Capstone Project -

The Battle of Neighborhoods

Introduction

Nowadays, people pay more attention to their health than any time before. How to keep fit/lose weight have been a hot topic all over the world. Generally, people do more exercise/workout and eat healthier to achieve this goal. In my experience, opening a fitness center is more expensive than opening a restaurant. So, I would like to focus on healthy food.

In this project, I want to open a restaurant that focuses on low GI diet. I choose Kaohsiung city as my target city. It is the second large city in Taiwan (the largest city is Taipei). I collected and analyzed some data to select the best location in the chosen city for a low GI restaurant.

Data

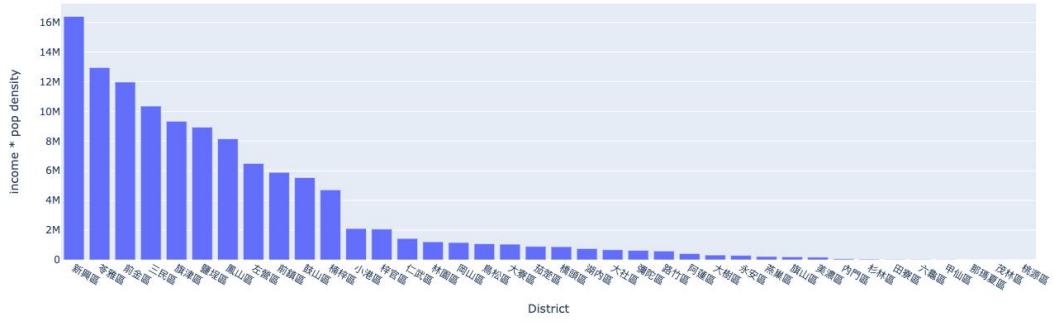
Basic districts data and basic census data is scraped from Wikipedia. Detailed census data is downloaded from website of Kaohsiung government and Taiwan government. Geospatial data is from geopy and Foursquare. The data is not up to date, but it is the latest version I can get.

Methodology

I do background research by google, and I also discuss my idea with a friend born in Kaohsiung city. Plus, I search information on Wikipedia and government websites. I analyze the data using data science methods and my personal experiences.

Firstly, I examine my business problem. Generally healthy food is more expansive than normal food, so I analyze tax information of each districts in Kaohsiung. I also notice that eating healthy is an everyday thing. My

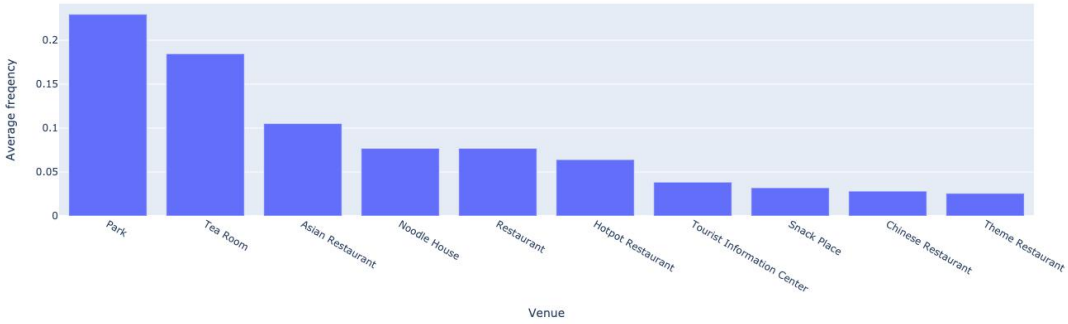
restaurant aims to provide low GI food for daily meals, so I analyze population information, too. Here I selected income median and population density as two most important factors and got this chart:

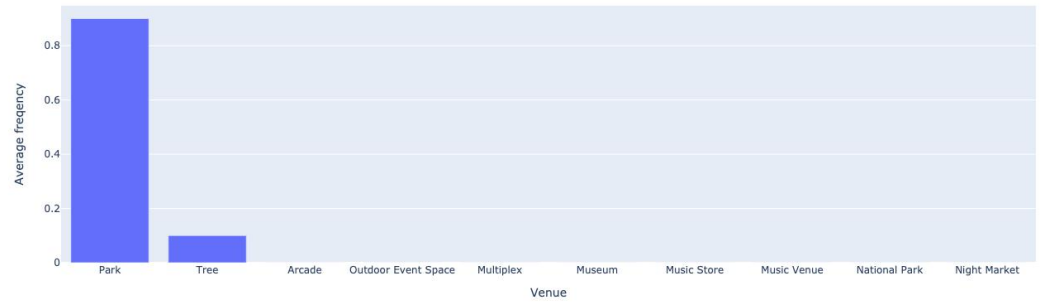


The result shows 新興區, 苓雅區, and 前金區 are better districts for me

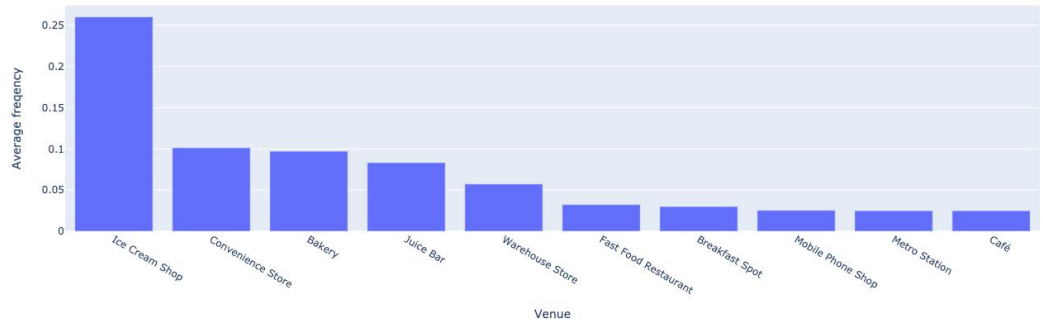
My first instinct is to find the “healthy areas”, here I mean the places with a lot of gym/fitness centers or sports related venues. I searched nearby venues of every neighborhoods and did Kmeans clustering on all 856 neighborhoods. I manually selected k value (30) and examined the result. I didn’t see the trend I hope to find, but I did see some intersting results:

1. Group 11, 27 seemed healthier because the top venue there is “park”.

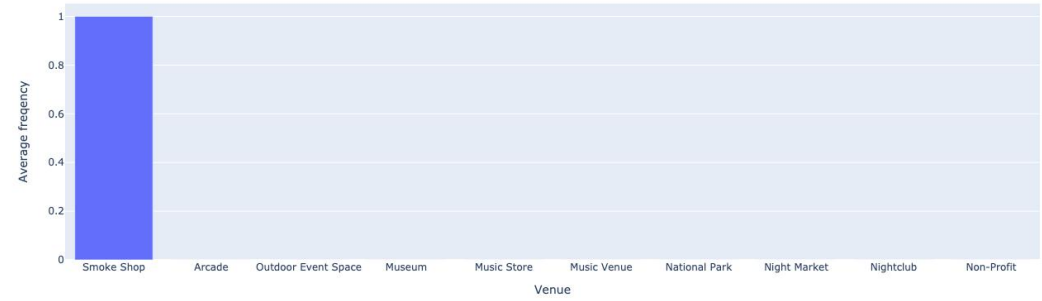




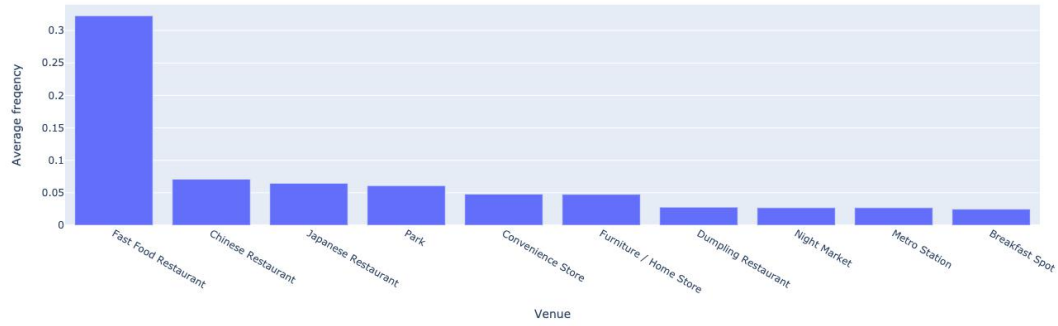
2. These groups seemed not so healthy:
    I. Group 1’s top venue is “Ice Cream Shop”



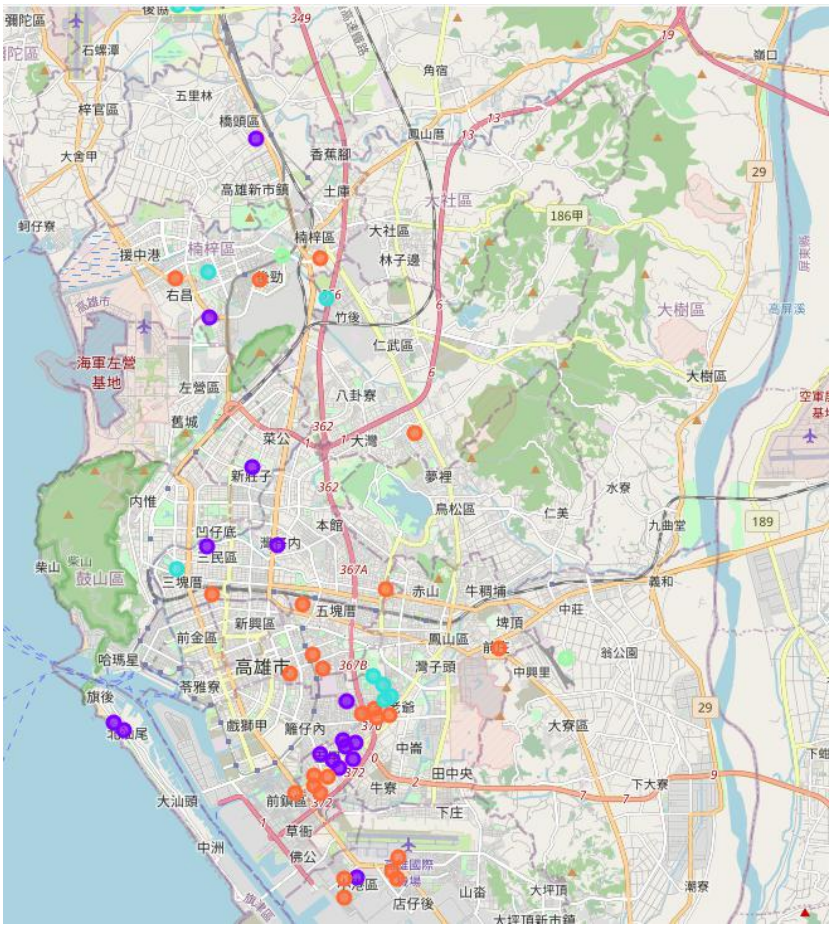
II. Group 25's top venue is "Smoke Shop"



III. Group 26's top venue is "Fast Food Restaurant"



According to the data, people around "not healthy" groups don't like healthy food very much. So I had better not to choose these places. The picture below is the visualization result of "not-healthy" groups.



According to some reports, I listed some attributes my customers may have:

1. People who go to gym/fitness centers a lot may love low GI foods.
2. Female may account substantial part of customers of healthy food.
3. Students don't like healthy food, but teathers and students' parents may like.

As a common sense, my restaurant should be close to public transportation facilities.

Base on assumptions above, I searched for certain kinds of venue categories:

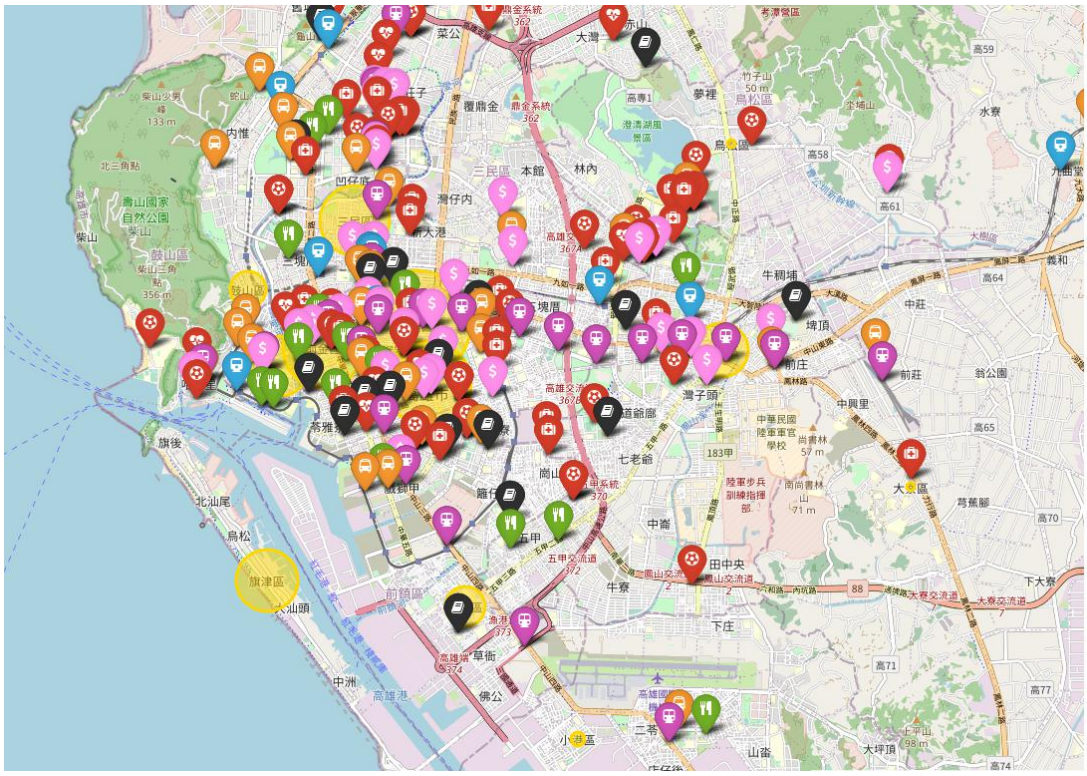
1. Gym/fitness center: People there would be my main customers.
2. Schools: Students usually don't like healthy food. They seldom buy healthy food even they like. But teathers and students' parents may like.
3. Metro stations, bus stops and train stations: stream of people means stream of customers.
4. Bank: In Taiwan, female workers are much more than male workers in

banks. What's more, all workers there are required to wear uniform. So I think they are also my potential customers.

5. Hospital: Workers there tend to pay more attention to health issues.

I also did some search about existing restaurant, and it turned out that there is no healthy food restaurant around.

Then I visualize all above venues to the map:

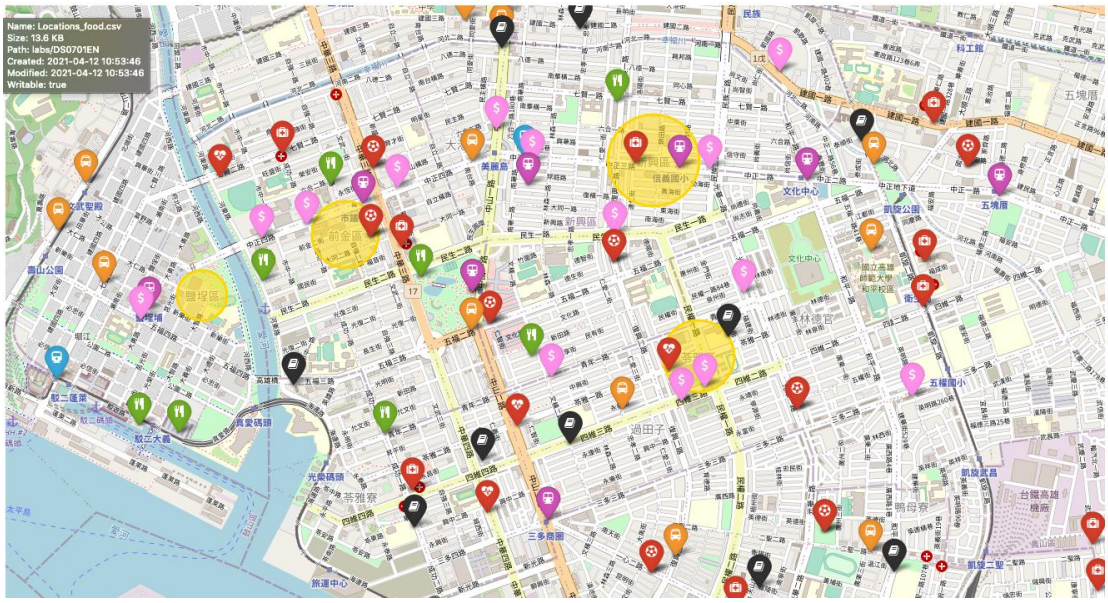


The golden circles belong to the bubble chart of income * population density of every district.

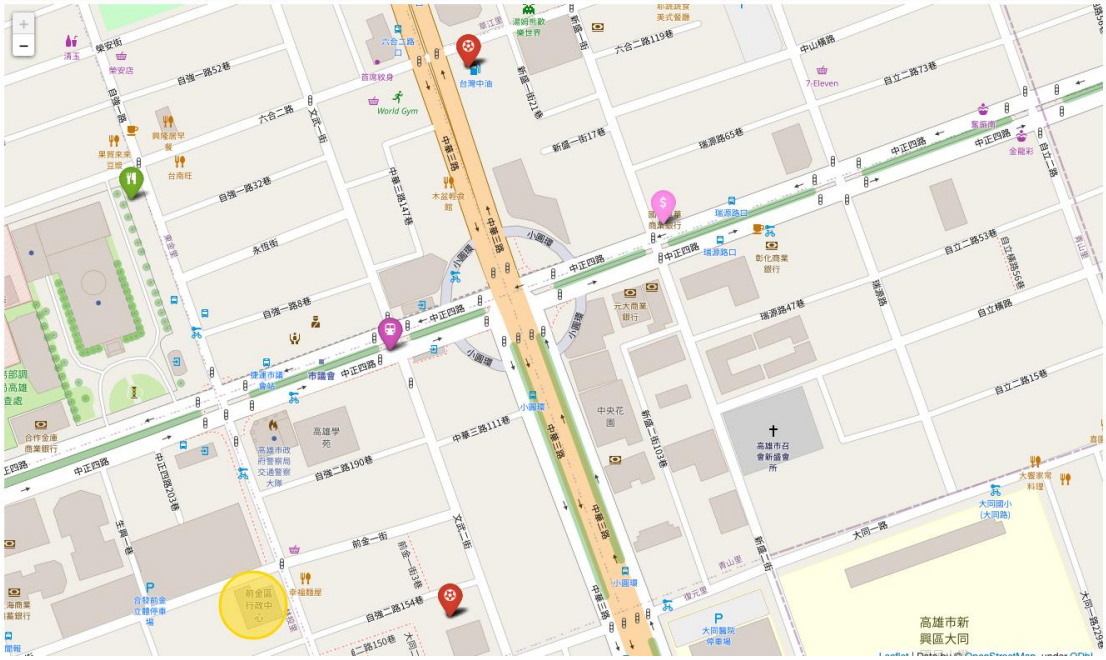
With all analysis above, I am able to select the best place for my restaurant. I leave that part at next section.

Result

Let’s zoom in to the districts with higher income * population density values



As we can see in the picture, top 3 income * population density districts are close to each other. And there is no not-healthy group around, so I decided to locate my restaurant here. After considering all important factors, I selected the final location as the picture below, the roundabout/circle at the center.



This place has several advantages:

1. Residents: Income median here is higher than other districts. And so is the population density.
2. Environment:
    I. A metro station is very close. It will bring me a big stream of people.
    II. In around 300m, there are several banks (potential customers 1).
    III. In around 500m, there are two fitness centers (potential customers 2).
    IV. In around 1 km, there are two elementary schools (potential customers 3).
    V. In around 1 km, there is one hospital (potential customers 4).

Discussion

I would like to discuss about several things I found during the process of this project. I believe all of them are important issues for data scientists.

Although we are working on data science, which should be very scientific and very “not personal”, I think the personal opinion of the data scientist plays an important role. In my project, you can see I personally believe that my restaurant should be close to/far from some place, but it is not always true, maybe some of the readers think my idea is totally wrong. For example, some reader may locate their healthy food restaurant near a place with a lot of fast food restaurants, because the people around there need more healthy food than other places.

The second issues is data accuracy. When I check the visualized map manually, I found that the information Foursquare API provides is somehow not completed or very old. A lot of venues on the map are missing, for example, the light-yellow rectangle at bottom right corner of my last screenshot is a school, but I cannot get it from Foursquare API. This kind of issue would make any analysis less accurate.

The last issue I want to discuss is that, throughout whole project, I used limited data only. In real cases, it’s necessary to consider more. Take my plan for instance, I should also consider the source of my food. Where can I buy a lot of fresh vegetables/meats/eggs every day may change the location of my restaurant. However, I cannot do such analysis since this kind of information is difficult to find, at least very difficult by using only Internet.

Conclusion

In this project, I defined a business problem, and I collected some useful data related to the problem. Then I perform some statistical methods and machine learning algorithms to narrow down my target area. In the end, I selected a very good place for my business problem. It is my first try to this kind of project, and I feel very good about it.

To make our business profitable, we have to analyze our plan again and again. As the goal of this project, selecting a better place is hard enough, but it would only be the first step if we are really going to start a business. It would be a strenuous and even costly process. But it also makes data scientists valuable.