↘ACTIVIDAD 5

EVALUACIÓN

RENATA PILAR GÓMEZ CASTILLO

A01351806

MAYO 2023

ACT. 5 EVALUACIÓN

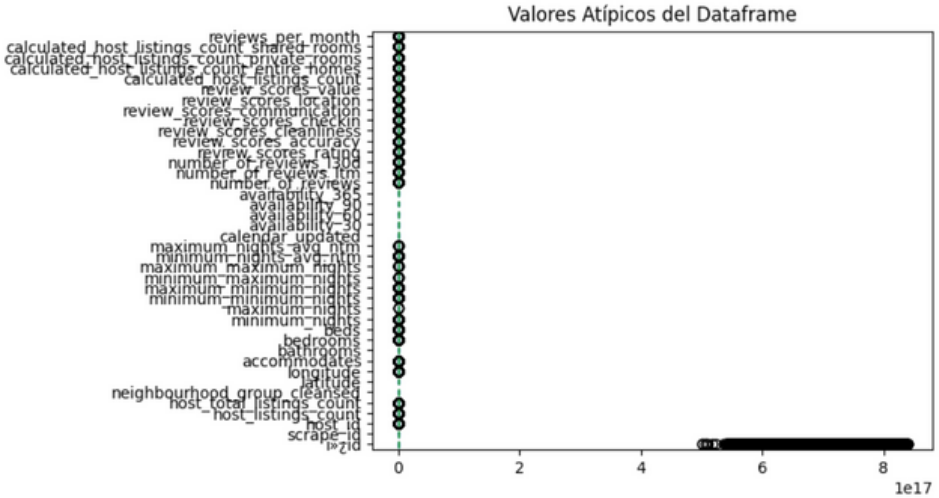
Contamos con tres Data Frames con los que estaré trabajando para el desarrollo de la actividad. Los Data Frames hacen referencia a datos obtenidos de Airbnb de tres diferentes ciudades, DF México, California EUA y Girona España.

Etapa 1 – procesamiento de datos

Los tres Data Frames cuentan con valores nulos y outliers los cuales tenían que ser tratados. Para esto primero separe cada Data Frame en dos, una para los valores numéricos y otra para los valores cualitativos.

Para los valores nulos decidí utilizar la función fillna(), donde los valores nulos de variables numéricas fueron reemplazados por un 0 y los de variables cualitativas con un string '--'. Esto con el fin de no alterar los datos e información que tienen los Data Frames originales. Ya que futuramente haré un análisis de las variables y no quiero que la información se vea afectada o que esté errónea.

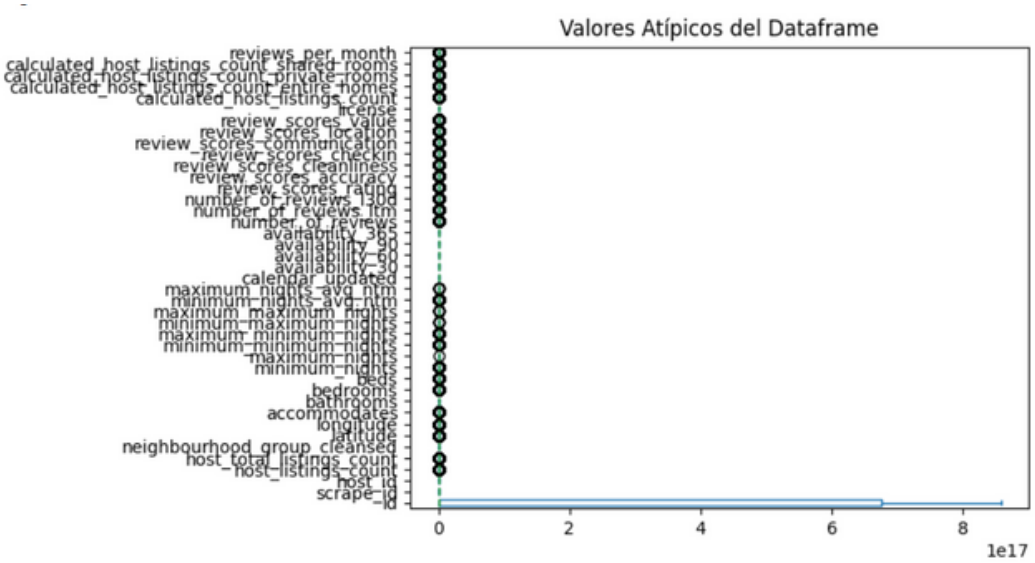
Mientras que para los valores atípicos decidí hacer uso de dos modelos que me ayuden a procesar los outliers.



Para el Data Frame de California decidí utilizar el método de desviación estándar. El cual utiliza la desviación estándar como una base que ayuda a identificar los valores atípicos de los datos numéricos que tienen las columnas del dataframe.

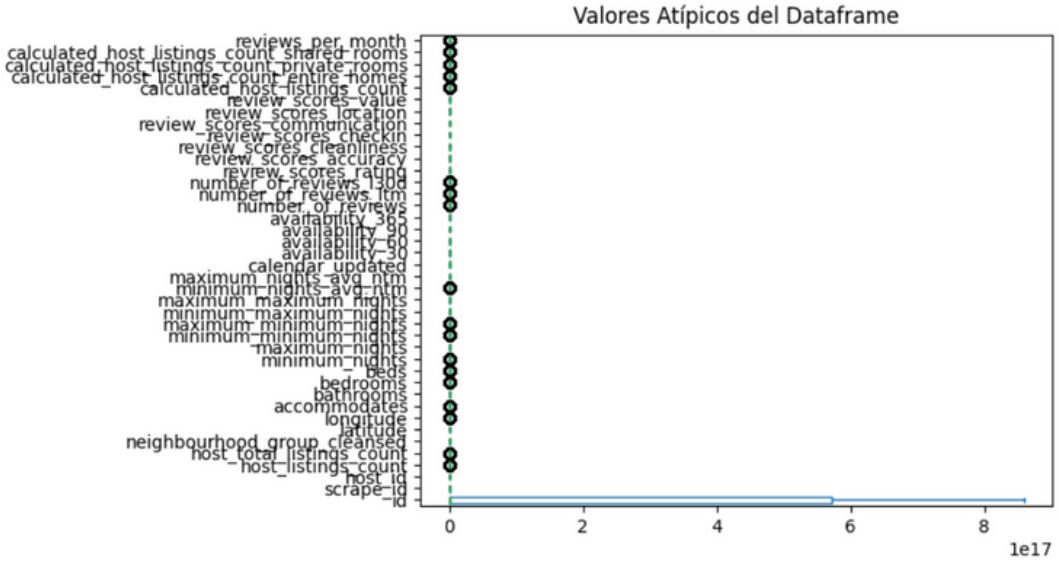
Decidí utilizar este método ya que es menos sensible con los valores atípicos, por lo que detectará valores extremos. Como podemos ver en el gráfico anterior hay una variable que cuenta con outliers extremos mientras que las otras variables no. Por lo que decidí utilizar este método.

Para el Data Frame de DF México decidí utilizar como método para eliminar los valores atípicos el método de los cuartiles. Donde el método reconozca a los valores outliers cuando estos son mayor que 1.5 veces el valor del rango intercuartil.



Decidí utilizar este método ya que es más sensible con los valores atípicos y me ayuda a identificar los valores atípicos que se puedan presentar en el dataframe. Ya que podemos observar que las variable no tienen valores tan extremos.

Por último tenemos el Data Frame de Girona para el cual nuevamente he decididó utilizar el método de cuartiles. Esto debido a que no observo valores extremos en las variables, por lo que utilizaré este método que es más sensible a los valores atípicos.



Etapa 2: Extracción de Datos

Posteriormente se paso a la siguiente etapa en donde realice 10 filtros para cada Data Frame. Estos filtros me permiten conocer y visualizar diferente información de cada ciudad.

Los filtros que se realizarón, son los siguiente:

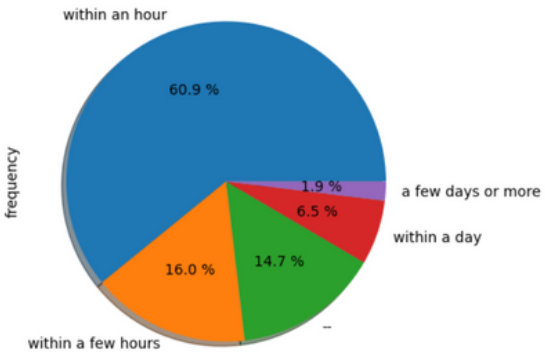
- Solo los registros con host_acceptance_rate> 50%
- Sólo los registros con categoría “superhost”
- Los registros que no hallan verificado identidad “ not identity_verified”
- Los registro cuyo property_type = “Private room” y “Hotel room”
- Los registros que cuenten con bathroom > 1
- Los registros cuyo precio sea mayor de $10,000 y que sean de tipo “Entire home”
- Los registros cuyo review_scores_cleanliness > 4.5
- Los registros cuyo review_scores_value > 4.9
- Los registros cuya availability_365< 100
- Los registros cuya host_response_time sea “within an hour”

Etapa 3: Extracción de Características

- **HOST_RESPONSE_TIME**

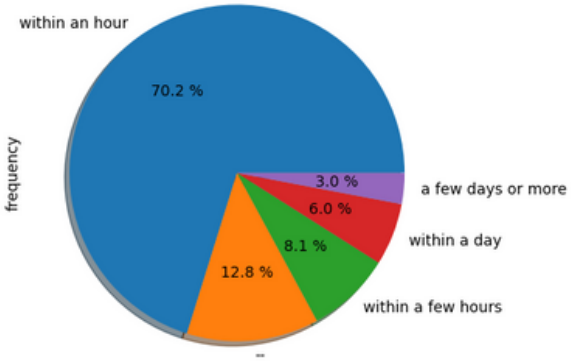
California

host_response_time	frequency	percentage	cumulative_perc
within an hour	4221	0.608564	0.608564
within a few hours	1111	0.160179	0.768743
--	1019	0.146915	0.915657
within a day	452	0.065167	0.980825
a few days or more	133	0.019175	1.000000



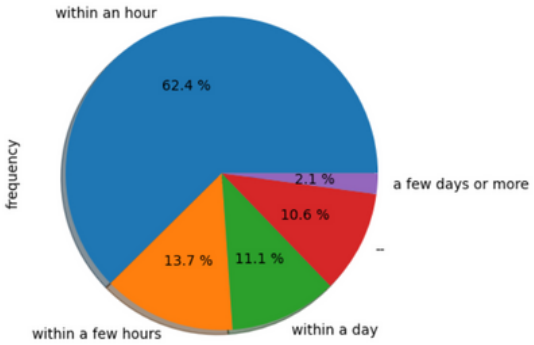
DF

host_response_time	frequency	percentage	cumulative_perc
within an hour	16996	0.701589	0.701647
--	3091	0.127595	0.829253
within a few hours	1966	0.081156	0.910416
within a day	1450	0.059856	0.970276
a few days or more	720	0.029721	1.000000



Girona

host_response_time	frequency	percentage	cumulative_perc
within an hour	12261	0.623811	0.623811
within a few hours	2700	0.137370	0.761180
within a day	2188	0.111320	0.872501
--	2088	0.106233	0.978733
a few days or more	418	0.021267	1.000000



Con esta variable podemos observar el tiempo que tarda el anfitrión en contestar. Para California la respuesta con mayor frecuencia es 'within an hour', para DF y Girona también es 'within an hour'. Las tres ciudades tienen la frecuencia de 'within an hour' entre 60% y 70%.
Esto indica que las tres ciudades tienen un desempeño de tiempo similar para responder a los huéspedes.

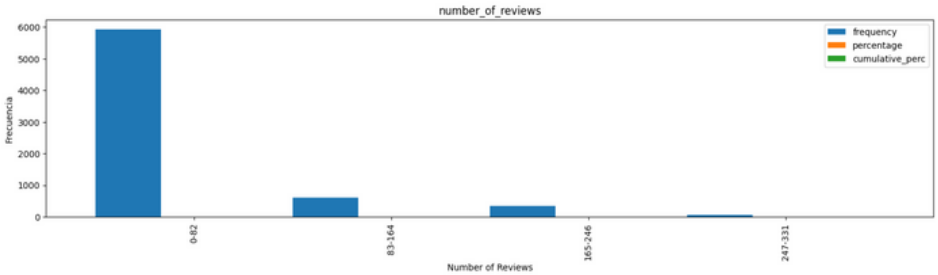
La segunda respuesta con mayor frecuencia es 'within a few hours' para las tres ciudades, por lo que están muy parecidas. Y también observamos, que la respuesta con menos frecuencias es 'a few days or more' lo que significa que los anfitriones regularmente no tardan más de un día.

- **NUMBER_OF_REVIEWS**

Esta variable no era categórica, por lo que para un mejor análisis cree una columna donde se pudieran englobar los datos y así conocer la frecuencia de las respuestas y el desempeño de las tres ciudades.

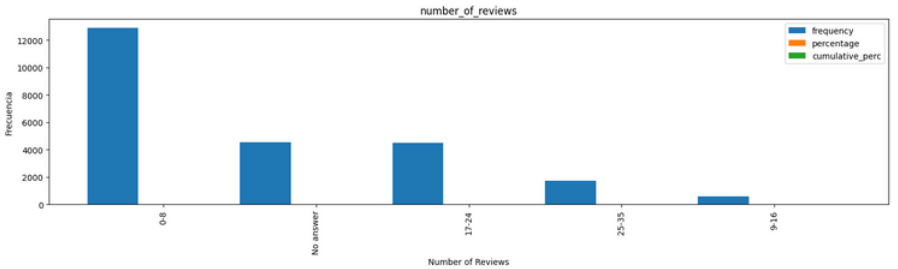
California

reviews	frequency	percentage	cumulative_perc
0-82	5933	0.855392	0.855392
83-164	604	0.087082	0.942474
165-246	341	0.049164	0.991638
247-331	58	0.008362	1.000000



DF

reviews	frequency	percentage	cumulative_perc
0-8	12885	0.531889	0.531889
No answer	4541	0.187451	0.719340
17-24	4477	0.184809	0.904149
25-35	1721	0.071042	0.975191
9-16	601	0.024809	1.000000



Girona

reviews	frequency	percentage	cumulative_perc
0-8	15322	0.779547	0.779547
17-24	3190	0.162300	0.941847
25-35	854	0.043450	0.985296
9-16	289	0.014704	1.000000

Se puede notar que las Girona y DF tienen, en su mayoría, entre 0 y 8 reseñas al tener la mayor frecuencia. Seguido por 17-24 y 25-35 número de reseñas.

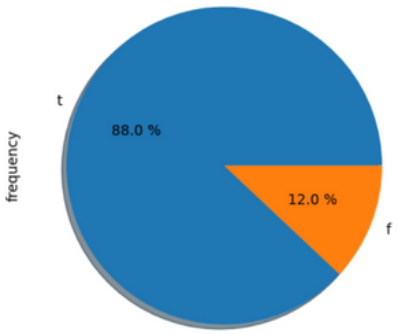
Mientras que para California el número de reseñas es de entre 0-82. Seguido por 83-164. Podemos notar que California tiene categorías de datos más grandes, en comparación a DF y Girona, esto debido a que California tiene una mayor cantidad de número de reseñas.

Por lo que podemos concluir que en California la mayoría de los huéspedes son más probables de dejar reseñas, a comparación de Girona y DF donde se deja una menor cantidad de número de reseñas.

- **HOST_IDENTITY_VERIFIED**

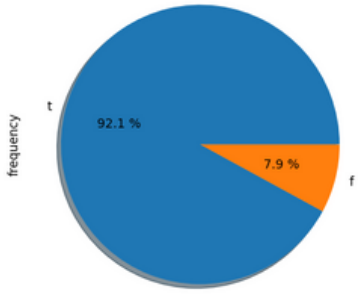
California

host_identity_verified	frequency	percentage	cumulative_perc
t	6106	0.880334	0.880334
f	830	0.119666	1.000000



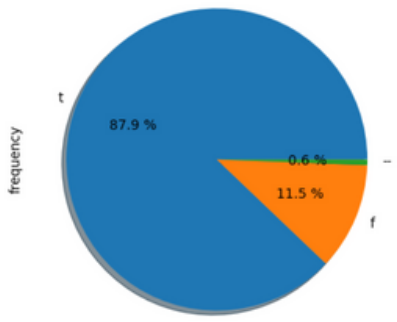
DF

host_identity_verified	frequency	percentage	cumulative_perc
t	22300	0.920537	0.920613
f	1923	0.079381	1.000000



Girona

host_identity_verified	frequency	percentage	cumulative_perc
t	17277	0.879013	0.879013
f	2261	0.115034	0.994047
--	117	0.005953	1.000000



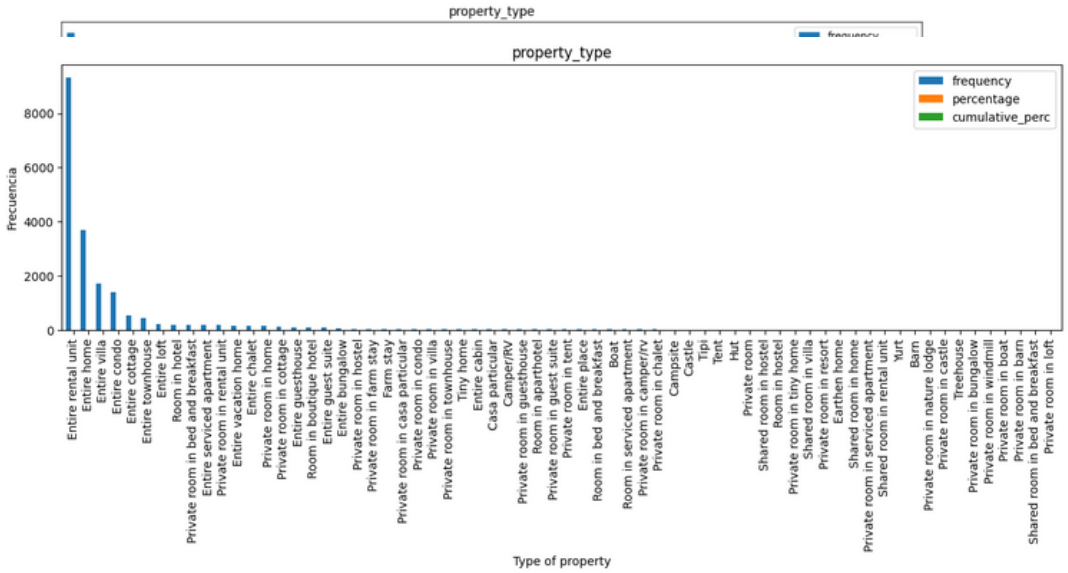
En las tres ciudades se puede observar que la mayoría tienen la identidad del anfitrión verificada, rondando entre un 87% y 93%. Por el otro lado, podemos ver que California es la ciudad con mayor número de anfitriones que no tienen su identidad verificada representando un 12%, seguido por Girona con un 11.5% y por último con DF con 7.9%.

DF es la ciudad donde la mayoría de los anfitriones tienen su identidad verificada representando el 92.1% y teniendo una frecuencia de 22300.

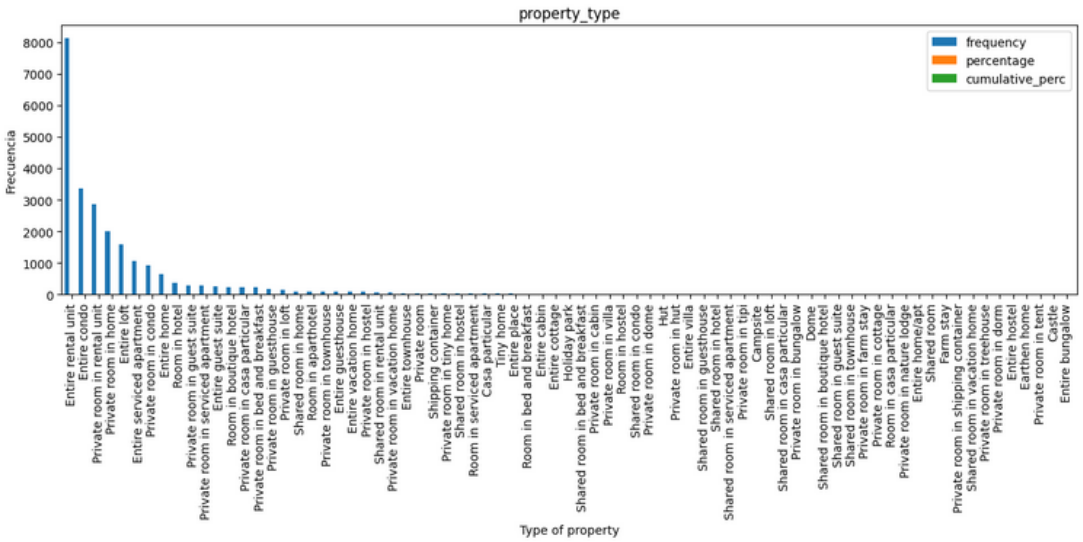
Es importante mencionar que en Girona podemos observar que hay algunos datos con dos guiones, estos eran los datos nulos que se procedieron a reemplazar con un string para no afectar las respuestas.

- **PROPERTY_TYPE**

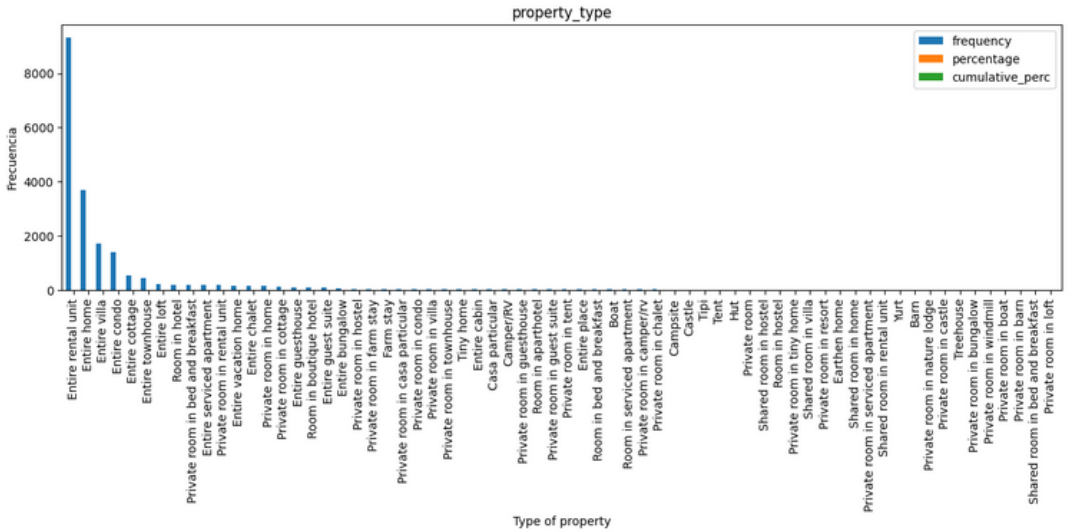
California



DF



Girona



Se puede concluir que el tipo de propiedad más ofrecido en las tres ciudades es el de 'Entire rental unit', por lo que se puede decir que es el más común de encontrar. Este tipo de propiedad representa aproximadamente el 26% en California, 33% en DF y 47% en Girona en los tipos de propiedad.
Después, California y Girona siguen con 'entire home', 'entire villa', 'entire condo' y 'entire cottage'. Podemos observar que estas dos ciudades tiene una distribución similar de los tipos de propiedades que se ofrecen. Puede esto significar que los clientes tienen gustos similares al momento de buscar rentar.

Mientras que para DF los siguientes con una mayor frecuencia son 'entire condo, 'private room in rental unit', 'private room in home', 'entire loft' y 'entired service aparment'. En DF podemos notar que hay más oferta de la renta de cuartos privados, en comparación con las otras ciudades.

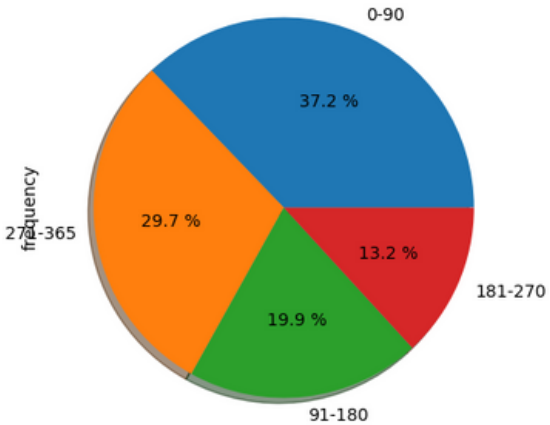
Esto puede ser un indicador de las preferencias y necesidades que tienen los clientes en cada ciudad.

- **AVAILABILITY_365**

Nuevamente, esta característica no es categórica por lo que procedí a crear categorías que engloban los datos de la información con la que se contaba. Para así hacer un mejor análisis y tener una mayor compresión de los datos.

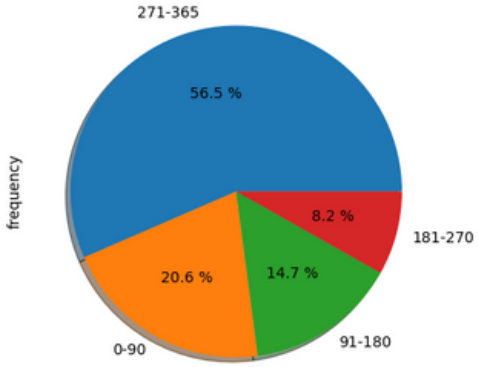
California

availability	frequency	percentage	cumulative_perc
0-90	2580	0.371972	0.371972
271-365	2062	0.297290	0.669262
91-180	1379	0.198818	0.868080
181-270	915	0.131920	1.000000



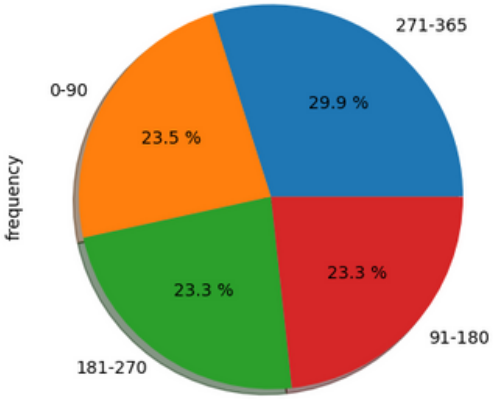
DF

availability	frequency	percentage	cumulative_perc
271-365	13689	0.565077	0.565077
0-90	4985	0.205779	0.770857
91-180	3569	0.147327	0.918184
181-270	1981	0.081775	0.999959



Girona

availability	frequency	percentage	cumulative_perc
271-365	5880	0.299161	0.299161
0-90	4622	0.235156	0.534317
181-270	4582	0.233121	0.767438
91-180	4571	0.232562	1.000000



Como última característica se tiene la disponibilidad 365 que hay en cada propiedad. DF es la ciudad con una mayor disponibilidad, ya que la mayoría de sus propiedades tiene una disponibilidad entre 271-365 días y representan el 56.5% de todos los datos. No obstante, la disponibilidad de 181-270 días es la que tiene menos frecuencia en DF representando el 8.2%.

Después está Girona con una disponibilidad de 271-365 días que representa el 29.9% de los datos, seguido por 0-90 días con 23.5% y 181-270 y 91-80 con 23.3%. Girona tiene una distribución similar, o muy cercana, entre todos las categorías de disponibilidad que hay.

Por último, está California con la mayor disponibilidad de 0-90 días representando el 37.2%, seguido por 271-365 días con el 29.7%.

California es la ciudad con la menor disponibilidad de días, mientras que DF la que más disponibilidad tiene. Esto podría ser un indicador de la demanda que se puede llegar a presentar en las ciudades. Por ejemplo, podríamos decir que California tiene mucha demanda mientras que DF tiene menos, lo que le permite tener mayor disponibilidad todo el año en la mayoría de las propiedades.