Opening Cappuccino Coffee Shop in San Francisco neighborhood Using Data Analysis Techniques.

Introduction.

In this project, we will use the Foursquare API to explore neighborhoods in San Francisco, get the most common venue categories in each neighborhood, use the *k*-means clustering algorithm to find similar neighborhoods, use the Folium library to visualize the neighborhoods in San Francisco. The venues in the neighborhood collected using Foursquare API and clustered (using K-mean) and analyzed based on the numbers of coffee shops in each cluster. K-mean cluster method was found for clusters. k-means is especially useful if you need to quickly discover insights from unlabeled data. Run k-means to cluster the neighborhood into 5 clusters. k-means will then partition our neighborhoods into 5 groups. The neighborhoods in each cluster are similar to each other in terms of the features included in the dataset. Based on the data analysis, the coffee shop can be open in the second cluster because the data analysis showed that numbers of coffee shops in the area are still small and can reduce the competition. However, it also found that the second cluster has the smallest number of area and have a potential smaller profit if the coffee shop is open in the first cluster.

Data used

We analyze the following page

http://www.healthysf.org/bdi/outcomes/zipmap.htm,

in order to obtain the data that is in the table of postal codes and to transform the data into a pandas data frame.

```
import requests # library to handle requests
from bs4 import BeautifulSoup
import pandas as pd
```

```
response = requests.get("http://www.healthysf.org/bdi/outcomes/zipmap.htm")
soup = BeautifulSoup(response.text, "lxml")
table = soup.find_all("table")
df = pd.read_html(str(table))
df = pd.DataFrame(df[4])
```

```
df.columns = df.iloc[0]
df = df.iloc[1:-1, :-1]
sf_data = df
sf_data.head()
```

	Zip Code	Neighborhood
1	94102	Hayes Valley/Tenderloin/North of Market
2	94103	South of Market
3	94107	Potrero Hill
4	94108	Chinatown
5	94109	Polk/Russian Hill (Nob Hill)

Convert Addresses into Latitude and Longitude

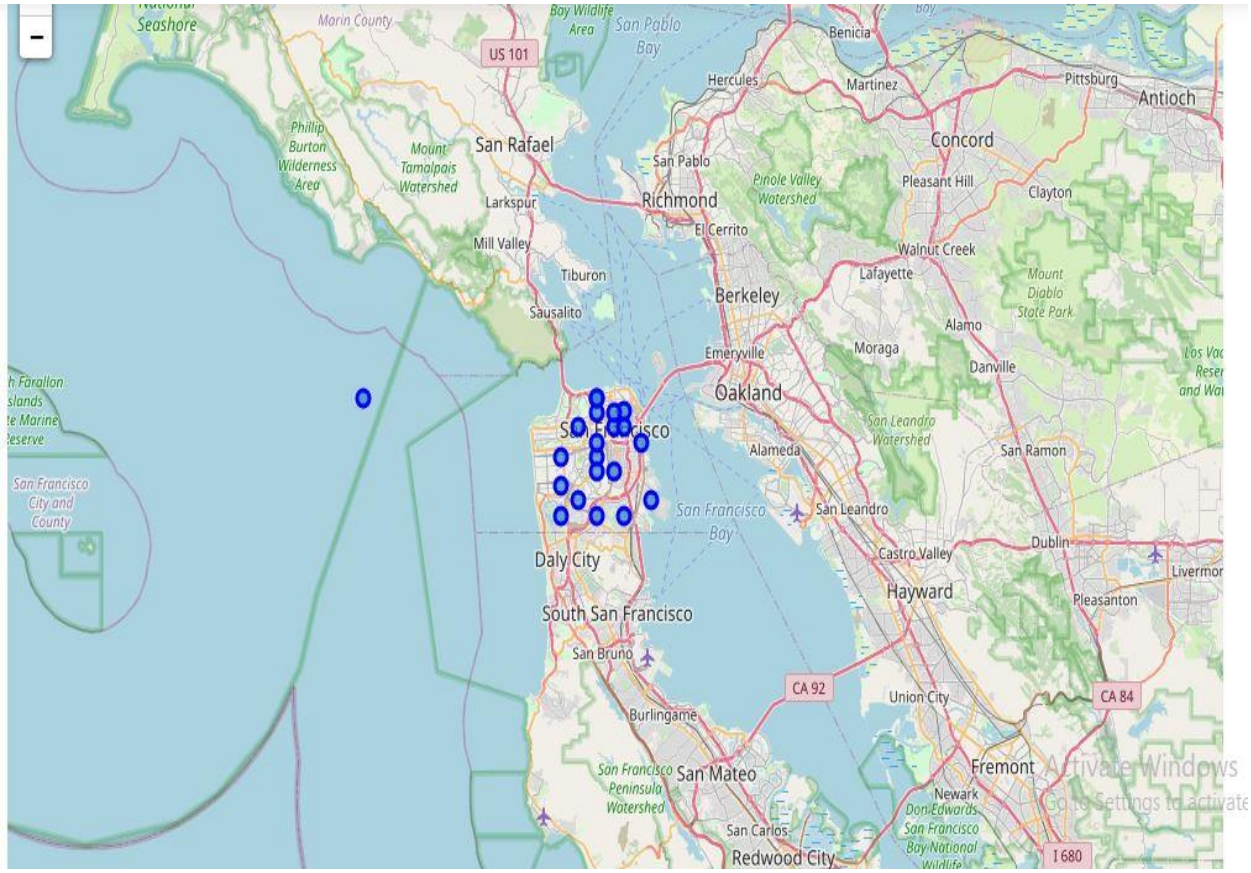
In order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighborhood

Explore Neighborhoods in San Francisco

Create a map of San Francisco with neighborhoods

Use geopy library to get the latitude and longitude values of San Francisco

We then use the folium library to plot the map. **folium** enables both the binding of data to a map as well as passing HTML visualizations as markers on the map.



Define Foursquare Credentials and Version

```
CLIENT_ID = 'PK4E3AX1HWYOAYAJXBCEN5FAIYYBI2YQMJTCM3DJTC0CUD2L' # your Foursquare ID
CLIENT_SECRET = 'GLWH10Z34GDBJB1J5T2UW5J0KSTXQIWBRLCFBD2LIM5LGGCA' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version
LIMIT = 100

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

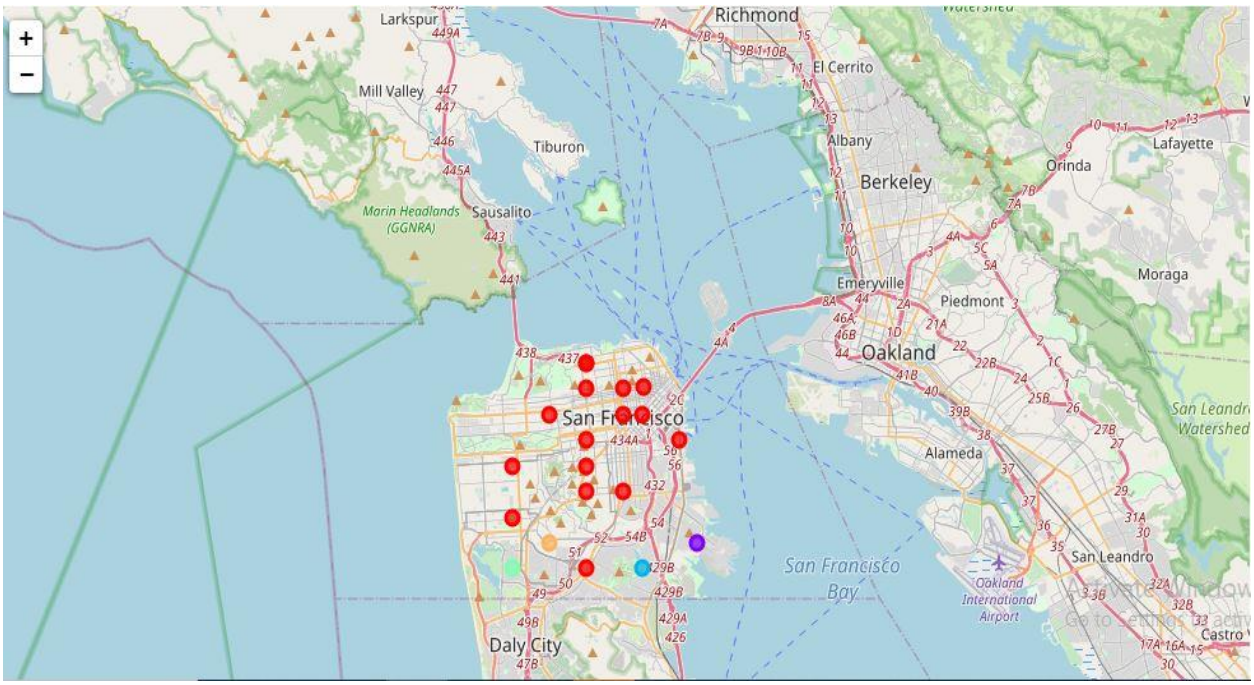
```
Your credentails:
CLIENT_ID: PK4E3AX1HWYOAYAJXBCEN5FAIYYBI2YQMJTCM3DJTC0CUD2L
CLIENT_SECRET:GLWH10Z34GDBJB1J5T2UW5J0KSTXQIWBRLCFBD2LIM5LGGCA
```

Analyze Each Neighborhood

Group rows by neighborhood and by taking the mean of the frequency of occurrence of each Venue Category.

Cluster Neighborhoods

k-means is especially useful if you need to quickly discover insights from unlabeled data. Run k-means to cluster the neighborhood into 5 clusters. k-means will then partition our neighborhoods into 5 groups. The neighborhoods in each cluster are similar to each other in terms of the features included in the dataset.



Examine Clusters

K-mean cluster method was found for clusters. k-means is especially useful if you need to quickly discover insights from unlabeled data. Run k-means to cluster the neighborhood into 5 clusters. k-means will then partition our neighborhoods into 5 groups. The neighborhoods in each cluster are similar to each other in terms of the features included in the dataset. Based on the data analysis, the coffee shop can be open in the second cluster because the data analysis showed that numbers of coffee shops in the area are still small and can reduce the competition. However, it also found that the second cluster has the smallest number of area and have a potential smaller profit if the coffee shop is open in the first cluster.