IMT2200

INTRODUCCIÓN A CIENCIA DE DATOS

2022-2

ANÁLISIS EXPLORATORIO DE DATOS

CLASE 16

ANÁLISIS EXPLORATORIO DE DATOS

EDA	Estadístico	Gráfico
Una variable	Media Mediana Desviación estándar Varianza Percentiles Rango intercuartil (IQR=Q3-Q1) Distribución de probabilidad	Histograma FDA KDE Boxplot Choroplet
Numérica		
Categórica	Cuentas por valor (value counts)	~~Gráfico de torta (pie chart)~~
Multi-variable	Coef. Pearson Matriz de correlación Regresión Agrupación (groupby, pivot)	Scatterplot Jointplot Pairplots Histograma múltiple Serie de tiempo Stacked area Heatmap Pairplots... etc
	`pandas` **`numpy`**	**`scipy`** **`sklearn`**	**`matplotlib, seaborn`** **`geopandas`**

09:27

Estadísticas de Resumen

- **Media**: es la suma de todos los valores, dividida por el número de puntos. $$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Mediana:** es el valor medio de un conjunto de datos. Es inmune a valores extremos o outliers. Para calcularla, se ordenan los datos y se elige el valor que queda en la mitad.
- **Percentiles:** el percentil p, corresponde al valor que es mayor al p % de los datos.

- **Varianza:** promedio de la distancia cuadrática de los datos a la media. Es una medida de la dispersión de los datos.

$$\text{Var} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Desviación Estándar:** es la raíz cuadrada de la varianza. Está en la misma escala de unidades que los datos.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Estadísticas de Resumen

- **Covarianza:** es una medida de cómo dos cantidades varían juntas. Es la media del producto entre las diferencias de los valores respecto a la media.

$$\mathrm{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

  - Si x e y tienden a estar ambas arriba, o ambas abajo de la media, la covarianza es positiva.
  - Esto quiere decir que hay una correlación positiva: cuando x es alta, y es alta.
  - Por el contrario, si x es alta cuando y es baja, la covarianza es negativa y los datos están anticorrelacionados.

Estadísticas de Resumen

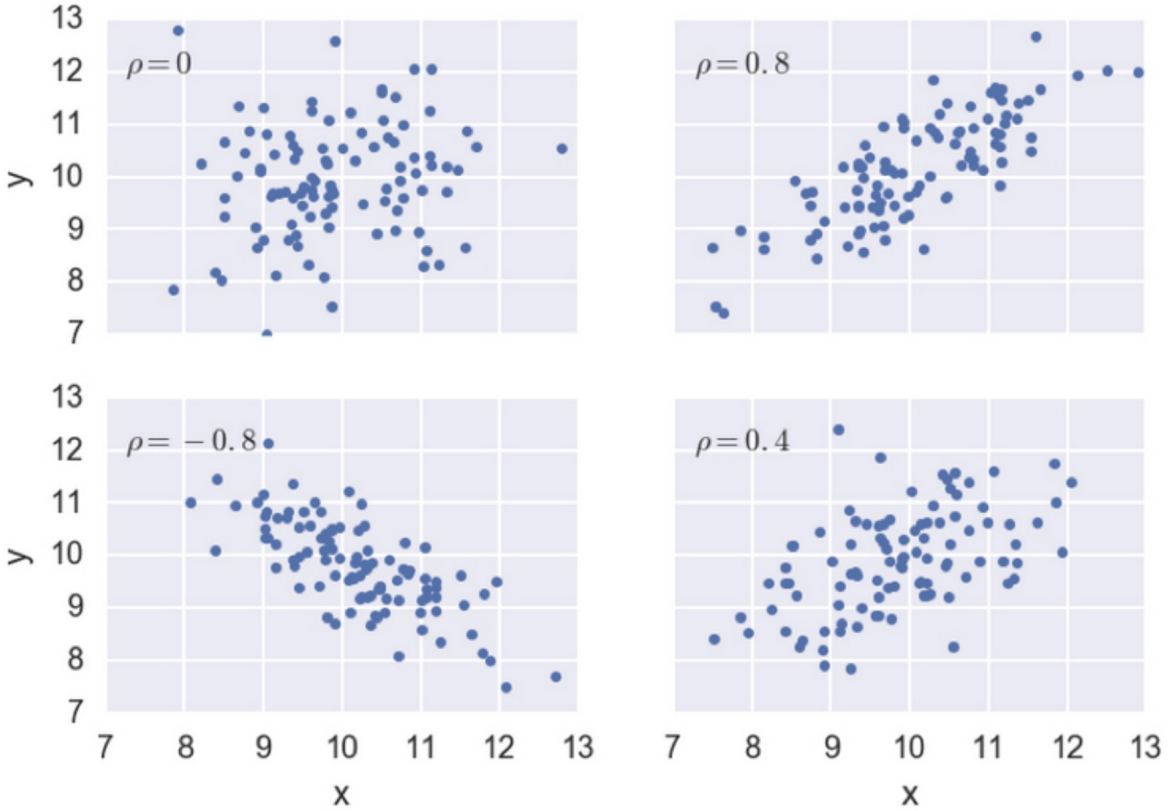
Coeficiente de Pearson (ρ): para tener una medida más general y aplicable de la correlación entre dos variables, necesitamos que sea adimensional. Por lo tanto dividimos la covarianza por las desviaciones estándar de x e y.

$$\rho = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

- Es la comparación de la variabilidad en los datos debido a una codependencia (covarianza), con la variabilidad inherente de cada variable (sus desviaciones estándar).

- Un valor **0** indica que **no hay correlación**, valor **-1/1** indica **alta correlación (negativa/positiva).**

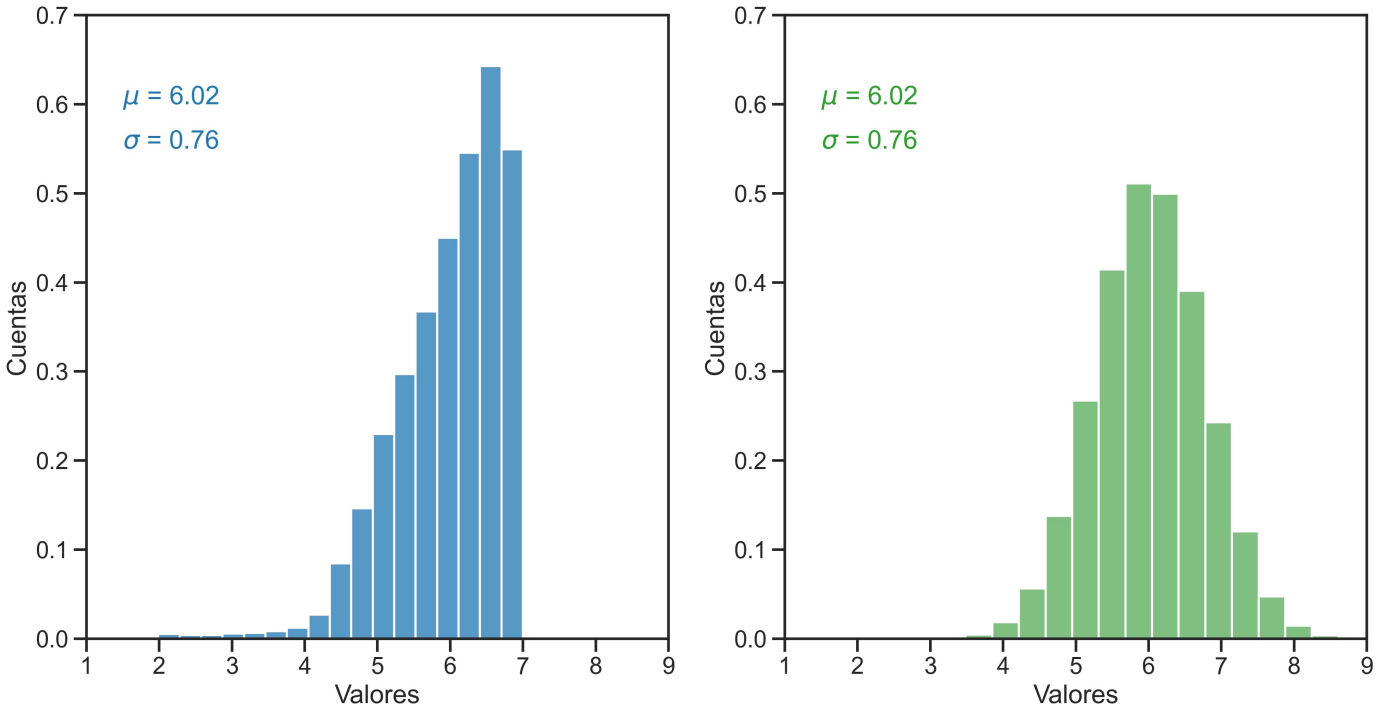
Estadísticas de Resumen



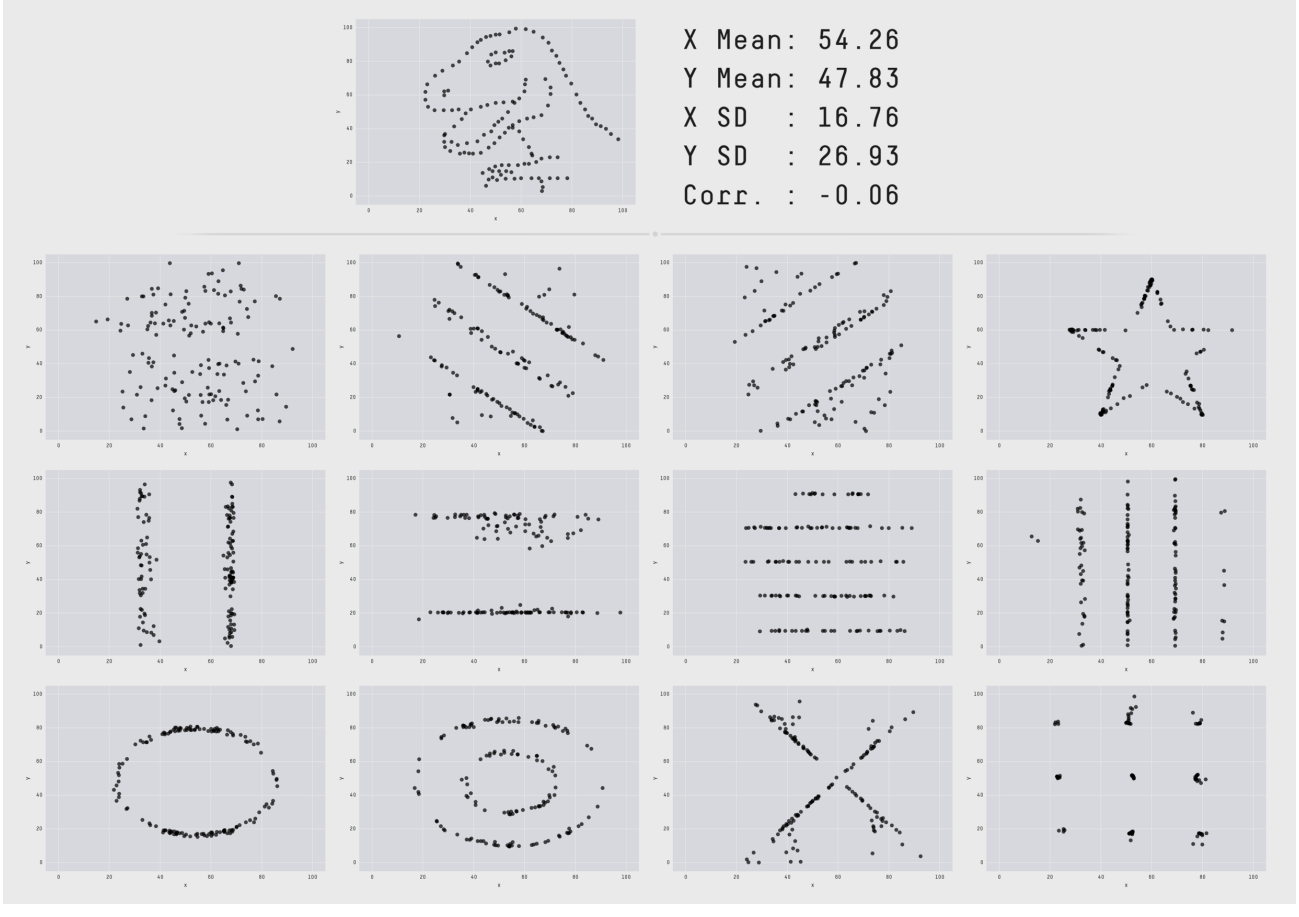
$$\rho = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

09:27

Distintas distribuciones pueden tener las mismas estadísticas de resumen



Distintas distribuciones pueden tener las mismas estadísticas de resumen



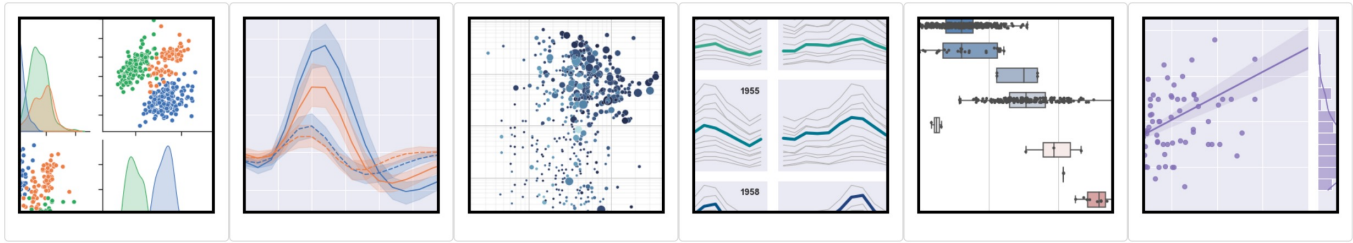
Además de analizar las estadísticas de resumen, es necesario GRAFICAR los datos para entender mejor su distribución.

Análisis Gráfico

Librerías

seaborn 0.11.1 Gallery Tutorial API Site Page Search

seaborn: statistical data visualization



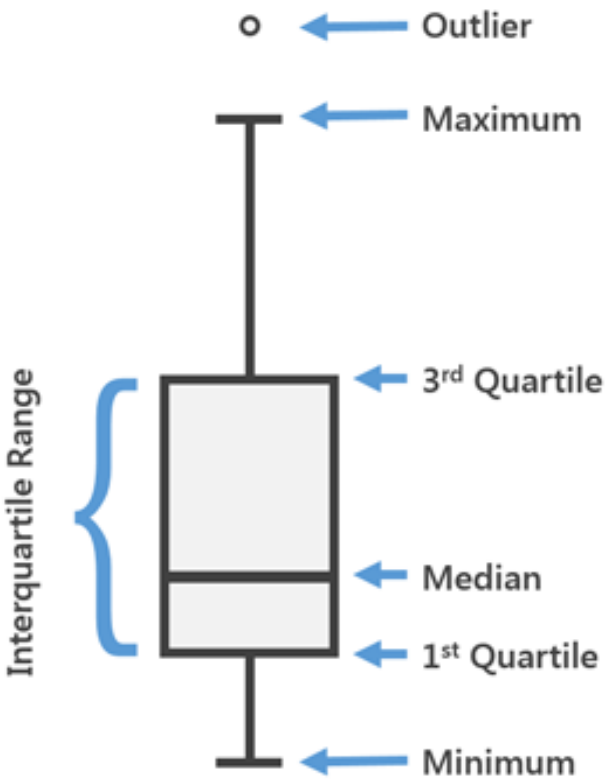
Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

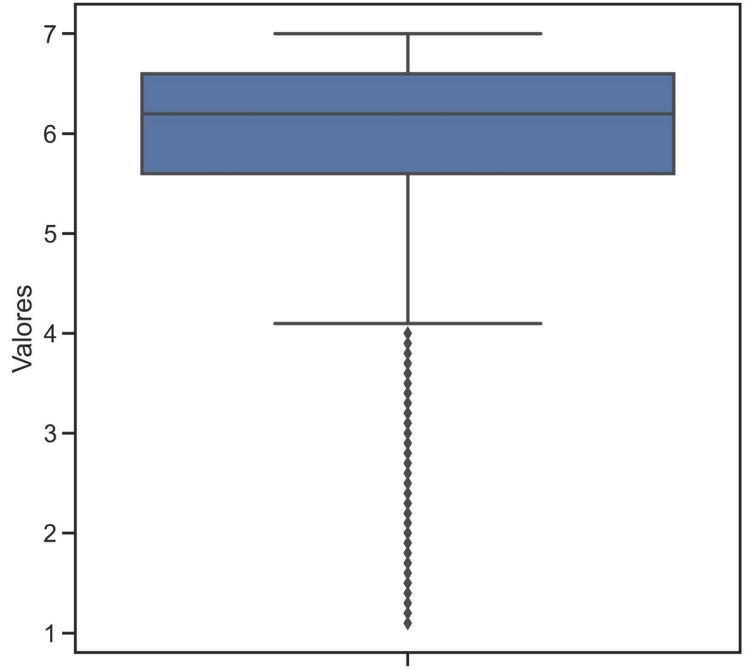
For a brief introduction to the ideas behind the library, you can read the introductory notes. Visit the installation page to see how you can download the package and get started with it. You can browse the example gallery to see what you can do with seaborn, and then check out the tutorial and API reference to find out how.

To see the code or report a bug, please visit the GitHub repository. General support questions are most at home on stackoverflow or discourse, which have dedicated channels for seaborn.

09:27

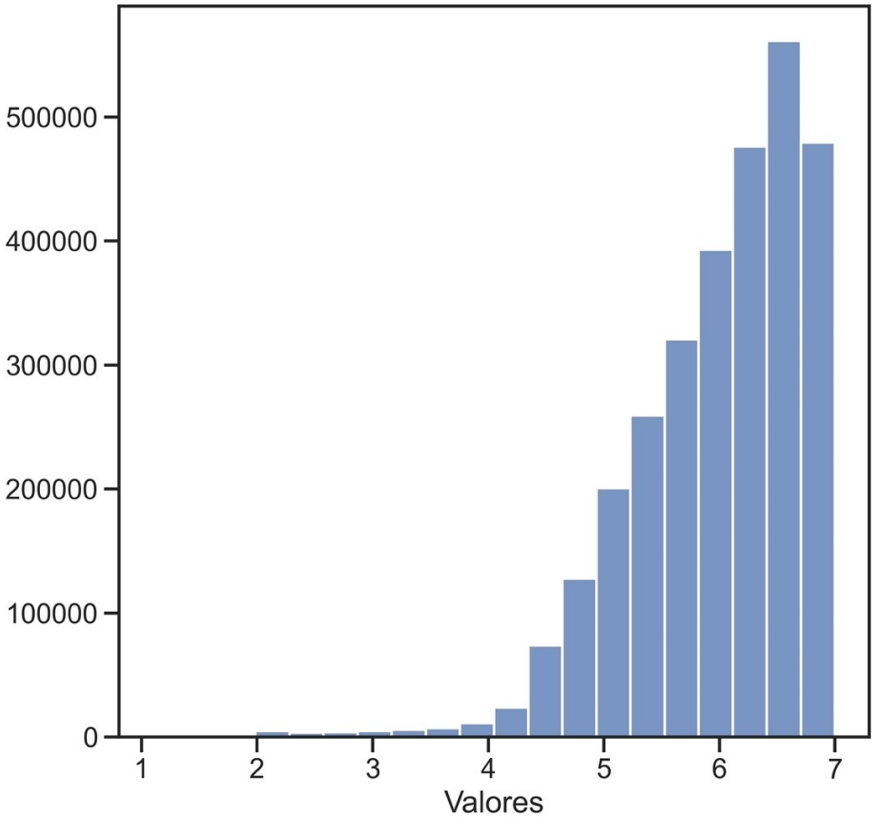
Boxplot





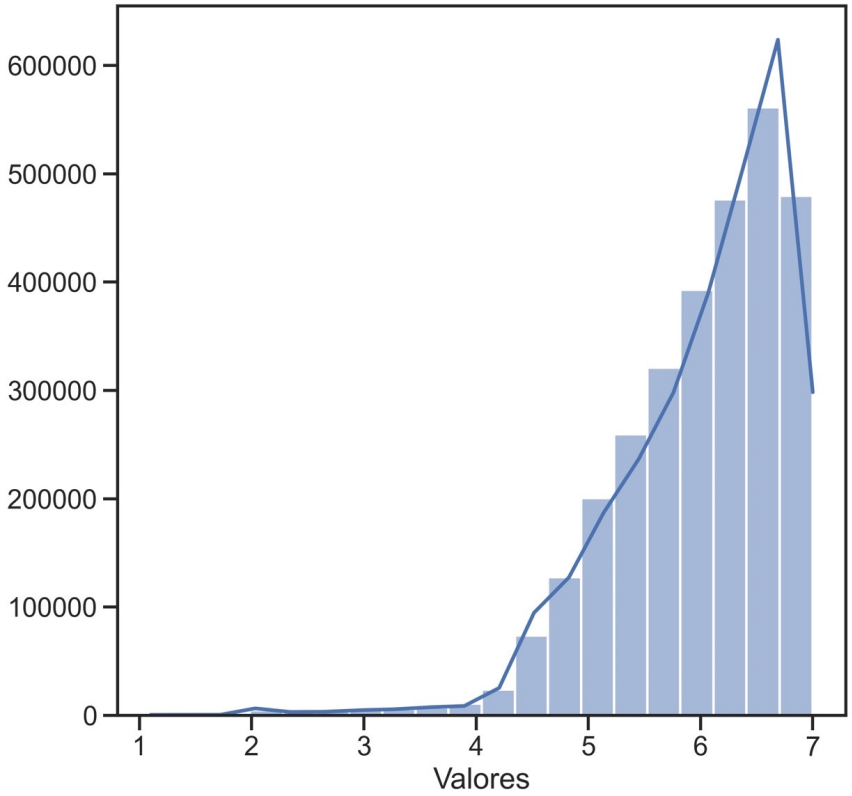
09:27

Histograma y Función de Densidad




```
sns.histplot(data=df,x='Valores'

,ax=ax,bins=20);
```

Histograma y Función de Densidad




```
sns.histplot(data=df,x='Valores'
,ax=ax,bins=20,kde=True);
```

09:27

Dataset de Ejemplo: Pasajeros del Titanic

- El dataset **`titanic.csv`** contiene datos para 887 pasajeros del Titanic, cada uno de los cuales representa una fila.
- Las columnas indican atributos de la persona: si sobrevivió, edad, clase, sexo, tarifa pagada, etc.

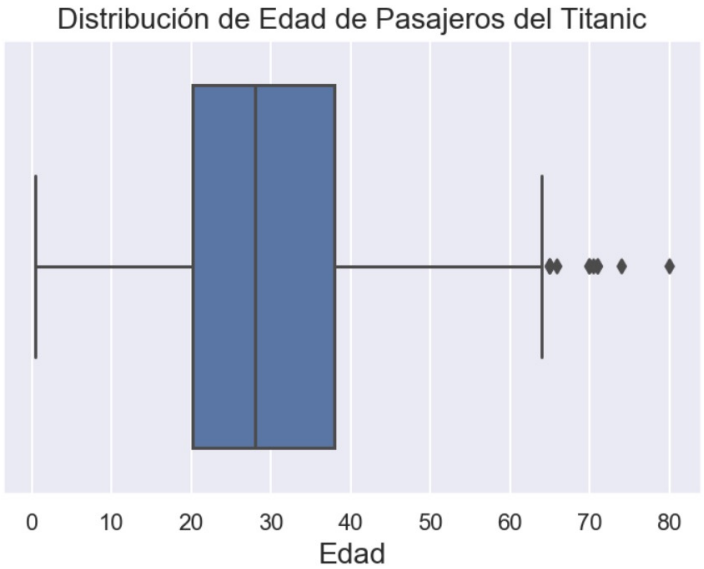
```
titanic = sns.load_dataset('titanic');
titanic.info();
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
survived       891 non-null int64
pclass         891 non-null int64
sex            891 non-null object
age            714 non-null float64
sibsp          891 non-null int64
parch          891 non-null int64
fare           891 non-null float64
embarked       889 non-null object
class          891 non-null category
who            891 non-null object
adult_male     891 non-null bool
deck           203 non-null category
embark_town    889 non-null object
alive          891 non-null object
alone          891 non-null bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.6+ KB
```

Boxplot

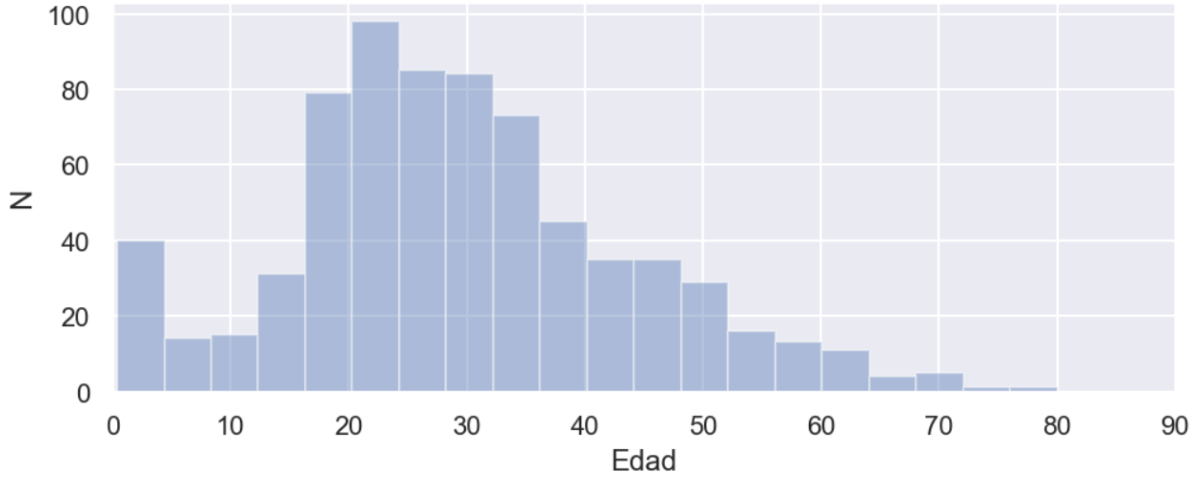
Una Variable

```
# seaborn
ax = sns.boxplot(x='age', data=titanic)
ax.set_ylabel(None);
ax.set_xlabel('Edad', fontsize=14);
ax.set_title('Distribución de Edad de Pasajeros del Titanic', fontsize=14);
```



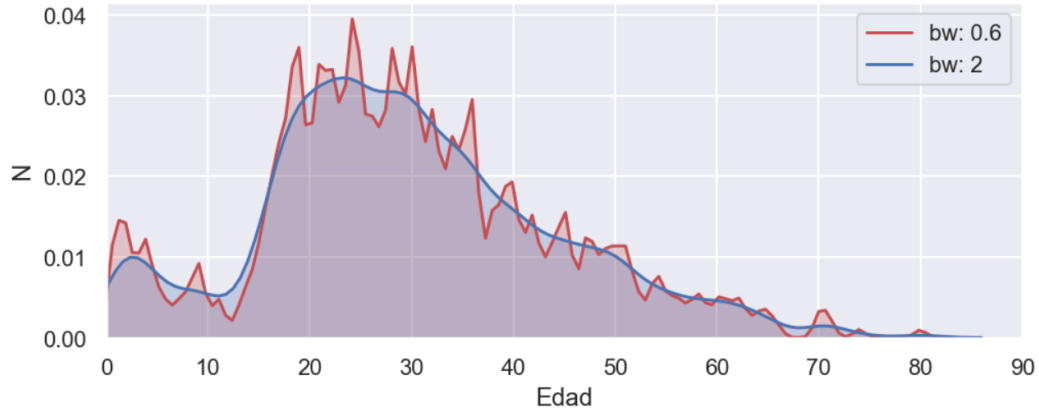
Histograma y Función de Densidad

```
# ¿Cuál es la distribución de edades de los pasajeros del Titanic?
import seaborn as sns
sns.set(color_codes=True)

f, ax = plt.subplots(1,1, figsize=(8, 3));
ax = sns.distplot(titanic.age, kde=False, bins=20)

ax.set(xlim=(0, 90));
ax.set_ylabel('N');
ax.set_xlabel('Edad');
```



Histograma y Función de Densidad

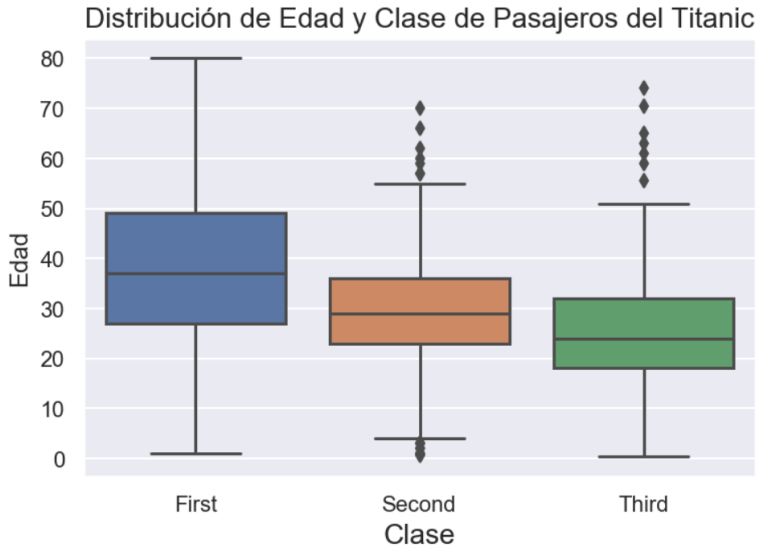
```
# ¿Cuál es la distribución de edades de los pasajeros del Titanic?
import seaborn as sns
sns.set(color_codes=True)

f, ax = plt.subplots(1,1, figsize=(8, 3));

sns.kdeplot(titanic.age, bw=0.6, label="bw: 0.6", shade=True, color="r",ax=ax);
sns.kdeplot(titanic.age, bw=2, label="bw: 2", shade=True,ax=ax);

ax.set(xlim=(0, 90));
ax.set_ylabel('N');
ax.set_xlabel('Edad');
```



Boxplot

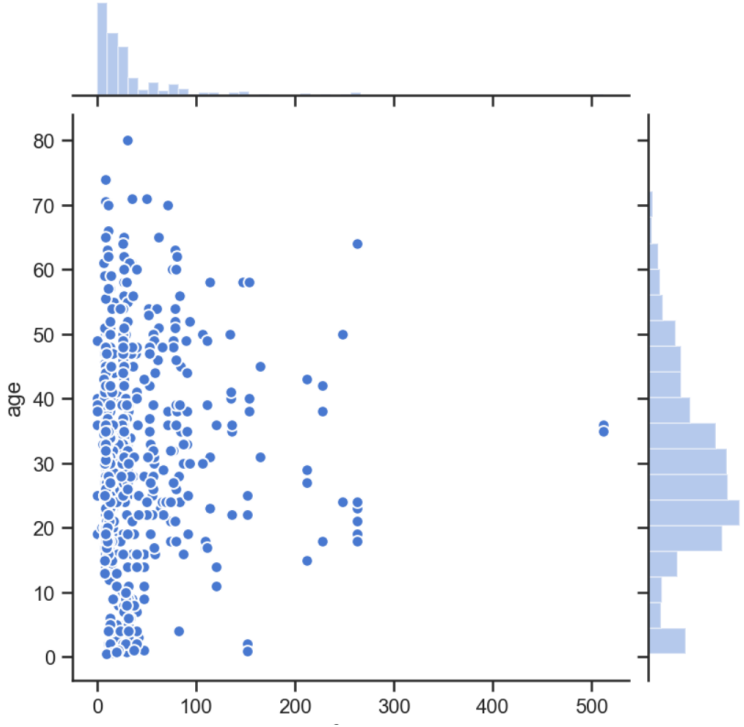
Dos variables: ¿la primera clase se componía de pasajeros más jóvenes o mayores?

```
:
1  ax = sns.boxplot(x='class', y='age', data=titanic)
2  ax.set_ylabel('Edad');
3  ax.set_xlabel('Clase', fontsize=14);
4  ax.set_title('Distribución de Edad y Clase de Pasajeros del Titanic', fontsize=14);
```



Scatterplot / Jointplot : Relación entre 2 variables

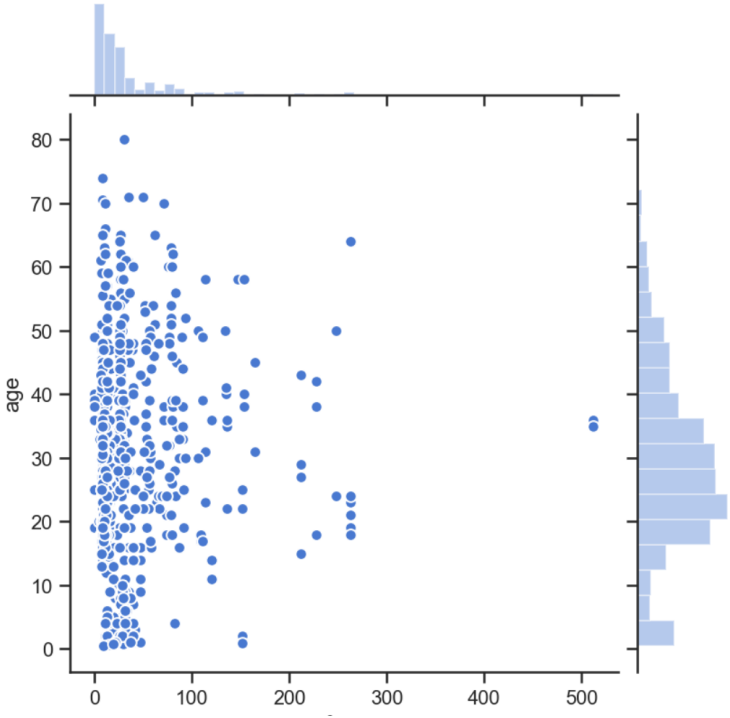
```
sns.jointplot("fare", "age", data=titanic, s=40, edgecolor="w", linewidth=1);
ax.set_ylabel('Tarifa');
ax.set_xlabel('Edad', fontsize=14);
ax.set_title('Relación entre Edad y Tarifa Pagada por Pasajeros del Titanic', fontsize=14);
```



09:27

Scatterplot / Jointplot : Relación entre 2 variables

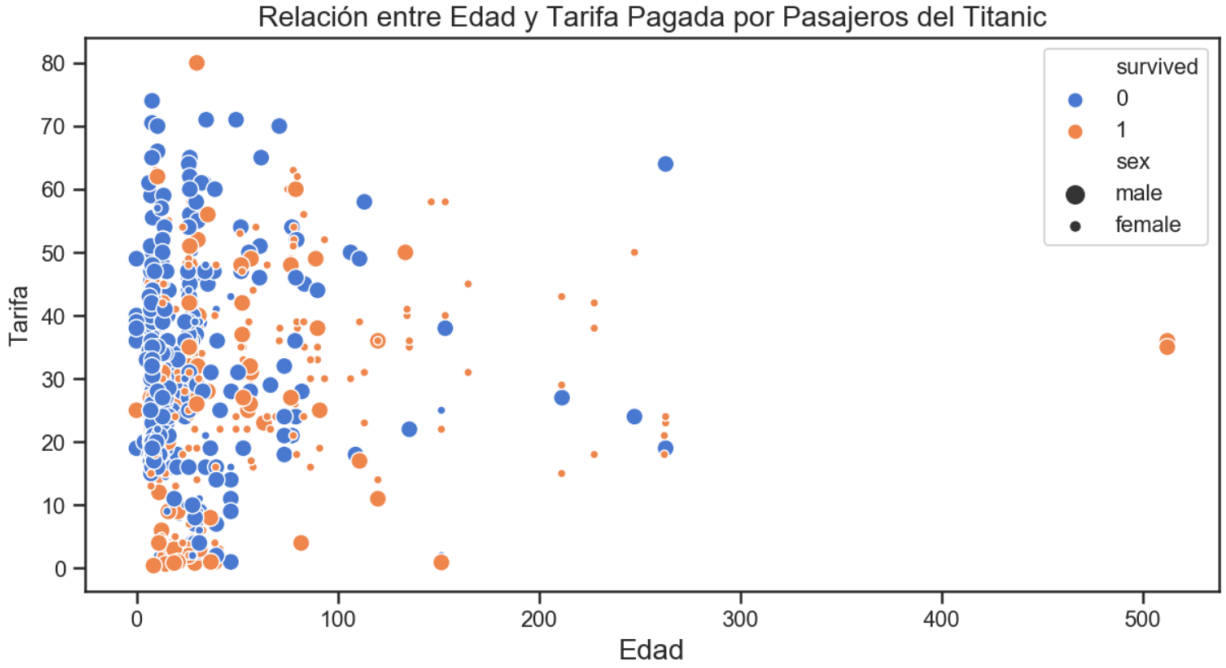
```
sns.jointplot("fare", "age", data=titanic, s=40, edgecolor="w", linewidth=1);
ax.set_ylabel('Tarifa');
ax.set_xlabel('Edad', fontsize=14);
ax.set_title('Relación entre Edad y Tarifa Pagada por Pasajeros del Titanic', fontsize=14);
```



09:27

Scatterplot y Distplot: Relación entre 2 o más variables

```
f, ax = plt.subplots(1,1, figsize=(10, 5))
sns.scatterplot(x="fare", y="age", hue="survived", size="sex", data=titanic, ax=ax);
ax.set_ylabel('Tarifa');
ax.set_xlabel('Edad', fontsize=14);
ax.set_title('Relación entre Edad, Tarifa Pagada, Sobrevivencia y Sexo', fontsize=14);
plt.show();
```



09:27

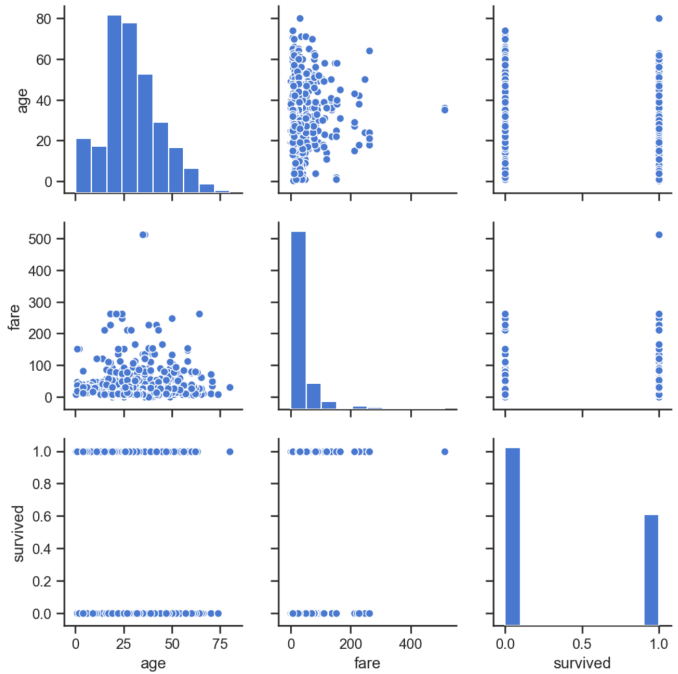
Pairplot: Análisis multivariable

```
1 titanic.columns
```

```
Index(['survived', 'pclass', 'sex', 'age', 'sibsp', 'parch', 'fare', 'class', 'deck'], dtype
='object')
```

```
1 graficar = ['age', 'fare', 'survived', 'deck']
```

```
1 titanic_plot = titanic.loc[:,graficar]
2 sns.pairplot(titanic_plot);
```



09:27

Pairplot: Análisis multivariable

```
In [ ]:
1 titanic.columns
```

```
Out[ ]:
Index(['survived', 'pclass', 'sex', 'age', 'sibsp', 'parch', 'fare', 'class', 'deck'], dtype
='object')
```

```
In [ ]:
1 graficar = ['age', 'fare', 'survived', 'deck']
```

```
In [ ]:
1 titanic_plot = titanic.loc[:,graficar]
2 sns.pairplot(titanic_plot,hue='deck');
```