1984年洛杉矶奥运会发布国家女子径赛记录如数据（主成分分析实验数据）所示，其中包含100m,200m,400m,800,1500m,3000m,Marathon(马拉松)的成绩以及国家作为指标变量，100m,200m,400m,的成绩是以秒为单位，其余成绩按分钟为单位。请对数据做如下分析：

**1、求得各项记录的样本相关阵。**

mydata<- read.csv("~/rstdio\_wsl/PRIN\_DATA.txt", sep="")

R=round(cor(mydata[,-8]),3)

R

X100m.s. X200m.s. X400m.min. X800m.min. X1500m.min. X3000m.min.

X100m.s. 1.000 0.953 0.835 0.728 0.728 0.742

X200m.s. 0.953 1.000 0.857 0.724 0.698 0.710

X400m.min. 0.835 0.857 1.000 0.898 0.788 0.778

X800m.min. 0.728 0.724 0.898 1.000 0.902 0.864

X1500m.min. 0.728 0.698 0.788 0.902 1.000 0.969

X3000m.min. 0.742 0.710 0.778 0.864 0.969 1.000

marathon.min. 0.686 0.686 0.705 0.779 0.878 0.900

marathon.min.

X100m.s. 0.686

X200m.s. 0.686

X400m.min. 0.705

X800m.min. 0.779

X1500m.min. 0.878

X3000m.min. 0.900

marathon.min. 1.000

**2、写出前两个主成分，为了解释标准化变量的变化情况，这两个主成分是否足够？**

PCA=princomp(mydata[,-8],cor=T)

summary(PCA,loadings=T)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6

Standard deviation 2.4094991 0.80848347 0.54761522 0.35422802 0.231984732 0.197608919

Proportion of Variance 0.8293837 0.09337793 0.04284035 0.01792536 0.007688131 0.005578469

Cumulative Proportion 0.8293837 0.92276161 0.96560196 0.98352731 0.991215445 0.996793914

Comp.7

Standard deviation 0.149808546

Proportion of Variance 0.003206086

Cumulative Proportion 1.000000000

Loadings:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

X100m.s. 0.368 0.490 0.286 0.319 0.231 0.620

X200m.s. 0.365 0.537 0.230 -0.711 -0.109

X400m.min. 0.382 0.247 -0.515 -0.347 -0.572 0.191 0.208

X800m.min. 0.385 -0.155 -0.585 0.620 -0.315

X1500m.min. 0.389 -0.360 0.430 -0.231 0.693

X3000m.min. 0.389 -0.348 0.153 0.363 -0.463 -0.598

marathon.min. 0.367 -0.369 0.484 -0.672 0.131 0.142

**两个成分累计贡献率已达到92%，故已足够**

**3、解释这两个主成分的含义。**

**第一主成分在各个指标上载荷值相当，反映了各国家运动员的综合实力；第二主成分在x1和x2上载荷较高，反映了各国家运动员的短期爆发力。**

**4、按照第一主成分的大小将55个国家排序，所排结果说明什么？**

scores=PCA$scores

scores1=as.data.frame(scores[,1])

country1=as.data.frame(mydata$country)

data4=cbind.data.frame(scores1,country1)

data4

order4=order(data4[,1])

data4[order4,2]

> data4[order4,2]

[1] gdr ussr usa czech frg gbni poland canada finland

[10] italy australi rumania france sweden netherla nz belgium norway

[19] hungary austria switzerl ireland denmark taipei kenya spain portugal

[28] israel brazil mexico japan columbia bermuda dprkorea argentin chile

[37] china greece india korea luxembou turkey philippi burma thailand

[46] singapor indonesi domrep malaysia costa guatemal png mauritiu cookis

[55] wsamoa

**所排情况表示各个国家运动员综合实力的排序**

**5、将55个国家或地区按照径赛成绩利用这两个主成分进行分类，给出大致的国家分类结果（可以用图示表示）。**

scores2=as.data.frame(scores[,2])

data5=cbind.data.frame(country1,scores1,scores2)

data5

> data5

mydata$country scores[, 1] scores[, 2]

1 argentin 0.53212195 -0.680937797

2 australi -2.11284700 -0.537709378

3 austria -1.39316656 -0.277530069

4 belgium -1.52390695 0.091802341

5 bermuda 0.39139200 -0.985408838

6 brazil -0.11948856 -0.919916756

7 burma 1.69754142 0.591594341

8 canada -2.63217070 -0.701696215

9 chile 0.55287936 1.182727959

10 china 0.64718368 0.989080326

11 columbia 0.14287696 0.155892166

12 cookis 6.13329130 1.412402548

13 costa 2.64336942 0.308776354

14 czech -3.08194526 -1.022064547

15 denmark -1.12666477 0.545109728

16 domrep 2.31659294 -0.621775472

17 finland -2.20194950 -0.676426245

18 france -1.90960964 -0.447922751

19 gdr -3.53833101 -1.213583461

20 frg -2.95274367 -0.441166540

21 gbni -2.80880784 -0.583680262

22 greece 0.82175014 0.235868928

23 guatemal 3.25704356 -0.927528451

24 hungary -1.49082853 0.059931588

25 india 1.02388750 0.255942992

26 indonesi 2.13183387 -0.381756355

27 ireland -1.12764937 0.517771847

28 israel -0.14428519 0.157303606

29 italy -2.15926044 0.355236135

30 japan -0.05977861 0.664038312

31 kenya -0.43486576 0.485040815

32 korea 1.24523372 0.821904700

33 dprkorea 0.46655772 1.733718848

34 luxembou 1.31374285 1.193070361

35 malaysia 2.36210756 -0.001271429

36 mauritiu 4.27286963 -1.191080643

37 mexico -0.06406697 0.574213788

38 netherla -1.81096545 -0.047519331

39 nz -1.52518788 0.381404100

40 norway -1.49667849 0.912528976

41 png 4.01755107 -0.343380195

42 philippi 1.65530683 -0.884117095

43 poland -2.69672473 -0.708796713

44 portugal -0.22635133 1.264207647

45 rumania -2.04853470 0.623743207

46 singapor 1.98828979 0.960756220

47 spain -0.35893084 0.934008388

48 sweden -1.84460096 -0.257169493

49 switzerl -1.35906564 0.518920084

50 taipei -0.50472889 -1.246032831

51 thailand 1.97117932 -0.141163113

52 turkey 1.62302658 0.599820487

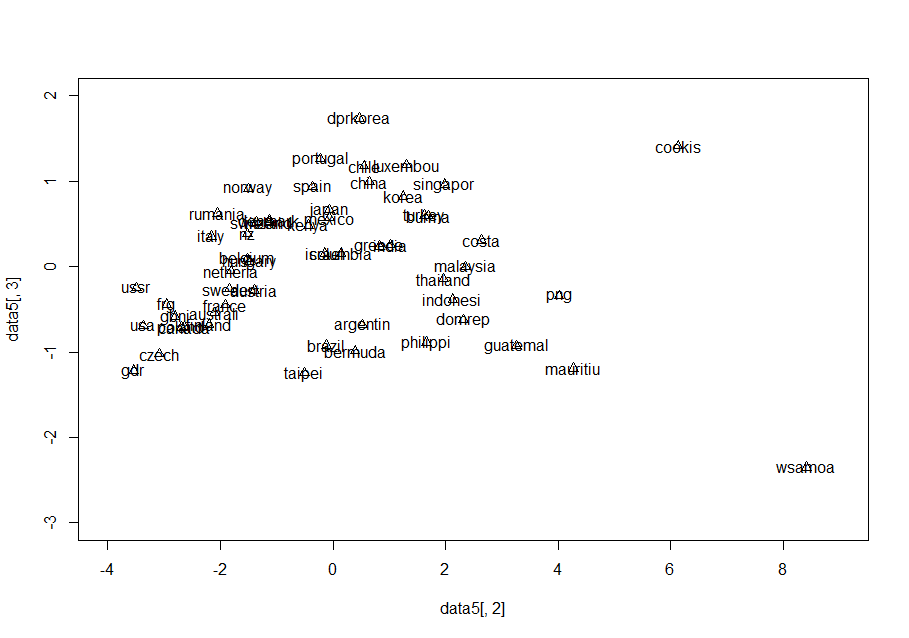
53 usa -3.36655736 -0.691419036

54 ussr -3.49662048 -0.247337282

55 wsamoa 8.40968394 -2.348426494

plot(data5[,2],data5[,3],xlim=c(-4,9),ylim=c(-3,2),pch=2)

text(data5[,2],data5[,3],data5[,1])



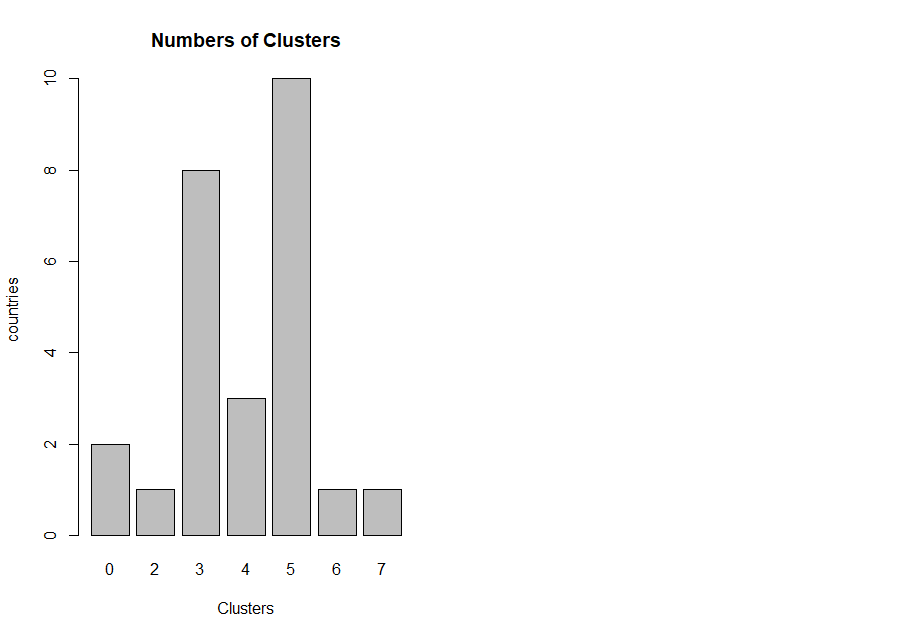
**接下来利用K-均值聚类法进行聚类**

library(NbClust)

df=scale(data5[-1])

nc=NbClust(df,distance="euclidean", min.nc=3, max.nc=7,method = 'kmeans')

print(table(nc$Best.n[1,]))

barplot(table(nc$Best.n[1,]),xlab='Clusters',ylab='countries',main='Numbers of Clusters')

**所以考虑分为5类**

|  |  |
| --- | --- |
| **1** | **australi austria brazil canada czech finland france gdr frg gbni poland sweden taipei usa ussr** |
| **2** | **Wsamoa** |
| **3** | **belgium columbia denmark greece hungary india ireland israel italy japan kenya mexico netherla nz norway rumania spain switzerl** |
| **4** | **argentin bermuda domrep guatemal indonesi malaysia mauritiu png philippi thailand** |
| **5** | **burma chile china cookis costa korea dprkorea luxembou Portugal singapor turkey** |

**6、利用你选定的主成分给出一个综合排名，说明该排名的意义。**

total=data5[,2]\*0.8294+data5[,3]\*0.0934

order2=order(total)

data5[order2,1]

综合排名：

[1] gdr ussr usa czech frg gbni poland canada finland

[10] australi italy rumania france sweden netherla belgium hungary nz

[19] austria norway switzerl ireland denmark taipei kenya spain brazil

[28] israel portugal mexico japan columbia bermuda argentin dprkorea chile

[37] china greece india korea luxembou philippi turkey burma thailand

[46] indonesi singapor domrep malaysia costa guatemal png mauritiu cookis

[55] wsamoa

**排名表明各国家女子径赛运动员的成绩从强到弱。其中gdr成绩最好，Wsamoa成绩最差。**

7**、判断是否有多重共线性，如果有应该如何利用主成分法消除多重共线性？**

> kappa(R)

[1] 153.4519

**条件数>100，有较强多重共线性**

**主成分回归法：**

**求由解释变量构成的主成分；选择前K个最重要的主成分，它能够反映解释变量的大部分(>90%)信息;用选择的主成分作为解释变量对被解释变量进行参数估计和检验，确定参数值；用通过主成分估计出的参数值通过变换求出原解释变量的参数值。**