Lab 2
Stat 215A, Fall 2017

SID: 3032126930

October 5, 2017

Plots are at the end of the report.

1 Kernel Density Plots and Smoothing

1.1 Kernel density estimation

If we take a look at the Figure 1, since the data has pretty concentrated values, choice of the kernel function didn't affect much determining the shape of the estimated density. The bandiwdth determined the smoothness of the density graph, but did not change the overall shape of the density much.

1.2 LOESS

If we take a look at the Figure 2, LOESS smoother fits well when I used linear function (blue line on the graph) even with the small span. However, when I used the polynomial with degree 3 (red line on the graph), it overfitted with the small span and even with the large span, it didn't converge to the well-fitted curve. From this, we can learn that the choice of the degree of the polynomials is very important to avoid the overfitting.

2 Linguistic Survey

2.1 Introduction

This part of the report will examine the linguistic data from a Dialect Survey conducted by Bert Vaux. We are going to study lexical differences between different parts of the country. Since there are many questions and many answers for those questions in the survey, we are going to use dimension reduction technique to analyze the data.

2.2 The Data

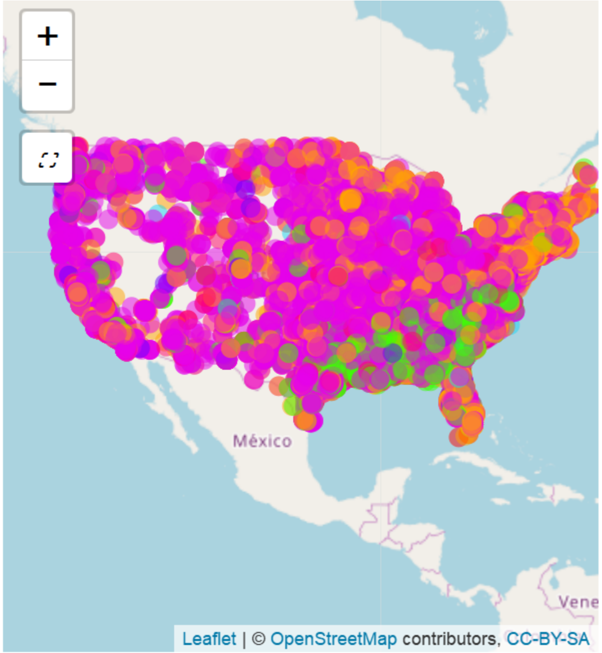
The data we are going to study was found from a Dialect Survey conducted by Bert Vaux. The data was processed accordingly by an intrepid STAT 215 student past. Since we are only interested in lexical differences not the phonetic differences, we are going to study the part of the survey questions.

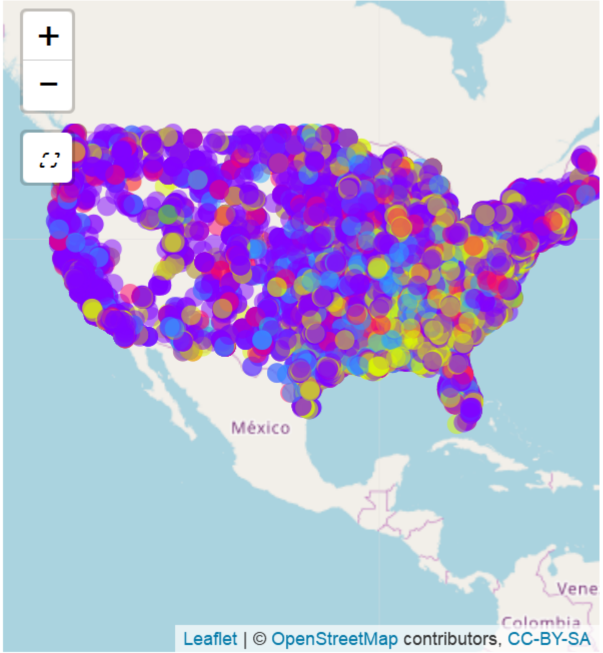
2.2.1 Data quality and cleaning

The dataset is already processed pretty fine, but there are still some issues. What I found out is that the some of the data include states that are not in the country. Since they don't have accurate geological information, I excluded them. Also a few questions in Q50 to Q121 were left out. I had to be careful about that when I used information from the complete dataset for the questions and the answers. Some of the data had NA for latitude and longtitude. since it's hard to get a visualization of them, I also excluded them.

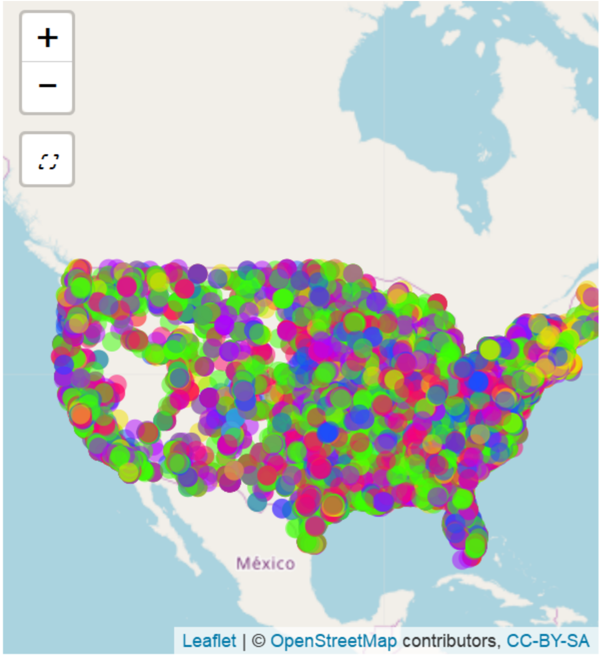
2.2.2 Exploratory Data Analysis

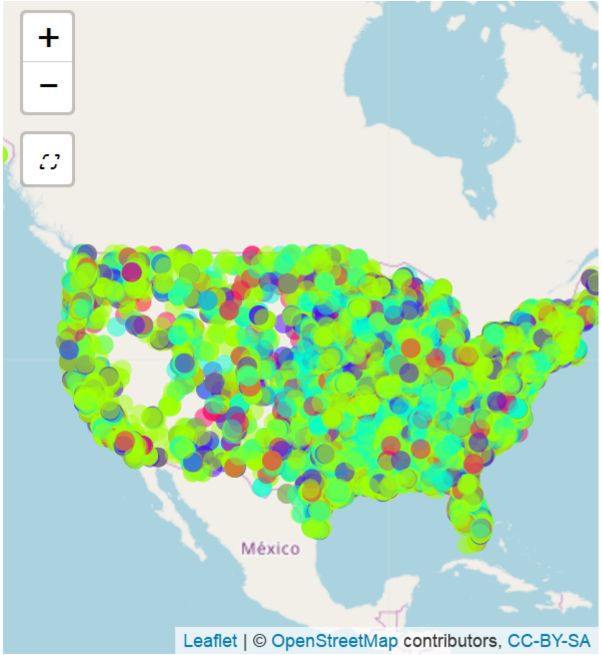
Most of the questions show tendency of most regions dominated by one answer and the small SouthEast regions dominated by another answer in the map. For example, this is the distributions of answers for Q080 and Q100.



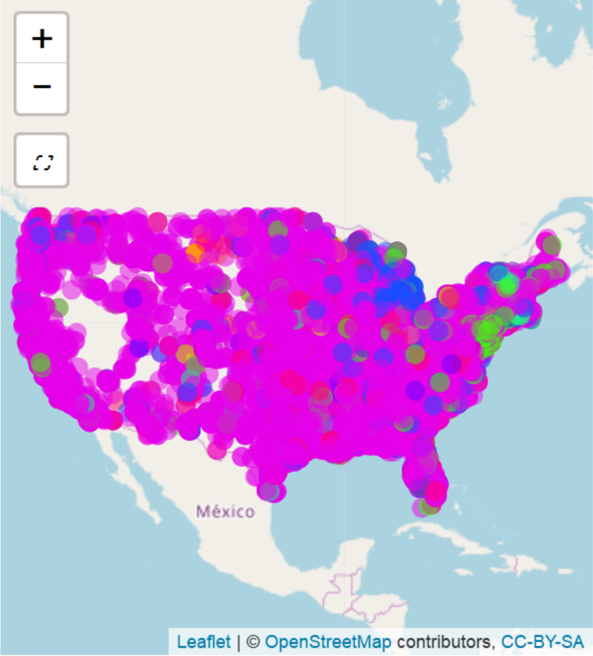


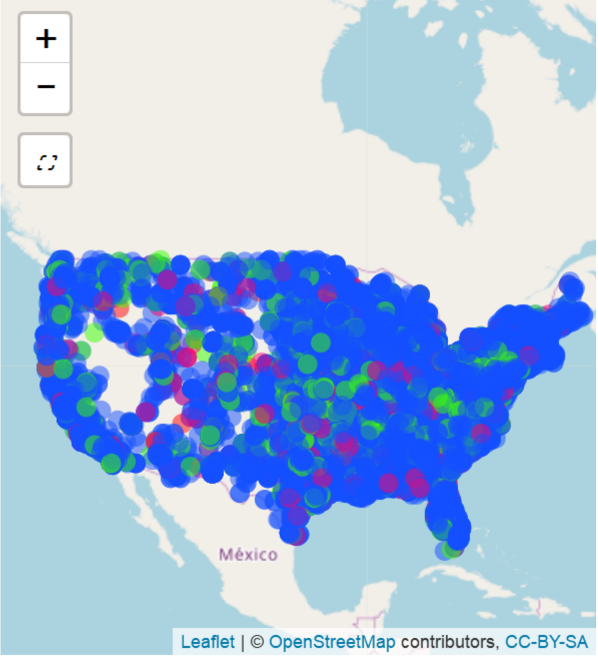
As you can see, SouthEast regions show clearly different color from the other regions. It also makes sense intuitively. Since we are living in one country, there is a specific word that's used by most of the people in the country. If one region uses dialect that is different from the other regions of the country, then that region is most likely to choose different answer for the survey questions.
Of course there are some questions that showed different tendency. For example, Q085 and Q090 showed many different answers spreaded across the country.





Q110 and Q055 had other regions than SouthEast region that showed 2nd popular answer.





But the most of the responses to the survey questions showed the similar patterns.

Since there is a tendency across the data, studying a response to one question will help predict the other questions in the survey.

I comment out the code to produce the map plots, so if you want, you can produce the same results using the code I comment out.

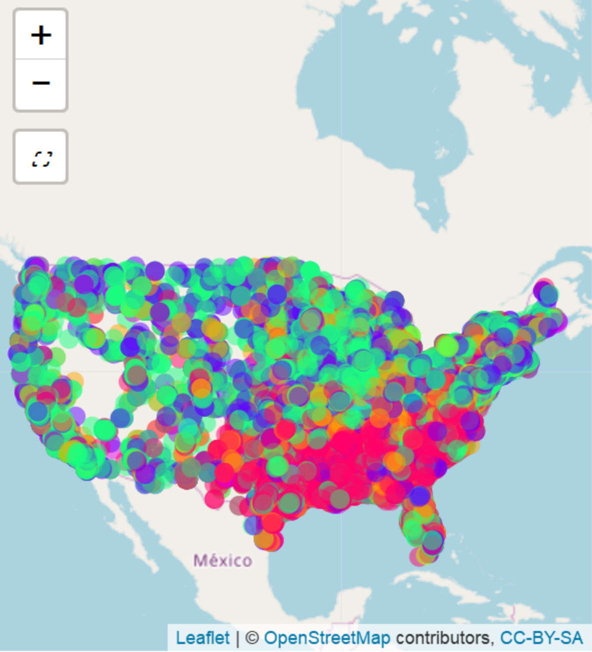
2.3 Dimension reduction methods

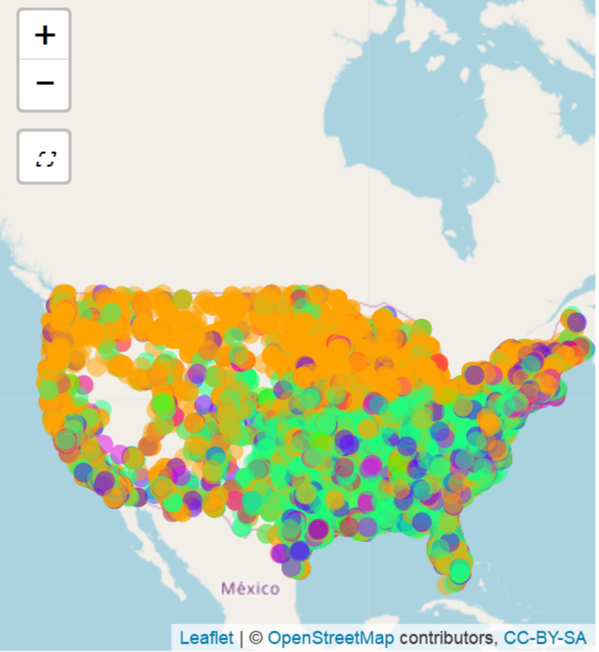
To do the analysis with the whole data, not just with one question at a time, I first changed data that had categorical responses to have binary responses and performed PCA on it to reduce the dimension of the data. After that, to determine how many principal components to use, I made the screeplot, which is Figure 3. However, if you take a look at Figure 3, we need much more than 2 principal components to successfully catch most of the variance of the data. But for the visualization purpose, I only used first two principal components.

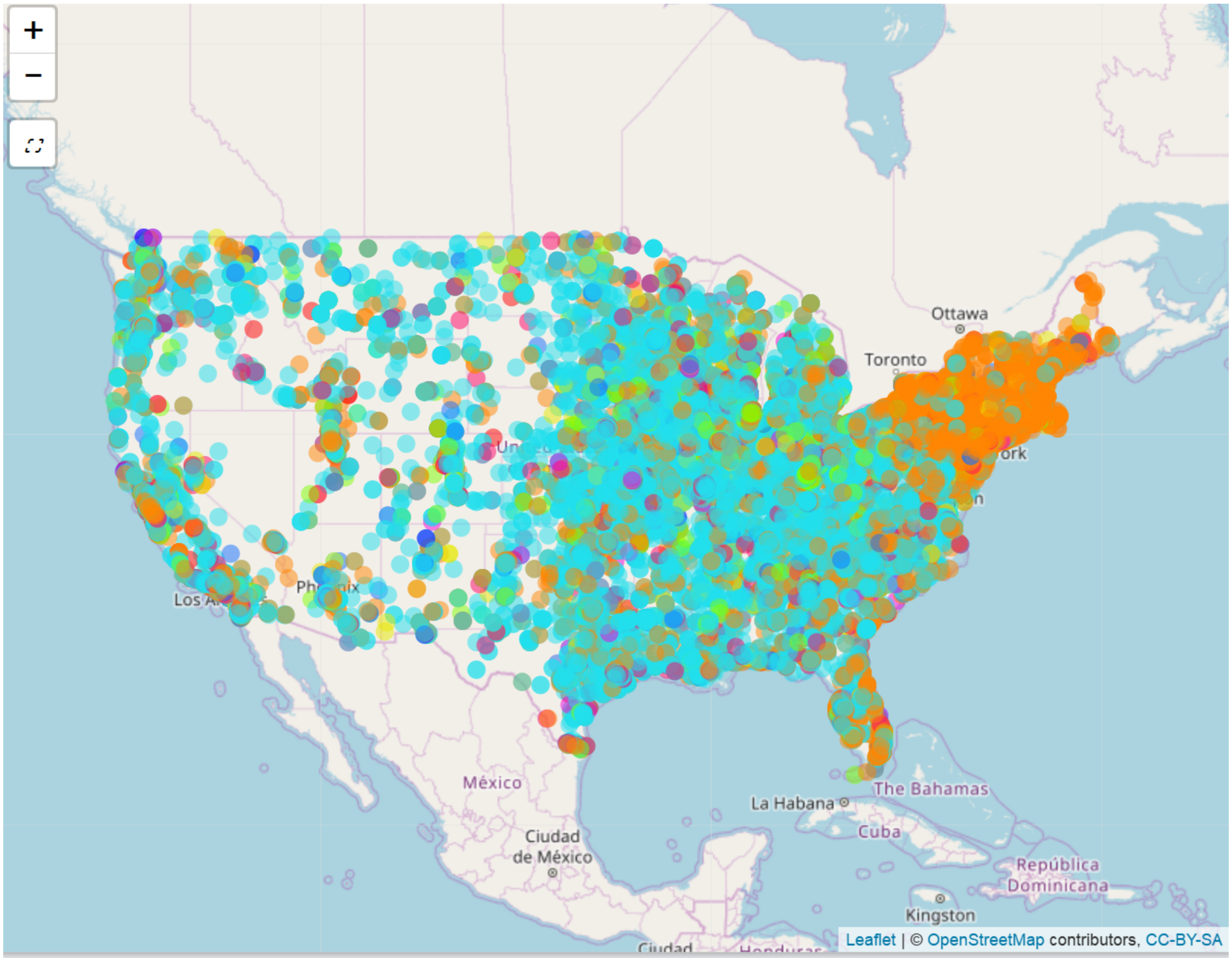
After that, to check how the geological factors affect the responses for the survey, I gave the colors to the dimension reduced data according to the latitude, longitude and zip code. I first guessed the longitude and the latitude information will explain the lexical differences between people. However, the plots given color according to the longitude and latitude didn't show clear explanation. Instead, when I gave color using zip codes to the plot using first two principal components, I could see the clear gradients of color throughout the data. Figure 4 contains the plot of this. The color changes smoothly as zip code goes bigger, which shows continuum. It means that the lexical difference between people is explained by zip code. Actually, this agrees with the exploratory data analysis we did on section 2.2.2 using plots. Zip code starts from SouthEast regions of the country and gets bigger as it goes to the Northwest regions of the country.

Then I tried the K-mean clustering to check clusters of data are related to the change of zip code. I used k = 2 since most of the plots we saw on the exploratory data analysis section showed two dominating answers for each question. Figure 5 shows the result of the clustering compared to the change of zip code. Sometimes, due to the random starting points, it gives wrong clusters, but most of the time, it gives clusters that are divided accordingly with the zip code as we expected.

To check which questions separated the groups, I plotted principal component scores. Figure 6 shows the plot of principal component scores. From the plot, we can see binary category 9, 198, 165, 170 affect mostly to the first two principal components. Q050 contains the binary category 9, Q076 contains the binary category 198, and Q073 contains both binary categories 165 and 170. So, I plotted Q050 vs. Q076 which affect mostly to the second principal component and plotted Q073 by itself which affects mostly to the first principal component.



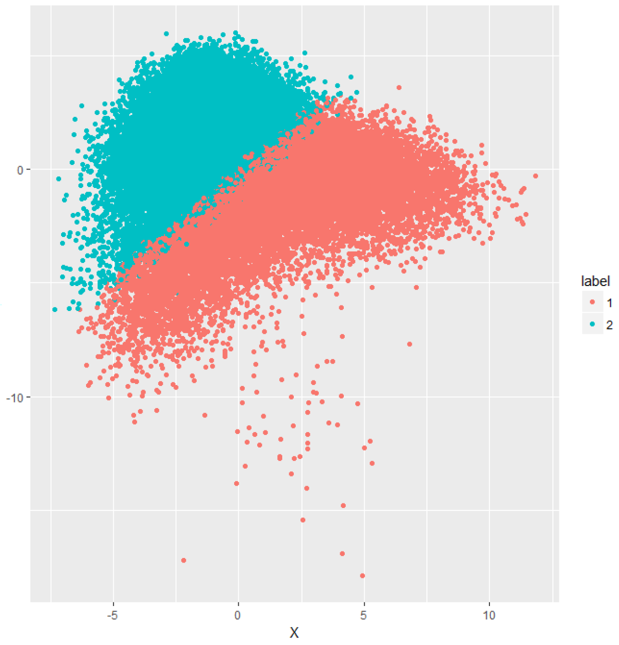


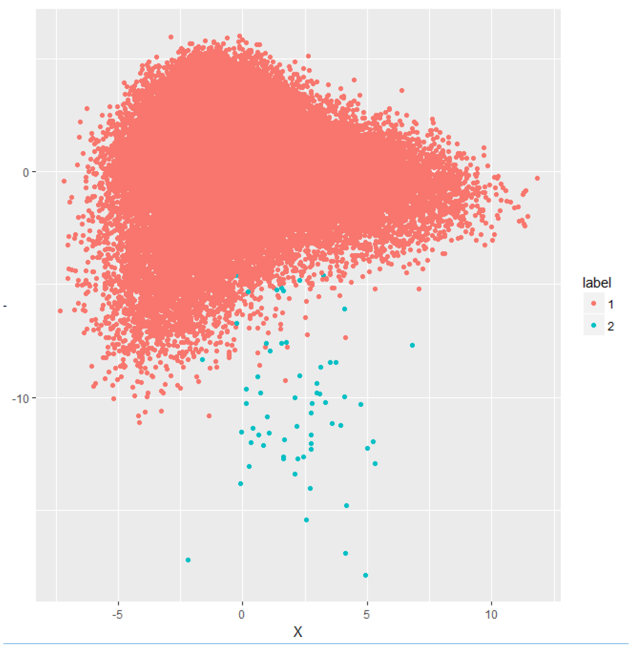


These questions show clearer distinction between two dominating answers than other questions. Q050 and Q076 show clear distinction between SouthEast part of the country and Northwest part of the country. Q073 shows small but clear distinction of the response at the East end of the country.

2.4 Stability of findings to perturbation

As I already mentioned above, k-mean cluster algorithm that I used to do the clustering uses different random starting points everytime and sometimes it gives bad clustering. Followings are two different clusterings that the algorithm with k = 2 gives:

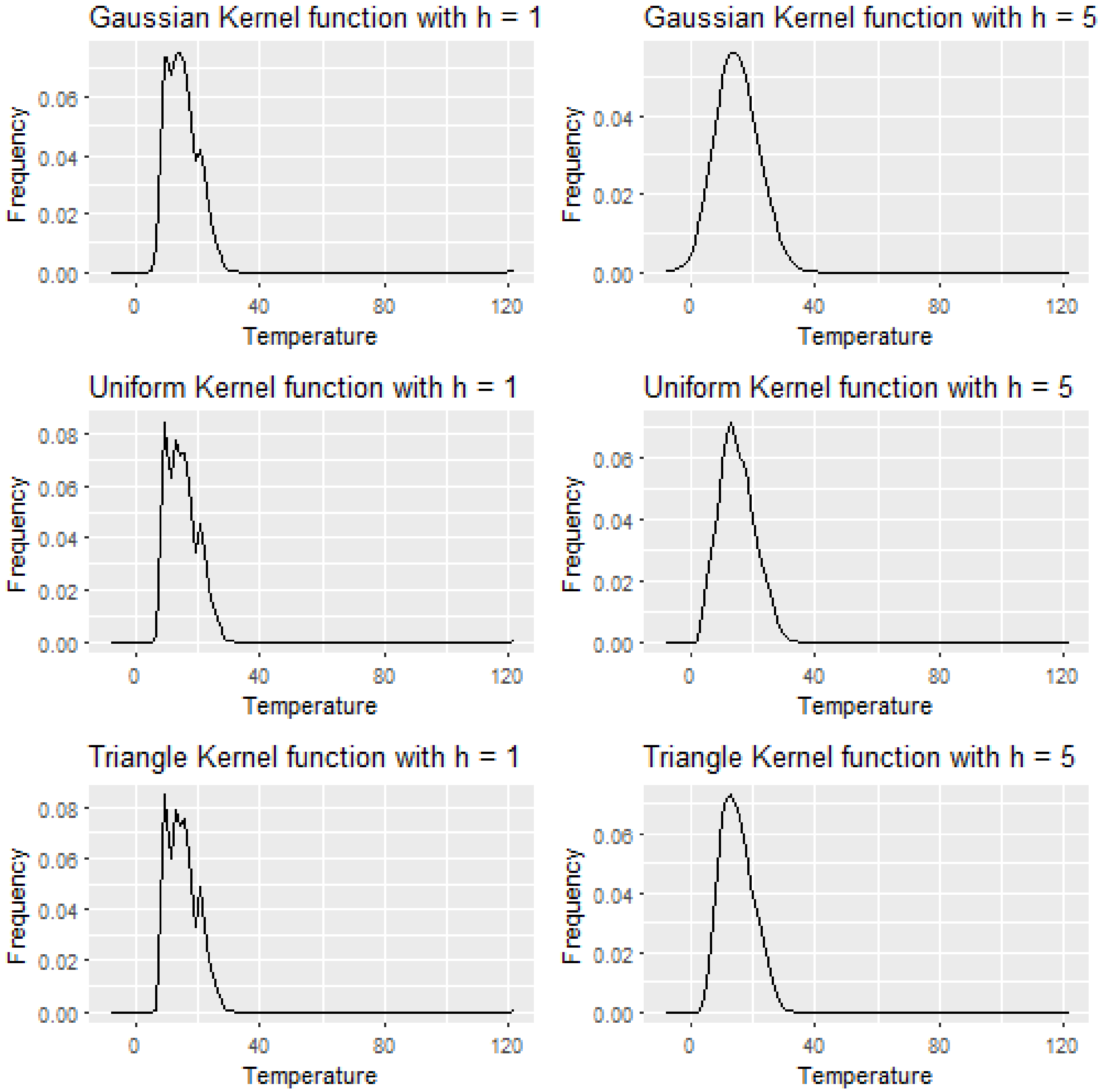




Also, I perturbed the dataset by subsampling it and tried to do PCA with it to check whether the perturbed dataset still shows smooth continuum with zip code. However, the covariance matrix of the subsampled data is often positive indefinite and is not able to be decomposed. When it decomposes successfully, it shows smooth change as we wanted, but it does not happen everytime. Therefore, the analysis I did in this report is not stable enough.

2.5 Conclusion

In this report, an analysis of the lexical difference across the country was performed. The data contained large dimension and the PCA, which is the dimension reduction algorithm, was conducted to deal with the large dimension issue. As a result of PCA, I could do visualization of the dimension reduced data and found out that the lexical difference across the country is described by the change of the zip code. Also the data clustered using k-mean algorithm withouth the zip code information often divided successfully agreeing with the change of the zip code. The analysis I performed on this report is not stable, but with the large enough data and well-tuned starting points, I could acquire the analysis that explains data well enough.



Gaussian Kernel function with h = 1

Gaussian Kernel function with h = 5

Uniform Kernel function with h = 1

Uniform Kernel function with h = 5

Triangle Kernel function with h = 1

Triangle Kernel function with h = 5

Figure 1: Kernel density estimations using different kernel functions and bandwidths

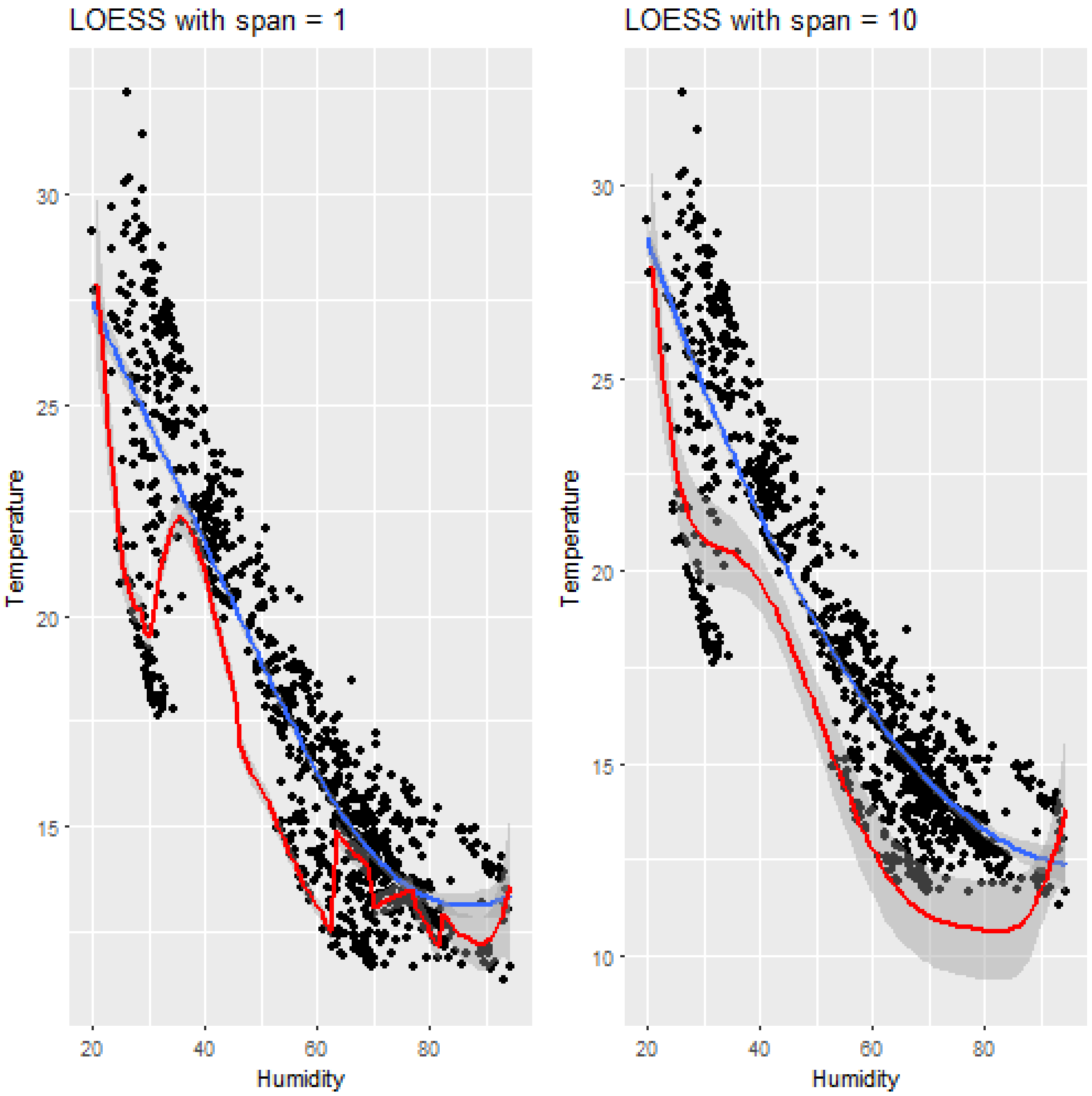


Figure 2: Temperature vs. Humidity at the same time of day and LOESS using linear(blue) and 3rd degree polynomial(red)

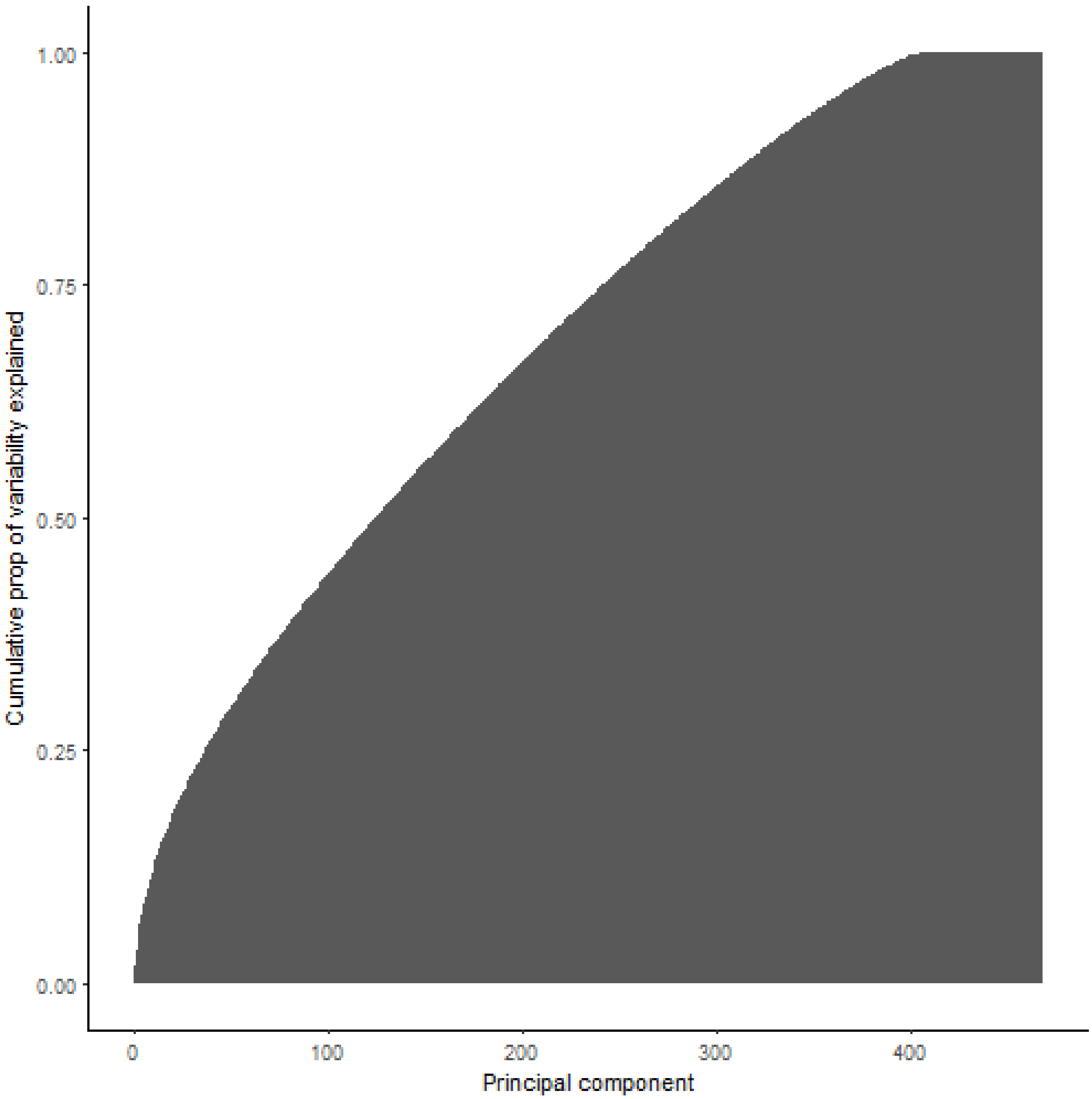


Figure 3: Screeplot of the linguistic data PCA

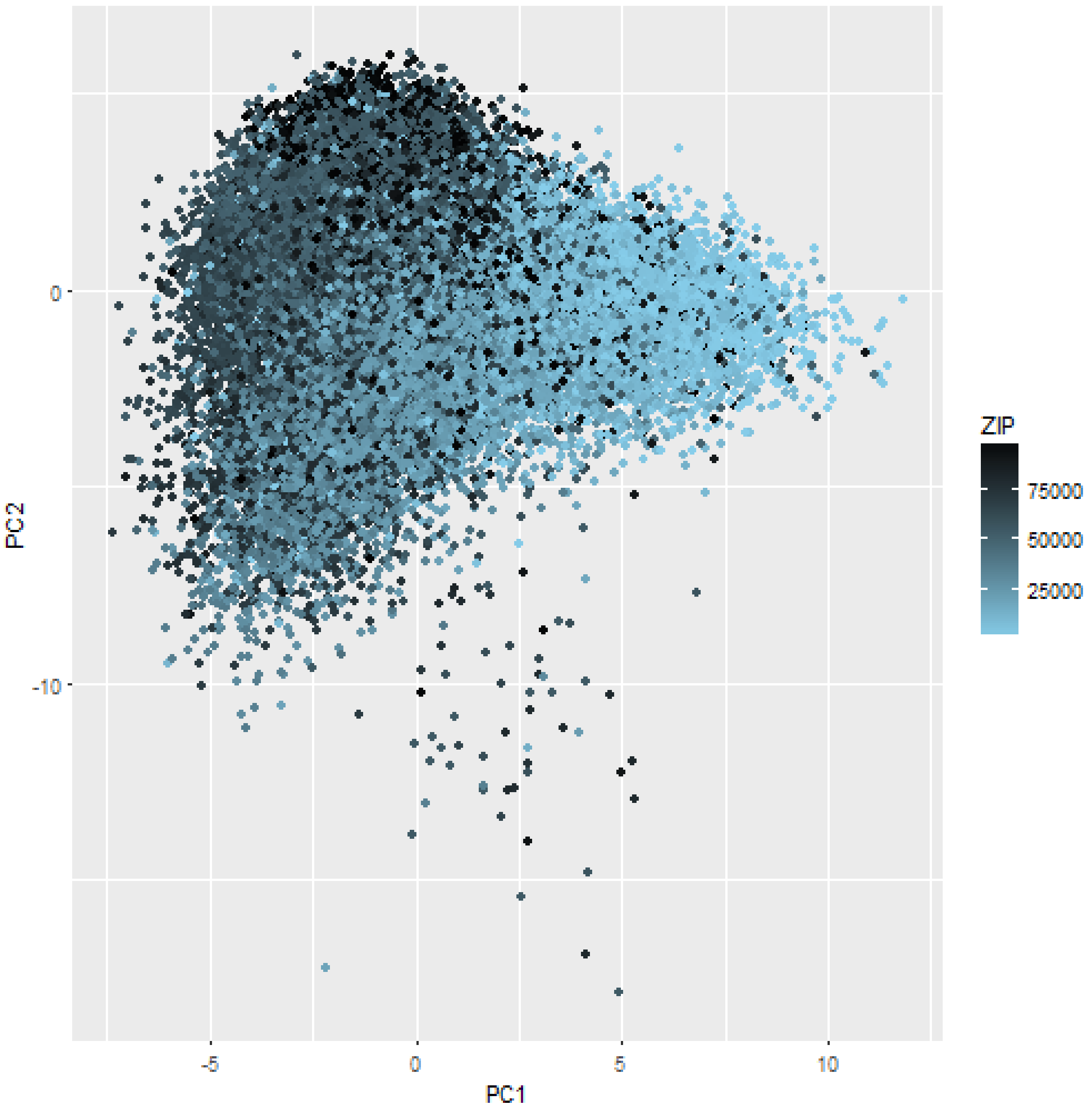


Figure 4: Principal Component 2 vs Principal Component 1 with color changing according to Zip code

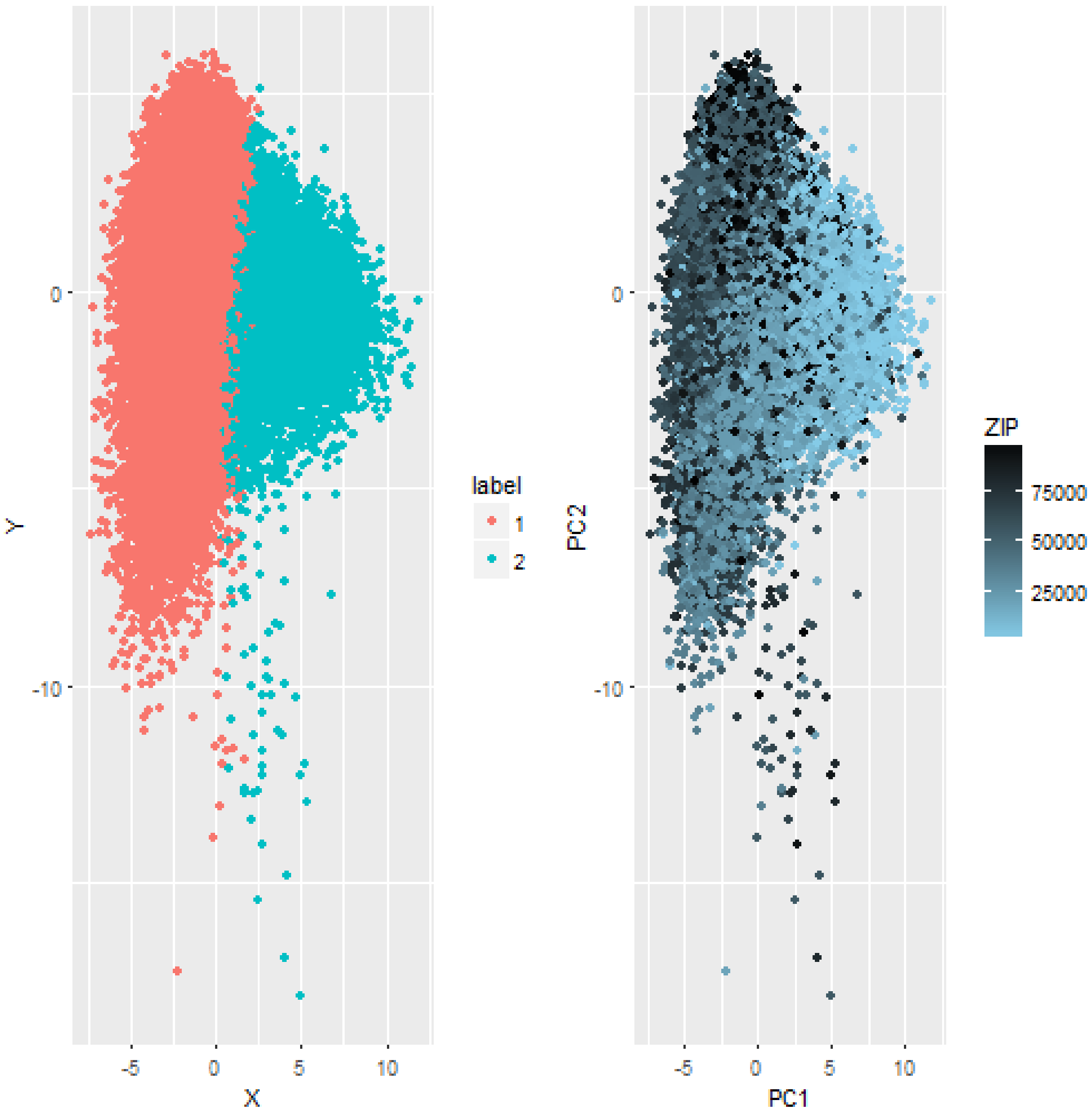


Figure 5: Clustering data projected onto the first two principal components vs Fig.4

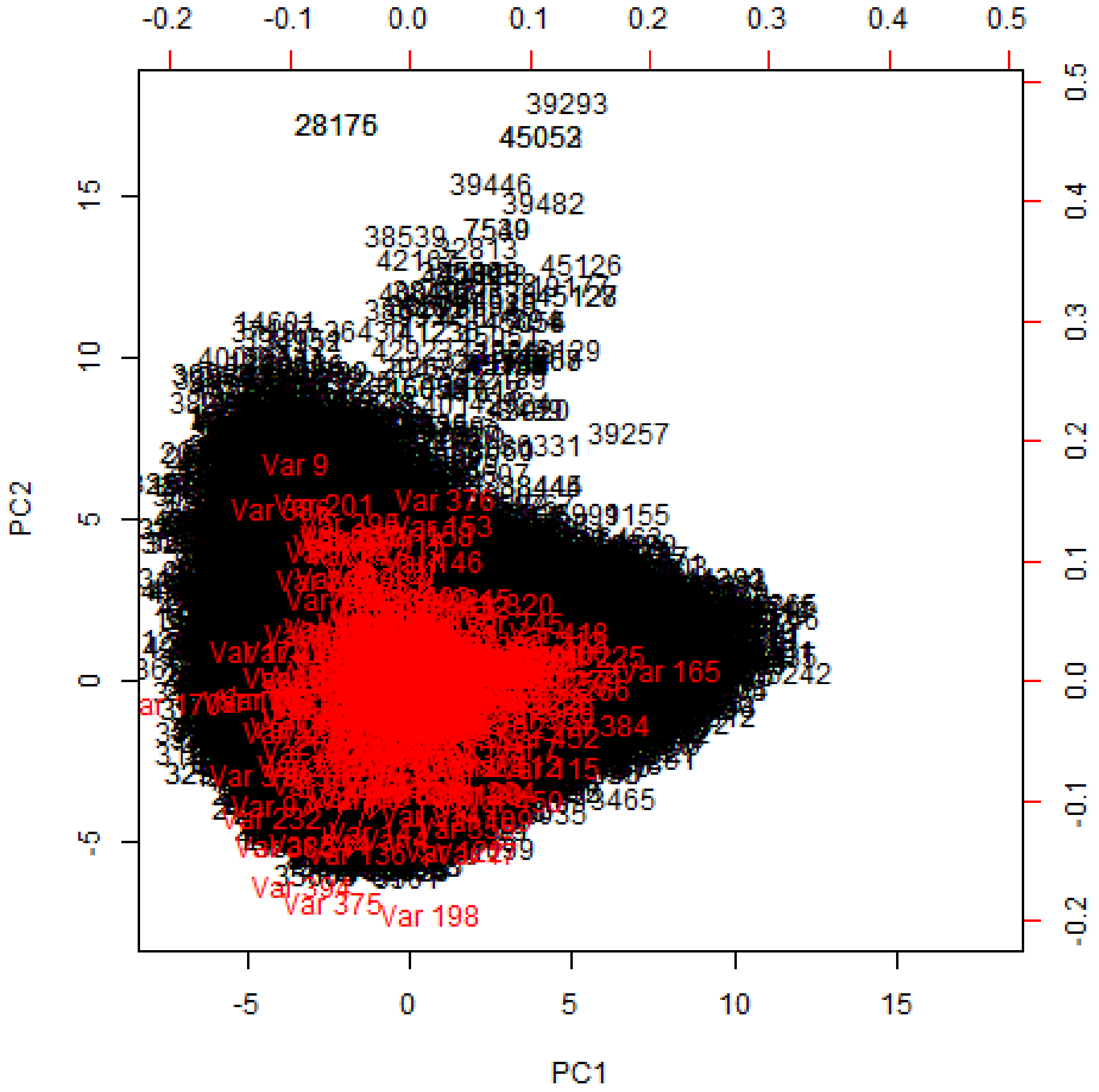


Figure 6: Principal Component Scores