Principal Component Analysis Mtcars Dataset and Iris Dataset

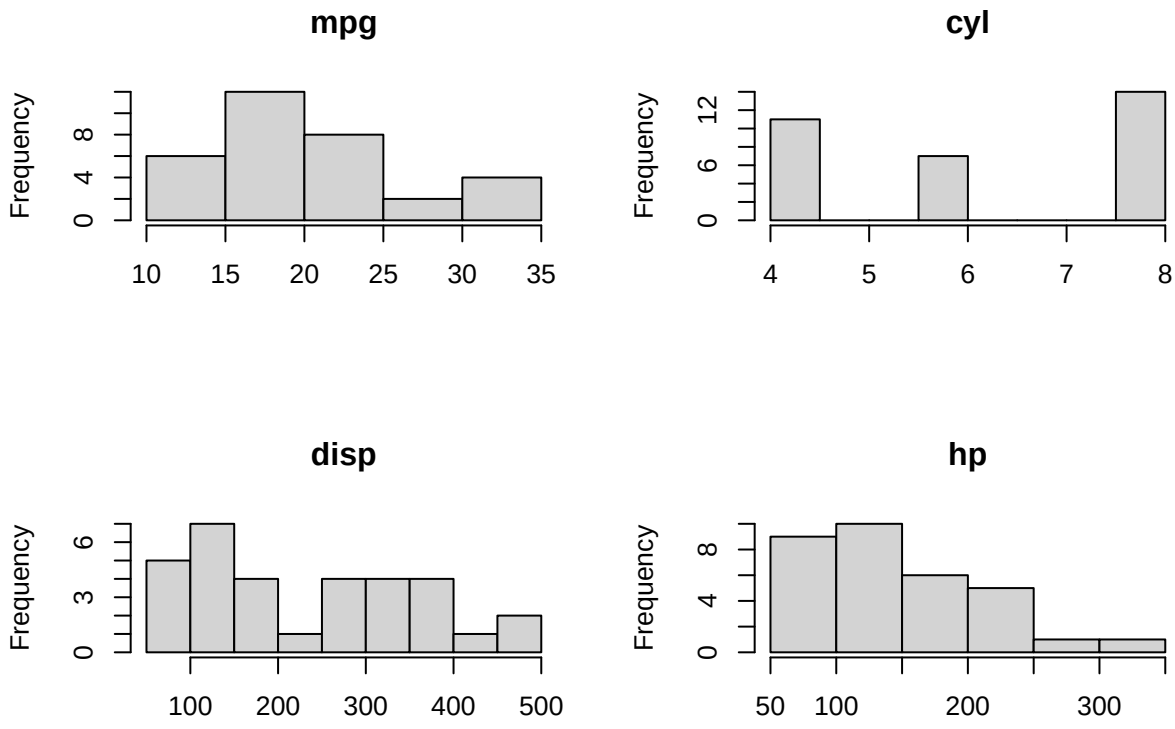
Mahesa Cadi Rajasa (19523122) Naufal Fadhlurohman (19523216)

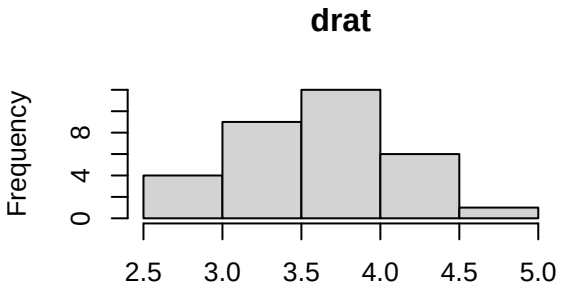
Latihan 1

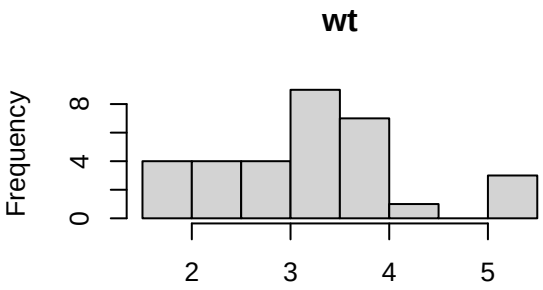
```
dataMPG <- mtcars[, -c(8,9)]
head(dataMPG)
```

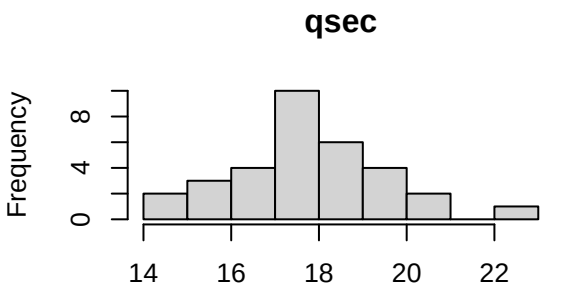
```
##                   mpg cyl disp  hp drat    wt  qsec gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22    3    1
```

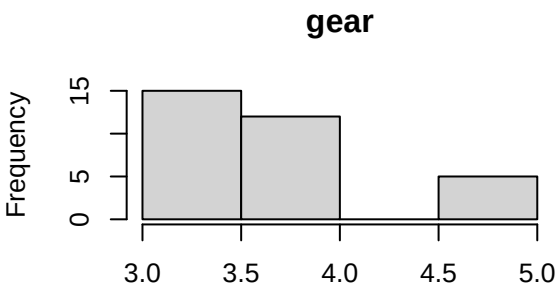
```
par(mfrow=c(2,2))
for(i in 1:ncol(dataMPG)) { hist(dataMPG[, i], main = paste(colnames(dataMPG[i])), xlab = "") }
```

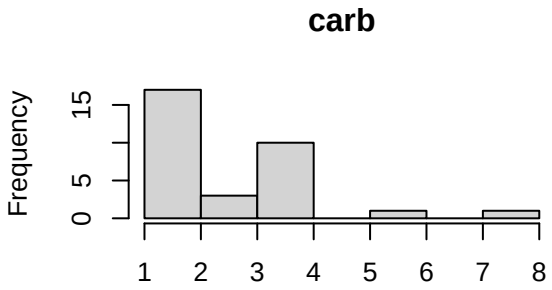



drat
Frequency
0
4
8
2.5
3.0
3.5
4.0
4.5
5.0


wt
Frequency
0
4
8
2
3
4
5


qsec
Frequency
0
4
8
14
16
18
20
22


gear
Frequency
0
5
15
3.0
3.5
4.0
4.5
5.0




```
mtcarsPca <- prcomp(dataMPG, scale. = TRUE, center=TRUE)
mtcarsPca$rotation
```

```
##             PC1         PC2         PC3          PC4        PC5         PC6
## mpg  -0.3931477  0.02753861 -0.22119309 -0.006126378 -0.3207620  0.72015586
## cyl   0.4025537  0.01570975 -0.25231615  0.040700251  0.1171397  0.22432550
## disp  0.3973528 -0.08888469 -0.07825139  0.339493732 -0.4867849 -0.01967516
## hp    0.3670814  0.26941371 -0.01721159  0.068300993 -0.2947317  0.35394225
## drat -0.3118165  0.34165268  0.14995507  0.845658485  0.1619259 -0.01536794
## wt    0.3734771 -0.17194306  0.45373418  0.191260029 -0.1874822 -0.08377237
## qsec -0.2243508 -0.48404435  0.62812782 -0.030329127 -0.1482495  0.25752940
## gear -0.2094749  0.55078264  0.20658376 -0.282381831 -0.5624860 -0.32298239
## carb  0.2445807  0.48431310  0.46412069 -0.214492216  0.3997820  0.35706914
##              PC7         PC8         PC9
## mpg  -0.38138068 -0.12465987  0.11492862
## cyl  -0.15893251  0.81032177  0.16266295
## disp -0.18233095 -0.06416707 -0.66190812
## hp    0.69620751 -0.16573993  0.25177306
## drat  0.04767957  0.13505066  0.03809096
## wt   -0.42777608 -0.19839375  0.56918844
## qsec  0.27622581  0.35613350 -0.16873731
## gear -0.08555707  0.31636479  0.04719694
## carb -0.20604210 -0.10832772 -0.32045892
```

```
library(factoextra)
```

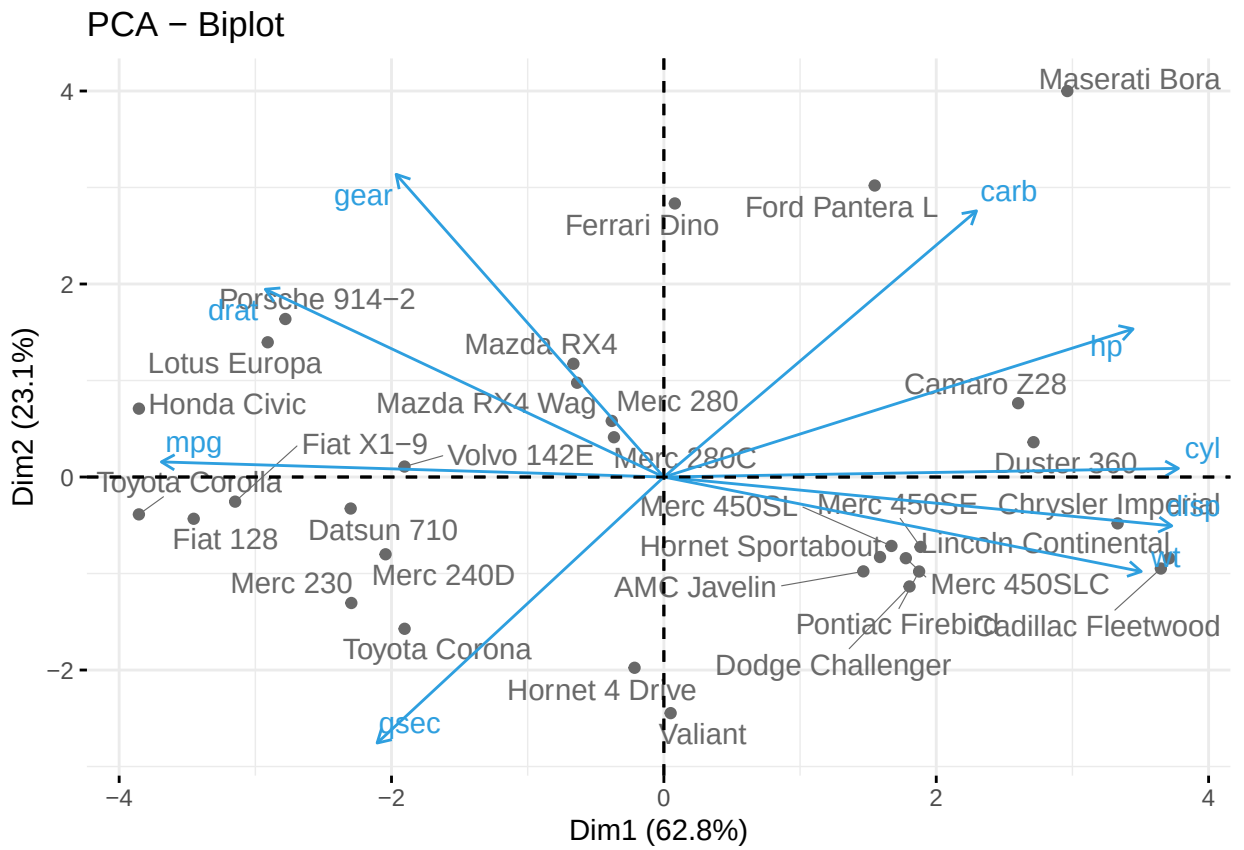
```
## Warning: package 'factoextra' was built under R version 4.0.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

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



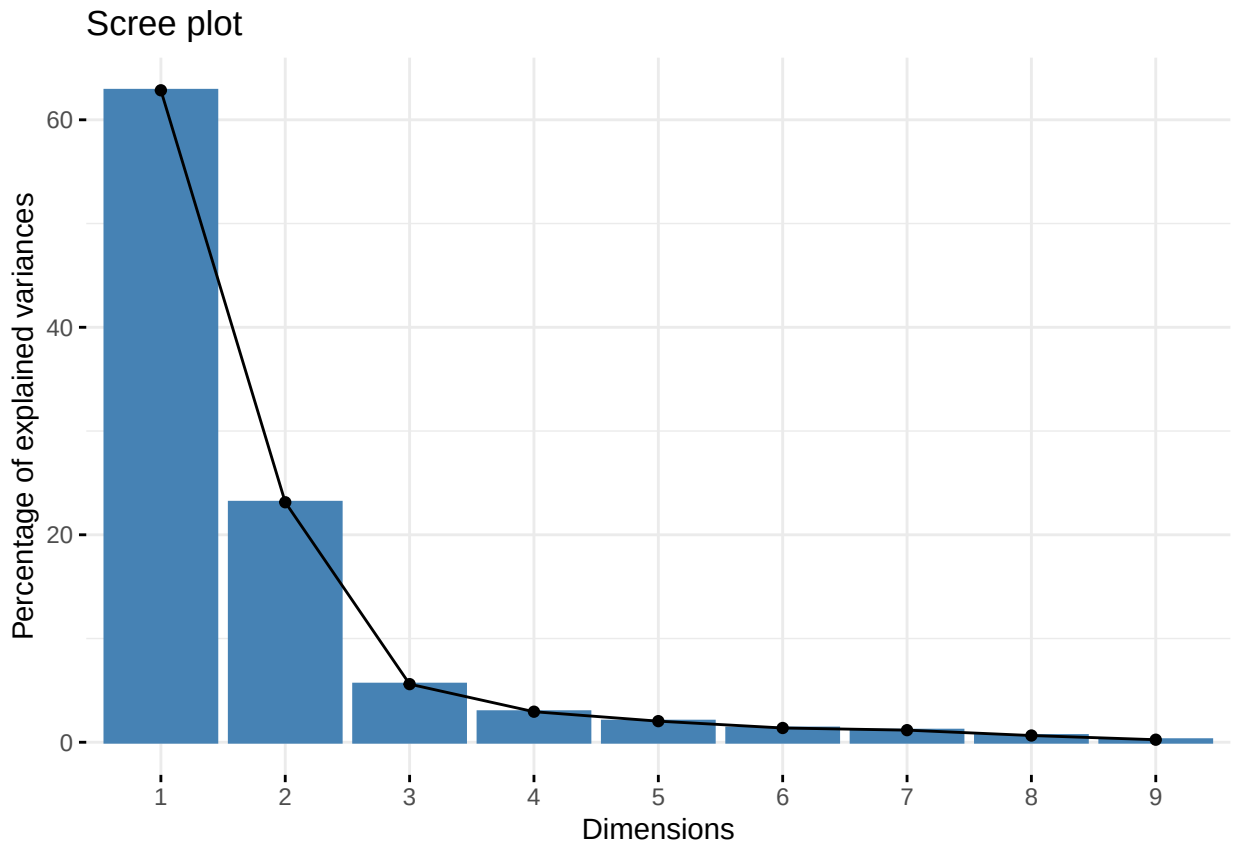
1. Dari biplot di atas, sumbu horizontal (Dim1) merepresentasikan principal component yang pertama, yang mengandung 62.8% variance dari seluruh data set; sumbu vertikal (Dim2) merepresentasikan principal compo- nent kedua, yang mengandung 23.1% variance. Dari dua component ini saja, 85.9% variance atau informasi yang dikandung data dapat dijelaskan.

2. Dari plot tersebut, dapat kita lihat arah vektor qsec dan gear cenderung vertikal seperti arah principal component yang pertama(Dim1). Ini mengindikasikan bahwa variable qsec dan gear lebih banyak dijelaskan/diwakili oleh principal komponen yang pertama. Sebaliknya, arah vektor drat, mpg, carb, hp, cyl, disp, dan wt lebih mendekati arah principal component yang kedua (Dim2). Ini mengindikasikan jika informasi yang dibawa variabel drat, mpg, carb, hp, cyl, disp, dan wt lebih banyak diwakili oleh principal component yang kedua.

3. Dalam pemakaian bahan bakar (mpg), Datsun 710 lebih hemat pemakaian bahan bakarnya dibandingkan Merc 450SLC karena lebih searah dengan vektor (mpg). Sebaliknya Merc 450SLC boros dalam pemakaian bahan bakar karena berlawanan arah dengan vektor (mpg).

4. Mobil-mobil yang daya pacunya besar (hp) dari mobil lain-lainnya yaitu: Camaro Z28, Duster 360.

```
fviz_eig(mtcarsPca)
```



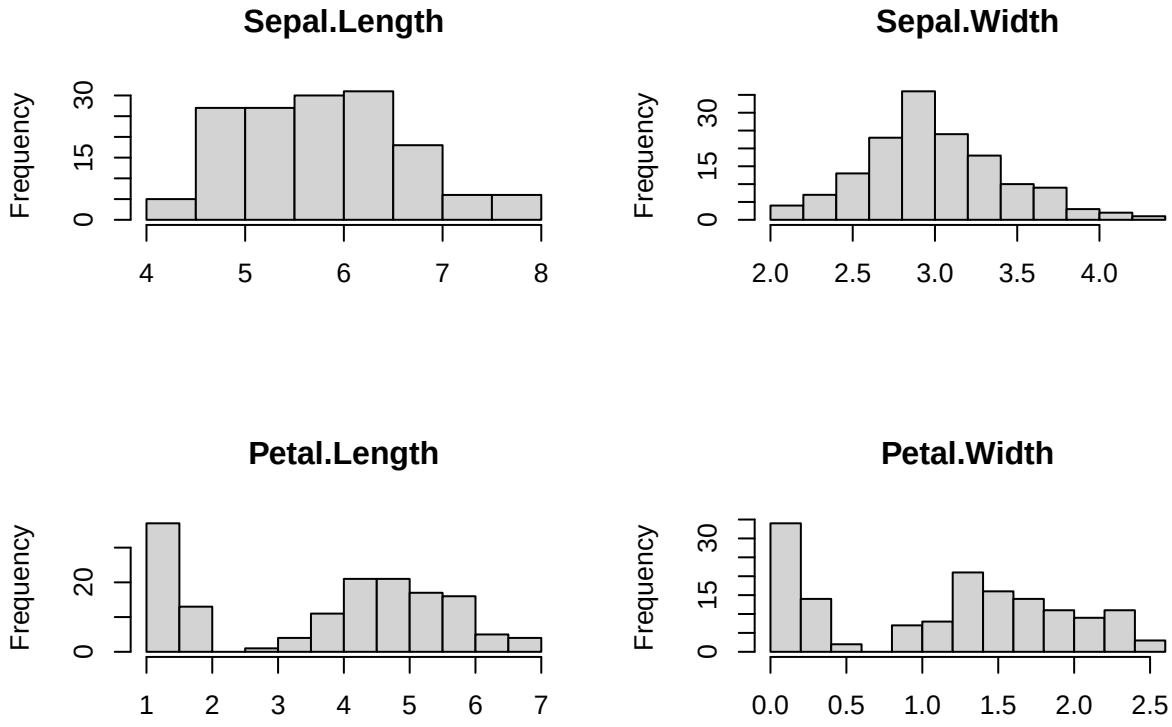
Latihan 2

Kami akan menggunakan data set Iris dari package datasets, data berisi 3 kelas masing-masing 50 instans, di mana setiap kelas mengacu pada jenis tanaman iris. Pertama, mari kita lihat bentuk dan sebaran datanya.

```
dataIris <- iris[, -c(5,6)]
head(dataIris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

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



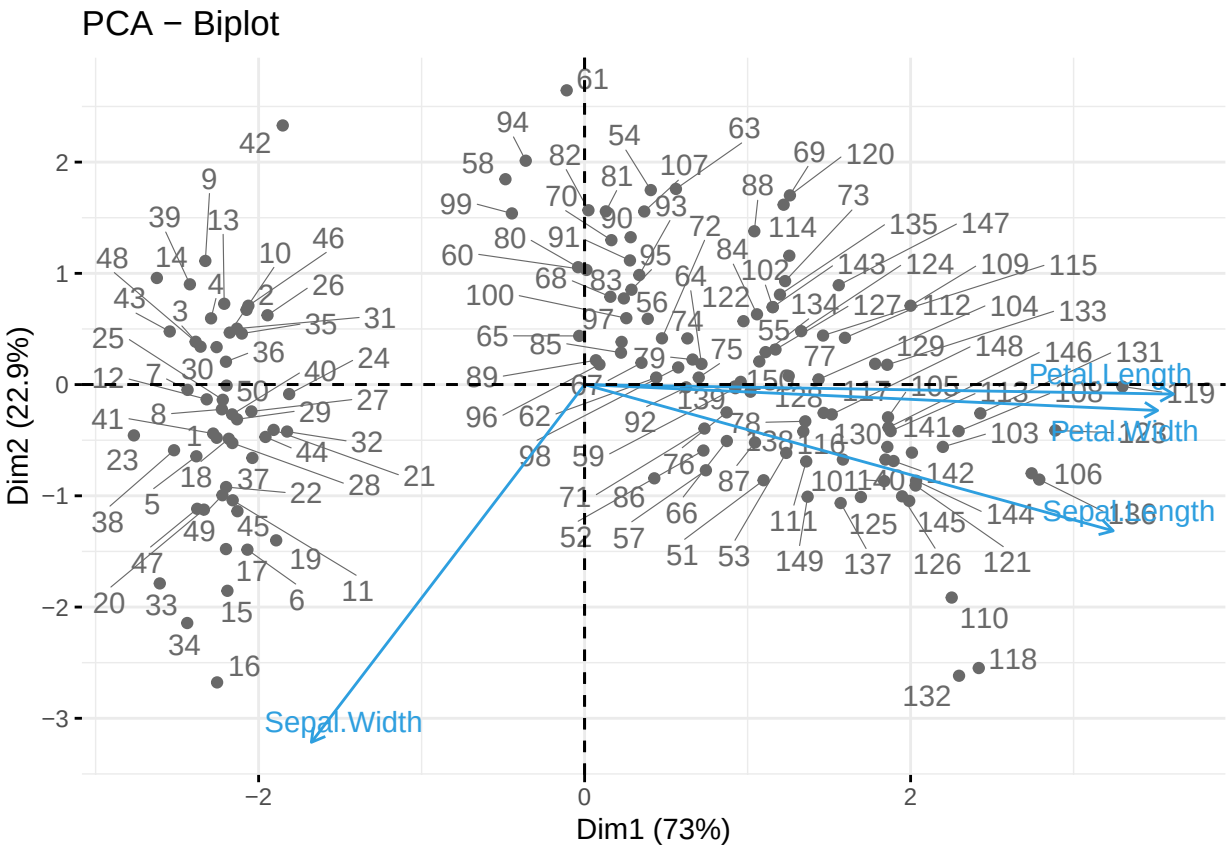
Selanjutnya, kami akan mengaplikasikan PCA pada data tersebut, menggunkaan fungsi prcomp() dari package stats. Parameter scale=TRUE, center=TRUE memastikan nilai setiap variabel distandarisasi, sehingga semua variabel memiliki skala nilai yang sama.

```
irisPCA <- prcomp(dataIris, scale. = TRUE, center=TRUE)
irisPCA$rotation
```

```
##                     PC1         PC2        PC3        PC4
## Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
## Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
## Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
## Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Selanjutnya, agar mudah kami memvisualisasikan biplot nya.

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

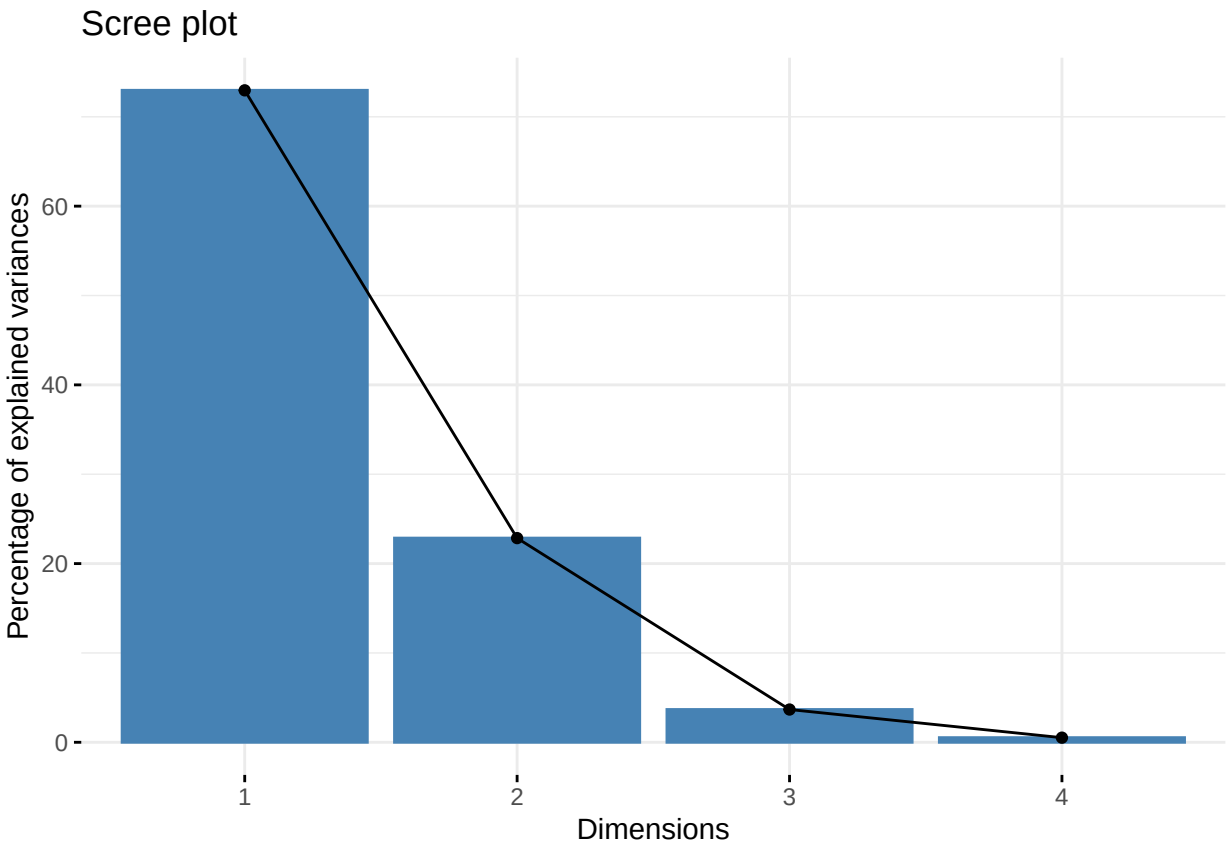


Dari biplot di atas, sumbu horizontal (Dim1) merepresentasikan principal component yang pertama, yang mengandung 73% variance dari seluruh data set; sumbu vertikal (Dim2) merepresentasikan principal component kedua, yang mengandung 22.9% variance. Dari dua component ini saja, 95.9% variance atau informasi yang dikandung data dapat dijelaskan.

Dari plot tersebut, dapat kita lihat arah vektor Sepal.Width cenderung vertikal seperti arah principal component yang pertama(Dim1). Ini mengindikasikan bahwa variable Sepal.Widthr lebih banyak dijelaskan/diwakili oleh principal komponen yang pertama. Sebaliknya, arah vektor Sepal.Length, Petal.Width, dan Petal.Length lebih mendekati arah principal component yang kedua (Dim2). Ini mengindikasikan jika informasi yang dibawa variabel Sepal.Length, Petal.Width, dan Petal.Length lebih banyak diwakili oleh principal component yang kedua.

Selanjutnya kami memvisualisasikan variance yang dibawa setiap component melalui screeplot sebagai berikut.

```
fviz_eig(irisPCA)
```



Kesimpulan Kita hanya butuh dua "variabel" (principal component) saja untuk merepresentasikan hampir 96% informasi yang ada di data. Ini adalah semangat dimensionality reduction yang menjadi tujuan PCA.