Start-up recommendation: Province of Rome, Italy

(Capstone project for IBM Data Science Professional Certification)

Introduction

The city of Rome has a population of 2856133 but the province of Rome has 1 486079 distributed in about 121 towns around Rome (radius of 50km).

Some of these towns around Rome has large populations that could make a business viable. Office space and taxes in Rome itself is high and capital needed for start-up would be prohibitive for a small business.

We explore the possibility of opening a business in the province of Rome.

Business question and approach:

The business question that attempt to answer is:

What would be the best business to start in the province of Rome (outside the capital) and which towns would be most appropriate?

<u>The approach will be:</u>

1. Find data for all the towns in the province of Rome
2. Analyse all the towns for the following aspects:
    a. Population
    b. Venues already recorded
3. find all the data on Foursquare for those towns
4. Use K-means to cluster the business venues
5. Select the towns with the highest population and propose as business the types that are the least common in those towns.

Data used for project

1. Get the geospatial co-ordinates for the towns surrounding Rome:

Publicly available at: http://www.dossier.net/utilities/coordinate-geografiche/provincia-roma.htm'
Scrape this website with beautifulsoup to get the data.

2. Get the demographical of each town:

Publicly available at: https://www.tuttitalia.it/lazio/provincia-di-roma/36-comuni/popolazione/'
Scrape this website with beautifulsoup to get the data.

3. Combine the **geospatial and demographical** data in a pandas dataframe

4. Draw a map of the Province

5. Get the **Foursquare data** for all towns in the province

6. Explore the venues in the Rome Province

7. Transform the data so that **kmeans clustering** can be applied to it.

8. Select the value of k most appropriate using the elbow method.

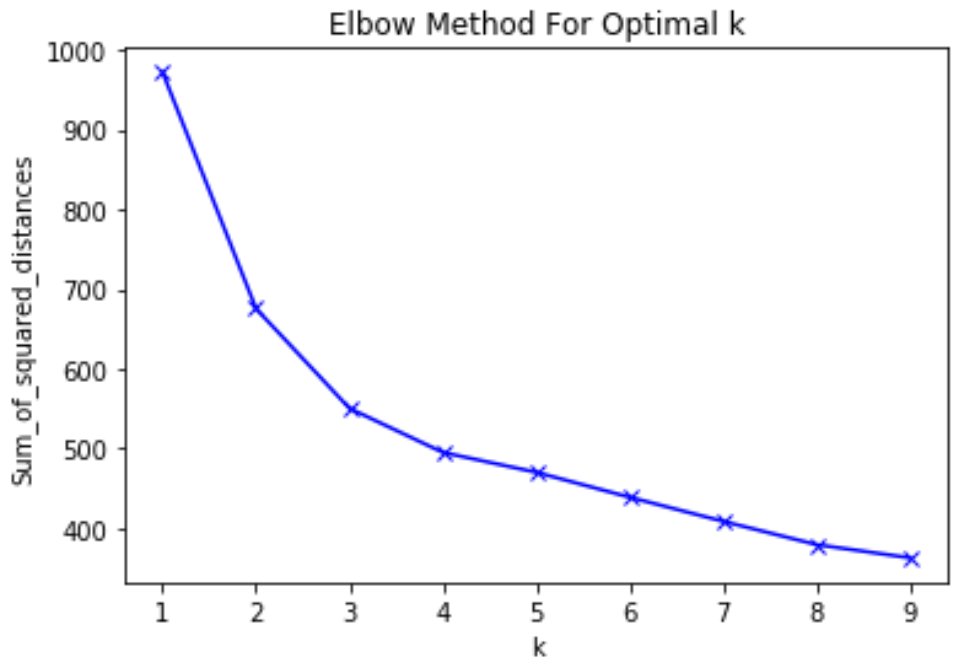
9. Apply **kmeans clustering** to get the most common venues in each town.

10. Map the clusters

11. Select the **least common venue** in the towns with the **highest population**. These will be the towns and businesses proposed to a prospective start-up.

Methodology

1. Scraped the website http://www.dossier.net/utilities/coordinate-geografiche/provincia-roma.htm to obtain the geospatial coordinates for the towns in the Province of Rome.
2. The next step was to get the demographic information for this towns. This was available on the website. https://www.tuttitalia.it/lazio/provincia-di-roma/36-comuni/popolazione/
3. Combined the data from the 2 websites in a one dataframe.
4. Fetched the venue data from Foursquare in a radius of 50km around Rome for each town (except Rome itself) in the combined dataframe. This yielded 559 venues and 118 different categories of venue.
5. I then prepared the data to be used in a kmeans algorithm by one-hot encoding the data retrieved from Foursquare.(Town/Neighborhood and 118 categories)
6. I applied various values of k to the kmeans algorithm to decide which would be the best. This method did not produce a very clear result but I decided to use a k of 7.



7. Next, we order the towns in order of descending population size and selected the top 5 towns. The result is as shown below.

	Comune	Pop	Area	Densitykm2	Alt	Lat	Long
0	Guidonia Montecelio	89671.0	79,47	1.128	105	41.9998179	12.7262970
1	Fiumicino	80470.0	213,89	376	1	41.7715258	12.2300032
2	Pomezia	63792.0	86,57	737	108	41.6692930	12.5017934
3	Tivoli	56472.0	68,65	823	235	41.9635786	12.7982722
4	Anzio	55101.0	43,65	1.262	3	41.4479468	12.6290520

8. Look at the venues in each of the towns selected.

	Comune	Long_y	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	9th Most Common Venue	10th Most Common Venue
0	Guidonia Montecelio	12.7262970	0	Supermarket	Bar	Playground	Pub	Food	Flower Shop	Fireworks Store	Financial or Legal Service	Fast Food Restaurant	Winery
1	Fiumicino	12.2300032	2	Seafood Restaurant	Italian Restaurant	Restaurant	Café	Hotel	Bed & Breakfast	Pub	Japanese Restaurant	Mediterranean Restaurant	Harbor / Marina
2	Pomezia	12.5017934	4	Café	Winery	Pizza Place	Outdoors & Recreation	Music Venue	Italian Restaurant	Food	Food Court	Electronics Store	Event Service
3	Tivoli	12.7982722	3	Italian Restaurant	Restaurant	Plaza	Historic Site	Park	Café	Scenic Lookout	Pub	Sandwich Place	Pizza Place
4	Anzio	12.6290520	6	Seafood Restaurant	Café	Italian Restaurant	Ice Cream Shop	Lighthouse	Pizza Place	Multiplex	Plaza	Movie Theater	Food

Results

We select the second and last least **common** venue in each of the towns to propose are a potential business to start. This yield the table below as result.

The proposal would therefore be to start a Fast Food restaurant or a winery in the town of Guidonia Montecelio.

	Comune	9th Most Common Venue	10th Most Common Venue
0	Guidonia Montecelio	Fast Food Restaurant	Winery
1	Fiumicino	Mediterranean Restaurant	Harbor / Marina
2	Pomezia	Electronics Store	Event Service
3	Tivoli	Sandwich Place	Pizza Place
4	Anzio	Movie Theater	Food

Discussion and Conclusion

For the town Anzio the model suggest a Movie theatre as one option, which is viable. The second option is too vague to be useful.

The data retrieved from Foursquare is not very consistent about types of venues. For instance, Mediterranean Restaurant and Italian restaurant are different categories.

It also includes things like “scenic lookout” that is not a business and categories like “food” which is too general.

For a further refinement of this project, I would recategorize some of the venues to get more consistent groupings.

I might also consider getting more detailed information regarding the population age distribution in each town to pick a town with the highest number of economically active residents.