Setting up a Café in Ho Chi Minh City

A final report for the course "Applied Data Science Capstone" given by IBM on Coursera

Vijay M

April 11,2020

1   Problem's description

If you have a chance to visit Ho Chi Minh City (HCMC), you will see that it's one of the most active cities in the world. People here love to talk, to chat everywhere. The catering services in this city, therefore, very thriving.

We need to clarify the differences between a *coffee shop* and a *café*. Coffee shop has no similar connotations. From personal experience in the United States, a café serves meals, while a coffee shop usually just sells snacks (muffins, scones, shortbread). This is not strictly the case, and both usually serve coffee. In this project, we suppose to work only on the café.

Although there are already a lot of cafés in HCMC, their density between districts is not uniform. There are some districts containing too many café while there are less in some others. If we have some knowledge about the *population density*, the *housing price* in each district coupling with an overview of *the number of cafe*, we can have a better idea to set up a new business there.

If we think of it by an investor, we expect to choose a place where the population density is high but less competitors. If the housing price in that place is low, it's more attractive to us.

By using Data Science and some geometric factors about the relation between districts in HCMC, we can give good answers of following questions to the investors so that they can have a better vision about not only the café but also about other venues in HCMC.

1. *How many venues in each district? Answering this question gives us a better understand- ing about the dynamic level of a district.*

2. *How many categories in each district? Answering this question helps us know about the diversity in business of a district.*

3. *How many venues in each category? This question shows the magnitude of a category in a district.*

4. *What are the most popular categories in each district? If investors change their mind to focus on other commercial fields instead of opening a café.*

5. *How many clusters we can use to categorize the districts based on the popularity of cafés?*

6. *In which districts, the average housing price is low and the number of cafés is low also?*

7. *Where there are many people but less cafés?*

8. Visualize all information on the map so that we can have a better look on what we want to find the answers!

2   Data presentation

In order to explore the questions, we need to use following data in the research.

1. List of Ho Chi Minh City administrative units from Wikipedia. It gives us a list of all urban districts of HCMC with their area (in Km^2), population (in 2015) and the density of each district (people/Km^2). The list is given in http://bit.ly/3or0yU8.

2. List of the coordinates (latitude, longitude) of all urban districts in HCMC. This list can be generated based on the name of each district and package *geopy.geocoders.Nominatim*.

3. List of average housing prices per m^2 in HCMC. The list is frequently updated in https: //mogi.vn/gia-nha-dat.

4. A *.json* file contains all coordinates where we use it to create a choropleth map of Hous- ing Sales Price Index of HCMC. I create this file by myself using https://nominatim. openstreetmap.org.