# **Applied Data Science Capstone**

## **Introduction**

The Socialist Republic of Vietnam is an S-shaped strip of land, located in the center of Southeast Asia, in the east of the Indochinese peninsula, on the north by China, on the west by Laos, Cambodia, and in the east. South overlooks the East Sea and Pacific Ocean. Vietnam coastline is 3 260 km long, land border is 4 510 km long. On the mainland, from the northernmost point to the southernmost point (according to the flight path), the length is 650km, from the easternmost point to the westernmost point where the widest place is 600km (North), 400 km (South), the narrowest place is 50km. (Quang Binh).

It is a country rich in cultural and economic traditions that are on the rise in recent years. Vietnam promises to be one of the tourism, shopping and entertainment centers of the region. With many landscapes and historical sites, each year tens of millions of people come to relax and visit here.

Realizing that when tourists come to Vietnam and especially Hanoi capital, there are still many difficulties in identifying and searching for places, entertainment services, eating and visiting here.

Just as it is difficult for investors and service developers to have an overview of the current status of services in each area in Hanoi city.

After completing the IBM Data Science course, I came up with an idea to build an application that provides tourists as well as investors with an overview, the status of services in the area to confirm. to decide where to visit, eat, and develop services here.

## **Data description**

- Data on administrative geography of the city of Hanoi

<https://diaocthinhvuong.vn/danh-sach-don-vi-hanh-chinh-thanh-pho-ha-noi/>

( I use BeatifulSoup to extract data from the website)

- Once I have the administrative geographic data I use the Foursquare API to discover the types of services that are active in the area.

- Foursquare account info to make a request

<https://developer.foursquare.com/>

## **Methodology**

In this section, I will guide how to get source data and perform data processing and analysis:

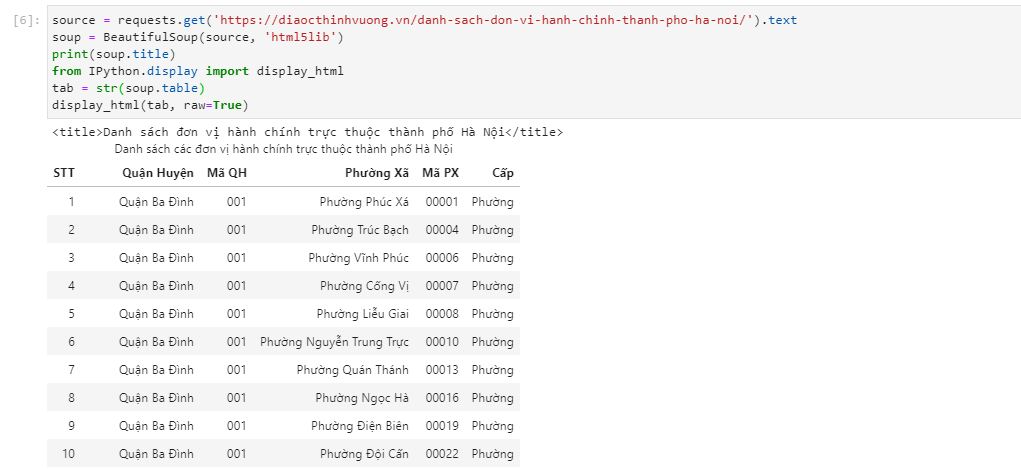
* Data retrieval, exploration and wrangling
* Performing K-means clustering algorithm to segment neighborhoods
* Visualizing population projections and neighborhood segments

**Data retrieval, exploration and wrangling**

- First use Beautifulsoup to get data from the website.

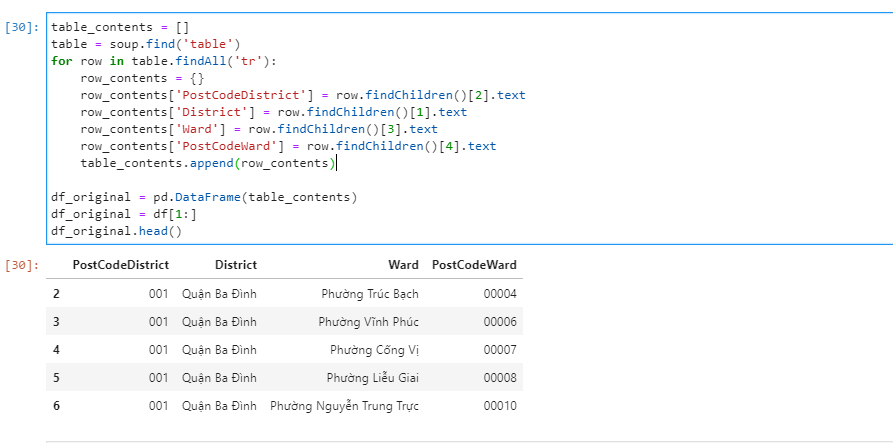
<https://pypi.org/project/beautifulsoup4/>

* The following results:



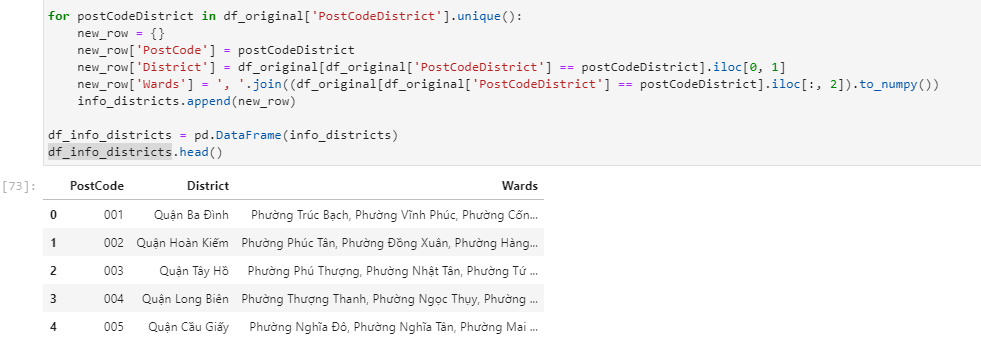
- After that, I select some necessary information for analysis such as postcode, name of district, commune and ward:

* The following results:



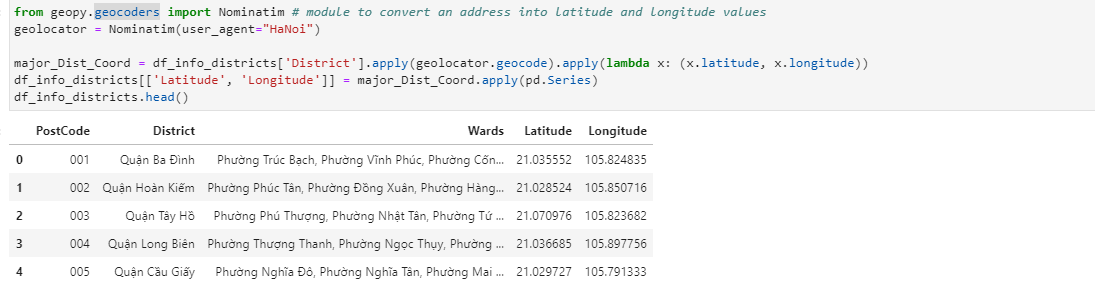
- Continue to filter data by district:

* The following results



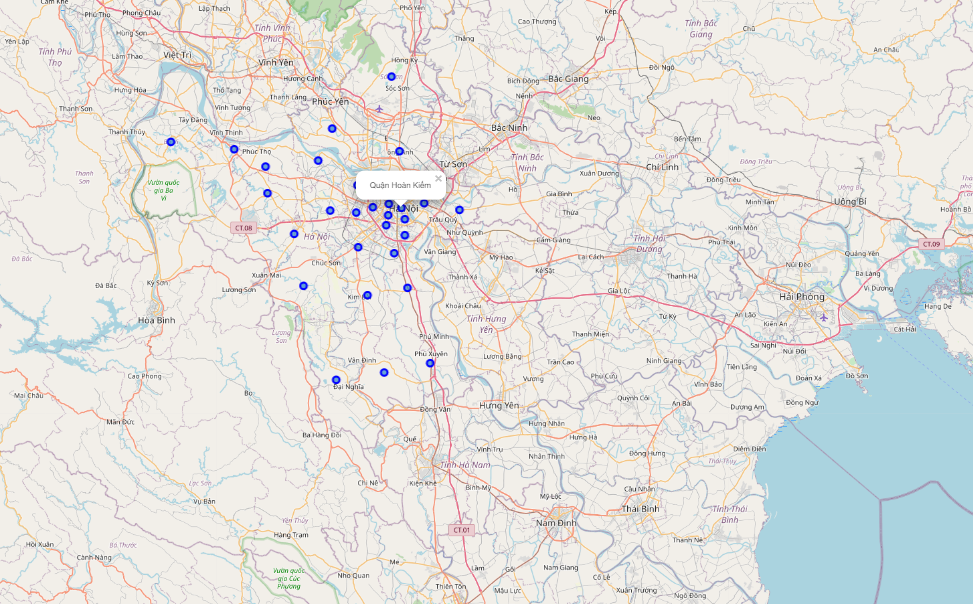
- After I have administrative geographic information, I use geocoders to get information about the geographical coordinates of each area.

* The following results:



- Perform data visualization to know the geographical distribution of each area:

* The following results:



- Use the Foursquare API to do a search for all types of related services in each area

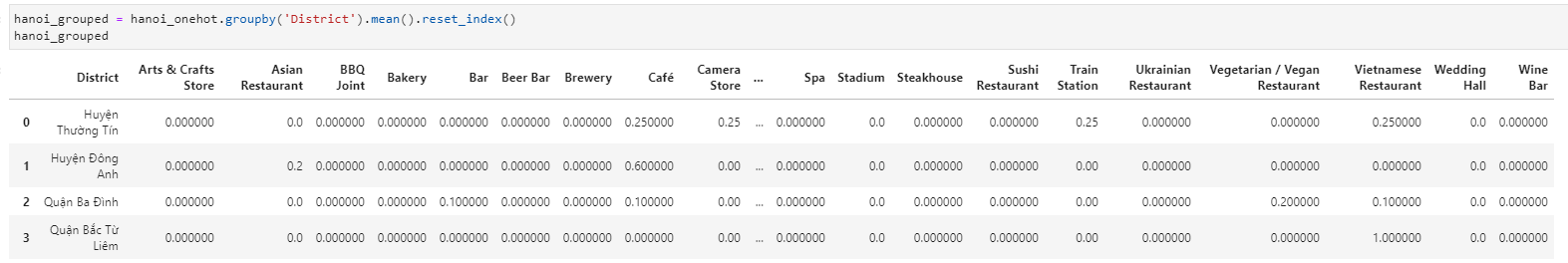
Foursquare API is used to explore types of venues in each area. Foursquare identifies 10 top level categories. There are multiple sub categories which will not be used it for the time

* Arts & Entertainment
* College & University
* Event
* Food
* Nightlife Spot
* Outdoors & Recreation
* Professional & Other Places
* Residence
* Shop & Service
* Travel & Transport
* The following results



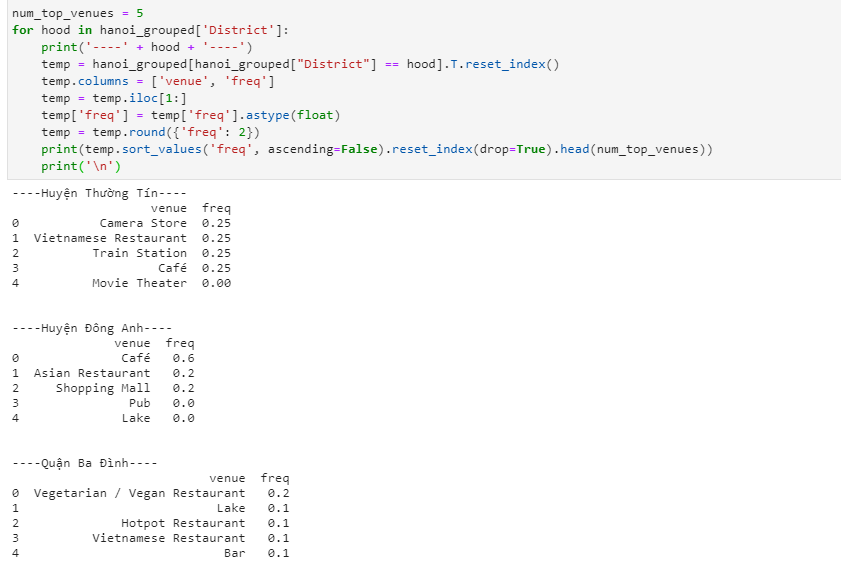
- Standardize data on each type of service to prepare for the training process

* The following results:



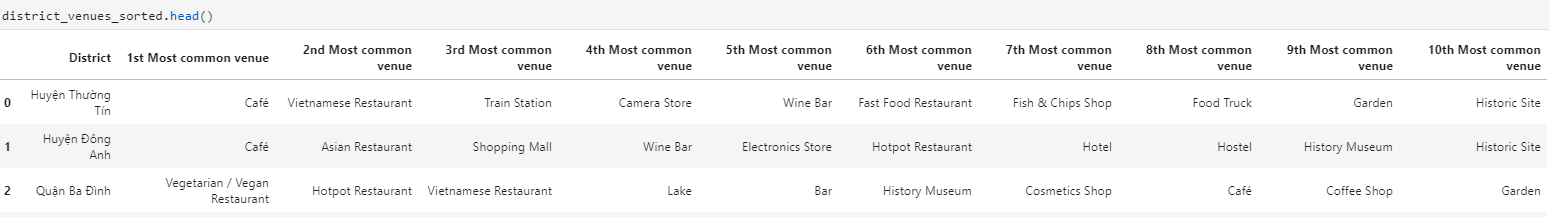
- Display the top services in each area for an overview:

* The following results:



- Get 10 services with high frequency in each area for training

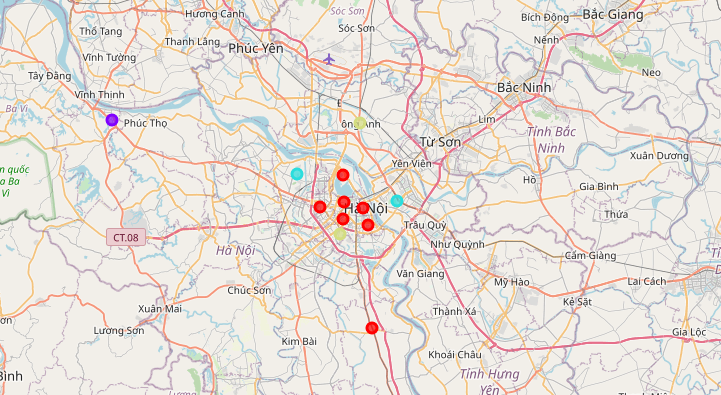
* The following results:



**Performing K-means clustering algorithm to segment neighborhoods**

****

- Perform model training and display results

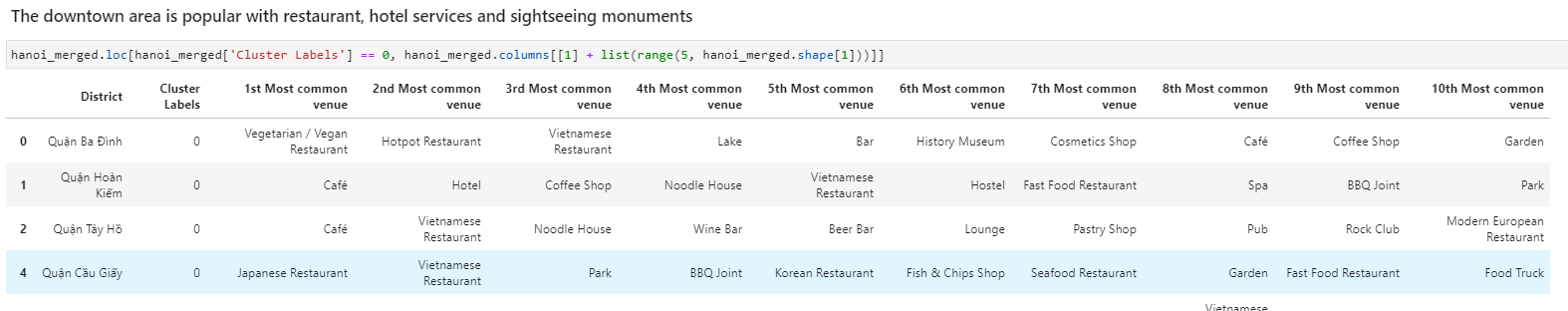


## Result

Categorization of development areas with typical services attached

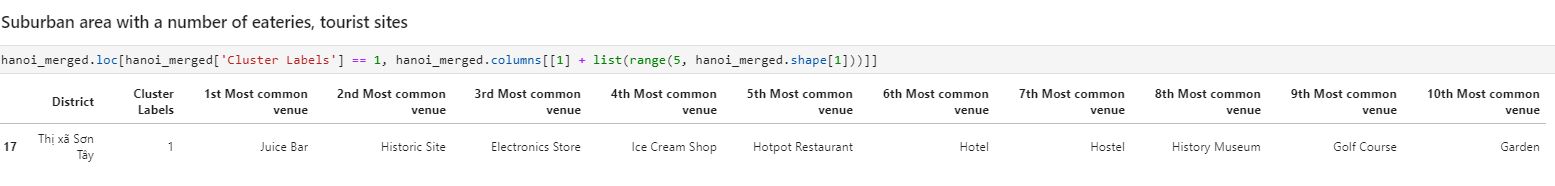
The downtown area is popular with restaurant, hotel services and sightseeing monuments

* The following result:



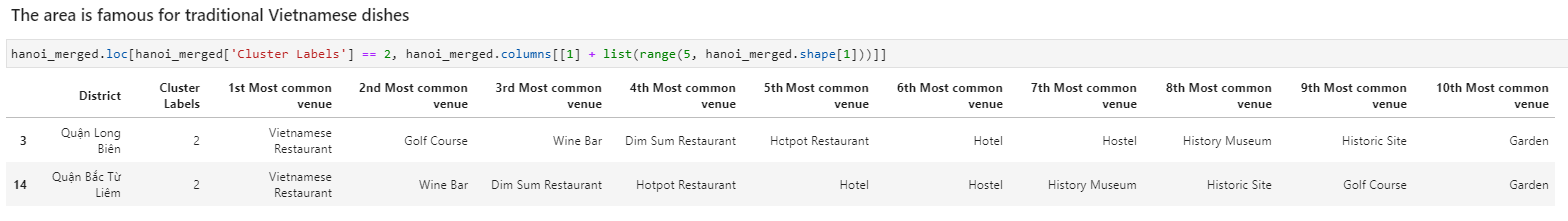
Suburban area with a number of eateries, tourist sites

* The following result:



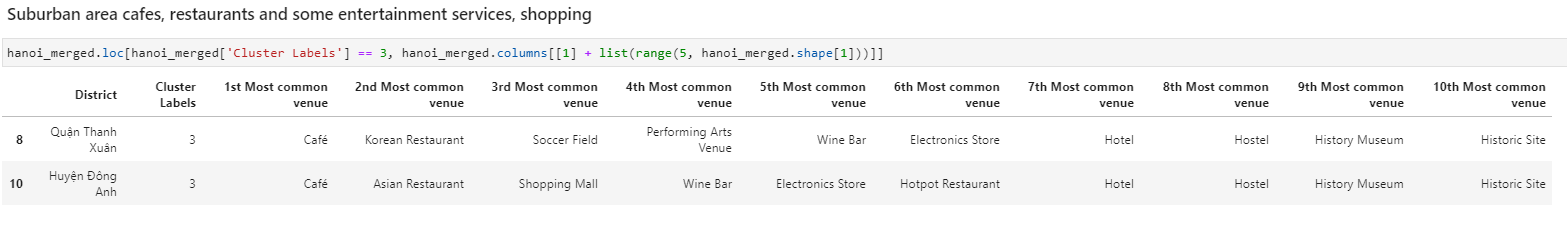
The area is famous for traditional Vietnamese dishes

* The following result:



Suburban area cafes, restaurants and some entertainment services, shopping

* The following result:



## **Discuss**

The application is for academic purposes only to support all knowledge gained from the course.

The results of the application are not really impressive because the available data is still very limited. The problem applied to Hanoi city is still very specific because most services are concentrated in the city center

## **Conclude**

I have accomplished the project's goal of giving visitors and developers an overview of the current state of service development in different regions.

## **Acknowledgments and sources**

I sincerely thank the IBM Data Science Professional Certificate course for having a scientific study path, providing me with a lot of knowledge about Data Science.

The course helps me to have an overview and detail of a Data Science project in practice with many specific and vivid lessons.

<https://www.coursera.org/professional-certificates/ibm-machine-learning>

Thank you for review my final project !