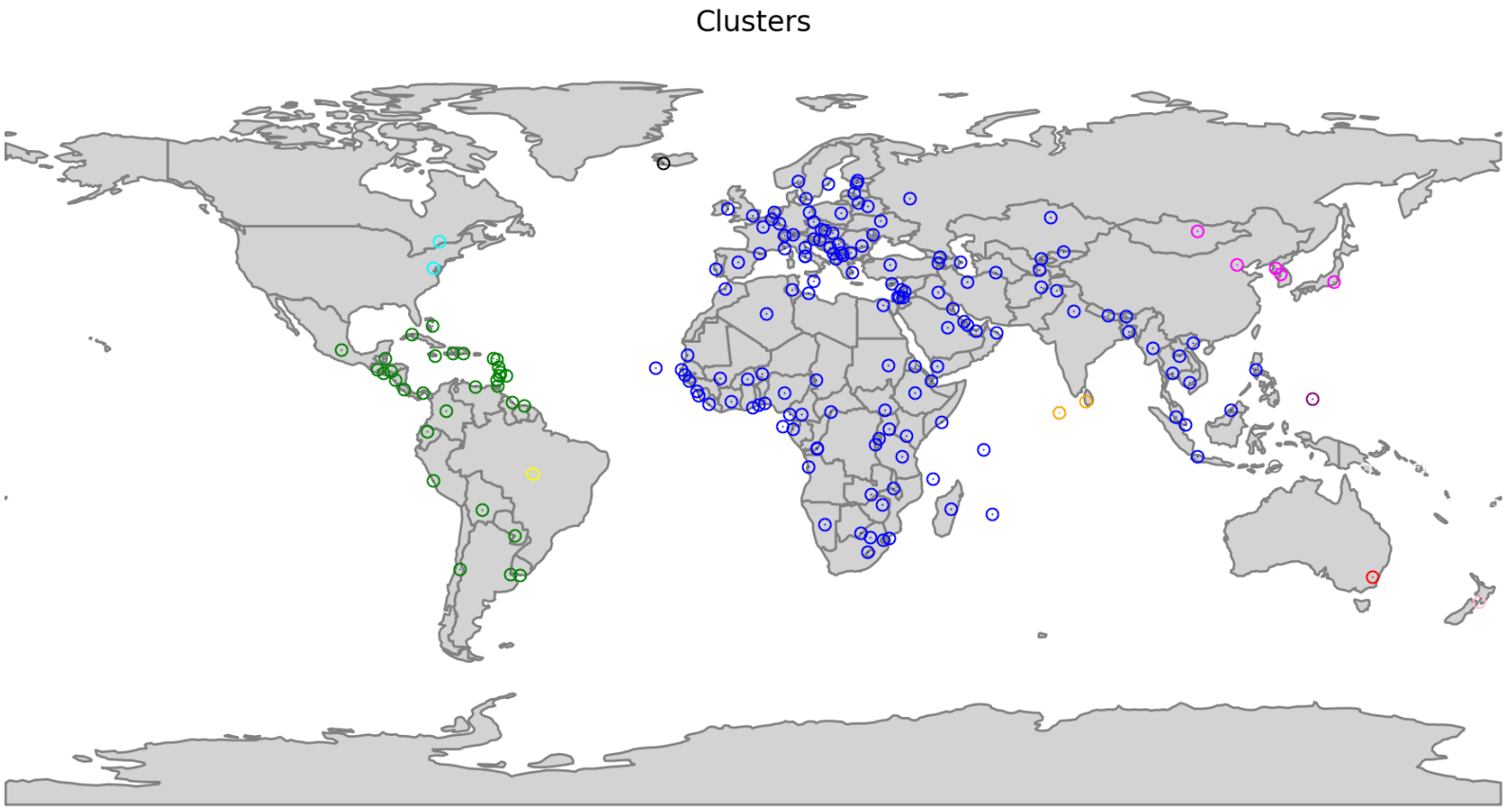
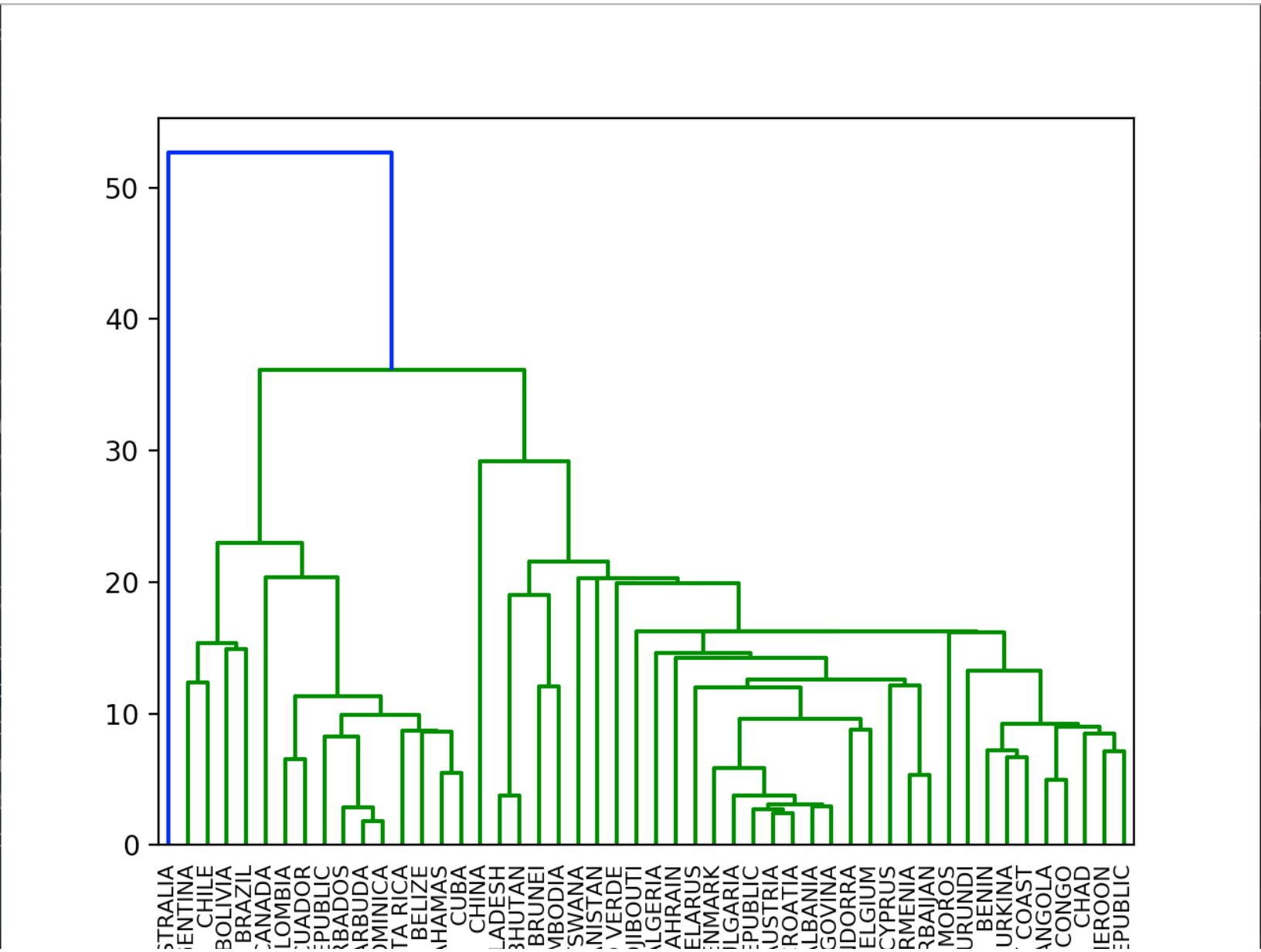
Pratishta Rao, Srikant Lakshminarayan

Big Data Analytics - Home Work 8

1. A map of the world withe the capital cities plotted is as follows:



2. Dendrogram showing the top 50 countries is as follows:



3. Who did What? How do what you did? Was anyone on quality assurance? How did you divide the tasks? What tasks where there?

Solution:

The following is how the tasks were divided between Srikant and Pratishta:

Srikant:

- Implemented the agglomeration clustering
- Plotted the plots on the world map
- Implemented the dendrogram

Pratishta:

- Implemented the function to get the gps points
- Implemented the code modularity/ Comments
- Testing/Debugging
- Writeup

Yes, there was someone to check the quality assurance. As described above the tasks were divided between the two of the team members.

The following are the tasks for the given homework:

- Get the latitude, longitude of the cities given
- Implementing the clustering algorithm
- Plotting the values obtained on
- Testing and debugging
- Write up

4. Write a conclusion showing what you learned and that you learned something. What issues did you face? What went wrong ? What worked earlier? Provide strong evidence of learning.

Solution:

Following are the things learned:

- Neither of us had worked with a few packages used like geopy, geopandas before. So this homework involved the both of us to go look into the packages and do a lot of trial and error before we got the things working.
- Understood the working of the agglomeration clustering. It took us implementing the algorithm thrice before we got it right.

- Learned how to plot the points on the world map since both of us had never done that before.

Some the issues we faced was trying and getting the gps points. We were getting the 404 error for too many requests. We tried various methods to fix it but it was a hit and miss. After we finally managed to get the points one time it worked, we saved these points in a csv file and use it to implement the rest of the code. In the testing/ debugging phase we weren't getting the right number of data points in the cluster and hence, the three different attempts to implement the agglomeration algorithm. We had problems with the gps points. It had missed to get the latitude and longitude points for a few cities. We didn't find this error until the last minute. But once we did, we put a try catch block to fix the error.
We also implemented the algorithm using the scikit package in order to have a better comparisons of the map and the dendrogram. But, the haversine distance matrix values were completely different from the one we had implemented. We didn't have enough time to look into the issue. But I think it was because, I had implemented a function to calculate haversine distance and using the affinity = precomputed. That's where I must have gone wrong.