IBM Capstone Project Assignment

Project: Where to locate a new Japanese restaurant in Munich?

1. Introduction

1.1 Background

Eating out is a popular pastime in Germany, with almost 51.3 million people sometimes doing so in 2019, and roughly 11.9 million people stating they often did so.[1] Many young people dream about starting their own restaurant, someday. The size of restaurant industry market is continuously increasing,

Jack lives at Munich. He has a strong interest in restaurant industry, and he has undergone a long process of training in Japan to learn how to cook Japanese cuisine. Now, he decided to start his own Japanese restaurant at Munich.

1.2 Business Problem

Munich is the capital and most populous city of Bavaria, the second most populous German state. With a population of around 1.5 million, it is the third-largest city in Germany. As a global city, Munich has a lot of restaurants that offer German, English and other continental cuisine. In these several years, with more and more drastic competition, the location of food service business will impact its success nearly as much as the menu.

Jack worries a lot about the location of his new restaurant, and he finds a market analysis agency to help him pick a district to start his new business.

2. Data

2.1 Data sources

There are three important factors in choosing a location for Jack's new restaurant, respectively, population base, foreigner ratio and venues categories within each district.

First, the population base data is required to make sure that there are adequate people to support his business. There are 25 districts in Munich with different population density. Those area with high population density should be preferred.

Second, the ratio of foreigner data is desired to estimate the potential customers for a Japanese food. Here we assume that the higher the ratio of foreigner population, the more popular the restaurant is in that district.

[1] Data source: https://www.statista.com/statistics/561124/eating-out-frequency-germany/

Third, the venues data at each district should be analyzed to determine the competitive level of restaurant industry in that area and estimate the restaurant exposure potential to pass-by customers.

2.2 Data acquisition and cleaning

In order to get above data, this research use web scrape to collect useful information from public website.

Impact Factors	Data
Population base	District data [1]
Foreigner population ratio	Foreigner ratio data[2]
Venues categories within each districts	Venues geographical data[3]

Data source:
1. District data
https://www.muenchen.de/int/en/living/postal-codes.html
https://www.all-zipcodes.com/postcode-munchen-in-germany
2. Foreigner data
http://www.total-munich.com/20160623888/blog/moving-to-munich/moving-to-munich-introduction-to-munich-s-boroughs.html
3. Venues data
https://api.foursquare.com

Firstly, we scrape the Munich district information, including district, surface, population, density and foreigners% from the district data source. However, within this dataframe, we don't have any geographical information like latitude and longitude. In order to retrieve the coordinates of these districts, we need the postal code for each district first.

	District	Surface (km²)	Population	Density (Persons/km²)	Foreigners (%)	latitude	longitude	Postal_Code
0	Allach-Untermenzing	15.45	31,353	2,029	19.5	0	0	80995
1	Altstadt-Lehel	3.15	20,806	6,614	25.8	0	0	80331
2	Au-Haidhausen	4.22	60,756	14,399	23.4	0	0	81541
3	Aubing-Lochhausen-Langwied	34.06	42,859	1,258	23.3	0	0	81243
4	Berg am Laim	6.31	44,002	6,971	29.9	0	0	81671

Secondly, according to postal code, we get the coordinates for each district. Since there is no public coordinates table, we need to search the coordinates for each district by using loop. The return result of this html is quite messy, the text manipulation is applied to attract the useful geographical information.

```
import requests
for i in range(0,Munich_df.shape[0]):
    post_data=Munich_df['Postal_Code'][i]
    URL3='https://www.all-zipcodes.com/postcode-munchen-in-germany-'+post_data+'/' #post_data is updated at each row
    r = requests.get(URL3)
    soup3 = BeautifulSoup(r.content, 'html.parser')
    content_list=soup3.find_all("script")[2].text
    Munich_df.loc[i,'latitude']=content_list.split(";")[5].split('setView')[1].split(',')[0].replace(r'([',"")
    Munich_df.loc[i,'longitude']=content_list.split(";")[5].split('setView')[1].split(',')[1].replace(r']',"")
Munich_df.head()
```

	District	Surface (km²)	Population	Density (Persons/km²)	Foreigners (%)	latitude	longitude	Postal_Code
0	Allach-Untermenzing	15.45	31,353	2,029	19.5	48.217157311	11.5149477676	80995
1	Altstadt-Lehel	3.15	20,806	6,614	25.8	48.1359642378	11.5729048503	80331
2	Au-Haidhausen	4.22	60,756	14,399	23.4	48.1204341448	11.5874447961	81541
3	Aubing-Lochhausen-Langwied	34.06	42,859	1,258	23.3	48.1452756567	11.4368110446	81243
4	Berg am Laim	6.31	44,002	6,971	29.9	48.1219040581	11.6182992419	81671

Thirdly, by using latitude and longitude for each district, we use Foursquare API to acquire the top 100 venues of within 1000m radius of each district. Within "Munich_venues dataframe", we get the detail information of venues within each district. These features can be used for clustering for later analysis.

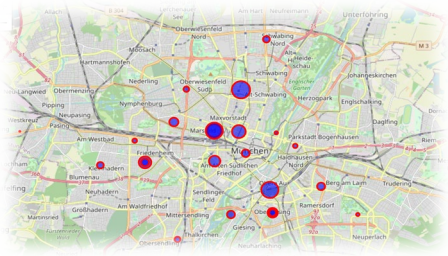
	District	latitude	longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allach-Untermenzing	48.217157311	11.5149477676	Seehaus Feldmoching	48.216965	11.517452	Greek Restaurant
1	Allach-Untermenzing	48.217157311	11.5149477676	Feldmochinger See	48.214687	11.514851	Lake
2	Allach-Untermenzing	48.217157311	11.5149477676	Andy's Seehäusl	48.215186	11.512046	BBQ Joint
3	Allach-Untermenzing	48.217157311	11.5149477676	Grüne Insel	48.215550	11.510013	Beer Garden
4	Allach-Untermenzing	48.217157311	11.5149477676	Hanika Autosattlerei	48.218914	11.521861	Auto Workshop

After this step, the tidy data work is finished. The *"Munich_df"* and "*Munich_venues"* dataframe will be used for further analysis.

3. Exploratory data analysis

3.1 Analyze the population data in Munich

Since each restaurant has its limited radius of service, instead of exploring the total population at each district, we pay more attention to the population density. Higher population density means that the restaurant can cover more potential customers within its radius of service. We create a map to show the district density on map. The larger the circle is, the higher the population density is.



In order to select the districts from high population density region, we store the top 10 density (Persons/km²) into a data frame named "*Density_District_candidate".*

3.2 Analyze the foreigner's ratio

Since Japanese food is an international food for Munich locals, a more diversity district could be more friendly for an external food culture. Thus, we assume that the higher ratio of foreigner is, the better the location is for Jack's new restaurant.

In order to select the districts from high foreigner(%) region, we store the top 10 foreigner(%) into a data frame named "*Foreigners_District_candidate".*

3.3 Analyze the restaurant industry in Munich

Besides the population data, we need to get an overview how restaurant industry is in Munich to get a better business sense. First, we draw a Pareto chart to observe the ranking of different countries' restaurant. As we can see, the diversity of restaurant is quite high in Munich, almost including the food around the world. It can be inferred that the restaurant industry is quite competitive in Munich.

But according to the ranking, the Japanese restaurant (sushi restaurant) only ranks at 9th in Munich. It means it has a relatively low market share and it is a positive signal that there is still a huge growth potential for Japanese restaurant. With 15 Japanese restaurants, we also needs to keep in mind that the new restaurant should keep a distance with existing Japanese restaurant to mitigate the competition.

By analyzing locations of existing Japanese restaurant, we find that the majority of them are located at the northwest side of Munich. So, it would be better to locate Jack's new restaurant in other side of

Munich. Besides, we would like to figure out the pattern of district at which these successful Japanese restaurants are located. Finding a similar district to locate a new restaurant will greatly reduce the risk of selecting a wrong place because all these patterns are valuable experience from past restaurant owners. For this topic, we will cover it in the methodology part by implementing the clustering method to figure the venues patterns.

Figure: The location of existing Japanese restaurant

4. Methodology: clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Since we don't know the exact pattern of districts that have Japanese restaurants, the unsupervised learning method is implemented here. Here, we choose the K-means clustering to group districts at which Japanese restaurants are located.

By trying different size of K, in the end, we set K=3 to group these districts. Based on clustering result, we select the top 10 venues within each district, there are three representative clusters.

Pattern I

Within first cluster, there are plenty of restaurants within each district. This could be agglomerative effect, high density of different restaurants can attract more customers and all these restaurants benefit from it.

```
1 district_venues_sorted_Japan[district_venues_sorted_Japan['Cluster']==0]
```

	District	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	Cluster
0	Bogenhausen	Italian Restaurant	Café	Hotel	German Restaurant	Restaurant	Plaza	Bar	Bakery	French Restaurant	Ice Cream Shop	0
3	Moosach	German Restaurant	Italian Restaurant	Café	Bakery	Plaza	Trattoria/Osteria	Gastropub	Restaurant	Concert Hall	Middle Eastern Restaurant	0
4	Neuhausen-Nymphenburg	German Restaurant	Italian Restaurant	Indian Restaurant	Café	Supermarket	Plaza	Sushi Restaurant	Vietnamese Restaurant	Bakery	Hotel	0
5	Schwabing-West	Italian Restaurant	Café	Vietnamese Restaurant	Trattoria/Osteria	Plaza	Greek Restaurant	Pizza Place	Gastropub	Thai Restaurant	Bar	0

Pattern II

Within second cluster, there are plenty of public transportation station and supermarkets. All these venues contribute high number of customers passby, which greatly increase the marketing exposure to potential customers.

```
1 district_venues_sorted_Japan[district_venues_sorted_Japan['Cluster']==1]
```

	District	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	Cluster
1	Laim	Supermarket	Bus Stop	Tram Station	Laundromat	Bank	Hotel	Pizza Place	Sushi Restaurant	Sandwich Place	Fast Food Restaurant	1
8	Sendling-Westpark	Supermarket	Bus Stop	Tram Station	Laundromat	Bank	Hotel	Pizza Place	Sushi Restaurant	Sandwich Place	Fast Food Restaurant	1

Pattern III

Within third cluster, there are a great number of hotel an café within the district.

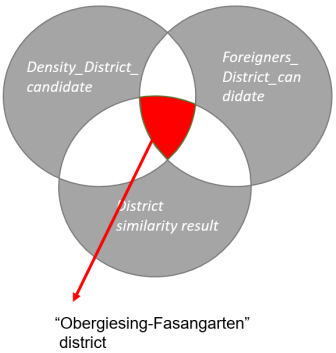
```
1 district_venues_sorted_Japan[district_venues_sorted_Japan['Cluster']==2]
```

	District	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	Cluster
2	Ludwigsvorstadt-Isarvorstadt	Hotel	Café	Italian Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant	German Restaurant	Steakhouse	Burger Joint	Restaurant	Drugstore	2
6	Schwanthalerhöhe	Hotel	Café	Italian Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant	German Restaurant	Steakhouse	Burger Joint	Restaurant	Drugstore	2
7	Sendling	Hotel	Café	German Restaurant	Bavarian Restaurant	Vietnamese Restaurant	Burger Joint	Italian Restaurant	Middle Eastern Restaurant	Asian Restaurant	Hotel Bar	2

5. Result

As mentioned at data source part, there are three important factors in choosing a location for Jack's new restaurant, respectively, population base, foreigner ratio and venues categories within each district. An ideal location should take all above factors into consideration, in other word, it should satisfy all above criteria.

First, we calculate the intersection set of Density_District_candidate and Foreigners_District_candidate, which respectively represent the top 10 candidates for each crucial factor. The "Obergiesing -Fasangarten" district is the only one ranks in top 10 in both criteria. But it doesn't mean this is a good place to locate a Japanese restaurant. We need to compare the venues within this district with existing district venues pattern for Japanese restaurant.



By comparing the Top 10 venues in “Obergiesing -Fasangarten” district with existing pattern we get in methodology part, we find that this candidate district share similar venues with Pattern I and Pattern III. In other word, it proves that “Obergiesing -Fasangarten” is ideal place to open a Japanese restaurant. In “Obergiesing -Fasangarten” district, there are a lot supermarket, hotel as well as Café. All these venues make Jack’s new restaurant get a high exposure to its potential customers. Besides, the similar pattern of surrounding venues could highly reduce the risk of selecting a wrong place cause the past experience has proven that Japanese restaurants are suitable for these similar district.

```
1 district_venues_sorted[district_venues_sorted['District']=='Obergiesing-Fasangarten']
```

	District	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
14	Obergiesing-Fasangarten	Supermarket	Hotel	Café	Greek Restaurant	Ice Cream Shop	Italian Restaurant	Bar	German Restaurant	Drugstore	Plaza

	District	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	Cluster
1	Laim	Supermarket	Bus Stop	Tram Station	Laundromat	Bank	Hotel	Pizza Place	Sushi Restaurant	Sandwich Place	Fast Food Restaurant	1
8	Sendling-Westpark	Supermarket	Bus Stop	Tram Station	Laundromat	Bank	Hotel	Pizza Place	Sushi Restaurant	Sandwich Place	Fast Food Restaurant	1

Pattern II

	District	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	Cluster
2	Ludwigsvorstadt-Isarvorstadt	Hotel	Café	Italian Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant	German Restaurant	Steakhouse	Burger Joint	Restaurant	Drugstore	2
6	Schwanthalerhöhe	Hotel	Café	Italian Restaurant	Vietnamese Restaurant	Middle Eastern Restaurant	German Restaurant	Steakhouse	Burger Joint	Restaurant	Drugstore	2
7	Sendling	Hotel	Café	German Restaurant	Bavarian Restaurant	Vietnamese Restaurant	Burger Joint	Italian Restaurant	Middle Eastern Restaurant	Asian Restaurant	Hotel Bar	2

Pattern III

6. Conclusion

Based on the analysis, by taking all crucial factors impact the location selection into consideration, we recommend Jack to open his new restaurant at “Obergiesing-Fasangarten” district. Here, the population density is high, and the ratio of foreigners is higher than majority districts. Besides, the venues within this district will bring a large number of potential customers passerby. The high exposure to customers commits a good business. In term of horizontal competition, this site is far from existing Japanese restaurants. All in all, we strongly recommend Jack to locate his new restaurant at this district.

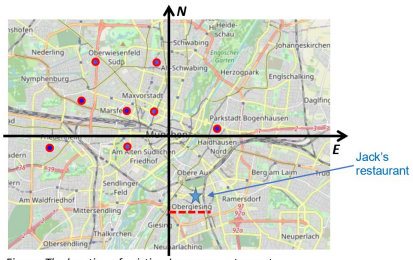


Figure: The location of existing Japanese restaurant

Further application of analysis

Besides selecting the location for restaurants, this analytical approach is appliable for other location decision making process. In this research, we take industrial information, geographical information and population into consideration. All these data can be retrieved from the public website.

Limitation

Due to the availability of data, we can only give the suggestion about location in district level. If the available storefront information is available, we can do further analysis to compare different sites within this district and give Jack more specific address to locate his new restaurant.