PCA

Thileepan Paulraj

18 joulukuuta 2018

UNDERSTANDING PRINCIPAL COMPONENT ANALYSIS (work reproduced from this webpage:https://goo.gl/Wgeieb)

Reading data

```
data = read.csv('diamonds.csv')
```

Viewing all the variable names in the dataset

```
colnames(data)
```

```
##  [1] "X"       "carat"   "cut"     "color"   "clarity" "depth"   "table"
##  [8] "price"   "x"       "y"       "z"
```

Taking only the numeric variables so we could use it in our analysis

```
library(dplyr)
```

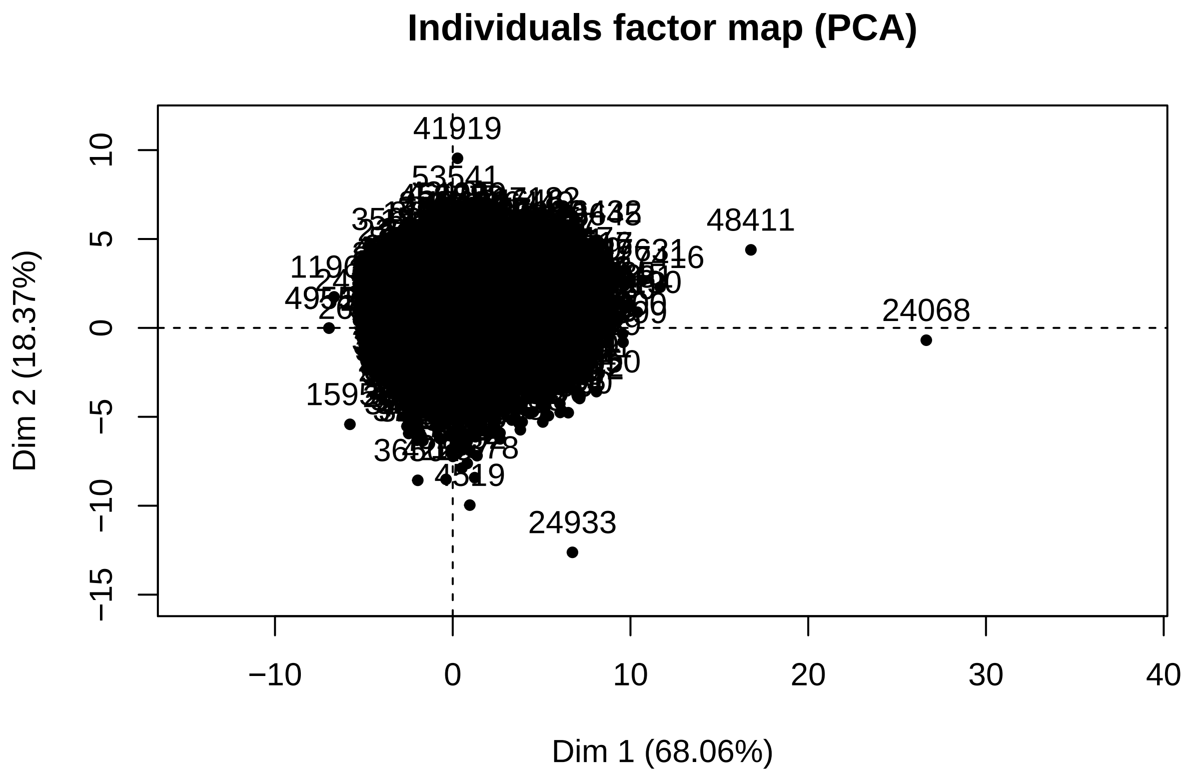
```
##
## Attaching package: 'dplyr'

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

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

```
data_for_pca <- select(data, -X, -cut, -color, -clarity)
```

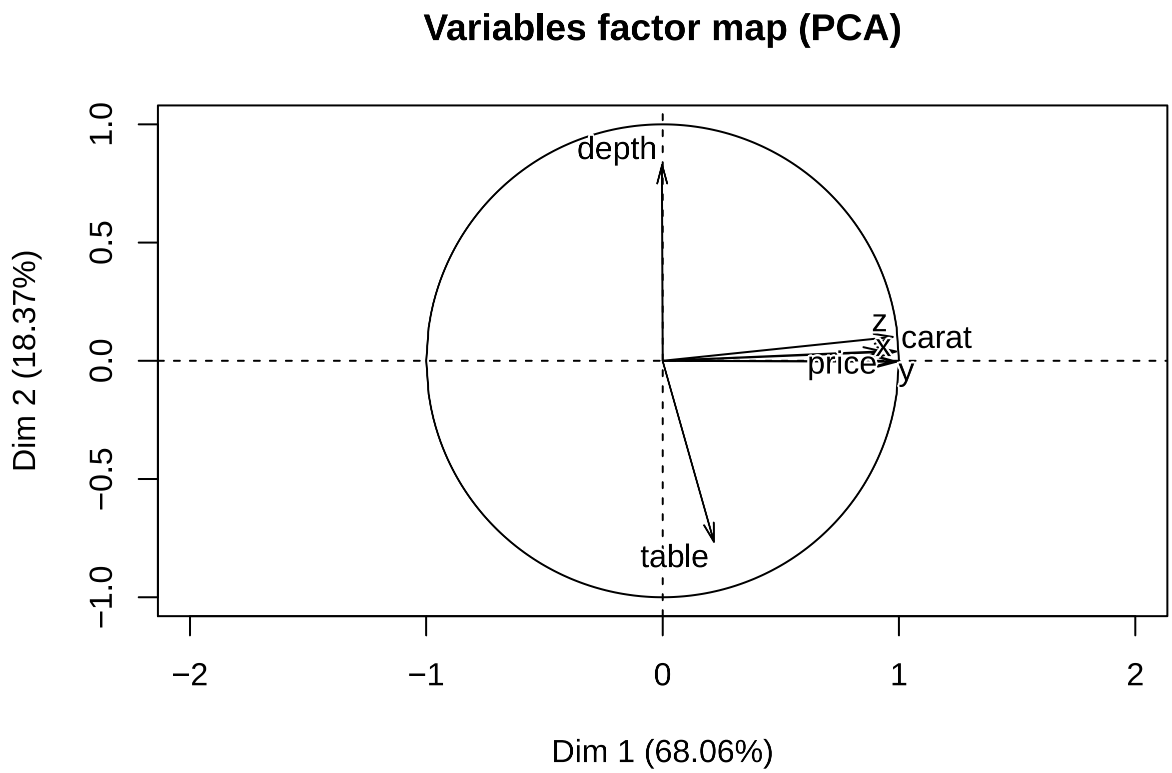
Installing the factominer package for PCA

```
library(FactoMineR)
```

```
pca = PCA(data_for_pca)
```



Individuals factor map (PCA)
41919
53541
48411
24068
24933
Dim 2 (18.37%)
Dim 1 (68.06%)
10
5
0
−5
−10
−15
−10
0
10
20
30
40



Here,the variables, 'price', 'carat', 'x', 'y', and 'z' form a composite variable called the Principal component 1 or Dim 1 which explains 68.06% of the variance in the data. Variable 'depth' explains 18.37% of the variance in the data along the second dimension. The variable 'table' is in the third dimension.

```
pca$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 4.76391480             68.0559258                          68.05593
## comp 2 1.28586808             18.3695440                          86.42547
## comp 3 0.69081126              9.8687323                          96.29420
## comp 4 0.17375333              2.4821905                          98.77639
## comp 5 0.04030722              0.5758174                          99.35221
## comp 6 0.03294659              0.4706656                          99.82288
## comp 7 0.01239871              0.1771245                         100.00000
```