Airbnb

November 11, 2025

```
[2]: import pandas as pd
```

```
[3]: import numpy as np
```

```
[10]: df = pd.read_csv("airbnb.csv", dtype="object")
```

```
[11]: df.shape
```

```
[11]: (31611, 26)
```

```
[12]: df.columns
```

```
[12]: Index(['id', 'NAME', 'host id', 'host_identity_verified', 'host name',
             'neighbourhood group', 'neighbourhood', 'lat', 'long', 'country',
             'country code', 'instant_bookable', 'cancellation_policy', 'room type',
             'Construction year', 'price', 'service fee', 'minimum nights',
             'number of reviews', 'last review', 'reviews per month',
             'review rate number', 'calculated host listings count',
             'availability 365', 'house_rules', 'license'],
           dtype='object')
```

```
[13]: df.columns = ['id', 'NAME', 'host_id', 'host_identity_verified', 'host_name',
             'neighbourhood_group', 'neighbourhood', 'lat', 'long', 'country',
             'country_code', 'instant_bookable', 'cancellation_policy', 'room_type',
             'construction_year', 'price', 'service_fee', 'minimum_nights',
             'number_of_reviews', 'last_review', 'reviews_per_month',
             'review_rate_number', 'calculated_host_listings_count',
             'availability_365', 'house_rules', 'license']
```

```
[14]: df.columns
```

```
[14]: Index(['id', 'NAME', 'host_id', 'host_identity_verified', 'host_name',
             'neighbourhood_group', 'neighbourhood', 'lat', 'long', 'country',
             'country_code', 'instant_bookable', 'cancellation_policy', 'room_type',
             'construction_year', 'price', 'service_fee', 'minimum_nights',
             'number_of_reviews', 'last_review', 'reviews_per_month',
             'review_rate_number', 'calculated_host_listings_count',
```

```
          'availability_365', 'house_rules', 'license'],
        dtype='object')
```

[15]: `df.isna().sum()`

```
[15]: id                                   0
      NAME                               114
      host_id                              1
      host_identity_verified             159
      host_name                          225
      neighbourhood_group                 28
      neighbourhood                       17
      lat                                  9
      long                                 9
      country                             88
      country_code                        80
      instant_bookable                    80
      cancellation_policy                 51
      room_type                            1
      construction_year                  147
      price                               99
      service_fee                         87
      minimum_nights                     179
      number_of_reviews                   35
      last_review                       5239
      reviews_per_month                 5225
      review_rate_number                 114
      calculated_host_listings_count      84
      availability_365                   179
      house_rules                      10496
      license                          31610
      dtype: int64
```

[16]:
```
df.columns.str.lower()
df.columns.str.replace(" ", "_")
```

```
[16]: Index(['id', 'NAME', 'host_id', 'host_identity_verified', 'host_name',
            'neighbourhood_group', 'neighbourhood', 'lat', 'long', 'country',
            'country_code', 'instant_bookable', 'cancellation_policy', 'room_type',
            'construction_year', 'price', 'service_fee', 'minimum_nights',
            'number_of_reviews', 'last_review', 'reviews_per_month',
            'review_rate_number', 'calculated_host_listings_count',
            'availability_365', 'house_rules', 'license'],
          dtype='object')
```

[17]: `df["reviews_per_month"] = df["reviews_per_month"].fillna(0)`

```
[19]: print("Antes:", df.shape)
```

Antes: (31611, 26)

```
[21]: df = df.drop_duplicates()
```

```
[22]: print("Después:", df.shape)
```

Después: (31611, 26)

```
[23]: # Ver si hay IDs repetidos
      df["id"].duplicated().sum()
```

```
[23]: 0
```

```
[24]: df[df.duplicated(subset=["host_name", "neighbourhood"], keep=False)]
```

```
[24]:              id                                        NAME       host_id  \
      5        1004098         Large Cozy 1 BR Apartment In Midtown East  45498551794
      7        1005202                              BlissArtsSpace!      90821839709
      11       1007411              Beautiful 1br on Upper West Side      18824631834
      12       1007964              Central Manhattan/near Broadway      88136055909
      13       1008516  Lovely Room 1, Garden, Best Area, Legal rental  26802410424
      ...         ...                                         ...           ...
      31596  18451807         BEAUTIFUL GREAT BEDROOM IN HARLEM      51000565775
      31598  18452911   ENJOY NYC IN  1 BDRM CONDO HEART OF MIDTOWN  17079982322
      31599  18453464  New Condo 2 bedrms /1,5bath elevator & balcony  70499550910
      31603  18455673         Private Bedroom in a Modern Building  92689244100
      31608  18458434                                Little Secret      31796844821

             host_identity_verified          host_name neighbourhood_group  \
      5                     verified            Michelle           Manhattan
      7                   unconfirmed                Emma            Brooklyn
      11                    verified                Alan           Manhattan
      12                    verified                 NaN           Manhattan
      13                    verified               Darcy             brookln
      ...                        ...                 ...                 ...
      31596                 verified  Maxime C And Winie C           Manhattan
      31598               unconfirmed             Deborah           Manhattan
      31599               unconfirmed                  Ha           Manhattan
      31603               unconfirmed                 Van            Brooklyn
      31608                 verified              Thomas           Manhattan

                neighbourhood       lat       long         country  ...  \
      5            Murray Hill  40.74767    -73.975  United States  ...
      7      Bedford-Stuyvesant  40.68688  -73.95596  United States  ...
      11        Upper West Side  40.80316  -73.96545  United States  ...
```

```
12           Hell's Kitchen   40.76076  -73.98867   United States  …
13             South Slope    40.66829  -73.98779   United States  …
…                     …           …          …             …   …
31596               Harlem    40.81639  -73.94414   United States  …
31598     Theater District    40.76205    -73.982   United States  …
31599            Chinatown    40.71409  -73.99434   United States  …
31603             Flatbush    40.63076  -73.95851   United States  …
31608       Upper West Side   40.77705  -73.98118   United States  …

       service_fee minimum_nights number_of_reviews last_review  \
5           $115                3                74   6/22/2019
7           $212               45                49   10/5/2017
11          $121                5                53   6/22/2019
12          $143                2               188   6/23/2019
13          $116                4               167   6/24/2019
…             …                 …                …          …
31596        $44                2                40   6/26/2019
31598       $219               30                 0         NaN
31599       $172                1                 2   10/7/2018
31603        $32                1                19  12/27/2018
31608       $222                2                81    7/7/2019

       reviews_per_month review_rate_number calculated_host_listings_count  \
5                   0.59                  3                              1
7                    0.4                  5                              1
11                  0.43                  4                              1
12                   1.5                  4                              1
13                  1.34                  4                              3
…                      …                  …                              …
31596               2.75                  3                              3
31598                  0                  3                              4
31599                0.2                  2                              4
31603               1.57                  2                              2
31608               5.63                  2                              1

       availability_365                                      house_rules  \
5                   374                 No smoking, please, and no drugs.
7                   219   House Guidelines for our BnB We are delighted …
11                  163   My ideal guests would be warm, friendly, and r…
12                  258   - One of the bedroom closets is not accessible…
13                   47                                               NaN
…                     …                                                 …
31596               177   NO PETS!  NO SMOKING ON PREMISES!  NO PARTY! A…
31598                 0                                               NaN
31599               191   Respect yourself respect others please clean u…
31603                98   No smoking at all. Keep the room and apartment…
31608                 6   No smoking in house. No Pets No parties or hos…
```

```
      license
5         NaN
7         NaN
11        NaN
12        NaN
13        NaN
…         …
31596     NaN
31598     NaN
31599     NaN
31603     NaN
31608     NaN

[15825 rows x 26 columns]
```

```python
# Anfitriones con más de un vivienda
hosts_con_multiples = df["host_name"].value_counts()
hosts_con_multiples[hosts_con_multiples > 1].head(10)
```

```
host_name
Michael    278
David      246
Sarah      176
John       162
Alex       162
Daniel     160
Mike       141
Andrew     138
Chris      134
Kara       131
Name: count, dtype: int64
```

```python
# DataFrame Agregado
# Limpiar símbolos y convertir a float
df["price"] = df["price"].replace('[\$,]', '', regex=True).astype(float)

df_hosts = df.groupby("host_name").agg({
    "id": "count",
    "price": "mean",
    "neighbourhood": pd.Series.nunique
}).reset_index()

df_hosts.columns = ["host_name", "num_propiedades", "precio_medio",
  ↪"num_barrios"]
display(df_hosts.head(10))
```

```
          host_name  num_propiedades  precio_medio  num_barrios
0       (Ari) HENRY LEE                1   1179.000000            1
1  (Email hidden by Airbnb)           4    488.500000            4
2           (Mary) Haiy               1    465.000000            1
3      @ Art House Monique            3    664.666667            1
4                     A               7    772.142857            6
5                 A & H               1    132.000000            1
6              A. Kaylee              2    641.000000            2
7                  A.B.               3    399.333333            2
8                  A.J.               5    594.600000            5
9                  A.R.               1    150.000000            1
```

[31]:
```python
# Normalizar textos
df["neighbourhood"] = df["neighbourhood"].str.strip().str.title()
df["room_type"] = df["room_type"].str.strip().str.lower()

# Correcciones específicas (por si existieran)
df["room_type"] = df["room_type"].replace({
    "entire home/apt": "entire home",
    "private room ": "private room",
    "shared room ": "shared room"
})

print(df["neighbourhood"].unique()[:10])
print(df["room_type"].unique())
```

```
['Kensington' 'Midtown' 'Harlem' 'Clinton Hill' 'East Harlem'
 'Murray Hill' 'Bedford-Stuyvesant' "Hell'S Kitchen" 'Upper West Side'
 'Chinatown']
['private room' 'entire home' 'shared room' nan]
```

[33]:
```python
#   EXPORTAR DATAFRAMES LIMPIOS Y RESÚMENES

# 1 Exportar dataset limpio completo
df.to_csv("airbnb_limpio.csv", index=False)
df.to_html("airbnb_limpio.html", index=False)

# 2 Exportar resumen de anfitriones
df_hosts.to_csv("airbnb_hosts_resumen.csv", index=False)
df_hosts.to_html("airbnb_hosts_resumen.html", index=False)

# 3 Mostrar confirmación final
print(" Cuaderno 2 terminado.")
print(f" Archivos exportados correctamente en el directorio de trabajo.\n")
print(f" airbnb_limpio.csv / airbnb_limpio.html")
print(f" airbnb_hosts_resumen.csv / airbnb_hosts_resumen.html")
print(f"\nFilas finales en dataset limpio: {df.shape[0]}")
```

```
print(f"Columnas totales: {df.shape[1]}")
print(f"Hosts únicos: {df['host_name'].nunique()}")
print(f"Media general de precios: {df['price'].mean():.2f}")
```

Cuaderno 2 terminado.
Archivos exportados correctamente en el directorio de trabajo.

airbnb_limpio.csv / airbnb_limpio.html
airbnb_hosts_resumen.csv / airbnb_hosts_resumen.html

Filas finales en dataset limpio: 31611
Columnas totales: 26
Hosts únicos: 8279
Media general de precios: 622.74

[ ]: