Isabel Osgood and Eric Browne
University of Denver: Masters of Science, Data Science
Final Capstone

Predicting COVID-19 Impact Using Clustering Analysis

Introduction:

The COVID-19 pandemic has had a terrible toll on the world, especially the United States populace. Experts have been warning the world that a global pandemic was imminent, but those warnings were ignored. The United States's unorganized and woefully unprepared response to the pandemic resulted in nearly 30 million cases and over 525,000 deaths as of March, 2021. The 2021 American Statistical Association challenged participants to use data to guide communities to help those affected by the pandemic. The goal of this project is to highlight which communities were most negatively affected by the pandemic and direct assistance to those communities most in need.

This project will attempt to identify the societal markers associated with reporting higher Covid cases and death levels in counties across America. Identifying these markers will provide necessary data that can help shape new policy to provide these communities the assistance needed to bolster their resilience against future public health crises.

Data Overview:

1. 2019 American Community Survey Single-Year Estimates (ACS)

    This is the required dataset for the data challenge expo. The ACS is a part of the U.S. census and contains one-year statistics covering a large range of topics including employment, income, health insurance, and age. Although the survey is only available in congressional districts or counties and places with populations of 65,000 or more, using University of Michigan's Institute of Social Research's crosswalk estimates these estimates were translated into county estimates.

2. Social Vulnerability Index (SVI)

   The Center for Disease Control releases the Social Vulnerability Index every two years with 2018 being the last release. Vulnerability is measured by socioeconomic, household composition and disability, minority status, and housing and transportation.

3. COVID-19 Data

   The New York Times corona virus data collection github was used to access the most up to date case and death counts per county.

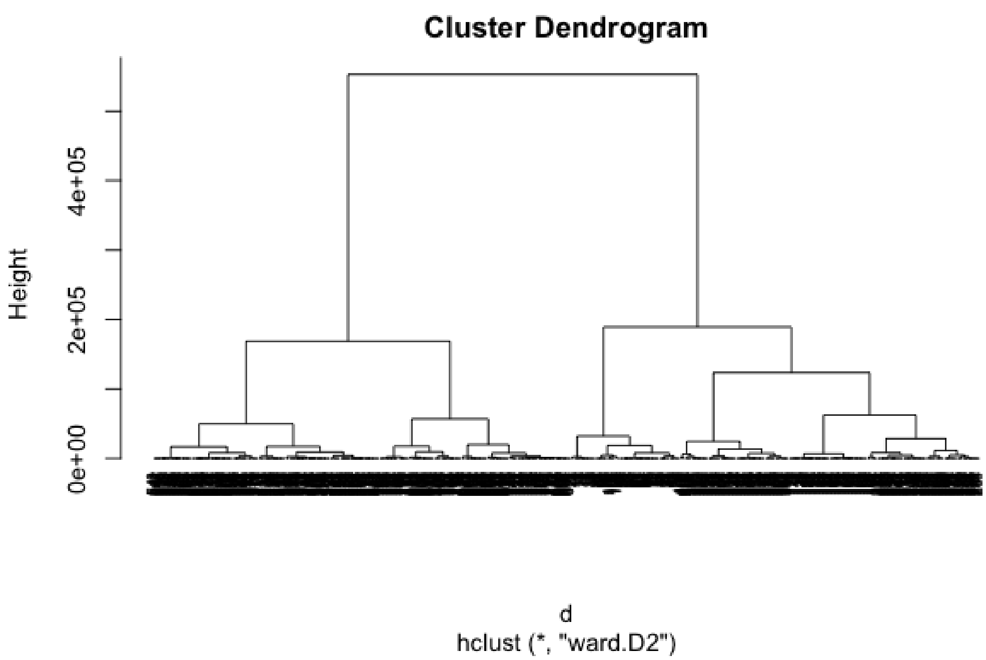
Data Preprocessing, Feature Selection:

The Social Vulnerability Index dataset was reduced to only contain the features that estimated the percentage of the county falling within a category (eg: estimated percentage of individuals without health insurance, estimated percentage of the county's population over the age 65). Every feature in the reduced SVI dataset were already scaled to the dataset creator's standards, and was not in need of further normalization as a preprocessing measure.

Methodology and Cluster Creation:

To identify the counties by risk of physical health impact of Covid, clustering was performed on the ACS and SVI datasets. Four different methods of clustering were performed and compared to achieve the best possible division of counties. Using traditional clustering methods, Ward and Complete Linkage were compared. Ward is the default method of comparing and separating clusters using sum of squared distances from the average observation. Complete Linkage was chosen because it computes groups by identifying the two least similar observations. This observation is appropriate for the given problem because differences between counties should be highlighted, not similarities. The other two clustering methods used hierarchical clustering, using both agglomerative and divisive techniques, specifically Agglomerative Nesting and Divisive Analysis.

The optimal number of clusters was chosen for each separate clustering technique using dendrograms. For example this is the dendrogram for Ward clustering. Two clusters were chosen for Ward clustering.

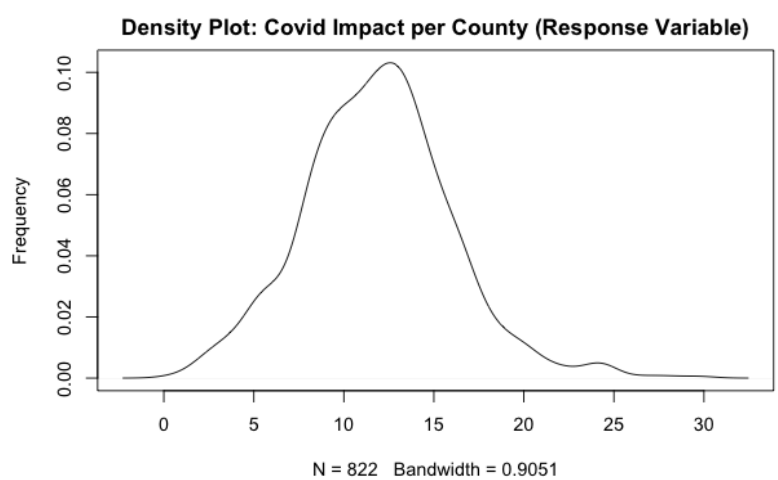


This process was repeated for each of the four techniques resulting in the following breakdown of clusters. The Cluster plots for Ward, Complete, Agglomerative Nesting, and Devising Analysis can be found in the Appendix as Figures: 1,2,3,4 respectively.

Method	Cluster 1 Size	Cluster 2 Size	Cluster 3 Size	Cluster 4 Size
Ward	409	418		
Complete	187	249	391	
Agglomerative Nesting	409	418		
Divisive Analysis	191	190	254	192

Results:

After completing each clustering task, the clusters need to be evaluated on their ability to predict Covid impact. Covid impact was measured by summing and standardizing Covid cases by population and Covid deaths by population. This data was not included in the clustering analysis to better highlight societal markers that lead to negative Covid responses as opposed to measuring which counties felt the greater negative impacts of the pandemic. First, a density plot of COVID impact was created to prove normalcy.



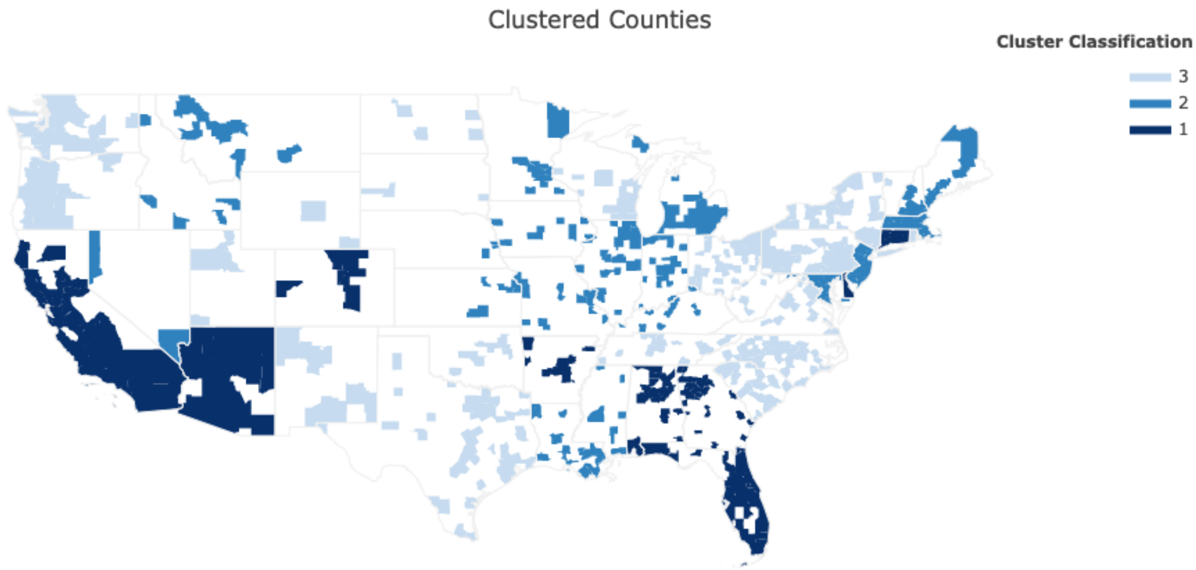
Next, a generalized linear model was created using each the Ward, Complete, Agglomerative Nesting, and Divisive Analysis clusters dataframes, which consisted of the assigned cluster of each instance and the standardized features from the SVI and ACS datasets. Importantly, only the Covid impact variable was included - total death, total cases, cases by population, and death by population were dropped. The cluster variable had a significant impact on the model for each method. The AIC (akaike information criterion) of each model was compared to determine the model with the best goodness of fit. The following shows that complete clustering created the best model according to AIC.

Model selection based on AICc:

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
complete_model	97	4113.44	0.00	0.99	0.99	-1946.59
ward_model	97	4123.67	10.23	0.01	0.99	-1951.70
agnes_model	97	4123.67	10.23	0.01	1.00	-1951.70
diana_model	97	4129.29	15.85	0.00	1.00	-1954.51
base_model	96	4137.03	23.59	0.00	1.00	-1959.67

Now that the complete clustering model has been selected, a deeper analysis may provide insights on the communities most devastated by the pandemic. Below is a map of the counties designated by cluster. Cluster 1 is associated with higher Covid impact and cluster 3 is associated with the lowest Covid impact. Interestingly, counties within the same states are clustered together, which suggests that state level policy did have an impact on the pandemic outcomes.

Isabel Osgood and Eric Browne
University of Denver: Masters of Science, Data Science
Final Capstone



The significant features (variables with a p-value of 0.001 or less) in this model can be broken up into Housing and Population Breakdown. The full model summary can be found in the appendix. For the full length of variable names within the codebook provided to us by the ASA, review the attached document.

Housing:

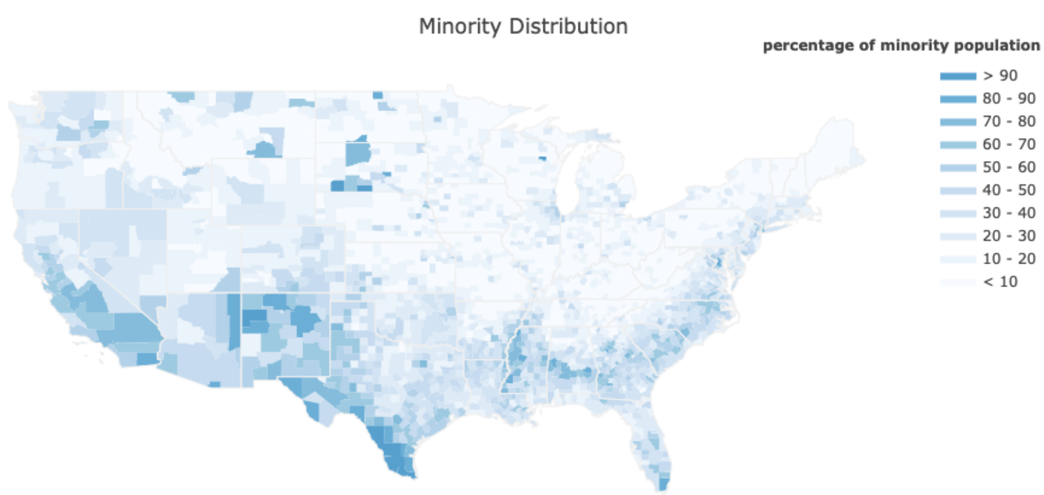
1. Estimated percentage of housing in structures with 10 or more units
2. Estimated percentage of mobile homes
3. Estimated percentage of occupied housing units with more people than rooms
4. Estimated percentile percentage households with no vehicle available

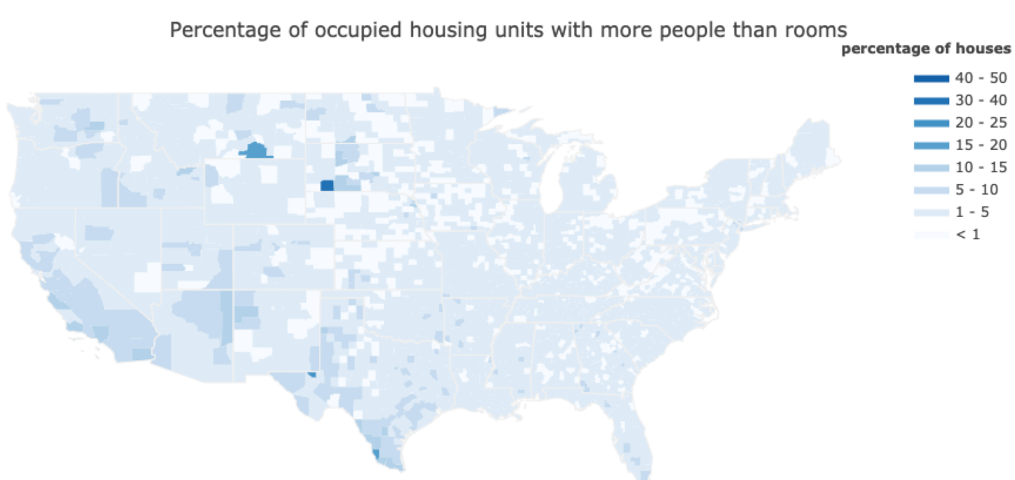
Population Breakdown:

1. Estimated percentage of population that identifies as a minority
2. Estimated percentile of unemployed

3. Estimated divorced or separated households,
4. Estimated number of unmarried women for 5 or more years
5. Estimated number of individuals born in the same state they currently reside
6. Estimated number of households with a primary language other than English.

Although these variables are often associated with low income or high/low populations, poverty and population were not significant in this model. Looking at a map of minority population and estimated percentage of houses with more occupants than rooms, a clear pattern begins to emerge.





Conclusion:

While the COVID-19 pandemic was catastrophic to our everyday life, public health, and social well-being, it is vital that the United States must learn from this historical event to better prepare for its future. Before undergoing the analysis of this project, one of the underlying assumptions was that features and characteristics of groups of people that were statistically significant in predicting the overall impact on their county, were also the most vulnerable to future pandemics. Additionally, we assumed that Social Vulnerability Index Data from 2 years prior (2018), would be relatively similar to that of the peak year of the COVID-19 pandemic (2020). After validating those assumptions, we feature engineered a variable to predict: COVID_IMPACT, that of which being the sum of the amount of cases and deaths divided by the population for that specific county. Then, we first clustered counties together based on features included in the 2019 ACS estimates and the Social Vulnerability Index estimates from the year prior. After conducting 4 different clustering methods: Ward, Complete, Agglomerative Clustering, and Diana, each of their cluster labels were evaluated using a GLM and Linear Regression model for predicting COVID_IMPACT; where the Complete Clustering method achieved the lowest AIC and RMSE from the respective models. From the Clustering GLM and Linear Model, we found the most significant features to be used in predicting COVID_IMPACT, listed above in the previous section.

The theme for the 2021 ASA Data Challenge Exposition is that of “Helping Families, Businesses, and Communities Respond to COVID-19”. We believe that the best way to help the citizens of the United States respond to the pandemic is to properly invest in its demographics that are most vulnerable to the impact of the COVID-19 pandemic. Therefore, our proposal to the United States legislature is to invest in and prioritize pandemic aid for the statistically significant categories displayed in our analysis, in order to achieve a more efficient welfare in the future public health emergencies.

Isabel Osgood and Eric Browne
University of Denver: Masters of Science, Data Science
Final Capstone

Appendix

Figure 1: WARD Clustering Chart

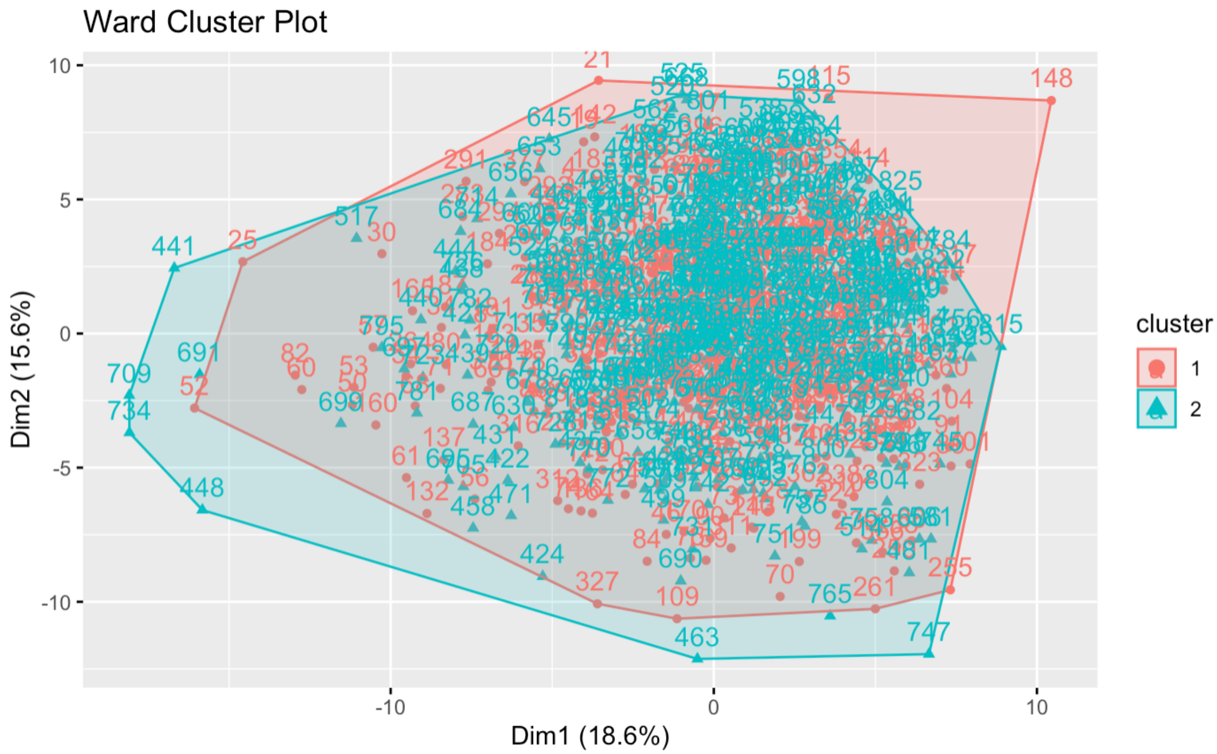


Figure 2: COMPLETE Clustering Chart

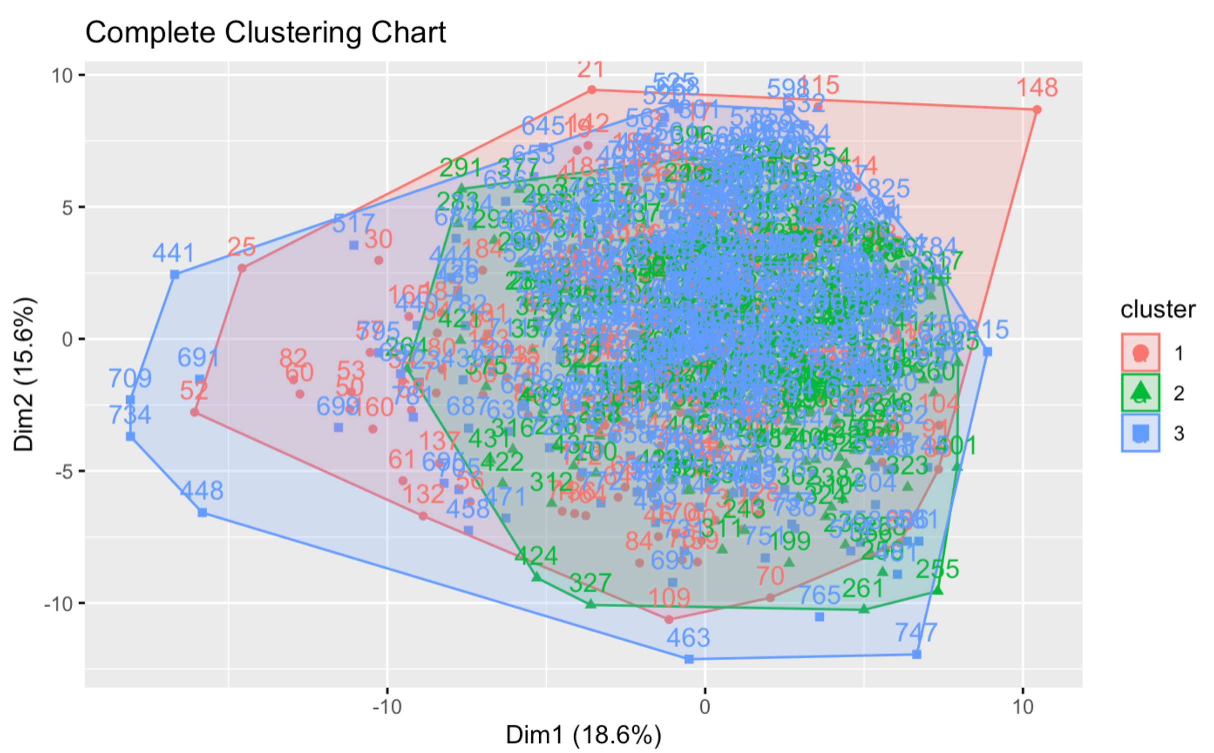


Figure 3: AGNES Clustering Chart

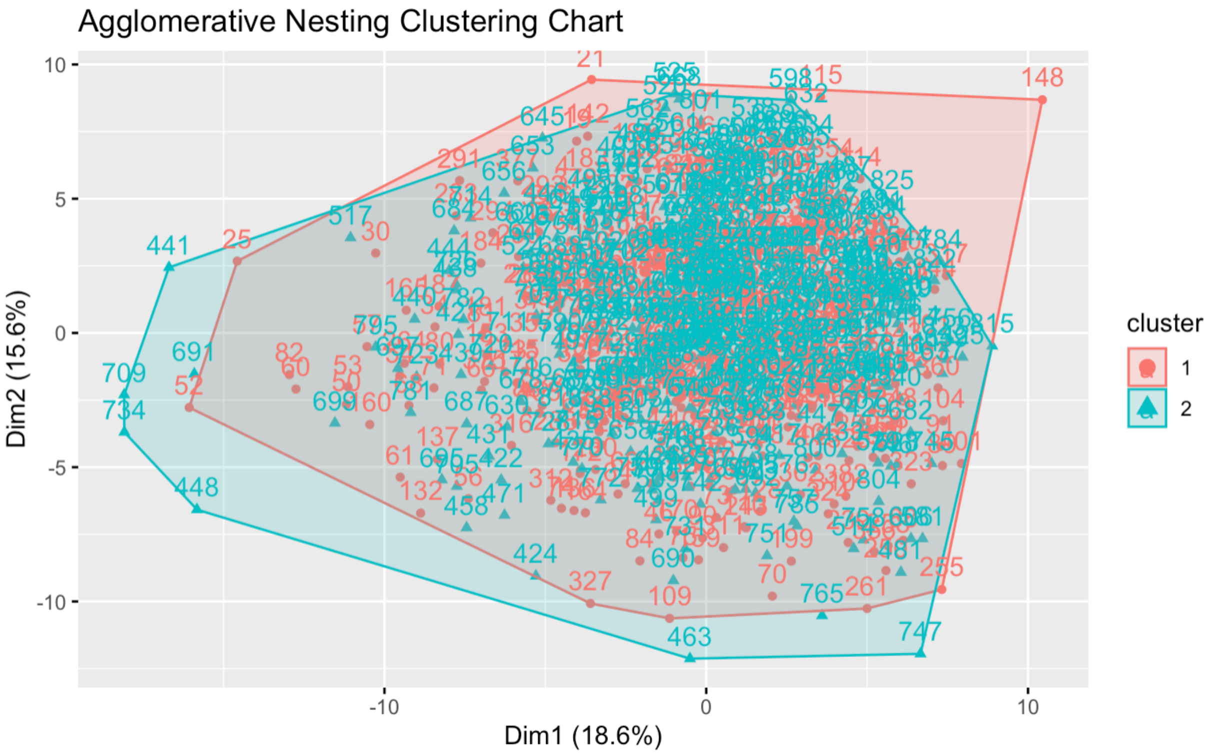


Figure 4: DIANA Clustering Chart

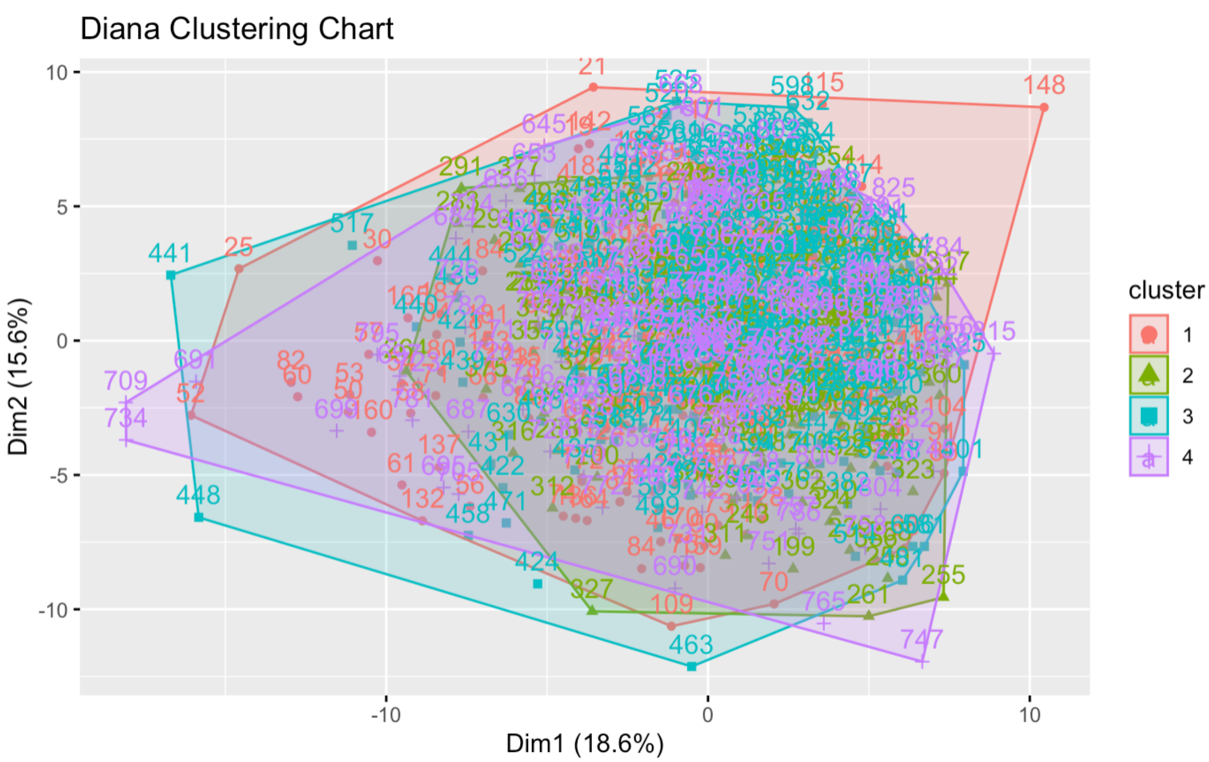


Figure 5: COVID_IMPACT Density Distribution

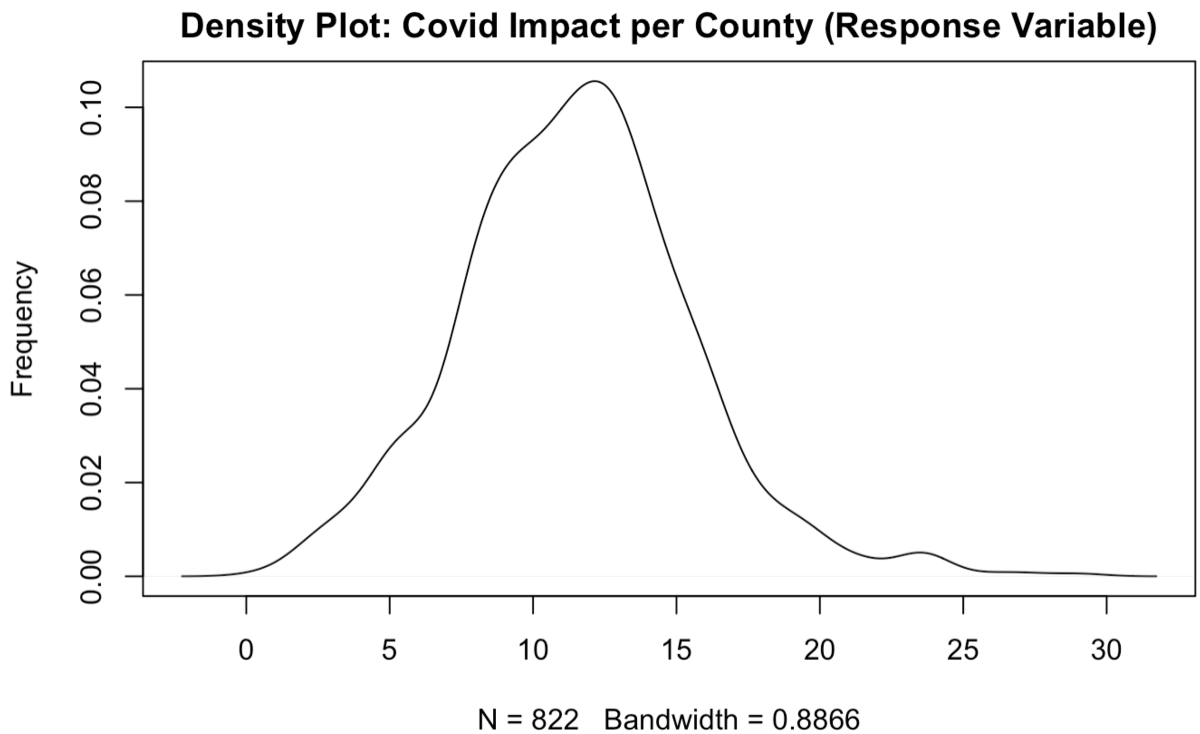


Figure 6: Complete GLM Summary

Call:
glm(formula = covid_impact ~ ., data = complete_dat)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-7.8574	-1.7718	-0.1822	1.6828	9.7956

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.371e+01	3.722e-01	36.825	< 2e-16	***
population	-2.750e-07	2.376e-07	-1.157	0.247528	
EP_POV	6.639e-01	6.343e-01	1.047	0.295647	
EP_AGE65	1.308e+00	7.992e-01	1.636	0.102177	
EP_AGE17	3.789e-01	7.278e-01	0.521	0.602817	
EP_DISABL	3.651e-02	8.420e-01	0.043	0.965428	
EP_SNGPNT	-3.615e-01	6.378e-01	-0.567	0.570999	
EP_MINRTY	-3.903e+00	5.646e-01	-6.913	1.04e-11	***
EP_LIMENG	8.931e-01	4.961e-01	1.800	0.072223	.
EP_MUNIT	1.184e+00	2.980e-01	3.973	7.80e-05	***
EP_MOBILE	1.326e+00	5.064e-01	2.618	0.009027	**
EP_CROWD	-1.151e+00	3.478e-01	-3.310	0.000980	***
EP_NOVEH	-1.850e-01	3.523e-01	-0.525	0.599617	
EP_GROUPQ	1.582e-01	2.534e-01	0.625	0.532452	
EPL_POV	3.965e-02	5.966e-01	0.066	0.947024	

EPL_UNEMP	-7.911e-01	2.086e-01	-3.792	0.000162	***
EPL_PCI	3.405e-02	3.195e-01	0.107	0.915173	
EPL_NOHSDP	7.400e-01	5.182e-01	1.428	0.153755	
EPL_AGE65	-5.129e-01	3.879e-01	-1.322	0.186543	
EPL_AGE17	2.806e-01	4.875e-01	0.576	0.565049	
EPL_DISABL	-8.129e-01	7.281e-01	-1.116	0.264623	
EPL_SNGPNT	1.416e-01	4.708e-01	0.301	0.763657	
EPL_MINRTY	1.950e+00	4.653e-01	4.190	3.14e-05	***
EPL_LIMENG	-5.785e-01	2.596e-01	-2.229	0.026142	*
EPL_MUNIT	-9.839e-02	2.150e-01	-0.458	0.647324	
EPL_MOBILE	-1.501e+00	5.323e-01	-2.820	0.004939	**
EPL_CROWD	5.837e-01	2.643e-01	2.209	0.027496	*
EPL_NOVEH	-1.012e+00	2.600e-01	-3.890	0.000109	***
EPL_GROUPQ	4.295e-01	2.000e-01	2.147	0.032114	*
EP_UNINSUR	1.890e-01	2.053e-01	0.921	0.357552	
With.own.children.of.the.householder.under.18.years	4.927e-01	5.561e-01	0.886	0.375981	
Cohabiting.couple.household	6.712e-01	8.694e-01	0.772	0.440329	
Male.householder..no.spouse.partner.present	-2.604e-01	6.858e-01	-0.380	0.704307	
Householder.living.alone	-7.919e-01	7.061e-01	-1.122	0.262434	
Female.householder..no.spouse.partner.present	2.180e+00	1.038e+00	2.100	0.036108	*
Households.with.one.or.more.people.under.18.years	-2.808e-01	6.071e-01	-0.463	0.643800	
Households.with.one.or.more.people.65.years.and.over	-6.178e-01	8.084e-01	-0.764	0.444986	
Average.household.size	-4.653e-01	1.476e+00	-0.315	0.752695	
Average.family.size	1.048e+00	1.291e+00	0.812	0.417021	
Householder	5.088e+00	9.967e+00	0.510	0.609892	
Spouse	9.334e-01	7.351e+00	0.127	0.899000	
Unmarried.partner	-5.211e-01	2.279e+00	-0.229	0.819227	
Child	4.624e+00	1.066e+01	0.434	0.664744	
Other.relatives	3.236e+00	7.109e+00	0.455	0.649098	
Other.nonrelatives	1.452e+00	6.099e+00	0.238	0.811897	
Never.married	1.497e-01	8.435e-01	0.177	0.859193	
Now.married..except.separated	1.779e+00	8.567e-01	2.076	0.038233	*
Separated	-6.804e-01	2.755e-01	-2.470	0.013749	*
Widowed	2.355e-01	4.141e-01	0.569	0.569675	
Divorced	-1.289e+00	3.668e-01	-3.513	0.000470	***
Number.of.women.15.to.50.years.old.who.had.a.birth.in.the.past.12.months	-7.984e-02	4.419e-01	-0.181	0.856676	
Unmarried.women..widowed..divorced..and.never.married.	1.349e-01	2.530e-01	0.533	0.594086	
Less.than.1.year	-4.274e-01	2.698e-01	-1.584	0.113541	
X1.or.2.years	-6.789e-02	2.863e-01	-0.237	0.812649	
X3.or.4.years	2.003e-01	2.574e-01	0.778	0.436810	
X5.or.more.years	9.804e-01	3.305e-01	2.966	0.003115	**
Number.of.grandparents.responsible.for.own.grandchildren.under.18.years	-4.990e-01	2.835e-01	-1.760	0.078806	.
Who.are.female	-5.139e-01	5.301e-01	-0.970	0.332614	
Who.are.married	-2.083e-01	4.334e-01	-0.481	0.630879	
Nursery.school..preschool	8.211e-01	5.109e+00	0.161	0.872365	
Kindergarten	8.541e-01	4.596e+00	0.186	0.852613	
Elementary.school..grades.1.8.	3.120e+00	1.995e+01	0.156	0.875765	
High.school..grades.9.12.	2.233e+00	1.225e+01	0.182	0.855421	
College.or.graduate.school	5.382e+00	3.270e+01	0.165	0.869303	
Less.than.9th.grade	2.140e+01	1.183e+01	1.810	0.070761	.

```
X9th.to.12th.grade..no.diploma                2.155e+01  1.138e+01   1.894 0.058598 .
High.school.graduate..includes.equivalency.  -1.750e+01  2.271e+01  -0.771 0.441247
Some.college..no.degree                      -1.030e+01  1.278e+01  -0.806 0.420697
Associate.s.degree                           -5.264e+00  7.021e+00  -0.750 0.453676
Bachelor.s.degree                             3.675e+01  2.491e+01   1.475 0.140581
Graduate.or.professional.degree               3.155e+01  2.201e+01   1.434 0.152109
High.school.graduate.or.higher                5.138e+01  2.285e+01   2.249 0.024813 *
Bachelor.s.degree.or.higher                  -9.328e+01  4.787e+01  -1.949 0.051732 .
Civilian.veterans                            -3.387e-01  2.312e-01  -1.465 0.143366
With.a.disability                            -1.365e-02  2.909e-01  -0.047 0.962592
Same.house                                    1.045e+00  1.825e+01   0.057 0.954384
Different.house.in.the.U.S.                   7.033e+00  2.039e+01   0.345 0.730242
Same.county                                  -3.295e+00  1.183e+01  -0.278 0.780713
Different.county                             -2.595e+00  1.081e+01  -0.240 0.810382
Same.state                                   -8.253e-01  3.174e-01  -2.600 0.009503 **
Different.state                               8.088e+01  5.017e+01   1.612 0.107388
Abroad                                       -5.101e-01  2.007e+00  -0.254 0.799440
Native                                        6.414e-01  4.229e-01   1.517 0.129794
Born.in.United.States                        -4.883e+01  3.058e+01  -1.597 0.110739
State.of.residence                            9.337e+01  5.789e+01   1.613 0.107178
Born.in.Puerto.Rico..U.S..Island.areas..or.born.abroad.to.American.parent.s.
                                             -1.557e-01  1.641e-01  -0.949 0.343071
Foreign.born                                  7.291e-01  4.926e-01   1.480 0.139297
Foreign.born.population                       1.798e+00  1.437e+00   1.251 0.211502
Naturalized.U.S..citizen                     -1.780e+00  1.138e+00  -1.564 0.118213
Not.a.U.S..citizen                           -1.411e+00  1.139e+00  -1.239 0.215674
Population.born.outside.the.United.States     1.757e+00  2.505e+00   0.701 0.483405
Entered.2010.or.later                        -1.710e-01  2.450e-01  -0.698 0.485293
Entered.before.2010                          -5.687e-01  2.991e-01  -1.901 0.057681 .
English.only                                 -1.901e-01  2.112e-01  -0.900 0.368361
Language.other.than.English                   2.645e+00  5.195e-01   5.092 4.51e-07 ***
cluster_complete                             -7.581e-01  1.564e-01  -4.845 1.55e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7.557495)

    Null deviance: 14125.8  on 821  degrees of freedom
Residual deviance:  5486.7  on 726  degrees of freedom
AIC: 4087.2

Number of Fisher Scoring iterations: 2
```