A project report on "Data Visualization and K-means clustering project for datasets on number of hospital beds all over the world currently available during CoVid-19 pandemic"

Dataset Information

The data contains 18 datasets and 14 datasets have been visualized here. The other 4 datasets were either incomplete, or they were same as any one of the 14s. All the datasets have the following columns:

> Country, State, County, lat for latitude, long for longitude, type for type of bed, measure, beds, population, year, source, source_url.

For each dataset, the name of countries or states are kept in one list variable called x1, and their beds capacity per thousand population is kept in another list variable y1.

We have only used the Country or State column and the beds column for the project.

Data Preprocessing

For some datasets, like India's, Italy's, there are states and not counties. But for some datasets, like Germany's, there are states, and for each state there are many counties(Or for each continent there are many countries). For the datasets of first kind, the name of the state is taken as X and the bed capacity is taken as Y. For the datasets of second kind, average bed capacity of states taken over the counties are taken as the Y variable and the name of the state is taken as X. For some datasets there are type as TOTAL and for some there are many types like TOTAL, ICU, PSYCHIATRIC, ACUTE. For datasets which have more than one type, only TOTAL is taken for the project.

Plots

Then State or Country(X) against the capacity of the bed for the State or Country(Y) are plotted as a bar diagram and a scatter diagram.

Algorithm Explanation

Then K-means algorithm is implemented. For that, all the countries(or, states) are sorted according to their bed capacity in ascending order. Then it was desirable to choose the initial K clusters as perfect as possible. As quantiles distribute a dataset evenly, for that K quantiles have been taken as the initial K clusters. These are K lists with only one item which is the cluster center.

A Z matrix is taken. It is a Nx2 matrix where N is the number of states and the 2 columns are sorted bed capacity and the corresponding state names.

Then for each data in Y other than the centers, their distance from the centers are stored in one list and from there, the minimum distance is found and thus it is decided in which cluster this data will go. As new data comes, the center is updated.

cluster_areas is the list of clusters we get by implementing the K-means algorithm and cluster_areas_sklearn is the list of clusters we get by implementing K-means using Scikit-learn. cluster_areas_sklearn is created by using the Kmeans labels.

Analysis Algorithm

Now to compare the two clusters, a dictionary "data" is created where the keys are country names(or, states) and the values are lists of two dictionaries. the first dictionary have cities as keys which are greater in order than the "data[key]" country and its value is either "Same" or "Different". "Same" means the two cities belong to the same cluster in cluster_areas, and "Different" means the two cities belong to different clusters in cluster_areas. The second dictionary also have keys which are greater in order than the "data[key]" country and its values are also either "Same" or "Different". Here "Same" or "Different" means whether the two cities belong to same or different clusters in cluster_areas_sklearn.

Then for each data[key],it is compared with other cities whether they both belong to same cluster in both cases(implementing k-means with algorithm and implementing k-means with sklearn) or they both belong to different clusters in both cases. If they both belong to same cluster or different cluster in both cases, then they are matched in both cases, otherwise they are not matched. Then a percentage value is calculated by dividing all possible cases from total number of matches. If it is more than 60% or 70%, then the clusters we get by implementing the algorithm is agreeing with the clusters we get using sklearn in more than 60% or 70% cases, which is reasonably good. It is called Match percentage.

Output Information

The output of each dataset contains the bar diagram, the scatter diagram, the clusters we get by implementing K-means algorithm, the clusters we get by implementing K-means using scikit-learn, the match percentage of the two set of clusters, and two scatter diagrams for the two set of clusters.