced degrees increases, so do COVID-19 cases. Given this information, we struggled to associate COVID-19 incidence per county with education level in relation to our proposed assumption/hypothesis. While two of four of the education subsets show some sort of correlation, they are not very strongly correlated. We believe that this relation may have more to do with the relationship between labor force and education level, as opposed to education level and cases. This is evidenced by the labor force being strongly positively correlated with COVID-19 cases and related deaths, and also being negatively correlated with high school diploma percentage and positively correlated with advanced degree percentage.

Finally, the highest correlation is between cases and deaths. This information is intuitive and likely valid because the more people that are diagnosed with COVID-19, the higher the number of people likely to die from COVID-19. In other words, if people are not catching the virus, then they are not going to die from it.

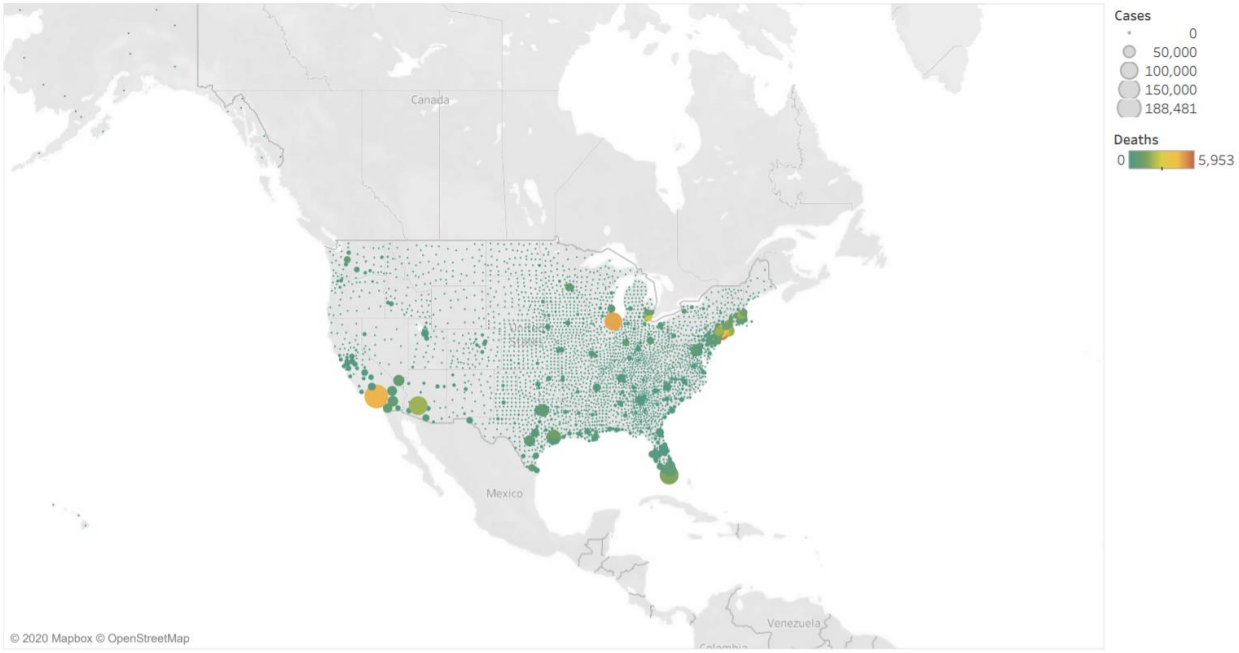
We found the labor force, number of cases, and number of deaths variables more valuable combined than separate. In addition, we found cases and deaths by county across the U.S. to be valuable when combined and visualized, as we could see the effect of COVID-19 in different regions across the country. As a whole, this merged data is valuable, as we are able to present some insight on the effects of COVID-19 and potential ways to mitigate the effects of future pandemics. Regions with a larger labor force have higher levels of COVID-19 incidence, including both cases and deaths. Businesses in these locations could be affected by COVID-19 in two ways. First, there could be an increased demand for supplies, such as masks and other PPE, food, and basic essentials. Second, as COVID-19 deaths and the unemployment rate increase, there could be a decrease in business for businesses that were forced to shut down due to the pandemic. If early on, supplies to mitigate the pandemic were allocated to and available in cities that needed the resources the most, the number of cases and deaths could potentially be decreased. If that was decreased and resources were consistently available, the effects of the pandemic could be controlled and alleviated to allow businesses to remain open.

Once the data was fully merged, visual representations were constructed to reveal patterns and emphasize possible correlations.



Graph 1 - Cases and Deaths by State

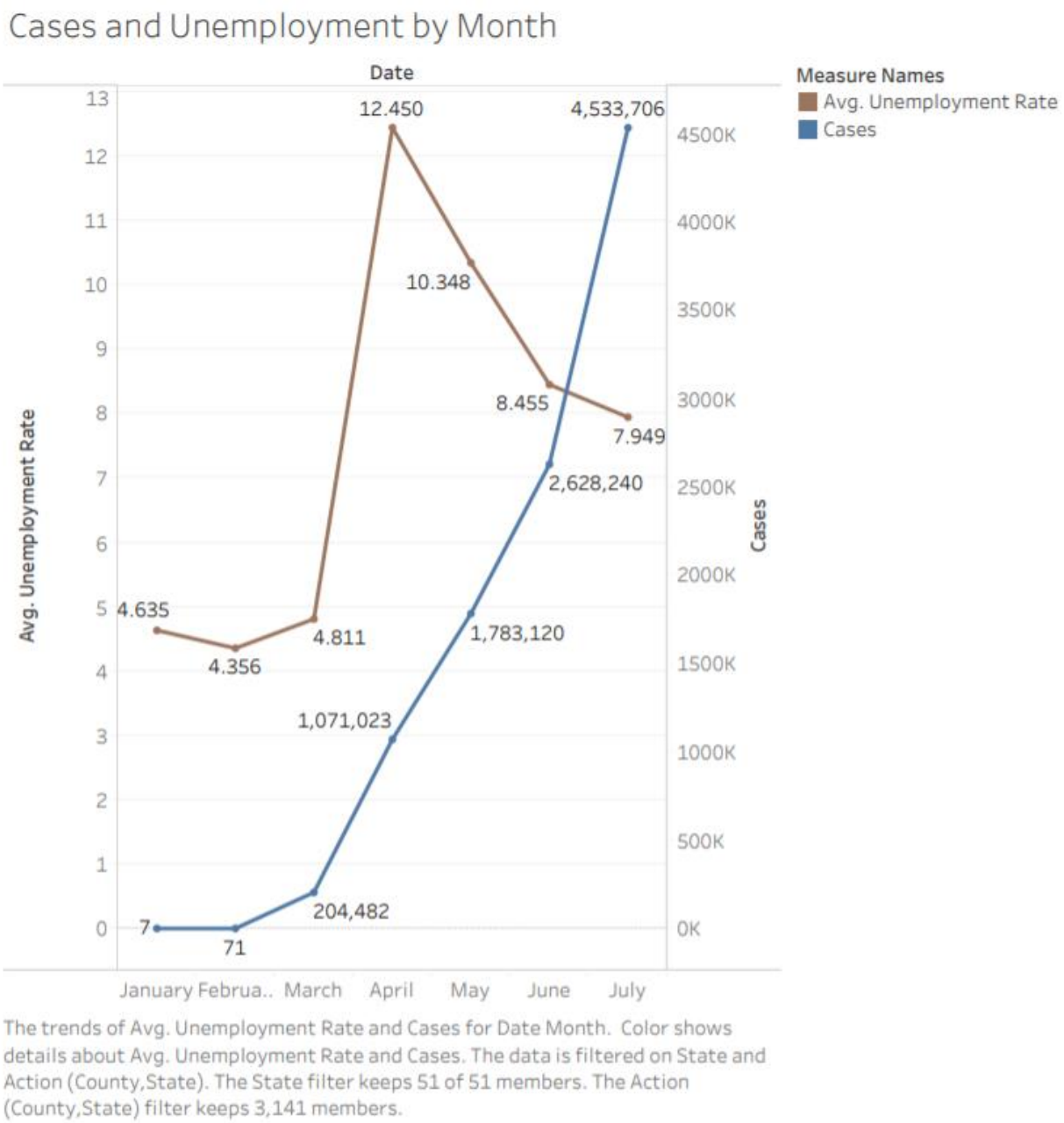
This graph demonstrates cases and deaths due to COVID-19 by each state in the contiguous United States. Based on the label size, you can see that California, Texas, Florida, New Jersey, and New York have the most cases as of July 2019. Based on the color spectrum, you can see that New York, New Jersey, and California have had the most deaths. The geographical map allows the user to easily interpret the information of cases and deaths by state.



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Deaths. Size shows sum of Cases. Details are shown for State and County. The data is filtered on Period and Action (MONTH(Date),State). The Period filter keeps Jul-20. The Action (MONTH(Date),State) filter keeps 357 members. The view is filtered on State, which keeps 51 of 51 members.

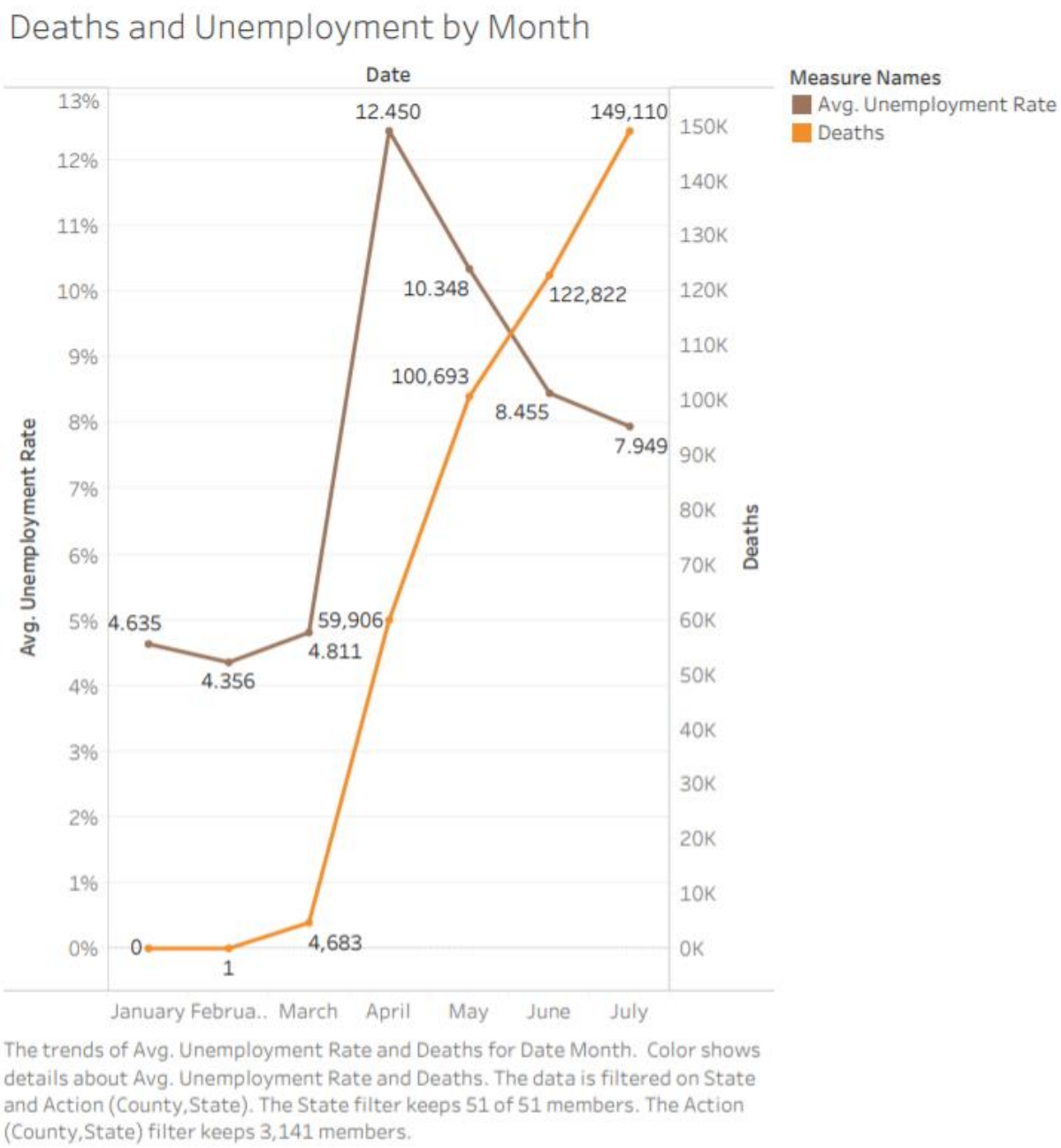
Graph 2 - Cases and Deaths by County

This graph depicts COVID-19 cases and related deaths by each county in the United States. Based on the size of the labels, the regions with the most cases can be seen in southern California, southwestern Arizona, the large Texas cities, most of Florida, northern Illinois, and much of the New England coast. Based on the color spectrum, the regions with the most deaths are those in southern California, southwestern Arizona, northern Illinois, southern Florida, and the New England coast nearest New York City and New Jersey. This graph draws attention to the fact that COVID-19 tends to be more prevalent in regions commonly known to be more densely populated, such as cities. The geographical graphs in Graph 1 and Graph 2 allow the readers to natively process the data, given the user likely has prior knowledge of the U.S. map. Additionally, Graph 2 invokes natural processing by demonstrating regions with more cases are more concentrated with circles and color.



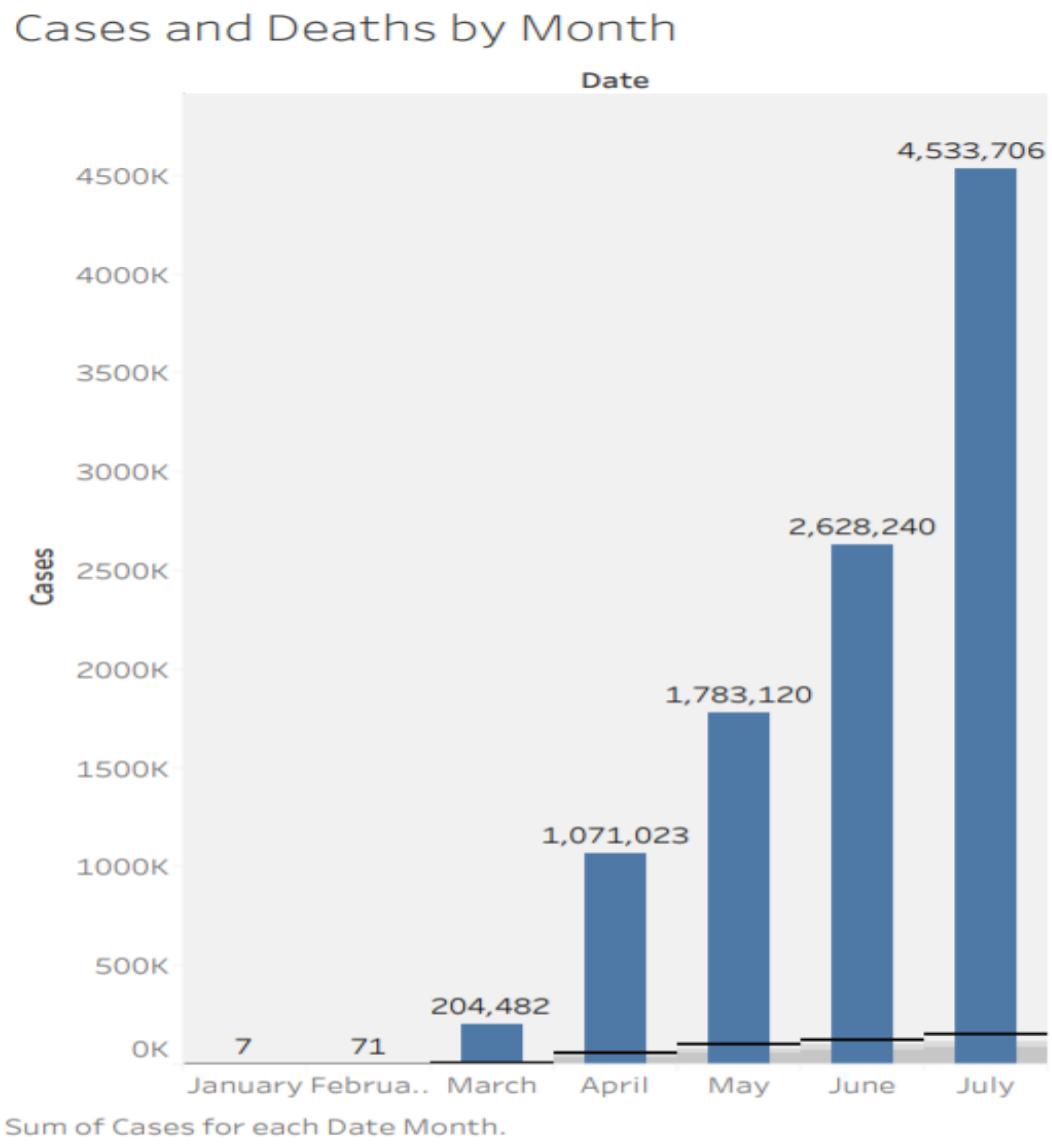
Graph 3 - Unemployment and Cases by Month

Graph 3 depicts the total number of COVID-19 cases in the US in relation to the average unemployment rate over time. We can see that in April, both the unemployment rate and number of COVID-19 cases increased significantly. This is expected, as when the virus first became prevalent, many businesses closed their doors. We can see that over time, the unemployment rate starts to decrease, but overall, it is still much higher than it was pre-pandemic.



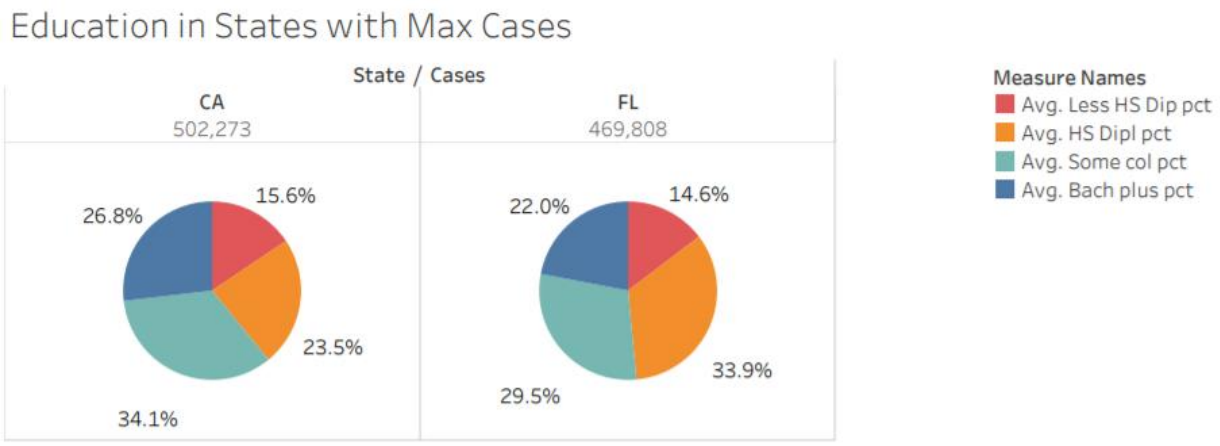
Graph 4 - Unemployment and Deaths by Month

Graph 4 depicts the total number of COVID-19 deaths in the US in relation to the average unemployment rate over time. We can see that in April, both the unemployment rate and number of COVID-19 related deaths increased significantly.



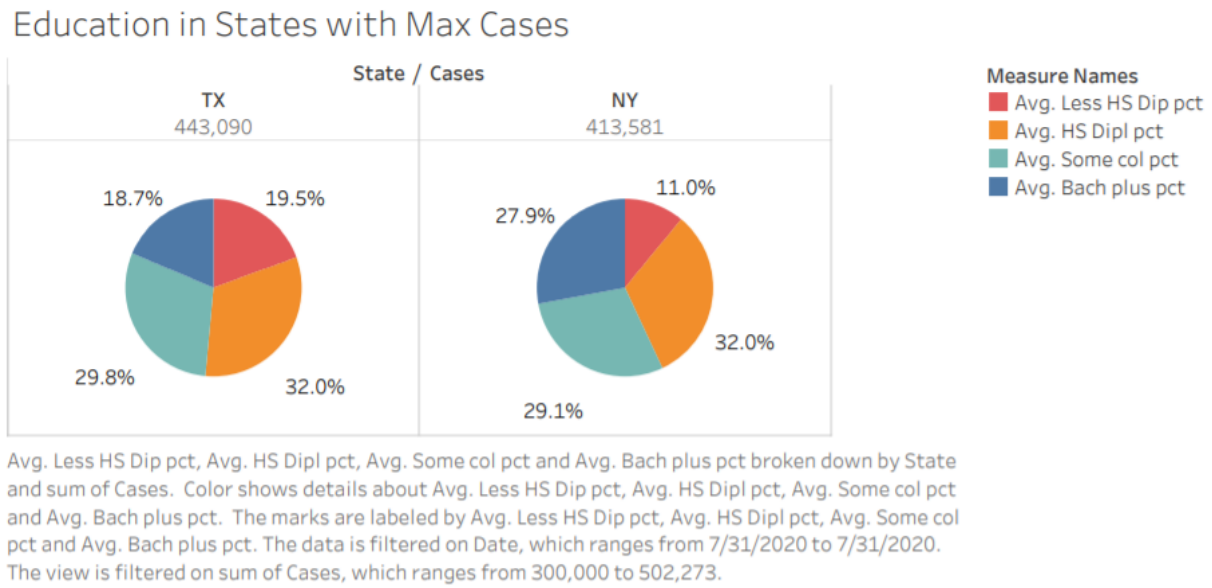
Graph 5 - Cases and Deaths by COVID-19 by Month

Graph 5 shows the relationship between COVID-19 cases and related deaths. As time progresses, the cases increase dramatically. (This can be demonstrated by the blue bars.) Additionally, as time progresses, the deaths increase. (This can be demonstrated by the black lines.) This demonstrates a positive correlation between cases and deaths. Graph 5 is a simple graph that may invoke natural processing because it uses both a simple bar and a simple line graph. The reader will be able to see the increase in cases and deaths over time.



Graph 6a - Education in States with the Most Cases (part 2 of 2)

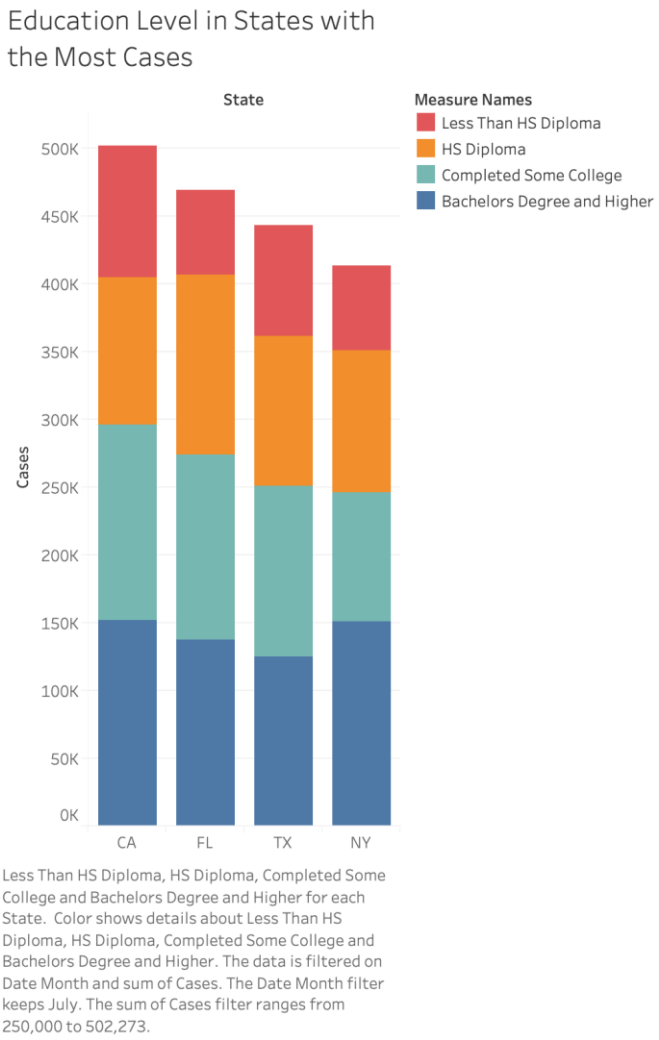
Graph 6a shows the two states with the highest number of COVID-19 cases in the U.S. and their corresponding education levels. As seen in the pie charts, both states have a large number of individuals with some college and individuals with only a high school diploma.



Graph 6b - Education in States with the Most Cases (part 2 of 2)

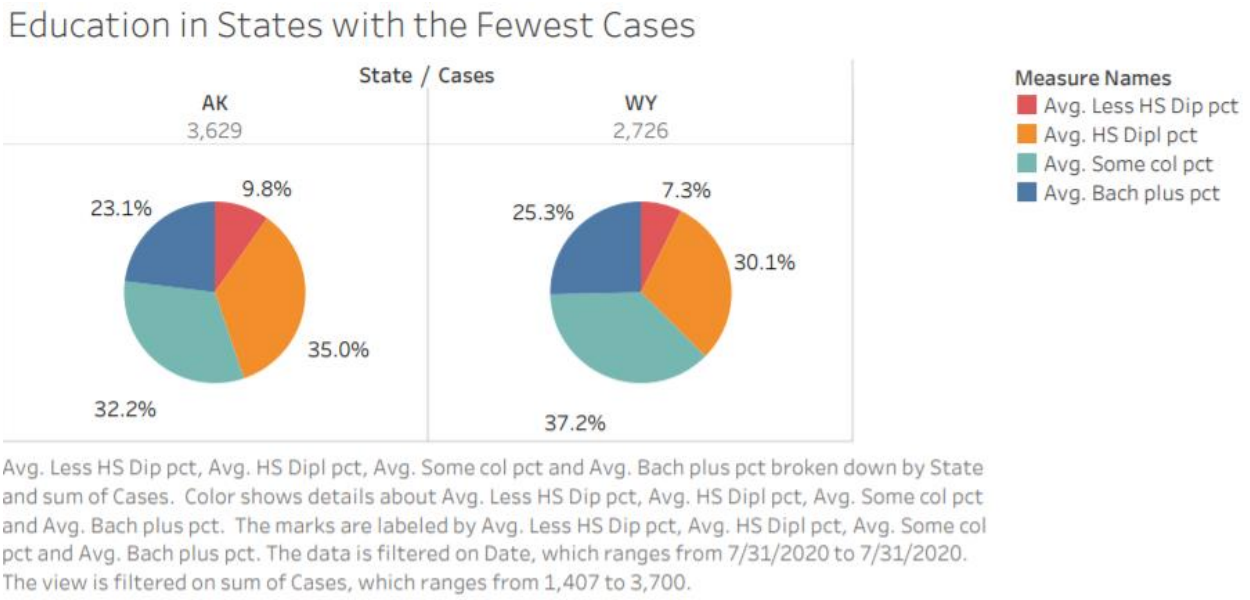
Graph 6b shows the two states with the third and fourth highest number of COVID-19 cases in the U.S. and their corresponding education levels. Compared to the pie charts in Graph 5a, the pie charts in Graph 6b also demonstrates a great number of individuals with only a high school diploma and some college. There may be a relationship between states with a high percentage

of people who have at least a high school degree and the amount of COVID-19 cases in the region. However, this may be a weak correlation, since there is noticeable variability in these graphs.



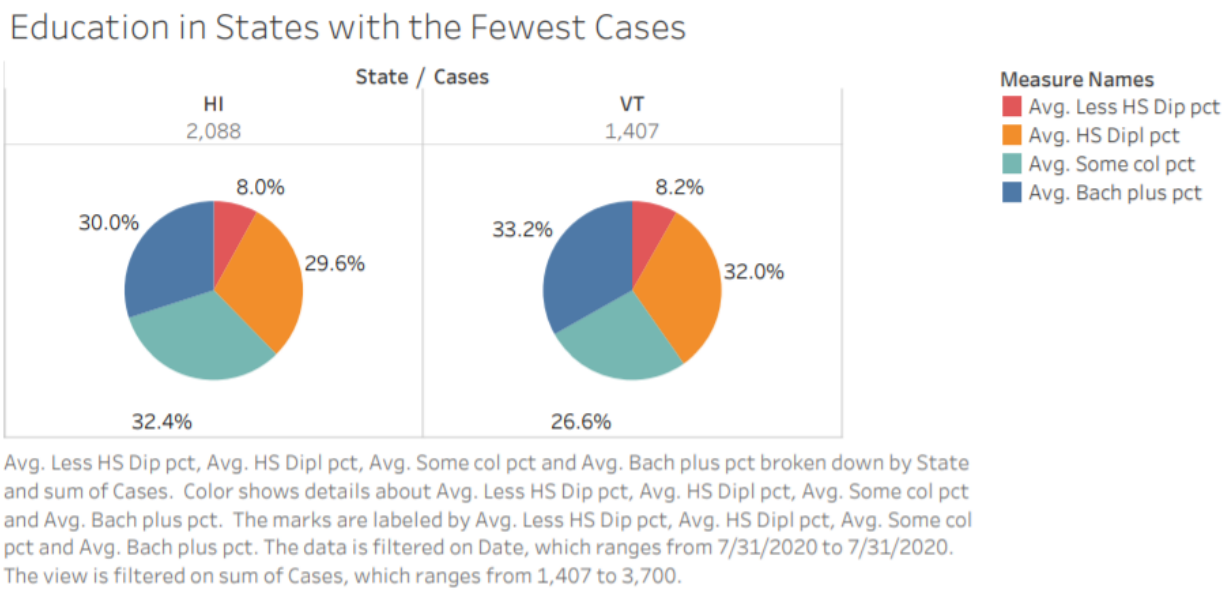
Graph 7 - Education in States with the Most Cases

Graph 7 depicts the states with the highest number of COVID-19 cases and the state's education level proportion. This graph depicts the same information as graph 6a/b but visualizes the relationship between education level and cases counts more clearly. As can be seen, education level is relatively constant for each of these, which is in line with the low correlation seen in the correlation matrix (Table 1).



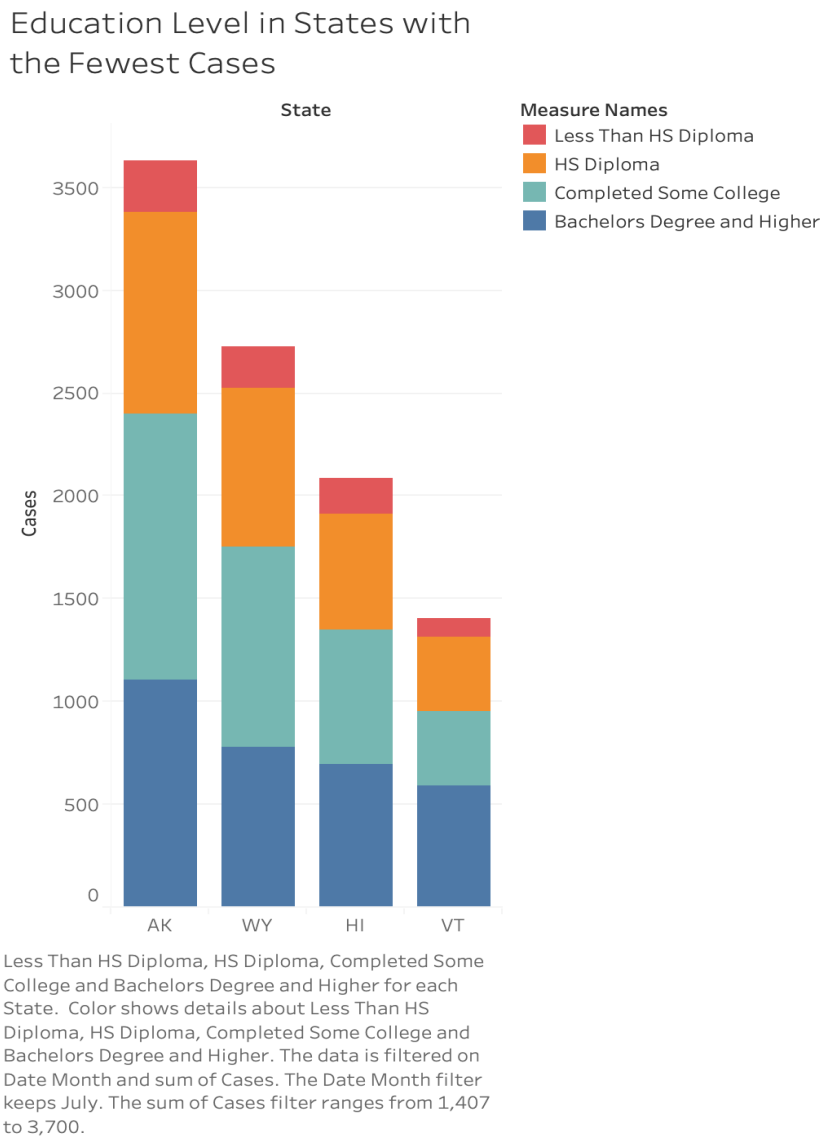
Graph 8a - Education in States with the Fewest Cases (part 1 of 2)

Graph 8a shows two of four of the states with the lowest COVID-19 numbers in the U.S. and their relationship with education levels. In Alaska and Wyoming, there are a great number of people with only a high school diploma and people with some college education.



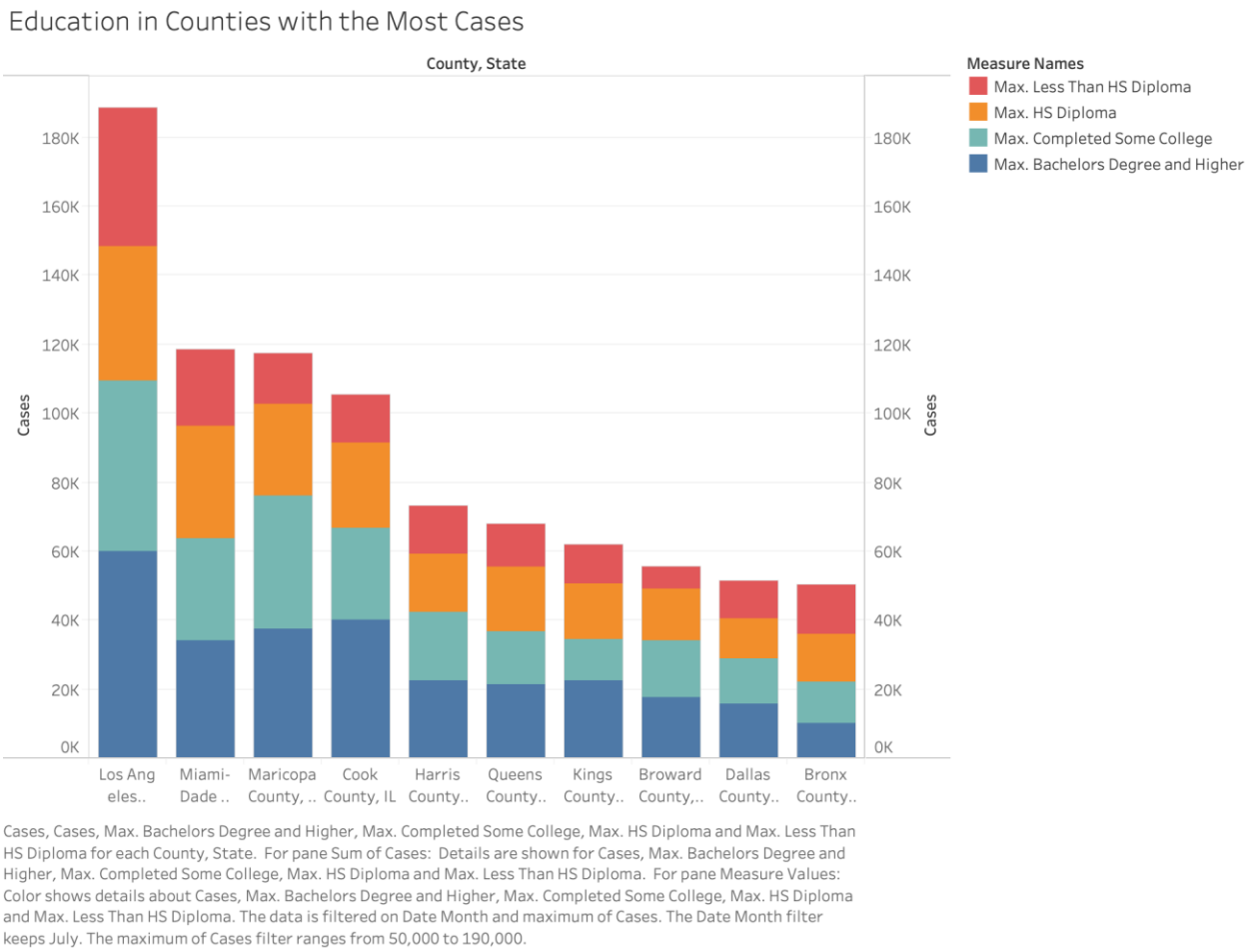
Graph 8b - Education in with the Fewest Cases (part 2 of 2)

Graph 8b shows the two states with the lowest number of COVID-19 cases in the U.S, and their relationship with education levels. In Hawaii and Vermont, the majority of the population and assumed cases are distributed amongst those with only a high school diploma, those with some college, and those with advanced degrees. The additional variability in these graphs shows that it is likely that there is low correlation between these variables. Additionally, states with the highest and the lowest number of COVID-19 cases had similar percentages of each education level, contributing to the assumption that there is a low correlation between these variables.



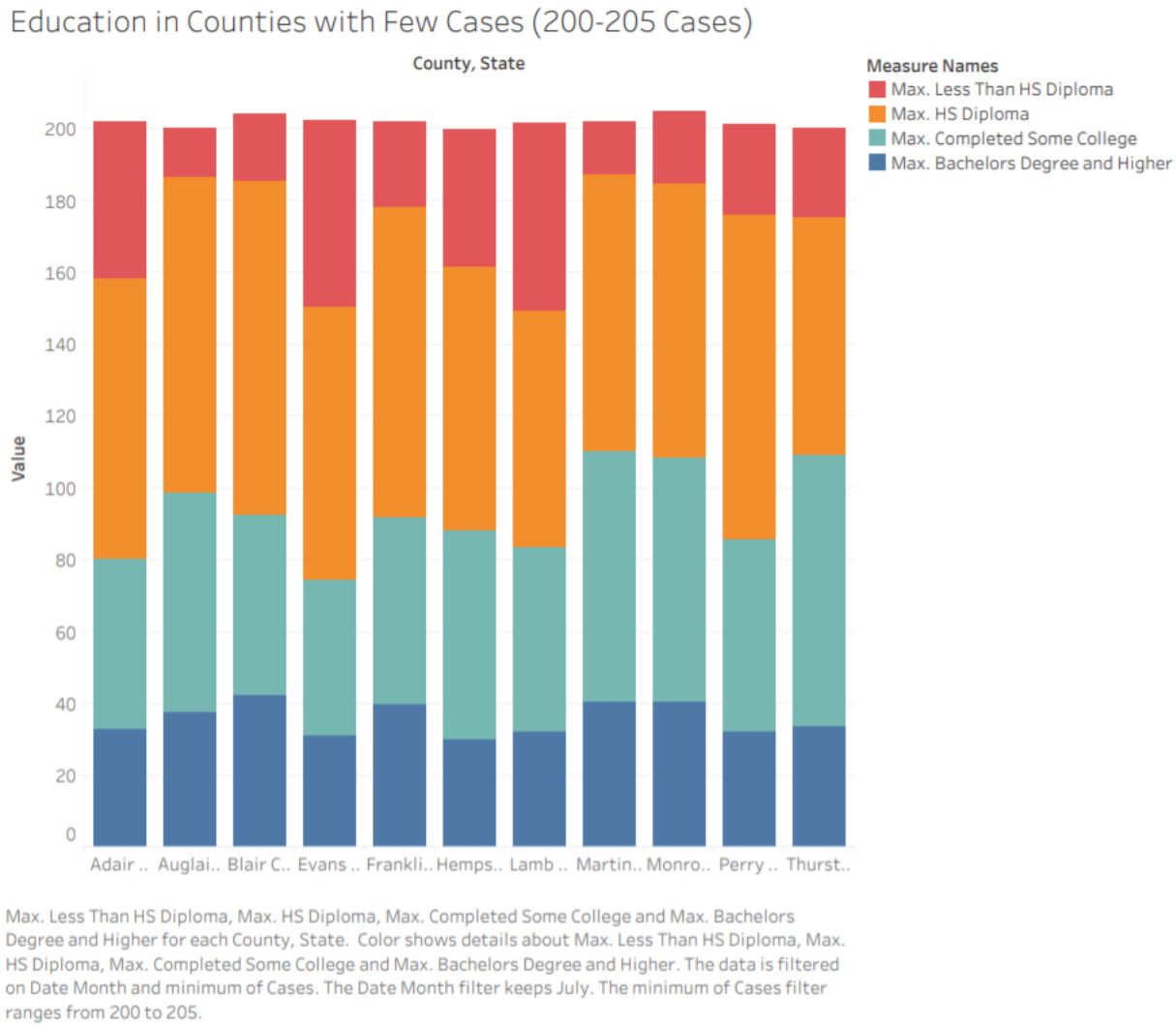
Graph 9 - Education in States with the Least Cases

Graph 9 depicts the relationship between the U.S. states with the lowest number of COVID-19 cases and education levels. This graph displays the same information as graph 8a/b but visualizes the relationship between education and cases counts more clearly. When compared with graph 7, it can be seen that states with a lower percentage of those with less than a high school diploma had fewer cases. However, these variables are not significantly correlated, and therefore, no assumptions should be made. Overall, education level and cases count's across states are varied, which line up with the low level of correlation displayed in the correlation matrix (Table 1).



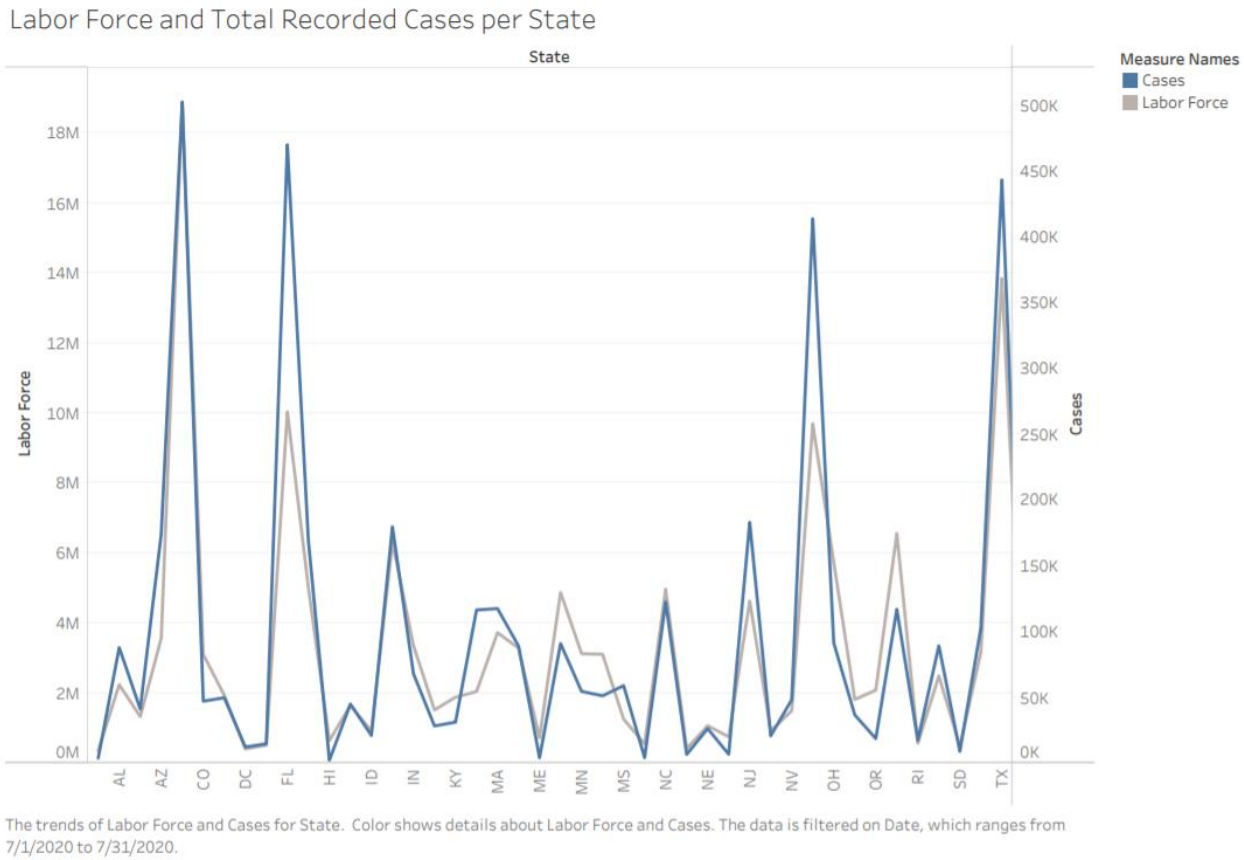
Graph 10 - Education Levels in Counties with the Most Cases

Graph 10 shows the relationship between U.S. counties with the most COVID-19 cases and education levels. The graph shows Los Angeles county with the highest number of cases, but the percentage of education levels appears to be relatively evenly distributed. The education in the other counties vary, hinting at a weak correlation.



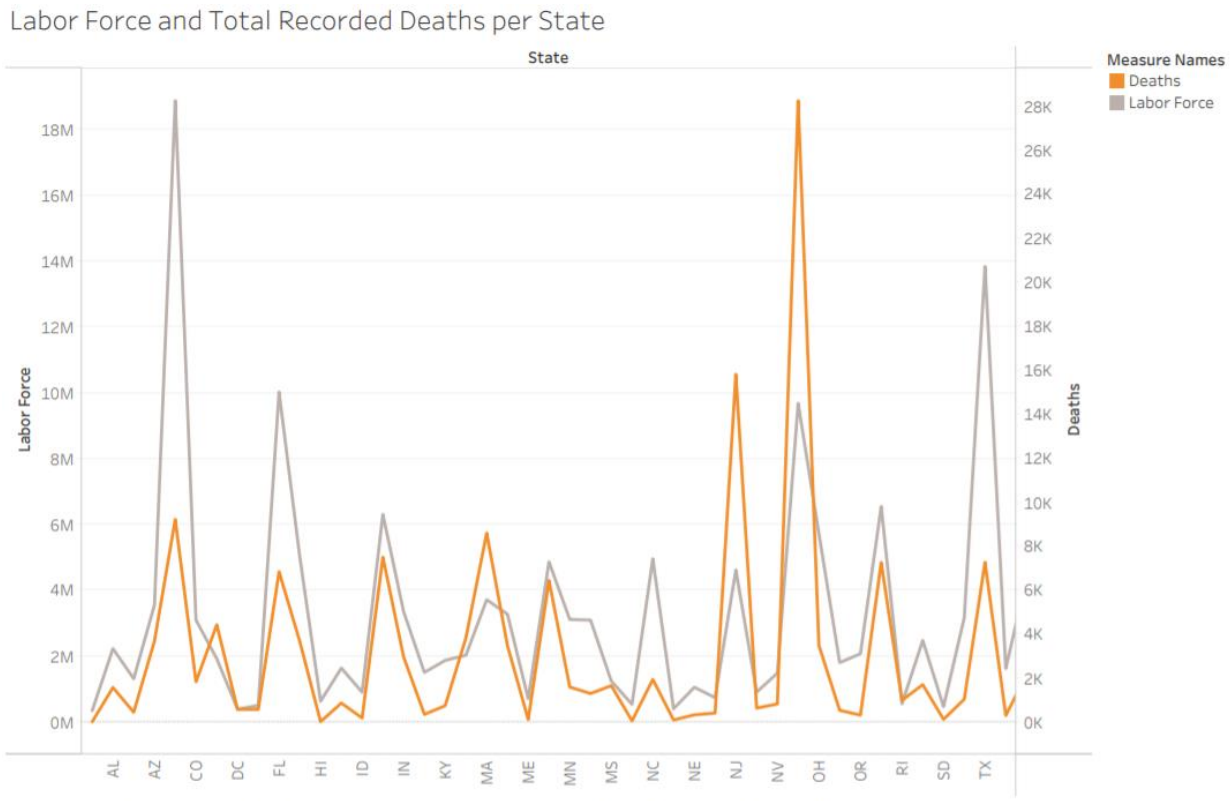
Graph 11 - Education in Counties with Few Cases

Graph 11 shows the relationship between U.S. counties with some of the lowest numbers of COVID-19 cases and education levels. The counties selected for have between 200-205 COVID-19 cases. Given the structure of the graph, counties with some cases must be selected for in order to show the corresponding relationship with education level. As shown in the graph, all eight counties have about the same number of cases with similar percentages of education levels. There could be a correlation between education levels and COVID-19 cases in the counties with the fewest number of cases. Although graphs 6a-11 do not perfectly adhere to the concept of natural processing due to the combination of colors, stacked bars, and pie charts, we did keep the colors of education levels constant across the graphs to aid in the reader's visual processing.



Graph 12 - Labor Force and Total Recorded Cases per State

Graph 12 demonstrates the high correlation (0.56) between labor force and COVID-19 cases across states. As you can see from the graph, states with high labor forces also tended to have a large number of reported COVID-19 cases. This backs up the assumption that the larger the labor force, the more COVID-19 cases in that area.



Graph 13 - Labor Force and Total Recorded Deaths per State

Much like Graph 12, Graph 13 shows the high positive correlation (0.49) between labor force and deaths from COVID-19. The graph shows that states with a large labor force also had a higher number of recorded deaths.

Conclusion

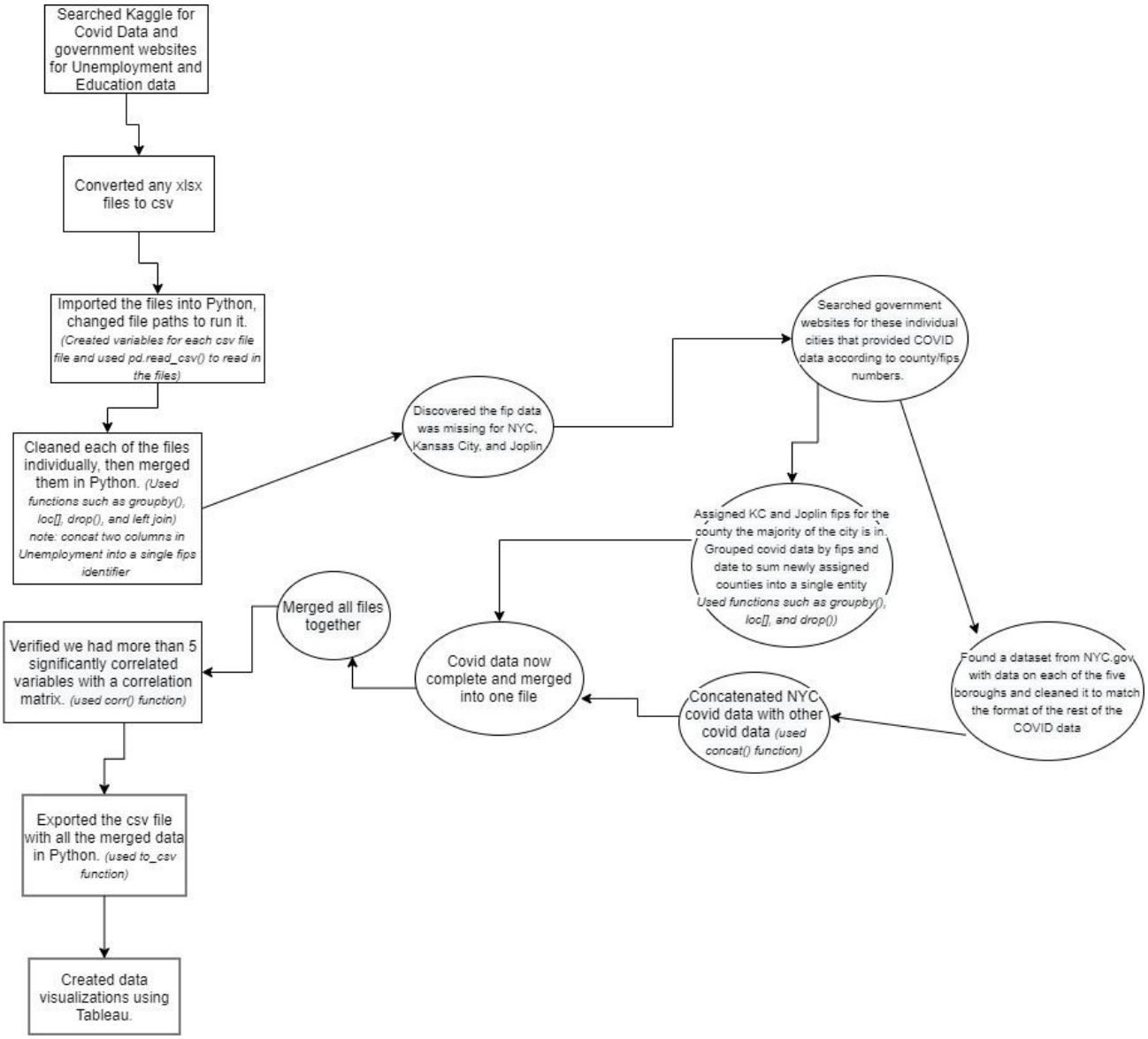
We obtained data for the prevalence of COVID-19 per county by month, unemployment rates per county by month, and education level per county in the US. We then cleaned and merged these datasets to evaluate the interactions between each.

We assumed that regions with larger labor forces would have a higher number of cases and deaths due to COVID-19. We were able to find evidence to support this assumption, as our correlation matrix depicted a strong positive correlation between labor force and both cases and deaths. In other words, as the labor force increases, so does the number of COVID-19 cases and deaths.

We also assumed that those with a four-year degree or higher would be less affected by COVID-19 than other education levels. Though we were not able to prove or disprove this statement, we did learn that there is little correlation between all four education levels and COVID-19 case numbers. There is a correlation between COVID-19 cases and related deaths for those with an advanced degree (bach_plus_pct) and a correlation between labor force and those with an advanced degree. However, not all education subsets supported this assumption, and therefore, further investigation would be needed to conclusively support this. Additionally, this correlation can potentially be explained by the strong correlation between labor force and cases.

Through the combination of these originally unrelated data sets, we were given the opportunity to provide insight on how to mitigate the effects of the pandemic. The correlation between labor force and COVID-19 cases and deaths could affect businesses positively or negatively. This depends on whether demand for a business's supply increases or decreases during a pandemic. Businesses such as grocery stores will likely experience demand increase as the population spends more time at home, while businesses such as movie theaters and bowling alleys will see their demand decrease for the same reason. Businesses with the capacity to adapt to an ever-changing pandemic related situation are more likely to remain afloat during this time. Additionally, at the beginning of the pandemic, resources can be allocated to the cities with the largest labor force to mitigate the effects of the pandemic before the situation worsens. Overall, the ability to combine and analyze multiple unrelated datasets in an effort to better see how COVID-19 affects the US may provide useful information to combat and alleviate the effects of possible future pandemics.

Appendix A: Flow diagram



This flow diagram illustrates the steps used to obtain data, munge and merge the data, and create data visualizations.

Appendix B: Code Instructions

1. Unzip the file that includes the code and data files. Move the files to the desired location on your personal machine, keeping the same filenames for the data sets. The necessary raw data files to run the code are in the raw datasets folder.
    a. We strongly recommend using the CSV's in the zip file provided. However, the user can pull the data from the corresponding website and export files to CSV. If this is done, there is no guarantee the data structure did not change from when used datasets were accessed
        i. 9/14/20 (COVID-19, unemployment, education) and 10/2/20 (NYC COVID-19)
2. Using your preferred python environment, set the path to that where the raw files have been saved. This is also where the output will be saved.
3. Within the python environment open the .py script included in the extracted zip file. Open the python script, the list of variables that can be manipulated are those at the beginning. A list of those variables is below for reference.
    a. The "filepath" variable is the only necessary change a user must make if using the provided datasets.

Variable	Default	Description
filepath	***Needs to be changed to a personal directory.**	The directory where the necessary files are located as well as where the data frame files will be saved.
raw_covid_file	"covid_data.csv"	Filename of the raw COVID data. Filename may be different if pulled from the website and exported to CSV by the user
raw_nyc_covid_file	"boroughs-case-hosp-death.csv"	Filename of the raw NYC COVID data. Filename may be different if pulled from the website and exported to CSV by the user
raw_unemployment_file	"unemployment.csv"	Filename of the raw unemployment data. Filename may be different if pulled from the website and exported to CSV by the user
raw_education_file	'education.csv'	Filename of the raw education data. Filename may be different if pulled from website and exported to CSV by the user
covid_cleaned_output	"covid_cleaned.csv"	Filename of the newly cleaned COVID data to be saved from the Python script.
nyc_covid_cleaned_output	"nyc_covid_cleaned.csv"	Filename of the newly cleaned NYC COVID data file to be saved

		from the python script.
unemployment_cleaned_output	"unemployment_clean.csv"	Filename of the newly cleaned unemployment data file to be saved from the python script.
education_cleaned_output	"edu_cleaned.csv"	Filename of the newly cleaned education file to be saved by the python script.
fully_merged_output	"completed_merged_data frame.csv"	Filename of the final merged data frame containing all previously separate datasets. This is the final output from the python script.
correl_matrix	"correl_matrix.csv"	Filename of the correlation matrix of all variables included in the merged data frame.

4. Once all necessary variables are changed, save the script. Run the newly saved .py python script with new variables.
5. Import the completed_merged_data frame into a data processing software such as Tableau to create visualizations.

Appendix C: Data Sources

New York Times, N. (2020). US counties COVID 19 dataset. Retrieved September 14, 2020, from https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset

NYC Health, N. (2020). COVID-19: Data. Retrieved October 02, 2020, from https://www1.nyc.gov/site/doh/covid/covid-19-data.page

US Dept of Agriculture, U. (2020). County-level Data Sets. Retrieved September 14, 2020, from https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

US Dept of Labor, U. (2020). Local Area Unemployment Statistics Home Page. Retrieved September 14, 2020, from https://www.bls.gov/lau/

COVID-19 -
Source: https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset
Direct Link: https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

NYC COVID-19 -
Source: https://www1.nyc.gov/site/doh/covid/covid-19-data.page
Direct Link: https://github.com/nychealth/coronavirus-data/blob/master/boro/boroughs-case-hosp-death.csv

Unemployment -
Source: https://www.bls.gov/lau/#tables
Access by: go to County Data > Table (Labor force data by county) > Zip file

Education -
Source: https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/
Access by: educational attainment file