Chisinau, Iasi, Odessa - 3 sister cities venues & neighborhood Analysis

# Business Problem

Initially I came with the idea to analyze road load statistics, cluster gas stations and give with a recommendation where to build a new one or an existing one isn’t necessary.

Since on Foursquare there are mostly point locations and road vectors are not present, I thought to take another challenge: to analyze the venues from the city I live in, Chisinau and compare them to other 2 major cities from nearby area, its sister-cities Iasi (Romania) & Odessa (Ukraine). Besides, I made also an analysis of the neighborhoods of these cities.

In this project, I’ll address the following questions:

* Which are the most venue types for these cities?
* How are they distributed among the neighborhoods?
* How are they segmented base on the similarity of the most common F&B venues?
* What are the location and type of cuisine recommendation for opening a new restaurant?

# Who would be interested?

* Start-ups who want to open a new restaurant.
* Investors who want to look into F&B business.
* City planners/ F&B Managers who need an overview on F&B venues distribution.

# Data

In this project I will use the following data source:

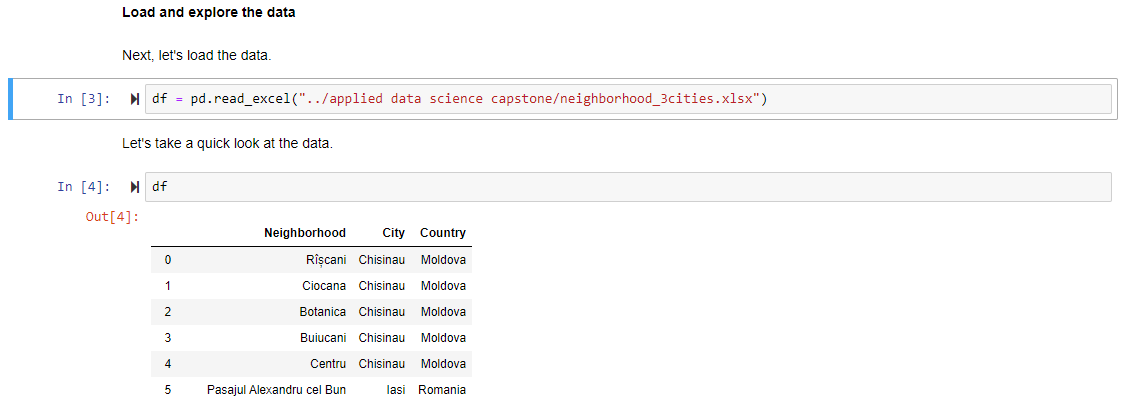
* I collected all neighborhood/sectors data from various data sources, but mostly on Wikipedia.
* Location (latitude/longitude) of each ward: using Geocoder on each ward location.
* FourSquare API to generate nearby venues for each location. Which represent venues in that ward area. I used the special category ID for F&B venue.
* K-means Clustering were used to cluster all F&B venue base on its category. Each cluster is then analyzed for location recommendation.

# Methodology

## 3 cities data - Web scraping

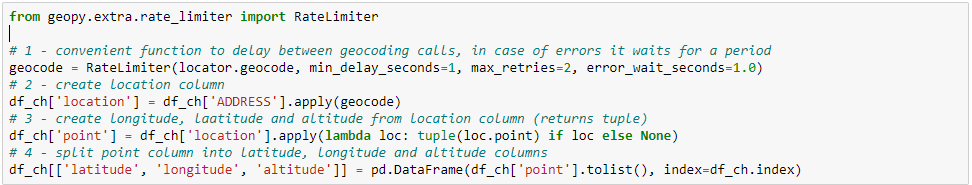
You can check the data at: <https://github.com/boombasteek/ADSC_COURSERA/blob/master/neighborhood_sector_3cities.xlsx>.

I’ve used read\_excel instead of BeautifulSoup because it’s faster and more efficient. The list contain more than 33 neighborhoods, all are available in Foursquare.

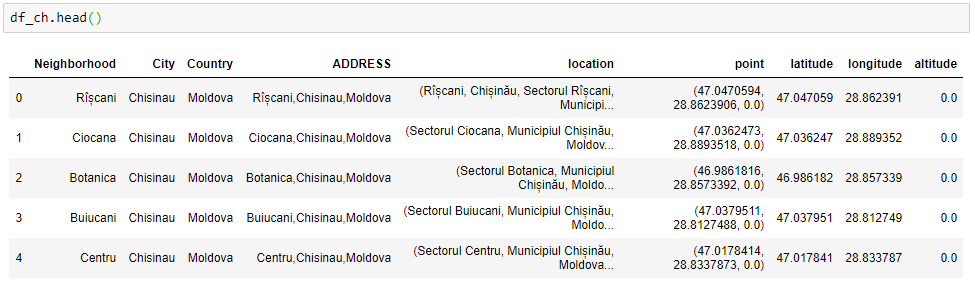


## Insert location data from geopy

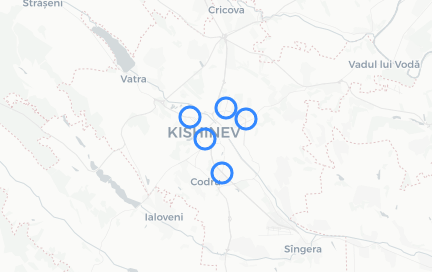
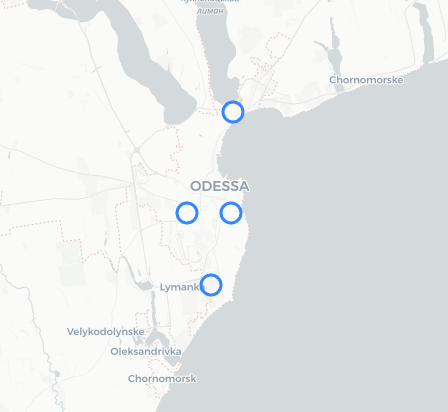
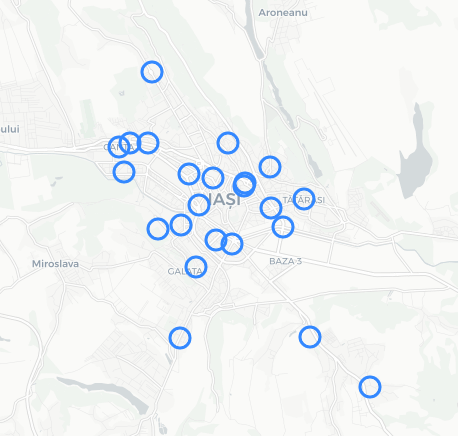
Next, I’ll get the coordinates of all venues by apply the geolocator loop. I will use RateLimiter, which is a convenient function to delay between geocoding calls, in case of errors it waits for a period.



Because FourSquare search engine is not sufficient for special language such as Vietnamese, several coordinates were not returned. I had to make some adjustment to the address name, both manually and by code loop. The final coordinates table is as follow:



Let’s visualize the location of neighborhoods within the city & a birds eye view of them:

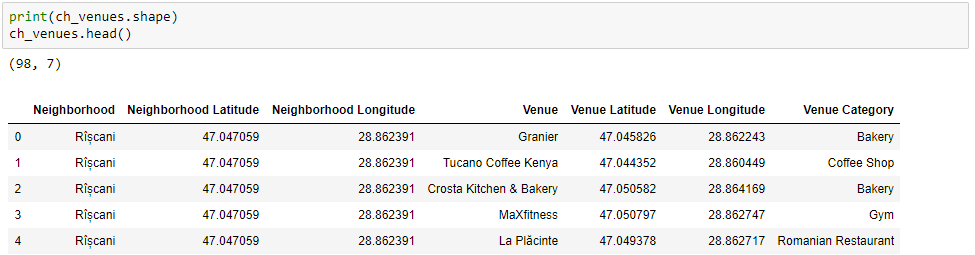
  



Get F&B venues in each ward using the “getNearbyVenues” function in the Lab sections. I made some changes as follow:

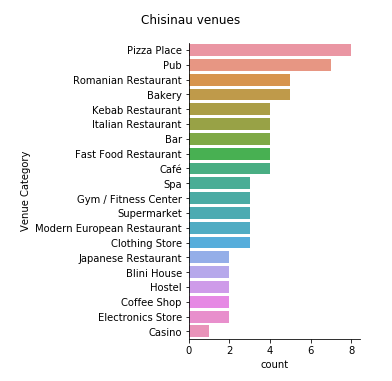
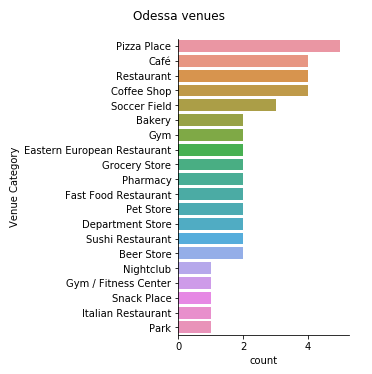
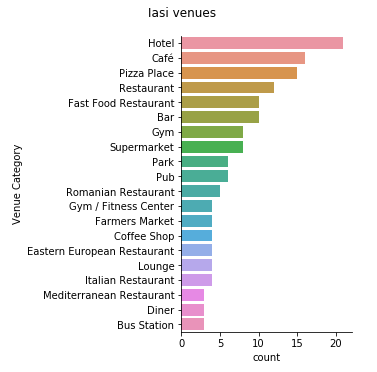
* I defined the radius of surrounding area as 2Km because it’s an average distance between each neighborhood, to make sure that I don’t miss any venue.

The final venues table is as follow:



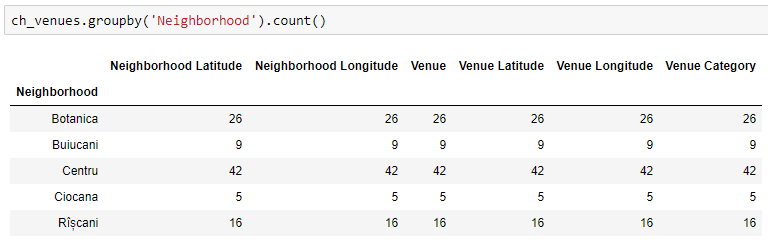
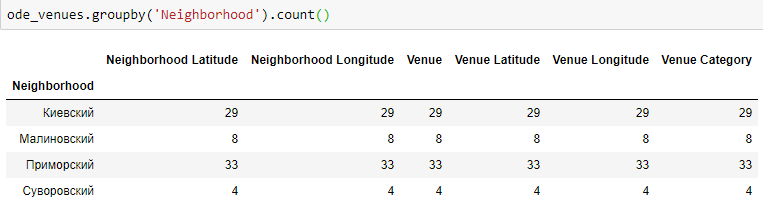
Let’s take a look at our data by finding out how many unique categories are there and frequency of each. I’ll plot the 20 most common F&B venues:

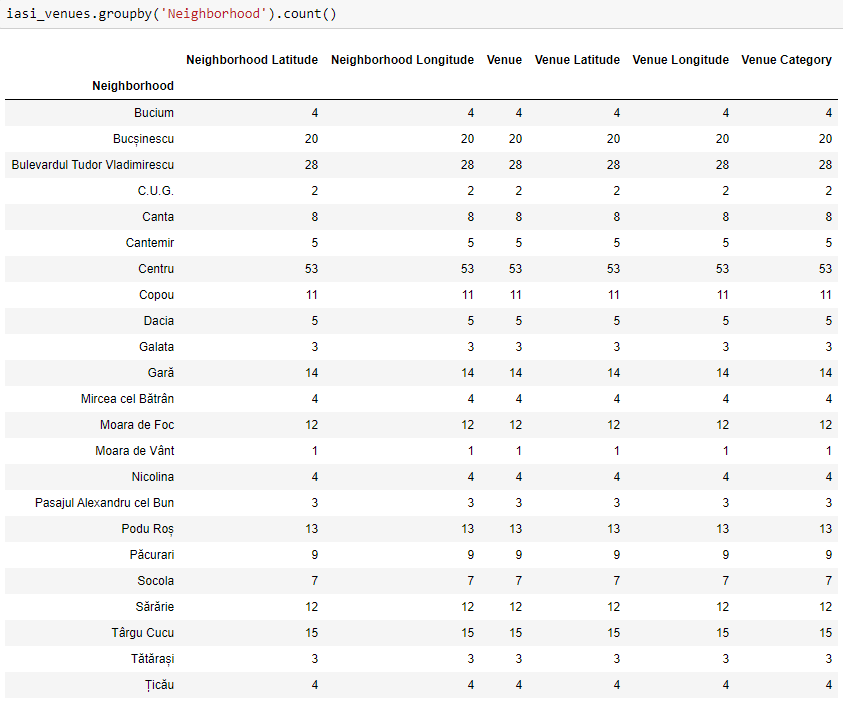


Pizza place is the most common venue in Chisinau & Odessa city, while in Iasi hotel is the most.

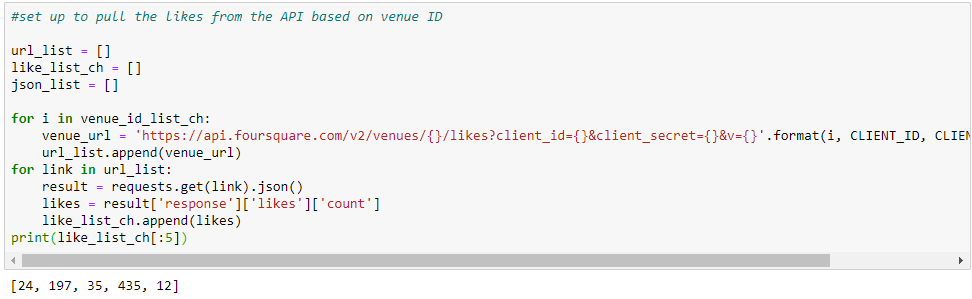
Let’s take a look at how some of the venue types are distributed over neighborhoods for the 3 cities:

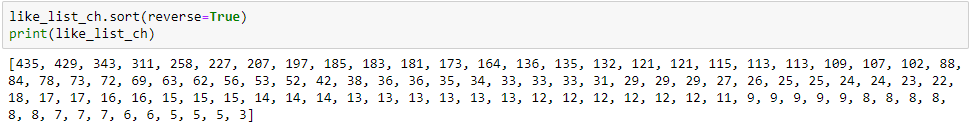


We can see that as expected, the central areas of each city have the most venues.

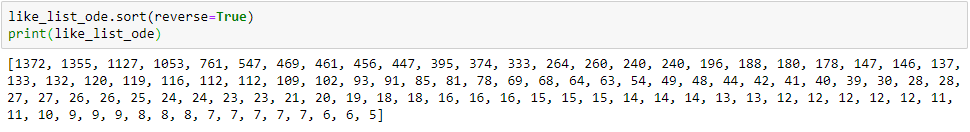
Now let's see the venue rating in the central cities neighborhood:



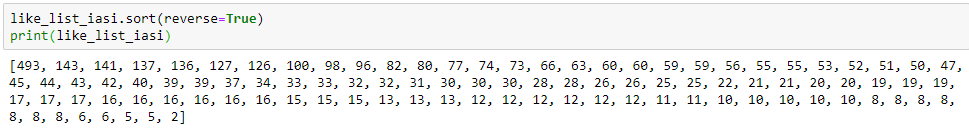
* Chisinau



* Odessa



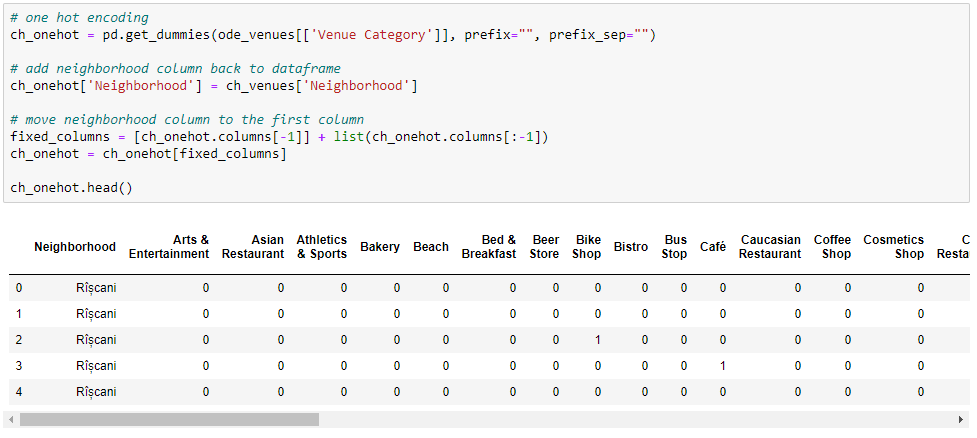
* Iasi



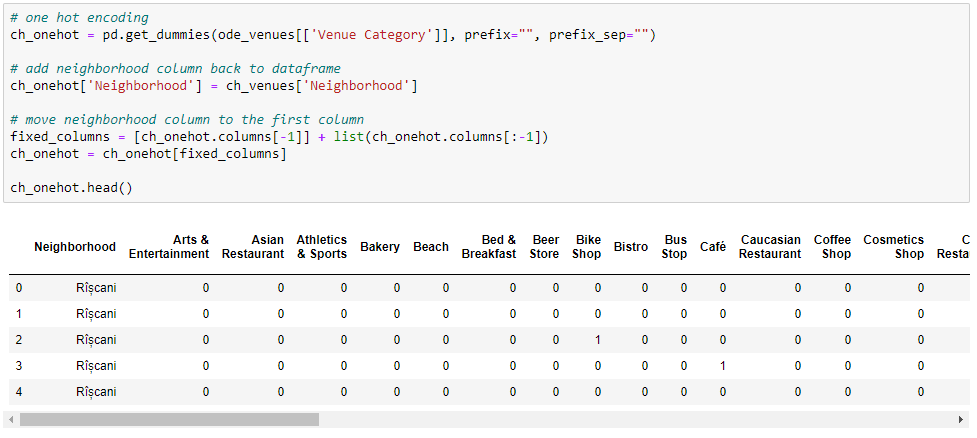
Next, we will see whether the venues that are being present in a city are also in the other one. For each type of F&B venues there are certain district with relatively less venue count than others, which are the potential location for new restaurant with less competition. However, because we’re looking at venues in each district, which are the “clusters” where borders were not originally created for our venue purposes. Therefore, let’s work on clustering these venues using machine learning (in this case, k-means clustering) to get better results.

## Preprocessing data for K-means Clustering

Create one hot encoding mask



Calculate the means of frequency of occurrence of each venue type in each ward.



Create a data frame and display the least 10 common venues from Odessa for each neighborhood from Chisinau

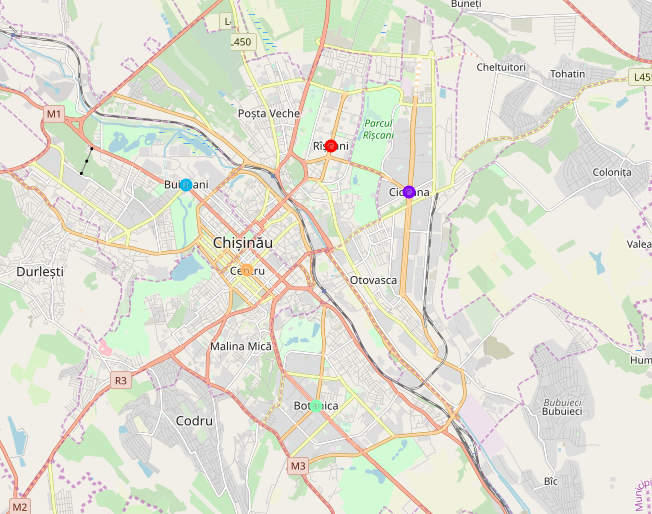


## K-means Clustering

I used this following code to choose the best K for K-means Clustering

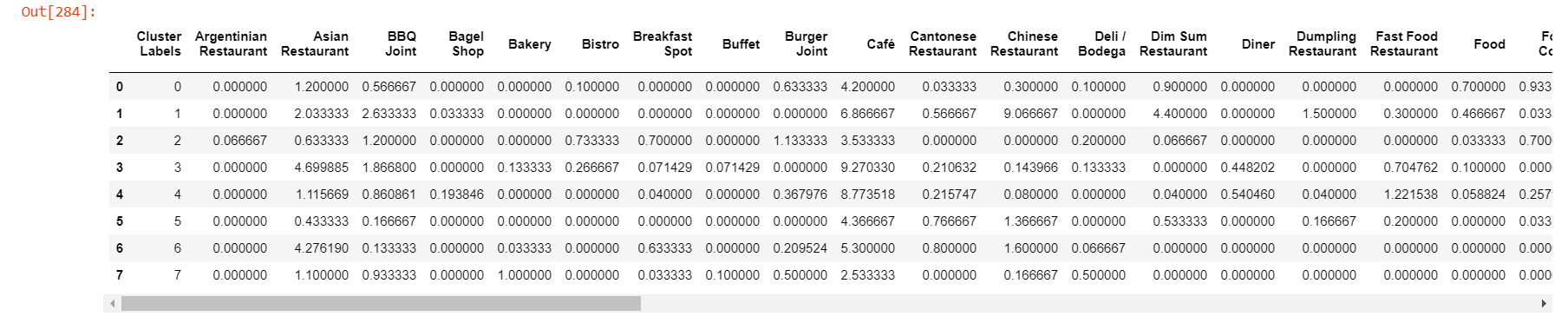


The larger the K, the more accurate the clustering method is, thus I chose K=5.

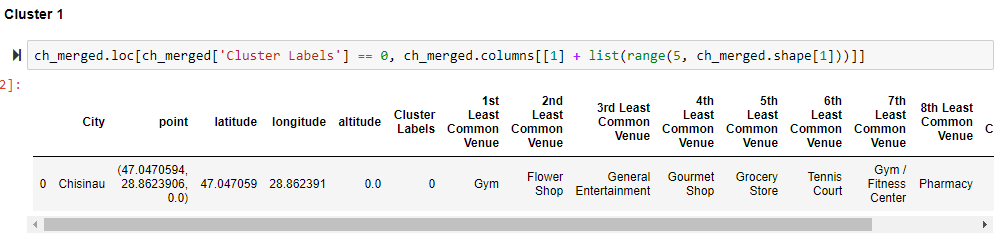


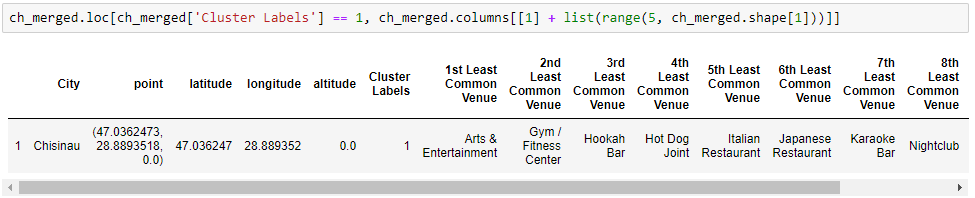
## Analyzing each cluster

Cluster labels were added to each single venue as follow:



# Now, we can examine each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, you can then assign a name to each cluster.





# Result & discussion

We’ve gone through step by step analysis of F&B venue location data in 3 sister cities – Chisinau, Iasi & Odessa. The idea was to get the overall picture of F&B venues location & types, cities neighborhoods and give recommendation to open new venue type that is already being present in one of these cities, but not in the other. By using web scraping, geopy library and Foursquare API, I’ve found out the following:

* Pizza place is the most common venue in Chisinau & Odessa city, while in Iasi hotel is the most.
* From a simple comparison of likes count of the venues in 3 cities, we can clearly see that in Odessa are most active users; while in Chisinau and Iasi, they are aproximately the same user patterns. Next, we will see whether the venues that are being present in a city are also in the other one.
* Number of clusters to be used in Chisinau is the same as the initial ones, as this corresponds well to geographic borders of them

From the outputs, Grocery Store is the venue to be deployed in Chisinau. As a city local, I can tell there is an issue with the recommendation. It is caused by the fact that there is not enough info on such venues in Foursquare.

# Conclusion

Although the Foursquare data is still limited to a certain level, the location analysis gave me an overview on how F&B venues are distributed among the 3 cities. Beside venues segmentation, this analysis can also be used for insights on specific type of venues that are planned to be opened. Despite some drawbacks and room for improvement as mentioned, some of the result suit the real-life observation. I hope this will be valuable for all those who are interested in F&B venues, and as an example of applied data science in actual problems.