STAT 43000/STAT 53001 Applied Statistics Spring 2023

Lab 5

Q.N. 1) A bank requires eight pieces of information from loan applicants: income, education level, age, length of time at current residence, length of time with current employer, savings, debt, and number of credit cards. A bank administrator wants to analyze this data to determine the best way to group and report it. The administrator collects this information for 30 loan applicants. The data is provided with this assignment

a. Import the data in R and print first 3 observations

```
> file.choose()
[1] "C:\\Users\\Zhang\\Downloads\\Loandata.xlsx"
> file.choose()
[1] "C:\\Users\\Zhang\\Desktop\\Loandata.csv"
> data=read.csv("C:\\Users\\Zhang\\Desktop\\Loandata.csv"
+ )
> head(data,3)
  Income Education Age Residence Employ Savings Debt Credit.cards
1  50000        16  28         2      2    5000 1200            2
2  72000        18  35        10      8   12000 5400            4
3  61000        18  36         6      5   15000 1000            2
> attach(data)
> head(data,3)
  Income Education Age Residence Employ Savings Debt Credit.cards
1  50000        16  28         2      2    5000 1200            2
2  72000        18  35        10      8   12000 5400            4
3  61000        18  36         6      5   15000 1000            2
```

b. Calculate the correlation matrix of the dataset

```
> cor(data)
                Income   Education         Age  Residence     Employ
Income      1.00000000  0.54871455  0.5150917 0.34711970 0.33377154
Education   0.54871455  1.00000000  0.2292770 0.10774778 0.04853683
Age         0.51509166  0.22927699  1.0000000 0.83771879 0.84802073
Residence   0.34711970  0.10774778  0.8377188 1.00000000 0.95221573
Employ      0.33377154  0.04853683  0.8480207 0.95221573 1.00000000
Savings     0.21010533  0.44735911  0.5524199 0.57004382 0.53879518
Debt       -0.19616432 -0.45658660  0.0323862 0.18581520 0.24667868
Credit.cards -0.05864313 -0.29617822 -0.1297580 0.05348158 0.02254348
               Savings       Debt Credit.cards
Income      0.2101053 -0.1961643  -0.05864313
Education   0.4473591 -0.4565866  -0.29617822
Age         0.5524199  0.0323862  -0.12975798
Residence   0.5700438  0.1858152   0.05348158
Employ      0.5387952  0.2466787   0.02254348
Savings     1.0000000 -0.3925015  -0.41046649
Debt       -0.3925015  1.0000000   0.47431532
Credit.cards -0.4104665  0.4743153   1.00000000
> 
```

c. Calculate the eigen values and eigen vectors of the correlation matrix

```
> eigen(cor(data))
eigen() decomposition
$values
[1] 3.54756813 2.13199433 1.04473533 0.53151226 0.41120264 0.16648803 0.12535593
[8] 0.04114334

$vectors
            [,1]        [,2]         [,3]        [,4]        [,5]        [,6]
[1,] -0.31390136  0.1446448  0.675860065  0.34694884  0.24134086  0.49387841
[2,] -0.23696085  0.4441021  0.400866616 -0.23983661 -0.62218299 -0.35695012
[3,] -0.48396049 -0.1350957  0.004136895  0.21151885  0.17495250 -0.48732954
[4,] -0.46639233 -0.2765863 -0.090660779 -0.11572183  0.03528727 -0.08519968
[5,] -0.45892284 -0.3044570 -0.121750395  0.01705887  0.01437731 -0.02308044
[6,] -0.40406743  0.2189397 -0.366484774 -0.43575313 -0.14314247  0.56809368
[7,]  0.06722312 -0.5850546  0.078009163  0.28096218 -0.68126042  0.24533067
[8,]  0.12319220 -0.4518659  0.468041622 -0.70346400  0.19487715 -0.02171165
            [,7]        [,8]
[1,] -0.01774284  0.03012301
[2,] -0.10308214 -0.05706354
[3,]  0.65725501  0.05187928
[4,] -0.48728920  0.66212713
[5,] -0.36786557 -0.73859130
[6,]  0.34796989  0.01731438
[7,]  0.19568741  0.07463574
[8,]  0.15786914 -0.05783743

> 
```

d. Find the principal components and determine the principal components if one wish to capture 90% variance.

```
> fit=prcomp(data,scale=TRUE)
> fit
Standard deviations (1, .., p=8):
[1] 1.8834989 1.4601350 1.0221230 0.7290489 0.6412508 0.4080294 0.3540564
[8] 0.2028382

Rotation (n x k) = (8 x 8):
                    PC1         PC2          PC3         PC4         PC5
Income       -0.31390136 -0.1446448  0.675860065  0.34694884  0.24134086
Education    -0.23696085 -0.4441021  0.400866616 -0.23983661 -0.62218299
Age          -0.48396049  0.1350957  0.004136895  0.21151885  0.17495250
Residence    -0.46639233  0.2765863 -0.090660779 -0.11572183  0.03528727
Employ       -0.45892284  0.3044570 -0.121750395  0.01705887  0.01437731
Savings      -0.40406743 -0.2189397 -0.366484774 -0.43575313 -0.14314247
Debt          0.06722312  0.5850546  0.078009163  0.28096218 -0.68126042
Credit.cards  0.12319220  0.4518659  0.468041622 -0.70346400  0.19487715
                    PC6          PC7          PC8
Income        0.49387841 -0.01774284  0.03012301
Education    -0.35695012 -0.10308214 -0.05706354
Age          -0.48732954  0.65725501  0.05187928
Residence    -0.08519968 -0.48728920  0.66212713
Employ       -0.02308044 -0.36786557 -0.73859130
Savings       0.56809368  0.34796989  0.01731438
Debt          0.24533067  0.19568741  0.07463574
Credit.cards -0.02171165  0.15786914 -0.05783743
> 
```

e. Interpret the coefficient and sign of PC1 for different variables.

While Debt and Credit.cards have positive loading on PC1, others have negative loading on PC1. The bigger the abs value is, the larger association the variable have with PC1. Here the largest two are Age(-0.487) and Savings(0.568) after do round(eigen(cor(data))$vectors,3)