Lab 2 - Linguistics Data, Stat 215A, Fall 2021

October 07, 2021

1 Introduction

The United States of America is a highly diverse and physically expansive country. San Francisco is more distant from New York City than Moscow is from Cairo. In addition, different regions of the country developed at much different times than others. Phoenix, AZ, is the fifth most populous city in the U.S. and it was founded only 140 years ago. At the same time, New York City was founded almost 400 years ago. Their respective cultures were formed by very different people at very different times.

One of the more obvious differences between regions in the U.S. can be found in vocabulary. This follows from the fact that different languages have influence in various regions. In the American Southwest, Latin American Spanish is almost an official language. In the Northeast, Western European countries have more sway. It is only logical to try to quantify these differences.

Such is our goal in this report. We are provided with the results from an online, self-reported, vocabulary survey. Before taking the survey, the user enters information regarding their location. Then, for each of the 67 questions, the user selects one of the options presented to them. The questions are along the lines of "What do you call it when rain falls while the sun is shining?" Potential answer options include "the wolf is giving birth", "the devil is beating his wife", and "monkey's wedding". Ideally, a user's response to that question would give us insight into where they are from.

In order to make that connection, we will use unsupervised learning techniques. Specifically, we will use k-means and non-negative matrix factorization clustering. If we are able to group users into geographically significant clusters, then we will have proven that this type of data contains useful information. However, the data must first be subject to dimension reduction. It is unlikely that every single variable will be informative, so techniques like PCA will give us more tractable data that is ideally just as useful.

2 The Data

We will almost exclusively use the raw linguistic data frame. It contains information for 47,471 users across 73 variables. The first contains the unique ID of each user. The second, third, and fourth columns present the user-supplied city, state, and zip code, respectively. The last two represent calculated latitude and longitude values calculated from each user's location information. Each of the remaining columns corresponds to a question in the survey. The values in each column represent the answer choices that users provided when prompted with that question.

Below is a description of the structure of this data set. Many columns similar to Q050 and Q121 have been omitted for the sake of visual presentation.

```
## [1] 47471    73
```

```
## 'data.frame':    47471 obs. of  8 variables:
##  $ ID   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ CITY : chr  "Boise" "Pittsfield" "Burlington" "Easton" ...
```

```
##  $ STATE: chr  "ID" "MA" "VT" "PA" ...
##  $ ZIP  : int  83704 1201 5401 18042 1730 77479 2066 21044 56150 1033 ...
##  $ Q050 : int  4 4 4 7 8 8 7 4 7 7 ...
##  $ Q121 : int  3 3 1 1 1 3 3 3 1 0 ...
##  $ lat  : num  43.6 42.5 44.5 40.7 42.5 ...
##  $ long : num  -116.3 -73.3 -73.2 -75.2 -71.3 ...
```

Another useful data set contains the actual questions. Here is a single row as an example of how the data is formatted:

```
##    qnum                                                      quest
## 80   80 What do you call it when rain falls while the sun is shining?
```

Here is the corresponding entry in the data set containing the answer choices:

```
##   qnum ans.let   per                                    ans
## 1   80       a 34.29                              sunshower
## 2   80       b  0.04             the wolf is giving birth
## 3   80       c  6.43         the devil is beating his wife
## 4   80       d  0.16                       monkey's wedding
## 5   80       e  0.15                          fox's wedding
## 6   80       f  0.03                        pineapple rain
## 7   80       g  0.74                           liquid sun
## 8   80       h 55.15 I have no term or expression for this
## 9   80       i  3.02                                other
```

Now that we have a decent idea of what has been given to us, it is time to whip the data into shape for further analysis.

2.1 Data Cleaning

2.1.1 Removing bad values

How many NA values are in each of the five data sets? The table below tells us that only the linguistic data set contains NA entries. We output the summaries for each column and see that the latitude and longitude columns are the main issue. The remaining 3 NA values come from the "STATE" column, which is populated by responses from the users. It is highly likely that a few respondents concerned with privacy simply typed "NA" when prompted for their location.

ling_data	loc_data	quest_mat	quest_use	answers
2043	0	0	0	0

```
##       lat             long
##  Min.   :19.23   Min.   :-161.42
##  1st Qu.:37.32   1st Qu.: -96.16
##  Median :40.34   Median : -87.63
##  Mean   :39.45   Mean   : -90.48
##  3rd Qu.:42.24   3rd Qu.: -77.28
##  Max.   :71.30   Max.   : -67.00
##  NA's   :1020    NA's   :1020
```

Because the location of each response will be important to later analysis, we will simply remove the entries corresponding to NA values. There is no discernible difference between the removed entries and the ones that remain.

In addition, a quick glance at a globe tells us that the contiguous United States roughly lies between -125 and -66 for longitude, and 24 to 50 for latitude. We will eliminate responses outside this range because we are interested in U.S. speech patterns. It is likely that people from outside the US took this test. We are not interested in their responses.

In total, we remove 1226 rows from the linguistic data: 1020 due to NA and 206 due to falling outside the boundaries. The dimension of the resulting data frame is

```
## [1] 46245    73
```

We aren't using all of the questions, so we will use the data set containing question information to get the indices of the questions of interest.

2.1.2 Binary encoding

We will one-hot encode the linguistic data so that the responses are binary instead of categorical. The reasoning for this is provided in a later section. Below, we see the dimension of the new binary data frame. There are 468 columns, which is how many we are supposed to have according to the lab instructions. The number of rows remained constant. In addition, we can see the summary statistics for the first 4 columns. These are included so that the data cleaning can be verified.

```
## [1] 46245   468
```

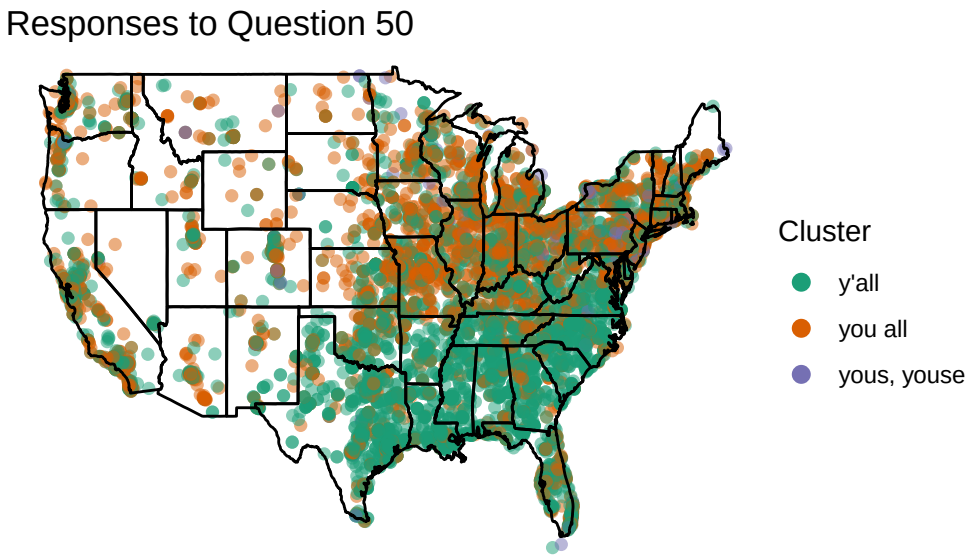
```
##      Q50.1            Q50.2              Q50.3             Q50.4
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.000000   Median :0.00000   Median :0.0000
##  Mean   :0.1231   Mean   :0.006293   Mean   :0.00147   Mean   :0.4082
##  3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000000   Max.   :1.00000   Max.   :1.0000
```

2.2 Exploratory Data Analysis

In this section, we will analyze three questions from the data set and see if we can see any geographical patterns. Furthermore, we will keep an eye out for any relationships between questions.

The first question is as follows:

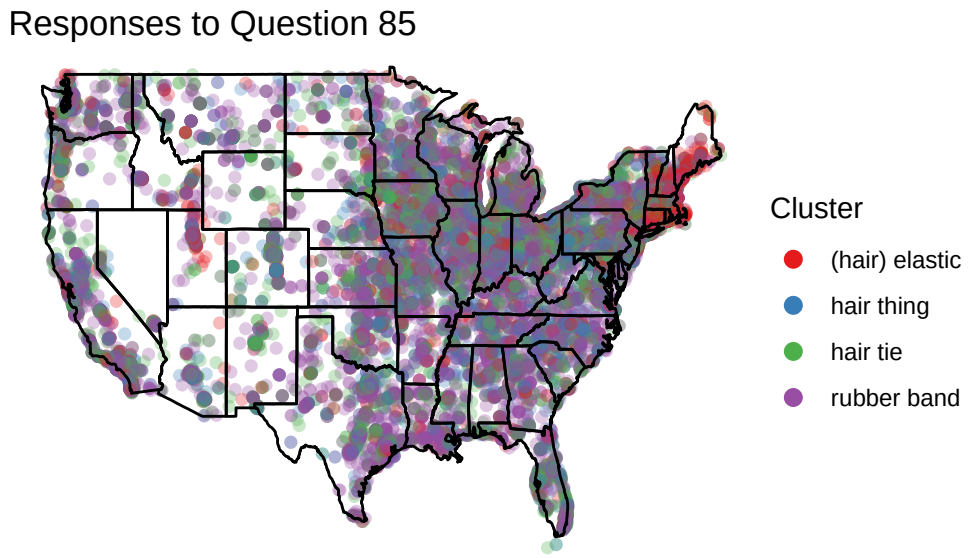
```
## [1] "What word(s) do you use to address a group of two or more people?"
```



Many of us probably guessed where the word "ya'll" is most often used. However, the fact that it's almost the only choice selected by users form that region is surprising. Furthermore, we can see the word "yous" being utilized around Philadelphia, PA. This result was unexpected.

Next, we analyze:

```
## [1] "What is the thing that women use to tie their hair?"
```



The results of this question would likely be hard to guess for anyone. The interesting pattern that we can observe in this visual is the concentration of users who say "elastic" in the Northeast. It would be difficult to determine exactly why that is, but it is likely that "elastic" is the term used in European nations like England.

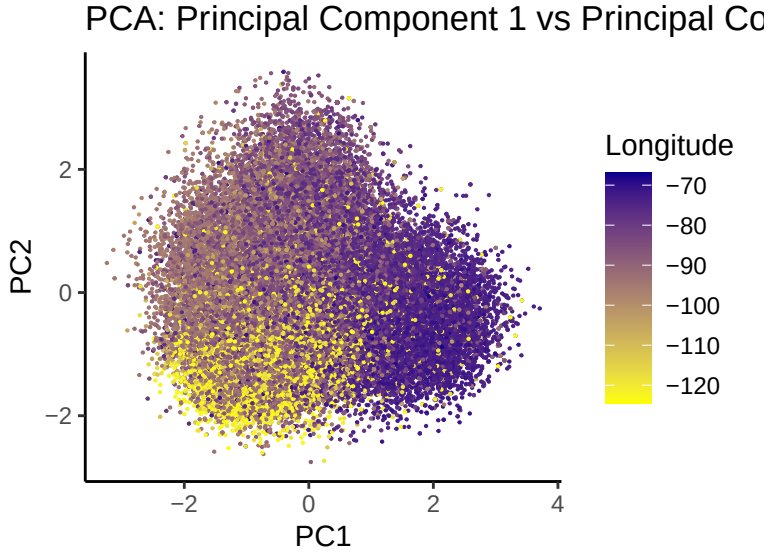
A similarity between the two data sets can be observed along the Gulf Coast. While most of the country lack definition when it comes to these questions, users from that area only use "ya'll" and "rubber band". If a large group of users seemed to only use those terms, we would have evidence suggesting that they are from the Gulf Coast region.

3 Dimension reduction methods

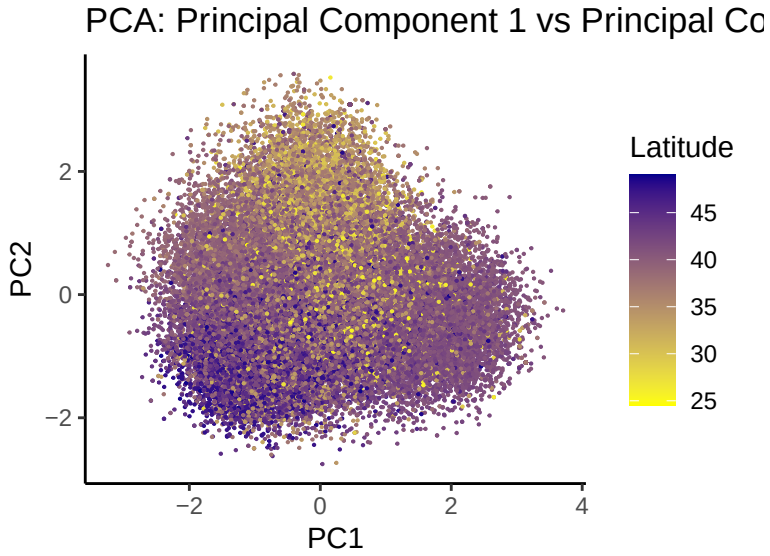
3.1 Principal Component Analysis (PCA)

When conducting principal component analysis, we choose to center but not scale the columns. We choose to center the data because we need the first principal component to be proportional to the maximum variance of our original data. Otherwise, the mean of the data set may be treated as an important variable. The columns are not scaled because we are more interested in columns with higher variance. Without scaling, PCA will prefer columns with more variance.

Those are the columns we are interested in because they suggest that there are many different responses to the question. Low variance suggests that either everyone selected that answer or nobody did. Either way, that column won't be very helpful when it comes to clustering.



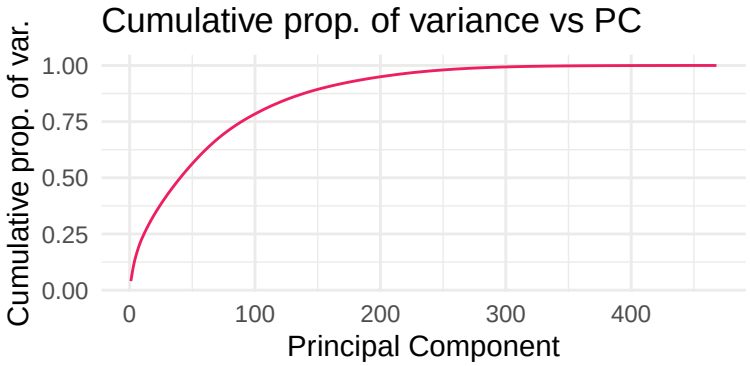
We can already see a pattern forming when comparing the first two principal components. Users in the Western United States receive lower PC1 and PC1 scores. Users in the Eastern Unites States clearly receive much higher PC1 scores. This tells us that PCA has discovered sources of variance that will facilitate later classification based on region.



The differences aren't quite as strong, but we can see that there is also a difference between users in the Northern and Southern United States. This time, users in the Northern U.S. receive lower PC1 and PC2 scores. Southern U.S. users receive higher PC2 scores. Putting the two plots together, we can already see three rough clusters forming: Northwest, mid-East coast, and Southeast U.S.

3.1.1 Eliminating components

In order for data reduction to be successful, we have to eliminate some of the principal components that we calculated. The following plot shows that the first 250 principal components describe nearly all of the variation. We have successfully reduced the number of variables from 468 to 250.



It would be a bad idea to perform PCA on the original data because it is categorical. This is a problem because it is difficult to compute the "distance" between different users and variables. In other words, it is much harder to quantify how different they are. The core idea of PCA is reducing variance, and it's not as clear how to interpret the variance of categorical variables.

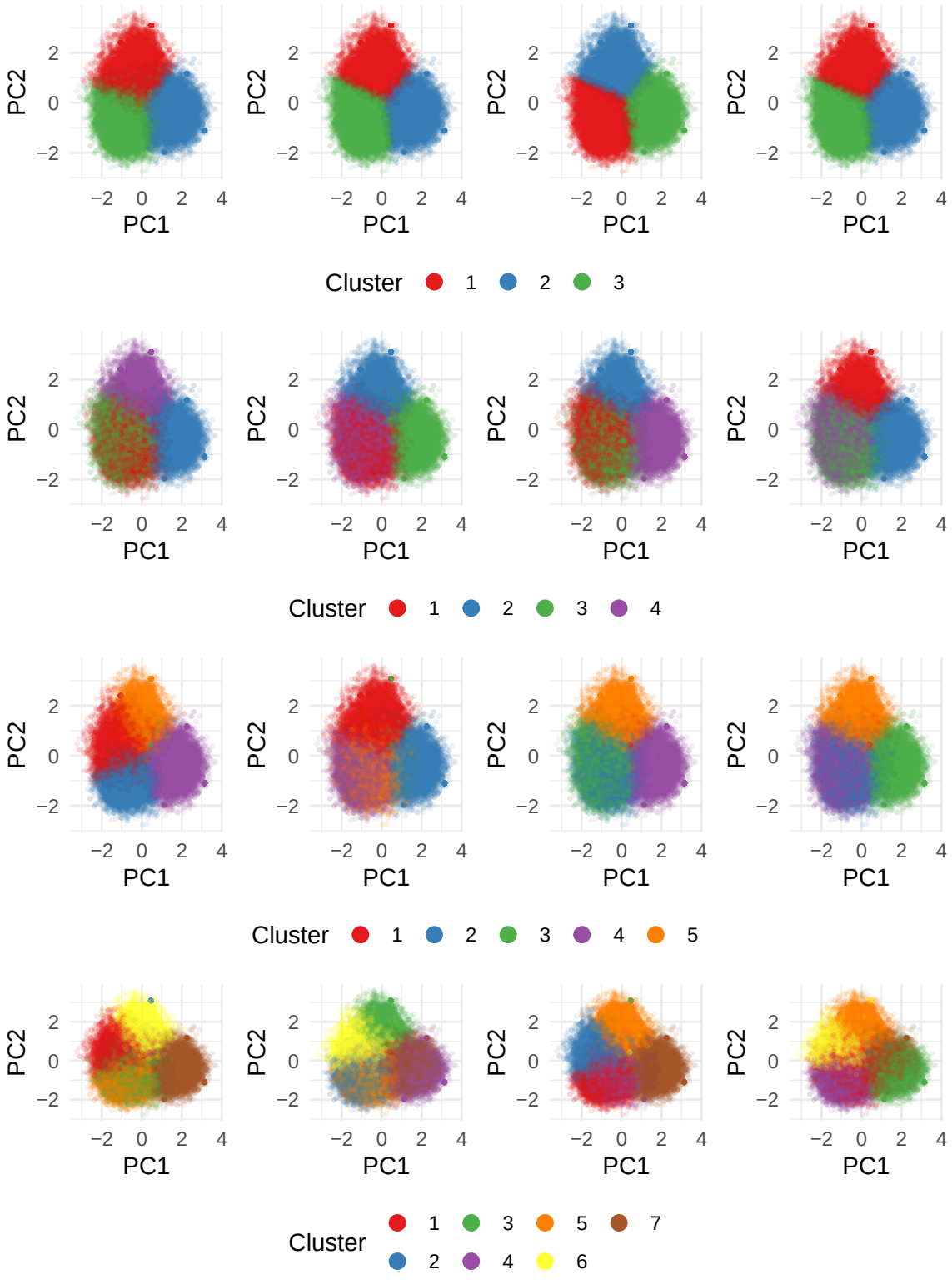
4 Clustering

We finally begin our statistical analysis. As mentioned in the introduction, we will use multiple supervised learning techniques to draw conclusions from our data: k-means clustering and non-negative matrix factorization. ## k-Means

We will perform k-Means clustering on the reduced data because this algorithm is known to perform poorly with high-dimensional data. We claimed to see three rough clusters after conducting PCA, so let's include $k = 3$ in the set of potential numbers of clusters. We will then take another look at our PCA plot and see if our guess was correct.

4.0.1 Choosing number of cluster centers

To select an accurate number of cluster centers, we will test a variety of values for stability. Below, we see multiple plots for each potential k. If all plots for a single k look similar, then we know it is likely more stable and truer to reality.

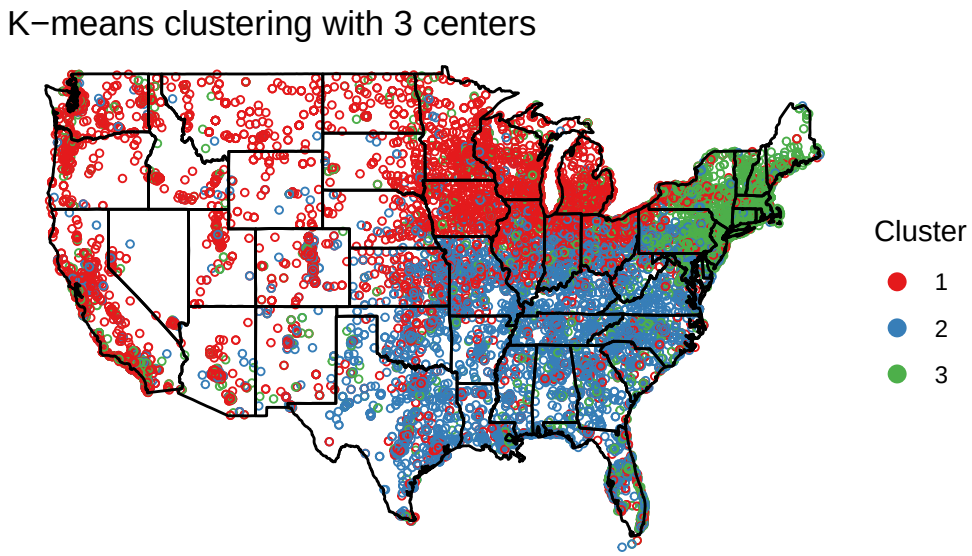


4.0.2 Geographic meaning

We can see that our predictions in the PCA section were correct. The data is most frequently clustered into three groups: Northeastern U.S., Southeastern U.S., and West Coast plus Midwest. In addition, using k=4

is often able to differentiate between the West Coast and the Midwest, but those results are not as stable. In the above plots, it can be seen as a wedge in the upper left side of the mass.

The following shows the data grouped into three categories and plotted on a map of the United States.



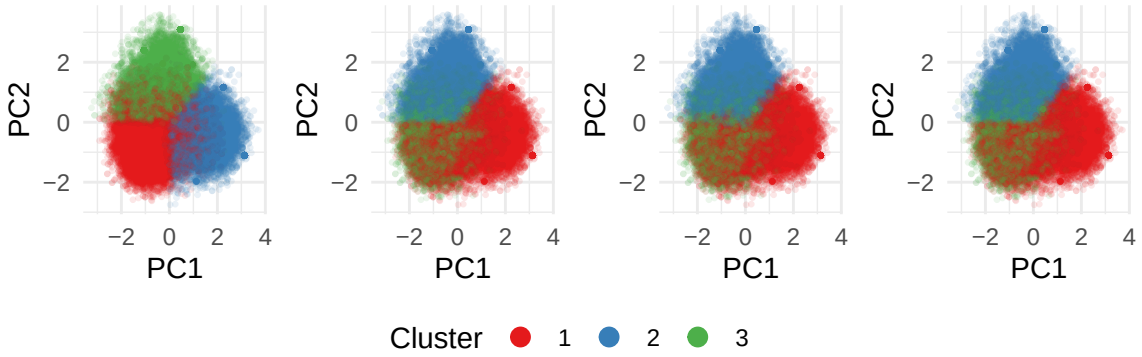
The strongest clusters correspond the the Southeast and Northeast. This makes sense because those regions have had a continuous culture for much longer than other regions of the U.S. Their way of life has endured in the same regions for centuries. On the other hand, much of the Western U.S. was settled relatively recently by a highly diverse assortment of people. As a result, it is harder to confidently identify someone from that area from speech alone.

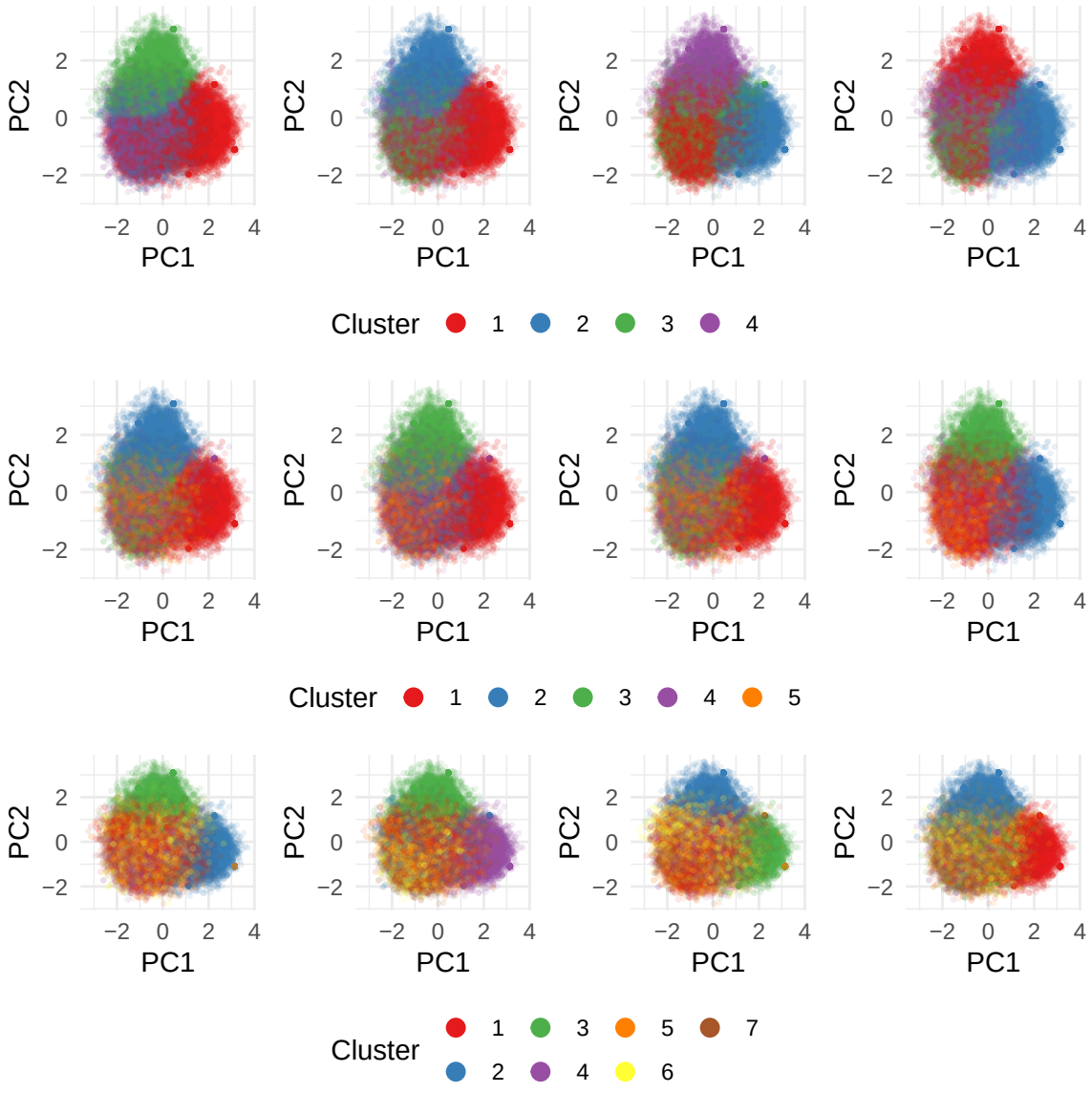
4.1 Non-negative matrix factorization (NMF)

To verify our results from k-means clustering, we will also run NMF on our reduced data.

4.1.1 Choosing number of cluster centers

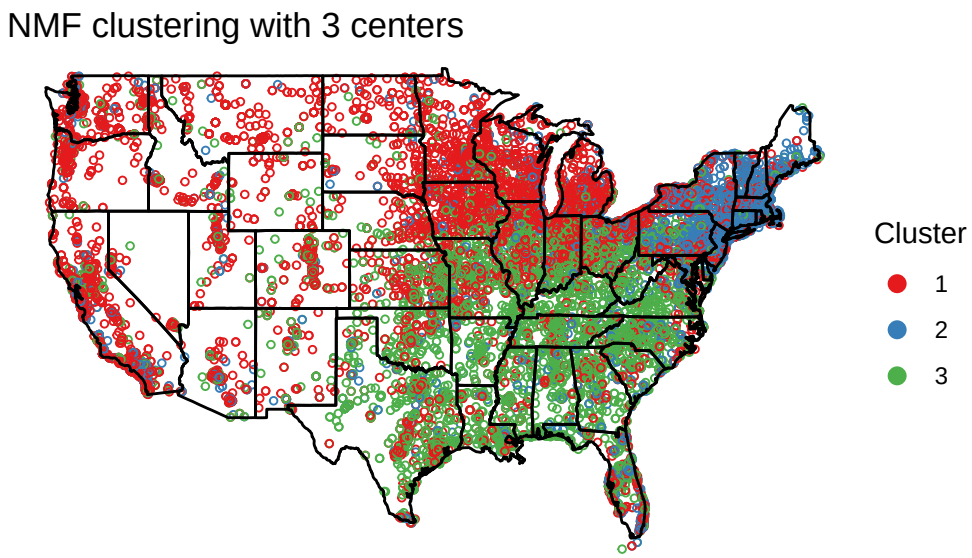
The process for choosing the number of cluster centers is almost identical to the one used in k-means clustering. Once again, we are looking for the value of k that provides the most stable predictions.





Again, k = 3 seems to be the winner. Even in plots with more clusters, the same 3 always make an appearance. This gives us a high degree of confidence in the generalizability of our clusters.

4.1.2 Geographic meaning



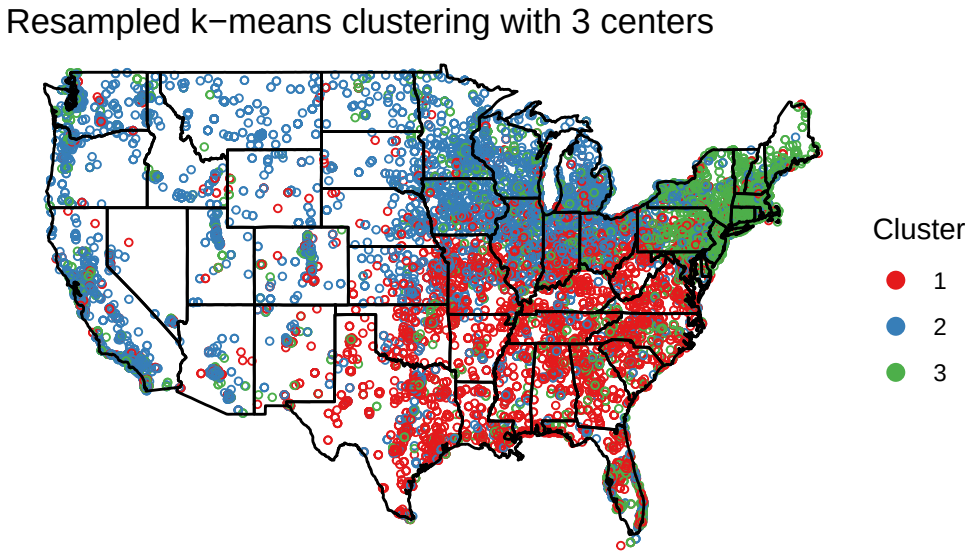
While we are again quite confident that there are three cluster centers, the results aren't as clear or stable as those of k-means clustering.

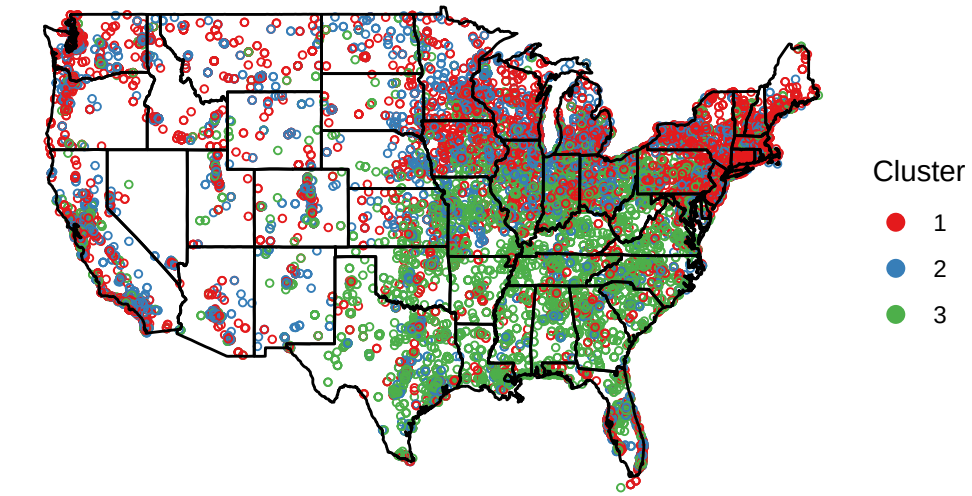
5 Stability of findings to perturbation

The results we have seen so far seem to pass the reality check and match what we know historically about the United States. I doubt anyone familiar with American culture would be surprised to see the groupings. However, we must ensure that we are not falling victim to confirmation bias. In the subsequent sections, we will test how much torture our results can endure.

5.1 Perturbing the data set

We will resample users with replacement from our original cleaned data set to determine if the clustering is stable. We will then run PCA again. Afterwards, the reduced data will be clustered using both methods and plotted on a map of the United States.





We can see that the three distinct regions still appear after performing both clustering methods. The main conclusion of our data analysis, that there are three major regions when it comes to speech patterns, still holds after several layers of perturbation. The results of PCA seem to have withstood the perturbation better than those from NMF. From those comparisons, we have evidence that PCA is the more stable method when it comes to problems similar to this one.

Another stability check was integrated into our analysis. For each method, we used many different starting points to pick the number of cluster centers. In the above figures, we can see that virtually every single one of the 32 subplots showed the same three major clusters. From this, we can conclude that our results are stable when it comes to starting points.

6 Conclusion

- Discuss the three realms of data science by answering the questions in the instructions pdf.
- Come up with a reality check that would help you to verify your clustering. You do not necessarily have to perform this reality check, but you can if doable.
- What are the main takeaways from your exploration/clustering/stability analysis?

6.1 Data

I think this data could be useful for certain decision-making purposes. This data set was collected anonymous users with no verification of their responses. If one takes the time to look at the CITY and STATE values, some are clearly jokes. The data is sufficient for satisfying someone's curiosity or practicing their data analysis skills, but it shouldn't be used to make any decisions of consequence. Ideally, each user's zip code could be recorded using the IP address that is provided to the website when they take the test. This would provide us with far more reliable data.

6.2 Algorithms and analysis

Despite the potential problems with the data, our clusters are representative of reality. As previously stated in this report, no person familiar with American culture is likely to disagree with our results. However, the problems with the data prevented us from drawing more insightful conclusions that do more than confirm what we already know.

6.3 Future data

We also showed that our data was stable to perturbations, which suggests that data collected in the future would behave similarly. A more stringent stability check could entail scraping social media posts in each of these regions and searching for specific terminology found in the provided questions. Many social media platforms already provide the general location of each user, which would be more reliable than what they provide to a short internet survey.

If I had more time, I would put more effort towards verifying the zip codes and checking how they relate to the reported cities and states. From my own testing, random 5-digit numbers often end up being real zip codes in the U.S., so it is likely that paranoid users entered random zip codes to prevent us from knowing their location. It would be difficult for us to differentiate between real and fake zip code values. Checking the cities and states could be a good place to start.

7 Academic Integrity Statement