Culinary Clustering

Tobias Splith

Coursera Capstone Project

10.09.2020

Table of Contents

Links: Data and Introduction-Notebook, Data Analysis-Notebook

Introduction

- venues form spatial clusters in cities:
  - restaurants and bars often in downtown- or in university-area
  - museums and parks in city center
  - commodity stores in malls
  - etc.
- it might be possible that:
  - specific venues may profit from being in the direct neighborhood of other specific venues.
  - a certain type of venue does not work well in a specific cluster.

These would be valuable information for an entrepreneur who is interested in opening a new venue or for a city that is interested in developing a specific area.

- This work will focus on:
  - **culinary venues** of three mayor European cities
  - the city of **Berlin** for an in-depth analysis of the culinary clusters

Data Sources

three data sources used:

1. simplemaps world cities database:

content:

- all prominent cities of the world
- population
- location

2. Foursquare explore endpoint:

returns:

- up to 100 recommended venues in a specified radius around location
- their location
- their category

3. Foursquare categories endpoint:

returns:

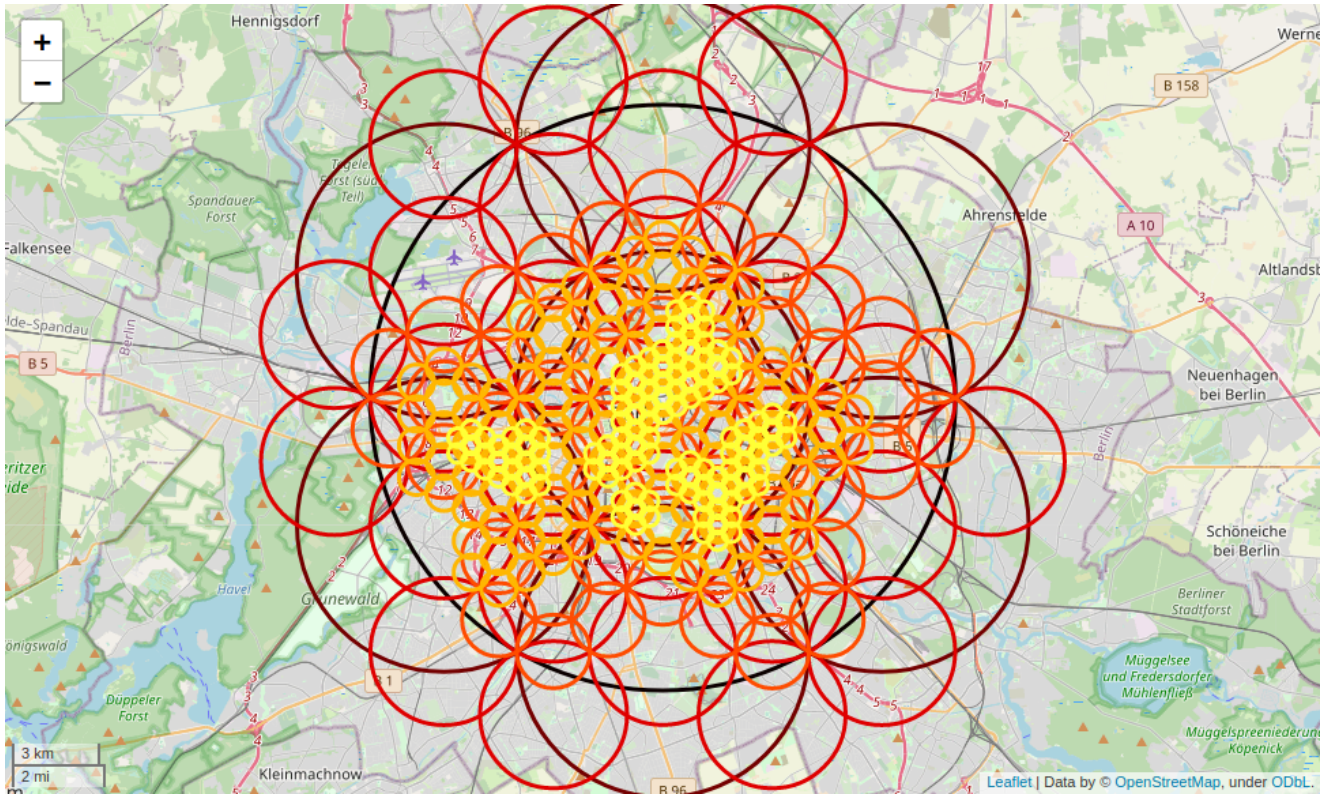
- all categories available in foursquare
- structured data-set, categories split in multiple subcategories

Data Acquisition

- number of items obtained with foursquare explore request limited to 100

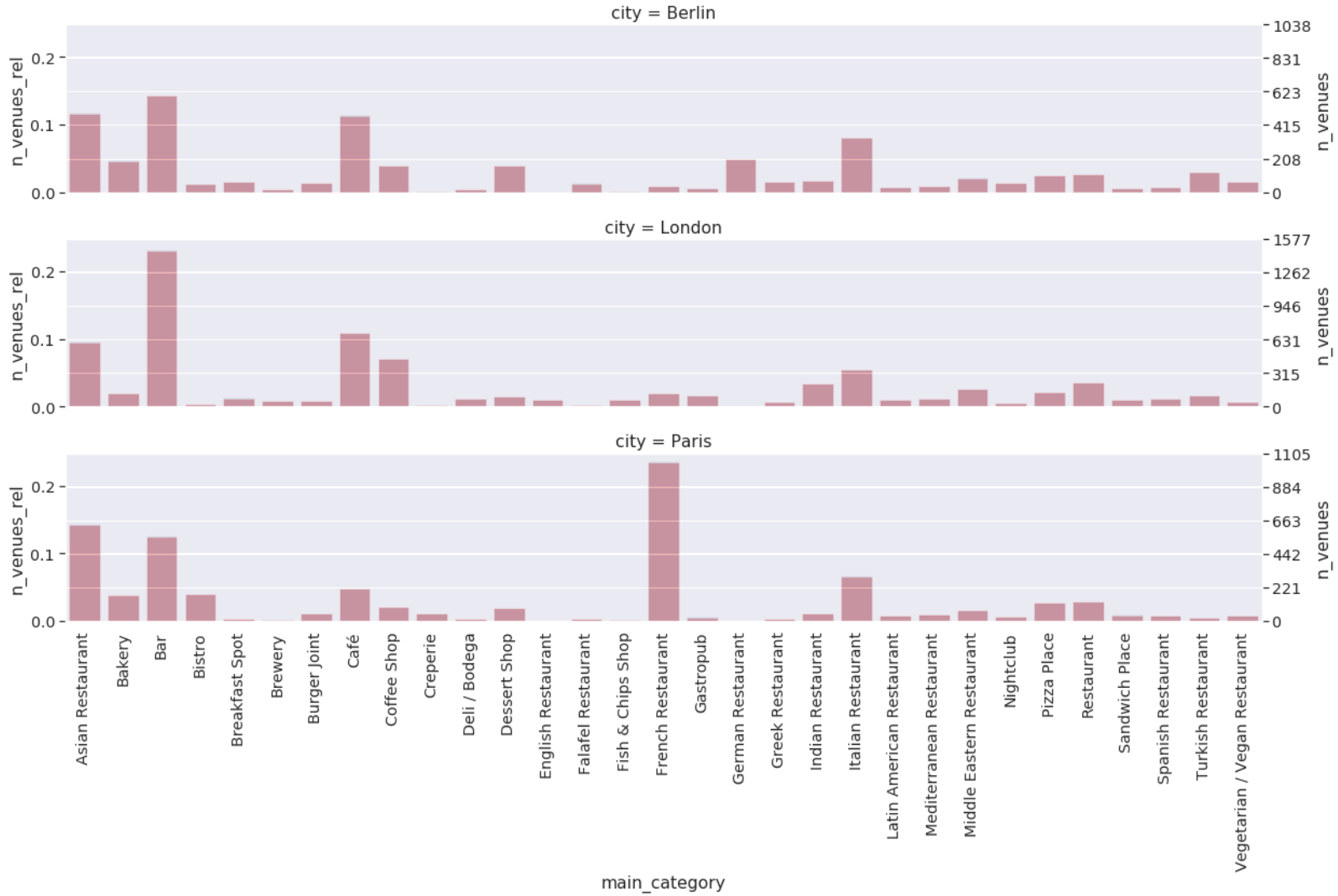
$\Rightarrow$ developed adaptive routine that decreases radius of request in areas of high venue density

$\Rightarrow$ can extracts all recommended venues of a city



Culinary Fingerprints

Comparison of composition of recommended venues in Berlin, London and Paris:

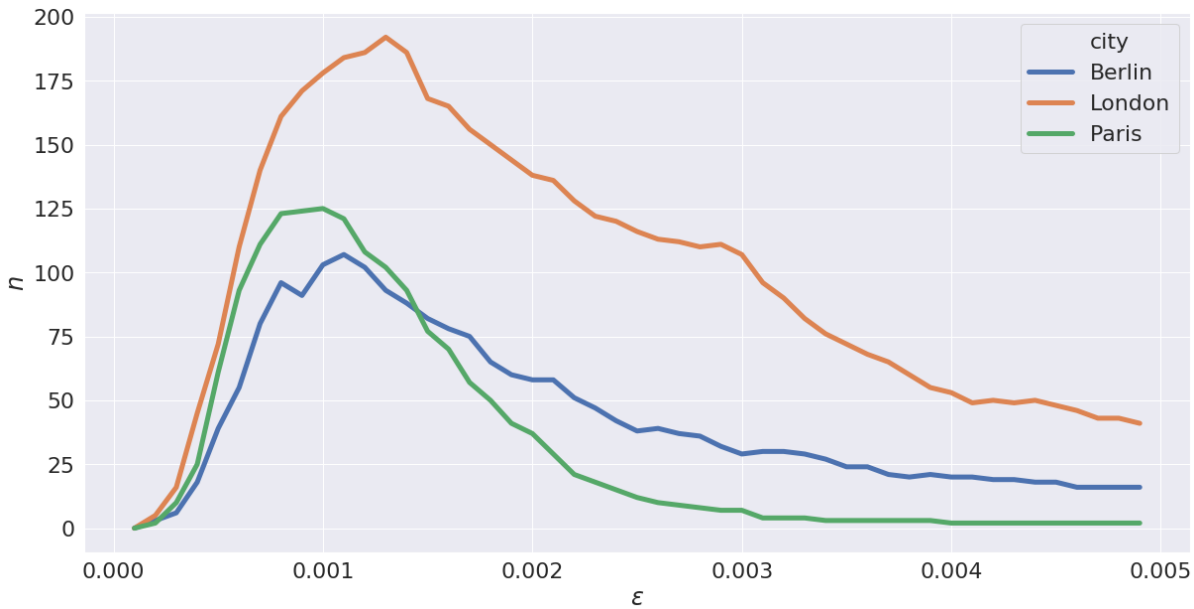


DBSCAN Parameters

DBSCAN uses two mandatory parameters:

- `min_samples` - number of items close to each other needed to start cluster
- ϵ (or `eps`) - radius in which this number of items must be found in order to produce cluster, all points in distance ϵ from existing cluster added to cluster

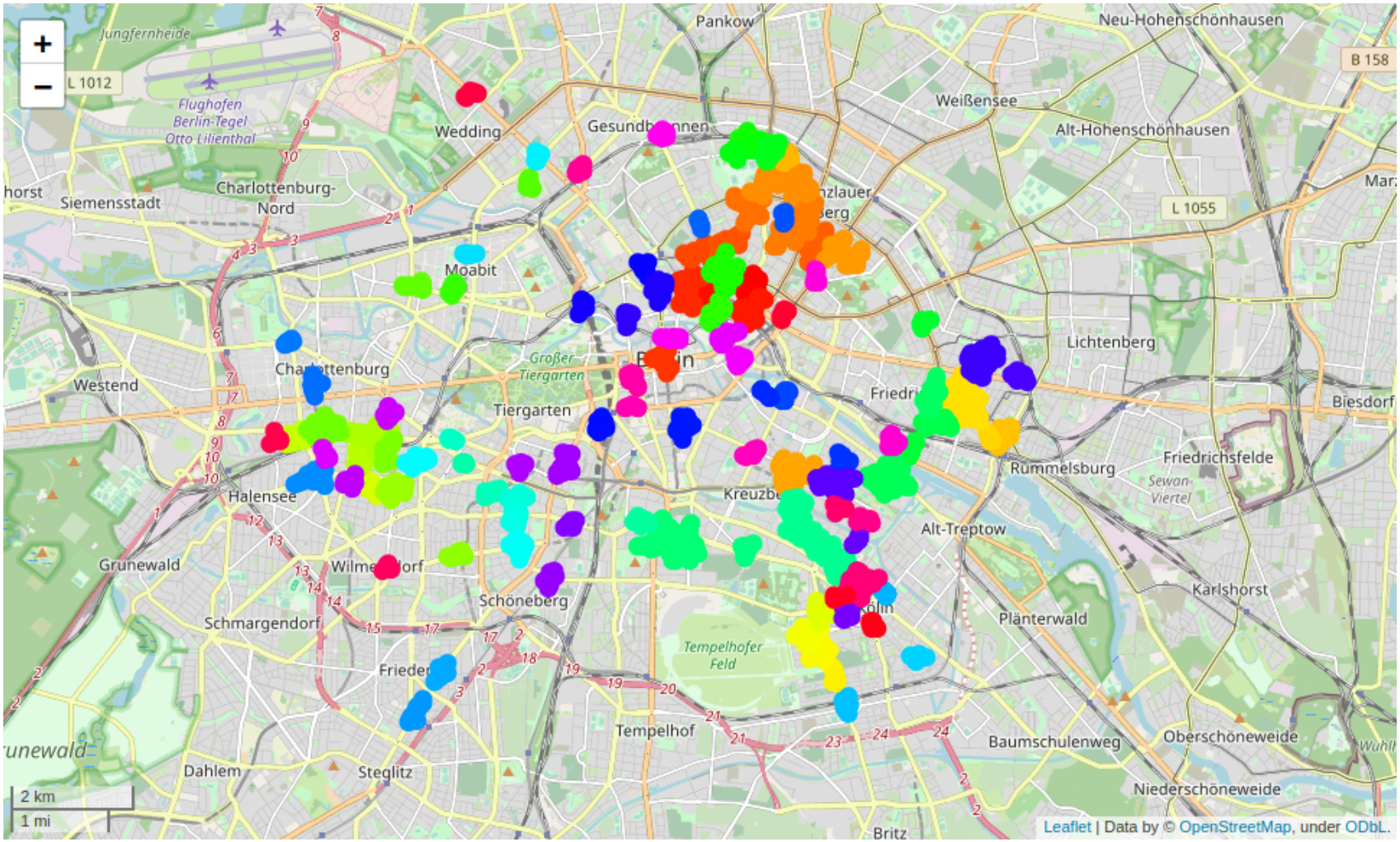
to have meaningfull data to analyse `min_samples`$= 7$



ϵ set to value with maximum number of clusters $\epsilon = 0.0011$

DBSCAN in Berlin

The DBSCAN algorithm produces 107 spatial clusters in Berlin



Analysis of unclustered Data

Difference from the category composition of venues not in clusters to the average category composition of Berlin

Analysis of unclustered Data

- number of Asian restaurants, bars, coffee shops and vegetarian/vegan restaurants is significantly above average in the clusters ⇒ these venues might be more successful in proximity of other culinary venues

- number of bakeries, cafés, German restaurants, Greek restaurants and Italian restaurants is above average outside the clusters ⇒ these venues are spread more evenly throughout the city

Pricipal-Component Analysis (PCA)

PCA used to:

- identify the categories with the greatest variance in the composition of the culinary clusters
- reduce dimensionality of the data-set from 369 features to 30 features.

Important categories:

```
['Asian Restaurant', 'Bakery', 'Bar', 'Bistro', 'Breakfast Spot',
'Burger Joint', 'Café', 'Cocktail Bar', 'Coffee Shop', 'Dessert
Shop', 'Doner Restaurant', 'French Restaurant', 'German
Restaurant', 'Hotel Bar', 'Ice Cream Shop', 'Indian Restaurant',
'Italian Restaurant', 'Korean Restaurant', 'Middle Eastern
Restaurant', 'Nightclub', 'Pizza Place', 'Pub', 'Restaurant',
'Seafood Restaurant', 'Steakhouse', 'Sushi Restaurant',
'Trattoria/Osteria', 'Turkish Restaurant', 'Vegetarian / Vegan
Restaurant', 'Vietnamese Restaurant', 'Wine Bar']
```

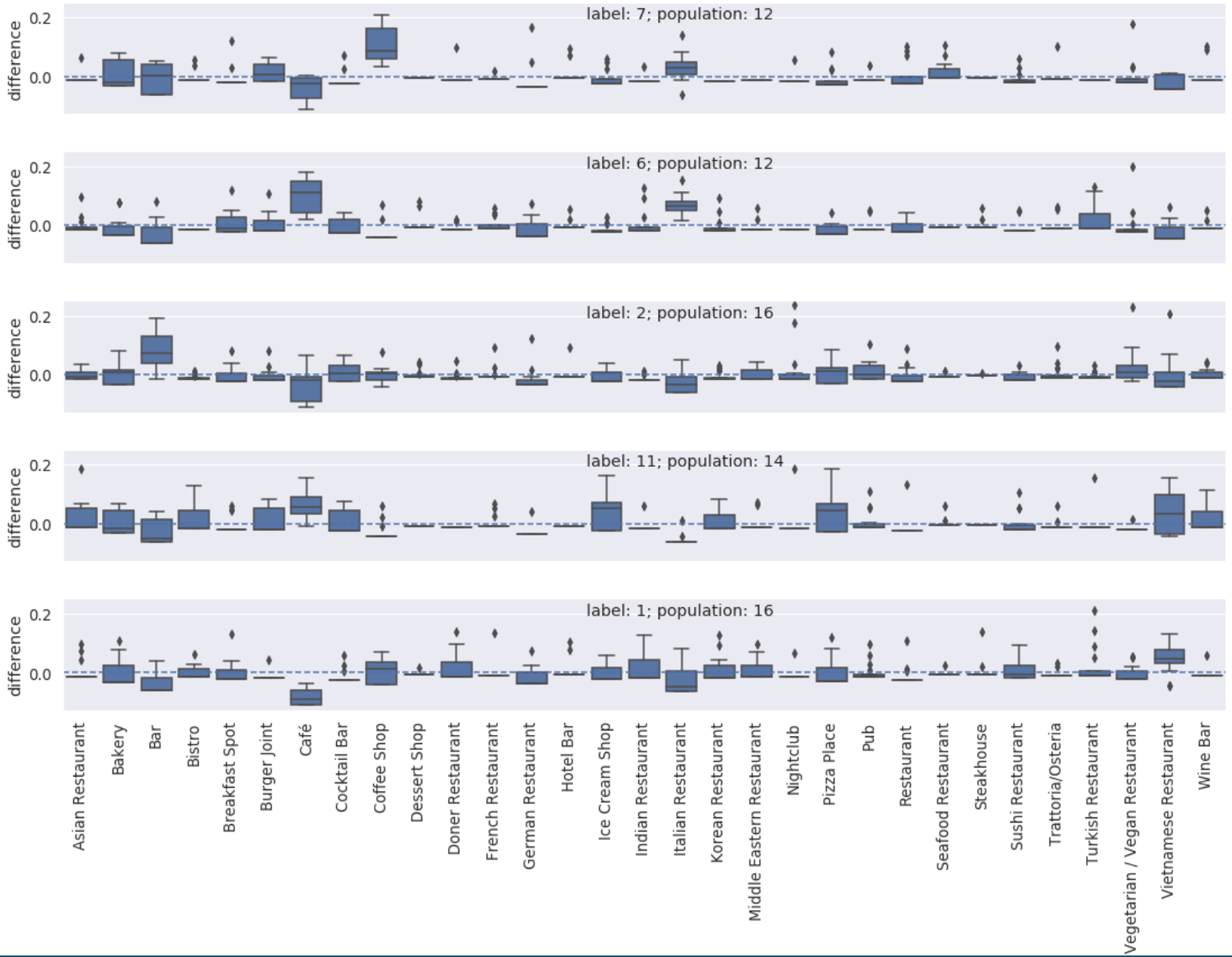
Kmeans-Clustering: parameter `n_clusters`

cluster spatial clusters with kmeans according to their category composition:

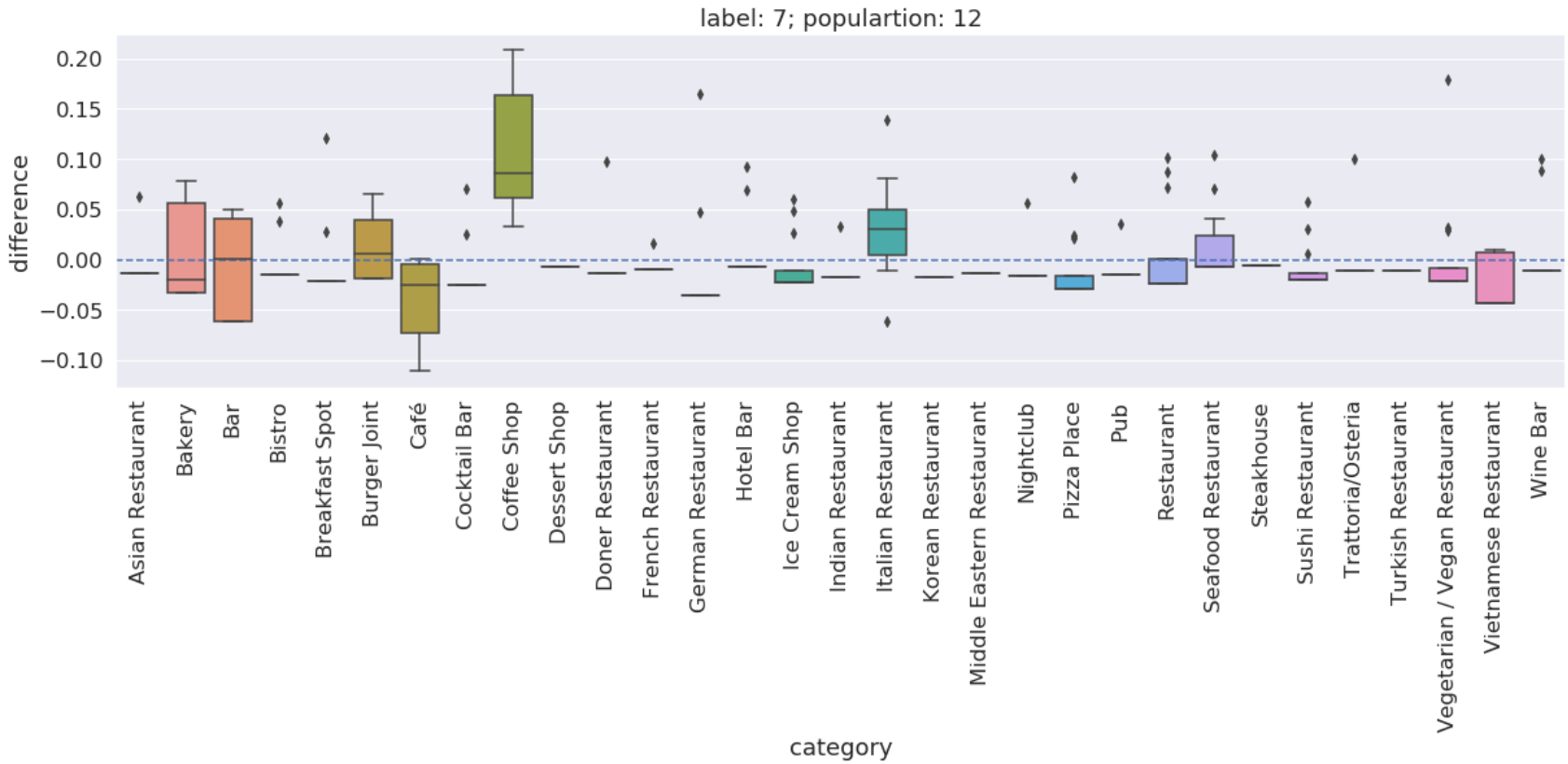
`n_clusters`	sizes of clusters
4	49, 19, 11, 28
5	3, 24, 27, 12, 41
6	16, 5, 47, 1, 27, 11
7	24, 1, 11, 3, 37, 1, 30
8	10, 30, 15, 24, 4, 11, 1, 12
9	11, 19, 13, 25, 3, 11, 1, 10, 14
10	10, 34, 4, 11, 2, 1, 10, 9, 3, 23
11	17, 16, 3, 5, 10, 19, 3, 1, 18, 12, 3
12	10, 33, 4, 11, 1, 1, 10, 9, 3, 23, 1, 1
13	10, 34, 3, 11, 1, 1, 10, 9, 3, 21, 1, 1, 2
14	9, 11, 1, 2, 21, 3, 13, 17, 6, 1, 1, 5, 6, 11
15	10, 33, 3, 11, 1, 1, 12, 8, 3, 19, 1, 1, 2, 1, 1
16	5, 16, 16, 2, 6, 2, 12, 12, 5, 1, 3, 14, 1, 1, 6, 5
17	5, 16, 16, 2, 6, 2, 10, 13, 5, 1, 3, 13, 1, 1, 6, 6, 1
18	5, 16, 16, 2, 6, 2, 8, 13, 2, 1, 3, 14, 1, 1, 6, 5, 1, 5
19	2, 14, 1, 2, 1, 8, 17, 11, 1, 5, 14, 6, 1, 10, 2, 9, 1, 1, 1

compromise on `n_clusters`$= 16$: no cluster with more than 20 items, only 6 clusters with less then 5 items.

Difference in mean composition of 5 largest kmeans-clusters and mean composition of all spatial clusters

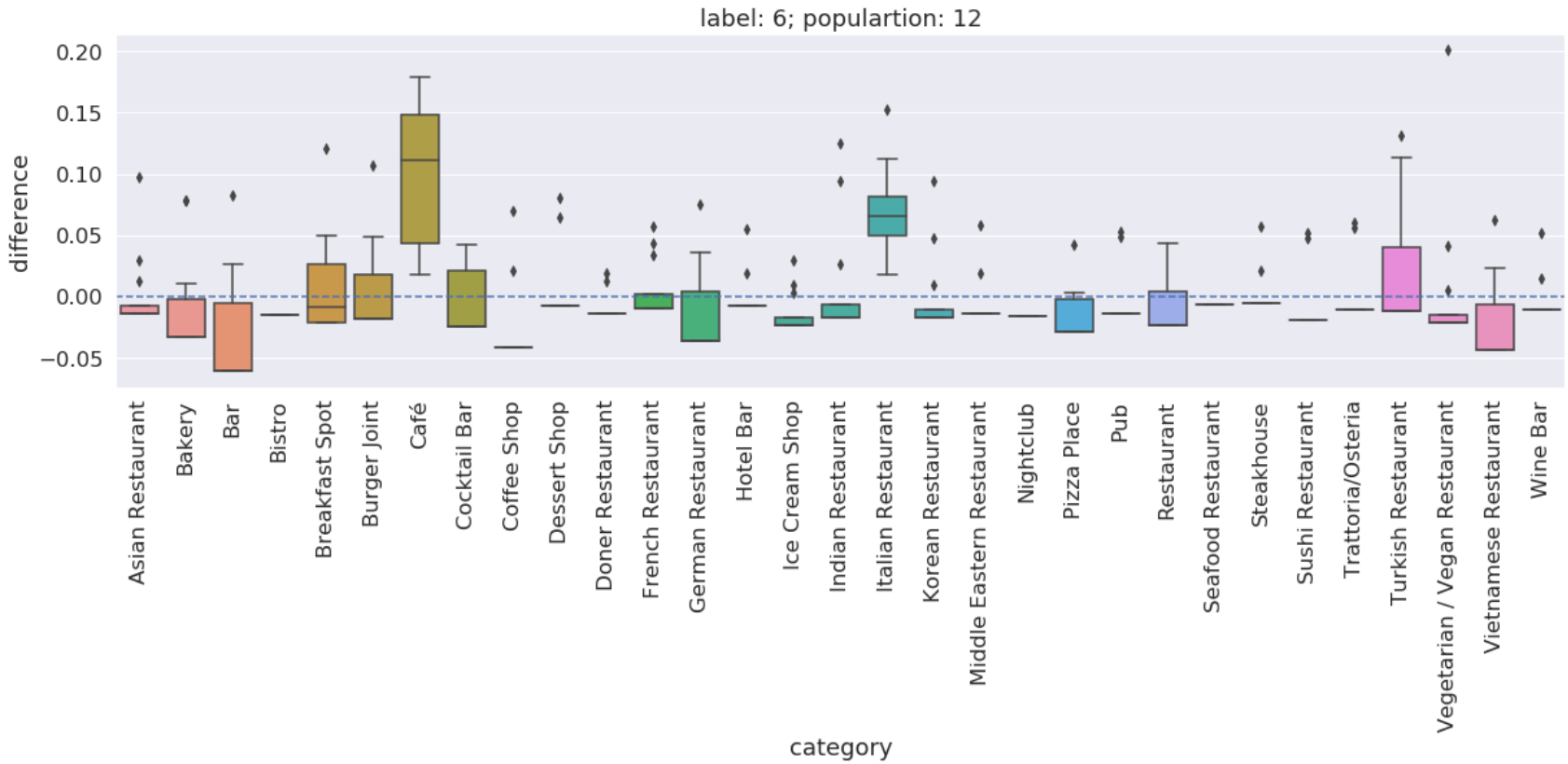


Cluster 7



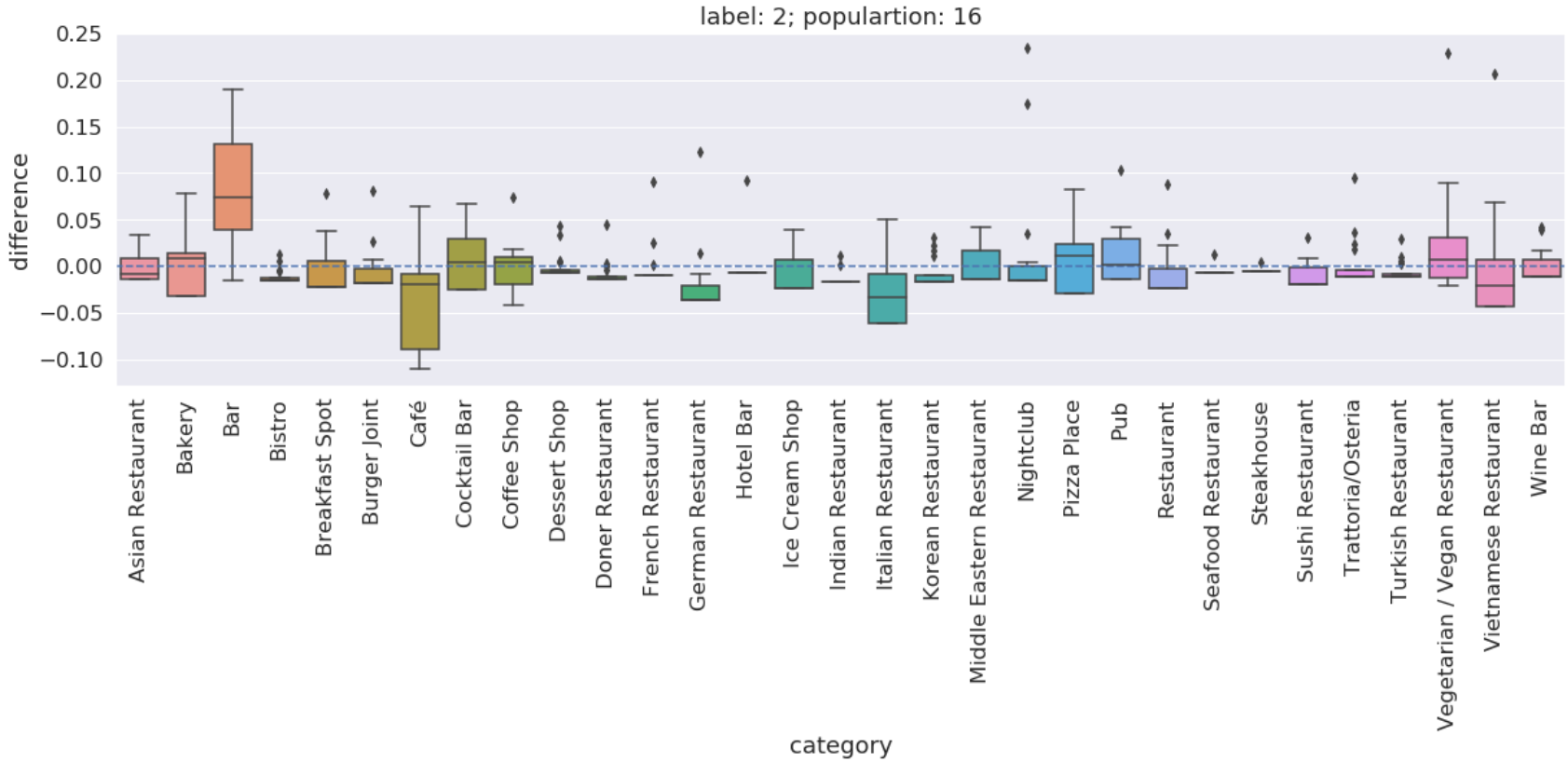
- large number of coffee shops and Italian restaurants
- number of cafes, bakeries and Vietnamese restaurants below average

Cluster 6



- number of cafes and Italian restaurants significantly above average
- number of bakeries, bars and Vietnamese restaurants significantly below average

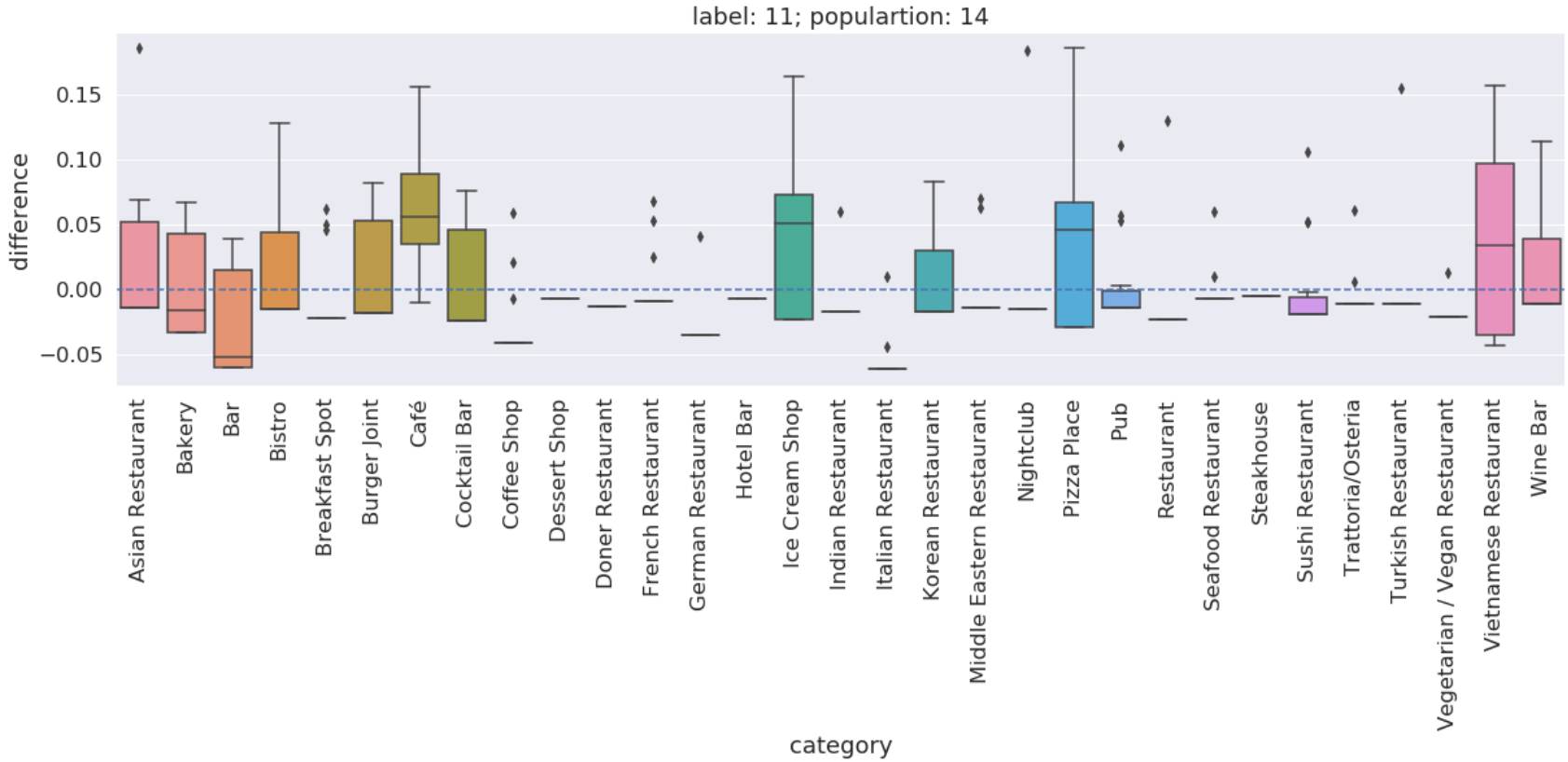
Cluster 2



- above average number of bars and Pizza places
- below average number of cafes and Italian restaurants
- large outliers in the "Nightclub"category

$\Rightarrow$ focused on nightlife

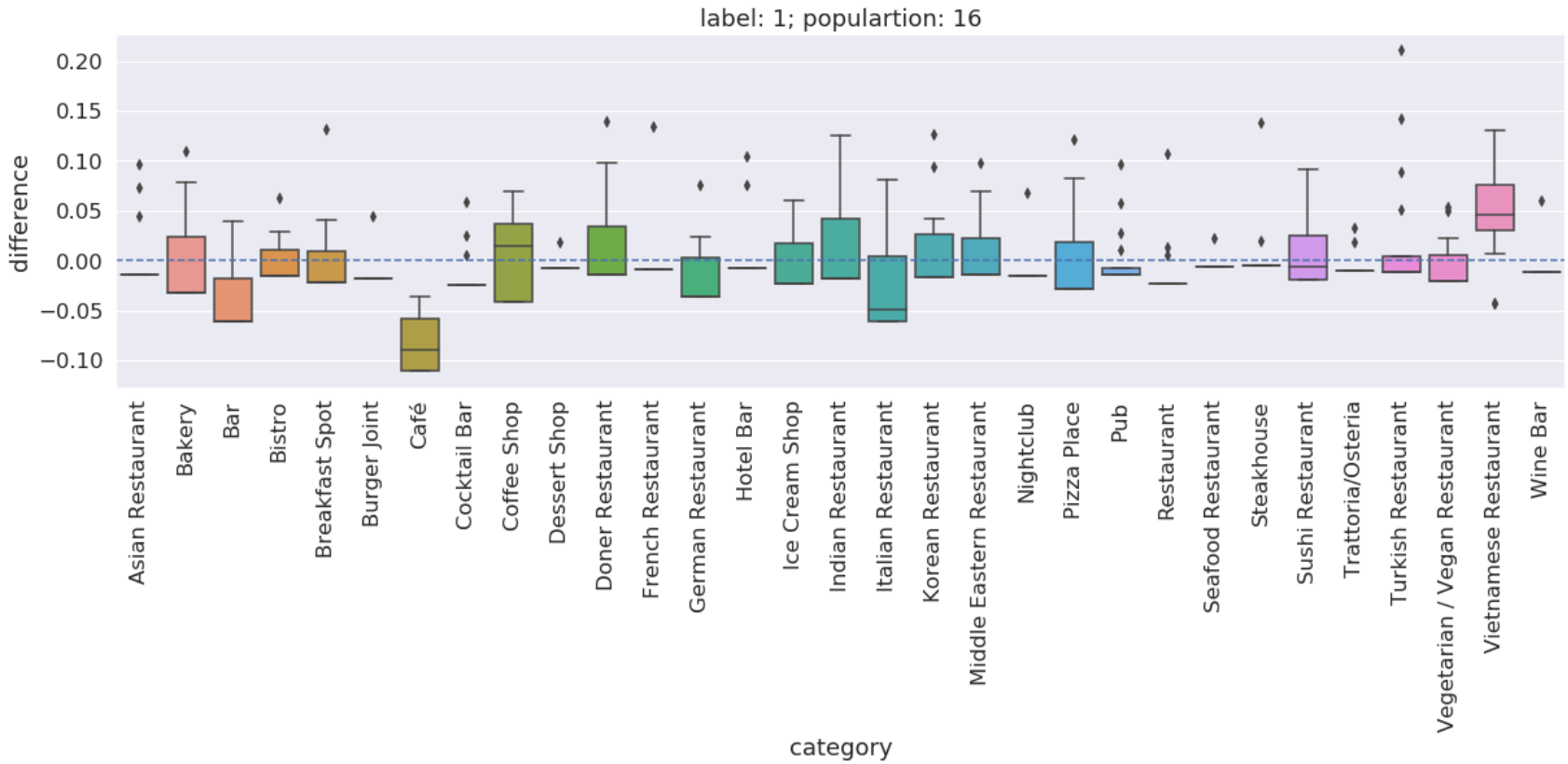
Cluster 11



- above average number of cafes, ice cream shops, pizza places and Vietnamese restaurants
- below average number of bars and Italian restaurants.

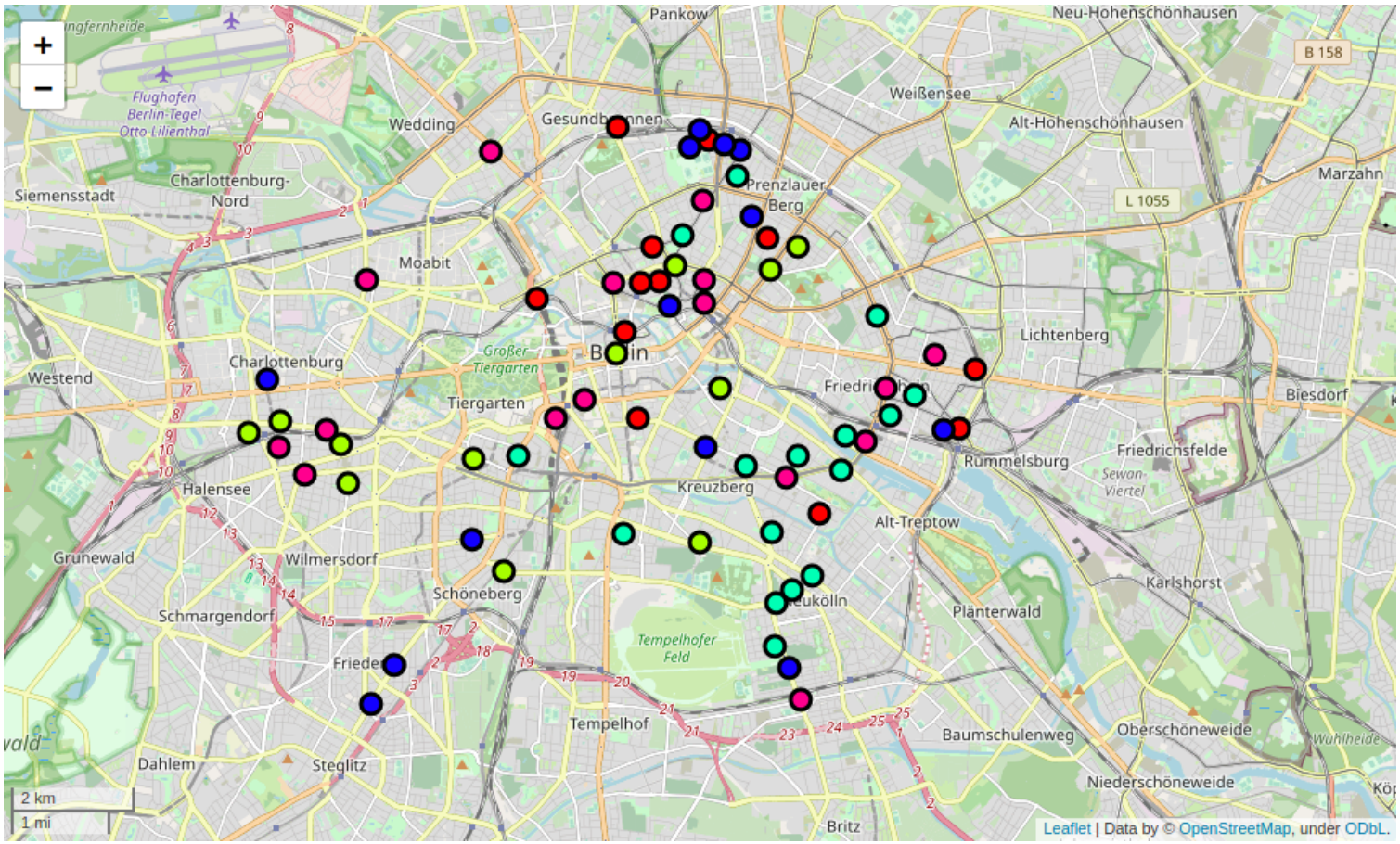
$\Rightarrow$ promising composition for a successful cluster

Cluster 1



- high variance in many categories
- low number of Bars, Cafes and Italian restaurants
- high number of Vietnamese restaurants

Distribution



●:7, ●:6, ●:2, ●:11, ●:1

- '2'-clusters (nightlife) tend to be in the south-east of Berlin
- '6'-clusters tend to be in the north-east of Berlin

Conclusion

- culinary fingerprints i.e. the venue category composition of Berlin, London and Paris identified
- data for Berlin studied further:
  - spatial venue clusters identified from the location data of venues
  - category composition of venues not included in any spatial clusters compared with the average category composition of Berlin:
    - Asian restaurants, bars, coffee shops and vegetarian/vegan restaurants can more often be found in spatial clusters
    - bakeries, cafés, German restaurants, Greek restaurants and Italian restaurants are more evently distributed in the city of Berlin
  - composition of spatial clusters was analysed further with pricipal-component anaylsis (pca):
    - pca yields 31 categories with highest contribution to variance of the data-set
  - and kmeans-clustering:
    - data quite noisy
    - some categories are more often found next to each other then other categories.
    - E.g. density of restaurants seems to be significantly lower in areas with a high density of bars and vice versa.