Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Shopping Mall in Lille, France

By: Soumia GOURRAM

Introduction

For many shoppers, visiting malls is a great way to relax and have fun during weekends and holidays. Shopping centers are like a single destination for all types of buyers. For the traders the central location and the large crowd in the shopping centers constitute an excellent distribution channel for marketing their products and services. As a result, there are many shopping centers in Lille. Of course, as with any business decision, opening a new shopping center requires serious thought and is much more complicated than it seems. In particular, the location of the mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

Business problem

The objective of this project is to analyze and select the best locations in the city of Lille, in France, to open a new shopping center. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: if a property developer is looking to open a new shopping center, where would you recommend that they open it?
This project is particularly useful for property developers and investors who wish to open or invest in new shopping centers.

Data

To solve the problem, we will need the following data:

- List of districts of Lille.
- Latitude and longitude coordinates of these districts. This is necessary to draw the map and also to obtain the data of the place.
- Site data, in particular data relating to shopping centers. We will use this data to cluster on neighborhoods.

Data sources and methods to extract it :

This web page (http://www.mapcrow.info/Lille-FR-suburbs) contains a list of neighbourhoods in Lille. We will use web scraping techniques to extract data from the web page, using queries Python and beautifulsoup packages. Then we will get the geographic coordinates of the neighborhoods using the Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.
After that, we will use the Foursquare API to get the location data for these neighborhoods. Foursquare has one of the largest databases of more than 105 million locations and is used by more than 125,000 developers.
The Foursquare API will provide many categories of data on the sites, we are particularly interested in the Shopping Mall category to help us solve the proposed business problem. This is a project that will use many skills in data science, from web scraping, to work with API (Foursquare), to data cleaning, to data manipulation, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps followed in this project, the data analysis we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighbourhoods in the city of Lille. Fortunately, the list is available in the Web page (http://www.mapcrow.info/Lille-FR-suburbs).
We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, 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 of Lille.
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 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.

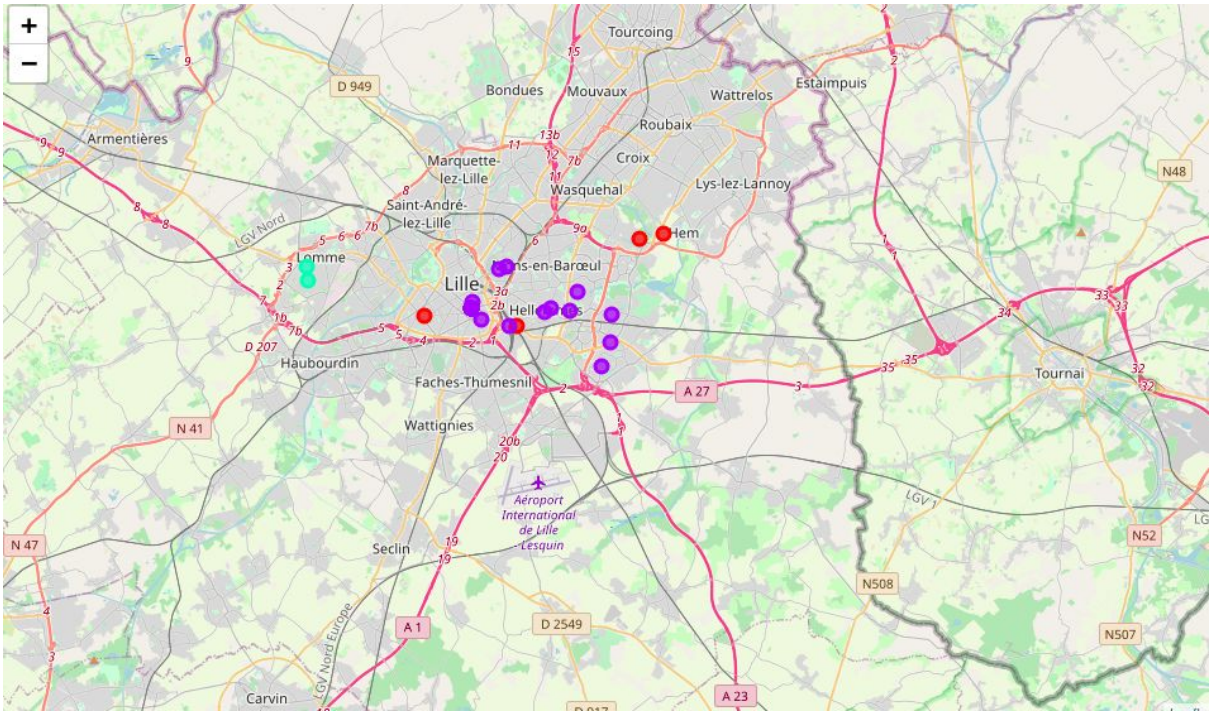
By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighbourhoods.
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies 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 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

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 “Shopping Mall”:

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

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



Discussion

As observations noted from the map in the Results section, most of the shopping malls are concentrated in the central area of Lille city, with the highest number in cluster 1 and moderate number in cluster 0. On the other hand, cluster 2 has very low number to no shopping mall in the neighbourhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls. Meanwhile, shopping malls in cluster 1 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. From another perspective, the results also show that the oversupply of shopping malls mostly happened in the central area of the city, with the suburb area still have very few shopping malls. Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighbourhoods in cluster 2 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in cluster 1 which already have high concentration of shopping malls and suffering from intense competition.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of shopping malls, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall. 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 account 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 shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 2 are the most preferred locations to open a new shopping mall. 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 shopping mall.