Introduction

Athens and Thessaloniki are the two major cities in Greece. Together, they are inhibited by the 50% of the Greek population. Both cities were struck by the financial crisis but recently have started to come back in all sorts of aspects such as the residential, tourism, education and job employment. They have also an extended history over the centuries and as such they attract tourist both from abroad and within Greece.

Athens is the capital and largest city of Greece. Athens dominates the Attica region and is one of the world's oldest cities, with its recorded history spanning over 3,400 years and its earliest human presence starting somewhere between the 11th and 7th millennium BC. The Athens Urban Area today consists of 40 municipalities, 35 of which make up what is referred to as the Greater Athens municipalities, located within 4 regional units (North Athens, West Athens, Central Athens, South Athens); and a further 5, which make up the Greater Piraeus municipalities, located within the regional unit of Piraeus as mentioned above. The densely built up urban area of the Greek capital sprawls across 412 km2 (159 sq mi) throughout the Attica Basin and has a total population of 3,074,160 (in 2011).

Thessaloniki is the second-largest city in Greece, with over 1 million inhabitants in its metropolitan area, and the capital of Greek Macedonia, the administrative region of Central Macedonia and the Decentralized Administration of Macedonia and Thrace.[Its nickname is "the co-capital", a reference to its historical status as the or "co-reigning" city of the Eastern Roman (Byzantine) Empire, alongside Constantinople.

Obejective

In this project, we will aim to to study:

1. the similarity or dissimilarity of both cities
2. classification of the areas inside the city to determine whether it is suitable for locals or it is more touristic. To achieve that we will utilize data from Foursquare and use Machine Learning techniques such as clustering and segmentation.

Data

The data was collected from a Greek webpage specializing in information 'https://www.xo.gr/greek-postal-codes-zips'.
As next step, the data was cleaned and saved in csv files in order to be easy to handle.
The data sets to be used can be found in the project repository:

1. Athens - https://github.com/amstersam/Coursera_CapStone_FInal_Project/blob/master/Athens.csv
2. Thessaloniki - https://github.com/amstersam/Coursera_CapStone_FInal_Project/blob/master/Thessaloniki.csv

To start, let's get and look at the data. I've already upload it on Jupyter, so let us have a look and load it to dataframe.

```
[112]: #import the required library
import numpy as np
import pandas as pd

#read csv file contain KL data
df_athens = pd.read_csv('Athens.csv')
df_athens.head()
```

[112]:

	District	Area	PostCode
0	North Suburbs	EKALI	14578
1	North Suburbs	NEA ERITHREA	14671
2	North Suburbs	PEFKI	15121
3	North Suburbs	MAROUSI	15122
4	North Suburbs	MELISIA	15127

We can explore the data with some basic queries in order to understand where we can focus and which areas may be more convenient to work with.

```
[113]: #examine data
print('Athens dataframe has {} districts and {} areas.'.format(
        len(df_athens['District'].unique()),
        df_athens.shape[0]
    )
)

#grouping data to find District with highest number of area
df_athens.groupby('District').count()
```

```
Athens dataframe has 6 districts and 50 areas.
```

[113]:

District	Area	PostCode
Center	4	4
East Suburbs	10	10
North Suburbs	11	11
South Suburbs	11	11
South Suburbs Suburbs	2	2
West Suburbs	12	12

As we can see below, we have loaded as well an extra csv file containing all Greek postal codes and their equivalent latitude and longitude values (https://github.com/MentatInnovations/grpostcodes).

[131]:

```
#read and load JB data
df_geo = pd.read_csv('postcode_lat_long_gr.csv')
df_geo.head()
```

[131]:

	tk	lat	lon
0	10021	37.974511	23.733071
1	10431	37.984438	23.728117
2	10432	37.987244	23.726373
3	10433	37.978381	23.780513
4	10434	37.993469	23.727507

We could have used the geocoder instead but as long as we have everything we need we shall proceed faster. Now, we are in a position to perform a join and assign coordinates to every point and we can take a quick look at the results.

[117]:

```
athens_full = pd.merge(df_athens, df_geo, how='left',
         left_on='PostCode', right_on='tk', validate="1:1")
athens_full.drop(labels='tk', axis=1, inplace=True)
athens_full.head()
```

[117]:

	District	Area	PostCode	lat	lon
0	North Suburbs	EKALI	14578	38.100877	23.835080
1	North Suburbs	NEA ERITHREA	14671	38.092649	23.820048
2	North Suburbs	PEFKI	15121	38.052949	23.790560
3	North Suburbs	MAROUSI	15122	38.054956	23.807655
4	North Suburbs	MELISIA	15127	38.057229	23.833602

[118]:

```
df_athens = df_athens.drop_duplicates().reset_index(drop=True)
```

```
[32]: df_athens
```

[32]:

	District	Area	PostCode
0	North Suburbs	EKALI	14578
1	North Suburbs	NEA ERITHREA	14671
2	North Suburbs	PEFKI	15121
3	North Suburbs	MAROUSI	15122
4	North Suburbs	MELISIA	15127
5	North Suburbs	HALANDRI	15231
6	North Suburbs	VRILISIA	15235
7	North Suburbs	PENTELI	15236
8	North Suburbs	FILOTHEI	15237
9	North Suburbs	HALANDRI - PATIMA	15238
10	North Suburbs	AGIA PARASKEVI	15341
11	East Suburbs	GERAKAS	15344
12	East Suburbs	STAVROS	15345
13	East Suburbs	ANTHOUSA	15349
14	East Suburbs	PALINI	15351
15	East Suburbs	GLIKA NERA	15354
16	East Suburbs	PSIHIKO	15451
17	East Suburbs	HOLARGOS	15561
18	East Suburbs	PAPAGOU	15669
19	Center	ZOGRAFOU	15771
20	Center	KESARIANI	16121
21	Center	VIRONAS	16231
22	South Suburbs	ILIOUPOLI	16341
23	South Suburbs	ARGIROUPOLI	16450
24	South Suburbs	GLIFADA	16561
25	South Suburbs	VOULIAGMENI	16671
26	South Suburbs Suburbs	VARI	16672

We shall do the same for Thessaloniki.

```
[132]: #read and load Thessaloniki data
df_thessaloniki = pd.read_csv('Thessaloniki.csv')
df_thessaloniki.head()
```

[132]:

	District	Area	PostCode
0	South-east Suburbs	HARILAOU	54248
1	South-east Suburbs	HARILAOU	54249
2	South-east Suburbs	KATO TOUMBA	54250
3	South-east Suburbs	KOSTADINOUPOLITIKA	54351
4	South-east Suburbs	ANO TOUMBA	54352

```
[116]: #examine
print('Thessaloniki dataframe has {} districts and {} areas.'.format(
        len(df_thessaloniki['District'].unique()),
        df_thessaloniki.shape[0]
    )
)

#group by district
df_thessaloniki.groupby('District').count()
```

```
Thessaloniki dataframe has 4 districts and 29 areas.
```

[116]:

District	Area	PostCode
CENTER	3	3
North Suburbs	2	2
South-east Suburbs	14	14
West Suburbs	10	10

We also change the columns names to be more homogenous from the aspect of detail.

```
[134]: thessaloniki_full = pd.merge(df_thessaloniki, df_geo, how='left',
            left_on='PostCode', right_on='tk', validate="1:1")
thessaloniki_full.drop(labels='tk', axis=1, inplace=True)
thessaloniki_full.head()
```

[134]:

	District	Area	PostCode	lat	lon
0	South-east Suburbs	HARILAOU	54248	39.677124	22.580364
1	South-east Suburbs	HARILAOU	54249	40.602356	22.965986
2	South-east Suburbs	KATO TOUMBA	54250	40.602126	22.972760
3	South-east Suburbs	KOSTADINOUPOLITIKA	54351	40.617032	22.974236
4	South-east Suburbs	ANO TOUMBA	54352	40.612080	22.983126

```
[135]: thessaloniki_full.rename(columns={'lat': 'Latitude'}, inplace=True)
```

```
[136]: thessaloniki_full.rename(columns={'lon': 'Longitude'}, inplace=True)
```

Methodology

In this project, we will try to use the basic methodology, which was presented during Week 3 and the respective lab.

Then we will use the Foursquare API to explore neighborhoods in both cities, Athens and Thessaloniki. After that, explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters K-means clustering algorithm will be use to complete this task. And also, the Folium library to visualize the neighborhoods in Athens and Thessaloniki and their emerging clusters.

Based on dataframe analysis above, we found out that North Suburbs area in Athens and West Suburbs area in Thessaloniki contain the largest area within and from an online research we found out that the population density there is quite high.

```
EKALI
NEA ERITHREA
PEFKI
MAROUSI
MELISIA
HALANDRI
VRILISIA
PENTELI
FILOTHEI
HALANDRI - PATIMA
AGIA PARASKEVI
(249, 7)
```

[155]:

	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	EKALI	38.100877	23.835080	Πλατεία Μουσσών	38.102017	23.832907	Plaza
1	EKALI	38.100877	23.835080	Nike Store	38.097062	23.836865	Sporting Goods Shop
2	NEA ERITHREA	38.092649	23.820048	Family Project	38.091214	23.819720	Bakery
3	NEA ERITHREA	38.092649	23.820048	Πανερυθραϊκός	38.090214	23.818901	Souvlaki Shop
4	NEA ERITHREA	38.092649	23.820048	Κοράλλι	38.092371	23.820113	Ouzeri

one_W5.ipynb
: Python 3

```
HARILAOU
HARILAOU
KATO TOUMBA
KOSTADINOUPOLITIKA
ANO  TOUMBA
TOUMBA
TOUMBA
NEA REDESTOS
ANALIPSI
ANTHEON
KALAMARIA
PANORAMA
TRIANDRIA
THERMI
(414, 7)
```

[159]:

	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	HARILAOU	40.602356	22.965986	Πατσάς 92	40.602313	22.964555	Patsa Restaurant
1	HARILAOU	40.602356	22.965986	Fregio	40.603232	22.969078	Ice Cream Shop
2	HARILAOU	40.602356	22.965986	Σείριος	40.600558	22.964317	Liquor Store
3	HARILAOU	40.602356	22.965986	Εστία	40.602769	22.969260	Bakery
4	HARILAOU	40.602356	22.965986	Θανάσης & Υιοί	40.603772	22.968866	Grilled Meat Restaurant

Analysis of Athens

Firstly, we will get the most common venues for Athens.

[163]:

	Area	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
0	AGIA PARASKEVI	Bakery	Clothing Store	Bar	Pharmacy	Pizza Place	Cosmetics Shop	Café	Mobile Phone Shop
1	EKALI	Plaza	Sporting Goods Shop	Grilled Meat Restaurant	Greek Restaurant	Dog Run	Donut Shop	Electronics Store	Farmers Market
2	FILOTHEI	Park	Plaza	Tennis Stadium	Stadium	Snack Place	French Restaurant	Dog Run	Donut Shop
3	HALANDRI	Gym / Fitness Center	Café	Bakery	Basketball Court	Burger Joint	Mobile Phone Shop	Coffee Shop	Furniture / Home Store
4	HALANDRI - PATIMA	Cupcake Shop	Park	Racetrack	Farmers Market	Steakhouse	Dog Run	Yoga Studio	French Restaurant

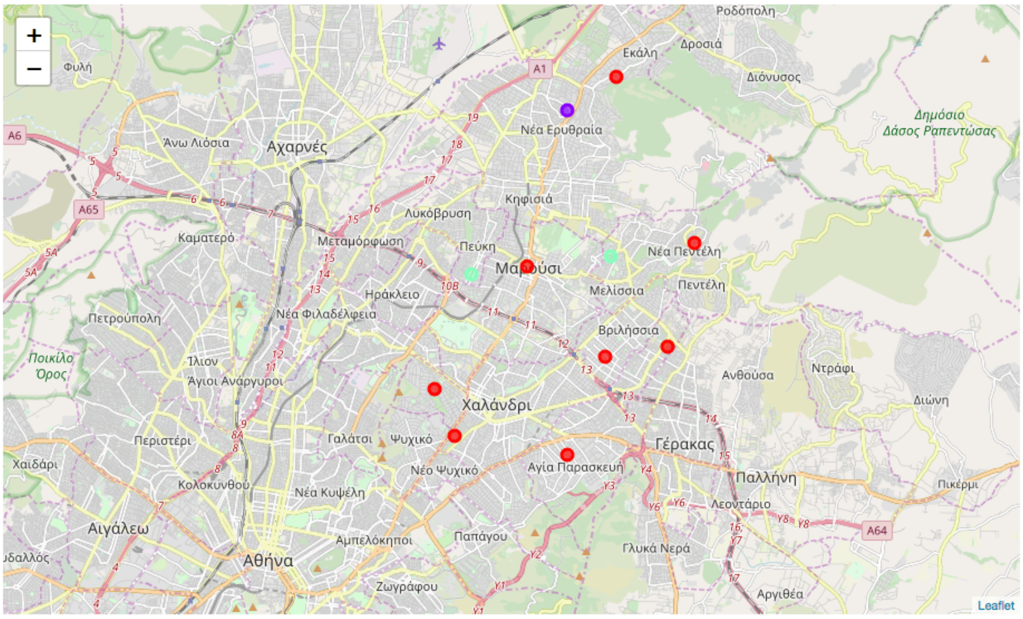
K-means for Athens

Secondly, we will move on to apply k-means.

[164]:

	District	Area	PostCode	Latitude	Longitude	Cluster Labels	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
0	North Suburbs	EKALI	14578	38.100877	23.835080	0	Plaza	Sporting Goods Shop	Grilled Meat Restaurant	Greek Restaurant	Dog Run	Donut Shop	Electronics Store
1	North Suburbs	NEA ERITHREA	14671	38.092649	23.820048	1	Dessert Shop	Bar	Nail Salon	Cocktail Bar	Italian Restaurant	Ice Cream Shop	Fire Station
2	North Suburbs	PEFKI	15121	38.052949	23.790560	2	Greek Restaurant	Bakery	Park	Café	Souvlaki Shop	Gym / Fitness Center	Grilled Meat Restaurant
3	North Suburbs	MAROUSI	15122	38.054956	23.807655	0	Café	Bakery	Bar	Coffee Shop	Souvlaki Shop	Multiplex	Plaza
4	North Suburbs	MELISIA	15127	38.057229	23.833602	2	Souvlaki Shop	Bakery	Café	Pharmacy	Seafood Restaurant	Grilled Meat Restaurant	Grocery Store

[166]:



Analysis of Thessaloniki

As far as Thessaloniki is concerned we will follow the same method.

[244]:

```python
#create the new dataframe and display the top 10 venues for each neighborhood
num_top_venues = 8

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Area']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
areas_venues_sorted = pd.DataFrame(columns=columns)
areas_venues_sorted['Area'] = thessaloniki_grouped['Area']

for ind in np.arange(thessaloniki_grouped.shape[0]):
    areas_venues_sorted.iloc[ind, 1:] = return_most_common_venues(thessaloniki_grouped.iloc[ind, :], num_top_venu

areas_venues_sorted.head()
```

[244]:

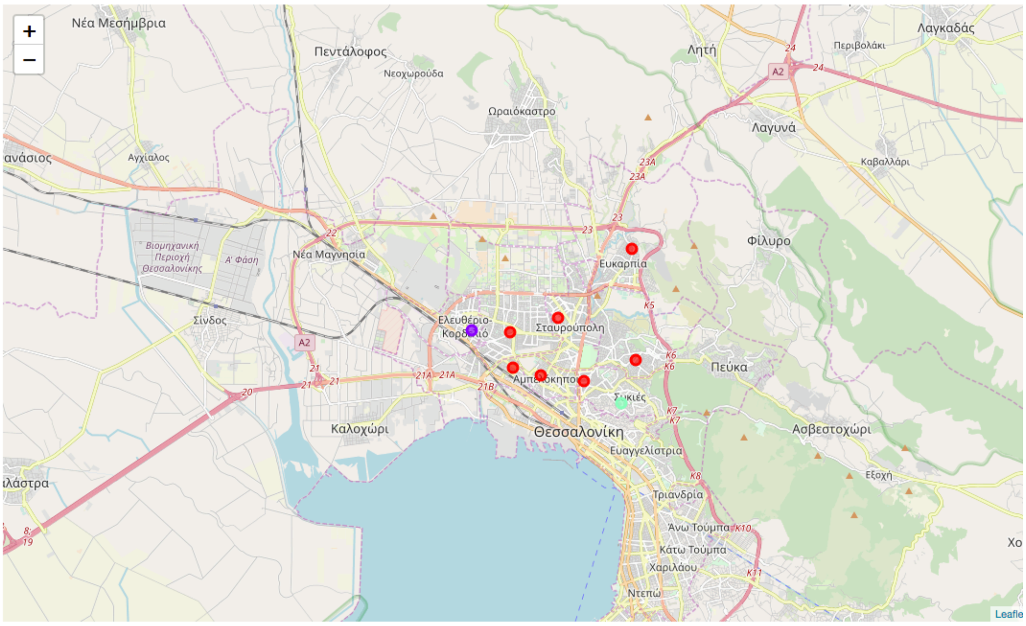
	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	AMPELOKIPI	Café	Bar	Mobile Phone Shop	Supermarket	Souvlaki Shop	Coffee Shop	Snack Place	Fast Food Restaurant
1	ELEUTHERIO - KORDELIO	Café	Supermarket	Coffee Shop	Fish Taverna	Jewelry Store	Meze Restaurant	Mobile Phone Shop	Electronics Store
2	EVOSMOS	Bar	Café	Snack Place	Meze Restaurant	Fast Food Restaurant	Greek Restaurant	Bakery	Mobile Phone Shop
3	MENEMENI	Pub	Soccer Field	Plaza	Pier	Seafood Restaurant	Café	Supermarket	Fast Food Restaurant
4	NEAPOLI	Café	Video Store	Gym	Electronics Store	Dessert Shop	Men's Store	Meze Restaurant	Mobile Phone Shop

K-means for Thessaloniki

[245]:

	District	Area	PostCode	Latitude	Longitude	Cluster Labels	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
0	West Suburbs	AMPELOKIPI	56121	40.655333	22.921503	0	Café	Bar	Mobile Phone Shop	Supermarket	Souvlaki Shop	Coffee Shop	Snack Place	Fast Food Restaurant
1	West Suburbs	MENEMENI	56122	40.657488	22.911517	0	Pub	Soccer Field	Plaza	Pier	Seafood Restaurant	Café	Supermarket	Fast Food Restaurant
2	West Suburbs	EVOSMOS	56224	40.667036	22.910492	0	Bar	Café	Snack Place	Meze Restaurant	Fast Food Restaurant	Greek Restaurant	Bakery	Mobile Phone Shop
3	West Suburbs	ELEUTHERIO - KORDELIO	56334	40.667525	22.896978	1	Café	Supermarket	Coffee Shop	Fish Taverna	Jewelry Store	Meze Restaurant	Mobile Phone Shop	Electronics Store
4	West Suburbs	STAVROUPOLI	56429	40.689740	22.954117	0	Gym	Grilled Meat Restaurant	Bakery	Greek Restaurant	Basketball Court	Plaza	Café	Salon / Barbershop

[250]:



[ ]:

Results

[251]:
```
#Cluster 1 for Athens
athens_merged.loc[athens_merged['Cluster Labels'] == 0, athens_merged.columns[[2] + list(range(5, athens_merged.shape[1]))]]
```

[251]:

	PostCode	Cluster Labels	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
0	14578	0	Plaza	Sporting Goods Shop	Grilled Meat Restaurant	Greek Restaurant	Dog Run	Donut Shop	Electronics Store	Farmers Market
3	15122	0	Café	Bakery	Bar	Coffee Shop	Souvlaki Shop	Multiplex	Plaza	Gaming Cafe
5	15231	0	Gym / Fitness Center	Café	Bakery	Basketball Court	Burger Joint	Mobile Phone Shop	Coffee Shop	Furniture / Home Store
6	15235	0	Park	Plaza	Convenience Store	Japanese Restaurant	Farmers Market	Pharmacy	Flower Shop	Coffee Shop
7	15236	0	Pharmacy	Plaza	Grilled Meat Restaurant	Greek Restaurant	Yoga Studio	Souvlaki Shop	Meze Restaurant	Pizza Place
8	15237	0	Park	Plaza	Tennis Stadium	Stadium	Snack Place	French Restaurant	Dog Run	Donut Shop
9	15238	0	Cupcake Shop	Park	Racetrack	Farmers Market	Steakhouse	Dog Run	Yoga Studio	French Restaurant
10	15341	0	Bakery	Clothing Store	Bar	Pharmacy	Pizza Place	Cosmetics Shop	Café	Mobile Phone Shop

[252]:
```
#Cluster 2 for Athens
athens_merged.loc[athens_merged['Cluster Labels'] == 1, athens_merged.columns[[2] + list(range(5, athens_merged.shape[1]))]]
```

[252]:

	PostCode	Cluster Labels	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
1	14671	1	Dessert Shop	Bar	Nail Salon	Cocktail Bar	Italian Restaurant	Ice Cream Shop	Fire Station	Health & Beauty Service

[253]:
```
#Cluster 3 for Athens
athens_merged.loc[athens_merged['Cluster Labels'] == 2, athens_merged.columns[[2] + list(range(5, athens_merged.shape[1]))]]
```

[253]:

	PostCode	Cluster Labels	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
2	15121	2	Greek Restaurant	Bakery	Park	Café	Souvlaki Shop	Gym / Fitness Center	Grilled Meat Restaurant	Pool
4	15127	2	Souvlaki Shop	Bakery	Café	Pharmacy	Seafood Restaurant	Grilled Meat Restaurant	Grocery Store	Gym

[254]:
```
#Cluster 1 for Thessaloniki
thessaloniki_merged.loc[thessaloniki_merged['Cluster Labels'] == 0, thessaloniki_merged.columns[[2] + list(range(5, thessaloniki_merged.sha
```

[254]:

	PostCode	Cluster Labels	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
0	56121	0	Café	Bar	Mobile Phone Shop	Supermarket	Souvlaki Shop	Coffee Shop	Snack Place	Fast Food Restaurant
1	56122	0	Pub	Soccer Field	Plaza	Pier	Seafood Restaurant	Café	Supermarket	Fast Food Restaurant
2	56224	0	Bar	Café	Snack Place	Meze Restaurant	Fast Food Restaurant	Greek Restaurant	Bakery	Mobile Phone Shop
4	56429	0	Gym	Grilled Meat Restaurant	Bakery	Greek Restaurant	Basketball Court	Plaza	Café	Salon / Barbershop
5	56431	0	Café	Taverna	Bar	Fast Food Restaurant	Plaza	Snack Place	Coffee Shop	Mobile Phone Shop
6	56532	0	BBQ Joint	Gym / Fitness Center	Snack Place	Park	Pizza Place	Playground	Plaza	Café
8	56727	0	Café	Video Store	Gym	Electronics Store	Dessert Shop	Men's Store	Meze Restaurant	Mobile Phone Shop
9	57002	0	Bar	Pub	Greek Restaurant	Video Store	Grilled Meat Restaurant	Frozen Yogurt Shop	Fried Chicken Joint	Fish Taverna

[255]:
```
#Cluster 2 for Thessaloniki
thessaloniki_merged.loc[thessaloniki_merged['Cluster Labels'] == 1, thessaloniki_merged.columns[[2] + list(range(5, thessaloniki_merged.sha
```

[255]:

	PostCode	Cluster Labels	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
3	56334	1	Café	Supermarket	Coffee Shop	Fish Taverna	Jewelry Store	Meze Restaurant	Mobile Phone Shop	Electronics Store

[ ]:

[256]:
```
#Cluster 3 for Thessaloniki
thessaloniki_merged.loc[thessaloniki_merged['Cluster Labels'] == 2, thessaloniki_merged.columns[[2] + list(range(5, thessaloniki_merged.sha
```

[256]:

	PostCode	Cluster Labels	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
7	56625	2	Café	Gym	Bougatsa Shop	Theater	Supermarket	Restaurant	Beer Bar	Scenic Lookout

Discussion

Based on cluster for each city above, we believe that classification for each cluster can be done better with calculation of venues categories (most common) in each city. Referring to each cluster, we can determine to some extend what represents each cluster by using Foursquare - Most Common Venue data.

We can summarize the following based on the clustering produced by our code.

Cluster	Athens	Thessaloniki
1	Residential area	Residential area
2	Residential area	Residential area
3	Residential area	Residential area

Unfortunately, the kind of data provided by Foursquare reaches to a certain extend, meaning that there are several limitations as far as the statistical or other quantitative aspect of the data is concerned. In other words, we have no information about the range of ages of the people visiting those venues, or the income and educational level of those people. Those categorical variables could have another kind of value to our project leading perhaps to several remarks or observations.

Conclusion

Using the Foursquare API, we can capture data of common places for the two major Greek cities. Going back to our initial objectives we could almost safely come to the conclusion that the two cities and more specifically the North suburbs of Athens and the West suburbs of Thessaloniki are similar and both be classified as residential areas.

However, to declare both cities are similar or dissimilar basef on common venues visited is quite difficult. Both cities are similar in some venues also dissimilar in certain venues. Consequently, both areas consist of many different districts that are quite attractive for locals mostly, since they have to offer venues to cover the daily needs of the people but also sports, entertainment and other.