North versus South: What makes a student city?

Joshua Weston

3rd August 2020

Abstract

Using the Foursquare API and the k-means++ algorithm we attempt to cluster student cities to verify whether geographic region is a good indicator of venue population according to perceived differences in student life between North England and South England. In this study we discuss the problem of the north-south divide among students, outline our dataset and methodology and include a brief summary on the k-means and k-means++ algorithms. We find over a range of clusters that via a combination of the model inertia and the silhouette score a model containing 10 clusters is an optimal fit, indicating a lowly-clustered, highly similar data population. From this we conclude that the idea that a student city will have differing venue populations based on its region, particularly north versus south, to be incorrect.

1.0 Introduction

"Student cities" are the communities surrounding universities or colleges dominated by a young, twenty-something student population. These communities are built on a diverse population hailing from different countries or even regions in the United Kingdom, which has in the past led to many heated debates regarding the 'North versus South' rivalry between English students [1]. Sides have often been found to accuse the other of harboring "dead" student cities that offer little in the way of nightlife or social activities.

To validate or invalidate these claims we require research on the makeup of the business communities in these student cities. By sourcing location data from several universities in the north, south, and midlands of England, as well as London, we can determine the most frequently established venues by city. A k-means clustering algorithm will attempt to group these communities into four different clusters based on these findings – if indeed there are differences in student cities based on geographic position, we should find that the clusters will tend to group geographically.

The findings will be of use to those looking to establish a business that will cater to the student community – what do students look for, and where do they look for it? It will also be of use to students who wish to look for universities with student communities that match their own interests.

2.0 Methodology

2.1 Data Origins

Our data is sourced using the Foursquare API [2] in a Python conda environment. For each university we take the latitude and longitude of the campus and calculate eight surrounding latitudes and longitudes (north, south, east, west, northwest, northeast, southwest and southeast) each a distance of 3.0 km from the origin. The reason for this is that the Foursquare API will only return 100 venue results for any given point. In order to cover a reliably large area around the campus we must feed multiple points into the search to obtain a full list of venues. Repeating venues can be dropped. Our search for each of these points will cover a circle of radius 3.0 km; thus, our maximum distance from the campus will be 6.0 km; a reasonable distance that would take approximately an hour for a student to walk.

The result then is a complete list of venues within the searched area (we know this to be a complete list as no result returns 800 venues, i.e. the search does not return the maximum number of venues and so no venues are dropped). This is fed into a Pandas data frame containing the labelled results for thirty different universities. This dataframe is one-hot encoded and a fractional proportion of the venues are returned for each city for normalization purposes.

The complete set of data is then two dataframes: one, containing the universities and their respective top ten most common venue types. The second is an unlabelled dataframe containing the fractional proportion of these venues for each university with the campuses unlabelled. The latter will be fed into the k-means clustering model and the former will be used by us to analyse the results of the fit.

2.2 K-Means Clustering

While a classification method may be suitable for determining the region an anonymous university may belong to it carries the caveat that we must be first able to define what defines the universities of each region. With no clear definition based on venue categories we instead use a clustering method.

K-Means clustering is a machine learning method of partitioning objects or observations into K clusters based on their closest cluster centroid, i.e. a method of quantizing a data space into Voronoi cells and determining which cells each data point belong to [3][4][5]. The standard algorithm (or Lloyd's Algorithm) uses an iterative process to refine the cluster centroids. First, mean points of each cluster as centroids. Each observation is assigned to a cluster based on which centroid is closest, typically (and in our case) the centroid with the least squared Euclidean distance. The centroids are then recalculated based on the observations assigned to each cluster. This process continues until the algorithm has converged and the cluster assignments are unchanged. These centroids will have minimal intra-class variances (the sum of the squared Euclidean distances from each point in the cluster to the centroid).

A primary issue of the algorithm is that the convergence can be poor in its accuracy. Many problems are found to have unbounded solutions, convergence to local minima or multiple centroids in the same cluster. The k-means++ algorithm designed by Arthur and Vassilvitskii [6][7] aims to tackle this issue through initialization, by choosing random centroids with specific probabilities. Random centroids are selected from the data points and their distances to each data point are computed. In the next iteration centroids are selected such that the probability of selecting a given data point as a centroid is directly proportional to its distance from the closest previously selected centroid (such that the most distance points are more likely to be selected). The result is centroids are less likely to be selected in the same cluster and all centroids are associated with data points by the end of the process.

The Python 2.7.16 module Scikit-learn (also known as sklearn) contains k-means clustering functions that can be used to run the k-means++ algorithm. It is additionally capable of providing metrics with which to explore the accuracy of our clustering model.

2.3 Clustering Student Cities

We select thirty universities at random from the top 100 universities in England: eight from The North, nine from The South, five from London, and eight from The Midlands (see Table 1). As outlined in 2.0 the coordinates of each campus are acquired before eight surrounding points are generated at a distance of 3.0 km. For each point we use the Foursquare API to obtain a list of surrounding venues within a 3.0 km radius, primarily venue names and the venue categories (e.g. Pub, Grocery Store, Park, etc.). These lists are concatenated into a single dataframe for each university. Each venue has a unique ID that allows us to drop additional instances of a given establishment to prevent venues on the boundary between overlapping points from being overweighed. Venue names are dropped and establishments are one-hot encoded by venue category for ease of analysis. The

REGION	UNIVERSITY	LATITUDE	LONGITUDE
South England	University of Oxford	51.7548	-1.2544
South England	University of Southampton	50.9351	-1.3958
South England	University of Sussex	50.8671	-0.0879
South England	University of Reading	51.4414	-0.9418
South England	University of Cambridge	52.2043	0.1149
South England	University of East Anglia	52.6219	1.2392
South England	University of Bristol	51.4584	-2.603
South England	University of Exeter	50.7371	-3.5351
South England	University of Bath	51.3782	-2.3264
Midlands	University of Warwick	52.3793	-1.5615
Midlands	University of Birmingham	52.4508	-1.9305
Midlands	Keele University	53.0034	-2.2721
Midlands	Coventry University	52.4072	-1.5037
Midlands	University of Nottingham	52.9386	-1.1952
Midlands	University of Leicester	52.6211	-1.1246
Midlands	Loughborough University	52.7651	-1.2321
Midlands	Nottingham Trent University	52.9581	-1.154
London	Imperial College London	51.4988	-0.1749
London	University College London	51.5246	-0.134
London	King's College London	51.5115	-0.116
London	London School of Economics and	51.5144	-0.1165
London	Queen Mary, University of Lond	51.5241	-0.0404
North England	University of Manchester	53.4668	-2.2339
North England	University of Liverpool	53.4048	-2.9653
North England	Lancaster University	54.0104	-2.7877
North England	University of Sheffield	53.3809	-1.4879
North England	University of Leeds	53.8067	-1.555
North England	University of York	53.9461	-1.0518
North England	Durham University	54.765	-1.5782
North England	Newcastle University	54.9792	-1.6147

Table 1. *(above) Universities selected for the study with their respective region and coordinates.*

establishments are then grouped into a single row by their mean fractional occurrence in the venue population.

Two dataframes are then returned for analysis, as explained above – the one hot encoded dataframe containing the venue category means for each university, and the dataframe containing the ten most common venue categories and their total population for each university (see Tables 3 and 4 in the appendices). The former we continue with for computational analysis via the k-means++ algorithm.

2.4 The Optimal Model

Rather than set our k-means++ model to fit four clusters as we expect it is best to instead attempt to model with a varying number of clusters and find the optimal fit. We attempt the model for a cluster number k ranging from 2 to 12 and select the optimal model based on two criteria.

The first criteria, the within-cluster sum-of-squared errors or inertia, is the sum of the square of the distance of each data point from its cluster centroid. When plotted against k the result is an elbow-like lineshape. Ideally the best k is one for which the inertia is at a minimum, or zero, i.e. at the vertex of the elbow [8]. Several issues arise when using inertia as the sole metric for optimizing a model. First, it responds poorly to clusters with non-convex, anisotropic shapes or irregular manifolds. It is not normalized, which can result in inflated distances in higher-dimension clusters. Finally, in many datasets the elbow curve may have multiple vertexes that can be interpreted as candidates for the optimal model k.

It is best to combine the results of the above 'elbow method' with an additional criterion known as the silhouette score [9][10]. The silhouette score is a normalized metric that measures the similarity of a data point to its assigned cluster compared to its similarity to others. A silhouette coefficient of +1 implies heavy similarities to the assigned cluster, while -1 implies heavy dissimilarity. A coefficient of zero indicates that the data point is close to the decision boundary of multiple clusters. The silhouette score gives an average value for every data point. Where a is the mean intra-cluster distance and b is the mean inter-cluster distance the silhouette score is given as:

$$\frac{b-a}{\max(a,b)} \quad (1)$$

Thus, when plotted against the number of clusters k, we observe the optimal k to correspond to the maxima of the curve. In an ideal model the maximum of the silhouette curve corresponds to the minimum of the elbow curve.

An additional method for verification may be gap statistics, which employs a comparison of cluster inertia to the inertia in a reference data distribution [11]. Three verification methods may be redundant, and as seen later we do not believe employing the latter technique would be a constructive use of time.

3.0 Results

We find an optimal k-means++ model where k=10, with a high silhouette score of 0.883 and low inertia of 0.145 indicating that this is a good cluster fit to the dataset (see Table 2, Figure A). The model seems unable to cluster the student cities into any distinguishable groups, with multiple clusters only containing a single data point (Table 5). It is unclear as to why this is. A potential reason is the large proportion of pubs and cafes in the complete venue population.

A low inertia and high silhouette score for 10 clusters indicates that this is the optimal k-means model for the dataset, and with a mean population of 3 universities per cluster we find that the population isn't particularly clustered. Analysing the dataset for other values of k however we find interesting results on the map. The population seems to contain a large cluster running from London to Birmingham that remains mostly consistent for all values of k. It may be that there are additional unknown variables that influence the venue population in this region.

Due to the implied presence of these additional variables, the low cluster population and the lack of expected cluster behaviour we are unable to find evidence that the university population clusters to a north-south divide.

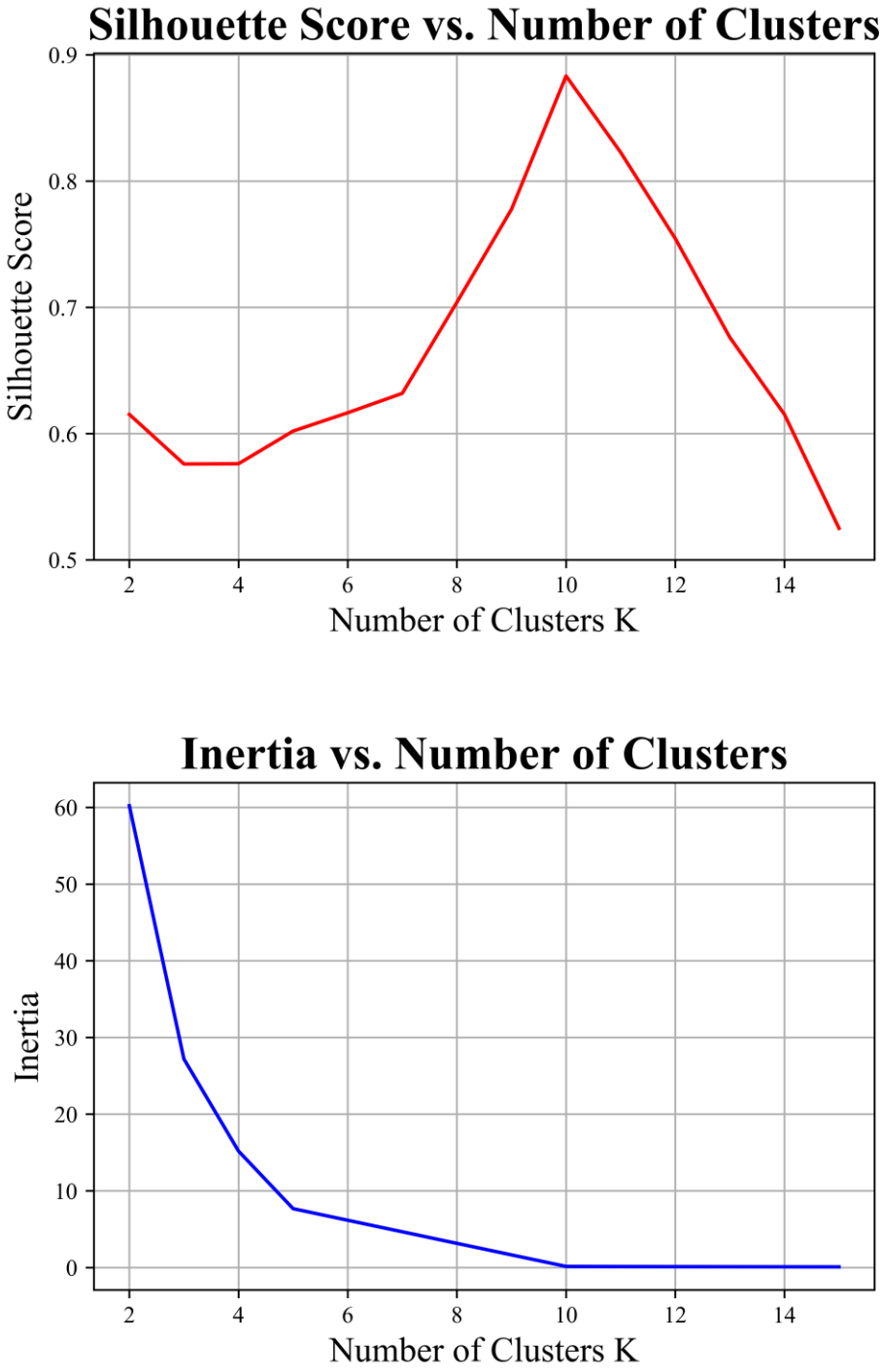

Figure A.1 *(above) Silhouette score and (below) inertia for a range of cluster values K.*

K	INERTIA	SILHOUETTE SCORE
2	60.21	0.615
3	27.19	0.576
4	15.18	0.576
5	7.68	0.602
6	6.17	0.617
7	4.66	0.632
8	3.15	0.704
9	1.65	0.778
10	0.145	0.883
11	0.13	0.823
12	0.117	0.755
13	0.107	0.676
14	0.0947	0.615
15	0.0852	0.523

Table 2. *(above) Cluster number K and their respective inertia and silhouette score.*

We have neglected to include gap statistic analysis due to the above-mentioned highly similar clusters, high silhouette score and low inertia value. While additional analysis may refine the model the model itself does not present worthwhile results that would benefit from refinement.

4.0 Conclusions

Using the Foursquare API and the k-means++ algorithm we attempted to cluster student cities to verify whether geographic region is a good indicator of venue population. We found via a combination of the elbow method and the silhouette score that a model containing 10 clusters was an optimal fit, which indicated a lowly-clustered, highly similar data population. When plotted with cluster labels on a map of England we find that there is no correlation between region and cluster label. We do however find a promising cluster that runs from London to Lancaster that reoccurs for models with various k-values, indicating that venue population is influenced by some form of geographic position. Nonetheless we must conclude that the idea that a student city will have differing venue populations based on its region, particularly north versus south, to be incorrect.

The study may be improved by the inclusion of additional venue statistics such as venue rating or age of venue to better indicate the popularity of the venue, or a larger number of universities with which to analyse. Gap statistics may also be employed to better narrow down the correct number of clusters.

References

[1] Me

[2] https://developer.foursquare.com/

[3] Lloyd S., 1982, IEEE Trans. Inf. Theory, 28

[4] MacKay D., Information Theory, Inference and Learning Algorithms, 2003, Ch.20, 284, 9780521642989

[5] Hartigan J., Wong M., 1979, JRSS, 28,100-108

[6] Arthur D., Vassilvitskii S., 2007, SODA ’07, 1027-1035

[7] Solis-Oba R., 2006, LNCS, 3483, 292-230

[8] Thorndike R., 1953, Psychometrika, 18, 267-276

[9] Rousseeuw P., 1987, JCAM, 20, 53-65

[10] Kaufman L., Rousseeuw P., 1990, Finding Groups in Data: An Introduction to Cluster Analysis, 9780471878766

[11] Tibshirani R., Walther G., Hastie T., 2001, JRSS, 63, 411-423

APPENDICES

Appendix A: Tables

CITY	MOST COMMON VENUE			
	1st	2nd	3rd	4th
University of Oxford	Pub	Hotel	Grocery Store	Coffee Shop
University of Southampton	Grocery Store	Pub	Coffee Shop	Supermarket
University of Sussex	Pub	Café	Coffee Shop	Park
University of Reading	Pub	Grocery Store	Coffee Shop	Hotel
University of Cambridge	Pub	Coffee Shop	Café	Grocery Store
University of East Anglia	Pub	Coffee Shop	Grocery Store	Café
University of Bristol	Pub	Café	Bar	Park
University of Exeter	Pub	Coffee Shop	Café	Supermarket
University of Bath	Pub	Hotel	Café	Gastropub
University of Warwick	Pub	Coffee Shop	Café	Indian Restaurant
University of Birmingham	Pub	Coffee Shop	Indian Restaurant	Grocery Store
Keele University	Pub	Coffee Shop	Grocery Store	Supermarket
Coventry University	Pub	Coffee Shop	Supermarket	Fast Food Restaurant
University of Nottingham	Pub	Grocery Store	Coffee Shop	Bar
University of Leicester	Pub	Grocery Store	Coffee Shop	Park
Loughborough University	Pub	Hotel	Grocery Store	Gym / Fitness Center
Nottingham Trent University	Pub	Grocery Store	Bar	Coffee Shop
Imperial College London	Pub	Café	Park	Hotel
University College London	Hotel	Pub	Coffee Shop	Café
King's College London	Hotel	Coffee Shop	Pub	Theater
London School of…	Hotel	Coffee Shop	Pub	Café
Queen Mary, University of London	Pub	Coffee Shop	Café	Park
University of Manchester	Pub	Bar	Coffee Shop	Café
University of Liverpool	Pub	Coffee Shop	Grocery Store	Bar
Lancaster University	Pub	Coffee Shop	Hotel	Sandwich Place
University of Sheffield	Pub	Café	Coffee Shop	Grocery Store
University of Leeds	Pub	Bar	Coffee Shop	Supermarket
University of York	Pub	Café	Hotel	Grocery Store
Durham University	Pub	Coffee Shop	Hotel	Grocery Store
Newcastle University	Pub	Supermarket	Coffee Shop	Grocery Store

5th	6th	7th	8th	9th	10th
Café	Park	Supermarket	Chinese Restaurant	Bakery	Restaurant
Fast Food Restaurant	Park	Pizza Place	Hotel	Bar	Italian Restaurant
Grocery Store	Gym / Fitness Center	Pizza Place	Gastropub	Indian Restaurant	Bakery
Supermarket	Gastropub	Clothing Store	Café	Gym / Fitness Center	Park
Hotel	Restaurant	Gastropub	Burger Joint	Supermarket	Gym / Fitness Center
Hotel	Supermarket	Bar	Fast Food Restaurant	Park	Gym / Fitness Center
Coffee Shop	Grocery Store	Restaurant	Burger Joint	Indian Restaurant	Pizza Place
Hotel	Bar	Furniture / Home Store	Grocery Store	Bakery	Bookstore
Grocery Store	Park	Coffee Shop	Bar	Supermarket	Restaurant
Hotel	Supermarket	Grocery Store	Park	Sandwich Place	Restaurant
Park	Bar	Café	Sandwich Place	Supermarket	Fast Food Restaurant
Bar	Sandwich Place	Bus Stop	Clothing Store	Fast Food Restaurant	Furniture / Home Store
Indian Restaurant	Hotel	Grocery Store	Café	Sandwich Place	Convenience Store
Supermarket	Café	Gym / Fitness Center	Indian Restaurant	Fast Food Restaurant	Park
Supermarket	Indian Restaurant	Pizza Place	Bar	Fast Food Restaurant	Clothing Store
Bar	Gastropub	Park	Café	Pizza Place	Coffee Shop
Café	Supermarket	Indian Restaurant	Park	Fast Food Restaurant	Gym / Fitness Center
Coffee Shop	Bakery	French Restaurant	Italian Restaurant	Pizza Place	Gastropub
Bakery	Theater	Park	Ice Cream Shop	Bookstore	Japanese Restaurant
Café	Park	Bookstore	Pizza Place	Art Gallery	Beer Bar
Theater	Park	Bookstore	Bakery	Pizza Place	Art Gallery
Hotel	Cocktail Bar	Bar	Gym / Fitness Center	Bakery	Italian Restaurant
Indian Restaurant	Park	Hotel	Italian Restaurant	Grocery Store	Middle Eastern Restaurant
Fast Food Restaurant	Café	Discount Store	Hotel	Park	Restaurant
Supermarket	Indian Restaurant	Pizza Place	Harbor / Marina	Café	Clothing Store
Park	Bar	Supermarket	Pizza Place	Gym / Fitness Center	Indian Restaurant
Café	Grocery Store	Gym / Fitness Center	Thai Restaurant	Italian Restaurant	Pizza Place
Coffee Shop	Bar	Historic Site	Bus Stop	Italian Restaurant	Sandwich Place
Café	Supermarket	Italian Restaurant	Bar	Asian Restaurant	Restaurant
Café	Hotel	Bar	Italian Restaurant	Pizza Place	Indian Restaurant

Table 3. *(Above) Most common venues for the universities examined in this study.*

CITY	MOST COMMON VENUE COUNT										TOTAL VENUES
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	
University of Oxford	42	15	13	11	11	8	8	7	7	7	279
University of Southampton	41	35	22	13	12	10	9	8	7	7	314
University of Sussex	15	13	13	9	5	4	3	3	3	3	139
University of Reading	35	16	14	12	12	10	8	8	7	7	290
University of Cambridge	39	13	12	11	10	8	7	6	6	6	238
University of East Anglia	36	11	9	9	7	7	6	6	5	4	195
University of Bristol	49	21	19	18	17	11	8	7	7	7	317
University of Exeter	21	11	8	6	6	4	3	3	3	3	137
University of Bath	17	15	10	9	9	7	5	5	4	4	179
University of Warwick	26	15	13	11	9	8	7	7	6	4	209
University of Birmingham	36	26	22	18	17	15	12	11	11	11	371
Keele University	14	11	11	9	5	4	4	4	4	4	132
Coventry University	28	16	16	12	12	12	10	9	9	8	261
University of Nottingham	42	22	21	16	10	9	9	9	8	8	334
University of Leicester	23	19	14	13	13	11	8	8	6	6	258
Loughborough University	13	8	6	3	3	3	3	3	3	3	86
Nottingham Trent University	37	18	16	16	13	11	10	10	8	7	312
Imperial College London	32	28	22	22	20	17	16	12	12	12	606
University College London	31	27	21	18	16	14	13	13	11	11	585
King's College London	35	33	24	16	15	15	13	11	10	10	552
London School of…	32	30	27	14	14	14	13	13	11	11	561
Queen Mary, University of London	51	38	28	24	17	15	13	12	12	11	544
University of Manchester	34	25	20	18	18	14	11	10	9	9	406
University of Liverpool	38	26	22	15	14	12	12	12	11	10	368
Lancaster University	27	5	4	4	4	4	3	3	3	3	128
University of Sheffield	45	16	16	14	14	11	7	7	6	5	261
University of Leeds	30	25	25	19	18	12	9	9	7	6	331
University of York	33	16	14	10	9	9	8	5	5	5	216
Durham University	13	10	9	6	6	5	5	4	3	3	132
Newcastle University	33	20	19	17	17	13	13	10	10	9	368

Table 4. *(Above) Most common venue counts for universities in examined in this study.*

UNIVERSITY	CLUSTER
University of Oxford	5
University of Southampton	5
University of Sussex	5
University of Reading	3
University of Cambridge	4
University of East Anglia	7
University of Bristol	8
University of Exeter	6
University of Bath	2
University of Warwick	0
University of Birmingham	5
Keele University	5
Coventry University	5
University of Nottingham	3
University of Leicester	4
Loughborough University	7
Nottingham Trent University	1
Imperial College London	9
University College London	2
King's College London	0
London School of Economics and	5
Queen Mary, University of Lond	5
University of Manchester	5
University of Liverpool	3
Lancaster University	4
University of Sheffield	7
University of Leeds	1
University of York	6
Durham University	2
Newcastle University	0

Table 5*. (left) Universities and their respective cluster label.*