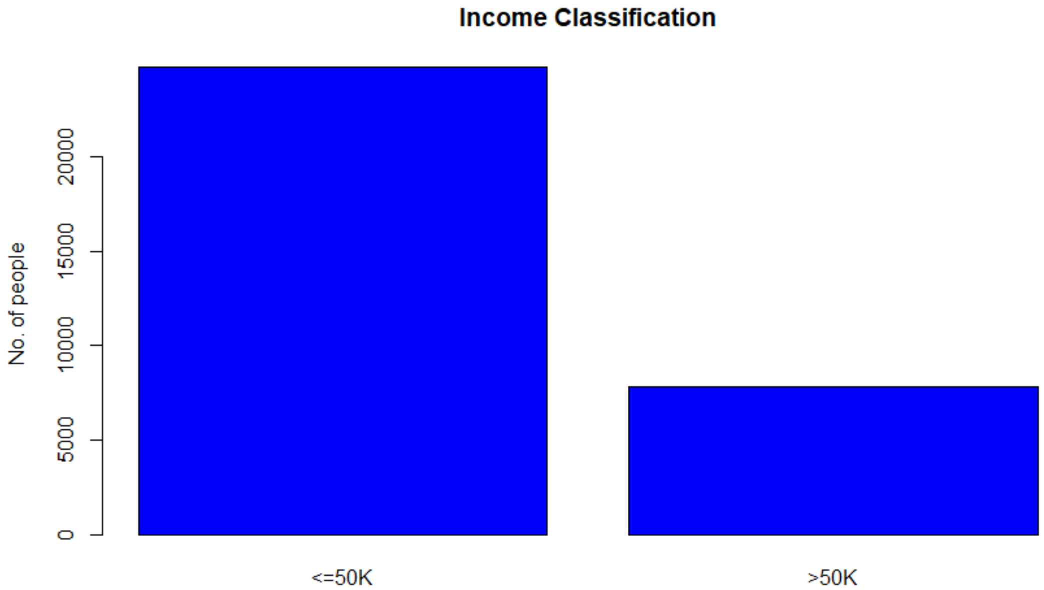
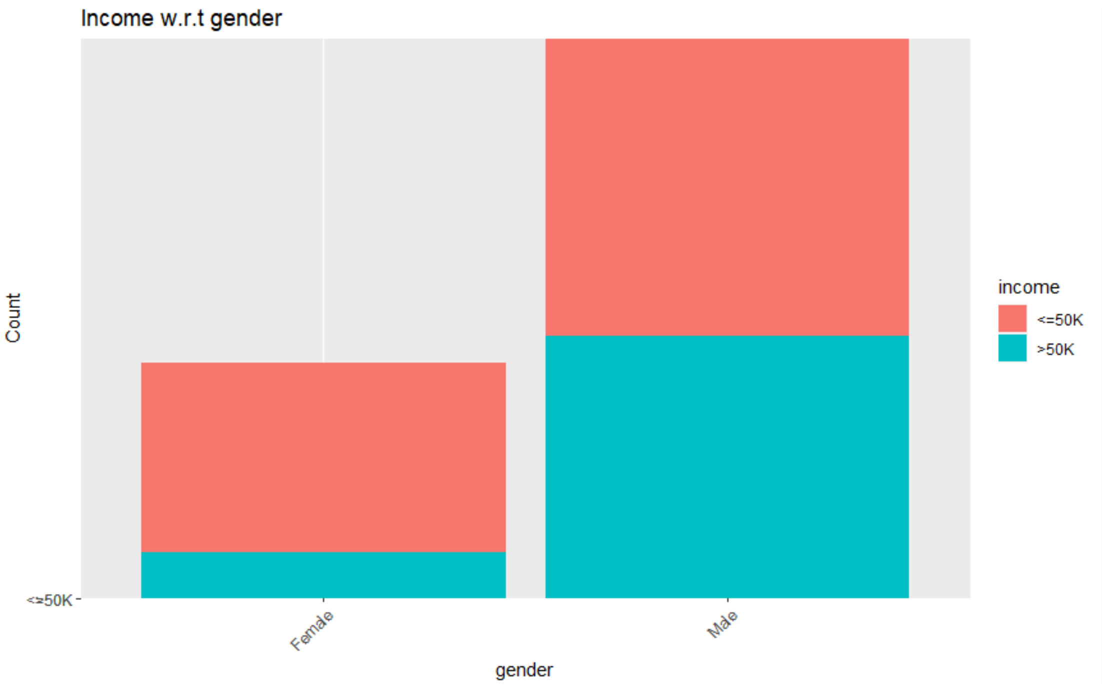
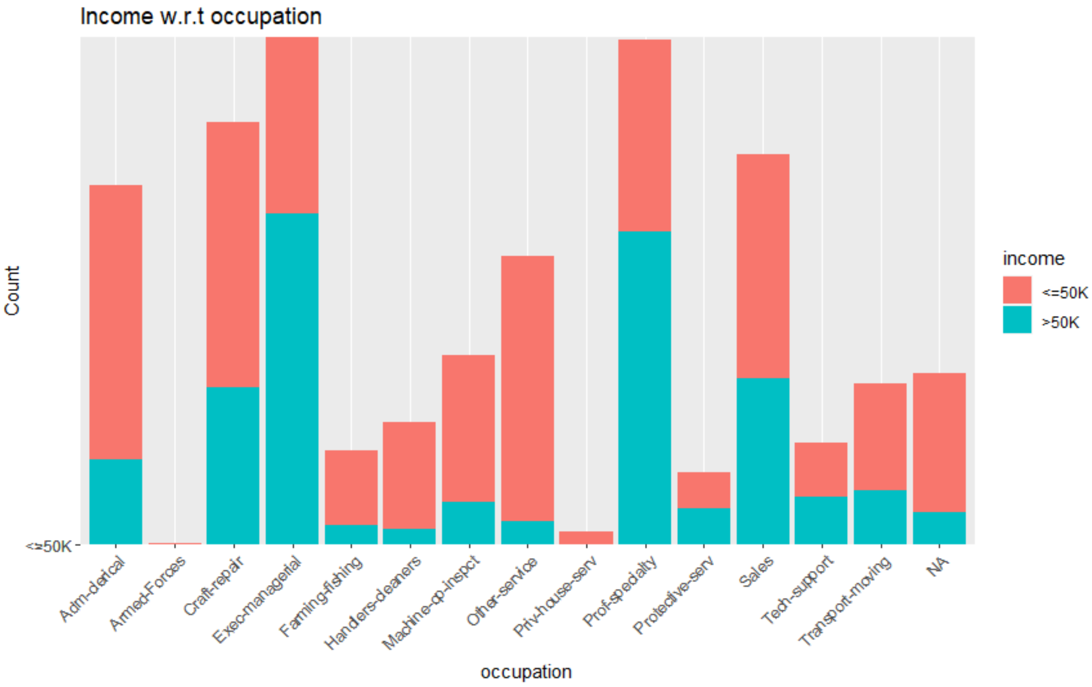
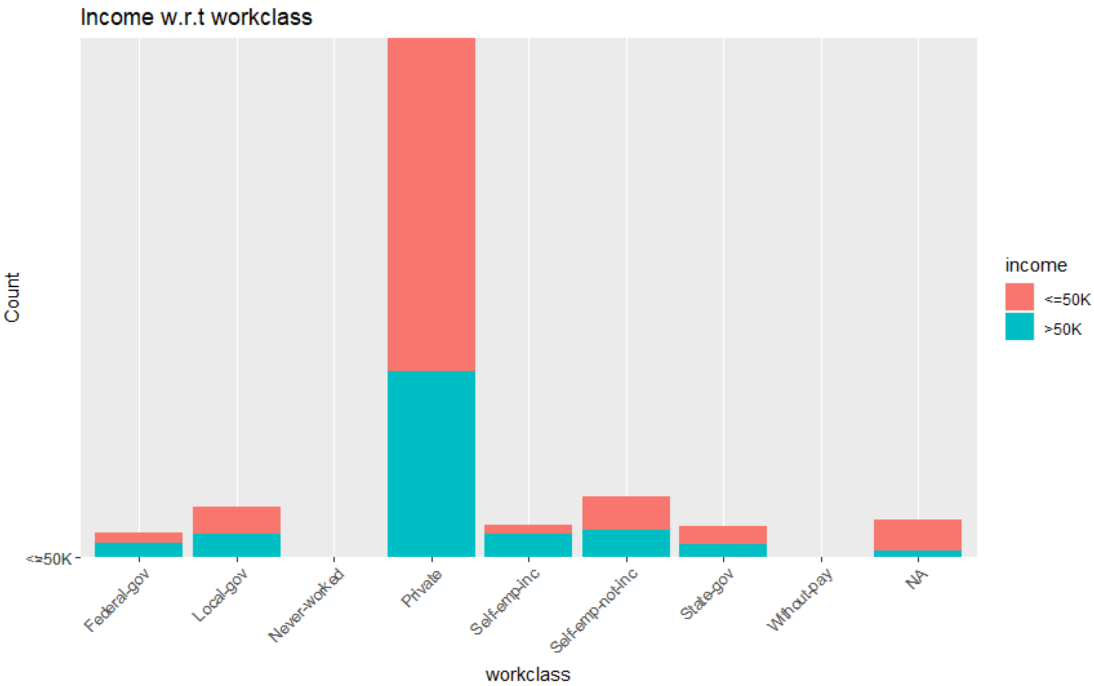
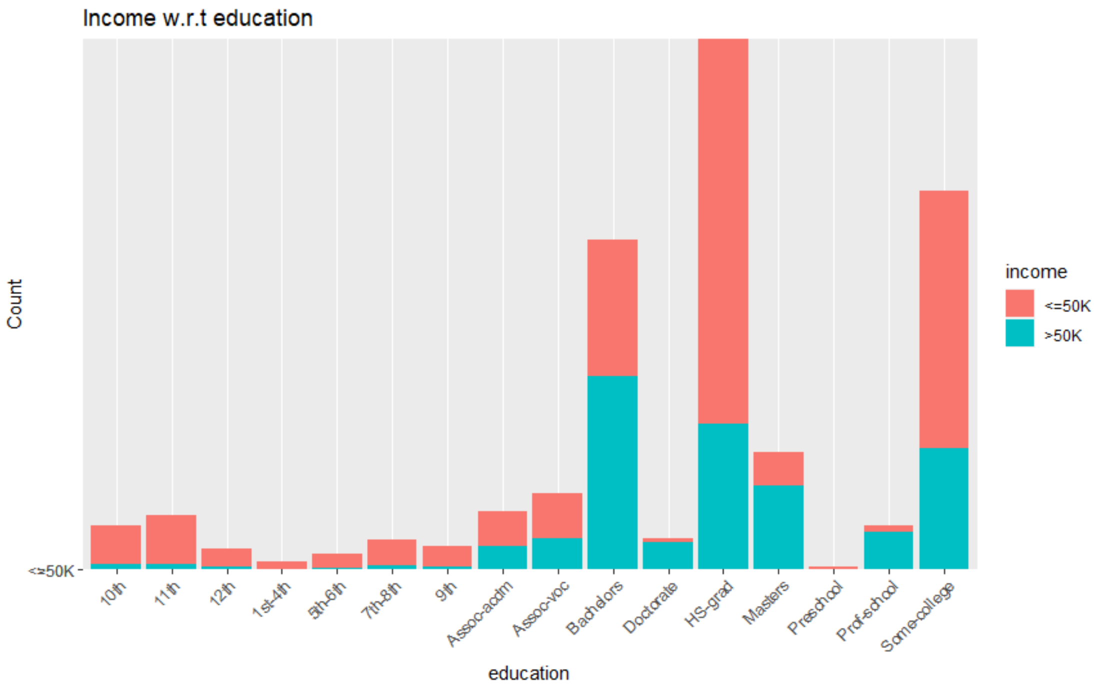
1. Pairs Plotting






Income w.r.t occupation
Count
<=50K
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
NA
occupation
income
<=50K
>50K


Income w.r.t workclass
Count
<=50K
Federal-gov
Local-gov
Never-worked
Private
Self-emp-inc
Self-emp-not-inc
State-gov
Without-pay
NA
workclass
income
<=50K
>50K


Income w.r.t education
Count
<=50K
10th
11th
12th
1st-4th
5th-6th
7th-8th
9th
Assoc-acdm
Assoc-voc
Bachelors
Doctorate
HS-grad
Masters
Preschool
Prof-school
Some-college
education
income
<=50K
>50K

2. 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.

```
#convert the non-numerical attributes to numbers
adult.na<-na.omit(adult)
adult.na$workclass<-as.factor(adult.na$workclass)
levels(adult.na$workclass)<-1:length(levels(adult.na$workclass))
adult.na$workclass<-as.numeric(adult.na$workclass)

adult.na$education<-as.factor(adult.na$education)
levels(adult.na$education)<-1:length(levels(adult.na$education))
adult.na$education<-as.numeric(adult.na$education)

adult.na$marital_status<-as.factor(adult.na$marital_status)
levels(adult.na$marital_status)<-1:length(levels(adult.na$marital_status))
adult.na$marital_status<-as.numeric(adult.na$marital_status)

adult.na$occupation<-as.factor(adult.na$occupation)
levels(adult.na$occupation)<-1:length(levels(adult.na$occupation))
adult.na$occupation<-as.numeric(adult.na$occupation)

adult.na$relationship<-as.factor(adult.na$relationship)
levels(adult.na$relationship)<-1:length(levels(adult.na$relationship))
adult.na$relationship<-as.numeric(adult.na$relationship)

adult.na$race<-as.factor(adult.na$race)
levels(adult.na$race)<-1:length(levels(adult.na$race))
adult.na$race<-as.numeric(adult.na$race)

adult.na$sex<-as.factor(adult.na$sex)
levels(adult.na$sex)<-1:length(levels(adult.na$sex))
adult.na$sex<-as.numeric(adult.na$sex)

adult.na$native_country<-as.factor(adult.na$native_country)
levels(adult.na$native_country)<-1:length(levels(adult.na$native_country))
adult.na$native_country<-as.numeric(adult.na$native_country)

adult.na$income<-as.factor(adult.na$income)
levels(adult.na$income)<-1:length(levels(adult.na$income))
adult.na$income<-as.numeric(adult.na$income)
```

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).

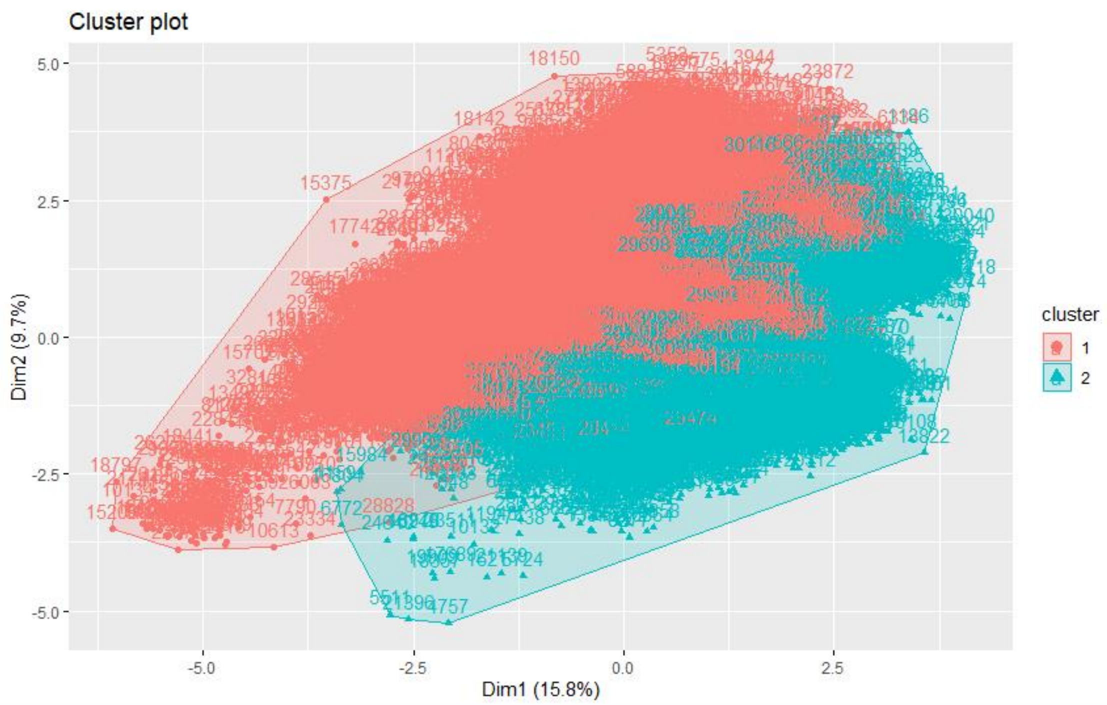
Native country:

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)

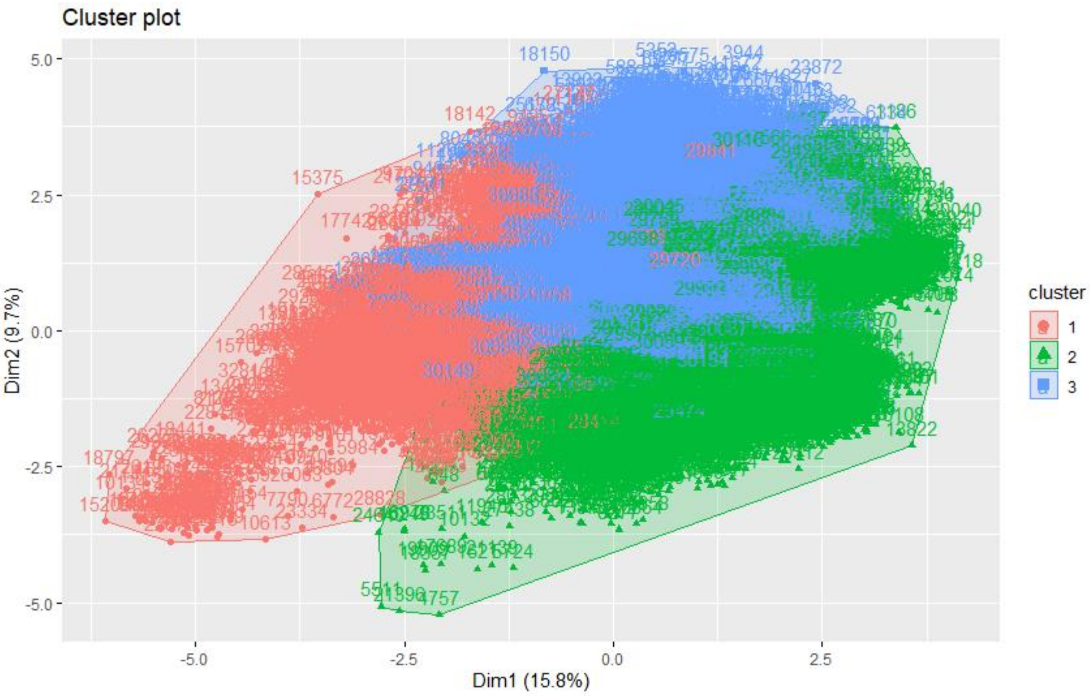
3. Clustering

1) Kmeans

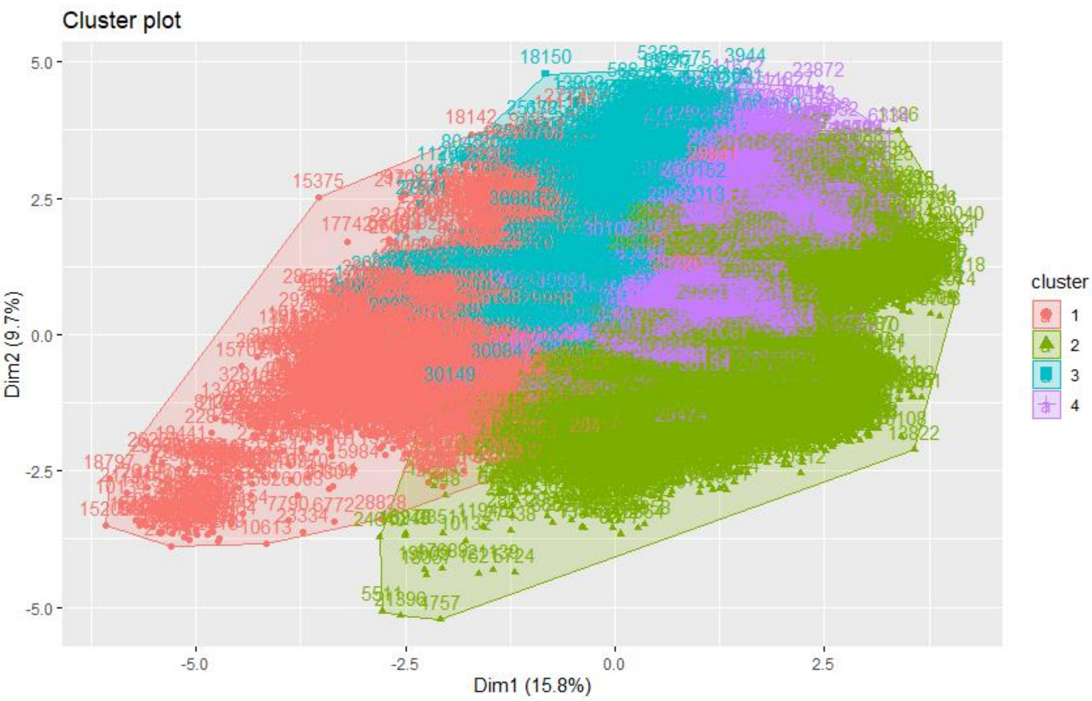
a) centers=2



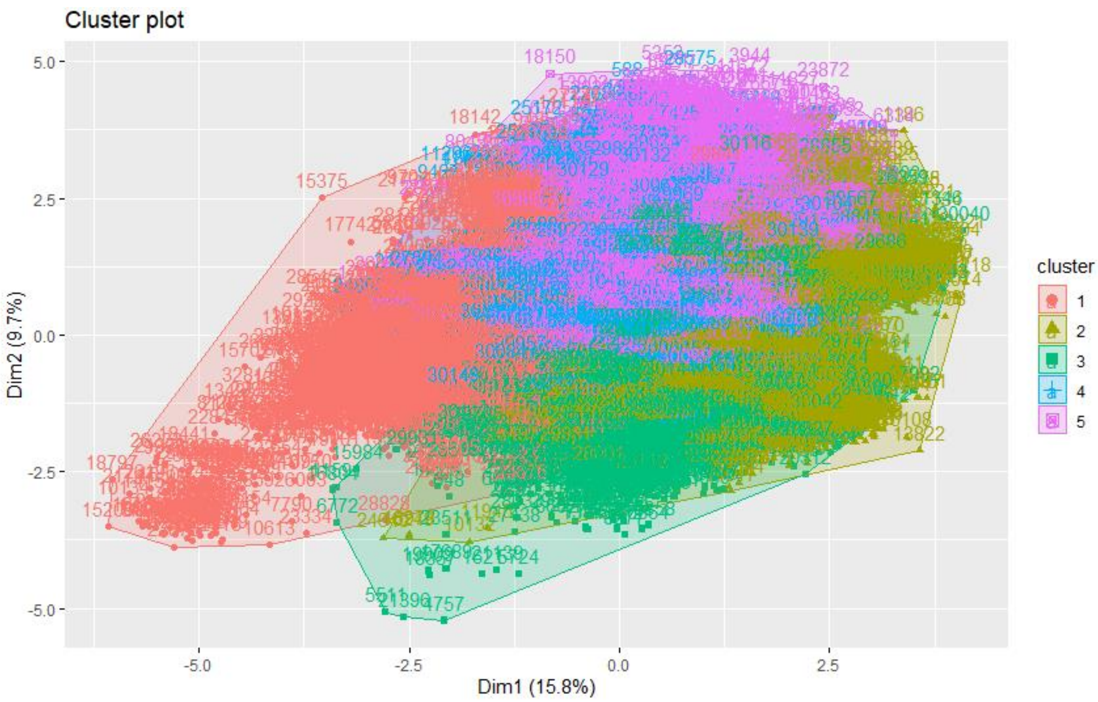
b) centers=3



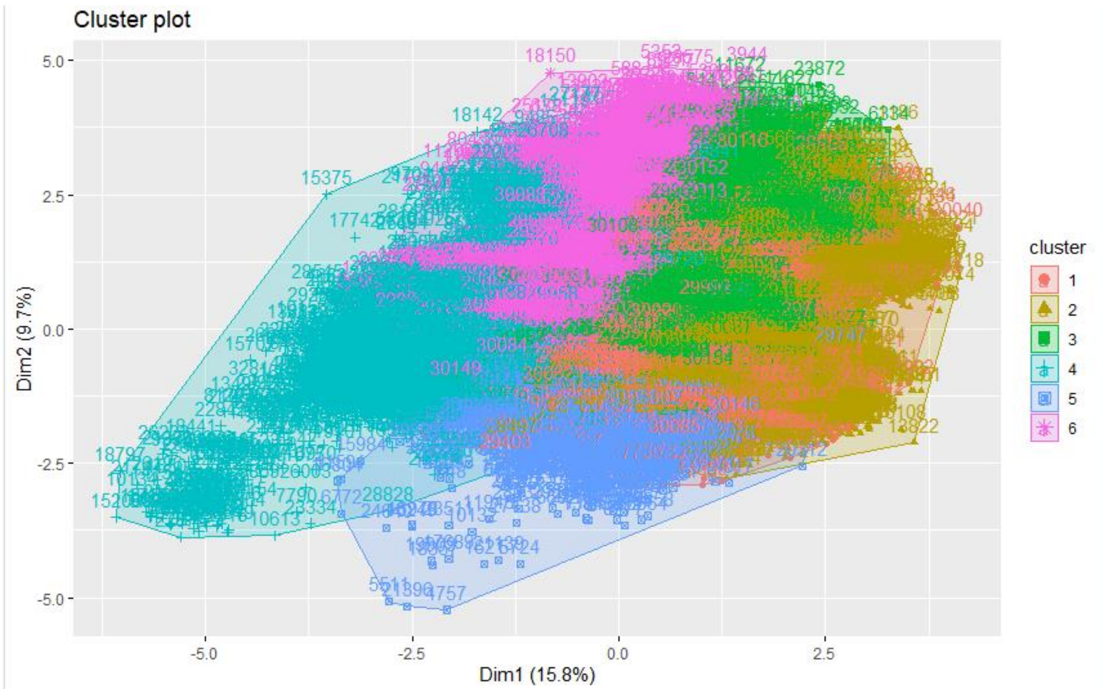
c) centers=4



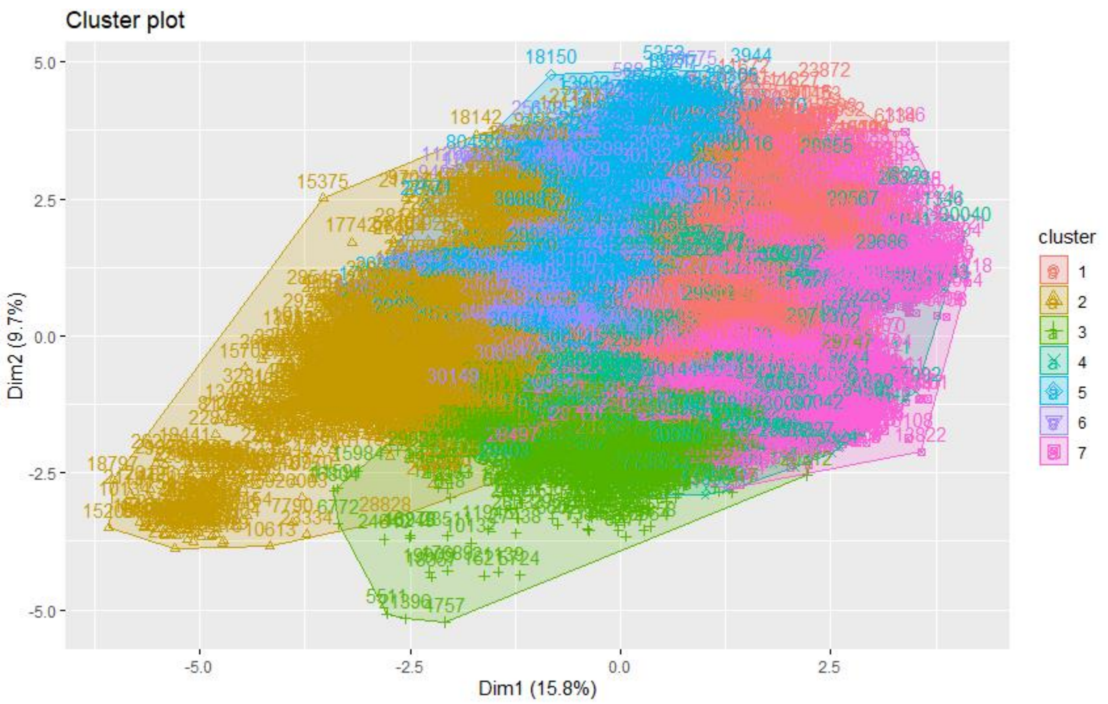
d) centers=5



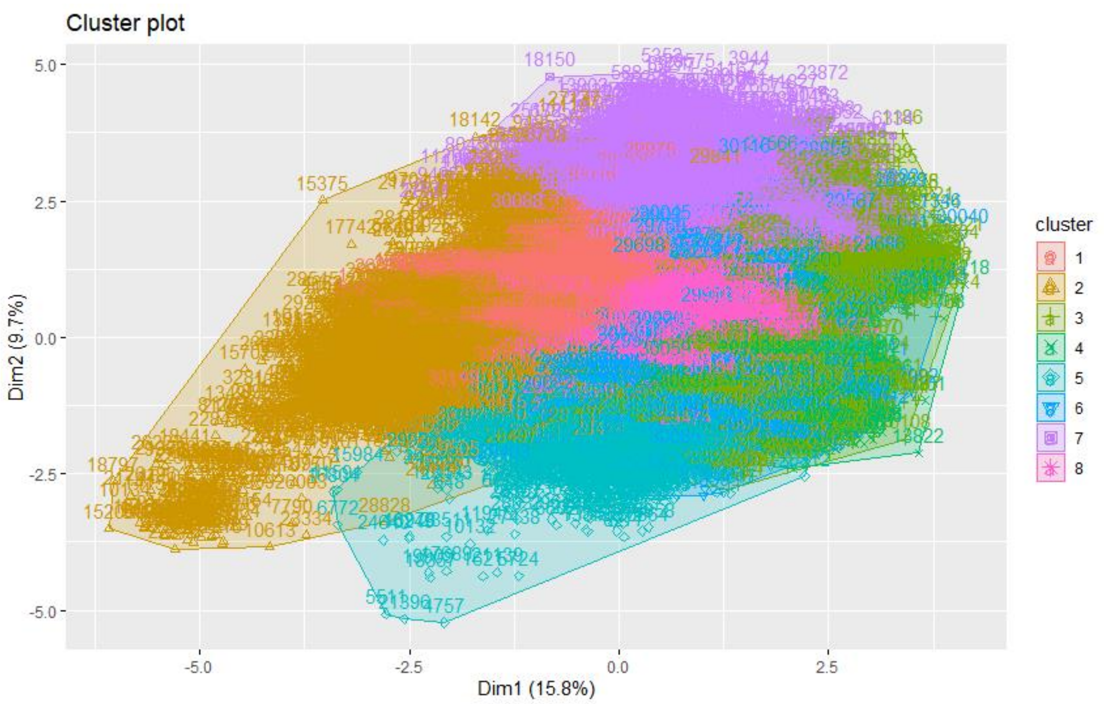
e) centers=6



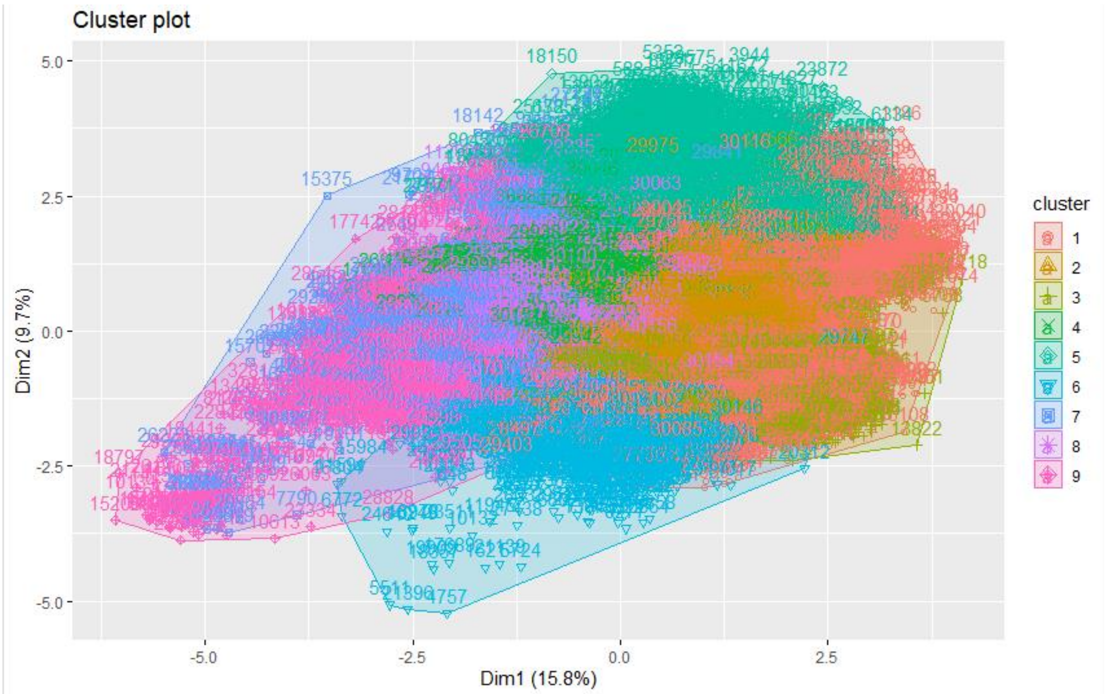
f) centers=7



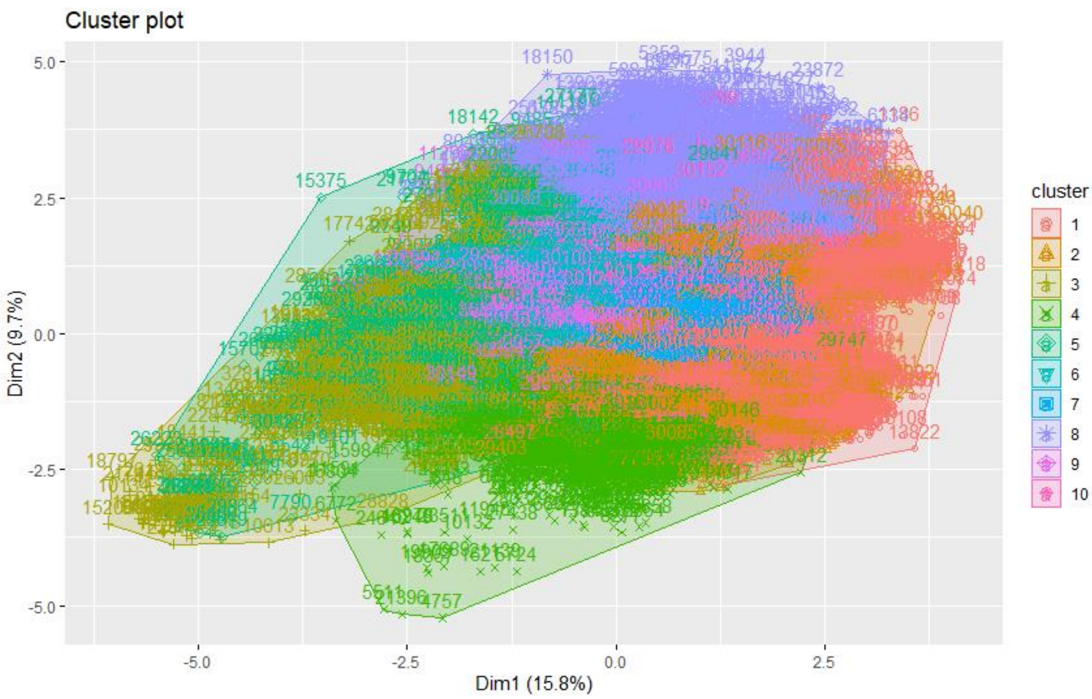
g) centers=8

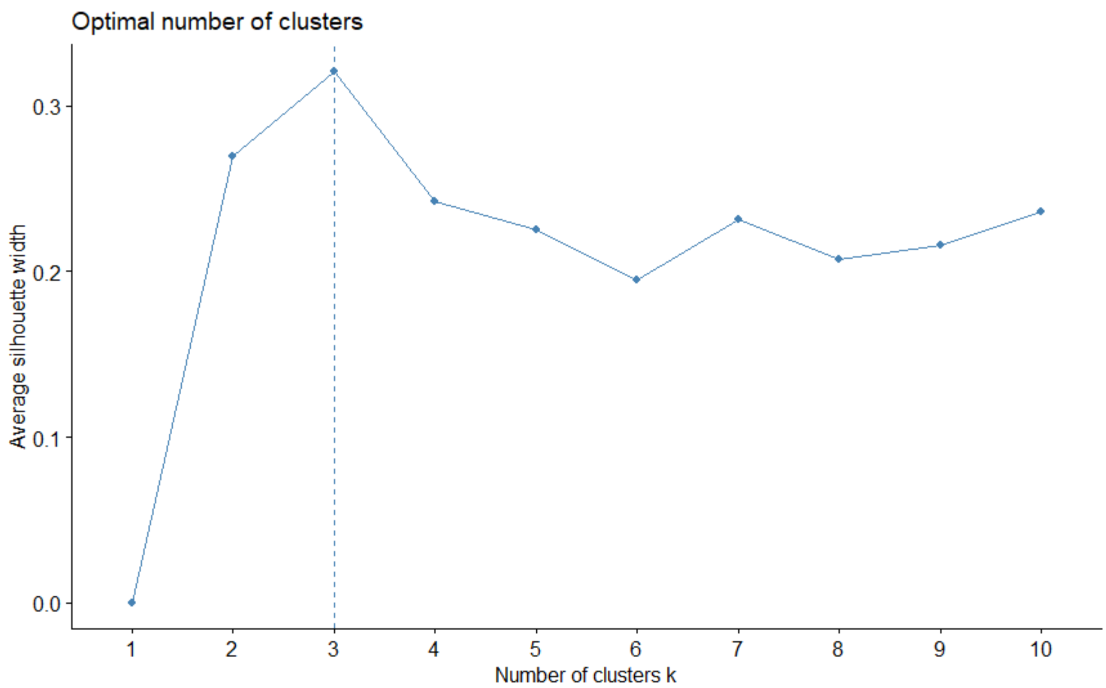


h) centers=9


Cluster plot
Dim1 (15.8%)
Dim2 (9.7%)
cluster
1
2
3
4
5
6
7
8
9

i) centers=10


Cluster plot
Dim1 (15.8%)
Dim2 (9.7%)
cluster
1
2
3
4
5
6
7
8
9
10



2) kNN

k = 2:

```
                        | adult.norm.test.pred
adult.norm.test.labels |           1 |           2 |           3 |   Row Total |
-----------------------|-------------|-------------|-------------|-------------|
                     1 |           2 |        2931 |           0 |        2933 |
                       |       0.001 |       0.999 |       0.000 |       0.324 |
                       |       0.001 |       0.997 |       0.000 |             |
                       |       0.000 |       0.324 |       0.000 |             |
-----------------------|-------------|-------------|-------------|-------------|
                     2 |           0 |           0 |        4199 |        4199 |
                       |       0.000 |       0.000 |       1.000 |       0.464 |
                       |       0.000 |       0.000 |       1.000 |             |
                       |       0.000 |       0.000 |       0.464 |             |
-----------------------|-------------|-------------|-------------|-------------|
                     3 |        1909 |           8 |           0 |        1917 |
                       |       0.996 |       0.004 |       0.000 |       0.212 |
                       |       0.999 |       0.003 |       0.000 |             |
                       |       0.211 |       0.001 |       0.000 |             |
-----------------------|-------------|-------------|-------------|-------------|
          Column Total |        1911 |        2939 |        4199 |        9049 |
                       |       0.211 |       0.325 |       0.464 |             |
-----------------------|-------------|-------------|-------------|-------------|
```

K = 3:

	adult.norm.test.pred			
adult.norm.test.labels	1	2	3	Row Total
1	1880 0.309 0.984 0.208	0 0.000 0.000 0.000	4199 0.691 1.000 0.464	6079 0.672
2	0 0.000 0.000 0.000	1739 1.000 0.592 0.192	0 0.000 0.000 0.000	1739 0.192
3	31 0.025 0.016 0.003	1200 0.975 0.408 0.133	0 0.000 0.000 0.000	1231 0.136
Column Total	1911 0.211	2939 0.325	4199 0.464	9049

K = 4:

	adult.norm.test.pred			
adult.norm.test.labels	1	2	3	Row Total
1	1880 0.309 0.983 0.208	0 0.000 0.000 0.000	4199 0.691 1.000 0.464	6079 0.672
2	3 0.002 0.002 0.000	1736 0.998 0.591 0.192	0 0.000 0.000 0.000	1739 0.192
3	29 0.024 0.015 0.003	1202 0.976 0.409 0.133	0 0.000 0.000 0.000	1231 0.136
Column Total	1912 0.211	2938 0.325	4199 0.464	9049

K = 5

```
                       | adult.norm.test.pred 
adult.norm.test.labels |           1 |           2 |           3 |   Row Total | 
-----------------------|-------------|-------------|-------------|-------------|
                     1 |           1 |        2932 |           0 |        2933 | 
                       |       0.000 |       1.000 |       0.000 |       0.324 | 
                       |       0.001 |       0.997 |       0.000 |             | 
                       |       0.000 |       0.324 |       0.000 |             | 
-----------------------|-------------|-------------|-------------|-------------|
                     2 |        1909 |           8 |           0 |        1917 | 
                       |       0.996 |       0.004 |       0.000 |       0.212 | 
                       |       0.999 |       0.003 |       0.000 |             | 
                       |       0.211 |       0.001 |       0.000 |             | 
-----------------------|-------------|-------------|-------------|-------------|
                     3 |           0 |           0 |        4199 |        4199 | 
                       |       0.000 |       0.000 |       1.000 |       0.464 | 
                       |       0.000 |       0.000 |       1.000 |             | 
                       |       0.000 |       0.000 |       0.464 |             | 
-----------------------|-------------|-------------|-------------|-------------|
          Column Total |        1910 |        2940 |        4199 |        9049 | 
                       |       0.211 |       0.325 |       0.464 |             | 
-----------------------|-------------|-------------|-------------|-------------|
```

K = 6:

```
                       | adult.norm.test.pred 
adult.norm.test.labels |           1 |           2 |           3 |   Row Total | 
-----------------------|-------------|-------------|-------------|-------------|
                     1 |           0 |           0 |        4199 |        4199 | 
                       |       0.000 |       0.000 |       1.000 |       0.464 | 
                       |       0.000 |       0.000 |       1.000 |             | 
                       |       0.000 |       0.000 |       0.464 |             | 
-----------------------|-------------|-------------|-------------|-------------|
                     2 |           1 |        2932 |           0 |        2933 | 
                       |       0.000 |       1.000 |       0.000 |       0.324 | 
                       |       0.001 |       0.997 |       0.000 |             | 
                       |       0.000 |       0.324 |       0.000 |             | 
-----------------------|-------------|-------------|-------------|-------------|
                     3 |        1909 |           8 |           0 |        1917 | 
                       |       0.996 |       0.004 |       0.000 |       0.212 | 
                       |       0.999 |       0.003 |       0.000 |             | 
                       |       0.211 |       0.001 |       0.000 |             | 
-----------------------|-------------|-------------|-------------|-------------|
          Column Total |        1910 |        2940 |        4199 |        9049 | 
                       |       0.211 |       0.325 |       0.464 |             | 
-----------------------|-------------|-------------|-------------|-------------|
```

K = 7:

	adult.norm.test.pred			
adult.norm.test.labels	1	2	3	Row Total
1	1 0.000 0.001 0.000	2932 1.000 0.997 0.324	0 0.000 0.000 0.000	2933 0.324
2	0 0.000 0.000 0.000	0 0.000 0.000 0.000	4199 1.000 1.000 0.464	4199 0.464
3	1909 0.996 0.999 0.211	8 0.004 0.003 0.001	0 0.000 0.000 0.000	1917 0.212
Column Total	1910 0.211	2940 0.325	4199 0.464	9049

K = 8:

	adult.norm.test.pred			
adult.norm.test.labels	1	2	3	Row Total
1	31 0.097 0.016 0.003	289 0.903 0.098 0.032	0 0.000 0.000 0.000	320 0.035
2	1880 1.000 0.984 0.208	0 0.000 0.000 0.000	0 0.000 0.000 0.000	1880 0.208
3	0 0.000 0.000 0.000	2650 0.387 0.902 0.293	4199 0.613 1.000 0.464	6849 0.757
Column Total	1911 0.211	2939 0.325	4199 0.464	9049

K = 9:

```
                       | adult.norm.test.pred
adult.norm.test.labels |           1 |           2 |           3 | Row Total |
-----------------------|-------------|-------------|-------------|-----------|
                     1 |           1 |        2932 |           0 |      2933 |
                       |       0.000 |       1.000 |       0.000 |     0.324 |
                       |       0.001 |       0.997 |       0.000 |           |
                       |       0.000 |       0.324 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-----------|
                     2 |        1907 |          10 |           0 |      1917 |
                       |       0.995 |       0.005 |       0.000 |     0.212 |
                       |       0.999 |       0.003 |       0.000 |           |
                       |       0.211 |       0.001 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-----------|
                     3 |           0 |           0 |        4199 |      4199 |
                       |       0.000 |       0.000 |       1.000 |     0.464 |
                       |       0.000 |       0.000 |       1.000 |           |
                       |       0.000 |       0.000 |       0.464 |           |
-----------------------|-------------|-------------|-------------|-----------|
          Column Total |        1908 |        2942 |        4199 |      9049 |
                       |       0.211 |       0.325 |       0.464 |           |
-----------------------|-------------|-------------|-------------|-----------|
```

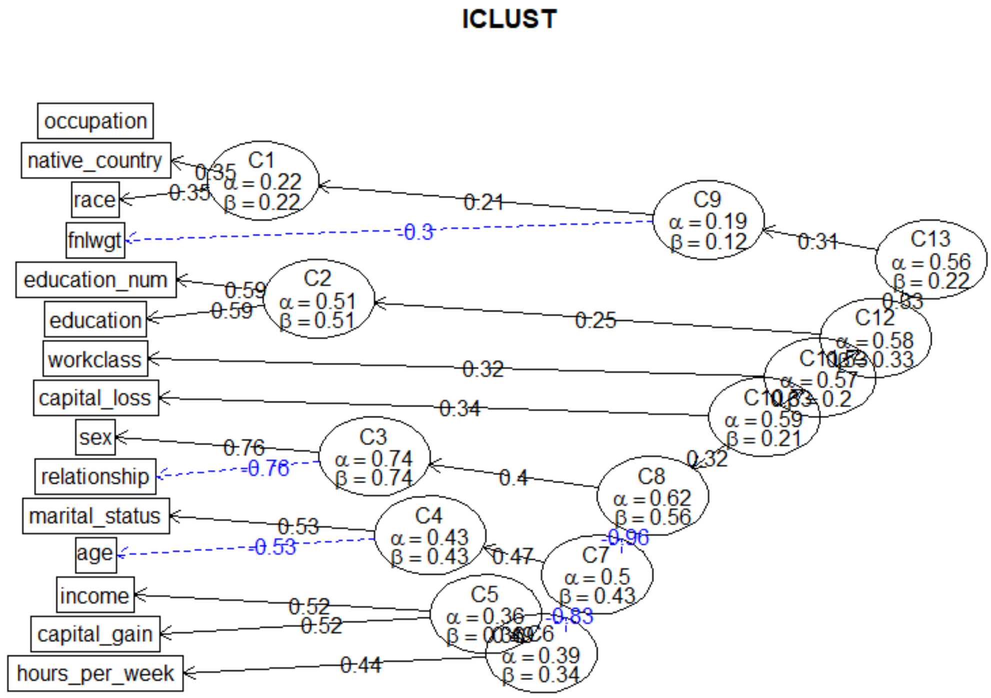
K = 10:

```
                       | adult.norm.test.pred
adult.norm.test.labels |           1 |           2 |           3 | Row Total |
-----------------------|-------------|-------------|-------------|-----------|
                     1 |        1907 |          10 |           0 |      1917 |
                       |       0.995 |       0.005 |       0.000 |     0.212 |
                       |       0.999 |       0.003 |       0.000 |           |
                       |       0.211 |       0.001 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-----------|
                     2 |           0 |           0 |        4199 |      4199 |
                       |       0.000 |       0.000 |       1.000 |     0.464 |
                       |       0.000 |       0.000 |       1.000 |           |
                       |       0.000 |       0.000 |       0.464 |           |
-----------------------|-------------|-------------|-------------|-----------|
                     3 |           2 |        2931 |           0 |      2933 |
                       |       0.001 |       0.999 |       0.000 |     0.324 |
                       |       0.001 |       0.997 |       0.000 |           |
                       |       0.000 |       0.324 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-----------|
          Column Total |        1909 |        2941 |        4199 |      9049 |
                       |       0.211 |       0.325 |       0.464 |           |
-----------------------|-------------|-------------|-------------|-----------|
```

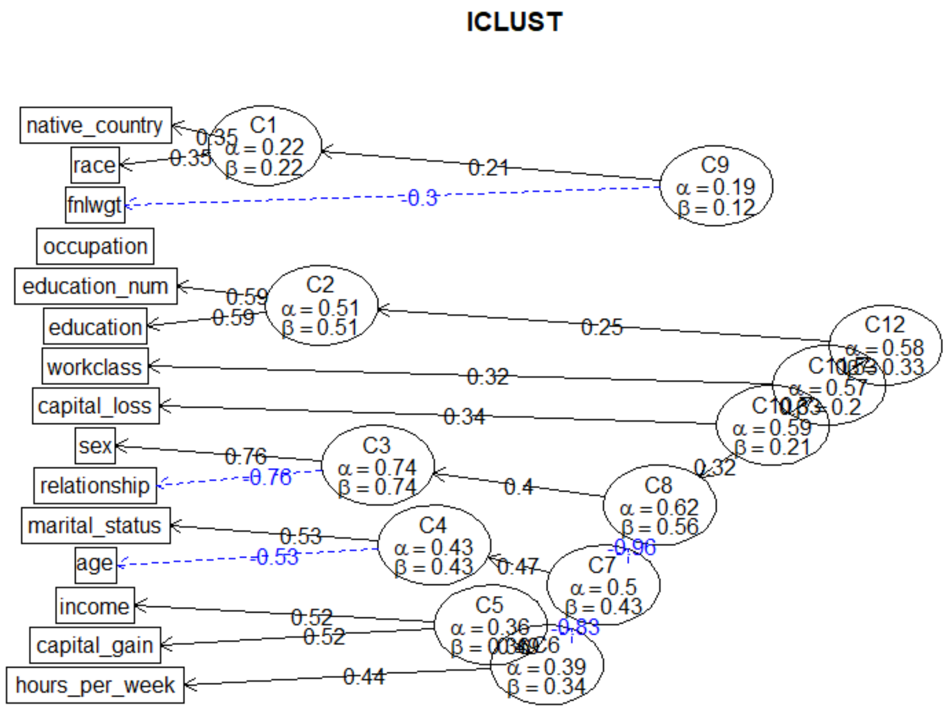
After we make k = 5 and let k increase, the prediction of clusters is getting more accurate. Although some accuracy is over 99%, some predictions label almost the whole cluster as another one, which means we need more clusters or some attributes are highly related.

3) iClust

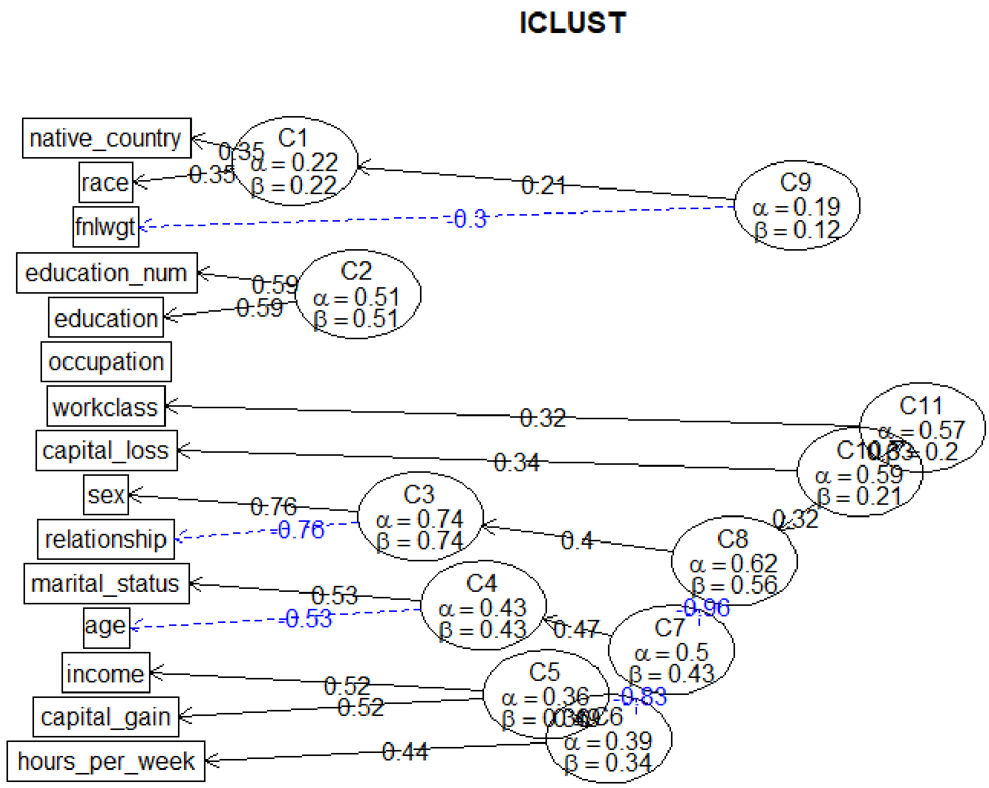
a) nclusters=2



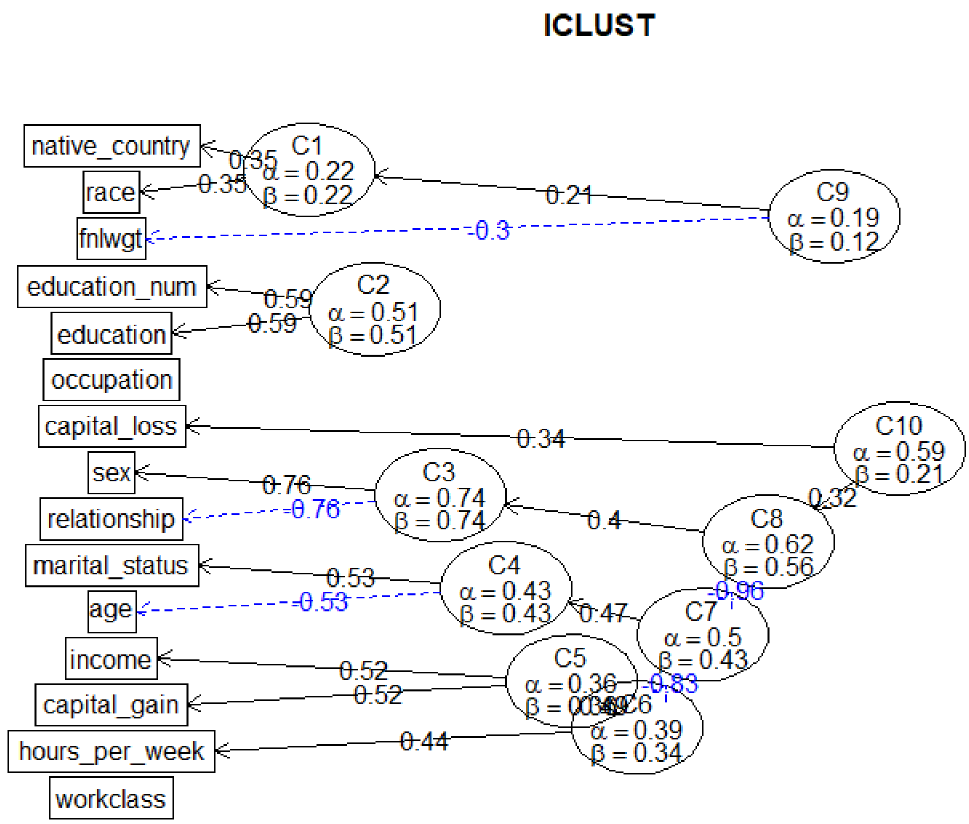
b) nclusters=3



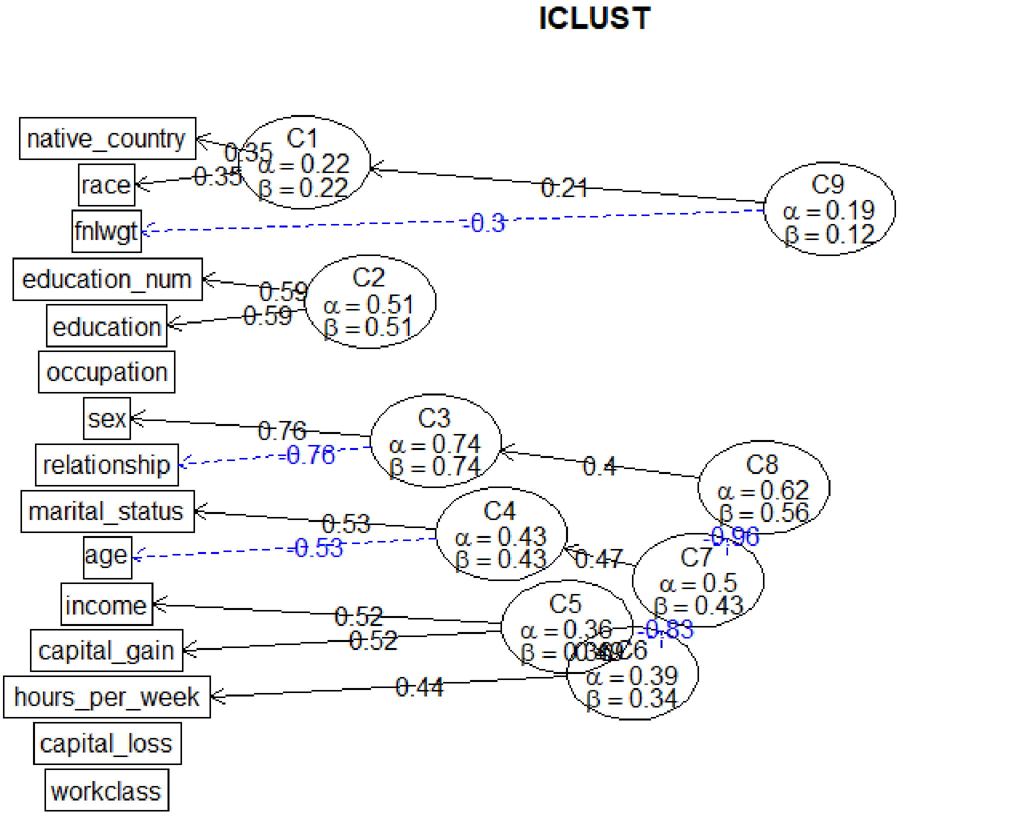
c) nclusters=4



d) nclusters=5

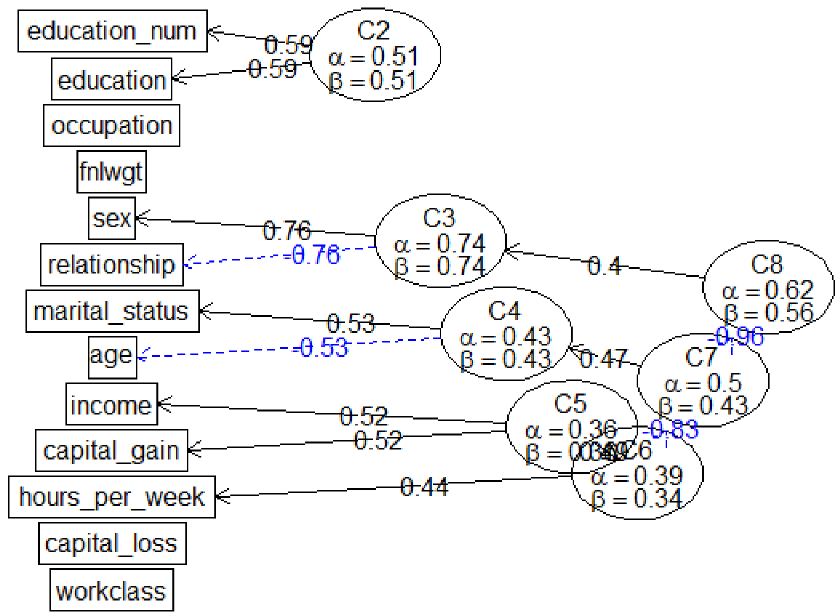


e) nclusters=6

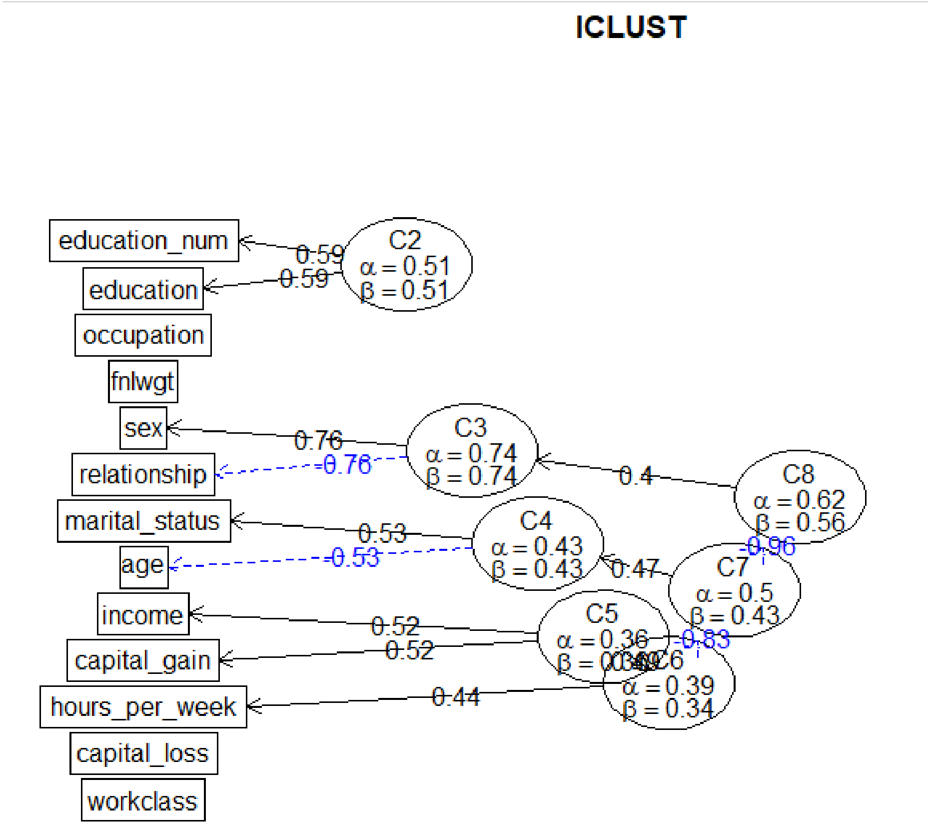


f) nclusters=7

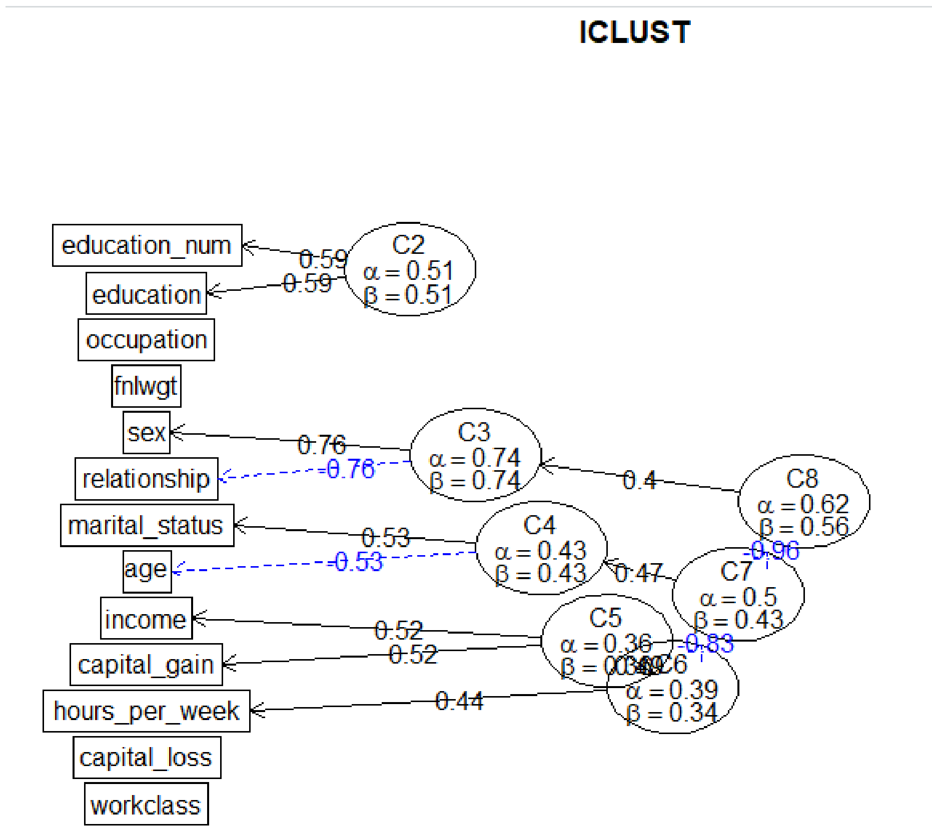
ICLUST



g) nclusters=8

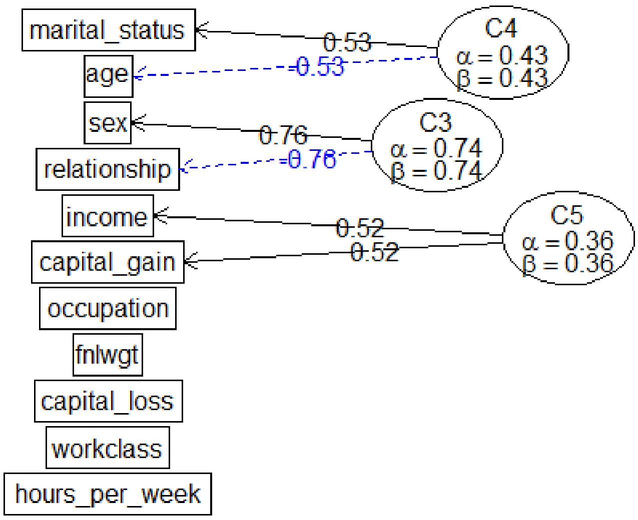


h) nclusters=9



i) nclusters=10

ICLUST



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.

4. Prediction

(a) Knn for prediction

1) 50-50

```
                      | adult.norm.test.pred
adult.norm.test.labels |           1 |           2 |           3 |           4 | Row Total |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     1 |           0 |        3306 |         971 |           0 |      4277 |
                       |       0.000 |       0.773 |       0.227 |       0.000 |     0.284 |
                       |       0.000 |       0.571 |       0.793 |       0.000 |           |
                       |       0.000 |       0.219 |       0.064 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     2 |           0 |        2485 |         253 |           0 |      2738 |
                       |       0.000 |       0.908 |       0.092 |       0.000 |     0.182 |
                       |       0.000 |       0.429 |       0.207 |       0.000 |           |
                       |       0.000 |       0.165 |       0.017 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     3 |           7 |           0 |           0 |        4835 |      4842 |
                       |       0.001 |       0.000 |       0.000 |       0.999 |     0.321 |
                       |       0.002 |       0.000 |       0.000 |       0.999 |           |
                       |       0.000 |       0.000 |       0.000 |       0.321 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     4 |        3219 |           0 |           0 |           5 |      3224 |
                       |       0.998 |       0.000 |       0.000 |       0.002 |     0.214 |
                       |       0.998 |       0.000 |       0.000 |       0.001 |           |
                       |       0.213 |       0.000 |       0.000 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
          Column Total |        3226 |        5791 |        1224 |        4840 |     15081 |
                       |       0.214 |       0.384 |       0.081 |       0.321 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
```

2) 60-40

```
                      | adult.norm.test.pred
adult.norm.test.labels |           1 |           2 |           3 |           4 | Row Total |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     1 |           0 |           6 |           2 |        3873 |      3881 |
                       |       0.000 |       0.002 |       0.001 |       0.998 |     0.322 |
                       |       0.000 |       0.002 |       0.040 |       0.998 |           |
                       |       0.000 |       0.000 |       0.000 |       0.321 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     2 |           0 |        2550 |          48 |           8 |      2606 |
                       |       0.000 |       0.979 |       0.018 |       0.003 |     0.216 |
                       |       0.000 |       0.998 |       0.960 |       0.002 |           |
                       |       0.000 |       0.211 |       0.004 |       0.001 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     3 |        3556 |           0 |           0 |           0 |      3556 |
                       |       1.000 |       0.000 |       0.000 |       0.000 |     0.295 |
                       |       0.638 |       0.000 |       0.000 |       0.000 |           |
                       |       0.295 |       0.000 |       0.000 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     4 |        2022 |           0 |           0 |           0 |      2022 |
                       |       1.000 |       0.000 |       0.000 |       0.000 |     0.168 |
                       |       0.362 |       0.000 |       0.000 |       0.000 |           |
                       |       0.168 |       0.000 |       0.000 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
          Column Total |        5578 |        2556 |          50 |        3881 |     12065 |
                       |       0.462 |       0.212 |       0.004 |       0.322 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
```

3) 70-30

```
                       | adult.norm.test.pred
adult.norm.test.labels |           1 |           2 |           3 |           4 | Row Total |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     1 |        1744 |           0 |           0 |           7 |      1751 |
                       |       0.996 |       0.000 |       0.000 |       0.004 |     0.194 |
                       |       0.979 |       0.000 |       0.000 |       0.003 |           |
                       |       0.193 |       0.000 |       0.000 |       0.001 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     2 |          38 |           0 |           0 |        2455 |      2493 |
                       |       0.015 |       0.000 |       0.000 |       0.985 |     0.276 |
                       |       0.021 |       0.000 |       0.000 |       0.997 |           |
                       |       0.004 |       0.000 |       0.000 |       0.271 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     3 |           0 |        2881 |           5 |           0 |      2886 |
                       |       0.000 |       0.998 |       0.002 |       0.000 |     0.319 |
                       |       0.000 |       0.999 |       0.003 |       0.000 |           |
                       |       0.000 |       0.318 |       0.001 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
                     4 |           0 |           3 |        1916 |           0 |      1919 |
                       |       0.000 |       0.002 |       0.998 |       0.000 |     0.212 |
                       |       0.000 |       0.001 |       0.997 |       0.000 |           |
                       |       0.000 |       0.000 |       0.212 |       0.000 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
          Column Total |        1782 |        2884 |        1921 |        2462 |      9049 |
                       |       0.197 |       0.319 |       0.212 |       0.272 |           |
-----------------------|-------------|-------------|-------------|-------------|-----------|
```

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

1) 50-50

1. Using all the attributes:

```
> adult.norm.train.lm<-lm(formula = adult.norm.train$income~adult.norm.train$fnlwgt+adult.norm.train$education+
+                             adult.norm.train$education_num+adult.norm.train$marital_status+adult.norm.train$occupation+
+                             adult.norm.train$relationship+adult.norm.train$race+adult.norm.train$sex+adult.norm.train$capital_gain+
+                             adult.norm.train$capital_loss+adult.norm.train$hours_per_week+adult.norm.train$native_country,
+                             data = adult.norm.train[1:14])
> adult.na.lm.pred<-predict.lm(adult.norm.train.lm)
> summary(adult.na.lm.pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4988  1.1104  1.2532  1.2509  1.3657  2.7190 
```

2. Using summary to check the probability of coefficient:

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.690973   0.029819  23.173  < 2e-16 ***
adult.norm.train$fnlwgt          0.052383   0.043045   1.217  0.22365    
adult.norm.train$education      -0.058927   0.013031  -4.522 6.17e-06 ***
adult.norm.train$education_num   0.736592   0.019947  36.927  < 2e-16 ***
adult.norm.train$marital_status -0.204657   0.012650 -16.179  < 2e-16 ***
adult.norm.train$occupation      0.023237   0.010051   2.312  0.02079 *  
adult.norm.train$relationship   -0.120056   0.012154  -9.878  < 2e-16 ***
adult.norm.train$race            0.048492   0.015032   3.226  0.00126 ** 
adult.norm.train$sex             0.102710   0.008225  12.488  < 2e-16 ***
adult.norm.train$capital_gain    0.997146   0.042475  23.476  < 2e-16 ***
adult.norm.train$capital_loss    0.525539   0.033663  15.612  < 2e-16 ***
adult.norm.train$hours_per_week  0.320451   0.026585  12.054  < 2e-16 ***
adult.norm.train$native_country -0.008689   0.020673  -0.420  0.67427    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
> adult.norm.train.lm<-lm(formula = adult.norm.train$income~adult.norm.train$education+
+                              adult.norm.train$education_num+adult.norm.train$marital_status+adult.norm.train$occupation+
+                              adult.norm.train$relationship+adult.norm.train$race+adult.norm.train$sex+adult.norm.train$capital_gain+
+                              adult.norm.train$capital_loss+adult.norm.train$hours_per_week,
+                            data = adult.norm.train[1:14])
> adult.na.lm.pred<-predict.lm(adult.norm.train.lm)
> summary(adult.na.lm.pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4993  1.1097  1.2530  1.2509  1.3650  2.7075 
```

4.After removing all discrete attributes:

```
> adult.norm.train.lm<-lm(formula = adult.norm.train$income~adult.norm.train$education_num+
+                              +adult.norm.train$capital_gain+
+                              adult.norm.train$capital_loss+adult.norm.train$hours_per_week,
+                            data = adult.norm.train[1:14])
> adult.na.lm.pred<-predict.lm(adult.norm.train.lm)
> summary(adult.na.lm.pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.6113  1.1679  1.2160  1.2509  1.3602  2.7354 
```

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

```
> adult.norm.train.glm<-glm(formula = adult.norm.train$income~adult.norm.train$education_num+
+                              +adult.norm.train$capital_gain+adult.norm.train$marital_status+adult.norm.train$sex
+                              +adult.norm.train$hours_per_week+adult.norm.train$sex+adult.norm.train$relationship,
+                            data = adult.norm.train[1:14])
> adult.na.glm.pred<-predict.lm(adult.norm.train.glm)
> summary(adult.na.glm.pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.5102  1.1148  1.2669  1.2471  1.3646  2.6791 
```

This is what the labels should be:

```
> summary(adult.norm.train$income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   1.000   1.247   1.000   2.000 
```

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

2)60-40

```
> adult.norm.train.glm<-glm(formula = adult.norm.train$income~adult.norm.train$education_num+
+                            +adult.norm.train$capital_gain+adult.norm.train$marital_status+adult.norm.train$sex
+                           +adult.norm.train$hours_per_week+adult.norm.train$sex+adult.norm.train$relationship,
+                           data = adult.norm.train[1:14])
> adult.na.glm.pred<-predict.glm(adult.norm.train.glm)
> summary(adult.na.glm.pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5297  1.1192  1.2718  1.2502  1.3685  2.6469
```

3)

```
> adult.norm.train.glm<-glm(formula = adult.norm.train$income~adult.norm.train$education_num+
+                            +adult.norm.train$capital_gain+adult.norm.train$marital_status+adult.norm.train$sex
+                           +adult.norm.train$hours_per_week+adult.norm.train$sex+adult.norm.train$relationship,
+                           data = adult.norm.train[1:14])
> adult.na.glm.pred<-predict.glm(adult.norm.train.glm)
> summary(adult.na.glm.pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5318  1.1176  1.2686  1.2497  1.3694  2.6756
```

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.