Final data Science Capstone Project Problem Definition and Data

Background

In this capstone project I will use the knowledge learned in the previous chapters to solve a practical problem and demonstrating the creation of value by applying the learned skills

Problem Definition

For this project, I chose a theoretical business problem. The question that we are trying to answer is the following.

My friend is studying in Shanghai University, he plans to start a business after graduation, therefor he decided to open a Tibetan restaurant in Shanghai,china.

Taking into account the price level at which the restaurant will operate, the intent is to find an optimal location in an area, where gastronomy is booming and which is easily accessible for tourists and for wealthier local citizens as well.

business logic

we can use unsupervised machine learning to create clusters of districts that will provide us with a list of areas for consideration for the restaurant, The intent is that the restaurant to be situated close to one of the gastronomical centers and high income area

Data collection

To perform this analysis, we will need the following data:

1. List of the districts of shanghai

   List of districts will be obtained from wikipedia (https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai)

2. Geo-coordinates of the districts in shanghai

   Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.

3. Top venues of districts

   Top venues data will be obtained from Foursquare through an API.

4.the income data (GDP per capita in 2019) for districts of shanghai

the income data for districts of shanghai will be obtained from Shanghai Bureau of Statistics(http://tjj.sh.gov.cn)

Use of Data and Methodology

After tidying up and exploring the data, we will apply the Unsupervised machine learning technique for creating clusters of districts. We will use the silhouette score for choosing the optimal number of clusters.