Homework 2

Bailey Ho | Daria Barbour-Brown | Warren Kennedy

Spring 2023

All group members contributed equally to this assignment.

Conceptual

Principal Component Analysis is pre-processing method used in unsupervised learning to project a high dimensional data set into in lower dimensional subspace to find a reduced representation of the features that best represent the relationships in the data.

Application

```
library(ISLR2)
usaa <- USArrests
names(usaa)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

Question 1: Scaled PCA

Part A

Principal Component **(Scaled)** We are sampling from 50 US states and are looking to test relationships between four features: Rape, Murder, Assault and Urban Population. After performing a principal component analysis on the data we get back four principal components, each explaining a portion of the variance in the data.

```
pca.sc <- prcomp(usaa, center=TRUE, scale = TRUE)
pca.sc$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
```

Part B

Each Principal Component column above is know as a loading vector. Each describe the direction within the feature space that points to features providing the most variation in our data.

Principal component 1 places the largest weights on Assault, Rape, and Murder. In other words PC1, describes the overall rates of serious crime. These three features account for the most variation in the data so they are given the larger magnitudes in the first principal component.

The Second Principal Component captures the most variance in features uncorrelated to PC1.

Principal component 2 places the most weight on urban population, and much less weight on the other three features. Therefore we interpret PC2 to explain urban population in a given state.

Part C

```
pve.s
```

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

PC1 explains 62% of the variation in the data, while PC2 explains roughly 25%. The first two principal components together account for roughly 87% of the variation in the data. The remaining 13% of the variation is explained by principal components 3 and 4.

Question 2: Unscaled PCA

Part A

When we run a principal component analysis on data that has not been standardized, we get different results.

```
pca.us <- prcomp(usaa, center=FALSE, scale = FALSE)
pca.us$rotation
```

Principal Component **(Unscaled)**

```
##                  PC1         PC2         PC3         PC4
## Murder   -0.04239181  0.01616262 -0.06588426  0.99679535
## Assault  -0.94395706  0.32068580  0.06655170 -0.04094568
## UrbanPop -0.30842767 -0.93845891  0.15496743  0.01234261
## Rape     -0.10963744 -0.12725666 -0.98347101 -0.06760284
```

Part B

The first two principal components are giving us the linear combination of our features that explain nearly all of the variability in the data. The first loading is telling us Assault accounts for the majority of the variability in the data. The loading in PC2 tells us that Urban Population accounts for most of the remaining variability.

Part C

PC1, which places nearly all of its weight on Assault, accounts for 98 percent of the variation in the data. PC2 accounts for nearly 2% of the remaining variability.

```
pca.us$sdev
```

```
## [1] 202.723056  27.832264   6.523048   2.581365
```

```
pcaVar <- pca.us$sdev^2
pcaVar
```

```
## [1] 41096.637613   774.634904    42.550158     6.663446
```

```
pve.u <- pcaVar / sum(pcaVar)
pve.u
```

```
## [1] 0.9803473532 0.0184786718 0.0010150206 0.0001589544
```

Question 3: Compare Results

The results from the two PCA's are very different. This largely due to the impact that non-standardized data has on the process of principal component analysis. Principal component analysis assumes that the data has a normal distribution. Data that has not been standardized will allow for features with larger unit magnitudes to dominate the analysis.

As you can see below, the mean and standard deviation of *Assault* and *Urban Population* are much higher than the values for Rape and Murder simply because each feature was measured using different units.

```
apply(usaa, 2, mean)
```

```
##   Murder  Assault UrbanPop     Rape 
##    7.788  170.760   65.540   21.232
```

```
apply(usaa, 2, sd)
```

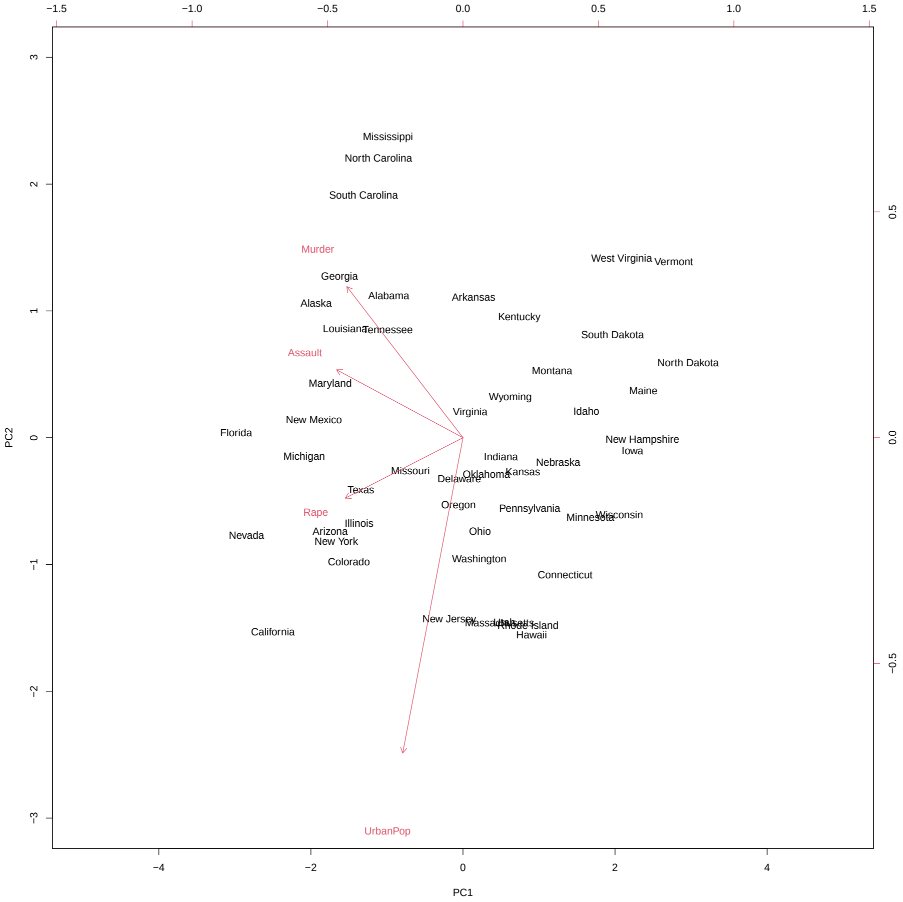
```
##    Murder   Assault  UrbanPop      Rape 
##  4.355510 83.337661 14.474763  9.366385
```

The larger values for Assault and Urban Population allow these features to dominate our second analysis.

In our first PCA, the initial loading used a balanced combination of features to attempt to maximize the amount of variability it captured. Comparatively, our initial loading from the non-standardized analysis placed the majority of the weight on just one feature: Assault. When we fail to scale the data, PCA will provide limited insight into patterns that may be present in the data

Question 4: Biplot for Scaled PCA

```
biplot(pca.sc, scale = 0,expand=2,cex=1,xlim=c(-5,5),ylim=c(-3,3))
```



Mississippi has the third lowest value for Urban Population in the data, and lies high on the bi-plot in the opposite direction of the Urban Population vector. It also holds the second highest rate for Murder and is therefore placed close to the tip of the Murder rate vector. This positioning is consistent with what we observe in the data.