Renaissance: задача №3

Георгий Шимановский

14 11 2017

Загрузка пакетов

```
require(readxl) # Пакет для чтения excel-файлов.
require(data.table) # Пакет для работы с таблицами в R.
require(ggbiplot) # Пакет для Biplot графика
# require(devtools); install_github("ggbiplot", "vqv")
require(ggplot2)
```

Загрузка данных из Excel файла.

```
path <- "Задача.xlsx" # путь к файлу.
dt.xls <- as.data.table(readxl::read_xlsx(path))
head(dt.xls)
```

Персона	Возраст, лет	Стаж вождения, лет	Убыточность, %	Уровень заработной платы, руб/год
6-LLJEH	20	1	263	716693
2-GLHFG	74	51	107	274393
6-FJFKL	27	1	165	723841
4-KJEJL	24	6	348	139419
5-JFFGH	26	3	286	650003
6-MFGJE	77	56	180	223249

Знакомство с данными

```
summary(dt.xls)
```

```
##    Персона           Возраст, лет   Стаж вождения, лет Убыточность, %
##  Length:484         Min.   :20.00   Min.   : 1.00      Min.   : 20.0
##  Class :character   1st Qu.:26.00   1st Qu.: 5.00      1st Qu.:103.0
##  Mode  :character   Median :39.50   Median :15.00      Median :173.0
##                     Mean   :45.32   Mean   :21.42      Mean   :227.1
##                     3rd Qu.:65.00   3rd Qu.:38.00      3rd Qu.:299.2
##                     Max.   :79.00   Max.   :58.00      Max.   :700.0
##  Уровень заработной платы, руб/год
##  Min.   : 120185
##  1st Qu.: 194998
##  Median : 441028
##  Mean   :1017901
##  3rd Qu.:1109668
##  Max.   :3982828
```

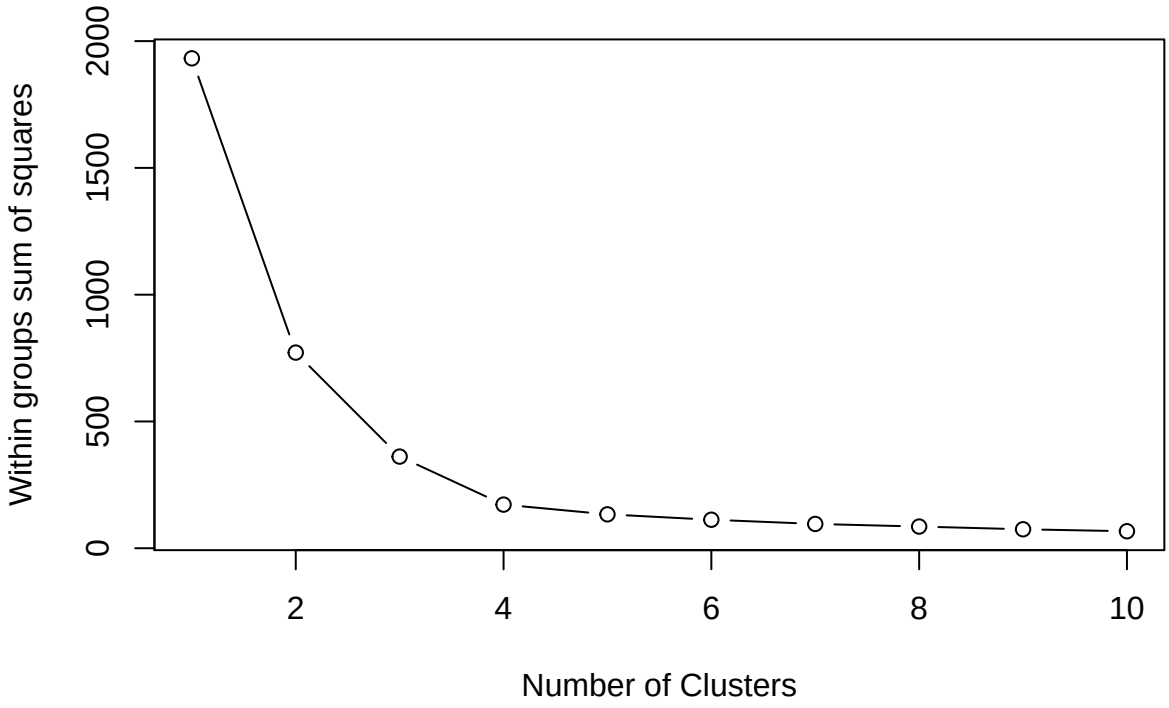
Нормализация данных

```
data.nrmlzd <- scale(dt.xls[, 2:5]) #Normalizing data features.
rownames(data.nrmlzd) <- dt.xls$Персона # Set the row names of data.nrmlzd
head(data.nrmlzd)
```

```
##         Возраст, лет Стаж вождения, лет Убыточность, %
## 6-LLJEH   -1.2897100         -1.1319647      0.2094040
## 2-GLHFG    1.4608295          1.6401002     -0.6998921
## 6-FJFKL   -0.9331586         -1.1319647     -0.3618205
## 4-KJEJL   -1.0859664         -0.8547582      0.7048538
## 5-JFFGH   -0.9840945         -1.0210821      0.3434669
## 6-MFGJE    1.6136372          1.9173067     -0.2743882
##         Уровень заработной платы, руб/год
## 6-LLJEH                         -0.2536391
## 2-GLHFG                         -0.6260881
## 6-FJFKL                         -0.2476199
## 4-KJEJL                         -0.7397461
## 5-JFFGH                         -0.3097969
## 6-MFGJE                         -0.6691551
```

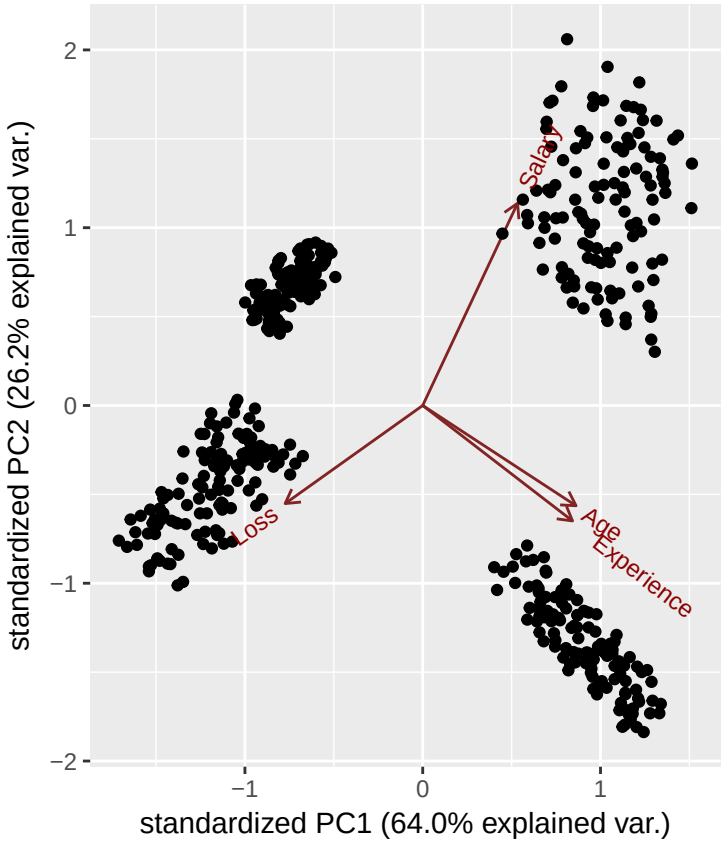
Определение оптимального количества кластеров

```
#Determine number of cluster by looping kmeans with cluter setting from 1 to 10.
set.seed(5) #for reproducability
wss.len <- 10L #Set length of the loop.
wss <- integer(wss.len) #Create integer vector (don't grow a vector for mem eff)

for (i in seq(wss.len)) {
        km.i <- kmeans(data.nrmlzd, centers = i, iter.max = 50, nstart = 20)
        # Save total within sum of squares to wss variable
        wss[i] <- km.i$tot.withinss
}

#Scree plot
plot(x = seq(wss.len), y = wss, type = "b",
     xlab = "Number of Clusters",
     ylab = "Within groups sum of squares")
```



PCA оптимизация на нормализованных данных

```
colnames(data.nrmlzd) <- c("Age", "Experience", "Loss", "Salary") #Имена столб.
pca.nrmlzd <- prcomp(data.nrmlzd) #PCA анализ
ggbiplot(pca.nrmlzd, obs.scale = 0, var.scale = 0) #Biplot график
```



Иерархическая клатеризация результатов PCA анализа

```
pca.hclust <- hclust(dist(pca.nrmlzd$x)) #H-clustering of pca data.

clust4 <- cutree(pca.hclust, k = 4) #h-clust cut tree at 4 clusters.
head(clust4)
```

```
## 6-LLJEH 2-GLHFG 6-FJFKL 4-KJEJL 5-JFFGH 6-MFGJE
##       1       2       1       3       1       2
```

Добавление кластеров к исходным данным

```
report <- cbind(dt.xls, clust4)
report.split <- lapply(split(report[, -1], report$clust4), summary)
```

Подгтовка отчета по кластерам

```
report <- cbind(dt.xls, clust4)
report.split <- split(report[, c(-1, -6)], clust4)
report.ranges <- lapply(report.split, apply, 2, range)
res.cols <- c("Мин", "Макс")
ranges.trans <- lapply(report.ranges, t)
```

```
for (i in seq_along(ranges.trans)) {
        colnames(ranges.trans[[i]]) <- res.cols
}
names(ranges.trans) <- paste("Кластер", names(ranges.trans))
```

Отчет по кластерам данных.

Кластер №1

```
as.data.frame(ranges.trans[[1]])
```

	Мин	Макс
Возраст, лет	20	28
Стаж вождения, лет	1	5
Убыточность, %	101	297
Уровень заработной платы, руб/год	602584	799461

Кластер №2

```
as.data.frame(ranges.trans[[2]])
```

	Мин	Макс
Возраст, лет	59	79
Стаж вождения, лет	33	58
Убыточность, %	92	209
Уровень заработной платы, руб/год	200043	279472

Кластер №3

```
as.data.frame(ranges.trans[[3]])
```

	Мин	Макс
Возраст, лет	22	38
Стаж вождения, лет	4	10
Убыточность, %	306	700
Уровень заработной платы, руб/год	120185	179863

Кластер №4

```
as.data.frame(ranges.trans[[4]])
```

	Мин	Макс
Возраст, лет	41	71
Стаж вождения, лет	20	40
Убыточность, %	20	116
Уровень заработной платы, руб/год	2040290	3982828

Визуализация взаимосвязи “Зарплаты” vs “Убыточность” по класстерам.

```
ggplot(dt.xls, aes(x = `Уровень заработной платы, руб/год` / 1000,
                   y = `Убыточность, %`,
                   group = clust4,
                   color = as.factor(clust4))) +
        labs(title = "Income / Loss",
             x = "Income in mln. rub / year",
             y = "Loss, %",
             color = "Cluster #\n") +
        theme_light() +
        geom_point()
```