Cluster Analysis

Jenn Le

12/4/2017

Abstract

This report outlines the results of various clustering methods on different featuresets of 2005 and 2013 federal employment data. Looking at how the clusters change between the two years help to detect differences between the terms of President Bush and President Obama's. There are noticeable differences between the two terms, including a shift towards higher pay early on in an employee's career with a higher level of education. An analysis of the Federal Aviation Administration reveals that while interesting exceptions have been found in this agency, it still follows the general structure of the government in general in regards to the observed features.

Contents

Business Understanding

Information regarding the differences between the terms of President Bush and President Obama assist us in understanding the trends over the years as well as the influence of each president on government employment. Observing the clusters will also show general trends and possibly reveal information that would not have been noticed otherwise. All of this information can be useful both to potential government employees as well as the government itself for future decision making. This can include potential employees figuring out how to make themselves more competitive when applying or the government changing recruitment tactics to better suit demographics they want to target.

Data Preparation

In preparation for different clustering methods, I first prepared the 2005 and 2013 employment data to suit the chosen methods. K-means clustering is a partitional clustering approach where each cluster is associated with a centroid and closeness is measured by Euclidean distance. Therefore, numerical data is best suited for k-means clustering. Hierarchical clustering also uses Euclidean distance by default so I aim to make the data suitable for that distance measure.

I made pay, age, length of service, and education continuous, choosing to keep the lower bound value where applicable. For supervisory status, I chose to make it binary to represent whether or not the employee has any type of authority. I then scaled all of these features. A description of the final features along with their appropriate distance measures are shown below in table 1.

I also decided to take a subset of one agency, the Federal Aviation Adminstration, that I noticed from previous projects to see how they differ between the presidential terms. From previous projects, it seemed like this particular agency has employees with higher compensation even with lower qualifications.

Feature	Scale	Distance Measure
Pay	Ratio	Euclidean Distance
Age	Ordinal	Euclidean Distance
Length of Service	Ordinal	Euclidean Distance
Education	Interval	Euclidean Distance
Supervisory Status	Binary	Euclidean Distance

Modeling

K-Means Clustering

For k-means clustering, I decided to use the features pay, education, age, length of service, and supervisory status since they are all already numeric and I wanted to keep the number of dimensions relatively low to avoid a loss of meaning with the Euclidean distances. With highly variable data such as the dataset we are looking at, using a higher number of clusters will allow more and more specific clusterings to emerge. As such, I intend to group the features into two employee characteristics in my analysis, general qualifications and compensation. In this case, that means age, length of service, and education versus pay and supervisory status. I use k = 4 clusters to try to get a good mix of low and high qualifications and compensation without getting too specific.

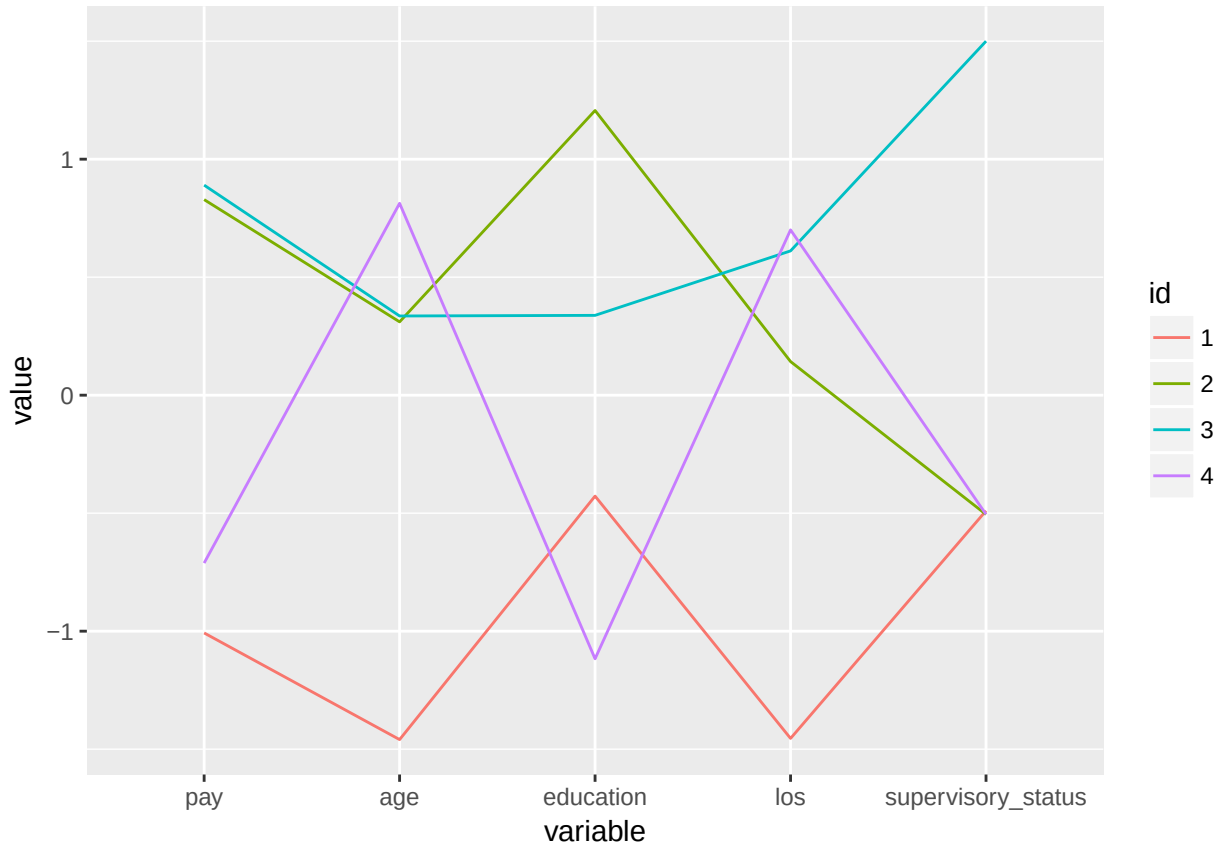

Figure 1: K-Means 2005

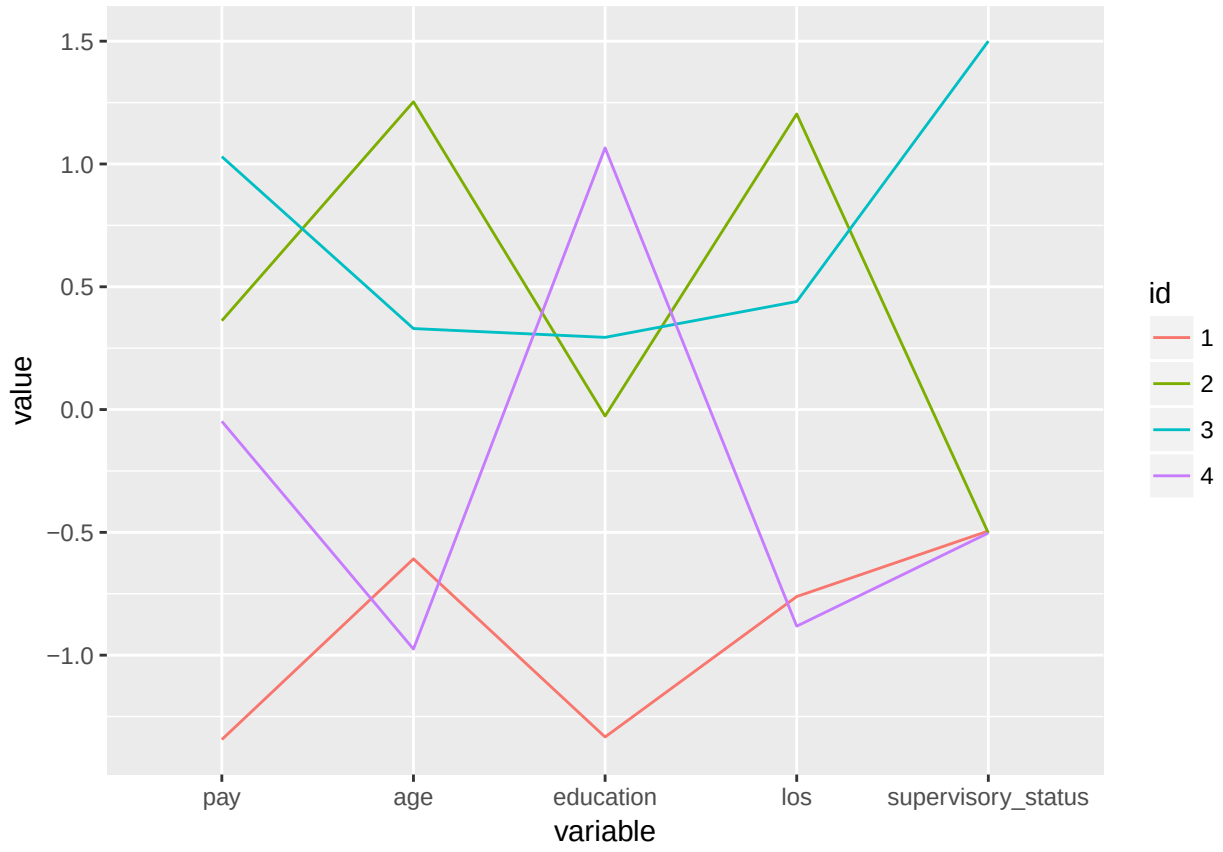

Figure 2: K-Means 2013

Figures 1 and 2 show the centroids for 2005 and 2013 clusters. The differences in these centroids reveal interesting information. In 2005, slightly above low pay is associated with low education when paired with higher age and length of service however in 2013, this is the opposite, a medium level salary is associated with high education but lower age and length of service. This could have been caused by an influx of newly hired, highly educated employees or by a shift in paradigm where salary is more reflective of education than in the past.

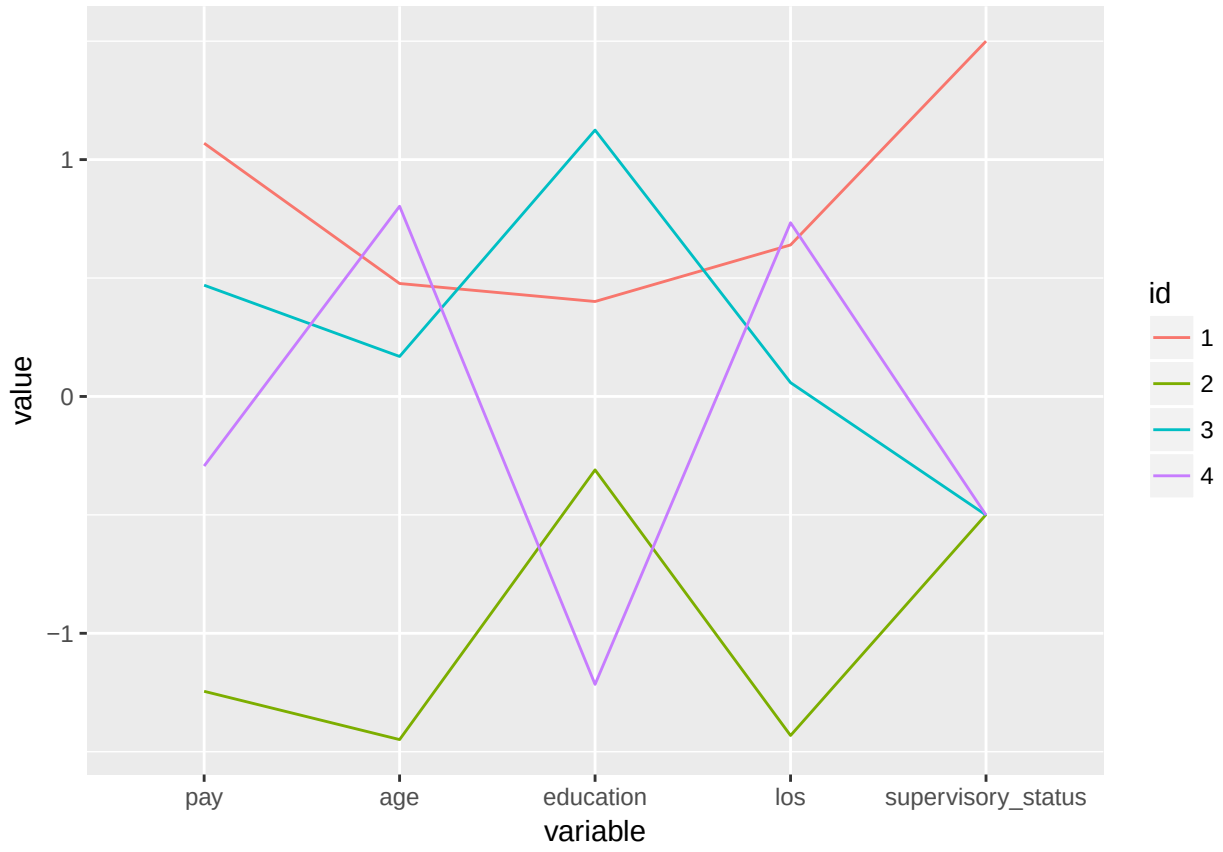

Figure 3: K-Means FAA 2005

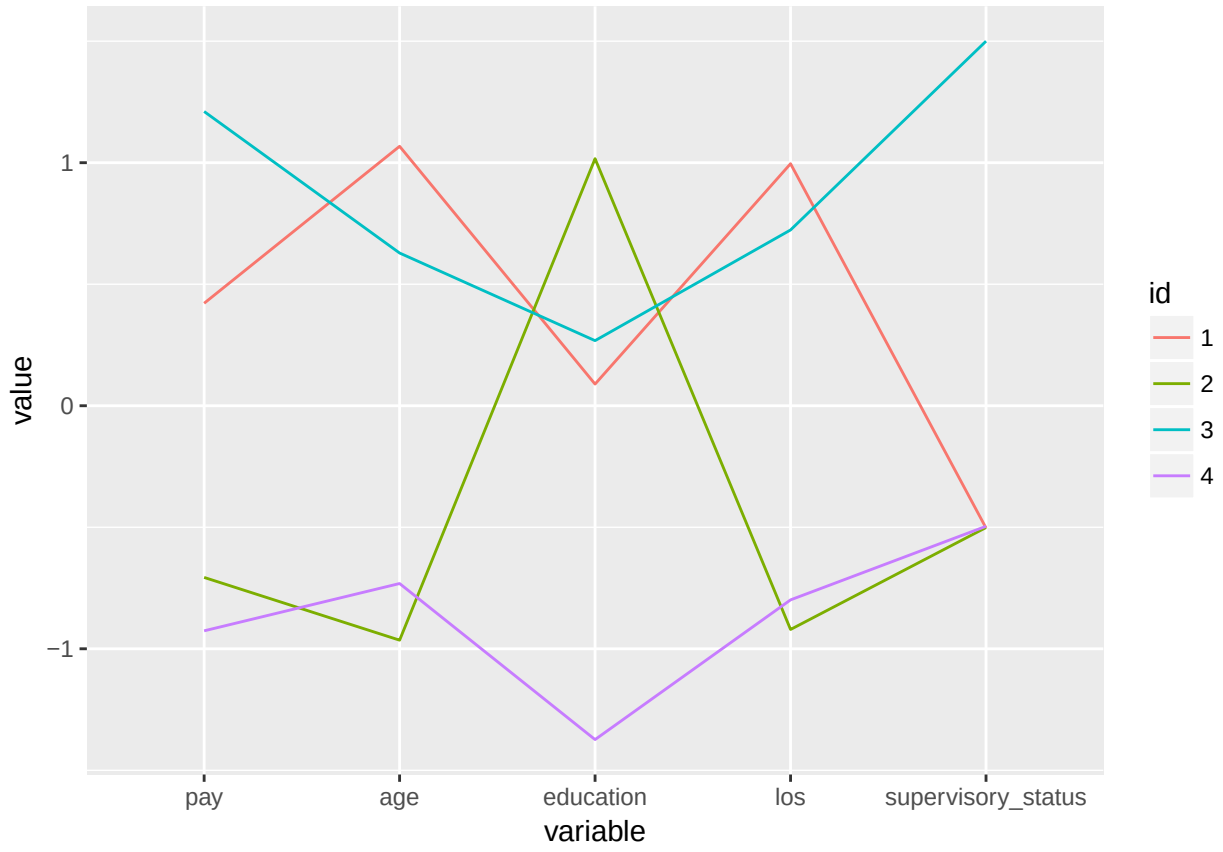

Figure 4: K-Means FAA 20013

Figures 3 and 4 show that the same general trends we saw also apply to the Federal Aviation Administration specifically. While we have seen in past projects that the FAA tends to pay higher wages even to less qualified employees, it still does follow the general clustering structure.

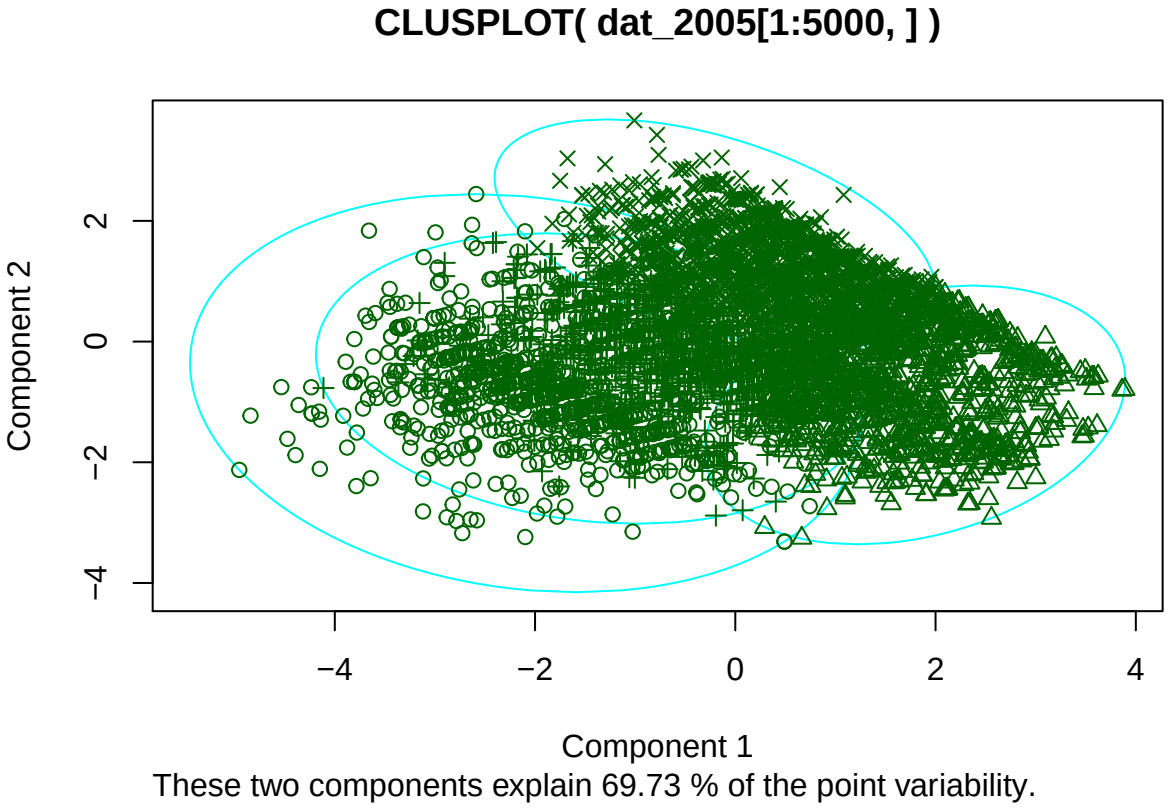

Figure 5: Cluster Plot K-Means 2005

Figure 5 shows a cluster plot of 5000 points on the 2005 data in order to give a better understanding of what the data generally looks like. As we can see from this small sample, there is a lot of overlap in the data and the clusters themselves reflect these overlaps.

Hierarchical Clustering

I used hierarchical clustering to cluster agencies in order to determine relationships between them. I will only be looking at agencies that have more than 20,000 employees because those are more relevant to the vast majority of potential government employees. It also makes it easier to analyze the clustering if we focus on fewer agencies.

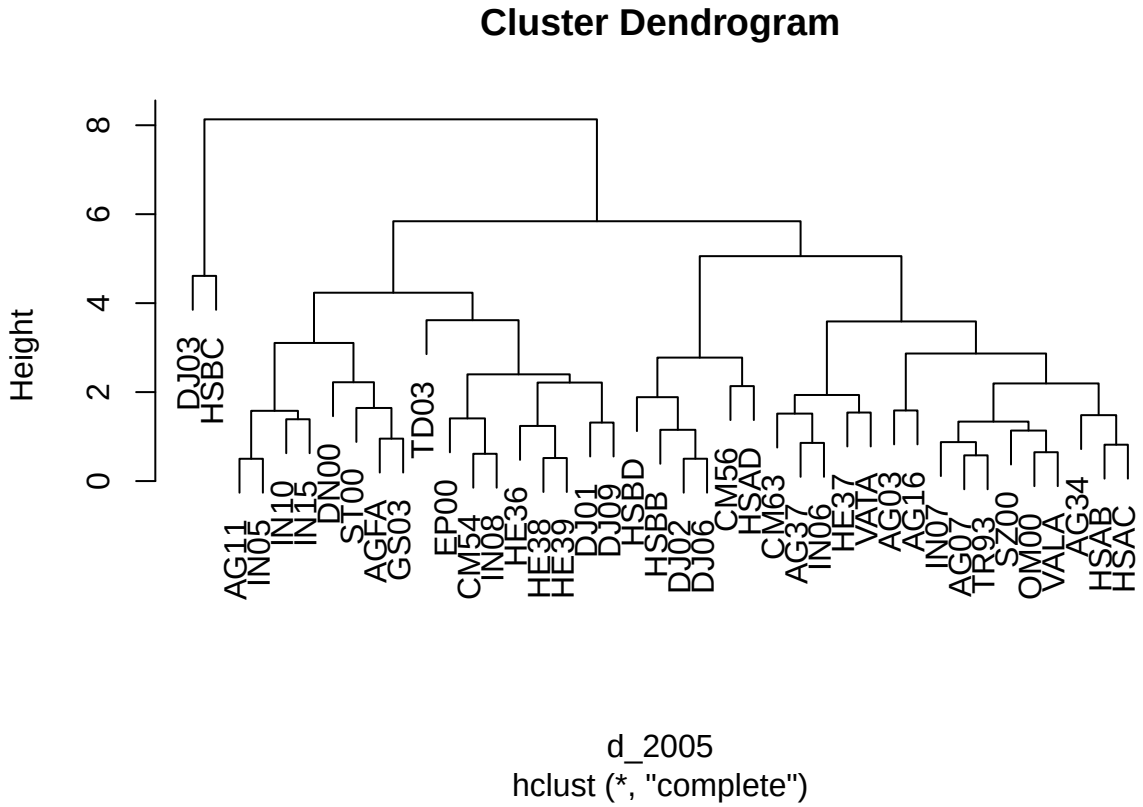

Figure 6: Agency Hierarchical Clustering 2005

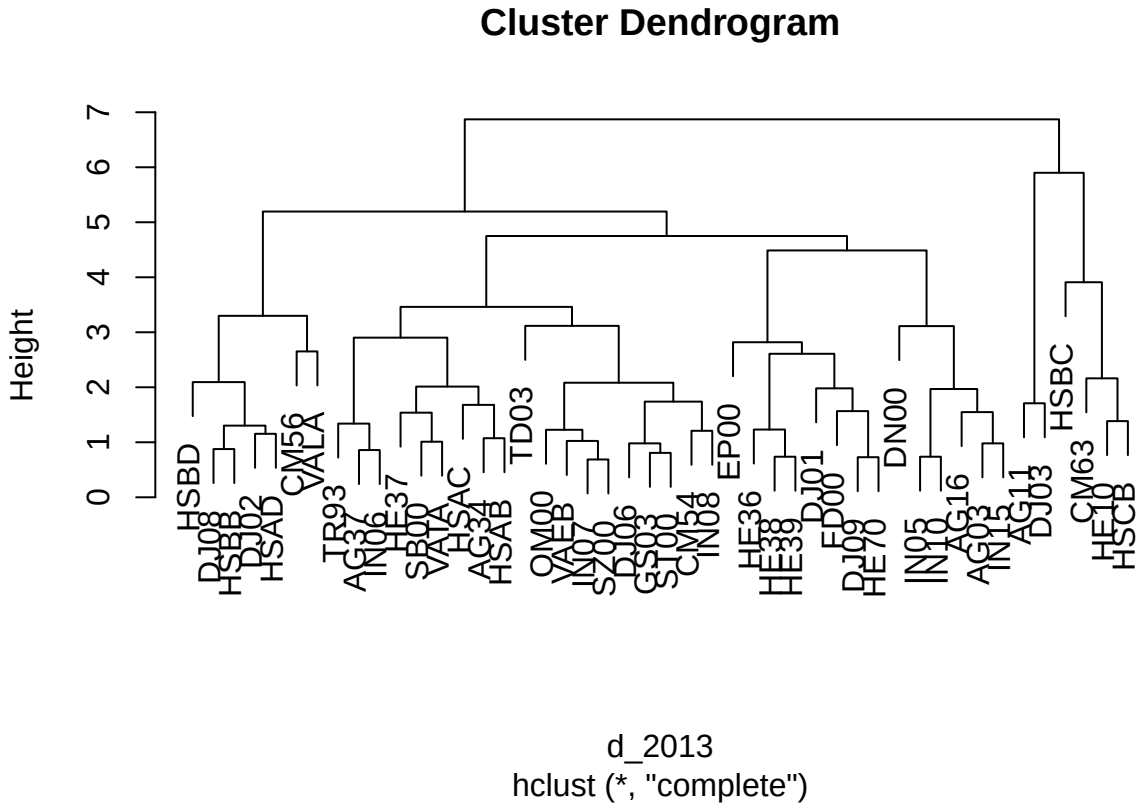

Figure 7: Agency Hierarchical Clustering 2013

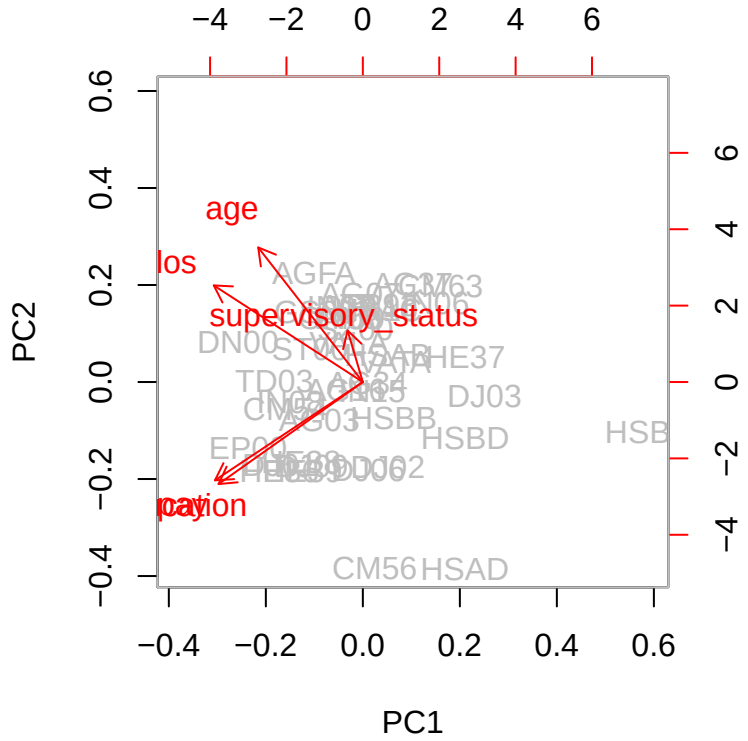

Figure 8: PCA 2005

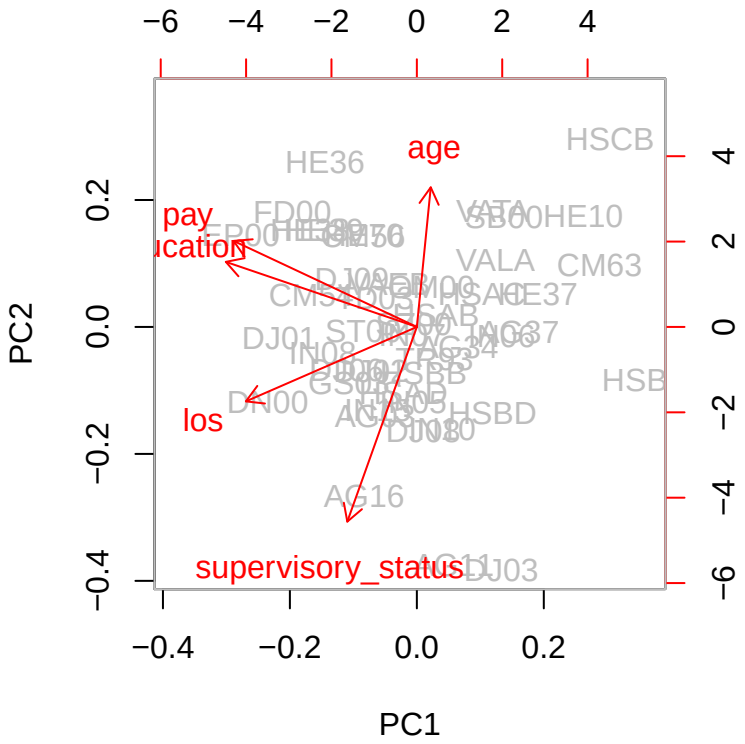

Figure 9: PCA 2013

Figures 6, 7, 8, and 9 show that the agencies are all very similar for the most part. This has to do with the fact that the data we are looking at is unique to individual employees rather than agencies therefore the clustering does not reveal anything about the struture or function of the agency. This clustering may have worked better if I had included information about states or if I had used a hamming distance measure on the agency codes. This could help reveal function as the agency code follows specific guidelines.

However, we do see differences between the 2005 and 2013 clusterings. In particular, age seems to be a lot more important in 2013. This is surprising considering the previous k-means clustering method shows the opposite.

Evaluation

The data that we examined is generally not suitable for clustering. Figure 5 shows that there are not really any distinct clusters that we can try to isolate. This is because the data we are looking at has a wide range of possible groups. While it is more likely that a person with higher education will have higher pay, there are also many instances where this is not the case and so this allows potential clusters to overlap.

The data I used to try to perform hierarchical clustering on agencies was also unsuitable for the task. This data included pay, education, and length of service, all features that are specific to individual employees and do not pertain to the structure or function of the agencies as a whole. Because of this, I was not able to get much information from this clustering.

Citations

Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Joseph Larmarange (2017). GGally: Extension to 'ggplot2'. R package version 1.3.2. https://CRAN.R-project.org/package=GGally

Hadley Wickham (2017). tidyverse: Easily Install and Load 'Tidyverse' Packages. R package version 1.1.1. https://CRAN.R-project.org/package=tidyverse

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.