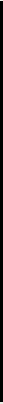



AN OPTIMAL LOCATION TO OPERATE A BAKERY IN SAN FRANCISCO



By Edison Li
September, 2020



Problem statement

- Mr Li operates a bakery in San Francisco that is famous for making the best sourdough bread in the area.
Currently, his business is providing fresh sourdough to 30 American restaurants in San Francisco downtown area.
- Mr Li is planning to move his bakery to a new location. He wants to find **a place that is right in the center all his current customers**. This would keep the total delivery time to his customers to a minimal
- Mr Li also has plans to expand his customer base to **90** in the next two years. Since he already knows the locations of the 60 potential customers, he wants to determine **the optimal number of kitchens** to meet the extra demand and **the approximate locations of these kitchens**.

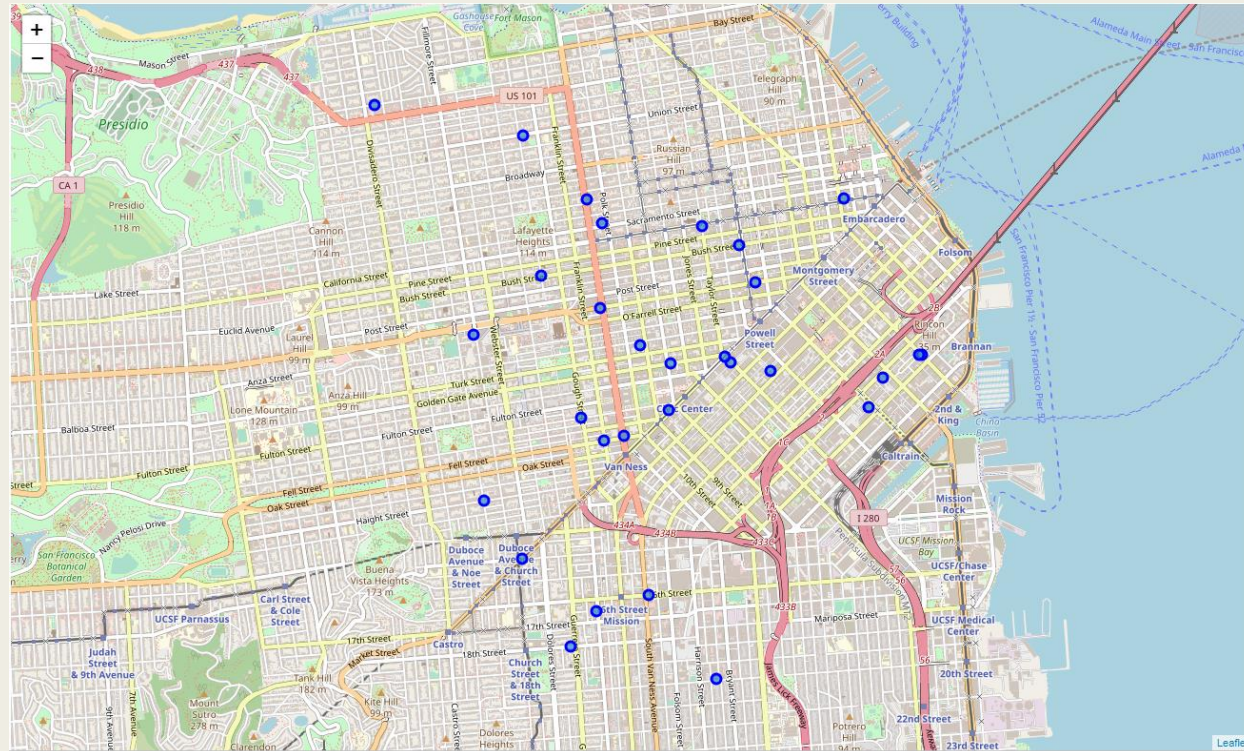
Data acquisition

- We are using foursquare API to get 90 restaurants around downtown San Francisco.
- In order to refine our search to American restaurants, we have modified our API call to include category id = “4bf58dd8d48988d14e941735”.
- we put the customer name, venue category, longitude and latitude into a dataframe.
- The data is complete with no missing data.

	Customer	Latitude	Longitude
0	The Progress	37.783745	-122.432972
1	Dottie's True Blue Cafe	37.781748	-122.409829
2	Box Kitchen	37.781158	-122.406243
3	Octavia	37.787935	-122.426934
4	Morty's Delicatessen	37.781710	-122.415243

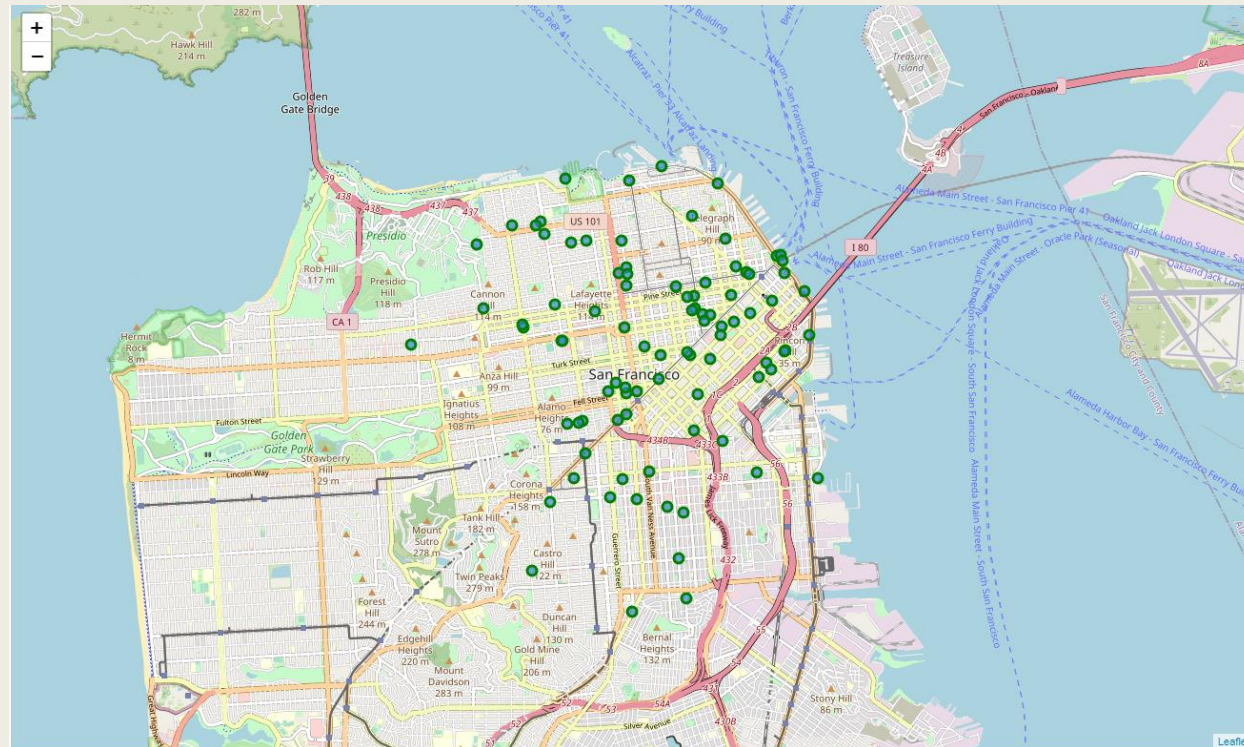
Data Cleaning

- We have assigned the first 30 venues as Mr Li's existing customer (as shown below)



Data Cleaning 2

- The other 60 venues are used as Mr Li's potential customers (as shown below)

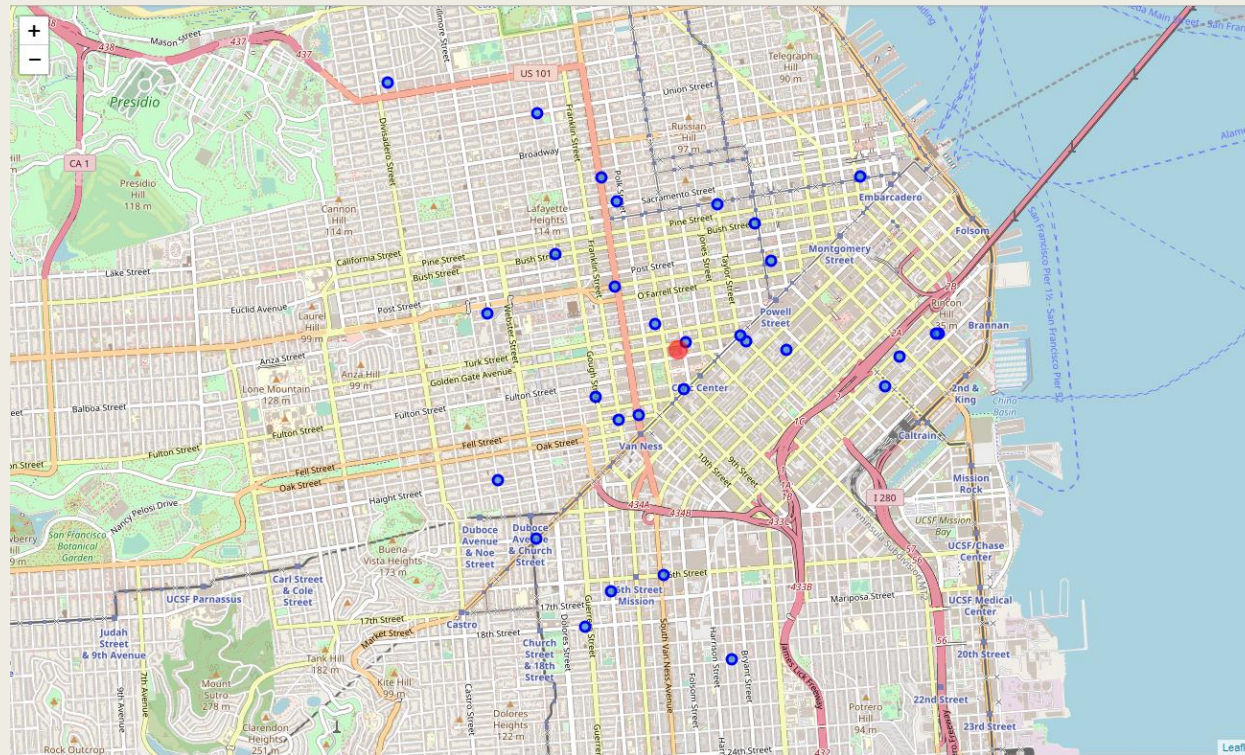


Problem 1: Clustering with 1 centroid

In order to find the most optimal location for Mr Li's bakery to serve his existing customers, we are doing a kmean clustering with a single centroid. That would give us the location that has the minimal mean distance to every other location in the cluster. In other word, the total travel distance to all 30 locations would be minimized.

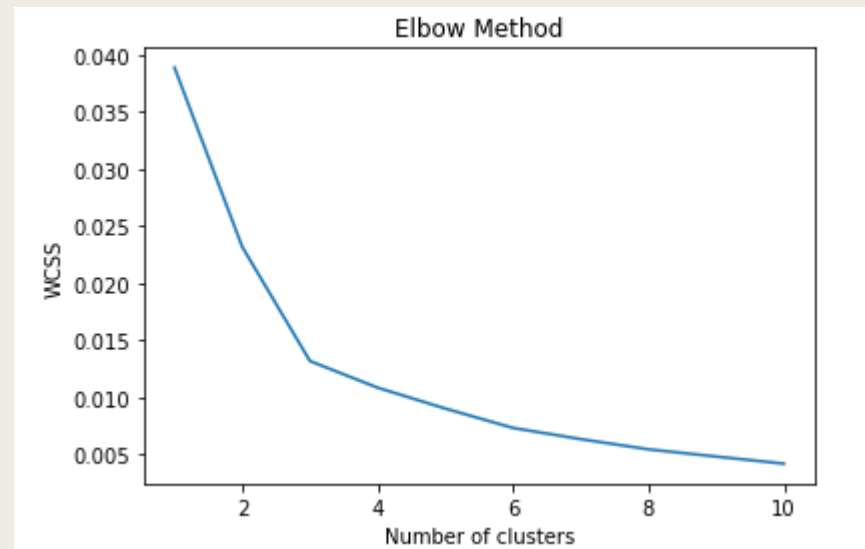
The resulting location is in this location (37.78118, -122.415941). A quick search in Google map has indicated that the address is **200 McAllister St, San Francisco, CA 94102**

Problem 1: Clustering with 1 centroid



Problem 2: Optimizing number of centroids

For the future list of customers, we have executed kmean clustering 10 times with increasing number of centroids. After each iteration, we would calculate the sum of squares of the distances of each data point in all clusters to their respective centroids (WCSS) and plot the result. Using the elbow method, we have found that **3 bakeries** would be the optimal number of location to serve 90 restaurants.



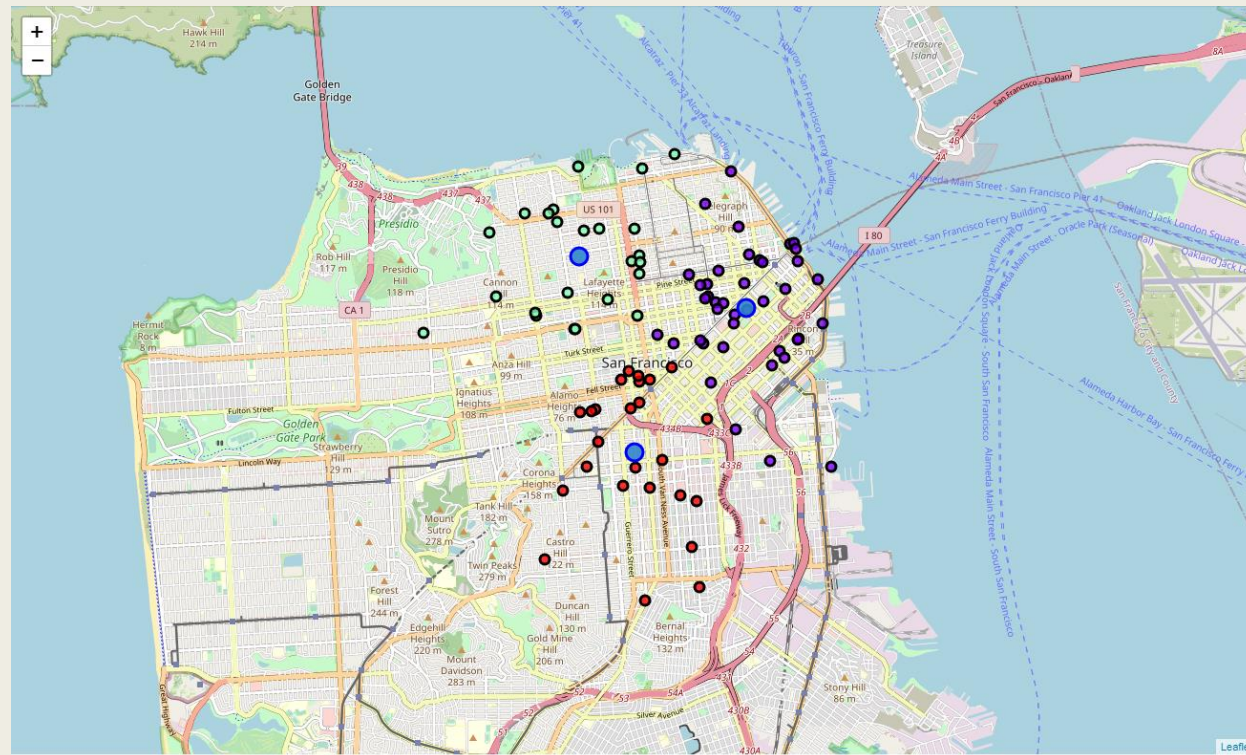
Problem 2: Optimizing number of centroids

By re-running kmean with 3 clusters, we have determined the locations are:

412 Valencia St, San Francisco, CA 94103 (37.766236, -122.422152)

680 Mission Street, San Francisco, CA 94105 (37.786743, -122.402110)

2200 Pacific Ave, San Francisco, CA 94115 (37.794098, -122.432219)



Recommendations

To serve his 30 existing customers, it would be best if Mr Li can move his bakery to **200 McAllister St, San Francisco, CA 94102 (37.78118, -122.415941)** right now.

If Mr Li can acquire 60 new customers in the next two years, it would be best for him to move open 3 new locations as follow:

412 Valencia St, San Francisco, CA 94103 (37.766236, -122.422152)

680 Mission Street, San Francisco, CA 94105 (37.786743, -122.402110)

2200 Pacific Ave, San Francisco, CA 94115 (37.794098, -122.432219)

Further analysis

Even though we are fairly satisfied with our recommendations, we could enhance our model by modifying how distance between two locations is calculated. In our current model, we are using Euclidian distance. In order to get more realistic clustering, we could evaluate distance by the actual travel time between two locations. This change will add real life elements like the traffic flow and traffic condition to our model. We would expect the recommendations would reflect the real-life conditions better.