post3-PCA

Rongyun Tang

December 11, 2018

Daily Demand Forecasting Orders Data

research outlines:

1. Objective:task is to determine what kind of order resources contribute most to the information, and then to reduce data dimentions.
2. Method: CPA
3. Data Source:this "Daily Demand Forecasting Orders Data(DDFOD)" dataset was downloaded from UCI Machine Learning website: https://archive.ics.uci.edu/ml/datasets/Daily+Demand+Forecasting+Orders. The dataset was collected during 60 days, this is a real database of a brazilian logistics company.Twelve predictive attributes and a target that is the total of orders for daily treatment.

==

Attribute Information:

==

The dataset was collected during 60 days, this is a real database of a brazilian logistics company. The dataset has twelve predictive attributes and a target that is the total of orders for daily treatment.

- Week_of_the_month (WM): {1.0, 2.0, 3.0, 4.0, 5.0}
- Day_of_the_week_(Monday_to_Friday)(DW): {2.0, 3.0, 4.0, 5.0, 6.0}
- Non_urgent_order(NUO): integer
- Urgent order(UO): integer
- Order type A(typeA): integer
- Order type B(typeB): integer
- Order type C(typeC):integer
- Fiscal sector orders(FO): integer
- Orders from the traffic controller sector(Traffic): integer
- Banking orders (1)(Bank1): integer
- Banking orders (2)(Bank2): integer
- Banking orders (3)(Bank3): integer
- Target(Total_orders)(Total):integer

1. Data Preprocessing

```
# Rename column names and Read data
rawdata <- read.csv('Daily_Demand_Forecasting_Orders2.csv')
head(rawdata)
```

```
##    WM DW     NUO      UO  typeA   typeB   typeC     FO Traffic Bank1  Bank2
## 1  1  4 316.307 223.270 61.543 175.586 302.448  0.000   65556 44914 188411
## 2  1  5 128.633  96.042 38.058  56.037 130.580  0.000   40419 21399  89461
## 3  1  6  43.651  84.375 21.826  25.125  82.461  1.386   11992  3452  21305
## 4  2  2 171.297 127.667 41.542 113.294 162.284 18.156   49971 33703  69054
## 5  2  3  90.532 113.526 37.679  56.618 116.220  6.459   48534 19646  16411
## 6  2  4 110.925  96.360 30.792  50.704 125.868 79.000   52042  8773  47522
##   Bank3   Total
## 1 14793 539.577
## 2  7679 224.675
## 3 14947 129.412
## 4 18423 317.120
## 5 20257 210.517
## 6 24966 207.364
```

```r
DOY<-as.factor(rawdata$WM*30+rawdata$DW)    # transfer week of month and day of week into day of year(D
data2<-cbind(DOY,rawdata[,3:13])
head(data2)
```

```
##   DOY     NUO      UO  typeA   typeB   typeC     FO Traffic Bank1  Bank2
## 1  34 316.307 223.270 61.543 175.586 302.448  0.000   65556 44914 188411
## 2  35 128.633  96.042 38.058  56.037 130.580  0.000   40419 21399  89461
## 3  36  43.651  84.375 21.826  25.125  82.461  1.386   11992  3452  21305
## 4  62 171.297 127.667 41.542 113.294 162.284 18.156   49971 33703  69054
## 5  63  90.532 113.526 37.679  56.618 116.220  6.459   48534 19646  16411
## 6  64 110.925  96.360 30.792  50.704 125.868 79.000   52042  8773  47522
##   Bank3   Total
## 1 14793 539.577
## 2  7679 224.675
## 3 14947 129.412
## 4 18423 317.120
## 5 20257 210.517
## 6 24966 207.364
```

2.Data Exploration

```r
# Missing value detection
sum(is.na(data2))
```

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

```r
print("Numbers of missing values : 0")
```

```
## [1] "Numbers of missing values : 0"
```

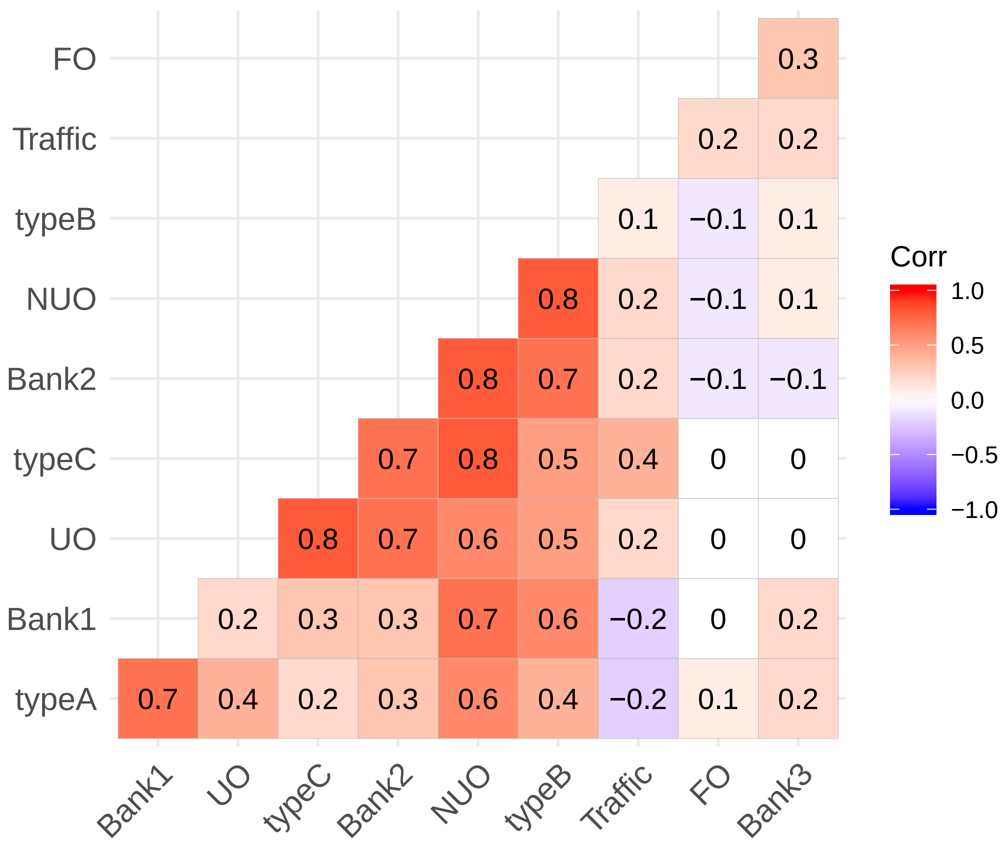
```r
# Data Correlationships

#install.packages("ggcorrplot")
library('ggcorrplot')
```

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

```r
corr <- round(cor(data2[,c(-1,-12)]), 1)
head(corr[, 1:6])
```

```
##         NUO  UO typeA typeB typeC   FO
## NUO     1.0 0.6   0.6   0.8   0.8 -0.1
## UO      0.6 1.0   0.4   0.5   0.8  0.0
## typeA   0.6 0.4   1.0   0.4   0.2  0.1
## typeB   0.8 0.5   0.4   1.0   0.5 -0.1
## typeC   0.8 0.8   0.2   0.5   1.0  0.0
## FO     -0.1 0.0   0.1  -0.1   0.0  1.0
```

```r
ggcorrplot(corr, hc.order = TRUE, type = "lower",lab = TRUE)
```




```r
# Resutls show that there are many variables mutually correlated, dimension could be reduced with PCA m
```

3. PCA and Visualization

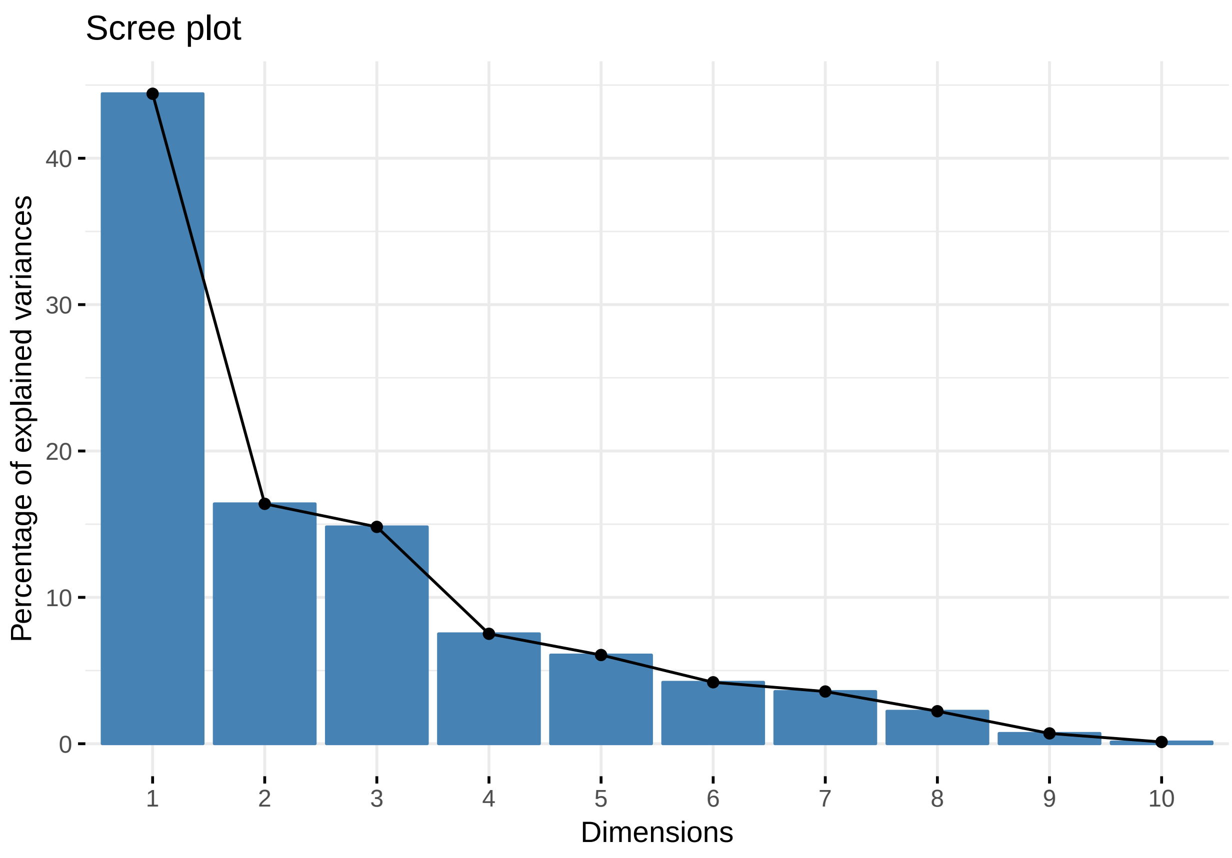
```r
# 3.1 apply PCA

library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

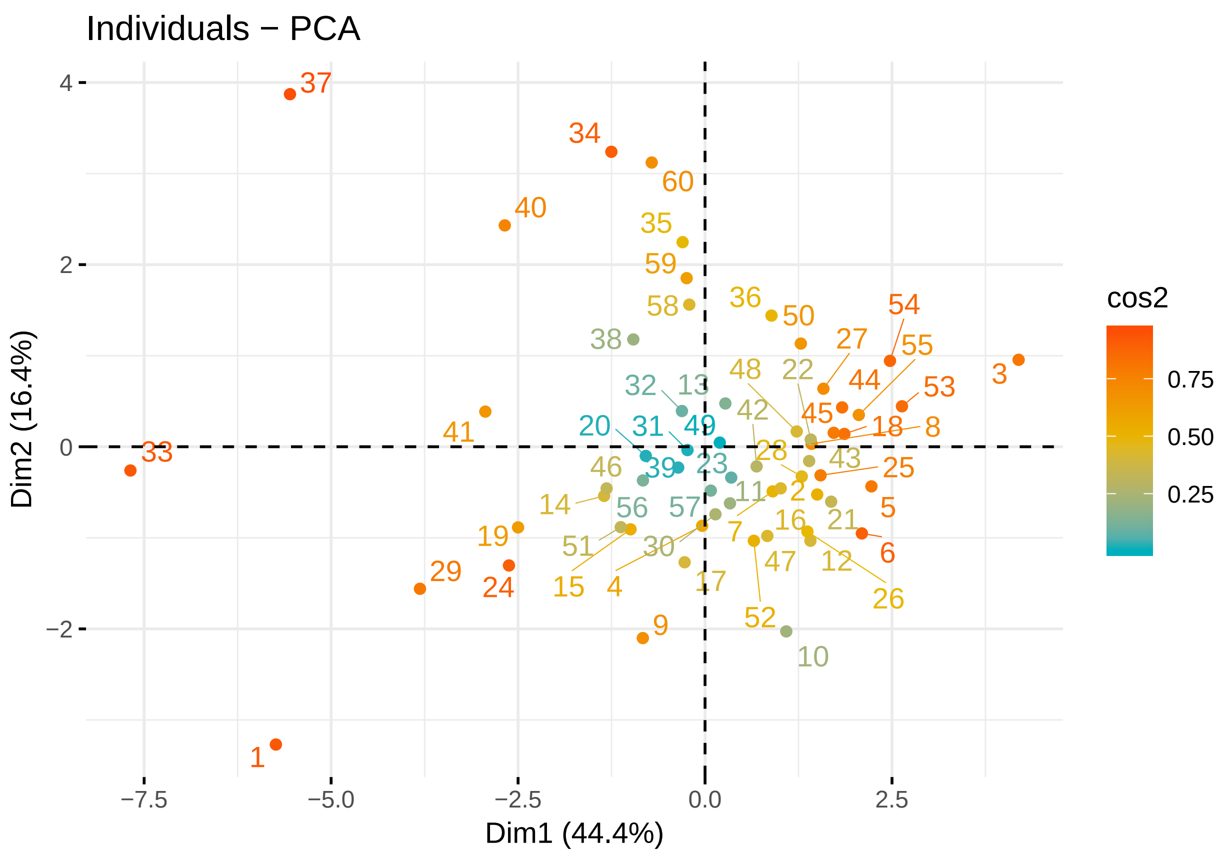
```r
data.pca <- prcomp(data2[,c(-1,-12)], center = TRUE,scale. = TRUE)
summary(data.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.1073 1.2802 1.2172 0.8666 0.77855 0.64800 0.59744
## Proportion of Variance 0.4441 0.1639 0.1482 0.0751 0.06061 0.04199 0.03569
## Cumulative Proportion  0.4441 0.6080 0.7561 0.8312 0.89183 0.93382 0.96951
##                            PC8     PC9    PC10
## Standard deviation     0.47140 0.26567 0.10986
## Proportion of Variance 0.02222 0.00706 0.00121
## Cumulative Proportion  0.99173 0.99879 1.00000
```

```r
# 3.12 results: PC1 explains 44% of the total variance,  PC2 explains 16% of the variance , PC3 expalin

fviz_eig(data.pca)
```



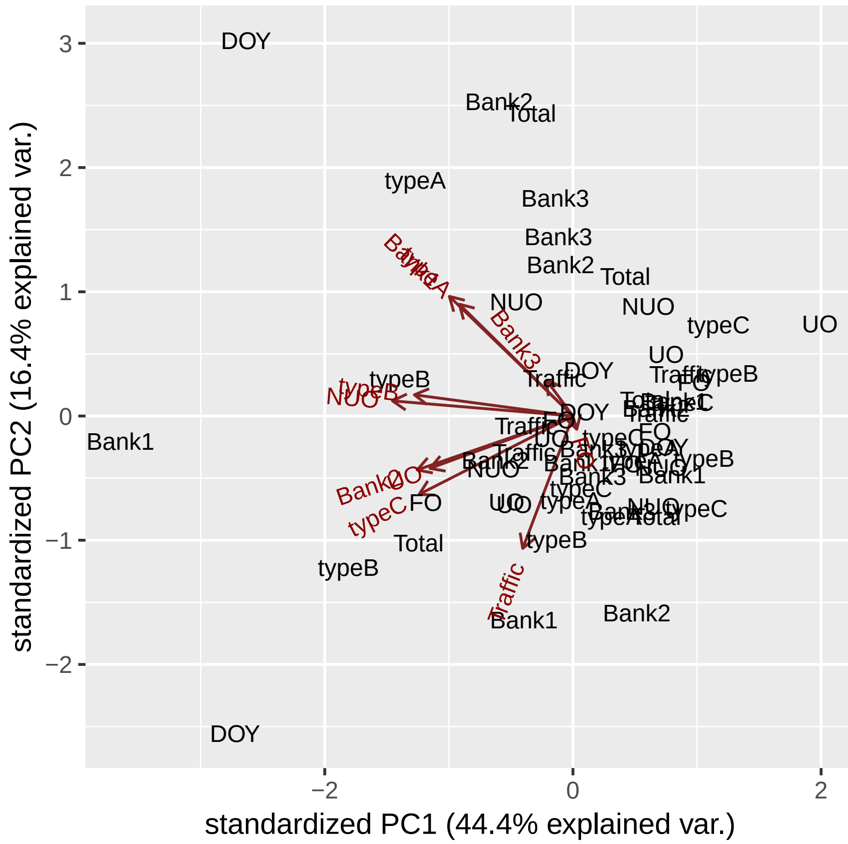

```r
#3.2 Individuals with a similar profile are grouped together
fviz_pca_ind(data.pca,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```



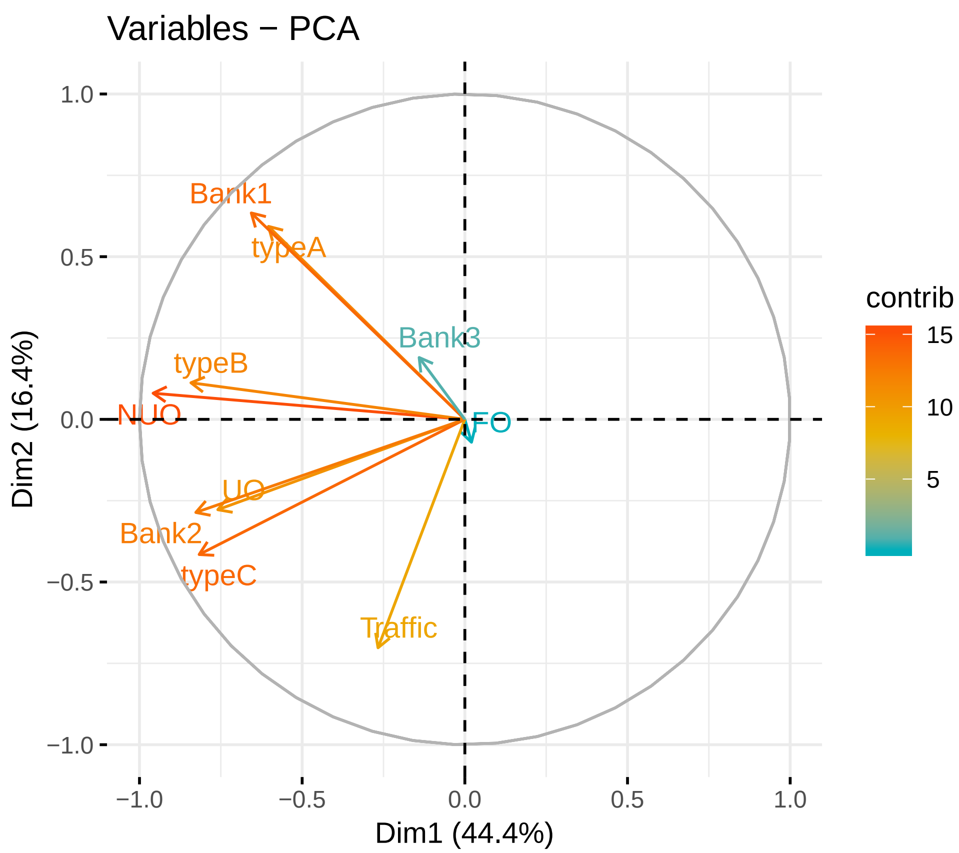

```r
library(devtools)
library(ggbiplot)
```

```
## Loading required package: plyr

## Loading required package: scales

## Loading required package: grid
```

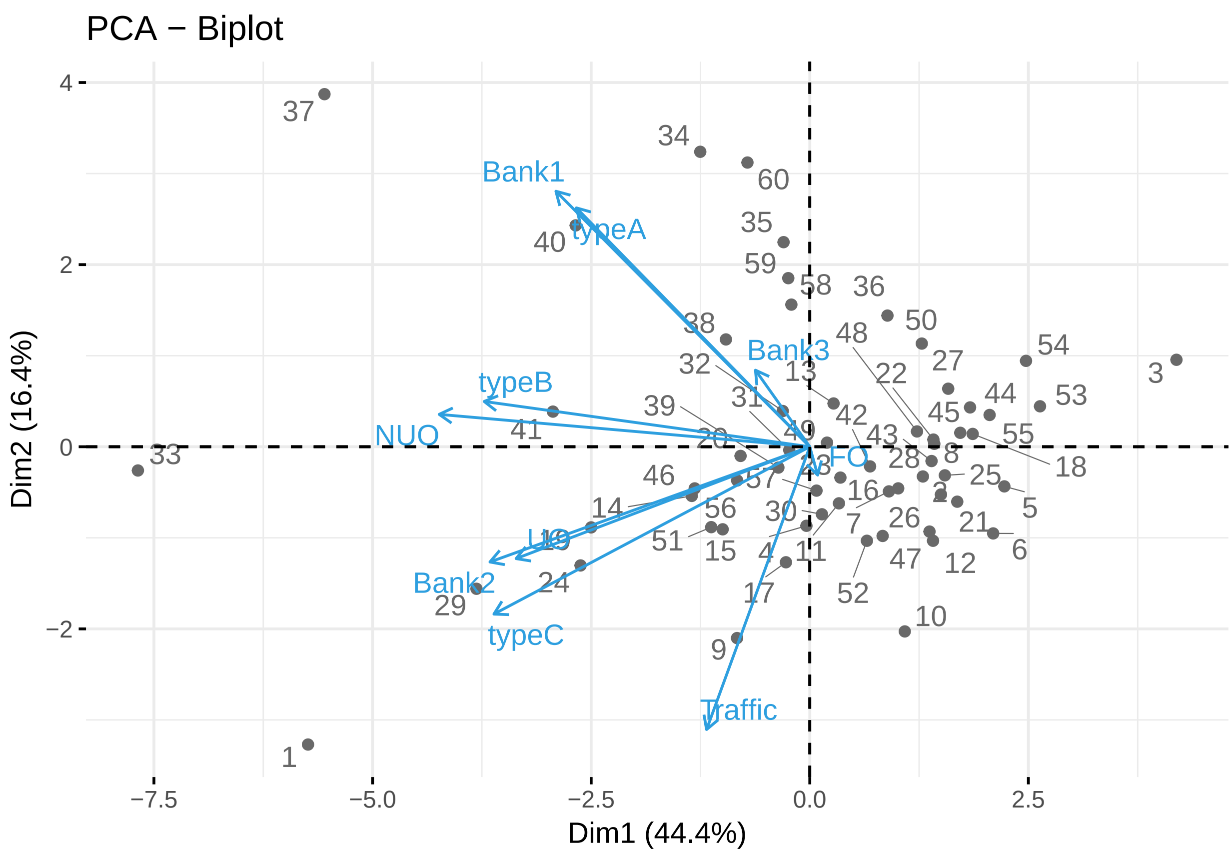
```r
ggbiplot(data.pca, labels=colnames(data2))
```




```
# 3.3 Variables with a similar profile are grouped together
#     Positive correlated variables point to the same side of the plot. Negative correlated variables p
fviz_pca_var(data.pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             )
```




```
# 3.32 results: NUO(not urgent orders) and Bank1 are two most strong positive-correlated variables. Ban

# 3.4 Biplot of individuals and variables
fviz_pca_biplot(data.pca, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969"  # Individuals color
                )
```



4. PCA results

```
library(factoextra)
# Eigenvalues
eig.val <- get_eigenvalue(data.pca)
eig.val
```

```
##         eigenvalue variance.percent cumulative.variance.percent
## Dim.1   4.44060587       44.4060587                    44.40606
## Dim.2   1.63896499       16.3896499                    60.79571
## Dim.3   1.48154899       14.8154899                    75.61120
## Dim.4   0.75102531        7.5102531                    83.12145
## Dim.5   0.60614507        6.0614507                    89.18290
## Dim.6   0.41990518        4.1990518                    93.38195
## Dim.7   0.35694039        3.5694039                    96.95136
## Dim.8   0.22221416        2.2221416                    99.17350
## Dim.9   0.07058052        0.7058052                    99.87930
## Dim.10  0.01206950        0.1206950                   100.00000
```

```
# Results for Variables
res.var <- get_pca_var(data.pca)
head(res.var$coord)          # Coordinates
```

```
##            Dim.1       Dim.2        Dim.3       Dim.4      Dim.5
## NUO   -0.9582251  0.08037040 -0.013984866  0.11687311 -0.1835161
## UO    -0.7592925 -0.27772052 -0.054364438 -0.25806876  0.4712286
## typeA -0.6035404  0.59306371  0.143750503 -0.18834192  0.2662002
## typeB -0.8418828  0.11243652 -0.138149132  0.15259525 -0.2576531
## typeC -0.8166467 -0.41549038 -0.003051467 -0.04273086  0.0953080
## FO     0.0204144 -0.07040925  0.748188673 -0.58733541 -0.2910611
##             Dim.6       Dim.7       Dim.8        Dim.9       Dim.10
## NUO    0.05933819 -0.03607775  0.12381078  0.002112668 -0.087498635
## UO    -0.04463561 -0.03527084 -0.20322410  0.098984015 -0.020014851
## typeA  0.16274400  0.33888380  0.09760983 -0.077519089  0.012258321
## typeB -0.26195698  0.12875595 -0.27644519 -0.089898302  0.011516406
## typeC  0.16252846 -0.31058414  0.06362295 -0.147667097  0.028382303
## FO    -0.04957464 -0.03153525 -0.04000742 -0.008340836 -0.006146227
```

```r
head(res.var$contrib)        # Contributions to the PCs
```

```
##              Dim.1      Dim.2        Dim.3      Dim.4     Dim.5      Dim.6
## NUO   20.677253378  0.3941146 1.320081e-02  1.8187568  5.556122  0.8385276
## UO    12.983027518  4.7059387 1.994866e-01  8.8678085 36.634203  0.4744732
## typeA  8.202957299 21.4601631 1.394770e+00  4.7232334 11.690692  6.3075216
## typeB 15.961033940  0.7713387 1.288191e+00  3.1004693 10.952014 16.3421321
## typeC 15.018488334 10.5330044 6.284945e-04  0.2431245  1.498588  6.2908249
## FO     0.009384926  0.3024752 3.778385e+01 45.9322581 13.976284  0.5852857
##            Dim.7      Dim.8        Dim.9     Dim.10
## NUO    0.3646559  6.8983492  0.006323793 63.4326954
## UO     0.3485265 18.5856905 13.881783928  3.3190617
## typeA 32.1740644  4.2876114  8.513977043  1.2450093
## typeB  4.6444998 34.3911220 11.450333320  1.0988654
## typeC 27.0248237  1.8216118 30.894603705  6.6743024
## FO     0.2786100  0.7202934  0.098567623  0.3129881
```

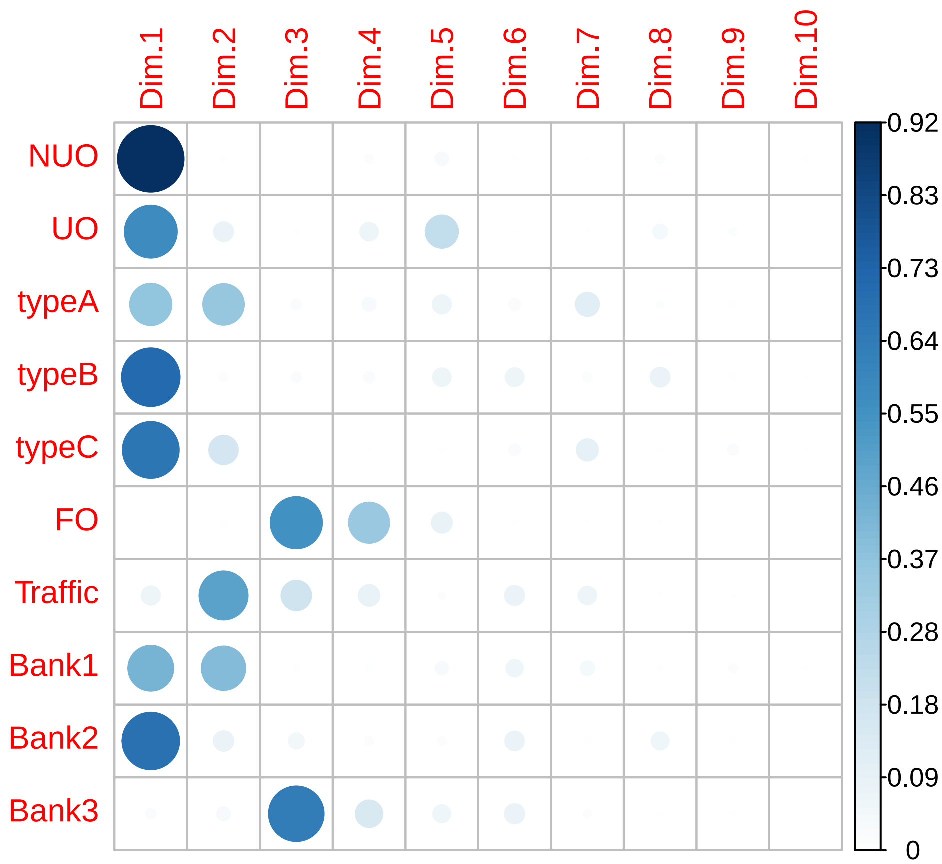
```r
head(res.var$cos2)  # Quality of representation
```

```
##              Dim.1       Dim.2        Dim.3       Dim.4       Dim.5
## NUO   0.9181953276 0.006459401 1.955765e-04 0.013659324 0.033678162
## UO    0.5765250823 0.077128687 2.955492e-03 0.066599486 0.222056417
## typeA 0.3642610035 0.351724561 2.066421e-02 0.035472678 0.070862554
## typeB 0.7087666103 0.012641971 1.908518e-02 0.023285309 0.066385095
## typeC 0.6669118748 0.172632255 9.311453e-06 0.001825926 0.009083615
## FO    0.0004167476 0.004957462 5.597863e-01 0.344962886 0.084716559
##             Dim.6        Dim.7       Dim.8        Dim.9       Dim.10
## NUO   0.003521021 0.0013016043 0.015329109 4.463366e-06 7.656011e-03
## UO    0.001992338 0.0012440320 0.041300037 9.797835e-03 4.005943e-04
## typeA 0.026485610 0.1148422307 0.009527680 6.009209e-03 1.502664e-04
## typeB 0.068621459 0.0165780957 0.076421944 8.081705e-03 1.326276e-04
## typeC 0.026415500 0.0964625111 0.004047880 2.180557e-02 8.055551e-04
## FO    0.002457645 0.0009944717 0.001600594 6.956954e-05 3.777611e-05
```

```r
library("corrplot")
```

```
## corrplot 0.84 loaded
```

```r
corrplot(res.var$cos2, is.corr=FALSE)
```




```r
# Results for individuals
res.ind <- get_pca_ind(data.pca)
head(res.ind$coord)          # Coordinates
```

```
##         Dim.1      Dim.2       Dim.3      Dim.4      Dim.5      Dim.6
## 1 -5.73817137 -3.2703564 -0.84411722 -0.7258498  1.3848215  0.4108160
## 2  1.50007633 -0.5246814 -1.15222491 -0.3009059 -0.3826778  0.2716387
## 3  4.19325338  0.9549023 -1.33114647 -0.5715037  0.2313122 -0.7561606
## 4 -0.03842312 -0.8681350 -0.33534062  0.2290695  0.1015195  0.3164103
## 5  2.22507563 -0.4353534  0.03135665  0.3470495  0.5796210  0.6940573
## 6  2.09736994 -0.9518579  0.44441549  0.4845820 -0.0591205  0.3029552
##        Dim.7       Dim.8       Dim.9      Dim.10
## 1 -0.5160051 -0.04136038 -0.14393324 -0.02771979
## 2 -0.1410312  0.75314566  0.15785284  0.26640646
## 3 -1.2922098  0.12175782 -0.03883196 -0.08548415
## 4 -0.2775529 -0.39365795 -0.18295331 -0.08440955
## 5 -0.1295618 -0.61871430  0.07246249 -0.04992735
## 6 -0.2214792  0.18399574  0.03878942  0.04663445
```

```r
head(res.ind$contrib)          # Contributions to the PCs
```

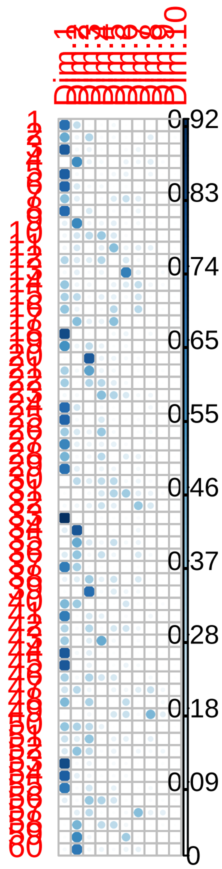
```
##          Dim.1      Dim.2       Dim.3      Dim.4       Dim.5      Dim.6
```

```
## 1 1.235815e+01 10.8760014 0.801564090 1.1691971 5.273023942 0.6698725
## 2 8.445653e-01  0.2799435 1.493506962 0.2009349 0.402660162 0.2928741
## 3 6.599465e+00  0.9272502 1.993353957 0.7248236 0.147119178 2.2694760
## 4 5.541045e-04  0.7663967 0.126504233 0.1164471 0.028338140 0.3973732
## 5 1.858215e+00  0.1927361 0.001106094 0.2672865 0.923762774 1.9120012
## 6 1.651036e+00  0.9213471 0.222183145 0.5211091 0.009610554 0.3642958
##        Dim.7      Dim.8      Dim.9    Dim.10
## 1 1.24325735 0.01283057 0.48919961 0.1061058
## 2 0.09287165 4.25436715 0.58839466 9.8005142
## 3 7.79684869 0.11119128 0.03560759 1.0090916
## 4 0.35970339 1.16229149 0.79039551 0.9838808
## 5 0.07838030 2.87115952 0.12399106 0.3442203
## 6 0.22904398 0.25391746 0.03552962 0.3003122
```

```
head(res.ind$cos2)          # Quality of representation
```

```
##         Dim.1      Dim.2        Dim.3      Dim.4        Dim.5       Dim.6
## 1 0.697048446 0.22641548 0.0150841713 0.01115346 0.0405978943 0.003572815
## 2 0.464259176 0.05679696 0.2739102586 0.01868078 0.0302134468 0.015223594
## 3 0.767413676 0.03979653 0.0773355207 0.01425494 0.0023352019 0.024954875
## 4 0.001132957 0.57836621 0.0862979951 0.04026827 0.0079091090 0.076829760
## 5 0.763220579 0.02921759 0.0001515723 0.01856707 0.0517903452 0.074259353
## 6 0.743170263 0.15306732 0.0333669588 0.03967097 0.0005904927 0.015505827
##         Dim.7        Dim.8        Dim.9       Dim.10
## 1 0.005636687 0.0000362147 4.385692e-04 1.626656e-05
## 2 0.004103589 0.1170285257 5.140893e-03 1.464277e-02
## 3 0.072877484 0.0006470249 6.581214e-05 3.189323e-04
## 4 0.059118016 0.1189231565 2.568674e-02 5.467786e-03
## 5 0.002587703 0.0590120673 8.094443e-04 3.842710e-04
## 6 0.008287129 0.0057194439 2.541938e-04 3.674108e-04
```

```
corrplot(res.ind$cos2,is.corr=FALSE)
```



5. Conclusion

- 13 variables of the raw data could be reduced by 2 PCA principles.
- NUO contributes most to the effective information, followed by type B and Bank2.
- individual comtribution shows that types C has seasonal attribute, and it reaches maxmiu contribution in April.