

# Análisis Descriptivo de Datos Multivariados



José A. Perusquía Cortés

Análisis Multivariado Semestre 2024-1



¿Qué es el análisis multivariado?

&

¿Qué tipo de datos nos interesan?

- El estudio de “muchas” variables

- El estudio de “muchas” variables **correlacionadas**.

- El estudio de “muchas” variables **correlacionadas**.
- Para  $n$  variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in S^p$ , i.e.,  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$  se tiene la notación

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- El estudio de “muchas” variables **correlacionadas**.
- Para  $n$  variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in S^p$ , i.e.,  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$  se tiene la notación

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

- El estudio de “muchas” variables **correlacionadas**.
- Para  $n$  variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in S^p$ , i.e.,  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$  se tiene la notación

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$$

$$\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(p)})$$

- ▶ Problemas de interés
  - Graficar/describir la estructura de los datos
  - Selección de variables
  - Aprendizaje supervisado, semi-supervisado y no supervisado
  - Analizar correlación entre variables
  - Etc...

- ▶ Problemas de interés
  - Graficar/describir la estructura de los datos
  - Selección de variables
  - Aprendizaje supervisado, semi-supervisado y no supervisado
  - Analizar correlación entre variables
  - Etc...
- ▶ Retos
  - $n >> 1, p >> 1$
  - $p > n$

- En R generalmente representados a través de data frames

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

- En R generalmente representados a través de data frames

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

- ¿Cómo podemos visualizar los datos?

# Diagramas de Dispersión y de Correlación

# Diagrama de Dispersion

- Graficar todas las variables contra todas las variables

# Diagrama de Dispersion

- Graficar todas las variables contra todas las variables
- Útil para:
  - Observar la relación por pares entre las variables (e.g. lineal, no lineal)
  - Identificar el tipo de correlación por pares entre ellas

# Diagrama de Dispersion

- Graficar todas las variables contra todas las variables

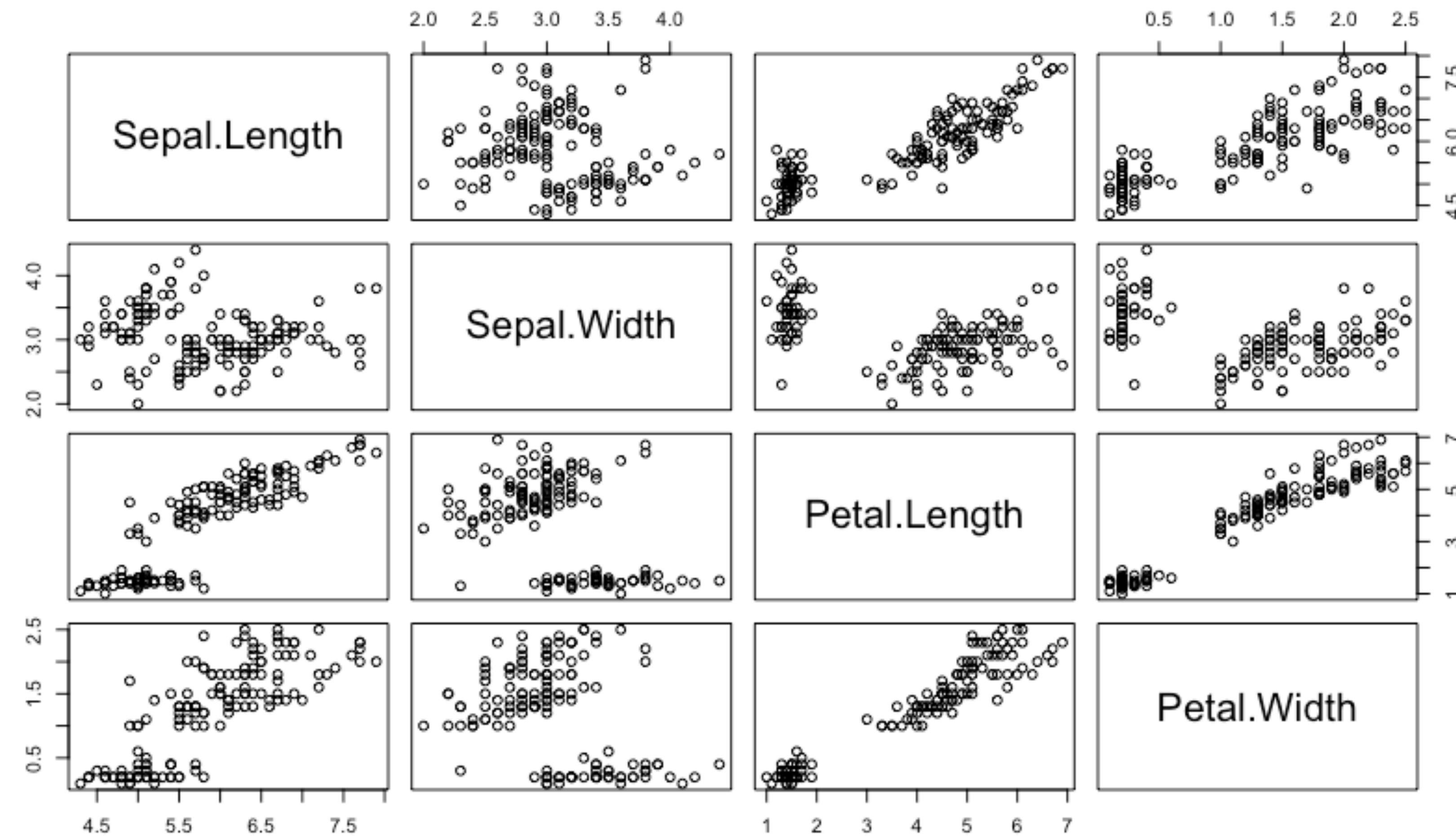
- Útil para:

- Observar la relación por pares entre las variables (e.g. lineal, no lineal)
- Identificar el tipo de correlación por pares entre ellas

- Desventajas:

- Solo se puede analizar a las variables por pares
- Muy difícil de graficar/analizar si se tienen muchas variables

# Diagrama de Dispersion



# Diagrama de Correlación

-Graficar la correlación por pares de las variables

# Diagrama de Correlación

- Graficar la correlación por pares de las variables
- Útil para:
  - Identificar el tipo y el grado de correlación por pares entre ellas

# Diagrama de Correlación

- Graficar la correlación por pares de las variables

- Útil para:

- Identificar el tipo y el grado de correlación por pares entre ellas

- Desventajas:

- Solo se puede analizar a las variables por pares
- Muy difícil de graficar/analizar si se tienen muchas variables

# Diagrama de Correlación

- Graficar la correlación por pares de las variables

- Útil para:

- Identificar el tipo y el grado de correlación por pares entre ellas

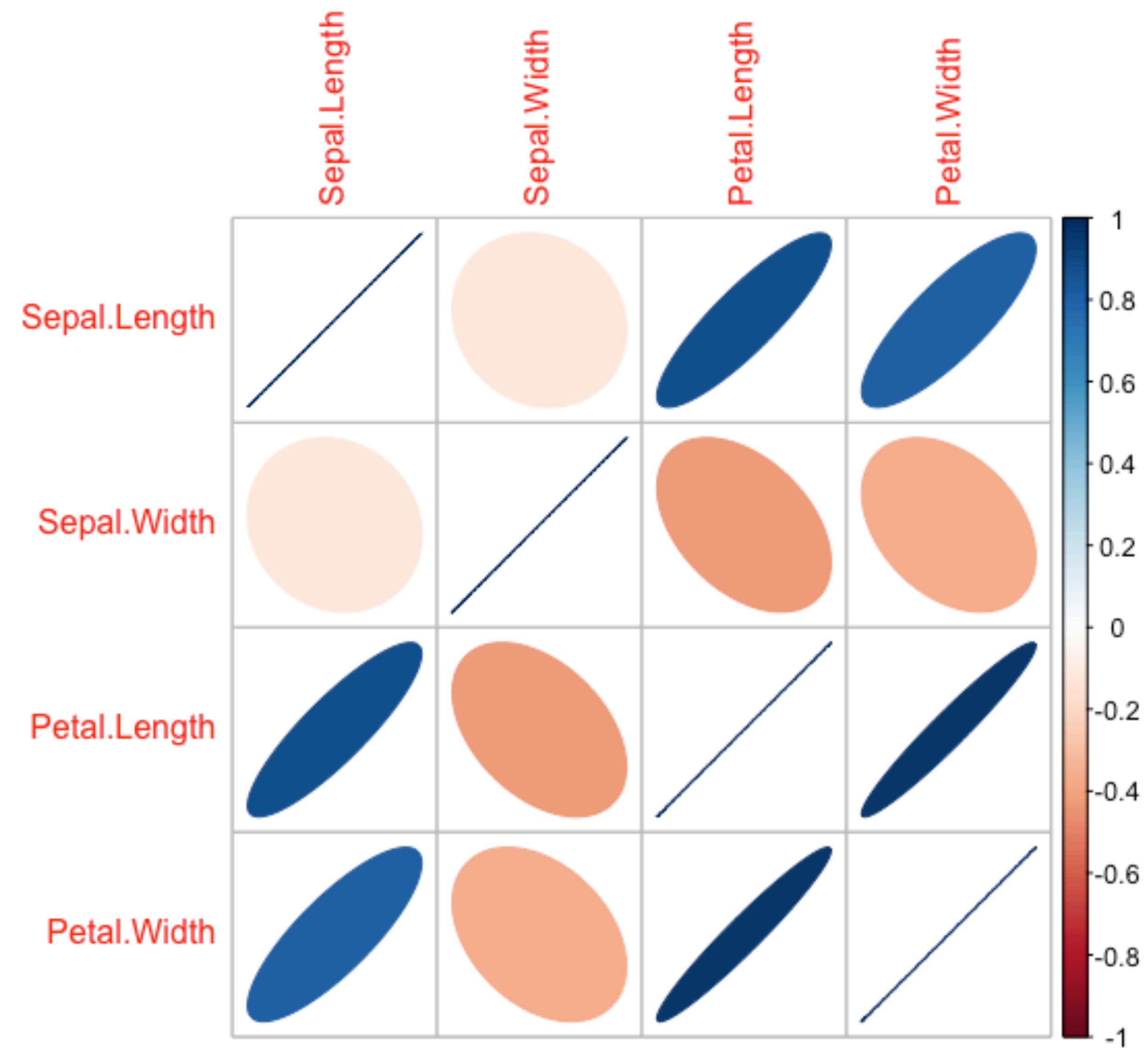
- Desventajas:

- Solo se puede analizar a las variables por pares
- Muy difícil de graficar/analizar si se tienen muchas variables

- En R:

- Librería: `corrplot`

# Diagrama de Correlación

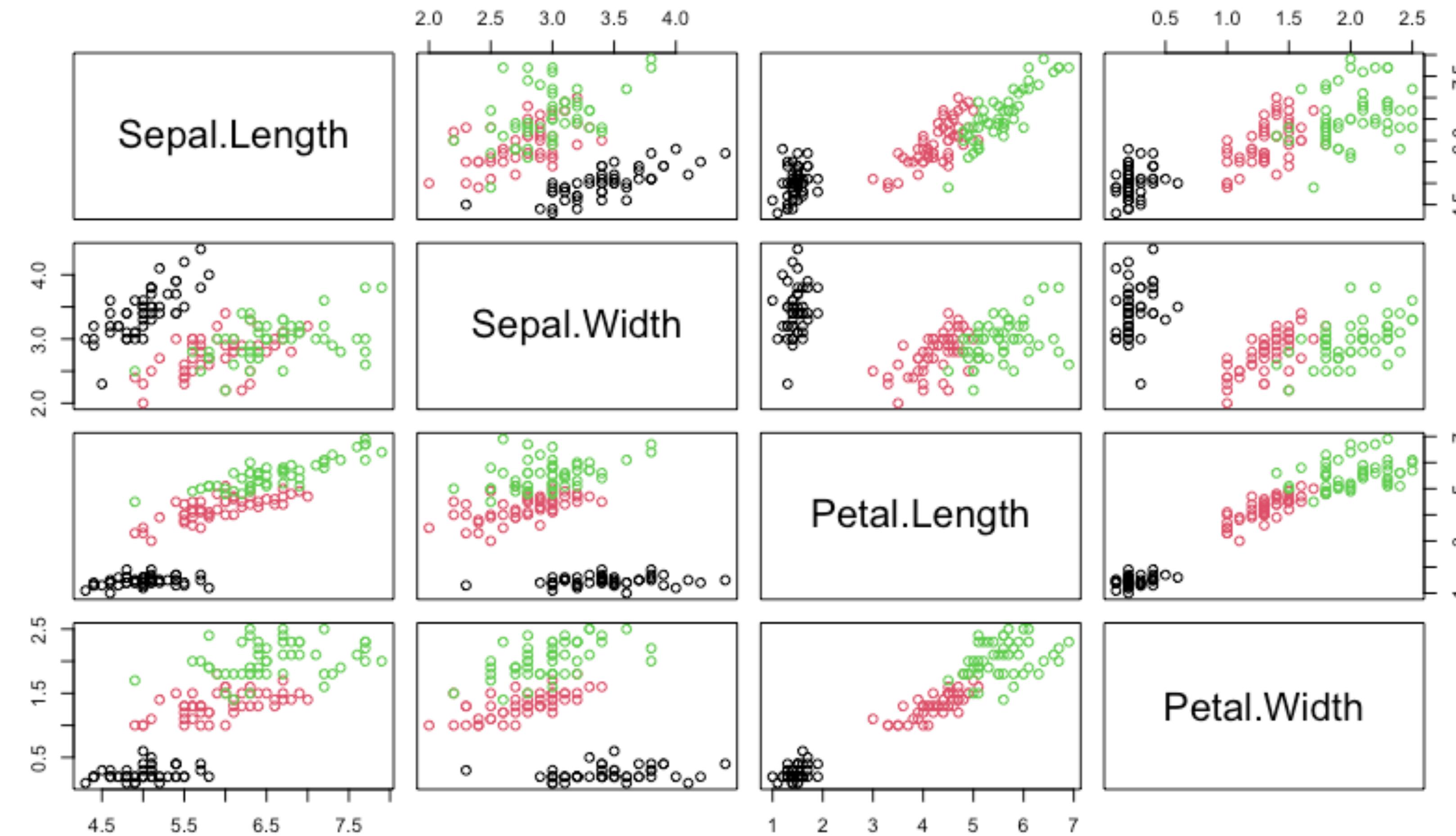


## Diagrama de Dispersión II

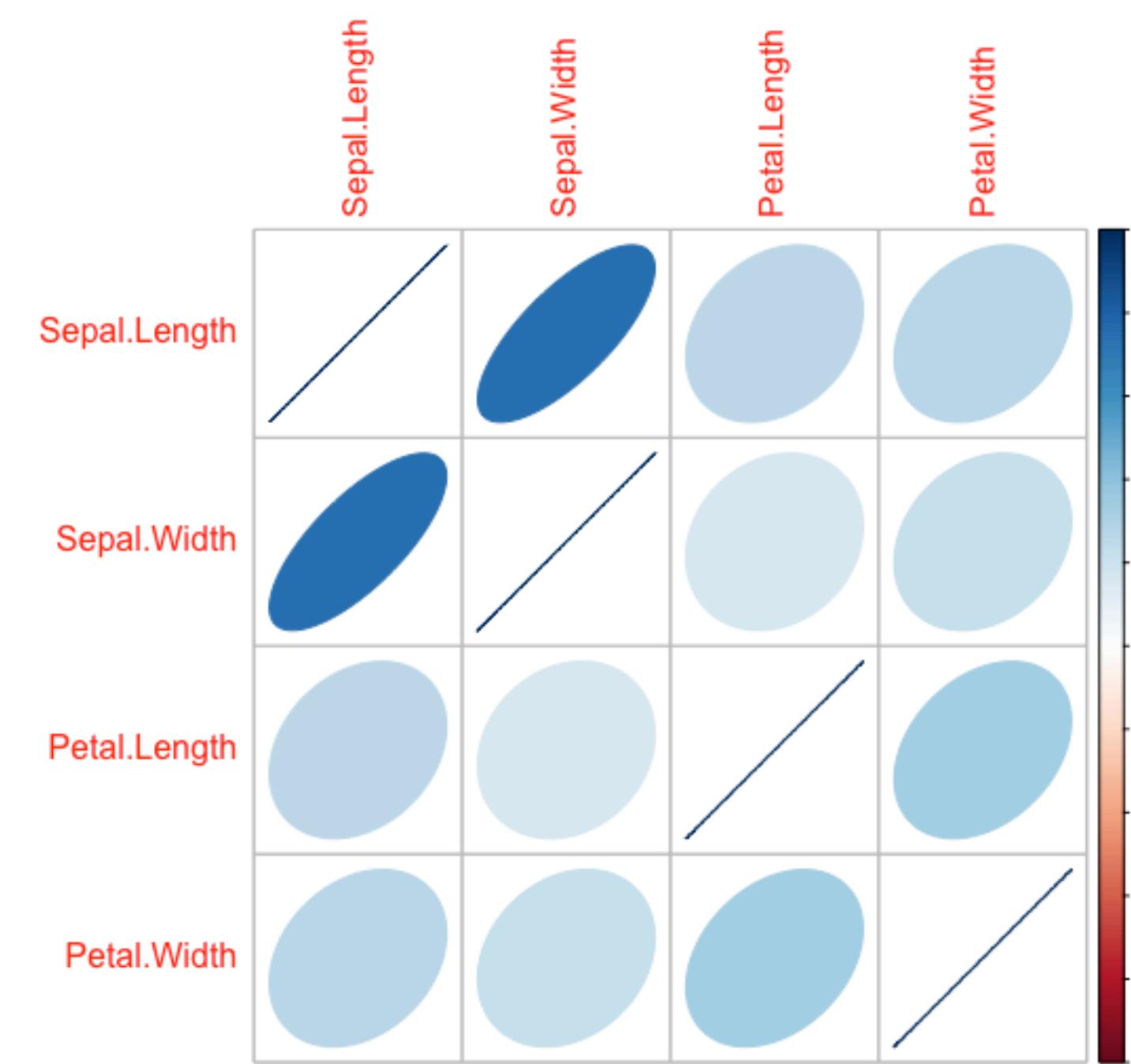
- ¿Qué sucede si se utiliza la información de la especie?

## Diagrama de Dispersion II

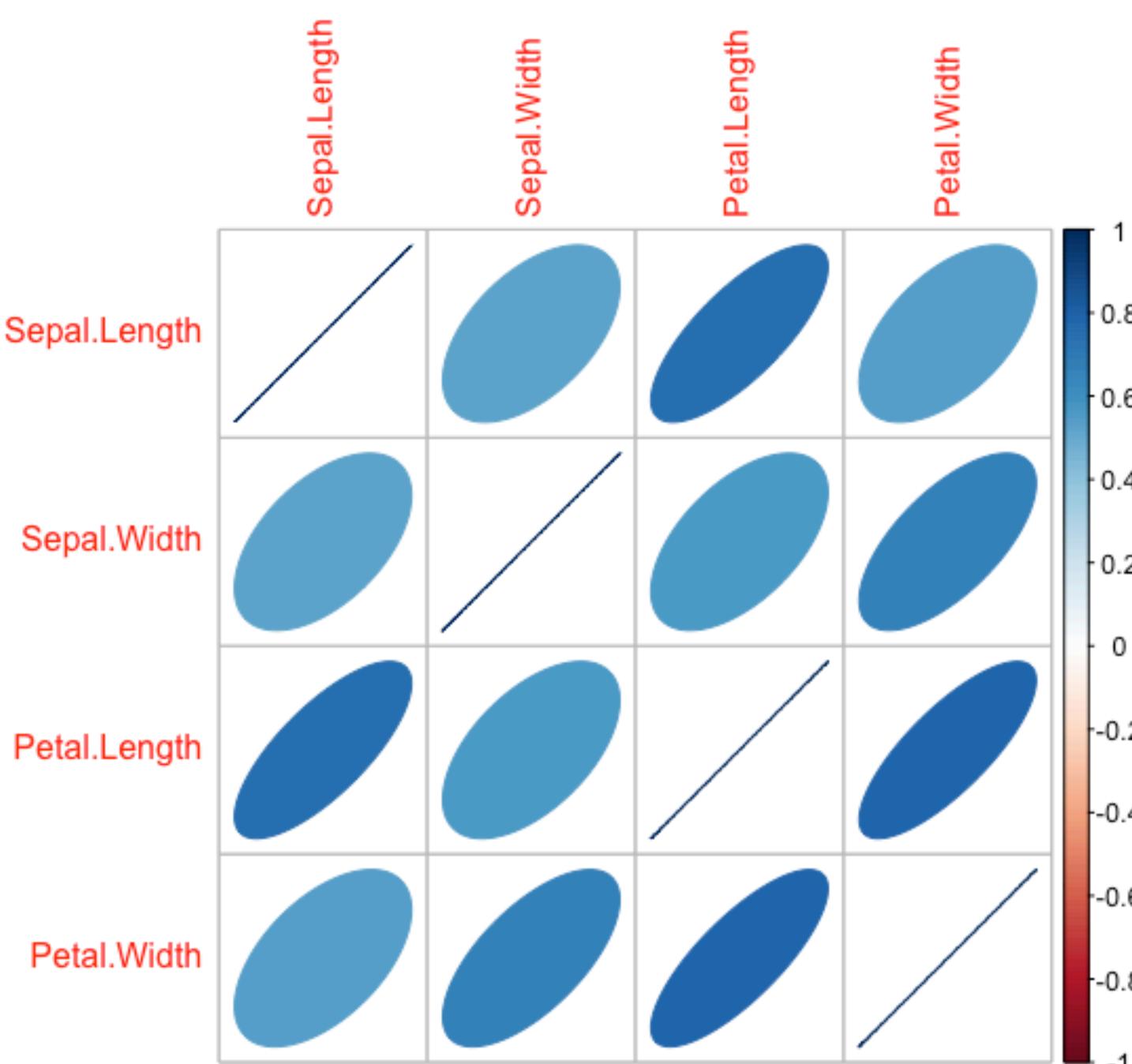
- ¿Qué sucede si se utiliza la información de la especie?



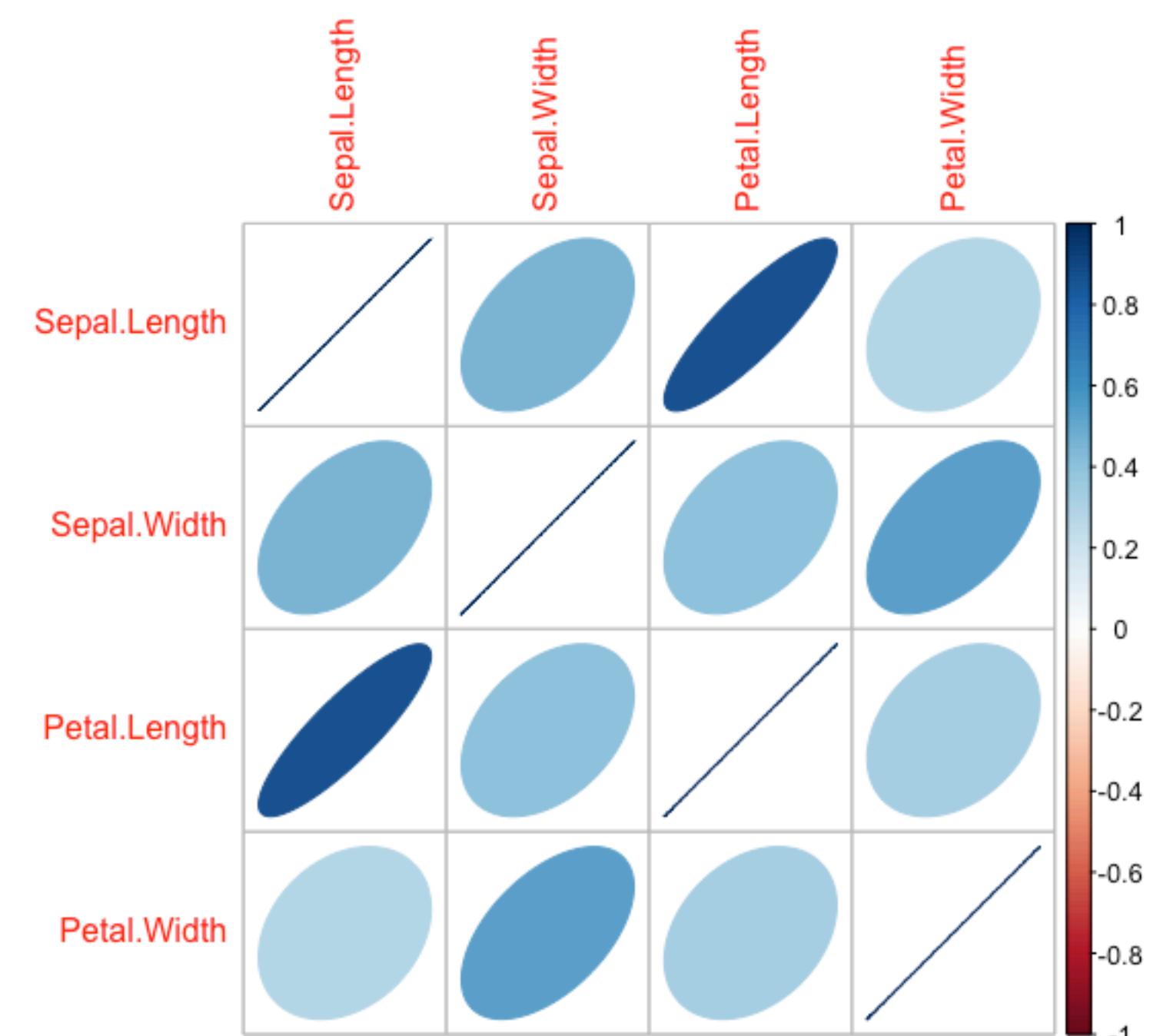
# Diagrama de Correlación



Setosa



Versicolor



Virginia

## Diagrama de Dispersión III

- Gráficas de R no son muy estéticas

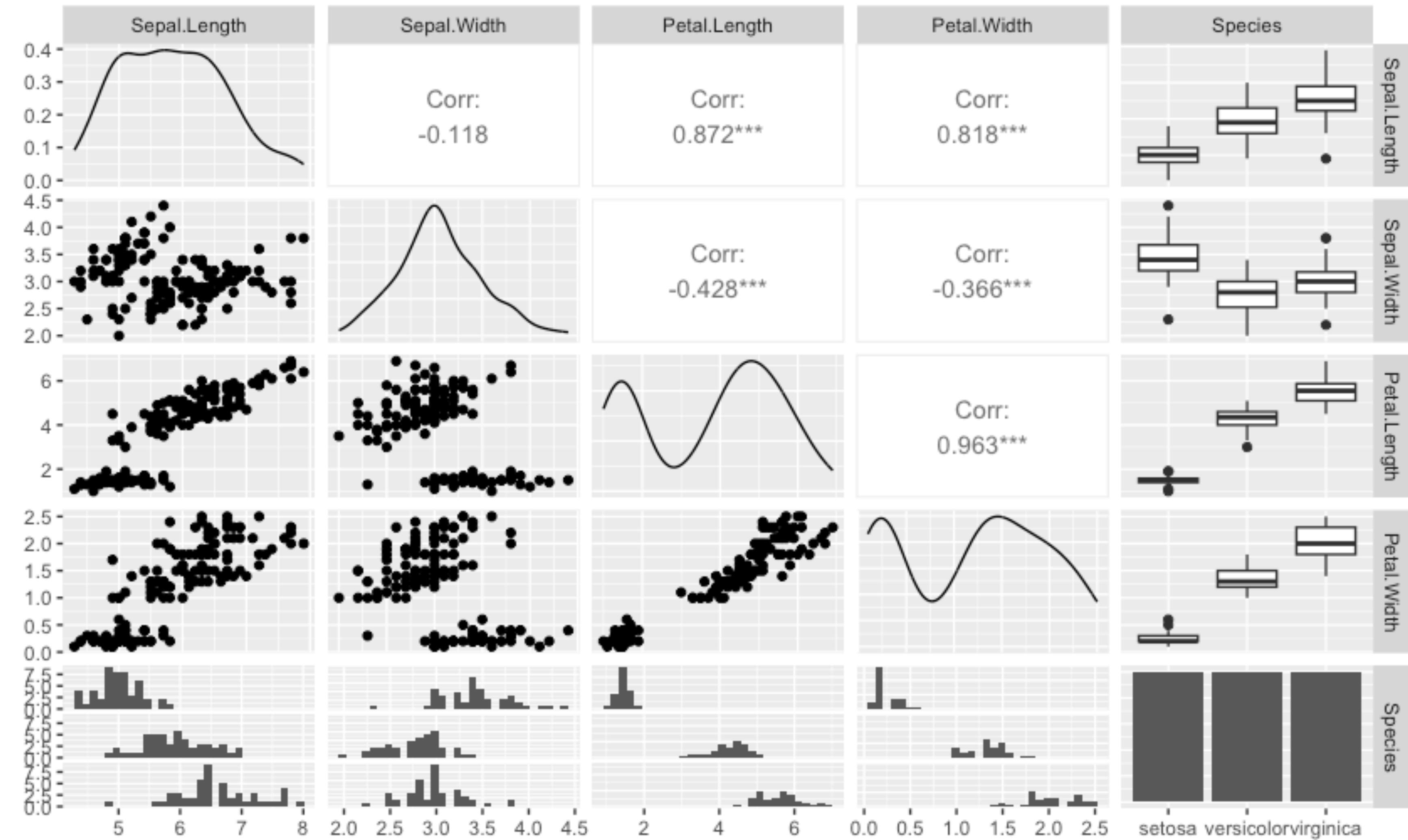
## Diagrama de Dispersión III

- Gráficas de R no son muy estéticas
- Podemos explotar las bondades de `ggplot2` para crear gráficas más estéticas e ilustrativas

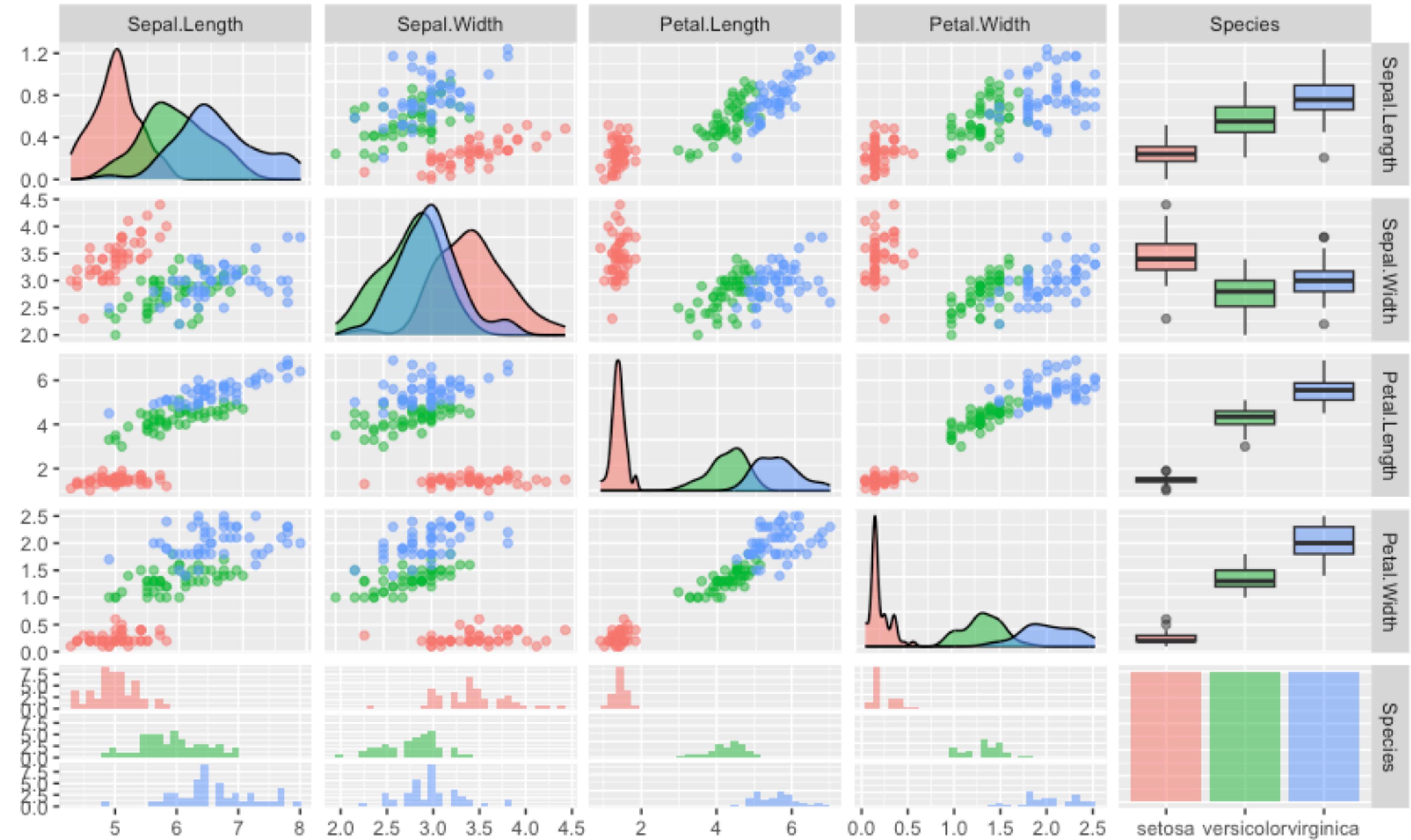
## Diagrama de Dispersión III

- Gráficas de R no son muy estéticas
- Podemos explotar las bondades de `ggplot2` para crear gráficas más estéticas e ilustrativas
- Para diagramas de dispersión y correlación:
  - ▶ Librería: `GGally`

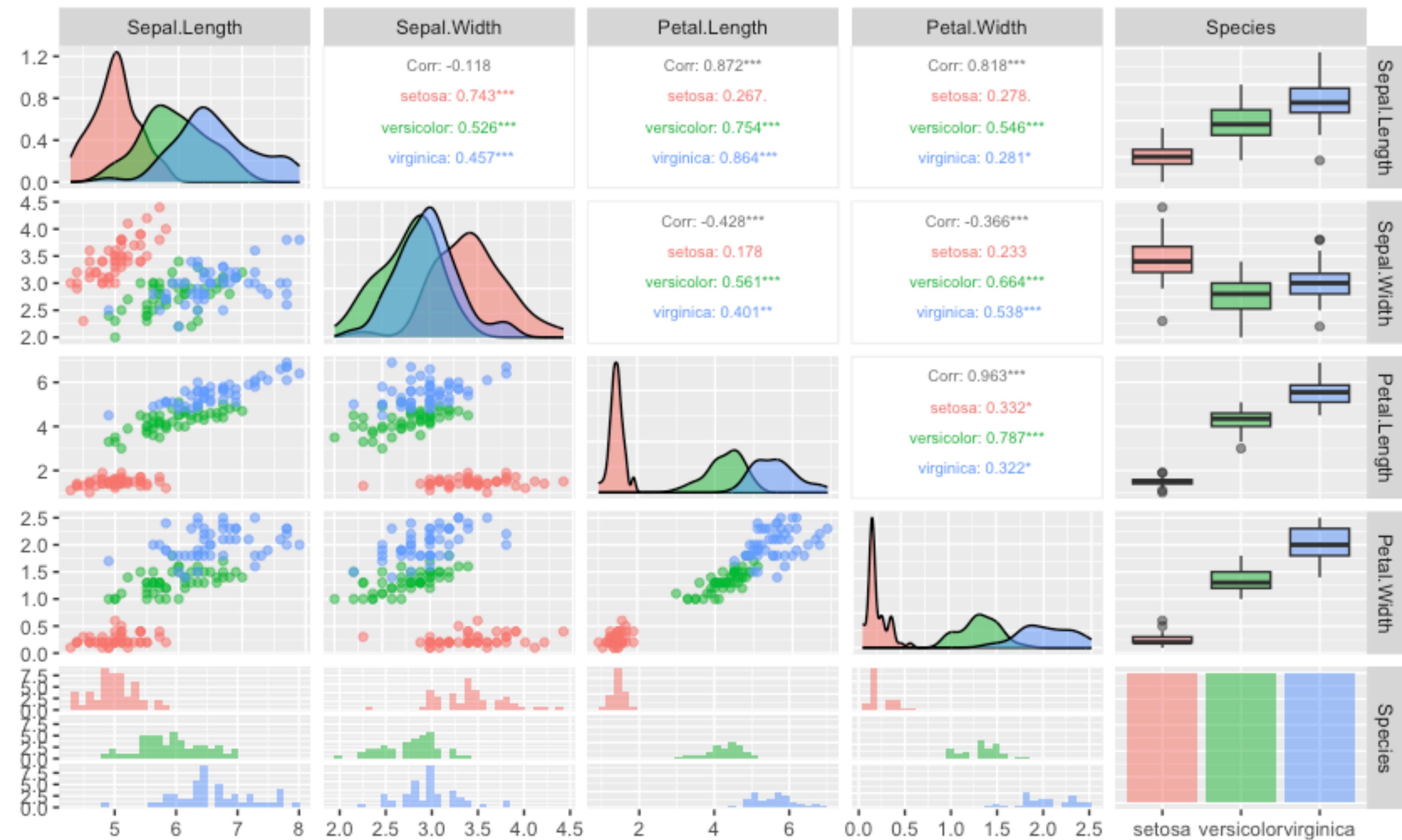
# Diagrama de Dispersion III



# Diagrama de Dispersion III



# Diagrama de Dispersion III



# Gráficas de Estrellas

# Gráficas de Estrellas

- Técnica para graficar datos multivariados en 2D (escalados a  $[0,1]$ )

# Gráficas de Estrellas

- Técnica para graficar datos multivariados en 2D (escalados a  $[0,1]$ )
- Se forma una “estrella” con  $p$  picos por cada una de las  $n$  observaciones

# Gráficas de Estrellas

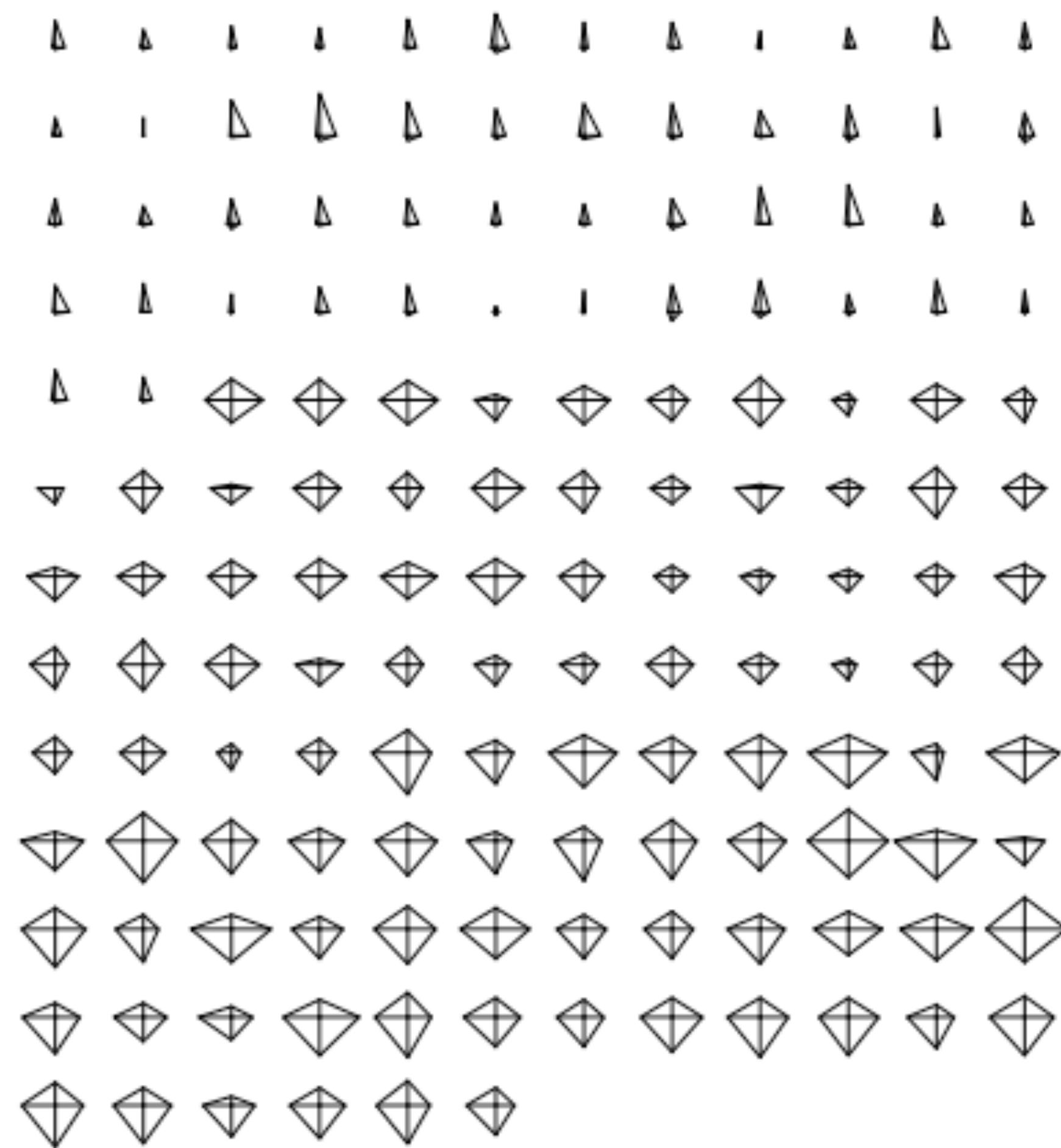
- Técnica para graficar datos multivariados en 2D (escalados a  $[0,1]$ )
- Se forma una “estrella” con  $p$  picos por cada una de las  $n$  observaciones
- Útil para:
  - Identificar clusters, outliers y variables importantes

# Gráficas de Estrellas

- Técnica para graficar datos multivariados en 2D (escalados a  $[0,1]$ )
- Se forma una “estrella” con  $p$  picos por cada una de las  $n$  observaciones
- Útil para:
  - Identificar clusters, outliers y variables “importantes”
- Desventajas:
  - Complicado de analizar si hay muchas observaciones y/o muchas variables

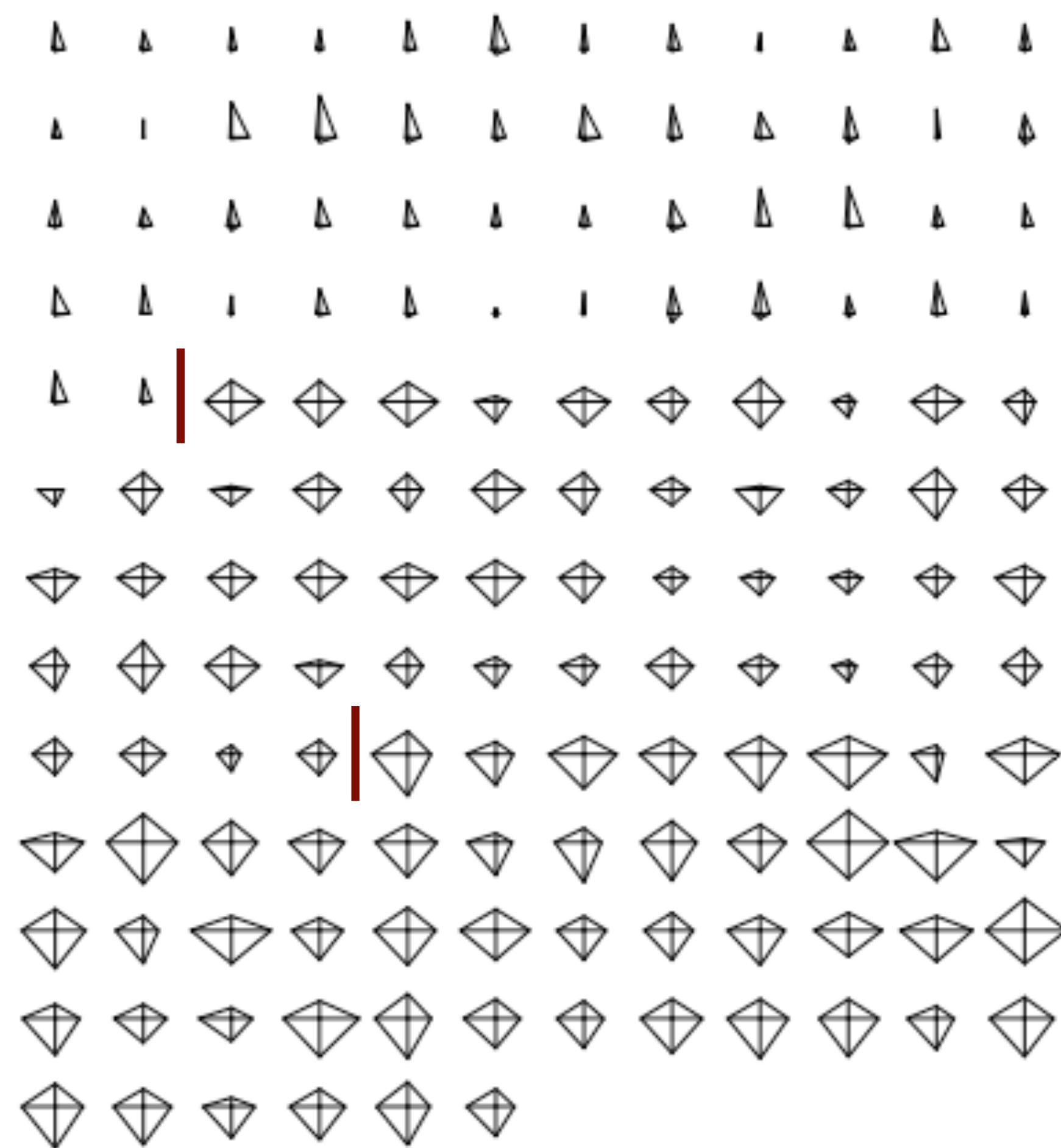
# Gráficas de Estrellas

- `stars(iris[,-5],lwd=1)`



# Gráficas de Estrellas

- `stars(iris[,-5],lwd=1)`



# Caras de Chernoff

## Caras de Chernoff

- Técnica para graficar datos multivariados (escalados a  $[0,1]$ ) similar a las estrellas pero usando caras

## Caras de Chernoff

- Técnica para graficar datos multivariados (escalados a  $[0,1]$ ) similar a las estrellas pero usando caras
- Desarrollado por Chernoff, Herman (1973). *The use of Faces to Represent Points in K-Dimensional Space Graphically*

## Caras de Chernoff

- Técnica para graficar datos multivariados (escalados a  $[0,1]$ ) similar a las estrellas pero usando caras
- Desarrollado por Chernoff, Herman (1973). *The use of Faces to Represent Points in K-Dimensional Space Graphically*
- Útil para:
  - Identificar rápidamente clusters, outliers y variables importantes

# Caras de Chernoff

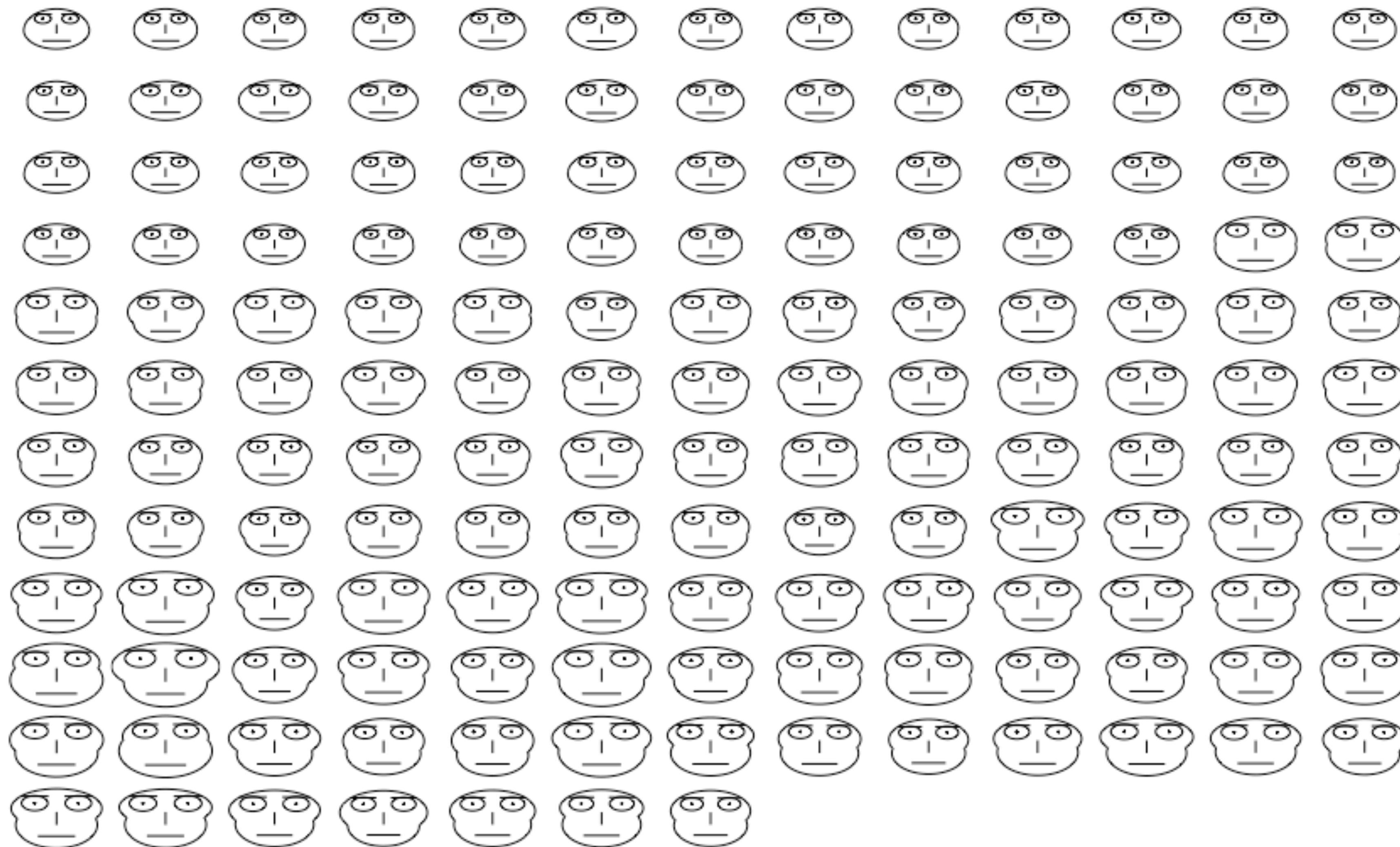
- Técnica para graficar datos multivariados (escalados a  $[0,1]$ ) similar a las estrellas pero usando caras
- Desarrollado por Chernoff, Herman (1973). *The use of Faces to Represent Points in K-Dimensional Space Graphically*
- Útil para:
  - Identificar rápidamente clusters, outliers y variables importantes
- Desventajas:
  - Limitado a  $p \leq 18$
  - El orden de las variables importa

# Caras de Chernoff

- Técnica para graficar datos multivariados (escalados a  $[0,1]$ ) similar a las estrellas pero usando caras
- Desarrollado por Chernoff, Herman (1973). *The use of Faces to Represent Points in K-Dimensional Space Graphically*
- Útil para:
  - Identificar rápidamente clusters, outliers y variables importantes
- Desventajas:
  - Limitado a  $p \leq 18$
  - El orden de las variables importa
- En R:
  - Librería TeachingDemos

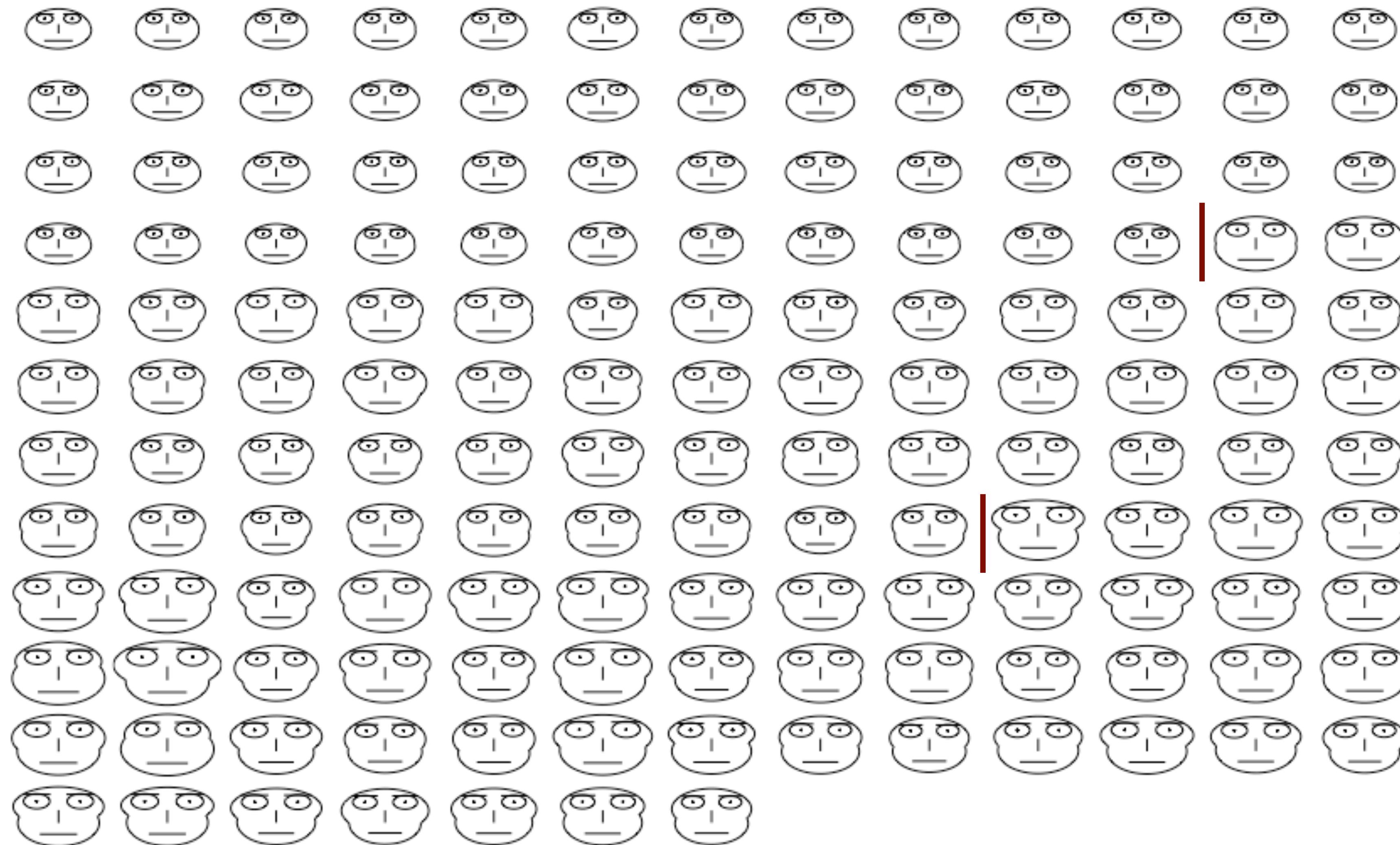
# Caras de Chernoff

- `faces2(iris[,-5])`



# Caras de Chernoff

- `faces2(iris[,-5])`



# Caras de Chernoff

- (El orden importa) `faces(iris[,c(4,3,2,1)])`



# Curvas de Andrews

# Curvas de Andrews

- Transformación ad hoc para graficar datos multivariados en el plano cartesiano o en coordenadas polares

## Curvas de Andrews

- Transformación ad hoc para graficar datos multivariados en el plano cartesiano o en coordenadas polares
- Desarrollado por Andrews, D.F. (1972). **Plots of High-Dimensional Data.**

## Curvas de Andrews

- Transformación ad hoc para graficar datos multivariados en el plano cartesiano o en coordenadas polares
- Desarrollado por Andrews, D.F. (1972). **Plots of High-Dimensional Data.**
- Cada punto  $\mathbf{x} = (x_1, \dots, x_p)$  es mapeado a

$$f_{\mathbf{x}}(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \dots, \quad -\pi < t < \pi$$

# Curvas de Andrews

- Transformación ad hoc para graficar datos multivariados en el plano cartesiano o en coordenadas polares
- Desarrollado por Andrews, D.F. (1972). **Plots of High-Dimensional Data.**
- Cada punto  $\mathbf{x} = (x_1, \dots, x_p)$  es mapeado a

$$f_{\mathbf{x}}(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \dots, \quad -\pi < t < \pi$$

- (Algunas) Propiedades:

i. Preserva medias, i.e.,

$$f_{\bar{\mathbf{x}}}(t) = \frac{1}{n} \sum_{i=1}^n f_{x_i}(t)$$

ii. Preserva distancias, i.e.,

$$\| f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t) \|_{L_2} = \int_{-\pi}^{\pi} [f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)]^2 dt = \pi \| \mathbf{x} - \mathbf{y} \|^2 \text{ (tarea)}$$

# Curvas de Andrews

- Otros posibles mapeos

$$f_{\mathbf{x}}(t) = x_1 \sin(n_1 t) + x_2 \cos(n_1 t) + x_3 \sin(n_2 t) + x_4 \cos(n_2 t) + \dots, \quad n_i \in \mathbb{N} \quad (\text{Andrews, 1972})$$

$$f_{\mathbf{x}}(t) = \frac{1}{\sqrt{2}} [x_1 + x_2(\sin(t) + \cos(t)) + x_3(\sin(t) - \cos(t)) + x_4(\sin(2t) + \cos(2t)) + \dots] \quad (\text{Khattree, R. (2002)})$$

# Curvas de Andrews

- Otros posibles mapeos

$$f_{\mathbf{x}}(t) = x_1 \sin(n_1 t) + x_2 \cos(n_1 t) + x_3 \sin(n_2 t) + x_4 \cos(n_2 t) + \dots, \quad n_i \in \mathbb{N} \quad (\text{Andrews, 1972})$$

$$f_{\mathbf{x}}(t) = \frac{1}{\sqrt{2}} [x_1 + x_2(\sin(t) + \cos(t)) + x_3(\sin(t) - \cos(t)) + x_4(\sin(2t) + \cos(2t)) + \dots] \quad (\text{Khattree, R. (2002)})$$

- En R
  - Librería `pracma` implementa la función definida por Khattree, R. & Naik, D. (2002) *Andrews plots for multivariate data: some new suggestions and applications.* ( $t \in [0, 2\pi]$ )

# Curvas de Andrews

- Ventajas

- No hay restricciones en el número de variables ni de observaciones.
- Detección de outliers y clusters
- No requiere datos escalados

# Curvas de Andrews

- Ventajas

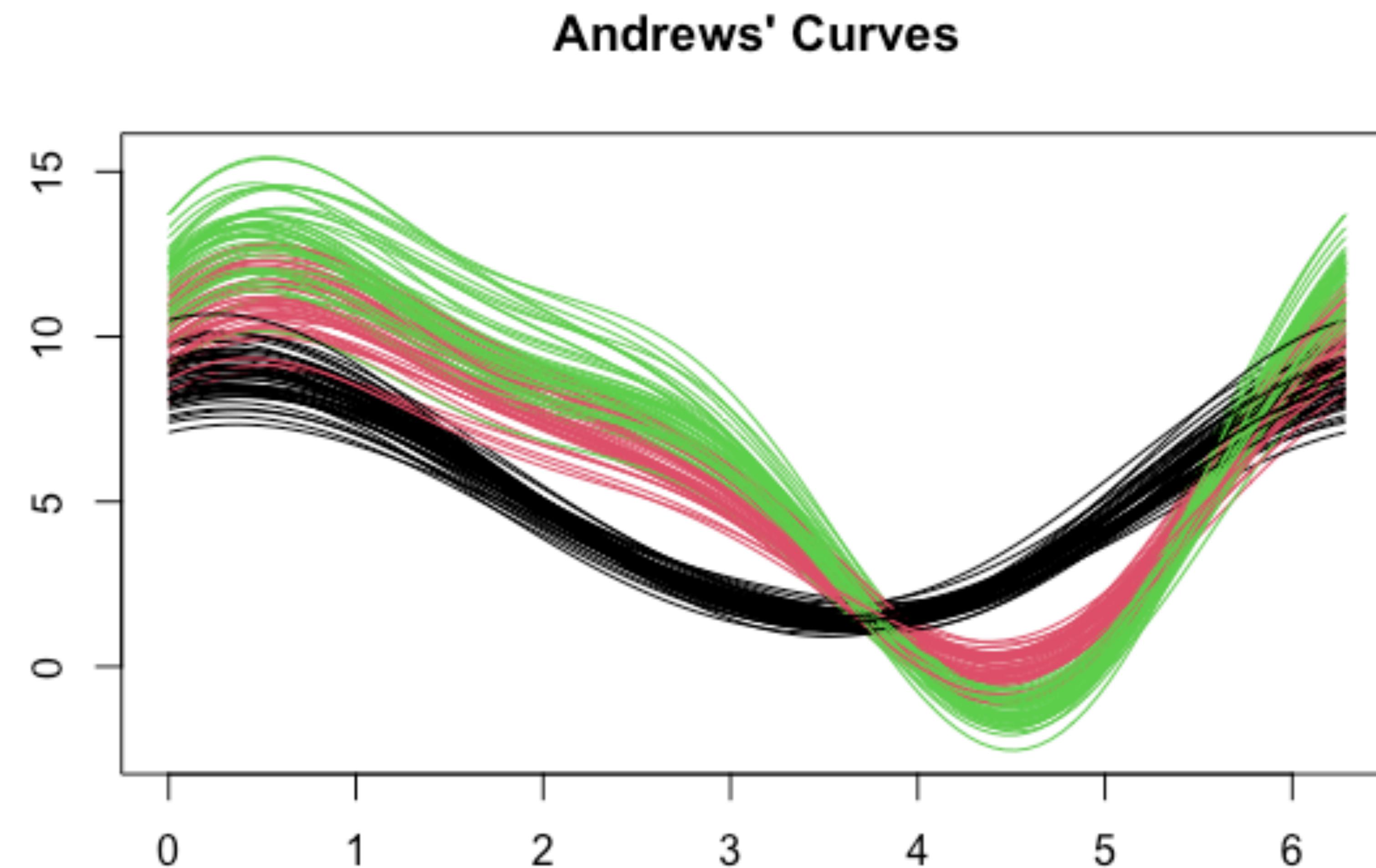
- No hay restricciones en el número de variables ni de observaciones.
- Detección de outliers y clusters
- No requiere datos escalados

- Desventajas

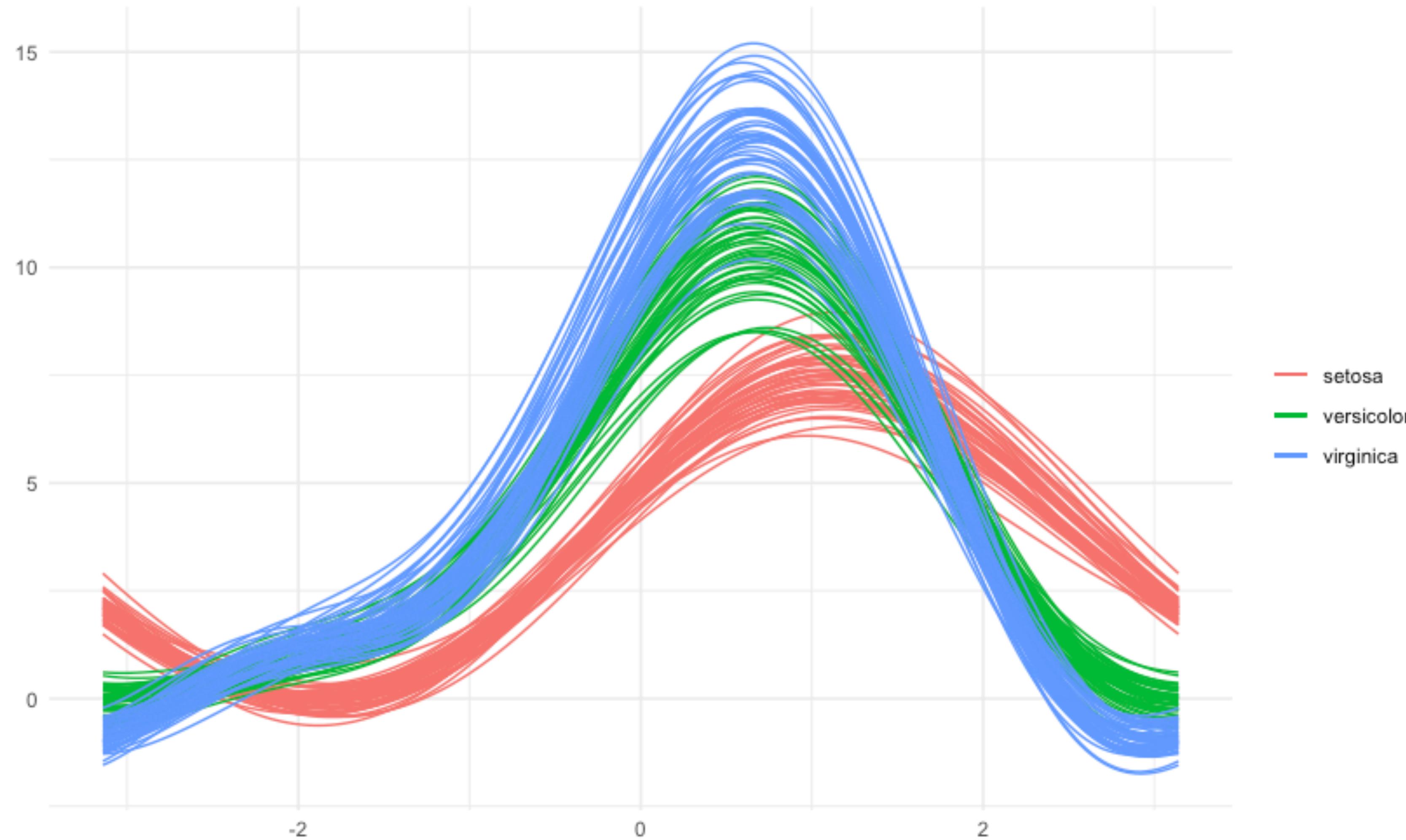
- El orden de las variables importa
- Mayor peso a las primeras variables.

## Curvas de Andrews

- `andrewsplot(as.matrix(iris[,-5]),iris[,5],style="cart")`

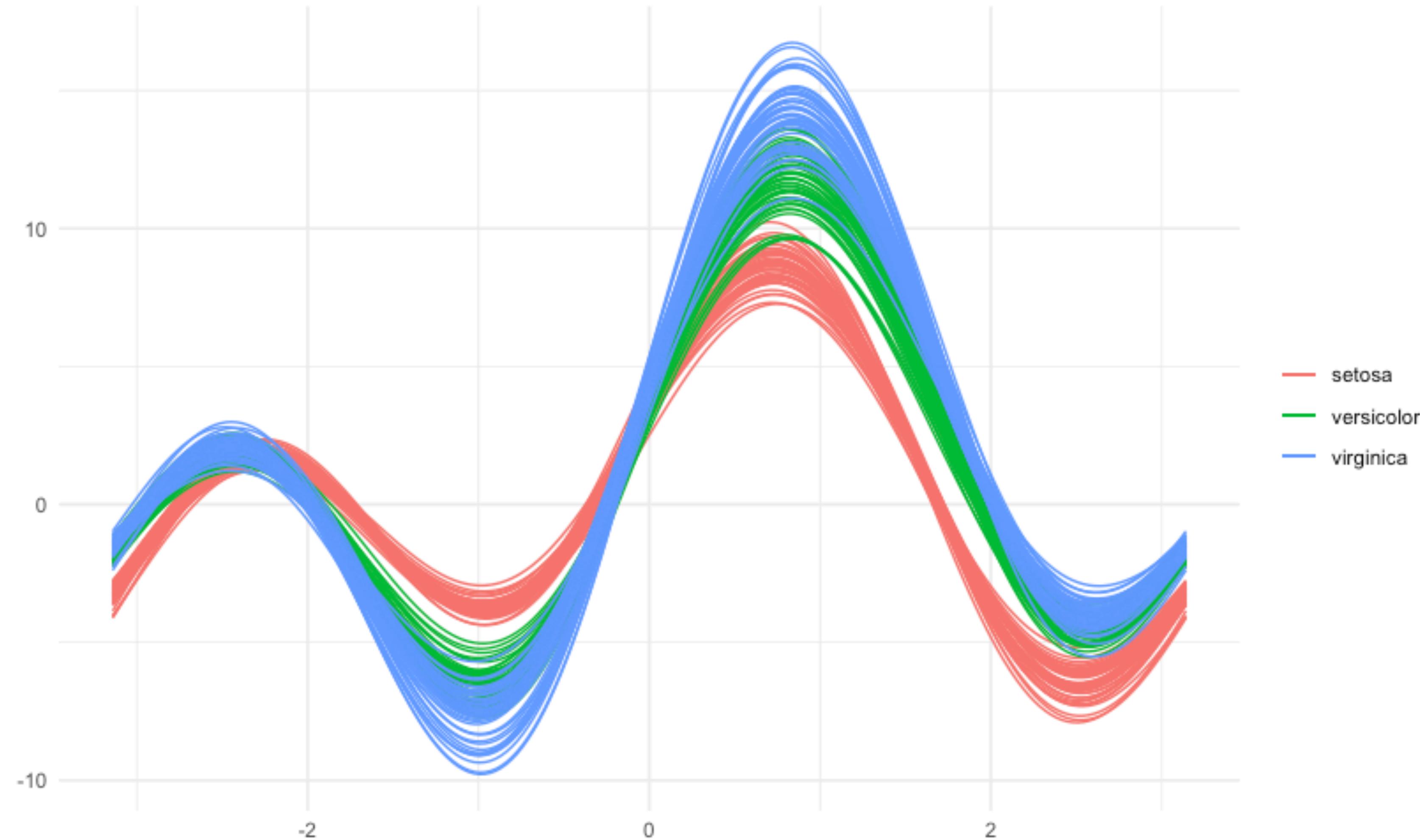


# Curvas de Andrews



# Curvas de Andrews

- El orden importa



# Estadísticas Descriptivas

## Media Muestral

- Para la matriz  $\mathbf{X}$  podemos obtener la media muestral para cada variable  $\mathbf{x}^{(j)}$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

- Para la matriz  $\mathbf{X}$  podemos obtener la media muestral para cada variable  $\mathbf{x}^{(j)}$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

- Para obtener el vector de medias

$$\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_p)$$

- Para la matriz  $\mathbf{X}$  podemos obtener la media muestral para cada variable  $\mathbf{x}^{(j)}$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

- Para obtener el vector de medias

$$\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_p)$$

- Formalmente, definimos al vector de medias como

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

- **Proposición 1**

La media muestral de una matriz de datos  $\mathbf{X}$  está dada por

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{X}^T \mathbf{1}_n,$$

donde

$$\mathbf{1}_n \equiv (1, 1, \dots, 1)^T.$$

- **Observaciones**

- $\mathbf{1}_n^T \mathbf{1}_n = n$

- $\mathbf{J}_{n \times p} = \mathbf{1}_n \mathbf{1}_p^T$

- En R:

- **summary()**

```
> summary(iris)
   Sepal.Length   Sepal.Width    Petal.Length   Petal.Width      Species
   Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa   :50
   1st Qu.:5.100  1st Qu.:2.800  1st Qu.:1.600  1st Qu.:0.300  versicolor:50
   Median :5.800  Median :3.000  Median :4.350  Median :1.300  virginica :50
   Mean   :5.843  Mean   :3.057  Mean   :3.758  Mean   :1.199
   3rd Qu.:6.400  3rd Qu.:3.300  3rd Qu.:5.100  3rd Qu.:1.800
   Max.   :7.900  Max.   :4.400  Max.   :6.900  Max.   :2.500
```

- En R:

- **summary()**

```
> summary(iris)
   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width      Species
   Min.    :4.300   Min.    :2.000   Min.    :1.000   Min.    :0.100   setosa    :50
   1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
   Median  :5.800   Median  :3.000   Median  :4.350   Median  :1.300   virginica :50
   Mean    :5.843   Mean    :3.057   Mean    :3.758   Mean    :1.199
   3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
   Max.    :7.900   Max.    :4.400   Max.    :6.900   Max.    :2.500
```

- **apply()**

```
> apply(iris[,-5],2,mean)
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
      5.843333     3.057333     3.758000     1.199333
```

- colMeans()

```
> colMeans(iris[,-5])
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
      5.843333    3.057333    3.758000    1.199333
```

- colMeans()

```
> colMeans(iris[,-5])
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
      5.843333    3.057333    3.758000    1.199333
```

- Hacer su propia función

```
> t(as.matrix(iris[,-5]))%*%rep(1,150)/150
[,1]
Sepal.Length 5.843333
Sepal.Width 3.057333
Petal.Length 3.758000
Petal.Width 1.199333
```

- Para la media muestral por grupos

- **by()**

```
> by(iris[,-5],iris[,5],colMeans)
iris[, 5]: setosa
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
      5.006        3.428       1.462        0.246
-----
iris[, 5]: versicolor
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
      5.936        2.770       4.260        1.326
-----
iris[, 5]: virginica
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
      6.588        2.974       5.552        2.026
```

# Varianza y Covarianza Muestral

- Varianza de  $\mathbf{x}^{(j)}$

$$s_j^2 = s_{jj} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

# Varianza y Covarianza Muestral

- Varianza de  $\mathbf{x}^{(j)}$

$$s_j^2 = s_{jj} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

- Covarianza entre  $\mathbf{x}^{(j)}$  y  $\mathbf{x}^{(k)}$

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

# Varianza y Covarianza Muestral

- Varianza de  $\mathbf{x}^{(j)}$

$$s_j^2 = s_{jj} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

- Covarianza entre  $\mathbf{x}^{(j)}$  y  $\mathbf{x}^{(k)}$

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

- Y así, la matriz de varianza y covarianza

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

# Varianza y Covarianza Muestral

- Formalmente definimos a  $\mathbf{S}$  como

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

# Varianza y Covarianza Muestral

- Formalmente definimos a  $\mathbf{S}$  como

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

- Considerando  $\mathbf{w}_i = \mathbf{x}_i - \bar{\mathbf{x}}$

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i^T$$

- Podemos pensar a  $\mathbf{w}_i$  como observaciones de una “nueva” matriz de datos  $\mathbf{W}$

# Varianza y Covarianza Muestral

- Observación

$$\begin{aligned}\mathbf{W} &= \mathbf{X} - \begin{pmatrix} \bar{\mathbf{x}}^T \\ \bar{\mathbf{x}}^T \\ \vdots \\ \bar{\mathbf{x}}^T \end{pmatrix} \\ &= \mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^T \\ &= \mathbf{X} - \mathbf{1}_n \left[ \frac{1}{n} \mathbf{X}^T \mathbf{1}_n \right]^T \\ &= \mathbf{X} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \mathbf{X} \\ &= \left( \mathbf{I}_{n \times n} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \mathbf{X} \\ &= \mathbf{H}_n \mathbf{X}\end{aligned}$$

# Varianza y Covarianza Muestral

## • Definición

A la matriz  $\mathbf{H}_n$  se le conoce como matriz de centrado (centring/centering matrix).

# Varianza y Covarianza Muestral

## • Definición

A la matriz  $\mathbf{H}_n$  se le conoce como matriz de centrado (centring/centering matrix).

## • Proposición (tareita)

- i.  $\mathbf{H}_n$  es simétrica, i.e.,  $\mathbf{H}_n^T = \mathbf{H}_n$
- ii.  $\mathbf{H}_n$  es idempotente, i.e.,  $\mathbf{H}_n^2 = \mathbf{H}_n$
- iii. Si  $\mathbf{X}$  es una matriz de datos de  $n$  observaciones y  $p$  variables entonces la matriz  $\mathbf{W} = \mathbf{H}_n \mathbf{X}$  tiene como media muestral al vector de ceros.
- iv. La matriz de varianza y covarianza se puede escribir como  $\mathbf{S} = \frac{1}{n-1} \mathbf{X}^T \mathbf{H}_n \mathbf{X}$

# Varianza y Covarianza Muestral

- De esta forma se tiene que

$$\mathbf{S} = \frac{1}{n - 1} \mathbf{W}^T \mathbf{W} = \mathbf{A}^T \mathbf{A}$$

# Varianza y Covarianza Muestral

- De esta forma se tiene que

$$\mathbf{S} = \frac{1}{n-1} \mathbf{W}^T \mathbf{W} = \mathbf{A}^T \mathbf{A}$$

- **Proposición (tareita).**

Sea  $\mathbf{S}$  una matriz cuadrada tal que  $\mathbf{S} = \mathbf{A}^T \mathbf{A}$ , donde  $\mathbf{A}_{n \times p}$  entonces

- i.  $\mathbf{S}$  es simétrica.
- ii.  $\mathbf{S}$  es semidefinida positiva, i.e.,  $\forall \alpha \in \mathbb{R}^p$  se cumple  $\alpha^T \mathbf{S} \alpha \geq 0$

# Varianza y Covarianza Muestral

- En R:
  - `var()` o `cov()`

```
> var(iris[,-5])
          Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.6856935 -0.0424340   1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794  -0.3296564  -0.1216394
Petal.Length   1.2743154 -0.3296564   3.1162779   1.2956094
Petal.Width    0.5162707 -0.1216394   1.2956094   0.5810063
```

# Varianza y Covarianza Muestral

- En R:

- `var()` o `cov()`

```
> var(iris[,-5])
          Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.6856935 -0.0424340   1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794  -0.3296564  -0.1216394
Petal.Length   1.2743154 -0.3296564   3.1162779   1.2956094
Petal.Width    0.5162707 -0.1216394   1.2956094   0.5810063
```

- Usar la función `sweep()` para construir la matriz **W**

```
> W=as.matrix(sweep(iris[,-5],2,colMeans(iris[,-5])))
> t(W)%*%W/(150-1)
          Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   0.6856935 -0.0424340   1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794  -0.3296564  -0.1216394
Petal.Length   1.2743154 -0.3296564   3.1162779   1.2956094
Petal.Width    0.5162707 -0.1216394   1.2956094   0.5810063
```

# Varianza y Covarianza Muestral

- Para la varianza y covarianza muestral por grupos

- **by()**

```
> by(iris[,-5],iris[,5],cov)
iris[, 5]: setosa
              Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  0.12424898 0.099216327 0.016355102 0.010330612
Sepal.Width   0.09921633 0.143689796 0.011697959 0.009297959
Petal.Length  0.01635510 0.011697959 0.030159184 0.006069388
Petal.Width   0.01033061 0.009297959 0.006069388 0.011106122
-----
iris[, 5]: versicolor
              Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  0.26643265 0.08518367 0.18289796 0.05577959
Sepal.Width   0.08518367 0.09846939 0.08265306 0.04120408
Petal.Length  0.18289796 0.08265306 0.22081633 0.07310204
Petal.Width   0.05577959 0.04120408 0.07310204 0.03910612
-----
iris[, 5]: virginica
              Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  0.40434286 0.09376327 0.30328980 0.04909388
Sepal.Width   0.09376327 0.10400408 0.07137959 0.04762857
Petal.Length  0.30328980 0.07137959 0.30458776 0.04882449
Petal.Width   0.04909388 0.04762857 0.04882449 0.07543265
```

## Correlación Muestral

- Finalmente, la correlación entre  $\mathbf{x}^{(j)}$  y  $\mathbf{x}^{(k)}$

$$r_{jk} = \frac{s_{jk}}{s_j s_k}, \quad s_j = \sqrt{s_{jj}}$$

# Correlación Muestral

- Finalmente, la correlación entre  $\mathbf{x}^{(j)}$  y  $\mathbf{x}^{(k)}$

$$r_{jk} = \frac{s_{jk}}{s_j s_k}, \quad s_j = \sqrt{s_{jj}}$$

- Usando  $\mathbf{w}^{(j)} = \mathbf{x}^{(j)} - \bar{x}_j \mathbf{1}_n$

$$r_{jk} = \frac{\mathbf{w}^{(j)} \cdot \mathbf{w}^{(k)}}{\|\mathbf{w}^{(j)}\| \|\mathbf{w}^{(k)}\|} \in [-1, 1] \quad (\text{Cauchy-Schwarz})$$

- Finalmente, la correlación entre  $\mathbf{x}^{(j)}$  y  $\mathbf{x}^{(k)}$

$$r_{jk} = \frac{s_{jk}}{s_j s_k}, \quad s_j = \sqrt{s_{jj}}$$

- Usando  $\mathbf{w}^{(j)} = \mathbf{x}^{(j)} - \bar{x}_j \mathbf{1}_n$

$$r_{jk} = \frac{\mathbf{w}^{(j)} \cdot \mathbf{w}^{(k)}}{\|\mathbf{w}^{(j)}\| \|\mathbf{w}^{(k)}\|} \in [-1,1] \quad (\text{Cauchy-Schwarz})$$

- La matriz de correlación dada por

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

# Correlación Muestral

- Otra representación útil

$$\mathbf{R} = \mathbf{D}^{-1} \mathbf{S} \mathbf{D}^{-1}$$

$$\mathbf{D} = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_p \end{pmatrix}$$

- Otra representación útil

$$\mathbf{R} = \mathbf{D}^{-1} \mathbf{S} \mathbf{D}^{-1}$$

$$\mathbf{D} = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_p \end{pmatrix}$$

- **Proposición (tarea).**

Sea  $\mathbf{R}$  la matriz de correlación muestral entonces

- i.  $\mathbf{R}$  es simétrica.
- ii.  $\mathbf{R}$  es semidefinida positiva.

# Correlación Muestral

- En R:

- `cor()`

```
> cor(iris[,-5])
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length   1.0000000 -0.1175698  0.8717538  0.8179411
Sepal.Width    -0.1175698  1.0000000 -0.4284401 -0.3661259
Petal.Length   0.8717538 -0.4284401  1.0000000  0.9628654
Petal.Width    0.8179411 -0.3661259  0.9628654  1.0000000
```

# Correlación Muestral

- Para la correlación muestral por grupos

- **by()**

```
> by(iris[,-5],iris[,5],cor)
iris[, 5]: setosa
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  1.0000000   0.7425467  0.2671758  0.2780984
Sepal.Width   0.7425467  1.0000000  0.1777000  0.2327520
Petal.Length  0.2671758  0.1777000  1.0000000  0.3316300
Petal.Width   0.2780984  0.2327520  0.3316300  1.0000000
-----
iris[, 5]: versicolor
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  1.0000000   0.5259107  0.7540490  0.5464611
Sepal.Width   0.5259107  1.0000000  0.5605221  0.6639987
Petal.Length  0.7540490  0.5605221  1.0000000  0.7866681
Petal.Width   0.5464611  0.6639987  0.7866681  1.0000000
-----
iris[, 5]: virginica
      Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length  1.0000000   0.4572278  0.8642247  0.2811077
Sepal.Width   0.4572278  1.0000000  0.4010446  0.5377280
Petal.Length  0.8642247  0.4010446  1.0000000  0.3221082
Petal.Width   0.2811077  0.5377280  0.3221082  1.0000000
```