Project 4

Jake Carlson

November 24, 2017

Abstract

Clustering

Contents

1 Business Understanding **3**

2 Data Preparation **3**

3 Modeling **3**
3.1 K-Means 4
3.2 Hierarchical 7

4 Evaluation **8**

Attribute	Scale	Values
Age	Interval	17, 22, 27, 32, 37, 42, 47, 52, 57, 62, 67, 72, 75
Education	Ratio	1 - 22
LOS	Interval	1, 3, 7, 12, 17, 22, 27, 32, 35
Pay	Ratio	1 - 401,000
SupervisoryStatus	Nominal	0, 1

Table 1: Final Data Set Attributes

1 Business Understanding

This report will focus on clustering analysis of the federal payroll data obtained by BuzzFeed News through the Freedom of Information Act. Specifically, k-means find groups of emplyees with similar attributes in the federal government and hierarchical clustering will be used to find agencies which have similar functions in the federal government.

Armed with an undestanding of what groups exist within each agency, agency leaders can work to create employee teams that are balanced with respect to the features each group holds. We can also see what features create the largest separation between the subgroups.

2 Data Preparation

I will prepare my data in a similar fashion to Project 3 [3]. I will take the middle of the age range for the age value of each employee. I will do the same thing for length of service. Education will be converted to an integer representing the number of years needed to achieve the degree the employee holds so that all education values are ratio scaled. I will also make supervisory status binary where a one indicates an employee is a supervisor and a zero indicates an employee is not a supervisor. The final list of attributes to be used for clustering is given in Table 1. Previous projects have shown that Age, Education, Length of Service, Pay, and Supervisory Status are the most relevant employee attributes.

All of these attributes are then scaled. With all of these attributes scaled, I can use Euclidean distance as my distance metric for clustering because all of the attributes are on the same scale.

3 Modeling

I will use two clustering methods to compare the structure of the government under Bush and Obama. First I will use k-means clustering. Then I will use hierarchical clusterging.

3.1 K-Means

K-means attempts to fit the data set by placing k cluster centers and moving them until they cease to move by an amount larger than a given tolerance. It is required that you specify the number of centers to use before the algorithm begins. I will start by first using $k = 3$. I believe this will form groups that represent upper management, middle management, and entry level positions.

The location of the cluster centers for the 2005 data set is given in Figure 1. Here, cluster 2 represents supervisors where the average Pay and Education for the group tends to be higher. The other two clusters represent non-supervisors. Cluster 1 represents employees who are newer and younger, as indicated by the lower Length of Service and Age values. Their Education and Pay is lower than that of cluster 2. Cluster 3 represents employees who are older and have worked for the government for longer, but are not supervisors. Their education is higher than cluster 1, but lower on average than cluster 2. These groupings indicate that employees in the middle age range have a higher likelihood of being supervisors than older employees. Supervisors also tend to have gone to school for longer than non-supervisors.

The cluster center locations for the 2013 data set is given in Figure 2. Again, there is one cluster, cluster 3, that represents the supervisors. Again, the supervisors have an age in the middle of the other two cluster centroids. Cluster 1 represents young employees with some education who are new to working for the federal government. Cluster 2 represents older employees who are not supervisors. This cluster centroid is placed at the lower end of the education axis. I would expect more of the older employees to have a higher level of education, so there is a posibility that the centroid for Cluster 2 is getting stuck in an akward position. To examine this further, I will rerun the clustering for 2013 using $k = 5$ to see if I can get clusters that are more representative of the underlying distribution of employees.

The cluster centroids for 2013 with $k = 5$ are given in Figure 4. We see that the oldest employees have been split by two clusters. In Cluster 4, employees habe a slightly higher education on average and have worked for the government for much longer than employees in Cluster 5. In this clustering, Cluster 3 contains the youngest employees, and Cluster 1 contains employees in the middle of the age range who have the highest levels of education.

It still feels like there are more groups hidden in these clusters. To examine how many groups is appropriate for k-means, I will run the algorithm for a varying number of of groups and look at the within sum of squares. By plotting these values and looking for the sharpest turn, we can determine the optimum value for the number of groups to use to cluster this data set using k-means. This plot is given in Figure 5. The sharpest turn is where $k = 8$, however, the slope increases fairly consistently over the number of clusters, indicating there is not a single number of clusters that emerges as best for this data set. This is likely due to a lack of structure in the data, and must be explored further.

The within cluster sum of squares for $k = 3, 5, 8$ is given in Table 2.

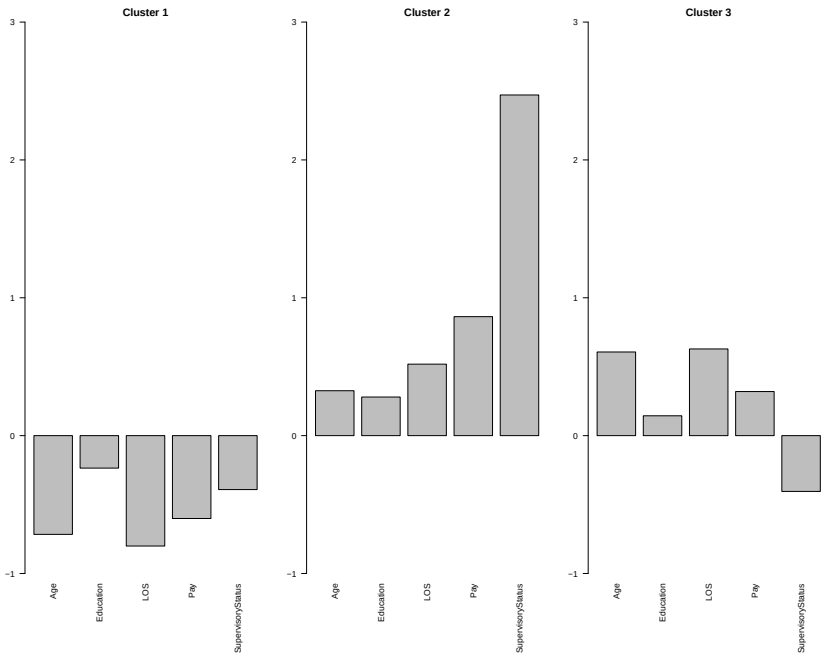

Figure 1: The Location of Cluster Centers for 2005 where k = 3

k	Within Sum of Squares
3	42.6%
5	59.3%
8	69.3%

Table 2: Within Cluster Sum of Squares for 2013

This shows that as the number of clusters increases, the variation within a cluster decreases. This makes sense, as increasing the number of centroids will make it easier for points to sit closer to an existing centroid. However, it is easier to understand the data if there are fewer clusters. That is why I would recommend using three clusters for k-means, as this produces clusters with easy to understand and identifyable employee attributes.

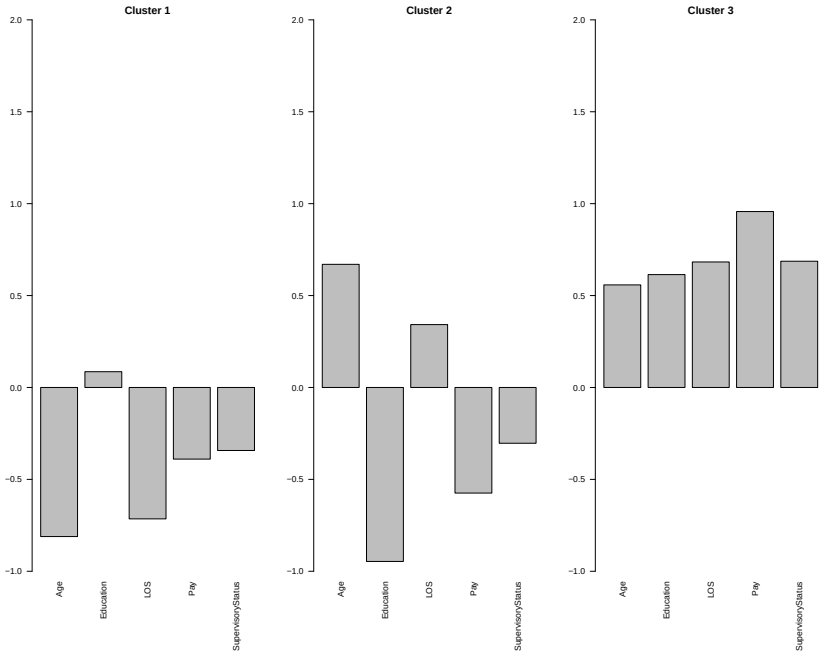

Figure 2: The Location of Cluster Centers for 2013 where k = 3

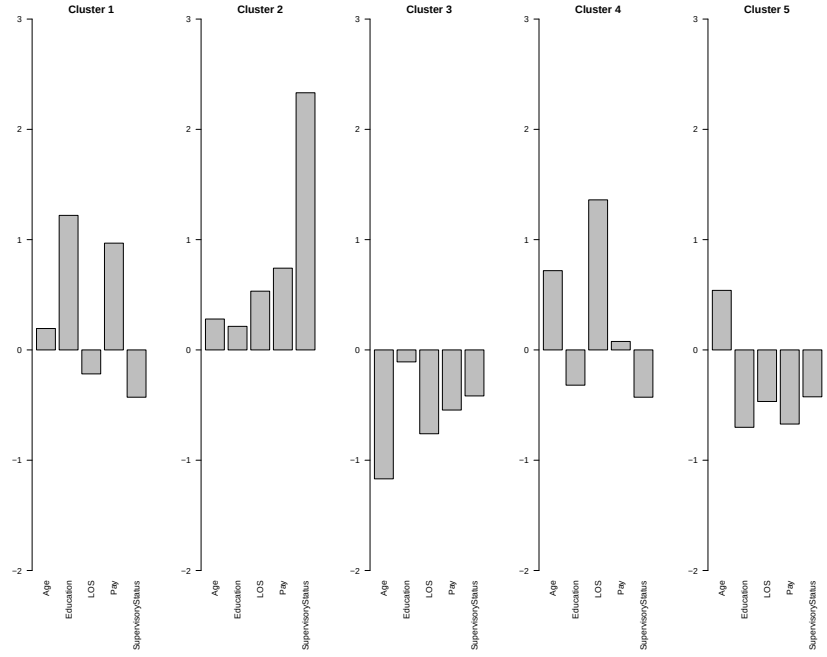

Figure 3: The Location of Cluster Centers for 2013 where k = 5

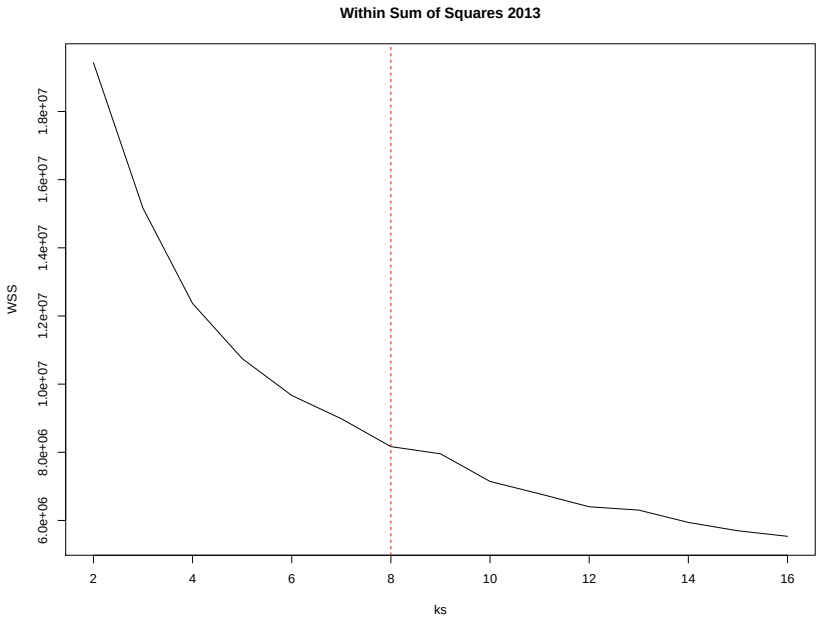

Figure 4: Within Sum of Squares for 2013

3.2 Hierarchical

In hierarchical clustering, a hierarchy of clusters can be formed by joining neighboring clusters in sequence. I will use complete link clustering, which assigns each data point its own group and joins the closest groups together. Here, closest is defined as the Euclidean distance between two points, or cluster centroids. For this process, I want to cluster agencies, so I will perform some additional processing to prepare agency data. I will take the mean Age, Education, and Pay for each of the three group identified using k-means. The goal here is to find clusters of agencies which carry out similar functions in the federal govenrment.

Looking at the dentrogram for 2005, given in Figure 5, it would appear that there are 8 groups that form. The agencies are projected onto two principal components in Figure 6 with the clusters drawn for visualization. Some of the groups contains agencies that are related. For example, Group 5 has the Office of Management and Budget, the Council of Economic Advisors, the Commodity Futures Trading Commission, the Federal Trade Commission, the Securities and Exchange Commission, and the Office of the U.S. Trade Representative. These agencies all have to do with financial regulation and policy, but this group also contains the National Science Foundation and the Office of Science and Technology Policy. These two agencies are also realted to each other, but they have little to do with the financial regulators. This grouping worked out fairly well. Unfortunately, the remaining groups have little to do with the purpose of each agency.

For example, Group 4 has the Arctic Research Comission, the Nuclear Waste Techinical Review Board, the Marine Mammal Commission, the national Council on disability, the Federal Mine Safety and Health Review Commission, and the Medicare Payment Advisory Commission. These agencies are loosly related at best.

Looking for agencies AN, AW, BK, BW, CX, FK, GO, GY, MA, NK, OS, RS, ZL (Group 4) in the first quarter of Figure 7 shows why this group is separated. These agencies have employees with slightly higher education levels, more pay, and that are older. This graph reveals the main issue with clustering agencies this way. The problem here is that the selected attributes for each agency represent features of employees, and have little to do with the structure or functionality of the agency.

It would be necessary to incorporate additional data about each agency that describes features such as funding, organization, the number of appointed positions, and some description of what each agency does for the nation. A more robust clustering could potentially be obtained by examining the agency names and clustering by common words, but this clustering would be crude.

The attributes I choose are not great for clustering agencies because all of the agencies are similar. Agencies have a similar distribution of education and age. Also, all of the agencies follow the General Schedule which clearly defines how much an employee is to be paid based on their attributes.

This clustering also determined some agencies to be outliers. Groups 7 and 8 only contain one agency each, and they fall far away from the remaining agencies. The Commission on Review of Overseas Military Structure (YA, Group 8) has average pay and education, but much older employees than the rest of the agencies. The International Boundary Commission: U.S and Canada (GX, Group 7) has lower paid and younger employees than all of the remaining agencies by a substantial margin.

4 Evaluation

References

[1] Jake Carlson *CSE 5331 - Data Mining Project 1* `https://github.com/jakecarlson1/data-mining-projects/blob/master/project-1/report/carlson-project-1.pdf`

[2] Jake Carlson *CSE 5331 - Data Mining Project 2* `https://github.com/jakecarlson1/data-mining-projects/blob/master/project-2/report/carlson-project-2.pdf`

Figure 5: Agency Dentrogram for 2005 with 8 Groups Highlighted

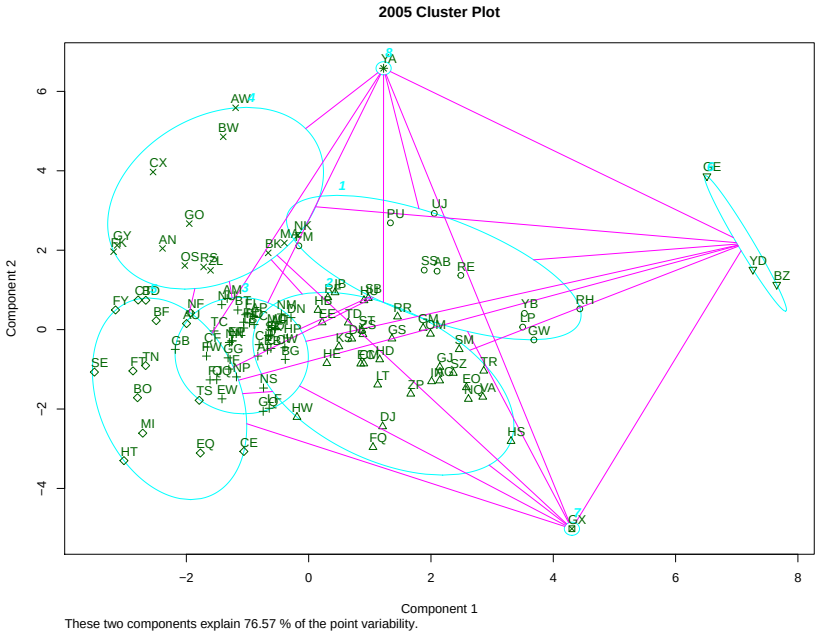

Figure 6: 2005 Agency Clusters Projected onto Two Principal Components

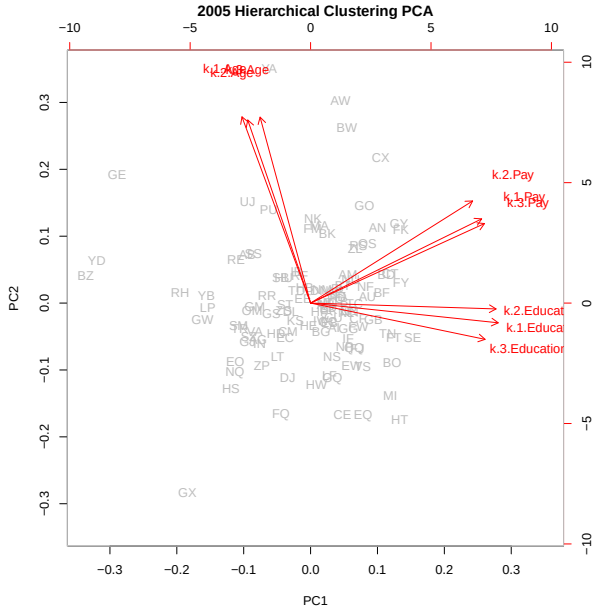

Figure 7: 2005 Agency Clusters Projected onto Two Principal Components

[3] Jake Carlson *CSE 5331 - Data Mining Project 3* `https://github.com/jakecarlson1/data-mining-projects/blob/master/project-3/report/carlson-project-3.pdf`

1. Group 1: AMERICAN BATTLE MONUMENTS COMMISSION, FED MEDIATION AND CONCILIATION SERVICE, INTERNAT BOUNDARY & WATER CMSN: US & MEX, GOVERNMENT PRINTING OFFICE, PEACE CORPS, OFC OF NAVAJO AND HOPI INDIAN RELOCATION, ARMED FORCES RETIREMENT HOME, SELECTIVE SERVICE SYSTEM, JAPAN-UNITED STATES FRIENDSHIP CMSN, ANTITRUST MODERNIZATION COMMISSION
2. Group 4: AFRICAN DEVELOPMENT FOUNDATION, ARCTIC RESEARCH COMMISSION, JAMES MADISON MEMORIAL FELLOWSHIP FOUND, NUCLEAR WASTE TECHNICAL REVIEW BOARD, NAT CMSN ON LIBRARIES AND INFO SCIENCE, FARM CREDIT SYSTEM INSURANCE CORPORATION, VIETNAM EDUCATION FOUNDATION, INTERNATIONAL JOINT CMSN: U.S. & CANADA, MARINE MAMMAL COMMISSION, NATIONAL COUNCIL ON DISABILITY, OCCUPATIONAL SAFETY & HEALTH REVIEW CMSN, FED MINE SAFETY AND HEALTH REVIEW CMSN, MEDICARE PAYMENT ADVISORY COMMISSION
3. Group 5: FEDERAL LABOR RELATIONS AUTHORITY, MERIT SYSTEMS PROTECTION BOARD, DEFENSE NUCLEAR FACILITIES SAFETY BOARD, OFFICE OF MANAGEMENT AND BUDGET, COUNCIL OF ECONOMIC ADVISERS, COMMODITY FUTURES TRADING COMMISSION, COUNCIL ON ENVIR QUAL/OFC OF ENVIR QUAL, FEDERAL TRADE COMMISSION, FEDERAL HOUSING FINANCE BOARD, HARRY S. TRUMAN SCHOLARSHIP FOUNDATION, MILLENNIUM CHALLENGE CORPORATION, NATIONAL SCIENCE FOUNDATION, SECURITIES AND EXCHANGE COMMISSION, OFFICE OF THE U.S. TRADE REPRESENTATIVE, OFFICE OF SCIENCE AND TECHNOLOGY POLICY
4. Group 6: CHRISTOPHER COLUMBUS FELLOWSHIP FOUNDATN, BARRY GOLDWATER SCHOL & EXCEL IN ED FOUN, HELP ENHANCE LIVELIHOOD OF PEOPLE CMSN
5. Group 7: INTERNAT BOUNDARY CMSN: U.S. AND CANADA
6. Group 8: CMSN ON REV OF OVERSEAS MIL STRUCTURE

Figure 8: Agencies in Each Hierarchical Cluster for 2005

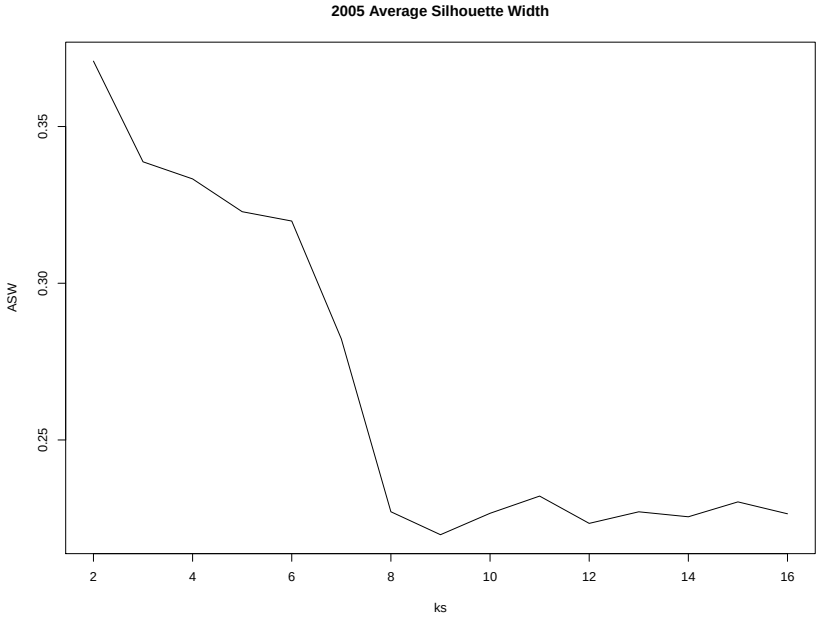

Figure 9: 2005 Average Silhouette Width