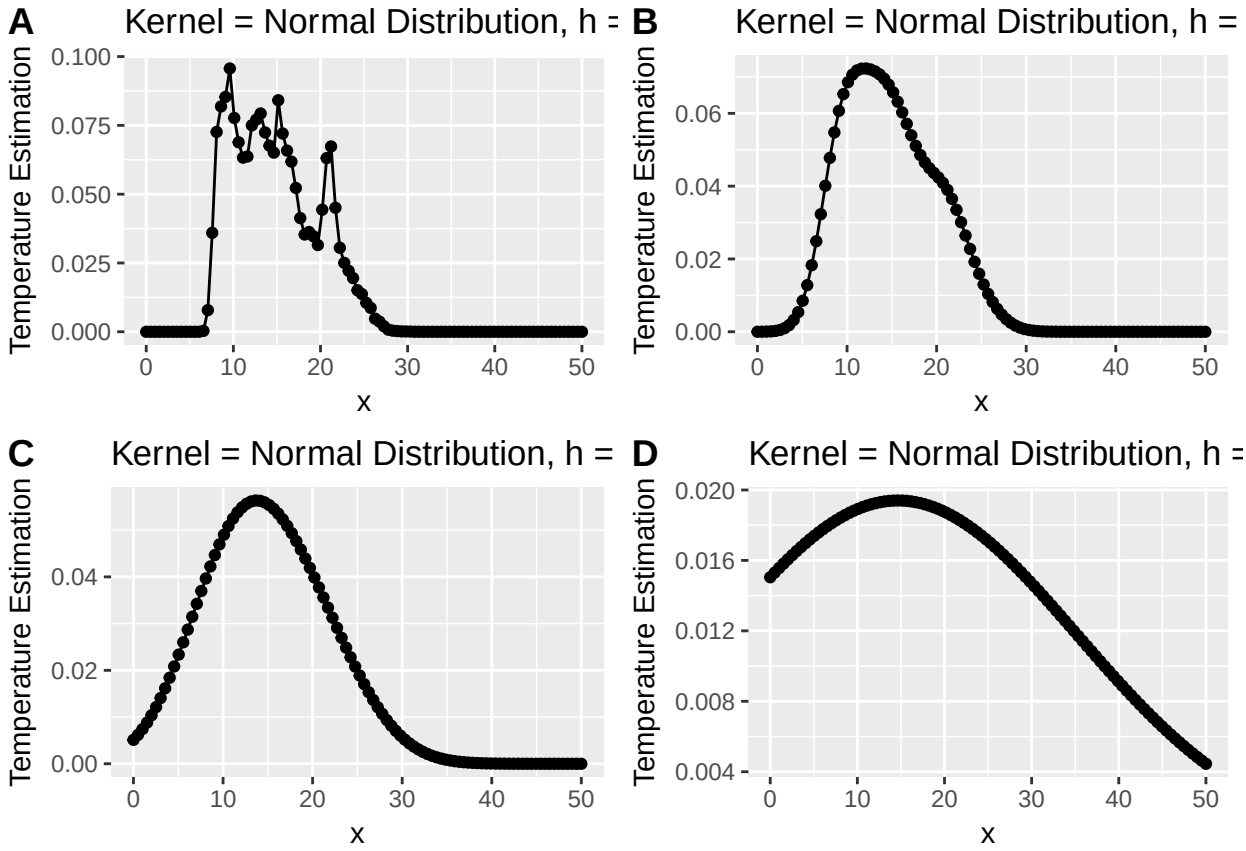
Lab 2 - Linguistic Survey Stat 215A, Fall 2017

Hongxu

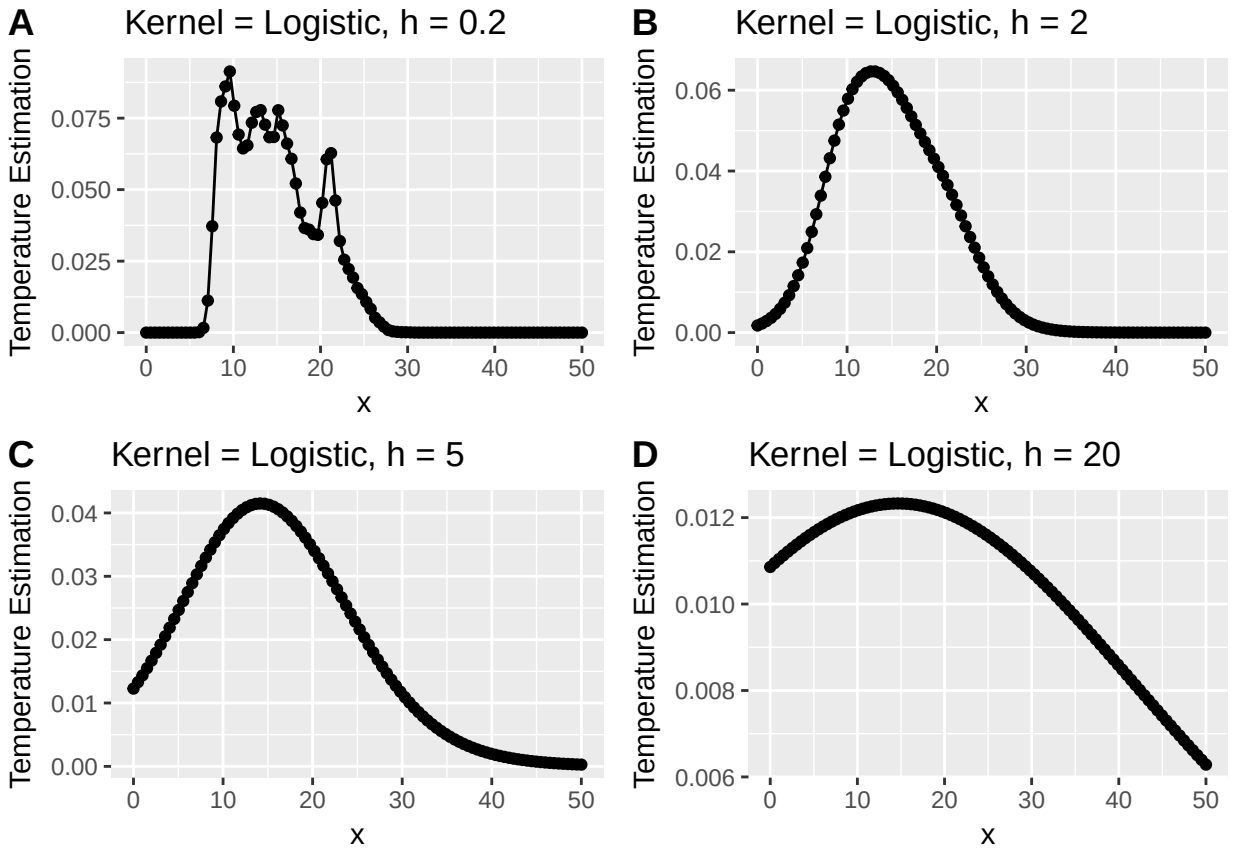
9/5/2017

1 Kernel density plots and smoothing

1.1 Plot a density estimate for the distribution of temperature over the whole dataset.



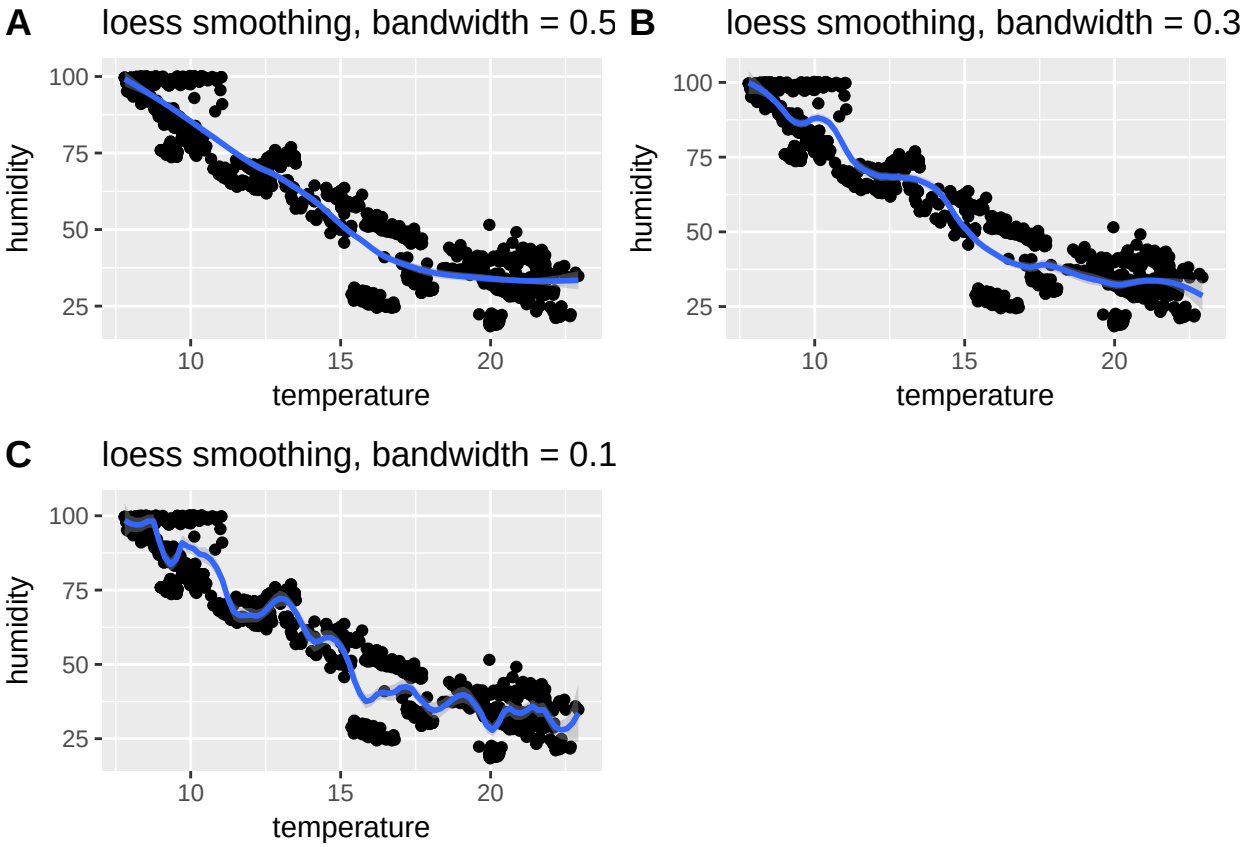
When h increase, the details of the temperature are masked. The curver is becoming more and more smooth. I also tried Logistic kernel.



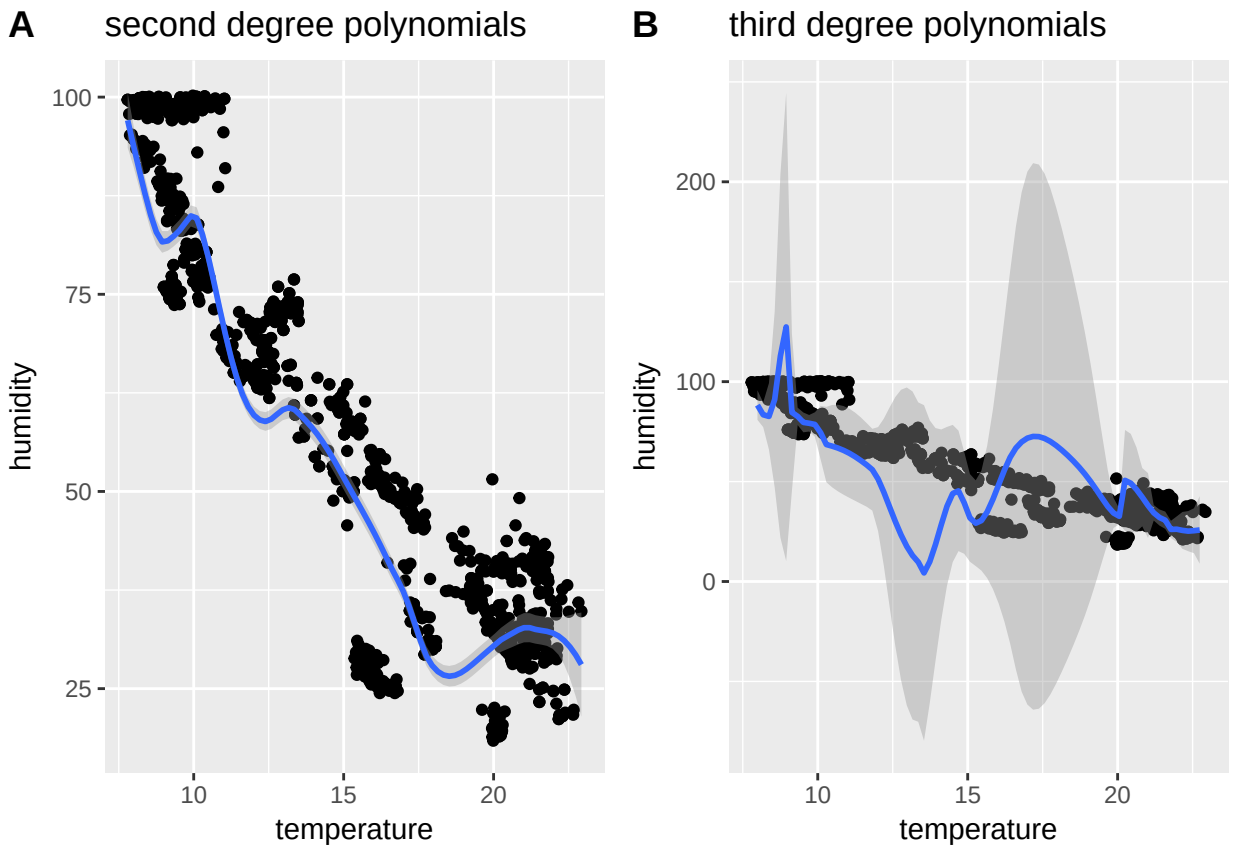
Logistic kernel density estimation looks similar to the normal distribution kernel density estimation. The only difference, for the same h, the logistic kernel density estimation is relatively lower and broader than normal distribution kernel density estimation.

1.2 Smoothing

I implement the loess smoother with different bandwidth on the dataset of temperature and humidity at the certain time of the day over the whole project period.



As bandwidth decrease, the smooth line will becomes more bumpy and try really hard to follow the data.



The second degree polynomial smooth paid more attention on the data points that group together than the first degree. But, I don't quiet understand why the figure becomes this wired with third degree polynomials smooth.

2 Linguistic Data

3 Introduction

In this project, I analysis the linguistic data collected from a Dialect Survey conducted by Bert Vaux. In the survey, 121 questiones are asked, and 67 of them are related to lexical differences. I picked up two questions and investigate the relationship to each other and then geography. Then I implement PCA on the dataset for dimension reduction.

4 The Data

I used three datasets for this project. The first dataset is named ling_data which contains the answers to 67 questions of each people, as well as his/her geographic location. The second dataset is named ling_location, which divided the United States into 1 degree * 1 degree cells and collect the answers wihtin each cells. This dataset is processed by former GSI. The last dataset is named question data, which contain all the questions and its answers.

4.1 Data quality and cleaning

To focus on the mainland United States, I selected the data that have longitude greater than -125.

To focus on the questions that related to lexical differences as opposed to phonetic differences, certain questions are selected from 212 total questions. The instruction says the question 50-121 are counted, however, question 108, 112, 113, 114 and 116 are not included.

I picked Question 70 and Question 71 to investigate their relationship. The answer of "other" is not included for analysis, because the people who choose "other" doesn't mean they have the same answer.

One of the answer to Question 70 is "I spell it 'grandpa' but pronounce it as 'grampa' ". Considering we are analysisng the lexical differences instead of phonetic differences. I group the people who choose this answer with the people who choose 'grandpa'.

To implement PCA, we have to normalized and center the data. In the ling_location dataset, I firstly divided the people who choose this certain answer by the total number of people in this region to make the row normalized. Secondly, before I implement the PCA on this dataset, I also normalized each columns.

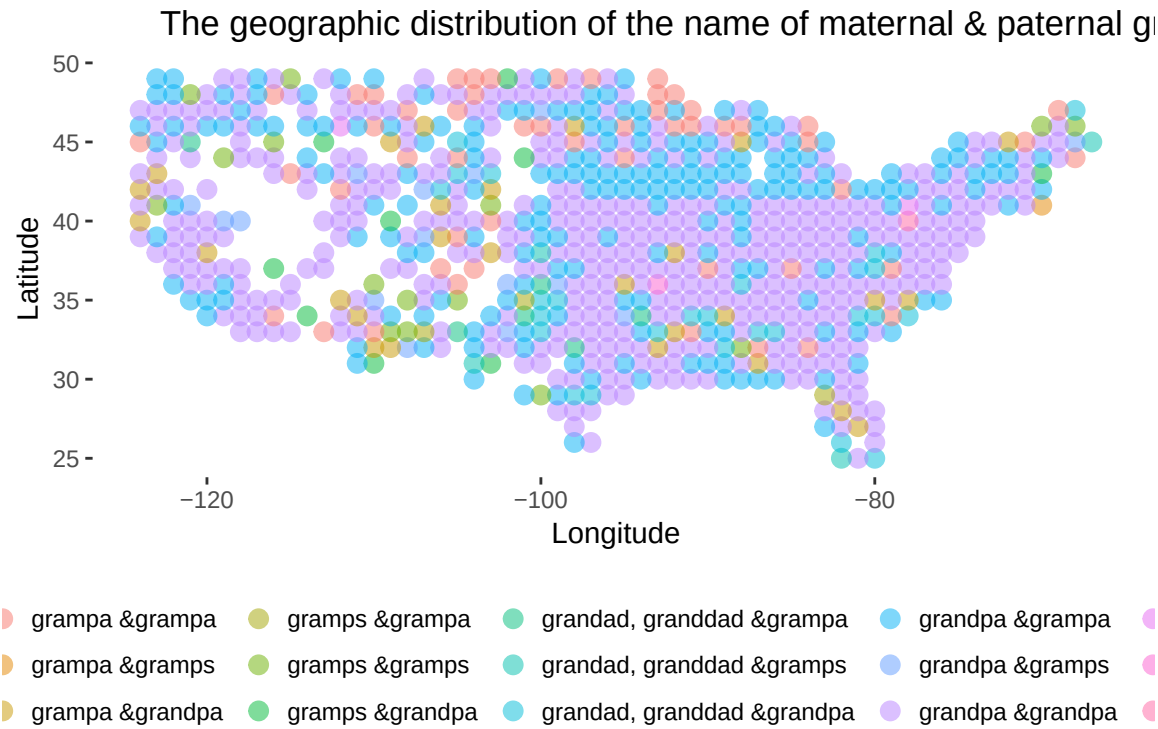
4.2 Exploratory Data Analysis

To investigate the relationship between two questions and their geographic correlation. I picked up question 70 and question 71.

Question 70:What do/did you call your maternal grandfather? Question 71:What do/did you call your paternal grandfather?

```
##   maternal & paternal grandfather Population Percentage
## 1               grandpa &grandpa         399 0.52638522
## 2                grandpa &grampa         198 0.26121372
## 3                 grampa &grampa          52 0.06860158
## 4                grampa &grandpa          27 0.03562005
## 5    grandad, granddad &grandpa          25 0.03298153
## 6                  gramps &gramps          15 0.01978892
```

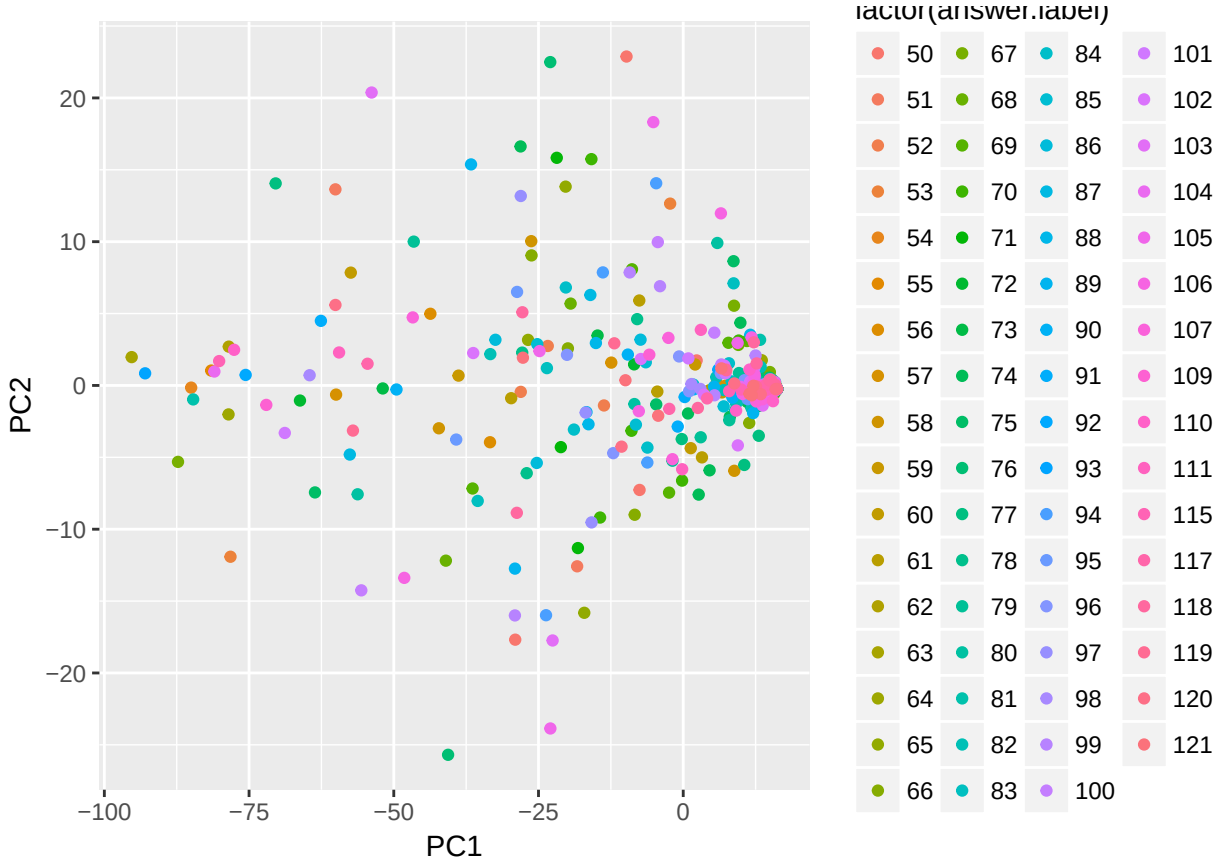
As shown in the previous table, people's answers to the name of maternal grandfather and paternal grandfather are highly correlated. More than 52.6% people call both of them grandpa and 26% people call his/her maternal grandfather grandpa and call his/her paternal grandfather grampa. This two pairs of answer already consist about 80% of responses. Further more, among the people who called his/her maternal grandfather grandpa, 65% will call his/her paternal grandfather grandpa and 31% will call his/her paternal grandfather grampa. We could say by knowing one answer will help with predicting the other answer.



These two questions are also geographically correlated. As shown in the figure above. In the south and north east United States, people will prefer to asked both grandfather as grandpa. In the midwest, people will prefer to call his/her maternal grandfather grandpa and paternal grandfather grampa.

5 Dimension reduction methods

I implement the PCA on the normalized ling_location dataset. Each row represents the summary of each region and each column represent the number of answers.

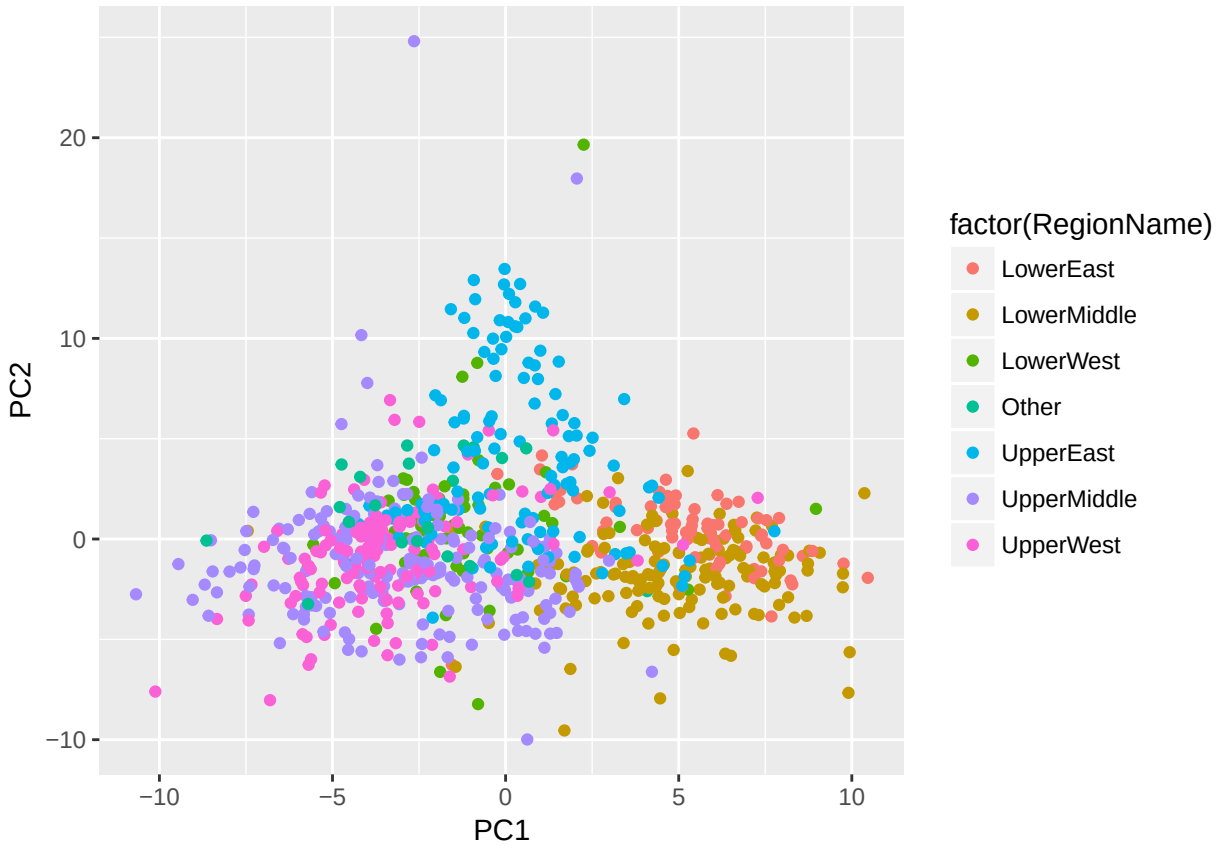


This figure is a little bit hard to figure out the patterns as we can not easily classify the type of questions. But we could use PCA to find the pair of questions that has highest geographic correlation.

```
##      row col
## V171 168 127
## V130 127 168
```

By calculating the distance matrix of the questions in the new space formed by PC1 and PC2, I find that the vector 171 which is the fourth answer to Question 73 is geographically correlated with the sixth answer to the Question 67.

I also transpose the dataset and implement the PCA on the new dataset. Considering we have 781 1 degree * 1 degree regions in the mainland United States, it is imposible to find any interesting pattern with such dataset. So I devided Unite State into 6 major regions. Upper West, Upper Middle, Upper East, Lower West, Lower Middle, Lower East and 1 minor region represent out of mainland.



This figure has some promising results. We could obviously find that the Lower Middle, Upper Middle and Upper East are distinguishable. The Upper Middle are well mixed with Upper West.

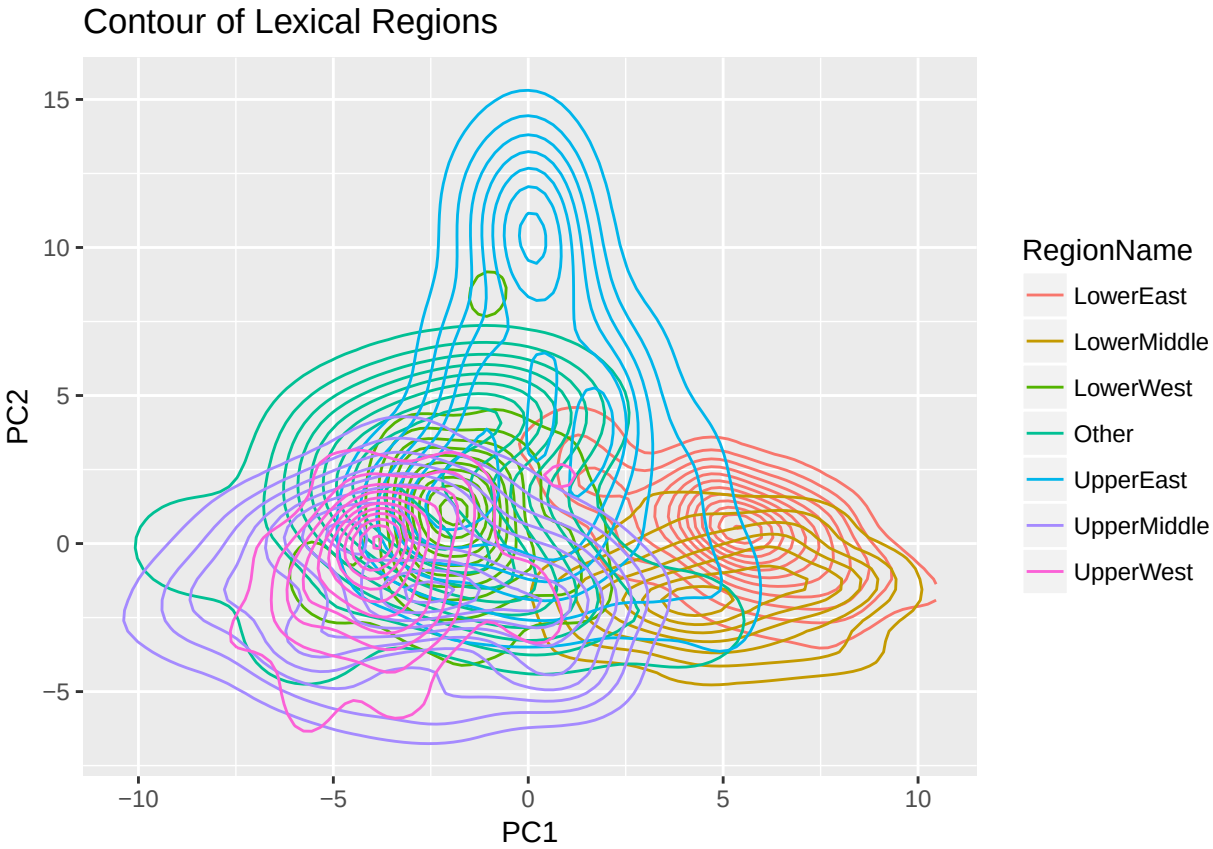
The top six questions that separate the groups are shown in the table below.

```
## [1] 63 93 67 54 81 55
```

The top six questions that produce the continuum are shown in the talbe below.

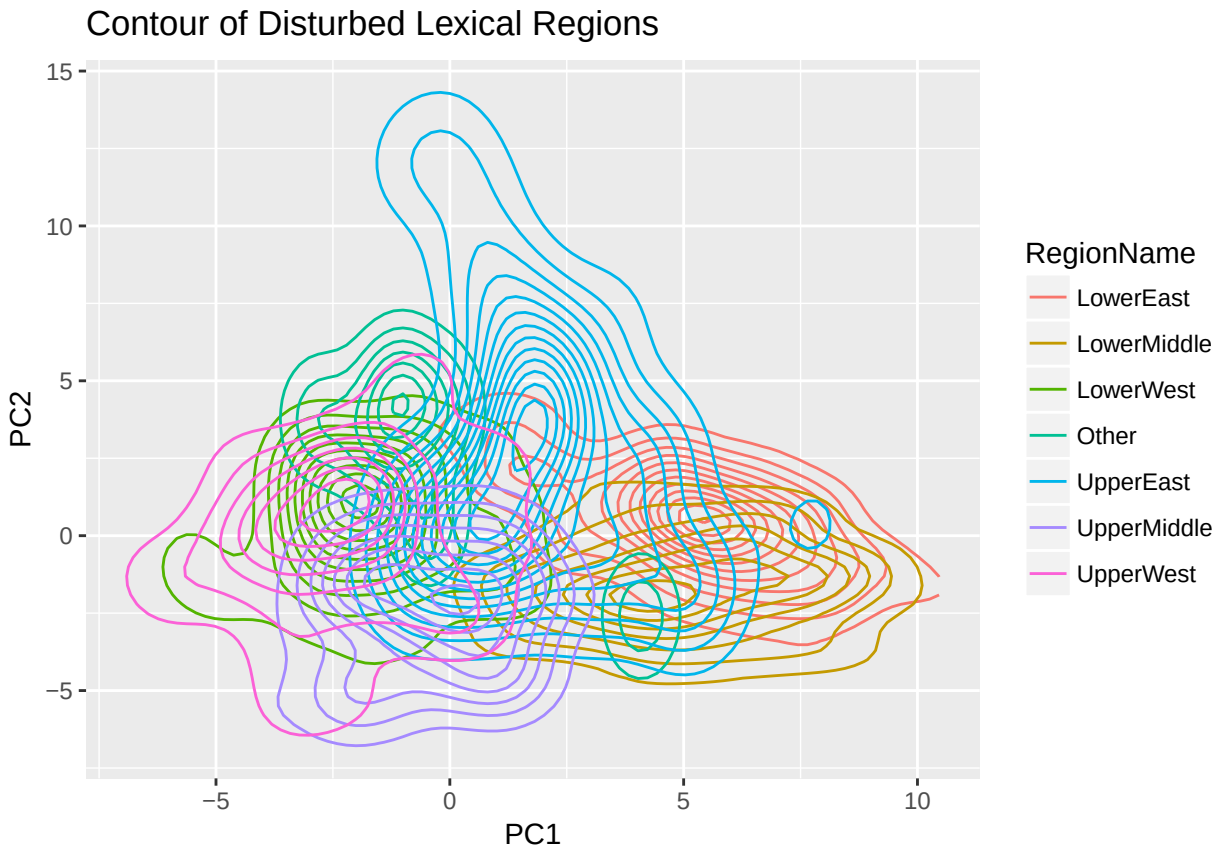
```
## [1] 111  70  91  77  96  74
```

6 Stability of findings to perturbation



This most interesting thing I have found is that the regions are highly geographically correlated. By plotting the contour lines, we are able to find the UpperEast, LowerEast and UpperWest are obviously seperated. We could also find taht the Upper Middle, Upper West and Lower West are well mixed. Meanwhile, the Lower Middle and Lower East are well mixed.

To test its stability, I subsamling about 60% of the original data and the result are shown below. We could confidently say our finding is relatively robusted to the perturb.



7 Conclusion

In this project, I found the lexical differences are highly associated with location. However, the scale of the space resolution is a critical criterion to be discussed. Limited by the sample size, the finer spatial resolution will not contribute to the analysis but mask the patterns. We need to downscale to find the big picture patterns. I seperated the mainland United States into six major regions, I find that we have three major lexical regions. The first region include Upper Middle, UpperWest and Lower West, the second region is Upper East and the last region is Lower Middle and Lower East.