San Francisco Bay Area New Bubble Tea Shop Classification

Rushil Jayant

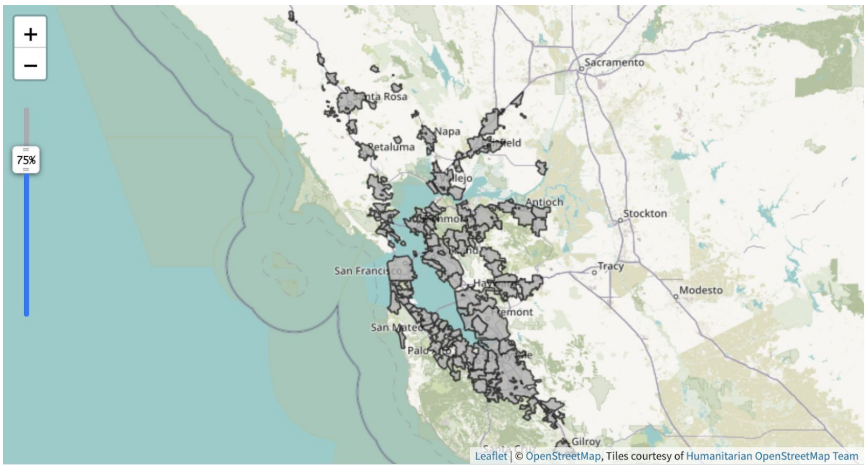
Introduction

Bubble Tea, known as Boba for short is a Taiwanese tea-based drink invented in Taiwan in the 1980s. Recepies are varied, and contain tea of some kind. Toppings are known as "pearls" or "boba" are tapioca balls are added to give the tea a unique taste.

In the San Francisco Bay Area, Bubble Tea is a very popular drink and a very profitable business endeavor. **An interested entrepreneur** (Target Audience) wants to open a new Bubble Tea Café in the area to have a share in this booming business. Since the market is very saturated, the entrepreneur wants to make sure that he is able to **maximize market share** by finding out the ideal City to open the new Bubble Tea Shop. The interested entrepreneur wants to make sure that the cit(ies) where he will open the new shop must have at least one shop already.

Data Acquisition

First, it uses the simplemaps.com's US Cities free data. This data provided several data points on about 30,000 cities in the USA, which included city, city's nickname, state ID, state name, county FIPS, county name, latitude, longitude, population, density, military, incorporation status, timezone, ranking, and zipcodes, and identification.



Next, it uses a GeoJson file from Stanford University of the Cities in the San Francisco Bay Area which is used in all the visualizations used in the project.

Then, the project used the wikipedia page on the cities and towns in the bay area that would later be used to filter the US Cities that are only in the San Francisco Bay Area. This also gave information into the different populations, densities, and incorporation dates of the cities that weren't used.

Finally, to meet the requirements of the project, the Foursquare data was used to get the different locations of Boba Shops around the Bay Area's cities. This data was used in the preprocessing visualizations and in the clustering machine learning model.

All these data sources were vital in making this project like how it is today and were helpful in bringing the results I have now.

Data Cleaning

First, the list of the cities in the Bay Area from this Wikipedia Page was scraped through the utilization of the Beautiful Soup (BS4) library. After in depth analysis of the HTML code in the Wikipedia Page, I was able to construct an algorithm that would scrape the table and return a list containing all the cities of the Bay Area.

Next, I was able to cross-reference that with the US Cities list which was cleaned by removing all the cities that were not in California and then removing all the cities that was not in that list.

Afterwards, I had constructed a looping algorithm that used a function that returned a DataFrame that contained the different Boba Shops in the area through parsing the Foursquare endpoint API, and after every iteration the loop added the new results to the DataFrame, which is now called *BobaShops*.

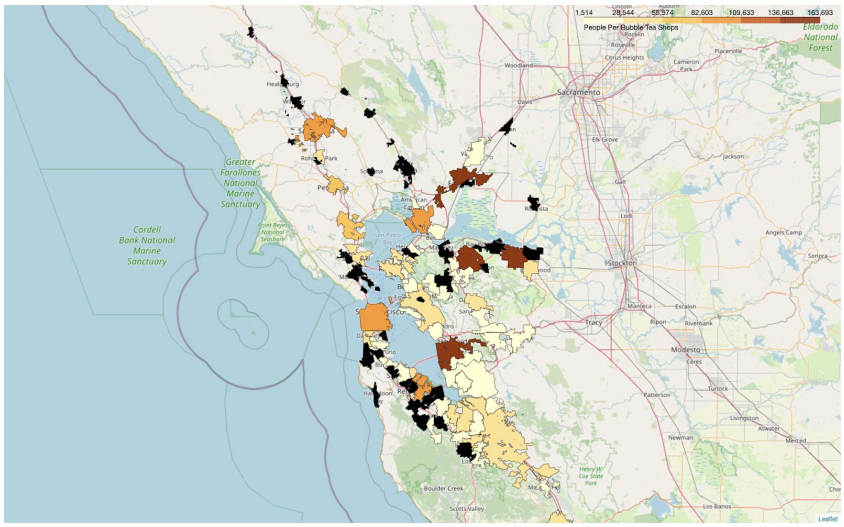
Then, the *BobaShops* DataFrame was removed of the duplicates, and checked again if the addresses were in the Bay Area using the US Cities data (now cleaned with only Bay Area Cities). Then the Indexes were reset and the data was prepared.

Feature Selection

In this project, after all the cleaning process was done, a new DataFrame was created called *BobaCityCount* which was containing the City's name, number of Bubble Tea Shops, population in the city, density, and the ratio between city population and the number of Bubble Tea Shops. This was then used in visualitions.

Exploratory Data Analysis Through Geospatial Visualization

In this project there were four geospatial maps containing data relating to population, density, number of Bubble Tea Shops in each city, and people per number of Bubble Tea Shops in each city.



The most important and striking geospatial visualization of all is the one on the left, which shows the number of people per Bubble Tea Shop in each city.

To start, the data used in this visualization was the People per Bubble Tea Shop column in the *BobaCityCount* DataFrame. This column gave the ratio between a city's population to the number of Boba Shops in the City.

The cities in Black are the Cities in the Bay Area that do not have any Bubble Tea Shops. The sites which are darker are cities with higher People per Bubble Tea Shop ratio than that of cities which are lighter. This is a good way of showing predicted demand in each city as it shows how many customers can come to the shop if all shops in the city get equal number customers. Cities with darker colors (not black) are cities with higher predicted demand, so they are better locations.

Predictive Modeling

In this project I used the KMeans algorithm using Sci-kit learn (Sklearn in python) that took all the cities that had at least one Bubble Tea Shop and clustered it into K groups.

I chose to cluster the cities into 8 groups, because I felt for the volume of data provided that many clusters were necessary. After the clustering process, I did in depth research on each cluster and wrote comments which conveyed how each cluster's data points were clustered, and from that I wrote if the cluster's cit(ies) in the cluster are one of the *Best, OK,* or *Bad* in the cities in the Bay Area.

Results

The purpose of this project was to determine the best cities to open a new Bubble Tea Franchise, and using various data sources, and Machine Learning Clustering, I was able to determine that the Best Cities are San Francisco, Concord, Fremont, Santa Rosa, and Antioch. The goal was to maximize population, density, and predicted demand, and find the clusters which try to do that. If for some reason these cities are not suitable for whatever reason, there are more cities that are OK to open a new Bubble Tea Shop. This project also details the cities to avoid in clusters 0 and 4.

Discussion

To expand this project further, one can apply this same clustering algorithm to other cities to determine where one can have a new Bubble Tea Shops. One can also use more data sources like traffic data, social media, and more in depth data of each store to take advantage. Also, It is possible to achieve the same result through different models like classification or logistic regression. There are plenty of ideas, but this is just a simple attempt to answer the question.

Conclusion

In summary, this project provides better reasoning ablity and insight to an interested investor who is willing to put his hard earned money to make this a success. It is very helpful to use Data Science to make research more useful, accurate, and predictable to make the best decisions that would not have otherwise been made. It is true that this project might not provide the perfect results, or be prefect, but is a better attempt that what would have been made without Data Science.