A01735502

Introducción

El presente documento tiene como objetivo justificar y validar las decisiones metodológicas adoptadas durante el proceso de limpieza, categorización y análisis de un conjunto de datos de Airbnb correspondiente a Dinamarca, compuesto por 21,722 registros y más de 50 variables. La finalidad principal del análisis es preparar la información para la identificación de outliers, la categorización de variables y la posterior visualización de tendencias, asegurando que los datos sean confiables, consistentes y adecuados para su interpretación.

Selección de Variables Cuantitativas y Cualitativas

El dataset original contenía múltiples columnas que incluían tanto variables cuantitativas como cualitativas, además de campos innecesarios o duplicados como Unnamed: 0 y Unnamed: 0.1. Para optimizar el análisis, se decidió separar las variables en dos subconjuntos principales:

Cuantitativas: Se seleccionaron variables numéricas continuas o discretas. Entre estas se incluyeron variables como estimated_revenue_l365d, review_scores_value, calculated_host_listings_count, accommodates, bathrooms, beds, price, maximum_nights_avg_ntm, availability_365 y number_of_reviews. La selección de estas variables respondió a criterios de relevancia para la evaluación del desempeño de los alojamientos, la capacidad de acomodación, la disponibilidad y la satisfacción de los huéspedes.

Cualitativas: Se seleccionaron variables de tipo categórico o textual, como neighbourhood, room_type y amenities, que permiten analizar la distribución de características no numéricas de los alojamientos. Estas variables son fundamentales para generar insights sobre patrones de localización, tipología de habitaciones y servicios ofrecidos, aspectos clave en estudios de benchmarking y segmentación de mercado.

Limpieza de Datos y Manejo de Outliers

Se adoptó el método de desviación estándar para la identificación y tratamiento de outliers en las variables cuantitativas, aprovechando el tamaño considerable del dataset (21,722 registros). El proceso consistió en calcular para cada variable:

Valor máximo (Max) y valor mínimo (Min), generando una lista de límites por variable.

Rango (R), definido como R = Max - Min, permitiendo observar la dispersión de cada variable y detectar valores atípicos.

Este enfoque es particularmente útil en variables como price o maximum_nights_avg_ntm, donde ciertos alojamientos presentan valores extremos. La limpieza mediante desviación estándar contribuye a generar estadísticas más representativas y reduce sesgos en análisis posteriores.

Se reemplazaron variables que podían contener strings en formato porcentaje (host_response_rate, host_acceptance_rate) por columnas completamente numéricas (estimated_revenue_l365d, review_scores_value), asegurando consistencia y evitando errores de tipo.

Categorización de Variables Cuantitativas mediante la Regla de Sturges

Para optimizar la visualización y el análisis comparativo, se decidió categorizarlas, aplicando la regla de Sturges para determinar el número óptimo de intervalos. La fórmula utilizada es:

es el número de intervalos sugerido y n=21,722 es el tamaño de la muestra.

Aplicando la regla, se obtiene aproximadamente 15 intervalos, lo que justifica la creación de hasta 15 categorías posibles para cada variable cuantitativa.

y luego se generaron los puntos de corte equidistantes con np.linspace(Min1, Max1, 15) para obtener 14 intervalos efectivos.

Esta metodología asegura que los intervalos:

Capturen adecuadamente la variabilidad de los datos, evitando la concentración de muchos registros en pocos intervalos.

No sean excesivos, evitando sobrefragmentación que dificulte la interpretación.

Permitan un análisis estadístico coherente, especialmente para la visualización de frecuencias y patrones de distribución.

Definición de 14 Categorías Nombradas y Limitación

Aunque la regla de Sturges sugiere 15 intervalos, en la práctica se definieron 14 categorías con nombre. Esto introduce una limitación importante:

No permite agregar o renombrar categorías adicionales, lo que restringe la identificación precisa de ciertos rangos, especialmente los valores extremos.

La falta de un nombre para el decimoquinto intervalo puede dificultar la interpretación de registros que caen fuera de los 14 intervalos nombrados.

Visualización de Intervalos

Se generó un resumen de intervalos para cada variable cuantitativa, mostrando límites inferior y superior de cada categoría:

```
intervalos_df = pd.DataFrame({
  'Categoria': catregorias,
  'Limite_inferior': intervalo1[:-1],
  'Limite_superior': intervalo1[1:]
})
```

Esto permite vincular la información categórica con los valores numéricos originales y mantener la integridad del análisis. Cada categoría tiene un rango definido, garantizando que los datos puedan interpretarse tanto de manera cuantitativa como categórica.

Visualización y Análisis Gráfico

Se generaron 10 gráficos distintos, uno por cada variable cuantitativa categorizada, utilizando diferentes tipos de visualización:

Gráficos de barras: muestran la frecuencia de registros por categoría.

Gráficos de área: permiten visualizar la acumulación de frecuencias y tendencias generales.

Gráficos de pastel: representan la proporción relativa de cada categoría respecto al total.

La elección de cada tipo de gráfico consideró la naturaleza de los datos y la claridad visual. Por ejemplo, estimated_revenue_l365d se representó con barras para mostrar la concentración de alojamientos por rango de ingresos, mientras que review_scores_value se representó con pastel para resaltar proporciones relativas.

Justificación

El proceso implementado garantiza confiabilidad estadística, ya que la limpieza de los datos mediante desviación estándar permite minimizar la distorsión causada por outliers. Asimismo, la categorización de las variables utilizando intervalos definidos y la aplicación de la regla de Sturges facilita la interpretación de las tendencias presentes en los datos, otorgando claridad analítica. La metodología asegura reproducibilidad, dado que la enumeración sistemática de variables, intervalos y categorías permite replicar los análisis de manera consistente. Además, la selección de variables clave, como price, accommodates, availability_365 y number_of_reviews, ofrece relevancia práctica al proporcionar información útil para evaluar el desempeño de los alojamientos y la satisfacción de los usuarios. No obstante, la definición de solo 14 categorías frente a los 15 posibles limita la flexibilidad para identificar valores extremos, subrayando la importancia de planificar cuidadosamente la cantidad y denominación de categorías en futuros análisis para mantener la integridad interpretativa de los resultados.