Grouping cities based on their Culinary Scene

Introduction:

As the World gets more and more globalized every year, people travel between and move to different cities at an unprecedented rate. While the benefits of this increased trade have been well established, the problems that come along with it are often less talked about. Here when I say problems, I refer to the problems faced by the people who are moving and not the politics that come with it.

While the cultural difference that each person faces due to this migration is humongous, I address one of the most important aspects of it – The Food.

This project is an attempt to use machine learning to cluster together the 200 most populous cities of the World based on the different kinds of restaurants in each of them.

Data:

I used the Urban agglomeration data from the department of Economic and Social Affairs of the United Nations to get the population data of the cities of the world.

I used the Foursquare database to get the data on the different kinds of restaurants in each of these cities.

https://github.com/arunkxip/Capstone-project/blob/master/Cities.xls

This is a link to the population data.

Methodology:

- The 200 most populous cities as of 2020 based on UN estimates were identified from the dataset of thousands of urban agglomerations in different regions of the world.
- Just the relevant features were filtered out from the above data set.
- Used the location coordinates in the above dataset to get the different kinds of restaurants in each of these cities using the Foursquare API.
- Used KMeans clustering algorithm to sort these cities into different groups.

Results and Discussions:

As is evident from the map attached with the notebook, a disproportionate number of groups are clubbed into a single cluster. Two possible reasons can be attributed to this error

- The KMeans algorithm: Each kind of a restaurant was an attribute in this unsupervised learning algorithm. Since equal weightage was given to all the attributes, the KMeans algorithm tends to club together data points that have a similar shape (A very high number of a particular attribute. Ex: Two cities will be clustered together if they have only one kind of a restaurant). Also, cities where the a most of features are zeroes tend to be clustered together.
- Unequal amount of data in certain cities. This is because of the open sourced nature of the foursquare database, It is only as good as its users' classifications. If there are more users in some areas, then those areas tend to have more data. This disproportionate amount of data both affects the quality of clustering in general and particularly in the KMeans algorithm.

Conclusion:

The algorithm works excellent on the cities for which ample data on the kinds of Restaurants in it are available. This is particularly the case with the regions of North America, Europe, the middle East and south-east Asia. If the problem was to take a more localized approach, being exclusively applied to one of the above said regions, the obtained results would be much more distinct and a lot more insights can be drawn from it.