US Arrests Analysis (PCA & Regression Model)

MINHCHAU

```
library(dplyr)
library(tidyr)
library(car)
library(psych) # for pairs.panels()
library(factoextra) # for fviz_cluster()
library(ggplot2)
```

Perform Principal Component Analysis (PCA), examine the results, and use the principal components in regression modeling.

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of a dataset while retaining as much variance (or information) as possible. It achieves this by transforming the original features (variables) into a smaller set of uncorrelated variables called **principal components**. These components are ordered so that the first few components capture the most significant variation in the data.

Dataset required: `Arrests.csv`

(1a) Load the Dataset and read the Arrests.csv file into a dataframe called df. The **US Arrests** dataset contains data about the number of arrests per 100,000 residents for different crimes in each US state.

- Use head() and names() functions to inspect the first few rows and column names of the dataset.
- What are the names of the columns in the dataset?
- Extract the columns Murder, Assault, UrbanPop, and Rape into a new dataframe called USArrests.

```
df = read.csv('Arrests.csv')
head(df)
```

```
##        State Murder Assault UrbanPop Rape GDP.in.dollars Urban
## 1    Alabama   13.2     236       58 21.2    1.05000e+11     n
## 2     Alaska   10.0     263       48 44.5    2.58108e+10     n
## 3    Arizona    8.1     294       80 31.0    1.32000e+11     y
## 4   Arkansas    8.8     190       50 19.5    5.98463e+10     n
## 5 California    9.0     276       91 40.6    3.02000e+11     y
## 6   Colorado    7.9     204       78 38.7    1.37000e+11     y
```

```
names(df)
```

```
## [1] "State"          "Murder"         "Assault"        "UrbanPop"      
## [5] "Rape"           "GDP.in.dollars" "Urban"
```

```
USArrests <- df[ , c('Murder', 'Assault', 'UrbanPop', 'Rape')]
```

(1b) Let’s take a quick look at the column means and variance of the data. Compute the mean and standard deviation of each feature in USArrests using the apply() function. We can use the apply() function to apply a function - in this case, the mean() function - to each row or column of the data set.

- What are the means and standard deviations of the features?
- Comment on any differences in scale.
- Standardize the variables to have a mean of zero and a standard deviation of one. Why is it important to center and scale the data before performing PCA?

```
#mena and standard deviation of the features
apply(USArrests , 2, mean)
```

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

```
apply(USArrests , 2, sd)
```

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

```
#standardize the data
USArrests_scale <- scale(USArrests)
```

(1c) Perform PCA with centering and scaling. You can do this by setting center = TRUE and scale = TRUE in the `prcomp()` function.

- View the summary of the PCA results, what percentage of variance is explained by the first two principal components?

A **scree plot** is a graphical representation that helps you determine how much variance is explained by each principal component.

- Create a scree plot to visualize the results.
- What do the scree plot reveal about the principal components?

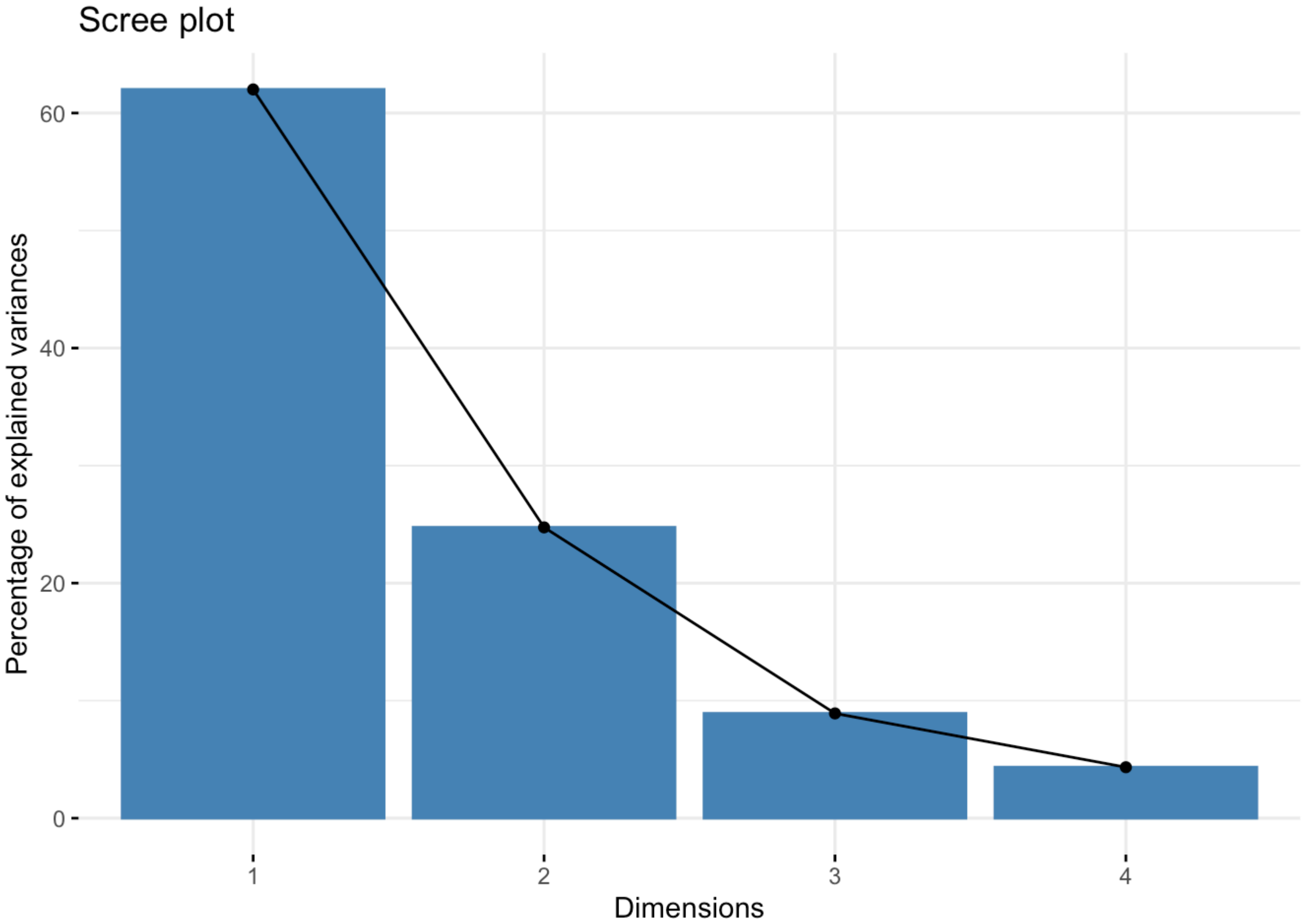
```
pr.out <- prcomp(~ Murder + Assault + UrbanPop + Rape, data=USArrests, center=TRUE, s
cale=TRUE)
#check the components of the prcomp object
names(pr.out)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"        "call"
```

```
summary(pr.out)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

```
#scree plot
library(factoextra)
fviz_eig(pr.out)
```



:::

{style="color: red"} 86.75% of variance is explained by the first 2 principal components PC1: The 1st principal component explains about 62.01% of the proportion of variance PC2: The 2nd principal component explains about 24.74% of the proportion of variance PC3 and PC4: explains about 8.91% and 4.34% of the proportion of variance :::

(1d) When you perform PCA, the goal is to find a set of new axes (principal components) that capture the maximum variance in the data. Each principal component is a linear combination of the original features (variables).

Loading vectors represent the coefficients of these linear combinations. In simple terms, a loading vector tells you how much each original variable "loads" or contributes to a particular principal component. Loading vectors defines a direction in feature space along which the data vary the most.

- You can use the pr.out$rotation to view the loadings for all principal components.
- Extract and inspect the loadings for the first two principal components (`pc1` and `pc2`).
- Provide the formula for PC1.
- Add the first two principal component scores as new columns to the USArrests dataframe.

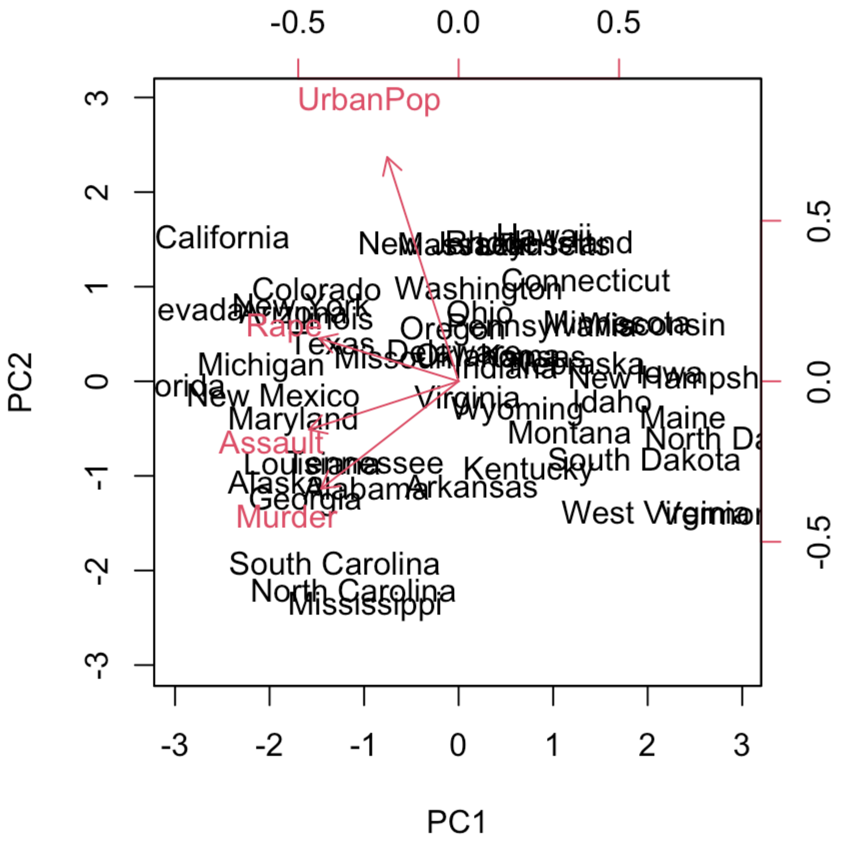
A **biplot** is a graphical representation that combines both the **scores** of the data points (the projections onto the principal components) and the **loadings** of the variables

- Create a biplot to visualize the results.
- What do the biplot reveal about the principal components?

```
head(pr.out$x, 10)
```

```
##             PC1         PC2         PC3          PC4
## 1  -0.97566045 -1.12200121  0.43980366  0.154696581
## 2  -1.93053788 -1.06242692 -2.01950027 -0.434175454
## 3  -1.74544285  0.73845954 -0.05423025 -0.826264240
## 4   0.13999894 -1.10854226 -0.11342217 -0.180973554
## 5  -2.49861285  1.52742672 -0.59254100 -0.338559240
## 6  -1.49934074  0.97762966 -1.08400162  0.001450164
## 7   1.34499236  1.07798362  0.63679250 -0.117278736
## 8  -0.04722981  0.32208890  0.71141032 -0.873113315
## 9  -2.98275967 -0.03883425  0.57103206 -0.095317042
## 10 -1.62280742 -1.26608838  0.33901818  1.065974459
```

```
USArrests$pc1 <- pr.out$x[, "PC1"]
USArrests$pc2 <- pr.out$x[, "PC2"]
biplot(pr.out, scale=0, xlabs=df$State)
```



Orientation of the vector: When a vector is parallel to a principal component axis, it means that the vector's direction closely follows that axis. This indicates that the vector contributes significantly to that particular principal component

PC1: The 1st principal component roughly corresponds to a measure of overall rates of serious crimes Interpretation: States with large negative scores on PC1, such as Florida, Nevada and California have high crime rates, while states like South Dakota ,with positive scores on PC1, have low crime rates

(1e) How can you use the principal component scores for further regression modeling?

- Perform a regression analysis using the principal component scores (`pc1` and `pc2`).
- Are the principal components significant in explaining the dependent variable (e.g., GDP) in this regression model?

```
USArrests$GDP <- df$GDP.in.dollars
#Fit a linear regression model
summary(lm(GDP ~ pc1+pc2, data=USArrests))
```

```
##
## Call:
## lm(formula = GDP ~ pc1 + pc2, data = USArrests)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.502e+11 -6.607e+10 -1.397e+10  4.653e+10  4.578e+11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.538e+11  1.854e+10   8.294 9.32e-11 ***
## pc1         -4.344e+10  1.189e+10  -3.653 0.000652 ***
## pc2          4.310e+10  1.883e+10   2.289 0.026592 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.311e+11 on 47 degrees of freedom
## Multiple R-squared:  0.2834, Adjusted R-squared:  0.2529
## F-statistic: 9.292 on 2 and 47 DF,  p-value: 0.0003978
```

Since the p-value for pc1 = 0.00652 < 0.05 and pc2 = 0.0265 < 0.05, both the principal components are significant at 5% level of significance to the GDP in this regression model