Bi-weekly Report (11/15) - Group 9

1. Dataset Selection

We found two datasets of Airbnb in Seoul.

The first dataset was from https://github.com/gogoj5896/2_teamproject_air_bnb_ligression. It is based on the second dataset, but added some features, such as bathroom and kitchens. This dataset included ['id', 'accommodates', 'bedrooms', 'city', 'host_id', 'overall_satisfaction', 'price', 'reviews', 'room_type', 'comfort', 'Bathrooms', 'Bedrooms', 'Beds', 'Cleaning Fee', 'Extra people', 'Property type', 'Room type', 'Total reviews', 'Weekend Price', 'others', 'review 1', 'review 2', 'review 3', 'review 4', 'review 5', 'review 6', 'superhost', 'Internet', 'Somke_detector', 'Family_kid_friendly', 'Kitchen'],

The second dataset was collected by Tom Slee(https://tomslee.net/category/airbnb-data). This included ['room_id', 'survey_id', 'host_id', 'room_type', 'country', 'city', 'borough', 'neighborhood', 'reviews', 'overall_satisfaction', 'accommodates', 'bedrooms', 'bathrooms', 'price', 'minstay', 'last_modified', 'latitude', 'longitude', 'location'].

The first dataset included more specific data about the accommodation. However, since we thought the most important feature about the airbnb accommodation is the location, we selected the second dataset.

2. Preprocessing

Among the Seoul Airbnb datasets, we used the largest and most recent dataset, composed of data collected in July, 2019.

- Excluded features

'room_id', 'survey_id', 'host_id': There were three id inputs in the dataset, which do not affect the price.

'country', 'city': The dataset we've decided to use is limited to Airbnb data in Seoul, so we concluded that 'country' and 'city' data would not be necessary.

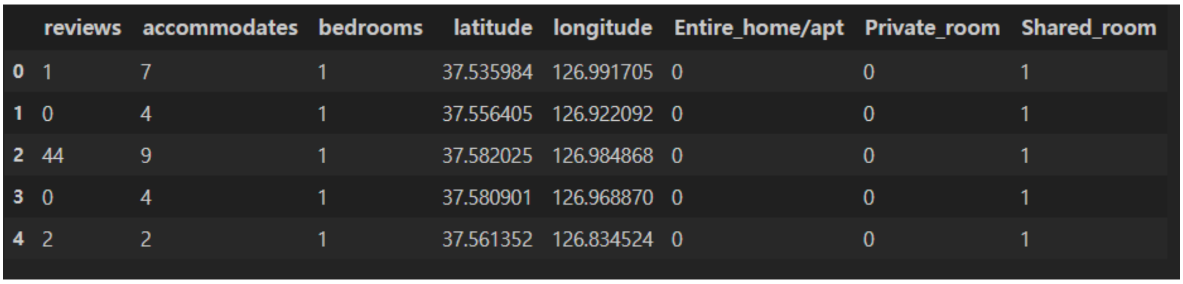
'borough', 'bathrooms, 'minstay': These features only have 'NaN' value.

'last_modified': This includes the date and time that the values were read from the Airbnb website.

'overall_satisfaction': Majority of the data were 0.0, and the rest were mostly 4.5 or 5.0. The gap between the instances is large, which could possibly lead the model to have wrong bias. Also, the difference in satisfaction did not affect the price much.

'location', 'neighborhood': These features are hard to process. 'Neighborhood' data is categorical, containing 438 categories. So encoding with one-hot encoder will have too many 0 values, leading to large error of the model. Also, 'latitude' and 'longitude' data is enough to indicate location data.

In conclusion, our features include 'room_type', 'reviews', 'accommodates', 'bedrooms', 'latitude', 'longitude' data.



	reviews	accommodates	bedrooms	latitude	longitude	Entire_home/apt	Private_room	Shared_room
0	1	7	1	37.535984	126.991705	0	0	1
1	0	4	1	37.556405	126.922092	0	0	1
2	44	9	1	37.582025	126.984868	0	0	1
3	0	4	1	37.580901	126.968870	0	0	1
4	2	2	1	37.561352	126.834524	0	0	1

'Room_type' data has categorical value. Using unique(), we can check three types of room types ('Shared room', 'Entire home/apt', 'Private room'). We encoded this column with one-hot encoder, and changed to three features ('Shared_room', 'Entire_home/apt', 'Private_room').

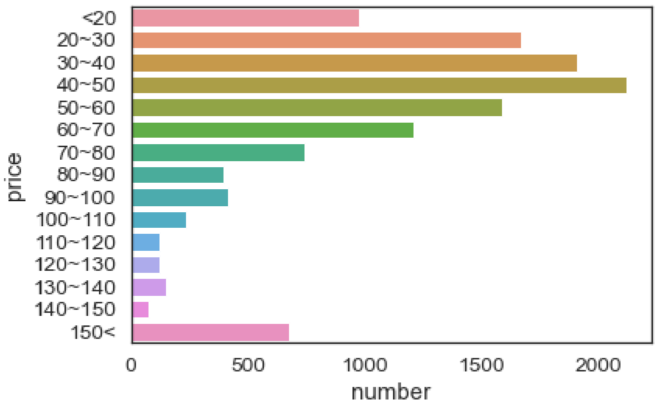
Although only a few instances had reviews, this feature can be important in some room types, such as 'shared room'. Therefore we did not eliminate them.

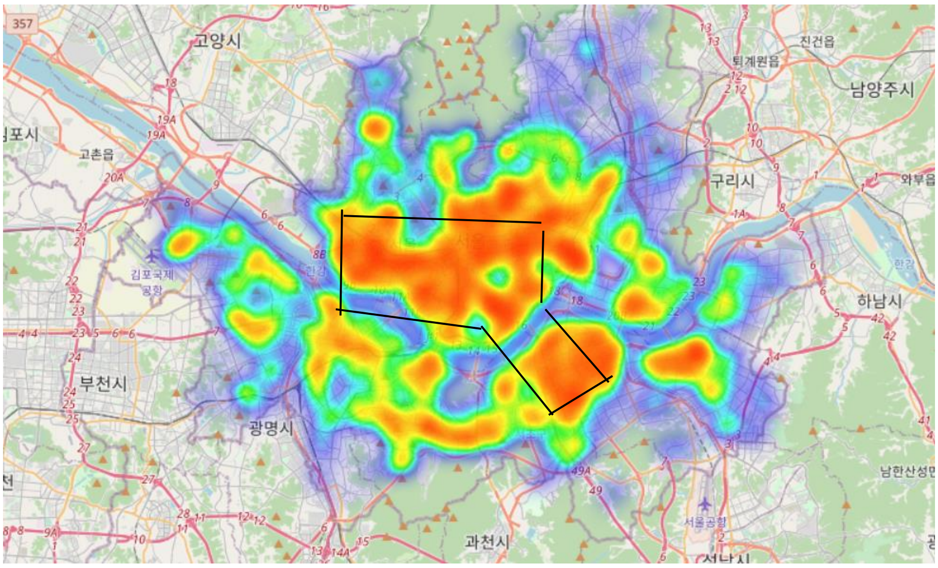
The numeric values will be scaled after the visualization step.

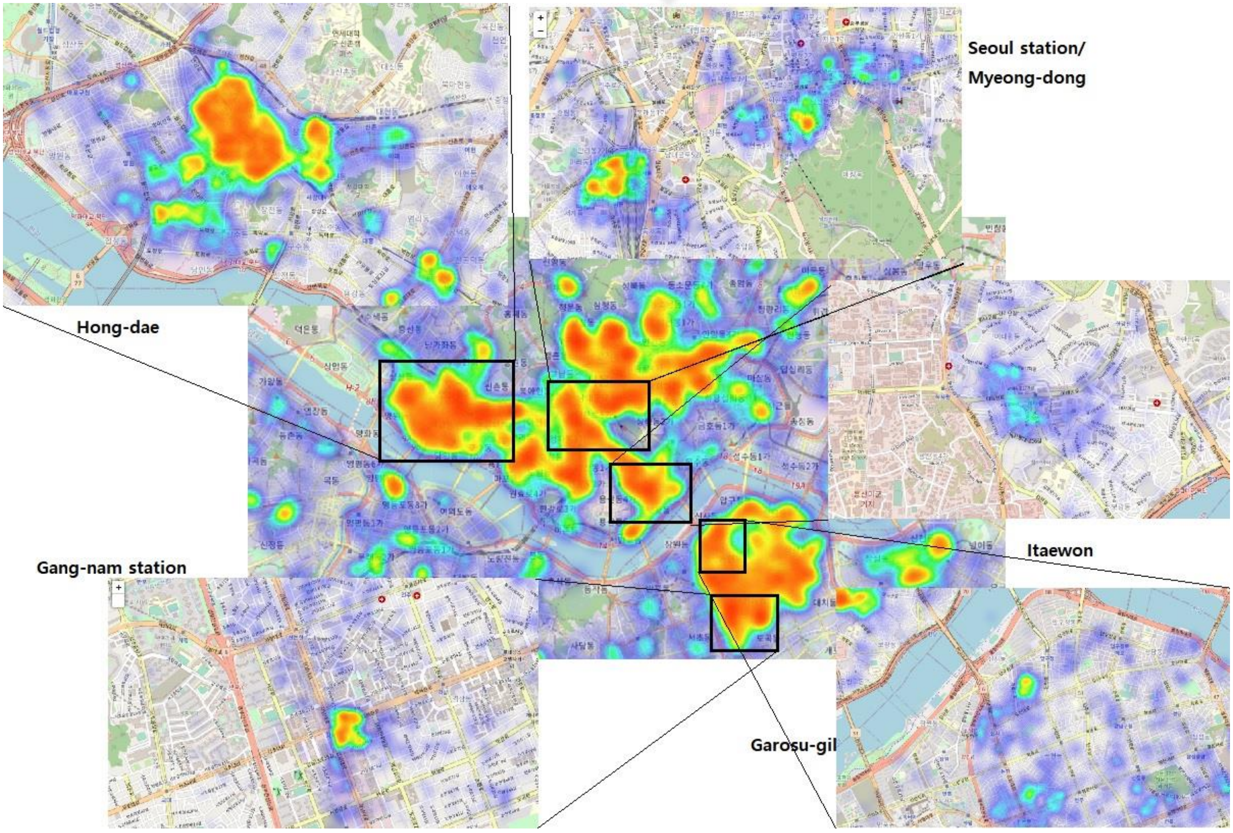
3. Visualization

The mean price was about 62.5. But as you can see on the right, most of the prices are between 20~70. So, we assumed that the price of over 150 is much higher than we expected. Therefore, we have to check the differences between group1(less150) and group2(over150.)

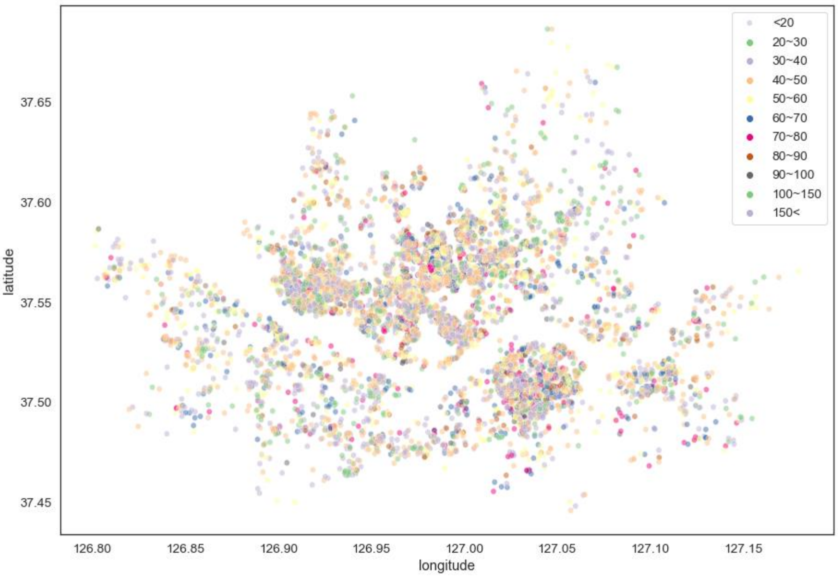
The overall number of airbnb are concentrated in some distinct places, including Gangnam, Hong-dae, Itaewon, etc.



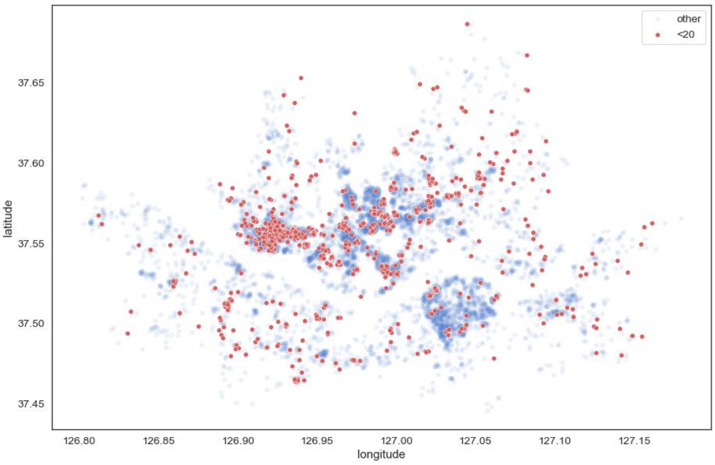




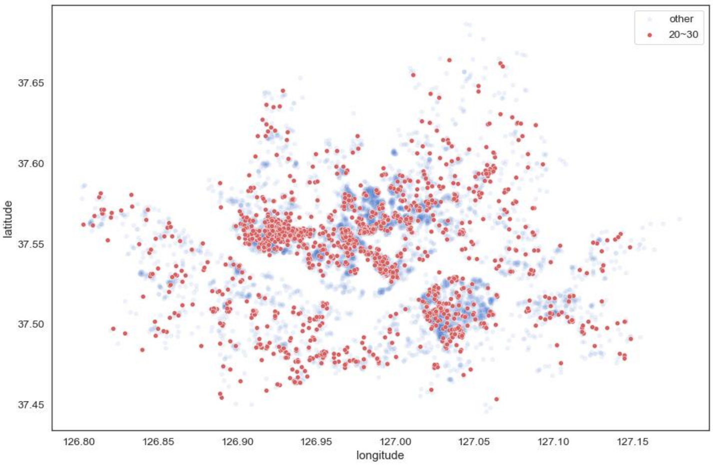
Overall population of airbnb according to prices. As the price goes higher, they are more concentrated in such famous places.



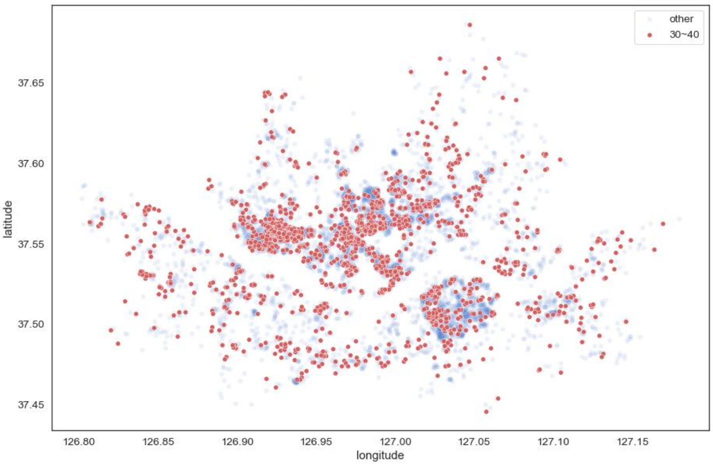
under 20



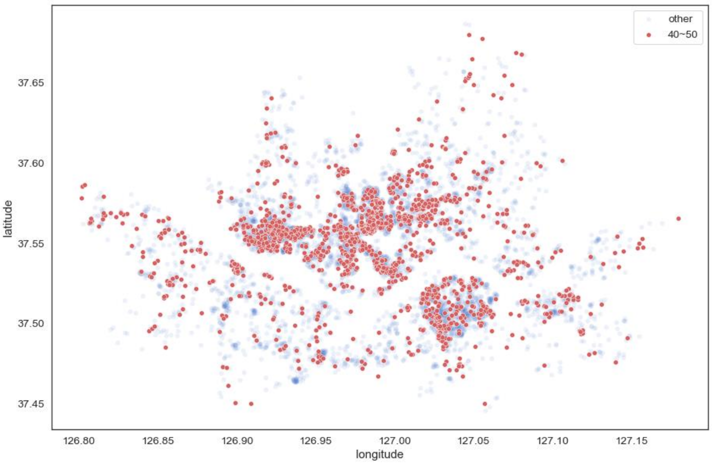
20~30



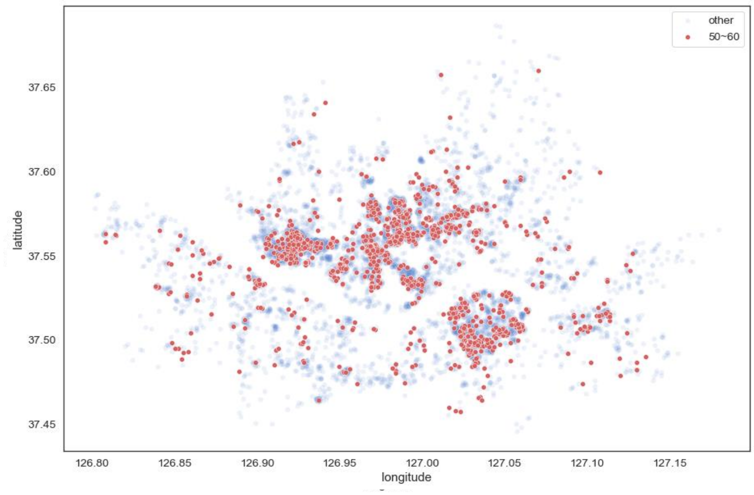
30~40



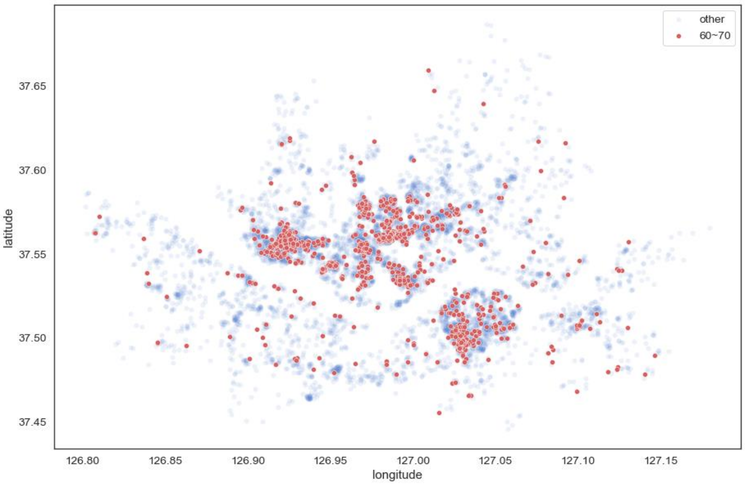
40~50



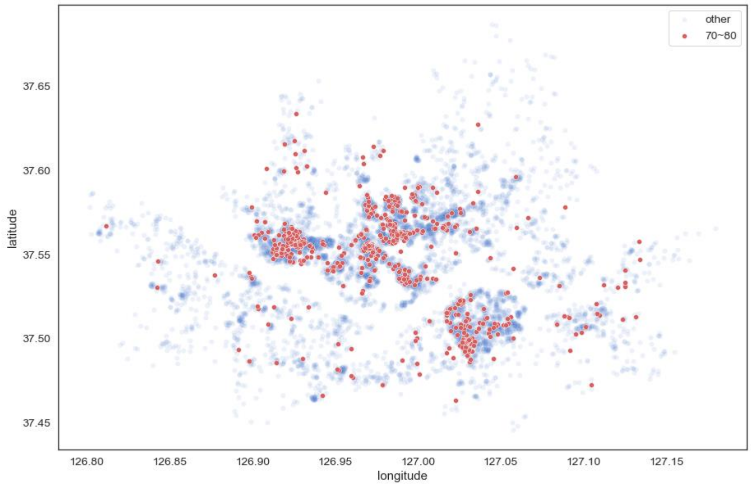
50~60



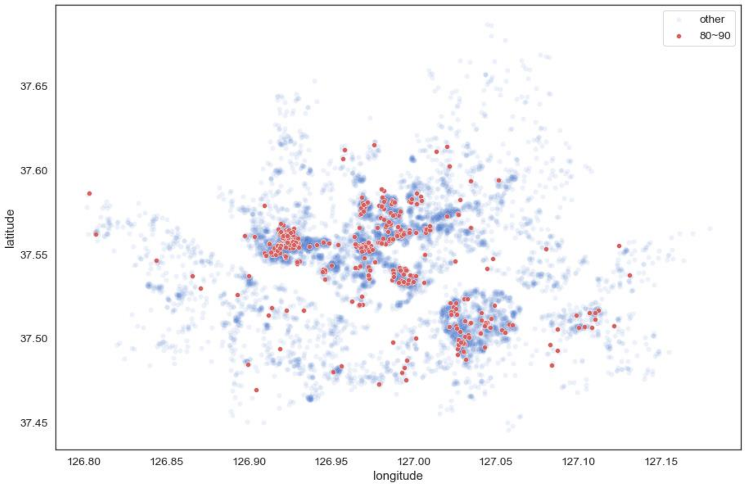
60~70



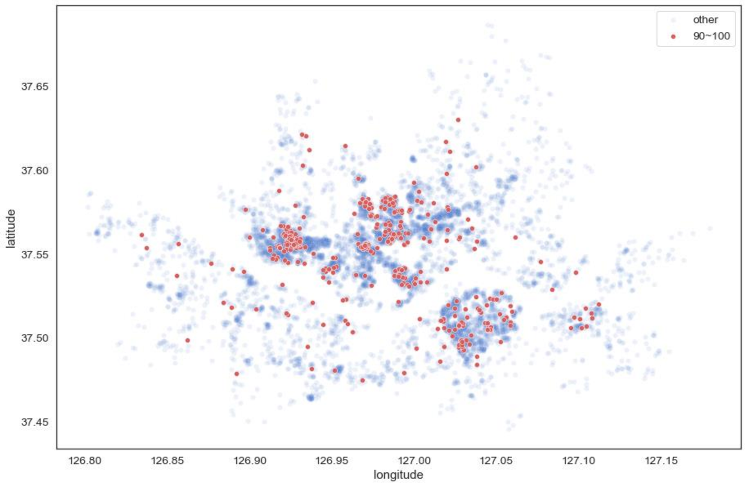
70~80



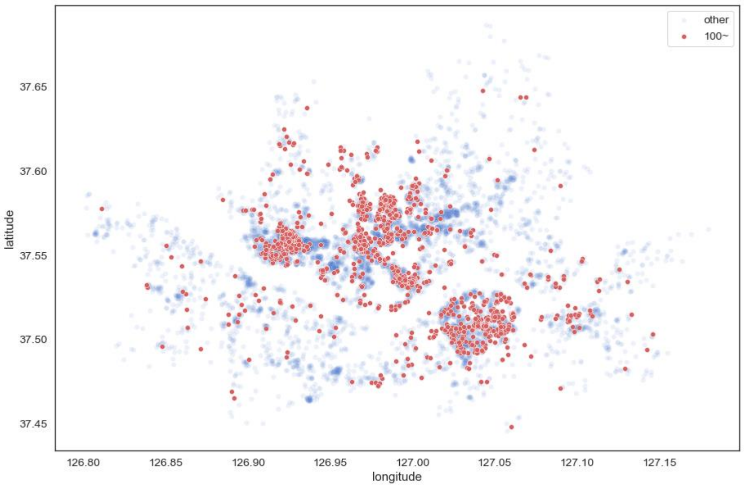
80~90



90~100

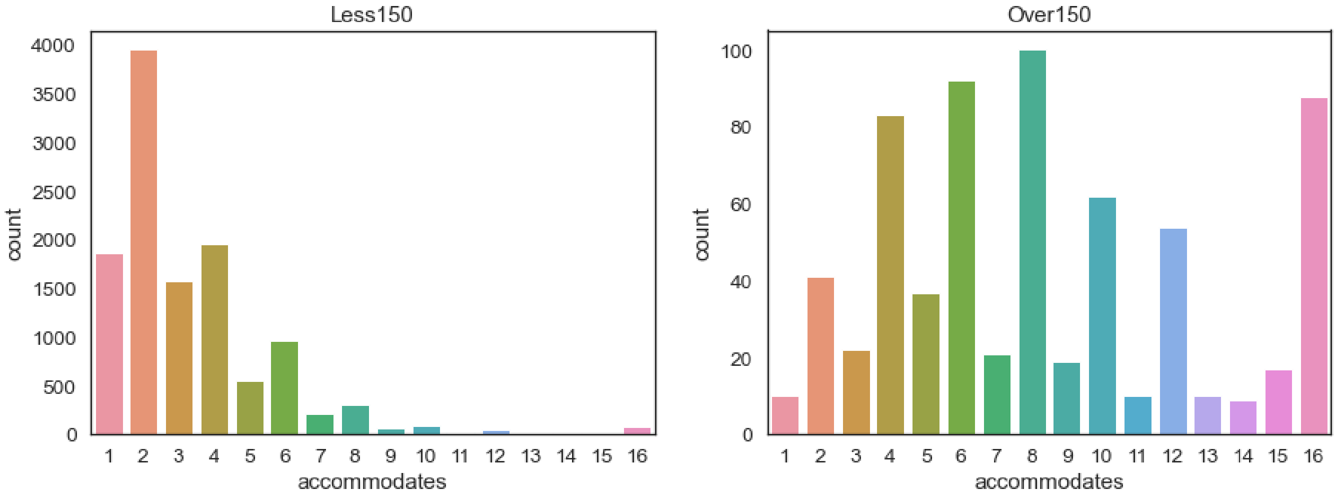


over 100



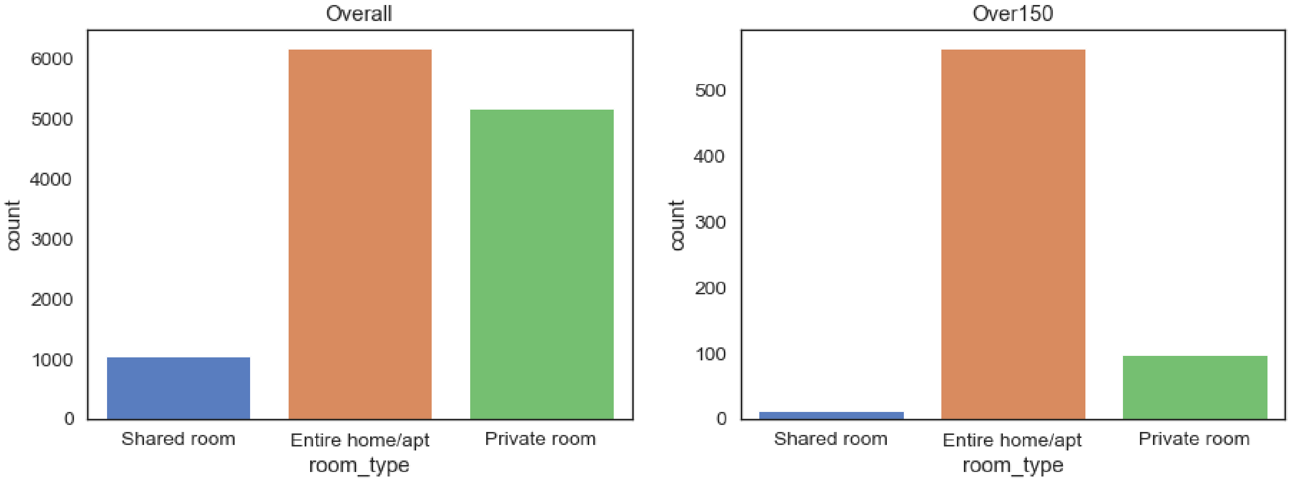
- Gap between less150 vs over150

a) Accommodates

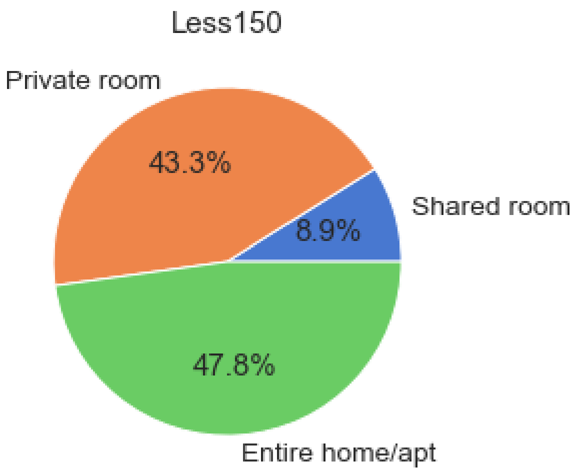


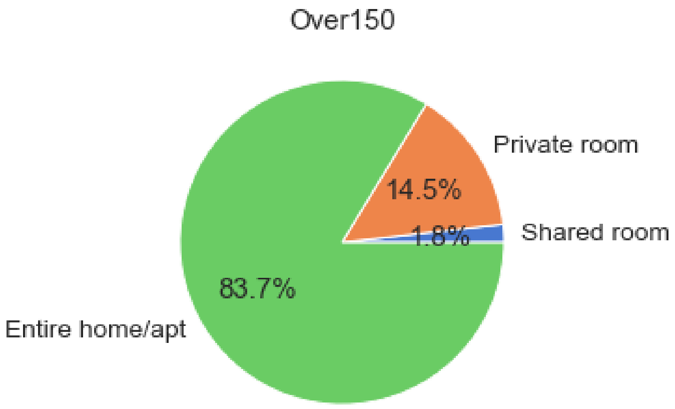
The number of accommodations between Airbnb listings listed for more than $150 and others shows huge differences. As the prices go higher, they tend to have more and better accommodations, which makes sense.

b) Room Type

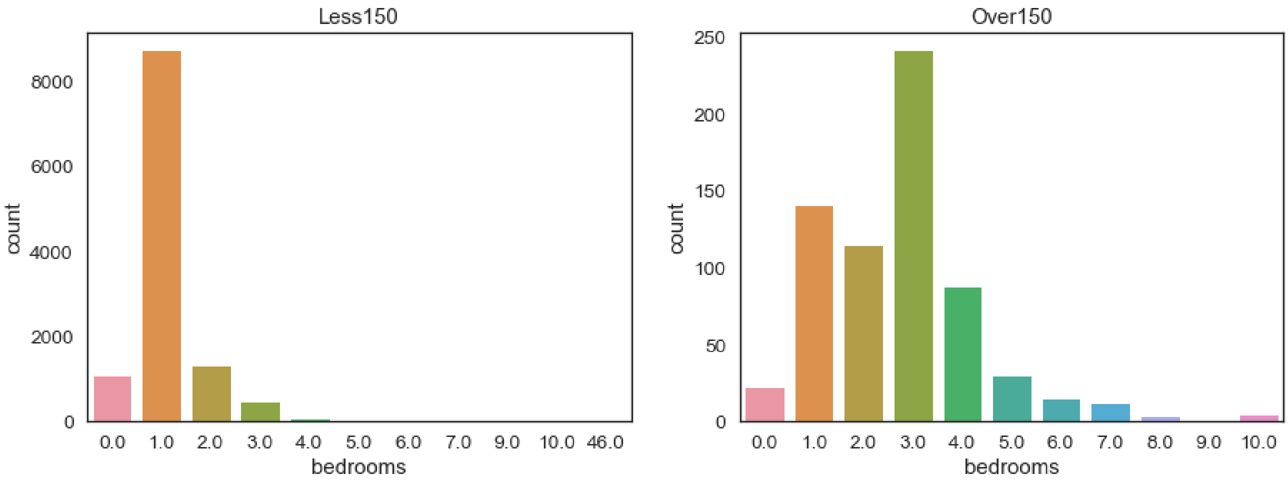


Room types don’t seem obviously different, but if converted to ratio rather than number, the difference between them looks distinct. The ratio of Entire home/apt is much higher in the over150 group, while the ratio of private room/shared room significantly decreased.





c) Bedrooms



The number of bedrooms also differs significantly based on the listing price. While the number of bedrooms are only one in the less150 group, lots of airbnb in the over150 group had 3 bedrooms. The mean value of bedrooms of over150 group is 2.75.

d) Reviews

For both of them, there were only a limited number of reviews. So we think that the number of reviews doesn't count that much of the price of airbnb than accommodations, room types, and bedrooms.

4. Building Models

Room types are what we consider first and the most when we look for a room on Airbnb. Room types don't only affect room types themselves; it affects everything besides room types as well.

Facilities and the extent to which guests are allowed to use them differ greatly based on the room type. Having access to the kitchen and a certain number of bathrooms in a shared-property setting, for instance, does not necessarily mean that guests will always have access to such facilities; for shared properties, namely private room and shared room listings, the availability of facilities may differ based on other guests. When one guest is using the bathroom or the kitchen in a shared property, then other guests cannot, whereas in an "entire home" setting, guests are guaranteed access to all of the facilities 100% of the time. This is the main reason that we've decided to differentiate price listing based on property type first. Also, the criteria we use are different for different room types. For example, when we search for a shared room, such as a hostel, the reviews are more important than other room types. Therefore, we decided to build models separately according to room types. Each of our members take one room type(Nahyun:Entire home/apt, Jangwoo:Private room, Jundong:Shared room), and will make regression models predicting prices accordingly.