Class Assignment 2- PCA

Problem 1- Use the TTU student evalauations data. See the attached file for an explanation of the variables. We analyze the result of 10 questions.

```
evaluation <- read.csv("https://raw.githubusercontent.com/asheikhz2/TTU_Zadeh/main/evals.csv")
mydata <- evaluation[,3:12]    # select variables to use
mydata <- na.omit(mydata) # you can do listwise cleaning for missing values
head(mydata)
```

```
##   RESP_1 RESP_2 RESP_3 RESP_4 RESP_5 RESP_6 RESP_7 RESP_8 RESP_9 RESP_10
## 1      4      5      4      4      4      4      4      4      4       4
## 2      4      5      5      5      5      5      4      5      5       5
## 3      4      5      5      5      5      5      5      5      5       5
## 4      5      5      5      5      5      5      5      5      5       5
## 5      5      5      5      5      5      5      5      5      5       5
## 6      4      4      4      4      4      4      4      4      4       4
```

a) Perform principal component analysis.

```
mydata_pca <- princomp(mydata, cor = T)
summary(mydata_pca, loading = T)
```

```
## Importance of components:
##                            Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     2.6694722 0.83361292 0.71509032 0.61390139 0.54649208
## Proportion of Variance 0.7126082 0.06949105 0.05113542 0.03768749 0.02986536
## Cumulative Proportion  0.7126082 0.78209923 0.83323464 0.87092213 0.90078749
##                            Comp.6    Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     0.51858543 0.4868983 0.42849089 0.40716864 0.36977489
## Proportion of Variance 0.02689309 0.0237070 0.01836044 0.01657863 0.01367335
## Cumulative Proportion  0.92768058 0.9513876 0.96974802 0.98632665 1.00000000
##
## Loadings:
##         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
## RESP_1   0.333  0.228         0.196  0.227  0.141  0.491  0.109  0.669  0.159
## RESP_2   0.273 -0.168 -0.929  0.105 -0.114
## RESP_3   0.334  0.171         0.157  0.325  0.431  0.219  0.258 -0.641 -0.136
## RESP_4   0.314 -0.501  0.131         0.297 -0.289                -0.153  0.651
## RESP_5   0.311 -0.530  0.228  0.128  0.154                        0.168 -0.698
## RESP_6   0.316 -0.258  0.194        -0.689  0.528                       0.174
## RESP_7   0.321  0.315  0.160  0.376 -0.243 -0.280        -0.665 -0.205
## RESP_8   0.325  0.313               -0.295 -0.468 -0.250  0.641
## RESP_9   0.321  0.306        -0.241  0.312  0.261 -0.705 -0.215  0.187
## RESP_10  0.311               -0.830        -0.239  0.357 -0.111
```

b) How many components represents more than 75% of the variability of the data? Answer : Two components represents more then 75% , Com1 : 71.26% , Com2: 6.95%

c) Attach a meaning to the PCs that cover more than 75% of the variablity.

Answer : For com1: Students expressed strong satisfaction with how the instructor encouraged their learning. For com2: Students are highly satisfied that the instructor presented the information clearly. However, they strongly disagree that the instructor treated all students with respect. **Problem 2**- Use the US state crime data.

```
crime <- read.csv("https://raw.githubusercontent.com/asheikhz2/TTU_Zadeh/main/crime.csv", row.names = "S
head(crime)
```

```
##            MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY  AUTO
## ALABAMA      14.2 25.2    96.8   278.3   1135.5  1881.9 280.7
## ALASKA       10.8 51.6    96.8   284.0   1331.7  3369.8 753.3
## ARIZONA       9.5 34.2   138.2   312.3   2346.1  4467.4 439.5
## ARKANSAS      8.8 27.6    83.2   203.4    972.6  1862.1 183.4
## CALIFORNIA   11.5 49.4   287.0   358.0   2139.4  3499.8 663.5
## COLORADO      6.3 42.0   170.7   292.9   1935.2  3903.2 477.1
```

a) Perform PCA using correlation matrix.

```
crime_PCA <- princomp(crime, cor = T)
summary(crime_PCA, loading = T)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.0285363 1.1129788 0.8519487 0.56252293 0.50791186
## Proportion of Variance 0.5878514 0.1769603 0.1036881 0.04520458 0.03685349
## Cumulative Proportion  0.5878514 0.7648116 0.8684997 0.91370429 0.95055778
##                            Comp.6     Comp.7
## Standard deviation     0.47121064 0.35221592
## Proportion of Variance 0.03171992 0.01772229
## Cumulative Proportion  0.98227771 1.00000000
##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## MURDER    0.300  0.629  0.178  0.232  0.538  0.259  0.268
## RAPE      0.432  0.169 -0.244         0.188 -0.773 -0.296
## ROBBERY   0.397         0.496  0.558 -0.520 -0.114
## ASSAULT   0.397  0.344        -0.630 -0.507  0.172  0.192
## BURGLARY  0.440 -0.203 -0.210         0.101  0.536 -0.648
## LARCENY   0.357 -0.402 -0.539  0.235                0.602
## AUTO      0.295 -0.502  0.568 -0.419  0.370         0.147
```

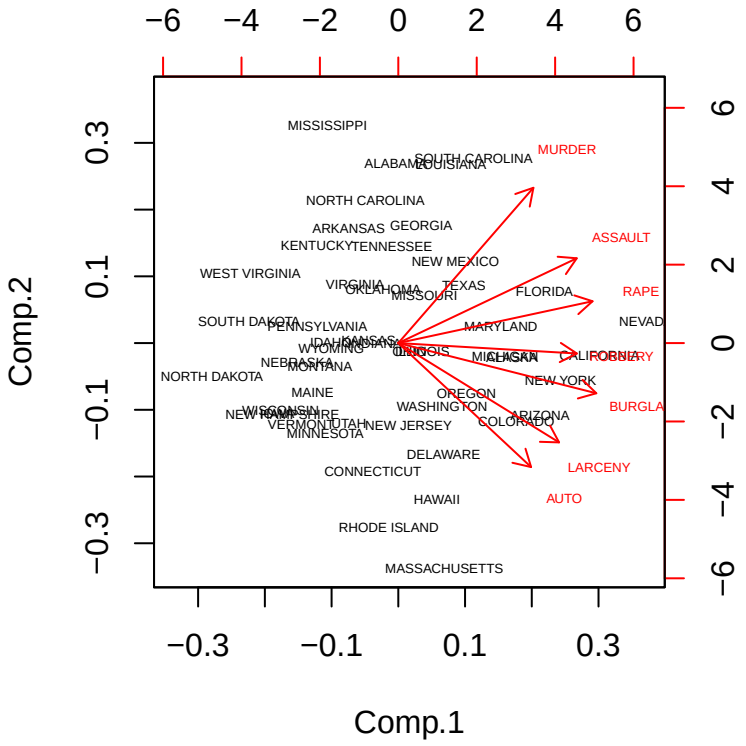
b) How many components represents the most variability of the data? Answer : Two components represents more then 75% , Com1 : 58.79% , Com2: 17.7%

c) By looking at the loadings, determine a meaning for PC1 and PC2 dimensions. Answer: for PC1: All these variables have positive loadings, meaning they all contribute in the same direction to Comp.1. The crimes RAPE, ROBBERY, ASSAULT, and BURGLARY have the highest loadings on this component.

for PC2: MURDER has the highest positive loading, while AUTO has the most negative. This indicates that areas with a higher value on Comp.2 might have higher murder rates but lower auto theft rates.

d) Construct the biplot graph of the crime data. interpret the resulting biplot graph for "MISSISSIPI", "NEVADA", and "HAWAII". You can validate your conclusions by looking at the actual standardized data values.

```
biplot(crime_PCA, col = c("Black", "red"), cex = .45)
```



MISSISSIPPI:

Located in the upper left quadrant of the plot. Closeness to vectors "MURDER" and "ASSAULT" indicates that MISSISSIPPI likely has higher standardized values for these crimes relative to its average. The state's position suggests that the occurrences of murder and assault might be higher in MISSISSIPPI compared to other states.

NEVADA:

Located in the right-middle part of the plot. Closest to the "ROBBERY" vector and not too far from the "RAPE" vector, implying that NEVADA might have a higher occurrence of robberies and rapes relative to its average. This positioning suggests that robberies (and to some extent, rapes) might be more prevalent in NEVADA compared to other states in the dataset.

HAWAII:

Located in the bottom right quadrant. Closeness to the vectors "LARCENY" and "AUTO" (presumably auto theft) suggests that HAWAII likely has higher standardized values for these crimes relative to its average. HAWAII might have a higher occurrence of larceny and auto theft compared to many other states in the dataset.