Districts of Bangkok

Rob Egrot

Bangkok Facts

- Capital city of Thailand.
- Thailand is classified as a middle income country.
- Population of Bangkok roughly 10,000,000 (out of 70,000,000 in Thailand).
- Divided into 50 districts.
- Sharp divide between rich and poor.
  - Some malls in Bangkok use more electricity than whole provinces elsewhere in the country.
  - Roughly 24% of Bangkok population live in slums.

Classification of Bangkok Districts

- Can we group districts of Bangkok into intuitively meaningful clusters based on socioeconomic factors?
- Do socioeconomic groups have venue/business profiles that can be witnessed by Foursquare data?
- Answer: Yes and yes, though with some caveats.

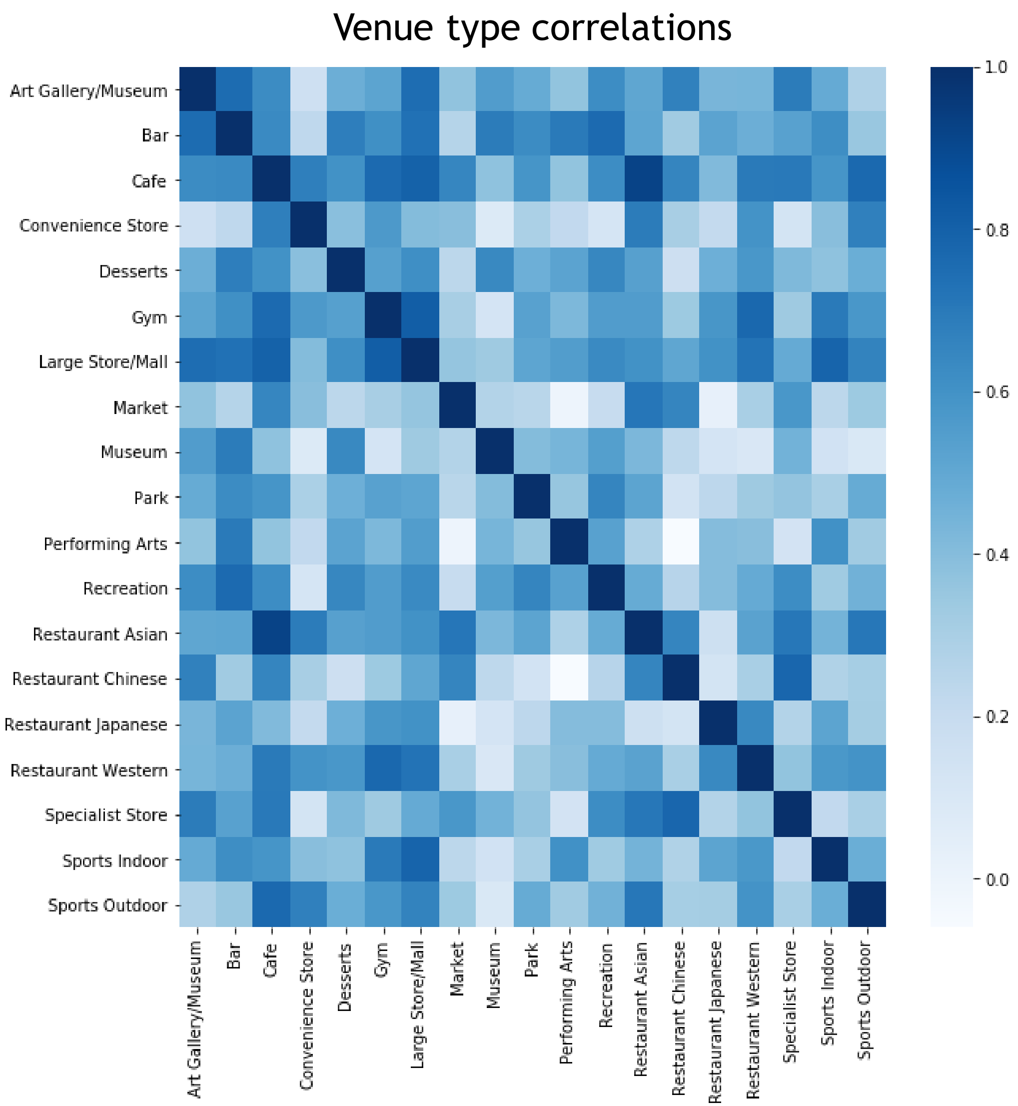
Socioeconomic data

- Basic district data (scraped from Wikipedia).
- Population data (obtained from Bangkok Metropolitan Administration – BMA).
- Community data (obtained from BMA).
- Data on schools (obtained from BMA and International Schools Association of Thailand – ISAT).
- Data on new businesses (obtained from BMA).
- Location of rapid transit stations (scraped from Wikipedia).

Socioeconomic features

	pop_total	area_km2	pop_density	C3_slum	C4_urban	C5_suburb	No. Government Schools	No. ISAT Schools	New retail capital	New wholesale capital	No. New Businesses	No. Rapid Transit Stations
count	50.000000	50.000000	50.000000	50.000000	50.000000	50.00000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000
mean	109346.040000	31.374640	6316.660000	13.160000	7.100000	8.52000	11.940000	3.500000	163.435345	333.261277	165.126667	2.300000
std	47761.066583	40.214508	4152.239718	12.179541	16.177901	13.87148	6.864312	5.218687	158.692747	522.558144	89.717139	2.991485
min	23655.000000	1.416000	732.000000	0.000000	0.000000	0.00000	2.000000	0.000000	28.517000	32.800000	33.000000	0.000000
25%	76342.000000	10.729000	4070.500000	3.000000	0.000000	0.25000	7.000000	1.000000	71.918792	92.610000	95.250000	0.000000
50%	103060.000000	19.027000	5293.500000	9.500000	0.000000	3.00000	10.000000	1.000000	103.953333	169.216667	138.166667	1.000000
75%	145878.250000	34.285750	8363.000000	18.000000	3.000000	9.75000	15.000000	6.000000	205.610000	296.325000	243.583333	4.000000
max	204532.000000	236.261000	23667.000000	47.000000	82.000000	73.00000	40.000000	25.000000	989.063047	3280.390000	382.666667	13.000000

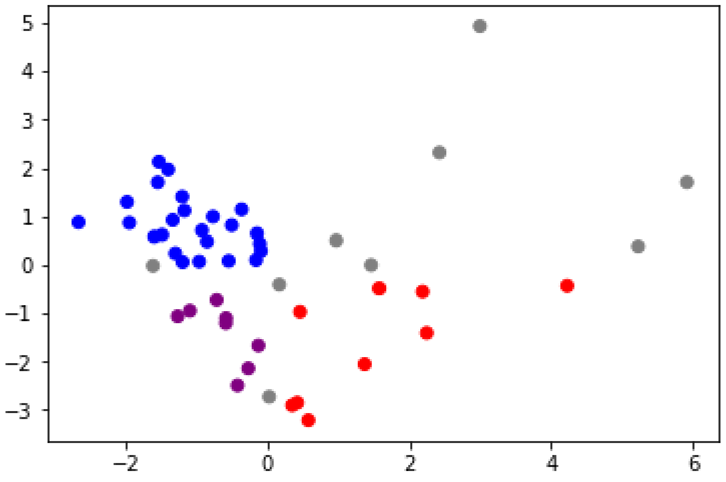
Foursquare data



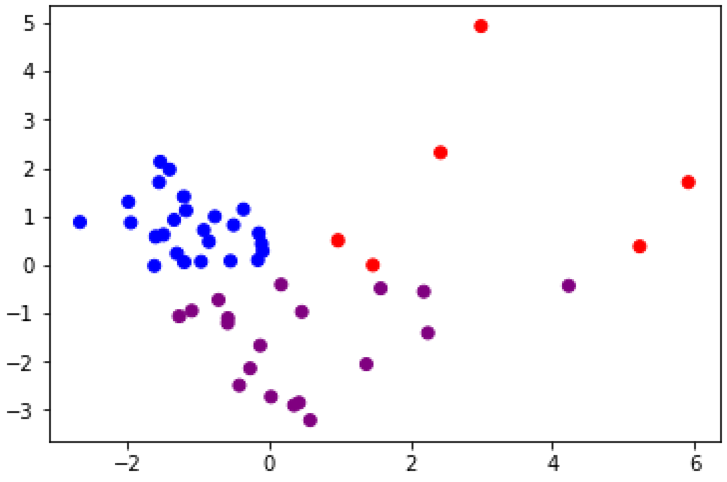
- Obtained for each district using Foursquare API.
- Manual editing to remove redundant venue types.
- 19 venue types obtained.

Clustering

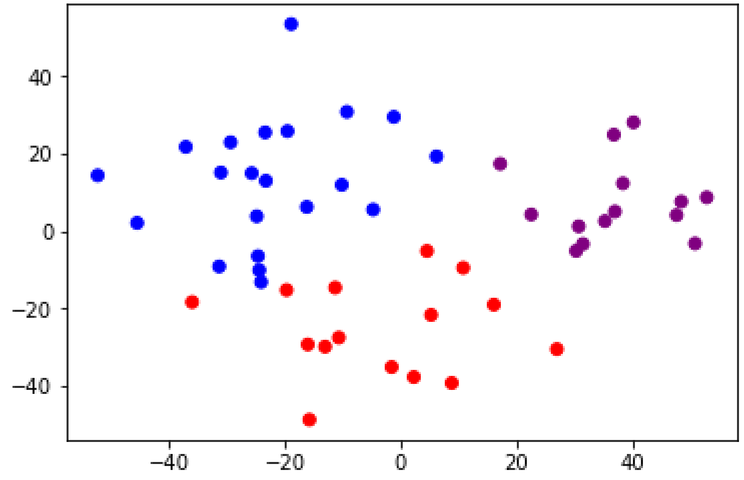
- PCA used to reduce to 2 dimensions for better defined clusters.
- Clustering performed using DBSCAN, K-Medoids and Agglomerative Clustering.
- K-Medoids performed on *ranked* data (so axis are different).
- Reasonable degree of consistency across methods.
- 3 clusters identified.



DBSCAN



Agglomerative



K-Medoids

Clusters Identified

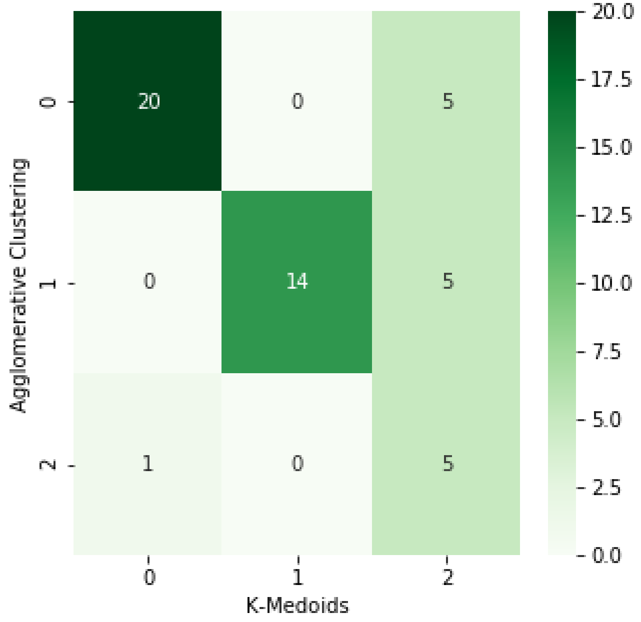
	DBSCAN	Agg	K-Med
DBSCAN	1	0.82	0.55
Agg	0.82	1	0.53
K-Med	0.55	0.53	1

Adjusted Rand scores of cluster agreements.

	DBSCAN	Agg	K-Med
0	24	24	21
1	8	19	14
2	9	6	15

Cluster sizes

- 3 categories of district identified.
- Type 0 (relatively suburban):
  - Low number of rapid transit links.
  - Generally average or a little below average business investment.
  - Below average population density.
- Type 1 (urban relatively poor):
  - High population density.
  - High proportion of the population living in slums.
  - Low investment.
- Cluster 2 (urban relatively affluent):
  - Similar to Cluster 1.
  - Higher capital investment
  - Better rapid transport links.
- Caveat: distinctions between clusters not always sharp.
  - Borderline districts.

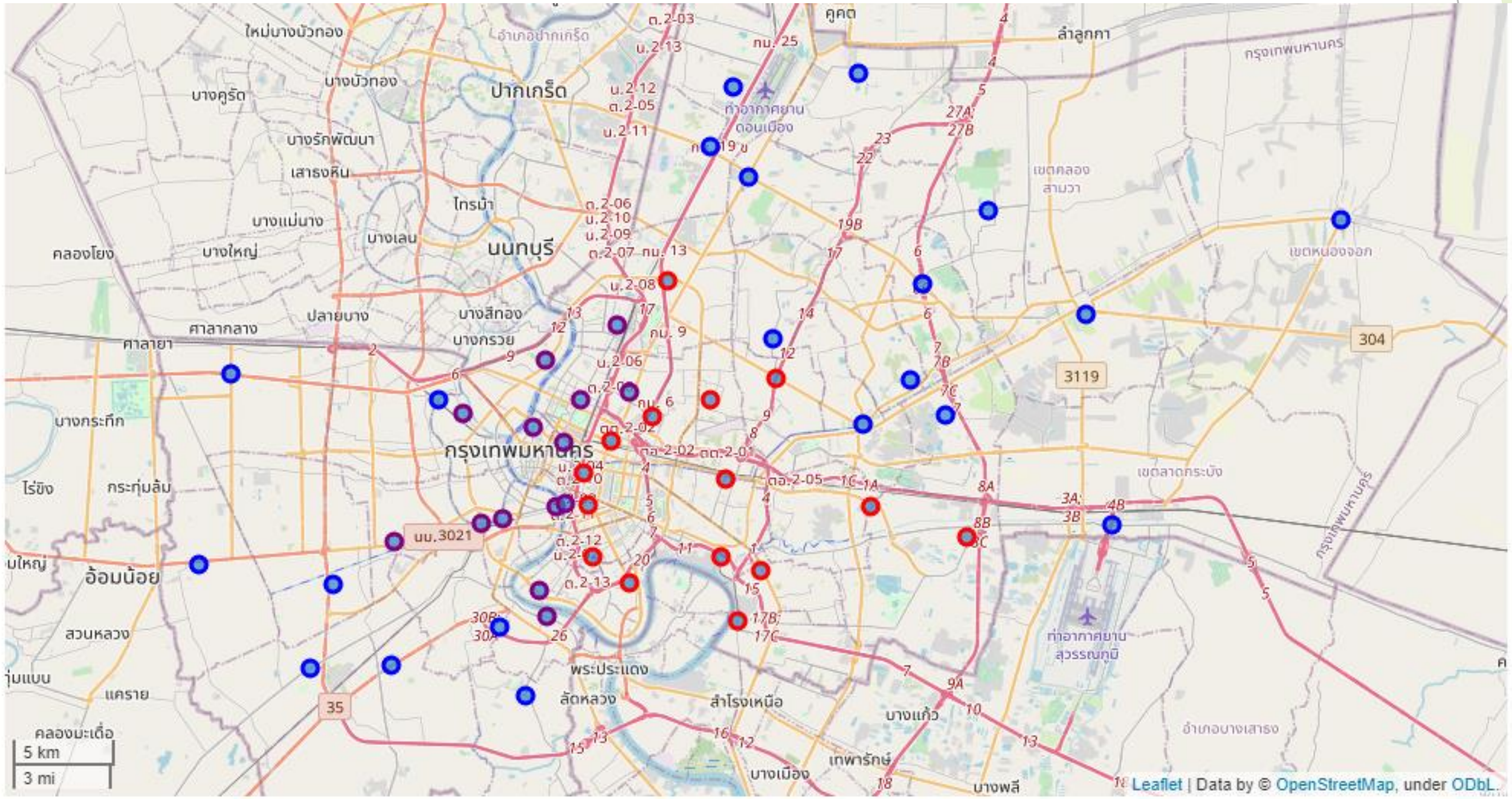


Confusion matrix Agg. vs K-Med.

K-Medoids Clusters

- Used K-Medoids clusters as these have good number of districts in each category.

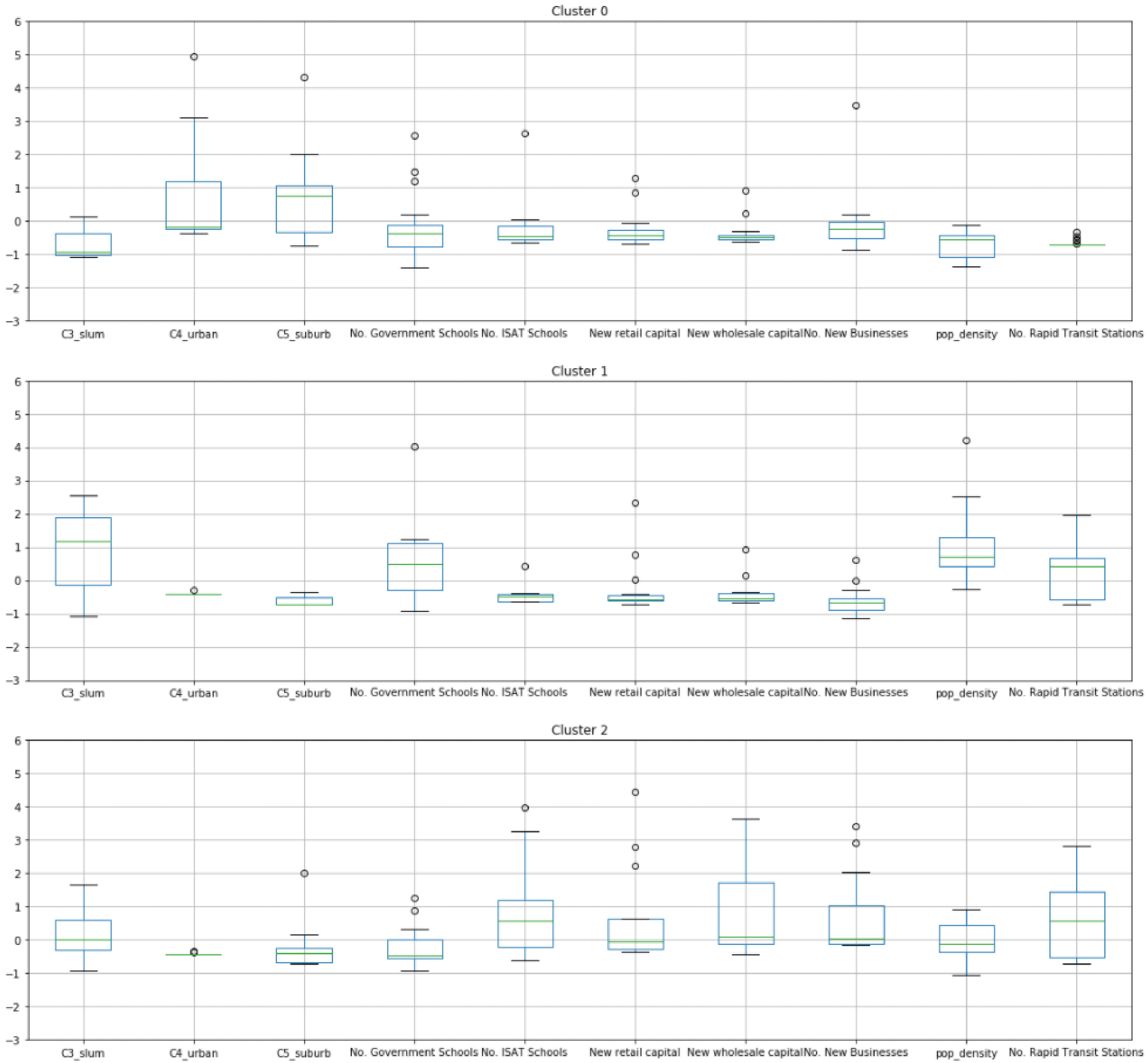
Type 0: Blue
Type 1: Purple
Type 2: Red



Clear geographical pattern visible

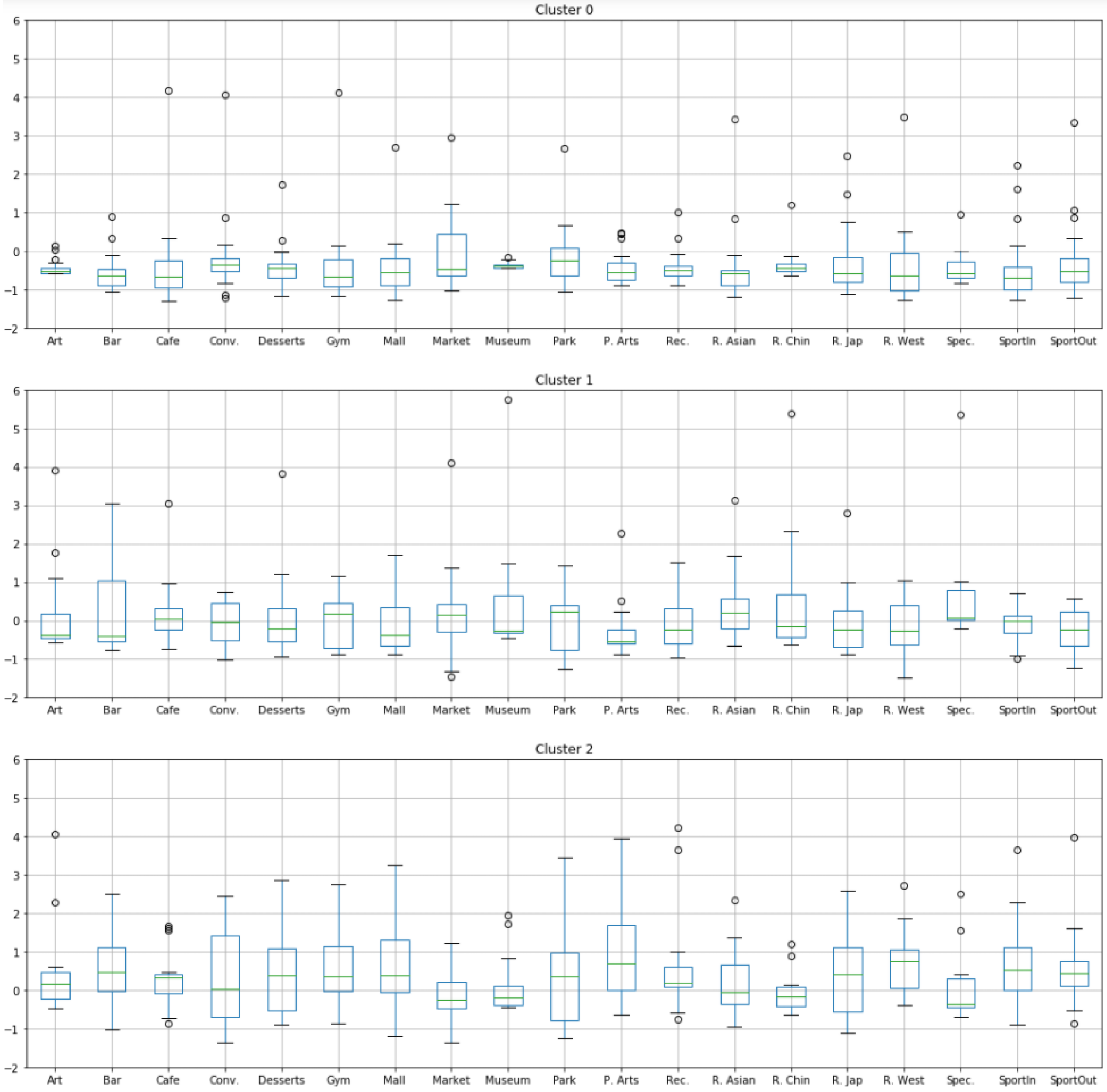
Feature Distributions

- Distributions shown for K-Medoids.
- Reasonably clear patterns distinguishable for each cluster.
- Caveat: Plenty of outliers.



Venue Distributions

- Distributions again shown for K-Medoids.
- Clear distinction between cluster 0 and others.
- Observed distinctions between Clusters 1 and 2 in some intuitive venue types. E.g.
  - Performing arts.
  - Western and Japanese Restaurant.
  - Indoor sports venues.
- Caveats:
  - Again plenty of outliers.
  - Some distinction possibly due to chance.



Conclusions

- Given problems with clustering algorithms in high dimensions with small datasets, surprising that we are able to extract 3 reasonably well defined and intuitive clusters from our mix of socioeconomic indicators.
- Striking how well defined the clusters appear on the geographical map.
- Interesting that these rather loose categories obtained from socioeconomic features seem to be identifiable from Foursquare venue data.

Limitations and further work

- Better data and preliminary analysis could improve selection of starting features for cluster analysis.
- The district level is possibly too broad. Maybe better to use subdistrict data.
- Support intuitions with statistical analysis.
  - More in depth statistical analysis on the observations of venue distributions for identified clusters should be performed.
- Some feature to venue analysis. E.g.
  - How do socioeconomic features relate to numbers of specific venues?
  - Can we predict cluster membership from venue data?
  - Use regression and classification algorithms for example.