Tecnológico de Monterrey

Campus Santa Fe

Desarrollo de proyectos de análisis de datos (Gpo 307)

ETAPA 2 – Extracción, Limpieza y Transformación de Datos

Nombre:

German Gedovius Dalloz - A01782383

Samuel Aranzola Monroy - A01786577

ETAPA 2 – Extracción, Limpieza y Transformación de Datos

Proyecto: Factores que influyen en el precio de alojamientos de Airbnb

En esta etapa se preparó la base listings.csv.gz para su uso en el análisis exploratorio. El objetivo fue obtener un conjunto de datos limpio, consistente y apto para responder las preguntas del Project Charter.

Primero se realizó una inspección general del dataset, identificando más de 70 variables entre numéricas y categóricas. Se encontró que varias columnas estaban completamente vacías (neighbourhood_group_cleansed, calendar_updated, license), por lo que fueron eliminadas. También se identificaron valores nulos en variables descriptivas, que se conservaron solo si eran potencialmente útiles.

Se revisaron duplicados tanto en todo el DataFrame como por id, eliminándolos para asegurar registros únicos. Además, se identificaron inconsistencias en variables como accommodates, bedrooms y beds, ajustando valores imposibles o dejándolos como faltantes.

La variable price requería limpieza por venir como texto con símbolos, por lo que se eliminaron caracteres especiales y se convirtió a numérica. Las tasas host_response_rate y host_acceptance_rate fueron transformadas de porcentajes a valores entre 0 y 1. Las columnas de fecha se convirtieron a formato datetime para facilitar análisis temporales.

Para evitar que valores extremos distorsionan el análisis, se aplicó recorte por percentiles 1 y 99 en price. Posteriormente se generaron variables derivadas alineadas al proyecto: price_per_guest, occupancy_rate y revenue_per_guest, permitiendo comparaciones más precisas entre alojamientos.

Al finalizar, se guardó el archivo limpio como listings_clean.csv. Esta base depurada constituye el insumo principal para los análisis de la Etapa 3 y garantiza que el trabajo posterior se realice con datos coherentes y confiables.