IBM Data Science Capstone Project blogpost

Find a suitable neighborhood for Chinese takeout in San Francisco

Hanna Lu September 14, 2020

Business Problem

Many restaurants in San Francisco had closed down due to the COVID-19 pandemics. Outdoor seating and take-outs become the normal way for restaurants to stay open.
Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to determine whether there is a location in San Francisco that:

1. Offer a convenient location for a Chinese takeout restaurant?
2. How dense are restaurants in different neighborhoods?
3. How many different types of restaurants and venues for socializing are in the neighborhoods?

For this particular analysis, we will focus on postal/zip code -level data, for the neighborhoods that fall within City of San Francisco. This may not be the best approach to take since San Francisco has a lot of distinct neighborhoods such as Nov Valley, Sunset and Marina, and the size of each neighborhood vary.

Target Audience

The primary audiences for this project are for potential investors to find a suitable location for a Chinese restaurant with just a take-out window and limit menu to reduce the start-up cost

Methodology (Synopsis). Please see Part 2 for a more detailed description

1. Review data specifications and availability. Steps include:
    - List data Requirements
    - Collect data – locate websites offering Zip code information that can be readily scraped.
    - Load Foursquare data for all Zip codes in the city of San Francisco.
    - Data understanding:
    - Clean data
    - Identify and describe the limitations of this research

Data used

Sources Included but not limited to:

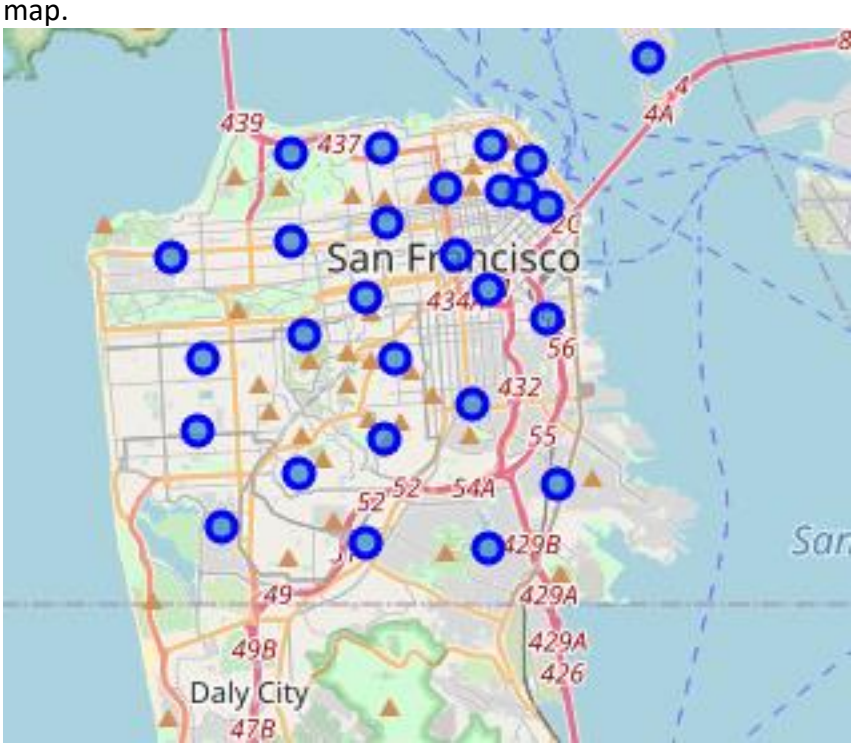
Yelp website
Foursquare Data
San Francisco neighborhood map
http://ciclt.net site to find San Francisco Zip codes
https://public.opendatasoft.com to get csv file with latitude and longitude for each zip code
U. S. Census

Methodology Detail

1. Our first step will be to get lists of neighborhoods for the cities of San Francisco The website http://ciclt.net/sn/clt/capitolimpact/gw_ziplist.aspx?ClientCode=capitolimpact&State=ca&StName=California&StFIPS=06&FIPS=06075 has this information.
2. We will use python's *beautifulsoup* library to extract the needed postal code lists. Then, we will get the geographical coordinates (latitude and longitude) so we can use them to query the Foursquare API database.

	PostalCode	City	Latitude	Longitude
0	94101	San Francisco	37.784827	-122.727802
1	94102	San Francisco	37.779329	-122.419150
2	94103	San Francisco	37.772329	-122.410870
3	94104	San Francisco	37.791728	-122.401900
4	94105	San Francisco	37.789228	-122.395700
5	94107	San Francisco	37.766529	-122.395770
6	94108	San Francisco	37.792678	-122.407930

A geocoder will allow us to do so. We will then be able to load this information into a pandas dataframe, then using *folium*, we will visualize each city's neighborhoods on the map.



3. Using the Foursquare API, we will subsequently get the top 100 venues that are within a radius of 500 meters from the center point of each Zip or Postal Code.

	PostalCode	City	Latitude	Longitude	Population
1	94102	San Francisco	37.779329	-122.41915	San Francisco
2	94103	San Francisco	37.772329	-122.41087	San Francisco
3	94104	San Francisco	37.791728	-122.40190	San Francisco
4	94105	San Francisco	37.789228	-122.39570	San Francisco
5	94107	San Francisco	37.766529	-122.39577	San Francisco
6	94108	San Francisco	37.792678	-122.40793	San Francisco
7	94109	San Francisco	37.792778	-122.42188	San Francisco

4. We do this by making API calls to Foursquare, passing the geographical coordinates until we are done via a Python loop. Foursquare then returns venue data to us in a JSON format, and we extract the venue name, category, latitude, and longitude. With these data, we will be able to check to see how many venues were returned for each neighborhood and to tally up the number of unique categories can be curated from all the returned venues.

	venue.name	venue.categories	venue.location.lat	venue.location.lng
0	Boba Guys	[{'id': '52e81612bcbc57f1066b7a0c', 'name': 'B...	37.766448	-122.397042
1	Bottom of the Hill	[{'id': '4bf58dd8d48988d1e9931735', 'name': 'R...	37.765116	-122.396218
2	Fitness Urbano	[{'id': '4bf58dd8d48988d175941735', 'name': 'G...	37.765684	-122.397009
3	Pawtrero Hill Bathhouse and Feed Company	[{'id': '4bf58dd8d48988d100951735', 'name': 'P...	37.764140	-122.394500
4	UCSF Bakar Fitness & Rec Center	[{'id': '4bf58dd8d48988d176941735', 'name': 'G...	37.768146	-122.393290
5	Daggett Plaza	[{'id': '4bf58dd8d48988d163941735', 'name': 'P...	37.766920	-122.396027

	name	categories	lat	lng
0	Boba Guys	Bubble Tea Shop	37.766448	-122.397042
1	Bottom of the Hill	Rock Club	37.765116	-122.396218
2	Fitness Urbano	Gym / Fitness Center	37.765684	-122.397009
3	Pawtrero Hill Bathhouse and Feed Company	Pet Store	37.764140	-122.394500
4	UCSF Bakar Fitness & Rec Center	Gym	37.768146	-122.393290
5	Daggett Plaza	Park	37.766920	-122.396027

5. The next step will be to conduct k-means clustering – using the mean frequency of occurrence of each venue category to create a centroid for each postal code. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is a simple and popular unsupervised machine learning algorithms.

	PostalCode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	94102	37.779329	-122.41915	Louise M. Davies Symphony Hall	37.777976	-122.420157	Concert Hall
1	94102	37.779329	-122.41915	War Memorial Opera House	37.778601	-122.420816	Opera House
2	94102	37.779329	-122.41915	Herbst Theater	37.779548	-122.420953	Concert Hall
3	94102	37.779329	-122.41915	San Francisco Ballet	37.778580	-122.420798	Dance Studio
4	94102	37.779329	-122.41915	War Memorial Court	37.779042	-122.420971	Park
5	94102	37.779329	-122.41915	Asian Art Museum	37.780178	-122.416505	Art Museum

```
sanfran_venues.groupby('PostalCode').count()
```

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
PostalCode						
94102	90	90	90	90	90	90
94103	77	77	77	77	77	77
94104	100	100	100	100	100	100
94105	81	81	81	81	81	81
94107	28	28	28	28	28	28
94108	93	93	93	93	93	93

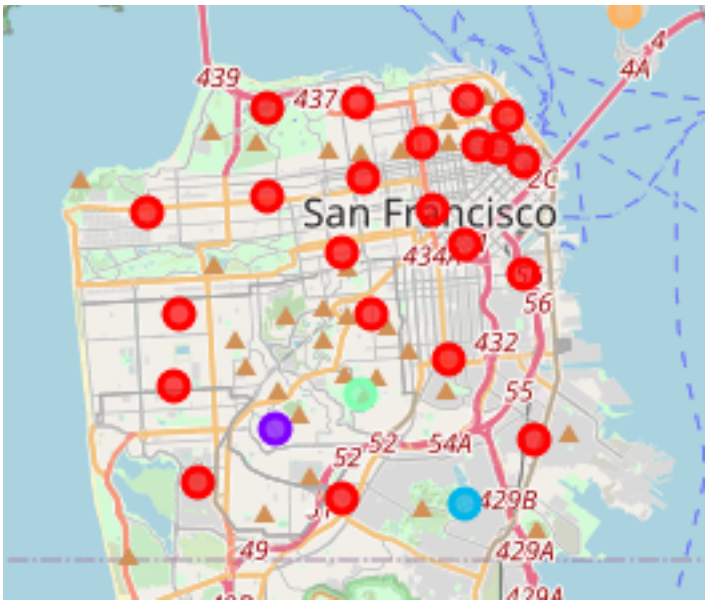
	PostalCode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	94102	Coffee Shop	Hotel	Café	Wine Bar	Vegetarian / Vegan Restaurant	French Restaurant	Boutique	Bakery	Optical Shop	Beer Bar
1	94103	Nightclub	Cocktail Bar	Gay Bar	Food Truck	Motorcycle Shop	Bar	Thai Restaurant	Cosmetics Shop	Furniture / Home Store	Sushi Restaurant
2	94104	Coffee Shop	Food Truck	Men's Store	Japanese Restaurant	Sandwich Place	Sushi Restaurant	Gym	Hotel	Italian Restaurant	Cosmetics Shop
3	94105	Coffee Shop	Food Truck	Café	Sandwich Place	Art Gallery	Gym / Fitness Center	Gym	Salad Place	New American Restaurant	Lounge
4	94107	Wine Shop	Breakfast Spot	Park	Café	Peruvian Restaurant	Pool	Bookstore	Coffee Shop	French Restaurant	Rock Club

The results will allow me to identify which neighborhoods are most likely to meet the need of potential investors for my Chinese takeout restaurant.

6.

Conclusion

The results will allow me to identify which neighborhoods are most likely to meet the need of potential investors for my Chinese takeout restaurant.



Foursquare lists out the most common venues of each Zip code.
We will be looking for area that have less Chinese restaurants or any other kinds of restaurant. For the following three Zip code, 94102, 94122 and 94127, the first two areas have too many restaurants. The third area has bus line and ATM, but no restaurants listed. It could potentially a suitable location to open a Chinese takeout restaurant. Of course, there are a lot more Zip codes we can investigate.

```
----94102----
                          venue  freq
0                   Coffee Shop  0.06
1                         Hotel  0.04
2                          Café  0.04
3              French Restaurant  0.03
4  Vegetarian / Vegan Restaurant  0.03
```

1.

```
----94122----
                venue  freq
0  Chinese Restaurant   0.2
1            Bus Line   0.1
2         Bus Station   0.1
3        Dessert Shop   0.1
4          Playground   0.1
```

2.

```
----94127----
           venue  freq
0      Bus Line   0.5
1         Trail   0.5
2           ATM   0.0
3  Neighborhood   0.0
4        Office   0.0
```

3.