Question 5

Georgios Pediaditis

3/5/2021

Question 5

Part 1

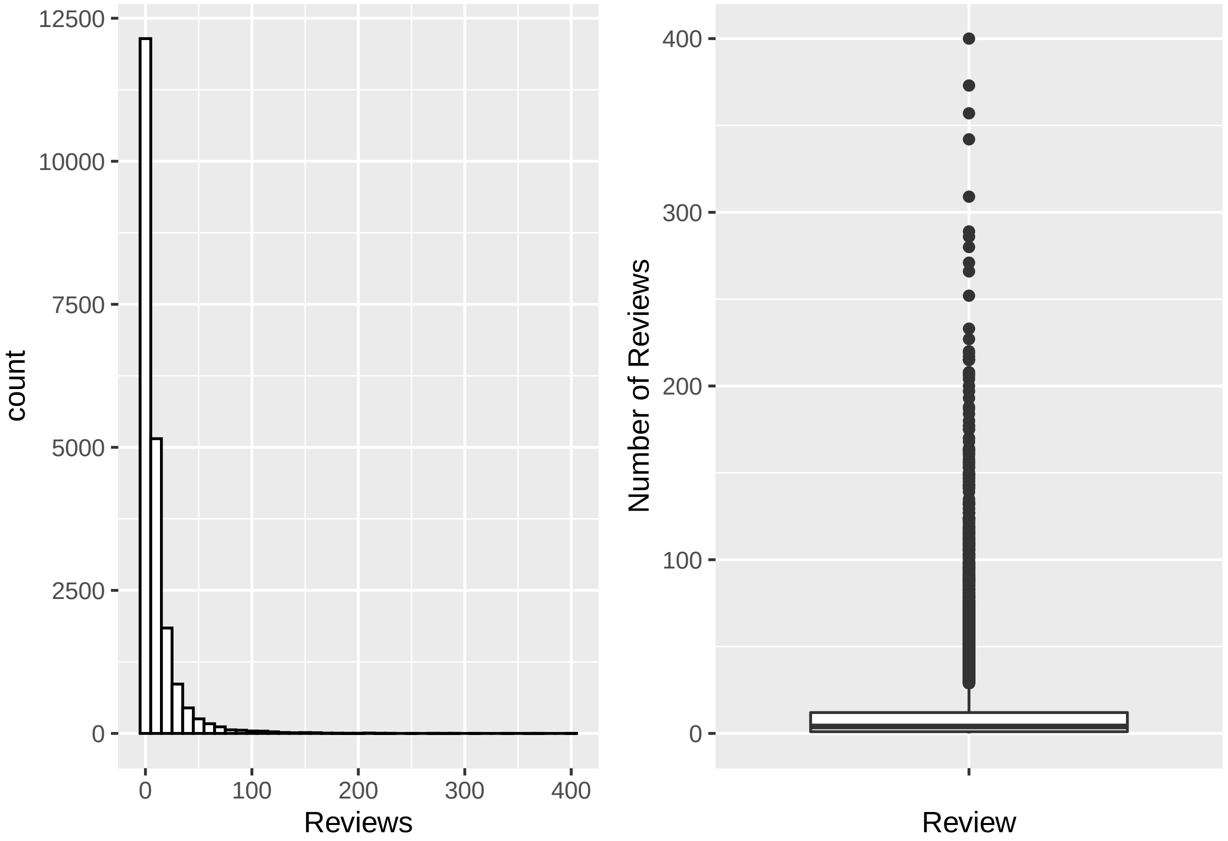
1.a

Selected cities are Copenhagen and London.

To begin with we start with summarizing the Review data.

Copenhagen

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    4.00   10.26   12.00  400.00
```



We could choose a data transformation with log or sqrt and a larger binwidth if we want to alter the visualization of our data although I find the current visualization sufficient.

From the above we can claim that.

Location

Mean is a lot bigger than Median. Since we know that Mean is greatly affected by outliers we can suspect we have a lot of Reviews for some apartments. Another conclusion that can be drawn from summary statistics is that the distribution of our data isn't symmetric.

Spread

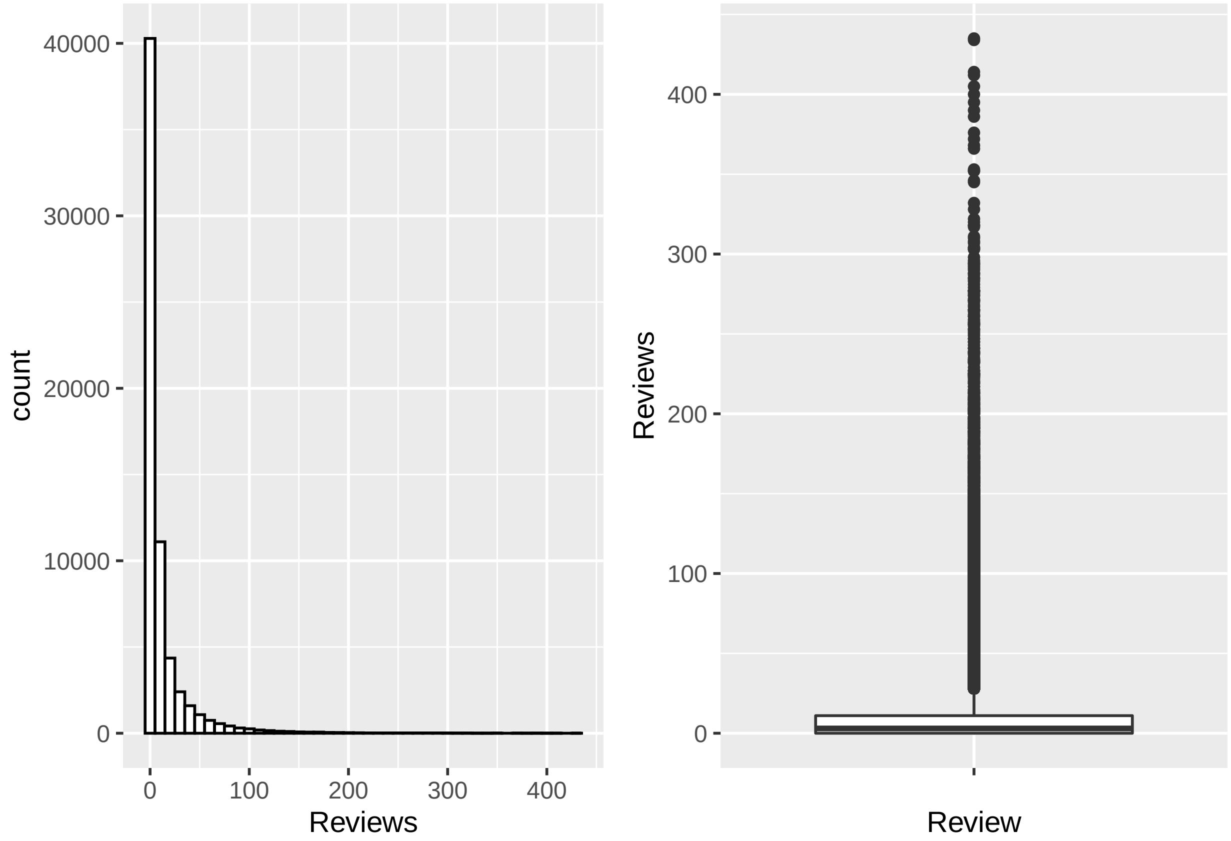
The most reliable measure of spread not effected by outliers is IQR. So we have IQR=12-1=11

Shape

Since median is much closer to lower quartile compared to upper quartile we can assume that our data are right skewed.

London

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0       0       3      12      11     435
```



Although the notes suggest to use a dotplot for discrete data and a histogram for continuous data I found using histogram more appropriate. The large range and number of data gives a visualization more close to continuous data than discrete. We could choose a data transformation with log or sqrt and a larger binwidth if we want to alter the visualization of our data although I find the current visualization sufficient.

From the above we can claim that.

Location

Mean is a lot bigger than Median. Since we know that Mean is greatly affected by outliers we can suspect we have a lot of Reviews for some apartments. Another conclusion that can be drawn from summary statistics is that the distribution of our data isn't symmetric since Mean is a lot bigger than Median.

Spread

The most reliable measure of spread not effected by outliers is IQR. So we have IQR=11-0=11

Shape

Since median is much closer to lower quartile compared to upper quartile we can assume that our data are right skewed.

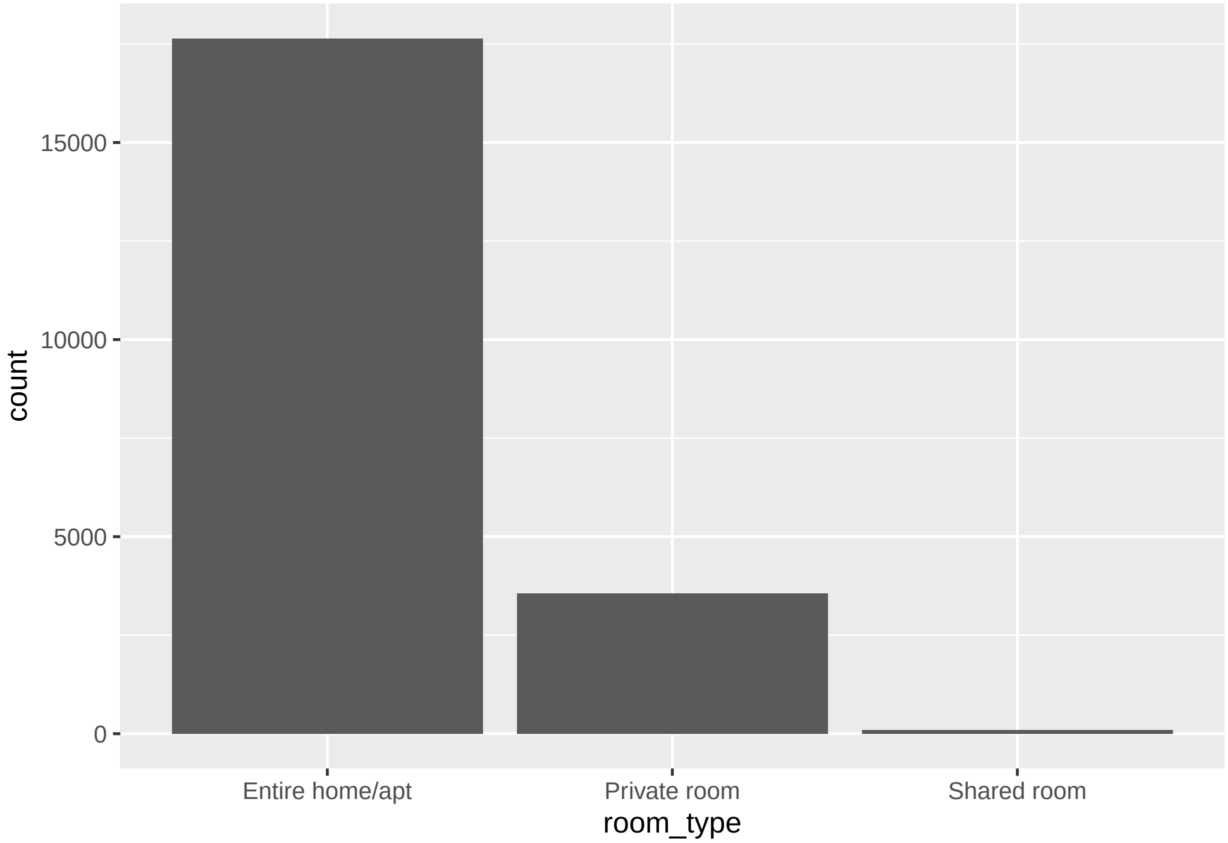
1.b

Summarise

Copenhagen

```
##
## Entire home/apt    Private room     Shared room
##           17640            3564              97
```

```
##
## Entire home/apt    Private room     Shared room
##     0.828130135     0.167316088     0.004553777
```

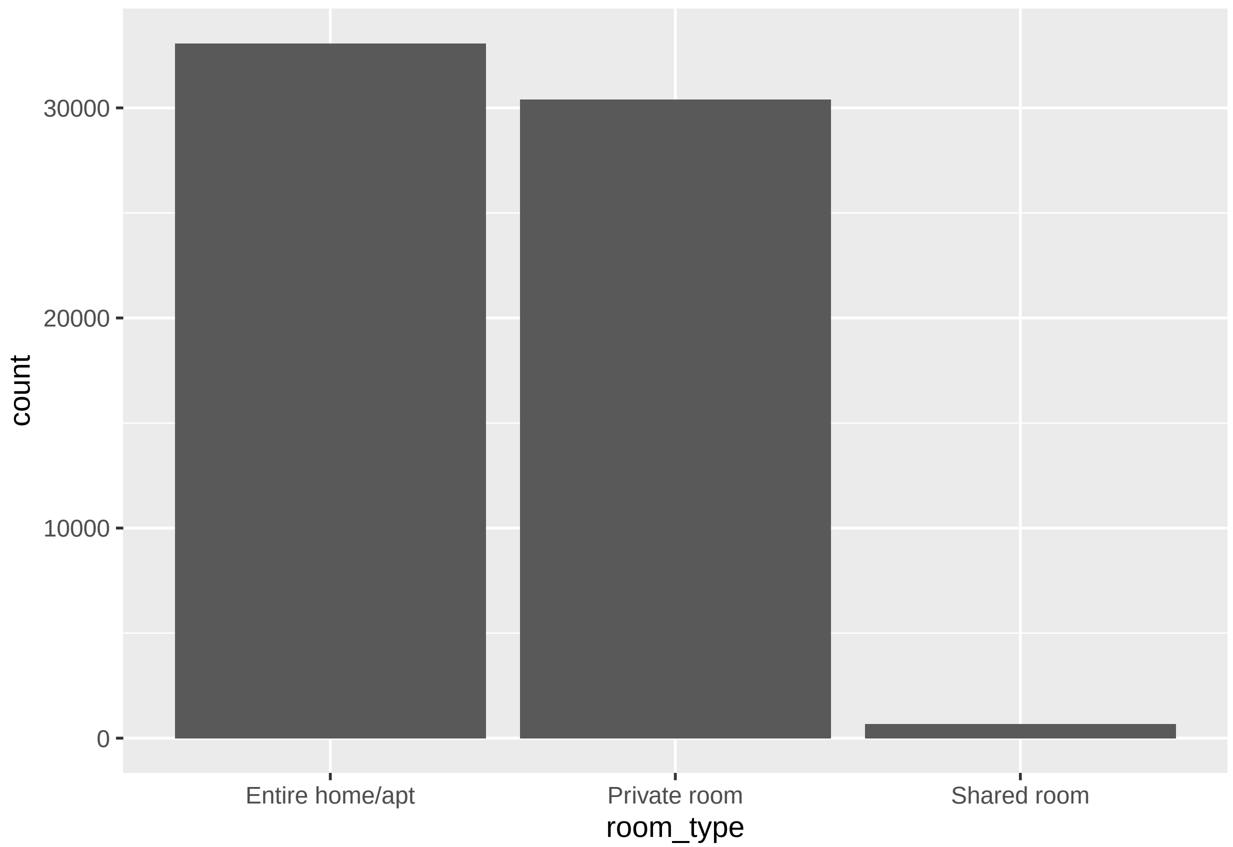


The above data shows that the vast majority of airbnb listings in Copenhagen consist mostly of entire homes. There is a small number of private rooms while shared rooms are almost non-existent.

London

```
##
## Entire home/apt    Private room     Shared room
##           33072           30399             673
```

```
##
## Entire home/apt    Private room     Shared room
##      0.51558992      0.47391806      0.01049202
```

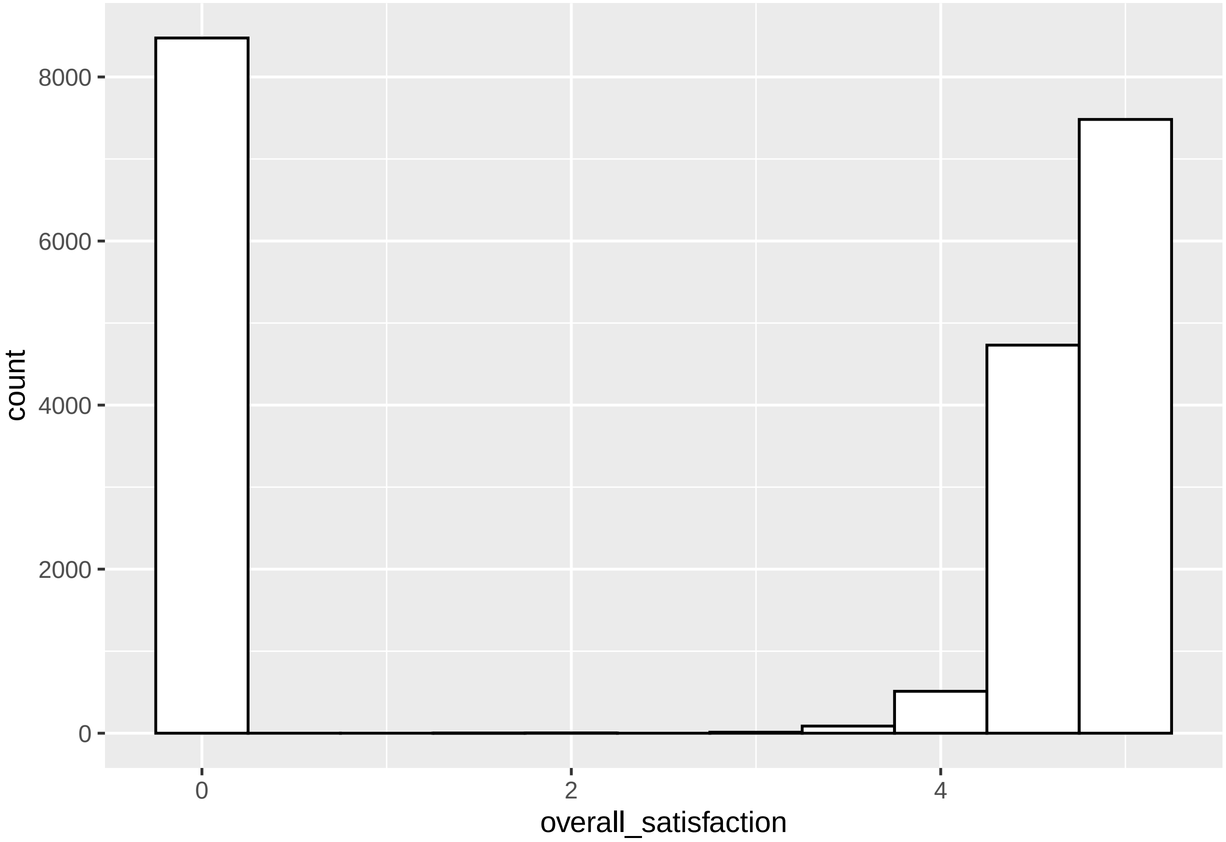


The above data shows that the number of entire home airbnb renting in London is almost equal to the number of private rooms. Share rooms are rare

1.c

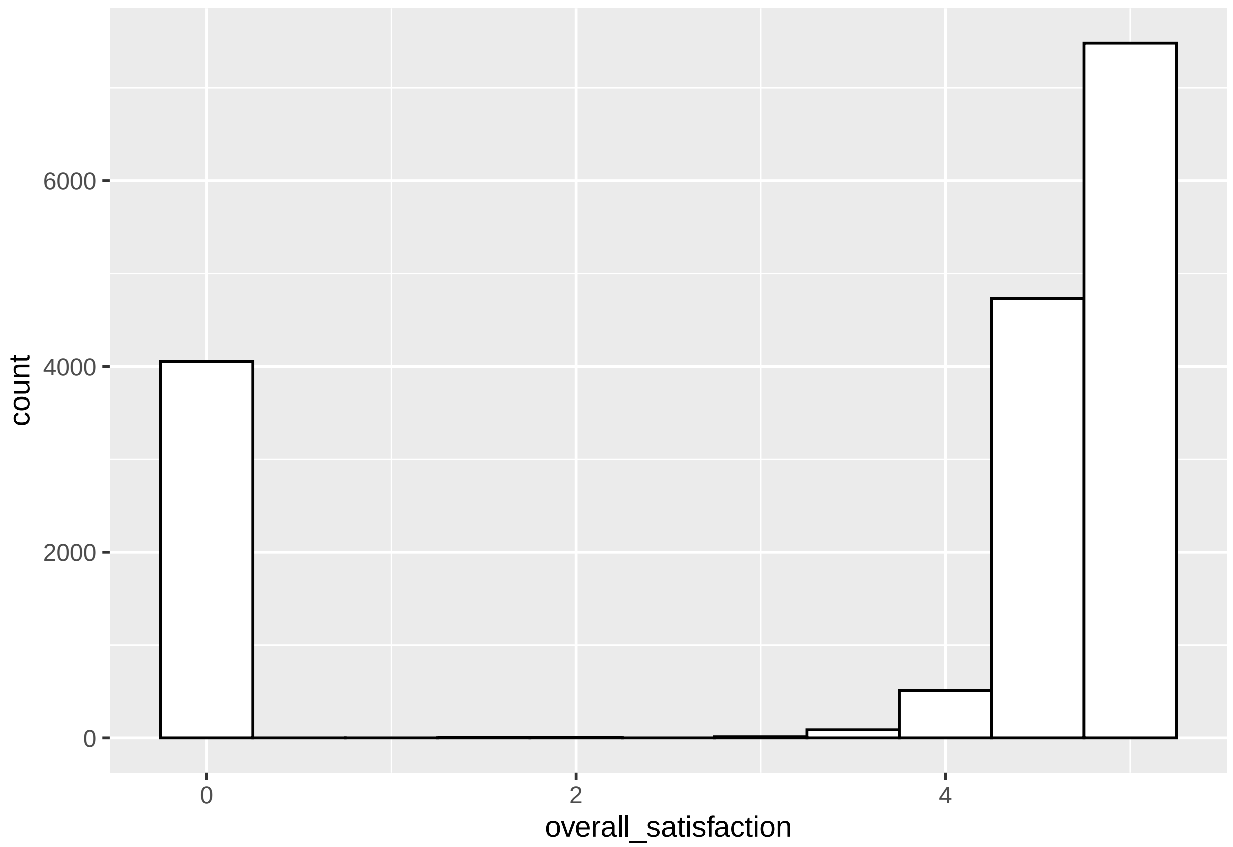
Plot

Copenhagen



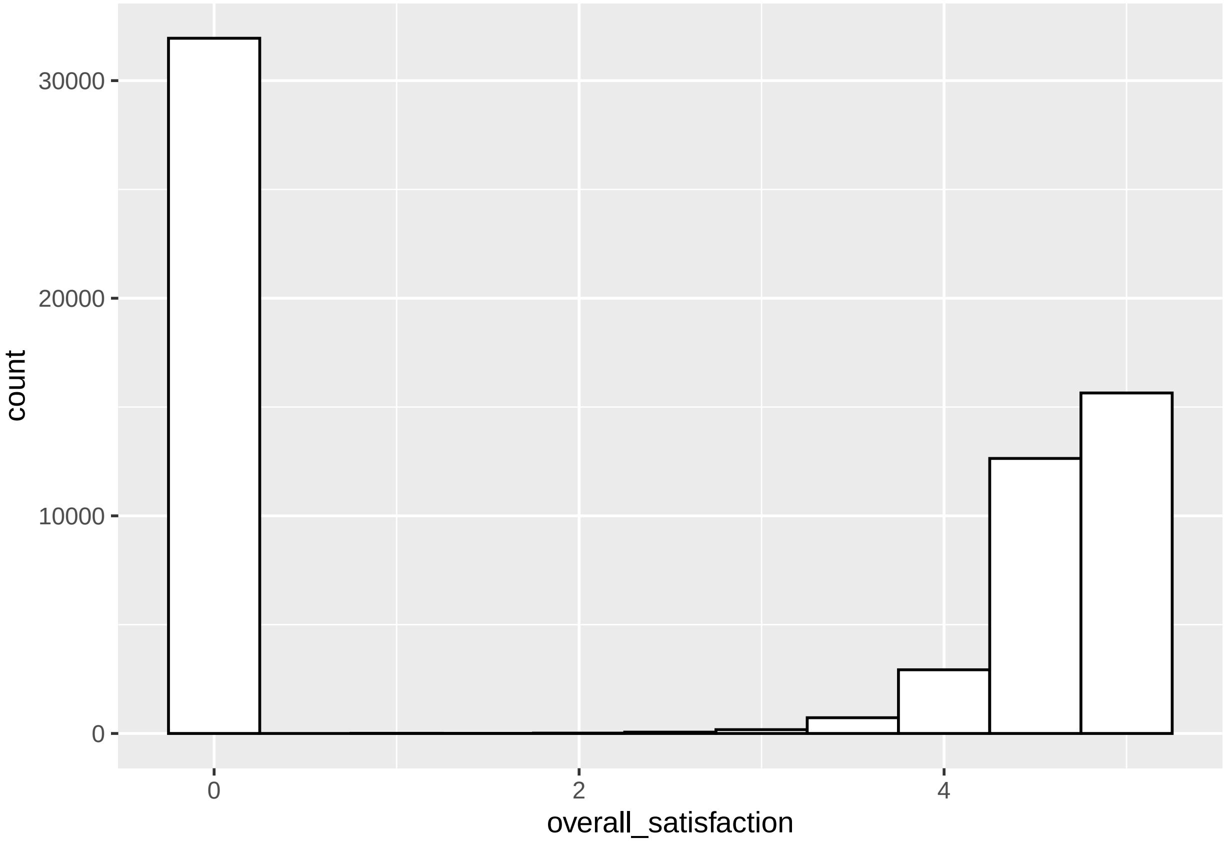
On the plot above we see a big number of 0 rating in our data. The rest of our data are left skewed. We observe from the data that there are a lot of apartments with 0 rating and 0 reviews. That means that it might be useful to ignore data that have 0 reviews.

That gives us the following plot.



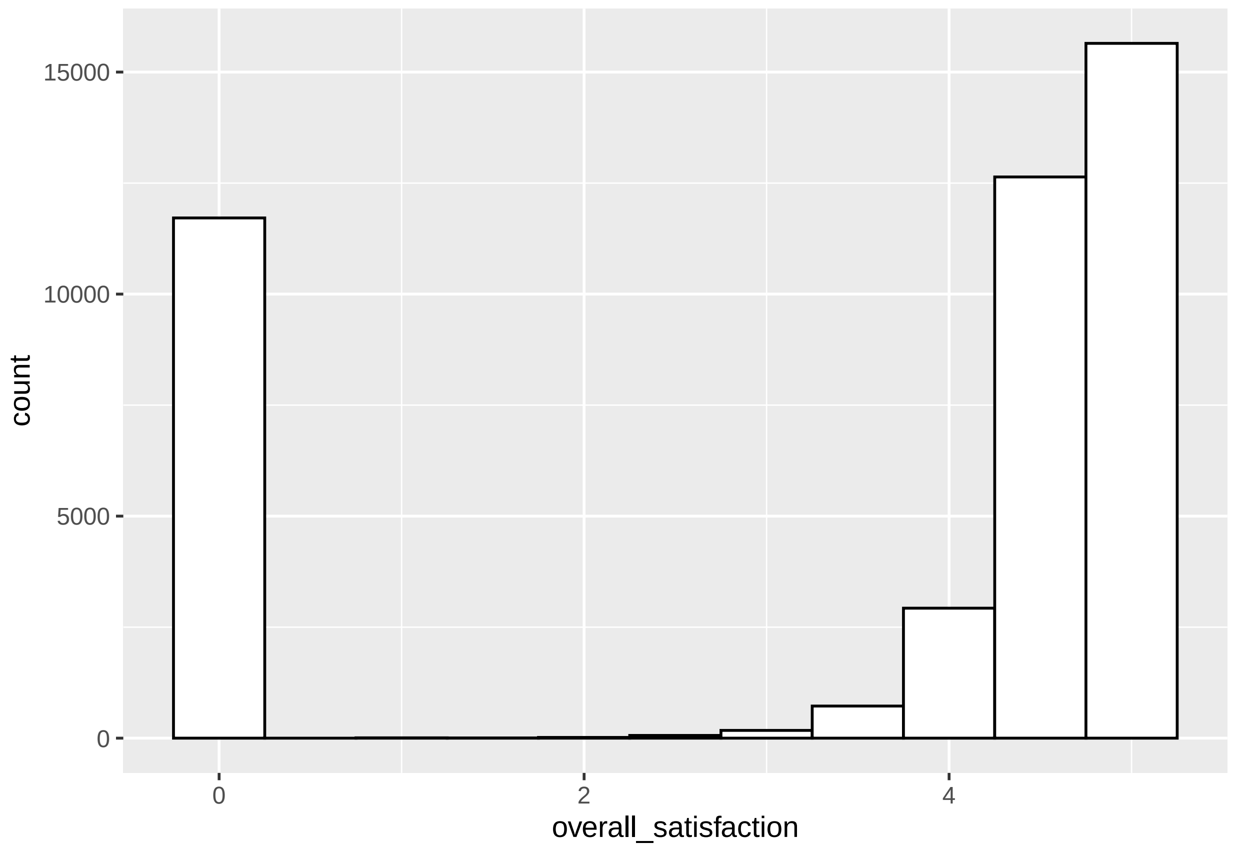
As we can see excluding data with 0 reviews reduce 0 ratings by half. The median overall satisfaction is 4.5 before and after excluding rows with 0 reviews

London



On the plot above we see a huge number of 0 ratings in our data. The rest of our data are left skewed. As we observed in Copenhagen data there are a lot of apartments with 0 rating and 0 reviews. That means that it might be useful to ignore data that have 0 reviews.

That gives us the following plot.



The new plot cuts by half 0 ratings for overall satisfaction. The new median is now 4.5 compared to the previous median 3

Part 2

Comparing the Reviews

- When we compare the reviews between Copenhagen and London we see a similar pattern. Both cities share a similar average number of reviews. The average number of reviews is 4 in Copenhagen while in London is slighty smaller at 3. Mean on both cities are quite larger than Median although the difference in Copenhagen is a bit smaller compared to London. Since upper outliers is based on (3/4)(n+1) that means that outliers in London tend to have more reviews compare to Copenhagen.
- Both cities show a similar spread. (IQR=11)
- Both cities show a similar shape. They are highly right skewed. That means that the majority of listings receive a small number of reviews.

Comparing room type.

- The main difference we see between Copenhagen and London is that in Copenhagen the vast majority of listings is Entire home/apt (~82%) followed by Private rooms (~16%) while in London the Entire home/apt and Private room listing are roughly equal with 51% and 47% respectively. Shared room listings on both cities are rare.

Comparing overall satisfaction

- Since in our data we notice that there are a lot of listings that lack reviews and their overall satisfaction is 0 we assume that a more representative presentation of our data would exclude these data.
- Excluding these data doesn't change the big picture. We notice that our ratings tend to accumulate on both sides of our spectrum. There could be various reasons for that. People choose to rate really good or bad experiences, it could just mean we have a scale that isn't representative of the way people perceive a situation, or that we are missing data due to error.
- Finally we notice that both cities have a median 4.5 and offer equally high overall satisfaction.