CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

By Pruthvi Reddy

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.

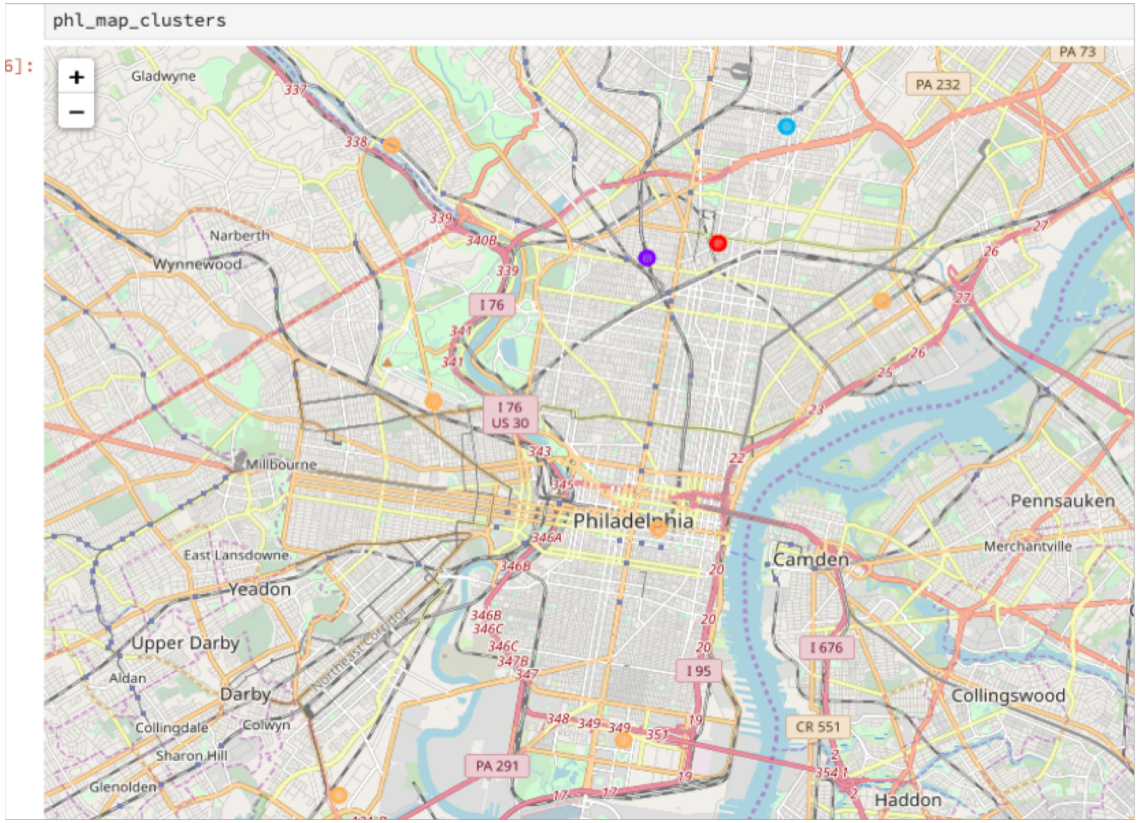
Once we have nearby venues we will use K-Means clustering (as in Labs) from the Scikit-learn library to cluster and identify/segment similar neighborhoods.

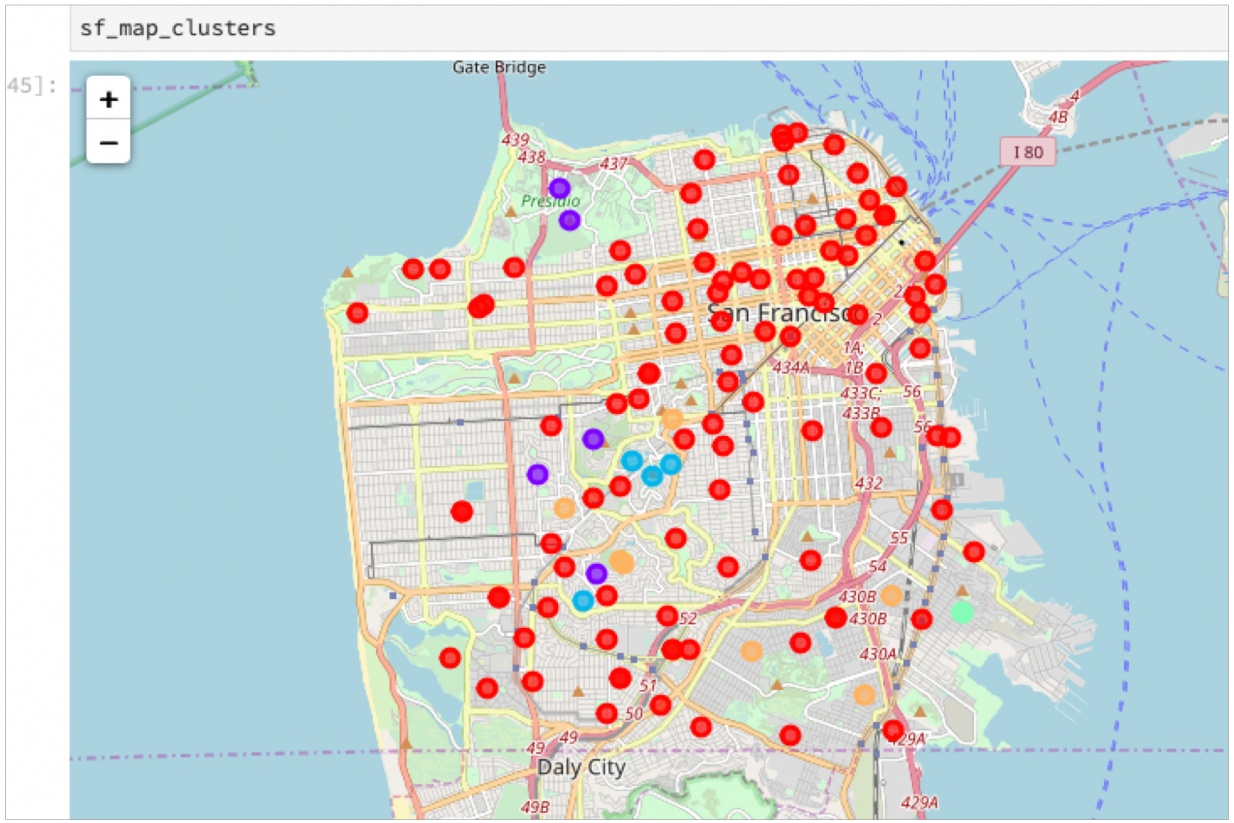
METHODOLOGY

Used K-Nearest Mean to perform segmentation and clustering on the Philadelphia and San Francisco neighborhood's.

Used the neighborhood characteristics to identify similar clusters.

METHODOLOGY





RESULTS

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.

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.

RESULTS

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

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