Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Puebla

M1 Actividad 2 (Preparación de la base de Datos)

Osvaldo Terrazas Sánchez A01276389

Mizuki Aranzazú Uscanga Pineda A01737787

Jasiel Guillermo García Añorve A01424128

Materia:
Gestión de proyectos de plataformas tecnológicas

Docente:
Alfredo García Suárez
María Luisa Gómez Barrios
Martín González Vásquez

Introducción

Para esta actividad se trabajó con bases de datos pertenecientes a la empresa Airbnb, la cual es una plataforma en línea que permite a las personas ofrecer, buscar y reservar alojamientos en todo el mundo. Fundada en 2008, la compañía ha revolucionado el mercado del hospedaje al conectar a anfitriones, que desean alquilar sus propiedades con huéspedes que buscan una alternativa a los hoteles tradicionales. Utilizando un sistema de reseñas y calificaciones, Airbnb proporciona una experiencia más personalizada y, a menudo, más económica, fomentando la interacción entre personas de diferentes culturas y estilos de vida.

Las ciudades con las que se trabajó durante esta actividad fueron CDMX (México), Río de Janeiro (Brasil), y Chicago (Estados Unidos).

Tratamiento de valores nulos

Lo primero que se realizó fue la limpieza de los valores nulos en las tres bases de datos. Para poder realizar este paso se utilizó la función dtypes( ) para conocer el tipo de dato que se encontraba en cada una de las columnas. Al realizar esto nos dimos cuenta de que la columna "price" aparecía con datos de tipo objeto, así que tuvimos que cambiar el tipo de dato a numérico para que no afectara en algún futuro paso en donde se deban sacar estadísticas y correlaciones. Una vez terminado este paso decidimos eliminar los valores nulos dependiendo del tipo de dato que incluía cada columna. Para las columnas de tipo objeto sustituímos los valores nulos con strings en concreto (Ejemplo: Para los valores nulos de la columna "host_name" se colocó "ANÓNIMO"). Para las columnas de tipo numérico decidimos sustituirlas ya sea por un número en concreto o por la media de los datos, de esta forma al seguir con el paso de identificar los outliers los datos numéricos importados en valores nulos no significarían ningún problema.

Selección de variables relevantes

Las bases de datos contenían muchas variables que en realidad no nos servían para realizar análisis, entonces para simplificarlo decidimos seleccionar aquellas que consideramos relevantes, las cuales fueron:

- Last_scraped
- Source
- Name
- Host_url
- Host_name
- Host_since
- Host_location
- Host_response_time
- Host_response_rate
- Host_acceptance_rate
- Host_is_superhost
- Host_neighbourhood
- Host_verifications
- Host_has_profile_pic
- Host_identity_verified

- Neighbourhood_cleansed
- Property_type
- Room_type
- Accommodates
- Bathrooms_text
- Bedrooms
- Beds
- Amenities
- Price
- Has_availability
- Number_of_reviews
- Review_scores_rating
- Instant_bookable
- Calculated_hosts_listings_count
- Reviews_per_month

Eliminación de outliers

Una vez sustituidos los valores nulos, el siguiente paso fue eliminar los outliers. Este paso se realiza para que al momento de hacer análisis estadísticos los resultados no se vean alterados por aquellos pocos valores que están muy fuera de rango. El método que utilizamos para eliminar los outliers fue definir los límites superiores e inferiores utilizando tres desviaciones estándar, aquellos valores fuera de rango se reemplazaron con la media. Con esto las bases de datos están listas para poder ser utilizadas para diversos análisis.