Research on the Rent of the Airbnb
by Ronnie Sin

by Ronnie Sin

Table of Content

by Ronnie Sin

1. Data Description(Sample of Data)

	0	1	2
neighborhood_overview	Not even 10 minutes by metro from Victoria Sta...	Finsbury Park is a friendly melting pot commun...	It is Chelsea.
host_name	Adriano	Alina	Philippa
host_location	London, England, United Kingdom	London, England, United Kingdom	Kensington, England, United Kingdom
host_about	Hello, I'm a friendly Italian man with a very ...	I am a Multi-Media Visual Artist and Creative ...	English, grandmother, I have travelled quite ...
host_response_time	within an hour	within a few hours	NaN
host_verifications	['email', 'phone', 'reviews', 'jumio', 'offlin...	['email', 'phone', 'facebook', 'reviews', 'off...	['email', 'phone', 'reviews', 'jumio', 'govern...
neighbourhood	London, United Kingdom	Islington, Greater London, United Kingdom	London, United Kingdom
property_type	Entire apartment	Private room in apartment	Entire apartment
room_type	Entire home/apt	Private room	Entire home/apt

by Ronnie Sin

1. Data Description(Sample of Data)

	0	1	2
amenities	["Refrigerator", "Kitchen", "Crib", "Dedicated...	["Long term stays allowed", "Lock on bedroom d...	["Refrigerator", "Wifi", "Heating", "Dishes an...
host_is_superhost	0	0	0
accommodates	4	2	2
bedrooms	1	1	1
beds	3	1	1
price	105	40	75
minimum_nights	2	1	10
number_of_reviews	192	21	89
review_scores_rating	91	97	96
instant_bookable	1	0	1

1. Data Description

by Ronnie Sin

	Variable	Type	Description
1	host_is_superhost	Binary	Binary indicator for whether or not the host is a superhost
2	accommodates	Numerical	Integer value for the number of people the property accomodates
3	bedrooms	Numerical	Integer value for the number of bedrooms the property has
4	beds	Numerical	Integer value for the number of beds the property has
5	price	Numerical	Float value for the price per day
6	minimum_nights	Numerical	Integer value for the minimum number of nights for a single stay
7	number_of_reviews	Numerical	Integer value for the number of reviews the host has received from renters.
8	review_scores_rating	Numerical	Average of the review scores across all of a host's reviews
9	instant_bookable	Binary	Binary indicator for whether or not the host allows instant bookings
10	neighborhood_overview	String	Free text description of the area provided by airbnb

by Ronnie Sin

1. Data Description

	Variable	Type	Description
11	host_name	String	The first name of the host
12	host_location	String	Where the host lives, provided either as the city, or country of residence.
13	host_about	String	Free text description of the host, provided by the host
14	host_response_time	String	String description of the average time it takes for the host to repsond
15	host_verifications	Categorical	List of strings for the methods airbnb have taken to verify the identity of the host
16	neighbourhood	Categorical	The location of the property, as specified by the host
17	property_type	Categorical	String description of the type of property
18	room_type	Categorical	String description of the type of room, taking one of four possible values
19	amenities	Categorical	List of strings for the amenities available within the property

by Ronnie Sin

2. Data Cleaning

- In this project, we would focus on the binary, categorical & numerical variables.
- When doing analysis, if there are NULL values in the corresponding features, we would drop the corrsponding rows.
- We would add a column “id” to indicate each renting house
- For “host_verifications” & “amenities”, we would:
    - Convert it from str to list
    - split the multiple values
- We would add a column “verification_count” to count the number of “host_verifications”
- For “host_verifications”, if the value = “[]”, we would replace it by “No”

by Ronnie Sin

2. Data Cleaning(Data Issue in “neighbourhood”)

In the variable -“neighbourhood”,

There are different spelling of place, but means the same place. Therefore, we need to regroup the data. We would just take the city of the place instead.

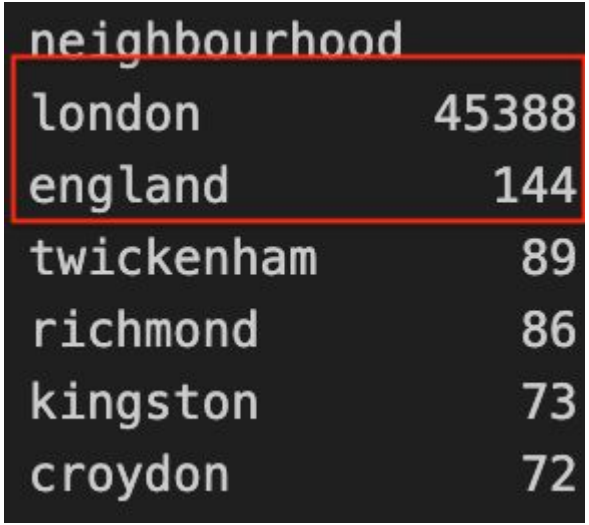
neighbourhood	
Greater London, England, United Kingdom	26731
London, United Kingdom	8431
London, England, United Kingdom	7420
London, Greater London, United Kingdom	977
Greater London, United Kingdom	328
London , England, United Kingdom	224
England, United Kingdom	144
Twickenham, United Kingdom	50
Croydon, England, United Kingdom	48
Richmond, England, United Kingdom	43
Kingston upon Thames, England, United Kingdom	35
Richmond, United Kingdom	34
Kingston upon Thames, United Kingdom	32
Wembley, England, United Kingdom	31
Twickenham, England, United Kingdom	30
London, UK, United Kingdom	30
Dagenham, England, United Kingdom	26
Edgware, England, United Kingdom	26
London , London, United Kingdom	24
Romford, England, United Kingdom	24
Londres, England, United Kingdom	24
Woodford, England, United Kingdom	23
London, ., United Kingdom	22
London, Uk, United Kingdom	21

by Ronnie Sin

2. Data Cleaning(Data Issue in “neighbourhood”)

After cleaning the data, we found that most renting place is in London. To make our analysis more representative, we need enough data in each category for each variable.

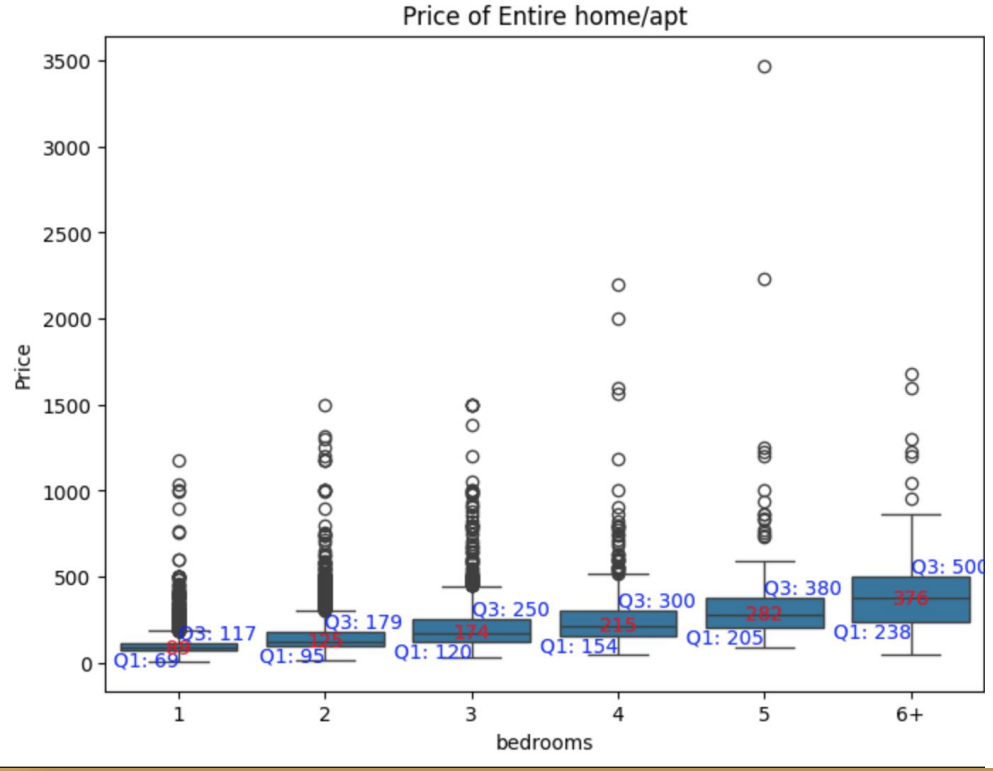
In this case, to make it simpler, we can just focus on the renting market in London

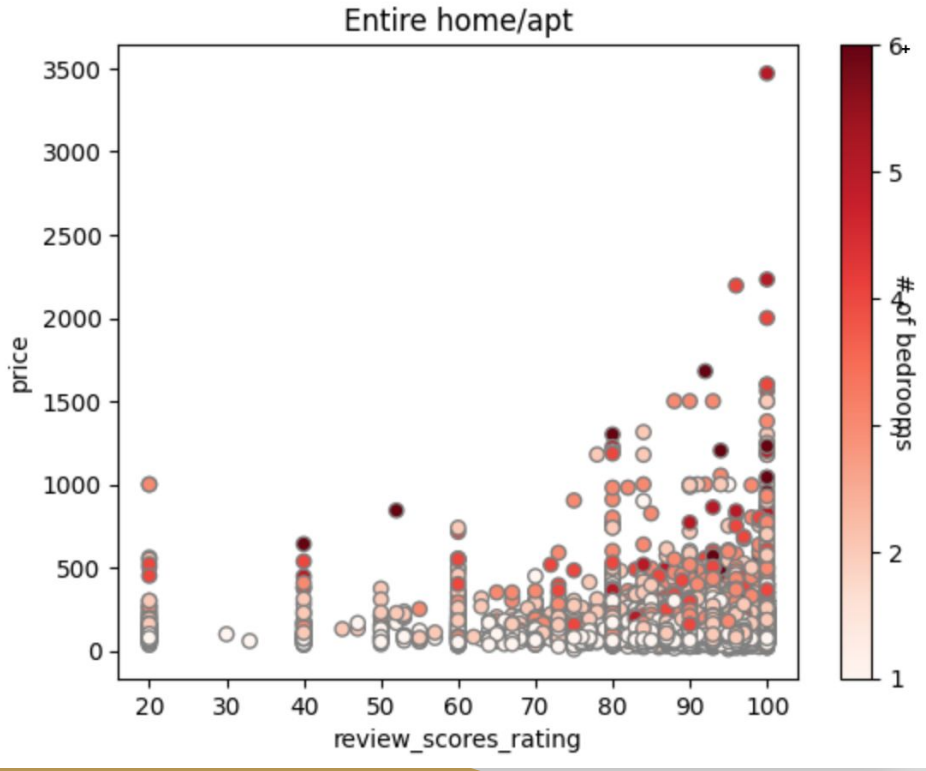


neighbourhood	
london	45388
england	144
twickenham	89
richmond	86
kingston	73
croydon	72

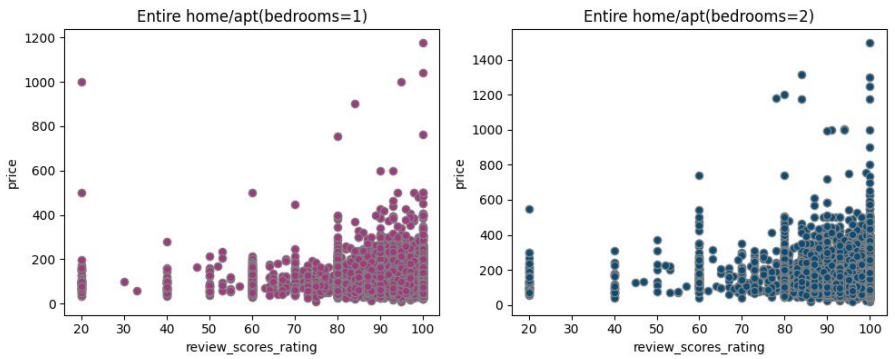
by Ronnie Sin

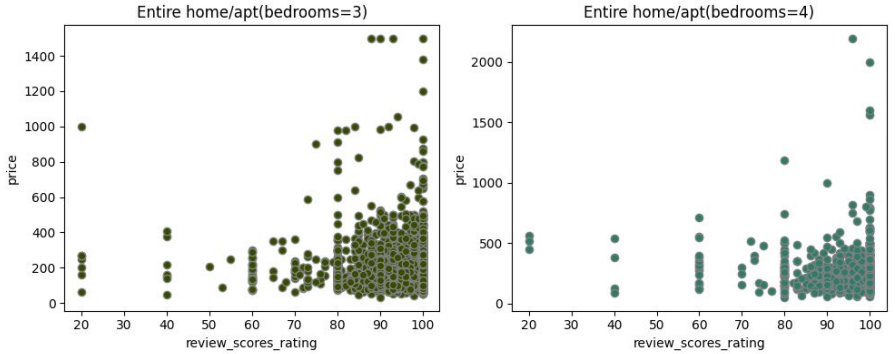
3. Data Overview, Grouped by Features

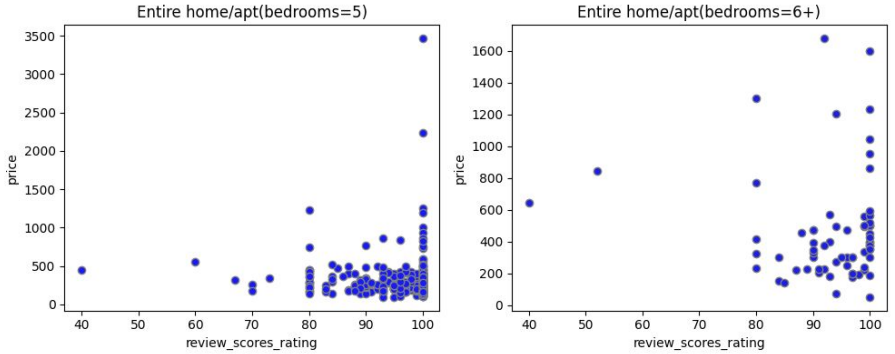




by Ronnie Sin

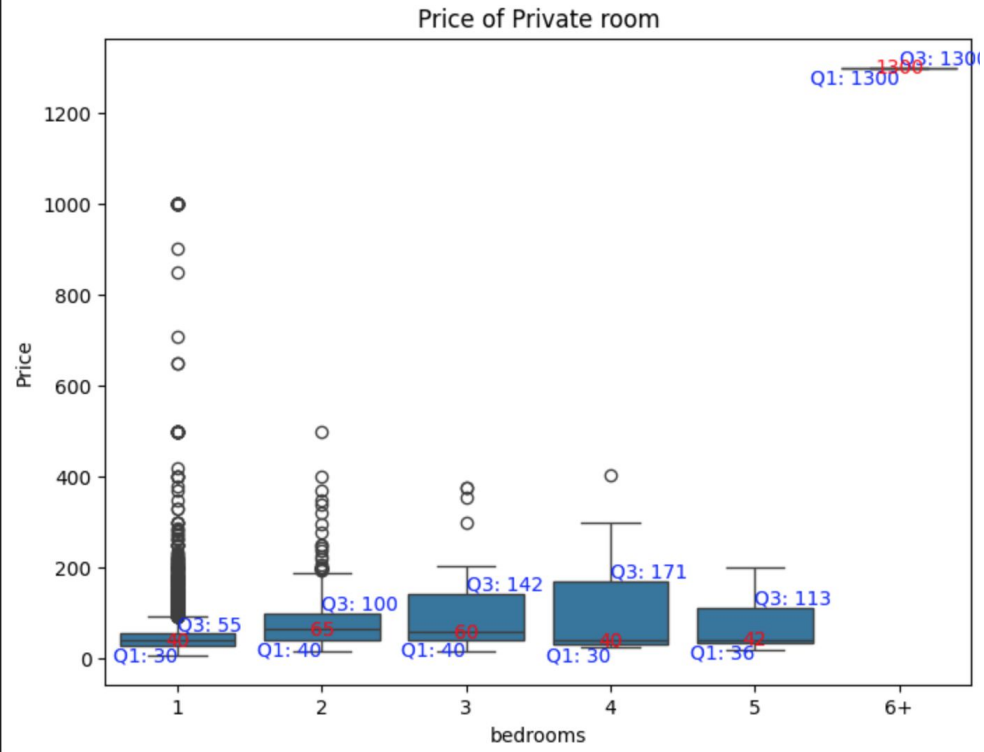


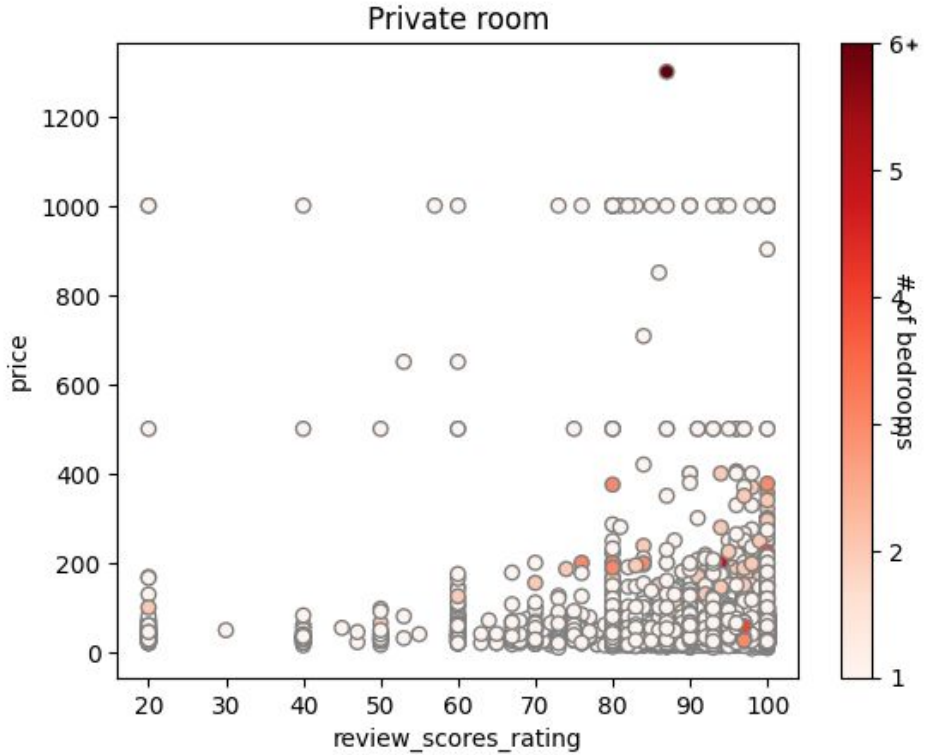




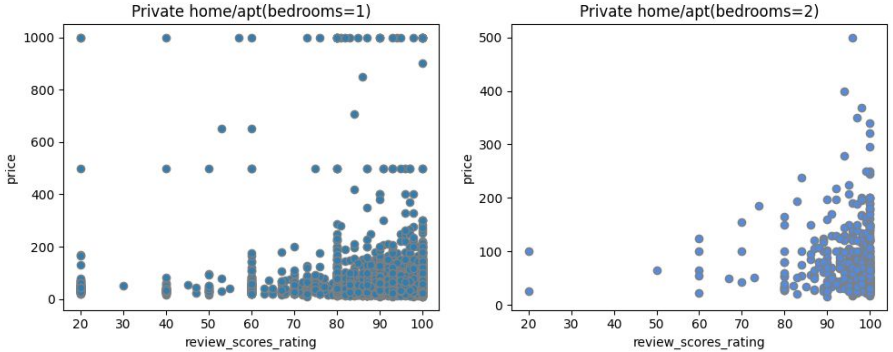
by Ronnie Sin

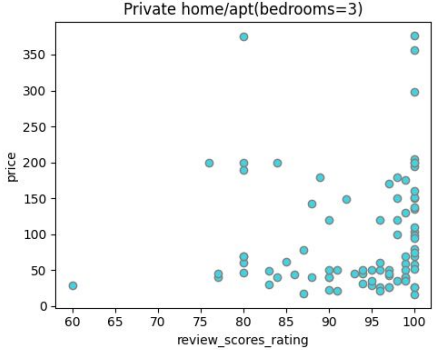
3. Data Overview, Grouped by Features

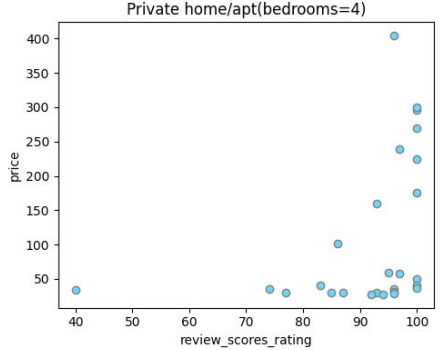


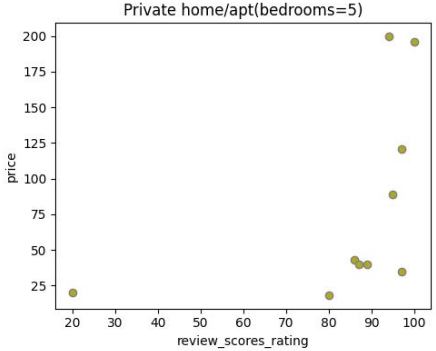


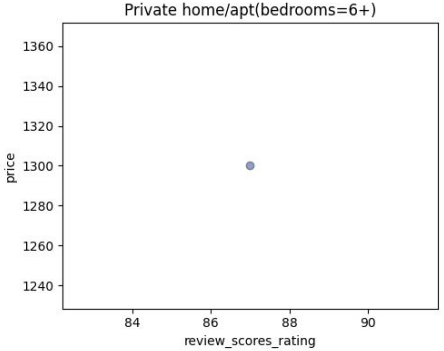
by Ronnie Sin





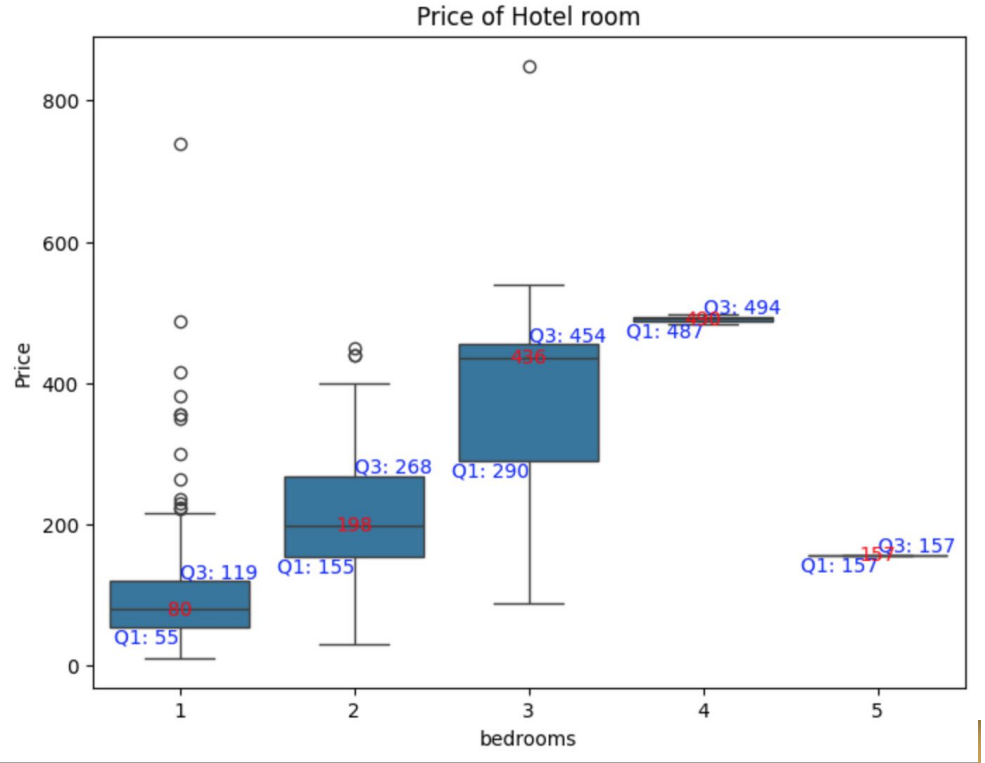


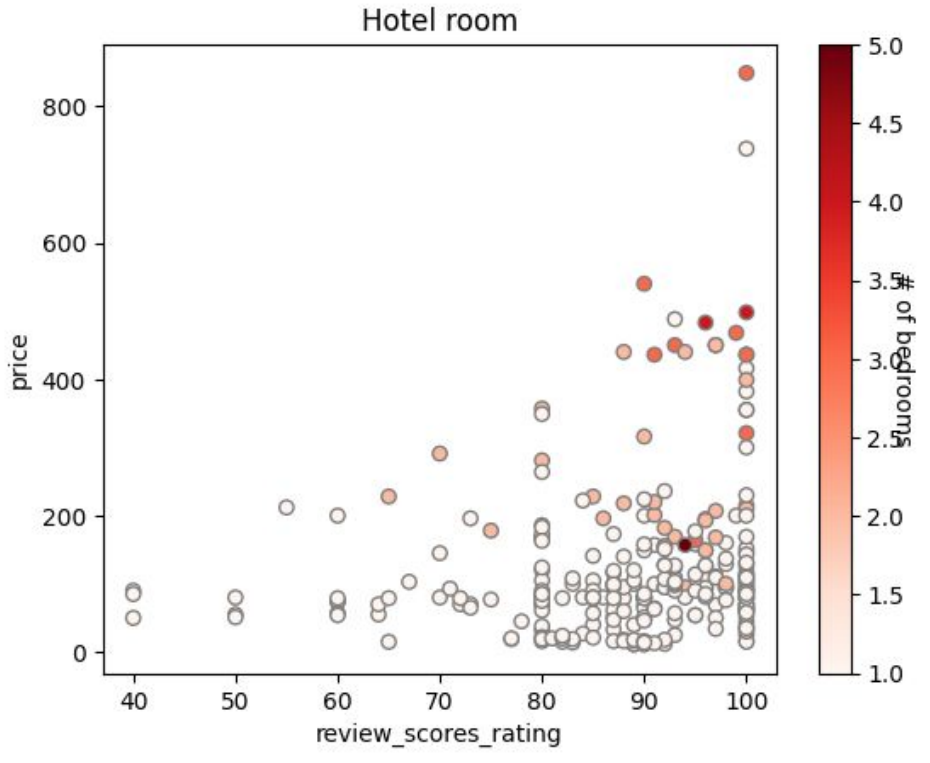




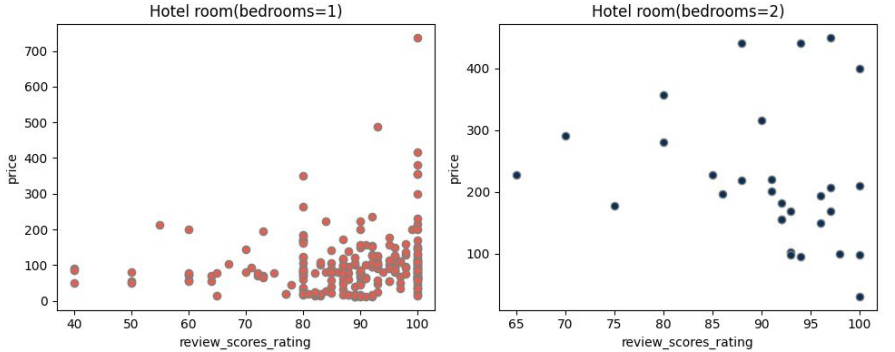
by Ronnie Sin

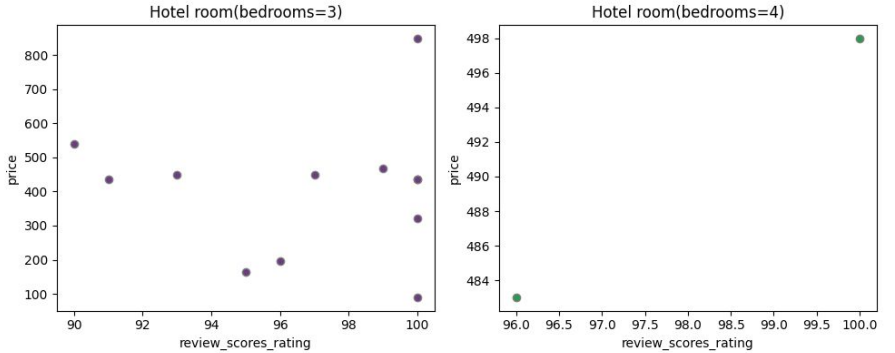
3. Data Overview, Grouped by Features

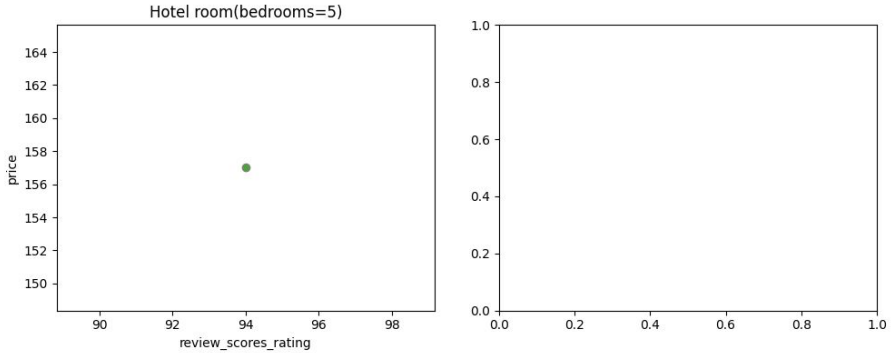




by Ronnie Sin

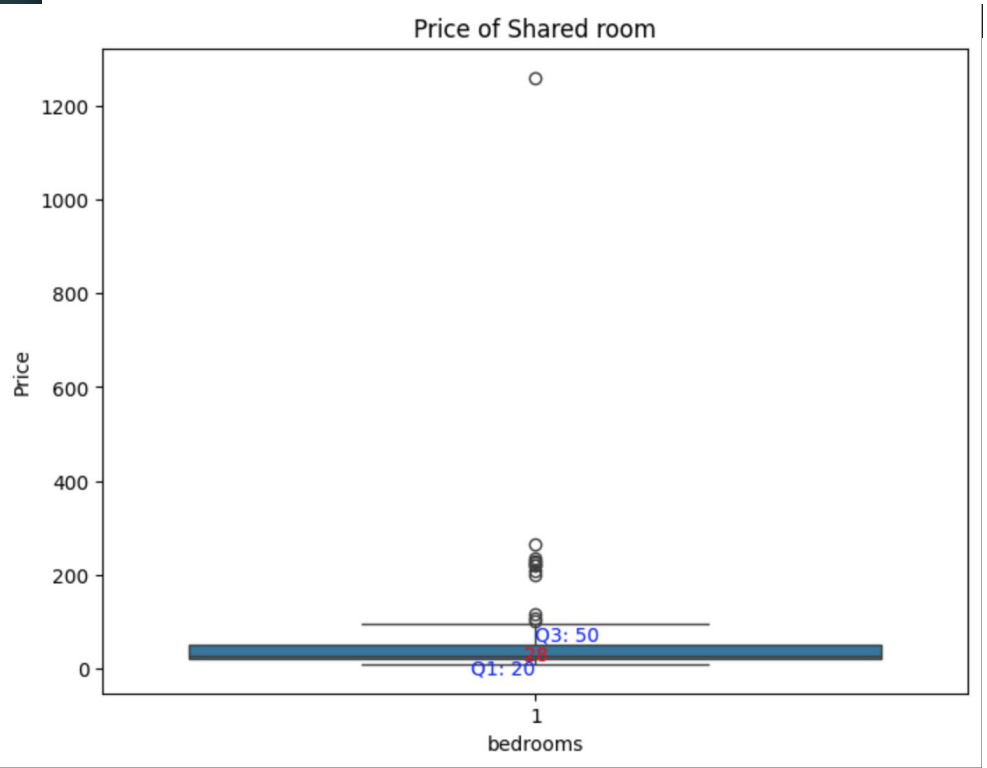


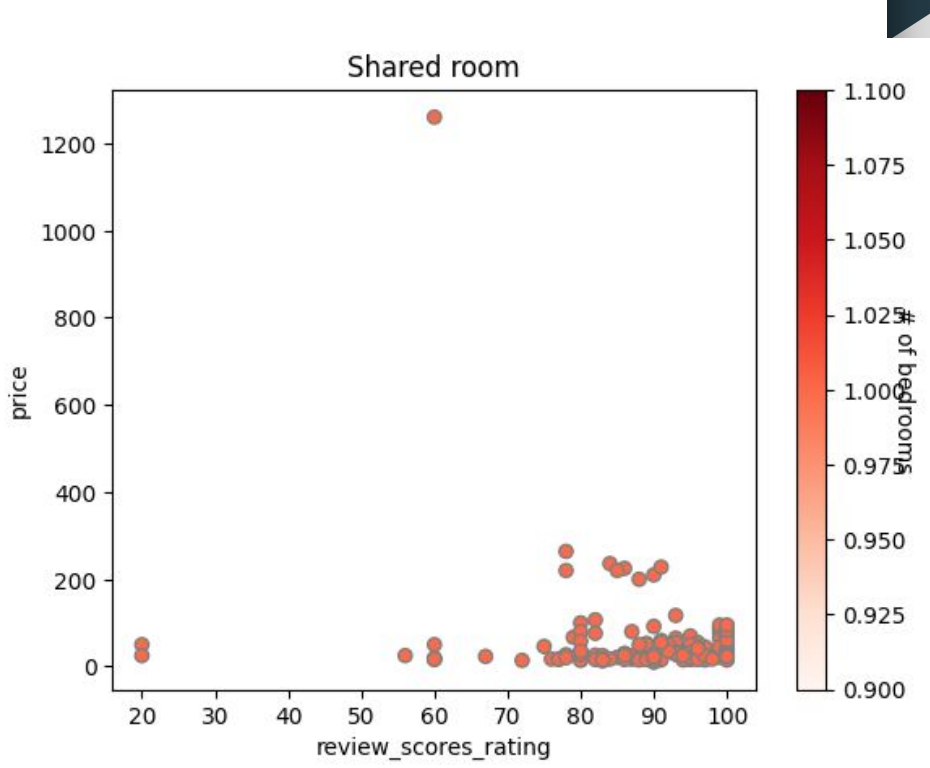




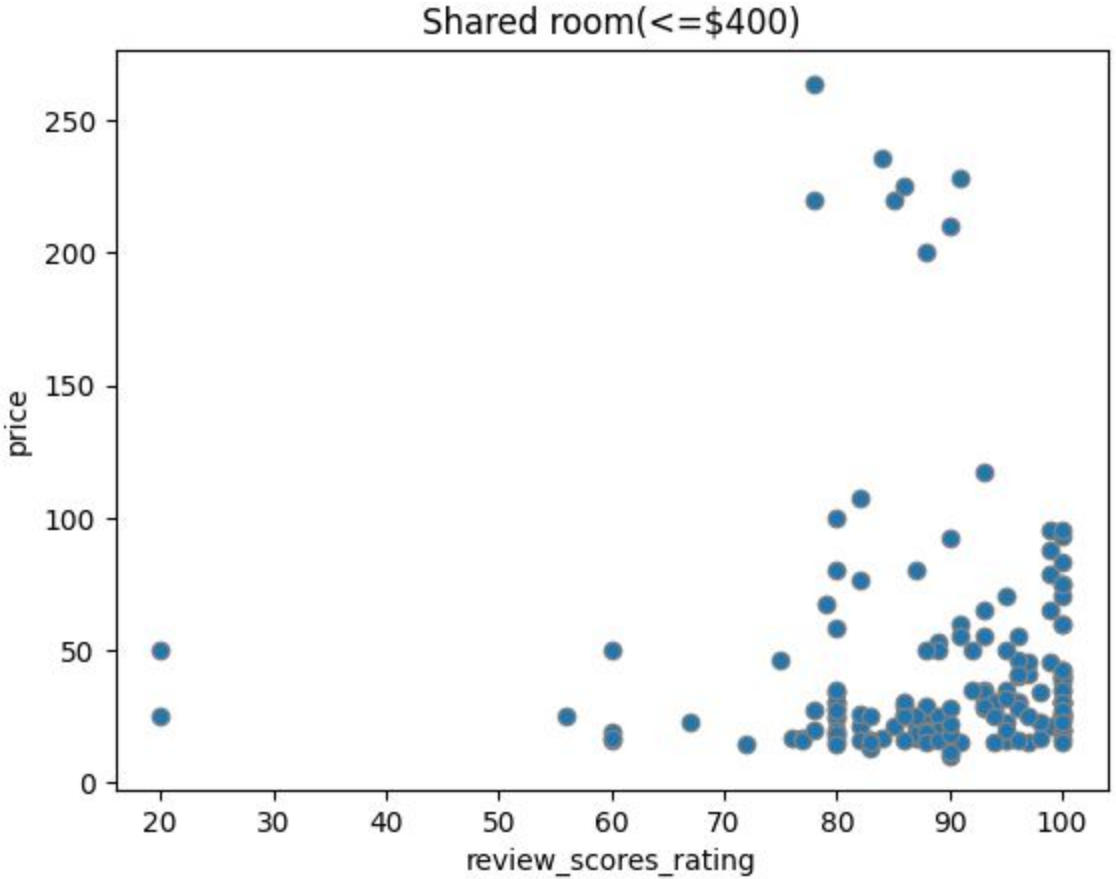
by Ronnie Sin

3. Data Overview, Grouped by Features





by Ronnie Sin



3. Data Overview, Grouped by Features

- More Bedrooms -> higher price
- Rating vs Price(Weekly positive correlation)
  - Generally, better rating would result in slightly higher price
  - Higher Rating -> higher price limit(greater chance, but not guarantee for higher price)

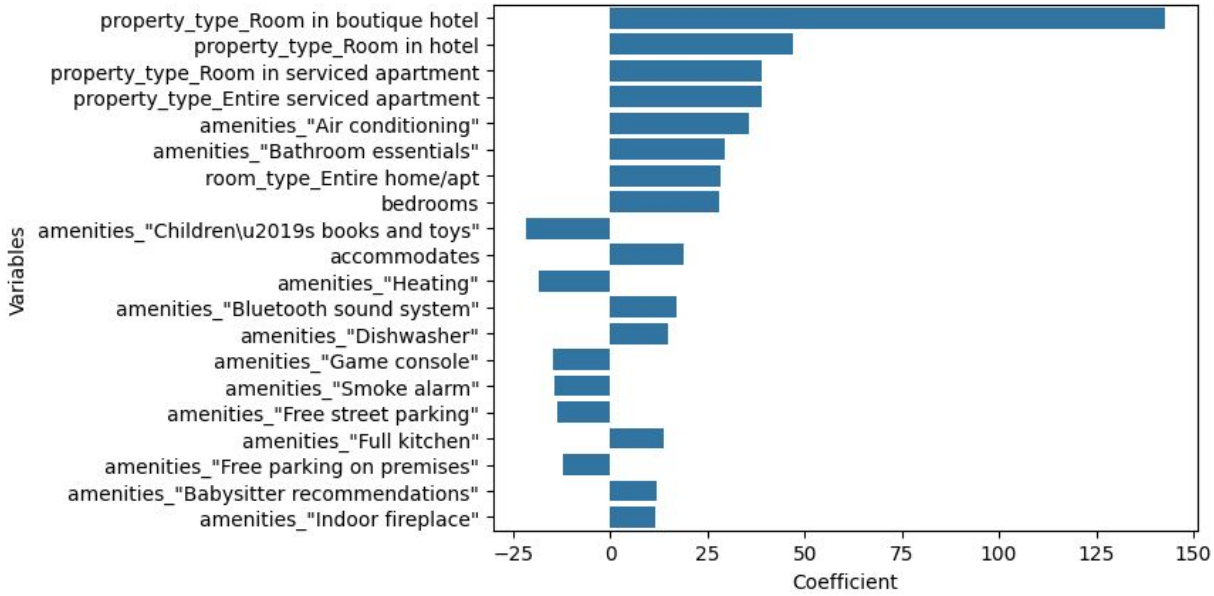
4. Lasso Regression(y: prices, x: features)

- Filter out the categorical value with count<100
- Convert categorical variables to binary variables
- Train-test-split : 7:3
- Scaling numerical variables
- 5-fold cross-validation for finding best alpha

by Ronnie Sin

4. Lasso Regression(Features Importance & r^2)

	Coefficient	Variables
198	142.890334	property_type_Room in boutique hotel
200	46.965118	property_type_Room in hotel
201	39.096004	property_type_Room in serviced apartment
166	38.822196	property_type_Entire serviced apartment
9	35.604696	amenities_"Air conditioning"
17	29.442608	amenities_"Bathroom essentials"
210	28.192312	room_type_Entire home/apt
2	28.131530	bedrooms
32	-21.842658	amenities_"Children\u2019s books and toys"
1	18.865961	accommodates
67	-18.563791	amenities_"Heating"
22	16.979325	amenities_"Bluetooth sound system"
43	14.941046	amenities_"Dishwasher"
61	-14.567776	amenities_"Game console"
112	-14.291320	amenities_"Smoke alarm"
58	-13.822101	amenities_"Free street parking"
60	13.688617	amenities_"Full kitchen"
57	-12.348310	amenities_"Free parking on premises"
14	12.036302	amenities_"Babysitter recommendations"
73	11.625336	amenities_"Indoor fireplace"




```
r_2_traing_set : 41.74
r_2_test_set : 43.55
mean_squared_error : 5657.951311358388
```