STA 380 Homework 1: WANG,YUWEN

Yuwen WANG

Wednesday, August 05, 2015

1. Exploratory Analysis

```
## Source: local data frame [4 x 2]
## 
##     equip avgrate
## 1   LEVER   0.012
## 2 OPTICAL   0.019
## 3   PAPER   0.009
## 4   PUNCH   0.017
```

We notice the undercount rates of different equipments are not the same, the undercount rate of 'PAPER' is the lowest, roughly one-half that of "OPTICAL" or "PUNCH".

```
## Source: local data frame [7 x 3]
## Groups: poor
## 
##   poor   equip  n
## 1    0   LEVER 29
## 2    0 OPTICAL 48
## 3    0   PUNCH 10
## 4    1   LEVER 45
## 5    1 OPTICAL 18
## 6    1   PAPER  2
## 7    1   PUNCH  7
```

At a brief glance, we notice that the number of poor counties using 'OPTICAL', which has the highest undercount rate, is about 1/3 the number of rich counties. In addition, rich communities have absolutely no access to 'PAPER', which has the lowest undercount ratio, at all.

```
## Source: local data frame [3 x 3]
## 
##     equip  n frac
## 1   LEVER 29 0.33
## 2 OPTICAL 48 0.55
## 3   PUNCH 10 0.11
```

```
## Source: local data frame [4 x 3]
## 
##     equip  n frac
## 1   LEVER 45 0.62
## 2 OPTICAL 18 0.25
## 3   PAPER  2 0.03
## 4   PUNCH  7 0.10
```

Breaking it down a little further, we notice that 55% of rich people have access to 'OPTICAL', the equipment with the highest undercount rate, while in contrast 'LEVEL' is used by the majority of poor group, which has a moderate undercount rate.

```
## Source: local data frame [2 x 2]
## 
##   poor avgrate
## 1    0    0.02
## 2    1    0.01
```

Further confirmed by the fact that rich communities have a undercount rate 2 times that of poor communities, we can therefore conclde that access to different equipment has a disparate impact on poor communities.

We can apply similar analysis to minority communities by looking at "perAA". I split "perAA" at 25% quantile and made it an indicator variable like "poor", and came to a similar conclusion that access to different equipment has a disparate impact on minority communities.

```
## Source: local data frame [4 x 3]
## 
##     equip  n frac
## 1   LEVER 61 0.51
## 2 OPTICAL 42 0.35
## 3   PAPER  2 0.02
## 4   PUNCH 14 0.12
```

```
## Source: local data frame [3 x 3]
## 
##     equip  n frac
## 1   LEVER 13 0.32
## 2 OPTICAL 24 0.60
## 3   PUNCH  3 0.08
```

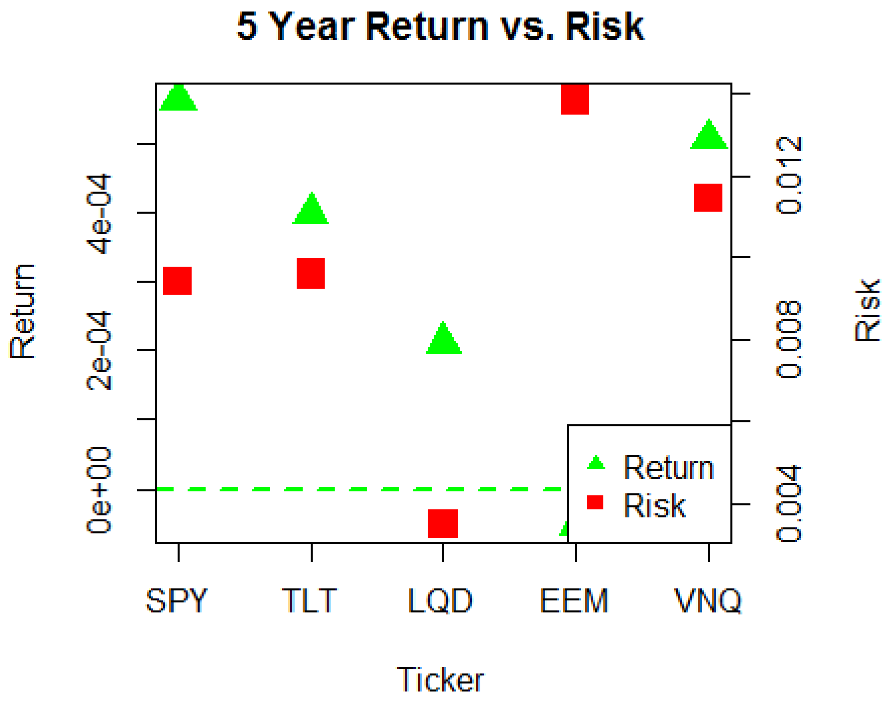
```
## Source: local data frame [2 x 2]
## 
##      AA avgrate
## 1 FALSE   0.014
## 2  TRUE   0.021
```

2. Portfolios

Building Portfolios

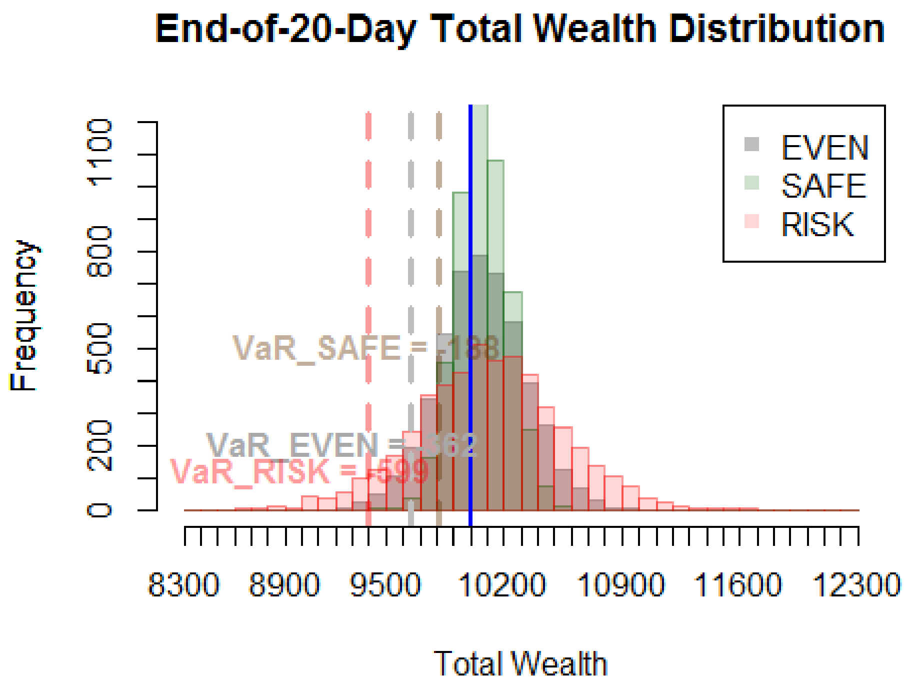
The following graph shows the average daily return and risk(measured by standard deviation of daily returns) of 5 assets in the past 5 years.

For a safer investment, I choose to invest 50% in LQD, 30% in SPY, and 20% in TLT. For a riskier investment, I choose to invest 60% in SPY and 40% in VNQ.



Simulation

The (AWESOME) histogram below shows the distribution of total wealth if we hold each of the 3 portfolios for 20 days. As a risk-reversion person, I would recomment the safer investment(50%LQD, 30% SPY, 20% TLT), with the lowest VaR(5% level) and highest probability of earning money.



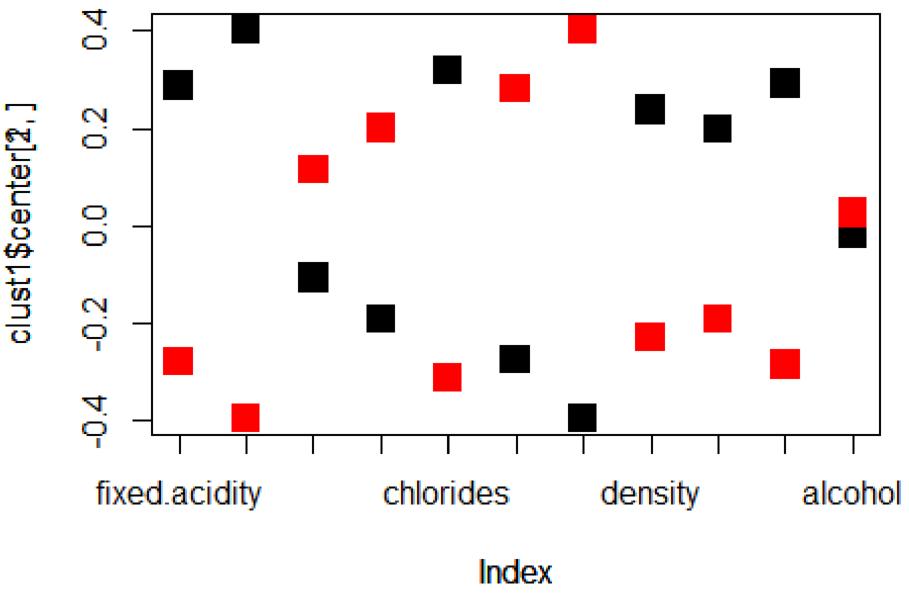
```
##              VaR pct_profitable
## EVEN -355.2738         0.5976
## RISK -602.9304         0.6098
## SAFE -182.8684         0.6706
```

3. Cluster

3.1 Red vs. White

First, I use all 11 chemicals to run a k-means clustering, separating all the records into two groups. From the tabulation result we can see that unsupervised learning can do a very good job to separate red wine and white wine.

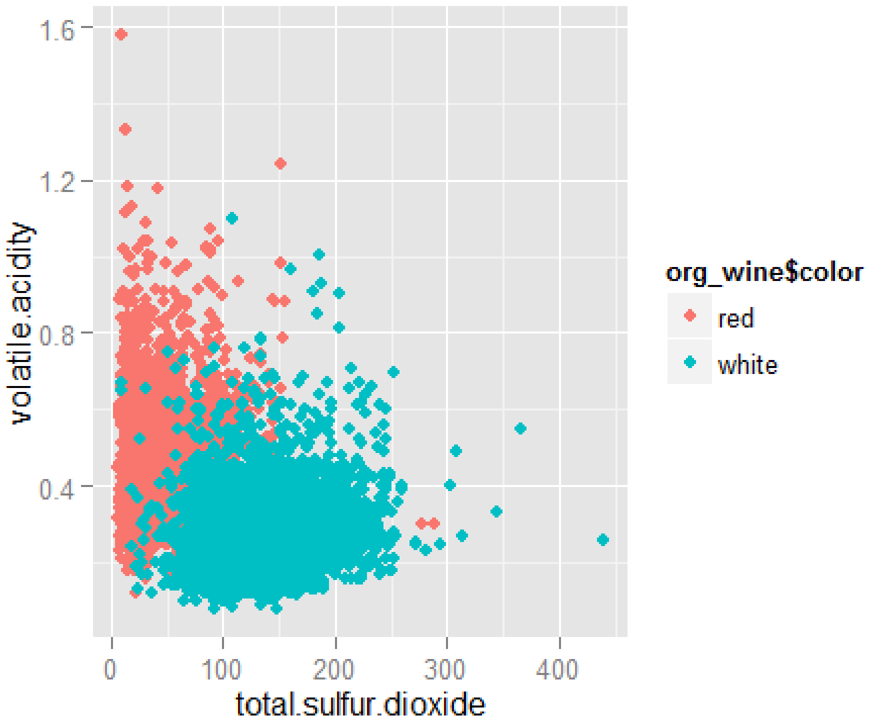
```
## color1
##   red white
##  1575    68

## color2
##   red white
##    24  4830
```



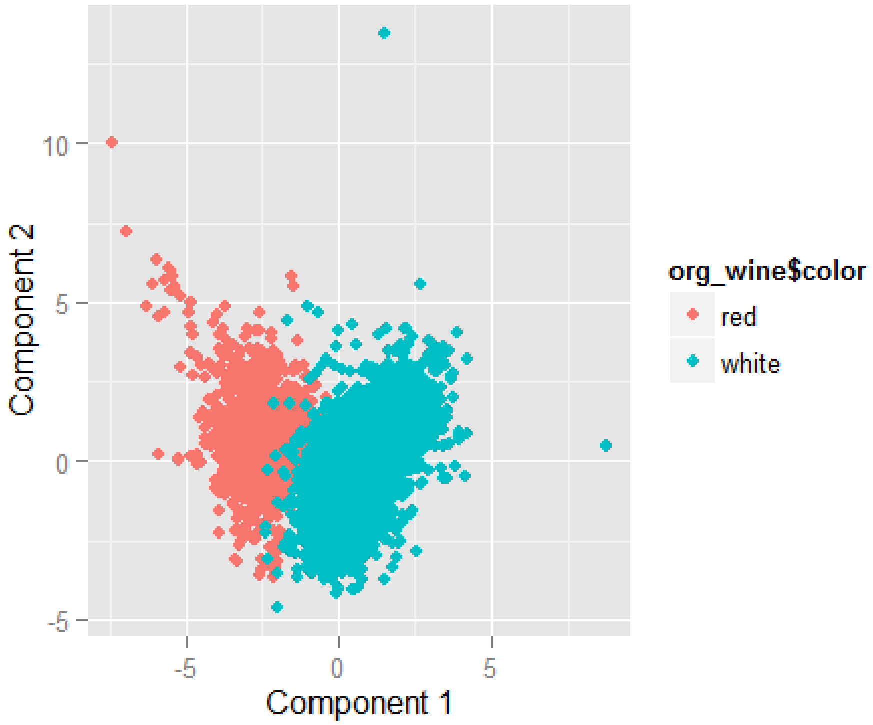
By plotting the two centers for each attribute, we notice that attribute alcohol is hardly distinguishable between two clusters. So I remove alcohol and redo a k-means clustering. The separating result is still satisfying (excatly the same, in fact).

```
##        fixed.acidity     volatile.acidity          citric.acid 
##           6.85145062           0.27479835           0.33513786 
##       residual.sugar            chlorides  free.sulfur.dioxide 
##           6.39407407           0.04518004          35.50874486 
## total.sulfur.dioxide              density                   pH 
##         138.48837449           0.99400769           3.18777366 
##            sulphates 
##           0.48897119 

##        fixed.acidity     volatile.acidity          citric.acid 
##           8.29554062           0.53224801           0.26963348 
##       residual.sugar            chlorides  free.sulfur.dioxide 
##           2.62034209           0.08825718          15.73029933 
## total.sulfur.dioxide              density                   pH 
##          48.22174710           0.99674202           3.30972511 
##            sulphates 
##           0.65684178 

## color1
##   red white 
##    24  4836 

## color2
##   red white 
##  1575    62 
```

We plot the two clusters agiainst the two most distinguishable attributes. We can see these two factors alone can separate the observations very well.



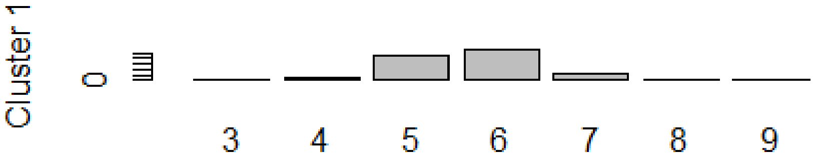
We can also use principal component analysis. From the graph below, we can see that the first component almost suffices to tell the white wine from the red wine.

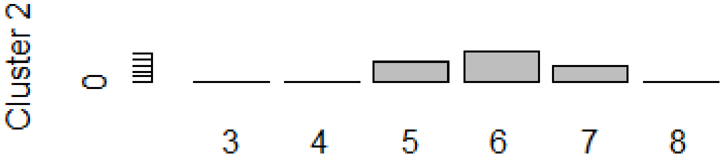


In this case, both clustering and PCA are doing a great job distinguishing red wine and white wine. However, I would recommend clustering over PCA, since with a similar result, not only has clustering got rid of an attribute but it is much easier to interpret than pca as well.

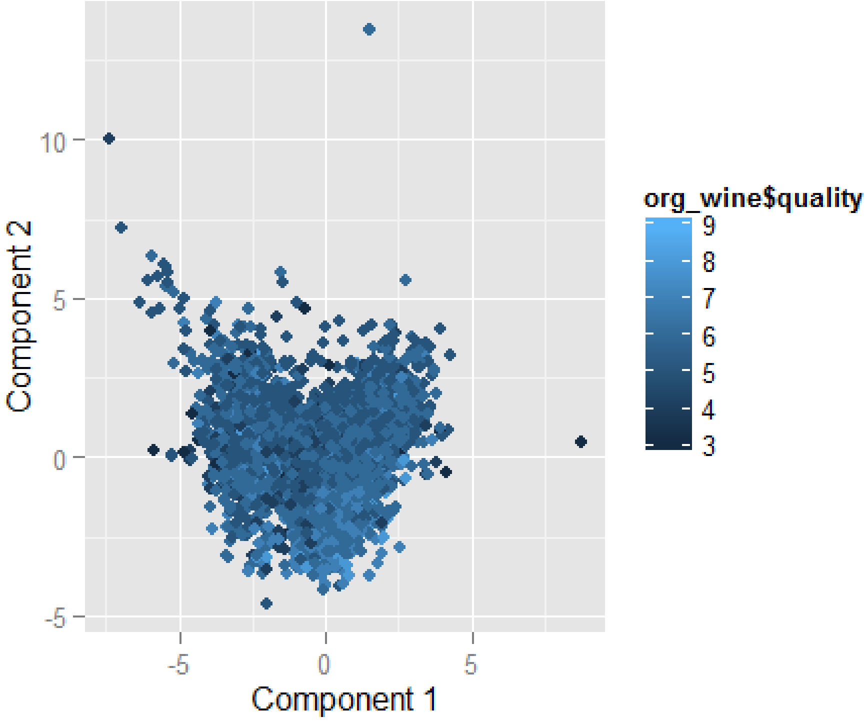
3.2 Quality

However, clusering doesn't work well for telling different quality apart. Since there're 7 levels of quality (3 to 9), I specify the number of clusters to be 7. However, if we contrast any two of the clusters, e.g. cluster 1 and cluster 6, they give a similar distribution of wine quality as shown in the bar plot below



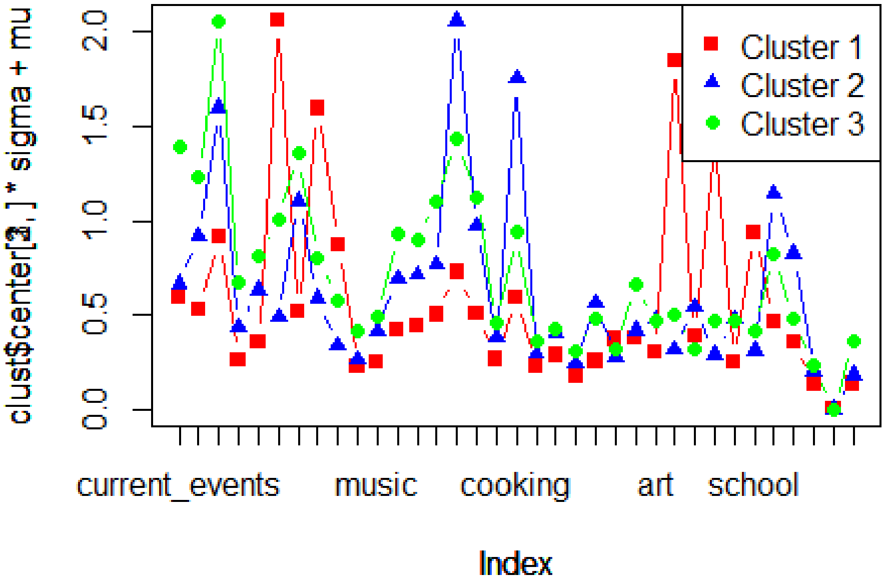


PCA does not working either.

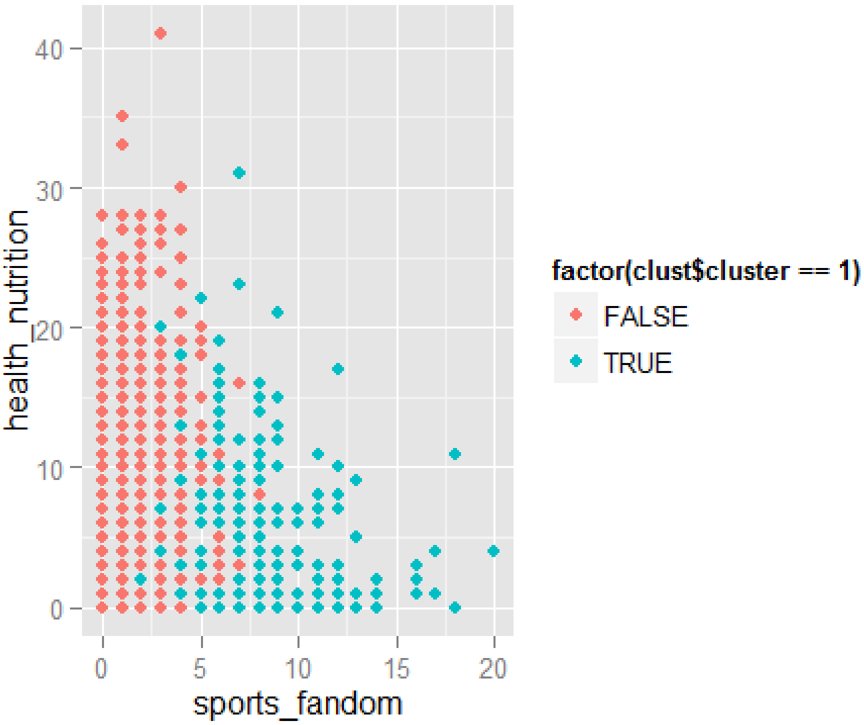


4. Market Segmentation

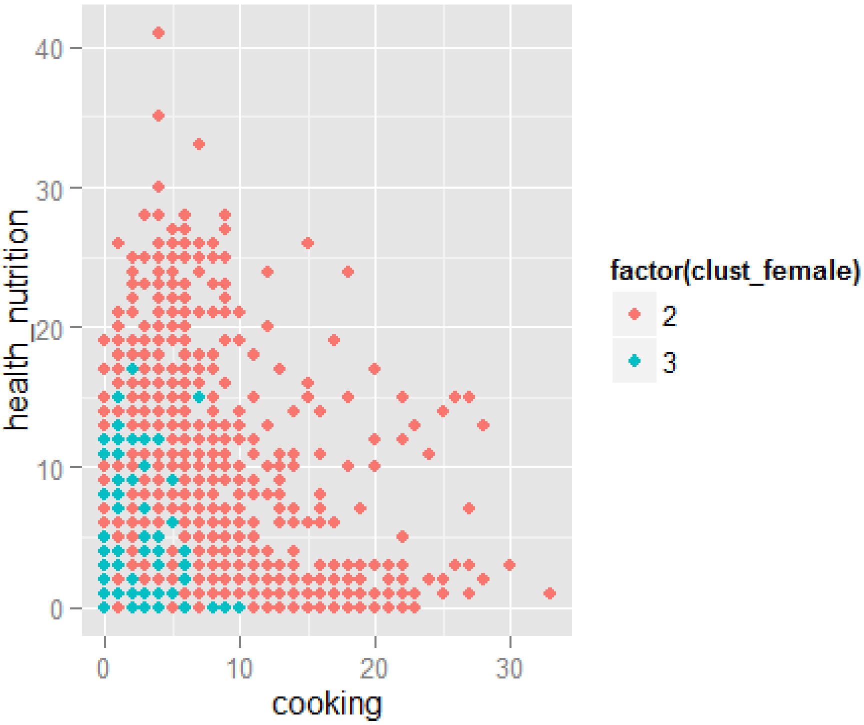
I first tried a clustering method with 3 clusters. The following graph shows the the 3 cluster centers broken down to each attribute. Taking a closer look at these e clusters, I could already picture three different user groups, so I sticked with 3 clusters. I noticed the 1st cluster(in red) showed significantly different behaviors from the 2nd and 3rd cluster(in green and blue).



The 1st cluster shows strong inclination to key words "sports_fandom, family, food, parenting, school,religion" etc., very clearly picturing a young father, while for the other two groups with rather girlish key words "photo sharing, shopping, health nutrition, cooking, personal fitness, fashion" are very likely to be female users. Plotting the clustering against two most distinguishable factors sports fandom and health nutrition, we can see the 1st cluster are well separated from the 2nd and 3rd.



Between the 2nd and 3rd clusters, the most distinguishable differences are the 3rd cluster cared more about cooking, health &nutrition, and fitness, while the 2nd twitted more about photo sharing, shopping. These clearly indicate that the 2nd cluster are younger females, very likely school girls compared to a more matured user group in the 3rd cluster. Plotting against the two most distinguishable attributes health nutrition and cooking, we can see that the 3rd cluster are centered at the bottom left corner.


health_nutrition
cooking
factor(clust_female)