**CS3907/CS6444 Big Data and Analytics**

**Fall 2018**

**Class Project #2**

**Due: October 22, 2018**

**Group 5**

**Group Member:**

**Yunfei Xu G21534523**

**Xiaohan Li G47313418**

**Siyuan Huang G43575455**

**Yanlong Shi G25529891**

Exploring Variations in Clustering and Predictive Analysis

1. Data Set: Adult (on Blackboard)

These data sets are described in adult.data, adult.names, adult.test

*Objective*: There are two classes: >50K and <=50K (as determined by the authors- see adult.names)

1. Your job is to determine which of the adults falls into which category by applying clustering, classification, and prediction techniques discussed in class as well as additional functions from the packages mentioned.

2. Using clustering techniques, determine if there are more than just two classes. How many are there?

…………………..

**Source:**

Donor:   
  
Ronny Kohavi and Barry Becker   
Data Mining and Visualization   
Silicon Graphics.

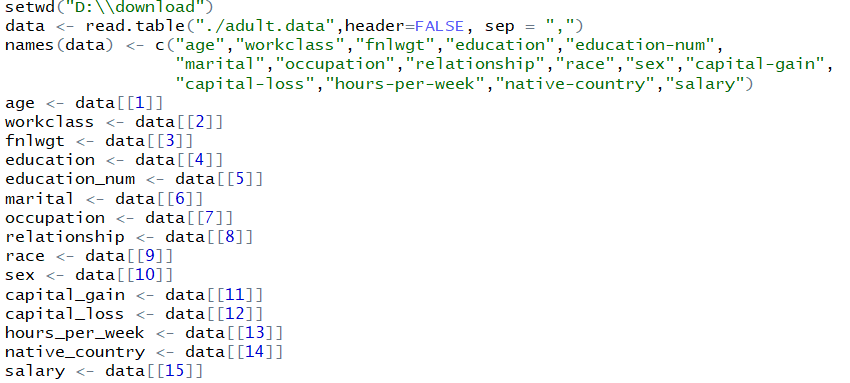
**Data Set Information:**

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))   
  
Prediction task is to determine whether a person makes over 50K a year.

**Attribute Information:**

Listing of attributes:   
  
>50K, <=50K.   
  
age: continuous.   
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.   
fnlwgt: continuous.   
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.   
education-num: continuous.   
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.   
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.   
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.   
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.   
sex: Female, Male.   
capital-gain: continuous.   
capital-loss: continuous.   
hours-per-week: continuous.   
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

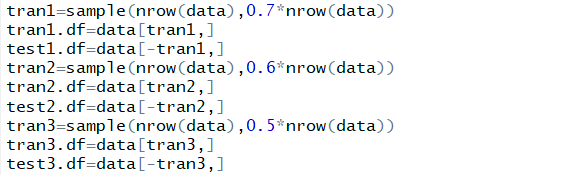
……………



At first, we need to load the adult.data file and name each attribute.

By the way, when you want to execute our source code, please don’t forget to change data direction in your own way.

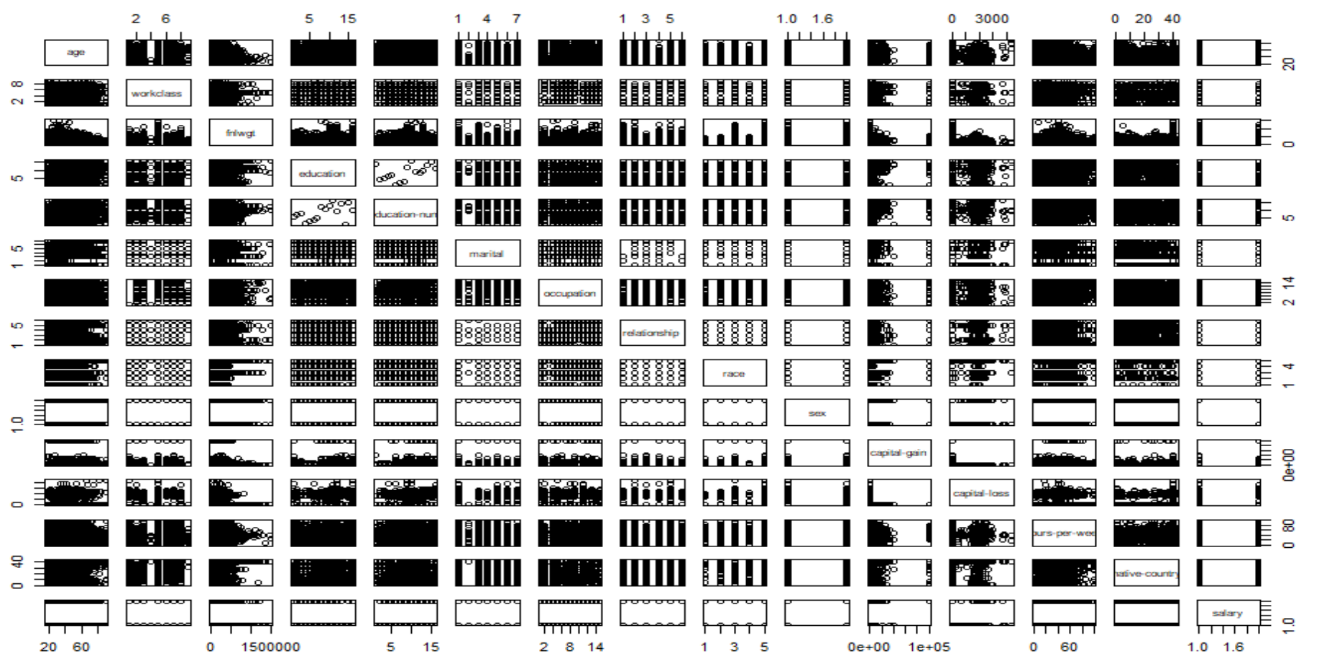
You will need to divide your data set into a training set and a test set. Use samples of 50-50, 60-40, and 70-30 for the training-test ratios. One test set, adult.test, is already provided.



Try plotting the data using several plotting functions to see what it looks like. Use pairs

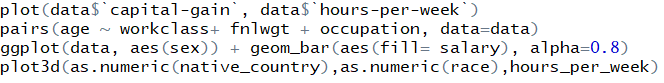
(e.g., 2D plots) or 3 variables (3D plots) based on the packages.

Using pairs(data) function, we can plot how data in one picture.



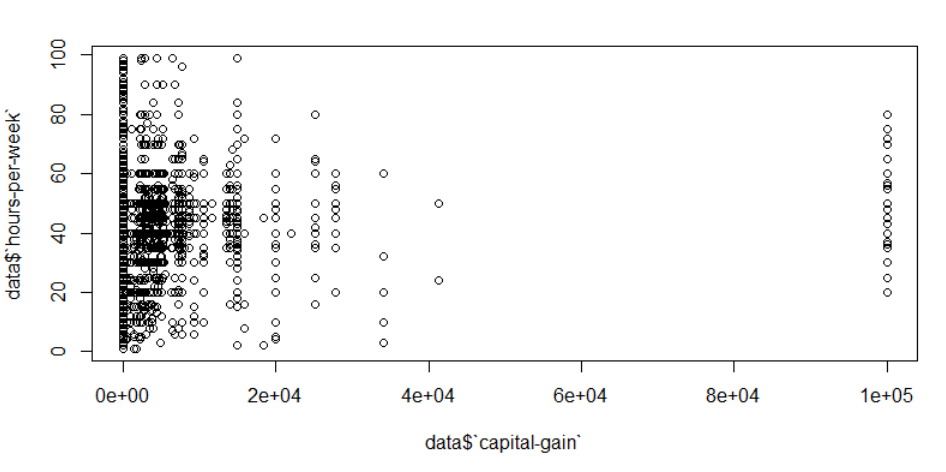
Try to filter the data by selecting samples with only certain attribute values and plotting them.

We can use several function to filter the data.



Firstly, we use traditional plots function

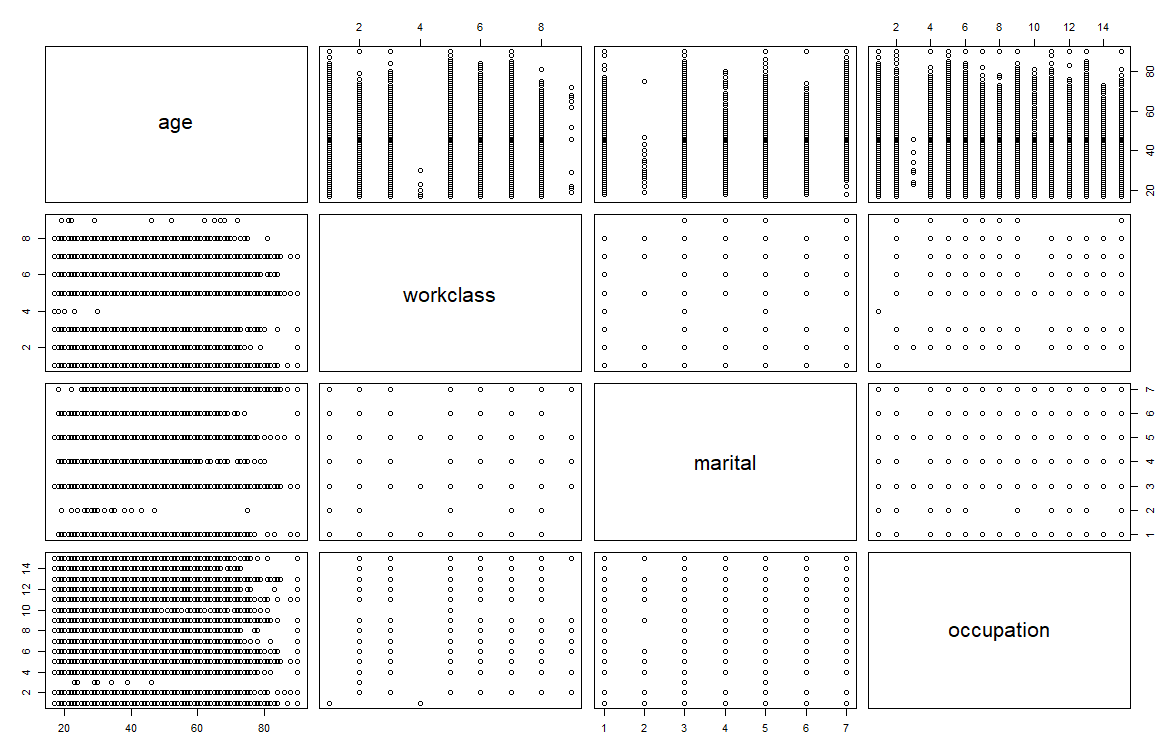
plot (data$`capital-gain`, data$`hours-per-week`)



Let us try pairs function.

pairs(age ~ workclass+ fnlwgt + occupation, data=data)

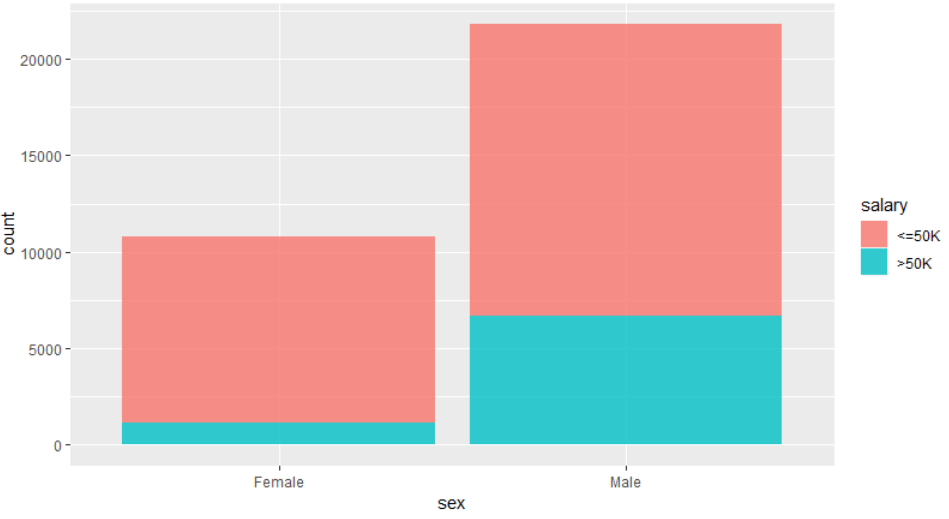
In this function, we can compare 4 specific attributes in data



Or ggplot in package ggplot2

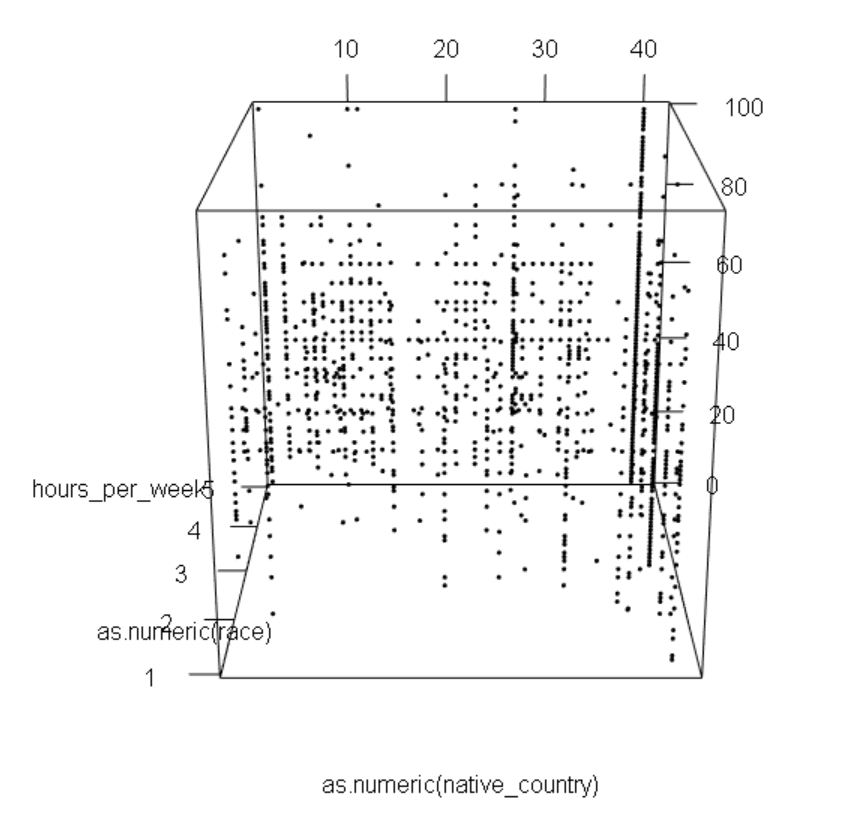
ggplot(data, aes(sex)) + geom\_bar(aes(fill= salary), alpha=0.8)

It can compare two attributes.



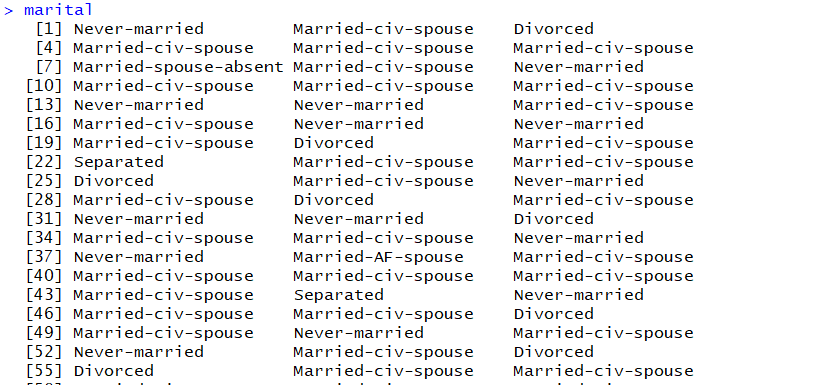
Or plot3d(as.numeric(native\_country),as.numeric(race),hours\_per\_week)

Plot 3 attribute using 3d.

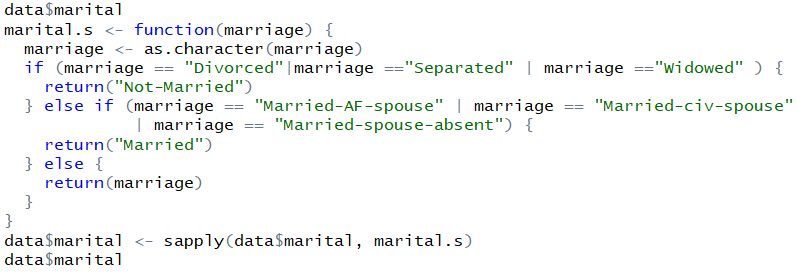


You should try data reduction to eliminate some attributes through Principal Components Analysis. The idea is to try and select N attributes that will help you focus on identifying the unsure samples.

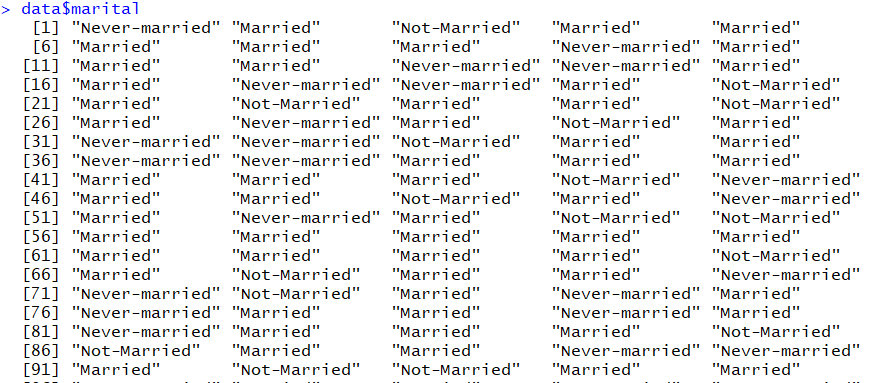
We try to eliminate marital attribute in data. You can find that the original marital contains several status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. We find that they are too much to be display, so we just divide them into two attributes, not-married and married.



Use function to set divorced, separated, widowed as not married and married-af-spouse, married-civ-spouse and married-spouse-absent as married.

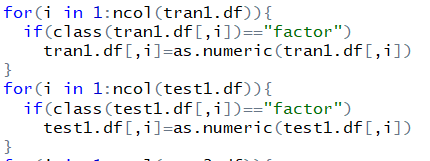


The new marital result shows below.



3.

At first, we need to translate all factor type in 6 datasets into numeric. These are the example to translate tran1.df and test1.df

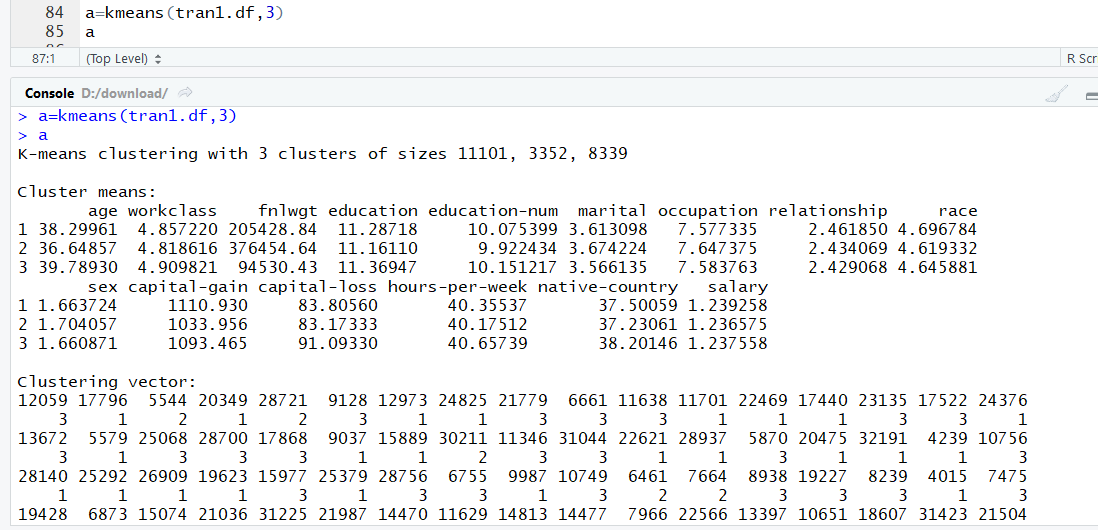


Then we use three different clustering methods, k-means, Cluster Analysis and K nearest neighbor function. For N=3,5,7.

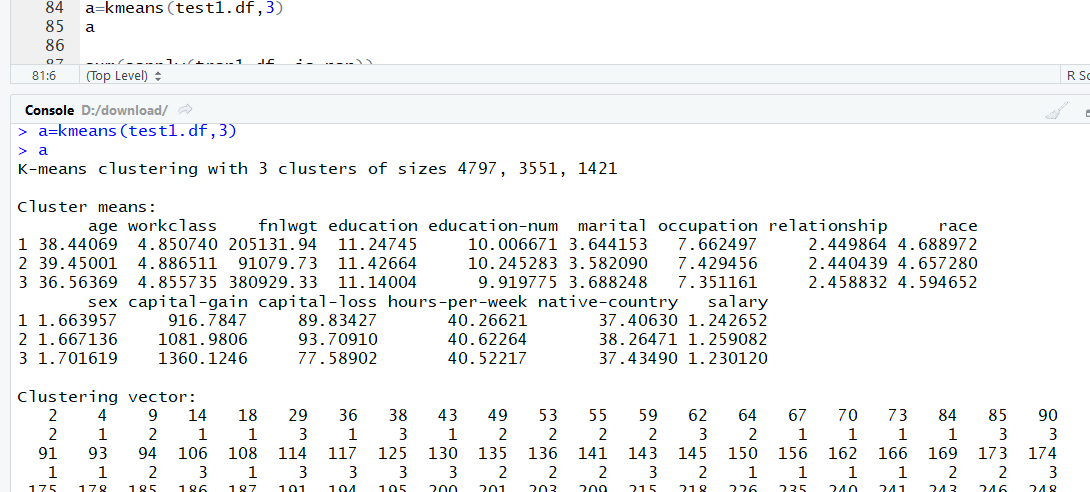
3.a

Now, we will perform k-means:

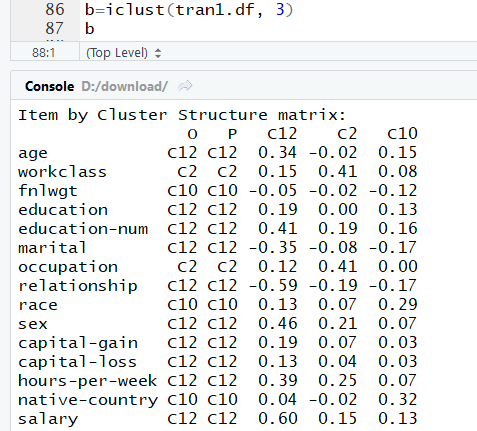
Kmeans with number of clusters (N)=3 on training data set with 70% of the values:



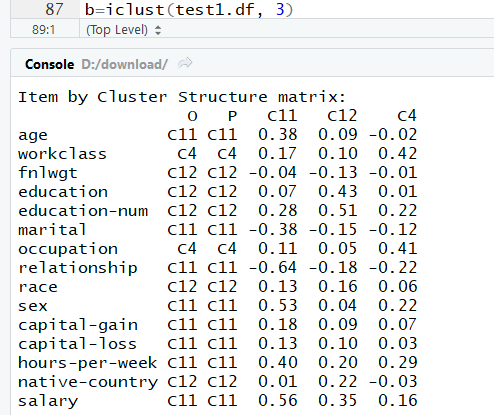
Kmeans with number of clusters (N)=3 on test data set with 30% of the values:



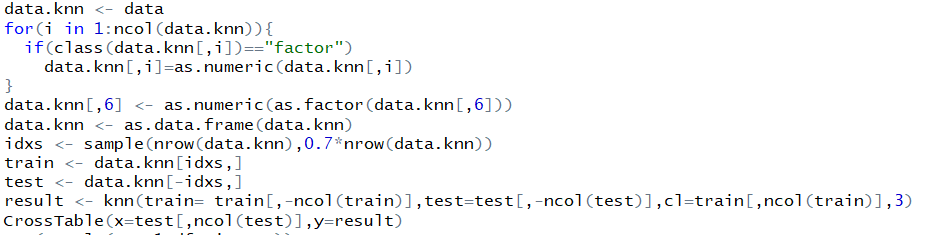
Cluster Analysis with number of clusters (N)=3 on training data set with 70% of the values:

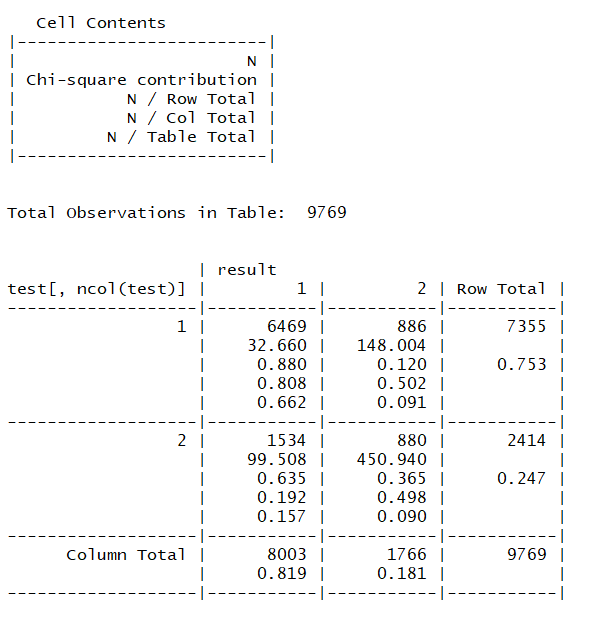


Cluster Analysis with number of clusters (N)=3 on test data set with 30% of the values:



K nearest neighbor on training data set with 70% of the values and test data set with 30% of the values, with N=3

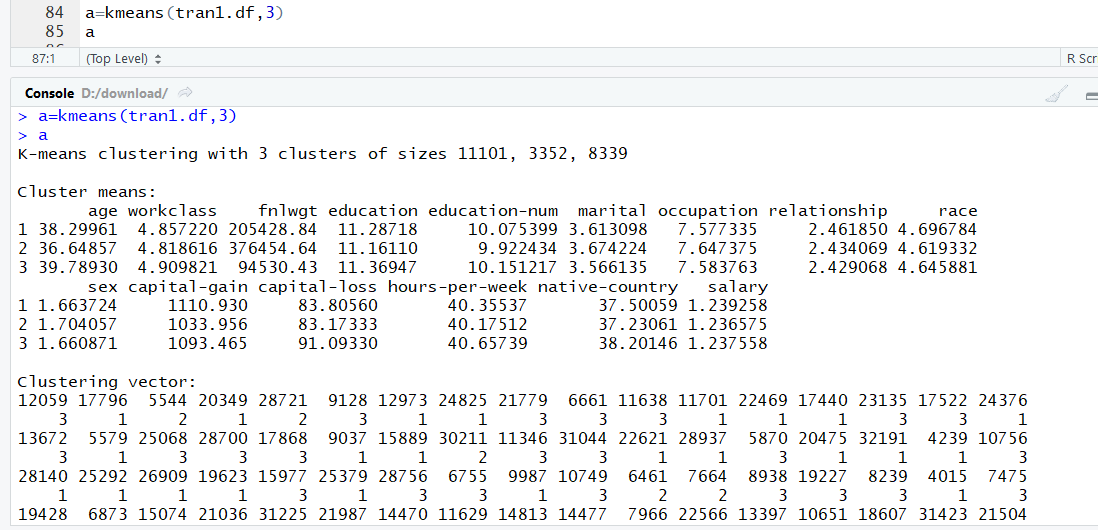




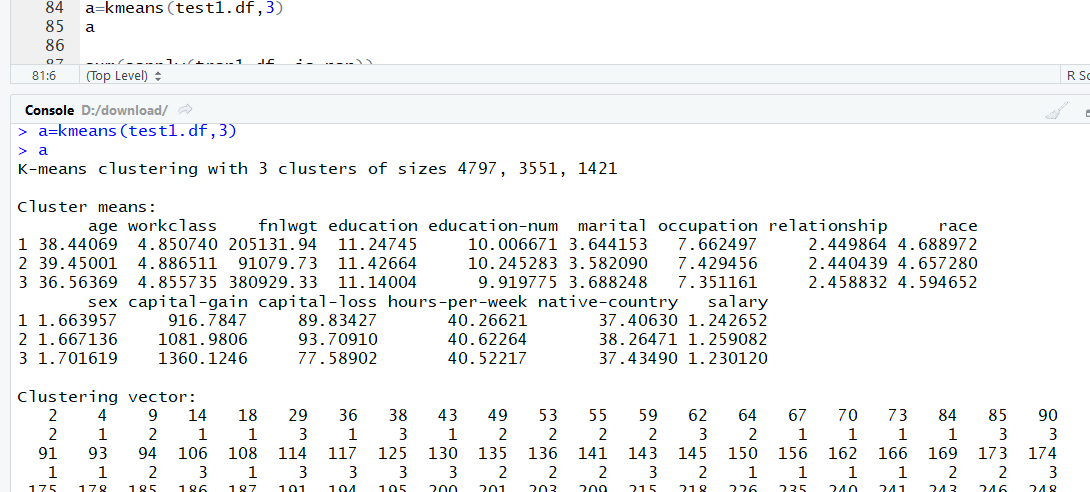
3.b We choose 70-30 for the training -test ratio

K-means:

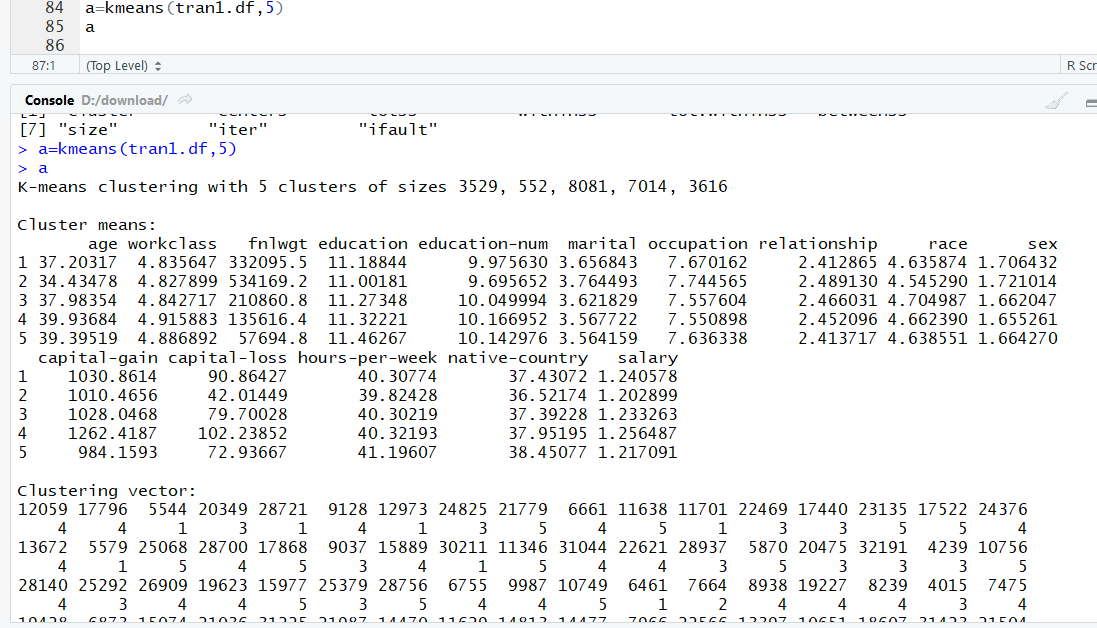
Kmeans with number of clusters (N)=3 on training data set with 70% of the values:



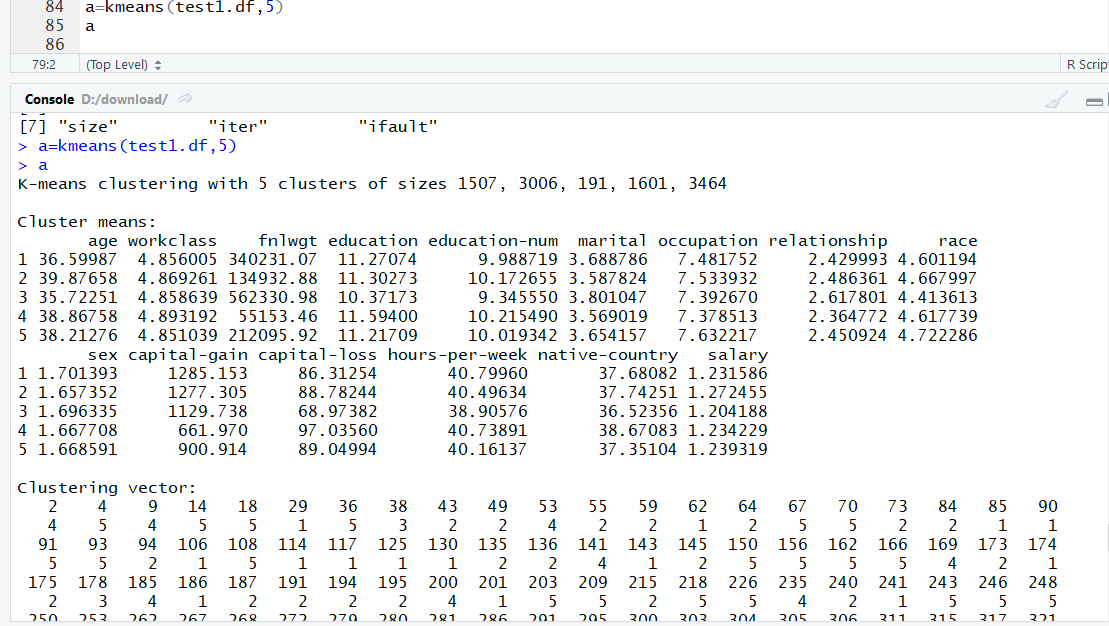
Kmeans with number of clusters (N)=3 on test data set with 30% of the values:



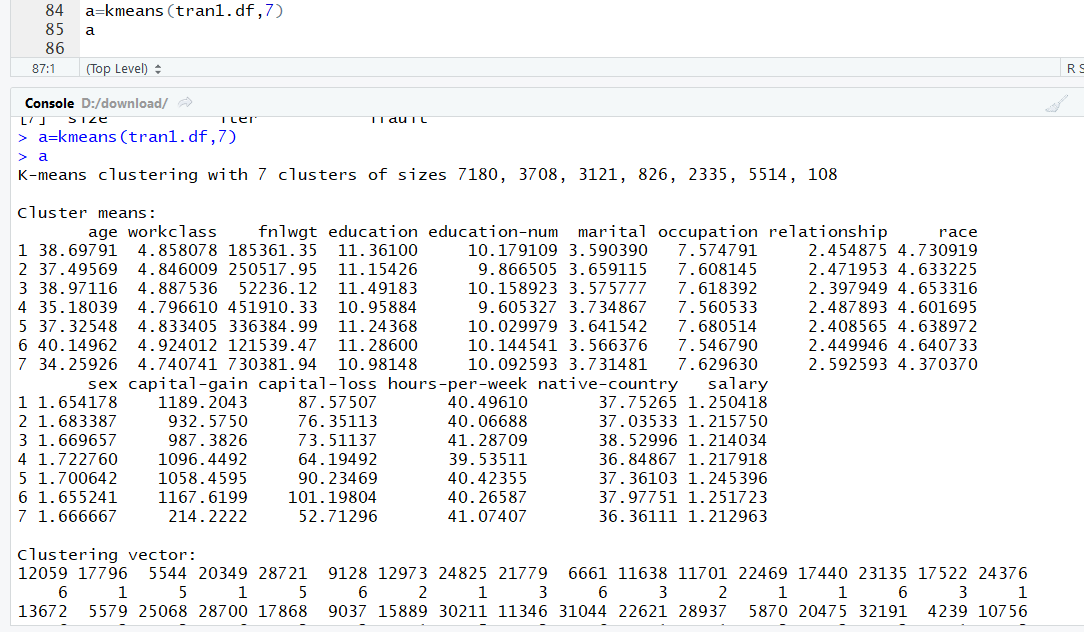
Kmeans with number of clusters (N)=5 on training data set with 70% of the values:



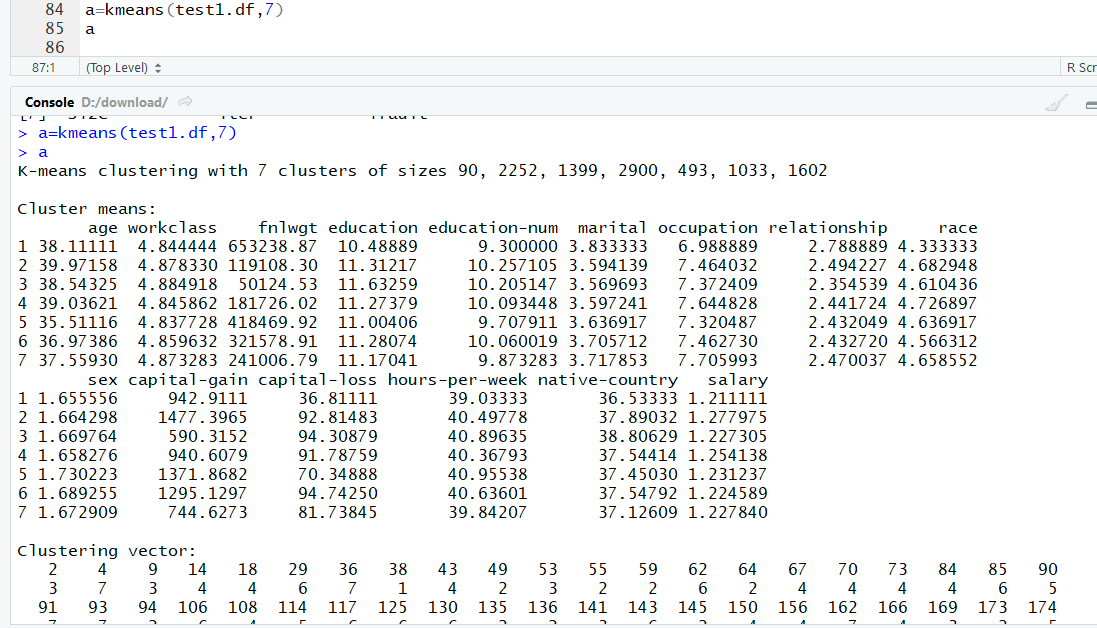
Kmeans with number of clusters (N)=5 on test data set with 30% of the values:



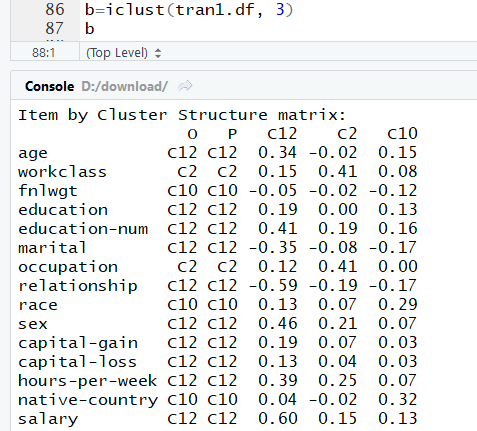
Kmeans with number of clusters (N)=7 on training data set with 70% of the values:



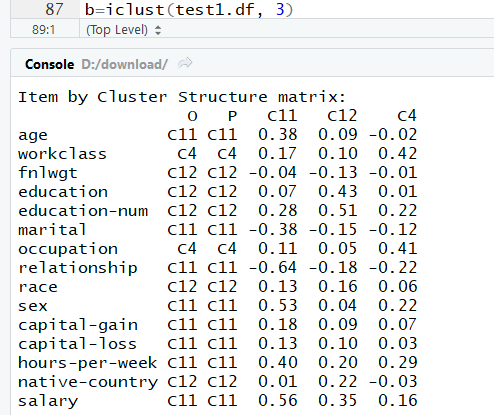
Kmeans with number of clusters (N)=7 on test data set with 30% of the values:



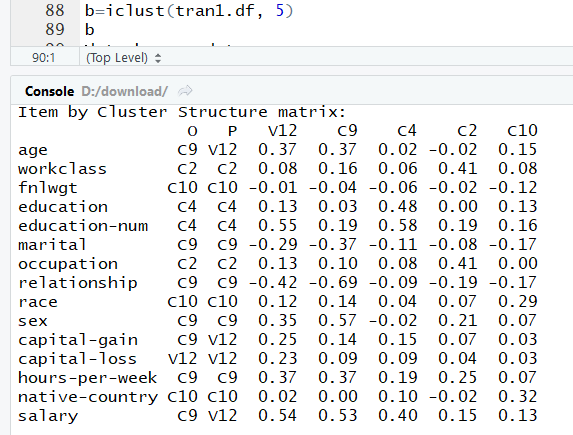
Cluster Analysis with number of clusters (N)=3 on training data set with 70% of the values:



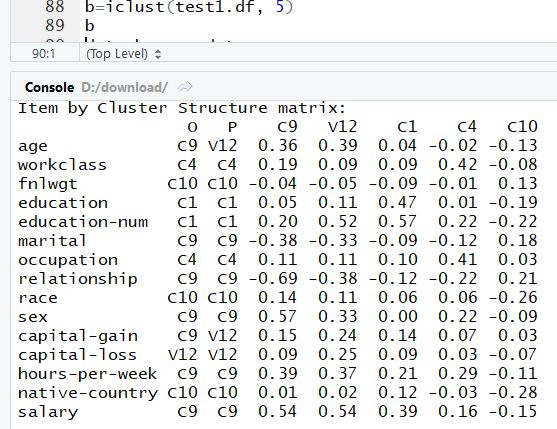
Cluster Analysis with number of clusters (N)=3 on test data set with 30% of the values:



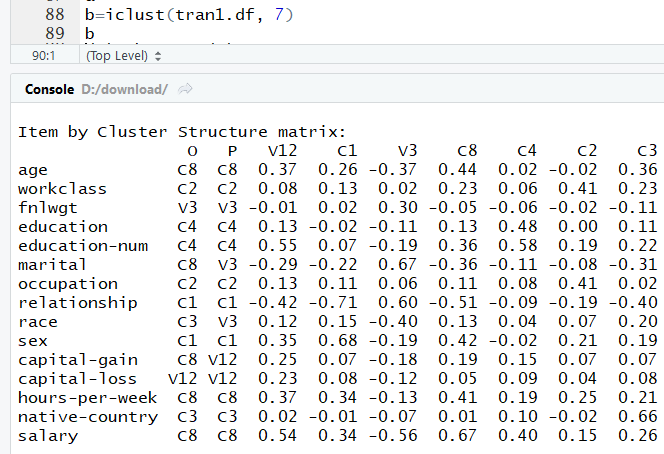
Cluster Analysis with number of clusters (N)=5 on training data set with 70% of the values:



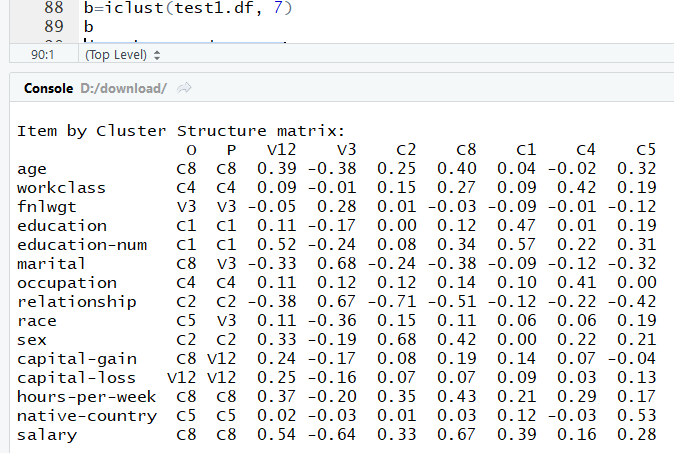
Cluster Analysis with number of clusters (N)=5 on test data set with 30% of the values:



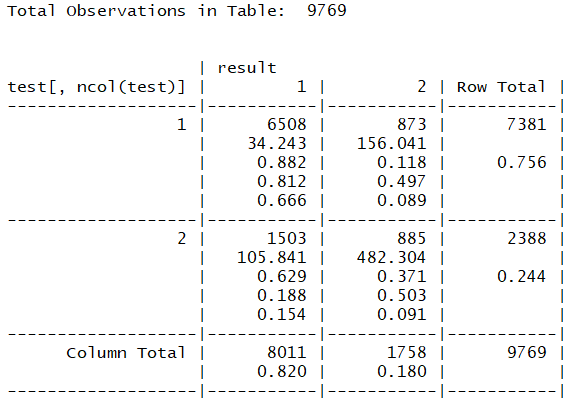
Cluster Analysis with number of clusters (N)=7 on training data set with 70% of the values:



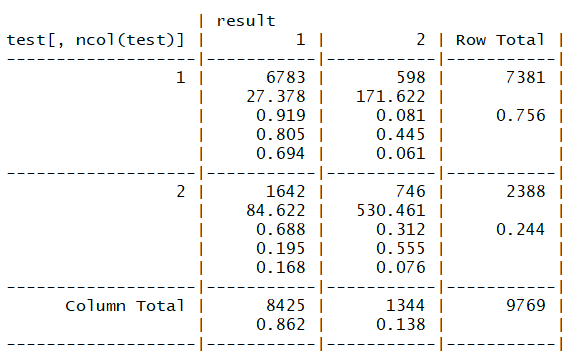
Cluster Analysis with number of clusters (N)=7 on test data set with 30% of the values:



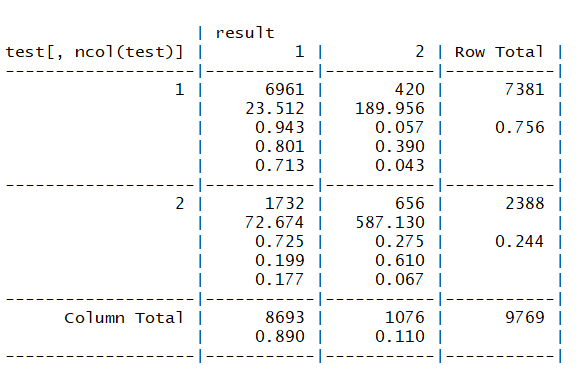
K nearest neighbor on training data set with 70% of the values and test data set with 30% of the values, with N=3



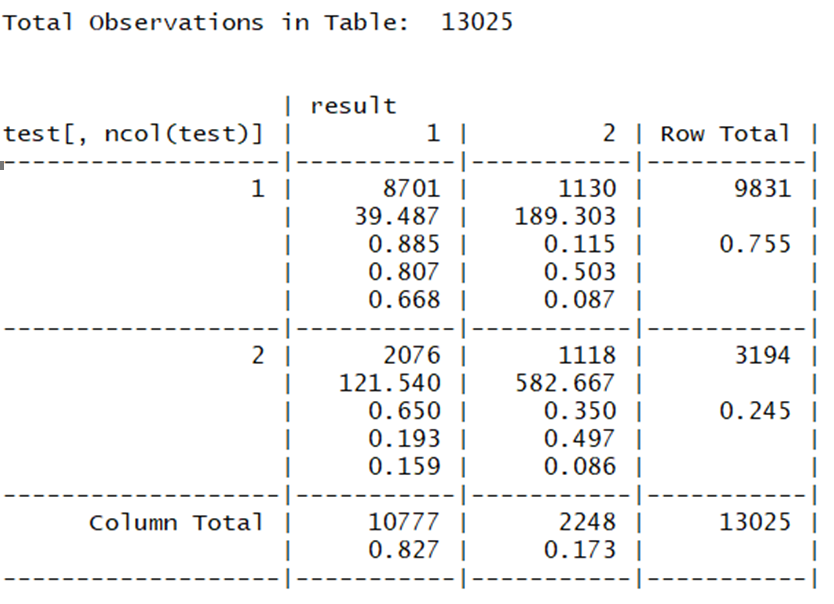
K nearest neighbor on training data set with 70% of the values and test data set with 30% of the values, with N=5



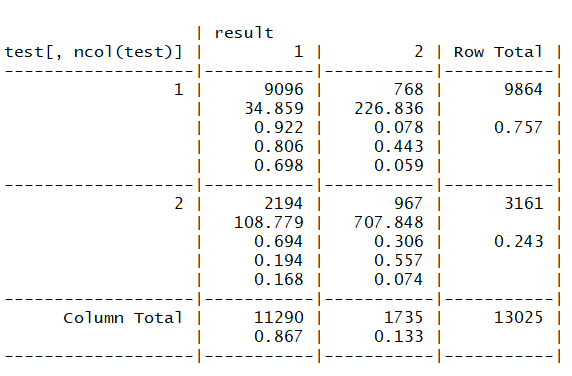
K nearest neighbor on training data set with 70% of the values and test data set with 30% of the values, with N=7



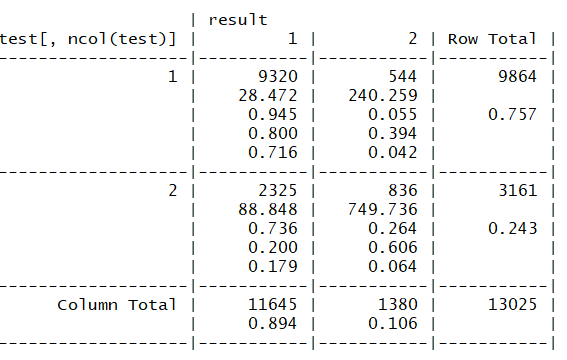
K nearest neighbor on training data set with 60% of the values and test data set with 40% of the values, with N=3



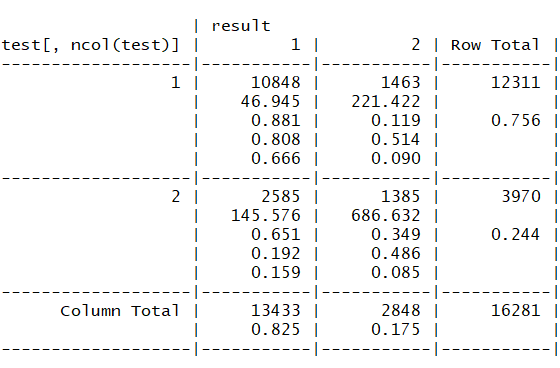
K nearest neighbor on training data set with 60% of the values and test data set with 40% of the values, with N=5



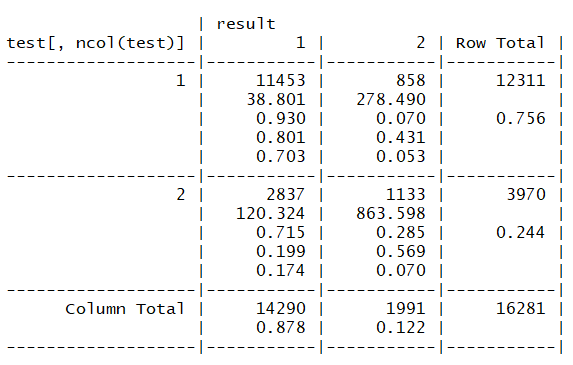
K nearest neighbor on training data set with 60% of the values and test data set with 40% of the values, with N=7



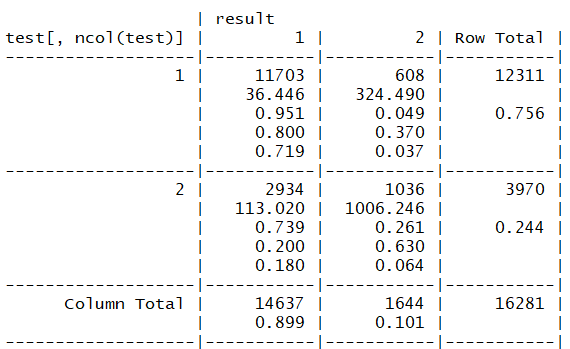
K nearest neighbor on training data set with 50% of the values and test data set with 50% of the values, with N=3



K nearest neighbor on training data set with 50% of the values and test data set with 50% of the values, with N=5



K nearest neighbor on training data set with 50% of the values and test data set with 50% of the values, with N=7



3.c Table (answer of 6.b)

Table matrix is as follows:

Kmeans accuracy

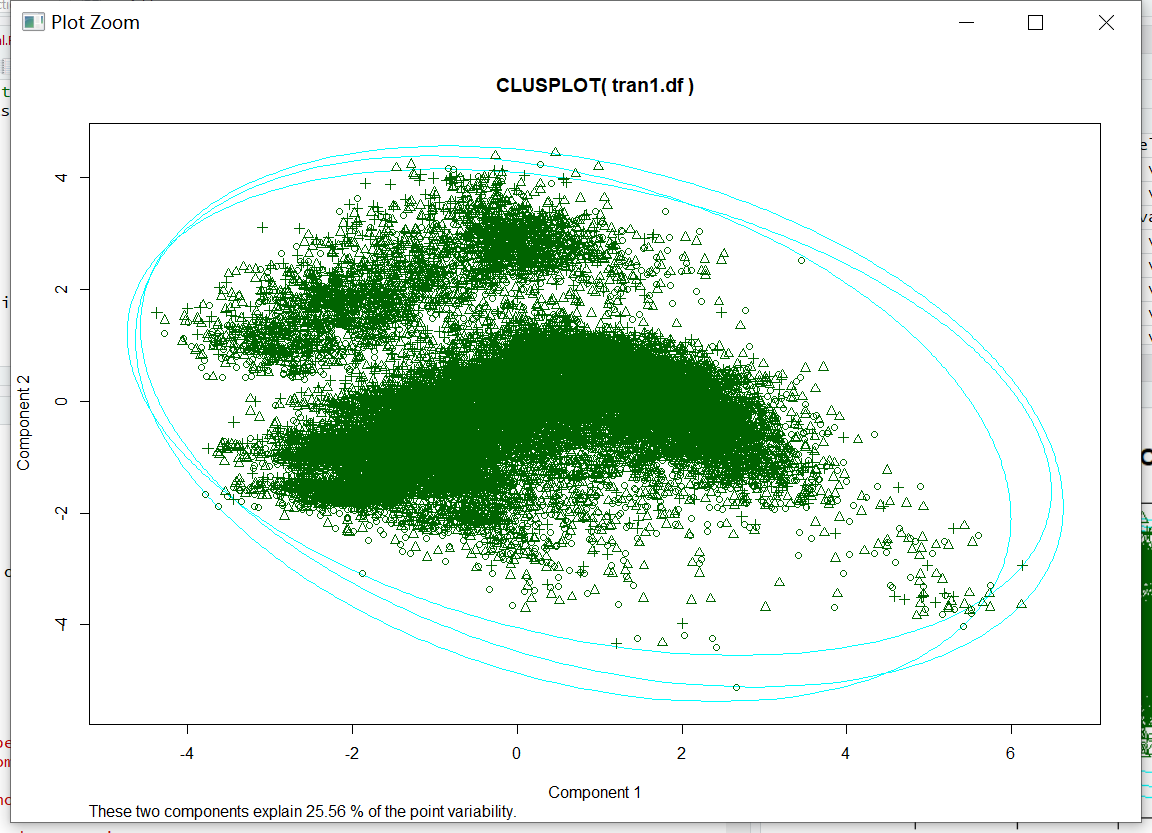
|  |  |  |  |
| --- | --- | --- | --- |
|  | 3 | 5 | 7 |
| 70-30 | 77.5% | 89.0% | 92.4% |
| 60-40 | 77.2% | 89.0% | 93.9% |
| 50-50 | 78.0% | 89.3% | 94.1% |

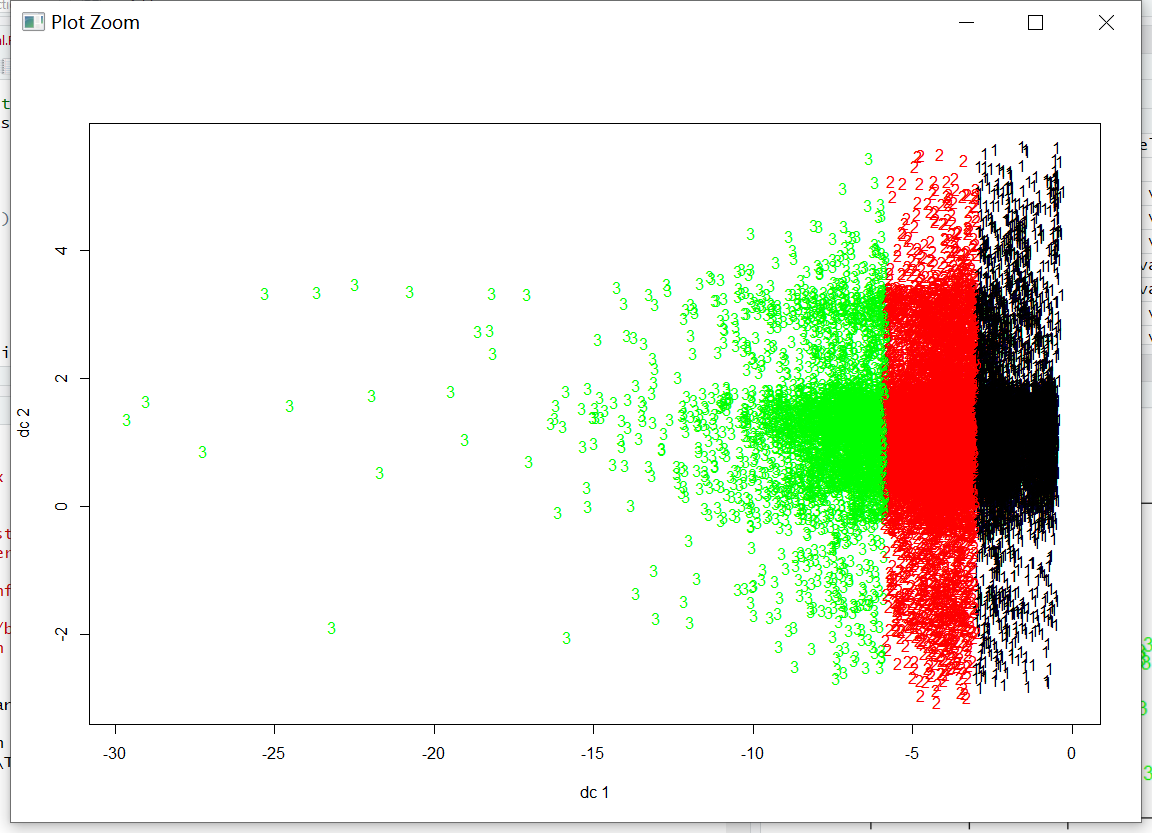
Knn miss rate

|  |  |  |  |
| --- | --- | --- | --- |
|  | 3 | 5 | 7 |
| 70-30 | 0.2503839 | 0.2297062 | 0.2204934 |
| 60-40 | 0.2459885 | 0.2284837 | 0.2148944 |
| 50-50 | 0.243474 | 0.2259689 | 0.2174314 |

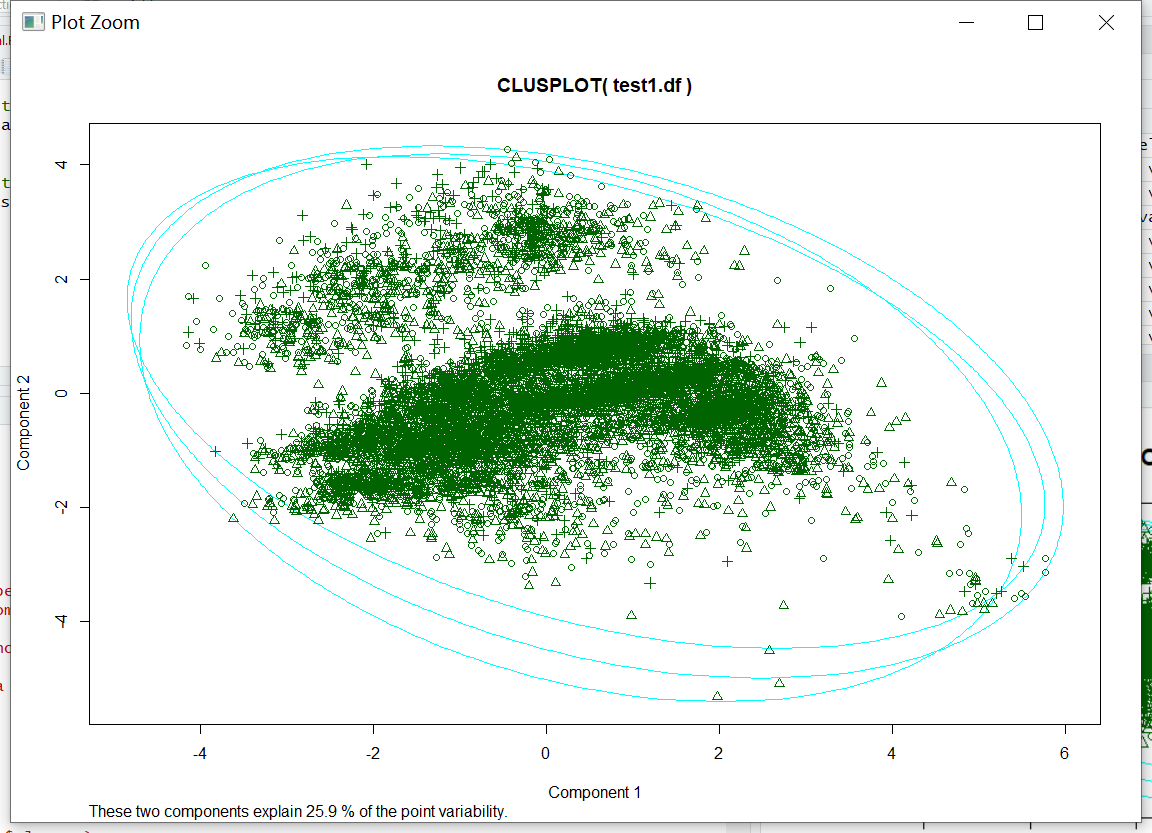
3.d Plot the result

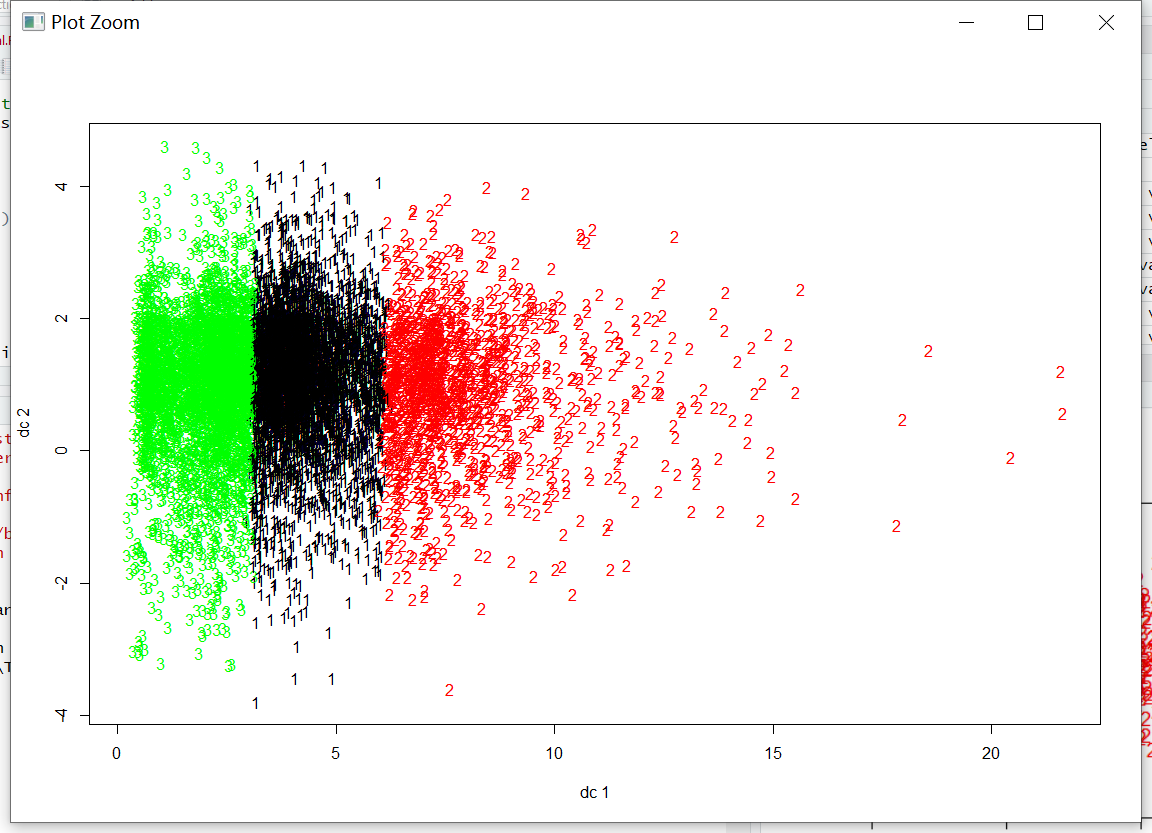
Kmeans with number of clusters (N)=3 on training data set with 70% of the values:



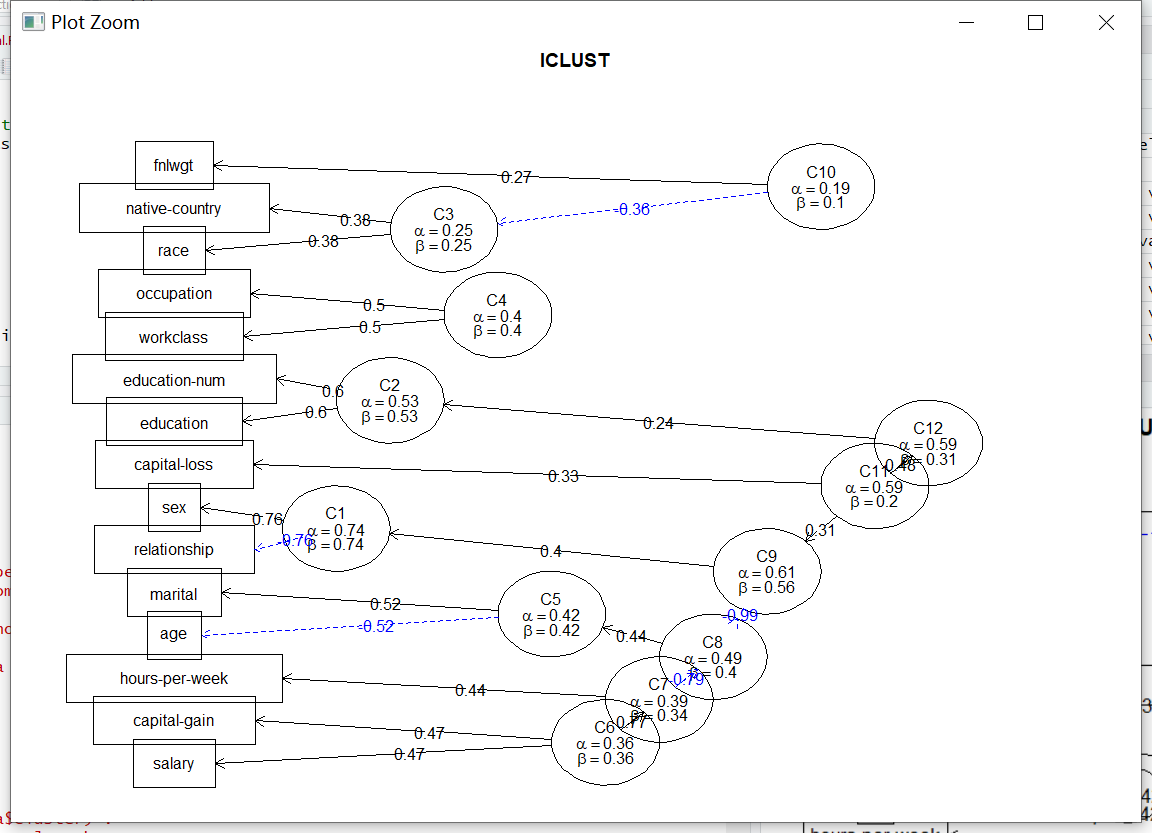


Kmeans with number of clusters (N)=3 on test data set with 30% of the values:

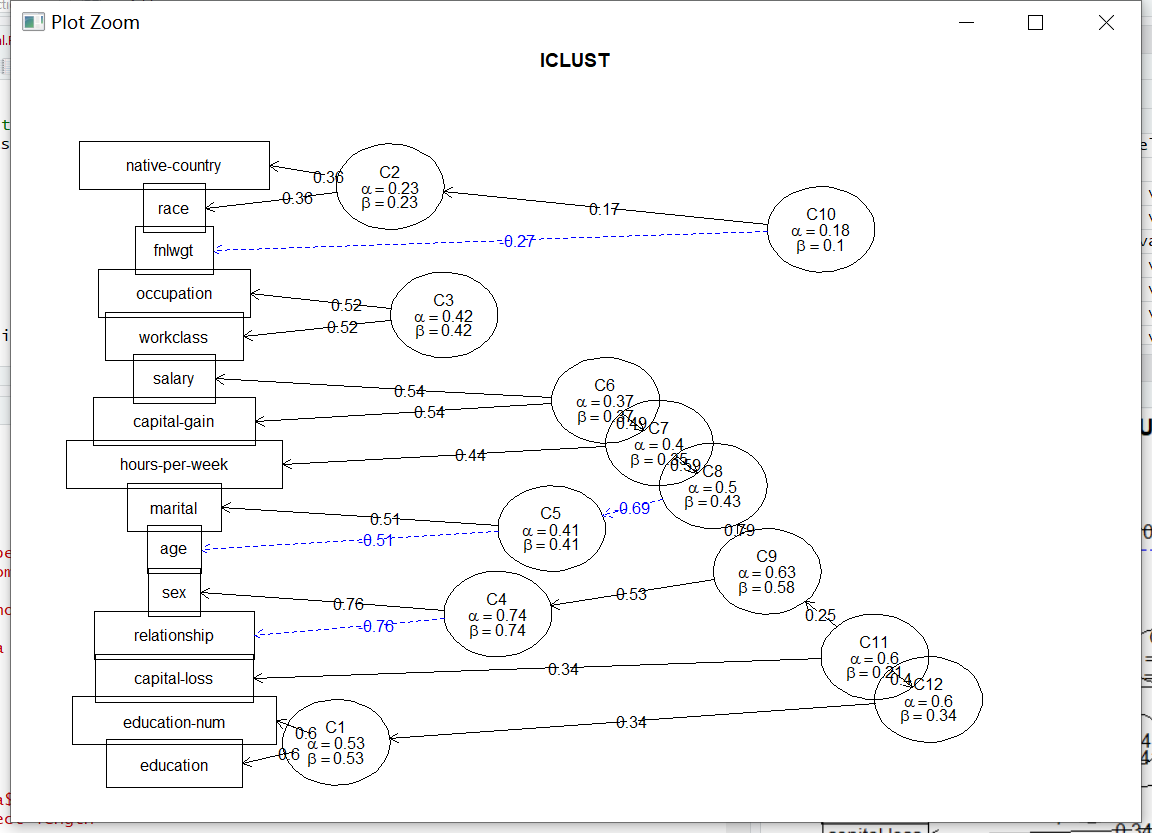




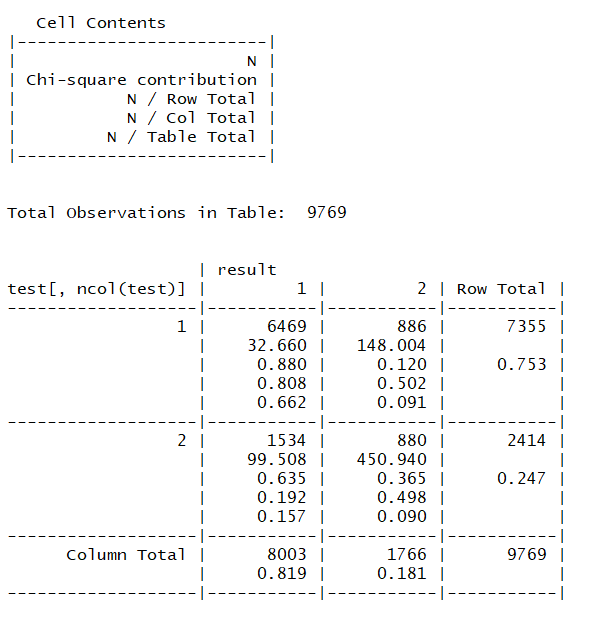
Cluster Analysis with number of clusters (N)=3 on training data set with 70% of the values:



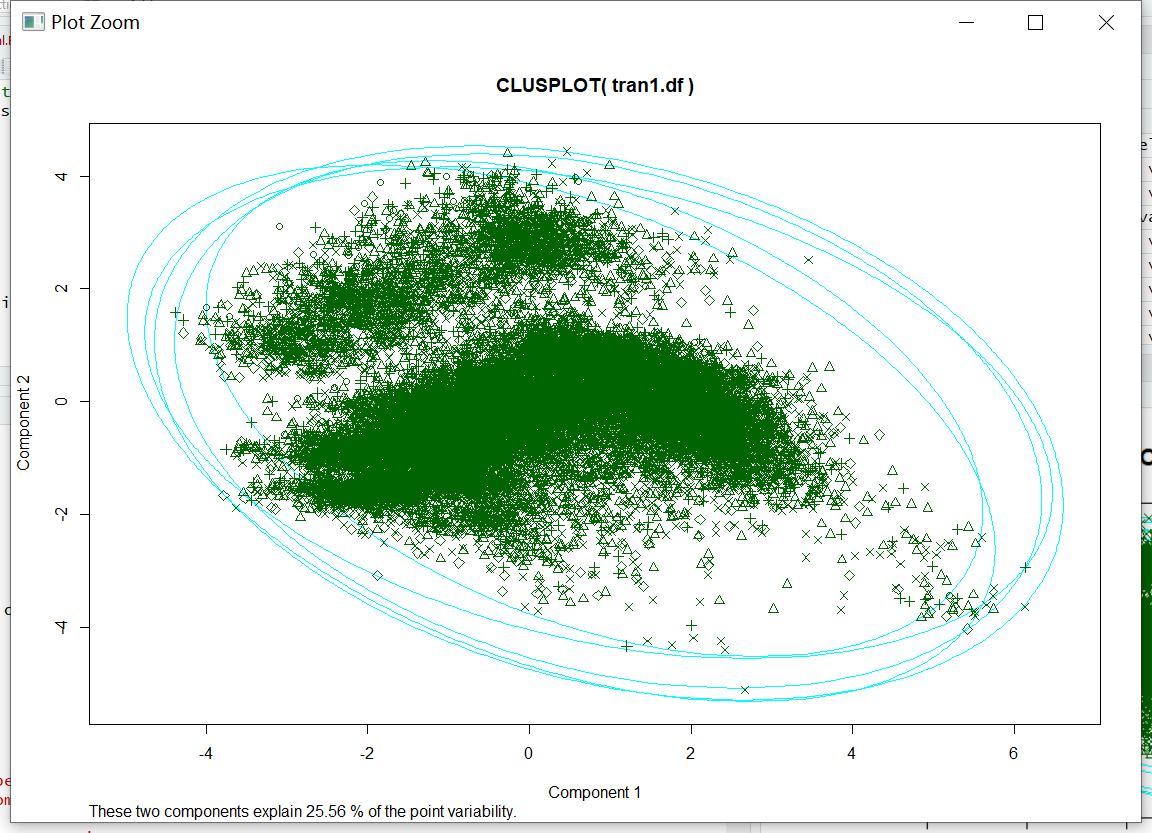
Cluster Analysis with number of clusters (N)=3 on test data set with 30% of the values:

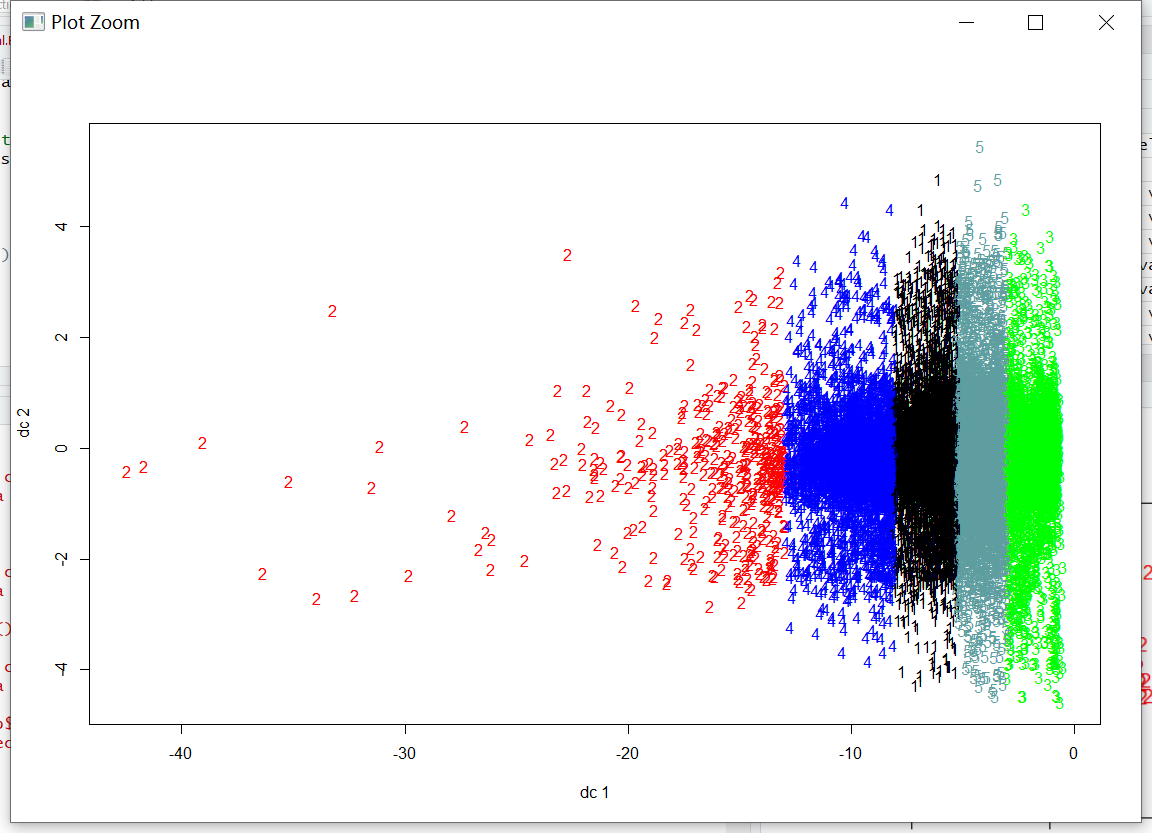


K nearest neighbor on training data set with 70% of the values and test data set with 30% of the values, with N=3

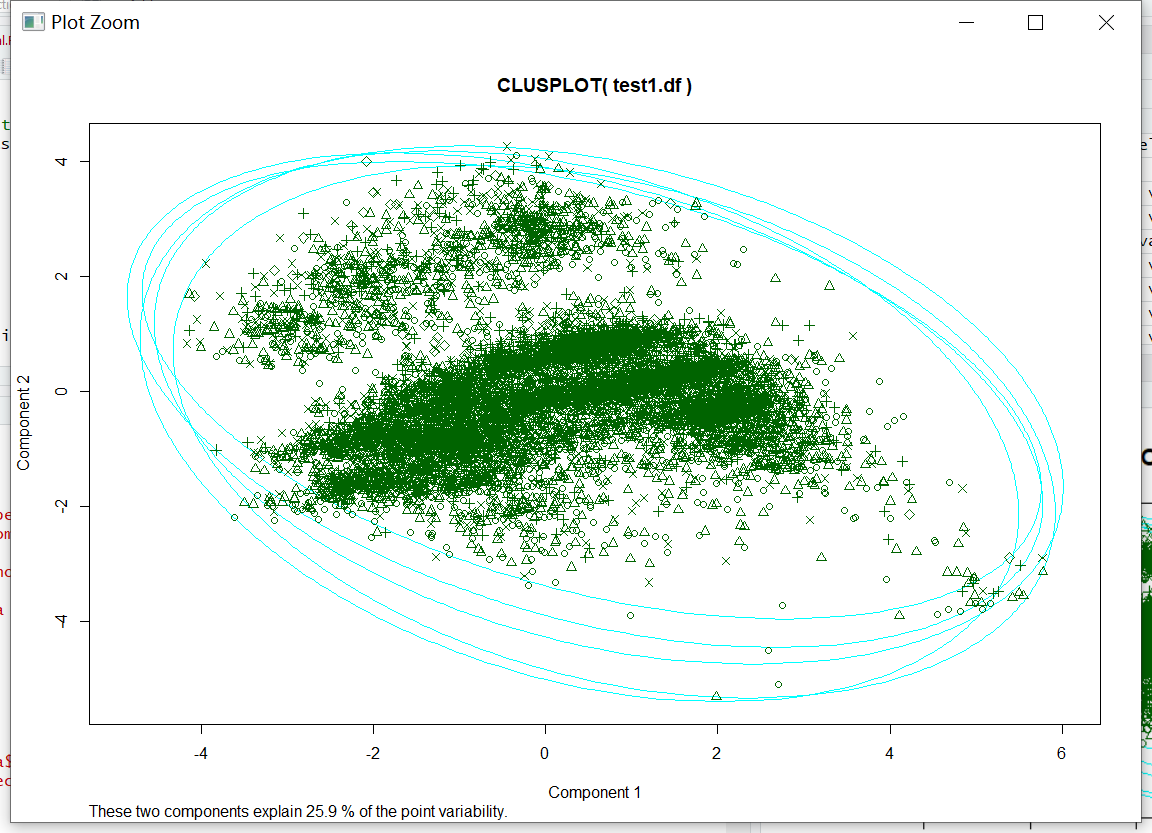


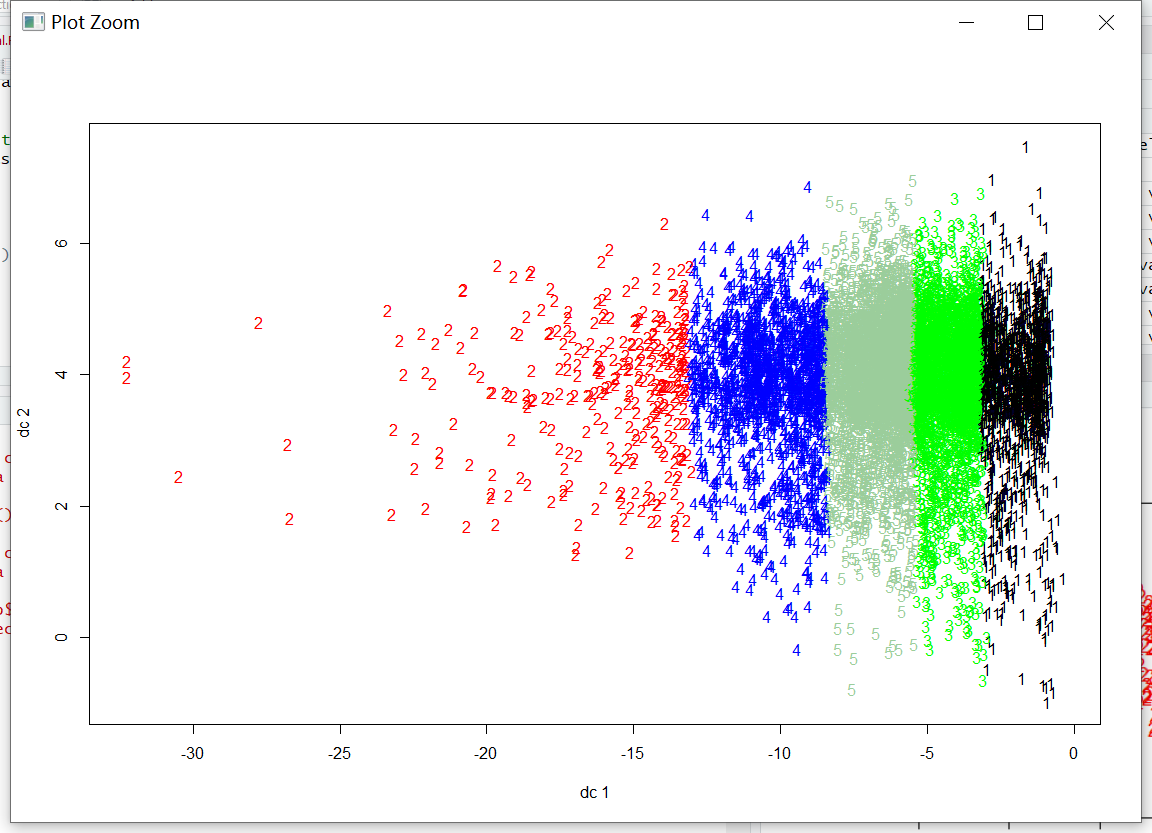
Kmeans with number of clusters (N)=5 on training data set with 70% of the values:



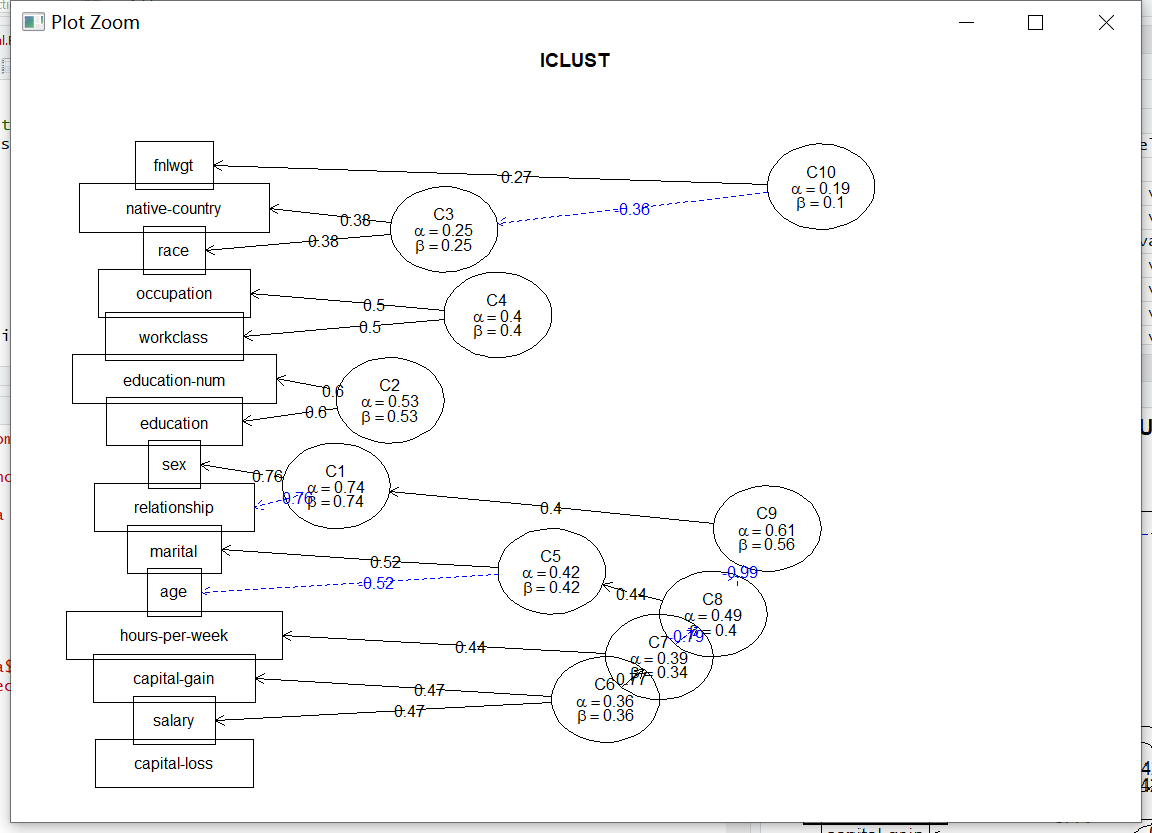


Kmeans with number of clusters (N)=5 on test data set with 30% of the values:

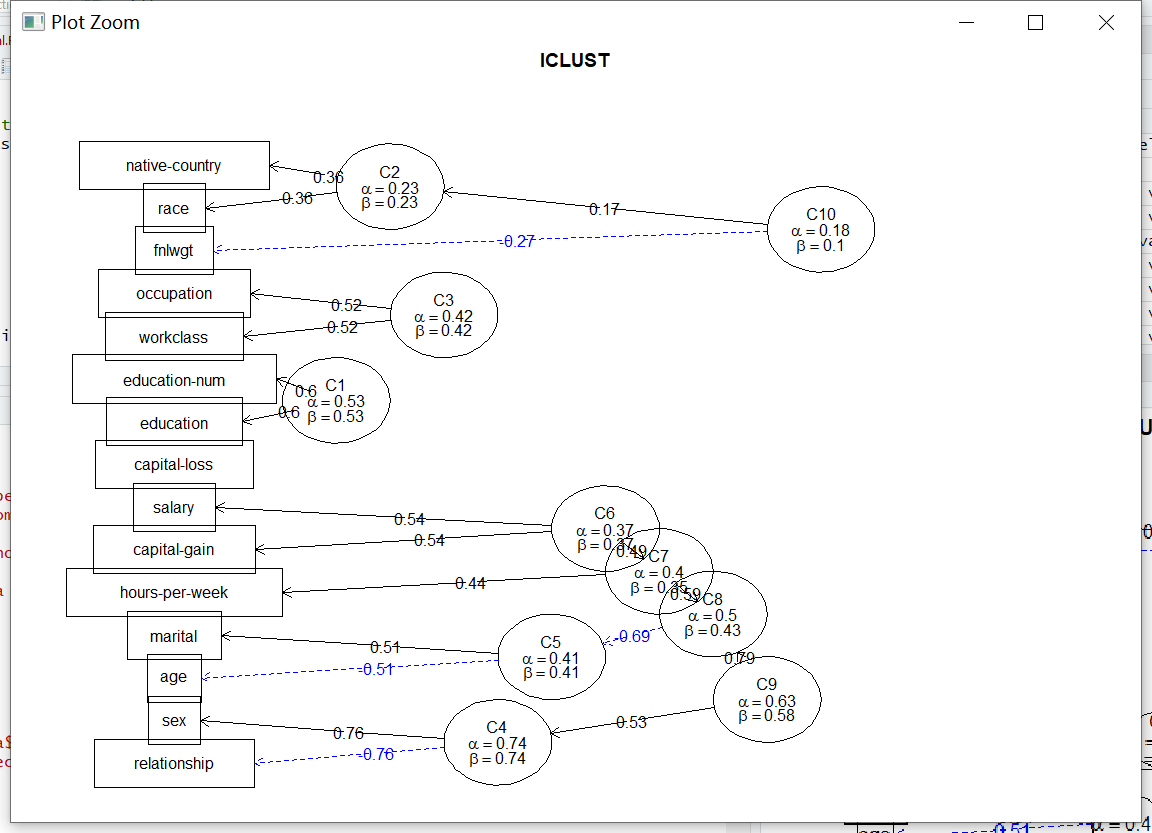




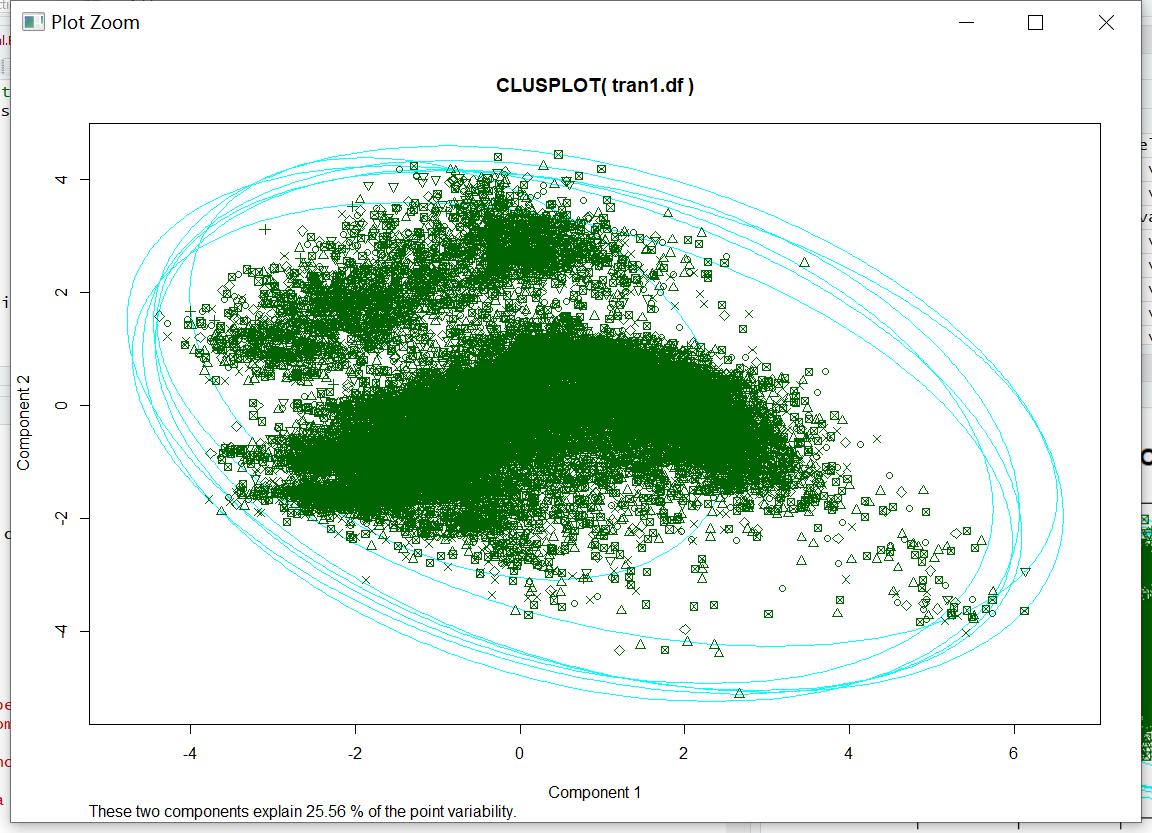
Cluster Analysis with number of clusters (N)=5 on training data set with 70% of the values:

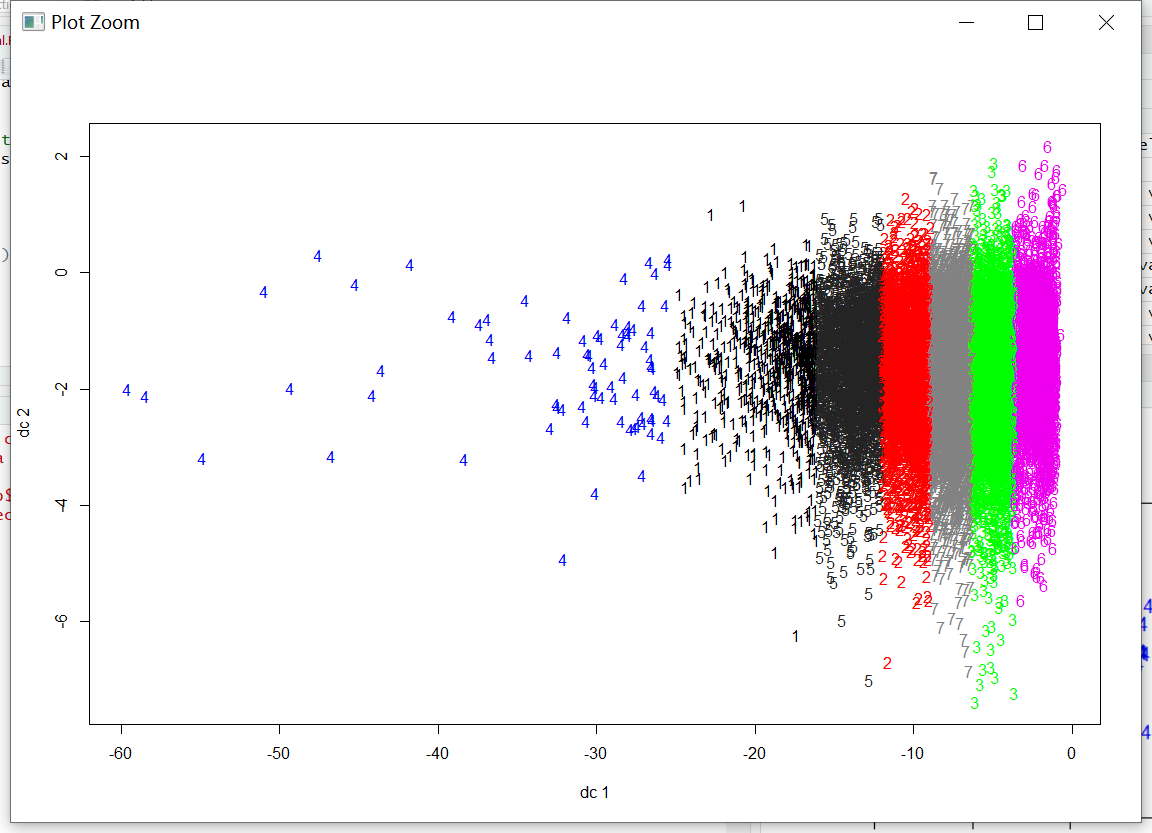


Cluster Analysis with number of clusters (N)=5 on test data set with 30% of the values:

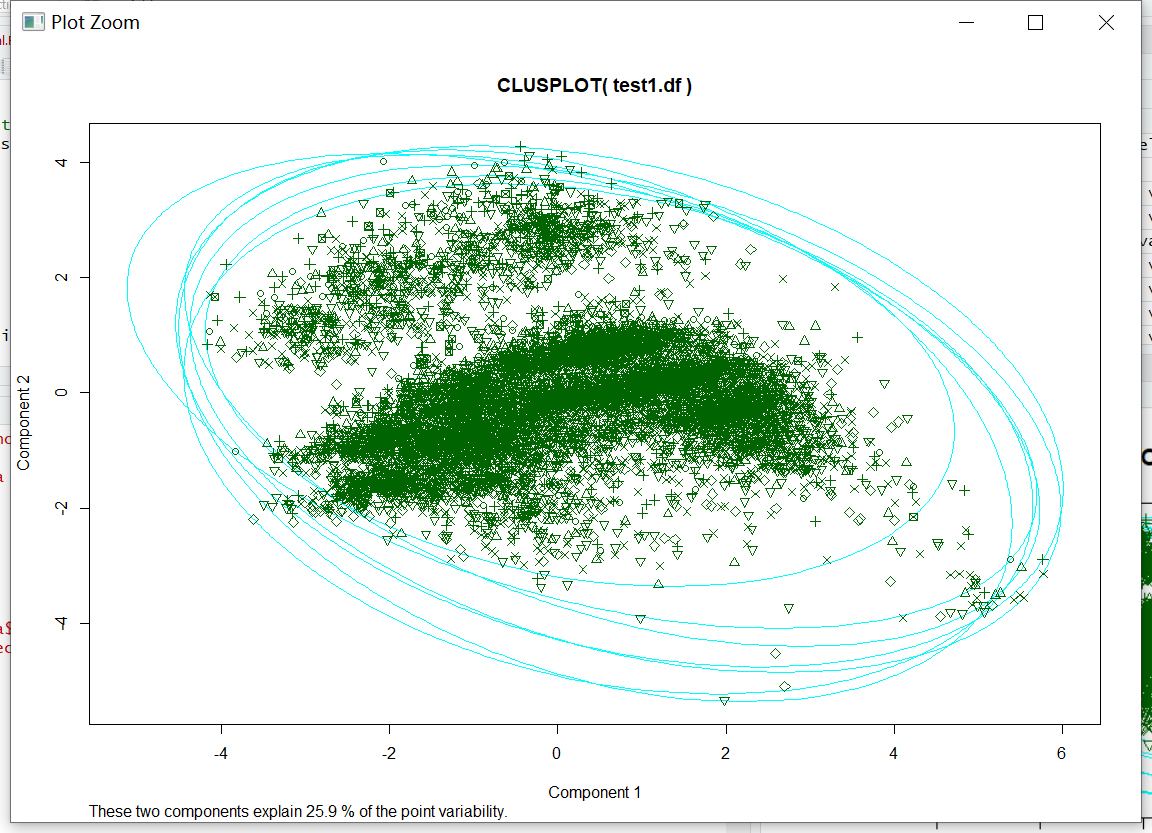


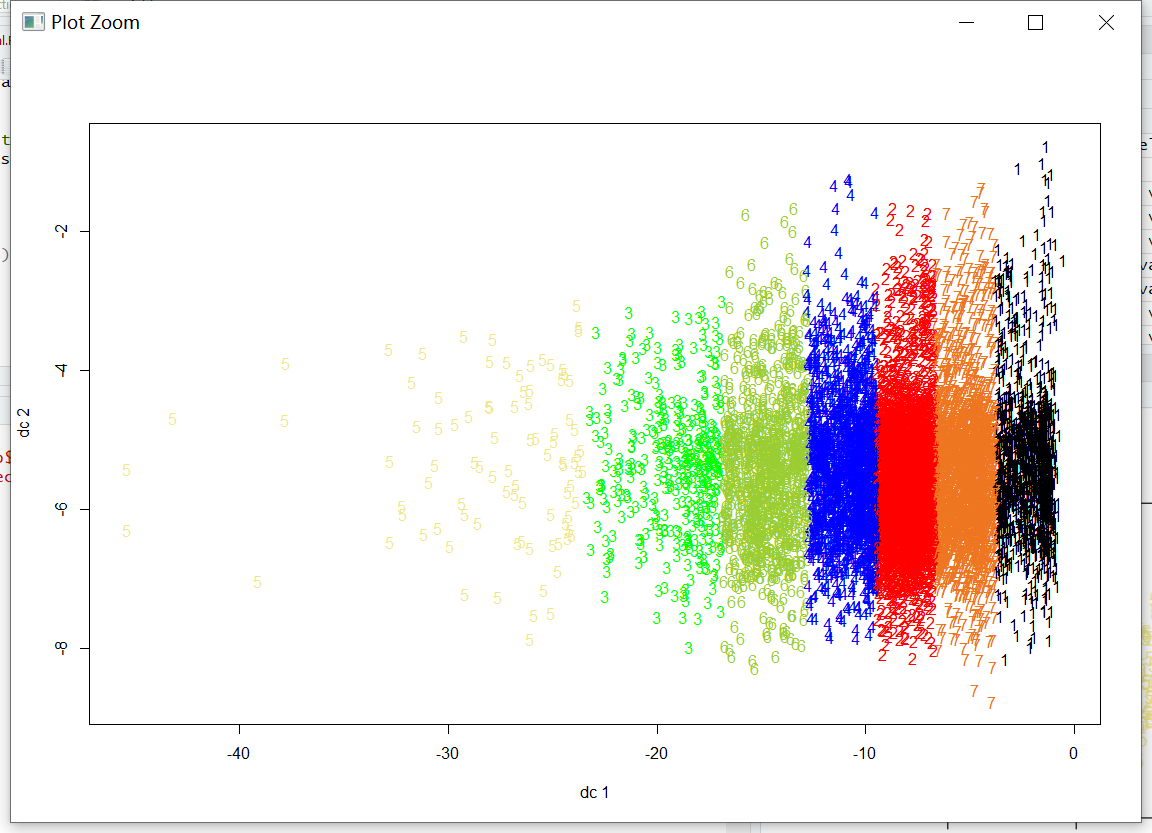
Kmeans with number of clusters (N)=7 on training data set with 70% of the values:



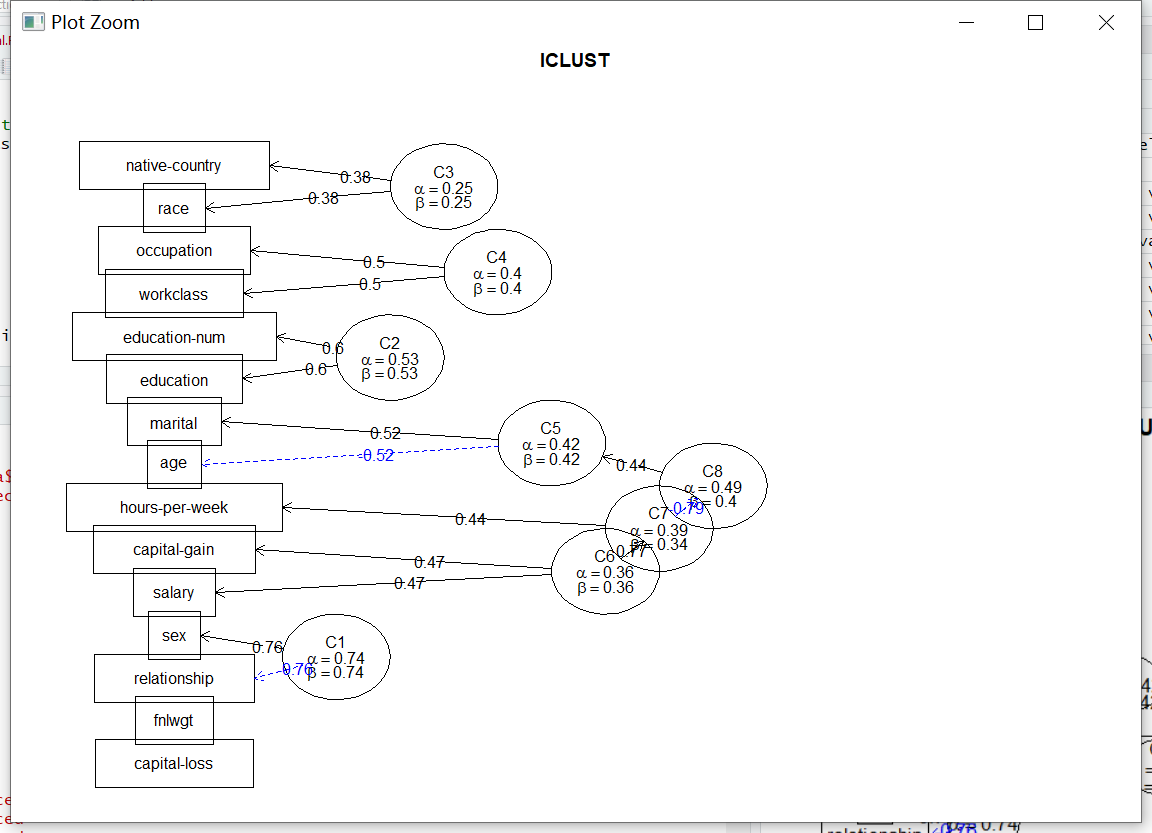


Kmeans with number of clusters (N)=7 on test data set with 30% of the values:

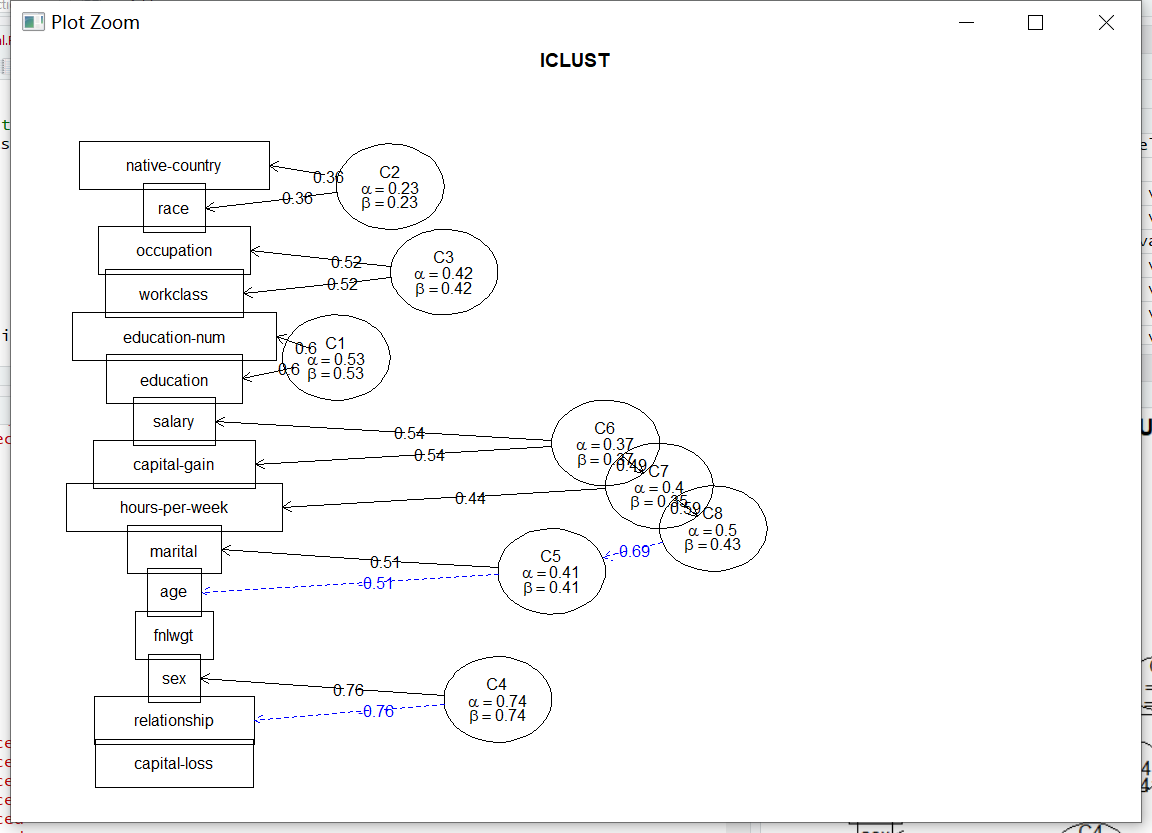




Cluster Analysis with number of clusters (N)=7 on training data set with 70% of the values:



Cluster Analysis with number of clusters (N)=7 on test data set with 30% of the values:



Answer of question 6.c:

Which method is better?

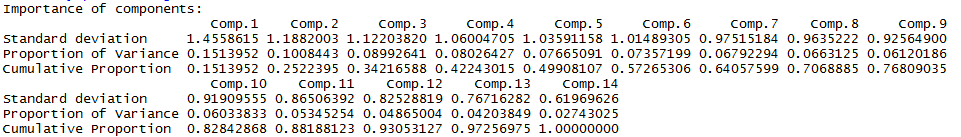
From the plots we get, we find that K-means is a better function, the character of different clustering shown in different style of plot function result, which let us understand the data in a great degree.

What seems to be the best number of clusters for each method?

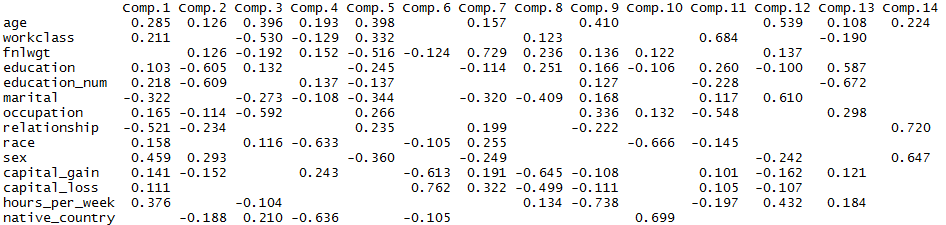
For all methods, it seems that as the number of clusters increase, the accuracy of knn and k-means increase at the same time. And the result of icluster more clearly to shows the feature of the data. So the best number of cluster in this assignment is N=7

4.

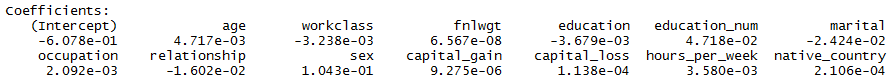
Before fit the data into lm or glm, we made some preprocessing. All features not numeric are turned into numeric. Firstly, we need to know what features are related, what are not. So, we did PCA analysis of all features:



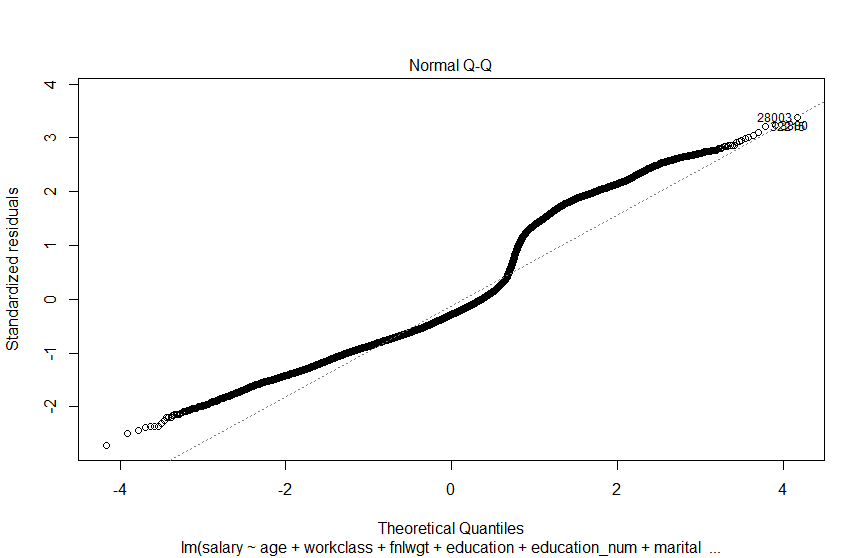
From this figure we can know that 13 features have got 97% cummulative proportion. Therefore, we can omit the 14th feature.



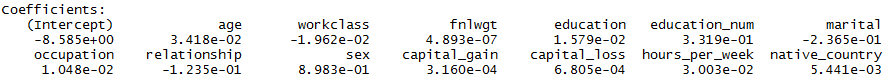
From this figure we can know that the most related to Comp.14 is feature “race”. Therefore, we don’t have to consider “race” feature in this question. Then, we can do lm for 13 features.



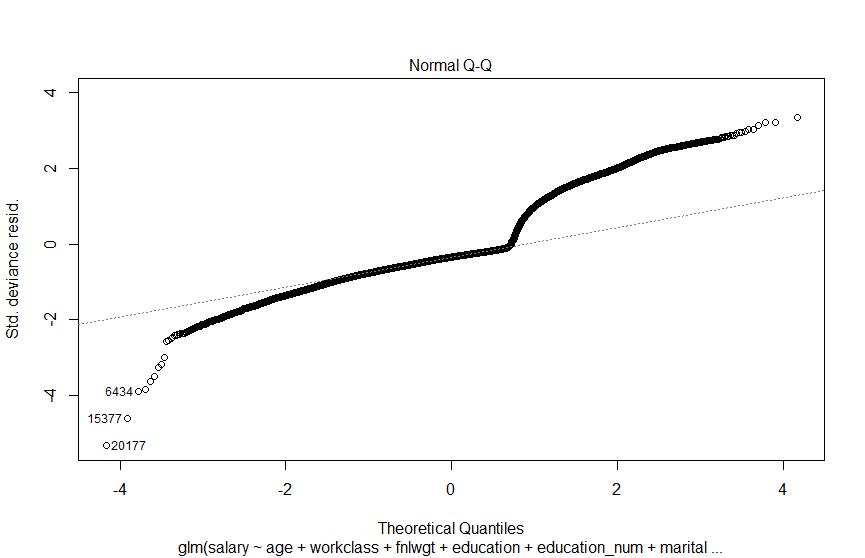
By this regression equation, the output should be around [1, 2]. 1 means “<=50K” and 2 means “>50K”. The regression visualization result Q-Q plot is shown as follows:



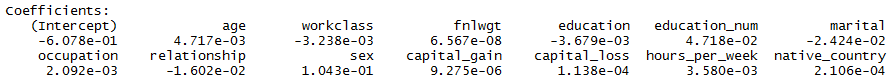
For glm, binomimal prior probability assumption:



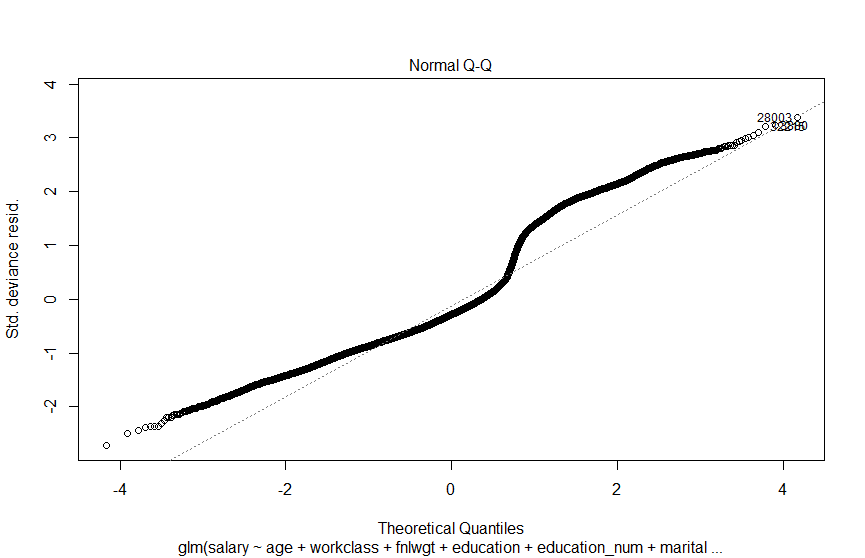
Q-Q plot:



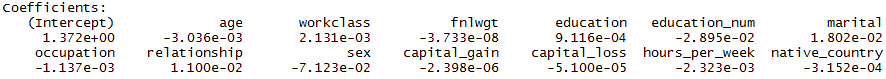
glm, gaussian prior probability assumption:



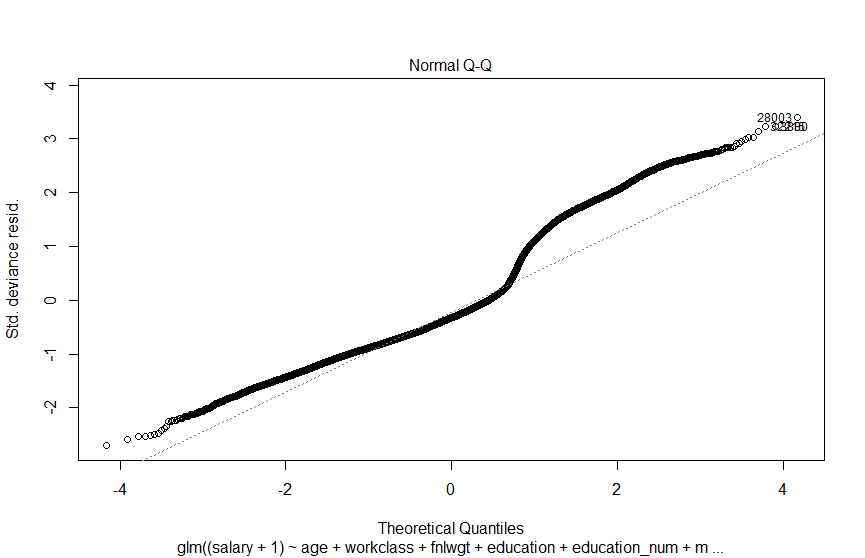
Q-Q plot:



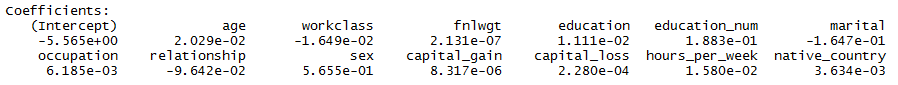
glm, gamma prior probability assumption:



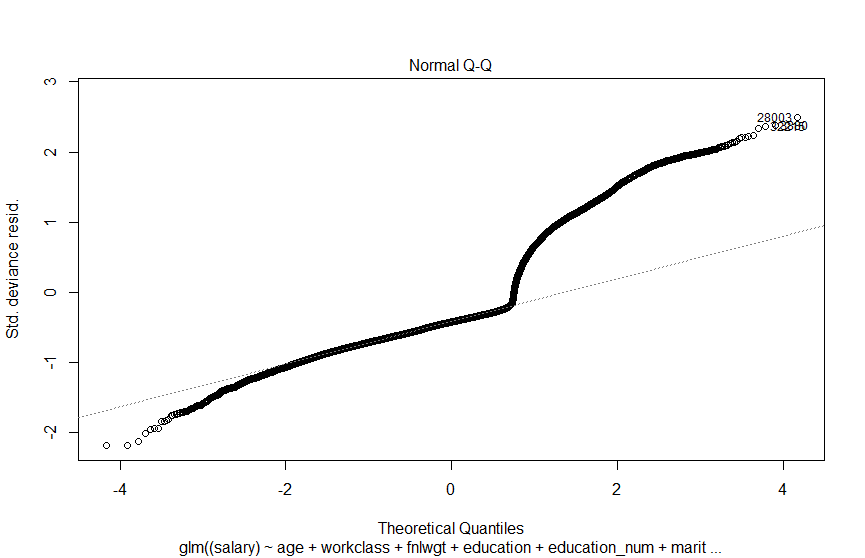
Q-Q plot:



glm, poisson prior probability assumption:



Q-Q plot:



Final summary table (anwser of 6.d):



From the above table and plot we can conclude the following patterns: (answer of 6.e)

1. glm with Gaussian prior probability is equal to lm, or we can say that normally linear regression is assuming that data distribution is Gaussian distribution.
2. fnlwgt, capital-gain, capital-loss and native-country are basically have no direct relation to a man’s salary. With the former race feature, we can say that these five features are unrelated features in this researching question.
3. glm with Binominal and Poisson prior probability have very high bias, which means men’s salary basically do not allow these 2 features. Since Gaussian distribution’s bias is lowest, we can say that men’s salary indeed allow Gaussian distribution, which fits our general intuition.
4. Different prior distribution can cause different coefficients for same feature, some even differs largely.

Answer of 6.f:

In all, from this data set analysis work, my teammate and I know that big data analysis can explore valuable information that cannot be reached by see the data itself. Using the clustering technologies such as k-means and knn and principal component analysis, we can filter some useless part of the data and analyze the core data into a new degree and know the connection of each value, the macro view of them. Furthermore, we even can make predictive analyze using such as glm function. We ensure that learning these can help us understand the truly meaning of the data we use.