# Project-2–Unsupervised-Learning.R

arpan

2023-06-17

```r
#Project 2: Unsupervised Learning


#Use the in-built "mtcars" data of R and do as follows in R studio with R script:

# 1. Perform the principal component analysis in the data and exact the
# dimensions based on components with eigenvalues >1, check it with screeplot
# as well and interpret the result carefully


# Load the mtcars dataset
data(mtcars)

# Principal Component Analysis (PCA)
pca <- princomp(mtcars, cor = TRUE)  # Perform PCA
eigenvalues <- pca$sdev^2  # Extract eigenvalues
eigenvalues
```

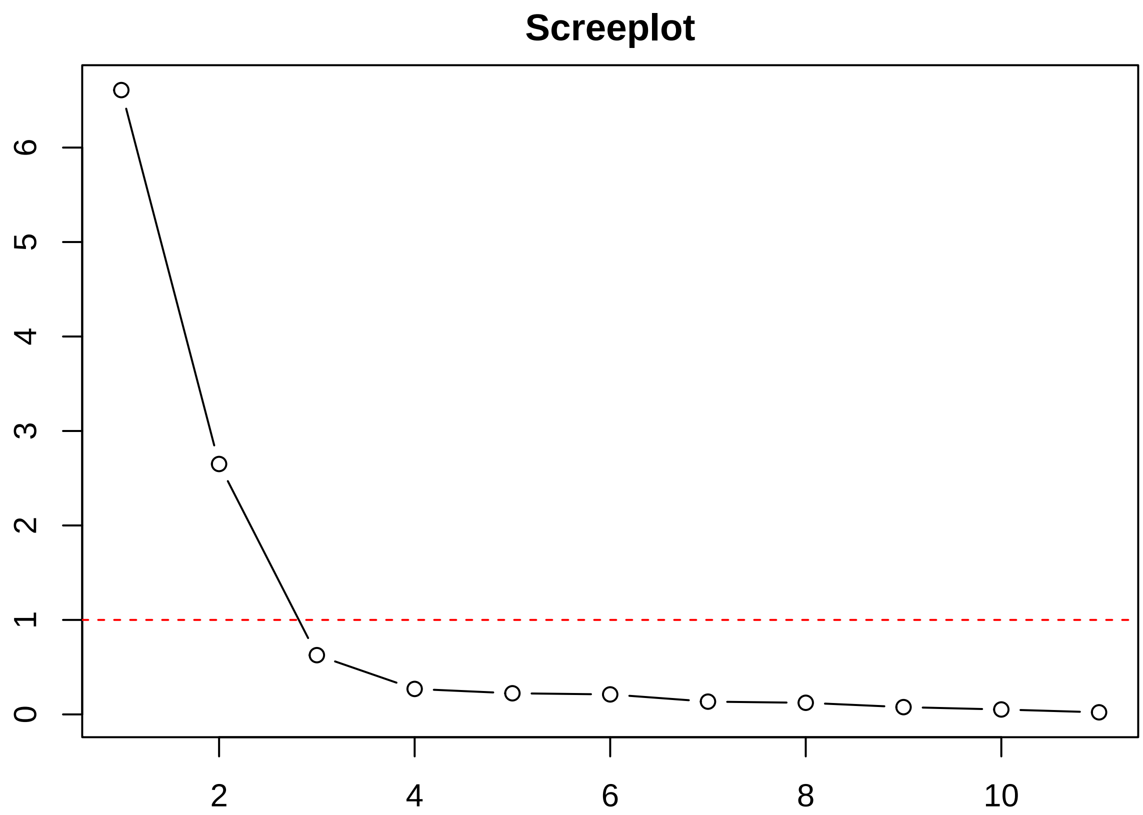
```
##     Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
## 6.60840025 2.65046789 0.62719727 0.26959744 0.22345110 0.21159612 0.13526199
##     Comp.8     Comp.9    Comp.10    Comp.11
## 0.12290143 0.07704665 0.05203544 0.02204441
```

```r
# Calculate cumulative variance
cumulative_variance <- cumsum(eigenvalues) / sum(eigenvalues)
cumulative_variance
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
## 0.6007637 0.8417153 0.8987332 0.9232421 0.9435558 0.9627918 0.9750884 0.9862612
##    Comp.9   Comp.10   Comp.11
## 0.9932655 0.9979960 1.0000000
```

```r
# Adjust figure margins
par(mar = c(3, 3, 2, 2))

# Screeplot
plot(1:length(eigenvalues), eigenvalues, type = "b", xlab = "Component",
     ylab = "Eigenvalue", main = "Screeplot")
abline(h = 1, lty = 2, col = "red")  # Add a line at eigenvalue = 1
```

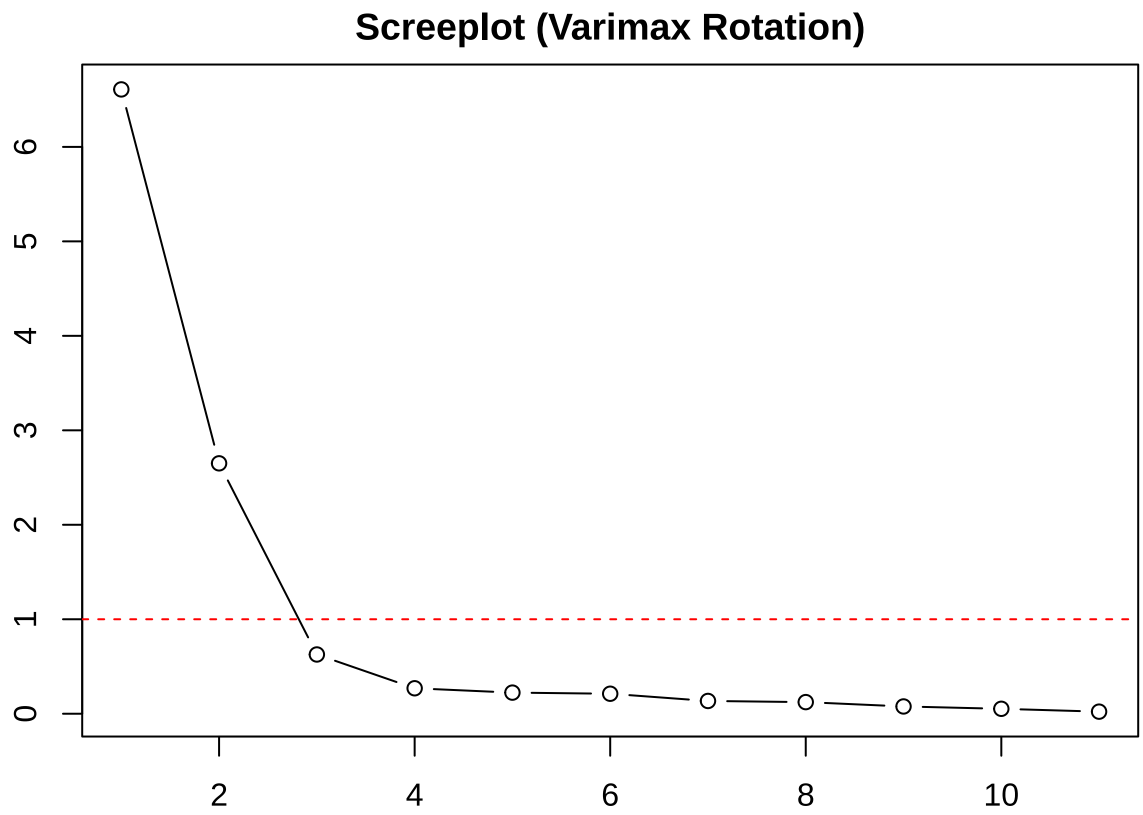
```
#Interpretation of PCA with eigenvalues > 1:
# The components with eigenvalues greater than 1 indicate dimensions that
# explain more variance in the data than a single variable. In this case,
# components with eigenvalues greater than 1 can be considered meaningful.
# It is recommended to retain the first two components (Comp.1 and Comp.2) as
# they explain a significant amount of variance in the data. These components
# are considered meaningful dimensions that summarize the key patterns and
# trends present in the original variables.

#2. Perform the principal component analysis with varimax rotation in the data
# and exact the dimensions based on eigenvalue >1 and check it with Screeplot
# as well and interpret the result carefully

# Principal Component Analysis with Varimax rotation
library(psych)
```




```
pca_rotated <- principal(mtcars, nfactors = length(eigenvalues), rotate = "varimax")  # Perform PCA wit
eigenvalues_rotated <- pca_rotated$values  # Extract eigenvalues
eigenvalues_rotated
```

```
##  [1] 6.60840025 2.65046789 0.62719727 0.26959744 0.22345110 0.21159612
##  [7] 0.13526199 0.12290143 0.07704665 0.05203544 0.02204441
```

```
# Screeplot
plot(1:length(eigenvalues_rotated), eigenvalues_rotated, type = "b", xlab = "Component", ylab = "Eigenva
abline(h = 1, lty = 2, col = "red")  # Add a line at eigenvalue = 1
```

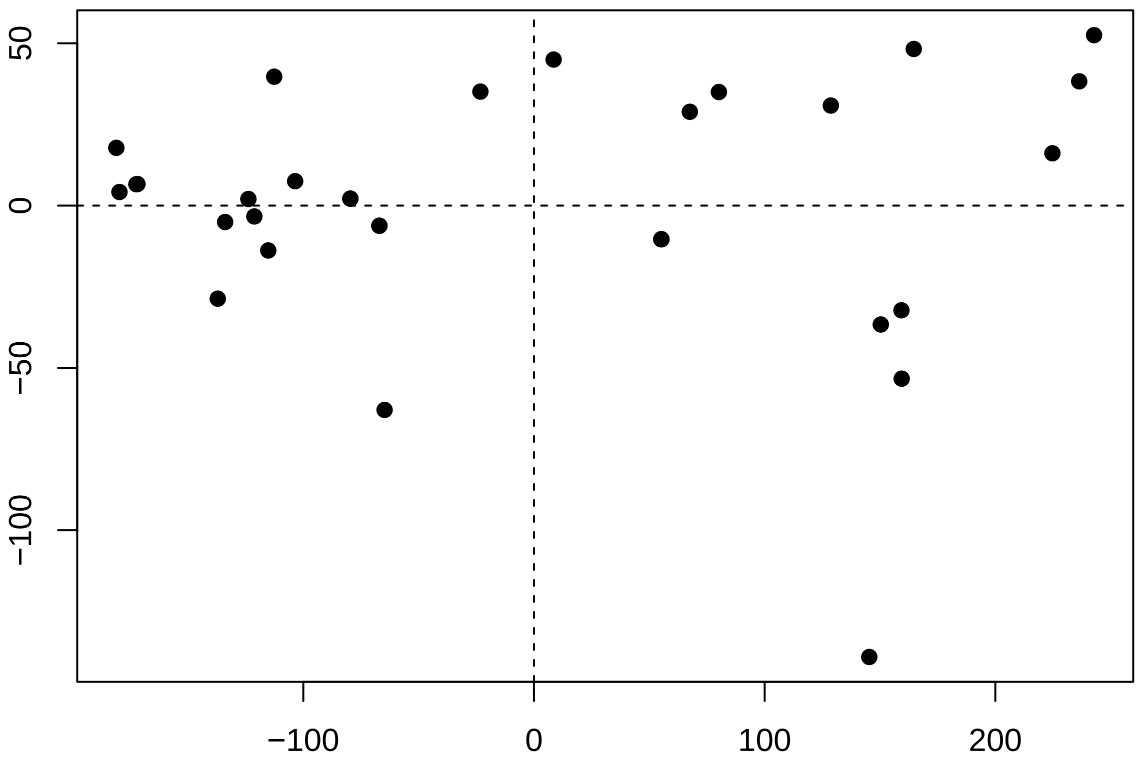



```
#Interpretation of PCA with varimax rotation and eigenvalues > 1:
# Similar to the previous analysis, components with eigenvalues greater than 1
# are considered significant. Based on the eigenvalues greater than 1 criterion
# and the varimax rotation, it is recommended to retain the first two components
# as they explain a substantial amount of variance in the data. These components,
# along with the varimax rotation, provide a meaningful representation of the
# key patterns and relationships in the original variables, capturing the most
# important information while reducing the complexity and correlation among the
# components.

#3. Perform the classical multidimensional scaling in the data, revise the
# results using stress values and interpret the result carefully

# Classical Multidimensional Scaling (MDS)
library(stats)

# Perform classical multidimensional scaling
mds <- cmdscale(dist(mtcars))
mds
```

```
##                                   [,1]         [,2]
```

```
## Mazda RX4            -79.596425    2.132241
## Mazda RX4 Wag        -79.598570    2.147487
## Datsun 710          -133.894096   -5.057570
## Hornet 4 Drive        8.516559   44.985630
## Hornet Sportabout   128.686342   30.817402
## Valiant             -23.220146   35.106518
## Duster 360          159.309025  -32.259197
## Merc 240D          -112.615805   39.702195
## Merc 230           -103.534591    7.513104
## Merc 280            -67.046877   -6.208536
## Merc 280C           -66.997514   -6.206387
## Merc 450SE           55.211672  -10.373509
## Merc 450SL           55.173910  -10.361893
## Merc 450SLC          55.251602  -10.370934
## Cadillac Fleetwood  242.814893   52.501758
## Lincoln Continental 236.369886   38.280788
## Chrysler Imperial   224.737944   16.111941
## Fiat 128           -172.363654    6.575522
## Honda Civic        -181.066911   17.783639
## Toyota Corolla     -179.697852    4.188212
## Toyota Corona      -121.224099   -3.345362
## Dodge Challenger     80.159386   34.983214
## AMC Javelin          67.572431   28.894067
## Camaro Z28          150.354631  -36.633575
## Pontiac Firebird    164.652522   48.239880
## Fiat X1-9          -171.897231    6.643746
## Porsche 914-2      -123.804988    2.033356
## Lotus Europa       -137.082789  -28.675647
## Ford Pantera L      159.413222  -53.318347
## Ferrari Dino        -64.762396  -62.954280
## Maserati Bora       145.361703 -139.049149
## Volvo 142E         -115.181783  -13.826313
```

```
plot(mds, pch = 19)
abline(h=0, v=0, lty=2)
```




```
# Calculate stress values
dissimilarity_matrix <- dist(mtcars)  # Dissimilarity matrix of the original data
reduced_distances <- as.dist(dist(mds))  # Distances in the reduced-dimensional space
stress_values <- sqrt(sum((dissimilarity_matrix - reduced_distances)^2)) / sqrt(sum(dissimilarity_matri

stress_values
```

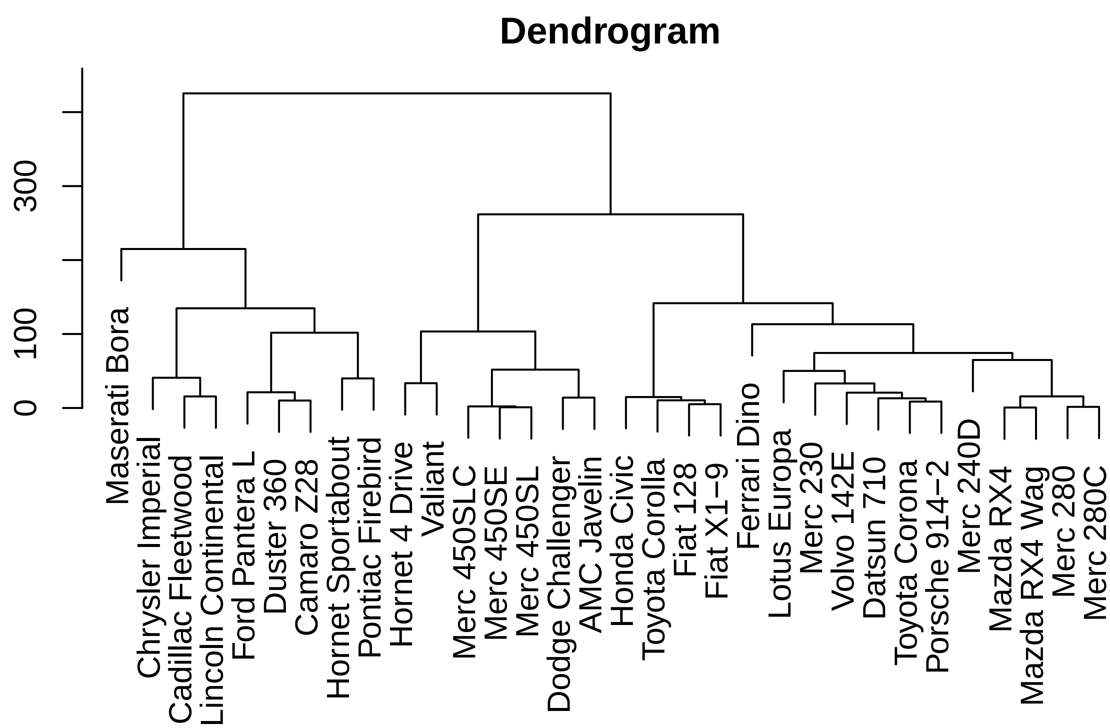
```
## [1] 0.001862287
```

```
# Interpretation of MDS using stress values:
# The stress value obtained from the classical multidimensional scaling (MDS)
# analysis is 0.001862287. The stress value is a measure of the discrepancy
# between the distances in the reduced-dimensional space and the original
# dissimilarity matrix.

#The stress values represent the goodness-of-fit of the MDS solution.
# Lower stress values indicate a better representation of distances in
# the reduced space. To interpret the MDS results, compare stress values across
# different MDS solutions and choose the solution with the lowest stress value
# as it provides the best representation of the original distances.


# 4. Perform the hierarchical cluster analysis in the data and determine the
# number of clusters to exact using the dendogram and cut at the various
# distances with justification
```

```
# Hierarchical Cluster Analysis
dist_matrix <- dist(mtcars)  # Calculate dissimilarity matrix
hclust_res <- hclust(dist_matrix, method = "complete")
hclust_res
```

```
##
## Call:
## hclust(d = dist_matrix, method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 32
```

```
# Dendrogram
plot(hclust_res, main = "Dendrogram")
```




```
# Determine the number of clusters using the dendrogram
cut_heights <- c(100, 150, 200)  # Adjust based on dendrogram visual inspection
cut_clusters <- cutree(hclust_res, h = cut_heights)
cut_clusters
```

```
##                       100 150 200
## Mazda RX4               1   1   1
```

```
## Mazda RX4 Wag         1   1   1
## Datsun 710            1   1   1
## Hornet 4 Drive        2   2   2
## Hornet Sportabout     3   3   3
## Valiant               2   2   2
## Duster 360            4   3   3
## Merc 240D             1   1   1
## Merc 230              1   1   1
## Merc 280              1   1   1
## Merc 280C             1   1   1
## Merc 450SE            5   2   2
## Merc 450SL            5   2   2
## Merc 450SLC           5   2   2
## Cadillac Fleetwood    6   3   3
## Lincoln Continental   6   3   3
## Chrysler Imperial     6   3   3
## Fiat 128              7   1   1
## Honda Civic           7   1   1
## Toyota Corolla        7   1   1
## Toyota Corona         1   1   1
## Dodge Challenger      5   2   2
## AMC Javelin           5   2   2
## Camaro Z28            4   3   3
## Pontiac Firebird      3   3   3
## Fiat X1-9             7   1   1
## Porsche 914-2         1   1   1
## Lotus Europa          1   1   1
## Ford Pantera L        4   3   3
## Ferrari Dino          8   1   1
## Maserati Bora         9   4   4
## Volvo 142E            1   1   1
```

```
# Justification:
# Cut height of 100: This results in a large number of clusters, with each
# observation assigned to its own cluster. It allows for a fine-grained analysis
# of individual cases and identification of unique patterns within the data.

# Cut height of 150: This leads to a moderate number of clusters, indicating a
# higher level of aggregation compared to the previous cut height. It captures
# broader similarities among observations and identifies groups with similar
# characteristics.

# Cut height of 200: This further reduces the number of clusters, suggesting a
# higher level of aggregation and grouping of more similar observations. It helps
# identify larger, distinct groups and provides a broader perspective on trends
# and patterns in the data.

#Interpretation of Hierarchical Cluster Analysis:
# The dendrogram provides a visual representation of the hierarchical
# clustering results. To determine the number of clusters, look for significant
# gaps between clusters in the dendrogram. Adjust the cut heights based on visual
# inspection to obtain a reasonable number of distinct clusters.

#5. Perform the k-means cluster analysis in the data based on the number of
```

```r
# clusters identified using dendogram and interpret the result carefully

# 5. K-means Cluster Analysis
library(cluster)

# Perform k-means clustering based on the number of clusters identified from the dendrogram
k <- 3  # Adjust based on the dendrogram analysis
kmeans_res <- kmeans(mtcars, centers = k)
kmeans_res
```

```
## K-means clustering with 3 clusters of sizes 14, 7, 11
##
## Cluster means:
##        mpg cyl     disp        hp     drat       wt     qsec        vs
## 1 15.10000   8 353.1000 209.21429 3.229286 3.999214 16.77214 0.0000000
## 2 19.74286   6 183.3143 122.28571 3.585714 3.117143 17.97714 0.5714286
## 3 26.66364   4 105.1364  82.63636 4.070909 2.285727 19.13727 0.9090909
##          am     gear     carb
## 1 0.1428571 3.285714 3.500000
## 2 0.4285714 3.857143 3.428571
## 3 0.7272727 4.090909 1.545455
##
## Clustering vector:
##           Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive
##                   2                   2                   3                   2
##   Hornet Sportabout             Valiant          Duster 360           Merc 240D
##                   1                   2                   1                   3
##            Merc 230            Merc 280           Merc 280C          Merc 450SE
##                   3                   2                   2                   1
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood Lincoln Continental
##                   1                   1                   1                   1
##   Chrysler Imperial            Fiat 128         Honda Civic      Toyota Corolla
##                   1                   3                   3                   3
##       Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28
##                   3                   1                   1                   1
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2        Lotus Europa
##                   1                   3                   3                   3
##      Ford Pantera L        Ferrari Dino       Maserati Bora          Volvo 142E
##                   1                   2                   1                   3
##
## Within cluster sum of squares by cluster:
## [1] 93643.90 13954.34 11848.37
##  (between_SS / total_SS =  80.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
# Interpretation of K-means Cluster Analysis:
# Number of clusters: The data has been divided into three clusters, with cluster
# sizes of 7, 11, and 14 observations, respectively.
```

```
#Cluster means: The cluster means represent the average values of each variable
# within each cluster. For example, in cluster 1, the average mpg is 19.74, the
# average cyl is 6, the average disp is 183.31, and so on. These cluster means
# provide insights into the typical characteristics of the observations within
# each cluster.

# Clustering vector: The clustering vector indicates the assignment of each
#observation to a specific cluster.

# Within cluster sum of squares: This metric measures the variability within each
# cluster. A lower value indicates that the observations within the cluster are
# more similar to each other. In this case, cluster 2 has the lowest
# within-cluster sum of squares, followed by cluster 1, and cluster 3 has the
# highest sum of squares.

#Between-cluster sum of squares: This metric measures the variability between
# the clusters. The percentage value (80.8%) indicates how much of the total
# variability in the data is accounted for by the differences between
# the clusters. A higher percentage suggests that the clusters are
# well-separated and distinct from each other.
```