

S&DS 563/S&DS 363 Final Project

15 May 2021

Introduction

With the nadir of the SARS-CoV-2 pandemic finally subsiding, the task of understanding the epidemiological factors contributing to the propagation of COVID-19 has only begun. Understanding relationships between COVID-19 spread prevention and socioeconomic variables will prove vital to inform us of how to mitigate the propagation of the next pandemic. This report aims to understand how the economic, education, behavioral, and population data for the 3141 U.S. counties relate to COVID-19 infection and death rate data.

We acknowledge that the impact of and response to COVID-19 has been very different from county to county in the United States. Looking at the current COVID-19 vaccination data in mid-May 2020, we note that the vaccination rate for ages 18+ ranges drastically from 11% in some counties in Louisiana to 74% in some counties in New York. Our guiding question is: Who is the most vulnerable to COVID-19 infection and death? This knowledge will guide public health efforts as we continue to fight against the spread of COVID-19. Knowledge of what socioeconomic factors put people at risk will allow us to prioritize our vaccination and education efforts from those who need it the most and will also let us take a step back to acknowledge the systemic health inequalities in our country.

Design and Primary Questions

This report deploys three multivariate techniques to examine the following questions:

1. How do the 3141 counties differ from one another, i.e., how do the socioeconomic and COVID-19 data relate to one another when distinguishing U.S. counties? Principal component analysis (PCA) will help to reduce the dimensionality of our large dataset, increasing interpretability of underlying trends between clusters of variables. This metric technique works on the columns of our dataset to reduce them into composite variables and make them more interpretable.
2. Which U.S. states are similar to one another? Cluster analysis will enable the clustering of states into a discrete number of groups based on similar socioeconomic and COVID-19 data. This metric technique works on the rows of our dataset to find similar groups of observations.
3. Which U.S. county variable pairings are similar to one another? Correspondence analysis is similar to PCA but applies to categorical rather than continuous variables. This nonmetric technique works on both the columns and rows of our dataset to visualize which rows and column points are similar in lower-dimensional space.

Using these techniques, we will be able to better understand our variables, our observations, and the interactions between our variables and observations. Who is most vulnerable to COVID-19 infection and death? This allows us to direct resources to protecting these vulnerable populations.

Data

The dataset referenced in this report includes COVID-19 infection and death statistics from U.S. counties (sourced from Johns Hopkins, as of 28 April 2021), combined with economic, education, and population data (sourced from various government agencies) and also survey responses about mask-wearing frequencies (sourced from NYT) for a total of 3141 complete observations on 10 continuous variables and 6 categorical variables. Continuous variables were rescaled as percentages of county population.

- **6 categorical variables:** FIPS, county name, state name, rural urban type, rural urban code, economic typology
- **9 continuous variables:** “Always” wear mask survey response percent, unemployment rate, median household income, percent poverty, percent of adults with less than a high school education, death rate, percent civilian labor force, percent of county population that has had confirmed COVID-19 cases, and percent of county population that has died from COVID-19.

[1] “FIPS” = State-County FIPS Code; Categorical (identifier)

[2] “County_Name” = US County Name; Categorical (identifier)

[3] “State_Name” = US State Name; Categorical

[4] “Rural_Urban_Type” = Regrouping of Rural-Urban Codes (2013) numbered 1-9 according to descriptions provided by the USDA. See variable [5]. Regroup codes 1 through 9 into three groups: (1) “Urban” for codes 1-3, (2) “Suburban” for codes 4-6, and (3) “Rural” for codes 7-9; Categorical (1-3)

[5] “Rural_Urban_Code_2013” = Rural-urban Continuum Code, 2013; (<https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/>); Categorical (1-9)

[6] “Economic_Typology_2015” = County economic types, 2015 edition (<https://www.ers.usda.gov/data-products/county-typology-codes/>); Non-overlapping economic-dependence county indicator. 0=Nonspecialized 1=Farm-dependent 2=Mining-dependent 3=Manufacturing-dependent 4=Federal/State government-dependent 5=Recreation; Categorical (0-5)

[7] “Always_Wear_Mask_Survey” = “Always” response. The New York Times administered a survey to 250,000 Americans from July 2 to July 14 asking the following question: How often do you wear a mask in public when you expect to be within six feet of another person?; Continuous (%)

[8] “Unemployment_Rate_2019” = Unemployment rate, 2019; Continuous (%)

[9] “Median_Household_Income_2019” = Estimate of median household Income, 2019; Continuous (\$)

[10] “Percent_Poverty_2019” = Estimate of people of all ages in poverty 2019; Continuous (%)

[11] “Percent_Adults_Less_Than_HS” = Percent of adults with less than a high school diploma, 2014-18

[12] “Death_Rate_2019” = Death rate in period 7/1/2018 to 6/30/2019; Continuous (%)

[13] “Civilian_Labor_Force_2019_as_pct” = Civilian labor force annual average, 2019, expressed as percent; Continuous (%)

[14] “Covid_Confirmed_Cases_as_pct” = Cumulative sum of COVID-19 cases expressed as percent. Reported from Johns Hopkins on 28 April 2021; Continuous (%)

[15] “Covid_Deaths_as_pct” = Cumulative sum of COVID-19 deaths expressed as percent. Reported from Johns Hopkins on 28 April 2021; Continuous (%)

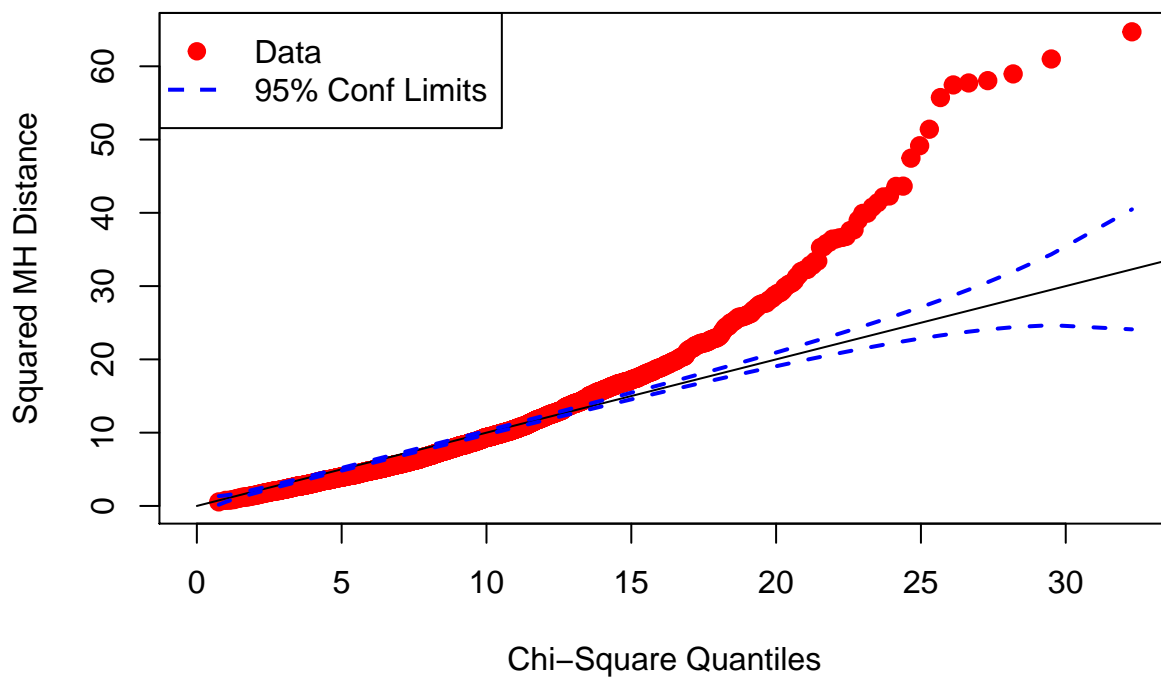
Descriptive Plots and Summary Statistics

Data Transformation

We made normal quantile plots for each of the 9 continuous variables in the dataset. This revealed that most variables initially did not have a univariate normal distribution. Taking the log-transform of the 10 continuous variables helped most variables have more linear quantile plots. Note that we also standardized the continuous variables since they were measured on different scales. Moreover, for death rate, percent COVID-19 cases, and percent COVID-19 deaths, a $1.5 \times \text{IQR}$ outlier exclusion method was applied to enable these variables to take on more normal univariate distributions. Note that the outlier exclusion method reduced the number of counties that we will analyze to 2,814 observation. Hence, outlier exclusion reduced the dataset by approximately 10%. This percent excluded is relatively substantial; however, we deemed the benefits of having univariate distributions outweighed this disadvantage. With these changes made, the 9 continuous variables all had univariate normal distributions.

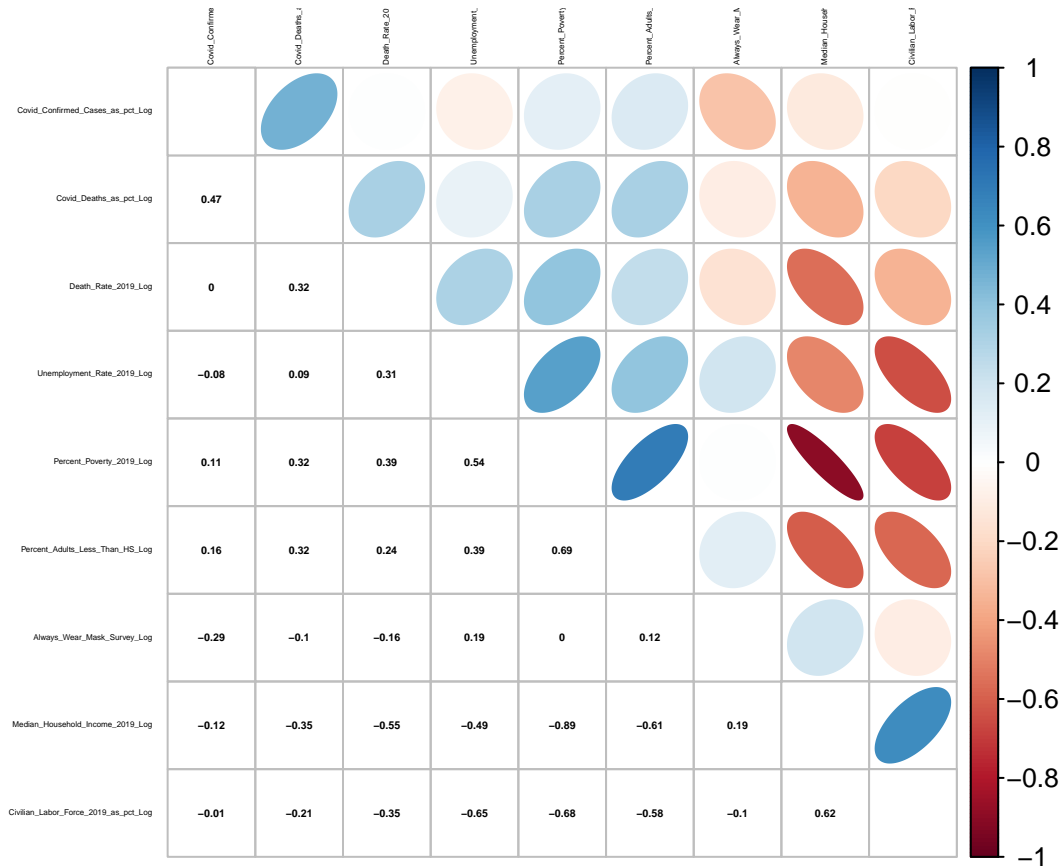
Lack of Multivariate Normality

Chi-Square Quantiles for Counties



A chi-square quantile plot (shown above) reflects that our data does not have a multivariate normal distribution. Thus, none of the techniques we use will require a multivariate normal distribution.

Variable Correlation



We note many variables highly correlated with other variables, which is appropriate for PCA. For instance, the correlation between the log of the unemployment rate and the labor force as a percent is -.065, the correlation between the log of the median household income and percent poverty is -.88, and the correlation between the percent of COVID-19 cases and the percent of COVID-19 deaths is 0.47. There appear to be underlying trends about the counties (about beliefs about COVID-19, about wealth/education, etc) that could be summarized in linear combinations of the 19 metric variables we have currently.

Summary Statistics

```
## [1] 2814 15
```

```
##
## Rural Suburban Urban
## 907 841 1066
```

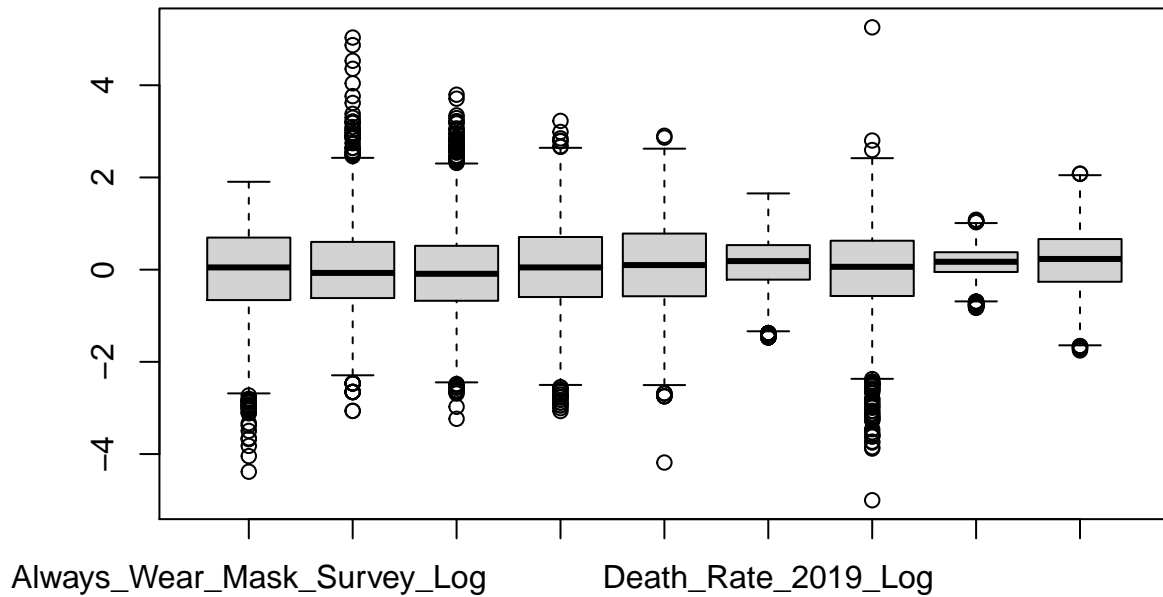
```
##
## 0 1 2 3 4 5
## 1146 392 196 480 351 249
```

```
## Economic_Typology_2015 Always_Wear_Mask_Survey_Log Unemployment_Rate_2019_Log
## Min. :0.000 Min. : -4.38292 Min. : -3.06445
## 1st Qu.:0.000 1st Qu.: -0.65971 1st Qu.: -0.61644
## Median :1.000 Median : 0.04834 Median : -0.07051
## Mean :1.732 Mean : -0.02779 Mean : 0.01558
## 3rd Qu.:3.000 3rd Qu.: 0.69544 3rd Qu.: 0.60164
## Max. :5.000 Max. : 1.90553 Max. : 5.03446
```

```

## Median_Household_Income_2019_Log Percent_Poverty_2019_Log
## Min.      :-3.23567              Min.      :-3.06671
## 1st Qu.: -0.67529              1st Qu.: -0.59390
## Median : -0.09013              Median :  0.04824
## Mean    : -0.05574              Mean     :  0.04623
## 3rd Qu.:  0.51653              3rd Qu.:  0.70815
## Max.     :  3.79491              Max.      :  3.22673
## Percent_Adults_Less_Than_HS_Log Death_Rate_2019_Log
## Min.      :-4.18428              Min.      :-1.4689
## 1st Qu.: -0.57692              1st Qu.: -0.2168
## Median :  0.09978              Median :  0.1861
## Mean     :  0.07065              Mean     :  0.1237
## 3rd Qu.:  0.78307              3rd Qu.:  0.5328
## Max.      :  2.90401              Max.      :  1.6538
## Civilian_Labor_Force_2019_as_pct_Log Covid_Confirmed_Cases_as_pct_Log
## Min.      :-5.00174              Min.      :-0.82906
## 1st Qu.: -0.57127              1st Qu.: -0.05023
## Median :  0.06175              Median :  0.17048
## Mean     : -0.03923              Mean     :  0.15011
## 3rd Qu.:  0.62656              3rd Qu.:  0.37812
## Max.      :  5.25580              Max.      :  1.08067
## Covid_Deaths_as_pct_Log
## Min.      :-1.7523
## 1st Qu.: -0.2623
## Median :  0.2318
## Mean     :  0.1925
## 3rd Qu.:  0.6637
## Max.      :  2.0837

```



From the 2,814 US counties that we are analyzing, there is a fairly equitable distribution of rural-urban type: 907 rural, 841 suburban, and 1066 urban. The distributions in the quantitative variables are consistent, as we expected considering our previous standardization operation of continuous variables.

Principle Component Analysis

We use PCA to reduce the dimensionality of our dataset to find composite variables that are linear combinations of our metric variables. Note that the multivariate normality and variable correlations were already assessed in the “Descriptive Plot and Summary Statistics” section and determined to be suitable for PCA, though parallel analysis may not be used due to the lack of multivariate normality.

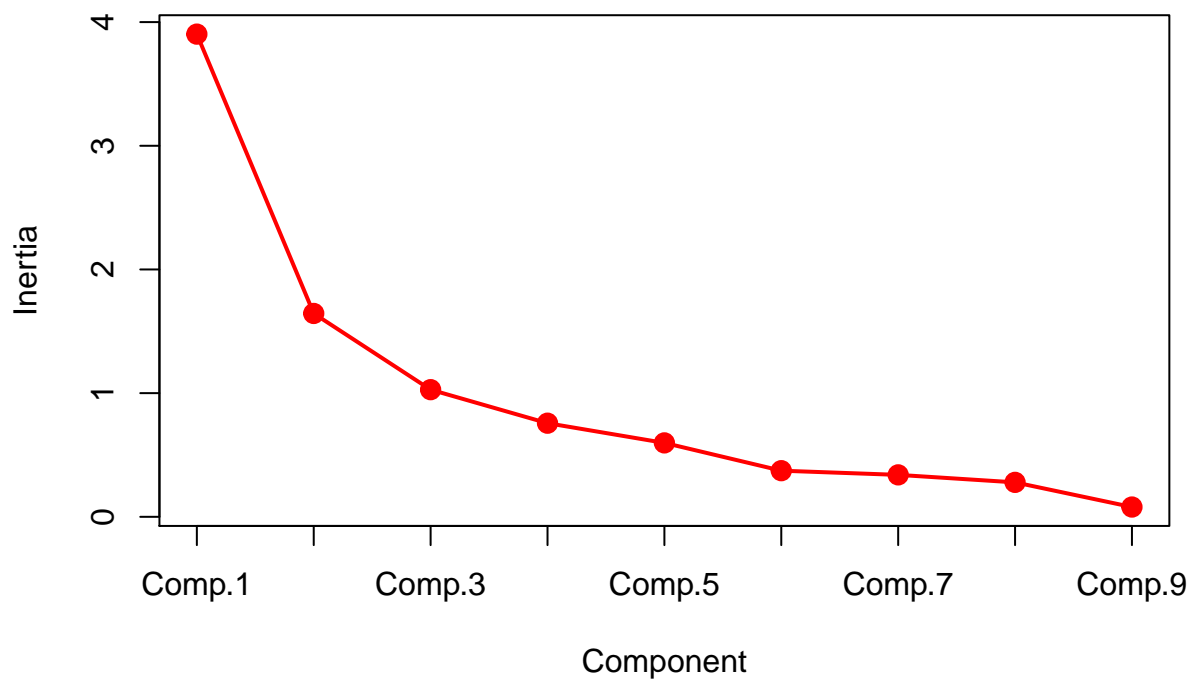
```
## Importance of components:
##               Comp.1   Comp.2   Comp.3   Comp.4   Comp.5
## Standard deviation    1.9755479 1.2823031 1.0139924 0.87008610 0.77310275
## Proportion of Variance 0.4336433 0.1827002 0.1142423 0.08411665 0.06640976
## Cumulative Proportion 0.4336433 0.6163434 0.7305857 0.81470236 0.88111212
##               Comp.6   Comp.7   Comp.8   Comp.9
## Standard deviation    0.61072705 0.58241257 0.52776316 0.281540505
## Proportion of Variance 0.04144306 0.03768938 0.03094822 0.008807228
## Cumulative Proportion 0.92255518 0.96024455 0.99119277 1.000000000

## Warning in if (loadings) {: the condition has length > 1 and only the first
## element will be used
```

```
##
## Loadings:
##
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Always_Wear_Mask_Survey_Log      0.01  0.52  0.57  0.45  0.11  0.25
## Unemployment_Rate_2019_Log     -0.34  0.32 -0.01  0.02 -0.70 -0.07
## Median_Household_Income_2019_Log  0.45  0.07  0.21  0.14 -0.21  0.03
## Percent_Poverty_2019_Log        -0.46  0.04  0.01 -0.22  0.22 -0.07
## Percent_Adults_Less_Than_HS_Log  -0.39  0.04  0.32 -0.20  0.44  0.23
## Death_Rate_2019_Log             -0.29 -0.11 -0.51  0.62  0.03  0.47
## Civilian_Labor_Force_2019_as_pct_Log 0.41 -0.20 -0.04  0.13  0.27  0.14
## Covid_Confirmed_Cases_as_pct_Log  -0.09 -0.60  0.40 -0.16 -0.38  0.52
## Covid_Deaths_as_pct_Log         -0.23 -0.46  0.31  0.52  0.02 -0.60
##
## Comp.7 Comp.8 Comp.9
## Always_Wear_Mask_Survey_Log      0.25  0.21  0.12
## Unemployment_Rate_2019_Log     0.15 -0.51  0.00
## Median_Household_Income_2019_Log -0.41 -0.10 -0.71
## Percent_Poverty_2019_Log        0.45  0.16 -0.68
## Percent_Adults_Less_Than_HS_Log  -0.46 -0.50  0.02
## Death_Rate_2019_Log            -0.14  0.01 -0.14
## Civilian_Labor_Force_2019_as_pct_Log 0.54 -0.62 -0.05
## Covid_Confirmed_Cases_as_pct_Log  0.12  0.14  0.02
## Covid_Deaths_as_pct_Log        -0.04 -0.08  0.00

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
##   3.90   1.64   1.03   0.76   0.60   0.37   0.34   0.28   0.08
```

Scree Plot of Transformed COVID-19 Data



1 2 3 4 5

## Percent_Poverty_2019_Log	-0.46	0.04	0.01	-0.22	0.22	-0.07
## Percent_Adults_Less_Than_HS_Log	-0.39	0.04	0.32	-0.20	0.44	0.23
## Death_Rate_2019_Log	-0.29	-0.11	-0.51	0.62	0.03	0.47
## Civilian_Labor_Force_2019_as_pct_Log	0.41	-0.20	-0.04	0.13	0.27	0.14
## Covid_Confirmed_Cases_as_pct_Log	-0.09	-0.60	0.40	-0.16	-0.38	0.52
## Covid_Deaths_as_pct_Log	-0.23	-0.46	0.31	0.52	0.02	-0.60
##	Comp.7	Comp.8	Comp.9			
## Always_Wear_Mask_Survey_Log	0.25	0.21	0.12			
## Unemployment_Rate_2019_Log	0.15	-0.51	0.00			
## Median_Household_Income_2019_Log	-0.41	-0.10	-0.71			
## Percent_Poverty_2019_Log	0.45	0.16	-0.68			
## Percent_Adults_Less_Than_HS_Log	-0.46	-0.50	0.02			
## Death_Rate_2019_Log	-0.14	0.01	-0.14			
## Civilian_Labor_Force_2019_as_pct_Log	0.54	-0.62	-0.05			
## Covid_Confirmed_Cases_as_pct_Log	0.12	0.14	0.02			
## Covid_Deaths_as_pct_Log	-0.04	-0.08	0.00			

Looking at PC1: This principle component seems to be related to wealth, employment status, and education. It combines log percent poverty (-0.46) with the log of the median household income (0.45), the log of the civilian labor force (0.41), and the percent of adults with a bachelor's degree or higher (-0.39). A higher value on this PC indicates more employment, more jobs, and more education.

Looking at PC2: This principle component seems to be a measure of masking behaviors and relation to COVID-19 infection. It combines the percentage of those who say they always mask (0.52), the log of the COVID-19 infection rate (-0.60), and the log of the COVID-19 death rate (-0.46). A higher value on this PC indicates more masking and less COVID-19 infection and death.

Looking at PC3: This principle component seems to be a combination of the first two and relates underlying traits about the county to COVID-19 infection. It combines the log of the 2019 death rate (-0.51) with the cumulative percentage of population with the log of the percentage of adults with less than a high school degree (0.32), the log of the COVID-19 infection rate (0.40), and the log of the COVID-19 death rate (0.31). A higher value on this PC indicates less education and more COVID-19 infection and death.

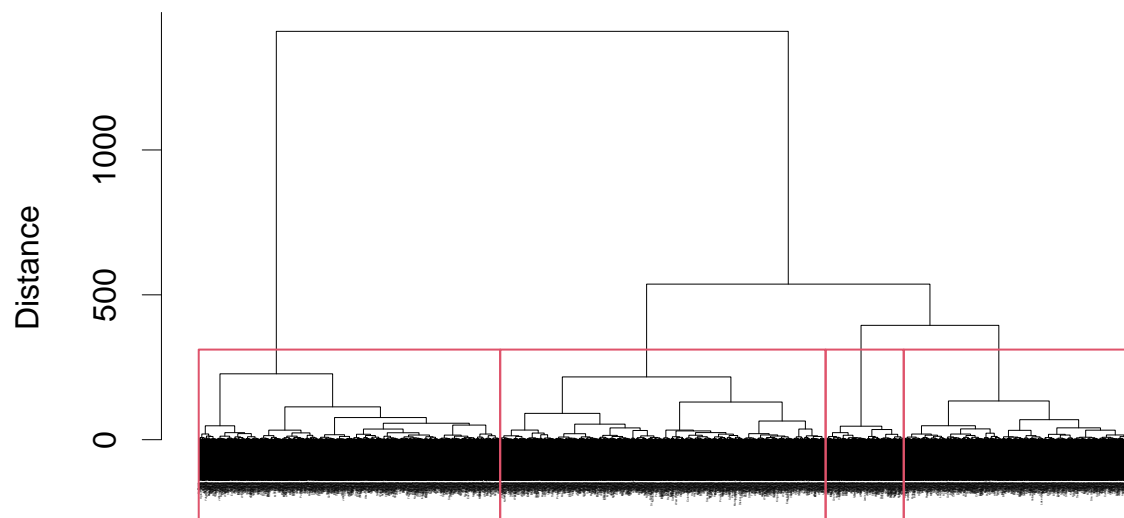
Using PCA, we can reduce 9 metric variables to 3 composite variables that are related to wealth and education, attitudes about masking, and population. These 3 PC's can account for 73% of the total variability. We note that COVID-19 infection and death rates are related to attitudes and behaviors, like the masking rate, but are also due to factors outside of the control of a county's population, like unemployment and education.

Cluster Analysis

We use cluster analysis to find groups of counties that are similar to each other but different from other counties across our metric variables. We are finding clusters of *observations*, unlike in PCA where we found clusters of *variables*.

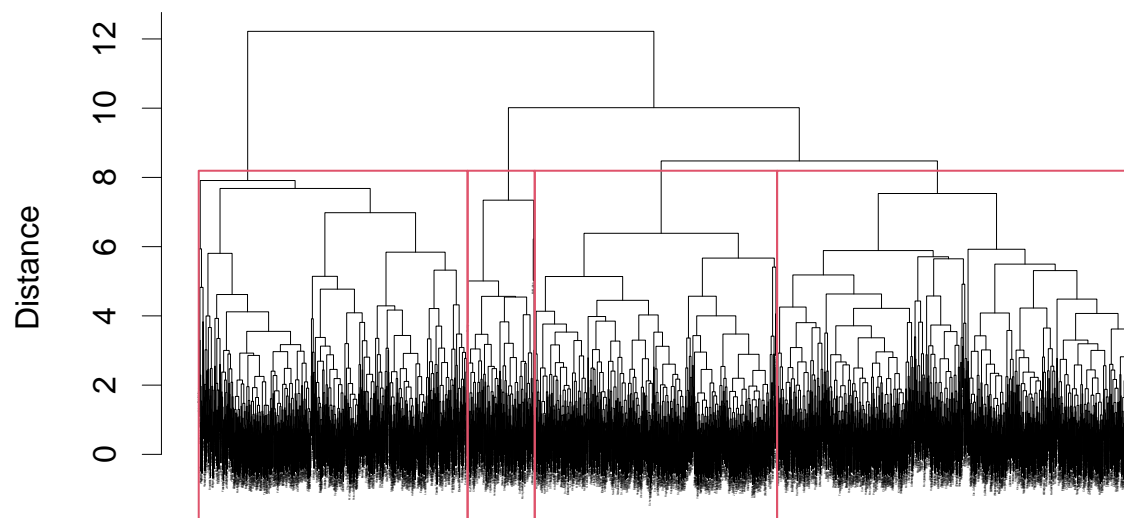
Our first task is to determine a distance and agglomeration method; we compare Euclidean versus Chebychev for distance and Ward versus complete for agglomeration. We choose a k value of 4 as has been indicated by our past pset work but will further explore cluster sizes after determining our preferred methods.

Euclidean/Ward



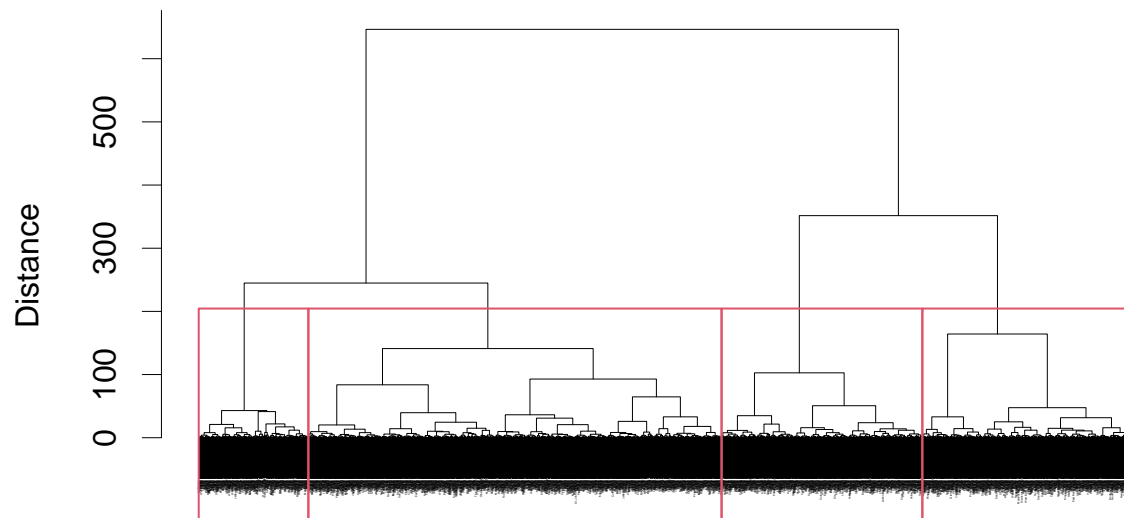
`hclust (*, "ward.D")`

Euclidean/Complete



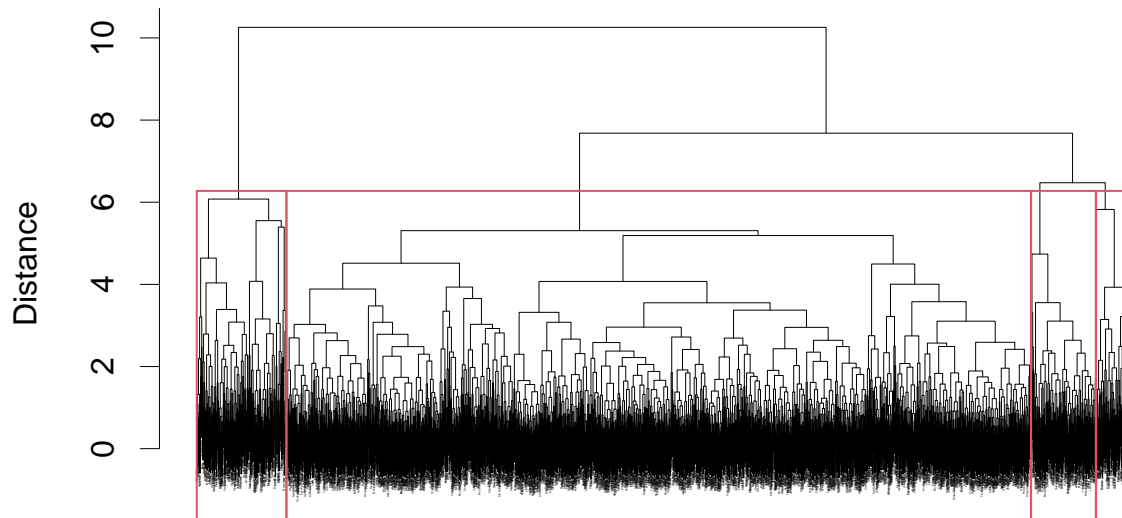
`hclust (*, "complete")`

Chebychev/Ward



`hclust (*, "ward.D")`

Chebychev/Complete



`hclust (*, "complete")`

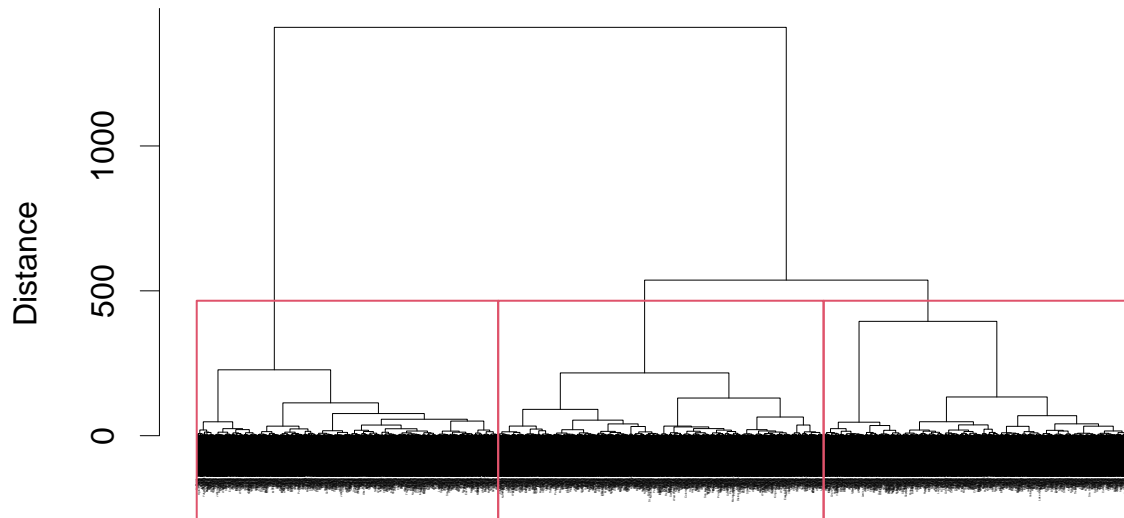
Explanation of agglomeration method: It's clear that the Ward's method gives much neater groupings than the complete linkage method. That is, the complete linkage clusters become very granular at much higher distances. It looks like Ward's might be the way to go, at least for these data.

Explanation of distance method: The distance methods are both Minowskis with very different exponents: Euclidean and a Chebychev approximation. It seems that Euclidean/Ward gives the cleanest, most sensible clusters, so we choose this combination moving forward.

Our next goal is to determine how many clusters would be appropriate with our Euclidean/Ward technique.

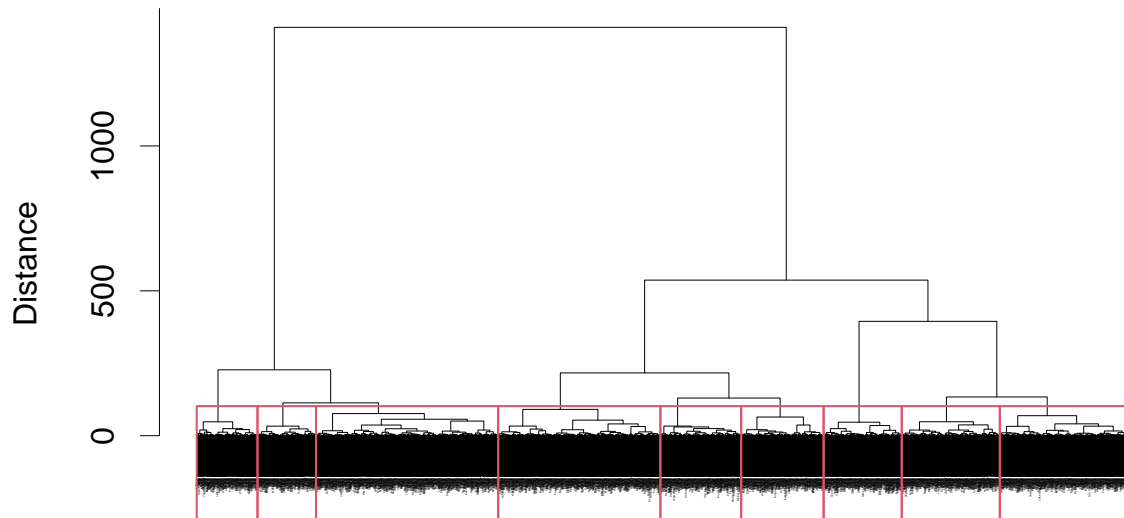
There are peaks in the RMSSTD at 3, 9, and most notably 12, indicating that these may be reasonable group counts. SPRSQ tapers at 9, supporting the idea that there may be 9 groups. However, the tapering is far more prominent at 2 and 4 and is echoed in the RSQ, so a lower group count may be indicated. Below, we examine the fits for 3, 9, and 12 clusters using dendograms and cluster plots in both principle component space.

Euclidean Distance & Ward Clustering



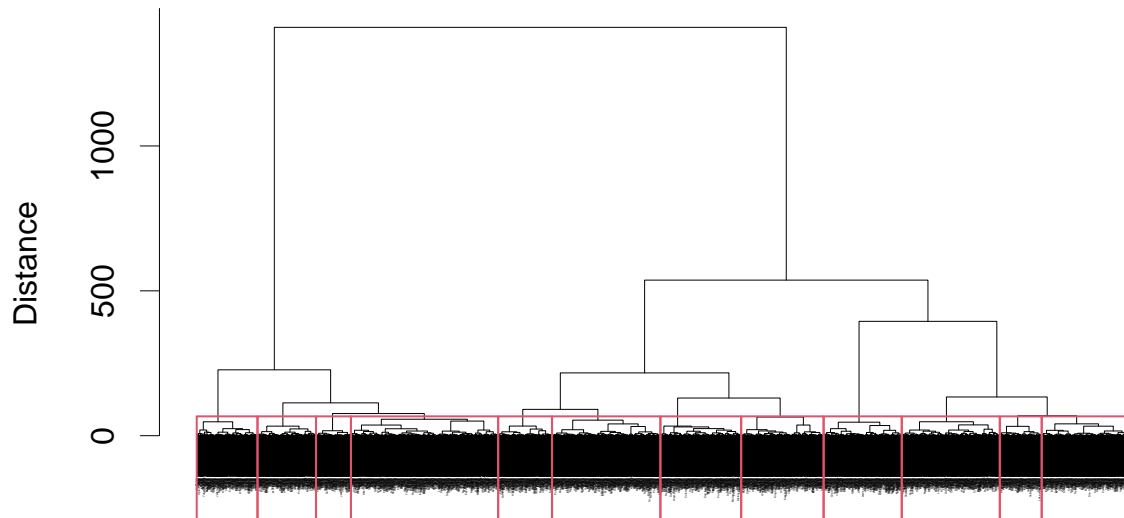
`hclust (*, "ward.D")`

Euclidean Distance & Ward Clustering



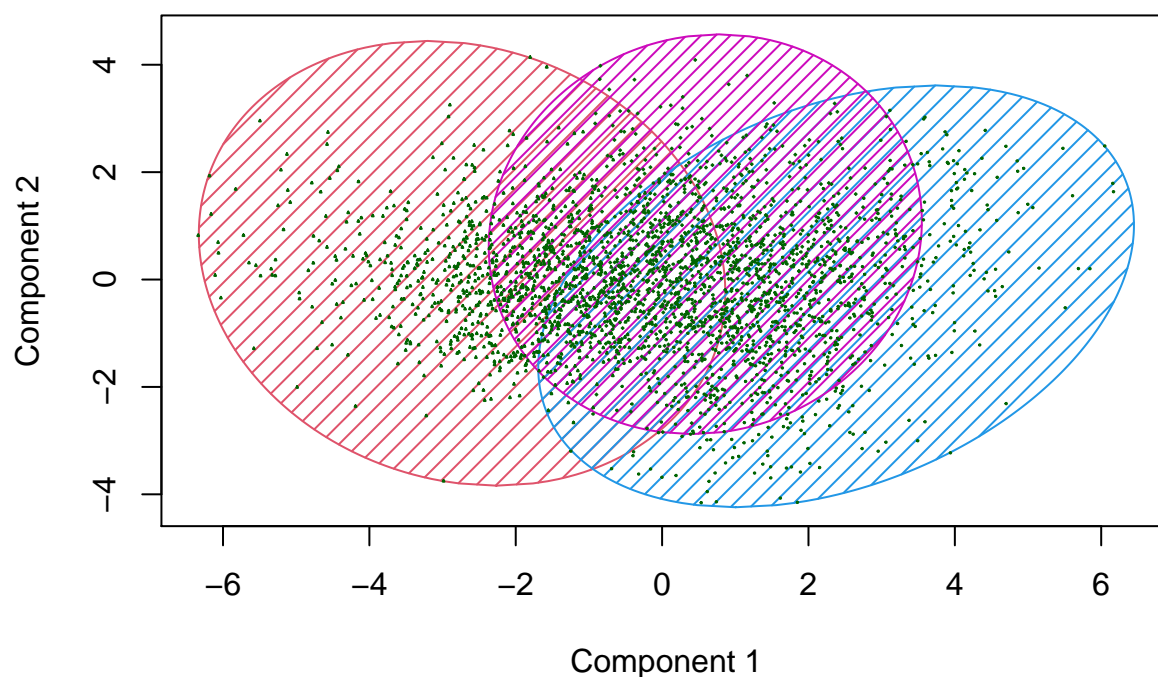
`hclust (*, "ward.D")`

Euclidean Distance & Ward Clustering



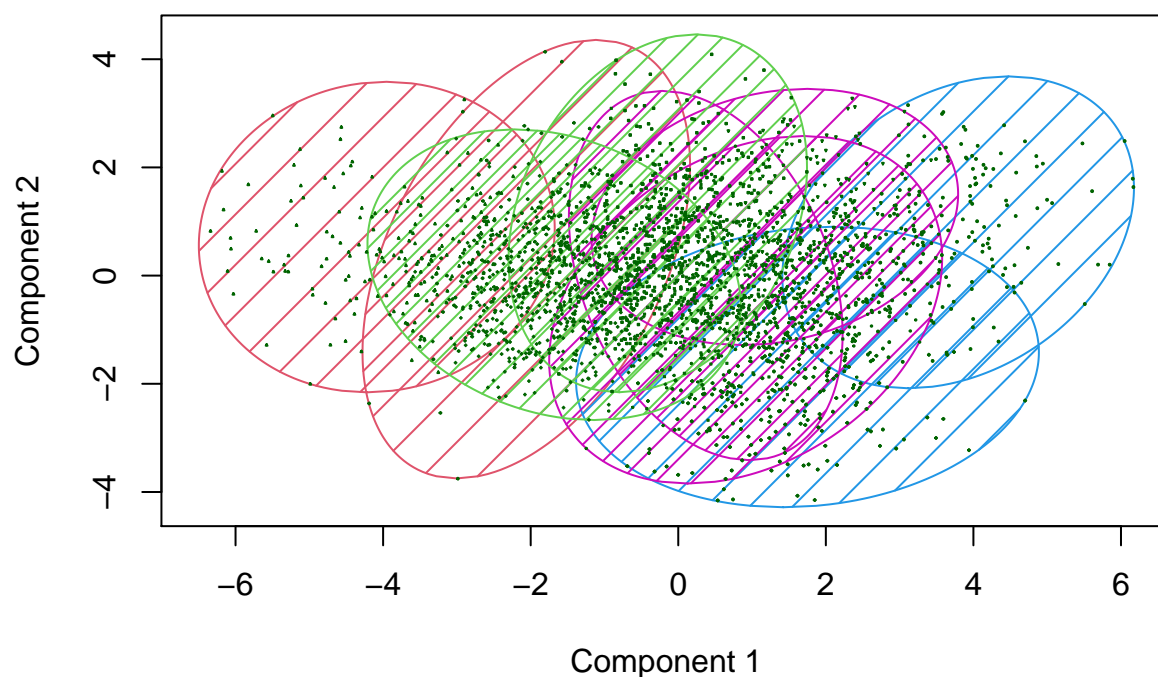
`hclust (*, "ward.D")`

County Cluster Plot, Ward's Method, First two PC, 3 groups



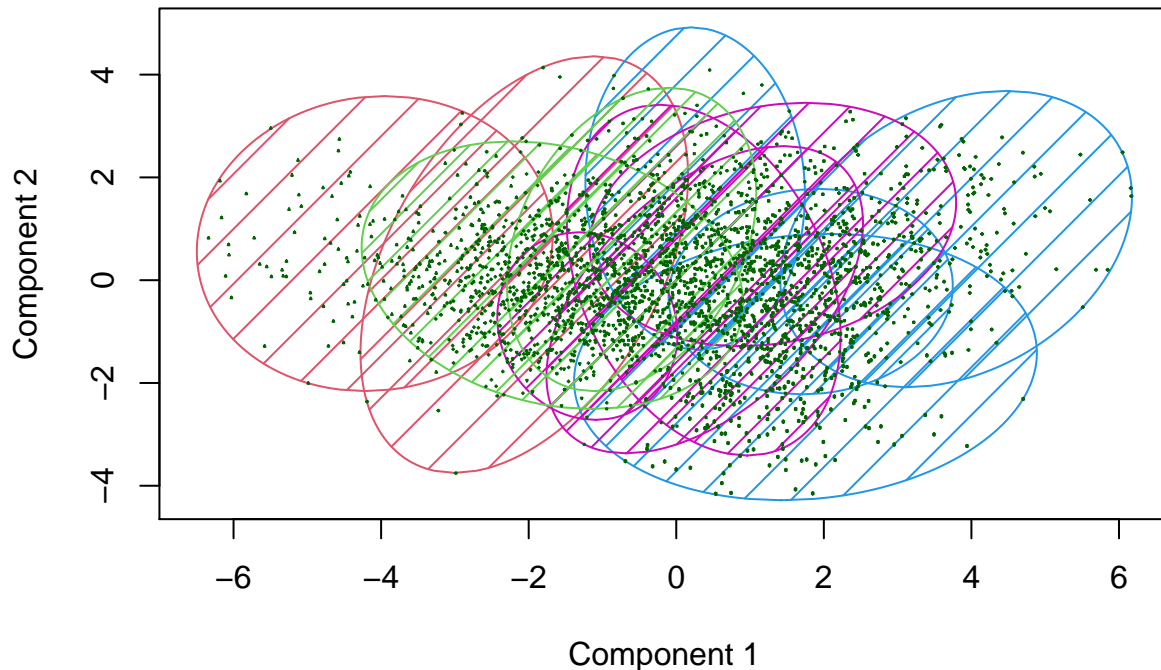
These two components explain 61.63 % of the point variability.

County Cluster Plot, Ward's Method, First two PC, 9 groups



These two components explain 61.63 % of the point variability.

County Cluster Plot, Ward's Method, First two PC, 12 groups



These two components explain 61.63 % of the point variability.

The large spike in RMSSTD is very promising for the largest cluster size of 12, and the dendrogram clustering looks apt as well, but we don't want to run the risk of having too many clusters! For the most parsimonious model, we examine 3 clusters that primarily differ on their wealth/education/employment and COVID-19 infection and death rates.

```
## Group.1 Always_Wear_Mask_Survey_Log Unemployment_Rate_2019_Log
## 1 1 -1.05432737 -0.6549327
## 2 2 0.74326479 -0.2676882
## 3 3 0.06360357 0.6206986
## Median_Household_Income_2019_Log Percent_Poverty_2019_Log
## 1 0.1357098 -0.4015587
## 2 0.8737143 -0.7273323
## 3 -0.8166837 0.8554088
## Percent_Adults_Less_Than_HS_Log Death_Rate_2019_Log
## 1 -0.4377790 0.08332214
## 2 -0.4726441 -0.21062228
## 3 0.7571856 0.38009164
## Civilian_Labor_Force_2019_as_pct_Log Covid_Confirmed_Cases_as_pct_Log
## 1 0.6435415 0.26990946
## 2 0.4384405 0.01800456
## 3 -0.7865695 0.16862940
## Covid_Deaths_as_pct_Log
## 1 0.2383466
## 2 -0.1854335
## 3 0.4264292
```

Clusters 1 and 3 are relatively well-off economically. Both have high household incomes, low poverty rates,

high civilian labor forces, low unemployment rates, and high education rates. Cluster 3 is more well-off than Cluster 1, perhaps representing those with higher paying jobs. The biggest difference, though, is in COVID-19 responses. Cluster 1 has high COVID-19 infection (0.53) and death rates (0.22), while Cluster 3 has low COVID-19 infection (-0.60) and death rates (-0.75). This difference can potentially be tied back to behavioral differences: counties in Cluster 3 always mask (0.68) while counties in Cluster 1 do not (-0.88). It is important to note that this difference is not necessarily due to any sort of moral gap but more likely due to a gap in resources - it is a privilege to be able to stay informed on scientific discoveries, purchase masks, and maintain social distancing. We only note connections but cannot specify any causal relationships.

In contrast to Clusters 1 and 3, Cluster 2 is underprivileged, with low household income (-0.82), high poverty (0.85), low civilian labor force (-0.82), and a high percent with less than a high school degree (0.76). Cluster 2 is hit the hardest by COVID-19, with the highest death rate (0.42). Here we most clearly see the connection between underlying economic factors and the impact of COVID-19. Members of these communities may have jobs as essential workers that require work outside of the home. They may have to take public transit. They may be unable to afford grocery delivery services. Affluent communities have the resources to avoid COVID-19 transmission, while impoverished communities may not. These communities are likely to have preexisting conditions that worsen the effects of COVID-19 and may lack quality healthcare or health insurance. We also note a potential gap in testing in these communities - death rates are high, but the reported infection rate is low (-0.01). As we continue to distribute COVID-19 vaccinations, special attention should be placed on supporting these communities most at risk for severe negative consequences associated with COVID-19.

Correspondence Analysis

With correspondence analysis, we are seeking to answer the question: Which U.S. states are similar to one another? Correspondence analysis is similar to PCA but applies to categorical rather than continuous variables. In this part of the report, we analyze aggregate statistics by state, i.e., taking means of continuous variables for all counties in each state. Organizing by the 50 states + D.C. will allow for more insightful visualizations.

Of our original 9 continuous variables, we assign 5 continuous variables for correspondence analysis: `Always_Wear_Mask_Survey`, `Median_Household_Income_2019`, `Percent_Poverty_2019`, `Percent_Adults_Less_Than_HS`, and `Covid_Confirmed_Cases_as_pct`.

For additional continuous variables, we make an environmental dataset. We look at four additional continuous variables describing each state: `Unemployment_Rate_2019`, `Death_Rate_2019`, `Civilian_Labor_Force_2019_as_pct`, and `Covid_Deaths_as_pct`.

Because correspondence analysis requires continuous variables to take on positive values, we applied a +2.5 pseudoshift to all values.

```
##
## Call:
## cca(X = stlm_CA_cont)
##
## Partitioning of scaled Chi-square:
##              Inertia Proportion
## Total          0.06632          1
## Unconstrained 0.06632          1
##
## Eigenvalues, and their contribution to the scaled Chi-square
##
## Importance of components:
##              CA1      CA2      CA3      CA4
## Eigenvalue    0.04169 0.01875 0.004208 0.001663
```

```

## Proportion Explained  0.62870 0.28277 0.063452 0.025071
## Cumulative Proportion 0.62870 0.91148 0.974929 1.000000
##
## Scaling 2 for species and site scores
## * Species are scaled proportional to eigenvalues
## * Sites are unscaled: weighted dispersion equal on all dimensions
##
##
## Species scores
##
##
##              CA1      CA2      CA3      CA4
## Always_Wear_Mask_Survey_Log      0.2281 -0.14957 -0.061806 -0.01840
## Median_Household_Income_2019_Log 0.2319  0.14978  0.062355  0.01428
## Percent_Poverty_2019_Log      -0.1937 -0.08139  0.079030 -0.05161
## Percent_Adults_Less_Than_HS_Log -0.1647 -0.09919  0.005033  0.07377
## Covid_Confirmed_Cases_as_pct_Log -0.1840  0.17579 -0.084069 -0.01569
##
##
## Site scores (weighted averages of species scores)
##
##              CA1      CA2      CA3      CA4
## Alabama      -1.27562 -0.77565  0.03062  0.511130
## Alaska        0.86500  0.05108  3.46625  0.705728
## Arizona      -0.54544 -0.78256 -0.89872 -0.693706
## Arkansas     -1.39194 -0.75744 -0.03250 -0.310587
## California    0.54172 -0.41110 -0.03193  1.529909
## Colorado      0.41547  0.49889 -0.21067 -2.038099
## Connecticut   1.44817  0.58045 -1.30918 -0.103394
## Delaware      0.67972 -0.16715 -1.53604 -0.445542
## District of Columbia 1.13499  0.28524  1.40745 -1.063718
## Florida      -0.35144 -0.43331 -0.54421  0.117424
## Georgia      -0.95583 -0.74232  0.02956  0.806844
## Hawaii        2.63864 -1.74329  3.12821  0.681807
## Idaho        -0.60216  0.80680  0.65669  0.739074
## Illinois     -0.11360  0.59270 -0.72492 -0.306071
## Indiana      -0.23631  0.74481 -0.53042  1.709596
## Iowa         -0.19778  1.68419 -0.24081 -0.652887
## Kansas       -0.55840  1.10964 -0.13704 -0.372341
## Kentucky     -1.15224 -0.89485  0.20784  0.743114
## Louisiana    -1.12621 -0.96672  0.35879 -0.115544
## Maine        0.77281 -1.16357  0.96986 -1.871170
## Maryland     1.28026  0.20364 -0.68981  0.946130
## Massachusetts 2.22026 -0.44335  0.75954  0.516414
## Michigan     0.06288 -0.24799 -0.75650 -1.907541
## Minnesota    0.10284  2.11032 -0.03714  0.427223
## Mississippi -1.48692 -1.36482 -0.14716 -0.610254
## Missouri     -1.25384  0.11380  0.93144  0.765571
## Montana     -1.12223  1.54644  1.80999 -3.137125
## Nebraska     -0.41762  1.63375  0.50684 -0.745166
## Nevada       0.45387 -0.22567  0.12975  1.000941
## New Hampshire 1.46716  0.48076 -0.62018 -0.061483
## New Jersey   1.32601  0.90977 -1.47912  1.149897
## New Mexico   -0.59647 -1.66974 -0.36360 -1.372719
## New York     0.64763 -0.52442 -0.81633 -0.824101

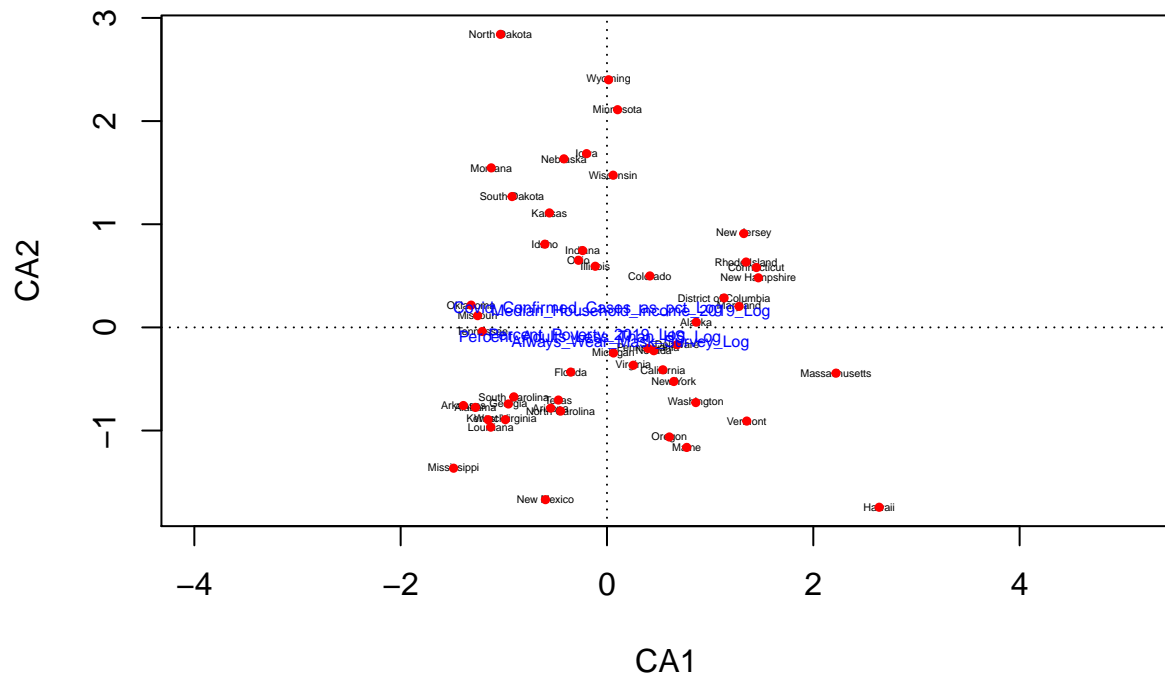
```

## North Carolina	-0.44976	-0.81121	-0.57800	0.239407
## North Dakota	-1.03218	2.84068	1.62301	1.597192
## Ohio	-0.27611	0.65059	0.18895	0.861486
## Oklahoma	-1.31688	0.21569	0.57945	0.212843
## Oregon	0.60453	-1.06235	0.78740	-0.115043
## Pennsylvania	0.39752	-0.20435	-1.28474	-0.829248
## Rhode Island	1.34831	0.63348	-1.80939	-0.013762
## South Carolina	-0.90282	-0.67346	-0.34530	-0.023101
## South Dakota	-0.92032	1.26830	0.22757	-1.093144
## Tennessee	-1.20916	-0.03982	-0.12689	1.410154
## Texas	-0.47112	-0.70415	-0.47830	1.766346
## Vermont	1.35482	-0.90803	0.35815	-0.706523
## Virginia	0.25485	-0.36695	-0.65106	1.044728
## Washington	0.86112	-0.72843	0.54599	-0.347208
## West Virginia	-0.98507	-0.89407	0.36090	-0.002092
## Wisconsin	0.05964	1.47528	-0.64736	-0.508003
## Wyoming	0.01554	2.40017	1.05651	-0.451485

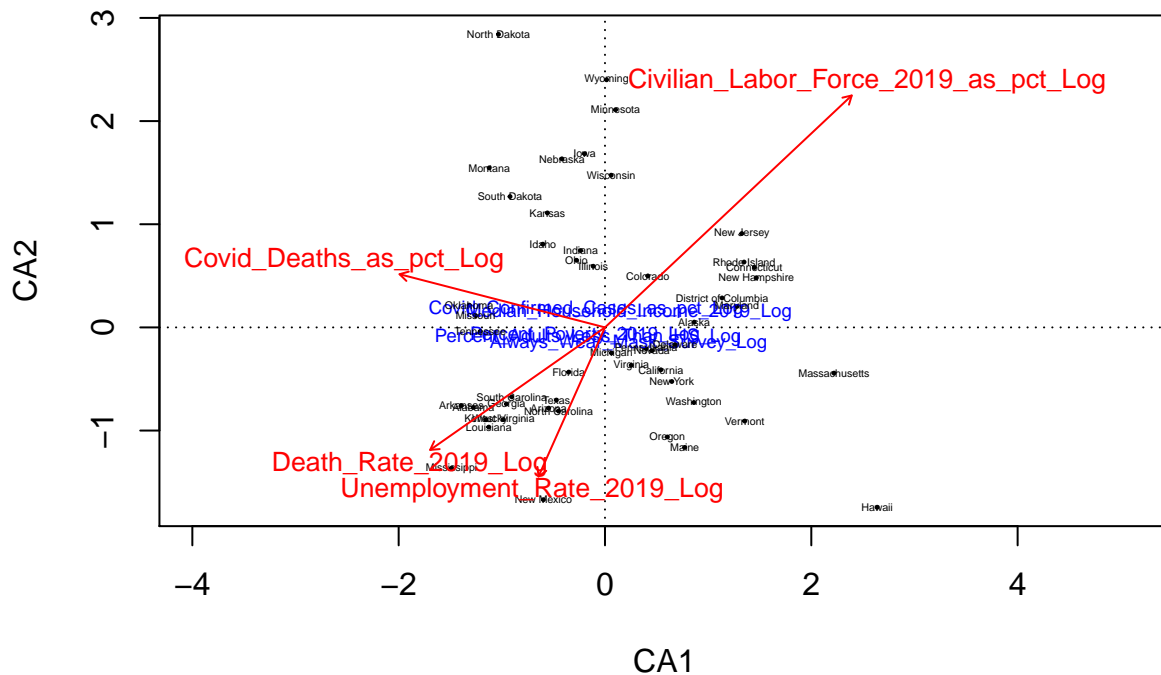
Equal to squared eigenvalues, inertia is like variance and measures departures from the independence model. We see that the inertia value is 0.06632. The magnitude of inertia does not reflect more or less variance; it is reflective of the magnitude of the data.

Deciding how many directions to keep: From the output data above in the “proportion explained” row, we can see that first direction explains ~62.9% of the relation. The “cumulative proportion” by the second direction is ~91.1%. Hence, the first two directions explain the vast majority of total inertia. The third and fourth directions have significantly smaller “proportion explained” values. This suggests that there are likely two major underlying discriminatory dimensions captured by the data of the 50 U.S. states (which reflect aggregate county data). To get a sense of these correspondence analysis directions, we subsequently plotted them.

Correspondence Analysis for U.S. States



Correspondence Analysis for U.S. States



```
##
## ***VECTORS
##
##
```

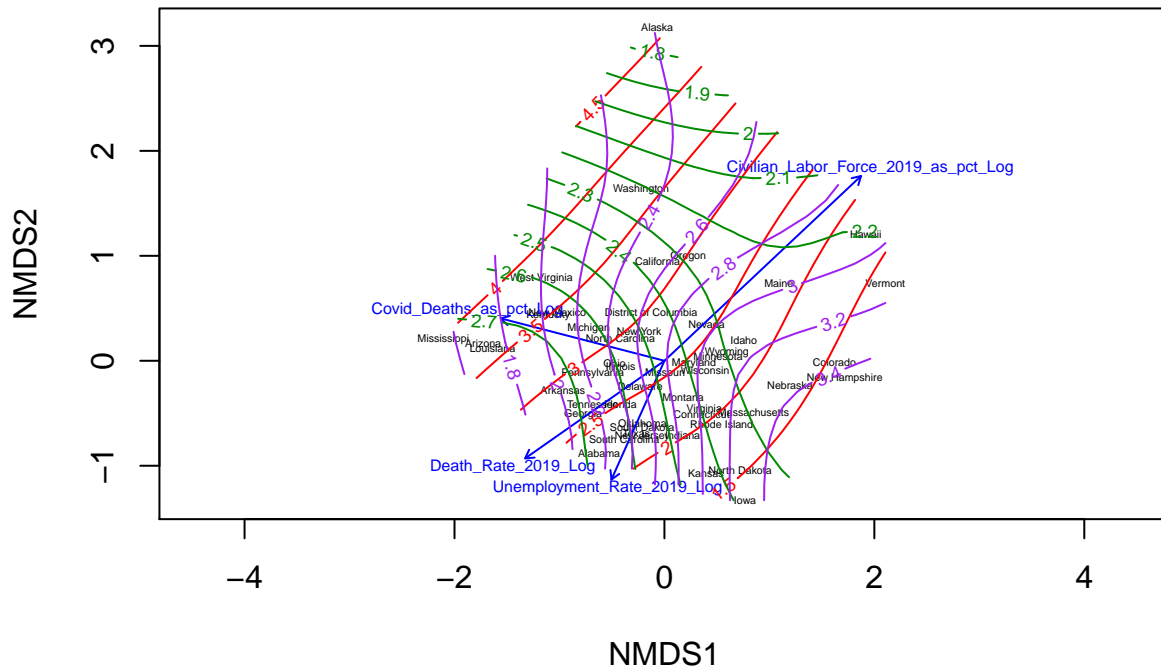
	CA1	CA2	r2	Pr(>r)	
Unemployment_Rate_2019_Log	-0.40737	-0.91326	0.1898	0.008991	**
Death_Rate_2019_Log	-0.81894	-0.57388	0.3247	0.000999	***
Civilian_Labor_Force_2019_as_pct_Log	0.72866	0.68487	0.8169	0.000999	***
Covid_Deaths_as_pct_Log	-0.96807	0.25067	0.3209	0.000999	***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 1000
```

We subsequently calculated p-values for the overlaid environmental variables, and found that the four overlaid environmental variables are all significant with $p < 0.05$.

Moreover, we applied non-metric multidimensional scaling (NMDS) to get a sense better sense of the distribution of U.S. states with respect to both the base and environmental continuous variables.

NMDS for COVID-19 Data



From our first correspondence analysis plot including the first two correspondence analysis directions, we are able to deduce the similarities and differences between states with respect to the applied continuous variables. Overall, the counties seem evenly and scattered between the four quadrants. Generally, the higher values for the first correspondence axis is associated with higher civilian labor force participation, fewer COVID-19 cases/deaths, lower death rate, and lower unemployment. The second correspondence axis is associated primarily with poor masking behaviors, greater COVID-19 cases/deaths, and greater unemployment. These results perhaps indicate two different types of counties that are associated with high COVID-19 rates (those in poorer, disadvantaged areas and also those with poor masking behaviors). Interestingly, higher civilian labor force participation is associated with slightly higher COVID-19 deaths, a trend whose causes (e.g., more workplace exposure to COVID-19) necessitates further research.

The NMDS plot reveals even more insight. We believe this plot with its contour lines optimally illustrates the distribution of the counties and their relations to the NMDS axes and environmental variables. Moreover, we can see that the contour lines are not exactly perpendicular to their respective blue dimensional axes, suggesting a more complex (non-linear) significant pattern of counties according to the environmental variables of unemployment, bachelor's percentage, and civilian labor force percentage. This graph moreover clearly illustrates how states form clusters according to these variables. There is a clear clustering of states with higher death rates, unemployment rates, and COVID-19 deaths. Other states exist along a spectrum of these variables.

Conclusions and Discussion

This report examined connections between the economic, education, behavioral, and population data for the 3141 U.S. counties and COVID-19 infection and death rate data.

Using PCA, we reduced the dimensionality of our large dataset to 3 principle components to explain 73% of

the total variability. We highlight that there are two main factors that are associated with risk of COVID-19 infection and death: 1) masking behaviors and 2) socioeconomic status. Masking and wealth are associated with lower COVID-19 infection and death rates. Using cluster analysis, we determined which counties are most similar to each other: there are very well-off counties with high mask compliance and low COVID-19 rates, moderately well-off counties with low-mask compliance and high COVID-19 rates, and impoverished counties with high COVID-19 rates. This clustering implies that differences in masking behaviors and COVID-19 infection may not be due to any sort of moral gap but more likely due to a gap in resources - it is a privilege to be able to stay informed on scientific discoveries, purchase masks, work from home, and maintain social distancing. We also note that impoverished counties have much higher COVID-19 death rates and may have preexisting conditions that worsen its effects and lack quality healthcare or health insurance. Moreover, correspondence analysis for states revealed similar trends: higher COVID-19 infection and death is associated with less proclivity to mask, higher unemployment, and overall higher death rate.

We can observe these connections, but we cannot make any cause-and-effect statements based on our current observational study. However, even without knowing the cause, we can say that vaccine and education efforts should be prioritized in underprivileged communities with lower masking rates - these communities are being hit the hardest by COVID-19.

Points for Further Analysis

We hope that studies of COVID-19 death and infection rates will continue, even as vaccination rates increase, so we can find the communities who can benefit from public health efforts both now and in the future. We also hope that these public efforts extend beyond just COVID-19 assistance; our work has highlighted the connection between socioeconomic factors and infection and death rates. While we are unable to examine the causal nature of this relationship with this dataset, hopefully future studies will probe at why this connection exists and present solutions.

We note that COVID-19 is a pandemic, impacting the entire world. Though we only studied counties in the United States, it would be worthwhile to study other countries to understand how to prioritize not only vaccination efforts in the U.S. but in the world. Vaccination is a world-wide effort, and none of us are protected until we are all protected.