Capstone Project Report – The Battle of Neighborhoods

Title: Recommending Ad Placement Sites for a Local Business

Introduction/Business Problem

A Dessert & Coffee Shop, located in a northern Suburb of Atlanta, GA wants to place three prominent Advertisement boards, within a few kilometers of their business. The goal is to place the ad boards at the center of (3) areas, which have the highest density of eateries/restaurants. The distance from Ad site to the client's shop should be relatively short, because the intention is to generate traffic directly from those areas, and most people won't consider driving too far on an impulse. The reasoning for such ad placement is that it has the potential to capture the attention of two key groups:

1. Dining guests who are having a meal in those areas, but who may be enticed to go a short distance for desserts &/or coffee afterwards at the client's Dessert Shop.
2. People who are in those areas because they are looking for a place to eat, but they are undecided. In that case, an attractive advertisement may persuade them to visit the nearby Dessert Shop.

So, the goal is to find central points of the three highest density eatery areas, within the local area of the Dessert Shop.

In addition to the client for this project, a similar methodology could produce useful results for other businesses, who want to determine the most advantageous locations for placing advertisements.

Data

We will need Foursquare location data for all eatery/restaurant venues within the target radius of the client's business. This may require exploring all venue categories at first, to make sure we are including all of those that are "eatery/restaurant" type. Each record will need to include the applicable business location, for the purpose of mapping, clustering, and finding the highest density areas. Once we get the correct dataset from Foursquare, including only the venues of selected categories, and within the specified radius, we need a DataFrame with VenueID, such as the Business Name, plus the Geo coordinates of each venue.

The dataset will therefore look like:

VenueName	Latitude	Longitude
Haru Ichiban	33.956893	-84.136399
Choong Man Chicken	33.953473	-84.142153
.	.	.
.	.	.

This is just an example, and VenueName could end up being either a unique ID or the place names, depending on whether duplicates are an issue.

Methodology

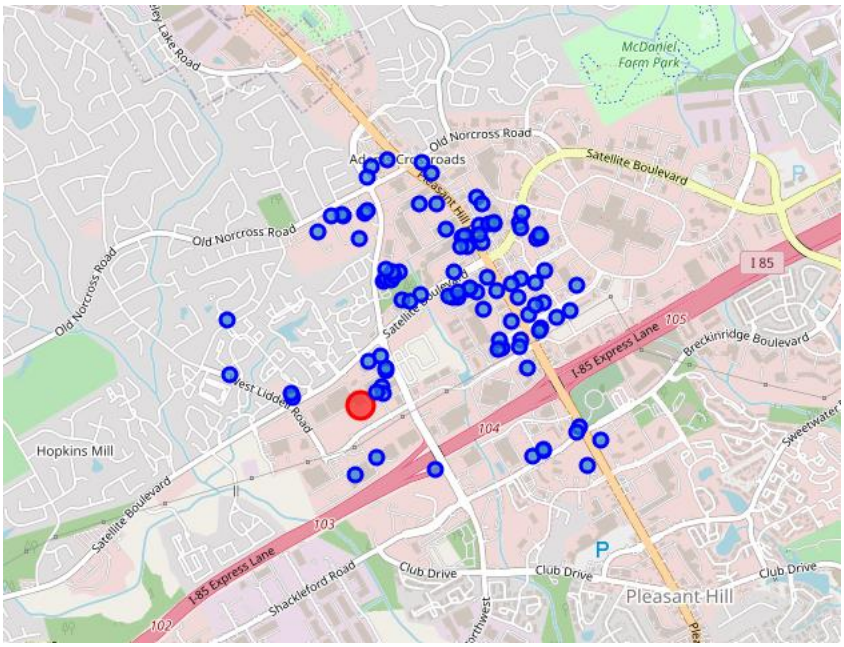
The steps are fairly straightforward, as follows:

1. Retrieve venue data for the target geographic area around the client's Dessert Shop. For this, I will use a Get request from Foursquare, with the 'Explore' endpoint. I will also use 'query=food', to get only restaurant type venues.
2. Load relevant parts of the venue data into a Pandas DataFrame. For this, I will use JSON_Normalize, then create a dataframe consisting of only the required fields.
3. Create a visualization of the data, using Folium maps.
4. Cluster the restaurant venues using K-Means: In this step, I will try different values for 'K', to isolate a result that achieves maximum exposure to restaurant customer foot traffic from (3) Ad Board placements. In other words, we want to create clusters such that the constituents of the (3) Ad Board Clusters maximizes [% of total restaurant venues that are either exposed to Ad Boards, or have direct exposure to the client's shop]. At the same time, we need to consider cluster density. Lower values of 'K' will result in a high percentage of coverage, but the clusters will not be dense enough to provide confidence of Ad exposure within the cluster area. Higher values of 'K' will result in smaller clusters, with the top three having higher density, but beyond a certain K-value, the coverage percentage of the top three will become sub-optimal. In the latter case, all three Ad Boards might end up being in close proximity to one another, due to a singularly dense area. We want to find the value of 'K' that results in good cluster density, plus high percentage of total Ad coverage.
5. Show the resulting cluster groups by color-coding their markers on our Folium map. Also, we will look at the numerical distribution, by examining cluster size.
6. Determine the three dominant (largest and densest) clusters of restaurant venues in the local area.
7. Use the centroids of the three dominant adjacent clusters, as recommended locations for the Ad Board placements.

Results

Foursquare results returned (96) restaurant venues within a 1.5-kilometer radius of the client's location.

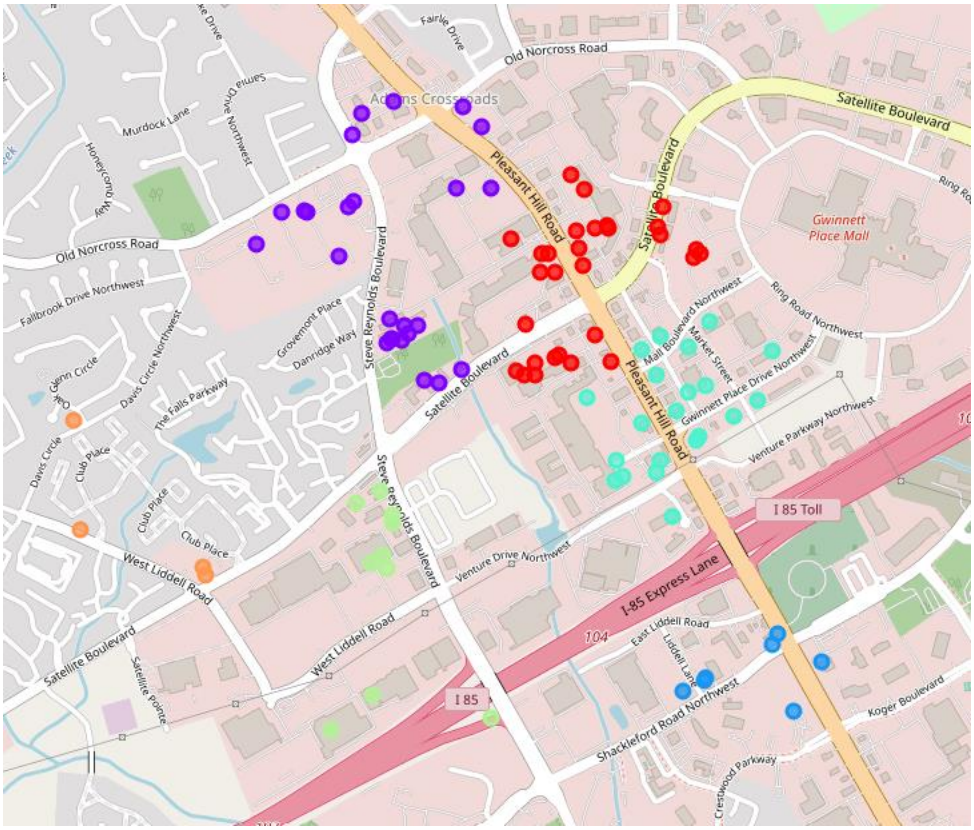
When placed on the Folium map, the restaurant venues seem to have a few high-density areas, as shown below:



** The Red marker indicates the client's location.

Several passes were made at Clustering with K-Means, using K values ranging from 4 to 8. Values below 6 showed only two relatively dense clusters, and other more dispersed clusters. Values above 6 began to show a tight grouping of more

than three clusters, with a couple of outlier clusters. Thus, K=6 proved to be the optimal parameter for clustering this group of restaurants by geographic proximity. The final result is shown below:



Numerically, this is how the distribution of clusters looks:

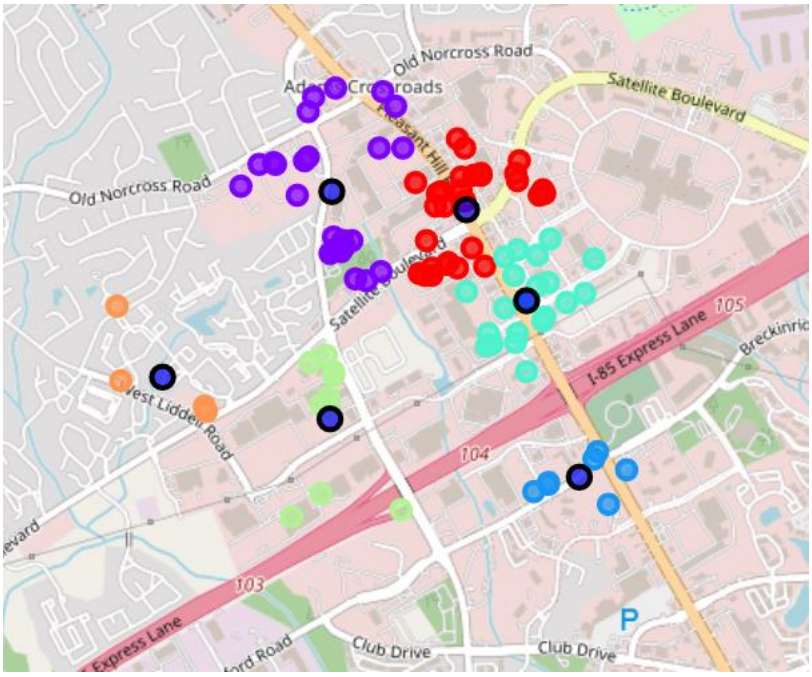
```
ad_venues.groupby('Clusters').count()
```

Out[17]:

Clusters	VenueName	Latitude	Longitude
0	29	29	29
1	25	25	25
2	7	7	7
3	21	21	21
4	10	10	10
5	4	4	4

Cluster #4 (Lime Green in the Folium map) is the group where the client's Dessert Shop is located.

With centroid markers added to the map, we can see the recommended Ad placement locations in the Red(0), Purple(1) and Aqua(3) clusters, which are the three dominant clusters, as shown below:



We will use the centroids of the three dominant adjacent clusters as the recommended locations for our client's three Ad Boards. We can see that those clusters represent 75 out of 96 restaurant venues in the local area. The client's Dessert Shop is also located in a cluster area that has 10 restaurants. Therefore, after placing their Ad Boards in the (3) recommended locations, they will have exposure to 88.5% (85/96) of restaurant customer foot traffic in their local area. This should be a satisfactory outcome for the client.

Clusters	Latitude	Longitude
0	33.959447	-84.134741
1	33.960172	-84.141112
2	33.948976	-84.129431
3	33.955902	-84.131912
4	33.951253	-84.141207
5	33.952928	-84.149090

So, we have three reference locations for Ad Board placements:

Centroid of Cluster 0: [33.959447, -84.134741] "2170 Pleasant Hill Rd"

Centroid of Cluster 1: [33.960172, -84.141112] "3093 Steve Reynolds Blvd"

Centroid of Cluster 3: [33.955902, -84.131912] "3550 Gwinnett Place Dr NW"

Discussion

This approach produced a very reasonable result. It seems to be a useful approach for similar business problems related to optimizing ad placement locations.

The data collection, cleaning and formatting steps were fairly straightforward and trouble-free.

When it came to K-Means clustering, finding the best result required multiple passes, using incremental changes to the K value. I think this is an unavoidable aspect of the approach. If the geographic area and distribution of target businesses had been more complex, the clustering step might have been a challenge. In this case, it was relatively simple to try K values starting from 4, and working higher until passing the peak at K=6.

With a more complex scenario, the K value optimization should probably be coded as a loop through a range, with outcomes measured and recorded at each iteration.

Conclusion

In this study, I retrieved location data from Foursquare for all venues of a particular type (restaurants) within a given radius of a local Dessert Shop. The business goal was to identify the most advantageous locations for three advertisement boards, by placing them in areas with the most restaurants. I analyzed the geographic distribution of the restaurant venues using K-Means clustering, to find the three densest areas within the overall radius. This met the stated goal, and provided a solution that would give the client exposure to 88.5% of all restaurant foot traffic in the target region.

This turned out to be a useful technique for locating optimal ad placement locations. In the future, this approach should work well for other businesses with similar needs.