Matthew Shinder

Capstone Two - Project Report

Introduction

COVID-19 is a contagious disease that was caused by the SARS-CoV-2 coronavirus. While first known to originate in Wuhan in December of 2019, the virus quickly spread around the world creating a global pandemic. The goal of this project is to look at trends of COVID-19 data, specifically between each US county.

Datasets

We will use the US counties COVID 19 dataset (from Kaggle) which contains COVID-19 case numbers dating from as early as February of 2020 and up to May of 2022. After data wrangling and modeling our COVID data, we can look into the US Census Bureau dataset to get demographic information by county. Connecticut data was lost from the main demographic file so an outside file was used to gather the data specifically for counties within that state.

Links:

https://www.kaggle.com/datasets/fireballbyedimyrnmom/us-counties-covid-19-dataset?resource=download

https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

https://data.census.gov/table?q=P6&t=Race+and+Ethnicity&g=010XX00US$0500000&y=2020&tid=DECENNIALDHC2020.P6

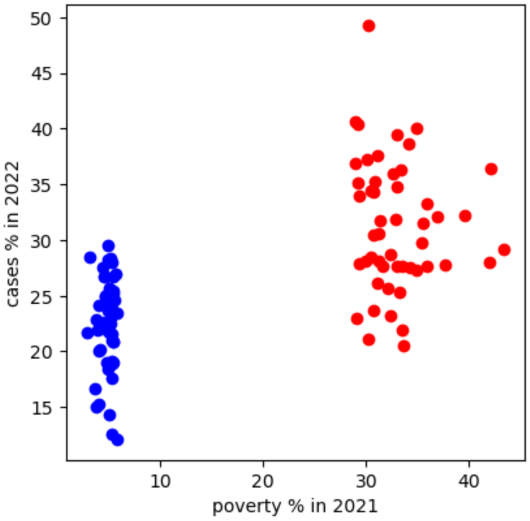
https://portal.ct.gov/DPH/Health-Information-Systems--Reporting/Population/Annual-Town-and-County-Population-for-Connecticut

Cleaning/Wrangling

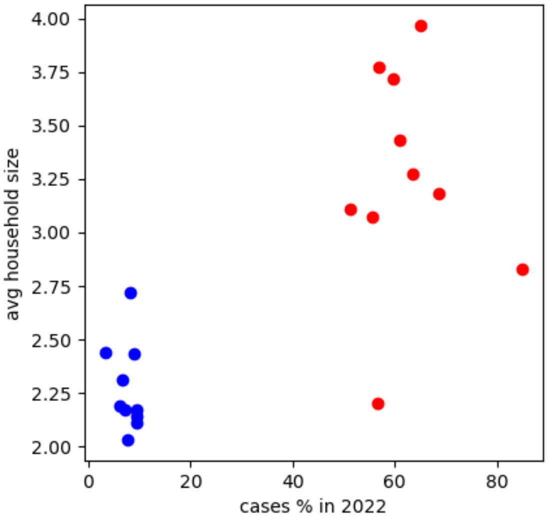
To clean the data, numerous steps were made to remove many values as we had over 2 million data rows of data in our COVID-19 cases dataset. Rows that had invalid data such as null values were removed as well all data associated with US territory locations such as Puerto Rico and the Virgin Islands. When it came to the demographics datasets, the main issue was the naming of columns and which columns from our data source that we would want. The hardest problem in all of these steps was the combination of all data sources. First, we had to make sure all the counties' demographic data sources matched by name and state in all data sources before merging. When doing this step, Connecticut as stated above was missing data, so an outside file was used in order to fill in the missing values as missing the whole state of Connecticut could be detrimental in the findings. The data from the newly formed demographics data frame was then checked with the COVID 19 data frame in order to only include counties that were listed in both. Those that were not, were removed from their respective dataset (very small amount). New York City was combined as the COVID 19 dataset listed the county as a whole vs. the demographic datasets that had them listed by NYC boroughs such as Queens and Brooklyn.

EDA

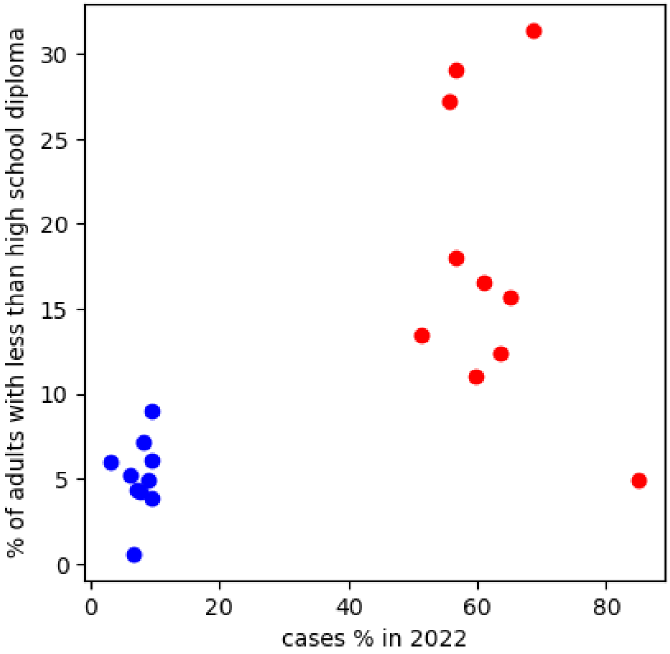
Two approaches were taken in order to find trends in our data. The first approach was to look at counties that have high covid rates and then compare their demographic stats whereas the second approach would do the same but vice versa. Finding out the most populous COVID 19 case counties proved to just be the most populous counties in the US. To combat this, a new column was created in order to show the COVID 19 case versus the total population of the county, referred to as covid case percentage. Three dates (one per year) were chosen to separate the data based on key events; however, the covid case percentage was an accumulated total by said date. The 2020 date was 11/13/20 as every US County had reported at least one case of COVID 19. For 2021, 6/1/21 was a good pace for COVID 19 as the reported cases had begun to dip down heavily due to the creation and use of the vaccine. Our 2022 was 5/13/22 as that was the end of the dataset given to us. The graphs below demonstrate either approach.



This first graph shows when the counties are sorted by highest and lowest poverty percentage counties, what their covid case percentage was in 2022.

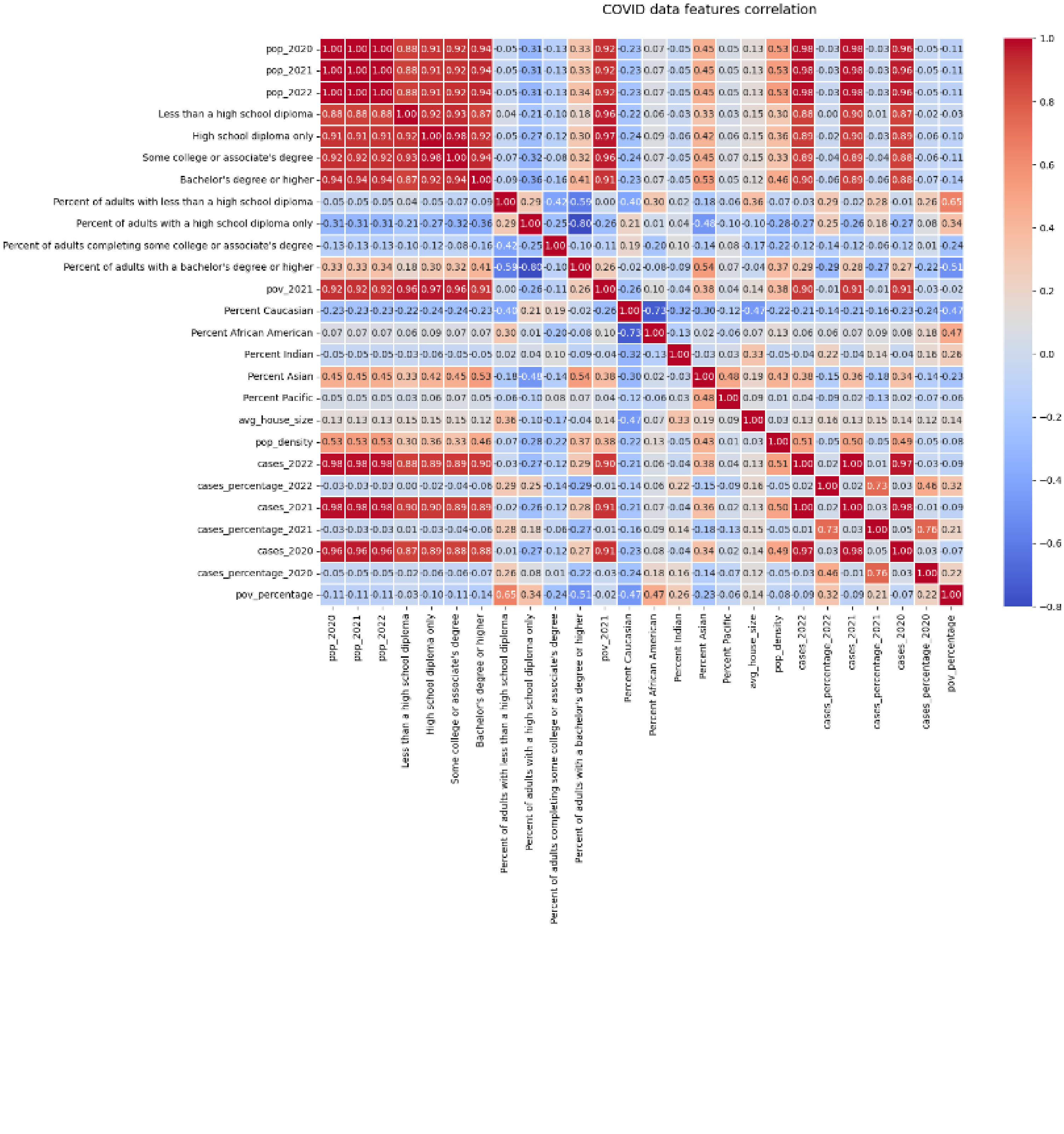


This second graph shows the same trend for the average household size in each county.

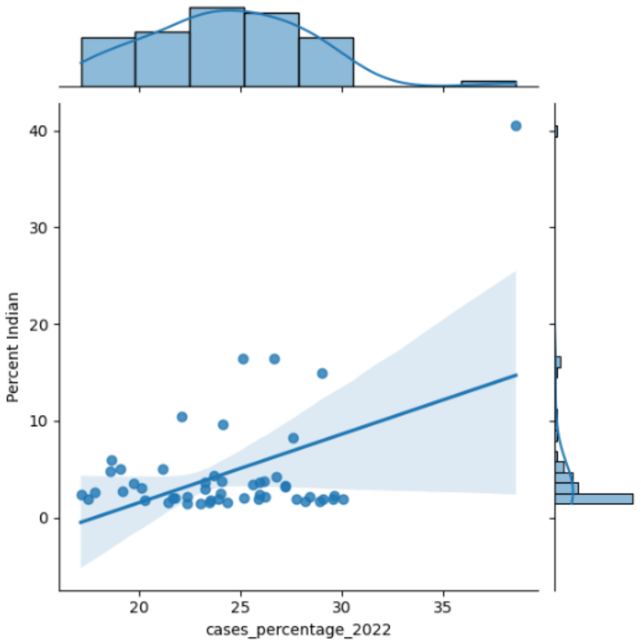


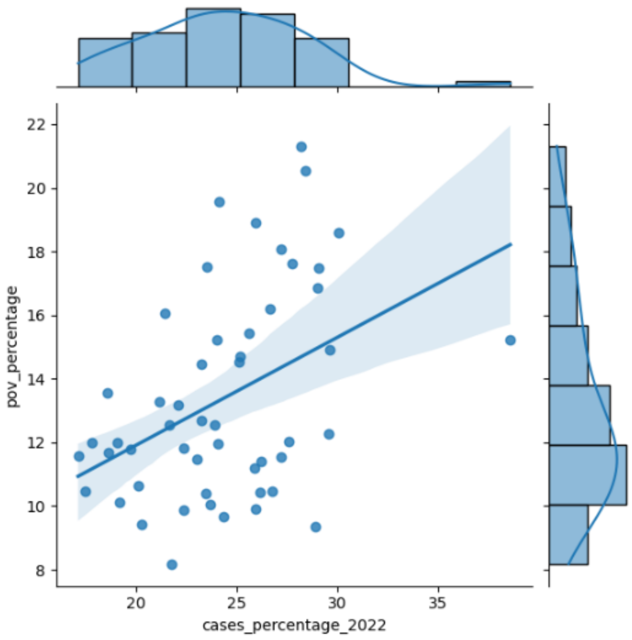
This last graph was sorted by COVID case percentage and then compared with an education stat.

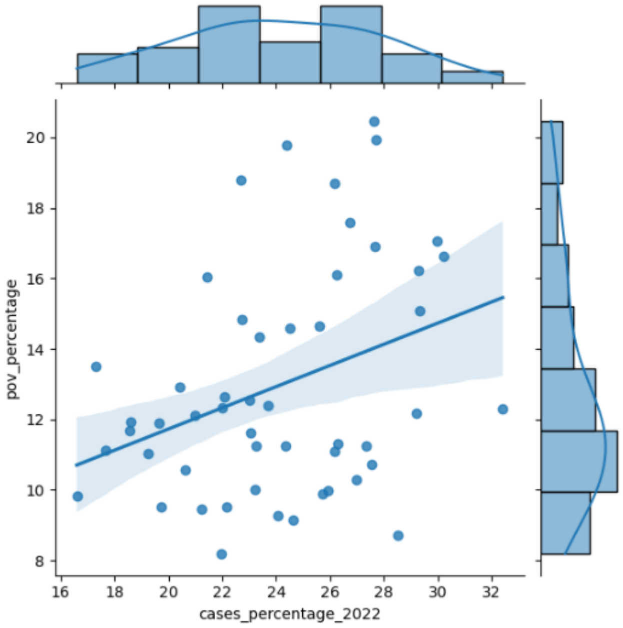
To get a better look at these trends, heatmaps were generated but, to no chagrin, the data did not lead to great results or positive / negative correlations rather regressed to showing no trend. The heatmap on the next page shows most values near 0 where the highly correlated values (dark red or dark blue) are correlated just based on population numbers compared to the percentage columns which is what we are most interested in.

COVID data features correlation
pop_2020 1.00 1.00 1.00 0.88 0.91 0.92 0.94 -0.05 -0.31 -0.13 0.33 0.92 -0.23 0.07 -0.05 0.45 0.05 0.13 0.53 0.98 -0.03 0.98 -0.03 0.96 -0.05 -0.11
pop_2021 1.00 1.00 1.00 0.88 0.91 0.92 0.94 -0.05 -0.31 -0.13 0.33 0.92 -0.23 0.07 -0.05 0.45 0.05 0.13 0.53 0.98 -0.03 0.98 -0.03 0.96 -0.05 -0.11
pop_2022 1.00 1.00 1.00 0.88 0.91 0.92 0.94 -0.05 -0.31 -0.13 0.34 0.92 -0.23 0.07 -0.05 0.45 0.05 0.13 0.53 0.98 -0.03 0.98 -0.03 0.96 -0.05 -0.11
Less than a high school diploma 0.88 0.88 0.88 1.00 0.92 0.93 0.87 0.04 -0.21 -0.10 0.18 0.96 -0.22 0.06 -0.03 0.33 0.03 0.15 0.30 0.88 0.00 0.90 0.01 0.87 -0.02 -0.03
High school diploma only 0.91 0.91 0.91 0.92 1.00 0.98 0.92 -0.05 -0.27 -0.12 0.30 0.97 -0.24 0.09 -0.06 0.42 0.06 0.15 0.36 0.89 -0.02 0.90 -0.03 0.89 -0.06 -0.10
Some college or associate's degree 0.92 0.92 0.92 0.93 0.98 1.00 0.94 -0.07 -0.32 -0.08 0.32 0.96 -0.24 0.07 -0.05 0.45 0.07 0.15 0.33 0.89 -0.04 0.89 -0.04 0.88 -0.06 -0.11
Bachelor's degree or higher 0.94 0.94 0.94 0.87 0.92 0.94 1.00 -0.09 -0.36 -0.16 0.41 0.91 -0.23 0.07 -0.05 0.53 0.05 0.12 0.46 0.90 -0.06 0.89 -0.06 0.88 -0.07 -0.14
Percent of adults with less than a high school diploma -0.05 -0.05 -0.05 0.04 -0.05 -0.07 -0.09 1.00 0.29 -0.42 -0.59 0.00 -0.40 0.30 0.02 -0.18 -0.06 0.36 -0.07 -0.03 0.29 -0.02 0.28 0.01 0.26 0.65
Percent of adults with a high school diploma only -0.31 -0.31 -0.31 -0.21 -0.27 -0.32 -0.36 0.29 1.00 -0.25 -0.80 -0.26 0.21 0.01 0.04 -0.48 -0.10 -0.10 -0.28 -0.27 0.25 -0.26 0.18 -0.27 0.08 0.34
Percent of adults completing some college or associate's degree -0.13 -0.13 -0.13 -0.10 -0.12 -0.08 -0.16 -0.42 -0.25 1.00 -0.10 -0.11 0.19 -0.20 0.10 -0.14 0.08 -0.17 -0.22 -0.12 -0.14 -0.12 -0.06 -0.12 0.01 -0.24
Percent of adults with a bachelor's degree or higher 0.33 0.33 0.34 0.18 0.30 0.32 0.41 -0.59 -0.80 -0.10 1.00 0.26 -0.02 -0.08 -0.09 0.54 0.07 -0.04 0.37 0.29 -0.29 0.28 -0.27 0.27 -0.22 -0.51
pov_2021 0.92 0.92 0.92 0.96 0.97 0.96 0.91 0.00 -0.26 -0.11 0.26 1.00 -0.26 0.10 -0.04 0.38 0.04 0.14 0.38 0.90 -0.01 0.91 -0.01 0.91 -0.03 -0.02
Percent Caucasian -0.23 -0.23 -0.23 -0.22 -0.24 -0.24 -0.23 -0.40 0.21 0.19 -0.02 -0.26 1.00 -0.73 -0.32 -0.30 -0.12 -0.47 -0.22 -0.21 -0.14 -0.21 -0.16 -0.23 -0.24 -0.47
Percent African American 0.07 0.07 0.07 0.06 0.09 0.07 0.07 0.30 0.01 -0.20 -0.08 0.10 -0.73 1.00 -0.13 0.02 -0.06 0.07 0.13 0.06 0.06 0.07 0.09 0.08 0.18 0.47
Percent Indian -0.05 -0.05 -0.05 -0.03 -0.06 -0.05 -0.05 0.02 0.04 0.10 -0.09 -0.04 -0.32 -0.13 1.00 -0.03 0.03 0.33 -0.05 -0.04 0.22 -0.04 0.14 -0.04 0.16 0.26
Percent Asian 0.45 0.45 0.45 0.33 0.42 0.45 0.53 -0.18 -0.48 -0.14 0.54 0.38 -0.30 0.02 -0.03 1.00 0.48 0.19 0.43 0.38 -0.15 0.36 -0.18 0.34 -0.14 -0.23
Percent Pacific 0.05 0.05 0.05 0.03 0.06 0.07 0.05 -0.06 -0.10 0.08 0.07 0.04 -0.12 -0.06 0.03 0.48 1.00 0.09 0.01 0.04 -0.09 0.02 -0.13 0.02 -0.07 -0.06
avg_house_size 0.13 0.13 0.13 0.15 0.15 0.15 0.12 0.36 -0.10 -0.17 -0.04 0.14 -0.47 0.07 0.33 0.19 0.09 1.00 0.03 0.13 0.16 0.13 0.15 0.14 0.12 0.14
pop_density 0.53 0.53 0.53 0.30 0.36 0.33 0.46 -0.07 -0.28 -0.22 0.37 0.38 -0.22 0.13 -0.05 0.43 0.01 0.03 1.00 0.51 -0.05 0.50 -0.05 0.49 -0.05 -0.08
cases_2022 0.98 0.98 0.98 0.88 0.89 0.89 0.90 -0.03 -0.27 -0.12 0.29 0.90 -0.21 0.06 -0.04 0.38 0.04 0.13 0.51 1.00 0.02 1.00 0.01 0.97 -0.03 -0.09
cases_percentage_2022 -0.03 -0.03 -0.03 0.00 -0.02 -0.04 -0.06 0.29 0.25 -0.14 -0.29 -0.01 -0.14 0.06 0.22 -0.15 -0.09 0.16 -0.05 0.02 1.00 0.02 0.73 0.03 0.46 0.32
cases_2021 0.98 0.98 0.98 0.90 0.90 0.89 0.89 -0.02 -0.26 -0.12 0.28 0.91 -0.21 0.07 -0.04 0.36 0.02 0.13 0.50 1.00 0.02 1.00 0.03 0.98 -0.01 -0.09
cases_percentage_2021 -0.03 -0.03 -0.03 0.01 -0.03 -0.04 -0.06 0.28 0.18 -0.06 -0.27 -0.01 -0.16 0.09 0.14 -0.18 -0.13 0.15 -0.05 0.01 0.73 0.03 1.00 0.05 0.76 0.21
cases_2020 0.96 0.96 0.96 0.87 0.89 0.88 0.88 -0.01 -0.27 -0.12 0.27 0.91 -0.23 0.08 -0.04 0.34 0.02 0.14 0.49 0.97 0.03 0.98 0.05 1.00 0.03 -0.07
cases_percentage_2020 -0.05 -0.05 -0.05 -0.02 -0.06 -0.06 -0.07 0.26 0.08 0.01 -0.22 -0.03 -0.24 0.18 0.16 -0.14 -0.07 0.12 -0.05 -0.03 0.46 -0.01 0.76 0.03 1.00 0.22
pov_percentage -0.11 -0.11 -0.11 -0.03 -0.10 -0.11 -0.14 0.65 0.34 -0.24 -0.51 -0.02 -0.47 0.47 0.26 -0.23 -0.06 0.14 -0.08 -0.09 0.32 -0.09 0.21 -0.07 0.22 1.00
pop_2020
pop_2021
pop_2022
Less than a high school diploma
High school diploma only
Some college or associate's degree
Bachelor's degree or higher
Percent of adults with less than a high school diploma
Percent of adults with a high school diploma only
Percent of adults completing some college or associate's degree
Percent of adults with a bachelor's degree or higher
pov_2021
Percent Caucasian
Percent African American
Percent Indian
Percent Asian
Percent Pacific
avg_house_size
pop_density
cases_2022
cases_percentage_2022
cases_2021
cases_percentage_2021
cases_2020
cases_percentage_2020
pov_percentage
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8

To fix this problem, the counties were grouped by state and then the same heatmaps were generated in hopes to see a stronger correlation as states had different rules for how they handled COVID 19. This isn't the perfect solution as external factors such as super spreader events and state border counties can influence this data but, we can generalize the data in order to see some more noticeable trends.





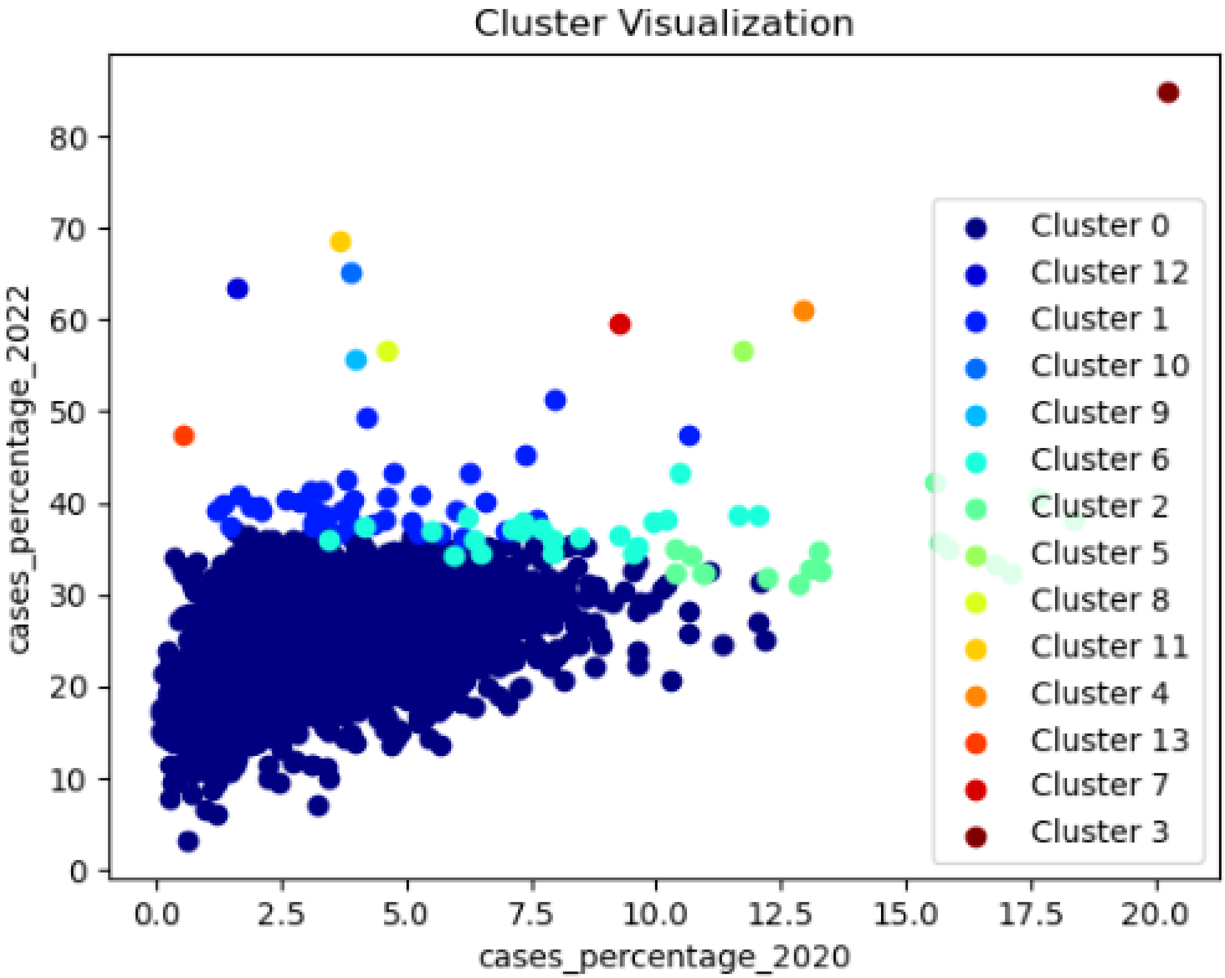


The first two graphs were sorted based on the mean average from each state (represented by a dot) while the bottom graph was sorted based on the median.

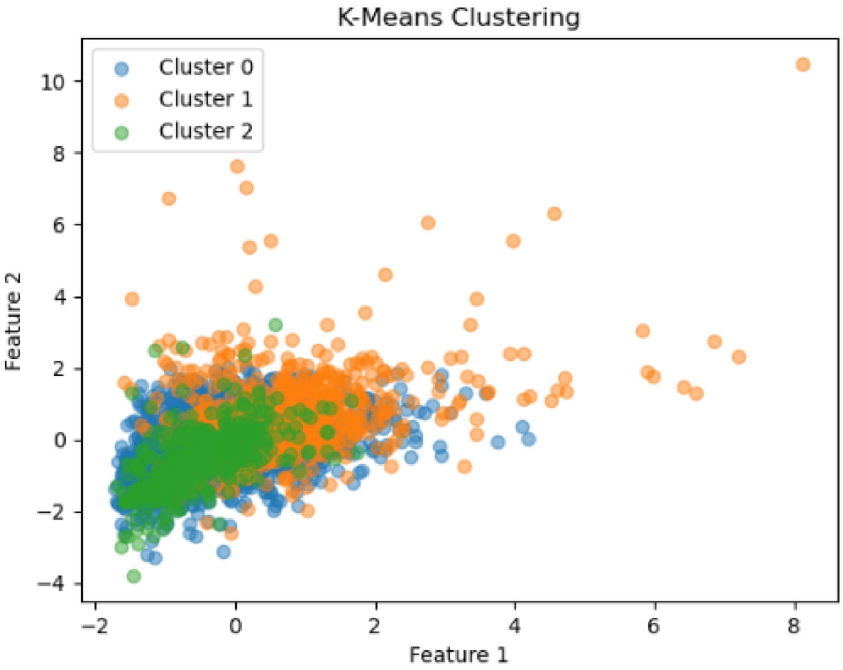
Preprocessing and Modeling

The goals of this step were to set up and model our data for classification models in order to find better trends within our data than what we saw with our EDA step.

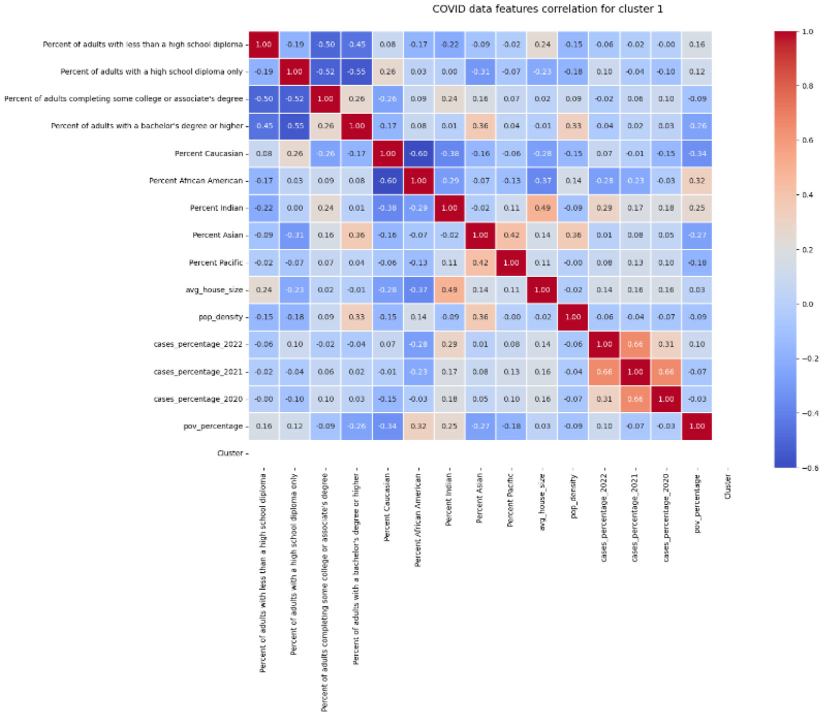
The first models built were with Mean Shift Clustering. Unfortunately, the models built from both approaches proved to be unsuccessful. As shown from the graph below, the models built clusters where one cluster clearly dominated the others when it came to the total number of data points (counties). Cluster 3 only had 1 data point!



With the combination of PCA and KMeans, we were able to find a better grouping of clusters by reducing the number of clusters. These initial models were clustered by COVID case percentage.

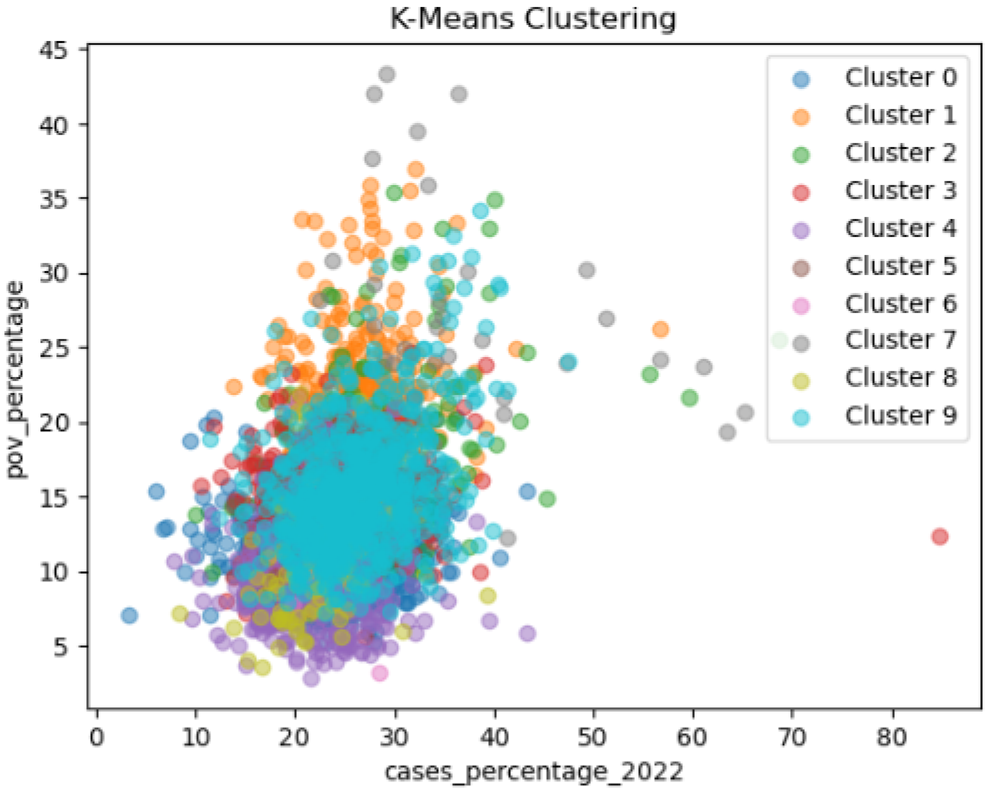


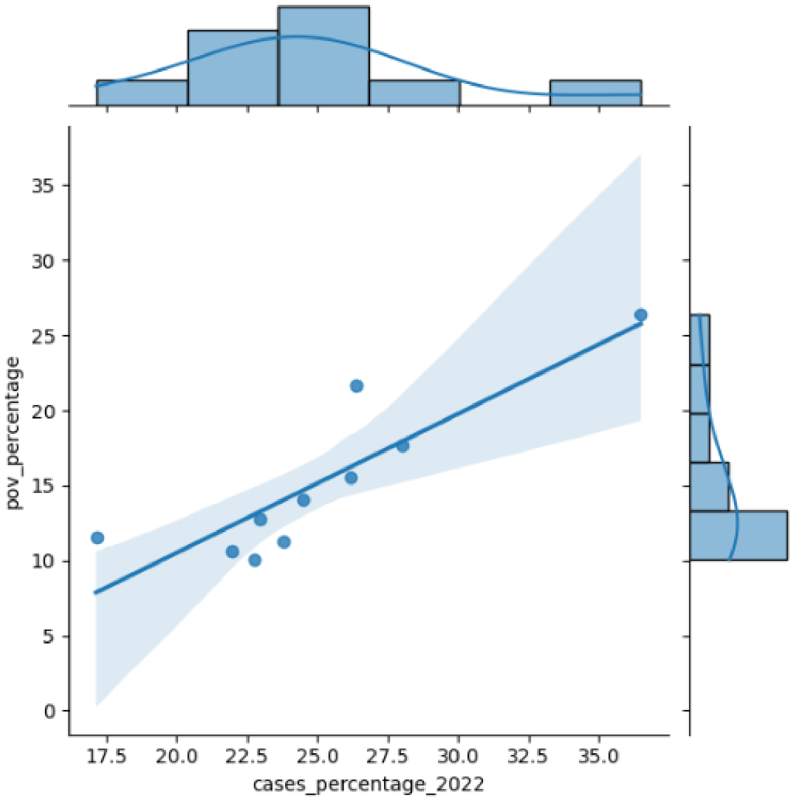
However, when looking at a heatmap of a specific cluster, most if not all of the relations showed no correlation.



	Percent of adults with less than a high school diploma	Percent of adults with a high school diploma only	Percent of adults completing some college or associate's degree	Percent of adults with a bachelor's degree or higher	Percent Caucasian	Percent African American	Percent Indian	Percent Asian	Percent Pacific	avg_house_size	pop_density	cases_percentage_2022	cases_percentage_2021	cases_percentage_2020	pov_percentage	Cluster
Percent of adults with less than a high school diploma	1.00	-0.19	-0.50	-0.45	0.08	-0.17	-0.22	-0.09	-0.02	0.24	-0.15	-0.06	-0.02	-0.00	0.16	
Percent of adults with a high school diploma only	-0.19	1.00	-0.52	-0.55	0.26	0.03	0.00	-0.31	-0.07	-0.23	-0.18	0.10	-0.04	-0.10	0.12	
Percent of adults completing some college or associate's degree	-0.50	-0.52	1.00	0.26	-0.26	0.09	0.24	0.16	0.07	0.02	0.09	-0.02	0.06	0.10	-0.09	
Percent of adults with a bachelor's degree or higher	-0.45	-0.55	0.26	1.00	-0.17	0.08	0.01	0.36	0.04	-0.01	0.33	-0.04	0.02	0.03	-0.26	
Percent Caucasian	0.08	0.26	-0.26	-0.17	1.00	-0.60	-0.38	-0.16	-0.06	-0.28	-0.15	0.07	-0.01	-0.15	-0.34	
Percent African American	-0.17	0.03	0.09	0.08	-0.60	1.00	-0.29	-0.07	-0.13	-0.37	0.14	-0.28	-0.23	-0.03	0.32	
Percent Indian	-0.22	0.00	0.24	0.01	-0.38	-0.29	1.00	-0.02	0.11	0.49	-0.09	0.29	0.17	0.18	0.25	
Percent Asian	-0.09	-0.31	0.16	0.36	-0.16	-0.07	-0.02	1.00	0.42	0.14	0.36	0.01	0.08	0.05	-0.27	
Percent Pacific	-0.02	-0.07	0.07	0.04	-0.06	-0.13	0.11	0.42	1.00	0.11	-0.00	0.08	0.13	0.10	-0.18	
avg_house_size	0.24	-0.23	0.02	-0.01	-0.28	-0.37	0.49	0.14	0.11	1.00	-0.02	0.14	0.16	0.16	0.03	
pop_density	-0.15	-0.18	0.09	0.33	-0.15	0.14	-0.09	0.36	-0.00	-0.02	1.00	-0.06	-0.04	-0.07	-0.09	
cases_percentage_2022	-0.06	0.10	-0.02	-0.04	0.07	-0.28	0.29	0.01	0.08	0.14	-0.06	1.00	0.66	0.31	0.10	
cases_percentage_2021	-0.02	-0.04	0.06	0.02	-0.01	-0.23	0.17	0.08	0.13	0.16	-0.04	0.66	1.00	0.66	-0.07	
cases_percentage_2020	-0.00	-0.10	0.10	0.03	-0.15	-0.03	0.18	0.05	0.10	0.16	-0.07	0.31	0.66	1.00	-0.03	
pov_percentage	0.16	0.12	-0.09	-0.26	-0.34	0.32	0.25	-0.27	-0.18	0.03	-0.09	0.10	-0.07	-0.03	1.00	
Cluster																

When trying from the other approach, it was noticeable that if you were to group the clusters by the mean, a stronger correlation would appear from the models.





Analysis

When looking at the clusters generated by the KMeans modeling algorithm, we can see some noticeable trends between our country clusters. Due to the nature of COVID 19 being a national problem however, these findings may not explain the whole picture due to a plethora of factors that may influence why COVID 19 case totals were high / low in different counties. These factors that can't be listed in statistics definitely played a major role in the outcome of the datasets. An example is whether or not an individual would even get tested in the first place because it was not a mandatory nor daily practice due to costs, so there is no way of knowing how many cases truly happened in the US. However, the biggest trend from our data that we can take away was that the COVID 19 case percentage was high in counties that have high poverty percentages. This might be explained by those who had access to clean and healthy resources during quarantine. While the models may not be perfect, we can look at the results to see what locations might be seen as dangerous when it comes to a new contagious and infectious disease in the future.