Tecnológico de Monterrey

Selección y limpieza de datos

Miranda Eugenia Colorado Arróniz A01737027

14 de septiembre de 2025

Analítica de datos y herramientas de inteligencia artificial I (Gpo 101)

Dr. Alfredo García Suárez

1. Filtrado de variables y registros

El proceso comenzó con la carga del archivo original **listings.csv.gz**, que contiene información de alojamientos en Hawaii. Se importaron las librerías necesarias para el análisis y se revisaron las dimensiones del dataset para conocer la cantidad de datos disponibles.

Posteriormente, se realizó una selección de columnas para quedarnos únicamente con las variables más relevantes para el análisis. De las decenas de columnas originales, se eligieron 50 que incluyen datos como el identificador del alojamiento, nombre, ubicación, información del anfitrión, precios, disponibilidad, calificaciones, y otros aspectos clave. Esta selección se hizo para enfocar el análisis en lo que realmente aporta valor y evitar trabajar con información innecesaria.

El nuevo DataFrame filtrado se guardó en un archivo CSV para facilitar su uso posterior. Además, se restablecieron los índices del DataFrame para asegurar que cada registro tuviera un identificador único y ordenado, lo que ayuda a evitar confusiones y facilita el manejo de los datos.

Como parte del filtrado, también se realizaron subconjuntos de registros para responder preguntas específicas, como:

- Filtrar los registros de 5 anfitriones diferentes.
- Seleccionar los anfitriones que se unieron a Airbnb después de 2020.
- Filtrar los anfitriones que responden en máximo 1 día.
- Seleccionar filas que son múltiplos de 200 para obtener una muestra representativa.
- Filtrar alojamientos con más de 200 días de disponibilidad anual.
- Seleccionar alojamientos con precio menor a 100 dólares por noche.
- Filtrar los que son superhost.
- Seleccionar únicamente las columnas impares del DataFrame.

Cada uno de estos filtros permitieron trabajar con subconjuntos de datos más pequeños y enfocados en diferentes aspectos del negocio. Finalmente, se guardaron los resultados de esta etapa en el archivo **Filtrado.csv** para su análisis individual.

2. Imputación de valores nulos

Una vez filtradas las variables y registros, se procedió a identificar y tratar los valores nulos, es decir, las celdas vacías o con información faltante. Este paso es fundamental para evitar errores en los análisis posteriores.

Se realizó un conteo de valores nulos por columna para saber en qué variables había más problemas de información faltante. Luego, se aplicaron diferentes estrategias de imputación según el tipo de variable:

- Variables numéricas: Se analizó la distribución de cada variable. Si la variable tenía valores extremos (outliers), se usó la mediana para rellenar los nulos, ya que la mediana no se ve afectada por estos valores. Si la variable era más simétrica y sin outliers, se usó la media. Ejemplo: precio, calificaciones, disponibilidad.
- Variables de texto: Para variables como el nombre del anfitrión, la respuesta del anfitrión, la ubicación, etc., se rellenaron los nulos con el texto "Sin dato" para dejar claro que no había información disponible.
- Variables de fecha: En variables como la fecha de registro del anfitrión o las fechas de las reseñas, se usó la técnica de "propagación hacia adelante" (ffill), que consiste en tomar el último valor válido y repetirlo en las celdas vacías siguientes. Esto es útil cuando los datos tienen un orden temporal y es razonable asumir que el valor se mantiene hasta que cambia.
- Variables con muchos nulos: Si una columna tenía más del 80% de valores nulos y no era relevante para el análisis, se eliminó para evitar que afectara los resultados.
- Otros casos: Si la variable no entraba en ninguna de las categorías anteriores, se usó la técnica de "propagación hacia atrás" (bfill) como refuerzo.

Todo este proceso se realizó de manera automatizada recorriendo cada columna y aplicando la estrategia más adecuada. Al final, se verificó que ya no quedaran valores nulos en el DataFrame y se guardó el resultado en un nuevo archivo llamado **Valores_Nulos.csv**. Esto garantiza que el análisis posterior se realice sobre datos completos y confiables.

3. Tratamiento de valores atípicos

Los valores atípicos (outliers) son datos que se alejan mucho del rango normal y pueden distorsionar los resultados del análisis. Para identificarlos y tratarlos, se siguió un procedimiento detallado:

- Se seleccionaron únicamente las columnas numéricas del DataFrame para analizar los valores atípicos.
- Se calcularon los cuartiles (Q1 y Q3) y el rango intercuartílico (IQR), que es la diferencia entre el tercer y el primer cuartil.
- Se definieron los límites superior e inferior permitidos usando la fórmula estándar: límite superior = Q3 + 1.5*IQR, límite inferior = Q1 - 1.5*IQR.
- Se filtraron los datos para quedarse únicamente con los valores que están dentro de estos límites, eliminando los que se consideran atípicos.
- Se verificó cuántos valores nulos quedaban tras eliminar los outliers y se graficaron los resultados usando diagramas de caja para visualizar la distribución final de los datos.
- El DataFrame resultante, ya sin valores atípicos, se guardó en un nuevo archivo CSV para su uso posterior.

Este proceso asegura que los análisis y modelos que se construyan con estos datos sean más precisos y representativos de la realidad, evitando que valores extremos distorsionen las conclusiones.

En resumen, el proceso de selección y limpieza de datos incluyó la reducción de variables a las más relevantes, la creación de subconjuntos de interés, la imputación cuidadosa de valores nulos y la eliminación de valores atípicos. Cada paso se realizó de manera sistemática y justificada, asegurando que el conjunto de datos final sea confiable, completo y adecuado para el análisis y la toma de decisiones.