Capstone Project - The Battle of Neighborhoods

---What kind of restaurant should l open in downtown Shanghai?

Issac Ma
March 2, 2021

Introduction: Business Problem

Shanghai, one of the four direct-controlled municipalities in China, has always been described as the "showpiece" of the booming economy only after the United States. As the financial and economic hub of China, Shanghai is a great place offering plenty of opportunities for people who want to succeed and thrive. On the other hand, greater opportunities come with fiercer competitions as starting a new business in Shanghai always means having hundreds and thousands of potential rivals from around the country and the world beyond. Therefore, it is of vital importance to be well-informed and better prepared before large financial resources are pulled in. This is especially true for people thinking of opening a new restaurant in the downtown area of Shanghai as there are just so many different restaurants catering to the various appetites of people working or living in the city center, most of whom could be white collar workers, foreigners and residents who are more likely to be middle-class or above.

This project, through data analysis, tries to provide insights for decision makers by answering the following questions:

- What kind of restaurant should l open?
- Where should my restaurant be located to ensure better profitability?

Data

Data Sources:

- Population and its density in Shanghai by district (preferrable if foreigners are included, including workers and permitted residents)
- Restaurant Info by Foursquare(restaurant type, rate, price range)
- More data may be required as we dig deeper into the data...

We need these data because we want to know:

- what are the most common restaurants in downtown Shanghai?
- After the area is identified, what is its characteristics?
- How about the price range and ratings regarding those popular restaurants?
- more insights could be generated as we dig deeper into the data...

Now we begin searching our data, our prime data source turned out to be:

Shanghai Statistical Yearbook 2019

(http://tjj.sh.gov.cn/tjnj/20200427/4aa08fba106d45fda6cb39817d961c98.html).

We can locate the data we mentioned above from:

2.2 LAND AREA, POPULATION AND DENSITY OF POPULATION IN DISTRICTS (2018)
http://tjj.sh.gov.cn/tjnj/nj19.htm?d1=2019tjnje/E0202.htm

And we find more potentially useful data to help us achieving our goal:
2.5 MIGRATION OF REGISTERED POPULATION IN DISTRICTS (2017～2018)
http://tjj.sh.gov.cn/tjnj/nj19.htm?d1=2019tjnje/E0205.htm

2.6 AGE STRUCTURE OF REGISTERED POPULATION IN DISTRICTS (2018)
http://tjj.sh.gov.cn/tjnj/nj19.htm?d1=2019tjnje/E0206.htm

2.11 RESIDENT FOREIGNERS IN SHANGHAI IN MAIN YEARS
http://tjj.sh.gov.cn/tjnj/nj19.htm?d1=2019tjnje/E0211.htm

GDP Per Papita By District (2019)
http://sh.people.com.cn/n2/2020/1103/c134768-34390634.html

Residents Monthly Salary and Average Housing Prices By Districts(2019)
https://www.tudinet.com/read/15198.html

Note: Some of the data may not come from official source, which could potentially lower the accuracy and objectiveness of our data. We tried our best to find the most reliable data, but it's just difficult to find district-level data since not all of them are made available by the government, or some of them are just not there. Given that most of data won't fluctuate too much within 1-2 years when the economy is generally stable, the trend they show can still generate valuable insights.
So far we have comfirmed the following features(or variables) that could affect the result:

- POPULATION
- POPULATION DENSITY
- MIGRATION OF REGISTERED POPULATION
- AGE STRUCTURE OF REGISTERED POPULATION
- GDP
- GDP PER CAPITA
- RESIDENTS MONTHLY SALARY
- AVERAGE HOUSING PRICE

Next, we also need to find data of neighborhoods in Shanghai by districts.

A neighborhood, by its definition, is an area where people live and interact with one another. Neighborhoods tend to have their own identity, or "feel" based on the people who live there and the places nearby.

In Shanghai's administrative division scheme, the concept of 'neighborhood' is similar to 'sub-district'('街道') , which indicates an area consisting of dozens of residential communities('小区') that is governed by the so-called 'subdistirct office'('街道办事处'), with a series of necessary life infrastructures like small-scale stores, restaurants, vegetable markets('菜市场', biscally equivalent to grocery store), parks, shopping malls etc. Other equivalent bodies include 'town'('镇'), 'township'('乡'), 'county'('县') usually governed by the government at its own level. When people are considering buying a new house in Shanghai, the overall quality of a sub-district is important. This is basically,in westerner's case, finding answers to questions like 'How is this community?', 'Is the neighborhood quiet and friendly?', etc.

However, as there are usually 10-20 subdistricts in one single district, for the ease of analysis, we will consider decompose the district into several representional blocks based on how most real estate information apps in China like Lianjia(链家), Fangtianxia(房天下) decompose a district. This division method is rather arbitrary but it is effective as the sub-divisions you find in these apps do reflect how local residents perceive the composition of the district they are living in conventionally.

Also, because we are focusing our analysis on downtown Shanghai, so we will only consider districts that are considered as the urban core of Shanghai historically, which include:

- Huangpu District
- Xuhui District
- Jingan District
- Changning District
- Hongkou District
- Yangpu District
- Putuo District

So far, l have explained what data we need and why they are important. Next, we will try to obtain all relevant data and manipulate them to eventually form 2 datasets. They are:
dist_info_center, containing all demographic and relevant information regarding the 7 central administrative districts in Shanghai mentioned above.
dist_neigh, containing neighbourhoods in central districts that are arbitrarily defined and their latitudes and longitudes.

These 2 datasets will be our primary data for later analysis.

Methodology
Now that the data is ready, it is time for data analysis. First, we will do some exploratory analyses by exploring the demographic information of the 7 central districts to understand them better. One of the important questions is: How are most people in this district different from the others?

After we have a general grasp of central districts and its people, we will use FourSquare API to fetch information of all restaurants that can be found in central districts, try to cluster them into different groups using K Means Clustering and further break them down using the same clustering method until we eventually find the best neighbourhood to open a new restaurant based on the following principles:

- Try to avoid locations where there are too many restaurants of the same kind, which means fierce competition;
- Prefer places where most residents have a relatively higher consuming capability.
- Prefer places where the rent is relatively affordable.

We will also try to decide what kind of restaurant it is (Chinese, Japanese, Italian, American, etc...) and its price range as going high-end does not necessarily mean high profitability.

Analysis

Based on the previous analyses, preliminary conclusions can be drawn that:
Huangpu, Hongkou and Jing'an, the 3 nearest districts to center of Shanghai tend to have more young people, higher population density and limited land areas. Meanwhile, Xuhui, Changning, Yangpu and Putuo tend to have more children and senior citizens within their respective areas. Different business strategies could be applied to these 2 cohorts when we are considering to open a new restaurant.

Yangpu, Xuhui and Jing'an seem to be more popular among people looking to settle and start a new life in downtown Shanghai, with larger population inflow from 2017 to 2018. Notice that housing prices in Yangpu are significantly lower than the other two, indicating that Yangpu could be the most affordable district for one to realize the dream of living in downtown Shanghai.

Yangpu, Putuo and Xuhui are the most populous areas among all central districts.

Taking housing prices into account, the 3 most affordable districts are Putuo, Hong'kou and Yangpu while residents in Putuo and Hongkou tend to have greater purchasing power than Yangpu residents. Meanwhile, Jing'an, Changning and Xuhui are the 3 most expensive districts with proportional purchasing power of residents in the area.

High GDP per capita does not necessarily indicate high income. While Jiang'an and Changning, both with high GDP per capita and high monthly salary, live up to people's presumption that high GDP per capita generates high income, Hongkou and Putuo stand out to prove that this is not always true with their average monthly salary close to 8,000 RMB. This could be an interesting point in later analysis and decision-making.

For the later part, see details in the Powerpoint which are presented with all necessary pictures.

Results and Discussions

Now that we have already decided to open a new Japanese restaurant in Changfeng neighbourhood, let's see if we could find more insights from everything we have done so far to ensure our business could survive and achieve maximum profit.

On the bright side, we are lucky to find the answers to the questions we ask at the beginning of this project. And the rationale behind how we land our final decision seems all reasonable. We also find a lot of interesting stories about the 7 districts and their residents in Shanghai.

However, we did not come to this stage easily, so let's discuss the limitations of our findings and any major problems we have encounted during this whole project.

First, the results are based on the data we gather from multiple channel, some of which official and others not. This means our results could be biased and far from executable since everything we've done in this project is based on data that is uncheck and a lot of assumptions that seem to be logically correct. Everything we have achieved in this project only serve to provide insights for us before we actually make a real business decision. And there are certainly more work to do like field explorations(average rents, price range), consulting experts to ensure our business stay healthy and thrive.

Second, we certainly spent a lot of time gathering, cleaning and formatting data before data analysis actually begun. This makes sense since everything we have done in this project is based on data and we need to try our best to ensure that the foundation of our project to be real and accurate and complete.

Third, we certainly see that the FourSquare API is far from perfect since even if we tries our best to ensure the name our neighbourhoods to be correct, we still have certain number of missing coordinates.

Finally, we should remember that results in this project need to be updated from time to time since our data are changing every day. We should always think more in terms of what tools to use to solve our problems. We should always be flexible to switch between tools and always ask the two most important questions in a data science project: Does this tool actually help me achieve what l need? Can l achieve the same goal more efficiently by using different tools or different methods?

Conclusion

In this project, we utilized multiple data science tools like Python, MySQL and FourSquare API to find out the best neighhouhood in downtown Shanghai to open a new restaurant. Along the way, we also generated useful insights by analyzing the demographic information of 7 central districts in Shanghai. By going through the full

cycle of data science using multiple tools, we formed a deeper understanding of what data, programming and business are all about.