Dimensionality Reduction

Alex

16/07/2021

IMPORTING OUR DATASET

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

Dataset<-read.csv(path, sep = ",", dec = ".",row.names = 1)

head(Dataset)
```

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

```
feature <- Dataset
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
```

Label encoding the categorical column Gender

```
Dataset$Gender <- ifelse(Dataset$Gender == "Male",1,2)
table(Dataset$Gender)
```

```
##
##   1   2
## 499 501
```

label encoding the customer type column

```
Dataset$Customer.type <- ifelse(Dataset$Customer.type == "Member",1,2)
table(Dataset$Customer.type)
```

```
##
##   1   2
## 501 499
```

label encoding the payment column

```
Dataset$Payment <- as.numeric(Dataset$Payment)
```

```
## Warning: NAs introduced by coercion
```

```
table(Dataset$Payment)
```

```
## < table of extent 0 >
```

label encoding the product line column

```
Dataset$Product.line <- as.numeric(Dataset$Product.line)
```

```
## Warning: NAs introduced by coercion
```

```
table(Dataset$Product.line)
```

```
## < table of extent 0 >
```

label encoding the branch column

```
Dataset$Branch <- as.numeric(Dataset$Branch)
```

```
## Warning: NAs introduced by coercion
```

```
table(Dataset$Branch)
```

```
## < table of extent 0 >
```

```
data2 <- select(Dataset, c(2,3,5,6,7,8,9,11,13,14,15))
```

```
head(data2)
```

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

```
data2 <- data2[ , unlist(lapply(data2, is.numeric))]
```

```
pca <- prcomp(data2, center = TRUE, scale. = TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6       PC7
## Standard deviation     2.2193 1.0331 1.0057 0.9931 0.9590 0.29992 5.124e-16
## Proportion of Variance 0.5473 0.1186 0.1124 0.1096 0.1022 0.00999 0.000e+00
## Cumulative Proportion  0.5473 0.6659 0.7782 0.8878 0.9900 1.00000 1.000e+00
##                            PC8       PC9
## Standard deviation     1.99e-16 1.192e-16
## Proportion of Variance 0.00e+00 0.000e+00
## Cumulative Proportion  1.00e+00 1.000e+00
```

PCA is not suitable for this data since some principal components do not convey most of the information of the data hence we can use tnse as an alternative method