**Analítica de datos y herramientas de inteligencia artificial II (Gpo 501)**

**Christian Jesus Soto-Vieyra Gil**

**A01707759**

**Actividad 5 (Evaluación)**

**ETAPA 1: Preprocesamiento de DATOS**

Lo primero que tuvimos que hacer es importar los tres datasets que se tenían de información respecto a las tres ciudades y convertir los tres en dataframes para poder trabajar posteriormente con estos datos, una vez que los cargamos observamos que nuestro primer DataFrame en este caso California, tiene un total de 6936 entradas y 75 columnas, en el caso del segundo df correspondiente a Distrito Federal, tenemos un total de 24224 entradas y 75 columnas y para el último df correspondiente a Girona, observamos un total de 19655 entradas y 75 columnas. Lo primero que teníamos que hacer es corroborar que no existían datos nulos que pudieran afectar nuestro análisis en el dataframe, sin embargo sí pudimos observar que existían columnas donde incluso toda la información era nula, por lo que procedimos a eliminar estas columnas, lo que nos dejó con 72 columnas en California, 71 en Distrito Federal , y 72 en Girona.

Posteriormente procedimos a buscar métodos para reemplazar los valores nulos de las otras columnas en nuestro caso, primeramente las que contenían información cuantitativa que en este caso decidimos emplear el metodo de el promedio de datos de la columna para sustituir el valor nulo de las variables, caso contrario para aquellas variables que eran cualitativas decidí aplicar métodos de backward y forward fill, es decir tomar el valor anterior o posterior a la instancia nula. de esta manera tomando no solo uno sino dos métodos corroboramos que ya no existían valores nulos y en efecto no existían.

**Aplicación de métodos (Desviación Estándar o Rango Intercuartil) para eliminar Outliers de cada dataframe y justificar el método aplicado**

Para poder analizar mejor nuestra información necesito eliminar los valores extremos que se presentan en nuestros dataframes, en este caso optamos por utilizar el método de Rangos Intercuartiles, esto lo que hace es detectar los valores por encima y por debajo de los percentiles 75 y 25. Escogí el método intercuartil debido a que este enfoque no es tan sensible a los valores extremos, por lo que esto nos puede proporcionar una mejor distribución de los datos. Esto también lo realizamos individualmente, para las variables cuantitativas y cualitativas, después de esto procedemos a unir nuevamente los datasets de ambas variables.

**ETAPA 2: Extracción de DATOS**

En este caso lo único que se hizo fue un filtrado de 10 condiciones para cada dataframe donde se impusieron diferentes condiciones: Obtener los registros que cumplan las siguientes condiciones para cada dataframe:

a) host\_acceptance\_rate> 50%

b) Sólo los registros con categoría “superhost”

c) Los registros que no hayan verificado identidad “ not identity\_verified”

d) Los registro cuyo property\_type = “Private room” y “Hotel room”

e) Los registros que cuenten con bathroom > 1

f) Los registros cuyo precio sea mayor de $10,000 y que sean de tipo “Entire home”

g) Los registros cuyo review\_scores\_cleanliness > 4.5

h) Los registros cuyo review\_scores\_value > 4.9

i) Los registros cuya availability\_365< 100

j) Los registros cuya host\_response\_time sea “within an hour”

**ETAPA 3: Extracción de Características y Visualización de DATOS**

Dentro de esta etapa tenemos nuevamente tres datasets de tres ciudades diferentes, California en Estados Unidos, Distrito Federal en México y Girona en España, lo primero que necesitamos realizar es un análisis comparativo de estas tres ciudades para el servicio AirBnb con las siguientes variables escogidas:

1. **host\_response\_time:**

Lo primero que tuvimos que hacer para esta variable es realizar una tabla de frecuencia de las tres ciudades empleando los datasets que anteriormente manipulamos en la etapa 1 y 2. Que nos dieron los siguientes resultados: Para verlo de mejor manera decidimos hacer tres gráficas de pastel de los datos:

Dentro de la información del tiempo de respuesta de la ciudad de California, observamos que el 69.2% de los anfitriones responden en un lapso de menos de una hora, seguidos por el 19.8% que responde en unas cuantas horas, 8.3% en un día y finalmente 2.6% responden alrededor de unos días.

En el caso del Distrito Federal, observamos que el 80.1% de los anfitriones responden en un lapso de menos de una hora siendo esta ciudad la de menor tiempo de respuesta, seguidos por el 9.4% que responde en unas cuantas horas, 7.1% en un día y finalmente 3.5% responden alrededor de unos días.

Finalmente para la ciudad de Girona observamos que el 70.3% de los anfitriones responden en un lapso de menos de una hora, seguidos por el 14.6% que responde en unas cuantas horas, 12.9% en un día y finalmente 2.2% responden alrededor de unos días.

1. **host\_acceptance\_rate**

En este caso debido a la cantidad de datos que se presentan decidimos hacer un rango de 0 a 100% pero separando los datos de 5 en 5. Así finalmente y con ayuda de nuestras gráficas podemos ver que:

En el caso de la ciudad de California arriba de 3000 anfitriones tienen un porcentaje de aceptación de entre 95 y 100%, seguido de alrededor de 600 con más de 90% de aceptación y teniendo 122 anfitriones con un porcentaje de aceptación de entre 0 y 5%.

Para el Distrito Federal observamos más de 18000 anfitriones con un nivel de aceptación de entre 95 y 100%, seguidos por 1594 anfitriones con más de 90%, y teniendo alrededor de 805 anfitriones con un porcentaje entre 0 y 5%.

Finalmente para el caso de Girona, observamos que 14796 anfitriones se colocan en un nivel de aceptación de entre 95 y 100%, seguidos de 1854 entre 95 y 100%, teniendo a final 434 anfitriones con un nivel bajo de aceptación de entre 0 y 5%.

1. **host\_identity\_verified**

Para la siguiente variable decidimos realizar además de tablas de frecuencias, una gráfica de pastel debido a que únicamente existen dos respuestas posibles, así finalmente pudimos observar que:

Para el caso de California 87.7% de los anfitriones están verificados por la plataforma mientras que el 12.3% no, siendo el número más bajo de verificaciones.

En el Distrito Federal 92.1% de los anfitriones están verificados, mientras 7.9% no, este siendo el número más alto de anfitriones verificados.

Finalmente en Girona 89% de los anfitriones están verificados, mientras el 11% no.

1. **property\_type**

En el caso de tipo de propiedad, existen demasiadas variables, sin embargo observamos que en el caso de las tres ciudades, la propiedad con más frecuencia es una Unidad Total en Renta. En casos individuales:

Las tres propiedades más frecuentes en California son: Entire Rental Unit (1518), Private Room in Home (664) y Entire Home (625).

Las tres propiedades más frecuentes en el Distrito Federal son: Entire Rental Unit (8145), Entire Condo (3363) y Private Room in Rental Unit (2860).

Las tres propiedades más frecuentes en Girona: Entire Rental Unit (10738), Entire Home (4227) y Entire Vila (1963).

1. **availability\_365**

En el caso de California observamos que 256 propiedades tienen disponibilidad prácticamente todo el año, mientras que 889 no lo tienen ningún día, esto la convierte en la ciudad con más nivel de ocupación por año, ya que es difícil encontrar alojamiento disponible.

En el caso del distrito federal 2078 propiedades están vacías todo el año, mientras que 1112 tienen disponibilidad agotada todo el año, números que nos dicen que prácticamente el doble de propiedades están disponibles que las agotadas.

Finalmente para Girona 416 propiedades están desocupadas todo el año y 1390 se mantienen ocupadas todo el año, lo que también vuelve a Girona como una ciudad con gran dificultad de disponibilidad de alojamiento.