4.2رتبه بندی دانشگاه ها.

مجموعه داده‌های رتبه‌بندی کالج‌ها و دانشگاه‌های آمریکا) موجود در (www.dataminingbook.com حاوی اطلاعاتی درباره ۱۳۰۲ کالج و دانشگاه آمریکایی است که برنامه‌ای در مقطع کارشناسی ارائه می‌کنند.

برای هر دانشگاه، ۱۷ اندازه گیری وجود دارد که شامل اندازه گیری های مستمر (مانند شهریه و نرخ فارغ التحصیلی) و اندازه گیری های طبقه بندی شده (مانند مکان بر اساس ایالت و خصوصی یا دولتی بودن مدرسه) است.

آ. تمام متغیرهای طبقه بندی را حذف کنید. سپس تمام رکوردهای دارای اندازه گیری های عددی از دست رفته را از مجموعه داده حذف کنید.

ب تجزیه و تحلیل اجزای اصلی را روی داده های پاک شده انجام دهید و در مورد نتایج نظر دهید.

آیا داده ها باید نرمال شوند؟ در مورد آنچه که مؤلفه‌هایی را که کلید می‌دانید مشخص می‌کند، بحث کنید

---

a.

ابتدا داده‌ها را فراخوانی میکنیم

```
> Universities <-read.csv("~/Universities.csv",header = T)
```

با دستور زیر میتوانیم اطلاعاتی از داده‌ها بدست آوریم

```
> str(Universities)
'data.frame':    1302 obs. of  20 variables:
 $ College.Name             : Factor w/ 1274 levels "Abilene Christian University",..
 $ State                    : Factor w/ 51 levels "AK","AL","AR",..: 1 1 1 1 2 2 2 2
 $ Public..1...Private..2.  : int  2 1 1 1 1 2 1 1 1 2 ...
 $ X..appli..rec.d          : int  193 1852 146 2065 2817 345 1351 4639 7548 805 ...
 $ X..appl..accepted        : int  146 1427 117 1598 1920 320 892 3272 6791 588 ...
 $ X..new.stud..enrolled    : int  55 928 89 1162 984 179 570 1278 3070 287 ...
 $ X..new.stud..from.top.10.: int  16 NA 4 NA NA NA 18 NA 25 67 ...
 $ X..new.stud..from.top.25.: int  44 NA 24 NA NA 27 78 NA 57 88 ...
 $ X..FT.undergrad          : int  249 3885 492 6209 3958 1367 2385 4051 16262 1376 .
 $ X..PT.undergrad          : int  869 4519 1849 10537 305 578 331 405 1716 207 ...
 $ in.state.tuition         : int  7560 1742 1742 1742 1700 5600 2220 1500 2100 11660
 $ out.of.state.tuition     : int  7560 5226 5226 5226 3400 5600 4440 3000 6300 11660
 $ room                     : int  1620 1800 2514 2600 1108 1550 NA 1960 NA 2050 ...
 $ board                    : int  2500 1790 2250 2520 1442 1700 NA NA NA 2430 ...
 $ add..fees                : int  130 155 34 114 155 300 124 84 NA 120 ...
 $ estim..book.costs        : int  800 650 500 580 500 350 300 500 600 400 ...
 $ estim..personal..        : int  1500 2304 1162 1260 850 NA 600 NA 1908 900 ...
 $ X..fac..w.PHD            : int  76 67 39 48 53 52 72 48 85 74 ...
 $ stud..fac..ratio         : num  11.9 10 9.5 13.7 14.3 32.8 18.9 18.7 16.7 14 ...
 $ Graduation.rate          : int  15 NA 39 NA 40 55 51 15 69 72 ...
```

از این خروجی اطلاعات زیر استخراج میشود :

داده‌های ما شامل ۲۰ متغیر برای ۱۳۰۲ مشاهده هستند.

در دیتای ما داده گمشده وجود دارد.

متغیرهایی مثل مکان بر اساس ایالت و خصوصی یا دولتی بودن مدرسه(**State , Public.1.Private.2**)متغیرهایی از جنس **categorical** هستند که ما باید انها را حذف کنیم و بقیه متغیرهای ما متغیرهایی عددی هستند.

اکنون داده‌های گمشده و متغیرهای طبقه بندی را از دیتای خود حذف میکنیم :

```
> # remove all records with missing numerical measurements from the dataset
> Universities.NAomit<-na.omit(Universities)

> # Remove all categorical variables
> Universities<-Universities[,-c(1,2,3)]
>
> # remove all records with missing numerical measurements from the dataset
> Universities.NAomit<-na.omit(Universities)

> str(Universities.NAomit)
'data.frame':    471 obs. of  17 variables:
 $ X..appli..rec.d          : int  193 146 805 608 4414 1797 708 823 605 1721 ...
 $ X..appl..accepted        : int  146 117 588 520 1500 1260 334 721 405 1068 ...
 $ X..new.stud..enrolled    : int  55 89 287 127 335 938 166 274 284 806 ...
 $ X..new.stud..from.top.10.: int  16 4 67 26 30 24 46 52 24 35 ...
 $ X..new.stud..from.top.25.: int  44 24 88 47 60 35 74 87 53 75 ...
 $ X..FT.undergrad          : int  249 492 1376 538 908 6960 530 954 961 3128 ...
 $ X..PT.undergrad          : int  869 1849 207 126 119 4698 182 6 99 213 ...
 $ in.state.tuition         : int  7560 1742 11660 8080 5666 2220 8644 8800 6398 5504
 $ out.of.state.tuition     : int  7560 5226 11660 8080 5666 4440 8644 8800 6398 5504
 $ room                     : int  1620 2514 2050 1380 1424 1935 2382 1935 1450 1650
 $ board                    : int  2500 2250 2430 2540 1540 3240 1540 1260 2222 1878
 $ add..fees                : int  130 34 120 100 418 291 120 325 148 1016 ...
 $ estim..book.costs        : int  800 500 400 500 1000 750 500 500 400 700 ...
 $ estim..personal..        : int  1500 1162 900 1100 1400 2200 800 1200 1350 910 ...
 $ X..fac..w.PHD            : int  76 39 74 63 56 96 79 82 68 71 ...
 $ stud..fac..ratio         : num  11.9 9.5 14 11.4 15.5 6.7 12.6 13.1 13.3 17.7 ...
 $ Graduation.rate          : int  15 39 72 44 46 33 54 63 75 73 ...
 - attr(*, "na.action")= 'omit' Named int  2 4 5 6 7 8 9 11 13 14 ...
  ..- attr(*, "names")= chr  "2" "4" "5" "6" ...
```

اکنون دیتای ما شامل**17** متغیر عددی برای ۴۷۱ مشاهده میباشد.(در اینجا متغیر اول ما نام دانشگاه بود که ان را نیز حذف کردیم).

در این قسمت ما تحلیل مولفه اصلی **PCA** را بر روی داده‌های که نرمال سازی نشده‌اند اجرا میکنیم:

```
> # PCA on non-normal data
```

```
> pca.nonnormal <- prcomp(Universities.NAomit)
> summary(pca.nonnormal)
Importance of components:
                            PC1    PC2          PC3     PC4     PC5
Standard deviation     7430.9140 5987.9890 1.855e+03 1.193e+03 967.42792
Proportion of Variance    0.5614    0.3645 3.497e-02 1.446e-02   0.00951
Cumulative Proportion     0.5614    0.9259 9.609e-01 9.753e-01   0.98484

                           PC6      PC7      PC8         PC9       PC10
Standard deviation       679.6527 596.97612 580.62990 417.61364 318.12719
Proportion of Variance   0.0047   0.00362   0.00343   0.00177   0.00103
Cumulative Proportion    0.9895   0.99316   0.99658   0.99836   0.99938

                          PC11     PC12      PC13   PC14  PC15   PC16  PC17
Standard deviation     188.86761 155.606   19.05  12.53  11.02  5.33  2.906
Proportion of Variance 0.00036   0.00025   0.00   0.00   0.00   0.00  0.000
Cumulative Proportion  0.99975   0.99999   1.00   1.00   1.00   1.00  1.000

> pca.nonnormal$rot[,1:2]
                                    PC1           PC2
X..appli..rec.d            2.718826e-01  0.5511833876
X..appl..accepted          1.941070e-01  0.3212993731
X..new.stud..enrolled      8.472979e-02  0.1015899308
X..new.stud..from.top.10. -8.984730e-04  0.0017322347
X..new.stud..from.top.25. -8.113414e-04  0.0019247328
X..FT.undergrad            4.581211e-01  0.4922634131
X..PT.undergrad            1.082532e-01  0.0734095353
in.state.tuition          -6.701873e-01  0.3824891315
out.of.state.tuition      -4.545345e-01  0.4286850581
room                      -3.342006e-02  0.0555839852
board                     -3.423588e-02  0.0408973641
add..fees                  1.320940e-02  0.0087460804
estim..book.costs         -5.792354e-05  0.0032905678
estim..personal..          3.755717e-02  0.0011851103
X..fac..w.PHD             -2.046899e-04  0.0015640592
stud..fac..ratio           2.954376e-04 -0.0001587084
Graduation.rate           -1.072320e-03  0.0013974456


> which.max(pca.nonnormal$rot[,1])
X..FT.undergrad
> max(pca.nonnormal$rot[,1])
[1] 0.4581211


> which.max(pca.nonnormal$rot[,2])
X..appli..rec.d
> max(pca.nonnormal$rot[,2])
[1] 0.5511834
```

طبق خروجی بالا مشخص میشود که دو مولفه اصلی اول یعنی **PC1** و **PC2** ، ۹۲ ٪ از واریانس کل را نمایش میدهند.

بیشتر از **PCA** برای درک ساختار داده ها استفاده میشود

این کار با بررسی وزن ها انجام می شود تا ببینید که چگونه متغیرهای اصلی به اجزای اصلی مختلف کمک می کنند.

در مثال ما، واضح است که اولین مولفه اصلی یعنی **PCA1** تحت سلطه محتوای **X..FT.undergrad** است، چون بالاترین مقدار واریانس را دارد یعنی ۰.۴۵۸ .

به طور مشابه، دومین مولفه اصلی به نظر می رسد تحت تسلط **X..appli..rec.d** است.چون بالاترین مقدار واریانس را دارد یعنی ۰.۵۵ .

دلیل این اتفاق این است که متغیرهای ما نرمال سازی نشده‌اند ، یعنی مقیاس‌های متفاوتی با یکدیگر دارند که این باعث بروز مشکلاتی میشود و زیاد مطلوب نیست.

یک راه حل این است که داده ها را قبل از انجام **PCA** نرمال کنید.

اکنون داده‌های خود را نرمال میکنیم سپس به تحلیل مولفه اصلی میپردازیم :

```
> # PCA on normal data
>
> pca.normal <- prcomp(Universities.NAomit,scale. = T)
> summary(pca.normal)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     2.2749 2.1426 1.09838 1.03247 0.97599
Proportion of Variance 0.3044 0.2700 0.07097 0.06271 0.05603
Cumulative Proportion  0.3044 0.5745 0.64542 0.70813 0.76416

                       PC6     PC7     PC8     PC9   PC10    PC11    PC12   PC13

Standard deviation      0.87284 0.80327 0.77279 0.70316 0.6622 0.62788 0.54973 0.4383
Proportion of Variance  0.04481 0.03796 0.03513 0.02908 0.0258 0.02319 0.01778 0.0113
Cumulative Proportion   0.80898 0.84693 0.88206 0.91115 0.9369 0.96013 0.97791 0.9892




                         PC14    PC15    PC16    PC17
Standard deviation     0.30389 0.20002 0.17428 0.14388
Proportion of Variance 0.00543 0.00235 0.00179 0.00122
Cumulative Proportion  0.99464 0.99700  0.99878 1.00000


> pca.normal$rot[,1:6]
                            PC1      PC2         PC3      PC4        PC5         PC6
X..appli..rec.d           0.07836 -0.42016  0.031982 -0.07262  0.016693 -0.112319
X..appl..accepted         0.023658 -0.434471  0.031422 -0.11812  0.08907 -0.11438
X..new.stud..enrolled     -0.02880 -0.4455  0.03865  0.03146  0.0759812 -0.0540786
```

```
X..new.stud..from.top.10.  0.354028 -0.09354  0.120128  0.37245 -0.16225  0.004445
X..new.stud..from.top.25.  0.34049  -0.11839  0.142719  0.38556 -0.15818 -0.092636
X..FT.undergrad           -0.04958  -0.44358  0.004012  0.05645  0.094780 -0.04350
X..PT.undergrad           -0.1063   -0.28769 -0.265769 -0.05349  0.343680  0.18804
in.state.tuition           0.37938   0.15024 -0.084350 -0.04106  0.172639  0.00053
out.of.state.tuition       0.40255   0.04872 -0.051577 -0.07765  0.158498  0.04440
room                       0.2731   -0.05227 -0.250577 -0.45441  0.004482  0.015063
board                      0.29043  -0.01005 -0.252096 -0.3016   0.199066  0.038473
add..fees                 -0.012350 -0.16949  0.249746 -0.44656 -0.648919  0.418437
estim..book.costs          0.05730  -0.05668 -0.652240  0.04435 -0.518    -0.421195
estim..personal..         -0.14490  -0.15683 -0.403735  0.40370 -0.103358  0.466598
X..fac..w.PHD              0.254200 -0.19685  0.189366  0.07460  0.017277  0.1806230
stud..fac..ratio          -0.278542 -0.10103  0.187598 -0.10522 -0.002999 -0.5221543
Graduation.rate            0.32530  -0.02426  0.181888  0.01260 -0.109137 -0.2153249
```

الان که نرمال سازی را انجام دادیم میبینم که ۶ مولفه اصلی اول ۸۰٪ از واریانس کل را نمایش میدهند. و هیچ کدام از مولفه‌های اصلی تحت تسلط متغیرهای خاصی نیستند، چون در اینجا با نرمال سازی و هم مقیاس سازی داده‌ها واریانس را کنترل کردیم و متغیری پیدا نمیشود که دارای وارایانس خیلی بزرگتر از بقیه باشد