



web
Scraping **Booking.com**



Introducción

- En este proyecto, realizamos un web scraping de Booking.com para extraer información sobre alojamientos en Madrid.
- Utilizamos Selenium para automatizar la extracción de datos y analizamos los resultados con Pandas y Plotly.
- Finalmente, construimos un Dashboard interactivo con Dash para visualizar y explorar los datos de forma dinámica.

Objetivo:

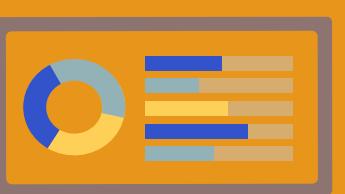
- Obtener información clave sobre precios, evaluación y disponibilidad de alojamientos.
- Identificar patrones en la oferta hotelera de Madrid.
- Diseñar un cuadro de mando para facilitar la toma de decisiones.



- 

1. Web Scraping con Selenium
Extracción de datos de Booking.com ★★★★★
- 

2. Limpieza y Transformación de Datos
Eliminación de datos inconsistentes ★★★★★
- 

3. Análisis Exploratorio de Datos (EDA)
Identificación de tendencias ★★★★★
- 

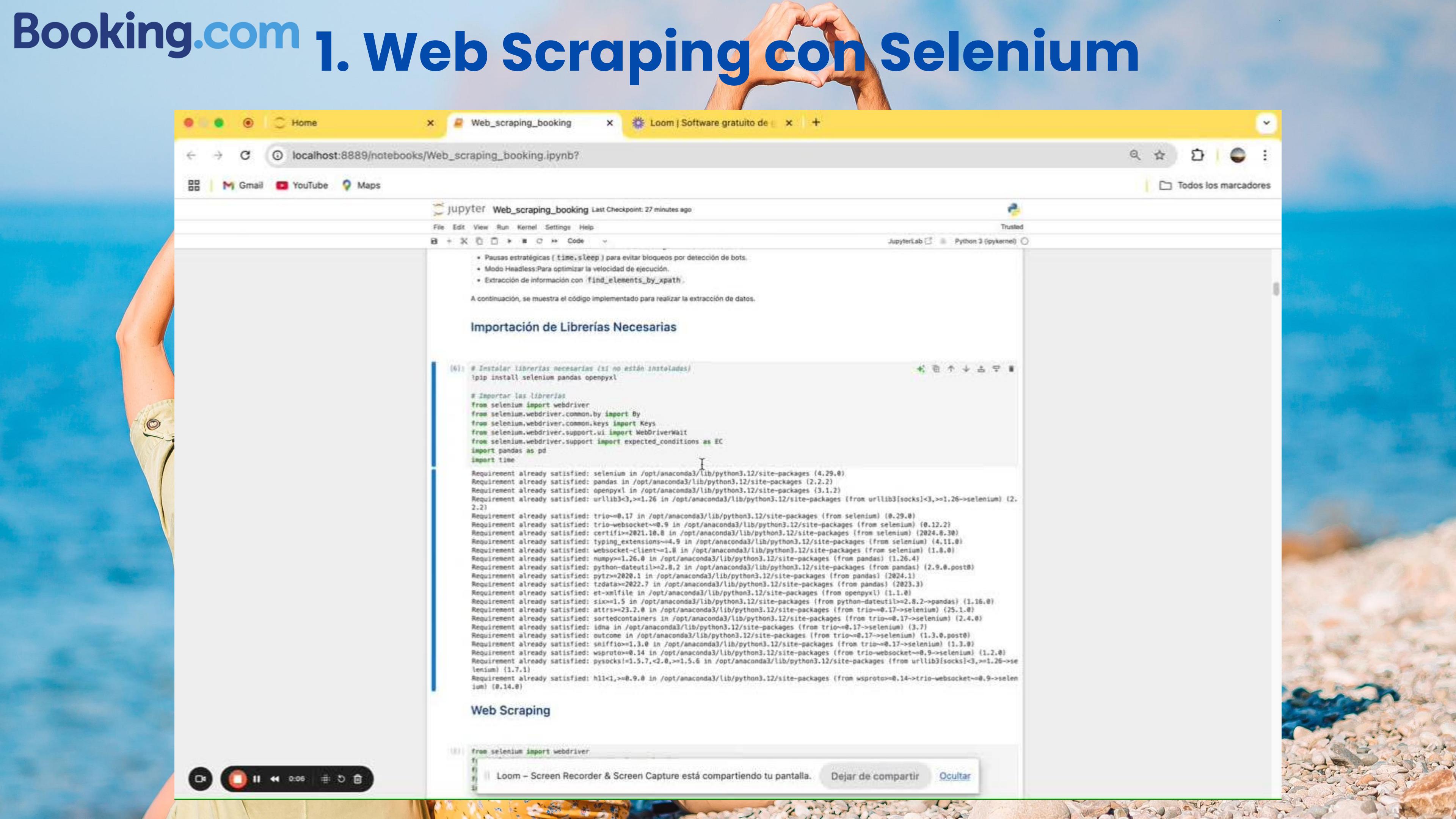
4. Dashboard Interactivo con Dash
Creación de visualizaciones dinámicas ★★★★★

Metodología

Tecnologías Utilizadas

Librería	Función
Selenium	Automatización de navegación y extracción de datos.
Pandas	Procesamiento y limpieza de datos
Plotly	Creación de visualizaciones interactivas
Dash	Desarrollo del cuadro de mando interactivo
Matplotlib	Generación de gráficos estáticos y personalizables

Booking.com 1. Web Scraping con Selenium



A screenshot of a Jupyter Notebook interface titled "jupyter Web_scraping_booking". The notebook is running on "localhost:8889/notebooks/Web_scraping_booking.ipynb?". The code cell contains Python code for web scraping, specifically for Booking.com. The code includes imports for selenium, pandas, and openpyxl, and defines a function for web scraping. A Loom sharing overlay is visible at the bottom.

```
# Instalar librerías necesarias (si no están instaladas)
!pip install selenium pandas openpyxl

# Importar las librerías
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import pandas as pd
import time

Requirement already satisfied: selenium in /opt/anaconda3/lib/python3.12/site-packages (4.29.0)
Requirement already satisfied: pandas in /opt/anaconda3/lib/python3.12/site-packages (2.2.2)
Requirement already satisfied: openpyxl in /opt/anaconda3/lib/python3.12/site-packages (3.1.2)
Requirement already satisfied: urllib3<3,>=1.26 in /opt/anaconda3/lib/python3.12/site-packages (from selenium) (2.2)
Requirement already satisfied: trio==0.17 in /opt/anaconda3/lib/python3.12/site-packages (from selenium) (0.29.0)
Requirement already satisfied: trio-websocket~=0.9 in /opt/anaconda3/lib/python3.12/site-packages (from selenium) (0.12.2)
Requirement already satisfied: certifi>=2021.10.8 in /opt/anaconda3/lib/python3.12/site-packages (from selenium) (2024.8.30)
Requirement already satisfied: typing_extensions~=4.9 in /opt/anaconda3/lib/python3.12/site-packages (from selenium) (4.11.0)
Requirement already satisfied: websocket-client~=1.8 in /opt/anaconda3/lib/python3.12/site-packages (from selenium) (1.8.0)
Requirement already satisfied: numpy>=1.26.0 in /opt/anaconda3/lib/python3.12/site-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/anaconda3/lib/python3.12/site-packages (from pandas) (2.9.0.post8)
Requirement already satisfied: pytz>=2020.1 in /opt/anaconda3/lib/python3.12/site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in /opt/anaconda3/lib/python3.12/site-packages (from pandas) (2023.3)
Requirement already satisfied: et-xmlfile in /opt/anaconda3/lib/python3.12/site-packages (from openpyxl) (1.1.0)
Requirement already satisfied: six>=1.5 in /opt/anaconda3/lib/python3.12/site-packages (from python-dateutil>=2.8.2>pandas) (1.16.0)
Requirement already satisfied: attrs>=23.2.0 in /opt/anaconda3/lib/python3.12/site-packages (from trio==0.17>selenium) (25.1.0)
Requirement already satisfied: sortedcontainers in /opt/anaconda3/lib/python3.12/site-packages (from trio==0.17>selenium) (2.4.0)
Requirement already satisfied: idna in /opt/anaconda3/lib/python3.12/site-packages (from trio==0.17>selenium) (3.7)
Requirement already satisfied: outcome in /opt/anaconda3/lib/python3.12/site-packages (from trio==0.17>selenium) (1.3.0.post0)
Requirement already satisfied: sniffio>=1.3.0 in /opt/anaconda3/lib/python3.12/site-packages (from trio==0.17>selenium) (1.3.0)
Requirement already satisfied: wsproto>=0.14 in /opt/anaconda3/lib/python3.12/site-packages (from trio-websocket~=0.9>selenium) (1.2.0)
Requirement already satisfied: pysocks!=1.5.7,<2.8,>=1.5.6 in /opt/anaconda3/lib/python3.12/site-packages (from urllib3[socks]<3,>=1.26>selenium) (1.7.1)
Requirement already satisfied: h1l<1,>=0.9.0 in /opt/anaconda3/lib/python3.12/site-packages (from wsproto>=0.14>trio-websocket~=0.9>selenium) (0.14.0)

from selenium import webdriver
```

Importación de Librerías Necesarias

Web Scraping

Loom - Screen Recorder & Screen Capture está compartiendo tu pantalla. Dejar de compartir Ocultar

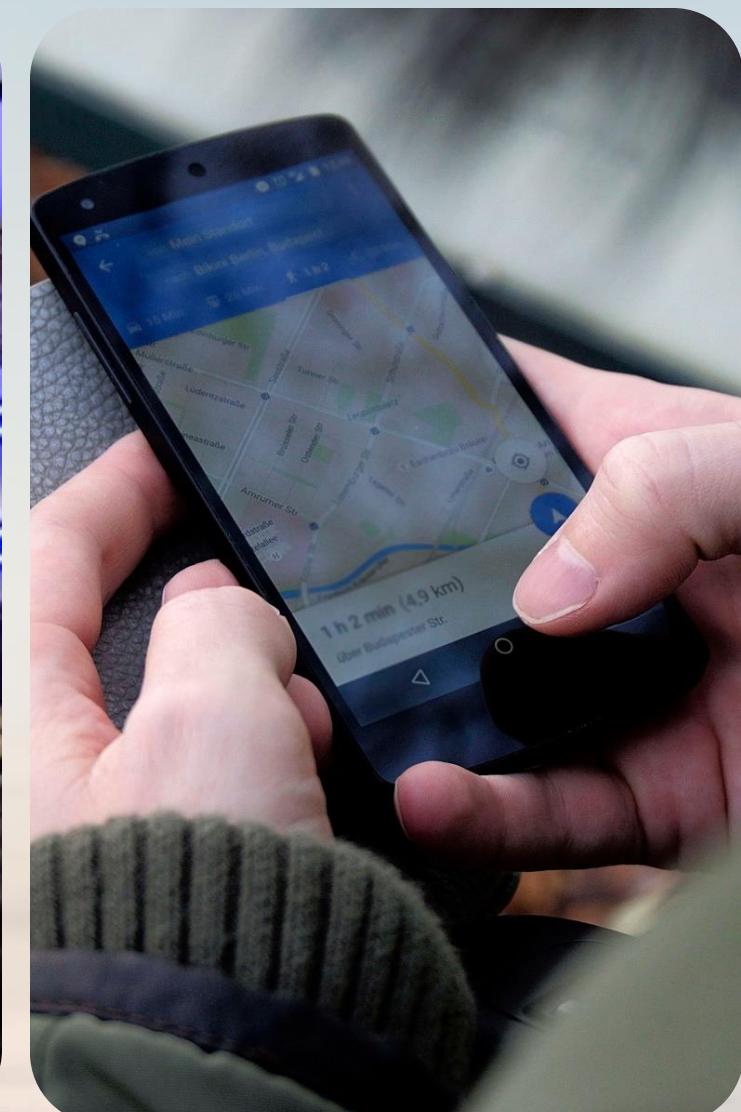
Obtener datos clave de los alojamientos en Booking.com



Extraer nombres de
alojamientos



Extraer Precios



Extraer Ubicación



Puntuación
Específica

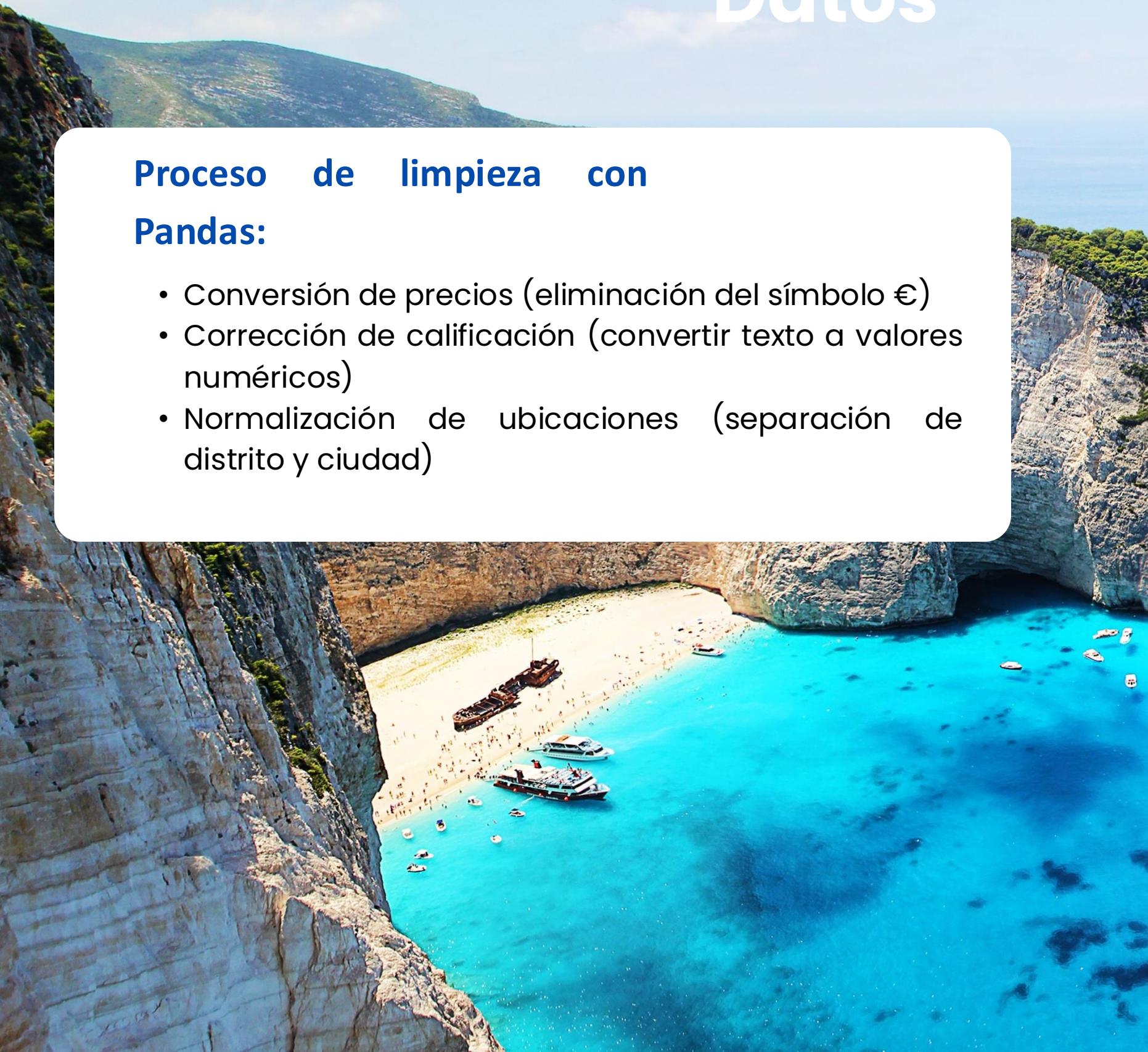


Número de estrellas/
Extraer Reseñas

2. Limpieza y Transformación de Datos

Proceso de limpieza con Pandas:

- Conversión de precios (eliminación del símbolo €)
- Corrección de calificación (convertir texto a valores numéricos)
- Normalización de ubicaciones (separación de distrito y ciudad)



DATOS LIMPIOS

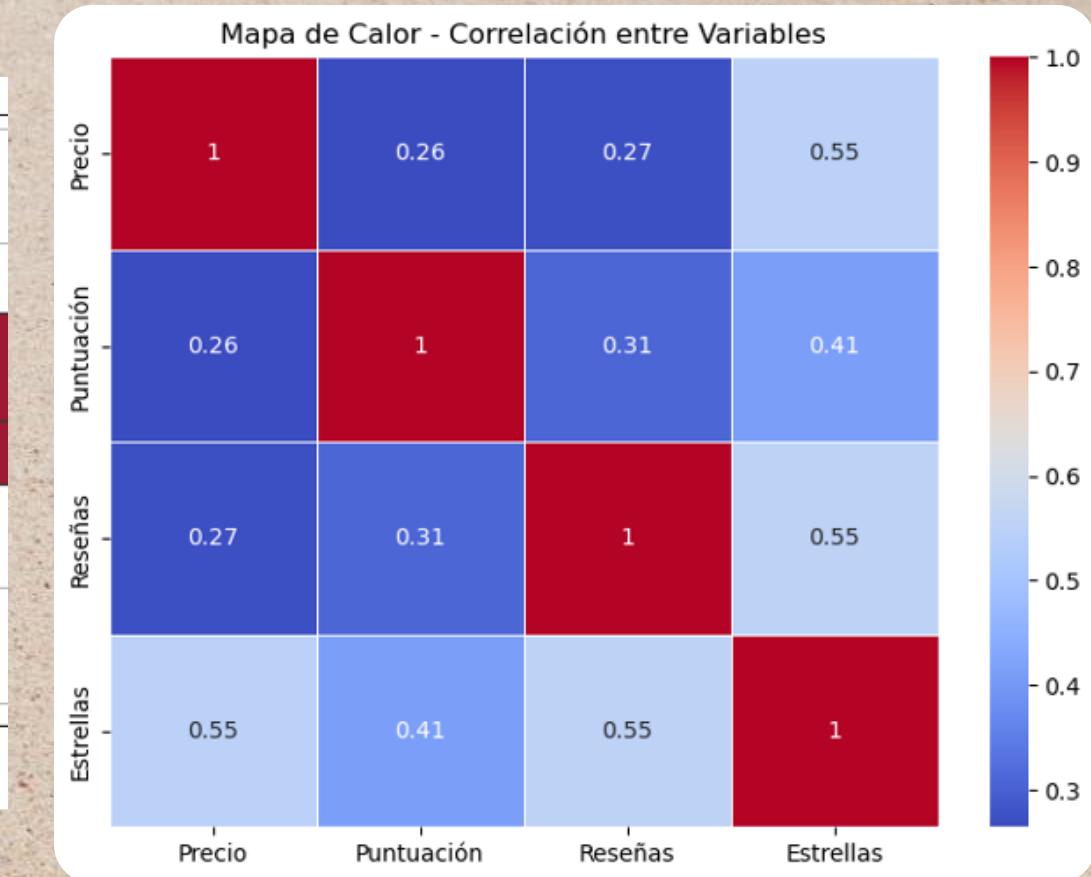
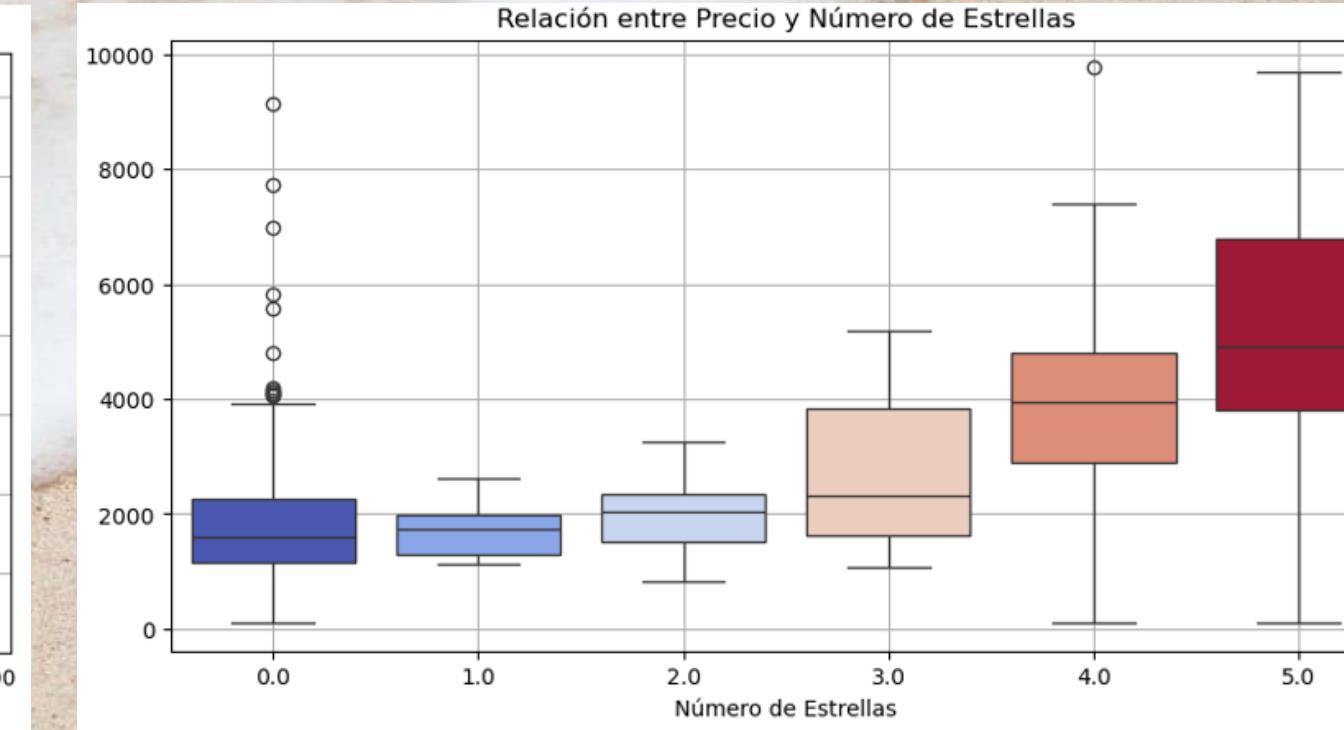
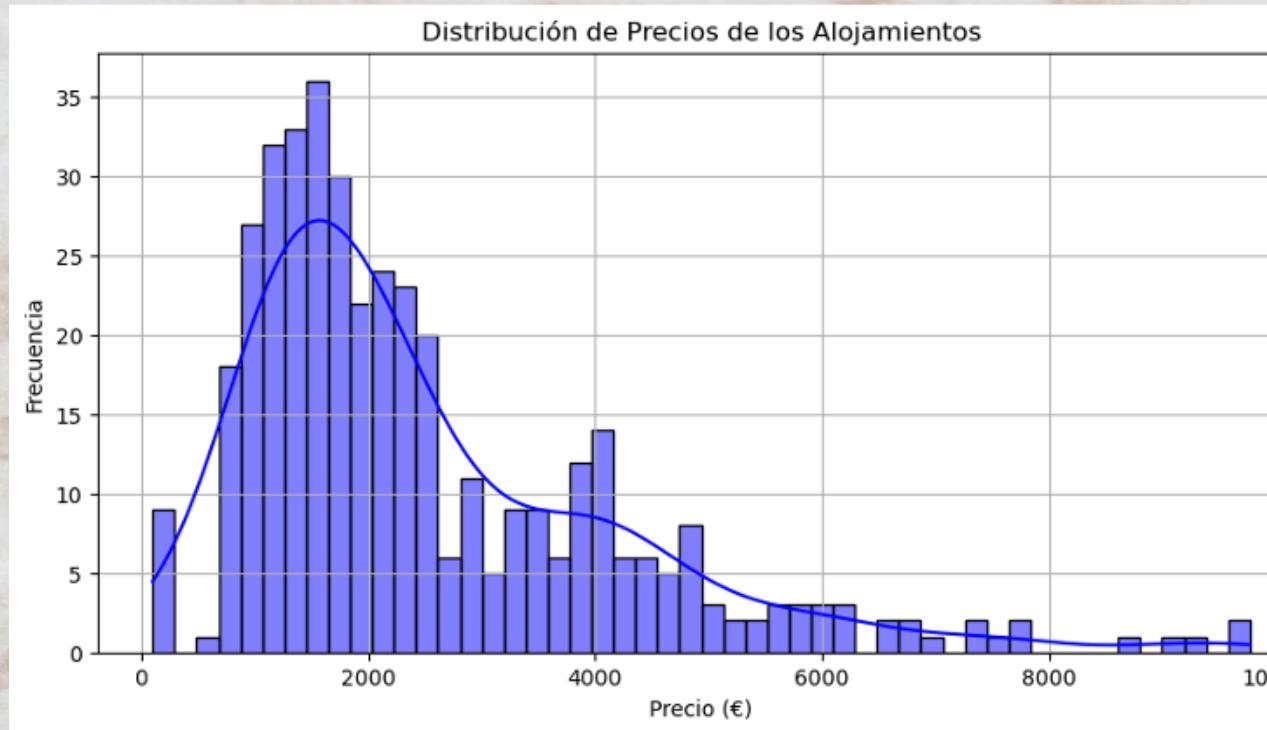
Primeras filas del DataFrame alojamientos_booking.csv

	Alojamiento	Precio	Puntuación	Reseñas	Ubicación	Estrellas
1	GO13D Apartamento Metro Bilbao WIFI	730,00 €	10	· 4 comentarios	Chamberí, Madrid	Sin estrellas
2	Hotel Fenix Gran Meliá - The Leading Hotels of the World	7.809,00 €	8.8	· 660 comentarios	Salamanca, Madrid	5
3	BM9 Piso 1 dormitorio Chamberí	795,00 €	7.3	· 7 comentarios	Chamberí, Madrid	Sin estrellas
4	VH75 Piso Patio Chamberí WIFI	735,00 €	5.8	· 5 comentarios	Chamberí, Madrid	Sin estrellas
5	Apartamentos Adelfas	943,00 €	7.6	· 273 comentarios	Retiro, Madrid	Sin estrellas
6	VR20I Estudio Centro WIFI	740,00 €	10	· 2 comentarios	Chamberí, Madrid	Sin estrellas
7	Apartamentos Cinco Rosas	760,00 €	7.4	· 57 comentarios	Carabanchel, Madrid	Sin estrellas

Primeras filas del DataFrame alojamientos_booking_limpio.csv

	Alojamiento	Precio	Puntuación	Reseñas	Ubicación	Estrellas	Distrito	Ciudad
1	GO13D Apartamento Metro Bilbao WIFI	730.0	10.0	4	Chamberí, Madrid	0.0	Chamberí	Madrid
2	Hotel Fenix Gran Meliá - The Leading Hotels of the World	7809.0	8.8	660	Salamanca, Madrid	5.0	Salamanca	Madrid
3	BM9 Piso 1 dormitorio Chamberí	795.0	7.3	7	Chamberí, Madrid	0.0	Chamberí	Madrid
4	VH75 Piso Patio Chamberí WIFI	735.0	5.8	5	Chamberí, Madrid	0.0	Chamberí	Madrid
5	Apartamentos Adelfas	943.0	7.6	273	Retiro, Madrid	0.0	Retiro	Madrid
6	VR20I Estudio Centro WIFI	740.0	10.0	2	Chamberí, Madrid	0.0	Chamberí	Madrid
7	Apartamentos Cinco Rosas	760.0	7.4	57	Carabanchel, Madrid	0.0	Carabanchel	Madrid

3. Análisis Exploratorio de Datos (EDA)

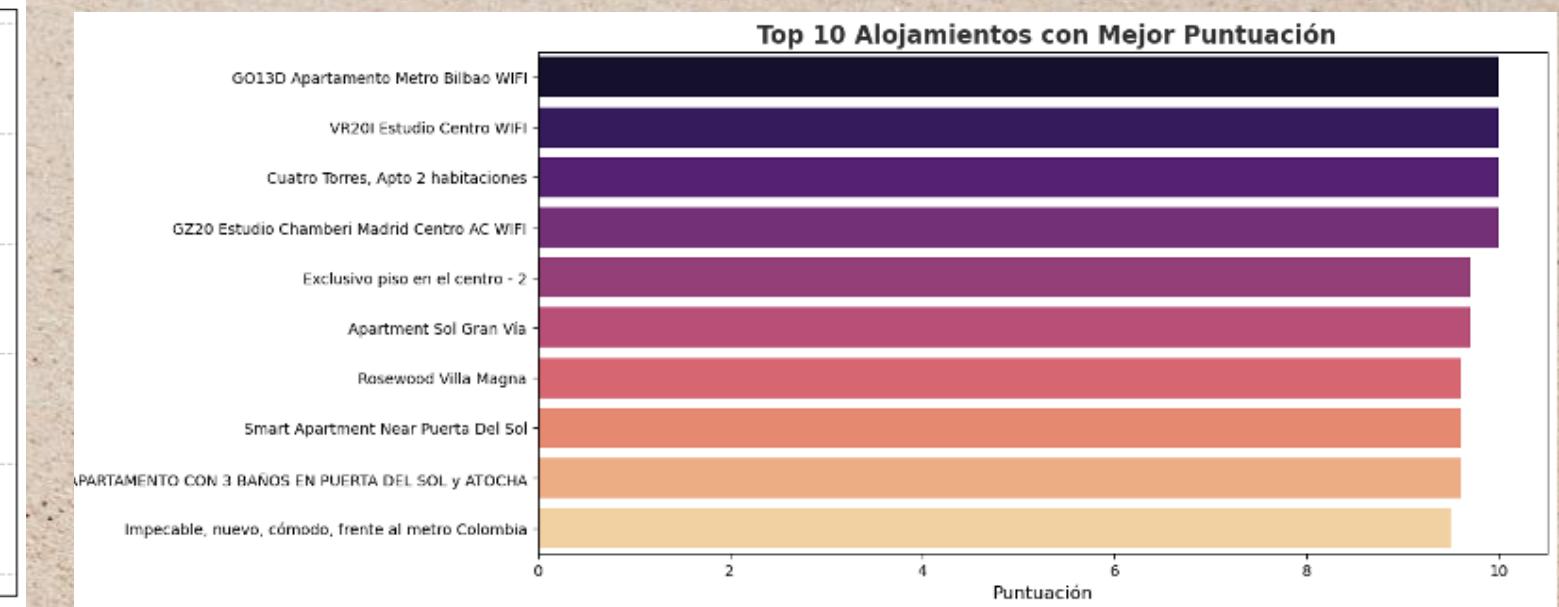
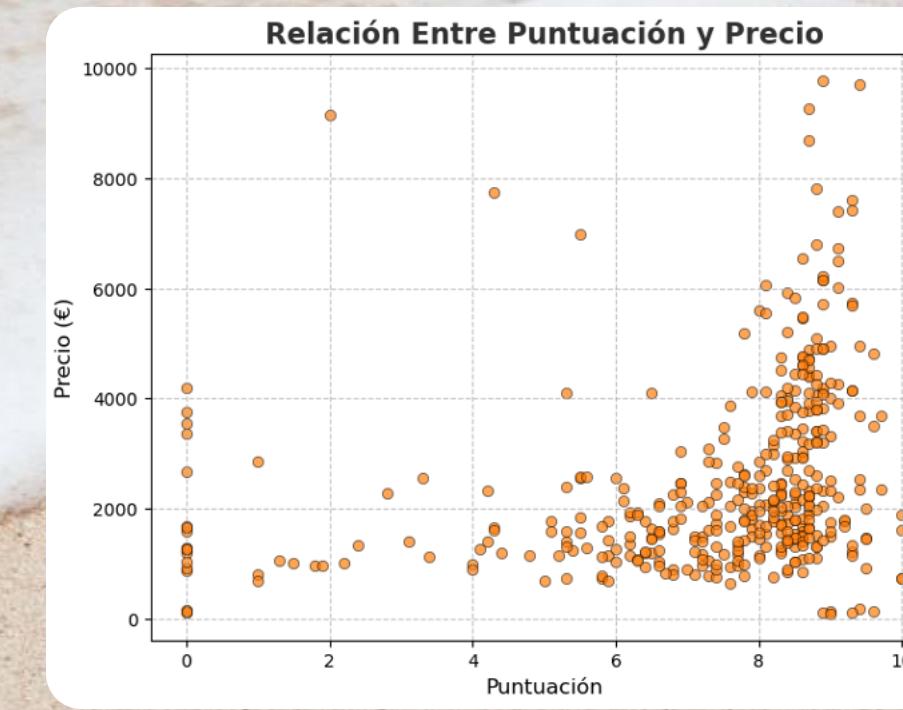
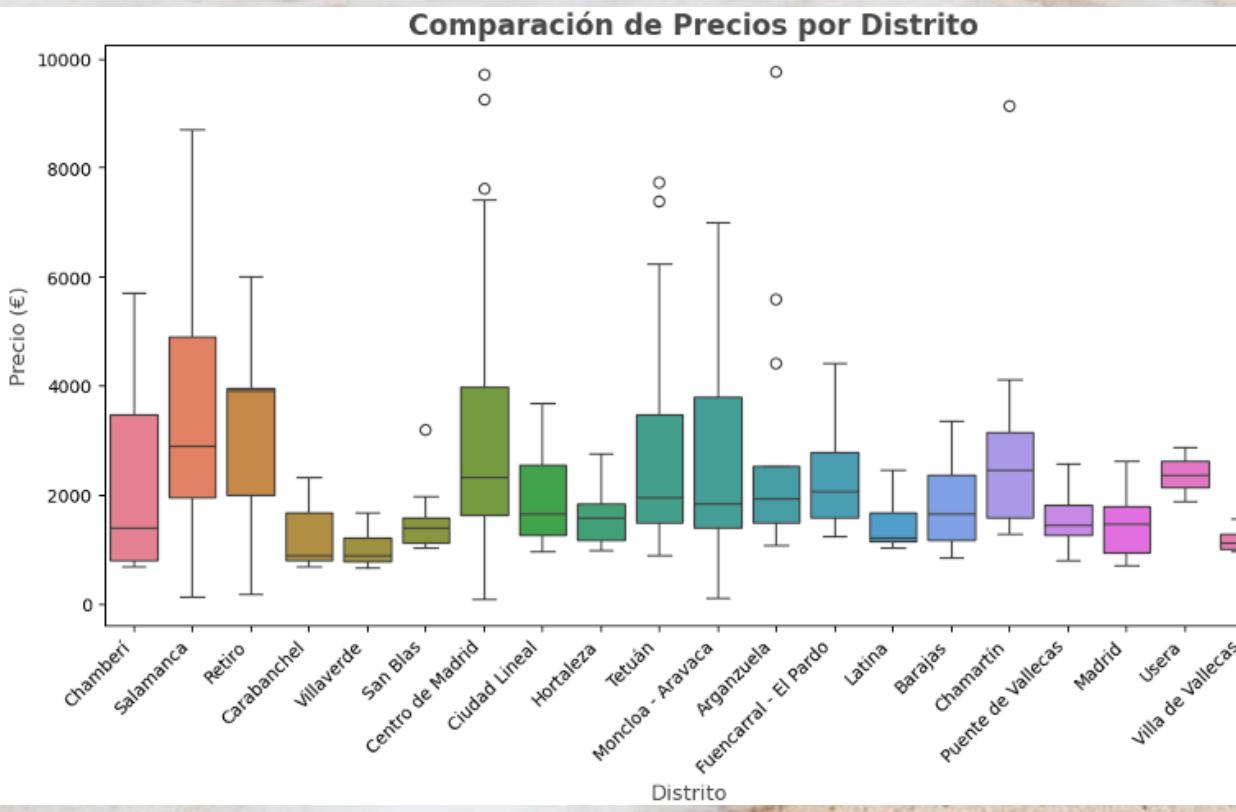


El histograma de distribución de precios de los alojamientos muestra una gran variabilidad en los costos. Se observa que la mayoría de los alojamientos se concentran en un rango de precios entre 1000€ y 3000€, lo que indica que este es el segmento más accesible para los viajeros.

El boxplot confirma que los alojamientos de mayor categoría tienden a ser más costosos. Los hoteles de 5 estrellas presentan las tarifas más altas, con una mediana superior a 4.000€, mientras que los de 1 y 2 estrellas tienen precios más bajos y menos variables. Por otro lado, los alojamientos sin clasificación muestran una amplia dispersión de precios.

El mapa de calor muestra que los alojamientos con más estrellas suelen ser más costosos (correlación de 0,55) y reciben más reseñas, probablemente por su mayor visibilidad. Sin embargo, la relación entre precio y puntuación es baja (0,26), lo que indica que un mayor costo no garantiza una mejor valoración.

3. Análisis Exploratorio de Datos (EDA)



El análisis de precios por distrito en Madrid muestra contrastes marcados: Salamanca, Centro y Chamberí destacan por sus altos costos debido a su exclusividad, mientras que Puente de Vallecas, Usera y Villa de Vallecas ofrecen opciones más costosas. Esta variabilidad refleja la influencia de la ubicación, la demanda turística y la categoría del alojamiento.

El gráfico de dispersión muestra que el precio y la puntuación no están fuertemente correlacionados. Aunque algunos alojamientos costosos tienen altas valoraciones, también hay opciones económicas bien puntuadas, lo que indica que los usuarios valoran factores como ubicación y comodidad más allá del precio.

El gráfico muestra los 10 alojamientos mejor valorados en Booking, con reducción cercana a 10. Destacan tanto apartamentos como hoteles , indicando que la alta valoración no depende del tipo de alojamiento. La ubicación céntrica y servicios como WiFi y comodidad parecen ser factores clave en la satisfacción de los huéspedes.

4. Panel Interactivo



- Mapa Interactivo → Ubicación de alojamientos
- Ranking de distritos → Mejores zonas para hospedarse
- Histograma de precios → Rango de precios más frecuenteRelación entre Puntuación y Precio

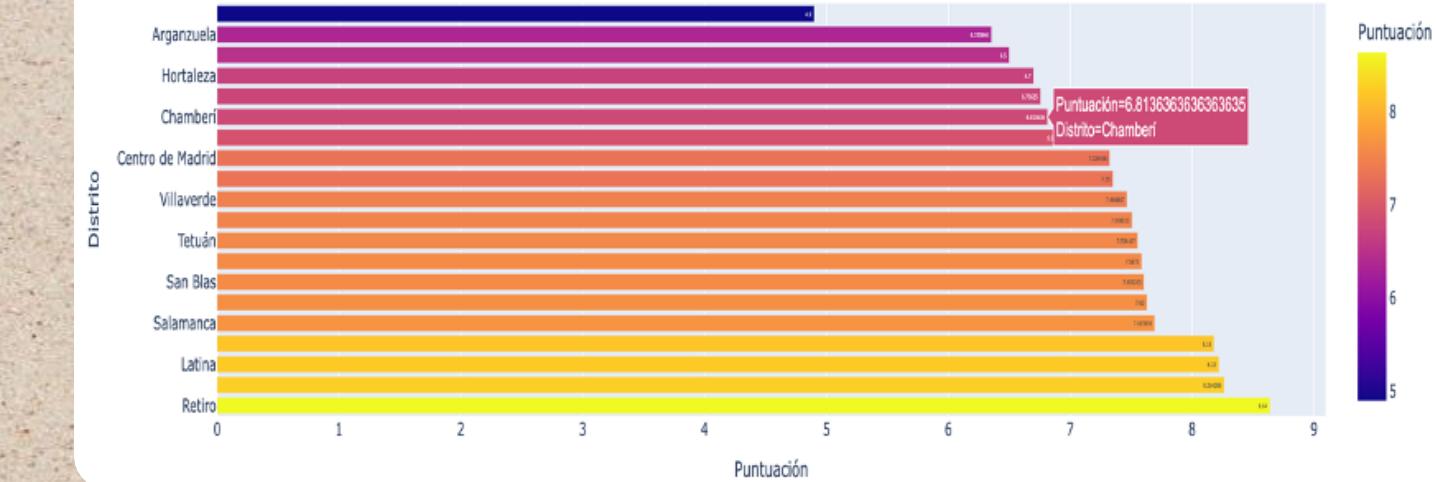
Cuadro de Mando - Alojamientos en Madrid

Selecciona un Distrito:
Todos

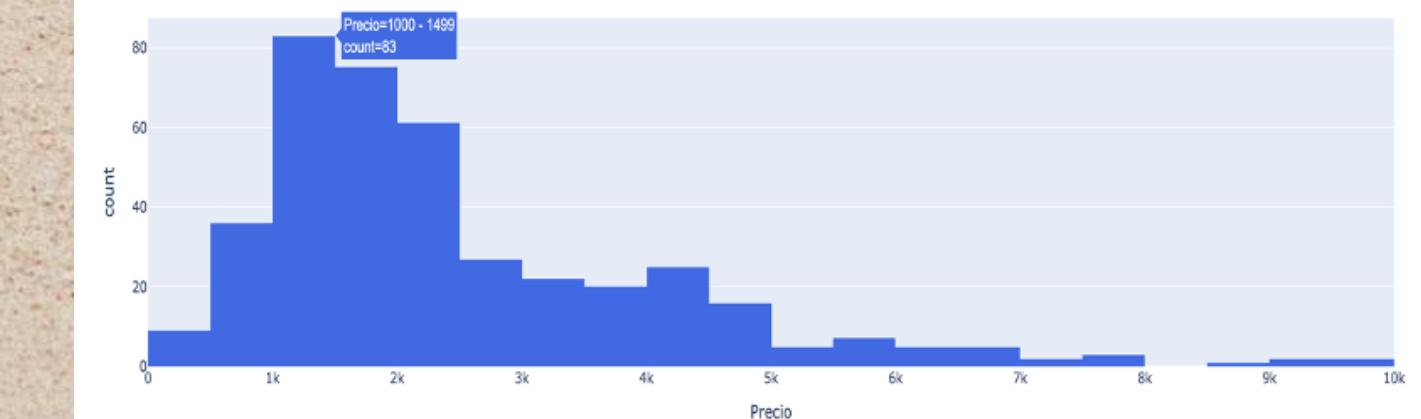
Ubicación de Alojamientos en Madrid



Ranking de Distritos con Mejores Puntuaciones



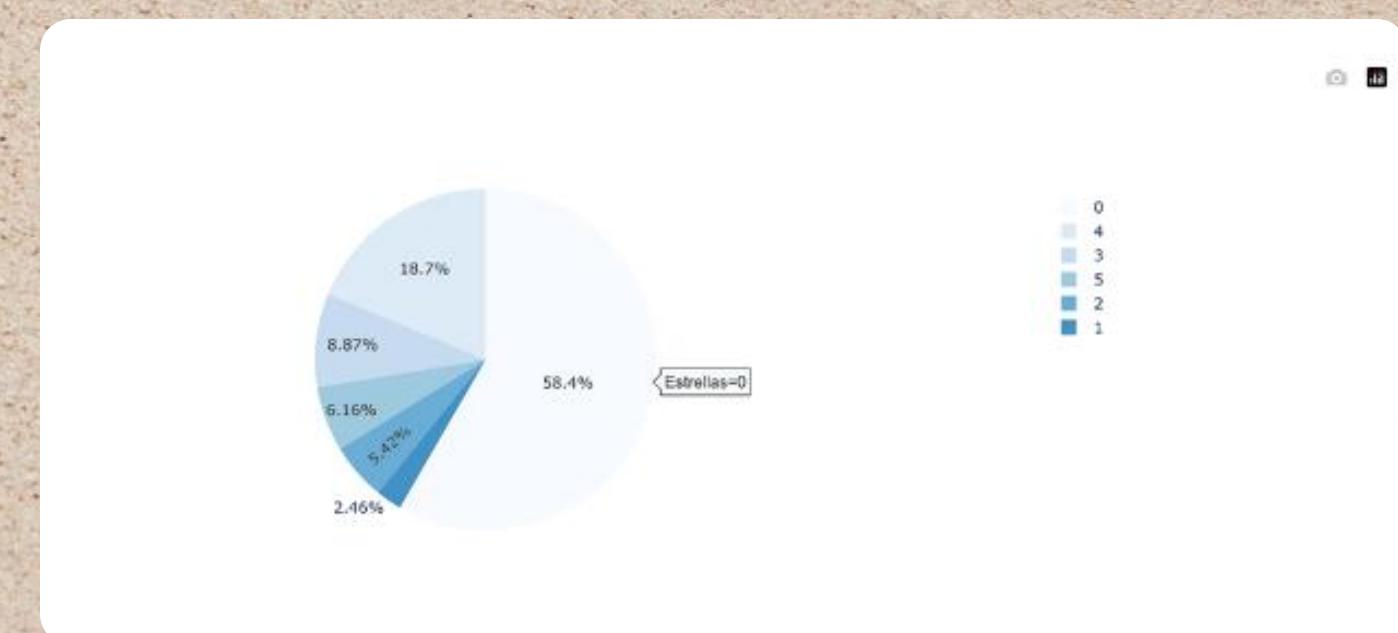
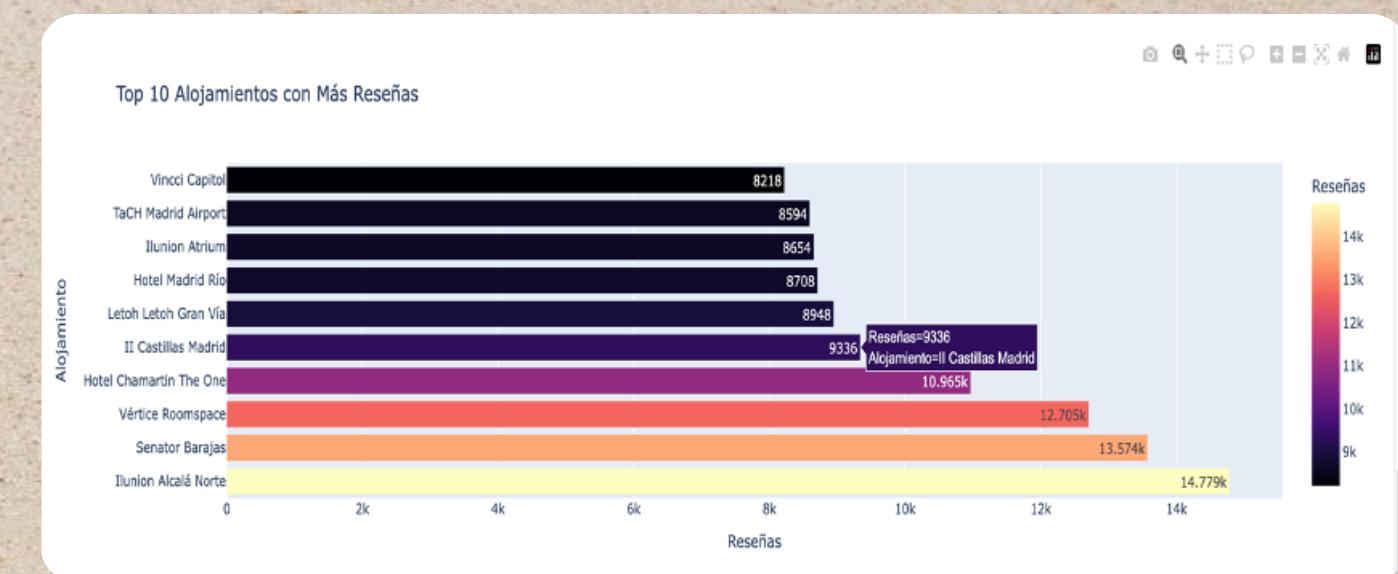
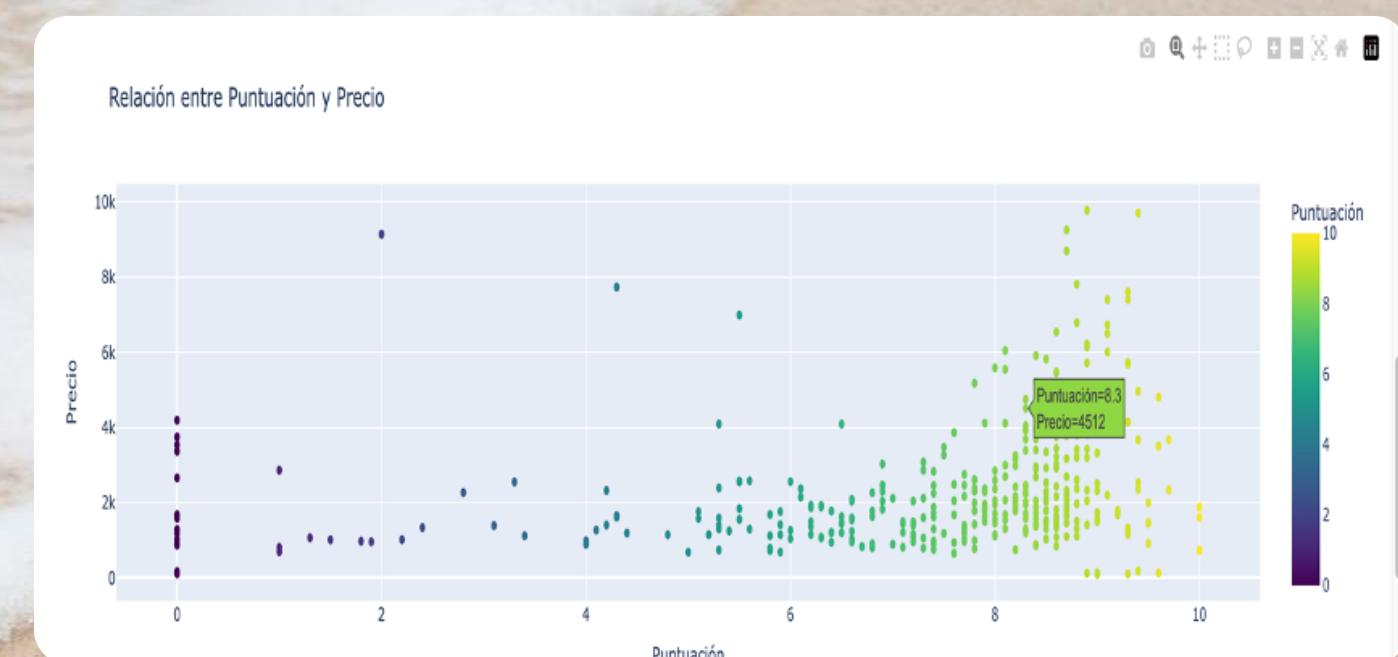
Distribución de Precios



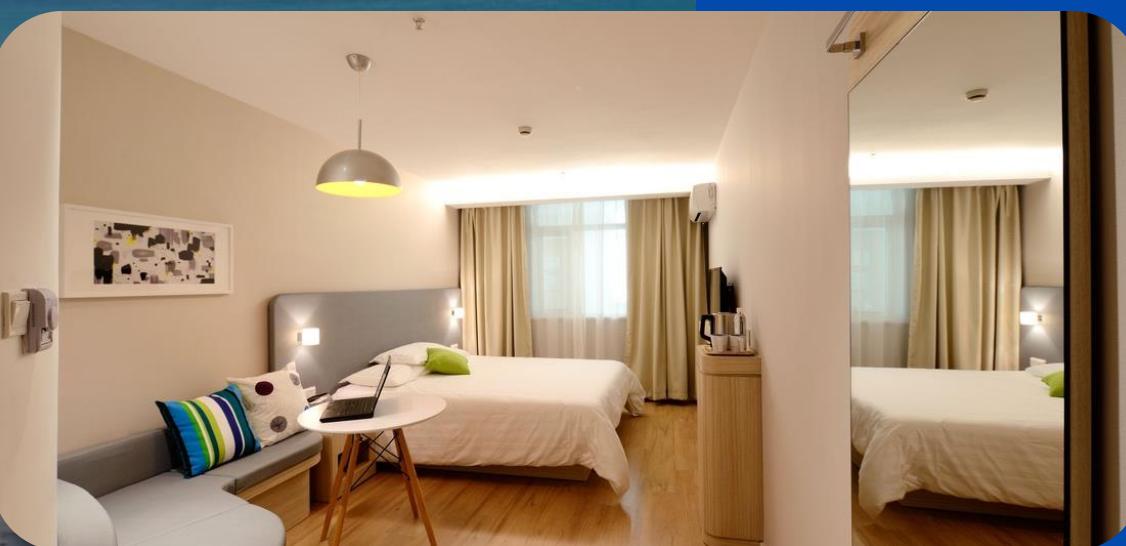
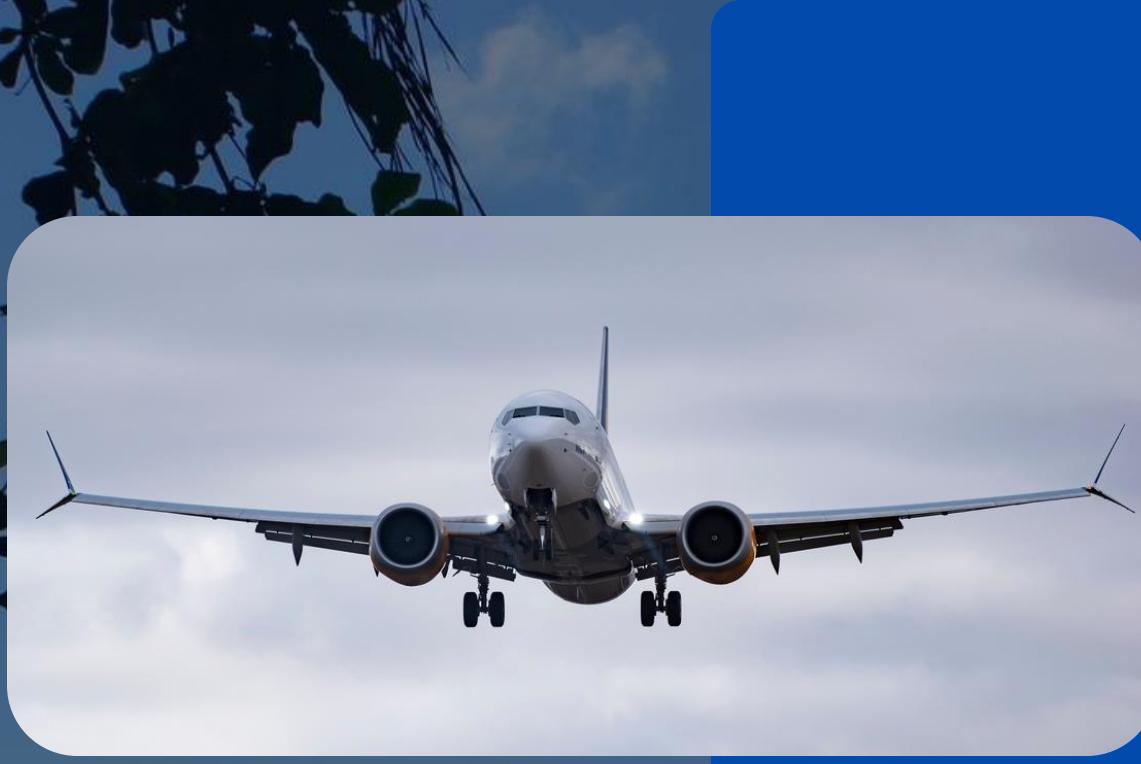
4. Panel Interactivo



- Relacion entre Puntuación y Precio
- Top 10 hoteles con más reseñas → Alojamientos más populares
- Distribución de Estrellas
- Top 10 Hoteles Más Caros
- Top 10 Hoteles con Mejor Puntuación
- Top 10 Hoteles con Más Estrellas



Conclusiones



El análisis de datos de alojamientos en Madrid permitió identificar tendencias claves en precios, reducción y distribución geográfica. La mayoría de los alojamientos tienen precios entre 1.000€ y 3.000€ , aunque algunos superan los 9.000€ . No se encontró una relación clara entre el precio y la puntuación, lo que indica que alojamientos económicos pueden ofrecer experiencias de calidad. Además, la mayoría de los hoteles tienen valoraciones superiores a 7.0 , lo que sugiere una percepción positiva por parte de los usuarios.

Durante la implementación del web scraping , se enfrentarán desafíos técnicos debido a las restricciones de Booking.com contra la automatización. Se utilizaron estrategias en Selenium para evitar bloqueos y se implementó desplazamiento automático y navegación entre páginas para obtener un conjunto de datos más amplio.

A pesar de estos desafíos, el tablero desarrollado permite explorar la oferta hotelera de manera dinámica e interactiva. Como mejoras futuras, se podría integrar API oficiales , optimizar el proceso de scraping y mejorar la geolocalización de los alojamientos. Este análisis proporciona información valiosa para turistas y empresarios del sector, facilitando la toma de decisiones basada en datos.

Gracias!

