Ejercicio 2

A) Cuantifique (en forma manual) la similaridad entre las variedades correspondientes a la primera y segunda fila en función del porcentaje de caracteres comunes respecto al número de caracteres totales. Idem las variedades asociadas a las filas 12 y 13 (incluir la variable TIPO)

Obtenemos dos vectores que representan a cada una de las dos primeras variedades y luego calculamos el la proporcion de variables donde ambas variedades coinciden.

```
variedad1 <- datos[1, ] %>% unlist() %>% unname()
variedad2 <- datos[2, ] %>% unlist() %>% unname()
similaridad_1_2 <- mean(variedad1 == variedad2, na.rm = TRUE) * 100
```

La similaridad entre las variedades de la fila 1 y 2 es del 88.89%. Si tenemos en cuenta que se tienen 9 variables, podemos notar que estas dos variedades coinciden en todas excepto 1. Luego, de manera analoga para el par de variedades 12 y 13

```
variedad12 <- datos[12, ] %>% unlist() %>% unname()
variedad13 <- datos[13, ] %>% unlist() %>% unname()
similaridad_12_13 <- mean(variedad12 == variedad13, na.rm = TRUE) * 100
```

La similaridad entre las variedades de la fila 12 y 13 es del 71.43%, lo que significa que difieren mas que el par de variedades 1 y 2.

B) Halle una matriz de similaridad entre variedades en función del coeficiente SM generalizado.

Para esta tarea utilizamos la funcion `daisy()` del paquete `cluster`, donde especificamos que la metrica a utilizar es `"gower"`, y nos devuelve la matriz de distancia entre las diferentes variedades. En la siguiente linea, convertimos la matriz de distancia a matriz de similaridad.

```
matriz_distancia <- daisy(datos_fct, metric = "gower")
matriz_similaridad <- 1 - matriz_distancia
```

Comprobemos, por ejemplo, si la similaridad entre las variedades de pepino de la primera y segunda fila computada mediante `daisy()` es igual a la que computamos a mano.

```
matriz_similaridad[1] == (similaridad_1_2 / 100)
```

```
## [1] TRUE
```

Por lo que ambos resultados son iguales.

C) Aplique Análisis de Coordenadas principales para representar en un espacio bidimensional la semejanza entre las variedades.

Aplicamos el analisis de coordenadas principales de la siguiente manera:

```
coordenadas_principales <- cmdscale(sqrt(matriz_distancia), k = 2, eig = TRUE)
```

Y graficamos a la caracterizacion cualitativa de las variedades de pepino en el plano principal.

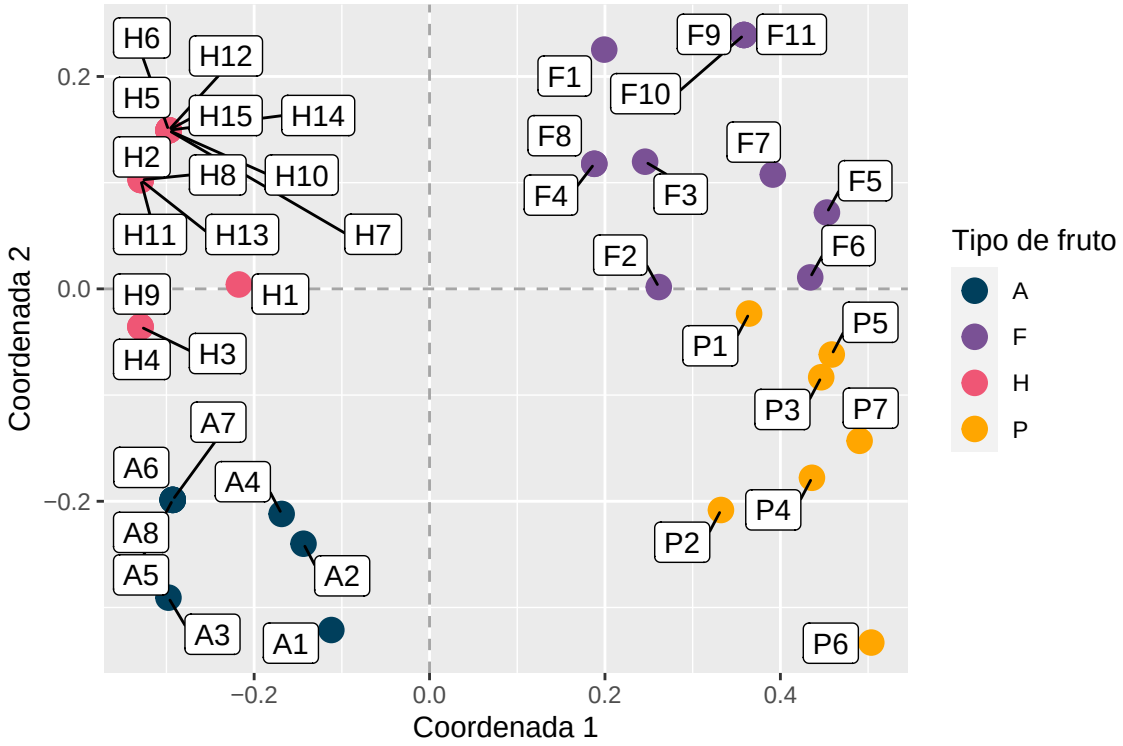

Figure 1: Caracterizacion cualitativa de las variedades de pepino en el plano principal.

En aquellos casos que mas de una etiqueta apunta hacia el mismo punto sucede que las variedades coinciden en terminos de las variables analizadas y en consecuencia los puntos que las representan estan encimados. Por ejemplo, las variedades `A5`, `A6`, `A7`, `A8` tienen idénticos valores para todas las variables.

	TIPO	SEXO	CORN...	TORN...	PEDUNC	VERR...	MOTEADO	CLADOSP	CMV
A5	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A6	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A7	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A8	Alf	Fem	Negra	Espin	Cue	No	No	No	No

D) Conforme grupos de variedades según su homogeneidad en la caracterización agronómica cualitativa.

En la Figura 1 se puede ver que las variedades de los tipos de fruto **A** y **H** se se agrupan de manera que respetan al tipo de fruto y se diferencian del resto. Tambien existe una agrupacion, ya no tan clara, para las variedades de los grupos **F** y **P**. En este caso, si no estuviera el color que diferencie a los tipos de frutos, no podriamos diferenciar a estos dos grupos claramente. Por ejemplo, las variedades **F2**, **F6** y **P1** estan muy cercanas en el plano y las podriamos haber tomado como parte de un mismo grupo.

Tambien podemos ver que la variabilidad de las variedades dentro de cada tipo de fruto difiere. Por ejemplo, para el tipo de fruto **H**, se tiene que casi todas las variedades se corresponden con dos categorizaciones particulares (por eso vemos tantos puntos encimados). Por otro lado, todas las variedades del tipo de fruto **P** se corresponden con una configuracion única de las variables cualitativas.

E) Encuentre el dendograma ultramétrico con ligamiento UPGMA correspondiente

Obtenemos el dendograma utilizando la funcion `hclust()`, a la que le pasamos la matriz de distancia previamente obtenida.

```
cluster_cualitativas <- hclust(matriz_distancia, method = "average")
```

Luego graficamos el dendograma y mostramos con diferentes colores a los clusters que se obtienen al especificar `k = 4`.

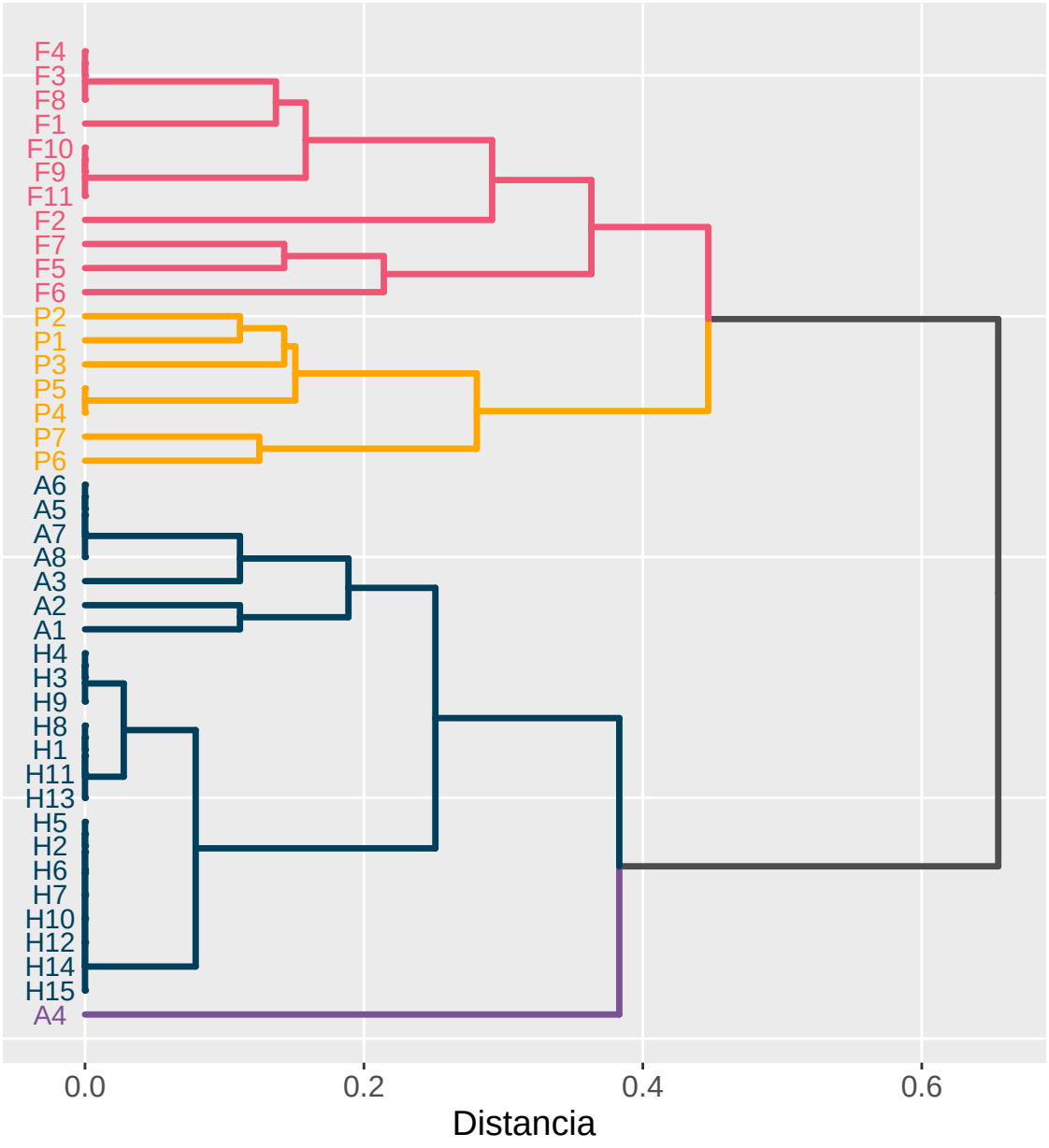

Figure 2: Dendograma Ultrametrico con ligamiento UPGMA.

En la Figura 2 se puede ver que el punto de corte `k = 4` no me permite separar a los tipos de pepinos de manera perfecta. El agrupamiento hace un buen trabajo al diferenciar a los pepinos de los tipos **F** y **P**, pero mezcla a las variedades de los tipos **A** y **H**. Estos dos ultimos tipos, son a su vez, los que mayor similaridad entre variedades presentan. Lo podemos ver en la cantidad de uniones que se presentan a una distancia de 0.

Si comparamos la Figura 1 y la Figura 2, puede sorprendernos que la variedad **A4** este tan distante

del resto de las variedades en la Figura 2, ya que se presenta cercano al resto de las variedades de tipo **A** en la Figura 1. Sin embargo, no debemos pasar por alto que la Figura 1 es una proyeccion de posicionamientos en un espacio de mayor dimensionalidad, pudiendo estos puntos estar distantes en ese espacio original.

Tomemos al tipo de fruto **A** y miremos, por ejemplo, a las variedades **A4** y **A2**, que parecen cercanos en la Figura 1 pero estan distantes en la Figura 2, y observemos sus datos crudos.

	TIPO	SEXO	CORN...	TORN...	PEDUNC	VERR...	MOTEADO	CLADOSP	CMV
A1	Alf	Fem	Blanca	Espin	Obt	No	No	No	No
A2	Alf	Fem	Negra	Espin	Obt	No	No	No	No
A3	Alf	Fem	Negra	Pelos	Cue	No	No	No	No
A4	Alf	Fem	Negra	Pelos	Agu	No	No	NA	NA
A5	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A6	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A7	Alf	Fem	Negra	Espin	Cue	No	No	No	No
A8	Alf	Fem	Negra	Espin	Cue	No	No	No	No

En la tabla podemos ver que la variedad **A4** es la unica que presenta extremo pedunculo agudo y es una de las dos unicas que presenta pelos como tipo de ornamentación, lo que alcanza para diferenciarla de las otras variedades, que son mucho mas similares entre si.

F) Mida a través de su matriz cofenética la concordancia con la matriz de distancias que le dio origen

Primero vamos a calcular la concordancia y luego obtenemos un grafico de dispersion donde se muestra la distancia original y la distancia cofenética.

```
distancia_cofenetica <- cophenetic(cluster_cualitativas)
concordancia <- cor(distancia_cofenetica, matriz_distancia)
```

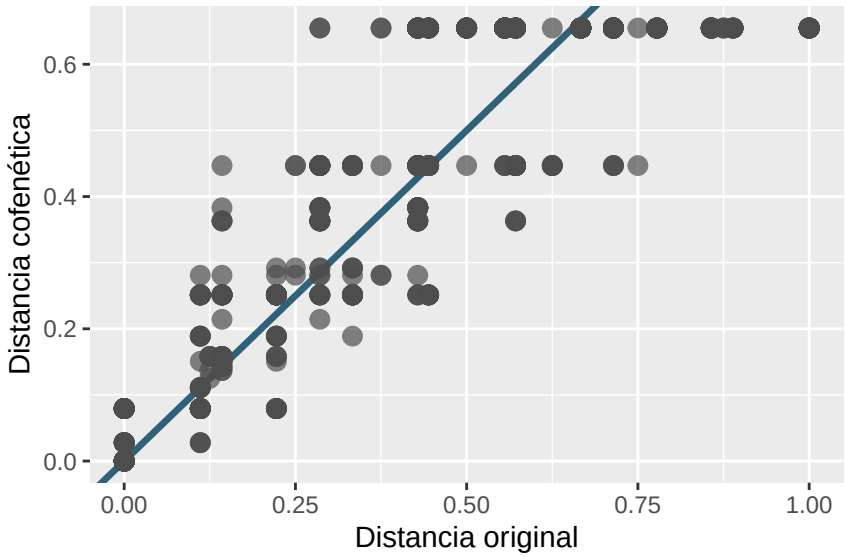

Figure 3: Grafico de dispersion entre distancia original y distancia cofenética a partir de cluster con ligamiento UPGMA. La linea azul representa a la recta identidad.

La concordancia entre la matriz de distancias cofenética y la matriz de distancia original es igual a 0.883, lo que indica una concordancia muy alta entre ambas representaciones.

En la Figura 3 se puede ver que la discrepancia entre estas distancias crece para valores mas altos de la distancia original.

G) Cuantifique concordancia entre plano principal de ACoordP y Cluster

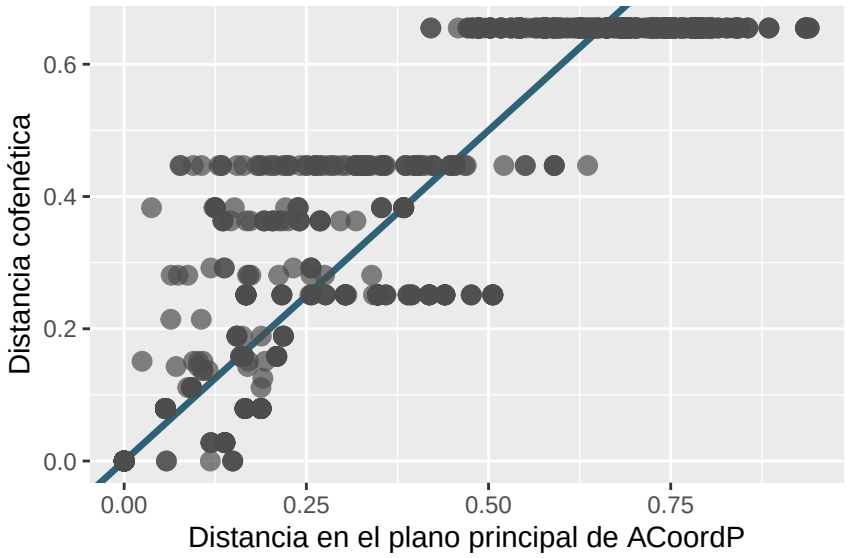

Figure 4: Grafico de dispersion entre distancia original y distancia cofenética a partir de cluster con ligamiento UPGMA. La linea azul representa a la recta identidad.

La concordancia concordancia entre plano principal de ACoordP y Cluster es 0.891, lo que nuevamente nos indica una concordancia muy alta entre ambas representaciones.