Capstone Project - The Battle of Neighborhoods

By Pruthvi Reddy

Table of Contents

Introduction

A friend of mine is relocating to San Francisco, CA. He currently lives in Philadelphia, PA in an apartment and uses public transportation to work. He goes to gym daily and loves eating out. He frequently visits parks and would like to live in an area which is similar to his life style. So we need to identity which location in San Francisco will best suit his current life style.

Objective

We will analyze Philadelphia and San Francisco area's by segmentation and clustering using Foursquare data. The aim of this project is to identify similar locations between San Francisco and Philadelphia, classify these areas based on accessibility, public transportation, type of restaurants etc.

Using machine learning methods like segmentation and clustering, we will identify similar neighborhoods in San Francisco based on the characteristics of my friends current Philadelphia neighborhood.

Data

Neighborhood and zip code data will be scrapped from Wikipedia pages and other sources–

https://en.wikipedia.org/wiki/List_of_Philadelphia_neighborhoods
https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco

We will utilize google geocoder or similar method to get latitude and longitude from the neighborhood addresses, then use Foursquare API to pull nearby venues based on latitude and longitude.

Once we have nearby venues we will use K-Means clustering (as in Labs) from the Scikit-learn library to cluster and identify/segment similar neighborhoods.

Methodology

I used Beautiful Soup library to scrape the Wikipedia pages of Philadelphia and San Francisco, then used pandas to create the neighborhood datasets.

```
phl_neigh_ds1.head()
```

Out[206]:

	neighborhood
0	Center City
1	South Philadelphia
2	Southwest Philadelphia
3	West Philadelphia
4	Lower North Philadelphia

```
sf_neigh.head()
```

Out[306]:

	neighborhood
0	Alamo Square
1	Anza Vista
2	Ashbury Heights
3	Balboa Park
4	Balboa Terrace

Then I used the google geocoder on the neighborhood datasets to retrieve the geological coordinates of the neighborhoods.

```
phl_neigh_ds1.head()
```

7]:

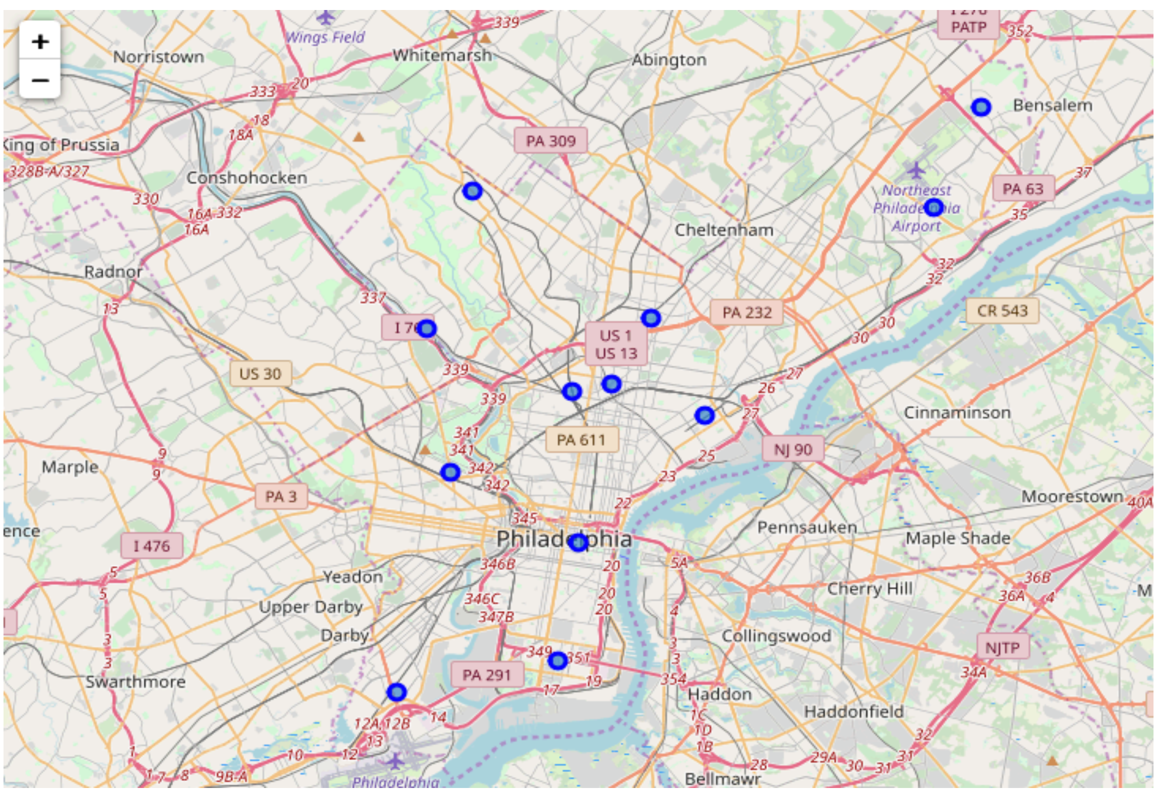
	neighborhood	latitude	longitude
0	Center City	39.950904	-75.157457
1	South Philadelphia	39.909315	-75.166212
2	Southwest Philadelphia	39.898299	-75.236238
3	West Philadelphia	39.975709	-75.212900
4	Lower North Philadelphia	40.006762	-75.142863

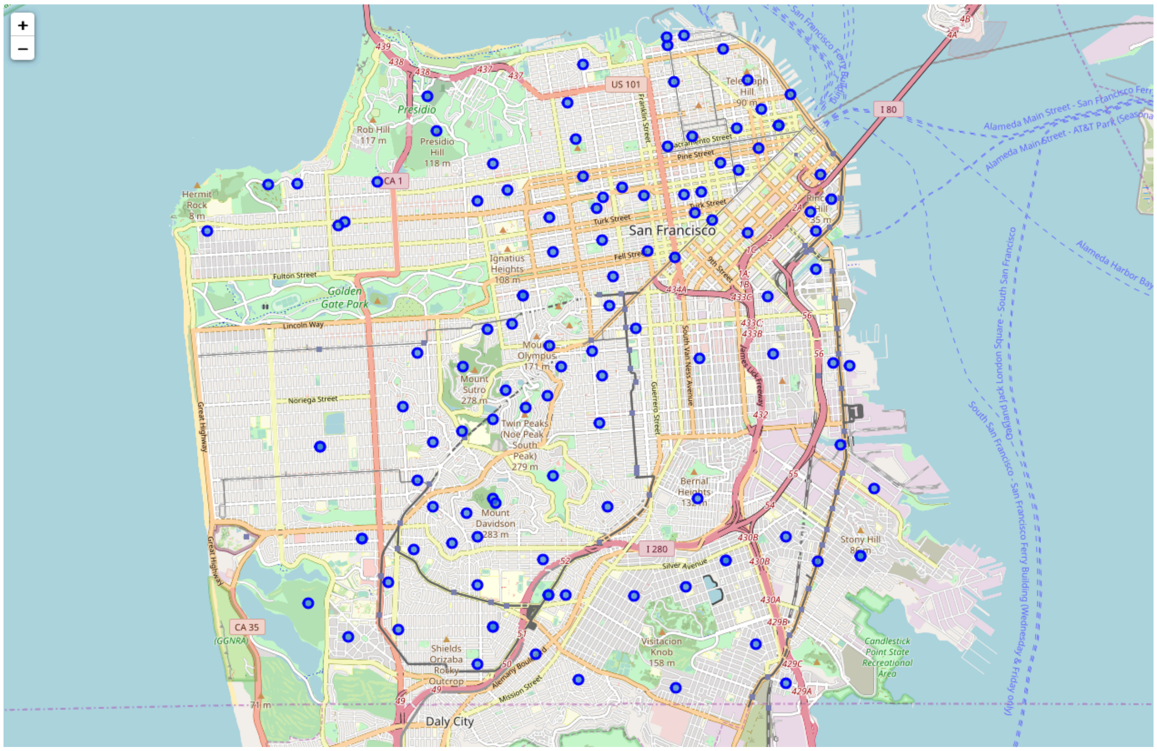
```
sf_neigh.head()
```

8]:

	neighborhood	latitude	longitude
0	Alamo Square	37.777499	-122.433252
1	Anza Vista	37.780868	-122.443185
2	Ashbury Heights	37.769220	-122.448139
3	Balboa Park	37.724569	-122.443357
4	Balboa Terrace	37.731333	-122.468661

Once I had the latitude and longitude of all the neighborhoods in Philadelphia and San Francisco, I used folium library to plot a map and analyze the neighborhoods.





After analyzing the neighborhoods, I used Foursquare API to pull 100 nearby venues for each of these locations and then cleaned JSON file using Pandas.

The resulting dataset is a location of all the venues based on the latitude and longitudes of the neighborhoods.

```
phl_venues.head()
```

(333, 7)

40]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Center City	39.950904	-75.157457	MOM's Organic Market	39.950918	-75.158815	Organic Grocery
1	Center City	39.950904	-75.157457	Luke's Lobster Market East	39.950857	-75.158476	Seafood Restaurant
2	Center City	39.950904	-75.157457	MilkBoy Philadelphia	39.950054	-75.158627	Music Venue
3	Center City	39.950904	-75.157457	Di Bruno Bros.	39.949148	-75.155587	Gourmet Shop
4	Center City	39.950904	-75.157457	Primo Hoagies	39.949216	-75.159052	Sandwich Place

44]:

```
# Check to see how many Thai Restaurant are near center city
phl_venues[phl_venues['Venue Category'].str.contains('Thai', case = False)].head(5)
```

44]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
34	Center City	39.950904	-75.157457	Xiandu Thai Fusion	39.948893	-75.160011	Thai Restaurant
36	Center City	39.950904	-75.157457	Little Thai Market	39.953202	-75.159499	Thai Restaurant
192	Roxborough-Manayunk	40.026001	-75.223111	Chabaa Thai Bistro	40.025885	-75.224442	Thai Restaurant

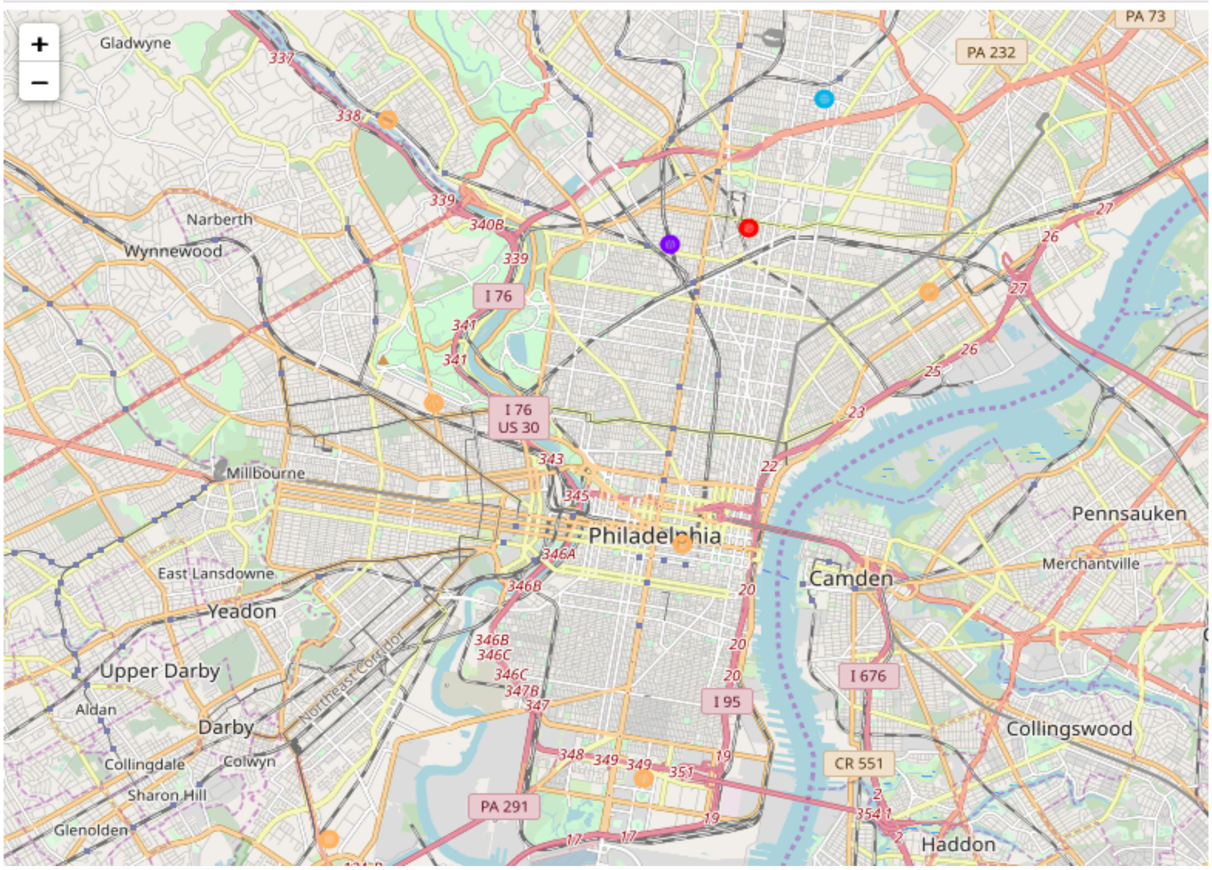
After retrieving and cleaning all the venues data, I analyzed the Philadelphia and San Francisco neighborhoods. I looked at the top 5 venues for each neighborhoods.

```
----Bridesburg-Kensington-Port Richmond----
                  venue  freq
0           Pizza Place  0.13
1        Clothing Store  0.13
2   American Restaurant  0.07
3             Bookstore  0.07
4     Mobile Phone Shop  0.07


----Center City----
             venue  freq
0           Bakery  0.06
1   Sandwich Place  0.04
2      Pizza Place  0.03
3              Bar  0.03
4              Pub  0.03


----Far Northeast Philadelphia----
                      venue  freq
0              Credit Union   0.2
1         Mobile Phone Shop   0.2
2                    Bakery   0.2
3                Smoke Shop   0.2
4   Health & Beauty Service   0.2
```

After analyzing the neighborhoods, I used the datasets to plot K-Nearest mean on Philadelphia and San Francisco datasets.

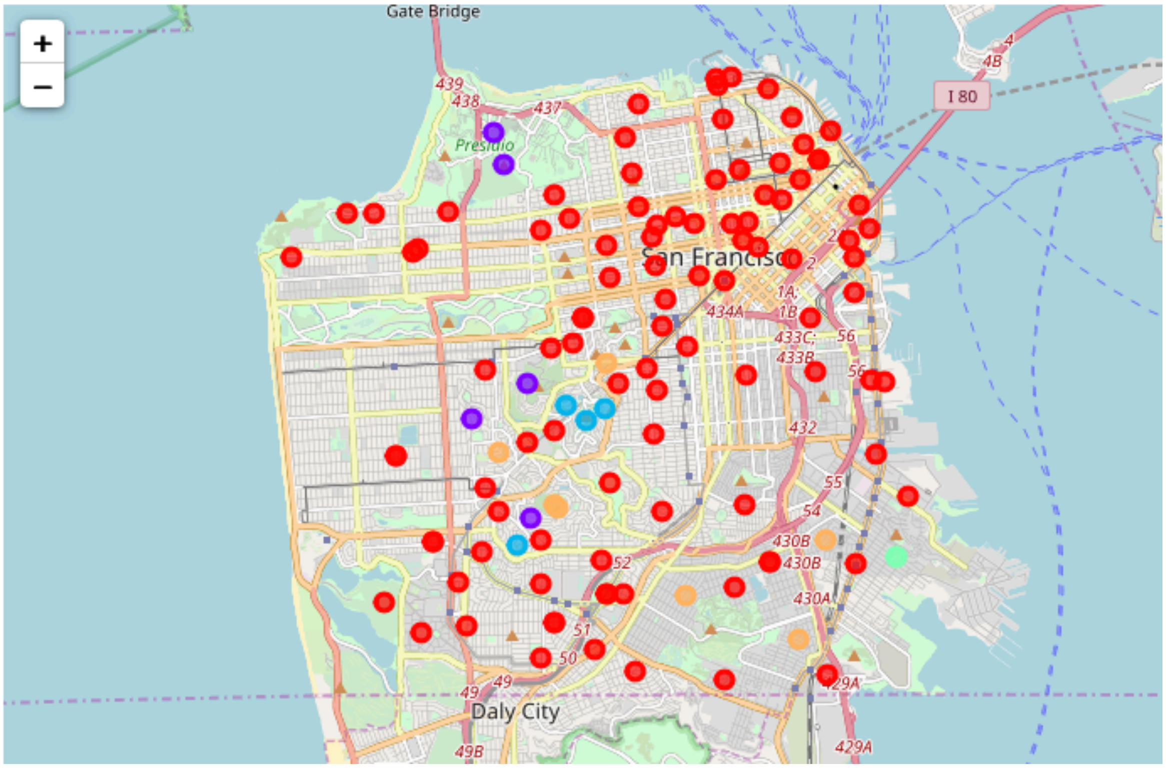
```
phl_map_clusters
```

6]:



```
sf_map_clusters
```

45]:



After plotting the clusters, I examined each cluster to identify their characteristics.

5. Examine Philadelphia Clusters

Lets examine each cluster and determine the discriminating venue categories that distinguish each cluster.

Cluster 1

```
phl_merged.loc[phl_merged['Cluster Labels'] == 0, phl_merged.columns[[1] + list(range(5, phl_merged.shape[1]))]]
```

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
4	40.006762	Cosmetics Shop	Bar	Smoke Shop	Chinese Restaurant	Pharmacy	Hardware Store
10	40.068629	Bar	Pharmacy	Discount Store	Chinese Restaurant	Liquor Store	Shopping Plaza

5. Examine San Francisco Clusters

Lets examine each cluster and determine the discriminating venue categories that distinguish each cluster.

Cluster 1

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

	Latitude	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	37.777499	Café	Park	Pizza Place	Seafood Restaurant	Liquor Store	Sushi Restaurant
1	37.780868	Coffee Shop	Health & Beauty Service	Sandwich Place	Tunnel	Big Box Store	Mexican Restaurant
2	37.769220	Coffee Shop	Thrift / Vintage Store	Pizza Place	Thai Restaurant	Shoe Store	Gift Shop
3	37.724569	Pool	BBQ Joint	Tennis Court	Sandwich Place	Light Rail Station	Dessert Shop
4	37.731333	Light Rail Station	Japanese Restaurant	Dessert Shop	Sushi Restaurant	Shoe Repair	Gym

Result

For the purpose of this project and to make our analysis simple, I assumed some basic information such as:

- Friend of mine lives in Center City
- He Loves Gym and going to parks
- He loves restaurants

Based on the these characteristics – I found that Cluster 5 in Philadelphia is more relevant, Center City is part of Cluster 5 and has nearest venues such as gym, restaurants and other highly active places.

Cluster 5

```
phl_merged.loc[phl_merged['Cluster Labels'] == 4, phl_merged.columns[[1] + list(range(5, phl_merged.shape[1]))]]
```

	Latitude	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
0	39.950904	Sandwich Place	Bar	Burger Joint	Pub	Pizza Place	Indian Restaurant	Salad Place
1	39.909315	Outdoor Sculpture	American Restaurant	Baseball Field	Sandwich Place	Lounge	Ice Cream Shop	BBQ Joint
2	39.898299	Asian Restaurant	Rental Service	Cosmetics Shop	Discount Store	Shoe Store	Fast Food Restaurant	Flower Shop
3	39.975709	Intersection	Board Shop	Art Gallery	Pet Store	Athletics & Sports	Museum	Sculpture Garden
6	39.995553	Clothing Store	American Restaurant	Bookstore	Mobile Phone Shop	Fast Food Restaurant	Donut Shop	Sandwich Place
7	40.026001	Pizza Place	New American Restaurant	Grocery Store	Bakery	Mexican Restaurant	Trail	Gym / Fitness Center
8	40.074334	American Restaurant	Boutique	Gym / Fitness Center	Ice Cream Shop	French Restaurant	Farmers Market	Park

I took these similar characteristics and looked for the clusters in the San Francisco data. I found out that Cluster 1 from San Francisco data also has similar venues nearby.

Cluster 1

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

	Latitude	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
0	37.777499	Café	Park	Pizza Place	Seafood Restaurant	Liquor Store	Sushi Restaurant	Hotel
1	37.780868	Coffee Shop	Health & Beauty Service	Sandwich Place	Tunnel	Big Box Store	Mexican Restaurant	Bus Station
2	37.769220	Coffee Shop	Thrift / Vintage Store	Pizza Place	Thai Restaurant	Shoe Store	Gift Shop	Bookstore
3	37.724569	Pool	BBQ Joint	Tennis Court	Sandwich Place	Light Rail Station	Dessert Shop	Fast Food Restaurant
4	37.731333	Light Rail Station	Japanese Restaurant	Dessert Shop	Sushi Restaurant	Shoe Repair	Gym	Park

Discussion

Some of the challenges I faced is that Foursquare data does not retrieve a high number of venues for some neighborhoods and as a result some of my neighborhoods had less number of venues which might have skewed the results for some clusters.

I also couldn't find enough neighborhoods for Philadelphia, data was not readily available and scrapping other websites would have increased the complexity dramatically. I had to settle with the Wikipedia data.

On the other hand Foursquare had lot of information on San Francisco, even Wikipedia had more neighborhoods listed and as a result had much better venue count. So having good data is definitely the corner stone for building an accurate machine learning model.

Conclusion

After analyzing all the data, I come to the conclusion that my friend should settle down in neighborhood from Cluster 1 in San Francisco.

Below are some of the neighborhoods in clusters 1 San Francisco.

- Alamo Square
- Anza Vista
- Ashbury Heights
- Balboa Park
- Balboa Terrace
- Cathedral Hill