Catalina Chavarría

  Curso: Machine Learning

Caso de análisis 1: EDA con pandas en PythonURL

Contenido

[**Introducción** 2](#_Toc149243284)

[1. Preparando datos 2](#_Toc149243285)

[2. Explorando la data 3](#_Toc149243286)

[3. Limpieza de datos 4](#_Toc149243287)

[4. Visualización de los datos 5](#_Toc149243288)

[5. Explorando las relaciones 7](#_Toc149243289)

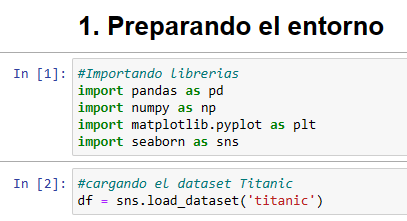
[**Conclusión** 8](#_Toc149243290)

# **Introducción**

Para este primer caso de análisis exploratorio de datos (EDA). Se utilizará el dataset Titanic en el cual se explorará sus datos y realizando análisis y creando mapas utilizando Python en este proyecto

## Preparando datos

Para este primer apartado importamos las librerías que necesitaremos para realizar este ejercicio.  
Seguido carga el dataset Titanic para realizar el análisis de la base de datos.



## Explorando la data

En este apartado se realizan operaciones como:

* Mostrar las primeras filas del dataset
* Verificar la dimensión del dataset
* Realizar resumen de las variables numéricas que tiene el dataset.
* Y verificamos las variables que no aparecen en el dataset.

Todo esto para comprender mejor la distribución de los datos.

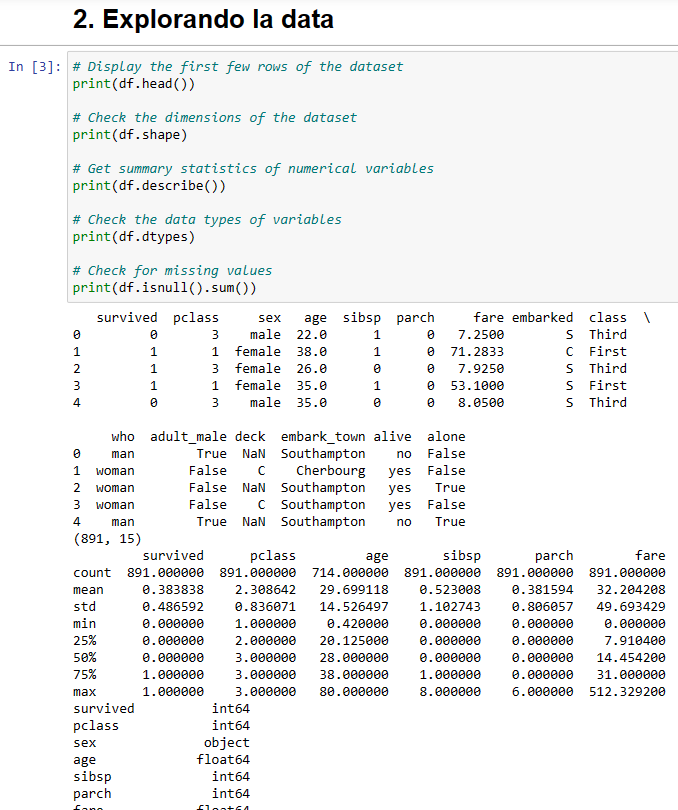


Imagen que contiene Texto

Descripción generada automáticamente

## Limpieza de datos

Aquí limpiamos los datos eliminando valores atípicos o en este caso datos que faltan. Esto garantiza que los datos sean precisos.

Texto

Descripción generada automáticamente

La sintaxis **dropna** nos ayuda eliminando de las filas los valores que falten

Y **fillna** rellena los valores que faltan con otro dato introducido en este caso, la variable es ‘valor’

## Visualización de los datos

Creamos gráficos de Matplotlib y Seaborn para obtener una primera impresión de la distribución de los datos. En este caso se harán varios ejemplos:

Este primer grafico mostrando cuantos sobrevivieron al titanic

Texto

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Un histograma distribuyendo la edad de los pasajeros a bordo del titanic.

Texto

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Diagrama de dispersión entre la edad vs la tarifa de pasaje por pasajero.

Texto

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Y por último este diagrama de cajas visualizando el status de sobrevivientes contra la tarifa

Texto

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Creando estas visualizaciones como gráficos de dispersión con colores o diagrama de caja nos ayuda para explorar más a fondo las relaciones y patrones en los datos.

## Explorando las relaciones

En este ultimo apartado lo que hacemos es examinar las relaciones entre diferentes variables utilizando matrices de correlación. Esto puede revelar patrones o asociaciones importantes.

Las correlaciones es una técnica usada en la estadística para evaluar la relación entre dos o más variables en un conjunto de datos.

En este caso muestra cómo estas dos variables se relacionan entre sí las columnas edad y tarifa. Los valores de correlación varían entre -1 y 1, donde -1 indica una correlación negativa, 1 indica una correlación positiva y 0 indica que no hay correlación.

La correlación entre edad y tarifa es aproximadamente -0.09073, lo que indica una correlación negativa entre la edad y la tarifa. A medida que la edad aumenta, la tarifa tiende a disminuir.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

# **Conclusión**

El EDA es un proceso crucial a la hora de realizar análisis de datos y poder explorar y comprender un conjunto de datos antes de realizar un análisis más profundo.   
Este enfoque nos ayuda a tomar las mejores decisiones y a obtener información valiosa de los datos.