Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

October 5, 2017

1 Introduction

This report will examine the dataset collected by the Harvard Dialect Survey (2003). In section 2 techniques for dealing with nonresponse are discussed and the relationship between two survey questions are investigated with illustrative maps plotted. In Section 3, the categorical data are encoded into binary variables and different dimension reduction and clustering methods are performed to gain insight into the underlying geographic groups of the respondents. I conclude the report with a discussion in Section 4. (The follow-up plots for lab 1 are presented in Section 5.)

2 The Data

The linguistic dataset contains the answers to the questions for 47,471 respondents across the United States. The dataset contains the variables ID, CITY, STATE, ZIP, Q50 - Q121 (a few questions in this range are left out), lat and long. The answers of the respondents are indexed by an integer. Some of the questions have as many as 21 choices while some have only 3. A zero indicates that the respondent did not answer the specific question.

2.1 Data quality and cleaning

This dataset isn't as bad as the redwood data, but there are still some issues, mostly because some of the the test takers did not answer all the questions. There are about 1020 rows with missing lat and long values. Since those rows would be automatically left out during plotting, I did not delete any of them. Almost all questions have a non-response rate of 3%, so I did not remove any of the questions. On the other hand, most respondents (over 40000 out of 47471) answered all the question while some of them answered less than a half. Missing more than 10 questions shows lack of care. Those respondents might not have taken the survey seriously and I removed the rows with more than 10 zeros (there are 1473 of them) and set other zeros to NA. I also checked that the range of the answers match the number of choices provided in all.ans.

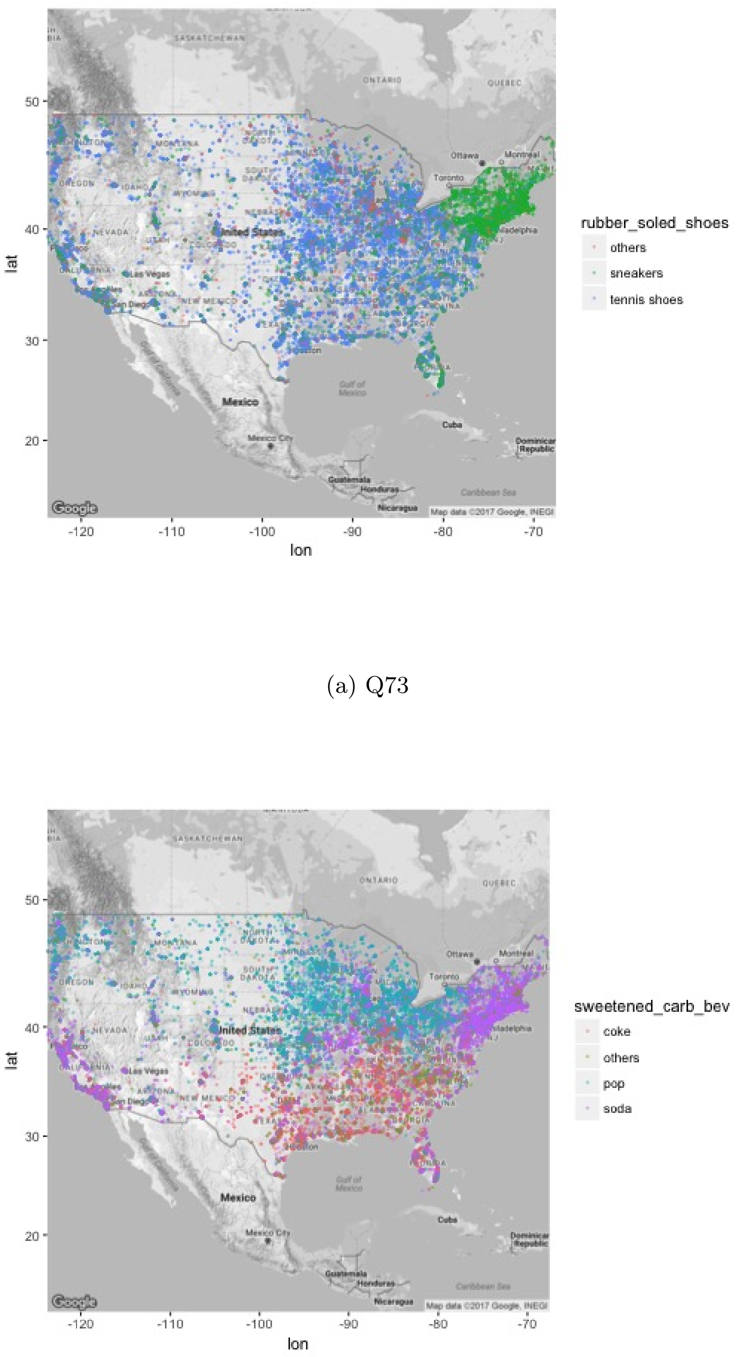
2.2 Exploratory Data Analysis

Let us look at the following two questions: Q73 and Q105. The 73rd question is

> *What is your *general* term for the rubber-soled shoes worn in gym class, for athletic activities, etc.?*

and choices include **a.sneakers** (45.50%), **c.gymshoes**(5.55%) and **f.tennis shoes** (41.34%), among others. The 105th question is

> *What is your generic term for a sweetened carbonated beverage?*



(a) Q73

(b) Q105

Figure 1: Distribution of answers to Q73 and Q105, shown on a US map

and choices include **a.soda** (52.97%), **b.pop**(25.08%) and **c.coke** (12.38%), among others.

The Kendall correlation for these two variables is 0.30, which is quite large for two categorical variables with more than 10 classes. The results of Q73 and Q105 are plotted on a U.S. map from the ggmap package, and is shown in Figure 1.

It can be seen that the two questions (especially Q73) do define distinct geographic groups. Almost all respondents from northeastern US (NY, PA, NJ, CT, MA, VT, NH, ME) as well as some from California call athletic footwear "sneakers", while respondents from other parts of the country call them "tennis shoes". As for carbonated beverages, respondents from northeast and California call them soda while other respondents refer to them as either pop , a term commonly used in the northern part of US, or coke, which is more widely used in the south.

As such, it is possible to predict the response of Q73 based on Q105 – if one chooses "soda" for Q105, he/she is likely to choose "sneakers" for Q73. If a respondent choose "tennis shoes" as their general term for rubber-soled shoes worn in gym class, he/she is likely to choose "pop" or "coke" for Q105.

There are several other cliques of questions closely associated with each other. For example, questions related to the use of "anymore" (Q54-57) and names for grandparents (Q68-71). These relationships will be examined more carefully in the following sections.

3 Dimension reduction methods

3.1 Dimension reduction via PCA

I proceed to encode categorical data into binary variables. This produces a new data frame with $p = 468$ variables.

For dimension reduction, I first used PCA. The eigenvalues of the covariance matrix is plotted in Figure 2. The elbow of the eigenvalue curve corresponds to the ≈ 50 largest eigenvalues. Due to high variance of the data, the first 2 principle components and the first 50 principle components can only explain 8% and 56% of total variance respectively.

Since displaying the whole dataset can be challenging, I use a subset of the data to present the result of PCA. Figure 3 plotted the first two principle components for respondents from two states –New York and Texas. These two states are chosen because they have a large number (over 2600) of respondents. It can be seen that samples from the two states are well separated after projecting the original dataset onto this two dimension space.

3.2 Dimension reduction via NMF

Next, Non-negative Matrix Factorization is used as an alternative approach for dimension reduction. For this dataset, the main advantage of non-negative matrix factorization is that the basis computed by NMF has direct interpretation.

Due to heavy computational complexity of NMF algorithms, I use a subset of the data for matrix factorization. I randomly select 5000 rows from the complete dataset, and compute a rank 10 approximation

$$X^T \approx WH,$$

where X is the 5000×468 data matrix, W is a 468×10 basis matrix and H is a 10×5000 coefficient matrix. Intuitively, the 10 column vectors in W can be understood as responses of 10 "representatives", each representing a potential geographical group. The (answers of) respondents can then be seen as a mixture , or linear combination, of the (answers of) the 10 representatives.

After fitting the model, I find that for almost all questions, only one of all possible answers corresponds to a value that is significantly larger than zero in each basis vector. For example, in the first basis vector, Q053b=2.02, Q053a=Q053c=0, Q058c=1.88, Q058a=Q058b=Q058d=. . . =Q058m=0, etc. This means that the basis can indeed be understood as a typical response to all the questions.

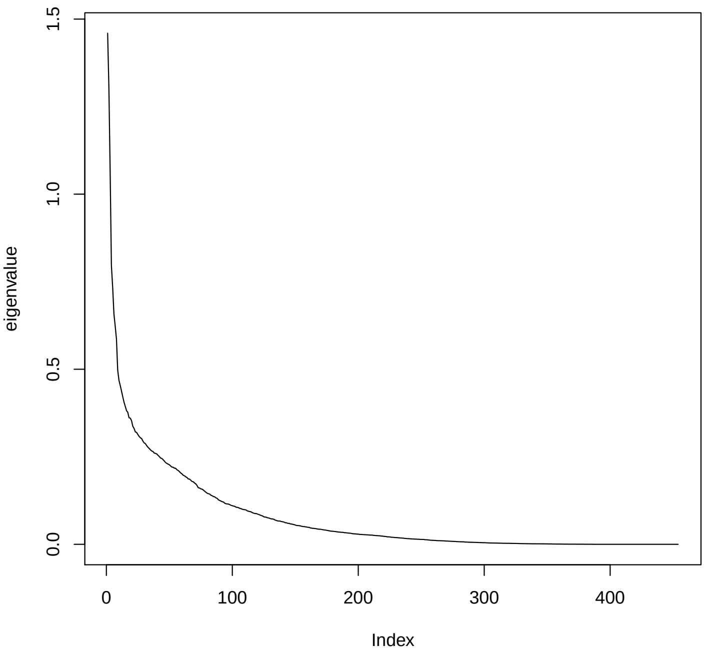


Figure 2: Eigenvalues of the covariance matrix in decreasing order

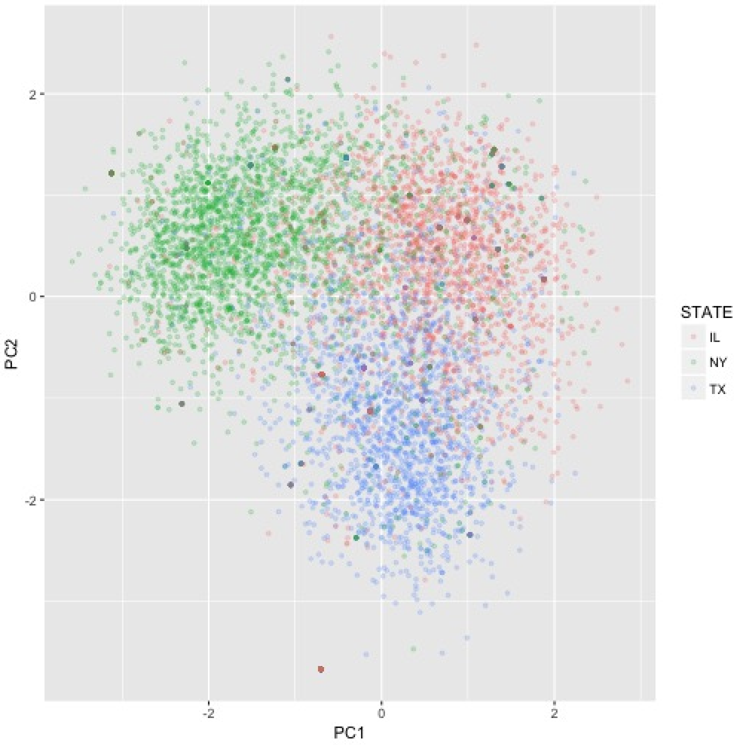


Figure 3: The first two principle components of respondents from TX, NY and IL.

	Question&Choice	Value		Question&Choice	Value
Basis vector 1	Q106a	2.47	Basis vector 2	Q073sneakers	3.48
	Q072a	2.20		Q105soda	2.46
	Q068c	2.20		Q109a	2.18
	Q069c	2.17		Q098a	2.09
	Q115a	2.09		Q080a	2.04
	Q100d	2.07		Q119a	2.00
	Q064a	2.05		Q056b	1.98
	Q081a	2.05		Q091b	1.93
	Q053b	2.02		Q079a	1.80
	Q054b	1.99		Q053b	1.76
Basis vector 3	Q076d	2.86	Basis vector 4	Q120a	2.77
	Q065a	2.44		Q098a	2.35
	Q103d	2.23		Q050g	1.98
	Q077a	1.92		Q097e	1.94
	Q089a	1.88		Q051b	1.84
	Q094b	1.87		Q052b	1.83
	Q097a	1.83		Q115a	1.68
	Q093b	1.80		Q083g	1.59
	Q051b	1.76		Q104a	1.57
	Q073tennis shoes	1.74		Q087a	1.55

Table 1: Leading non-negative components in the first four basis vectors of the NMF fit. Each basis vector can be understood as the response of a "representative".

Moreover, The result of non-negative matrix factorization echoes our findings in Section 2. The leading terms in the first four basis vectors are presented in Table 1. We can see that the question&choice combinations "Q073sneakers" and "Q105soda" appears simultaneously as leading features in the second basis vector. Figure 4 shows the coefficients of these two basis vectors for respondents from two states – New York and Texas. We can see that the second coefficient NMF_2(corresponding to basis vector2 in Table 1) is zero for most respondents from Texas, which indicates that few of them answered "sneakers" for Q73 or "soda" for Q105. Similarly, a large proportion of respondents from the state of New York did not select choice **a.tp'ing** for Q106: *What do you call the act of covering a house or area in front of a house with toilet paper?*

3.3 Clustering via kmeans

Although the dimension of the data $p = 468$ is not small, the Euclidean distance between two rows is bounded by $\sqrt{2 \times \text{Number of questions}} < 12$. Therefore we can directly apply the kmeans algorithm to the full binary data. If we use PCA for dimension reduction before applying kmeans, we would have to keep the first 50 principle components to account for half the variance and the first 300 for 90% of the variance, and it would be hard to identify the questions which separates the clusters. So PCA+k-means will only be used for validation.

Figure 5 shows the average squared distance between each row and their corresponding cluster center. There is an elbow at $K = 3$, so kmeans with number of cluster $K = 3$ is applied. It will be shown in the following section that this choice also meets stability requirements.

75% of the data is used to train the kmeans algorithm and obtain cluster centers, while the other 25% is left for validation. Cluster assignments are shown in Figure 6. The three groups shown on the map relate closely to geography. Specificly, group 1 (red) includes most of the southern states. Group 3 (blue) corresponds to northeastern states along the east coast while group 2 (green) mainly consists of states in the western and midwestern regions.

Which responses define these groups? To answer this question, let's look at the coordinates of the cluster centers. These are the percentage of each choice for each question in the three groups. Figure 7 shows the

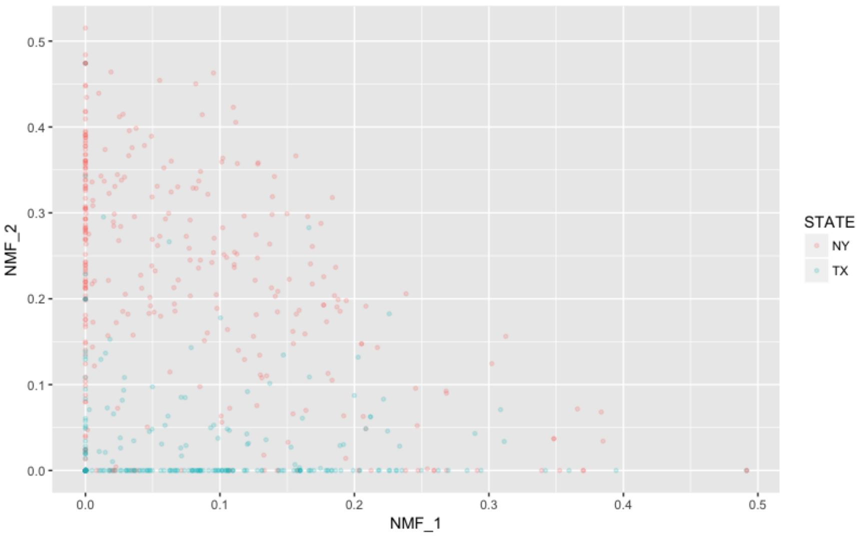


Figure 4: The first two coefficients of the NMF fit, plotted for respondents from TX and NY.

	Q103d	Q076d	Q089a	Q050i	Q105c	Q076a	Q103c	Q105b	Q089c
Group 1	86.4%	70.3%	76.8%	39.2%	34.3%	12.5%	11.7%	22.6%	15.7%
Group 2	27.9%	15.1%	36.8%	3.55%	4.20%	74.3%	61.3%	59.9%	46.1%

Table 2: (Question, choice) pairs that produce the continuum. The percentage of each answer in the boundary region (latitude between 35 and 45 and longitude between -92 and -80) is shown for group 1 and 2. For example, in this region, 86.4 percent of respondents who belong to group 1 answered d for Q103.

top ten question&answer pairs for which the three groups respond most differently. We can tell from the figure that Q50, Q73, Q76, Q80, Q103 and Q105 contributes most to the separation of the groups.

There is a continuum from Missouri and Illinois to Indiana and Ohio, as shown in Figure 8. In particular, respondents from the cities of St. Louis, Indianapolis, Cincinatti and Columbus are a mixture of the three groups. For example, in Indianapolis, 54.1%, 35.0% and 10.9% of the the respondents belong to group 1 (red), 2 (green) and 3 (blue) respectively. Compare this to the city of New York, where 78% of the respondents belong to group 3 (blue).

To examine the questions that cause the continuum, let's focus on the subset of the data with latitude between 35 and 45 and longitude between -92 and -80, an area identified as the boundary of groups 1 and 2. As before, difference in the percentage of each answer is calculated. Table 2 shows the questions with the most significant differences. We can see that the continuum is produced by Q76 (kitty-corner (Group 2) vs. catty-corner (Group 1)), Q103 (drinking fountain (Group 2) vs. water fountain (Group 1)), Q89 (can you call coleslaw 'slaw'? yes (Group 1) vs. no (Group 2)) and Q105(coke (Group 1) vs. pop (Group 2)).

4 Stability of findings to perturbation

To analyze the robustness of kmeans, I draw $n = 100$ samples of 75% of the data and run the kmeans algorithm. This results in 100 sets of cluster centers. One of the cluster centers is treated as an oracle and the distance to this oracle is calculated for the other 99 cluster centers. Figure 9 shows the histogram of this distance. We see that the distance is typically around 0.1, which is fairly small. Note that the starting point of the kmeans algorithm is chosen randomly from the rows (a default setting of the R kmeans function), so sensitivity of the result to different starting points is included here. We conclude that the result of the kmeans algorithm is fairly robust. In general, this will be the case when the number of observations is much larger than the number of variables.

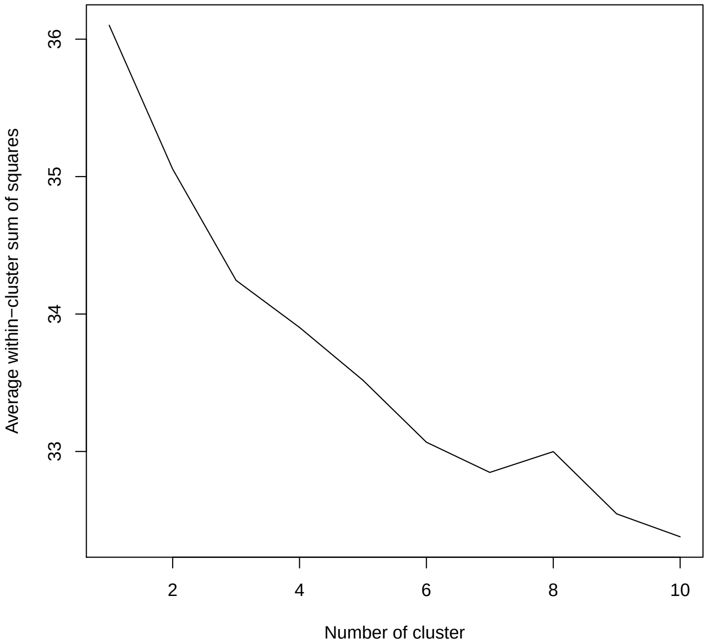
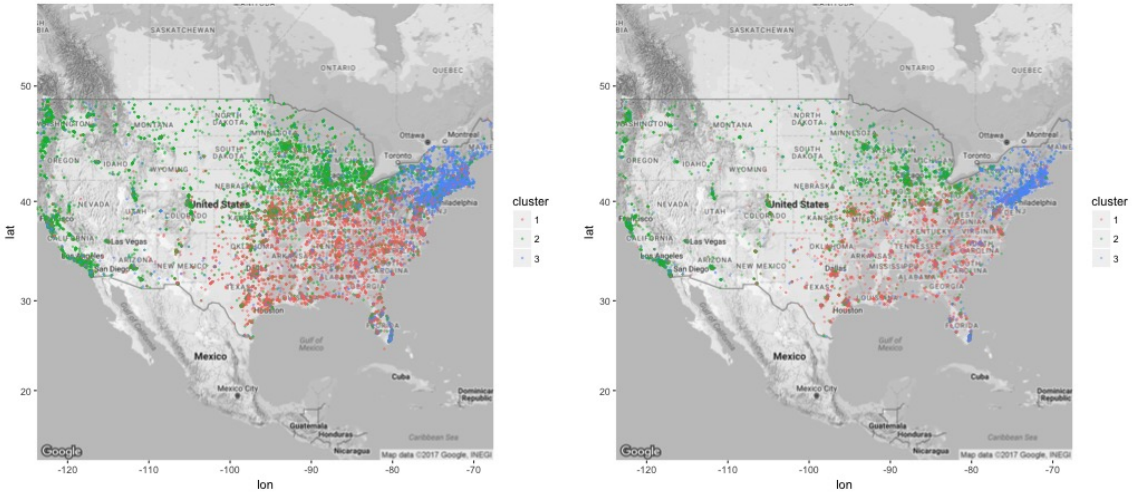


Figure 5: Average within cluster sum of squares for different number of clusters.



(a) Training

(b) Validation

Figure 6: The result of kmeans with $K = 3$ clusters. 3/4 of the data is used for training (left panel), and the other 1/4 for validation (right panel)

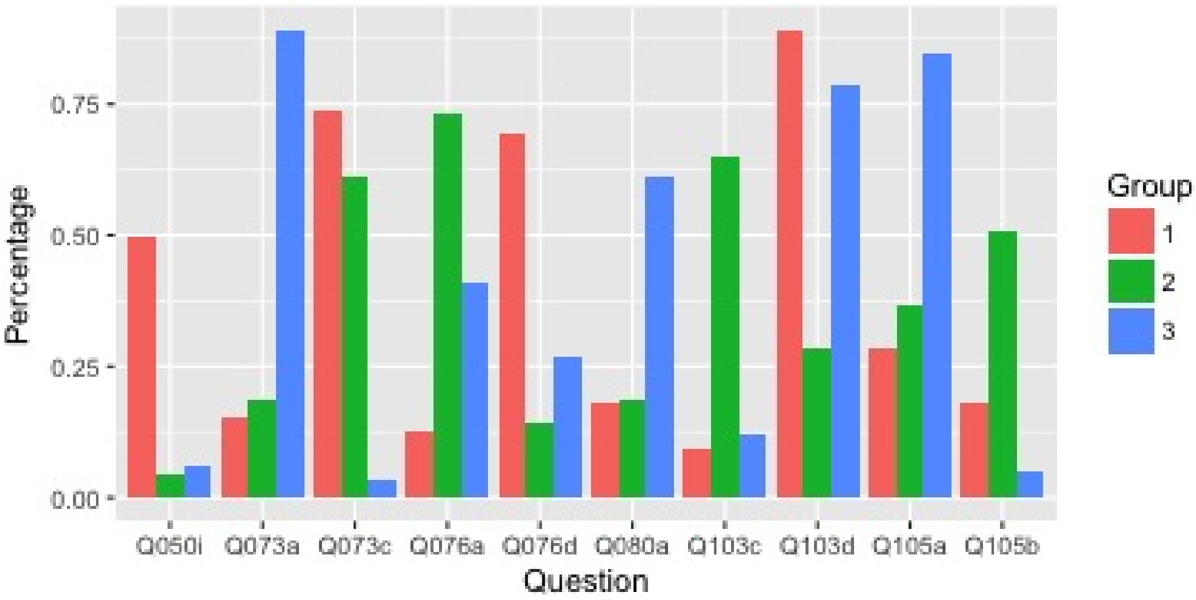
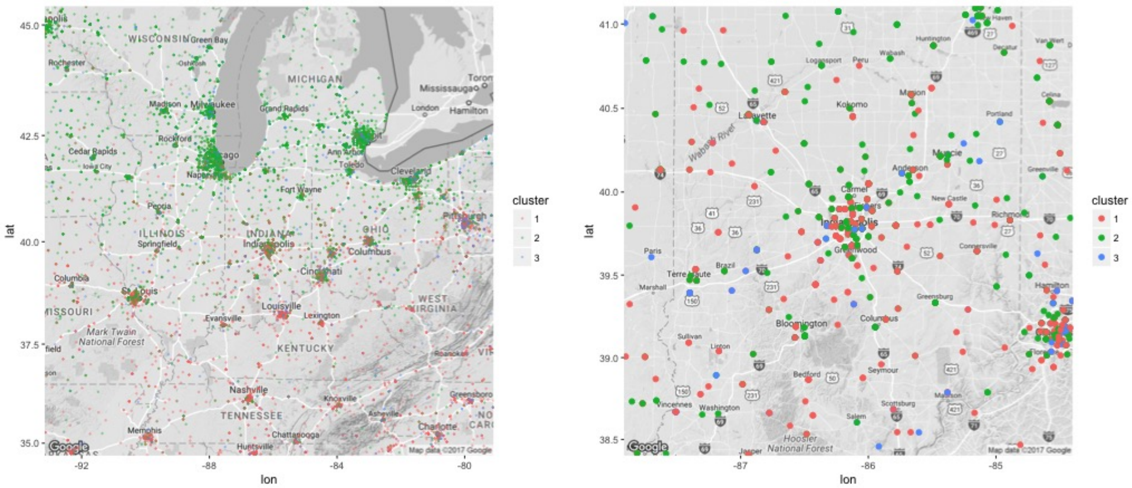


Figure 7: (Question,answer) pairs that define the groups



(a) The whole region–MO, IL, IN, OH

(b) The city of indianapolis– balanced mixture of group 1 and group 2.

Figure 8: The boundary of groups 1 and 2

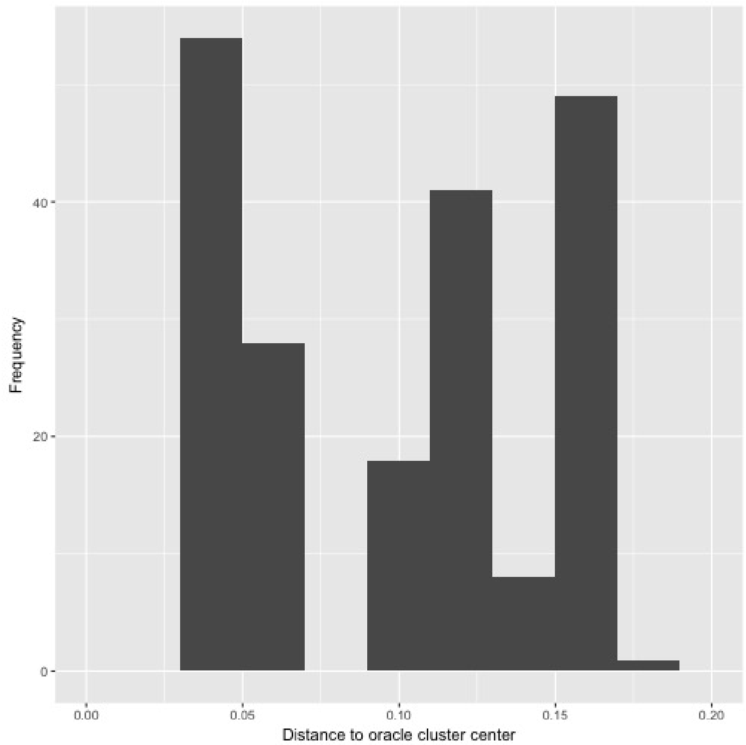


Figure 9: Histogram of distance from cluster centers to the oracle center for different subset of the data. 3/4 of the whole dataset is drawn randomly for n=100 times.

5 Discussion

We can see that Q73 and Q105 play a very important role in both dimension reduction and clustering. It would be helpful to apply dimension reduction algorithms designed specificly for binary data. Due to time and space limitations, this would be left for future research.

6 Kernel density plots and smoothing

The kernel density estimation is shown in Figure 10. As bandwidth increases, the estimation becomes smoother. This can be explained by the following two properties of KDE. First, we know for each fixed x, the variance of the kernel density estimator at x is proportional to $\dfrac{1}{nh}$ for small h, so when h increases the variance drops. Second, the absolute value of the second derivative of KDE with respect to x is given by the In fact, we have

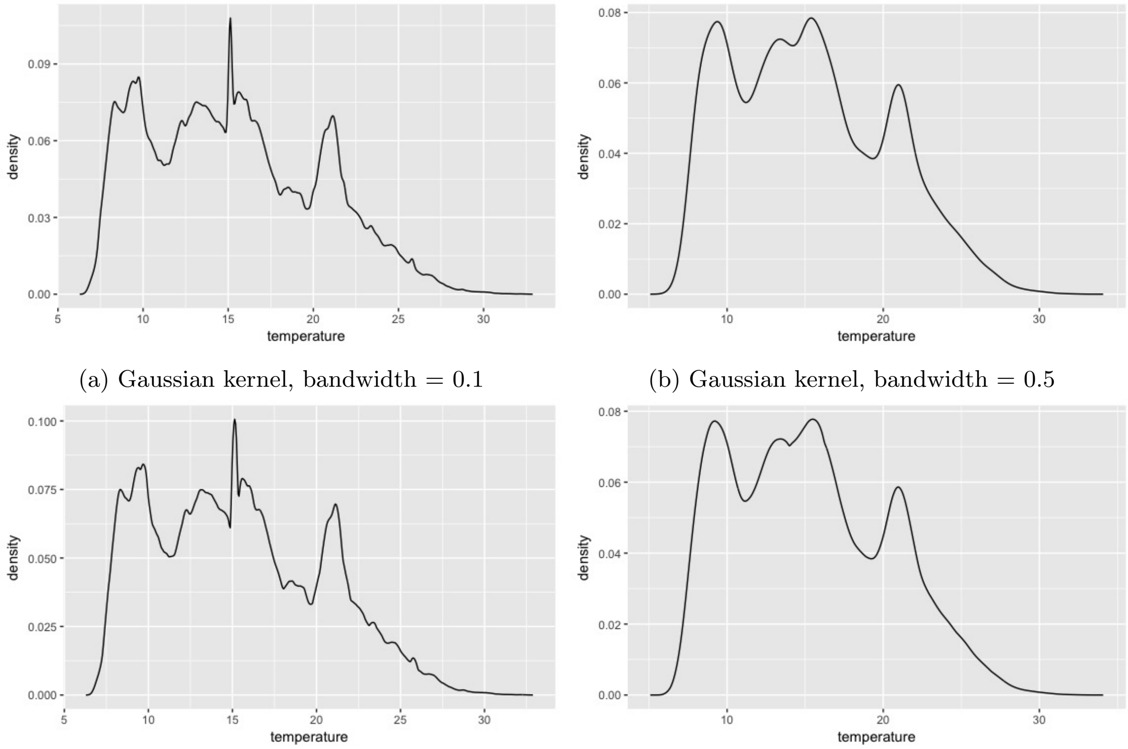
$$\hat{f}_h''(x) = \frac{1}{nh^3}\sum_{i=1}^{n} K''(\frac{x-x_i}{h}) \to \frac{1}{h^3}\int_{-\infty}^{\infty} K''(\frac{x-y}{h})f(y)dy$$

when n is large, where f is the true density of the data. When h is small,

$$\begin{aligned} |\frac{1}{h^3}\int_{-\infty}^{\infty} K''(\frac{x-y}{h})f(y)dy| &= |\frac{1}{h^2}\int_{-\infty}^{\infty} K''(z)f(x-zh)dz| \\ &= |\frac{1}{h^2}\int_{-\infty}^{\infty} K''(z)(f(x)-zhf'(x)+o(h))dz| \\ &= \frac{1}{h}|f'(x)\int zk''(z)dz| + o(\frac{1}{h}). \end{aligned}$$

Therefore when h is small, the second derivative of KDE is roughly inversely proportional to h, and KDE is less smooth.

For the loess smoother (Figure 11), we find that it fits the data better when a second degree polynomial is used and when the span parameter (smoothness parameter) is small. This is because using second degree polynomials introduces more parameters, and using a small span parameter reduces the smoothness and therefore the bias, as smoothing can be understood as a kind of regularization.



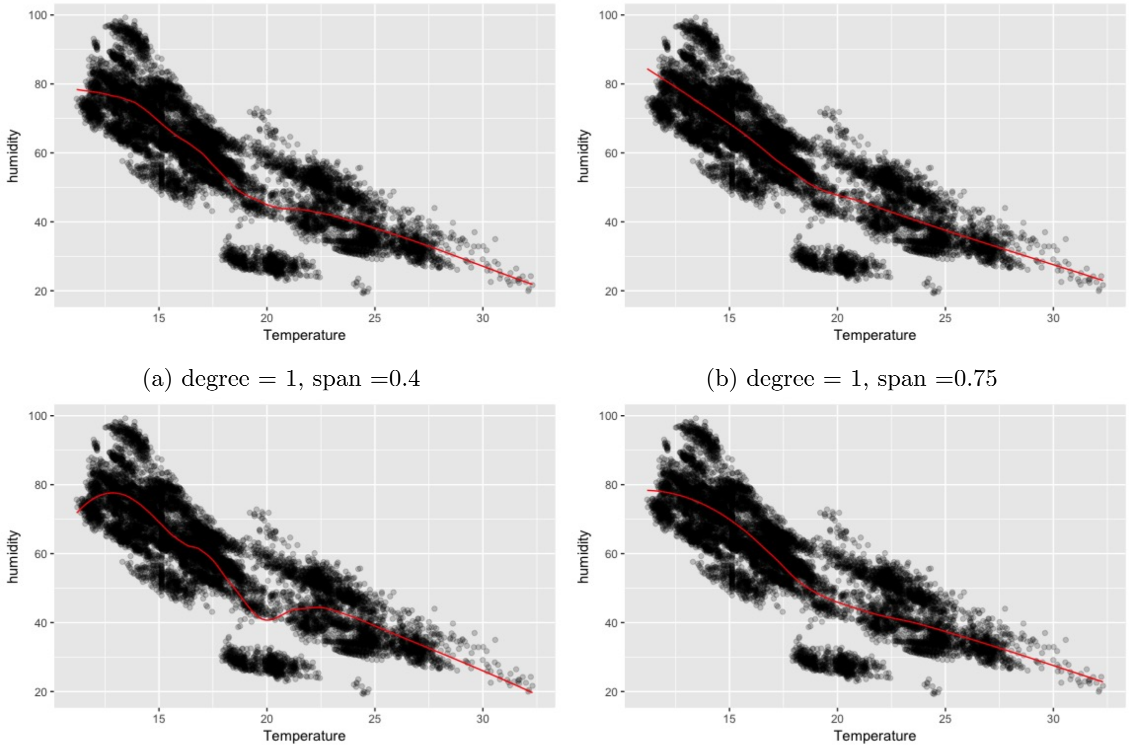
(a) Gaussian kernel, bandwidth = 0.1

(b) Gaussian kernel, bandwidth = 0.5

(c) Epanechnikov kernel, bandwidth = 0.1

(d) Epanechnikov kernel, bandwidth = 0.5

Figure 10: Kernel density estimation for the distribution of temperature with different kernels and different bandwidths.



(a) degree = 1, span =0.4

(b) degree = 1, span =0.75

(c) degree = 2, span =0.4

(d) degree = 2, span =0.75

Figure 11: Temperature against humidity at 2pm, with a loess smoother with different bandwidths and the degrees of the polynomials