Capstone Project - The Battle of Neighborhoods (Week 2)

Purpose: This document is designed to server below purposes.
1) Describe the problem that I am going to solve
2) Identify the sources from where rich data can be downloaded.
3) Methodology: It describes the exploratory data analysis and statistical testing.
4) Results
5) Discussion
6) Conclusion

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.

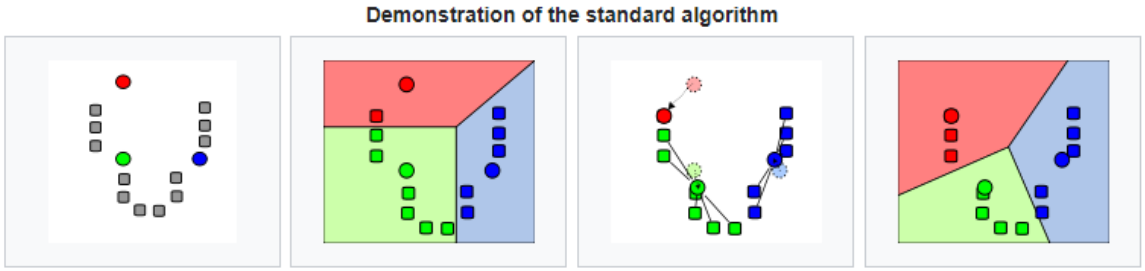
I'll also explore the population demographics of Edison city and compare with existing restaurants to identify how many restaurants per thousand population are available and how high will be the chances of success for a new restaurant.

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.



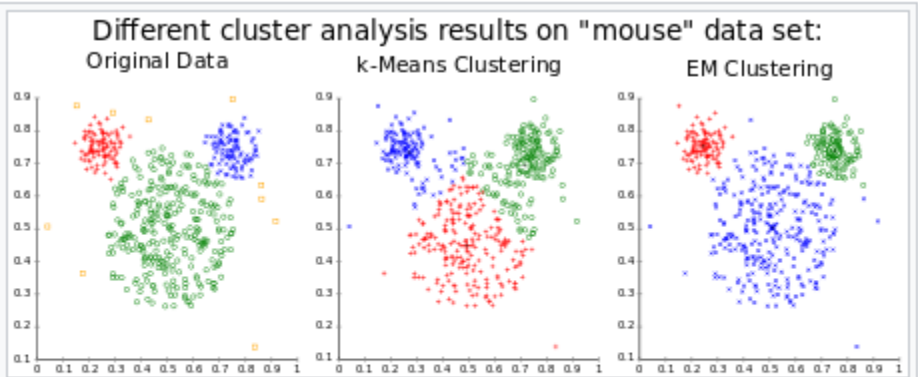
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

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.

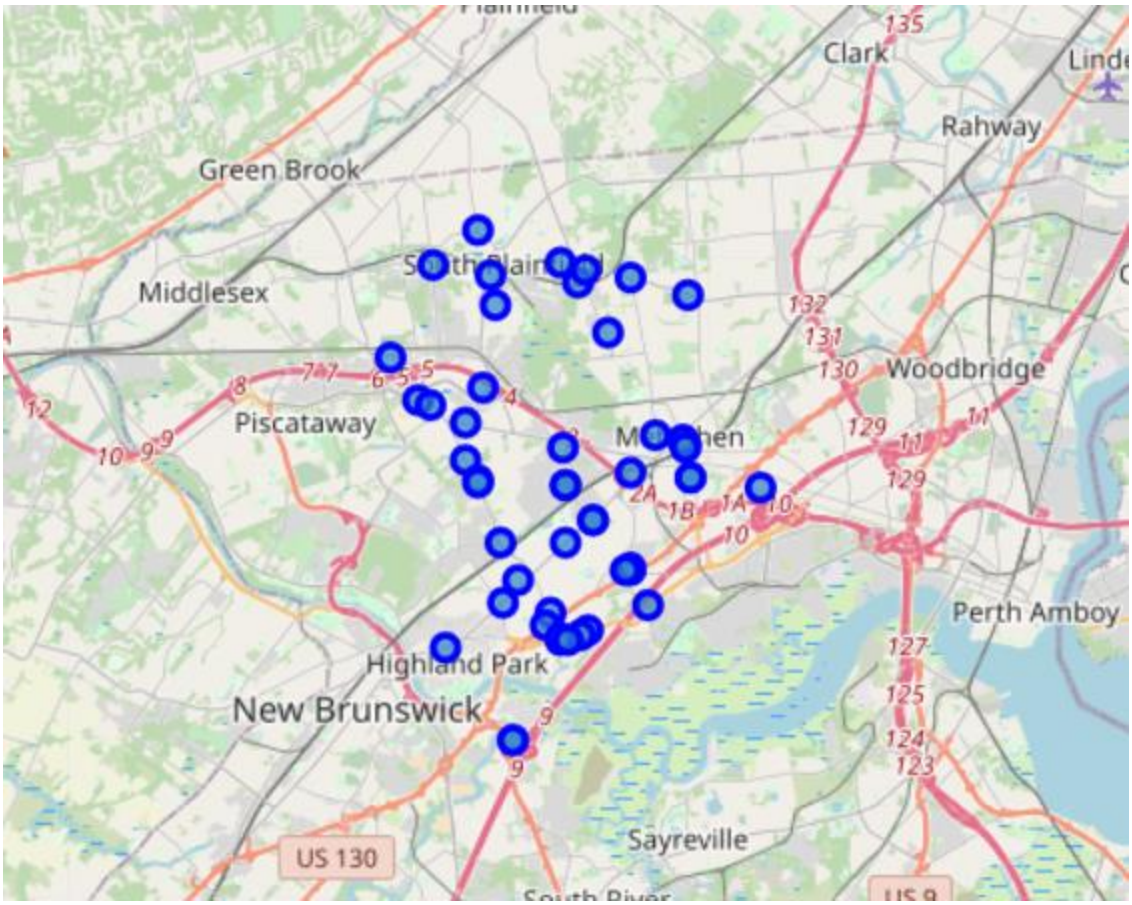
With all above considerations, end up having two different set of data sources, Foursquare for restaurants related data and New Jersey government Data center site that works in partnership with US Census bureau, to download demographic profile of Edison city. The link of both sites are given as below

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

3) **Methodology:** My research is on identify and suggest a restaurant type that can be more profitable in Edison based on existing demography of the city.

To achieve this, I need the demography data besides the restaurants data. Hence, I retrieved restaurant data from Foursquare site and studied under the map folium. This data was not enough for my research as I have to recommend a restaurant based on the demographic of Edison city. So, I also downloaded the Excel file with all demographic details from www.nj.gov site.

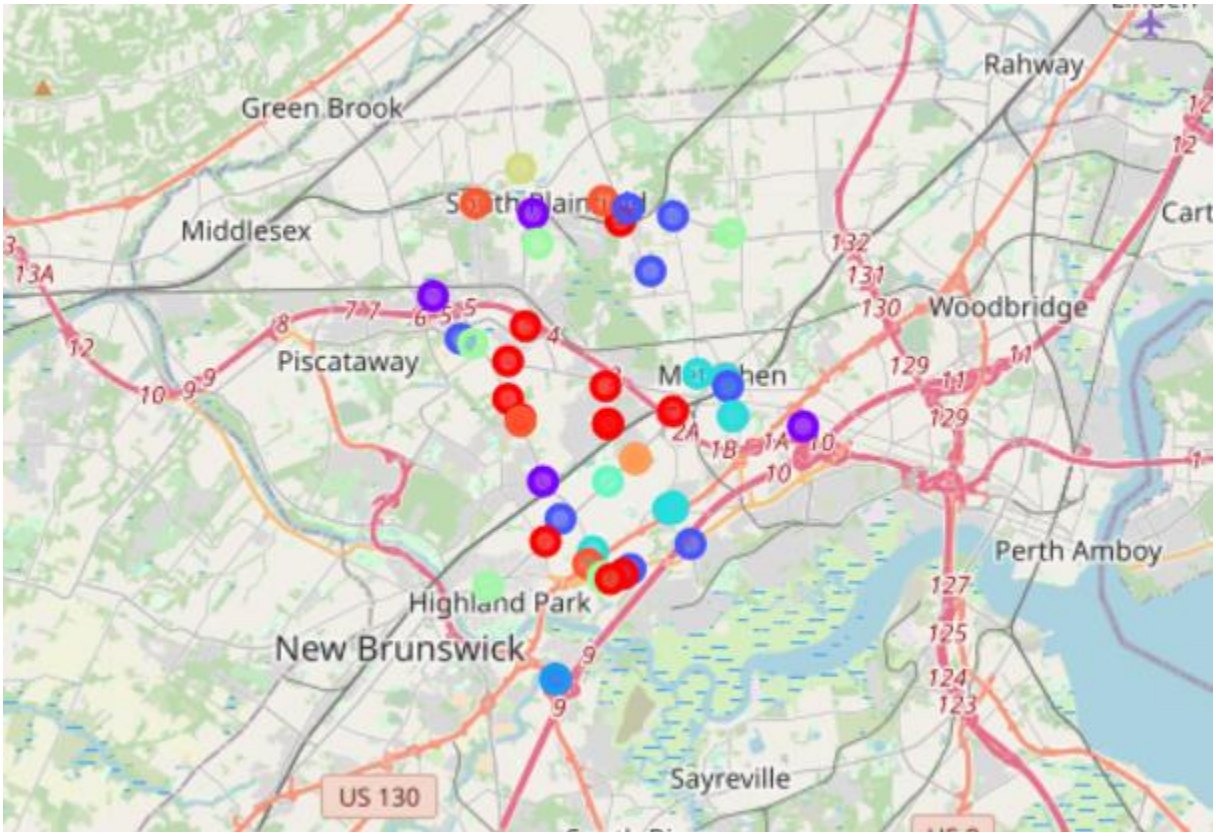
Restaurants in Edison can be visualized in below folium.



Next task was to identify the categories for which I used the group by clause to my data frame and derived below rsults.

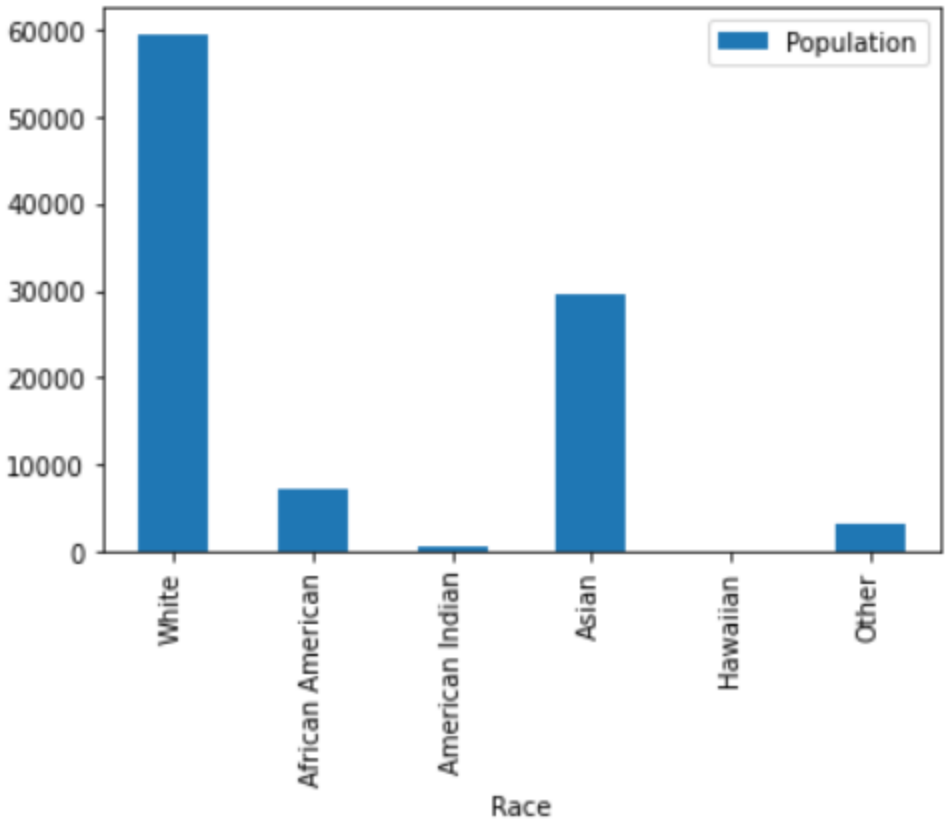
	categories	lat	lng	distance
0	Asian Restaurant	40.544740	-74.392365	3744.00
1	Brazilian Restaurant	40.586483	-74.418775	5749.00
2	Breakfast Spot	40.479919	-74.409005	6606.00
3	Business Service	40.553336	-74.417479	2568.00
4	Chinese Restaurant	40.539805	-74.394199	3336.25

Below visualization is the concentration of specific cuisine type in Edison and nearby area.



Next is to study the demography of Edison city. For that, I plot a car chart of population of Edison by Race. See below chart for more details.

: <AxesSubplot:xlabel='Race'>



4) **Results and Conclusion**: It has been identified that Edison is mostly populated by White Americans (60%) and Asians (29%) comprising to 90% of the population. Also, it has been observed that Indian restaurants are less comparative to American diners as seen from the above map.
   This makes Edison an ideal spot for running an Indian restaurant.

5) **Discussion and Observation**: While the Indian restaurant is most suitable in Edison and is highly recommended based on the collected data from Foursquare and based on the above model designed, I also observed that there's only one restaurant available for breakfast in whole Edison city. This gives an

alternative opportunity to have a spot for breakfast restaurant in Edison.

	categories	lng	labeledLatLngs	distance	postalCode	cc	city	state	country	formattedAddress
22	Breakfast Spot	-74.409005	[{'label': 'display', 'lat': 40.47991872670491...	6606	08816	US	East Brunswick	NJ	United States	[2 Tower Center Blvd, East Brunswick, NJ 08816...