# 1.Overview

In the context of project 1, we put forward the hypothesis that people's willingness to get vaccinated is related to ICU capacity, death rate, sentiment on social media, local policies, and the trend of the pandemic. In this project, we did some basic statistical analysis on the data of attributes listed above to find out the patterns, and also implemented four clustering methods trying to build relationships between those datasets. Furthermore, we did topic modeling on our text dataset trying to extract the factors that may have influence on the result.

# 2. Data Cleaning

## 2.1 Covid Act Now Api

We decided to preprocess the raw dataset because the original dataset (written in the csv file from api) is in the sequence of all the states of America. So, we separate the whole dataset into 53 subsets. While separating the raw data, we replace the outliers with the median of its two neighbors. The reason we set this principle is that all of the data actually is describing a trend the attributes of the pandemic, so we would believe that if a data equals to 0 while both of its neighbors doesn't equal to 0, then it's an outlier, otherwise, it's just a result describing the fact. After preprocess, the code will generate "modified.csv" which is the cleaned data of the raw dataset.

## 2.2 Covid Tracking Project API

For Covid Tracking Project API, there are three datasets: States_Daily, US_Daily and States_Info. States_Info only contains notes of all stats in the US, is a text dataset with no null values so we do no clean operation on it. For States_Daily and US_Daily, in Project 1 we found there are many null values in them. So, at first, we delete the attributes deprecated by the API website and attributes which have more than 0.4 percent of null values. Then computed the values for the remaining null values equals the mean of its four adjacent values. Finally, we have a dataset with no null values and saved them as States_Daily_clean and US_Daily_clean.

## 2.3 twitter-COVID-dataset

We found no missing values in the twitter-COVID-dataset, so what we have done is remove the data whose sentiment values are out of range. As mentioned in project 1, the intensity score may be exceeding the [0,1] range, which indicates that these tweets the researchers collected represented extreme intensities beyond the algorithms they used. Therefore, we delete 2757 data to avoid their impact on subsequent analysis. After cleaning, 54723594 data remain and we can be sure all data are in the right range.

Since the size of the original data is more than 4GB, considering the running time, it is difficult for us to use the entire data set at one time to do complex and time-consuming analysis problems. Fortunately, this data set is collected from January 2020 to August 2021, so we divided the data set by month and obtained 20 separate files. Due to computer memory, our code cannot store all cleaned sentiment data into one file. Therefore, we decided that when observing the sentiment during the entire pandemic, we use the original complete data (including a small part of emotional values that are out of range); when observing the sentiment per month, we use reshaped sentiment data.

### 2.4 MongoDB data set

For bad values or some data containing missing values, we just remove these values. Because we count these bad values and find that they only account for a small part in this dataset (for about 0.05%). What' we remove some useless attributes in our data, MongoDB data set provides us with 15 attributes, however, some of them are useless for our analysis, so we remain 6 attributes--combined name (county name and state name), location(longitude and latitude), population, date, confirmed, deaths, confirmed daily, deaths daily. And we transfer combined name values into numeric values (for each value, transfer it into a continued integer).

## 3. Basic Statistical Analysis and data cleaning insight

### 3.1 Covid Act Now Api

Firstly, we calculate the mean, median, and standard deviation of almost all of our attributes while doing the data cleaning and separating the raw data.We write the result in the last 3 lines of each file under the path ". /clean" (path ". /clean" will be generated after running the covid_act_now_p2.python). We're listing 10 of those results below.

| | testPositivity Ratio | Case Density | ICU Capacity Ratio | Vaccinations CompletedRatio | hospital beds capacity |
|---|---|---|---|---|---|
| **mean** | 0.040 | 2.110 | 0.449 | 0.151 | 999.014 |
| **median** | 0.028 | 2 | 0.64 | 0 | 1479.500 |
| **std** | 0.032 | 1.396168 | 0.347 | 0.196 | 764.444 |

| | icuBeds currentUsageCovid | New Cases | New Deaths | Vaccinations Initiated | cdcTransmission Level |
|---|---|---|---|---|---|
| **mean** | 11.33727 | 259.4636 | 1.477789 | 135515.2 | 2.10118 |
| **median** | 7 | 126.5 | 0.75 | 0 | 3 |
| **std** | 13.10361 | 373.7367 | 2.618401 | 164849.3 | 1.107009 |

The reasons that these attributes are listed are: (1) they represent almost all the types of data in the dataset, (2) they are the most important attributes to describe the topic of our project. The median of vaccinations completed ratio, vaccinations Initiated equals to zero is because in the early stage of the pandemic, the vaccines were not available and the data was collected after a period of time.

Secondly, while running the code, it will output the results of LOF of all of the attributes in the dataset. You might observe that most of the results equals to 1, it is because most of the data collected from api has been preprocessed to rule out the outliers. After experiments, we set the neighbors of the LOF equals to 3 which finds the outliers the best. Since we already dealt with the outliers before in the process of data cleaning, we decided not to remove the outliers.

### 3.2 twitter-COVID-dataset

As shown in the following table, we can find that the mean and median values of fear, anger, and sadness are all concentrated around 0.45, while the value of happiness is concentrated around 0.2. As mentioned in project 1, these values indicate the intensity of a certain sentiment in a tweet. We can conclude that during the entire pandemic, most of the tweets related to COVID-19 did not contain or contained a very small amount of happiness, more of fear, anger, and sadness. The value of valence indicates that most tweets are relatively negative.

| | valence | fear | anger | happiness | sadness |
|---|---|---|---|---|---|
| **mean** | 0.459 | 0.442 | 0.441 | 0.299 | 4.144 |
| **median** | 0.459 | 0.440 | 0.437 | 0.295 | 0.408 |
| **std** | 0.095 | 0.095 | 0.088 | 0.094 | 0.080 |

**3.3 MongoDB data set**

First of all, loc_ids represent different areas, and it turns out our dataset collects information from 1996 regions. Coordinate-x and y give more details about these regions' location. All of the means, medians, and deviations of those variables mentioned above plus population variables enable us to have a basic overview about the regions' nature. For other attributes, we can see that for most of the reports, death, confirmed_daily, deaths_daily for people was low, but peaked in some time. It is reasonable that it peaked when the epidemic broke out. Confirmed numbers are similar, most reports reflect there are hundreds of confirmed people, and the highest numbers are up to 10 thousand.

| | loc_id | coordinate-x | coordinate-y | population |
|---|---|---|---|---|
| mean | 1034 | -93.34 | 39.01 | 125292 |
| median | 1036 | -90.20 | 39.23 | 32399 |
| std | 574 | 14.94 | 5.38 | 379414 |

| | confirmed | deaths | confirmed_daily | deaths_daily |
|---|---|---|---|---|
| mean | 5847 | 122 | 30 | 0.52 |
| median | 975 | 18 | 0 | 0.00 |
| std | 26110 | 548 | 172 | 4.18 |

We choose 5, 50, 100 as k values to utilize LOF to detect outliers on this dataset. It turns out to be a negative outlier factor matrix, which reflects the score for each point. Most of the point's score is around -1, so we think there are few outliers inside it. For this outlier, it might be caused by regions or date or something, we can't figure it out if they are useless information.  Thus, we don't remove this outlier.

### 3.4 Covid Tracking Project API

For Covid Tracking project dataset, we compute the mean, median, and standard deviation of all attributes in States_Daily and US_Daily, because this dataset has similar attributes with MongoDB data set，we don't put the results here, but it has result in the output of the code. We choose 2, 20 and 35 as k values to utilize LOF on these two files. When the k value is 35, it shows there is little '-1' in metrics. Since we have processed this data before, and we believe that data such as "death" or "positive" rate may indeed have a sudden increase and decrease, we feel that a small number of outliers are normal.

## 4.bin

We use the attribute "death" in this dataset to do the bin operation. Mean,median, and standard deviation of "death"are 174729, 154802 and 515151.So we bin the death data by 0-10000, 10000-150000, 150000-60000 as low, regular and high. We output the date with a new bin line to a new file called US_Daily_bin. The method we use is pandas.cut, we use this method because it is easy to use, and the data we need to bin is exactly One-dimensional. By the result of bin, we count the number and percent of these three values we have. We are sad to find that the number of deaths greater than 150,000 in the United States is the largest，accounting for 67%.  The number of deaths less than 10000 per day is the least, only 8%. In a way, this proves that the epidemic is really serious.

We also made a binning strategy for the sentiment data to classify their emotional intensity more clearly. For the valence intensity, if the value is close to 0, it shows a very negative sentiment, and if the value is close to 1, it shows a very positive sentiment. Therefore, we bin the valence attributes for 5 parts: the range [0,0.3] means very negative, [0.3,0.48] means negative, [0.48, 0.52] means neutral or mixed, [0.52, 0.7] means positive, and [0.7,1] means very positive. Actually, the classification obtained by this division is nearly the same as the distribution of the 'sentiment' attribute in the database.

In addition, we divided each of the other four sentiments (fear, anger, happiness, and sadness) into 4 parts: [0, 0.3], [0.3, 0.5], [0.5, 0.7] and [0.7,1], which indicates the degree of a certain kind of emotion contained in the tweet. For example, when a happiness value falls

within the range [0, 0.3], it shows that this tweet nearly did not express any happiness at all. When a fear value falls within the range [0.7,1], it shows that this tweet expresses a very obvious fear. This binning strategy is very helpful to get the overall emotional attitude of tweets during the entire pandemic period or a certain month, which is an important factor for the willingness to vaccinate. These boundaries are determined based on the data distribution shown in the histogram in the third part of this report.

## 5.Histograms and Correlations

In this part, we select MongoDB dataset as our analyzed object, and we plot date, loc_id, confirmed, deaths four variables' histograms. Here are the results:

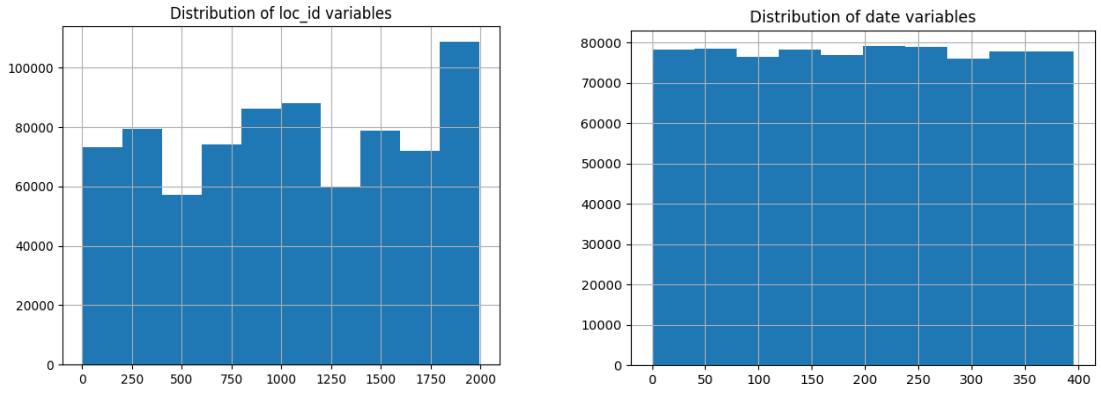


Figure 1: Distribution of loc_id variables and date variables

For the first two histograms, they show us the distribution of the data set in regions and dates. The number of reports from different regions have differences, and it might influence the results of the cluster. For the second histogram, the number of statistical reports on different dates is relatively similar in quantity, which avoids differences due to the number of reports on different dates. For distribution of confirmed variables and deaths, the results are the same as basic analysis for those two variables in the basic analysis part. In most reports, the number of confirmed people and deaths are relatively low.

Correlation

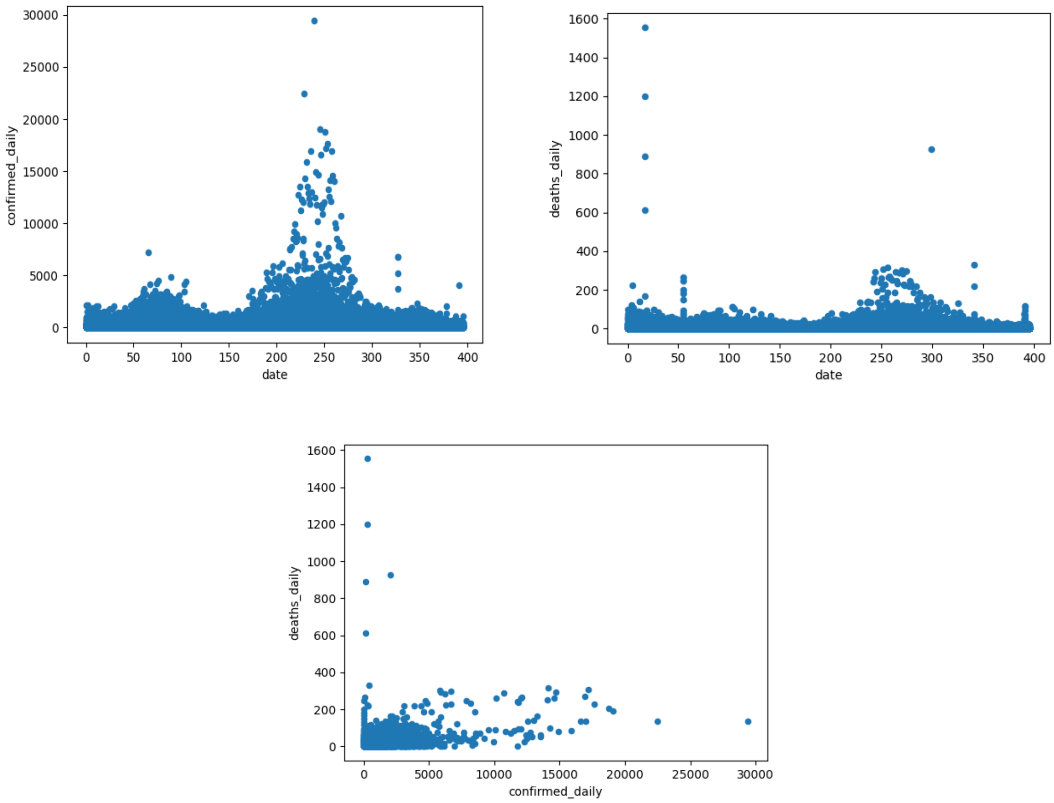


Figure 2: Scatter diagrams for confirmed_daily, deaths_daily and date variables

Those three tables are the correlations between all the pairs of dates, deaths_daily, confirmed_daily variables, The first two tables show the epidemic situation changes over time, and have two peaks, the dates of deaths_daily peak is a little behind the dates of confirmed_daily. It is reasonable in reality. And the third table reflects the relation between confirmed deaths_daily, the two of them are positively correlated most of the time. However, sometimes, the number of confirmed deaths daily is low, the number of deaths daily is high. It might be because of the delay between the confirmation and deaths.

For addition, we also plot the correlations between deaths, confirmed and date. The number of deaths and confirmed people increases over time, and deaths and confirmed are positively correlated.

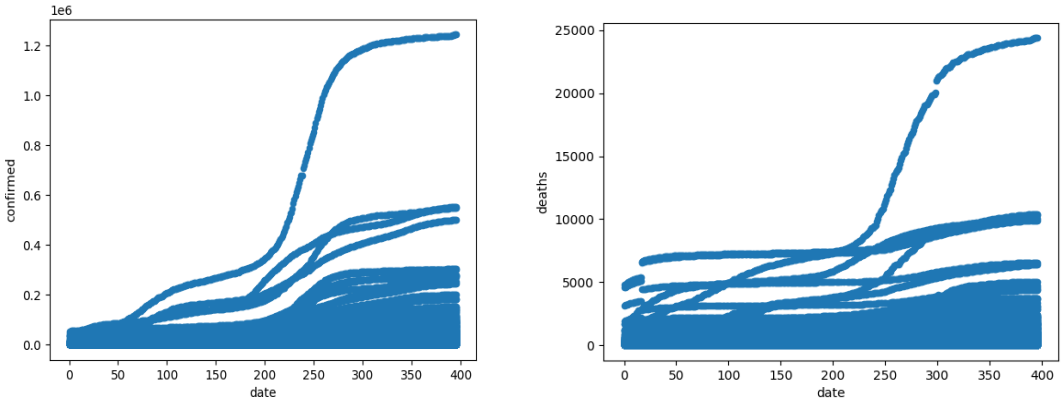


Figure 3: Correlation between date and confirmed and deaths

We drew a picture of the sentiment distribution in the sentiment data set to determine the boundary value of the binning strategy. As is shown in the valence intensity histogram, throughout the pandemic period, the proportion of negative emotions (around 0.4) is the largest. For the other four emotions, most of the data are concentrated between 0.3 and 0.6.

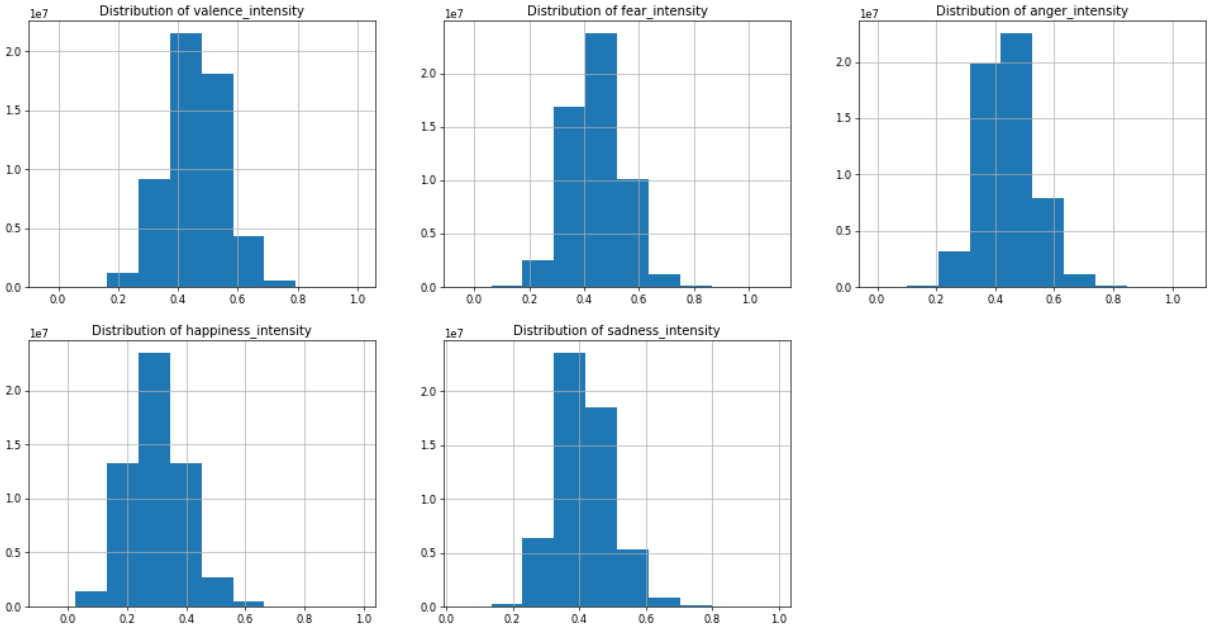


Figure 4: Distribution of different emotion

# 6.Cluster Analysis

## 6.1 statistical method

We implemented GMM for the dataset Covid Act Now. We chose to cluster the data of “vaccines administered” and “ICU Capacity Ratio” to show the correlation between people’s willingness to get vaccinated and the capacity of ICU beds. After running the code, the result will be saved as the image file “GMM_result.png”. To cluster the data, we read the data from csv files and then store them in a two-dimensional array, and then we do the clustering with the array.

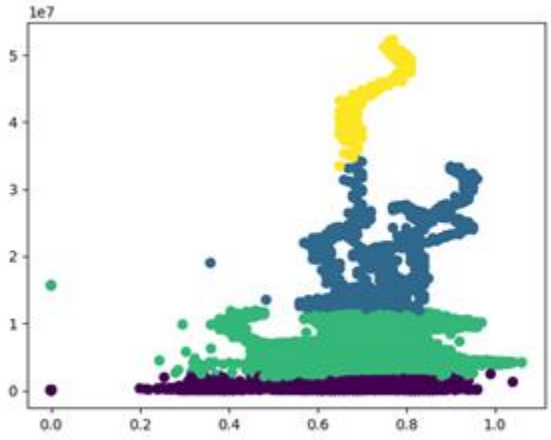


Figure 5: the result of GMM clustering

In the result, the x-axis represents the ICU capacity ratio and the y-axis represents vaccines that have been administered. From the image, we can observe that there is a trend that while the ratio of ICU capacity increases, people are more willing to get vaccinated. Although the result is slightly influenced by the fact that the vaccines were not available in the beginning of the pandemic, it is still clear for us to draw the conclusion.

After clustering, we used Silhouette to assess the quality of the clusters. Due to the memory constraints, the program cannot calculate the whole dataset of the two attributes all at once, so we calculate the mean score and output in the console at the end of the program. The result is around 0.50 and it is also because of the incomplete early data.

We utilize Calinski harabasz score to evaluate the clusters. The score is 12468.14, which means the performance of DBSCAN is pretty good.

## 6.3 partition clustering method

We put the bin labels of the valence intensity and the other four emotions into k-means for clustering. We set the value of K to 4, and then use PCA to reduce the dimensionality to 2.

For example, in the figure below, we clustered all tweets related to COVID-19 in February 2020 into 4 parts: positive, neutrality, anger and sadness. We think that the intensity of sadness and fear is very similar and it is hard to distinguish, so when clustering, these two sentiments are combined together. Neutrality means no sentiment is particularly high or low.

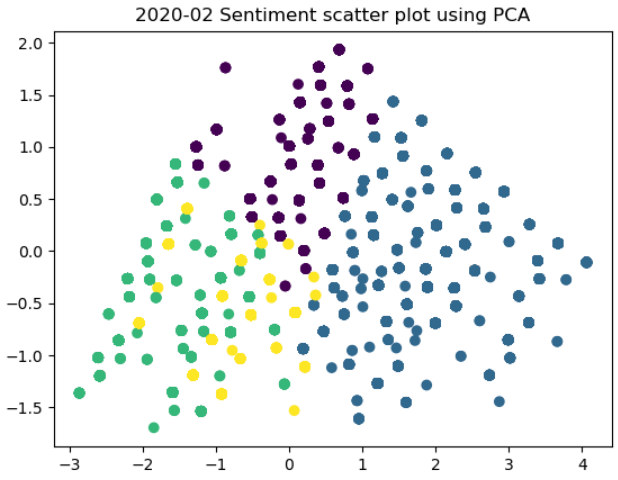


Figure 6: 2020-02 Sentiment scatter plot using PCA

**6.4 hierarchical clustering method**

We used hierarchical_cluster to analyse data “inIcuCurrently” and “death” in US_Daily_clean. We’ve tried several attributes and other methods on this dataset(We commented out those codes), but it seems many attributes of it have linear relation. For these two attributes， we got the graph below:

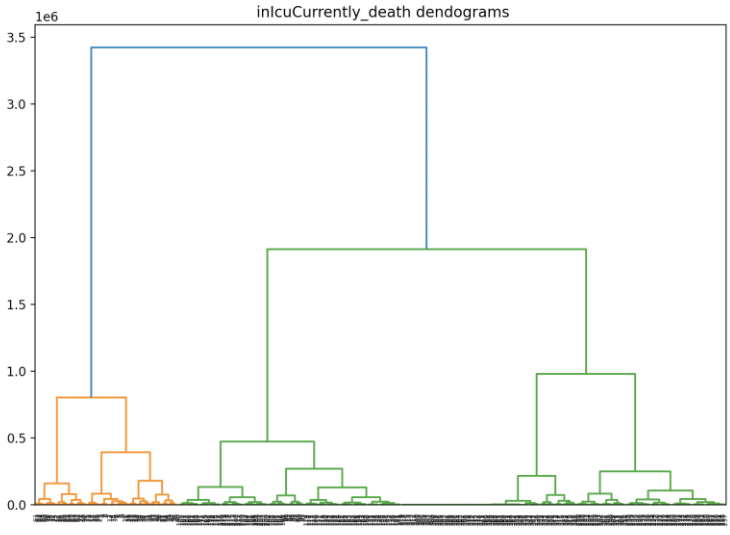


Figure 7 : Dendrogram1

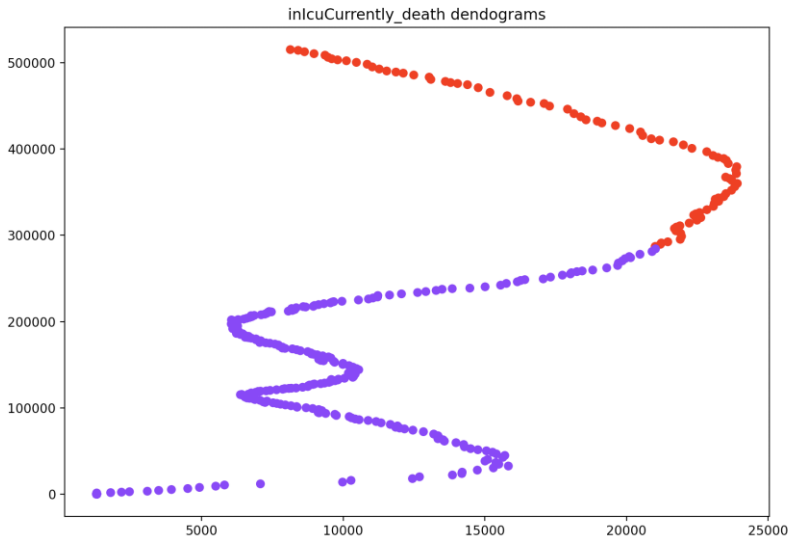


Figure 8: Dendograms

The average silhouette_score is: 0. 6245755016597423.From these graphs we found that when there are not many people in ICU, there are many different distributions of death rate, the highest reached when there is no more than 10000 people in ICU. So we think this may be the early and mid-stage of the epidemic.People still don't have a perfect system to identify and solve the covid problem.At the turning point on the far right of the figure, there are a large number of people in the ICU, but the death rate is not so high. It may be that the epidemic has stabilized and the hospital has experience in responding.

## 7.Topic Modeling

We did our topic modeling on Covid Tracking Project dataset, State_info.json of it is the text dataset we have. It contains some information about state news and COVID-related Twitter handles. We analyse the topics on state policies.

We  utilize LDA to generate the results of our data, and we set the number of topic parameter as five, it turns out 5 prevalent topics :

```
[(0,
  '0.039*"data" + 0.032*"idaho" + 0.022*"report" + 0.021*"2020" + '
  '0.019*"death" + 0.017*"updat" + 0.016*"probabl" + 0.013*"juli" + '
  '0.013*"day" + 0.012*"19"'),
 (1,
  '0.035*"2020" + 0.032*"test" + 0.030*"data" + 0.027*"case" + 0.022*"report" '
  '+ 0.015*"total" + 0.014*"updat" + 0.014*"2021" + 0.014*"decemb" + '
  '0.013*"pcr"'),
 (2,
  '0.031*"test" + 0.028*"2020" + 0.025*"data" + 0.024*"report" + 0.023*"south" '
  '+ 0.018*"case" + 0.018*"2021" + 0.016*"carolina" + 0.015*"result" + '
  '0.012*"twitter"'),
 (3,
  '0.040*"test" + 0.039*"2020" + 0.032*"data" + 0.023*"report" + 0.023*"2021" '
  '+ 0.021*"case" + 0.019*"total" + 0.019*"pcr" + 0.015*"decemb" + '
  '0.014*"updat"'),
 (4,
  '0.042*"test" + 0.041*"2020" + 0.028*"data" + 0.024*"report" + 0.023*"case" '
  '+ 0.023*"total" + 0.017*"pcr" + 0.016*"updat" + 0.016*"2021" + '
  '0.013*"novemb"')]
```

In this case, we can see many common words for covid-19-- test, data , report and something like that. Thus, we manually select some words as a new stop list to extend the older stop list, and it turns out:

```
[(0,
  '0.025*"decemb" + 0.017*"death" + 0.015*"idaho" + 0.015*"novemb" + '
  '0.015*"pcr" + 0.015*"confirm" + 0.014*"day" + 0.012*"januari" + '
  '0.012*"announc" + 0.012*"hospit"'),
 (1,
  '0.022*"pcr" + 0.021*"washington" + 0.019*"death" + 0.017*"decemb" + '
  '0.017*"day" + 0.016*"specimen" + 0.014*"result" + 0.014*"januari" + '
  '0.014*"februari" + 0.012*"novemb"'),
 (2,
  '0.021*"pcr" + 0.021*"peopl" + 0.021*"confirm" + 0.020*"neg" + 0.016*"posit" '
  '+ 0.015*"oregon" + 0.013*"utah" + 0.013*"dashboard" + 0.013*"probabl" + '
  '0.012*"march"'),
 (3,
  '0.022*"pcr" + 0.021*"result" + 0.017*"decemb" + 0.015*"novemb" + '
  '0.015*"januari" + 0.015*"death" + 0.015*"neg" + 0.014*"peopl" + '
  '0.014*"confirm" + 0.013*"number"'),
 (4,
  '0.017*"pcr" + 0.017*"florida" + 0.015*"hospit" + 0.015*"number" + '
  '0.014*"peopl" + 0.013*"decreas" + 0.012*"probabl" + 0.011*"explan" + '
  '0.010*"confirm" + 0.010*"valu"')]
```

These topics partly reflect the situation about the COVID-19, and topic 0 mentions that in December, November, January, people began to be hospitalized and some people were dead. Also, in topic one, it shows the similar situation, but it adds a new month February and no key word ‘hospit ‘, but more percentage of death (0.019 > 0.017), it might mean more or less people were dead so that it was mentioned more frequently. For topic 2, there are three key words negetive,positive, and  confirmed related to covid-19 situation. It means the epidemic gets better. Topic 4 reflects the number of confirmed people decreasing in Florida in many news reports. Here’s the table of the distribution of topics for each “document”:

| | Topic_no | Documents_number |
|---|---|---|
| 0 | 2 | 15 |
| 1 | 3 | 14 |
| 2 | 0 | 12 |
| 3 | 1 | 8 |
| 4 | 4 | 7 |

Here's an example of the output CSV file of the results of LDA.

| Document_No | Dominant_Topic | Topic_Perc_Contrib | Keywords |
|---|---|---|---|
| 0 | 0 | 0.7922 | decemb, death, idaho, novemb, pcr, confirm, day, januari, announc, hospit |
| 1 | 3 | 0.9986 | pcr, result, decemb, novemb, januari, death, neg, peopl, confirm, number |
| 2 | 3 | 0.9942 | pcr, result, decemb, novemb, januari, death, neg, peopl, confirm, number |
| 3 | 2 | 0.9662 | pcr, peopl, confirm, neg, posit, oregon, utah, dashboard, probabl, march |
| 4 | 3 | 0.9977 | pcr, result, decemb, novemb, januari, death, neg, peopl, confirm, number |

Document_No is the number of each document. Domiant_topic is the most relevant topic for each document. Topic_Perc_Contribution is the percentage contribution of the topic in a document. Keywords are the key words of the topic. There is another column of text, which represents the most relevant to the topic in each document. These data enable us to know which is the most relevant to a document and to see what text is similar to other text based on similar topics.