INVESTING A REAL ESTATE FOR THE STARTUP "BLACK GARLIC CAFE CHAIN" IN HUE CITY, VIETNAM

1. Introduction

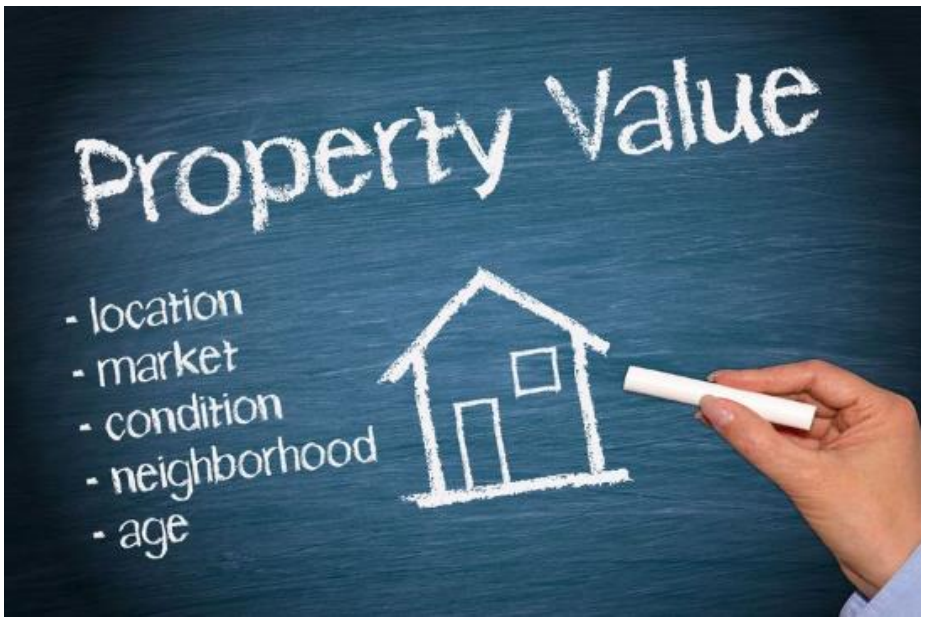
Thừa Thiên-Hue is a province in the North Central Coast region of Vietnam, approximately in the centre of the country. It borders Quảng Trị Province to the north and Đà Nẵng to the south, Laos to the west and the East Sea to the east. The province has 128 km of coastline, 22,000 ha of lagoons and over 200,000 ha of forest. There is an extensive complex of imperial tombs and temples in Hue.

Hue is intimately connected to the imperial Nguyễn Dynasty, based in Hue, who ruled from 1802 to 1945, when the Emperor Bao Dai abdicated in favor of Ho Chi Minh's revolutionary government.

Hue is also famous for it delicious sweet desserts such as Lotus seeds sweet soups, Lotus seed wrapped in logan sweet soup, Areca flower sweet soup, Grilled pork wrapped in cassava flour sweet soup, and Green sticky rice sweet soup.

Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Hue, Vietnam to open a new Cafe chain. However, the choices are severely affected by the restrictions placed by misunderstanding and speculative bubbles.



Source: Internet

Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Hue, Vietnam, if a property developer is looking to open a new coffee chain, where would you recommend that they open it?

Target Audience

This project is particularly useful to property developers and investors looking to open or invest in new coffee chain in the capital city of Hue. This project is timely as the city is currently suffering from oversupply of coffee shop.

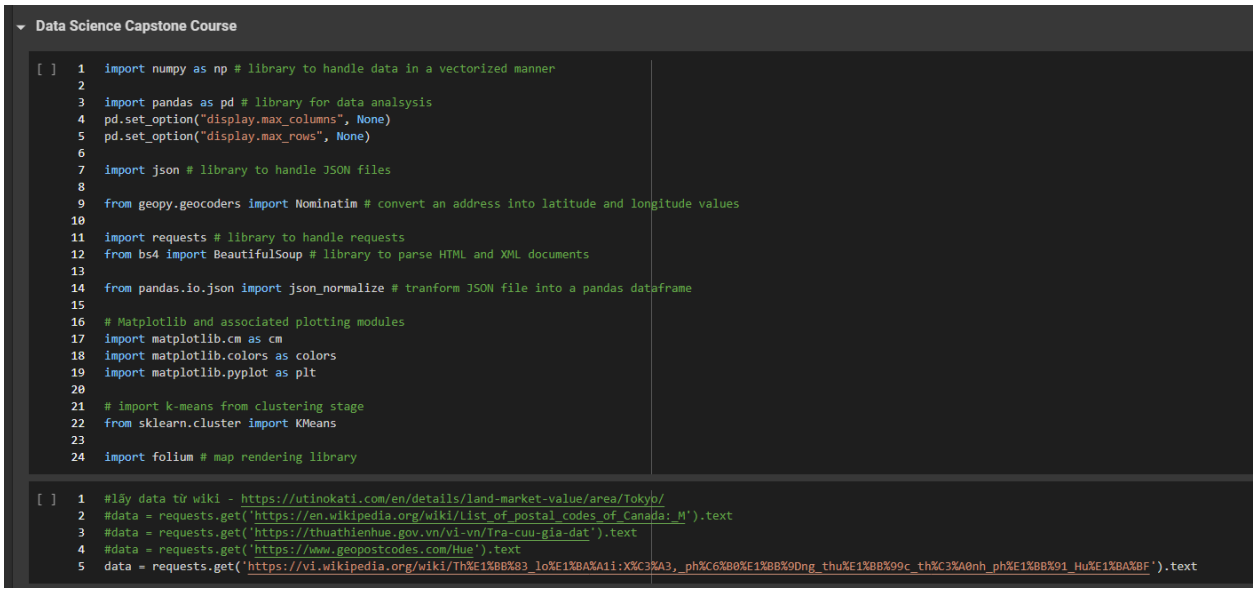
- Business personnel who wants to invest or open a cafe chain. This analysis will be a comprehensive guide to start or expand cafe business targeting the large pool of office workers in Hue during working hours.

- Freelancer who loves to have their own cafe shop as a side business. This analysis will give an idea, how beneficial it is to open a cafe store and what are the pros and cons of this business.

- Budding Data Scientists, who wants to implement some of the most used Exploratory Data Analysis techniques to obtain necessary data, analyze it and, finally be able to tell a story out of it.

Conclusion 1st Week:

Description of Problem and Data Preparation



Data Science Capstone Course

```
[ ]
import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows", None)

import json # library to handle JSON files

from geopy.geocoders import Nominatim # convert an address into latitude and longitude values

import requests # library to handle requests
from bs4 import BeautifulSoup # library to parse HTML and XML documents

from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors
import matplotlib.pyplot as plt

# import k-means from clustering stage
from sklearn.cluster import KMeans

import folium # map rendering library
```

```
[ ]
#lấy data từ wiki - https://utinokati.com/en/details/land-market-value/area/Tokyo/
#data = requests.get('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M').text
#data = requests.get('https://thuathienhue.gov.vn/vi-vn/Tra-cuu-gia-dat').text
#data = requests.get('https://www.geopostcodes.com/Hue').text
data = requests.get('https://vi.wikipedia.org/wiki/Th%E1%BB%83_lo%E1%BA%A1i:X%C3%A3,_ph%C6%B0%E1%BB%9Dng_thu%E1%BB%99c_th%C3%A0nh_ph%E1%BB%91_Hu%E1%BA%BF').text
```

We get the Initial Data-Frame with Names of Major Wards, and corresponding wards in those Major Wards

```
[ ]
#chuyển data thành đối tượng object của gói beautifulsoup
soup = BeautifulSoup(data,'html.parser')
```

```
[ ]
#Tìm tất cả các [hàng (dòng) thẻ 'tr':table row] - thẻ 'td':table data: là các cell
#table = soup.find('tbody')
#table
#new_feed = soup.find('table', class_='list2')
#new_feed.find('thead').find_all('tr')[0].find_all('th')[0].text.rstrip('\r\n')
new_feed = soup.find_all('li')
```

```
[ ]
phuongs = []
for p in range(0,28):
  text=new_feed[p].text.split()
  if(len(text)>1):
    text = text[0]+" "+text[1]
    phuongs.append(text.rstrip(","))
  #text1 = text.rstrip("(phường)")
  #text2 = text.rstrip("Huế")
#phuongs.append('Thành Phố Huế')
phuongs[24]='Vỹ Dạ'
```

```
[ ]
df=pd.DataFrame(phuongs,columns=['Ward'])
df.index = np.arange(1, len(df) + 1) # reset the index so that it starts from 1.
df.head()
```

	Ward
1	An Cựu
2	An Đông
3	An Hòa
4	An Tây
5	Hương Long

+ Code + Text

```
[ ]
wards = []
latitudes = []
longitudes = []
geolocator = Nominatim(user_agent="Thua Thien Hue, Vietnam")
for p in phuongs:
  if(p not in ('An Tây','Phú Cát','Phú Hậu','Phú Hiệp','Phú Hội','Phú Nhuận','Phước Vĩnh','Phường Đúc','Trường An','Xuân Phú')):
    address = p+',Thừa Thiên Huế, Việt Nam'#'thanh pho Hue, Thua Thien Hue, Vietnam' - 'Phường Kim Long, Thành phố Huế, Thừa Thiên Huế, Việt Nam'
    #print(address)
    location = geolocator.geocode(address)
    latitude = str(location.latitude)
    longitude = str(location.longitude)
    wards.append(p)
    latitudes.append(latitude)
    longitudes.append(longitude)
    #print('The geograpical coordinates of {} are {}, {}.'.format(address,latitude, longitude))

df=pd.DataFrame(wards,columns=['Ward'])
df['Latitude']=latitudes#str(latitudes)#map(str,latitudes)
df['Longitude']=longitudes#str(longitudes)#map(str,longitudes)
#df.index = np.arange(1, len(df) + 1) # reset the index so that it starts from 1.
#df.dtypes
```

```
[ ]
df
```

	Ward	Latitude	Longitude
0	An Cựu	16.4574892	107.6002023
1	An Đông	16.4248801	107.6469586
2	An Hòa	16.4248801	107.6469586
3	Hương Long	16.4248801	107.6469586
4	Hương Sơ	16.4924722	107.5119899
5	Kim Long	16.4642271	107.559751
6	Phú Bình	16.4885976	107.5762144
7	Phú Hòa	16.471216	107.5868958
8	Phú Thuận	16.4248801	107.6469586
9	Tây Lộc	16.4772416	107.5637146
10	Thuận Hòa	16.4674072	107.5710217
11	Thuận Lộc	16.4831435	107.572616

So as the next step we will use Foursquare data and obtain information on coffee chains. With these, we can start with our battle of neighborhoods for opening a coffee shop in Hue city.

2. Initial Data Preparation

2.1. Web-Scrapping and Cleaning

2.1.1. Get The Names of Wards from Wikipedia

The Wikipedia page of Hue Wards contains the table of 27 wards of Hue. I have used Beautifulsoup4 and pandas library to create the initial data-frame. For a clean and understandable data-frame some of the wards are renamed for example ‘Vĩ Dạ’ to Vỹ Dạ. Here I have taken the first entry of the major ward column sin the Wikipedia table to concentrate on. Even though not complete but it gives us quite a detailed picture of the corresponding ward, as later on I have considered top most venues within 500 meter radius of the major ward. After this inital preparation, I moved on to the next step to obtain coordinates using Geopy library.

2.1.2. Get the Coordinates of the Major Wards

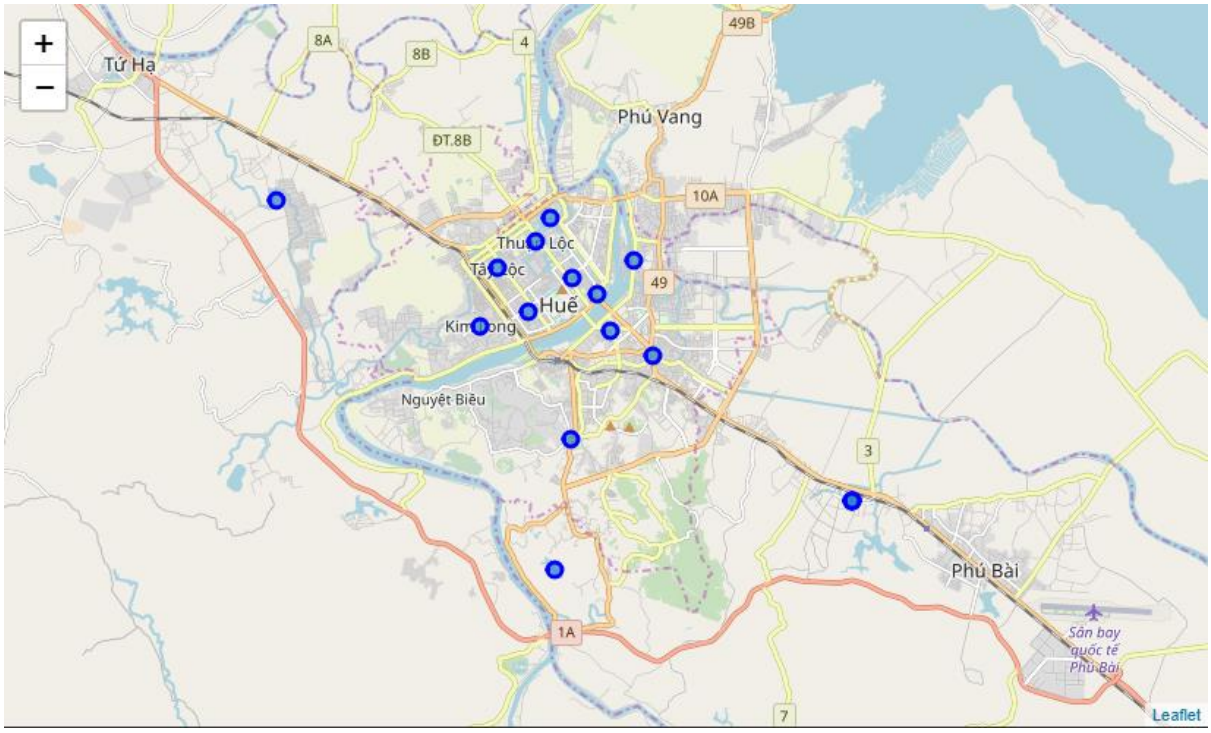
Some of the coo-ordinates of the major wards returned by Geopy are wrong and I have figured out the reason for this is the name of the major wards in the data-frame are different from their actual names. In these cases (4 of them), I had to google search and replace using pandas library. After little manipulation the obtained data-frame looks as below after remove several wards which could not receive locations from code “geolocator.geocode(address)”, I guess that geocoders package did update it’s data.

```
[ ]  1  df
```

	Ward	Latitude	Longitude
0	An Cựu	16.4574892	107.6002023
1	An Đông	16.4248801	107.6469586
2	An Hòa	16.4248801	107.6469586
3	Hương Long	16.4248801	107.6469586
4	Hương Sơ	16.4924722	107.5119899
5	Kim Long	16.4642271	107.559751
6	Phú Bình	16.4885976	107.5762144
7	Phú Hòa	16.471216	107.5868958
8	Phú Thuận	16.4248801	107.6469586
9	Tây Lộc	16.4772416	107.5637146
10	Thuận Hòa	16.4674072	107.5710217
11	Thuận Lộc	16.4831435	107.572616
12	Thuận Thành	16.4748538	107.5812484
13	Thủy Biều	16.4094987	107.5770314
14	Thủy Xuân	16.4388	107.5809431
15	Vỹ Dạ	16.4790295	107.5958008
16	Vĩnh Ninh	16.4631912	107.5901125

2.2. Foursquare Data

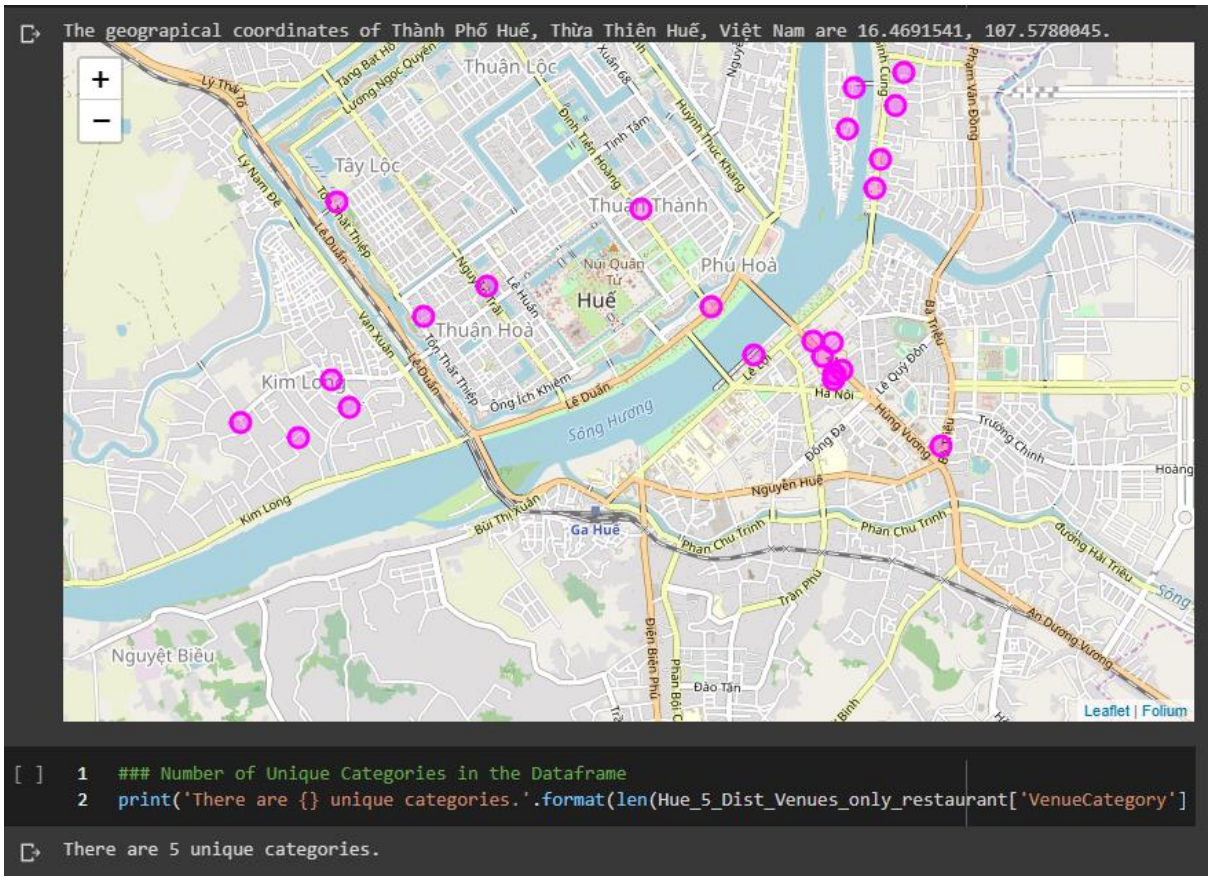
Finally, I make use of Foursquare API to obtain the 100 most common venues within 1000 meter of each wards.



3. Exploring the Data

3.1. Exploring the Wards of Hue

From the Foursquare data, we could see that there are 5 unique categories of Restaurant and 2 unique catagories of coffee



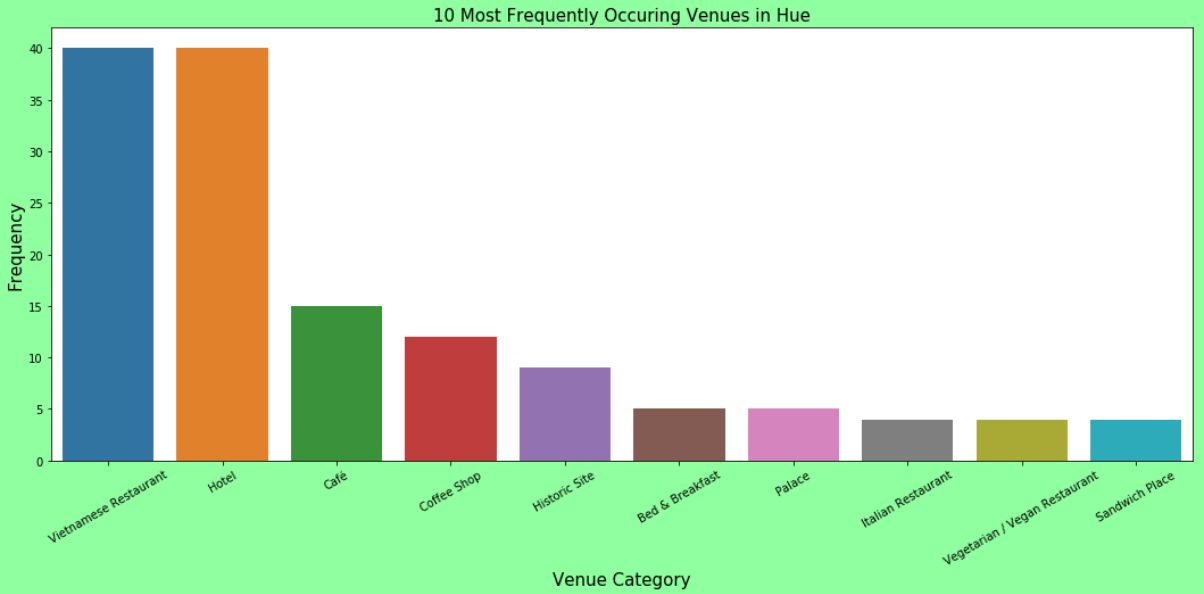
```
[40]  1  print (Hue_Venues_Coffee_Cafe ['VenueCategory'].value_counts())

Café           15
Coffee Shop    12
```

Here we have found out that

• Vietnamese restaurants are top the charts of most common venues in top 10 categories, followed by hotels, café and Coffee Shop

• A plot of the ten most frequent venues in Hue are as below



Since this reminded of Hue restaurant and coffee so you need to take a break


CỬA
HÀNG
CAFE

Next step was to obtain information about the top 5 venues of each district. And to do that, I proceed as follows

• Create a data-frame with pandas one hot encoding for the venue categories.

• Use pandas groupby on District column and obtain the mean.

• Transpose the data-frame at step 2 and arrange in descending order.

Implementing them in Pandas outputs the following:

---------An Cựu---------

	Venue	Freq
0	Café	0.25
1	Movie Theater	0.25
2	Hotel	0.12
3	Vegetarian / Vegan Restaurant	0.12
4	Bed & Breakfast	0.12

---------Hương Sơ---------

	Venue	Freq
0	Coffee Shop	1.0
1	Art Gallery	0.0
2	Hotel	0.0
3	Indian Restaurant	0.0
4	Italian Restaurant	0.0

---------Thuận Hòa---------

	Venue	Freq
0	Café	0.20
1	Vietnamese Restaurant	0.13
2	Palace	0.13
3	Juice Bar	0.07
4	Historic Site	0.07

---------Thuận Lộc---------

	Venue	Freq
0	Café	0.33
1	Vietnamese Restaurant	0.11
2	Vegetarian / Vegan Restaurant	0.11
3	Souvenir Shop	0.11
4	Historic Site	0.11

---------Thuận Thành---------

	Venue	Freq
0	Historic Site	0.17
1	Café	0.12
2	Coffee Shop	0.12
3	Vietnamese Restaurant	0.08
4	Palace	0.08

---------Vỹ Dạ---------

	Venue	Freq
0	Hotel	0.37
1	Vietnamese Restaurant	0.21
2	Coffee Shop	0.11
3	Resort	0.05
4	BBQ Joint	0.05

---------Vĩnh Ninh---------

	Venue	Freq
0	Hotel	0.27
1	Vietnamese Restaurant	0.23
2	Coffee Shop	0.07
3	Café	0.04
4	Noodle House	0.03

---------Kim Long---------

	Venue	Freq
0	Vietnamese Restaurant	0.75
1	Noodle House	0.25
2	Vegetarian / Vegan Restaurant	0.00
3	Movie Theater	0.00
4	Indian Restaurant	0.00

---------Phú Bình---------

	Venue	Freq
0	Vietnamese Restaurant	0.25
1	Historic Site	0.12
2	Lottery Retailer	0.12
3	Boat or Ferry	0.12
4	Sandwich Place	0.12

---------Phú Hòa---------

	Venue	Freq
0	Hotel	0.34
1	Vietnamese Restaurant	0.23
2	Italian Restaurant	0.06
3	Historic Site	0.03
4	Sports Bar	0.03

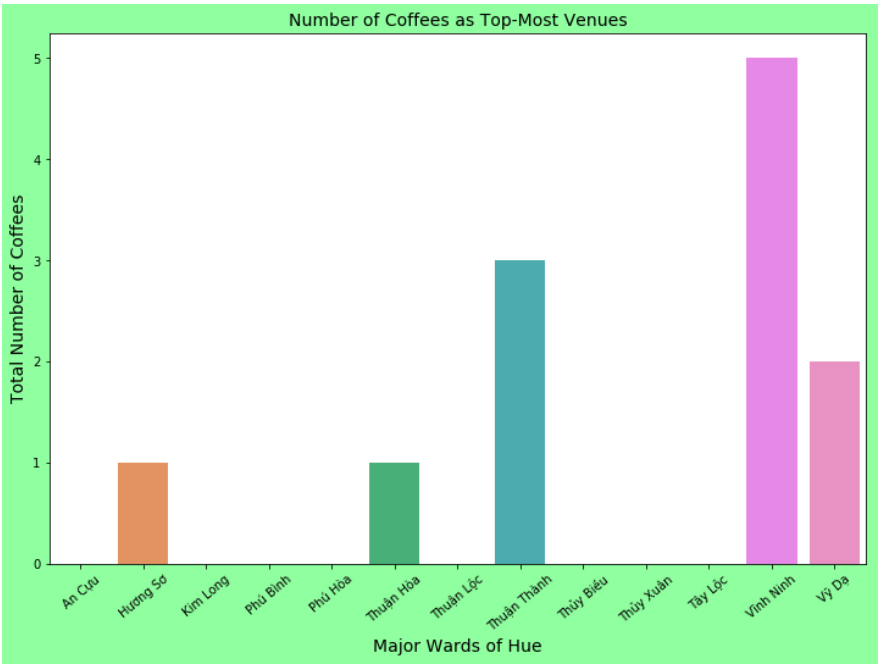
---------Thủy Biều---------

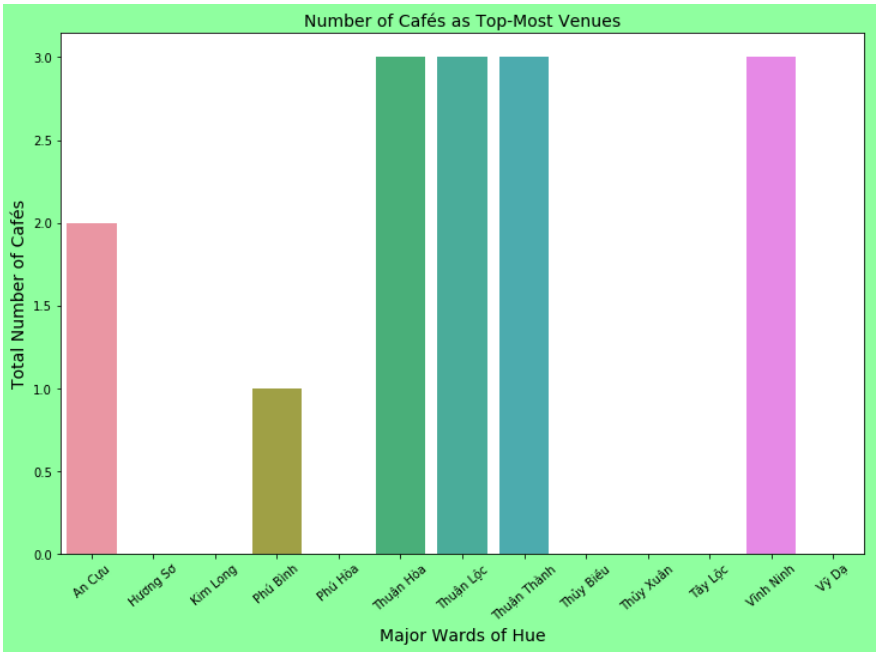
	Venue	Freq
0	Art Gallery	0.33
1	Lake	0.33
2	Campground	0.33
3	Night Market	0.00
4	Indian Restaurant	0.00

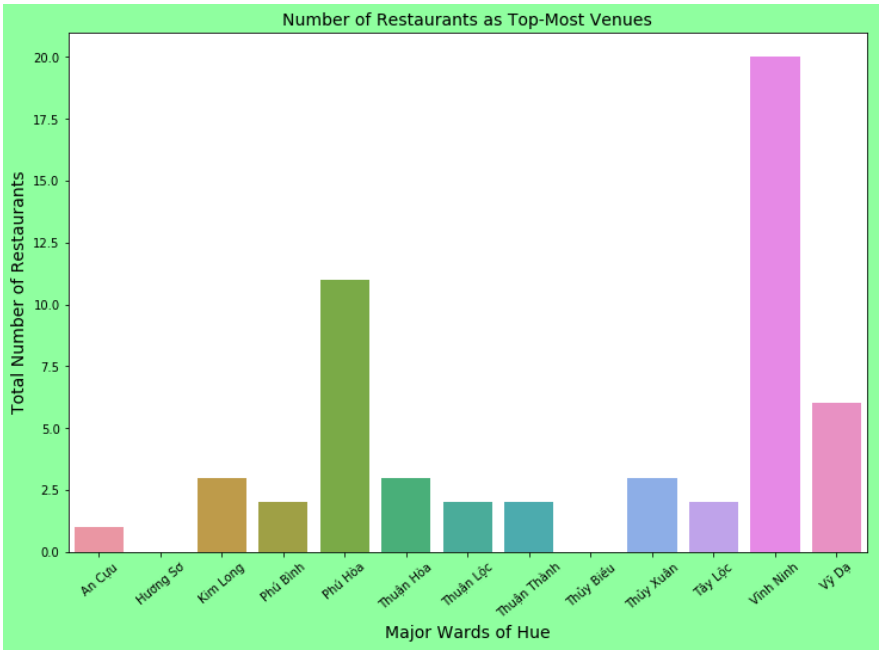
---------Thủy Xuân---------

	Venue	Freq
0	Vietnamese Restaurant	0.50
1	Historic Site	0.25
2	Seafood Restaurant	0.25
3	Movie Theater	0.00
4	Indian Restaurant	0.00

To explore which wards has the highest number of Cafés, Coffee shops and restaurants as the most common venue and the plot below is the answer



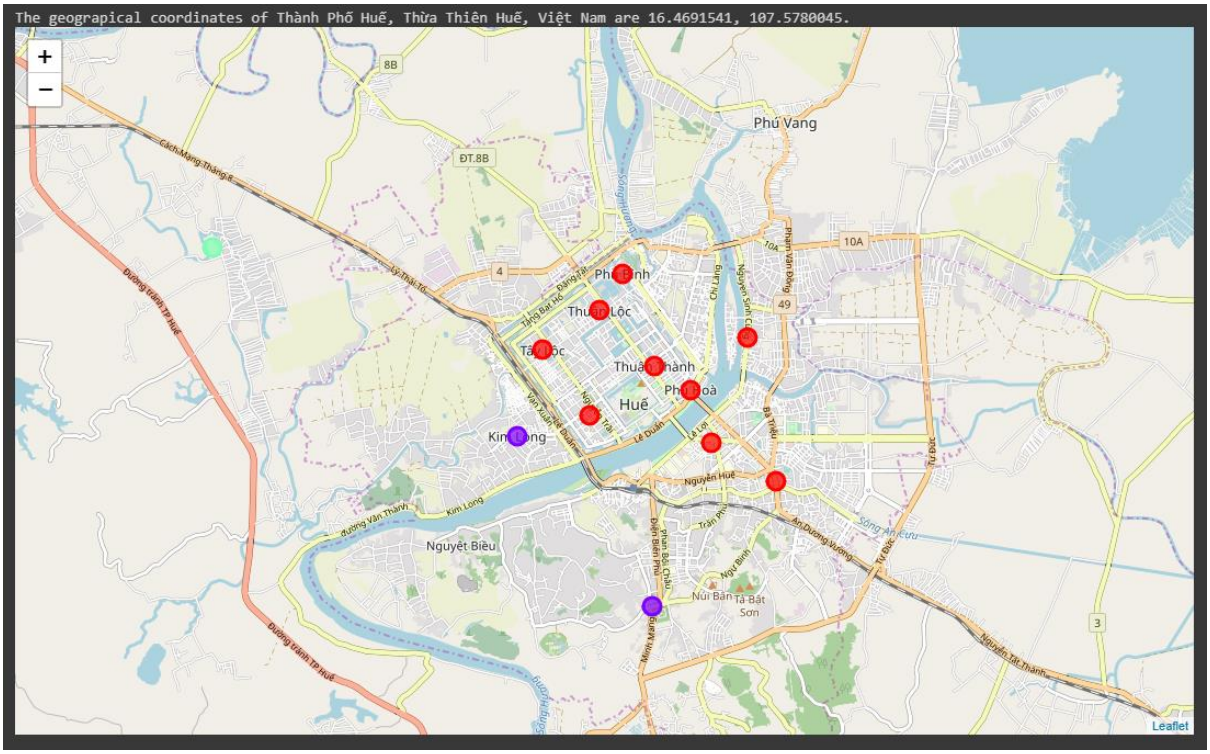




3.2. Clustering the Major Wards of Hue

Finally, we try to cluster these wards based on the frequency of venue categories and, use K-Means clustering. So our expectation would be based on the similarities of venue categories, these wards will be clustered. Using K-Means algorithm from Scikit-learn library we obtain 3 clusters as shown below.

The geograpical coordinates of Thành Phố Huế, Thừa Thiên Huế, Việt Nam are 16.4691541, 107.5780045.



Cluster 1

```
[ ]  1  Hue_merged_not_NAN.loc[Hue_merged_not_NAN['Cluster Label'] == 0, Hue_merged_not_NAN.columns[[0]+list(range(4, Hue_merged_not_NAN.shape[1]))]]
```

	Ward	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	An Cựu	Movie Theater	Café	Vegetarian / Vegan Restaurant	Hotel	Bed & Breakfast	Bus Station	Castle	Food Truck	Food Service	Food Court
6	Phú Bình	Vietnamese Restaurant	Sandwich Place	Food Service	Lottery Retailer	Boat or Ferry	Café	Historic Site	Bakery	French Restaurant	Food Truck
7	Phú Hòa	Hotel	Vietnamese Restaurant	Italian Restaurant	BBQ Joint	Bakery	Bar	Bed & Breakfast	Beer Garden	Breakfast Spot	Historic Site
9	Tây Lộc	Food Truck	Shopping Mall	Hotpot Restaurant	Food Court	Fast Food Restaurant	Noodle House	Breakfast Spot	Vietnamese Restaurant	Campground	Food Service
10	Thuận Hòa	Café	Vietnamese Restaurant	Palace	Historic Site	Juice Bar	Coffee Shop	Castle	French Restaurant	Monument / Landmark	Shopping Mall
11	Thuận Lộc	Café	Vietnamese Restaurant	Sandwich Place	Vegetarian / Vegan Restaurant	Food Service	Historic Site	Souvenir Shop	Bar	Coffee Shop	BBQ Joint
12	Thuận Thành	Historic Site	Café	Coffee Shop	Vietnamese Restaurant	Palace	Monument / Landmark	Bookstore	Castle	Creperie	Dessert Shop
13	Thủy Biều	Art Gallery	Lake	Campground	Bakery	Coffee Shop	French Restaurant	Food Truck	Food Service	Food Court	Fast Food Restaurant
15	Vỹ Dạ	Hotel	Vietnamese Restaurant	Coffee Shop	Lounge	BBQ Joint	Vegetarian / Vegan Restaurant	Bed & Breakfast	Resort	Italian Restaurant	Food Service
16	Vĩnh Ninh	Hotel	Vietnamese Restaurant	Coffee Shop	Café	Bakery	Bed & Breakfast	Noodle House	Breakfast Spot	Dive Bar	Vegetarian / Vegan Restaurant

Cluster 2

```
[ ]  1  Hue_merged_not_NAN.loc[Hue_merged_not_NAN['Cluster Label'] == 1, Hue_merged_not_NAN.columns[[0]+list(range(4, Hue_merged_not_NAN.shape[1]))]]
```

	Ward	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Kim Long	Vietnamese Restaurant	Noodle House	Campground	French Restaurant	Food Truck	Food Service	Food Court	Fast Food Restaurant	Dive Bar	Dessert Shop
14	Thủy Xuân	Vietnamese Restaurant	Seafood Restaurant	Historic Site	Campground	French Restaurant	Food Truck	Food Service	Food Court	Fast Food Restaurant	Dive Bar

Cluster 3

```
[ ]  1  Hue_merged_not_NAN.loc[Hue_merged_not_NAN['Cluster Label'] == 2, Hue_merged_not_NAN.columns[[0]+list(range(4, Hue_merged_not_NAN.shape[1]))]]
```

	Ward	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	Hương Sơ	Coffee Shop	Vietnamese Restaurant	Campground	French Restaurant	Food Truck	Food Service	Food Court	Fast Food Restaurant	Dive Bar	Dessert Shop

4. Results

The results of the exploratory data analysis and clustering are summarized below:

- Most of the neighborhoods fall into Cluster 2 which are mostly business areas with Vietnamese Restaurant. Cluster 1 is only 2 1st Most Common Venue with Cafe. In cluster 3, Coffee Shop is the number one of most common Venue
- Restaurants top the charts of most common venues in the wards.
- Phú Binh, Kim Long and Thủy Xuân wards are dominated by Vietnamese restaurants as the the most common venues.
- Vỹ Dạ, Vĩnh Ninh and Phú Hòa areas are dominated by hotel as most common venues.

5. Discussion

According to this analysis, Vinh Ninh area will provide most competition for an upcoming café chain as hotel is the most common venue in this area and the frequency of café as common venue are very low compared to the remaining wards.

Some drawbacks of this analysis are the clustering is completely based on the most common venues obtained from Foursquare data. Since land price, distance of the venues from closest stations, number of potential customers, benefits and drawbacks of Shinagawa being a port region, could all play a major role and thus, this analysis is definitely far from being conclusory. However, it definitely gives us some very important preliminary information on possibilities of opening restaurants around the major wards of Hue.

Also, another pitfall of this analysis could be consideration of only one major district of each ward of Hue, taking into account of all the areas, wards would give us an even more realistic picture.

Furthermore, this results also could potentially vary if we use some other clustering techniques like DBSCAN.

6. Conclusion

Finally to conclude this project, we have got a small glimpse of how real life data-science projects look like. I have made use of some frequently used python

libraries to scrap web-data, use Foursquare API to explore the major wards of Hue and saw the results of segmentation of wards using Folium leaflet map. Potential for this kind of analysis in a real life business problem is discussed in great detail. Also, some of the drawbacks and chance for improvements to represent even more realistic pictures are mentioned. Finally, since my analysis were mostly concentrated on the possibilities of opening a café chain targeting the huge pool of office workers, some of the results obtained are surprisingly what I have expected after staying in Hue.