Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Pub in Milano

April 2021

Introduction

Pubs are one of the main ways of socialization, especially during night time. Milan is a city that has been increasingly developed its appeal towards tourism and night-life, especially after the 2015 expo.
Besides a strong recovery is expected as soon as the current covid emergency will be solved. In this perspective being able to make an informed choice about the location of a new pub becomes crucial.

Business Problem

The aim of this project is then to guide in the selection of the best locations in Milan to open a new pub.
Using what has been learned during the course this exercise has the ambition to help possible entrepreneurs in the choice.

Target Audience of this project

The main audience would be of course people interested in opening a new pub in the city, but we can aim this project to the entire community and to the administration of the city as a helpful guide on the matter, in the perspective of developing night-life in new neighborhoods.

Data

To solve the problem, we will need the following:
• List of neighbourhoods in Milan
• Latitude and longitude coordinates of those neighbourhoods. Required to get the venue data and plot the resulting data.
• Data related to pubs, in particular their locations.

Sources of data and methods to extract them

This Wikipedia page contains a list of all neighbourhoods in Milan, from Affori to Zona Tortona. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages.

Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods.

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Milano available in the Wikipedia page.

We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data.

We then need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API.

We will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude.

After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package.

This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters.

We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a 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 use in clustering.

Since we are analysing the “Pubs” data, we will filter only the related categories for the neighbourhoods.

Lastly, 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.

We will cluster the neighbourhoods into 4 clusters based on their frequency of occurrences. The results will allow us to identify which neighbourhoods have higher or lowest concentration of pubs.

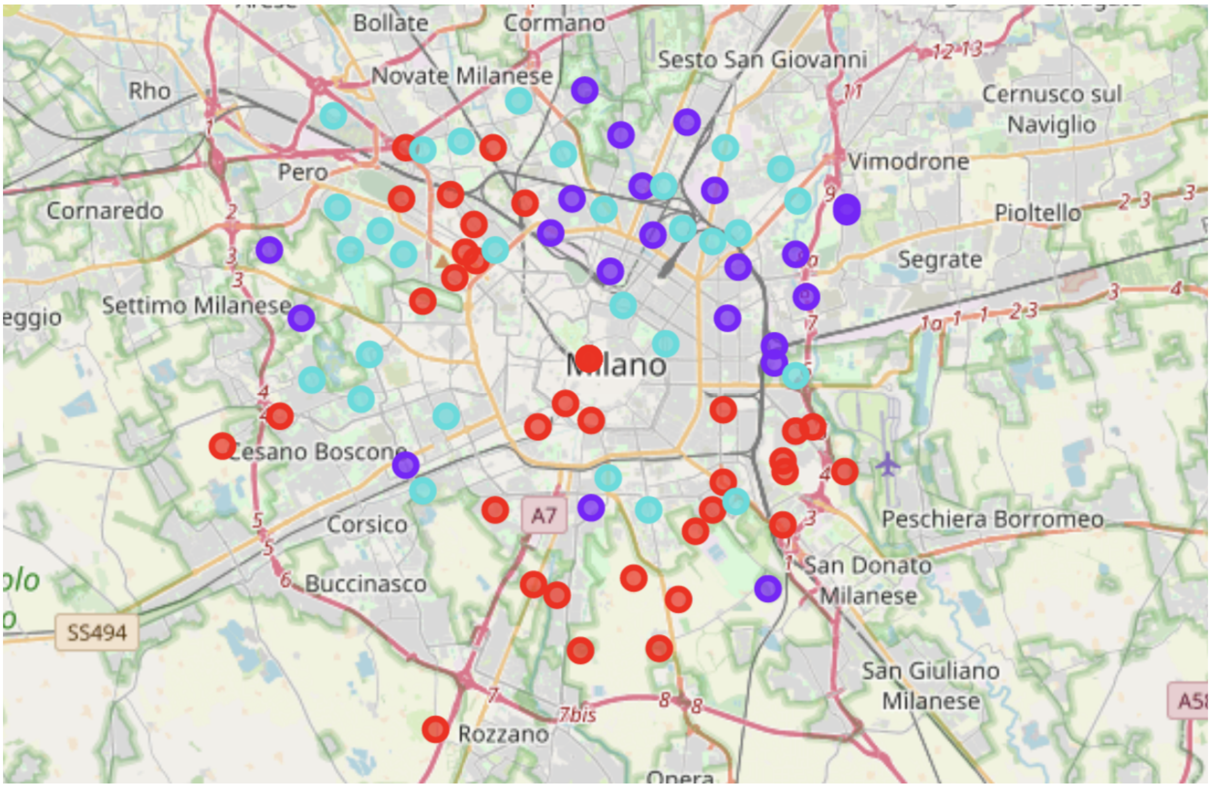
Based on the occurrence of pubs in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable for a new business.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Pubs”:

- Cluster 0: Neighbourhoods with low number to no existence
- Cluster 1: Neighbourhoods with high concentration
- Cluster 2: Neighbourhoods with moderate number

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



Discussion

As observations noted from the map in the Results section, most of the pubs seem to be located in the north-east area of the city, with the highest number in cluster 1 and moderate number in cluster 2.
On the other hand, cluster 0 has very low number to no pubs in the neighbourhoods. This may represent neighborhood with little or zero nightlife.

From this perspective it may be advisable to pick either cluster 1 or 2 as the neighborhood of choice depending of the USP of the pub and joining this results with other data available to the possible entrepreneur, such as rent costs.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence, as already stated there are other factors such as population and rent costs.
However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project.
In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid subscription to bypass these limitations and obtain more results.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new pub. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 1 or 2 are the most preferred locations to open a new pub. 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 to open a new business.

References:

Quartieri di Milano. Wikipedia. Retrieved from
https://it.wikipedia.org/wiki/Categoria:Quartieri_di_Milano

Foursquare Developers Documentation. Foursquare. Retrieved from
https://developer.foursquare.com/docs