Where to establish a Start-up Company

Applied Data Science Capstone by IBM/Coursera – “The Battle of Neighbourhoods”

Author: Henrique M. L. Pereira

Problem → Finding a suitable Neighbourhood

- Start-up concerns & needs:
  - Affordability of office and living;
  - Multicultural environment;
  - Near essential commodities;
  - Near places to relax and meet people
  - In a buzzing city but without stressing situations;

- Country: Portugal
- Cities to choose from:
  - Lisbon
  - Porto
  - Coimbra
  - Braga
  - Aveiro

Why Portugal?

- Web Summit conference
- Start-up community is thriving
- Healthy support from the Portuguese government
- Rise of desirability of Portugal as a good tech hub
- Quality and low-cost services

https://www.forbes.com/sites/heatherfarmbrough/2018/02/28/all-roads-lead-to-lisbon-why-startups-are-booming-in-the-portuguese-capital/#5399ed1177ea

https://www.entrepreneur.com/article/307526

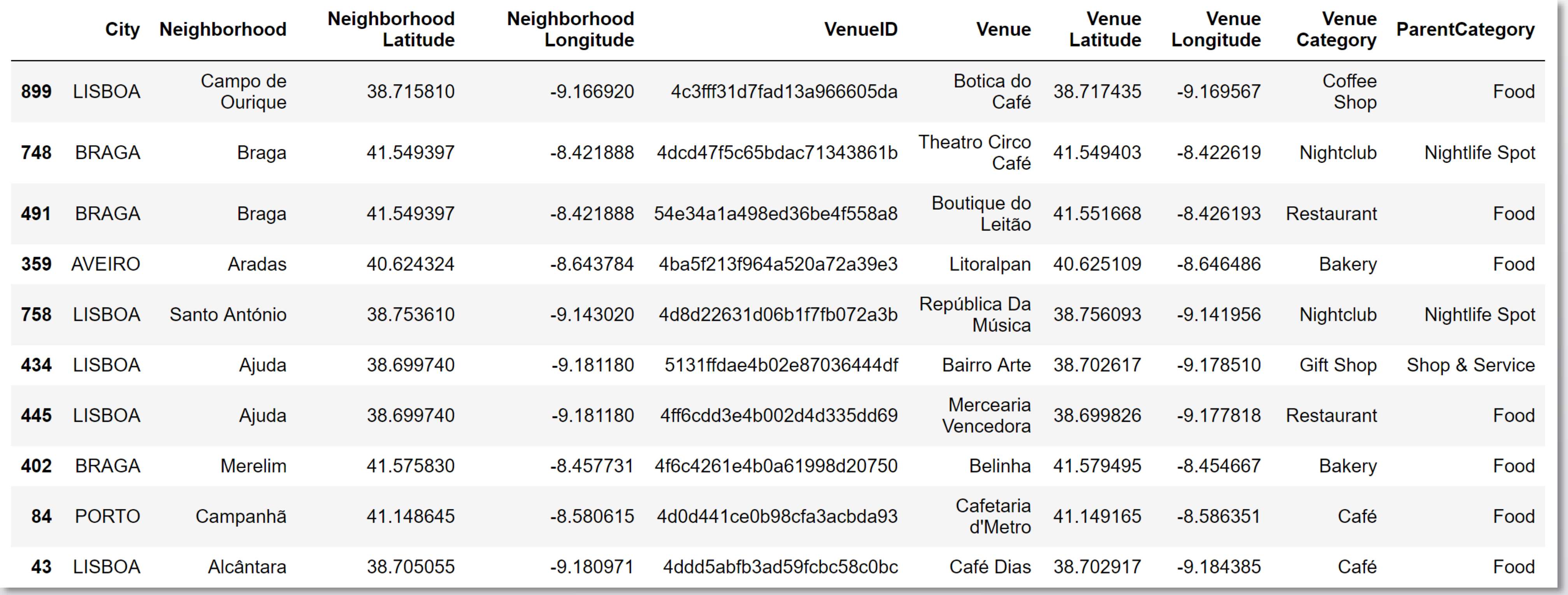
Data retrieval and Cleaning

- General Features from each city:
  - Web Scrapping of Nomad List (https://nomadlist.com/)
  - To rank final candidates
- Portuguese cities neighbourhood’s and location data
  - Portuguese Government open source data
  - *Geocoder* Package

- Foursquare Explore API
  - Center on each Neighbourhood
  - Radius = 500 meters
  - Features collected:
    - Venue Category
    - Venue Parent Category

Data retrieval and Cleaning

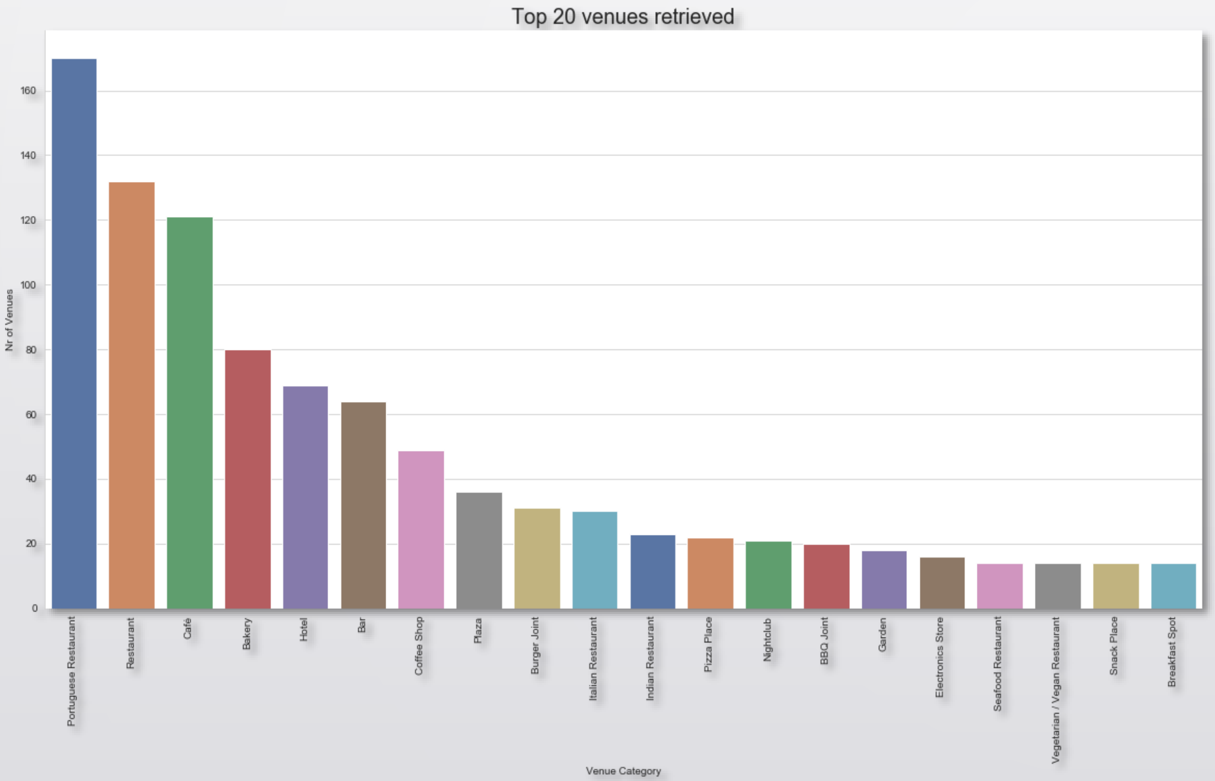
- Merge data into single Dataframe (sample below):

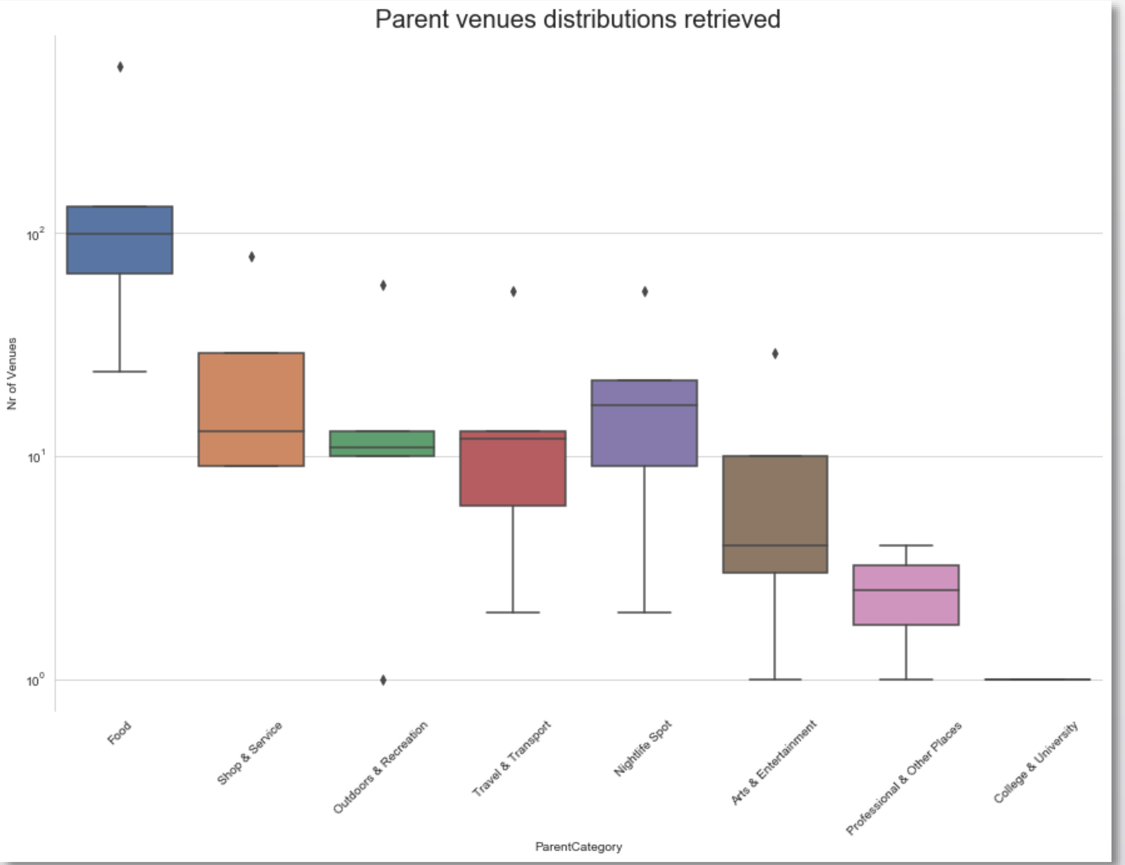


	City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	VenueID	Venue	Venue Latitude	Venue Longitude	Venue Category	ParentCategory
899	LISBOA	Campo de Ourique	38.715810	-9.166920	4c3fff31d7fad13a966605da	Botica do Café	38.717435	-9.169567	Coffee Shop	Food
748	BRAGA	Braga	41.549397	-8.421888	4dcd47f5c65bdac71343861b	Theatro Circo Café	41.549403	-8.422619	Nightclub	Nightlife Spot
491	BRAGA	Braga	41.549397	-8.421888	54e34a1a498ed36be4f558a8	Boutique do Leitão	41.551668	-8.426193	Restaurant	Food
359	AVEIRO	Aradas	40.624324	-8.643784	4ba5f213f964a520a72a39e3	Litoralpan	40.625109	-8.646486	Bakery	Food
758	LISBOA	Santo António	38.753610	-9.143020	4d8d22631d06b1f7fb072a3b	República Da Música	38.756093	-9.141956	Nightclub	Nightlife Spot
434	LISBOA	Ajuda	38.699740	-9.181180	5131ffdae4b02e87036444df	Bairro Arte	38.702617	-9.178510	Gift Shop	Shop & Service
445	LISBOA	Ajuda	38.699740	-9.181180	4ff6cdd3e4b002d4d335dd69	Mercearia Vencedora	38.699826	-9.177818	Restaurant	Food
402	BRAGA	Merelim	41.575830	-8.457731	4f6c4261e4b0a61998d20750	Belinha	41.579495	-8.454667	Bakery	Food
84	PORTO	Campanhã	41.148645	-8.580615	4d0d441ce0b98cfa3acbda93	Cafetaria d'Metro	41.149165	-8.586351	Café	Food
43	LISBOA	Alcântara	38.705055	-9.180971	4ddd5abfb3ad59fcbc58c0bc	Café Dias	38.702917	-9.184385	Café	Food

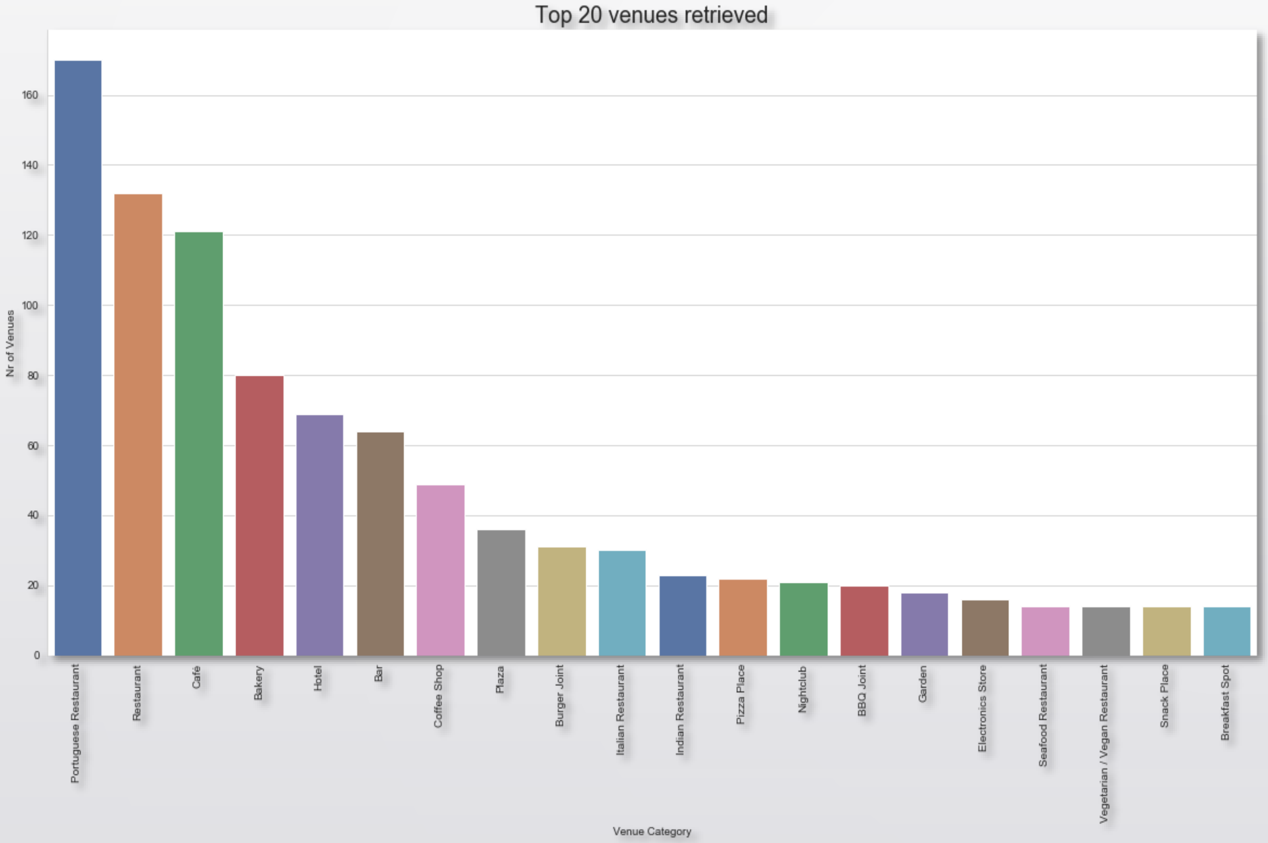
Exploratory analysis

- I found in the dataframe 139 unique Venue categories and a total of 8 unique Parent Venue categories…





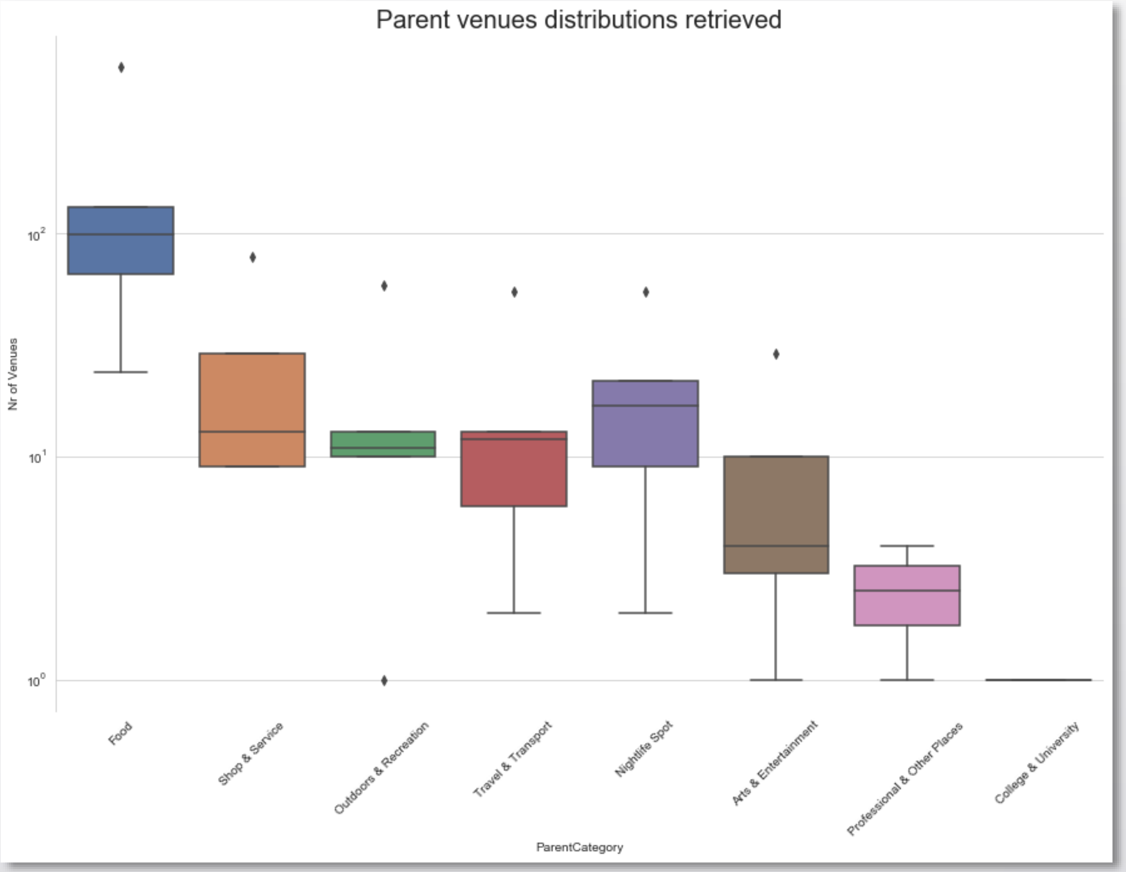
Exploratory analysis



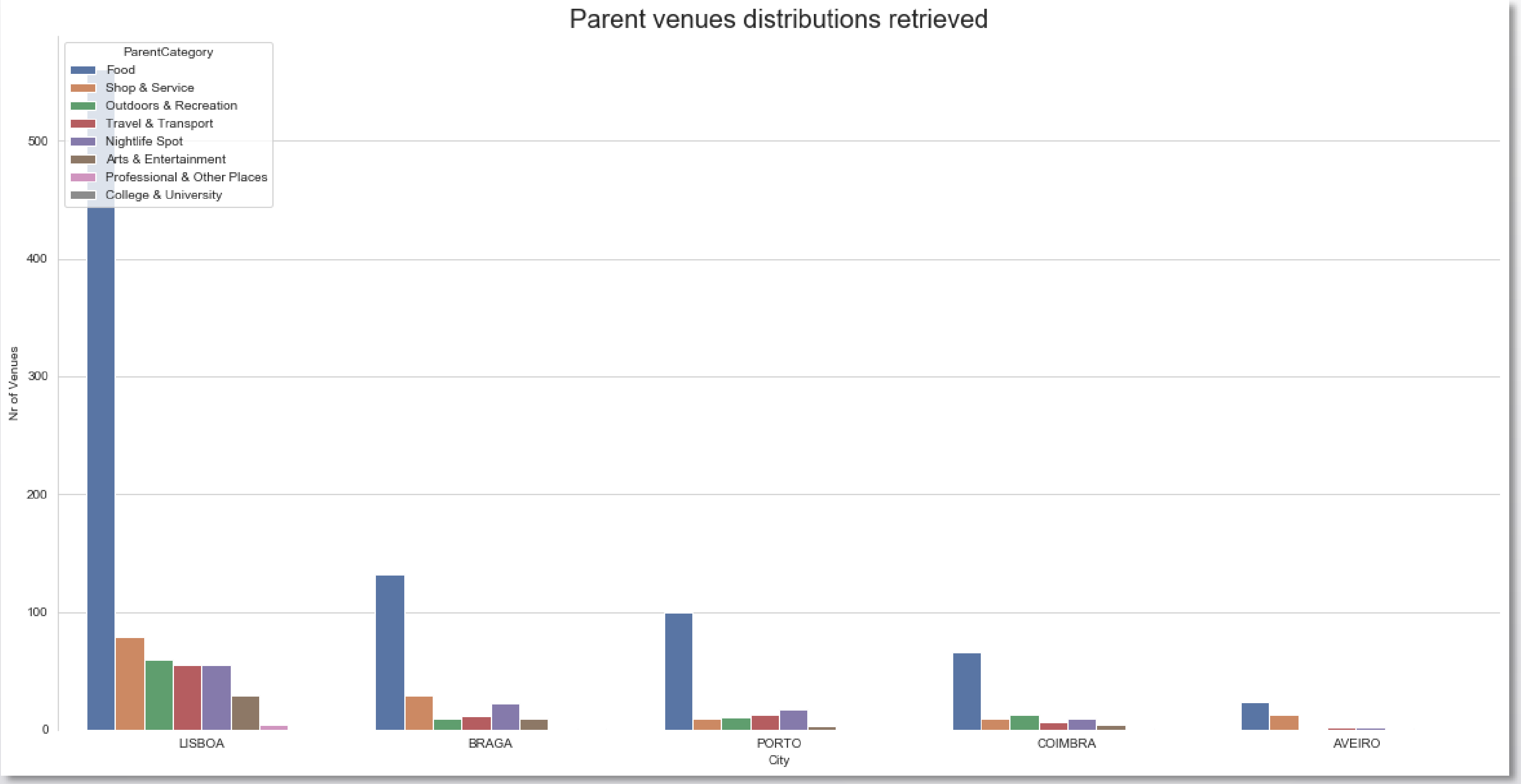
- 139 unique Venue categories mostly food related…
- Too many features to be selected…
  - ‘Curse of dimensionality’ when using unsupervised learning…

Exploratory analysis

- 8 unique Parent Venue categories
  - Loss of information… but,
  - Less features,
  - Best for next steps:
    - K-Means
    - OPTICS

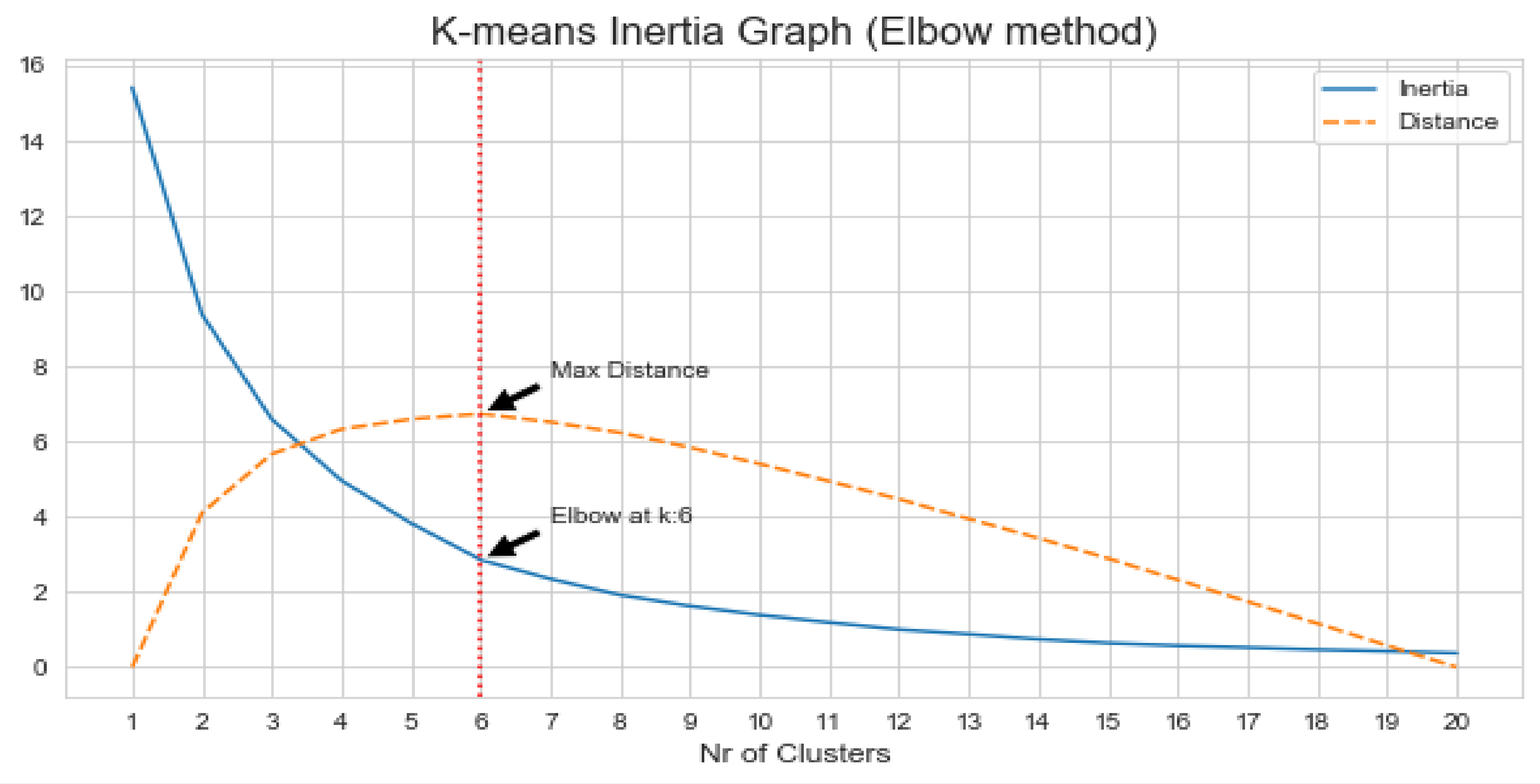

Parent venues distributions retrieved

Exploratory analysis



- Cities are different…
- Proportion of non food related venues differ form city to city.

K-Means

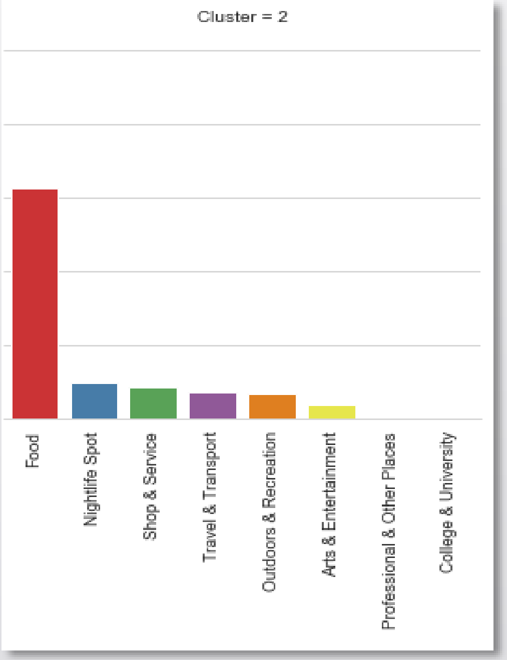


- Best nr of Clusters is 6
- Elbow not very clear…

K-Means

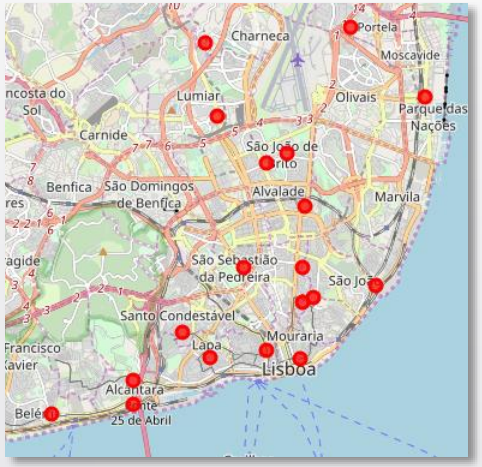
- Selected only the neighbourhoods that had values for the following important categories:
  - 'Arts & Entertainment',
  - 'Outdoors & Recreation',
  - 'Professional & Other Places',
  - 'Shop & Service',
  - 'Travel & Transport’

- Best cluster to these conditions:

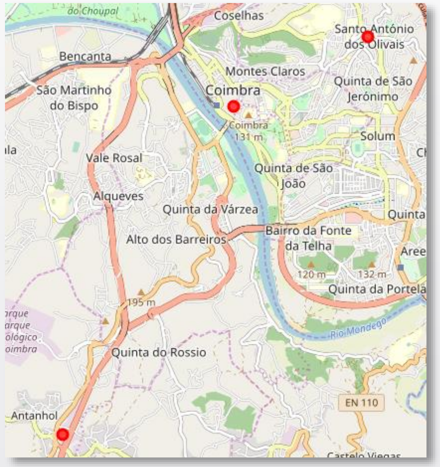


K-Means → Cities with Cluster nr 2

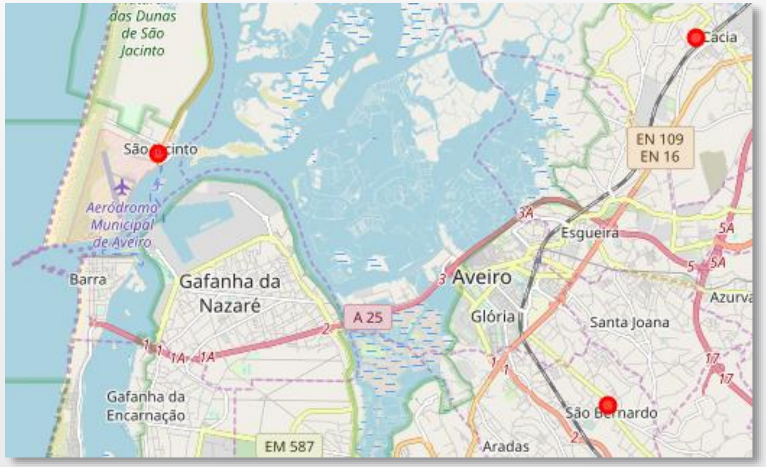
Lisbon



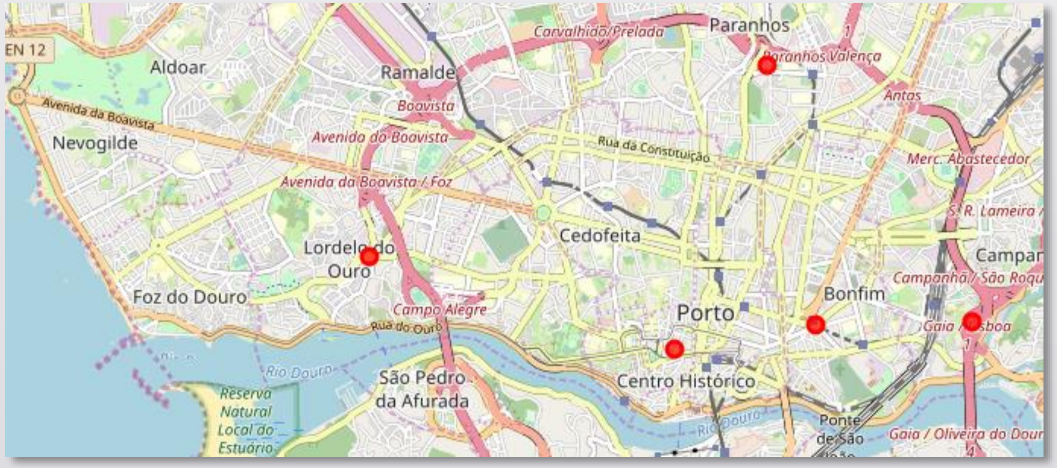
Coimbra



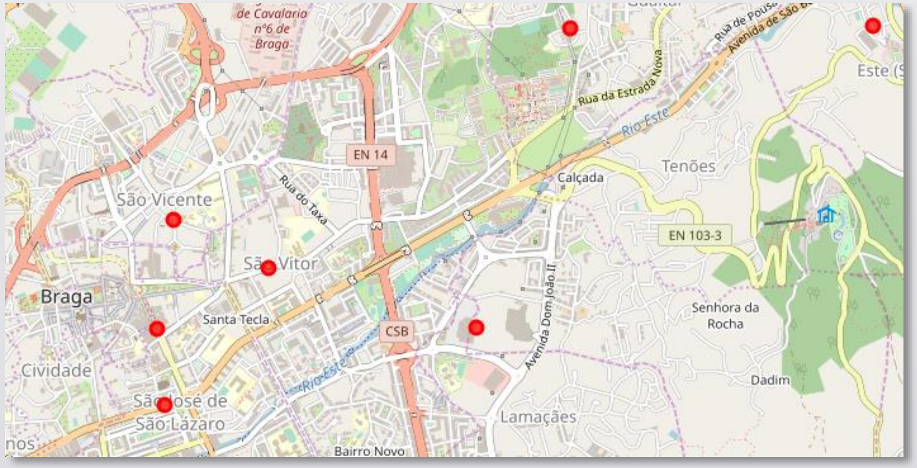
Aveiro



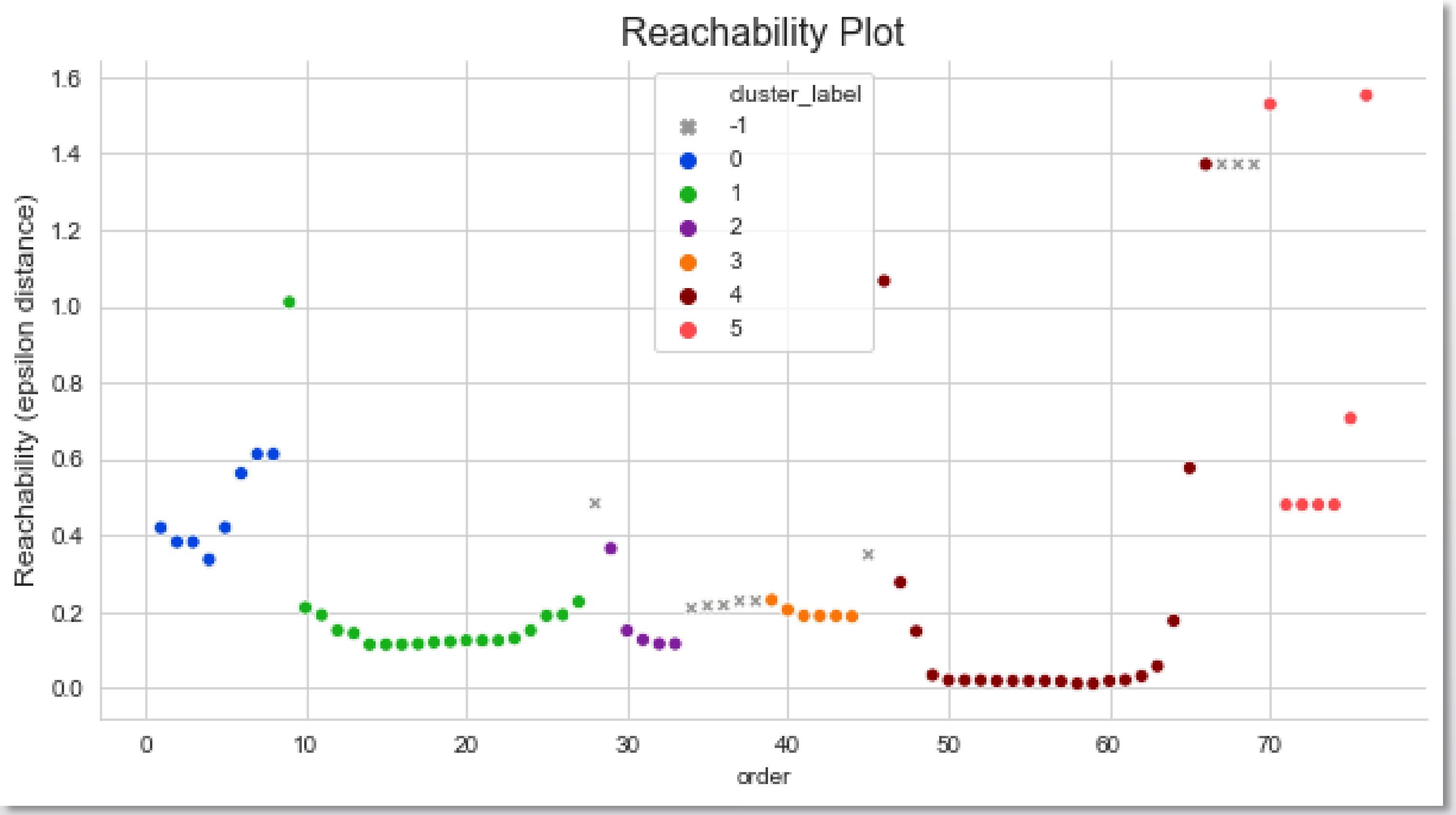
Porto



Braga



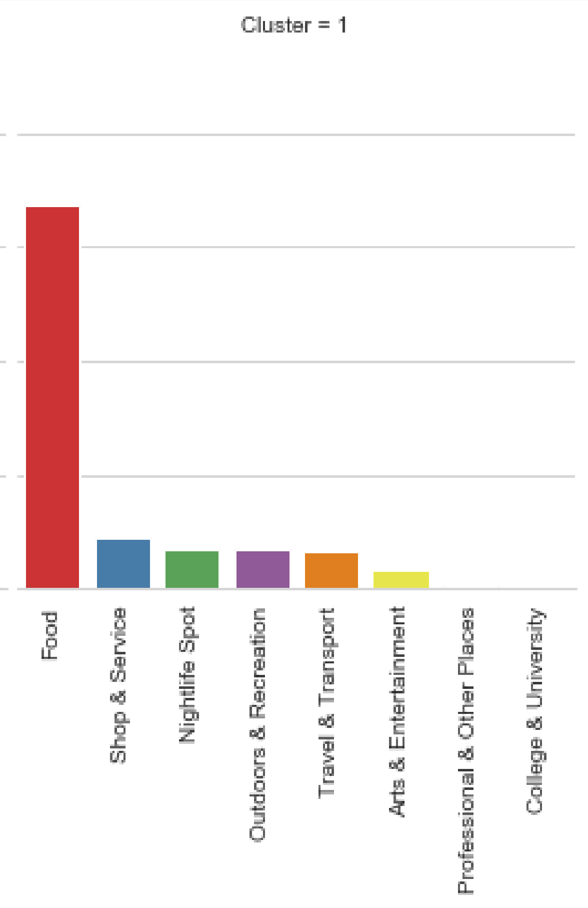
OPTICS

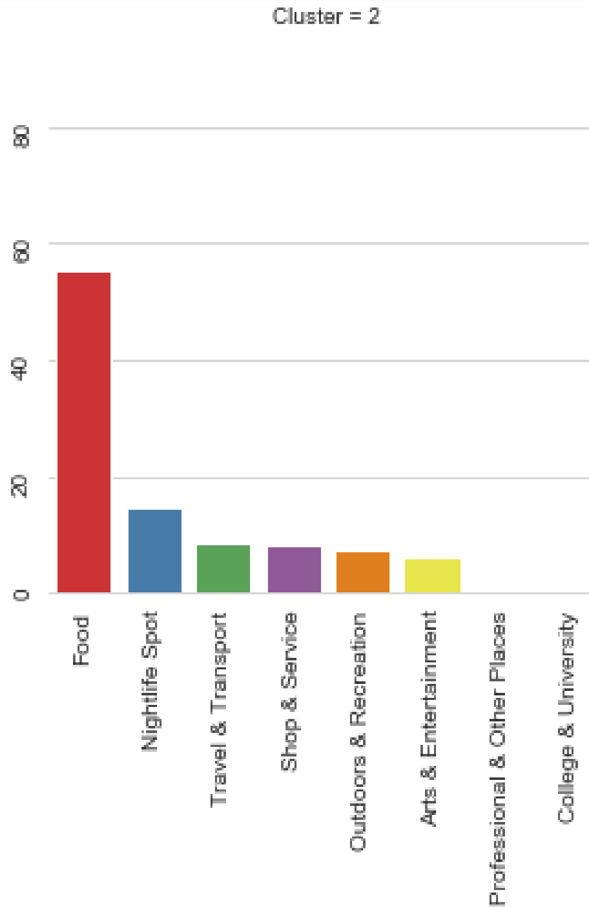


- Identification of:
  - Cluster -1: Outliers
  - 6 Clusters of interest
- Further important categories to be selected:
  - 'Arts & Entertainment',
  - 'Outdoors & Recreation',
  - 'Professional & Other Places',
  - 'Shop & Service',
  - 'Travel & Transport'

OPTICS

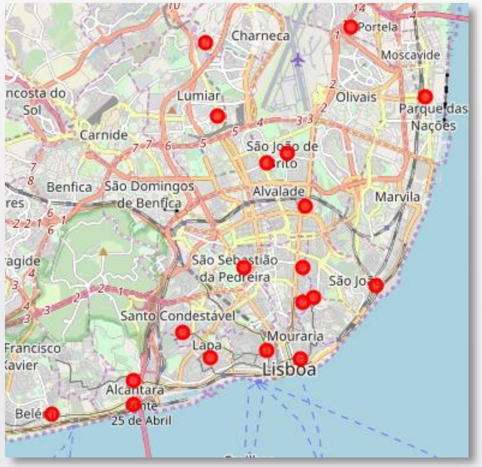
- Further important categories **selected**:
    - 'Arts & Entertainment',
    - 'Outdoors & Recreation',
    - 'Professional & Other Places',
    - 'Shop & Service',
    - 'Travel & Transport'
- **Applying Data** from Nomad List
    - Cities **Ranked**
- Identification of:
    - Cluster 1
    - Cluster 2



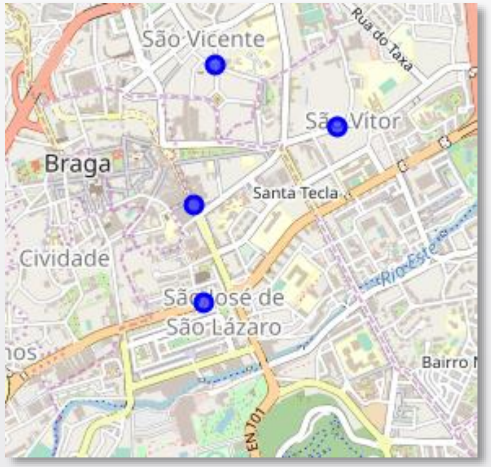


OPTICS → Cities with Cluster nr 1 and 2

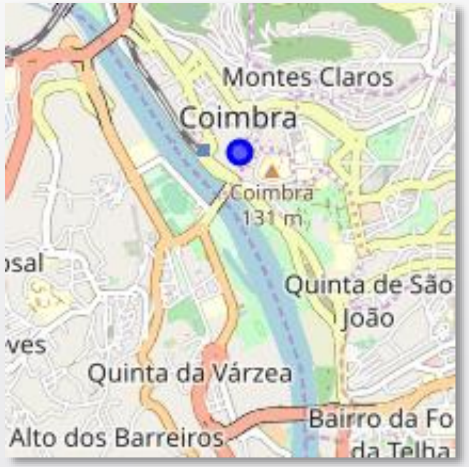
Lisbon



Braga



Coimbra



	Lisbon	Braga	Coimbra
Cluster	***1***	**2**	
Rank	**1st**	**2nd**	**3rd**

When comparing both clusters it is obvious that the cities of Coimbra and Braga have higher number of Nightlife Spots comparing to Lisbon.

Conclusion

- Objective achieved: List of good candidate Neighbourhoods to establish the Start-up.
  - Using OPTICS and with ranking based on Nomad List Features
- Smaller cities might have less venues reported to Foursquare
  - Skew data in favour of the main city (the capital of Portugal - Lisbon)
- Lack of more information at the neighbourhood level
  - the real population for each neighbourhood, the price of housing and commerce per square meter, the overall condition of the neighbourhood in various dimensions, etc

Future

- The data retrieved in this project with the previous additional data could be presented, by UI:
    1. As a list of variables from which the Client could chose the most important ones
    2. Feed these features to our OPTICS model,
    3. Ranking by Nomad List and, finally,
    4. Presentation of a more refined and shorter list of the best neighbourhood to settle the Start-up