Independent Project Week 14- Feature Selection

Billiah

2022-04-04

```
library (caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(clustvarsel)
```

```
## Loading required package: mclust
```

```
## Package 'mclust' version 5.4.9
## Type 'citation("mclust")' for citing this R package in publications.
```

```
## Package 'clustvarsel' version 2.3.4
```

```
## Type 'citation("clustvarsel")' for citing this R package in publications.
```

```
library(mclust)
library(wskm)
```

```
## Loading required package: latticeExtra
```

```
##
## Attaching package: 'latticeExtra'
```

```
## The following object is masked from 'package:ggplot2':
##
##     layer
```

```
## Loading required package: fpc
```

```r
library("cluster")
```

```r
path<-"http://bit.ly/CarreFourDataset"
data <- read.csv(path)
head(data)
```

```
##    Invoice.ID Branch Customer.type Gender           Product.line Unit.price
## 1 750-67-8428      A        Member Female      Health and beauty      74.69
## 2 226-31-3081      C        Normal Female Electronic accessories      15.28
## 3 631-41-3108      A        Normal   Male     Home and lifestyle      46.33
## 4 123-19-1176      A        Member   Male      Health and beauty      58.22
## 5 373-73-7910      A        Normal   Male      Sports and travel      86.31
## 6 699-14-3026      C        Normal   Male Electronic accessories      85.39
##   Quantity     Tax      Date  Time     Payment   cogs gross.margin.percentage
## 1        7 26.1415  1/5/2019 13:08     Ewallet 522.83                4.761905
## 2        5  3.8200  3/8/2019 10:29        Cash  76.40                4.761905
## 3        7 16.2155  3/3/2019 13:23 Credit card 324.31                4.761905
## 4        8 23.2880 1/27/2019 20:33     Ewallet 465.76                4.761905
## 5        7 30.2085  2/8/2019 10:37     Ewallet 604.17                4.761905
## 6        7 29.8865 3/25/2019 18:30     Ewallet 597.73                4.761905
##   gross.income Rating    Total
## 1      26.1415    9.1 548.9715
## 2       3.8200    9.6  80.2200
## 3      16.2155    7.4 340.5255
## 4      23.2880    8.4 489.0480
## 5      30.2085    5.3 634.3785
## 6      29.8865    4.1 627.6165
```

```r
dim(data)
```

```
## [1] 1000   16
```

The dataset has 1000 rows and 16 columns.

```r
str(data)
```

```
## 'data.frame':    1000 obs. of  16 variables:
##  $ Invoice.ID             : chr  "750-67-8428" "226-31-3081" "631-41-3108" "123-19-1176" ...
##  $ Branch                 : chr  "A" "C" "A" "A" ...
##  $ Customer.type          : chr  "Member" "Normal" "Normal" "Member" ...
##  $ Gender                 : chr  "Female" "Female" "Male" "Male" ...
##  $ Product.line           : chr  "Health and beauty" "Electronic accessories" "Home and lifestyle" "I
##  $ Unit.price             : num  74.7 15.3 46.3 58.2 86.3 ...
##  $ Quantity               : int  7 5 7 8 7 7 6 10 2 3 ...
##  $ Tax                    : num  26.14 3.82 16.22 23.29 30.21 ...
##  $ Date                   : chr  "1/5/2019" "3/8/2019" "3/3/2019" "1/27/2019" ...
##  $ Time                   : chr  "13:08" "10:29" "13:23" "20:33" ...
##  $ Payment                : chr  "Ewallet" "Cash" "Credit card" "Ewallet" ...
##  $ cogs                   : num  522.8 76.4 324.3 465.8 604.2 ...
##  $ gross.margin.percentage: num  4.76 4.76 4.76 4.76 4.76 ...
##  $ gross.income           : num  26.14 3.82 16.22 23.29 30.21 ...
##  $ Rating                 : num  9.1 9.6 7.4 8.4 5.3 4.1 5.8 8 7.2 5.9 ...
##  $ Total                  : num  549 80.2 340.5 489 634.4 ...
```

```
sum(is.na(data))
```

```
## [1] 0
```

There are no null values.

```
sum(duplicated(data))
```

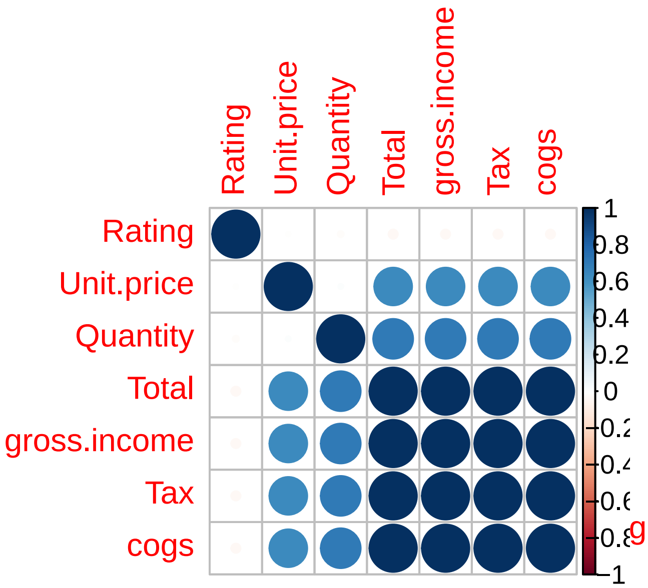
```
## [1] 0
```

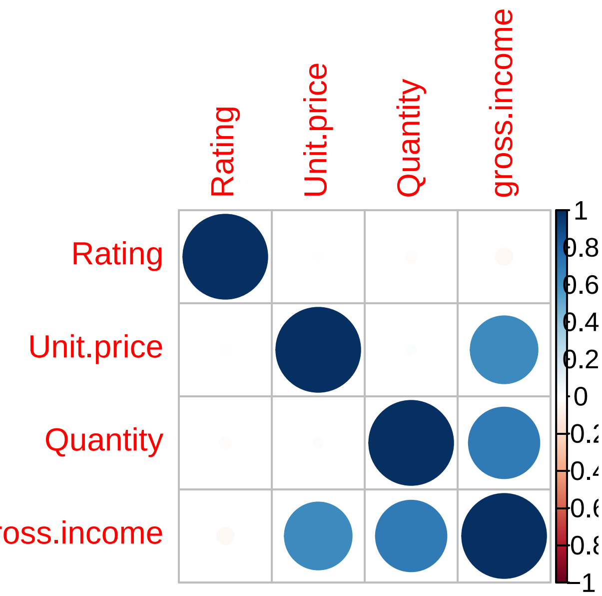
There are no duplicates in the dataset.

```
# Getting numeric columns
n_data <- data[c(6:8,12,14:16)]
# Finding correlation matrix
corr_m <- cor(n_data)
# Features that are highly correlated
h_corr <- findCorrelation(corr_m, cutoff=0.75)
names(n_data[,h_corr])
```

```
## [1] "cogs"  "Total" "Tax"
```

```
# Removing variables with high correlation
data_1<-n_data[-h_corr]
# Graphical comparison
par(mfrow = c(1, 2))
corrplot(corr_m, order = "hclust")
corrplot(cor(data_1), order = "hclust")
```





The highly correlated features(Tax and cogs) have been eliminated.

```r
# Sequential search
data_2 = clustvarsel(n_data, G = 1:5)
data_2
```

```
## ------------------------------------------------------
## Variable selection for Gaussian model-based clustering
## Stepwise (forward/backward) greedy search
## ------------------------------------------------------
##
##  Variable proposed Type of step   BICclust Model G     BICdiff Decision
##                Tax          Add  -7382.354     V 4    389.0238 Accepted
##       gross.income          Add  55117.386   VEV 3   2502.9883 Accepted
##           Quantity          Add -16164.602   VVI 5 -66967.5199 Rejected
##                Tax       Remove  -7392.222     V 3   2512.8564 Rejected
##
## Selected subset: Tax, gross.income
```

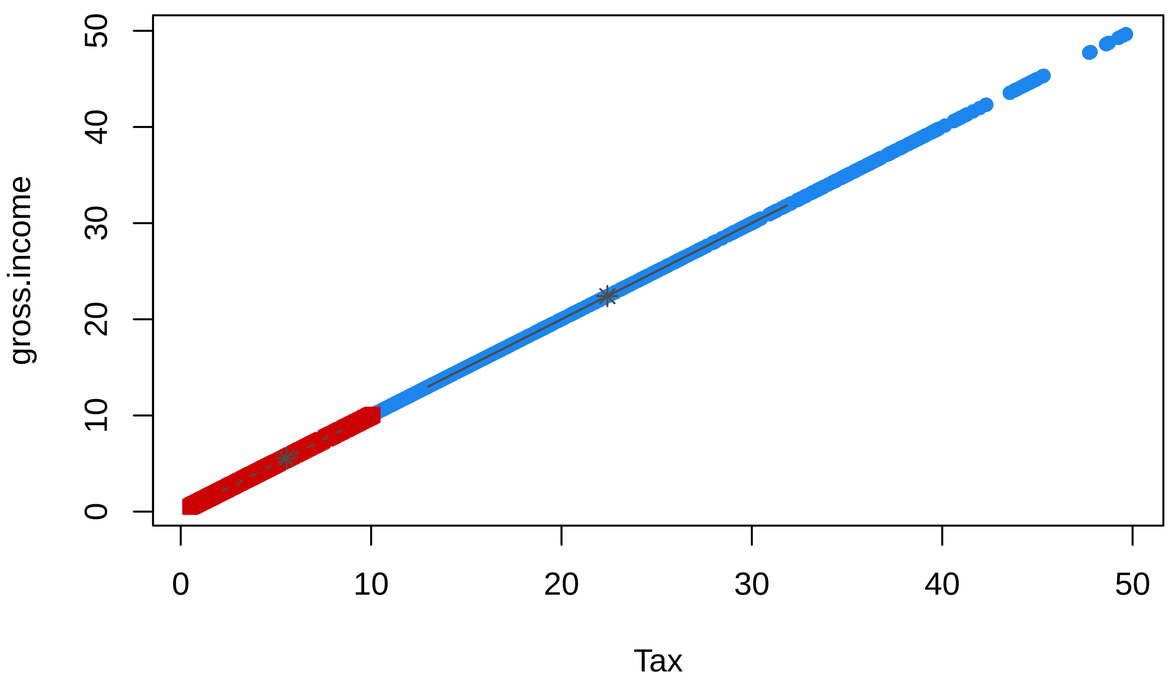
Gross income has been accepted, as well as tax. These are the optimal variables in this dataset.

```r
Subset1 = n_data[,data_2$subset]
mod = Mclust(Subset1, G = 1:5)
summary(mod)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VEV (ellipsoidal, equal shape) model with 2 components:
##
##  log-likelihood    n df      BIC      ICL
##        27364.17 1000 10 54659.26 54524.45
##
## Clustering table:
##   1   2
## 564 436
```

```r
# PLotting
plot(mod,c("classification"))
```

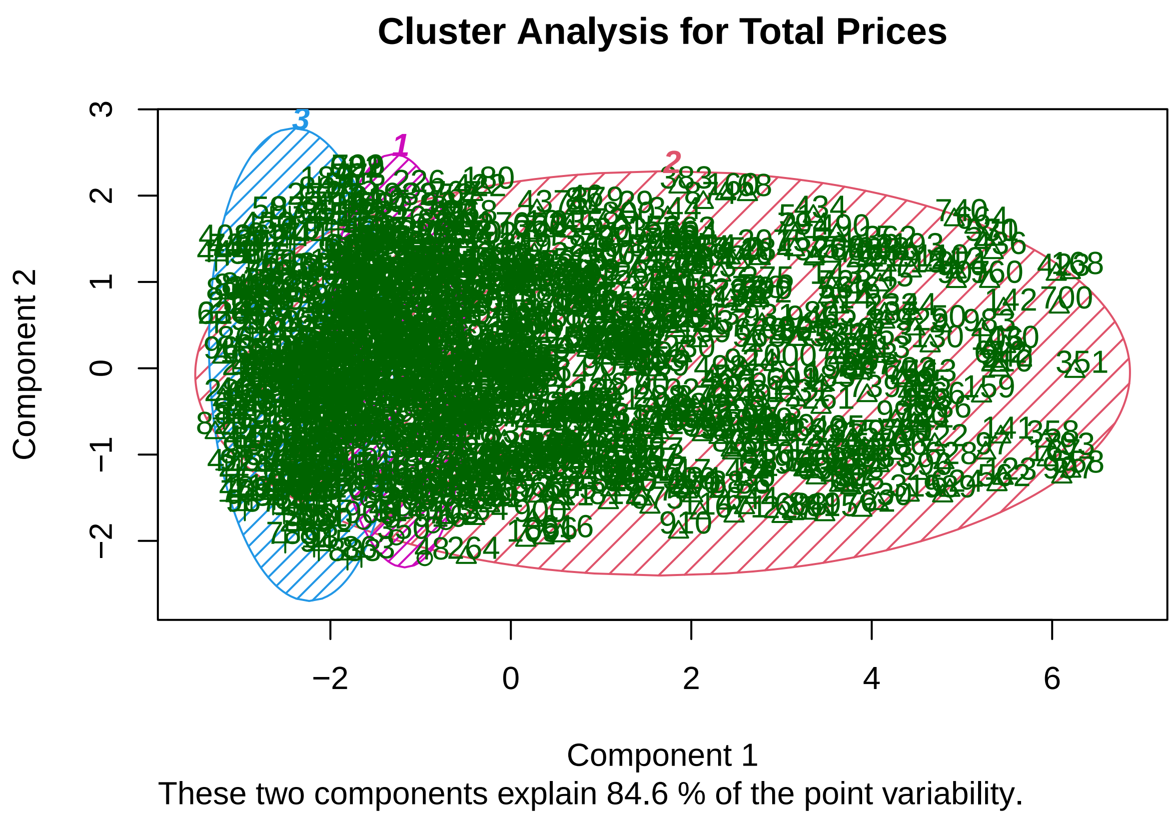
```
## Warning in sqrt(rev(sort(ev$values))): NaNs produced
```



The optimal features are gross income and tax, which have a linear correlation. An increase in gross income influences an increase in tax.

```r
# The ewkm function from the wskm package will be used.
# This is a weighted subspace clustering algorithm that is well suited to very high dimensional data
set.seed(2)
model <- ewkm(n_data,3, lambda=2, maxiter=1000)
```

```r
#  Cluster Plot against 1st 2 principal components
clusplot(n_data, model$cluster, color=TRUE, shade=TRUE,
         labels=2, lines=1,main='Cluster Analysis for Total Prices')
```



The two components cumulatively explain 84.6% variability in the data. Therefore, the two components capture alot of information in the data.

To measure importanc of each element, weight have to be calculated, incorporated in the distance function.

```r
# Checking for weights
round(model$weights*100,2)
```

```
##   Unit.price Quantity Tax cogs gross.income Rating Total
## 1          0        0  50    0           50   0.00     0
## 2          0        0   0    0            0  99.99     0
## 3          0        0  50    0           50   0.00     0
```

Tax has more weight in cluster 1 and 3, gross income has more weight in cluster 1 and 3. Rating has more weight in the second cluster.

CONCLUSION

Gross income plays an important role in the total value of items. It is an important variable in this dataset.

To increase the total prices, the gross income has to be evaluated first.