Coursera Capstone Project in Applied Data Science

Opening a new Chinese restaurant in Milan

Alessandro Bello

January 20, 2021

Contents

1 Introduction

People in Italy love to dine in restaurants and enjoy the moment of being in touch with other ones. But since the COVID-19 pandemic striked into the country things started to get out of control. Fake news were popping out ever more often. Sadly, most of them were about Chinese people. Demonizing and putting them on the edge. To the point were, not only shops, but even restaurants started to close [1] and have fewer and fewer reservations [2] hitting a -43% decrease due to all these fake news.

In Italy there was an 18% increase in people coming from Mainland China from 2013 to 2019 [3]. So the interest in buying properties, and to spend their life in Italy, is actually rising. Therefore we can bet that there will still be people that want to open a new restaurant in a wonderful city like Milan. Opening a business in such a big city can be quite burdensome, given that the position is crucial for two reasons. Firstly, it gives visibility to the place in such a way to keep up with the competition. Secondly, since Milan it's one of the biggest city in Italy, with approximately 1.4 million of residents, and an average density of about 7000 ab/km^2, the housing costs can skyrocket the closer the place is to the city center. Thus the location is actually one of the key parameters to look after.

1.1 Business Problem & Target Audience

The objective of this project is to find the best location in the city of Milan to open a new Chinese restaurant. We are going to use some machine learning techinques, such as K-means for clustering, and some Data Science methods to clean up and prepare the dataset. Thus this project is aimed not only to business owners, but to investor too that could ask themselves *"Where should I open my traditional chinese cusine restaurant? "*

1.2 Data

In order to solve this problem we need:

1. List of neighborhoods in Milan. This is the focus of the project. Milan is one of the most multi-cultural cities across Italy, situated in Nothern part;

2. Latitude and longitude of these neighborhoods to plot a map and get the venues data;

3. Venues data about the actual number of chinese restaurant in the city. This will allow us to cluster the neighborhoods and get a classification based on the presence or not of similar restaurants.

We will use some web scraping techniques to get the neighborhoods from a Wikipedia page [4]. Using the beatifulsoup library we will handle the request and then clean it. Afterwards we are going to use the Python Geocoder package to associate to each neighborhood a precise latitude and longitude. Finally we will use Foursquare APIs to get the venues data that we need proceding thus to use K-means to cluster the data. All of this will be represented graphically using Folium.

2 Methodology

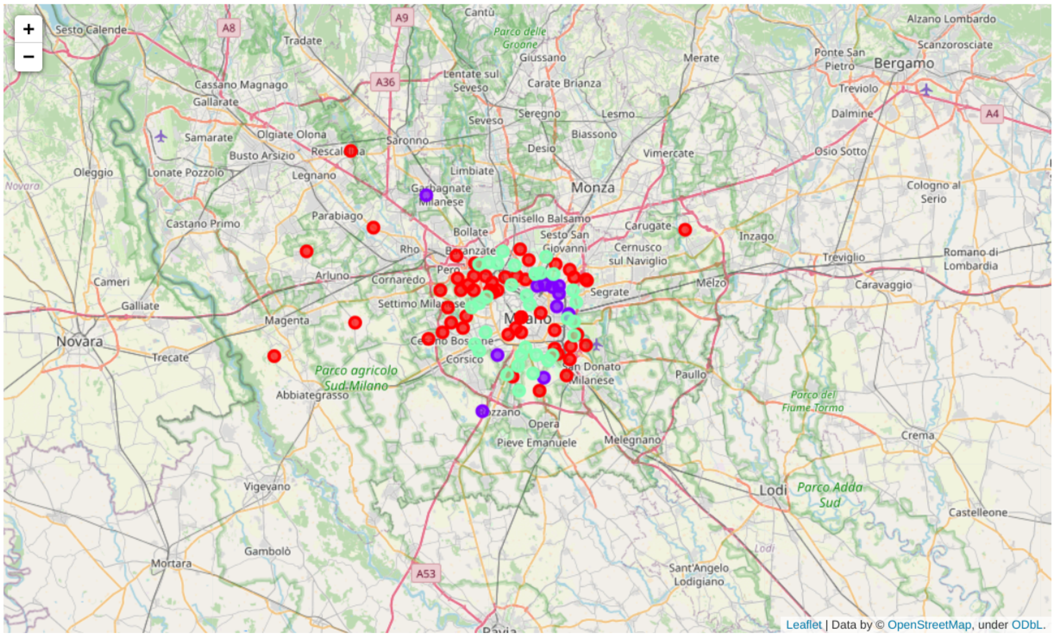
Firstly, we need to get the list of neighbourhoods in the city of Kuala Lumpur. Fortunately, the list is available in the Wikipedia page in the reference [4]. We are going to do some web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. We need to get then the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. In order to do that, we will use the Geocoder package. This will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering all the data, we will parse them into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to also check that the coordinates are actually faithful and compare them in the city of Milan. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We just need our Client and Secret ID precedently created. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for the clustering. Since we are analysing the "Chinese Restaurant" data, we will filter for this string as venue category for the neighbourhoods. Finally, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. In order to decide the optimum value k we will use the *elbow method.* Then, we will cluster the neighbourhoods into k clusters based on their frequency of occurrency for "Chinese Restaurant". The results will allow us to identify which neighbourhoods have a higher concentration of those restaurants and which neighbourhoods have the fewest. Based on the occurrence of chinese restaurant in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to opening a new one.

3 Results

The result of the elbow method gave as the optimal k the value 5. I choose the value $k = 3$ due to the empirical fact that when I inspected the clusters of neighborhoods with k=4, 5, there were classes with only one restaurant in very improbable location (on the far West site of the city of Milan). From the k-means clustering we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for "Chinese Restaurant":

- Cluster 0: Neighbourhoods with virtually no chinese restaurants;
- Cluster 1: Neighbourhoods with high number of chinese restaurants;
- Cluster 2: Neighbourhoods with medium concentration of chinese restaurants;

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in violet, and cluster 2 in mint green.



Leaflet | Data by © OpenStreetMap, under ODbL.

4 Discussion

Looking at the results of the clustering we can clearly see how the restaurants are spread across the city. The 0 labeled cluster represents the neighborhood with virtually no Chinese restaurants open. The 2 labeled clusters represents a medium concentration and 1 the highest one. More than 50% of the datas are labeled as 0. This gives us create opportunities when we want to open a new local. Moreover many of the "free" places are in an interesting position in the northern part of the city with little to no competition. Finally the we can see a higher concentration of similar restaurants in the North-East part of the city and thus should be avoided in order to have the least competition.

4.1 Critique

Obviously there are many limitations to this study. The assumption that one investor/entrepreneur should base its investment/opening on a new venture, such as opening a chinese restaurant, based only on the location it's silly. There are many more variables that have to be taken into account for such as: rent, expenses, fees etc. Future research should take into account even these factors. One thing to mantion also is the restricted number of API calls that can be done using a Free account in Foursquare. This is not the most limiting factor, but it is one of them [5].

5 Conclusions

In this project, we have gone through the process of identifying a business problem, specifying the data required, extracting and cleaning the data, using a machine learning algorithm (clustering the data into 3 clusters based on their similarities), and lastly providing recommendations to the relevant stakeholders like entrepeneurs and investors regarding the best locations to open a new restaurant. We can now answer the question at the beginning of our journey saying that the neighbourhoods in the cluster 0 nearby cluster 2 are the most attractive locations to open a new chinese restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions.

References

[1] *Milanotoday* Coronavirus psicosis, chinese restaurants obliged to close

[2] *The Fork* Drastic decrease in chinese restaurants closure

[3] ISTAT Data of chinese population in Italy

[4] Wikipedia source data for Neighborhoods

[5] Foursqaure API