airbnb

January 12, 2023

1 Portfolio Project: Analyze Airbnb Data with Python

1.1 Overview

In this project, you will step into the shoes of an entry-level data analyst at Airbnb, helping interpret real-world data to help make a key business decision.

1.2 Case Study

The team at Airbnb is trying to increase their profits from their rentals across the US. To do this, they want to explore what factors encourage renters to pay more for a particular listing. Is it the location? Walkability? The property's ratings?

They want you to provide insights and recommendations by analyzing a dataset containing information on current rental prices, rental locations, and a slew of other details. The team will use your analysis in the future to provide property owners with a suggested price to charge renters. This feature will help hosts (and Airbnb) maximize their profits from each listing.

1.3 Business Objectives

1. Explore the prices of current Airbnb listings
2. Determine important factors that may influence the price of listings
3. Provide analytic insights and data-driven recommendations

1.4 Your Task

Your task will be to conduct an exploratory data analysis to investigate if there are any patterns or themes that may influence the pricing of rentals on Airbnb. To do this, you will load, clean, process, analyze, and visualize data. You will also pose questions, and seek to answer them meaningfully using the dataset provided.

In this project, we'll use data from Airbnb's New York City dataset (listings.csv) however, to keep this project unique and open-ended please feel free to choose whichever major city and datasets you'd prefer - which can be found from Inside Airbnb data.

1.5 Getting Task-by-Task Guidance

If you'd like a little more support while completing this project, explore this step-by-step resource to get additional hints and resources to help you along each task of this project.

If you don't need much support and feel comfortable enough to start building right away, you can take a look at the general steps below and jump right into your project.

1.6 Task 0: Import Python Modules

First, import the primary modules that will be used in this project. Remember as this is an open-ended project, feel free to make use of any of your favorite libraries that you feel may be useful for this data analysis project. Some common examples may include: - pandas - numpy - maplotlib - seaborn

[2]:
```
# Import required packages

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

1.7 Task 1: Load the Data

The `listings.csv` dataset contains Airbnb listing activity and metrics within New York City. This dataset contains:

39881 rows - each row is a different Airbnb listing **18 columns**

Column name	Description
id	Listing id
name	Name of listing
host_id	Host id
host_name	Name of host
neighbourhood_group	Neighbourhood group the listing is in
neighbourhood	Neighbourhood the listing is in
latitude	Latitude coordinate of listing location
longitude	Longitude coordinate of listing location
room_type	Room type of the listing
price	Price of the listing
minimum_nights	Minimum number of nights stay for listing
number_of_reviews	Number of reviews for listing
last_review	Date of the latest review
reviews_per_month	Number of reviews per month of listing
calculated_host_listings_count	Number of listings the host has
availability_365	The availability of the listing in the next 365 days
number_of_reviews_ltm	Number of reviews of listing in last 12 months

Column name	Description
license	If host is licensed

Load the dataset `listings.csv` into a dataframe `listings` and display the first five rows.

[3]:
```
# Load the data for airbnbs in Hawaii
listings =pd.read_csv('Hawaii_Data/listings.csv')
listings.head()
```

[3]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365	number_of_reviews_ltm	license
0	5269	Upcountry Hospitality in the 'Auwai Suite	7620	Lea & Pat	Hawaii	South Kohala	20.02740	-155.70200	Entire home/apt	149	5	24	2022-07-13	0.17	3	212	10	119-269-5808-01R
1	34843927	Simply Paradise Glamping	262664392	Adriano And Julia	Hawaii	North Kona	19.66220	-155.95681	Private room	83	1	194	2022-08-25	5.90	3	334	96	NaN
2	35066424	Spacious 3 Bedroom 3 Bath + Loft at Alii Cove	264152810	Robyn	Hawaii	North Kona	19.62768	-155.98543	Entire home/apt	175	30	2	2022-02-19	0.25	1	197	2	NaN
3	35067513	Noah's Hideaway Maui, Luxury B&B, Walk to Beach!	264162605	Fadi	Maui	Lahaina	20.91764	-156.68840	Entire home/apt	622	1	70	2022-06-24	1.80	2	191	5	440090330000, TA-197-216-9216-01
4	5387	Hale Koa Studio & 1 Bedroom Units!!	7878	Edward	Hawaii	South Kona	19.43081	-155.88069	Entire home/apt	91	5	201	2022-09-03	1.31	3	166	19	NaN

1.8 Task 2: Explore the Data

In this task, let's explore the Airbnb listing data further.

[4]:
```
# Check dtypes for each column and missing values
listings.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28580 entries, 0 to 28579
Data columns (total 18 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              28580 non-null  int64
 1   name                            28580 non-null  object
 2   host_id                         28580 non-null  int64
 3   host_name                       28446 non-null  object
 4   neighbourhood_group             28580 non-null  object
 5   neighbourhood                   28580 non-null  object
 6   latitude                        28580 non-null  float64
 7   longitude                       28580 non-null  float64
 8   room_type                       28580 non-null  object
 9   price                           28580 non-null  int64
 10  minimum_nights                  28580 non-null  int64
 11  number_of_reviews               28580 non-null  int64
 12  last_review                     23041 non-null  object
 13  reviews_per_month               23041 non-null  float64
 14  calculated_host_listings_count  28580 non-null  int64
 15  availability_365                28580 non-null  int64
 16  number_of_reviews_ltm           28580 non-null  int64
 17  license                         23875 non-null  object
dtypes: float64(3), int64(8), object(7)
memory usage: 3.9+ MB
```

[6]:
```
# Get the size of the data
listings.shape
```

[6]:
```
(28580, 18)
```

Insights - 18 columns , 28580 entries
- The columns seem to match their datatypes. - Columns 'host_name', 'last_review', 'reviews_per_month' and 'license' have a few missing values.

[23]:
```
# Get a summary of the data
listings.describe()
```

[23]:
```
                 id       host_id      latitude     longitude         price  \
count  2.858000e+04  2.858000e+04  28580.000000  28580.000000  28580.000000
mean   1.298151e+17  1.310257e+08     20.903315   -157.184148    476.107558
```

```
std    2.569452e+17  1.329351e+08      0.796710      1.254248   1304.712077
min    5.269000e+03  8.840000e+02     18.920250   -159.714620      0.000000
25%    2.459821e+07  2.797933e+07     20.697357   -157.837951    167.000000
50%    4.444412e+07  9.021966e+07     20.963640   -156.690550    266.000000
75%    5.336420e+07  1.891050e+08     21.288513   -156.437914    432.000000
max    7.144702e+17  4.788538e+08     22.229178   -154.822930  26774.000000

       minimum_nights  number_of_reviews  reviews_per_month  \
count    28580.000000       28580.000000       23041.000000
mean         8.595311          32.411127           1.271445
std         25.744771          59.443697           1.391117
min          1.000000           0.000000           0.010000
25%          1.000000           1.000000           0.310000
50%          3.000000           9.000000           0.810000
75%          5.000000          36.000000           1.810000
max       1000.000000        1025.000000          44.350000

       calculated_host_listings_count  availability_365  number_of_reviews_ltm
count                    28580.000000      28580.000000           28580.000000
mean                        79.051015        185.122778              11.199895
std                        133.673483        112.795197              16.999292
min                          1.000000          0.000000               0.000000
25%                          2.000000         89.000000               0.000000
50%                         17.000000        194.000000               4.000000
75%                         91.000000        279.000000              15.000000
max                        660.000000        365.000000             306.000000
```

Insights - The maximun airbnb price in Hawaii is $26774 - There seems to be an outlier in the 'minimum_nights' column. It doesn't seen possible for someone to stay a minimum of 1000 nights at an airbnb. - 'id' and 'host_id' can be removed because there wouldn't have a correlation to 'price'.

1.9 Task 3: Clean and Validate the Data

When loading and previewing your data you might have come across a few NaN or null values in your data. In this task, consider a few methods of finding and dealing with null or missing values. Feel free to explore any other data cleaning methods you feel may be useful.

[24]:
```
# Find duplicates
listings.duplicated().sum()
```

[24]: 0

[25]:
```
# Find missing values in each column
listings.isna().sum()
```

[25]:
```
id                                  0
name                                0
host_id                             0
host_name                         134
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
last_review                      5539
reviews_per_month                5539
calculated_host_listings_count      0
availability_365                    0
number_of_reviews_ltm               0
license                          4705
dtype: int64
```

There is no duplicates in this dataset but there are a few missing values.

[26]:
```
# Check for inconsistencies in categorical values
obj_cols = listings.dtypes[listings.dtypes=='object'].index

# create a loop to examine the object type columns
for col in obj_cols:
  print(f'Column: {col}')
  print(listings[col].value_counts(dropna=False))
  print('\n')
```

```
Column: name
Ka Eo Kai Studio                                    34
Bali Hai Villas 1 Bedroom                           24
Wyndham at Waikiki Beach Walk® - 2 Bedroom Deluxe   22
Ka'anapali Beach Club- 1 bedroom scenic view        17
Mauna Loa Village 1 Bedroom                         16
                                                    ..
Ocean-View Maui Penthouse w/ Balcony & Pool Access   1
Majestic Volcano Villas ®                            1
The Aloha Kona House (2bed/2bath, rooftop deck)      1
Napili Bay Condo - 1 Bdrm                            1
On the beach in Maui, condo B-304 at Maui Sunset     1
Name: name, Length: 27535, dtype: int64


Column: host_name
Vacasa Hawaii          697
```

```
RoomPicks              696
Crystal                473
Maui Condo             416
Maui Resort Rentals    351
                       …
John & Maggie            1
Ira                      1
Wendyliza                1
Kd                       1
Ujwol                    1
Name: host_name, Length: 3705, dtype: int64


Column: neighbourhood_group
Maui        9142
Honolulu    8652
Hawaii      6575
Kauai       4211
Name: neighbourhood_group, dtype: int64


Column: neighbourhood
Primary Urban Center     6221
Lahaina                  4421
Kihei-Makena             4064
North Kona               3011
North Shore Kauai        1817
Koloa-Poipu              1389
South Kohala             1249
Puna                     1113
Kapaa-Wailua              697
Koolauloa                 568
Ewa                       516
South Hilo                513
Koolaupoko                455
North Shore Oahu          420
Lihue                     284
South Kona                267
Waianae                   241
Paia-Haiku                222
Kau                       220
East Honolulu             190
Molokai                   169
Wailuku-Kahului           139
North Kohala               93
Hana                       75
Hamakua                    66
Makawao-Pukalani-Kula      45
```

```
North Hilo                 43
Central Oahu               41
Waimea-Kekaha              24
Lanai                       7
Name: neighbourhood, dtype: int64


Column: room_type
Entire home/apt    25147
Private room        3324
Hotel room            73
Shared room           36
Name: room_type, dtype: int64


Column: last_review
NaN           5539
2022-09-05     654
2022-08-28     586
2022-08-29     575
2022-09-06     521
              …
2018-11-10       1
2020-10-30       1
2017-01-04       1
2016-08-28       1
2019-07-26       1
Name: last_review, Length: 1105, dtype: int64


Column: license
NaN                                                                  4705
Exempt                                                               2269
540050360000                                                          118
050555494401, 050555494401, TA-050-555-4944-01                         62
540050050000                                                           57
                                                                     …
390080110083, TA-083-880-1920-01                                        1
390060040045, TA-032-900-7104-01                                        1
210080640049, TA-168-659-5584-01                                        1
TA-021-737-8816-01                                                      1
Property Permit ID: 690070350125 Property Tax ID: TA-128-958-3104-01    1
Name: license, Length: 18120, dtype: int64
```

- 'license' & 'last_review' columns can be removed since majority of the entries are NaN and won't be helpful for our analysis.

[27]:
```
# Remove unnecessary columns
listings.drop(columns= [ 'license', 'last_review'], inplace=True)
```

[28]:
```
# Find outliers and remove them
num_cols = listings.dtypes[listings.dtypes=='int64'].index

# Create a loop to examine the numerical columns
for col in num_cols:
  print(f'Column: {col}')
  print(listings[col].value_counts(dropna=False))
  print('\n')
```

```
Column: id
697119635946801844    1
53626690              1
30963492              1
11632421              1
51582760              1
                     ..
698395308513479757    1
25265093              1
41718727              1
50330570              1
198657                1
Name: id, Length: 28580, dtype: int64


Column: host_id
5615582      660
132087088    416
39073224     351
111808435    338
107293305    306
            ...
248364094      1
312958018      1
413119567      1
68435038       1
44472323       1
Name: host_id, Length: 8110, dtype: int64


Column: price
150     288
250     261
200     257
199     252
```

```
299     241
       …
7476      1
5333      1
1159      1
3206      1
0         1
Name: price, Length: 1852, dtype: int64


Column: minimum_nights
1       7816
2       5419
3       5158
5       3329
4       1985
30      1633
7       1406
180      452
6        404
29       326
90       218
10        81
14        78
31        75
28        46
8         19
21        18
15        12
9         12
60        10
20         9
25         7
27         6
12         6
181        6
365        5
13         5
91         4
182        3
45         3
89         3
100        2
185        2
80         2
16         2
85         1
360        1
```

```
32          1
48          1
183         1
160         1
23          1
17          1
61          1
86          1
19          1
35          1
22          1
200         1
184         1
1000        1
44          1
120         1
Name: minimum_nights, dtype: int64


Column: number_of_reviews
0       5539
1       2259
2       1528
3       1154
4        973
        …
327        1
247        1
902        1
550        1
487        1
Name: number_of_reviews, Length: 443, dtype: int64


Column: calculated_host_listings_count
1       5364
2       2550
3       1500
4       1036
660      660
        …
52        52
51        51
48        48
46        46
40        40
Name: calculated_host_listings_count, Length: 108, dtype: int64
```

```
Column: availability_365
0      1550
365     444
364     303
90      218
7       209
       …
27       36
55       36
96       36
37       36
54       32
Name: availability_365, Length: 366, dtype: int64


Column: number_of_reviews_ltm
0      7520
1      2832
2      1837
3      1477
4      1171
       …
115       1
184       1
162       1
306       1
101       1
Name: number_of_reviews_ltm, Length: 157, dtype: int64
```

1000 is an outlier it doesnt seem possible to rent an airbnb for 1000 nights. I am also going to remove the row where the price is equal to 0 because there is no free stays for airbnb.

[43]:
```
# Remove the outlier
i=listings[(listings.price == 0)].index
listings=listings.drop(i)
```

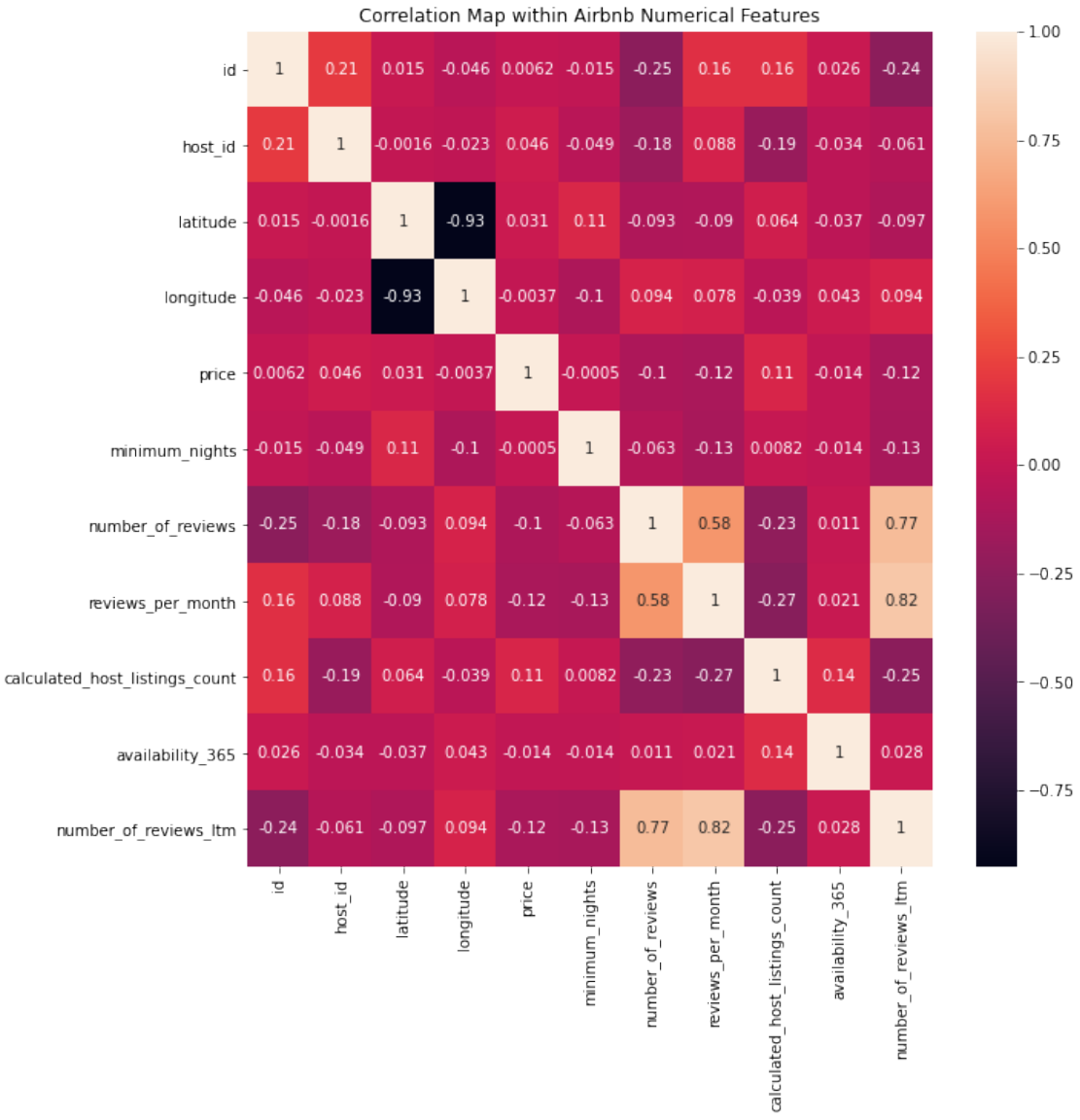
[45]:
```
# Remove the imposible number
i2=listings[(listings.minimum_nights == 1000)].index
listings=listings.drop(i2)

# code adapted from https://stackoverflow.com/questions/43136137/
 ↪drop-a-specific-row-in-pandas
```

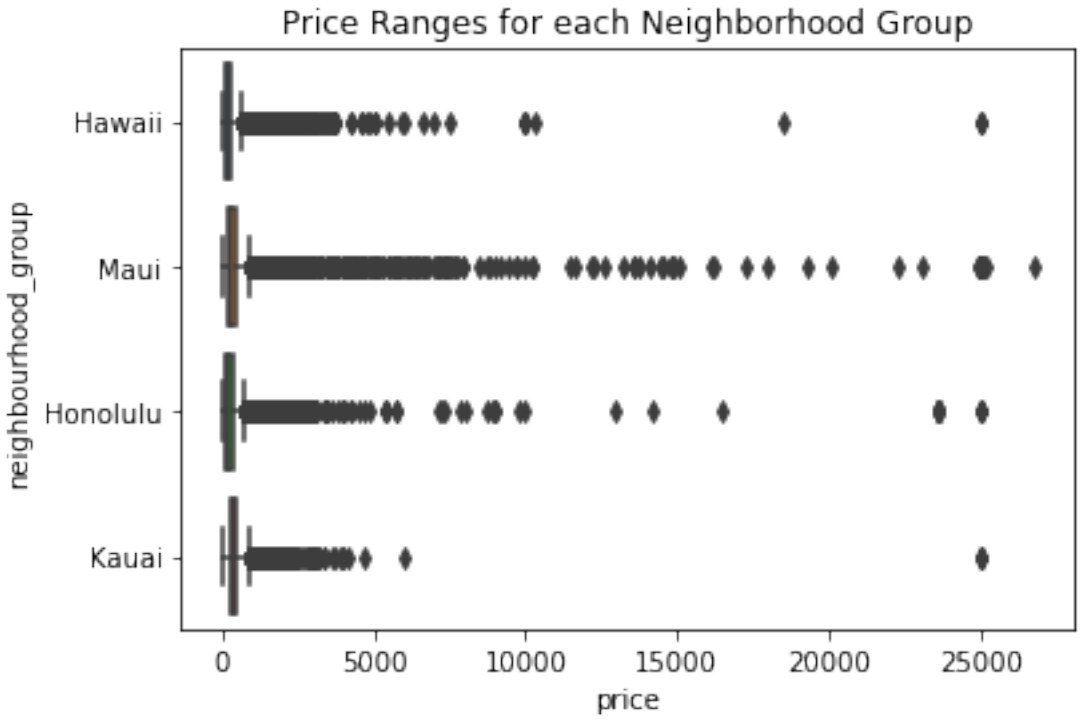
1.10 Task 4: Analyze the Data

Now that your data has been loaded and explored, continue to analyze the data further. With our key question being how various factors affect the pricing of Airbnb rentals, you might want to start your analysis with how something like location or availability affects the price of rentals.

[28]:
```
# Find correlations withtin the data
correlation= listings.corr()
plt.figure(figsize=(10,10))
plt.title('Correlation Map within Airbnb Numerical Features')
sns.heatmap(correlation, annot=True,);
```

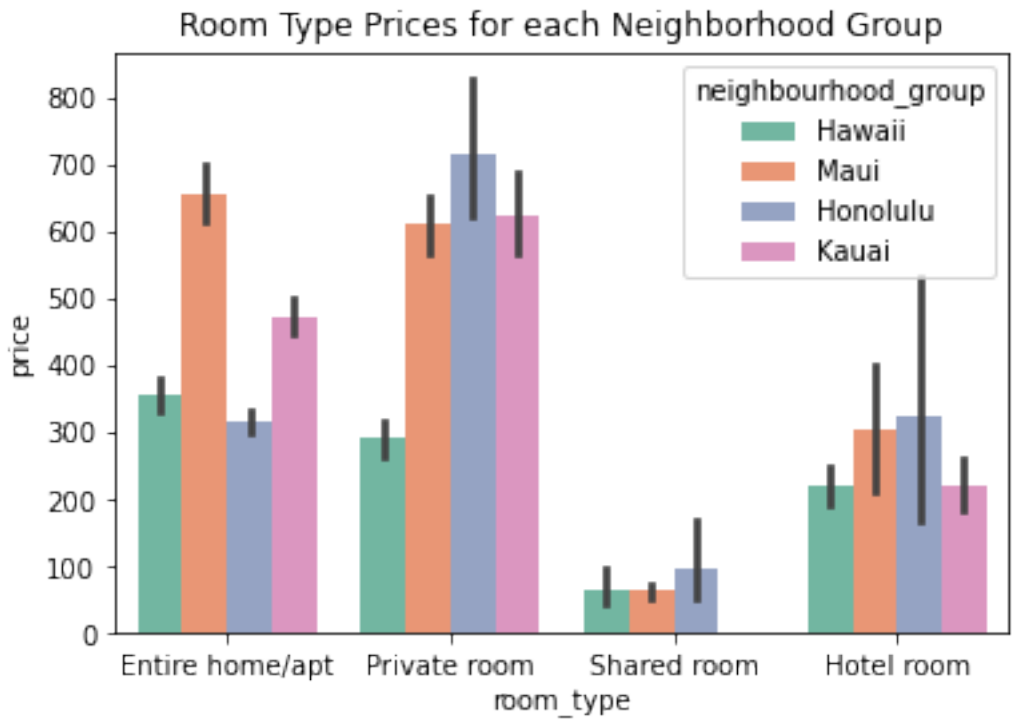


[44]:
```
# Graph a boxplot for the price ranges in each neighbourhood_group
sns.boxplot(data=listings, x="price", y="neighbourhood_group")
plt.title('Price Ranges for each Neighborhood Group');
```

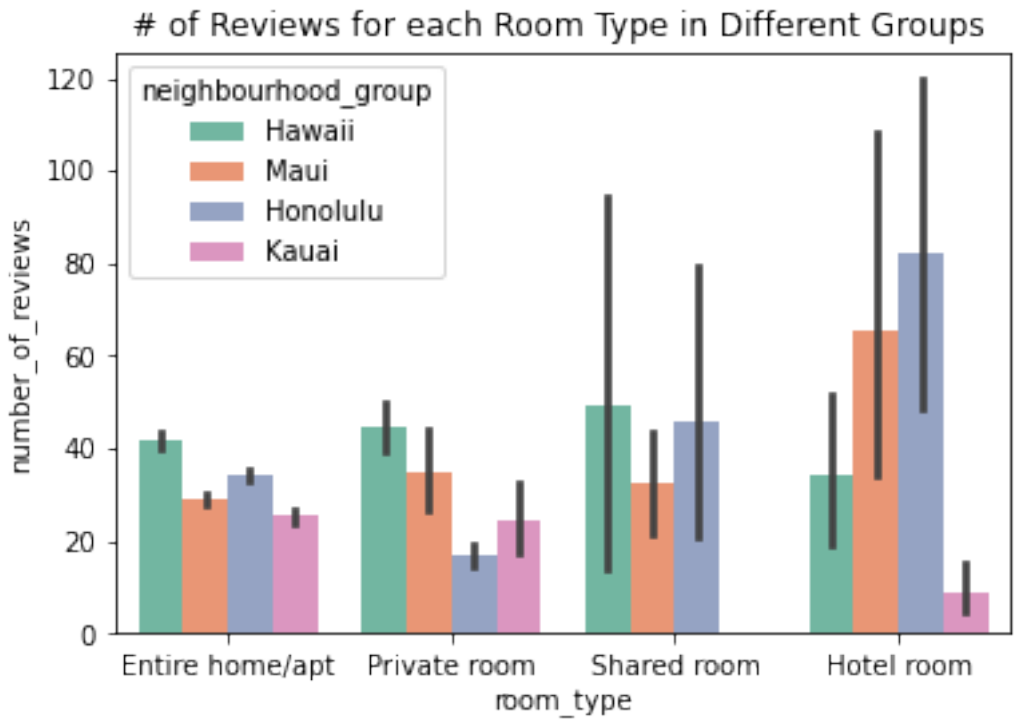


[45]:
```
# Change the room_type column from an integer to an object to be able to graph␣
 ↪it
listings['room_type'] = listings['room_type'].astype(object)
```

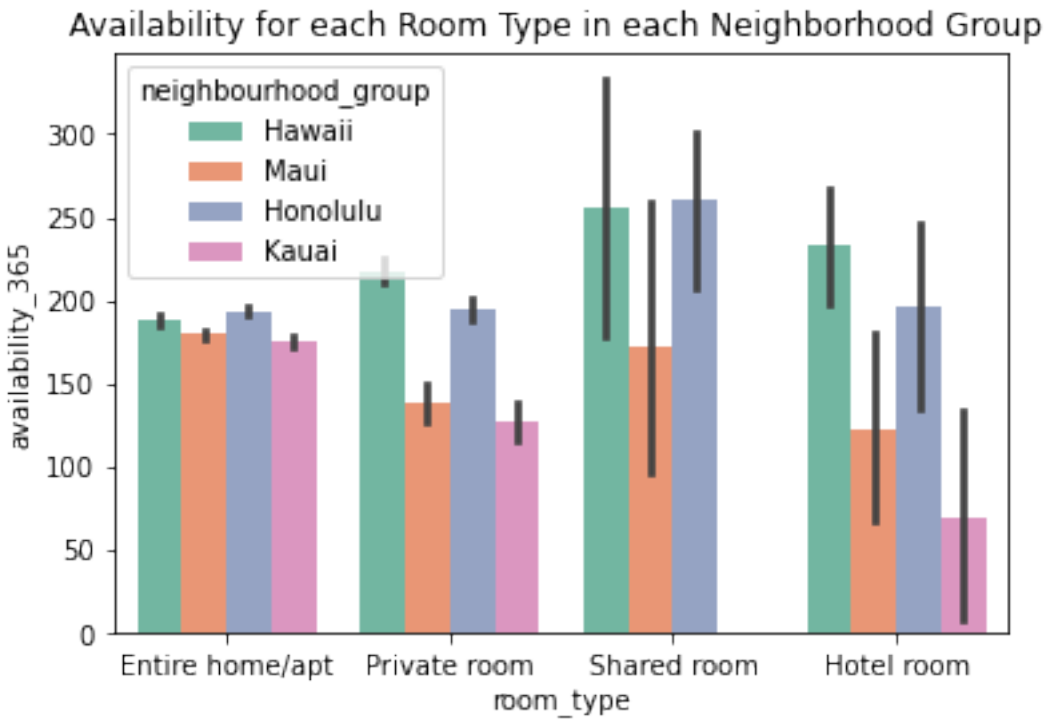
[46]:
```
# Graph a barplot to compare the room type prices in each neighborhood in Hawaii
sns.barplot(data=listings, x='room_type', y='price', hue='neighbourhood_group',␣
 ↪palette='Set2')
plt.title('Room Type Prices for each Neighborhood Group');
```



[47]:
```
# Graph a barplot to compare which rooms get rated the most in each␣
 ↪neighborhood group
sns.barplot(data=listings, x='room_type', y='number_of_reviews',␣
 ↪hue='neighbourhood_group', palette='Set2')
plt.title('# of Reviews for each Room Type in Different Groups ');
```



[49]:
```
# Graph a barplot to compare availability for each room type in each␣
 ↪neighborhood group
sns.barplot(data=listings, y='availability_365', x='room_type',␣
 ↪hue='neighbourhood_group', palette='Set2')
plt.title('Availability for each Room Type in each Neighborhood Group ');
```



1.11 Task 5: Findings and Conclusions

'?:"Write a conclusion about your process and any key findings here. You will be asked to submit these findings later in this experience when you upload your work. Make sure to limit your response to 500 words.

When completing this section, consider including the following: - What did you learn throughout the process? - Are the results what you expected? - What are the key findings and takeaways?

Add your Findings and Conclusions as markdown text here

After completing this project, I learned that the different room types of Airbnbs in Hawaii will vary in price depending on the island. Maui's, Honolulu's and Hawaii's most expensive listings are private rooms, while Kauai's most expensive listings are entire homes/apartments. Shared rooms are the least expensive for all the islands except Kauai since they don't have any. According to the number of reviews, hotel rooms in Honolulu and Maui are the most popular, which could be because they are the second most affordable option across all four islands. When comparing the price ranges between islands, Maui has the largest range of prices and Kauai has the smallest. A major takeaway here is that Honolulu has more options for guests who are planning to get an airbnb.

1.12 Your process and reflections

In this section, include any additional information you think employers should know about your analysis. Discuss your thought process and the rationale behind how you conducted your analysis. You can also include anything that might set you apart from other learners who completed the same case study. Make sure to limit your response to 500 words.

Add your process and reflections as markdown text here

Since there aren't any shared rooms in Kauai, I would recommend finding some places to list on Airbnb. Shared rooms are popular across other islands so it would be a good idea to be able to provide this room type in Kauai. Perhaps more people would consider coming to Kauai after finding out that this option is available to them.