Act 3.4 Extracción de Característica

Instituto Tecnológico de Estudios Superiores de Monterrey

ANALÍTICA DE DATOS Y HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL I

Grupo: 101

Profesor:

Alfredo Garcia Suárez

Autores:

César Alejandro Rivera Guzmán A01567012

Julio Alejandro Sotero Montiel A01656310

Diego Soto Camacho A01732608

Fecha de entrega:

30 de septiembre 2024

Rio de Janeiro

Con los métodos para convertir variables cuantitativa en variables categóricas, se realizó este procedimiento para 12 variables que son las siguientes:

1. host_response_rate
2. host_acceptance_rate
3. host_total_listings_count
4. accommodates
5. bathrooms
6. beds
7. price
8. 'availability_60'
9. 'availability_90'
10. 'availability_365'
11. 'number_of_reviews_ltm'
12. 'reviews_per_month'

De las cuales se utilizó la variable **price** cómo variable de estudio, se aplicó el método de categorización, dándonos un total de 12 categorías, al tratarse de los precios de los airbnb´s, se decidió categorizar los precios dependiendo el rango de precio, quedando de la siguiente manera:

```
#Creamos las categorias
categorias = [
    '0-19', '20-40', '41-60', '61-81', '82-101','102-122','123-142','143-163','164-183','184-205'
]
```

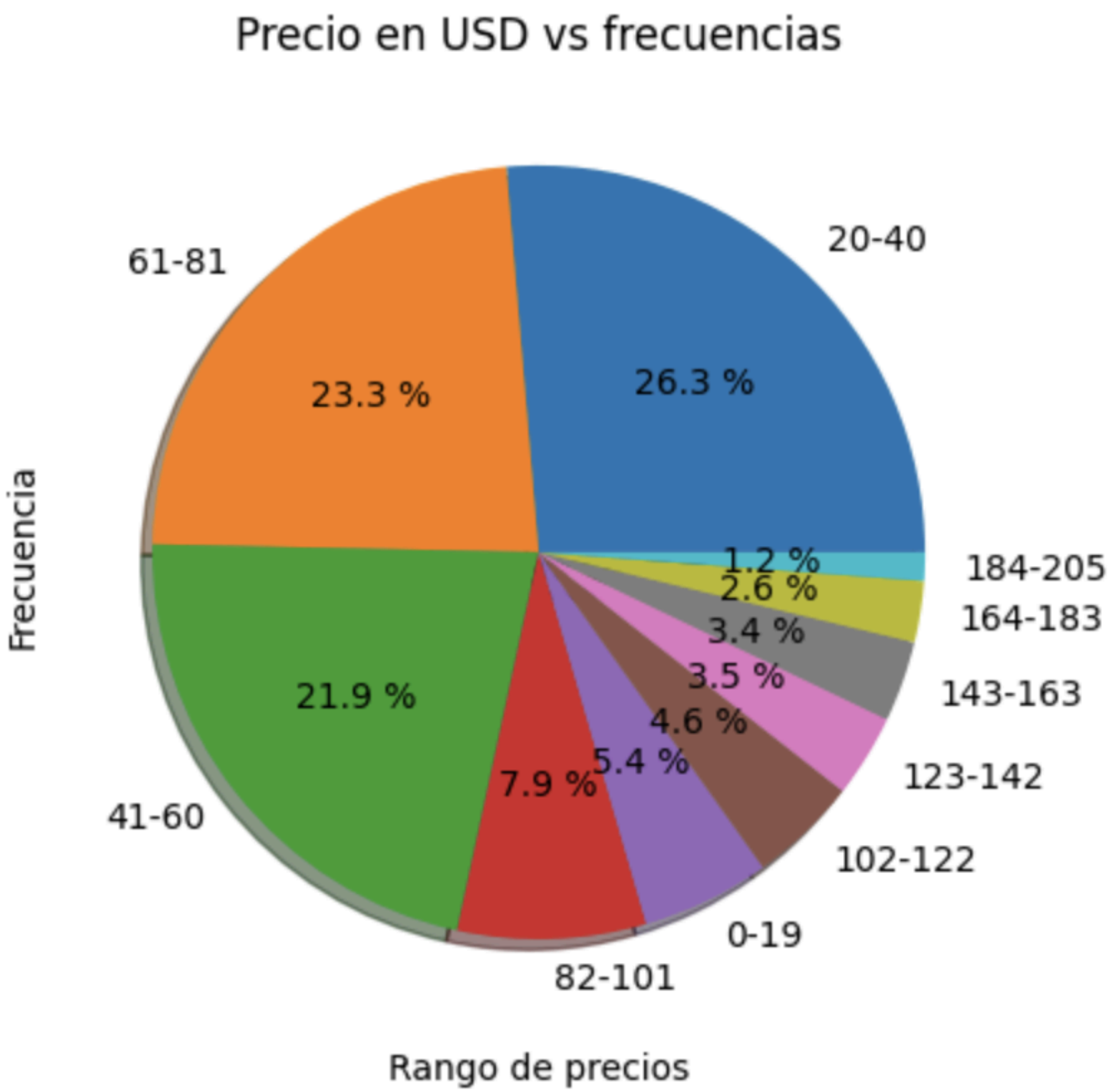
	price	frequency	percentage	cumulative_perc
0	20-40	9129	0.263357	0.263357
1	61-81	8090	0.233383	0.496740
2	41-60	7603	0.219334	0.716074
3	82-101	2734	0.078871	0.794946
4	0-19	1860	0.053658	0.848604
5	102-122	1579	0.045552	0.894155
6	123-142	1207	0.034820	0.928975
7	143-163	1167	0.033666	0.962641
8	164-183	887	0.025589	0.988230
9	184-205	408	0.011770	1.000000

Posterior a eso se eliminaron las columnas de percentage y cumulative_perc para únicamente trabajar con el rango de precios y las frecuencias, se anexó el precio con la finalidad de poder ordenar el DF de manera que sea fácil graficar, el proceso quedó de la siguiente manera:

price	frequency
20-40	9129
61-81	8090
41-60	7603
82-101	2734
0-19	1860
102-122	1579
123-142	1207
143-163	1167
164-183	887
184-205	408

Tabla de precios USD Tabla 1.1

Una vez limpiada el nuevo DF, se utilizaron diversas gráficas para poder observar el comportamiento de los datos, a continuación se mostrarán las gráficas utilizadas



Hallazgos:

Con estas gráficas se pudo observar el comportamiento de los precios de las habitaciones en Río de Janeiro.

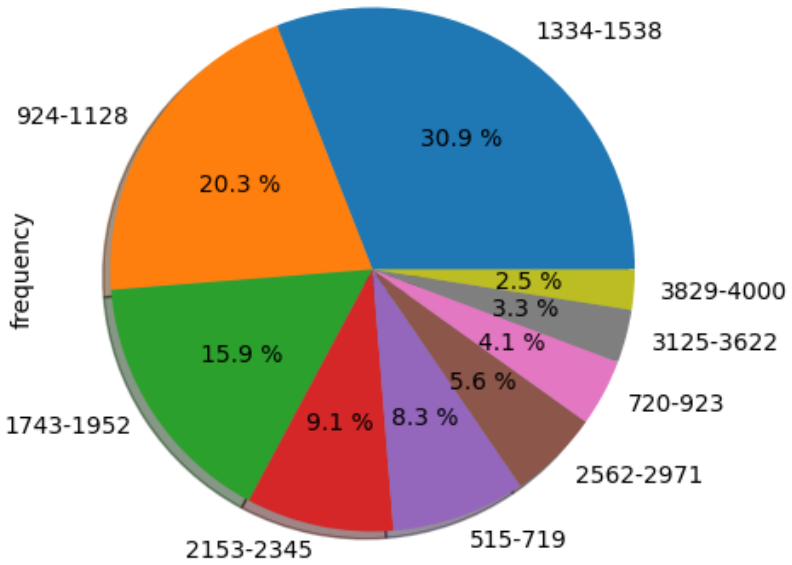
1. El rango de precio más común se encuentra entre $20 dlls y $40 dlls por noche.

2. En segundo lugar, está el rango de $61 dlls a $81 dlls
3. El tercer lugar se encuentra entre los $41 dlls a $160 dlls por noche.

Málaga

Se utilizaron 14 columnas, de las cuales la mayoría eran cuantitativas, sin embargo, se utilizó la regla de sturges para convertirlas a cualitativas. Se utilizó la librería funpymodeling para obtener las tablas de frecuencia de cada categoría y se hicieron los gráficos de barras de cada tabla eliminando los porcentajes y los porcentajes acumulativos. A continuación se muestran los resultados.

1. host_response_rate
2. host_acceptance_rate
3. host_total_listings_count
4. accommodates
5. beds
6. price
7. maximum_nights_avg_ntm
8. availability_60
9. availability_90
10. availability_365
11. number_of_reviews
12. number_of_reviews_ltm
13. review_scores_value
14. reviews_per_month



Hallazgos:
El análisis indica que los precios más bajos tienen una mayor frecuencia en Malaga es el rango de 1334 a 1538, mientras que los precios más altos tienen una menor frecuencia junto con los precios más bajos.

México

Trabajamos con 12 columnas, las cuales categorizamos:

- host_response_rate
- host_acceptance_rate
- host_total_listings_countaccommodates
- bathrooms
- beds
- price
- 'availability_60'
- 'availability_90'
- 'availability_365'
- 'number_of_reviews_ltm'
- 'reviews_per_month'

host_response_rate

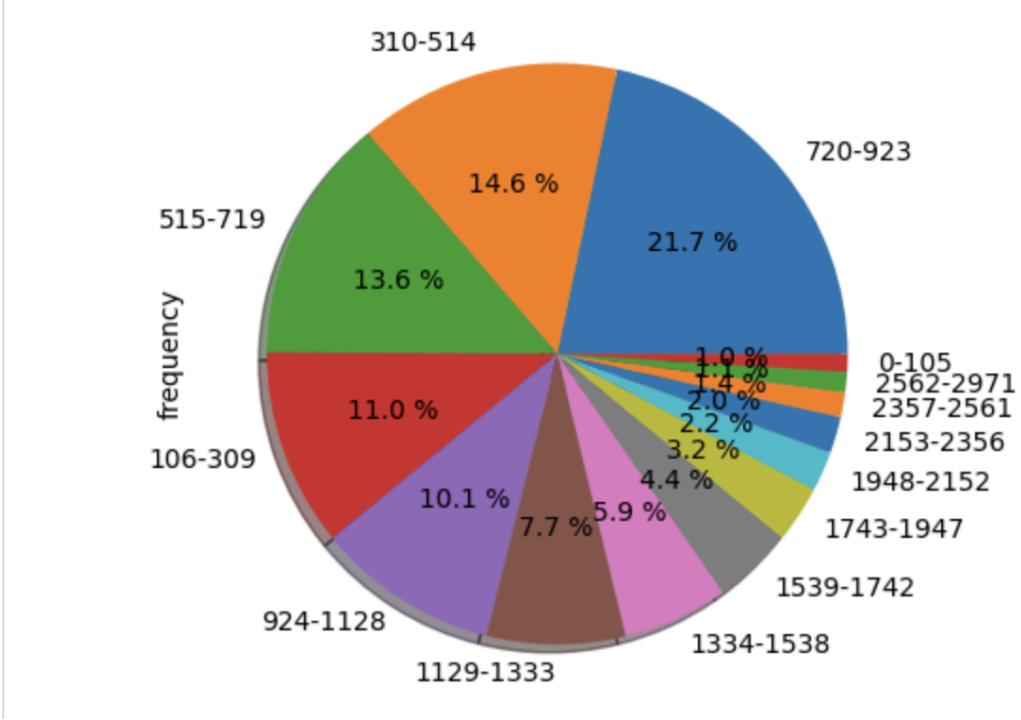
```
x='host_response_rate'
#Calculamos el numero total de la poblacion "n"
city[x].info()
n1=len(city)
#Obtenemos el limite superior y el limite inferior de la colmna objetivo
max=city[x].max()
min=city[x].min()
limites= [min, max]
#Calculamos el Rango R
r=max-min
#Calculamos el numero de intervalos de clase "ni", aplicando la regla de Sturges
ni= 1+3.32*np.log10(max)
#Calculamos el Ancho del Intervalo "i"
i=r/ni
```

Decidimos hacer un solo cuadro de código para el modelo explicativo de variables numéricas, esto para evitar alargar el documento, simplificarlo y comprender de mejor manera el código

donde:

- calculamos el número total de la población "n"
- obtuvimos el límite superior e inferior de la columna con la que queríamos trabajar
- calculamos R
- calculamos el número de intervalos de clase ni, usando Sturges
- calculamos el ancho del intervalo i

Gracias a la librería funpymodeling obtuvimos algunas gráficas, donde mostramos la columna de “price”



Hallazgos:

En general, podemos apreciar el precio comparado con la frecuencia, donde encontramos que hay bastante volatilidad en el rango de precios. Siendo de $720-$923 la frecuencia más común en México. Algo interesante es que encontramos más caro rentar una habitación en CDMX que en Río de Janeiro. Tenemos en cuenta que el precio es en pesos mexicanos. Más adelante, trabajaremos en la corrección, conversión y comparación de las monedas, para un mejor análisis comparativo entre las ciudades.