Descripción de las transformaciones y la limpieza de datos – StaySpain

Conexión y extracción de datos: los datos se extrajeron directamente de una base de datos MySQL. La tabla *Tourist_Accommodation* se cargó como un DataFrame de pandas.

Conversión de tipos de datos

Se realizaron conversiones para asegurar que las columnas tuvieran los tipos de datos adecuados para el análisis:

Las fechas (first_review_date, last_review_date, insert_date) se convirtieron utilizando pd.to_datetime y errors='coerce' para asegurar que los valores no válidos se convirtieran en NaT (null).

Los campos booleanos, como `has_availability` e `is_instant_bookable`, se estandarizaron (True) y se convirtieron al tipo booleanos.

Las columnas numéricas (baños, dormitorios, camas) se convirtieron al tipo Int64 para permitir valores nulos y garantizar la compatibilidad con el análisis estadístico.

Justificación: El uso de los tipos correctos facilita la validación, permite el uso de funciones y evita errores en análisis o modelos posteriores.

Eliminación de duplicados

Los duplicados se eliminan según el campo `apartment_id`, conservando únicamente el registro más reciente (`insert_date`). Esto garantiza que exista una única versión de cada apartamento, con los datos más actualizados.

Limpieza de texto – Barrios: `neighborhood_name` y `neighborhood_district` se sometieron a una función de limpieza: reemplazo de caracteres problemáticos como  , eliminación de símbolos especiales, conservando únicamente letras, números, espacios, guiones y apóstrofes. Normalización del texto a mayúsculas iniciales (primera letra en mayúscula).

Justificación: Garantiza la estandarización visual y evita inconsistencias como «Centro» vs. «centro» o «El-Centro».

Estandarización de ciudades: creado un diccionario de correspondencia para unificar las variaciones en los nombres de las ciudades, considerando errores tipográficos y diferencias en la acentuación o el uso de mayúsculas. Solo se conservaron en el conjunto de datos las ciudades válidas y que existían en dataset.

Justificación: Garantiza la coherencia en el análisis geográfico y evita la duplicación de categorías debido a diferencias ortográficas.

Estandarización de país: normalización del texto a mayúsculas iniciales (primera letra en mayúscula).

Validación de registros: se creó una columna `is_valid` para identificar los registros con información esencial faltante (precio, ciudad, tipo de habitación). Solo se conservan los registros válidos para el análisis. En este caso todos.

Gestión de reseñas: los registros con `number_of_reviews` igual a 0, todas las columnas de reseñas se forzaron a NaN, respetando la lógica de las reseñas ausentes. El `reviews_per_month` se calculó dinámicamente utilizando la fecha de la primera o última reseña, proporcional al tiempo transcurrido.

Justificación: Garantiza la coherencia entre el número de reseñas y los datos de reseñas agregados, además de evitar imputaciones artificiales.

Normalización amenities: se convirtió en listas vacías cuando el valor estaba ausente o en blanco.

Validaciones

Se verificó que `minimum_nights` no fuera mayor que `maximum_nights`.
Consistencia de disponibilidad: Se verificó en los campos availability_30, availability_60, availability_90 y availability_365. Se corrigieron las fechas invertidas entre first_review_date y last_review_date.

Correcciones de is_instant_bookable

Se implementó la condición de que una propiedad solo puede ser de "reserva inmediata" si tiene disponibilidad (has_availability = True) y un precio mayor que cero.

Normalización de puntuaciones

Las puntuaciones de las reseñas (review_scores_*) se dividieron entre 10 para ajustarlas a una escala de 0 a 10.

Imputación de campos nulos

Se realizaron imputacione:

dormitorios = 1 para habitación privada.
baños = 1 si el valor es nulo.
camas = dormitorios si el valor es nulo.
reseñas_por_mes = 0 cuando número_de_reseñas = 0.
Los demás campos de reseñas se rellenaron con 0 cuando no había reseñas.

Creación de la métrica de ocupación

Se creó la columna ocupación_mes = 30 - disponibilidad_30, que representa una estimación de cuántos días estuvo ocupada la propiedad el mes pasado.

Detección de valores atípicos

Se revisaron las propiedades con precio superior a 6000 y capacidad superior a 25 camas. Se inspeccionaron estos registros para su posible eliminación o tratamiento, pero al final no fueron eliminados o tratados, se han mantenido.

Exportación final: El conjunto de datos limpio se preparó para guardarse en formato .csv.