**Project Report: Analysing Localities based on data from Zomato API, and recommending localities based on given parameters.**

**Introduction:**

The problem statement states that given a dataset consisting restaurants, cuisines served, and their coordinates, we need to identify similar localities, based on which localities prefer a similar kind of cuisines. Similar cuisines add to the similarity of two localities, and using K Means clustering, we can identify clusters with similar localities. A cluster can have localities with different “1st popular cuisine”, but there can be other less popular cuisines that contribute to their similarity.

Restaurants can often be used to analyse localities. Our analysis can be extended to recommending localities to new and upcoming restaurants, and giving them alternative cuisines to try in these localities. The analysis can also be extended to analyse the average spending or the localities that have similar restaurants in terms of probable average spending.

Food being a very integral part of our cultures, should give us neighbourhoods/localities that are similar to each other, with an acceptable accuracy.

Most of the restaurant businesses would be interested in this model and its analysis.

**Data:**

The data we use has the following columns. Every Restaurant contains the following variables:

• Restaurant Id: Unique id of every restaurant across various cities of the world  
• Restaurant Name: Name of the restaurant  
• Country Code: Country in which restaurant is located  
• City: City in which restaurant is located  
• Address: Address of the restaurant  
• Locality: Location in the city  
• Locality Verbose: Detailed description of the locality  
• Longitude: Longitude coordinate of the restaurant's location  
• Latitude: Latitude coordinate of the restaurant's location  
• Cuisines: Cuisines offered by the restaurant  
• Average Cost for two: Cost for two people in different currencies  
• Currency: Currency of the country  
• Has Table booking: yes/no  
• Has Online delivery: yes/ no  
• Is delivering: yes/ no  
• Switch to order menu: yes/no  
• Price range: range of price of food  
• Aggregate Rating: Average rating out of 5  
• Rating colour: depending upon the average rating colour  
• Rating text: text on the basis of rating of rating  
• Votes: Number of ratings casted by people

The collected data has been stored in the Comma Separated Value file Zomato.csv. Each restaurant in the dataset is uniquely identified by its Restaurant Id.

Zomato API Analysis is one of the most useful analysis for foodies who want to taste the best cuisines of every part of the world which lies in their budget. This analysis is also for those who want to find the value for money restaurants in various parts of the country for the cuisines.

Dataset credits: Shruti Mehta, Kaggle.

**Methodology:**

We use K Means clustering to work on the project for getting meaningful ideas.

The reason for using K Means Clustering is that:

k-means clustering is a method of [vector quantization](https://en.wikipedia.org/wiki/Vector_quantization), originally from [signal processing](https://en.wikipedia.org/wiki/Signal_processing), that aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) n observations into k clusters in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) with the nearest [mean](https://en.wikipedia.org/wiki/Mean) (cluster centers or cluster [centroid](https://en.wikipedia.org/wiki/Centroid)), serving as a prototype of the cluster. This results in a partitioning of the data space into [Voronoi cells](https://en.wikipedia.org/wiki/Voronoi_cell). It is popular for [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) in [data mining](https://en.wikipedia.org/wiki/Data_mining). k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult [Weber problem](https://en.wikipedia.org/wiki/Weber_problem): the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, Better Euclidean solutions can be found using [k-medians](https://en.wikipedia.org/wiki/K-medians_clustering) and [k-medoids](https://en.wikipedia.org/wiki/K-medoids).

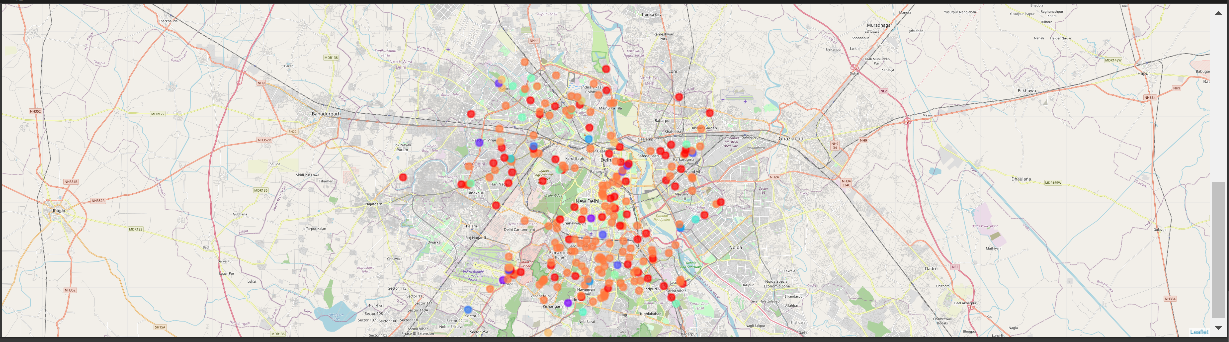
The problem is computationally difficult ([NP-hard](https://en.wikipedia.org/wiki/NP-hardness)); however, efficient [heuristic algorithms](https://en.wikipedia.org/wiki/Heuristic_algorithm) converge quickly to a [local optimum](https://en.wikipedia.org/wiki/Local_optimum). These are usually similar to the [expectation-maximization algorithm](https://en.wikipedia.org/wiki/Expectation-maximization_algorithm) for [mixtures](https://en.wikipedia.org/wiki/Mixture_model) of [Gaussian distributions](https://en.wikipedia.org/wiki/Gaussian_distribution) via an iterative refinement approach employed by both k-means and Gaussian mixture modeling. They both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the [k-nearest neighbor classifier](https://en.wikipedia.org/wiki/K-nearest_neighbor), a popular [machine learning](https://en.wikipedia.org/wiki/Machine_learning) technique for classification that is often confused with k-means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by k-means classifies new data into the existing clusters. This is known as [nearest centroid classifier](https://en.wikipedia.org/wiki/Nearest_centroid_classifier) or [Rocchio algorithm](https://en.wikipedia.org/wiki/Rocchio_algorithm" \o "Rocchio algorithm).

**Results:**

We obtain a result map, where the clusters are plotted. Further, we can also define a target cuisine or average spending, according to which a cluster can be recommended to the user. We can also derive inferences, as to which localities are related based on average restaurant budgets and cuisines.

An image of the output is as shown here:



**Discussion:**

We can recommend clusters and localities, based on the defined target cuisine or average spending, and the result list is present in the notebook with sample example, for which results for these recommendations have been derived. We can also derive inferences, as to which localities are related based on average restaurant budgets and cuisines.

**Conclusion:**

We hereby conclude, that our model identifies similar localities/neighbourhoods based on the aforementioned characteristics/parameters. We can also use the analysis to recommend localities for a new or an upcoming restaurant business.