PCA

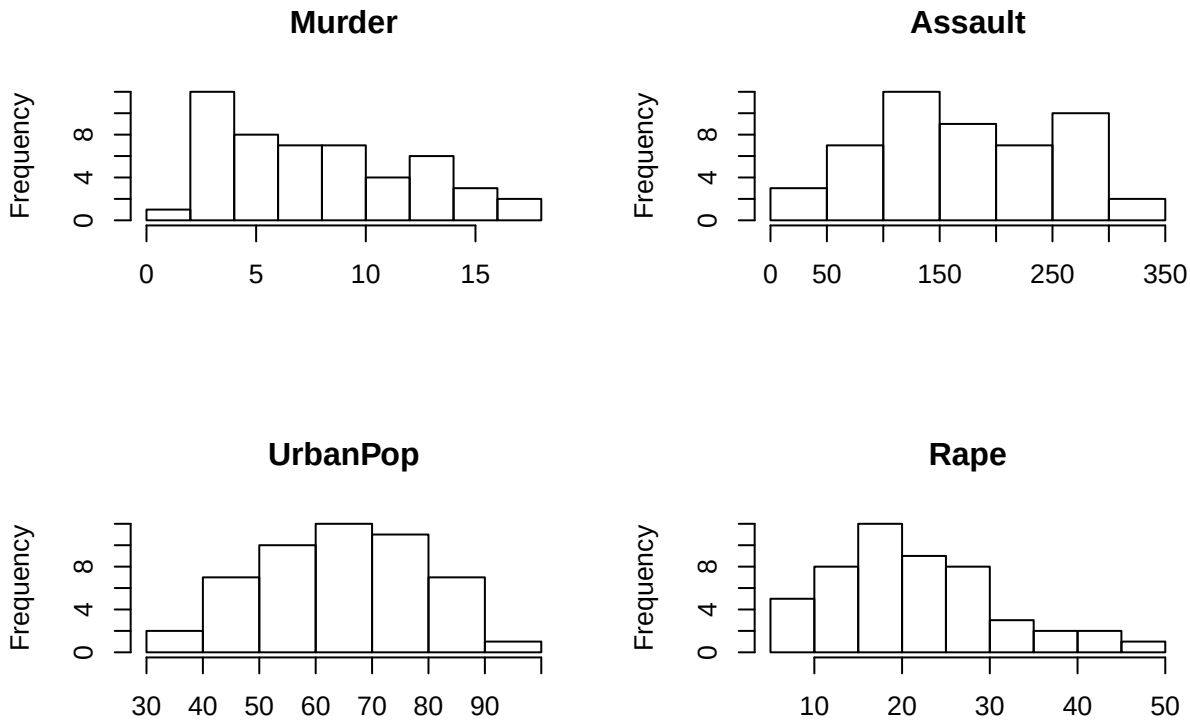
Tutorial ini adalah bagian dari kuliah Fundamen Sains Data, Informatika UII.

Contoh 1 Pada contoh ini, kita akan menggunakan data set `USArrest` dari package `datasets`, yang berisi 50 observasi dari negara-negara bagian di Amerika Serikat tentang presentase penangkapan tindak kejahatan per 100.000 orang. Tindak kejahatan tersebut terbagi ke dalam beberapa macam: `Murder`, `Assault`, dan `Rape`. Selain itu, data ini juga mencatat `UrbanPop` presentase populasi yang melakukan urbanisasi. Pertama, mari kita lihat bentuk dan sebaran datanya.

```
data("USArrests")
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
par(mfrow=c(2,2))
for(i in 1:ncol(USArrests)) {
  hist(USArrests[, i], main = paste(colnames(USArrests[i])), xlab = "")
}
```



Selanjutnya, kita mengaplikasikan PCA pada data tersebut, menggunkaan fungsi `prcomp()` dari package `stats`. Parameter `scale=TRUE, center=TRUE` memastikan nilai setiap variabel distandarisasi, sehingga semua variabel memiliki skala nilai yang sama.

```
1 pcaModel <- prcomp(USArrests, scale. = TRUE, center = TRUE)
2 pcaModel$rotation
```

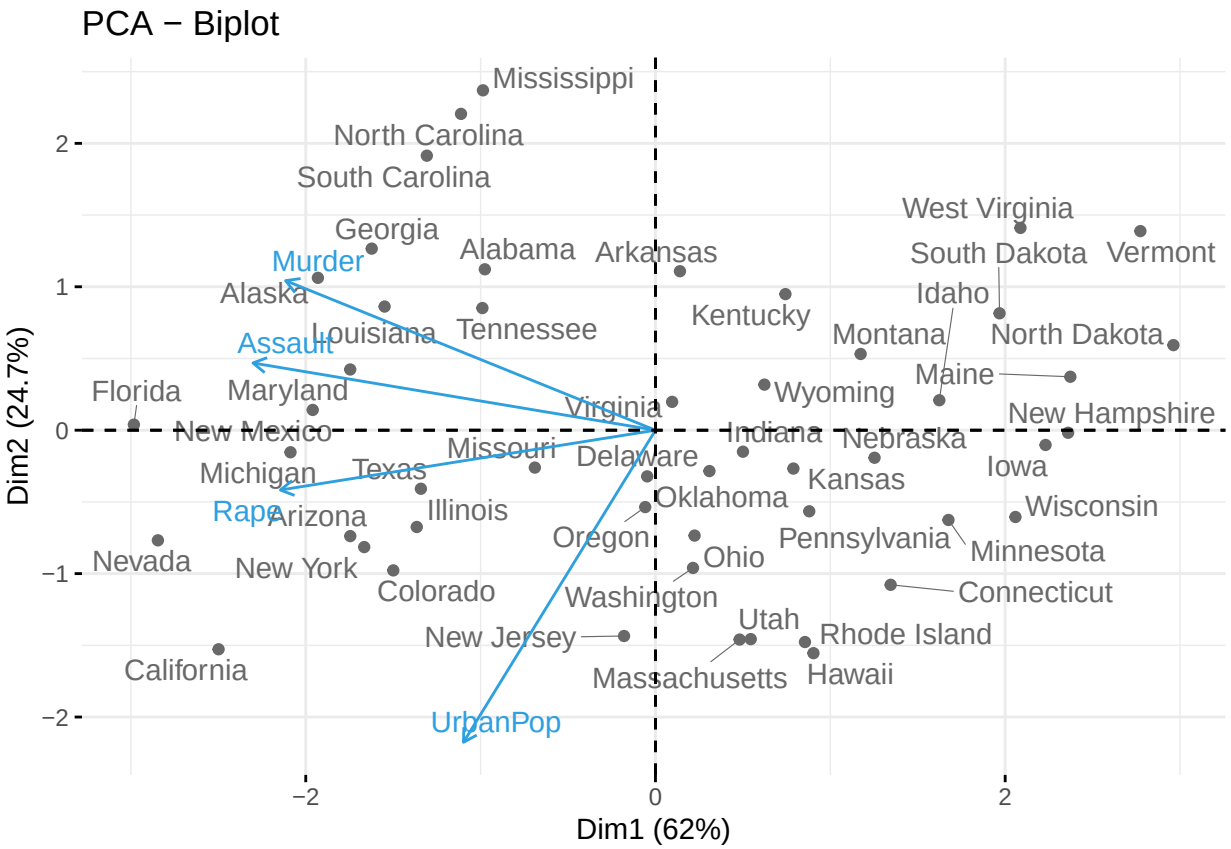
```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

Baris 2 mengembalikan `matrix rotation` yang setiap kolomnya madalah eigenvector yang mengindikasikan principal component pertama hingga `p` (dimana `p` adalah jumlah variabel di data set). `pcaModel$x` memberikan proyeksi setiap data point ke semua principal component.

Untuk kemudahan interpretasi, kita dapat memvisualisasikan *biplot* nya.

```
library(factoextra)

fviz_pca_biplot(pcaModel, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969"  # Individuals color
)
```



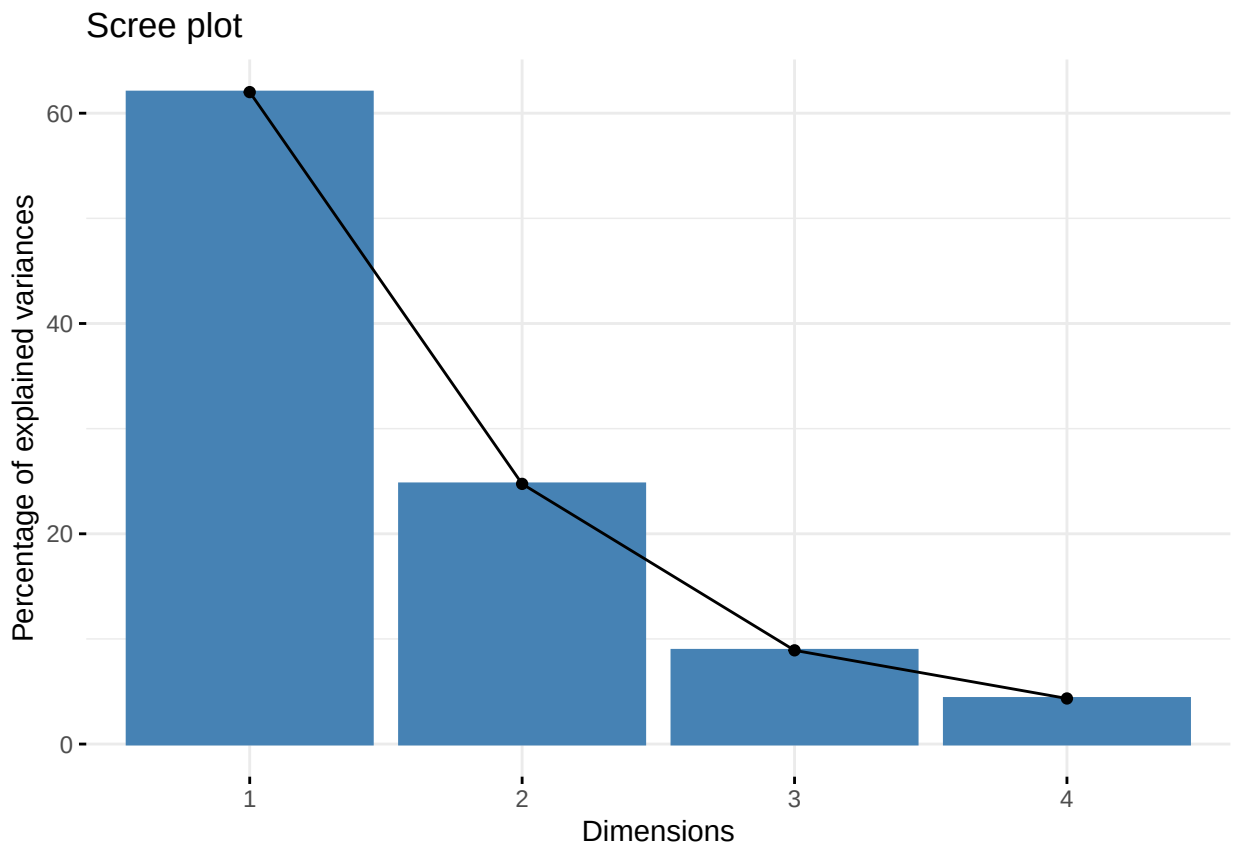
Dari biplot di atas, sumbu horizontal (Dim1) merepresentasikan principal component yang pertama, yang mengandung 62% variance dari seluruh data set; sumbu vertikal (Dim2) merepresentasikan principal component kedua, yang mengandung 24.7% variance. Dari dua component ini saja, 86.7% variance atau informasi yang dikandung data dapat dijelaskan. Di sini dapat kita lihat, dari contoh yang relatif kecil dengan empat variabel saja, kita hanya butuh dua "variabel" (principal component) saja untuk merepresentasikan hampir 90% informasi yang ada di data. Ini adalah semangat dimensionality reduction yang menjadi tujuan PCA.

Kemudian pada biplot di atas, juga divisualisasikan setiap variabel sebagai bentuk vektor (panah biru). Dari plot tersebut, dapat kita lihat arah vektor `Murder`, `Assault`, dan `Rape` cenderung horizontal seperti arah principal component yang pertama (Dim1). Ini mengindikasikan bahwa variable `Murder`, `Assault`, dan `Rape` lebih banyak dijelaskan/diwakili oleh principal komponen yang pertama. Sebaliknya, arah vektor `UrbanPop` lebih mendekati arah principal component yang kedua (dim2). Ini mengindikasikan jika informasi yang dibawa variabel `UrbanPop` lebih banyak diwakili oleh principal component yang kedua.

Dari biplot di atas kita juga dapat menarasikan temuan, sebagai contoh, bahwa angka kriminal (`murder`, `assault`, dan `rape`) lebih banyak terjadi di negara bagian Florida (karena lebih searah dengan vektor `murder`, `assault`, dan `rape`) dibanding dengan Pennsylvania. Kita juga dapat lihat jika angka urbanisasi di New Jersey lebih tinggi (karena lebih searah dengan vektor `UrbanPop`) jika dibandingkan dengan Arkansas.

Selanjutnya kita dapat memvisualisasikan variance yang dibawa setiap component melalui screeplot sebagai berikut. Seperti dijelaskan di atas, dapat kita lihat component 1 dan 2 menjelaskan hampir 90% variance di data.

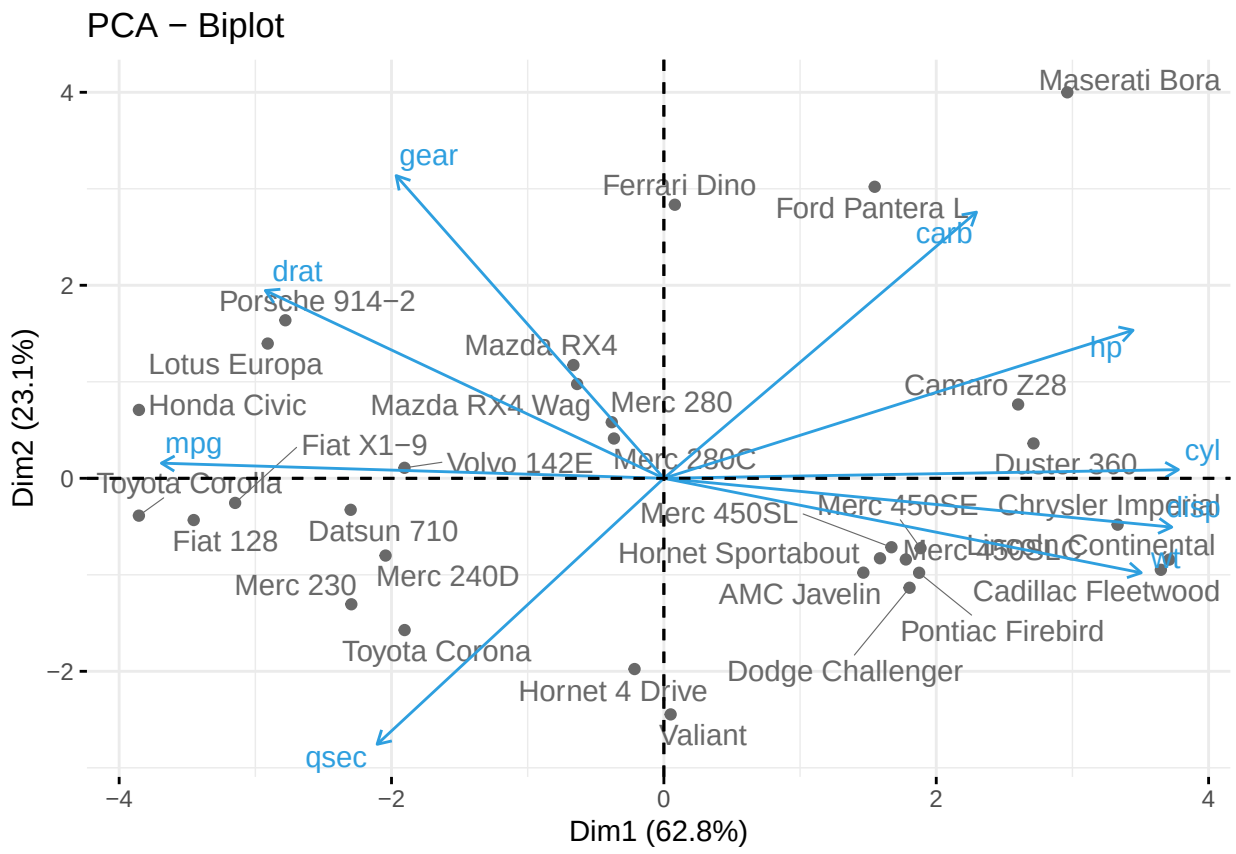
```
fviz_eig(pcaModel)
```



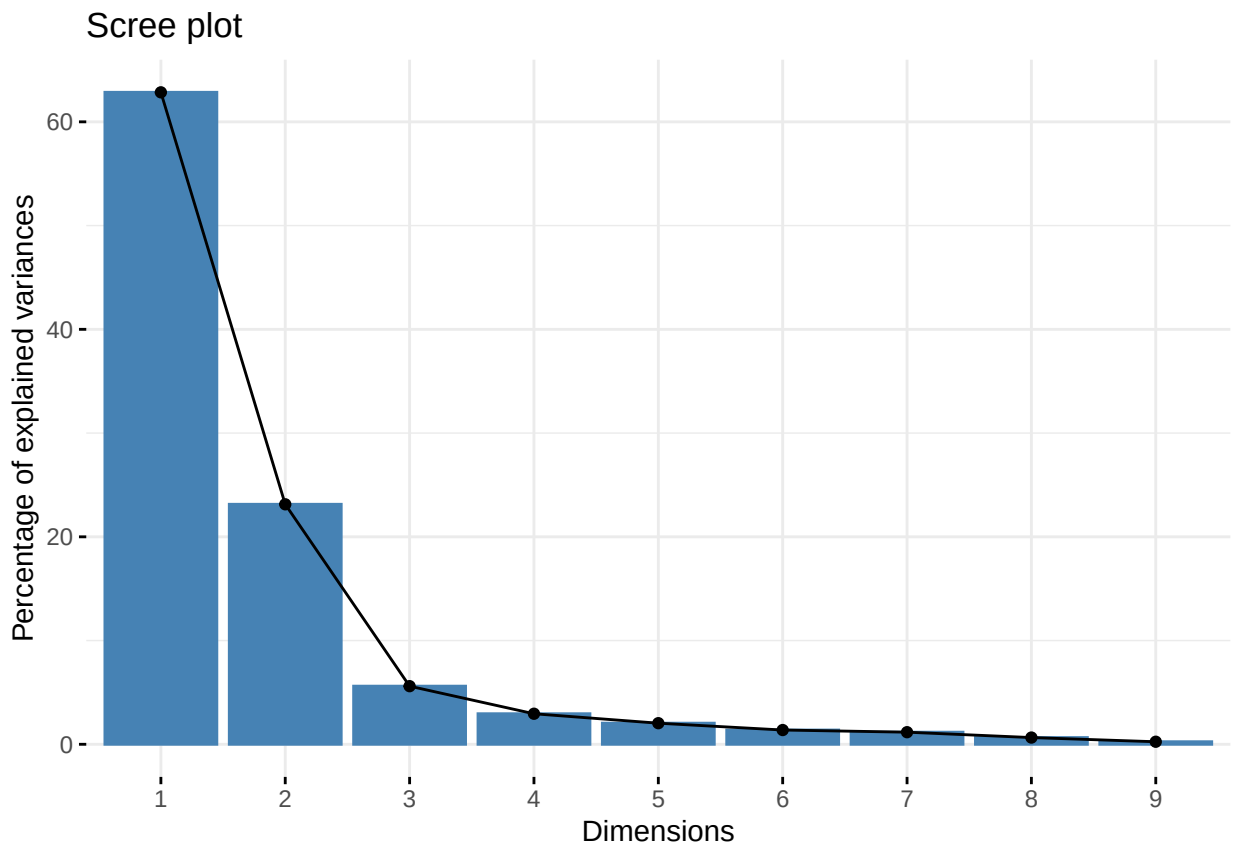
Contoh 2 Pada contoh ini kita akan mengaplikasikan PCA pada data set `mtcars` dari package `datasets` yang berisi 32 observasi (mobil) yang dihitung berdasarkan 11 variabel, lihat `?mtcars` untuk lebih detilnya. Di sini kita akan menggunakan semua variabel kecual variabel 8 dan 9 (tipe mesin dan tipe transmisi yang bersifat diskrit/kategori).

```
dataMPG <- mtcars[, -c(8,9)]
mtcarsPca <- prcomp(dataMPG, scale. = TRUE, center=TRUE)

fviz_pca_biplot(mtcarsPca, repel = TRUE,
                col.var = "#2E9FDF",
                col.ind = "#696969"
)
```




```
fviz_eig(mtcarsPca)
```



Latihan 1: dari Contoh 2

1. Berapa total variance yang dijelaskan dua principal component pertama?
2. Jelaskan variabel apa saja yang dijelaskan oleh principal component pertama dan kedua?
3. Dalam hal pemakaian bahan bakar (`mpg`), manakah yang lebih hemat antara Datsun 710 dan Merc 450SLC? Jelaskan dalam konteks biplot di atas.
4. Sebutkan mobil-mobil yang daya pacunya (`hp`) lebih besar ketimbang mobil lainnya.

Latihan 2

1. Lakukan analisis PCA terhadap `iris dataset`!
2. Berikan penjelasan pada setiap langkah yang anda lakukan dan buat visualisasinya!
3. Berikan kesimpulan anda dari hasil yang diperoleh!