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

Armandas Bartas, Alex Romanus, Braeden Norman, Gabriel Lanzaro

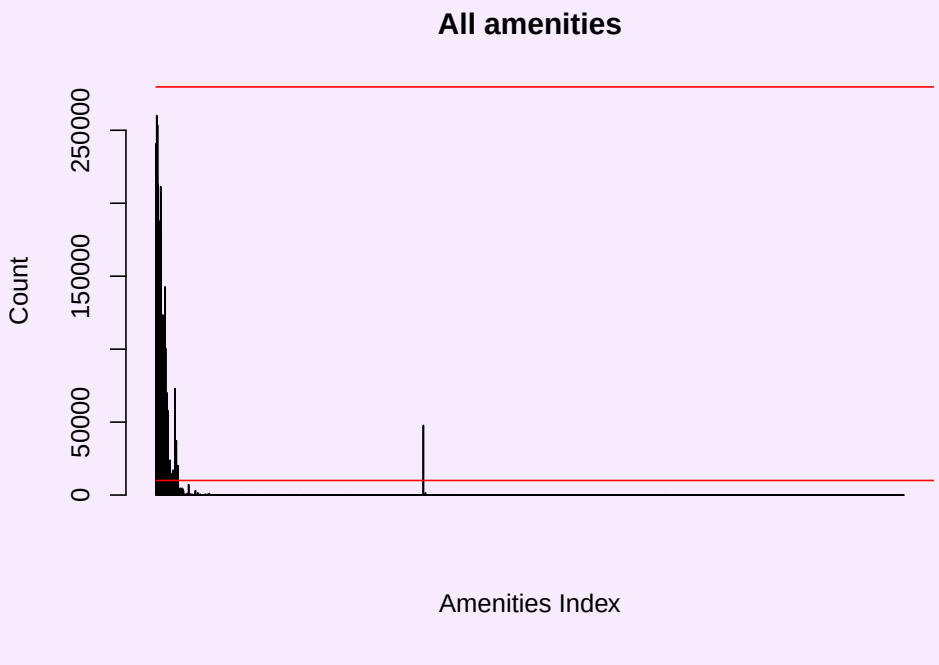
2021-11-08

```
library(tidyverse)
library(testthat)
listings <- read.csv("data/Listings_updated.csv")
amenitiesCount <- read.csv("data/amenities_count.csv")
```

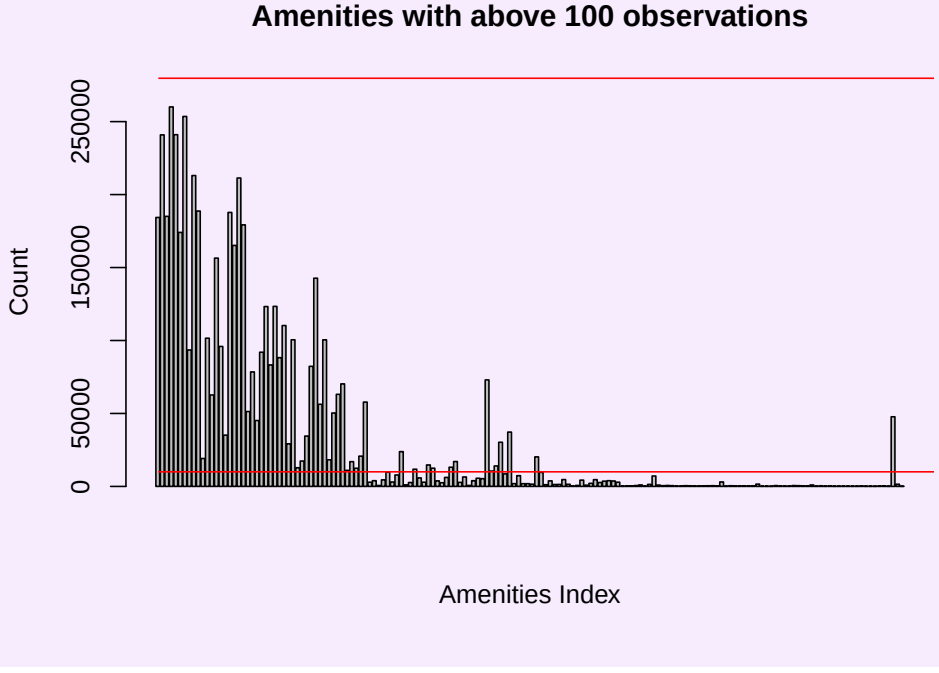
Exploring Amenities

In the original dataset there is a columns that has a list of string of all amenities a listing has. Here we explored these lists to extract useful information for our predictions.
["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”]
This is the first 5 observation. We found the list of all the amenities and graphically determined which to include in update dataset. For each included amenities, we will add a new columns labeling whether this listing has this amenities or not. (1/0)

```
barplot(amenitiesCount$V2, ylim = c(0, 280000), main = "All amenities"
        , xlab = "Amenities Index", ylab = "Count")
lines((integer(279712) + 1)*279712, col = "red")
lines((integer(279712) + 1)*10000, col = "red")
```




```
noOnes <- amenitiesCount$V2[-which( amenitiesCount$V2 < 100)]
barplot(noOnes, ylim = c(0, 280000), main = "Amenities with above 100 observations"
        , xlab = "Amenities Index", ylab = "Count")
lines((integer(279712) + 1)*279712, col = "red")
lines((integer(279712) + 1)*10000, col = "red")
```



```
reducedAmenities <- amenitiesCount$V1[which(amenitiesCount$V2 > 10000)]
reducedAmenitiesCount <- amenitiesCount$V2[which(amenitiesCount$V2 > 10000)]
```

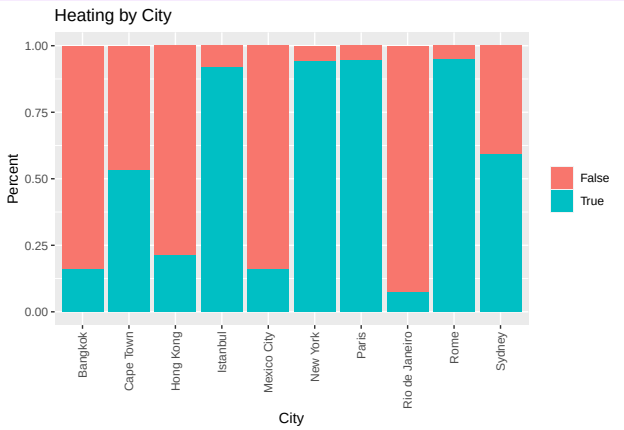
The first graph shows the count of all amenities. There is 2865 different amenities in the dataset. A lot of them only have a count of 1, so to better view the distribution we removed all amenities that had below 100 observations. The second graph gives all amenities with a count over 100. The top red 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.

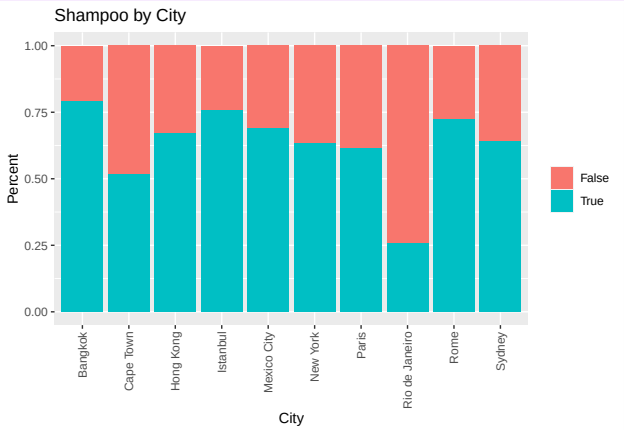
```
  listings[1:5,c(2,4,16,35:38)]
```

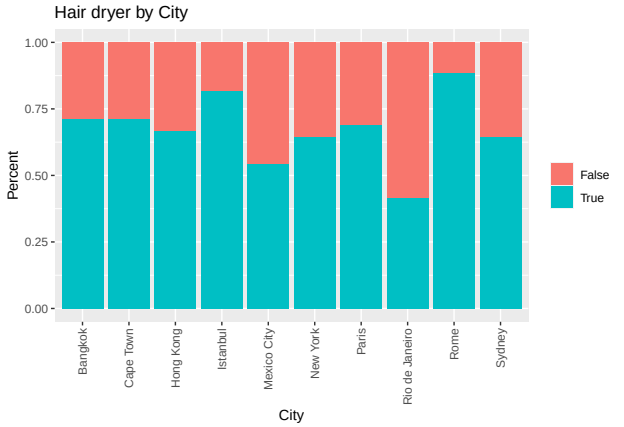
```
##   listing_id  host_id  city Heating Kitchen Washer Wifi
## 1     281420  1466919 Paris       1       1      1    1
## 2    3705183 10328771 Paris       1       1      1    1
## 3    4082273 19252768 Paris       1       1      1    1
## 4    4797344 10668311 Paris       1       1      0    1
##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

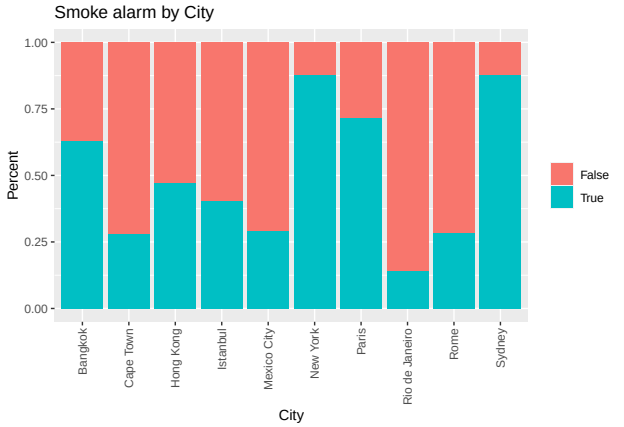
Quick preview of what the updated dataset looks like.
Below are the percent TRUE/FALSE values of the selected amenities for each city. Since we are trying to determine the city, based on the listings, seeing the difference of amenities by city will give us an understanding of how useful these amenities 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
The usefulness was determine by seeing the distinct difference between each city. 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 Heating, we can be almost positive the listing is Mexico City.

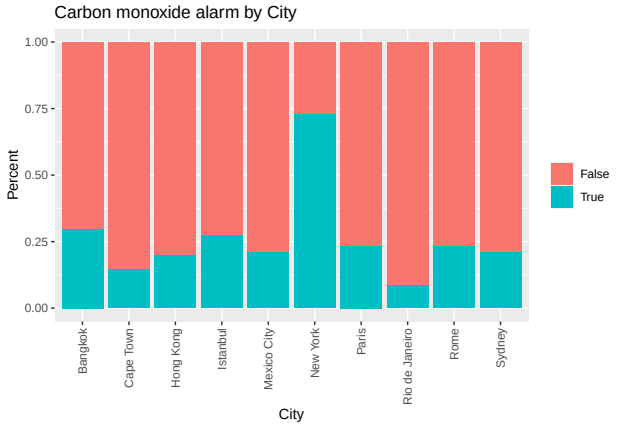
```
selected <- c(1,6,10,14,22,36,54,58)
knitr::opts_chunk$set(fig.width=unit(18,"cm"), fig.height=unit(5,"cm"))
for (ii in selected) {
   plt <- ggplot(listings, aes(x = city,
      fill = factor(listings[,str_replace_all(reducedAmenities[ii], " ", ".")],
      levels = c(0, 1), labels = c("False", "True")))) +
      geom_bar(position = "fill") +
      labs(y = "Percent", fill = "", x = "City",
      title =  paste(reducedAmenities[ii],"by City")) +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
    print(plt)
}
```

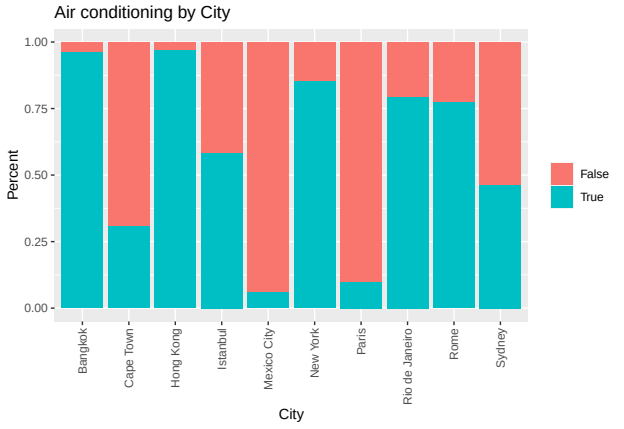

Heating by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True

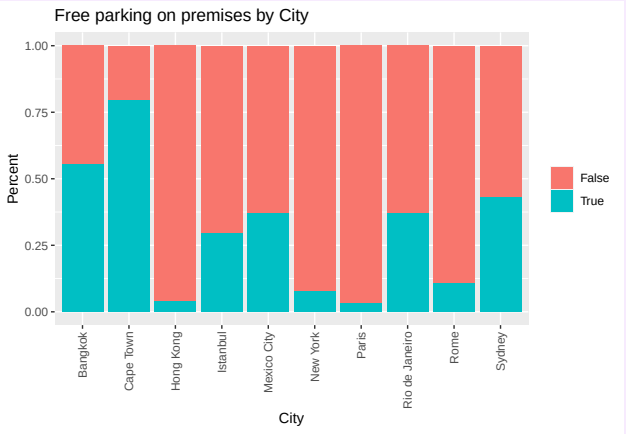

Shampoo by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True

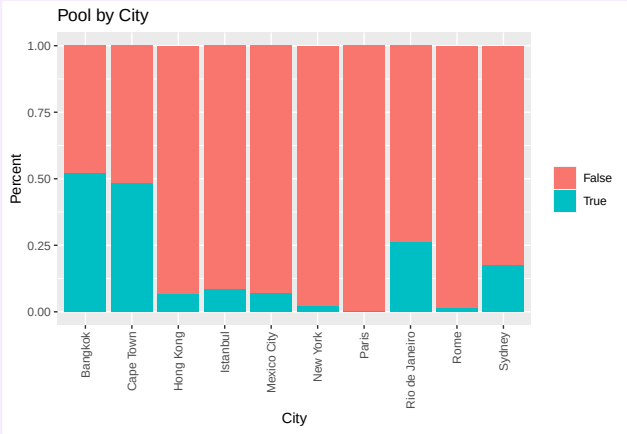

Hair dryer by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Smoke alarm by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Carbon monoxide alarm by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Air conditioning by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Free parking on premises by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Pool by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True