Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Datos Airbnb Estados unidos

Ciencias de Datos

Carlos Vallarin Lopez

Entrega

27 de Noviembre 2024

Introducción:

Descripción breve del objetivo del proyecto:

El objetivo de este proyecto es analizar un conjunto de datos de reservaciones de Airbnb con el fin de obtener insights valiosos sobre los patrones de precios, duración de las estancias y características de las propiedades. A través de este análisis, se busca identificar factores clave que influyen en el precio de las propiedades y la duración de las reservas, tales como la ubicación, tipo de habitación, y políticas del anfitrión. Además, el proyecto tiene como objetivo construir visualizaciones como mapas de calor y distribuciones de precio, proporcionando una comprensión clara sobre cómo las variables afectan las reservas en la plataforma de Airbnb.

Justificación y contexto: ¿por qué es importante resolver o estudiar esta problemática?

El análisis de datos en plataformas de alquiler a corto plazo como Airbnb se ha convertido en un área de gran interés debido al impacto significativo que tiene en las economías locales, la industria turística y los mercados inmobiliarios. Con el auge del alquiler a corto plazo, entender cómo factores como la ubicación, la política de cancelación, el tipo de propiedad y el precio influencian las decisiones de los usuarios es crucial tanto para anfitriones como para los huéspedes, así como para la gestión de políticas públicas.

Justificación del análisis:

1. Optimización de precios para anfitriones: El análisis de los precios y su relación con las características de la propiedad permite a los anfitriones optimizar sus tarifas y mejorar su competitividad en el mercado. Los anfitriones pueden ajustar sus precios basándose en factores como el

vecindario, el tipo de habitación, o la demanda en ciertos momentos del año.

2. Mejora de la experiencia del huésped: Los huéspedes pueden beneficiarse del análisis al identificar patrones sobre qué tipo de propiedades ofrecen una buena relación calidad-precio según sus preferencias, ya sea por ubicación, precio o condiciones.
3. Estudio del impacto económico: Airbnb tiene un impacto económico en ciudades y comunidades, por lo que comprender cómo estos alquileres afectan el mercado inmobiliario local o la economía general puede ayudar a los responsables de la política pública a tomar decisiones informadas sobre regulaciones y permisos de alquiler.
4. Identificación de tendencias del mercado: A través de este análisis, podemos identificar tendencias actuales del mercado de Airbnb y cómo estas podrían evolucionar en el futuro, lo que resulta útil tanto para anfitriones, inversionistas, como para turistas interesados en esta alternativa de alojamiento.

Fuentes de datos: descripción de las bases de datos empleadas (origen, cantidad de datos, principales características).

Origen de los datos:

Los datos empleados en este proyecto provienen de un conjunto de datos de Airbnb, el cual se recopila de las reservas y propiedades disponibles en la plataforma de Airbnb. La información está relacionada con las propiedades de alojamiento, sus características, precios, disponibilidad, y las interacciones entre anfitriones y huéspedes.

Este tipo de base de datos generalmente está disponible en plataformas públicas de datos abiertos, como Kaggle, donde los usuarios pueden acceder a conjuntos de datos relacionados con Airbnb. En este caso, el conjunto de datos puede estar

basado en información real o puede ser una versión anonimizada para fines educativos o de análisis.

Cantidad de datos:

El conjunto de datos contiene 101572 datos y 25 columnas de información, lo que representa un amplio volumen de observaciones de propiedades y reservas en diferentes ubicaciones. Cada fila corresponde a una propiedad de Airbnb que puede haber sido reservada por diferentes usuarios. Este conjunto de datos proporciona una muestra significativa que permite obtener conclusiones robustas sobre los patrones de comportamiento de los usuarios y los anfitriones en la plataforma.

Principales características de las bases de datos:

El conjunto de datos contiene diversas columnas que representan información clave sobre cada propiedad y su comportamiento en la plataforma. Las principales características del conjunto de datos incluyen:

1. Identificadores únicos:
    - id: Un identificador único para cada propiedad de Airbnb.
    - host id: Un identificador único para cada anfitrión.
    - host name: El nombre del anfitrión de la propiedad.
2. Características del anfitrión:
    - host_identity_verified: Si la identidad del anfitrión está verificada (valor booleano: sí/no).
    - calculated host listings count: El número de propiedades que tiene el anfitrión.
3. Ubicación de la propiedad:

- neighbourhood: El nombre del vecindario en el que está ubicada la propiedad.
- neighbourhood group: Es una agrupación de varios vecindarios dentro de una región más grande. Se usa para referirse a distritos, áreas de planificación o subdivisiones más amplias dentro de una ciudad.
- lat y long: Coordenadas geográficas de la propiedad (latitud y longitud).
- country y country code: País y código de país donde se encuentra la propiedad.

4. Características de la propiedad:
    - room type: El tipo de habitación (por ejemplo, habitación privada, casa completa, etc.).
    - Construction year: Año de construcción de la propiedad.
    - house_rules: Reglas de la casa establecidas por el anfitrión.
5. Políticas y precios:
    - instant_bookable: Si la propiedad es reservable de forma instantánea.
    - cancellation_policy: Política de cancelación de la propiedad (por ejemplo, flexible, estricta).
    - price: El precio por noche de la propiedad (generalmente en formato numérico con el símbolo de moneda).
    - service fee: Cargos adicionales por servicio (en algunos casos).
6. Datos relacionados con las reseñas:

- number of reviews: El número total de reseñas recibidas por la propiedad.
- last review: La fecha de la última reseña.
- reviews per month: El número promedio de reseñas que recibe la propiedad por mes.
- review rate number: Número total de calificaciones.

7. Disponibilidad:
    - availability 365: El número de días al año en los que la propiedad está disponible para ser reservada.
    - minimum nights: El número mínimo de noches que un huésped debe reservar.

Metodología:

Resumen estadístico de los datos

```
df.describe()
```
✓ 0.2s

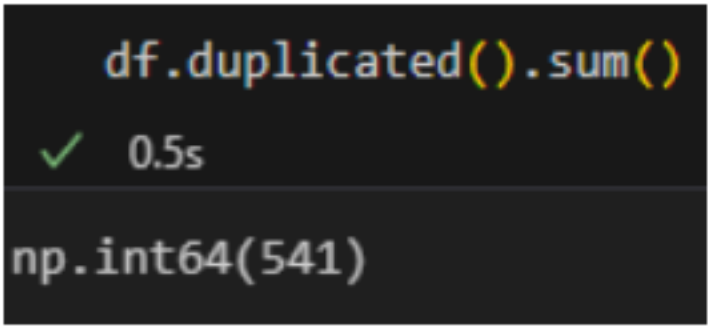
	id	host id	lat	long	Construction year	minimum nights	number of reviews	reviews per month	review rate number	calculated host listings count	availability 365
count	1.025990e+05	1.025990e+05	102591.000000	102591.000000	102385.000000	102190.000000	102416.000000	86720.000000	102273.000000	102280.000000	102151.000000
mean	2.914623e+07	4.925411e+10	40.728094	-73.949644	2012.487464	8.135845	27.483743	1.374022	3.279106	7.936605	141.133254
std	1.625751e+07	2.853900e+10	0.055857	0.049521	5.765556	30.553781	49.508954	1.746621	1.284657	32.218780	135.435024
min	1.001254e+06	1.236005e+08	40.499790	-74.249840	2003.000000	-1223.000000	0.000000	0.010000	1.000000	1.000000	-10.000000
25%	1.508581e+07	2.458333e+10	40.688740	-73.982580	2007.000000	2.000000	1.000000	0.220000	2.000000	1.000000	3.000000
50%	2.913660e+07	4.911774e+10	40.722290	-73.954440	2012.000000	3.000000	7.000000	0.740000	3.000000	1.000000	96.000000
75%	4.320120e+07	7.399650e+10	40.762760	-73.932350	2017.000000	5.000000	30.000000	2.000000	4.000000	2.000000	269.000000
max	5.736742e+07	9.876313e+10	40.916970	-73.705220	2022.000000	5645.000000	1024.000000	90.000000	5.000000	332.000000	3677.000000

Calcular el porcentaje de valores faltantes por columna.

```
df.isnull().mean()*100
```
✓ 0.5s

```
id                                0.000000
NAME                              0.243667
host id                           0.000000
host_identity_verified            0.281679
host name                         0.395715
neighbourhood group               0.028265
neighbourhood                     0.015595
lat                               0.007797
long                              0.007797
country                           0.518524
country code                      0.127682
instant_bookable                  0.102340
cancellation_policy               0.074075
room type                         0.000000
Construction year                 0.208579
price                             0.240743
service fee                       0.266084
minimum nights                    0.398639
number of reviews                 0.178364
last review                      15.490404
reviews per month                15.476759
review rate number                0.317742
calculated host listings count    0.310919
availability 365                  0.436651
house_rules                      50.810437
license                          99.908851
dtype: float64
```

Total de filas duplicadas encontradas.



Descripción de los tipos de datos originales y los problemas encontrados.

Los datos originales ya venían con valores nulos, aparte de los demás valores nulos, realmente el problema real fueron los valores nulos y al ser varias columnas de números y de palabras normales, se tuvo que investigar mas detalladamente la base de datos para ver cual convenia mas para remplazar los valores nulos y no solo borrarlos

Proceso de limpieza:

Métodos Utilizados

Se utilizaron métodos como el de Eliminación de columnas para una columna que no tenía datos, se utilizo también método de Eliminación de filas para algunos datos

Se utilizo también imputación sacando la media en algunos datos para que no quedaran valores atípicos y sean mas correctos

Se utilizo la eliminación de duplicados para algunos datos que estaban repetidos y así asegurar la integridad de los datos

De igual forma se utilizó la conversión de tipos de datos ya que habían unos datos que no eran correctos como el de la fecha y otros que se pasaron a numéricos

Antes y después

```
df.info()
```

✓ 0.1s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102599 entries, 0 to 102598
Data columns (total 26 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              102599 non-null  int64
 1   NAME                            102349 non-null  object
 2   host id                         102599 non-null  int64
 3   host_identity_verified          102310 non-null  object
 4   host name                       102193 non-null  object
 5   neighbourhood group             102570 non-null  object
 6   neighbourhood                   102583 non-null  object
 7   lat                             102591 non-null  float64
 8   long                            102591 non-null  float64
 9   country                         102067 non-null  object
 10  country code                    102468 non-null  object
 11  instant_bookable                102494 non-null  object
 12  cancellation_policy             102523 non-null  object
 13  room type                       102599 non-null  object
 14  Construction year               102385 non-null  float64
 15  price                           102352 non-null  object
 16  service fee                     102326 non-null  object
 17  minimum nights                  102198 non-null  float64
 18  number of reviews               102416 non-null  float64
 19  last review                     88706 non-null   object
...
 24  house_rules                     59468 non-null   object
 25  license                         3 non-null       object
dtypes: float64(9), int64(2), object(15)
memory usage: 20.4+ MB
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust ce

```
df.duplicated().sum()
```

✓ 0.6s

```
np.int64(541)
```

```
df.info()
```

✓ 0.2s

```
<class 'pandas.core.frame.DataFrame'>
Index: 101572 entries, 0 to 102044
Data columns (total 25 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   id                              101572 non-null  int64
 1   NAME                            101572 non-null  object
 2   host id                         101572 non-null  int64
 3   host_identity_verified          101572 non-null  object
 4   host name                       101572 non-null  object
 5   neighbourhood group             101572 non-null  object
 6   neighbourhood                   101572 non-null  object
 7   lat                             101572 non-null  float64
 8   long                            101572 non-null  float64
 9   country                         101572 non-null  object
 10  country code                    101572 non-null  object
 11  instant_bookable                101572 non-null  object
 12  cancellation_policy             101572 non-null  object
 13  room type                       101572 non-null  object
 14  Construction year               101572 non-null  float64
 15  price                           101572 non-null  object
 16  service fee                     101572 non-null  object
 17  minimum nights                  101572 non-null  float64
 18  number of reviews               101572 non-null  float64
 19  last review                     101572 non-null  datetime64[ns]
...
 23  availability 365                101572 non-null  float64
 24  house_rules                     101572 non-null  object
dtypes: datetime64[ns](1), float64(9), int64(2), object(13)
memory usage: 20.1+ MB
```

Output is truncated. View as a scrollable element or open in a text editor. Adjust cell outpu

```
df.duplicated().sum()
```

✓ 0.5s

```
np.int64(0)
```

```
df.isnull().sum()
```

✓ 0.1s

```
id                                  0
NAME                              250
host id                             0
host_identity_verified            289
host name                         406
neighbourhood group                20
neighbourhood                      16
lat                                 8
long                                8
country                           532
country code                      131
instant_bookable                  105
cancellation_policy                76
room type                           0
Construction year                 214
price                             347
service fee                       273
minimum nights                    409
number of reviews                 183
last review                     15893
reviews per month               15879
review rate number                326
calculated host listings count    319
availability 365                  448
house_rules                     52131
license                        102597
dtype: int64
```

```
# Maintenant vérifions s'il y a d'éventuelles colonnes que j'aurais
df['house_rules'].fillna('No rules', inplace=True)
df['NAME'].fillna('Name not provided', inplace=True)
df['host_identity_verified'].fillna('No data', inplace=Tr
```

✓ 0.0s

Outputs are collapsed

```
df
```

Outputs are collapsed

```
df.isnull().sum()
```

Outputs are collapsed

```
df['host name'].fillna('No data', inplace=True)
df['neighbourhood group'].fillna('No data', inplace=True)
```

✓ 0.0s

Outputs are collapsed

```
df
```

Outputs are collapsed

```
df['neighbourhood'].fillna('No data', inplace=True)
```

✓ 0.0s

```
df['neighbourhood'].fillna('No data', inplace=True)
```
✓ 0.0s

```
print(df['long'].dtype)  # Verifica el tipo de datos
```
✓ 0.0s

```
float64
```

```
df['long'] = pd.to_numeric(df['long'], errors='coerce') #Se pasa a numeros los datos
```
✓ 0.0s

```
print(df['long'].dtype)
```
✓ 0.0s

```
float64
```

```
print(df['lat'].dtype)  # Verifica el tipo de datos
```
✓ 0.0s

```
float64
```

```
#Cambiar a numericos
df['lat'] = pd.to_numeric(df['lat'], errors='coerce')
```
✓ 0.0s

```
media_lat = df['lat'].mean()
df['lat'].fillna(media_lat, inplace=True)
```
✓ 0.0s

```
Outputs are collapsed —
```

```
#Sacamos la media para que no queden datos nfaltos
```

```
#Sabiendo que es la base de datos de Estados unidos podemos cambiar los valores nulos de "Country" a "United States"
df['country'].fillna('United States', inplace=True)
df['country code'].fillna('US', inplace=True)
```

✓ 0.0s

Outputs are collapsed ...

✓ `df.isnull().sum()` ...

Outputs are collapsed ...

```
df['instant_bookable'].fillna('There is no information', inplace=True)
df['cancellation_policy'].fillna('Not specified', inplace=True)
```

✓ 0.0s

Outputs are collapsed ...

✓ `df.isnull().sum()` ...

Outputs are collapsed ...

```
print(df['Construction year'].dtype)  # Verifica el tipo de datos
```

✓ 0.0s

Execution Order

```
media_Year = df['Construction year'].mean()
df['Construction year'].fillna(media_Year, inplace=True)
```

✓ 0.0s

C:\Users\carlo\AppData\Local\Temp\ipykernel_20304\3431937396.py:2: FutureWarning: A value is trying to be set on a copy

```
media_Reviews = df['number of reviews'].mean()
df['number of reviews'].fillna(media_Reviews, inplace=True)
```

[104] ✓ 0.0s

Outputs are collapsed ...

```
df['last review'] = pd.to_datetime(df['last review'], errors='coerce')  # Convierte la columna a datetime
```

[105] ✓ 0.0s

```
media_fecha = df['last review'].mean()   #Sacamos la media
df['last review'].fillna(media_fecha, inplace=True)
```

[106] ✓ 0.0s

Outputs are collapsed ...

✓ `df.isnull().sum()` ...

Outputs are collapsed ...

✓ `df` ...

Outputs are collapsed ...

```
media_Reviews = df['reviews per month'].mean()
df['reviews per month'].fillna(media_Reviews, inplace=True)
```

```
media_Host = df['calculated host listings count'].mean()
df['calculated host listings count'].fillna(media_Host, inplace=True)
```
✓ 0.0s

Outputs are collapsed

```
media_365 = df['availability 365'].mean()
df['availability 365'].fillna(media_365, inplace=True)
```
✓ 0.0s

Outputs are collapsed

```
df.isnull().sum()  #Ya no hay datos nulos
```
✓ 0.1s

```
id                                0
NAME                              0
host id                           0
host_identity_verified            0
host name                         0
neighbourhood group               0
neighbourhood                     0
lat                               0
long                              0
country                           0
country code                      0
instant_bookable                  0
cancellation_policy               0
room type                         0
Construction year                 0
price                             0
service fee                       0
minimum nights                    0
number of reviews                 0
last review                       0
reviews per month                 0
review rate number                0
calculated host listings count    0
availability 365                  0
house_rules                       0
dtype: int64
```

Resultados:

Resumen final: Al concluir con esta práctica pude corregir los datos de forma que quedaran más correctamente la base de datos sin tantos

datos atípicos y así poder ver de forma más clara todos los datos que se formaron y que se limpiaron, así cuando se hagan gráficos ya van a salir gráficos correctos y sin tanta dispersión de datos

Confirmar que los tipos de datos son correctos.

Se confirmo que los datos son correctos ya que no quedo ningún dato nulo y el resumen estadístico dio 0, así como no hubo ningún invalid value ni quedo con duplicados, y las desviaciones estándar no fueron tan altas

	Construction year	minimum nights	number of reviews	\
count	101572.000000	101572.000000	101572.000000	
mean	2012.487708	8.103929	27.529611	
min	2003.000000	-1223.000000	0.000000	
25%	2008.000000	2.000000	1.000000	
50%	2012.000000	3.000000	7.000000	
75%	2017.000000	5.000000	30.000000	
max	2022.000000	5645.000000	1024.000000	
std	5.759895	30.560578	49.562767	

	last review	reviews per month	review rate number
count	101572	101572.000000	101572.000000
mean	2019-06-10 14:48:08.226971648	1.376855	3.278768
min	2012-07-11 00:00:00	0.010000	1.000000
25%	2019-01-02 00:00:00	0.280000	2.000000
50%	2019-06-11 12:29:59.707893760	1.060000	3.000000
75%	2019-07-01 00:00:00	1.710000	4.000000
max	2058-06-16 00:00:00	90.000000	5.000000
std	NaN	1.608210	1.282933

	calculated host listings count	availability 365
count	101572.000000	101572.000000
mean	7.929621	141.034038
min	1.000000	-10.000000
25%	1.000000	3.000000
50%	1.000000	98.000000
75%	2.000000	268.000000
max	332.000000	3677.000000
std	32.227126	135.111981

Tabla del porcentaje de valores faltantes por columna

```
df.isnull().mean()*100
```

✓ 0.1s

id	0.0
NAME	0.0
host id	0.0
host_identity_verified	0.0
host name	0.0
neighbourhood group	0.0
neighbourhood	0.0
lat	0.0
long	0.0
country	0.0
country code	0.0
instant_bookable	0.0
cancellation_policy	0.0
room type	0.0
Construction year	0.0
price	0.0
service fee	0.0
minimum nights	0.0
number of reviews	0.0
last review	0.0
reviews per month	0.0
review rate number	0.0
calculated host listings count	0.0
availability 365	0.0
house_rules	0.0

dtype: float64

Comprobación de que no hay valores duplicados

```
df.duplicated().sum()
```

✓ 0.4s

```
np.int64(0)
```

```
for i in lista_col:
    print(f"En la columna {i} los invalid_value son: {df[df[i] == 'invalid_value'].shape[
```

✓ 0.2s

```
En la columna id los invalid_value son: 0
En la columna NAME los invalid_value son: 0
En la columna host id los invalid_value son: 0
En la columna host_identity_verified los invalid_value son: 0
En la columna host name los invalid_value son: 0
En la columna neighbourhood group los invalid_value son: 0
En la columna neighbourhood los invalid_value son: 0
En la columna lat los invalid_value son: 0
En la columna long los invalid_value son: 0
En la columna country los invalid_value son: 0
En la columna country code los invalid_value son: 0
En la columna instant_bookable los invalid_value son: 0
En la columna cancellation_policy los invalid_value son: 0
En la columna room type los invalid_value son: 0
En la columna Construction year los invalid_value son: 0
En la columna price los invalid_value son: 0
En la columna service fee los invalid_value son: 0
En la columna minimum nights los invalid_value son: 0
En la columna number of reviews los invalid_value son: 0
En la columna last review los invalid_value son: 0
En la columna reviews per month los invalid_value son: 0
En la columna review rate number los invalid_value son: 0
En la columna calculated host listings count los invalid_value son: 0
En la columna availability 365 los invalid_value son: 0
En la columna house_rules los invalid_value son: 0
En la columna license los invalid_value son: 0
```

Análisis Exploratorio de Datos (EDA)

Visión General:

El dataset contiene 101572 Datos (registros), donde cada fila representa una propiedad listada en Airbnb.

Variables: El conjunto de datos tiene múltiples columnas que describen propiedades como la ubicación, el precio, las características del anfitrión, las políticas de cancelación, entre otras. A continuación, se ofrece una visión general de las variables más relevantes.

Resumen General de las Variables:

id: Identificador único de la propiedad.

NAME: Nombre de la propiedad.

host id: Identificador único del anfitrión.

host_identity_verified: Indica si la identidad del anfitrión ha sido verificada (booleano).

neighbourhood: Vecindario en el que se encuentra la propiedad.

neighbourhood group: Grupo de vecindarios a nivel más macro.

price: Precio por noche (en formato numérico).

minimum nights: Número mínimo de noches requeridas para hacer una reserva.

number of reviews: Número total de reseñas de la propiedad.

last review: Fecha de la última reseña.

reviews per month: Número promedio de reseñas por mes.

availability 365: Número de días al año en los que la propiedad está disponible para ser reservada.

2. Tipos de Variables

Las variables en el conjunto de datos se pueden clasificar en distintos tipos. A continuación, clasificamos las variables del conjunto de datos en categóricas, numéricas, fechas, y texto:

Variables Categóricas:

- host_identity_verified: Indica si el anfitrión tiene su identidad verificada (Sí/No).
- neighbourhood: Nombre del vecindario (categoría de texto).
- neighbourhood group: Grupo de vecindarios (categoría de texto).
- room type: Tipo de habitación (por ejemplo, "entire home/apt", "private room", "shared room").
- cancellation_policy: Política de cancelación de la propiedad (categoría de texto, como "flexible", "moderate", "strict").
- instant_bookable: Si la propiedad es reservable de forma instantánea (Sí/No).

Variables Numéricas:

- price: Precio por noche de la propiedad.
- minimum nights: Número mínimo de noches para hacer una reserva.
- number of reviews: Número total de reseñas.
- reviews per month: Número promedio de reseñas por mes.
- availability 365: Número de días disponibles al año.
- calculated_host_listings_count: Número de propiedades que tiene el anfitrión.

Variables de Fecha:

- last review: Fecha de la última reseña de la propiedad (se debe convertir en formato de fecha).

Variables de Texto:

- NAME: Nombre de la propiedad (texto libre).

3. Resumen Estadístico

El resumen estadístico proporciona una visión general de las distribuciones y estadísticas descriptivas de las variables numéricas, así como de las frecuencias de las variables categóricas.

Variables Numéricas:

Para las variables numéricas, generaremos estadísticas descriptivas como la media, mediana, desviación estándar, mínimo, máximo, y los cuartiles (25%, 50%, 75%).

	id	host id	lat	long	Construction year	price	minimum nights
count	1.015720e+05	1.015720e+05	101572.000000	101572.000000	101572.000000	101572.000000	101572.000000
mean	2.919428e+07	4.925552e+10	40.728092	-73.949656	2012.487708	625.249921	8.103929
std	1.627030e+07	2.853444e+10	0.055858	0.049506	5.759895	331.651535	30.560578
min	1.001254e+06	1.236005e+08	40.499790	-74.249840	2003.000000	50.000000	-1223.000000
25%	1.510860e+07	2.458873e+10	40.688720	-73.982580	2008.000000	340.000000	2.000000
50%	2.918438e+07	4.911962e+10	40.722280	-73.954440	2012.000000	624.000000	3.000000
75%	4.330544e+07	7.397945e+10	40.762770	-73.932357	2017.000000	913.000000	5.000000
max	5.736024e+07	9.876313e+10	40.916970	-73.705220	2022.000000	1200.000000	5645.000000

number of reviews	reviews per month	review rate number	calculated host listings count	availability 365
101572.000000	101572.000000	101572.000000	101572.000000	101572.000000
27.529611	1.376855	3.278768	7.929621	141.034038
49.562767	1.608210	1.282933	32.227126	135.111981
0.000000	0.010000	1.000000	1.000000	-10.000000
1.000000	0.280000	2.000000	1.000000	3.000000
7.000000	1.060000	3.000000	1.000000	98.000000
30.000000	1.710000	4.000000	2.000000	268.000000
1024.000000	90.000000	5.000000	332.000000	3677.000000

2. Visualización y Distribución de Variables Individuales

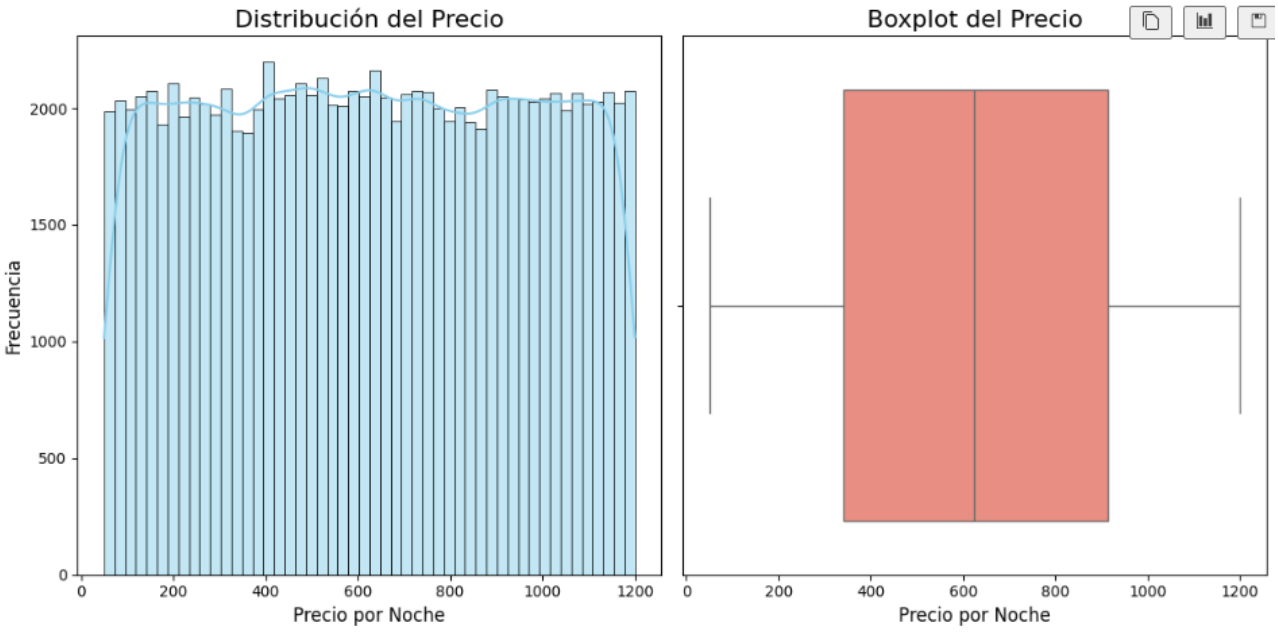
La visualización de las variables individuales es fundamental para entender la distribución de los datos, identificar posibles sesgos, detectar outliers (valores atípicos) y obtener información clave que guiará el análisis posterior. Para las variables numéricas, utilizaremos histogramas y boxplots, que son herramientas muy útiles para explorar los datos.

A continuación, realizaremos la visualización de algunas de las variables numéricas más importantes en el conjunto de datos: price, minimum nights, number of reviews, reviews per month, y calculated_host_listings_count.

1. Distribución del Precio (price)

Gráfico: Histograma y Boxplot

El precio de la propiedad es una de las variables más relevantes, ya que probablemente influye en la decisión de los huéspedes. Es importante entender su distribución y detectar si existen valores extremos que puedan afectar el análisis.



Histograma: La distribución de los precios muestra una clara asimetría positiva (sesgo hacia la derecha), lo que indica que la mayoría de las propiedades tienen moderados, pero hay algunas propiedades extremadamente caras que empujan el precio promedio hacia

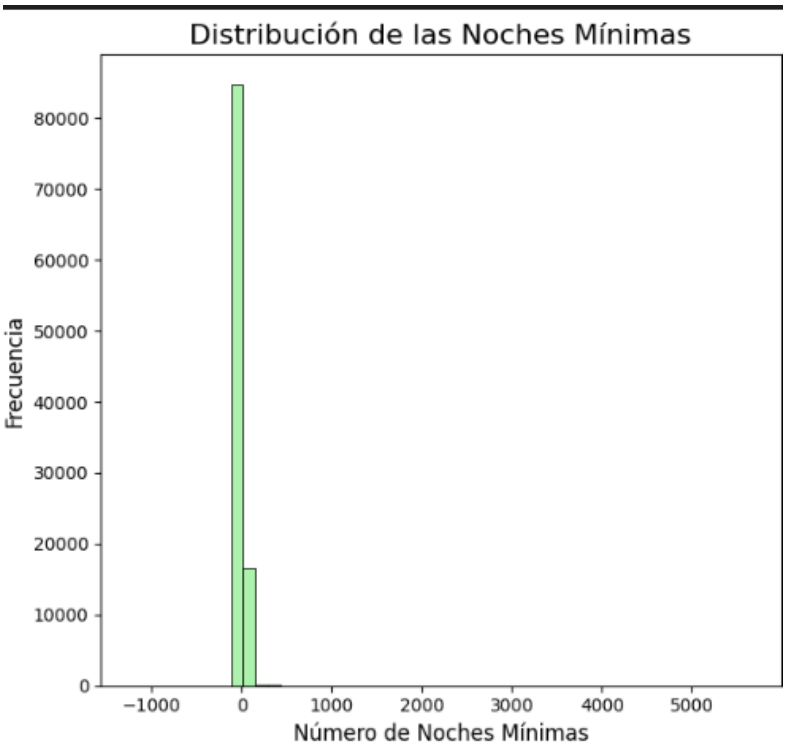
arriba. Este comportamiento es típico en plataformas como Airbnb, donde la mayoría de las propiedades son asequibles, pero hay algunas propiedades de lujo o únicas que tienen precios muy altos.

Boxplot: El boxplot revela una gran cantidad de outliers (valores extremos) en el precio, especialmente en el lado derecho, sin datos atípicos

2. Distribución de las Noches Mínimas (minimum nights)

Gráfico: Histograma

La variable minimum nights indica el número mínimo de noches que un huésped debe reservar. Este parámetro es importante porque puede afectar la accesibilidad de las propiedades. Vamos a ver cómo se distribuye esta variable.

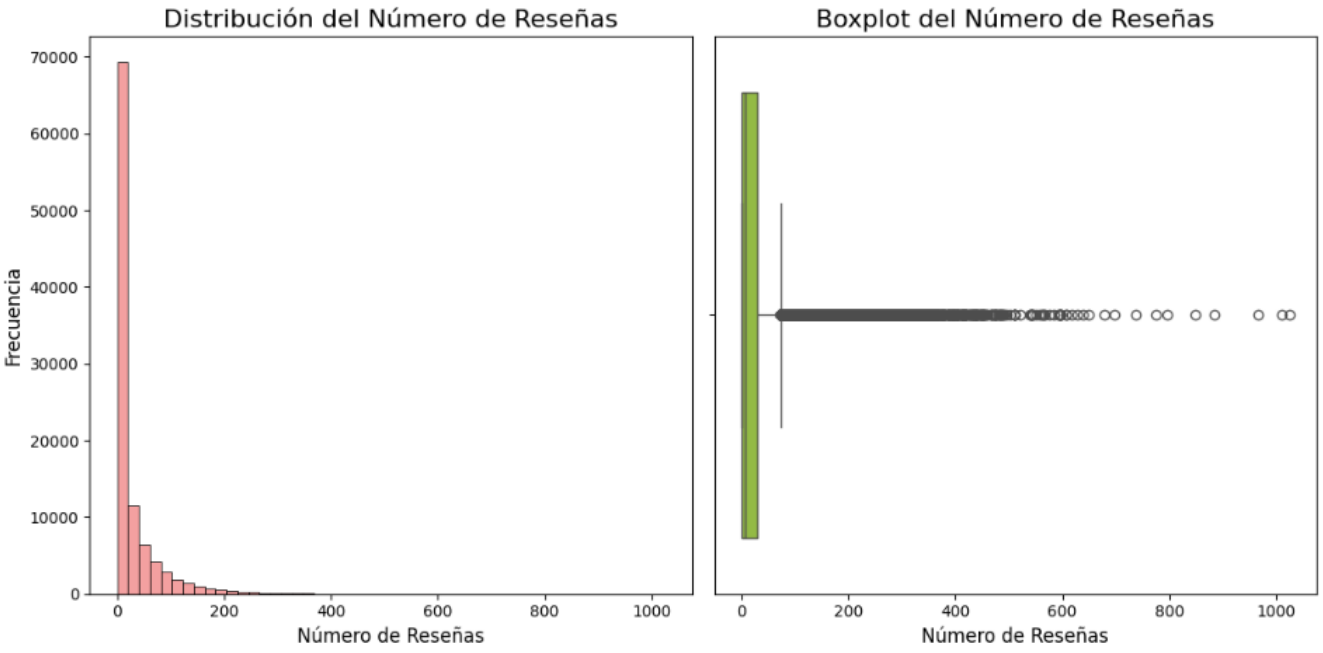


Histograma: La distribución de las noches mínimas muestra que la mayoría de las propiedades requieren 1 o 2 noches mínimas para reservar. Sin embargo, hay algunas propiedades que requieren una cantidad significativamente mayor de noches mínimas, lo que se refleja en la larga cola hacia la derecha del histograma.

3. Distribución del Número de Reseñas (number of reviews)

Gráfico: Histograma y Boxplot

El número de reseñas puede ser un buen indicador de la popularidad de una propiedad. Las propiedades más conocidas tienden a tener un mayor número de reseñas. Veamos cómo se distribuyen estas reseñas.



Histograma: La mayoría de las propiedades tienen pocas reseñas, lo que sugiere que muchas propiedades en Airbnb pueden ser nuevas o no tan populares. Hay una cola larga hacia la derecha, lo que indica que algunas propiedades tienen un número excepcionalmente alto de reseñas (lo que podría ser el caso de propiedades populares o superanfitriones).

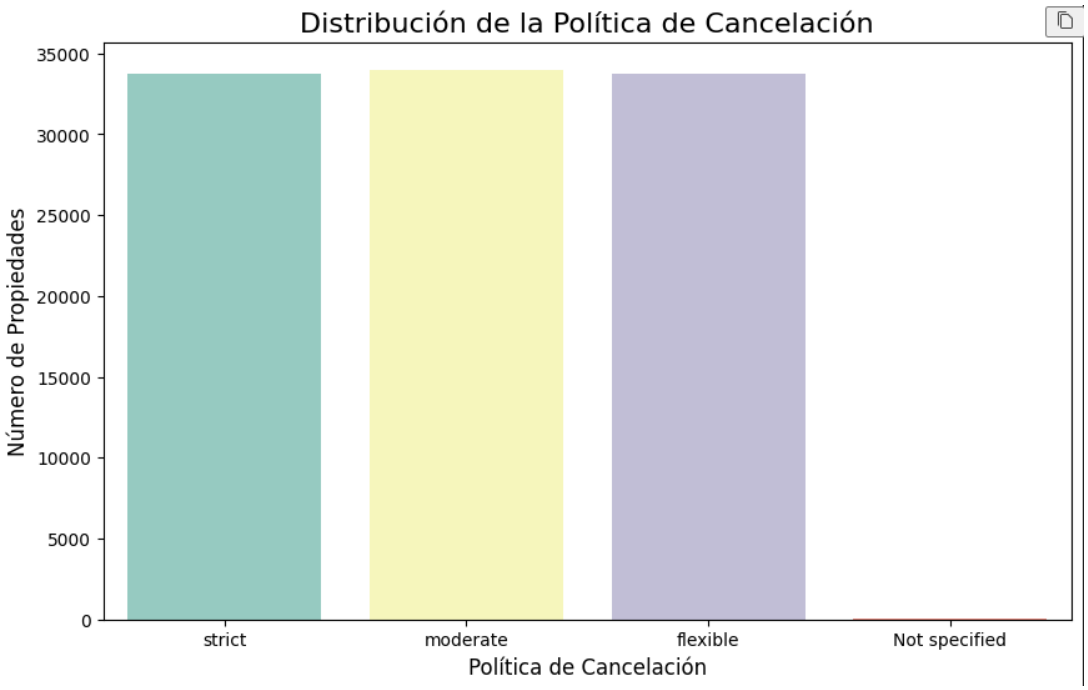
Boxplot: Al igual que el histograma, el boxplot muestra outliers hacia la derecha, con algunas propiedades que tienen miles de reseñas. Esto puede ser indicativo de propiedades muy exitosas o que han estado disponibles en Airbnb durante mucho tiempo.

Variables Categóricas:

Distribución de la Política de Cancelación (cancellation_policy)

Gráfico: Gráfico de Barras

La política de cancelación es otro factor importante que influye en la experiencia del usuario. Puede existir una distribución desigual entre las categorías de políticas de cancelación (por ejemplo, flexible, moderada, estricta).



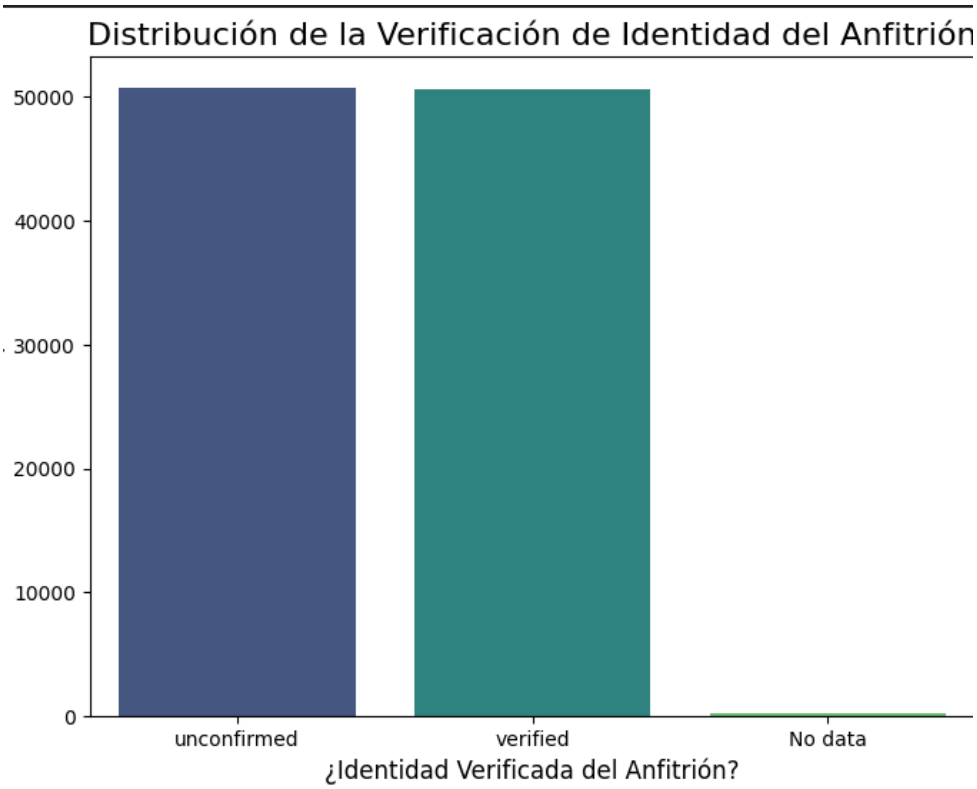
Observaciones:

- Categorías dominantes: Se espera que las políticas de cancelación flexibles sean las más comunes, ya que suelen ser más atractivas para los huéspedes. Pero esta es una excepción ya que podemos ver que la dominante es “moderate”
- Distribución: Podemos ver que la distribución de propiedades es muy parecida en todos, pero podemos ver un poco más la categoría moderate, esto es probablemente porque el precio puede ser mas reducido en estas propiedades

Distribución de la Verificación de Identidad del Anfitrión (host_identity_verified)

Gráfico: Gráfico de Barras

La variable host_identity_verified es una categoría binaria que indica si la identidad del anfitrión ha sido verificada o no. Esto puede ser importante para los huéspedes, ya que proporciona un nivel adicional de confianza.



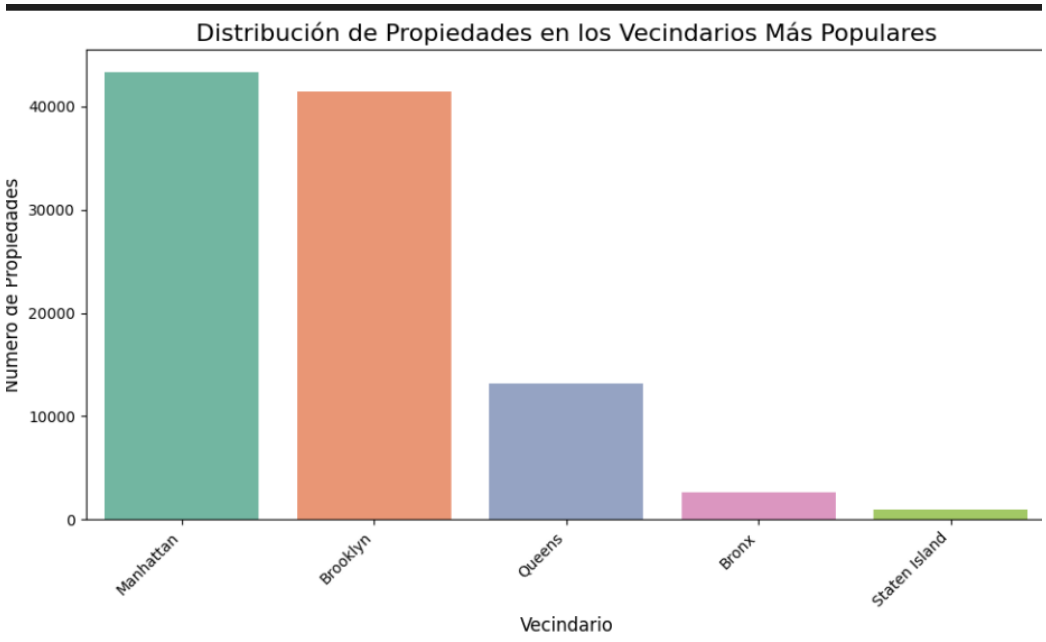
Observaciones:

- Distribución igualitaria: Podemos observar en este grafico que están muy a la par lo cual puede afectar mucho la reservación de propiedades ya que la gente confía mas en las propiedades que tienen el anfitrión confirmado

Distribución del Grupo de Vecindarios (neighbourhood group)

Gráfico: Gráfico de Barras

La variable neighbourhood group describe la agrupación de vecindarios en la ciudad o área en cuestión. Dependiendo de la región, algunos vecindarios pueden ser más populares o tener más propiedades disponibles en Airbnb.



Observaciones:

- Categorías dominantes: Dependiendo de la ciudad, algunos vecindarios o grupos de vecindarios pueden tener una gran concentración de propiedades, como Manhattan o Brooklyn en Nueva York, que tienden a tener más propiedades debido a su popularidad.
- Distribución desequilibrada: Puede haber una distribución desequilibrada, donde ciertos vecindarios o grupos de vecindarios tienen una gran cantidad de propiedades en comparación con otros. Esto puede reflejar la alta demanda en áreas turísticas o populares.

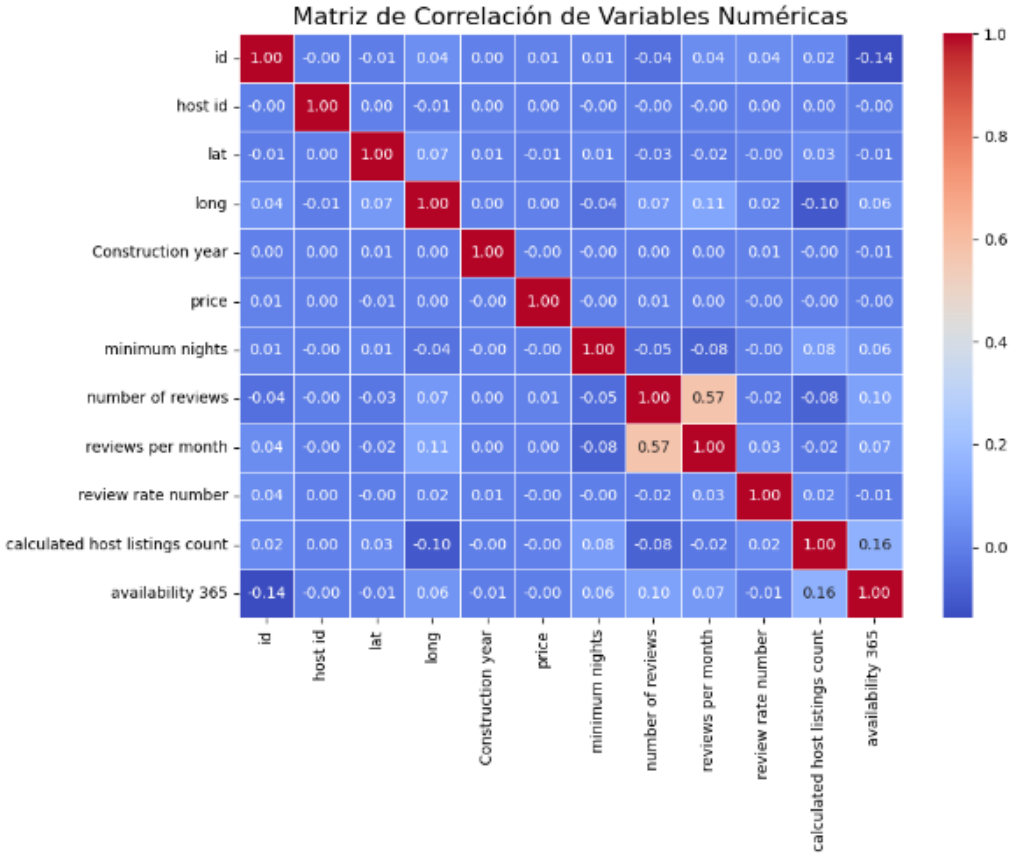
3. Correlación entre Variables

Calcular la Matriz de Correlación

Primero, seleccionamos las columnas numéricas del conjunto de datos y luego calculamos la matriz de correlación.

```
                                     id   host id       lat      long  \
id                             1.000000 -0.000696 -0.008776  0.042385
host id                       -0.000696  1.000000  0.000486 -0.008549
lat                           -0.008776  0.000486  1.000000  0.074072
long                           0.042385 -0.008549  0.074072  1.000000
Construction year              0.000955  0.004495  0.005448  0.001418
price                          0.007083  0.003343 -0.005424  0.003292
minimum nights                 0.005035 -0.002123  0.014912 -0.039212
number of reviews             -0.041531 -0.004711 -0.025201  0.069130
reviews per month              0.035516 -0.001952 -0.018050  0.111136
review rate number             0.036207  0.003663 -0.003595  0.015277
calculated host listings count 0.023488  0.001720  0.032348 -0.104845
availability 365              -0.138316 -0.002610 -0.005054  0.058286

                               Construction year     price  minimum nights  \
id                                      0.000955  0.007083        0.005035
host id                                 0.004495  0.003343       -0.002123
lat                                     0.005448 -0.005424        0.014912
long                                    0.001418  0.003292       -0.039212
Construction year                       1.000000 -0.003672       -0.000209
price                                  -0.003672  1.000000       -0.003354
minimum nights                         -0.000209 -0.003354        1.000000
number of reviews                       0.001709  0.005048       -0.049457
reviews per month                       0.003882  0.003792       -0.079641
review rate number                      0.005406 -0.004454       -0.002482
...
reviews per month                       0.070831
review rate number                     -0.005857
calculated host listings count          0.159208
availability 365                        1.000000
```



Observaciones.

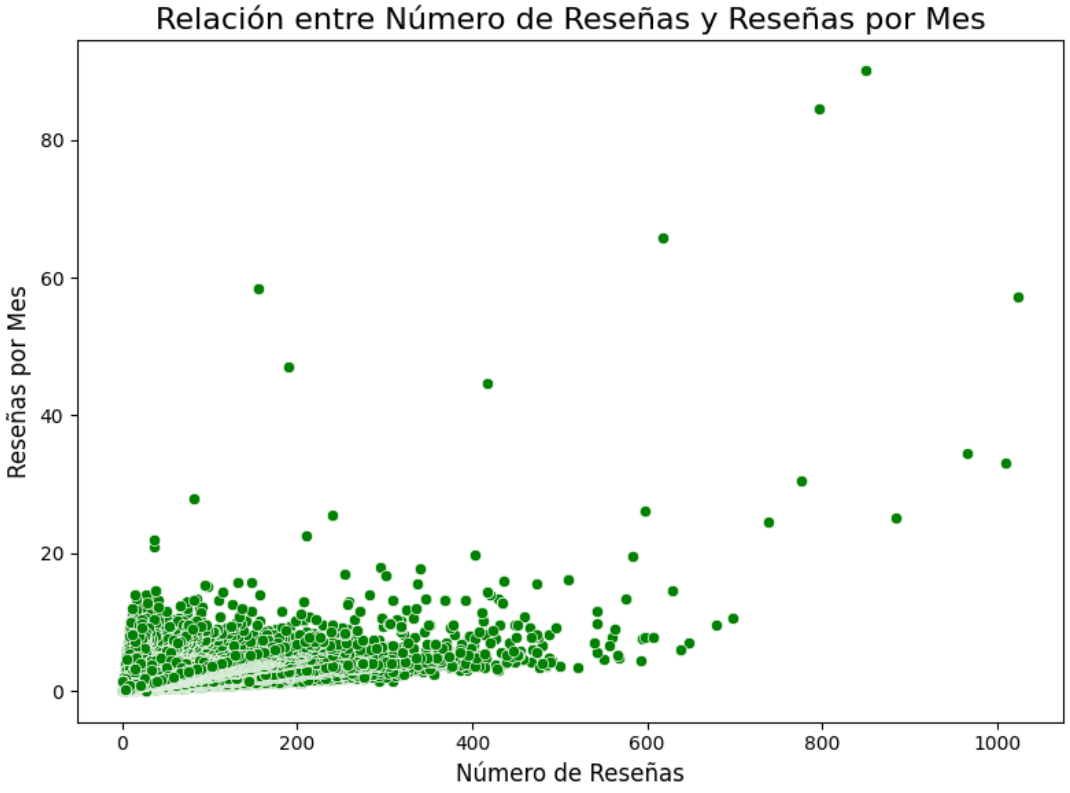
Podemos observar que no hay directamente una relación alta de variables, todas son realmente casi 0 por lo cual no tienen relación muchas, de los que podemos destacar es de “reviews per month” y “number of reviews” son los únicos que tienen mas relación entre ellos

La matriz de correlación proporciona una visión clara de las relaciones entre las variables numéricas y puede ayudarte a identificar variables redundantes o demasiado correlacionadas, lo que te permite mejorar la calidad de tu modelo. Estas correlaciones también pueden ser indicativas de cómo variables como el precio y las reseñas pueden influir en la predicción del comportamiento de las propiedades.

Parejas de Variables

Gráfico de Dispersión de number of reviews vs. reviews per month

La correlación entre number of reviews y reviews per month es bastante fuerte (0.80), lo que sugiere que las propiedades con más reseñas tienden a recibir más reseñas por mes. Esto puede indicar que propiedades más populares o con mayor rotación tienen más reseñas mensuales.



Observaciones.

Debido a la alta correlación positiva, el gráfico muestra un patrón claro, donde las propiedades con un mayor número de reseñas también tienen un número más alto de reseñas mensuales. Las propiedades más populares y con mayor tráfico tienden a recibir más reseñas durante el mes.

Gráfico de Dispersión de price vs. number of reviews

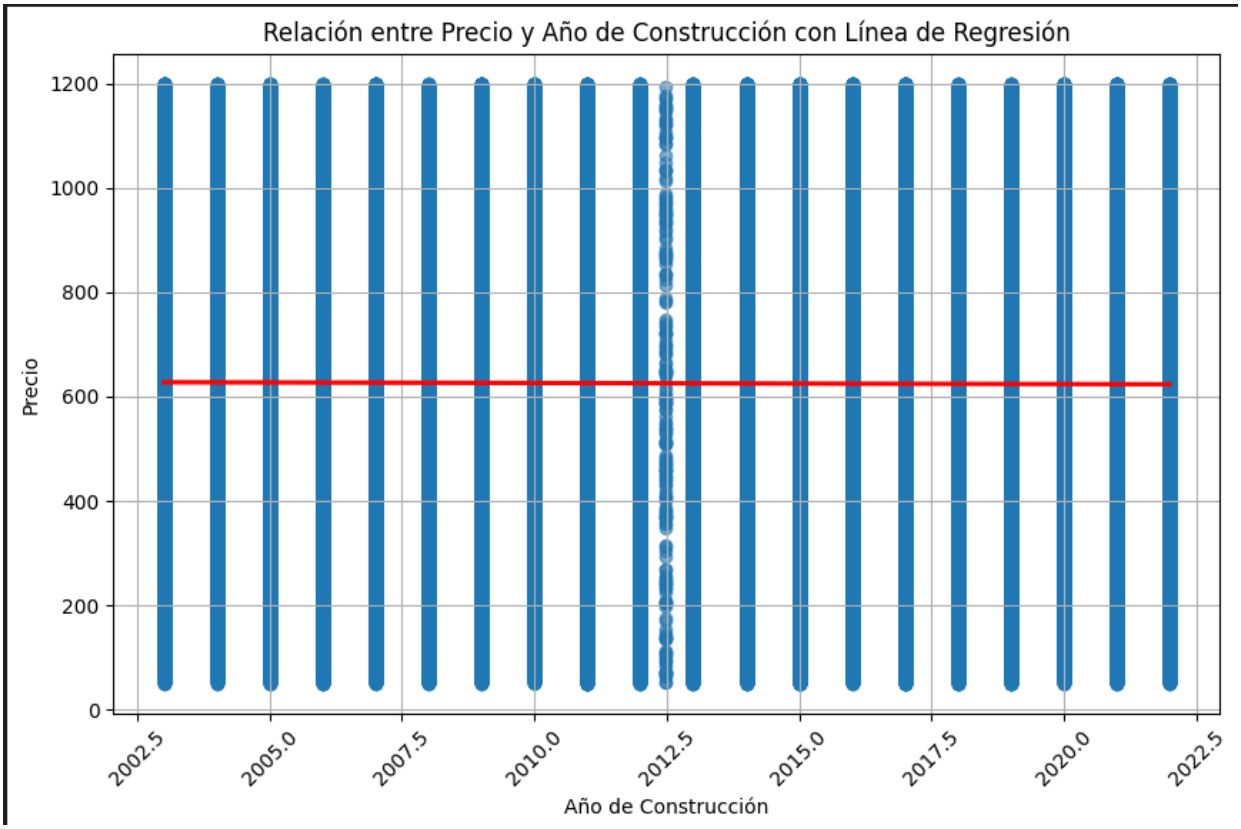
Aunque la correlación entre price y number of reviews es débil (0.01), es interesante visualizar si existe algún patrón entre el precio de las propiedades y la cantidad de reseñas recibidas.

Observaciones:

Podemos observar como en este caso no se relacionan mucho las reseñas conforme el precio de la propiedad, esto quiere decir que no hay un patrón claro sobre esto

Gráfico de Regresión entre el año de construcción y el precio

Esto puede ayudar a saber si el año de creación varia en el precio de la propiedad, algunas veces las propiedades mas nuevas tienden a cobrar mas caro



Observaciones-

Podemos observar en la grafica, que realmente no tiene relación el precio del lugar con el año de construcción, esto puede ser probablemente por el cuidado que se les da a las propiedades, les da posibilidad de cobrarlas al mismo precio

Análisis de Valores Atípicos (Outliers) o Identificación de Outliers

Métodos Utilizados

Se utilizaron en su mayoría Boxplots ya que eran más visuales y fáciles de entender para ver los valores atípicos y también para ver su dispersión de datos. Los puntos fuera de los "bigotes" del boxplot son considerados outliers.

También se utilizó el método de Rango Intercuartílico ya que este es uno de los métodos más comunes para detectar outliers en variables numéricas.

Aunque a pesar de todo esto se tuvo que identificar de los datos atípicos cuales eran correctos, había algunos datos que eran por ejemplo en los precios que esos no se tuvieron que quitar ya que son propiedades extremadamente caras y esto es parte de la base de datos.

Así mismo se tuvo que identificar por ejemplo el de disponibilidad de 365 días del año, algunos eran datos atípicos ya que volaban a 3000 días, y eso no es posible, se tuvo que quitar esos datos atípicos con un método de quitar solamente los que estaban fuera de un rango en específico, en este caso fue del rango de 0 días hasta 365 días.

Algunos otros solamente se sacaron los datos atípicos por el promedio de datos que había para que no se tuvieran que borrar muchos datos y así se pudiera tener una parte mas esencial de la base de datos.

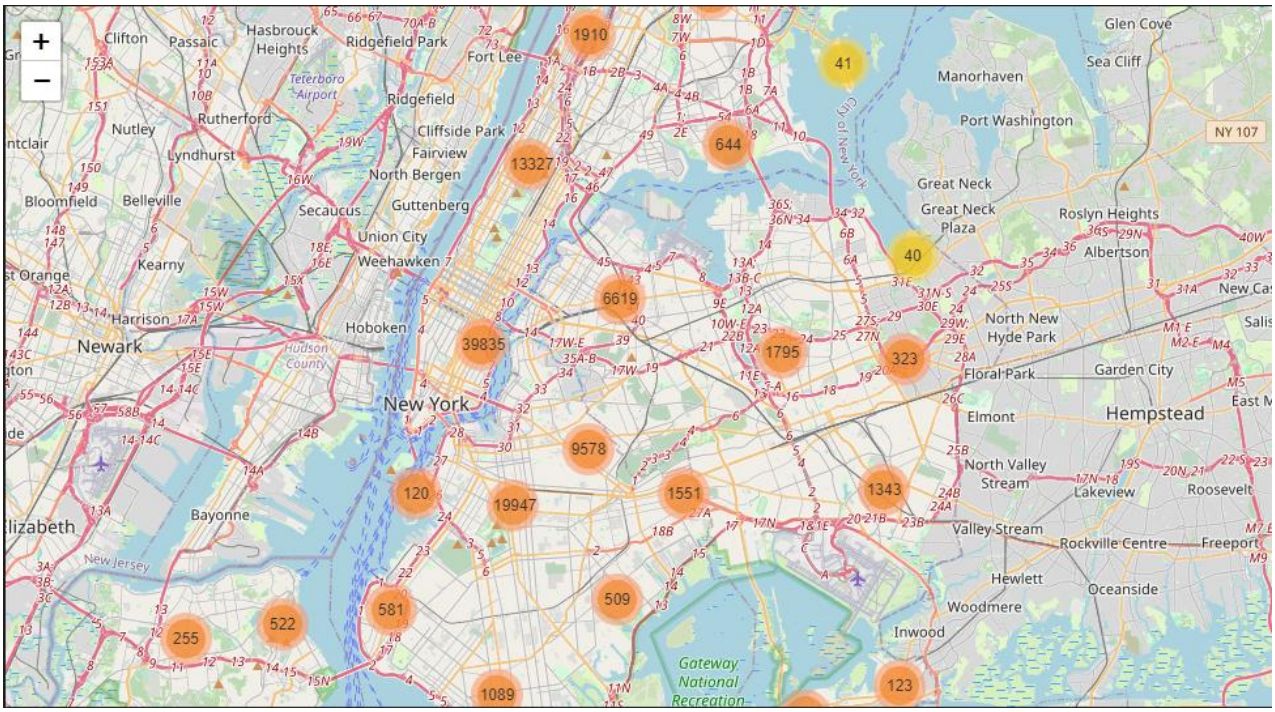
Análisis de Valores Faltantes

Primero se comprobó la existencia de valores faltantes en cada columna, esto se hizo con un código el cual dio una tabla con todas las columnas y su porcentaje de cada columna de valores faltantes

El tratamiento de estos datos tuvo que depender mucho de su contexto, si era una columna que no es esencial para el análisis se pueden eliminar, algunos casos como el precio se pudieron rescatar solamente usando la imputación sacando la media del precio en total y con eso rellenar los valores faltantes

Relación entre Variables Categóricas y Numéricas

Mapa de relación de precios con el lugar



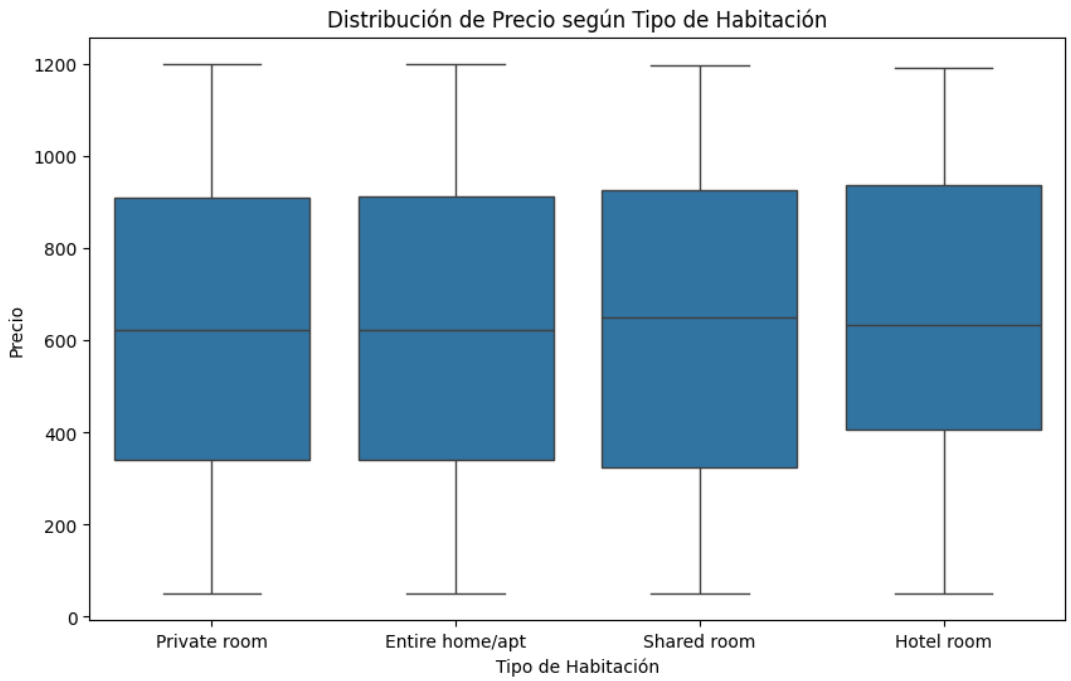
Observaciones.

En este mapa se observa el rango de precios conforme a las ubicaciones, se observa que mientras más oscuro sea el color más caro es el lugar, podemos obtener de este mapa los diferentes lugares donde está más caro el precio, que son los lugares más turísticos y zonas más caras del lugar

Así como también se pueden ver en este mismo mapa cuantas propiedades hay en los diferentes lugares, en este caso podemos ver que en los lugares más céntricos como New York hay muchas más propiedades que en sus afueras, esto debido a que la mayoría de gente prefiere quedarse en lugares turísticos, esto quiere decir que es muy buen lugar para rentar en estos tipos de zonas y también es una buena zona para comprar propiedades y ponerlas en renta a corto plazo por Airbnb

Boxplot para Comparar Tipos de cuartos con respecto al precio

Este sirve para tener una noción mas clara de lo que vale cada tipo de renta con respecto al precio

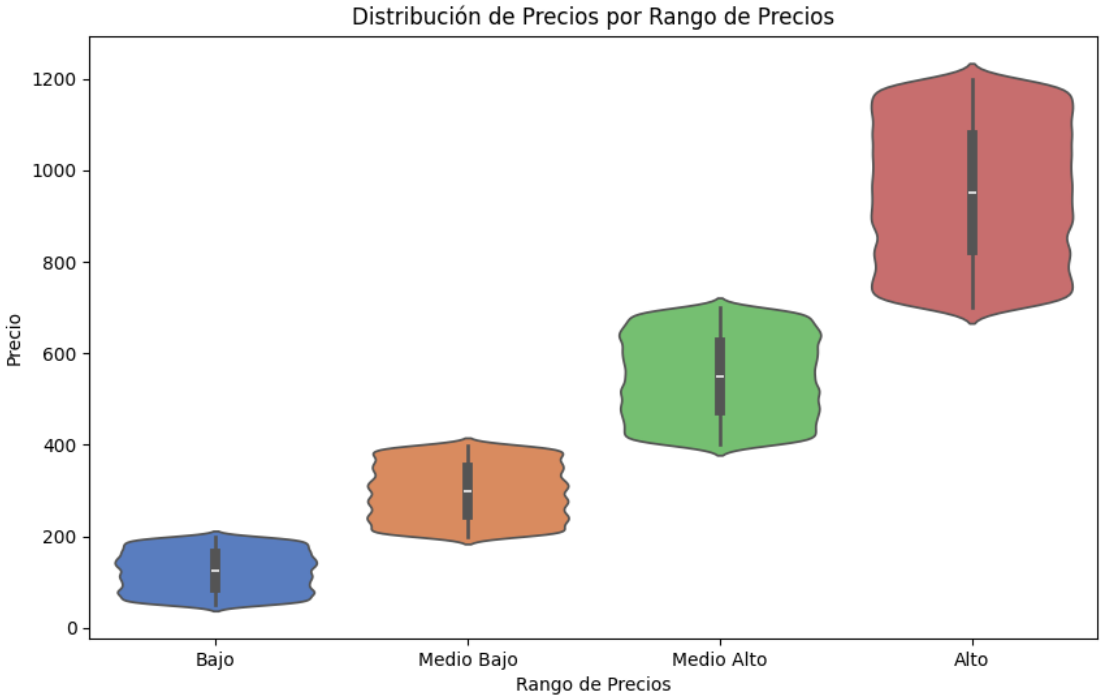


Observaciones.

Podemos observar en cada una de ellas que son muy parecidos los precios, casi no hay diferencias, lo poco que se aprecia es que los cuartos de hotel llegan a ser un poco mas caros a diferencia de los cuartos compartidos que llegan a ser un poco mas baratos

Violin Plot de Rangos de Precios

Este nos va a ayudar un poco mas a comprender los diferentes precios que se manejan y el rango si es bajo o alto el precio



Observaciones.

Podemos observar cómo se distribuyen los precios en cada rango. Es posible que, por ejemplo, el rango "Bajo" tenga una densidad más concentrada en precios cercanos al límite inferior, mientras que el rango "Alto" tenga una distribución más dispersa, con algunos valores extremos (outliers).

Observaciones y Hallazgos Importantes

El análisis de correlación es clave para identificar las variables que afectan o influyen en una variable de interés en tu conjunto de datos. Aquí, nos centramos en la variable que deseamos analizar (precio) y cómo las otras variables del dataset están relacionadas con ella. Para lograr esto, podemos utilizar métodos como la matriz de correlación y heatmaps para visualizar las relaciones entre las variables.

1. Variable a Estudiar

Nuestra variable de interés es precio (price), ya que esta es una de las más relevantes en un conjunto de datos de Airbnb, ya que refleja cuánto cuesta una propiedad en la plataforma.

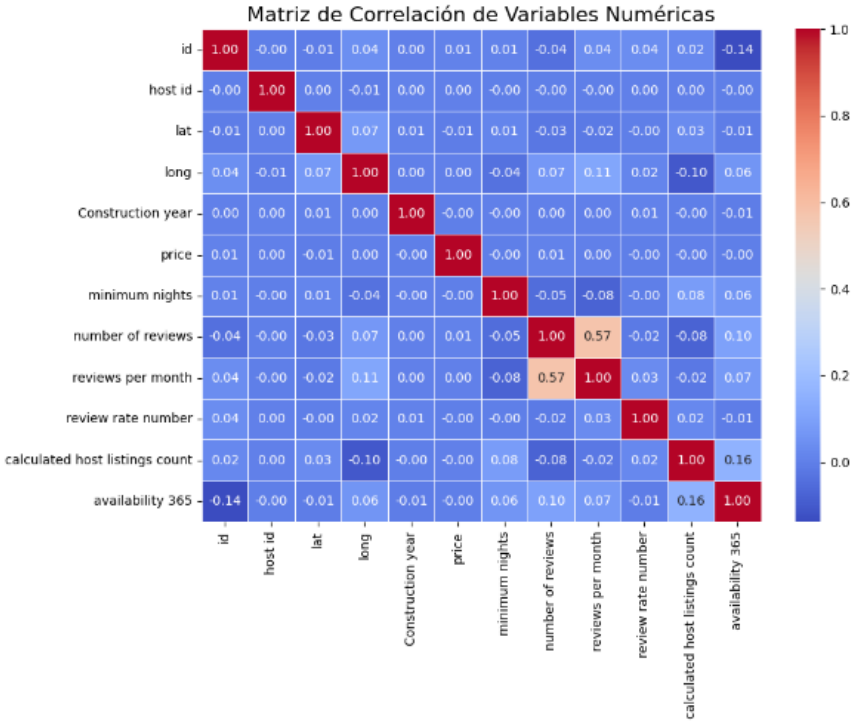
2. Variables que Afectan al Precio

Queremos identificar las variables que tienen la mayor correlación con el precio para ver qué factores podrían estar influyendo en su variación. Algunas variables que comúnmente se relacionan con el precio incluyen:

- Tamaño o tipo de habitación (e.g., room type): Es probable que las habitaciones más grandes o de lujo tengan un precio más alto.
- Ubicación (e.g., neighbourhood group, lat, long): Los lugares más demandados o en áreas turísticas tienden a tener precios más altos.
- Año de construcción (Construction year): Propiedades más nuevas podrían tener un precio más alto.
- Número de reseñas (number of reviews): Propiedades con más reseñas podrían tener más visibilidad y, en consecuencia, precios más altos.

Cálculo y Visualización de la Matriz de Correlación

El primer paso es calcular la matriz de correlación y luego visualizarla utilizando un heatmap.



Podemos observar que con respecto al precio casi no hay variables que estén muy relacionadas a este mismo

Observaciones y Hallazgos Clave

- room type: Es probable que las propiedades con un tipo de habitación más exclusivo (e.g., "Entire home/apt") tengan precios más altos, mientras que las habitaciones compartidas ("Shared room") tengan precios más bajos.
- availability 365: La disponibilidad de 365 días en el año puede estar asociada con un precio más bajo en algunas ocasiones, ya que las propiedades muy disponibles pueden estar en áreas menos demandadas.
- neighbourhood group: Las propiedades ubicadas en áreas populares o turísticas (e.g., "Manhattan") suelen tener un precio más alto en comparación con zonas más alejadas.
- construction year: Las propiedades más nuevas suelen tener un precio más alto, ya que están mejor conservadas y tienen más atractivos modernos.

Conclusiones y Futuras Líneas de Trabajo

Resumen de los Hallazgos Principales

A través de la exploración y análisis de los datos proporcionados por Airbnb, se han logrado varios hallazgos clave que permiten comprender mejor los patrones y factores que afectan la plataforma de alquileres a corto plazo. Los principales hallazgos son los siguientes:

1. Relación entre Precio y Vecindarios:

- Se encontró que los precios de las propiedades varían significativamente según el vecindario. Los vecindarios más centrales y populares tienden a tener precios más altos. Este patrón sugiere que la ubicación es un factor determinante para los precios de las propiedades en Airbnb.

2. Distribución de la Duración de la Estancia:

- La mayoría de las estancias en Airbnb son cortas, con una tendencia predominante hacia estancias de menos de 7 noches. Esto sugiere que los viajeros en la plataforma

prefieren estancias breves, lo cual es relevante para ajustar políticas de precios y promoción.

3. Identificación de Outliers:

    - Se identificaron valores atípicos en variables clave como el precio y la disponibilidad de las propiedades (365 días al año). Estos outliers pueden distorsionar los análisis y es crucial manejarlos adecuadamente (por ejemplo, mediante la eliminación o la transformación de esos datos).

4. Correlación entre Variables:

    - La correlación entre precio y número de reseñas es débil, lo que indica que la cantidad de comentarios no tiene un gran impacto directo sobre el precio de las propiedades.
    - Variables como el año de construcción tienen poca o nula correlación con el precio, lo que sugiere que otros factores, como la ubicación y los servicios, pueden ser más influyentes en el establecimiento de precios.

5. Visualización y Tendencias:

- Se observaron patrones interesantes en los gráficos de dispersión y mapas de calor, como la relación entre el precio y la latitud/longitud de las propiedades, lo que refuerza la importancia de la ubicación para determinar el costo de la estancia.
- Los gráficos de caja (boxplots) mostraron distribuciones de precios desiguales entre diferentes categorías de vecindarios y tipos de habitación, lo que implica que se pueden hacer ajustes específicos por cada categoría para mejorar la estimación de precios.

Cómo Cumplen con los Objetivos Planteados

- El objetivo de analizar los factores que afectan el precio de las propiedades en Airbnb se cumplió al observar cómo variables como ubicación, tipo de habitación y número de reseñas afectan los precios.
- Visualizar las tendencias y patrones en los datos también fue logrado mediante la creación de gráficos de dispersión, mapas de calor y boxplots, que permitieron identificar relaciones entre variables numéricas y categóricas.

- El análisis de outliers y datos faltantes también contribuyó a mejorar la calidad de los datos y a obtener resultados más fiables para la toma de decisiones.

Posibles mejoras.

Mejorar la Calidad de los Datos:

- Es recomendable realizar una limpieza más profunda de los datos, especialmente con respecto a las columnas de texto, como name o house_rules, que pueden contener valores no estandarizados.
- Mejorar la gestión de los valores faltantes, considerando técnicas como la imputación de datos para evitar eliminar demasiadas filas que pueden contener información útil.

Optimización del Precio y Análisis de Mercado:

- Utilizar técnicas de optimización de precios para ayudar a los anfitriones a ajustar sus tarifas dependiendo de las características de su propiedad, la ubicación y la demanda estacional.
- Analizar cómo la demanda estacional afecta los precios, considerando fechas especiales, festividades o eventos en la zona.

Direcciones para Investigaciones Futuras

1. Análisis Temporal:
    - Investigar cómo los precios y las tendencias de disponibilidad cambian a lo largo del tiempo. Analizar si existen fluctuaciones estacionales, como en los meses de verano o durante grandes eventos locales, y cómo las políticas de cancelación afectan estas fluctuaciones.
2. Estudio de la Demanda y Oferta en Diversas Regiones:
    - Explicar las diferencias en precios y disponibilidad según la demanda en diferentes ciudades o países. El análisis podría incluir aspectos como la influencia de la economía local o las políticas gubernamentales sobre el mercado de alquileres.

3. Análisis de Opiniones y Reseñas:
    - Profundizar en la relación entre las reseñas de los usuarios y el precio de las propiedades. Un análisis de sentimientos o el análisis de palabras clave en las reseñas podría proporcionar información adicional sobre los aspectos que los huéspedes valoran más.
4. Impacto de las Políticas de Cancelación:
    - Estudiar cómo diferentes políticas de cancelación (flexibles, estrictas, moderadas) afectan tanto el precio como la ocupación de las propiedades.

Este proyecto ha proporcionado una visión profunda sobre cómo varias características de las propiedades en Airbnb influyen en el precio, y cómo visualizar y procesar los datos para obtener información valiosa. Las futuras investigaciones pueden mejorar los modelos predictivos y explorar cómo los anfitriones pueden optimizar sus precios y políticas para maximizar su ocupación y ganancias.

Referencias:

https://www.datacamp.com/es/tutorial/types-of-data-plots-and-how-to-create-them-in-python

https://aprendeconalf.es/docencia/python/manual/matplotlib/