Homework 2

Chance Kang / A13605546 / csk025@ucsd.edu
Jayden Kim / A16271107 / s0k003@ucsd.edu
Sia Sheth / A16357789 / snsheth@ucsd.edu
Math 189
Spring 2023

Conceptual Question

1. Principal Component Analysis (PCA) takes multiple categories and sum them up into new "components" so that one can observe data in less complex dimensions (2-D,3-D). PCA chooses these categories that minimizes that data loss and maximizes the data separability. One can conceptualize PCA as the following: there are multiple measurements taken account for making a t-shirt. PCA takes these measurements and produces new components sizes S, M, and L.

Application Question

1a.

```
data(USArrests)
states <- row.names(USArrests)
arrests <- USArrests
pcaScale <- prcomp(arrests, scale = TRUE)
pcaScale$rotation
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

1b. From part a, the largest coefficients of PC1 responds to the three categories: Murder, Assault and Rape. Hence, high PC1 score may suggest high rate of violent crime or crime against persons.
The largest coefficients of PC2 responds to Urban Population, hence high PC2 score directly correlates to the urban population of the state.
Note that both PC1, PC2 have negative coefficient, meaning lesser the value of PC1, PC2, the higher the score of PC1, PC2.
1c.

```
pcaScaleVar <-pcaScale$sdev^2
pveScale<-pcaScaleVar/sum(pcaScaleVar)
pveScale
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

2a.

```
pcaScaleNot <- prcomp(arrests, scale = FALSE)
pcaScaleNot$rotation
```

```
##                 PC1         PC2         PC3         PC4
## Murder   0.04170432 -0.04482166  0.07989066 -0.99492173
## Assault  0.99522128 -0.05876003 -0.06756974  0.03893830
## UrbanPop 0.04633575  0.97685748 -0.20054629 -0.05816914
## Rape     0.07515550  0.20071807  0.97408059  0.07232502
```

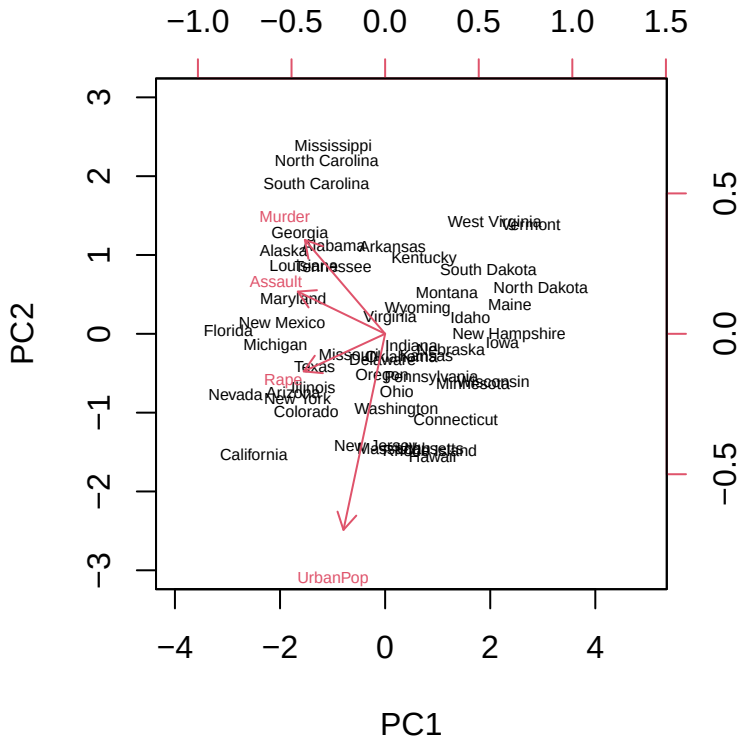
2b. The same argument as Part 1 is used to derive the following: PC1 responds to Assault, PC2 responds to Urban Population. High PC1 score of a state means higher arrests of murder in the state, high PC2 score of a state means higher percent of urban population in the state. Such is because the data set was not scaled before performing PCA.
2c.

```
pcaScaleNotVar <-pcaScaleNot$sdev^2
pveScaleNot<-pcaScaleNotVar/sum(pcaScaleNotVar)
pveScaleNot
```

```
## [1] 0.9655342206 0.0278173366 0.0057995349 0.0008489079
```

3. Results from part 1 and 2 seem to include all of the original variables among the PCs. First, PC2 of part 1 represents the Urban Population variable and PC1 represents the rest of the variables. Each PC (1~4) of part 2 seems to represent a single variable. However, the proportion of variability explained in PC1 of part 1 takes 62% and PC2 takes 24% of the PVE whereas PC1 of part 2 takes up about 96% of the PVE. Hence, PCA of part 2 cares mostly about only the Assault variable. Such is because PCA maximizes variance and some variables may be prioritized for its high magnitude (hence high variance). The USArrests data set shows relatively higher arrests of assault compared to the rest of the variables. Hence, part 2, PCA without scaling, showed PCA prioritizing the Assault variable (for it was represented by PC1). Overall, part 1 successfully reduces the dimension of USArrests data set with substantial results from PCA with scaled data whereas part 2 shows misleading results from PCA with absence of scaling data.
4.

```
biplot(pcaScale, scale = 0,expand = 2, cex = 0.5, xlim = c(-4,5), ylim = c(-3,3))
```




```
USArrests[,c(1,3)]
```

```
##                Murder UrbanPop
## Alabama          13.2       58
## Alaska           10.0       48
## Arizona           8.1       80
## Arkansas          8.8       50
## California        9.0       91
## Colorado          7.9       78
## Connecticut       3.3       77
## Delaware          5.9       72
## Florida          15.4       80
## Georgia          17.4       60
## Hawaii            5.3       83
## Idaho             2.6       54
## Illinois         10.4       83
## Indiana           7.2       65
## Iowa              2.2       57
## Kansas            6.0       66
## Kentucky          9.7       52
## Louisiana        15.4       66
## Maine             2.1       51
## Maryland         11.3       67
## Massachusetts     4.4       85
## Michigan         12.1       74
## Minnesota         2.7       66
## Mississippi      16.1       44
## Missouri          9.0       70
## Montana           6.0       53
## Nebraska          4.3       62
## Nevada           12.2       81
```

```
## New Hampshire     2.1       56
## New Jersey        7.4       89
## New Mexico       11.4       70
## New York         11.1       86
## North Carolina   13.0       45
## North Dakota      0.8       44
## Ohio              7.3       75
## Oklahoma          6.6       68
## Oregon            4.9       67
## Pennsylvania      6.3       72
## Rhode Island      3.4       87
## South Carolina   14.4       48
## South Dakota      3.8       45
## Tennessee        13.2       59
## Texas            12.7       80
## Utah              3.2       80
## Vermont           2.2       32
## Virginia          8.5       63
## Washington        4.0       73
## West Virginia     5.7       39
## Wisconsin         2.6       66
## Wyoming           6.8       60
```

```
USArrests[c("Mississippi"),c(1,3)]
```

```
##             Murder UrbanPop
## Mississippi   16.1       44
```

Yes it does make sense that Mississippi is placed where the problem suggests. As the above shows, Mississippi shows the highest reported murder arrests among the states with one of the smallest urban population percentage. Hence it is close to the "Murder" vector and is far from "UrbanPop" vector.

Contribution

Our group homework process goes as the following:
1. Each member attempts to complete the homework
2. Compare and discuss the answers
3. Complete a finalized version to submit

All members have contributed about the equal amount to complete this homework.