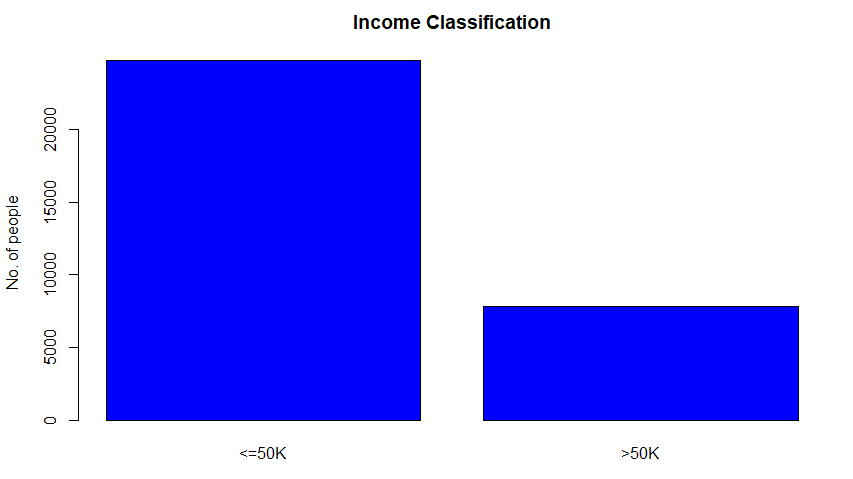
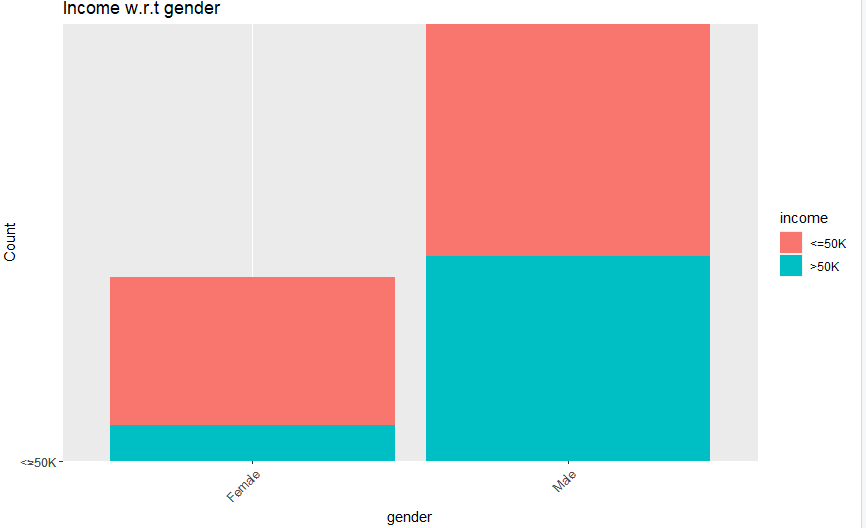
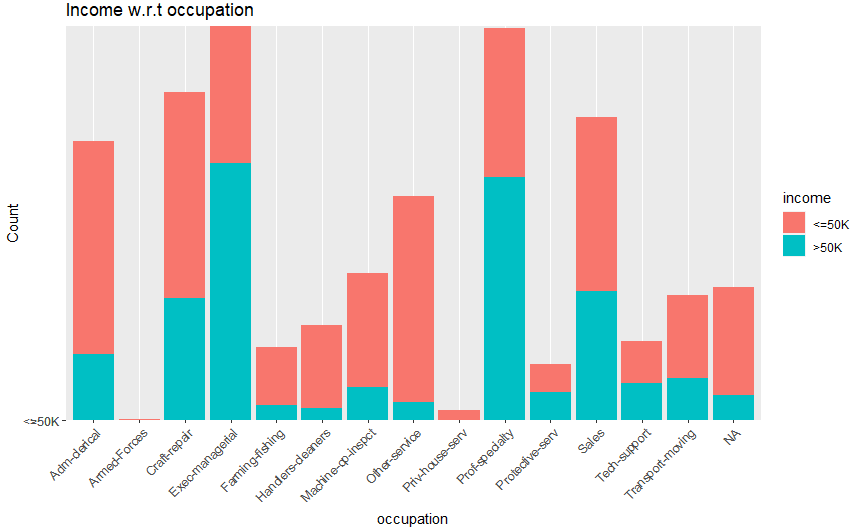
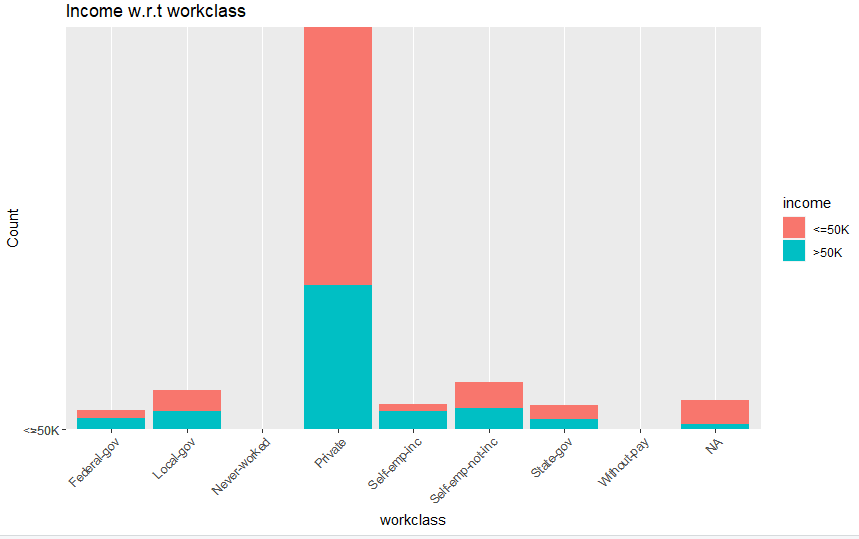
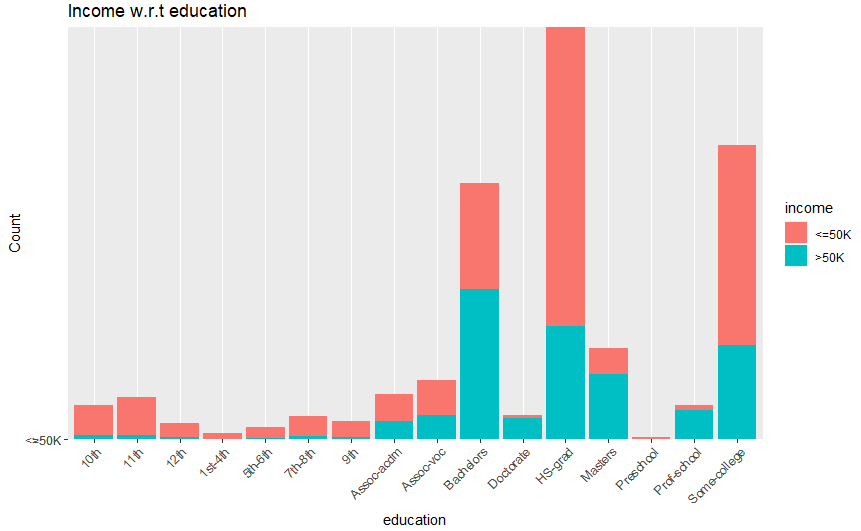
1. Pairs Plotting





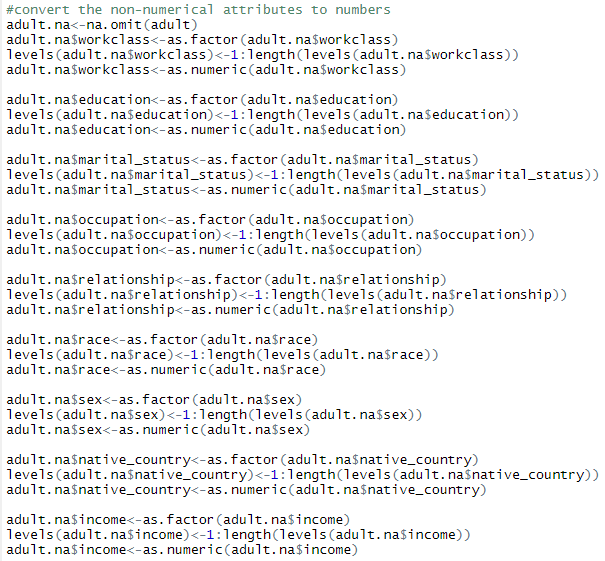






1. Prepare the Data

Some of the features are not numeric values, so changing is needed. Function level() and numeric() are used to get the numeric values.



Here are mappings of discrete values and numbers:

workclass: Private(3), Self-emp-not-inc(5), Self-emp-inc(4), Federal-gov(1), Local-gov(2), State-gov(6), Without-pay(7), Never-worked(out).

education: Bachelors(10), Some-college(16), 11th(2), HS-grad(12), Prof-school(15), Assoc-acdm(8), Assoc-voc(9), 9th(7), 7th-8th(6), 12th(3), Masters(13), 1st-4th(4), 10th(1), Doctorate(11), 5th-6th(5), Preschool(14).

marital-status: Married-civ-spouse(3), Divorced(1), Never-married(5), Separated(6), Widowed(7), Married-spouse-absent(4), Married-AF-spouse(2).

occupation: Tech-support(13), Craft-repair(3), Other-service(8), Sales(12), Exec-managerial(4), Prof-specialty(10), Handlers-cleaners(6), Machine-op-inspct(7), Adm-clerical(1), Farming-fishing(5), Transport-moving(14), Priv-house-serv(9), Protective-serv(11), Armed-Forces(2).

relationship: Wife(6), Own-child(4), Husband(1), Not-in-family(2), Other-relative(3), Unmarried(5).

race: White(5), Asian-Pac-Islander(2), Amer-Indian-Eskimo(1), Other(4), Black(3).

sex: Female(1), Male(2).

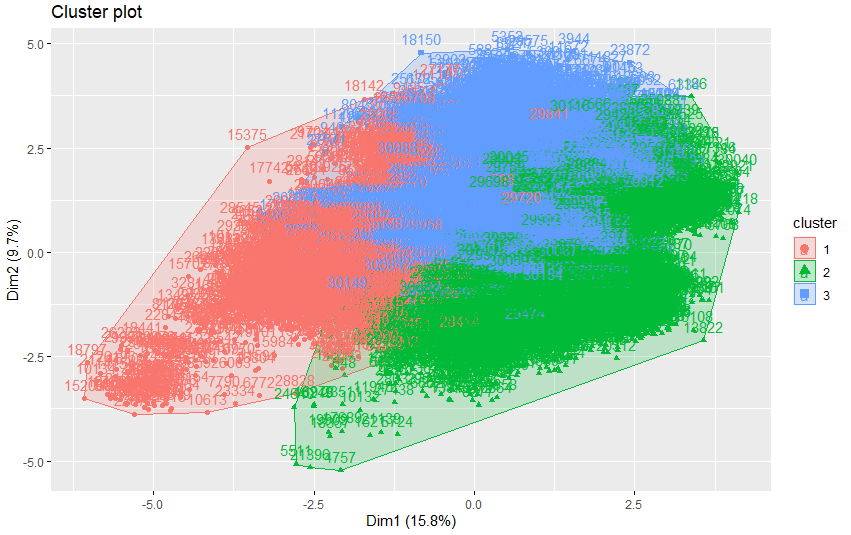
Native country:

Cambodia(1), Canada(2), China(3) , Columbia(4) , Cuba(5) , Dominican-Republic(6), Ecuador(7), El-Salvador(8), England(9), France(10), Germany(11), Greece(12), Guatemala(13), Haiti(14), Holand-Netherlands(15), Honduras(16), Hong(17), Hungary(18), India(19), Iran(20), Ireland(21), Italy(22) Jamaica(23), Japan(24), Laos(25), Mexico(26), Nicaragua(27), Outlying-US(Guam-USVI-etc)(28), Peru(29), Philippines(30), Poland(31), Portugal(32), Puerto-Rico(33), Scotland(34), South(35), Taiwan(36), Thailand(37), Trinadad&Tobago(38), United-States(39), Vietnam(40), Yugoslavia(41)

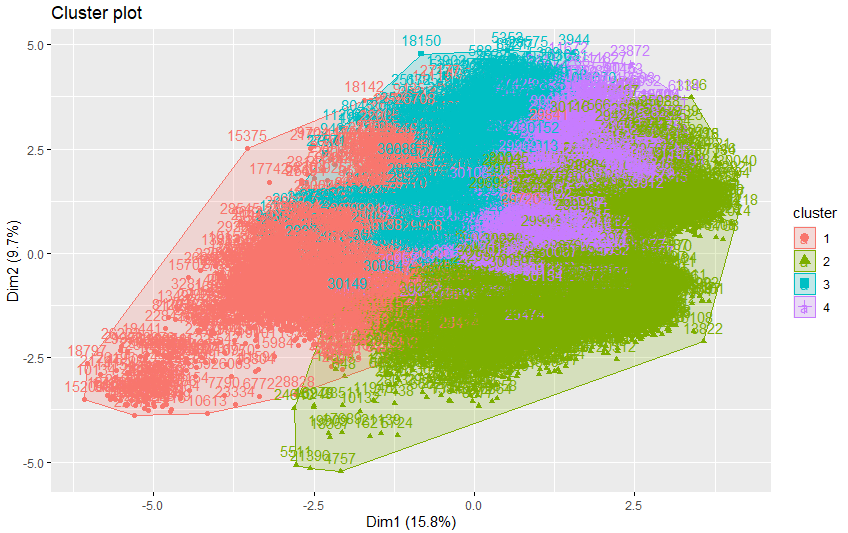
1. Clustering
2. Kmeans
3. centers=2



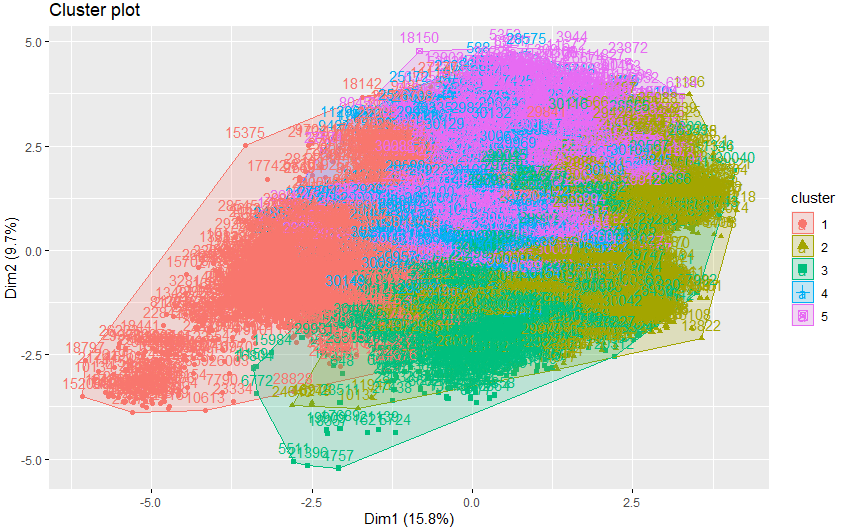
1. centers=3



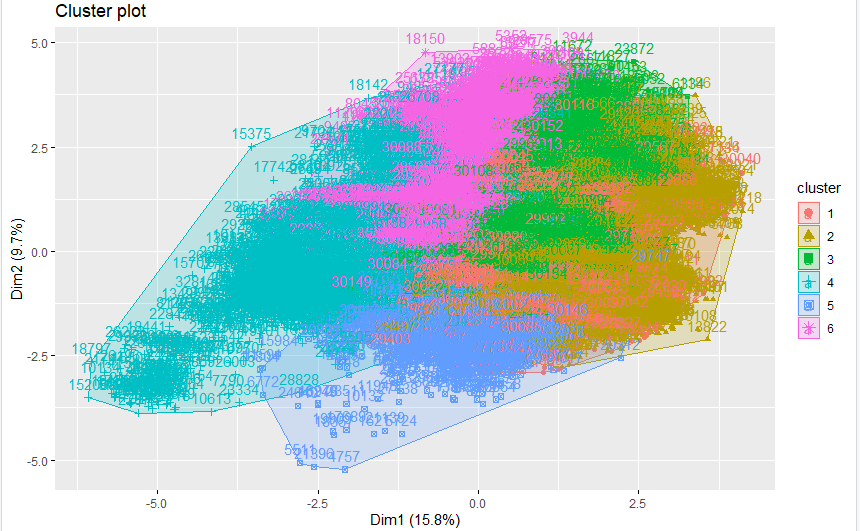
1. centers=4



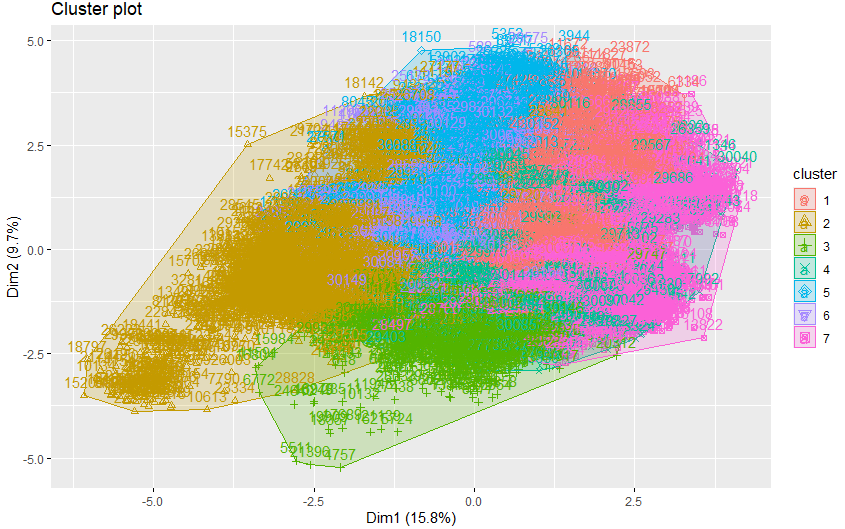
1. centers=5



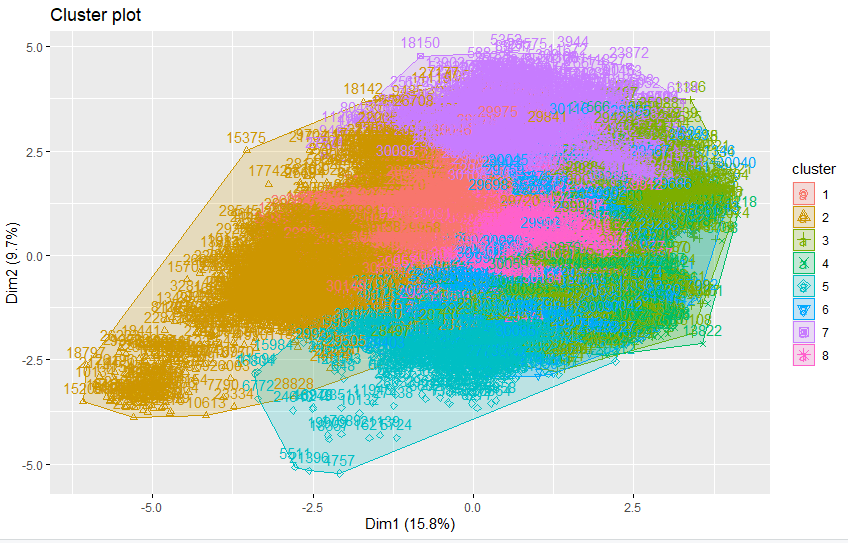
1. centers=6



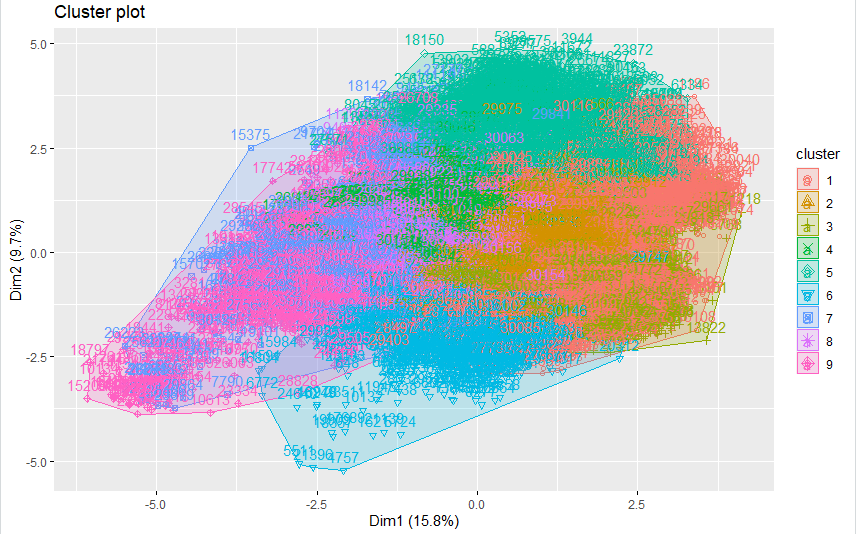
1. centers=7



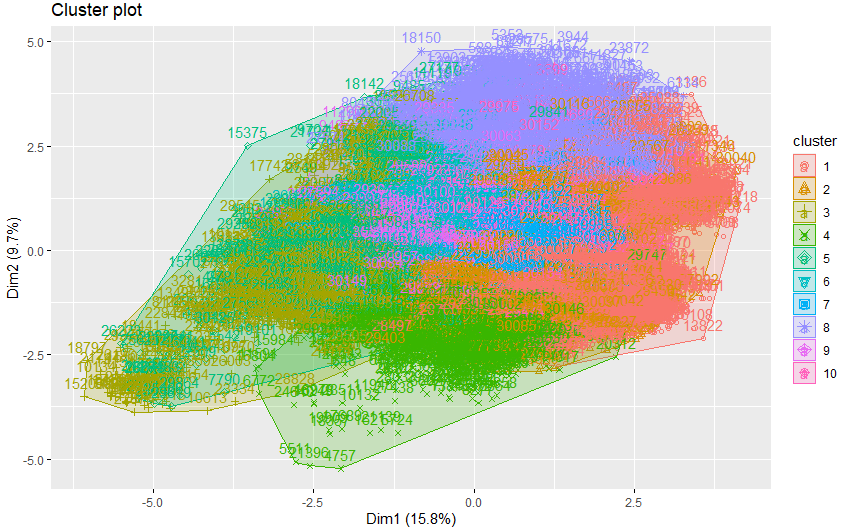
1. centers=8

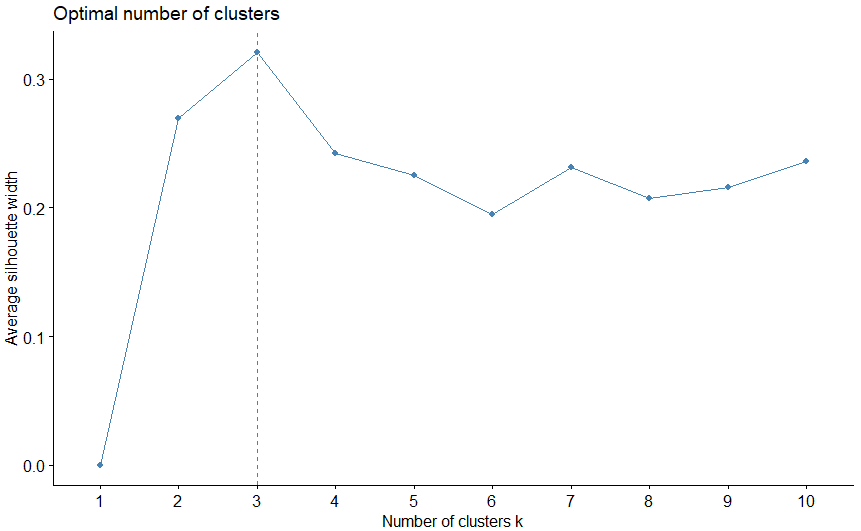


1. centers=9



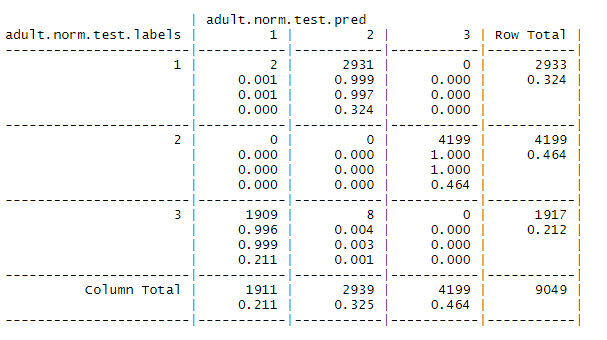
1. centers=10



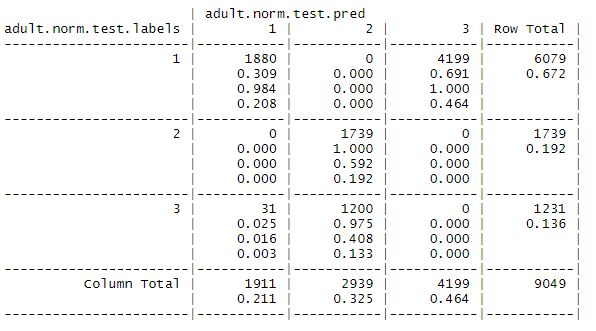


1. kNN

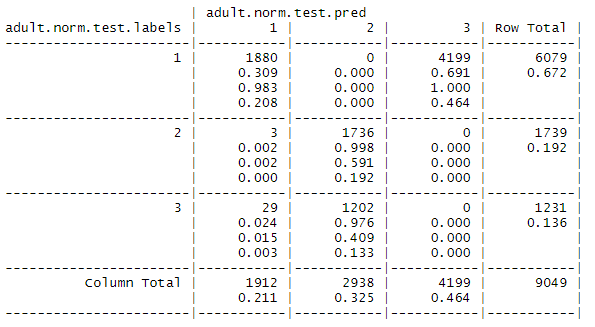
k = 2:



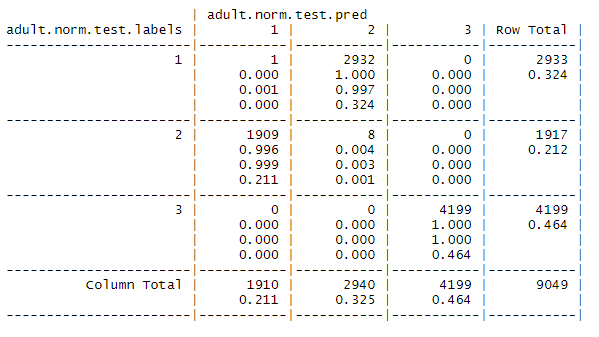
K = 3:



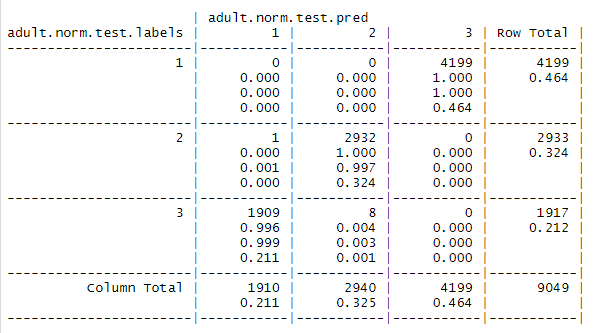
K = 4:



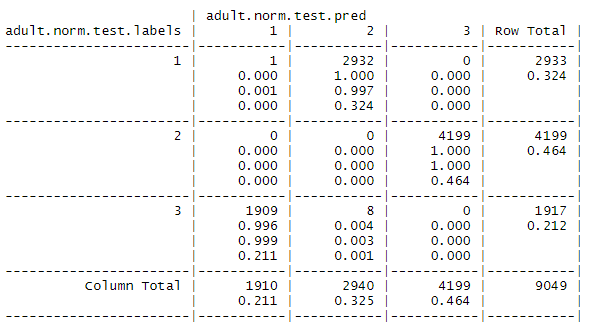
K = 5



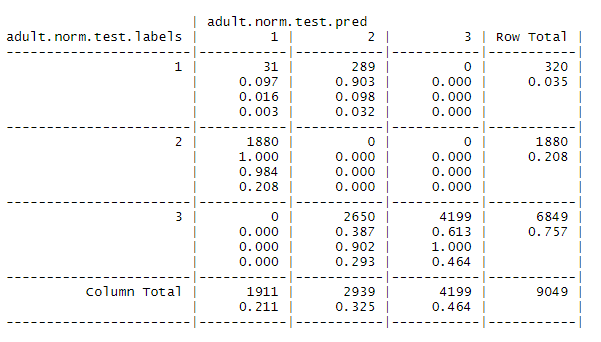
K = 6:



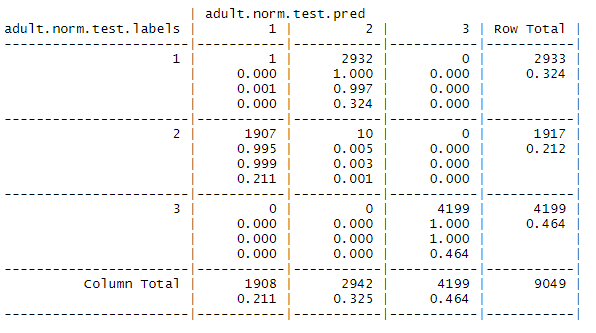
K = 7:



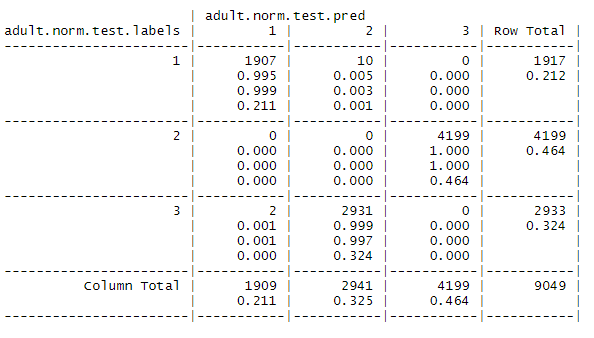
K = 8:



K = 9:

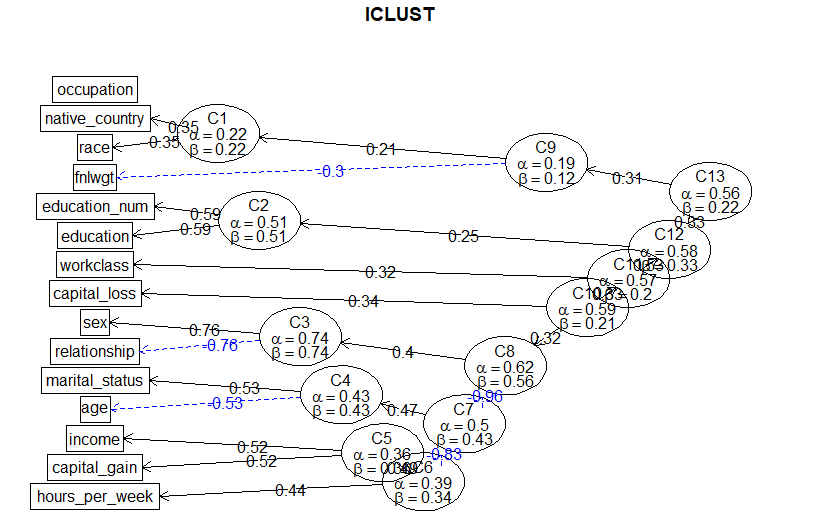


K = 10:

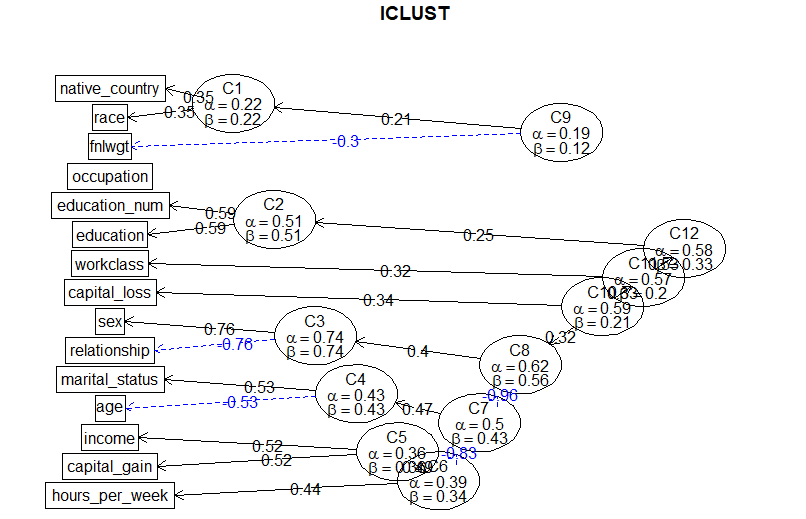


After we make k = 5 and let k increase, the prediction of clusters is getting more accurate. Although some accuracy is over 99%, some predictions label almost the whole cluster as another one, which means we need more clusters or some attributes are highly related.

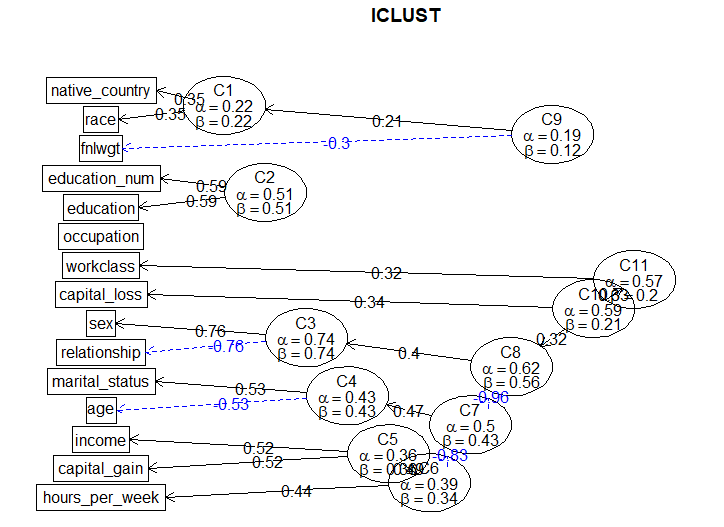
1. iClust
2. nclusters=2



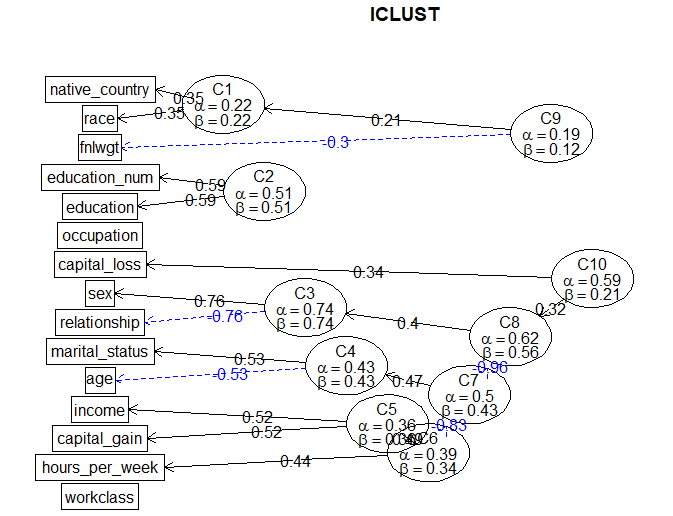
1. nclusters=3



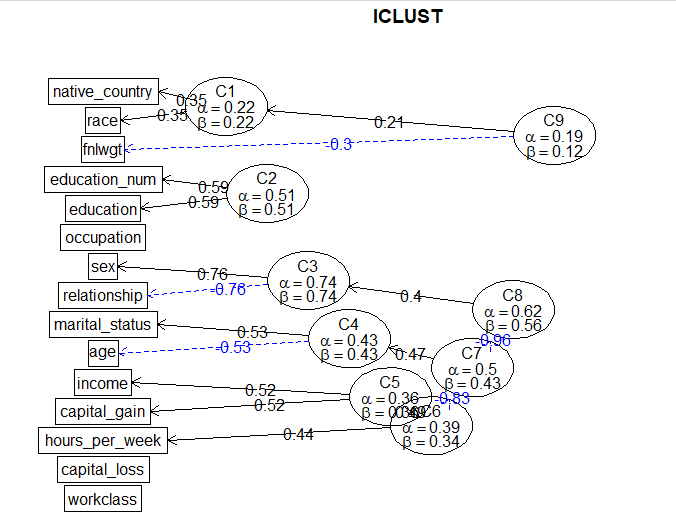
1. nclusters=4



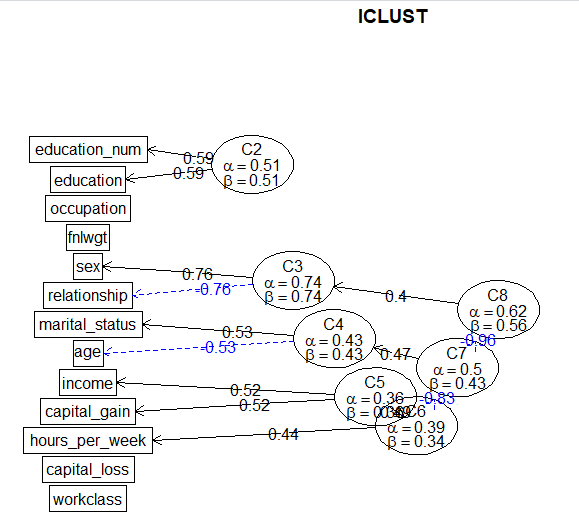
1. nclusters=5



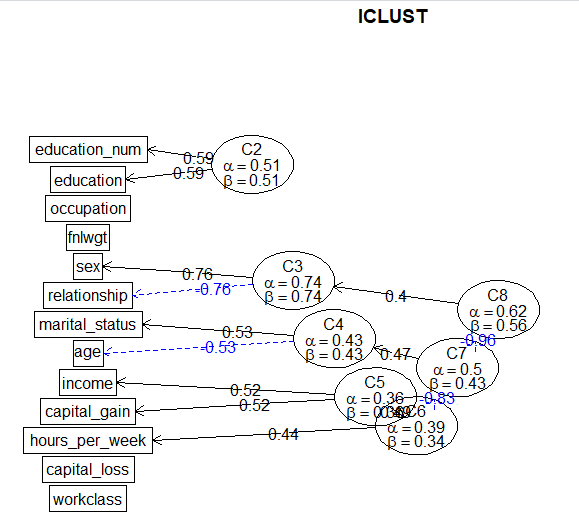
1. nclusters=6



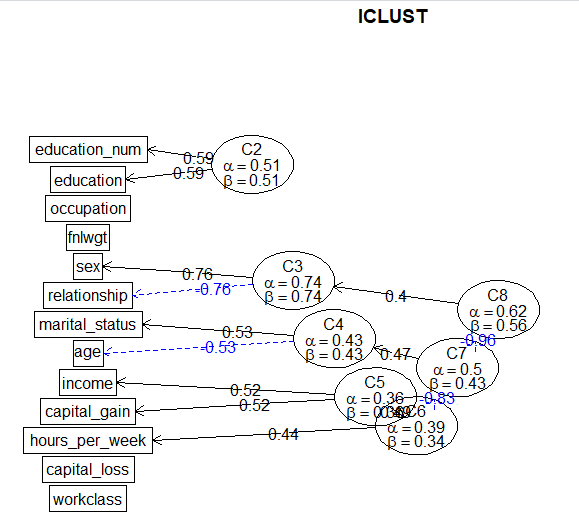
1. nclusters=7



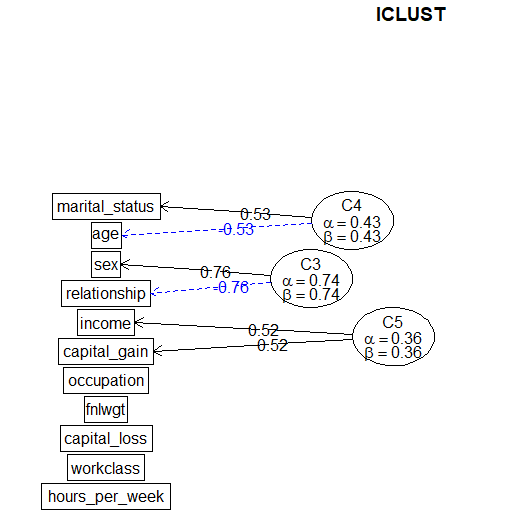
1. nclusters=8



1. nclusters=9



1. nclusters=10

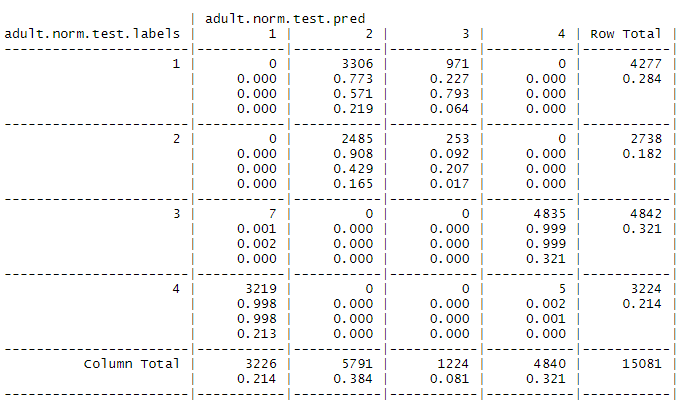


Our observation: the earlier two clusters combine, the more relative they are. As we can see from those graphs above. The hour per week, capital gain, relationship are closely connected with income. Also the education. However, the race and native country have nothing to do with the income.

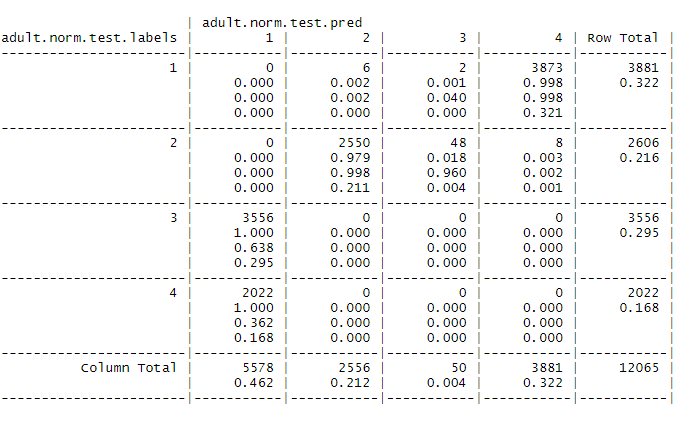
1. Prediction

(a) Knn for prediction

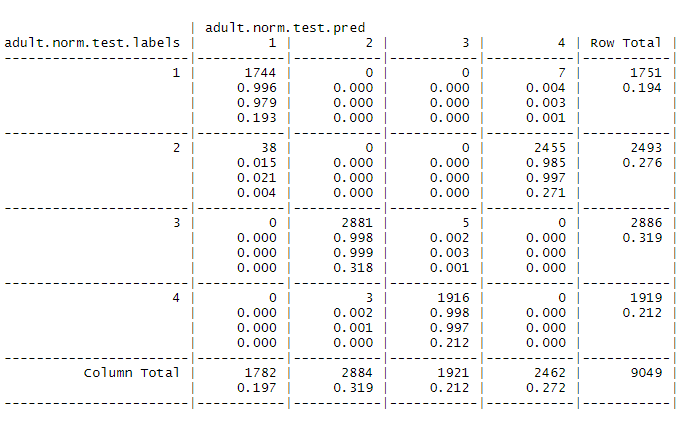
1. 50-50



1. 60-40



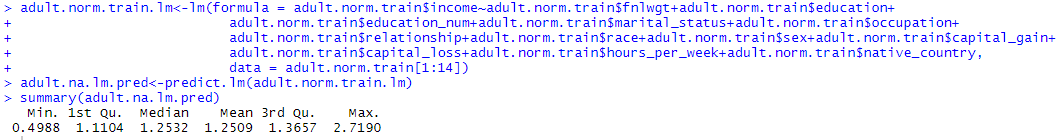
1. 70-30



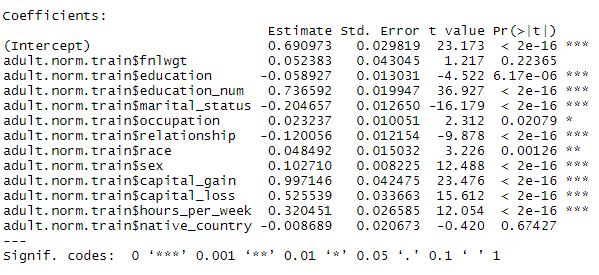
(b) Using lm() and glm():

1) 50-50

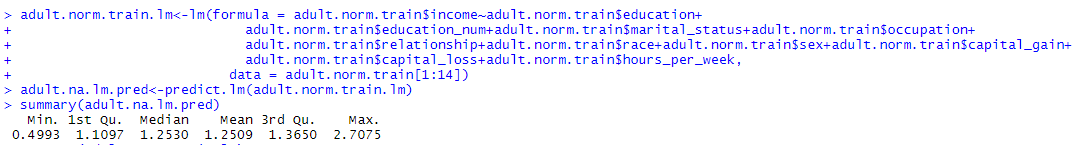
1. Using all the attributes:



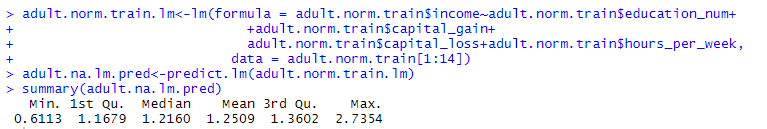
2. Using summary to check the probability of coefficient:



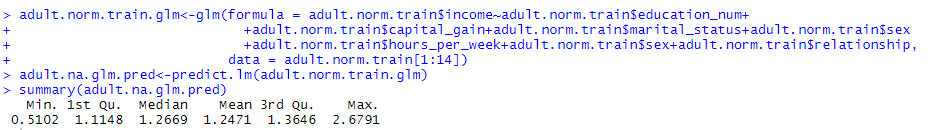
3. After getting rid of attributes ‘fnlwgt’ and ‘native country’:



4.After removing all discrete attributes:



5.Try several most related attributes which are concluded by iclust graphs:

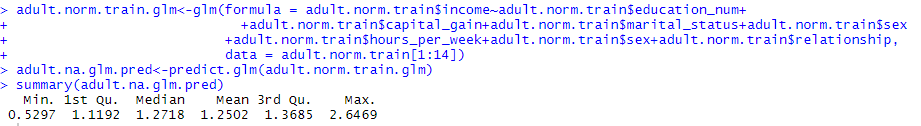


This is what the labels should be:

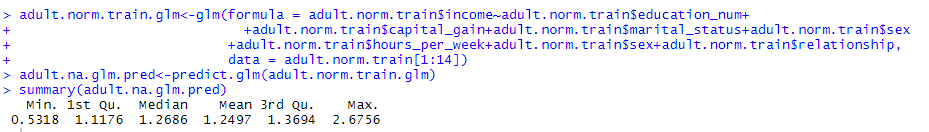


Seems that this prediction is closer to the real number. Now we try 60/40 and 70/30 on these attributes.

2)60-40



3)



Now, according to the summaries, 50-50 have the most accurate prediction among them.

5.Functions we wrote:

(1) normalize:

normalize<-function(x){((x-min(x))/(max(x)-min(x)))}

(2)foo—used to try different number of clusters of Kmeans:

foo<-function(data=adult.norm,feats,nn=8,low=2,high=10){

adult.n<-data[,feats]

adult.norm.nrows<-nrow(adult.n)

adult.norm.sample<-0.7

adult.norm.train.index<-sample(adult.norm.nrows,adult.norm.sample\*adult.norm.nrows)

adult.norm.train<-adult.n[adult.norm.train.index,]

adult.norm.test<-adult.n[-adult.norm.train.index,]

for(nc in low:high){

print("###############################")

print("")

print(nc)

print("###############################")

adult.norm.train.k4<-kmeans(adult.norm.train,centers=nc)

adult.norm.train.labels<-adult.norm.train.k4$cluster

adult.norm.test.k4<-kmeans(adult.norm.test,centers=nc)

adult.norm.test.labels<-adult.norm.test.k4$cluster

adult.norm.test.pred<-knn(adult.norm.train,adult.norm.test,adult.norm.train.k4$cluster,k=nn)

str(adult.norm.test.pred)

adult.norm.ct<-CrossTable(adult.norm.test.labels, adult.norm.test.pred,prop.chisq=FALSE)

##confusionMatrix(adult.norm.test.pred,adult.norm.test.labels)

}

}

6.What we have learnt from this project

First, we know how to convert text into numbers so that they can be used for calculation. The normalization and scale are helpful for making calculation more efficient. Then, we learnt how to kmeans function of R to help with clustering and we realized that more number of clusters does not mean better. To find the best number of clusters for clustering, we used factoextra::fviz\_nbclust() function. Later we applied different k for knn. We also learned that bigger k does not mean better classification. There might be noise data which may influence which cluster the element should be in. Last, we used linear regression to predict the income for test dataset. We divided data into 50-50, 60-40 and 70-30. We found that 50-50 actually do the best prediction among them. From this we learnt that bigger training set may not produce better model for prediction because the model may be more fitted with training set rather than general data.