Lab 2 - Linguistic Survey
Stat 215A, Fall 2017

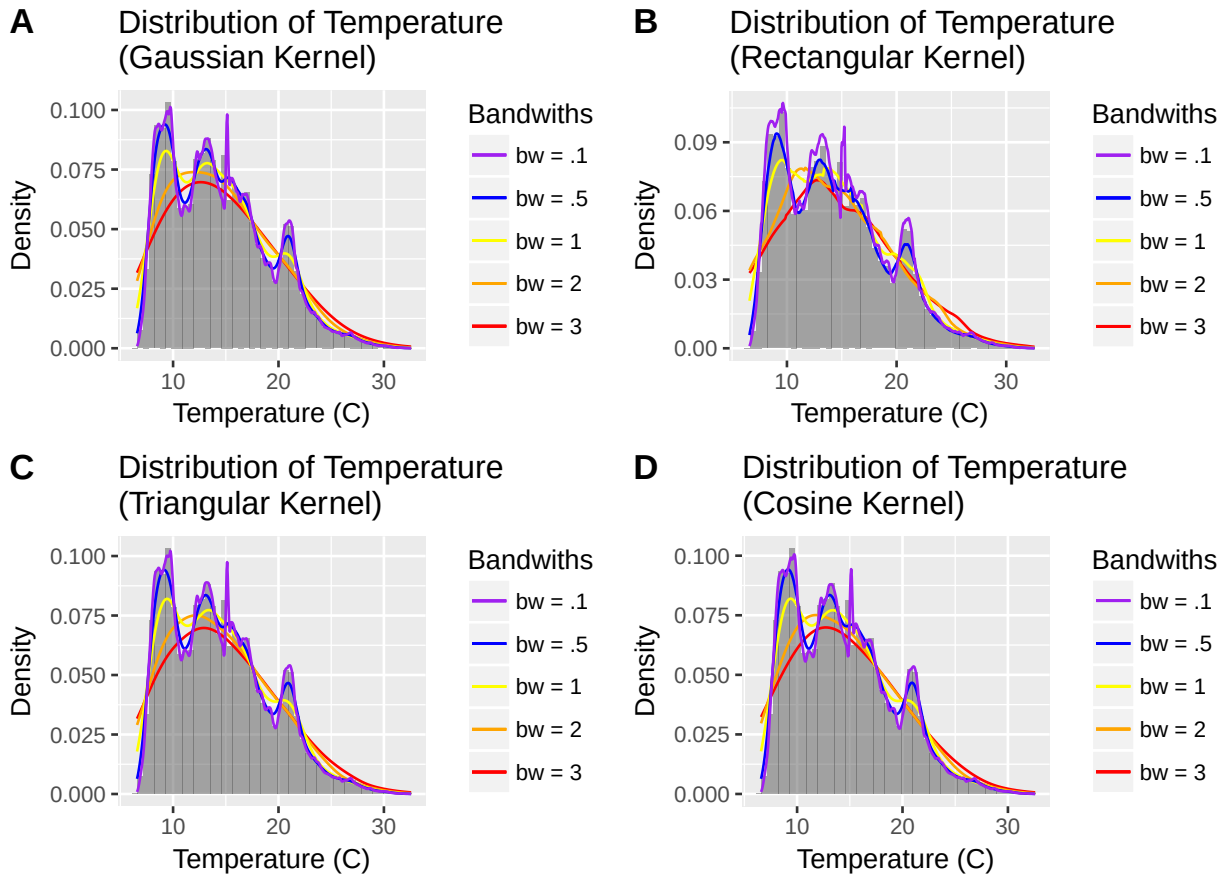
SID: ****2167

October 5, 2017

1 Kernel Density Plots and Smoothing

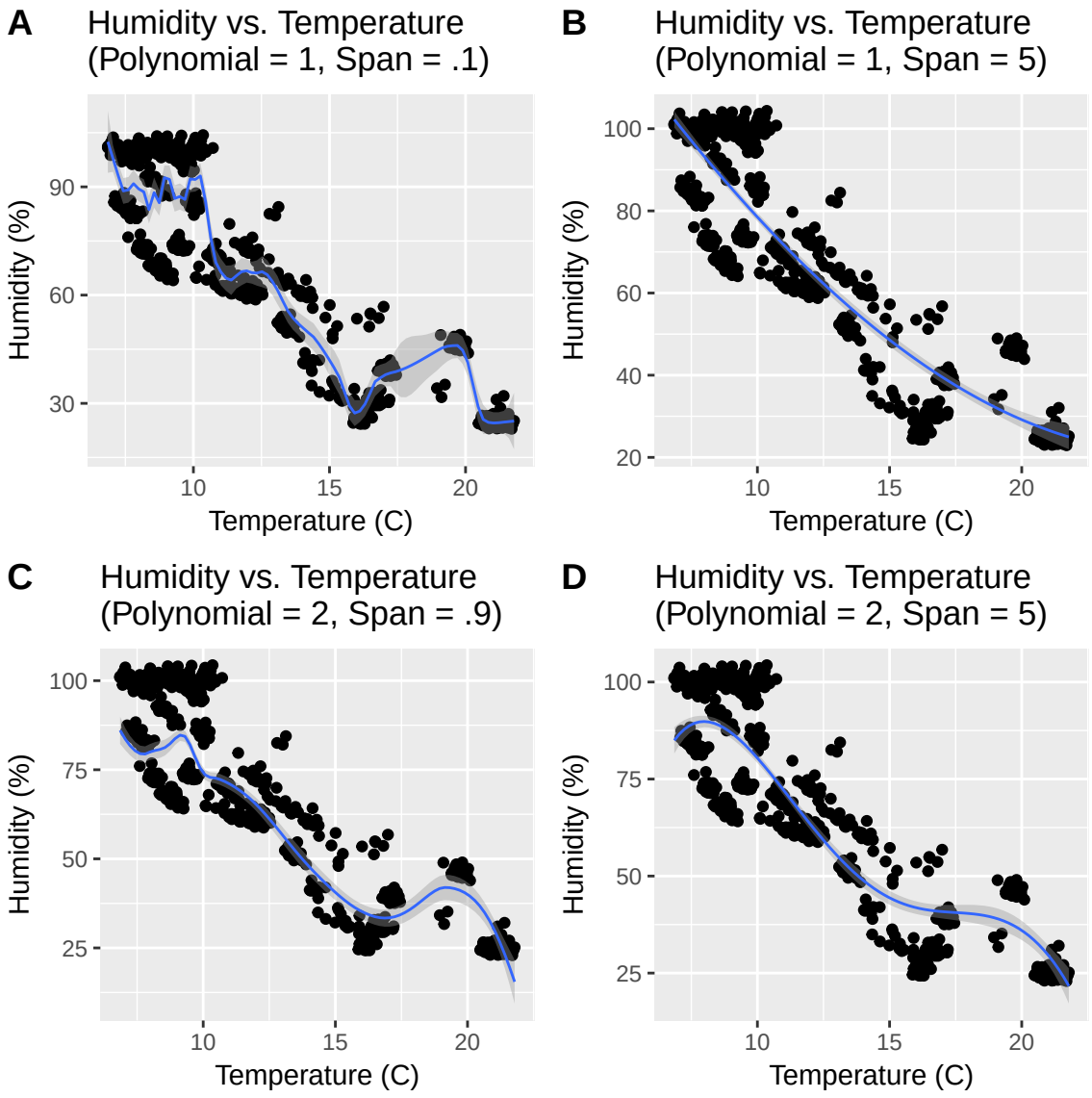
1.1 Density Plots

When attempting to fit the histogram of temperature values for the Redwood Experiment data, there are clear trade-offs between the kernel function shape and the bandwidth. For bandwidth that is too small, the combined distributions become noisy and perhaps over fit. For bandwidth that is too large, the natural variance of the data is poorly explained. In the case of this data, $bw = .1$ seems to work very well for explaining the data without over fitting. Several shapes could be acceptable candidates, besides the square function, though there seems to be no obvious reason to stray from the default (Gaussian) option.



1.2 Smoothing

Attempting to smooth the data with the loess module, we find that several options can be used to optimize the fit. The 2nd degree polynomials and higher are not appropriate, as their standard error and general appearance suggest. The first degree polynomial fits quite well, and given a sufficiently well selected bandwidth (here, span= 5), the fit is better behaved.



2 Introduction

In this section we will analyze the data from a 2002 dialectic survey, focusing on regional differences between speech and word patterns in the United States. Our analysis will be focused on a subset of questions on pronunciation and word choice.

3 The Data

3.1 Data quality and cleaning

Our data cleaning protocol is fairly straightforward. Given that the linguistic location data (lingLocation.txt) is already cleaned by a previous member of the class, it falls to us to clean up the raw file (lingData.txt). We notice there are a number of erroneous state ID's. We can clean those up by selecting only complete rows (rows without NA's), and remove them from the factor list. Our data is now ready for processing. We will convert our (cleaned) categorical responses into a binary response table, but this will come in a later section.

3.2 Exploratory Data Analysis

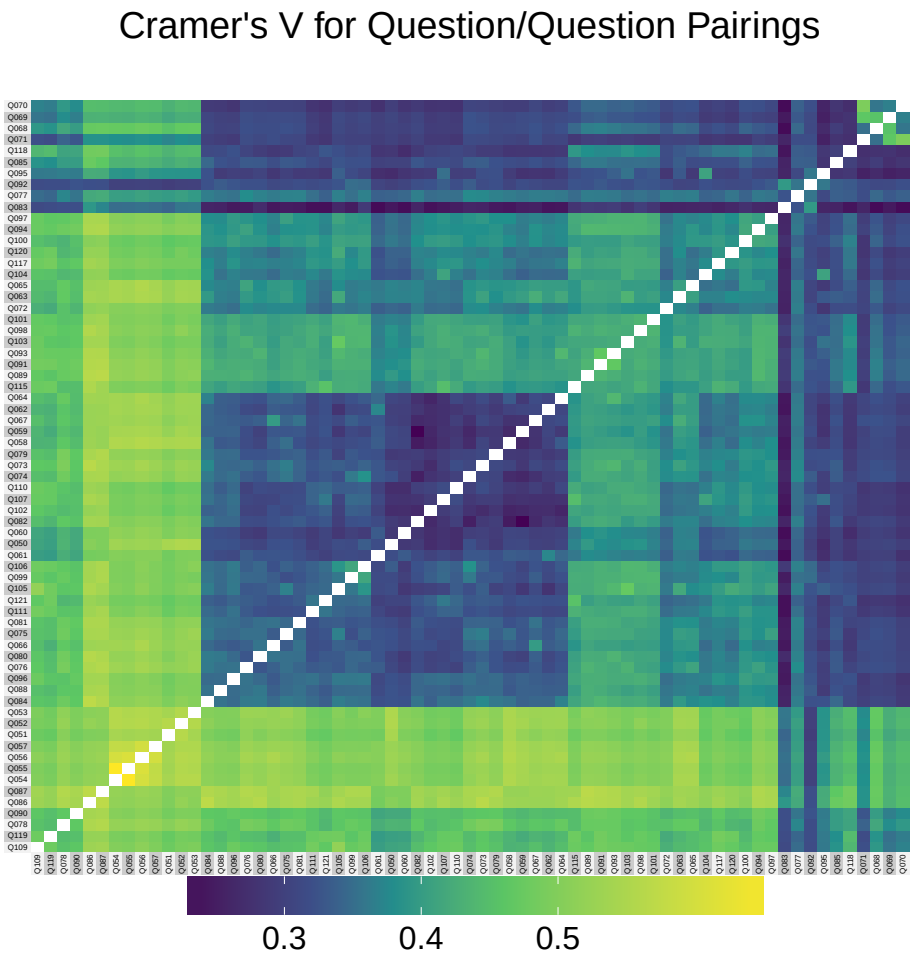
To begin, we take a general exploration approach in order to see which pairs of questions have the most relevance to one another. More generally stated: some pairs of questions will probably have more similarity to one another than others. What is the best way to compute this question to question interaction? We remember that our data is nominal (i.e., categorical), so while various responses are coded as integers, these integers in and of themselves have no meaning.

A good way to investigate categorical data is with Cramer's V. Cramer's V is a variant of the chi-squared statistic that takes the form:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

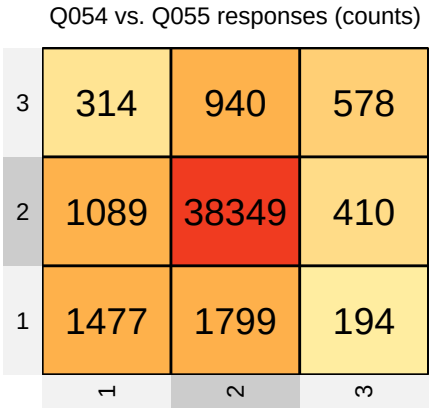
Where n is the number of observations, k is the number of columns in the data frame, and r is the number of rows. Cramer's V is bounded from 0 to 1, and thus lends itself to a somewhat intuitive interpretation. For reasons that will no be discussed here, it has the propensity towards over biasing associative claims.

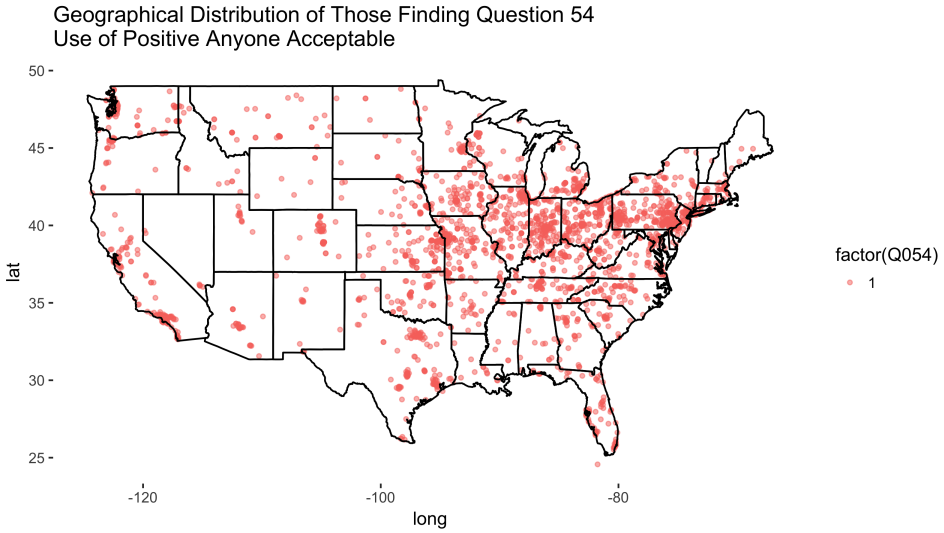
Here we will create a heat map of $V(Q_i, Q_j)$ for all i, j. Then, using the superheat heat map package, we will do a simple hierarchical clustering of the results, so strongly correlated questions will be grouped together. For my investigation, this was presented as an interactive heat map, but for the purposes of creating a report that can be efficiently recompiled, it will be represented as a static image, with the results from some of the highest Q/Q pairings reported below. The Cramer's V value is reported for each pair as well.

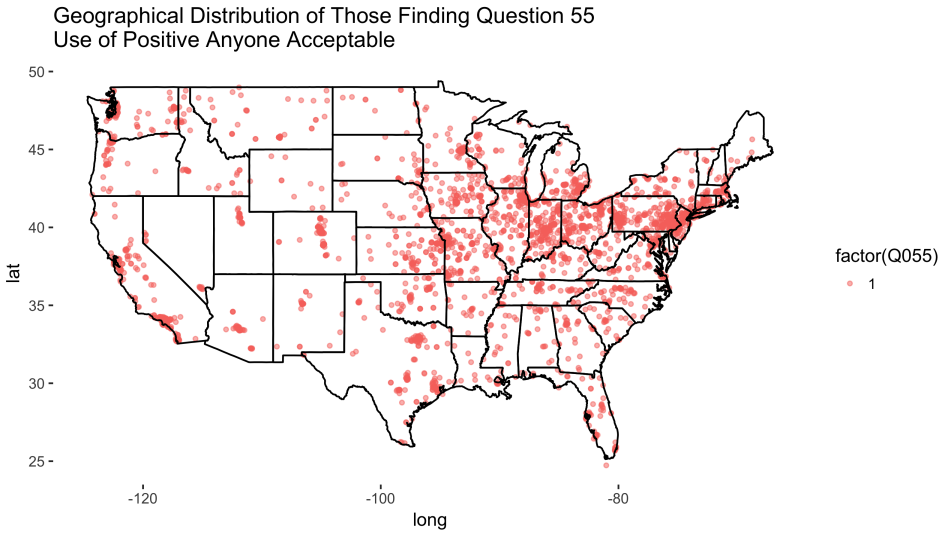


3.2.1 Q54-Q55 ($V = 0.65$)

We investigate the relationship between the acceptability of the following statements: "He used to nap on the couch, but he sprawls out in that new lounge chair anymore" (Q54) and "I do exclusively figurative paintings anymore" (Q55). Both are examining the use of the "positive anymore." Normally it is used in the negative context "I don't eat like that anymore" but here it is used both times in the positive sense. It makes sense that this was the strongest of our correlative pairs, since the questions are interrogating very similar ideas. From our table/heat map we see that most find both of these statements unacceptable. Let's try to examine then, where both contexts of the positive anymore are acceptable.


Q054 vs. Q055 responses (counts)
3
2
1
314
940
578
1089
38349
410
1477
1799
194
1
2
3

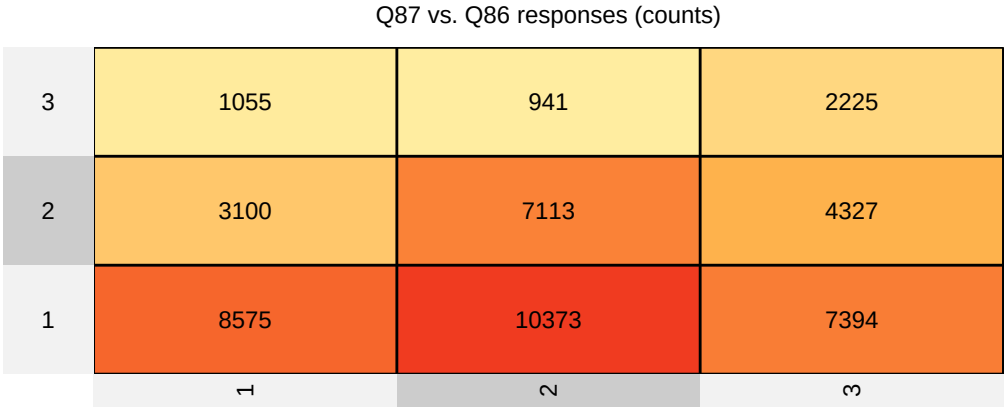

Geographical Distribution of Those Finding Question 54
Use of Positive Anyone Acceptable
lat
long
50
45
40
35
30
25
-120
-100
-80
factor(Q054)
1

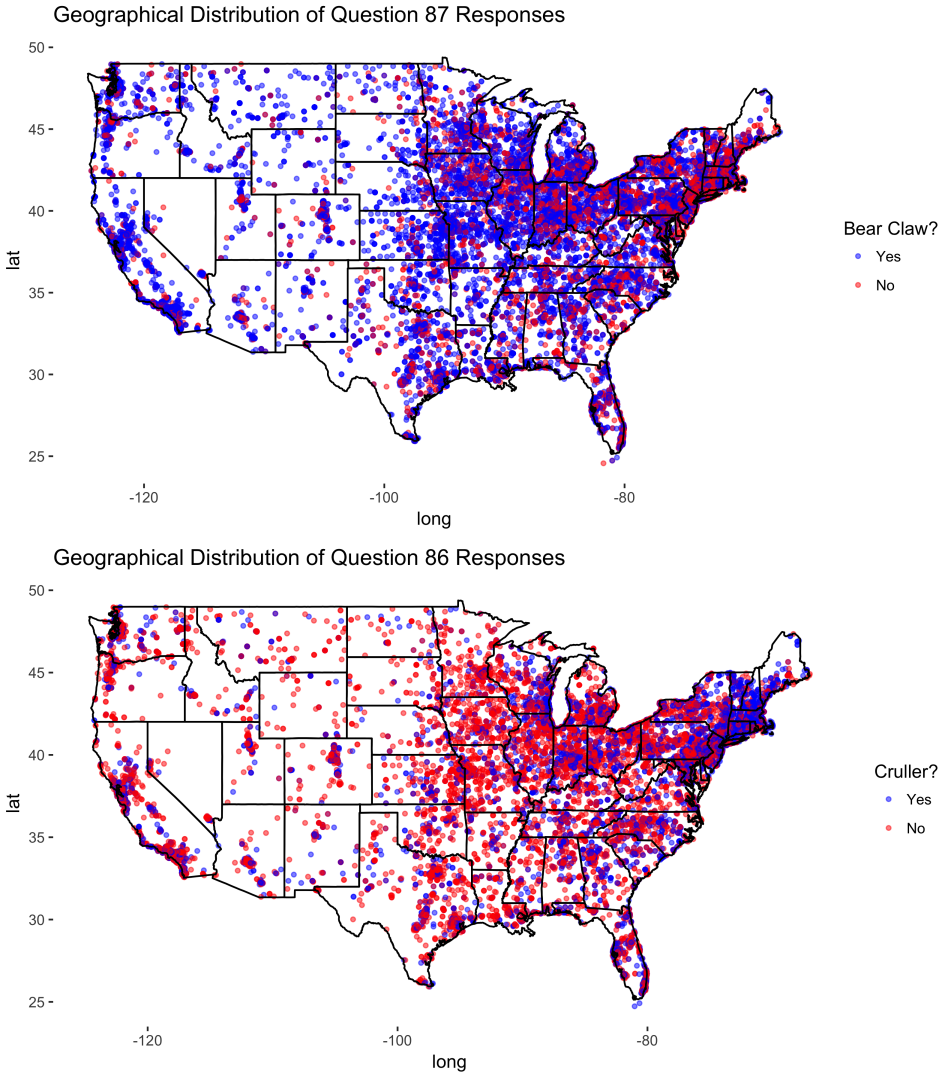

Geographical Distribution of Those Finding Question 55
Use of Positive Anyone Acceptable
lat
long
50
45
40
35
30
25
-120
-100
-80
factor(Q055)
1

We see the use of the "positive anymore" is mostly visible in the Midwest and Appalachia, which might fit with our intuition, as some theorize it brought the the United States by Irish immigrants ("Irish English Volume 2: The Republic of Ireland"), and these regions have traditionally house Irish immigrants and would allow for immigrants to maintain their linguistic identity more than in large cities.

3.2.2 Q87-Q86 ($V = 0.57$)

We then try to investigate a less significant association between "Do you use the term "bear claw" for a kind of pastry? " (Q87) and "Do you use the word cruller?" (Q86).



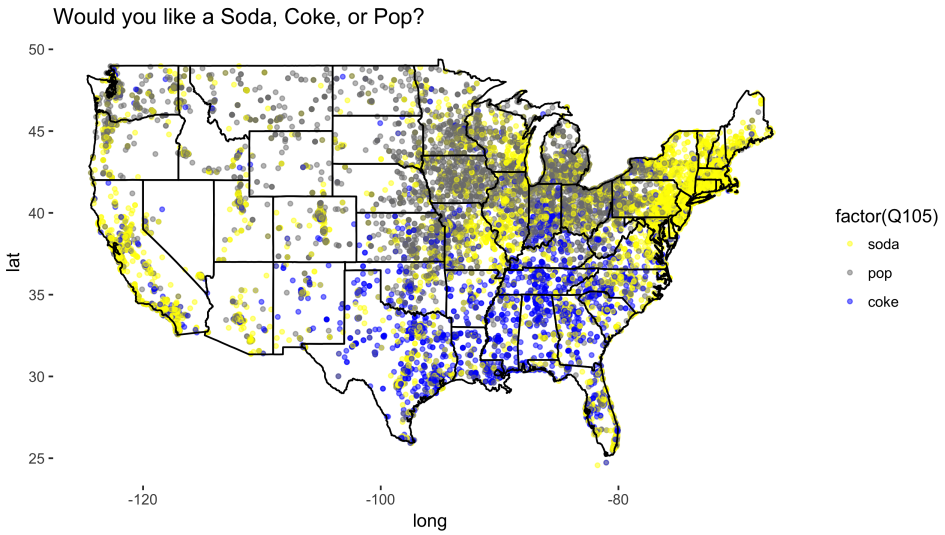


A cruller is a traditional type of New England pastry, whereas bear claw pastries are more traditionally found in the Western united states. We have limited our answer selection to those who know both types of pastries, but have usage preference, since they are in the majority. The cruller is indeed used more often as a word in New England, especially compare to the bear claw. The bear claw has more prevalence in the Midwestern states when compared to the cruller.

These findings confirm anecdotal evidence of word-of-mouth food history.

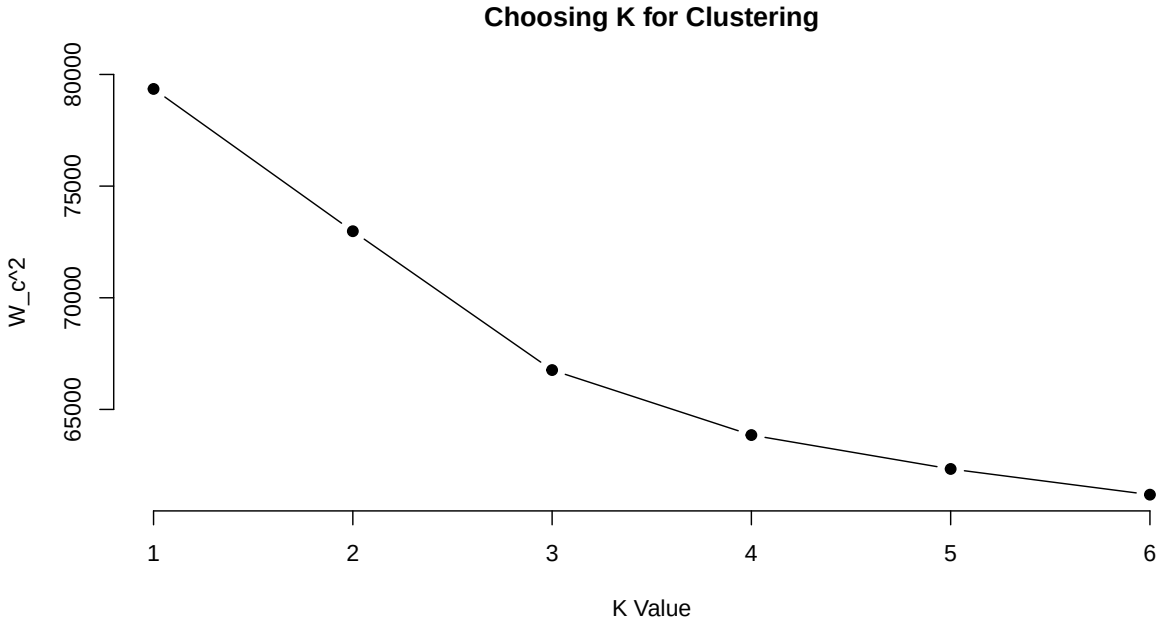
3.2.3 Investigating Language Directly

Rather than looking at question pairs, we can also explore the data with direct geographical separation. One such example is what people call sweetened carbonated beverages. There were several options for this question, so we will restrict ourselves to three of the most common words: “soda,” “coke,” and “pop.”



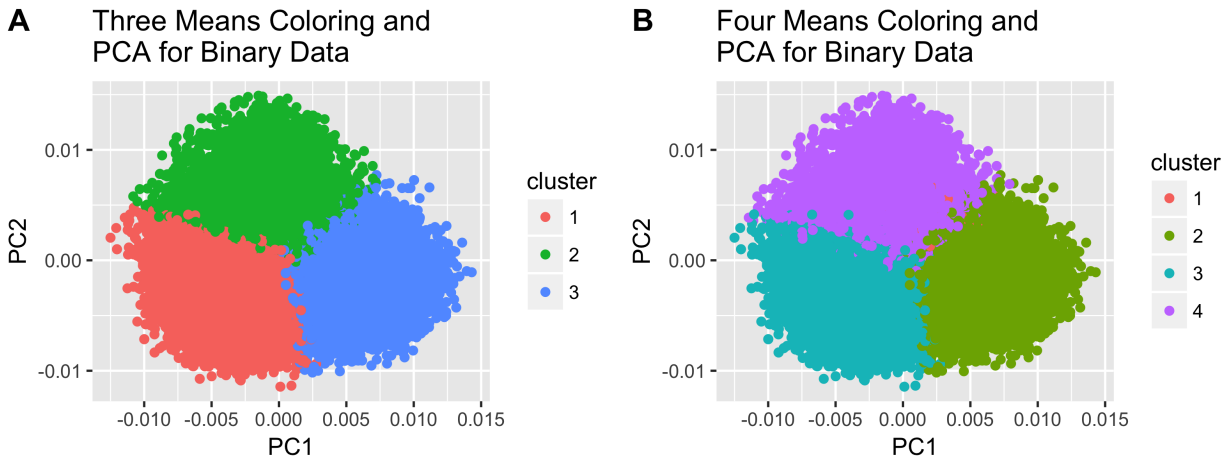
4 Dimension reduction methods

From the outset we attempt to use k-means clustering in order to make sense of the relationship between survey answers and geographical grouping. The first question with k-means that one must address is how to set k (i.e., how many groups to decide). A fairly crude but useful metric is looking at the within clusters sum of squares for multiple values of k. Often good choices for k can be found at the "elbow" or inflection of these graphs. Please note that some of the code used to generate this diagnostic graph came directly from R-Bloggers website. From our graph, we see that $k = 3$ or $k = 4$ would be an appropriate choice for our clustering algorithm.

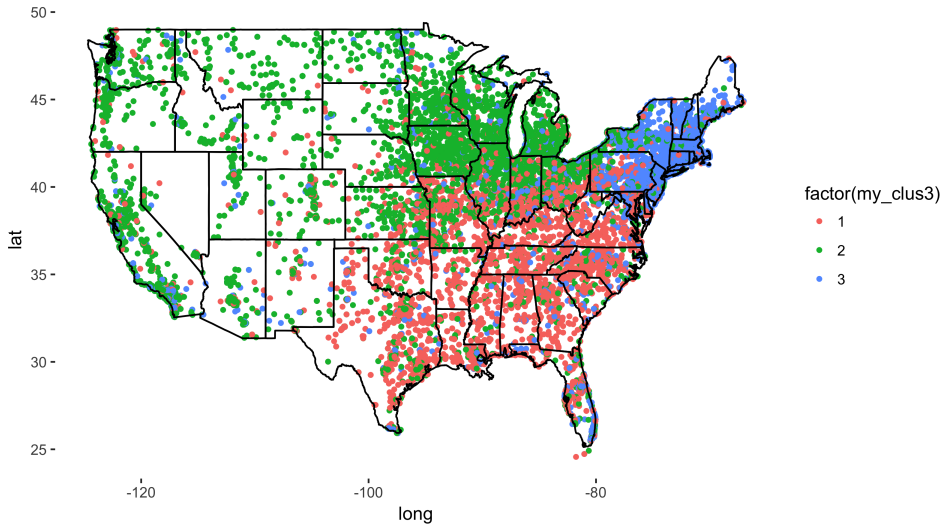


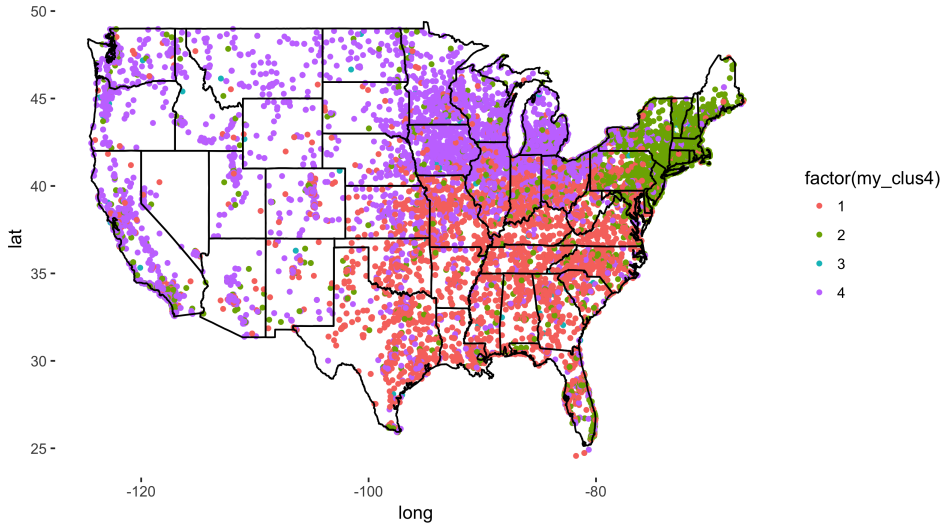
We then attempt to visualize this clustering initially by coloring a principle component axis plot by cluster. While the PCA itself is fairly unremarkable in terms of explanatory power, it does seem to detect three distinct sub-parts of the overall point mass. It also shows that $k = 4$ doesn't provide much more insight

$k = 3$ when it comes to explaining our data.



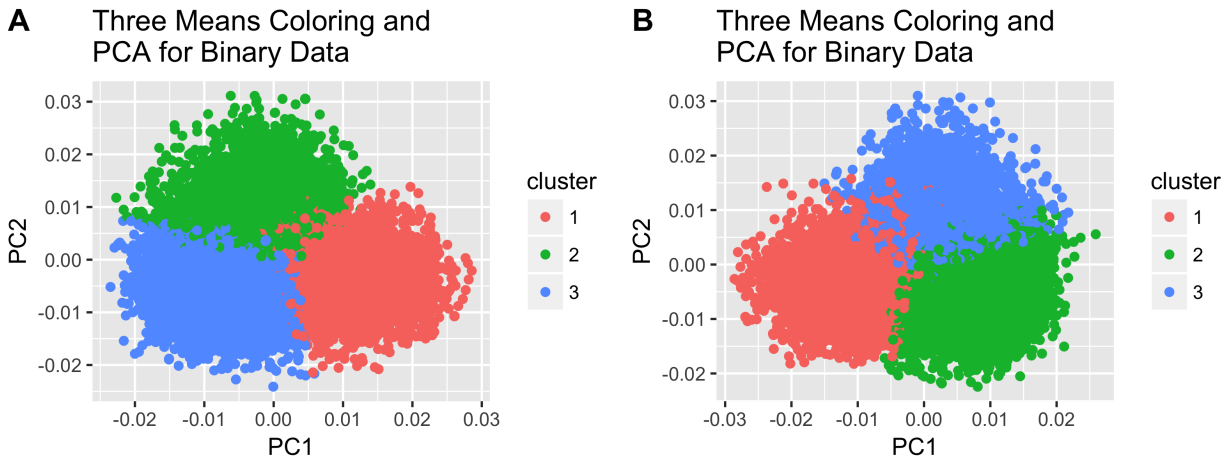
Finally we use our clustering to color our map of the continental United States. The k-means derived groups seem to identify areas that fit with common intuition about regions dialectical differences: the Southern United States, the North Eastern United States, and the Midwestern/Western United States. There are exceptions to these trends, especially around major urban centers (which makes sense, given that they are often hubs for drastic geographic relocation).

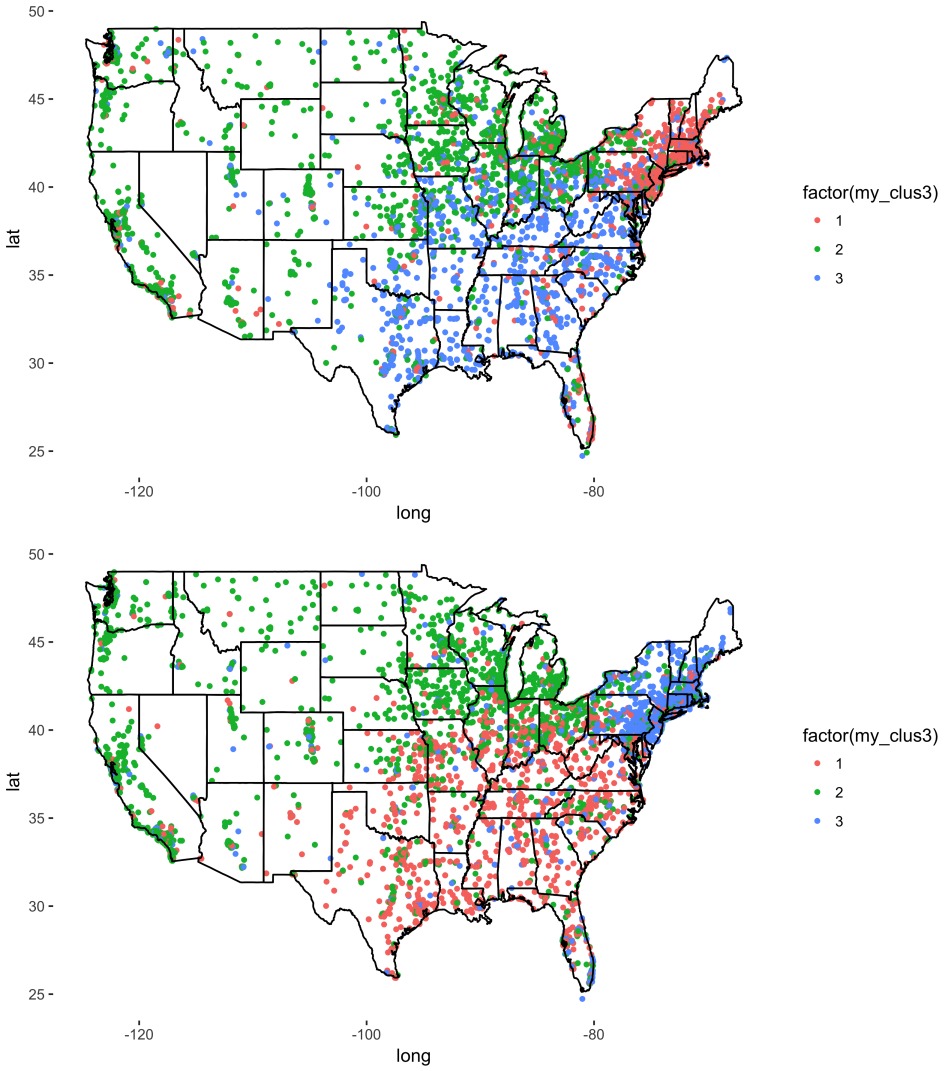




5 Stability of findings to perturbation

In order to check that our findings are robust, we attempt to sub-sample, and then re-examine our PCA/clustering results to see if we can find similar conclusions. We will randomly sample 10,000 data points twice, comparing their results against one another as well as the entire data set.





From the above results we can confirm that our conclusions drawn from the PCA plots and k-means clustering are stable for simple random samples drawn from the larger data set, and are effective are recapitulating our results.

6 Conclusion

Our analysis has generated diverse conclusions for several different types of questions. For the re-examination of our redwood data set, we see that different kernel functional forms with different bandwidths can bring about different qualitative understandings of the data. The most plausible explanation seems to be that it is a three or four peaked histogram that is well fit by the default Gaussian shape, given an appropriate bandwidth.

Similarly, our LOESS smoothing examination shows value in larger regions (relative to the whole data set), and single polynomial term fittings. Though it was interesting to view the ways in which the trend line could be perturbed by changing to a squared polynomial term fitting with short bandwidths.

We examined the use of Cramer's V in order to look at the association of certain question answer to question answer associations. Though our heat-map showed that the associations were weak in general, the interactive version of these heat-map was useful for identifying the highest association pairs so we could

analyze their relationship to one another as well as geography. We looked at an interesting relationship between uses of the "positive anymore" and examined prefered nomenclature for pastries. Another food nomenclature examination involved qualitative descriptions of word choice in the US by looking at how the North Eastern, Midwestern, and Southern United States preferred the terms soda, pop, and coke (respectively) which describing carbonated beverages. These regional groupings concerning carbonated beverages and pastries would foreshadow results from our k-means analyses in the following section.

Our PCA showed a tight grouping, and we selected our value for k in k-means clustering by looking at the within clusters sum of squares graph for various k's, and attempting to find the "elbow" or slope change point of the graph. We experimented further with k values of $k = 3$ and $k = 4$ and found $k = 3$ to be sufficient. When graphed, we see these clusters associate with the North East, Midwest/Western, and Southern United States.

Finally, to confirm our findings, we sub-sampled our data and saw that for independent sampling of the whole data set, our conclusions drawn from our clustering could be recapitulated.

7 Acknowledgements

The author gratefully acknowledges the assistance of *** ******* in discussing various aspects of this lab.

8 Note to Grader

If you are attempting to recompile this document, it may take a while. Please also make sure that you have all of the necessary packages installed in your system, since there are quite a few and they will probably not be on your system.