Analysis of porto city areas

A study of area types near metro stations

Introduction

Porto Metro is one of the largest public transit systems in Portugal, being one of the crucial means of transportation for work and leisure trips in the city.

For this project, we want to look at the neighborhoods surrounding metro stations and classify them. Some neighborhoods are mostly residential, some have more business or commercial spaces surrounding them. The venues closest to a station determine why and how people use it. E.g. if there are no professional places in a neighborhood its residents are likely to travel to other areas for work. This creates daily migrations of people.

By analyzing this data we can classify stations by primary usage. This data is useful for city planners to determine where from and where to people are most likely to travel for work and leisure. This can help plan further extension of the network and find places for new development.

Data

We'll need data on the location of stations and on the venues closest to them:

1. List of stations and their geographical coordinates using Foursquare API, with category = Metro Station.
2. Foursquare API to explore venue and venue types surrounding each station.

We'll be querying the number of venues of each type in each category in a radius around each station.

Methodology

To obtain the metro location data, we can use the Foursquare API with category ID corresponding to metro stations in a radius from Porto center, obtaining their name and location.

We can then use the Foursquare API with the most common category ID's to query the venues of each category in a specific radius around each station.

Finally, with both group of data, we can create clusters classifying each metro station area, providing insights to the area types in Porto.

Obtaining metro data

Using Foursquare API, we can access the Search Endpoint, and, passing the parameter categoryId ('4bf58dd8d48988d1fd931735'), corresponding to metro stations, the Porto city location and a radius around the city, we can obtain a list of all the metro stations.

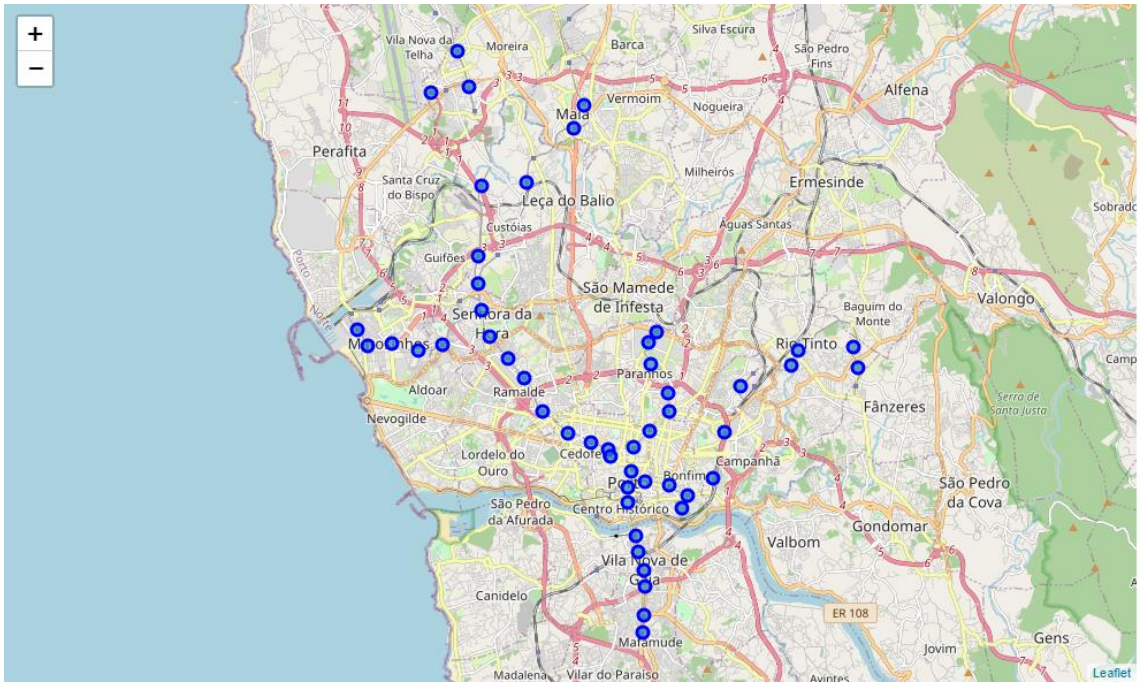
Response consists of an array, with each position having the following data structure:


```
{'id': '4bbc46a0e436ef3bc35e5664',
 'name': 'Metro Trindade [A,B,C,D,E,F]',
 'location':
        {'address': 'R. de Camões',
         'crossStreet': 'Lg. Dr. Tito Fontes',
         'lat': 41.15211447651871,
         'lng': -8.609718140360703,
         'labeledLatLngs': [{'label': 'display', 'lat': 41.15211447651871, 'lng': -8.609718140360703}],
         'distance': 298,
         'cc': 'PT',
         'city': 'Porto',
         'state': 'Porto',
         'country': 'Portugal',
         'formattedAddress': ['R. de Camões (Lg. Dr. Tito Fontes)', 'Porto', 'Portugal']},
 'categories':
        [{'id': '4bf58dd8d48988d1fd931735',
         'name': 'Metro Station',
         'pluralName': 'Metro Stations',
         'shortName': 'Metro',
         'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/travel/subway_', 'suffix': '.png'}, 'primary': True}],
         'referralId': 'v-1593600245',
         'hasPerk': False}
```


From this output array, we can clean the data and obtain the following dataframe:

	Name	Latitude	Longitude
0	Metro Trindade [A,B,C,D,E,F]	41.152114	-8.609718
1	Metro Aliados [D]	41.148357	-8.610925
2	Metro São Bento [D]	41.144993	-8.610923
3	Metro Faria Guimarães [D]	41.157524	-8.609059
4	Metro Bolhão [A,B,C,E,F]	41.149782	-8.605819

Using Folium library, we can create markers on a map for each station.



Obtaining venue data

First, we need to retrieve the most common venue types in Foursquare, using the categories endpoint, in order to avoid multiple smaller categories, as Foursquare has groups corresponding to macro category types.

- Arts & Entertainment (4d4b7104d754a06370d81259)
- College & University (4d4b7105d754a06372d81259)
- Event (4d4b7105d754a06373d81259)
- Food (4d4b7105d754a06374d81259)
- Nightlife Spot (4d4b7105d754a06376d81259)
- Outdoors & Recreation (4d4b7105d754a06377d81259)
- Professional & Other Places (4d4b7105d754a06375d81259)
- Residence (4e67e38e036454776db1fb3a)
- Shop & Service (4d4b7105d754a06378d81259)
- Travel & Transport (4d4b7105d754a06379d81259)

Having this category list and metro station list, we can make requests in the Foursquare API to get the venues of each type for each station, obtaining this dataframe.

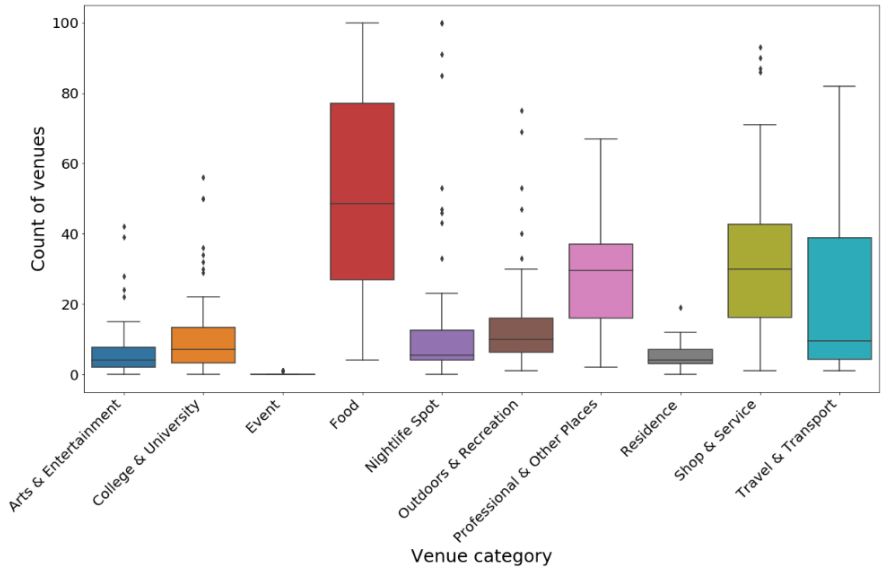
	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue SubCategory
0	Metro Trindade [A,B,C,D,E,F]	41.152114	-8.609718	Cinema Trindade	41.150515	-8.611822	Arts & Entertainment	Indie Movie Theater
1	Metro Trindade [A,B,C,D,E,F]	41.152114	-8.609718	Coliseu do Porto	41.147241	-8.605157	Arts & Entertainment	Music Venue
2	Metro Trindade [A,B,C,D,E,F]	41.152114	-8.609718	Fantasporto	41.147710	-8.609245	Arts & Entertainment	Indie Movie Theater
3	Metro Trindade [A,B,C,D,E,F]	41.152114	-8.609718	Maus Hábitos	41.146657	-8.605855	Arts & Entertainment	Music Venue
4	Metro Trindade [A,B,C,D,E,F]	41.152114	-8.609718	Teatro Carlos Alberto	41.148972	-8.615094	Arts & Entertainment	Theater

Using one hot encoding and summing the columns, we can create an easier to analyze dataframe.

	Station	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Metro 24 de Agosto [A,B,C,E,F]	15	14	0	100	23	16	37	3	61	76
1	Metro Aeroporto [E]	0	0	0	27	2	3	8	0	24	41
2	Metro Aliados [D]	39	36	0	100	100	69	58	9	69	80
3	Metro Araújo [C]	1	3	0	5	1	4	18	2	11	2
4	Metro Bolhão [A,B,C,E,F]	28	29	0	76	53	53	51	6	70	82

Exploratory Analysis

Looking at the data, we can see for example that Aliados station has the highest number of Nightlife Spots (100) while the Airport station has only 2. It is consisted with the goal for each station, with Airport being oriented for travel and Aliados station being oriented to support one of the city main leisure and tourist spots.



Looking at the number of venues as boxplots (showing the average count, spread and outliers), we can see that the most frequent venue categories are Food, Shop & Service, Professional & Other Places and Travel & Transport.

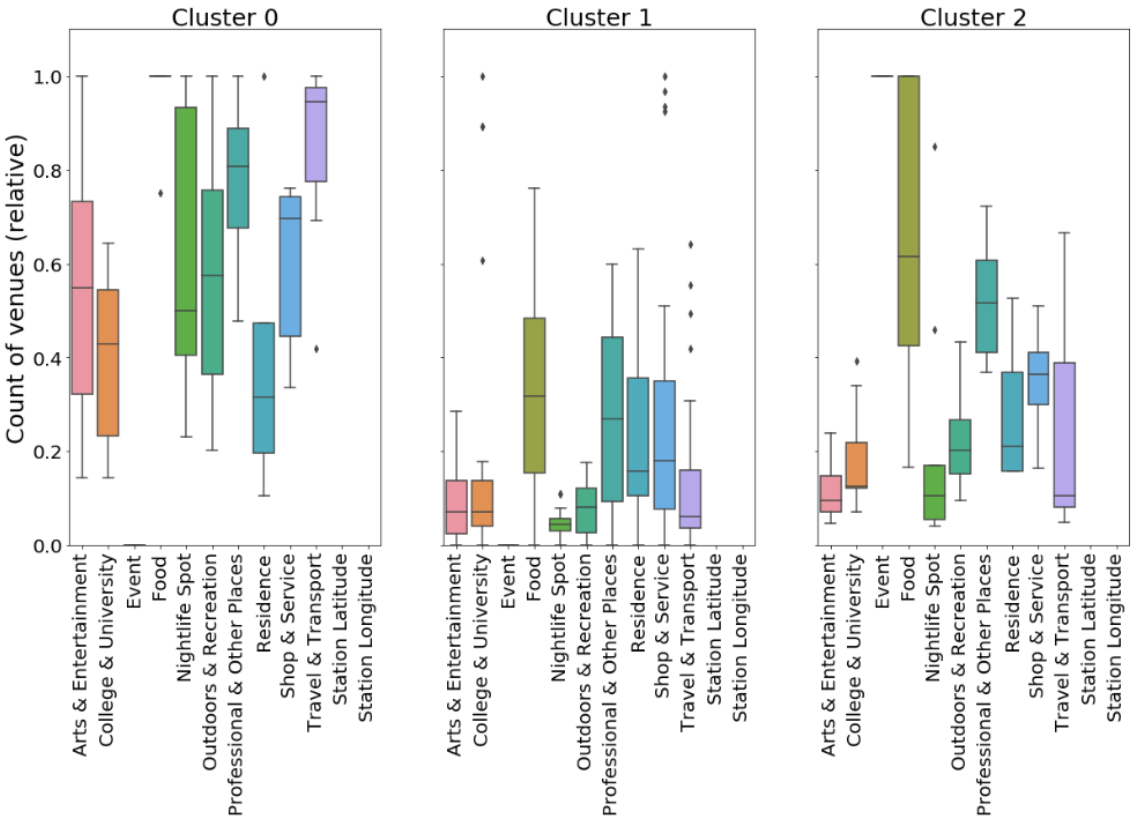
Clustering

We'll be using k-means clustering. Before that, we normalize the data using min-max scaling (scale count of venues from 0 to 1 where 0 is the lowest value in a set and 1 is highest). This both normalizes the data and provides an easy to interpret score at the same time.

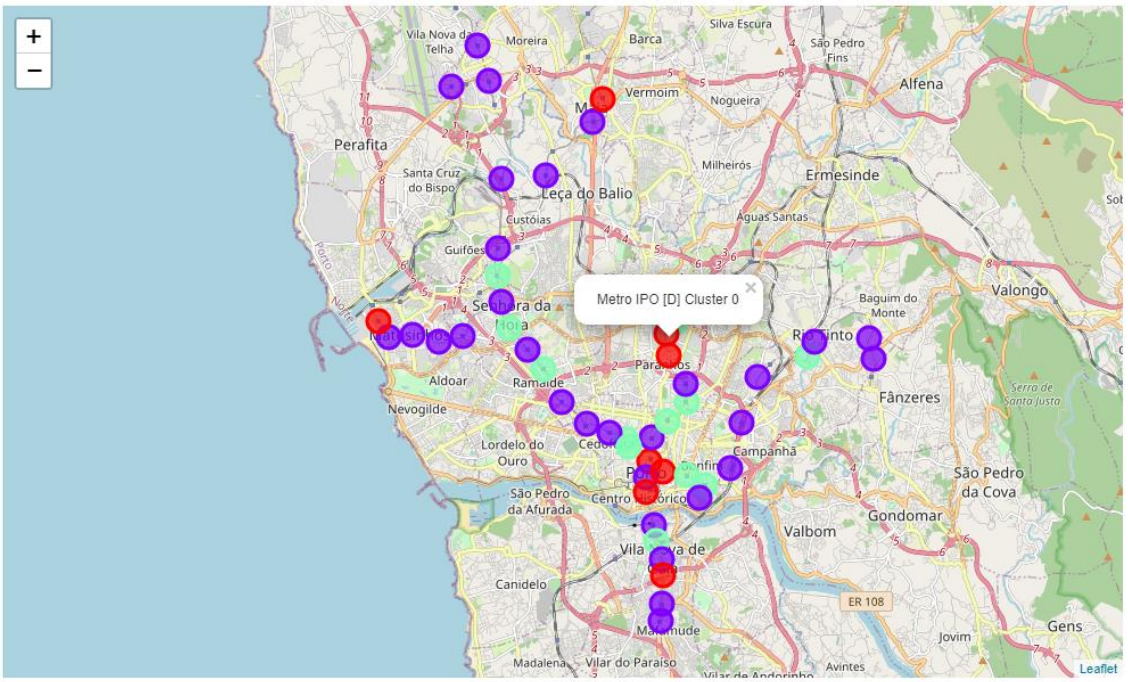
These were the preliminary results with different number of clusters:

- 2 clusters show the main tourist and nightlife/leisure spots
- 3 clusters allow to distinguish some professional oriented areas
- 4 and more clusters are difficult to interpret

For the final analysis we settled on 3 clusters (0 to 2).



We can then use the Folium map to easily visualize each station type, after association with previous tables with location.



Results

Here is how we can characterize the clusters by looking at venue scores

- Cluster 0 (Red) has consistently high scores for all venue categories. This is the most diversely developed part of the city, with lots of restaurants and night life
- Cluster 1 (Purple) has balanced marks between residential, professional, food and services, being the most common type of area.
- Cluster 2 (Green) is similar to Cluster 1, but with more food places and professional venues.

Plotting the clusters on a map shows us that:

- Cluster 0 places are the most high-density spots, where people tend to converge for all activities, but, in particular night life;
- Cluster 1 is scattered all over the city. As Porto doesn't have very accentuated business or residential areas, we see the mix described by this cluster to be the most common;

- Cluster 2 show places more focused on professional and food places all over the city, being areas a bit more specialized on workplaces than cluster 2.

Discussion

Foursquare data isn't all-encompassing. The highest number of venues are in the Food and Shop & Service categories. Data doesn't consider a venue's size (e.g. a university building attracts a lot more people that a hot dog stand – each of them is still one Foursquare "venue").

Conclusion

Foursquare data is limited but can provide insights into a city's characteristics. This data could be combined with other sources (e.g. city data on number of residents) to provide more accurate results and other insights. Also, comparative analysis can be done with other cities.