Autor:
Iván García Santillán

# Análisis exploratorio de datos

## Introducción

El análisis exploratorio de datos (EDA por sus siglas en inglés, Exploratory Data Analysis) es una parte fundamental de cualquier proyecto de análisis de datos que tiene varios propósitos importantes, tales como: Comprender los datos, incluyendo la estructura de los datos, sus características, distribuciones, tendencias y patrones subyacentes); Detectar Anomalías y Errores; Seleccionar variables relevantes; Generar hipótesis iniciales sobre relaciones entre variables; Tomar decisiones iniciales sobre cómo abordar el proyecto de análisis de datos (elección de algoritmos), Optimización de recursos eliminando variables irrelevantes o reducir la dimensionalidad de los datos; Preparación para un modelado avanzado, etc.

Aquí se mostrará cómo realizar un análisis exploratorio de datos básico en el conjunto de datos del *Titanic* utilizando Python, centrándonos en estadísticas descriptivas y visualización de datos.

## Contenido

### Análisis exploratorio de datos (estadísticas y visualización)

En esta lectura se mostrará cómo realizar un análisis exploratorio de datos básico en el conjunto de datos del *Titanic* utilizando Python, centrándonos en estadísticas descriptivas y visualización de datos. Para este ejemplo, asumimos que se tiene el conjunto de datos del *Titanic* en un archivo CSV llamado "titanic.csv".
Primero, debes importar las bibliotecas necesarias, cargar los datos y observar las primeras filas para tener una idea de cómo se ven los datos:

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

# Cargar el conjunto de datos del Titanic
titanic_df = pd.read_csv('titanic.csv')

# Ver las primeras filas del conjunto de datos
print(titanic_df.head())
```

A continuación, se obtiene estadísticas descriptivas básicas de las variables numéricas en el conjunto de datos:

```
# Estadísticas descriptivas
print(titanic_df.describe())
```

Para obtener estadísticas de las variables categóricas, se puede utilizar el siguiente código:

```
# Estadísticas de variables categóricas
print(titanic_df.describe(include=['object']))
```

Para visualizar los datos, puedes crear varios tipos de gráficos. Aquí hay algunos ejemplos:

Histogramas de Edades (Figura 7):

```
plt.figure(figsize=(8, 6))
sns.histplot(titanic_df['Age'].dropna(), bins=30, kde=True)
plt.xlabel('Edad')
plt.ylabel('Frecuencia')
plt.title('Distribución de Edades en el Titanic')
plt.show()
```

Figura 7. Distribución de edades en el dataset Titanic

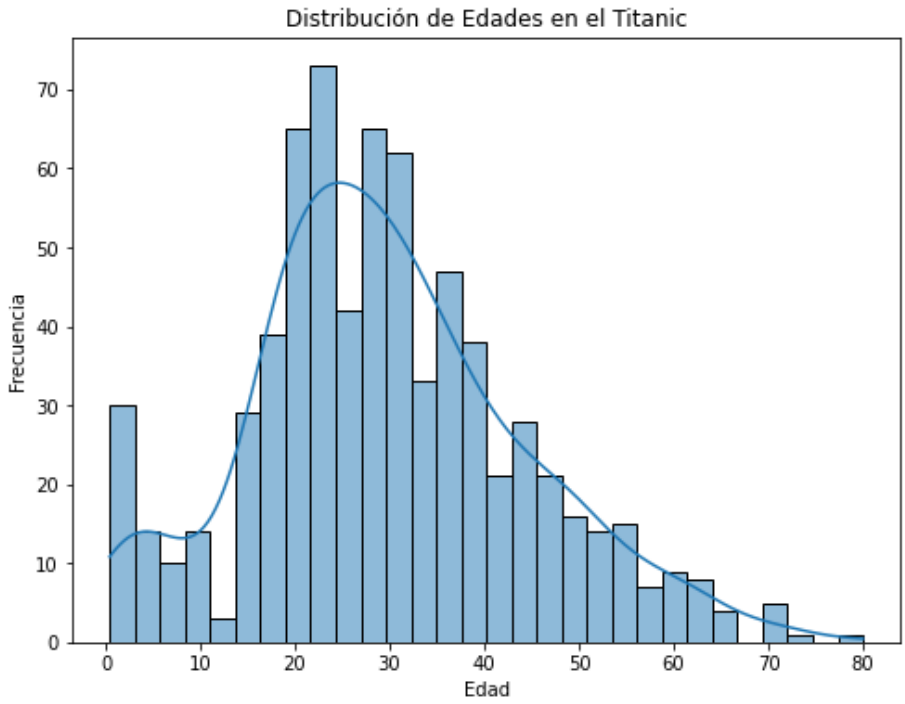



Gráfico de Barras de Supervivencia por Clase (Figura 8):

```
plt.figure(figsize=(8, 6))
sns.countplot(data=titanic_df, x='Pclass', hue='Survived')
plt.xlabel('Clase')
plt.ylabel('Cantidad')
plt.title('Supervivencia por Clase en el Titanic')
plt.legend(title='Sobreviviente', labels=['No', 'Sí'])
plt.show()
```

Figura 8. Supervivencia de pasajeros por clase en el dataset Titanic

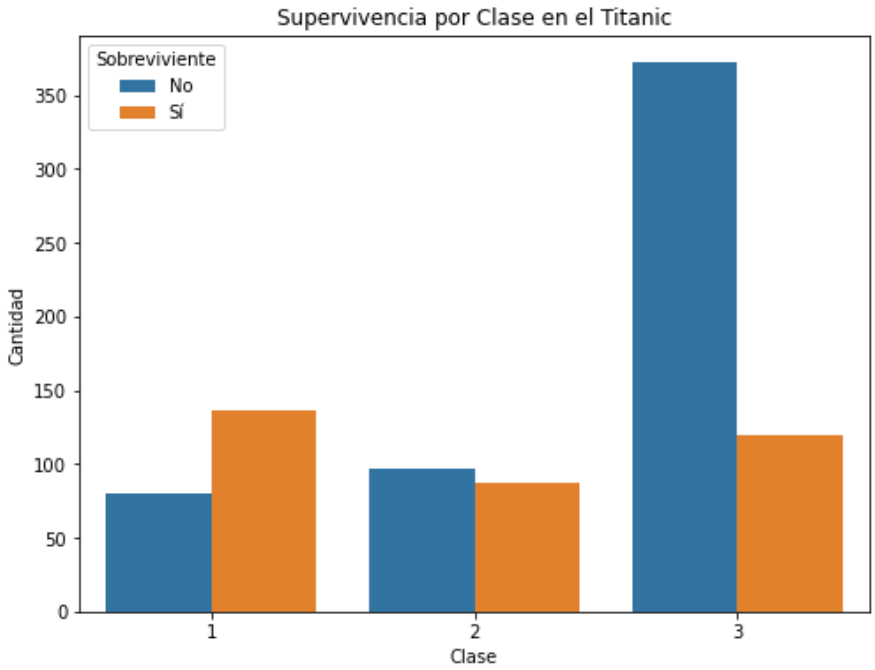
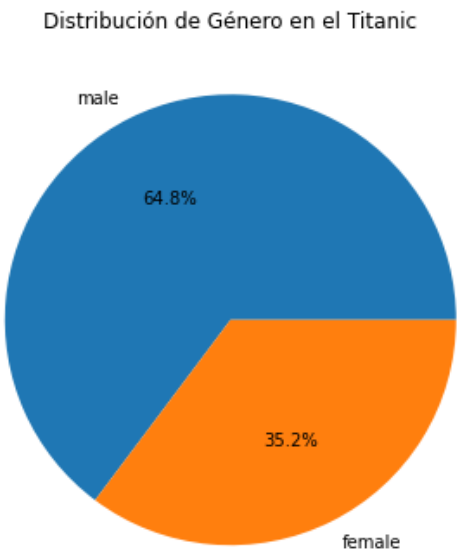



Gráfico de Torta de Género (Figura 9):

```
plt.figure(figsize=(8, 6))
titanic_df['Sex'].value_counts().plot(kind='pie', autopct='%1.1f%%')
plt.title('Distribución de Género en el Titanic')
plt.ylabel('')
plt.show()
```

Figura 9. Distribución del género de pasajeros en el dataset Titanic




Se puede continuar explorando y visualizando más aspectos de los datos, como la correlación entre variables, la distribución de tarifas, etc., según los objetivos de análisis. La biblioteca ***seaborn*** es especialmente útil para la creación de gráficos informativos y visualmente atractivos.

Se va a calcular la matriz de correlación y luego visualizarla de manera efectiva, utilizando la biblioteca ***panda*** para manejar los datos y ***seaborn*** para la visualización.

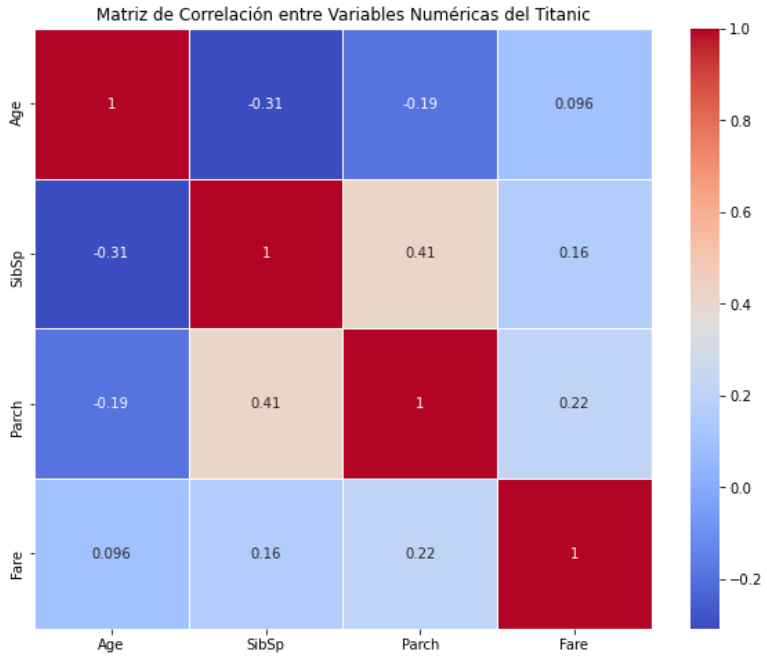
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Cargar el conjunto de datos del Titanic
titanic_df = pd.read_csv('titanic.csv')

# Seleccionar las variables numéricas principales para el análisis de correlación
variables_numericas = ['Age', 'SibSp', 'Parch', 'Fare']

# Crear una submatriz de correlación
correlation_matrix = titanic_df[variables_numericas].corr()

# Crear un mapa de calor de correlación
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=.5)
plt.title('Matriz de Correlación entre Variables Numéricas del Titanic')
```

```
plt.show()
```

En este código:

- Se carga el conjunto de datos del Titanic en un DataFrame de pandas.
- Se selecciona las variables numéricas que se desea incluir en el análisis de correlación, en este caso, 'Age' (edad), 'SibSp' (número de hermanos/cónyuges a bordo), 'Parch' (número de padres/hijos a bordo) y 'Fare' (tarifa).
- Se calcula la matriz de correlación utilizando el método .corr() del DataFrame.
- Se visualiza la matriz de correlación en un mapa de calor usando seaborn. Los valores de correlación se muestran en cada celda del mapa de calor (Figura 10).

Figura 10. Matriz de correlación entre las variables numéricas del dataset Titanic




El mapa de calor de correlación muestra la fuerza y la dirección de las correlaciones entre estas variables numéricas. Los valores de correlación pueden variar entre -1 y 1, donde -1 indica una correlación negativa perfecta, 1 indica una correlación positiva perfecta y 0 indica que no hay correlación. Además, los colores del mapa de calor ayudan a identificar visualmente las relaciones. Para dar un mayor significado a la correlación se puede utilizar los valores de la Tabla 4.

Tabla 4. Significado de la correlación

| VALOR | SIGNIFICADO |
|---|---|
| -1 | Correlación negativa grande y perfecta |
| -0,9 a -0,99 | Correlación negativa muy alta |
| -0,7 a -0,89 | Correlación negativa alta |
| -0,4 a -0,69 | Correlación negativa moderada |
| -0,2 a -0,39 | Correlación negativa baja |
| -0,01 a -0,19 | Correlación negativa muy baja |
| 0 | Correlación nula |
| 0,01 a 0,19 | Correlación positiva muy baja |
| 0,2 a 0,39 | Correlación positiva baja |
| 0,4 a 0,69 | Correlación positiva moderada |
| 0,7 a 0,89 | Correlación positiva alta |
| 0,9 a 0,99 | Correlación positiva muy alta |
| 1 | Correlación positiva grande y perfecta |

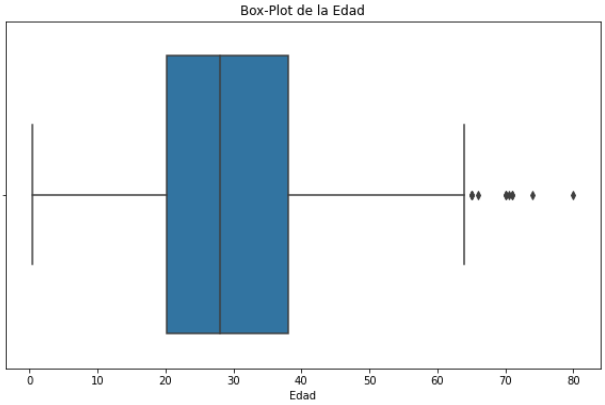
Fuente: Lind et al. (2015)

Este análisis de correlación permitirá entender mejor cómo las variables numéricas principales se relacionan entre sí en el conjunto de datos del Titanic y puede ser útil para futuros análisis o modelado de datos.

También, se visualiza un diagrama de cajas (box-plot) de la variable edad, donde se aprecia los valores atípicos.

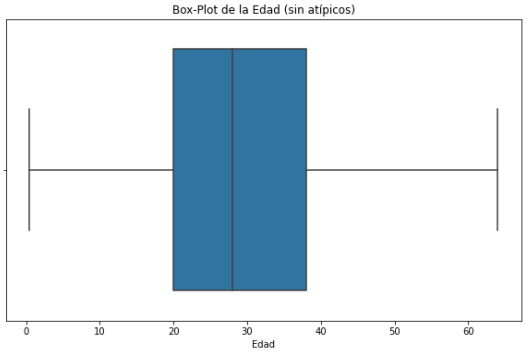
```
# Crear el box-plot de la variable "edad"
plt.figure(figsize=(10, 6))
sns.boxplot(x=titanic_df['Age'])
plt.title('Box-Plot de la Edad')
plt.xlabel('Edad')
plt.show()
```

Figura 11. Diagrama de cajas de la variable edad




Los **outliers** pueden llevar al modelo a aprender patrones que no son representativos de la población general, lo que puede resultar en un sobreajuste del modelo. Esto significa que el modelo puede funcionar bien en el conjunto de datos de entrenamiento, pero no generalizarse bien a nuevos datos. Por esto, se podría experimentar eliminando los valores atípicos:

```
# Eliminar valores atípicos (filas)
Q1 = titanic_df['Age'].quantile(0.25)
Q3 = titanic_df['Age'].quantile(0.75)
IQR = Q3 - Q1 # rango intercuartil
limite_superior = Q3 + 1.5*IQR
limite_inferior = Q1 - 1.5*IQR
filtered_titanic_df = titanic_df[(titanic_df['Age'] >= limite_inferior) & (titanic_df['Age'] <=
limite_superior)]

# Crear el box-plot de la variable "edad"
plt.figure(figsize=(10, 6))
sns.boxplot(x=filtered_titanic_df['Age'])
plt.title('Box-Plot de la Edad')
plt.xlabel('Edad')
plt.show()
```



## Conclusiones

En esta unidad se ha revisado los fundamentos de la ciencia de datos, respecto al Análisis exploratorio de datos (estadísticas y visualización) que es una parte fundamental de cualquier proyecto de análisis de datos que tiene varios propósitos importantes, tales como: Comprender los datos, incluyendo la estructura de los datos, sus características, distribuciones, tendencias y patrones subyacentes; Detectar Anomalías y Errores; Seleccionar variables relevantes; Generar hipótesis iniciales sobre relaciones entre variables; Tomar decisiones iniciales sobre cómo abordar el proyecto de análisis de datos (elección de algoritmos), Optimización de recursos eliminando variables irrelevantes o reducir la dimensionalidad de los datos; Preparación para un modelado avanzado de datos, etc. Se motiva a los estudiantes a revisar material y ejemplos adicionales disponibles libremente en Internet.

# ANEXOS

## Referencias:

Lind et al., (2015). Estadística aplicada a los negocios y la economía. 16° ed. McGraw-Hill: México.