Diplomatura Universitaria en Ciencia de Datos

https://exa.unne.edu.ar/diplomatura/

Módulo 3. Análisis Exploratorio de Datos

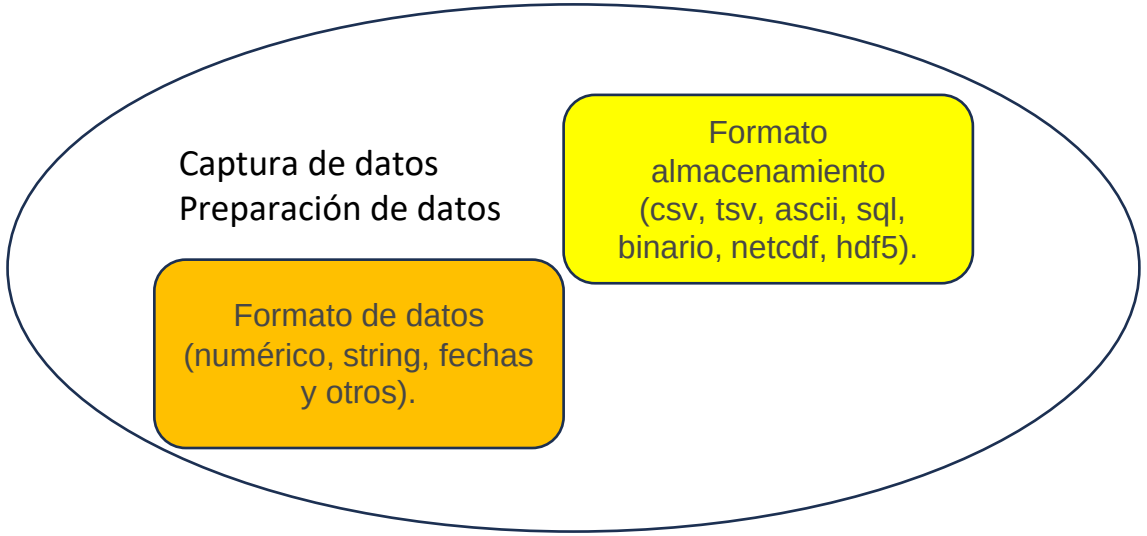
Equipo Docente:

Dra. Sonia I. Mariño

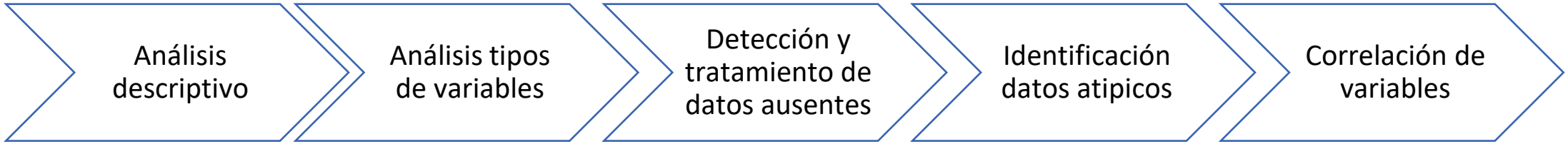
Lic. Lucia del Valle Ledezma

Lic. Rafael Perez

Proceso EDA



ANALISIS EXPLORATORIO DE DATOS



evaluación y corrección de datos.

EDA univariado

EDA bivariado
EDA multivariado

Documentar las decisiones en el proceso

EDA, niveles

Nivel 1 – EDA descriptivo, univariado

- Centrado en el valor de un solo indicador

Nivel 2 – EDA inferencial, bivariado

- Relaciona dos o más variables,
- Estudia una variable en función de otra.

Nivel 3 – EDA modelización, multivariado

- Centrado en los indicadores disponible para estudiar un fenómeno determinado.
- Clasificación cruzada, análisis de varianza y regresiones simples.

EDA, Visualización de datos

- Simplifica la complejidad:
  - existencia de numerosas variables y puntos de datos.
  - presenta la información en un formato fácil de comprender.
- Reconocimiento de patrones:
  - facilidad para identificar patrones y relaciones en los datos,
  - facilidad en la generación y validación de hipótesis.
- Mejora la comunicación:
  - simplificación en transmitir información o hallazgos
  - una imagen dice más que mil palabras
- Detección de anomalías:
  - detección de valores atípicos, valores nulos, datos inusuales
- Eficiencia del tiempo:
  - representación o visión general,
  - «conocer» los datos, previo al procesamiento

EDA, Herramientas

Funciones y técnicas estadísticas específicas para realizar con las herramientas EDA se :

- Según representaciones gráficas
  - No gráficas estadísticas de resumen
  - Gráficas, distintas representaciones visuales
- Según número de variables
  - Univariante, de cada variable del conjunto de datos sin procesar, con estadísticas de resumen.
  - Bivariante, evaluar la relación entre cada variable del conjunto de datos y la variable de destino.
  - Multivariante, mapeo y comprensión de las interacciones entre diferentes variables de los datos
- Técnicas de agrupamiento y reducción de dimensiones, visualizaciones gráficas de datos de alta dimensión con muchas variables.
  - K-means
  - PCA
  - Otros

Representaciones gráficas.

Histogramas:

representan la distribución de una variable en un conjunto de datos.

Diagramas de sectores:

representan proporciones o porcentajes.

Diagramas de caja y bigotes:

representan la distribución y los valores atípicos en un conjunto de datos.

Diagramas de barras:

representan datos en forma de barras verticales u horizontales.

Diagramas de violín:

combinan histogramas y diagramas de caja para mostrar la distribución y densidad de datos.

Diagramas de dispersión o puntos:

representan la relación entre dos variables, muestra cuánto afecta una variable a otra.

Diagramas de líneas:

muestran tendencias a lo largo del tiempo o en secuencias.

Diagramas de áreas:

resaltan la acumulación de valores a lo largo de un eje.

Mapa de calor,

representación gráfica de datos, los valores se representan por color.

EDA, univariado

Medidas

- De tendencia central
- De variabilidad
- De distribución
- De posición

Representaciones

- Tablas
- Gráficos:
  - Histograma.
  - Diagrama de Cajas (Boxplot).

```
df.describe()
df.describe().T
```

```
# Resumen estadístico del conjunto de datos
df.describe().T
```

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

```
Número de filas del dataset 344
Número de columnas  del dataset 7
```

```
[49]
# datos balanceados en especies ?
### Ejemplo de variable categórica
## df.species.value_counts()
print ("Número de filas por variable categorica", df.species.value_counts())
```

```
Número de filas por variable categorica species
Adelie       152
Gentoo       124
Chinstrap     68
Name: count, dtype: int64
```

```
## datos balanceados sexo de especies ?
### Ejemplo de variable categórica
## df.sex.value_counts()
print ("Número de filas por variable categorica", df.sex.value_counts())
```

```
Número de filas por variable categorica sex
Male      168
Female    165
Name: count, dtype: int64
```

EDA univariado - Visualización

Histogramas

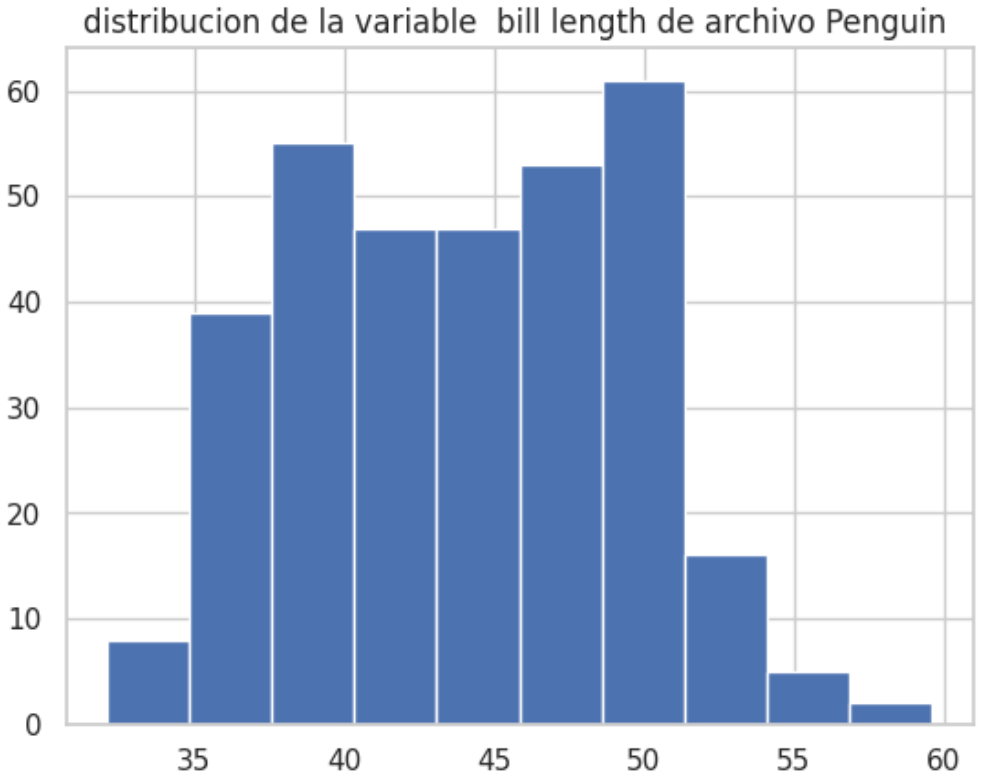
- Utilizado en variables discretas y continuas.
- visualiza distribuciones de datos
- Informan sobre
  - la tendencia central,
  - la variabilidad y
  - la asimetría de los datos.
- Organiza los datos :
  - Eje x: diferentes subgrupos (o bins) de ancho proporcional a la amplitud del intervalo
  - Eje y: altura proporcional a la frecuencia o cantidad de apariciones de los valores del intervalo en un conjunto de datos.

EDA bivariado – Visualización

Histograma

- Representa la distribución de frecuencias de los valores de una variable numérica.
- Divide los datos en grupos de rangos de valores o bins
- Bin, tiene una altura proporcional al valor de su recuento
- Desventaja, no muestra valores atípicos

```
# Creating a histogram
plt.hist(df['bill_length_mm'])
plt.title('distribucion de la variable
bill length de archivo Penguin ')
plt.show()
```



EDA univariado - Visualización

Gráfico de barras

```
# Gráfico de barras
plt.figure(figsize=(10, 6))
sns.countplot(x='species', data=df)
plt.title('Cantidad de Pingüinos por Especie')
plt.xlabel('Especie')
plt.ylabel('Cantidad')
plt.show()
```

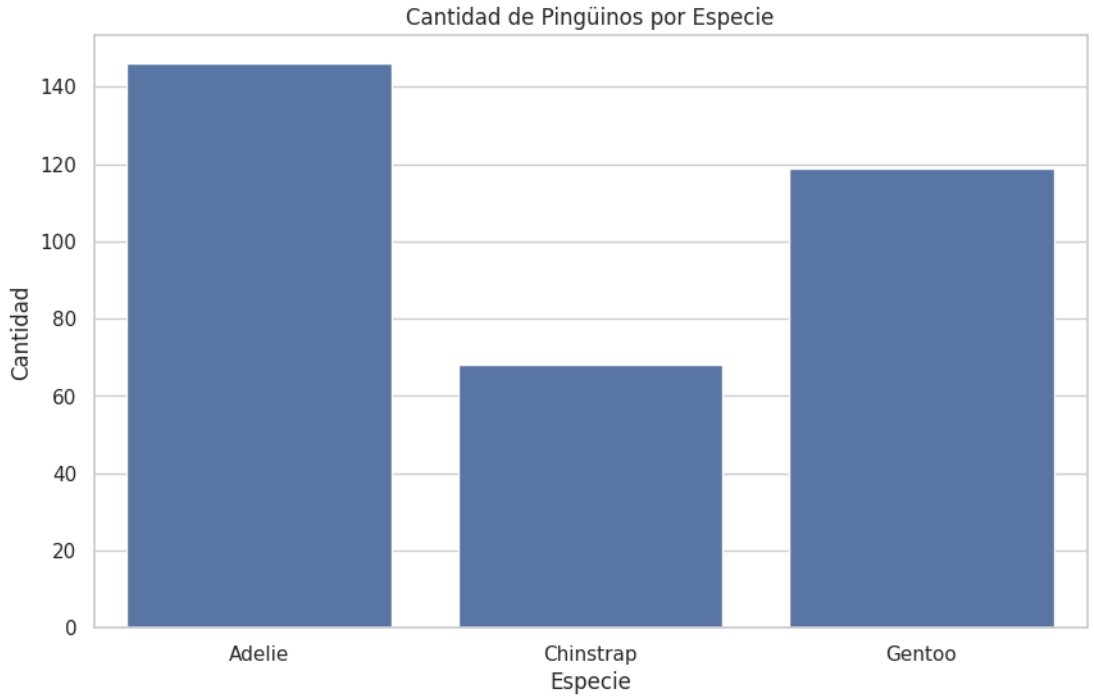
Variables categóricas y numéricas discretas

- variables numéricas discretas con pocos valores (número de hijos, otros).

representación visual clara y precisa.

tantas barras como categorías tiene la variable,

la altura de cada barra es proporcional a la frecuencia o porcentaje de casos en cada clase.



similar a los gráficos de sectores.

EDA BIVARIADO / MULTIVARIADO

- comparar categorías diferentes entre sí.

Balanceo de clases

EDA univariado - Visualización

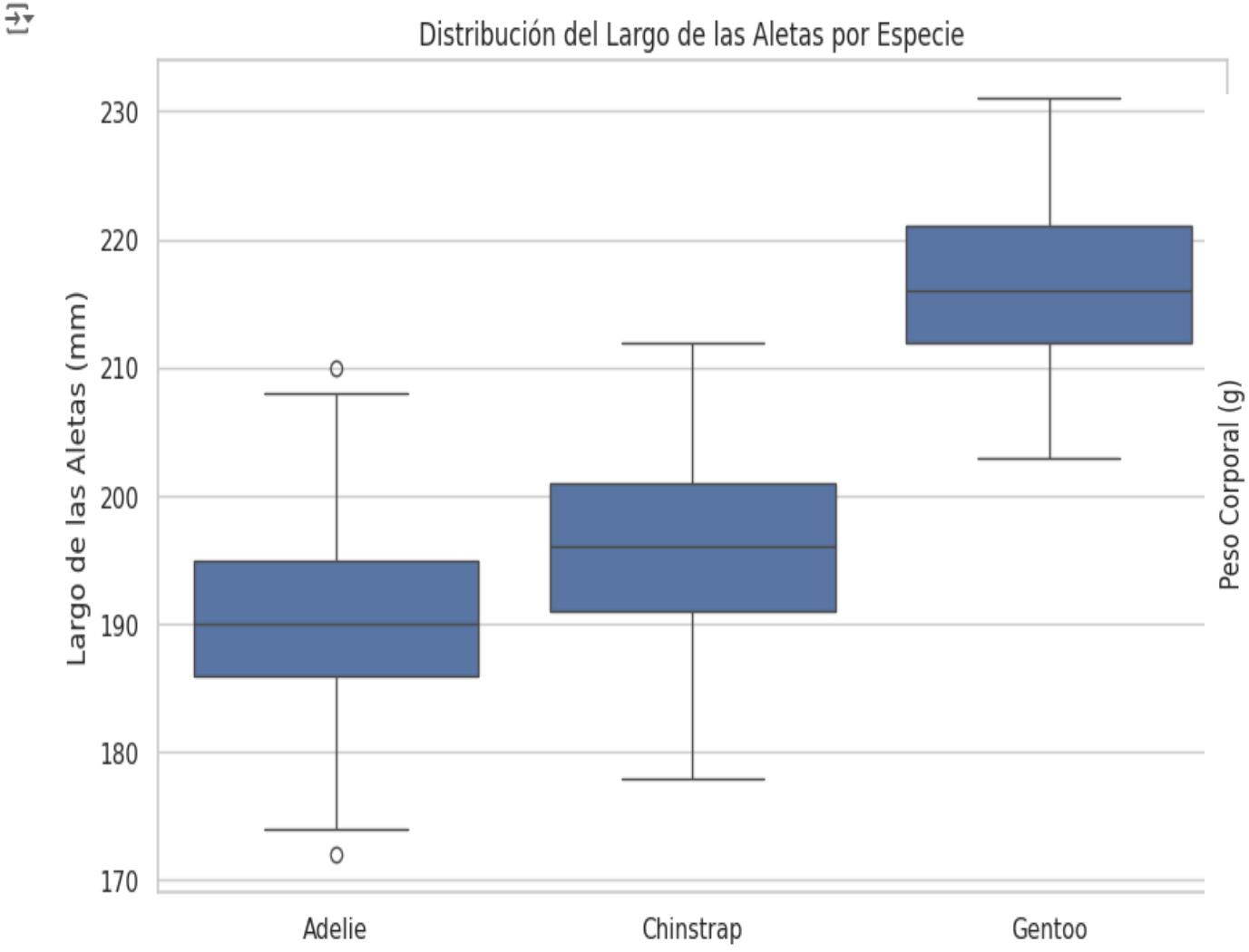
Boxplots

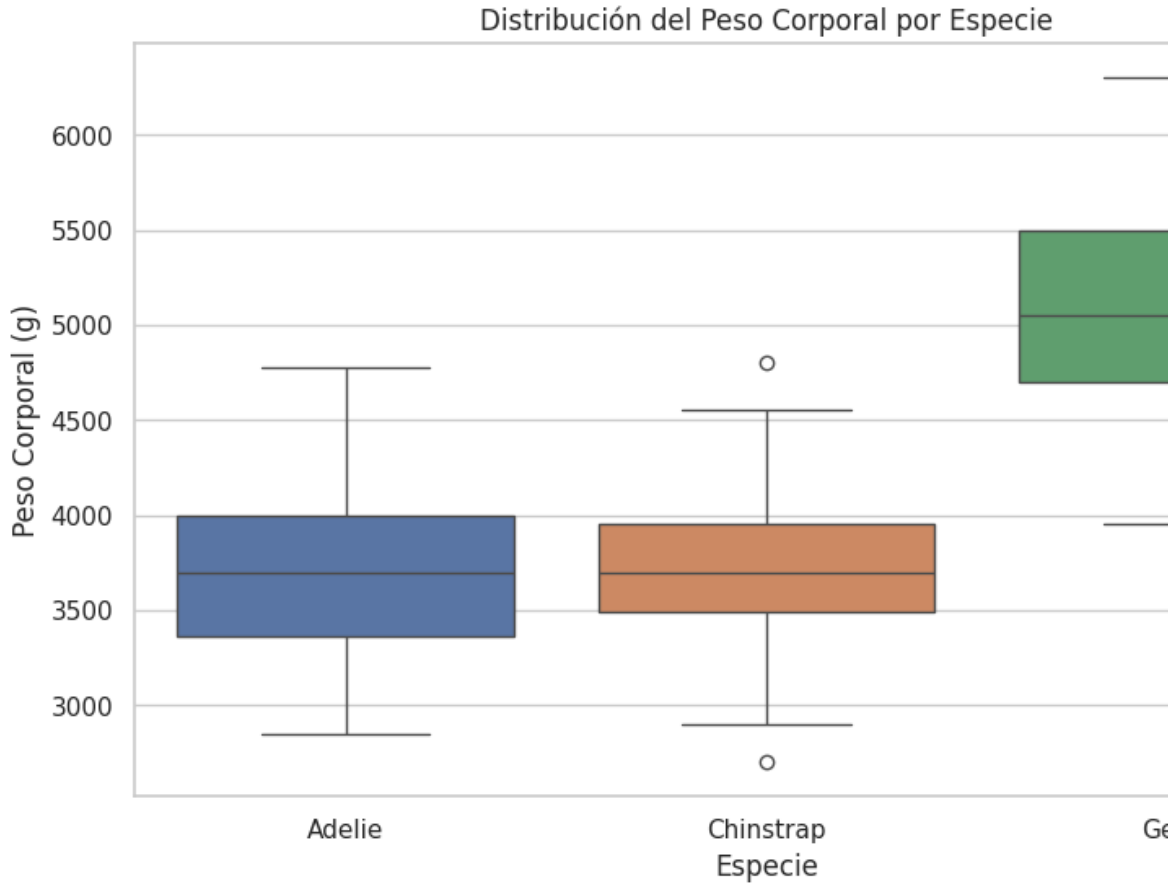
- variables discretas y continuas.
- representa la distribución de los datos,
- permite detectar valores outliers o atípicos y comprender las tendencias centrales.

Contiene:

- los percentiles: la barras superior e inferior corresponden a los percentiles 75 y 25,
- la mediana, línea en medio de la caja
- los bigotes –whiskers- o líneas fuera de la caja, equivale al percentil 75 o 25 por 1.5
- rango intercuartil, diferencia entre el primer cuartil (Q1) y el tercer cuartil (Q3) de un conjunto de datos.
- valores mínimos (LI) y máximo (LS)
- valores **atípicos**, son aquellos menores a Min (LI) y mayores a Max (LS).

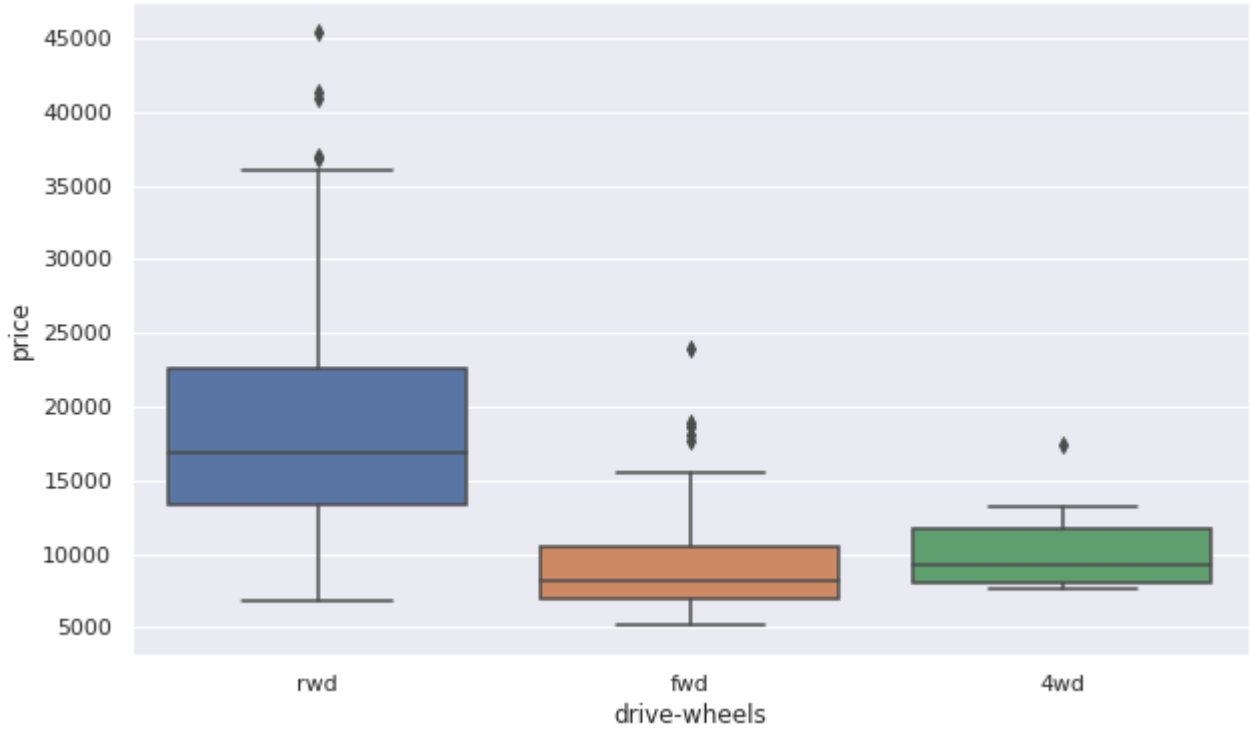
```python
# Box plot de la variable 'flipper_length_mm' por especie
plt.figure(figsize=(10, 6))
sns.boxplot(x='species', y='flipper_length_mm', data=df)
plt.title('Distribución del Largo de las Aletas por Especie')
plt.xlabel('Especie')
plt.ylabel('Largo de las Aletas (mm)')
plt.show()
```





EDA - Visualización

```
sns.boxplot(x="drive-wheels", y="price", data=df)
```



EDA bivariado – Visualización

Variables categóricas y numéricas discretas

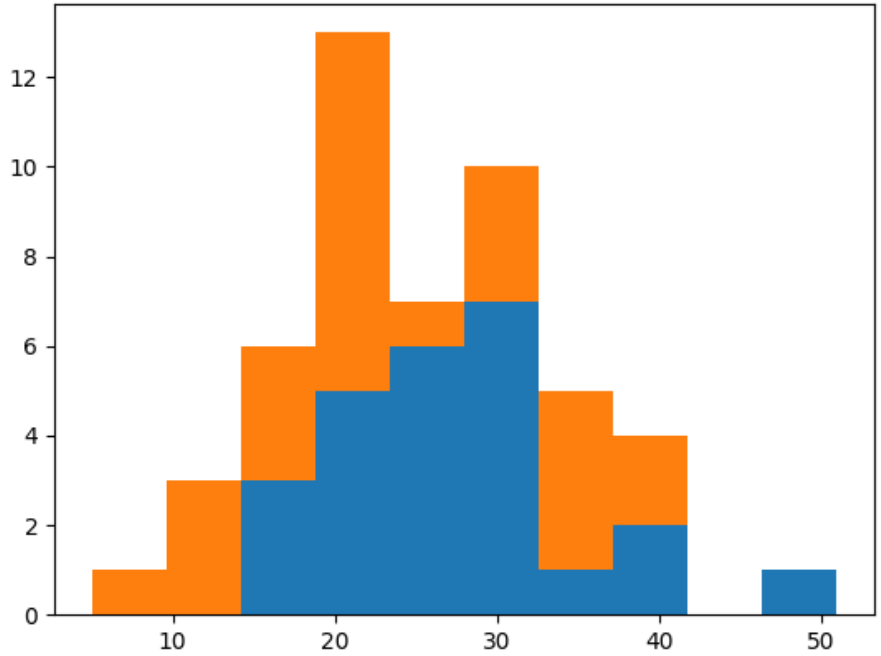
EDA BIVARIADO / MULTIVARIADO

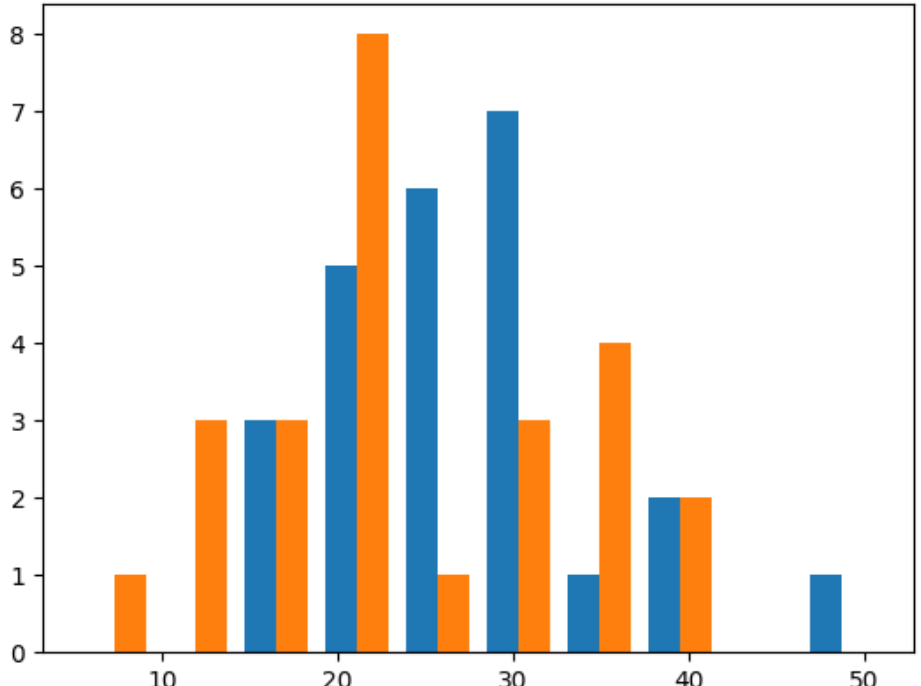
- comparar categorías diferentes entre sí.

  representa las cantidades en ejes verticales y horizontales,

  La altura de cada barra es proporcional a la frecuencia o porcentaje de casos en cada clase.

```
import matplotlib.pyplot as plt
valoresA =
[23,22,28,32,24,28,32,15,26,22,24,24,26,28,32,41,20,39,51,18,23,
28,26,34,17]
valoresB =
[23,20,30,15,10,25,30,36,20,21,20,23,34,15,14,38,34,17,38,5,34,20,21,
30,10]
fig, ax = plt.subplots()
plt.hist([valoresA,valoresB])
plt.hist([valoresA,valoresB], stacked=True)
plt.show()
plt.savefig(«compara.png")
```



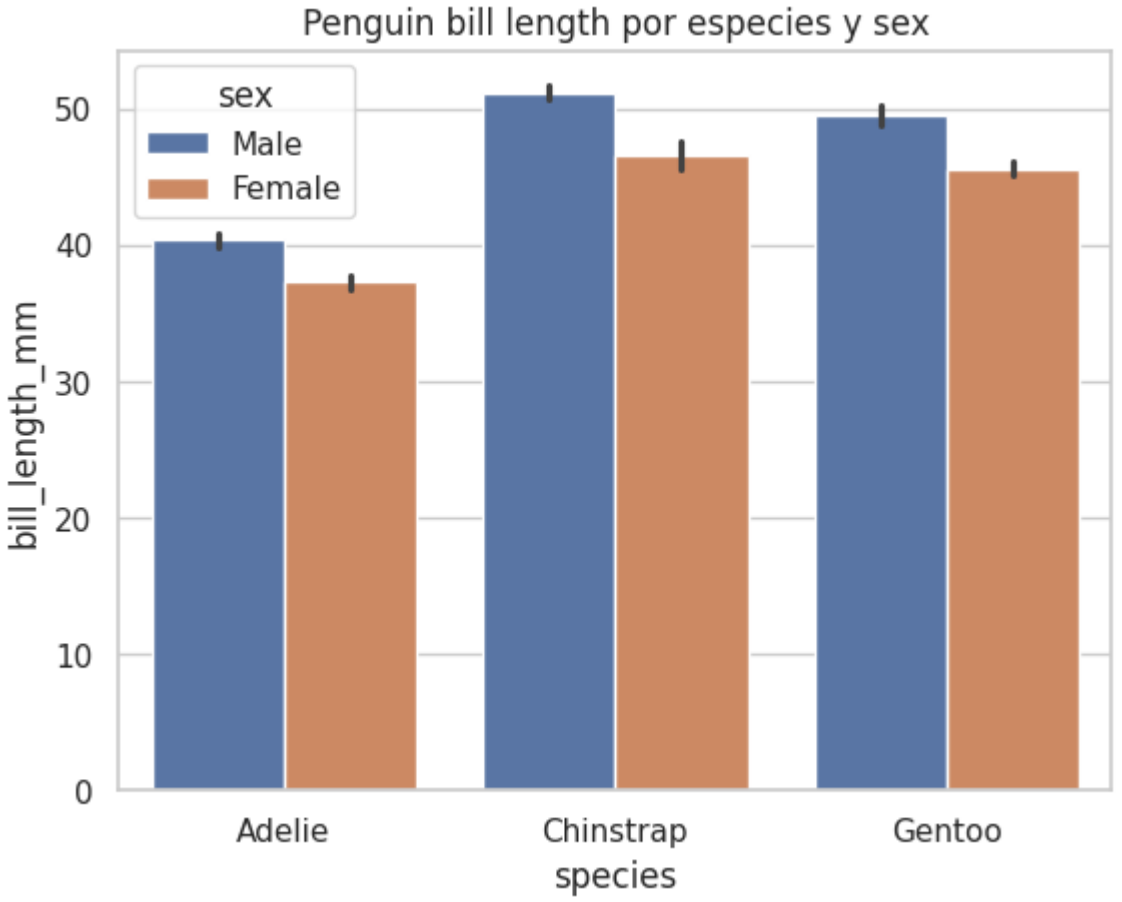


EDA bivariado – Visualización

Gráfico de barras agrupadas

Representa gráficamente los valores numéricos de datos categóricos
Compara categorías entre sí
Categorías, se representan mediante barras rectangulares de la misma anchura
Las alturas o longitudes, son proporcionales a los valores numéricos a los que corresponden.

```
sns.barplot(df, x='species',
y='bill_length_mm', hue='sex')
plt.title('Penguin bill length por especie y
sexo')
plt.show()
```



EDA bivariado - Diagrama de dispersión

Representa la relación entre dos variables.

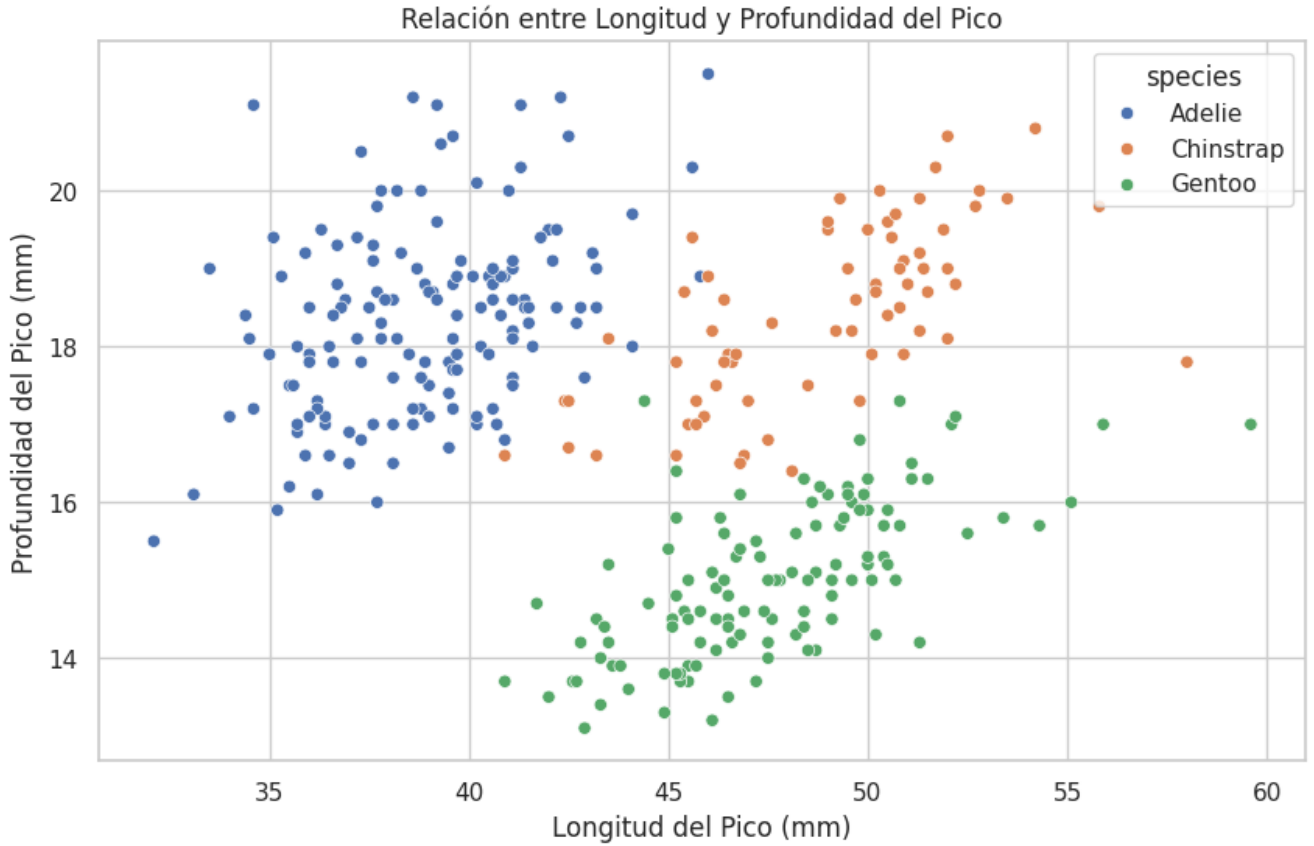
Los puntos se dispersan en un plano cartesiano, y se identifican posibles correlaciones entre las variables.

Los ejes representan las variables que se comparan

cada punto trazado en el gráfico corresponde a una observación de datos específica.

Escala del eje permite interpretar con precisión los datos;

Línea de tendencia, representa la dirección general de los puntos de datos, y puede indicar la fuerza de la relación entre las variables



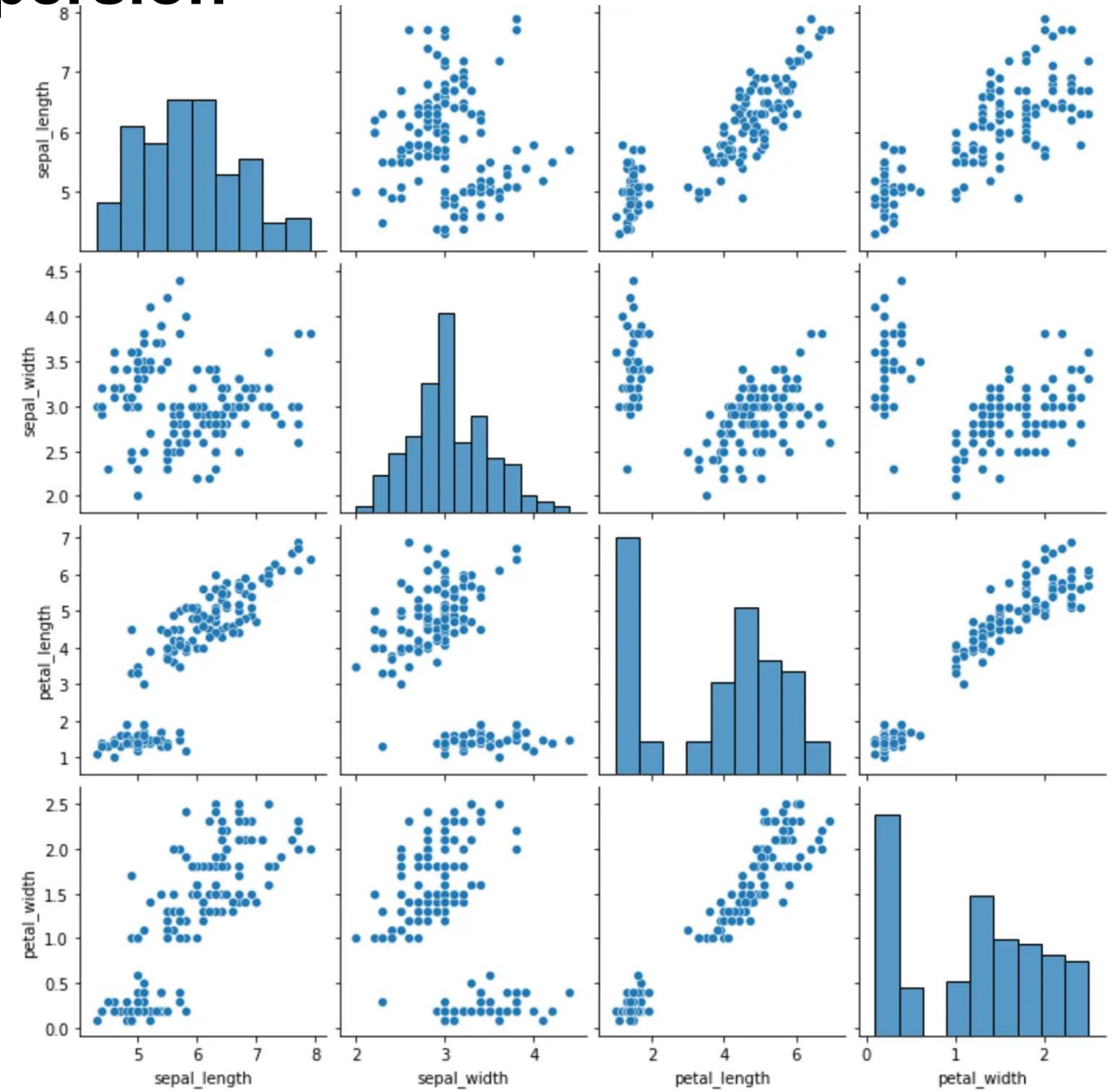

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10, 6))
  sns.scatterplot(x='bill_length_mm', y='bill_depth_mm',
hue='species', data=df)
  plt.title('Relación entre Longitud y Profundidad del Pico')
  plt.xlabel('Longitud del Pico (mm)')
  plt.ylabel('Profundidad del Pico (mm)')
  plt.show()
```

EDA bivariado - Diagrama Dispersión

Muestra la relación entre dos variables.

Los puntos se dispersan en un plano cartesiano, permitiendo identificar posibles correlaciones entre las variables.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar el conjunto de datos
"iris"
df1 = sns.load_dataset("iris")

# Crear gráfico de pares para todas
las variables numéricas
sns.pairplot(df1.dropna())
sns_plot = sns.pairplot(df1)
sns_plot.savefig('cross_plots_iris.
png')
```



```
sns.pairplot(df1.dropna())
sns_plot = sns.pairplot(df1,
hue='species')

sns_plot.savefig('cross_plots_iris.
png')
```

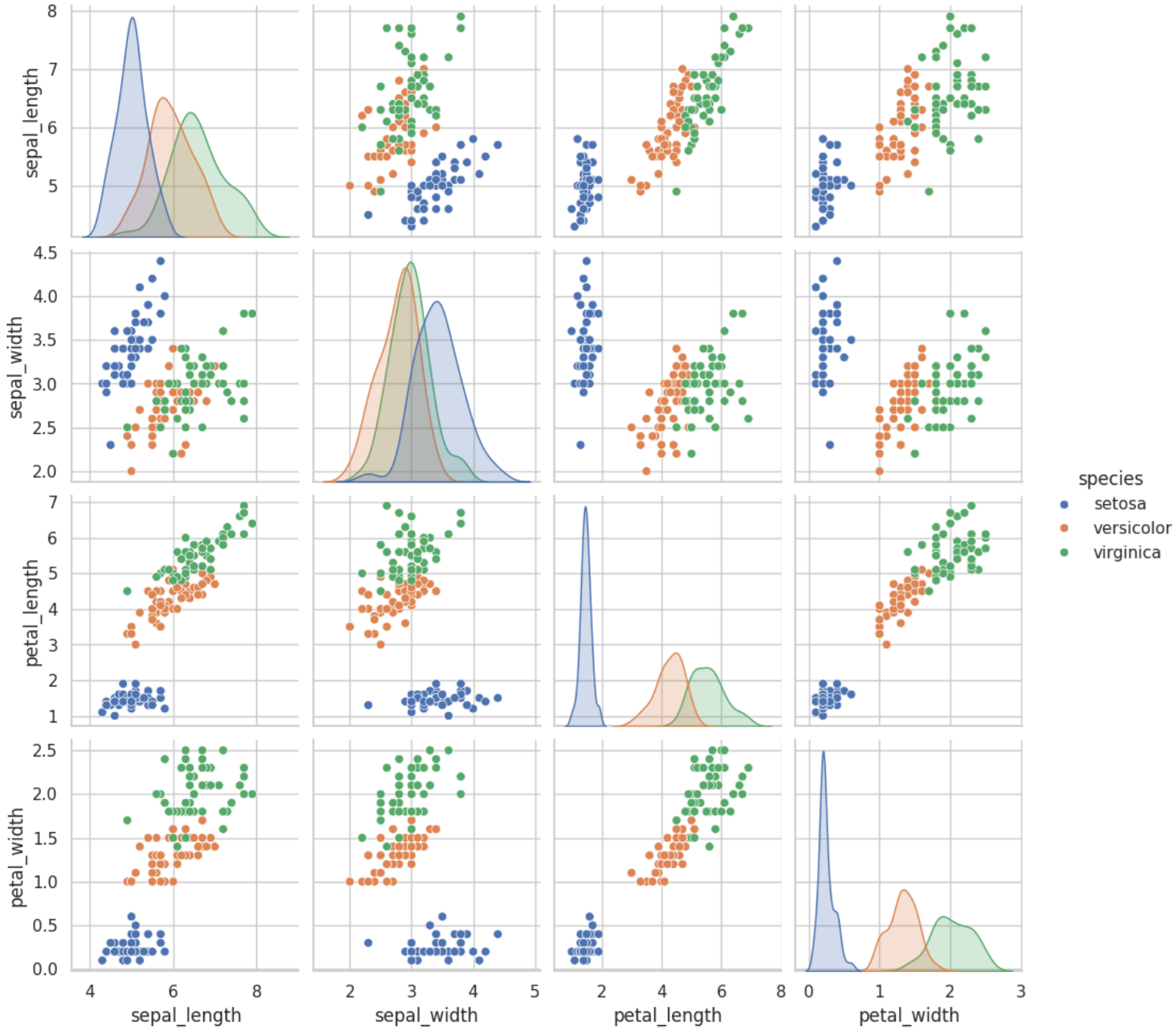
Pairplot

representa la relación entre todas las combinaciones de
variables

ayuda a identificar las relaciones

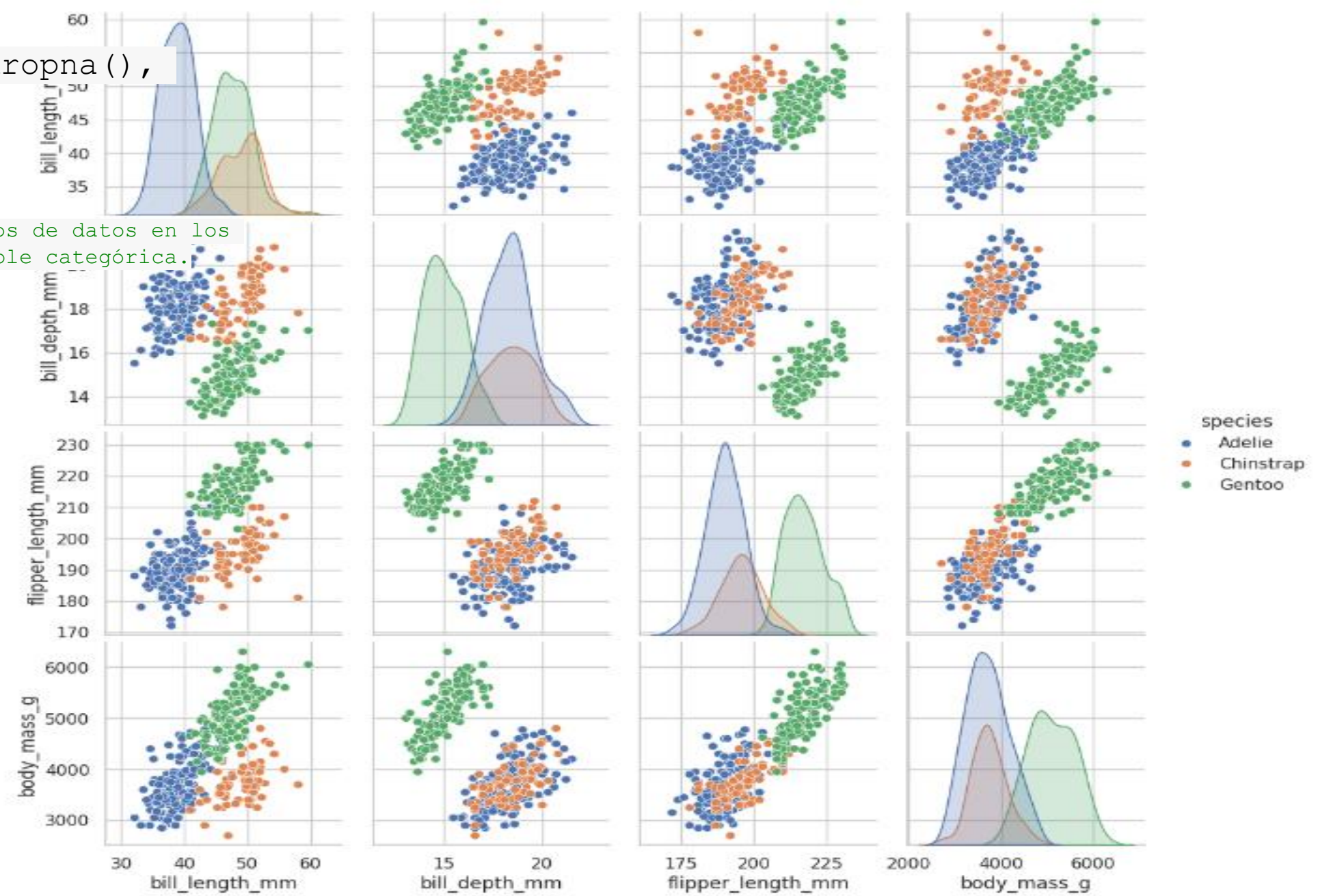
entre las diferentes variables y su

impacto en el resultado.



```
sns.pairplot(df.dropna(), hue='species')
plt.show()
```

```
#hue diferencia los puntos de datos en los gráficos según una variable categórica.
```



Correlación

Correlación entre variables

herramienta valiosa en EDA, para análisis estadístico

facilita identificar patrones y relaciones entre variables.

Notas:

- Distinguir correlación y causalidad
    - una variable NO necesariamente causa el cambio en la otra.
- Verificar la linealidad:
    - La correlación de Pearson sólo mide relaciones lineales.
    - Otras medidas si las variables tienen una relación no lineal (Ej. Sperman).
- Primero se debe aplicar técnicas de limpieza y transformación de los datos.

Correlación de variables

EDA – Análisis bivariado

- gráfico de pares de variables

EDA – Análisis multivariado

- visualizaciones, mapea relaciones entre variables,
- permite comprender las interacciones entre diferentes variables.

Correlación entre variables

Medida de correlación, depende del tipo de variables y la distribución de los datos.

Coeficiente de Pearson:

- medida de la relación lineal entre dos variables continuas, que siguen distribución normal.
- coeficiente de correlación más comúnmente utilizado. Varía entre -1 y 1

Coeficiente de Spearman:

- medida de la relación no lineal entre dos variables.
- mide la relación entre las posiciones relativas de los datos en lugar de los valores de las variables.
- adecuado para datos ordinales o continuos no normales. Varía entre -1 y 1

Coeficiente de Kendall:

- mide la relación ordinal entre dos variables.
- similar a coeficiente de Spearman, específico para variables ordinales. Varía entre -1 y 1

Correlación entre variables

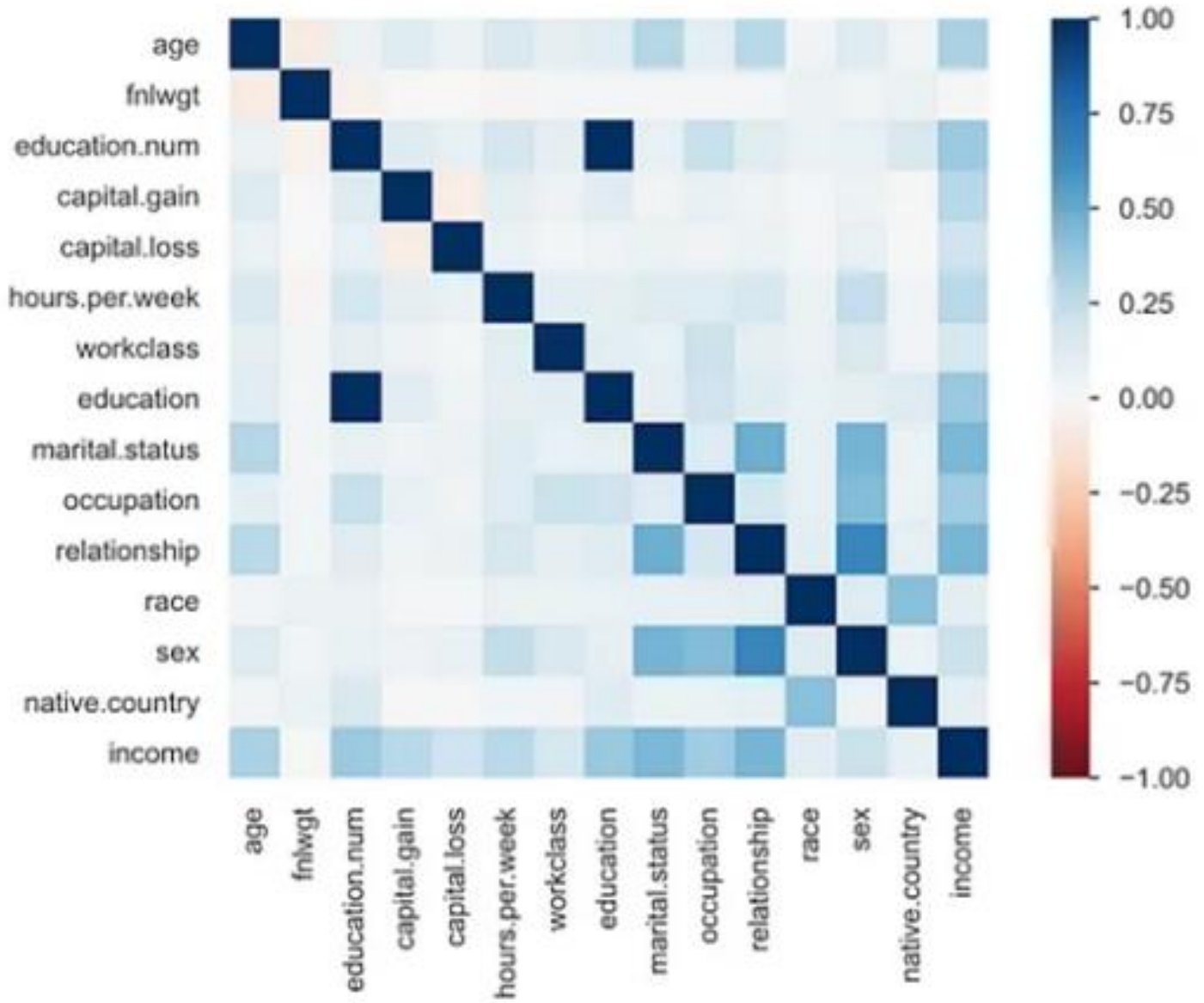
correlación
medida estadística
indica la relación entre dos o más variables. Valores entre 1 a -1

correlación positiva
ambas variables aumentan
valor coeficiente cercano a 1

correlación negativa
una variable aumenta y otra variable disminuye.
valor coeficiente cercano a -1

correlación neutra o nula
valor coeficiente cercano a 0

correlación no implica causalidad



Correlación entre v

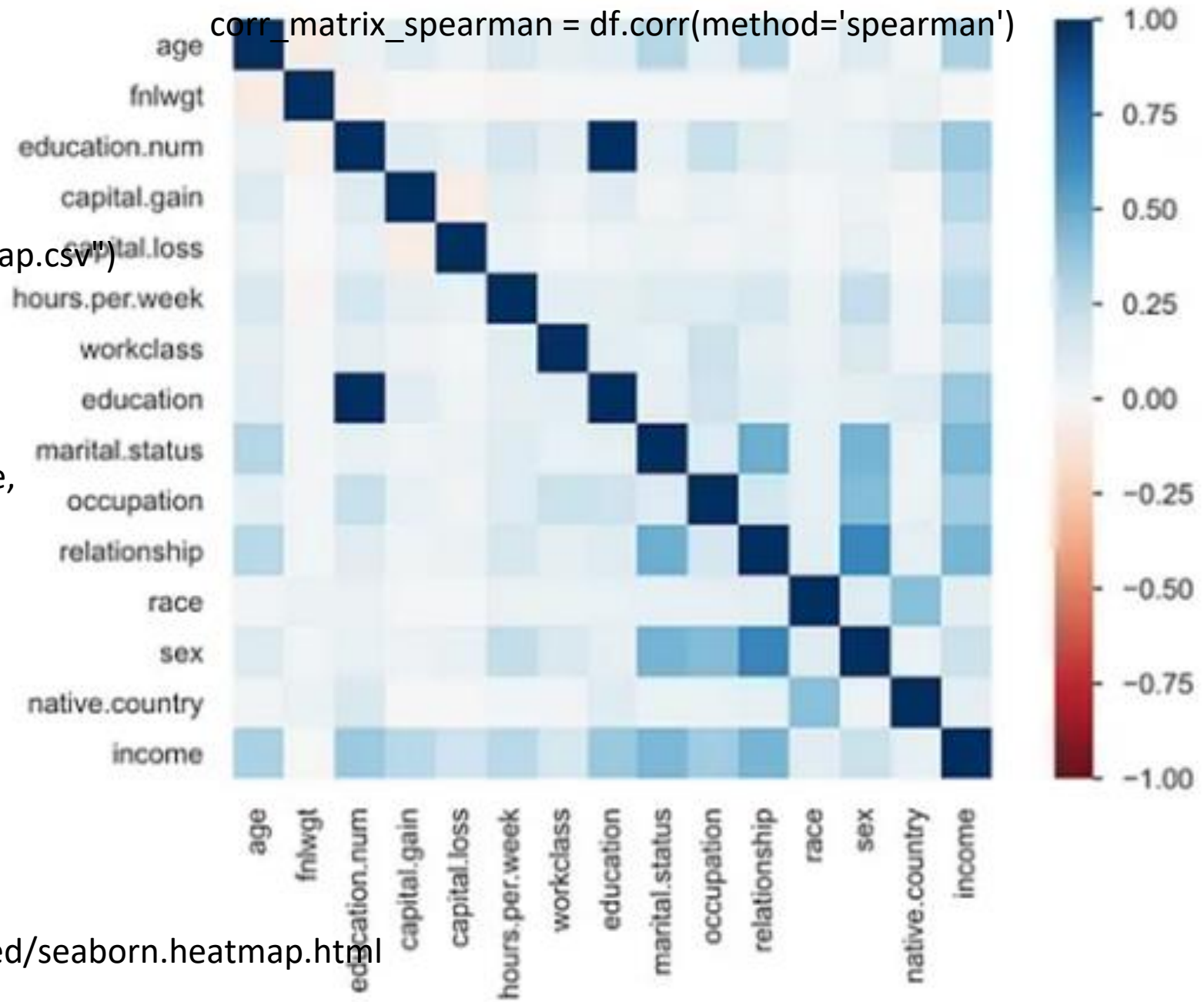
```
import pandas as pd
df =
pd.read_csv("/content/DatasetHeatMap.csv")

#Correlación lineal
corr_matrix = df.corr() print(corr)

sns.heatmap(corr_matrix , annot=True,
cmap='YlGnBu', vmax=1,vmin=-1)
plt.title('Gráfico de Calor')
plt.show()
plt.savefig (matriz.png)
```

mapa de calor permite determinar :
variables mayor poder predicción:
estado civil o relación

variables que carecen de impacto :
fnlwgt

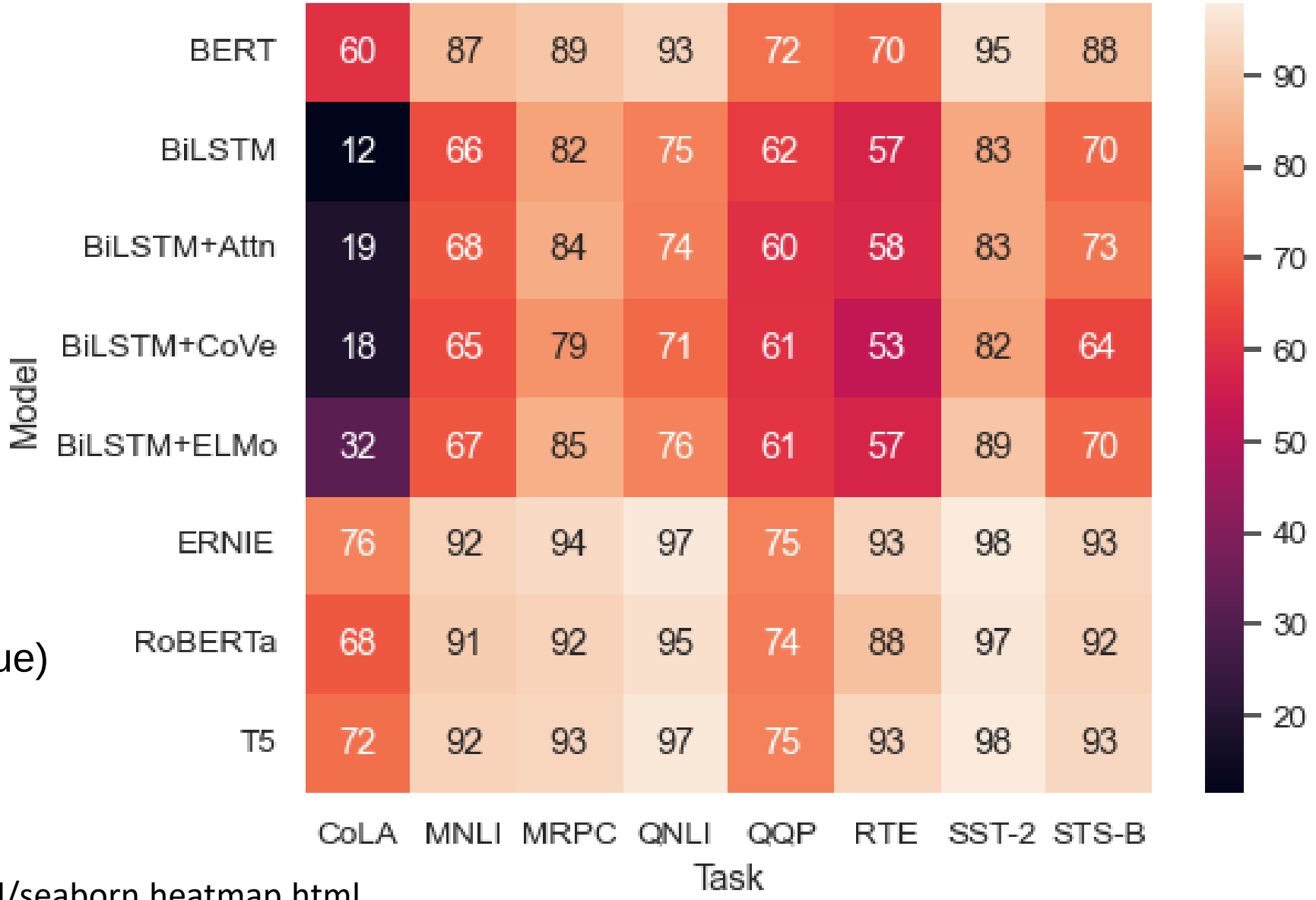
https://seaborn.pydata.org/generated/seaborn.heatmap.html

```
corr_matrix_kendall = df.corr(method='kendall')
corr_matrix_spearman = df.corr(method='spearman')
```



Correlación entre variables

```
glue =
sns.load_dataset("glue").pivot(index
="Model", columns="Task",
values="Score")
sns.heatmap(glue)
```



Incluir en las celdas, el valor

```
sns.heatmap(glue, annot=True)
```

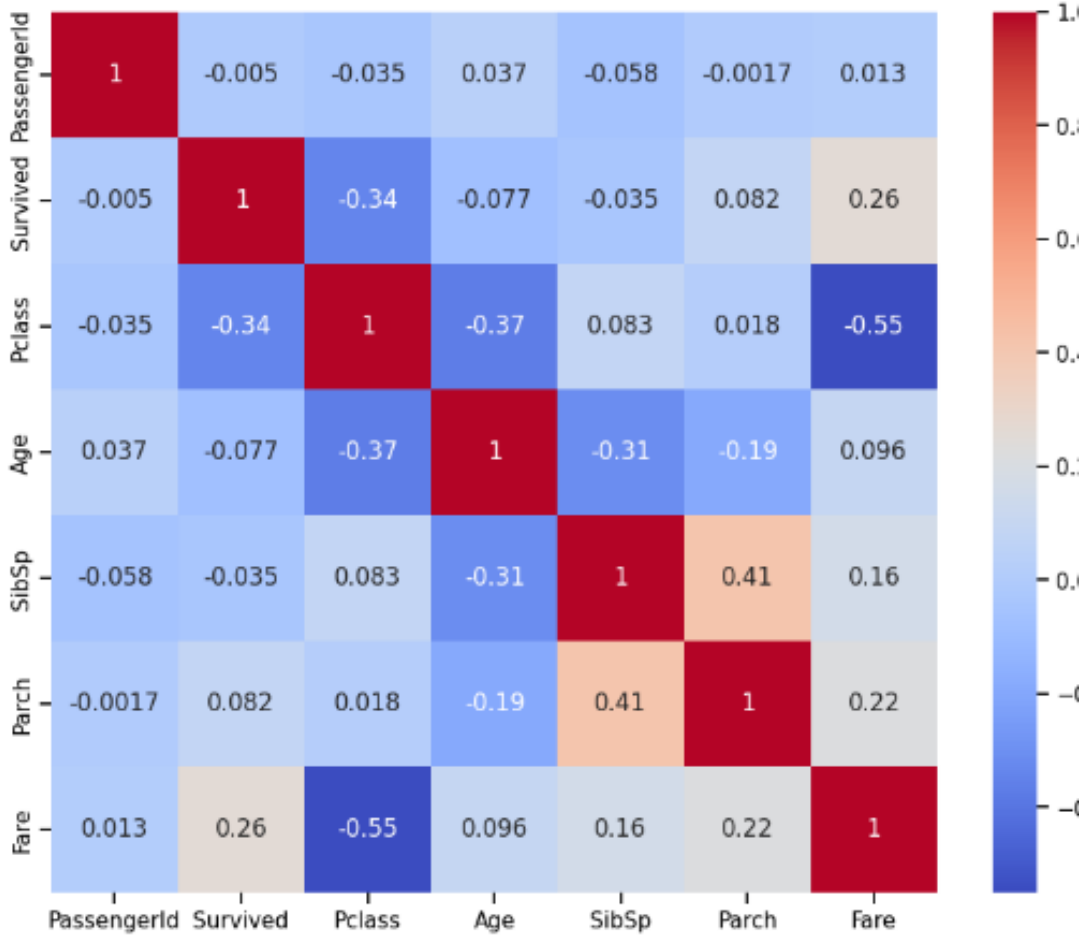
https://seaborn.pydata.org/generated/seaborn.heatmap.html

EDA multivariado - Visualización

Mapas de Calor (Heatmaps)

Visualiza la correlación entre características numérica. Facilita descubrir dependencias en los datos.

```
import seaborn as sns
import matplotlib.pyplot as plt
correlation_matrix = data.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True,
cmap="coolwarm")
```



Otro ejemplo

	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

correlación entre “horas dedicadas a estudiar” y “puntaje del examen” = **0.82** , indica
Correlación fuertemente positiva. + horas dedicadas a estudiar está fuertemente asociado con + puntajes en exámenes.

	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

correlación entre «horas dedicadas a estudiar» y «horas dedicadas a dormir» = **-0,22** ,
Correlación débilmente negativa. +horas dedicadas al estudio se asocia con - horas dedicadas a dormir.

la correlación entre «horas dedicadas a dormir» y «puntaje de CI» = **0.06** ,

Poca correlación o asociación entre cantidad de horas que duerme y su puntaje I.

	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00