

# AirBnB Listings: An in depth dive into the world of short-term sublets

Armandas Bartas, Alex Romanus, Braeden Norman, Gabriel Lanzaro

2021-11-09

## Motivation

This dataset is interesting to us because it combines our love for statistics with our love for vacation planning. Statistical analysis of this data will provide insights while comparing prices and booking accommodations.

## Introduction

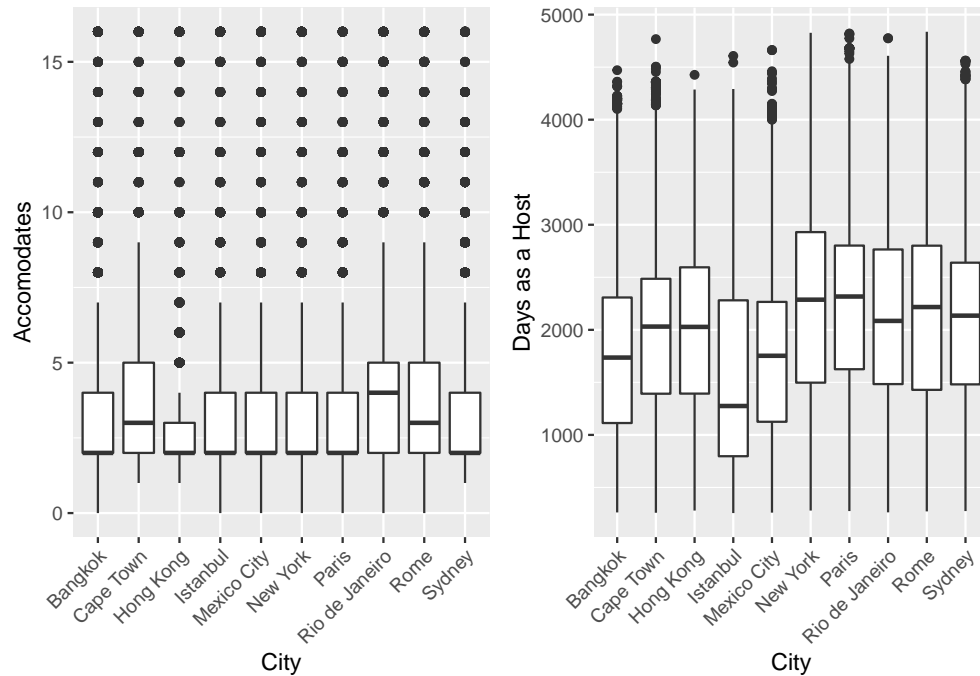
This project aims to investigate AirBnb listings and obtain insights into the most important features of short-term sublets. More specifically, the goal of this project is to classify the cities based on different attributes. The dataset, which was obtained from Kaggle, contains 10 cities from very distinct parts of the world: Bangkok, Cape Town, Hong Kong, Istanbul, Mexico City, New York, Paris, Rio de Janeiro, Rome, and Sydney. The Airbnb data contains 280 000 listings including, but not limited to: host info, geographical data, price, number of bedrooms, amenities, and review scores.

The analysis can then reveal important aspects regarding how different attributes may characterize each city, for example:

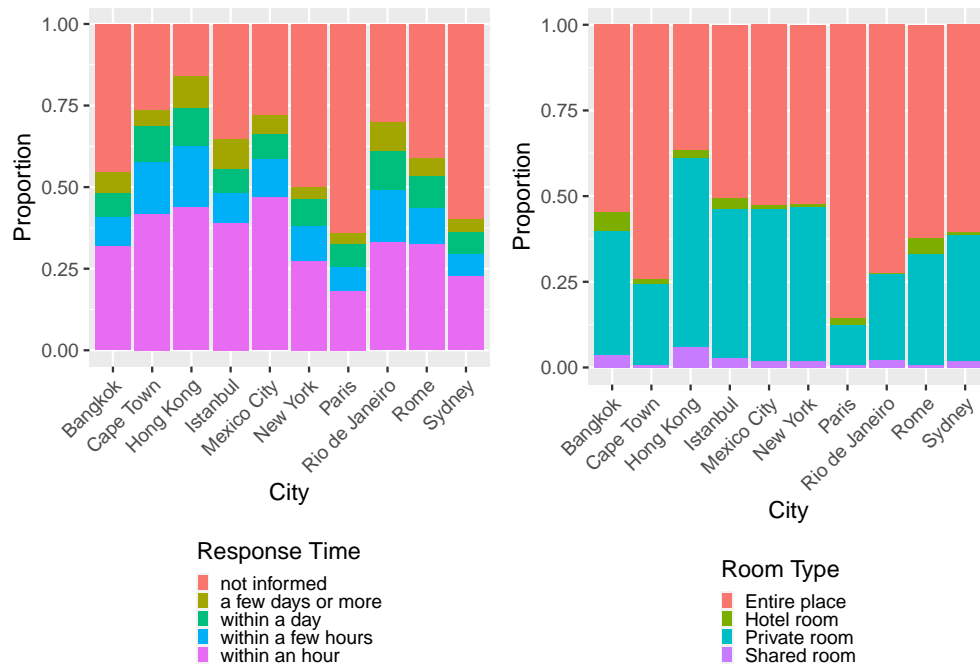
- Which amenities are more important for each city when selecting a property?
- Does the host profile differ among different cities?
- Which types of accommodation are more common depending on the city?
- Can we predict the city based on different preferences related to the place to stay?

## Exploratory Analysis

Several insights can be obtained by plotting different variables against the cities. For example, the next figure shows boxplots that present (1) the number of guests the listing accommodates and (2) for how long the host has been renting properties in AirBnB. The first boxplot indicates that cities such as Cape Town, Rio de Janeiro, and Rome tend to offer listings with more guests, which might be suitable for group or family trips. For Hong Kong, however, the accommodations tend to be for fewer guests, which shows that listings might be tiny and that the city is more appropriate for business trips. In addition, the second boxplot shows that AirBnb has been used in some cities for more time than in others. For example, New York and Paris have an average for the number of days as a host variable that is considerably higher than the average for Istanbul. It might show that AirBnB has only been widely used in Istanbul for a shorter amount of time.



The next plot shows the proportion per city of (1) response time and (2) room type for different categories. The first plot shows that hosts in Mexico City and Hong Long tend to have the highest response times, whereas hosts in Paris and New York have the lowest response times. The second plot shows that most of the accommodations in Paris and Cape Town are for the entire place, and most of the accommodations in Hong Kong are for private rooms. This room type analysis for Hong Kong is consistent with the previous plots (i.e., the number of guests a property can accommodate). The room type variable can also provide information regarding the trip purpose (e.g., business, family, group). Cities such as Paris are preferred for group trips, whereas Hong Kong is more appropriate for business trips.



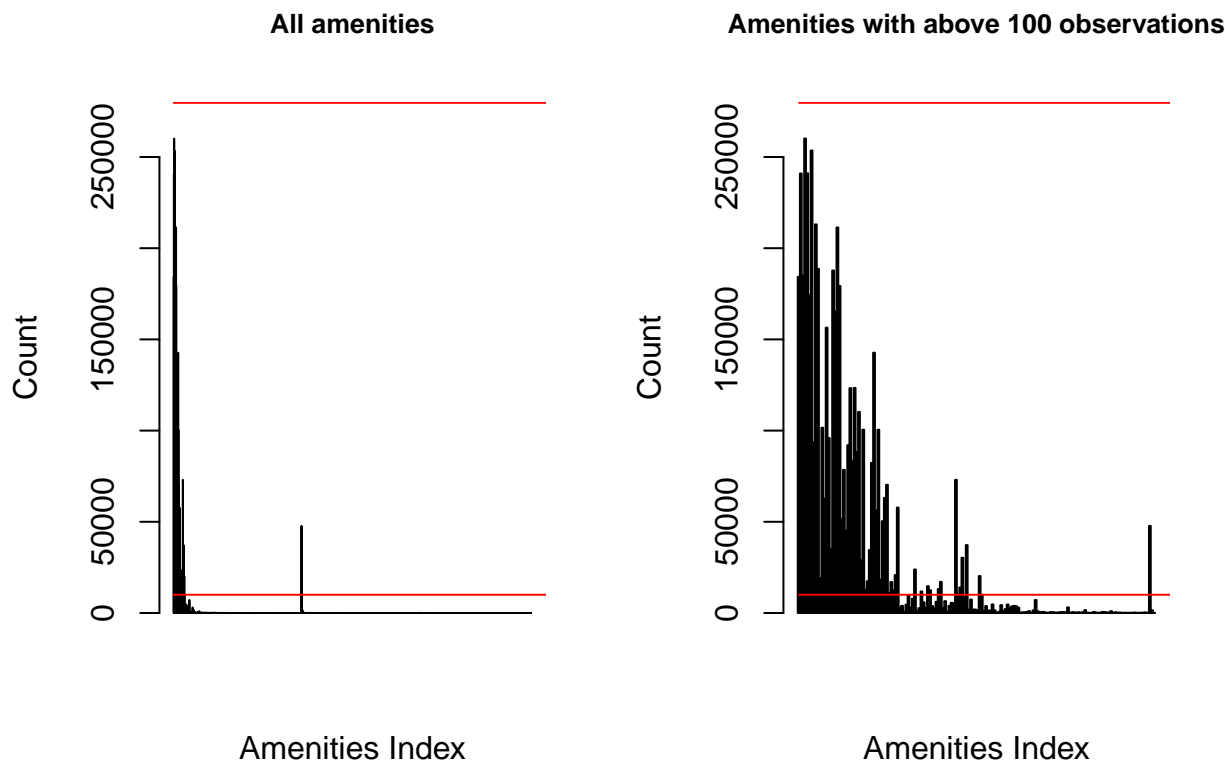
## Exploring Amenities

In the original dataset, there is a column with a list of string of possible amenities a listing has. Here we explored these lists to extract useful information for our predictions. For example, these are the first 5 observations:

```
["Heating", "Kitchen", "Washer", "Wifi", "Long term stays allowed"], ["Shampoo", "Heating", "Kitchen",  
"Essentials", "Washer", "Dryer", "Wifi", "Long term stays allowed"], ["Heating", "TV", "Kitchen",  
"Washer", "Wifi", "Long term stays allowed"], ["Heating", "TV", "Kitchen", "Wifi", "Long term stays  
allowed"], ["Heating", "TV", "Kitchen", "Essentials", "Hair dryer", "Washer", "Dryer", "Bathtub", "Wifi",  
"Elevator", "Long term stays allowed", "Cable TV"]
```

We found the list of all amenities and graphically determined which to include in the updated dataset. For each included amenity, we added new columns labeling whether this listing has these amenities or not.

The first graph below shows the count of all amenities. There are 2865 different amenities in the dataset. A lot of them only have a count of 1 so, to better view the distribution, we removed all amenities that had below 100 observations. The second graph gives all amenities with a count over 100. The top red line is at the total number of observations in the dataset, and the bottom line is at 10,000. 10k was decided to be a good number to remove all amenities with less observations. This would leave us with a more reasonable size of amenities to add to our dataset as new columns (60 new columns after reduction).



All observations were updated with new columns: 1 for has amenity and 2 for not. Here is a quick preview of what the updated dataset looks like.

The first graph shows the count of all amenities. There is 2865 different amenities in the dataset. A lot of them only have a count of 1, so to better view the distribution we removed all amenities that had below

100 observations. The second graph gives all amenities with a count over 100. The top red line is at total observation of dataset, and the bottom line is at 10,000. 10k was decided to be a good number to remove all amenities with less observations, because this would leave us with a more reasonable size of amenities to add to our dataset as new columns (60 after reduction). All observations we updated with a new columns, 1 for has amenity and 2 for not.

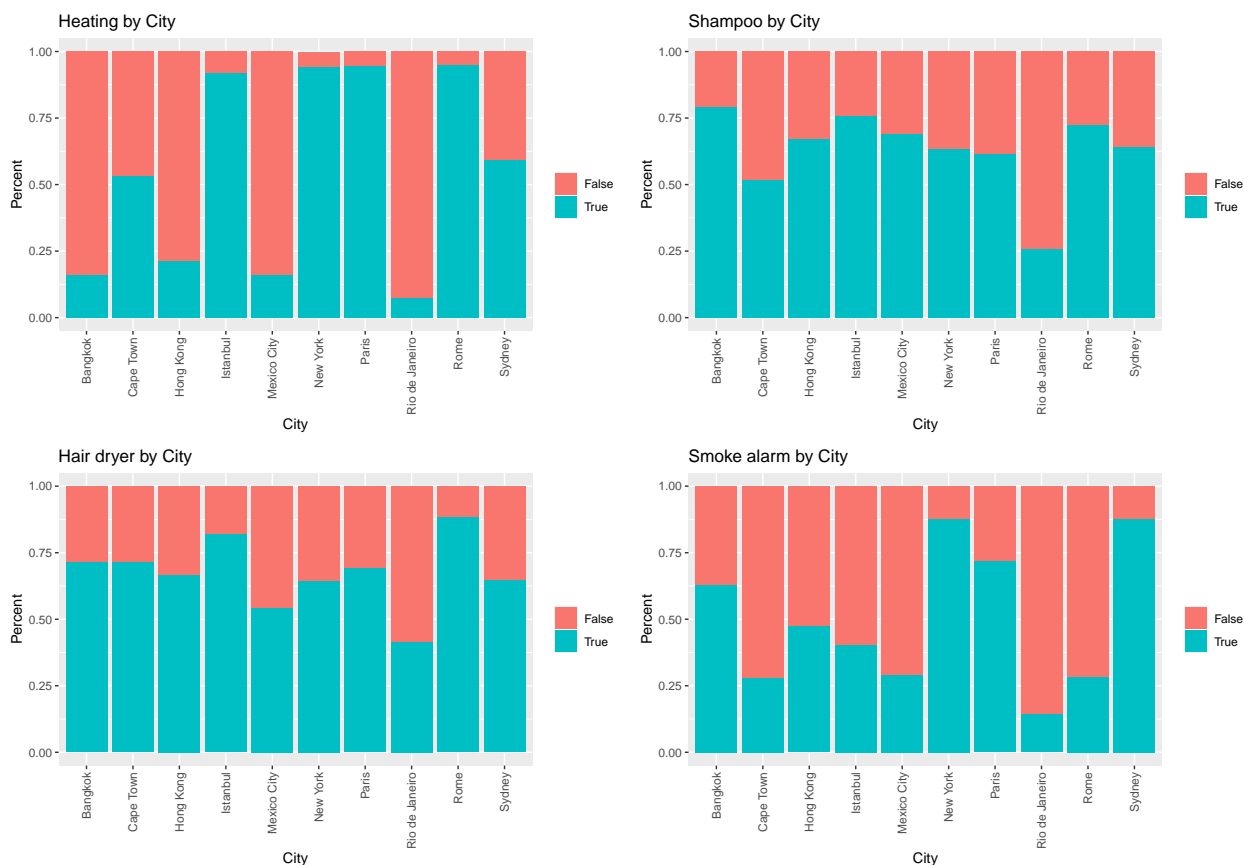
```
listings[1:5,c(2,4,16,35:38)]
```

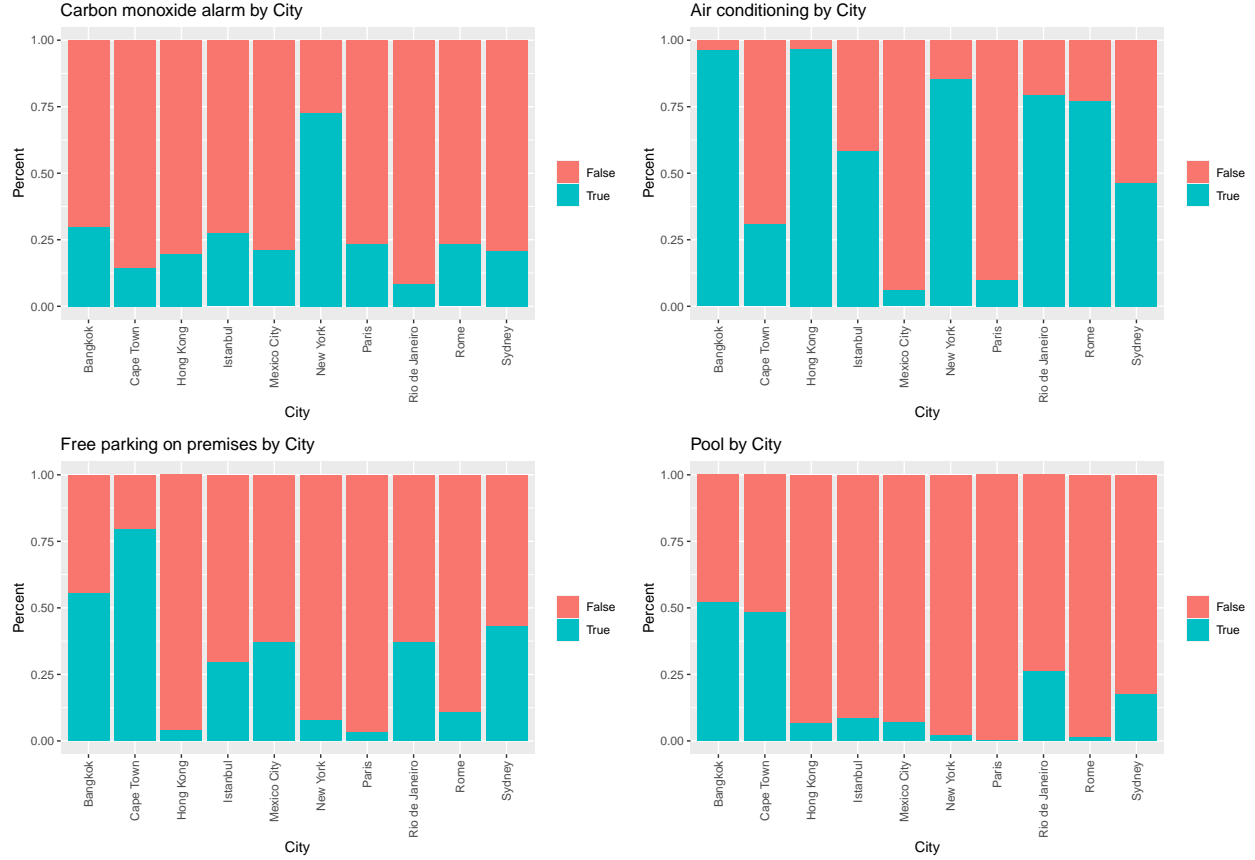
##	listing_id	host_id	city	Heating	Kitchen	Washer	Wifi
## 1	281420	1466919	Paris	1	1	1	1
## 2	3705183	10328771	Paris	1	1	1	1
## 3	4082273	19252768	Paris	1	1	1	1
## 4	4797344	10668311	Paris	1	1	0	1
## 5	4823489	24837558	Paris	1	1	1	1

Below are the percent TRUE/FALSE values of the selected amenities for each city. Since we are trying to determine the city, based on the listings, seeing the different amenities by city will give us an understanding of how useful these amenities will be in our model. The graphs selected were:

Heating, Shampoo, Hair dryer, Smoke alarm, Carbon monoxide alarm, Air conditioning, Free parking on premises, Pool

For example, in Heating, we only have 4 cities that have almost no listings with heating (Bangkok, Hong Kong, Mexico City, and Rio de Janeiro) If we also look at Air conditioning, we see that only 2, maybe, 3 have little to no listings with AC (Cape Town, Mexico City, and Paris). Now, if we take these two into account, we can see that given no AC and no Heating, we can be almost positive the listing is Mexico City.





## Exploring Interactions

The addition of interaction terms is a form of basis expansion. Interaction terms model the change in response variable against the change in a product or quotient of two or more predictor variables. They help in the case where the added effects of two predictor variables do not stack linearly, rather, they compound on each other. The right interaction term can improve a model, but adding too many terms, or the wrong term, increases variance and can actually worsen out of sample performance.

## Method

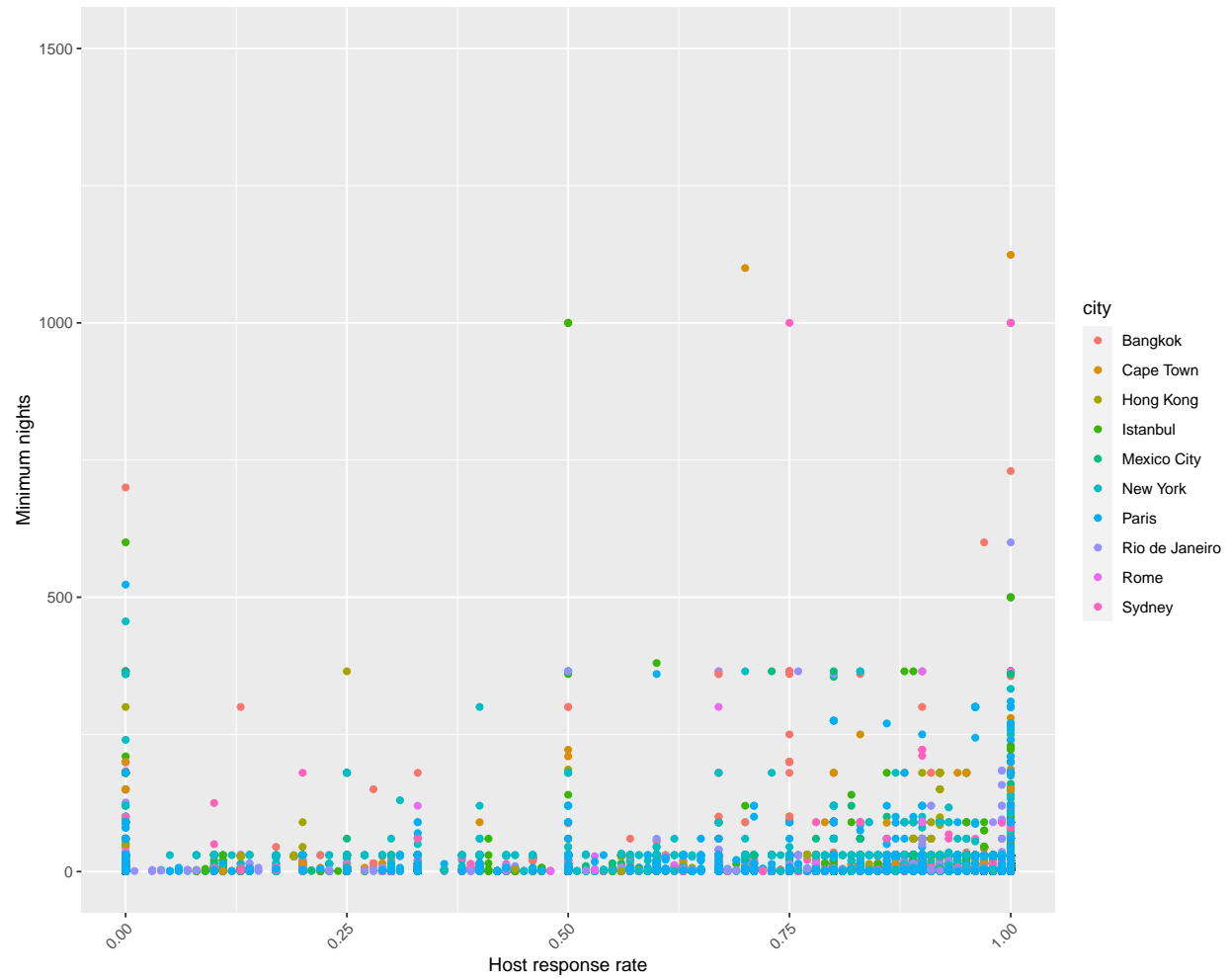
The method was to systematically fit a separate model for each interaction term. The interaction terms used were products of pairs of predictor variables. As such, the categorical variables were not included in the analysis of interactions. The true/false variables were converted into binary 0/1. Some data cleaning had to be done: ListingID and name were removed as they uniquely identify the corresponding observation, geographical information such as neighbourhood, longitude, latitude, district, and location were removed as they also eliminate the challenge of predicting the city. Amenities was removed for this analysis as it provided too much trouble. Each combination of amenities is considered a unique categorical variable. Overcoming this challenge is covered within this report, but it was not necessary for this particular analysis. The data set was cut down from ~300k observations to 1k observations, blocked by city and randomized, with 100 observations for each one. This was done to expediate the process of fitting so many models. The model used was a basic LDA model over all of the variables, minus the aforementioned removals. For each model, one unique product of two numeric variables was added to the predictor space, and the misclassification rate was recorded.

## Results

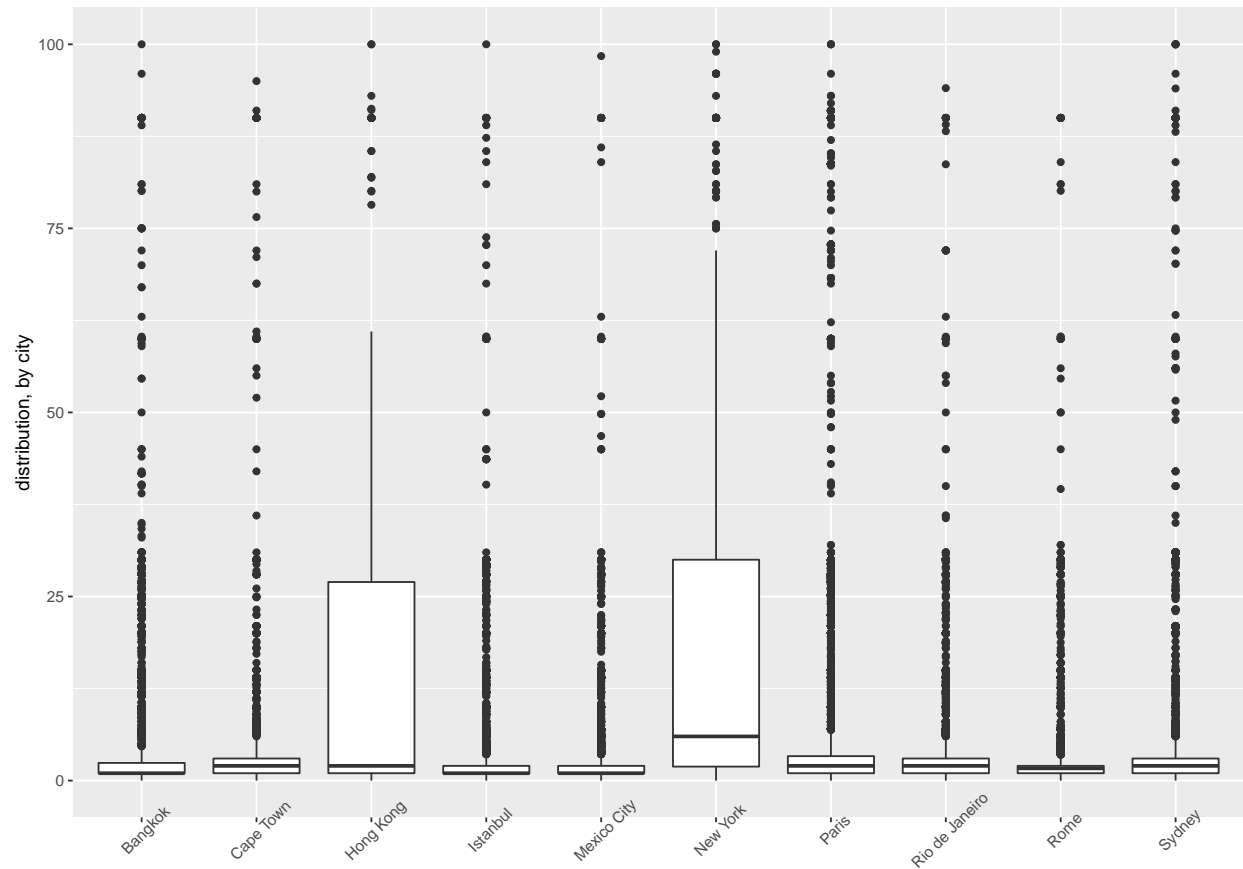
The best performing models are seen in the table below. The default (no interaction terms added) misclassification rate was 0.013. There were two interaction terms that provided a rate of 0.008. The first was `host_response_rate * minimum_nights`, and the second was `host_identity_verified * review_scores_communication`. Five interaction terms gave a misclassification rate of 0.009, three of which contained the variable `host_is_superhost`, and the other two contained the variable `bedrooms`. Of these seven interaction terms, four contained a binary 0/1 variable, essentially meaning that the other term in the interaction had a different effect on the city depending on some binary condition.

```
##      [,1]
## [1,] "misclass error"
## [2,] "0.008"
## [3,] "0.008"
## [4,] "0.009"
## [5,] "0.009"
## [6,] "0.009"
## [7,] "0.009"
## [8,] "0.009"
##      [,2]
## [1,] "model formula"
## [2,] "city ~ . + host_response_rate * minimum_nights"
## [3,] "city ~ . + host_identity_verified * review_scores_communication"
## [4,] "city ~ . + host_is_superhost * host_total_listings_count"
## [5,] "city ~ . + host_is_superhost * price"
## [6,] "city ~ . + host_is_superhost * review_scores_location"
## [7,] "city ~ . + bedrooms * maximum_nights"
## [8,] "city ~ . + bedrooms * review_scores_location"

## Warning: Removed 128782 rows containing missing values (geom_point).
```



## Warning: Removed 129205 rows containing non-finite values (stat\_boxplot).

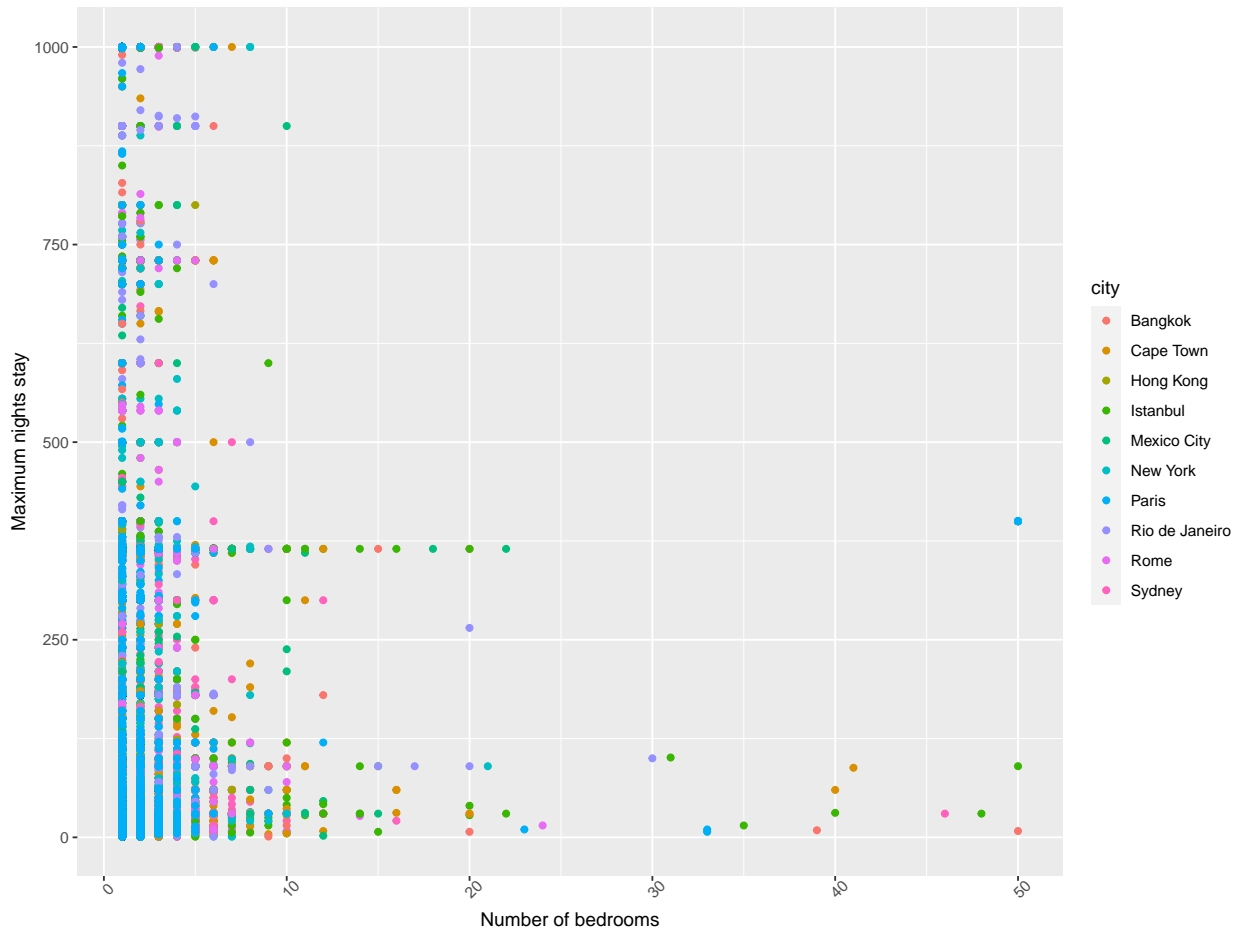


#### Host response rate \* minimum nights interaction term

As we can see, this interaction term helps identify observations as New York or Hong Kong Airbnbs; because their distributions are much higher than the other cities in the dataset. Adding this interaction term to the model may improve the models ability to correctly identify these two cities, as well as deter it from misidentifying other cities as New York or Hong Kong.

```
## Warning: Removed 12101 rows containing missing values (geom_point).
```

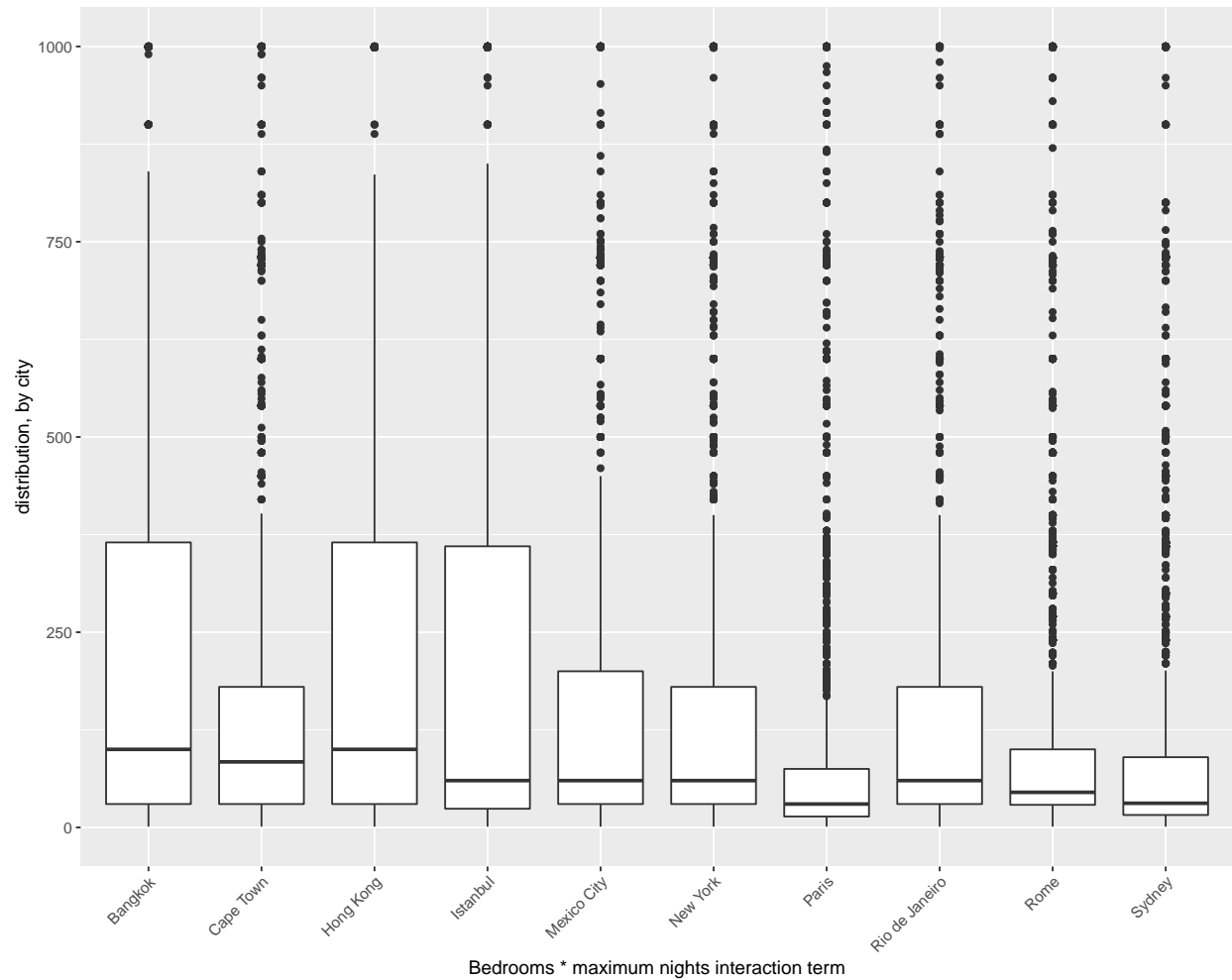




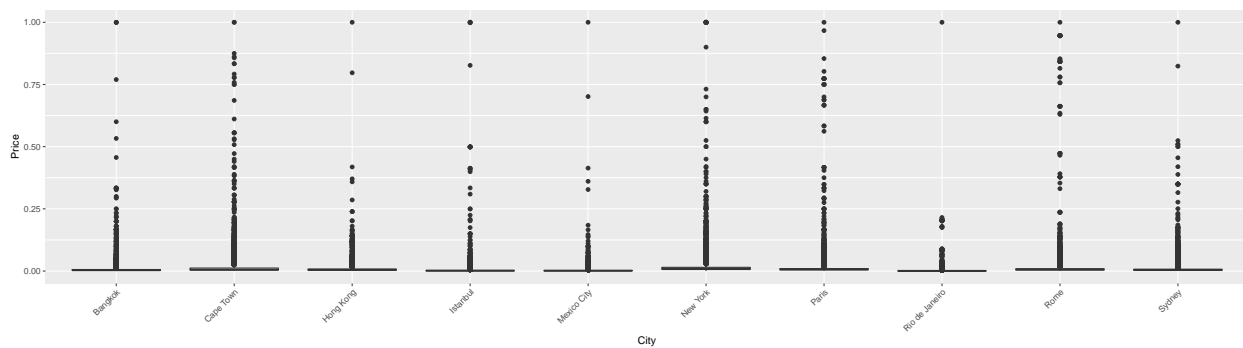
We can see that the observations cluster along the axes of the plot, resulting on low product interaction terms. We see a cluster of observations coloured 'Paris' in the bottom corner, implying that Parisian Airbnbs have low values for their interaction terms. Looking at the next plot, a boxplot of the interaction term by city, we see this is confirmed.

```
## Warning in bedrooms * maximum_nights: NAs produced by integer overflow
```

```
## Warning: Removed 174726 rows containing non-finite values (stat_boxplot).
```

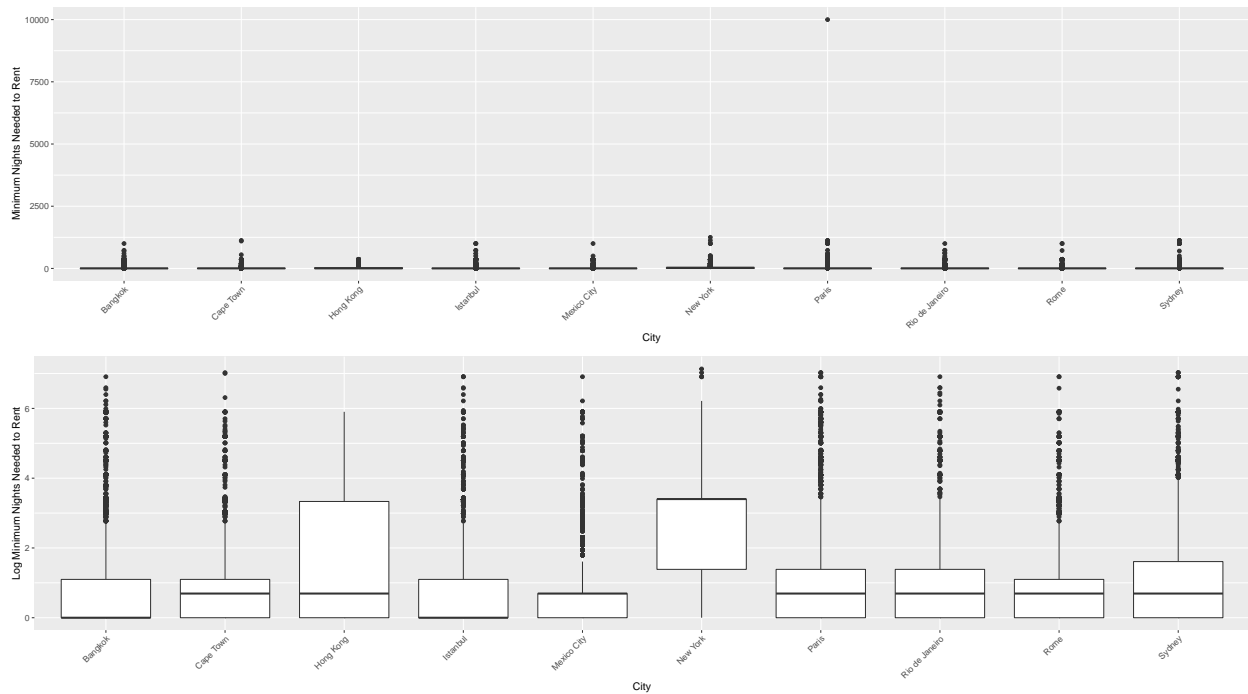


## Exploring Price



Since prices of the Airbnb's are recorded in each city's country's own currency, to make the prices comparable, it makes sense to standardize the variable. The best way to make all listings' prices comparable is to divide each observation's price by it's city's costliest Airbnb. That way, the spreads of the data separated by city are preserved while making the price variable unitless to allow for comparison of observations between cities.

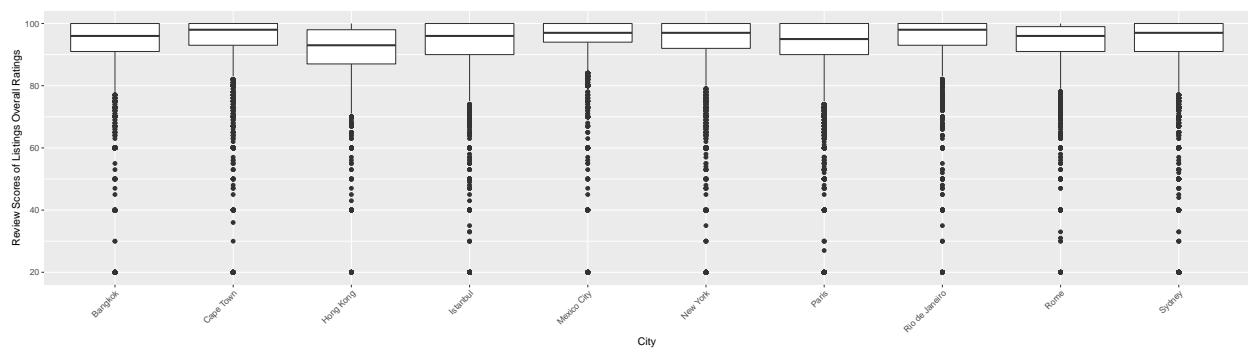
## Exploring Minimum Nights



Since no other observations has anywhere near 10000 minimum nights required to rent, which no person would realistically want to rent anyway, it is reasonable to remove this observation from the data set. Furthermore, there are still many outliers, i.e values above the whiskers, preventing the spread from being reasonably assessed, so we apply a log transformation to `minimum_nights` to make it more readable.

Upon transforming the data, it is clear that most cities have many listings with very few minimum nights required to rent, as many of their boxes show Q1's hovering right at 0. New York is the only exception, with a Q1 value of about 1.5 log days, or about 4.5 days, and a median of about 12 days. Conversely, Istanbul and Bangkok have medians hovering just above 0. These distinct spreads of values of minimum days may make Istanbul, Bangkok, and especially New York much easier to predict if this variable is significant in our model.

## Exploring Review Score of Overall Rating



Many observations had to be removed from the data to accommodate the large amount of NaNs in the `review_score_ratings` variable.

The high medians of and Q1 values of Cape Town and Rio suggest these cities may be easier to predict than

the rest. Hong Kong also seems to have a significantly lower rating spread of ratings than the rest of the cities, suggesting it may also be easier to predict than the rest.

However, unless the model used is sensitive to subtle differences in the values of these ratings, it like won't be very useful for classification. This is because the spreads of the data for these reviews are all very similar, with almost all quantiles above a score of 90. This contradicts what we originally thought, as we predicted that overall reviews may differ significantly from city to city. Generally, it seems Airbnb customers seem to give high reviews.

## Exploring Review Scores of Location

##	city	10	2	3	4	5	6	7	8	9
##	Bangkok	5598	62	0	33	15	181	217	1230	3832
##	Cape Town	10900	50	1	10	9	94	59	426	1847
##	Hong Kong	2698	29	0	12	4	33	24	175	804
##	Istanbul	7886	147	2	52	28	212	137	713	2042
##	Mexico City	12674	70	0	8	2	78	42	233	1336
##	New York	18819	64	0	27	14	200	151	1245	6220
##	Paris	36662	76	1	22	12	243	239	1882	8833
##	Rio de Janeiro	13657	46	1	20	7	114	63	497	1701
##	Rome	13842	29	1	27	21	138	130	1076	5548
##	Sydney	17759	62	1	23	8	187	89	988	3186

##	city	10	2	3	4	5	6	7	8	9
##	Bangkok	3.98%	9.76%	0.00%	14.10%	12.50%	12.23%	18.85%	14.53%	10.84%
##	Cape Town	7.76%	7.87%	14.29%	4.27%	7.50%	6.35%	5.13%	5.03%	5.23%
##	Hong Kong	1.92%	4.57%	0.00%	5.13%	3.33%	2.23%	2.09%	2.07%	2.27%
##	Istanbul	5.61%	23.15%	28.57%	22.22%	23.33%	14.32%	11.90%	8.42%	5.78%
##	Mexico City	9.02%	11.02%	0.00%	3.42%	1.67%	5.27%	3.65%	2.75%	3.78%
##	New York	13.39%	10.08%	0.00%	11.54%	11.67%	13.51%	13.12%	14.71%	17.60%
##	Paris	26.09%	11.97%	14.29%	9.40%	10.00%	16.42%	20.76%	22.23%	24.99%
##	Rio de Janeiro	9.72%	7.24%	14.29%	8.55%	5.83%	7.70%	5.47%	5.87%	4.81%
##	Rome	9.85%	4.57%	14.29%	11.54%	17.50%	9.32%	11.29%	12.71%	15.69%
##	Sydney	12.64%	9.76%	14.29%	9.83%	6.67%	12.64%	7.73%	11.67%	9.01%

Similar to overall review scores, many observations had to be removed to analyze `review_scores_location` due to the high volume of NaNs. Some cities, such as Hong Kong, were significantly effected, as it is far fewer location reviews than, say, Paris.

The largest proportion of high reviews, i.e 9s and 10s, belong to Paris at 24.99% and 26.09% respectively. Also, Istanbul has large shares of low reviews, with 23.15%, 28.57%, and 22.22% of 2s, 3s, and 4s respectively, compared to th other cities. These heavy tails may make it easier to classify these two cities if review scores based on location is a significant predictor in our model.