CLUSTER ANALYSIS

CSE7331 Project 4

Zhang, Shiyang

SMU ID:47319809

Content

Content of table and figure

0.Abstract

In this project, I use some clustering method to cluster data set. I use data 2013-03. I mainly use K-MEANS and Hierarchical clustering to analyze. At the end of each model, I use some methods to do numbers of clustering selection and internal validation comparison for different method.

Here are some interesting findings:

1. For most agencies, the promotion principle or salary treatment are similar. Only few are special. So most agencies in PCA image gather together.
2. There are different ways to determine number of clusters. Parameters are also different. Sometimes silwidth is high, but BSS/TSS is small. Sometimes is opposite. Therefore, determine the number of clusters depends on different situation.
3. People in different age have different opportunities to get to work. For example, some agencies are well come to young people. But others are not so welcomed. Older people are more popular, maybe for they are more experienced.
4. Older people earn more than young people.

1.Data preparation

1.1 Feature selection

At first, do basic cleaning for invalid data set. Using NA replace unknown characters like '########',
'UNSP', '*'. Then impute NA-Age with median age of the same agency. Impute NA-Pay with median
pay for the Age of the employee at that agency. Then drop NA pays. Finally, dealing with values have
duplicate IDs:
1. Select rows with duplicate IDs
2. Order selection by ID, then Agency, then descending Pay
3. Select rows where the ID and Agency are duplicated (same employee at same agency),
4. Get row numbers for rows with the lowest pay for each grouping in the above selection
5. Reselect from data frame where rows are not required to removed

Secondly, I set some features into null since there are not so important. Like ID, Name, Date, AgencyName, Schedule and NSFTP.

The features I want to use for clustering is Age, LOS, Education, SupervisoryStatus, Pay, Agengy. Since these features are quite important features for pay according to previous projects. In addition, Many of these attributes are ratio or ordinal variables. It is convenient for computing distance between variables. For pure numeric values, Minkovsky distance (Manhattan distance, Euclidean
Distance) is appropriate, especially Euclidean distance. For binary variables, Jaccard index is appropriate. For mix data(nominal, ratio, ordinal), Gower's distance can be used to compute.

	Scale	Description
Age	ratio	Use the first number represents the range
Education	ordinal	Use numbers represents education background
SupervisoryStatus	ordinal	Consists of manager, supervisor and other
Pay	ritio	The income of people
Agency	nominal(binary)	Sometimes are dummy binary values
LOS	ratio	Use the first number represents the range

Table 1 Data preparation

2.Modeling

2.1 K-Means clustering

In this part, I use Age, LOS, Education, Supervisory Status, Pay as features to compute distance and build model. Since these features are numeric, Age is ratio, LOS is ratio, Education is interval, Supervisory status is nominal attribute but in the form of ordinal. Pay is ratio. So the clustering method I use is k-means. Because of the compexity and time comsuming for data computation, I sampled 10,000 data for clustering and evaluation.

I set number of clusters k as 4. The center for each cluster:

	Age	LOS	Education	SupervisoryStatus	Pay
1	0.2837730	0.5542264	0.225180927	2.4729276	0.83814735
2	-0.5243351	-0.6465299	0.788329263	-0.3049075	0.04651584
3	0.8153422	1.2259931	-0.002217123	-0.3758486	0.34499103
4	-0.2273863	-0.5175828	-0.986957071	-0.3952813	-0.70504680

Table 2 Center of clusters

SSE validation:

WSS	23512
BSS	26340
TSS	49852

Table 3 SSE validation

BSS/TSS = 0.52836

Relationship between variables of four clusters:

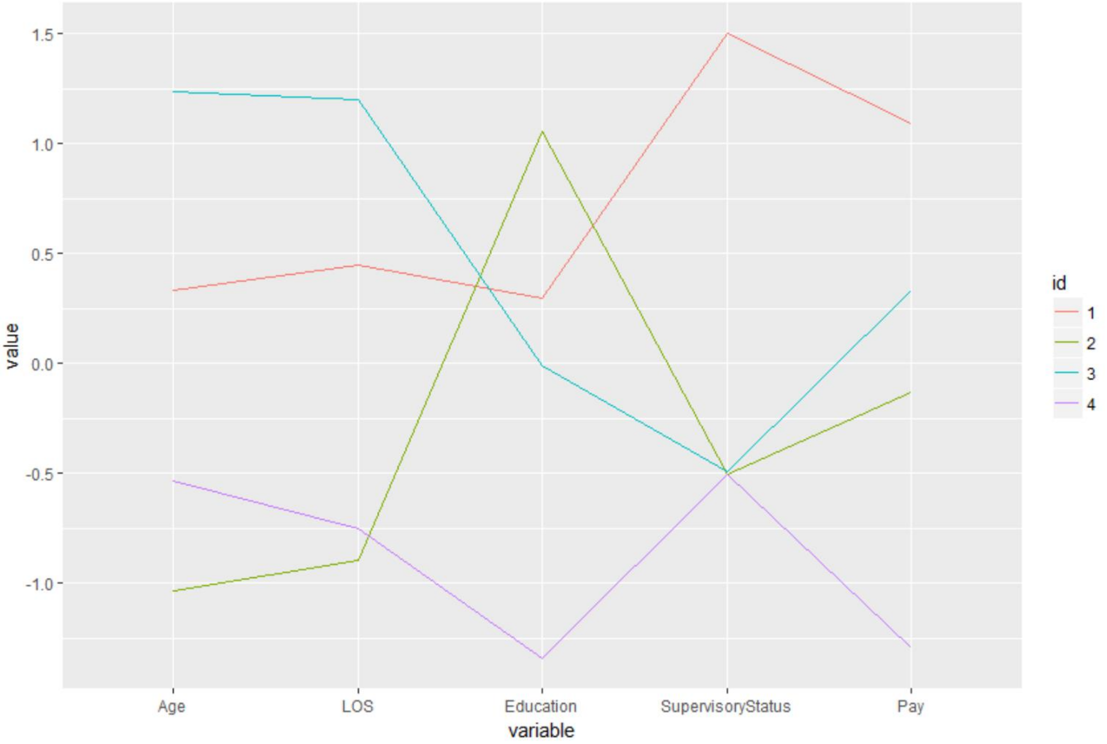


Fig 1 Four clusters' comparison

In this image, we can see that values are divided into 4 clusters, green represents people with small age, purple represents people with second smallest age, red one represents people with second largest age, blue one represents people are oldest. In this figure, LOS of each cluster perform normal, it seems like that old people have larger LOS, young people have smaller LOS. The Education shows that young people has higher education, this result show the same conclusion I made in project 1. Young people lay more emphasis on education background. It is very interesting that three clusters has the same supervisory Satus with relatively lower salary. One cluster with different supervisory status has higher salary. It may because this group has more manager or supervisor as it has higher value(opposite to original value).

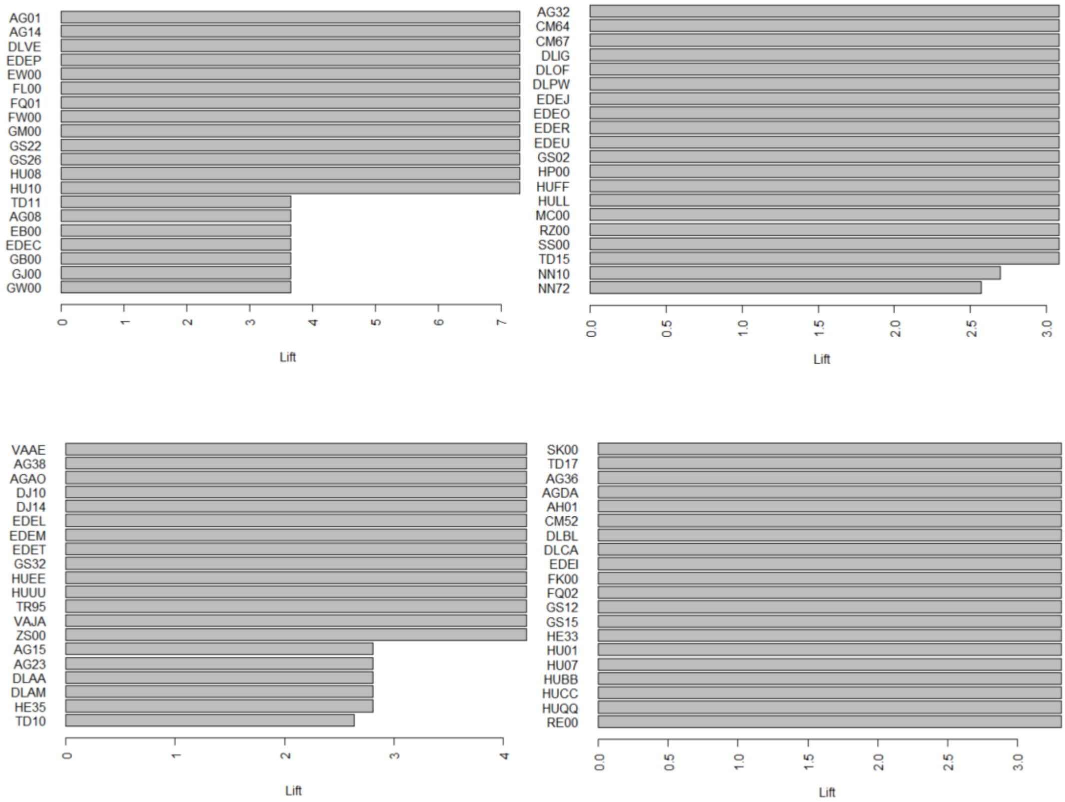


Fig 2 Work chance in different clusters

This figure shows work chance for people in different clusters. From left to right, top to down represents cluster1, cluster2, cluster3, cluster4. It shows that different agencies have different working chance. In this case, I sampled 10,000 elements that are divided into 4 clusters. It seems like that many agencies have similar and high lift. Especially in cluster 4. It may because that those people are experienced that many agencies like them. Meanwhile, maybe the 10,000 is not quite accurate, the distribution of agencies does not vary a lot.

Model validation and cluster determine

In this part, I use several ways to do internal validation to compare modeling methods and to determine number of clusters.

Number of clusters:

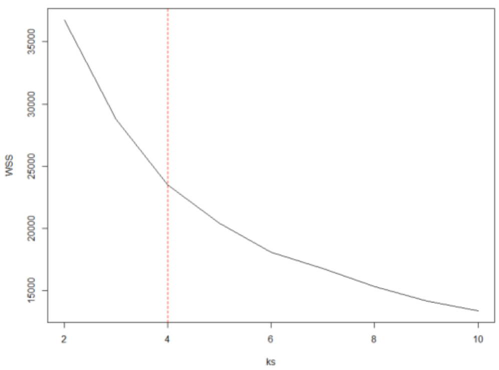


Fig 3 WSS validation

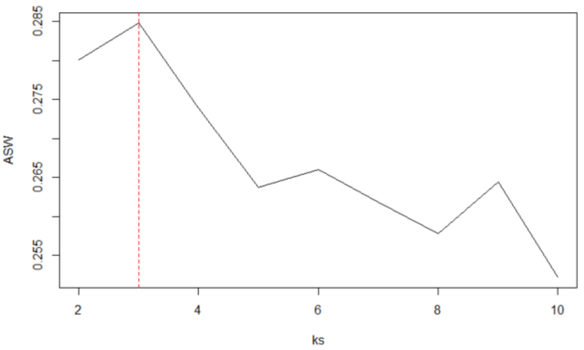


Fig 4 ASW validation

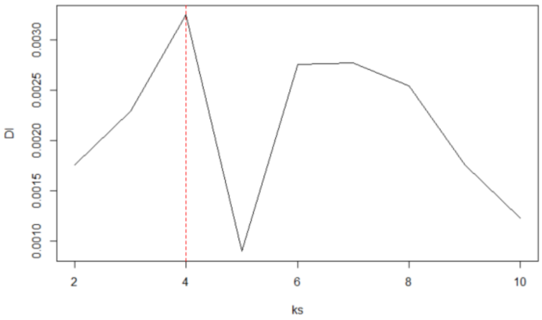


Fig 5 DUNN validation

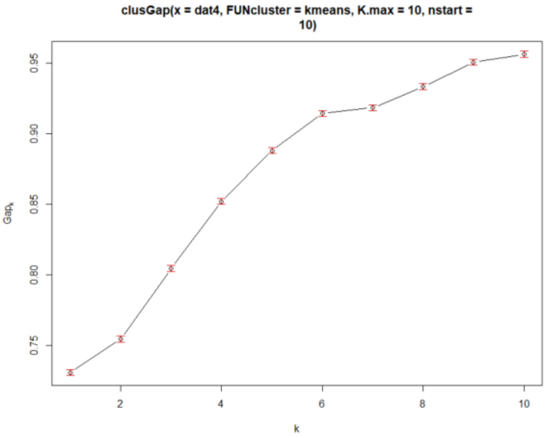


Fig 6 Gap validation

In these four figures, it shows that 3 and 4 clusters are suggested numbers. In figure 1, I do not thinks that it is meaningful because the curve is very smooth without a steep turn. In figure 4, it also does not show some key point. I think ASW and DUNN are more meaningful. So, both three or four clusters are appropriate.

When cluster = 3, the cluster quality is shown as follow:

Cluster size	3531,1390,5079
Cluster average silwidths	0.1864772,0.3161878,0.3444955
Average between	3.461446
Average within	2.132397
Average silwidth	0.2847645

Table 4 some cluster quality features when cluster=3

Average BSS/TSS = 3.461446/5.593834=0.6188

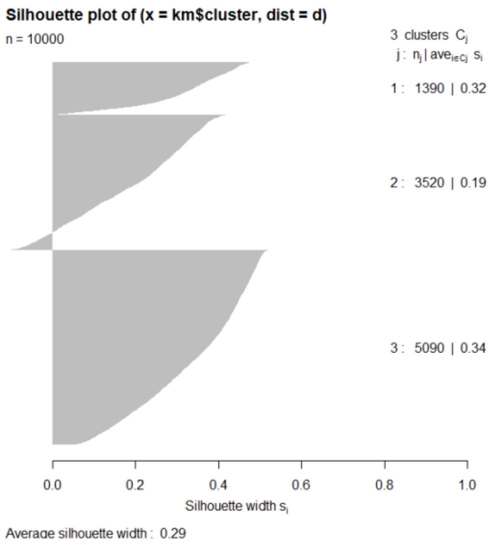
When cluster = 4, the cluster quality is shown as follow:

Cluster size	2334,2703,1362,3601
Cluster average silwidths	0.1891050,0.2453329,0.2914007,0.3438305
Average between	3.31479
Average within	1.909239
Average silwidth	0.2739527

Table 5 some cluster quality features when cluster=4

Average BSS/TSS = 3.31479/5.224029=0.6345

Two silhouette plots are shown as follows:



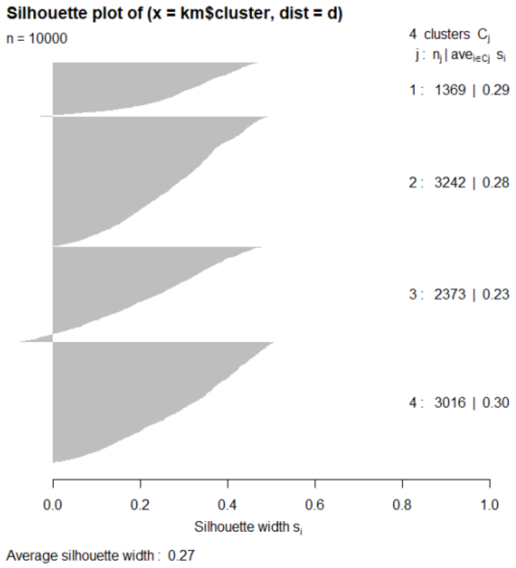
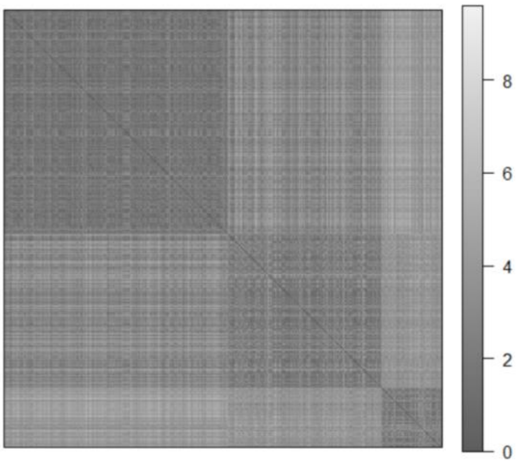


Fig 7 silhouette plot when cluster = 3 Fig 8 silhouette plot when cluster = 4

The distance of points is visualized as follow:



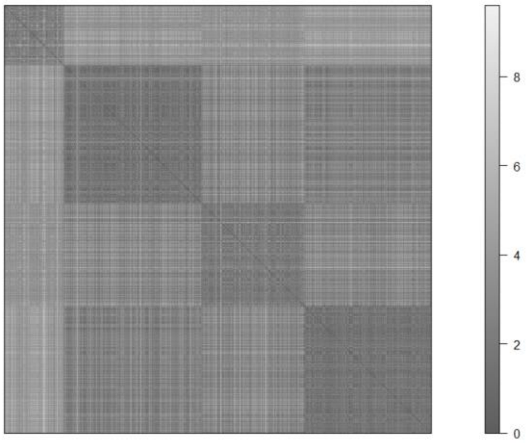


Fig 9 distance when cluster=3 Fig 10 distance when cluster=4

From above statistics, we can see that 3 cluster's BSS/TSS is smaller than 4 cluster's. But the average silwidth of 3 clusters is bigger than 4. The quantity distribution of 4 is more even. In conclusion, I think 4 cluster is better.

Following table shows cluster quality that different clustering methods used for this subset:

	km	Hc_complete	m
Within.cluster.ss	23511.75	36818.08	22307.52

Avg.silwidth	0.274577	0.1985241	0.1431737

Table 6 Different methods for clusters=4

In table 6, km represents K-means Clustering, hc_complete represents Hierarchical Clustering, m represents Gaussian Mixture Models. From WSS and silwidth, K-means is the best method used to do this model.

2.2 Hierarchical clustering

In this part, I use Age, Education, Grade, LOS, Pay, Supervisory status, agency, size as features to do the clustering. It mainly focus on different agencies, and compute each agency's median age, education, grade, los, pay and supervisory status as its value. I think median is better than mean because some decimals do not mean anything. Such like supervisory status and grade. Also, median can get rid of some influence of unremoved outliers.

Since there are only 373 agencies in this data set. I use hierarchical clustering because this model cannot hold large quantity numbers. The cluster dendrogram shows as follow:

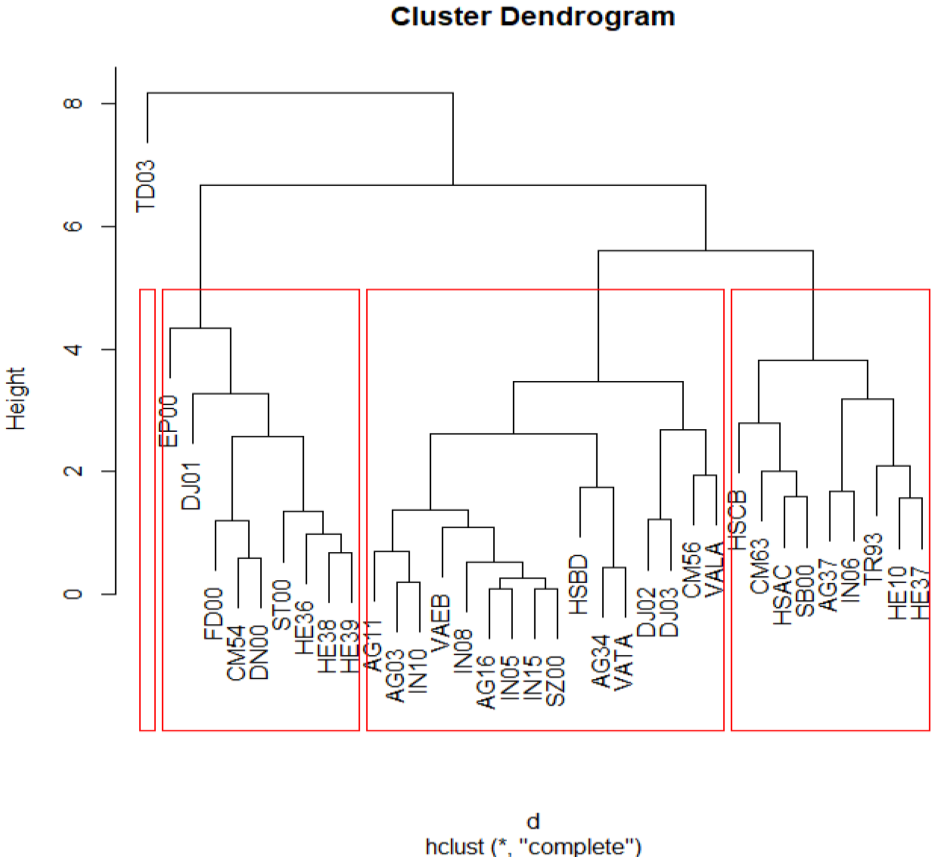


Fig 11 Hierarchical clustering dendrogram

In fig 11, I get rid of ‘size’ feature, and only grab agencies have more than 5000 people. It shows that many agencies with similar name close to each other. Like HE10 and HE37. Maybe people with the same age, los and education condition will have similar pay in these two agencies. The red line shows that I separate this data into four cluster. Focus on the left most agency, itself belongs to a cluster. This agency is far from other agencies. Maybe this agency is quite different from others.

Model validation and cluster determine

I use ASW, DUNN to determine the number of clusters.

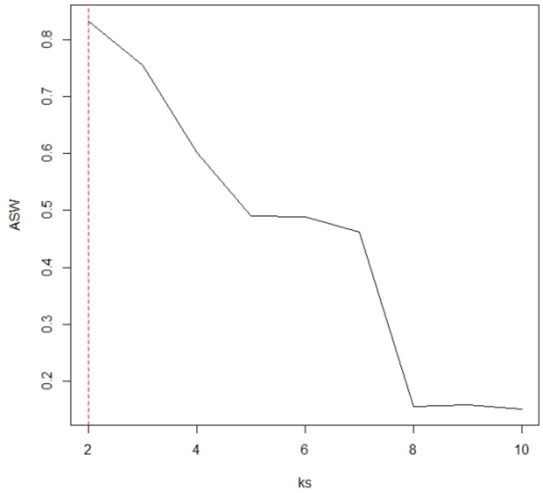


Fig 12 ASW validation

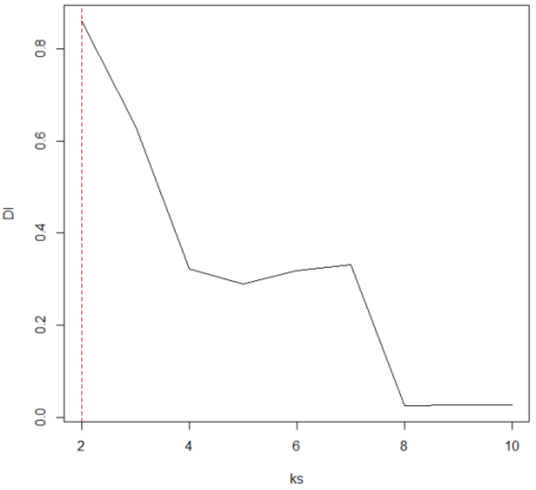


Fig 13 DUNN validation

Both ASW and DUNN show 2 is an appropriate number for clustering. But the distribution of size of each cluster is weird. The statistic is shown as follow:

Cluster size	372,1
Cluster average silwidths	0.8353962, 0
Average between	17.98316
Average within	2.972399
Average silwidth	0.8331565

Table 7 some cluster quality features

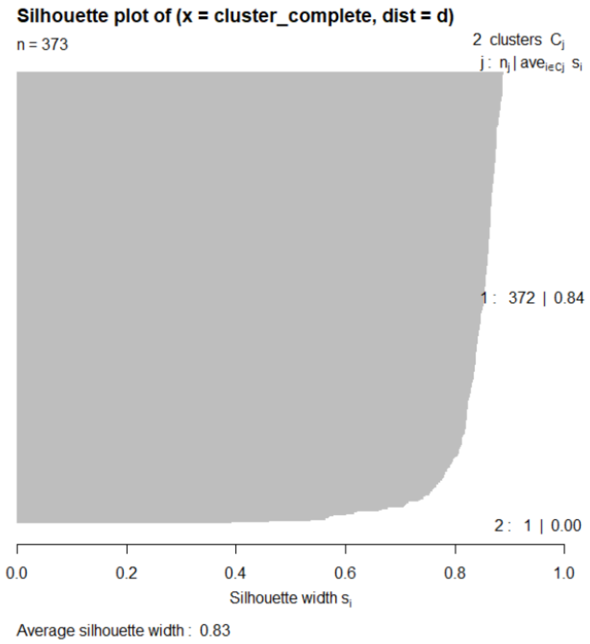


Fig 14 silhouette plot

	km	Hc_complete
Within.cluster.ss	2335.345	2287.244
Avg.silwidth	0.152753	0.8331565

Table 8 Clustering quality comparison

From above statistic, cluster of 2 is quite strange but the silwidth is pretty high. Maybe it because one agency is quite far from other agencies.

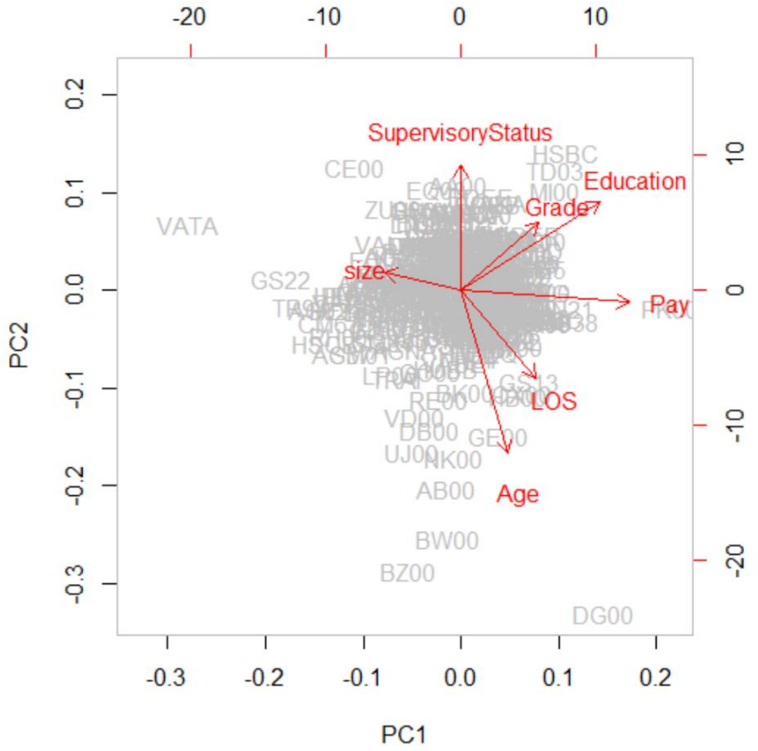


Fig 15 PCA image

In fig 15, this image shows the correlation between different values. It looks like that grade and education has some positive relation. The pay and size have negative relation. Maybe small company pays people more. The LOS and pay have positive relationship. And most agencies gather together at one point. Only few agencies are separate. So maybe most agencies have similar principle for promotion.

3. Conclusion

In this project, we use some clustering method to cluster points. I have some conclusions:

1.For most agencies, the promotion principle or salary treatment are similar. Only few are special. So most agencies in PCA image gather together.

2.There are different ways to determine number of clusters. Parameters are also different. Sometimes silwidth is high, but BSS/TSS is small. Sometimes is opposite. Therefore, determine the number of clusters depends on different situation.

3.People in different age have different opportunities to get to work. For example, some agencies are well come to young people. But others are not so welcomed. Older people are more popular, maybe for they are more experienced.

4.Older people earn more than young people.

4.Reference

1. https://smu.instructure.com/courses/37393/discussion_topics/53763
2.
http://michael.hahsler.net/SMU/EMIS7331/data/federal_employment/project4.html
3.
https://rawgit.com/mhahsler/Introduction_to_Data_Mining_R_Examples/master/chap8.html#internal-cluster-validation
4.
http://michael.hahsler.net/SMU/EMIS7331/slides/chap8_basic_cluster_analysis.pdf