R Notebook

FEATURE SELECTION

1. DEFINING THE QUESTION

a) Specifying the Question

Performing feature selection and provide insights on the features that contribute the most information to the dataset.

b) Defining the Metrics of Success

To perform feature selection through the use of the unsupervised learning methods.

c) Understanding the context

You are a Data analyst at Carrefour Kenya and are currently undertaking a project that will inform the marketing department on the most relevant marketing strategies that will result in the highest no. of sales (total price including tax). Your project has been divided into four parts where you'll explore a recent marketing dataset by performing various unsupervised learning techniques and later providing recommendations based on your insights.

d) Recording the Experimental Design

1. Defining the question, the metric for success, the context and the experimental design.
2. Reading and exploring the dataset.
3. Performing feature selection and providing insights on the features that contribute the most information to the dataset.

e) Relevance of the data

The data used will inform the marketing department on the most relevant marketing strategies that will result in the highest number of sales and total price including tax. The dataset link: http://bit.ly/CarreFourDataset

2. DATA ANALYSIS

a) Checking the Data

```
# Loading libraries
library(relaimpo)
```

```
## Loading required package: MASS

## Loading required package: boot

## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##
## Attaching package: ’survival’

## The following object is masked from ’package:boot’:
##
##     aml

##
## Attaching package: ’survey’

## The following object is masked from ’package:graphics’:
##
##     dotchart

## Loading required package: mitools

## This is the global version of package relaimpo.

## If you are a non-US user, a version with the interesting additional metric pmvd is available

## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.
```

```
library(data.table)
library(ggplot2) # Data visualization
library(ggthemes) # Plot themes
library(plotly) # Interactive data visualizations
```

```
##
## Attaching package: ’plotly’

## The following object is masked from ’package:ggplot2’:
##
##     last_plot
## The following object is masked from ’package:MASS’:
##
##     select

## The following object is masked from ’package:stats’:
##
##     filter

## The following object is masked from ’package:graphics’:
##
##     layout
```

```r
library(dplyr) # Data manipulation
```

```
##
## Attaching package: ’dplyr’

## The following objects are masked from ’package:data.table’:
##
##     between, first, last

## The following object is masked from ’package:MASS’:
##
##     select

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```r
library(psych) # Will be used for correlation visualization
```

```
##
## Attaching package: ’psych’

## The following objects are masked from ’package:ggplot2’:
##
##     %+%, alpha

## The following object is masked from ’package:boot’:
##
##     logit
```

```r
# Importing the data
df <- fread('http://bit.ly/CarreFourDataset')
df
```

```
##        Invoice ID Branch Customer type Gender           Product line Unit price
##            <char> <char>        <char> <char>                 <char>      <num>
##    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
##   ---
##  996: 233-67-5758      C        Normal   Male      Health and beauty      40.35
##  997: 303-96-2227      B        Normal Female     Home and lifestyle      97.38
##  998: 727-02-1313      A        Member   Male     Food and beverages      31.84
##  999: 347-56-2442      A        Normal   Male     Home and lifestyle      65.82
## 1000: 849-09-3807      A        Member Female    Fashion accessories      88.34
##       Quantity     Tax      Date   Time     Payment   cogs
##          <int>   <num>    <char> <char>      <char>  <num>
##    1:        7 26.1415  1/5/2019  13:08     Ewallet 522.83
##    2:        5  3.8200  3/8/2019  10:29        Cash  76.40
##    3:        7 16.2155  3/3/2019  13:23 Credit card 324.31
##    4:        8 23.2880 1/27/2019  20:33     Ewallet 465.76
##    5:        7 30.2085  2/8/2019  10:37     Ewallet 604.17
##   ---
##  996:        1  2.0175 1/29/2019  13:46     Ewallet  40.35
##  997:       10 48.6900  3/2/2019  17:16     Ewallet 973.80
##  998:        1  1.5920  2/9/2019  13:22        Cash  31.84
##  999:        1  3.2910 2/22/2019  15:33        Cash  65.82
## 1000:        7 30.9190 2/18/2019  13:28        Cash 618.38
##       gross margin percentage gross income Rating     Total
##                         <num>        <num>  <num>     <num>
##    1:                4.761905      26.1415    9.1  548.9715
##    2:                4.761905       3.8200    9.6   80.2200
##    3:                4.761905      16.2155    7.4  340.5255
##    4:                4.761905      23.2880    8.4  489.0480
##    5:                4.761905      30.2085    5.3  634.3785
##   ---
##  996:                4.761905       2.0175    6.2   42.3675
##  997:                4.761905      48.6900    4.4 1022.4900
##  998:                4.761905       1.5920    7.7   33.4320
##  999:                4.761905       3.2910    4.1   69.1110
## 1000:                4.761905      30.9190    6.6  649.2990
```

b) Data Checking

```
# Previewing the dataset
View(df)
```

```
# Previewing the column names
colnames(df)
```

```
##  [1] "Invoice ID"              "Branch"
##  [3] "Customer type"           "Gender"
##  [5] "Product line"            "Unit price"
##  [7] "Quantity"                "Tax"
##  [9] "Date"                    "Time"
## [11] "Payment"                 "cogs"
## [13] "gross margin percentage" "gross income"
## [15] "Rating"                  "Total"
```

```
# Previewing the datatypes of the dataset
sapply(df, class)
```

```
##              Invoice ID                  Branch           Customer type
##             "character"             "character"             "character"
##                  Gender            Product line              Unit price
##             "character"             "character"               "numeric"
##                Quantity                     Tax                    Date
##               "integer"               "numeric"             "character"
##                    Time                 Payment                    cogs
##             "character"             "character"               "numeric"
## gross margin percentage            gross income                  Rating
##               "numeric"               "numeric"               "numeric"
##                   Total
##               "numeric"
```

```
# Previewing the head of the dataset
head(df, n = 5)
```

```
##     Invoice ID Branch Customer type Gender           Product line Unit price
##         <char> <char>        <char> <char>                 <char>      <num>
## 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
##    Quantity     Tax      Date   Time     Payment   cogs gross margin percentage
##       <int>   <num>    <char> <char>      <char>  <num>                   <num>
## 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
##    gross income Rating    Total
##           <num>  <num>    <num>
## 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
```

```
# Previewing the tail of the dataset
tail(df, n = 5)
```

```
##     Invoice ID Branch Customer type Gender        Product line Unit price
##         <char> <char>        <char> <char>              <char>      <num>
## 1: 233-67-5758      C        Normal   Male   Health and beauty      40.35
## 2: 303-96-2227      B        Normal Female  Home and lifestyle      97.38
## 3: 727-02-1313      A        Member   Male  Food and beverages      31.84
## 4: 347-56-2442      A        Normal   Male  Home and lifestyle      65.82
## 5: 849-09-3807      A        Member Female Fashion accessories      88.34
##    Quantity     Tax      Date   Time Payment   cogs gross margin percentage
##       <int>   <num>    <char> <char>  <char>  <num>                   <num>
## 1:        1  2.0175 1/29/2019  13:46 Ewallet  40.35                4.761905
## 2:       10 48.6900  3/2/2019  17:16 Ewallet 973.80                4.761905
## 3:        1  1.5920  2/9/2019  13:22    Cash  31.84                4.761905
## 4:        1  3.2910 2/22/2019  15:33    Cash  65.82                4.761905
## 5:        7 30.9190 2/18/2019  13:28    Cash 618.38                4.761905
##    gross income Rating     Total
##           <num>  <num>     <num>
## 1:       2.0175    6.2   42.3675
## 2:      48.6900    4.4 1022.4900
## 3:       1.5920    7.7   33.4320
## 4:       3.2910    4.1   69.1110
## 5:      30.9190    6.6  649.2990
```

```
# Checking the structure of the data
str(df)
```

```
## Classes ’data.table’ and ’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" "
##  $ 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 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
# Checking the shape of the data
dim(df)
```

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

1000 rows and 16 columns

c) Data Cleaning

Missing Values

```
# Checking for missing values
sum(is.na(df))
```

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

There are no missing values in the data

Duplicates

```
# Checking for duplicates
duplicated_rows <- df[duplicated(df),]
duplicated_rows
```

```
## Empty data.table (0 rows and 16 cols): Invoice ID,Branch,Customer type,Gender,Product line,Unit pric
```

There are no duplicates in the data

```
# Displaying unique items and assigning them to a variable unique_items below
unique_items <- df[!duplicated(df), ]
unique_items
```

```
##        Invoice ID Branch Customer type Gender           Product line Unit price
##            <char> <char>        <char> <char>                 <char>      <num>
##    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
##   ---
##  996: 233-67-5758      C        Normal   Male      Health and beauty      40.35
##  997: 303-96-2227      B        Normal Female     Home and lifestyle      97.38
##  998: 727-02-1313      A        Member   Male     Food and beverages      31.84
##  999: 347-56-2442      A        Normal   Male     Home and lifestyle      65.82
## 1000: 849-09-3807      A        Member Female    Fashion accessories      88.34
##       Quantity     Tax      Date   Time     Payment   cogs
##          <int>   <num>    <char> <char>      <char>  <num>
##    1:        7 26.1415  1/5/2019  13:08     Ewallet 522.83
##    2:        5  3.8200  3/8/2019  10:29        Cash  76.40
##    3:        7 16.2155  3/3/2019  13:23 Credit card 324.31
##    4:        8 23.2880 1/27/2019  20:33     Ewallet 465.76
##    5:        7 30.2085  2/8/2019  10:37     Ewallet 604.17
##   ---
##  996:        1  2.0175 1/29/2019  13:46     Ewallet  40.35
##  997:       10 48.6900  3/2/2019  17:16     Ewallet 973.80
##  998:        1  1.5920  2/9/2019  13:22        Cash  31.84
##  999:        1  3.2910 2/22/2019  15:33        Cash  65.82
## 1000:        7 30.9190 2/18/2019  13:28        Cash 618.38
##       gross margin percentage gross income Rating     Total
##                         <num>        <num>  <num>     <num>
##    1:                4.761905      26.1415    9.1  548.9715
##    2:                4.761905       3.8200    9.6   80.2200
##    3:                4.761905      16.2155    7.4  340.5255
##    4:                4.761905      23.2880    8.4  489.0480
##    5:                4.761905      30.2085    5.3  634.3785
##   ---
##  996:                4.761905       2.0175    6.2   42.3675
##  997:                4.761905      48.6900    4.4 1022.4900
##  998:                4.761905       1.5920    7.7   33.4320
##  999:                4.761905       3.2910    4.1   69.1110
## 1000:                4.761905      30.9190    6.6  649.2990
```

```
# Displaying the numerical data columns
df1 <- df %>% select_if(is.numeric)
colnames(df1)
```

```
## [1] "Unit price"              "Quantity"
## [3] "Tax"                     "cogs"
## [5] "gross margin percentage" "gross income"
## [7] "Rating"                  "Total"
```

```
# Renaming columns for an easy analysis
df1 <- df1 %>% rename(Unit_price = "Unit price")
df1 <- df1 %>% rename(gross_income = "gross income")

# Selecting needed columns
df2 <- subset(df1, select = c("Unit_price", "Quantity", "Tax", "cogs", "gross_income", "Rating", "Total
colnames(df2)
```

```
## [1] "Unit_price"   "Quantity"     "Tax"          "cogs"         "gross_income"
## [6] "Rating"       "Total"
```

3.FEATURE SELECTION

Using filter methods

```
# Loading libraries
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:boot':
##
##     melanoma
```

```
##
## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##
##     cluster
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
colnames(df2)
```

```
## [1] "Unit_price"   "Quantity"     "Tax"          "cogs"         "gross_income"
## [6] "Rating"       "Total"
```

```
# Calculating the correlation matrix
correlationMatrix <- cor(df2)
```

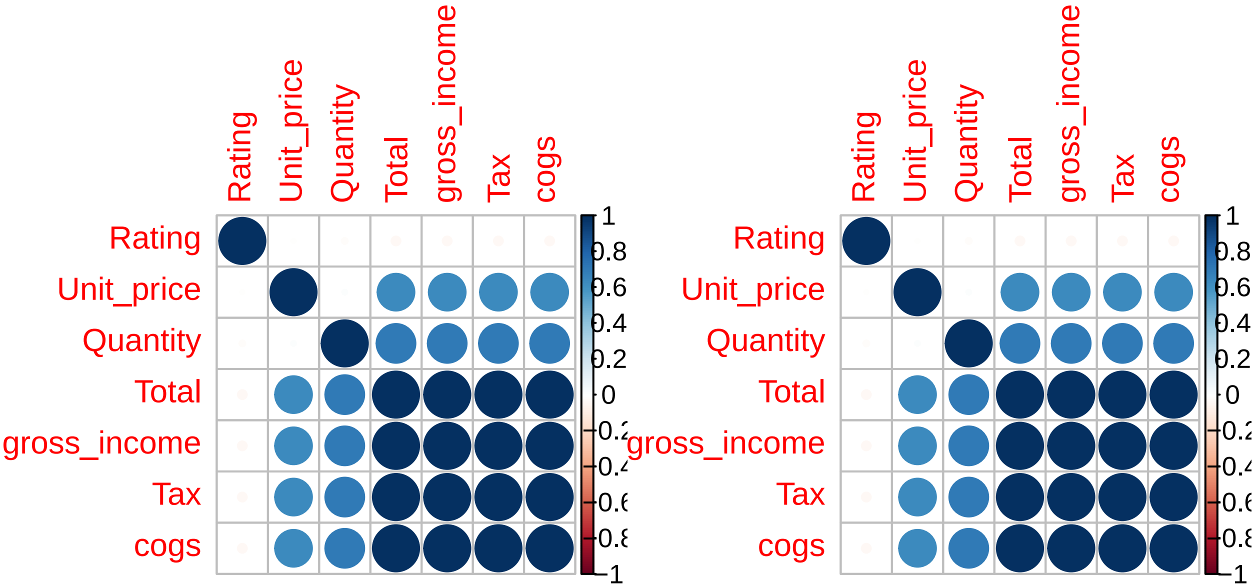
```
# Attributes that are highly correlated
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.75)
highlyCorrelated
```

```
## [1] 4 7 3
```

Highly correlated attributes.

```
# Removing the variables with a higher correlation
df3<-df2[-highlyCorrelated]
```

```
# Graphical comparison
par(mfrow = c(1, 2))
corrplot(correlationMatrix, order = "hclust")
corrplot(cor(df3), order = "hclust")
```



Graphical comparison.

Using Wrapper Methods

```
# Installing and loading our clustvarsel package
suppressWarnings(
    suppressMessages(if
                     (!require(clustvarsel, quietly=TRUE))
        install.packages("clustvarsel")))
library(clustvarsel)
```

```
# Installing and loading our mclust package

suppressWarnings(
    suppressMessages(if
                     (!require(mclust, quietly=TRUE))
        install.packages("mclust")))
library(mclust)
```

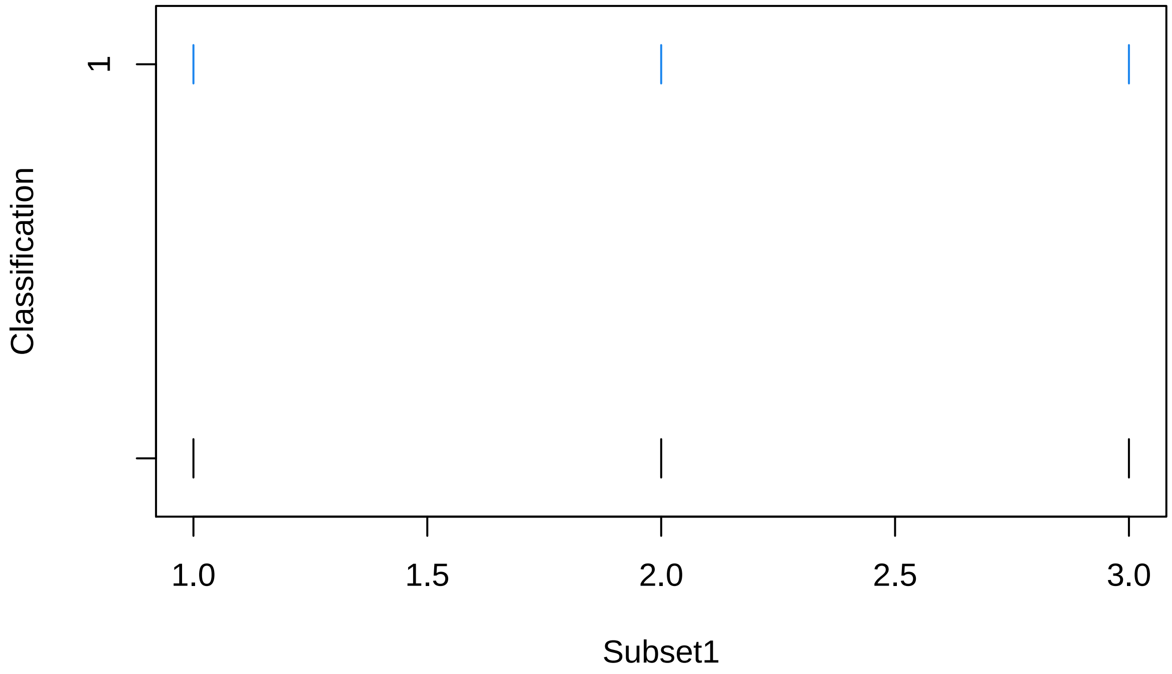
```
# Sequential forward greedy search (default)
out = clustvarsel(df3, G = 1:5)
out
```

```
## ------------------------------------------------------
## Variable selection for Gaussian model-based clustering
## Stepwise (forward/backward) greedy search
## ------------------------------------------------------
##
##  Variable proposed Type of step  BICclust Model G   BICdiff Decision
##                Tax          Add  -7359.02     V 4  391.4098 Accepted
##           Quantity          Add -11021.89   VEE 5  640.9594 Accepted
##         Unit_price          Add -16279.78   VVV 5 2620.0483 Accepted
##         Unit_price       Remove -11021.89   VEE 5 2620.0483 Rejected
##             Rating          Add -20603.86   EVV 5 -400.3689 Rejected
##         Unit_price       Remove -11021.89   VEE 5 2620.0483 Rejected
##
## Selected subset: Tax, Quantity, Unit_price
```

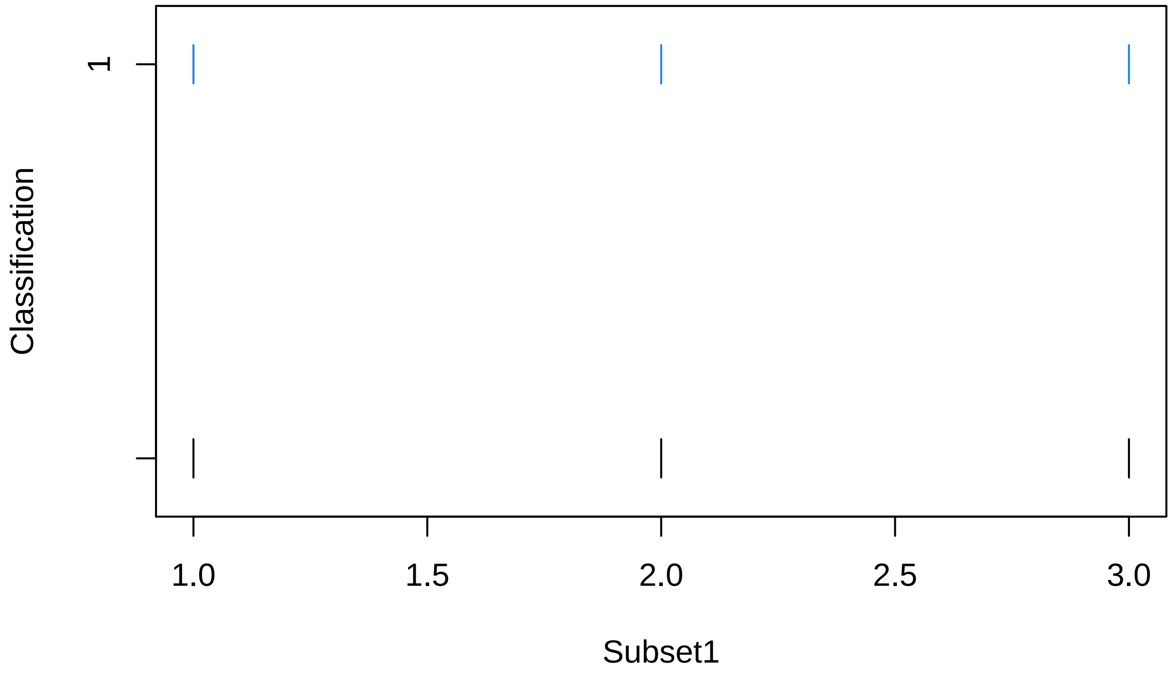
```
# Creating the clustering model:
Subset1 = df2[,out$subset]
mod = Mclust(Subset1, G = 1:5)
summary(mod)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust X (univariate normal) model with 1 component:
##
##  log-likelihood n df      BIC      ICL
##       -3.648618 3  2 -9.49446 -9.49446
##
## Clustering table:
## 1
## 3
```

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



```
plot(mod,c("classification"))
```



Using Embedded Methods

```
library(wskm)
```

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

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

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

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

```
df4 <- df[,apply(df2, 2, var, na.rm=TRUE) != 0]
df4=prcomp(df4)
model <- ewkm(df2[1:4], 3, lambda=2, maxiter=1000)
```

```
#checking weights
round(model$weights*100,2)
```

```
##   Unit_price Quantity   Tax  cogs gross_income Rating Total
## 1      14.29    14.29 14.29 14.29        14.29  14.29 14.29
## 2       0.00    40.46  6.79  0.00         6.79  45.96  0.00
## 3      14.29    14.29 14.29 14.29        14.29  14.29 14.29
```

The following were the most important variables: tax, cogs, quantity, total, gross income and the rating.