AIRBNB DATA MINING

AN EXPLORATION OF RATINGS OF AIRBNB IN PARIS

RAN TAO

Why study the ratings of Airbnb in Paris?

- Nowadays, more and more travelers are choosing Airbnb over hotel because of its low cost, convenient location and household amenities especially in Europe.

- Paris, one of the most popular touristic city in Europe, has highest number of Airbnbs.

- I want to study the ratings of Airbnbs in Paris to give better recommendation to people who are visiting Paris and also offer advice to Airbnbs host in Paris

What's the goal of this study?

1. Discover the factors that affect the ratings of Airbnbs in Paris

2. Discover the distribution of ratings between districts and within districts

What's the implication of this study?

For Airbnb hosts: have a better understanding of the preference of customers

For Airbnb guests: improve of their knowledge of choosing Airbnb in different districts

Methods

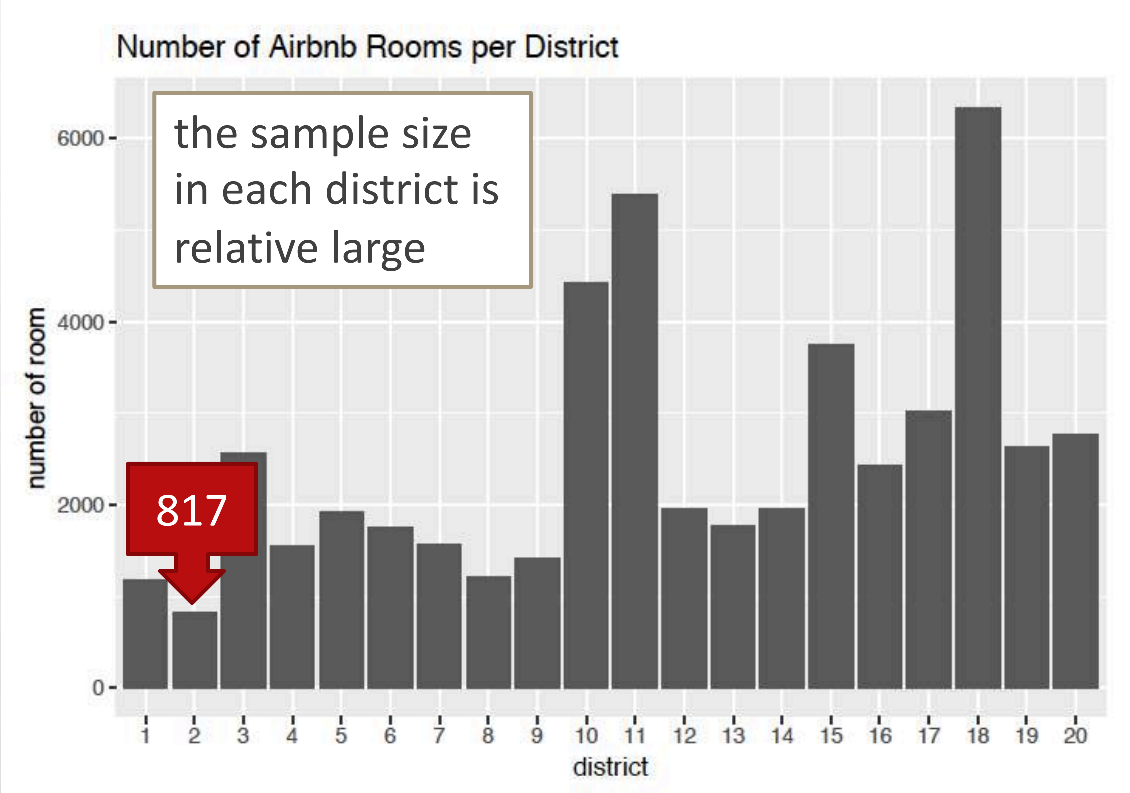
Data source: single survey for Paris with 70,158 listing properties as of 25th July 2017, which is collected from the public Airbnb website

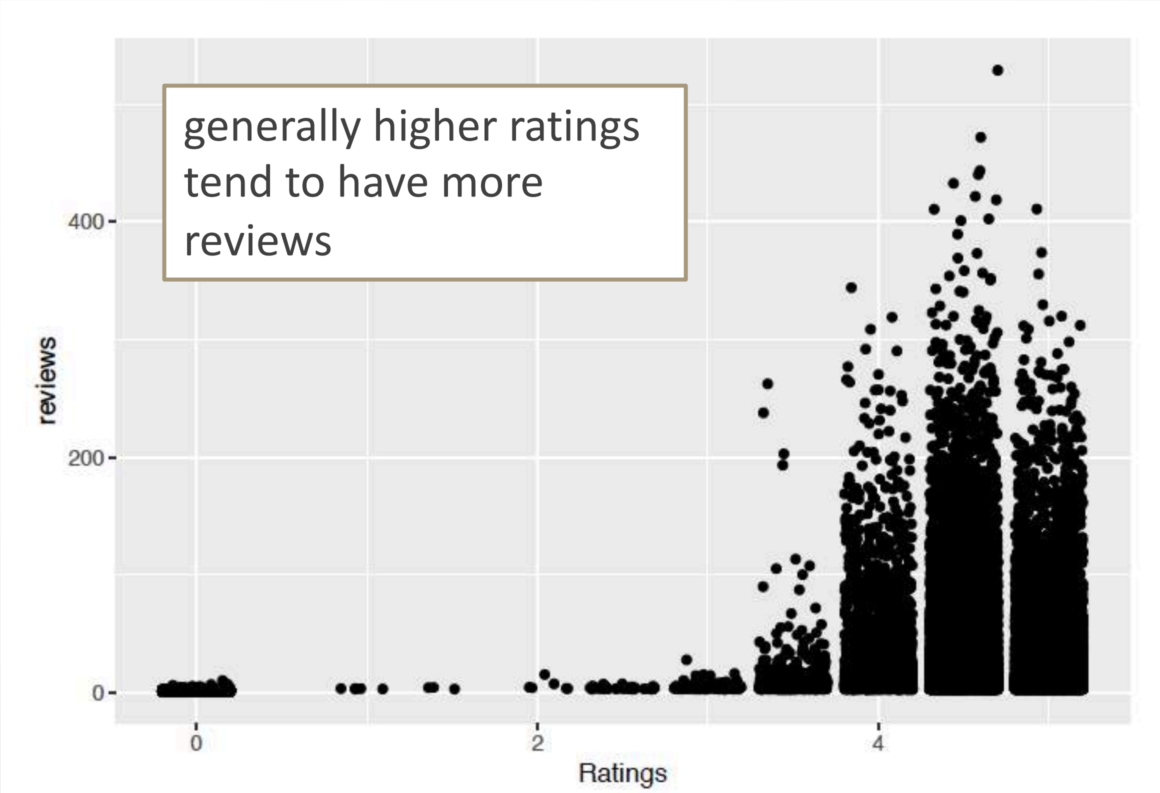
Study population: 50,406 listing Airbnb properties in Paris, we exclude the 19,752 listing properties which have no reviews in our dataset

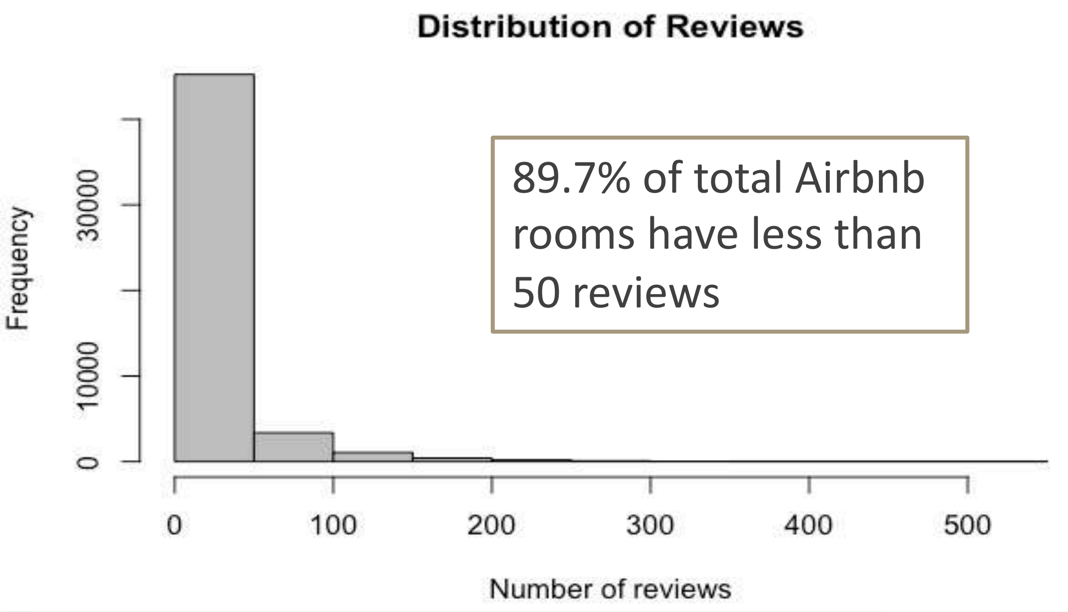
Outcome definition:
Satisfaction: YES (weighted average >=2); No (weighted average <2)
Weighted average ratings=average overall satisfaction * weight
Weight= sigmoid function (number of reviews)

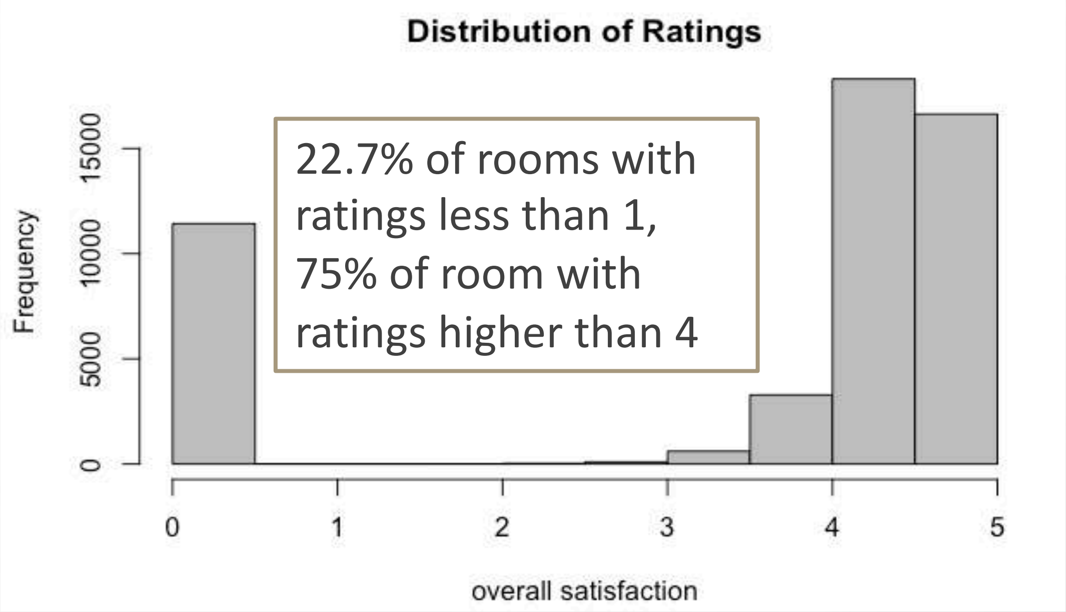
Room id	A unique number identifying an Airbnb listing
Host id	A unique number identifying an Airbnb host
Room type	One of Entire home/apt, Private room, or Shared room
Neighborhood	a sub-region of the city or search area for which the survey is carried out
Reviews	The number of reviews that a listing has received
Overall satisfaction	The average rating (out of five) that the listing has received from those visitors who left a review
Accommodates	The number of guests a listing can accommodate
Bedrooms	The number of bedrooms a listing offers
price	The price (in $US) for a night stay
Latitude and longitude	The latitude and longitude of the listing as posted on the Airbnb

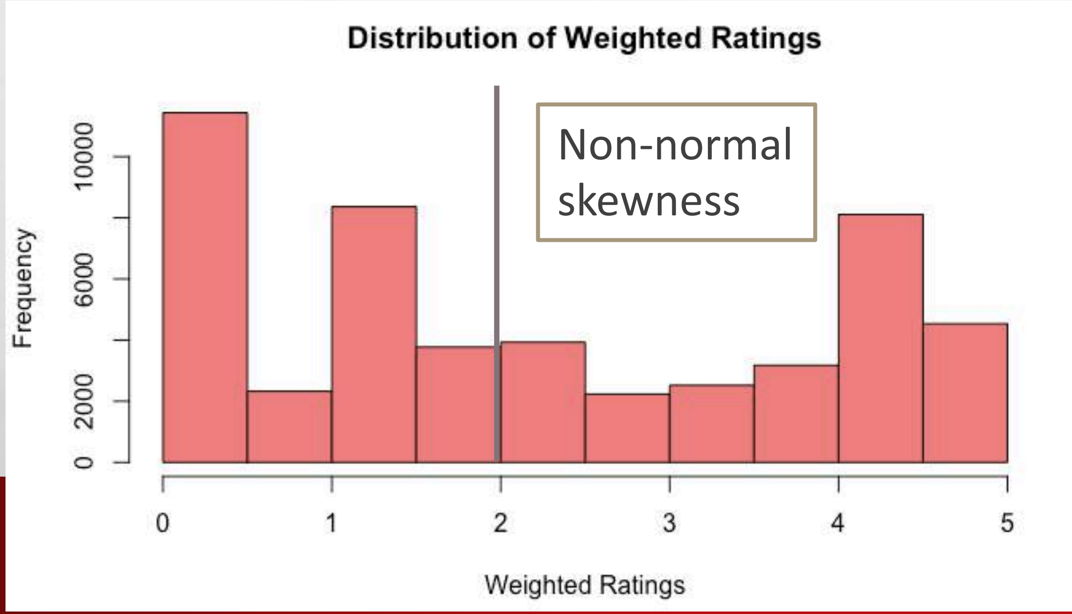
EDA

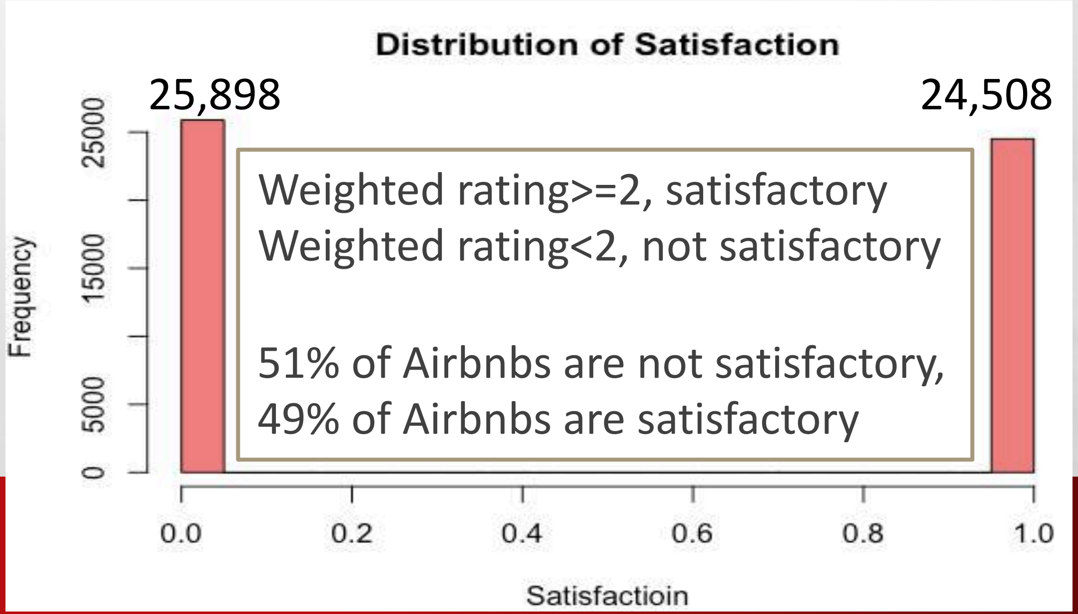

Number of Airbnb Rooms per District
the sample size in each district is relative large
817
number of room
6000
4000
2000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
district


generally higher ratings tend to have more reviews
reviews
400
200
0
0 2 4
Ratings

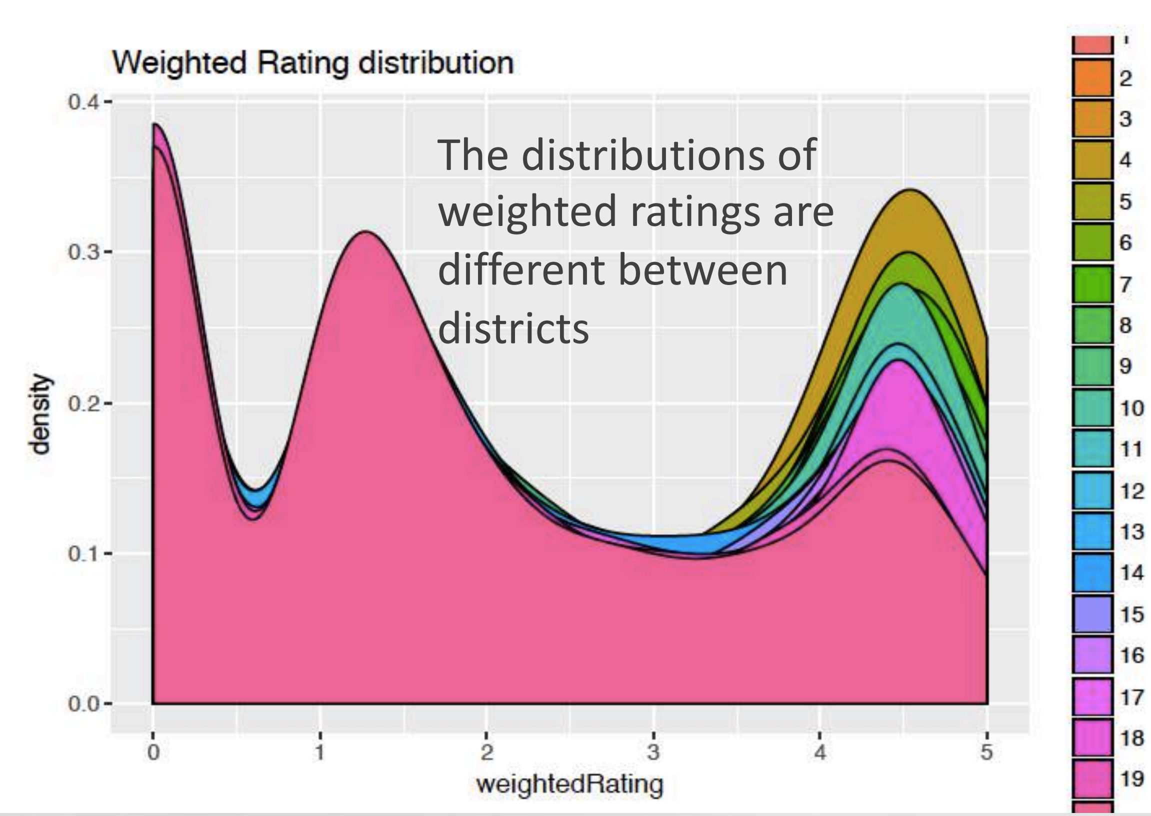

Distribution of Reviews
89.7% of total Airbnb rooms have less than 50 reviews
Frequency
0
10000
30000
0
100
200
300
400
500
Number of reviews


Distribution of Ratings
22.7% of rooms with ratings less than 1,
75% of room with ratings higher than 4
Frequency
0
5000
10000
15000
0
1
2
3
4
5
overall satisfaction

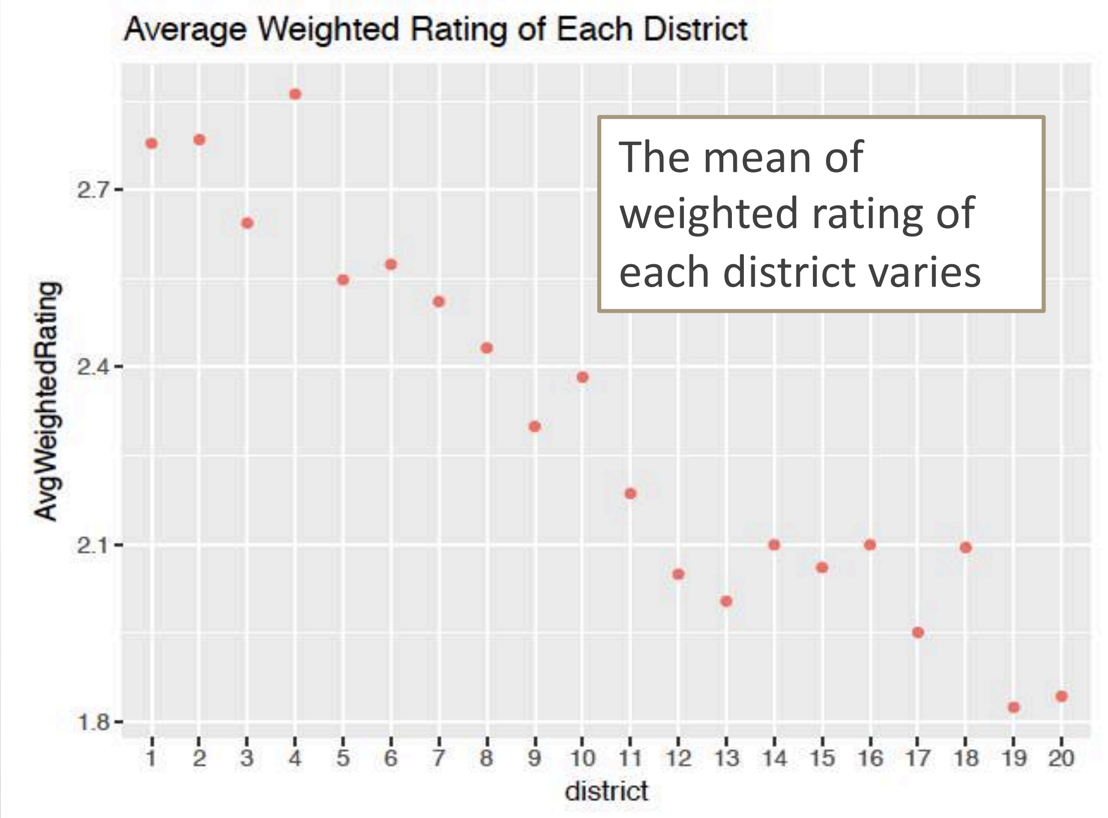

Distribution of Weighted Ratings
Non-normal skewness
Frequency
0
2000
6000
10000
0
1
2
3
4
5
Weighted Ratings


Distribution of Satisfaction
25,898
24,508
Weighted rating>=2, satisfactory
Weighted rating<2, not satisfactory
51% of Airbnbs are not satisfactory,
49% of Airbnbs are satisfactory
Frequency
0
5000
15000
25000
0.0
0.2
0.4
0.6
0.8
1.0
Satisfactioin

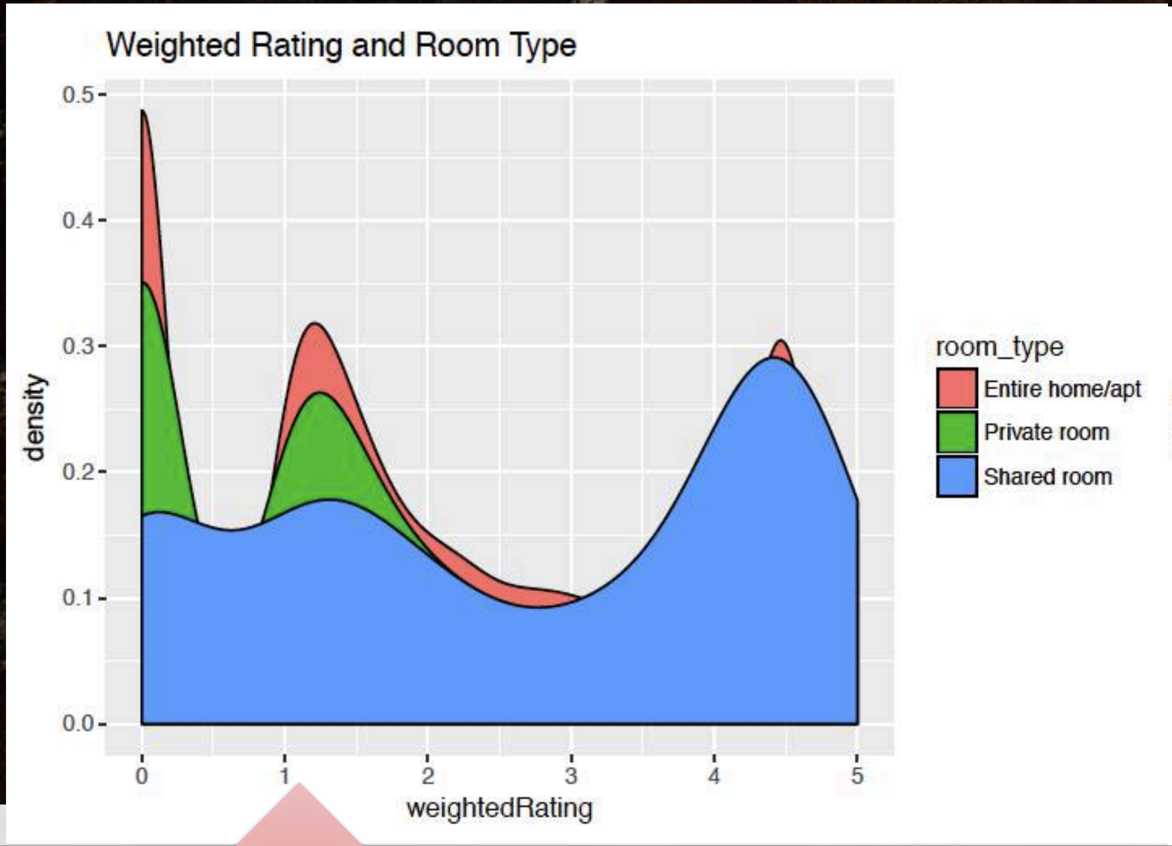
EDA

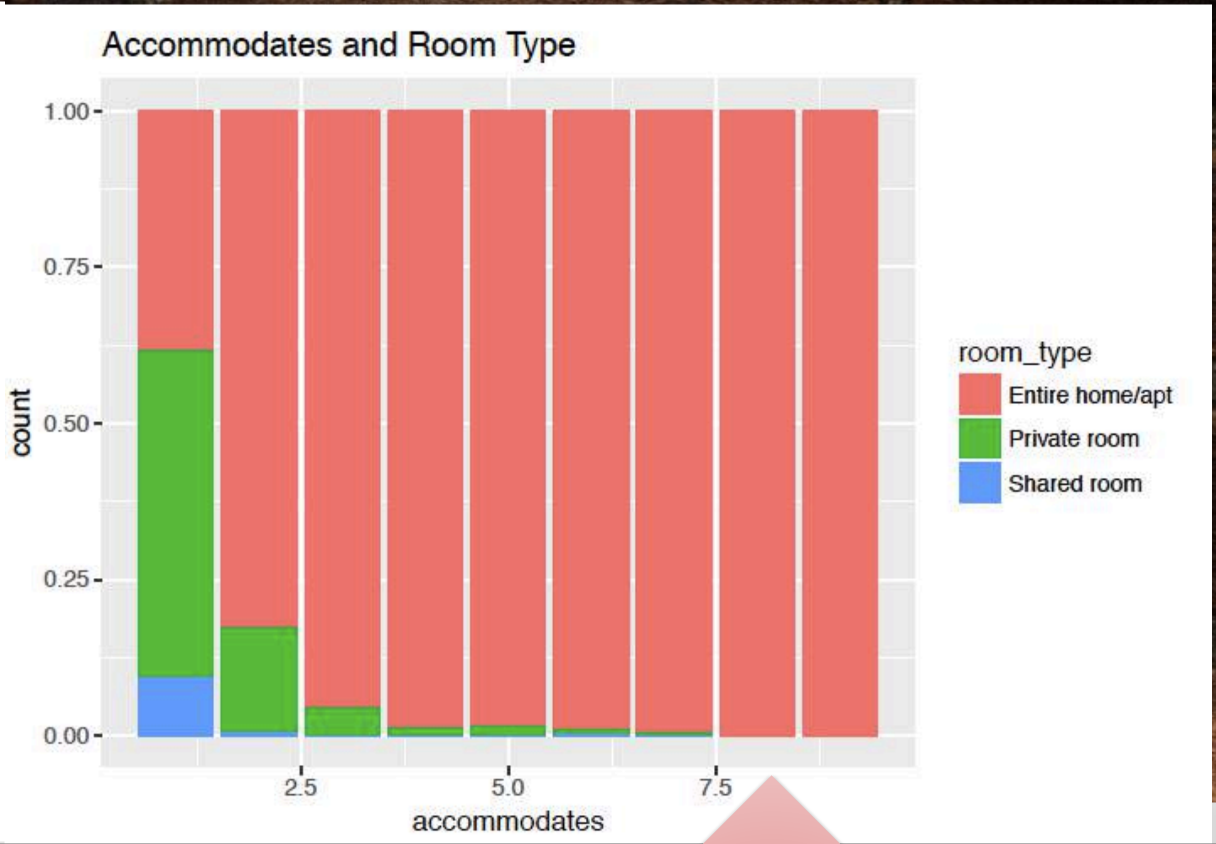


Weighted Rating distribution
The distributions of weighted ratings are different between districts
density
weightedRating
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19



Average Weighted Rating of Each District
The mean of weighted rating of each district varies
AvgWeightedRating
district
1.8
2.1
2.4
2.7
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

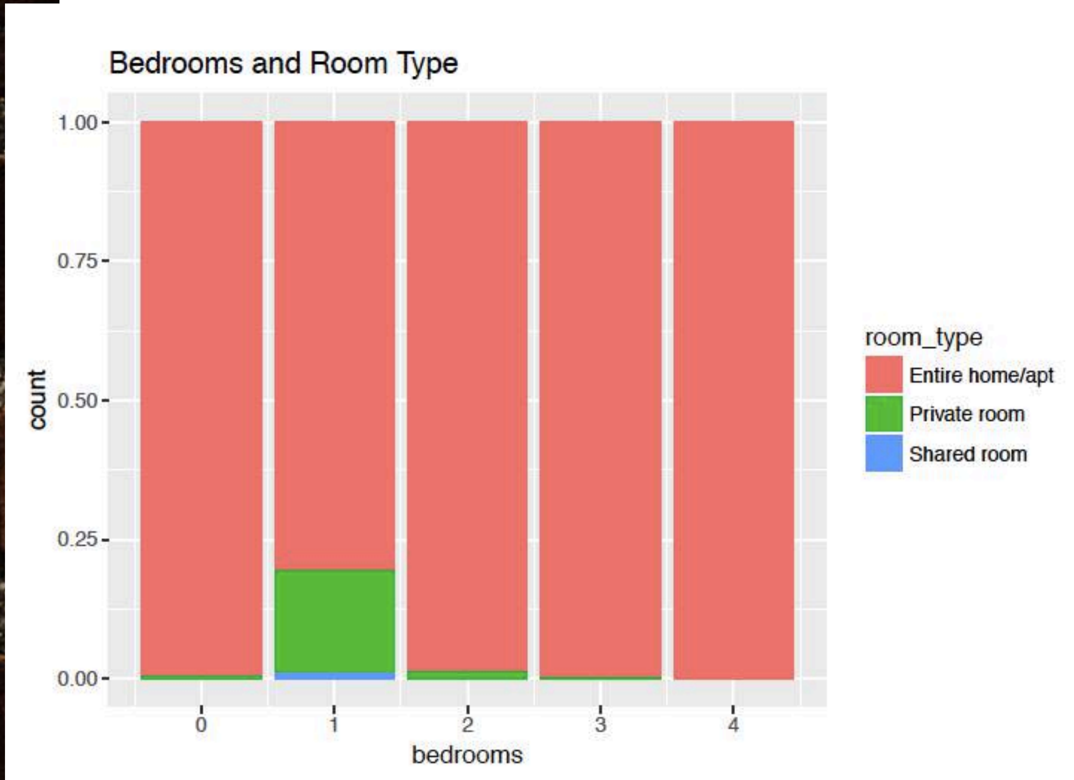


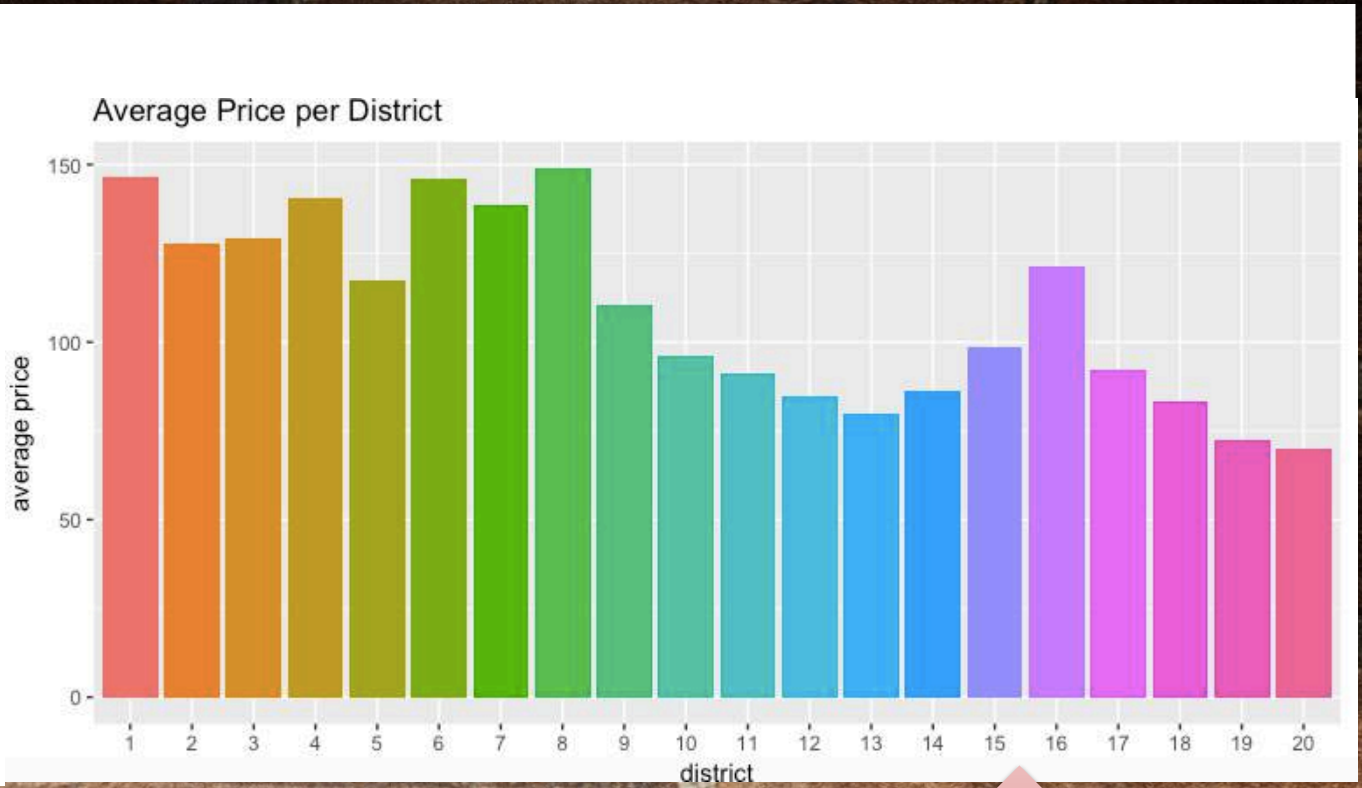


Shared room tends to have higher weighted ratings

EDA

Interaction between room type and accommodate
Entire room tends to allow more accommodates





EDA

No obvious interaction between room type and bedrooms

Price varies between district

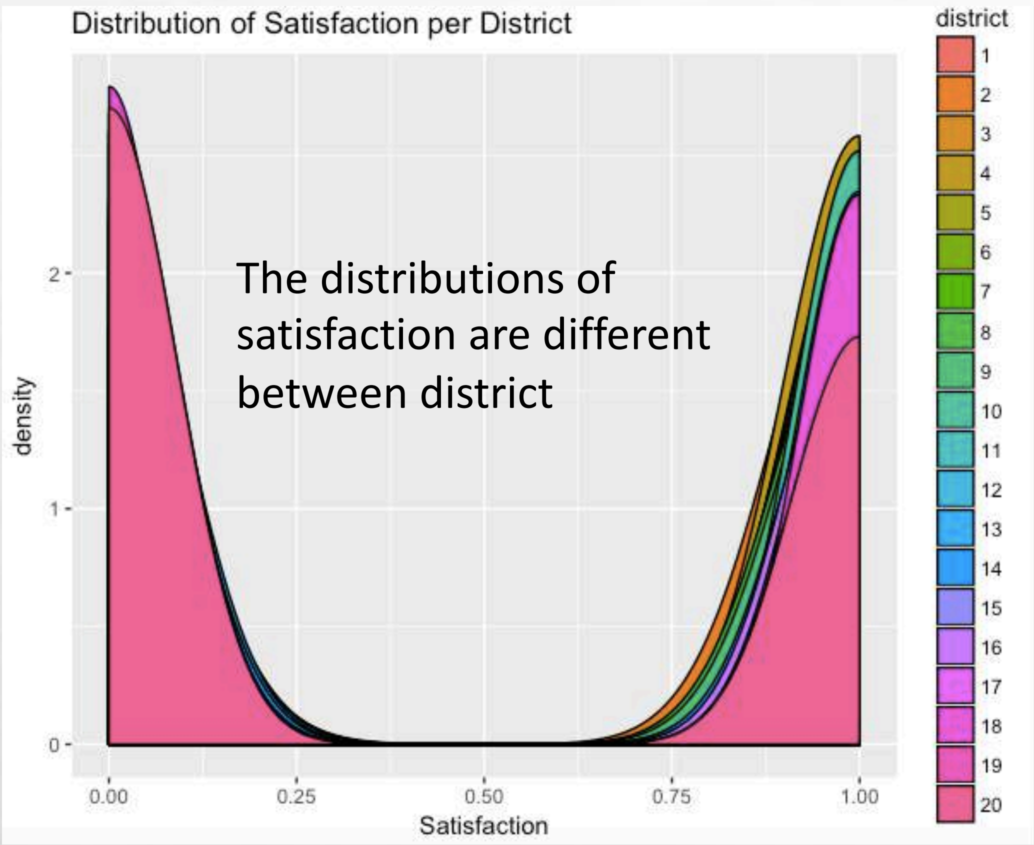
Analysis - Descriptive statistics

Satisfaction	No	Yes
Binary	25,898	24,508

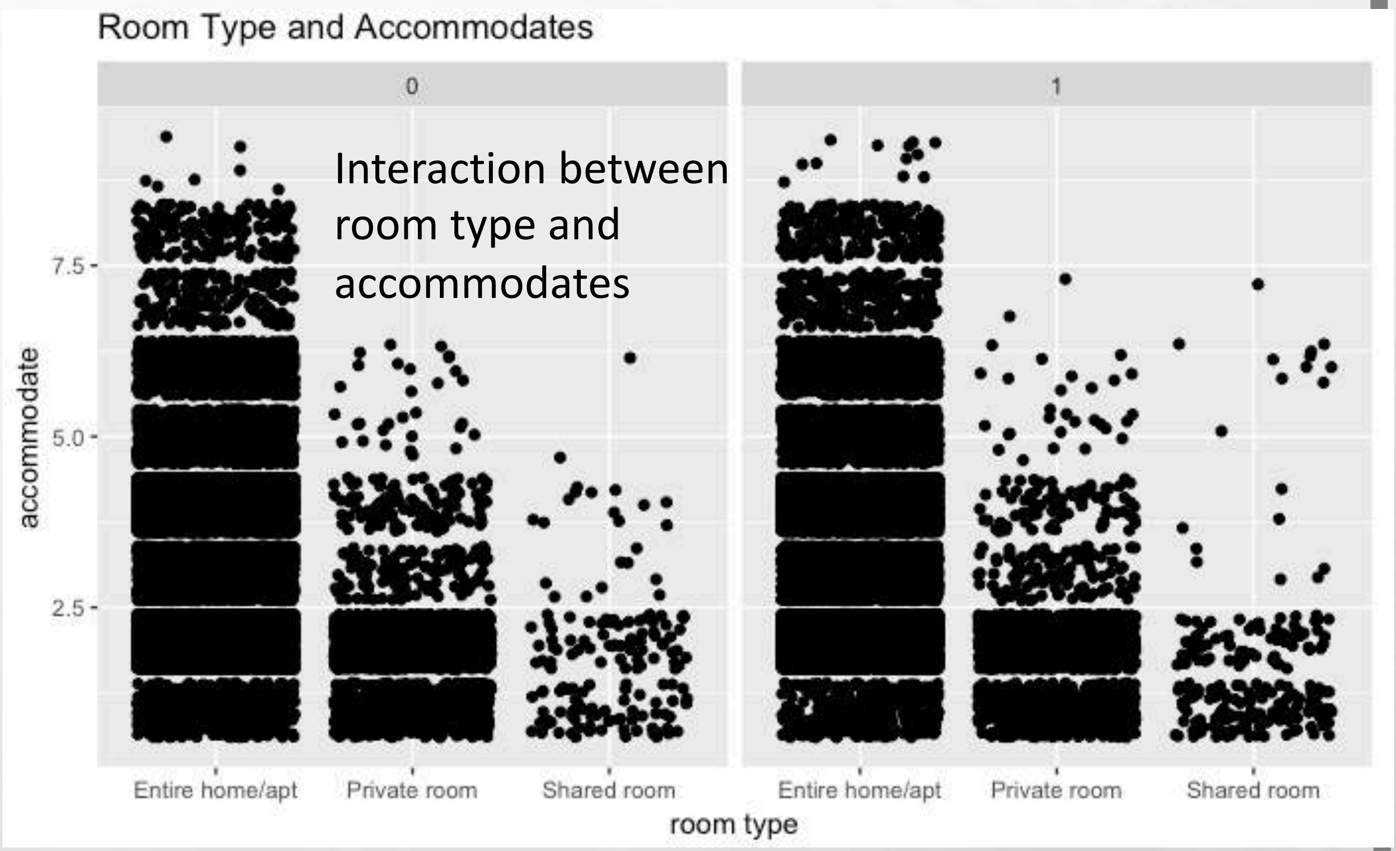
Room type	Entire room/Apt	Private room	Shared room
Category	44,418	5,612	376

Accommodate	Min	1st Qu	Median	Mean	3rd Qu	Max
Numeric	1	2	2	3	4	9
Bedrooms	**Min**	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**
Numeric	0	1	1	1	1	4
Price	**Min**	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**
Numeric	11.0	60.0	84.0	101.3	119.0	555.0
Violence rate	**Min**	**1st Qu**	**Median**	**Mean**	**3rd Qu**	**Max**
Numeric	3.3	4.6	5.7	7.5	8.7	33.3

EDA



Distribution of Satisfaction per District
density
Satisfaction
0.00
0.25
0.50
0.75
1.00
0
1
2
district
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
The distributions of satisfaction are different between district



Room Type and Accommodates
0
1
accommodate
2.5
5.0
7.5
Entire home/apt
Private room
Shared room
Entire home/apt
Private room
Shared room
room type
Interaction between room type and accommodates

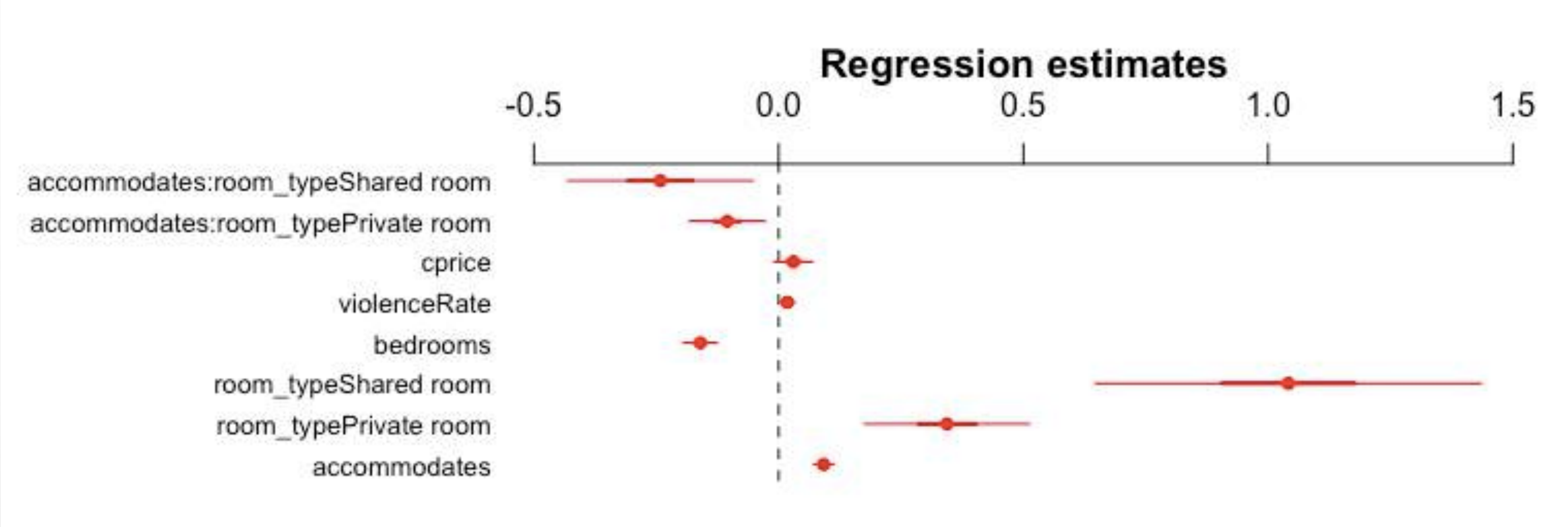
Analysis

Analytical Approach: multilevel logistic regression
Group level: districts
Individual level: per Airbnb

Dependent variable	**Satisfaction: binary variable** **Yes: weighted average rating >=2** **No: weighted average rating<2**
Independent variable Group level	Violence rate: per 1000 inhabitants of each district Data source: www.lefigaro.fr
Independent variable Individual level	Room type, Accommodates Bedrooms Price

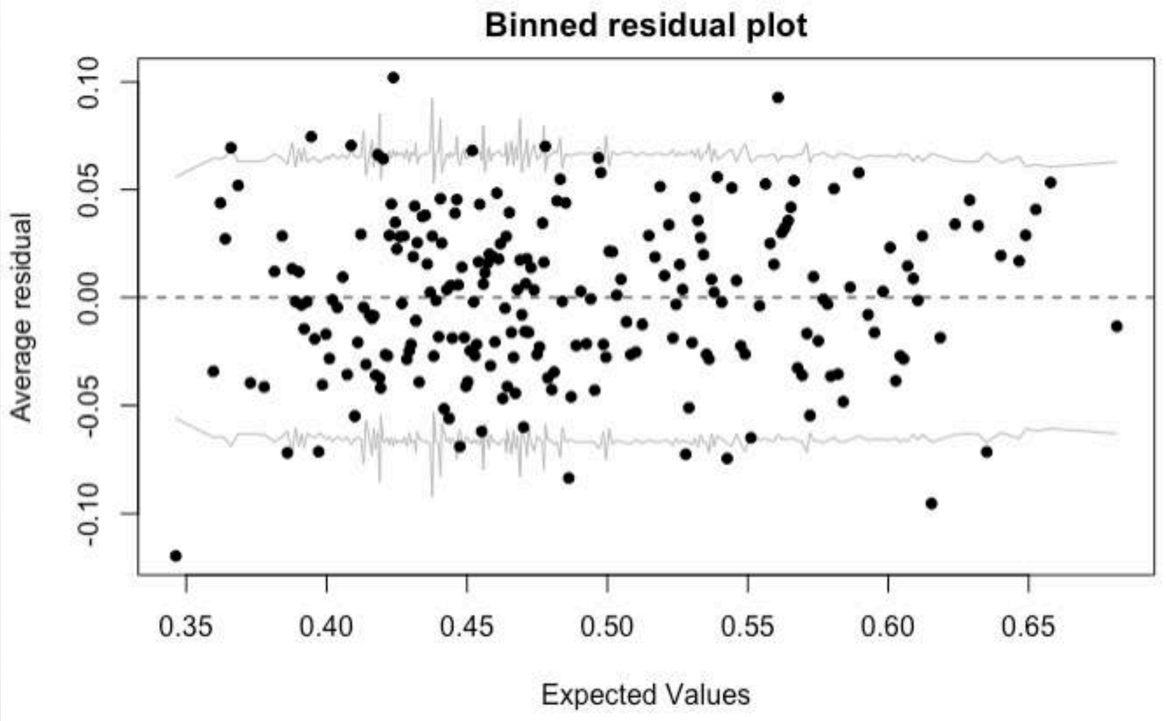
```
glmer(formula = Satisfaction ~ accommodates * room_type + bedrooms +
    violenceRate + cprice + (1 + cprice | district), data = pardata,
    family = binomial)
```

Standardized price

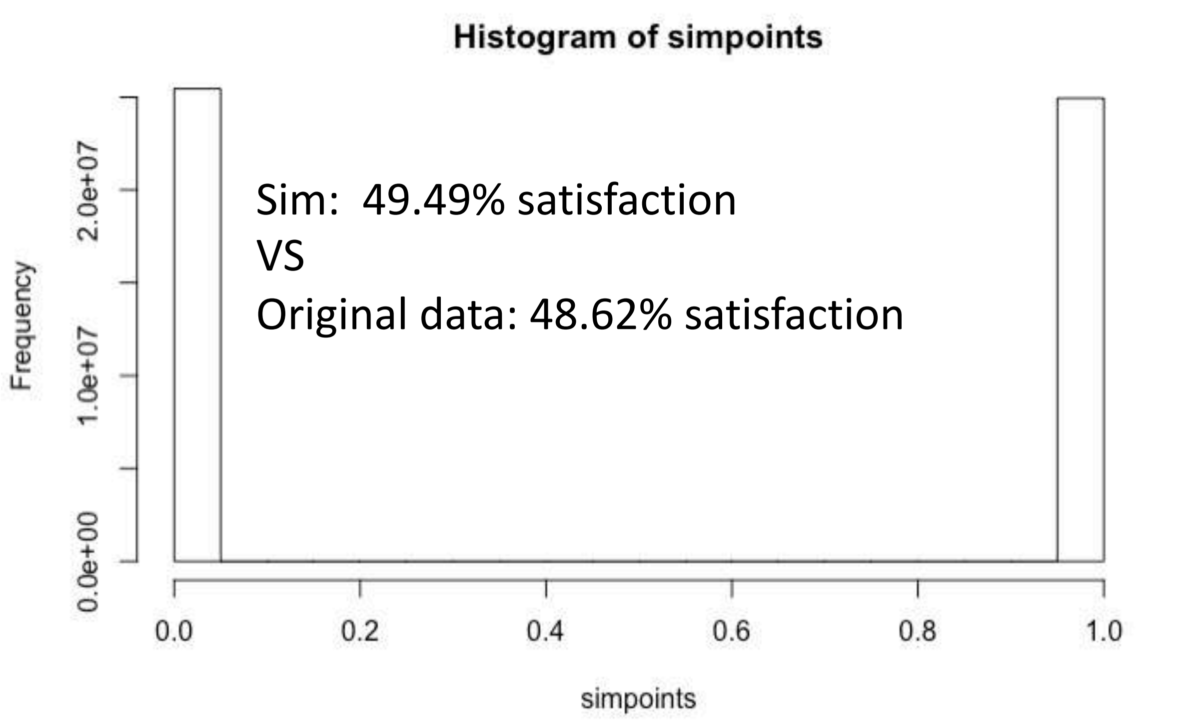


Model Diagnosis

Residual Analysis

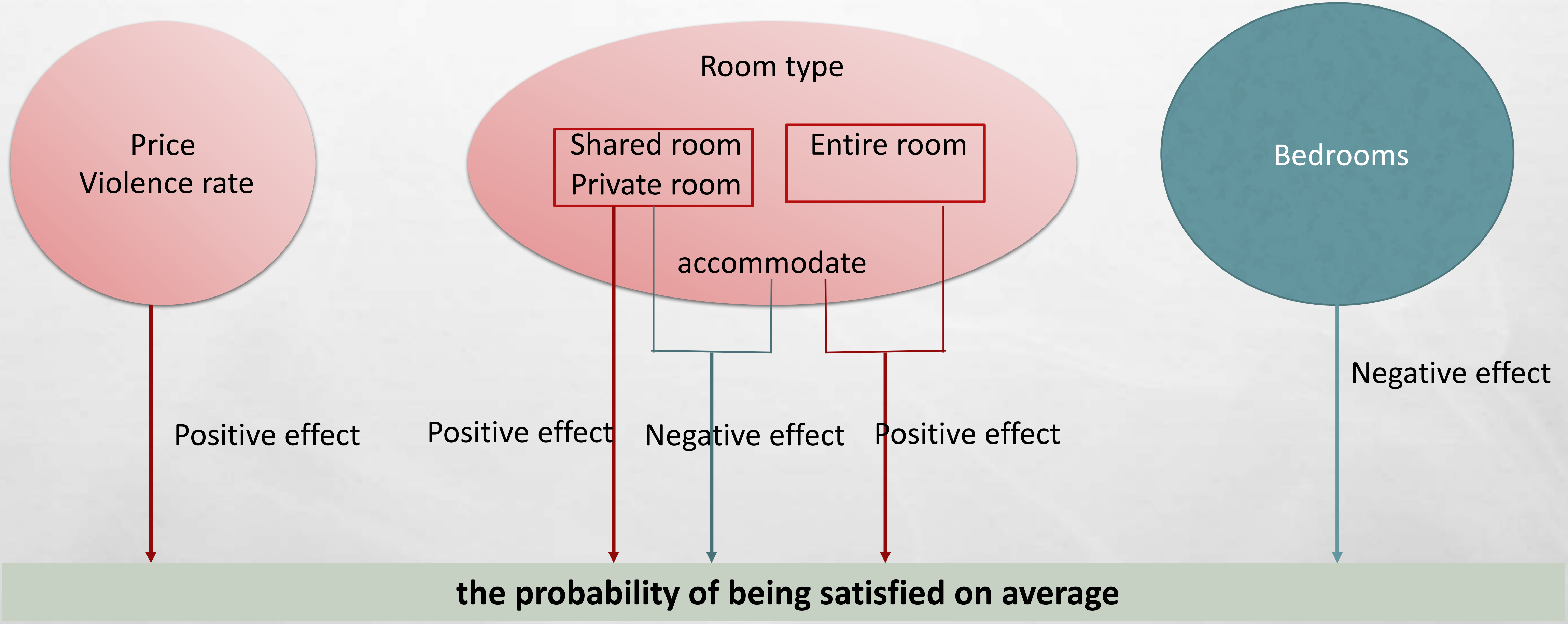


Predictive Checking

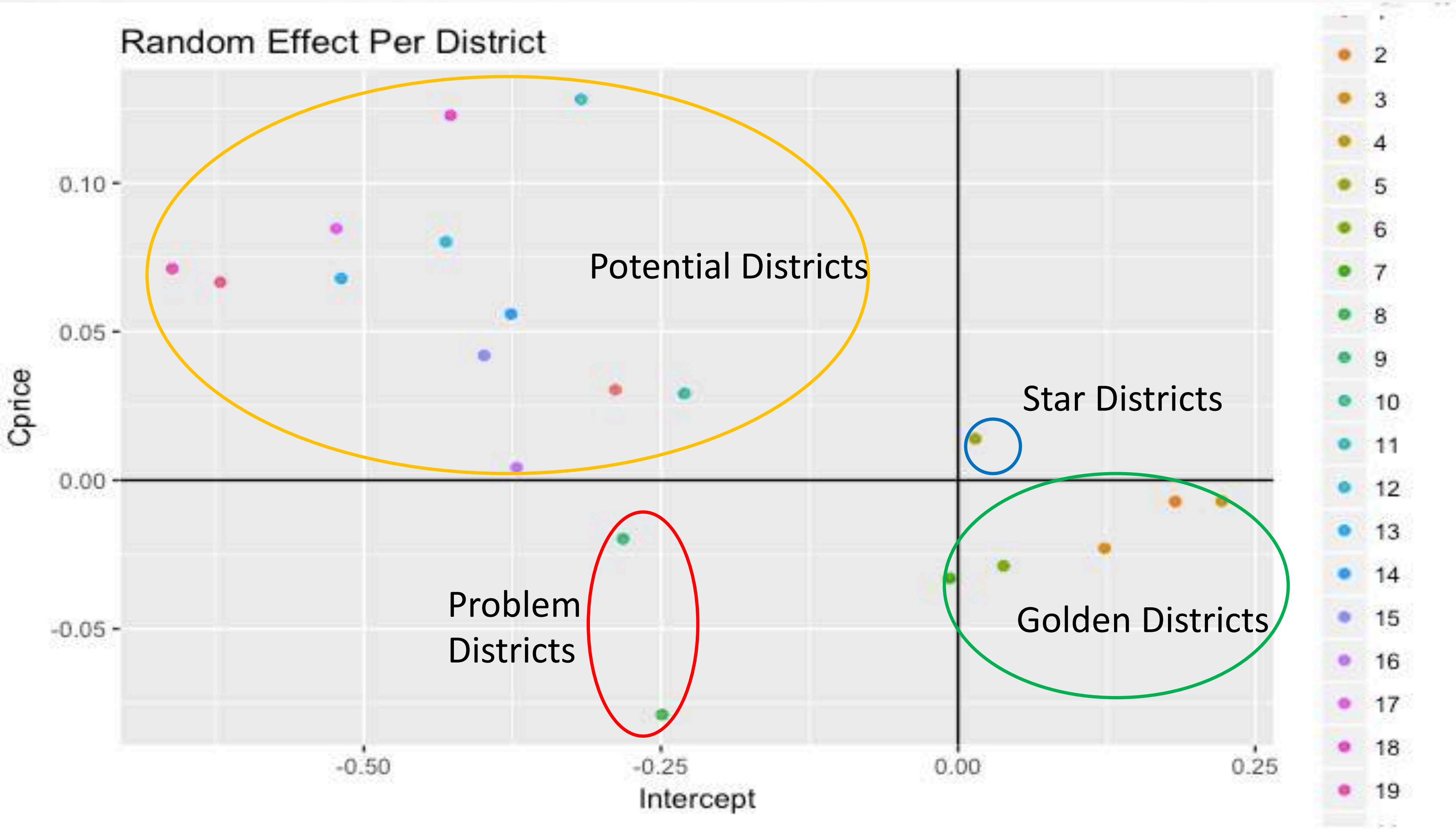


Pretty good fit of the data

Implications-Fixed Effect



Implications-Random Effect



(1) Random intercept: we can see 2,3,4,5,6 districts have positive intercept, which means they have higher probability of being satisfied with average price than other districts when other variables are same

(2) Random slope: we can see 2,3,4,6,7,8,9 districts have negative slope, which means each unit increase of price deviating from average price over standard error will decrease probability of being satisfied, while other districts have positive slope

Star District
(positive random intercept and random slope):
district 5th

Airbnbs in this district have higher probability of being satisfied generally, and there is still potential for Airbnb hosts to increase the price

Golden District
(positive random intercept but negative random slope):
district 2nd, 3rd, 4th, 6th, 7th

Airbnbs in this district have higher probability of being satisfied generally, but they are already highly priced, further increase in price will decrease the ratings

Potential District
(negative random intercept and random slope):
district 1st, 10th, 11th, 12th, 13th, 14th, 15th, 16th, 17th, 18th, 19th, 20th

Airbnbs in this district have lower probability of being satisfied generally, but there is still potential for Airbnb hosts to increase the price

Problem District
(negative random intercept and random slope):
district 8th, 9th

Airbnbs in this district have lower probability of being satisfied generally, but they are already highly priced, further increase in price will decrease the ratings

Recommendations

For Airbnb host:

1) Shared room and private room tend to be favored by guest
2) Increasing the accommodate of entire room could attract guests
3) Downtown area is preferred by guests despite of the high violence rate
4) Property with many bedrooms are not welcomed by guests
5) Airbnb hosts in star and potential district can increase the price
6) Airbnb hosts in problem and golden district better not increase price, otherwise ratings would be negatively impacted

For Airbnb guest:

1) If you have enough budget and prefer a comfortable place for stay, you can choose Airbnb in star districts
2) If you don't mind high price and prefer a comfortable place for stay, you can choose Airbnb in golden district
3) If you have limited budget and don't have preference for the place to stay, you can choose Airbnb in potential districts

Study limitations

1) Time limitation

The data only contained the listing properties as of 25th July, which couldn't display a full picture of the ratings in terms of trend over time.

2) Data limitation

We don't have the text data per review, with which we can carry out further text analysis.

Appendix

1. http://tomslee.net/airbnb-data-collection-get-the-data
2. http://www.developintelligence.com/blog/2017/06/practical-neural-networks-keras-classifying-yelp-reviews/
3. Andrew Gelman, Data Analysis Using Regression and Multilevel/Hierarchical Models, 1st Edition, 2006
4. http://www.lefigaro.fr/actualite-france/2017/01/02/01016-20170102ARTFIG00290-decouvrez-la-carte-des-crimes-et-delits-en-france-et-dans-le-grand-paris.php