IBM Data Science Capstone Project

Find a suitable neighborhood for Chinese takeout in San Francisco

Part 1: Research Plan

Hanna Lu September 14, 2020

Research Plan
Submitted for Week 1 of 2: Data Science Capstone Project

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. 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. 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.
4. 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.
5. The results will allow me to identify which neighborhoods are most likely to meet the need of potential investors for my Chinese takeout restaurant.