

**UNIVERSIDAD COMPLUTENSE DE MADRID**

**MÁSTER EN  
BIG DATA, DATA SCIENCE & ARTIFICIAL INTELLIGENCE**



**Minería de datos y modelización predictiva**

---

**TAREA DE EVALUACIÓN:**

**ACP + Clustering aplicado a pingüinos**

**Trabajo realizado por  
Diego López Escobar**

**Profesor  
PABLO ARCADIO FLORES  
MADRID – 11/01/2024**

## Análisis de los datos:

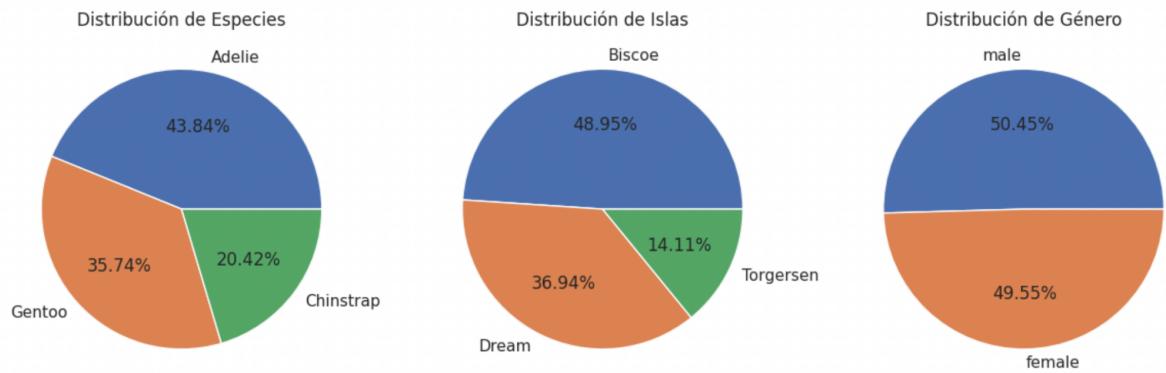
Descripción de los datos del dataset.

Nombre	Descripción
species	Es la especie de pingüino. Hay tres especies en el conjunto de datos: 'Adelie', 'Chinstrap' y 'Gentoo'.
island	Representa la isla donde se recopilaron los datos. Las islas son 'Biscoe', 'Dream' y 'Torgersen'.
bill length mm	Longitud del pico en milímetros.
bill depth mm	Profundidad del pico en milímetros.
flipper length mm	Longitud de la aleta en milímetros.
body mass g	Masa corporal del pingüino en gramos.
sex	Género del pingüino, con las categorías 'Male' (macho), 'Female' (hembra) o 'NaN' si la información no está disponible.

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	Adelie	Torgersen	39.1	18.7	181.0	3750.0	male	2007
1	Adelie	Torgersen	39.5	17.4	186.0	3800.0	female	2007
2	Adelie	Torgersen	40.3	18.0	195.0	3250.0	female	2007
3	Adelie	Torgersen	36.7	19.3	193.0	3450.0	female	2007
4	Adelie	Torgersen	39.3	20.6	190.0	3650.0	male	2007
...	...	...	...	...	...	...	...	...
328	Chinstrap	Dream	55.8	19.8	207.0	4000.0	male	2009
329	Chinstrap	Dream	43.5	18.1	202.0	3400.0	female	2009
330	Chinstrap	Dream	49.6	18.2	193.0	3775.0	male	2009
331	Chinstrap	Dream	50.8	19.0	210.0	4100.0	male	2009
332	Chinstrap	Dream	50.2	18.7	198.0	3775.0	female	2009

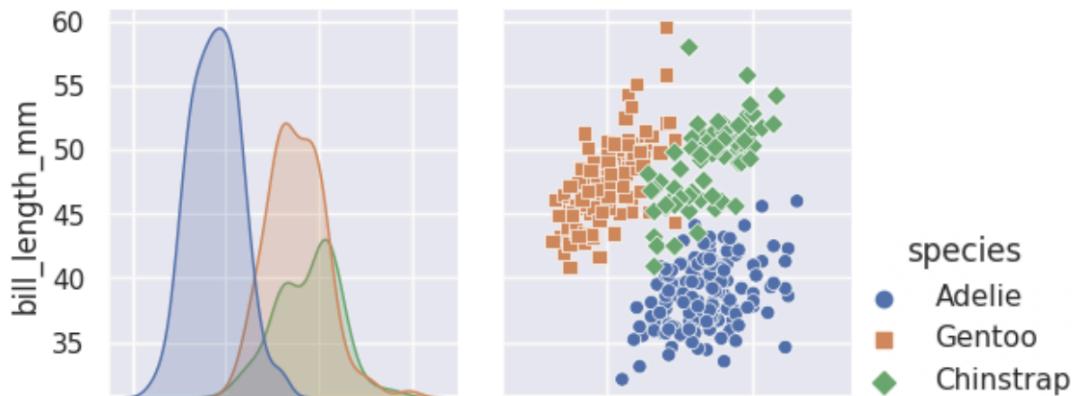
La distribución inicial de los datos es la siguiente:

#### Análisis de Distribuciones

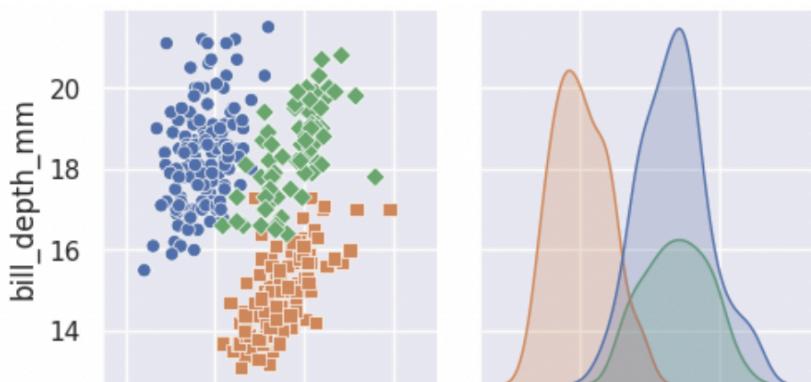


En donde se puede apreciar algunos datos superficiales como que dentro de las especies la mayor muestra representativa es la de Adelie, en la distribución de Islas la mayor representación la tiene Biscoe y en género se puede decir que prácticamente es mitad y mitad.

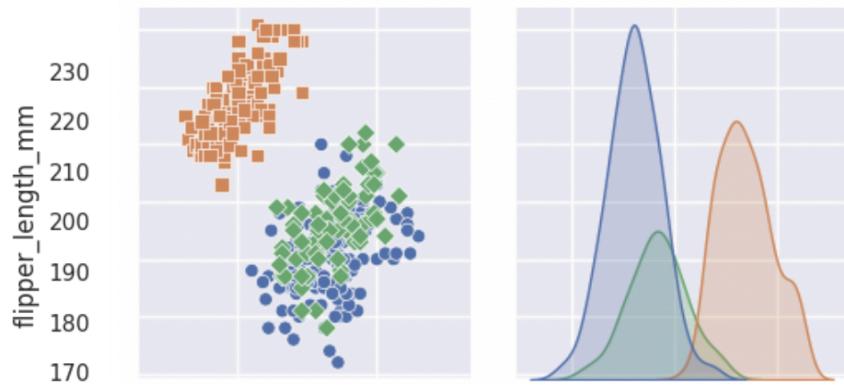
Y algunos de los gráficos exploratorios de los datos:



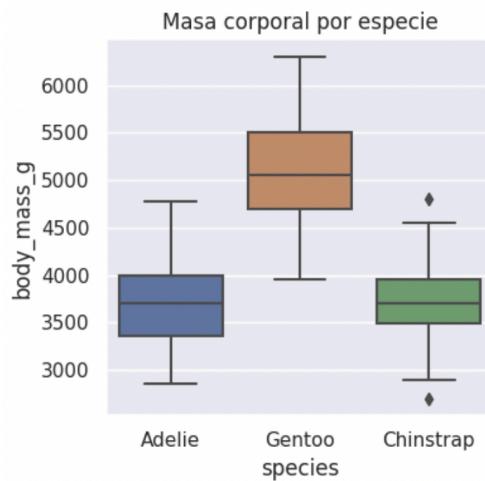
De la longitud del pico de los pingüinos podemos observar que la especie Adelie es la ganadora.



De la profundidad del pico de los pingüinos podemos observar que la especie Adelie también es la ganadora.



De la longitud de la aleta de los pingüinos podemos observar que la especie Adelie es la ganadora.

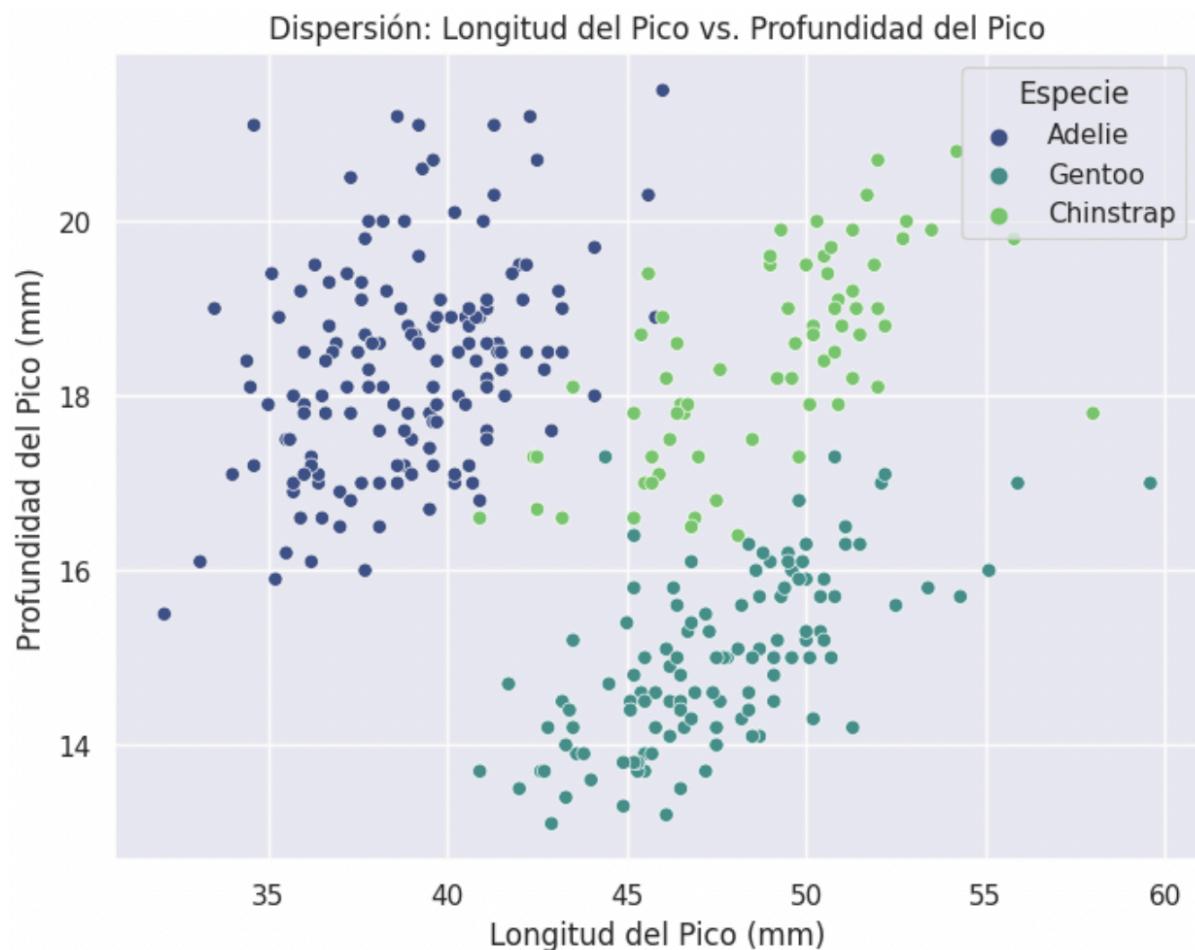


Sin embargo podemos observar que en cuanto a la masa corporal la especie de Gentoo es la ganadora.

Por lo tanto podemos decir a primera vista que la especie Adelie físicamente parece ser la más alargada, sin embargo la especie Gentoo es la más pesada.

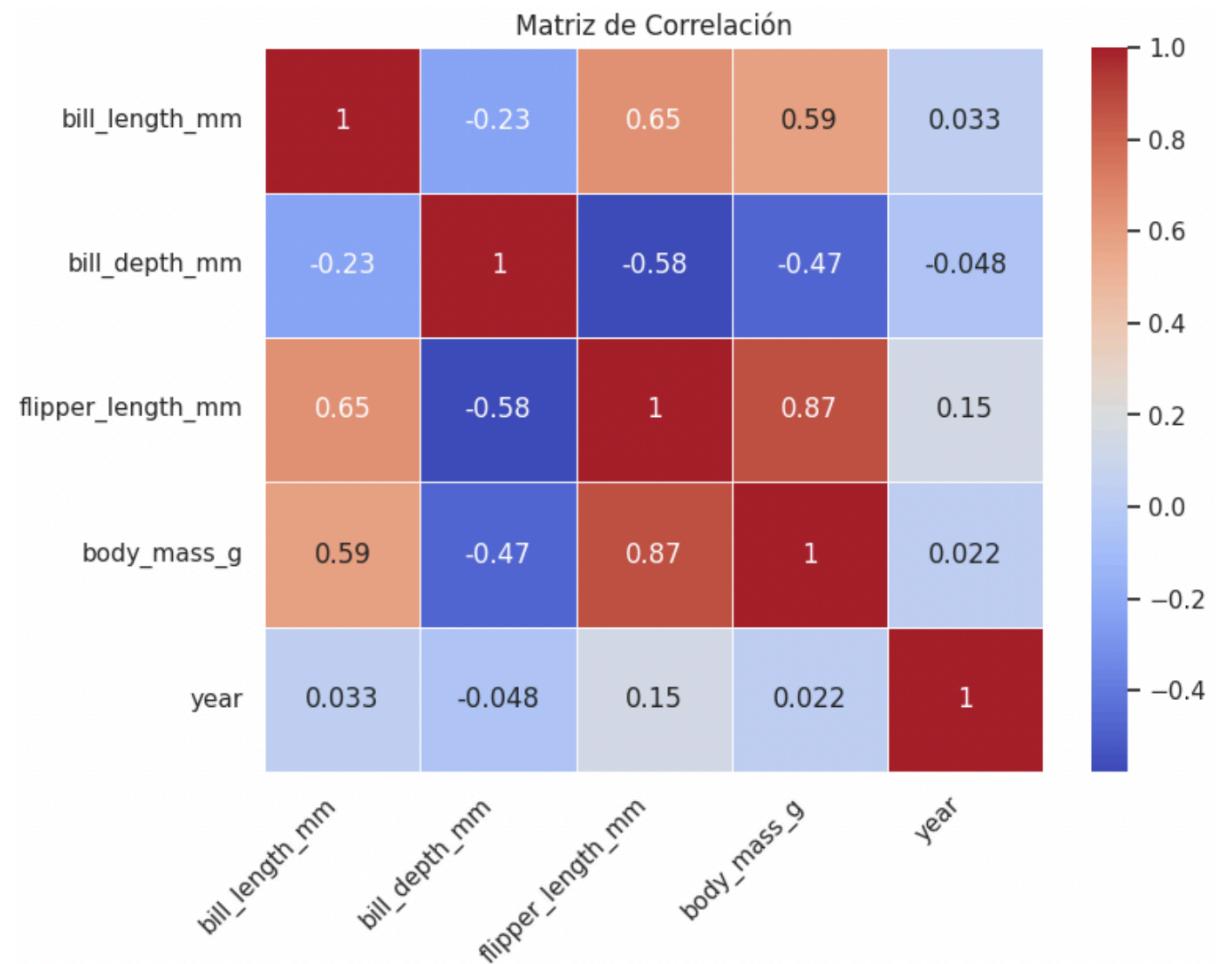
## Gráfico de dispersión (Scatter Plot)

Se puede observar que existen 3 nubes de datos que justamente coinciden con las 3 especies a las que estamos analizando, la diferenciación de los colores de los puntos ayuda a ver gráficamente los grupos, con esto podemos inferir que cada especie contiene características similares, al menos de las características que se están analizando ahora mismo.



### Matriz de correlación:

Observemos que las variables con mayor correlación positiva son flipper length y body mass, bill length y flipper length y también bill length y body mass. Por otro lado, se observa una destacable correlación negativa entre bill depth y flipper length y también entre bill depth y body mass.

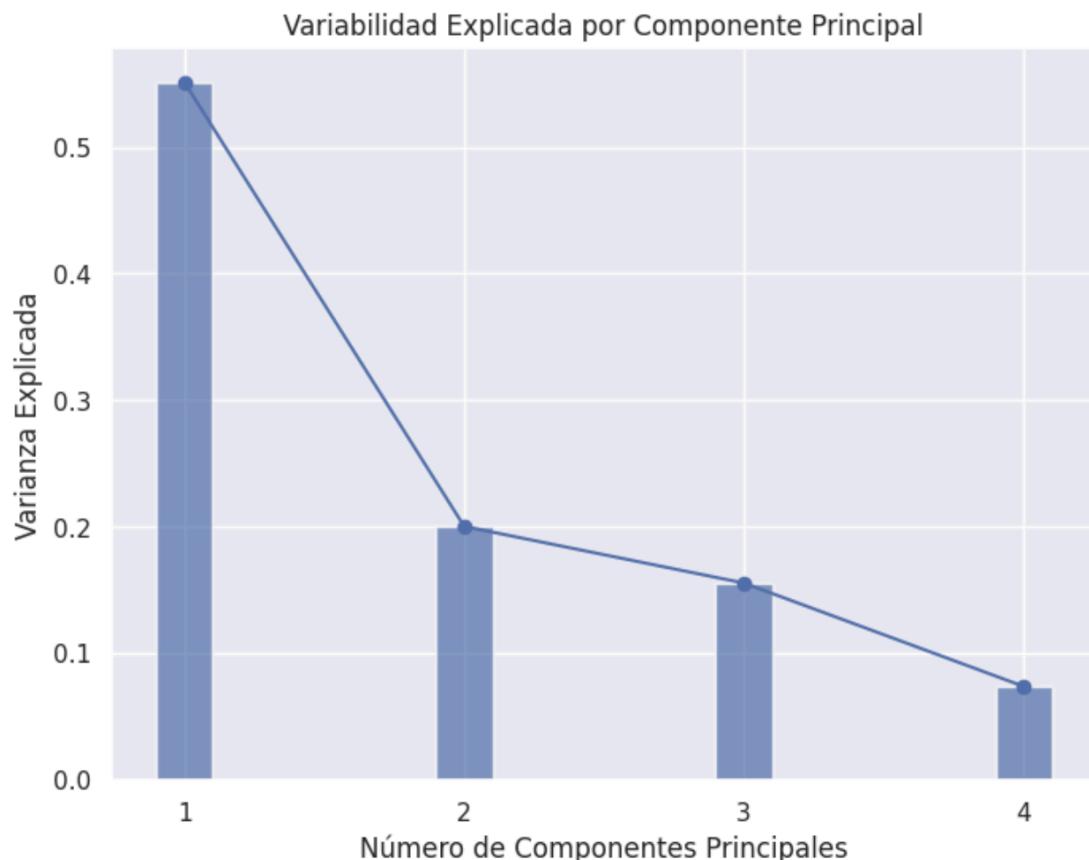


## Análisis de Componentes Principales (PCA):

¿Cuál es el número adecuado de componentes para representar eficientemente la variabilidad de las especies de pingüinos?

Luego de llevar a cabo el Análisis de Componentes Principales (PCA) en donde se decidió utilizar un número de 4 componentes para realizar el estudio, podemos deducir que la respuesta a la pregunta puede variar, en este caso consideramos que la respuesta sería que con el uso de **2 componentes principales** de los 4 sería suficiente, pues si utilizamos las primeras 2 viendo los datos de la variabilidad explicada, en la componente 1 tenemos un 55,11% y en el componente 2 tenemos un 20,02%, lo cual sumado nos da como resultado 75,13% de la variabilidad. Dependiendo del caso y de la exactitud que necesitemos podemos subir o bajar el porcentaje, de momento considero que tener arriba de 70% es más que suficiente.

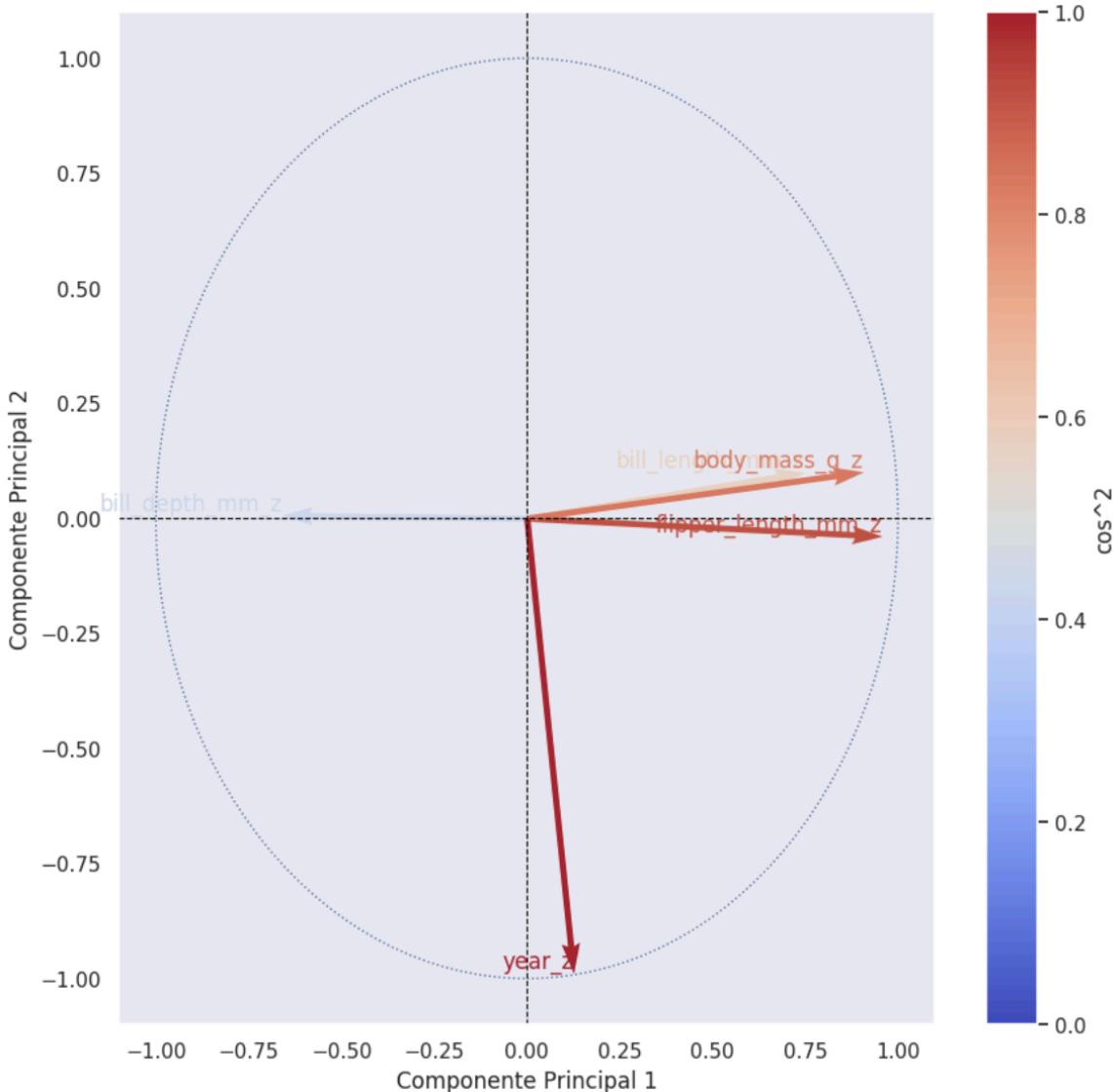
	Autovalores	Variabilidad Explicada	Variabilidad Acumulada
Componente 1	2.763782	0.551096	0.551096
Componente 2	1.003772	0.200152	0.751248
Componente 3	0.779300	0.155392	0.906640
Componente 4	0.369676	0.073713	0.980353



Además podemos observar que la variabilidad explicada por cada componente, decrece de una manera menos violenta a partir de la segunda. Recordando que los dos autovalores correspondientes a estas componentes seleccionadas son 2.76 y 1.00, representando el 55,11% y el 20,02% de la variabilidad, respectivamente.

PCA indicando el número de componentes principales (2):

Vale, pues de momento lo que hemos realizado hasta ahora es que hemos reducido la dimensionalidad de nuestra base de datos original, sustituyendo un total de 5 variables por tan solo, dos componentes, las cuales, explican aproximadamente un 75% de la variabilidad original. Pero para responder a la pregunta de explicar lo que representa cada componente en términos de las características físicas de los pingüinos y saber cuáles son las que más destacan. podemos observar el gráfico de correlaciones entre variables y componentes principales.



Lo que podemos observar es que se representa un vector por cada descripción de los pingüinos, siendo los ejes cada una de las componentes seleccionadas, de tal forma que la longitud del vector en cada uno de los ejes representa la correlación en dicha componente, el color del vector es la suma de estas correlaciones al cuadrado ( $\cos^2$ ).

En el gráfico podemos visualizar que las descripciones físicas de los pingüinos están fuertemente relacionadas con la componente del eje x, siendo la componente principal 2, y la descripción del año está fuertemente relacionada con el componente principal 1.

Pero además podemos observar que las variables con el color rojo más intenso son el año y la longitud de la aleta, lo cual indica que son las variables mejor explicadas por el conjunto de componentes, por el lado contrario observamos que la variable menos explicada por el conjunto de componentes es la profundidad del pico.

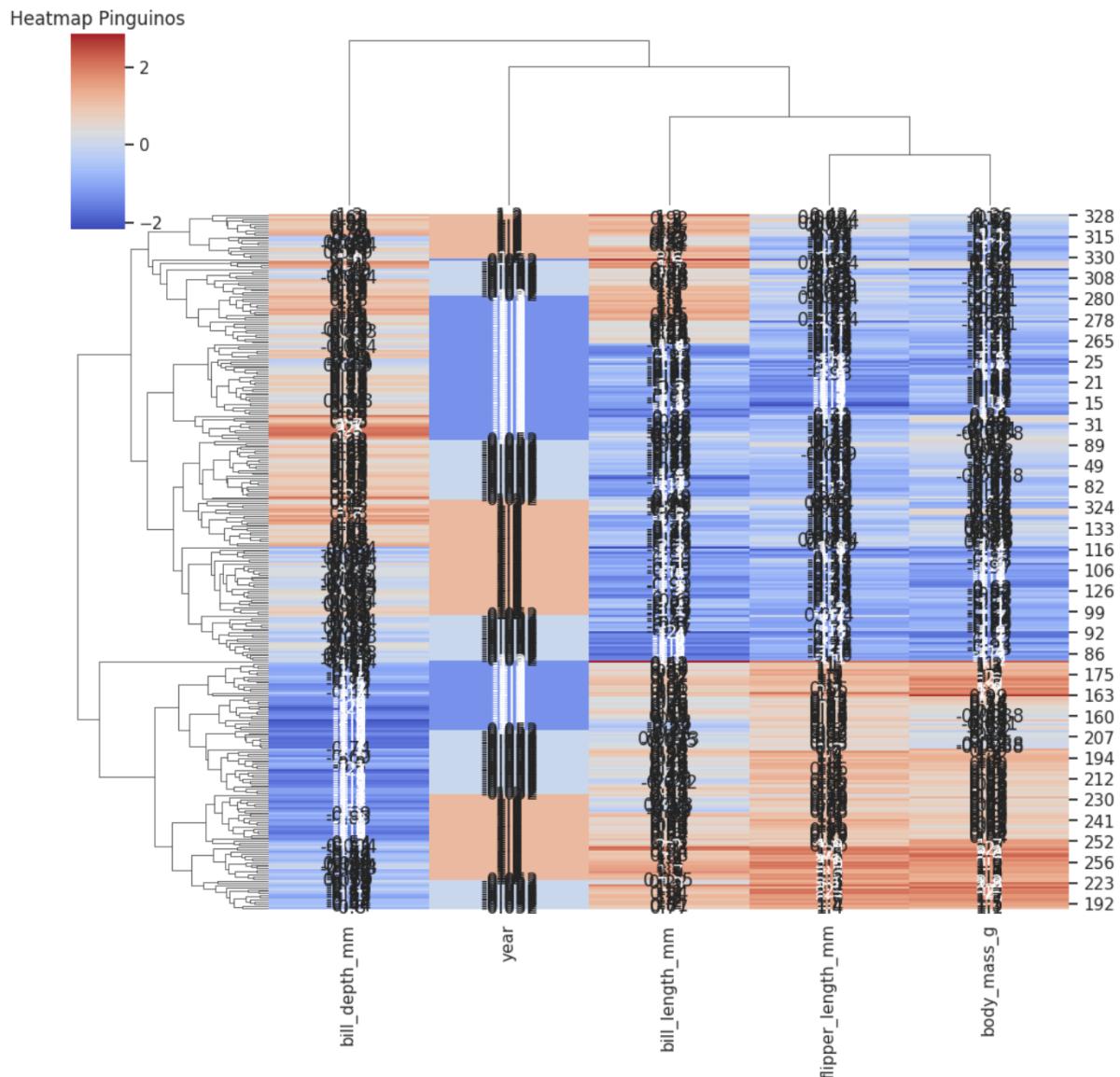
Para construir el índice del conjunto de características físicas del pingüino, basado en los datos y los gráficos vistos anteriormente podría decirse que para la especie Adelie la longitud del pico debería de estar entre 53mm y 59mm y es más probable que se encuentre en la isla de Biscoe, para la especie Gentoo la longitud del pico debería de estar entre 45 mm y 53 mm y es más probable que se encuentre en la isla de Dream y para la especie Chinstrap la longitud del pico debería de estar entre 43 mm y 37 mm y es más probable que se encuentre en la isla de Torgersen.

## Clustering:

Determinar el número de grupos:

Después de aplicar el agrupamiento jerárquico al conjunto de datos y utilizar un dendrograma para obtener el número de grupos podemos decir lo siguiente:

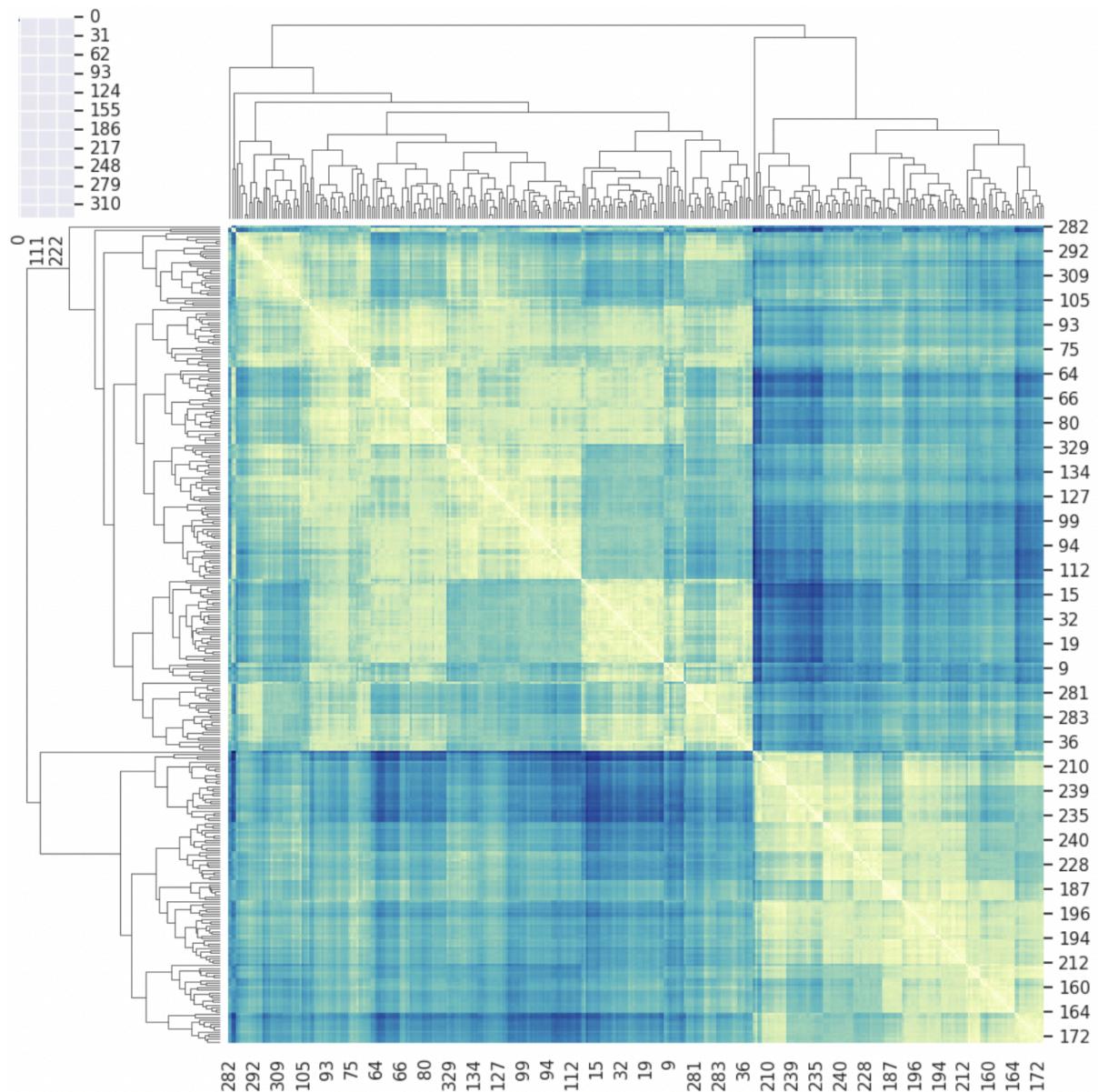
Al comienzo se obtuvo el heatmap así:



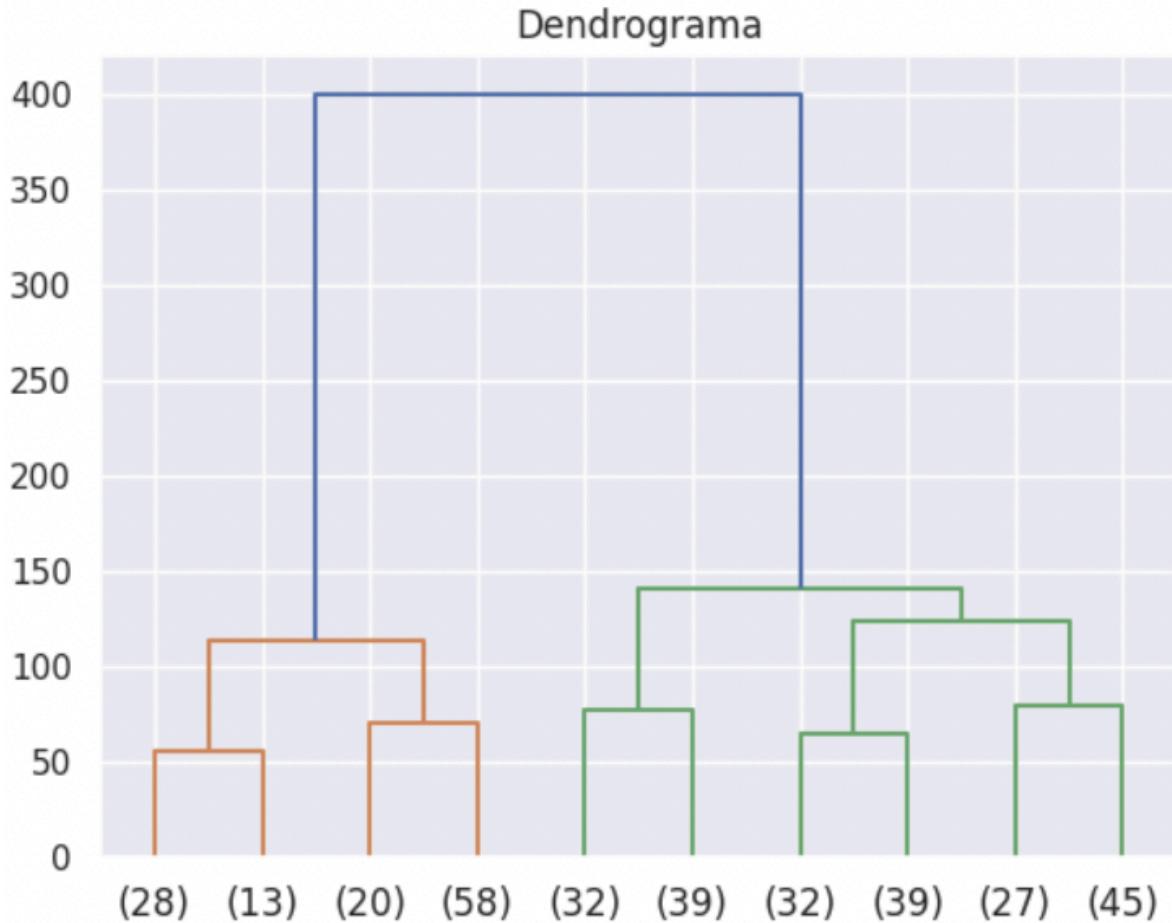
Lo cual no nos proporcionó mucha información de la cual sacar conclusiones. Luego se utilizó una matriz de distancia en la cual calculamos las distancias Euclídeas entre las observaciones (con los valores estandarizados):

	0	1	2	3	4
0	0.00	0.76	1.25	1.08	1.17
1	0.76	0.00	1.00	1.28	1.66
2	1.25	1.00	0.00	0.98	1.47
3	1.08	1.28	0.98	0.00	0.88
4	1.17	1.66	1.47	0.88	0.00

Representemos ahora mediante escalas de color la distancia entre todas las observaciones y se utilizo métodos de reordenación para conseguir una visualización de los grupos óptima de cara a reconocer patrones y similitudes en los datos. Esto ayudó a detectar grupos de observaciones con grandes distancias entre sí.



Y como es frecuente presentar los resultados del análisis clúster jerárquico con un dendrograma. Este es el resultado obtenido:



Observando la estructura del dendrograma podemos hacernos una idea de cuál podría ser el número más adecuado de cluster. En nuestro caso podría ser **cuatro**, ya que se puede apreciar como las uniones de esos cuatro cluster se realizan "pronto" la línea de corte que utilizo es alrededor de 125 (arriba del grupo naranja).

## Agrupamiento K-Means:

Luego de aplicar el algoritmo K-Means y experimentar con diferentes valores obtuve estos resultados, los cuales fueron muy sorprendentes.

### Gráfico utilizando 4 clusters:

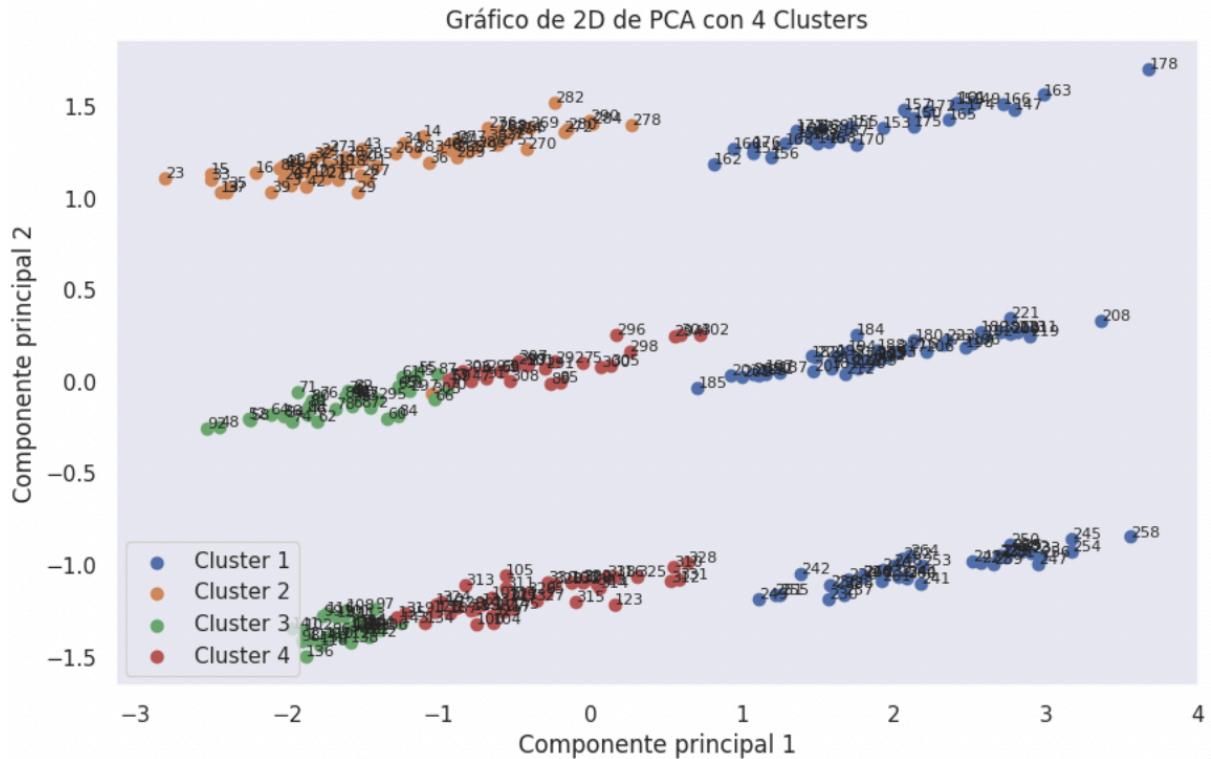


Gráfico utilizando 3 clusters:

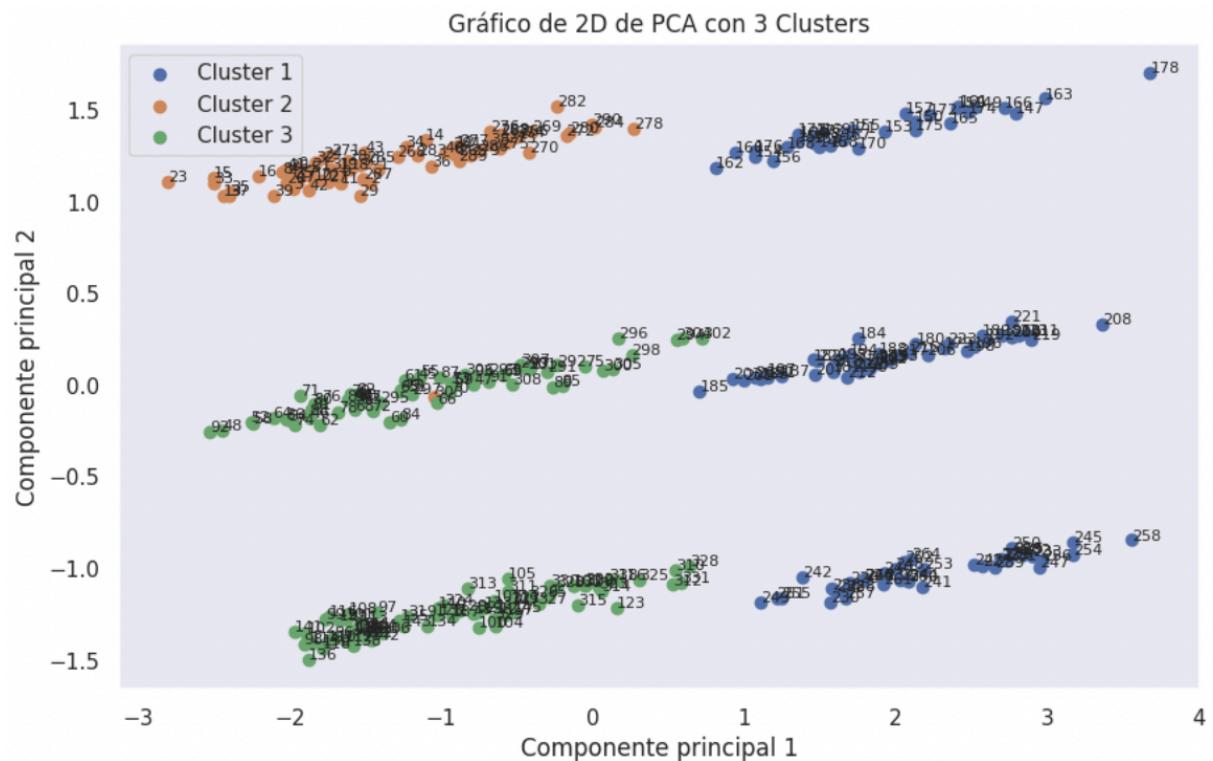
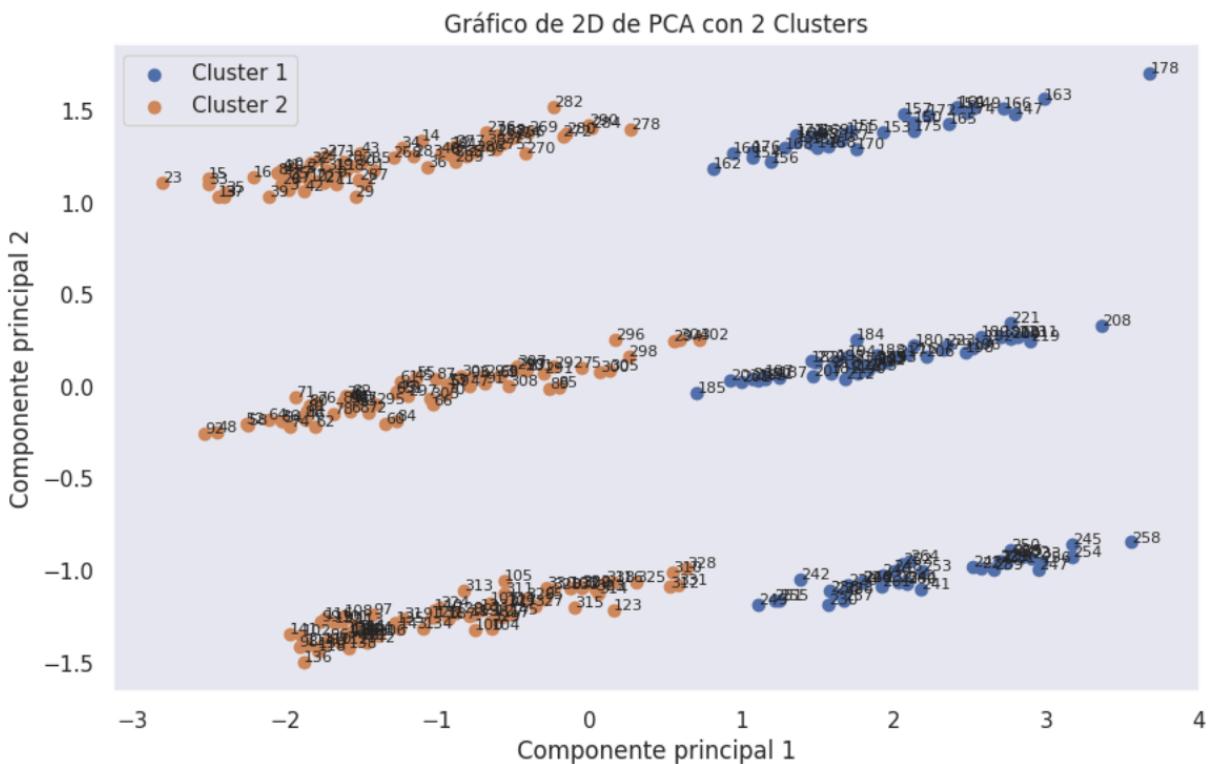
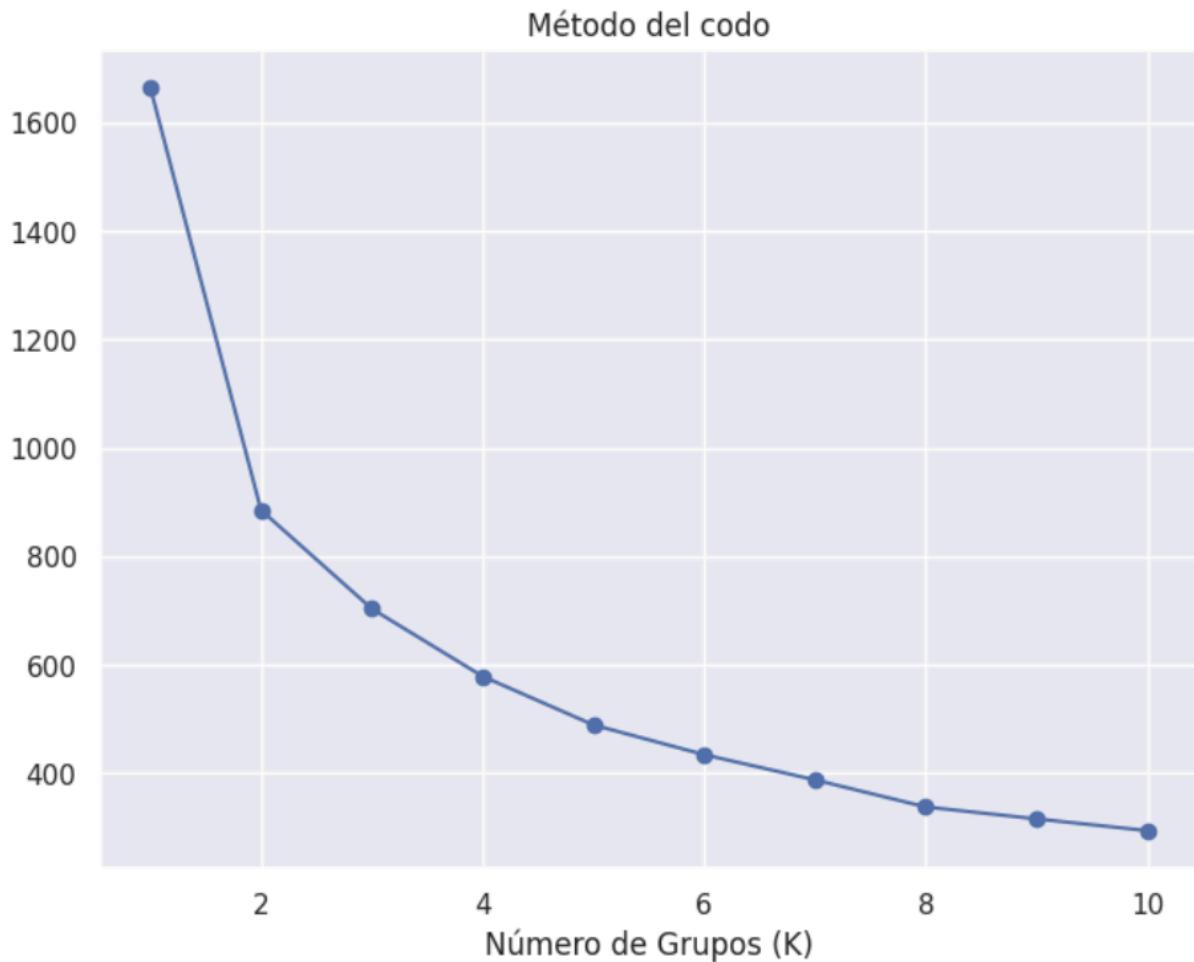


Gráfico utilizando 2 clusters:



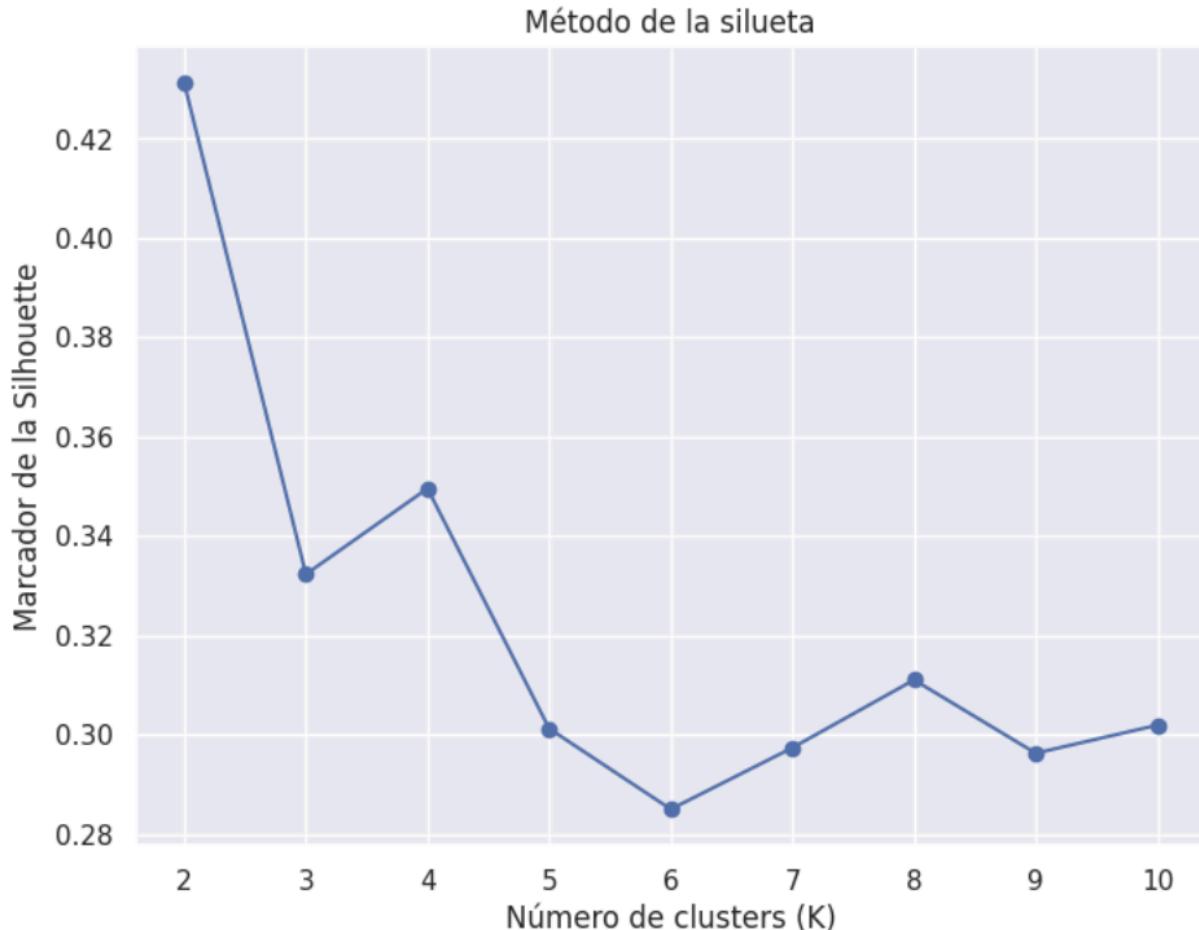
Como se puede observar, en el primer gráfico con 4 clusters los grupos no terminan de encajar del todo, y en el de 2 clusters no me parece lo adecuado, por lo tanto considero que el número óptimo de grupos es el cluster de 3, pues considero que las nubes de datos se adaptan de una mejor manera.

Luego tenemos el método del codo con el cual nos ayuda a confirmar la decisión de cuál es el número adecuado de clusters, se observa que la caída deja de ser tan violenta cuando  $k=3$ , lo cual confirma la decisión inicial.



## Validación del Agrupamiento:

Se aplicaron las métricas de validación de agrupamiento, se utilizó el método de silueta para evaluar la calidad de los resultados. Y se obtuvo la siguiente gráfica:



Se sabe que la puntuación de silueta varía entre -1 y 1. Una puntuación cercana a 1 indica que la observación coincide bien con su propio grupo y no coincide con los grupos vecinos. Una puntuación cercana a 0 indica grupos superpuestos, donde la observación podría estar en cualquiera de los grupos adyacentes. Una puntuación cercana a -1 indica que la observación probablemente esté asignada al grupo incorrecto.

Se observa que en la gráfica solo se presentan valores 0.28 hasta 0.50 aproximadamente, y si nos fijamos en  $k=3$  que es el número de cluster que nos interesa, vemos que el valor es aproximadamente 0.31 lo cual nos da una idea que existen grupos superpuestos, donde la observación podría estar en cualquiera de los grupos adyacentes. Por lo tanto no **podemos decir que la efectividad del algoritmo es perfecta para agrupar grupos, pues existe una posibilidad de que hayan grupos superpuestos.**

### **Similitudes y/o diferencias entre grupos:**

La principal diferencia que existe entre jerárquico versus K-Means es el número de grupos que obtuve como resultado, pues en el jerárquico obtuve como resultado que había 4 grupos y en el k-means luego de probar con diferentes valores de k, pude observar que la k que mejor se adapta al problema era 3.

La similitud es que si nos vamos por un rango ambos métodos me daban que la respuesta se ubicaba entre 2-4 grupos.

**¿Qué representan los grupos identificados en el contexto de las especies de pingüinos? ¿Existen patrones o tendencias significativas?**

Lo que representa cada grupo identificado es una especie de pingüino, pues cada uno de ellos tiene características diferentes las cuales se encontraron y se demostraron luego de aplicar las técnicas de agrupamiento. El resultado final fue que existían 3 grupos diferentes, y claro que existen patrones dentro de cada especie por ejemplo la especie Adeline suele ser la más grande en medidas, mientras que la especie Gentoo a pesar de no ser la más grande en tamaño es la más pesada.

### **Conclusiones:**

- El gráfico de dispersión del comienzo nos da una muy buena aproximación de la cantidad de grupos que hay con tan solo ver la distribución de las nubes de puntos de diferentes colores.
- Solo se necesitan 2 componentes principales para explicar un 75,13% de la variabilidad, lo cual es bueno porque se reducen de 5 variables a solo 2.
- Si se quisiera un mejor porcentaje de variabilidad con tomar un tercer componente principal se podría explicar un 90,67%.
- Con el método de agrupación jerárquico se encontraron 4 grupos después de ver el dendograma, con el método de agrupación K-Means el resultado final fue 3, lo cual indica una diferencia en el valor de k, en donde optamos como respuesta final que existen 3 grupos por la mejor representación de los datos del método K-Means.
- La validación por agrupamiento nos dice que podemos decir que la efectividad del algoritmo es perfecta para agrupar grupos, pues existe una posibilidad de que hayan grupos superpuestos, pues obtuvimos un valor de 0,31.