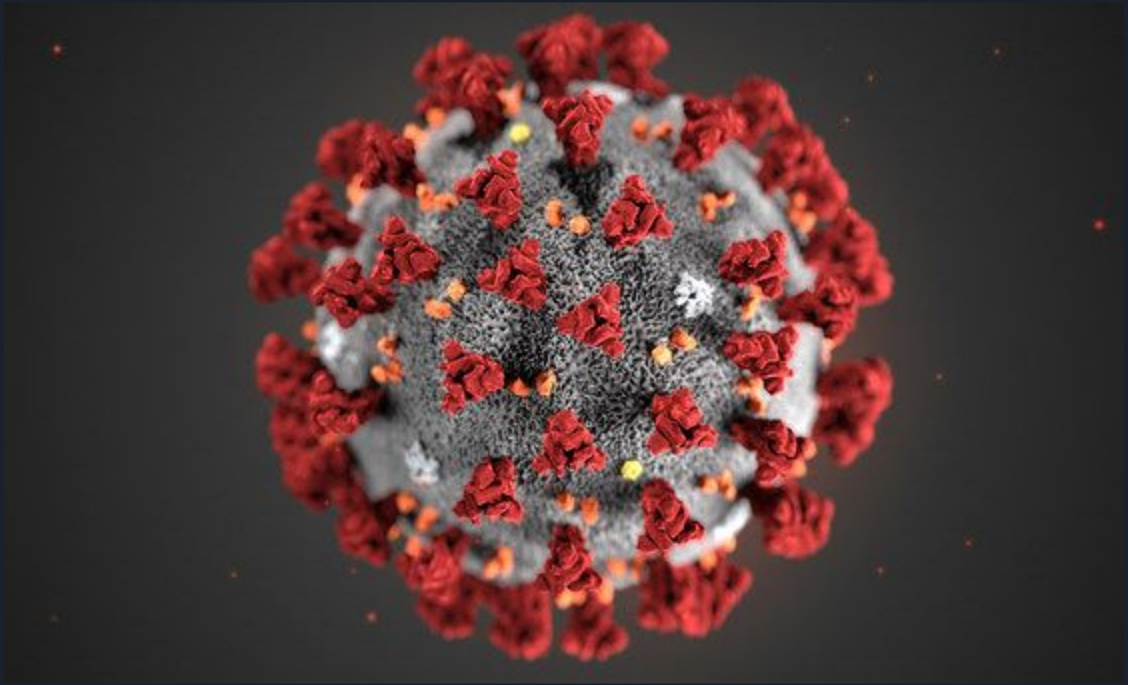
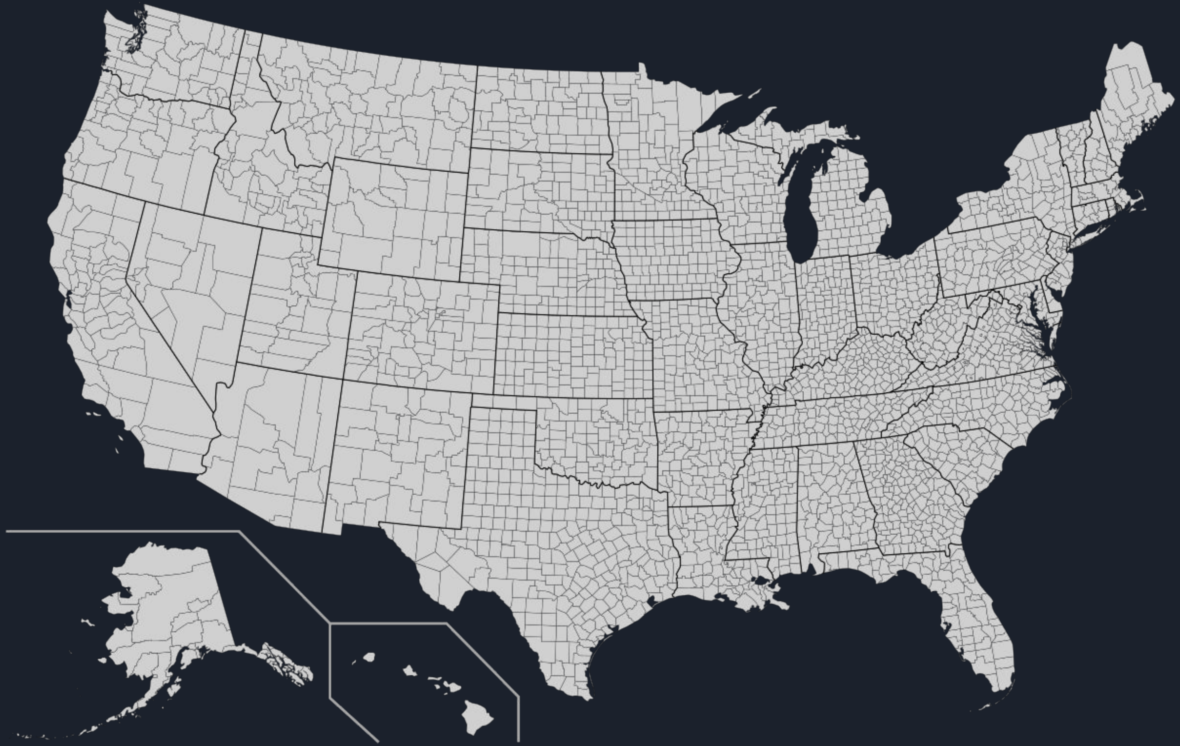
COVID-19 County Correlations

Matthew Shinder

COVID-19



US Counties



Goals of this Project

Find trends of COVID-19 data, specifically between US counties

Data Sources

COVID-19 County Data

kaggle

US County Demographic Information

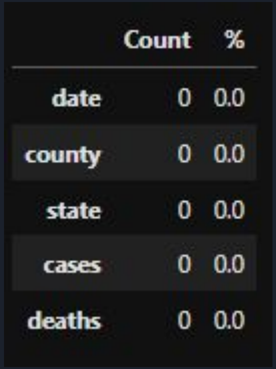
United States® Census Bureau

USDA

Data Wrangling / Cleaning: COVID-19

Number of entries: 2,502,832

Null data



	Count	%
date	0	0.0
county	0	0.0
state	0	0.0
cases	0	0.0
deaths	0	0.0

Unwanted Data

US territories

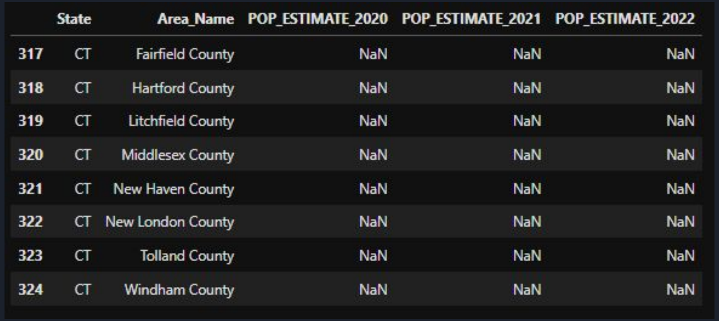
- Puerto Rico
- Northern Mariana Islands
- Virgin Islands

Data Wrangling / Cleaning: Demographics

Number of entries: ~3285 per file (4 files)

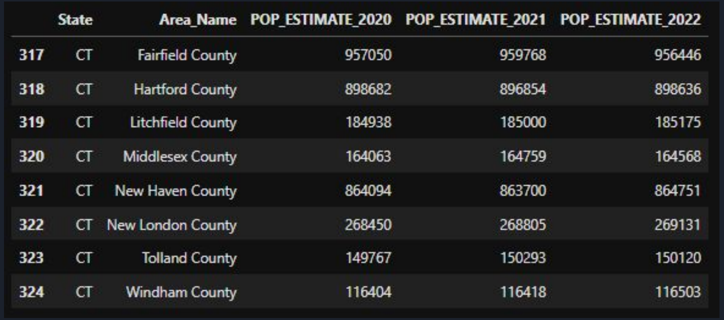
NaN and Puerto Rico Data: Dropped

Connecticut Population Issue



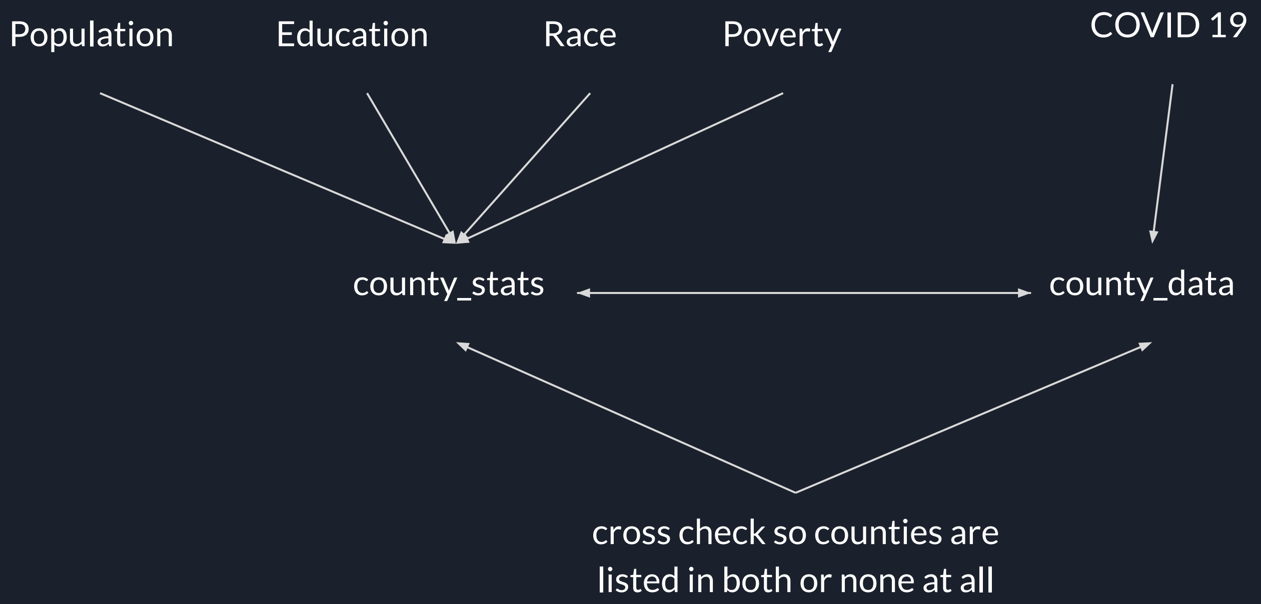
	State	Area_Name	POP_ESTIMATE_2020	POP_ESTIMATE_2021	POP_ESTIMATE_2022
317	CT	Fairfield County	NaN	NaN	NaN
318	CT	Hartford County	NaN	NaN	NaN
319	CT	Litchfield County	NaN	NaN	NaN
320	CT	Middlesex County	NaN	NaN	NaN
321	CT	New Haven County	NaN	NaN	NaN
322	CT	New London County	NaN	NaN	NaN
323	CT	Tolland County	NaN	NaN	NaN
324	CT	Windham County	NaN	NaN	NaN

Revised:



	State	Area_Name	POP_ESTIMATE_2020	POP_ESTIMATE_2021	POP_ESTIMATE_2022
317	CT	Fairfield County	957050	959768	956446
318	CT	Hartford County	898682	896854	898636
319	CT	Litchfield County	184938	185000	185175
320	CT	Middlesex County	164063	164759	164568
321	CT	New Haven County	864094	863700	864751
322	CT	New London County	268450	268805	269131
323	CT	Tolland County	149767	150293	150120
324	CT	Windham County	116404	116418	116503

Data Merging



Population

Education

Race

Poverty

COVID 19

county_stats

county_data

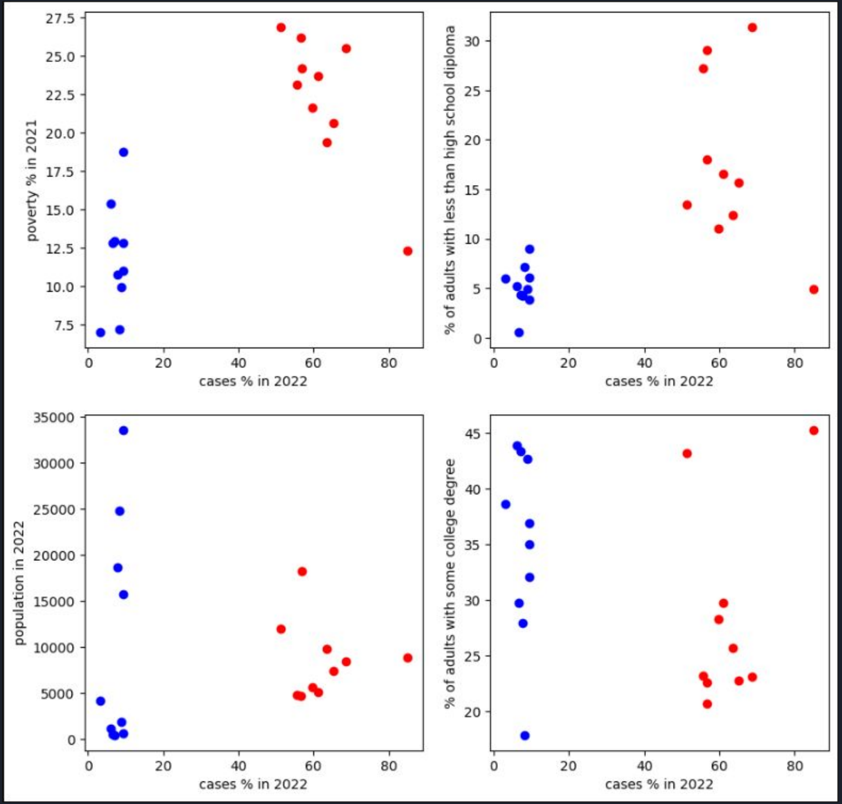
cross check so counties are listed in both or none at all

EDA

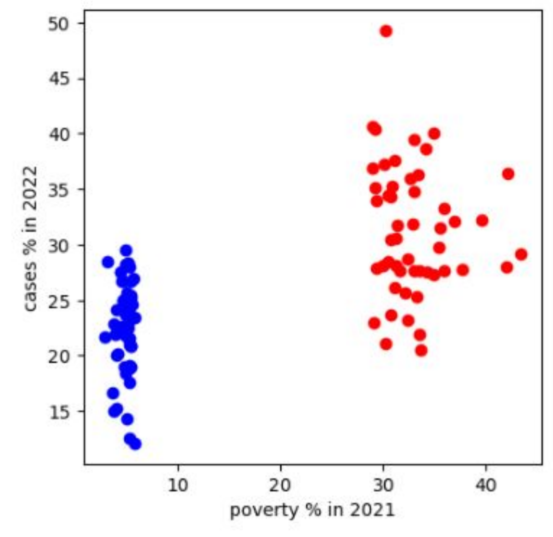
Two Approaches

1. Look at counties with high / low covid statistics and compare their demographics
2. Look at counties with high / low demographics and compare their covid statistics

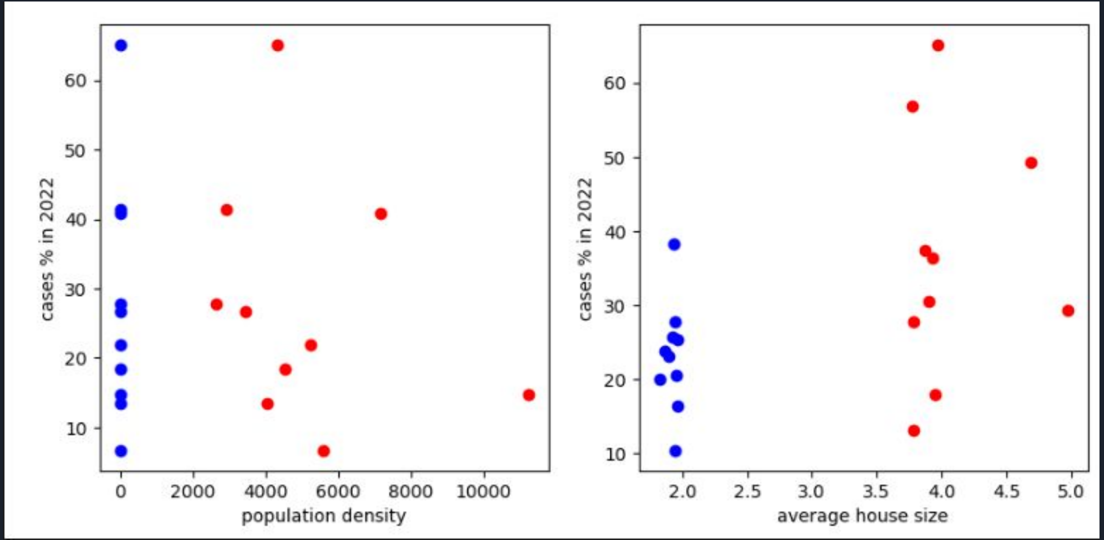
Approach 1


poverty % in 2021
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
% of adults with less than high school diploma
30
25
20
15
10
5
0
population in 2022
35000
30000
25000
20000
15000
10000
5000
0
% of adults with some college degree
45
40
35
30
25
20
0
20
40
60
80
cases % in 2022

Approach 2



cases % in 2022
poverty % in 2021
50
45
40
35
30
25
20
15
10
20
30
40



cases % in 2022
population density
60
50
40
30
20
10
0
2000
4000
6000
8000
10000
cases % in 2022
average house size
60
50
40
30
20
10
2.0
2.5
3.0
3.5
4.0
4.5
5.0

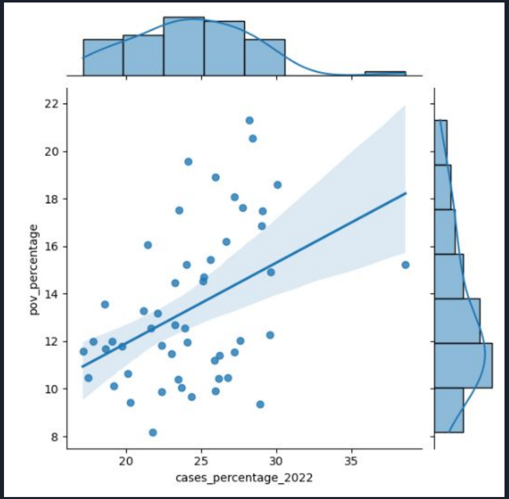
Choosing Dates

Three Dates chosen to separate key events

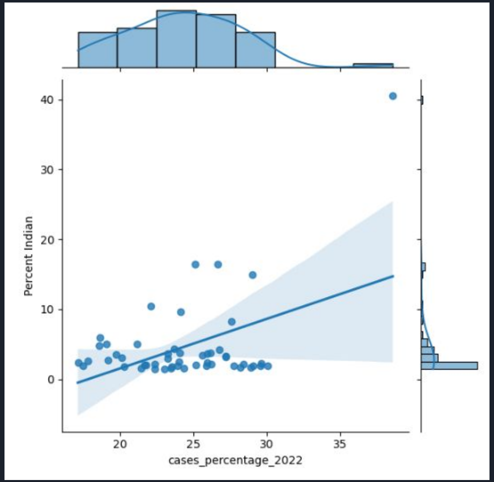
1. 11/13/20 : Every US County recorded at least 1 COVID 19 case
2. 6/1/21: US COVID 19 cases surging down after release of vaccine public to all
3. 5/13/22: End of given data

Note: These dates were used to make the 20XX_case_percentage columns (running totals)

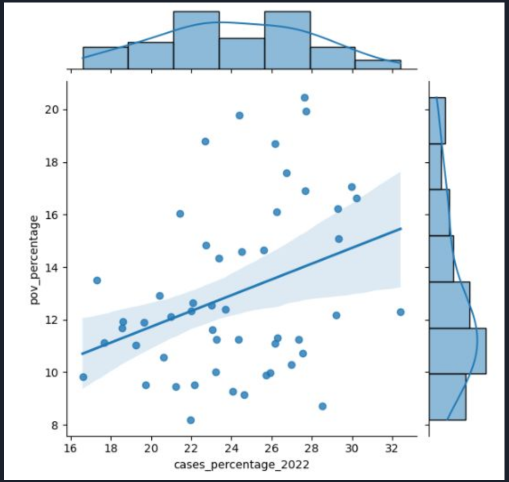
Grouping by State



By Mean



By Mean



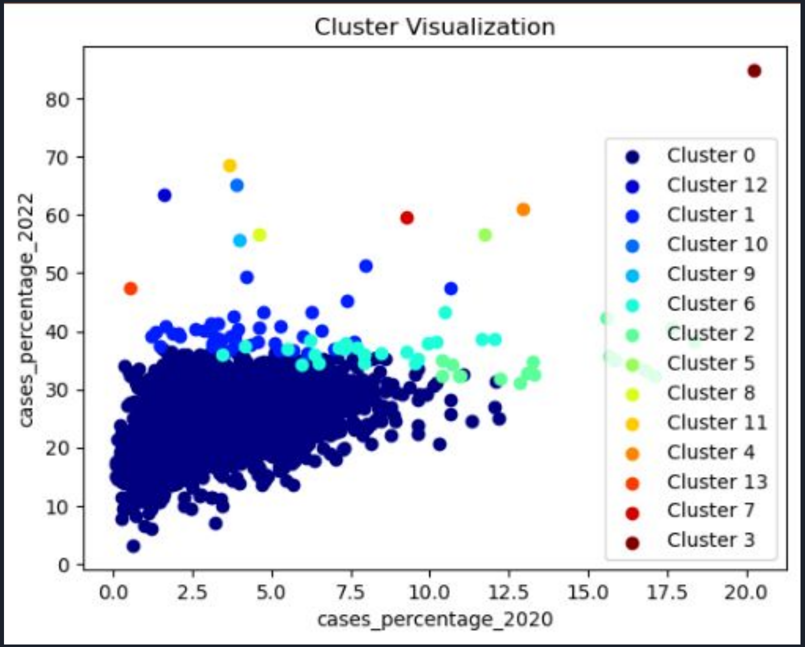
By Median

Modeling

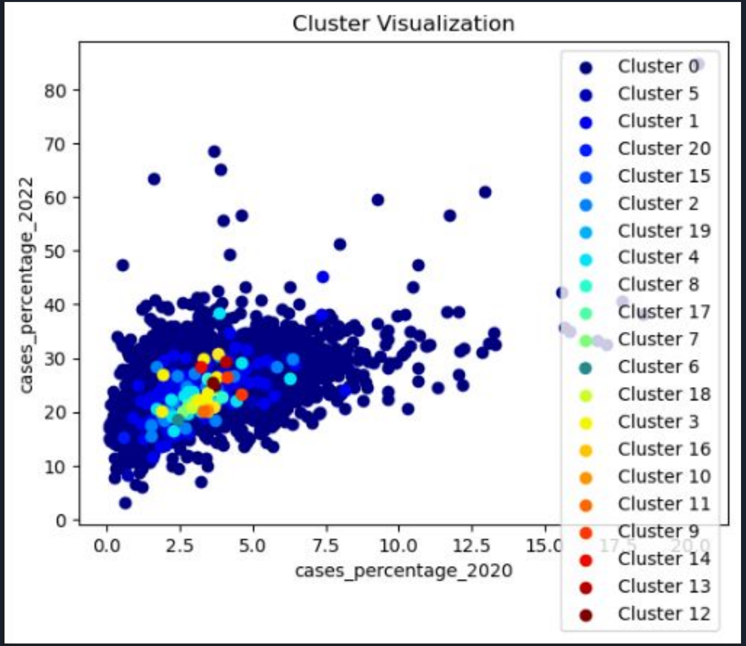
Two Approaches

1. Cluster counties around their covid statistics and compare their demographics
2. Cluster counties around their demographics and compare their covid stats

Mean Shift Clustering

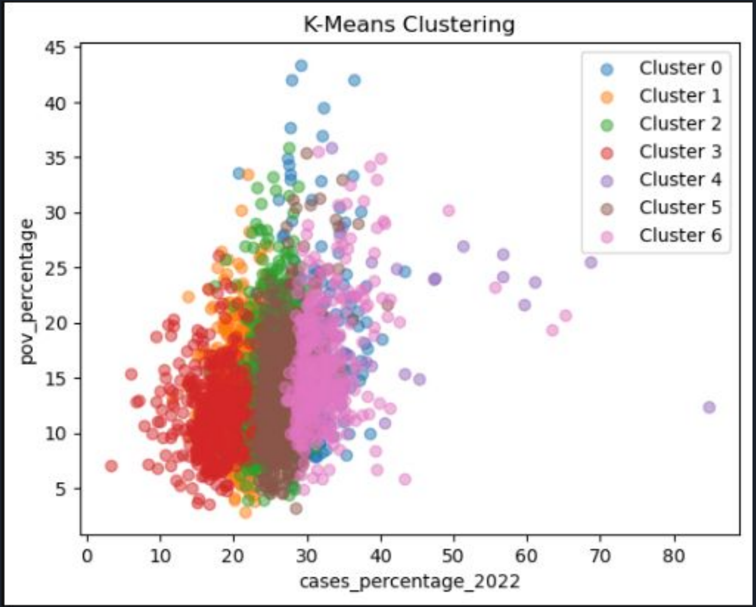


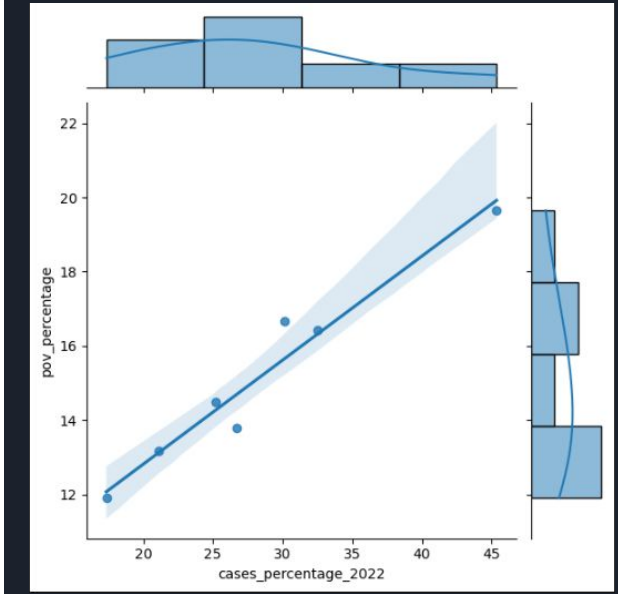
Approach 1

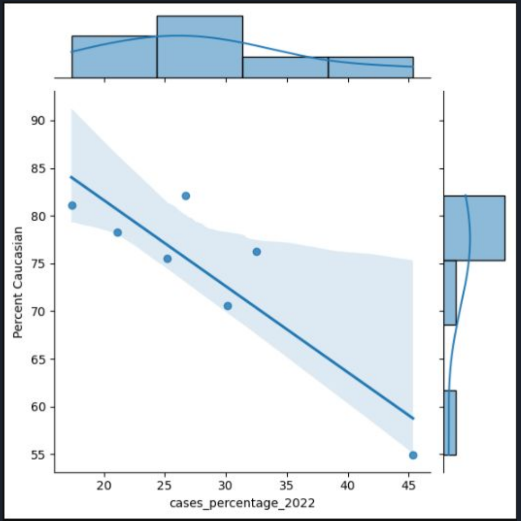


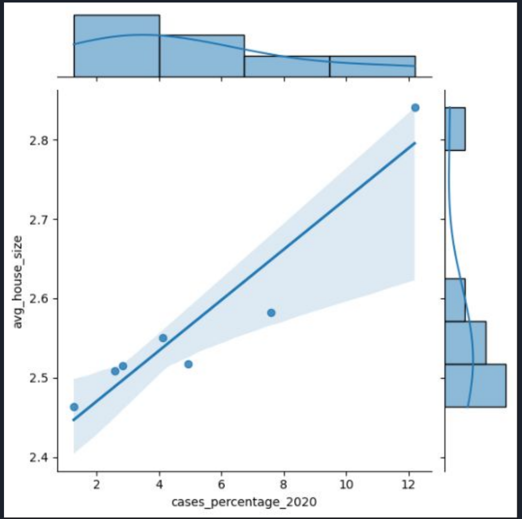
Approach 2

K Means Clustering: Approach 1

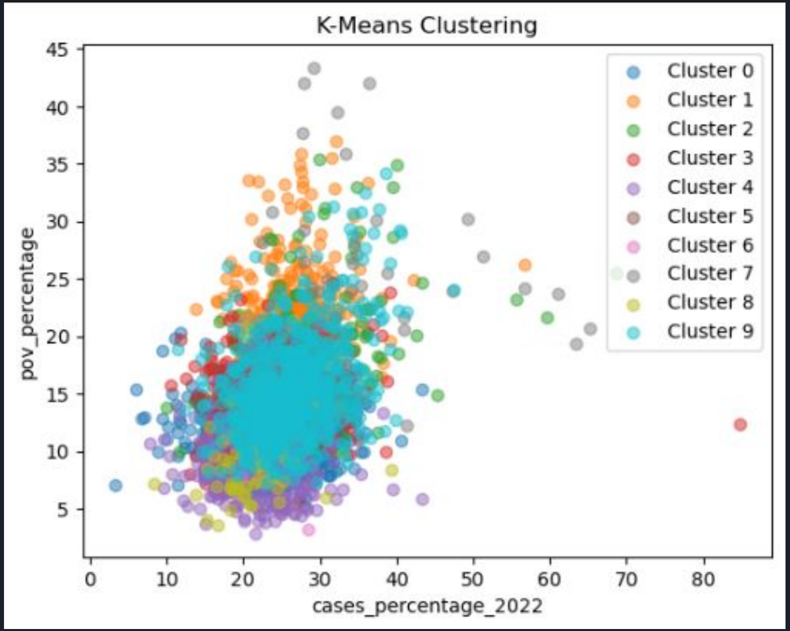


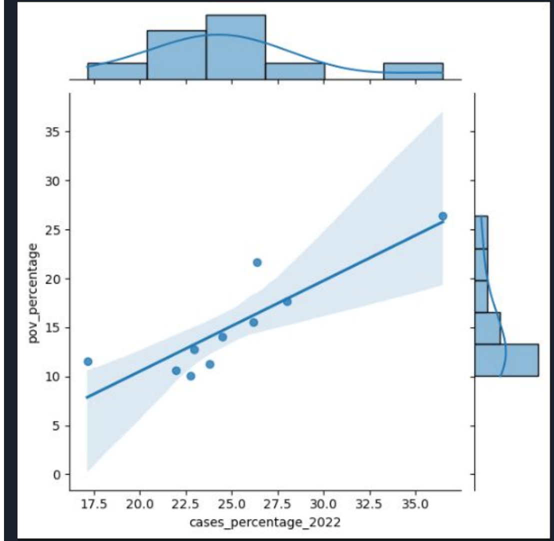


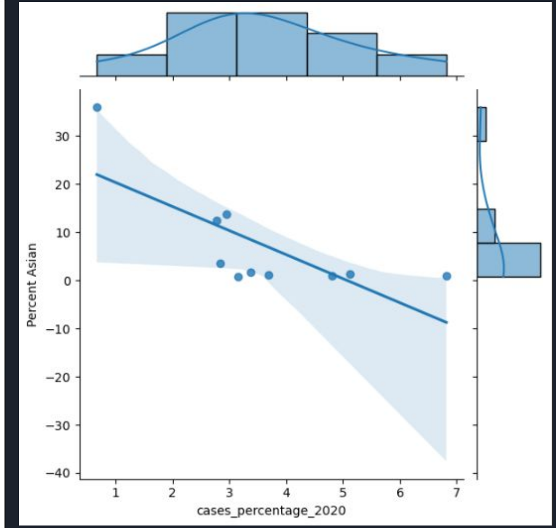




K Means Clustering: Approach 2







Analysis

- Models are not perfect
- Many external factors can influence data
- Poverty had strongest correlation with covid case percentage
- Future advice

Thank you!