Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

Melody(Mengling) Liu

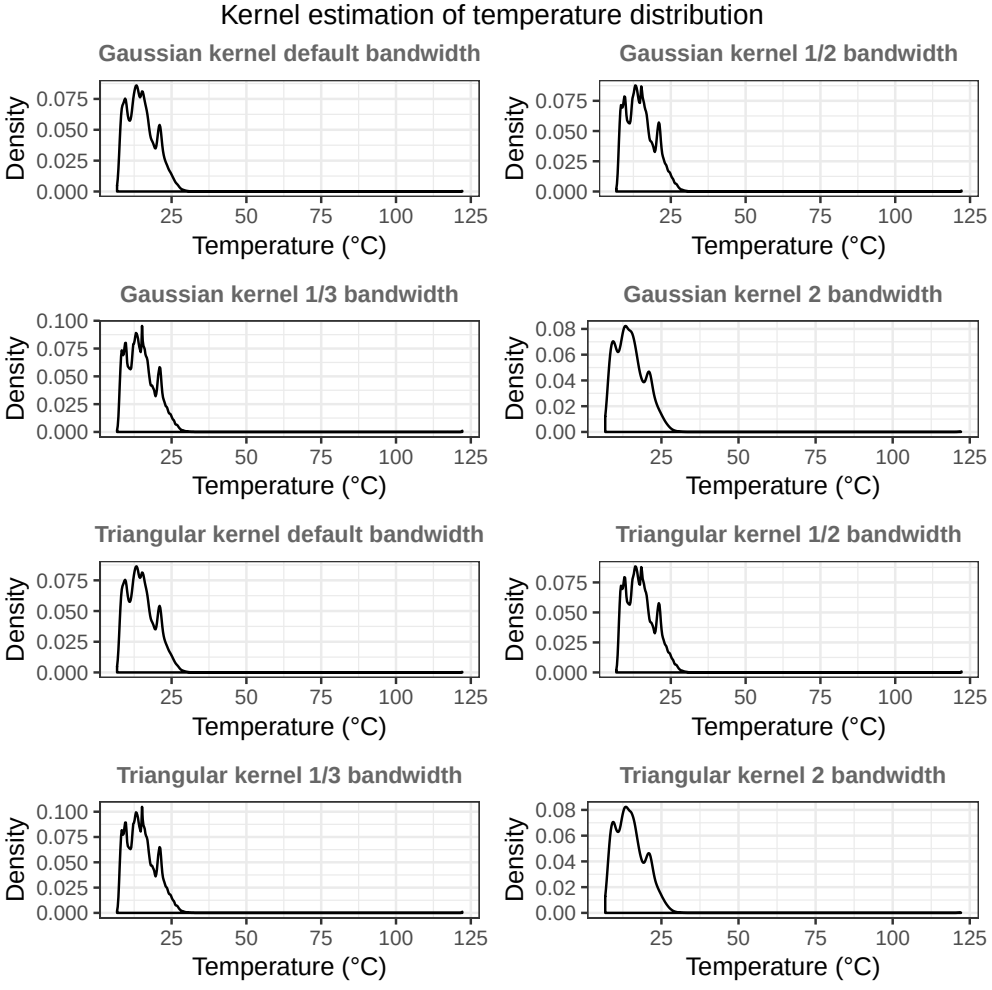
October 8, 2018

1 Kernel density plots and smoothing

These tasks use the redwood data from the previous lab. These tasks are focused on experimenting with parameters in kernel smoothers.

1. Plot a density estimate for the distribution of temperature over the whole dataset. Experiment with different kernels and bandwidth. Explain your findings.

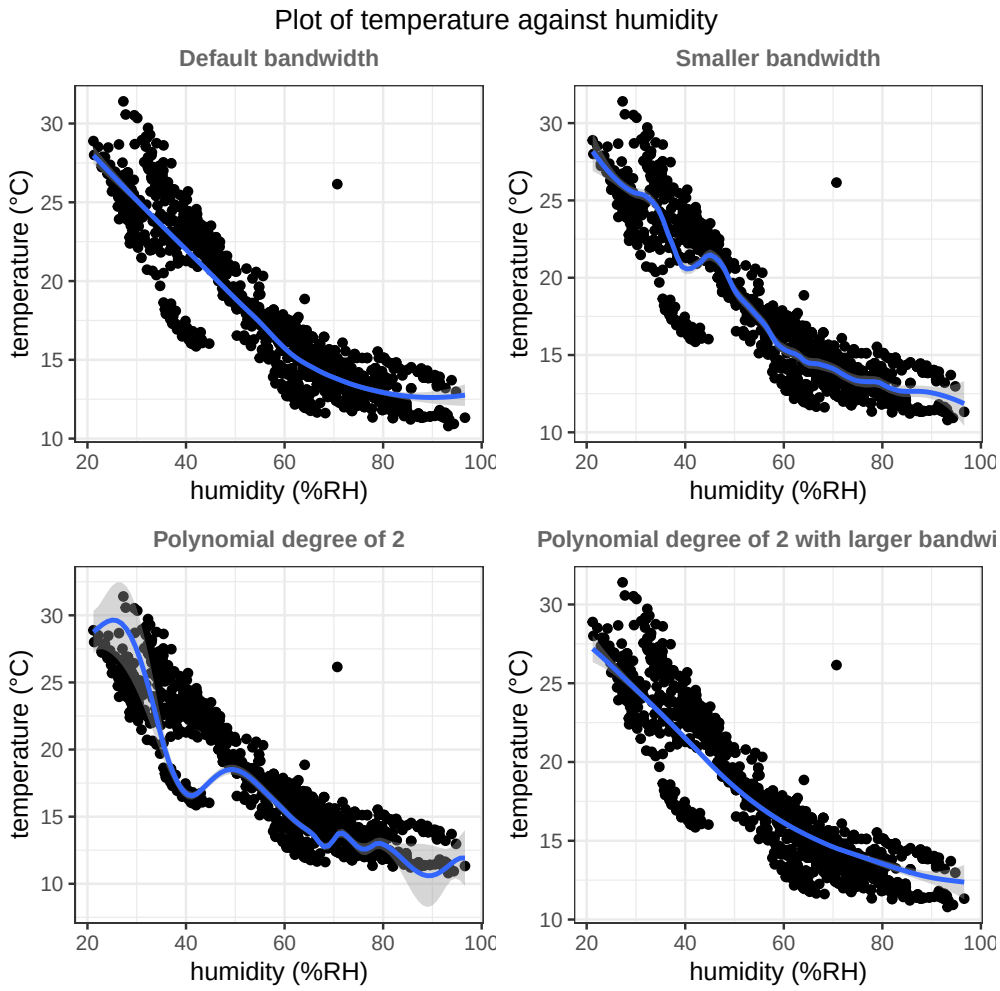
I used both gaussian and triangular kernel to estimate the distribution. I discovered that the larger the bandwidth, the more smooth the plot will be and the plot will be closer to the original kernel distribution. Therefore there is trade off between variance and bias.



2. Choose a time of day and plot the temperature against the humidity for all nodes at that time for the entire project period. Add a loess smoother to the plot. Experiment with bandwidth and the degree of the polynomials. Explain your findings.

I chose 00:30 of the day and plotted the temperature against the humidity of all nodes at that time for the entire project period.

I observed from the plots below that the smaller the bandwidth is, the wigglier the line is. Also the higher degree of polynomials, the wigglier the line is.



2 Introduction

Dialectology is the study of dialects, and dialectometry is the measurement of dialect differences, i.e. linguistic differences whose distribution is determined primarily by geography. Dialectology may be classified within the more general study of how languages vary not only along geographical, but also social lines or along lines of age and gender. In addition to dialectology, the study of linguistic variation as it correlates with social class, age, sex, and occupation.

In this r eport, we studied the linguistic varieties that primarily determined by geography. We analyzed the linguistic survey data that contains the answers of 72 questions regarding to daily colloquial expressions from more than forty thousands respondent across North America.

Here I first performed dimension reduction using principal component analysis and then applied k-means clustering on the dimension-reduced form of data.

3 The Data

3.1 Data quality and cleaning

We have five data frames loaded: lingData, lingLocation, question.mat, quest.user, all.ans

lingData: contains 47,471 observations and 73 features, including observation ID, city, state, zipcode, answers for question 50 to question 121, the latitude and longitude of the location.

lingLocation: contains 781 locations and 471 features, including the one-hot coding for 72 questions, numbers of observations at this location, and latitude and longitude for this location.

quest.mat: all the questions that were asked in the survey, 121 in total.

quest.use: all the questions that are included in the lingData and lingLocation

I first check whether there are missing values for each column. There are no missing value for ID, City and Zip, but there are three missing values for state, so I filled in the state name corresponding to the zip code.

3.2 Exploratory Data Analysis

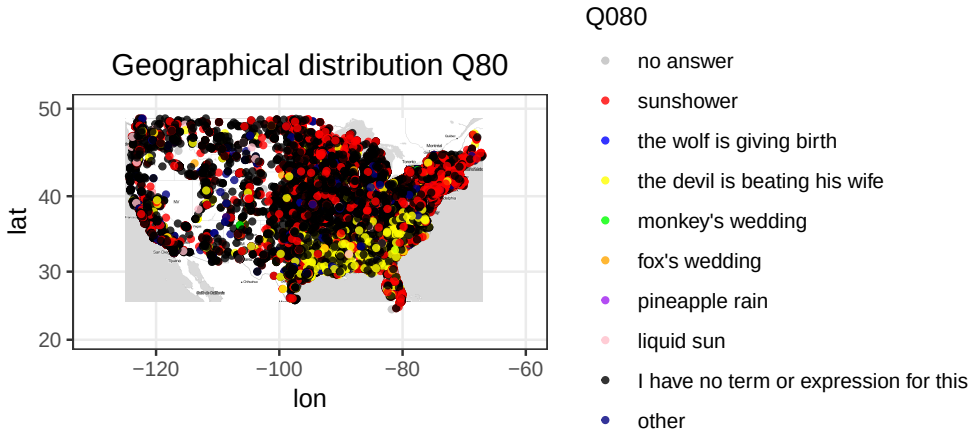
I am interested in the geographical distribution of answers for question 80 and question 70.

Question 80: What do you call it when rain falls while the sun is shining?

Question 70: What do did you call your maternal grandfather??

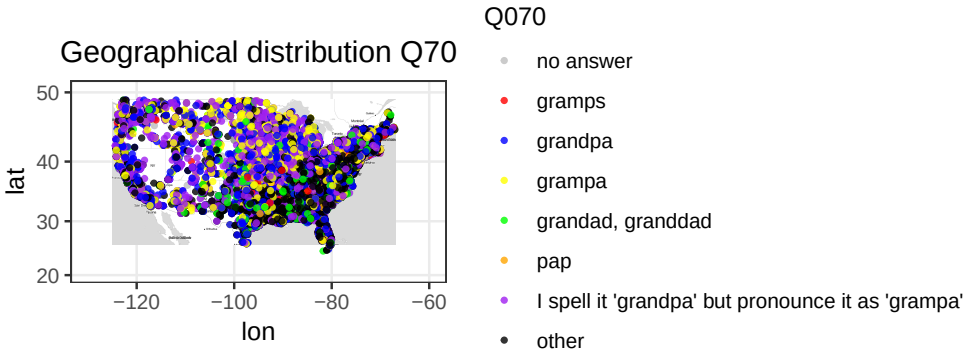
First, regarding question 80. 55.15 percent of people came up with answer "I have no term or expression for this", 34.29 percent have answer "sunshower", 6.43 percentage have answer "the devil is beating his wife", 3.02 percentage have answer "other", etc.

I narrowed the geographical scope to US mainland only. Then I plotted each answer as a dot on the US map with colors representing different answers. We observe from the graph that people from different areas of the US gave different answers. People from New England Area tend to give sunshower as an answer, while people from the South came up with answer "the devil is beating his wife". The black dots representing "I have no term or expression for this" have the most wide distribution across the West and North of the US.



Second, I also explore the answers to question 70 - ”What do did you call your maternal grandfather?”. Regarding question 70. Gramps: 1.02 percentage; Grandpa: 21.05 percentage; Grampa: 13.86 percentage; Grandad,granddad: 5.07 percentage; pap: 0.84 percentage; I spell it ’grandpa’ but pronounce it as grampa: 25.90 percentage; other: 32.26 percentage.

Similarly, I plotted each answer as a dot on the US map with colors representing different answers. We observe from the graph that people from different areas of the US gave slightly different answers. People from the North tend to come up with answer such as ”grandpa” or ”I spell it as grandpa but pronounce it as grampa”, on the other hand, people from the South and West tend to give more ”other” answers.

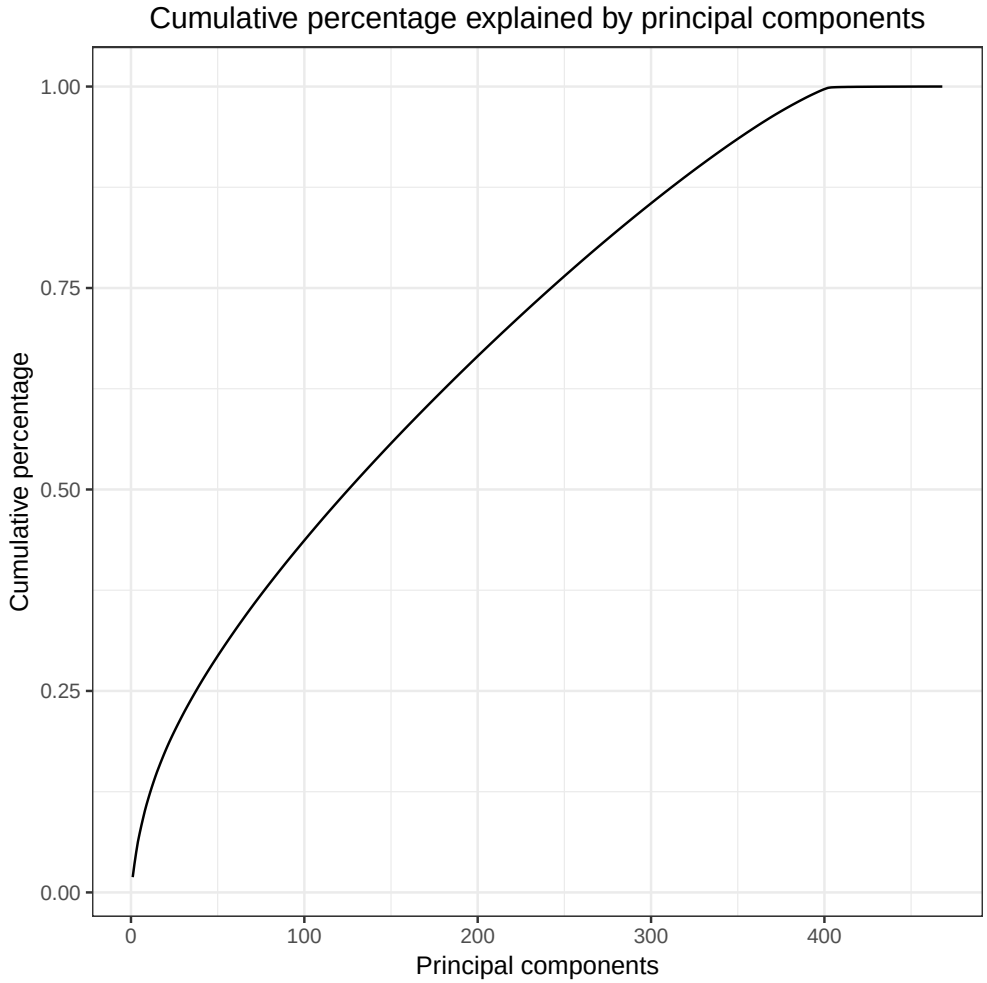


4 Dimension reduction methods

I first encoded the answer data to one-hot and I removed the features with No.0 because No.0 means this individual didn't have a response for this question.

Principal Component Analysis:

In order to implement dimensional reduction, I ran PCA on the binary encoding matrix with normalization. The following plots show the cumulative percentage explained by principal components.



According to the plot, the first 150 PCs explained around 55 percent of the variation of the observations, and the first 250 PCs explained around 75 percent of the variation of the observations. The first PC only explained around 1.9 percent of variation, which is pretty small. The top ten PCs also were not able to explain large percentage of variation in observations.

Based on the discovery from above on the relationship between location and question answers, it is interesting to see if there is any pattern of principal components associated with locations.

First, I grouped the states into several regions, according to their locations.

Northeast: NY, NH, MA, ME, VT, CT, RI, PA, NJ, DE, MD, DC

Pacific: CA, OR, WA, AK, HI

Midwest: MI, OH, IN, WI, IL, IA, MN, ND, SD, NE, KS, MT

South: TN,AL, AR, KY, LA, GA, FL, MS, NC, SC, VA, WV, MO, OK

Frontier: TX, NM, AZ, WY, CO, ID, NV, UT

Western Canada: BC, AB,MB,SK, YT, NT, NU

Eastern Canada: ON, QC, PE, NL, NB, NS

Other: States not belonging to any regions above

After grouping by Region, there are 14431 observations under Midwest, 12540 under Northeast, 6397 observations under Pacific, 4374 observations under Frontier, and 13 observations under Canada.

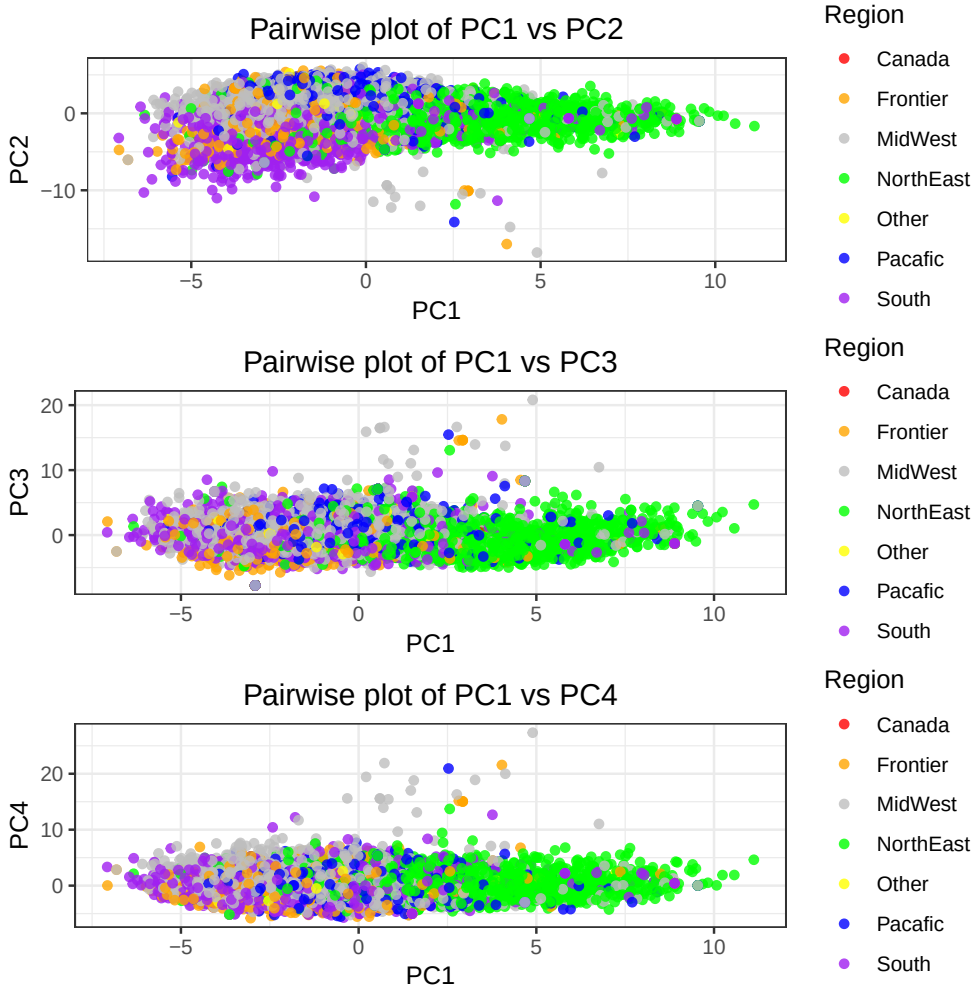
I conducted pairwise plot on PCs, including PC1 vs. PC2, PC1 vs. PC3, and PC1 vs. PC4. I selected 8000 samples out to create the plot.

The first two plots showed fairly clear clustering patterns.

On the pairwise plot of PC1 vs. PC2, PC1 separates Northeast fairly well from the South and MidWest. Northeast mostly located from 0 to 10 on PC1 and South located mostly from -7 to 0 on PC1. Also I observe that PC2 separates the South well from MidWest: the South sits in the range of -10 to 0 on PC2, and MidWest on the range of -2 to 3 on PC2. This indicates that NorthEast has a larger positive PC1 coefficient, and South has a larger negative PC2 coefficient.

On the pairwise plot of PC1 vs. PC3, we observe pretty clear pattern of PC1 separating Northeast out. NorthEast has a larger positive PC1 coefficient. However, PC3 doesn't capture strong pattern on location segmentation here.

On the pairwise plot of PC1 vs. PC4, we observe the same pattern as PC1 vs. PC3 that PC1 clearly separates NorthEast out from the graph. Similarly, PC4 doesn't capture strong pattern on location segmentation.



Interesting findings:

I want to find out which factor contributes the most to PC1. I looked into the linear combination of PC1 and located the maximum absolute number of PC1 rotation to tbe the 73rd question. The 73rd question goes "What is your general term for the rubber-soled shoes worn in gym class, for athletic activities?" "Sneakers" as an answer takes up 45.5 percent and therefore "sneakers" as an answer to the 73rd question contributed the most to PC1.

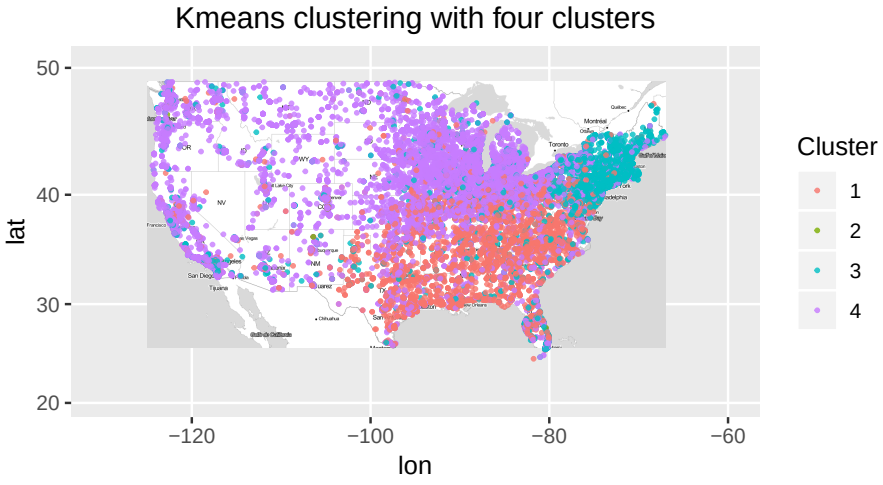
```
Q073_1
   165

   qnum ans.let   per                                ans
1    73       a 45.50                           sneakers
```

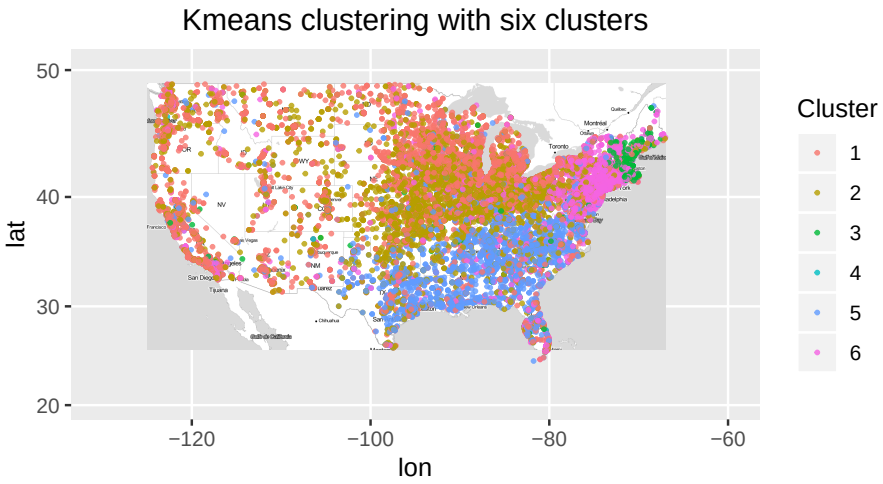
```
2    73       b  1.93                          shoes
3    73       c  5.55                       gymshoes
4    73       d  0.03                     sand shoes
5    73       e  0.01                        jumpers
6    73       f 41.34                   tennis shoes
7    73       g  1.42                  running shoes
8    73       h  0.17                        runners
9    73       i  0.23                       trainers
10   73       j  0.89 I have no general word for this
11   73       k  2.95                          other
```

K Means Clustering:

After the dimension reduction with PCA, we understand the first 250 PCs explained around 75 percent of the variation of the observations. Then I reduced the dimension from 468 to 200 and then applied K-means clustering on reduced matrix.

We can observe from the map plot that no matter whether we choose to adopt four or six clusters, the end results are fairly close. Northeast New England area, Southeast, and Midwest are the major clusters. Geographical location as a factor does matter to the question answers. The difference is that the six cluster further separate the New England Area out from the Northeast area, and MidEast area also stood out as a new segmentation as well.

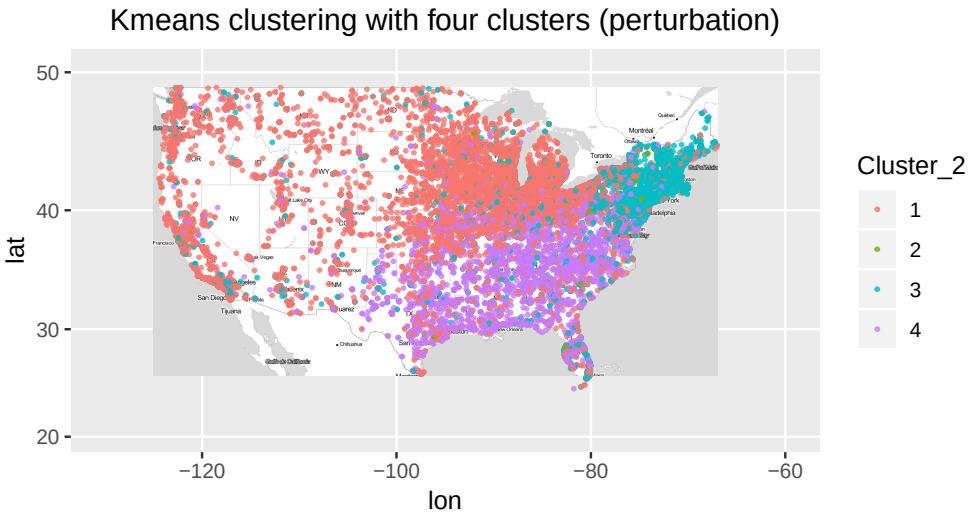




5 Stability of findings to perturbation

I manually perturbated the answers with randomly chosen questions and answers. I randomly changed the encoding of problem 59, 69, 79, 89, 99, 109. To test the robustness of the above findings, I ran PCA on the new encoding matrix and applied k-means clustering to check if similar results can be achieved.

From the new plot on the perturbated data, we can conclude that the above conclusions are robust and stable. Question 73 is still the one that contribute most to PC1. The clustering results are also very similar as well with Northeast, Southeast and Midwest being the major clusters.



6 Conclusion

We can conclude from the above analysis that there are geographical location contributes significantly to the dialect varieties. Question 73 is still the one that contribute most to PC1. I performed principal component analysis to reduce dimension from more than 400 to 250 and then applied k means clustering to group the geographical locations to 4 or 6 groups. Northeast, Southeast and Midwest are the three major clusters. The clustering result is robust with perturbated data as well.

References

[1] J.Nerbonne, W.Kretzschmar *University of Groningen, University of Georgia.* Introducing Computational Techniques in Dialectometry.

[2] J.Nerbonne, W.Kretzschmar *University of Groningen, University of Georgia.* Progress in Dialectometry: Toward Explanation