CLUSTERING LOCATION HISTOGRAMS

By: Le Yan Koh
Supervisor: Dr Grigorios Loukides

Outline

Introduction
Motivation, Contributions, Domain and Dataset

Background
Data, Clustering Methods, Software Libraries

Approach
K-means clustering, Multidimensional Scaling, Layered Clustering Approach

Results
Silhouette Scoring, Cluster Distribution, Objective Function and Runtime

Conclusion
And Questions

Introduction: Motivation

- Increasing use of Location Sharing Services (LSS) and Location Based Social Networks (LBSN)
  - Facebook Places, Foursquare
- Generating interest in user movement and mobility patterns
- **Uses:** Cost effective planning of urban spaces, prediction of socialisation, collaborative filtering, targeted advertising
- **Needs:** A meaningful way of grouping users via their check-ins

Introduction: Contributions

- Test existing clustering methods on clustering users based on their check-in histograms (location histograms)
- Create and test new, combined clustering methods for clustering users
- Hypothesis: **Users tend to visit similar *types* of spaces, rather than specific locations themselves**

Introduction: Domain and Dataset

- Open dataset on Kaggle: 227,428 check-ins in New York
- User ID, location coordinates, venue category (different taxonomies – on Foursquare developer site)
- For baseline algorithm testing, histograms are represented by matrices of their frequency – Each user has their own histogram
  - 1 row = 1 user
  - 1 column = 1 venue category
  - Column values (x, y) = Frequency the user *y* has visited location *x*

Background: Clustering Techniques

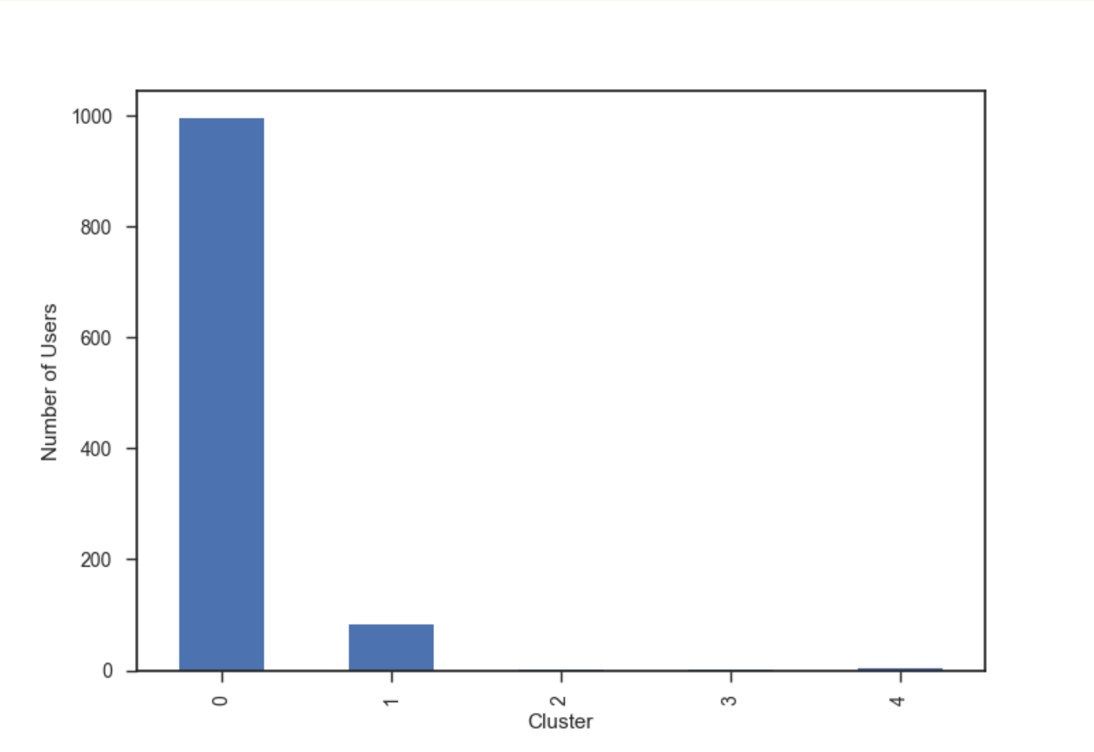
- **Purpose:** Similarity between users are maximised; Given matrix D of M objects and N attributes, find an optimal partitioning of m objects using features described by N attributes.
  - Information retrieval, pattern recognition
- **Typical Steps:** Deciding which attributes are able to distinguish the data the most; Picking measure of similarity (e.g. Euclidean distance), then grouping is done by type of clustering algorithm
- **Types of Clustering Algorithm:** Hierarchical, Partitional, Fuzzy
- **Clustering Algorithms:** *k*-means clustering, manifold learning, spectral clustering, DBSCAN, bi-clustering…
- **Software Libraries:** Scikit-learn (SKLearn)

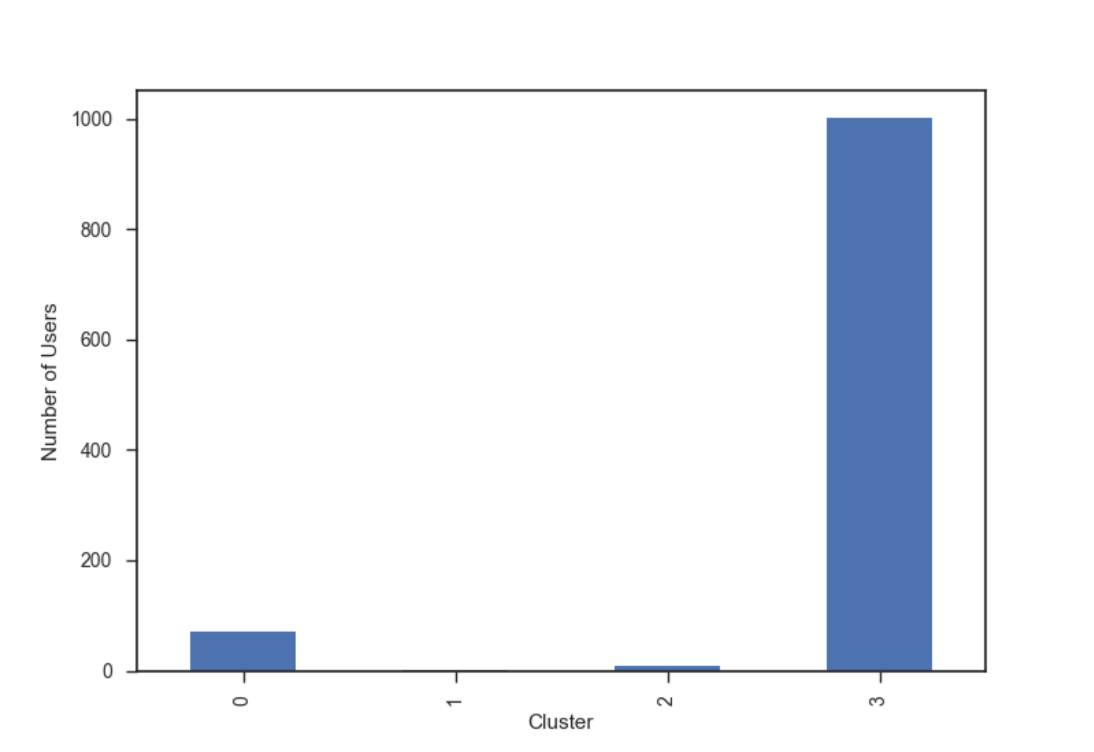
Approach: *k*-means clustering

- Example Representation of Matrix:

User ID	Asian Restaurant	Cafe	Zoo
1	10	5	0
2	5	10	3
3	1	4	10

- 1083 users total, 251 venue categories or 9 venue categories (upper hierarchy)
- Silhouette score used to determine the optimal number of *k* to pick
- Euclidean distance measure used to minimise within cluster sum of squares
- Problems:
  - Curse of dimensionality
  - Extremely high value of objective function
  - Extremely skewed clusters, which does not give meaningful user groupings





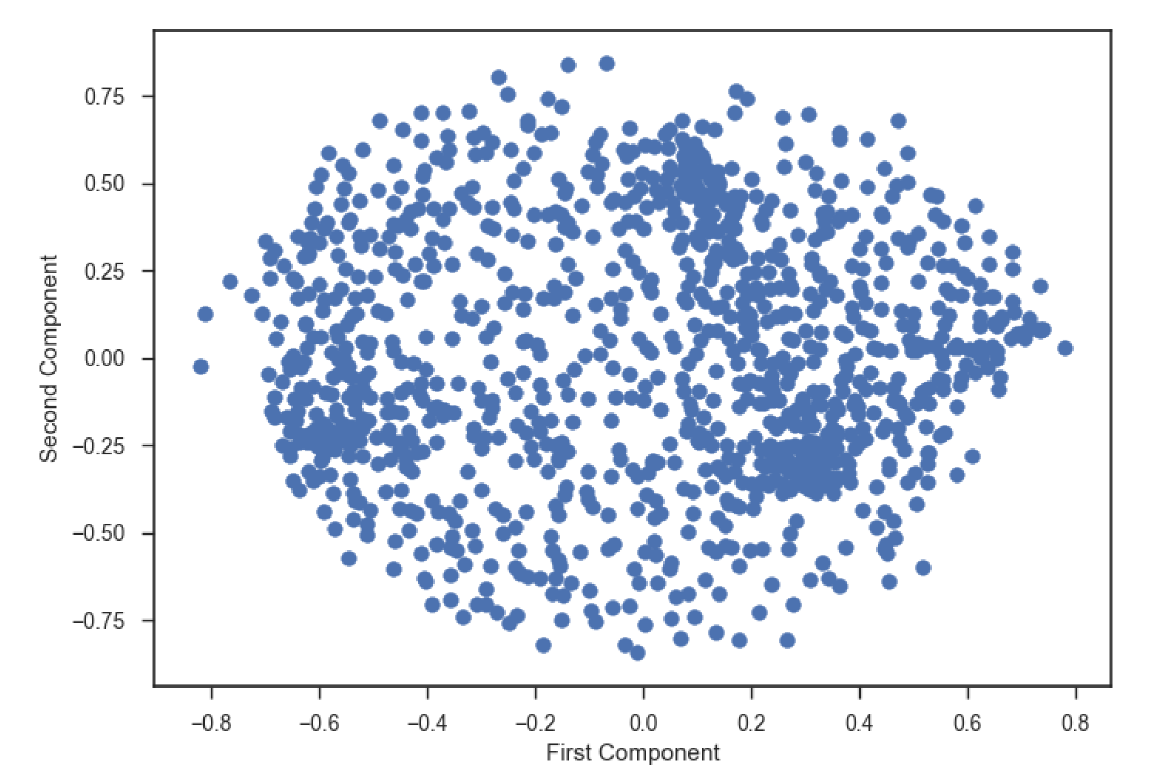
k-means clustering, 251 venue category histograms
K = 5, silhouette score = 0.52

k-means clustering, 9 venue category histograms
K = 4, silhouette score = 0.6

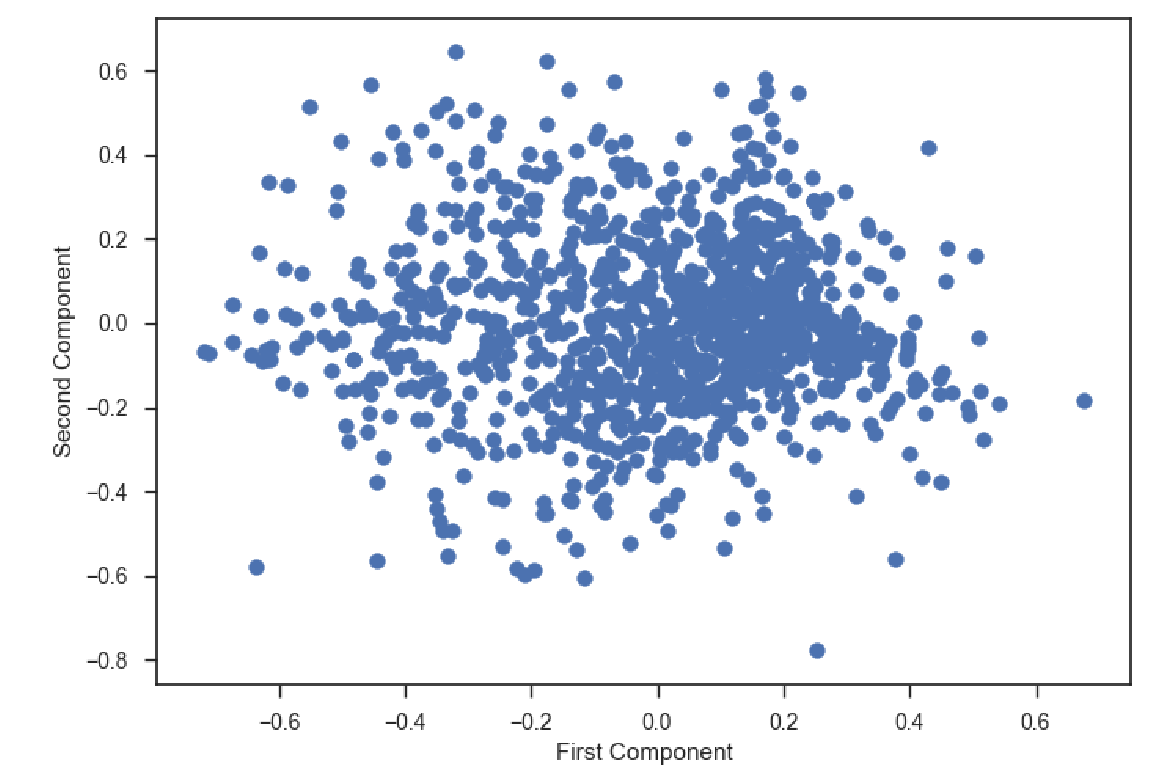
Problem: No specific, well-distributed clusters; Poor user segmentation, perhaps assumptions made about user distribution makes it unsuitable for k-means clustering

Approach: Multi-dimensional Scaling

- Helps to avoid the “curse of dimensionality” by projecting data into lower-dimensional subspace; preserves non-linearity
- 2 dimensions for the ease of visualisation
- Apply k-means clustering after reducing dimensions
- Dissimilarity measure: cosine dissimilarity (also used in high dimensional clustering domains, such as document clustering)



MDS, 251 venue category histograms
Best K = 5, silhouette score = 0.41



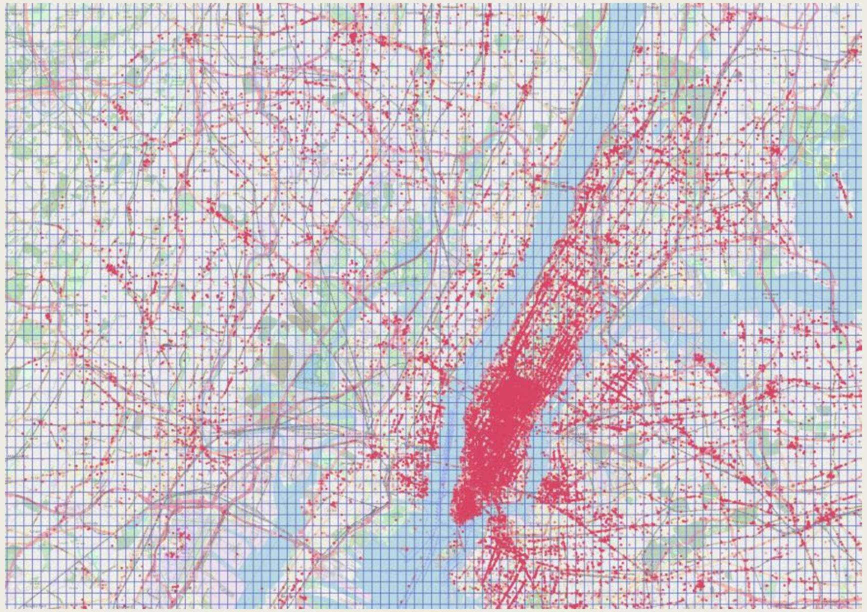
MDS, 9 venue category histograms
Best K = 4, silhouette score = 0.36

Poorer silhouette scores, better user cluster distributions…
But, there might be a better way of segmenting users *semantically*

Approach: Layered Clustering

- **Goal:** Cluster users according to the semantic associations of the land parcels that users are visiting
- **Summary:** Create histograms of users based on *primary semantic value of land parcels they checked into*, rather than specific venue category
- **Steps:**

  1) Data Preprocessing: Divide geographic space into 500m by 500m grids. Create histograms of *user activity* in each grid.



Grid ID	Asian Restaurant	Cinema	Zoo
1	320	5	0
2	5	432	3
3	1	4	763

Each row = 1 Histogram
Each column = Feature/attribute (venue category)

2) Spectral Clustering: Use eigengap heuristic to determine the optimal number of clusters. Apply spectral clustering to the matrix using **optimal clusters.**

3) Generate user histograms from area clusters. Hence, instead of venue categories, *area clusters* will be used as the attribute.

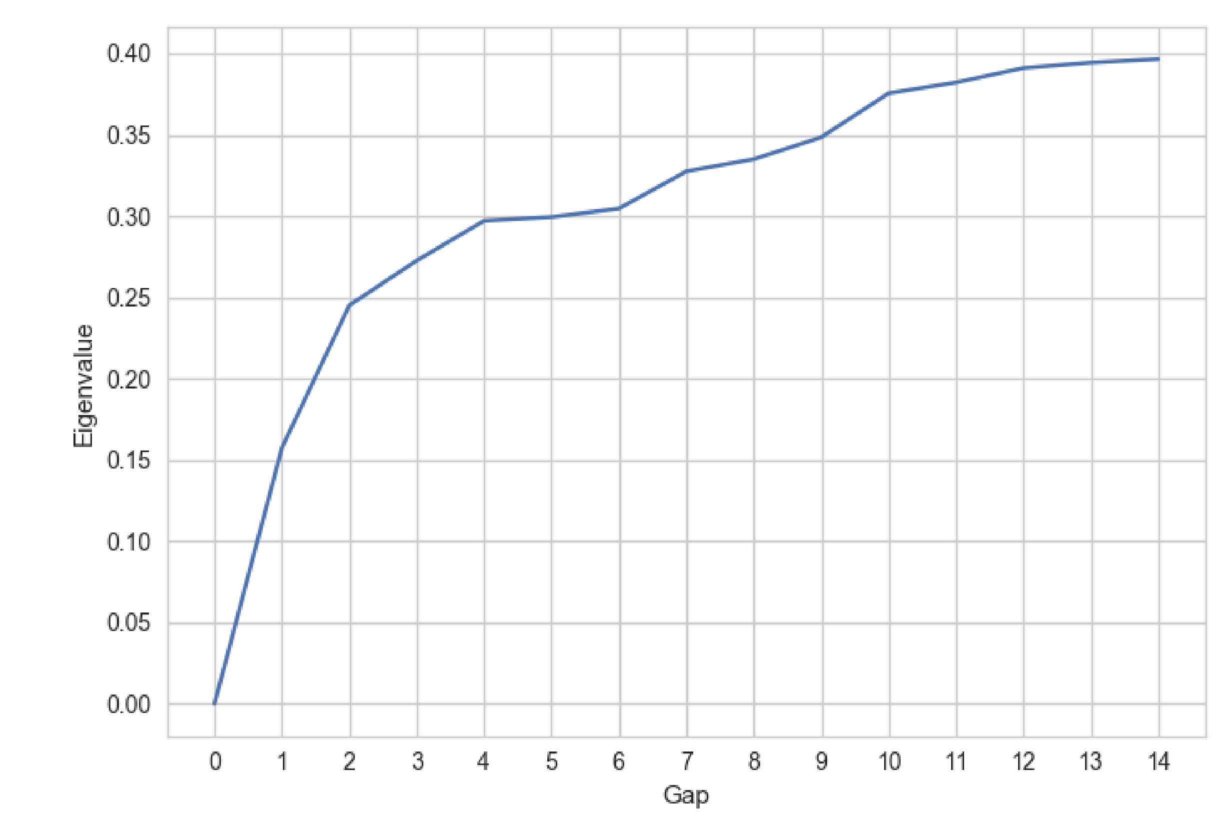
User ID	0	1	2	3	4	5	6	7	8
1	9	56	3	5	17	0	10	1	5
2	43	31	3	30	5	8	6	0	16
3	12	50	7	10	17	3	10	1	4

Example of new user histograms and their clusters

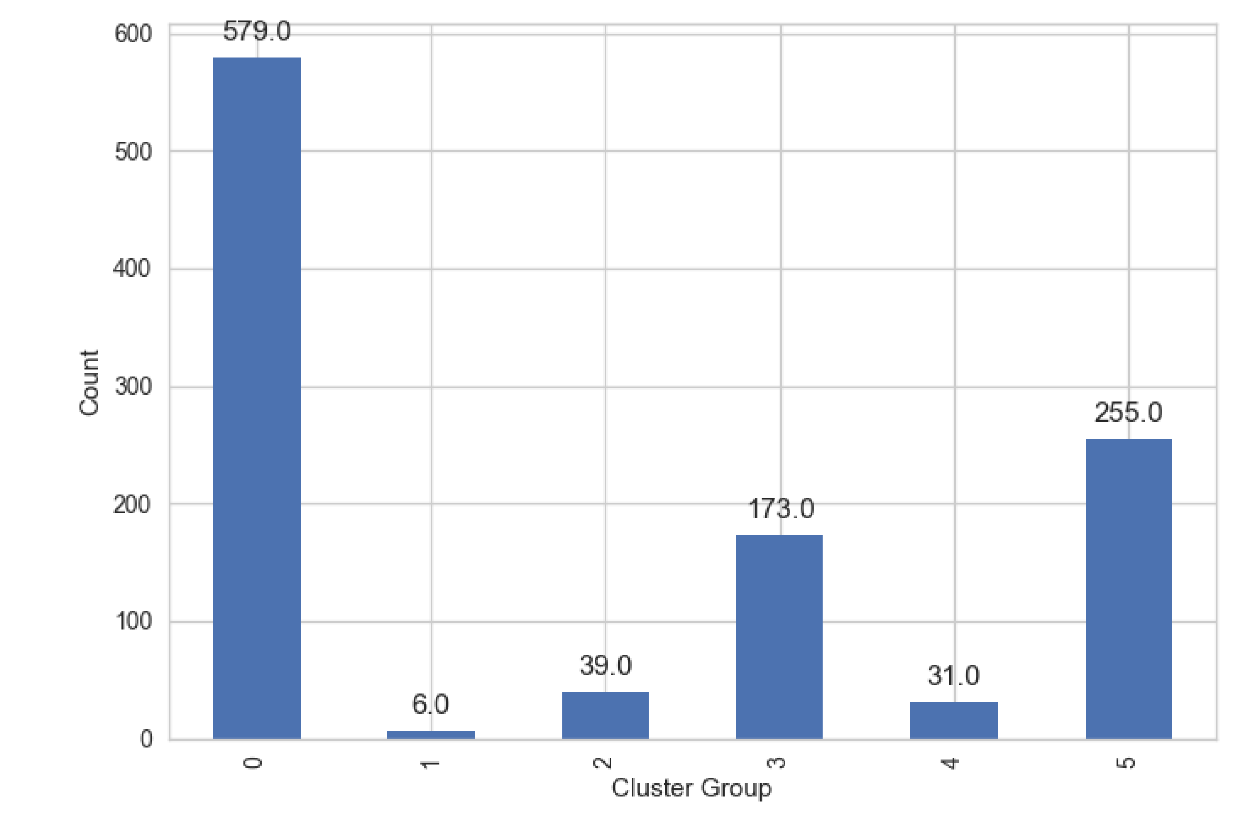
4) Either: Directly apply k-means clustering to the results, or perform multidimensional scaling first (if dimensionality is too high) then apply k-means clustering.

Results: Layered Clustering Method (251 venue categories)

- **Not effective.** Why? Eigengap heuristic was unable to detect the optimal number of clusters for area clustering:

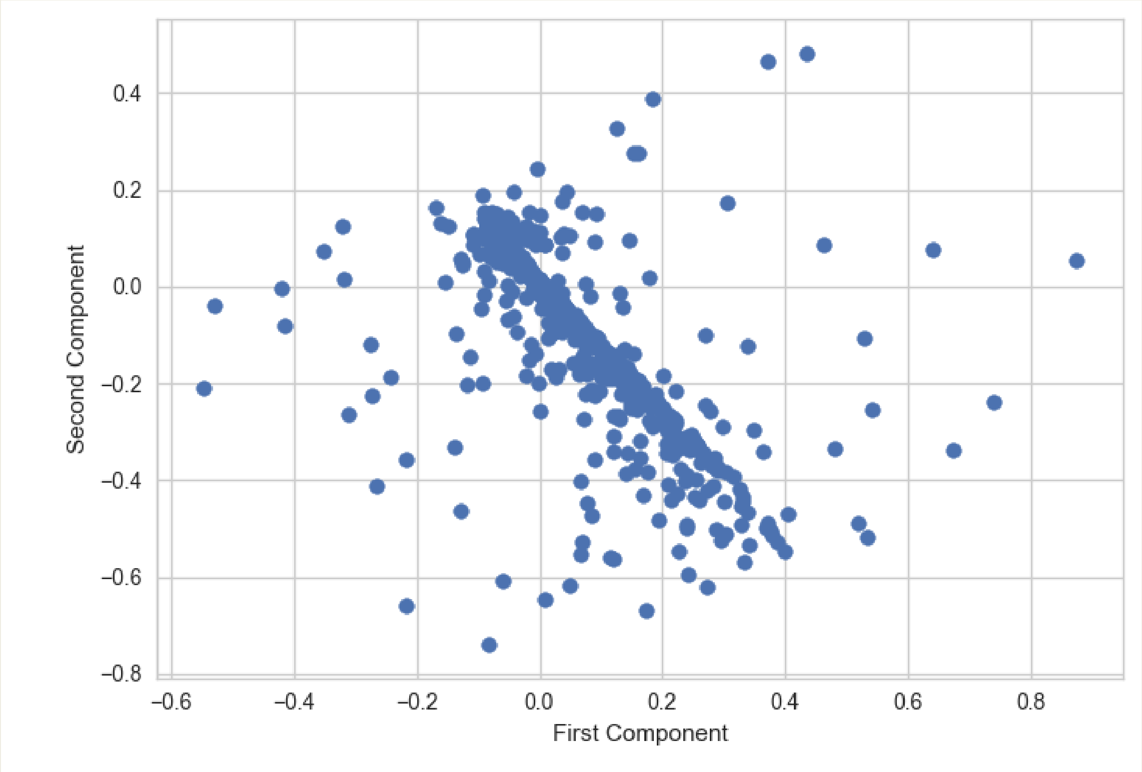


- As a result, applying *k*-means clustering to user histograms generated from these area histograms would have poor returns. For example below: Incredibly skewed **user** clusters if we pressed on with using non-optimal clusters for **area** histograms



- Possible reasons: At this level of granularity, the venue categories are unable to describe splits in areas. A “vegan restaurant vs Greek restaurant” may not produce any meaningful discrimination in semantic area clustering.

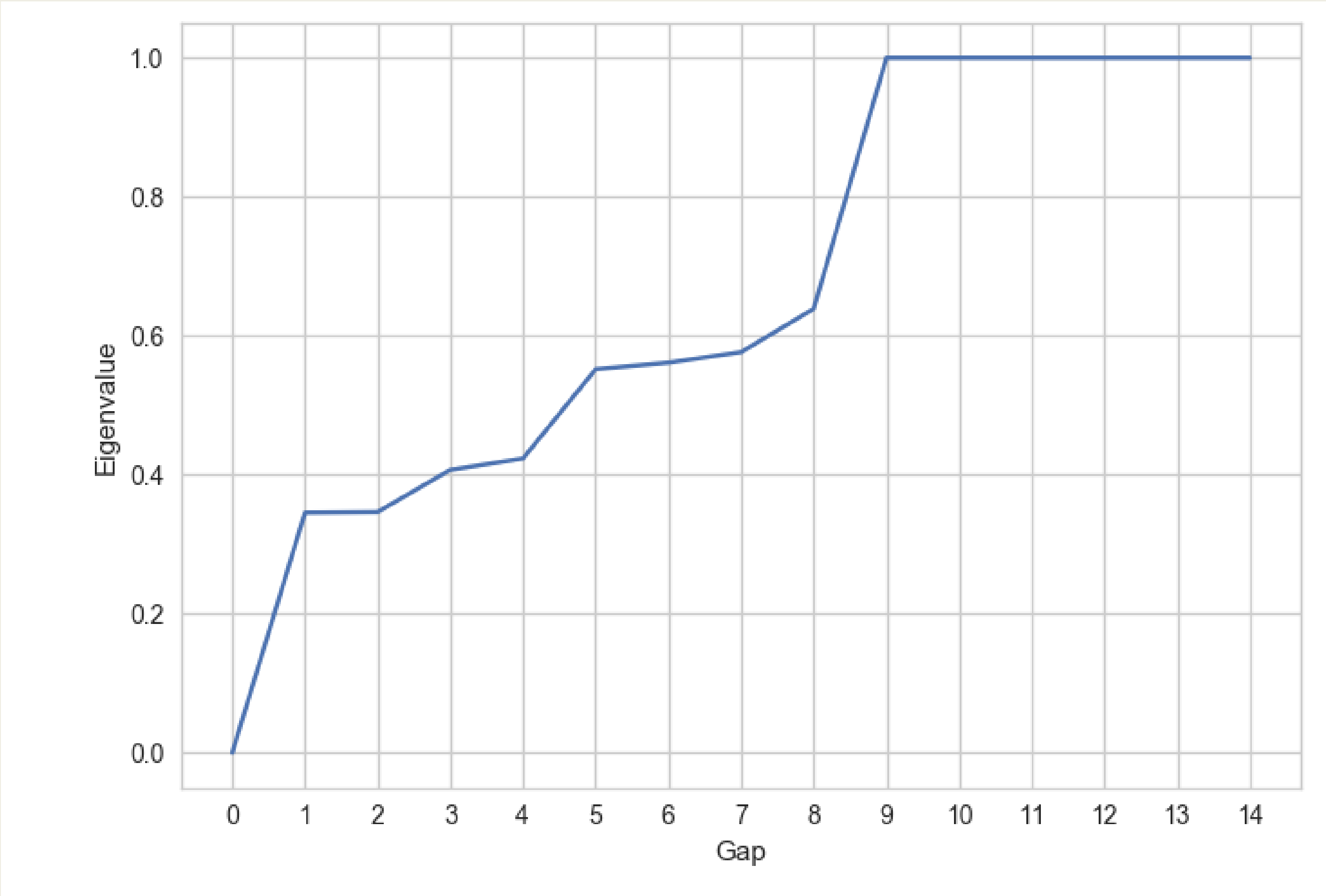
- MDS results are also strange as well:



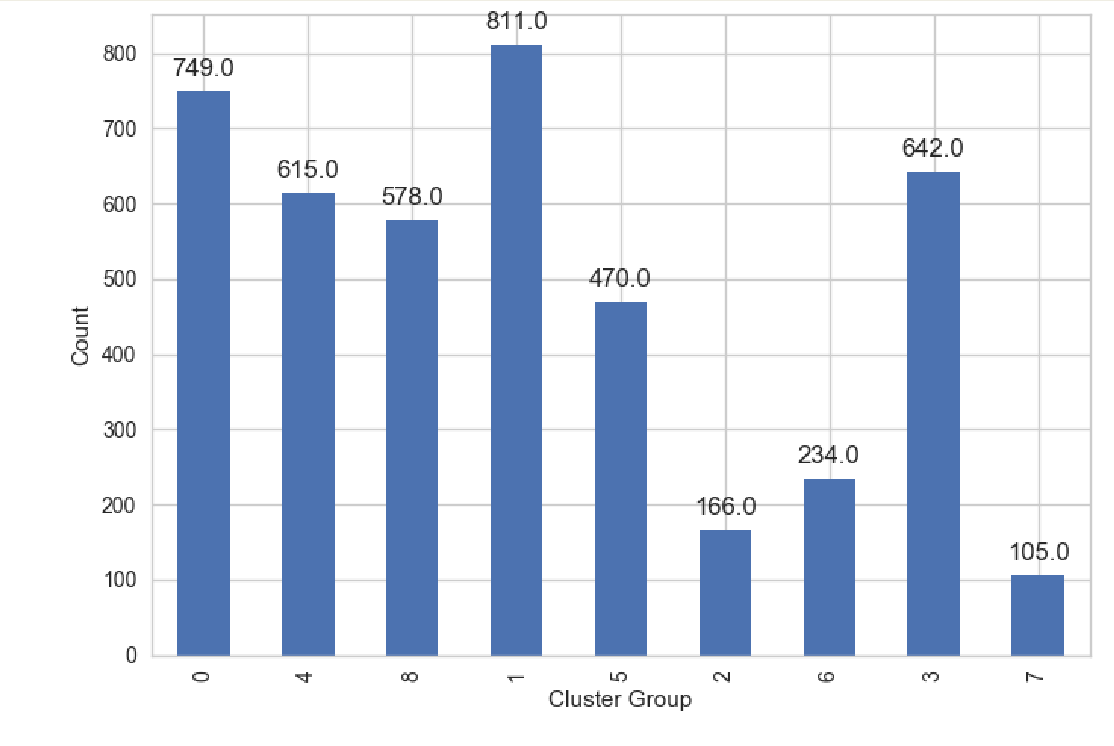
- Users are clustered according to their number of check-ins into **Cluster 3,** because majority of area histograms were divided into that cluster. Hence, this distribution is generally **not that useful**
- **Lesson learnt:** Attributes/venue categories need to be able to meaningfully describe areas, not necessarily a problem with **dimensionality**

Results: Layered Clustering Method (9 venue categories

- Able to find appropriate splits with eigengap heuristic: k = 9



- Able to find adequate distribution and description of each cluster:



Cluster 0
1. Shops & Services (68.6%)
2. Food (12.8%)
3. Professional & Other Places (5.5%)

Cluster 1
1. Food (69.9%)
2. Shop & Services (9.6%)
3. Professional & Other Places (5.1%)

Cluster 2
1. Arts & Entertainment (73.9%)
2. Outdoors & Recreation (5.3%)
3. Food (4.8%)

Cluster 3
1. Travel & Transport (77.8%)
2. Food (6%)
3. Outdoors & Recreation (4.7%)

Cluster 4
1. Professional & Other Places (77.9%)
2. Food (5.6%)
3. Travel & Transport (4.6%)

Cluster 5
1. Residence (82.2%)
2. Food (4.1%)
3. Outdoors & Recreation (3.3%)

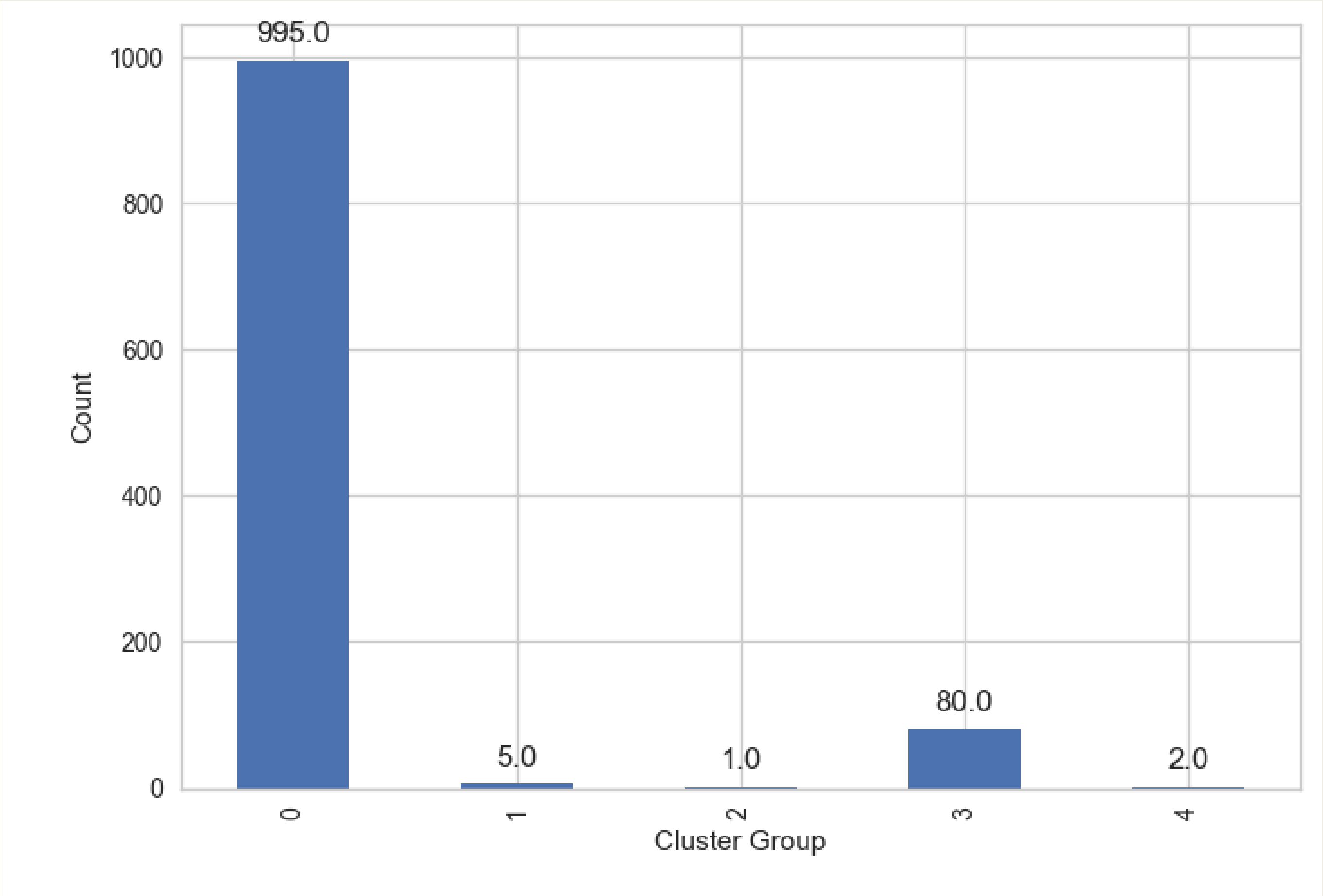
Cluster 6
1. Nightlife Spot (63.9%)
2. Food (12%)
3. Shop & Services (6%)

Cluster 7
1. College & University (74.7%)
2. Professional & Other Places (8.7%)
3. Food (4.2%)

Cluster 8
1. Outdoors & Recreation (85.9%)
2. Food (3.5%)
3. Professional & Other Places (2.6%)

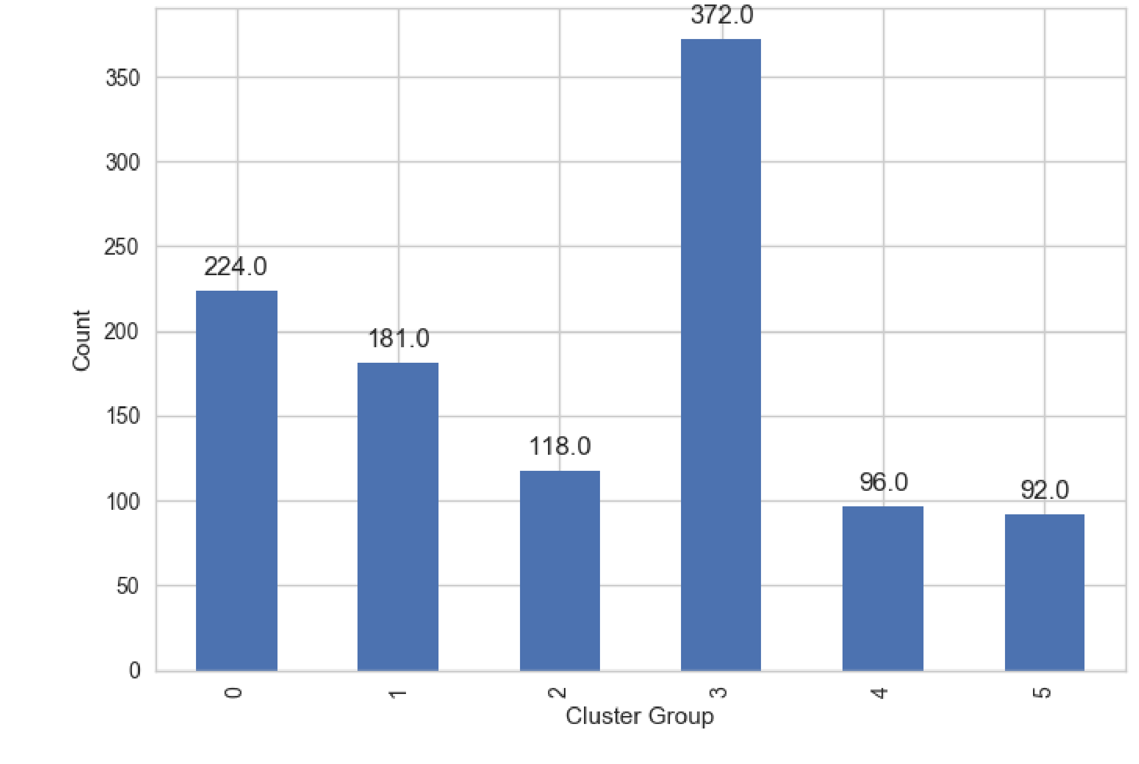
Results: Applying k-means directly after spectral clustering

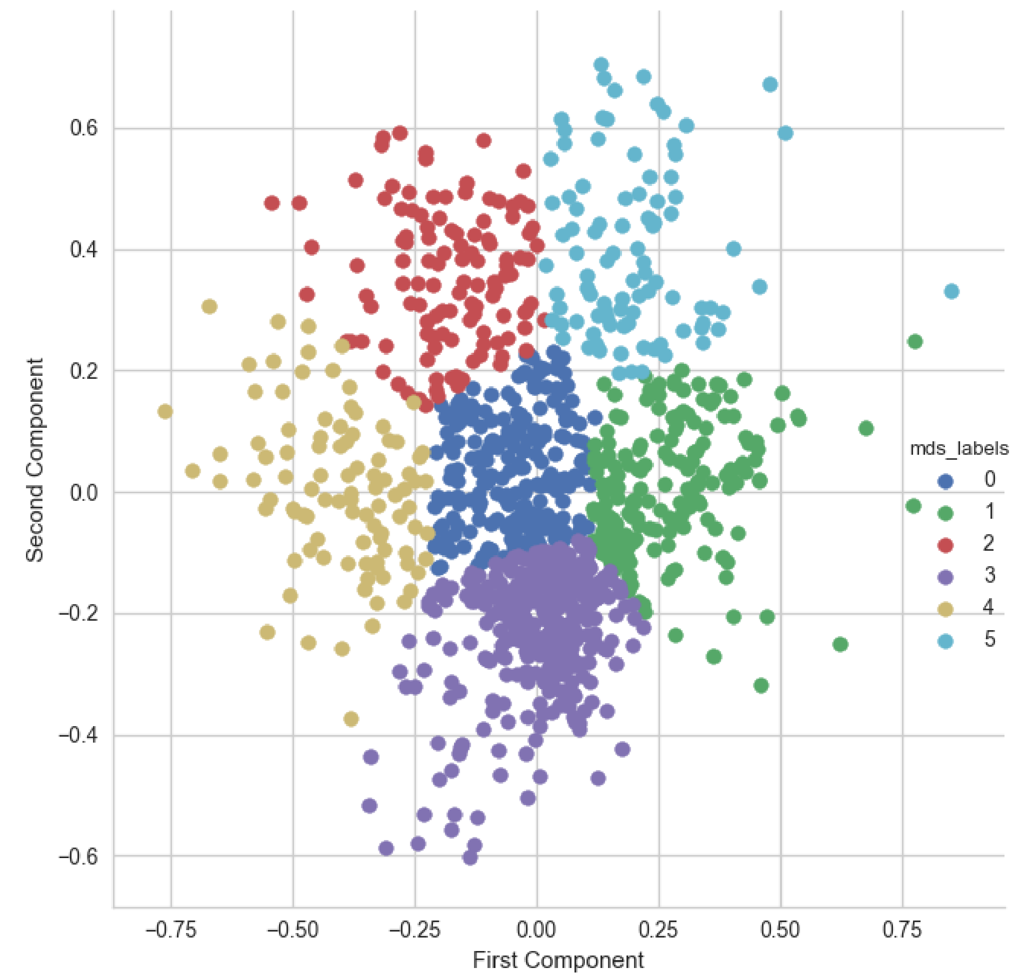
- Optimal silhouette score is 0.59, at $k = 5$
- However…
  - Poor cluster distribution
  - Incredibly high objective function
- Possible reasons:
  - Euclidean space still works poorly in this case
  - Shape of user data is not suitable for *k*-means clustering
- Does not mean this method will be unsuitable for other datasets!



Results: Using MDS before *k*-means clustering

- Better cluster distribution, poorer silhouette coefficient (k=6, coefficient = 0.38)





Cluster Descriptions

User Cluster 0	User Cluster 1	User Cluster 2	User Cluster 3
1. Food (25.8%) 2. Travel & Transport (19.7%) 3. Shops & Services (15.2%)	1. Shops & Services (38.4%) 2. Food (22.1%) 3. Travel & Transport (7.7%)	1. Travel & Transport (38.8%) 2. Residence (12.5%) 3. Professional & Other Places (11.3%)	1. Food (43.5%) 2. Nightlife Spot (16.2%) 3. Shops & Services (12.4%)
User Cluster 4	**User Cluster 5**		
1. Professional & Other Places (41.8%) 2. Food (15.3%) 3. Outdoors & Recreation (10.0%)	1. Residence (35.4%) 2. Shops & Services (18.3%) 3. Food (9.7%)		

Description of user clustering. Dominant categories of each area cluster used instead of cluster name (e.g. cluster 0…8)

Hence…

- We were able to produce user clusters from the semantic associations of locations users visit
  - *E.g. a user visiting a restaurant in a work-dominated area will be classed differently from a user visiting a restaurant in a nightclub-dominated area*

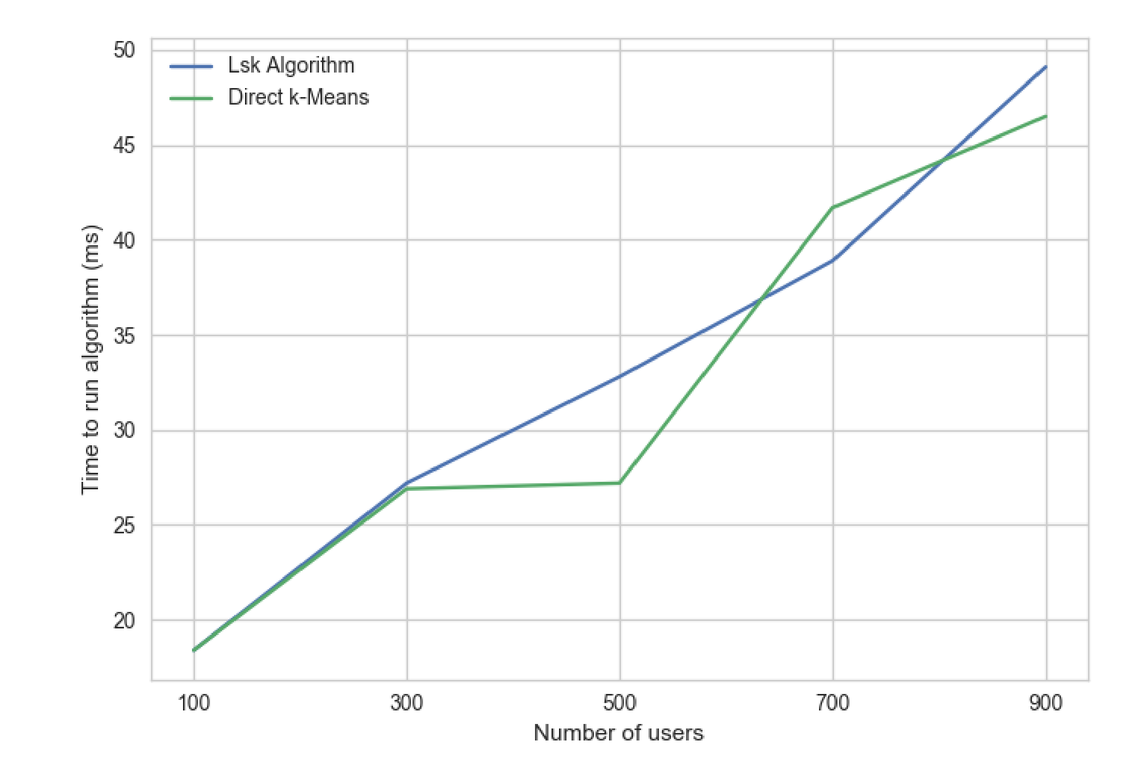
- Venue category descriptions are *important* to describing the space

- Next… run-time of algorithm

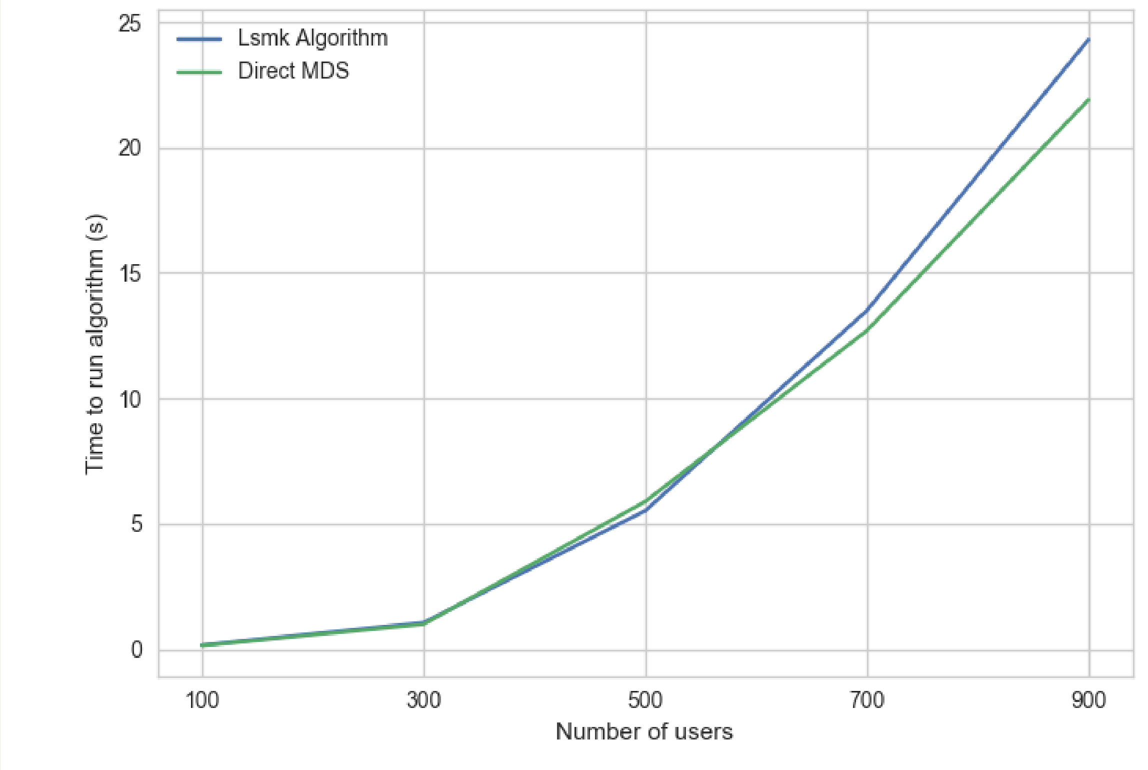
Results: Runtime

- Direct *k*-means – Very fast, varying from **0.267s** (251 dimensions) to **0.048s** (9 dimension)
- Direct *k*-means **after** MDS – Much slower, **43.4s** (251 dimensions) and **36.1s** (9 dimensions)
- Layered clustering – Total time (*k*-means only)
  - 251 Dimensions: **3.87s**
  - 9 Dimensions: **3.66s**
- Layered clustering – Total time (MDS + *k*-means)
  - 251 Dimensions: **20.23s**
  - 9 Dimensions: **39.31s**

Results: Runtime with increasing data size



Layered clustering algorithm: *k-means only comparison*

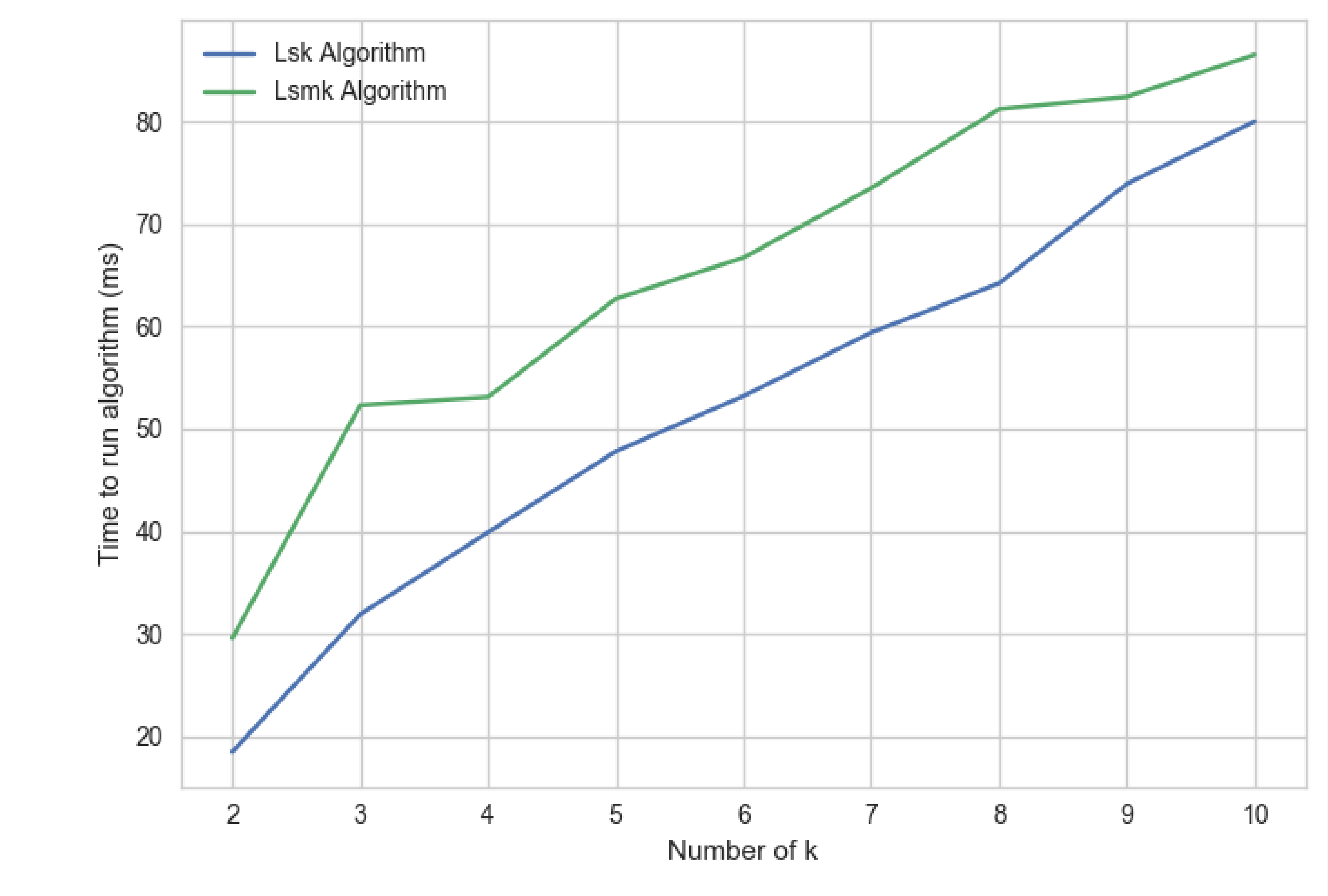


Layered clustering algorithm: *k-means with MDS comparison*

For most part: the new algorithm performs *slightly* slower. However, we could argue that the new algorithm also produces new results that might be more meaningful – Possibly worth the trade-off.

Results: Runtime with increasing *k*

- Comparison of 9 venue categories only (since this was the one that produced the most effective results)
- Does not count spectral clustering step, only second layer of clustering (user histogram clustering)
- As expected, using MDS in the algorithm makes it perform slower – but it is more effective
- Runtime increases with increasing amount of *k*



Conclusion

- **Contributions:** Devising and testing a new algorithm that can better separate users based on geographical principles; Qualitatively and quantitatively assess clusters given from baseline algorithms and new algorithm
- **Future work:**
  - *Testing on other datasets;*
  - *Weighing area histogram check-ins to improve accuracy of area semantics (to account for locations that are less checked-into but of some importance).*
  - *Optimise algorithm*
- **Lessons learnt:**
  - *Algorithm may not necessarily be good for online learning and using as of yet*
  - *However, poor results does not mean the algorithm performed poorly – but the dataset may not fit the method*

QUESTIONS AND DEMONSTRATION