Download the listings.csv.gz from the websit

Data science process

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Some possible important features

- ✧ calculated_host_listings_count,
- ✧ host_is_superhost,
- ✧ accommodates,
- ✧ bathrooms,
- ✧ bedrooms,
- ✧ beds,
- ✧ price,
- ✧ number_of_reviews,
- ✧ review_scores_rating,
- ✧ review_scores_value

Some exploration in HK_listings.csv

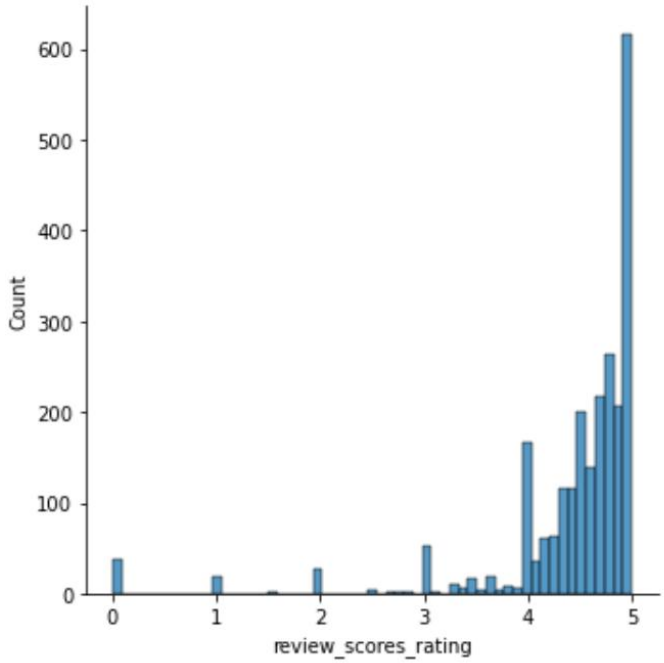
1. 5057 records, 75 features(cloumns)

```
    # in listings, drop the row with the null price
    listings.dropna(axis=0,subset=['price'],inplace=True)
    listings.shape
```

✓ 0.1s

```
(5057, 75)
```

2. Revies score is relative high in total.



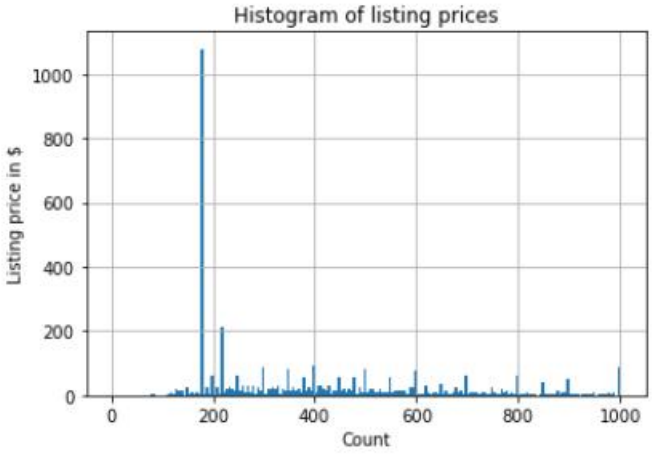
3. Price check in statistic

```
listings['price'].describe()
```

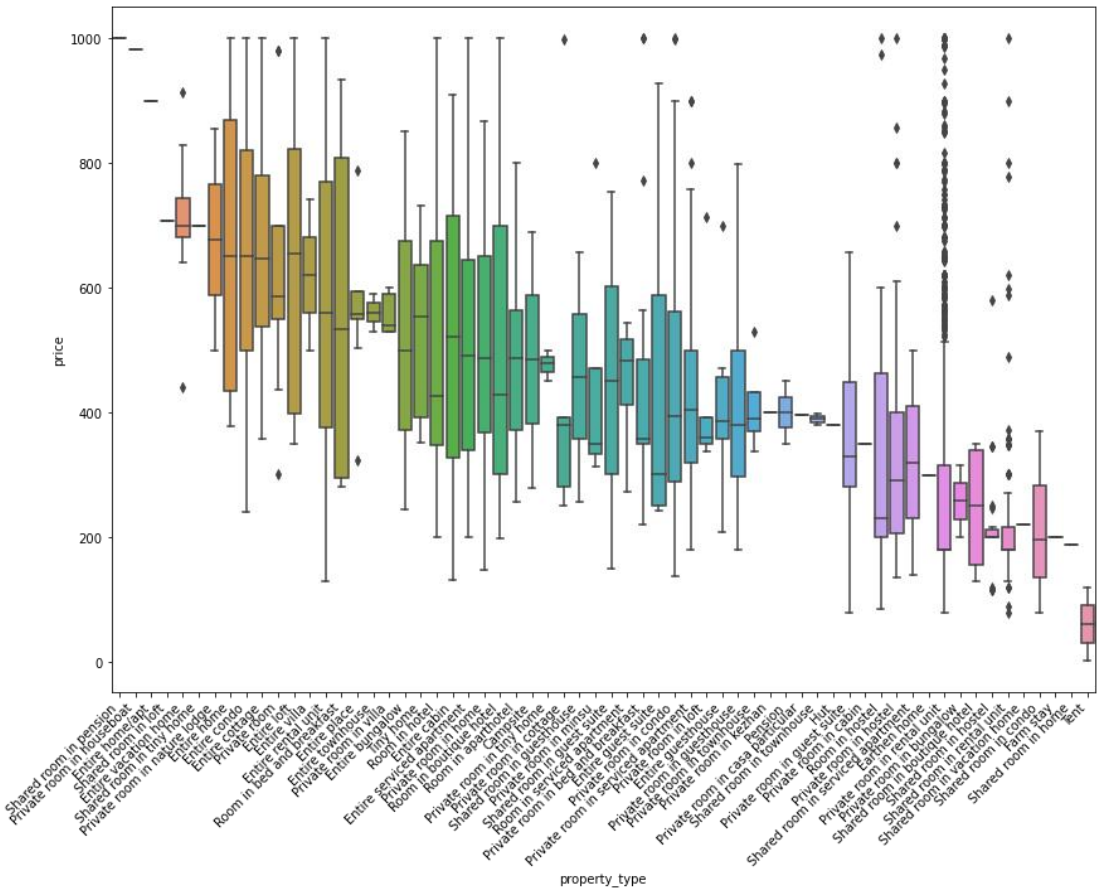
✓ 0.1s

```
count      5057.000000
mean        878.140597
std        2841.561223
min           0.000000
25%         200.000000
50%         414.000000
75%         800.000000
max      117631.000000
Name: price, dtype: float64
```

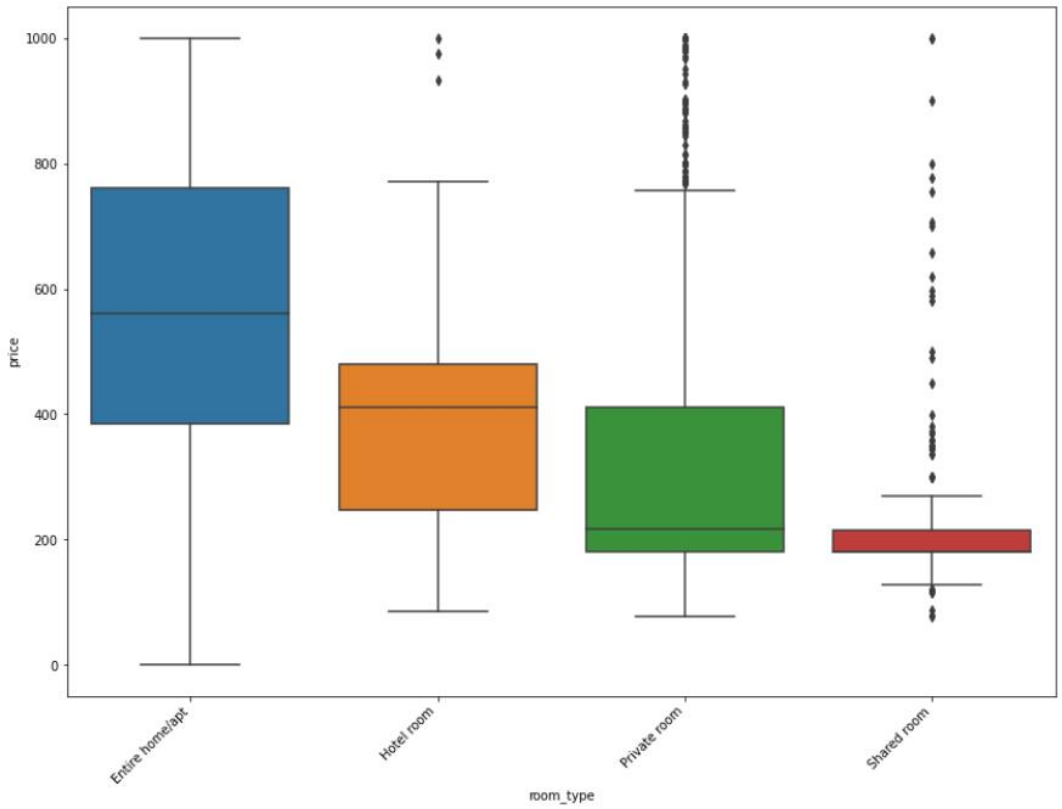
4. Set a filter condition: take the part where the house price is between 0-600$
5. Price histogram



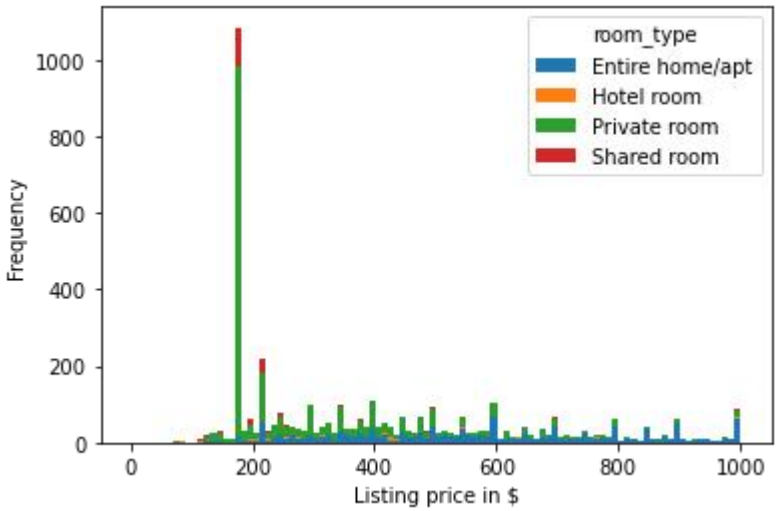
6. the price boxplot of each type in property_type



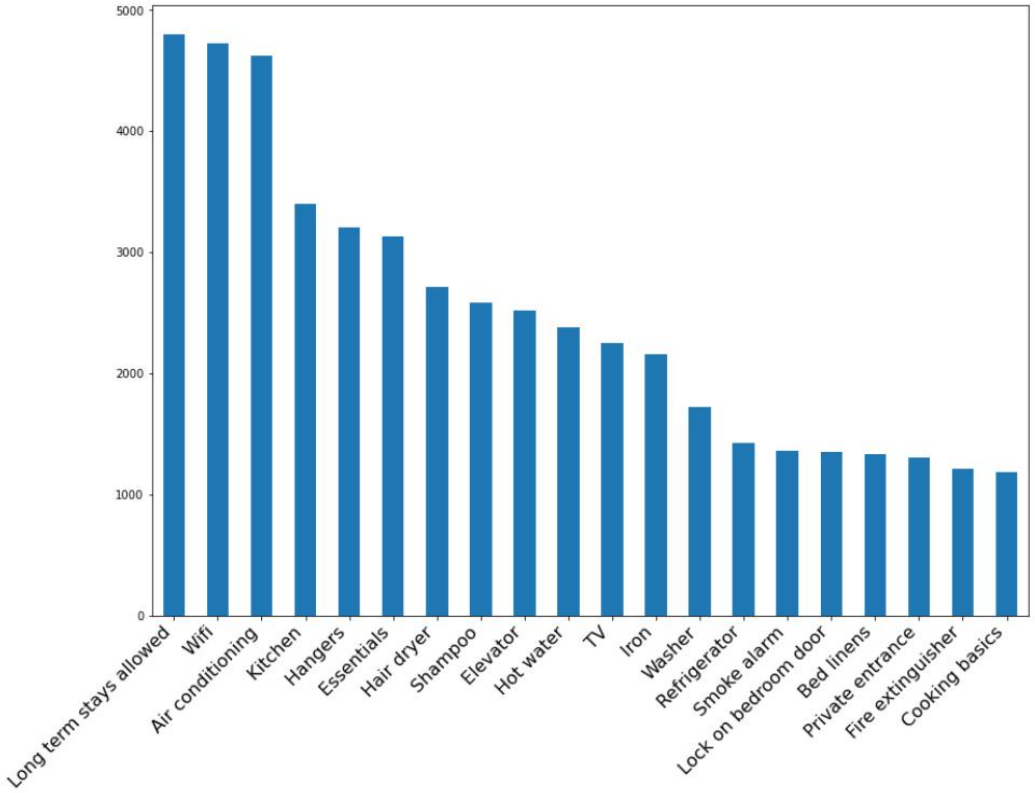
7. the price boxplot of each type in room_type



8. the cumulative histogram of each type of room_type price



9. the top20 devices with the most occurrences


5000
4000
3000
2000
1000
0
Long term stays allowed
Wifi
Air conditioning
Kitchen
Hangers
Essentials
Hair dryer
Shampoo
Elevator
Hot water
TV
Iron
Washer
Refrigerator
Smoke alarm
Lock on bedroom door
Bed linens
Private entrance
Fire extinguisher
Cooking basics