Capstone Project - The Battle of Neighborhoods (Week 1)

Purpose: This document is designed to server two main purposes.
1) Describe the problem that I am going to solve
2) Identify the sources from where rich data can be downloaded.

1) **Problem Description:**

Identify the type of restaurant to open in Edison, NJ which is highly populated with multi cultured population. To get a good start for any new business, it is very important to identify right business type and right location, like a well-known proverb said in photography world – Right time, right place.

Here I am going to explore what type of restaurant should be good for city of Edison in New Jersey and identify the location to get maximum benefits in respect to easy transportation, customers, raw material supply and labor.

I'll also explore the population demographics of Edison city and compare with existing restaurants to identify how many restaurants per thousand population are available and how high will be the chances of success for a new restaurant.

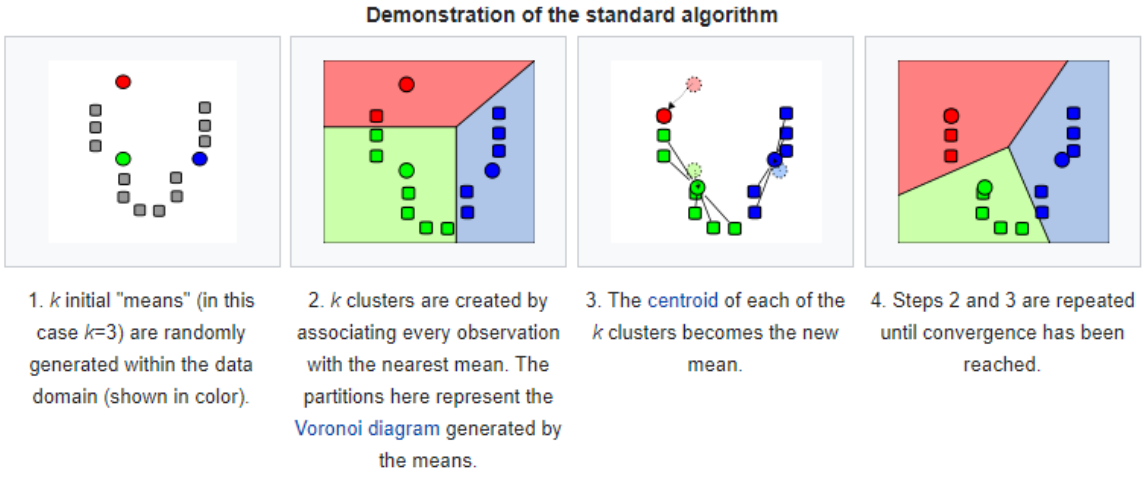
To recommend from my research and reach to a conclusion, I'll be heavily using the K-Means model of Data science. k-means clustering is a method of vector quantization, originally from signal processing, that

aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. It is popular for cluster analysis in 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: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

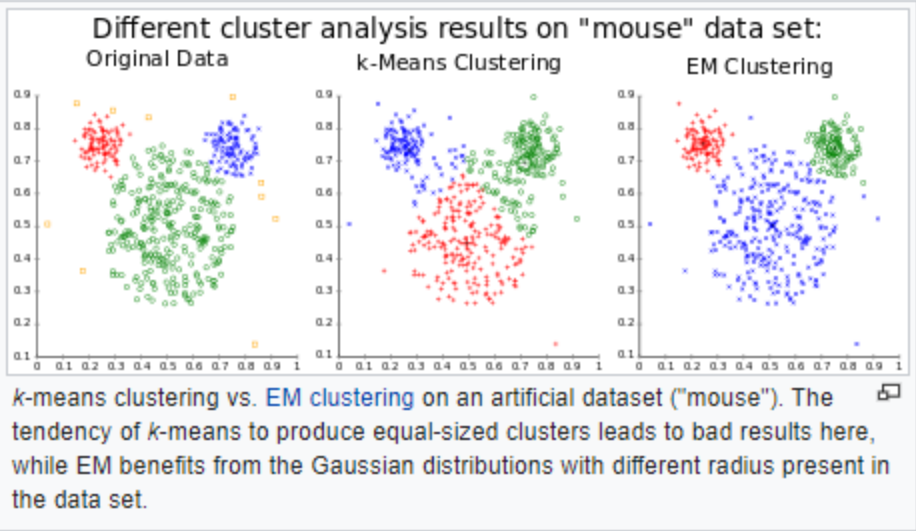
How K-Means works

Below is the step by step explanation on how the clustering is formed using this model, where k stands for number of clusters you want to create and how it propagates towards the centroid of a cluster.



The outcome of K-Means model

Below screenshot shows how the cluster will look once the data is feed in and after completion of the model.



k-means clustering vs. EM clustering on an artificial dataset ("mouse"). The tendency of *k*-means to produce equal-sized clusters leads to bad results here, while EM benefits from the Gaussian distributions with different radius present in the data set.

2) **Identify the Data source:** Getting the data is the most important part of the project because the whole prediction and recommendation is based on the available data on which whole model is designed. So while choosing the data, I have below major concerns
    a. Data should be from very reliable and from authentic source
    b. Data should be latest
    c. Data should be complete
    d. Data should be readily available on demand.
    e. Data should be rich with many attributes so that it can be sliced and diced as per the needs.

    With all above considerations, end up having two different set of data sources, Foursquare for restaurants related data and New Jersey government Data center site that works in partnership with US Census bureau, to download demographic profile of Edison city. The link of both sites are given as below

1) **Foursquare** – Data is available online on demand through the API. I need to register to this site to get the live data. Site link is https://foursquare.com/developers/apps
2) **NJ govt site** – Data is available online in Excel sheet format. I have to download the Census data from this site and use in my project. The site URL is https://www.nj.gov/labor/lpa/census/2kcensus/mid_ndx.html