HW11

xw-zeng

2022-12-20

载入 R 包。

```
library(readxl) # load excel data
library(ggplot2) # plot beautiful graphs
library(corrplot) # plot loadings
library(factoextra) # visualization of clustering
```

NBA 数据降维分析

分析任务 1

读入 NBA 数据集。

```
data <- read_excel('NBA.xlsx')
```

对数据集进行标准化，消除量纲的影响。

```
data_sc <- data.frame(scale(data[, -1]))
```

分析任务 2

使用数据相关系数矩阵的特征值绘制崖底碎石图，在第 3 个特征值后逐渐变得平缓，因此我们可以选择前 3 个主成分。

```
evalues = eigen(cor(data_sc))$values
ggplot(mapping = aes(x = 1:length(evalues), y = evalues)) +
  geom_line(linewidth = 0.6) + theme_light() +
  labs(title = 'Scree Plot', x = 'n', y = 'eigen value')
```

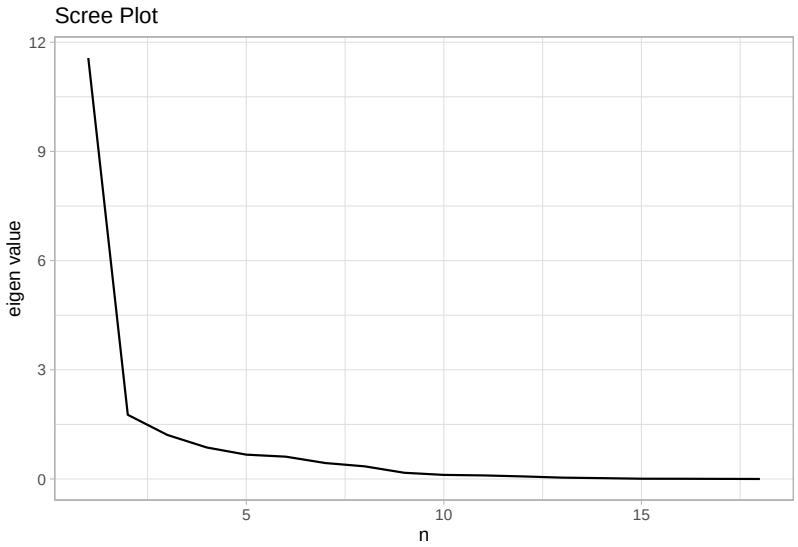

图 1: 崖底碎石图

对所有的自变量进行主成分分析，查看主成分的结果。前 3 个主成分的累计方差贡献率约为 80.78%，能够比较好地解释数据信息。

```
pca <- princomp(data_sc)
summary(pca)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     3.4002103 1.32777134 1.09987159 0.93006426 0.81746056
## Proportion of Variance 0.6425642 0.09798318 0.06723399 0.04807628 0.03713971
## Cumulative Proportion  0.6425642 0.74054734 0.80778134 0.85585762 0.89299733
##                           Comp.6     Comp.7     Comp.8      Comp.9     Comp.10
## Standard deviation     0.7826010 0.66183135 0.58897439 0.412905912 0.335182108
## Proportion of Variance 0.0340397 0.02434443 0.01927959 0.009475609 0.006244053
## Cumulative Proportion  0.9270370 0.95138146 0.97066105 0.980136656 0.986380709
##                           Comp.11     Comp.12     Comp.13     Comp.14
## Standard deviation     0.31327850 0.265702944 0.192276045 0.152756082
## Proportion of Variance 0.00545464 0.003923717 0.002054733 0.001296886
## Cumulative Proportion  0.99183535 0.995759067 0.997813799 0.999110686
##                            Comp.15      Comp.16     Comp.17      Comp.18
## Standard deviation     0.085461376 0.0757499972 0.054375173 1.657537e-03
## Proportion of Variance 0.000405924 0.0003189115 0.000164326 1.526974e-07
## Cumulative Proportion  0.999516610 0.9998355213 0.999999847 1.000000e+00
```

分析任务 3

查看前 3 个主成分的载荷，可以认为第一个主成分表示球员的综合属性；第二个主成分在三分投球率、三分命中次数、三分出手次数有很高的载荷，表示三分球的能力；第三个主成分在投篮率、罚球率、三分投球率上有很高的正载荷，在三分命中次数、三分出手次数上有较高的负载荷，表示投篮的效率。其中第二个主成分与三分球能力是负相关关系，第三主成分与投篮效率是负相关关系，因此我们考虑在接下来的分析中对第二、第三主成分得分取相反数，以更好地进行解释。

```
corrplot(t(pca$loadings[, 1:3]), is.corr=TRUE, number.cex=0.5, method='square',
         addCoef.col="grey30", tl.col="black", tl.cex=0.7, cl.cex=0.7, win.asp = 1.25)
```

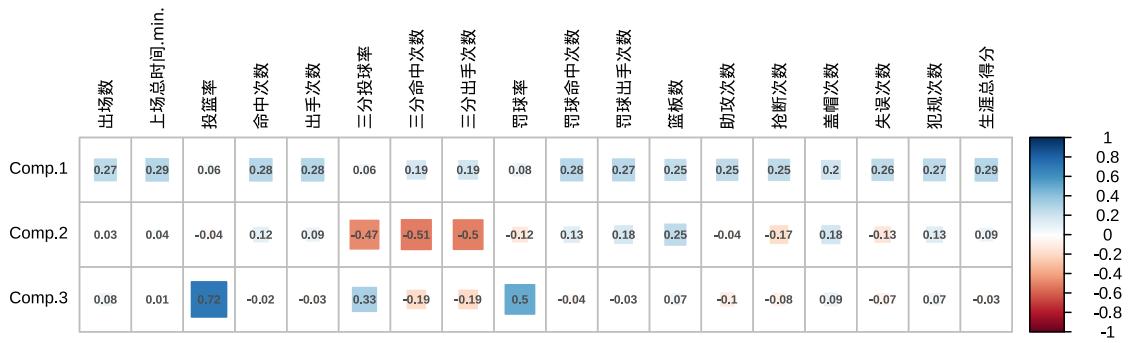

图 2: 前三个主成分载荷图

计算每一个球员的主成分得分。

```
pca_score = data.frame(data$球员, pca$scores[, 1:3])
colnames(pca_score) = c('球员', '综合能力', '三分球能力', '投篮效率')
pca_score$三分球能力 = - pca_score$三分球能力
pca_score$投篮效率 = - pca_score$投篮效率
head(pca_score)
```

```
##              球员 综合能力   三分球能力    投篮效率
## 1 勒布朗-詹姆斯 36.09607   4.5178213  5.6813132
## 2   迈克尔-乔丹 26.47592  -0.7996777  2.5200051
## 3 卡里姆-贾巴尔 20.23115  -9.0431130 -0.5897240
## 4 科比-布莱恩特 27.97238   3.8009598  4.1159774
## 5 沙奎尔-奥尼尔 25.64926 -10.3344415 -0.5986237
## 6     蒂姆-邓肯 27.07261  -9.8564313 -1.0496198
```

挑选几位我稍微听说过的球员进行分析。

```
pca_score[c(5, 36, 37, 364), ]
```

```
##              球员 综合能力 三分球能力    投篮效率
## 5   沙奎尔-奥尼尔 25.64926 -10.334441 -0.5986237
```

```
## 36     斯蒂芬-库里 13.90347  10.569519  4.0099620
## 37     詹姆斯-哈登 14.70380   6.344809  3.1918803
## 364          姚明  1.09382  -1.103508 -0.8258598
```

奥尼尔和姚明都是中锋，因此很少参与三分球的进攻，三分球能力的得分自然就要弱一些，而库里三分球众所周知非常厉害，因此得分特别高；就投篮效率而言，库里和哈登比较高；最后从综合能力看，奥尼尔的得分最高，其次是库里和哈登，最后是姚明。但姚明分数比较低可能是因为他打季后赛的次数比较少。

分析任务 4

首先查看最佳 kmeans 聚类数，发现聚为 4 类最为合适。

```
fviz_nbclust(pca_score[, -1], kmeans) + geom_vline(xintercept = 4, linetype = 2)
```

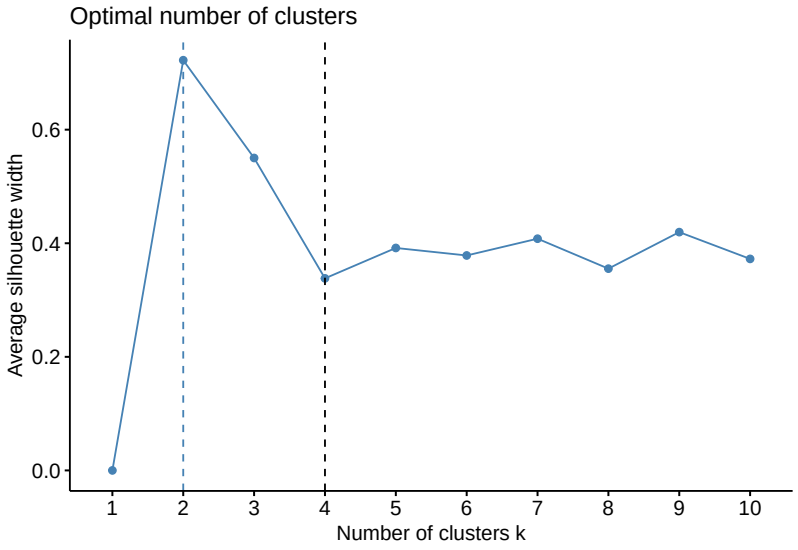

图 3: 目标函数与聚类个数折线图

使用主成分得分对 NBA 球员进行 K-means 聚类。

```
set.seed(5201314)
result <- kmeans(scale(pca_score[, -1]), iter.max = 1000, center = 4, nstart = 25)
```

对聚类结果进行可视化，可以发现区分效果是不错的。红色类别的球员三分球能力在 0 附近，综合能力为负，投篮效率较高，可以命名为轮换球员；绿色类别的球员综合能力较低，三分球能力为负，投篮效率也为负，可以命名为板凳球员；紫色类别的球员三分球能力、综合能力、投篮效率都较高，可以命名为进攻球员；蓝色类别的球员三分球能力为负，但综合能力较高，投篮效率一般，可以命名为防守球员。

```
fviz_cluster(result, pca_score[, -1], labelsize = 0,
             choose.vars = colnames(pca_score)[2:3]) + theme_light()
```

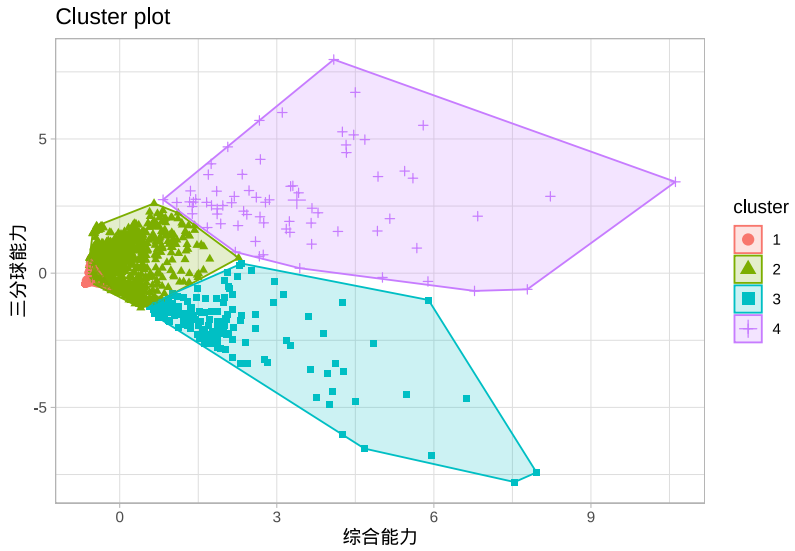

图 4: 综合能力与三分球能力

```
fviz_cluster(result, pca_score[, -1], labelsize = 0,
             choose.vars = colnames(pca_score)[c(2, 4)]) + theme_light()
```

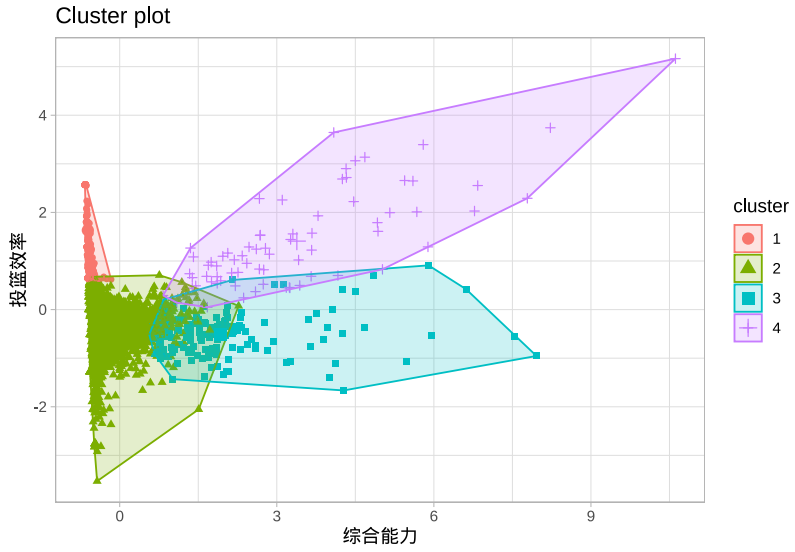

图 5: 综合能力与投篮效率

```
fviz_cluster(result, pca_score[, -1], labelsize = 0,
             choose.vars = colnames(pca_score)[c(3, 4)]) + theme_light()
```

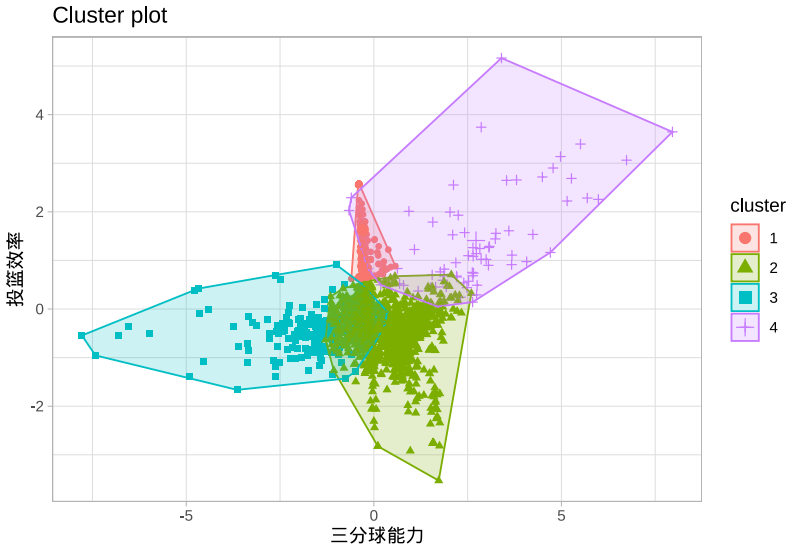

图 6: 三分球能力与投篮效率