

Clustering Lisbon neighbourhoods

Marta da Silva Luis

December 2019

1 Introduction

The objective of this report is to describe the capstone project named “Clustering Lisbon neighbourhoods” developed in Jupyter Notebooks.

The aim of this project is to recommend which neighbourhood of Lisbon is the best location to open a new office, analysing location data – to explore and compare neighborhoods, using the following tools:

- Foursquare API: to explore the neighbourhoods in Lisbon;
- K-means clustering algorithm: to cluster the neighbourhoods;
- Folium library: to visualize the neighbourhoods and clusters.

2 Data

Lisbon is divided in 24 neighbourhoods and for this project we use the list of neighbourhoods (‘freguesias’ in Portuguese) from Wikipedia (https://pt.wikipedia.org/wiki/Lista_de_freguesias_de_Lisboa). In order to use the Foursquare location data, we need the latitude and the longitude of each neighbourhood and this data is also available at Wikipedia.

The latitude and longitude data are an input to the Foursquare API in order to explore the venues of each neighbourhood.

2.1 Required libraries

In order to analyze the data, it’s necessary to import the following Python libraries:

```
[1] import numpy as np
import pandas as pd
from geopy.geocoders import Nominatim # convert an address into latitude and longitude values
import folium # map rendering library
import json # library to handle JSON files
import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas dataframe
import matplotlib.cm as cm # Matplotlib and associated plotting modules
import matplotlib.colors as colors
from sklearn.cluster import KMeans # import k-means from clustering stage
import folium # map rendering library
```

Figure 1 – Python libraries

2.2 Data scrapped from Wikipedia

Data scrapped from Wikipedia were transformed into a pandas dataframe.

```
[3] import wikipedia as wp
wp.set_lang("pt")
html = wp.page("Lista_de_freguesias_de_Lisboa").html().encode("UTF-8")
df = pd.read_html(html)[0]
df.to_csv('lisboa_data.csv', header=1, index=False)

[4] # transform the data into a pandas dataframe

df = pd.read_csv('lisboa_data.csv')
df.head(3)
```

Figure 2 – Transform the Wikipedia data into a pandas dataframe

The cleaning operations of the dataframe were:

- remove the useless columns like old name of the current neighbourhoods;
- remove duplicates;
- new column names in English (Wikipedia table is in Portuguese).

2.3 Latitude and Longitude from a csv file

The latitude and longitude were also scrapped from Wikipedia but in this case no table was available, so we need to create one and upload it to Github.

```
[9] url = 'https://raw.githubusercontent.com/mdsl22/Coursera_Capstone/master/Geo_Lisbon.csv'

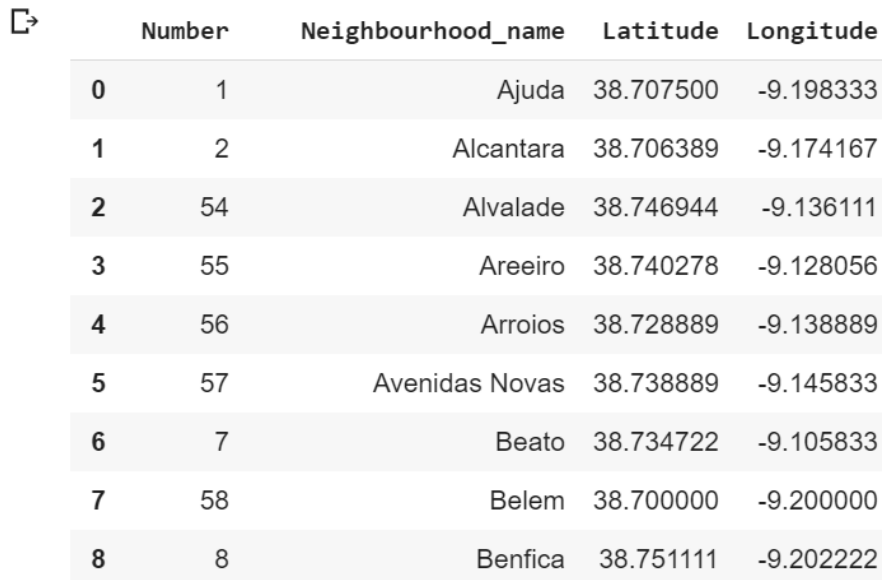
[10] df_geo = pd.read_csv(url)
df_geo.head(3)
```

	Number	Latitude	Longitude	Neighbourhood_name
0	1	38.707500	-9.198333	Ajudá
1	2	38.706389	-9.174167	Alcantara
2	54	38.746944	-9.136111	Alvalade

Figure 3 – Latitude and Longitude data

In the last step we need to find the coordinates of each neighbourhood and assign it to the dataframe. The result is the dataframe named `df_lisbon` that we will use to explore Foursquare venues.

```
[14] df_lisbon=df_geo[['Number','Neighbourhood_name','Latitude','Longitude']]
df_lisbon
```



	Number	Neighbourhood_name	Latitude	Longitude
0	1	Ajuda	38.707500	-9.198333
1	2	Alcantara	38.706389	-9.174167
2	54	Alvalade	38.746944	-9.136111
3	55	Areeiro	38.740278	-9.128056
4	56	Arroios	38.728889	-9.138889
5	57	Avenidas Novas	38.738889	-9.145833
6	7	Beato	38.734722	-9.105833
7	58	Belem	38.700000	-9.200000
8	8	Benfica	38.751111	-9.202222

Figure 4 – Dataframe df_lisbon

3 Methodology

3.1 Exploratory data analysis: Map of Lisbon

In order to visualize the data, we need the geographical coordinates of Lisbon which are 38.7077507 and -9.1365919.

```
[16] # Geographical coordinates of Lisbon

address = 'Lisbon'

geolocator = Nominatim(user_agent="coursera")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The georgapical coordinates of Lisbon are {}, {}'.format(latitude, longitude))
```

Figure 5 – Lisbon coordinates

To visualize the 24 neighbourhoods in Lisbon we create a map using folium library as follows:

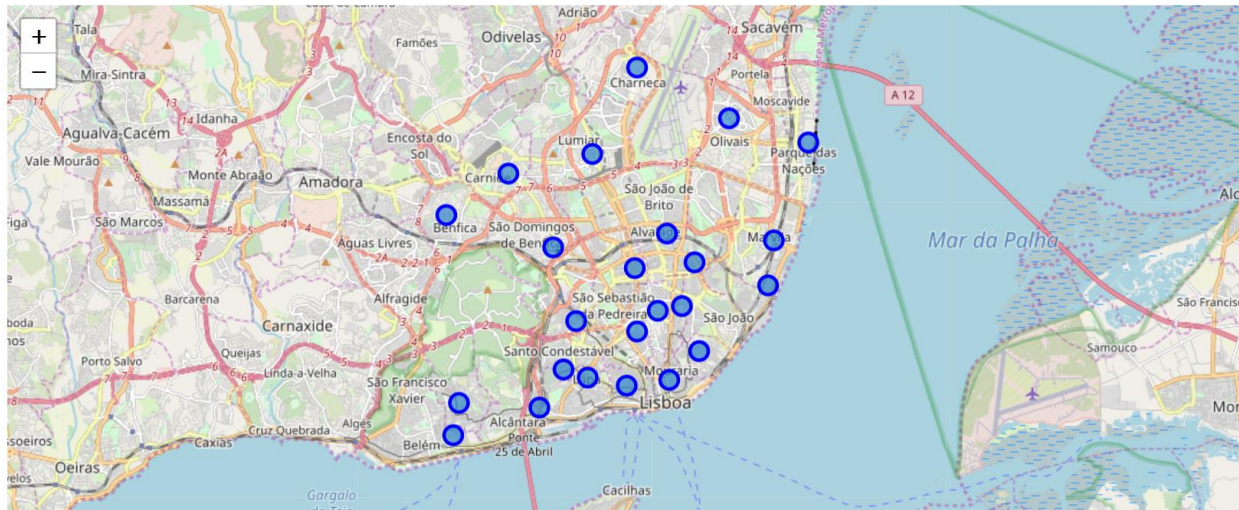


Figure 6 – Map of Lisbon with neighbourhoods

3.2 Foursquare

Communicating with the Foursquare database is really very simple, all thanks to their API. We create a uniform resource identifier (URI) and append it with extra parameters depending on the data that we are taking from the database. Any call request is composed of:

- base URI: `api.foursquare.com/v2`;
- request data: venues, users, or tips;
- developer account credentials: Client ID and Client Secret as well the version of the API, which is simply a date.

We make the call to the database and in return we get a JSON file of the venues that match our query. For each venue, we get mostly its name, unique ID, location, and category.

3.3 Explore the venues of Benfica (Number 8 of the dataframe)

To proceed with the exploratory data analysis, we explore Benfica neighbourhood (within 500 m radius) with Foursquare API. The call returned 46 venues (dataframe `nearby_venues`). Each venue has a name, category, latitude and longitude as shown in the Figure 7.

	name	categories	lat	lng
0	A Padaria Portuguesa	Bakery	38.750727	-9.201863
1	Teatro Turim	Theater	38.750702	-9.202288
2	Mata de Benfica - Parque Silva Porto	Park	38.749030	-9.204727
3	Com Calma	Coffee Shop	38.751835	-9.199237
4	A Travessa do Rio	Portuguese Restaurant	38.750511	-9.198650

Figure 7 – 5 Venues of Benfica

3.4 Analyse all neighbourhoods

As we did for Benfica, we need to do the same for all neighbourhoods in order to map the venues of each neighbourhood based in the defined parameters.

We run a function that repeat the same process to all neighbourhoods:

1. Create the API request URL;
2. Make the GET request;
3. Return only relevant information for each nearby venue: name, location and category;
4. Run the function on each neighbourhood and create a new dataframe called `venues`.

```
[74] print(venues.shape)
venues.head()
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ajuda	38.7075	-9.198333	Palácio Nacional da Ajuda	38.707653	-9.197758	Historic Site
1	Ajuda	38.7075	-9.198333	Restaurante Andorinhas	38.704911	-9.199349	Restaurant
2	Ajuda	38.7075	-9.198333	Páteo Alfacinha	38.706537	-9.194202	Restaurant
3	Ajuda	38.7075	-9.198333	Jardim Botânico da Ajuda	38.706430	-9.201222	Botanical Garden
4	Ajuda	38.7075	-9.198333	Churrasqueira do Marquês	38.703996	-9.199402	BBQ Joint

Figure 8 – Venues dataframe

Using pandas groupby method, we group by neighbourhood name and count the number of venues using counts function.

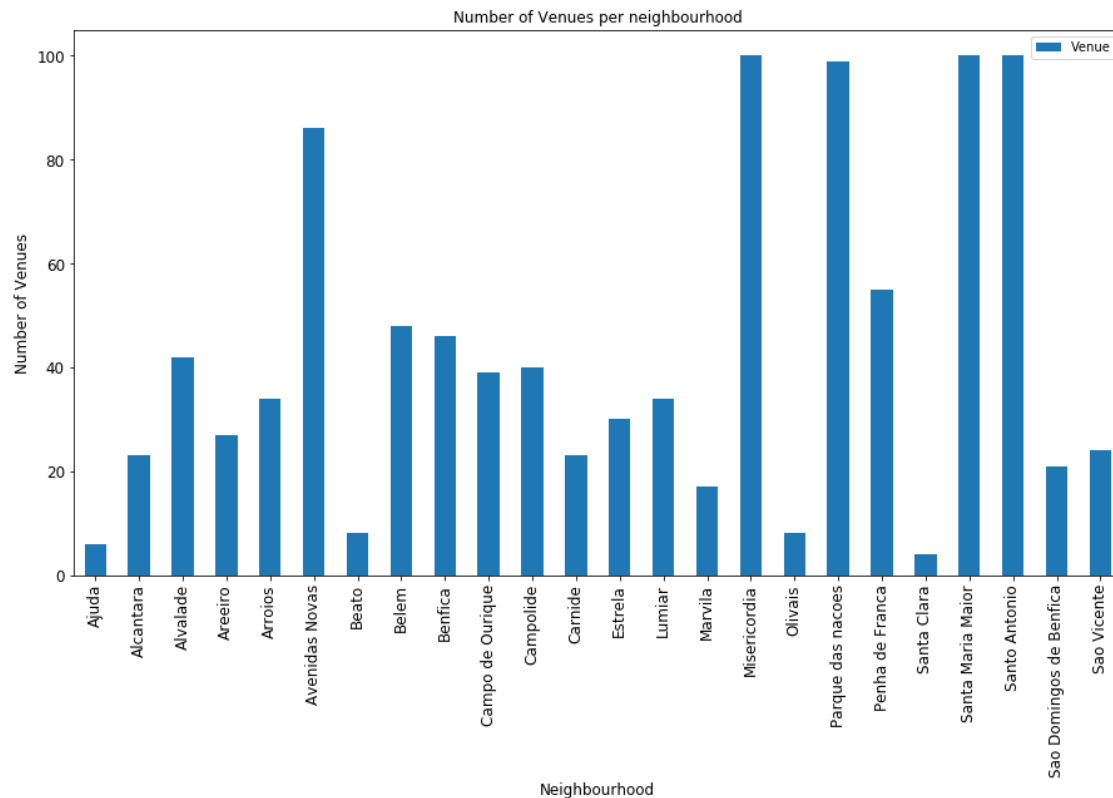


Figure 9 – Number of Venues per neighbourhood

It's clear from the previous graph that the neighbourhoods with greater number of venues are: Avenidas Novas, Misericórdia, Parque das Nações, Santa Maria Maior and Santo António.

There are 1022 venues and 169 unique categories of venues.

The next step is to perform one hot encoding which is a process by which categorical variables are converted into a form that could be provided to Machine Learning algorithms like k-means clustering to do a better job in prediction.

The process is:

1. One hot encoding: column 'Venue Category'
2. Dataframe `onehot`: 1022 rows × 169 columns
3. Group by neighbourhood `onehot` dataframe
4. Dataframe `grouped`: 24 rows × 169 columns

We proceed to create a function that calculates the 10 top venues for each neighbourhood and create the dataframe `neighborhoods_venues_sorted`.

	Neighborhood	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	Ajuda	Restaurant	Historic Site	Botanical Garden	BBQ Joint	Exhibit	Food	Flower Shop	Flea Market	Fish & Chips Shop	Fast Food Restaurant
1	Alcantara	Portuguese Restaurant	Mediterranean Restaurant	Plaza	Beer Bar	Pizza Place	Coffee Shop	Park	Eastern European Restaurant	Nightclub	Restaurant
2	Alvalade	Portuguese Restaurant	Café	BBQ Joint	Bookstore	Ice Cream Shop	Electronics Store	Bakery	Hotel	Coffee Shop	Restaurant
3	Areiro	Portuguese Restaurant	Hotel	Electronics Store	Bakery	Café	Fountain	Restaurant	Chinese Restaurant	Sports Club	Stadium

Figure 10 – Topmost common venue for each neighbourhood

4 Cluster neighbourhoods

K-means clustering, which is one of the simplest unsupervised learning algorithms, will allow us to find the best location to open a new office in Lisbon considering the top common venues of each neighbourhood. It is a partitioning clustering that divides the data into non overlapping subsets (clusters), the algorithm minimizes the intra-cluster distances and maximizes the inter-cluster distances. The steps are the following:

1. Set number of clusters: 5
2. Run k-means clustering
3. Check cluster labels generated for each row in the dataframe
4. Add clustering labels
5. Merge `neighborhoods_venues_sorted` with `df_lisbon` to add latitude/longitude for each neighborhood
6. Create a map with folium
7. Set color scheme for the clusters
8. Add markers to the map

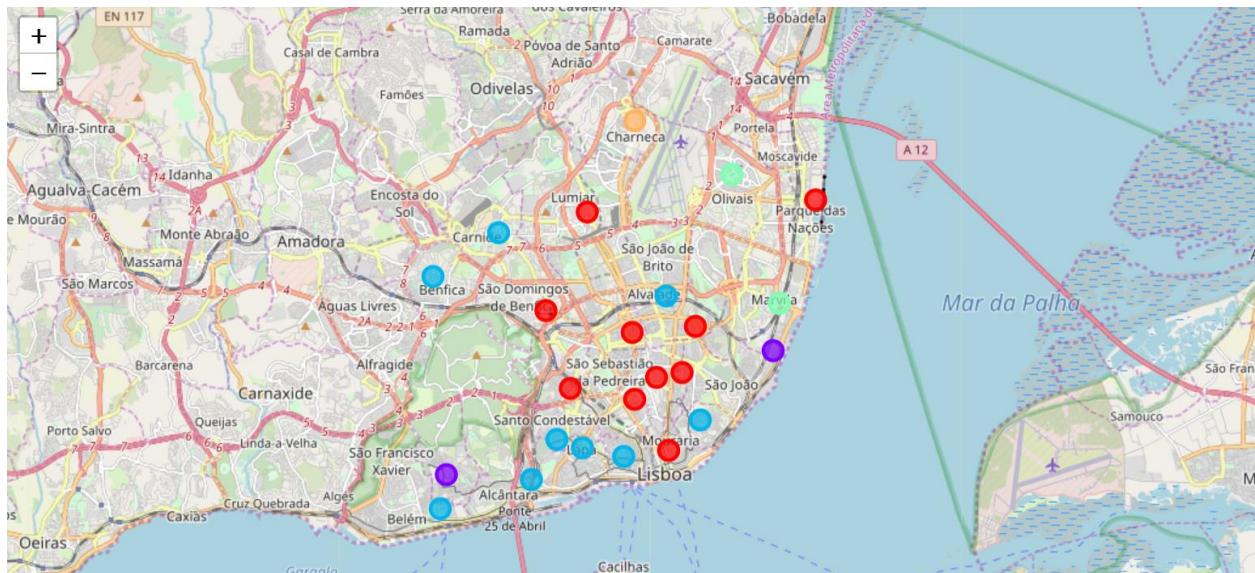


Figure 11 – Clusters map

5 Examine Clusters

5.1 Cluster 0

For Cluster 0, the top 10 venues of each neighbourhood are:

Neighbourhood_name	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
55	Areeiro	Hotel	Portuguese Restaurant	Electronics Store	Coffee Shop	Stadium	Italian Restaurant	Fountain	Flower Shop	Pizza Place
56	Arroios	Hotel	Portuguese Restaurant	Indian Restaurant	Hostel	Vegetarian / Vegan Restaurant	Bakery	Pawn Shop	Chinese Restaurant	Restaurant
57	Avenidas Novas	Portuguese Restaurant	Hotel	Restaurant	Italian Restaurant	Bakery	Café	Gym / Fitness Center	Vegetarian / Vegan Restaurant	Sushi Restaurant
10	Campolide	Restaurant	Hotel	Portuguese Restaurant	Bakery	Seafood Restaurant	Electronics Store	Scenic Lookout	Chocolate Shop	Clothing Store
18	Lumiar	Café	Pizza Place	Gym / Fitness Center	Japanese Restaurant	Fast Food Restaurant	Soccer Stadium	Supermarket	Chinese Restaurant	Restaurant
62	Parque das nacoes	Portuguese Restaurant	Restaurant	Sushi Restaurant	Ice Cream Shop	Burger Joint	Coffee Shop	Electronics Store	Chinese Restaurant	Hotel
63	Penha de Franca	Portuguese Restaurant	Hotel	Café	Hostel	Supermarket	Bakery	Chinese Restaurant	Scenic Lookout	Indian Restaurant
65	Santa Maria Maior	Portuguese Restaurant	Hotel	Café	Restaurant	Wine Bar	Indian Restaurant	Ice Cream Shop	Coffee Shop	Scenic Lookout
66	Santo Antonio	Hotel	Portuguese Restaurant	Café	Hostel	Restaurant	Bakery	Pizza Place	Clothing Store	Supermarket
39	Sao Domingos de Benfica	Japanese Restaurant	Theme Park	BBQ Joint	Zoo	Office	Metro Station	Mediterranean Restaurant	Café	Food Truck

Figure 12 – Cluster 0

5.2 Cluster 1

For Cluster 1, the top 10 venues of each neighbourhood are:

Neighbourhood_name	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
1	Ajuda	Restaurant	Historic Site	Botanical Garden	BBQ Joint	Exhibit	Food	Flower Shop	Flea Market	Fish & Chips Shop
7	Beato	Restaurant	Historic Site	Tapas Restaurant	Cantonese Restaurant	Brewery	Gym / Fitness Center	Event Space	Flea Market	Fish & Chips Shop

Figure 13 – Cluster 1

5.3 Cluster 2

For Cluster 2, the top 10 venues of each neighbourhood are:

	Neighbourhood_name	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
Number											
2	Alcantara	Portuguese Restaurant	Mediterranean Restaurant	Indian Restaurant	Museum	Restaurant	Nightclub	Supermarket	Sushi Restaurant	Eastern European Restaurant	Beer Bar
54	Alvalade	Portuguese Restaurant	Café	Bookstore	Bakery	Hotel	Ice Cream Shop	Coffee Shop	BBQ Joint	Electronics Store	Pizza Place
58	Belem	Portuguese Restaurant	Garden	Bakery	BBQ Joint	Café	History Museum	Sandwich Place	Restaurant	Pizza Place	Plaza
8	Benfica	Café	Portuguese Restaurant	Seafood Restaurant	Coffee Shop	Bakery	Breakfast Spot	Pharmacy	Restaurant	Park	Soccer Field
59	Campo de Ourique	Portuguese Restaurant	Restaurant	Coffee Shop	Bar	Bakery	Seafood Restaurant	Furniture / Home Store	Burger Joint	Café	Sandwich Place
11	Camide	Portuguese Restaurant	Restaurant	Sushi Restaurant	Café	Bakery	Tapas Restaurant	Theater	Burger Joint	Mediterranean Restaurant	Garden
60	Estrela	Portuguese Restaurant	Café	Grocery Store	Steakhouse	Hostel	Supermarket	Ice Cream Shop	Garden	Japanese Restaurant	Miscellaneous Shop
61	Misericordia	Portuguese Restaurant	Bar	Café	Wine Bar	Hostel	Cocktail Bar	Restaurant	Breakfast Spot	Bed & Breakfast	Hotel
67	Sao Vicente	Portuguese Restaurant	Bakery	Café	Mediterranean Restaurant	Plaza	Bistro	Other Nightlife	Coffee Shop	Event Space	Flea Market

Figure 14 – Cluster 2

5.4 Cluster 3

For Cluster 3, the top 10 venues of each neighbourhood are:

	Neighbourhood_name	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
Number											
21	Marvila	Restaurant	Portuguese Restaurant	Art Gallery	Mediterranean Restaurant	Pizza Place	Argentinian Restaurant	Café	Train Station	Thrift / Vintage Store	Motorcycle Shop
33	Olivais	Restaurant	Hostel	Café	Metro Station	BBQ Joint	Furniture / Home Store	Chinese Restaurant	Falafel Restaurant	Flower Shop	Flea Market

Figure 15 – Cluster 3

5.5 Cluster 4

For Cluster 4, the top 10 venues of each neighbourhood are:

	Neighbourhood_name	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
Number											
64	Santa Clara	Portuguese Restaurant	Brewery	Gas Station	Exhibit	Food	Flower Shop	Flea Market	Fish & Chips Shop	Fast Food Restaurant	Farmers Market

Figure 16 – Cluster 4

6 Discussion and Conclusion

The following table summarizes the k-means clustering results:

Clusters	Number of neighbourhoods	Neighbourhoods Names	Color
Cluster 0	10	Areeiro, Arroios, Avenidas Novas, Campolide, Lumiar, Parque das nacoes, Penha de Franca, Santa Maria Maior, Santo Antonio, Sao Domingos de Benfica	Red
Cluster 1	2	Ajuda, Beato	Purple
Cluster 2	9	Alcantara, Alvalade, Belem, Benfica, Campo de Ourique, Carnide, Estrela, Misericordia, Sao Vicente	Blue
Cluster 3	2	Marvila, Olivais	Green
Cluster 4	1	Santa Clara	Yellow

Figure 17 – Clusters Summary table

After analysing the above 5 clusters we can recommend that the neighbourhoods of Cluster 0 are the best to open a new office. The 5th most common venue in the neighbourhood Sao Domingos de Benfica is the category office, so I would say this neighbourhood is a good location for a new office. Parque das Nações, Santa Maria Maior and Santo António from Cluster 0 are the ones with greater number of venues which means they are good locations as well.

The decision of which location is the best to open a new office in Lisbon depends on many factors but with this project we want to focus on the advantages of using location data from Foursquare and Machine Learning algorithms to solve problems.