Lab 10 Principal Component Analysis

Problem Statement

We are going to study how to perform PCA on a dataset.

Dataset

`USArrests` dataset (in the base R package) The rows of the data set contain the 50 states, in alphabetical order.

```
states <- row.names(USArrests)
states[1:10]
```

```
##  [1] "Alabama"     "Alaska"      "Arizona"     "Arkansas"    "California"
##  [6] "Colorado"    "Connecticut" "Delaware"    "Florida"     "Georgia"
```

The columns of the data set contain the four variables.

```
names(USArrests)
```

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

Questions

1) Calculate the mean and variance of each column, by using `apply()` function.

   Hint: `apply(dataset, 1, func)` is to apply the `func` to each row of dataset, and

   `apply(dataset, 2, func)` is to apply the func to each column of dataset.

2) What conclusions can you draw from 1)? And consequently what transformation would you do to your dataset?

3) Perform principal component analysis using the `prcomp()` function.

4) Check the results, report the number of PCs and their center, scale, and rotation.

5) Plot the first two PCs.

6) What are the standard deviation of each principal component? Based on this result, calculate the variance explained by each PC and the proportion of variance explained by each PC.

7) Plot the PVE explained by each component as well as the cumulative PVE.

   Hint: the cumulative PVE can be obtained by the `cumsum()` function.