
Alexander Penagos Muñoz A01738144

Domingo 7 de septiembre de 2025

**Actividad 5 (Extracción de Características)**

## Introducción

El objetivo de la actividad es aplicar técnicas de extracción de características para poder comprender la estructura de los datos listados de numéricos a categóricos. Este análisis es un análisis univariado de variables categóricas y categorización con Sturges de variables numéricas para ver distribuciones, concentración y tendencias centrales para la toma de decisiones.

## Datos

Se utilizó el archivo df_limpio el cual fue depurado de nulos y de datos atípicos.

Se crearon variables con columnas duplicadas en la cual se hizo una limpieza y se eliminaron estas columnas.

## Preparación de datos

Se hizo una normalización de las variables como porcentajes en host_response_rate, host_acceptance_rate en el cual venían con porcentaje y se convirtieron a decimal, además de convertir todo a escala 0 a 100 para poder hacer una comparación de todo y la categorización de Sturges.

En la variable Price se quitaron símbolos y separadores por comas para tener mejores datos y poder analizarlos.

De igual forma, se contó el número de faltantes y el porcentaje de valores nulos por columna sin rellenar registros. En las variables numéricas se aplicaron los intervalos de Sturges y se eliminaron los datos nulos.

## Metodología de análisis

Se eligieron 0 columnas de tipo objeto en las cuales se les hicieron un análisis de frecuencias, moda en categorías, categorías únicas y % de datos faltantes.

En las variables categóricas se revisaron los nulos y luego se procesaron con el método de Sturges en la que las variables tienen frecuencias por intervalos y gráficas de barras con sus estadísticas descriptivas.

## Resultados y hallazgos

**Variables categóricas**

La variable host_response_time tiene la mayor parte de los anfitriones con within an hour lo que indica una rapidez alta de respuesta a los usuarios. En segundo lugar, aparece con within a few hours y tercer lugar con within a day.

En la variable de host_is_superhost predomina el valor t que es verdadero el cual nos indica que hay una propoción significativa en superhosts lo cual indica que por lo general dan buena experiencia y tienen antigüedad.

Para la variable de host_identity_verified predomina que gran parte de los anfitriones están verificados lo que da seguridad y confianza a los huéspedes.

En la variable de property_type / room_type predomina Entire rental unit, con menor participación guest suites y townhouses lo que nos indica que hay personas que buscan privacidad.

En la variable de host_location y neighbourhood_cleansed todas las ubicaciones están en Washington, DC, y la oferta se concentra en barrios centrales (p. ej., Union Station–Stanton Park–Kingman Park, Capitol Hill) lo que implica que hay zonas de oferta mayores y atractivos urbanos.

**Variables numéricas**

En la variable de host_response_rate se cambió la escala de 0-1 a 0-100 la cual tiene una distribución en 100 y la mediana en 100 y sesgo hacia arriba ya que la mayoría se concentra en este valor.

En la variable de host_acceptance_rate tenemos una alta concentración en la cola alta con más dispersión en la cual la aceptación es elevada pero no todas las respuestas se representan en un 100%.

En la variable de host_total_listings_count hay asimetrías a la derecha en la que la mayoría de los anfitriones tiene pocos anuncios (1–3) y algunos operan más de 10 el cual indica que

el mercado prevalece anfitirones individuales o con pocas ofertas y con pocos casos de operadores grandes.

En la variable de accommodates la mayoría de los húespedes es entre 2 y 4 huéspedes con mediana en 3, lo que la oferta está para parejas o pequeñas familias.

En la variable de beds y bathrooms lo que predomina es 1 cama y 1 baño aunque algunos tienen 0 camas o 0.5 baños lo que confirma que la oferta es para pocas personas o familias pequeñas.

La variable de Price la mayoría de los precios se concentran en 90 y 155 USD, con mediana en 115 con pocos casos con precios más elevados los cuales serían alojamientos premium de 475 USD.

La variable de maximum_nights_avg_ntm el cual tiene una moda en el valor máximo de 1125 noches el cual nos dice que las estadías pueden ser cortas con mínimos de 1 noche hasta estadías largas que nos da disponibilidad de 3 años.

En la variable de availability_365 los valores varían mucho y tienen una mediana de 185 noches el cual indica que hay mucha disponibilidad a lo largo del año y otros con poca disponibilidad.

En la variable de number_of_reviews y reviews_per_month temenos una Buena cantidad de reseñas con más de la mitad en 128 reseñas y con un ritmo mensual de 1 a 3 reseñas en promedio.

En la variable de review_scores_value la mayoría de las calificaciones con 5 o cercanos a este el cual nos dice que la calidad de la satisfacción de los huéspedes es muy alta y consistente.

**Conclusiones**

el dataset de hospedajes en Washington, DC, muestra que los anfitriones en su mayoría son altamente receptivos y confiables, con tasas de respuesta y aceptación cercanas al cien por ciento en el que los tiempos de respuesta y la disposición a aceptar las reservas es algo muy importante. Esto junto a las altas proporciones de superhosts e identidades verificadas.

La clasificación de las variables indica que la mayor parte de los alojamientos están hechos para grupos pequeños, presentando precios que se encuentran en un rango accesible. Esto sugiere una oferta clara y homogénea. No obstante, la existencia de propiedades con tarifas más altas y numerosas reseñas indica que también hay segmentos establecidos de alojamientos de lujo que añaden variedad al mercado.

Un hallazgo clave es la capacidad de adaptación en las operaciones: los calendarios de disponibilidad son muy diversos y muchos anfitriones no imponen restricciones estrictas sobre la duración de las estancias. Esto convierte a la ciudad en un destino apto tanto para estancias turísticas cortas como para reservas más largas.

Por último, las valoraciones cercanas al máximo confirman que la experiencia de los huéspedes es constante y satisfactoria, lo que aumenta la competitividad de la oferta. En resumen, los resultados muestran que este mercado integra confiabilidad, accesibilidad y flexibilidad, tres elementos que explican su dinamismo y atractivo.

Se agarraron las siguientes variables categóricas las cuales son las siguientes: property_type", "property_type.1", "room_type", "host_is_superhost",

"host_identity_verified", "neighbourhood_cleansed", "host_response_time",

"instant_bookable", "has_availability", "host_location".

En la parte A, se hizo un dataframe de las frecuencias de las categorías de tipo de cuarto que más se renta en esta ciudad, con el fin de consrguir insights valiosos. r

Se realize la categorización de las variables numéricas con Sturges en la cual las variables previamente mencionadas se le hizo una limpieza a los datos que tengan caracteres especiales como la coma, signos de dinero y porcentajes.