Background

For the background, assume I have a client who is looking to open a restaurant in San Francisco and needs to find the best possible places for it. The purpose of this project is to find the best area with the potential to open a tea garden mixed with a Mediterranian restaurant (this idea was developed before the coronavirus outbreak, so please ignore the current market conditions for such establishments.) After acquiring the data for the neighborhood and clustering them by their venues, I will try to find a cluster where the restaurants and food courts are concentrated but do not have many Mediterranean restaurants and tea gardens. If such a cluster does not exist, I hope to find a cluster where the parks are concentrated.

Data Acquisition and Cleaning

The data will be acquired in three steps. These are:

1- Get the neighborhoods in San Francisco from Wikipedia, read and parse the data using beautifulsoup library and turn it into a data frame.

2- Get the latitude and longitude information for these neighborhoods using the geopy library

3- Get the top 10 venues for the neighborhood using the Foursquare API

The above raw data will be turned into valuable information in two steps. These are:

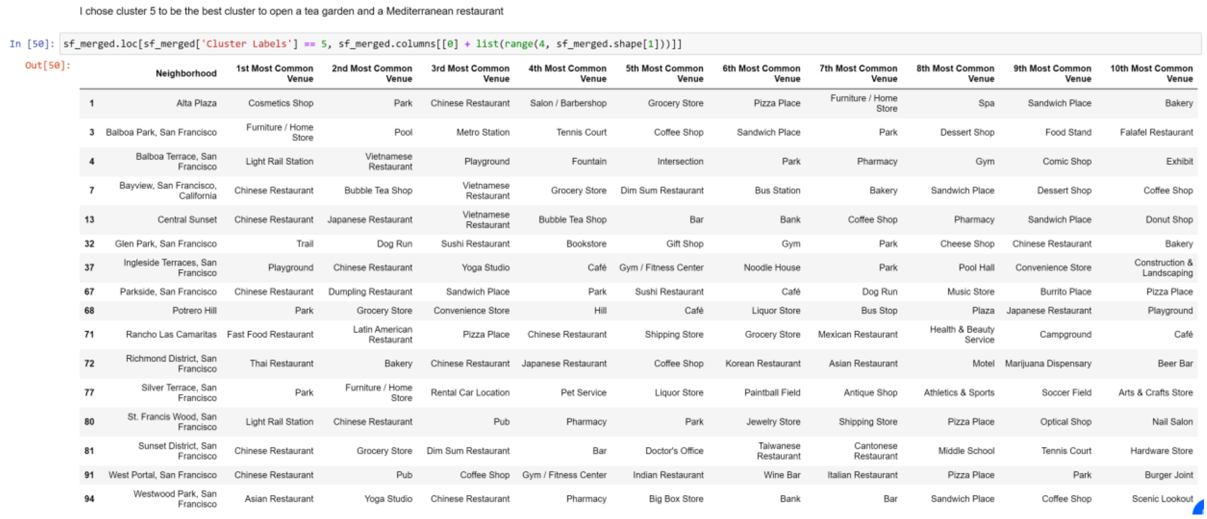
1- After the data is acquired and cleaned using the scikit-learn library to cluster the data.

2- Use the folium library to create a visualization of the San Francisco and its neighborhood clusters.

Methodology

From the analysis I have performed I have found 95 neighborhoods in San Francisco. By creating a visual in folium to se the distribution of the neighborhoods I expected to see 8 to 13 clusters. After creating multiple clusters from 5 to 15 I have decided the best number of clusters I can use to be 10. I have used k-means clustering algorithm because the data I had was best fitted for this particular algorithm. The data is unlabeled, so a clustering algorithm was used. Sci-kit learn library was used because of the simplicity of the project and the library is very easy to use compared to libraries like TensorFlow or PyTorch.

Results



I chose cluster 5 to be the best cluster to open a tea garden and a Mediterranean restaurant

```
In [50]: sf_merged.loc[sf_merged['Cluster Labels'] == 5, sf_merged.columns[[0] + list(range(4, sf_merged.shape[1]))]]
```

Out[50]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Alta Plaza	Cosmetics Shop	Park	Chinese Restaurant	Salon / Barbershop	Grocery Store	Pizza Place	Furniture / Home Store	Spa	Sandwich Place	Bakery
3	Balboa Park, San Francisco	Furniture / Home Store	Pool	Metro Station	Tennis Court	Coffee Shop	Sandwich Place	Park	Dessert Shop	Food Stand	Falafel Restaurant
4	Balboa Terrace, San Francisco	Light Rail Station	Vietnamese Restaurant	Playground	Fountain	Intersection	Park	Pharmacy	Gym	Comic Shop	Exhibit
7	Bayview, San Francisco, California	Chinese Restaurant	Bubble Tea Shop	Vietnamese Restaurant	Grocery Store	Dim Sum Restaurant	Bus Station	Bakery	Sandwich Place	Dessert Shop	Coffee Shop
13	Central Sunset	Chinese Restaurant	Japanese Restaurant	Vietnamese Restaurant	Bubble Tea Shop	Bar	Bank	Coffee Shop	Pharmacy	Sandwich Place	Donut Shop
32	Glen Park, San Francisco	Trail	Dog Run	Sushi Restaurant	Bookstore	Gift Shop	Gym	Park	Cheese Shop	Chinese Restaurant	Bakery
37	Ingleside Terraces, San Francisco	Playground	Chinese Restaurant	Yoga Studio	Café	Gym / Fitness Center	Noodle House	Park	Pool Hall	Convenience Store	Construction & Landscaping
67	Parkside, San Francisco	Chinese Restaurant	Dumpling Restaurant	Sandwich Place	Park	Sushi Restaurant	Café	Dog Run	Music Store	Burrito Place	Pizza Place
68	Potrero Hill	Park	Grocery Store	Convenience Store	Hill	Café	Liquor Store	Bus Stop	Plaza	Japanese Restaurant	Playground
71	Rancho Las Camaritas	Fast Food Restaurant	Latin American Restaurant	Pizza Place	Chinese Restaurant	Shipping Store	Grocery Store	Mexican Restaurant	Health & Beauty Service	Campground	Café
72	Richmond District, San Francisco	Thai Restaurant	Bakery	Chinese Restaurant	Japanese Restaurant	Coffee Shop	Korean Restaurant	Asian Restaurant	Motel	Marijuana Dispensary	Beer Bar
77	Silver Terrace, San Francisco	Park	Furniture / Home Store	Rental Car Location	Pet Service	Liquor Store	Paintball Field	Antique Shop	Athletics & Sports	Soccer Field	Arts & Crafts Store
80	St. Francis Wood, San Francisco	Light Rail Station	Chinese Restaurant	Pub	Pharmacy	Park	Jewelry Store	Shipping Store	Pizza Place	Optical Shop	Nail Salon
81	Sunset District, San Francisco	Chinese Restaurant	Grocery Store	Dim Sum Restaurant	Bar	Doctor's Office	Taiwanese Restaurant	Cantonese Restaurant	Middle School	Tennis Court	Hardware Store
91	West Portal, San Francisco	Chinese Restaurant	Pub	Coffee Shop	Gym / Fitness Center	Indian Restaurant	Wine Bar	Italian Restaurant	Pizza Place	Park	Burger Joint
94	Westwood Park, San Francisco	Asian Restaurant	Yoga Studio	Chinese Restaurant	Pharmacy	Big Box Store	Bank	Bar	Sandwich Place	Coffee Shop	Scenic Lookout

After creating 10 clusters cluster 6 is the best place to open the tea garden and restaurant. This cluster was chosen because of the high concentration of the Asian restaurants. Since tea is an important part of the Asian culture it is believed that the traffic will be higher. The high concentration of the Asian restaurants also means that the Mediterranean restaurant will be stand-out from its competition. So the customers will be able to enjoy a nice fresh cup of tea that they are used to and have Mediterranean desserts as side that they are not used to. This will provide them a unique experience of tasting nostalgic drink with an unusual dessert.

Discussion

Although 10 clusters were used to be able to create distinct clusters, San Franciscos diverse culture made it nearly impossible to create trully distinct groups. Parkside, San Francisco was chosen to be the optimal location because of the lack of competition in the neighborhood for the tea shops. One very interesting observation is Forest Knolls, San Francisco. This is such a unique place that even when only two clusters are created the algotrihm clusters everything but the Forest Knolls together, so Forest Knolls is the most unique place in terms of the venues in San Francisco.

Conclusion

In conclusion clusters were created in the San Francisco area to find an optimal location for a Tea Garden and Mediterranean Restaurant. For future improvements more data such as the average price of land, how much land is being sold on the market, what is the traffic flow to the neighborhood and things such as these could be included in the data to make a more clear clustering to analyze where a restaurant should be opened.

Thank you very much for your time!