Práctica

Imagina que estamos interesados en predecir las especies de pingüinos basándonos en dos de sus medidas corporales: la longitud y profundidad del pico. Primero queremos hacer una exploración de los datos para tener una idea de ellos.

¿Cuáles son las características? ¿Cuál es el objetivo?

Los datos se encuentran en `penguins.csv`, cárgalos con `pandas` en un `DataFrame`.

```
import pandas as pd

df = pd.read_csv("penguins.csv")
```

Muestra las primeras 5 filas de los datos.

¿Cuántas características son numéricas? ¿Cuántas características son categóricas?

```
# 5 caracteristicas numericas y 3 categoricas
df.head()
```

	rowid	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass
0	1	Adelie	Torgersen	39.1	18.7	181.0	375
1	2	Adelie	Torgersen	39.5	17.4	186.0	380
2	3	Adelie	Torgersen	40.3	18.0	195.0	325
3	4	Adelie	Torgersen	NaN	NaN	NaN	N
4	5	Adelie	Torgersen	36.7	19.3	193.0	345

Next steps: Generate code with `df` View recommended plots New interactive sheet

¿Cuáles son las diferentes especies de pingüinos disponibles en el conjunto de datos y cuántas muestras de cada especie hay? Sugerencia: selecciona la columna correcta y utiliza el método `value_counts`.

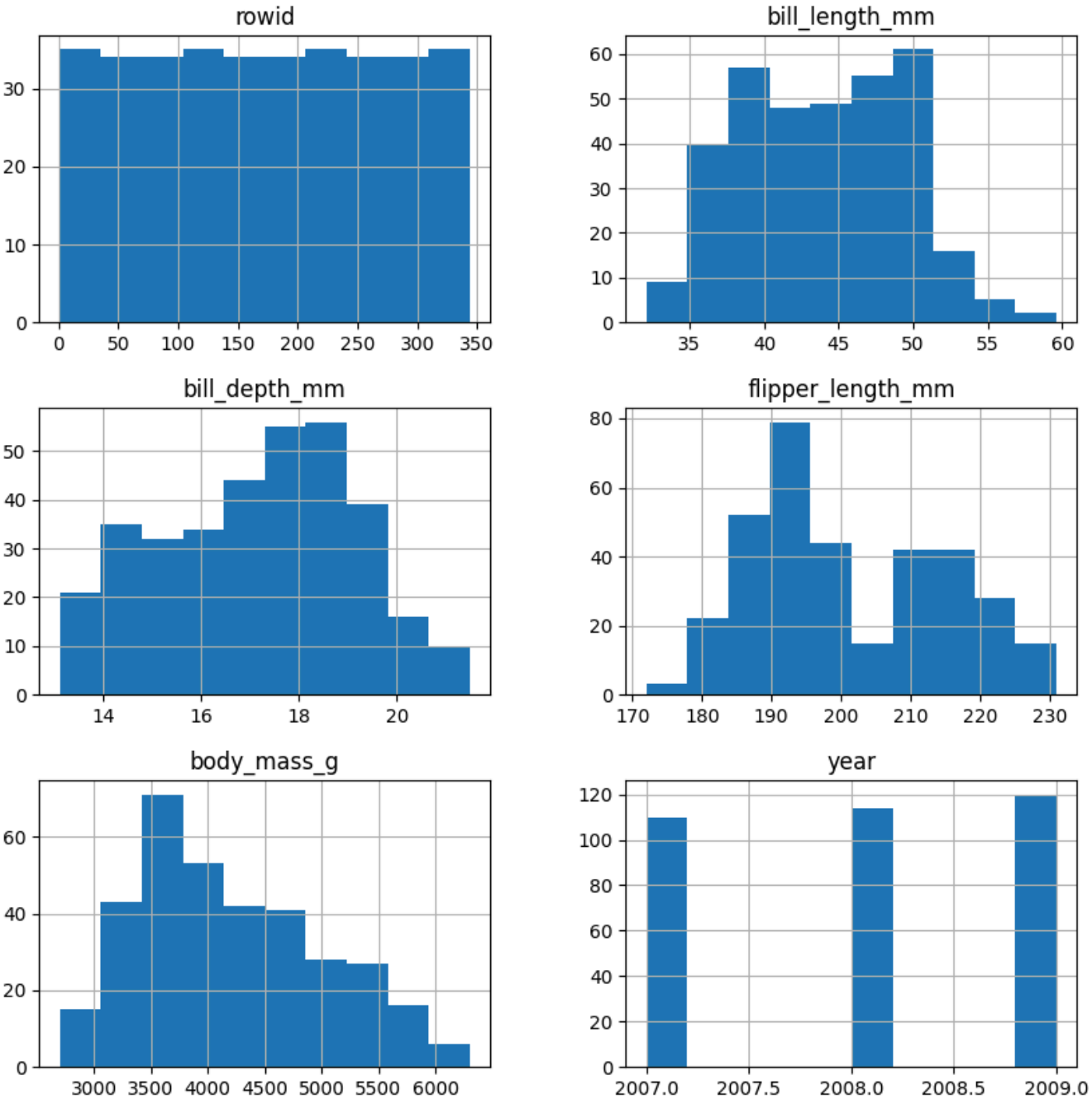
```
# Hay 3 especies diferentes. Hay 152 de Adelie, 124 de Gentoo y 68 de Chinstrap
df["species"].value_counts()
```

species	count
Adelie	152
Gentoo	124
Chinstrap	68

dtype: int64

Dibuja los histogramas para las características numéricas.

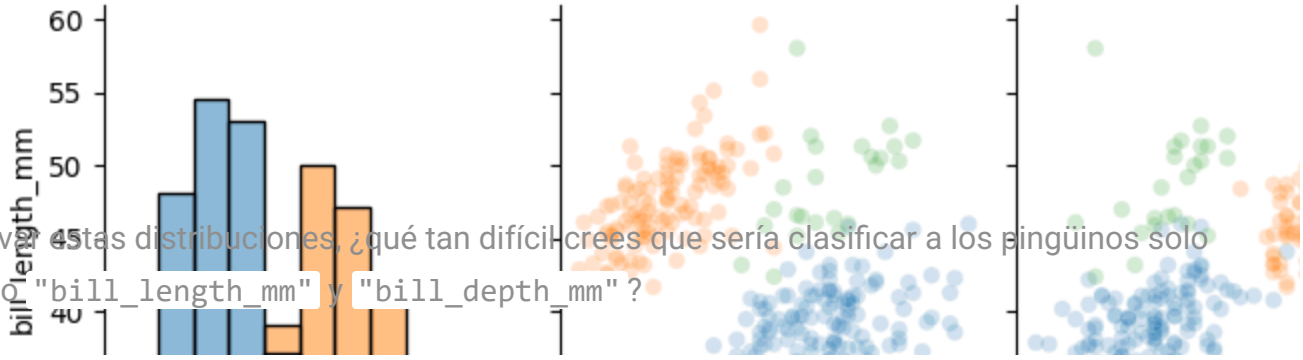
```
_ = df.hist(figsize=(10, 10))
```



Muestra la distribución de las características para cada clase. Sugerencia: utiliza `seaborn.pairplot` .

```
import seaborn as sns

n_samples = 300
columns = ["bill_length_mm",    "bill_depth_mm",    "flipper_length_mm",    "body_mass_g"]
_ = sns.pairplot(data=df[:n_samples],vars=columns,hue="species",plot_kws={"alpha": 0.2},diag
```



Al observar estas distribuciones, ¿qué tan difícil crees que sería clasificar a los pingüinos solo utilizando `"bill_length_mm"` y `"bill_depth_mm"` ?

```
# No muy difícil, los conjuntos están definidos
columns = ["bill_length_mm",  "bill_depth_mm"]
_ = sns.pairplot(data=df[:n_samples],vars=columns,hue="species",plot_kws={"alpha": 0.2},diag_
```