Assignment 15

Anthony Cunningham

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)

# Change working dir in RMarkdown cell
knitr::opts_knit$set(root.dir =
'C:/Users/AC069015/kumc_applied_stats/data_824_data_viz_and_acquisition'
)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
library(readr)
library(datasets)
library(readxl)
library(circlize)
```

```
## Warning: package 'circlize' was built under R version 4.2.1
```

```
## ========================================
## circlize version 0.4.15
## CRAN page: https://cran.r-project.org/package=circlize
## Github page: https://github.com/jokergoo/circlize
## Documentation: https://jokergoo.github.io/circlize_book/book/
##
## If you use it in published research, please cite:
## Gu, Z. circlize implements and enhances circular visualization
##   in R. Bioinformatics 2014.
##
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(circlize))
## ========================================
```

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.1
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.1
```

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

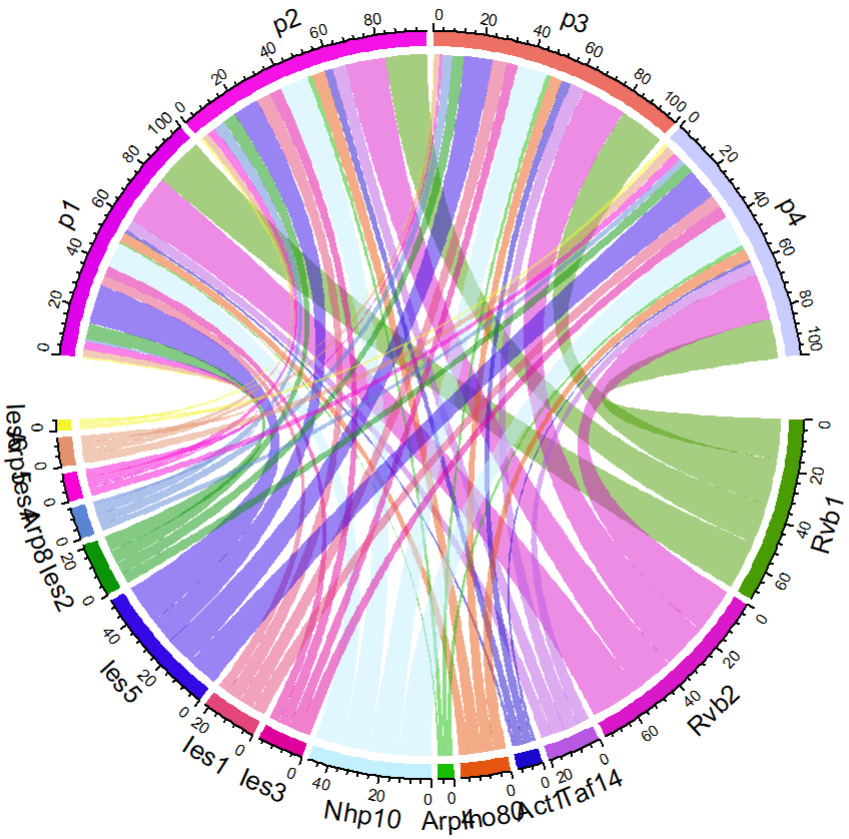
```
library(pheatmap)
```

```
## Warning: package 'pheatmap' was built under R version 4.2.1
```

```
library(cluster)
library(factoextra)
```

Exercise 1

```
dta <- read_excel("datasets/Interaction_proteins.xls", skip=1)

dta_new <- dta %>% mutate(
    p1 = rowMeans(select(., contains("_P1")))*100,
    p2 = rowMeans(select(., contains("_P2")))*100,
    p3 = rowMeans(select(., contains("_P3")))*100,
    p4 = rowMeans(select(., contains("_P4")))*100
) %>% select(p1, p2, p3, p4)

rownames(dta_new) <-dta$Proteins
chordDiagram(as.matrix(dta_new))
```



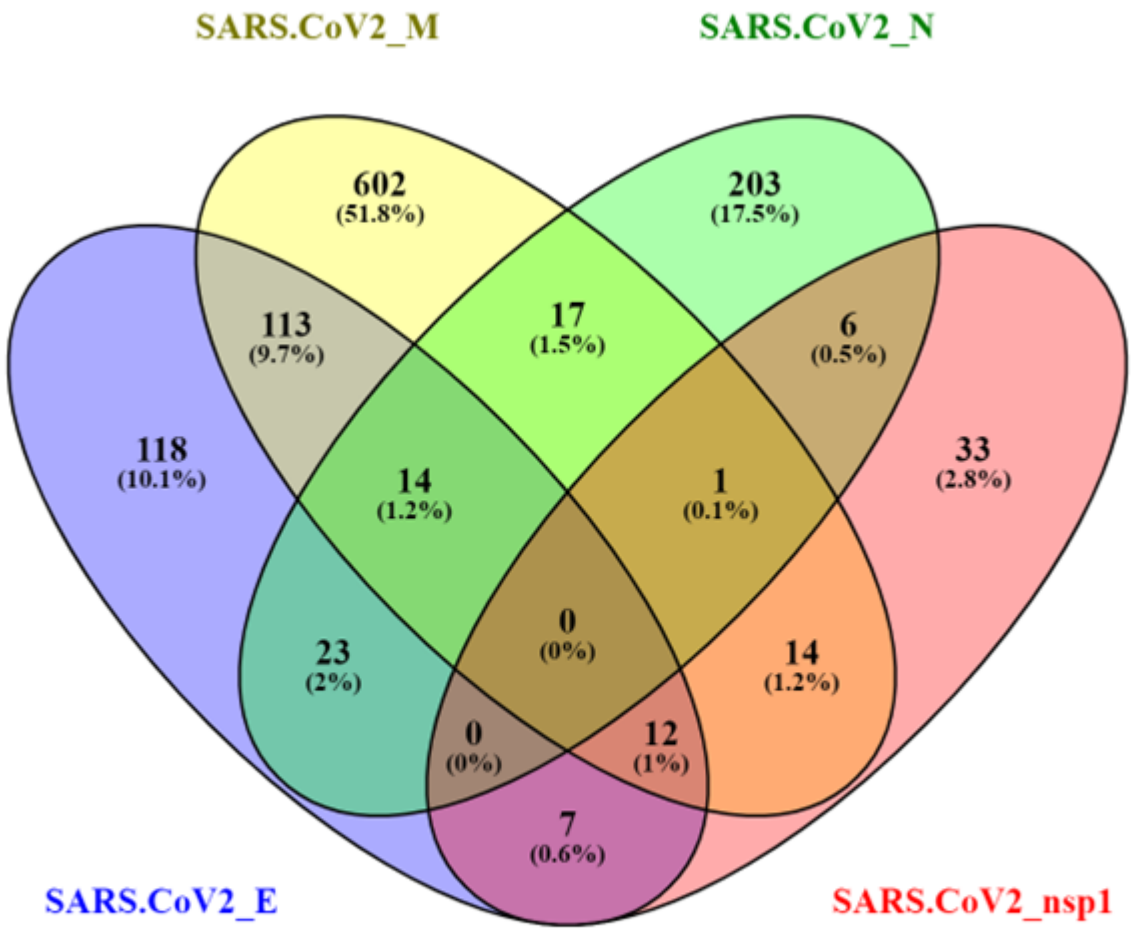
This circle diagram visually depicts the average value for 16 proteins (labeled at the bottom-half) across 4 separate replicates. Larger widths indicate higher averages for a particular protein. From the diagram, we notice that the proteins with the highest average values are Rvb1 and Rvb2, followed by Ies5 and Nhp10. On the other hand, proteins Ies3 and Ies6 consistently displayed the smalled average values.

Exercise 2

Examples of Missing Data Structures:

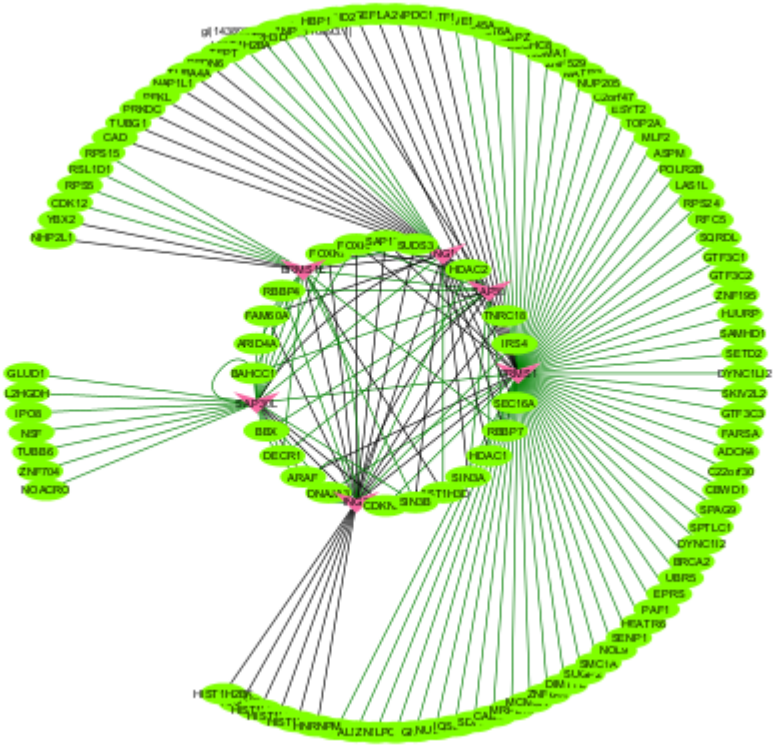
- MCAR: Blood pressure monitor malfunctions at random times, resulting in blood pressure measures not being recorded for some routine patient visits.
- MAR: An overworked nurse in the emergency department doesn't capture patient weight or blood pressure measurements during very busy periods in order to maximize patient throughput. While missingness of these measures aren't related to the values themselves, they are missing for non-random patients (associated with time and missingness of the other measures that are also not captured).
- MNAR: An outpatient clinic asks all patients about whether (and to what extent) they use tobacco, marijuana, alcohol and hard drugs. It's plausible that users of substances that are banned or illegal in the clinic's area are more likely to not answer this question than non-users (i.e. response is missing), for fear of disclosure to law enforcement agencies.

Exercise 3



Venn Diagram

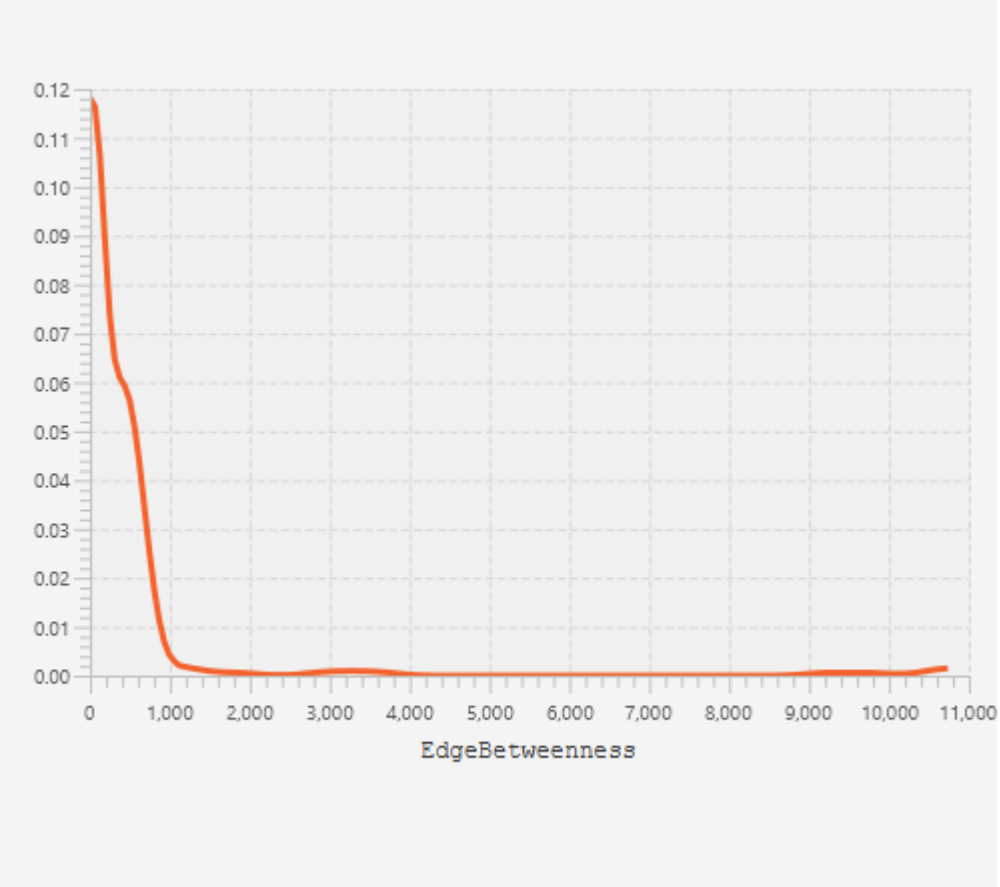
Exercise 4



Network Graph – Circular Layout

Network Properties:

- Clustering Coefficient = 0.111 – this indicates a large degree of separation between nodes, on average (i.e. a randomly-selected node is unlikely to have triangular connections)
- Degree Distribution: This network appears to be a scale-free network, with only a few nodes with many edges; most nodes only have a few edges:

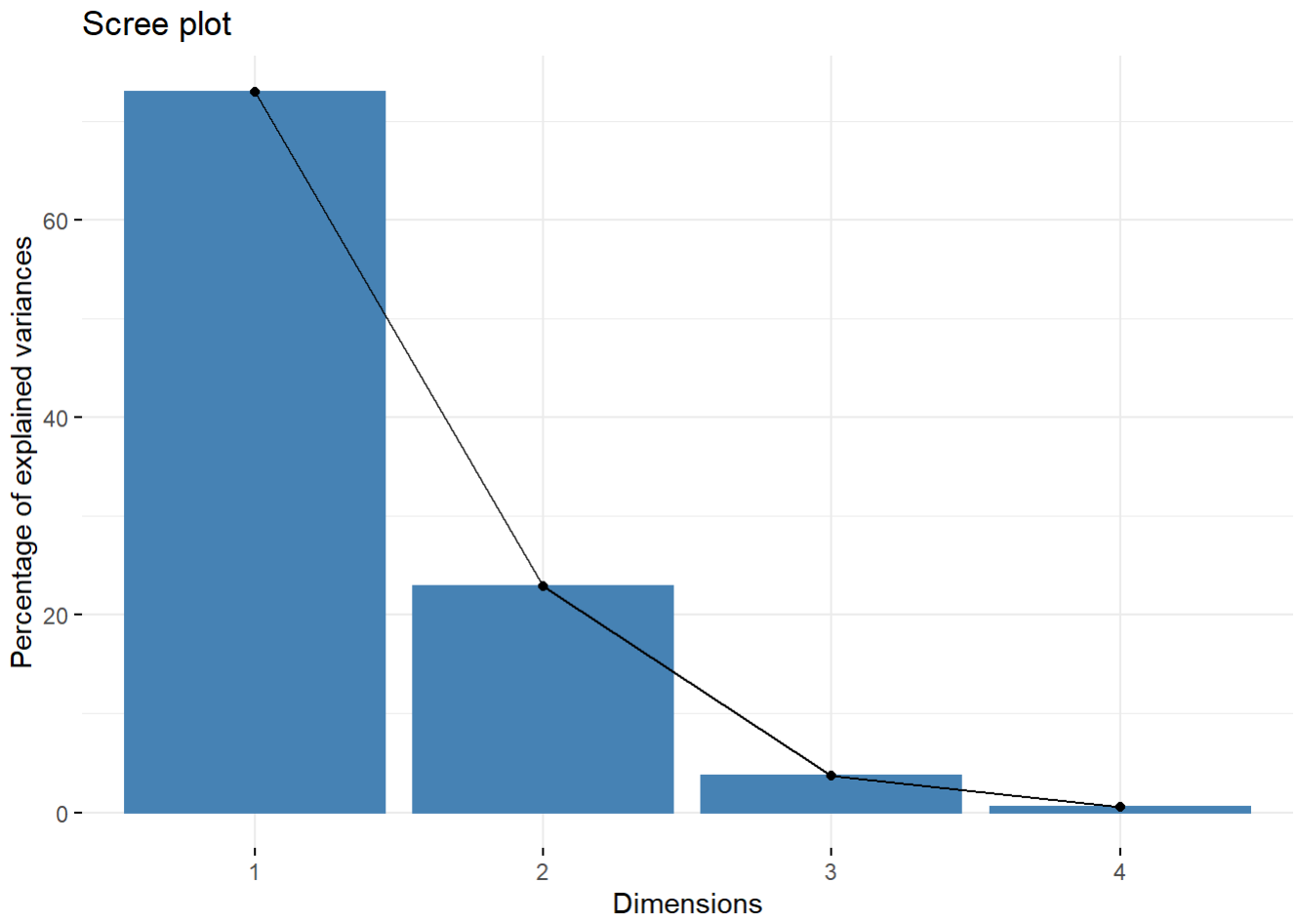


- Diameter = 4, which indicates that this network is not very extensive: the maximum number of nodes that the 2 nodes furthest away from each other is is 4.
- Avg. Path Length = 2.267, again, indicating a shallow network.

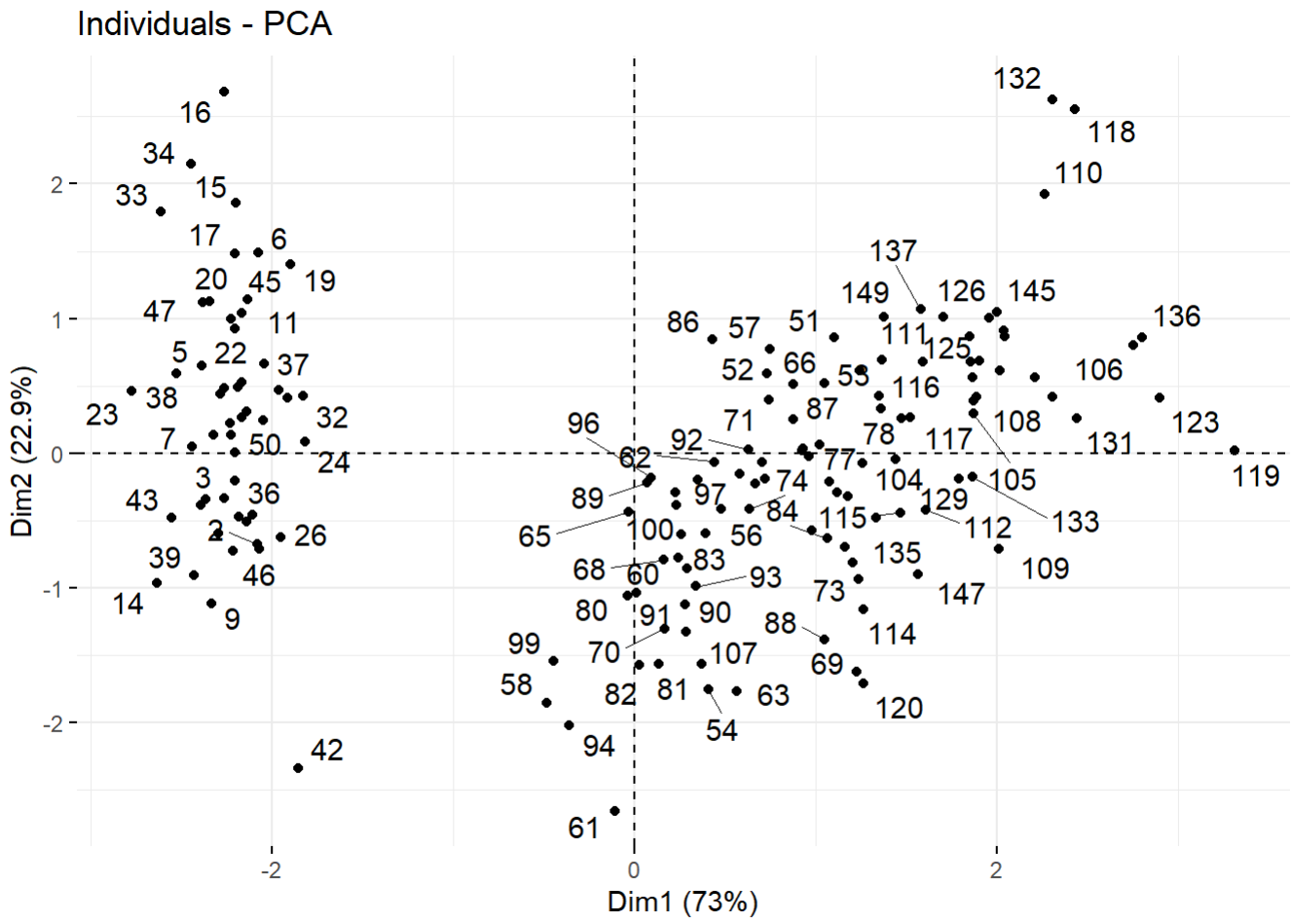
Exercise 5

- Use any data (iris is a good data) to visualize the relationship between variables using PCA, CA, or MCA.

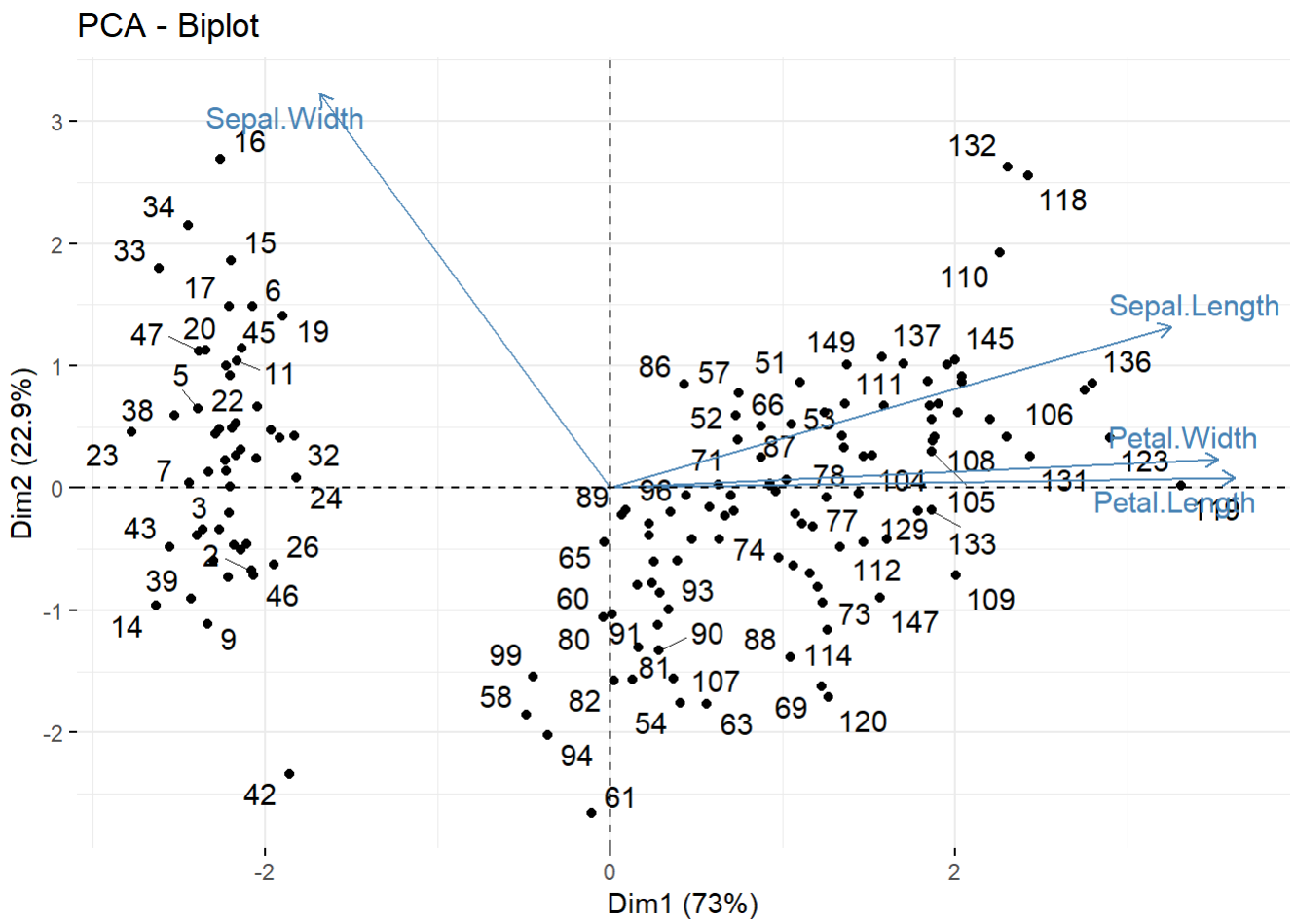
```
data("iris")

res.PCA <- PCA(iris %>% select(-Species), graph = FALSE)
fviz_eig(res.PCA)
```




```
fviz_pca_ind(res.PCA, repel = TRUE)
```

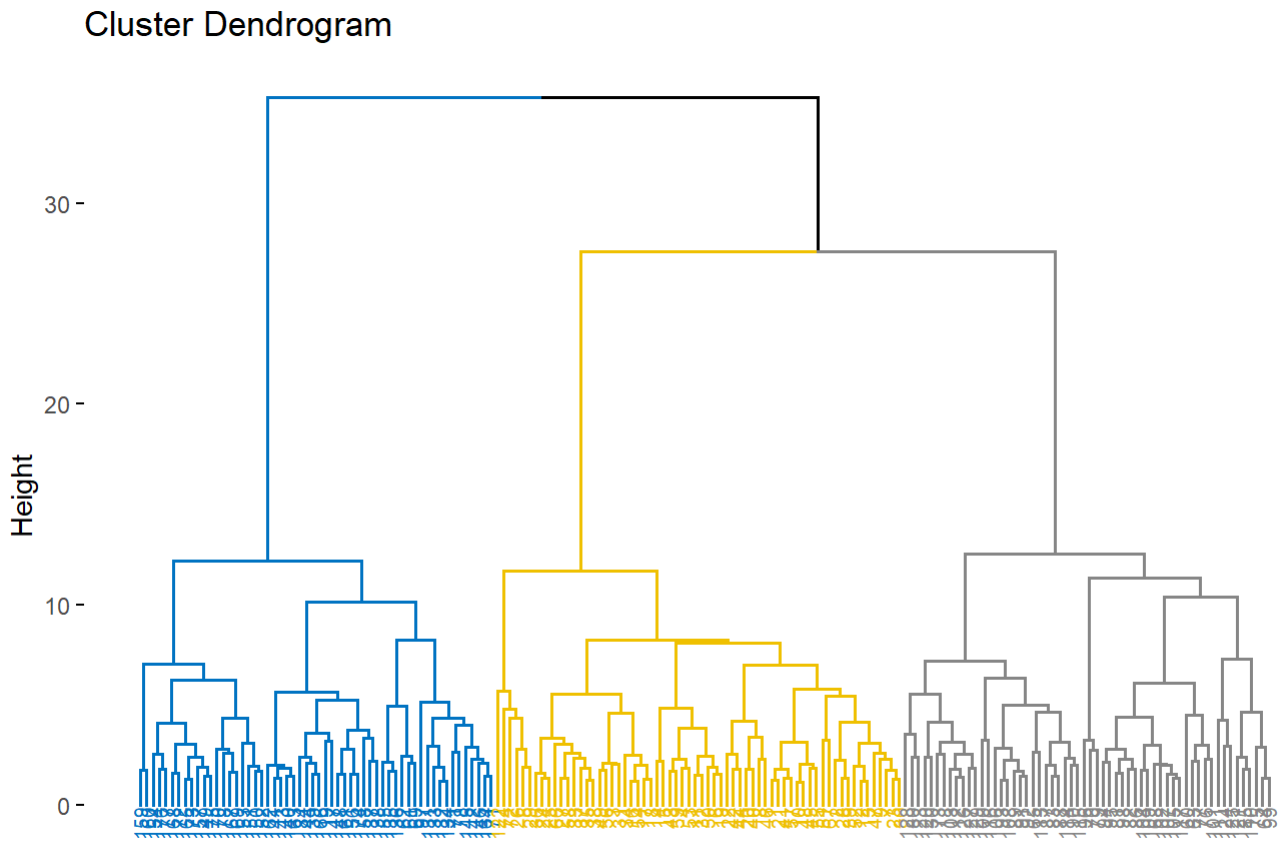



```
fviz_pca_biplot(res.PCA, repel = TRUE)
```

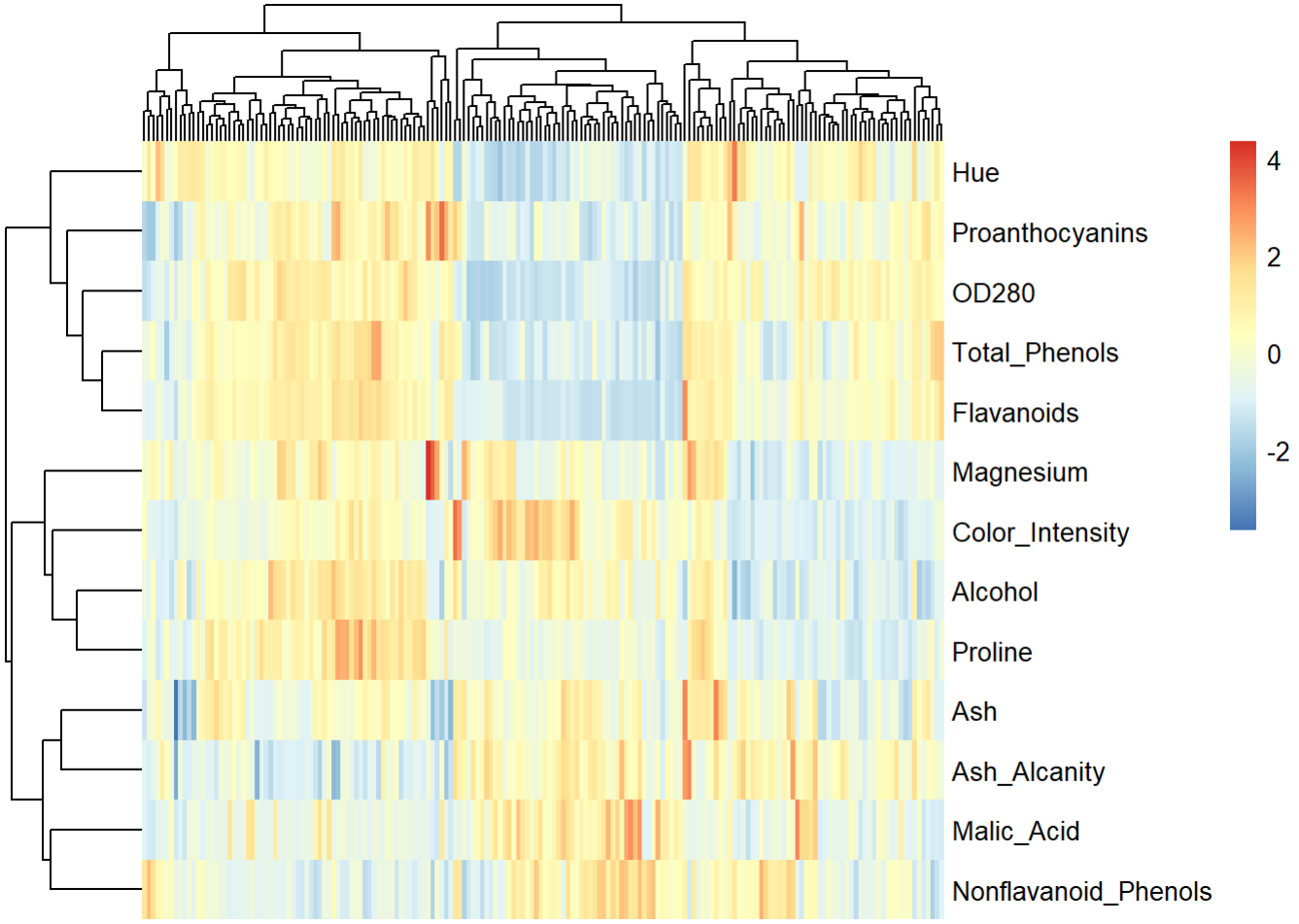


- Use any data to visualize variables and observations using heat map and hierarchical clustering method. You can use the same data that you used when doing clustering in Clustvis.

```
wine <- read_csv("datasets/wine-clustering.csv")
wine_scaled <- scale(wine)

res.hc <- hclust(dist(wine_scaled), method = "ward.D2")
fviz_dend(res.hc, cex=0.5, k=3, palette="jco")
```




```
pheatmap(t(wine_scaled), cuttree_cols=3)
```


Hue
Proanthocyanins
OD280
Total_Phenols
Flavanoids
Magnesium
Color_Intensity
Alcohol
Proline
Ash
Ash_Alcanity
Malic_Acid
Nonflavanoid_Phenols
4
2
0
-2