Finding factors of the COVID-19 pandemic by Clustering Methods

October 21, 2020

Abstract

We used K-means clustering, Mini Batch K-Means, Fazzy-c means, and Gaussian mixtures for our clustering. In this paper, we would introduce clustering techniques, evaluation methods that we used for each clusters and interpretation of our clusters.

0.1 K-Means

The K-Means algorithm clusters data by trying to separate samples in n groups of equal variance. The k-Means algorithm divides a set of samples into disjoint clusters, each described by the mean of the samples in the cluster, called 'centroids'. The centroid are not necessary samples in the cluster. K-Means algorithm aims to choose centroids that minimise the within-cluster sum-of-squares, called inertia, which defined as: $\sum_{i=1}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$, where x_i is a samples in the cluster of samples X, μ_j is the centroid of a cluster, and C represent the set of clusters. Specifically, to achieve the aim, the K-Means algorithm has three steps: firstly, randomly initialize samples from the dataset as centroids; secondly, assign the rest of the sample in the dataset to its nearest centroid; third, update the centroid by taking the mean value of all of the samples assigned to each previous centroid; fourth, repeat the second and third steps until the value in the third step is less than a fixed threshold.

K-means performed poorly on elongated clusters, or manifolds with irregular shapes, because within-cluster sum-of-squares is small only when clusters are convex and isotropic. In addition, within-cluster sum-of-squares are large in high-dimensional spaces, to address it we'd better do Principal component analysis(PCA) prior to K-means clustering.

The parameters that we can use to improve the performance of K-Means are "init='k-means++"', which initializes the centroids to be distant from each other, in order to avoid converging to a local minimum[1], and "sample weight", which assigns more weight to some samples for computing cluster centers.

0.2 Mini Batch K-Means

The MiniBatchKMeans is a variant of the K-Means algorithm. The algorithm iterates between two major steps until converges. Firstly, b samples are drawn randomly from the dataset, to form a mini-batch, which are assigned to the nearest centroid; secondly, update the centroids by taking the streaming average of the sample and all previous samples assigned to that centroid. The results of MiniBatchKMeans are relatively the same as K-Means, but it converges faster than K-Means.

0.3 Fazzy-c means

Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster. The algorithm is very similar to the K-Means algorithm: First, randomly initialize samples from the dataset as centroids. Second, Assign coefficients randomly to each sample for being in each cluster. Third, compute the new centroid for each cluster. Fourth, compute its coefficients of being in the clusters. Fiveth, repeat the third and fourth step, until the algorithm converged.

0.4 Gaussian mixtures

A Gaussian mixture model is a probabilistic model that assumes all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Each Gaussian distribution in the mixture is comprised of the following parameters: A mean μ defines its centre; A covariance $\sum$ defines its width. A mixing probability π defines how big the Gaussian function will be.

aussian Mixture Model Selection concerns both the covariance type and the number of components in the model. BIC and AIC are two criteria for model selection concerning both the covariance type and the number of components in the model.

1 Evaluation Methods

To evaluate the clusters, we selected the methods that could performed using the model itself without ground truth. Based on the properties of each clustering methods, we further choose the ones that could effectively evaluate the quality of each clusters individually. Let's briefly explain what each method means.

Silhouette Coefficients is defined for each samples and is composed of two scores: a is the mean distance between a sample and all other points in the same class; b is the mean distance between a sample and all other points in the nearest class, other than its own class. The Silhouette Coefficient for a single sample is defined as: $s = \frac{b-a}{\max(a,b)}$.

Silhouette Score is ranging from -1, incorrect clustering, to +1, highly dense clustering. Zero indicates overlapping clusters.

The silhouette score is generally high for convex clusters, such as density based clusters like those obtained through DBSCAN.

Calinski-Harabasz Index is defined as the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters.

The index is high, when clusters are dense and well separated. (The index is not ranging from -1 to +1.)

The Calinski-Harabasz index is high for convex clusters, such as density based clusters.

Davies-Bouldin Index indicates the average 'similarity' between clusters by comparing the distance between clusters with the size of the clusters themselves. Zero is the lowest possible score. Values closer to zero indicates a better cluster. The Davies-Boulding index is high for convex cluster, such as density based clusters. In addition, the usage of centroid distance limits the distance metric to Euclidean space.

Elbow method is also called **Distortion** in some cases. It is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

BIC is a criterion for model selection among a finite set of models, sepecially for Gaussian mixtures parameter selection. The model with the lowest BIC is preferred.

AIC estimates the information lost by a given model. The less info a model loses, the higher the quality of that model.

2 Interpretation Approaches

To explain the clustering results we need to find which features influence the most on clusters. One way is that if the average value of a feature ordered by clusters differs significantly among each other, that variable is likely important in creating the clusters. **Random Forest Feature Selection** is another way of finding which features are important in the generation of the clusters. We can set the clusters' labels as the target value, and apply Random Forest Feature Selection Method to select important features. We used Random Forest Feature Selection to select features after tried both.

Further more, to find exactly if the variation pattern of those selected features are the same as the variation pattern of the clusters, we need to use **Jenks Optimization Method** to cluster each feature individually. Jenks Optimization Method is doing single variable clustering. This method uses an iterative approach to group data by gradually minimizing variance within classes and maximizing variance between classes, which is similar with K-Means clustering on multiple features.

we used **v measure score** to test the agreement of each feature clustering labels and the clustering labels generated by using all features. A higher v measure score indicates a higher agreement of two clusters. Which means the variate pattern of the feature cluster is similar with the variate pattern of the cluster generated by all features.

So the steps briefly are: First, find the most variate features though Forest Feature Selection. Second, use Jenks Optimization Method to cluster those features individually, with initializing the number of components to the number that we find in the overall clustering. Third, using v measure score to evaluate the differences between the feature clusters and the overall clusters. A high score indicates a high influence of the feature to the overall clusters.

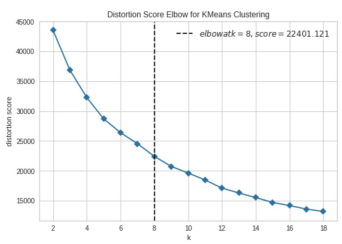
3 Contributions and Results

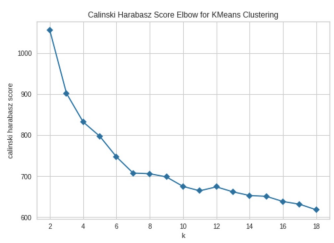
3.1 Clustering by K-means

Since the draw back of the K-means clustering method mentioned above, we did PCA to visualize the data, before clustering. We calculate the eigenvectors and eigenvalues of covariance matirx, to select the best principle components for PCA. The result shows best principle components is 8, where the explained variance ratio is approximately 95 percent.

Secondly, we evaluate all of the possible number of components of the clusters, by elbow score, silhouette score, and calinski-harabasz score. The package that we used to calculate and visualize the scores is "yellowbrick". It has an indicator that could automatically show the optimal number of components for each evaluation methods. The optimal number of components is 8 based on the indicator of Elbow score, which is 22401.121. However, Calinski Harabasz score and silhouette score continuous decrease dramatically. Which means that K-means is not a good clustering method in our case. There isn't a number of components where all of the scores are the best, even trade off.

The visualization of evaluations ranging from two to nineteen clusters are showing hear:





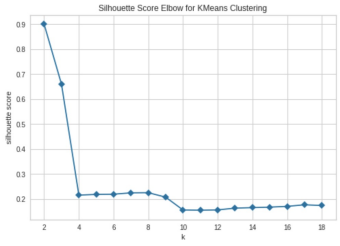


Figure 1: K-Means Evaluation Scores

The clustering result can be visualized across the country:

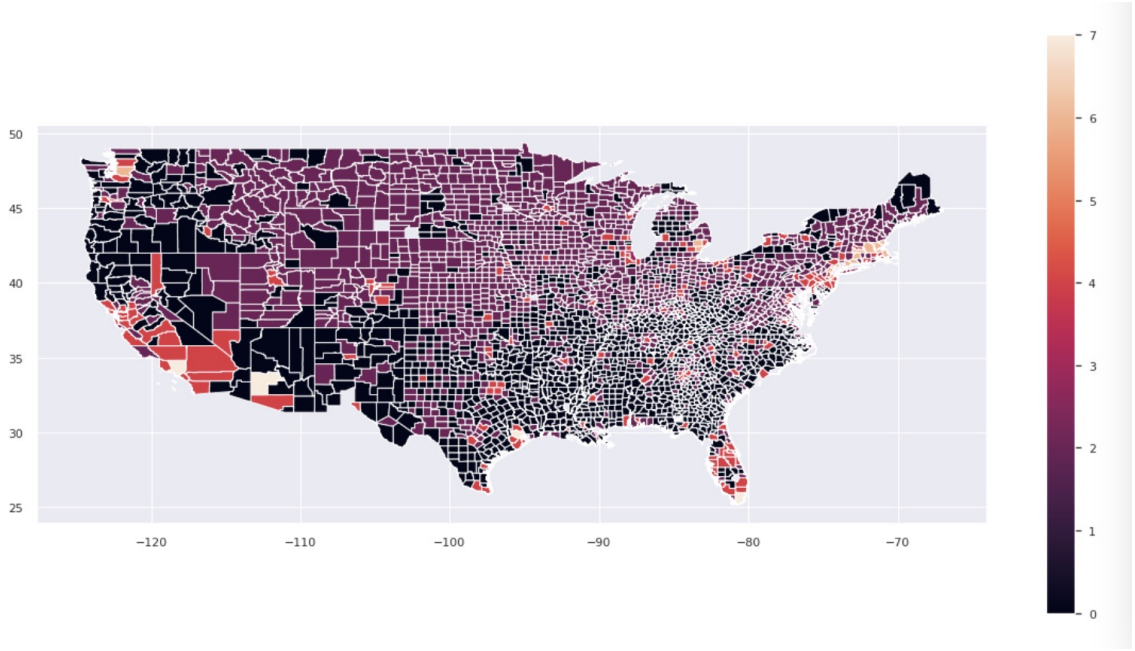


Figure 2: K-Means Map

To interpret the clusters, firstly, we selected the important features using Random Forest Feature selection by initializing the clustering labels as the target value. The top seven important features are listed hereand beside it is the corresponding plot show of the average variation of each features among different clusters.

- 'ranking socioeconomic(svi)',
- 'ranking householdcomp(svi)',
- 'ranking housingtransport(svi)',
- 'ranking minoritylang(svi)',
- 'population(svi)',
- 'cumulative cases august8',
- 'icu beds(kaiser)',

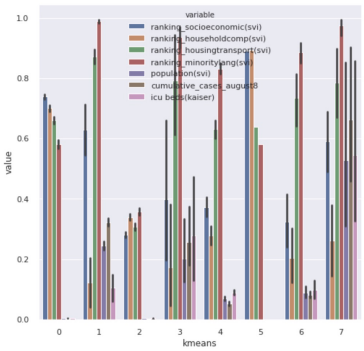


Figure 3: K-Means clusters Important Features

We used Jenks Optimization Method to cluster each features individually. Then we used V measure score between each feature clusters and the overall clusters to test the homogeneity between each feature clusters and overall clusters. A high score indicates a high homogeneity. The v measure score for each selected features:

ranking_socioeconomic(svi)	ranking_householdcomp(svi)	ranking_housingtransport(svi)	ranking_minoritylang(svi)	population(svi)	cumulative_cases_august8	icu beds(kaiser)
0.272713	0.169207	0.15193	0.100293	0.268914	0.283737	0.24517

Figure 4: V Measure Score of each selected features

The result shows that the variate pattern of feature clusters: 'ranking socioeconomic(svi)', 'population(svi)', 'cumulative cases august8', and 'icu beds(kaiser)' are closer to the K-Means clustering result. Therefore, those features make great influence on K-Means Clustering.

We proceed the same steps to find the optimal number of clusters for Mini Batch K-Means, Fazzy-c means,and Gaussian mixtures. But, the evaluation methods are different depending on their properties.

3.2 Clustering by Fazzy-c means

For Fazzy-c means, we used silhouette score, calinski harabasz score, and davies bouldin score. The optimal number of clusters is 3. The corresponding silhouette score is 0.17947678910037573, calinski harabasz score is 297.3418802201032, and davies bouldin score is 1.943634882800222, they are the best score in each of the three evaluation methods.

The visualization of evaluations ranging from two to nineteen clusters are showing hear:

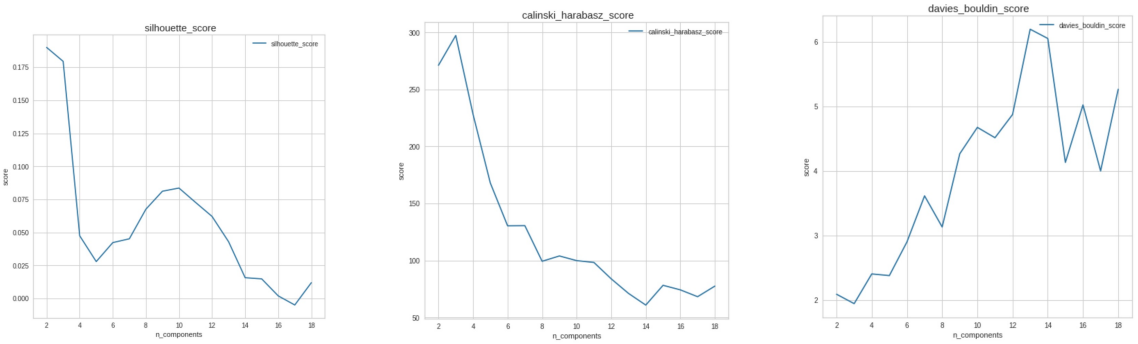


Figure 5: Fazzy-c Means Evaluation Scores

The clustering result can be visualized across the country:

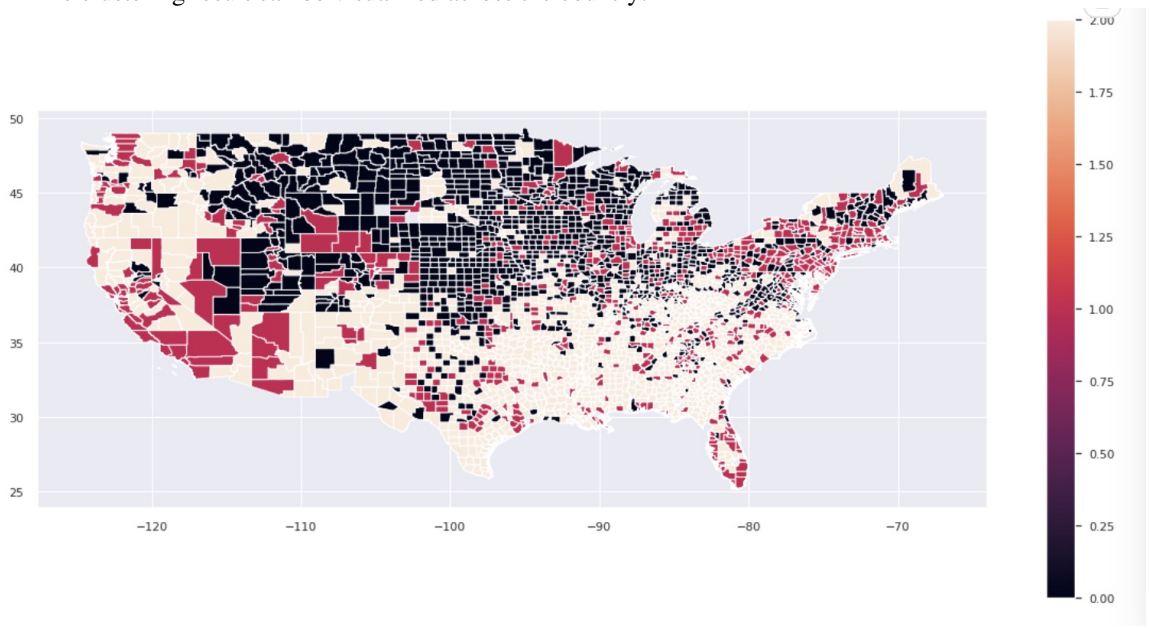


Figure 6: Fazzy-c Means Map

The top seven important features by Random Forest Feature selection are listed below. Beside it is the corresponding plot show of the average variation of each selected features among different clusters.

- “ranking socioeconomic(svi)”,
- “ranking housingtransport(svi)”,
- “ranking householdcomp(svi)”,
- “ranking minoritylang(svi)”,
- “population(svi)”,
- “rurality(irr)”,
- “cumulative cases august8’.

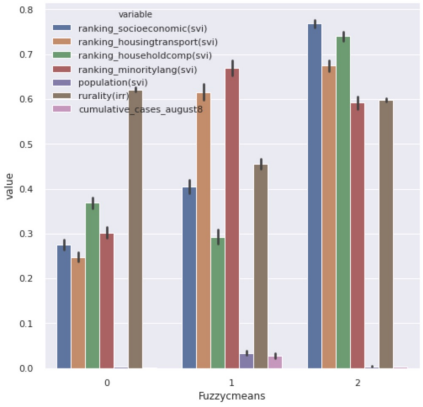


Figure 7: Fazzy-c Means clusters Important Features

The v measure score for each selected features:

ranking_socioeconomic(svi)	ranking_housingtransport(svi)	ranking_householdcomp(svi)	ranking_minoritylang(svi)	population(svi)	rurality(irr)	cumulative_cases_august8
0.336379	0.239229	0.233125	0.14406	0.081151	0.119135	0.055716

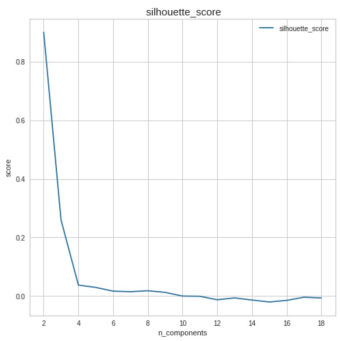
Figure 8: V Measure Score of each selected features

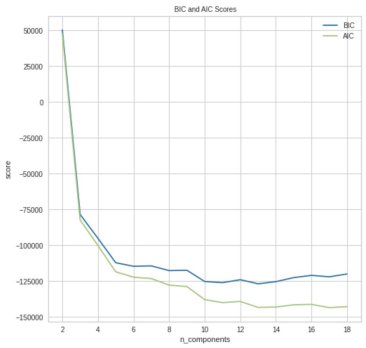
The result shows that the variate pattern of feature clusters: “ranking socioeconomic(svi)”, “ranking housingtransport(svi)”,and “ranking householdcomp(svi)” are closer to the Fazzy-c Means clustering result. Therefore, those features make great influence on Fazzy-c Means Clustering.

3.3 Clustering by Gaussian Mixture

For Gaussian Mixture, the evaluation methods we used are silhouette score, BIC and AIC. For BIC and AIC, both 3 and 5 clusters are optimal choices. But when the number of clusters is 5, the silhouette score is 0.028490858445860506 less than 3 clusters, where the silhouette score is 0.25964033557217897. The silhouette score is closer to zero, which means the generated clusters are overlapping with each other. Thus, the optimal number of clusters is 3.

The visualization of evaluations ranging from two to nineteen clusters are showing hear:





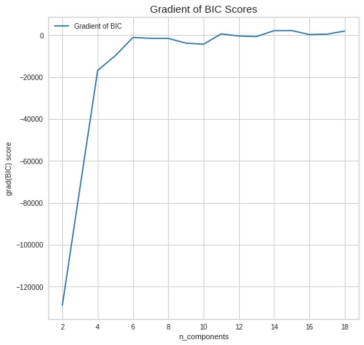


Figure 9: Gaussian Mixture Evaluation Scores

The clustering result can be visualized across the country:

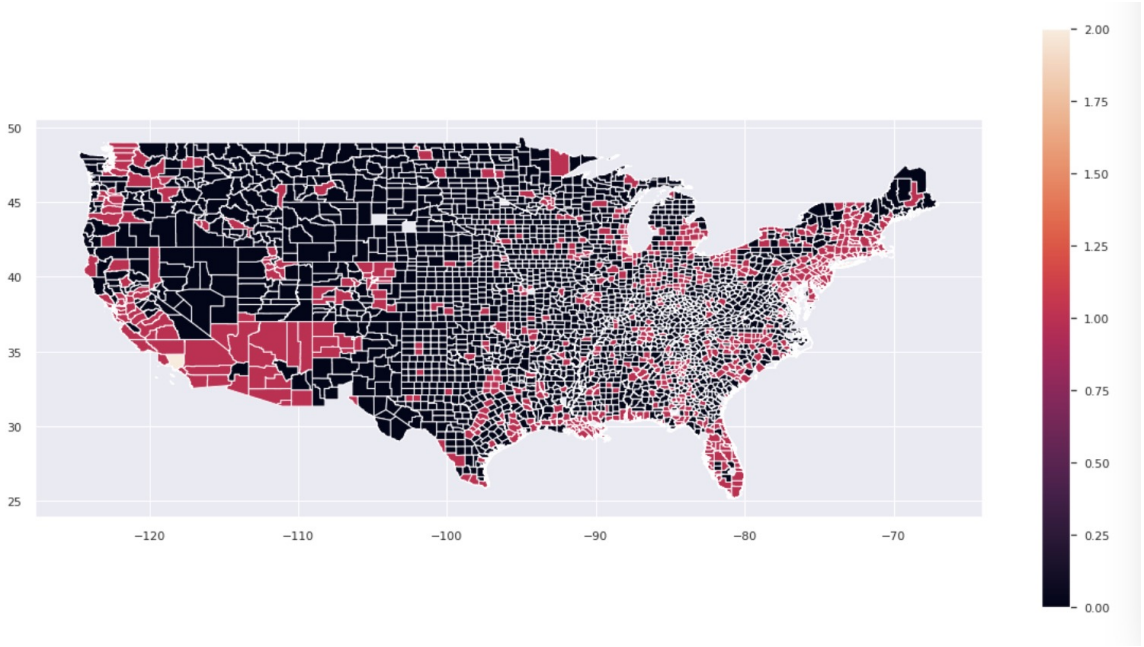


Figure 10: Gaussian Mixture Map

The top seven important features by Random Forest Feature selection are listed below. Beside it is the corresponding plot show of the average variation of each selected features among different clusters.

- 'new cases july23',
- 'slope to first peak',
- 'population(svi)',
- 'cumulative cases august8',
- 'cumulative deaths august8',
- 'icu beds(kaiser)',
- 'rurality(irr)'.

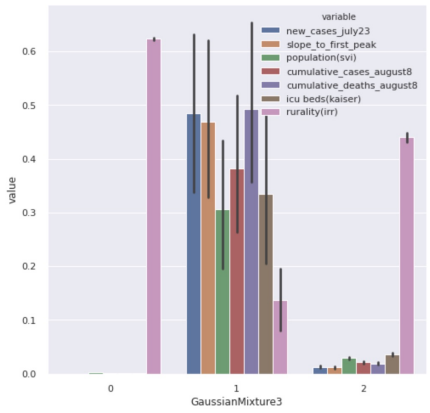


Figure 11: Gaussian Mixture Important Features

The v measure score for each selected features:

new_cases_july23	slope_to_first_peak	population(svi)	cumulative_cases_august8	cumulative_deaths_august8	icu beds(kaiser)	rurality(irr)
0.09065	0.086219	0.169633	0.139323	0.117541	0.13449	0.298137

Figure 12: V Measure Score of each selected features

The result shows that the variate pattern of feature clusters: 'rurality(irr)' is closest to the Gaussian Mixture clustering result. Therefore, this feature greatly influence on Gaussian Mixture Clustering.

3.4 Clusterng by MiniBatchKMeans

The evaluation methods we used for MiniBatchKMeans are silhouette score, calinski harabasz score, and elbow score. We can refer to the figure below, both silhouette score and calinski harabasz score indicate that 4 is the number of optimal clusters. However, elbow score indicates that 12 is the number of optimal clusters. But, if we see the plot show of the elbow score, when number of cluster is 4, the corresponding elbow score is local minimum, which is a good choice. Thus, the optimal number of clusters is 4.

The visualization of evaluations ranging from two to nineteen clusters are showing hear:

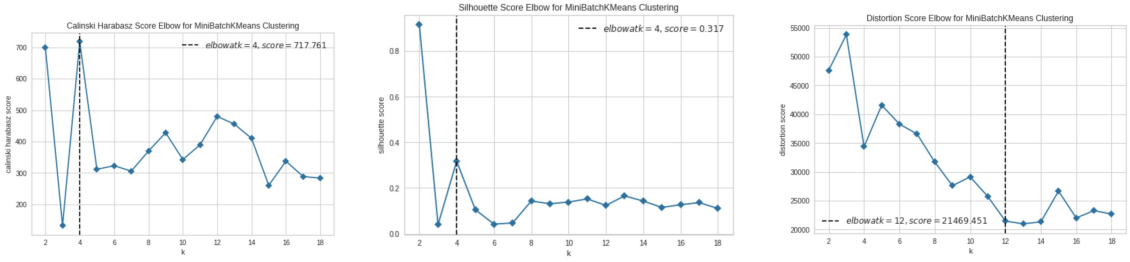


Figure 13: MiniBatchKMeans Evaluation Scores

The clustering result can be visualized across the country:

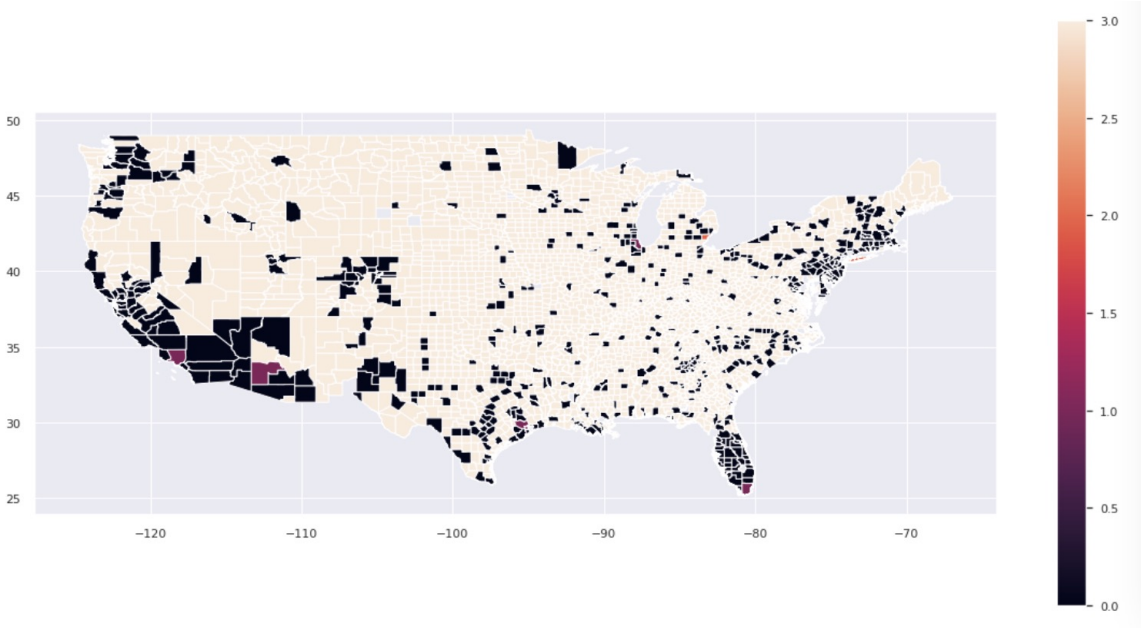


Figure 14: MiniBatchKMeans Map

The top seven important features by Random Forest Feature selection are listed below. Beside it is the corresponding plot show of the average variation of each selected features among different clusters.

- 'population(svi)',
- 'cumulative cases August8',
- 'overall score',
- 'rurality(irr)',
- 'icu beds(kaiser)',
- 'slope to first peak',
- 'ranking minoritylang(svi)'.

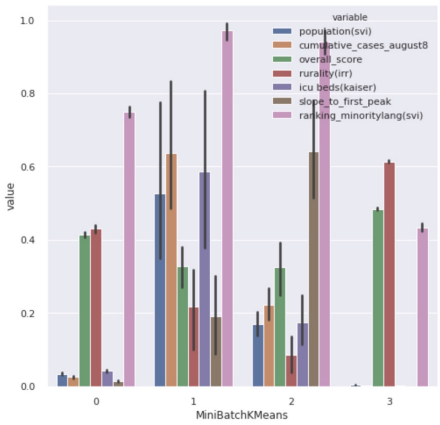


Figure 15: MiniBatchKMeans Important Features

The v measure score for each selected features are:

population(svi)	cumulative_cases_august8	overall_score	rurality(irr)	icu beds(kaiser)	slope_to_first_peak	ranking_minoritylang(svi)
0.289316	0.277041	0.079274	0.244794	0.317856	0.110003	0.110972

Figure 16: V Measure Score of each selected features

The result shows that the variate pattern of feature clusters: 'population(svi)', 'cumulative cases August8', 'rurality(irr)', 'icu beds(kaiser)', are closer to the MiniBatchKMeans clustering result. Therefore, those features greatly influence on MiniBatchKMeans Clustering.

4 Conclusion

- K-Means is not an optimal clustering method in our case. Fazzy-c Means and MiniBatchKMeans are optimal clustering methods in our case.
- According to the important features selected by each of the clustering methods, include, K-Means, Fazzy-c Means, Gaussian Mixture, and MiniBatchKMeans, features such as, 'population(svi)', 'rurality(irr)', 'icu beds(kaiser)', 'cumulative cases august8', and 'ranking socioeconomic(svi)' are metioned more frequently. Thus we concluded that those features greatly impact on the spread of COVID-19 pandemic in America. Features 'ranking housingtransport(svi)', and 'ranking householdcomp(svi)' also impact on the spread of COVID-19 pandemic in America to some extent.

References

[1] Arthur, David, and Sergei Vassilvitskii, *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics (2007).