Lindsey Gao
Mentor: Sarah Teichman

Multivariate Data Analysis

This quarter I had the opportunity to study multivariate data analysis methods with Sarah Teichman; we covered principal component analysis, classical and non-metric multidimensional scaling, exploratory and confirmatory factor analysis, and cluster analysis. STAT 311 was the only statistics course that I took prior to this project, so we reviewed a lot of preliminary statistical concepts prior to tackling the methods. I also wanted to apply these methods to a data set after reading about them, so Sarah and I spent time finding a suitable data set that interested me. I am passionate about climate change and sustainability issues, so we decided on a 400+ feature data set from the NASA Earth Space database, which contained various indicators for how sustainable a country is. The initial data set had 30% of missing data so I had to tidy the data. Since each observation was a single country, I suspected that there were certain small/obscure countries/islands that made up a lot of the missingness. I dropped all of the observations that contained more than 300 missing values and then dropped features for the remaining observations that had more than 25% missingness. Depending on whether the range of possible values for a variable was large or small, I imputed either the median or mean for the remaining missing values. This process was tedious and I had never performed a large-scale imputation process, so this was great practice to learn.

I already had some exposure to PCA before starting this project, but I did not have much experience applying the method. I discovered by using this method on the data that the first principal component seemed to indicate how developed a country is. I also learned about multidimensional scaling, another type of spatial dimensionality reduction technique. Specifically, we read about classical multidimensional scaling and non-metric MDS. The former closely resembled PCA, but the latter was very new to me because it utilized a stress function. Prior to the reading project, I was very unfamiliar with concepts such as stress functions and maximum likelihood estimators. This project acquainted me not only with multivariate data analysis methods, but also statistical concepts in general, which I think was one of the most valuable parts of the program. I thoroughly enjoyed just having the opportunity to converse with Sarah every week about statistics. Through applying these methods, I learned also when certain models are not a great fit for the data. This was the case when I tried to apply factor analysis to my data, which gave me results that any number of factors between 1 and 20 were sufficient. It led me to think about why this may be the case and reexamine the assumptions of the model. I realized that factor analysis assumes that the manifest variables in the data are independent and depend only on these assumed factors themselves - which was not the case for my data.

Through presenting my findings on my Shiny application, I gained additional coding skills in debugging with Shiny as well as R. I initially went into this project thinking that I would mainly learn about multivariate data analysis methods, but I came out with a stronger comprehensive knowledge as a statistician and scientist.