Principal Component Analysis

Devan Morehouse

2023-09-19

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a valuable technique in data analysis, especially when you want to reduce the dimensionality of high-dimensional data while preserving the most critical information. It is commonly applied to numeric data types, typically using continuous variables, such as real numbers or integers. PCA is beneficial when dealing with datasets containing a large number of features or variables, like in genomics, image analysis, or finance. By transforming the original variables into a new set of orthogonal variables (principal components), PCA simplifies data representation while retaining as much variance as possible. This can aid in visualizing data, identifying dominant patterns, and improving computational efficiency in subsequent analysis tasks, such as clustering, classification, or regression.

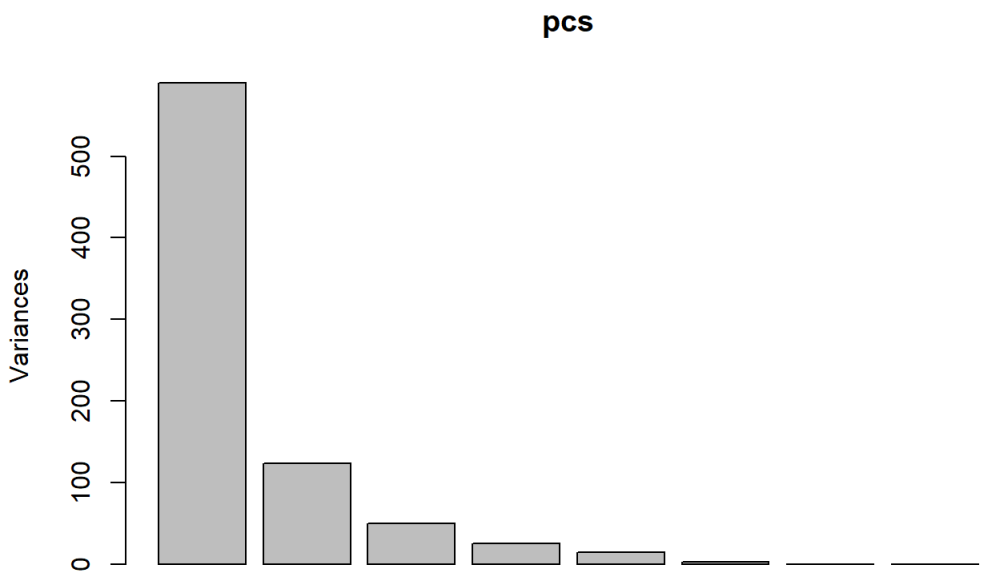
```
cereals.df <- read.csv("C:\\Users\\dev46\\OneDrive\\Desktop\\School Documents\\Spring 2023\\MGQ 408 Bus. Analytics & Data Science\\Data\\Cereals.csv")
pcs <- prcomp(data.frame(cereals.df$calories, cereals.df$rating))
## Note: 86% of the total variance is accounted for by component
summary(pcs)
```

```
## Importance of components:
##                             PC1    PC2
## Standard deviation     22.3165 8.8844
## Proportion of Variance  0.8632 0.1368
## Cumulative Proportion   0.8632 1.0000
```

```
## Non-Normalized
### Note: na.omit will omit null values
boston.housing.df <- read.csv("C:\\Users\\dev46\\OneDrive\\Desktop\\School Documents\\Spring 2023\\MGQ 408 Bus. Analytics & Data Science\\Data\\BostonHousing.csv")
pcs <- prcomp(na.omit(boston.housing.df[,-c(5:10)]))
summary(pcs)
```

```
## Importance of components:
##                             PC1     PC2     PC3     PC4    PC5     PC6     PC7
## Standard deviation     24.2908 11.1168 7.03073 5.01897 3.7437 1.75941 0.24610
## Proportion of Variance  0.7326  0.1534 0.06137 0.03127 0.0174 0.00384 0.00008
## Cumulative Proportion   0.7326  0.8860 0.94735 0.97863 0.9960 0.99987 0.99995
##                            PC8
## Standard deviation     0.20697
## Proportion of Variance 0.00005
## Cumulative Proportion  1.00000
```

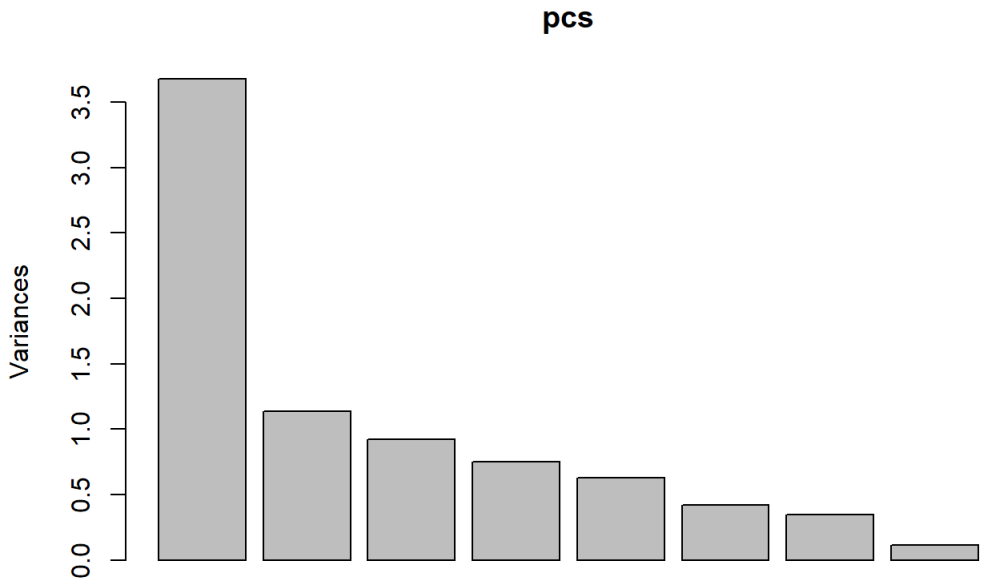
```
#pcs$sd^2
plot(pcs)
```




```
## Normalized
pcs <- prcomp(na.omit(boston.housing.df[,-c(5:10)]),scale. = T)
summary(pcs)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.9179 1.0671 0.9602 0.86769 0.79389 0.64772 0.58830
## Proportion of Variance 0.4598 0.1423 0.1153 0.09411 0.07878 0.05244 0.04326
## Cumulative Proportion  0.4598 0.6021 0.7174 0.81150 0.89029 0.94273 0.98599
##                           PC8
## Standard deviation     0.33478
## Proportion of Variance 0.01401
## Cumulative Proportion  1.00000
```

```
#pcs$sd^2
plot(pcs)
```


pcs
Variances
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5