Lab 2 - Linguistics Data, Stat 215A, Fall 2024

2024-09-28

Introduction

The study of language variation has long been a crucial area of linguistic research, offering valuable insights into the historical, social, and geographical factors that shape language use in society. Linguists have traditionally sought to understand how language varieties—such as dialects—are distributed across different regions and communities. A central question in twentieth-century dialectology concerns the analysis of geographic coherence in language variation: given the imperfect geographic boundaries of dialects, how should this variation be analyzed? Traditionally, dialectology has focused on identifying dialect areas, defined as regions with relatively limited internal linguistic variation, distinguishing them from neighboring areas.

However, the existence of such dialect areas, which we intuitively perceive, has proven difficult to demonstrate rigorously. Kretzschmar (1998) offers a potential explanation, differentiating between attributive dialects—linguistic varieties specific to a given place—and blind dialects, or areas whose distinctiveness cannot easily be captured by conventional methods. Field linguists can often catalog the linguistic features of a particular location, identifying an attributive dialect, but struggle to compare it meaningfully with varieties from other places. This challenge arises because traditional non-computational approaches typically analyze a small number of linguistic features, making it difficult to capture aggregate levels of variation.

Recent advances in computational techniques have transformed this landscape, enabling linguists to study language variation at scale. For example, simple techniques such as counting differences between linguistic features allowed Seguy to aggregate individual differences over large datasets. Other researchers, such as Speelman, Grondelaers, and Geeraerts, have utilized relative frequency analyses to measure variation, exploring how pairs of alternative lexical choices—such as *car* vs. *automobile* or *quiet* vs. *still*—can reveal deeper insights into linguistic differences. These frequency-based approaches, often referred to as linguistic profiles, have become essential tools in quantifying linguistic distance.

Building on these computational approaches, this paper explores the application of Principal Component Analysis (PCA) and clustering techniques for analyzing large-scale linguistic datasets. PCA is employed to reduce the dimensionality of the linguistic variables, capturing the most significant patterns of variation. Clustering methods, such as K-Means and Agglomerative Clustering, are then applied to group linguistic varieties based on their aggregate features. These techniques enable the identification of meaningful linguistic groupings and the exploration of dialect areas at a finer granularity than traditional methods. By incorporating computational tools like PCA and clustering, we hope to address longstanding challenges in dialectology, providing a more robust and scalable framework for analyzing language variation across regions.

The Data

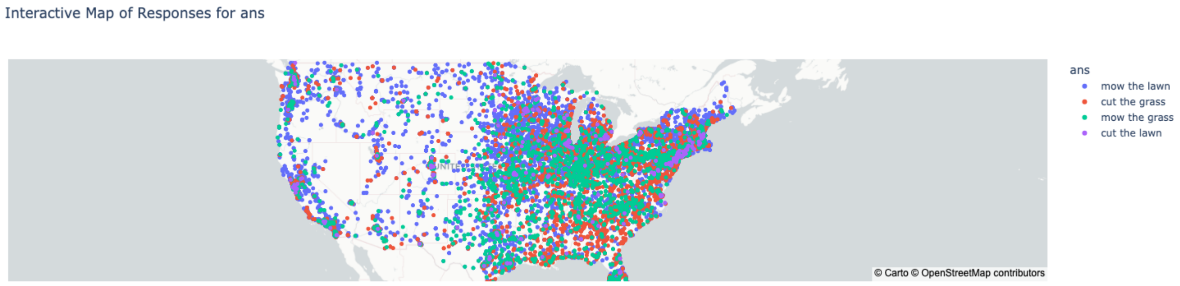
The data are from a Dialect Survey conducted by Bert Vaux: https://www.dialectsofenglish.com/. The questions and answers was found and processed from the http://dialect.redlog.net/index.html by a past intrepid STAT215A student. The dataset contains the answers to the survey questions for 47, 471 respondents across the United States as well as the variables ID, CITY, STATE, ZIP, Q50 - Q121, lat and long. ID is a number identifying the respondent. CITY and STATE were self reported by respondents. Former GSIs found the latitude and longitude for the center of each zipcode and added the lat and long variables based on the reported city and state. Then the data was binned into one degree latitude by one degree longitude squares. In this paper, we will focus on the questions that look at lexical differences as opposed to phonetic differences, which are numbered 50-121.

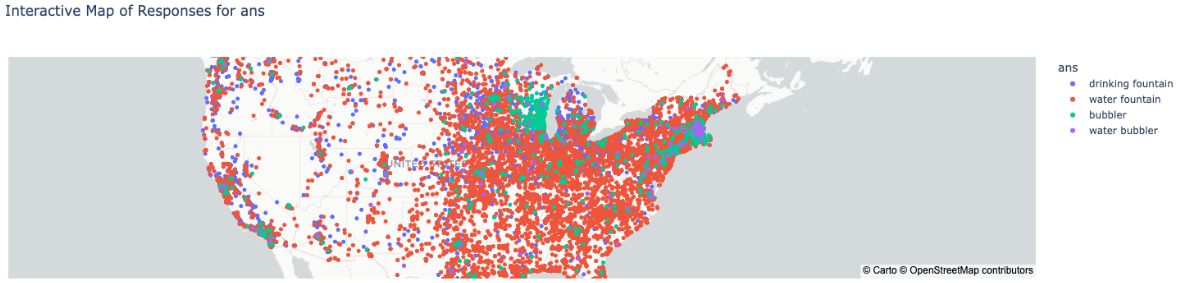
Data Cleaning

In the dataset, “QXXX” are the responses to the corresponding question on the survey website. A value of 0 indicates no response. The other numbers should directly match the responses on the website, i.e. a value of 1 should match a response of (a). Data cleaning will drop rows with answers to 25 questions or more as “0”, which reduced the record from 47471 to 46136.

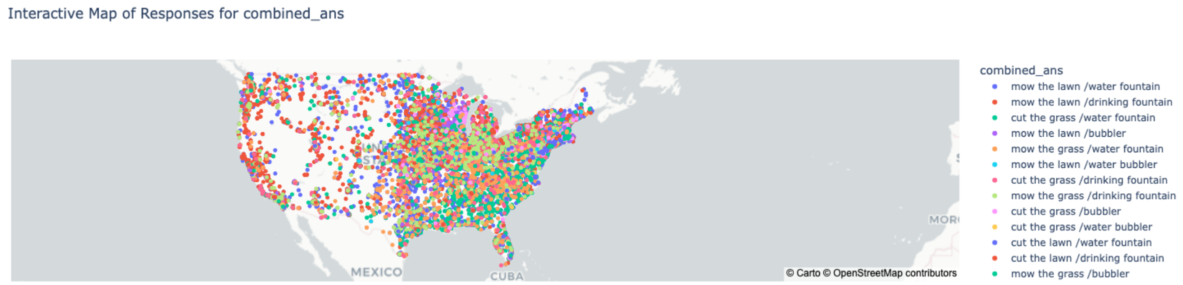
Exploratory Data Analysis

We picked two survey questions and investigate their relationship to each other and geography: what word do you use to refer ‘mow the lawn’ and what word do you use to refer ‘water fountain’? We created interactive maps to examine the geographical relationships. It seems “mow the lawn” and “cut the grass” are most common for the first question while “water fountain” and “drinking fountain” are most common for the second question. However, the popular answers do not define geopgraphical groups, as they spread across US. However, the infrequent answers (“mow the grass” and “bubbler”) seem to have a more distinct clusters (central and south US), consistent with Goebl’s observation when he assigned more heavy weightings which count overlap in infrequent words.


Interactive Map of Responses for ans
ans
● mow the lawn
● cut the grass
● mow the grass
● cut the lawn
© Carto © OpenStreetMap contributors


Interactive Map of Responses for ans
ans
● drinking fountain
● water fountain
● bubbler
● water bubbler
© Carto © OpenStreetMap contributors

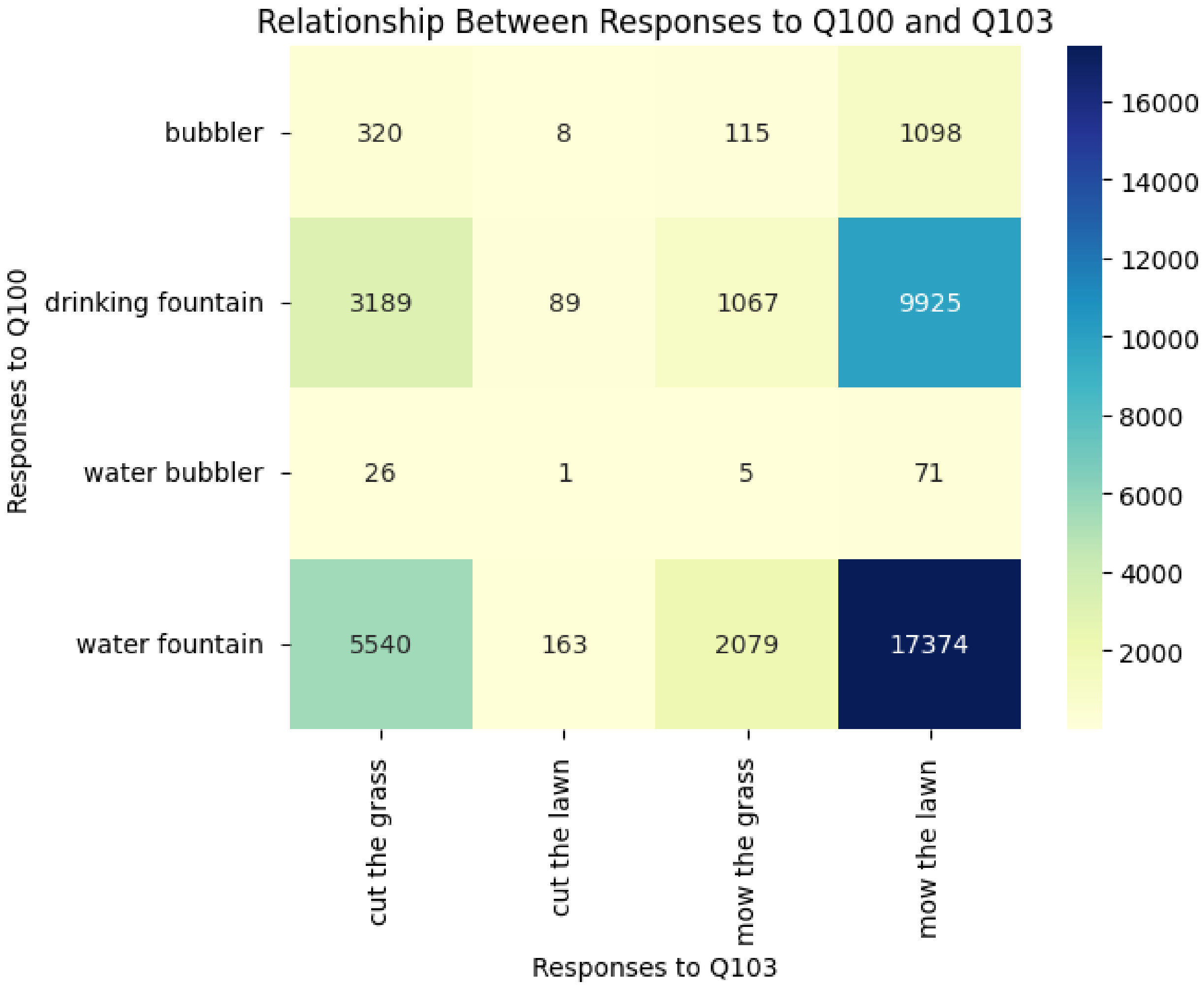
Next we explore if the answers to the two questions define any distinct geographical groups. Similarly to previous finding, the infrequent pair such as “mow the grass” and “bubbler” seem to form a cluster than the common phrases, based on the interactive map below.


Interactive Map of Responses for combined_ans
combined_ans
mow the lawn /water fountain
mow the lawn /drinking fountain
cut the grass /water fountain
mow the lawn /bubbler
mow the grass /water fountain
mow the lawn /water bubbler
cut the grass /drinking fountain
mow the grass /drinking fountain
cut the grass /bubbler
cut the grass /water bubbler
cut the lawn /water fountain
cut the lawn /drinking fountain
mow the grass /bubbler
MEXICO
CUBA
MOR
© Carto © OpenStreetMap contributors

We explore if a response to one question help predict the other by creating a contigency table and a heat map below, which confirms each other. The heatmap visualizes the relationship between responses to two different questions, Q100 (water fountain) and Q103 (mow the lawn). On the x-axis, we see the responses to Q103, and on the y-axis, the responses to Q100. The color intensity in the heatmap represents the frequency of respondents choosing each combination of answers for the two questions. The color bar on the right shows the frequency scale, where darker blue represents a higher number of respondents, and lighter shades represent fewer respondents. Darker squares indicate the most common response combinations between Q100 and Q103. "Water fountain" + "mow the lawn" is the most frequent combination, with 17,374 respondents answering these two options. This suggests that many people who use "water fountain" also prefer "mow the lawn." "Drinking fountain" + "mow the lawn" is the second most common combination, with 9,925 respondents, indicates that "drinking fountain" is also often paired with "mow the lawn."

```
ans                 cut the grass  cut the lawn  mow the grass  \
ans
bubbler                       320             8            115
drinking fountain            3189            89           1067
water bubbler                  26             1              5
water fountain               5540           163           2079
All                          9075           261           3266

ans                 mow the lawn    All
ans
bubbler                     1098   1541
drinking fountain           9925  14270
water bubbler                 71    103
water fountain             17374  25156
All                        28468  41070
```



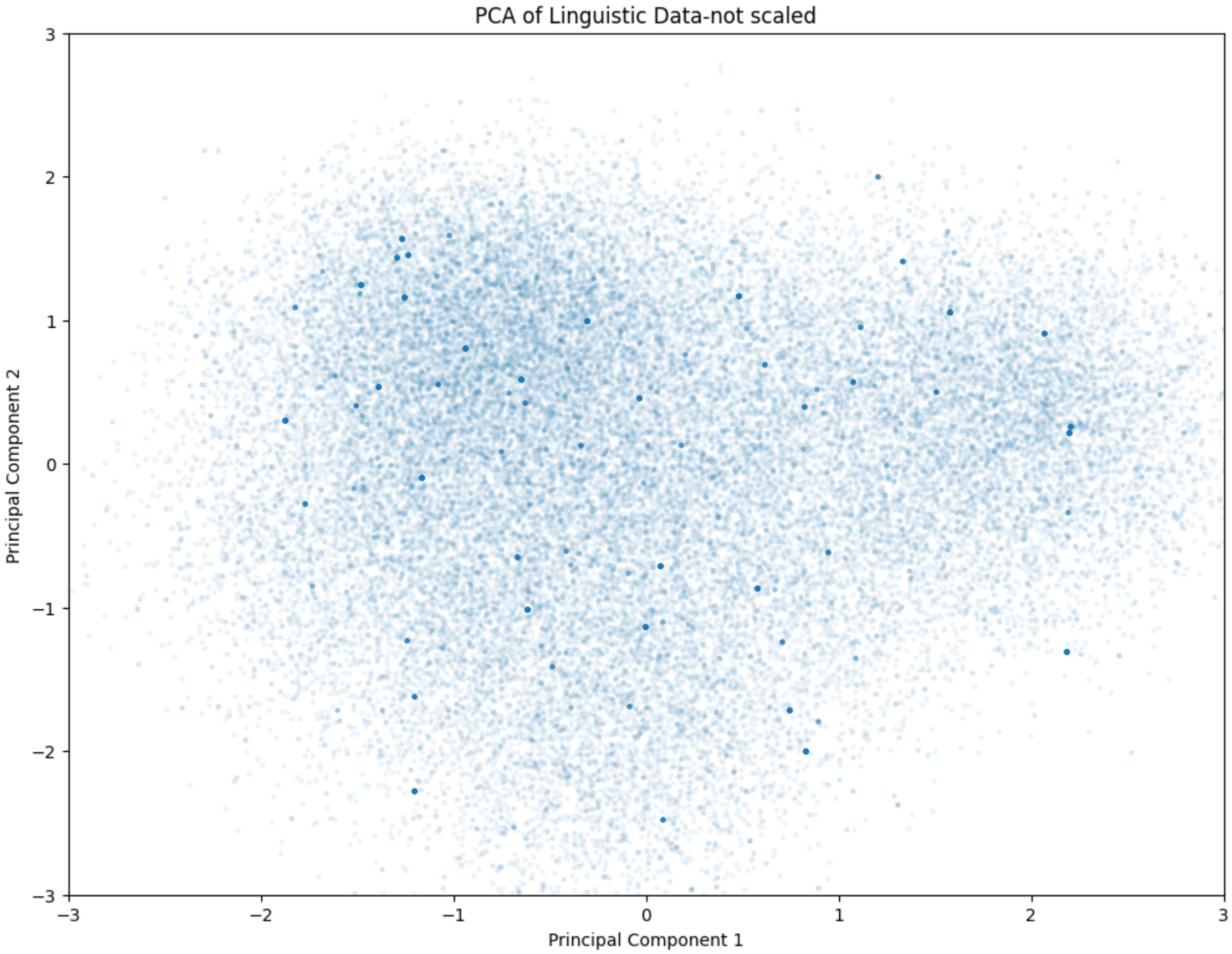
Quantitatively, we ran chi squared test which showed no significant association found between the responses to the two questions. See output below.

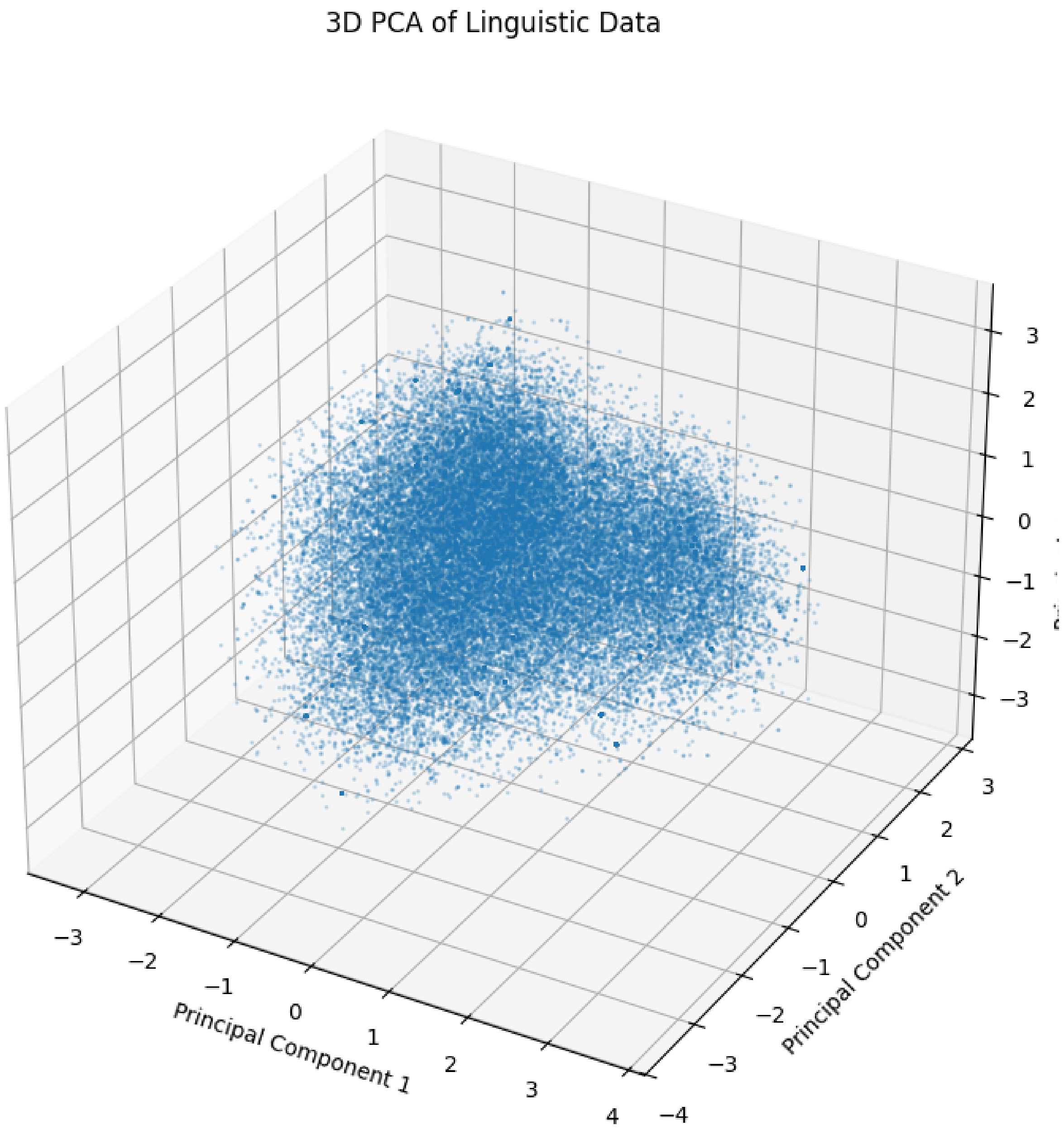
```
Chi-squared Test Statistic: 12.7872682247342
P-value: 0.17247076692033975
No significant association found between the responses.
```

Dimension Reduction

- This is where you discuss and show plots about the results of whatever dimension reduction techniques you tried—PCA, variants of PCA, t-SNE, NMF, random projections, etc.
- What do you learn from your dimension reduction outputs
- Discuss centering and scaling decisions It is very likely the two questions are not sufficient as features to characterize aggragate levels as discussed in Nerbonne et. al (2003). We are using all Q50-Q121 to see if that provide enough features. First step is to encode the data so that the response is binary instead of categorical. This makes p = 541 and n = 46,136. We used k instead of k-1 during the encoding. Next, we used PCA to reduce the dimensions. We started with PCA without scaling and plotted the PC1 and PC2 in 2-D. The dimension reduction didn't seem to maximize the variance of the dataset in PC1 or PC2. Project PC1, PC2 and PC3 to 3-D space and issue persists.

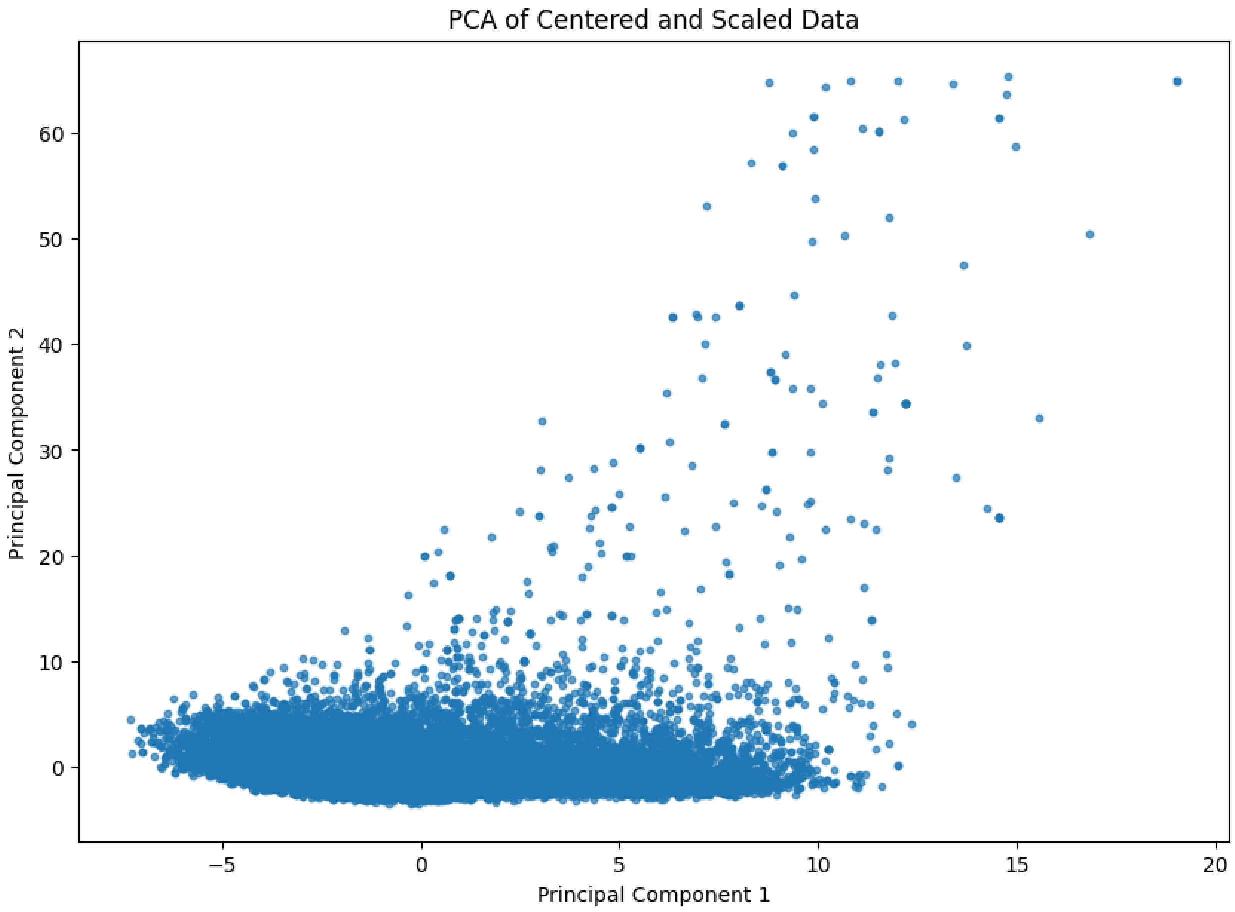
```
Feature matrix shape: (46136, 535)
```


PCA of Linguistic Data-not scaled
Principal Component 1
Principal Component 2
−3
−2
−1
0
1
2
3
−3
−2
−1
0
1
2
3

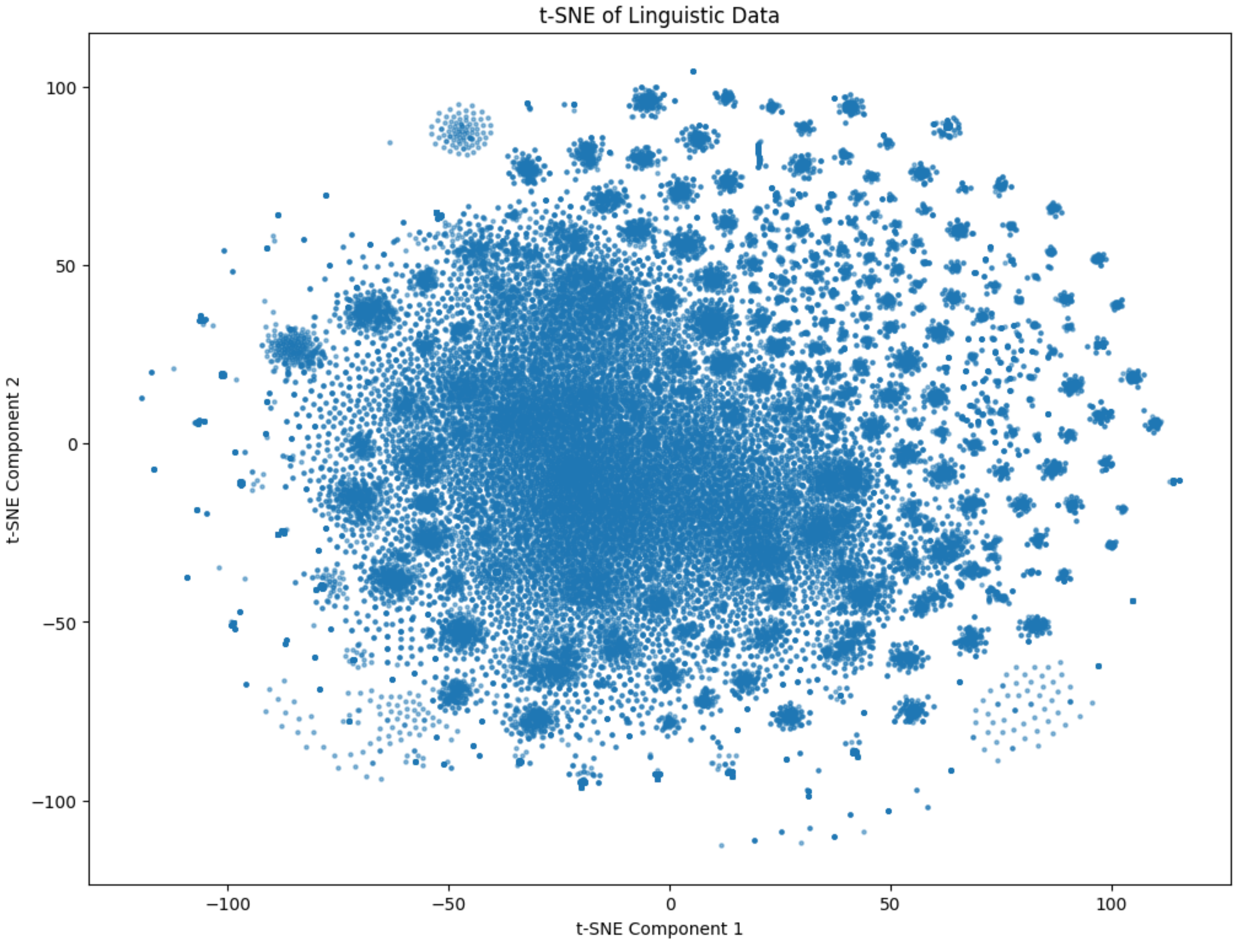


Next, we scaled the encoded dataset, ran PCA and projected the results to 2-D as shown below. The outcome has been significantly improved. Centering ensures that the principal components are computed based on the direction of maximum variance from the mean of the data. Scaling (standardizing to unit variance) is especially important if our features have different units or ranges. Without scaling, features with larger ranges could dominate the PCA, which may lead to incorrect results. However, in our case, scaling helped with our encoded dataset.

```
Feature matrix shape: (46136, 535)
```

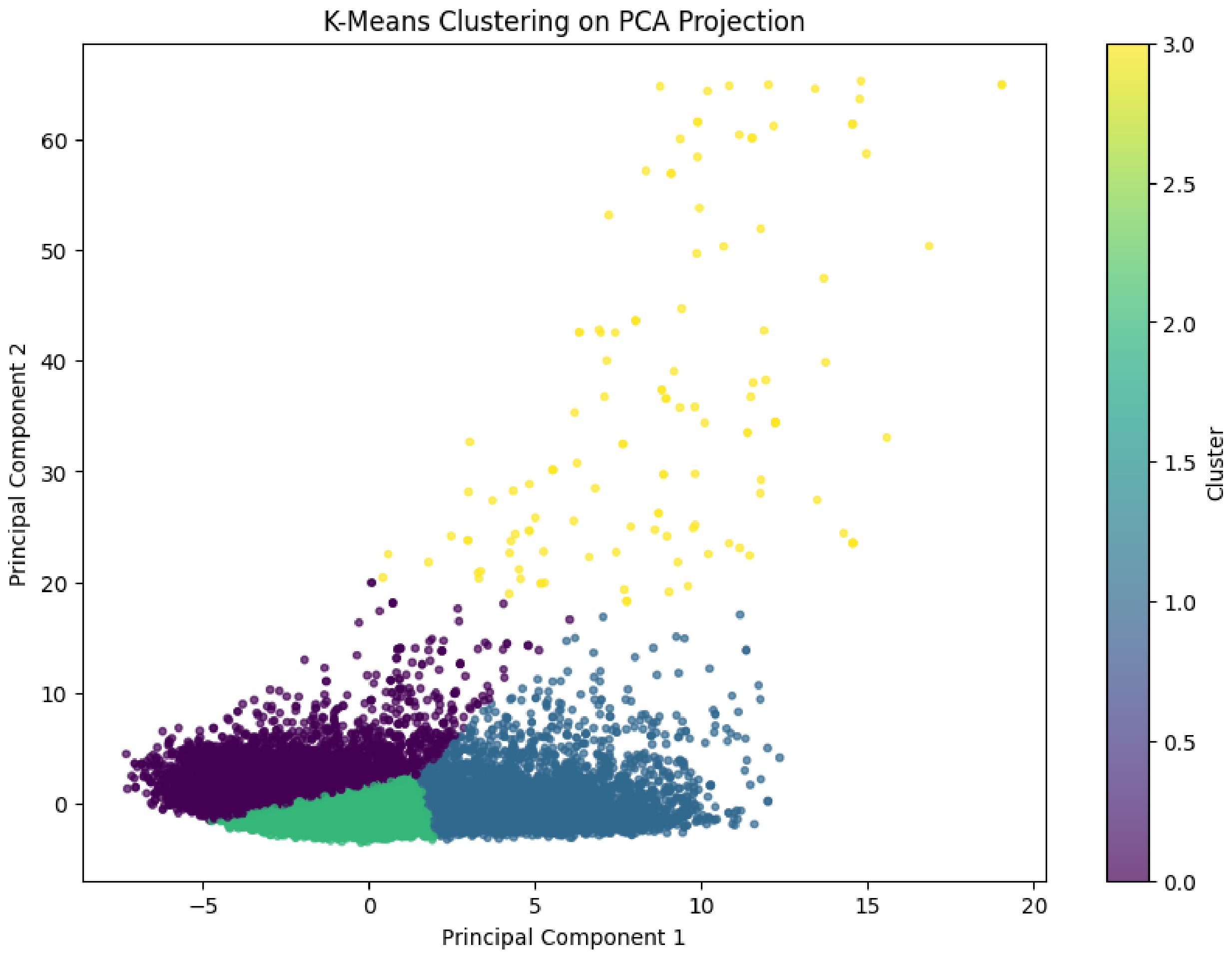


T-SNE was also experimented as a dimension reduction technique, with the encoded and scaled dataset. The dataset was projected onto the primary component and secondary component space as shown below. PCA with scaled data performs better than T-SNE in our case.

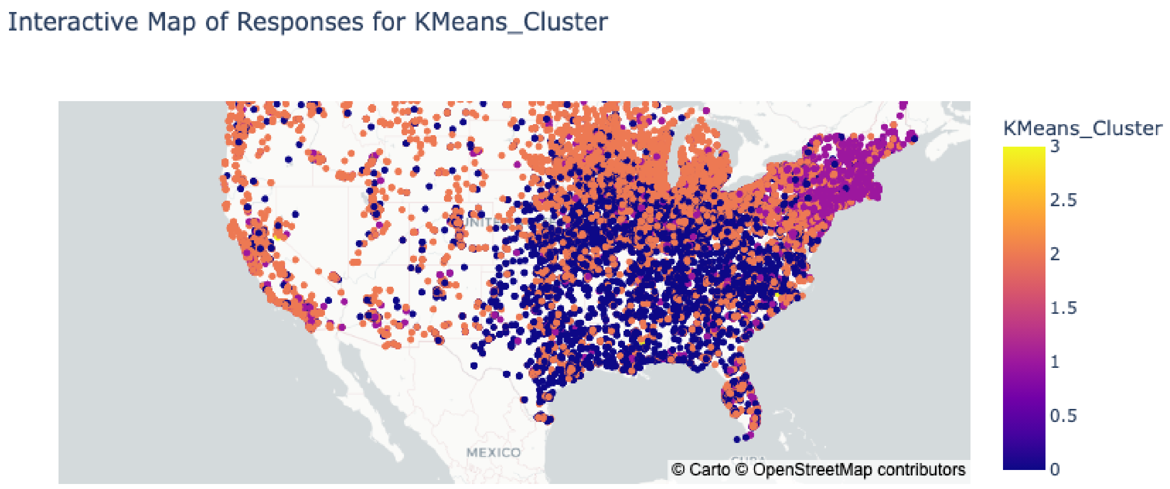


Clustering

After the dataset being projected on to PC1 and PC2 after PCA with scaled data, two different clustering methods are performed- K-means and Agglomerative Clustering. K=4 was selected using adjusted_rand_score as an evaluation mechanism, discussed in the Stability section. K-Means Clustering on PCA Projection is shownn below.



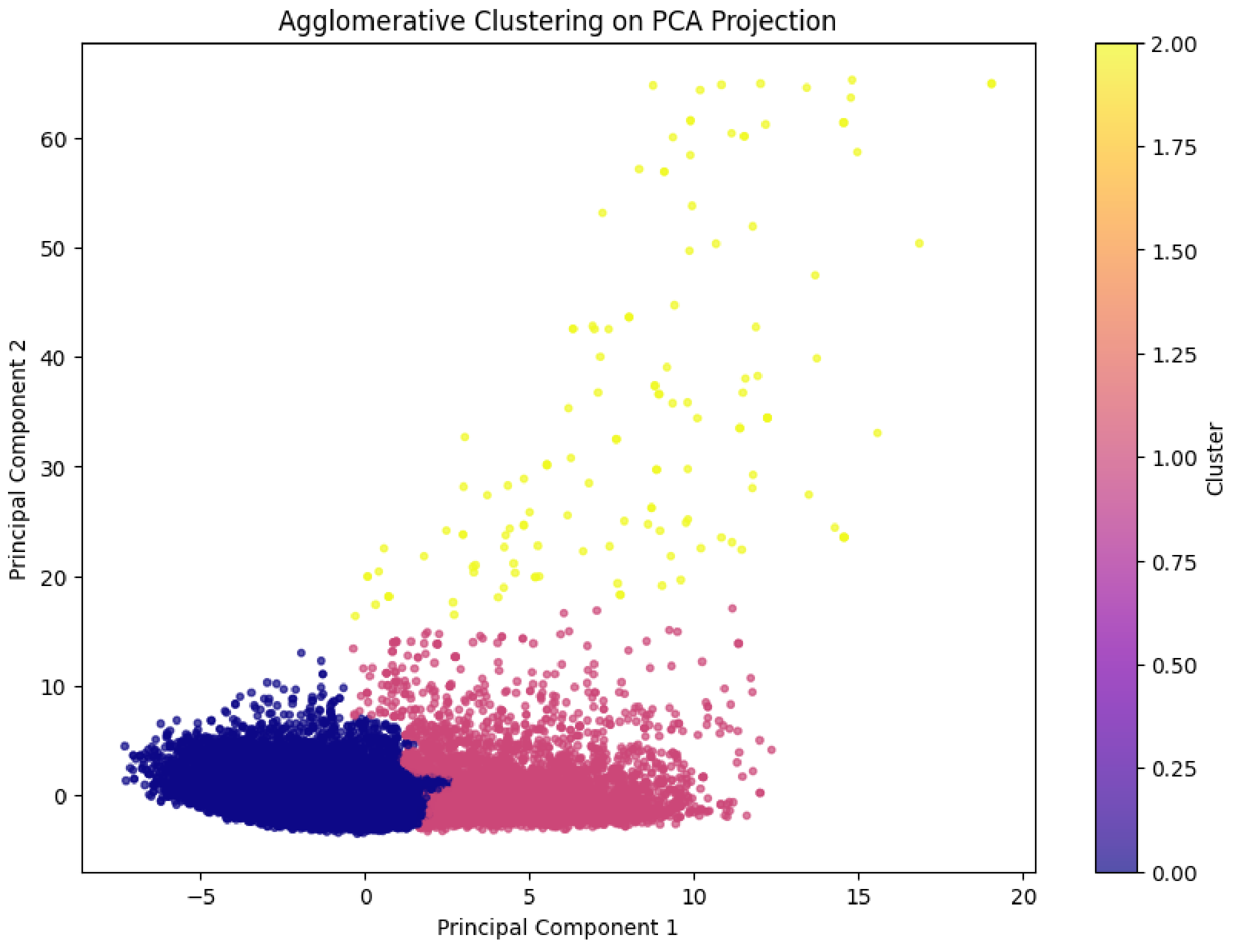
We then added the K-means labels to the dataset and plot them on the interactive map below. Consistent with the previous discussion on "mow and lawn" and "water fountain", the common label is widely spread across US while the blue and purple cluster seem to cluster in central/south US and east US. However, we cannot say the blue and purple clusters are completely isolated in a certain geographic area. There seems to be a continuum with a transition in density at central US.


Interactive Map of Responses for KMeans_Cluster
KMeans_Cluster
3
2.5
2
1.5
1
0.5
0
MEXICO
© Carto © OpenStreetMap contributors

We evaluated Which questions produce this continuum or separate the clusters by examing the PCA loading below, in descending order of the absolute value of PC1 weight. The top two weights are assigned to the answers to question 73, as shown below. Question 73 is about the word for "sneakers" or "tennis shoes".

```
            PC1       PC2
Q073_1  0.209652 -0.030818
Q073_6 -0.189739  0.033216
Q105_1  0.148928 -0.057069
Q080_1  0.147106 -0.020493
Q084_1  0.137032 -0.030482
Q093_2 -0.132329  0.001311
Q086_1  0.130709 -0.012003
Q066_5 -0.123389  0.026494
Q106_7  0.121860  0.002081
Q083_1  0.121532 -0.010367
```

Agglomerative Clustering is also examed, as shown below. Agglomerative Clustering is a type of hierarchical clustering where clusters are formed by recursively merging smaller clusters. It starts with each data point as its own cluster and iteratively merges the closest clusters until all points are in one large cluster, or a predefined number of clusters is reached. This "bottom-up" approach contrasts with divisive clustering, which starts with all data points in a single cluster and splits them recursively. Agglomerative Clustering builds a hierarchy of clusters, and the results can be represented using a dendrogram—a tree-like diagram that shows the merging process. It is very computational expense compared to K-means clustering.



	PC1	PC2	KMeans_Cluster
0	-0.861760	3.715114	0
1	5.163210	2.005029	1
2	4.562783	0.724209	1
3	4.446937	2.496730	1
4	4.212887	3.322333	1
...	...	...	...
46131	-1.778819	-0.378576	2
46132	-1.324144	0.055743	2
46133	-0.764036	-1.132271	2
46134	-0.695840	-1.154920	2
46135	1.119769	-1.624910	2

Stability of findings to perturbation

We analyzed the robustness of the clusters by bootstrap samping and evaluated the Adjusted Rand Index. After K-means clustering assigned the label for the original dataset and the bootstraped sample dataset, the labels for the same records are compared for consistency. The adjusted Rand Index approaching 1 means the results are consistent and the model is stable. We also used ARI to optimize k, the number of clusters. For k=2, 3 and 5, ARI is about 0.4; for k=4, using different starting point, ARI is consistenly between 0.96 and 0.98, see one instant below. Therefore we used k = 4 and the model is regarded as stable.

```
Adjusted Rand Index: 0.9658040528630559
```

Conclusion

Using PCA as a dimension reduction technique and performing k-means clustering has derived some insights on using word choices to identify the geopgraphic regions where the respondents are located. However, common language associates with a shared culture and heritage which might not shared by people clustering physically, because people in US move frequently. Because of this limitation among others, I am not confident that we can use this model to predict the future data. If we have the opportunity to collect the data from the survey, I wonder if we can think of labels to identify the culture and social clusters for respondents to choose from, instead of using their physical location as a proxy.

As far as reality check, the challenges with PCA is that the new projections are hard to explain intuitively. However, the questions about "water foundation" and "mow the lawn" identified the frequent and infreqent choices of words, which seems to align with the reality. "Sneakers" vs "tennis shoes" also seems to be used by people from distinct regions.

Academic Honesty

Statement

I affirm that the work presented in this report is my own, and I have not received unauthorized assistance from any individual or source. All external references and sources used have been appropriately cited and acknowledged, and I have followed the guidelines set forth by Prof. Yu regarding academic integrity and honesty. I understand that any violation of these principles may result in disciplinary action.

LLM Usage

Collaborators

Bibliography

[1] John Nerbonne and William Kretzschmar. "Introducing computational techniques in dialectometry". In: Computers and the Humanities 37 (2003), pp. 245–255.

[2] John Nerbonne and William Kretzschmar. "Progress in dialectometry: toward explanation". In: Literary and Linguistic Computing 21.4 (2006), pp. 387–397.