Feature selection

Alex

16/07/2021

Feature Selection

```
path<-"http://bit.ly/CarreFourDataset"

Dataset<-read.csv(path, sep = ",", dec = ".",row.names = 1)
Dataset<-Dataset[-4]
head(Dataset,3)
```

```
##             Branch Customer.type Gender Unit.price Quantity     Tax     Date
## 750-67-8428      A        Member Female      74.69        7 26.1415 1/5/2019
## 226-31-3081      C        Normal Female      15.28        5  3.8200 3/8/2019
## 631-41-3108      A        Normal   Male      46.33        7 16.2155 3/3/2019
##              Time     Payment   cogs gross.margin.percentage gross.income
## 750-67-8428 13:08     Ewallet 522.83                4.761905      26.1415
## 226-31-3081 10:29        Cash  76.40                4.761905       3.8200
## 631-41-3108 13:23 Credit card 324.31                4.761905      16.2155
##             Rating    Total
## 750-67-8428    9.1 548.9715
## 226-31-3081    9.6  80.2200
## 631-41-3108    7.4 340.5255
```

Installing and loading our caret package

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

Installing and loading the corrplot package for plotting

```
suppressWarnings(
        suppressMessages(if
```

```
                        (!require(corrplot, quietly=TRUE))
                install.packages("corrplot")))
library(corrplot)
```

Calculating the correlation matrix

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
head(Dataset)
```

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

```
dataset2 <- select(Dataset, c(5, 10,  12, 13, 14))

correlationMatrix <- cor(dataset2)
```

```
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.75)


highlyCorrelated
```

```
## [1] 2 3
```

```
names(Dataset[,highlyCorrelated])
```

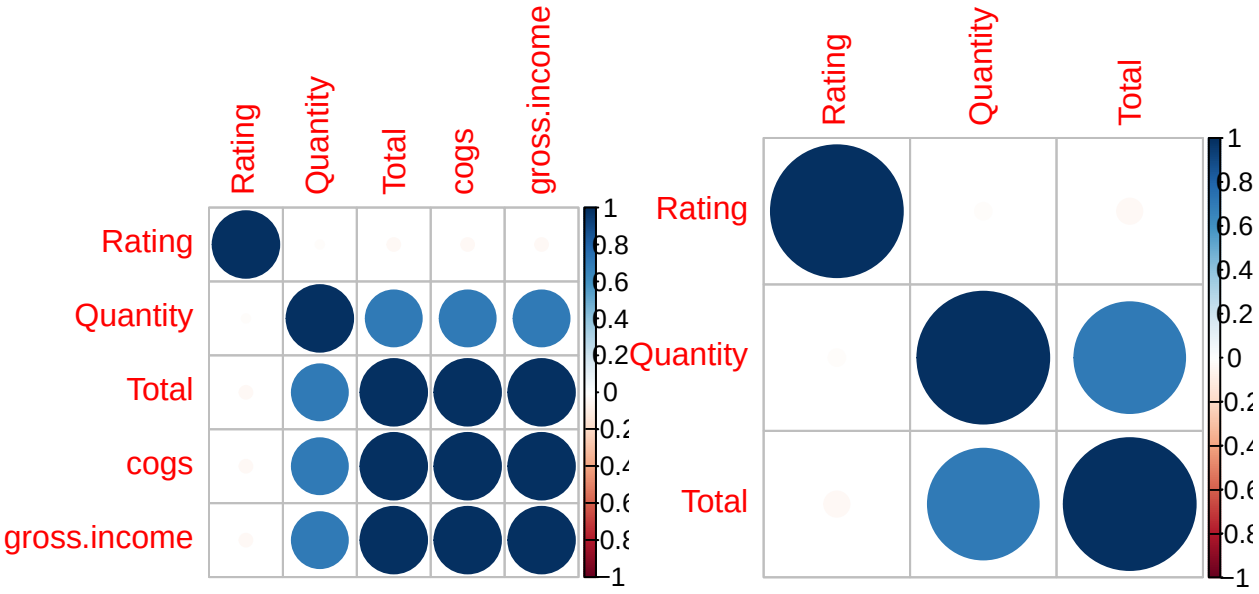
```
## [1] "Customer.type" "Gender"
```

```
final_features <- dataset2[-highlyCorrelated]
```

```
library(corrplot)

par(mfrow = c(1, 2))

corrplot(correlationMatrix, order = "hclust")
corrplot(cor(final_features), order = "hclust")
```



Conclusion

The selected features that contribute most of the information in the dataset are Rating, Quantity, Total, cogs and gross income