Hamin Han

Machine Learning and COVID-19 Data

Over the last few years, everyone around the globe has been affected by COVID-19. I want to use machine learning to find out if there is any visible pattern related to population and geography. Specifically, I want to use clustering on COVID-19 data and see if it visualize clear clusters formed by the geography and population of the regions. I would assume that the regions with massive populations, especially near populous cities, would have far more COVID cases. But what would that look like on a local scale?

Related Work:

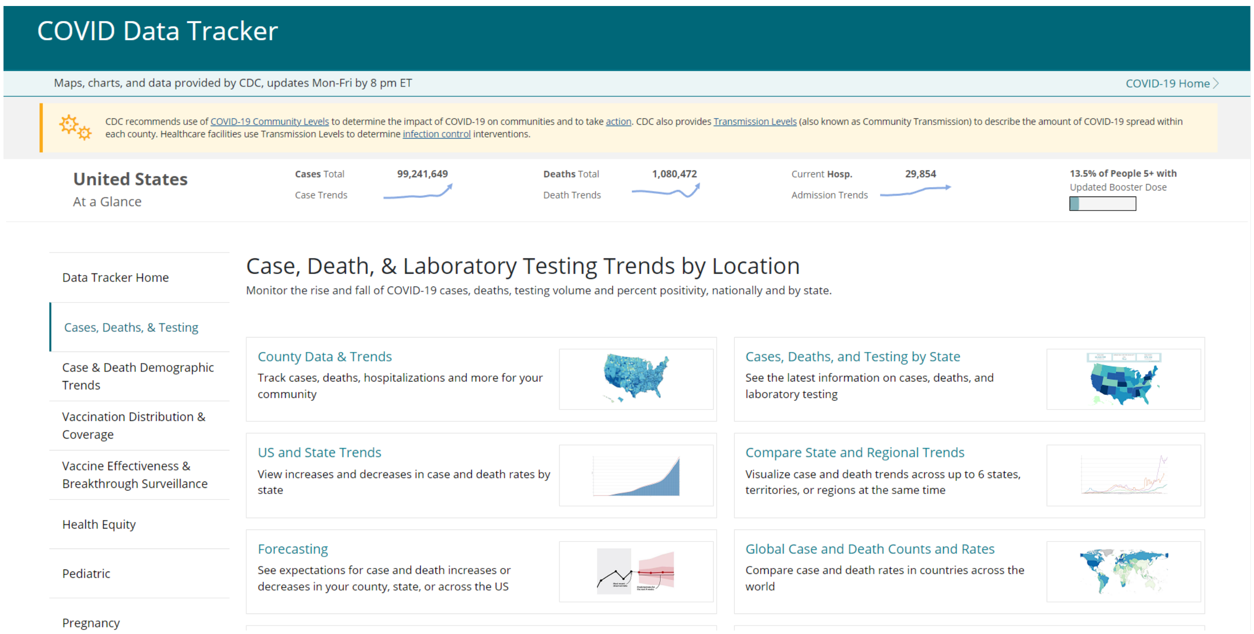
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7998460/

From a published academic journal called *The Geography of the Covid-19 Pandemic: A Data-Driven Approach to Exploring Geographical Driving Forces,* the authors concluded that “the disease initially spread in the densely and heavy populated capital region and over time moved to more distant regions of the country, furthermore single events of large spread and counter measures of these showed a space-time ripple effect in decreasing infection rates”. From their investigation of the relationship of geographic factors and COVID-19, they were able to first prove the direct relationship between population size and the spread of the virus. This was also the case for my findings when running clustering on US states and local counties.

Methodology:

First and foremost, it is important to have accurate and organized data to use. For my project, there was an abundance of COVID-19 data out there but I used the data provided by CDC, the Centers for Disease Control and Prevention.

LINK: https://covid.cdc.gov/covid-data-tracker/#cases-deaths-testing-trends



From there, I was able to get specific data for recent covid cases in state counties.

(Maryland Counties Covid Cases Data)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	County Level Latest Data for Maryland																					
2	Date generated: Tue Dec 13 2022 17:35:38 GMT-0500 (Eastern Standard Time)																					
3	FIPS code	State Nam	County	Weekly Ca	Weekly Ca	% Change	Test posit	Test posit	Total diag	Total diag	Total diag	Weekly D	Total adm	Weekly D	% Change	Populatio	Average h	Percent u	Poverty ra	Percent p	SVI	CCVI
4	24001	Maryland	Allegany (	74	105.09	21.31	5.81	-1.56	1314	1875.62	256.1	suppresse	9.2	suppresse	N/A	166.79	2.3	4.8	16.4	20.55	0.621	0.3
5	24003	Maryland	Anne Arun	541	93.4	68.01	12.64	1.56	4419	758.27	152.8	suppresse	11.5	suppresse	0	1396.38	2.65	4.7	6	15.02	0.1783	0.1
6	24510	Maryland	Baltimore	781	131.59	43.04	8.34	-0.76	9347	1594.69	210.22	suppresse	11.5	suppresse	-50	7331.86	2.48	7.2	21.8	14.52	0.8312	0.8
7	24005	Maryland	Baltimore	744	89.92	57.29	10.52	0.71	8144	985.94	170.65	13	11.5	1.57	18.18	1382.73	2.58	5.6	9.2	17.56	0.3818	0.3
8	24009	Maryland	Calvert Co	69	74.57	46.81	9.29	-1.42	612	657.56	245.76	suppresse	11.5	suppresse	N/A	434	2.85	4.2	5.1	15.48	0.0471	0.1
9	24011	Maryland	Caroline C	23	68.85	53.33	8.59	2.55	342	1021.14	147.83	0	0.7	0	0	104.58	2.68	6.4	14.7	16.69	0.5535	0.5
10	24013	Maryland	Carroll Co	92	54.62	27.78	12	3.37	1060	626.88	161.73	suppresse	11.5	suppresse	-50	376.31	2.71	3	5.3	17.28	0.0404	0.1
11	24015	Maryland	Cecil Coun	65	63.2	-20.73	9.07	1.4	860	831.57	179.22	0	6.7	0	0	297.01	2.74	4.5	9.4	16.21	0.2933	0.2
12	24017	Maryland	Charles Co	210	128.63	77.97	14.6	0.23	1530	930.45	190.87	0	9.7	0	-100	356.61	2.78	3.6	6.1	12.87	0.1723	0.2
13	24019	Maryland	Dorcheste	41	128.41	105	12.31	3.34	408	1280.88	191.43	0	0.7	0	0	59.04	2.39	5.5	15.8	22.12	0.7506	0.7
14	24021	Maryland	Frederick	233	89.77	54.3	10.35	2.15	2345	884.37	223	suppresse	6.7	suppresse	-33.33	392.94	2.67	4.8	7.1	14.82	0.1589	0.1
15	24023	Maryland	Garrett Co	21	72.38	10.53	5.85	-0.9	411	1424.51	328.13	0	9.2	0	-100	44.7	2.38	7.3	9.7	23.14	0.1812	0.1
16	24025	Maryland	Harford Co	216	84.56	36.71	10.8	-1.14	2172	845.78	183.92	0	6.7	0	-100	584.37	2.67	3.7	7.6	16.56	0.0933	0.1
17	24027	Maryland	Howard Co	354	108.69	69.38	13.28	1.77	2787	849.18	174.85	suppresse	11.5	suppresse	N/A	1297.84	2.77	4	5.4	14.28	0.1306	0.1
18	24029	Maryland	Kent Coun	16	82.38	77.78	6.76	-3.92	151	786.79	106.85	0	0	0	0	70.11	2.27	5.4	12.3	27.1	0.4277	0.5
19	24031	Maryland	Montgom	1399	133.15	64.01	12.72	0.98	12829	1219.7	168.61	11	9.7	1.05	266.67	2130.94	2.79	7.4	6.9	16.06	0.3255	0.2
20	24033	Maryland	Prince Ge	1268	139.44	83.24	12.66	1.8	11450	1258.78	165.97	suppresse	9.7	suppresse	0	1884.02	2.87	10.8	8.9	13.89	0.6261	0.5
21	24035	Maryland	Queen An	44	87.33	18.92	14.79	-0.74	358	699.67	211.3	0	0.7	0	0	135.55	2.69	4.4	5.5	19.22	0.022	0.1
22	24039	Maryland	Somerset	19	74.17	72.73	12.26	-2.78	154	605.04	120	0	15.9	0	0	80.11	2.34	6.8	20.4	17.32	0.8978	0.5
23	24037	Maryland	St. Mary's	130	114.53	78.08	19.25	4.48	807	703.65	241.95	0	9.7	0	-100	316.45	2.69	5.8	8.3	13.39	0.2086	0.2
24	24041	Maryland	Talbot Cou	30	80.69	-18.92	10.26	1.25	360	973.71	260	suppresse	0.7	suppresse	N/A	138.45	2.21	4.8	9.5	29.74	0.1331	0.3
25	24043	Maryland	Washingto	105	69.51	101.92	7.08	0.65	1311	867.37	131.22	suppresse	7.2	suppresse	100	329.97	2.52	6.2	12.7	17.5	0.614	0.5
26	24045	Maryland	Wicomico	110	106.17	86.44	10.77	-1.16	1226	1178.96	208.04	0	15.9	0	0	276.72	2.61	6.9	15.2	16.35	0.7739	0.
27	24047	Maryland	Worcester	23	44	43.75	12.09	6.09	381	727.06	252.78	0	15.9	0	0	111.61	2.34	5.9	9.3	28.23	0.2803	0.4

And then from there, I do the clustering. First import all the libraries and data

```
# import libraries

import pandas as pd
import numpy as np
import random as rd
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn import metrics
from sklearn.metrics import pairwise_distances
from sklearn.cluster import KMeans
```

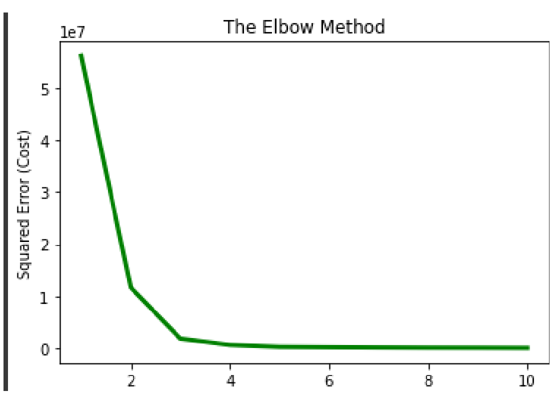
```
# import the data csv file

data = pd.read_csv('county_level_latest_data_for_maryland.csv', skiprows=2, on_bad_lines='skip')
data.columns = [x.replace("\n", " ") for x in data.columns.to_list()]
data.head()
```

	FIPS code	State Name	County	Weekly Cases	Weekly Case rate /100k	% Change in weekly Cases from past week	Test positivity rate - last 7 days	Test positivity rate - absolute change	Total diagnostic tests - last 7 days
0	24001	Maryland	Allegany County	74	105.09	21.31	5.81	-1.56	1314
1	24003	Maryland	Anne Arundel County	541	93.40	68.01	12.64	1.56	4419

Run K-means looking at the Population Density and Weekly Cases columns. But first decide the optimal k-value using the elbow method. (4 in this case)

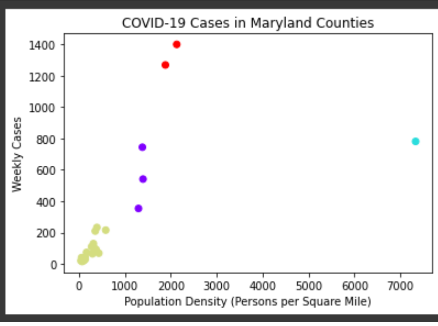
For reference, “The elbow method is a graphical representation of finding the optimal 'K' in a K-means clustering. It works by finding WCSS (Within-Cluster Sum of Square) i.e. the sum of the square distance between points in a cluster and the cluster centroid.” (towardsdatascience)

```
cost =[]
for i in range(1, 11):
    KM = KMeans(n_clusters = i, max_iter = 500)
    KM.fit(X)

    # calculates squared error
    # for the clustered points
    cost.append(KM.inertia_)

# plot the cost against K values
plt.plot(range(1, 11), cost, color ='g', linewidth ='3')
plt.title('The Elbow Method')
plt.xlabel("Value of K")
plt.ylabel("Squared Error (Cost)")
plt.show() # clear the plot
```



Now implement K-means clustering using the Sklearn library.

```
kmeans = KMeans(4) # From the value taken from the elbow method
kmeans.fit(X)
identified_clusters = kmeans.fit_predict(X)

data_with_clusters = data.copy()
data_with_clusters['Clusters'] = identified_clusters
plt.scatter(data_with_clusters['Population density - county'],data_with_clusters['Weekly Cases'],c=data_with_clusters['Clusters'],cmap='rainbow')
plt.title('COVID-19 Cases in Maryland Counties')
plt.xlabel('Population Density (Persons per Square Mile)')
plt.ylabel('Weekly Cases')
plt.show()
```



Also test the silhouette score and average it to see how good the clusters are.

For reference, “The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.” (Wikipedia)

```
[ ]  score = metrics.silhouette_score(X, kmeans.labels_, metric='euclidean')
     print('Average Silhouette Score over All Samples: %.3f' % score)

     Average Silhouette Score over All Samples: 0.770
```

I did this for the local counties in Maryland and the nearby state Virginia. I also thought to run the same clustering algorithm for the four very populous states, California, Texas, Florida, and New York. As well as doing a similar approach for clustering the 50 states of America.

All of the ipynb files and csv files are included. And kept separate for each clustering.

Analysis of Results:

I made a separate powerpoint slides that organized the results and analysis.

But most of the results verified the direct relationship between population density and the number of COVID-19 cases. Additionally, the clusters formed were divided into geographical regions, initially grouping around large populous cities and a second group forming that are near these cities, and the other set of cluseters that are far away. Even if multiple clusters with multiple cities were formed, they would group up near the matching city or regions. The results proved what was being said in the academic journal: “the disease initially spread in the densely and heavy populated capital region and over time moved to more distant regions of the country.”

Check the slides for the indepth analysis.

I will also add parts of the slides as reference at the end of the report.

Limitations:

This might be due to my lack of experience with the coding, but it’s hard to identify which county or state the point on the graph represents. I had to check the values through the csv file and match it up to find out what the data point was and where it’s located on the actual map. Also, for the USA states clustering, the data didn’t have the population density of the states so I just had to use the number of total cases as a general benchmark for the amount of people there are in the states.

Potential Follow-up Work:

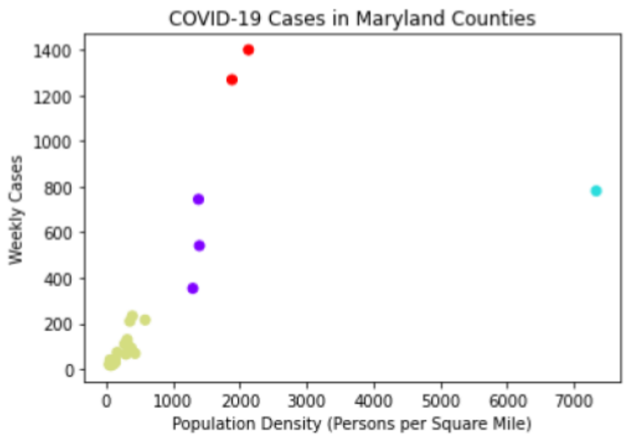
One thing is to add more functionality to the code. Be able to see what each data points are and how that might look geographically as well. Maybe fine tune the elbow method and the silhouette score code as well. Just more depth and analysis with the coding since I would consider myself a novice when it comes to Machine Learning and python in general. I’m sure there are a lot of more useful libraries that I could’ve implemented to make the quality of the project better.

As for a long-term improvement, perhaps going beyond and trying to predict how the new variants will spread geographically using Machine Learning. Just predicting the effects and spread of the virus is impressive but tying it into geographical regions, especially locally in Maryland would be a cool idea to explore in the future.

Hamin Han

EXTRA REFERENCES FOR RESULTS (Check the slides included)

Maryland Results

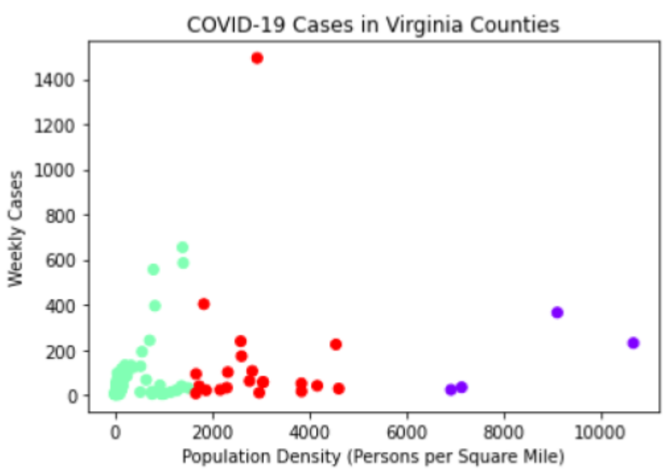


Average Silhouette Score over All Samples: 0.770

- **Yellow:** Less populated counties that have less cases as expected.
- All the other counties are nearby DC or Baltimore.
- **Purple:** Counties near Baltimore City. The closer the county is to the city, the more cases is has.
- **Red:** Montgomery County and Prince George's County that are right next to DC.
- **Cyan:** Baltimore City. Still a lot of cases but expected more with the population.

Clusters that are easily identifiable and shows the direct relationship between population and COVID cases.

Virginia Results



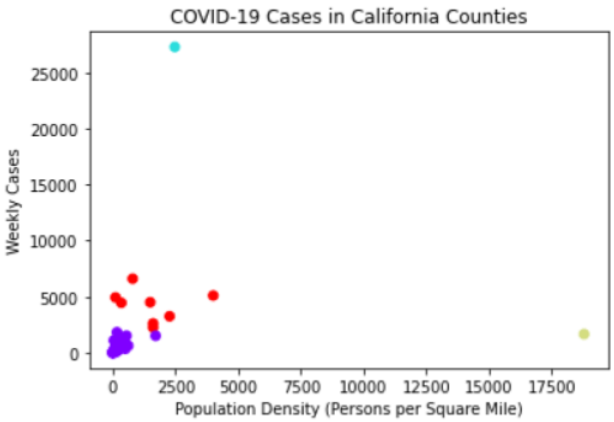
Average Silhouette Score over All Samples: 0.779

- **Purple:** Mostly counties near DC.
- **Green:** A lot of the other counties that aren't close to nearby cities.
- **Red:** A mixture of counties near cities such as Virginia Beach, Chesapeake City, and Richmond.

There are clusters being formed and divided into geographical regions but doesn't show a great relationship between population and COVID cases.

This could be due to the data being collected in the recent 7 day period so there might be a lot of outliers.

California Results



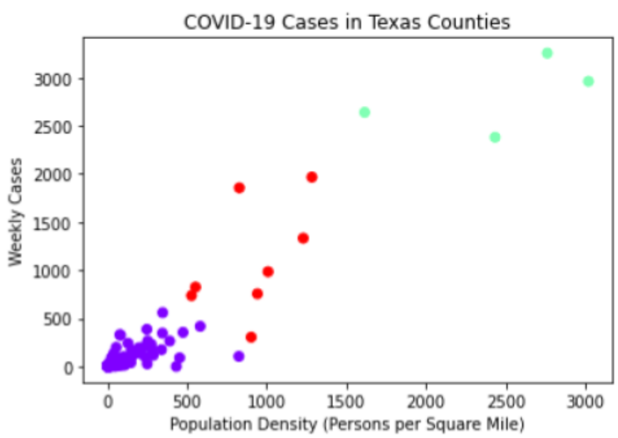
Average Silhouette Score over All Samples: 0.763

- **Yellow:** San Francisco County with the most population density.
- **Cyan:** Los Angeles County with the incomparably high number of COVID cases.
- **Red:** These are mostly counties in the southern parts of California, parts of them are near LA and a handful are near San Francisco.
- **Purple:** The rest of California. Mostly on the northern parts of California or the outer borders of California where it's far from the big cities.

Clusters that are easily identifiable by geographical location, particularly around LA and San Francisco.

Surprisingly, San Francisco County has a small number cases compared to its population density while LA county is on another level with COVID cases

Texas Results

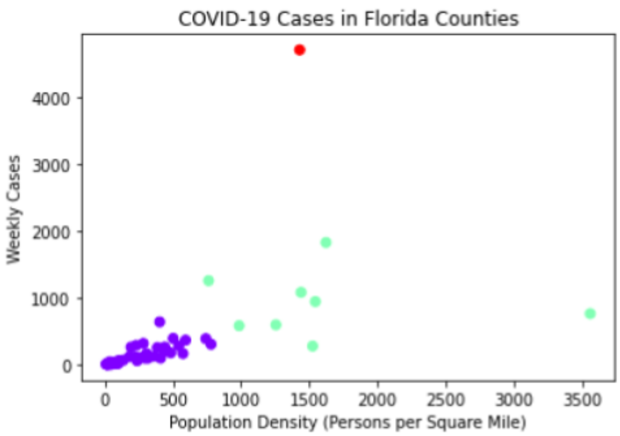


- **Green:** These counties aren't exactly near each other but they all contain big cities such as Dallas, Houston, Austin, and El Paso.
- **Red:** This is the second group of counties that are still nearby one of the big cities.
- **Purple:** All the rest of counties that aren't exactly near any populous cities.

Clusters that are easily identifiable and shows a clear and direct relationship between population and COVID cases.

Average Silhouette Score over All Samples: 0.893

Florida Results

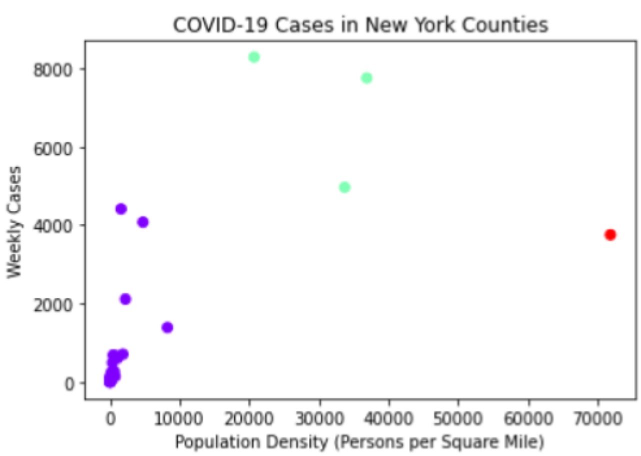


Average Silhouette Score over All Samples: 0.754

- **Red:** Miami-Dade County with the most cases. Most likely due to having a big city like Miami
- **Green:** Counties with or near big cities such as Orlando, Hollywood, and Tampa. These are also the counties with more population density.
- **Purple:** Rest of the counties that are far from big cities and less population densities.

Clusters that are easily identifiable and shows the direct relationship between population and COVID cases. Although there is an outlier with Miami-Dade with the enormous amount of cases while not having the most population density. Most likely due to a big city like Miami being there and spreading COVID easily.

New York Results

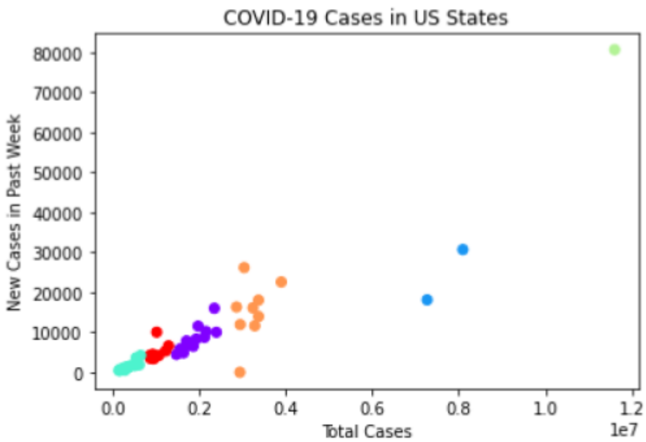


Average Silhouette Score over All Samples: 0.932

- **Red and Green** are the counties of NYC
- **Red:** New York County with Manhattan with the most population
- **Green:** Bronx, Kings, Queens counties with Bronx, Brooklyn, and Queens.
- **Purple:** All the other counties.

Clusters that are easily identifiable and shows the direct relationship between population and COVID cases. Large population density with a lot of cases where NYC is. Although New York County with the most population seem to be an outlier, could be because the data is only from the recent 7 days.
For the purple cluster, more cases as the counties get close to NYC towards south of the state

US States Results



Average Silhouette Score over All Samples: 0.661

- **Green:** California
- **Blue:** Texas and Florida
- **Orange:** New York, New Jersey, Pennsylvania and other States. Mostly in the Mideast and Midwest.
- **Cyan:** Mountain-Prairie area
- **Red and Purple:** Everything in between. WIth more cases for states with or near large cities

Clusters that are easily identifiable and shows the direct relationship between population and COVID cases.

Although California is skyrocketing with the number of cases.