Capstone: Find the best neighborhood in Da Nang to open a Coffe Shop
Introduction:

To give effect to this program, I wanted to see what would be the best location to open a coffee shop in Da Nag.

Example Company:

https://www.foody.vn/da-nang/doll-coffee

Problem Description:

Which neighborhood should be open a new Coffe Shop in Da Nang to guarantee

a sustainable business over time?

Requirements:
1. Store needs to be strategically located inside the biggest concentration of restaurants in Da Nang area.
2. Confirm any assumption by means of modeling and testing the data. Specifically, visually cluster common restaurants in Da Nang by neighborhood.
3. Additionally determine that a good number people can frequent these restaurants with sustainable frequency inside these neighborhoods.
    a.) Is the neighborhood populous?
    b.) Is the neighborhood average salary close to the Canadian National Average?

I wants to be able to judge which neighborhoods also may be poised to grow in Coffe Shops numbers in coming years.

Locating the Coffe Shop according to these requirements will ensure the following:
- lowest cost for delivery
- shortest travel time to his store for his clients
- overall lower run costs
- increase in overall business
- overall greater customer satisfaction

Data:
You can follow along in my Capstone Notebook located here: https://github.com/alemarchan/Coursera_Capstone/blob/main/capstone_project.ipynb

Data Wrangling

A lot of hard work went into creating the working data set. I had to combine the following disparate data sources. The order of events went like this

1. Load all the Data from all the various sources.

1.1 Da Nang information

https://en.wikipedia.org/wiki/Da_Nang

1.1.1 Load Da Nang geospatial coordinates and merge to Da Nang Postal Code Data

http://cocl.us/Geospatial_data

Next, I joined geo spatial to the Da Nang Data.

1.2 Da Nang neighborhoods populations broken down by postal code

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/File.cfm?T=1201&SR=1&RPP=9999&PR=0&CMA=0&CSD=0&S=22&O=A&Lang=Eng&OFT=CSV

Use Pandas to grab the csv

1.2.1 Merge Da Nang Neighbourhood populations

https://www.unescap.org/sites/default/d8files/2020-08/DaNangCity_0.pdf

Next, I joined population data to the Da Nang Data.

1.3 Da Nang average after tax income

Here we must manually download these from Stats Canada and load them.

https://www.payscale.com/research/VN/Location=Da-Nang/Salary

1.3.1 Merge Da Nang Neighbourhood income data with Da Nang populations

Next, I joined income data to the Da Nang Data.

At this time I also saved a copy of the data set as my friend had asked for it in his list of requirements.

1.4 What is the Canadian National Average After Tax Income

Here I must also manually download this from Stats Vietnam and load them.

https://www.unescap.org/sites/default/d8files/2020-08/DaNangCity_0.pdf

1.5 Da Nang list of Restaurants or Venues that could potentially use Restaurant Equipment

4SQUARE API

https://api.foursquare.com

1.5.1 Get all the Venues in Da Nang.

1.5.2 Only add Restaurants as Venue Categories

Use this list to Extract Restaurants and only include Coffe Shops in our Data Set.

1.5.3 OneHot encode and count Coffe Shops

Prepare the data for clustering

Combine all of those into a working Data Set to cluster and geo spatial map of the results showing the best neighborhood to open a Coffe Shop

Combining all of these disparate data sets will clearly demonstrate the following:

- which neighborhoods in Da Nang have clusters of like Restaurants
- how populated each neighborhoods is
- the average after tax income is all of these neighborhoods
- which neighborhood should he target to open his Coffe Shop.
-

Methodology:
Choice of Algorithms
I chose K-Means Clustering.

https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac
A backgrounder on K-Means clustering

"K-means clustering is an iterative clustering algorithm where the number of clusters K is predetermined and the algorithm iteratively assigns each data point to one of the K clusters based on the feature similarity."
Key Observation: And for my project feature similarity means Coffe Shops similarity in Neighborhoods
Choosing the correct number of clusters.
https://www.jeremyjordan.me/grouping-data-points-with-k-means-clustering/
Here I use Silhouette analysis to determine the optimum number of clusters to use.

A backgrounder on Silhouette analysis.
"We can use Silhouette analysis to evaluate each model. A Silhouette coefficient is calculated for observation, which is then averaged to determine the Silhouette score.
The coefficient combines the average within-cluster distance with average nearest-cluster distance to assign a value between -1 and 1. A value below zero denotes that the observation is probably in the wrong cluster and a value closer to 1 denotes that the observation is a great fit for the cluster and clearly separated from other clusters. This coefficient essentially measures how close an observation is to neighboring clusters, where it is desirable to be the maximum distance possible from neighboring clusters. We can automatically determine the best number of clusters, k, by selecting the model which yields the highest Silhouette score."
Key Observation: My highest score was 2.
2.1 Run K means and segment data into clusters and generate labels
2.2 Merge the Da Nang data with geo coordinates data and make sure it's the right shape
Here I reshape the Da Nang data so that it's shape matches the clustered data.
2.3 Add the KMeans Labels
Determine the largest cluster in this case it was cluster number 2 with a shape of (76, 15)
3. Cluster 2 Contains the highest cluster density. We need to find the geographic centroid for this cluster. This is the optimum location for a new Coffe Shop.
Here we take the average latitude and longitude to be the centroid.
3.1 Install opencage to reverse lookup the coordinates

Opencage allows me to reverse lookup the geo coordinates.
Key Observation: This is the optimum location for a new Coffe Shop.
Results:
4.1 Plot the clusters on a Map of the Da Nang and Super Impose the best location of a Store

4.2 Exact Address of desired Location
Based on a reverse Lookup

Discussion:
5.1 Explaining the results
As we built our list of neighborhoods with Coffe Shops venues exclusively we discovered most neighborhoods were similar and the greatest concentration of Coffe Shops was in Central Da Nang and downtown Da Nang. This might seem obvious but it would also appear that these are some of the most affluent neighborhoods in Da Nang so there appears to be correlation. By Locating in the general vicinity of the Exact location my friend could be geographically centered in this cluster and poised to service his restaurant customer base with greatest efficiency.

When we built our our K-Means dataset we used Silhouette analysis to tell us there was a lot of similarity between neighborhoods and the most common restaurants contained with in. Really there was only 2 types of cluster or neighborhoods in greater Da Nang. The vast majority of those were in 1 cluster. So Da Nang restaurants might be many but they are very homogeneously located near the center of Da Nang.

We decided to keep all neighborhoods in the dataset regardless of income of population as the majority were close enough.

Conclusion:

Based on the study carried out and the calculations provided, the potential of the cafeteria as a business is evident. Much more inference can be obtained with more work. A potential side business for my friend might be assisting new restaurant owners where they might locate a new restaurant, who their competition is and who their clientele might be.