

IDS Project

Airbnb Seattle Data

Team members:

T. NaveenKumar (PES1201801673)

Arpit Kumar (PES1201800406)

K. Sreesh Reddy (PES1201801580)

► Objective:

To clean and analyze the seattle.csv dataset for understanding of how the prices and reviews are listed for the rentals.

► Overview:

Since 2008, guests and hosts have used Airbnb to travel in a more unique personalized way.

As part of the Airbnb Inside initiative, this dataset describes the listing activity of homestays in Seattle, WA

► Steps done:

✓ Data preprocessing:

1. Checking if missing values are present.
2. Filling numerical missing values with mean of the column.
3. Filling categorical missing values with the previous row values using ffill.

✓ Visualizations:

1. Plotting various types of graphs.
2. Drawing inferences from the graphs.
3. Finding correlations.

Filling the missing numerical and categorical values

BEFORE

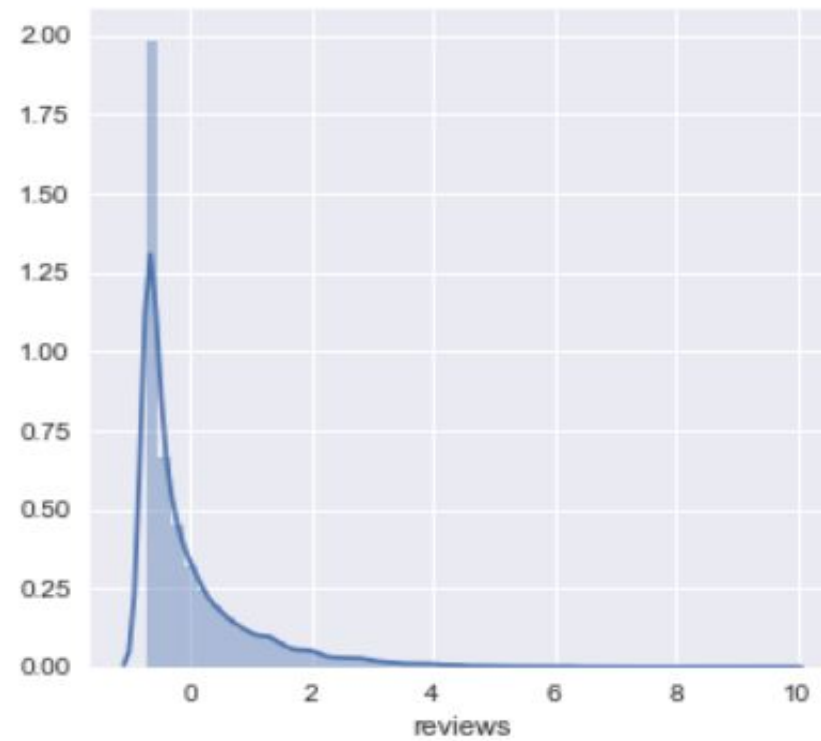
```
room_id      0
host_id      0
room_type    1
address      0
reviews      0
overall_satisfaction 1473
accommodates 0
bedrooms     0
bathrooms    2
price        0
last_modified 0
latitude     0
longitude    0
location     0
name         0
currency     0
rate_type    0
dtype: int64
```

AFTER

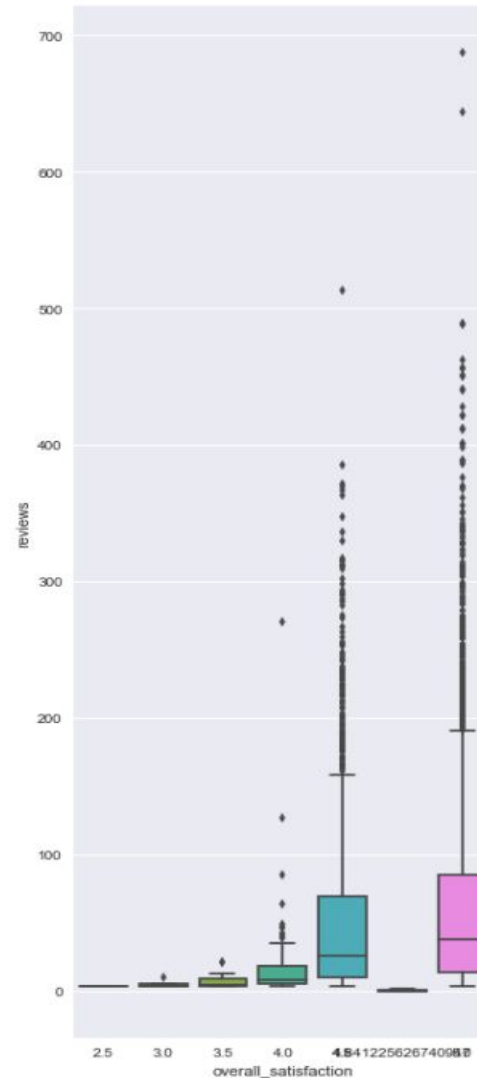
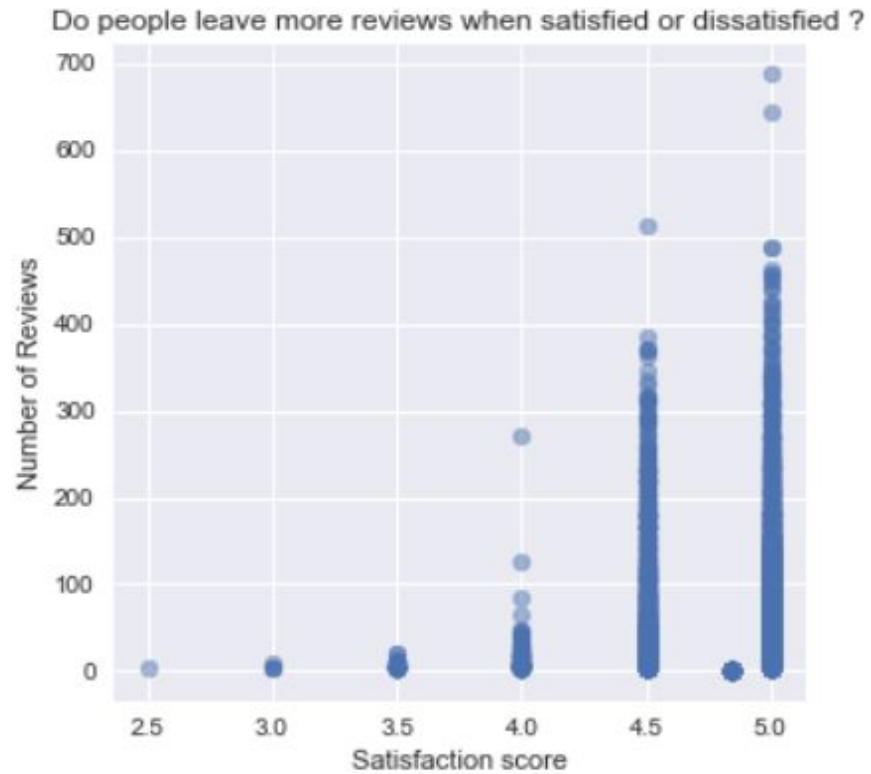
```
room_id      0
host_id      0
room_type    0
address      0
reviews      0
overall_satisfaction 0
accommodates 0
bedrooms     0
bathrooms    0
price        0
last_modified 0
latitude     0
longitude    0
location     0
name         0
currency     0
rate_type    0
dtype: int64
```

Data Visualization

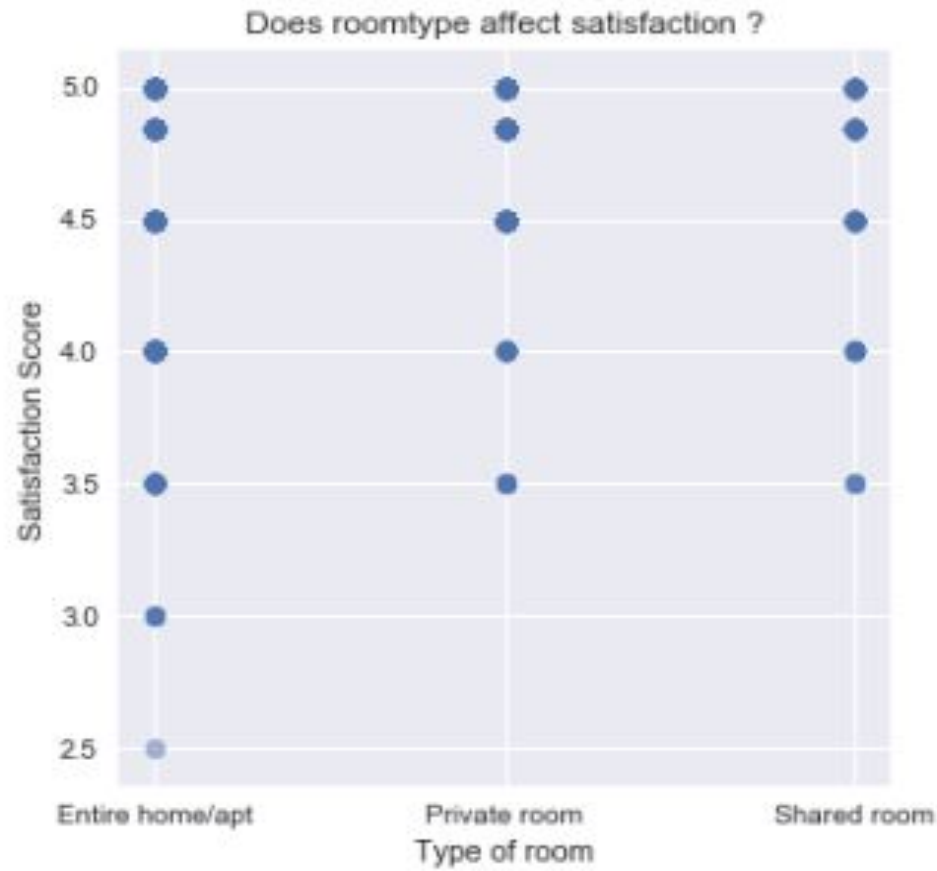
Reviews Normalized Graph



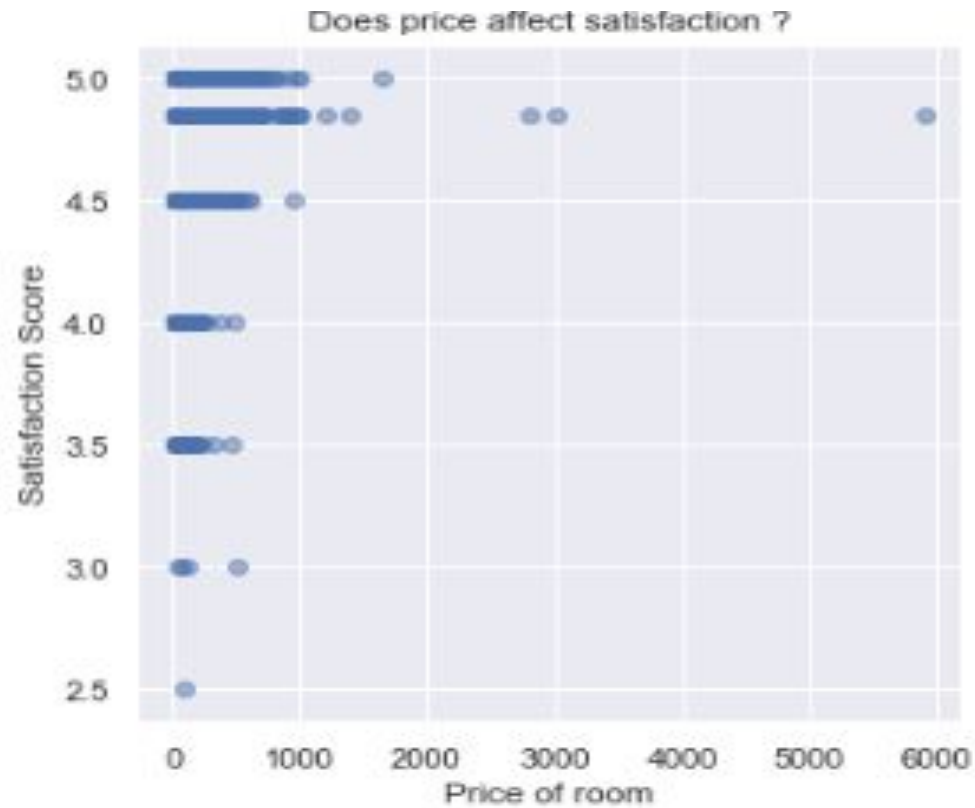
No. of reviews vs. satisfaction score



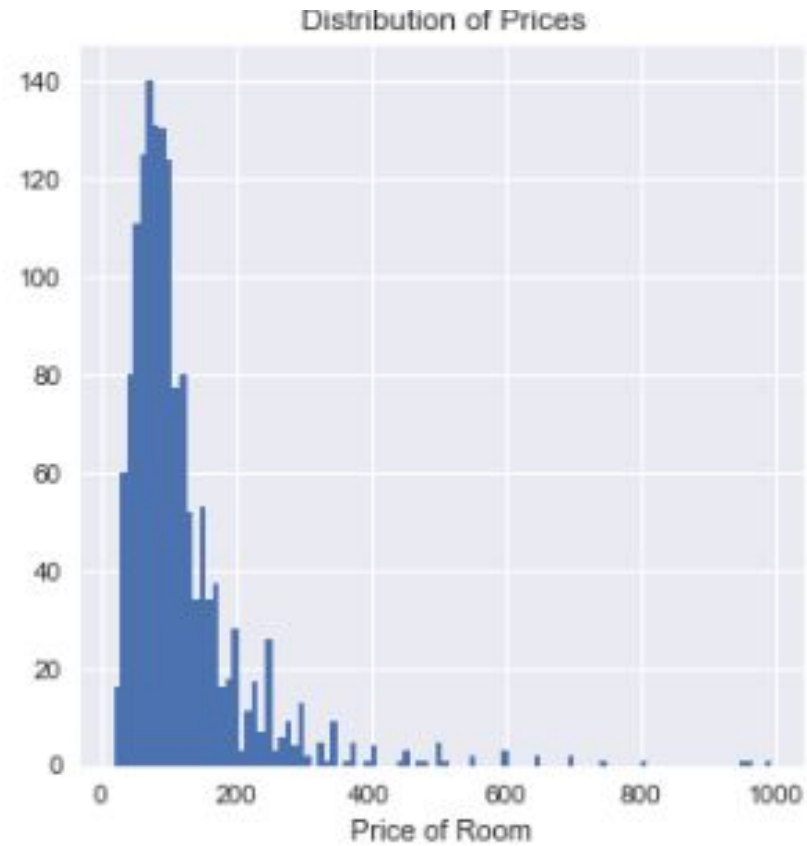
Room type vs Satisfaction score



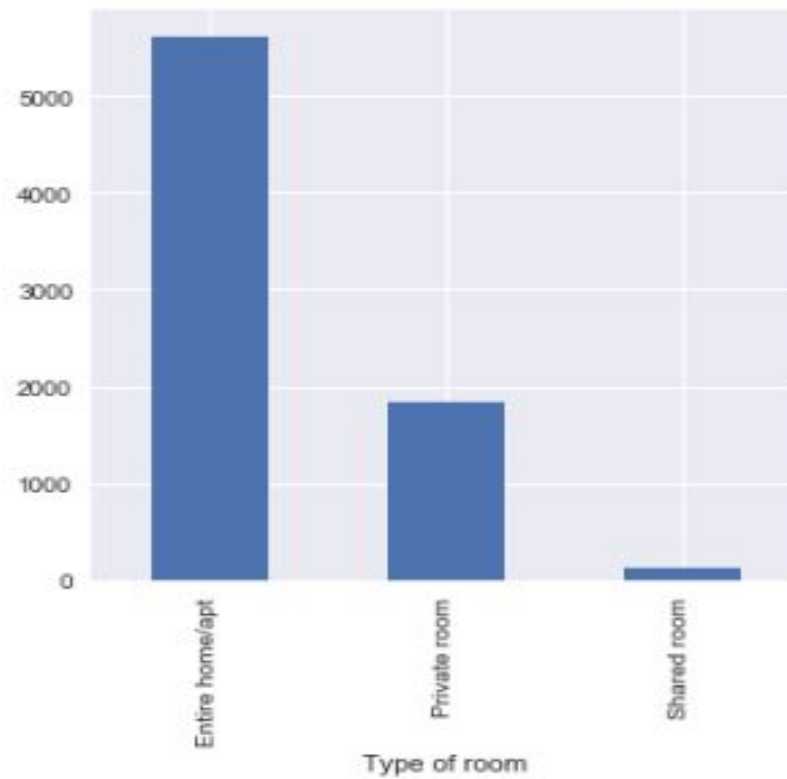
Price vs satisfaction score



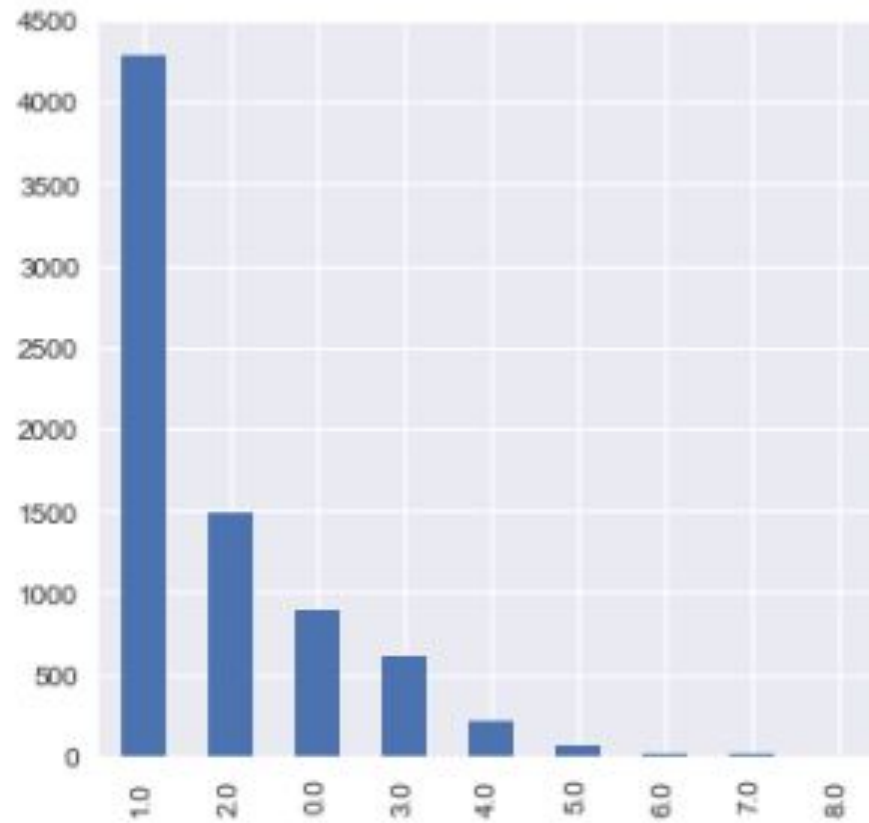
Price Distribution



No. of houses up for rental vs. type of room




No. of houses up for rental vs. no. of bedrooms



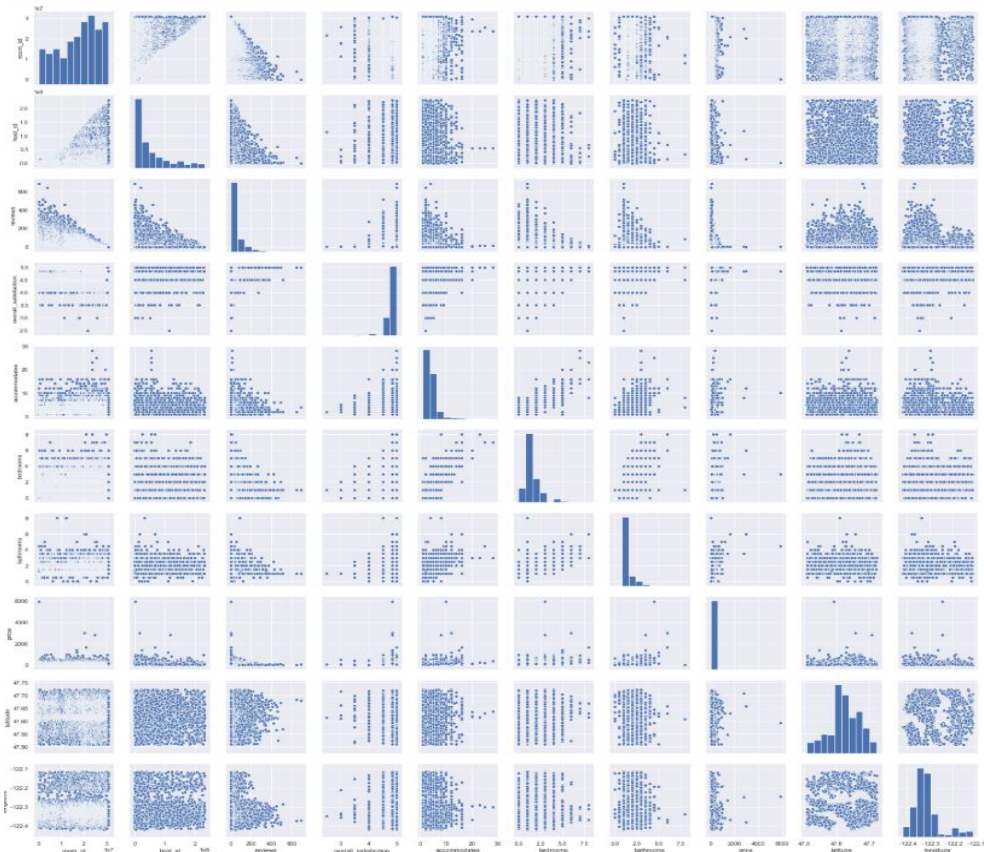
Inferences we can draw from the visualizations

1. People are leaving many more reviews when they are **satisfied** as compared to when they are dissatisfied.
2. The satisfaction score is **not particularly affected by the room type** since all the three types of rooms have similar scatter plot.
3. Likewise, the satisfaction plot is **not affected by the price**, since for the same price, all types of scores are given.

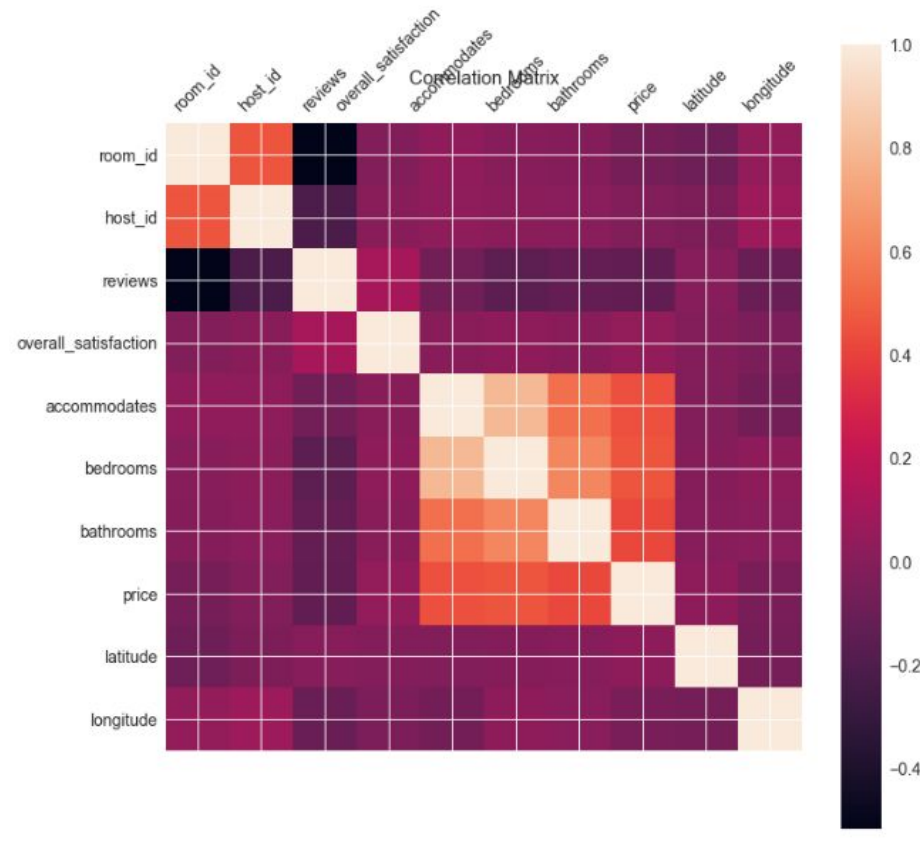
- 
- 4. Majority of the prices of the houses lie in between 0\$ and 200\$.
 - 5. An apartment/Entire room is available more than a private room or a shared room.
 - 6. Similarly, a house with 1 bedroom is available the most.

Correlation

Correlation Plot between each variable



Correlation coefficient
between each variable



Inferences

1. High correlation between no. of accommodates and the no. of bedrooms since more people will require more rooms to live in.
2. Low correlation between the price and the latitude of the location of the house since the latitude is irrelevant to the price.
3. Average correlation between price and the no. of accommodates.
4. No correlation between reviews and room id.

Hypothesis Testing

Let us assume our null hypothesis to be H_0 = Price and no. of bedrooms are not correlated.

Therefore, our alternate hypothesis will be H_1 = Price and no. of bedrooms are correlated.

We import pingouin which will help us give the r (correlation coefficient) , p -value etc. respectively.


```
In [20]: import pingouin as pg
pg.corr(x=airbnb["price"], y=airbnb["overall_satisfaction"])
```

```
In [21]: pg.corr(x=airbnb["price"], y=airbnb["bedrooms"])
```

```
Out[21]:
```

	n	r	CI95%	r2	adj_r2	p-val	BF10	power
pearson	7576	0.462	[0.44, 0.48]	0.214	0.214	0.0	inf	1.0

- ▶ Since, the p-value is 0.0 (which is less than 0.05), we can reject our null hypothesis which was that the price and the no. of bedrooms are not correlated.
- ▶ The r (correlation coefficient) too is 0.462 which is quite high and therefore supports the fact that they are correlated.