Capstone project: The Battle of Neighborhoods

Table of contents

Introduction/Business Problem

In this project I would like to make a description of the activities of the most important coastal cities of southern Italy and how they affect their economy.
For this purpose I will consider the most important provinces and the most famous tourist resorts of Campania, Puglia, Calabria and the two big islands Sicily and Sardinia.

Data description

Regarding the data on the activities of the cities I will use the API of Foursquare within the geocoder Nominatim.While regarding the economic aspect the data will be taken from the website of 'Sole 24 ore', the most important Italian economic newspaper, year 2016. This is the reference link: https://lab24.ilsole24ore.com/mappaRedditi/redditiTabelle.html

Metodhology

After having loaded the libraries that we will use for the project and having defined the various functions to speed up the procedures, we use the FourSquare query that will give us the data of all the various activities near the various cities, which are the Sicilian Palermo, Catania, Siracusa, Ragusa, Trapani, Agrigento and Messina, the Sardinian Cagliari and Sassari, the Campanian Naples and Salerno and the Apulian Bari, Lecce, Brindisi and Taranto. We will also add the data of the annual income of the various cities (which in Italy is called 'PIL')

[3]:
```
import numpy as np # library to handle data in a vectorized manner
import pandas as pd # library for data analsysis
import json # library to handle JSON files
from pandas.io.json import json_normalize
import requests
!conda install -c conda-forge geopy --yes # uncomment this line if you haven't completed the Foursqua
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values
# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors
# import k-means from clustering stage
from sklearn.cluster import KMeans
#!conda install -c conda-forge folium=0.5.0 --yes # uncomment this line if you haven't completed the
import folium # map rendering library
print("Intallation done")
```

E261 at least two spaces before inline comment (pycodestyle)

[18]:

	PIL	Loc	Airport	Bar	Beach	Bed & Breakfast	Cafeteria	Café	Castle	Chocolate Shop	...	Gym	Mobile Phone Shop	Sports Bar	Hotel Bar
0	17925	Ragusa	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
1	17925	Ragusa	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
2	17925	Ragusa	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
3	17925	Ragusa	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
4	17925	Ragusa	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1363	21058	Taranto	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
1364	21058	Taranto	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
1365	21058	Taranto	0.0	0.0	0.0	1.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
1366	21058	Taranto	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0
1367	21058	Taranto	0.0	0.0	0.0	0.0	0.0	0	0.0	0.0	...	0.0	0.0	0.0	0.0

1368 rows × 149 columns

At this point we obtain the table that gives us for each city the average of the various activities

At this point we obtain the table that gives us for each city the average of the various activities

[19]:
```
or_df_grouped=or_df.groupby('Loc').mean().reset_index()
or_df_grouped
```

9]:

	Loc	PIL	Airport	Bar	Beach	Bed & Breakfast	Cafeteria	Café	Castle	Chocolate Shop	...	Gym	Mobile Phone Shop
0	Agrigento	20881	0.000000	0.000000	0.159420	0.000000	0.000000	0.086957	0.000000	0.000000	...	0.00	0.00
1	Bari	22947	0.000000	0.030000	0.020000	0.000000	0.000000	0.080000	0.000000	0.000000	...	0.01	0.01
2	Brindisi	19816	0.000000	0.013699	0.082192	0.000000	0.000000	0.013699	0.013699	0.000000	...	0.00	0.00
3	Cagliari	25681	0.000000	0.000000	0.056818	0.000000	0.011364	0.045455	0.000000	0.000000	...	0.00	0.00
4	Catania	20179	0.000000	0.000000	0.020000	0.030000	0.000000	0.040000	0.020000	0.010000	...	0.00	0.00
5	Catanzaro	21487	0.000000	0.022222	0.122222	0.000000	0.000000	0.088889	0.000000	0.000000	...	0.00	0.00
6	Cosenza	21131	0.000000	0.000000	0.142857	0.020408	0.020408	0.102041	0.000000	0.000000	...	0.00	0.00
7	Lecce	23420	0.000000	0.021505	0.075269	0.000000	0.000000	0.086022	0.010753	0.000000	...	0.00	0.00
8	Messina	21534	0.000000	0.000000	0.016393	0.000000	0.000000	0.180328	0.000000	0.000000	...	0.00	0.00
9	Napoli	22434	0.000000	0.000000	0.020000	0.010000	0.000000	0.060000	0.030000	0.000000	...	0.00	0.00
10	Palermo	22264	0.000000	0.031250	0.083333	0.020833	0.000000	0.052083	0.000000	0.000000	...	0.00	0.00
11	Ragusa	17925	0.010417	0.010417	0.041667	0.020833	0.010417	0.104167	0.010417	0.010417	...	0.00	0.00
12	Reggio Calabria	20079	0.000000	0.000000	0.129630	0.000000	0.000000	0.092593	0.018519	0.000000	...	0.00	0.00
13	Salerno	23888	0.000000	0.020408	0.081633	0.020408	0.000000	0.040816	0.020408	0.000000	...	0.00	0.00
14	Sassari	22165	0.000000	0.000000	0.125000	0.000000	0.000000	0.125000	0.000000	0.000000	...	0.00	0.00
15	Siracusa	20395	0.000000	0.000000	0.122449	0.010204	0.000000	0.091837	0.000000	0.000000	...	0.00	0.00
16	Taranto	21058	0.000000	0.000000	0.000000	0.022727	0.000000	0.068182	0.022727	0.000000	...	0.00	0.00
17	Trapani	18318	0.000000	0.010000	0.130000	0.020000	0.000000	0.050000	0.010000	0.000000	...	0.00	0.00

8 rows × 149 columns

Results

We are interested in knowing which activity has the greatest impact on income, so we use the corr() function that Phyton gives us to measure the correlation between PIL and the other components

[29]:
```
or_df_grouped.corr()['PIL'].sort_values().head(10)
```

[29]:
```
Comfort Food Restaurant   -0.628433
Deli / Bodega             -0.612401
City                      -0.544690
Diner                     -0.484220
Sicilian Restaurant       -0.483076
Vineyard                  -0.456317
Hobby Shop                -0.456317
Airport                   -0.456317
Chocolate Shop            -0.455104
Neighborhood              -0.455104
Name: PIL, dtype: float64
```

This data concerns the activities that negatively affect income, strangely enough we notice that among the 10 activities with greater negative influence there are those related to food at various levels. Let's see now which are the 10 activities with the greatest positive influence on income

[31]:
```
or_df_grouped.corr()['PIL'].sort_values(ascending=False).head(11)
```

[31]:
```
PIL                    1.000000
Coffee Shop            0.680327
Dive Bar               0.565296
Rock Climbing Spot     0.555653
Beach Bar              0.555653
College Soccer Field   0.555653
Flower Shop            0.555653
Noodle House           0.555653
Art Gallery            0.555653
Garden                 0.508428
Tea Room               0.498639
Name: PIL, dtype: float64
```

We note that activities such as bars and tea rooms and activities related to floriculture have a positive influence on income.

Conclusion

It is clear that 'light' activities with little investment but of good interest have a positive impact on income, unlike activities that are too costly.