In the process of data mining, we use the patient's mobility level, age group, and acquisition group to determine whether a specific location is related to outbreak or not. We extracted data from fact table, patient dimension, mobility dimension such as retail_and_recreation, grocery, parks, transit_stations, workplaces, age_group, acquisition_group, outbreak_related, etc. Moreover we use three different data mining algorithms by analyzing and comparing the relevant data.

The decision tree is the first algorithm we use, because its main advantage is the availability of the model and the speed of classification. For learning, we use the training data to build a decision tree model based on the principle of minimizing the loss function; for prediction, we use the decision tree model to classify new data.

The second algorithm we use is "Gradient Boosting". The advantage of Gradient Boosting is that the Gradient Boosting algorithm first calculates the negative gradient of the current model on all samples in each iteration, and then trains a new weak classifier with this value as the target to fit and calculate the weight of the weak classifier, and finally realizes the update of the model.

The last algorithm we used is random forest, according to the definition of random forest: a random forest is composed of many decision trees, and there is no association between different decision trees. When we perform the classification task, new input samples enter and let each decision tree in the forest judge and classify separately, each decision tree will get a classification result of its own, which one of the classification results of the decision tree has the most classification, then the random forest will take this result as the final result.

After analyzing the covid-19 data by using three different algorithms, we concluded that only a small number of patients are associated with outbreak.Secondly, we used the data to build different models and graphs to observe the mobility level of patients in different regions to infer whether the region was relevant to outbreak. Although all three algorithms produced some bias in the results, the data results were within acceptable range.