Enhanced location information for potential home buyers

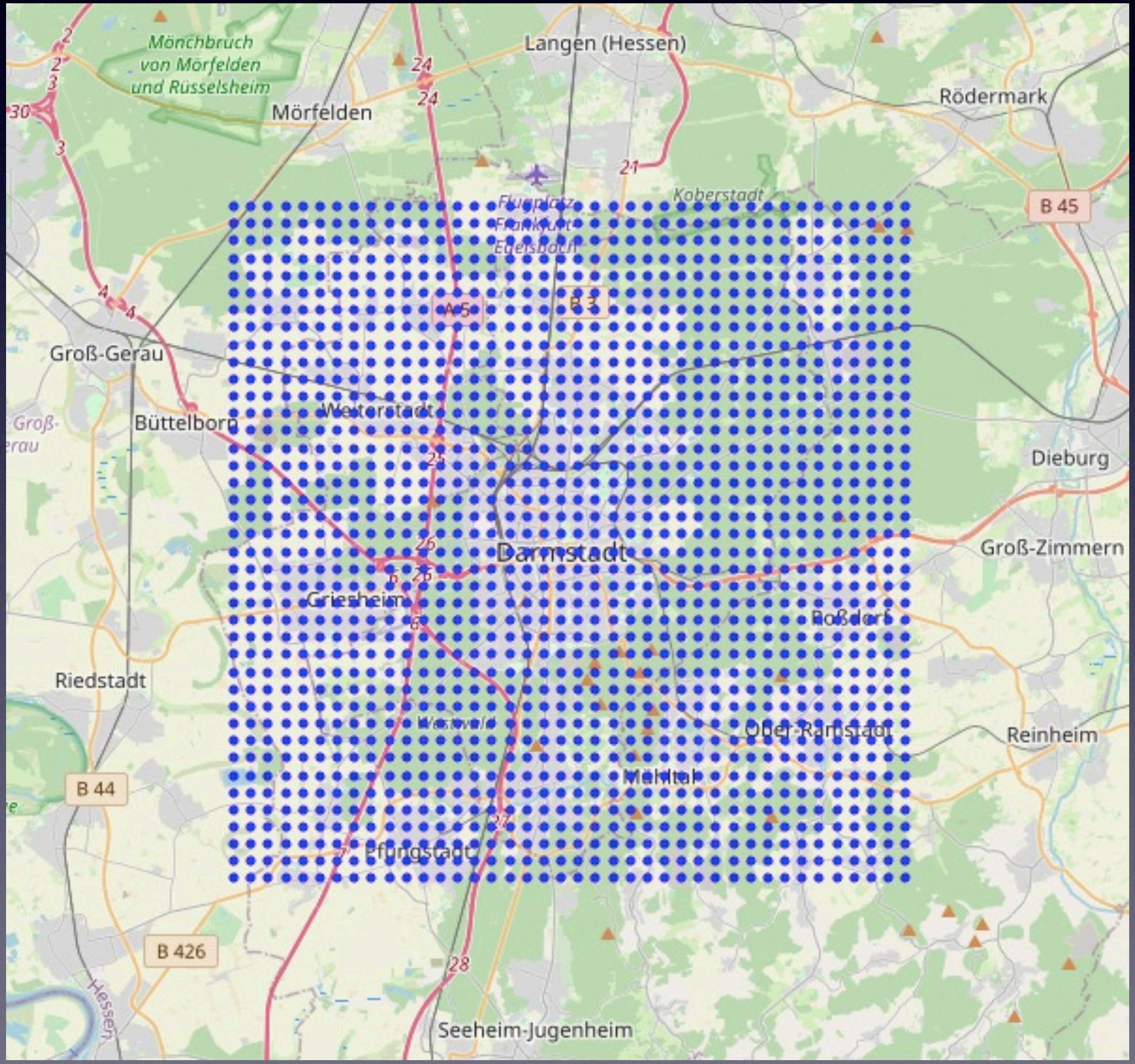
Application to a medium german city (Darmstadt) using machine learning

Location, location, location

- For most people, home purchase is a huge project
- Determines life quality for many years/decades
- It is important to be close to (as an example):
    - Transportation
    - Schools, local amenities
    - Nature

Application on medium german city (Darmstadt)

- Basic advantage: I have good knowledge of the area
- I can check that the analysis produces good results
- Create spatial grid
  - segment size ~ 500x500 m

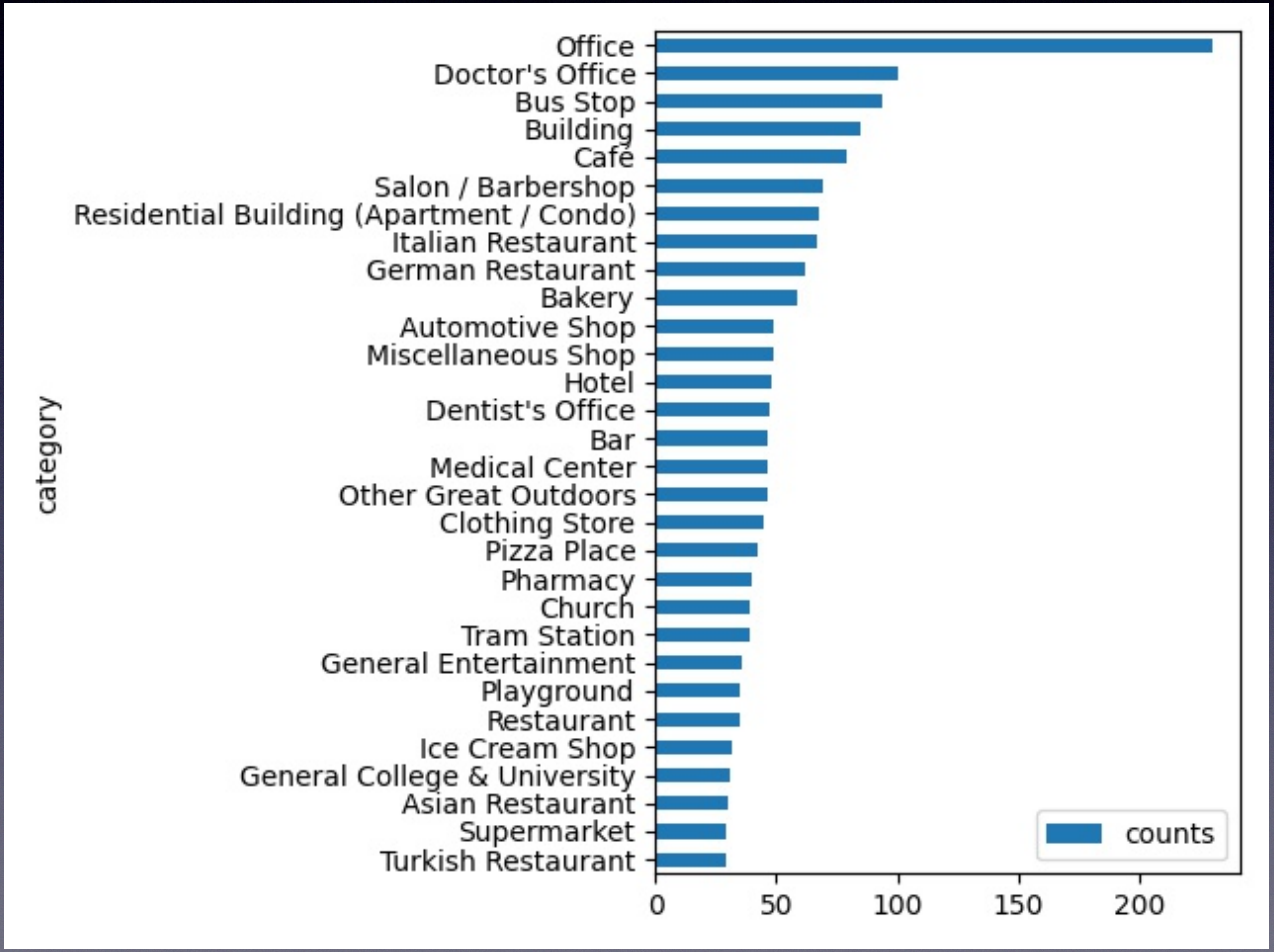


Data processing

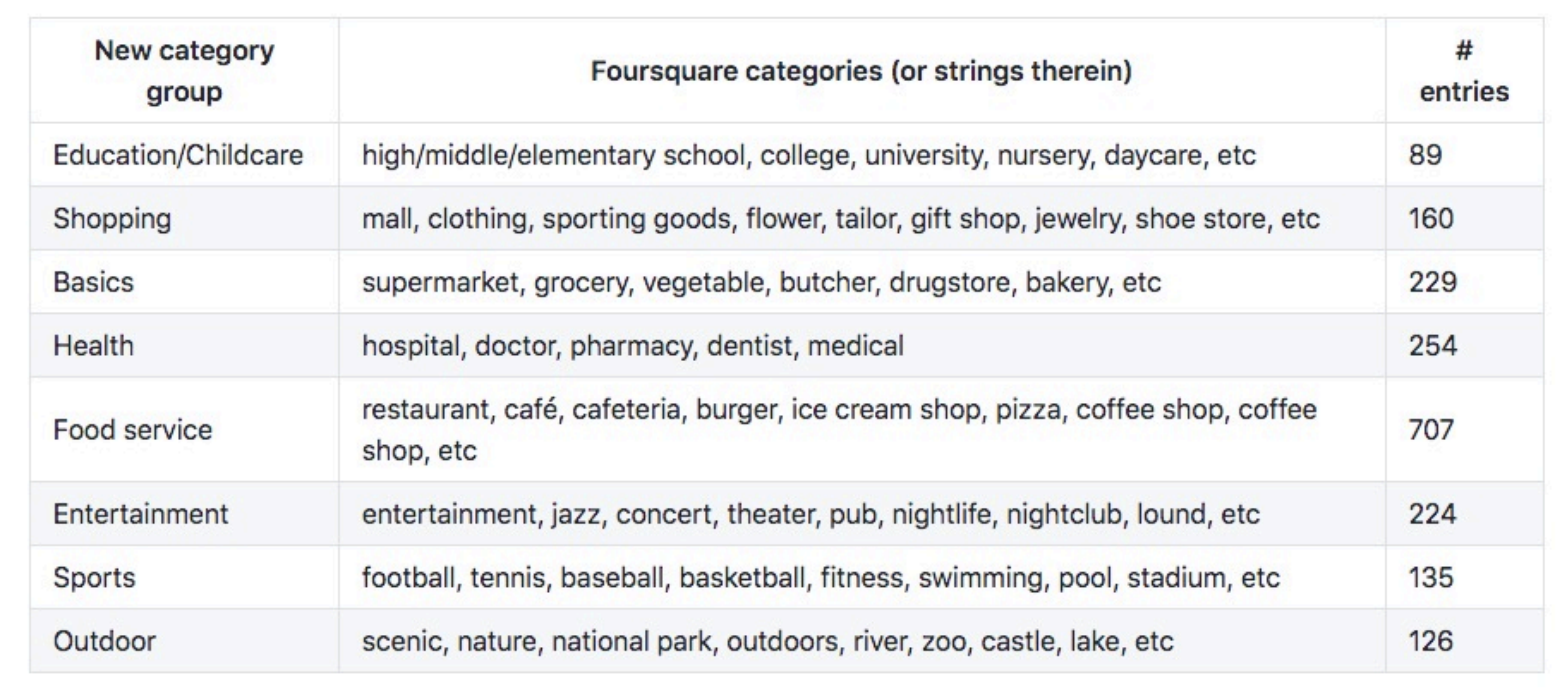
- Use Foursquare API to get data for venues
- Map data onto the grid using the Haversine distance
- Process data
    - Filter
    - Group venue categories
    - Agregate

Too many venue categories

- Venue categories:
  - too many
  - too detailed
- Need to group them together into supersets
- Some venue categories are not used



Group into supersets



New category group	Foursquare categories (or strings therein)	# entries
Education/Childcare	high/middle/elementary school, college, university, nursery, daycare, etc	89
Shopping	mall, clothing, sporting goods, flower, tailor, gift shop, jewelry, shoe store, etc	160
Basics	supermarket, grocery, vegetable, butcher, drugstore, bakery, etc	229
Health	hospital, doctor, pharmacy, dentist, medical	254
Food service	restaurant, café, cafeteria, burger, ice cream shop, pizza, coffee shop, coffee shop, etc	707
Entertainment	entertainment, jazz, concert, theater, pub, nightlife, nightclub, lound, etc	224
Sports	football, tennis, baseball, basketball, fitness, swimming, pool, stadium, etc	135
Outdoor	scenic, nature, national park, outdoors, river, zoo, castle, lake, etc	126

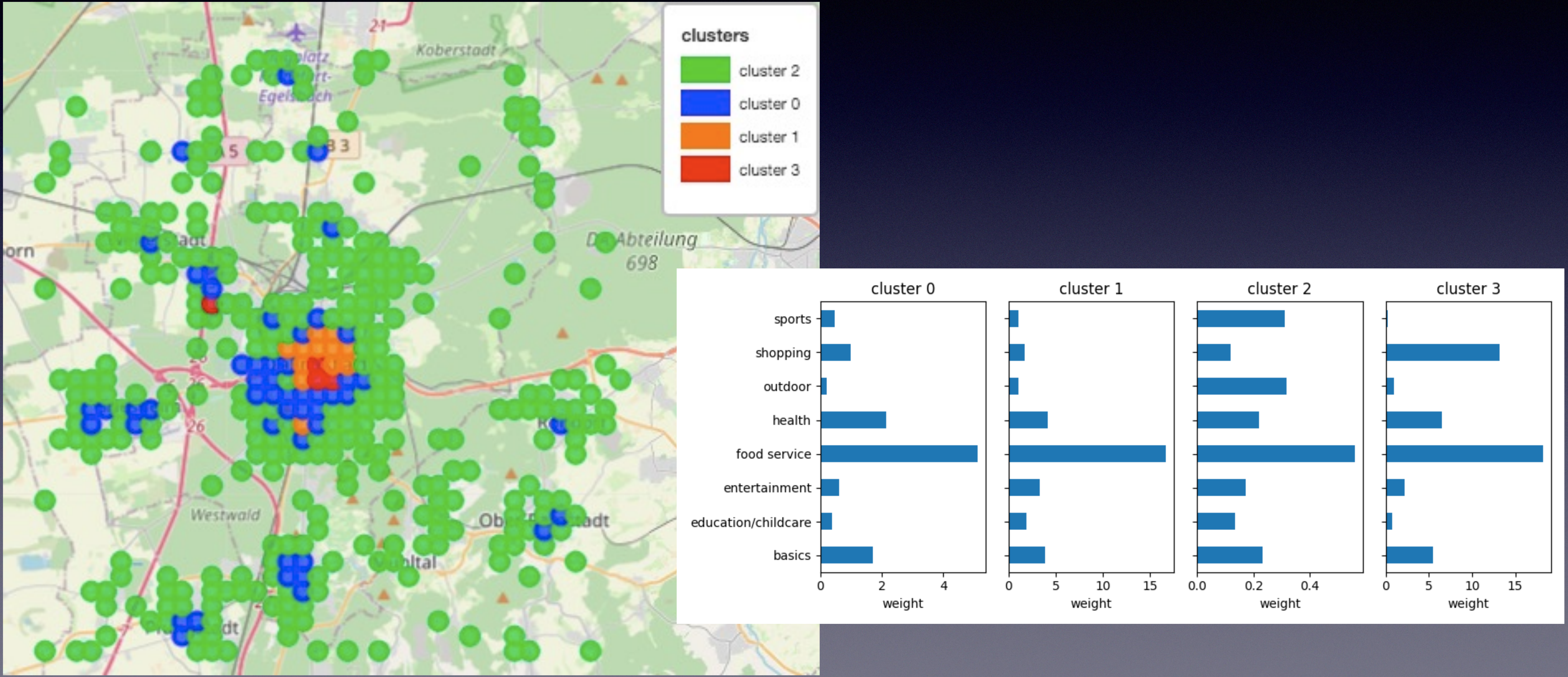
Agregate data in grid segments

- An aggregation operation must be defined:
- Considered:
  - sum (venue categories are added)
  - mean (venue categories are averaged)
  - median

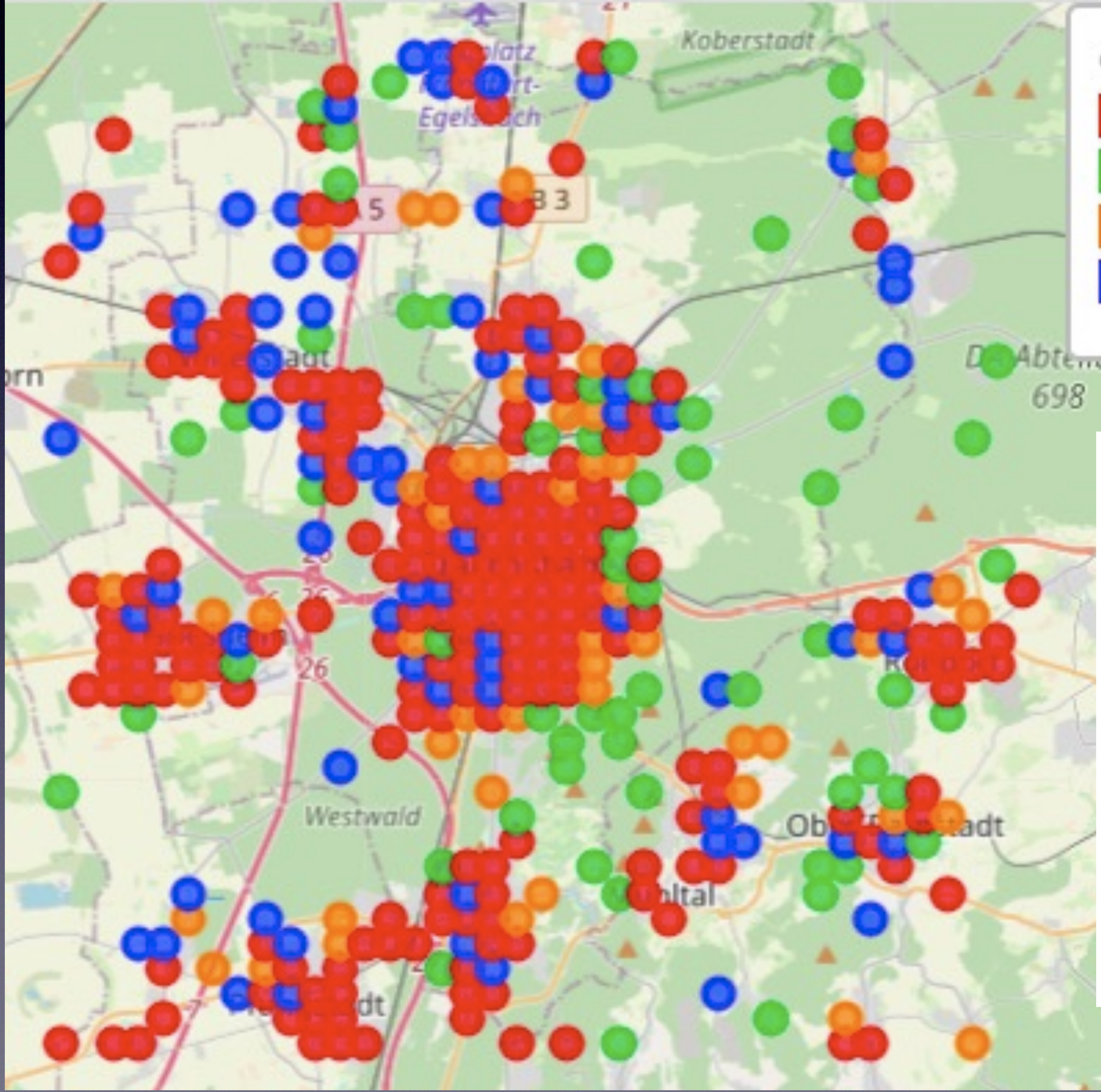
Clustering

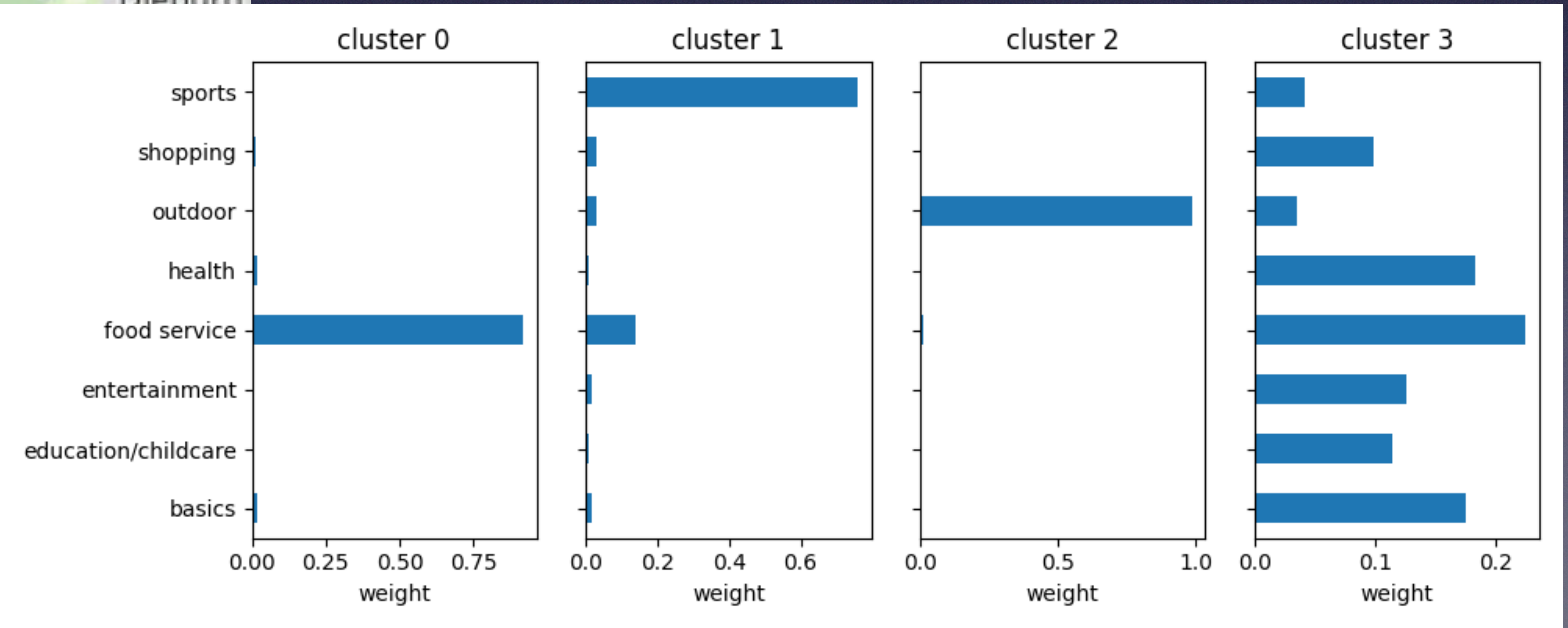
- Unsupervised
- k-means
- Effect of k
- Effect of the aggregation operator
- Silhouette score to evaluate clustering quality

Operator=sum, k=4: city centres shown clearly, otherwise granularity is inadequate

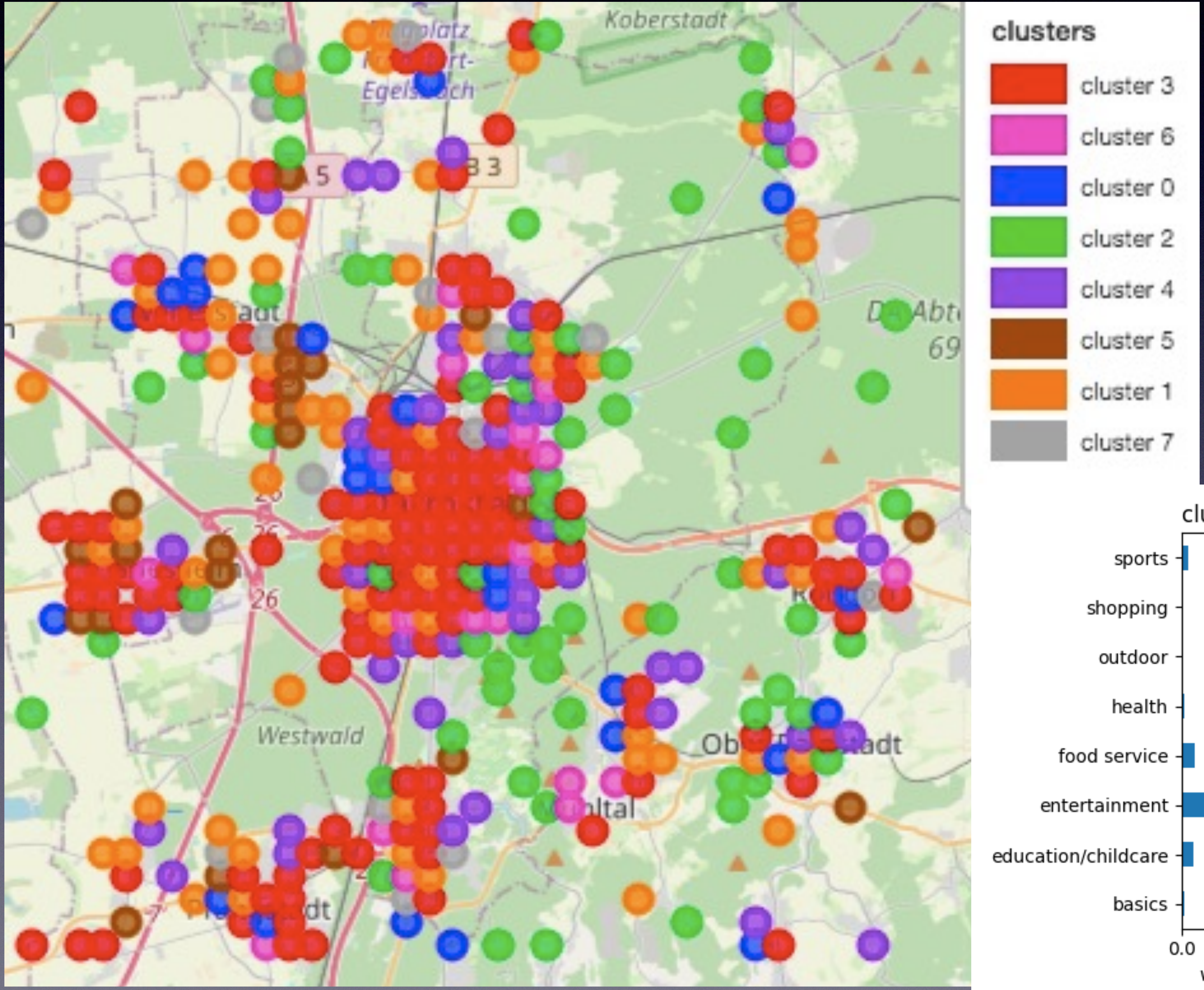


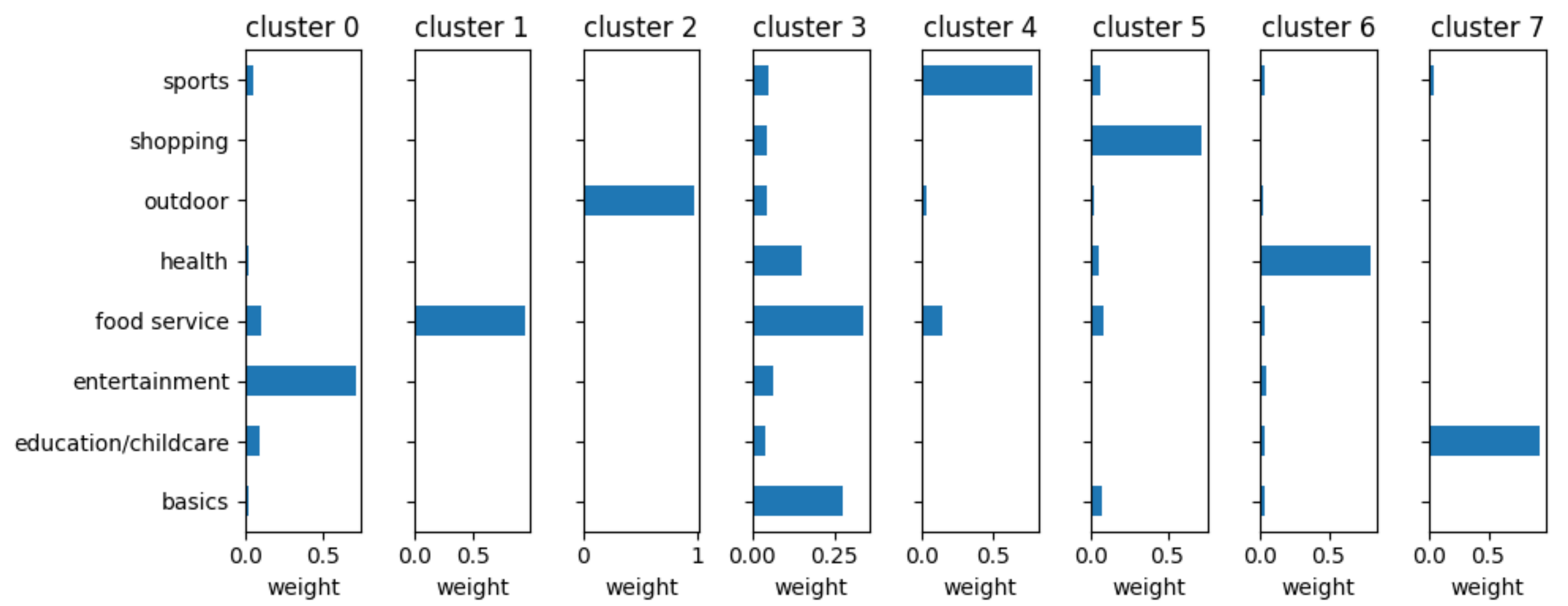
Operator=mean, k=4: clusters are quite far from realistic representation



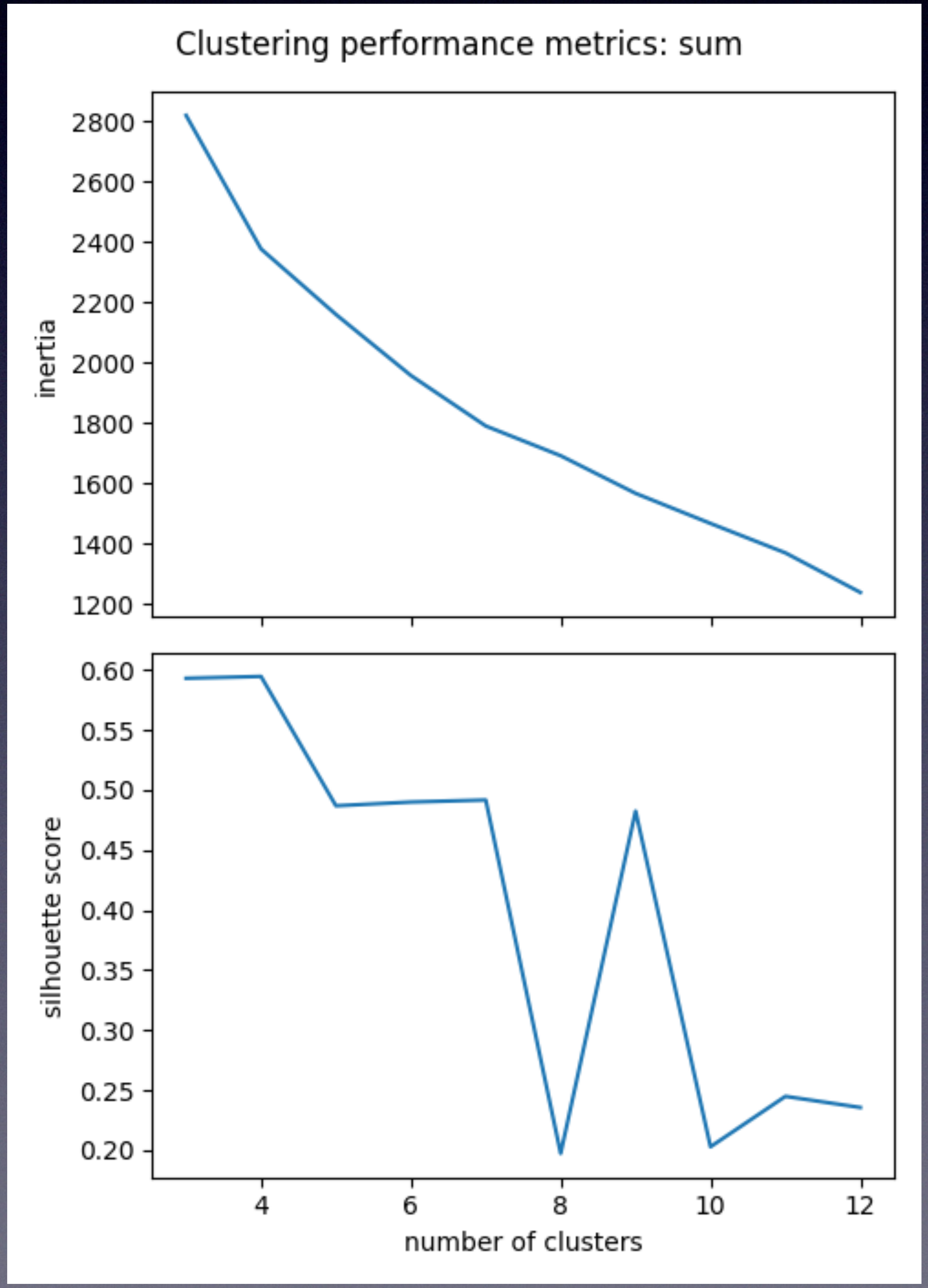


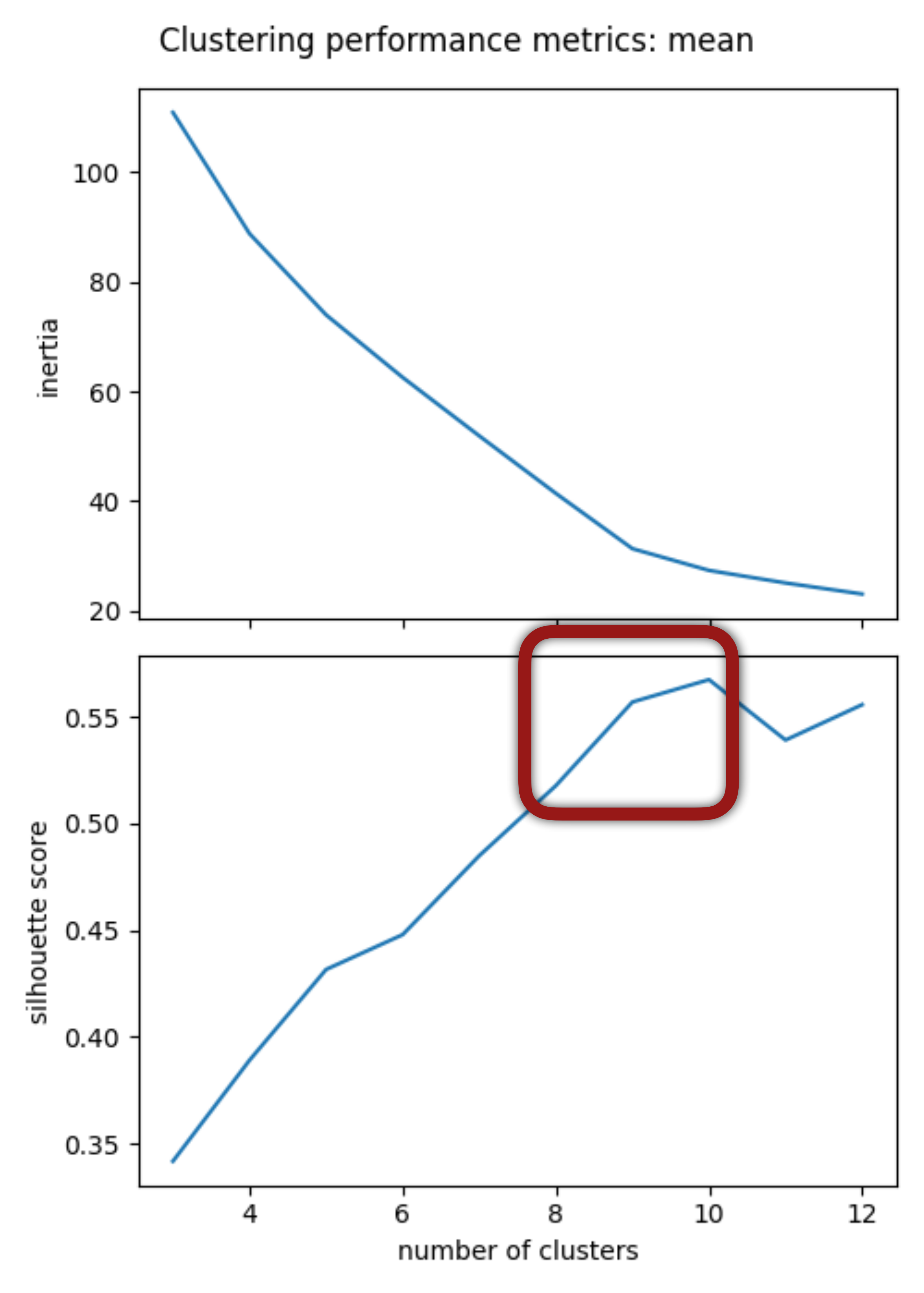
Operator=mean, k=8: clusters more cohesive and respresentative





Clustering quality measured with silhouette score is better for operator=mean, k=8-9





Conclusion

- Performed clustering analysis to provide enhanced location information for potential home buyers
- Sensitive dependence on k and aggregation operator
- Found good representation when k=8, operator=“mean”
- Analysis is generic: can be applied on other areas