**Introduction:**

The COVID-19 pandemic has been wreaking havoc globally for the larger part of 2020. My home country of India is one of the most severely affected countries in the world and, while many countermeasures are currently underway, the virus still seems to be on the winning side.

One problematic issue that India faces is the enormous percentage of rural population that, most often, do not have access to proper medical facilities. This creates significant problems for anyone attempting to reach out to the rural population and these sections of the country almost always require special planning in terms of distribution due to the lack of passable roadways. I believe that it would be immensely helpful for anyone planning a rollout of supplies and eventually, a vaccine, to be able to understand not only the break-up of urban/rural population across states but also the number of hospitals in each state. I believe that this information can be an immensely beneficial input to any strategic rollout of any resources/medication/vaccines.

**Description of the data:**

Through this project, I aim to use the FourSquare API to discover the density of hospitals in a state using their capital cities as a reference point.

In addition to the data collected from FourSquare, I have also collected data on the Urban and Rural population in each state.

The data used contains the following variables:

State (str): The name of the state/Union Territory

Population (num): The population of the state as per the 2011 Census

Capital (str): The capital city of the state/Union Territory in question

U to R (num): The ratio of Urban to Rural population in the state

Latitude (num): The latitude value of the state capital (captured using geolocator)

Longitude (num): The latitude value of the state capital (captured using geolocator)

Area (num): The area of the state in square kilometers.

Active Cases (num): The number of currently active COVID cases in each state.

The geographical coordinates will be used to discover hospitals within a given radius of the capital city of each state. All of these locations will be represented on a map of India.

The final outcome of this project will be a map containing all of the states (using the capital city as reference). Each capital will be color-coded based on the cluster of which it is a part.

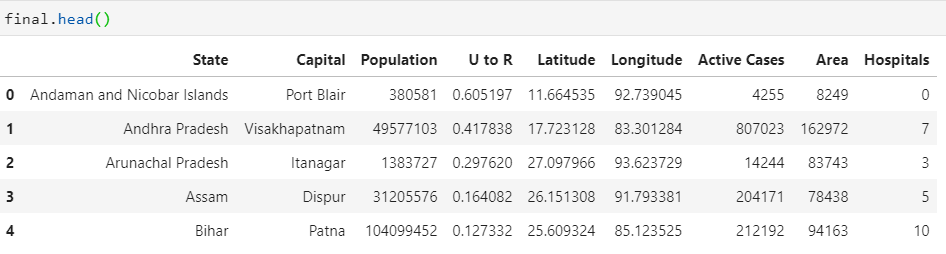
**Sources:**

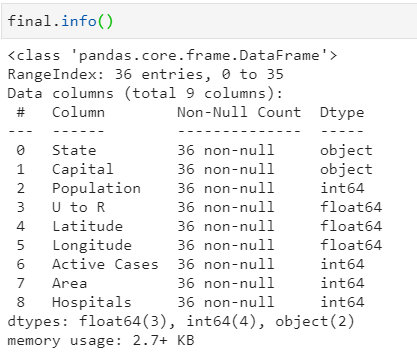
1. <https://en.wikipedia.org/wiki/List_of_state_and_union_territory_capitals_in_India>
2. <https://en.wikipedia.org/wiki/List_of_states_and_union_territories_of_India_by_population>
3. <https://covidindia.org/>
4. <https://en.wikipedia.org/wiki/List_of_states_and_union_territories_of_India_by_area>
5. FourSquare API

**Methodology:**

For the purpose of this project: creating clusters within the 36 states and union territories of India, we will make use of the K-means clustering algorithm in the scikit-learn package.

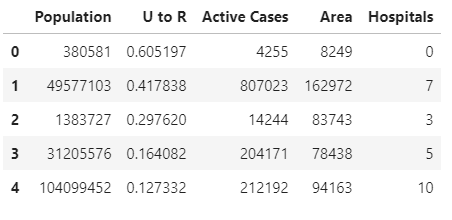
Let’s take a look at our compiled dataset:



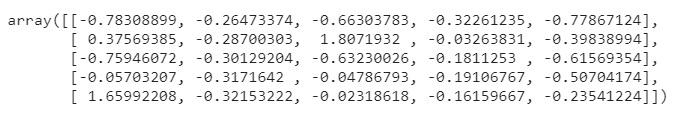


As we can see, the majority of the data is continuous and, therefore, we remove the State and Capital columns so that we can go ahead with our clustering.

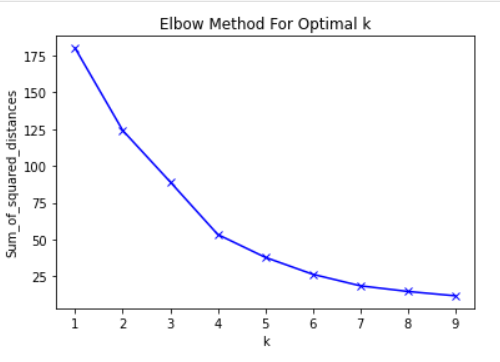
The final data set to be used for creating clusters is as follows:



In order to perform clustering, we first require our data to be scaled. In order to do this, we make use of the StandardScaler function from scikit-learn. After scaling, the data looks as follows:



Now that our data is preprocessed, we can go ahead with our clustering. The first step is to determine the optimal number of clusters using the elbow method which requires us to plot the WSS values against the number of clusters. The graph looks as follows:



As we can see, there is a clear elbow at k=4. Therefore, we proceed with 4 clusters.

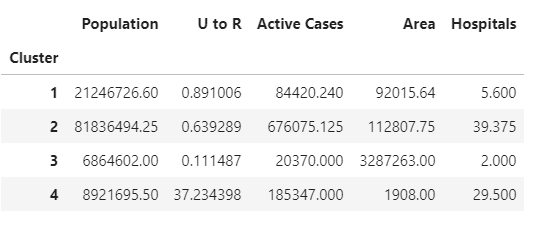
Having split the data into 4 clusters, we can now see the mean and median values of each variable for each of the clusters.

Figure 1: Mean values for each cluster on each of the variables.

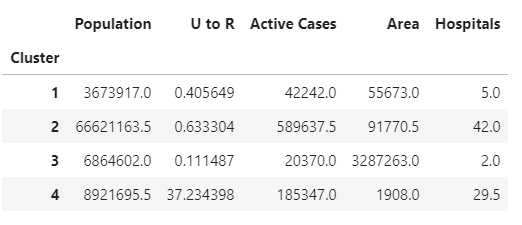
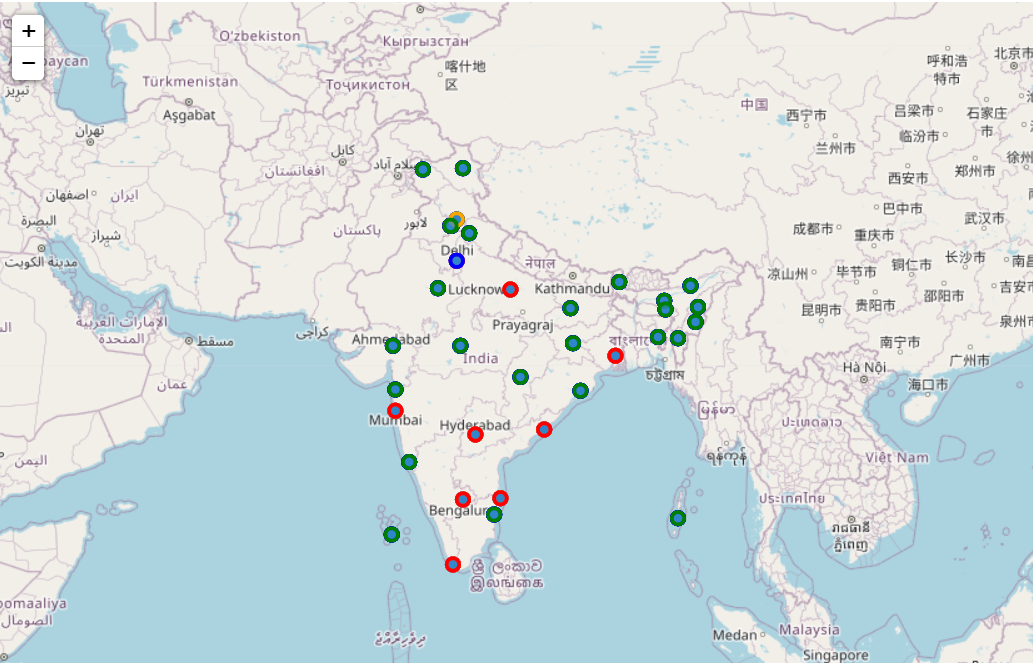


Figure 2: Median values for each cluster on each of the variables.

Finally, we create an interactive map using the Folium package. It contains all the states, using their capital cities as a reference point and is interactive and color-coded as mentioned in the cluster analysis.



**Results:**

Cluster 1 is the largest of all the clusters and cluster 4 is the smallest (containing only one state).

The attributes and names of each cluster are listed below.

Cluster 1:

Name: Mildly affected tier-2 states.

Color: Green

Attributes:

* Second highest population.
* Almost equal urban and rural population.
* Relatively low case count.
* Smallest area on average.
* Low hospital count.

Cluster 2:

Name: Severely affected developed states

Color: Red

Attributes:

* Highest population.
* Low urban population.
* Very high case count.
* Fairly large in area.
* Highest hospital count.

Cluster 3:

Name: Relatively unaffected tier-3 states.

Color: Orange

Attributes:

* Lowest population.
* Extremely low urban population.
* Lowest case count.
* Largest area on average.
* Extremely low hospital count.

Cluster 4:

Name: Small urban states.

Color: Blue

Attributes:

* Relatively low population.
* Overwhelmingly urban population.
* Second highest case count.
* Smallest area on average.
* High hospital count.

**Discussion:**

In terms of drawing up a distribution plan for any medical supplies or, eventually, a vaccine, it is imperative that the highest priority be given to clusters 1 and 2. These are the mildly affected tier-2 states and the severely affected developed states for the following reasons:

1. The developed states are hubs of population and commerce. It is imperative to get these states up and running as soon as possible so as to minimize the economic impact of the pandemic on India.
2. The Tier-2 states appear to have a very low level of medical infrastructure development and will require the most effort not only in terms of providing the required healthcare but also in reaching them as they are underdeveloped and are difficult to reach via established supply lines.

Delhi can be used as a hub to reach out to many of the states mentioned above and, since it is the smallest cluster, using it in this fashion would allow for a very swift response to a confirmed vaccine.

**Conclusion:**

Establishing supply lines in India is incredibly difficult given the variance in infrastructure and levels of urbanization. However, if prioritized correctly by looking at each state individually, these challenges can be overcome to a large extent.