Análisis de Datos de Pingüinos para National Geographic

Introducción

Este documento presenta un análisis detallado del conjunto de datos 'penguins' de la librería 'seaborn' en Python, con el objetivo de aplicar técnicas de reducción de dimensionalidad y agrupamiento. Se busca explorar las relaciones entre las características físicas de diferentes especies de pingüinos y entre las propias especies.

Descripción del Conjunto de Datos

El conjunto de datos incluye información sobre varias especies de pingüinos, recopilada en diferentes islas, y abarca las siguientes características:

- `species`: Especie del pingüino.
- `island`: Isla de recopilación de datos.
- `bill length mm`: Longitud del pico.
- `bill depth mm`: Profundidad del pico.
- `flipper length mm`: Longitud de la aleta.
- `body mass g`: Masa corporal.
- `sex`: Género del pingüino.

Apartados a Realizar

Carga y Exploración del Dataset

Tareas Iniciales (0.5 puntos)

1. **Carga del Conjunto de Datos** Para cargar el conjunto de datos 'penguins' desde Seaborn, utilizamos el siguiente código:

```
import seaborn as sns

# Cargando el conjunto de datos de Palmer Penguins
penguins_data = sns.load_dataset("penguins")
print(penguins_data.head())
```

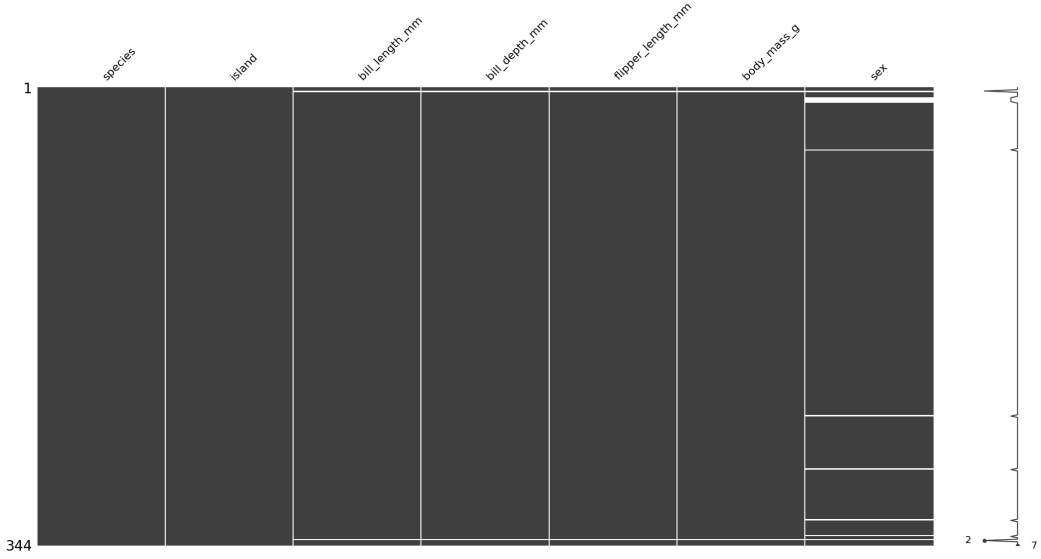
Esto nos da las primeras cinco filas del conjunto de datos para tener una primera impresión del tipo de datos con los que trabajaremos:

```
    species     island  bill_length_mm  bill_depth_mm  flipper_length_mm
body_mass_g     sex
0  Adelie  Torgersen            39.1           18.7              181.0
3750.0    Male
```

```
1  Adelie  Torgersen           39.5           17.4              186.0
3800.0  Female
2  Adelie  Torgersen           40.3           18.0              195.0
3250.0  Female
3  Adelie  Torgersen            NaN            NaN                NaN
NaN     NaN
4  Adelie  Torgersen           36.7           19.3              193.0
3450.0  Female
```

2. **Estadísticas Descriptivas**

   En primer lugar, usando la librería `missingno` de python vamos a visualizar gráficmente los valores perdidos del dataset:



   Como vemos, los valores perdidos suelen darse en su mayoría en la columna "sex" del dataframe, o en algunas filas simultáneamente para "bill_length_mm" , "bill_depth_mm", "flipper_length_mm", "body_mass_g". Para conocer el número exacto de valores perdidos ejecutamos:

```
penguins_data.isna().sum()
```

   Obteniendo:

```
species              0
island               0
bill_length_mm       2
bill_depth_mm        2
flipper_length_mm    2
body_mass_g          2
sex                 11
```

Para limpiar estos valores perdidos vamos a proceder de la siguiente forma:

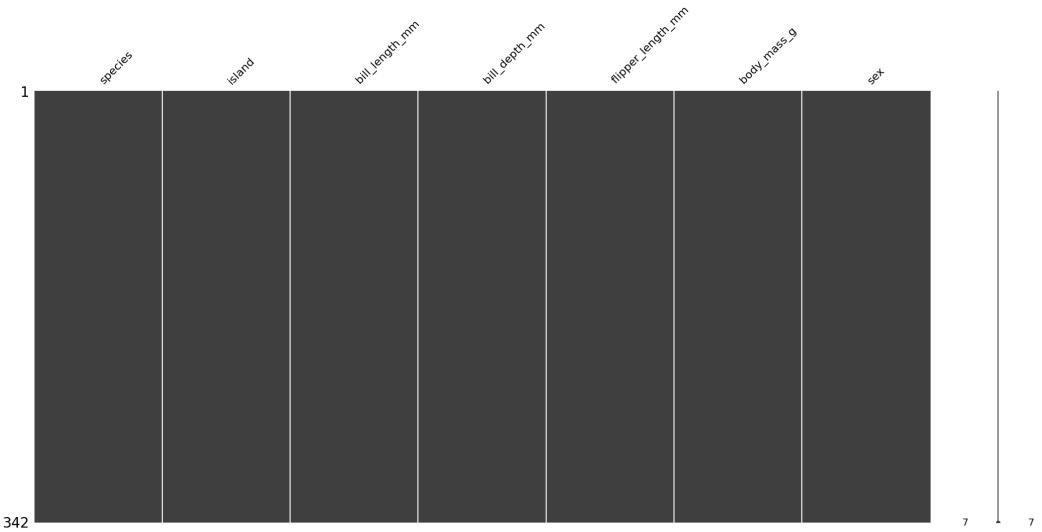
- Eliminamos las filas que simultáneamente tienen valores perdidos en "bill_length_mm" , "bill_depth_mm", "flipper_length_mm", "body_mass_g", pues solamente son dos filas y no supone una pérdida de datos relevante.
- Imputamos el valor de sex con el siguiente método: Para cada fila con el valor sex como valor perdido, calculamos la distancia euclídea teniendo en cuenta "bill_length_mm" , "bill_depth_mm", "flipper_length_mm", "body_mass_g" a la media para los ejemplares "hembras" y "Machos". Si está más próximo a la media de los "Machos" se imputa como este valor, en caso contrario como "hembra". El código es el siguiente:

```
# Imputación del valor de 'sex'
# Calcula las medias para machos y hembras
means_male = penguins_data[penguins_data['sex'] == 'Male']
[columnas_a_verificar].mean()
means_female = penguins_data[penguins_data['sex'] == 'Female']
[columnas_a_verificar].mean()

# Función para calcular la distancia a las medias
def impute_sex(row):
    if pd.isna(row['sex']):
        dist_to_male = euclidean(row[columnas_a_verificar], means_male)
        dist_to_female = euclidean(row[columnas_a_verificar],
means_female)
        return 'Male' if dist_to_male < dist_to_female else 'Female'
    else:
        return row['sex']

# Aplicar la función para imputar 'Sex'
penguins_data['sex'] = penguins_data.apply(impute_sex, axis=1)
```

Tras esto, se eliminan los valores perdidos del dataset:



Por otro lado se calculan estadísticos para las columnas numéricas Obteniendo los siguientes resultados:

```
                   count         mean         std     min       25%
50%     75%     max  Datos Perdidos
bill_length_mm     342.0    43.921930    5.459584    32.1    39.225    44.45
48.5    59.6               0
bill_depth_mm      342.0    17.151170    1.974793    13.1    15.600    17.30
18.7    21.5               0
flipper_length_mm  342.0   200.915205   14.061714   172.0   190.000   197.00
213.0   231.0               0
body_mass_g        342.0  4201.754386  801.954536  2700.0  3550.000  4050.00
4750.0  6300.0               0
```

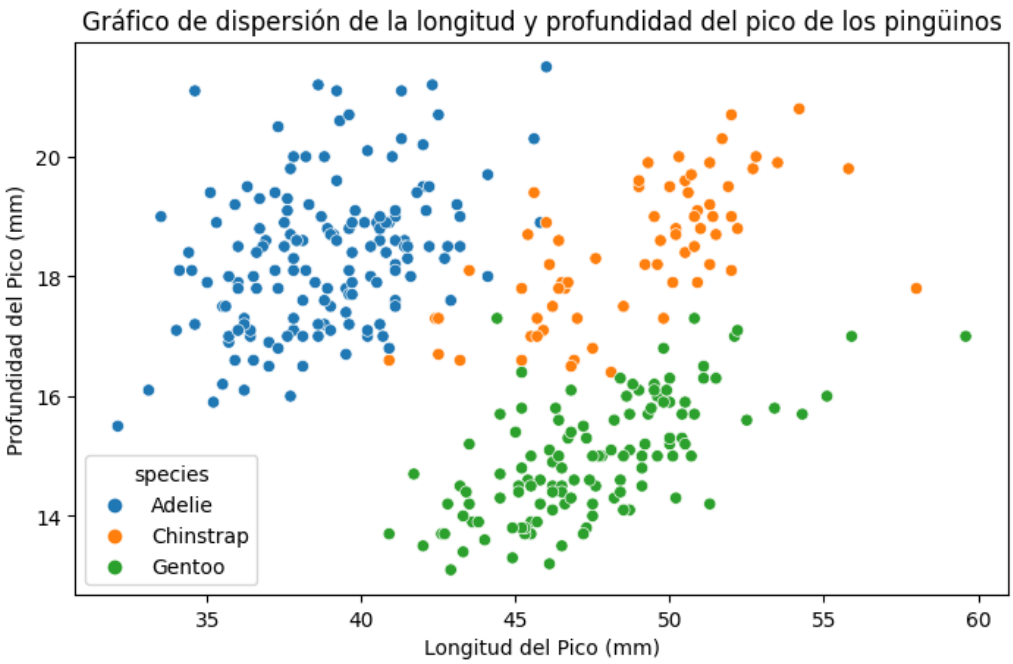
Las estadísticas descriptivas del conjunto de datos 'penguins' muestran que:

- La longitud del pico tiene una media de aproximadamente 43.92 mm y una variabilidad moderada.
- La profundidad del pico es menos variable que la longitud y tiene una media de 17.15 mm.
- La longitud de la aleta muestra una dispersión similar a la longitud del pico con una media de 200.91 mm.
- La masa corporal de los pingüinos varía sustancialmente, con una media de 4201.75 g. Se presenta un posible valor atípico pues el máximo es de 6300.

3. **Gráficos de Dispersión y Visualizaciones Relevantes**

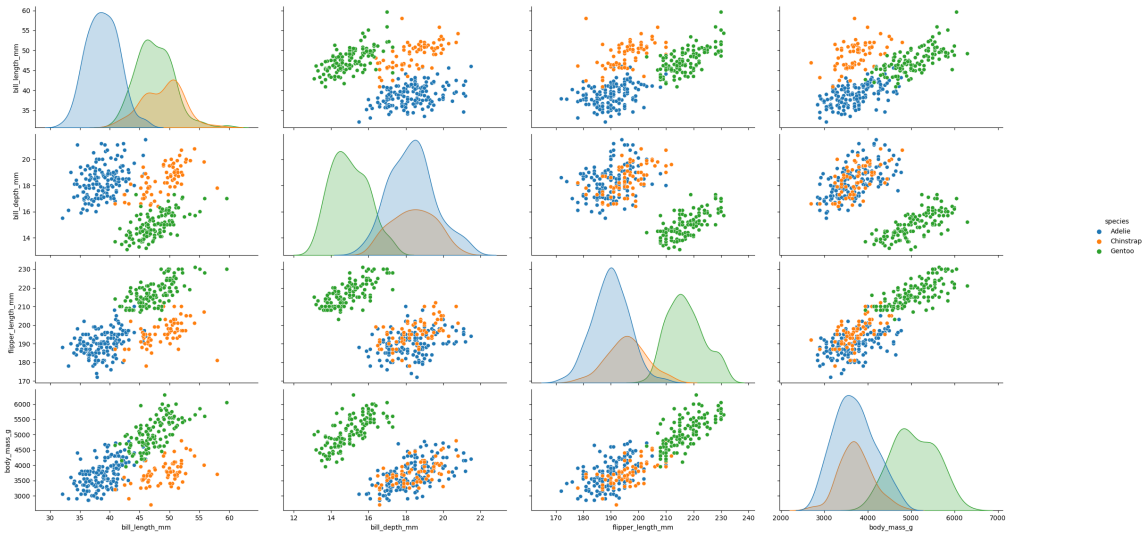
La exploración gráfica del conjunto de datos de pingüinos, utilizando gráficos de dispersión y pairplots, proporciona información valiosa sobre la estructura y las relaciones en los datos:

- **Gráfico de Dispersión de Longitud y Profundidad del Pico**



Este gráfico muestra una clara separación entre las tres especies de pingüinos basándose en la longitud y la profundidad de sus picos. Los pingüinos Adelie tienden a tener picos más cortos y profundos, los Chinstrap tienen picos más largos y menos profundos, y los Gentoo se distinguen por tener los picos más largos y menos profundos de los tres. Esta distinción sugiere que la longitud y la profundidad del pico podrían ser buenos predictores para la clasificación de las especies.

- **Pairplot de Características Físicas**



El pairplot muestra las distribuciones y correlaciones entre la longitud del pico, la profundidad del pico, la longitud de la aleta y la masa corporal. Se observa que los Gentoo, en general, tienen las aletas más largas y la mayor masa corporal, mientras que los Adelie son más pequeños en ambos aspectos. Los Chinstrap se ubican en un rango intermedio en la mayoría de las características. La masa corporal parece aumentar con la longitud de la aleta, lo que indica una posible correlación entre estas dos variables.

Ambos gráficos destacan diferencias morfológicas significativas entre las especies, lo que puede reflejar adaptaciones a diferentes nichos ecológicos. Estas visualizaciones también sugieren que las medidas morfológicas de los pingüinos son multidimensionales y que la variación dentro de cada especie es considerable, lo que justifica un análisis más profundo para comprender las dinámicas de la población y la adaptación de las especies.

Reducción de Dimensionalidad

Análisis de Componentes Principales (ACP) (1.5 puntos)

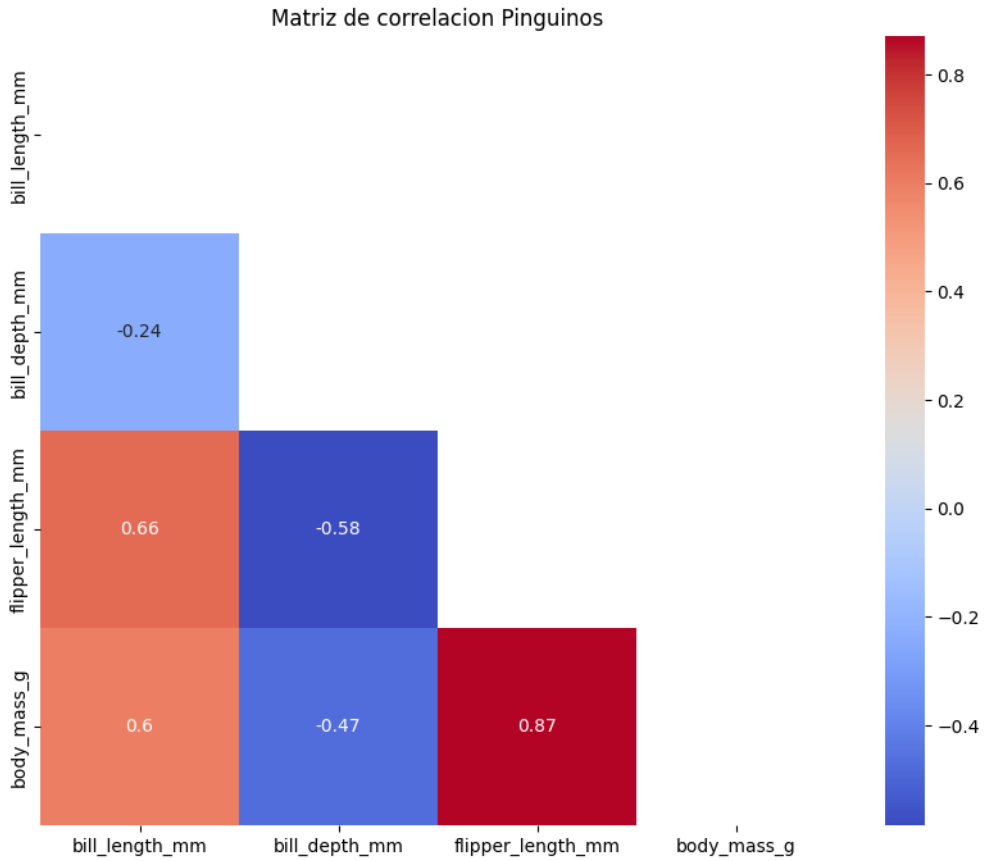
1. **Matriz de Correlaciones y Representación Gráfica (0.5 pts)** En primer lugar vamos a calcular la matriz de correlaciones y representarla gráficamente, para ello empleamos el siguiente código:

```
# Matriz de correlaciones
    matriz_correlaciones = penguins_data.corr(numeric_only=True)

    # Crear una máscara para la matriz triangular superior
    mask = np.triu(np.ones_like(matriz_correlaciones, dtype=bool))
    matriz_correlaciones[mask] = np.nan

    # Crear un mapa de calor para la matriz de correlación
    plt.figure(figsize=(10, 8))
    sns.heatmap(matriz_correlaciones, annot=True, cmap='coolwarm')
    plt.title("Matriz de correlacion Pinguinos")
    plt.show()
```

Obteniendo la siguiente matriz de correlaciones:



La matriz de correlación proporcionada muestra las siguientes correlaciones entre las medidas físicas de los pingüinos:

- **Correlación Directa más Fuerte (positiva)**: Entre la longitud de la aleta (flipper_length_mm) y la masa corporal (body_mass_g) con un coeficiente de 0.87. Esto indica que los pingüinos con aletas más largas tienden a tener una masa corporal mayor, lo cual indica una relación lógica entre el tamaño del pingüino y su peso.
- **Correlación Inversa más Fuerte (negativa)**: Entre la longitud de la aleta (flipper_length_mm) y la profundidad del pico (bill_depth_mm) con un coeficiente de -0.58. Esto sugiere que los pingüinos con aletas más largas tienden a tener picos menos profundos, lo que podría reflejar adaptaciones específicas entre diferentes especies o dentro de una población en respuesta a sus hábitats y estilos de alimentación.

Estas correlaciones pueden tener implicaciones ecológicas y evolutivas significativas, ya que las diferentes medidas pueden estar relacionadas con estrategias de alimentación, hábitats preferidos y otras presiones de selección ambiental.

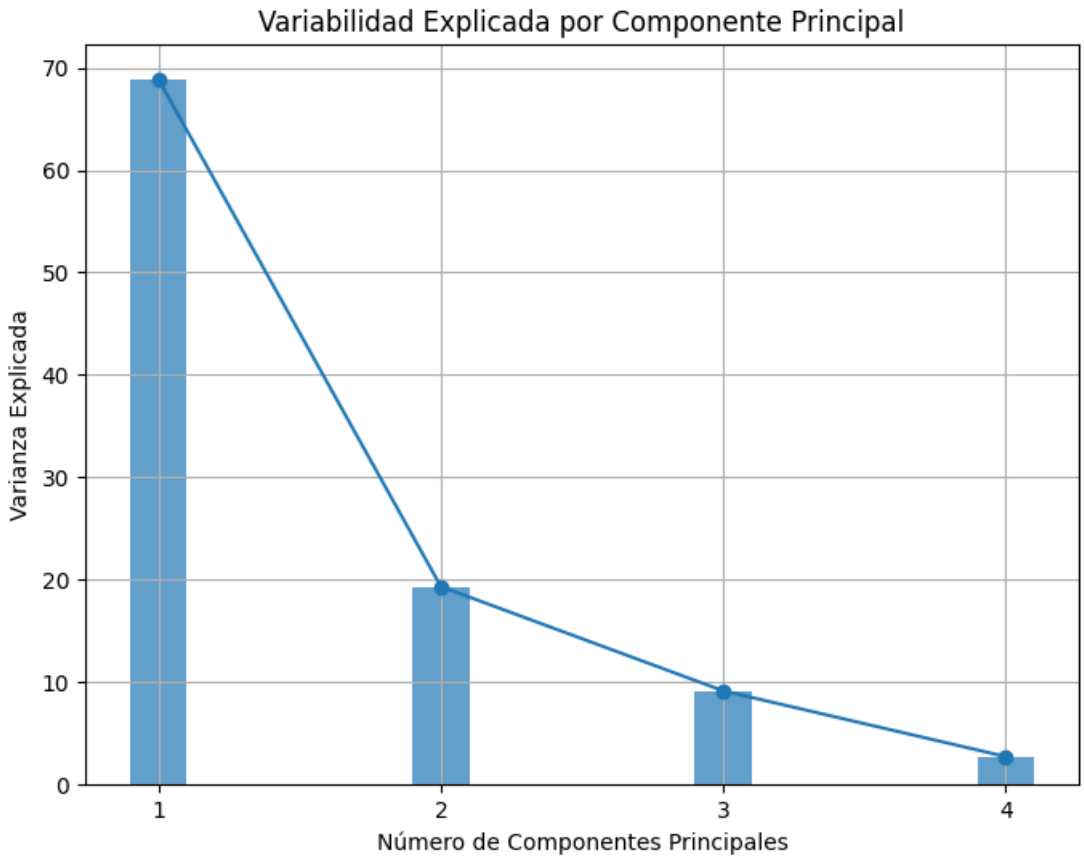
2. **PCA con Datos Estandarizados (1 pt)**

Cálculo de componentes y análisis de autovalores: Los autovalores obtenidos de un PCA proporcionan una medida de la variabilidad que cada componente principal captura del conjunto de datos. En este caso, los autovalores y la variabilidad explicada por cada componente son:

- *Componente 1*: Autovalor de 2.761831, explicando el 68.843878% de la variabilidad.
- *Componente 2*: Autovalor de 0.774782, explicando el 19.312919% de la variabilidad.
- *Componente 3*: Autovalor de 0.366307, explicando el 9.130898% de la variabilidad.
- *Componente 4*: Autovalor de 0.108810, explicando el 2.712305% de la variabilidad.

Estos valores sugieren que el primer componente principal captura la mayoría de la variabilidad, seguido por el segundo componente. Los componentes 3 y 4 contribuyen significativamente menos a la explicación de la variabilidad en los datos.

Gráficas resumen y decisión sobre el número de componentes:



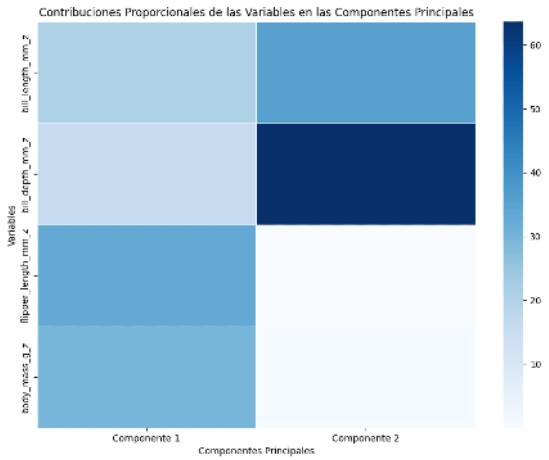
En la gráfica obtenida se aprecia un fuerte "codo" después del segundo componente principal, lo que indica que la inclusión de más componentes no añade mucha información. La variabilidad acumulada hasta el segundo componente es del 88.156797%, lo que significa que estos dos componentes juntos capturan la gran mayoría de la información en el conjunto de datos.

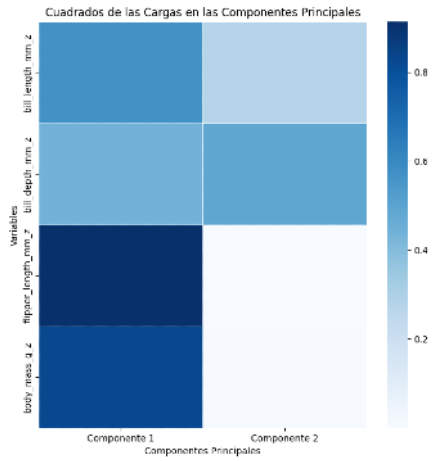
Decisión sobre el número de componentes: En la práctica, una regla común es elegir el número de componentes que suman una variabilidad acumulada cercana al 70-90%. Dado que los dos primeros componentes explican aproximadamente el 88% de la variabilidad, es razonable seleccionar solo estos dos para una representación eficiente de los datos. Esto reduce la dimensionalidad del conjunto de datos mientras se retiene la mayoría de la información, lo que facilita la visualización y el análisis subsiguiente de los datos.

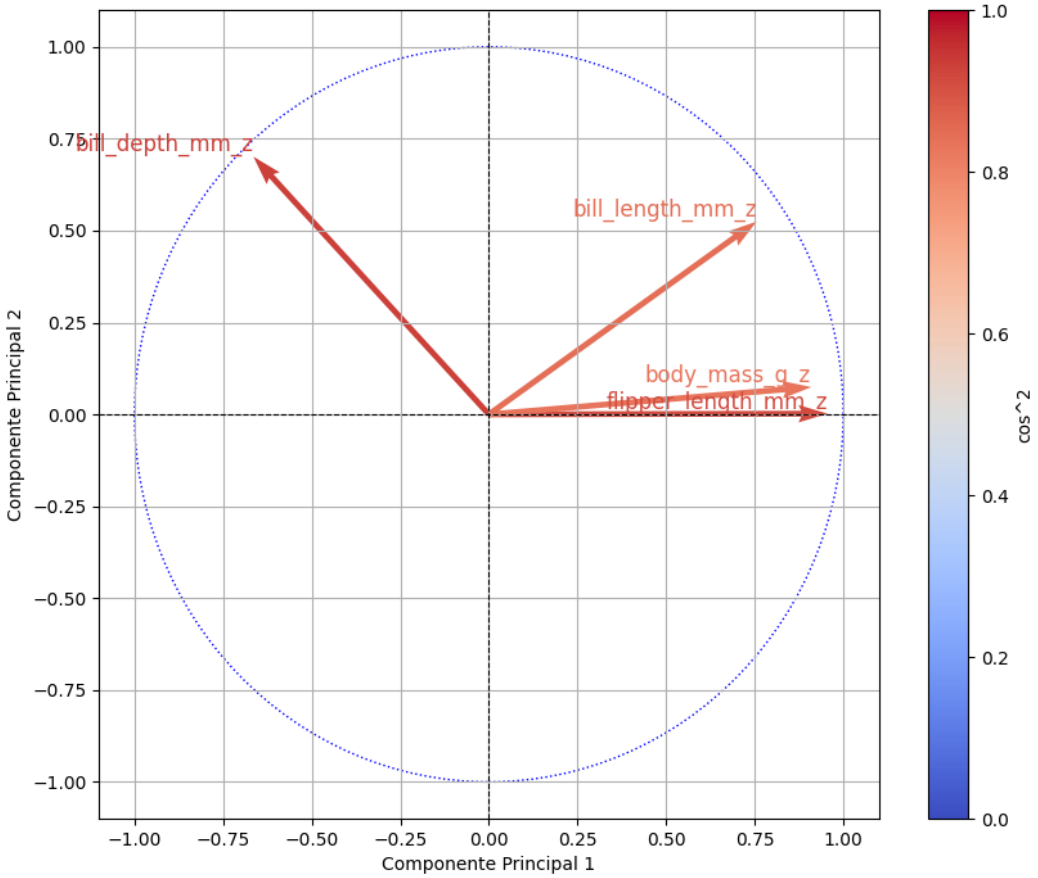
Por lo tanto, se recomienda utilizar dos componentes principales para representar este conjunto de datos de pingüinos.

Interpretación de Componentes Principales (2 pts)

1. **Representación Gráfica de Variables en Componentes**



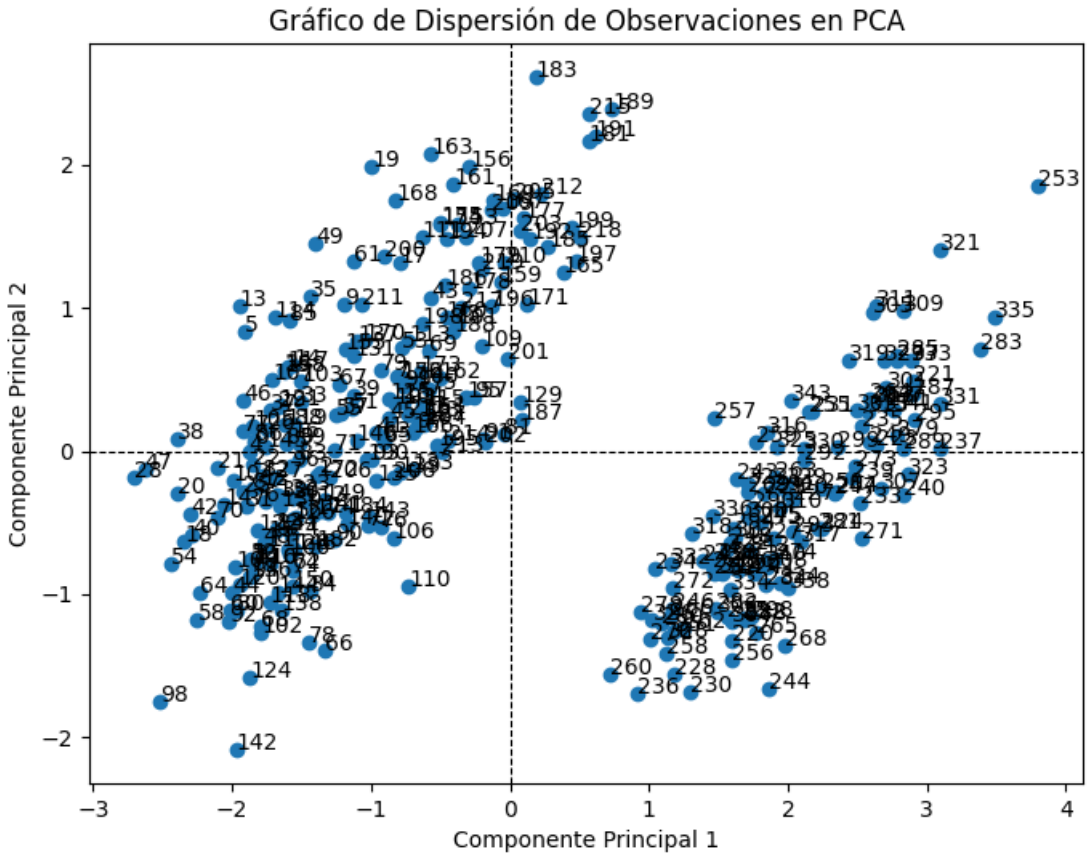


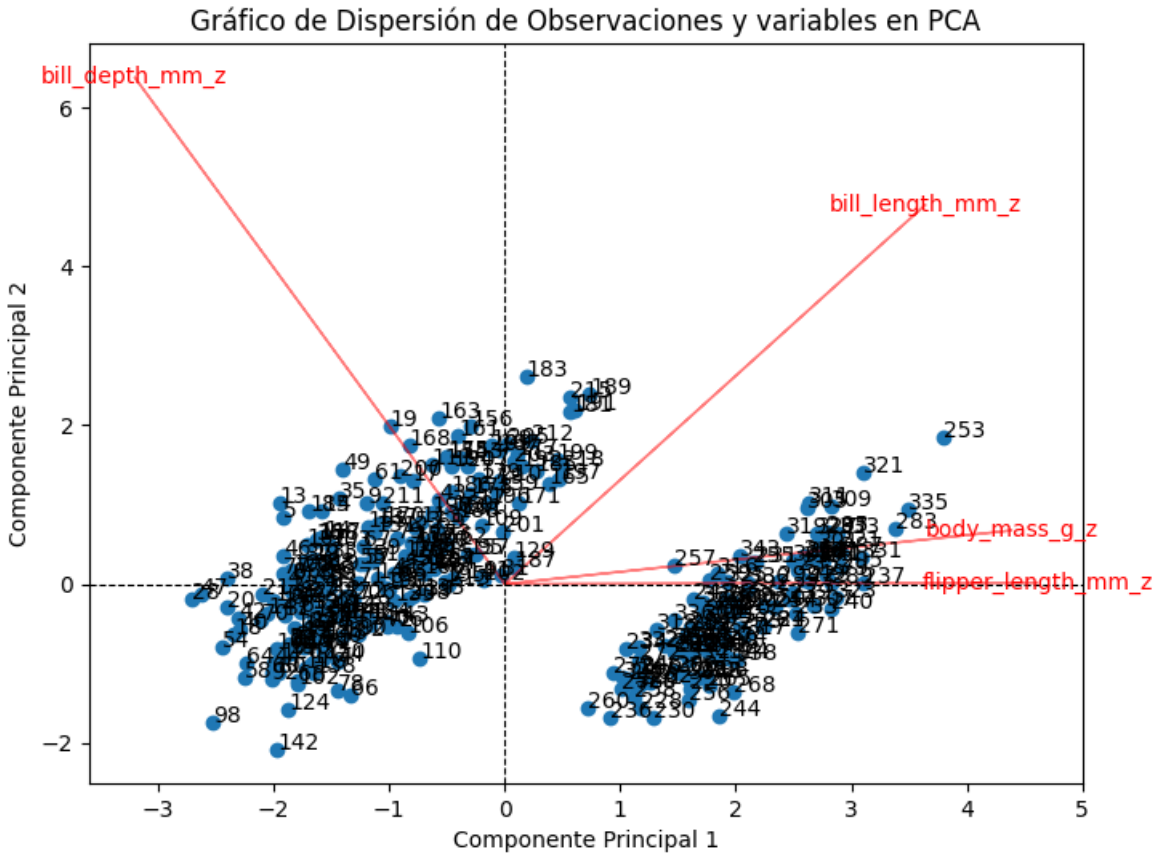


La representación gráfica de las variables en los componentes principales, conocida como biplot, permite observar cómo cada variable contribuye a los componentes. Las imágenes que muestran las contribuciones proporcionales y los cuadrados de las cargas en las componentes principales revelan la importancia relativa de cada medida física en los componentes seleccionados.

- *Componente 1*: Este componente captura la mayor varianza y parece estar fuertemente influenciado por todas las variables físicas, como la longitud de la aleta y la masa corporal, lo que sugiere que puede representar el tamaño general o la 'corpulencia' de los pingüinos.
- *Componente 2*: El segundo componente, aunque captura menos varianza que el primero, parece estar más relacionado con la longitud del pico y la profundidad del pico, lo que podría reflejar adaptaciones específicas relacionadas con el comportamiento alimenticio de los pingüinos o diferencias entre las especies.

2. **Representación de Observaciones en Nuevos Ejes**





Las especies de pingüinos que se destacan en cada componente se pueden inferir de la posición de las observaciones en el gráfico de dispersión de PCA.

- Las observaciones que tienen valores altos en la Primera Componente Principal son aquellos pingüinos que son grandes en términos de masa corporal y tamaño del pico.
- Las observaciones que tienen valores altos o bajos en la Segunda Componente Principal son aquellos pingüinos que tienen características distintivas de pico y aletas que no están directamente relacionadas con el tamaño general. Por ejemplo, una aleta más larga o un pico más profundo en relación con su masa corporal.

3. **Construcción de un Índice para Características Físicas** La construcción de un índice compuesto para representar las características físicas de los pingüinos puede proporcionar una única medida descriptiva que resuma las variables más significativas. Utilizando los componentes principales, podemos desarrollar un índice que capture la esencia de la variabilidad física de los pingüinos.

   El **índice de características físicas (ICF)** podría definirse como una combinación lineal de las variables estandarizadas, ponderadas por su contribución al primer componente principal, que es el que más varianza explica. Este índice se calcularía de la siguiente manera para cada pingüino:

```
ICF = (carga_1 * bill_length_mm_z) + (carga_2 * bill_depth_mm_z) +
    (carga_3 * flipper_length_mm_z) + (carga_4 * body_mass_g_z)
```

Donde carga_n es la carga de la variable en el Componente Principal 1 y variable_z es el valor estandarizado de dicha variable.

Para evaluar el índice de cada especie, calcularíamos el promedio del ICF para todas las observaciones de cada especie. Esto nos daría una puntuación única que reflejaría las tendencias generales en las características físicas para las especies Adelie, Chinstrap y Gentoo. Los valores altos del índice indicarían pingüinos con características físicas que corresponden positivamente con el primer componente principal, mientras que los valores bajos indicarían lo contrario.

Este índice podría tener aplicaciones prácticas, por ejemplo, en estudios de conservación para monitorear la salud y el bienestar de las poblaciones de pingüinos en relación con sus hábitats y disponibilidad de recursos.

Este enfoque proporciona una herramienta simplificada para analizar y comparar las características físicas de las diferentes especies de pingüinos, lo cual es particularmente útil cuando se manejan múltiples variables y se desea tener una visión integrada del fenotipo de los organismos estudiados.

Técnicas de Agrupamiento (Clustering)

Determinación del Número de Grupos (1 pt)

1. **Agrupamiento Jerárquico y Dendrograma**
    - Elección y justificación del número de grupos.

Agrupamiento K-Means (1 pt)

1. **Implementación y Experimentación con K-Means**
    - Determinación del número óptimo de grupos usando métricas apropiadas.

Validación del Agrupamiento (1 pt)

1. **Métricas de Validación de Agrupamiento**
    - Evaluación de la calidad de los resultados del agrupamiento.

Comparación de Métodos de Agrupamiento (1 pt)

1. **Jerárquico vs. K-Means**
    - Comparación y contraste de los resultados obtenidos.

Interpretación de los Grupos (1 pt)

1. **Análisis de los Grupos Identificados**
    - Interpretación y análisis de patrones y tendencias.

Conclusión

1. **Resumen de Hallazgos**
    - Discusión de limitaciones o desafíos encontrados.
    - Conclusiones finales del análisis.

Pregunta 3

a) Comentario sobre los gráficos que representan las variables en los planos formados por las componentes: Las imágenes que muestran las contribuciones de las variables a las componentes principales nos ayudan a entender qué característica física de los pingüinos es capturada por cada componente principal. Generalmente, una componente principal que tiene altas cargas (valores absolutos grandes en sus vectores propios) para ciertas variables significa que esas variables contribuyen significativamente a la variabilidad en esa dirección.

```
La Primera Componente Principal parece capturar la mayor parte de la variabilidad
asociada con el tamaño general de los pingüinos, ya que variables como la longitud
del pico, la profundidad del pico, y la masa corporal tienen grandes
contribuciones. Esto sugiere que esta componente podría interpretarse como un
factor de "tamaño general" o "masa corporal" de los pingüinos.

La Segunda Componente Principal podría estar capturando aspectos relacionados con
la morfología específica, posiblemente diferenciando entre las proporciones del
pico y la longitud de las aletas en relación con el tamaño corporal.
```

b) Comentario sobre los gráficos que representan las observaciones en los nuevos ejes: Las especies de pingüinos que se destacan en cada componente se pueden inferir de la posición de las observaciones en el gráfico de dispersión de PCA.

```
Las observaciones que tienen valores altos en la Primera Componente Principal son
aquellos pingüinos que son grandes en términos de masa corporal y tamaño del pico.

Las observaciones que tienen valores altos o bajos en la Segunda Componente
Principal son aquellos pingüinos que tienen características distintivas de pico y
aletas que no están directamente relacionadas con el tamaño general. Por ejemplo,
una aleta más larga o un pico más profundo en relación con su masa corporal.
```

c) Construcción de un índice utilizando una combinación lineal de todas las variables: Un índice que valore de forma conjunta las características físicas de un pingüino podría construirse tomando los pesos de las cargas de las variables en las componentes principales y sumándolos para cada pingüino. Este índice sería esencialmente una puntuación compuesta basada en las componentes principales que hemos decidido retener.

Para construirlo, podríamos calcular la suma ponderada de las variables estandarizadas para cada pingüino, utilizando los pesos de las cargas de las componentes principales que hemos retenido. Por ejemplo, si retenemos dos componentes principales, el índice para un pingüino podría ser:

Índice_pingüino = (carga_CP1 × valor_variable_1) + (carga_CP2 × valor_variable_2) + ... + (carga_CPn × valor_variable_n)

El valor del índice para una especie de pingüino representada por el conjunto de datos sería el promedio de los índices de todos los pingüinos dentro de esa especie.

```
Para la especie 'Adelie', calcularíamos el valor del índice utilizando los valores
medios de sus características físicas multiplicados por las cargas
correspondientes de las componentes principales retenidas.

Para la especie 'Chinstrap', haríamos lo mismo con los valores medios de las
características físicas de los pingüinos de esa especie.
```

Estos índices nos darían una puntuación que refleja las características físicas predominantes de las especies basadas en las componentes retenidas del PCA. Estos valores serían únicos para cada especie y podrían servir para diferenciar entre ellas basándonos en las características físicas medidas.

Para realizar estos cálculos y obtener los valores de índice específicos para las especies mencionadas, necesitaríamos los datos crudos y los pesos de las cargas de las componentes de PCA.