Statistics 133 Final Project Analysis

We began our analysis by first writing some functions to map out every single question, color-coded by response, with the hopes that some obvious trends between geographic location and response might jump out at us. For this we used the original response data, rather than the binary data, since it allowed us to easily map each response to a different color, with palettes being generated by the RColorBrewer package. The results of that process can be found in the file all\_maps.pdf, and though it worked as we had wanted, we discovered that the huge number of responses made it nearly undecipherable. What was immediately clear, however, was that there were many questions, such as Q054 and Q081, where one answer was by far the most common across the entire country.

With this in mind, we decided to proceed by using the binary data to find which response was most common for each question, and to then plot it and see what trends emerged. The resulting file (most\_frequent.pdf) revealed that for many of the questions, the most frequent response would arise in the same approximate geographic location. For example, Q054.2, Q055.2, Q056.2, and Q057.2 are each the most popular responses to the questions, and each has a geographic distribution extremely similar to the others.

In order to get more rigorous results, we decided to apply hierarchical cluster analysis to these most frequent results, and see how stable the resulting clusters are. While running our code we discovered, unfortunately, that performing Euclidean distance and clustering calculations on all responses took an infeasible amount of time, and so decided to run it on a random sample of 3000 responses. We then ran the code with three levels of clustering: 8, 6, and 4 clusters. The resulting maps, most\_frequent\_k8.pdf, most\_frequent\_k6.pdf, and most\_frequent\_k4.pdf reveal a number of very stable clusters.