Ji Ahn, SID: 22981009
GitHub: jiahn

For the linguistics dataset, I chose questions that showed signs of regional separation. Midwestern areas seemed to yield very little responses in general across most questions, so I focused on comparing responses of western, eastern, and southeastern areas.

A good comparison question would be 119 and 64. 119 has data that points to 2 main clusters in the NY area and general mideast. When compared to 64 it is clear to see from their PCA scoring that although they both exhibit similar behavior in predicting each other, the regions are different thus their PCA scoring are on different quadrants (119 showing a general increase, where as 64 shows a fairly linear and obvious correlation.) When you sort the scoring for 64 and 119, 119 shows a clearly more evenly distributed weighting in components, whereas it is clear to see in 64 that the first few questions hold the majority of weight (first 3-4 PCs). When done with other questions, the first 2-3 questions tend to hold the majority of weight (and responses). Thus we decide to focus on the first 3 responses for our tests.

When we cluster our binary response data, we see the same pattern in questions with multiple clusters. Problem 119's cluster labels are evenly distributed with a mean of 1.72, sd of 0.448 while 64's labels indicate a mean of 1.081 and sd 0.273. When compared to the regional data we can see that there is one prominent regional cluster for 64 whereas there are 2 main clusters that overlap in 119 (with a skew towards cluster near NY, thus the mean of 1.72). Upon increasing k to higher levels, we see that the cluster groups increase (along with SD and means, as more labels are introduced.) However even when we increase k to great numbers, the mean stays relatively low (2~3 at k=25). What this indicates is that although there are many groups that can be organized and labeled differently, the majority of the weight lies on the first 1-3 clusters - as we predicted with our PCA. Thus, the clusters seem to be stable and reliable to run tests with, at least for the first 3 clusters, or questions with the most amount of responses.

When comparing regional factors (latitude, longitude) for continuum, responses can be divided across 4 quadrants. I've split them at the mean latitude and longtitudes (respectively 38.6, -88.) Quadrant 1 (upper right) I've labeled 1 (and then increasing to 4 clockwise.) The response groups correspond less strongly with increase k values, but provide satisfactory results when clustered with the original dataset, which encompasses all possible values and responses (thus leading to more stable clusters with an increased k value.) For predicting other questions' responses most data stayed in Quadrant 1, with occasional Quadrant 2 clusters (the southern area near Florida, Alabama.)However with subsamples of the dataset we can more accurately predict the predominant cluster group of a different question (ie. subset). However some questions have most of their responses located in areas other than Quadrant 1, which account for most of the inconsistencies provided when clustering using the entire dataset. With lower values of k, the clusters become more stable (evidenced by a more accurate linear regression on our PCA subsets.) The plots display a more uniform slope and regression line with a decreased value of k. As our k value increases the relationship between problems and their respective responses become hazier. Visually we can identify general clusters but when calculating explicit relations with binary response data, an increased number of groups in PCA and

clustering result in unclear data except for problems with two very distinct groups (ie. problem 64, 110).

An interesting factor to consider when looking at our linguistic data is population, and the lack of responses in the midwest. Because there is a vast lack of responses in the midwest and the west coast, we need to weigh responses from western areas with less weight, or even filter them out to accurately gauge and predict responses. This isn't completely necessary but I tried subsetting the data to exclude data in Quadrant 3/4 (midwest, west coast) when considering PCA and kmeans clustering. With the west coast dataset excluded, I got much stronger results in predicting responses of other questions from the response of another. Stronger clusters were seen in the NY, Florida, and Alabama area, which were previously less strongly weighted when calculated with the entire dataset. When I did the same calculations for the west coast data only, I got much weaker results due to most responses being clustered in the eastern coast. This regional separation provides us with an idea of how people will respond differently depending on eastern areas of living, but is not as applicable to the west coast.