

# Nursing Home COVID-19 Risk Assessment

## ***Group 10:***

**Marianna Carini**

**Eric Chen**

**Allen Lee**

**Stella Lee**

**Smitha Kannanaikkal**

## **Introduction**

The ongoing pandemic has shot the world into a paralyzed state, its impact affecting the day-to-days of millions around the globe. Particularly in nursing homes where a record number of outbreaks is fueling the coronavirus (COVID-19) cases spike. While the COVID-19 surge threatens to overwhelm hospitals and economies, the biggest breakthrough of 2020 came when Pfizer and BioNTech announced its first result of vaccine for the COVID-19. A Centers for Disease Control and Prevention advisory panel group recommended that “some of the first vaccines for COVID-19 go to nursing home residents as well as front-line health care workers” (DeMio & Behrens, 2020). The question then arises: which nursing homes are at higher risk from COVID-19? In our project, we will be performing exploratory analysis and applying machine learning techniques to examine important features for predicting the groups of people with the highest to lowest risk of coronavirus infections by combining data sets from multiple sources – Nursing Home, Census, and Policy data. We strongly feel that our analysis can provide actionable items that could help us better protect our seniors for the future.

## **Data**

The three main data sources we used for this project are Covid-19 nursing home, Covid-19 policy, and census data.


## *COVID-19 Nursing Home Data*

We obtained the nursing home data through the data.cms.gov API. The API is hosted by Socrata and has a limit of 100,000 rows without a token. With a token you are limited to 1,000 calls per hour. Data from May through mid-October was roughly 330,000 rows, so we created a token to collect as much data as possible.

The nursing home data set was recorded on a weekly basis and we started off by analyzing provider name and federal provider number. There were a total of 15,118 providers and 1,315 providers had more than one nursing home while only 111 providers had more than one nursing home in more than one state. For clustering purposes we aggregated the data to achieve single records per nursing home. We considered the latest entry for the attributes which indicated total cumulative count. Also, for the attributes related to supplies and tests having a binary response, we calculated the percentage to indicate how often the nursing home faced the situation where there was a shortage of masks, a shortage of hand sanitizer, shortage of staff, tested residents with new signs or symptoms, etc.

To examine the quality of data, we examined the features and noticed that the feature indicating if the nursing home has done point-of-care (poc) test had around 14% missing values. Through our analysis, we realized that the missing values were presented when the nursing homes had no poc machine. Furthermore, we observed that nursing homes were least likely to do the test if they do not have a machine. Therefore we filled in “no” for the missing poc test values. Another step we took for data cleaning was that we engineered a new feature which indicates the percentage deaths due to COVID-19 among the nursing home residents who had COVID-19, both detected and suspected by COVID-19. We call this the mortality rate.

## *Census Data*

 We wanted to enhance the nursing home data with each county's demographics since characteristics like population density and age play a factor in how widespread COVID-19 is. Like the nursing home data, the Census data is also available through an API. This API had the largest amount of reports and was also the most difficult to

connect to. Without a key, one is limited to 50 calls per day. We also created a key for the Census API, so we did not have to be concerned about the limit. The API allows the user to create their own query from decades of Census reports. This process was time consuming due to the fact that the variables are labeled with codes, so one has to read through documentation to decipher which codes to include in the query. There are a number of different reports with different types of information as well. We chose the 2019 American Community Survey since it was the latest full-year data available with demographics information.

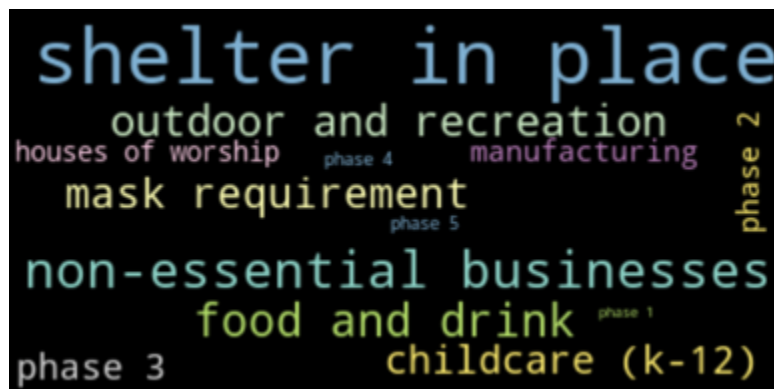
While creating the query was time consuming, the time was well spent. Once we were able to connect with the API and capture the data in a pandas dataframe, we found that the data was almost entirely clean. The only item that needed to be cleaned was replacing “-999999999.0” with a blank. This value is used as N/A in the reports and was easily detected in the percentage variables. Next, we had to construct a way to connect the nursing home data with the Census data. The Federal Information Processing Standard code, or FIPS code, is a government issued unique identifier for each county. We were able to create a reference table manually in Excel. We merged the county and state into one field in each file, which we called “ID”. We included the state to account for situations where there are two counties with the same name, but in different states (ex. Orange County, CA, Orange County, FL, and Orange County NJ). Then we used the Excel VLOOKUP function and found matches for a majority of the counties. The nursing home county field had a number of typos and formatting differences, so we included multiple versions of the county name to fix these variances(ex. AlbemarleVA and AlbermarleVA). We loaded this table into our Python code in order to add FIPS to the nursing home dataframe.

### *COVID-19 Policy Data*

As we did for the other two data sets, we collected our data from connecting with government health data API. We used regular expressions to turn byte data into a dataframe and cleaned up the data using regex because the content from our API connection was formatted as one long string with various special characters throughout.

However, we found out that this API had the limit of 100 rows without an account. We needed to extract 3,500+ rows. However, the Health Data API does not allow for accounts to be created. Thus, we decided to manually download the data from HealthData.gov. We kept the code so that we can connect to the API live if the issue is ever fixed.

The policy data contained 10 columns including county, FIPS code, policy level, policy type and so on and so forth. Since we wanted to compare the data county-wise, we filtered policy level as county and changed the column name for FIPS code to match with the column name from nursing home data. Our data consisted of a total of 23 policy types for counties and after examining the description of each policy type, we merged the policy types with similar behavior and the total number of policy types has reduced to 13. As shown in Figure 1, we made a word cloud to catch a glimpse of policy types' frequencies and we noticed that shelter in place was the most frequent policy type that appeared.



(Figure 1)

The next step for data cleaning for policy data was to create binary variables for each policy type to indicate whether that policy type has been implemented or not. The purpose of this step was to count the number of each policy executed by each county-level FIPS code. Before merging our policy data to cluster data, we created another dataframe containing only the variables we needed – county-level FIPS variable and all the binary variables for each policy type. Then we merged the new policy dataframe to our cluster data on county-level FIPS code. Finally, we grouped by clusters and found

the proportion of each policy type by clusters by calculating the average of the number of each policy type for each cluster. (Figure 2) The four policy types, 'house of worship', 'childcare', 'phase 1', and 'phase 2', were removed because no county in our cluster data implemented these policies.

cluster	policy_type: _food and drink	policy_type: _manufacturing	policy_type: _mask requirement	policy_type: _non- essential businesses	policy_type: _outdoor and recreation	policy_type: _phase 2	policy_type: _phase 3	policy_type: _phase 4	policy_type: _shelter in place
0	0.031782	0.005959	0.143965	0.051646	0.025823	0.023837	0.172815	0.007946	0.536606
1	0.036693	0.107786	0.055531	0.028667	0.005733	0.038987	0.084853	0.004587	0.464617
2	0.048463	0.026498	0.146102	0.033122	0.014295	0.013597	0.166805	0.003487	0.491417

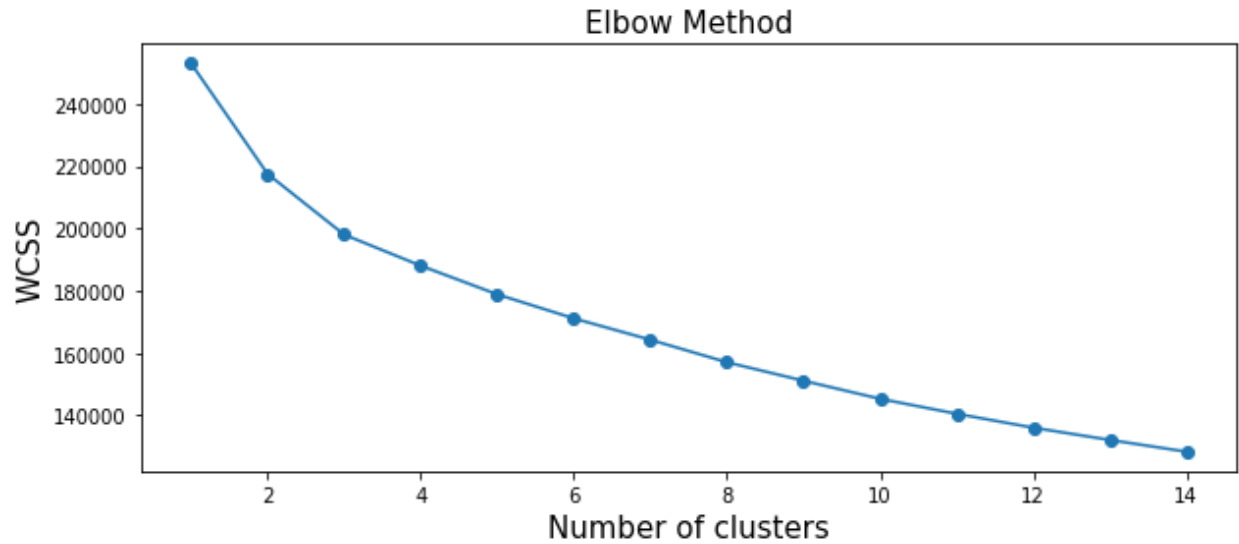
(Figure 2)

## K-means Clustering Model

We used K-means clustering, an unsupervised machine learning technique, to divide the nursing home into like groups. Since the clustering algorithm uses distance based measurement to determine the similarity between the data instances, we standardized the dataset. This helps get all the data on the same scale, which in turn improves the accuracy of clustering. In order to determine the optimum number of clusters we ran fourteen iterations of K means with K ranging from 1 to 14 and plotted the elbow plot as shown below. We used the K-means algorithm from sklearn package.

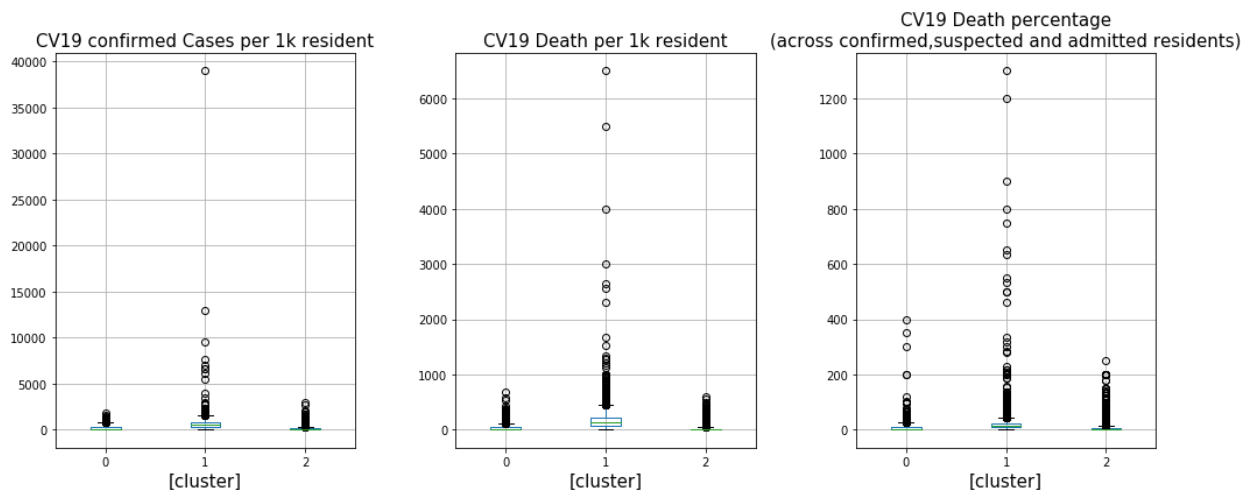
Following are the parameters used while running the K-means algorithm-

1. max\_iter: sets the number of maximum iterations for each initialization of the K-means algorithm
2. n\_init: Number of times the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n\_init consecutive runs in terms of inertia.

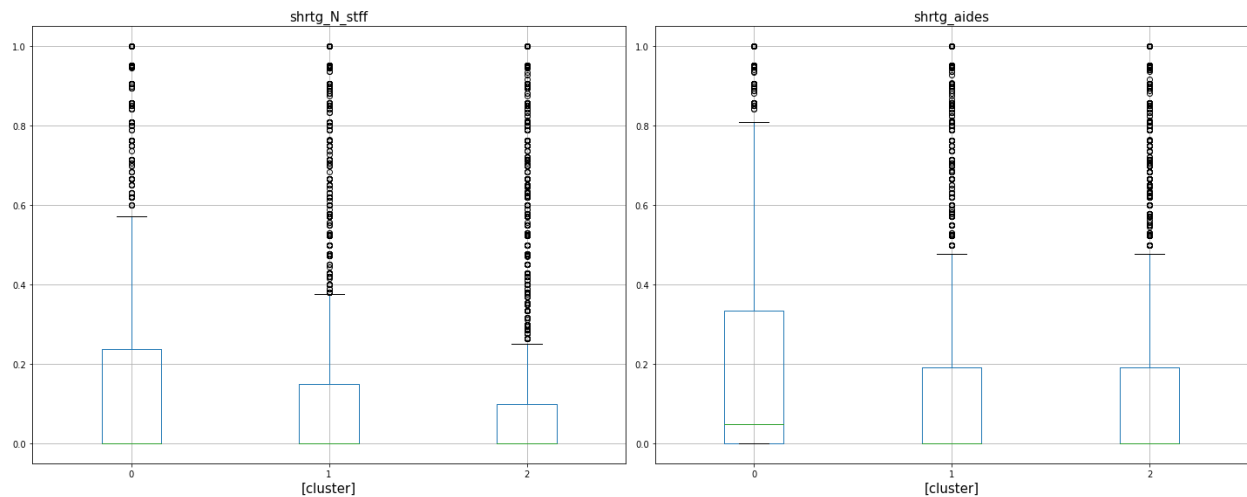


The elbow plot above shows that  $k=3$  is a good choice for the number of clusters to be considered for the final model. After which we ran the final K-means model using  $k=3$  and used the resultant clusters for our further analysis.

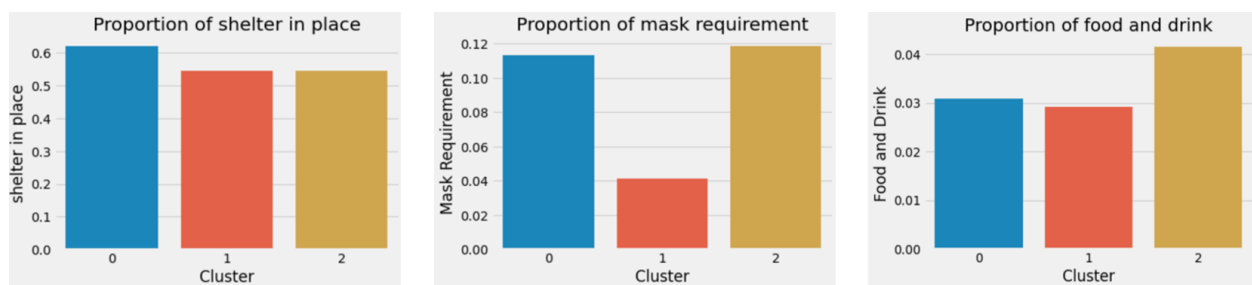
The final cluster analysis are as follows



From the above box plot we can see that cluster 1 has the nursing homes where the highest COVID-19 confirmed case and mortality due to COVID-19 was observed, which is then followed by clusters 0 and cluster 2. We classify these clusters based on the impact due to COVID-19. Thus cluster 1 is classified as 'high impact', cluster 0 as 'medium impact' and cluster 2 as 'low impact'.



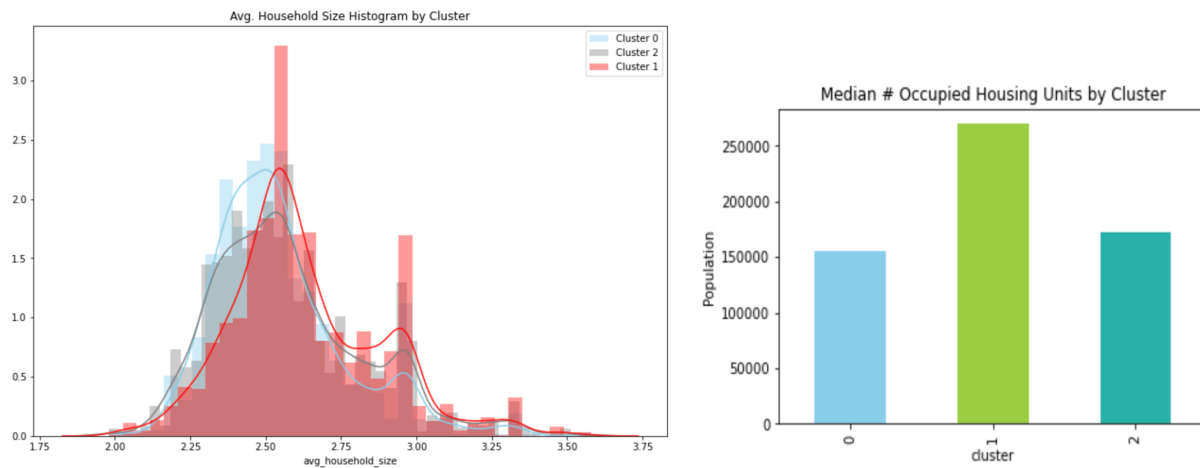
We were surprised to see that cluster zero did not have the highest mortality rate despite having the highest shortages of supplies and staff. We looked deeper into the policy data to see if the types of restrictions varied between clusters. We found that the policies did in fact vary between the clusters. The graphs below show the percentage of counties in each cluster that implemented a policy type due to COVID-19. We can see that 54% of cluster one was under shelter in place restrictions versus 62% in cluster zero. We also observed that only 4% of cluster one was in a county that had mask requirements versus 11% in cluster zero. For food and drink restrictions, 0.29% of cluster one had limitations compared to 0.31% of cluster zero.



(Figure 3)

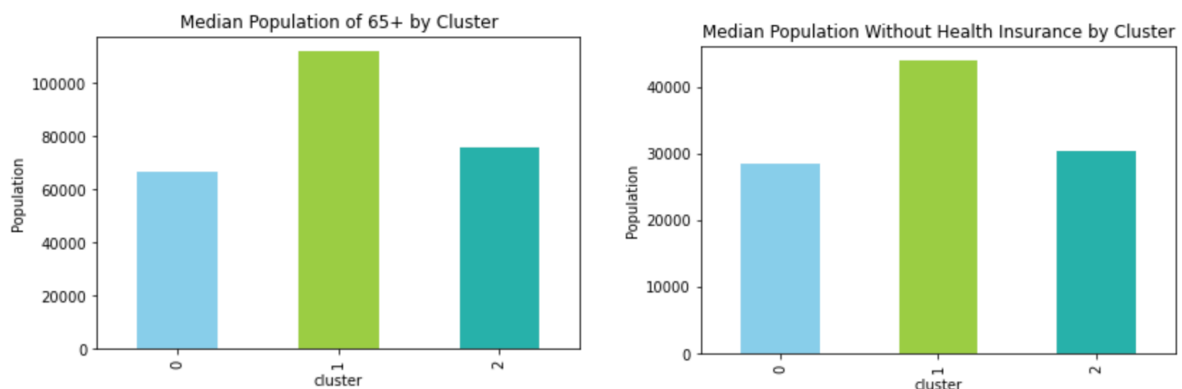
Similarly, our group was surprised to see that cluster one had a high mortality rate despite having an adequate amount of staff and supplies when compared to clusters zero and two. We studied the demographics and found differences between the

clusters in a few key areas. Cluster one consists of higher density regions. We can see from the left chart in figure 4 that cluster one, shown in red, has a higher number of people living in each household on average when compared with clusters zero and two. We can also see in the right chart of figure 4 that cluster one also has a higher number of occupied housing units than the other two clusters.



(Figure 4)

We discovered that cluster one also had the highest population of people ages 65 and over as well as the population of people without health insurance from studying the demographic makeup of each cluster. These three key differences likely played a role in why cluster one saw a high mortality rate.



(Figure 6)



## Conclusion/ Recommendations

The newly-developed COVID-19 vaccine is a source of hope for many. Because producing vaccines takes time, the pharmaceutical companies cannot produce vaccines for everyone initially. The first vaccines will go to healthcare workers and nursing homes. Even with that small subset, the manufacturers will not be able to send doses to each nursing home immediately.

Based on our findings, we recommend the COVID-19 vaccinations for nursing homes be prioritized in three waves. The first addressing facilities in cluster one. This would cover 295,000 nursing home residents. We recommend treating cluster one first because of their high mortality rate, lack of policies, high density, and higher at-risk population. Next, we recommend sending vaccines to cluster zero which contains 152,000 residents. Their high amount of shortages is a point of concern. The providers may not be able to adequately contain outbreaks due to the lack of supplies and staff. Lastly, we recommend sending vaccines to the 652,000 residents of cluster two. They are fairing the best with the lowest mortality rates, adequate supplies, and various policies in place.

While having a viable vaccine is a big step forward, the logistics of dispersing the vaccine are equally as daunting. In a report on the H1N1 vaccine deployment from the World Health Organization, the task of shipping the vaccines proved to be a massive undertaking with a number of issues arising (*World Health Organization*). Much like the COVID-19 vaccines, the H1N1 vaccines needed to be kept at low temperatures. This meant the vaccines needed to be shipped in cold-chain packaging. This packaging is large and bulky and requires specific handling procedures. This limits the number of transportation providers equipped with the space and training to transport these vaccines. Additionally, the demand for H1N1 was much higher than that of the seasonal flu. This put further strain on the transportation providers. It is reasonable to assume that the COVID-19 vaccines will face the same challenges due to the high demand and the use of cold-chain packaging. There are a vast number of moving pieces in the disbursement of vaccines. We recommend having an entity like the World Health

Organization create a team to oversee the logistics. This team could help develop and communicate a plan for nursing home facilities and transportation providers to receive vaccines. The team could also help prevent logistics problems from arising during the transportation.

From our analysis, we found that shelter-in-place policies and mask requirements were likely attributed to the lower COVID-19 cases and mortality rates in clusters zero and two. We advise that counties in cluster one consider implementing shelter-in-place restrictions and mask requirements, if they have not done so already, in order to control mortality rate.

## Works Cited

DeMio and Behrens, T. (2020, December 02). COVID-19: Rising resident, worker cases put new strain on nursing homes. Retrieved December 07, 2020, from <https://www.cincinnati.com/story/news/2020/12/02/covid-19-rising-resident-worker-cases-put-strain-nursing-homes-ohio/6347681002/>

“Report of the WHO Pandemic Influenza A(H1N1) Vaccine Deployment Initiative.” *WHO.int*, World Health Organization. [https://www.who.int/influenza\\_vaccines\\_plan/resources/h1n1\\_deployment\\_report.pdf](https://www.who.int/influenza_vaccines_plan/resources/h1n1_deployment_report.pdf). Accessed 2 Dec. 2020.