Final_Project_Spanish

December 22, 2023

Proyecto final

IBM SkillsBuild Europa - Análisis de datos

1 Requisitos

- Conocimientos de Python, Power BI o Tableau.
- Comprensión de la limpieza de datos.
- Comprensión de la visualización de datos.

Nivel de ejercicio: Intermedio

Duración: aproximadamente 3 horas

1.0.1 Análisis de datos de Airbnb:

Objetivo: En este ejercicio, practicarás el análisis de datos sobre un conjunto de datos abiertos procedentes de Airbnb. Algunas de las tareas incluyen:

- limpieza de datos,
- transformación de datos y
- visualización de datos.

Resumen sobre los datos de Airbnb: Los criterios principales de la gente cuando visita lugares nuevos son contar con alojamiento y comida a precios razonables. Airbnb (Air-Bed-Breakfast) es un mercado en línea creado para satisfacer esta necesidad, pues permite a la gente alquilar sus casas durante plazos cortos. Estos servicios se ofrecen a un precio relativamente inferior al de los hoteles y en diversas ubicaciones geográficas. Además, personas de todo el mundo prefieren el servicio hogareño y económico.

Fuente de los datos Puedes obtener el dataset para realizar este proyecto en el siguiente enlace: https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata

Este dataset contiene información sobre los alojamientos, tal como el barrio, el tipo de habitación, el precio, la disponibilidad, las opiniones, los gastos de servicio, la política de cancelación y las normas de uso de la casa.

¡Te deseamos lo mejor en tu análisis de los datos de Airbnb!

1.1 Tarea 1: Carga de datos (Python)

1. Lee el fichero csv y cárgarlo en un dataframe de pandas.
2. Visualiza las cinco primeras filas de tu dataframe.
3. Visualize los tipos de datos de las columnas.

[2]:
```
## Leer directorio
import os
dir = os.getcwd()
os.listdir(dir)
```

[2]:
```
['Final_Project_Spanish.ipynb', 'Airbnb_Open_Data.csv']
```

[3]:
```
## Lee el fichero csv.
import pandas as pd

df = pd.read_csv('Airbnb_Open_Data.csv')
```

```
/tmp/ipykernel_7767/1697345662.py:4: DtypeWarning: Columns (25) have mixed
types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv('Airbnb_Open_Data.csv')
```

[4]:
```
## Muestra las primeras 5 filas.
df.head()
```

[4]:

	id	NAME	host id	\
0	1001254	Clean & quiet apt home by the park	80014485718	
1	1002102	Skylit Midtown Castle	52335172823	
2	1002403	THE VILLAGE OF HARLEM…NEW YORK !	78829239556	
3	1002755	NaN	85098326012	
4	1003689	Entire Apt: Spacious Studio/Loft by central park	92037596077	

	host_identity_verified	host name	neighbourhood group	neighbourhood	\
0	unconfirmed	Madaline	Brooklyn	Kensington	
1	verified	Jenna	Manhattan	Midtown	
2	NaN	Elise	Manhattan	Harlem	
3	unconfirmed	Garry	Brooklyn	Clinton Hill	
4	verified	Lyndon	Manhattan	East Harlem	

	lat	long	country	…	service fee	minimum nights	\
0	40.64749	-73.97237	United States	…	$193	10.0	
1	40.75362	-73.98377	United States	…	$28	30.0	
2	40.80902	-73.94190	United States	…	$124	3.0	
3	40.68514	-73.95976	United States	…	$74	30.0	
4	40.79851	-73.94399	United States	…	$41	10.0	

	number of reviews	last review	reviews per month	review rate number	\
0	9.0	10/19/2021	0.21	4.0	

```
1              45.0   5/21/2022              0.38                4.0
2               0.0         NaN               NaN                5.0
3             270.0    7/5/2019              4.64                4.0
4               9.0  11/19/2018              0.10                3.0

  calculated host listings count  availability 365  \
0                            6.0             286.0
1                            2.0             228.0
2                            1.0             352.0
3                            1.0             322.0
4                            1.0             289.0

                                        house_rules license
0  Clean up and treat the home the way you'd like…     NaN
1  Pet friendly but please confirm with me if the…     NaN
2  I encourage you to use my kitchen, cooking and…     NaN
3                                              NaN     NaN
4  Please no smoking in the house, porch or on th…     NaN

[5 rows x 26 columns]
```

[5]:
```
## Muestra los tipos de datos.
df.dtypes
```

[5]:
```
id                               int64
NAME                            object
host id                          int64
host_identity_verified          object
host name                       object
neighbourhood group             object
neighbourhood                   object
lat                            float64
long                           float64
country                         object
country code                    object
instant_bookable                object
cancellation_policy             object
room type                       object
Construction year              float64
price                           object
service fee                     object
minimum nights                 float64
number of reviews              float64
last review                     object
reviews per month              float64
review rate number             float64
calculated host listings count float64
```

```
availability 365                float64
house_rules                      object
license                          object
dtype: object
```

1.2 Tarea 2a: Limpieza de datos (cualquier herramienta)

1. Elimina las columnas no deseadas del dataframe, entre ellas se incluyen `host id`, `id`, `country` y `country code`.
2. Indica la razón por la cual se han eliminado estas columnas para tu análisis de datos.

Si utilizas Python para este ejercicio, incluye el código que hayas utilizado en las celdas siguientes. Si utilizas cualquier otra herramienta, incluye capturas de pantalla tomadas antes y después de eliminar las columnas.

[6]:
```
## Columnas a eliminar
colums_to_drop = ['host id', 'id', 'country', 'country code']


## Eliminar
df = df.drop(colums_to_drop, axis=1)
```

[7]:
```
df.head()
```

[7]:
```
                                              NAME host_identity_verified  \
0                Clean & quiet apt home by the park            unconfirmed
1                             Skylit Midtown Castle               verified
2               THE VILLAGE OF HARLEM…NEW YORK !                   NaN
3                                              NaN            unconfirmed
4  Entire Apt: Spacious Studio/Loft by central park               verified

  host name neighbourhood group neighbourhood       lat      long  \
0  Madaline            Brooklyn    Kensington  40.64749 -73.97237
1     Jenna           Manhattan       Midtown  40.75362 -73.98377
2     Elise           Manhattan        Harlem  40.80902 -73.94190
3     Garry            Brooklyn  Clinton Hill  40.68514 -73.95976
4    Lyndon           Manhattan   East Harlem  40.79851 -73.94399

  instant_bookable cancellation_policy        room type  …  service fee  \
0            False              strict     Private room  …         $193
1            False            moderate  Entire home/apt  …          $28
2             True            flexible     Private room  …         $124
3             True            moderate  Entire home/apt  …          $74
4            False            moderate  Entire home/apt  …          $41

  minimum nights number of reviews  last review  reviews per month  \
0           10.0               9.0   10/19/2021               0.21
1           30.0              45.0    5/21/2022               0.38
```

```
2            3.0               0.0          NaN                NaN
3           30.0             270.0      7/5/2019               4.64
4           10.0               9.0    11/19/2018               0.10

  review rate number  calculated host listings count  availability 365  \
0                4.0                            6.0             286.0
1                4.0                            2.0             228.0
2                5.0                            1.0             352.0
3                4.0                            1.0             322.0
4                3.0                            1.0             289.0

                                          house_rules  license
0  Clean up and treat the home the way you'd like…      NaN
1  Pet friendly but please confirm with me if the…      NaN
2  I encourage you to use my kitchen, cooking and…      NaN
3                                                NaN      NaN
4  Please no smoking in the house, porch or on th…      NaN

[5 rows x 22 columns]
```

[8]:
```
"""Motivos por los cuales se eliminan las columnas:
1) Irrelevancia para el análisis. 'country' y 'country code' no son relevantes␣
 ↪para el  análisis debido a que se centran en datos específicos de una␣
 ↪localización geográfica concreta.
Como todos pertenecen a la misma localización, no aporta información.
'host id' y 'id' pueden ser identificadores únicos que no aportan información␣
 ↪sustancial al análisis.

2) Protección de la privacidad. Al eliminar identificadores únicos como 'host␣
 ↪id' y 'id' podemos argumentar que es una medida de privacidad, especialmente␣
 ↪si se comparte o publica el análisis.
"""
```

[8]:
```
"Motivos por los cuales se eliminan las columnas:\n1) Irrelevancia para el
análisis. 'country' y 'country code' no son relevantes para el  análisis debido
a que se centran en datos específicos de una localización geográfica
concreta.\nComo todos pertenecen a la misma localización, no aporta
información.\n'host id' y 'id' pueden ser identificadores únicos que no aportan
información sustancial al análisis.\n\n2) Protección de la privacidad. Al
eliminar identificadores únicos como 'host id' y 'id' podemos argumentar que es
una medida de privacidad, especialmente si se comparte o publica el análisis.\n"
```

1.3 Tarea 2b: Limpieza de datos (Python)

- Compueba si hay valores nulos y muestra el recuento en orden ascendente. **Si faltan valores, imputa los valores como consideres.**
- Comprueba si hay valores duplicados y elimínalos.

- Muestra el número total de registros antes y después de eliminar los duplicados.

[9]:
```
## Comprueba si hay valores nulos y muestra el recuento en orden ascendente.
df.isnull().sum().sort_values()

## Imputar valores nulos con el valor medio de la columna
"""
## Identificar columnas numericas
num_columns = df.select_dtypes(include=['int64']).columns

## Imputar valores nulos solo en columnas numericas con el valor medio de la␣
 ↪columna
df[num_columns] = df[num_columns].fillna(df[num_columns].mean())"""
```

[9]:
```
"\n## Identificar columnas numericas\nnum_columns =
df.select_dtypes(include=['int64']).columns\n\n## Imputar valores nulos solo en
columnas numericas con el valor medio de la columna\ndf[num_columns] =
df[num_columns].fillna(df[num_columns].mean())"
```

[10]:
```
## Comprueba si hay valores duplicados y elimínalos.
df.duplicated()
```

[10]:
```
0         False
1         False
2         False
3         False
4         False
          …
102594     True
102595     True
102596     True
102597     True
102598     True
Length: 102599, dtype: bool
```

[11]:
```
## Comprueba si hay valores duplicados y elimínalos.
df.drop_duplicates()
```

[11]:
```
                                                    NAME  \
0                     Clean & quiet apt home by the park
1                                  Skylit Midtown Castle
2                    THE VILLAGE OF HARLEM…NEW YORK !
3                                                   NaN
4         Entire Apt: Spacious Studio/Loft by central park
…                                                   …
102053                Cozy bright room near Prospect Park
102054          Private Bedroom with Amazing Rooftop View
```

```
102055     Pretty Brooklyn One-Bedroom for 2 to 4 people
102056        Room & private bathroom in historic Harlem
102057                                   Rosalee Stewart

       host_identity_verified host name neighbourhood group  \
0                unconfirmed  Madaline            Brooklyn
1                   verified     Jenna           Manhattan
2                        NaN     Elise           Manhattan
3                unconfirmed     Garry            Brooklyn
4                   verified    Lyndon           Manhattan
…                       …       …                  …
102053           unconfirmed    Mariam            Brooklyn
102054              verified      Trey            Brooklyn
102055              verified   Michael            Brooklyn
102056           unconfirmed   Shireen           Manhattan
102057              verified   Stanley           Manhattan

            neighbourhood       lat      long instant_bookable  \
0              Kensington  40.64749 -73.97237            False
1                 Midtown  40.75362 -73.98377            False
2                  Harlem  40.80902 -73.94190             True
3             Clinton Hill  40.68514 -73.95976             True
4              East Harlem  40.79851 -73.94399            False
…                     …       …       …                …
102053            Flatbush  40.64945 -73.96108             True
102054            Bushwick  40.69872 -73.92718            False
102055  Bedford-Stuyvesant  40.67810 -73.90822             True
102056              Harlem  40.81248 -73.94317             True
102057              Harlem  40.81315 -73.94747            False

       cancellation_policy        room type  …  service fee minimum nights  \
0                   strict     Private room  …        $193            10.0
1                moderate  Entire home/apt  …         $28            30.0
2                flexible     Private room  …        $124             3.0
3                moderate  Entire home/apt  …         $74            30.0
4                moderate  Entire home/apt  …         $41            10.0
…                     …              …  …          …             …
102053           moderate     Private room  …          NaN            7.0
102054           flexible     Private room  …          NaN            1.0
102055           moderate  Entire home/apt  …          NaN            2.0
102056              strict     Private room  …          NaN            2.0
102057           flexible  Entire home/apt  …          NaN            4.0

       number of reviews  last review  reviews per month review rate number  \
0                    9.0   10/19/2021               0.21                4.0
1                   45.0    5/21/2022               0.38                4.0
2                    0.0          NaN                NaN                5.0
```

```
3                  270.0     7/5/2019               4.64                 4.0
4                    9.0   11/19/2018               0.10                 3.0
…                    …           …                  …                    …
102053              12.0    3/27/2019               0.44                 5.0
102054              19.0    8/31/2017               0.72                 3.0
102055              50.0    6/26/2019               3.12                 4.0
102056               0.0          NaN                NaN                 1.0
102057              22.0    6/15/2019               0.85                 4.0

        calculated host listings count  availability 365  \
0                                  6.0             286.0
1                                  2.0             228.0
2                                  1.0             352.0
3                                  1.0             322.0
4                                  1.0             289.0
…                                  …               …
102053                             1.0               0.0
102054                             2.0               0.0
102055                             2.0             235.0
102056                             1.0               0.0
102057                             1.0             238.0

                                              house_rules  license
0       Clean up and treat the home the way you'd like…       NaN
1       Pet friendly but please confirm with me if the…       NaN
2       I encourage you to use my kitchen, cooking and…       NaN
3                                                   NaN       NaN
4       Please no smoking in the house, porch or on th…       NaN
…                                                   …       …
102053  Shoes off Clean After yourself Turn Lights and…       NaN
102054                                          #NAME?        NaN
102055  * Check out: 10am * We made an effort to keep …       NaN
102056  Each of us is working and/or going to school a…       NaN
102057  Please remember that this is a residential bui…       NaN

[99163 rows x 22 columns]
```

[12]:
```
## Muestra el número total de registros antes y después de eliminar los␣
 ↪duplicados.

duplicate_data = df.duplicated()
print(df[duplicate_data])

df_without_duplicate = df.drop_duplicates()

# Verificar el número de registros antes y después de eliminar duplicados
```

```
print("\nNúmero total de registros antes de eliminar duplicados:",␣
 ↪len(duplicate_data))
print("Número total de registros después de eliminar duplicados:",␣
 ↪len(df_without_duplicate))
```

```
                                                    NAME  \
70825                    Penthouse Designer Loft Brooklyn
71877                  Brooklyn, Clinton Hill, Private rm
73083   Enormous and illuminated on top floor (long term)
73137          Manhattan Club 1 Bedroom (Sleeps 4 adults)
73191    SHARE;CHEAPEST;Pure;Cozy;Safe;Silent IN NEW YORK
…                                                     …
102594                       Spare room in Williamsburg
102595                     Best Location near Columbia U
102596                    Comfy, bright room in Brooklyn
102597                   Big Studio-One Stop from Midtown
102598                               585 sf Luxury Studio

       host_identity_verified    host name neighbourhood group  \
70825              unconfirmed       Robert            Brooklyn
71877              unconfirmed      Cameron            Brooklyn
73083              unconfirmed         Alek              Queens
73137              unconfirmed      Sujatha           Manhattan
73191                 verified      Erin V.              Queens
…                          …            …                    …
102594                verified         Krik            Brooklyn
102595             unconfirmed        Mifan           Manhattan
102596             unconfirmed        Megan            Brooklyn
102597             unconfirmed  Christopher              Queens
102598             unconfirmed      Rebecca           Manhattan

              neighbourhood       lat      long instant_bookable  \
70825           Sunset Park  40.63947 -74.01888            False
71877          Clinton Hill  40.68405 -73.96681            False
73083              Elmhurst  40.74524 -73.88760            False
73137               Midtown  40.76422 -73.98188            False
73191             Sunnyside  40.73272 -73.91826            False
…                       …       …        …                …
102594         Williamsburg  40.70862 -73.94651            False
102595  Morningside Heights  40.80460 -73.96545             True
102596           Park Slope  40.67505 -73.98045             True
102597     Long Island City  40.74989 -73.93777             True
102598      Upper West Side  40.76807 -73.98342            False

       cancellation_policy        room type  …  service fee minimum nights  \
70825              moderate  Entire home/apt  …        $118             2.0
71877              flexible     Private room  …        $229             3.0
```

```
73083             moderate  Entire home/apt  …        $218            60.0
73137               strict     Private room  …        $193             1.0
73191             moderate      Shared room  …        $166             1.0
…                    …              …  …          …            …
102594            flexible     Private room  …        $169             1.0
102595            moderate     Private room  …        $167             1.0
102596            moderate     Private room  …        $198             3.0
102597              strict  Entire home/apt  …        $109             2.0
102598            flexible  Entire home/apt  …        $206             1.0

       number of reviews  last review  reviews per month review rate number  \
70825              220.0    6/27/2019               3.10                 1.0
71877                2.0    9/14/2015               0.04                 5.0
73083                2.0     1/2/2019               0.17                 5.0
73137                0.0          NaN                NaN                 3.0
73191               36.0     7/1/2019               2.29                 4.0
…                    …           …                  …                 …
102594               0.0          NaN                NaN                 3.0
102595               1.0     7/6/2015               0.02                 2.0
102596               0.0          NaN                NaN                 5.0
102597               5.0   10/11/2015               0.10                 3.0
102598               0.0          NaN                NaN                 3.0

        calculated host listings count  availability 365  \
70825                              2.0             287.0
71877                              1.0               0.0
73083                              1.0             132.0
73137                              1.0               0.0
73191                              3.0             140.0
…                                  …                 …
102594                             1.0             227.0
102595                             2.0             395.0
102596                             1.0             342.0
102597                             1.0             386.0
102598                             1.0              69.0

                                              house_rules  license
70825                                                 NaN      NaN
71877                                                 NaN      NaN
73083                                                 NaN      NaN
73137                                                 NaN      NaN
73191                                                 NaN      NaN
…                                                       …      …
102594  No Smoking No Parties or Events of any kind Pl…      NaN
102595  House rules: Guests agree to the following ter…      NaN
102596                                                NaN      NaN
102597                                                NaN      NaN
102598                                                NaN      NaN
```

```
[3436 rows x 22 columns]

Número total de registros antes de eliminar duplicados: 102599
Número total de registros después de eliminar duplicados: 99163
```

1.4 Tarea 3: Transformación de datos (cualquier herramienta)

- Cambia el nombre de la columna `availability 365` a `days_booked`.
- Convierte todos los nombres de columna a minúsculas y sustituye los espacios en los nombres de columna por un guión bajo "_".
- Elimina el signo de dólares y la coma de las columnas `price` y `service_fee`. Si es necesario, convierte estas dos columnas al tipo de datos adecuado.

Si utilizas Python para este ejercicio, incluye el código que hayas utilizado en las celdas siguientes. Si utilizas cualquier otra herramienta, incluye capturas de pantalla de tu trabajo.

[13]:
```
## Sustituir los esapcios en blanco por guiones bajos
df_without_duplicate.columns = df_without_duplicate.columns.str.replace(' ',␣
 ↪'_')

df_without_duplicate.head()
```

[13]:

	NAME	host_identity_verified	host_name	neighbourhood_group	neighbourhood	lat	long	instant_bookable	cancellation_policy	room_type	…	service_fee	minimum_nights	number_of_reviews	last_review	reviews_per_month
0	Clean & quiet apt home by the park	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	False	strict	Private room	…	$193	10.0	9.0	10/19/2021	0.21
1	Skylit Midtown Castle	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	False	moderate	Entire home/apt	…	$28	30.0	45.0	5/21/2022	0.38
2	THE VILLAGE OF HARLEM…NEW YORK !	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	True	flexible	Private room	…	$124	3.0	0.0	NaN	NaN
3	NaN	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	True	moderate	Entire home/apt	…	$74				
4	Entire Apt: Spacious Studio/Loft by central park	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	False	moderate	Entire home/apt	…	$41				

```
3           30.0             270.0     7/5/2019               4.64
4           10.0               9.0   11/19/2018               0.10

  review_rate_number  calculated_host_listings_count  availability_365  \
0                4.0                             6.0             286.0
1                4.0                             2.0             228.0
2                5.0                             1.0             352.0
3                4.0                             1.0             322.0
4                3.0                             1.0             289.0

                                        house_rules  license
0  Clean up and treat the home the way you'd like…      NaN
1  Pet friendly but please confirm with me if the…      NaN
2  I encourage you to use my kitchen, cooking and…      NaN
3                                              NaN      NaN
4  Please no smoking in the house, porch or on th…      NaN

[5 rows x 22 columns]
```

[14]:
```
## Cambia el nombre de la columna.
df_without_duplicate.rename(columns={'availability_365': 'days_booked'},␣
 ↪inplace=True)

df_without_duplicate.head()
```

```
/tmp/ipykernel_7767/2503747552.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_without_duplicate.rename(columns={'availability_365': 'days_booked'},
inplace=True)
```

[14]:
```
                                             NAME host_identity_verified  \
0               Clean & quiet apt home by the park            unconfirmed
1                            Skylit Midtown Castle               verified
2            THE VILLAGE OF HARLEM…NEW YORK !                    NaN
3                                             NaN            unconfirmed
4  Entire Apt: Spacious Studio/Loft by central park               verified

  host_name neighbourhood_group neighbourhood      lat      long  \
0  Madaline            Brooklyn    Kensington  40.64749 -73.97237
1     Jenna           Manhattan       Midtown  40.75362 -73.98377
2     Elise           Manhattan        Harlem  40.80902 -73.94190
3     Garry            Brooklyn  Clinton Hill  40.68514 -73.95976
4    Lyndon           Manhattan   East Harlem  40.79851 -73.94399
```

```
  instant_bookable cancellation_policy        room_type  …  service_fee  \
0            False              strict     Private room  …         $193
1            False            moderate  Entire home/apt  …          $28
2             True            flexible     Private room  …         $124
3             True            moderate  Entire home/apt  …          $74
4            False            moderate  Entire home/apt  …          $41

  minimum_nights number_of_reviews  last_review  reviews_per_month  \
0           10.0               9.0   10/19/2021               0.21
1           30.0              45.0    5/21/2022               0.38
2            3.0               0.0          NaN                NaN
3           30.0             270.0     7/5/2019               4.64
4           10.0               9.0   11/19/2018               0.10

  review_rate_number  calculated_host_listings_count  days_booked  \
0                4.0                             6.0        286.0
1                4.0                             2.0        228.0
2                5.0                             1.0        352.0
3                4.0                             1.0        322.0
4                3.0                             1.0        289.0

                                        house_rules  license
0  Clean up and treat the home the way you'd like…      NaN
1  Pet friendly but please confirm with me if the…      NaN
2  I encourage you to use my kitchen, cooking and…      NaN
3                                              NaN      NaN
4  Please no smoking in the house, porch or on th…      NaN

[5 rows x 22 columns]
```

[15]:
```
## Convierte todos los nombres de columna a minúsculas y sustituye los espacios␣
 ↪por un guión bajo "_".
## NOTA: La conversion de los esapcios fue realizada en un paso anterior.
df_without_duplicate.columns = df_without_duplicate.columns.str.lower()
df_without_duplicate.head()
```

[15]:
```
                                          name host_identity_verified  \
0               Clean & quiet apt home by the park           unconfirmed
1                           Skylit Midtown Castle              verified
2              THE VILLAGE OF HARLEM…NEW YORK !                   NaN
3                                           NaN           unconfirmed
4  Entire Apt: Spacious Studio/Loft by central park              verified

  host_name neighbourhood_group neighbourhood       lat      long  \
0  Madaline            Brooklyn    Kensington  40.64749 -73.97237
1     Jenna           Manhattan       Midtown  40.75362 -73.98377
2     Elise           Manhattan        Harlem  40.80902 -73.94190
```

```
3     Garry            Brooklyn  Clinton Hill  40.68514 -73.95976
4    Lyndon           Manhattan   East Harlem  40.79851 -73.94399

  instant_bookable cancellation_policy        room_type  …  service_fee  \
0            False              strict     Private room  …         $193
1            False            moderate  Entire home/apt  …          $28
2             True            flexible     Private room  …         $124
3             True            moderate  Entire home/apt  …          $74
4            False            moderate  Entire home/apt  …          $41

  minimum_nights number_of_reviews  last_review  reviews_per_month  \
0           10.0               9.0   10/19/2021               0.21
1           30.0              45.0    5/21/2022               0.38
2            3.0               0.0          NaN                NaN
3           30.0             270.0    7/5/2019               4.64
4           10.0               9.0   11/19/2018               0.10

  review_rate_number  calculated_host_listings_count  days_booked  \
0                4.0                             6.0        286.0
1                4.0                             2.0        228.0
2                5.0                             1.0        352.0
3                4.0                             1.0        322.0
4                3.0                             1.0        289.0

                                        house_rules  license
0  Clean up and treat the home the way you'd like…      NaN
1  Pet friendly but please confirm with me if the…      NaN
2  I encourage you to use my kitchen, cooking and…      NaN
3                                              NaN      NaN
4  Please no smoking in the house, porch or on th…      NaN

[5 rows x 22 columns]
```

[16]:
```
## Elimina el signo de dólares y la coma de las columnas. Si es necesario,␣
 ↪convierte estas dos columnas al tipo de datos adecuado.

money_columns = ['price', 'service_fee']

for money_column in money_columns:
    df_without_duplicate[money_column] = df_without_duplicate[money_column].
 ↪replace('[\$,]', '', regex=True).astype(float)

df_without_duplicate.head()
```

```
/tmp/ipykernel_7767/453405683.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_without_duplicate[money_column] =
df_without_duplicate[money_column].replace('[\$,]', '',
regex=True).astype(float)
/tmp/ipykernel_7767/453405683.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_without_duplicate[money_column] =
df_without_duplicate[money_column].replace('[\$,]', '',
regex=True).astype(float)
```

[16]:

	name	host_identity_verified	host_name	neighbourhood_group	neighbourhood	lat	long	instant_bookable	cancellation_policy	room_type	…	service_fee	minimum_nights	number_of_reviews	last_review	reviews_per_month	review_rate_number	calculated_host_listings_count	days_booked
0	Clean & quiet apt home by the park	unconfirmed	Madaline	Brooklyn	Kensington	40.64749	-73.97237	False	strict	Private room	…	193.0	10.0	9.0	10/19/2021	0.21	4.0	6.0	286.0
1	Skylit Midtown Castle	verified	Jenna	Manhattan	Midtown	40.75362	-73.98377	False	moderate	Entire home/apt	…	28.0	30.0	45.0	5/21/2022	0.38	4.0	2.0	228.0
2	THE VILLAGE OF HARLEM…NEW YORK !	NaN	Elise	Manhattan	Harlem	40.80902	-73.94190	True	flexible	Private room	…	124.0	3.0	0.0	NaN	NaN			
3	NaN	unconfirmed	Garry	Brooklyn	Clinton Hill	40.68514	-73.95976	True	moderate	Entire home/apt	…	74.0	30.0	270.0	7/5/2019	4.64			
4	Entire Apt: Spacious Studio/Loft by central park	verified	Lyndon	Manhattan	East Harlem	40.79851	-73.94399	False	moderate	Entire home/apt	…	41.0	10.0	9.0	11/19/2018	0.10			

```
2                5.0                                1.0        352.0
3                4.0                                1.0        322.0
4                3.0                                1.0        289.0

                                           house_rules  license
0  Clean up and treat the home the way you'd like…       NaN
1  Pet friendly but please confirm with me if the…       NaN
2  I encourage you to use my kitchen, cooking and…       NaN
3                                                NaN       NaN
4  Please no smoking in the house, porch or on th…       NaN

[5 rows x 22 columns]
```

1.4.1 Tarea 4: Análisis exploratorio de datos (cualquier herramienta)

- Enumera los tipos de habitaciones disponibles en el dataset.
- ¿Qué tipo de habitación tiene la política de cancelación más estricta?
- Enumera el precio medio por barrio y señala cuál es el conjunto de barrios más caro para alquilar.

Si utilizas Python para este ejercicio, incluye el código que hayas utilizado en las celdas siguientes. Si utilizas cualquier otra herramienta, incluye capturas de pantalla de tu trabajo.

[17]:
```
df_without_duplicate.columns
```

[17]:
```
Index(['name', 'host_identity_verified', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'lat', 'long', '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', 'days_booked', 'house_rules',
       'license'],
      dtype='object')
```

[18]:
```
## Enumera los tipos de habitaciones disponibles en Airbnb.
room_types = df_without_duplicate['room_type'].unique()

print("Tipos de habitaciones en el dataset: ")
for type in room_types:
    print(type)
```

```
Tipos de habitaciones en el dataset:
Private room
Entire home/apt
Shared room
Hotel room
```

[19]:
```
## ¿Qué tipo de habitación se adhiere a una política de cancelación más␣
 ↪estricta?
cancellation_policy_by_type = df_without_duplicate.
 ↪groupby('room_type')['cancellation_policy'].value_counts()

strict_cancellation_policy = cancellation_policy_by_type.groupby('room_type').
 ↪idxmax().apply(lambda x: x[1])

print("Politicas de cancelacion segun el tipo de habitacion: ")
print(strict_cancellation_policy)
```

```
Politicas de cancelacion segun el tipo de habitacion:
room_type
Entire home/apt    flexible
Hotel room         flexible
Private room       moderate
Shared room          strict
Name: count, dtype: object
```

[20]:
```
## Enumera los precios por barrio y menciona también cuál es el grupo de␣
 ↪barrios con alquileres más caros.

df_without_duplicate['price'] = pd.to_numeric(df_without_duplicate['price'].
 ↪replace('[\$,]', '', regex=True))

price_by_hood = df_without_duplicate.groupby('neighbourhood')['price'].mean()

print("Precios por barrio:")
print(price_by_hood)

most_expensive_hood = price_by_hood.nlargest(5)
print("\nBarrios con alquileres mas caros: ")
print(most_expensive_hood)
```

```
Precios por barrio:
neighbourhood
Allerton           633.923913
Arden Heights      804.888889
Arrochar           612.734694
Arverne            649.052632
Astoria            638.993385
                        …
Windsor Terrace    578.810127
Woodhaven          619.227027
Woodlawn           588.370370
Woodrow            709.333333
Woodside           638.454867
```

```
Name: price, Length: 224, dtype: float64

Barrios con alquileres mas caros:
neighbourhood
New Dorp                   1045.333333
Chelsea, Staten Island     1042.000000
Fort Wadsworth             1024.000000
Little Neck                 817.750000
Jamaica Hills               812.904762
Name: price, dtype: float64
```

```
/tmp/ipykernel_7767/2270480487.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_without_duplicate['price'] =
pd.to_numeric(df_without_duplicate['price'].replace('[\$,]', '', regex=True))
```

1.5 Tarea 5a: Visualización de datos (Cualquier herramienta)

- Enumerar los distintos tipos de habitaciones disponibles en Airnb
- Qué tipo de habitación se adhiere a una política de cancelación más estricta.
- Enumere los precios por grupo de barrios y mencione también cuál es el grupo de barrios más caro para los alquileres.
- Enumere los 10 barrios más caros por orden creciente de precio con la ayuda de un gráfico de barras horizontales. ¿Cuál es el barrio más barato?
- Enumere los barrios que ofrecen alquileres a corto plazo de menos de 10 días. Ilustrar con un gráfico de barras
- Enumere los precios con respecto al tipo de habitación utilizando un gráfico de barras y exponga también sus inferencias.
- Cree un gráfico circular que muestre la distribución de los días reservados para cada grupo de barrios.

Si utiliza Python para este ejercicio, por favor incluya el código en las celdas de abajo. Si utiliza cualquier otra herramienta, por favor incluya pantallazos de su trabajo.

[21]:
```python
## Enumera los tipos de habitaciones disponibles en Airbnb.
room_types = df_without_duplicate['room_type'].unique()

print("Tipos de habitaciones en el dataset: ")
for type in room_types:
    print(type)
```

```
Tipos de habitaciones en el dataset:
Private room
Entire home/apt
Shared room
```

```
Hotel room
```

[22]:
```
## ¿Qué tipo de habitación se adhiere a una política de cancelación más estricta?
cancellation_policy_by_type = df_without_duplicate.groupby('room_type')['cancellation_policy'].value_counts()

strict_cancellation_policy = cancellation_policy_by_type.groupby('room_type').idxmax().apply(lambda x: x[1])

print("Politicas de cancelacion segun el tipo de habitacion: ")
print(strict_cancellation_policy)
```

```
Politicas de cancelacion segun el tipo de habitacion:
room_type
Entire home/apt    flexible
Hotel room         flexible
Private room       moderate
Shared room          strict
Name: count, dtype: object
```

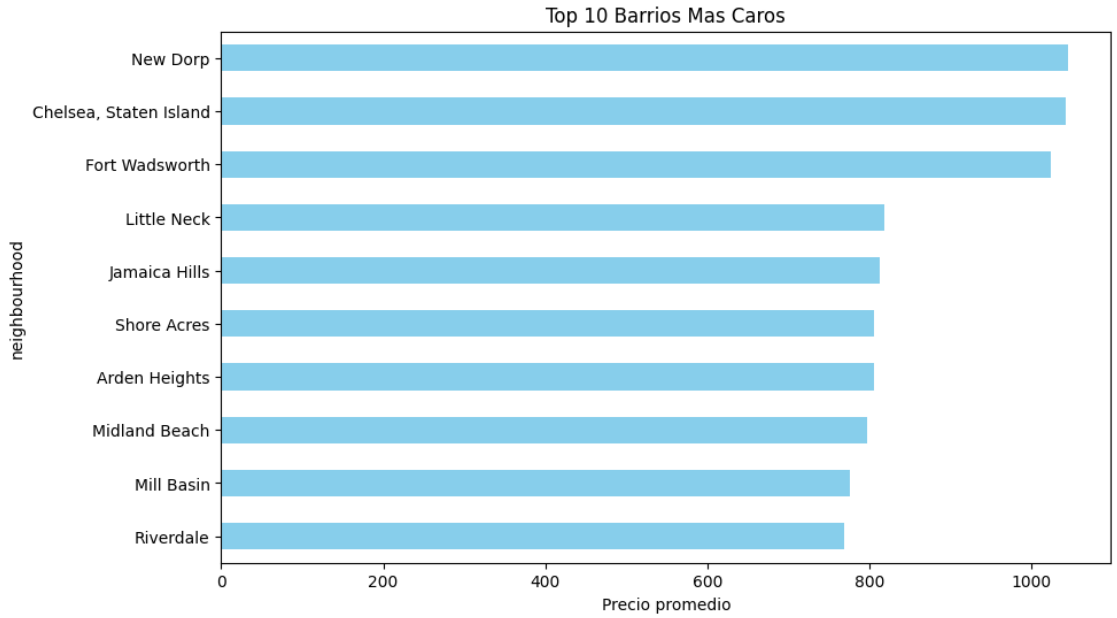
[23]:
```
## Enumera los precios por barrio y menciona también cuál es el grupo de barrios con alquileres más caros.

df_without_duplicate['price'] = pd.to_numeric(df_without_duplicate['price'].replace('[\$,]', '', regex=True))

price_by_hood = df_without_duplicate.groupby('neighbourhood')['price'].mean()

print("Precios por barrio:")
print(price_by_hood)

most_expensive_hood = price_by_hood.nlargest(5)
print("\nBarrios con alquileres mas caros: ")
print(most_expensive_hood)
```

```
Precios por barrio:
neighbourhood
Allerton           633.923913
Arden Heights      804.888889
Arrochar           612.734694
Arverne            649.052632
Astoria            638.993385
                         …
Windsor Terrace    578.810127
Woodhaven          619.227027
Woodlawn           588.370370
Woodrow            709.333333
```

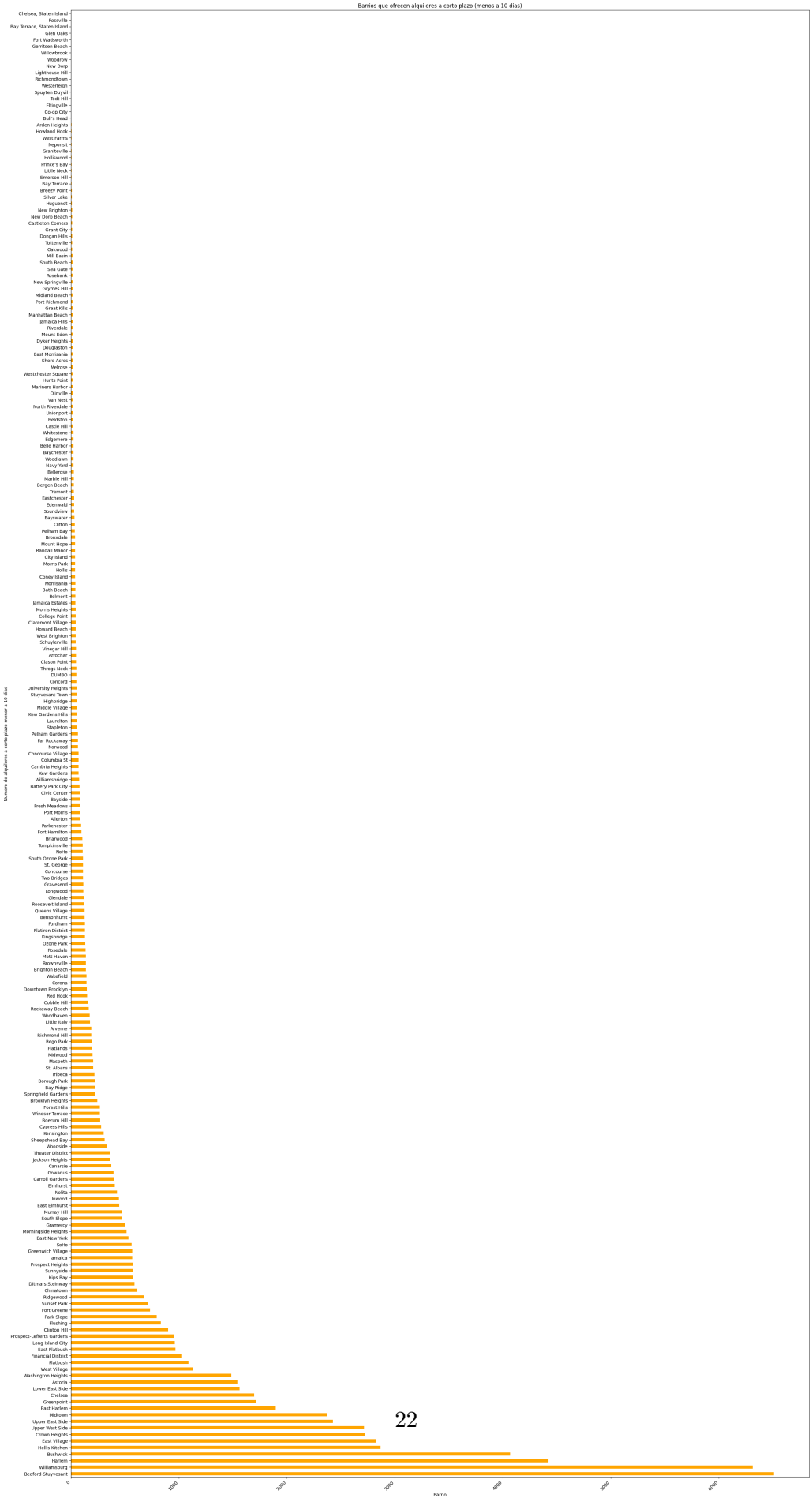
```
Woodside            638.454867
Name: price, Length: 224, dtype: float64

Barrios con alquileres mas caros:
neighbourhood
New Dorp                   1045.333333
Chelsea, Staten Island     1042.000000
Fort Wadsworth             1024.000000
Little Neck                 817.750000
Jamaica Hills               812.904762
Name: price, dtype: float64
```

```
/tmp/ipykernel_7767/2270480487.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_without_duplicate['price'] =
pd.to_numeric(df_without_duplicate['price'].replace('[\$,]', '', regex=True))
```

[24]:
```python
import matplotlib.pyplot as plt

## Agrupar por barrio y calcular el precio medio
price_by_hood = df_without_duplicate.groupby('neighbourhood')['price'].mean()

## Seleccionar los 10 barrios mas caros
most_expensive_hood = price_by_hood.nlargest(10).sort_values()

##Crear un grafico de barras horizontales
plt.figure(figsize=(10, 6))
most_expensive_hood.plot(kind='barh', color='skyblue')
plt.xlabel('Precio promedio')
plt.title('Top 10 Barrios Mas Caros')
plt.show()
```



[25]:
```
## Enumerar los barrios que ofrecen alquileres a corto plazo de menos de 10
↪días.
short_time_rent = df_without_duplicate[df_without_duplicate['minimum_nights'] <
↪10]

## Contar la frecuencia de cada barrio en el nuevo DF
short_time_rent_hoods = short_time_rent['neighbourhood'].value_counts()

## Grafico de barras
plt.figure(figsize=(30, 60))
short_time_rent_hoods.plot(kind='barh', color='orange')
plt.xlabel('Barrio')
plt.ylabel('Numero de alquileres a corto plazo menor a 10 dias')
plt.title('Barrios que ofrecen alquileres a corto plazo (menos a 10 dias)')
plt.xticks(rotation=45, ha='right')
plt.show()
```


Barrios que ofrecen alquileres a corto plazo (menos a 10 dias)
Numero de alquileres a corto plazo menos a 10 dias
Barrio
Chelsea, Staten Island
Rossville
Bay Terrace, Staten Island
Glen Oaks
Fort Wadsworth
Gerritsen Beach
Willowbrook
Woodrow
New Dorp
Lighthouse Hill
Richmondtown
Westerleigh
Spuyten Duyvil
Todt Hill
Eltingville
Co-op City
Bull's Head
Arden Heights
Howland Hook
West Farms
Neponsit
Graniteville
Hollswood
Prince's Bay
Little Neck
Emerson Hill
Bay Terrace
Breezy Point
Silver Lake
Huguenot
New Brighton
New Dorp Beach
Castleton Corners
Grant City
Dongan Hills
Tottenville
Oakwood
Mill Basin
South Beach
Sea Gate
Rosebank
New Springville
Grymes Hill
Midland Beach
Port Richmond
Great Kills
Manhattan Beach
Jamaica Hills
Riverdale
Mount Eden
Dyker Heights
Douglaston
East Morrisania
Shore Acres
Melrose
Westchester Square
Hunts Point
Mariners Harbor
Olinville
Van Nest
North Riverdale
Unionport
Fieldston
Castle Hill
Whitestone
Edgemere
Belle Harbor
Baychester
Woodlawn
Navy Yard
Bellerose
Marble Hill
Bergen Beach
Tremont
Eastchester
Edenwald
Soundview
Bayswater
Clifton
Pelham Bay
Bronxdale
Mount Hope
Randall Manor
City Island
Morris Park
Hollis
Coney Island
Morrisania
Bath Beach
Belmont
Jamaica Estates
Morris Heights
College Point
Claremont Village
Howard Beach
West Brighton
Schuylerville
Vinegar Hill
Arrochar
Clason Point
Throgs Neck
DUMBO
Concord
University Heights
Stuyvesant Town
Highbridge
Middle Village
Kew Gardens Hills
Laurelton
Stapleton
Pelham Gardens
Far Rockaway
Norwood
Concourse Village
Columbia St
Cambria Heights
Kew Gardens
Williamsbridge
Battery Park City
Civic Center
Bayside
Fresh Meadows
Port Morris
Allerton
Parkchester
Fort Hamilton
Briarwood
Tompkinsville
NoHo
South Ozone Park
St. George
Concourse
Two Bridges
Gravesend
Longwood
Glendale
Roosevelt Island
Queens Village
Bensonhurst
Fordham
Fiairon District
Kingsbridge
Ozone Park
Rosedale
Mott Haven
Brownsville
Brighton Beach
Wakefield
Corona
Downtown Brooklyn
Red Hook
Cobble Hill
Rockaway Beach
Woodhaven
Little Italy
Arverne
Richmond Hill
Rego Park
Flatlands
Midwood
Maspeth
St. Albans
Tribeca
Borough Park
Bay Ridge
Springfield Gardens
Brooklyn Heights
Forest Hills
Windsor Terrace
Boerum Hill
Cypress Hills
Kensington
Sheepshead Bay
Woodside
Theater District
Jackson Heights
Canarsie
Gowanus
Carroll Gardens
Elmhurst
Nolita
Inwood
East Elmhurst
Murray Hill
South Slope
Gramercy
Morningside Heights
East New York
SoHo
Greenwich Village
Jamaica
Prospect Heights
Sunnyside
Kips Bay
Ditmars Steinway
Chinatown
Ridgewood
Sunset Park
Fort Greene
Park Slope
Flushing
Clinton Hill
Prospect-Lefferts Gardens
Long Island City
East Flatbush
Financial District
Flatbush
West Village
Washington Heights
Astoria
Lower East Side
Chelsea
Greenpoint
East Harlem
Midtown
Upper East Side
Upper West Side
Crown Heights
East Village
Hell's Kitchen
Bushwick
Harlem
Williamsburg
Bedford-Stuyvesant
0
1000
2000
3000
4000
5000
6000

[33]:
```
## En virtud de poder mejorar la visualizacion
import plotly.express as px

fig = px.bar(short_time_rent_hoods,
             orientation ='h',
             labels={'value': 'Número de Alquileres a Corto Plazo (< 10 días)',␣
 ↪'index': 'Barrio'},
             title='Barrios que ofrecen alquileres a corto plazo (< 10 días)')

fig.update_layout(width=2120, height=600, margin=dict(l=0, r=0, b=0, t=30 ))

fig.show()
```

1.6 Tarea 5b: Visualización de datos (Cualquier herramienta)

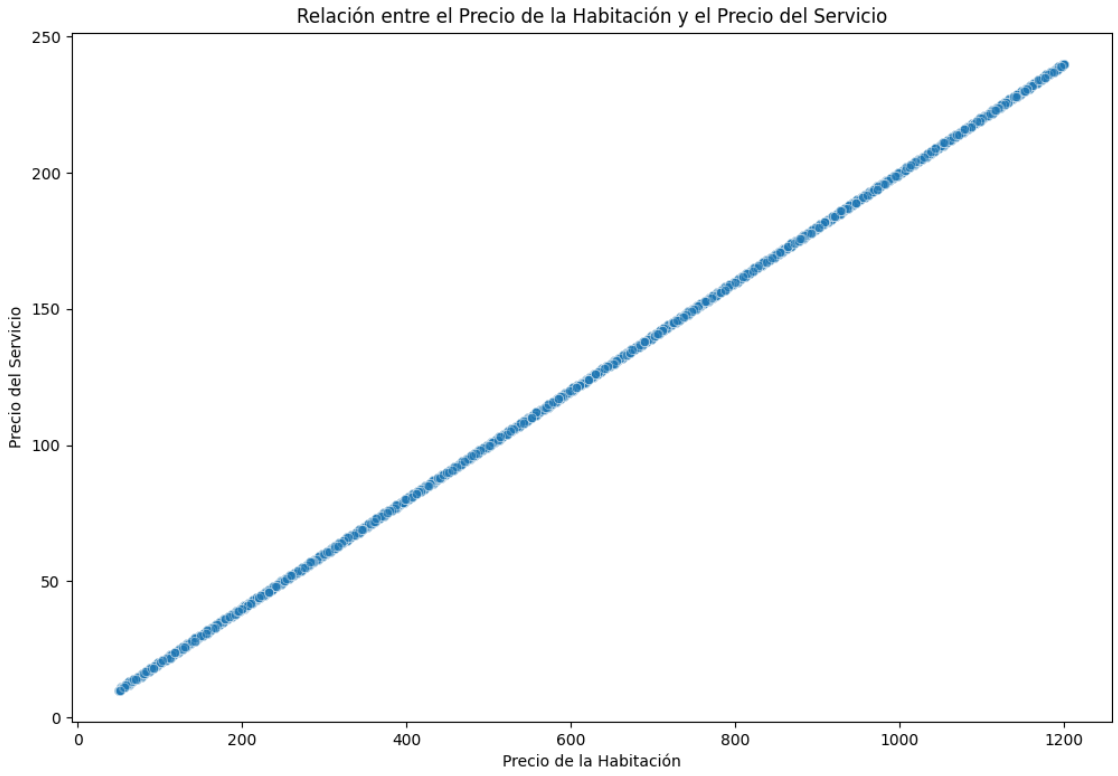
- ¿El precio del servicio y el precio de la habitación tienen un impacto mutuo? Ilustre esta relación con un gráfico de dispersión e indique sus inferencias
- Utilizando un gráfico lineal muestre en qué año tuvo lugar la máxima construcción de habitaciones.

Si utiliza Python para este ejercicio, incluya el código en las celdas siguientes. Si utiliza cualquier otra herramienta, incluya capturas de pantalla de su trabajo.

[34]:
```
import seaborn as sns
## Convertir las columnas de precios a tipo de dato numerico
df_without_duplicate['price'] = pd.to_numeric(df_without_duplicate['price'].
 ↪replace('[\$,]', '', regex=True))
df_without_duplicate['service_fee'] = pd.
 ↪to_numeric(df_without_duplicate['service_fee'].replace('[\$,]', '',␣
 ↪regex=True))

## Crear grafico de dispersion
plt.figure(figsize=(12, 8))
sns.scatterplot(x='price', y='service_fee', data=df_without_duplicate, alpha= 0.
 ↪5)
plt.xlabel('Precio de la Habitación')
plt.ylabel('Precio del Servicio')
plt.title('Relación entre el Precio de la Habitación y el Precio del Servicio')
plt.show()
```

```
/tmp/ipykernel_7767/442962489.py:3: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

/tmp/ipykernel_7767/442962489.py:4: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

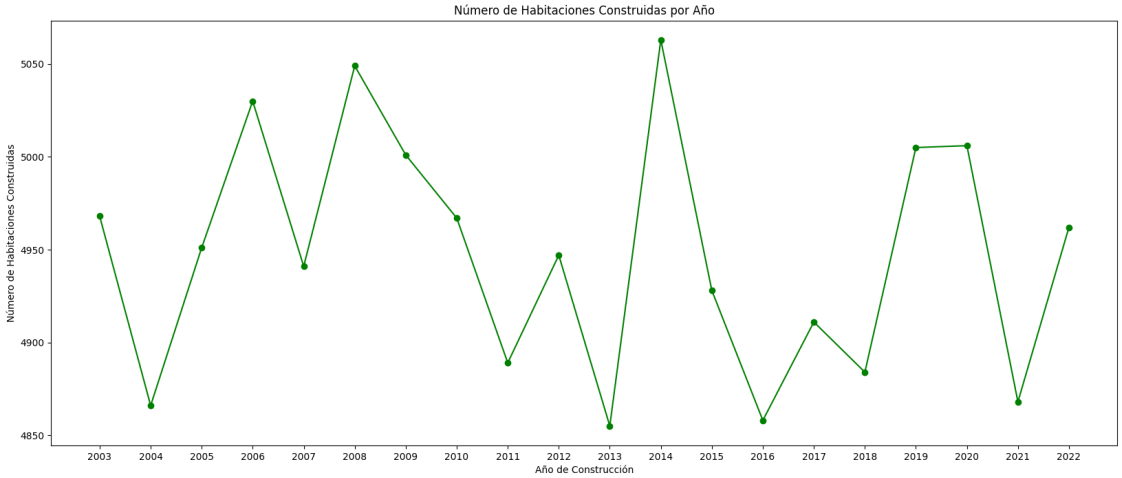


[ ]:
```
"""Como puede oibservarse, claramente SÍ existe una correlación entre el precio␣
 ↪de la habitación y el precio del servicio
"""
```

[47]:
```
import matplotlib.ticker as ticker

# En qué año tuvo lugar la máxima construcción de habitaciones
```

```
df_without_duplicate['construction_year'] =␣
 ↪df_without_duplicate['construction_year'].astype('Int64')
rooms_per_year = df_without_duplicate.groupby('construction_year').size()

# Crear un gráfico lineal
plt.figure(figsize=(20, 8))
rooms_per_year.plot(kind='line', marker='o', color='green')

# Establecer las marcas del eje x con todos los años únicos
plt.xticks([int(year) for year in rooms_per_year.index])

plt.xlabel('Año de Construcción')
plt.ylabel('Número de Habitaciones Construidas')
plt.title('Número de Habitaciones Construidas por Año')
plt.show()
```

```
/tmp/ipykernel_7767/3487650443.py:4: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

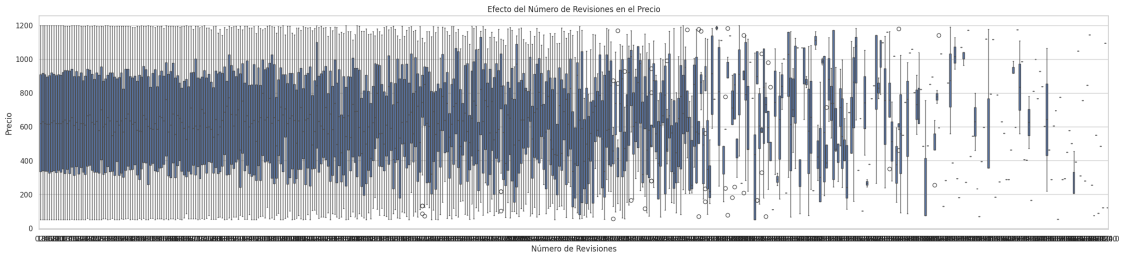


1.7 Tarea 5c: Visualización de datos (Cualquier herramienta)

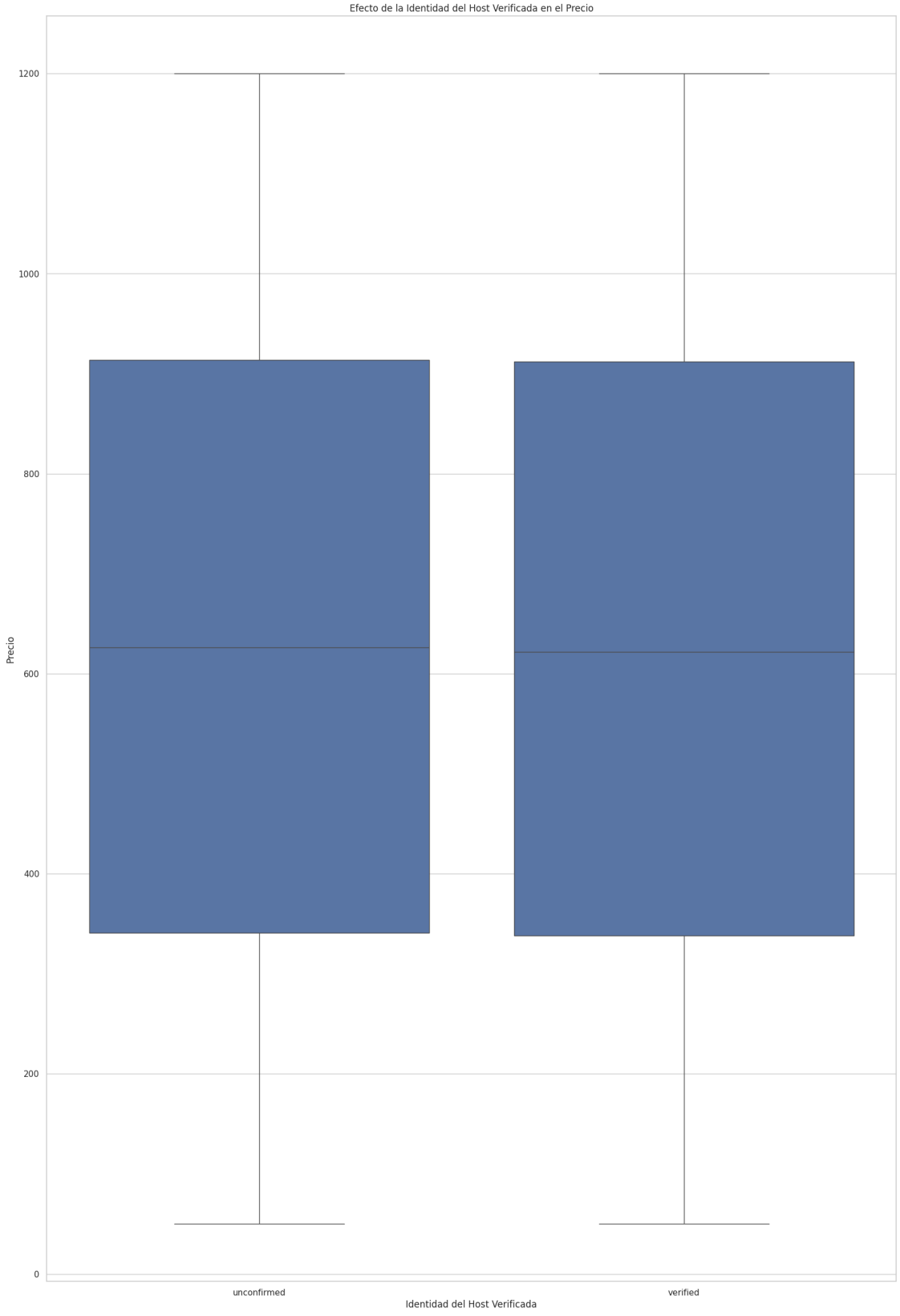
- Con la ayuda de gráficos de caja ilustra lo siguiente
- Efecto del número de tasa de revisión en el precio
- Efecto de la identidad del host verificada en el precio

Si utiliza Python para este ejercicio, por favor incluya el código en las celdas de abajo. Si utiliza cualquier otra herramienta, por favor incluya capturas de pantalla de su trabajo.

[51]:
```
import seaborn as sns
import matplotlib.pyplot as plt

## COnfiguracion del estilo
sns.set(style="whitegrid")

## Grafico de caja para el efecto del numero de tasas de revision en el precio
plt.figure(figsize=(30, 6))
sns.boxplot(x='number_of_reviews', y='price', data=df_without_duplicate)
plt.xlabel('Número de Revisiones')
plt.ylabel('Precio')
plt.title('Efecto del Número de Revisiones en el Precio')
plt.show()
```



[53]:
```
# Gráfico de caja para el efecto de la identidad del host verificada en el␣
 ↪precio
plt.figure(figsize=(20, 30))
sns.boxplot(x='host_identity_verified', y='price', data=df_without_duplicate)
plt.xlabel('Identidad del Host Verificada')
plt.ylabel('Precio')
plt.title('Efecto de la Identidad del Host Verificada en el Precio')
plt.show()
```


Efecto de la Identidad del Host Verificada en el Precio
Precio
1200
1000
800
600
400
200
0
unconfirmed
verified
Identidad del Host Verificada

[ ]: