Commuter Friendly Neighborhoods of San Francisco, CA, USA

Golden Gate Bridge, San Francisco

IBM Applied Data Science Capstone Project
The Battle of Neighborhoods

Usha Manoharan

July 2020

Business Problem

San Francisco (fondly called "The City" by the locals), is a beautiful and culturally diverse city. It has many parks with lush greenery, coastline trails with views of the Pacific Ocean, world class museums and ample options for dining. People from the Chinese, Japanese, Latino, LGBTQ, Hippie, Italian, Southeast Asian cultures have all made San Francisco their home for many decades resulting in culturally vibrant neighborhoods. Above all, many industries and particularly high-tech companies like Airbnb, Dropbox, Facebook, Salesforce, Uber…to name a few, have their presence in San Francisco. Top notch universities like Stanford and Berkeley are in the proximity which help in bringing talented individuals closer to the city. All these factors attract people from all walks of life to the city to live and make a living.

To provide a smooth and stress free transition to a person moving into San Francisco, I would like to address the problem of identifying the best "commuter friendly" neighborhoods where people can live and use economical options such as buses and trains to commute to the neighborhoods for work, pleasure, dining and other activities of interest within the city.

Introduction

The goal of this capstone project is to identify the neighborhoods of San Francisco that are best suited for people to commute using public transportation.

My approach is to search the web and obtain a list of zip codes of San Francisco, then find their geographical co-ordinates to use as input to the Foursquare API to retrieve location data related to transit venues. To keep it simple, I will be looking for only venues that relate to buses or trains. I will use the K-means algorithm to segment the zip codes into clusters with similar venue categories. I will also use the Folium library to visually represent the zip codes and the commuter friendly clusters in a map of San Francisco.

The outcome of this analysis should lead to the identification of zip codes in San Francisco that provide similar options to commute by bus or train. I think this analysis will benefit a wide variety of people who is seeking economical modes of transportation in the city of San Francisco. Fresh graduates taking up new jobs, tourists visiting the city, immigrants trying to move in and establish themselves, city/transportation officials interested in knowing about the distribution and frequency of bus/train stations in the city are all some of the examples that I can think of.

Data

Let's take a closer look at the data and its source(s), used in this project.

Zip codes of San Francisco

I downloaded an excel file from the web at https://www.melissa.com/v2/lookups/cityzip/city/?name=san+francisco, which contains the list of zip codes of San Francisco.

	ZIP
45	94143
46	94137
47	94080
48	94083

Figure 1: Sample list of zip codes of San Francisco

Geographical co-ordinates

Next, I used the "pgeocode" library from https://pypi.org/project/pgeocode/ to obtain the latitude and longitude for each of the listed zip codes.

	zipcode	latitude	longitude
45	94143	37.7631	-122.4586
46	94137	37.7749	-122.4194
47	94080	37.6574	-122.4235
48	94083	37.6547	-122.4077

Figure 2: zip codes with latitude and longitude information

Location Data from Foursquare

Foursquare is a location technology platform that provides real-time access to location based venue data and user content.

I used the strings "bus" and "train" in foursquare's Places API, to obtain the venues related to buses and trains for each of the zip codes.

	Zipcode	Zipcode Latitude	Zipcode Longitude	Venue Name	Venue Category	Venue Address	Venue Latitude	Venue Longitude
0	94163	37.7749	-122.4194	MUNI Bus Stop - Van Ness & Market	Bus Station	Van Ness Ave.	37.775657	-122.419270
1	94163	37.7749	-122.4194	Muni Bus 5474	Bus Line	NaN	37.774795	-122.420017
2	94163	37.7749	-122.4194	MUNI bus 8143	Bus Line	NaN	37.774355	-122.420631
3	94163	37.7749	-122.4194	MUNI bus #8198	Bus Line	NaN	37.776567	-122.419050
4	94163	37.7749	-122.4194	MUNI Bus Stop - Mission & 11th	Bus Line	Mission St	37.774149	-122.417307
...	...	...	...	...	...	...	...	...

Figure 3: Venues obtained from Foursquare API for search query "bus"

Methodology

Acquire, Explore and Prepare Data

In the first step, I searched the internet and collected San Francisco's zipcode data and stored it into a dataframe. I got 48 unique zip codes for San Francisco. To the same dataframe, I added the latitude and longitude values obtained from the "pgeocode" library for each of the zipcodes.

Next, I used the Foursquare API to search for "bus" and "train" venues within a radius of 500 meters of each zipcode and a maximum limit of 30 venues per zipcode. I skipped the zipcode from further analysis if there were no "bus" or "train" venues returned in the API results. I also dropped the venues which did not have a "Venue Category" listed in the API results.

Then, I filtered the results to contain only these columns that will be useful in our analysis - "Venue Name", "Venue Category", "Venue Address", "Venue Latitude" and "Venue Longitude". See Figure 3 above.

Finally, I used one-hot encoding to pivot the venue categories from rows to columns and calculated the mean of frequency of occurrence for each category.

	Zipcode	Art Gallery	Automotive Shop	Bar	Beer Garden	Building	Bus Line	Bus Station	Bus Stop	Business Center	...	Thai Restaurant	Theater	Trade School	Train	Tra Stati
0	94083	0.000000	0.00	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.3333
1	94102	0.000000	0.00	0.000000	0.0	0.000000	0.333333	0.256410	0.153846	0.000000	...	0.000000	0.000000	0.000000	0.128205	0.0000
2	94103	0.024390	0.00	0.000000	0.0	0.000000	0.268293	0.268293	0.073171	0.000000	...	0.000000	0.000000	0.000000	0.146341	0.0000
3	94104	0.000000	0.00	0.000000	0.0	0.021277	0.170213	0.446809	0.021277	0.021277	...	0.000000	0.000000	0.000000	0.106383	0.0000
4	94105	0.000000	0.00	0.000000	0.0	0.000000	0.437500	0.187500	0.062500	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.0000
5	94107	0.000000	0.00	0.000000	0.0	0.000000	0.263158	0.315789	0.263158	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.0000
6	94108	0.000000	0.00	0.027778	0.0	0.000000	0.194444	0.472222	0.027778	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.0000
7	94109	0.000000	0.00	0.000000	0.0	0.030303	0.303030	0.424242	0.090909	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.0303
8	94110	0.000000	0.00	0.000000	0.0	0.000000	0.235294	0.235294	0.294118	0.000000	...	0.000000	0.058824	0.000000	0.000000	0.0000
9	94111	0.000000	0.00	0.000000	0.0	0.023256	0.279070	0.279070	0.023256	0.023256	...	0.000000	0.000000	0.000000	0.116279	0.0000

Figure 4: Dataframe after one-hot encoding of Venue Categories and calculating the mean of frequency of occurrence

Analyze Data

After preparing the data, I used the K-means algorithm to segment the zip codes into clusters that have similar bus and train venues. I used 5 for the number of clusters. The resulting data frame is shown below.

	zipcode	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	94163	37.7749	-122.4194	1.0	Bus Line	Bus Station	Train	Bus Stop	Moving Target
1	94177	37.7749	-122.4194	1.0	Bus Line	Bus Station	Train	Bus Stop	Moving Target
2	94109	37.7917	-122.4186	0.0	Bus Station	Bus Line	Bus Stop	Other Nightlife	General Travel
3	94142	37.7749	-122.4194	1.0	Bus Line	Bus Station	Train	Bus Stop	Moving Target
4	94115	37.7856	-122.4358	1.0	Bus Line	Bus Station	Bus Stop	Transportation Service	Art Gallery

Figure 5: Clusters of zip codes with transit venues like "bus" and "train"

Results

The resulting 5 clusters of my analysis are shown in different colored dots in the map below.

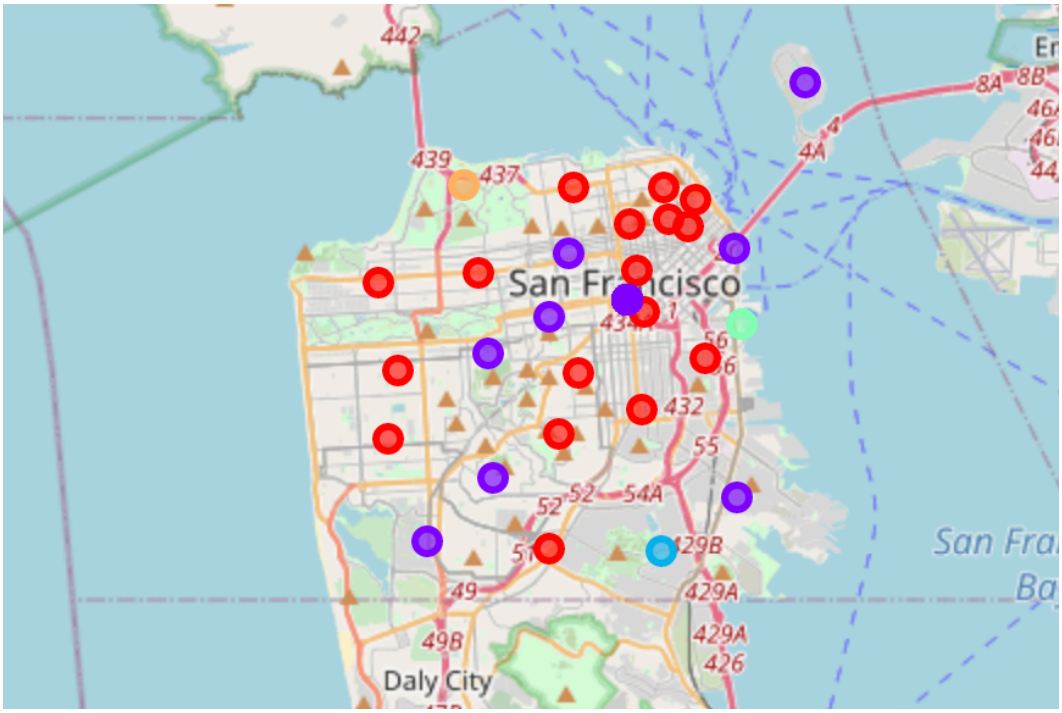

Figure 6: Resulting 5 commuter friendly clusters of San Francisco

Cluster 0 Cluster 1 Cluster 2 Cluster 3 Cluster 4

From the above picture we see that almost all parts of the city is marked by colored dots which indicate that most parts of the city have some sort of bus/train station. Particularly the red and purple dots representing “Cluster 0” and “Cluster 1” respectively are more dominant than the other clusters.

Discussion

Further, I analyzed the clusters to get the count of the top 5 common venues in each cluster. The pictures below show a summary of my findings.

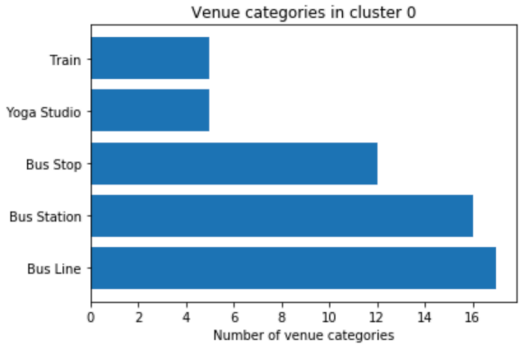

Figure 7: Cluster 0 with mostly Bus venues and few Train venues

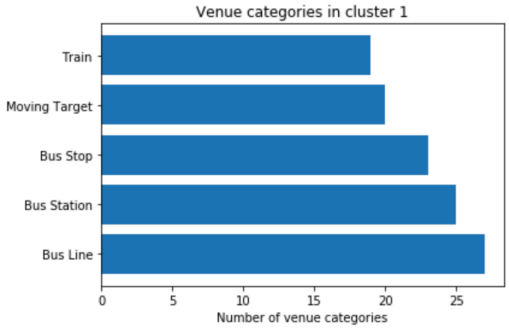

Figure 8: Cluster 1 with the most Bus venues and fairly close number of Train venues

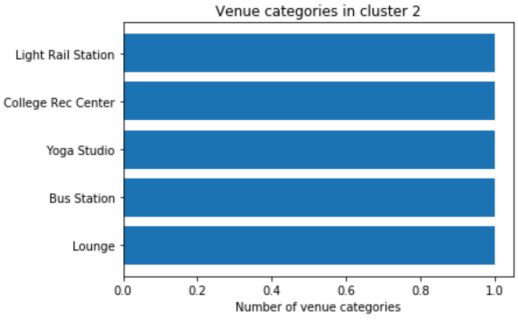

Figure 9: Cluster 2 with only one bus/train venue

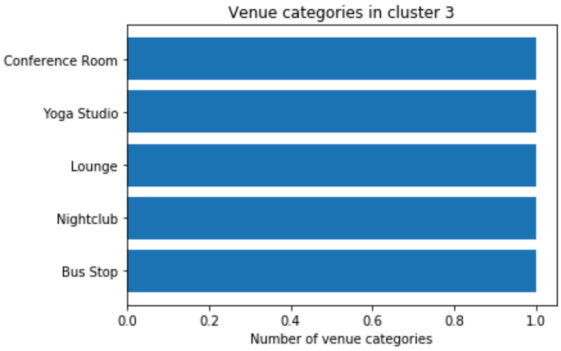

Figure 10: Cluster 3 with only one bus venue

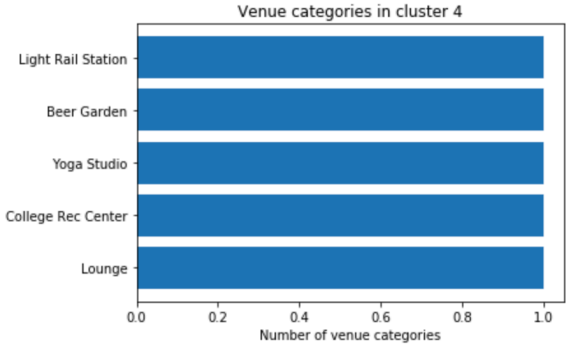

Figure 11: Cluster 4 with only one train venue

Cluster 0 - has about 5 train venues and nearly 45 venues identified as either “Bus Line”, “Bus Station” or “Bus Stop”.

Cluster 1 - has about 20 train venues and nearly 75 venues identified as “Bus Line”, “Bus Station” or “Bus Stop”.

Cluster 2 - has exactly 1 bus and 1 train venue.

Cluster 3 - has exactly 1 bus venue.

Cluster 4 - has exactly one train venue.

Next, I counted the number of zip codes in each of the clusters. I found that about half (27 out of 48) the zip codes in San Francisco falls into “Cluster 1” and nearly the other half (18 out of 48) falls into “Cluster 0”. There is one zip code each in cluster 2, 3 and 4.

	zipcode
Cluster Labels	
0	18
1	27
2	1
3	1
4	1

Figure 12: Number of zip codes in each cluster

Conclusion

From the above analysis of the zip codes, it appears that San Francisco is well connected by Buses and Trains. The neighborhoods shown in **Cluster 0** and **Cluster 1** are both good options for anyone wanting to use buses and trains to commute within the city. Clearly **Cluster 1** is the best neighborhood in the city of San Francisco to commute by both train and bus, followed by **Cluster 0**